Math 222: Exam 2 **Name:** *Caleb M^c^Whorter — Solutions*

Fall – 2019

11/07/2019

80 Minutes

Write your name on the appropriate line on the exam cover sheet. This exam contains 12 pages (including this cover page) and 5 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	16	
2	16	
3	27	
4	31	
5	10	
Total:	100	

Race	White	Hispanic/ Latino	African American	American Indian/ Alaska Native	Asian	Hawaiian/ Pacific Islander	Other	Two or More Races
Percentage of US Population	56.1	16.3	12.6	0.9	4.8	0.2	6.2	2.9

1. (16 points) Is the Syracuse University a microcosm of the United States population, or is population at SU different, i.e. more/less diverse? According to the 2010 US Census, the population of the United States, broken down by race, is given in the table at the top of the page.[1] On the other hand, according to the 2018 Syracuse University Fall Census,[2] the breakdown of the 13,175 domestic undergraduate students by race is as in the table below. To determine if the distribution of races at SU is representative of the US race distribution, perform a Chi-Square Goodness of Fit Test. Be sure to state your null and alternative hypotheses, the test statistic, p-value, and conclusion at the 1% significance level.

Race	White	Hispanic/ Latino	African American	American Indian/ Alaska Native	Asian	Hawaiian/ Pacific Islander	Other	Two or More Races
Number of SU Students	8662	1393	987	80	1043	9	503	498

We have

$$\begin{cases} H_0 : \textit{domestic SU undergrads fit the US race distribution} \\ H_a : \textit{domestic SU undergrads do not fit the US race distribution} \end{cases}$$

First, we calculate a table of expected values by taking the probability of each race and multiplying by the number of domestic undergraduates:

Race	White	Hispanic/ Latino	African American	American Indian/ Alaska Native	Asian	Hawaiian/ Pacific Islander	Other	Two or More Races
Expected SU Students	7391.18	2147.53	1660.05	118.575	632.4	26.35	816.85	382.075

Then we compute the contribution to χ^2:

Race	White	Hispanic/ Latino	African American	American Indian/ Alaska Native	Asian	Hawaiian/ Pacific Islander	Other	Two or More Races
Expected SU Students	218.503	265.1	272.881	12.5493	266.591	11.424	120.587	35.1727

Summing these, we find $X^2 = 1202.81$, so that with degrees of freedom $8 - 1 = 7$, we find $p \approx 0$. Therefore, we reject the null hypothesis. There is sufficient evidence to suggest the racial demographics at SU is different than the racial demographics of the US population.

[1] https://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf

[2] http://institutionalresearch.syr.edu/wp-content/uploads/2019/02/02-Syracuse-University-Student-Enrollment-by-Career-and-Ethnicity-Fall-2018-Census.pdf

2. The Kaiser Family Foundation regularly polls Americans to track opinions on the Affordable Care Act (ACA). They ask the following "As you may know, a health reform bill was signed into law in 2010. Given what you know about the health reform law, do you have a generally favorable or generally unfavorable opinion of it?" According to their October 2019 poll,[3] if one took a survey of 100 Americans, you would obtain data as in Table 1 on the next page.

   (a) (2 points) Write the null and alternative hypotheses for a Chi-Square Test for Association in the context of the problem.

   $$\begin{cases} H_0: & \text{there is no association between age and support for the ACA} \\ H_a: & \text{there is an association between age and support for the ACA} \end{cases}$$

   (b) (4 points) Fill in the missing values on the tables on the next page.

   (c) (3 points) What are the assumptions for a Chi-Square Test for Association? Does this test meet these requirements?

   The average expected value be at least 5, with each expected value at least 1. This test certainly meets those requirements.

   (d) (5 points) Write the statistics for the Chi-Square Test at the 5% significance level below:

   Degrees of Freedom: 6

   Critical Value: 12.59

   Test Statistic: $X^2 = 4.44$

   p-value range: $p > 0.25$

   (e) (2 points) Write your conclusion for the Chi-Square Test in the context of the problem, using $\alpha = 0.05$.

   We fail to reject the null hypothesis. The data is consistent with the fact that there is no association between age and support for the ACA.

Table 1: Counts of responses to the survey, broken down by age and opinion

Age/Opinion	Favorable	Unfavorable	Don't Know	Total
18–29	51	37	12	100
30–49	54	38	7	99
50–64	53	41	7	101
65+	43	44	11	98
Total	201	160	37	398

Table 2: Expected counts for the survey, assuming no association.

Age/Opinion	Favorable	Unfavorable	Don't Know
18–29	50.50	40.20	9.30
30–49	50.00	*39.80*	9.20
50–64	51.01	40.60	9.39
65+	*49.49*	39.40	9.11

Table 3: Contribution to χ^2

Age/Opinion	Favorable	Unfavorable	Don't Know
18–29	0.00	0.25	0.79
30–49	*0.32*	0.08	0.53
50–64	0.08	0.00	0.61
65+	0.85	*0.54*	0.39

[3] https://www.kff.org/interactive/kff-health-tracking-poll-the-publics-views-on-the-aca/

3. A Geriatric Rehabilitation Facility works with elderly patients that have suffered injuries from falls. As part of the rehabilitation, the patients perform balancing exercises. To understand how long the 'healthy' elderly patient should be able to perform the exercise, researchers at the facility administer the exercise to a number of people from a variety of ages and try to determine if one can predict how long one should be able to hold the balance pose using the patient's age. The results from their simple linear regression are found below.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	10514.8	10514.8	380.96	0.000
Age	1	10514.8	10514.8	380.96	0.000
Error	25	690.0	27.6		
Total	26	11204.8			

Model Summary

S	R-sq	R-sq (adj)	R-sq (pred)
5.25361	93.84%	93.60%	92.73%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	85.51	2.81	30.43	0.000	
Age	-1.0135	0.0519	−19.52	0.000	1.00

(a) (5 points) Fill in the missing entries in ANOVA table.

(b) (2 points) How many observations did the researchers use?

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

(c) (2 points) What percentage of the variation in 'Balance Time' is explained by the variable 'Age'?

This is the coefficient of determination, R^2*, which is 93.84%.*

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

We know $R = \sqrt{R^2} = \sqrt{0.9384} = \pm 0.96871$. *But because* $b_1 < 0$*, we know* $R = -0.96871$.

(e) (2 points) What is the least-square regression equation for predicting 'Balance Time' using 'Age'?

$\texttt{Balance Time} = 85.51 - 1.0135 \ \texttt{Age}$

(f) (2 points) What is the average predicted balance time for someone who is 74 years old?

Balance Time $= 85.51 - 1.0135 \cdot 74 = 10.511$

(g) (3 points) Use SE_{b_1} to show $\sum(x_i - \overline{x})^2 = 10246.6$.

We know $SE_{b_1} = \dfrac{s}{\sqrt{\sum(x_i - \overline{x})^2}}$. *Then* $\sum(x_i - \overline{x})^2 = \left(\dfrac{s^2}{SE_{b_1}}\right)^2$. *Then*

$$\sum(x_i - \overline{x})^2 = \left(\frac{5.25361}{0.0519}\right)^2 = 10246.6$$

(h) (4 points) Construct a 95% confidence interval for the average predicted balance time for someone who is 74 years old. [Note: the average age of the participants was 50.5.]

We have

$$SE_{\hat{\mu}} = 5.25361\sqrt{\frac{1}{27} + \frac{(74 - 50.5)^2}{10246.6}} = 1.58423$$

For a 95% confidence interval, using degrees of freedom 25, we have $t^* = 2.060$. *Then*

$$10.511 \quad \pm \quad 2.060(1.58423) \rightsquigarrow (7.25, 13.77)$$

(i) (5 points) At the 10% significance level, test the hypothesis $H_0 : \beta_1 = 0$ against $H_a : \beta_1 < 0$. For this test, state your critical value, test statistic, p-value, and conclusion. From this test, are 'Balance Time' and 'Age' positively or negatively correlated, or neither?

Using degrees of freedom 25, we have critical value -1.316. *From the table, we know the test statistic is* $t = -19.52$ *so that we have* $p \approx 0.000$. *Therefore, we reject the null hypothesis. There is sufficient evidence to suggest that* $\beta_1 < 0$*, i.e. that 'Balance Time' and 'Age' are negatively correlated.*

4. A university is reviewing the types of students that it admits to try to accept the best possible students. They examine whether a student's HS average, SAT Reading/Writing score, SAT Math score, and SAT Essay scores can be used to predict a student's success, measured by their final college GPA. They examine 22 students averages and create a multilinear regression, the model summary of which is found below.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	2.47957	0.619893	25.16	0.000
HS Average	1	0.02235	0.022345	0.91	0.354
Reading/Writing	1	0.00426	0.004259	0.17	0.683
Math	1	0.01442	0.014421	0.59	0.455
Essay	1	0.00296	0.002964	0.12	0.733
Error	17	0.41881	0.024636		
Total	21	2.89838			

Model Summary

S	R-sq	R-sq (adj)	R-sq (pred)
0.156959	85.55%	82.15%	75.45%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	−0.88	1.86	-0.47	0.642	
HS Average	0.0356	0.0374	0.95	0.354	34.71
Reading/Writing	0.00076	0.00182	0.42	0.683	22.33
Math	0.000715	0.000934	0.77	0.455	5.07
Essay	0.042	0.122	0.35	0.733	8.47

The regression equation is

GPA $= -0.88 + 0.0356$ HS Average $+ 0.00076$ Reading/Writing $+ 0.000715$ Math $+ 0.042$ Essay

(a) (9 points) Fill in the missing entries in the table for this model.

(b) (1 point) What is the coefficient of determination for this model?

The coefficient of determination is $R^2 = 0.8555$.

(c) (5 points) Create a 95% confidence interval for β_3. Interpret your result.

This is the third variable, Math Score. We know $b_3 = 0.000715$ *and* $SE_{b_3} = 0.000934$. *We have degrees of freedom 17 so that* $t^* = 2.110$. *Then*

$$0.000715 \pm 2.110(0.000934) = (-0.00125574, 0.00268574)$$

[Note scaling this interval by 100, we have $(-0.125574, 0.268574)$*.] Therefore, we are 95% certain that, on average, every 100 points more a student had in the SAT Math score resulted in between a 0.126 decrease to a 0.269 increase in their final college GPA.*

(d) (2 points) Predict the final GPA of an admitted student whose HS average was 93 with an SAT Reading/Writing, Math, and Writing scores of 620, 720, and 6, respectively.

`Balance Time` $= -0.88 + 0.0356(93) + 0.00076(620) + 0.000715(720) + 0.042(6) = 3.669$

(e) (1 point) If a student with scores in (d) actually had a college GPA of 3.232, find the residual.

$$y - \hat{y} = 3.232 - 3.669 = -0.437$$

(f) (7 points) Perform the ANOVA F-test for the regression. Be sure to state the null and alternative hypotheses, test statistic, p-value, and conclusion at the 5% significance level.

We have

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

We have $F = 25.16$ *with degrees of freedom* $(4, 17)$*, which gives a* p*-value of* $p \approx 0.000$*. Therefore, we reject the null hypothesis. Not all the* $\beta_i = 0$*, i.e. the model is significant.*

(g) (3 points) What variables in the model are significant in this model? Which variables are insignificant in this model?

Examining the p*-values for the coefficients, we see that none of the coefficients are significant. Therefore, all the variables are insignificant.*

(h) (3 points) Does the answer to (g) conflict with the answer to (f)? Explain why or why not.

They do not conflict. It is possible for the model to be significant while each of the variables are insignificant.

5. (10 points) Mark the following statements T (True) or F (False):

(a) ___F___ In a simple linear regression, a 95% confidence interval for the mean response at x^* has the largest width when $x^* = \overline{x}$.

(b) ___F___ In a multilinear regression, it is impossible to have the p-value for the F-statistic be less than 0.05 but have the p-values for every t-statistic for the coefficients be larger than 0.05.

(c) ___F___ If one rejects the null hypothesis in an ANOVA F-test for a multilinear regression, then all the model coefficient parameters β_i are nonzero.

(d) ___T___ A residual plot helps assess the fit of a regression line.

(e) ___T___ In a simple linear regression, the ANOVA F-statistic is always equal to the square of the t-statistic for b_1.

(f) ___T___ If the residual for one of the data points in a simple linear regression is negative, then the point lies below the regression line.

(g) ___F___ If one performs a linear regression and the model is not significant, that means there is no relationship between the response and explanatory variables.

(h) ___T___ The value of s is the estimate of the standard deviation about the regression line.

(i) ___T___ Adding more variables to a linear regression does not necessarily improve the model.

(j) ___F___ If one of the coefficients in a multivariable linear regression is insignificant, then removing it from the regression will improve the model.

BONUS. (10 points) Below is a partial ANOVA table for a linear regression model.

Source	DF	Adj SS	Adj MS	F-Value
Regression	*69*	11583.6	*167.8780*	3.884
Error	*155*	6700.18	*43.2270*	
Total	224			

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

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

where have defined $\alpha := SSE/SSM \cdot F = 6700.18/11583.6 \cdot 3.884 = 2.24658$. *But we know also that DFM + DFE = DFT. However,*

$$DFE = \alpha\, DFM$$

Therefore, using substitution

$$\begin{aligned} DFM + DFE &= DFT \\ DFM + \alpha\, DFM &= 224 \\ DFM\,(1+\alpha) &= 224 \\ DFM &= \frac{224}{1+\alpha} \\ DFM &= \frac{224}{1+2.24658} \\ DFM &= 68.9957 \end{aligned}$$

Then DFM $= 69$, *so that DFE* $= 155$. *Using MS– = SS–/DF–, we easily fill in the remaining two entries.*