STATISTICS 110/201, FALL 2017 LECTURE A, MIDTERM EXAM

NAME:_____ Homework code :_____ Seat: _____

Open notes, calculator required. Your exam should have 6 pages, and a page of R output that will be handed out separately. Make sure you have them all. Each part of each problem is worth 4 points unless specified otherwise. Use the back of the pages if you need more space, but *tell us to turn the page over and look*.

1. A study investigated 1,213 pregnancies between 1960 and 1967 among women in the San Francisco East Bay area. Three of the variables measured and the notation we will use for them are:

Y = bwt = Birth weight (in ounces)
X_1 = gestation = Gestational age (in days, i.e. the length of the pregnancy)
X_2 = smoke = 1 if the mother smoked and 0 if she did not smoke during pregnancy

a. *For part (a) only, use the notation with the names of the variables instead of Y and Xs.* Write the population model that specifies a linear relationship between Y = bwt and X_1 = gestation *and* for which that relationship has the same slope and the same intercept for smokers and nonsmokers. Include information about the normality assumption. (The left hand side of the model is provided, to get you started.)

$$bwt = \beta_0 + \beta_1(gestation) + \varepsilon,$$
$$\text{where we assume } \varepsilon \sim N(0, \sigma_\varepsilon)$$

For parts (b) and (c) you don't need to include information about the normality assumption. *For these two parts use the Y and X notation rather than names of variables. (That will save you some writing!)*

b. Write the population model for the linear relationship between bwt and gestation that includes the same *intercept* but different *slopes* for smokers and nonsmokers.

$$Y = \beta_0 + \beta_1(X_1) + \beta_2(X_1X_2) + \varepsilon$$

c. Write the population model for the linear relationship between bwt and gestation that includes different intercepts *and* different slopes for smokers and nonsmokers.

$$Y = \beta_0 + \beta_1(X_1) + \beta_2(X_2) + \beta_3(X_1X_2) + \varepsilon$$

d. Using the notation from your model in part (c), write the null and alternative hypotheses that would be used to test whether the population regression lines are the same for smokers and nonsmokers, versus that they are not the same in some way. Write these in terms of the coefficients.

H_0: $\beta_2 = \beta_3 = 0$

H_a: β_2 *and* β_3 *are not both 0*

The R output handed out separately includes regression results for various models for the situation described in Question 1. Use it to answer Questions 2 through 7.

2. The model Smokemod includes *smoke* as the only explanatory variable. The coefficients for this model represent simple summary statistics learned in Statistics 7. Using the output for Smokemod, explain what summary statistic each of the following represents. In other words, if you had the data, what summaries of it could you compute to get these values?

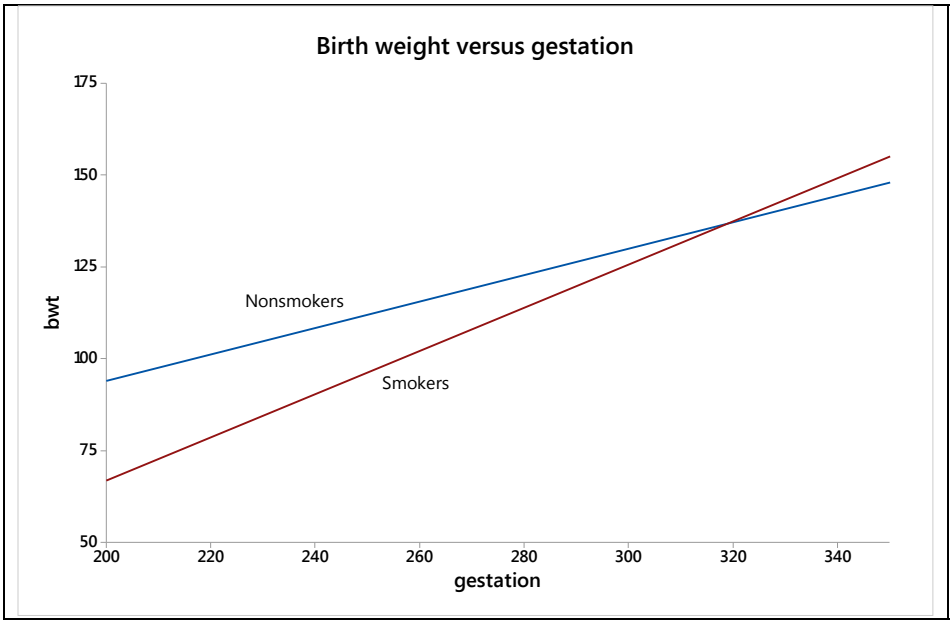
a. The intercept of 123.047

The intercept of 123.047 is the mean of the birthweights for the babies in the sample whose mothers did not smoke.

b. The "smoke" coefficient of –8.938

The smoke coefficient is the difference in the sample mean birthweights for the babies born to mothers who smoked and babies born to mothers who did not smoke, i.e. $\bar{y}_{nonsmokers} - \bar{y}_{smokers}$.

3. [8 pts total] The model SmokeGestInt includes *smoke*, *gestation*, and the *interaction* between them. Using the results in the output, draw two regression lines for the relationship between gestation and btw: one for women who smoked during pregnancy and one for women who did not smoke. Label the two lines to show which is which, and write the equation of each line in the space provided. You can round the coefficients to 2 decimal places. *Try to place the lines in the right place and note that the axes do not start at 0.*



Equations:

For Nonsmokers:

$\hat{Y}$ = 21.99 + 0.36(*gestation*)

For Smokers:

$\hat{Y}$ = (21.99 – 73.28)
+ (.36 + .23)(*gestation)*
= – 51.29 + 0.59*(gestation)*

4. Using the SmokeGestInt model, the four intervals below are all for gestation = 290 days. They represent the following (not in order):

- A 95% confidence interval for the mean birth weight of babies whose mothers smoked
- A 95% prediction interval for birth weight of a randomly selected baby whose mother smoked
- A 95% confidence interval for the mean birth weight of babies whose mothers did not smoke
- A 95% prediction interval for birth weight of a randomly selected baby whose mother did not smoke

a. For each interval, put a check mark under either Prediction or Confidence and under either Smoker or Nonsmoker, illustrating which combination of those the interval represents.

fit lwr upr	Prediction	Confidence	Smoker	Nonsmoker
126.6946 95.08699 158.3023	X			X
121.2372 119.3929 123.0815		X	X	
121.2372 89.60481 152.8696	X		X	
126.6946 125.34 128.0492		X		X

b. The mean gestation for all babies in the data set was 279.3 days. For what values of gestation would the 95% confidence interval for mean birthweight be *narrower* than it was for gestation = 290 days? Explain how you know.

The width of the interval is smallest when $x^* = \bar{x}$, and gets larger as those two values move apart. So the intervals that would be narrower would be the ones with gestation closer to 279.3 than 290 is. That would include gestation values that are within (290 – 279.3) = 10.7 of the mean of 279.3. That includes values between 268.6 and 290.

5. (1 pt each) For the SmokeGestInt model, fill in the blanks in the ANOVA table below, where F is the test statistic for H_0: $\beta_1 = \beta_2 = \beta_3 = 0$. Hint: All of the information you need is in the output, but you will need to do some arithmetic to get some of the values.

Source	Df	SumSq	MeanSq	F	*p*-value
Model	3	88749	29583	114.2	2.2×10^{-16}
Error	1209	313217	259		
Total	1212	401966			

6. [2 pts each blank] The model SmokeGest includes the two explanatory variables *gestation* and *smoke* but not the interaction. Using that model, fill in numerical values in each of the blanks where possible. Write NA (not available) if a numerical value cannot be determined from the R output or from general statistical knowledge. (An example of general statistical knowledge would be that the sum of the residuals is 0, even though that's not shown anywhere in the R output.) No extensive computations are required, but if you need to compute something you can show your work on the side. (If you make a mistake in computation, you might get you partial credit if we can see where you went wrong.)

a. $\hat{\beta}_0$ = ___-0.93166_____

b. The standard error of $\hat{\beta}_0$ = __8.15239_____

c. $\sqrt{MSE}$ = _16.19_____

d. β_1 = ____NA_____ (*This is a population value, unknown*)

e. The standard error of β_1 = ____0_____ (β_1 *is a constant, so it has 0 standard error*)

f. The value of the test statistic for testing H_0: $\beta_1 = \beta_2 = 0$ is ___F = 161.9_____

g. SS*Model* = __65988 + 18890 = 84878__

7. For the 3 models shown in the output (Smokemod, SmokeGestInt and SmokeGest), which one would you recommend using to predict birth weight? Explain why you chose that model.

Use the one with the highest Adjusted R-squared or the lowest Residual standard error, which is the one with the interaction, SmokeGestInt. You could also justify it by noting that for that model all of the coefficients are significantly different from 0, so none of them should be removed.

8. A study recently reported in the *New York Times* stated the following: *"A Finnish study suggests that regular sauna visits can reduce the risk for high blood pressure. The study included 1,621 middle-aged men with normal blood pressure who were followed for an average of 25 years. During that time, 251 developed hypertension. Compared with those who reported one sauna session a week or less, those who reported two to three sessions were 24 percent less likely to have hypertension, and four to seven visits a week reduced the risk by 46 percent."*

a. Do you think this was a randomized experiment, or an observational study? Explain why you think that.

1. *This is almost surely an observational study. You can't randomly assign people to different weekly sauna use over a 25-year period. (You could try, but they are unlikely to cooperate!)*

b. (Question 8 continued from previous page.) The headline accompanying the report was "Saunas May Lower Health Risk." Do you think that headline is appropriate? Explain why or why not.

The headline uses the word "may" so it's okay. It is true that it could be a cause and effect relationship. The headline would not be appropriate if it said "Saunas lower health risk" without the "may" used in the headline.

9. [2 pts each] For each of the following situations, specify whether the statement provided is *always* true, *could be* true for some populations and/or samples, or is *never* true. (Circle your answer.)

a. When X and Y have a deterministic linear relationship, the correlation is +1.

Always true **Could be true** Never true

b. In simple linear regression, if the population intercept, slope and error standard deviation (i.e. β_0, β_1 and σ_ε) are all known and non-zero, then the width of a prediction interval for Y at any value of X will be 0.

Always true Could be true **Never true**

c. Consider a regression situation with Y as the response, and 2 possible predictors X_1 and X_2. Assuming there are no missing values, SS*Total* will be the same for the model with X_1 and X_2 as predictors as it is for the model with only X_1 as a predictor.

Always true Could be true Never true

d. The slope of the regression line for Y versus X would be the same if the roles of X and Y are reversed.

Always true **Could be true** Never true

e. When the correlation between X and Y is positive (and not 0) the slope of the least square regression line for simple linear regression is also positive.

Always true Could be true Never true

f. In a simple linear regression setting the numerical values of β_1 and $\hat{\beta}_1$ are equal.

Always true **Could be true** Never true

MULTIPLE CHOICE (3 pts each) *Circle the best choice*

1. In simple linear regression, which of the following cannot be checked by a plot of the residuals versus fitted values?
A. The relationship between Y and X is approximately linear.
B. The standard deviation of the errors remains constant across the *x* values.
C. The *n* pairs of observations are all independent.
D. There are no major outliers.

2. Which of the following is <u>always</u> true about the nested F test for comparing a full and reduced model?
A. The value for the numerator degrees of freedom must be one.
B. The denominator of the F statistic is the MSE for the full model.
C. The sum of squares for the numerator is computed using only a subset of the *n* units in the data set.
D. The value for the denominator degrees of freedom is $n - 1$.

3. In class an applet for constructing 95% confidence intervals for the mean body temperature of 18 to 30 year-old adults was illustrated. Which of the following was true for the confidence intervals generated by the applet?
A. When 20 different confidence intervals were constructed, 19 of them always covered the population mean and one did not cover the population mean.
B. When 100 different confidence intervals were constructed, 95 of them always covered the population mean and the other 5 did not cover the population mean.
C. Both A and B above were true.
D. Neither A nor B above was true.

4. In simple linear regression, $\sqrt{MSE}$ is used as an estimate of σ. In this context, what is σ?
A. The standard deviation of the population of X values at each value of Y.
B. The standard deviation of the population of Y values at each value of X.
C. The standard deviation of the population of all Y values combined, across all of the values of X.
D. The standard deviation of the residuals from the sample.

For Question 2:

```
> Smokemod <- lm(bwt ~ smoke, data = babies)
> summary(Smokemod)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  123.047      0.649 189.597   <2e-16 ***
smoke         -8.938      1.033  -8.653   <2e-16 ***

Residual standard error: 17.68 on 1224 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.05764,  Adjusted R-squared:  0.05687
F-statistic: 74.87 on 1 and 1224 DF,  p-value: < 2.2e-16
```

For Questions 3 through 5

```
> SmokeGestInt<-lm(bwt ~ gestation + smoke + gestation:smoke, data = babies)
> summary(SmokeGestInt)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      21.98517   10.04233   2.189 0.028769 *
gestation         0.36107    0.03578  10.092  < 2e-16 ***
smoke           -73.28346   16.89161  -4.338 1.55e-05 ***
gestation:smoke   0.23388    0.06050   3.866 0.000117 ***

Residual standard error: 16.1 on 1209 degrees of freedom
  (23 observations deleted due to missingness)
Multiple R-squared:  0.2208,   Adjusted R-squared:  0.2189
F-statistic: 114.2 on 3 and 1209 DF,  p-value: < 2.2e-16

> anova(SmokeGestInt)
Analysis of Variance Table

Response: bwt
                Df Sum Sq Mean Sq F value    Pr(>F)
gestation        1  65988   65988 254.708 < 2.2e-16 ***
smoke            1  18890   18890  72.914 < 2.2e-16 ***
gestation:smoke  1   3871    3871  14.944 0.0001167 ***
Residuals       1209 313217     259
```

For Question 6:

```
> SmokeGest<-lm(bwt ~ gestation + smoke, data = babies)
> summary(SmokeGest)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.93166    8.15239  -0.114    0.909
gestation    0.44286    0.02902  15.262   <2e-16 ***
smoke       -8.08830    0.95266  -8.490   <2e-16 ***

Residual standard error: 16.19 on 1210 degrees of freedom
  (23 observations deleted due to missingness)
Multiple R-squared:  0.2112,   Adjusted R-squared:  0.2099
F-statistic: 161.9 on 2 and 1210 DF,  p-value: < 2.2e-16

> anova(SmokeGest)
Analysis of Variance Table

Response: bwt
            Df Sum Sq Mean Sq F value    Pr(>F)
gestation    1  65988   65988 251.807 < 2.2e-16 ***
smoke        1  18890   18890  72.083 < 2.2e-16 ***
Residuals 1210 317089     262
```