Assignment 6

```r
library(MASS)
library(tidyverse)
```

1.

a).

```r
# import data
prestige <- read_csv("prestige.csv", col_types = cols(type = col_factor()))
```

```r
ols.lm <- lm(data = prestige, prestige ~ type*education + type*income)
summary(ols.lm)
```

```
##
## Call:
## lm(formula = prestige ~ type * education + type * income, data = prestige)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.2629  -5.5337  -0.2431   5.1065  22.5198
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        28.057263  12.365744   2.269   0.0294 *
## typewc            -39.051009  23.080172  -1.692   0.0993 .
## typebc            -32.007806  14.109231  -2.269   0.0294 *
## education           0.338214   0.151904   2.226   0.0323 *
## income              0.414268   0.156445   2.648   0.0119 *
## typewc:education    0.088180   0.275596   0.320   0.7508
## typebc:education   -0.018591   0.318369  -0.058   0.9538
```

```
## typewc:income      0.008834   0.273425   0.032   0.9744
## typebc:income      0.369143   0.203880   1.811   0.0786 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.647 on 36 degrees of freedom
## Multiple R-squared:  0.9233, Adjusted R-squared:  0.9063
## F-statistic: 54.17 on 8 and 36 DF,  p-value: < 2.2e-16
```

We notice that the interaction terms are all insignificant, so we will fit the model without it.

```r
ols.lm <- lm(data = prestige, prestige ~ type + education + income)
summary(ols.lm)
```

```
##
## Call:
## lm(formula = prestige ~ type + education + income, data = prestige)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.890  -5.740  -1.754   5.442  28.972
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.47249    9.53329   1.728  0.09172 .
## typewc      -31.31865    5.07854  -6.167 2.75e-07 ***
## typebc      -16.65751    6.99301  -2.382  0.02206 *
## education     0.34532    0.11361   3.040  0.00416 **
## income        0.59755    0.08936   6.687 5.12e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.744 on 40 degrees of freedom
## Multiple R-squared:  0.9131, Adjusted R-squared:  0.9044
## F-statistic:   105 on 4 and 40 DF,  p-value: < 2.2e-16
```

Hypothesis Test for Significance of Regression

$$
\begin{aligned}
&H_0: \ \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \\
&\text{vs.} \\
&H_A: \ \text{at least one } \beta \neq 0
\end{aligned}
$$

p-value is almost 0, we reject the null hypothesis and conclude that there is a linear relationship.

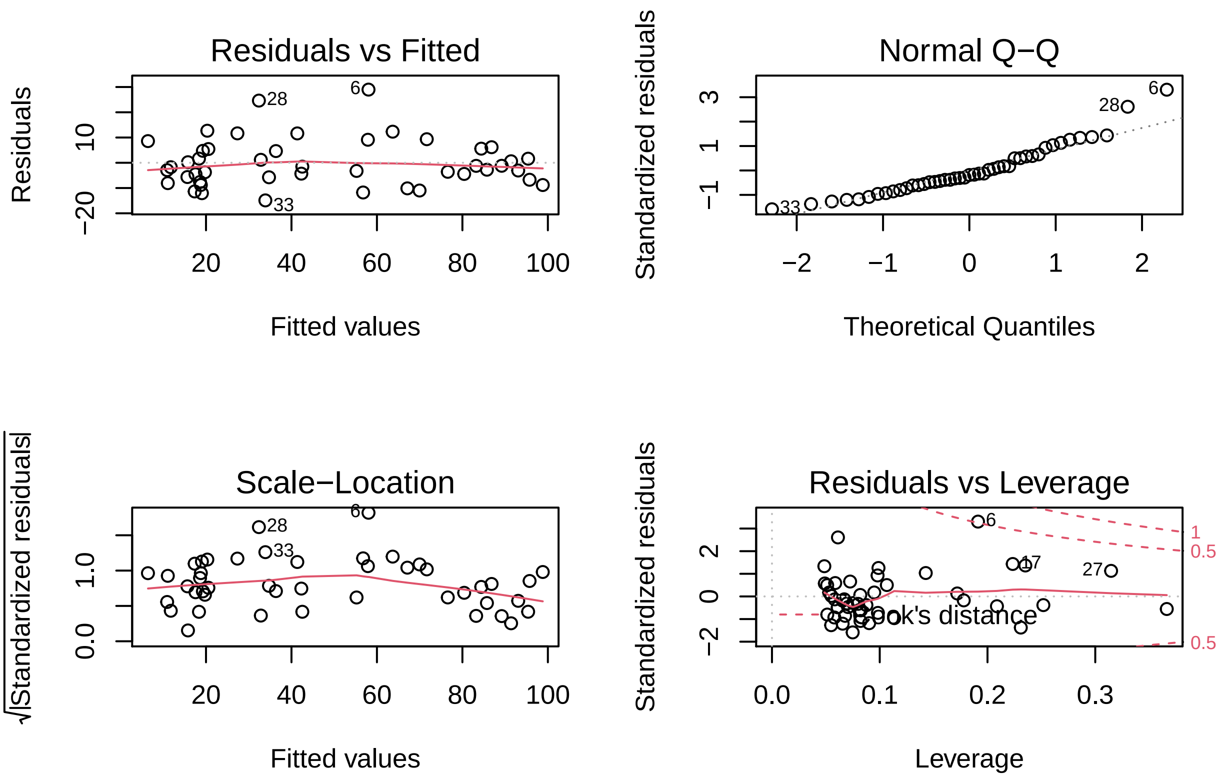
Hypothesis Test for a single β_j

$$
\begin{aligned}
&H_0: \ \beta_j = 0 \\
&\text{vs.} \\
&H_A: \ \beta_j \neq 0, \ \ \text{given other regressors in the model}
\end{aligned}
$$

p-values for type prof, type wc, education, and income are $< \alpha = 0.05$, we conclude that they are all significant.

b).

```
par(mfrow = c(2,2))
plot(ols.lm)
```




```
rstandard(ols.lm)
```

```
##           1           2           3           4           5           6
## -0.12878934 -0.29243659 -0.32717002 -0.47012563  0.59098092  3.30611267
##           7           8           9          10          11          12
##  0.65747539 -0.96002953 -0.38349498 -0.12678878 -1.08436414 -0.72706201
##          13          14          15          16          17          18
##  0.17453779 -1.18121952 -0.38451290 -0.55543070  1.43450883  1.03488294
##          19          20          21          22          23          24
## -1.37658430  0.06455453  1.36757853  0.13210431 -0.17392632 -0.43249334
##          25          26          27          28          29          30
##  1.33561081  1.25760823  1.12994535  2.60806656  0.57270149 -0.61498160
##          31          32          33          34          35          36
## -0.92963241  0.92766602 -1.58877526 -0.18621577 -0.48116260  0.49836617
```

```
##          37          38          39          40          41          42
##  0.17358804 -1.27150150 -0.85706893  0.02347531 -1.20447008 -0.80143853
##          43          44          45
## -0.30899084  0.50361636 -0.60438060
```

Looking the residual plot and standardized residuals, point 6 is a problematic observation. the profeesion for observation 6 is minister.

c).

```r
rr.lm <- rlm(data = prestige, prestige ~ type + income + education, psi = psi.huber)
sum.rr.lm <- summary(rr.lm)
sum.rr.lm
```

```
##
## Call: rlm(formula = prestige ~ type + income + education, data = prestige,
##     psi = psi.huber)
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.676  -6.646  -1.073   5.626  33.098
##
## Coefficients:
##             Value    Std. Error t value
## (Intercept)  14.4580   8.6380     1.6738
## typewc      -30.6474   4.6016    -6.6601
## typebc      -15.7434   6.3363    -2.4846
## income        0.6691   0.0810     8.2639
## education     0.3023   0.1029     2.9367
##
## Residual standard error: 8.797 on 40 degrees of freedom
```

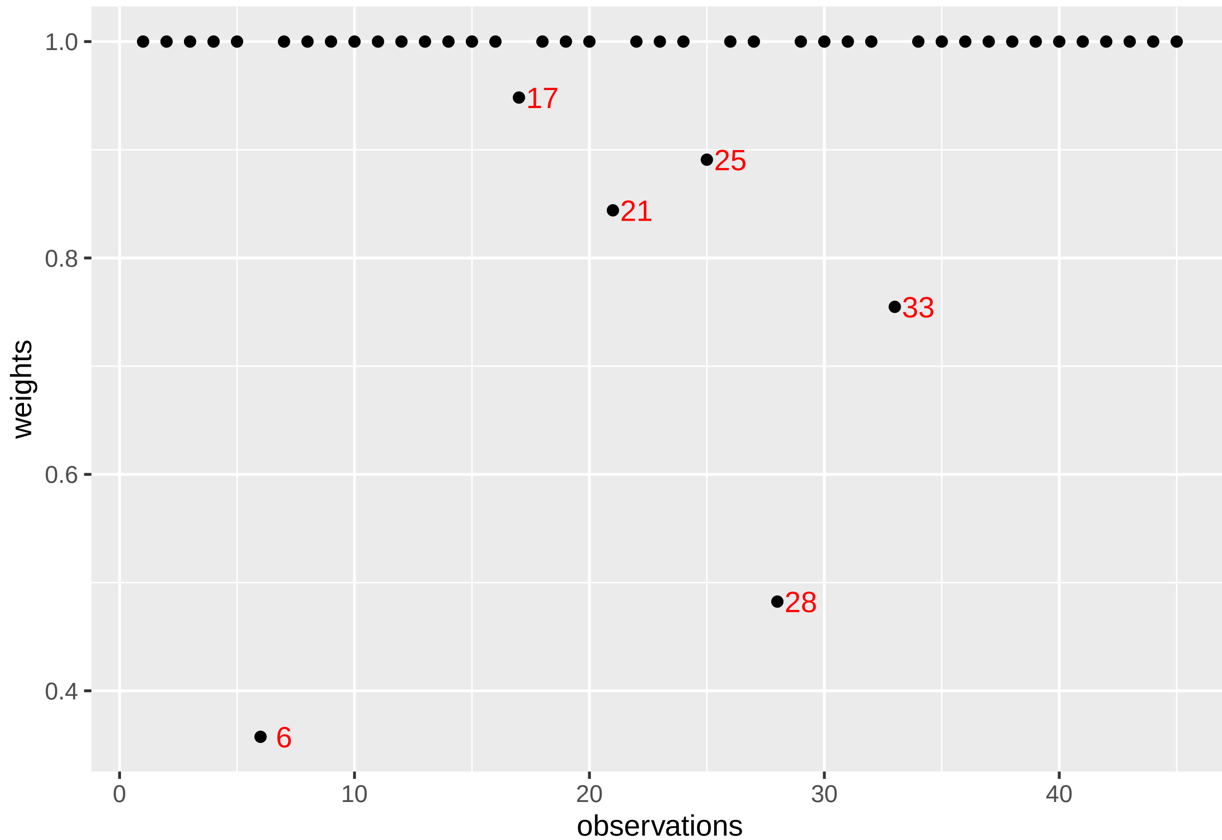
```r
tval <- sum.rr.lm$coefficients[,3]
pval <- 2*pt(abs(tval),40, lower = FALSE)
pval
```

```
##  (Intercept)       typewc       typebc       income    education
## 1.019822e-01 5.592769e-08 1.725200e-02 3.481343e-10 5.478526e-03
```

According to the p-value we calculated, the same variables appear significant when compared to the usual linear model fit.

d).

```
plot.data <- as_tibble(rr.lm$w) %>%
  rename(weights = value) %>%
  mutate(observations = row_number())
point.data <- filter(plot.data, weights != 1)
ggplot(data = plot.data, aes(x = observations, y = weights)) +
  geom_point() +
  geom_text(data = point.data,aes(label=observations), color = "red", nudge_x = 1)
```



the profession for observation 6 is minister has the smallest weights which is expected to see.

2.

```r
# import data
assignment6 <- read_csv("assignment6.csv")
```

a).

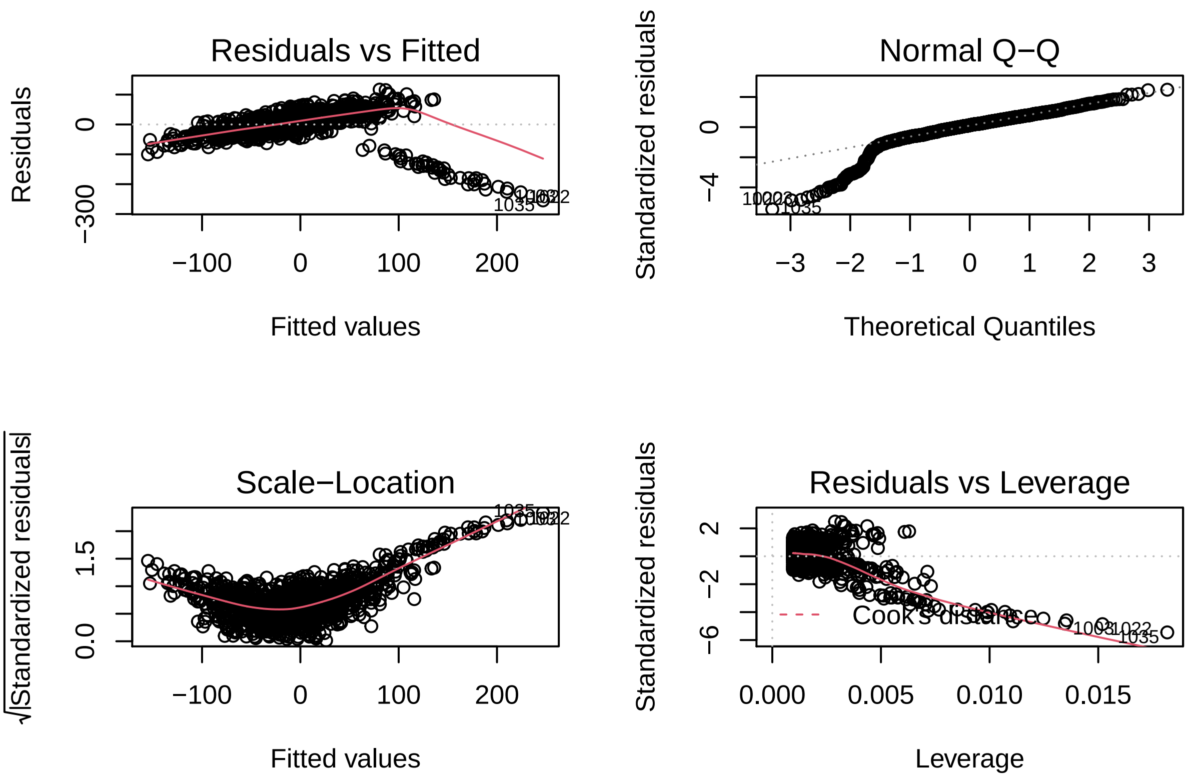
```r
lm_mdl <- lm(data = assignment6, y ~ x)
summary(lm_mdl)
```

```
##
## Call:
## lm(formula = y ~ x, data = assignment6)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -254.295  -18.438    6.238   27.523  116.715
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -9.067      1.460   -6.21 7.64e-10 ***
## x             49.850      1.232   40.46  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47.08 on 1048 degrees of freedom
## Multiple R-squared:  0.6097, Adjusted R-squared:  0.6093
## F-statistic:  1637 on 1 and 1048 DF,  p-value: < 2.2e-16
```

The fitted model is $y = -9.067 + 49.850x$.

b).

```r
par(mfrow = c(2,2))
plot(lm_mdl)
```



According to the residual plots, i do not feel comfortable constructing a confidence interval for the slope. Because we do not see the points distributed around 0 in Residuals vs Fitted and Scale-Location plots, and the lower-tail in Normal Q-Q plot do not fall on the straight line.

c).

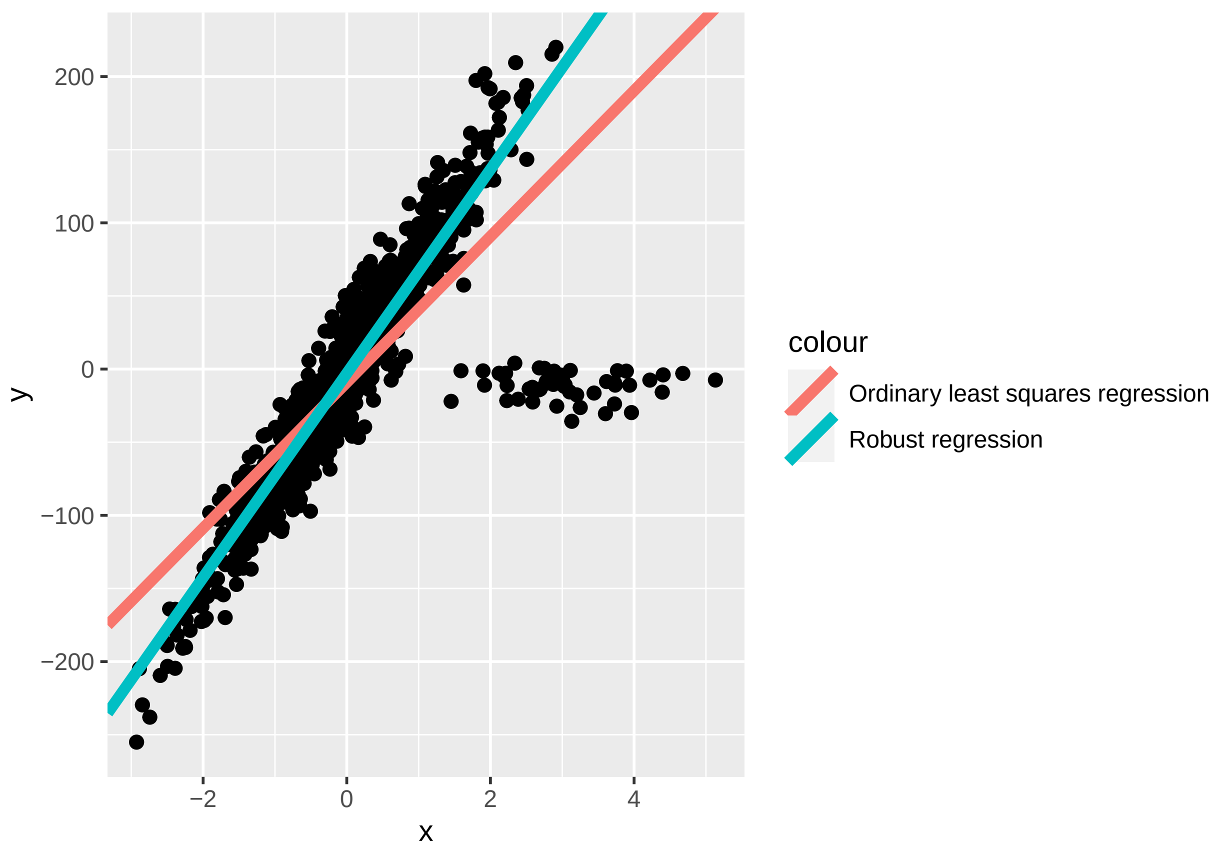
```r
rlm_mdl <- rlm(data = assignment6, y ~ x, psi = psi.huber)
summary(rlm_mdl)
```

```
##
## Call: rlm(formula = y ~ x, data = assignment6, psi = psi.huber)
## Residuals:
##       Min        1Q    Median        3Q       Max
## -362.5174  -14.0263    0.6103   14.2769   74.8127
##
## Coefficients:
##             Value    Std. Error t value
## (Intercept)  -2.9548   0.7092    -4.1662
## x            69.7441   0.5984   116.5426
##
## Residual standard error: 21.04 on 1048 degrees of freedom
```

The fitted model is $y = -2.9548 + 69.7441x$.

d).

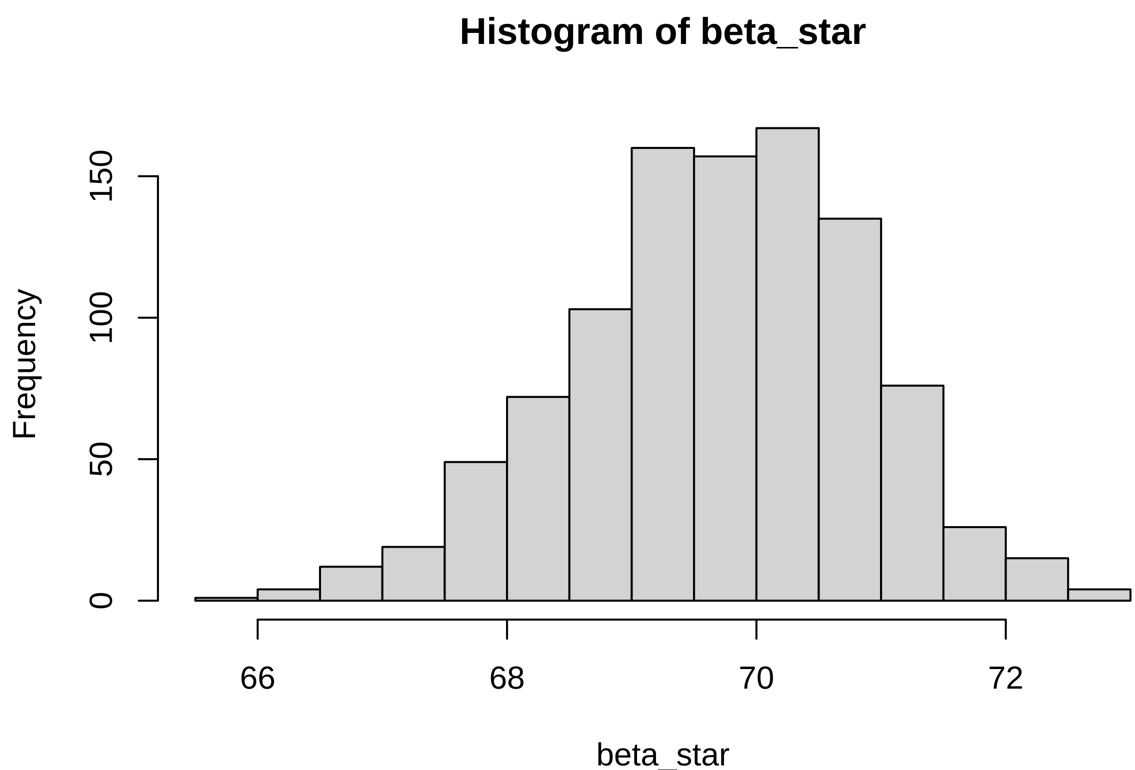
```
ggplot(data = assignment6, aes(x = x, y = y)) +
  geom_point(size = 2) +
  geom_abline(aes(slope = lm_mdl$coefficients[2],
                  intercept = lm_mdl$coefficients[1],
                  color = "Ordinary least squares regression"),
              size = 2) +
  geom_abline(aes(slope = rlm_mdl$coefficients[2],
                  intercept = rlm_mdl$coefficients[1],
                  color = "Robust regression"),
              size = 2)
```



Robust regression fits the data better in terms of the majority of data.

e).

```r
set.seed(123)
n <- 1050
m <- 1000
beta_star <- NULL
for (i in 1:m){
  index <- sample(n, replace = TRUE)
  xstar <- assignment6$x[index]
  ystar <- assignment6$y[index]
  fit <- rlm(ystar ~ xstar , psi = psi.huber)
  beta_star[i] <- coef(fit)["xstar"]
}
hist(beta_star)
```



f).

```
CI <- quantile(beta_star, c(0.025,0.975))
CI
```

```
##     2.5%    97.5%
## 67.29611 71.84983
```

The estimated 95% confidence interval is [67.29611 , 71.84983].

g).

$$H_0: \text{ the slope difference is } 0$$
$$\text{vs.}$$
$$H_A: \text{ the slope difference is not } 0$$

```r
rlm_mdl$coefficients[2]
```

```
##        x
## 69.74409
```

```r
CI
```

```
##     2.5%    97.5%
## 67.29611 71.84983
```

The estimated slope we calculated previous is 69.74409 which is inside the 95% confidence interval, so we fail to reject H_0, and conclude that the slope difference is 0.