# p8130_hw6_ps3194

Pangsibo Shen

12/3/2020

```
library(tidyverse)
```

**Problem 1**

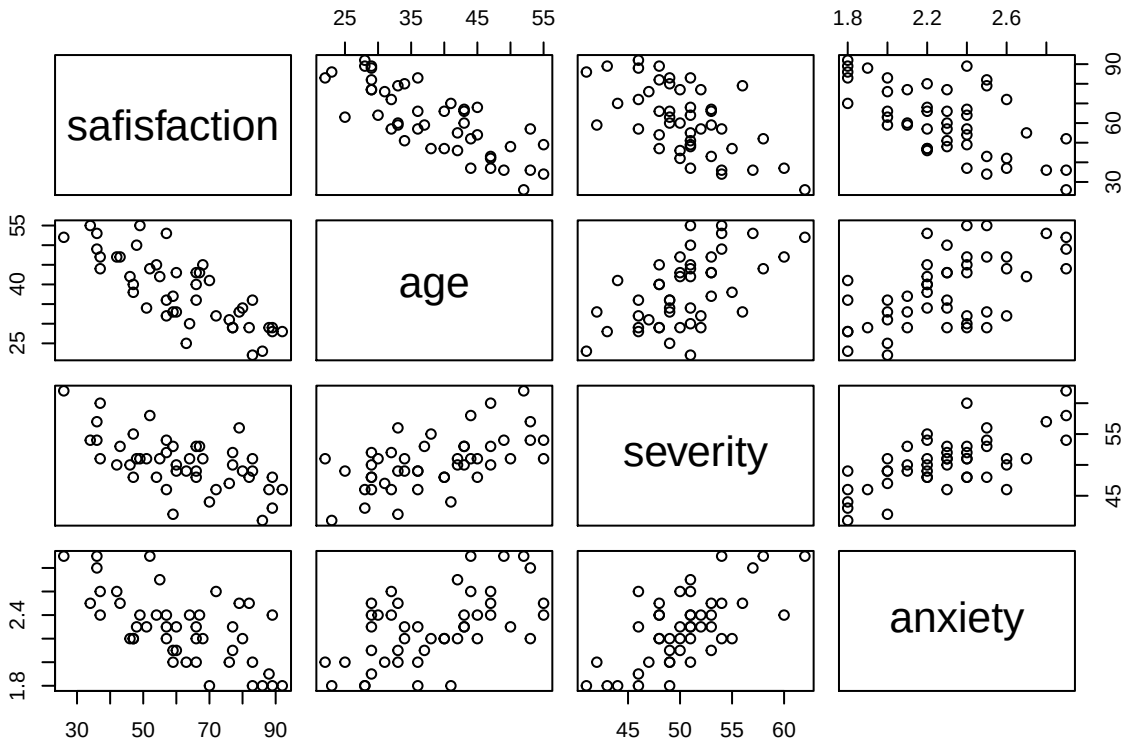
```
#load the data
pat_satisfaction = read_csv("./data/PatSatisfaction.csv") %>%
  janitor::clean_names()
```

**1) Create a correlation matrix for all variables and interpret your findings. Focus on the correlation values between each predictor and the outcome of interest.**

```
round(cor(pat_satisfaction),3)
```

```
##              safisfaction    age severity anxiety
## safisfaction        1.000 -0.787   -0.603  -0.645
## age                -0.787  1.000    0.568   0.570
## severity           -0.603  0.568    1.000   0.671
## anxiety            -0.645  0.570    0.671   1.000
```

```
# Scatter plot matrix for all variables
pairs(pat_satisfaction)
```




The estimated correlation between satisfaction and age is -0.787; between satisfaction and severity is -0.603; between satisfaction and anxiety is -0.643; between age and severity is 0.568; between age and anxiety is 0.570; between severity and anxiety is 0.671. Among three predictors, age has the largest correlation coefficient, which means this predictor has the strongest correlation with satisfaction among all predictors.Among three predictors, severity has the smallest correlation coefficient, which means this predictor has the weakest correlation with satisfaction among all predictors.In conclusion, there are strong negative linear relationships between satisfaction(outcome) and age; between satisfaction(outcome) and severity and between satisfaction(outcome) and anxiety.

**2) Fit a multiple regression model including all three predictors and test whether at least one of these variables is significant. State the hypotheses, test-statistic, decision rule and conclusion.**

For this question, we are going to use a global F-test to test whether at least one of these variables is significant and the hypotheses are stated below:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$
$$H_1 : \textit{the least one } \beta \textit{ is not zero}$$
$$F = \frac{MSR}{MSE} = \frac{3040.155}{101.1629} = 30.05208$$
$$F(1-\alpha; p, n-p-1) = 2.816466$$

```
fit_1 = lm(safisfaction~age + severity + anxiety, data = pat_satisfaction)
anova(fit_1)
```

```
## Analysis of Variance Table
##
## Response: safisfaction
##           Df Sum Sq Mean Sq F value    Pr(>F)
## age        1 8275.4  8275.4 81.8026 2.059e-11 ***
## severity   1  480.9   480.9  4.7539   0.03489 *
## anxiety    1  364.2   364.2  3.5997   0.06468 .
## Residuals 42 4248.8   101.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
MSR = sum(anova(fit_1)[1:3, 'Sum Sq'])/3
MSE = anova(fit_1)['Residuals', 'Mean Sq']
F_stats = MSR/MSE
```

For the F statistics, we got 30.0520779 which is greater than the F critical value (2.8164658). Hence we reject the null hypothesis and conclude that at significance level of 0.05, there are some estimated slope coefficients are not zero.

**3) Show the regression results for all estimated slope coefficients with 95% CIs. Interpret the coefficient and 95% CI associated with 'severity of illness'.**

The formula for a(1-$\alpha$)100% confidence for the true slope is given by below:

$$\hat{\beta}_i \pm t_{n-2,1-\alpha/2} * se(\hat{\beta}_i)$$

$$where\ se(\hat{\beta}_i) = \sqrt{\frac{MSE}{\sum_{j=1}^{n}(x_j - \bar{x})}}$$

```
summary(fit_1)
```

```
##
## Call:
## lm(formula = safisfaction ~ age + severity + anxiety, data = pat_satisfaction)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.3524  -6.4230   0.5196   8.3715  17.1601
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 158.4913    18.1259   8.744 5.26e-11 ***
## age          -1.1416     0.2148  -5.315 3.81e-06 ***
## severity     -0.4420     0.4920  -0.898   0.3741
## anxiety     -13.4702     7.0997  -1.897   0.0647 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.06 on 42 degrees of freedom
## Multiple R-squared:  0.6822, Adjusted R-squared:  0.6595
## F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

```r
confint(fit_1,level=0.95)
```

```
##                   2.5 %      97.5 %
## (Intercept) 121.911727 195.0707761
## age          -1.575093  -0.7081303
## severity     -1.434831   0.5508228
## anxiety     -27.797859   0.8575324
```

The estimated slope coefficients with 95% CIs for 'age': (-1.575093, -0.7081303); for 'severity': (-1.434831, 0.5508228); for 'anxiety': (-27.797859, 0.8575324). The estimated slope coefficient for 'severity of illness' is -0.442 which means that for everyone one unit increase in patient's severity of illness, the average satisfaction score will drop by 0.442. And at the significance level of 5%, we estimate that the true slope coefficient for 'severity of illness' lies between -1.434831 and 0.5508228.

**4) Obtain an interval estimate for a new patient's satisfaction with the following characteristics: Age=35, Severity=42, Anxiety=2.1. Interpret the interval.**

We are going to use prediction interval to Obtain an interval estimate for a new patient's satisfaction using the formula below:

$$\hat{\beta}_0 + \hat{\beta}_i * X_n \pm t_{n-2,1-\alpha/2} * se(\hat{\beta}_0 + \hat{\beta}_i * X_n)$$

$$where\ se(\hat{\beta}_0 + \hat{\beta}_i * X_n) = \sqrt{MSE(1/n + \frac{(X_n - \bar{X}^2)}{\sum_{j=1}^{n}(x_j - \bar{x})^2} + 1)}$$

```r
newdata = data.frame(age = 35, severity = 42, anxiety = 2.1)
predict.lm(fit_1, newdata, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 71.68332 50.06237 93.30426
```

Therefore, at the significant level of 0.05, an interval estimate for a new patient's satisfaction with Age = 35, Severity = 42 and Anxiety = 2.1 is (50.06237, 93.30426). We're 95% confident that the true satisfaction score of a new patient whose age is 35, severity of illness is 42 and anxiety level is 21 will lay between 50.06237 and 93.30426.

**5a) Test whether 'anxiety level' can be dropped from the regression model, given the other two. covariates are retained. State the hypotheses, test-statistic, decision rule and conclusion.**

For the question, we are going to first create a new (small) model with anxiety level dropped and we are going to use 'partial' F-test to test whether the small model or the original (large) model is superior. The hypotheses are stated below:

$$H_0 : \beta_{anx} = 0,\ small\ model$$
$$H_1 : \beta_{anx} \neq 0,\ large\ model$$
$$F = \frac{(SSE_L - SSE_S)/(df_L - df_S)}{SSE_L/df_L} = \frac{364.1595}{107.2791} = 3.5997$$
$$F(1-\alpha; df_L - df_S, df_L) = 4.067047$$

```r
#create a small model excluding anxiety level
fit_2 = lm(safisfaction~age + severity, data = pat_satisfaction)

anova(fit_1,fit_2)
```

```
## Analysis of Variance Table
##
## Model 1: safisfaction ~ age + severity + anxiety
## Model 2: safisfaction ~ age + severity
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     42 4248.8
## 2     43 4613.0 -1   -364.16 3.5997 0.06468 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
qf(0.95,1,43)
```

```
## [1] 4.067047
```

From the 'partial' F-test, we got the F statistic 3.5997 which is less the critical value 4.0670474. Hence, we fail to reject the null hypothesis and we can conclude that at significance level of 0.05, the small model is 'superior.' Therefore, the variable 'anxiety level' can be dropped from the regression model, given the other two covariates are retained.

### 5b) How are R2/R2-adjusted impacted by the action that you took in part 5-a)?

```r
summary(fit_1)
```

```
##
## Call:
## lm(formula = safisfaction ~ age + severity + anxiety, data = pat_satisfaction)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.3524  -6.4230   0.5196   8.3715  17.1601
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 158.4913    18.1259   8.744 5.26e-11 ***
## age          -1.1416     0.2148  -5.315 3.81e-06 ***
## severity     -0.4420     0.4920  -0.898   0.3741
## anxiety     -13.4702     7.0997  -1.897   0.0647 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.06 on 42 degrees of freedom
## Multiple R-squared:  0.6822, Adjusted R-squared:  0.6595
## F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

```r
summary(fit_2)
```

```
##
## Call:
## lm(formula = safisfaction ~ age + severity, data = pat_satisfaction)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.1662  -8.5462  -0.4595   7.1342  17.2364
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 156.6719    18.6396   8.405 1.27e-10 ***
## age          -1.2677     0.2104  -6.026 3.35e-07 ***
## severity     -0.9208     0.4349  -2.117   0.0401 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.36 on 43 degrees of freedom
## Multiple R-squared:  0.655,  Adjusted R-squared:  0.6389
## F-statistic: 40.81 on 2 and 43 DF,  p-value: 1.16e-10
```

As we dropped the 'anxiety level' from the regression model, both the $R^2$ and the adjusted $R^2$ decreased from 0.6822 to 0.655 and from 0.6595 to 0.6389 respectively. Such effect makes sense as we include more predictors in the regression model, the coefficient of determination will always increase.

---

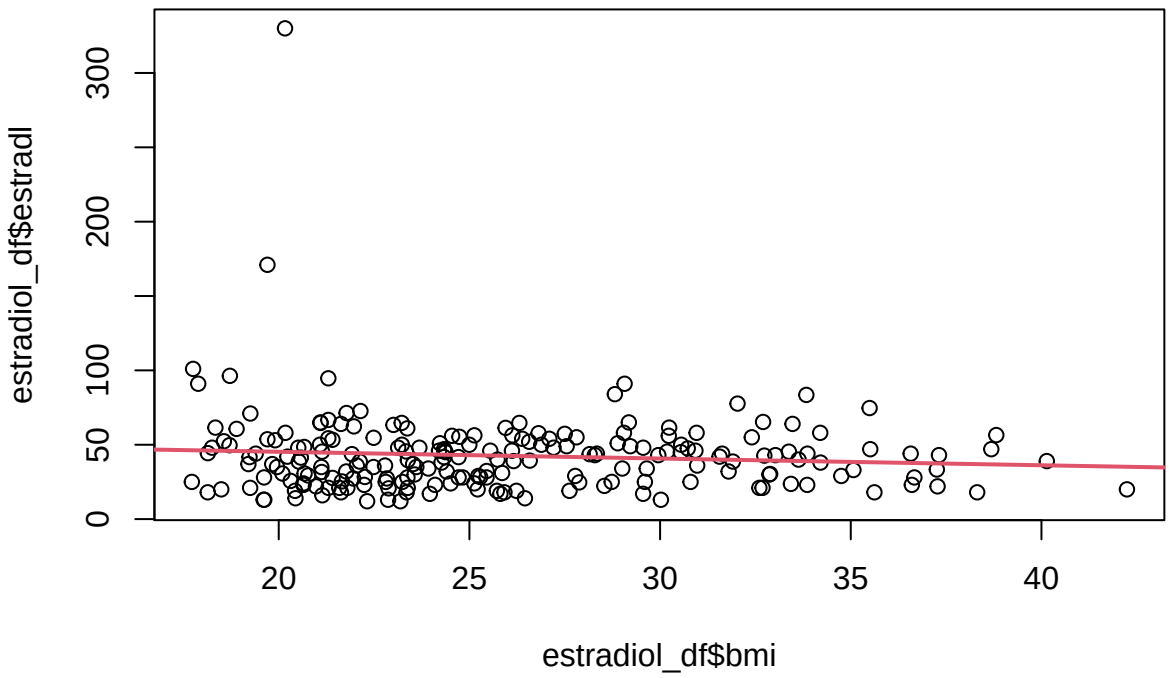
**Problem 2**

```r
#load the data
estradiol_df = read_csv("./data/Estradl.csv") %>%
  janitor::clean_names()
```

**1) Is there a crude association between BMI and serum estradiol**

**a) Generate a scatter plot with the overlaid regression line. Comment.**

```r
fit_3 = lm(estradl~bmi, data = estradiol_df)
plot(estradiol_df$bmi, estradiol_df$estradl)
abline(fit_3,lwd=2,col=2)
```




From the scatterplot, the fitted line has a negative slope but the line is almost horizontal which means that the estimate slope coefficient is closed to 0. Hence it is hard to identify an association between BMI and serum estradiol.

**b) Provide the summary regression output and comment on the nature of the relationship (i.e.,sign, magnitude, significance).**

```
summary(fit_3)
```

```
##
## Call:
## lm(formula = estradl ~ bmi, data = estradiol_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.432 -15.903  -2.209   8.758 284.822
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  54.3095     9.5054   5.714  3.8e-08 ***
## bmi          -0.4529     0.3605  -1.256     0.21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.19 on 208 degrees of freedom
```

```
## Multiple R-squared:  0.007529,   Adjusted R-squared:  0.002758
## F-statistic: 1.578 on 1 and 208 DF,  p-value: 0.2105
```

From the regression summary, the estimate coefficient for bmi is -0.4529 which means that there is a negative relationship between bmi and serum estradiol and as bmi increase by 1, the average serum estradiol level will decrease by 0.4529. However, the p-value for bmi is 0.21 which is greater than 0.05.As a result, we can conclude that at the significant level of 0.05, there is no significant linear association between BMI and serum estradiol.

**2) How does the relationship between BMI and serum estradiol change after controlling for all the other risk factors listed above? Provide the summary regression output and comment on the relationships observed for each of the predictors.**

```
#MLR model with all risk factors
fit_4 = lm(estradl~ bmi + factor(ethnic) + entage + numchild + agemenar , data = estradiol_df)
summary(fit_4)
```

```
##
## Call:
## lm(formula = estradl ~ bmi + factor(ethnic) + entage + numchild +
##     agemenar, data = estradiol_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.561 -15.279  -4.652   9.962 271.230
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       42.2147    12.5117   3.374 0.000887 ***
## bmi               -0.1066     0.3702  -0.288 0.773727
## factor(ethnic)1 -16.0579     4.4492  -3.609 0.000386 ***
## entage            0.5180     0.3587   1.444 0.150259
## numchild         -0.4906     1.2444  -0.394 0.693788
## agemenar          0.1073     0.1691   0.635 0.526429
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.4 on 204 degrees of freedom
## Multiple R-squared:  0.08063,    Adjusted R-squared:  0.0581
## F-statistic: 3.578 on 5 and 204 DF,  p-value: 0.004007
```

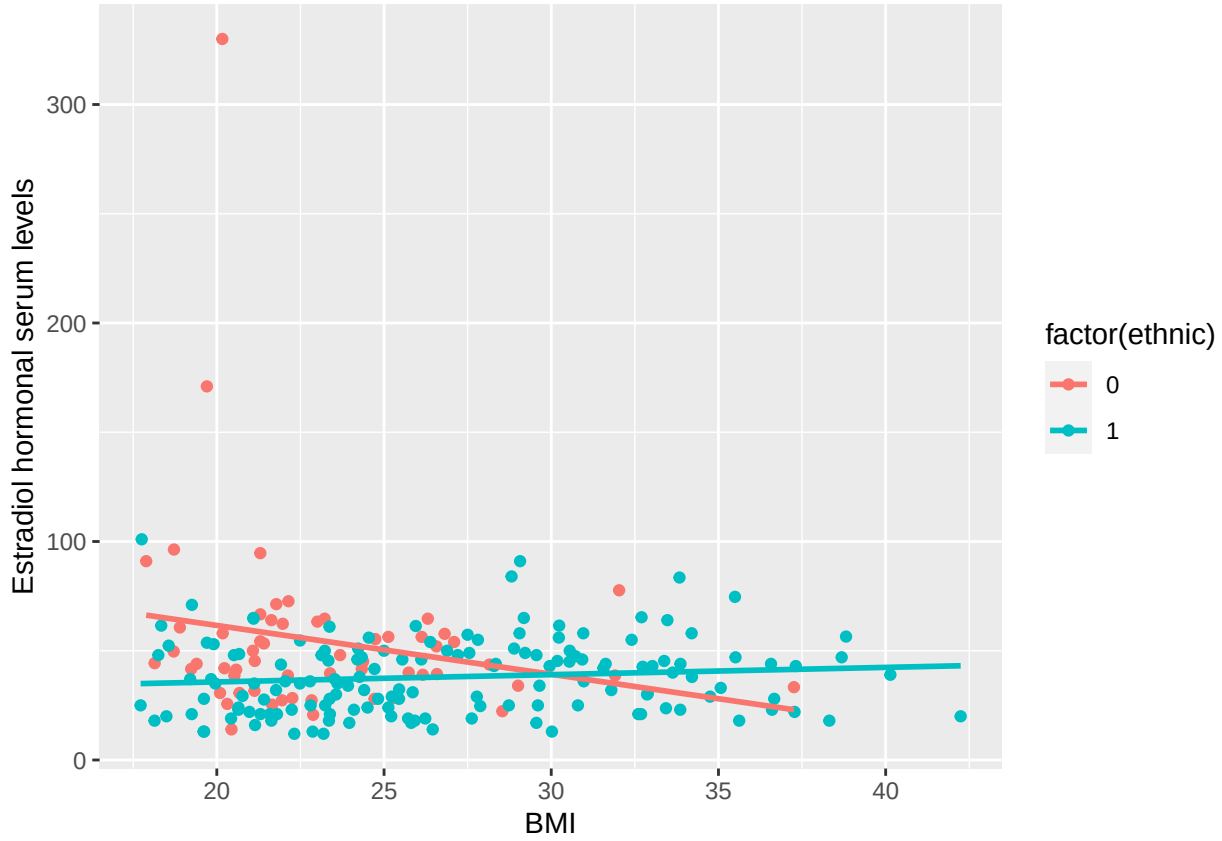
After we controlled for all the other risk factors, the coefficient of BMI increases from -0.4529 to -0.1066, and the p-value for BMI is 0.773727 which is still greater than 0.05. We could conclude that there is still no significant linear association between Estradiol and BMI after controlled for all the other risk factors. As for the new MLR model,both $R^2$ and adjusted $R^2$ slightly increased comparing to the SLR model and the p-value for the MLR model is 0.004007 for the F-statistic which indicates that there is a relationship between outcome and the set of covariate. On the other hand, ethnicity is the only predictor with p-value less than 0.05 which indicates there is a positive association between the 'estradiol' and 'ethnic' with a coefficient of 0.5180. Other predictors such as entage, numchild and agemenar have no significant linear association with serum estradiol.

**3) Now focus only the relationship between BMI and serum estradiol by ethnicity. Is there any evidence that these relationships vary for African American and Caucasian women?**

**a) Use graphical displays and numerical summaries to sustain your conclusion.**

```
#graphical display for interaction
qplot(x = bmi, y = estradl, data = estradiol_df, color = factor(ethnic)) +
     geom_smooth(method = "lm", se=FALSE) +
     labs(x="BMI", y="Estradiol hormonal serum levels")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
#regression on the interaction term
fit_5 = lm(estradl~ bmi*factor(ethnic), data = estradiol_df)
summary(fit_5)
```

```
##
## Call:
## lm(formula = estradl ~ bmi * factor(ethnic), data = estradiol_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -46.60 -15.21  -3.38  10.12 268.79
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          106.2850    22.3276   4.760 3.64e-06 ***
```

```
## bmi                  -2.2352     0.9507  -2.351   0.0197 *
## factor(ethnic)1     -77.2104    24.7838  -3.115   0.0021 **
## bmi:factor(ethnic)1   2.5679     1.0285   2.497   0.0133 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.03 on 206 degrees of freedom
## Multiple R-squared:  0.09631,    Adjusted R-squared:  0.08315
## F-statistic: 7.318 on 3 and 206 DF,  p-value: 0.0001099
```

After we fit separate regression models for both ethnicity,from the graphic display, we observed two fitted line intersected with each other which is the indication for interaction. From the numerical summaries, we saw that the p-value for the interaction term (bmi*ethnic) and the MLR model are both less then 0.05.Once again, we observed interaction between bmi and ethnic. In other words, the relationship between serum estradiol and BMI vary for African American and Caucasian women.

**b) Based on your response in part 3-a), take additional steps to quantify the relationship between BMI and serum estradiol by ethnicity. Comment on your findings.**

Based on the results from part 3-a), we are going to do a stratified analysis to quantify the relationship between BMI and serum estradiol by ethnicity.

```
caucasian_df = estradiol_df %>%
  filter(ethnic == 0)

African_American_df = estradiol_df %>%
  filter(ethnic == 1)

fit_6 = lm(estradl~bmi, data = caucasian_df)
broom::tidy(fit_6)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)   106.       35.7       2.98 0.00427
## 2 bmi            -2.24      1.52     -1.47 0.147
```

```
fit_7 = lm(estradl~bmi, data = African_American_df)
broom::tidy(fit_7)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   29.1       6.84       4.25 0.0000374
## 2 bmi            0.333     0.250      1.33 0.184
```

From the stratified analysis, for Caucasian stratum,, the coefficient of BMI is -2.235 and its p-value is 0.14702 which is bigger than 0.05. Hence we conclude that for Caucasian women there is no statistically significant association between BMI and Estradiol; for African American stratum, the coefficient of BMI is 0.3327 and its p-value is 0.184 which is bigger than 0.05. Hence we conclude that for African American women there is no statistically significant association between BMI and Estradiol. There is possibility that ethnicity is a confounder which is associated with BMI and is also associated with Estradiol and ethnicity is not on the causal pathway between BMI and Estradiol. Next we are going identify if ethnicity is a confounder.

```r
#check association between estradiol and ethnicity
fit_8 = lm(estradl~ factor(ethnic), data = estradiol_df)
broom::tidy(fit_8)
```

```
## # A tibble: 2 x 5
##   term            estimate std.error statistic  p.value
##   <chr>              <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)         54.4      3.55     15.3  5.88e-36
## 2 factor(ethnic)1    -16.4      4.19     -3.92 1.19e- 4
```

```r
#check association between bmi and ethnicity
fit_9 = lm(bmi~factor(ethnic), data = estradiol_df)
broom::tidy(fit_9)
```

```
## # A tibble: 2 x 5
##   term            estimate std.error statistic  p.value
##   <chr>              <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)        23.2      0.673     34.5  5.77e-88
## 2 factor(ethnic)1     3.64     0.793      4.59 7.65e- 6
```

```r
#check if ethnicity is in causal pathway
fit_10 = lm(estradl~bmi + factor(ethnic), data = estradiol_df)
broom::tidy(fit_10)
```

```
## # A tibble: 3 x 5
##   term            estimate std.error statistic       p.value
##   <chr>              <dbl>     <dbl>     <dbl>         <dbl>
## 1 (Intercept)      55.4        9.23      6.00  0.00000000869
## 2 bmi              -0.0412     0.367    -0.112 0.911
## 3 factor(ethnic)1 -16.3        4.41     -3.70  0.000281
```

From SLR model fit_8 and fit_9, we observed significant associations between estradiol and ethnicity and between bmi and ethnicity. After we included ethnicity as second predictor into the regression model of estradiol on bmi(fit_10), the coefficient of BMI predictor is greatly distorted comparing to the coefficient of BMI predictor in SLR model(fit_3). Therefore, we can conclude that ethnicity is a confounder of relationship between BMI and Estradiol hormonal serum levels.In conclusion, the ethnicity variable not only interact with BMI variable, but also is a confounder to the relationship between BMI and serum estradiol.