p8130_hw6_rw2844

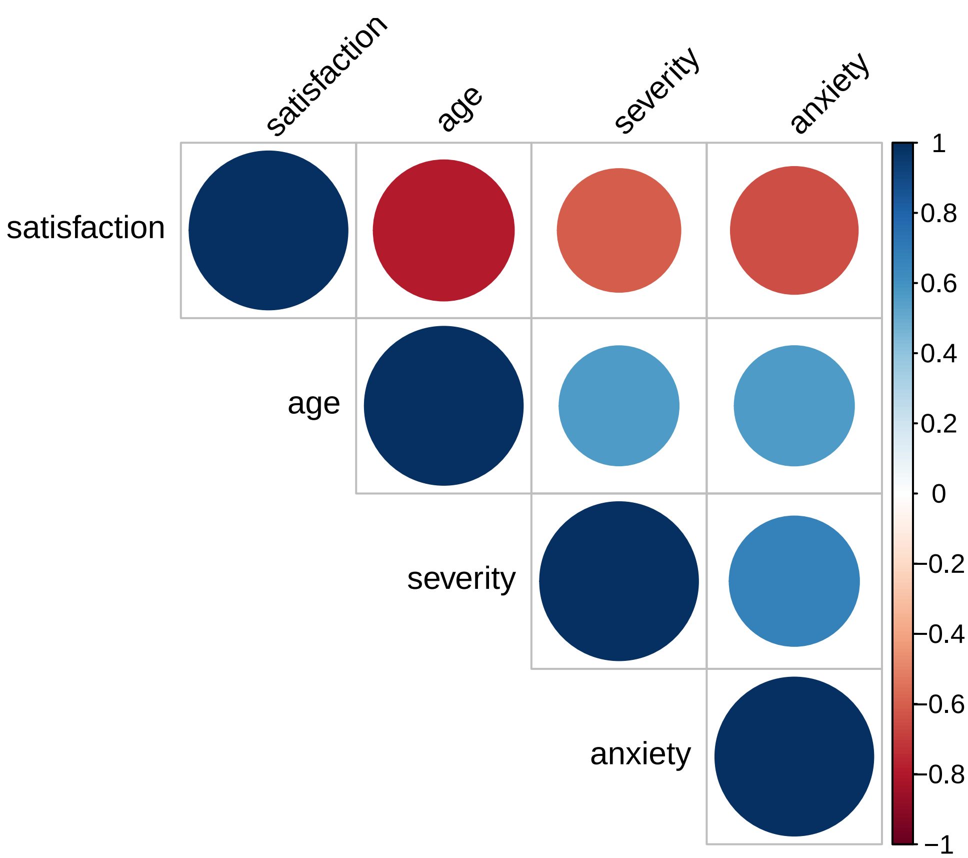
Renjie Wei

11/30/2020

Problem 1

1. Create a correlation matrix

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Safisfaction = col_double(),
##   Age = col_double(),
##   Severity = col_double(),
##   Anxiety = col_double()
## )
```



It seems that each predictor has a negative correlation with the outcome of interest.

2. Fit a multiple regression model.

Let's set up the test hypotheses:

$H_0 \ : \ \beta_i = 0, \ i = 1, 2, 3$

$H_1 : at \ least \ one \ of \ the \ coefficient \ is \ not \ equal \ to \ 0$

And our model to be test is:

$$Model_{test} : \ Statisfaction = 158.49 - 1.14age - 0.44severity - 13.47anxiety$$

We can test this model against the model only with intercept:

$$Model_{null} : \ Statisfaction = 61.57$$

And we do the ANOVA test:

The F statistic is calculated by:

$$F_{stat} = \frac{MSR}{MSE} \sim F_{p, \ n-p-1}, \ where \ p = 3, \ n = 46$$

```
## Analysis of Variance Table
##
## Model 1: satisfaction ~ 1
## Model 2: satisfaction ~ age + severity + anxiety
##   Res.Df   RSS Df Sum of Sq    F  Pr(>F)
## 1     45 13369
## 2     42  4249  3      9120 30.1 1.5e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Decision Rules: $Reject \ H_0 \ if \ F_{stat} > F_{1-\alpha; \ p, \ n-p-1}$

In our situation, $F_{stat} = 30.1 > F_{0.95; \ 3, \ 42} = 2.827$

So, we reject the null hypothesis with 95% confidence and conclude that at least one coefficient is not equal to zero.

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

term	estimate	2.5 %	97.5 %
(Intercept)	158.491	121.91	195.071
age	-1.142	-1.57	-0.708
severity	-0.442	-1.44	0.551
anxiety	-13.470	-27.80	0.858

Interpretation:

The 95% CI for severity of illness is $(-1.435, \ 0.551)$, that means at 95% confidence level, the mean change of patient satisfaction given all the same except for severity of illness per unit is between $(-1.435, \ 0.551)$, noticing that 0 is in this interval.

4. Obtain an interval estimate for a new patient's satisfaction with Age=35, Severity=42, and Anxiety=2.1.

The 95% PI for this new patient is :

lwr	upr
50.1	93.3

Interpretation:

This means at 95% confidence level, the true estimate of patient satisfaction is between $(50.062, 93.304)$.

5.

a) Test whether 'anxiety level' can be dropped from the regression model, given the other two covariates are retained.

First, we set up the hypotheses:

$H_0 \ : \ \beta_{anxiety} = 0$

$H_1 \ : \ \beta_{anxiety} \neq 0$

And we conduct the ANOVA test:

$$F_{stat} = \frac{MSR(X3|X1X2)}{MSE(X1X2X3)} \sim F_{df_L - df_S, df_L}, \ where \ df_L = 43, \ df_S = 42$$

```
## Analysis of Variance Table
##
## Model 1: satisfaction ~ (age + severity + anxiety) - anxiety
## Model 2: satisfaction ~ age + severity + anxiety
##   Res.Df  RSS Df Sum of Sq   F Pr(>F)
## 1     43 4613
## 2     42 4249  1       364 3.6  0.065 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Decision Rules: $Reject \ H_0 \ if \ F_{stat} > F_{1-\alpha; df_L - df_S, df_L}$

In our situation, $F_{stat} = 3.6 < F_{0.95; \ 1, \ 42} = 4.073$

So, we cannot reject the null hypothesis with 95% confidence and conclude that 'anxiety level' can be dropped from the regression model.

b) How are R2/R2-adjusted impacted

Model	R_square	Adjusted_R_square
With Anxiety	0.682	0.659
Without Anxiety	0.655	0.639

We can see that the R square and Adjusted R square are higher in model with anxiety level.

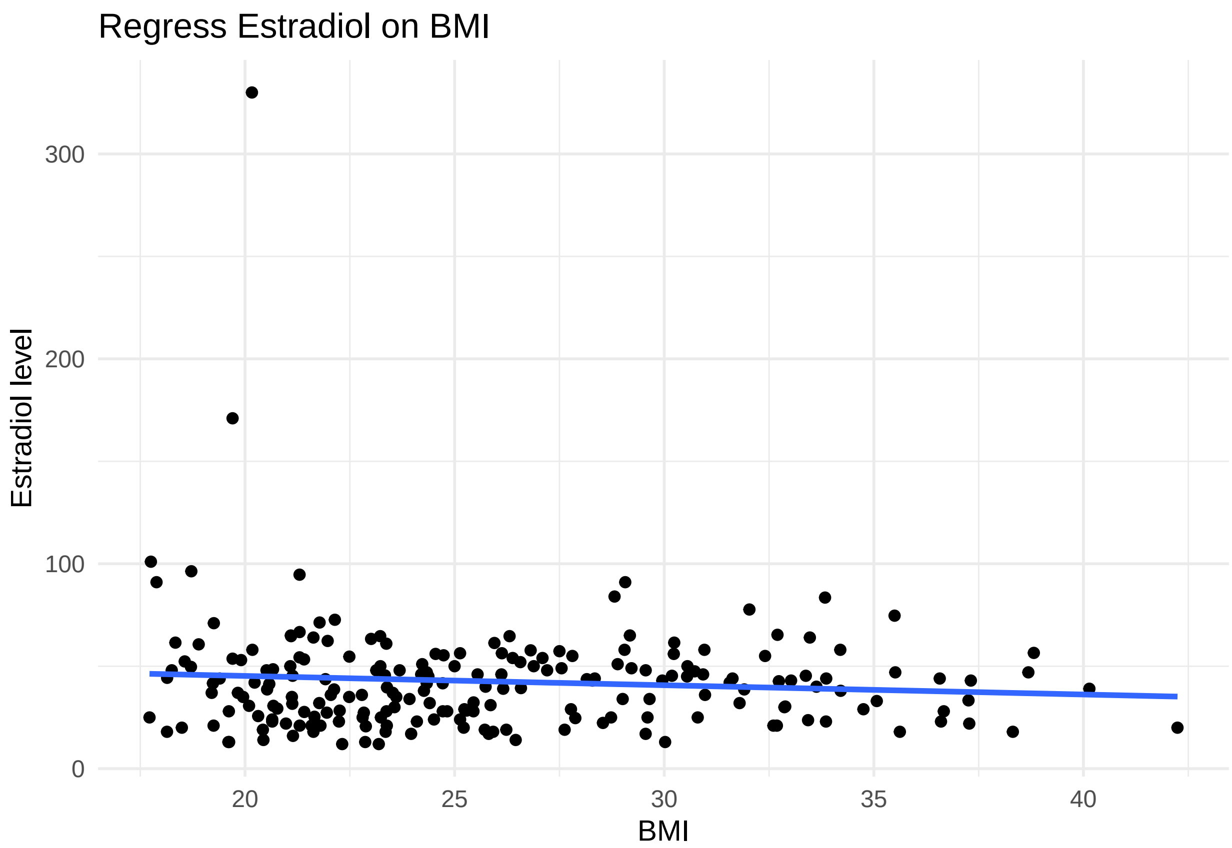
Problem 2

1. Crude association between BMI and serum estradiol

a) Scatter plot with regression line.

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Id = col_double(),
##   Estradl = col_double(),
##   Ethnic = col_double(),
##   Entage = col_double(),
##   Numchild = col_double(),
##   Agemenar = col_double(),
##   BMI = col_double()
## )

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



Comments:

As the plot shown above, the regression line has a very small slope, and the slope is negative. And there are some outliers when BMI is low.

b) Regression output

term	estimate	std.error	statistic	p.value
(Intercept)	54.310	9.51	5.71	0.00
bmi	-0.453	0.36	-1.26	0.21

Comments:

The coefficient of BMI is -0.453, and the p-value is 0.21. There is little evidence showing a strong relation ship between BMI and Estradiol level.

2. Relationship between BMI and serum estradiol change after controlling for all the other risk factors

term	estimate	std.error	statistic	p.value
(Intercept)	26.157	13.072	2.001	0.047
ethnicCaucasian	16.058	4.449	3.609	0.000
entage	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
bmi	-0.107	0.370	-0.288	0.774

Comments:

The coefficient of BMI after controlling for all the other risk factors changed from -0.453 to -0.107, and the p-value changed from 0.21 to 0.774 The relationship between BMI and Estradiol level seems to be more insignificant after controlling.

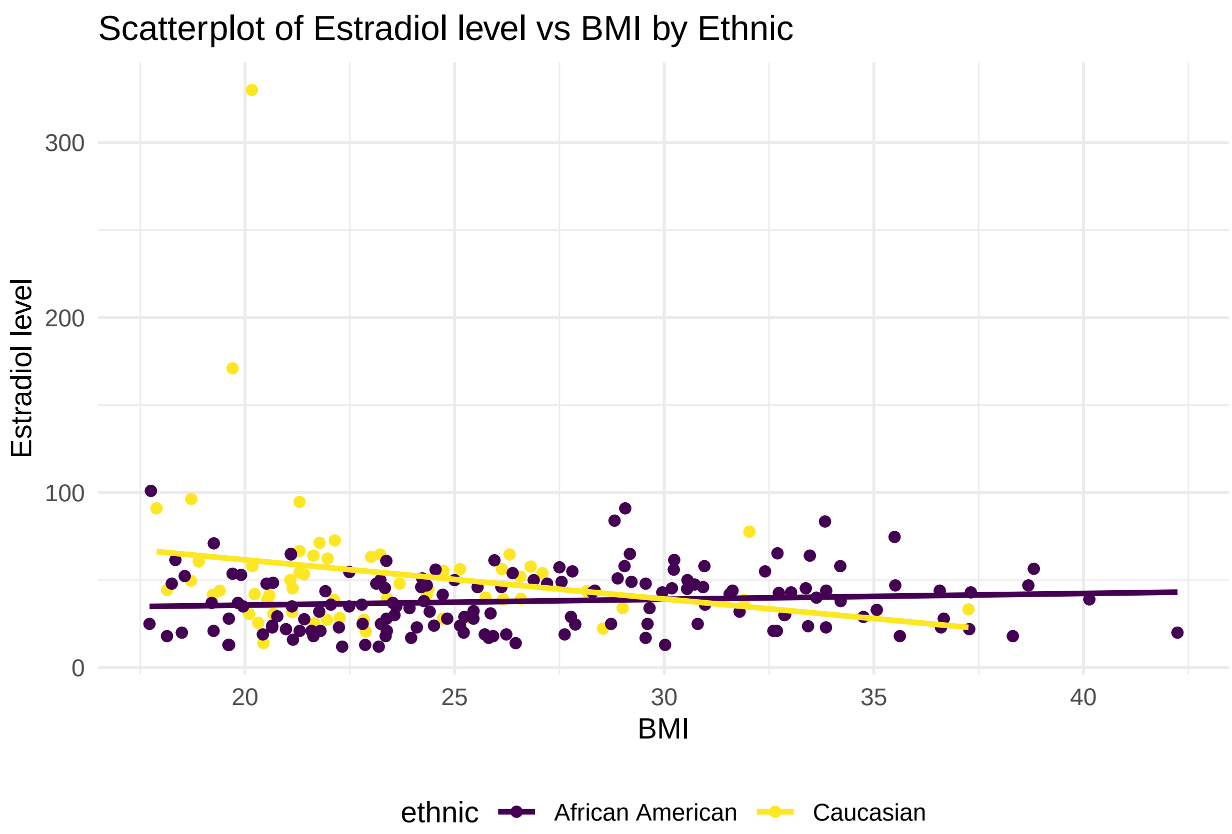
The p-value of `entage`, `nunchild`,`agemenar` and `bmi` are relatively high and there coefficient is small in magnitude, which implies there might not be a strong relationship between estradiol level. However, the p-value of `ethnicCaucasian` is small, and its coefficient is large in magnitude, there might be a relationship between ethnic and estradiol level.

3. Focus on BMI, ethnic and estradiol level

a) Graphical displays and numerical summaries

First I will draw a scatter plot about Estradiol level vs BMI by Ethnic:

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



From the plots, we can see a cross over two regression line, that is a indication of interactions between `BMI` and `ethnic`.

Let's build a model with this interaction and see if it is significant.

```
##
## Call:
## lm(formula = estradl ~ bmi * ethnic, data = estradl_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)            29.075     10.757    2.70   0.0074 **
## bmi                     0.333      0.392    0.85   0.3976
## ethnicCaucasian        77.210     24.784    3.12   0.0021 **
## bmi:ethnicCaucasian    -2.568      1.029   -2.50   0.0133 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27 on 206 degrees of freedom
## Multiple R-squared:  0.0963, Adjusted R-squared:  0.0831
## F-statistic: 7.32 on 3 and 206 DF,  p-value: 0.00011
```

From the summary of the model, we can see that the interaction is significant. With p-value 0.0133.

b) Additional steps

Since there is a significant interaction, we need to do stratified analysis.

```r
caucasian_df =
  estradl_df %>%
  filter(ethnic == "Caucasian")

aamerican_df =
  estradl_df %>%
  filter(ethnic == "African American")

strat_reg_cau = lm(estradl ~ bmi, data = caucasian_df)
strat_reg_aam = lm(estradl ~ bmi, data = aamerican_df)
```

So in Caucasian, a negative, relatively large in magnitude association b/w BMI and estradiol level

```
##
## Call:
## lm(formula = estradl ~ bmi, data = caucasian_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -46.60 -20.79  -6.80   8.14 268.79
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   106.29      35.71    2.98   0.0043 **
## bmi            -2.24       1.52   -1.47   0.1470
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.2 on 57 degrees of freedom
## Multiple R-squared:  0.0365, Adjusted R-squared:  0.0196
## F-statistic: 2.16 on 1 and 57 DF,  p-value: 0.147
```

And in African American, a positive association b/w BMI and estradiol level.

```
##
## Call:
## lm(formula = estradl ~ bmi, data = aamerican_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -26.1  -14.0   -1.1   11.0   66.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   29.075      6.839    4.25  3.7e-05 ***
## bmi            0.333      0.250    1.33     0.18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 17.2 on 149 degrees of freedom
## Multiple R-squared:  0.0118, Adjusted R-squared:  0.00516
## F-statistic: 1.78 on 1 and 149 DF,  p-value: 0.184
```

However, both associations are not statistically significant.

And to see if Ethnic is a confounder of BMI, it need to meet 3 conditions

Condition 1) Associated with the outcome:

```
##
## Call:
## lm(formula = estradl ~ ethnic, data = .)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -40.5  -15.1   -3.5   10.0  275.6
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)        38.00       2.22   17.10  < 2e-16 ***
## ethnicCaucasian    16.45       4.19    3.92  0.00012 ***
## ---
## 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:  0.0689, Adjusted R-squared:  0.0644
## F-statistic: 15.4 on 1 and 208 DF,  p-value: 0.000119
```

Yes, ethnic is associated with estradiol level with a very small p-value.

Condition 2) Associated with the exposure:

```
##
## Call:
## lm(formula = bmi ~ ethnic, data = .)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9.11  -3.60  -1.06   3.39  15.41
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       26.832      0.420   63.82  < 2e-16 ***
## ethnicCaucasian   -3.641      0.793   -4.59  7.7e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.17 on 208 degrees of freedom
## Multiple R-squared:  0.092,  Adjusted R-squared:  0.0876
## F-statistic: 21.1 on 1 and 208 DF,  p-value: 7.65e-06
```

Yes, ethnic is associated with BMI with a extreme small p-value.

Condition 3) Not on the causal pathway b/w exposure and outcome, which is obvious.

term	estimate	std.error	statistic	p.value
(Intercept)	54.310	9.51	5.71	0.00
bmi	-0.453	0.36	-1.26	0.21

term	estimate	std.error	statistic	p.value
(Intercept)	39.105	10.104	3.870	0.000
bmi	-0.041	0.367	-0.112	0.911
ethnicCaucasian	16.297	4.410	3.696	0.000

And by comparing model controlling `ethnic` and not, we can see that the BMI coefficient reduced from -0.453 in SLR to -0.041 in MLR after adjusting for `ethnic` (~90% reduction). We can conclude that `ethnic` confounds the relationship b/w `BMI` and `estradiol level`.