Name (Last, First):

Student ID:

Assignment 13

METCS544A3A4_F2024

Instructions:

1. For answering programming questions, please use Adobe Acrobat to edit the pdf file in two steps **[See Appendix: Example Question and Answer]**:
    a. Copy and paste your R code as text in the box provided (so that your teaching team can run your code);
    b. Screenshot your R console outputs, save them as a .PNG image file, and paste/insert them in the box provided.
    c. Show all work—credit will not be given for code without showing it in action, including a screenshot of R console outputs.
2. To answer non-programming questions, please type or handwrite your final answers clearly in the boxes. Show all work - credit will not be given for numerical solutions that appear without explanation in the space above the boxes. **You're encouraged to use R to graph/plot the data and produce numerical summaries; please append your code and screenshot of the outputs at the end of your PDF submission.**

[Total 120 pts = 105 pts + 15 Extra Credit pts]

Grading Rubric

Each question is worth 3 points and will be graded as follows:

3 points: Correct answer with work shown

2 points: Incorrect answer but attempt shows some understanding (work shown)

1 point: Incorrect answer but an attempt was made (work shown), or **correct answer without explanation (work not shown)**

0 points: Left blank or made little to no effort/work not shown

Reflective Journal [3 pts]

(Copy and paste the link to your live Google doc in the box below)

Part I. Goodness of Fit Test (12 pts)

Are more babies born on a specific day of the week than others? To determine if the distribution of births across the week happened in different proportions than expected, a researcher took a random sample of 84 births in the year from a local hospital and recorded what day they were born on. The data is given in the following table:

Day of the week	Sun	Mon	Tue	Wed	Thu	Fri	Sat
# of births	6	7	9	14	19	17	12

Based on these data, is it reasonable to conclude that the proportion of births is not the same for all days of the week? Use α = 0.05.

Answer:

Part II. Chi-Square Test for Homogeneity (12 pts)

A study at a university wanted to see if there was a gender difference with respect to drinking behavior. Two independent random samples of male and female college students at the university asked them to rate their drinking behavior as none, low, moderate, or high. The results are shown in the table below.

	Drinking Level			
	None	Low	Moderate	High
Male	140	478	300	63
Female	180	580	285	40

a) What would be the null and alternative hypotheses to test to see if there was a gender difference with respect to drinking behavior?

b) What would the expected counts be for a male with a moderate drinking level? Show your work!

c) You run the test and get a chi-square statistic of 15.157. What is the p-value?

d) What can you conclude at the 5% significance level?

Part III. Chi-Square Test and the Follow-Up Analysis (15 pts)

Does the treatment of a stress fracture in a foot affect the success or failure of healing the bone? A recent experiment in a medical journal took four separate random samples of various treatment methods used to treat a fractured foot. In each of these random samples, they recorded whether the patient saw success in the healing of the fracture. A Chi-Square test for Homogeneity was performed and the follow-up analysis is given below.

	Success	Failure
Surgery	54 50.471 0.247	12 15.529 0.802
Weight-Bearing Cast	41 51.235 2.045	26 15.765 6.645
Non-Weight Bearing Cast for Less Than 6 Weeks	17 19.118 0.235	8 5.882 0.762
Non-Weight Bearing Case for 6 Weeks	70 61.176 1.273	10 18.824 4.136

Key:
Observed
Expected
Contribution

a) What would be the null and alternative hypotheses in this situation?

b) Show how the value of “15.529” under “surgery, failure” was obtained.

c) What would be the chi-square statistic and p-value for this test?

d) What is your conclusion in the context of the problem, at the 1% significance level?

e) Identify the two largest contributions to the chi-square statistic. What do these contributions imply in the context of the problem?

Part IV. Chi-Square Test for Association/Independence (21 pts)

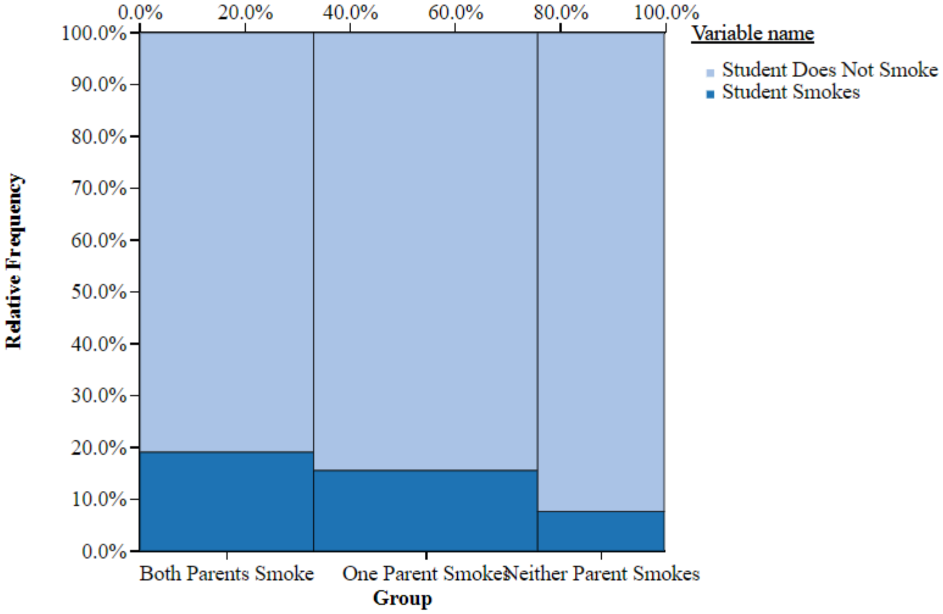
Are the smoking habits of college students related to their parents' smoking habits? Below are data from a survey of students gathered across five different colleges.

	Both Parents Smoke	One Parent Smokes	Neither Parent Smokes
Student Smokes	300	316	88
Student Does Not Smoke	1280	1723	1068

a) How can the data be gathered so that we would perform a chi-square test for homogeneity?

b) How can the data be gathered so that we would perform a chi-square test for association/independence?

c) Below is a mosaic plot of the data. In a mosaic plot, the width of each vertical bar represents the relative size of each parent smoking habit category. Are the categories (Both Parents Smoke, One Parent Smokes, Neither Parent Smokes) equally represented in the survey? Explain.



d) What would be the null and alternative hypotheses from a chi-square test for association/independence?

e) Using you calculator, find the expected counts, chi-square statistic, degrees of freedom, and the p-value.

Expected Counts	Both Parents Smoke	One Parent Smokes	Neither Parent Smokes
Student Smokes			
Student Does Not Smoke			

χ^2 = ___________

df = ___________

P-value = _______

f) Based on your p-value, what can you conclude?

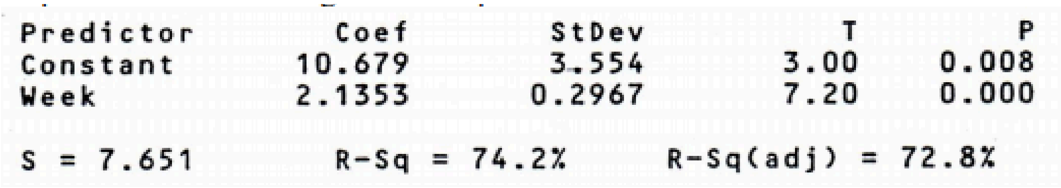
g) Which cell in the table contributed the most to the chi-square statistic? How does this information expand on your solution in part (f)?

Part V. Inference for Quantitative Data: Slopes (45 pts)

1) **(12 pts)** A service center for kitchen appliances has hired a large group of new technicians. During the first 20 weeks of their employment, data were collected documenting the number of appliances serviced by a random sample of these technicians.

Week	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Number	12	21	18	13	28	20	29	35	26	20	36	40	28	52	32	32	55	56	49	60

Here is part of the R regression output for these data:



```
Predictor        Coef       StDev          T         P
Constant       10.679       3.554       3.00     0.008
Week           2.1353      0.2967       7.20     0.000

S = 7.651       R-Sq = 74.2%     R-Sq(adj) = 72.8%
```

(a) Interpret the slope of the regression in the context of the problem.

Answer:

(b) What is the equation for the LSRL line?

Answer:

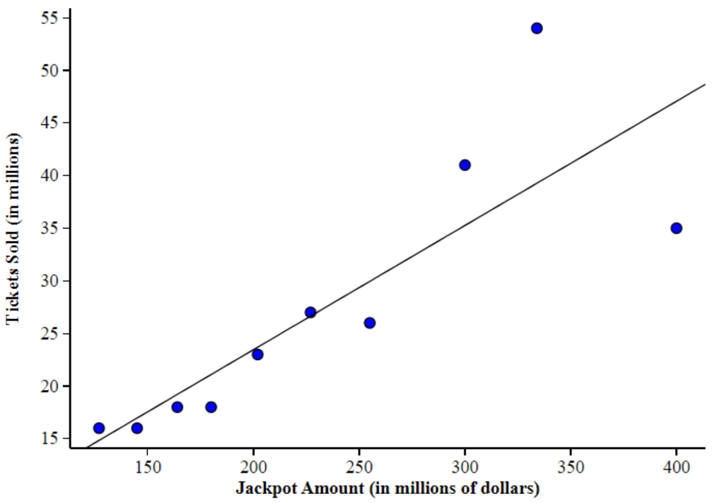
(c) Using the R output, determine a 99% confidence interval for the true slope of the regression line. Just carry out the calculation

Answer:

(d) Interpret the meaning of the confidence interval in the context of the problem. What does this interval suggest about the employment of the new technicians?

Answer:

2) **(12 pts)** The largest winning jackpot in Mega Millions history was $1.537 billion, for the October 23, 2018, drawing, in which a single winning jackpot ticket was sold in South Carolina. Ticket sales leading up to the Jackpot being won were unprecedented. A statistician was interested in determining if the number of tickets sold for a Mega Millions drawing depended on how large the Jackpot was. They randomly selected 10 Mega Millions drawings in history and recorded how many ticket sales there were for that drawing. The data below shows the result of the investigation.



r = 0.836

$s_x = 88.57$

Regression Analysis

Equation	n	s
$\hat{y} =- 0.147 + 0.118$	10	7.288

T-Test for Slope

$H_0 : \beta = 0$

$H_a : \beta > 0$

(a) What is the standard error of the slope of the regression line? Interpret it in the context of the problem.

Answer:

(b) What do you know (given the information about) or what would you need to know in order to perform a T-Test for the slope of the regression line?

Answer:

(c) Assuming that the conditions needed for doing inference for regression are present, what is the correct t-test statistic for the listed hypothesis?

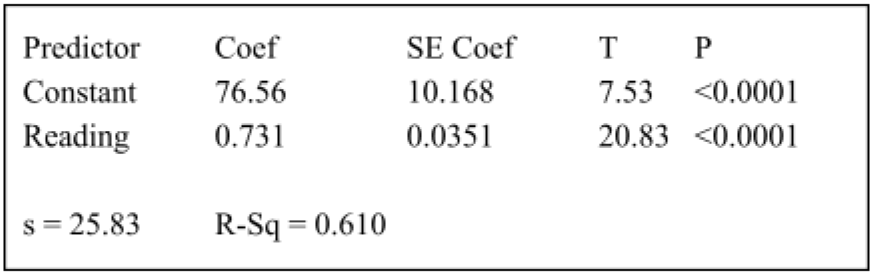
Answer:

(d) Find and interpret the p-value with the test statistic you calculated in part (c). What is the conclusion to this test in the context of the problem? Use a significance level of 5%.

Answer:

3) **(21 pts)**

Data was gathered on a random sample of 25 sophomores at your school. Each sophomore’s score on a standardized chemistry exam and their score on a standardized reading exam were taken. School officials wanted to see if a sophomore’s reading score (in points) could help predict their chemistry score (in points).



```
Predictor    Coef       SE Coef     T       P
Constant     76.56      10.168      7.53    <0.0001
Reading      0.731      0.0351      20.83   <0.0001

s = 25.83    R-Sq = 0.610
```

a) Identify the value of the standard deviation of the residuals and interpret this value in the context of the problem.

b) Identify the value of the estimated standard deviation of the slope and interpret this value in the context of the problem.

c) What would be a null and alternative hypothesis for a hypothesis test based on the computer output above?

d) What would the value of the parameter, β, measure in this test?

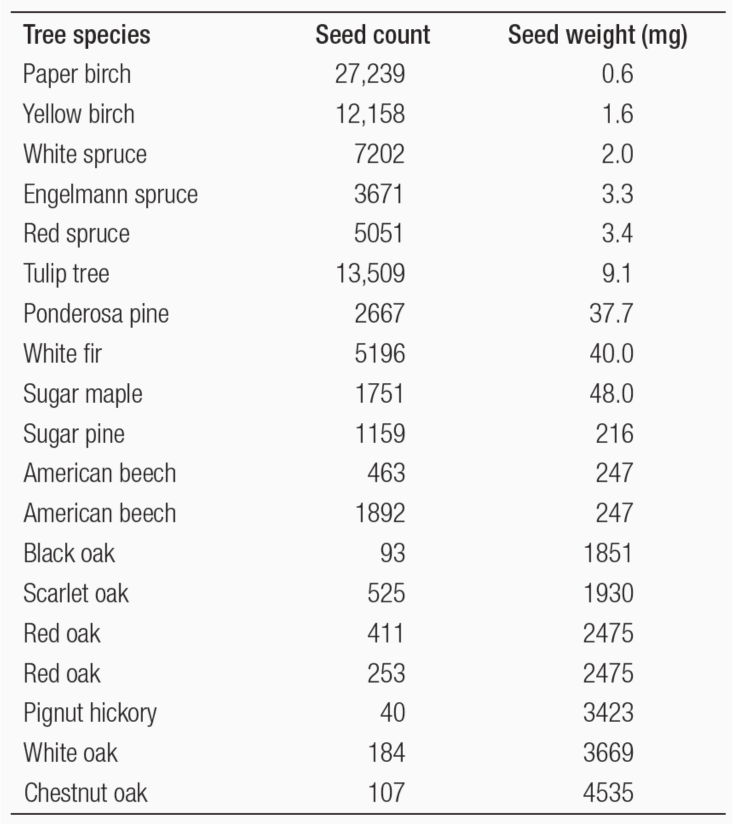
e) Explain how a t-test statistic of 20.83, and a p-value of <0.0001 were obtained.

f) Create a 99% confidence interval using the data from the computer output above.

g) What would be the correct conclusion of your significance test in the context of the problem? Explain how your results in (e) and (f) agree on this conclusion.

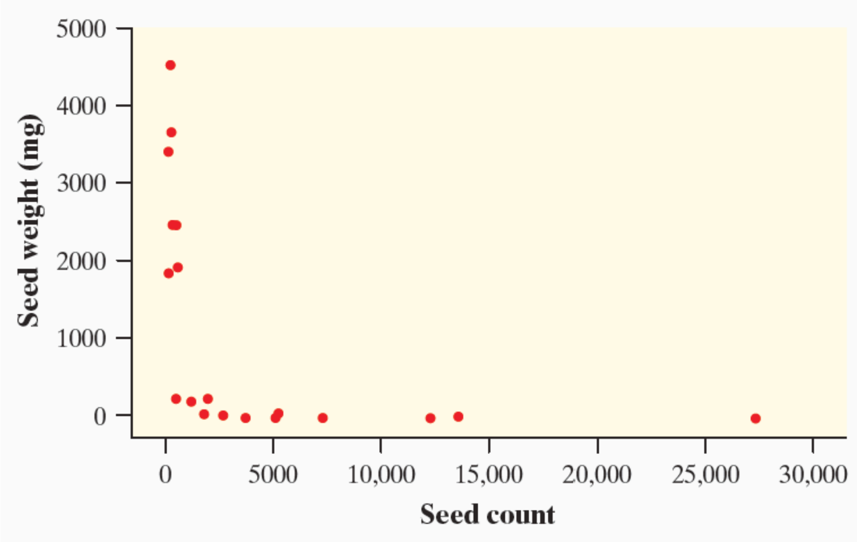
Part VI. Extra Credit Question (12 pts)

The following table gives data on the mean number of seeds produced in a year by several common tree species and the mean weight (in milligrams) of the seeds produced. Two species appear twice because their seeds were counted in two locations. We might expect trees with heavy seeds to produce fewer of them, but what mathematical model best describes the relationship?



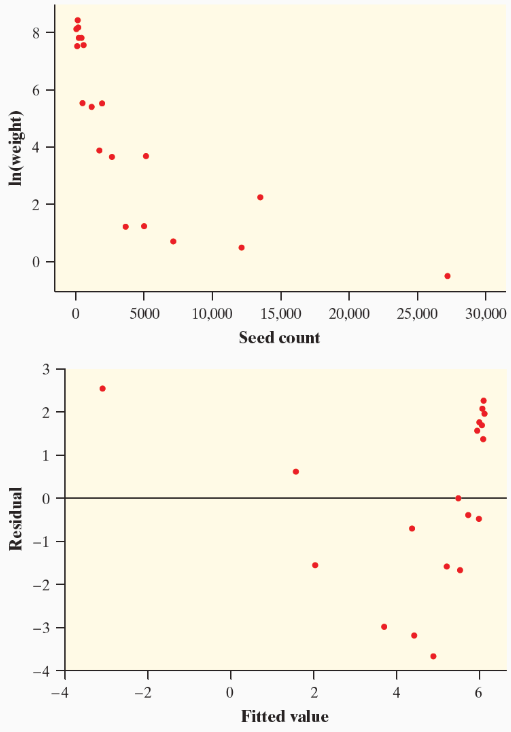
Tree species	Seed count	Seed weight (mg)
Paper birch	27,239	0.6
Yellow birch	12,158	1.6
White spruce	7202	2.0
Engelmann spruce	3671	3.3
Red spruce	5051	3.4
Tulip tree	13,509	9.1
Ponderosa pine	2667	37.7
White fir	5196	40.0
Sugar maple	1751	48.0
Sugar pine	1159	216
American beech	463	247
American beech	1892	247
Black oak	93	1851
Scarlet oak	525	1930
Red oak	411	2475
Red oak	253	2475
Pignut hickory	40	3423
White oak	184	3669
Chestnut oak	107	4535

(a) Based on the scatterplot below, is a linear model appropriate to describe the relationship between seed count and seed weight? Explain.



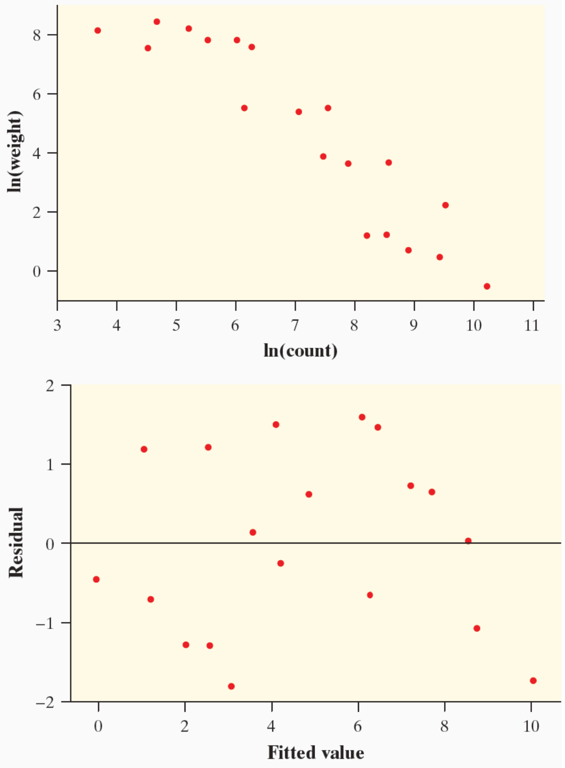
(b) Two alternative models based on transforming the original data are proposed to predict the seed weight from the seed count. Graphs and computer output from a least-squares regression analysis on the transformed data are shown below.

Model A:



```
Predictor Coef          SE Coef    T      P
Constant    6.1394      0.5726     10.72  0.000
Seed       −0.00033869 0.00007187 −4.71  0.000
Count
S=2.08100    R-Sq=56.6%    R-Sq(adj)=54.1%
```

Model B:



```
Predictor  Coef     SE Coef  T        P
Constant   15.491   1.081     14.33   0.000
ln(count)  −1.5222  0.1470   −10.35   0.000
S=1.16932    R-Sq=86.3%    R-Sq(adj)=85.5%
```

Which model, A or B, is more appropriate for predicting seed weight from seed count? Justify your answer.

(c) Using the model you chose in part (b), predict the seed weight if the seed count is 3700.

(d) Interpret the R-squared value of your model.

Answer:

THE END