BIOS 145 Lab #5 10/1/97

- 1st hour: KKM Ch.8: #2, 3
- 2nd hour: questions
- Other:

    1. Problems with web page/downloading?
    2. Lab overheads now on web page & at HSL
    3. Pone dataset on disk at SPH lab
    4. Starting tomorrow, extra office hour:
       2303 MCG Thurs. 1-2pm.
    5. HW#5 assigned?

Multiple Regression:

Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + e$$

Or:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + e_i,$$

$$i = 1, 2, \ldots, n$$

ANOVA Table:

Source	df	SS	MS	F	R^2
Regression	k	$\Sigma_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$	$\frac{\text{SSR}}{k}$	$\frac{\text{MSR}}{\text{MSE}}$	$\frac{\text{SSR}}{\text{SST}}$
Residual	$n-k-1$	$\Sigma_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$	$\frac{\text{SSE}}{n-k-1}$		
Total	$n-1$	$\Sigma_{i=1}^{n}(Y_i - \overline{Y})^2$			

KKM Ch.8 #2:

> Psychologist wanted to know if level of pathology (Y) in psychotic patients 6 mo. after trt. could be predicted from pretrt. symptom ratings of thinking disturbance (X_1) and hostile suspiciousness (X_2).

a. Least-squares eqn. with both variables involved is given by

$$\widehat{Y}_5 = -0.628 + 23.639 X_1 - 7.147 X_2$$

Determine $\widehat{Y}$ for pt. with pretrt. scores of $X_1 = 2.80$ and $X_2 = 7.0$.

$$\begin{aligned}\widehat{Y}_5 &= -0.628 + 23.639(2.80) - 7.147(7.0) \\ &\approx 15.53\end{aligned}$$

Observed value for pt.#5 $= Y_5 = 25$, so the predicted value doesn't compare well to the observed.

b. Using ANOVA tables, do overall F-tests for models containing: 1) X_1 and X_2, 2) X_1 alone, and 3) X_2 alone.

$\underline{X_1 \& X_2}$:

$$F = \frac{\text{MSR}(X_1, X_2)}{\text{MSE}(X_1, X_2)} = \frac{2,784/2}{11,008/50} \approx 6.32$$

Here, $F \sim F_{2,50}$ and from back of KKM,

$$\begin{aligned}F_{2,50,.95} &= 3.18 \\ F_{2,50,.975} &= 3.97 \\ F_{2,50,.99} &= 5.06 \\ F_{2,50,.995} &= 5.90 \\ F_{2,50,.999} &= 7.96\end{aligned}$$

so $.001 < p$-value $< .005$. Thus, we reject H_0: $\beta_1 = \beta_2 = 0$ at the $\alpha = .05$ level.

$\underline{X_1}$:

$$F = \frac{\text{MSR}(X_1)}{\text{MSE}(X_1)} = \frac{1,546/1}{12,246/51} \approx 6.44$$

Here, $F \sim F_{1,51}$ and from back of KKM,

$$F_{1,51,.975} \approx 5.33$$
$$F_{1,51,.99} \approx 7.16$$

so $.01 < p$-value $< .025$. Thus, we reject H_0: $\beta_1 = 0$ at the $\alpha = .05$ level.

$\underline{X_2}$:

$$F = \frac{\text{MSR}(X_2)}{\text{MSE}(X_2)} = \frac{160/1}{13,632/51} \approx 0.60$$

Here, $F \sim F_{1,51}$ and from back of KKM,

$$F_{1,51,.75} \approx 1.34$$

so p-value > 0.25. Thus, we can't reject H_0: $\beta_2 = 0$ at the $\alpha = .05$ level.

c. Based on results in (b), how would you rate the importance of the two variables in predicting Y?

d. What are the R^2-values for the three regressions referred to in (b)?

$$R^2_{X_1,X_2} = \frac{\text{SSR}(X_1, X_2)}{\text{SST}(X_1, X_2)} = \frac{2,784}{13,792} \approx 0.20$$

$$R^2_{X_1} = \frac{\text{SSR}(X_1)}{\text{SST}(X_1)} = \frac{1,546}{13,792} \approx 0.11$$

$$R^2_{X_2} = \frac{\text{SSR}(X_2)}{\text{SST}(X_2)} = \frac{160}{13,792} \approx 0.01$$

Note:

$$\begin{aligned}
\text{adj. } R^2_{X_1,X_2} &= 1 - \frac{\text{MSE}(X_1, X_2)}{\text{MST}(X_1, X_2)} \\
&= 1 - \frac{11,008/50}{13,792/52} \\
&\approx .17 \\
\text{adj. } R^2_{X_1} &= 1 - \frac{\text{MSE}(X_1)}{\text{MST}(X_1)} \\
&= 1 - \frac{12,246/51}{13,792/52} \\
&\approx .09 \\
\text{adj. } R^2_{X_2} &= 1 - \frac{\text{MSE}(X_2)}{\text{MST}(X_2)} \\
&= 1 - \frac{13,632/51}{13,792/52} \\
&\approx -.008!
\end{aligned}$$

e. What is the best model involving either one or both the independent variables?

3. Sample of 25 patients with hyperlipoproteinemia before drug therapy. Given weight (X_1), age (X_2), and plasma lipid levels of total cholesterol (Y)

a. Given the ANOVA tables, which of the two indep. vars would you say is more important predictor of Y?

b. Given

$$\widehat{Y}_{X_1,X_2} = 77.983 + 0.417X_1 + 5.217X_2$$

$$\widehat{Y}_{X_1} = 199.2975 + 1.622X_1$$

$$\widehat{Y}_{X_2} = 102.5751 + 5.321X_2$$

Find $\widehat{Y}$ for patient 4 using each model.

$$\widehat{Y}_{X_1,X_2} = 77.983 + 0.417(70) + 5.217(30) \approx 263.68$$

$$\widehat{Y}_{X_1} = 199.2975 + 1.622(70) \approx 312.84$$

$$\widehat{Y}_{X_2} = 102.5751 + 5.321(30) \approx 262.21$$

and the observed value for pt.4 is $Y_4 = 263$.

Comment.

c. Given ANOVA table, carry out overall F-test for the two-variable model and the partial F-test for the addition of X_1 to the model, given that X_2 is already in the model

Overall:

$$F = \frac{102,570.8/2}{42,806.2/22} = 26.36$$

Here, $F \sim F_{2,22}$ and $F_{2,22,.999} = 9.61$ so p-value <<< 0.001. Thus, we reject H_0: $\beta_1 = \beta_2 = 0$ at the $\alpha = .05$ level.

Partial $(X_1 | X_2)$:

$$\begin{aligned} F &= \frac{\{\text{SSR}(X_1, X_2) - \text{SSR}(X_2)\}/1}{\text{SSE}(X_1, X_2)/22} \\ &= \frac{102,570.8 - 101,932.7}{42,806.2/22} \\ &\approx 0.33 \end{aligned}$$

Here, $F \sim F_{1,22}$ and $F_{1,22,.75} = 1.40$ so p-value > 0.25. Thus, we cannot reject H_0: $\beta_1 = 0$ at the $\alpha = .05$ level.

d. Compute and compare R^2 values for each of the three models in (b).

Check that $R^2_{X_1,X_2} \approx .71$, $R^2_{X_1} \approx .07$, and $R^2_{X_2} \approx .70$

e. Based on (a)-(d), what do you consider to be the best model?