Math 484 Exam 1 is on Wednesday, Sept. 21 and covers sections 2.1, 2.2, 2.4, 2.6, 2.7, homeworks 1-3 and quizzes 1-3. You are allowed 7 sheets of notes and a calculator. Any needed tables will be provided. CHECK FORMULAS: YOU ARE RESPONSIBLE FOR ANY ERRORS ON THIS HANDOUT!

For the exam and final know the meaning of the least squares regression output.

The MLR model is $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ or

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

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 $V(e_i) = \sigma^2 < \infty$. 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.

Assume that the errors are independent of the predictor variables $\boldsymbol{x}_i$. (If $x_2, ..., x_p$ are random variables, then the model is conditional on the $x_j's$. Hence the $x_j's$ are still treated as constants.) Sometimes it is also assumed that the errors are symmetric. If the errors are iid $N(0, \sigma^2)$, then $Y|\boldsymbol{x}^T\boldsymbol{\beta} \sim N(\boldsymbol{x}^T\boldsymbol{\beta}, \sigma^2)$.

The OLS estimators are $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ 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* $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{Y}$ 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 $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$.

The **response variable** is the variable that you want to predict. The **predictor** (or explanatory or independent) variables are used to predict the response variable.

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 $Y|\boldsymbol{x}^T\boldsymbol{\beta}$. Suppose $n \geq 5p$ and that the errors are roughly symmetric (so not highly skewed). 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 (right or left opening megaphone) or a curved pattern, then the MLR model may be inadequate.

Response = Y, Label or predictor, Estimate or coef, Std. Error or SE, pvalue or Pr > t

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
```

Analysis of Variance Table, Regression or Model, Residual or Error, pvalue or Pr > F

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

```
Response       = brnweight
Coefficient Estimates
Label      Estimate        Std. Error    t-value    p-value
Constant   99.8495         171.619         0.582     0.5612
size       0.220942        0.0357902       6.173     0.0000
sex        22.5491         11.2372         2.007     0.0458
breadth   -1.24638         1.51386        -0.823     0.4111
circum     1.02552         0.471868        2.173     0.0307
R Squared:                0.749755
Sigma hat:                 82.9175
Number of cases:               267
Degrees of freedom:            262
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
```

The above output is in symbols and from Arc.

Assume that the MLR model contains a constant $x_{i1} \equiv 1$ unless told otherwise. Types of problems likely to appear on Exam 1:

1) **Know for final:** The least squares (OLS) **regression equation** for a model containing a constant is $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$. See HW1 Ba, Cb, Dc, Q1 1a?, 2a?.

2) **Know for final:** Given $x_2, ..., x_p$ find $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$.
See HW1 Bb, Cc, Ee, HW3 Aa, Cb, Q1 1b?, 2b?, Q2 2?.

3) **Know for final:** 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.

See HW2 Ab, Ec, HW3 Cc, Q2 1?, 3?.

4) **Know for final:** 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 $z_{1-\delta/2}$ (then the 90% CI uses 1.645, the 95% CI uses 1.96 and the 99% CI uses 2.576).

See HW3cd, Q3.

5) **Know for final:** The corresponding 4 step (Wald) 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.

See HW3 B, Cfgh, Q3.

Full model

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

Reduced model

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

6) **Know for final:** 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.

Variant: Use R output `anova(Red,Full)` to get ii) and iii) where Red corresponds to the reduced model and Full to the full model.

See HW3 Ab, Dc, Dg, Q3.

7) Given data or given Y_i, find the residual $r_i = Y_i - \hat{Y}_i$ where $\hat{Y}_i$ is found using 2). See HW1 Ef.

8) **Know for final:** Be able to recognize whether a response plot is near its ideal shape of scatter about the identity line with no other pattern.

Gaps, curvature, nonconstant variance and outliers (cases far from the bulk of the data) are cause for concern.

Given several response plots, you should be able to pick out the worst one (if all but one are good) or the best one (if all but one are bad).

See HW2 Acd, HW3Clm, Dd, Dj.

9) **Know for final:** Be able to recognize whether a residual plot is near its ideal shape of scatter about the $r = 0$ line with no other pattern.

Gaps, curvature, nonconstant variance and outliers (cases far from the bulk of the data) are cause for concern.

Given several residual plots, you should be able to pick out the worst one (if all but one are good) or the best one (if all but one are bad).

See HW2 Aef, HW3 Clm, De, Dj.

10) A gap is usually bad, but if the fitted values from the MLR fit only to the bulk of the data fit the small cluster of data fairly well, then the small cluster of data are called *good leverage points*. This happened with the brainweight data `cbrain.lsp` in HW1.

11) If the MLR model contains a constant, then given two of SSTO, SSR and SSE, be able to find the third using SSTO = SSE + SSR where $SSTO = \sum_{i=1}^n (Y_i - \overline{Y})^2$, the regression sum of squares $SSR = \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2$ and error (or residual) sum of squares $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n r_i^2$.

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

13) From a story problem be able to determine which variable is the response variable and which variables are the predictor = explanatory variables.

Know: For testing, use $\delta = 0.05$ if δ is not given.

t–table for CIs: For t CIs find the df $= \nu = n-p$. If $n-p > 30$ use the $\nu = \infty$ row and 1.645, 1.96 or 2.576 depending on whether a 90, 95 or 99% CI is wanted. Otherwise intersect the appropriate column (90%, 95% or 99%) with the $\nu = n-p$ row. So $n-p = 14$ and a 90% CI uses $t_{14,0.95} = 1.761$.

F–table for pval: If Den df $= n - p > 60$ use Den df $= \infty$. Otherwise take the table Den df closest to $n - p$. Intersect the Num df column with the Den df row to find the 0.50, ..., 0.999 percentiles. If the statistic F_R is close to 0 and less than the 0.50 percentile, then pval $> 0.5 = 1 - 0.5$. If the test statistic $F_R >$ the 0.999 percentile, then pval $= 0.000 < 0.001 = 1 - 0.999$. If the test statistic F_R is between to percentiles then $1-$ largest percentile area$<$ pval $< 1-$ smallest percentile area. So if Den df $= 20$, Num df $= 7$, and $F_R = 4.00$, then $1 - 0.995 = 0.005 < pval < 0.01 = 1 - 0.99$.

Know: For MLR, the MSE $= SSE/(n-p)$ is an unbiased estimator of the error variance σ^2.