Chapter 2

Multiple Linear Regression

2.1 The MLR Model

Definition 2.1. The **response variable** is the variable that you want to predict. The **predictor variables** are the variables used to predict the response variable.

Notation. In this text the response variable will usually be denoted by Y and the p predictor variables will often be denoted by $x_1, ..., x_p$. The response variable is also called the dependent variable while the predictor variables are also called independent variables, explanatory variables or covariates. Often the predictor variables will be collected in a vector $\boldsymbol{x}$. Then $\boldsymbol{x}^T$ is the transpose of $\boldsymbol{x}$.

Definition 2.2. Regression is the study of the conditional distribution $Y|\boldsymbol{x}$ of the response variable Y given the vector of predictors $\boldsymbol{x} = (x_1, ..., x_p)^T$.

Definition 2.3. A **quantitative variable** takes on numerical values while a **qualitative variable** takes on categorical values.

Example 2.1. Archeologists and crime scene investigators sometimes want to predict the height of a person from partial skeletal remains. A model for prediction can be built from nearly complete skeletons or from living humans, depending on the population of interest (eg ancient Egyptians or modern US citizens). The response variable Y is *height* and the predictor variables might be $x_1 \equiv 1$, $x_2 =$ *femur length* and $x_3 =$ *ulna length*. The

heights of individuals with $x_2 = 200mm$ and $x_3 = 140mm$ should be shorter on average than the heights of individuals with $x_2 = 500mm$ and $x_3 = 350mm$. In this example Y, x_2 and x_3 are quantitative variables. If $x_4 =$ *gender* is a predictor variable, then gender (coded as male = 1 and female = 0) is qualitative.

Definition 2.4. Suppose that the response variable Y and at least one predictor variable x_i are quantitative. Then the **multiple linear regression (MLR) model** is

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{2.1}$$

for $i = 1, \ldots, n$. Here n is the *sample size* and the random variable e_i is the ith *error*. Suppressing the subscript i, the model is $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$.

In matrix notation, these n equations become

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{2.2}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. Equivalently,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}. \tag{2.3}$$

Often the first column of $\boldsymbol{X}$ is $X_1 = \mathbf{1}$, the $n \times 1$ vector of ones. The ith case $(\boldsymbol{x}_i^T, Y_i)$ corresponds to the ith row $\boldsymbol{x}_i^T$ of $\boldsymbol{X}$ and the ith element of $\boldsymbol{Y}$. In the MLR model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$, the Y and e are random variables, but we only have observed values Y_i and $\boldsymbol{x}_i$. If the e_i are iid (independent and identically distributed) with zero mean and variance σ^2, then regression is used to estimate the unknown parameters $\boldsymbol{\beta}$ and σ^2.

Definition 2.5. The **iid error MLR model** uses the assumption that the errors $e_1, ..., e_n$ are iid with $E(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2 < \infty$. Also assume that the errors are independent of the predictor variables $\boldsymbol{x}_i$. The predictor variables $\boldsymbol{x}_i$ are assumed to be fixed and measured without error. The cases $(\boldsymbol{x}_i^T, Y_i)$ are independent for $i = 1, ..., n$.

If the predictor variables are random variables, then the above MLR model is conditional on the observed values of the $\boldsymbol{x}_i$. That is, observe the $\boldsymbol{x}_i$ and then act as if the observed $\boldsymbol{x}_i$ are fixed.

Definition 2.6. The **iid symmetric error MLR model** has the same assumptions as the iid error MLR model but adds the assumption that the iid errors come from a symmetric distribution.

Definition 2.7. The *normal MLR model* or **Gaussian MLR model** has the same assumptions as the iid error MLR model but adds the assumption that the errors $e_1, ..., e_n$ are iid $N(0, \sigma^2)$ random variables. That is, the e_i are iid normal random variables with zero mean and variance σ^2.

The unknown coefficients for the above 3 models are usually estimated using (ordinary) least squares.

Notation. The symbol $A \equiv B = f(c)$ means that A and B are equivalent and equal, and that $f(c)$ is the formula used to compute A and B.

Definition 2.8. Given an estimate $\boldsymbol{b}$ of $\boldsymbol{\beta}$, the corresponding vector of *predicted* or *fitted values* is $\widehat{\boldsymbol{Y}} \equiv \widehat{\boldsymbol{Y}}(\boldsymbol{b}) = \boldsymbol{X}\boldsymbol{b}$. Thus the ith fitted value

$$\hat{Y}_i \equiv \hat{Y}_i(\boldsymbol{b}) = \boldsymbol{x}_i^T \boldsymbol{b} = x_{i,1} b_1 + \cdots + x_{i,p} b_p.$$

The vector of *residuals* is $\boldsymbol{r} \equiv \boldsymbol{r}(\boldsymbol{b}) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}(\boldsymbol{b})$. Thus ith residual $r_i \equiv r_i(\boldsymbol{b}) = Y_i - \hat{Y}_i(\boldsymbol{b}) = Y_i - x_{i,1} b_1 - \cdots - x_{i,p} b_p$.

Most regression methods attempt to find an estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ which minimizes some criterion function $Q(\boldsymbol{b})$ of the residuals.

Definition 2.9. The *ordinary least squares (OLS) estimator* $\hat{\boldsymbol{\beta}}_{OLS}$ minimizes

$$Q_{OLS}(\boldsymbol{b}) = \sum_{i=1}^{n} r_i^2(\boldsymbol{b}), \tag{2.4}$$

$$\text{and } \hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}.$$

The vector of *predicted* or *fitted values* $\widehat{\boldsymbol{Y}}_{OLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{H}\boldsymbol{Y}$ where the *hat matrix* $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T$ provided the inverse exists. Typically the subscript OLS is omitted, and the least squares *regression equation* is $\hat{Y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$ where $x_1 \equiv 1$ if the model contains a constant.

There are many statistical models besides the MLR model, and you should learn how to quickly recognize an MLR model. A "*regression*" model has a response variable Y and the conditional distribution of Y given the predictors $\boldsymbol{x} = (x_1, ..., x_p)^T$ is of interest. Regression models are used to predict Y and to summarize the relationship between Y and $\boldsymbol{x}$. If a constant $x_{i,1} \equiv 1$ (this notation means that $x_{i,1} = 1$ for $i = 1, ..., n$) is in the model, then $x_{i,1}$ is often called the trivial predictor, and the MLR model is said to have a constant or intercept. All nonconstant predictors are called nontrivial predictors. The term "*multiple*" is used if the model uses one or more nontrivial predictors. The simple linear regression model is a special case that uses exactly one nontrivial predictor. Suppose the response variable is Y and data has been collected on additional variables $x_1, ..., x_p$.

An MLR model is "*linear*" in the unknown coefficients $\boldsymbol{\beta}$. Thus the model is an MLR model in Y and $\boldsymbol{\beta}$ if we can write $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ or $Y_i = \boldsymbol{w}_i^T\boldsymbol{\beta} + e_i$ where each w_i is a function of $x_1, ..., x_p$. Symbols other than $\boldsymbol{w}$ or $\boldsymbol{x}$ may be used. Alternatively, the model is linear in the parameters $\boldsymbol{\beta}$ if $\partial Y/\partial\beta_i$ does not depend on the parameters. If $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e = x_1\beta_1 + \cdots + x_p\beta_p + e$, then $\partial Y/\partial\beta_i = x_i$. Similarly, if $Y = \boldsymbol{w}^T\boldsymbol{\beta} + e$, then $\partial Y/\partial\beta_i = w_i$.

Example 2.2. a) Suppose that interest is in predicting a function of Z from functions of $w_1, ..., w_k$. If $Y = t(Z) = \boldsymbol{x}^T\boldsymbol{\beta} + e$ where t is a function and each x_i is some function of $w_1, ..., w_k$, then there is an MLR model in Y and $\boldsymbol{\beta}$. Similarly, $Z = t(Y) = \boldsymbol{w}^T\boldsymbol{\beta} + e$ is an MLR model in Z and $\boldsymbol{\beta}$.

b) To see that $Y = \beta_1 + \beta_2 x + \beta_3 x^2 + e$ is an MLR model in Y and $\boldsymbol{\beta}$, take $w_1 = 1$, $w_2 = x$ and $w_3 = x^2$. Then $Y = \boldsymbol{w}^T\boldsymbol{\beta} + e$.

c) If $Y = \beta_1 + \beta_2 \exp(\beta_3 x) + e$, then the model is a nonlinear regression model that is not an MLR model in Y and $\boldsymbol{\beta}$. Notice that the model can not be written in the form $Y = \boldsymbol{w}^T\boldsymbol{\beta} + e$ and that $\partial Y/\partial\beta_2 = \exp(\beta_3 x)$ and $\partial Y/\partial\beta_3 = \beta_2 x \exp(\beta_3 x)$ depend on the parameters.

2.2 Checking Goodness of Fit

It is crucial to realize that an MLR model is not necessarily a useful model for the data, even if the data set consists of a response variable and several predictor variables. For example, a nonlinear regression model or a much more complicated model may be needed. Let p be the number of predictors and n the number of cases. Assume that $n > 5p$, then plots can

be used to check whether the MLR model is useful for studying the data. This technique is known as checking the goodness of fit of the MLR model.

Notation. Plots will be used to simplify regression analysis, and in this text a plot of W versus Z uses W on the horizontal axis and Z on the vertical axis.

Definition 2.10. A **scatterplot** of X versus Y is a plot of X versus Y and is used to **visualize the conditional distribution** $Y|X$ of Y given X.

Definition 2.11. A **response plot** is a plot of a variable w_i versus Y_i. Typically w_i is a linear combination of the predictors: $w_i = \boldsymbol{x}_i^T \boldsymbol{\eta}$ where $\boldsymbol{\eta}$ is a known $p \times 1$ vector. The most commonly used response plot is a plot of the fitted values $\widehat{Y}_i$ versus the response Y_i.

Proposition 2.1. Suppose that the regression estimator $\boldsymbol{b}$ of $\boldsymbol{\beta}$ is used to find the residuals $r_i \equiv r_i(\boldsymbol{b})$ and the fitted values $\widehat{Y}_i \equiv \widehat{Y}_i(\boldsymbol{b}) = \boldsymbol{x}_i^T \boldsymbol{b}$. Then in the response plot of $\widehat{Y}_i$ versus Y_i, the vertical deviations from the identity line (that has unit slope and zero intercept) are the residuals $r_i(\boldsymbol{b})$.

Proof. The identity line in the response plot is $Y = \boldsymbol{x}^T \boldsymbol{b}$. Hence the vertical deviation is $Y_i - \boldsymbol{x}_i^T \boldsymbol{b} = r_i(\boldsymbol{b})$. $\square$

Definition 2.12. A **residual plot** is a plot of a variable w_i versus the residuals r_i. The most commonly used residual plot is a plot of $\hat{Y}_i$ versus r_i.

Notation: For MLR, "the residual plot" will often mean the residual plot of $\hat{Y}_i$ versus r_i, and "the response plot" will often mean the plot of $\hat{Y}_i$ versus Y_i.

If the iid error MLR model as estimated by least squares is useful, then in the response plot the plotted points should scatter about the identity line while in the residual plot of $\hat{Y}$ versus r the plotted points should scatter about the $r = 0$ line (the horizontal axis) with no other pattern. Figures 1.2 and 1.3 show what a response plot and residual plot look like for an artificial MLR data set where the MLR regression relationship is rather strong in that the sample correlation $\text{corr}(\hat{Y}, Y)$ is near 1. Figure 1.4 shows a response plot where the response Y is independent of the nontrivial predictors in the model. Here $\text{corr}(\hat{Y}, Y)$ is near 0 but the points still scatter about the identity line. When the MLR relationship is very weak, the response plot will look like

Figure 1.4.

The above ideal shapes for the response and residual plots are for when the iid symmetric error MLR model gives a good approximation for the data. If the plots have the ideal shapes and $n \geq 5p$, then expect inference, except for prediction intervals, to be approximately correct.

If the response and residual plots suggest a MLR model with iid skewed errors, then add lowess to both plots. The scatterplot smoother tries to estimate the mean function $E(Y|\hat{Y})$ or $E(r|\hat{Y})$ without using any model. If the lowess curve is close to the identity line in the response plot and close to the $r = 0$ line in the residual plot, then the iid error MLR model may be a good approximation to the data, but sample sizes much larger than $n = 5p$ may be needed before inference is approximately correct. Such skewed data sets seem rather rare, but see Chen, Bengtsson and Ho (2009) and see Problem 2.27.

Remark 2.1. For any MLR analysis, always make the response plot and the residual plot of $\hat{Y}_i$ versus Y_i and r_i, respectively.

Definition 2.13. An **outlier** is an observation that lies far away from the bulk of the data.

Remark 2.2. For MLR, **the response plot is the single most important plot that can be made** because MLR is the study of the conditional distribution of $Y|\boldsymbol{x}^T\boldsymbol{\beta}$, and **the response plot is used to visualize the conditional distribution** of $Y|\boldsymbol{x}^T\boldsymbol{\beta}$ since $\hat{Y} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ is a good estimator of $\boldsymbol{x}^T\boldsymbol{\beta}$ if $\hat{\boldsymbol{\beta}}$ is a good estimator of $\boldsymbol{\beta}$.

If the MLR model is useful, then the plotted points in the response plot should be linear and scatter about the identity line with no gross outliers. Suppose the fitted values range in value from w_L to w_H with no outliers. Fix the fit $= w$ in this range and mentally add a narrow vertical strip centered at w to the response plot. The plotted points in the vertical strip should have a mean near w since they scatter about the identity line. Hence $Y|fit = w$ is like a sample from a distribution with mean w. The following example helps illustrate this remark.

Example 2.3. Tremearne (1911) presents a data set of about 17 measurements on 115 people of Hausa nationality. We deleted 3 cases (107, 108 and 109) because of missing values and used *height* as the response variable Y. Along with a constant $x_{i,1} \equiv 1$, the five additional predictor variables used

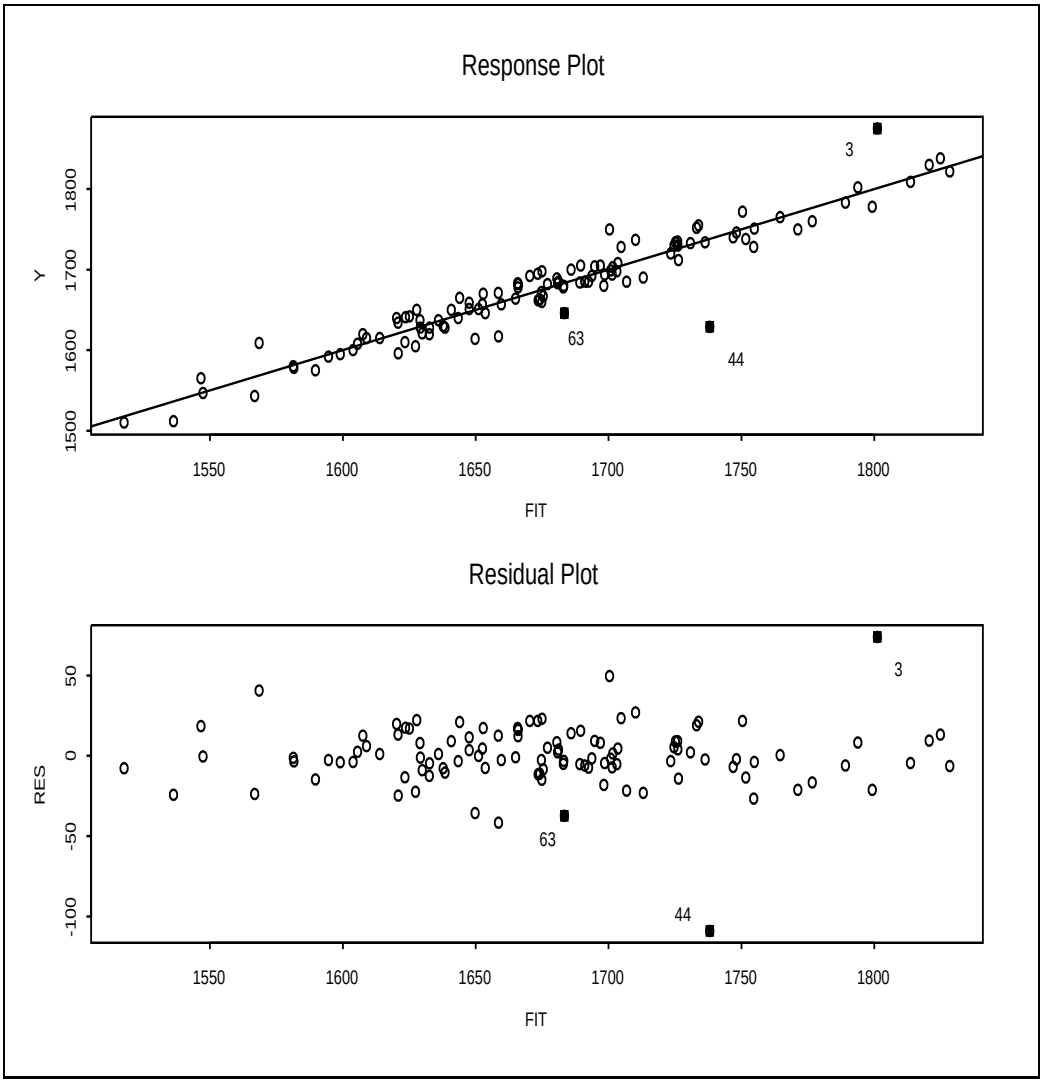

Figure 2.1: Residual and Response Plots for the Tremearne Data

were *height when sitting*, *height when kneeling*, *head length, nasal breadth*, and *span* (perhaps from left hand to right hand). Figure 2.1 presents the OLS response and residual plots for this data set. These plots show that an MLR model should be a useful model for the data since the plotted points in the response plot are linear and follow the identity line while the plotted points in the residual plot follow the $r = 0$ line with no other pattern (except for a possible outlier marked 44).

To use the response plot to visualize the conditional distribution of $Y|\boldsymbol{x}^T\boldsymbol{\beta}$, use the fact that the fitted values $\hat{Y} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$. For example, suppose the height given fit = 1700 is of interest. Mentally examine the plot about a narrow vertical strip about fit = 1700, perhaps from 1675 to 1725. The cases in the narrow strip have a mean close to 1700 since they fall close to the identity line. Similarly, when the fit $= w$ for w between 1500 and 1850, the cases have heights near w, on average.

Cases 3, 44 and 63 are highlighted. The 3rd person was very tall while the 44th person was rather short. Beginners often label too many points as outliers. Mentally draw a box about the bulk of the data ignoring any outliers. Double the width of the box (about the identity line for the response plot and about the horizontal line for the residual plot). Cases outside of this imaginary doubled box are potential outliers. Alternatively, visually estimate the standard deviation of the residuals in both plots. In the residual plot look for residuals that are more than 5 standard deviations from the $r = 0$ line. In Figure 2.1, the standard deviation of the residuals appears to be around 10. Hence cases 3 and 44 are certainly worth examining.

The identity line can also pass through or near an outlier or a cluster of outliers. Then the outliers will be in the upper right or lower left of the response plot, and there will be a large gap between the cluster of outliers and the bulk of the data. See Figure 3.14.

2.3 Checking Lack of Fit

The response plot may look good while the residual plot suggests that the iid error MLR model can be improved. Examining plots to find model violations is called checking for lack of fit. Again assume that $n > 5p$.

The iid error MLR model often provides a useful model for the data, but the following assumptions do need to be checked.
i) Is the MLR model appropriate?

ii) Are outliers present?
iii) Is the error variance constant or nonconstant? The constant variance assumption VAR$(e_i) \equiv \sigma^2$ is known as homoscedasticity. The nonconstant variance assumption VAR$(e_i) = \sigma_i^2$ is known as heteroscedasticity.
iv) Are any important predictors left out of the model?
v) Are the errors $e_1, ..., e_n$ iid?
vi) Are the errors e_i independent of the predictors $\boldsymbol{x}_i$?

Make the response plot and the residual plot to check i), ii) and iii). An MLR model is reasonable if the plots look like Figures 1.2, 1.3, 1.4 and 2.1. A response plot that looks like Figure 1.13 suggests that the model is not linear. If the plotted points in the residual plot do not scatter about the $r = 0$ line with no other pattern (ie if the cloud of points is not ellipsoidal or rectangular with zero slope), then the iid error MLR model is not sustained.

The ith residual r_i is an estimator of the ith error e_i. The constant variance assumption may have been violated if the variability of the point cloud in the residual plot depends on the value of $\hat{Y}$. Often the variability of the residuals increases as $\hat{Y}$ increases, resulting in a right opening megaphone shape. (Figure 4.1b has this shape.) Often the variability of the residuals decreases as $\hat{Y}$ increases, resulting in a left opening megaphone shape. Sometimes the variability decreases then increases again (like a stretched or compressed bone), and sometimes the variability increases then decreases again.

2.3.1 Residual Plots

Remark 2.3. Residual plots *magnify departures* from the model while the response plot emphasizes *how well the MLR model fits the data.*

Since the residuals $r_i = \hat{e}_i$ are estimators of the errors, the residual plot is used to visualize the conditional distribution $e|SP$ of the errors given the sufficient predictor SP $= \boldsymbol{x}^T\boldsymbol{\beta}$, where SP is estimated by $\widehat{Y} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$. For the iid error MLR model, there should not be any pattern in the residual plot: as a narrow vertical strip is moved from left to right, the behavior of the residuals within the strip should show little change.

Notation. A *rule of thumb* is a rule that often but not always works well in practice.

Rule of thumb 2.1. If the residual plot would look good after several

points have been deleted, and if these deleted points were not gross outliers (points far from the point cloud formed by the bulk of the data), then the residual plot is probably good. Beginners often find too many things wrong with a good model. For practice, use the computer to generate several MLR data sets, and make the response and residual plots for these data sets. This exercise will help show that the plots can have considerable variability even when the MLR model is good.

Rule of thumb 2.2. If the plotted points in the residual plot look like a left or right opening megaphone, the first model violation to check is the assumption of nonconstant variance. (This is a rule of thumb because it is possible that such a residual plot results from another model violation such as nonlinearity, but nonconstant variance is much more common.)

The residual plot of $\hat{Y}$ versus r should always be made. It is also a good idea to plot each nontrivial predictor x_j versus r and to plot potential predictors w_j versus r. If the predictor is quantitative, then the residual plot of x_j versus r should look like the residual plot of $\hat{Y}$ versus r. If the predictor is qualitative, eg gender, then interpreting the residual plot is much more difficult; however, if each category contains many observations, then the plotted points for each category should form a vertical line centered at $r = 0$ with roughly the same variability (spread or range).

Rule of thumb 2.3. Suppose that the MLR model uses predictors x_j and that data has been collected on variables w_j that are not included in the MLR model. To check whether important predictors have been left out, make residual plots of x_j and w_j versus r. If these plots scatter about the $r = 0$ line with no other pattern, then there is no evidence that x_j^2 or w_j are needed in the model. If the plotted points scatter about a parabolic curve, try adding x_j^2 or w_j and w_j^2 to the MLR model. If the plot of the potential predictor w_j versus r has a linear trend, try adding w_j to the MLR model.

Rule of thumb 2.4. To check that the errors are independent of the predictors, make residual plots of x_j versus r. If the plot of x_j versus r scatters about the $r = 0$ line with no other pattern, then there is no evidence that the errors depend on x_j. If the variability of the residuals changes with the value of x_j, eg if the plot resembles a left or right opening megaphone, the errors may depend on x_j. Some remedies for nonconstant variance are considered in Chapter 4.

To study residual plots, some notation and properties of the least squares estimator are needed. MLR is the study of the conditional distribution of $Y_i|\boldsymbol{x}_i^T\boldsymbol{\beta}$, and the MLR model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ is an $n \times p$ matrix of full rank p. Hence the number of predictors $p \leq n$. The ith row of $\boldsymbol{X}$ is $\boldsymbol{x}_i^T = (x_{i,1}, ..., x_{i,p})$ where $x_{i,k}$ is the value of the ith observation on the kth predictor x_k. We will denote the jth column of $\boldsymbol{X}$ by $X_j \equiv \boldsymbol{x}^j$ which corresponds to the jth variable or predictor x_j.

Example 2.4. If Y is *brain weight* in grams, $x_1 \equiv 1$, x_2 is *age* and x_3 is the *size* of the head in $(mm)^3$, then for the Gladstone (1905-6) data

$$\boldsymbol{Y} = \begin{bmatrix} 3738 \\ 4261 \\ \vdots \\ 3306 \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & 39 & 149.5 \\ 1 & 35 & 152.5 \\ \vdots & \vdots & \vdots \\ 1 & 19 & 141 \end{bmatrix}.$$

Hence the first person had *brain weight* $= 3738$, $age = 39$ and $size = 149.5$. After deleting observations with missing values, there were $n = 267$ cases (people measured on brain weight, age and size), and $\boldsymbol{x}_{267} = (1, 19, 141)^T$. The second predictor $x_2 = age$ corresponds to the 2nd column of $\boldsymbol{X}$ and is $X_2 = (39, 35, ..., 19)^T$. Notice that $X_1 \equiv \boldsymbol{x}^1 = \boldsymbol{1} = (1, ..., 1)^T$ corresponds to the constant x_1.

The results in the following proposition are properties of least squares (OLS), not of the underlying MLR model. Definitions 2.8 and 2.9 define the hat matrix $\boldsymbol{H}$, vector of fitted values $\hat{\boldsymbol{Y}}$ and vector of residuals $\boldsymbol{r}$. Parts f) and g) make residual plots useful. If the plotted points are linear with roughly constant variance and the correlation is zero, then the plotted points scatter about the $r = 0$ line with no other pattern. If the plotted points in a residual plot of w versus r do show a pattern such as a curve or a right opening megaphone, zero correlation will usually force symmetry about either the $r = 0$ line or the $w = \text{median}(w)$ line. Hence departures from the ideal plot of random scatter about the $r = 0$ line are often easy to detect.

Warning: If $n > p$, as is usually the case, $\boldsymbol{X}$ is not square, so $(\boldsymbol{X}^T\boldsymbol{X})^{-1} \neq \boldsymbol{X}^{-1}(\boldsymbol{X}^T)^{-1}$ since $\boldsymbol{X}^{-1}$ does not exist.

Proposition 2.2. Suppose that $\boldsymbol{X}$ is an $n \times p$ matrix of full rank p. Then

a) $\boldsymbol{H}$ is symmetric: $\boldsymbol{H} = \boldsymbol{H}^T$.

b) $\boldsymbol{H}$ is idempotent: $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$.

c) $\boldsymbol{X}^T\boldsymbol{r} = \boldsymbol{0}$ so that $X_j^T\boldsymbol{r} = (\boldsymbol{x}^j)^T\boldsymbol{r} = 0$.

d) If there is a constant $X_1 \equiv \boldsymbol{x}^1 = \boldsymbol{1}$ in the model, then the sum of the residuals is zero: $\sum_{i=1}^n r_i = 0$.

e) $\boldsymbol{r}^T\hat{\boldsymbol{Y}} = 0$.

f) If there is a constant in the model, then the sample correlation of the fitted values and the residuals is 0: $\text{corr}(\boldsymbol{r}, \hat{\boldsymbol{Y}}) = 0$.

g) If there is a constant in the model, then the sample correlation of the jth predictor with the residuals is 0: $\text{corr}(\boldsymbol{r}, \boldsymbol{x}^j) = 0$ for $j = 1, ..., p$.

Proof. a) $\boldsymbol{X}^T\boldsymbol{X}$ is symmetric since $(\boldsymbol{X}^T\boldsymbol{X})^T = \boldsymbol{X}^T(\boldsymbol{X}^T)^T = \boldsymbol{X}^T\boldsymbol{X}$. Hence $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is symmetric since the inverse of a symmetric matrix is symmetric. (Recall that if $\boldsymbol{A}$ has an inverse then $(\boldsymbol{A}^T)^{-1} = (\boldsymbol{A}^{-1})^T$.) Thus using $(\boldsymbol{A}^T)^T = \boldsymbol{A}$ and $(\boldsymbol{ABC})^T = \boldsymbol{C}^T\boldsymbol{B}^T\boldsymbol{A}^T$ shows that

$$\boldsymbol{H}^T = \boldsymbol{X}^T[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]^T(\boldsymbol{X}^T)^T = \boldsymbol{H}.$$

b) $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{H}$ since $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_p$, the $p \times p$ identity matrix.

c) $\boldsymbol{X}^T\boldsymbol{r} = \boldsymbol{X}^T(\boldsymbol{I}_p - \boldsymbol{H})\boldsymbol{Y} = [\boldsymbol{X}^T - \boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{Y} = [\boldsymbol{X}^T - \boldsymbol{X}^T]\boldsymbol{Y} = \boldsymbol{0}$. Since $\boldsymbol{x}^j$ is the jth column of $\boldsymbol{X}$, $(\boldsymbol{x}^j)^T$ is the jth row of $\boldsymbol{X}^T$ and $(\boldsymbol{x}^j)^T\boldsymbol{r} = 0$ for $j = 1, ..., p$.

d) Since $\boldsymbol{x}^1 = \boldsymbol{1}$, $(\boldsymbol{x}^1)^T\boldsymbol{r} = \sum_{i=1}^n r_i = 0$ by c).

e) $\boldsymbol{r}^T\hat{\boldsymbol{Y}} = [(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}]^T\boldsymbol{H}\boldsymbol{Y} = \boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{H}\boldsymbol{Y} = \boldsymbol{Y}^T(\boldsymbol{H} - \boldsymbol{H})\boldsymbol{Y} = 0$.

f) The sample correlation between W and Z is $\text{corr}(W, Z) =$

$$\frac{\sum_{i=1}^n (w_i - \overline{w})(z_i - \overline{z})}{(n-1)s_w s_z} = \frac{\sum_{i=1}^n (w_i - \overline{w})(z_i - \overline{z})}{\sqrt{\sum_{i=1}^n (w_i - \overline{w})^2 \sum_{i=1}^n (z_i - \overline{z})^2}}$$

where s_m is the sample standard deviation of m for $m = z, w$. So the result follows if $A = \sum_{i=1}^n (\hat{Y}_i - \overline{\hat{Y}})(r_i - \overline{r}) = 0$. Now $\overline{r} = 0$ by d), and thus

$$A = \sum_{i=1}^n \hat{Y}_i r_i - \overline{\hat{Y}} \sum_{i=1}^n r_i = \sum_{i=1}^n \hat{Y}_i r_i$$

by d) again. But $\sum_{i=1}^n \hat{Y}_i r_i = \boldsymbol{r}^T\hat{\boldsymbol{Y}} = 0$ by e).

g) Following the argument in f), the result follows if $A = \sum_{i=1}^{n}(x_{i,j} - \overline{x}_j)(r_i - \overline{r}) = 0$ where $\overline{x}_j$ is the mean of the jth predictor. Now $\overline{r} = 0$ by d), and thus

$$A = \sum_{i=1}^{n} x_{i,j} r_i - \overline{x}_j \sum_{i=1}^{n} r_i = \sum_{i=1}^{n} x_{i,j} r_i$$

by d) again. But $\sum_{i=1}^{n} x_{i,j} r_i = (\boldsymbol{x}^j)^T \boldsymbol{r} = 0$ by c). QED

2.3.2 Other Model Violations

Without loss of generality, $E(e) = 0$ for the iid error MLR model with a constant, in that if $E(\tilde{e}) = \mu \neq 0$, then the MLR model can always be written as $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ where $E(e) = 0$ and $E(Y) \equiv E(Y|\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$. To see this claim notice that

$$Y = \tilde{\beta}_1 + x_2\beta_2 + \cdots + x_p\beta_p + \tilde{e} = \tilde{\beta}_1 + E(\tilde{e}) + x_2\beta_2 + \cdots + x_p\beta_p + \tilde{e} - E(\tilde{e})$$

$$= \beta_1 + x_2\beta_2 + \cdots + x_p\beta_p + e$$

where $\beta_1 = \tilde{\beta}_1 + E(\tilde{e})$ and $e = \tilde{e} - E(\tilde{e})$. For example, if the errors $\tilde{e}_i$ are iid exponential (λ) with $E(\tilde{e}_i) = \lambda$, use $e_i = \tilde{e}_i - \lambda$.

For least squares, it is crucial that σ^2 exists. For example, if the e_i are iid Cauchy(0,1), than σ^2 does not exist and the least squares estimators tend to perform very poorly.

The performance of least squares is analogous to the performance of $\overline{Y}$. The sample mean $\overline{Y}$ is a very good estimator of the population mean μ if the Y_i are iid $N(\mu, \sigma^2)$ and $\overline{Y}$ is a good estimator of μ if the sample size is large and the Y_i are iid with mean μ and variance σ^2. This result follows from the cental limit theorem, but how "large is large" depends on the underlying distribution. The $n > 30$ rule tends to hold for distributions that are close to normal in that they take on many values and σ^2 is not huge. Errors distributions that are highly nonnormal with tiny σ^2 often need $n >> 30$. For example, if $Y_1, ..., Y_n$ are iid Gamma$(1/m, 1)$, then $n > 25m$ may be needed. Another example is distributions that take on one value with very high probability, eg a Poisson random variable with very small variance. Bimodal and multimodal distributions and highly skewed distributions with large variances also need larger n.

There are central limit type theorems for the least squares estimators that depend on the error distribution of the iid errors e_i. We always assume that the e_i are continuous random variables with a probability density function. Error distributions that are close to normal may give good results for moderate n if $n > 10p$ and $n - p > 30$ where p is the number of predictors. Error distributions that need large n for the CLT to apply for $\overline{e}$, will tend to need large n for the limit theorems for least squares to apply (to give good approximations).

Checking whether the errors are iid is often difficult. The iid assumption is often reasonable if measurements are taken on different objects, eg people. In industry often several measurements are taken on a batch of material. For example a batch of cement is mixed and then several small cylinders of concrete are made from the batch. Then the cylinders are tested for strength. Experience from such experiments suggests that objects (eg cylinders) from different batches are independent, but objects from the same batch are not independent.

One check on independence can also be made if the time order of the observations is known. Let $r_{[t]}$ be the residual where $[t]$ is the time order of the trial. Hence [1] was the 1st and $[n]$ was the last trial. Plot the time order t versus $r_{[t]}$ if the time order is known. Again, trends and outliers suggest that the model could be improved. A box shaped plot with no trend suggests that the MLR model is good. A plot similar to the Durbin Watson test plots $r_{[t-1]}$ versus $r_{[t]}$ for $t = 2, ..., n$. Linear trend suggests serial correlation while random scatter suggests that there is no lag 1 autocorrelation. As a rule of thumb, if the OLS slope b is computed for the plotted points, $b > 0.25$ gives some evidence that there is positive correlation between $r_{[t-1]}$ and $r_{[t]}$.

If it is assumed that the error distribution is symmetric, make a histogram of the residuals. Check whether the histogram is roughly symmetric or clearly skewed. If it is assumed that the errors e_i are iid $N(0, \sigma^2)$ again check whether the histogram is mound shaped with "short tails." A commonly used alternative is to make a normal probability plot of the residuals. Let $r_{(1)} < r_{(2)} < \cdots < r_{(n)}$ denote the residuals ordered from smallest to largest. Hence $r_{(1)}$ is the value of the smallest residual. The normal probability plot plots the $\tilde{e}_{(i)}$ versus $r_{(i)}$ where the $\tilde{e}_{(i)}$ are the expected values of the order statistics from a sample of size n from a $N(0, 1)$ distribution. (Often the $\tilde{e}_{(i)}$ are the standard normal percentiles that satisfy $P(Z \leq \tilde{e}_{(i)}) = (i - 0.5)/n$

where $Z \sim N(0,1)$.)

Rules of thumb: i) if the plotted points scatter about some straight line in the normal probability plot, then there is no evidence against the normal assumption. ii) if the plotted points have an "ess shape" (concave up then concave down) then the error distribution is symmetric with lighter tails than the normal distribution. iii) If the plot resembles a cubic function, then the error distribution is symmetric with heavier tails than the normal distribution. iv) If the plotted points look concave up (eg like x^2 where $x > 0$), then the error distribution is right skewed.

2.4 The ANOVA F TEST

After fitting least squares and checking the response and residual plot to see that an MLR model is reasonable, the next step is to check whether there is an MLR relationship between Y and the nontrivial predictors x_2, ..., x_p. If at least one of these predictors is useful, then the OLS fitted values $\hat{Y}_i$ should be used. If none of the nontrivial predictors is useful, then $\overline{Y}$ will give as good predictions as $\hat{Y}_i$. Here the *sample mean*

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{2.5}$$

In the definition below, SSE is the sum of squared residuals and a residual $r_i = \hat{e}_i =$ "errorhat." In the literature "errorhat" is often rather misleadingly abbreviated as "error."

Definition 2.14. Assume that a constant is in the MLR model.
a) The *total sum of squares*

$$SSTO = \sum_{i=1}^{n} (Y_i - \overline{Y})^2. \tag{2.6}$$

b) The *regression sum of squares*

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2. \tag{2.7}$$

c) The residual sum of squares or *error sum of squares* is

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} r_i^2. \tag{2.8}$$

The result in the following proposition is a property of least squares (OLS), not of the underlying MLR model. An obvious application is that given any two of SSTO, SSE and SSR, the 3rd sum of squares can be found using the formula $SSTO = SSE + SSR$.

Proposition 2.3. Assume that a constant is in the MLR model. Then $SSTO = SSE + SSR$.

Proof.

$$SSTO = \sum_{i=1}^{n} (Y_i - \hat{Y}_i + \hat{Y}_i - \overline{Y})^2 = SSE + SSR + 2\sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y}).$$

Hence the result follows if

$$A \equiv \sum_{i=1}^{n} r_i(\hat{Y}_i - \overline{Y}) = 0.$$

But

$$A = \sum_{i=1}^{n} r_i \hat{Y}_i - \overline{Y} \sum_{i=1}^{n} r_i = 0$$

by Proposition 2.2 d) and e). $\square$

Definition 2.15. Assume that a constant is in the MLR model and that SSTO $\neq 0$. The **coefficient of multiple determination**

$$R^2 = [\mathrm{corr}(\mathrm{Y_i}, \hat{\mathrm{Y}}_\mathrm{i})]^2 = \frac{\mathrm{SSR}}{\mathrm{SSTO}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SSTO}}$$

where $\mathrm{corr}(\mathrm{Y_i}, \hat{\mathrm{Y}}_\mathrm{i})$ is the sample correlation of Y_i and $\hat{Y}_i$.

Warnings: i) $0 \leq R^2 \leq 1$, but small R^2 does not imply that the MLR model is bad.

ii) If the MLR model contains a constant then there are several equivalent formulas for R^2. If the model does not contain a constant, then R^2 depends on the software package.

iii) R^2 does not have much meaning unless the response plot and residual plot both look good.

iv) R^2 tends to be too high if n is small.

v) R^2 tends to be too high if there are two or more separated clusters of data in the response plot.

vi) R^2 is too high if the number of predictors p is close to n.

vii) In large samples R^2 will be large (close to one) if σ^2 is small compared to the sample variance S_Y^2 of the response variable Y. R^2 is also large if the sample variance of $\hat{Y}$ is close to S_Y^2. Thus R^2 is sometimes interpreted as the proportion of the variability of Y explained by conditioning on $\boldsymbol{x}$, but warnings i) - v) suggest that R^2 may not have much meaning.

The following 2 propositions suggest that R^2 does not behave well when many predictors that are not needed in the model are included in the model. Such a variable is sometimes called a noise variable and the MLR model is "fitting noise." Proposition 2.5, appears, for example, in Cramér (1946, p. 414-415), and suggests that R^2 should be considerably larger than p/n if the predictors are useful.

Proposition 2.4. Assume that a constant is in the MLR model. Adding a variable to the MLR model does not decrease (and usually increases) R^2.

Proposition 2.5. Assume that a constant β_1 is in the MLR model, that $\beta_2 = \cdots = \beta_p = 0$ and that the e_i are iid $N(0, \sigma^2)$. Hence the Y_i are iid $N(\beta_1, \sigma^2)$. Then

a)R^2 follows a beta distribution: $R^2 \sim \text{beta}(\frac{\text{p}-1}{2}, \frac{\text{n}-\text{p}}{2})$.

b)

$$E(R^2) = \frac{p-1}{n-1}.$$

c)

$$\text{VAR}(R^2) = \frac{2(p-1)(n-p)}{(n-1)^2(n+1)}.$$

Notice that each SS/n estimates the variability of some quantity. $SSTO/n \approx S_Y^2$, $SSE/n \approx S_e^2$ and $SSR/n \approx S_{\hat{Y}}^2$.

Definition 2.16. Assume that a constant is in the MLR model. Associated with each SS in Definition 2.14 is a degrees of freedom (df) and a mean square $= SS/df$. For SSTO, df $= n-1$ and $MSTO = SSTO/(n-1)$. For SSR, df $= p-1$ and $MSR = SSR/(p-1)$. For SSE, $df = n-p$ and $MSE = SSE/(n-p)$.

Seber and Lee (2003, p. 44–47) show that when the MLR model holds, MSE is often a good estimator of σ^2. Under regularity conditions, the MSE is one of the best unbiased quadratic estimators of σ^2. For the normal MLR model, MSE is the uniformly minimum variance unbiased estimator of σ^2. Seber and Lee also give the following theorem that shows that the MSE is an unbiased estimator of σ^2 under very weak assumptions if the MLR model is appropriate.

Theorem 2.6. If $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ is an $n \times p$ matrix of full rank p, if the e_i are independent with $E(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2$, then $\hat{\sigma}^2 = MSE$ is an unbiased estimator of σ^2.

The ANOVA F test tests whether any of the nontrivial predictors $x_2, ..., x_p$ are needed in the OLS MLR model, that is, whether Y_i should be predicted by the OLS fit $\hat{Y}_i = \hat{\beta}_1 + x_{i,2}\hat{\beta}_2 + \cdots + x_{i,p}\hat{\beta}_p$ or with the sample mean $\overline{Y}$. ANOVA stands for analysis of variance, and the computer output needed to perform the test is contained in the ANOVA table. Below is an ANOVA table given in symbols. Sometimes "Regression" is replaced by "Model" and "Residual" by "Error."

Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Regression	p-1	SSR	MSR	Fo=MSR/MSE	for Ho:
Residual	n-p	SSE	MSE		$\beta_2 = \cdots = \beta_p = 0$

Remark 2.4. Recall that for a 4 step test of hypotheses, the p–value is the probability of getting a test statistic as extreme as the test statistic actually observed and that Ho is rejected if the p–value $< \delta$. As a benchmark for this textbook, use $\delta = 0.05$ if δ is not given. The 4th step is the nontechnical conclusion which is crucial for presenting your results to people who are not familiar with MLR. Replace Y and $x_2, ..., x_p$ by the actual variables used in the MLR model. Follow Example 2.5.

Notation. The p–value $\equiv$ pvalue given by output tends to only be correct for the normal MLR model. Hence the output is usually only giving an estimate of the pvalue, which will often be denoted by *pval*. Often

$$\text{pval} - \text{pvalue} \xrightarrow{\text{P}} 0$$

(converges to 0 in probability) as the sample size $n \to \infty$. Then the computer

output pval is a good estimator of the unknown pvalue.

Be able to perform the 4 step ANOVA F test of hypotheses:
i) State the hypotheses Ho: $\beta_2 = \cdots = \beta_p = 0$ Ha: not Ho
ii) Find the test statistic $F_o = MSR/MSE$ or obtain it from output.
iii) Find the p–value from output or use the F–table: p–value =

$$P(F_{p-1,n-p} > F_o).$$

iv) State whether you reject Ho or fail to reject Ho. If Ho is rejected, conclude that there is an MLR relationship between Y and the predictors x_2, ..., x_p. If you fail to reject Ho, conclude that there is not a MLR relationship between Y and the predictors x_2, ..., x_p.

Example 2.5. For the Gladstone (1905-6) data, the response variable Y = *brain weight*, $x_1 \equiv 1$, x_2 = *size* of head, x_3 = *sex*, x_4 = *breadth* of head, x_5 = *circumference* of head. Assume that the response and residual plots look good and test whether at least one of the nontrivial predictors is needed in the model using the output shown below.

```
Summary Analysis of Variance Table
Source          df       SS            MS           F     p-value
Regression       4  5396942.      1349235.     196.24     0.0000
Residual       262  1801333.       6875.32
```

Solution: i) Ho: $\beta_2 = \cdots = \beta_5 = 0$ Ha: not Ho
ii) $F_o = 196.24$ from output.
iii) p–value = 0.0 from output.
iv) The p–value $< \delta$ (= 0.05 since δ was not given). So reject Ho. Hence there is an MLR relationship between brain weight and the predictors size, sex, breadth, and circumference.

Remark 2.5. There is a close relationship between the response plot and the ANOVA F test. If $n > 10p$ and $n - p > 30$ and if the plotted points follow the identity line, typically Ho will be rejected if the identity line fits the plotted points better than any horizontal line (in particular, the line $Y = \overline{Y}$). If a horizontal line fits the plotted points about as well as the identity line, as in Figure 1.4, this graphical diagnostic is inconclusive (sometimes the ANOVA F test will reject Ho and sometimes fail to reject Ho), but the MLR relationship is at best weak. In Figures 1.2 and 2.1, the

ANOVA F test should reject Ho since the identity line fits the plotted points better than any horizontal line.

Definition 2.17. An **RR plot** is a plot of residuals from 2 different models or fitting methods.

Remark 2.6. If the RR plot of the residuals $Y_i - \overline{Y}$ versus the OLS residuals $r_i = Y_i - \hat{Y}_i$ shows tight clustering about the identity line, then the MLR relationship is weak: $\overline{Y}$ fits the data about as well as the OLS fit.

Example 2.6. Cook and Weisberg (1999a, p. 261, 371) describe a data set where rats were injected with a dose of a drug approximately proportional to body weight. The response Y is the fraction of the drug recovered from the rat's liver. The three predictors are the *body weight* of the rat, the *dose* of the drug, and the *liver weight*. A constant was also used. The experimenter expected the response to be independent of the predictors, and 19 cases were used. However, the ANOVA F test suggested that the predictors were important. The third case was an outlier and easily detected in the response and residual plots (not shown). After deleting the outlier, the response and residual plots looked ok and the following output was obtained.

```
Summary Analysis of Variance Table
Source          df       SS            MS           F     p-value
Regression       3  0.00184396   0.000614652    0.10     0.9585
Residual        14  0.0857172    0.00612265
```

The 4 step ANOVA F test is
i) Ho: $\beta_2 = \cdots = \beta_4 = 0$ Ha: not Ho
ii) $F_o = 0.10$.
iii) p–value = 0.9585.
iv) The p–value $> \delta$ (= 0.05 since δ was not given). So fail to reject Ho. Hence there is not an MLR relationship between fraction of drug recovered and the predictors body weight, dose, and liver weight. (More accurately, there is not enough statistical evidence to conclude that there is an MLR relationship: failing to reject Ho is not the same as accepting Ho; however, it may be a good idea to keep the nontechnical conclusions nontechnical.)

Figure 2.2 shows the RR plot where the residuals from the full model are plotted against $Y_i - \overline{Y}$, the residuals from the model using no nontrivial predictors. This plot reinforces the conclusion that the response Y is independent of the nontrivial predictors. The identity line and the OLS line from

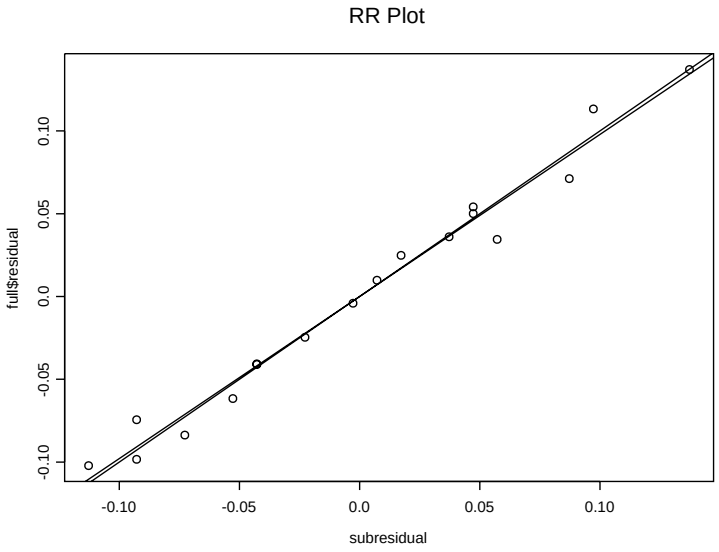

Figure 2.2: RR Plot With Outlier Deleted, Submodel Uses No Predictors with $\hat{Y} = \overline{Y}$

regressing r_i on $Y_i - \overline{Y}$ (that is, use $\tilde{Y}_i = r_i$, a constant and $\tilde{x}_{i,2} = Y_i - \overline{Y}$, find the OLS line and then plot it) are shown as visual aids. If the OLS line and identity line nearly coincide in that it is difficult to tell that the two lines intersect at the origin, then the 2 sets of residuals are "close."

Some assumptions are needed on the ANOVA F test. Assume that both the response and residual plots look good. It is crucial that there are no outliers. Then a rule of thumb is that if $n - p$ is large, then the ANOVA F test p–value is approximately correct. An analogy can be made with the central limit theorem, $\overline{Y}$ is a good estimator for μ if the Y_i are iid $N(\mu, \sigma^2)$ and also a good estimator for μ if the data are iid with mean μ and variance σ^2 if n is large enough. More on the robustness and lack of robustness of the ANOVA F test can be found in Wilcox (2005).

If all of the $\boldsymbol{x}_i$ are different (no replication) and if the number of predictors $p = n$, then the OLS fit $\hat{Y}_i = Y_i$ and $R^2 = 1$. Notice that Ho is rejected if the statistic F_o is large. More precisely, reject Ho if

$$F_o > F_{p-1,n-p,1-\delta}$$

where

$$P(F \leq F_{p-1,n-p,1-\delta}) = 1 - \delta$$

when $F \sim F_{p-1,n-p}$. Since R^2 increases to 1 while $(n-p)/(p-1)$ decreases to 0 as p increases to n, Theorem 2.7a below implies that if p is large then the F_o statistic may be small even if some of the predictors are very good. It is a good idea to use $n > 10p$ or at least $n > 5p$ if possible.

Theorem 2.7. Assume that the MLR model has a constant β_1.
a)

$$F_o = \frac{MSR}{MSE} = \frac{R^2}{1-R^2} \quad \frac{n-p}{p-1}.$$

b) If the errors e_i are iid $N(0, \sigma^2)$, and if Ho: $\beta_2 = \cdots = \beta_p = 0$ is true, then F_o has an F distribution with $p-1$ numerator and $n-p$ denominator degrees of freedom: $F_o \sim F_{p-1,n-p}$.

c) If the errors are iid with mean 0 and variance σ^2, if the error distribution is close to normal and if $n-p$ is large enough, and if Ho is true, then $F_o \approx F_{p-1,n-p}$ in that the p-value is approximately correct.

Remark 2.7. When a constant is not contained in the model (ie $x_{i,1}$ is not equal to 1 for all i), then the computer output still produces an ANOVA table with the test statistic and p–value, and nearly the same 4 step test of hypotheses can be used. The hypotheses are now Ho: $\beta_1 = \cdots = \beta_p = 0$ Ha: not Ho, and you are testing whether or not there is an MLR relationship between Y and $x_1, ..., x_p$. An MLR model without a constant (no intercept) is sometimes called a "regression through the origin." See Section 2.10.

2.5 Prediction

This section gives estimators for predicting a future or new value Y_f of the response variable given the predictors $\boldsymbol{x}_f$, and for estimating the mean $E(Y_f) \equiv E(Y_f|\boldsymbol{x}_f)$. This mean is conditional on the values of the predictors $\boldsymbol{x}_f$, but the conditioning is often suppressed.

Warning: All too often the MLR model seems to fit the data

$$(Y_1, \boldsymbol{x}_1), ..., (Y_n, \boldsymbol{x}_n)$$

well, but when new data is collected, a very different MLR model is needed to fit the new data well. In particular, the MLR model seems to fit the data

$(Y_i, \boldsymbol{x}_i)$ well for $i = 1, ..., n$, but when the researcher tries to predict Y_f for a new vector of predictors $\boldsymbol{x}_f$, the prediction is very poor in that $\hat{Y}_f$ is not close to the Y_f actually observed. **Wait until after the MLR model has been shown to make good predictions before claiming that the model gives good predictions!**

There are several reasons why the MLR model may not fit new data well. i) The model building process is usually iterative. Data Z, $w_1, ..., w_r$ is collected. If the model is not linear, then functions of Z are used as a potential response and functions of the w_i as potential predictors. After trial and error, the functions are chosen, resulting in a final MLR model using Y and $x_1, ..., x_p$. Since the same data set was used during the model building process, biases are introduced and the MLR model fits the "training data" better than it fits new data. Suppose that Y, $x_1, ..., x_p$ are specified before collecting data and that the residual and response plots from the resulting MLR model look good. Then predictions from the prespecified model will often be better for predicting new data than a model built from an iterative process.

ii) If $(Y_f, \boldsymbol{x}_f)$ come from a different population than the population of $(Y_1, \boldsymbol{x}_1), ..., (Y_n, \boldsymbol{x}_n)$, then prediction for Y_f can be arbitrarily bad.

iii) Even a good MLR model may not provide good predictions for an $\boldsymbol{x}_f$ that is far from the $\boldsymbol{x}_i$ (extrapolation).

iv) The MLR model may be missing important predictors (underfitting).

v) The MLR model may contain unnecessary predictors (overfitting).

Two remedies for i) are a) use previously published studies to select an MLR model before gathering data. b) Do a trial study. Collect some data, build an MLR model using the iterative process. Then use this model as the prespecified model and collect data for the main part of the study. Better yet, do a trial study, specify a model, collect more trial data, improve the specified model and repeat until the latest specified model works well. Unfortunately, trial studies are often too expensive or not possible because the data is difficult to collect. Also often the population from a published study is quite different from the population of the data collected by the researcher. Then the MLR model from the published study is not adequate. If the data set is large enough, using a random sample of $< n/4$ of the cases to build a model may help reduce biases.

Definition 2.18. Consider the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ and the hat

matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Let $h_i = h_{ii}$ be the ith diagonal element of $\boldsymbol{H}$ for $i = 1, ..., n$. Then h_i is called the ith **leverage** and $h_i = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$. Suppose new data is to be collected with predictor vector $\boldsymbol{x}_f$. Then the leverage of $\boldsymbol{x}_f$ is $h_f = \boldsymbol{x}_f^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_f$. **Extrapolation** occurs if $\boldsymbol{x}_f$ is far from the $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$.

Rule of thumb 2.5. Predictions based on extrapolation are not reliable. A rule of thumb is that extrapolation occurs if $h_f > \max(h_1, ..., h_n)$. This rule works best if the predictors are linearly related in that a plot of x_i versus x_j should not have any strong nonlinearities. If there are strong nonlinearities among the predictors, then $\boldsymbol{x}_f$ could be far from the $\boldsymbol{x}_i$ but still have $h_f < \max(h_1, ..., h_n)$.

Example 2.7. Consider predicting $Y = \textit{weight}$ from $x = \textit{height}$ and a constant from data collected on men between 18 and 24 where the minimum height was 57 and the maximum height was 79 inches. The OLS equation was $\hat{Y} = -167 + 4.7x$. If $x = 70$ then $\hat{Y} = -167 + 4.7(70) = 162$ pounds. If $x = 1$ inch, then $\hat{Y} = -167 + 4.7(1) = -162.3$ pounds. It is impossible to have negative weight, but it is also impossible to find a 1 inch man. This MLR model should not be used for x far from the interval $(57, 79)$.

Definition 2.19. Consider the iid error MLR model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ where $E(e) = 0$. Then **regression function** is the hyperplane

$$E(Y) \equiv E(Y|\boldsymbol{x}) = x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p = \boldsymbol{x}^T\boldsymbol{\beta}. \tag{2.9}$$

Assume OLS is used to find $\hat{\boldsymbol{\beta}}$. Then the **point estimator** of Y_f given $\boldsymbol{x} = \boldsymbol{x}_f$ is

$$\hat{Y}_f = x_{f,1}\hat{\beta}_1 + \cdots + x_{f,p}\hat{\beta}_p = \boldsymbol{x}_f^T\hat{\boldsymbol{\beta}}. \tag{2.10}$$

The **point estimator** of $E(Y_f) \equiv E(Y_f|\boldsymbol{x}_f)$ given $\boldsymbol{x} = \boldsymbol{x}_f$ is also $\hat{Y}_f = \boldsymbol{x}_f^T\hat{\boldsymbol{\beta}}$. Assume that the MLR model contains a constant β_1 so that $x_1 \equiv 1$. The large sample 100 $(1 - \delta)\%$ confidence interval (CI) for $\mathrm{E}(Y_f|\boldsymbol{x}_f) = \boldsymbol{x}_f^T\boldsymbol{\beta} = E(\hat{Y}_f)$ is

$$\hat{Y}_f \pm t_{n-p,1-\delta/2}se(\hat{Y}_f) \tag{2.11}$$

where $P(T \leq t_{n-p,\delta}) = \delta$ if T has a t distribution with $n - p$ degrees of freedom. Generally $se(\hat{Y}_f)$ will come from output, but

$$se(\hat{Y}_f) = \sqrt{MSE\ h_f} = \sqrt{MSE\ \boldsymbol{x}_f^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_f}.$$

Recall the interpretation of a 100 $(1-\delta)$% CI for a parameter μ is that if you collect data then form the CI, and repeat for a total of k times where the k trials are independent from the same population, then the probability that m of the CIs will contain μ follows a binomial$(k, \rho = 1-\delta)$ distribution. Hence if 100 95% CIs are made, $\rho = 0.95$ and about 95 of the CIs will contain μ while about 5 will not. Any given CI may (good sample) or may not (bad sample) contain μ, but the probability of a "bad sample" is δ.

The following theorem is analogous to the central limit theorem and the theory for the t–interval for μ based on $\overline{Y}$ and the sample standard deviation (SD) S_Y. If the data $Y_1, ..., Y_n$ are iid with mean 0 and variance σ^2, then $\overline{Y}$ is asymptotically normal and the t–interval will perform well if the sample size is large enough. The result below suggests that the OLS estimators $\hat{Y}_i$ and $\hat{\boldsymbol{\beta}}$ are good if the sample size is large enough. The condition $\max h_i \to 0$ in probability usually holds if the researcher picked the design matrix $\boldsymbol{X}$ or if the $\boldsymbol{x}_i$ are iid random vectors from a well behaved population. Outliers can cause the condition to fail. Convergence in probability, $Y_n \xrightarrow{P} c$, is similar to other types of convergence: Y_n is likely to be close to c if the sample size n is large enough.

Theorem 2.8: Huber (1981, p. 157-160). Consider the MLR model $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ and assume that the errors are independent with zero mean and the same variance: $E(e_i) = 0$ and VAR$(e_i) = \sigma^2$. Also assume that $\max_i(h_1, ..., h_n) \to 0$ in probability as $n \to \infty$. Then

a) $\hat{Y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} \to E(Y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta}$ in probability for $i = 1, ..., n$ as $n \to \infty$.

b) All of the least squares estimators $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ are asymptotically normal where $\boldsymbol{a}$ is any fixed constant $p \times 1$ vector.

Definition 2.20. A large sample $100(1-\delta)$% *prediction interval* (PI) has the form $(\hat{L}_n, \hat{U}_n)$ where $P(\hat{L}_n < Y_f < \hat{U}_n) \xrightarrow{P} 1-\delta$ as the sample size $n \to \infty$. For the Gaussian MLR model, assume that the random variable Y_f is independent of $Y_1, ..., Y_n$. Then the 100 $(1-\delta)$% PI for Y_f is

$$\hat{Y}_f \pm t_{n-p,1-\delta/2}se(pred) \tag{2.12}$$

where $P(T \leq t_{n-p,\delta}) = \delta$ if T has a t distribution with $n-p$ degrees of freedom. Generally $se(pred)$ will come from output, but

$$se(pred) = \sqrt{MSE\ (1+h_f)}.$$

The interpretation of a 100 $(1-\delta)$% PI for a random variable Y_f is similar to that of a CI. Collect data, then form the PI, and repeat for a total of k times where k trials are independent from the same population. If Y_{fi} is the ith random variable and PI_i is the ith PI, then the probability that $Y_{fi} \in PI_i$ for m of the PIs follows a binomial$(k, \rho = 1 - \delta)$ distribution. Hence if 100 95% PIs are made, $\rho = 0.95$ and $Y_{fi} \in PI_i$ happens about 95 times.

There are two big differences between CIs and PIs. First, the length of the CI goes to 0 as the sample size n goes to ∞ while the length of the PI converges to some nonzero number J, say. Secondly, the CI for $E(Y_f|\boldsymbol{x}_f)$ given in Definition 2.19 tends to work well for the iid error MLR model if the sample size is large while the PI in Definition 2.20 is made under the assumption that the e_i are iid $N(0, \sigma^2)$ and may not perform well if the normality assumption is violated.

To see this, consider $\boldsymbol{x}_f$ such that the heights Y of women between 18 and 24 is normal with a mean of 66 inches and an SD of 3 inches. A 95% CI for $E(Y|\boldsymbol{x}_f)$ should be centered at about 66 and the length should go to zero as n gets large. But a 95% PI needs to contain about 95% of the heights so the PI should converge to the interval $66 \pm 1.96(3))$. This result follows because if $Y \sim N(66, 9)$ then $P(Y < 66 - 1.96(3)) = P(Y > 66 + 1.96(3)) \approx 0.025$. In other words, the endpoints of the PI estimate the 97.5 and 2.5 percentiles of the normal distribution. However, the percentiles of a parametric error distribution depend heavily on the parametric distribution and the parametric formulas are violated if the assumed error distribution is incorrect.

Assume that the iid error MLR model is valid so that e is from some distribution with 0 mean and variance σ^2. Olive (2007) shows that if $1-\gamma$ is the asymptotic coverage of the classical nominal $(1-\delta)100\%$ PI (2.12), then

$$1 - \gamma = P(-\sigma z_{1-\delta/2} < e < \sigma z_{1-\delta/2}) \geq 1 - \frac{1}{z^2_{1-\delta/2}} \tag{2.13}$$

where the inequality follows from Chebyshev's inequality. Hence the asymptotic coverage of the nominal 95% PI is at least 73.9%. The 95% PI (2.12) was often quite accurate in that the asymptotic coverage was close to 95% for a wide variety of error distributions. The 99% and 90% PIs did not perform as well.

Let ξ_δ be the δ percentile of the error e, ie, $P(e \leq \xi_\delta) = \delta$. Let $\hat{\xi}_\delta$ be the sample δ percentile of the residuals. Then the results from Theorem 2.8 suggest that the residuals r_i estimate the errors e_i, and that the sample percentiles of the residuals $\hat{\xi}_\delta$ estimate ξ_δ. For many error distributions,

$$E(MSE) = E\left(\sum_{i=1}^{n} \frac{r_i^2}{n-p}\right) = \sigma^2 = E\left(\sum_{i=1}^{n} \frac{e_i^2}{n}\right).$$

This result suggests that

$$\sqrt{\frac{n}{n-p}}\ \ r_i \approx e_i.$$

Using

$$a_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n}{n-p}}\sqrt{(1+h_f)}, \tag{2.14}$$

a large sample semiparametric $100(1-\delta)\%$ PI for Y_f is

$$(\hat{Y}_f + a_n\hat{\xi}_{\delta/2}, \hat{Y}_f + a_n\hat{\xi}_{1-\delta/2}). \tag{2.15}$$

This PI is very similar to the classical PI except that $\hat{\xi}_\delta$ is used instead of σz_δ to estimate the error percentiles ξ_δ. The large sample coverage $1-\gamma$ of this nominal $100(1-\delta)\%$ PI is asymptotically correct: $1-\gamma = 1-\delta$.

Example 2.8. For the Buxton (1920) data suppose that the response Y = *height* and the predictors were a constant, *head length*, *nasal height*, *bigonal breadth* and *cephalic index*. Five outliers were deleted leaving 82 cases. Figure 2.3 shows a response plot of the fitted values versus the response Y with the identity line added as a visual aid. The plot suggests that the model is good since the plotted points scatter about the identity line in an evenly populated band although the relationship is rather weak since the correlation of the plotted points is not very high. The triangles represent the upper and lower limits of the semiparametric 95% PI (2.15). For this example, 79 (or 96%) of the Y_i fell within their corresponding PI while 3 Y_i did not. A plot using the classical PI (2.12) would be very similar for this data.

Given output showing $\hat{\boldsymbol{\beta}}_i$ and given $\boldsymbol{x}_f$, $se(pred)$ and $se(\hat{Y}_f)$, Example 2.9 shows how to find $\hat{Y}_f$, a CI for $E(Y_f|\boldsymbol{x}_f)$ and a PI for Y_f. Below is shown typical output in symbols. Sometimes "Label" is replaced by "Predictor" and "Estimate" by "coef" or "Coefficients."

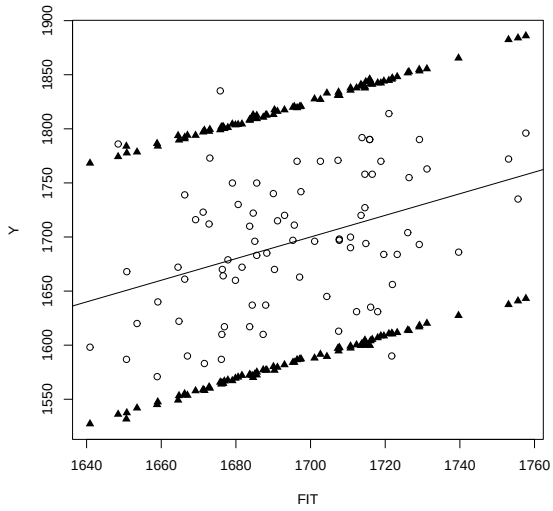

Figure 2.3: 95% PI Limits for Buxton Data

Label	Estimate	Std. Error	t-value	p-value
Constant	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$t_{o,1}$	for Ho: $\beta_1 = 0$
x_2	$\hat{\beta}_2$	$se(\hat{\beta}_2)$	$t_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$	for Ho: $\beta_2 = 0$
$\vdots$				
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$t_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	for Ho: $\beta_p = 0$

Example 2.9. The Rouncefield (1995) data are female and male life expectancies from $n = 91$ countries. Suppose that it is desired to predict female life expectancy Y from male life expectancy X. Suppose that if $X_f = 60$, then $se(\text{pred}) = 2.1285$, and $se(\hat{Y}_f) = 0.2241$. Below is some output.

```
Label      Estimate       Std. Error    t-value    p-value
Constant  -2.93739        1.42523        -2.061     0.0422
mlife      1.12359        0.0229362      48.988     0.0000
```

a) Find $\hat{Y}_f$ if $X_f = 60$.

Solution: In this example, $\boldsymbol{x}_f = (1, X_f)^T$ since a constant is in the output above. Thus $\hat{Y}_f = \hat{\beta}_1 + \hat{\beta}_2 X_f = -2.93739 + 1.12359(60) = 64.478$.

b) If $X_f = 60$, find a 90% confidence interval for $E(Y) \equiv E(Y_f|\boldsymbol{x}_f)$.

Solution: The CI is $\hat{Y}_f \pm t_{n-2,1-\delta/2} se(\hat{Y}_f) = 64.478 \pm 1.645(0.2241) = 64.478 \pm 0.3686 = (64.1094, 64.8466)$. To use the t–table on the last page of Chapter 17, use the 2nd to last row marked by Z since $d = df = n - 2 = 89 > 30$. In the last row find CI = 90% and intersect the 90% column and the Z row to get the value of $t_{89,0.95} \approx z_{.95} = 1.645$.

c) If $X_f = 60$, find a 90% prediction interval for Y_f.

Solution: The PI is $\hat{Y}_f \pm t_{n-2,1-\delta/2} se(pred) = 64.478 \pm 1.645(2.1285) = 64.478 \pm 3.5014 = (60.9766, 67.9794)$.

2.6 The Partial F or Change in SS TEST

Suppose that there is data on variables Z, $w_1, ..., w_r$ and that a useful MLR model has been made using $Y = t(Z)$, $x_1 \equiv 1$, $x_2, ..., x_p$ where each x_i is some function of $w_1, ..., w_r$. This useful model will be called the full model. It is important to realize that the full model does not need to use every variable w_j that was collected. For example, variables with outliers or missing values may not be used. Forming a useful full model is often very difficult, and it is often not reasonable to assume that the candidate full model is good based on a single data set, especially if the model is to be used for prediction.

Even if the full model is useful, the investigator will often be interested in checking whether a model that uses fewer predictors will work just as well. For example, perhaps x_p is a very expensive predictor but is not needed given that $x_1, ..., x_{p-1}$ are in the model. Also a model with fewer predictors tends to be easier to understand.

Definition 2.21. Let the **full model** use Y, $x_1 \equiv 1$, $x_2, ..., x_p$ and let the **reduced model** use Y, x_1, $x_{i_2}, ..., x_{i_q}$ where $\{i_2, ..., i_q\} \subset \{2, ..., p\}$.

The change in SS F test or partial F test is used to test whether the reduced model is good in that it can be used instead of the full model. It is crucial that the reduced model be selected before looking at the data. If the reduced model is selected after looking at output and discarding the worst variables, then the p–value for the partial F test will be too high. For (ordinary) least squares, usually a constant is used, and we are assuming that both the full model and the reduced model contain a constant. The partial F test has null hypothesis $Ho : \beta_{i_{q+1}} = \cdots = \beta_{i_p} = 0$, and alternative

hypothesis H_A : at least one of the $\beta_{i_j} \neq 0$ for $j > q$. The null hypothesis is equivalent to Ho: "the reduced model is good." Since only the full model and reduced model are being compared, the alternative hypothesis is equivalent to H_A: "the reduced model is not as good as the full model, so use the full model," or more simply, H_A : "use the full model."

To perform the change in SS or partial F test, fit the full model and the reduced model and obtain the ANOVA table for each model. The quantities df_F, SSE(F) and MSE(F) are for the full model and the corresponding quantities from the reduced model use an R instead of an F. Hence SSE(F) and SSE(R) are the residual sums of squares for the full and reduced models, respectively. Shown below is output only using symbols.

Full model

Source df	SS	MS	Fo and p-value
Regression $p-1$	SSR	MSR	Fo=MSR/MSE
Residual $df_F = n-p$	SSE(F)	MSE(F)	for Ho:$\beta_2 = \cdots = \beta_p = 0$

Reduced model

Source df	SS	MS	Fo and p-value
Regression $q-1$	SSR	MSR	Fo=MSR/MSE
Residual $df_R = n-q$	SSE(R)	MSE(R)	for Ho: $\beta_2 = \cdots = \beta_q = 0$

Be able to perform the 4 step change in SS F test = partial F test of hypotheses: i) State the hypotheses. Ho: the reduced model is good Ha: use the full model
ii) Find the test statistic. $F_R =$

$$\left[\frac{SSE(R) - SSE(F)}{df_R - df_F}\right] / MSE(F)$$

iii) Find the p–value = P($F_{df_R - df_F, df_F} > F_R$). (On exams typically an F table is used. Here $df_R - df_F = p - q =$ number of parameters set to 0, and $df_F = n - p$).
iv) State whether you reject Ho or fail to reject Ho. Reject Ho if the p–value $< \delta$ and conclude that the full model should be used. Otherwise, fail to reject Ho and conclude that the reduced model is good.

Sometime software has a shortcut. For example the *R/Splus* software uses the `anova` command. As an example, assume that the full model uses x_2 and x_3 while the reduced models uses x_2. Both models contain a constant. Then the following commands will perform the partial F test. (On the computer screen the 1st command looks more like red <- lm(y~x1).)

```
full <- lm(y~x2+x3)
red <- lm(y~x2)
anova(red,full)
```

For an $n \times 1$ vector $\boldsymbol{a}$, let

$$\|\boldsymbol{a}\| = \sqrt{a_1^2 + \cdots + a_n^2} = \sqrt{\boldsymbol{a}^T \boldsymbol{a}}$$

be the Euclidean norm of $\boldsymbol{a}$. If $\boldsymbol{r}$ and $\boldsymbol{r}_R$ are the vector of residuals from the full and reduced models, respectively, notice that $SSE(F) = \|\boldsymbol{r}\|^2$ and $SSE(R) = \|\boldsymbol{r}_R\|^2$.

The following proposition suggests that Ho is rejected in the partial F test if the change in residual sum of squares $SSE(R) - SSE(F)$ is large compared to $SSE(F)$. If the change is small, then F_R is small and the test suggests that the reduced model can be used.

Proposition 2.9. Let R^2 and R_R^2 be the multiple coefficients of determination for the full and reduced models, respectively. Let $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{Y}}_R$ be the vectors of fitted values for the full and reduced models, respectively. Then the test statistic in the partial F test is

$$F_R = \left[\frac{SSE(R) - SSE(F)}{df_R - df_F}\right] / MSE(F) =$$

$$\left[\frac{\|\hat{\boldsymbol{Y}}\|^2 - \|\hat{\boldsymbol{Y}}_R\|^2}{df_R - df_F}\right] / MSE(F) =$$

$$\frac{SSE(R) - SSE(F)}{SSE(F)} \ \frac{n-p}{p-q} = \frac{R^2 - R_R^2}{1 - R^2} \ \frac{n-p}{p-q}.$$

Definition 2.22. An **FF plot** is a plot of fitted values from 2 different models or fitting methods.

Six plots are useful diagnostics for the partial F test: the RR plot with the residuals from the full model on the vertical axis, the FF plots with the fitted values from the full model on the vertical axis, and always make the response and residual plots for the full and reduced models. Suppose that the full model is a useful MLR model. If the reduced model is good, then the response plots from the full and reduced models should be very similar, visually. Similarly, the residual plots (of the fitted values versus the residuals) from the full and reduced models should be very similar, visually. Finally, the correlation of the plotted points in the RR and FF plots should be high, ≥ 0.95, say, and the plotted points in the RR and FF plots should cluster tightly about the identity line. Add the identity line to both the RR and FF plots as a visual aid. Also add the OLS line from regressing $\boldsymbol{r}$ on $\boldsymbol{r}_R$ to the RR plot (the OLS line is the identity line in the FF plot). If the reduced model is good, then the OLS line should nearly coincide with the identity line in that it should be difficult to see that the two lines intersect at the origin, as in Figure 2.2. If the FF plot looks good but the RR plot does not, the reduced model may be good if the main goal of the analysis is to predict Y.

In Chapter 3, Example 3.8 describes the Gladstone (1905-1906) data. Let the reduced model use a constant, $(size)^{1/3}$, *sex* and *age*. Then Figure 3.7 shows the response and residual plots for the full and reduced models, and Figure 3.9 shows the RR and FF plots.

```
Summary Analysis of Variance Table for the Full Model
Source         df       SS            MS           F    p-value
Regression      6   260467.       43411.1      87.41    0.0000
Residual       69   34267.4       496.629

Summary Analysis of Variance Table for the Reduced Model
Source         df       SS            MS           F    p-value
Regression      2   94110.5       47055.3      17.12    0.0000
Residual       73   200623.       2748.27
```

Example 2.10. For the Buxton (1920) data, $n = 76$ after 5 outliers and 6 cases with missing values are removed. Assume that the response variable Y is *height*, and the explanatory variables are $x_2 =$ *bigonal breadth*, $x_3 =$ *cephalic index*, $x_4 =$ *finger to ground*, $x_5 =$ *head length*, $x_6 =$ *nasal height*, $x_7 =$ *sternal height*. Suppose that the full model uses all 6 predictors plus a

constant (x_1) while the reduced model uses the constant, *cephalic index* and *finger to ground*. Test whether the reduced model can be used instead of the full model using the above output.

Solution: The 4 step partial F test is shown below.
i) Ho: the reduced model is good Ha: use the full model
ii)

$$F_R = \left[\frac{SSE(R) - SSE(F)}{df_R - df_F}\right] / MSE(F) = \left[\frac{200623.0 - 34267.4}{73 - 69}\right] / 496.629$$

$= 41588.9/496.629 = 83.742$.
iii) p–value $= P(F_{4,69} > 83.742) = 0.00$.
iv) The p–value $< \delta$ ($= 0.05$, since δ was not given), so reject Ho. The full model should be used instead of the reduced model. (Bigonal breadth, head length, nasal height, and sternal height are needed in the MLR for height given that cephalic index and finger to ground are in the model.)

Using a computer to get the p–value makes sense, but for exams you may need to use a table. In *ARC*, you can use the *Calculate probability* option from the *ARC* menu, enter 83.742 as the value of the statistic, 4 and 69 as the degrees of freedom, and select the F distribution. To use the table near the end of Chapter 17, use the bottom row since the denominator degrees of freedom $69 > 30$. Intersect with the column corresponding to $k = 4$ numerator degrees of freedom. The cutoff value is 2.37. If the F_R statistic was 2.37, then the p–value would be 0.05. Since $83.472 > 2.37$, the p–value < 0.05, and since $83.472 >> 2.37$, we can say that the p–value ≈ 0.0.

Example 2.11. Now assume that the reduced model uses the constant, *sternal height*, *finger to ground* and *head length*. Using the output below, test whether the reduced model is good.

```
Summary Analysis of Variance Table for Reduced Model
Source          df        SS             MS           F     p-value
Regression       3    259704.         86568.      177.93     0.0000
Residual        72    35030.1        486.528
```

Solution: The 4 step partial F test follows.
i) Ho: the reduced model is good Ha: use the full model
ii)

$$F_R = \left[\frac{SSE(R) - SSE(F)}{df_R - df_F}\right] / MSE(F) = \left[\frac{35030.1.0 - 34267.4}{72 - 69}\right] / 496.629$$

$= 254.2333/496.629 = 0.512$.
iii) The p–value $= P(F_{3,69} > 0.512) = 0.675$.
iv) The p–value $> \delta$, so reject fail to reject Ho. The reduced model is good.

To use the F table near the end of Chapter 17, use the bottom row since the denominator degrees of freedom $69 > 30$. Intersect with the column corresponding to $k = 3$ numerator degrees of freedom. The cutoff value is 2.61. Since $0.512 < 2.61$, the p–value > 0.05, and this is enough information to fail to reject Ho.

2.7 The Wald t Test

Often investigators hope to examine β_k in order to determine the importance of the predictor x_k in the model; however, β_k is the coefficient for x_k given that the other predictors are in the model. Hence β_k depends strongly on the other predictors in the model. Suppose that the model has an intercept: $x_1 \equiv 1$. The predictor x_k is highly correlated with the other predictors if the OLS regression of x_k on $x_1, ..., x_{k-1}, x_{k+1}, ..., x_p$ has a high coefficient of determination R_k^2. If this is the case, then often x_k is not needed in the model given that the other predictors are in the model. If at least one R_k^2 is high for $k \geq 2$, then there is multicollinearity among the predictors.

As an example, suppose that $Y =$ *height*, $x_1 = 1$, $x_2 =$ *left leg length*, and $x_3 =$ *right leg length*. Then x_2 should not be needed given x_3 is in the model and $\beta_2 = 0$ is reasonable. Similarly $\beta_3 = 0$ is reasonable. On the other hand, if the model only contains x_1 and x_2, then x_2 is extremely important with β_2 near 2. If the model contains x_1, x_2, x_3, $x_4 =$ *height at shoulder*, $x_5 =$ *right arm length*, $x_6 =$ *head length* and $x_7 =$ *length of back*, then R_i^2 may be high for each $i \geq 2$. Hence x_i is not needed in the MLR model for Y given that the other predictors are in the model.

Definition 2.23. The 100 $(1 - \delta)$ % CI for β_k is $\hat{\beta}_k \pm t_{n-p,1-\delta/2}\ se(\hat{\beta}_k)$. If the degrees of freedom $d = n - p > 30$, use the N(0,1) cutoff $z_{1-\delta/2}$.

Know how to do the 4 step Wald t–test of hypotheses.
i) State the hypotheses Ho: $\beta_k = 0$ Ha: $\beta_k \neq 0$.
ii) Find the test statistic $t_{o,k} = \hat{\beta}_k / se(\hat{\beta}_k)$ or obtain it from output.
iii) Find the p–value from output or use the t–table: p–value $=$

$$2P(t_{n-p} < -|t_{o,k}|).$$

Use the normal table or $\nu = \infty$ in the t–table if the degrees of freedom $\nu = n - p > 30$.
iv) State whether you reject Ho or fail to reject Ho and give a nontechnical sentence restating your conclusion in terms of the story problem.

Recall that Ho is rejected if the p–value $< \delta$. As a benchmark for this textbook, use $\delta = 0.05$ if δ is not given. If Ho is rejected, then conclude that x_k is needed in the MLR model for Y given that the other predictors are in the model. If you fail to reject Ho, then conclude that x_k is not needed in the MLR model for Y given that the other predictors are in the model. Note that x_k could be a very useful individual predictor, but may not be needed if other predictors are added to the model. It is better to use the output to get the test statistic and p–value than to use formulas and the t–table, but exams may not give the relevant output.

Definition 2.24. Assume that there is a constant $x_1 \equiv 1$ in the model, and let $\boldsymbol{x}_{(k)} = (x_1, ..., x_{k-1}, x_{k+1}, ..., x_p)^T$ be the vector of predictors with the kth predictor x_k deleted. Let $\boldsymbol{r}_{(k)}$ be the residuals from regressing Y on $\boldsymbol{x}_{(k)}$, that is, on all of the predictor variables except x_k. Let $\boldsymbol{r}(x_k|\boldsymbol{x}_{(k)})$ denote the residuals from regressing x_k on $\boldsymbol{x}_{(k)}$. Then an **added variable plot** for x_k is a plot of $\boldsymbol{r}(x_k|\boldsymbol{x}_{(k)})$ versus $\boldsymbol{r}_{(k)}$ for $k = 2, ..., p$.

The added variable plot (also called a partial regression plot) is used to give information about the test $Ho : \beta_k = 0$. The points in the plot cluster about a line through the origin with slope $= \hat{\beta}_k$. An interesting fact is that the residuals from this line, ie the residuals from regressing $\boldsymbol{r}_{(k)}$ on $\boldsymbol{r}(x_k|\boldsymbol{x}_{(k)})$, are exactly the same as the usual residuals from regressing Y on $\boldsymbol{x}$. The range of the horizontal axis gives information about the collinearity of x_k with the other predictors. Small range implies that x_k is well explained by the other predictors. The $\boldsymbol{r}(x_k|\boldsymbol{x}_{(k)})$ represent the part of x_k that is not explained by the remaining variables while the $\boldsymbol{r}_{(k)}$ represent the part of Y that is not explained by the remaining variables.

An added variable plot with a clearly nonzero slope and tight clustering about a line implies that x_k is needed in the MLR for Y given that the other predictors $x_2, ..., x_{k-1}, x_{k+1}, ..., x_p$ are in the model. Slope near zero in the added variable plot implies that x_k may not be needed in the MLR for Y given that all other predictors $x_2, ..., x_{i-1}, x_{k+1}, ..., x_p$ are in the model.

If the zero line with 0 slope and 0 intercept and the OLS line are added to the added variable plot, the variable is probably needed if it is clear that the

two lines intersect at the origin. Then the point cloud should be tilted away from the zero line. The variable is probably not needed if the two lines nearly coincide near the origin in that you can not clearly tell that they intersect at the origin.

Shown below is output only using symbols and the following example shows how to use output to perform the Wald t–test.

Response = Y
Coefficient Estimates

Label	Estimate	Std. Error	t-value	p-value
Constant	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$t_{o,1}$	for Ho: $\beta_1 = 0$
x_2	$\hat{\beta}_2$	$se(\hat{\beta}_2)$	$t_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$	for Ho: $\beta_2 = 0$
$\vdots$				
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$t_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	for Ho: $\beta_p = 0$

```
 Label      Estimate        Std. Error    t-value    p-value
Constant  -7736.26         2660.36       -2.908     0.0079
x2         0.180225        0.00503871    35.768     0.0000
x3        -1.89411         2.65789       -0.713     0.4832

R Squared: 0.987584, Sigma hat: 4756.08, Number of cases: 26

Summary Analysis of Variance Table
Source         df       SS            MS           F     p-value
Regression      2  41380950140.  20690475070. 914.69     0.0000
Residual       23  520265969.    22620260.
```

Example 2.12. The output above was collected from 26 districts in Prussia in 1843. See Hebbler (1847). The goal is to study the relationship between Y = the *number of women married to civilians* in the district with the predictors x_2 = the *population* of the district and x_3 = *military women* = number of women married to husbands in the military.

a) Find a 95% confidence interval for β_2 corresponding to *population*.

The CI is $\hat{\beta}_k \pm t_{n-p,1-\delta/2}\ se(\hat{\beta}_k)$. Since $n = 26$, $df = n - p = 26 - 3 = 23$. From the t–table at the end of Chapter 17, intersect the df = 23 row with the column that is labelled by 95% on the bottom. Then $t_{n-p,1-\delta/2} = 2.069$.

Using the output shows that the 95% CI is $0.180225 \pm 2.069(0.00503871) = (0.16980, 0.19065)$.

b) Perform a 4 step test for Ho: $\beta_2 = 0$ corresponding to *population*.

i) Ho: $\beta_2 = 0$ $H_A : \beta_2 \neq 0$

ii) $t_{o2} = 35.768$

iii) p–value = 0.0

iv) Reject Ho, the population is needed in the MLR model for the number of women married to civilians if number of military women is in the model.

c) Perform a 4 step test for Ho: $\beta_3 = 0$ corresponding to *military women*.

i) Ho: $\beta_3 = 0$ $H_A : \beta_3 \neq 0$

ii) $t_{o2} = -0.713$

iii) p–value = 0.4883

iv) Fail to reject Ho, the number of military women is not needed in the MLR model for the number of women married to civilians if population is in the model.

Figure 2.4 shows the added variable plots for x_2 and x_3. The plot for x_2 strongly suggests that x_2 is needed in the MLR model while the plot for x_3 indicates that x_3 does not seem to be very important. The slope of the OLS line in a) is 0.1802 while the slope of the line in b) is -1.894.

If the predictor x_k is categorical, eg gender, the added variable plot may look like two spheres, but if the OLS line is added to the plot, it will have slope equal to $\hat{\beta}_k$.

2.8 The OLS Criterion

The OLS estimator $\hat{\boldsymbol{\beta}}$ minimizes the OLS criterion

$$Q_{OLS}(\boldsymbol{\eta}) = \sum_{i=1}^{n} r_i^2(\boldsymbol{\eta})$$

where the residual $r_i(\boldsymbol{\eta}) = Y_i - \boldsymbol{x}_i^T\boldsymbol{\eta}$. In other words, let $r_i = r_i(\hat{\boldsymbol{\beta}})$ be the OLS residuals. Then $\sum_{i=1}^{n} r_i^2 \leq \sum_{i=1}^{n} r_i^2(\boldsymbol{\eta})$ for any $p \times 1$ vector $\boldsymbol{\eta}$, and the equality holds iff $\boldsymbol{\eta} = \hat{\boldsymbol{\beta}}$ if the $n \times p$ design matrix $\boldsymbol{X}$ is of full rank $p \leq n$. In particular, if $\boldsymbol{X}$ has full rank p, then $\sum_{i=1}^{n} r_i^2 < \sum_{i=1}^{n} r_i^2(\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2$ even if the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ is a good approximation to the data.

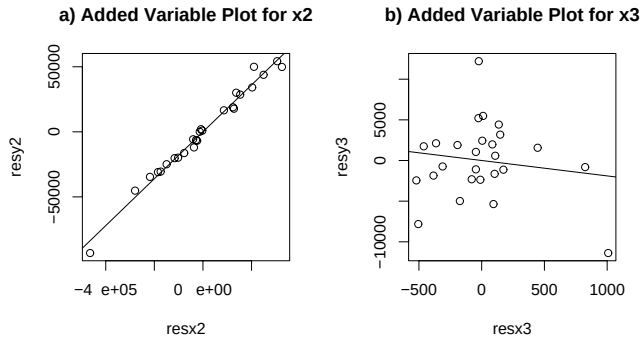

Figure 2.4: Added Variable Plots for x_2 and x_3

Example 2.13. When a model depends on the predictors $\boldsymbol{x}$ only through the linear combination $\boldsymbol{x}^T\boldsymbol{\beta}$, then $\boldsymbol{x}^T\boldsymbol{\beta}$ is called a sufficient predictor and $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ is called an estimated sufficient predictor (ESP). For OLS the model is $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$, and the fitted value $\hat{Y} = ESP$. To illustrate the OLS criterion graphically, consider the Gladstone (1905-6) data where we used *brain weight* as the response. A constant, $x_2 = age$, $x_3 = sex$ and $x_4 = (size)^{1/3}$ were used as predictors after deleting five "infants" from the data set. In Figure 2.5a, the OLS response plot of the OLS ESP $= \hat{Y}$ versus Y is shown. The vertical deviations from the identity line are the residuals, and OLS minimizes the sum of squared residuals. If any other ESP $\boldsymbol{x}^T\boldsymbol{\eta}$ is plotted versus Y, then the vertical deviations from the identity line are the residuals $r_i(\boldsymbol{\eta})$. For this data, the OLS estimator $\hat{\boldsymbol{\beta}} = (498.726, -1.597, 30.462, 0.696)^T$. Figure 2.5b shows the response plot using the ESP $\boldsymbol{x}^T\boldsymbol{\eta}$ where $\boldsymbol{\eta} = (498.726, -1.597, 30.462, 0.796)^T$. Hence only the coefficient for x_4 was changed; however, the residuals $r_i(\boldsymbol{\eta})$ in the resulting plot are much larger on average than the residuals in the OLS response plot. With slightly larger changes in the OLS ESP, the resulting $\boldsymbol{\eta}$ will be such

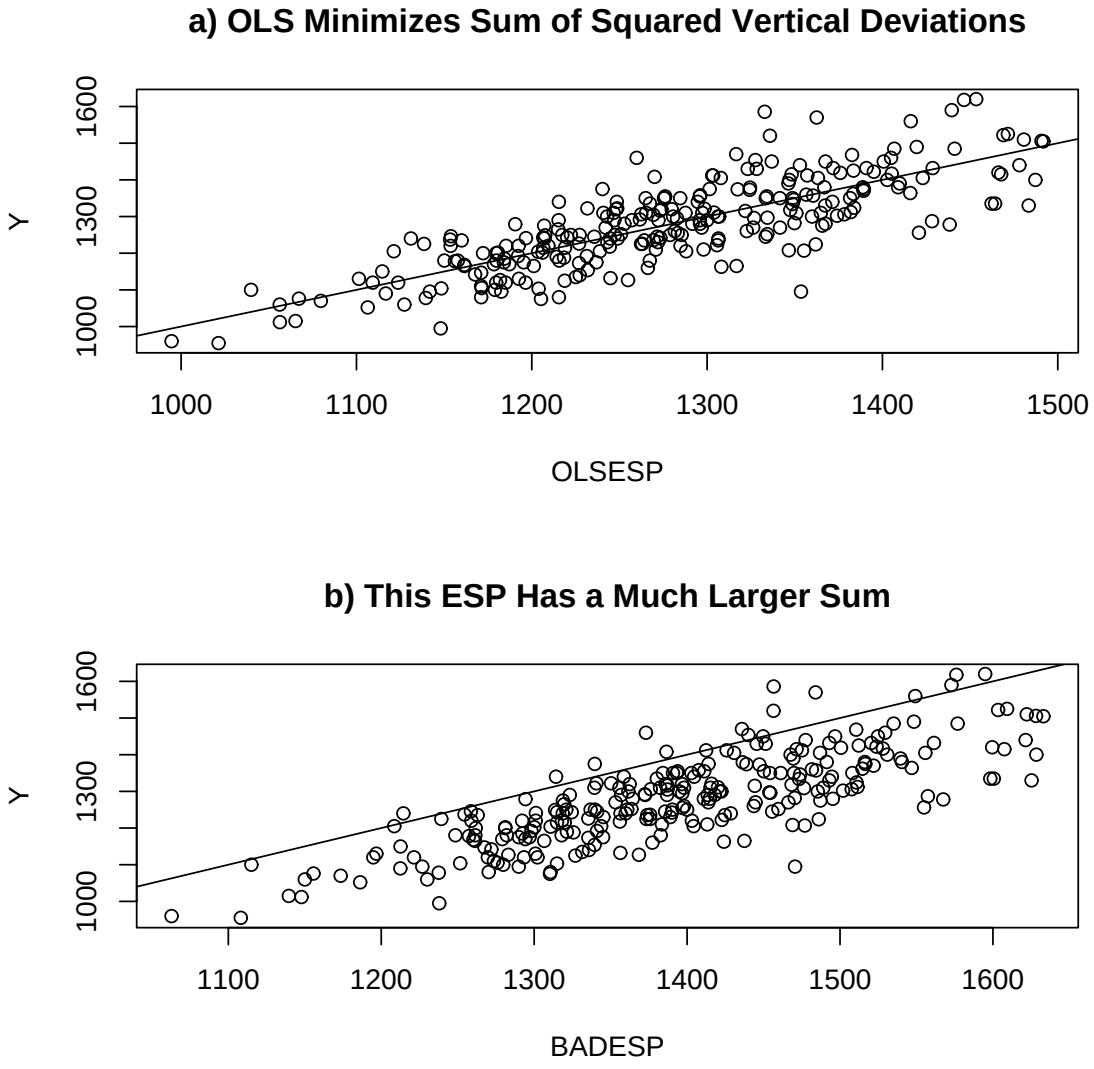

Figure 2.5: The OLS Fit Minimizes the Sum of Squared Residuals

that the squared residuals are massive.

Proposition 2.10. The OLS estimator $\hat{\boldsymbol{\beta}}$ is the unique minimizer of the OLS criterion if $\boldsymbol{X}$ has full rank $p \leq n$.

Proof: Seber and Lee p. 36-37. Recall that the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ and notice that $(\boldsymbol{I}-\boldsymbol{H})^T = \boldsymbol{I}-\boldsymbol{H}$, that $(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{H} = \boldsymbol{0}$ and that $\boldsymbol{HX} = \boldsymbol{X}$. Let $\boldsymbol{\eta}$ be any $p \times 1$ vector. Then

$$(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\eta}) = (\boldsymbol{Y}-\boldsymbol{HY})^T(\boldsymbol{HY}-\boldsymbol{HX}\boldsymbol{\eta}) =$$

$$\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{H}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\eta}) = \boldsymbol{0}.$$

Thus $Q_{OLS}(\boldsymbol{\eta}) =$

$$\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\eta}\|^2 = \|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}+\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\eta}\|^2 =$$

$$\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 + \|\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\eta}\|^2 + 2(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\eta}).$$

Hence

$$\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\eta}\|^2 = \|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 + \|\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\eta}\|^2. \tag{2.16}$$

So

$$\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\eta}\|^2 \geq \|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2$$

with equality iff

$$\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\eta}) = \boldsymbol{0}$$

iff $\hat{\boldsymbol{\beta}} = \boldsymbol{\eta}$ since $\boldsymbol{X}$ is full rank. □

Alternatively calculus can be used. Notice that $r_i(\boldsymbol{\eta}) = Y_i - x_{i,1}\eta_1 - x_{i,2}\eta_2 - \cdots - x_{i,p}\eta_p$. Recall that $\boldsymbol{x}_i^T$ is the ith row of $\boldsymbol{X}$ while $\boldsymbol{x}^j$ is the jth column. Since $Q_{OLS}(\boldsymbol{\eta}) =$

$$\sum_{i=1}^{n}(Y_i - x_{i,1}\eta_1 - x_{i,2}\eta_2 - \cdots - x_{i,p}\eta_p)^2,$$

the jth partial derivative

$$\frac{\partial Q_{OLS}(\boldsymbol{\eta})}{\partial \eta_j} = -2\sum_{i=1}^{n} x_{i,j}(Y_i - x_{i,1}\eta_1 - x_{i,2}\eta_2 - \cdots - x_{i,p}\eta_p) = -2(\boldsymbol{x}^j)^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\eta})$$

for $j = 1, ..., p$. Combining these equations into matrix form, setting the derivative to zero and calling the solution $\hat{\boldsymbol{\beta}}$ gives

$$\boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{0},$$

or

$$\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}. \tag{2.17}$$

Equation (2.17) is known as the **normal equations**. If $\boldsymbol{X}$ has full rank then $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. To show that $\hat{\boldsymbol{\beta}}$ is the global minimizer of the OLS criterion, use the argument following Equation (2.16).

2.9 Two Important Special Cases

When studying a statistical model, it is often useful to try to understand the model that contains a constant but no nontrivial predictors, then try to understand the model with a constant and one nontrivial predictor, then the model with a constant and two nontrivial predictors and then the general model with many predictors. In this text, most of the models are such that Y is independent of $\boldsymbol{x}$ given $\boldsymbol{x}^T\boldsymbol{\beta}$, written

$$Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}^T\boldsymbol{\beta}.$$

Then $w_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ is a scalar, and trying to understand the model in terms of $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ is about as easy as trying to understand the model in terms of one nontrivial predictor. In particular, the plot of $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ versus Y_i is essential.

For MLR, the two main benefits of studying the MLR model with one nontrivial predictor X are that the data can be plotted in a scatterplot of X_i versus Y_i and that the OLS estimators can be computed by hand with the aid of a calculator if n is small.

2.9.1 The Location Model

The *location model*

$$Y_i = \mu + e_i, \quad i = 1, \ldots, n \tag{2.18}$$

is a special case of the multiple linear regression model where $p = 1$, $\boldsymbol{X} = \boldsymbol{1}$ and $\boldsymbol{\beta} = \beta_1 = \mu$. This model contains a constant but no nontrivial predictors.

In the location model, $\hat{\boldsymbol{\beta}}_{OLS} = \hat{\beta}_1 = \hat{\mu} = \overline{Y}$. To see this, notice that

$$Q_{OLS}(\eta) = \sum_{i=1}^{n}(Y_i - \eta)^2 \quad \text{and} \quad \frac{dQ_{OLS}(\eta)}{d\eta} = -2\sum_{i=1}^{n}(Y_i - \eta).$$

Setting the derivative equal to 0 and calling the solution $\hat{\mu}$ gives $\sum_{i=1}^{n} Y_i = n\hat{\mu}$ or $\hat{\mu} = \overline{Y}$. The second derivative

$$\frac{d^2 Q_{OLS}(\eta)}{d\eta^2} = 2n > 0,$$

hence $\hat{\mu}$ is the global minimizer.

2.9.2 Simple Linear Regression

The **simple linear regression** (SLR) model is

$$Y_i = \beta_1 + \beta_2 X_i + e_i = \alpha + \beta X_i + e_i$$

where the e_i are iid with $E(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2$ for $i = 1, ..., n$. The Y_i and e_i are **random variables** while the X_i are treated as known **constants**. The parameters β_1, β_2 and σ^2 are **unknown constants** that need to be estimated. (If the X_i are random variables, then the model is conditional on the X_i's provided that the errors e_i are independent of the X_i. Hence the X_i's are still treated as constants.)

The SLR model is a special case of the MLR model with $p = 2$, $x_{i,1} \equiv 1$ and $x_{i,2} = X_i$. The normal SLR model adds the assumption that the e_i are iid $\text{N}(0, \sigma^2)$. That is, the error distribution is normal with zero mean and constant variance σ^2. The response variable Y is the variable that you want to predict while the predictor variable X is the variable used to predict the response.

For SLR, $E(Y_i) = \beta_1 + \beta_2 X_i$ and the line $E(Y) = \beta_1 + \beta_2 X$ is the regression function. $\text{VAR}(Y_i) = \sigma^2$.

For SLR, the **least squares estimators** $\hat{\beta}_1$ and $\hat{\beta}_2$ minimize the least squares criterion $Q(\eta_1, \eta_2) = \sum_{i=1}^{n} (Y_i - \eta_1 - \eta_2 X_i)^2$. For a fixed η_1 and η_2, Q is the sum of the squared vertical deviations from the line $Y = \eta_1 + \eta_2 X$.

The least squares (OLS) line is $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$ where the slope

$$\hat{\beta}_2 \equiv \hat{\beta} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

and the intercept $\hat{\beta}_1 \equiv \hat{\alpha} = \overline{Y} - \hat{\beta}_2 \overline{X}$.

By the **chain rule,**

$$\frac{\partial Q}{\partial \eta_1} = -2 \sum_{i=1}^{n} (Y_i - \eta_1 - \eta_2 X_i)$$

and

$$\frac{\partial^2 Q}{\partial \eta_1^2} = 2n.$$

Similarly,

$$\frac{\partial Q}{\partial \eta_2} = -2\sum_{i=1}^{n} X_i(Y_i - \eta_1 - \eta_2 X_i)$$

and

$$\frac{\partial^2 Q}{\partial \eta_1^2} = 2\sum_{i=1}^{n} X_i^2.$$

Setting the first partial derivatives to zero and calling the solutions $\hat{\beta}_1$ and $\hat{\beta}_2$ shows that the OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ satisfy the **normal equations**:

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} X_i \text{ and}$$

$$\sum_{i=1}^{n} X_i Y_i = \hat{\beta}_1 \sum_{i=1}^{n} X_i + \hat{\beta}_2 \sum_{i=1}^{n} X_i^2.$$

The first equation gives $\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}$.

There are several equivalent formulas for the slope $\hat{\beta}_2$.

$$\hat{\beta}_2 \equiv \hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{1}{n}(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{\sum_{i=1}^{n} X_i^2 - \frac{1}{n}(\sum_{i=1}^{n} X_i)^2}$$

$$= \frac{\sum_{i=1}^{n}(X_i - \overline{X})Y_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\ \overline{Y}}{\sum_{i=1}^{n} X_i^2 - n(\overline{X})^2} = \hat{\rho} s_Y / s_X.$$

Here the sample correlation $\hat{\rho} \equiv \hat{\rho}(X, Y) = \text{corr}(X, Y) =$

$$\frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)s_X s_Y} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

where the sample standard deviation

$$s_W = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(W_i - \overline{W})^2}$$

for $W = X, Y$. Notice that the term $n-1$ that occurs in the denominator of $\hat{\rho}, s_Y^2$ and s_X^2 can be replaced by n as long as n is used in all 3 quantities.

Also notice that the slope $\hat{\beta}_2 = \sum_{i=1}^n k_i Y_i$ where the constants

$$k_i = \frac{X_i - \overline{X}}{\sum_{j=1}^n (X_j - \overline{X})^2}. \tag{2.19}$$

2.10 The No Intercept MLR Model

The *no intercept MLR model*, also known as *regression through the origin*, is still $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, but there is no intercept β_1 in the model, so $\boldsymbol{X}$ does not contain a column of ones $\mathbf{1}$. Software gives output for this model if the "no intercept" or "intercept = F" option is selected. For the no intercept model, the assumption $E(\boldsymbol{e}) = \mathbf{0}$ is important, and this assumption is rather strong.

Many of the usual MLR results still hold: $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, the vector of *predicted fitted values* $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{H}\boldsymbol{Y}$ where the *hat matrix* $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ provided the inverse exists, and the vector of residuals is $\boldsymbol{r} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$. The response plot and residual plot are made in the same way and should be made before performing inference.

The main difference in the output is the ANOVA table. The ANOVA F test in Section 2.4 tests $Ho : \beta_2 = \cdots = \beta_p = 0$. The test in this section tests $Ho : \beta_1 = \cdots = \beta_p = 0 \equiv Ho : \boldsymbol{\beta} = \mathbf{0}$. The following definition and test follows Guttman (1982, p. 147) closely.

Definition 2.25. Assume that $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where the e_i are iid. Assume that it is desired to test $Ho : \boldsymbol{\beta} = \mathbf{0}$ versus $H_A : \boldsymbol{\beta} \neq \mathbf{0}$.

a) The *uncorrected total sum of squares*

$$SST = \sum_{i=1}^n Y_i^2. \tag{2.20}$$

b) The *model sum of squares*

$$SSM = \sum_{i=1}^n \hat{Y}_i^2. \tag{2.21}$$

c) The residual sum of squares or *error sum of squares* is

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} r_i^2. \tag{2.22}$$

d) The degrees of freedom (df) for SSM is p, the df for SSE is $n-p$ and the df for SST is n. The mean squares are MSE = SSE/$(n-p)$ and MSM = SSM/p.

The ANOVA table given for the "no intercept" or "intercept = F" option is below.

Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Model	p	SSM	MSM	Fo=MSM/MSE	for Ho:
Residual	n-p	SSE	MSE		$\boldsymbol{\beta} = \mathbf{0}$

The 4 step no intercept ANOVA F test for $\boldsymbol{\beta} = \mathbf{0}$ is below.
i) State the hypotheses Ho: $\boldsymbol{\beta} = \mathbf{0}$, Ha: $\boldsymbol{\beta} \neq \mathbf{0}$.
ii) Find the test statistic $F_o = MSM/MSE$ or obtain it from output.
iii) Find the p–value from output or use the F–table: p–value =

$$P(F_{p,n-p} > F_o).$$

iv) State whether you reject Ho or fail to reject Ho. If Ho is rejected, conclude that there is an MLR relationship between Y and the predictors x_1, ..., x_p. If you fail to reject Ho, conclude that there is not a MLR relationship between Y and the predictors x_1, ..., x_p.

Warning: Several important models can be cast in the no intercept MLR form, but often a different test than $Ho : \boldsymbol{\beta} = \mathbf{0}$ is desired. For example, when the generalized or weighted least squares models of Chapter 4 are transformed into no intercept MLR form, the test of interest is Ho: $\beta_2 = \cdots = \beta_p = 0$. The one way ANOVA model of Chapter 5 is equivalent to the cell means model, which is in no intercept MLR form, but the test of interest is $Ho : \beta_1 = \cdots = \beta_p$.

Proposition 2.11. Suppose $Y = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ may or may not contain a column of ones. Then the partial F test of Section 2.6 can be used for inference.

Example 2.14. Consider the Gladstone (1905-6) data described in Example 2.5. If the file of data sets `regdata` is downloaded into *R/Splus*, then the ANOVA F statistic for testing $\beta_2 = \cdots = \beta_4 = 0$ can be found with the following commands. The command *lsfit* adds a column of ones to x which contains the variables *size*, *sex*, *breadth* and *circumference*. Three of these predictor variables are head measurements. Then the response Y is *brain weight*, and the model contains a constant (intercept).

```
> y <- cbrainy
> x <- cbrainx[,c(11,10,3,6)]
> ls.print(lsfit(x,y))
F-statistic (df=4, 262)=196.2433
```

The ANOVA F test can also be found with the no intercept model by adding a column of ones to *R/Splus* matrix x and then performing the partial F test with the full model and the reduced model that only uses the column of ones. Notice that the "intercept=F" option needs to be used to fit both models. The residual standard error = RSE = $\sqrt{MSE}$. Thus SSE = $(n-k)(RSE)^2$ where $n-k$ is the denominator degrees of freedom for the F test and k is the numerator degrees of freedom = number of variables in the model. The column of ones *xone* is counted as a variable. The last line of output computes the partial F statistic and is again ≈ 196.24.

```
> xone <- 1 + 0*1:267
> x <- cbind(xone,x)
> ls.print(lsfit(x,y,intercept=F))
Residual Standard Error=82.9175
F-statistic (df=5, 262)=12551.02

        Estimate  Std.Err t-value Pr(>|t|)
xone     99.8495 171.6189  0.5818   0.5612
size      0.2209   0.0358  6.1733   0.0000
sex      22.5491  11.2372  2.0066   0.0458
breadth  -1.2464   1.5139 -0.8233   0.4111
circum    1.0255   0.4719  2.1733   0.0307

> ls.print(lsfit(x[,1],y,intercept=F))
Residual Standard Error=164.5028
```

```
F-statistic (df=1, 266)=15744.48

  Estimate Std.Err t-value Pr(>|t|)
X 1263.228 10.0674 125.477        0

> ((266*(164.5028)^2 - 262*(82.9175)^2)/4)/(82.9175)^2
[1] 196.2435
```

2.11 Summary

1) The response variable is the variable that you want to predict. The predictor variables are the variables used to predict the response variable.

2) **Regression** is the study of the conditional distribution $Y|\boldsymbol{x}$.

3) The MLR model is

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$$

for $i = 1, \ldots, n$. Here n is the *sample size* and the random variable e_i is the ith **error**. Assume that the errors are iid with $E(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2 < \infty$. Assume that the errors are independent of the predictor variables $\boldsymbol{x}_i$.

4) In matrix notation, these n equations become

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e},$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors.

5) The OLS estimators are $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ and $\hat{\sigma}^2 = MSE = \sum_{i=1}^{n} r_i^2/(n-p)$. Thus $\hat{\sigma} = \sqrt{MSE}$. The vector of *predicted* or *fitted values* $\hat{\boldsymbol{Y}}_{OLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{H}\boldsymbol{Y}$ where the *hat matrix* $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. The ith fitted value $\hat{Y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$. The ith residual $r_i = Y_i - \hat{Y}_i$ and the vector of residuals $\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$. The least squares regression equation for a model containing a constant is $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$.

6) Always make the response plot of $\hat{Y}$ versus Y and residual plot of $\hat{Y}$ versus r for any MLR analysis. The response plot is used to visualize the MLR model, that is, to visualize the conditional distribution of $Y|\boldsymbol{x}^T\boldsymbol{\beta}$. If the iid constant variance MLR model is useful, then i) the plotted points in the

response plot should scatter about the identity line with no other pattern, and ii) the plotted points in the residual plot should scatter about the $r = 0$ line with no other pattern. If either i) or ii) is violated, then the iid constant variance MLR model *is not sustained.* In other words, if the plotted points in the residual plot show some type of dependency, eg increasing variance or a curved pattern, then the multiple linear regression model may be inadequate.

7) Use $x_f < \max h_i$ for valid predictions.

8) If the MLR model contains a constant, then SSTO = SSE + SSR where $SSTO = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$, $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ and $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} r_i^2$.

9) If the MLR model contains a constant, then $R^2 = [\mathrm{corr}(\mathrm{Y_i}, \hat{\mathrm{Y}}_\mathrm{i})]^2 = \dfrac{\mathrm{SSR}}{\mathrm{SSTO}} = 1 - \dfrac{\mathrm{SSE}}{\mathrm{SSTO}}$.

Source	df	SS	MS	F	p-value
Regression	p-1	SSR	MSR	Fo=MSR/MSE	for Ho:
Residual	n-p	SSE	MSE		$\beta_2 = \cdots = \beta_p = 0$

10) Be able to perform the 4 step ANOVA F test of hypotheses:
i) State the hypotheses Ho: $\beta_2 = \cdots = \beta_p = 0$ Ha: not Ho.
ii) Find the test statistic $F_o = MSR/MSE$ or obtain it from output.
iii) Find the p–value from output or use the F–table: p–value =

$$P(F_{p-1,n-p} > F_o).$$

iv) State whether you reject Ho or fail to reject Ho. If Ho is rejected, conclude that there is an MLR relationship between Y and the predictors x_2, ..., x_p. If you fail to reject Ho, conclude that there is a not a MLR relationship between Y and the predictors x_2, ..., x_p.

11) The large sample 100 $(1-\delta)$% CI for $\mathrm{E}(Y_f|\boldsymbol{x}_f) = \boldsymbol{x}_f^T\boldsymbol{\beta} = E(\hat{Y}_f)$ is $\hat{Y}_f \pm t_{n-p,1-\delta/2} se(\hat{Y}_f)$ where $P(T \leq t_{n-p,\delta}) = \delta$ if T has a t distribution with $n-p$ degrees of freedom.

12) The 100 $(1-\delta)$% PI for Y_f is $\hat{Y}_f \pm t_{n-p,1-\delta/2} se(pred)$.

Full model

Source df	SS	MS	Fo and p-value
Regression $p-1$	SSR	MSR	Fo=MSR/MSE
Residual $df_F = n-p$	SSE(F)	MSE(F)	for Ho:$\beta_2 = \cdots = \beta_p = 0$

Reduced model

Source df	SS	MS	Fo and p-value
Regression $q-1$	SSR	MSR	Fo=MSR/MSE
Residual $df_R = n-q$	SSE(R)	MSE(R)	for Ho: $\beta_2 = \cdots = \beta_q = 0$

13) Be able to perform the 4 step **partial F test = change in SS F test** of hypotheses: i) State the hypotheses Ho: the reduced model is good Ha: use the full model.
ii) Find the test statistic $F_R =$

$$\left[\frac{SSE(R) - SSE(F)}{df_R - df_F}\right] / MSE(F)$$

iii) Find the p–value = P($F_{df_R - df_F, df_F} > F_R$). (On exams typically an F table is used. Here $df_R - df_F = p - q =$ number of parameters set to 0, and $df_F = n - p$).
iv) State whether you reject Ho or fail to reject Ho. Reject Ho if the p–value $< \delta$ and conclude that the full model should be used. Otherwise, fail to reject Ho and conclude that the reduced model is good.

Label	Estimate	Std. Error	t-value	p-value
Constant	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$t_{o,1}$	for Ho: $\beta_1 = 0$
x_2	$\hat{\beta}_2$	$se(\hat{\beta}_2)$	$t_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$	for Ho: $\beta_2 = 0$
$\vdots$				
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$t_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	for Ho: $\beta_p = 0$

14) The 100 $(1-\delta)$ % CI for β_k is $\hat{\beta}_k \pm t_{n-p,1-\delta/2}\ se(\hat{\beta}_k)$. If the degrees of freedom $d = n - p > 30$, use the N(0,1) cutoff $z_{1-\delta/2}$.

15) The corresponding 4 step t–test of hypotheses has the following steps: i) State the hypotheses Ho: $\beta_k = 0$ Ha: $\beta_k \neq 0$.

ii) Find the test statistic $t_{o,k} = \hat{\beta}_k / se(\hat{\beta}_k)$ or obtain it from output.
iii) Find the p–value from output or use the t–table: p–value =

$$2P(t_{n-p} < -|t_{o,k}|).$$

Use the normal table or $\nu = \infty$ in the t–table if the degrees of freedom $\nu = n - p > 30$.
iv) State whether you reject Ho or fail to reject Ho and give a nontechnical sentence restating your conclusion in terms of the story problem. If Ho is rejected, then conclude that x_k is needed in the MLR model for Y given that the other predictors are in the model. If you fail to reject Ho, then conclude that x_k is not needed in the MLR model for Y given that the other predictors are in the model.

16) Given $\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$, $\sum_{i=1}^{n}(X_i - \overline{X})^2$, $\overline{X}$, and $\overline{Y}$, find the least squares line $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$ where

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

and $\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}$.

17) Given $\hat{\rho}$, s_X, s_Y, $\overline{X}$, and $\overline{Y}$, find the least squares line $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$ where $\hat{\beta}_2 = \hat{\rho} s_Y / s_X$ and $\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}$.

2.12 Complements

Under regularity conditions, the least squares (OLS) estimator $\hat{\boldsymbol{\beta}}$ satisfies

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(0, \sigma^2 \, \boldsymbol{W}) \tag{2.23}$$

when

$$\frac{\boldsymbol{X}^T \boldsymbol{X}}{n} \rightarrow \boldsymbol{W}^{-1}.$$

This large sample result is analogous to the central limit theorem and is often a good approximation if $n > 5p$ and the error distribution has "light tails," ie, the probability of an outlier is nearly 0 and the tails go to zero at an exponential rate or faster. For error distributions with heavier tails, much larger samples are needed, and the assumption that the variance σ^2 exists is crucial, eg, Cauchy errors are not allowed.

Under the regularity conditions, much of the inference that is valid for the normal MLR model is approximately valid for the iid error MLR model when the sample size is large. For example, confidence intervals for β_i are asymptotically correct, as are t tests for $\beta_i = 0$ (see Li and Duan 1989, p. 1035), the MSE is an estimator of σ^2 by Theorem 2.6 and variable selection procedures perform well (see Chapter 3 and Olive and Hawkins 2005).

Algorithms for OLS are described in Datta (1995), Dongarra, Moler, Bunch and Stewart (1979), and Golub and Van Loan (1989). See Harter (1974a,b, 1975a,b,c, 1976) for a historical account of multiple linear regression. Draper (2000) provides a bibliography of more recent references.

Cook and Weisberg (1997, 1999 ch. 17) call a plot that emphasizes model agreement a *model checking plot*.

Anscombe (1961) and Anscombe and Tukey (1963) suggested graphical methods for checking multiple linear regression and experimental design methods that were the "state of the art" at the time.

The rules of thumb given in this chapter for residual plots are not perfect. Cook (1998, p. 4–6) gives an example of a residual plot that looks like a right opening megaphone, but the MLR assumption that was violated was linearity, not constant variance. Ghosh (1987) gives an example where the residual plot shows no pattern even though the constant variance assumption is violated. Searle (1988) shows that residual plots will have parallel lines if several cases take on each of the possible values of the response variable, eg if the response is a count.

Several authors have suggested using the response plot to visualize the coefficient of determination R^2 in multiple linear regression. See for example Chambers, Cleveland, Kleiner, and Tukey (1983, p. 280). Anderson-Sprecher (1994) provides an excellent discussion about R^2. Kachigan (1982, p. 174 – 177) also gives a good explanation of R^2. Also see Kvålseth (1985) and Freedman (1983).

Hoaglin and Welsh (1978) discuss the hat matrix $\boldsymbol{H}$, and Brooks, Carroll and Verdini (1988) recommend using $x_f < \max h_i$ for valid predictions. Simultaneous prediction intervals are given by Sadooghi-Alvandi (1990). Olive (2007) suggests three large sample prediction intervals for MLR that are valid under the iid error MLR model. Also see Schoemoyer (1992).

Sall (1990) discusses the history of added variable plots while Darlington (1969) provides an interesting proof that $\hat{\boldsymbol{\beta}}$ minimizes the OLS criterion.

2.12.1 Lack of Fit Tests

Label	Estimate	Std. Error	t-value	p-value
Constant	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$t_{o,1}$	for Ho: $\beta_1 = 0$
x_2	$\hat{\beta}_2$	$se(\hat{\beta}_2)$	$t_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$	for Ho: $\beta_2 = 0$
$\vdots$				
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$t_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	for Ho: $\beta_p = 0$

R Squared: R^2
Sigma hat: $\sqrt{MSE}$
Number of cases: n
Degrees of Freedom : $n - p$

Source	df	SS	MS	F	p-value
Regression	p-1	SSR	MSR	Fo=MSR/MSE	for Ho:
Residual	n-p	SSE	MSE		$\beta_2 = \cdots = \beta_p = 0$

The typical "relevant OLS output" has the form given above, but occasionally software also includes output for a lack of fit test as shown below.

Source	df	SS	MS	Fo
Regression	$p-1$	SSR	MSR	Fo=MSR/MSE
Residual	$n-p$	SSE	MSE	
lack of fit	$c-p$	SSLF	MSLF	F_{LF} = MSLF/MSPE
pure error	$n-c$	SSPE	MSPE	

The lack of fit test assumes that

$$Y_i = m(\boldsymbol{x}_i) + e_i \tag{2.24}$$

where $E(Y_i|\boldsymbol{x}_i) = m(\boldsymbol{x}_i)$, m is some possibly nonlinear function, and that the e_i are iid $N(0, \sigma^2)$. Notice that the MLR model is the special case with $m(\boldsymbol{x}_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$. The lack of fit test needs at least one *replicate*: 2 or more Ys with the same value of predictors $\boldsymbol{x}$. Then there a c "replicate groups" with n_j observations in the jth group. Each group has the vector of predictors $\boldsymbol{x}_j$, say, and at least one $n_j > 1$. Also, $\sum_{j=1}^{c} n_j = n$. Denote the Ys in the jth group by Y_{ij}, and let the sample mean of the Ys in the jth group be $\overline{Y}_j$.

Then

$$\frac{1}{n_j - 1} \sum_{i=1}^{n_j} (Y_{ij} - \overline{Y}_j)^2$$

is an estimator of σ^2 for each group with $n_j > 1$. Let

$$SSPE = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (Y_{ij} - \overline{Y}_j)^2.$$

Then $MSPE = SSPE/(n-c)$ is an unbiased estimator of σ^2 when model (2.24) holds, regardless of the form of m. The PE in SSPE stands for "pure error."

Now $SSLF = SSE - SSPE = \sum_{j=1}^{c} n_j(\overline{Y}_j - \hat{Y}_j)^2$. Notice that $\overline{Y}_j$ is an unbiased estimator of $m(\boldsymbol{x}_j)$ while $\hat{Y}_j$ is an estimator of m if the MLR model is appropriate: $m(\boldsymbol{x}_j) = \boldsymbol{x}_j^T \boldsymbol{\beta}$. Hence SSLF and MSLF can be very large if the MLR model is not appropriate.

The 4 step lack of fit test is i) Ho: no evidence of MLR lack of fit, H_A: there is lack of fit for the MLR model.
ii) $F_{LF} = MSLF/MSPE$.
iii) The p–value $= P(F_{c-p,n-c} > F_{LF})$.
iv) Reject Ho if p–value $< \delta$ and state the H_A claim that there is lack of fit. Otherwise, fail to reject Ho and state that there is not enough evidence to conclude that there is MLR lack of fit.

Although the lack of fit test seems clever, examining the response plot and residual plot is a much more effective method for examining whether or not the MLR model fits the data well provided that $n > 10p$. A graphical version of the lack of fit test would compute the $\overline{Y}_j$ and see whether they scatter about the identity line in the response plot. When there are no replicates, the range of $\hat{Y}$ could be divided into several narrow nonoverlapping intervals called slices. Then the mean $\overline{Y}_j$ of each slice could be computed and a step function with step height $\overline{Y}_j$ at the jth slice could be plotted. If the step function follows the identity line, then there is no evidence of lack of fit. However, it is easier to check whether the Y_i are scattered about the identity line. Examining the residual plot is useful because it magnifies deviations from the identity line that may be difficult to see until the linear trend is removed. The lack of fit test may be sensitive to the assumption that the errors are iid $N(0, \sigma^2)$.

When $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}^T\boldsymbol{\beta}$, then the response plot of the estimated sufficient predictor (ESP) $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ versus Y is used to visualize the conditional distribution of $Y|\boldsymbol{x}^T\boldsymbol{\beta}$, and will often greatly outperform the corresponding lack of fit test. When the response plot can be combined with a good lack of fit plot such as a residual plot, using a one number summary of lack of fit such as the test statistic F_{LF} makes little sense.

Nevertheless, the literature for lack of fit tests for various statistical methods is enormous. See Joglekar, Schuenemeyer and LaRiccia (1989), Cheng and Wu (1994), Kauermann and Tutz (2001), Peña and Slate (2006) and Su and Yang (2006) for references.

For the following homework problems, Cody and Smith (2006) is useful for *SAS*, Cook and Weisberg (1999) for *Arc*. Becker, Chambers and Wilks (1988) and Crawley (2007) are useful for *R* and *Splus*.

2.13 Problems

Problems with an asterisk * are especially important.

```
Output for Problem 2.1
Full Model Summary Analysis of Variance Table
Source         df       SS            MS           F    p-value
Regression      6   265784.       44297.4     172.14    0.0000
Residual       67   17240.9       257.327

Reduced Model Summary Analysis of Variance Table
Source         df       SS            MS           F    p-value
Regression      1   264621.       264621.     1035.26    0.0000
Residual       72   18403.8       255.608
```

2.1. Assume that the response variable Y is *height*, and the explanatory variables are $X_2 =$ *sternal height*, $X_3 =$ *cephalic index*, $X_4 =$ *finger to ground*, $X_5 =$ *head length*, $X_6 =$ *nasal height*, $X_7 =$ *bigonal breadth*. Suppose that the full model uses all 6 predictors plus a constant ($= X_1$) while the reduced model uses the constant and *sternal height*. Test whether the reduced model can be used instead of the full model using the output above. The data set had 74 cases.

```
Output for Problem 2.2
Full Model Summary Analysis of Variance Table
Source     df       SS     MS           F        p-value
Regression  9   16771.7   1863.52   1479148.9     0.0000
Residual  235   0.29607   0.0012599

Reduced Model Summary Analysis of Variance Table
Source     df       SS     MS           F        p-value
Regression  2   16771.7   8385.85   6734072.0     0.0000
Residual  242  0.301359   0.0012453

Coefficient Estimates, Response = y, Terms  = (x2 x2^2)
Label      Estimate     Std. Error    t-value    p-value
Constant   958.470      5.88584       162.843     0.0000
x2        -1335.39     11.1656       -119.599     0.0000
x2^2       421.881      5.29434        79.685     0.0000
```

2.2. The above output comes from the Johnson (1996) STATLIB data set *bodyfat* after several outliers are deleted. It is believed that $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_2^2 + e$ where Y is the person's bodyfat and X_2 is the person's density. Measurements on 245 people were taken. In addition to X_2 and X_2^2, 7 additional measurements $X_4, ..., X_{10}$ were taken. Both the full and reduced models contain a constant $X_1 \equiv 1$.

a) Predict Y if $X_2 = 1.04$. (Use the reduced model $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_2^2 + e$.)

b) Test whether the reduced model can be used instead of the full model.

2.3. The output on the next page was produced from the file *mussels.lsp* in *Arc*. See Cook and Weisberg (1999a). Let Y = log(M) where M is the muscle mass of a mussel. Let $X_1 \equiv 1$, $X_2 = \log(\mathrm{H})$ where H is the height of the shell, and let $X_3 = \log(\mathrm{S})$ where S is the shell mass. Suppose that it is desired to predict Y_f if $\log(H) = 4$ and $\log(S) = 5$, so that $\boldsymbol{x}_f^T = (1, 4, 5)$. Assume that $se(\hat{Y}_f) = 0.410715$ and that se(pred) = 0.467664.

a) If $\boldsymbol{x}_f^T = (1, 4, 5)$ find a 99% confidence interval for $E(Y_f)$.

b) If $\boldsymbol{x}_f^T = (1, 4, 5)$ find a 99% prediction interval for Y_f.

```
Output for Problem 2.3
Label      Estimate   Std. Error    t-value    p-value
Constant  -5.07459    1.85124       -2.741     0.0076
log[H]     1.12399    0.498937       2.253     0.0270
log[S]     0.573167   0.116455       4.922     0.0000

R Squared: 0.895655 Sigma hat: 0.223658 Number of cases: 82
(log[H] log[S]) (4 5)
Prediction = 2.2872, s(pred) = 0.467664,
Estimated population mean value  = 2.2872, s = 0.410715

Output for Problem 2.4 Coefficient Estimates Response = height
Label              Estimate  Std. Error    t-value    p-value
Constant           227.351   65.1732         3.488     0.0008
sternal height     0.955973   0.0515390     18.549     0.0000
finger to ground   0.197429   0.0889004      2.221     0.0295

R Squared: 0.879324    Sigma hat:  22.0731

Summary Analysis of Variance Table
Source         df       SS        MS           F    p-value
Regression      2   259167.    129583.     265.96    0.0000
Residual       73   35567.2    487.222
```

2.4. The output above is from the multiple linear regression of the response $Y =$ *height* on the two nontrivial predictors *sternal height* = height at shoulder and *finger to ground* = distance from the tip of a person's middle finger to the ground.

a) Consider the plot with Y_i on the vertical axis and the least squares fitted values $\hat{Y}_i$ on the horizontal axis. Sketch how this plot should look if the multiple linear regression model is appropriate.

b) Sketch how the residual plot should look if the residuals r_i are on the vertical axis and the fitted values $\hat{Y}_i$ are on the horizontal axis.

c) From the output, are *sternal height* and *finger to ground* useful for predicting *height*? (Perform the ANOVA F test.)

2.5. Suppose that it is desired to predict the weight of the brain (in

grams) from the cephalic index measurement. The output below uses data from 267 people.

```
predictor   coef          Std. Error    t-value    p-value
Constant   865.001        274.252         3.154     0.0018
cephalic   5.05961        3.48212         1.453     0.1474
```

Do a 4 step test for $\beta_2 \neq 0$.

2.6. Suppose that the scatterplot of X versus Y is strongly curved rather than ellipsoidal. Should you use simple linear regression to predict Y from X? Explain.

2.7. Suppose that the 95% confidence interval for β_2 is $(-17.457, 15.832)$. In the simple linear regression model, is X a useful linear predictor for Y? If your answer is no, could X be a useful predictor for Y? Explain.

2.8. Suppose it is desired to predict the yearly return from the stock market from the return in January. Assume that the correlation $\hat{\rho} = 0.496$. Using the table below, find the least squares line $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$.

variable	mean $\overline{X}$ or $\overline{Y}$	standard deviation s
January return	1.75	5.36
yearly return	9.07	15.35

2.9. Suppose that $\sum(X_i - \overline{X})(Y_i - \overline{Y}) = 70690.0$, $\sum(X_i - \overline{X})^2 = 19800.0$, $\overline{X} = 70.0$ and $\overline{Y} = 312.28$.

a) Find the least squares slope $\hat{\beta}_2$.

b) Find the least squares intercept $\hat{\beta}_1$.

c) Predict Y if $X = 80$.

x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})(y_i - \overline{y})$	$(x_i - \overline{x})^2$
38	41				
56	63				
59	70				
64	72				
74	84				

2.10. In the above table, x_i is the length of the femur and y_i is the length of the humerus taken from five dinosaur fossils (*Archaeopteryx*) that preserved both bones. See Moore (2000, p. 99).

a) Complete the table and find the least squares estimators $\hat{\beta}_1$ and $\hat{\beta}_2$.

b) Predict the humerus length if the femur length is 60.

2.11. Suppose that the regression model is $Y_i = 7 + \beta X_i + e_i$ for $i = 1, ..., n$ where the e_i are iid $N(0, \sigma^2)$ random variables. The least squares criterion is $Q(\eta) = \sum_{i=1}^{n} (Y_i - 7 - \eta X_i)^2$.

a) What is $E(Y_i)$?

b) Find the least squares estimator $\hat{\beta}$ of β by setting the first derivative $\frac{d}{d\eta} Q(\eta)$ equal to zero.

c) Show that your $\hat{\beta}$ is the global minimizer of the least squares criterion Q by showing that the second derivative $\frac{d^2}{d\eta^2} Q(\eta) > 0$ for all values of η.

2.12. The location model is $Y_i = \mu + e_i$ for $i = 1, ..., n$ where the e_i are iid with mean $E(e_i) = 0$ and constant variance $\text{VAR}(e_i) = \sigma^2$. The least squares estimator $\hat{\mu}$ of μ minimizes the least squares criterion $Q(\eta) = \sum_{i=1}^{n} (Y_i - \eta)^2$. To find the least squares estimator, perform the following steps.

a) Find the derivative $\frac{d}{d\eta}Q$, set the derivative equal to zero and solve for η. Call the solution $\hat{\mu}$.

b) To show that the solution was indeed the global minimizer of Q, show that $\frac{d^2}{d\eta^2}Q > 0$ for all real η. (Then the solution $\hat{\mu}$ is a local min and Q is convex, so $\hat{\mu}$ is the global min.)

2.13. The normal error model for simple linear regression through the origin is

$$Y_i = \beta X_i + e_i$$

for $i = 1, ..., n$ where $e_1, ..., e_n$ are iid $N(0, \sigma^2)$ random variables.

a) Show that the least squares estimator for β is

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}.$$

b) Find $E(\hat{\beta})$.

c) Find VAR$(\hat{\beta})$.

(Hint: Note that $\hat{\beta} = \sum_{i=1}^{n} k_i Y_i$ where the k_i depend on the X_i which are treated as constants.)

2.14. Suppose that the regression model is $Y_i = 10 + 2X_{i2} + \beta_3 X_{i3} + e_i$ for $i = 1, ..., n$ where the e_i are iid $N(0, \sigma^2)$ random variables. The least squares criterion is $Q(\eta_3) = \sum_{i=1}^{n} (Y_i - 10 - 2X_{i2} - \eta_3 X_{i3})^2$. Find the least squares estimator $\hat{\beta}_3$ of β_3 by setting the first derivative $\frac{d}{d\eta_3}Q(\eta_3)$ equal to zero. Show that your $\hat{\beta}_3$ is the global minimizer of the least squares criterion Q by showing that the second derivative $\frac{d^2}{d\eta_3^2}Q(\eta_3) > 0$ for all values of η_3.

Minitab Problems

"Double click" means press the rightmost "mouse" button twice in rapid succession. "Drag" means hold the mouse button down. This technique is used to select "menu" options.

After your computer is on get into *Minitab*, often by double clicking an icon marked "shortcut to math programs" or "math progs" and then double clicking on the icon marked "Student Minitab."
i) In a few seconds, the *Minitab* session and worksheet windows fill the screen. At the top of the screen there is a menu. The upper left corner has the menu option "File." Move your cursor to "File" and drag down the option "Open Worksheet." A window will appear. Double click on the icon "Student." This will display a large number of data sets.
ii) In the middle of the screen there is a "scroll bar," a gray line with left and right arrow keys. Use the right arrow key to make the data file " Prof.mtw" appear. Double click on "Prof.mtw." A window will appear. Click on "OK."
iii) The worksheet window will now be filled with data. The top of the screen has a menu. Go to "Stat" and drag down "Regression." Another window will appear: drag down Regression (write this as Stat>Regression>Regression).
iv) A window will appear with variables to the left and the response variable and predictors (explanatory variables) to the right. Double click on "instrucrs" to make it the response. Double click on "manner" to make it the (predictor) explanatory variable. Then click on "OK."
v) The required output will appear in the session window. You can view the output by using the vertical scroll bar on the right of the screen.
vi) Copy and paste the output into *Word*, or to print your single page of output, go to "File," and drag down the option "Print Session Window." A window will appear. Click on "ok." Then get your output from the printer.

Use the **F3** key to clear entries from a dialog window if you make a mistake or want a new plot.

To get out of *Minitab*, move your cursor to the "x" in the upper right corner of the screen. When asked whether to save changes, click on "no."

2.15 (*Minitab* problem.) See the instructions above for using *Minitab*. Get the data set *prof.mtw*. Assign the response variable to be *instrucr* (the instructor rating from course evaluations) and the explanatory variable (predictor) to be *manner* (the manner of the instructor). Run a regression on these variables.

a) Place the computer output into *Word*.

b) Write the regression equation.

c) Predict *instrucr* if *manner* = 2.47.

d) To get residual and response plots you need to store the residuals and fitted values. Use the menu commands "Stat>Regression>Regression" to get the regression window. Put *instrucr* in the **Response** and *manner* in the **Predictors** boxes. The click on **Storage**. From the resulting window click on **Fits** and **Residuals**. Then click on **OK** twice.

To get a response plot, use the commands "Graph>Plot," (double click) place *instrucr* in the **Y** box, and *Fits1* in the **X** box. Then click on **OK**. Print the plot by clicking on the graph and then clicking on the printer icon.

e) To make a residual plot, use the menu commands "Graph>Plot" to get a window. Place "Resi1" in the **Y** box and "Fits1" in the **X** box. Then click on **OK.** Print the plot by clicking on the graph and then clicking on the printer icon.

2.16. a) Enter the following data on the *Minitab* worksheet:

```
 x  y
30 73
20 50
60 128
80 170
40 87
50 108
60 135
30 60
70 148
60 132
```

To enter the data click on the **C1** column header and enter **x**. Then click on the **C2** header and enter **y**. Then enter the data. Alternatively, copy the data from Problem 2.17 obtained from (www.math.siu.edu/olive/regsas.txt). Then in *Minitab*, use the menu commands "Edit>Paste Cells" and click on "OK." Obtain the regression output from *Minitab* with the menu commands "Stat>Regression>Regression".

b) Place the output into *Word.*

c) Write down the least squares equation.

To save your output on your diskette, use the *Word* menu commands "File > Save as." In the **Save in** box select "3 1/2 Floppy a:" and in the "File name box" enter *HW2d16.doc*. To get a *Word* printout, click on the printer icon or use the menu commands "File>Print."

d) To get residual and response plots you need to store the residuals and fitted values. Use the menu commands "Stat>Regression>Regression" to get the regression window. Put Y in the **Response** and X in the **Predictors** boxes. The click on **Storage**. From the resulting window click on **Fits** and **Residuals**. Then click on **OK** twice.

To make a response plot, use the menu commands "Graph>Plot" to get a window. Place "Y" in the **Y** box and "Fits1" in the **X** box. Then click on **OK.** Print the plot by clicking on the graph and then clicking on the printer icon.

e) To make a residual plot of the fitted values versus the residuals, use the menu commands "Graph>Plot" to get a window. Place "Resi1" in the **Y** box and "Fits1" in the **X** box. Then click on **OK.** Print the plot by clicking on the graph and then clicking on the printer icon.

f) To save your *Minitab* data on your diskette, use the menu commands "File>Save Current Worksheet as." In the resulting dialog window, the top box says **Save in** and there is an arrow icon to the right of the top box. Click several times on the arrow icon until the **Save in** box reads "My computer", then click on 3 1/2 Floppy(A:). In the **File name** box, enter *H2d16.mtw*. Then click on **OK**.

SAS Problems

SAS is a statistical software package widely used in industry. You will need a disk. Referring to the program in Problem 2.17, the semicolon ";" is used to end *SAS* commands and the "options ls = 70;" command makes the output readable. (An "*" can be used to insert comments into the *SAS* program. Try putting an * before the options command and see what it does to the output.) The next step is to get the data into *SAS*. The command "data wcdata;" gives the name "wcdata" to the data set. The command "input x y;" says the first entry is variable x and the 2nd variable y. The command "cards;" means that the data is entered below. Then the data in entered and the isolated semicolon indicates that the last case has been entered. The command "proc print;" prints out the data. The command "proc corr;" will give the correlation between x and y. The commands "proc

plot; plot y*x;" makes a scatterplot of x and y. The commands "proc reg; model y=x; output out = a p =pred r =resid;" tells *SAS* to perform a simple linear regression with y as the response variable. The output data set is called "a" and contains the fitted values and residuals. The command "proc plot data = a;" tells *SAS* to make plots from data set "a" rather than data set "wcdata." The command "plot resid*(pred x);" will make a residual plot of the fitted values versus the residuals and a residual plot of x versus the residuals. The following plot command makes a response plot.

To use *SAS* on windows (PC), use the following steps.

i) Get into *SAS*, often by double clicking on an icon for programs such as a "*Math Progs*" icon and then double clicking on a *SAS* icon. If your computer does not have *SAS*, go to another computer.

ii) A window should appear with 3 icons. Double click on *The SAS System for ....*

iii) Like *Minitab*, a window with a split screen will open. The top screen says *Log-(Untitled)* while the bottom screen says *Editor-Untitled1*. Press the spacebar and an asterisk appears: *Editor-Untitled1**.

2.17. a) Copy and paste the program for this problem from (www.math.siu.edu/olive/reghw.txt), or enter the *SAS* program given below in *Notepad* or *Word*. The *ls* stands for linesize so *l* is a lowercase *L*, not the number one.

When you are done entering the program, save your file as *h2d17.sas* on your diskette (A: drive). (On the top menu of the editor, use the commands "File > Save as". A window will appear. Use the upper right arrow to locate "31/2 Floppy A" and then type the file name in the bottom box. Click on OK.)

```
options ls = 70;
data wcdata;
input x y;
cards;
30 73
20 50
60 128
80 170
40 87
50 108
60 135
30 60
70 148
60 132
;
proc print;
proc corr;
proc plot; plot y*x;
proc reg;
  model y=x;
  output out =a p = pred r = resid;
proc plot data = a;
plot resid*(pred x);
plot y*pred;
run;
```

b) Get back into *SAS*, and from the top menu, use the "File> Open" command. A window will open. Use the arrow in the upper right corner of the window to navigate to "31/2 Floppy(A:)". (As you click on the

arrow, you should see My Documents, C: etc, then 31/2 Floppy(A:).) Double click on **h2d17.sas**. (Alternatively cut and paste the program into the *SAS* editor window.) To execute the program, use the top menu commands "Run>Submit". An output window will appear if successful.

If you were not successful, look at the *log window* for hints on errors. A single typo can cause failure. Reopen your file in *Word* or *Notepad* and make corrections. Occasionally you can not find your error. Then find your instructor or wait a few hours and reenter the program.

c) To copy and paste relevant output into *Word* or *Notepad*, click on the output window and use the top menu commands "Edit>Select All" and then the menu commands "Edit>Copy".

In *Notepad* use the commands "Edit>Paste". Then use the mouse to highlight the relevant output. Then use the commands "Edit>Copy".

Finally, in *Word*, use the commands "Edit>Paste". You can also cut output from *Word* and paste it into *Notepad*.

You may want to save your *SAS* output as the file *HW2d17.doc* on your disk.

d) To save your output on your disk, use the *Word* menu commands "File > Save as." In the **Save in** box select "3 1/2 Floppy a:" and in the "File name box" enter *HW2d17.doc*. To get a *Word* printout, click on the printer icon or use the menu commands "File>Print."

Save the output giving the least squares coefficients in *Word*.

e) Predict Y if $X = 40$.

f) What is the residual when $X = 40$?

2.18. This problem shows how to use *SAS* for MLR. The data are from Kutner, Nachtsheim, Neter and Li (2005, problem 6.5). The response is "brand liking," a measurement for whether the consumer liked the brand. The variable $X1$ is "moisture content" and the variable $X2$ is "sweetness." Enter the program below as file *h2d18.sas*, or copy and paste the program for this problem from (www.math.siu.edu/olive/reghw.txt).

```
options ls = 70;
data brand;
input  y x1 x2;
cards;
   64.0    4.0    2.0
   73.0    4.0    4.0
   61.0    4.0    2.0
   76.0    4.0    4.0
   72.0    6.0    2.0
   80.0    6.0    4.0
   71.0    6.0    2.0
   83.0    6.0    4.0
   83.0    8.0    2.0
   89.0    8.0    4.0
   86.0    8.0    2.0
   93.0    8.0    4.0
   88.0   10.0    2.0
   95.0   10.0    4.0
   94.0   10.0    2.0
  100.0   10.0    4.0
;
proc print;
proc corr;
proc plot; plot y*(x1 x2);
proc reg;
  model y=x1 x2;
  output out =a p = pred r = resid;
proc plot data = a;
plot resid*(pred x1 x2);
plot y*pred;
run;
```

a) Execute the *SAS* program and copy the output file into *Notepad*. Scroll down the output that is now in *Notepad* until you find the regression coefficients and ANOVA table. Then cut and paste this output into *Word*.

b) Do the 4 step ANOVA F test.

You should scroll through your *SAS* output to see how it made the response plot and various residual plots, but cutting and pasting these plots is tedious. So we will use *Minitab* to get these plots. Find the program for this problem from (www.math.siu.edu/olive/regsas.txt). Then copy and paste the numbers (between "cards;" and the semicolon ";") into *Minitab*. Use the mouse commands "Edit>Paste Cells". This should enter the data in the Worksheet (bottom part of *Minitab*). Under **C1** enter **Y** and under **C2** enter **X1** under **C3** enter **X2**. Use the menu commands "Stat>Regression>Regression" to get a dialog window. Enter Y as the response variable and $X1$ and $X2$ as the predictor variable. Click on **Storage** then on **Fits**, **Residuals** and **OK OK**.

c) To make a response plot, enter the menu commands "Graph>Plot" and place "Y" in the Y–box and "FITS1" in the X–box. Click on **OK**. Then use the commands "Edit>Copy Graph" to copy the plot. Include the plot in *Word* with the commands "Edit> Paste." If these commands fail, click on the graph and then click on the printer icon.

d) Based on the response plot, does a linear model seem reasonable?

e) To make a residual plot, enter the menu commands "Graph>Plot" and place "RESI 1" in the Y–box and "FITS1" in the X–box. Click on **OK**. Then use the commands "Edit>Copy Graph" to copy the plot. Include the plot in *Word* with the commands "Edit> Paste." If these commands fail, click on the graph and then click on the printer icon.

f) Based on the residual plot does a linear model seem reasonable?

Problems using ARC

To quit *Arc*, move the cursor to the **x** in the upper right corner and click.

2.19*. (Scatterplot in *Arc*.) Get *cbrain.lsp* from (www.math.siu.edu/olive/regbk.htm), and save the file on a disk. Activate the *cbrain.lsp* dataset with the menu commands "File > Load > 3 1/2 Floppy(A:) > cbrain.lsp." Scroll up the screen to read the data description.

a) Make a plot of *age* versus brain weight *brnweight*. The commands

"Graph&Fit > Plot of" will bring down a menu. Put *age* in the **H** box and *brnweight* in the **V** box. Put *sex* in the **Mark by** box. Click *OK*. Make the **lowess bar** on the plot read .1. Open *Word*.

In *Arc*, use the menu commands "Edit > Copy." In *Word*, use the menu commands "Edit > Paste." This should copy the graph into the *Word* document.

b) For a given age, which gender tends to have larger brains?

c) At what age does the brain weight appear to be decreasing?

2.20. (SLR in *Arc*.) Activate *cbrain.lsp* as in Problem 2.19. Brain weight and the cube root of size should be linearly related. To add the cube root of size to the data set, use the menu commands "cbrain > Transform." From the window, select *size* and enter 1/3 in the **p:** box. Then click *OK*. Get some output with commands "Graph&Fit > Fit linear LS." In the dialog window, put *brnweight* in **Response,** and $(size)^{1/3}$ in **terms**.

a) Cut and paste the output (from *Coefficient Estimates* to *Sigma hat*) into *Word*. Write down the least squares equation $\hat{Y} = b_1 + b_2 x$.

b) If $(size)^{1/3} = 15$, what is the estimated brnweight?

c) Make a residual plot of the fitted values versus the residuals. Use the commands "Graph&Fit > Plot of" and put "L1:Fit-values" in **H** and "L1:Residuals" in **V**. Put sex in the **Mark by** box. Move the OLS bar to 1. Put the plot into *Word*. Does the plot look ellipsoidal with zero mean?

d) Make a response plot of the fitted values versus y = brnweight. Use the commands "Graph&Fit > Plot of" and put "L1:Fit-values in **H** and *brnweight* in **V**. Put *sex* in **Mark by.** Move the OLS bar to 1. Put the plot into *Word*. Does the plot look linear?

2.21. In *Arc* enter the menu commands "File>Load>Data>ARCG" and open the file *mussels.lsp*. This data set is from Cook and Weisberg (1999a).

The response variable Y is the mussel muscle mass M, and the explanatory variables are $X_2 = S =$ shell mass, $X_3 = H =$ shell height, $X_4 = L =$ shell length and $X_5 = W =$ shell width.

Enter the menu commands "Graph&Fit>Fit linear LS" and fit the model: enter S, H, L, W in the "Terms/Predictors" box, M in the "Response" box

and click on *OK*.

a) To get a response plot, enter the menu commands "Graph&Fit>Plot of" and place *L1:Fit-Values* in the H–box and M in the V–box. Copy the plot into *Word*.

b) Based on the response plot, does a linear model seem reasonable?

c) To get a residual plot, enter the menu commands "Graph&Fit>Plot of" and place *L1:Fit-Values* in the H–box and *L1:Residuals* in the V–box. Copy the plot into *Word*.

d) Based on the residual plot, what MLR assumption seems to be violated?

e) Include the regression output in *Word*.

f) Ignoring the fact that an important MLR assumption seems to have been violated, do any of predictors seem to be needed given that the other predictors are in the model?

g) Ignoring the fact that an important MLR assumption seems to have been violated, perform the ANOVA F test.

2.22. Get *cyp.lsp* from (www.math.siu.edu/olive/regbk.htm), and save the file on a disk: you can open the file in *Notepad* and then save it on a disk using the *Notepad* menu commands "File>Save As" and clicking the top checklist then click "Floppy 3 1/2 A:". You could also save the file on the desktop, load it in *Arc* from the desktop, and then delete the file (sending it to the Recycle Bin).

a) In *Arc* enter the menu commands "File>Load>3 1/2 Floppy(A:)" and open the file *cyp.lsp*. This data set consists of various measurements taken on men from Cyprus around 1920. Let the response Y = *height* and X = *cephalic index* = 100(head breadth)/(head length). Use *Arc* to get the least squares output and include the relevant output in *Word*.

b) Intuitively, the cephalic index should not be a good predictor for a person's height. Perform a 4 step test of hypotheses with Ho: $\beta_2 = 0$.

2.23. a) In *Arc* enter the menu commands "File>Load>3 1/2 Floppy(A:)" and open the file *cyp.lsp* (obtained as in Problem 2.22).

The response variable Y is *height*, and the explanatory variables are a constant, X_2 = *sternal height* (probably height at shoulder) and X_3 = *finger to ground*.

Enter the menu commands "Graph&Fit>Fit linear LS" and fit the model: enter *sternal height* and *finger to ground* in the "Terms/Predictors" box, *height* in the "Response" box and click on *OK*.

Include the output in *Word*. Your output should certainly include the lines from "Response = height" to the ANOVA table.

b) Predict Y if $X_2 = 1400$ and $X_3 = 650$.

c) Perform a 4 step ANOVA F test of the hypotheses with
Ho: $\beta_2 = \beta_3 = 0$.

d) Find a 99% CI for β_2.

e) Find a 99% CI for β_3.

f) Perform a 4 step test for $\beta_2 = 0$.

g) Perform a 4 step test for $\beta_3 = 0$.

h) What happens to the conclusion in g) if $\delta = 0.01$?

i) The *Arc* menu "L1" should have been created for the regression. Use the menu commands "L1>Prediction" to open a dialog window. Enter 1400 650 in the box and click on *OK*. Include the resulting output in *Word*.

j) Let $X_{f,2} = 1400$ and $X_{f,3} = 650$ and use the output from i) to find a 95% CI for $E(Y_f)$. Use the last line of the output, that is, se = $S(\hat{Y}_f)$.

k) Use the output from i) to find a 95% PI for Y_f. Now se(pred) = s(pred).

l) Make a residual plot of the fitted values versus the residuals and make the response plot of the fitted values versus Y. Include both plots in *Word*. (See Problem 2.24.)

m) Do the plots suggest that the MLR model is appropriate? Explain.

2.24. In *Arc* enter the menu commands "File>Load>3 1/2 Floppy(A:)" and open the file *cyp.lsp* (obtained as in Problem 2.22).

The response variable Y is *height*, and the explanatory variables are X_2 = *sternal height* (probably height at shoulder) and X_3 = *finger to ground*.

Enter the menu commands "Graph&Fit>Fit linear LS" and fit the model: enter *sternal height* and *finger to ground* in the "Terms/Predictors" box, *height* in the "Response" box and click on *OK*.

a) To get a response plot, enter the menu commands "Graph&Fit>Plot of" and place *L1:Fit-Values* in the H–box and *height* in the V–box. Copy the plot into *Word.*

b) Based on the response plot, does a linear model seem reasonable?

c) To get a residual plot, enter the menu commands "Graph&Fit>Plot of" and place *L1:Fit-Values* in the H–box and *L1:Residuals* in the V–box. Copy the plot into *Word.*

d) Based on the residual plot, does a linear model seem reasonable?

2.25. In *Arc* enter the menu commands "File>Load>3 1/2 Floppy(A:)" and open the file *cyp.lsp* (obtained as in Problem 2.22).

The response variable Y is *height*, and the explanatory variables are X_2 = *sternal height*, X_3 = *finger to ground*, X_4 = *bigonal breadth* X_5 = *cephalic index* X_6 = *head length* and X_7 = *nasal height*. Enter the menu commands "Graph&Fit>Fit linear LS" and fit the model: enter the 6 predictors (in order: X_2 1st and X_7 last) in the "Terms/Predictors" box, *height* in the "Response" box and click on *OK*. This gives the *full model*. For the *reduced model*, only use predictors 2 and 3.

a) Include the ANOVA tables for the full and reduced models in *Word.*

b) Use the menu commands "Graph&Fit>Plot of..." to get a dialog window. Place *L2:Fit-Values* in the H–box and *L1:Fit-Values* in the V–box. Place the resulting plot in *Word*.

c) Use the menu commands "Graph&Fit>Plot of..." to get a dialog window. Place *L2:Residuals* in the H–box and *L1:Residuals* in the V–box. Place the resulting plot in *Word*.

d) Both plots should cluster tightly about the identity line if the reduced model is about as good as the full model. Is the reduced model good?

e) Perform the 4 step partial F test (of Ho: the reduced model is good) using the 2 ANOVA tables from part a).

2.26. a) Activate the *cbrain.lsp* data set in *ARC*. Fit least squares with *age, sex*, $size^{1/3}$, and *headht* as terms and *brnweight* as the response. Assume that the multiple linear regression model is appropriate (this may be a reasonable assumption, 5 infants appear as outliers but the data set has hardly any cases that are babies. If *age* was uniformly represented, the babies might not be outliers anymore). Assuming that *ARC* makes the menu "L1" for this regression, select "AVP-All 2D." A window will appear. Move the OLS slider bar to 1 and click on the "zero line box". The window will show the added variable plots for *age, sex*, $size^{1/3}$, and *headht* as you move along the slider bar that is below "case deletions". Include all 4 added variable plots in *Word*.

b) What information do the 4 plots give? For example, which variables do not seem to be needed?

(If it is clear that the zero and OLS lines intersect at the origin, then the variable is probably needed, and the point cloud should be tilted away from the zero line. If it is difficult to see where the two lines intersect since they nearly coincide near the origin, then the variable may not be needed, and the point cloud may not tilt away from the zero line.)

R/Splus Problem

2.27. a) Use the command *source("A:/regdata.txt")* **to download the data. See Preface or Section 17.1.** You may also copy and paste `regdata.txt` from (www.math.siu.edu/olive/regdata.txt) into *R*. You can copy and paste the *R* following commands for this problem from (www.math.siu.edu/olive/reghw.txt).

For the Buxton (1920) data suppose that the response $Y = height$ and the predictors were a constant, *head length*, *nasal height*, *bigonal breadth* and *cephalic index*. There are 87 cases.

Type the following commands

```
zbux <- cbind(buxx,buxy)
zbux <- as.data.frame(zbux)
zfull <- lm(buxy~len+nasal+bigonal+cephalic,data=zbux)
zred <- lm(buxy~len+nasal,data=zbux)
anova(zred,zfull)
```

b) Include the output in *Word*: press the *Ctrl* and *c* keys as the same time. Then use the menu commands "Edit>Paste" in *Word* (or copy and paste the output).

c) Use the output to perform the partial F test where the full model is described in a) and the reduced model uses a constant, *head length* and *nasal height*. The output from the `anova(zred,zfull)` command produces the correct partial F statistic.

d) Use the following commands to make the response plot for the reduced model. Include the plot in *Word*

```
plot(zred$fit,buxy)
abline(0,1)
```

e) Use the following command to make the residual plot for the reduced model. Include the plot in *Word*.

```
plot(zred$fit,zred$resid)
```

f) The plots look bad because of 5 massive outiers. The following commands remove the outliers. Include the output in *Word*.

```
zbux <- zbux[-c(60,61,62,63,64,65),]
zfull <- lm(buxy~len+nasal+bigonal+cephalic,data=zbux)
zred <- lm(buxy~len+nasal,data=zbux)
anova(zred,zfull)
```

g) Redo the partial F test.

h) Use the following commands to make the response plot for the reduced model without the outliers. Include the plot in *Word*.

```
plot(zred$fit,zbux[,5])
abline(0,1)
```

i) Use the following command to make the residual plot for the reduced model without the outliers. Include the plot in *Word*.

```
plot(zred$fit,zred$resid)
```

j) Do the plots look ok?

2.28. Get the R commands for this problem from (www.math.siu.edu/olive/reghw.txt). The data is such that $Y = 2 + x_2 + x_3 + x_4 + e$ where the zero mean errors are iid [exponential(2) - 2]. Hence the residual and response plots should show high skew. Note that $\boldsymbol{\beta} = (2, 1, 1, 1)^T$. The R code uses 3 nontrivial predictors and a constant, and the sample size $n = 1000$.

a) Copy and paste the commands for part a) of this problem into R. Include the response plot im *Word*. Is the lowess curve fairly close to the identity line?

b) Copy and paste the commands for part b) of this problem into R. Include the residual plot im *Word*: press the *Ctrl* and c keys as the same time. Then use the menu commands "Edit>Paste" in *Word*. Is the lowess curve fairly close to the $r = 0$ line?

c) The output out$coef gives $\hat{\boldsymbol{\beta}}$. Write down $\hat{\boldsymbol{\beta}}$. Is $\hat{\boldsymbol{\beta}}$ close to $\boldsymbol{\beta}$?

2.29. a) Download the *R/Splus* functions `piplot` and `pisim` from *regpack.txt*.

b) The command `pisim(n=100, type = 1)` will produce the mean length of the classical, semiparametric, conservative and asymptotically optimal PIs when the errors are normal, as well as the coverage proportions. Give the simulated lengths and coverages.

c) Repeat b) using the command `pisim(n=100, type = 3)`. Now the errors are EXP(1) - 1.

d) Download `regdata.txt` and type the command `piplot(cbrainx,cbrainy)`. This command gives the semiparametric PI limits for the Gladstone data. Include the plot in *Word*.

e) The infants are in the lower left corner of the plot. Do the PIs seem to be better for the infants or the bulk of the data. Explain briefly.