Name: ______________________________________

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?

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

(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 $r < 0$?

(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.]

(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.]

(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.

(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 $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ with degrees of freedom df $= n - 2$. [True/False.]

(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.]

(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?

(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.]

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  _______    71.26    0.000
              Tone Length __  _____   _______    _____    _____
            Error         __  _____     9.357
            Total         14  _____

Model Summary

                      S       R-sq  R-sq (adj)  R-sq (pred)
               ________  _______%      83.38%       78.27%

Coefficients

            Term           Coef  SE Coef  T-Value  P-Value   VIF
            Constant      24.23     1.66    _____    _____
            Tone Length  -1.543    _____    -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?

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

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

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

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

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

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

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

(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?

(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.

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

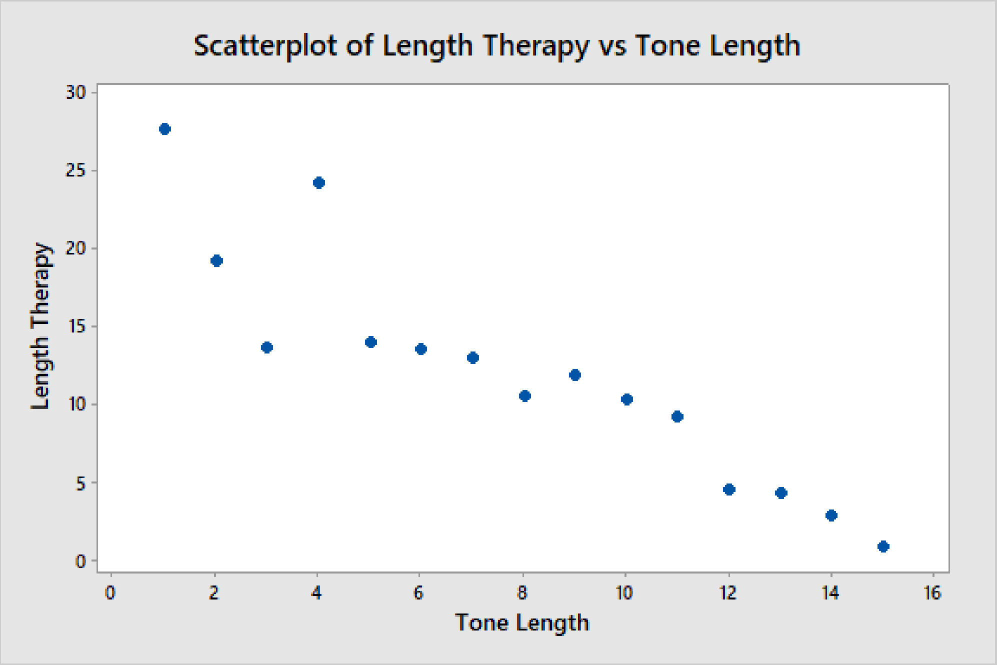
(l) Find $\sum(x_i - \overline{x})^2$. [Hint: $s_x^2 = \frac{1}{n-1}\sum(x_i - \overline{x})^2$.]

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

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

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

(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?



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            __   473.92  _______     3.06    0.038
                Age                 __   ______   20.425     0.53    0.475
                BMI                 __   289.90  289.899     ____    0.012
                Weekly PT Hours     __     3.74   ______     0.10    0.759
                Leptin Level (ng/mL) __  176.23  176.233     4.55    0.044
              Error                 __   851.91   ______
              Total                 26  _______

Model Summary

                    S      R-sq  R-sq (adj)  R-sq (pred)
             _______  _______%      24.06%       10.67%

Coefficients

               Term                    Coef  SE Coef  T-Value  P-Value   VIF
               Constant                58.5     14.4     ____    0.001
               Age                    0.216   ______     0.73    0.475  1.26
               BMI                  _______    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?

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

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

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

(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.

(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.

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

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

(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.]

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

(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.

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

Problem 4: Answer the following questions:

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

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

(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 ($p > 0.05$)?

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

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

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