# EDMS 646: Homework 2

*Minoo Ahmadi*

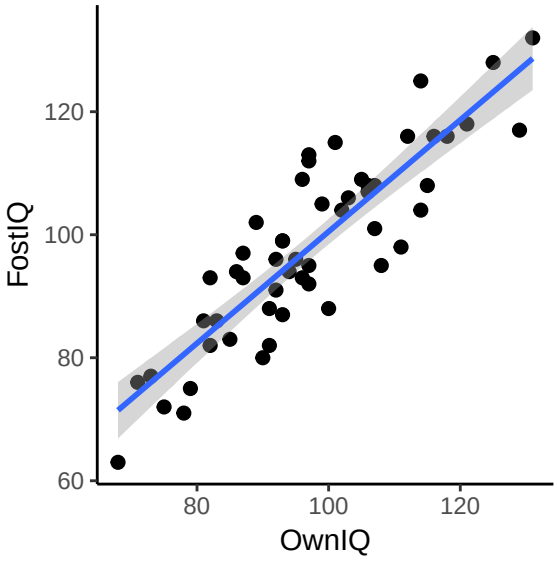
*February 23, 2017*

## Part 1: Correlation

1. There's a relatively strong positive linear relationship between IQ score of identical twins raised in foster homes (FostIQ) and that of their siblings whom were raised in the biological parent's homes (QwnIQ).

```
ggplot(burt.data, aes(OwnIQ, FostIQ)) + geom_point(size = 2) + geom_smooth(method = "lm") +
   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(),

    )
```




2.

```
cor.test(burt.data$OwnIQ, burt.data$FostIQ)
```

```
##
##  Pearson's product-moment correlation
##
## data:  burt.data$OwnIQ and burt.data$FostIQ
## t = 13.016, df = 51, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7947517 0.9272713
## sample estimates:
##       cor
```

```
## 0.8767131
```

correlation coefficient (r): 0.8767131 p $< .001$

3.

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.8767131\sqrt{\frac{51}{0.2313741}} = 13.01623$$
$$\nu = n - 2 = 53 - 2 = 51$$
$$\rightarrow t(51) = 13.01$$

4. The critical t-value is $\pm 2.008$ and the observed t-value exceeds this value. Hence, r is statistically significant at the $\alpha = .05$.

5. A test of the Pearson correlation was used to address the linear relation between IQ score of identical twins raised in foster homes (M = 98.11, SD = 15.21) and that of their siblings whom were raised in the biological parent's homes (M = 97.36, SD = 14.69). Using an alpha level of 0.05, this test was found to be statistically significant($\hat{\rho} = 0.87$, $t(51) = 13.01$, $p < .05$ (two-tailed)) indicating that these two variables are positively linearly related.

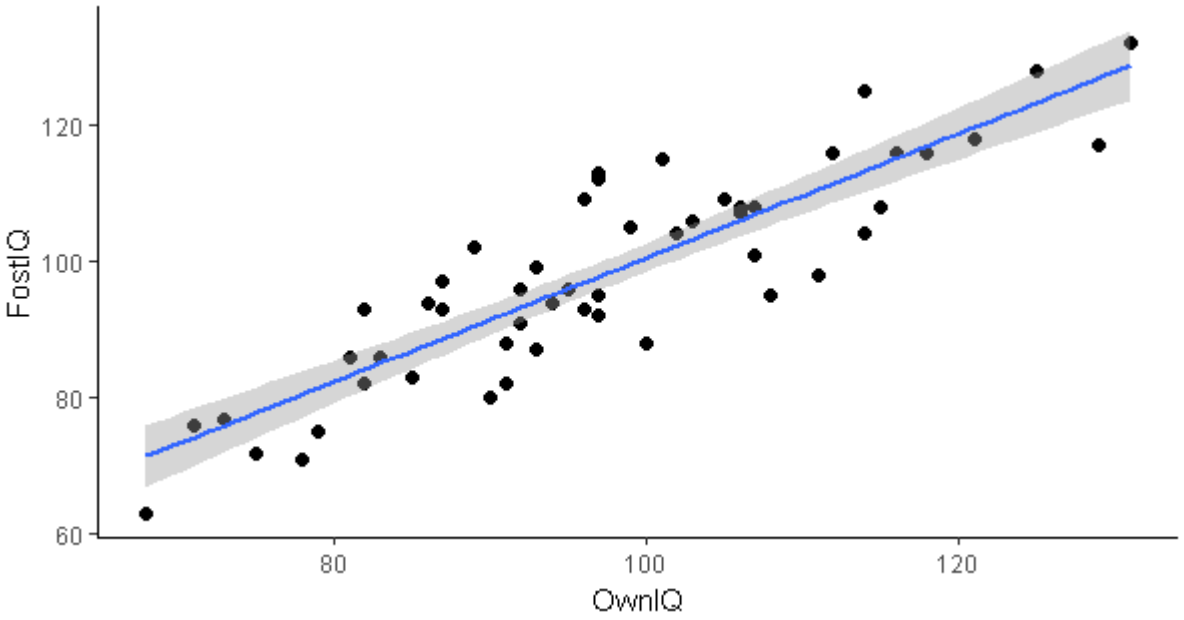



Figure 1: Correlation between IQ scores of the 2 groups.

# Part 2: Simple Regression

1.

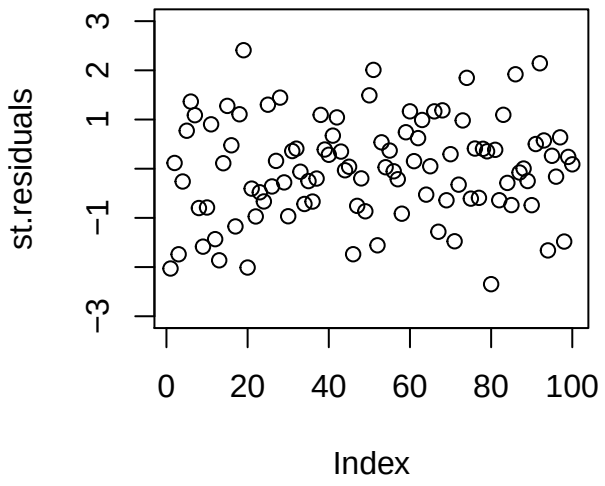
```
reg.lm<- lm (Y~X, data = reg.data)
summary(reg.lm)
```

```
##
## Call:
## lm(formula = Y ~ X, data = reg.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -17.8164  -5.0267   0.2628   4.7536  18.1045
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  44.2801     6.6253   6.684 1.43e-09 ***
## X             1.8573     0.2194   8.467 2.54e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.638 on 98 degrees of freedom
## Multiple R-squared:  0.4225, Adjusted R-squared:  0.4166
## F-statistic: 71.69 on 1 and 98 DF,  p-value: 2.535e-13
```

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

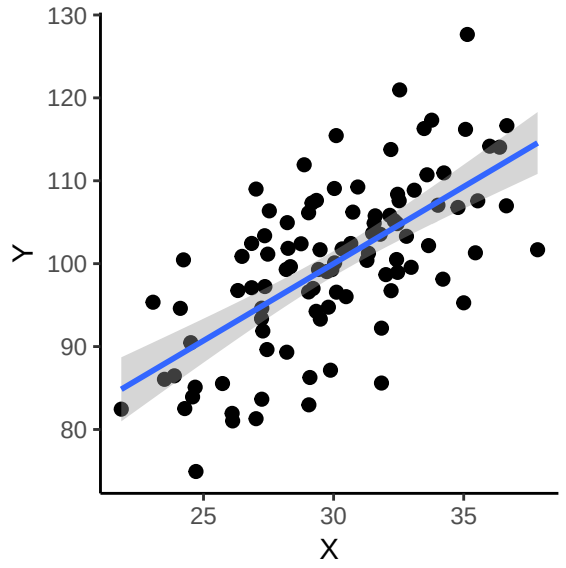
```
plot(st.residuals, ylim = c(-3, 3))
```




3. (a)

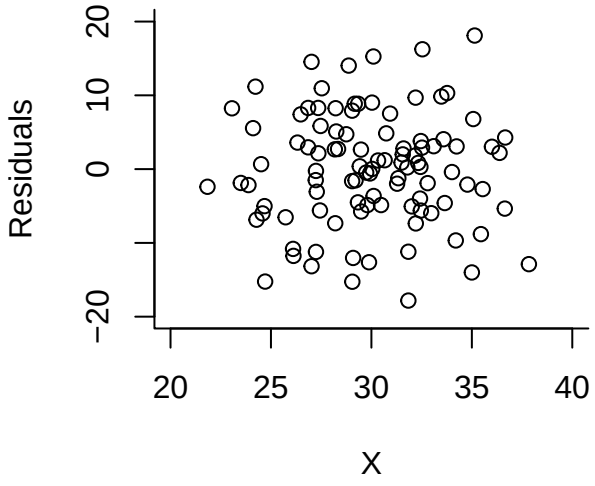
```
ggplot(reg.data, aes(X, Y))+ geom_point(size = 2) + geom_smooth(method = "lm") +
  theme(
       panel.grid.major = element_blank(),
       panel.grid.minor = element_blank(),
       panel.background = element_blank(),
       panel.border = element_blank(),
       axis.line = element_line(),
       axis.ticks = element_line(),

    )
```



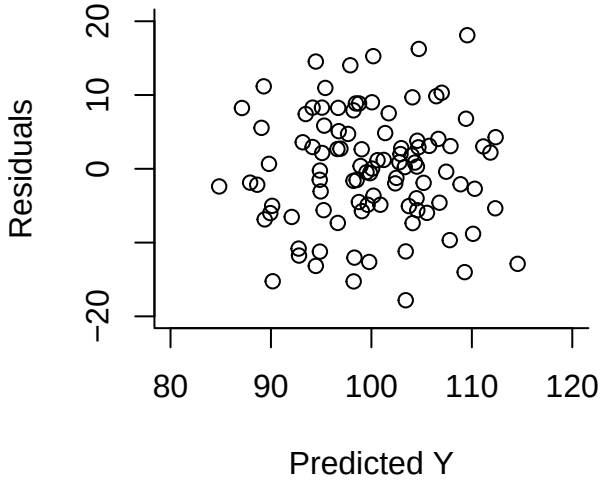

(b)

```
plot(residuals(reg.lm)~reg.data$X, col="black",
pch=1, bty="l", xlab="X",
ylab="Residuals",
xlim=c(20,40),ylim=c(-20,20))
```




(c)

```
plot(residuals(reg.lm)~fitted.values(reg.lm), col="black",
pch=1, bty="l", xlab="Predicted Y",
ylab="Residuals",
xlim=c(80,120),ylim=c(-20,20))
```




4.

Independence of observarions: Not violated. Observations seem to be independent from each otehr.

Linearity of the function relating X and Y: A positive linear relationship can be seen between X and Y from the plot in 3(a).

Homoscedasticity of Y or residuals: I can't see any obvious fanning in the residuals distribuion plot.

Normality of Y or residuals: I'm not sure if this is a normal distribution. It looks more uniform rather than bell-shaped. So, normality might have been violated, but we need better statistical analysis to test this.

5.

$$\hat{Y} = 44.2801 + 1.8573X$$

When X is equal to zero, Y has a value of 44.2801 (the baseline value of Y). The regression line has a slope of 1.8573, which indicates 1 unit increase in X increases Y by 1.8573 units. $r = 0.65$

6. `(a)`

$r = 0.65$ There's a medium to large correlation between X and Y.

`(b)`

$r^2 = 0.4225$. This indicates the shared variance between X and Y. In other words, X explains 42% of the variablity in Y.

`(c)`

$SS_{\text{regression}} = 4182.5$. It indicates the variablity in Y due to its relationship with X. It seems large, which makes sense because they have a significant correlation.

`(d)`

$SS_{\text{residual}} = 5717.8$. It indicates the variablity in Y not explained by its relationship with X. This is the value that is minimized by the least squares procedure.

`(e)`

$S_b$ = standard error of the slope = 0.2194. Variablity of the sampling distribution for slope.

(f)

$t_{\text{observed}} = 8.467$, $t_{\text{critical}} = 1.984$ the observed t exceeds the critical t, which indicates X significantly predicts Y.

(g)

$F(1, 98) = 71.69$. This is a large F-value and shows that a large portion of variabiliy in Y is explained by our model.

(h)

p-value for the slope< .001. It indicates a significant relationship between X and Y.

(i)

$\%95CI = [1.421976 \quad 2.292618]$. Our slope is outside the $\%95CI$ range and we can reject the null hypothesis.

7. (a)

This research was designed to determine the influence of hours of sleep at night on college students' performance in a test the day after.

(b)

A simple linear regression was performed to evaluate the relationship between sleep hours and performance and to see whether hours of sleep can predict students' performance on the test.

(c)

Students' performance was regressed on the average sleep hours the night before. The overall multiple regression was statistically significant ($R^2 = 0.42, F(1, 98) = 71.69, p < 0.001$). Sleep hours accounted for 42% of the variance in students' performance on the test. The unstandardized regression coefficient ($\beta$) for sleep hours was 1.85 ($t(98) = 2.266, p < 0.001$), meaning that for each additional hour of sleep, students' performance on the test increased by 1.85 points. This finding suggests that for each additional hour students sleep the night before the test, their performance will improve by 1.85 points.

(d)

These results suggest that sleeping well the night before a test is indeed an important influence on students' performance on the test. Students who want to improve their performance on the test may do so by sleeping for longer hours the night before the test. These findings suggest that each additional hour of sleep the night before the test should result in close to a 2-point increase in students' performance on the test.

# Part 3: ANOVA Table

1.

| Source | SS | df | MS | F | sig. |
|---|---|---|---|---|---|
| Regression | 300 | 4 | 75 | 59.25 | <.001 |
| Residuals | 500 | 5 | 1.26 | | |
| Total | 800 | 9 | | | |

Table 1: The ANOVA table.