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. We will explore this further.

Problem 1.2.

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:

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

In our case, the test statistic for all of the predictors is $F = 3.5997$, and the critical value is given by `qf(0.95, 3, 46-3-1)`, $F_{1-\alpha;p,n-p-1} = 2.827049$

Therefore, at the alpha $= 0.05$ we reject the null hypothesis and conclude that at least one of the predictors (beta's) in the model is 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)
```

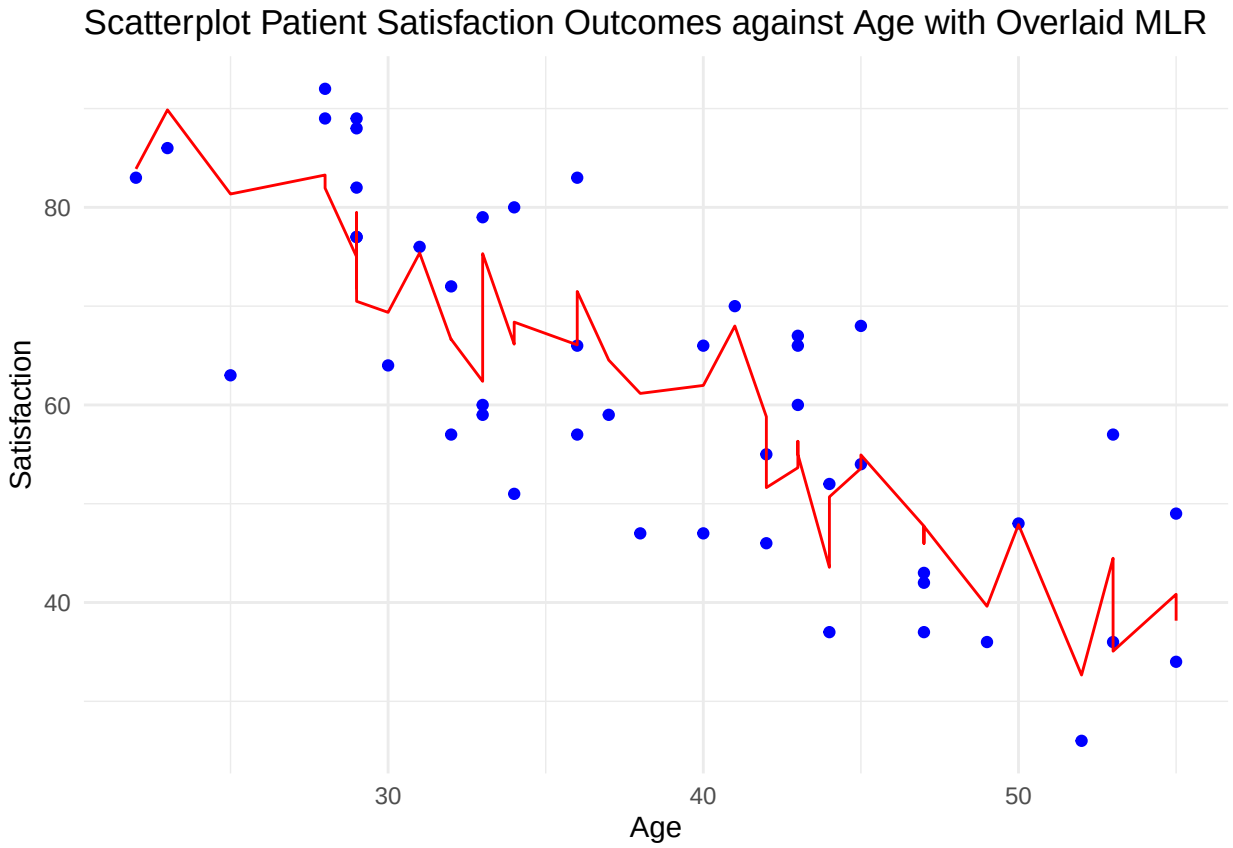
```
##
## Call:
## lm(formula = satisfaction ~ age + severity + anxiety, data = pat_sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.3524  -6.4230   0.5196   8.3715  17.1601
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 158.4913   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



The 95% confidence intervals for each of the coefficients from the regression are as follows.

```
confint(fit_patsat, 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
```

From the regression output, 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).

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 $se(\widehat{\beta_2}) = \sqrt{MSE/\Sigma_{i=1}^{n}}(X_i - \overline{X})^2$.

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 (-1.434831, 0.5508228) 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.

We are examining the the 95% interval for a specific new patient's satisfaction when that patient has age = 35, severity of illness = 42, and anxiety = 2.1. As such, we are calculating the prediction interval given by

The 95% prediction interval for Anne's freshman GPA is calculated as below.

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

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

In context, this means that with 95% confidence we predict the true value of the specific new patient's satisfaction when they have age = 35, severity of illness = 42, and anxiety = 2.1 to be between 50.06237 and 93.30426 units (50.06237, 93.30426). Notice how wide a prediction interval is, versus a comparable confidence interval for the mean value of any new patient's satisfaction who meet those criteria, because the prediction interval focuses on one specific new value of Y_h, and since we do not calculate an expected mean the errors do not reduce to 0, and so the SE formula for prediction includes a +1 in the denominator, widening the interval overall.

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

data_to_predict_from = data.frame(age = 35, severity = 42, anxiety = 2.1)

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

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

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

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

```
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 and p = 0.06467813, 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. Said otherwise, the larger model including the anxiety variable produces a very marginal increase in the R^2 and adjusted R^2 such that it is not strong evidence to retain the factor either. We can use the 5% change heuristic as presented in lecture to interpret this change as very small.

```
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>
```

```
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.

First, let's build the scatterplot and overlay the SLR line.

From this plot, it appears there is perhaps a small negative relationship between the predictor, BMI, and the outcome, estradiol. There are a few large outliers, however, and we have not examined other influences.

Problem 2.1.b.

Now, let's look at the SLR output itself.

```r
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 the SLR output, we can see that there is a negative, small in magnitude, and statistically insignificant relationship ($p = 0.21$) between BMI and serum estradiol level. This is to say, if BMI increases by 1 unit, there is an associated minor decrease in serum estradiol levels (by 0.4529), though this relationship is, again, not significant.

Problem 2.2.

Now, let's fit an MLR using the other predictors. How does the BMI-estradiol relationship change after adjusting for the other factors?

```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
```

Well, 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 instead, 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; a lower decrease than previously anticipated. However, this is also a statistically insignificant relationship.

Let's look at the other predictors as well.

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). This is notable and the only statistically significant predictor.

We see a one unit increase in age (years) associated with a 0.5180 increase in serum estradiol (statistically insignificant), when adjusting for other predictors. 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) (when adjusting for other predictors). Lastly, when the age of menarche increases by one year, we see an associated increase in serum estradiol of 0.1073 units (when adjusting for other predictors).

Problem 2.3.

Problem 2.3.a.

Now, let's focus back on just the relationship primarily between BMI and serum estradiol.

We saw a significant interaction for ethnicity, indicating that the relationship between BMI and serum estradiol may differ for African American and Caucasion women. There is evidence for this, though not conclusive. Let's investigate further.

First, is ethnicity associated with the outcome?

```r
#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
```

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.

Now, we examine whether the relationship between BMI and serum estradiol varies by ethnicity by stratifying our previous regression by ethnicity status. We may note that when we fit with collinear variables, we inflate the standard errors for each collinear variable; this consequently decreases the test statistic, which further clouds the significant level reached. Since ethnicity had such a strong relationship with the outcome variable in the SLR performed prior, let's investigate by stratifying on ethnicity using an SLR otherwise.

```r
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 among Caucasian women, where each increase in BMI unit is associated with a 2.235 decrease in serum estradiol level, though this relationship is not statistically significant.

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

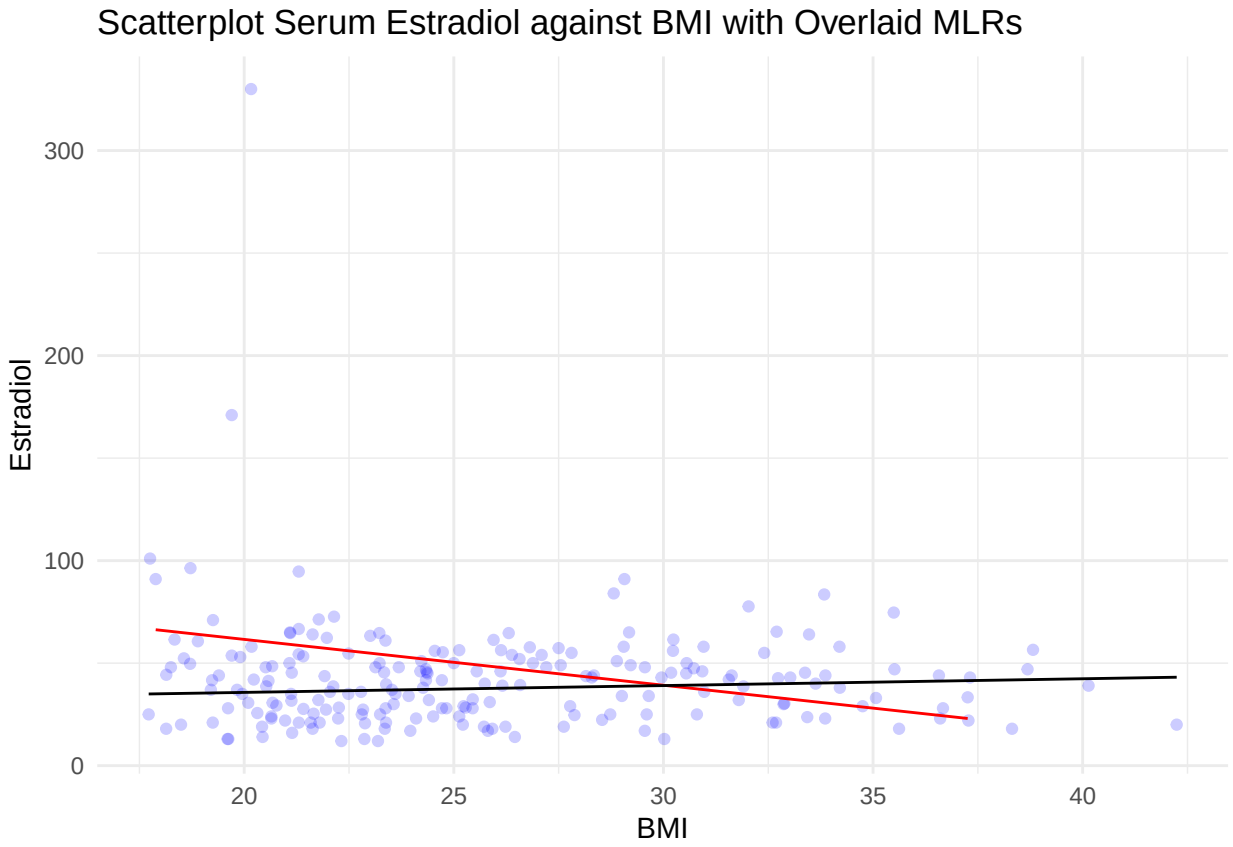
```
##
## Call:
## lm(formula = estradiol ~ bmi, data = estradiol_aa_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -26.06 -13.99  -1.10  11.00  66.02
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29.0746     6.8392   4.251 3.74e-05 ***
## bmi           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 among African American women, where each increase in BMI unit is associated with a 0.3327 increase in serum estradiol levels; this is in contrast to the decrease among Caucasian women, 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 predictors; while we do observe different values for the relationship between BMI and serum estradiol among Caucasian women (-2.235) and African American women (+0.3327), these coefficients are not significant.

Of course, 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",
       x = "BMI",
       y="Estradiol")
```



Notice that when we plot the relationship between BMI and serum estradiol stratified by ethnicity, we can see that these regression lines cross, indicative of an important interaction. However, we saw previously that the interaction term is significant, but when we stratify we do not see a significant association between BMI and serum estradiol among Caucasian or African American women.

Problem 2.3.b.

As noted earlier, we could quantify this relationship further but may be concerned about collinearity when fitting an MLR. Let's examine the collinearity.

```
cor(estradiol_df) %>%
  knitr::kable(
              align = "cccc",
              digits = 3)
```

	id	estradiol	ethnic	age	numchild	agemenar	bmi
id	1.000	-0.247	0.747	0.144	0.038	0.063	0.297
estradiol	-0.247	1.000	-0.262	0.096	-0.024	0.030	-0.087
ethnic	0.747	-0.262	1.000	0.005	0.114	0.061	0.303
age	0.144	0.096	0.005	1.000	0.185	0.110	0.105
numchild	0.038	-0.024	0.114	0.185	1.000	0.343	0.018
agemenar	0.063	0.030	0.061	0.110	0.343	1.000	0.033
bmi	0.297	-0.087	0.303	0.105	0.018	0.033	1.000

We can see from this correlation matrix that while ethnicity has some correlation with BMI, the other predictors do not tend have strong relationships with each other. Thus, let's proceed with fitting MLRs.

We explore the relationship between BMI and serum estradiol by utilizing MLRs stratified on ethnicity.

```
fit_estradiol_white_mlr = lm(estradiol ~ bmi + age + numchild + agemenar, data = estradiol_white_df)
summary(fit_estradiol_white_mlr)
```

```
##
## Call:
## lm(formula = estradiol ~ bmi + age + numchild + agemenar, data = estradiol_white_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.162 -17.827  -3.233   6.698 244.993
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   37.978     60.355   0.629   0.5318
## bmi           -2.856      1.519  -1.880   0.0655 .
## age            2.102      1.038   2.025   0.0478 *
## numchild      -5.834      4.075  -1.432   0.1580
## agemenar       2.351      3.634   0.647   0.5204
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.13 on 54 degrees of freedom
## Multiple R-squared:  0.1327, Adjusted R-squared:  0.06844
## F-statistic: 2.065 on 4 and 54 DF,  p-value: 0.09814
```

We can see that the association between BMI and serum estradiol among Caucasian women, when adjusting for the other predictors, is estimated at a 2.856 decrease in anticipated serum estradiol per increase in BMI unit. However, this relationship is not significant. We obtain the 95% confidence interval for this relationship as (-5.90195911, 0.1899121), when adjusting for other factors. This value overlaps the null value of 0.

```
fit_estradiol_aa_mlr = lm(estradiol ~ bmi + age + numchild + agemenar, data = estradiol_aa_df)
summary(fit_estradiol_aa_mlr)
```

```
##
## Call:
## lm(formula = estradiol ~ bmi + age + numchild + agemenar, data = estradiol_aa_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.842 -13.751  -1.481  11.486  66.724
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.36299    9.35386   3.139  0.00205 **
## bmi          0.33670    0.25218   1.335  0.18390
## age         -0.07076    0.27070  -0.261  0.79416
## numchild     0.70467    0.89141   0.791  0.43051
## agemenar     0.05957    0.10878   0.548  0.58481
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.28 on 146 degrees of freedom
## Multiple R-squared:  0.02142,    Adjusted R-squared:  -0.005387
## F-statistic: 0.7991 on 4 and 146 DF,  p-value: 0.5276
```

We can see that the association between BMI and serum estradiol among African American women, when adjusting for the other predictors, is estimated at a 2.856 decrease in anticipated serum estradiol per increase in BMI unit. However, this relationship is not significant. We obtain the 95% confidence interval for this relationship as (-0.1616919, 0.8351005), when adjusting for other factors. This value overlaps the null value of 0.

We can also graph these stratified MLRs, as we did previously for the SLRs.

In conclusion, when we stratify by ethnicity, both in an SLR setting and when adjusting for other covariates, we do not see significant relationship between BMI and serum estradiol.

```
fit_mlr_estradiol_tester = lm(estradiol ~ bmi*ethnic + ethnic + age + numchild + agemenar, data = estrac
summary(fit_mlr_estradiol_tester)
```

```
##
## Call:
## lm(formula = estradiol ~ bmi * ethnic + ethnic + age + numchild +
##     agemenar, data = estradiol_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.744 -15.684  -3.432  10.041 263.967
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  93.8068    23.5326   3.986 9.36e-05 ***
## bmi          -2.3801     0.9555  -2.491  0.01354 *
## ethnic      -78.9271    24.8070  -3.182  0.00169 **
## age           0.5637     0.3543   1.591  0.11319
## numchild     -0.6459     1.2290  -0.526  0.59977
## agemenar      0.1079     0.1668   0.647  0.51846
## bmi:ethnic    2.6538     1.0306   2.575  0.01074 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.03 on 203 degrees of freedom
## Multiple R-squared:  0.1097, Adjusted R-squared:  0.0834
## F-statistic: 4.169 on 6 and 203 DF,  p-value: 0.0005637
```