p8130 HW6

JingYao Geng

2020-12-03

Problem 1

1.1 Correlation matrix for all variables:

	safisfaction	age	severity	anxiety
safisfaction	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

- Based on the correlation matrix for all variables above, we find that patients' satifaction score `safisfaction` are negative correlated with the three potential predictors: `age`, `severity of illness` and `anxiety level`. A `negative correlation` is a relationship between two variables in which an increase in one variable is associated with a decrease in the other. Moreover, there is a positive correlation between each pair of the predictors. All values are symmetric about the diagonal line, this is true because corr(X,Y)=corr(Y,X). All values in the diagonal are 1, this is true because corr(X, X) = 1.

1.2 Multiple Regression Model:

$$Satisfaction_i = \beta_0 + \beta_{age}Age + \beta_{severity}Severity + \beta_{anxiety}Anxiety + \varepsilon_i$$

```
##
## Call:
## lm(formula = safisfaction ~ age + severity + anxiety, data = hospital_df)
##
## 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
```

Hypothesis:

$H_0 : \beta_{age} = \beta_{severity} = \beta_{anxiety} = 0$ $H_1 : at\ least\ one\ \beta\ is\ not\ 0.$

Test Statistics: $F = \frac{MSR}{MSE} = \frac{SSR/p}{SSE/(n-p-1)} = \frac{SSR/3}{SSE/42} = 30.05 \sim F_{3,42}\ under\ null$

Test Critical

$F_{\ 1-\alpha,\ p,\ n-p-1} = F_{\ 0.95,\ 3,\ 42} = 2.8270487$

Decision Rule: $Reject\ H_0 : F > FF_{\ 0.95;\ 3,\ 42}$ $Fail\ to\ Reject\ H_0 : F \leq F_{\ 0.95,\ 3,\ 42}$

Conclusion: Based on both summary table above, we find: $F = 30.05 > F_{\ 1-0.05;\ 3,\ 42} = 2.827049$, and we reject H_0 and conclude that at least one slope/coefficient is not zero.

1.3 95% C.I.

	2.5 %	97.5 %
(Intercept)	121.911727	195.0707761
age	-1.575093	-0.7081303
severity	-1.434831	0.5508228
anxiety	-27.797859	0.8575324

- We are 95% confident that for every one unit increase of severity of illness, the estimated patient's satisfaction scores on average would change between -1.4348 and 0.5508.

1.4 Interval estimate for a new patient's satisfaction with `age=35`, `severity=42`, `anxiety=2.1`

fit	lwr	upr
71.68332	64.23592	79.13071

- We are 95% that the estimated mean patients' satisfaction score for a new patient with age = 35, severity = 42, and anxiety = 2.1 is between 64.23592 and 79.13071.

1.5 Test whether `anxiety level` can be dropped from the regression model, given the other two covariates are retained:

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
43	4613.000	NA	NA	NA	NA
42	4248.841	1	364.1595	3.599735	0.0646781

We use `Partial F-test` for nested models:

Small Model: $Satisfaction_i = \beta_0 + \beta_{age}Age + \beta_{severity}Severity + \varepsilon_i$

Large Model: $Satisfaction_i = \beta_0 + \beta_{age}Age + \beta_{severity}Severity + \beta_{anxiety}Anxiety + \varepsilon_i$

$H_0 : small\ model$ $H_1 : large\ model$

test statistics:

$F = \frac{(SSE_S - SSE_L)/(df_S - f_L)}{SSE_L/df_L} \sim F_{df_l - df_s, df_l}$, where $df_s = n - p_s - 1, df_l = n - p_l - 1$

Small: $Y_i = \beta_0 + \beta_{anxiety} + \varepsilon_i$ large: $Y_i = \beta_0 + \beta_{age} + \beta_{severity} + \beta_{anxiety} + \varepsilon_i$

We are testing that:

$H_0 : \beta_{anxiety} = 0$ $H_1 : \beta_{anxiety} \neq 0$

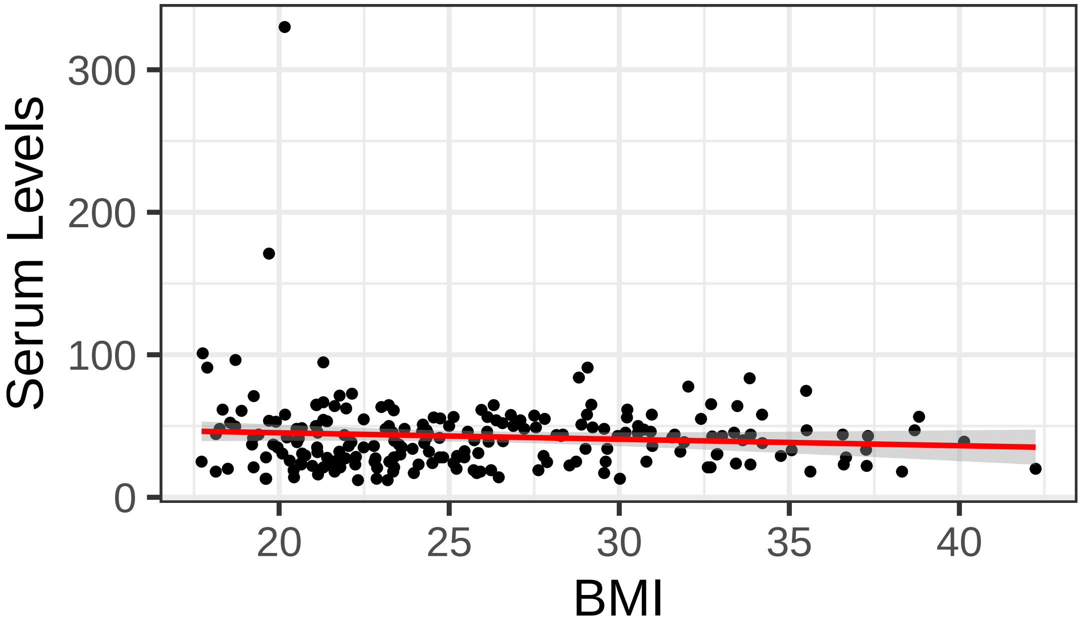
Conclusion:

- F = 3.5997, p-value = 0.06468 > $\alpha = 0.05$, we fail to reject H_0 and conclude that the larger model is not superior. And we conclude that the model with predictors 'age', 'severity', 'anxiety' is actually not providing more information than the model that only containing 'age' and 'severity'. Therefore, based on alpha = 0.05, we think it's better to DROP `anxiety`, given the other two variables retained.
- When we take a look at the R^2 and adjusted R^2 for both models. We find that large model (same as part a) has a R^2 of 0.6822 and a adjusted R^2 of 0.6595. The small model (without 'anxiety') has a R^2 of 0.655 and a adjusted R^2 of 0.6389. Both R^2 and adjusted R^2 are slightly bigger in the large model, however the increases are not meaningful (not >5%). Therefore, we would agree with the previous conclusion and conclude it's better to **DROP** the variable `anxiety`.

Problem 2

2.1 Is there a crude association between BMI and serum estradiol?

2.1.(a) Scatter plot:



- Based on the scatter plot above, we find there is not a strong positive or negative association between bmi and serum estradiol. The fitted regression line is slightly decreasing, but it's quite parallel to the x-axis. Also, there are 2 potential outliers in the plot.

2.1.(b)

```
##
## Call:
## lm(formula = estradl ~ bmi, data = obs_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
```

$Y_i = \beta_0 + \beta_1 BMI + \varepsilon_i$

- Based on the summary regression output, we find a weak negative (-0.4529) association between bmi and serum estradiol. For one unit increase in bmi, the estimated estradl level will decrease by 0.4529 unit on average. However, the p-value of 0.21 is high compared with $\alpha = 0.05$. This might cause us to conclude that bmi is not a significant predictor of the estradiol hormonal serum levels. And both R^2 and adjusted R^2 are really small. It indicates that this model does not fit well.

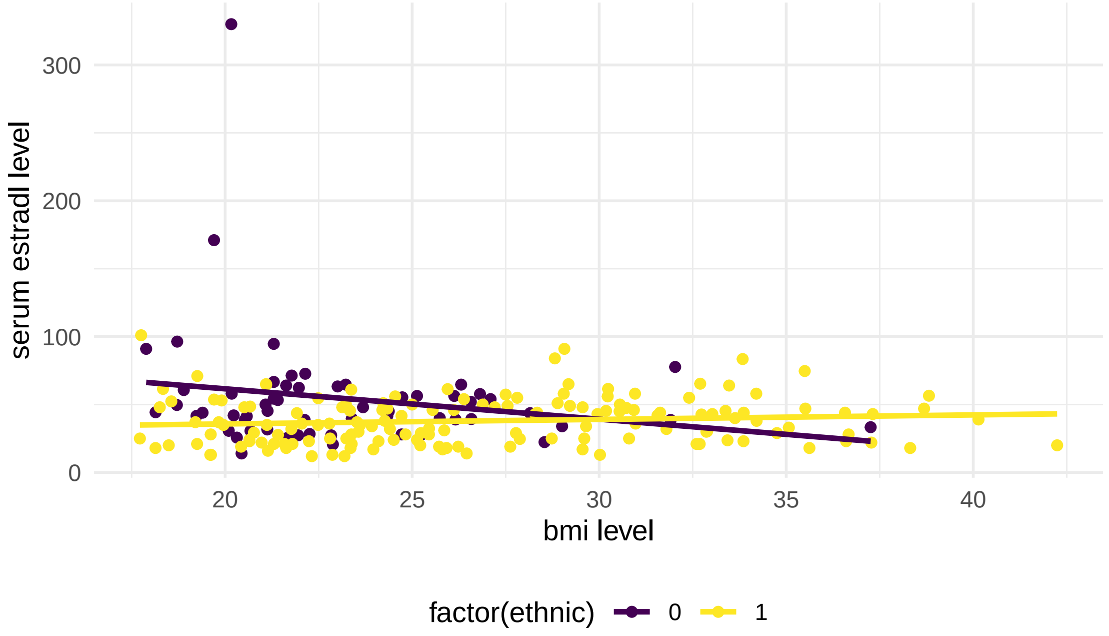
2.2 How does the relationship between BMI and serum serum estradiol change after controlling for all the other risk factors listed above?

```
##
## Call:
## lm(formula = estradl ~ bmi + ethnic + entage + numchild + agemenar,
##     data = obs_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 ***
## entage        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
```

- After adjusting for other variables, we still observe a negative correlation between bmi and estradl but weaker. And we conclude that for one unit increase in bmi level, the estimated serum estradiol will decrease by 0.1066 on average, adjusting/ controlling for 'ethnic', 'entage', 'numchild', and 'agemenar'. At significance level of 0.05, we conclude that bmi is not significant predictor for serum estradiol.
- ethnic is negative correlated with serum estradiol. However, at significance level of 0.05, a p-value of 0.000386 indicates that ethnic is a significant predictor for serum estradiol.
- entage is positive correlated with serum estradiol. However, at significance level of 0.05, a p-value of 0.150259 indicates that entage is not a significant predictor for serum estradiol.
- numchild is negative correlated with serum estradiol. However, at significance level of 0.05, a p-value of 0.693788 indicates that numchild is not a significant predictor for serum estradiol.
- agemenar is positive correlated with serum estradiol. However, at significance level of 0.05, a p-value of 0.526429 indicates that agemenar is not a significant predictor for serum estradiol.

2.3 Is there any evidence that these relationships vary for African American and Caucasian women?

Scatter Plot:



- Based on the scatter plot, we find for African American women (ethnic = 0), the bmi level is negative related to the serum estradl level; but for Caucasian women (ethnic = 1), the bmi level is slightly positive related to the serum estradl level. Therefore, we conclude that there is some evidence that the relationship between 'bmi' level and 'serum estradl' level varied by 'ethnic'. There are some sort of interactions.

Numerical summary:

- African American Women:

```
##
## Call:
## lm(formula = estradl ~ bmi, data = a)
##
```

```
## 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
```

- Caucasian Women:

```
##
## Call:
## lm(formula = estradl ~ bmi, data = b)
##
## 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 stratify the dataset by 'ethnic', and fit a regression model with the same predictor 'bmi' and response 'estradl'. We find the coefficients are different. For model with African American women, the summary shows a positive slope of 0.3327 and p-value of 0.184. For Caucasian women, the summary shows a negative slope of -2.235 and p-value of 0.14702. This further strengthens the findings we observed from the plot.

Based on your findings in 2.3.a , take additional steps to quantify the relationship between BMI and serum estradl by ethnicity.

without ethnic:

```
##
## Call:
## lm(formula = estradl ~ bmi, data = obs_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
```

with ethnic:

```
##
## Call:
## lm(formula = estradl ~ bmi + ethnic, data = obs_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.561 -15.239  -3.585   9.719 275.428
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  55.40203    9.23344   6.000 8.69e-09 ***
## bmi          -0.04115    0.36731  -0.112 0.910899
## ethnic      -16.29666    4.40960  -3.696 0.000281 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.37 on 207 degrees of freedom
## Multiple R-squared:  0.06896,    Adjusted R-squared:  0.05997
## F-statistic: 7.666 on 2 and 207 DF,  p-value: 0.000614
```

- We think `ethnic` might be a potential confounder. Then we fit 2 regression models for `bmi` and `estradl`: one contains `ethnic` and the other doesn't contain `ethnic`. We examine the coefficient of `bmi` for both models, and we find there is a bid difference in magnitude (from -0.4529 to -0.04115). Therefore, we conclude: `ethnic` is confounding the relationship between bmi and estradl.

Moreover: Interaction

term	estimate	std.error	statistic	p.value
(Intercept)	106.285012	22.3275525	4.760262	0.0000036
bmi	-2.235219	0.9507284	-2.351060	0.0196646
ethnic	-77.210424	24.7838258	-3.115355	0.0020990
bmi:ethnic	2.567900	1.0285388	2.496648	0.0133211

- We take look at the interaction term, and we find a significant p-value of 0.013 (compared to 0.05) of the for `bmi:ethnicCaucasian`. This might indicate that including `ethic` the out model is somehow necessary.

ethnic	term	estimate	p_value	conf_low	conf_high
0	(Intercept)	106.2850120	0.0042711	34.7849286	177.7850953
0	bmi	-2.2352193	0.1470183	-5.2797608	0.8093221
1	(Intercept)	29.0745881	0.0000374	15.5602335	42.5889427
1	bmi	0.3326803	0.1844442	-0.1603404	0.8257009

- We find that for Caucasian group, the mean of change in estradol per unit bmi is -2.235 with a p-value of 0.147, and for African American group, the mean of change in estradol per unit bmi is 0.333 with a p-value of 0.184.