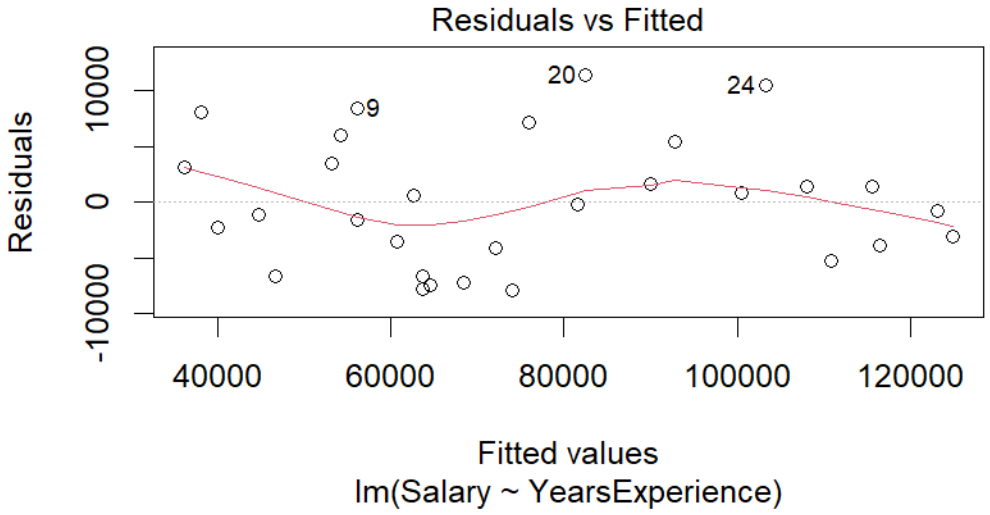
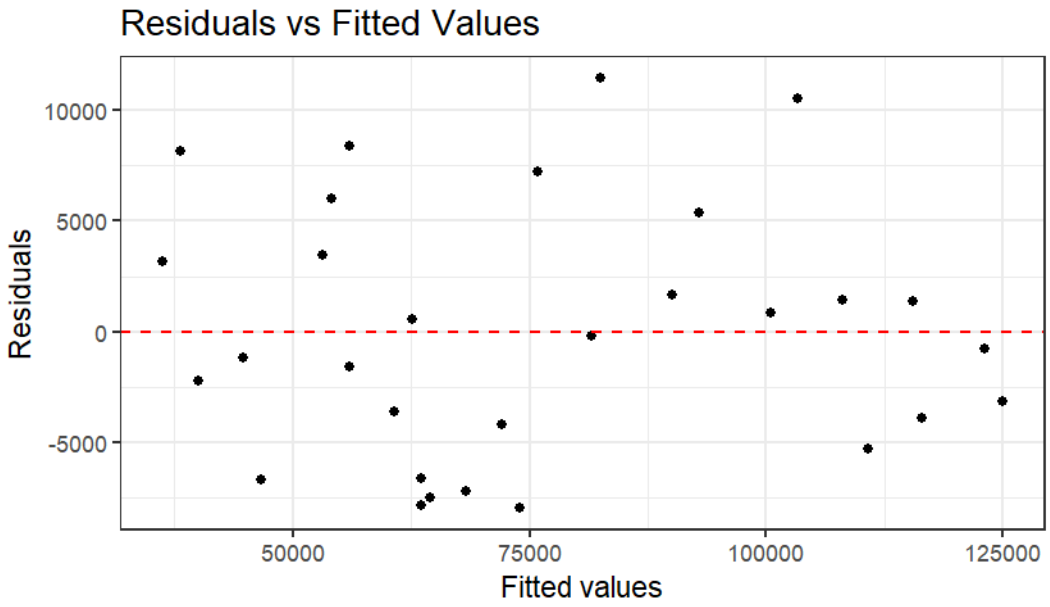
Visualize Represention of the Regression analysis

Assumptions:

1. **A linear relationship between the explanatory and outcome variable**



The residuals should be scattered randomly around the horizontal axis (which represents a residual value of 0). If the points are symmetrically distributed around a horizontal line without distinct patterns, that's a good sign of linearity. If you see a systematic pattern or a curve in the residuals, it suggests that the relationship between the predictor(s) and the response might be non-linear. For instance, a U-shaped or inverted U-shaped pattern often suggests missing polynomial terms (e.g., squared terms) in the model. The methods above can be subject to personal interpretation.



The correlation coeffcient can also be looked at. Again there is not definitive cut off for what is considered to be suffcient but in general if the correlation coeffcient is more than 0.3 (or less than -0.3 if negatively correlated) then there is moderate evidence of linearity.

A statistically rigorous method of establishing linearity is to use the Harvey-Collier Multiplier Test for Linearity. To do this we'll need to install and call a the lmtest package

Harvey-Collier test

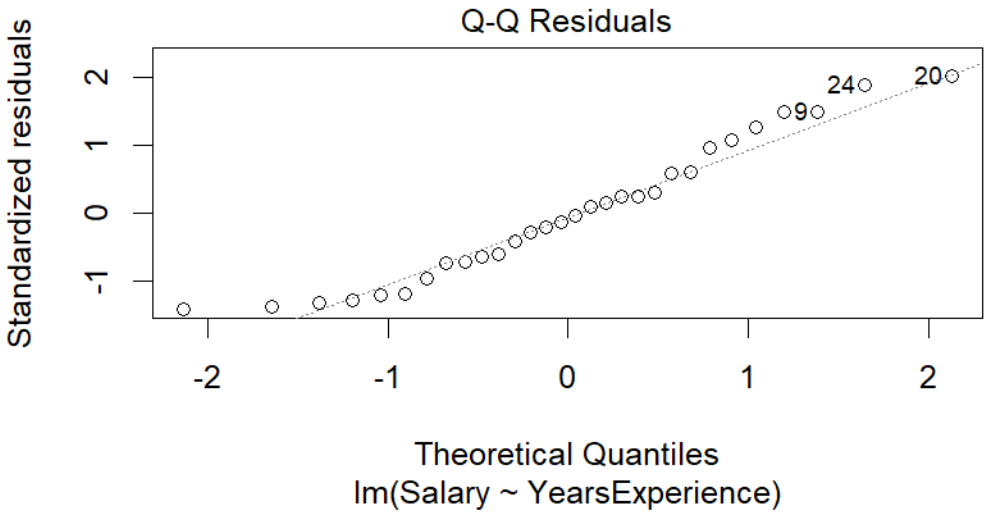
data: model

HC = 0.55664, df = 27, p-value = 0.5824

The interpretation of this test result is a little unusual. The null hypothesis is that there is a linear relationship. So a small *p* value (less than 0.05) would cause us to reject the assumption that the relationship between the two variables is linear. We're looking for a *p* value larger than 0.05 (as is the case in this model).

2. **The residuals follow a normal distribution**

Remember, the residuals are the distance between the actual and predicted values. If there are any major outliers in the data or an unusual relationship between the predictor and outcome variables in certain intervals (perhaps related to a third variable) then the distribution of residual values won't be normally distributed.



We can use the Shapiro-Wilk test as a formal test of normality. The null hypothesis of this test is that the data is normally distributed. If the *p*-value is less than a chosen alpha level (commonly 0.05), the null hypothesis is rejected, indicating that the data deviates from a normal distribution.

Shapiro-Wilk normality test

data: residuals(model)

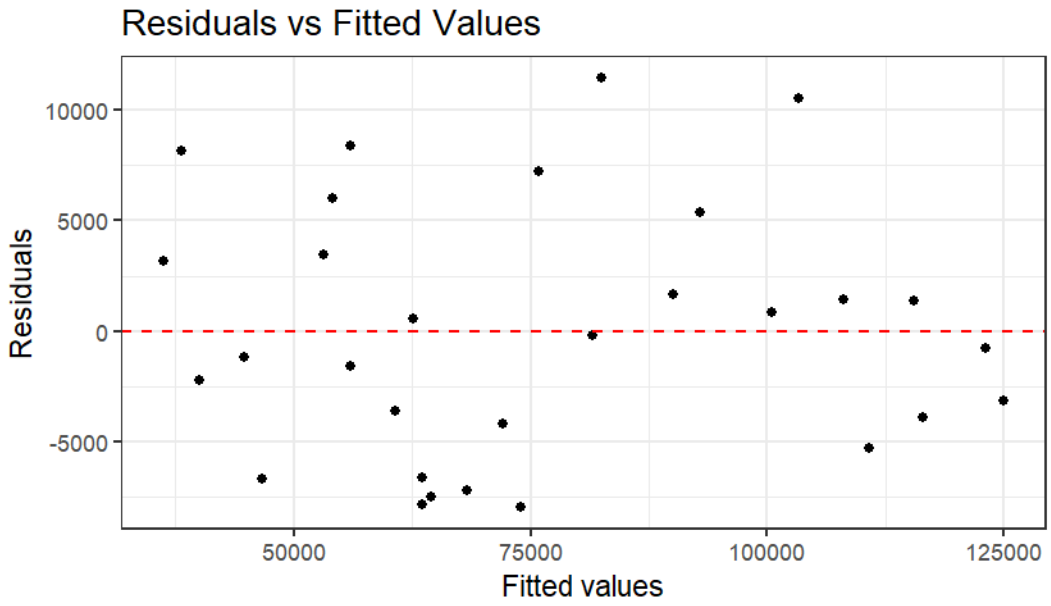
W = 0.95234, p-value = 0.1952

In this case the p value is large. We usually like small p values but not in this case. For the Shapiro-Wilk test a small p value suggests that the residuals are not normally distributed. A large p value (larger than the alpha value cut off of 0.05, for example) suggests that the residuals are normally distributed and our model assumption is met.

3. **Residuals are homoscedastic**

This refers to the assumption in regression analysis that the variance of the residuals is constant across all levels of the explanatory variables. To illustrate this, think about the case of the Salary dataset: we want the model residuals to vary from one point to next in a way that is consistent as we look at the graph from left to right (or across the values of the x-axis).

By plotting the fitted values against the residuals, one can look a pattern in the plot. This could be a funnel shape or clumping of the data points. This suggests that there is a problem and the residuals are not homoscedastic (the assumption is not met). In chase of the Salary dataset, we've already created this plot and as you can see below, the funnel shape doesn't



So that you know what it is that you're looking for (and can recognize the funnel shape, here is an example of the assumption not being met:

There are other reasons for seeing unusual patterns in the plot of residuals vs fitted values (like non-linearity) so it is important to look at the whole picture when interpreting these plots.

While visual confirmation might be sufficient, you can use formal statistical methods to test for heteroscedasticity. The Breusch-Pagan test (or the Cook-Weisberg test) can be used

studentized Breusch-Pagan test

data: model

BP = 0.39905, df = 1, p-value = 0.5276

The null hypothesis assumes homoscedasticity. In other words, you'd like to see a *p* value of more than 0.05 to support the inclusion of the variable in your model. In this case we're very happy with a p value of 0.52Assumption: residual values are independent.