Testing the Assumptions of OLS using R

1. **The regression model is linear in parameters**

   - Let ***X*** be an n x k matrix, with observations on k independent variables for n observations.
   - Let ***y*** be an n x 1 vector of observations on the dependent variables.
   - Let ***a*** be an n x 1 vector of errors.
   - Let ***B*** be an k x 1 vector of unknown population parameters that we want to estimate.

   Therefore, **Ordinary Least Squares** regression model can be written as:

   $$y = XB + a$$

   Here, B_1, B_2 ... B_k are the parameters, and they are clearly linear in nature. Therefore, the regression model is linear in parameters. **Assumption holds true for this model.**

2. **The mean of residuals is zero**

```
ols <- lm(price ~ ., data = Housing);
mean(ols$residuals);
[1] -8.984169e-13
```

   The mean of residuals is a very small number extremely close to zero, therefore, approximately equal to zero. **Assumption holds true for this model.**

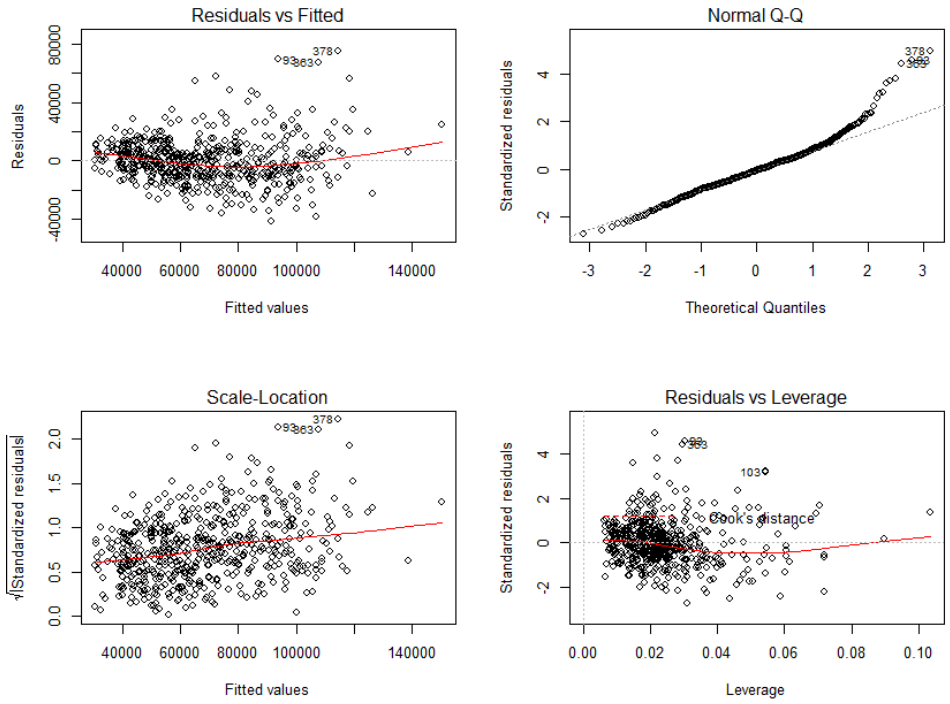
3. **Homoscedasticity of residuals or equal variance**

```
par(mfrow=c(2,2))
plot(ols);
```

   From the first plot (top-left), as the fitted values along x increase, the residuals decrease and then increase. This pattern is indicated by the red line, which should be approximately flat if the disturbances are homoscedastic. The plot on the bottom left also checks this, and is more convenient as the disturbance term in Y axis is standardized.

   In the plots (top-left, bottom-left) below, the points appear random and the line looks fairly flat, with no increasing or decreasing trend. So, **the condition of homoscedasticity holds true.**

Oishi Bhattacharyya
PRN: 14060321050



4. No autocorrelation of residuals

```
lawstat::runs.test(ols$residuals)
```

```
        Runs Test - Two sided

data:  ols$residuals
Standardized Runs Statistic = -2.9128, p-value = 0.003582
```

At a significance level of 0.05, p-value is less than 0.05, therefore we reject the null hypothesis. There is autocorrelation of residuals present in the data **(assumption does not hold true)**. In order to correct it, we add lag1 of residual as an X variable to the original model.

```
library(DataCombine)
h_data <- data.frame(Housing, resid_mod1=ols$residuals)
h_data_1 <- slide(h_data, Var="resid_mod1", NewVar = "lag1", slideBy = -1)
h_data_2 <- na.omit(h_data_1)
ols2 <- lm(price ~ . + lag1, data=h_data_2)
```

Testing again, we have:

```
lmtest::dwtest(ols2)
```

```
        Durbin-Watson test

data:  ols2
DW = 2.2925, p-value = 0.9995
alternative hypothesis: true autocorrelation is greater than 0
```

Oishi Bhattacharyya
PRN: 14060321050

At a level of significance 0.05, and a p-value of 0.9995, since the p-value is higher than the significance level, we fail to reject the null hypothesis that the autocorrelation is zero. Therefore, **this assumption also holds true.**

5. **The X variables and residuals are uncorrelated**

```
        Pearson's product-moment correlation

> cor.test(Housing$lotsize, ols$residuals)

data:  Housing$lotsize and ols$residuals
t = 6.2128e-16, df = 544, p-value = 1

> cor.test(Housing$bedrooms, ols$residuals)

data:  Housing$bedrooms and ols$residuals
t = 3.4792e-14, df = 544, p-value = 1

> cor.test(Housing$bathrms, ols$residuals)

data:  Housing$bathrms and ols$residuals
t = -1.1503e-15, df = 544, p-value = 1

> cor.test(Housing$stories, ols$residuals)

data:  Housing$stories and ols$residuals
t = 1.4883e-14, df = 544, p-value = 1

> cor.test(Housing$garagepl, ols$residuals)

data:  Housing$garagepl and ols$residuals
t = -1.2594e-18, df = 544, p-value = 1
```

In all the above correlation tests with ols residuals and X variables, for an alternative hypothesis: true correlation is not equal to 0 since the p-values are 1 (very high) we fail to reject the null hypothesis.

Therefore, **the assumption that X variables are not correlated with residuals holds true for this model.**

6. **The number of observations must be greater than number of Xs**

This is directly observed from the data. The number of observations is higher than the number of Xs. **This assumption holds true.**

Oishi Bhattacharyya
PRN: 14060321050

7. **The variability in X values is positive**

```
var(Housing$lotsize)
[1] 4700912
var(Housing$bedrooms)
[1] 0.543741
var(Housing$bathrms)
[1] 0.2521625
var(Housing$stories)
[1] 0.7537756
var(Housing$garagepl)
[1] 0.741849
```

The variability in all the X values are positive, although the last 4 variables are closer to zero than the variable lotsize. While there is still a fair bit of variance in bedrooms, stories, and garagepl, the variance of bathrms is simply not significant enough. **The model should be re-fitted by excluding these variables to better make the assumption hold true.**

8. **The regression model is correctly specified**

In this case, the Y and X variables have a direct (and not inverse) relationship, which has been specified in the regression equation for OLS. Therefore, **this assumption too, holds true.**

9. **No perfect multicollinearity**

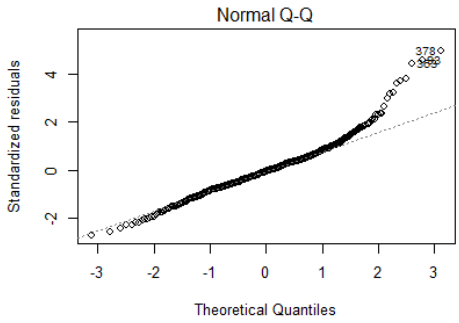
```
library(car)
vif(ols)
```

lotsize	bedrooms	bathrms	stories	driveway	recroom	fullbase	gashw	airco	garagepl	prefarea
1.407528	1.365692	1.282608	1.523841	1.199104	1.210641	1.331593	1.038258	1.201398	1.202007	1.491928

The Variance Inflation Factors (VIFs) for all variables are lesser than 2, and therefore will not cause multicollinearity. **This assumption is satisfied.**

10. **Normality of residuals:**

The residuals follow the normal distribution, except at the very ends where it deviates from the normal curve by large amounts. Therefore, **this assumption is not satisfied.**



Oishi Bhattacharyya
PRN: 14060321050