Name: *Caleb McWhorter — Solutions*
MAT 222
Fall 2019
Chapter 10–11 Worksheet

"Fool me once, shame on you. But teach a man to fool me and I'll be fooled for the rest of my life"

–BoJack Horseman, BoJack Horseman

Problem 1: Answer the following questions about simple linear regressions (SLRs) and multilinear regressions (MLRs).

(a) What is the (Pearson) correlation coefficient? What are its possible values?

The correlation coefficient is r and $-1 \leq r \leq 1$.

(b) What is the coefficient of determination? What does it represent? What are its possible values?

The coefficient of determination is r^2. It represents the proportion of variability in the data explained by the model. We always have $0 \leq r^2 \leq 1$.

(c) What does it mean if $r^2 = 1$, i.e. if $r = 1$ or $r = -1$? For a SLR, what does it mean if $r > 0$? What about if $r < 0$?

If $r^2 = 1$, then all the data points in the sample lie perfectly along some line. For a SLR, if $r > 0$ then the variables are positively correlated, and if $r < 0$ the variables are negatively correlated.

(d) The value of s is the estimate of the standard deviation of the errors, ϵ, for a linear regression, i.e. the standard deviation for the values $y_i - \hat{y}_i$. This value also represents the estimated standard deviation about the regression line. [True/False.]

The statement is True. Note if the residual $y - \hat{y} > 0$ then the value is above the regression line, and otherwise is below the regression line.

(e) If there are outliers in the data used to construct a linear regression, the conclusions from the model must be interpreted with care. [True/False.]

The statement is True.

(f) Confidence intervals for a response (or mean response) are narrowest when they are nearest to which of the following: near the ends of the data, near the middle, or they are the same width throughout a data set.

The are narrowest near the middle of the data: note for responses, we have $SE_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum(x_i - \overline{x})^2}}$, *which is smallest if* x^* *is 'close' to* $\overline{x}$*, and for mean responses,* $SE_{\hat{\mu}} = \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum(x_i - \overline{x})^2}}$, *and this term is smallest if* x^* *is 'close' to* $\overline{x}$.

(g) One of the ways of determining how 'good' a linear model is by examining the correlation coefficient r. The corresponding parameter is ρ. To test $H_0 : \rho = 0$ against $H_a : \rho \neq 0$, one uses $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ with degrees of freedom df $= n - 2$. [True/False.]

The statement is True.

(h) For a MLR, if one performs an F-test and rejects the null hypothesis, this means all the coefficients in the model are nonzero. [True/False.]

No. This means at least one of the coefficients in the model is nonzero.

(i) If one is performing a MLR constructed from 34 observations and using 6 predictors, what is the degrees of freedom for the numerator and denominator for the corresponding F-test?

The degrees of freedom of the numerator are the number of variables used in the model, i.e. the DFM, 6. The degrees of freedom of the denominator are the degrees of freedom of the error, i.e. the DFE, $n - p - 1 = 34 - 6 - 1 = 27$.

(j) The first step in performing a SLR is plotting the data and seeing if the data is approximately linear, i.e. that there is no curvature in the data. Furthermore, the model should be primarily used for values 'between' values used to construct the model. [True/False.]

The statement is True. Note using data 'between' the values used to construct the model to predict values is called interpolation while values outside this range is called extrapolation.

Problem 2: Many patients suffering from a throat injury experience speech pathologies. Researchers at a rehabilitation facility are trying to estimate the total recovery time for patients based on exercises that they have the patients perform. For instance, after 3 weeks of therapy, they have patients perform pitch exercises where measure how long, on average, patients are able to maintain certain pitches. At the end of a patient's treatment, they record how long the patients total rehabilitation time was. The researchers then try to use a SLR to predict the total recovery time from this early treatment exercise. The data from their model is summarized in the table below.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	666.8	*666.771*	71.26	0.000
Tone Length	*1*	*666.8*	*666.771*	*71.26*	*0.000*
Error	*13*	*121.6*	9.357		
Total	14	*788.4*			

Model Summary

S	R-sq	R-sq (adj)	R-sq (pred)
3.05888	*84.57%*	83.38%	78.27%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	24.23	1.66	*14.58*	*0.000*	
Tone Length	-1.543	*0.183*	-8.44	0.000	1.00

(a) Fill in the missing entries in the ANOVA table above. [You may show your work for these entries below.]

(b) How many observations were used to create the model?

We know DFT $= n - 1$*, so a total of 15 observations were used.*

(c) According to this model, is 'tone length' positively or negatively correlated with recovery time? Explain.

Note that $b_1 < 0$ *so the model suggests that these are negatively correlated.*

(d) What percent of variability in the response variable 'recovery time' is explained by this model?

This is the coefficient of determination, r^2*. So this is 84.57%.*

(e) In this case, the general model is recovery time $= \beta_0 + \beta_1 \cdot$ tone length $+ \epsilon$. [True/False]

The statement is True.

(f) Report the least-square regression equation for predicting recovery time from tone length.

We have $\hat{y} = b_0 + b_1 x^*$*, so recovery time* $= 24.23 - 1.543$ *tone length.*

(g) Does the constant term β_0 have meaning in this context? Explain.

The constant term is when all the variables are set to 0*. [So if* 0 *is not a possible value for all the variables, the constant term cannot have meaning in the context.] Here this is when the tone pitch is sustained for* 0 *seconds. Then* β_0 *represents the average recovery time for either a patient that cannot sustain this pitch at all or all patients with general throat injuries (because value* 0 *could also mean they did not take the test.*

(h) Construct a 98% confidence interval for β_0.

We have $b_0 = 24.23$ *and* $SE_{b_0} = 1.66$*. The degrees of freedom are* $DFE = 13$ *so that* $t^* = 2.650$*. Then we have*

$$24.23 \pm 2.650(1.66) = 24.23 \pm 4.399 \rightsquigarrow (19.831, 28.629)$$

(i) Construct a 98% confidence interval for β_1. Interpret the result.

We have $b_1 = -1.543$ *and* $SE_{b_1} = 0.183$*. The degrees of freedom are* $DFE = 13$ *so that* $t^* = 2.650$*. Then we have*

$$-1.543 \pm 2.650(0.183) = -1.543 \pm 0.485 \rightsquigarrow (-2.03, -1.06)$$

We are 98% certain that, on average, for every second longer a patient can sustain the pitch, their total recovery time will be between 2.03 and 1.06 weeks less.

(j) Test the hypothesis $H_0 : \beta_1 = 0$ against $H_a : \beta_1 < 0$ using a significance level of 5%. Interpret your results. What is the critical value for this test?

We have

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_a : \beta_1 < 0 \end{cases}$$

The degrees of freedom are $DFE = 13$*, so that with* $\alpha = 0.01$*, we have critical value* -1.771*. We have* $t = -8.44$ *(less than the critical value) with* p*-value 0.000. Therefore, we reject the null hypothesis. There is sufficient evidence to suggest that recovery time is negatively correlated with tone length.*

(k) Write the null and alternative hypotheses for the F-test for this regression. Then perform this F-test using $\alpha = 0.05$, being sure to give your degrees of freedom, critical value, test statistic, p-value, and conclusion. Compare this to the previous part.

We have

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_a : \beta_1 \neq 0 \end{cases}$$

The degrees of freedom are $(1, 13)$*, i.e. the degrees of freedom of the numerator are 1 and the degrees of freedom of the denominator are 13. Using a significance value of* $\alpha = 0.05$*, we have critical value* 4.75*. The* F*-statistic is* $F = 71.26$*, which is greater than the critical value, with* p*-value 0.000. Therefore, we reject the null hypothesis. There is sufficient evidence to suggest that the coefficient for tone length is not zero, i.e. there is some (linear) relationship between tone length and recovery time. The* F*-test would be the same as the* t*-test* $H_0 : \beta_1 = 0$ *against* $H_a : \beta_1 \neq 0$ *with* $\alpha = 0.05$*.*

A summary of the data obtained by the researchers is given below.

Variable, x	N	Mean, $\overline{x}$	StDev, s	Variance, s^2	Sum, $\sum x_i$	Sum of Squares, $\sum x_i^2$
Tone Length	15	8.00	4.47	20.00	120.00	1240.00
Length Therapy	15	11.88	7.50	56.31	178.26	2906.78

(l) Find $\sum(x_i - \overline{x})^2$. [Hint: $s_x^2 = \frac{1}{n-1}\sum(x_i - \overline{x})^2$.]

We know $s_x^2 = 20.00$ *and* $n = 16$*. Then* $\sum(x_i - \overline{x})^2 = 15 \cdot 20.00 = 300$*.*

(m) Construct a 95% confidence interval for the average recovery time for patients that can only sustain a pitch for 5 seconds.

We have $x^* = 5$*. The model says* $\hat{\mu} = 24.23 - 1.543(5) = 16.515$*. We also know* $n = 16$*,* $s = 3.05888$*,* $(x^* - \overline{x})^2 = (5-8)^2 = 9$*, and* $\sum(x_i - \overline{x})^2 = 300$*. Then*

$$SE_{\hat{\mu}} = 3.05888\sqrt{\frac{1}{16} + \frac{9}{300}} = 0.930$$

We have degrees of freedom DFE $= 13$ *so that* $t^* = 2.160$*. Then*

$$16.515 \pm 2.160(0.930) = 16.515 \pm 2.009 \rightsquigarrow (14.506, 18.524)$$

Therefore, we are 95% certain that patients that can sustain the pitch for 5 seconds, on average, have a recovery time of between 14.5 and 18.5 weeks.

(n) Construct a 95% confidence interval for the recovery time for a patient that can only sustain a pitch for 5 seconds.

We have $x^* = 5$*. The model says* $\hat{\mu} = 24.23 - 1.543(5) = 16.515$*. We also know* $n = 16$*,* $s = 3.05888$*,* $(x^* - \overline{x})^2 = (5-8)^2 = 9$*, and* $\sum(x_i - \overline{x})^2 = 300$*. Then*

$$SE_{\hat{\mu}} = 3.05888\sqrt{1 + \frac{1}{16} + \frac{9}{300}} = 3.197$$

We have degrees of freedom DFE $= 13$ *so that* $t^* = 2.160$*. Then*

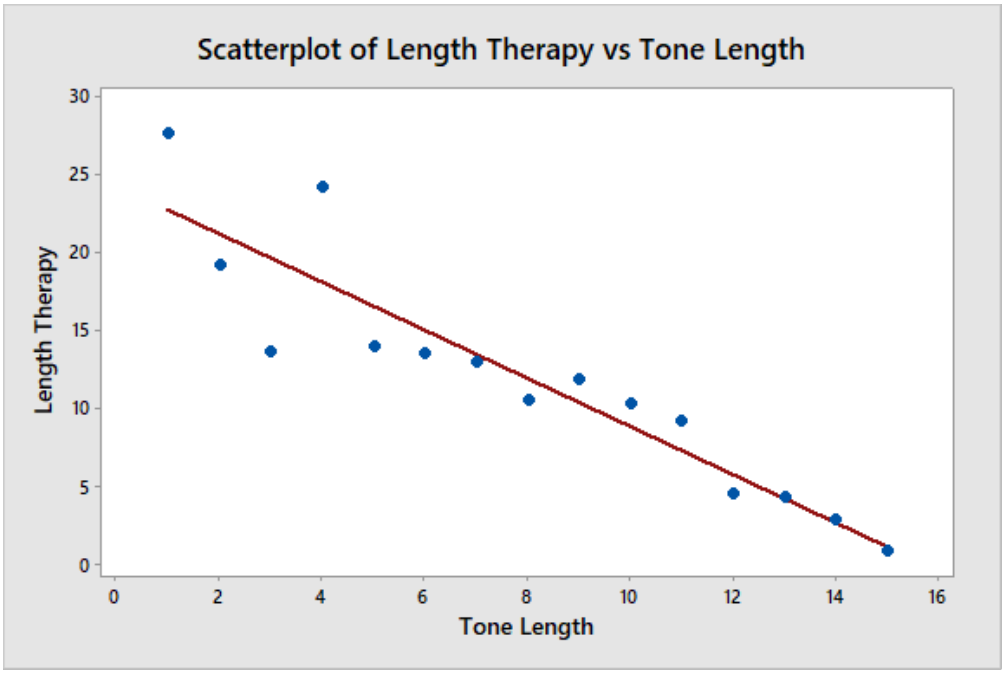
$$16.515 \pm 2.160(3.197) = 16.515 \pm 6.906 \rightsquigarrow (9.609, 23.421)$$

Therefore, we are 95% certain that the recovery time for a patient that can sustain the pitch for 5 seconds have a recovery time of between 9.61 and 23.4 weeks.

(o) What is the difference between (m) and (n)? Explain.

Part (m) predicts the average recovery time while (n) predicts the recovery time, not on average across all patients, for individual patients. This explains why the confidence interval in (n) is larger than the one in (m)—individuals differ more than groups.

(p) Below is a scatterplot of the data. Is a SLR appropriate? Explain. Sketch the line of best fit. From this sketch, is the data positively or negatively correlated? Do you expect r^2 to be close to 0 or 1? Do you expect r to be close to -1, 1, or 0?



The data seems to be approximately linear, with no curvature; therefore, a SLR seems to be appropriate. Examining the regression line, we expect the data to be negatively correlated. Because the line fits the data "well", we expect r^2 to be close to 1. This means that r is close to ± 1. Because the variables are negatively correlated, we expect r to be close to -1.

Problem 3: A sports rehabilitation facility frequently treats female athletes with leg issues relating to muscle injury. Researchers at the facility are trying to predict the total recovery time using a number of factors including age, BMI, total weekly hours of physical therapy (PT), and their leptin levels (a hormone related to body fat and the angiogenesis, aka the development of blood vessels). Their model data is summarized below.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	*4*	473.92	*118.480*	3.06	0.038
Age	*1*	*20.42*	20.425	0.53	0.475
BMI	*1*	289.90	289.899	*7.49*	0.012
Weekly PT Hours	*1*	3.74	*3.741*	0.10	0.759
Leptin Level (ng/mL)	*1*	176.23	176.233	4.55	0.044
Error	*22*	851.91	*38.723*		
Total	26	*1325.83*			

Model Summary

S	R-sq	R-sq (adj)	R-sq (pred)
6.22278	*35.75*%	24.06%	10.67%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	58.5	14.4	*4.08*	0.001	
Age	0.216	*0.298*	0.73	0.475	1.26
BMI	*-1.840*	0.672	-2.74	0.012	1.20
Weekly PT Hours	-0.92	2.95	-0.31	0.759	1.23
Leptin Level (ng/mL)	-0.705	0.330	-2.13	0.044	1.11

(a) Fill in the missing entries in the ANOVA table above. [You may show your work for these entries below.]

(b) How many observations were used to create the model?

We know that DFT $= n - 1$ *so that* $n = 27$.

(c) What proportion in the response variable 'recovery time' is explained by this model?

This is the coefficient of determination, r^2. *Therefore, this is 0.3575.*

(d) Report the least-square regression equation (EBF) for predicting recovery time from age, BMI, weekly PT hours, and leptin levels.

$$\textit{Recovery time} = 58.5 + 0.216\,\textit{Age} - 1.840\,\textit{BMI} - 0.92\,\textit{PT Hours} - 0.705\,\textit{Leptin level}$$

(e) Construct a 95% confidence interval for β_3. Interpret your results.

The coefficient β_3 *corresponds to the variable 'weekly PT hours'. We know that* $b_3 = -0.92$ *and* $SE_{b_3} = 2.95$. *The degrees of freedom are DFE* $= 22$ *so that we have* $t^* = 2.074$. *Then*

$$-0.92 \pm 2.074(2.95) = -0.92 \pm 6.1183 \rightsquigarrow (-7.038, 5.198)$$

Therefore, we are 95% certain that, on average and all other factors equal, every extra hour of weekly PT time spent causes between a 7 week decrease to a 5.2 week increase in recovery time.

(f) Test the hypothesis $H_0 : \beta_4 = 0$ against $H_a : \beta_4 \neq 0$ using $\alpha = 0.05$. Be sure to state the critical value, test statistic, p-value, and conclusion.

The coefficient β_4 *corresponds to the variable 'leptin level'. The degrees of freedom are DFE* $= 22$ *so that the critical value is 2.074. We know* $b_4 = -0.705$, $t = -2.13$, *and the* p*-value is 0.044. Therefore, we reject* H_0. *There is sufficient evidence to suggest that the variable 'leptin level' is significant.*

(g) Perform the F-test for this regression at a significance level of 10%. Be sure to state the null and alternative hypotheses, the critical value, test statistic, p-value, and the conclusion.

We have

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \\ H_a : \textit{Not all } \beta_i = 0 \end{cases}$$

The degrees of freedom are $(4, 22)$, i.e. the degrees of freedom of the numerator are 4 and the degrees of freedom of the denominator are 22. Using a significance level of $\alpha = 0.10$, we have critical value 2.22. We know that $F = 3.06$ with p-value 0.038. Therefore, we reject the null hypothesis. There is sufficient evidence to suggest that not all the regression coefficients are 0.

(h) Which variables for this model are significant? [Use $\alpha = 0.05$.] Which are not? Explain.

Examining the p-values for the individual coefficients, it seems that 'BMI' and 'leptin level' are significant while 'age' and 'weekly PT hours' are not.

(i) Compare parts (g) and (h), explaining how they are similar, and how they are different.

In part (g), we found that at least one of the variables in the model is significant, but not which. In part (h), we found which variables were significant.

(j) Note that the p-value for 'weekly PT hours' in the multilinear regression reported above is 0.759. This implies that if another researchers gathers their own data, using the same experimental procedure, and runs the SLR model *Recovery time* $= \beta_0 + \beta_1 \cdot$ *Weekly PT hours* $+ \epsilon$, then the p-value for the t-test for β_1 must always be greater than 0.05. [True/False.]

False. The significance (or lack thereof) depends on the other variables present. So if more variables are included or if variables are removed, the p-value for a given coefficient can change in either direction. Some variables are only good predictors in the presence of other variables, and other times variables are bad predictors in the presence of other variables.

(k) For the test described in the previous part, what are the degrees of freedom?

We would still have $n = 27$ but the number of variables would be 1. Therefore, the degrees of freedom would be DFE $= n - p - 1 = 27 - 1 - 1 = 27 - 2 = 25$.

(l) Find the predicted average recovery time for a patient aged 24, with BMI 18.1 and leptin levels 0.201, and that spends 3 hours a week in PT.

$$\textit{Recovery time} = 58.5 + 0.216(24) - 1.840(18.1) - 0.92(3) - 0.705(0.201) = 27.48 \textit{ weeks}$$

(m) If a patient with the values described in (l) has an average recovery time of 26 weeks, find the residual.

$$e = y - \hat{y} = 26 - 27.48 = -1.48$$

Problem 4: Answer the following questions:

(a) What plot can be used to assess the fit of a regression line?

The residual plot.

(b) For a SLR, the ANOVA F-value is equal to the square of which value?

$$F = t_1^2 = \left(\frac{b_1}{SE_{b_1}}\right)^2$$

(c) Is it possible in a MLR to have an F-statistic with p-value $p < 0.05$, but none of the t-tests significant ($p > 0.05$)?

Yes, this is possible. It could be that none of the variables are significant, i.e. 'good' predictors, but somehow the model as a whole is a 'good' predictor.

(d) Does adding more variables to a linear regression always improve the model?

No. Adding more variables could make the previous predictors worse or better. There is no set effect for all models.

(e) If $\beta_i \approx 0$, then there is no relationship between the variable corresponding to β_i and the response variable. [True/False]

The statement is False. It could be that the true value for β_i is simply very close to 0. It could also be that the variables are strongly related but just not linearly related, i.e. that there is some nonlinear relationship between the variables.

(f) A linear regression is most appropriate when points in the residual plot are randomly dispersed about the horizontal line at 0. [True/False]

The statement is True.