Math 222: Exam 2
Spring – 2019
04/12/2019
50 Minutes

Name: *Caleb M^{c}Whorter — Solutions*

Write your name on the appropriate line on the exam cover sheet. This exam contains 9 pages (including this cover page) and 4 questions. Check that you have every page of the exam. Answer the questions in the spaces provided on the question sheets. Be sure to answer every part of each question and show all your work. If you run out of room for an answer, continue on the back of the page — being sure to indicate the problem number.

Question	Points	Score
1	12	
2	16	
3	20	
4	32	
Total:	80	

1. (12 points) Mark each of the following statements as True (T) or False (F).

(a) ___T___ A χ^2-analysis is used to analyze categorical variables.

(b) ___T___ If a variable used in a regression is not significant, then the value of its coefficient must be 'approximately' 0.

(c) ___F___ If the coefficient for a variable in a regression is approximately 0, then the variable is insignificant for the model.

(d) ___F___ The degrees of freedom for a χ^2-test with 8 rows and 6 columns is 48.

(e) ___T___ In performing an F-test for a regression which examined 5 variables and 31 people, if the F-statistic was 3.72, then using the F-table the p-value would be 0.025.

(f) ___T___ For an F-test with 2 degrees of freedom in the numerator and 366 degrees of freedom in the denominator, if the F-statistic was 3.72, then the p-value is 0.050.

(g) ___F___ For a regression with an F-test having 6 degrees of freedom in the numerator and 37 degrees of freedom in the denominator, 43 observations must have been used in making the regression.

(h) ___T___ For a χ^2-goodness of fit test, there is only one row in the expected value table.

(i) ___F___ For a χ^2-goodness of fit, the degrees of freedom is the number of columns.

(j) ___F___ The null hypothesis for an F-test with two variables is $H_0 : \beta_1 = \beta_2 = 0$ while the alternative hypothesis is $H_a : \beta_1 = \beta_2 \neq 0$.

2. A guidance counselor at a high school is trying to determine if there is a relationship between gender and whether a graduating student ends up choosing a STEM or a non-STEM major upon graduating and entering college. They collect data which is summarized in the table below.

	STEM	Non-STEM	Total
Female	91	75	166
Male	70	84	154
Total	161	159	320

(a) (8 points) Complete the expected value table (on the left) and the χ^2-contribution table (on the right).

	STEM	Non-STEM
Female	*83.52*	*82.48*
Male	*77.48*	*76.52*

	STEM	Non-STEM
Female	*0.6701*	*0.6786*
Male	*0.7224*	*0.7314*

(b) (6 points) Perform a χ^2-test for this analysis by giving the null and alternative hypothesis (in the problem context), the degrees of freedom, test statistic, p-value, and stating your conclusion. Use $\alpha = 0.05$.

$$\begin{cases} H_0: & \text{there is no association between gender and choice of major} \\ H_a: & \text{there is an association between gender and choice of major} \end{cases}$$

degrees of freedom: 1

test statistic: $X^2 = 2.803$

p-value: 0.094

Conclusion: *We fail to reject the null hypothesis. There is probably not an association between gender and choice of major.*

(c) (2 points) What are the assumptions for a χ^2-test? Does this χ^2-test meet these assumptions?

The assumptions are that the average expected value is at least 5 and that each expected value is at least 1. Because all of the values in the expected value table from (a) are greater than 5, these conditions are certainly met.

3. Certain isotopes of sodium are used in cardiovascular traces. A researcher has an unknown isotope of sodium. Starting with 1 g of the substance, they record the natural log of how much remains over a period of approximately 36 hours. Using algebraic manipulation, the slope of the linear model can be used to determine the half-life, and thus the isotope. The data for their linear fit is found below.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	6.5721	6.57207	1663.86	0.000
Time	1	6.5721	6.57207	1663.86	0.000
Error	27	0.1066	0.00395		
Total	28	6.6787			

Model Summary

S	R-sq	R-sq (adj)	R-sq (pred)
0.0628482	98.40%	98.34%	98.11%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.0178	0.0254	0.70	0.488	
Time	-0.04742	0.00116	-40.79	0.000	1.00

(a) (1 point) How many data points were used in constructing the model?

We know DFT$= n - 1$*. Therefore,* $n - 1 = 28$ *so that* $n = 29$*.*

(b) (1 point) What percent of variability in the response variable 'Amount' is explained by this model?

This is the R^2 *value, which is* 98.40%*.*

(c) (2 points) What is value of the correlation coefficient?

We know that $r = \pm\sqrt{r^2} = \pm\sqrt{0.9840} = 0.992$*. But because* $b_1 < 0$*, we know that* $r < 0$ *(because* $r = b_1 s_x/s_y$ *and* $s_x, s_y > 0$*). Therefore,* $r = -0.992$*.*

(d) (2 points) What is the equation of the linear regression for this data?

$\texttt{Amount} = 0.0178 - 0.04742\ \texttt{Time}$

(e) (5 points) Construct a 95% confidence interval for β_0.

From the table, we know that $b_0 = 0.0178$ *and* $SE_{b_0} = 0.0254$*. Because the degrees of freedom* $n - 2 = 29 - 2 = 27$ *(the degrees of freedom of the error), we know that* $t^* = 2.052$*. Therefore, we have*

$$\begin{aligned} b_1 &\pm t^* SE_{b_1} \\ 0.0178 &\pm 2.052(0.0254) \\ 0.0178 &\pm 0.0024 \end{aligned}$$

which gives 95% confidence interval $(-0.0343, 0.0699)$.

(f) (3 points) Show that $\sum(x_i - \overline{x})^2 \approx 2935.4$.

We know $SE_{b_1} = \dfrac{s}{\sqrt{\sum(x_i - \overline{x})^2}}$*. From the table, we know* $SE_{b_1} = 0.00116$ *and* $s = 0.0628482$*. Therefore,*

$$\begin{aligned} SE_{b_1} &= \frac{s}{\sqrt{\sum(x_i - \overline{x})^2}} \\ 0.00116 &= \frac{0.531042}{\sqrt{\sum(x_i - \overline{x})^2}} \\ \sqrt{\sum(x_i - \overline{x})^2} &= \frac{0.0628482}{0.00116} \\ \sum(x_i - \overline{x})^2 &= 54.179^2 \\ \sum(x_i - \overline{x})^2 &= 2935.4 \end{aligned}$$

(g) (6 points) Construct a 96% confidence interval for the natural log of the amount of sodium left after 12 hours. [The average amount of hours used in the model was 19.4]

We have degrees of freedom $n - 2 = 27$ *(the degrees of freedom of the error), so that* $t^* = 2.158$*. Using the model, we have*

$$\widehat{y} = 0.0178 - 0.04742 \cdot 12 = -0.55124.$$

Finally, using the table and the previous parts, we know we have $s = 0.0628482$*,* $n = 29$*,* $x^* = 12$*,* $\overline{x} = 19.4$*, and* $\sum(x_i - \overline{x})^2 = 2935.4$*. But then*

$$SE_{\widehat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum(x_i - \overline{x})^2}} = 0.062848\sqrt{1 + \frac{1}{29} + \frac{(12 - 19.4)^2}{2935.4}} = 0.0645$$

This gives

$$\begin{aligned} \widehat{y} &\pm t^* SE_{\widehat{y}} \\ -0.55124 &\pm 2.158\,(0.0645) \\ -0.55124 &\pm 0.139 \end{aligned}$$

Therefore, a 96% confidence interval for $\widehat{y}$ *when* $x = 12$ *is* $(-0.690, -0.412)$.

4. An engineer is trying to predict the milage per gallon (MPG) of a new car they are designing. Using data collected from other cars, they create a predictive model based on the number of cylinders, the horsepower, and the weight of a car (all variables which can be controlled during construction). The model is (partially) summarized below.[1]

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	*3*	*16855.6*	5618.52	*313.06*	*0.000*
Cylinders	*1*	30.4	30.41	1.69	0.194
Horsepower	*1*	242.7	*242.7*	13.52	0.000
Weight	*1*	1209.1	1209.05	67.37	0.000
Error	*388*	6963.4	17.95		
Total	391	23819.0			

Model Summary

S	R-sq	R-sq (adj)	R-sq (pred)
4.23639	*70.77%*	70.54%	70.20%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	*45.737*	0.796	57.46	0.000	
Cylinders	-0.389	*0.299*	-1.30	0.194	5.66
Horsepower	-0.0427	0.0116	*-3.68*	0.000	4.36
Weight	-0.005272	0.000642	-8.21	*0.000*	6.49

The regression equation is

$$\texttt{MGP} = 45.737 - 0.389\ \texttt{Cylinders} - 0.0427\ \texttt{Horsepower} - 0.005272\ \texttt{Weight}$$

[1] Dua, D. and Graff, C. (2019) UCI Machine Learning Repository. http://archive.ics.edu/ml. Irvine, CA. University of California School of Information and Computer Science. Quinlan,R. (1993). Combining Instance-Based and Model-Based Learning. In Proceedings on the Tenth International Conference of Machine Learning, 236-243, University of Massachusetts, Amherst. Morgan Kaufmann.

(a) (15 points) Fill in the missing table entries for this model.

(b) (2 points) For this model, what is the value of $\sum(y_i - \overline{y})^2$?

This is exactly SST, which is given on the table as 23819.0.

(c) (2 points) For this model, what is the value of $\sum(y_i - \widehat{y}_i)^2$?

This is exactly SSE, which is given on the table as 6963.4.

(d) (6 points) Construct a 95% confidence interval for β_2.

Because the degrees of freedom is $n - p - 1 = 392 - 3 - 1 = 388$ (the degrees of freedom of the error), we have $t^* = 1.984$. Using the table, we have

$$\begin{aligned} b_2 &\pm t^* SE_{b_2} \\ -0.0427 &\pm 1.984(0.0116) \\ -0.0427 &\pm 0.0230 \end{aligned}$$

Therefore, a 95% confidence interval for the coefficient of 'Horsepower' is $(-0.066, -0.020)$.

(e) (3 points) Of the variables used, which are significant predictors for this regression model and which are not. Explain your answer.

From the p-values for the t-test, we know that 'Horsepower' and 'Weight' are significant predictors but 'Cylinders' is not.

(f) (4 points) Is this regression model significant? To answer this, perform an F-test for this regression model. Be sure to state the null and alternative hypotheses, the degrees of freedom of the numerator and denominator, F-value, p-value, and your conclusions at the 10% significance level.

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

The degrees of freedom of the numerator is the number of variables while the degrees of freedom of the denominator is the DFE. Therefore, we have degrees of freedom $(3, 388)$. We have already calculated that $F = 313.06$ and $p = 0.000$. Therefore, we reject the null hypothesis so that at least one of the coefficients is nonzero. Therefore, the model is significant.

BONUS. (8 points) Below is a partial ANOVA table for a linear regression model.

Source	DF	Adj SS	Adj MS	F-Value
Regression	*2*	7362.7	*3681.35*	16.43
Error	*48*	10753.8	*224.04*	
Total	50	18116.5		

Complete the table above. For credit, you must show all your computations in the space below.

We know that

$$F = \frac{MSM}{MSE} = \frac{SSM/DFM}{SSE/DFE} = \frac{SSM}{DFM} \cdot \frac{DFE}{SSE}$$

From this, we have $\frac{DFE}{DFM} = \frac{SSE}{SSM} F$. *But then we have*

$$DFE = \frac{SSE}{SSM} \cdot F \cdot DFM = \alpha\, DFM$$

where have defined $\alpha := SSE/SSM \cdot F = 10753.8/7362.7 \cdot 16.43 = 23.997$. *But we know also that* $DFM + DFE = DFT$. *However,*

$$\frac{DFE}{DFM} = \alpha$$
$$DFE = \alpha\, DFM$$

Therefore, using substitution

$$DFM + DFE = DFT$$
$$DFM + \alpha\, DFM = 50$$
$$DFM\,(1 + \alpha) = 50$$
$$DFM = \frac{50}{1 + \alpha}$$
$$DFM = \frac{50}{1 + 23.997}$$
$$DFM = 2.002$$

Then $DFM = 2$, *so that* $DFE = 48$. *Using MS– = SS–/DF–, we easily fill in the remaining two entries.*