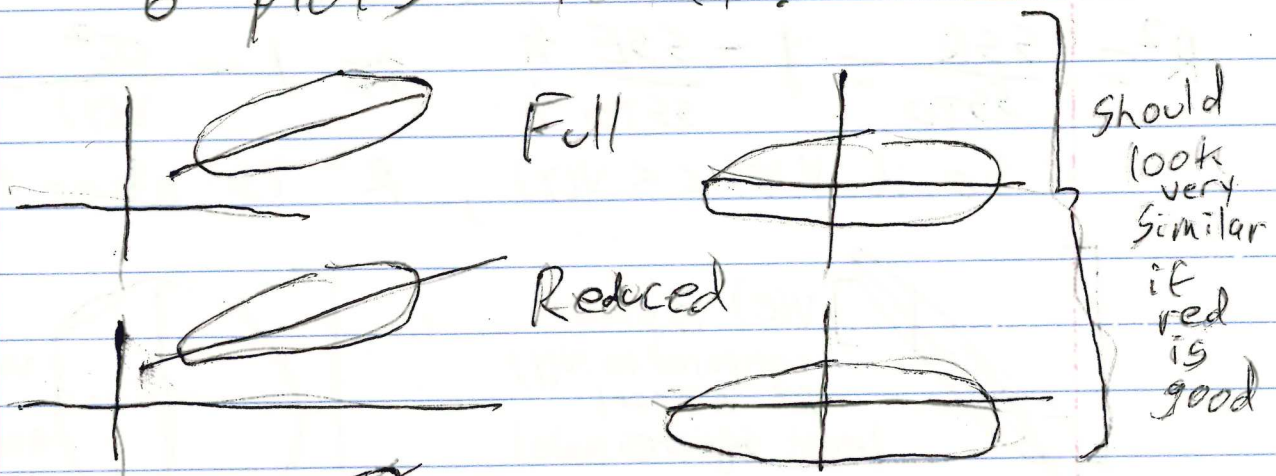
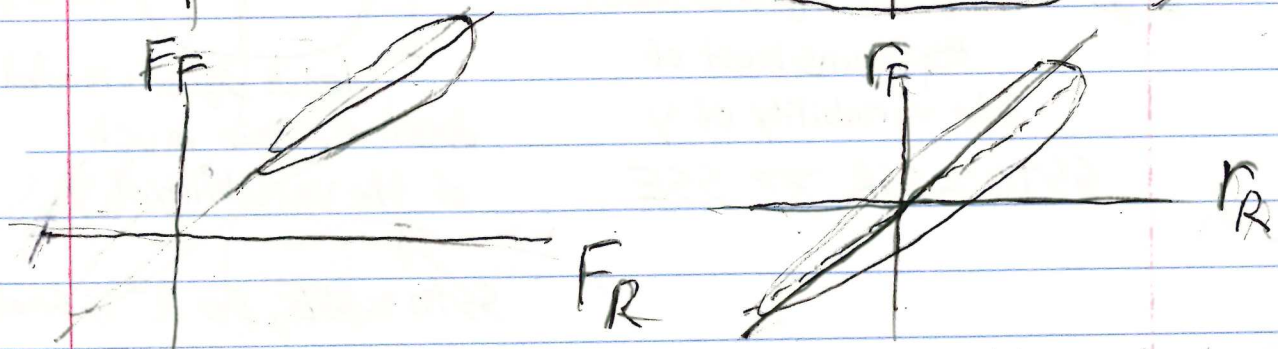
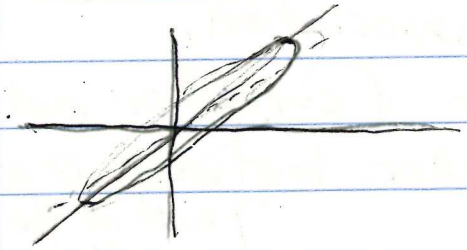
67] p 47 Make the response and residual plots for the full and reduced model, the RR plot and the FF plot: 6 plots total.





Make these plots for variable selection in ch3.

68] p 34 For the Anova F test, if you fail to reject H_0, the reduced model residuals are $r_{Ri} = Y_i - \bar{Y}$ and full model residuals are $r_i = Y_i - \hat{Y}$



see RR plot on p 35

69] p 47 If the FF plot looks good, but the RR plot is poor, the reduced model

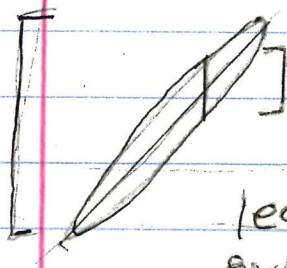
may be good if the main goal of the analysis is to predict Y (or description).

70] If a constant is in the model,

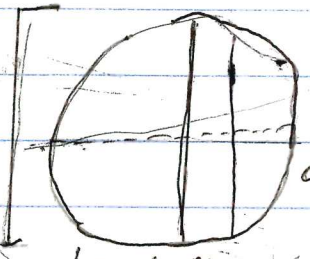
$SSTO = SSE + SSR$ and

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE\,\frac{1}{n}}{SSTO\,\frac{1}{n}} \approx 1 - \frac{\sigma^2}{V(Y)} = \frac{1 - V(e)}{V(Y)}$$

So if $V(e) << V(Y)$, R^2 is near 1.

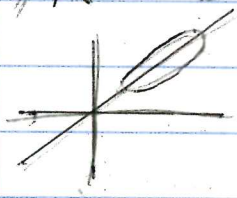


V(e) is small compared to V(Y)
R^2 is near 1
least squares model explains most of the variability of Y
$SSTO \approx SSR >> SSE$



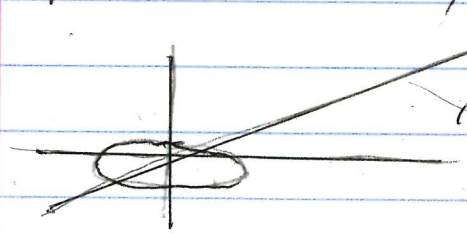
V(e) = V(Y) and $R^2 \approx 0$
least squares model doesn't explain much of the variability of Y
$SSTO \approx SSE$, so R^2 is small

71] i) R^2 is best for a good response plot



ii) If response plot is linear with gaps, R^2 is usually too high and over emphasizes linear association.

iii) If plot is nonlinear (curved or outliers) R^2 can be any number between 0 and 1



identity line

most of the variability of Y is due to the cluster of outliers

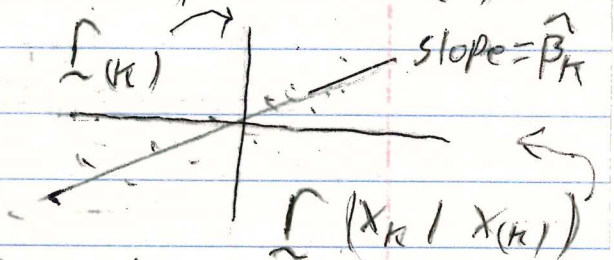
If $Y=X^2$ on an interval symmetric about $X=0$, could have $R^2=0$. But given X, you know Y. R^2 measures linear association $\neq$ association.

end exam 1 material

72] pg 1 Let $X_1 \equiv 1$ and

$$\underset{\sim}{X}_{(k)} = (X_1, X_2, \ldots, X_{k-1}, X_{k+1}, \ldots, X_p)^T$$

be the vector of predictors with X_k deleted. Let $\underset{\sim}{r}_{(k)}$ be the vector of residuals from regressing Y on $\underset{\sim}{X}_{(k)}$. Let $\underset{\sim}{r}(X_k | \underset{\sim}{X}_{(k)})$ be the vector of residuals from regressing X_k on $\underset{\sim}{X}_{(k)}$. An added variable plot for X_k is a plot of $\underset{\sim}{r}(X_k | \underset{\sim}{X}_{(k)})$ vs $\underset{\sim}{r}_{(k)}$ for $k=2, \ldots, p$.



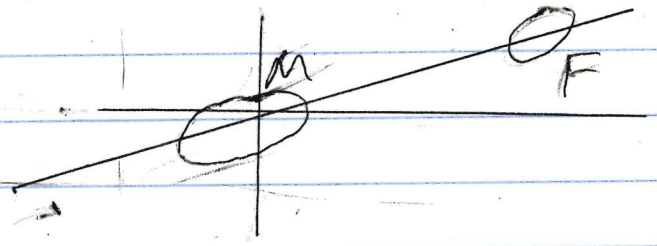
73] This plot gives information about Ho $\beta_k = 0$. The plotted points cluster about a line through the origin with slope $\hat{\beta}_k$. Small range on the horiz axis means X_k is well explained by the other predictors. $\underset{\sim}{r}(X_k | X_{(k)})$ represents

the part of x_k not explained by the remaining variables while $\underset{\sim}{r}_{(k)}$ represents the part of Y that is not explained by the remaining variables.

74] p51 An added variable plot with clearly nonzero slope and tight clustering about some line through the origin implies x_k is needed in the MLR model for Y given the other variables are in the model.

Slope near zero implies that x_k may not be needed, see Fig 2.4.

75] p53 For a categorical variable like gender, the plot tends to have clusters.



slope $= \hat{\beta}_k$

x_k = gender

§23 76] p25 If residual and response plots look good after several points are moved slightly, then the plots are ok. (If the points have to be moved a lot, they may be outliers.)

77] p27 The sample correlation between w_i and z_i is $\text{corr}(w,z) = \hat{\rho}(w,z) = \frac{\sum_{i=1}^{n}(w_i-\bar{w})(z_i-\bar{z})}{\sqrt{\sum(w_i-\bar{w})^2 \sum(z_i-\bar{z})^2}}$

$= \frac{\sum w_i z_i - n\bar{w}\bar{z}}{\sqrt{\sum(w_i-\bar{w})^2 \sum(z_i-\bar{z})^2}}$.

$\left(\text{num} = \sum w_i z_i - \sum w_i \bar{z} - \sum z_i \bar{w} + n\bar{w}\bar{z} = \sum w_i z_i - n\bar{w}\bar{z}\right)$

78] p26-7 Suppose $\underset{\sim}{X}$ is $n \times p$ with full rank p, $n > p$, and a constant is in the model. Fact if $\hat{x} = \bar{x}$, then $\hat{y} = \bar{y}$.

i) $\sum r_i = 0$ so $\bar{r} = 0$

ii) $\sum r_i \hat{y}_i = \underset{\sim}{r}^T \underset{\sim}{\hat{y}} = 0$ so $\text{corr}(r, \hat{y}) = 0$

iii) $\underset{\sim}{X}^T \underset{\sim}{r} = \underset{\sim}{0}$ so $\underset{\sim}{x}^{jT} \underset{\sim}{r} = 0$ so $\text{corr}(r, x_j) = 0$ (x^j: jth column of X; x_j: jth predictor)

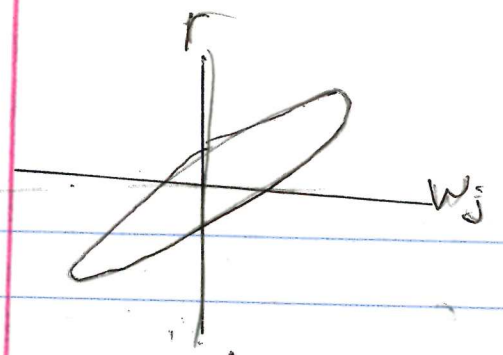
iv) $\sum y_i = \sum \hat{y}_i = \sum \hat{y}_i + \sum r_i$

79] * p25 i) The residual plot of $\hat{y}$ vs r should always be made.

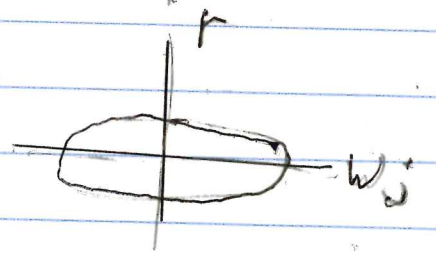
ii) Residual plots of the jth predictor x_j ($= x^j$) vs $\underset{\sim}{r}$ should look ellipsoidal about the $r=0$ line for $j = 2, \ldots, p$ if the errors are independent of the predictors and the predictors are quantitative.

iii) Residual plots of potential predictors w_j vs r may be made. If w_j is not important, the plot should look ellipsoidal about $r=0$ if w_j is quantitative.

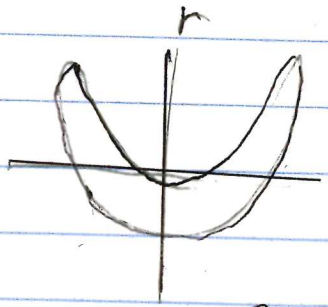
iv) If the plot is parabolic, add x_j^2 to the model or add w_j and w_j^2 to the model.



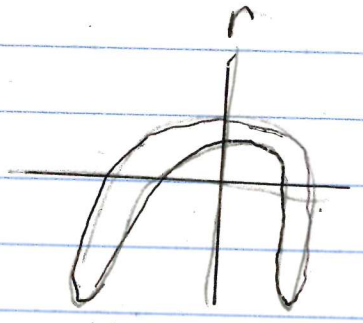
add w_j to model



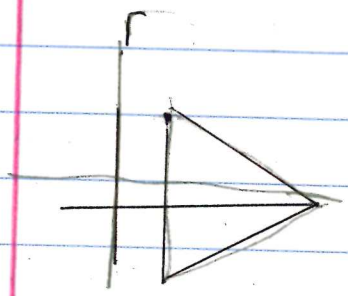
don't add w_j to model

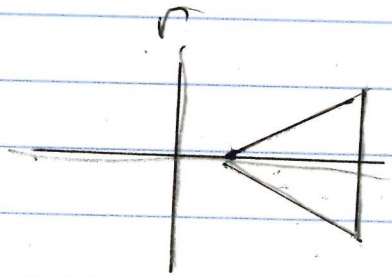


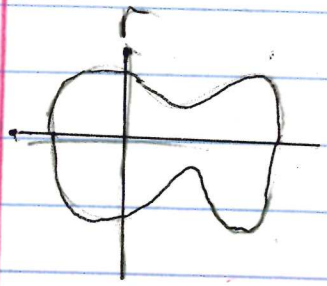
add x_j^2 to model



add w_j and w_j^2 to model







check constant variance assumption if $\hat{Y}$ is used

NOW

80] Response plots are good for visualizing the MLR model ($Y | x^T\beta$) in the background of the data, for checking goodness of fit, linearity, strength of relationship as measured by R^2 or signal to noise ratio, for detecting