PID: _____

Last Name, First Name: _____

Section: _____

Approximate time spent to complete this assignment: _____ hour(s)

Homework 2
Math 11, UCSD, Winter 2018
Due on Tuesday, 23rd January

Readings: Chapters 7, 8 and 9.

Exercise 1

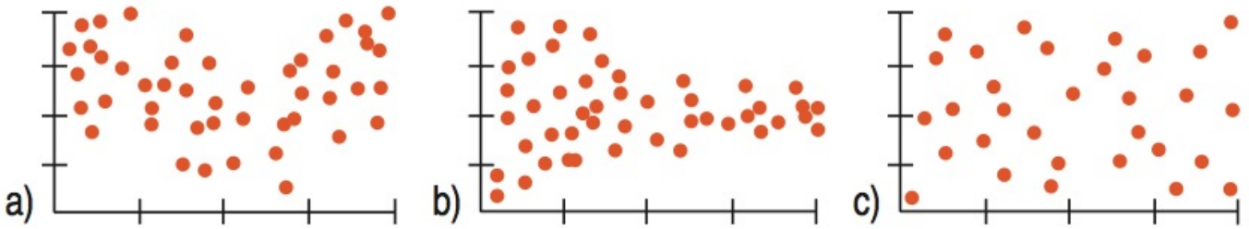
We examine the relationship between the fuel economy (*mpg*) and *Engine Size* for 35 models of cars. We produce the regression model $\widehat{mpg} = 36.25 - 3.867 Engine\ Size$.

1. In this context, what does it mean to say that a certain car has a positive residual?

2. Explain what the slope means.

3. The correlation between a car's engine size and its fuel economy (in mpg) is $R = -0.8476$. What fraction of the variability in fuel economy is accounted for by the engine size?

4. Tell what each of the residual plots below indicates about the appropriateness of the linear model that was fit to the data.



(a)

(b)

(c)

Exercise 2

A Sociology student investigated the association between a country's *Literacy Rate* and *Life Expectancy*, and then drew the conclusions listed below. Explain why each statement is incorrect. (Assume that all the calculations were done properly.)

1. The R^2 of 64% means that the *Literacy Rate* determines 64% of the *Life Expectancy* for a country.

2. The slope of the line shows that an increase of 5% in *Literacy Rate* will produce a 2-year improvement in *Life Expectancy*.

Exercise 3

Players in any sport who are having great seasons, turning in performances that are much better than anyone might have anticipated, often are pictured on the cover of *Sports Illustrated*. Frequently, their performances then falter somewhat, leading some athletes to believe in a "*Sports Illustrated* jinx." Similarly, it is common for phenomenal rookies to have less stellar second seasons (the so-called "sophomore slump"). While fans, athletes, and analysts have proposed many theories about what leads to such declines, a statistician might offer a simpler (statistical) explanation. Explain.

Exercise 4

Colleges use SAT scores in the admissions process because they believe these scores provide some insight into how a high school student will perform at the college level. Suppose the entering freshmen at a certain college have mean combined SAT Scores of 1222, with a standard deviation of 123. In the first semester, these students attained a mean GPA of 2.66, with a standard deviation of 0.56. A scatterplot showed the association to be reasonably linear, and the correlation between SAT score and GPA was 0.47.

1. Write the equation of the regression line.

2. Explain what the y-intercept of the regression line indicates.

3. Interpret the slope of the regression line.

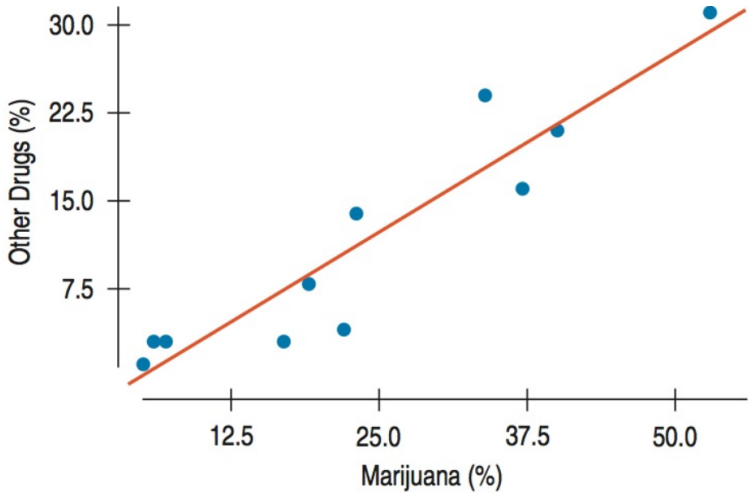
4. Predict the GPA of a freshman who scored a combined 1400.

5. Based upon these statistics, how effective do you think SAT scores would be in predicting academic success during the first semester of the freshman year at this college? Explain.

6. As a student, would you rather have a positive or a negative residual in this context? Explain.

Exercise 5

We examine results of a survey conducted in the United States and 10 countries of Western Europe to determine the percentage of teenagers who had used marijuana and other drugs. Below is the scatterplot. Summary statistics showed that the mean percent that had used marijuana was 23.9%, with a standard deviation of 15.6%. An average of 11.6% of teens had used other drugs, with a standard deviation of 10.2%.



1. Do you think a linear model is appropriate? Explain.

2. For this regression, R^2 is 87.3%. Interpret this statistic in this context.

3. Write the equation you would use to estimate the percentage of teens who use other drugs from the percentage who have used marijuana.

4. Explain in context what the slope of this line means.

5. Do these results confirm that marijuana is a "gateway drug", that is, that marijuana use leads to the use of other drugs?

Exercise 6

Wildlife researchers monitor many wildlife populations by taking aerial photographs. Can they estimate the weights of alligators accurately from the air? Here is a regression analysis of the Weight of alligators (in pounds) and their Length (in inches) based on data collected about captured alligators.

Dependent variable is Weight
R-*squared* = 83.6%
s = 54.01

Variable	Coefficient	SE(Coeff)	t-Ratio	P-Value
Intercept	−393.3	47.53	−8.27	<0.0001
Length	5.9	0.5448	10.8	<0.0001

1. Did they choose the correct variable to use as the dependent variable and the predictor? Explain.

2. What is the correlation between an alligator's length and weight?

3. Write the regression equation.

4. Interpret the slope of the equation in this context.

5. Do you think this equation will allow the scientists to make accurate predictions about alligators? What part of the regression analysis indicates this? What additional concerns do you have?

Exercise 7

Using data from 20 compact cars, a consumer group develops a model that predicts the stopping time for a vehicle by using its weight. You consider using this model to predict the stopping time for your large SUV. Explain why this is not advisable.

Exercise 8

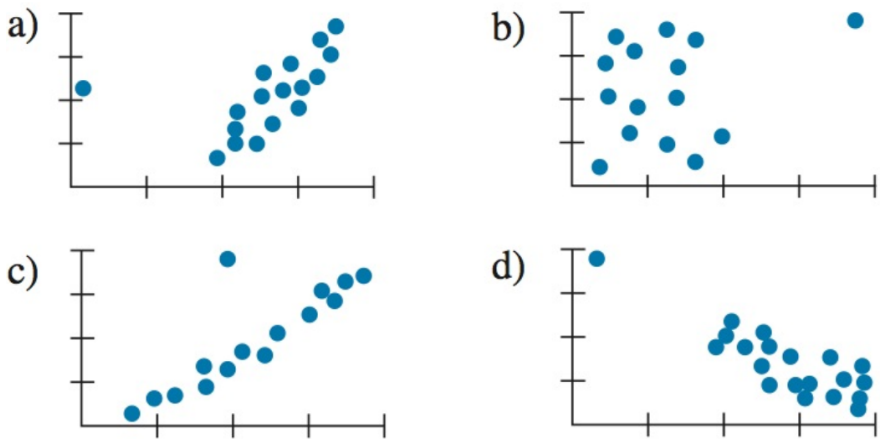
A student who has created a linear model is disappointed to find that her R^2 value is a very low 13%.

1. Does this mean that a linear model is not appropriate? Explain.

2. Does this model allow the student to make accurate predictions? Explain.

Exercise 9

Each of the following scatterplots shows a cluster of points and one "stray" point. For each, answer these questions:



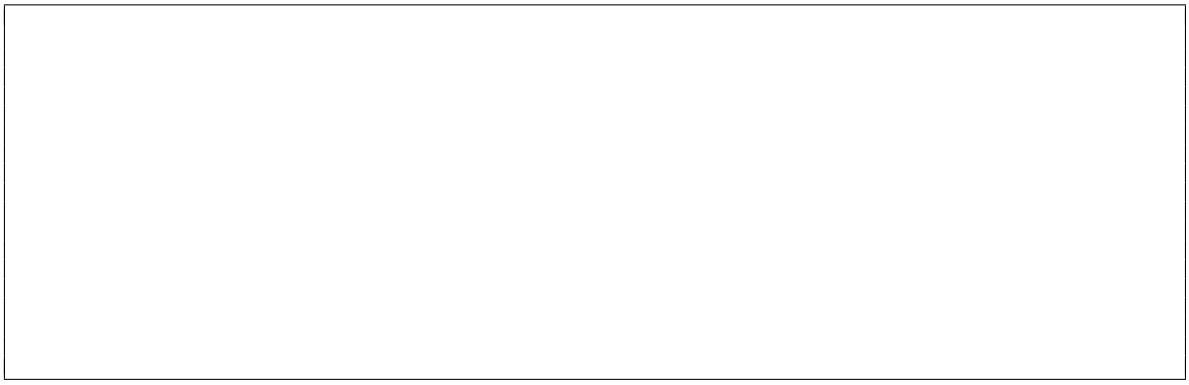
1. In what way is the point unusual? Does it have high leverage, a large residual, or both?

2. Do you think that point is an influential point?

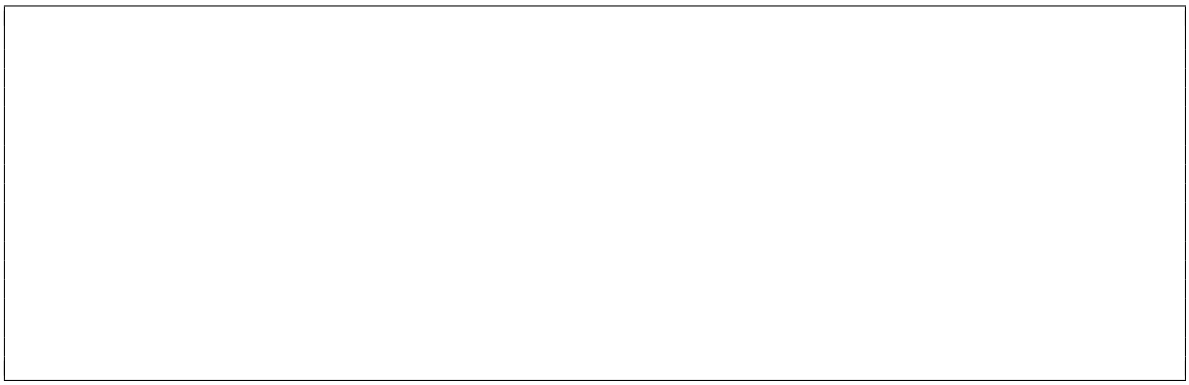
3. If that point were removed, would the correlation become stronger or weaker? Explain.

4. If that point were removed, would the slope of the regression line increase or decrease?

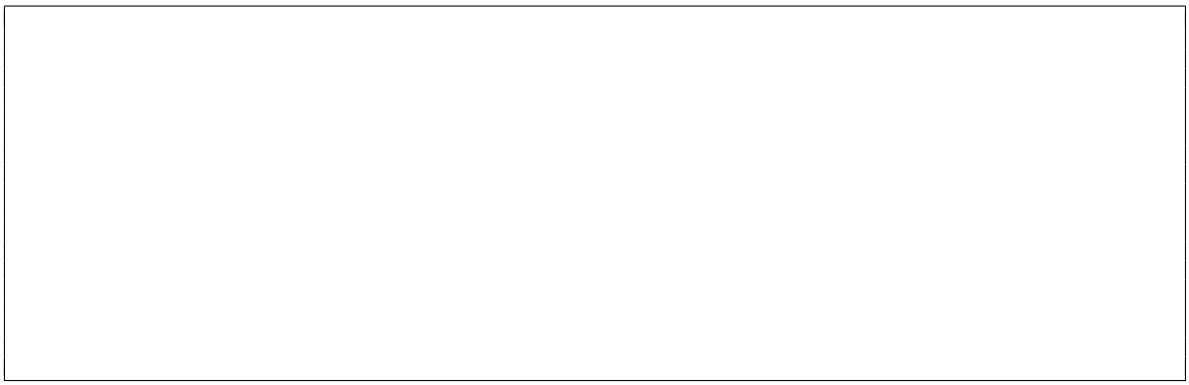
(a)



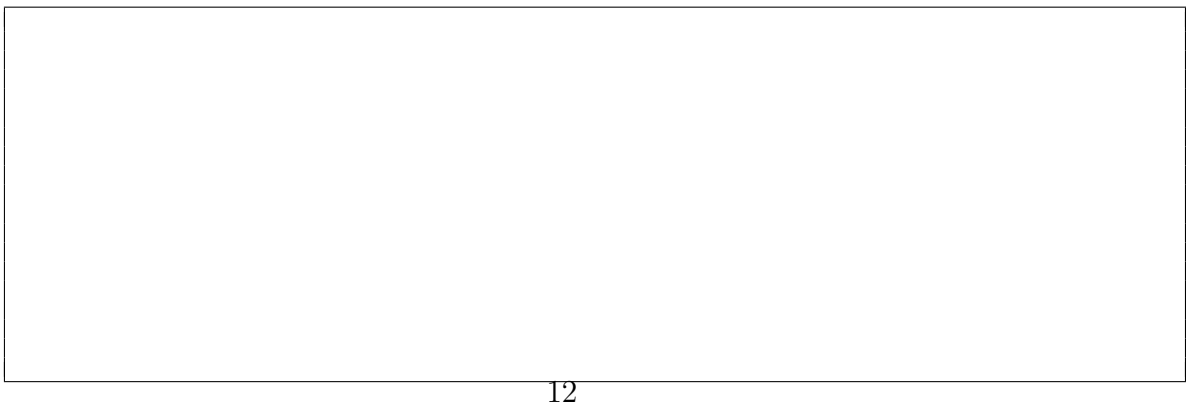
(b)



(c)

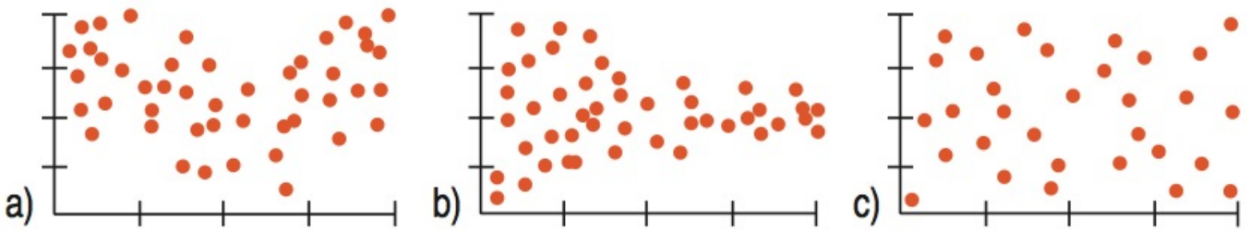


(d)



Exercise 10

Suppose you have fit a linear model to some data (like that of Exercise 1) and now take a look at the residuals. For each of the following possible residuals plots, tell whether you would try a re-expression and, if so, why.



(a)

(b)

(c)

Exercise 11

For each of the models listed below, predict y when $x = 2$.

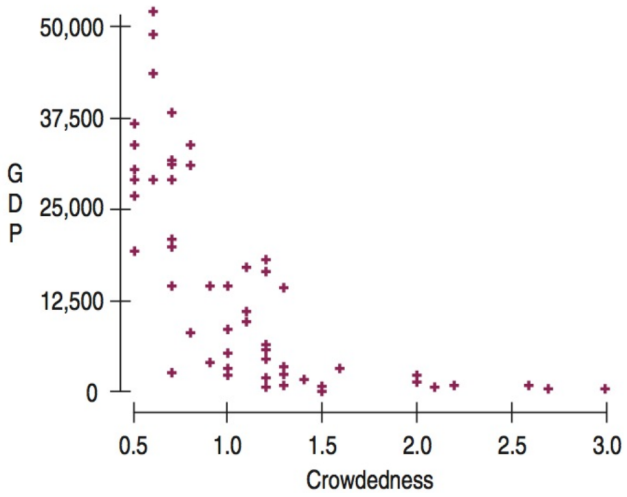
1. $\ln \hat{y} = 1.2 + 0.8 \ln x$

2. $\hat{y}^2 = 1.2 + 0.8x$

3. $1/\sqrt{\hat{y}} = 1.2 + 0.8x$

Exercise 12

In a Chance magazine article (Summer 2005), Danielle Vasilescu and Howard Wainer used data from the United Nations Center for Human Settlements to investigate aspects of living conditions for several countries. Among the variables they looked at were the country's per capita gross domestic product (*GDP*, in $) and *Crowdedness*, defined as the average number of persons per room living in homes there. This scatterplot displays these data for 56 countries:



1. Explain why you should re-express these data before trying to fit a model.

2. What re-expression of *GDP* would you try as a starting point?

Exercise 13

Are good grades in high school associated with family togetherness? A random sample of 142 high school students was asked how many meals per week their families ate together. Their responses produced a mean of 3.78 meals per week, with a standard deviation of 2.2. Researchers then matched these responses against the students' grade point averages (GPAs). The scatterplot appeared to be reasonably linear, so they created a line of regression. No apparent pattern emerged in the residuals plot. The equation of the line was

$$\widehat{GPA} = 2.73 + 0.11 Meals.$$

1. Interpret the y-intercept in this context.

2. Interpret the slope in this context.

3. What was the mean GPA for these students?

4. If a student in this study had a negative residual, what did that mean?

5. Upon hearing of this study, a counselor recommended that parents who want to improve the grades their children get should get the family to eat together more often. Do you agree with this interpretation? Explain.

Exercise 14

A Statistics instructor created a linear regression equation to predict students' final exam scores from their midterm exam scores. The regression equation was $\widehat{Fin} = 10 + 0.9Mid$.

1. If Susan scored a 70 on the midterm, what did the instructor predict for her score on the final?

2. Susan got an 80 on the final. How big is her residual?

3. If the standard deviation of the final was 12 points and the standard deviation of the midterm was 10 points, what is the correlation between the two tests?

4. How many points would someone need to score on the midterm to have a predicted final score of 100?

5. Suppose someone scored 100 on the final. Explain why you can't estimate this student's midterm score from the information given.

6. One of the students in the class scored 100 on the midterm but got overconfident, slacked off, and scored only 15 on the final exam. What is the residual for this student?

7. No other student in the class "achieved" such a dramatic turnaround. If the instructor decides not to include this student's scores when constructing a new regression model, will the R^2 value of the regression increase, decrease, or remain the same? Explain.

8. Will the slope of the new line increase or decrease?

Exercise 15

The ranges inhabited by the Indian gharial crocodile and the Australian saltwater crocodile overlap in Bangladesh. Suppose a very large crocodile skeleton is found there, and we wish to determine the species of the animal. Wildlife scientists have measured the lengths of the heads and the complete bodies of several crocs (in centimeters) of each species, creating the regression analyses below:

Indian Crocodile
Dependent variable is IBody
R-squared = 97.2%

Variable	Coefficient
Intercept	−69.3693
IHead	7.40004

Australian Crocodile
Dependent variable is ABody
R-squared = 98.1%

Variable	Coefficient
Intercept	−21.3429
AHead	7.82761

1. Do the associations between the sizes of the heads and bodies of the two species appear to be strong? Explain.

2. The crocodile skeleton found had a head length of 62 cm and a body length of 380 cm. Which species do you think it was? Explain why.

Exercise 16

Data from 50 large U.S. cities show the mean *January Temperature* (degrees Fahrenheit), *Altitude* (feet above sea level), and *Latitude* (degrees north of the equator) for 55 cities. Here's the correlation matrix:

	Jan. Temp	Latitude	Altitude
Jan. Temp	1.000		
Latitude	−0.848	1.000	
Altitude	−0.369	0.184	1.000

1. Which seems to be more useful in predicting *January Temperature*: *Altitude* or *Latitude*? Explain.

2. If the *Temperature* were measured in degrees Celsius, what would be the correlation between *Temperature* and *Latitude*?

3. If the *Temperature* were measured in degrees Celsius and the *Altitude* in meters, what would be the correlation? Explain.

4. What would you predict about the *January Temperature* in a city whose *Altitude* is two standard deviations higher than the average *Altitude*?