Name: *Caleb McWhorter — Solutions*
MAT 222
Fall 2019
Homework 8

"[Miranda] You have no style or sense of fashion. [Andy] I think that depends on— [Miranda] No, no, that wasn't a question."
–Miranda Priestly & Andy Sachs, Devil Wears Prada

Problem 1: Education researchers are interested in studying equity in the SAT and ACT in college acceptances. The SAT and ACT are both supposedly exams to test students mastery of High School material and thus gauge their college readiness. Because both these exams test similar material, researchers are wondering how strong the relationship is between them. Specifically, researchers are testing if there is a predictive linear relationship between the two by testing if one can use a student's SAT score to predict their ACT score. To test this hypothesis, the group takesx a SRS of 60 students who took both the SAT and ACT. The data is summarized below:

$$\begin{array}{lll} \overline{x} = 912.667 & s_x = 180.112 & \sum(x_i - \overline{x})^2 = 1913973 \\ \overline{y} = 21.133 & s_y = 4.714 & \sum(y_i - \overline{y})^2 = 1310.93 \\ r = 0.8167 & s = 2.74353 & \\ b_0 = 1.63 & b_1 = 0.02137 & \end{array}$$

(a) For this sample, what was the average SAT score? What was the average ACT score?

The average SAT score was 912.7 *while the average ACT score was* 21.133.

(b) What was the resulting linear model for their statistical analyses?

$$\texttt{ACT} = 1.63 + 0.02137\ \texttt{SAT}$$

(c) Use s_x to confirm the value $\sum(x_i - \overline{x})^2 = 1913973$.

$$\begin{aligned} s_x^2 &= \frac{1}{n-1}\sum(x_i - \overline{x})^2 \\ 180.112^2 &= \frac{1}{9-1}\sum(x_i - \overline{x})^2 \\ 180.112^2 \cdot 59 &= \sum(x_i - \overline{x})^2 \\ \sum(x_i - \overline{x})^2 &= 1913980 \end{aligned}$$

(d) What is the standard error for b_0?

$$SE_{b_0} = s\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum(x_i - \overline{x})^2}} = 2.74353\sqrt{\frac{1}{60} + \frac{912.667^2}{1913973}} = 1.84423$$

(e) Create a 99% confidence interval for the coefficient b_1 and interpret the result.

We have $n = 60$ so that the degrees of freedom are $n - 2 = 58$ (so we use dof 50) (this gives $t^* = 2.678$). We also have standard error

$$SE_{b_1} = \frac{s}{\sqrt{\sum (x_i - \overline{x})^2}} = \frac{2.74353}{\sqrt{1913973}} = 0.00198309$$

Then we compute $b_1 \pm t^* SE_{b_1} = 0.02137 \pm 2.678(0.00198309)$ to find confidence interval $(0.016, 0.027)$. Therefore, we are 99% confident that the coefficient b_1 is between 0.016 and 0.027, i.e. we are 99% confident that for every 100 point increase in a student's SAT score the student will see a corresponding 1.6 to 2.7 increase in their ACT score, on average.

(f) What is the value of the coefficient of determination for this model? What does it tell you?

The coefficient of determination is $r^2 = 0.8167^2 = 0.6670$. This gives the 'percent linearity' in the data, i.e. the percent of the variability in the data explained by the model.

(g) What is the value of the MSE for this model?

We have MSE= $s^2 = 2.74353^2 = 7.52696$.

(h) What is the value of SSE for this model?

We know that $MSE = \frac{SSE}{DFE}$ so that $SSE = MSE \cdot DFE$. But then $SSE = 7.52696 \cdot 58 = 436.564$.

(i) What is the value of SST?

We know SST= $\sum (y_i - \overline{y})^2$ so that we have SST= 1310.93.

(j) What is the value of SSM?

We know that $SST = SSM + SSE$. But then $1310.93 = SSM + 436.564$ so that $SSM = 874.366$.

(k) Use SE_{b_1} to confirm that $\sum(x_i - \overline{x})^2 = 1913973$.

We have $SE_{b_1} = \dfrac{s}{\sqrt{\sum(x_i - \overline{x})^2}}$. *But then we have* $\sum(x_i - \overline{x})^2 = \dfrac{s^2}{SE_{b_1}^2}$. *Then*

$$\sum(x_i - \overline{x})^2 = \frac{2.74353^2}{0.00198^2} = 1919950$$

(l) Create a 90% confidence interval for the ACT score of a student who receives an SAT score of 1100. Interpret the result.

Using the model, we have predicted take-home exam score of $\hat{y} = 1.63 + 0.02137(1100) = 25.137$. *Because* $n = 60$, *we have degrees of freedom* $n - 2 = 58$ *(so we use dof 50), giving* $t^* = 1.676$. *We have standard error*

$$SE_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum(x_i - \overline{x})^2}} = 2.74353\sqrt{1 + \frac{1}{60} + \frac{(1100 - 912.667)^2}{1913973}} = 2.79113$$

Computing $\hat{y} \pm t^* SE_{\hat{y}} = 25.137 \pm 1.676(2.79113)$ *gives confidence interval* $(20.46, 29.81)$. *Therefore, we are 90% certain that, on average, a student receiving an SAT score of 1100 will receive an ACT score between 20.46 and 29.81.*

(m) What is the *mean* ACT score for a student who receives an SAT score of 1100?

This is precisely what the model predicts, $\hat{\mu} = 1.63 + 0.02137(1100) = 25.137$.

(n) If the sample consisted of students who only received between 500 and 1200, should one use the model to predict the ACT scores for students whom received an SAT score of 1600? Explain.

No. Linear models should be used on values which are 'in the range of' values used to construct the model. Using values far from these numbers can lead to false conclusions. This is interpolation vs extrapolation.

Problem 2: College loan debt is a major drain on the US national economy. In an effort to try to predict future collection student loan debt, a government research group is attempting to predict the average amount of debt students exiting from different universities will have. They begin by analyzing debt from public universities. Using data from Kiplinger's "Best Values in Public Colleges", they try to predict average debt based on admittance rate, graduation rate, total in-state cost & aid, and total out-of-state cost & aid using data from 25 different schools. All currency was measured in tens of thousands of dollars. A summary of the model data is found below.[1]

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	*5*	2.38156	*0.47631*	2.43	0.073
Admit	*1*	1.29861	1.29861	6.62	0.019
GradRat	*1*	*0.29454*	0.29454	1.50	0.235
TCostInS	*1*	0.28563	0.28563	*1.46*	0.242
InCostAid	*1*	0.00085	0.00085	0.00	0.948
TCostOut	*1*	0.33604	0.33604	1.71	0.206
Error	*19*	*3.72817*	*3.72817*		
Total	*24*	6.10974			

Model Summary

S	R-sq	R-sq (adj)	R-sq (pred)
0.442967	*38.98%*	22.92%	0.00%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.6819	*0.728*	0.94	0.361	
Admit	1.803	0.701	2.57	0.019	1.20
GradRate	*0.925*	0.755	1.23	0.235	1.67
TCostInS	0.585	0.484	*1.21*	0.242	4.33
InCostAid	-0.027	0.410	-0.07	0.948	1.90
TCostOut	-0.275	0.210	-1.31	0.206	2.79

The regression equation is

$$\texttt{AvgDebt} = 0.6819 + 1.803\ \texttt{Admit} + 0.925\ \texttt{GradRat} + 0.585\ \texttt{TCostInS} - 0.027\ \texttt{InCostAid} - 0.275\ \texttt{TCostOut}$$

[1] Note because there is a perfect linear relationship between out-of-state aid and the other variables, Minitab removes it from the model.

(a) Fill in the missing entries in the table.

(b) What does the model say the average loan debt is for a student that went to a public school with a admittance rate of 47 percent, graduation rate of 67 percent, total in-state cost of $19,000, in-state aid of $15,000, total out-of-state cost of $31,500, and total out-of-state aid of $24,000?

$$\textit{AvgDebt} = 0.6819 + 1.803(0.47) + 0.925(0.67) + 0.585(19) - 0.027(15) - 0.275(31.5) = 4.197$$

That is, the predicted average debt for this type of student is $41,970.

(c) What is the coefficient of determination for this model?

$$r^2 = 0.3898$$

(d) Construct a 95% confidence interval for β_3.

We have $n = 25$ and 5 variables so that gives degrees of freedom $n - p - 1 = 25 - 5 - 1 = 19$, the degrees of freedom of the error (DFE). This gives $t^* = 2.093$. We know that β_3 is the variable 'TCostInS' and that $b_3 = 0.585$ with standard error $SE_{b_2} = 0.484$. Therefore, we compute $b_3 \pm t^* SE_{b_2} = 0.585 \pm 2.093(0.484)$ to find confidence interval $(-0.428012, 1.598)$

(e) Perform the F-test for this model using a significance level of $\alpha = 0.05$.. State your null and alternative hypotheses, F-statistic, degrees of freedom of the numerator/denominator, p-value, and conclusion.

We have hypotheses

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_a : \text{ not all } \beta_i = 0 \end{cases}$$

We already found F-statistic 2.43 with p-value 0.073. [Note the degrees of the freedom of the numerator is $p = 5$ and the degrees of freedom of the denominator is $n - p - 1 = 19$.] Therefore, we fail to reject the null hypothesis, i.e. the model is not statistically significant.