Linear regression

SSSA - Applied Statistics - Chiara Seghieri and Costanza Tortù

2023-10-03

Idea

Scenario:

Linear regression is a statistical method used in social sciences to model the relationship between a dependent variable and one or more independent variables. However, it comes with certain assumptions that need to be met for the results to be valid. Imagine you want to examine the relationship between hours spent studying (independent variable) and exam scores (dependent variable) through a \textbf{linear regression model]. Before interpreting your results you have to check whether the assumptions behind the model are met.

Recap

Here are the key assumptions of linear regression.

Linearity:

The relationship between the independent variable(s) and the dependent variable is assumed to be linear. This means that the change in the mean of the dependent variable is proportional to a change in one unit of the independent variable, holding other variables constant.

In the example above, the assumption of linearity implies that the change in exam scores is consistent for each additional hour of study.

Independence of Errors:

The errors (residuals), which are the differences between the observed and predicted values, should be independent of each other. In other words, the value of the error for one observation should not provide information about the value of the error for another observation.

If the residuals are not independent, it might suggest that there is some unaccounted factor influencing the dependent variable, violating the assumption. For instance, if the residuals from one student's exam score are related to the residuals of another student, this assumption may be violated.

Homoskedasticity:

The variance of the errors should be constant across all levels of the independent variable(s). This implies that the spread of the residuals should remain roughly the same as the values of the independent variable(s) change.

If the variance of the residuals increases or decreases as the level of study hours changes, homoscedasticity is violated. It's like saying that the variability in exam scores is not constant across different amounts of study time.

No Perfect Multicollinearity:

There should not be perfect linear relationships among the independent variables. Perfect multicollinearity occurs when one independent variable is a perfect linear function of another, leading to difficulties in estimating the individual coefficients.

If you are studying the factors affecting job performance and you have both hours worked per week and income as independent variables, perfect multicollinearity would occur if one variable is always a fixed multiple of the other.

No Endogeneity:

The independent variables are assumed to be exogenous, meaning they are not influenced by the errors. Endogeneity arises when there is a two-way causation between the independent and dependent variables.

Normality of Errors:

The errors are assumed to be normally distributed. This is important for making statistical inferences, such as hypothesis testing and constructing confidence intervals. However, this assumption is less critical for large sample sizes due to the Central Limit Theorem.

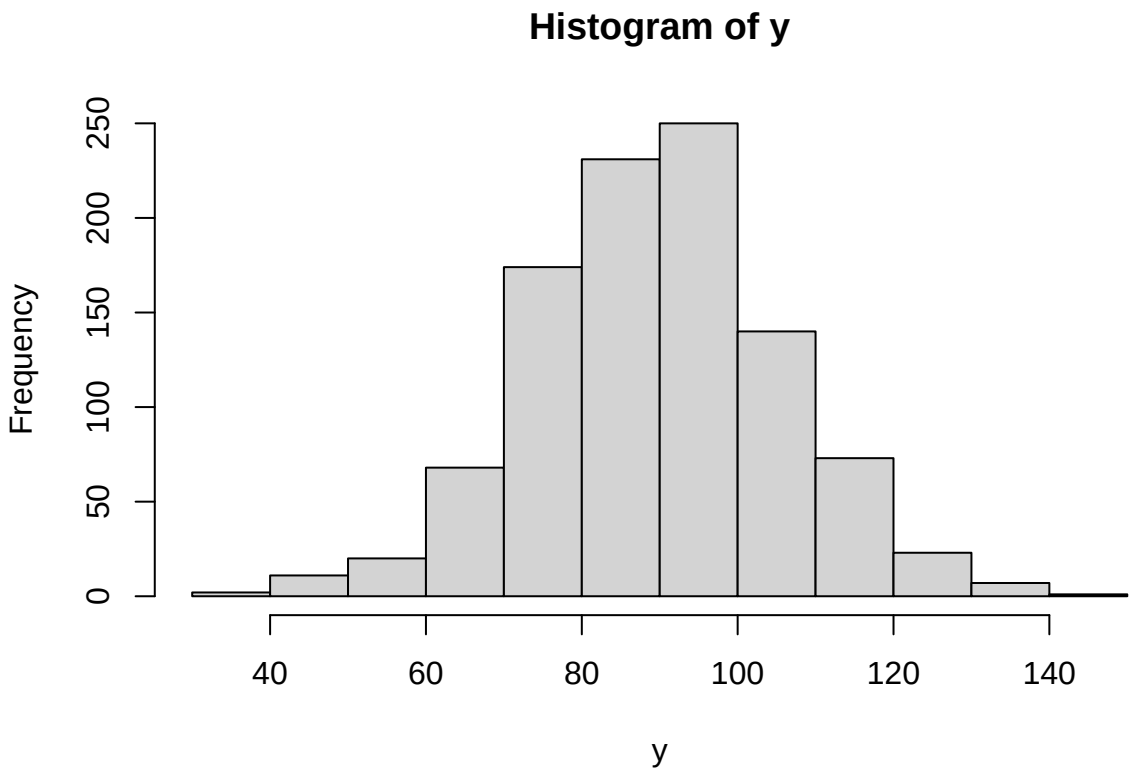
If the residuals are not normally distributed, it might indicate that there are other factors influencing the dependent variable that are not accounted for in the model.

Get the idea through simulations

Simulate your data

```
set.seed(1234)

N=1000 #number of observations
x=rnorm(N,20,4)
error=rnorm(N,0,2)
b0 = 10
b1 = 4
y= b0 + b1 * x + error
hist(y)
```



Estimate your model

```
model = lm(y~x)
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.332 -1.288  0.029  1.307  6.137
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.47485    0.31542   30.04   <2e-16 ***
## x            4.02786    0.01555  259.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.96 on 998 degrees of freedom
## Multiple R-squared:  0.9854, Adjusted R-squared:  0.9853
## F-statistic: 6.713e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```

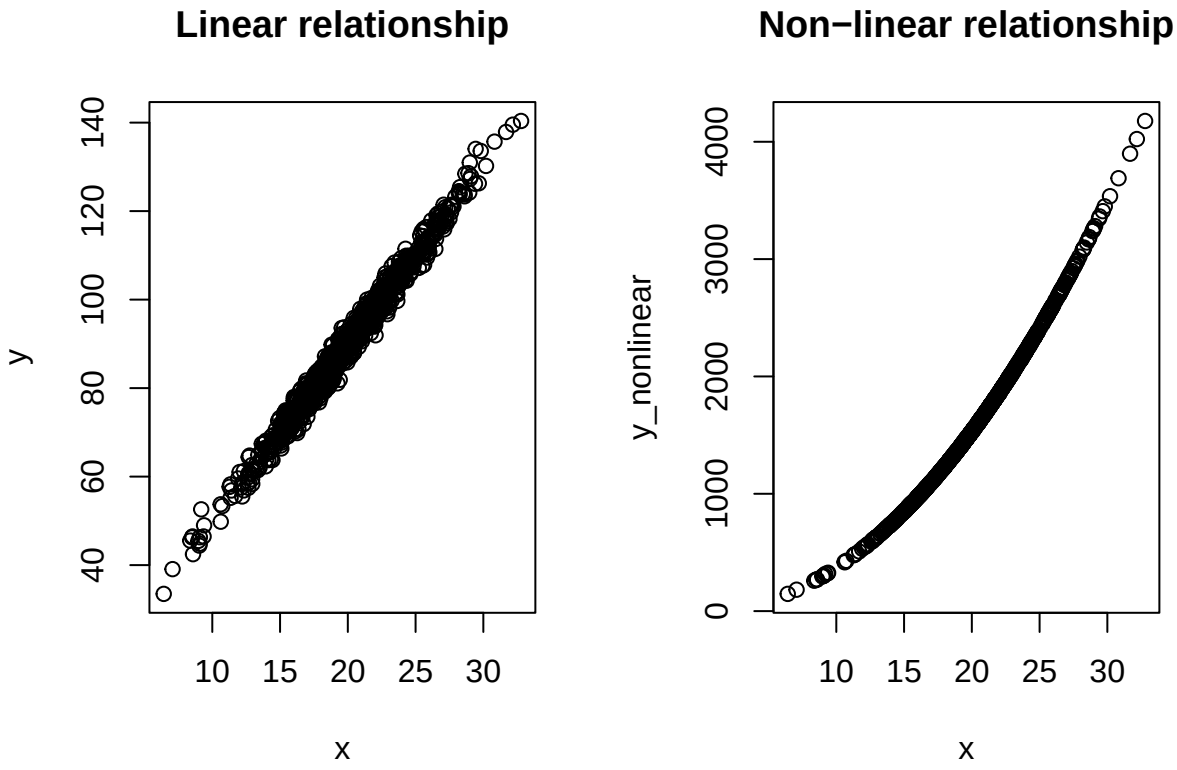
```
res = residuals(model)
```

Check the assumptions

Linearity

```
b2 = 4
y_nonlinear =  b0 - b1*x  + b2 * (x^2)   + error

par(mfrow=c(1,2))
plot(x,y, main = "Linear relationship")
plot(x,y_nonlinear, main = "Non-linear relationship")
```



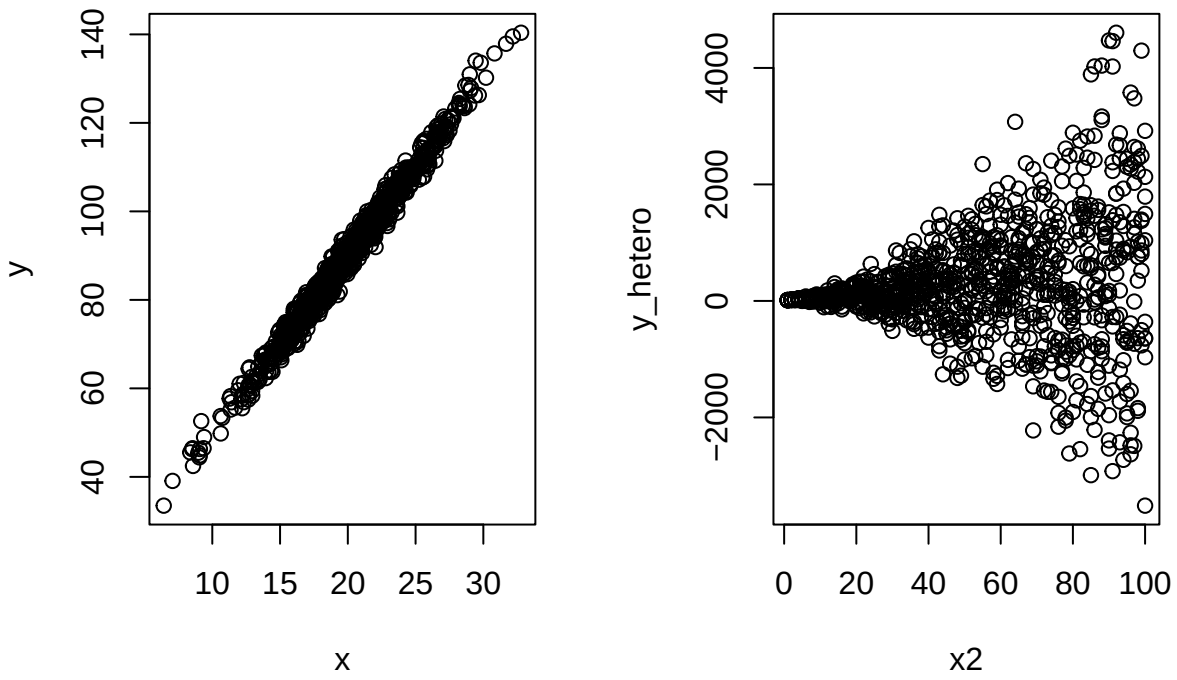
Homoskedasticity

```
x2      = rep(1:100,10)

sigma2_hetero =  x2^3.3
error_hetero  = rnorm(N,mean=0,sd=sqrt(sigma2_hetero))
y_hetero      = b0 + b1*x2 + error_hetero
model_hetero  = lm(y_hetero ~ x2)
res_hetero    = residuals(model_hetero)


par(mfrow=c(1,2))
```

```
plot(x,y)
plot(x2,y_hetero)
```



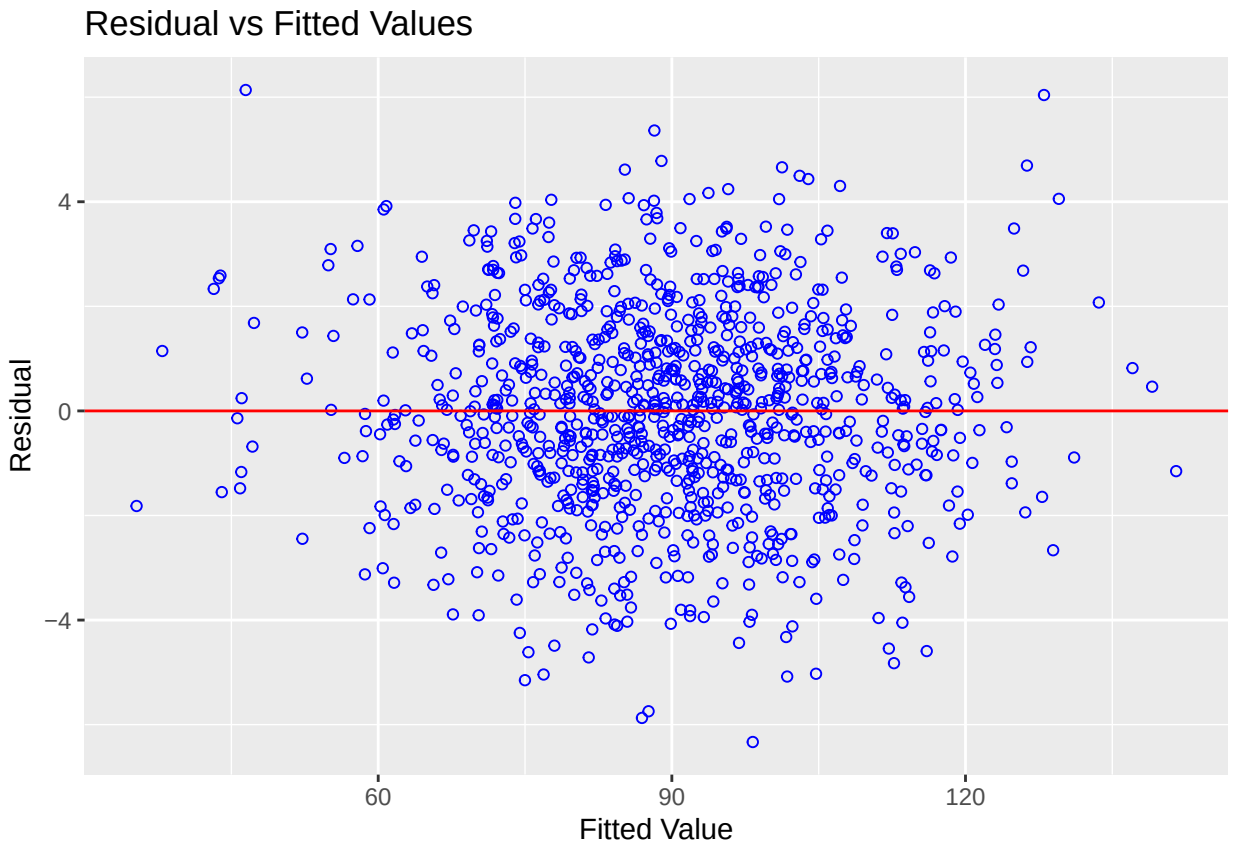
Normality

```
library("olsrr")
```

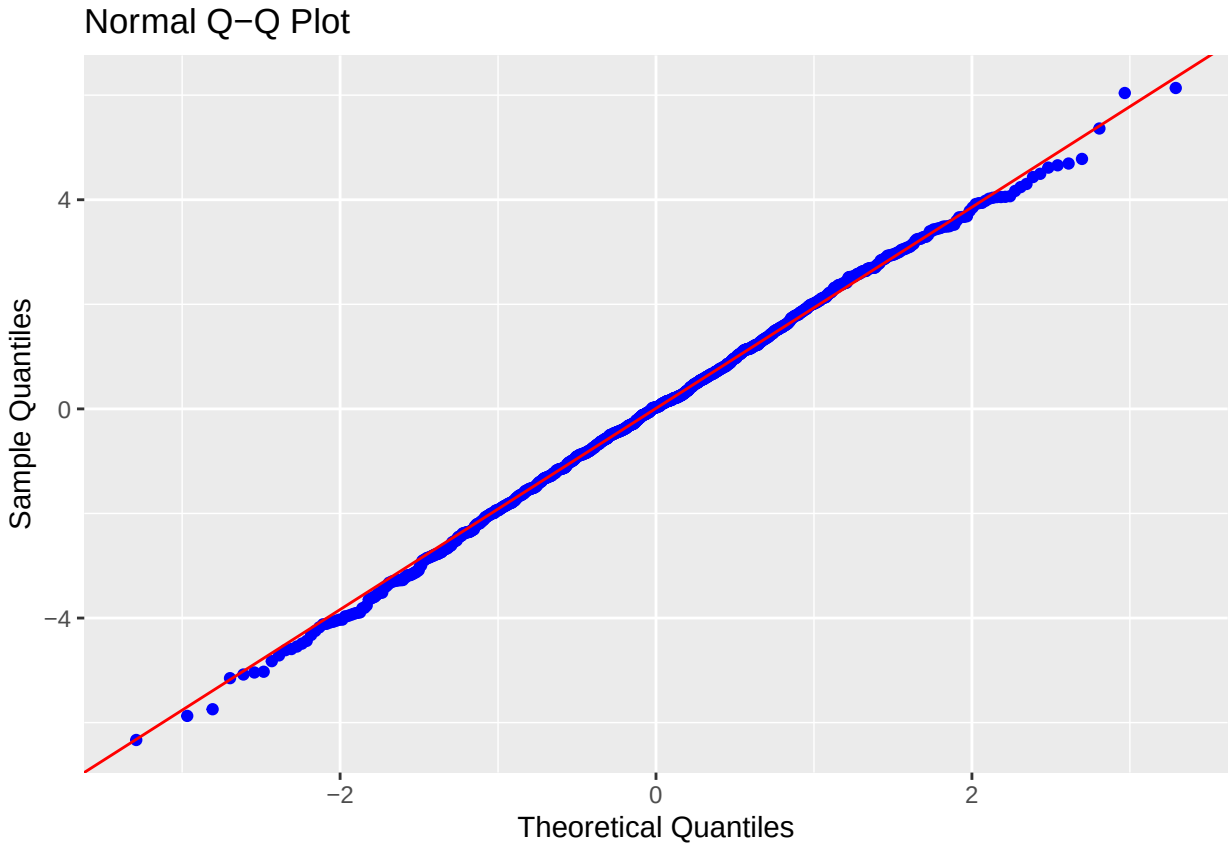
```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
par(mfrow=c(1,2))
ols_plot_resid_fit(model)
```




```
ols_plot_resid_qq(model)
```


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
4
0
−4
−2
0
2