STATISTICS 110/201, FALL 2017 LECTURE B, 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 1148 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)
$X_1$ = height = mother's height in inches
$X_2$ = parity = 1 if the mother has given birth before and 0 if she has not

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 $X_1$ = height. Do not include parity in the model. Include information about the normality assumption as part of the model specification. (The left hand side of the model is provided, to get you started.)

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

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 height that includes the same *slope* but different *intercepts* for mothers with parity = 1 and parity = 0.

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

c. Write the population model for the linear relationship between bwt and height that includes different intercepts *and* different slopes for the two parity groups.

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

d. For your model in part (c), give an interpretation of the coefficient for $X_1$.

*$\beta_1$ is the population slope of the regression line for Y = bwt and X = height for mothers with parity = 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 6. Note: Height was missing for 17 cases, so the models involving height have 17 fewer cases than the model not involving height.**

**2.** The model ParModel includes *parity* as the only explanatory variable. Using the results given by the summary and anova commands for that model, provide numerical values for each of the following. Hint: One or more require you to do a small amount of arithmetic to find the answer. Show your work.

a. The sample mean of the birth weights for mothers with parity = 0.

*This is* $\hat{\beta}_0$ *= 119.8767.*

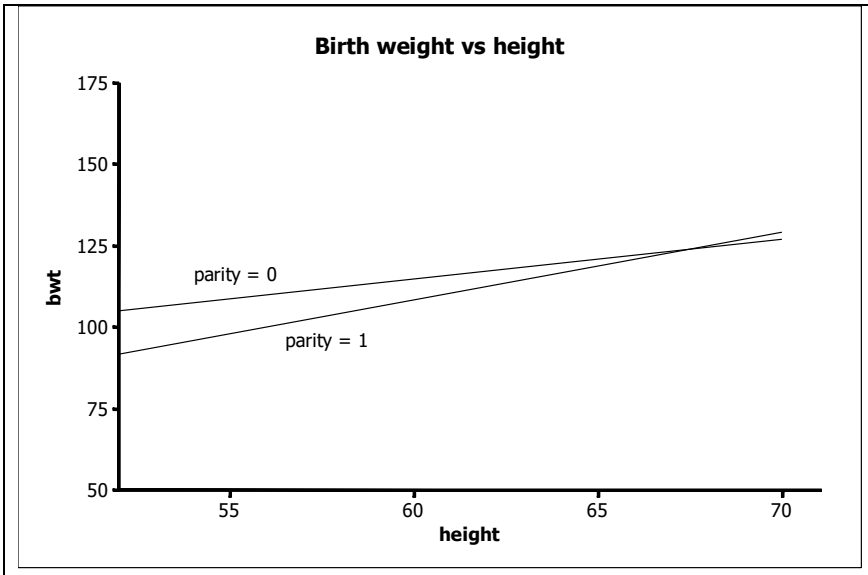
b. The sample mean of the birth weights for mothers with parity = 1.

*This is* $\hat{\beta}_0 + \hat{\beta}_1$ *= 119.8767 – 2.5642 = 117.3125.*

c. SSY = $\sum(Y - \bar{Y})^2$ where the sum is over all of the cases in the data set (i.e. not just one parity group).

*This is SSTotal = SSModel + SSE = 1419 + 383143 = 384,562.*

**3.** [8 pts total] The model HtParInt includes *height*, *parity*, and the *interaction* between them. Using the results in the output, draw two regression lines for the relationship between height and btw: one for women who had parity = 0 and one for women who had parity = 1. 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 parity = 0:

$\hat{Y}$ = 41.74 + 1.22(*height*)

For parity = 1:

$\hat{Y}$ = (41.74 – 57.97)
+ (1.22 + .86)(*height)*
= –16.23 + 2.08(*height*)

**4.** The HtOnly model is for simple linear regression using height to predict bwt. The intervals below were all created using that model. They represent the following (not in order):

- A 95% confidence interval for the mean birth weight of babies whose mothers were 60 inches tall
- A 95% prediction interval for birth weight of a randomly selected baby whose mother was 60 inches tall
- A 95% confidence interval for the mean birth weight of babies whose mothers were 66 inches tall
- A 95% prediction interval for birth weight of a randomly selected baby whose mother was 66 inches tall

a. For each interval, put a check mark under either Prediction or Confidence and under either 60 inches or 66 inches, illustrating which combination of those the interval represents.

| fit lwr upr | Prediction | Confidence | 60 inches | 66 inches |
|---|---|---|---|---|
| 121.9811 86.61498 157.3471 | X | | | X |
| 113.4554 78.05916 148.8517 | X | | X | |
| 113.4554 111.4783 115.4326 | | X | X | |
| 121.9811 120.65 123.3122 | | X | | X |

b. The mean height of all mothers used in the data set to create the HtOnly model was 64 inches. Use that information to explain why a confidence interval for the mean birth weight of babies whose mothers are 70 inches tall would be wider than the corresponding confidence interval for mothers who are 60 inches tall.

*The width includes a term with* $(x^* - \bar{x})^2$ *in the numerator.* $(60 - 64)^2 = 16$*, which is smaller than* $(70 - 64)^2 = 36$*. So the part added and subtracted to* $\hat{Y}$ *to get the interval is smaller for height = 60 than for height = 70.*

**5.** (1 pt each) The HtPar model has height and parity, but not interaction. For that model, fill in the blanks in the ANOVA table below, where F is the test statistic for $H_0$: $\beta_1 = \beta_2 = 0$. 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 | 2 | 14622+1648=16270 | 8135 | 25.17 | $2.031 \times 10^{-11}$ |
| Error | 1128 | 364641 | 323.3 | | |
| Total | 1130 | 380911 | | | |

**6.** [2 pts each blank] Continuing to use the output from the HtPar model *only* (i.e. not output from any of the other models), 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. SSModel/SSTotal = __$R^2 = .04271$_____

b. The standard error of $\hat{\beta}_0$ = __13.5404_____

c. The point estimate for $\sigma_\varepsilon$ = __17.98_____

d. $\beta_2$ = __NA_____ (This is a population value, unknown.)

e. The *p*-value for testing height, given that parity is in the model = ___$1.67 \times 10^{-11}$_____

f. The *p*-value for testing height, before adding parity to the model = __$2.777 \times 10^{-11}$_____

g. The sample size used for the analysis of the HtPar model = ____1131_____

**7.** A study recently reported in the *New York Times* was described as follows: "*Researchers told 49 volunteers that they were testing two anti-itch creams – one costly and one cheap – that contained the same ingredient known to reduce itch, but that the ingredient sometimes increased sensitivity to heat. They then showed them the two creams, one in an expensive-looking brand-name box, and the other in a generic-looking container. They did not tell them that neither contained any medicine. They then randomly assigned them to try one of the two creams. All participants knew which cream they were using. When exposed to heat, the volunteers using the expensive-looking cream felt consistently more pain than those using the cheap-looking one, and the effect increased over time.*"

a. Was this a randomized experiment or an observational study? Explain how you know.

*This was a randomized experiment. The volunteers were randomly assigned to receive one cream or the other.*

b. Explain what it would mean to make a "cause and effect" conclusion for this study, and whether you think such a conclusion is justified based on how the study was done.

*It would mean that seeing an expensive product leads people to think it's more potent that a cheaper one. Yes, a cause and effect conclusion is justified because this was a randomized experiment.*

**8.** [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. The least squares line in simple linear regression goes through the point $(\bar{x}, \bar{y})$.

**Always true** Could be true Never true

b. In simple linear regression, if the population intercept, slope and error standard deviation (i.e. $\beta_0$, $\beta_1$ and $\sigma_\varepsilon$) are all known, 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. For simple linear regression, the terms labeled Multiple $R^2$ and Adjusted $R^2$ (in R) are the same.

Always true Could be true **Never true**

d. The correlation between X and Y will be the same if the roles of X and Y are reversed.

**Always true** Could be true Never true

e. The p-value for a one-sided alternative hypothesis is ½ of the p-value for a two-sided alternative.

Always true **Could be true** Never true

f. In simple linear regression if a 95% confidence interval for the population slope does not cover 0, then the *p*-value for the test of $H_0$: $\beta_1 = 0$ versus $H_a$: $\beta_1 \neq 0$ will be less than 0.05.

**Always true** Could be true Never true

*For part (b), if there was a deterministic relationship this would be true, but in that case we would not use simple linear regression. However, if you stated that as the reason, or that $\sigma_\varepsilon = 0$, then you would get credit for this part if said "could be true."

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

**1.** In simple linear regression, if Multiple $R^2 = .25$, which of the following must be true?
A. The correlation between X and Y is 0.5.
B. The slope of the regression line is either 0.5 or –0.5.
**C. About 25% of the variability in Y can be explained by the regression model.**
D. The relationship between X and Y must be linear.

**2.** When comparing a full and reduced model using a nested F test which of the following must be true?
A. SSE for the reduced model is less than or equal to SSE for the full model.
**B. SSModel for the reduced model is less than or equal to SSModel for the full model.**
C. MSE for the reduced model is less than or equal to MSE for the full model.
D. MSModel for the reduced model is less than or equal to MSModel for the full model.

**3.** Consider a scatter plot of data for simple linear regression that shows the individual points and the least squares regression line. Which of the following is true for an individual with a negative residual?
**A. The point for that individual falls below the regression line.**
B. The point for that individual falls above the regression line.
C. If the slope is negative, the point for that individual falls below the line, but if the slope is positive, the point falls above the line.
D. If the slope is negative, the point for that individual falls above the line, but if the slope is positive, the point falls below the line.

**4.** In the simple linear regression situation, suppose we know the population values for $\beta_0$, $\beta_1$ and $\sigma_\varepsilon$. A confidence interval is found for $\mu_Y$ at a given value $X = x^*$, and a prediction interval is found for Y at a given value $X = x^*$. Which interval would have a width of 0?
**A. The confidence interval but not the prediction interval**
B. The prediction interval but not the confidence interval
C. Both intervals would have a width of 0.
D. Neither interval would have a width of 0.

**For Question 2**

```
> ParModel <- lm(bwt ~ parity, data = babies2)
> summary(ParModel)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 119.8767     0.6235  192.26   <2e-16 ***
parity       -2.5642     1.2448   -2.06   0.0396 *

Residual standard error: 18.28 on 1146 degrees of freedom
Multiple R-squared:  0.003689, Adjusted R-squared:  0.00282
F-statistic: 4.243 on 1 and 1146 DF,  p-value: 0.03963

> anova(ParModel)
Analysis of Variance Table
Response: bwt
            Df Sum Sq Mean Sq F value  Pr(>F)
parity       1   1419 1418.63  4.2432 0.03963 *
Residuals 1146 383143  334.33
```

**For Question 3**

```
> HtParInt <- lm(bwt ~ height + parity + height:parity, data = babies2)
> summary(HtParInt)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    41.7438    15.6504   2.667  0.00776 **
height          1.2202     0.2444   4.992 6.92e-07 ***
parity        -57.9670    31.1741  -1.859  0.06322 .
height:parity   0.8604     0.4857   1.772  0.07673 .

Residual standard error: 17.96 on 1127 degrees of freedom
  (17 observations deleted due to missingness)
Multiple R-squared:  0.04537,  Adjusted R-squared:  0.04283
F-statistic: 17.85 on 3 and 1127 DF,  p-value: 2.518e-11
```

**For Question 4**

```
> HtOnly <- lm(bwt ~ height, data = babies2)
> summary(HtOnly)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  28.1992    13.5639   2.079   0.0378 *
height        1.4209     0.2117   6.713 3.01e-11 ***

Residual standard error: 18.01 on 1129 degrees of freedom
  (17 observations deleted due to missingness)
Multiple R-squared:  0.03839,  Adjusted R-squared:  0.03753
F-statistic: 45.07 on 1 and 1129 DF,  p-value: 3.008e-11
```

**For Questions 5 and 6**

```
> HtPar <- lm(bwt ~ height + parity, data = babies2)
> summary(HtPar)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  27.8012    13.5404   2.053   0.0403 *
height        1.4381     0.2114   6.802 1.67e-11 ***
parity       -2.7824     1.2322  -2.258   0.0241 *

Residual standard error: 17.98 on 1128 degrees of freedom
  (17 observations deleted due to missingness)
Multiple R-squared:  0.04271,  Adjusted R-squared:  0.04102
F-statistic: 25.17 on 2 and 1128 DF,  p-value: 2.031e-11

> anova(HtPar)
Analysis of Variance Table
Response: bwt
            Df Sum Sq Mean Sq F value    Pr(>F)
height       1  14622 14621.6  45.231 2.777e-11 ***
parity       1   1648  1648.3   5.099   0.02413 *
Residuals 1128 364641   323.3
```