BIOS 145 Lab #8 10/15/97

1st hour: KKM Ch.13: #1, 4 (rev.) Ch.14: #2a-c, 9a-c

2nd hour: questions

Other:

1. New additions to web page
2. Recent HW labels - check web page (news link)
3. Computer Exercise #2
4. Tomorrow: Ch. 15, 11

KKM Ch.13: #1

a. left as exercise

b. 1. Estimated equation of straight-line regression of Y on X:

$$\widehat{Y} = -1.9318 + .24597X$$

2. Estimated equation (quadratic)

$$\widehat{Y} = 3.1721 - .78102X + .04668X^2$$

3. Estimated equation (ln$Y$ on X)

$$\widehat{\ln} Y = -6.2096 + .45117X$$

4. left as exercise

c. Complete the table for straight-line regression:

Source	df	SS	MS	F	Pr $> F$
Regression	1	12.705	12.705	43.7	$<$ .001
Residual					
Lack of fit	4	4.4196	1.1049	57.0	$<$ .001
Pure Error	12	.2325	.0194		
Total	17	17.3571			

d. Table for quadratic regression:

Source	df	SS	MS	F	Pr $> F$
Regression					
X	1	12.705	12.705	43.7	$< .001$
$X^2 \mid X$	1	3.9051	3.9051	78.4	$< .001$
Residual					
Lack of fit	3	.51466	.17155	8.86	(.001,.005)
Pure Error	12	.23248	.01937		
Total	17	17.3571			

e. $r^2_{XY} = .732$, R^2(quadratic) $= .957$

f. See part (c)

g. See part (d)

KKM Ch.13: #4 (revised to be like #5)

For the date on VOTE (Y) and TVEXP (X) in Ch.5 #5,

a. Test for significance of straight-line regression and adequacy of fit of straight-line model.

Est. eqn.: $\widehat{Y} = 2.174 + 1.177X$

Source	df	SS	MS	F	Pr $> F$
Regression	1	2023.205	2023.205	182.3	$< .001$
Residual					
Lack of fit	13	169.253	13.019	2.13	(.10,.25)
Pure Error	5	30.560	6.112		
Total	19	2223.018			

b. Test for significance of addition of X^2 to the model.

Source	df	SS	MS	F	Pr $> F$
Regression					
X	1	2023.205	2023.205	182.3	$< .001$
$X^2 \mid X$	1	2.450	2.450	.211	$> .25$
Residual					
Lack of fit	12	166.803	13.900	2.27	(.10,.25)
Pure Error	5	30.560	6.112		
Total	19	2223.018			

c. Which model is more appropriate?

Note: $r^2_{XY} = .910$, R^2(quadratic) $= .911$

<u>KKM Ch.14: #2</u>

Tables display values of per capita income (PCI) and population percentage under age 15 (YNG) for a hypothetical sample of developing countries in Latin America and Africa.

a. Determine l.s. line of PCI (Y) on YNG (X) separately for Latin American and African countries.

See attached SAS printouts from PROC GLM.

$$\widehat{Y}_{\text{LA}} = 2170.67 - 42.0X \text{ and } \widehat{Y}_{\text{AF}} = 893.571 - 17.28X$$

b. Test H_0: "the slopes are the same for the Latin Am. countries as the African countries being sampled" vs. H_A: "the slopes are different"

We calculate from equation 14.6,

$$\begin{aligned} S^2_{P,Y|X} &= \frac{(n_{\text{LA}} - 2)S^2_{Y|X_{\text{LA}}} + (n_{\text{AFR}} - 2)S^2_{Y|X_{\text{AFR}}}}{n_{\text{LA}} + n_{\text{AFR}} - 4} \\ &= \frac{(26-2)(1391.756) + (26-2)(157.238)}{26 + 26 - 4} \\ &= 774.497 \end{aligned}$$

So,

$$\begin{aligned} S^2_{\widehat{\beta}_{1,\text{LA}} - \widehat{\beta}_{1,\text{AFR}}} &= S^2_{P,Y|X}\left[\frac{1}{(n_{\text{LA}} - 1)S^2_{X_{\text{LA}}}} + \frac{1}{(n_{\text{AFR}} - 1)S^2_{X_{\text{AFR}}}}\right] \\ &= (774.497)\left[\frac{1}{(26-1)(21.633)} + \frac{1}{(26-1)(20.611)}\right] \\ &= 2.935 \end{aligned}$$

and

$$T = \frac{\widehat{\beta}_{1,\text{LA}} - \widehat{\beta}_{1,\text{AFR}}}{S_{\widehat{\beta}_{1,\text{LA}} - \widehat{\beta}_{1,\text{AFR}}}} = \frac{(-42.0) - (-17.28)}{1.713} = -14.43$$

c. Test H_0: "the intercepts are the same for the LA countries as the African countries being sampled"

Here, from equation 14.8,

$$\begin{aligned} S^2_{\widehat{\beta}_{0,\text{LA}} - \widehat{\beta}_{0,\text{AFR}}} &= S^2_{P,Y|X}\left[\frac{1}{n_{\text{LA}}} + \frac{1}{n_{\text{AFR}}} + \frac{\overline{X}^2_{\text{LA}}}{(n_{\text{LA}}-1)S^2_{X_{\text{LA}}}} + \frac{\overline{X}^2_{\text{LA}}}{(n_{\text{LA}}-1)S^2_{X_{\text{LA}}}}\right] \\ &= (774.497)\left[\frac{1}{26} + \frac{1}{26} + \frac{(40.277)^2}{(26-1)(21.633)} + \frac{(40.931)^2}{(26-1)(20.611)}\right] \\ &= 4900.9 \end{aligned}$$

so

$$T = \frac{\widehat{\beta}_{0,\text{LA}} - \widehat{\beta}_{0,\text{AFR}}}{S_{\widehat{\beta}_{0,\text{LA}} - \widehat{\beta}_{0,\text{AFR}}}} = \frac{(2170.67) - (893.571)}{70.01} = 18.24$$

<u>KKM Ch14: #9</u>

See Problem 2 of Chapter 5

Here, let $X_1 =$ AGE, $X_2 =$ QUET, $Z =$ SMK

a. State the appropriate regression model incorporating both equations for smokers and nonsmokers:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 Z + \beta_4 X_1 Z + \beta_5 X_2 Z + e$$

The model for nonsmokers is:

$$Y_{\text{NS}} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$$

and the model for smokers is:

$$Y_{\text{S}} = (\beta_0 + \beta_3) + (\beta_1 + \beta_4) X_1 + (\beta_2 + \beta_5) X_2 + e$$

b. Determine the fitted regression equations for smokers and nonsmokers separately using the fitted regression model.

$$\widehat{Y}_{\text{NS}} = 48.613 + 1.029 X_1 + 10.451 X_2$$

and

$$\widehat{Y}_{\text{S}} = 48.076 + 1.467 X_1 + 6.744 X_2$$

c. Test for parallelism of the two models, stating the null hypothesis in therms of appropriate regression coefficients.

How can Y_{NS} and Y_{S} be parallel?

We test H_0: $\beta_4 = \beta_5 = 0$ vs. H_A: $\beta_4 = 0$ and/or $\beta_5 = 0$ using the multiple-partial F-test:

$$
\begin{aligned}
& F(X_1Z, X_2Z | X_1, X_2, Z) \\
&= \frac{\text{SSR}(X_1Z, X_2Z | X_1, X_2, Z)/2}{\text{SSE}(X_1, X_2, Z, X_1Z, X_2Z)/26} \\
&= \frac{\{\text{SSR}(X_1, X_2, Z, X_1Z, X_2Z) - \text{SSR}(X_1, X_2, Z)\}/2}{\text{SSE}(X_1, X_2, Z, X_1Z, X_2Z)/26} \\
&= \frac{\{4,915.63033 - 4,889.82563\}/2}{1,510.33842/26} \\
&\approx 0.136
\end{aligned}
$$

which compared to $F_{2,26}$ gives p-value > 0.25, so we don't reject H_0 at the $\alpha = .05$ level.