EDMS 646: Homework 3

Minoo Ahmadi

March 9, 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.

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. As opposed to ANOVa, which is an omnibus test, regression table is reporting t-test 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.

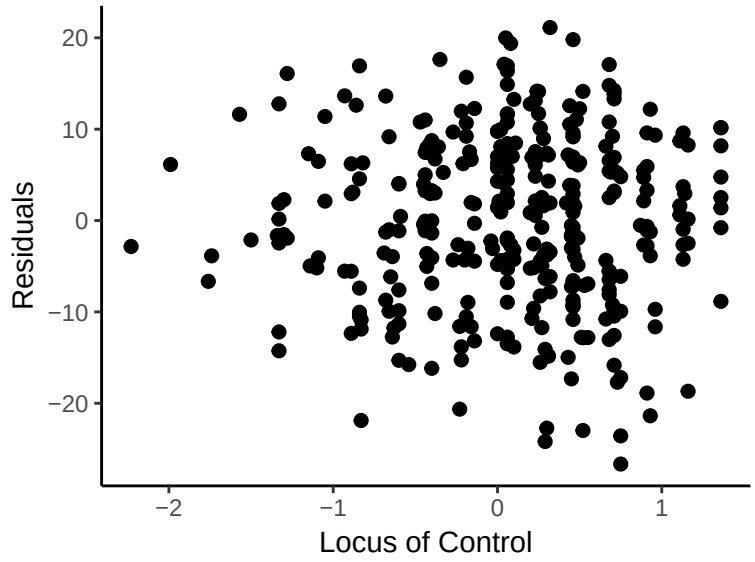
4. Independence: Having a look at plots for resicuals vs individual predictors, seems that the independence assumption is violated for all the predictors, specially for motivation and self-concept. Several groups of dependent data points can be seen in these plots.

Normality: The histogram for standarized residulas seems a little bit left-skewed (skewness = -.25). But both PP plot and QQ plot seems fairly normal. The K-S test returns significant results though (p<.001), which is not much reliable since our sample size is >50 (we have 300 subjects). All in all, I think normality can be assumed.

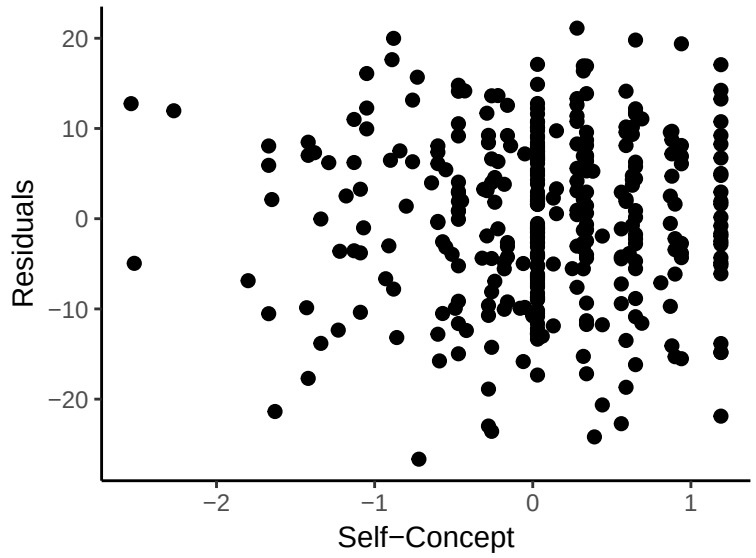
Linearity: The plot for residuals vs predicted Y shows fair linearity. The data points seem to be equally distributed above and below the y = 0 line.

Homoscedasticicy: Some level of fanning out is observed in the residulads vs predicted Y plot. This supports potential violation of homogenity of variances.

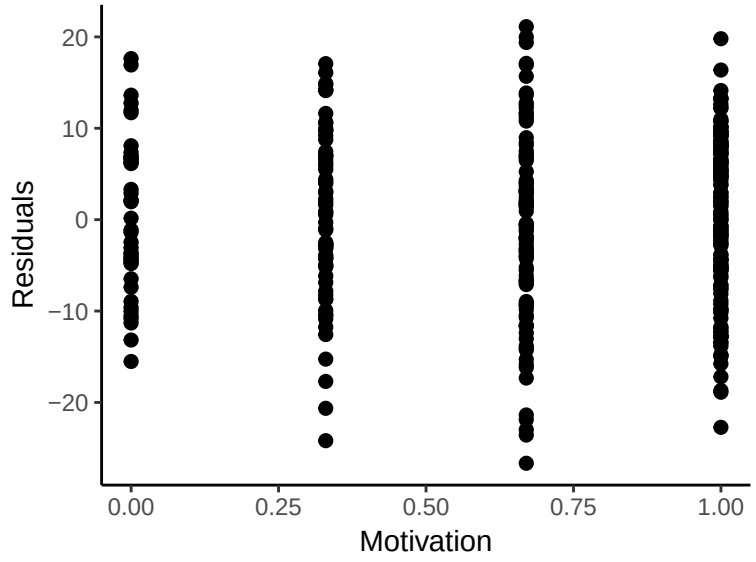
```
#Plot: Residuals vs X
ggplot(hsb1.data, aes(locus, residuals(hsb1.lm)))+ geom_point(size = 2) + xlab("Locus of Control") + yla
   theme(
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        panel.border = element_blank(),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

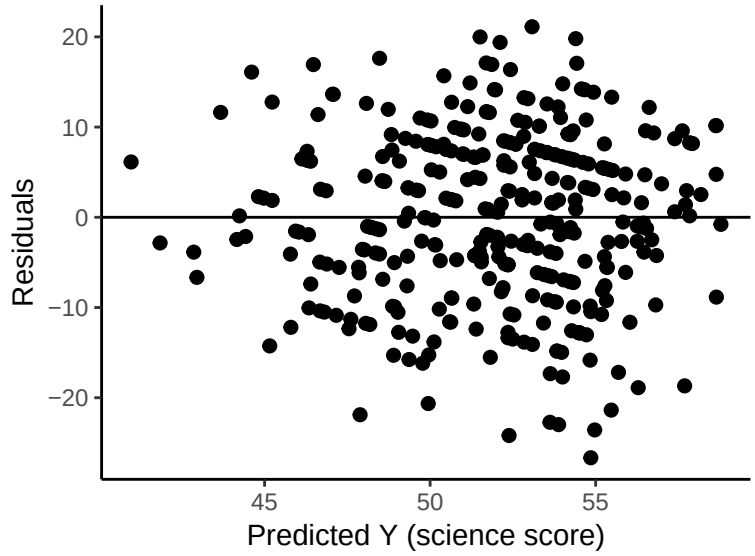



```
ggplot(hsb1.data, aes(mot, residuals(hsb1.lm)))+ geom_point(size = 2) + xlab("Motivation") + ylab("Resi
   theme(
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        panel.border = element_blank(),
        axis.line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

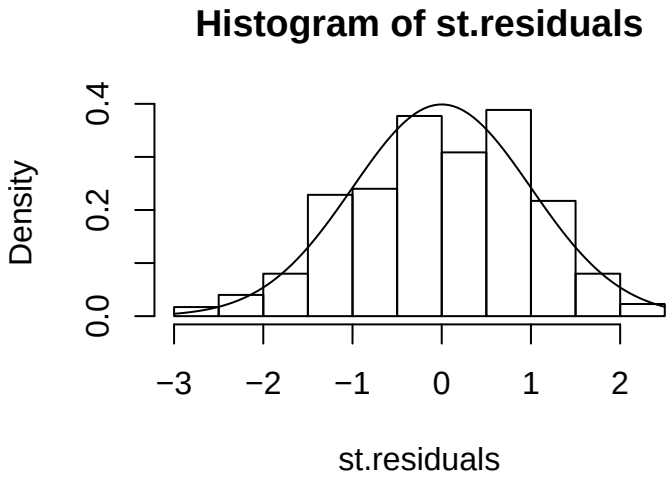



```
#Plot: Resisulas vs Predicted Y
ggplot(hsb1.data, aes(fitted.values(hsb1.lm), residuals(hsb1.lm)))+ geom_point(size = 2) + xlab("Predic
   theme(
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        panel.border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

```
        axis.line = element_line(),
        axis.ticks = element_line(),
    )
```

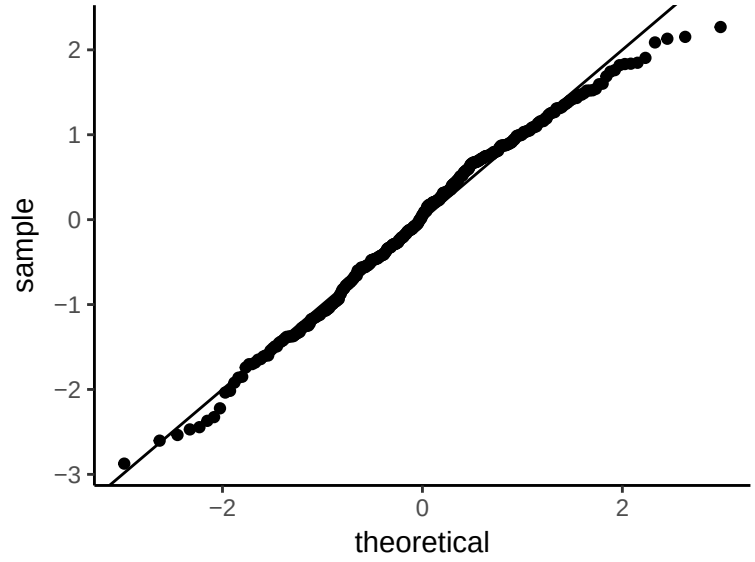



```
# getting stadardized residuals
st.residuals <- stdres(hsb1.lm)
# create residuals hist
hist(st.residuals, freq = FALSE)
curve(dnorm, add = TRUE)
```

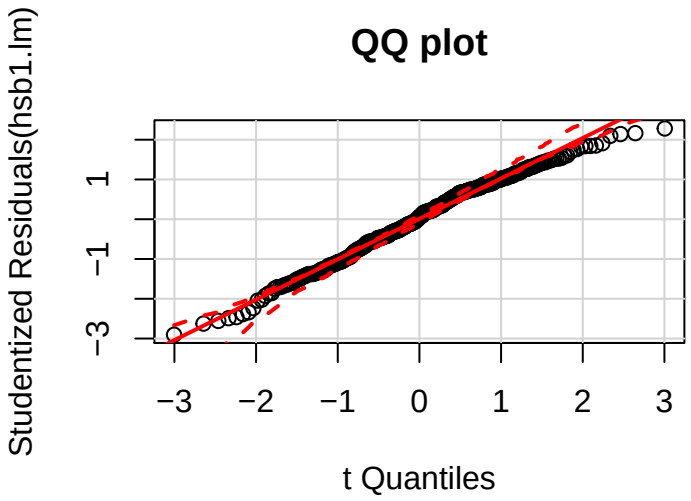



```
# create PP plot
ggplot(hsb1.data, aes(sample = st.residuals))+ stat_qq() + geom_abline(intercept = 0, slope = 1) +
   theme(
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        panel.border = element_blank(),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

```
        axis.line = element_line(),
        axis.ticks = element_line(),
    )
```




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



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

1.

2. It looks evenly distributed to me and I don't seem to detect any outliers here.

Part 3: ANOVA using Multiple Regression

1.

2.

3.