# Homework 6

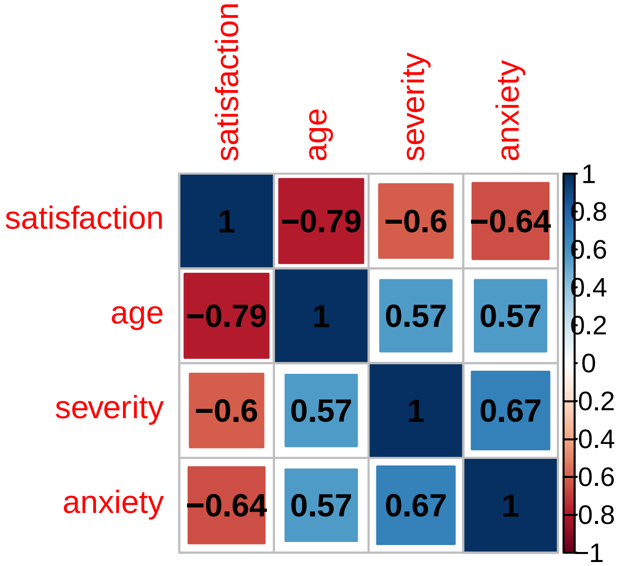
Jeffrey Liang

11/27/2020

## Problem 1(15p)

A hospital administrator wishes to test the relationship between 'patient's satisfactionscore' (Y) and three potential predictors:'age', 'severity of illness', and 'anxiety level' (see dataset'PatSatisfaction.csv'). The administrator randomly selected 46 patients, collected the data, and asked for your help with the analysis.

1. Create a correlation matrix for all variables and interpret your findings. Focus on the correlation values between each predictor and the out come of interest. (2p)



Satisfaction is negative correlated with all other variable, and other variables are postively correlated with each others. The correlation of anxiety and severity is 67%, there's may be collinearity if introducing both variable into model.

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. (3p)

To test if at least one of these model is significant, we propose hypothesis:

$H_0 \ : \ \beta_1 = \beta_2 = \beta_3 = 0$

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

The model we fit is:

$$statisfaction = 158.49 - 1.14age - 0.44severity - 13.47anxiety$$

We campared with the model with only intercept, and with ANOVA, we have

$$F^* = \frac{MSR(0|X1X2X3)}{MSE(X1X2X3)} \sim F_{df_L - df_S, df_L}$$

$Reject\ H_0\ if\ F > F_{df_L - df_S, df_L, 1-\alpha}$

$Fail\ reject\ H_0\ if\ F < F_{df_L - df_S, df_L, 1-\alpha}$

```
## 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 1.5e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With F statistics of 30, at 95% confidence level, we have critical value of 2.827, we reject the null hypothesis and conclude that at least one predictor have coefficent not equal to 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'. (5p)

| term | conf.low | conf.high |
|---|---:|---:|
| (Intercept) | 121.91 | 195.071 |
| age | -1.57 | -0.708 |
| severity | -1.44 | 0.551 |
| anxiety | -27.80 | 0.858 |

The CI for coefficient of disease severity is ( -1.43, 0.55), we interpret as: at 95% confidence level, the mean change of satisfaction give all the same except for disease severity per unit is somewhere within this interval, and 0 is included,adjusted for all other predictors.

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. (2p)

The 95% confidence prediciton interval for Age=35, Severity=42, Anxiety=2.1 is (50.06, 93.3) we interpret as: at 95% confidence level the true estimate of satisfaction for Age=35, Severity=42, Anxiety=2.1 is somewhere within this interval.

5. a) 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. (1.5p)

We campared the full model with model without anxiety, and with ANOVA, we have

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

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

$$F^* = \frac{MSR(X3|X1X2)}{MSE(X1X2X3)} \sim F_{df_L - df_S, df_L}$$

$Reject\ H_0\ if\ F > F_{df_L - df_S, df_L, 1-\alpha}$

$Fail\ reject\ H_0\ if\ F < F_{df_L - df_S, df_L, 1-\alpha}$

```
## 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
```

At 95% confidence level, with F statistics of 3,6 less than 4.073, we can't reject the null hypothesis and conclude that the model include *anxiety* is not superior or different from model without *anxiety*.

`b) How areR2/R2-adjusted impacted by the action that youtook in part 5-a)? (1.5p)`

| model | r_square | adj_r_square |
|---|---:|---:|
| full model | 0.682 | 0.659 |
| model withour anxiety | 0.655 | 0.639 |

We see that the Coefficient of Determine decrease from the full model to the smaller model, because of the sum of square error increaseing as we take out one predictor. But the difference is less than 6%. As much as $R^2$, the Adjusted $R^2$ also decreases and less than 6% difference.

# Problem 2(15p)

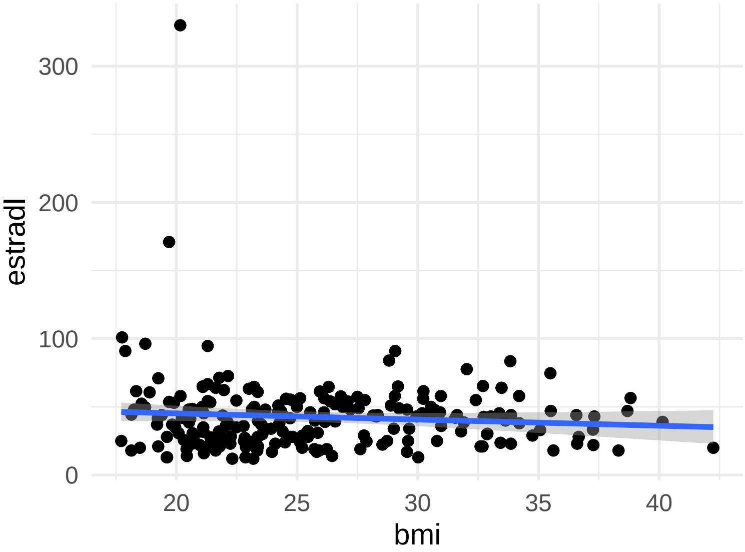
Obesity is very common in American society and is a risk factor for breast cancer in postmenopausal women. One mechanism explaining why obesity is a risk factor is that it may raise estrogen levels in women. In particular, one biomarker of estrogen, serum estradiol, is a strong risk factor for breast cancer. To better assess these relationships, researchers studied a group of 210 premenopausal women and recorded the following information('Estradl.csv'):

- Estradiol hormonal serum levels(Estradl);
- BMI = weight (kg)/height2(m2); measure of overall adiposity(used to indicate obesity, e.g., BMI>30; note that for this analysis we will use the continuous measurements);
- Ethnicity (Ethnic= 1 if African American, = 0 if Caucasian)
- Age (Entage);
- Number of children (Numchild);
- Age at menarche (Agemenar= age when menstrual periods began).

1. Is there a crude association between BMI and serum estradiol?

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

Looking at the fig below, we saw that the regression line is a line almost parallel to the x-axis, and there's some potential outliers.



`b)    Provide the summary regression output and comment on`

`the nature of the relationship (i.e., sign, magnitude, significance).(2.5p)`

| 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 |

As shown, the coefficient of BMI is -0.453, with negative sign but also with a p.value of 0.21.

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)

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 42.215 | 12.512 | 3.374 | 0.001 |
| ethnicAfrican American | -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 |

[1] "After controlling other factors, the mean change per unit of entage is 0.518 with p.value of 0.1503; After controlling other factors, the mean change per unit of numchild is -0.4906 with p.value of 0.6938; After controlling other factors, the mean change per unit of agemenar is 0.1073 with p.value of 0.5264; After controlling other factors, the mean change per unit of bmi is -0.1066 with p.value of 0.7737"

After adjusting all other factors, African American group have a nagative coefficient of -16.058 compared to the Caucasian groups, with a p.value of $3.865 \times 10^{-4}$

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.(2.5p)

On the first look of fig, we see that the regression line of estradol and bmi different trend by ethnic groups. We hypothesis that there's difference in relationship between BMI and serum estradiol by ethnicity, that is ethinicity is a confounder.

Then we test this hypothesis with ANOVA.

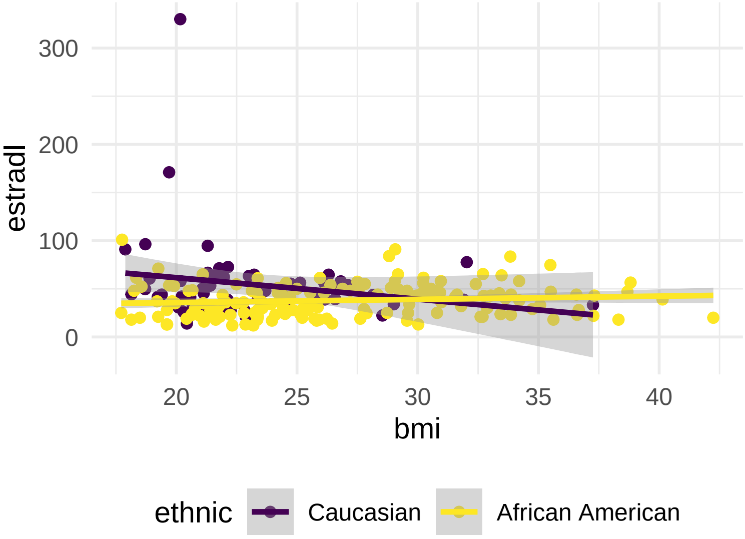
$H_0 \ : \ \beta_{ethnic} = 0$

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

$$F^* = \frac{MSR(X2|X1)}{MSE(X1X2)} \sim F_{df_L - df_S, df_L}$$

$Reject\ H_0\ if\ F > F_{df_L - df_S, df_L, 1-\alpha}$

$Fail\ reject\ H_0\ if\ F < F_{df_L - df_S, df_L, 1-\alpha}$



```
## Analysis of Variance Table
##
## Model 1: estradl ~ bmi
## Model 2: estradl ~ bmi + ethnic
##   Res.Df    RSS Df Sum of Sq    F  Pr(>F)
## 1    208 165281
## 2    207 155050  1     10231 13.7 0.00028 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

At 95% confidence level, because the F statistics of 13.658 is greater than the critical value of 3.887, we reject the null hypothesis and conclude that the coefficient of ethnic is a confounder.

```
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.(2.5p)
```

Also from the fig, we saw that the regression line is crossing at some point in the range of BMI, we hypothesis that the BMI might have interaction with ethnic groups.

| term | estimate | std.error | statistic | p.value |
|---|---:|---:|---:|---:|
| (Intercept) | 106.28 | 22.328 | 4.76 | 0.000 |
| bmi | -2.23 | 0.951 | -2.35 | 0.020 |
| ethnicAfrican American | -77.21 | 24.784 | -3.12 | 0.002 |
| bmi:ethnicAfrican American | 2.57 | 1.029 | 2.50 | 0.013 |

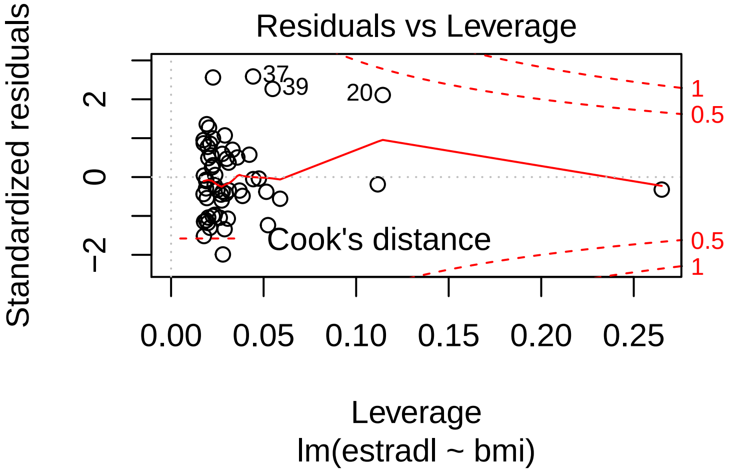
The interaction term in the model is statistically significant, so we need to include the interaction into the model, to interpret the true asssociation of BMI and estradol by ethnic group, we proform stratified analysis on ethnic groups.

| ethnic | term | estimate | p_value | conf_low | conf_high |
|---|---|---:|---:|---:|---:|
| Caucasian | (Intercept) | 106.285 | 0.004 | 34.78 | 177.785 |
| Caucasian | bmi | -2.235 | 0.147 | -5.28 | 0.809 |
| African American | (Intercept) | 29.075 | 0.000 | 15.56 | 42.589 |
| African American | bmi | 0.333 | 0.184 | -0.16 | 0.826 |

[1] "For Caucasian group, the mean of change in estradol per unit bmi is -2.235 with a p.value of 0.147; For African American group, the mean of change in estradol per unit bmi is 0.333 with a p.value of 0.184"

As we saw in problem 1, there's outlier in the observation, Using the Cook-leverge plot, we saw one observation with high leverge, and 3 observations close the 0.5 Cook's distance. After taking out these outliers,

```
## Model filterred possible outlier
```



| term | estimate | std.error | statistic | p.value |
|---|---:|---:|---:|---:|
| (Intercept) | 63.021 | 15.217 | 4.14 | 0.000 |
| bmi | -0.663 | 0.645 | -1.03 | 0.308 |

[1] "For Caucasian group, the mean of change in estradol per unit bmi is -0.663 with a p.value of 0.3083"