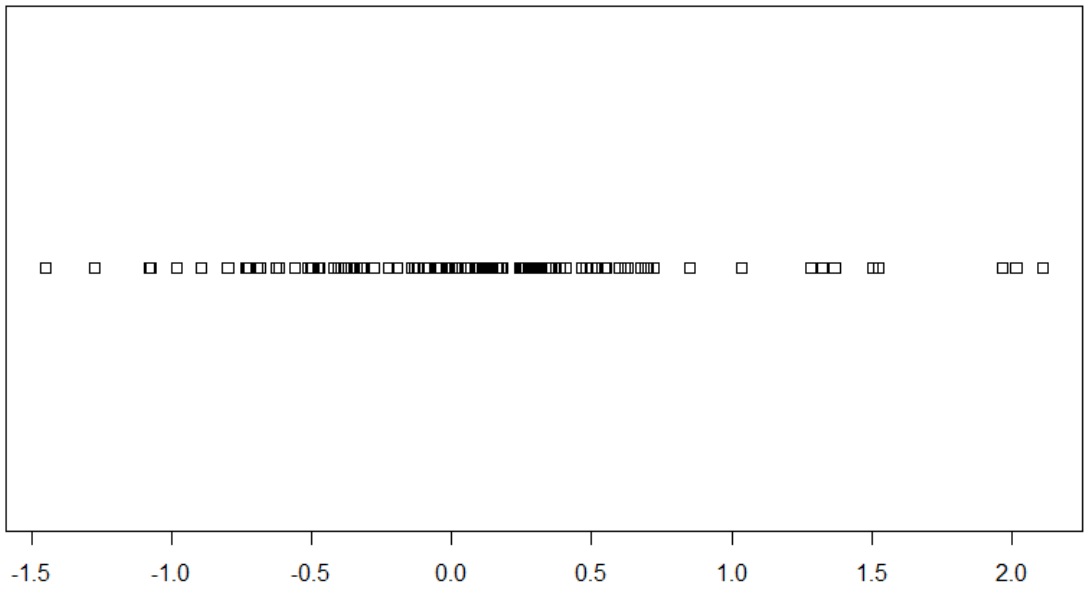
Tuan Dang
3/30/2021

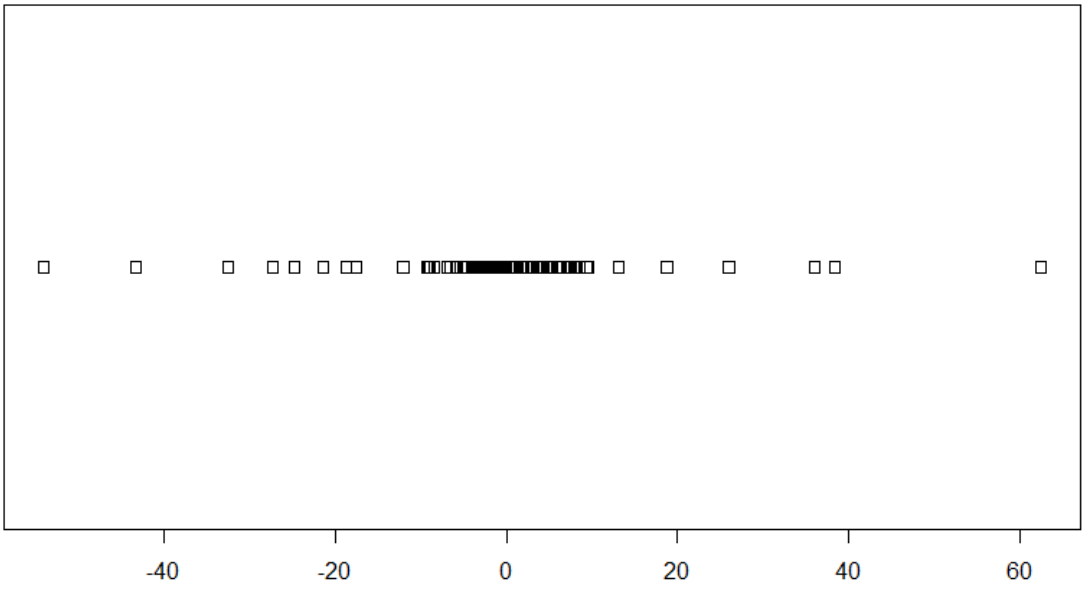
Project 1: Predicting Increase in Salary based on Increase in Education (Twins Sample Data)

2.

R Command:

```
Twins <- read_excel("C:/Users/Tuan/Downloads/Twins.xlsx")
Twins$DHRWAGE <- Twins$HRWAGEL - Twins$HRWAGEH
stripchart(Twins$DLHRWAGE)
stripchart(Twins$DHRWAGE)
```





After looking at the strip charts, it seems that taking the logs of wages gives us a smaller variance. In the first strip chart, the data points are distributed over a very small interval, whereas in the second strip chart, the range is much bigger, and the data points are focused into a big cluster in the middle with a few outliers on the sides, which is harder to look at.

3.

```
model1 <- lm(Twins$DLHRWAGE ~ Twins$DEDUC1)
summary(model1)

Call:
lm(formula = Twins$DLHRWAGE ~ Twins$DEDUC1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.07859    0.04547   1.728 0.086022 .
Twins$DEDUC1  0.09157    0.02371   3.862 0.000168 ***
```

Multiple R-squared: 0.09211

We can see the probability that the coefficient for the difference in self-reported education is equal to 0 is very small (0.000168). So, increased education is a significant predictor of increased wages.
The expected increase in log hourly wage per additional year of education is equal to the coefficient b1, which in this case is 0.09157.

4.
```
model2 <- lm(Twins$DLHRWAGE ~ Twins$DEDUC2)
summary(model2)
```

```
Call:
lm(formula = Twins$DLHRWAGE ~ Twins$DEDUC2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.07806    0.04529   1.724   0.0869 .
Twins$DEDUC2  0.09234    0.02297   4.020 9.28e-05 ***
```

Multiple R-squared: 0.09903

We can see the probability that the coefficient for the difference in cross-reported education is equal to 0 is very small (9.28e-05). So, increased education is a significant predictor of increased wages.
The expected increase in log hourly wage per additional year of education is equal to the coefficient b1, which in this case is 0.09234.

5.
R Command:
```
fits1 <- fitted(model1)
resids1 <- residuals(model1)
fits2 <- fitted(model2)
resids2 <- residuals(model2)
```

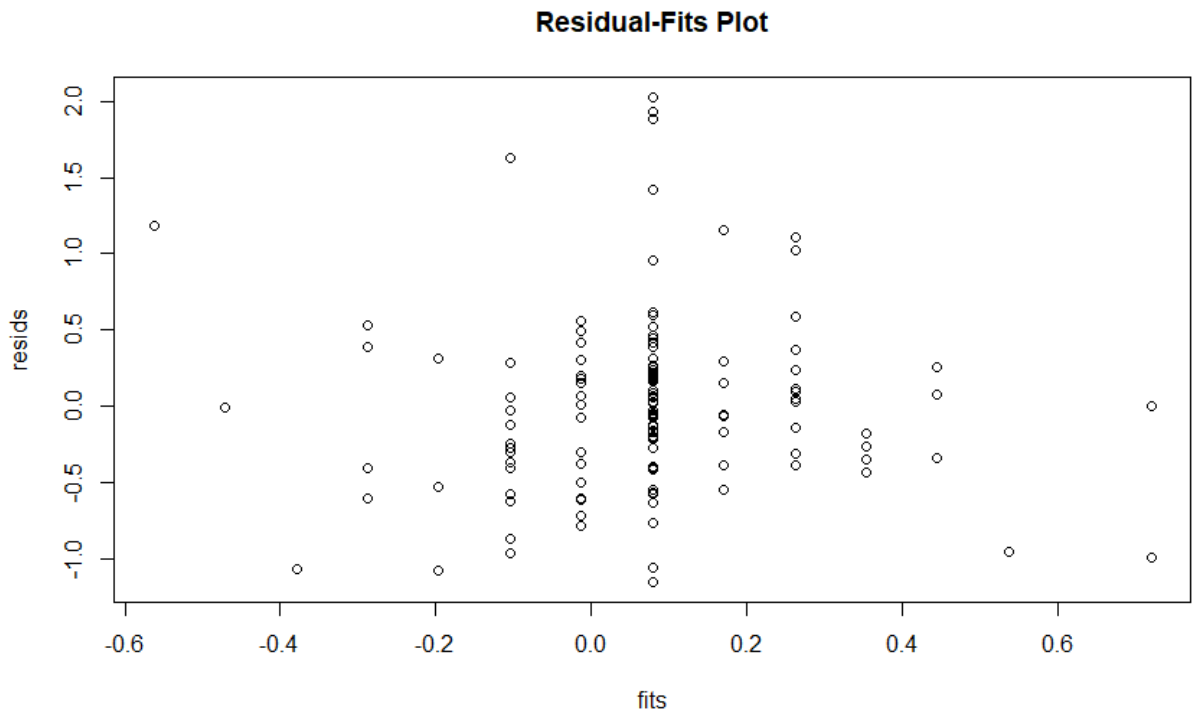
```
plot(fits1, resids1, main = "Residual-Fits Plot", xlab = "fits", ylab = "resids")
plot(fits2, resids2, main = "Residual-Fits Plot", xlab = "fits", ylab = "resids")
plot(resids1, main = "Residuals Time Series Plot")
plot(resids2, main = "Residuals Time Series Plot")
hist(resids1)
hist(resids2)
```

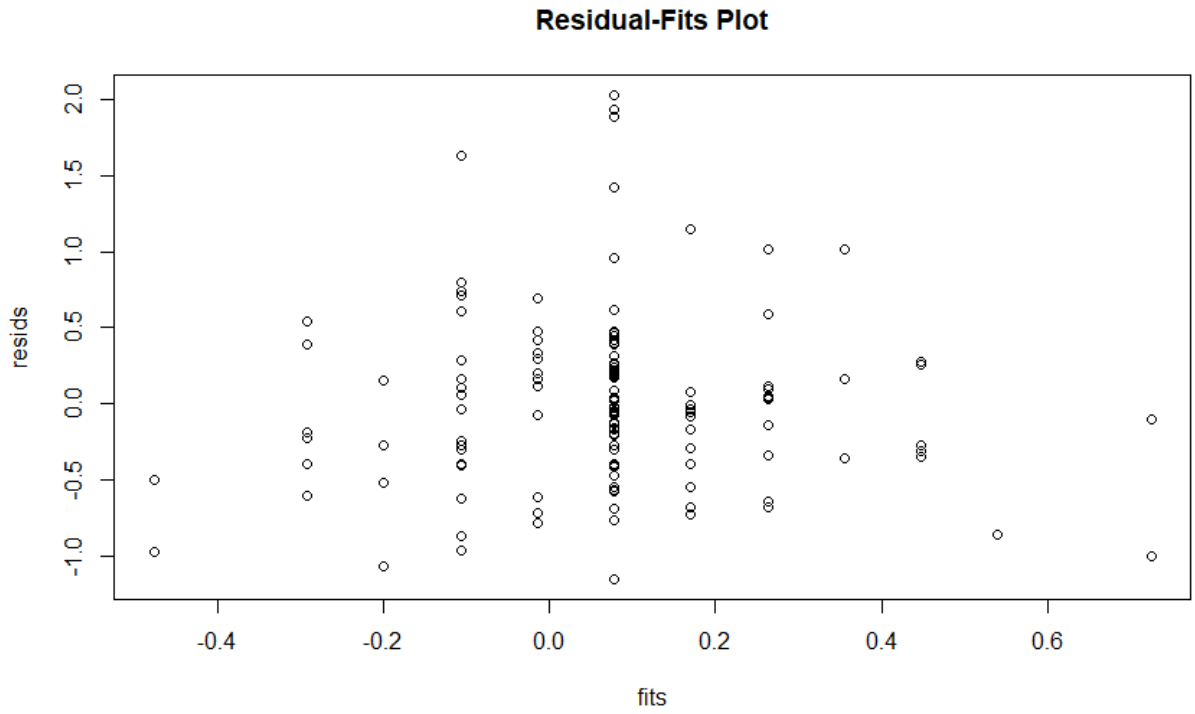
We will check each assumption one by one:

a. Validity of predictor variables and regression model:

From the summaries of the linear models from each case, we can see that the R^2 values are 0.09211 and 0.09903, respectively. This is not significant, since each one is only about 0.1. This might suggest that the model requires other predictor variables to be complete. This in turn will give us a different looking model than right now. So these two assumptions do not hold.

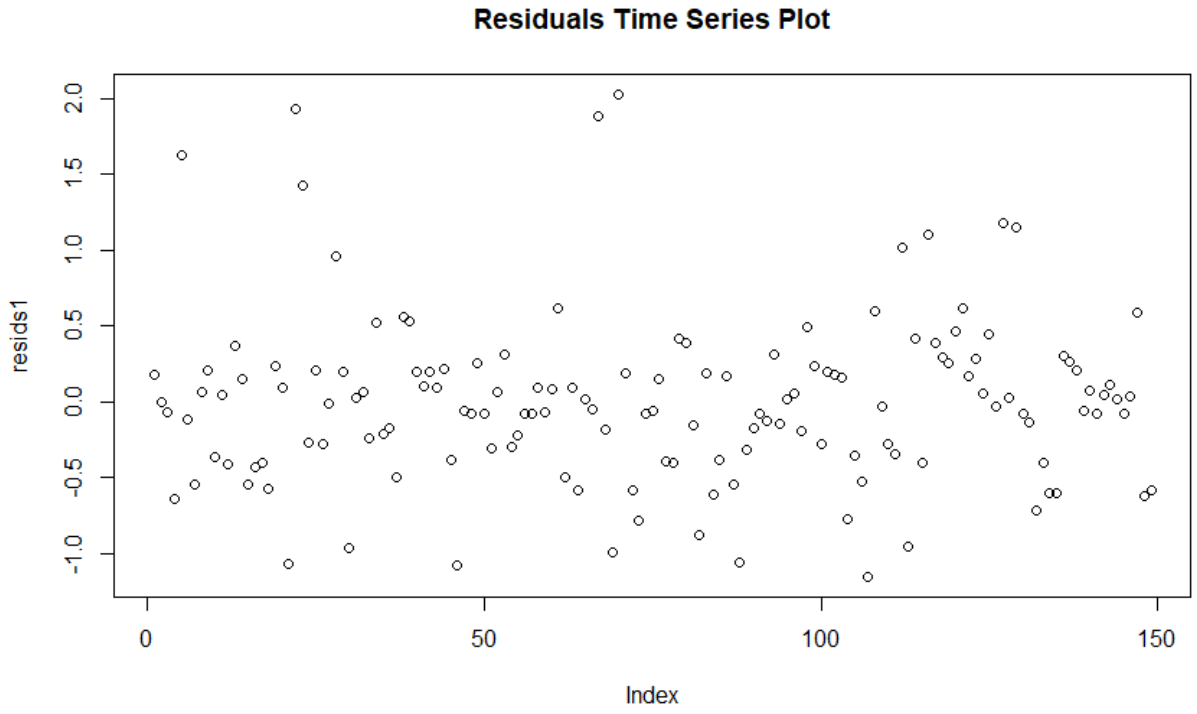
b. Homoscedasticity:

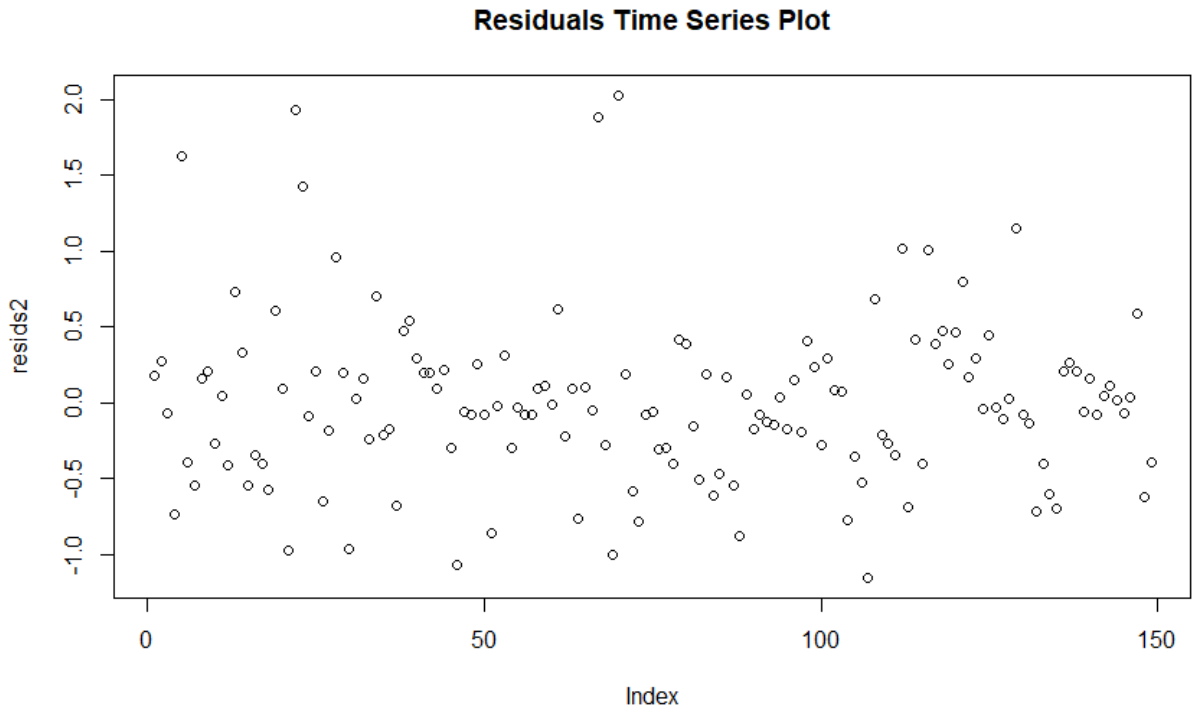




The residuals vs fits plots above do not show any particular pattern, and there does not seem to be any “fanning out” effect. So we can assume that the data does have homoscedasticity.

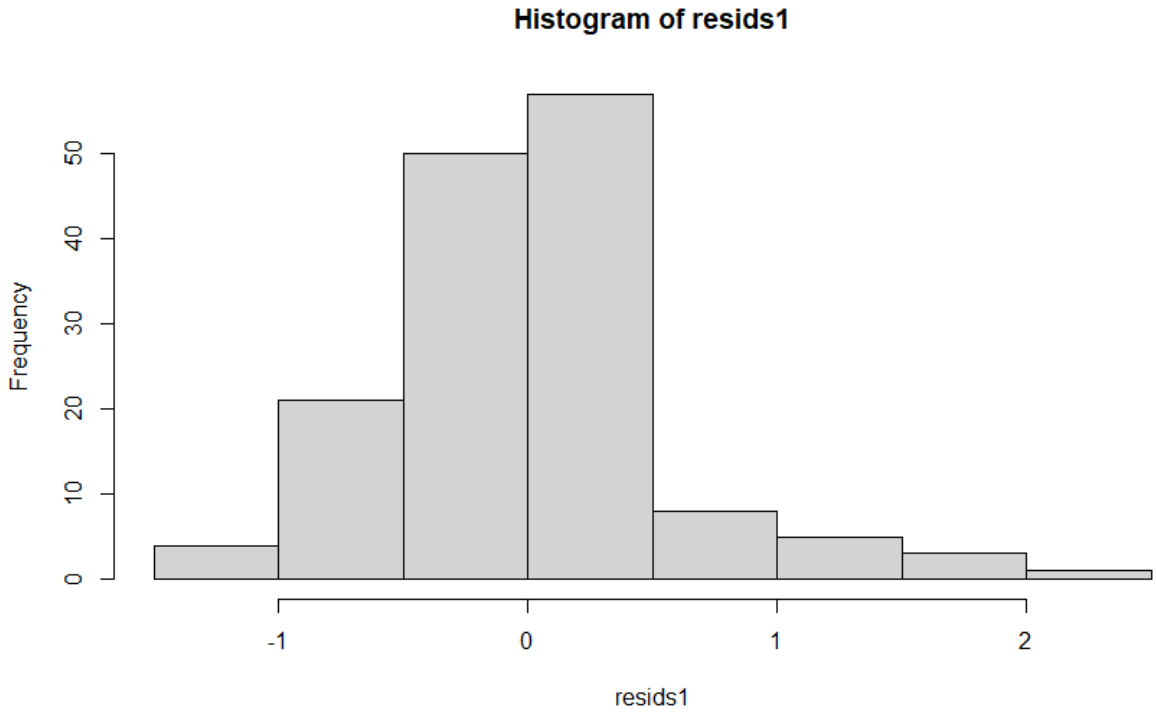
c. Residuals Independence:

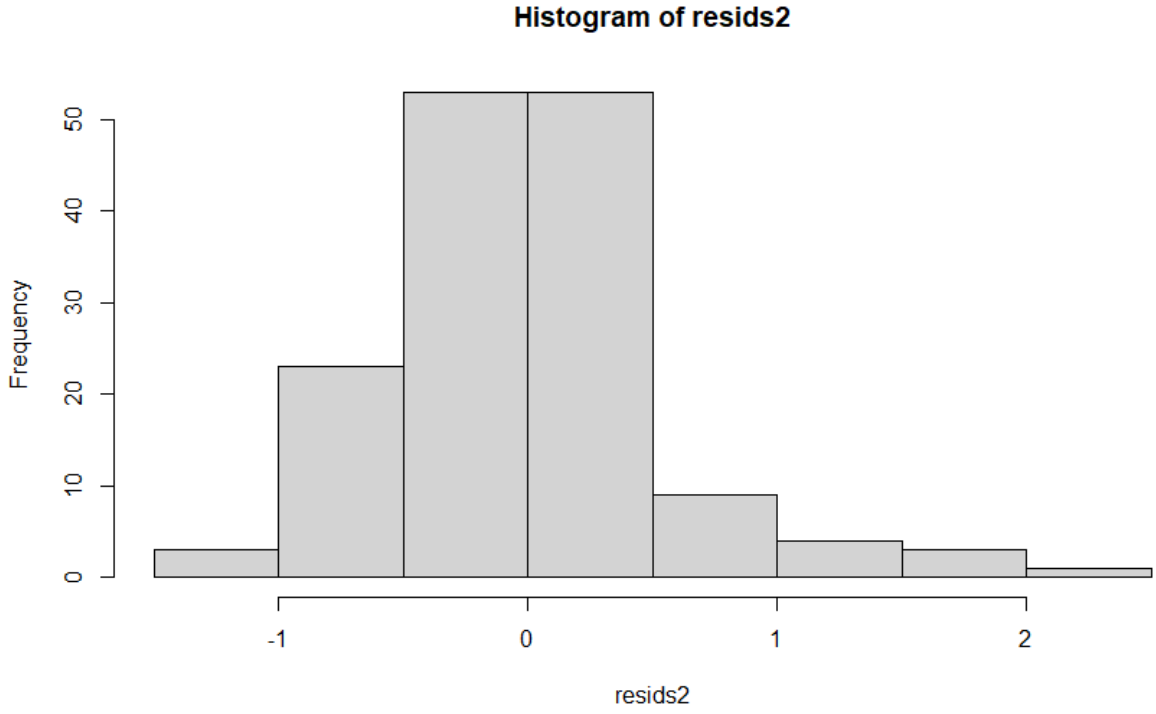




These are the residuals time series plots for both models. The pattern looks random, so we can say that the residuals are independent of each other.

d. Residuals Normality:





Based on these two histograms, we can see that the residuals seem to be normally distributed with a mean of 0 (a right-skewed bell shape centered at 0).