Print Name: KEY

Math 127 – Exam 2 – Summer 2017

Version 61

REGRESSION PART – TAKE HOME

Oath: ***"I will not discuss the exam contents with anyone on planet Earth until the answer key is posted to Blackboard."***

Sign Name: Key

The penalty for cheating on this Exam is a grade of 0% for Math 127 Exam 2.

Student Instructions

1. This test is graded out of 50 points and counts for 10% of your Math 127 grade. Points are in parentheses for each question.
2. You may not work together. You may not use the Math Lab. Any clarifications need to be directed to your instructor. I know who hangs out together. I know who your friends are. I have spies. Don't work together.
3. Show work or points will be deducted. If you only report an answer and it is wrong, you will receive no credit.

DUE DATE: Wednesday, July 12th, when I get in my car around 7:52 PM. Turn in hard copy.

1. Use the "***2010-2012 Earnings by College Major***" dataset for this question.

"***Employed***" is the number of people who are employed (for each "***Major***")

"***Employed Full Time Year Round***" is the number of people who have full time jobs all year long (for each "***Major***")

1a. (1) Give the linear regression equation for predicting "***Employed Full Time Year Round***" based on "***Employed***".

$\hat{y}$ x

$\widehat{EFTYR} = -4215.85 + 0.786(EMPLOYED)$

1b. (3) Explain why the *y*-intercept is not an interpretable predicted value: If $x=0$ Employed, we would not expect a NEGATIVE number of people to employed year-round! Nonsense.

2. Use the "**2010 Hurricanes**" dataset for this question. Stronger hurricanes typically have higher winds and lower pressures.

"***Max Wind***" is the maximum wind speed measured in miles per hour

"***Pressure***" is the lowest recorded pressure measured in millibars

2a. (1) Give the linear regression equation for predicting the "***Pressure***" based on the "***Max Wind***".

y x

$\widehat{Pressure} = 1030.57 - 0.732(MAX\ WIND)$

2b. (3) Interpret the slope with a sentence in the context of the problem: For each extra 1 mph increase in wind, we expect Pressure to decrease by 0.732 mb.

zzz Retired

3. Use our "**Calendar Year 2017 Library Data**" dataset for this one.

"***Thickness***" is measured in inches

"***Pages***" is measured in, well, pages.

3a. (1) Give the linear regression equation if we use "***Thickness***" to predict the "***Pages***".

$\times$ 288.74 y

$\widehat{Pages} = \cancel{17.03} + \cancel{228.92}(Thickness)$

37.15

3b. (3) Interpret with a sentence in context, the value of R^2: ~~21.71%~~ 31.84% of the variation in Pages can be explained by knowing x = Thickness. ~~78.29%~~ 68.16% is (still unexplained.)

4. Use the "**Marvel vs. DC at the Box Office**" dataset for #4.

"***Foreign***" is the Box Office Revenue reported in millions of U.S. dollars.

Example: "*Avengers*" made $895.237 million = $895,237,000 at the box office in foreign countries.

"***Domestic***" is the box office revenue reported in millions of U.S. dollars for movie theatres in the U.S.A.

4a. (1) Cook up the linear regression equation if the explanatory variable is "***Domestic***" and the response variable is "***Foreign***".

$\hat{Foreign} = -15.78 + 1.334(Domestic)$

4b. (3) Interpret the value of S_e with a sentence in context.

On average, our predicted Foreign Revenues are off by 95.87 million when using x = Domestic Rev.

5. Use the "**Neighborhood**" dataset for this question.

"***Square Footage***" is the size of the house, measured in square feet.

"***Zillow Value***" is how much the website Zillow thinks a home is worth

5a. (1) Give the linear regression equation. We will use "***Square Footage***" to predict the "***Zillow Value***".

$\hat{Zillow} = 90340.38 + 62.94(Sq\ FT)$

5b. (3) Professor Kupe's coming back to Elkton! He buys the vacant lot in the ol' neighborhood and will build a 3000 square foot house. Predict its "***Zillow Value***" based on your equation.

Answer: $279,145.76

6. Use the "**Roller Coasters**" dataset on this one.

"***Height***" is the ground-to-top-of-first-hill measurement, in feet.

"***Speed***" is the maximum speed, measured in miles per hour.

6a. (1) Give the equation of the regression line. We will predict the "***Speed***" based on the "***Height***".

$\hat{Speed} = 36.62 + 0.184(Height)$

6b. (3) "***Magnum XL-200***" at Cedar Point has a residual of –2.36 mile per hour. Does this coaster have an unusually slow "***Speed***" for its "***Height***"? Yes / no and explain clearly for full credit.

No. The Studentized Residual is −0.49, does not exceed ±2.

ZZZ Retired

7. Use our "**Calendar Year 2017 Large Survey**" dataset for this problem.

We will use x = "***Work Time***" (in weekly hours) to predict your y = "***Salary***" (in yearly dollars).

7a. (1) We will run the regression analysis using all $n = 228$ data points. ~~But, based on a scatterplot, one student's reported values should probably be verified for accuracy (I hope it true and not just a comedian taking stats).~~

~~Suspicious Data Value~~ - "***Work Time***" = 40 "***Salary***" = $200,000

7b. (2) Linear Equation: ~~Salary = 1350.63 + 619.48 (Work Time)~~ 2248.45 + 584.45 (Work Time)

7c. (2) Using the correct units for both x and y, interpret the slope with a sentence in context:

For each extra 1 hour worked per week, we expect yearly salary to increase by ~~$619.48~~ $584.45

7d. (2) The y-intercept has meaning in the context of the problem. Interpret this point in context.

Someone who works 0 hours/week is expected to have a yearly salary of ~~$1350.63~~ $2248.45.

7e. (2) The point on the red line ~~(40, $26,129.83)~~ (40, $25626.52) has real meaning in the context of this problem. Interpret in context.

Someone who works 40 hours/week is expected to have a yearly salary of ~~$26,129.83~~ $25,626.52.

7f. (2) Person #94 didn't report a "***Salary***", but you can predict her "***Salary***". Do it. Show calculation here:

$\widehat{Salary} = $ ~~1350.63 + 619.48 (35)~~ 2248.45 + 584.45

$\approx$ ~~$23,032.43~~ $22,704.26

7g. (2) The person in row 223 has a residual of ~~$32,675.38~~ 33528.97. Interpret the value of this residual with a sentence in context:

Her salary of $65,000 is ~~$32,675.38~~ 33528.97 higher than expected for working a $x = 50$ hours per week. Nice.

7g. (2) After the survey was closed and the dataset was posted, Professor Kupe reveals his "***Salary***" to be $74,000.

He Works 40 Hours!

Show the calculation to arrive at the residual for his data point. Do not add this data point and recalculate your regression. Just use the regression equation as is:

Kupe's $x = 40$ hours gives $\hat{y} =$ ~~$26,129.83~~ 25626.52

So $e = y - \hat{y} = 74{,}000 - $ ~~$26,129.83~~

$e =$ ~~$47,870.17~~ $48373.48 ← Nice.

7h. (2) Interpret the value of S_e with a sentence in context: On average, our Predicted Salaries are off by ~ ~~$17,175~~ when using x = Hours Worked. 19777.98

7i. (2) Interpret the value of R^2 with a sentence in context: ~~25.73%~~ 18.92% of the variation in Salary is explained by knowing x = Work Time. ~~74.27%~~ 81.08% is still unexplained.

7k. (2) Describe with bullet points the relationship between "***Work Time***" and "***Salary***".

- Linear (ish) positive, weak, r = 0.435
- with the ~~one~~ 4 ish glaring outlier @ (~~40 hours, $200,000~~)

7l. (1) How many students have large positive Studentized residuals? ~~4~~ 9

7m. (1) How many students have large negative Studentized residuals? 0

7n. (1) How many students are official outliers for "***Work Time***"? ~~0~~ 1

7o. (1) How many students are official outliers for "***Salary***"? ~~9~~ 17

7p. (1) How many students have large Cook's Distances? ~~0~~ 14 ($4/433 = 0.009\,24$ ~~$= 0.017544$~~)

7q. (1) Why is it unwise to predict "***Salary***" for "***Work Time***" = 80 hours? Give the statistical reason.

x = 80 is extrapolation (essentially)