NAME:

Statistics 111 Summer Session II

Homework Nine & Ten

You are allowed to discuss problems with other students, but the final answers must be your own work.

For all problems that require calculation, YOU MUST ATTACH SEPARATE PAGES, NEATLY WRITTEN, THAT SHOW YOUR WORK.

Please mark your answer in the space provided. As a general rule, each blank counts for one point unless otherwise specified. If necessary work is not shown, or if that work is substantially wrong, then you will not get credit even if the answer is correct. (The obvious purpose of this is to prevent students from mindlessly copying each other's answers.)

Report all numerical answers to at least two correct decimal places.

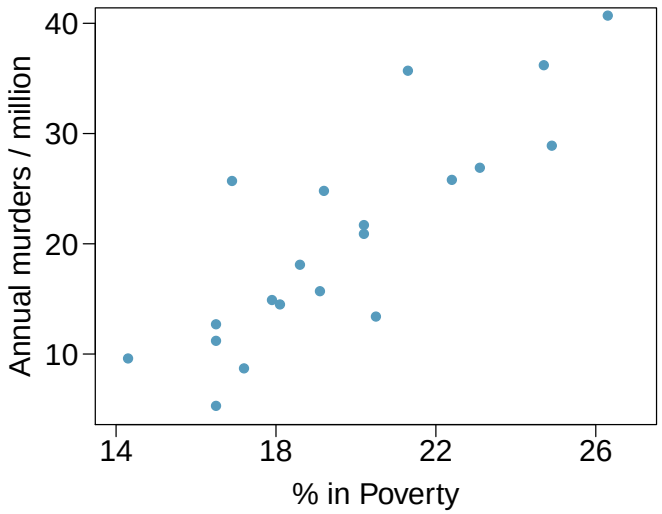
This homework contains some additional practice questions to help you better preparing for the final exam.

DUE DATE: START of class on Thursday, August 9, 2018.

1. (**Inference for Simple Linear Model**) The following regression output is for predicting annual murders per million from percentage living in poverty in a random sample of 20 metropolitan areas. The model output is also provided below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-29.901	7.789	-3.839	0.001
poverty%	2.559	0.390	6.562	0.000

$s = 5.512 \quad R^2 = 70.52\% \quad R^2_{adj} = 68.89\%$



(a) Write out the linear model.

(b) Interpret the intercept and the slope.

(c) Interpret R^2.

(d) What are the hypotheses for evaluating whether poverty percentage is a significant predictor of murder rate?

(e) State the conclusion of the hypothesis test from the hypothesis test in context of the data.

(f) Calculate a 95% confidence interval for the slope of poverty percentage, and interpret it in context of the data.

2. **(Prediction Interval)** In 1966 Cyril Burt published a paper called "The genetic determination of differences in intelligence: A study of monozygotic twins reared apart?" The data consist of IQ scores for [an assumed random sample of] 27 identical twins, one raised by foster parents, the other by the biological parents. Note that the average IQ score of 27 biological twins in the sample is 95.3 points, with a standard deviation of 15.74 points. The R output for simple linear regression is as below.

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      9.20760    9.29990   0.990    0.332
bioIQ            0.90144    0.09633   9.358  1.2e-09

Residual standard error: 7.729 on 25 degrees of freedom
```

(a) Construct the 95% confidence interval for the average IQ score of foster twins whose biological twins have IQ scores of 100 points.

(b) Construct the 95% prediction interval for the future value of IQ score of foster twins whose biological twins have IQ scores of 100 points.

3. (**Multiple Linear Regression**) The Child Health and Development Studies investigate a range of topics. One study considered all pregnancies between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco East Bay area. Here, we study the relationship between various explanatory variables and weight of the baby. The variable `smoke` is coded 1 if the mother is a smoker, and 0 if not. The variability within the smokers and non-smokers are about equal and the distributions are symmetric. Another variable we consider is `parity`, which is 0 if the child is the first born, and 1 otherwise. Other variables of interest include length of pregnancy in days (`gestation`), mother's age in years (`age`), mother's height in inches (`height`), and mother's pregnancy weight in pounds (`weight`). We consider all these possibly related variables at once. Below are three observations from this data set.

	bwt	gestation	parity	age	height	weight	smoke
1	120	284	0	27	62	100	0
2	113	282	0	33	64	135	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1236	117	297	0	38	65	129	0

The summary table below shows the results of a regression model for predicting the average birth weight of babies based on all of the variables included in the data set.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-80.41	14.35	-5.60	0.0000
gestation	0.44	0.03	15.26	0.0000
parity	-3.33	1.13	-2.95	0.0033
age	-0.01	0.09	-0.10	0.9170
height	1.15	0.21	5.63	0.0000
weight	0.05	0.03	1.99	0.0471
smoke	-8.40	0.95	-8.81	0.0000

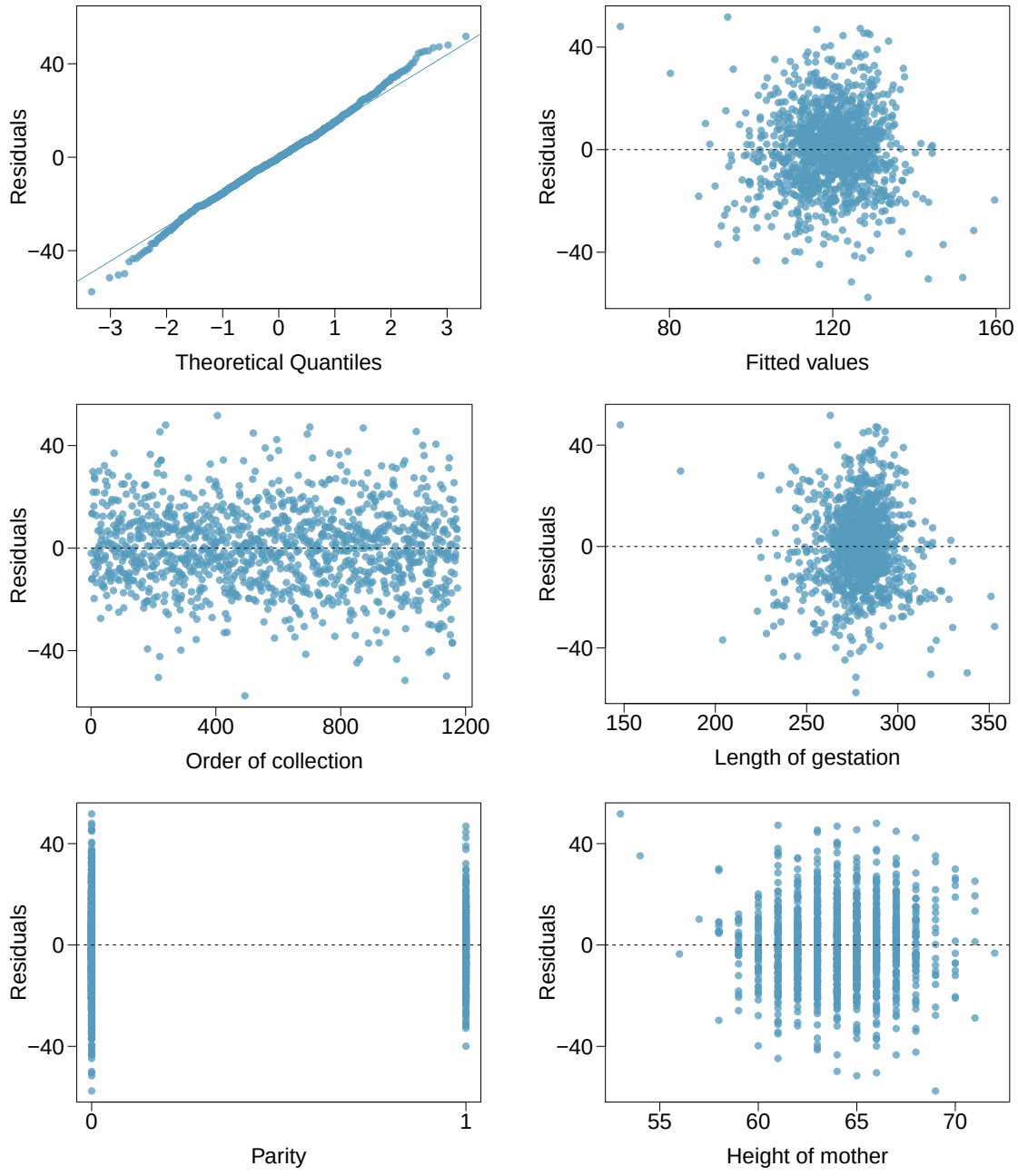
(a) Write the equation of the regression line that includes all of the variables.

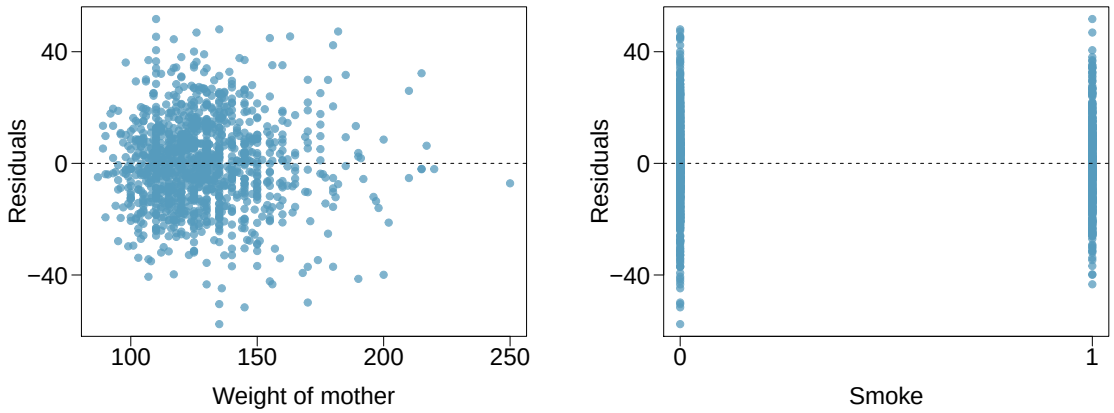
(b) Interpret the slopes of `gestation` and `parity` in this context.

(c) Calculate the residual for the first observation in the data set.

(d) The variance of the residuals is 249.28, and the variance of the birth weights of all babies in the data set is 332.57. Calculate the R^2 and the adjusted R^2. Note that there are 1,236 observations in the data set.

4. (**Condition Checking**) In last problem, we present a regression model for predicting the average birth weight of babies based on length of gestation, parity, height, weight, and smoking status of the mother.





Determine if the model assumptions (nearly normal residuals, constant variability of residuals, independent residuals, linear relationship between the response variable and numerical explanatory variables) are met using the plots below. If not, describe how to proceed with the analysis. (2 points)

5. (**Practice Question**) Let $X_1, \ldots, X_n$ be a random sample from the distribution with probability mass function,

$$f(x;\theta) = \frac{\theta^x}{x!}\exp(-\theta), \quad \text{for } x = 0, 1, ... \tag{1}$$

and 0 else, where $\theta > 0$

(a) Write the likelihood function

(b) Find the maximum likelihood estimate of θ

(c) What is the bias of the estimate $\hat{\theta} = 3\bar{X}$?

(d) What is the variance of the estimate $\hat{\theta} = 3\bar{X}$?

6. (**Practice Question**) You want to test whether the mean breaking strength of thread manufactured by Damocles Industries exceeds 3 pounds of force. Assume that you know that the standard deviation in breaking strengths is 0.5 pounds and let $\bar{x} = 3.6$, $n = 60$.

   (a) In symbols, what are your null and alternative hypotheses?

   (b) What is your test statistic?

   (c) For $\alpha = 0.01$, what is your critical value?

   (d) What is your p-value?

(e) Suppose the true mean breaking strength is 3.2 pounds. What is the power of a 0.01 level test with 100 observations?

(f) Suppose the true mean breaking strength is 3.2 pounds. How large a sample size would you need to have probability 0.9 of rejecting the null hypothesis at level 0.05?

(g) Suppose the true mean breaking strength is 3.2 pounds. You have a test with power 0.8 and sample size 49. What is the probability of falsely rejecting the null hypothesis if the null hypothesis were true (i.e., your α level)?

7. (**Practice Question**) Suppose Durham has 25 fast-food chain restaurants, and you visit 10 of them, and find that the average sanitation score is 90.2 with a sample standard deviation of 5. Set a one-sided lower 95% confidence interval on the average sanitation score in such restaurants. (Hint: Is FPCF necessary?)

8. (**Practice Question**) Let $X_1, \ldots, X_n$ be a random variable from the exponential distribution with parameter $\lambda = 4$. Write density function for the sample maximum and indicate support.

9. (**Practice Question**) Someone asserts that at least 10% more women than men support gay marriage. You want to test this at the 0.05 level. You sample 100 women and find that 60 are supporters, and among 80 men, 30 are supporters.

   (a) Write the alternative hypothesis in symbols. (Subtract men from women.)

   (b) What is the value of your test statistic?

   (c) What is your p-value?

This is your last homework, hopefully you have enjoyed the semester and learned a lot. It has been a lot of work, but the intent was to make it rewarding and worthwhile. What aspects of this course do you like best? Are there any areas where you think the course (or section) could be improved?