EDMS 646: Homework 3

Minoo Ahmadi

March 30, 2017

PART 1: Multiple Regression - Initial model: Use data set HSB1

1.

$$\hat{Y} = 50.3514 + 4.7188(\text{locus}) + 0.2549(\text{concept}) + 1.7113(\text{mot})$$

Controlling for self-concept and motivation, one unit increase in locus of control will increase the science score by 4.7188 units. Controlling for locus of control and motivation, one unit increase in self-concept will increase the science score by 0.2549 units. Controlling for locus of control and self-concept, one unit increase in motivation will increase the science score by 1.7113 units.

Residuals:				
Min	1Q	Median	3Q	Max
-26.6536	-6.1518	0.5052	7.0151	21.1206

Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	50.3514	1.0900	46.192	< 2e-16 ***
locus	4.7188	0.7776	6.068	3.39e-09 ***
concept	0.2549	0.7606	0.335	0.738
mot	1.7113	1.4978	1.143	0.254

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘’ 1

Residual standard error: 9.326 on 346 degrees of freedom. Multiple R-squared: 0.1139, Adjusted R-squared: 0.1062. F-statistic: 14.83 on 3 and 346 DF, p-value: 4.192e-09.

2.

H_0: $\beta_{locus} = \beta_{concept} = \beta_{mot} = 0$
H_1: $\beta_j \neq 0$

The null hypothesis states that the model has no predictive capability and all the regression coefficients equal to zero. In other words, none of the 3 predictors (locus of control, self-concept and motivation) can predict the outcome (science score).

The alternative hypothesis on the other hand, states that at least one of the predictors has the capability of predicting the outcome. In other words, at least one of the regression coefficients is significantly different from 0.

Looking at the results from the ANOVA table, our F-value exceeds the critical F-value and the p-value is less than .001. Therefore, at least one of our predictors has a significant regression coefficient and we reject the null hypothesis.

3. H_0: $\beta_j = 0$
   H_1: $\beta_j \neq 0$

As opposed to ANOVA, which is an omnibus test, regression table is reporting t-tests for each of the predictors.

The null hypothesis for each of the predictors is that this specific predictor (let it be locus of control, self-concept or motivation) has no predictive capability for the outcome (science score) when controlling for other predictors in the model. In other words, the true slope is 0.

The alternative hypothesis on the other hand states that, when the other predictors in the model are controlled for, this specific predictor predicts the outcome and its regression coefficient is significantly different from 0.

Based on the t-tests, it seems that locus of control is the only predictor in our model that can significantly predict the science score when accounting for self-concept and motivation. For locus of predictor, the t-value exceeds the critical t-value and our p-value is less than .001. Hence, we reject the null hypothesis.

The t-test is not significant for the concept and motivation and we cannot conclude that they have the capability of predicting science score.

Standard error of the regression coefficients show the variability of the sampling distribution for the slope for each of the predictors when accounting for the other predictors.

4. Normality: The histogram for standarized residulas seems left-skewed.

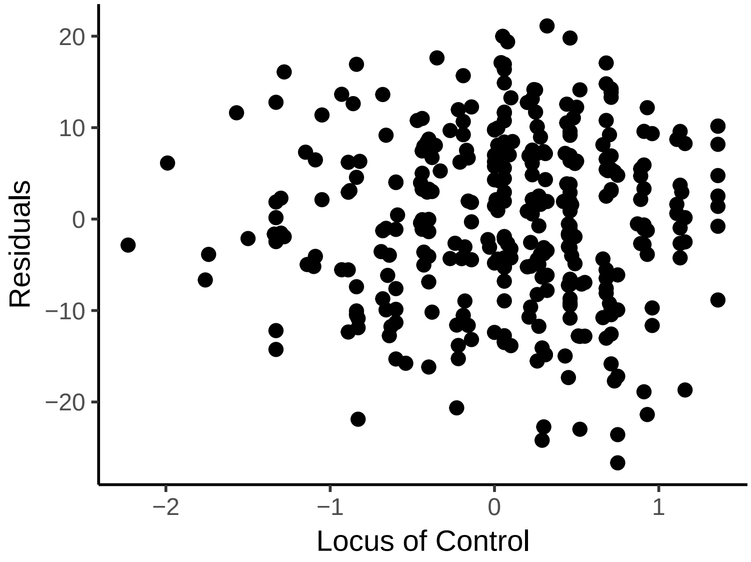
	Statistic	SE	t-val	p.
Skewness	-0.2597293	0.1309307	-1.983716	0.02364378
Kurtosis	-0.3921500	0.2618615	-1.497548	0.06712540

The K-S test returns significant results (p<.001), which is not much reliable since our sample size is >50 (we have 300 subjects). Yet QQ plot seems fairly normal. Overall, I'm suspicious and think that normality assumption has been violated.

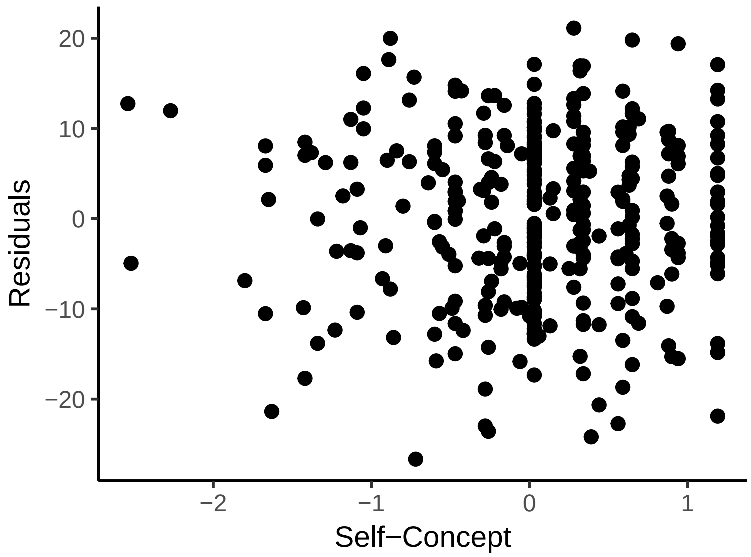
Linearity: The plot for residuals vs predicted Y shows fair linearity. The data points seem to be equally distributed above and below the y = 0 line. However, when we look at the residual vs. individual predictor plots, we can see grouping and dependence issues. So, linearity might have been violated.

Homoscedasticicy: Some level of fanning out is observed in the residulads vs predicted Y plot. It's even more visible in residulads vs Locus of control.

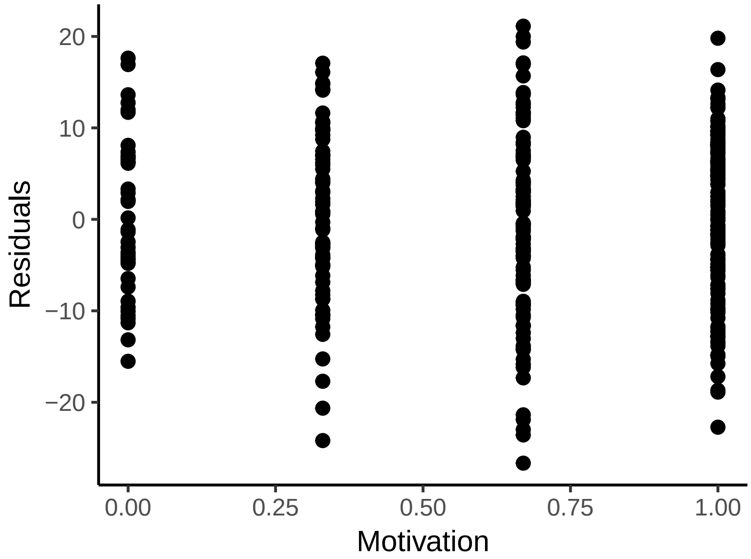
```r
#Plot: Residuals vs X
ggplot(hsb1.data, aes(locus, residuals(hsb1.lm)))+ geom_point(size = 2) + xlab("Locus of Control") + yla
   theme(
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        panel.border = element_blank(),
        axis.line = element_line(),
        axis.ticks = element_line(),
     )
```



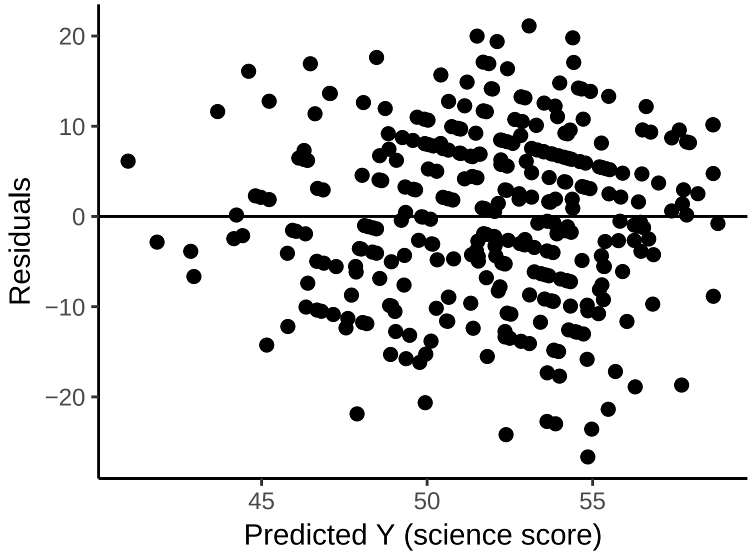
```
ggplot(hsb1.data, aes(concept, residuals(hsb1.lm)))+ geom_point(size = 2) + xlab("Self-Concept") + ylab
   theme(
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        panel.border = element_blank(),
        axis.line = element_line(),
        axis.ticks = element_line(),
     )
```



```
ggplot(hsb1.data, aes(mot, residuals(hsb1.lm)))+ geom_point(size = 2) + xlab("Motivation") + ylab("Resi
   theme(
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        panel.border = element_blank(),
        axis.line = element_line(),
        axis.ticks = element_line(),
     )
```

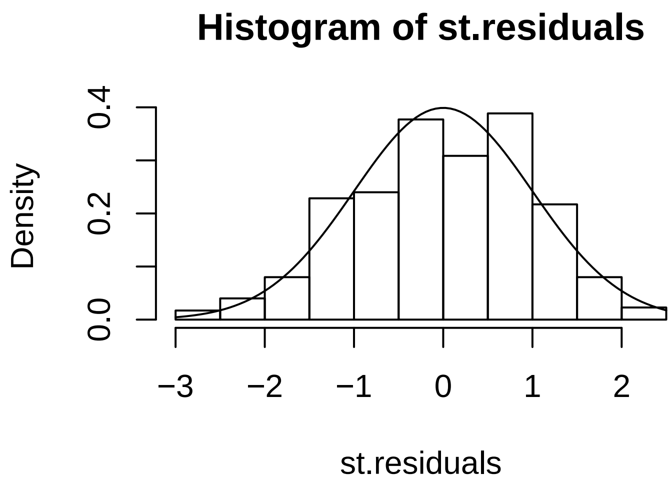


```
#Plot: Resisulas vs Predicted Y
ggplot(hsb1.data, aes(fitted.values(hsb1.lm), residuals(hsb1.lm)))+ geom_point(size = 2) + xlab("Predic
   theme(
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        panel.border = element_blank(),
        axis.line = element_line(),
        axis.ticks = element_line(),
     )
```

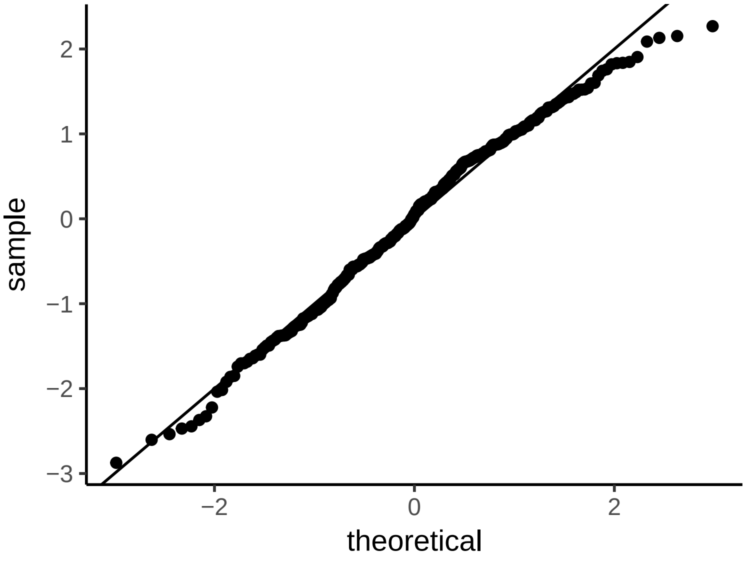


```
# getting stadardized residuals
st.residuals <- stdres(hsb1.lm)
# create residuals hist
```

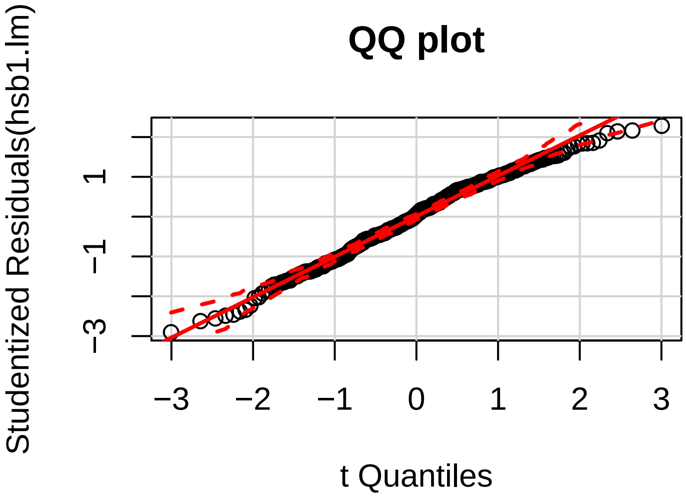
```r
hist(st.residuals, freq = FALSE)
curve(dnorm, add = TRUE)
```



```r
# create PP plot
ggplot(hsb1.data, aes(sample = st.residuals))+ stat_qq() + geom_abline(intercept = 0, slope = 1) +
   theme(
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        panel.border = element_blank(),
        axis.line = element_line(),
        axis.ticks = element_line(),
     )
```



```r
# qq plot for studentized residuals
qqPlot(hsb1.lm, main = "QQ plot")
```



Part 2: Multiple Regression - Final model: Use data set HSB1

1. I tried BoxCox and BoxTidwell to achieve linearity, also tried taking logs to correct for the skewness, but neither of these changed the t-value for the two non-significant predictors (concept and motivation) significantly and they still remained non-significant. For the other types of regression (WLS) I couldn't find helpful resources on the internet, mainly because they were too advanced for me.

$$\hat{Y} = 50.3514 + 4.7188(\text{locus}) + 0.2549(\text{concept}) + 1.7113(\text{mot})$$

2. This research was designed to determine the influence of locus of control, self-concept and motivation on science score. Students' science scores were regressed on their scores for locus of control, self-concept and motivation. The overall multiple regression was statistically significant ($R^2 = 0.1139, F(3,346) = 14.83, p < 0.001$). The three variables (locus of control, self-concept and motivation) accounted for 11% of the variance in science score. From among these three variable, only locus of attention had a statistically significant effect on the science score. The unstandardized regression coefficient (β) for locus of control was 4.7188 ($t(346) = 6.068, p < 0.001$), meaning that for each additional score on locus of control, students' science score increased by 4.7188 points, controlling for on self-concept and motivation. These results suggest that locus of control is indeed an important influence on students' science grade and that this effect holds even after students' self-concept and motivation are taken into account. Students who want to improve their grades in science may do so by increasing their locus of contorl.

Table 1: Summary of Regression Analysis for Variables Predicting Science Score ($N = 300$)

Variable	B	$SE\ B$	β
Locus	4.71**	0.77	0.31
Concept	0.25	0.76	0.01
Motivation	1.71	1.49	0.06

Part 3: ANOVA using Multiple Regression

1.

$$\hat{Y} = 97.8 + 7.7(\text{Tretment A}) + 2.7(\text{Tretment B}) + 5.7(\text{Tretment C})$$

Call: lm(formula = Score ~ Treatment.ct, data = my.data)

Residuals: Min 1Q Median 3Q Max -20.500 -8.000 -2.000 6.275 29.500

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 97.800 4.031 24.263 <2e-16 *** Treatment.ct1 7.700 5.700 1.351 0.185
Treatment.ct2 2.700 5.700 0.474 0.639
Treatment.ct3 5.700 5.700 1.000 0.324
— Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘’ 1

Residual standard error: 12.75 on 36 degrees of freedom Multiple R-squared: 0.05534, Adjusted R-squared: -0.02338 F-statistic: 0.703 on 3 and 36 DF, p-value: 0.5565

```
A     B     C     D
```

105.5 100.5 103.5 97.8

The intercept shows the mean score for Treatment D (our reference). Each of the other coefficients in the model represents the difference between mean score for that treatment with the mean score for Treatment D.

2.

H_0: $\mu_{\text{treatment A}} = \mu_{\text{treatment B}} = \mu_{\text{treatment C}} = \mu_{\text{treatment D}}$

H_1: At least one of group means is different from $\mu_{\text{treatment D}}$.

We failed to reject the null, because of all the below: The mean scores in the treatment group A was not significantly different from the mean score in Treatment D. (p>.05) The mean scores in the treatment group B was not significantly different from the mean score in Treatment D. (p>.05) The mean scores in the treatment group C was not significantly different from the mean score in Treatment D. (p>.05)

3.

$$\hat{Y_D} = 97.8 + 7.7(0) + 2.7(0) + 5.7(0) = 97.8$$

The mean score in Treatment D is 97.8.

$$\hat{Y_B} = 97.8 + 7.7(0) + 2.7(1) + 5.7(0) = 100.5$$

The mean score for Treatment B is 100.5.