Homework 6, Biostatistical Methods

Emil Hafeez (eh2928)

11/29/2020

First, let's read in the data.

```
pat_sat =
  read.csv("./data/PatSatisfaction.csv") %>%
  janitor::clean_names() %>%
  rename(satisfaction = safisfaction)
```

Problem 1 (15p)

Problem 1.1.

The correlation matrix refers to the array of numbers where r_{jk} is the pearson correlation coefficient between variables x_j and x_j such that

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1p} \\ r_{21} & 1 & r_{23} & \cdots & r_{2p} \\ r_{31} & r_{32} & 1 & \cdots & r_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & r_{p3} & \cdots & 1 \end{pmatrix}$$

As such, the correlation coefficient for the all variables "patient's satisfaction score" (the outcome), "age", "severity of illness", and "anxiety level" is as follows.

```
# Correlation matrix for all variables
cor(pat_sat) %>%
  knitr::kable(
            align = "cccc",
            digits = 3)
```

	satisfaction	age	severity	anxiety
satisfaction	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

In regards to these values, the predictors each show moderate to strong negative correlation with the outcome variable. As such, it appears that an increase in age, severity of illness, or anxiety level is correlated with a decrease in satisfaction. We may also make a note that there is correlation between the predictors, a multicollinearity concern.

Problem 1.2. THIS NEEDS REVISITING

Fit a MLR with all 3 predictors and test whether at least one is significant.

```
fit_patsat = lm(satisfaction ~ age + severity + anxiety, data = pat_sat)
anova(fit_patsat)
```

```
## Analysis of Variance Table
##
## Response: satisfaction
##           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
```

Note that in this code, the tests for each term are conditioned for everything else above them in the output.

Hypotheses: $H_0 : \beta_1 = \beta_2 = \beta_3$

H_A : at least one β is not zero.

Test statistic and decision rule is given by:

$$\begin{array}{l} F = \frac{MSR}{MSE} > F_{1-\alpha;p,n-p-1}, \text{ reject } H_0 \\ F = \frac{MSR}{MSE} \leq F_{1-\alpha;p,n-p-1}, \text{ fail to reject } H_0 \end{array}$$

In our case, the test statistic for all of the predictors is $F = 3.5997$, and the critical value is given by `qf(0.99, 3, 46-3)`, $F_{1-\alpha;p,n-p-1} = 4.27265$

Therefore, we fail to reject the null hypothesis and conclude that at least one of the predictors in the model is not significant in association with outcome variable (satisfaction).

Problem 1.3.

Show the regression results for all estimated slope coefficients with 95% CIs.

```
summary(fit_patsat)
```

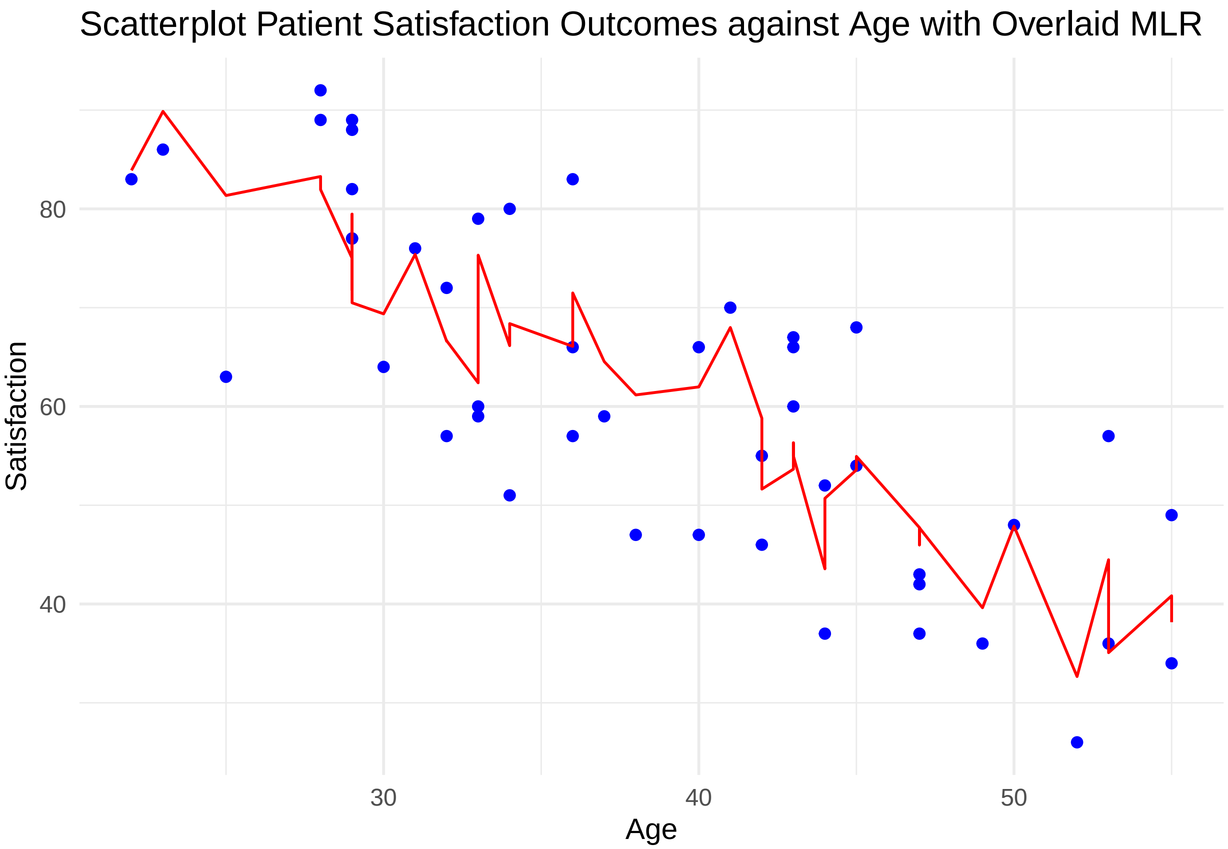
```
##
## Call:
## lm(formula = satisfaction ~ age + severity + anxiety, data = pat_sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.3524  -6.4230   0.5196   8.3715  17.1601
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 158.4913    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
```

```
predicted_df = data.frame(patsat_pred = predict(fit_patsat, data = pat_sat), age=pat_sat$age)

pat_sat %>%
  ggplot(aes(x = age, y = satisfaction)) +
  geom_point(color = 'blue') +
  geom_line(color='red', data = predicted_df, se=TRUE, aes(x = age, y = patsat_pred)) +
  labs(title =
         "Scatterplot Patient Satisfaction Outcomes against Age with Overlaid MLR",
       x = "Age",
       y="Satisfaction")
```

```
## Warning: Ignoring unknown parameters: se
```



Here we can see that the coefficient for the "severity of illness" variable is equal to -0.4420, implying that a one unit increase in the severity of illness variable is associated with a 0.4420 decrease in patient satisfaction rating (the outcome). Another way to think of this is that as a patient's severity of illness rating increases by 1, there is an associated decrease of 0.4420 unit of patient satisfaction as an outcome.

The 95% confidence interval for true slope of the "severity of illness" coefficient β_2 is given by $\widehat{\beta_2} \pm t_{n-2,1-(\alpha/2)} \cdot se(\widehat{\beta_2})$where $se(\widehat{\beta_2}) = \sqrt{MSE/\Sigma_{i=1}^{n}}(X_i - \overline{X})^2$.

```
tidy(fit_patsat)
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  158.       18.1       8.74  5.26e-11
## 2 age           -1.14      0.215    -5.31  3.81e- 6
## 3 severity      -0.442     0.492    -0.898 3.74e- 1
## 4 anxiety      -13.5       7.10     -1.90  6.47e- 2
```

```
qt(0.975,44)
```

```
## [1] 2.015368
```

Seeing as $t_{n-2,1-(\alpha/2)} = 2.015368$, in our context, the 95% confidence interval for the true slope is equal to $-0.4420043 \pm 2.015368 \cdot 0.4919657 = (-1.433496, 0.5494876)$. As such, we are 95% confident that as patient severity of illness increases by one unit, the true value of the associated change in satisfaction is between (-1.433496, 0.5494876) points. This overlaps the null value of 0, implying that there may be no true association between patient severity of illness and satisfaction.

Problem 1.4.

The 95% confidence interval for a new patient's satisfaction when they have age = 35, severity of illness = 42, and anxiety = 2.1 is given by:

$$\widehat{\beta_0}+\widehat{\beta_1}+\widehat{\beta_2}+\widehat{\beta_3}X_h \pm t_{n-2,1-\alpha/2}\cdot \text{se}\left(\widehat{\beta_0}+\widehat{\beta_1}X_h\right)$$

$$\text{se}\left(\widehat{\beta_0}+\widehat{\beta_1}X_h\right)=\sqrt{MSE\left\{\frac{1}{n}+\left[\left(X_h-\bar{X}\right)^2/\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right]\right\}}$$

```
fit_patsat = lm(satisfaction ~ age + severity + anxiety, data = pat_sat)

data_to_predict_from = data.frame(age = 28, severity = 42, anxiety = 2.1)

predict(fit_patsat, data_to_predict_from, interval = "confidence")
```

```
##       fit      lwr      upr
## 1 79.6746 72.39005 86.95914
```

In context, this means that with 95% confidence we estimate the mean value of a new patient's satisfaction when they have age = 35, severity of illness = 42, and anxiety = 2.1 is given by between 72.39005 and 86.95914 units (72.39005, 86.95914).

Problem 1.5.a.

First, we fit the two nested models.

```
small_patsat_fit = lm(satisfaction ~ age + severity, data = pat_sat)
large_patsat_fit = lm(satisfaction ~ age + severity + anxiety, data = pat_sat)
```

Note that we are comparing the two models here:

Model 1, without the anxiety variable: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$

Model 2, with the anxiety variable: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}$

Note that Model 1 is a subset of Model 2.

The null hypothesis is to retain the the smaller model, and the alternate hypothesis is to utilize the larger model. This is also to say, $H_0 : \beta_3 = 0$, and $H_A : \beta_3 \neq 0$

The test statistic F is given by the following,

$F = \frac{(SSR_L - SSR_S)/(df_L - df_S)}{\frac{SSE_L}{df_L}} \sim F_{df_L - df_S, df_L}$

where $df_S = n - p_s - 1, df_L = n - p_L - 1$.

This can also be written as

$F = \frac{(SSE_S - SSE_L)/(df_S - df_L)}{\frac{SSE_L}{df_L}}$.

The decision rule is given by

$$\begin{array}{ll} F = \frac{MSR}{MSE} > F_{1-\alpha;df_L - df_S, df_L}, & \text{reject } H_0 \\ F = \frac{MSR}{MSE} \leq F_{1-\alpha;df_L - df_S, df_L}, & \text{fail to reject } H_0 \end{array}$$

```r
anova(small_patsat_fit, large_patsat_fit) %>%
  tidy()
```

```
## # A tibble: 2 x 6
##   res.df   rss    df sumsq statistic p.value
##    <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl>
## 1     43 4613.    NA    NA      NA     NA
## 2     42 4249.     1  364.      3.60  0.0647
```

Given F = 3.599735, we fail to reject the null hypothesis and conclude that we retain the smaller model, and so do not include the anxiety variable in our MLR. We discard it.

Problem 1.5.b.

The R^2 and adjusted R^2 in the former, larger model are respectively 0.6821943 and 0.6594939

The R^2 and adjusted R^2 in the latter, smaller model where we do not include the anxiety variable are respectively 0.6549559 and 0.6389073.

Therefore, the action we took (dropping the anxiety variable) produces a marginally lower R^2 and adjusted R^2 than previously.

```r
lm(satisfaction ~ age + severity + anxiety, data = pat_sat) %>%
  glance()
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.682         0.659  10.1      30.1 1.54e-10     3  -169.  349.  358.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```r
lm(satisfaction ~ age + severity, data = pat_sat) %>%
  glance()
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.655         0.639  10.4      40.8 1.16e-10     2  -171.  351.  358.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Problem 2 (15p)

First let's read in the data.

```r
estradiol_df =
  read.csv("./data/ESTRADL.csv") %>%
  janitor::clean_names() %>%
  rename(estradiol = estradl) %>%
  rename(age = entage)
```

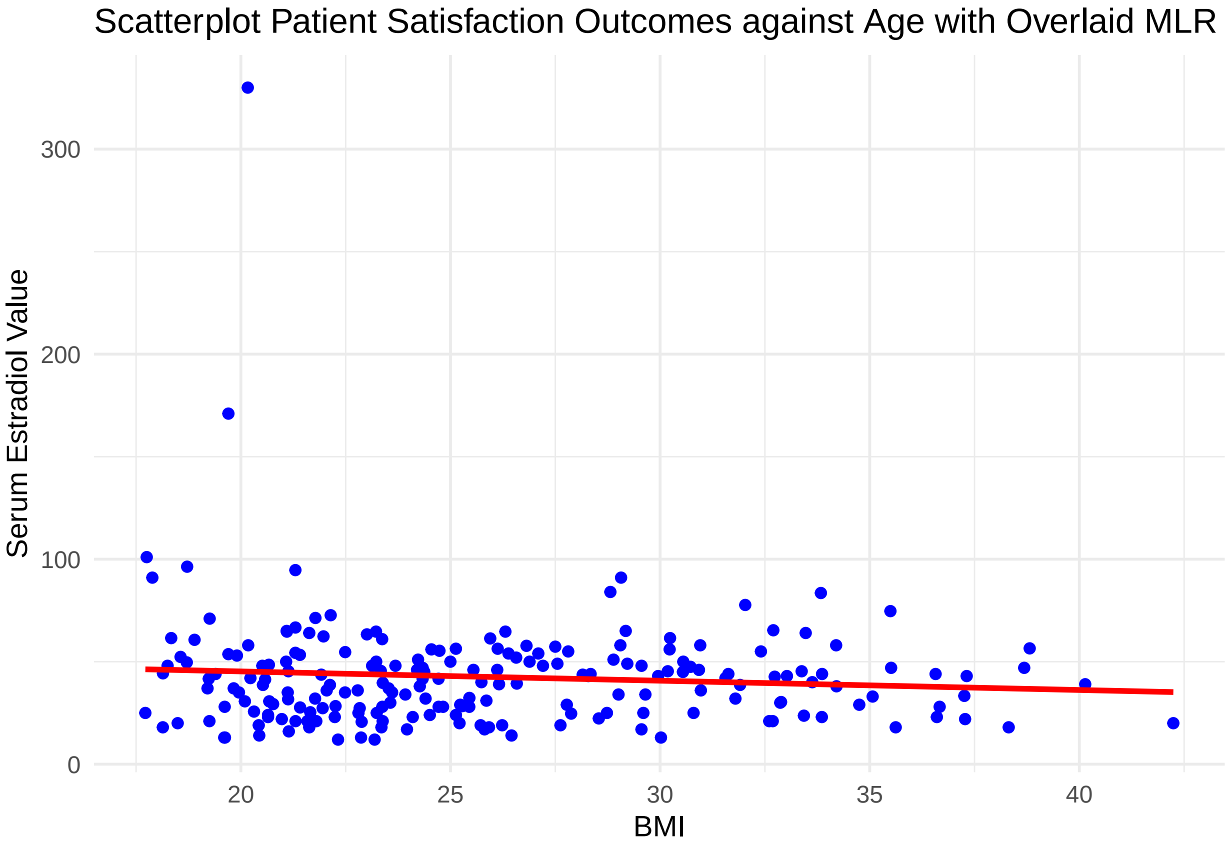
Problem 2.1.

Problem 2.1.a.

Generate a scatter plot with the overlaid regression line. Comment. (2.5p)

```
# Scatter plot with regression line overlaid
estradiol_df %>%
  ggplot(aes(x = bmi, y = estradiol)) +
  geom_point(color = 'blue') +
  geom_smooth(method = 'lm', color = 'red', se = F) +
  labs(title =
         "Scatterplot Patient Satisfaction Outcomes against Age with Overlaid MLR",
       x = "BMI",
       y = "Serum Estradiol Value")
```

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



From this plot, we can see that there is perhaps a small or unclear relationship between the predictor, BMI, and the outcome, estradiol. There are a few large outliers potentially obscuring the regression, however.

Problem 2.1.b.

```
fit_estradiol = lm(estradiol ~ bmi, data = estradiol_df)
summary(fit_estradiol)
```

```
##
## Call:
## lm(formula = estradiol ~ 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 this SLR output, we can see that there is a negative, small in magnitude, and statistically insignificant relationship between BMI and serum estradiol level. This is to say, if BMI increases by, there is a minor decrease in serum estradiol levels (0.4529) (though this relationship is, again, not significant).

Problem 2.2.

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. (5p)

```r
fit_mlr_estradiol = lm(estradiol ~ bmi + ethnic + age + numchild + agemenar, data = estradiol_df)

summary(fit_mlr_estradiol)
```

```
##
## Call:
## lm(formula = estradiol ~ bmi + ethnic + age + 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
## ethnic      -16.0579     4.4492  -3.609 0.000386 ***
## age           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
```

Now, we can see a different relationship between BMI and serum estradiol. Previously, we saw a statistically insignificant decrease that when BMI increases by one unit, there was to be an associated reduction in serum estradiol levels of 0.4529 units. Now, when we adjust for other predictors (namely ethnicity, age, the number of children the woman has had, and their age at menarche), when BMI increases by one unit we anticipate a decrease in serum estradiol of 0.1066 units; this is also a statistically insignificant relationship.

Regarding other predictors: we show a statistically significant relationship between ethnic status (African American versus Caucasian) and serum estradiol, where we anticipate someone of African American ethnicity will have 16.0579 serum estradiol units lower than someone of Caucasian ethnicity (controlling for other predictors).

We see a one unit increase in age (years) associated with a 0.5180 increase in serum estradiol (statistically insignificant). We also see that when the number of children the woman has had increases by one, we associate a 0.4906 units decrease in serum estradiol (also statistically insignificant). Lastly, when the age of menarche increases by one year, we see an associated increase in serum estradiol of 0.1073 units.

See these relationships printed more neatly below.

```r
tidy(fit_mlr_estradiol) %>%
  knitr::kable(
              col.names =
                c("Term Name", "Coeff Estimate", "Std Err", "Test Statistic", "p Value"),
              align = "ccccc",
              digits = 3)
```

Term Name	Coeff Estimate	Std Err	Test Statistic	p Value
(Intercept)	42.215	12.512	3.374	0.001
bmi	-0.107	0.370	-0.288	0.774
ethnic	-16.058	4.449	-3.609	0.000
age	0.518	0.359	1.444	0.150
numchild	-0.491	1.244	-0.394	0.694
agemenar	0.107	0.169	0.635	0.526

Problem 2.3.

Problem 2.3.a.

```r
#Examining a stratified analysis, we could fit separate regressions for each of the 2 ethnicities in th

#first, is ethnicity associated with the outcome?
fit_ethnic_estradiol = lm(formula = estradiol ~ factor(ethnic), data = estradiol_df)
summary(fit_ethnic_estradiol)
```

```
##
## Call:
## lm(formula = estradiol ~ factor(ethnic), data = estradiol_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.448 -15.089  -3.501   9.999 275.552
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       54.448      3.555  15.317  < 2e-16 ***
## factor(ethnic)1  -16.447      4.192  -3.923 0.000119 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.3 on 208 degrees of freedom
## Multiple R-squared:  0.06891,    Adjusted R-squared:  0.06443
## F-statistic: 15.39 on 1 and 208 DF,  p-value: 0.0001186
```

There is a clear relationship between ethnicity and serum estradiol. Now, we examine whether the relationship between BMI and serum estradiol varies by ethnicity by stratifying our previous regression by ethnicity status.

Now, we examine whether the relationship between BMI and serum estradiol differs by the levels of ethnicity variable.

```
fit_estradiol_inter = lm(estradiol ~ bmi * ethnic, data = estradiol_df)
summary(fit_estradiol_inter)
```

```
##
## Call:
## lm(formula = estradiol ~ bmi * 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 *
## ethnic      -77.2104    24.7838  -3.115   0.0021 **
## bmi:ethnic    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
```

Here, we see a significant interaction; the relationship between BMI and serum estradiol seems to differ by levels of ethnicity, such that status in the African American ethnic group is associated with a decrease in serum estradiol. We investigate further using a stratified analysis.

```
fit_estradiol_white = lm(estradiol ~ bmi, data = estradiol_white_df)
summary(fit_estradiol_white)
```

```
##
## Call:
## lm(formula = estradiol ~ bmi, data = estradiol_white_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.600 -20.786  -6.804   8.138 268.787
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  106.285     35.706   2.977  0.00427 **
## bmi           -2.235      1.520  -1.470  0.14702
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.22 on 57 degrees of freedom
## Multiple R-squared:  0.03653,    Adjusted R-squared:  0.01963
## F-statistic: 2.161 on 1 and 57 DF,  p-value: 0.147
```

We see a low, negative association between BMI and serum estradiol levels, where each increase in BMI unit is associated with a 2.235 decrease in serum estradiol levels, though this relationship is not statistically significant.

```
fit_estradiol_aa = lm(estradiol ~ bmi, data = estradiol_aa_df)
summary(fit_estradiol_aa)
```

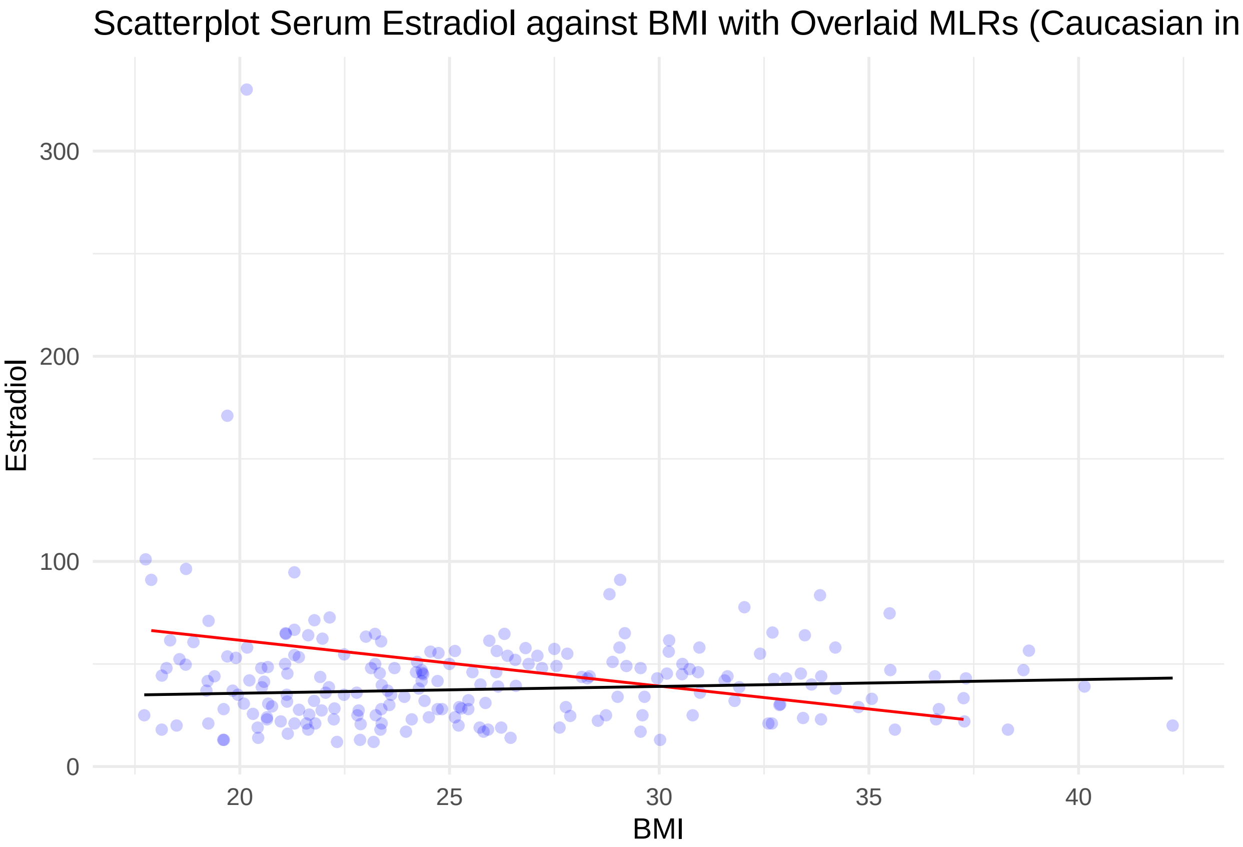
```
##
## Call:
## lm(formula = estradiol ~ bmi, data = estradiol_aa_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -26.06 -13.99  -1.10  11.00  66.02
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29.0746     6.8392   4.251 3.74e-05 ***
## bmi           0.3327     0.2495   1.333    0.184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.18 on 149 degrees of freedom
## Multiple R-squared:  0.01179,    Adjusted R-squared:  0.005159
## F-statistic: 1.778 on 1 and 149 DF,  p-value: 0.1844
```

We see a low, positive association between BMI and serum estradiol levels, where each increase in BMI unit is associated with a 0.3327 increase in serum estradiol levels, though this relationship is not statistically significant.

There is not strong evidence to support the relationship between BMI and serum estradiol varies for African American and Caucasian women when only examining these three variables.

We can graph these relationships; let's take a look at serum estradiol as a function of BMI, with separate regressions for Caucasians (in red) and African Americans (in Black).

```
estradiol_df %>%
  ggplot(aes(x = bmi, y = estradiol)) +
  geom_point(color = 'blue', alpha = 0.2) +
  geom_line(aes(x = bmi, y = estradiol_white_predicted), color='red', data = predicted_white_pred_df) +
  geom_line( aes(x = bmi, y = estradiol_aa_predicted), color='black', data = predicted_aa_pred_df) +
  labs(title =
         "Scatterplot Serum Estradiol against BMI with Overlaid MLRs (Caucasian in Red)",
       x = "BMI",
       y="Estradiol")
```


Scatterplot Serum Estradiol against BMI with Overlaid MLRs (Caucasian in
Estradiol
300
200
100
0
20
25
30
35
40
BMI

Problem 2.3.b.