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:

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

In our case, the test statistic for all of the predictors is $F = 3.5997$, and the critical value is given by `qf(0.99, 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)
```

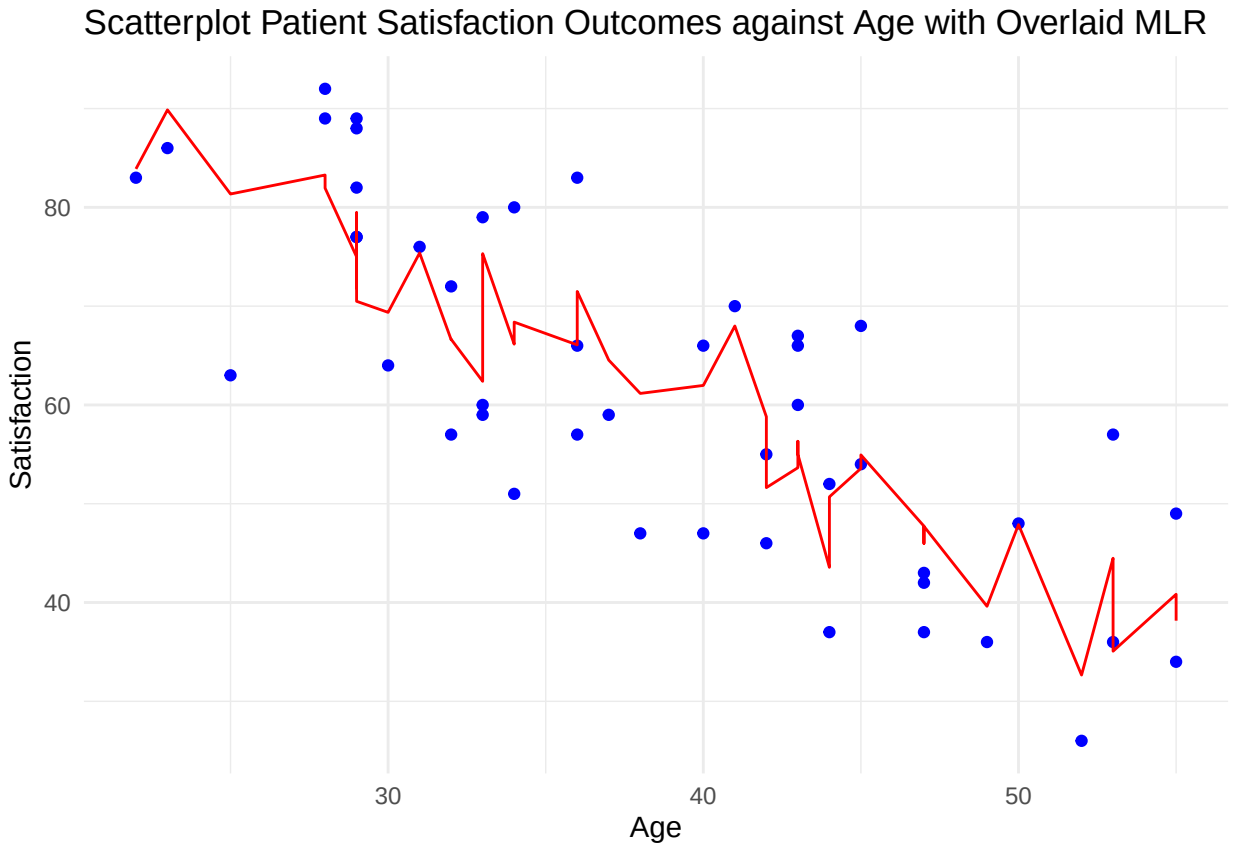
```
##
## Call:
## lm(formula = satisfaction ~ age + severity + anxiety, data = pat_sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.3524  -6.4230   0.5196   8.3715  17.1601
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 158.4913   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

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

```
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, 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.

```
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.

```
estradiol_df =
  read.csv("./data/ESTRADL.csv") %>%
  janitor::clean_names() %>%
```

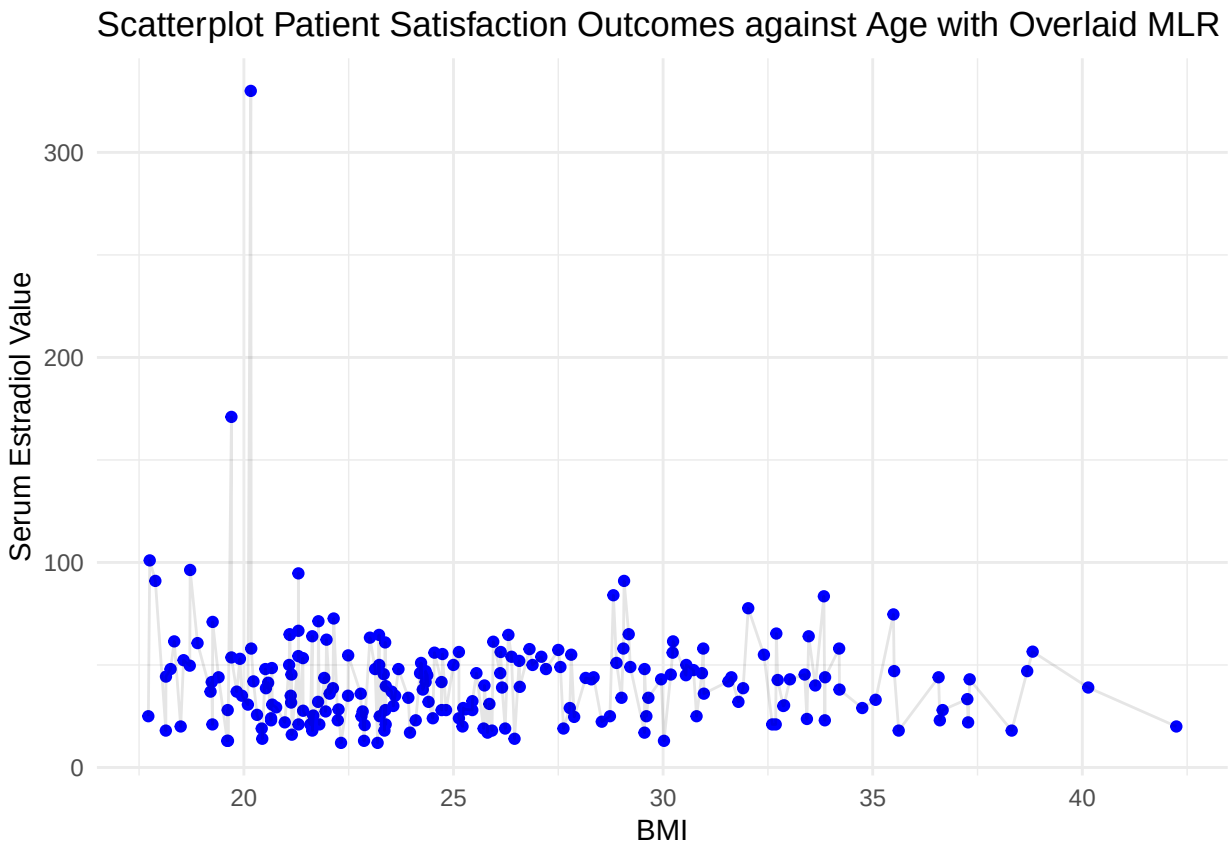
```
  rename(estradiol = estradl) %>%
  rename(age = entage)
```

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_line(alpha = .1) +
  labs(title =
         "Scatterplot Patient Satisfaction Outcomes against Age with Overlaid MLR",
       x = "BMI",
       y = "Serum Estradiol Value")
```



Problem 2.1.b.

Problem 2.2.

Problem 2.3.

Problem 2.3.a.

Problem 2.3.b.