BIOS 145 **Lab #5** **2/11/98**

- 1st hour: KKMN Ch.7: #3, 5; Ch.8: #2, 10
- 2nd hour: questions

KKMN Ch.7: #3 p.108:

Using the data of Pr. 2, Ch.5:

a) Determine the ANOVA table for QUET (Y) regressed on AGE (X).

Since only one predictor (X) in the model, $k = 1$ model d.f.

Since there are $n = 32$ observations, $n - 1 = 31$ total d.f.

So there are $31 - 1 = 30$ error d.f. Then,

$$\text{MSE} = S^2_{Y|X} = .09079 \Rightarrow \text{SSE} = (.09079)(30) = 2.7237$$

$$\begin{aligned} \text{SSY} = 7.65968 &\Rightarrow \text{SSR} = 7.65968 - 2.7237 = 4.93598 \\ &\Rightarrow \text{MSR} = 4.93598/1 = 4.93598 \end{aligned}$$

So the ANOVA table is given by:

Source	DF	SS	MS	Var. ratio
Regression	1	4.94	4.94	54.36
Error	30	2.72	.091	
Total	31	7.66		

b) Use the tables in (a) to perform the F-test for the significance of the straight-line regression.

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{4.93598}{.09079} \approx 54.36 > 4.16 = F_{1,31,.95}$$

Thus, we reject H_0: $\beta_1 = 0$ at the $\alpha = .05$ level.

KKMN Ch.7: #5 p.108:

Using the data of Pr. 3, Ch. 5:

a) Determine the ANOVA table for the regression of TIME (Y) on INC (X)

Since only one predictor (X) in the model, $k = 1$ model d.f.

Since there are $n = 20$ observations, $n - 1 = 19$ total d.f.

So there are $19 - 1 = 18$ error d.f. Then,

$$\text{MSE} = S^2_{Y|X} \approx 110.16 \Rightarrow \text{SSE} = (110.16)(18) = 1982.88$$

$$\begin{aligned} \text{SSY} \approx 2433.78 &\Rightarrow \text{SSR} = 2433.78 - 1982.88 = 450.90 \\ &\Rightarrow \text{MSR} = 450.90/1 = 450.90 \end{aligned}$$

So the ANOVA table is given by:

Source	DF	SS	MS	Var. ratio
Regression	1	450.90	450.90	4.09
Error	18	1982.88	110.16	
Total	19	7.66		

b) Use the tables in (a) to perform the F-test for the significance of the straight-line regression.

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{450.90}{110.16} \approx 4.09 < 4.41 = F_{1,18,.95}$$

Thus, we cannot reject H_0: $\beta_1 = 0$ at the $\alpha = .05$ level.

c) Compare the value of the F-statistic obtained in (b) with the value of T^2, the square of the test statistic required in part (e) of Pr. 3, Ch. 5.

$$T = \frac{\widehat{\beta}_1 - 0}{S_{\widehat{\beta}_1}} = \frac{.000714}{.00035281} \approx 2.023 < 2.101 = t_{18,.975}$$

so $T^2 \approx 4.09 = F$

d) The p-values for the F-test in part (b) and for the t-test in part (e) of Pr. 3, Ch. 5 are the same. Intuitively, why does this make sense?

See back of book.

KKMN Ch.8: #2 p.125:

For 53 patients, level of pathology (Y) regressed on thinking disturbance (X_1) and hostile suspiciousness (X_2).

a) The least-squares equation is

$$\widehat{Y} = -.0635 + 23.451X_1 - 7.073X_2$$

Using this, determine predicted level of pathology for a pt. with 2.80 on thinking disturbance and 7.0 on hostile suspiciousness.

$$\widehat{Y}_5 = -.0635 + 23.451(2.80) - 7.073(7.0) \approx 16.09$$

How does this compare to the value actually obtained for pt. 5?

$\widehat{Y}_5 \approx 16.09$ and $Y_5 = 25$ so it was not a very good estimate.

b) Determine R^2 values for each of these models. If you use R^2 and model simplicity as selection criteria, which model appears to be best?

(1) Y on X_1: SSY $= 13791.1698$, SSE $= 12255.3128$

model d.f. $= 1$, error d.f. $= 53 - (1 + 1) = 51$

$$\text{SSR} = 13791.1698 - 12255.3128 = 1535.857$$

$$R^2 = \frac{\text{SSR}}{\text{SSY}} = \frac{1535.857}{13791.1698} \approx .1114$$

(2) Y on X_2: $\text{SSE} = 13633.3225$

$$\text{model d.f.} = 1, \text{ error d.f.} = 53 - (1 + 1) = 51$$

$$\text{SSR} = 13791.1698 - 13633.3225 = 157.8473$$

$$R^2 = \frac{\text{SSR}}{\text{SSY}} = \frac{157.8473}{13791.1698} \approx .0114$$

(3) Y on X_1 and X_2: $\text{SSE} = 11037.2985$

$$\text{model d.f.} = 2, \text{ error d.f.} = 53 - (2 + 1) = 50$$

$$\text{SSR} = 13791.1698 - 11037.2985 = 2753.8713$$

$$R^2 = \frac{\text{SSR}}{\text{SSY}} = \frac{2753.8713}{13791.1698} \approx .20$$

Model (3) seems to be the "best" model here. While not the simplest, it does explain almost twice the variation as the next best model (model (1)).

<u>KKMN Ch.8: #10 p.131:</u>

House price (Y) regressed on house size (X_1) and no. of bedrooms (X_2).

(a) Determine the l.s. estimates for the model with both predictors.

From the SAS output given,

$$\widehat{Y} = -16.093 + 5.722X_1 - 1.173X_2$$

so the predicted average price of a 2,500 sq. ft. home that has four bedrooms.

Here, $X_1 = 25$ and $X_2 = 4$ so

$$\widehat{Y} = -16.093 + 5.722(25) - 1.173(4) = 122.27$$

(b) Compare the R^2-value for the regression in (a) with the r^2 value in Pr. 22 of Ch. 6.

$$R^2 = \frac{\text{SSR}}{\text{SSY}} = \frac{5733.32}{6305.71} \approx .9092$$

$$r^2 = .9091 \text{ (see p.83)}$$

Since adding X_2 only slightly increases R^2, it does not appear useful to add X_2 to a model already containing X_1 in predicting the house price.