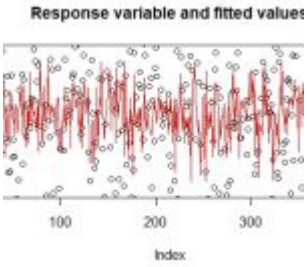
Regression Model Assumptions Tutorial

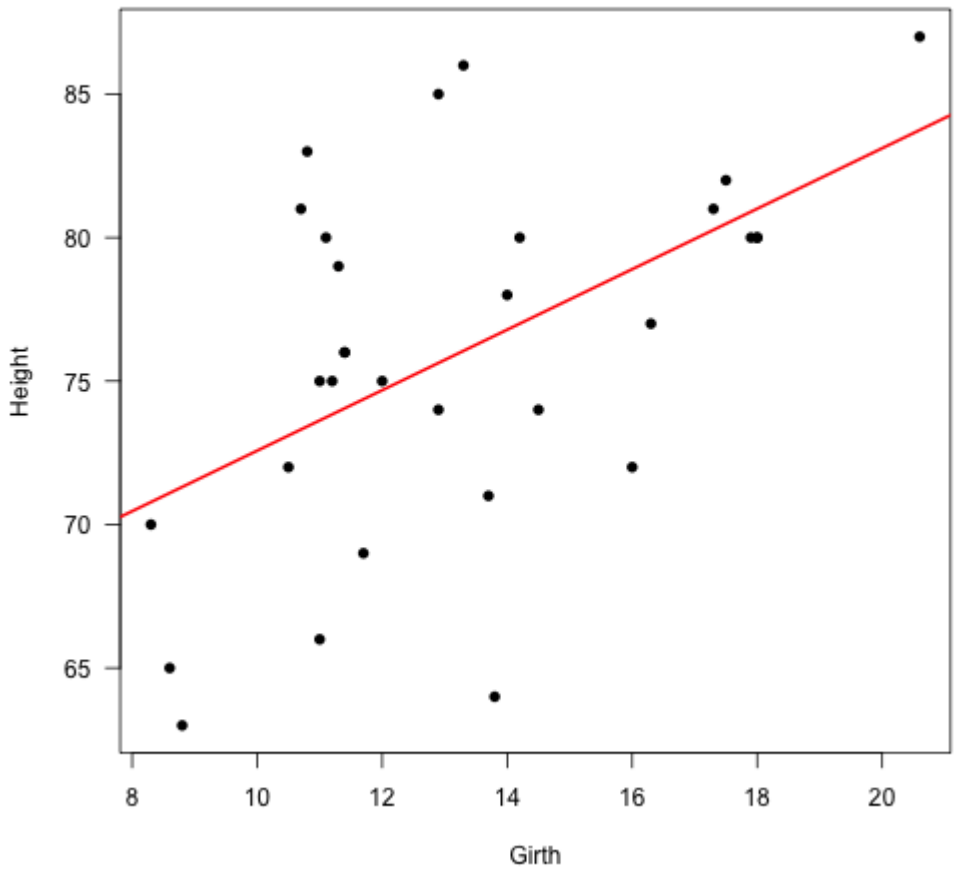


Regression is used to explore the relationship between one variable (often termed the response) and one or more other variables (termed explanatory). Several exercises are already available on simple linear regression or multiple regression. These are fantastic tools that are used frequently. However, each has a number of assumptions that need to be met. Unfortunately, people often conduct regression analyses without checking their assumptions.

In this tutorial, we will focus on how to check assumptions for simple linear regression.

We will use the *trees* data already found in R. The data includes the girth, height, and volume for 31 Black Cherry Trees. The following code loads the data and then creates a plot of volume versus girth. The red line is the line of best fit from linear regression.

```
data("trees")
attach(trees)
plot(Girth,Height,pch=16,las=1)
lm_model <- lm(Height~Girth)
abline(lm_model,lwd=2,col='red')
```



We can also use summary() to to obtain the model estimates from linear regression.

```
summary(lm_model)
```

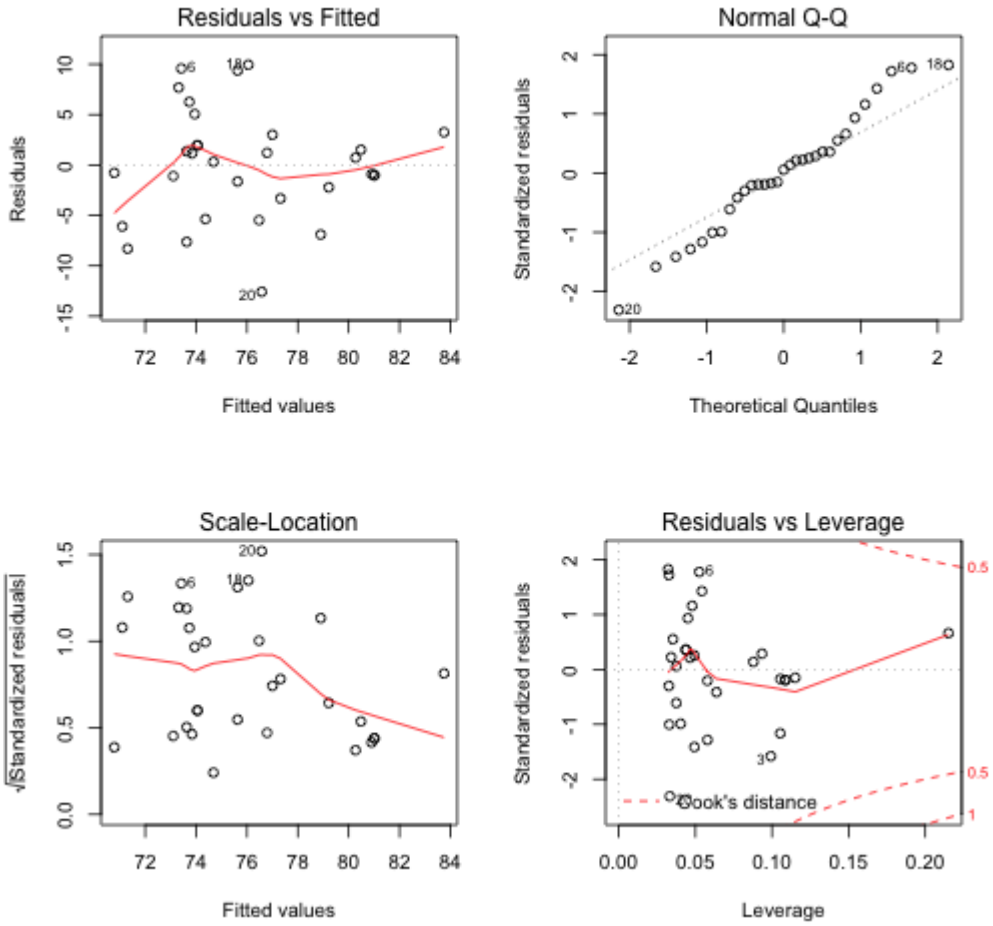
```
##
## Call:
## lm(formula = Height ~ Girth)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.5816  -2.7686   0.3163   2.4728   9.9456
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.0313     4.3833  14.152 1.49e-14 ***
## Girth         1.0544     0.3222   3.272  0.00276 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

```
' 1
##
## Residual standard error: 5.538 on 29 degrees of freedom
## Multiple R-squared:  0.2697, Adjusted R-squared:  0.2445
## F-statistic: 10.71 on 1 and 29 DF,  p-value: 0.002758
```

We obtain parameter estimates, a R-squared value, and other useful bits of information. Great. Using several graphs, let's examine if our model has met the assumptions of linear regression. We are going to examine model residuals. A residual is simply the difference between each observed value (black points in first graph) and the predicted value (the red line).

Learn more about evaluating different statistical models in the online courses Linear regression in R for Data Scientists and Structural equation modeling (SEM) with lavaan. These courses cover different statistical models that can help you choose the right design for your solution.
Thankfully, the plot() command provides several useful plots to check our model assumptions about residuals.

```
par(mfrow=c(2,2))
plot(lm_model)
```



Okay, there is a lot going on in these graphs. Let's break down each assumption and how they relate to these four graphs.

1. Homoscedasticity of residuals

The residuals should be homoscedastic. This means that the distribution of errors should be the same for all values of the explanatory variable. In the top left figure, we see residuals plotted against the fitted values. If the residuals are homoscedastic, we would expect the red line to fall on zero for the entire range of fitted values. There is some discrepancy, but the assumption looks to hold overall. The bottom left plot can also be used to examine homoscedasticity. Here, standardized residuals are used with the same reasoning that the red line should be horizontal. It is important to remember that examining model assumptions graphically is more of an art than science.

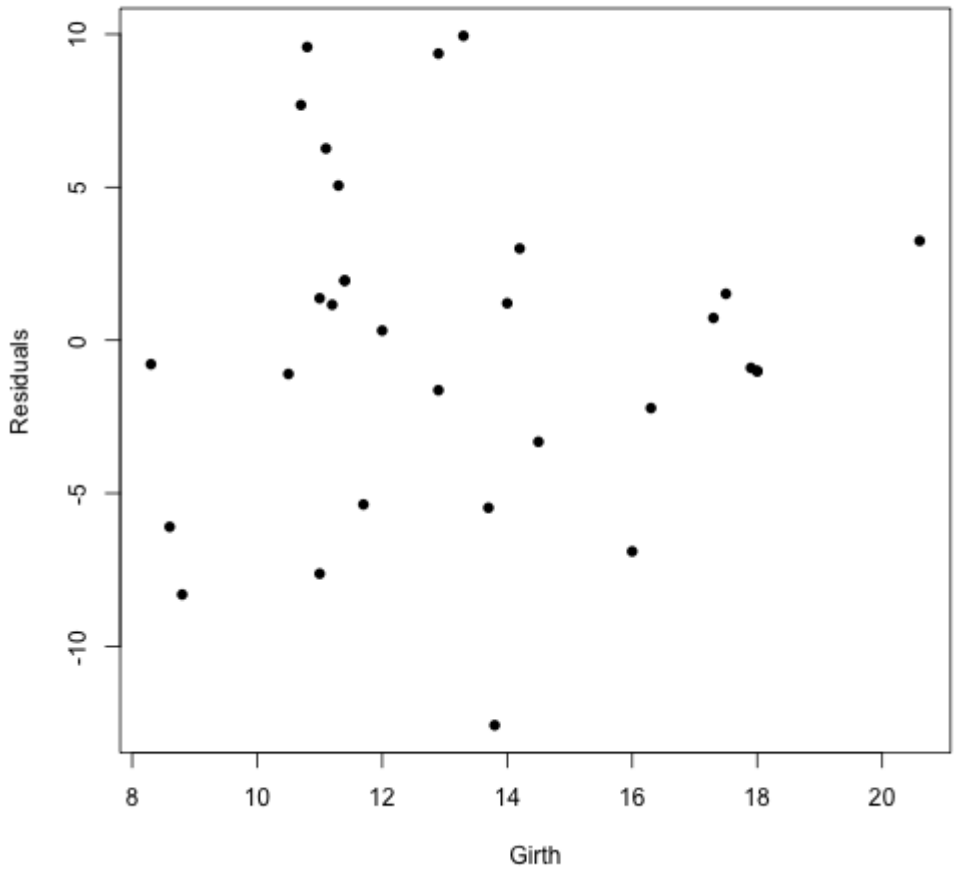
2. Normality of the residuals

The residuals should also be normally distributed. To check for normality, examine the qqnorm() plot on the top right of the 4-panel figure. Perfectly normal distributed data would all fall on the dashed line. This data looks fairly close to normal with some deviance at the tails.

3. Residuals should not be correlated with the explanatory variables

Another important assumption is that residuals should not be correlated with any explanatory variables. Previously, we let the plot() command handle the residuals, but we need the residuals themselves now. To extract model residuals we use lm_model$residuals. We can then use this command to plot the model residuals versus Girth, the explanatory variable.

```
par(mfrow=c(1,1))
plot(Girth, lm_model$residuals, ylab='Residuals',pch=16)
```



Here we see there is very little correlation between our residuals and the Girth variable. Therefore, this assumption is met.

4. The mean of the residuals should equal zero

We can see the mean of residuals is close to zero from either our plots, or simply using mean(lm_model$residuals). Our assumption is met as the mean is very close to zero.

5. Little or no multicollinearity

If we had more than one explanatory variable, we would have a multiple regression model. In that case, it would be important to verify that none of the explanatory variables are strongly correlated with one another. We only have one explanatory variable in our regression, so this assumption is not relevant here.