Regression Analysis
Chapter 8 and 9: Transformations and Regression Diagnostics

Shaobo Jin

Department of Mathematics

Diagnostics

Regression diagnostics are used after model fitting to check if a fitted mean function and assumptions are consistent with observed data.

Recall that the residuals are

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\,\boldsymbol{y}.$$

- If OLS is used to estimate $\boldsymbol{\beta}$, then

$$\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T.$$

- If WLS is used to estimate $\boldsymbol{\beta}$, then

$$\boldsymbol{H} = \boldsymbol{W}^{1/2}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}^{1/2}.$$

Properties of Hat Matrix

$$\text{Hat matrix} \quad \boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T.$$

- The hat matrix is symmetric and idempotent. Hence it is an orthogonal projection matrix to the column space of $\boldsymbol{X}$.
  - Let $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{X}_1 & \boldsymbol{X}_2\end{bmatrix}$ and $\boldsymbol{H}_1 = \boldsymbol{X}_1\left(\boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1^T$. Then,

$$\boldsymbol{H}\boldsymbol{H}_1\boldsymbol{y} = \boldsymbol{H}_1\boldsymbol{y}.$$

- The hat matrix is not full rank: $\operatorname{rank}(\boldsymbol{H}) = \operatorname{tr}(\boldsymbol{H}) = p$, if $\boldsymbol{X}$ is a $n \times p$ matrix.
- The diagonal entries satisfy $h_{ii} \leq 1$ for all i.
  - If the intercept is included, then we also have $1/n \leq h_{ii} \leq 1$.

Leverage

Suppose that an intercept is included in the model. Then,

$$\sum_{i=1}^{n} h_{ij} = 1, \qquad \sum_{j=1}^{n} h_{ij} = 1.$$

The predicted value is $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$. If h_{ii} is close to 1, then

1. the predicted value tends to be close to y_i, i.e.,

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j$$

2. and $\operatorname{Var}(\hat{e}_i \mid \boldsymbol{X}) = \sigma^2 (1 - h_{ii})$ is close to 0.

For this reason, h_{ii} is called the leverage of the ith observation.

Leverage on $\hat{\boldsymbol{\beta}}$

Let $\boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \boldsymbol{X}_{-1} \end{bmatrix}$ and $\bar{\boldsymbol{x}}_{-1} = \frac{1}{n}\boldsymbol{X}_{-1}^T\mathbf{1}$. Let

$$\mathcal{X} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{1,p-1} - \bar{x}_{p-1} \\ \vdots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n,p-1} - \bar{x}_{p-1} \end{bmatrix}$$

be the demeaned data matrix. It can be shown that

$$h_{ii} = \frac{1}{n} + \left(\boldsymbol{x}_i^* - \bar{\boldsymbol{x}}_{-1}\right)^T \left(\mathcal{X}^T\mathcal{X}\right)^{-1} \left(\boldsymbol{x}_i^* - \bar{\boldsymbol{x}}_{-1}\right),$$

where $\boldsymbol{x}_i^T = \begin{bmatrix} 1 & (\boldsymbol{x}_i^*)^T \end{bmatrix}$. Hence, if h_{ii} is close to 1, it means that $\boldsymbol{x}_i^*$ is far from the average $\bar{\boldsymbol{x}}_{-1}$ and the influence of the ith observation to the estimation of our regression model is large.

Different Residuals

Suppose that the estimator is the minimizer of

$$\mathrm{RSS}\left(\boldsymbol{\beta}\right) = \sum_{i=1}^{n} w_i \left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}\right)^2 .$$

1. The residual is $\hat{e}_i = y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$.
2. The Pearson residual is $\sqrt{w_i}\left(y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}\right)$.
   - The sum of squared Pearson residuals is the same as $\mathrm{RSS}\left(\hat{\boldsymbol{\beta}}\right)$.
3. Under some assumptions, we can show that

   $$\mathrm{Var}\left(\hat{\boldsymbol{e}} \mid \boldsymbol{X}\right) = \sigma^2 \boldsymbol{W}^{-1/2}\left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{W}^{-1/2}.$$

   A standardized residual is

   $$\frac{\hat{e}_i}{\hat{\sigma}\sqrt{\left(1 - h_{ii}\right)/w_{ii}}}.$$

   Its variance is closer to 1 than $\hat{e}_i$.

Properties of Residual

Suppose that the intercept is included in our model and OLS is used. Then, regardless of whether our is model is correct or not,

1. we always have $\sum_i \hat{e}_i = 0$.
2. the sample correlation between residual and regressors is zero.
3. the sample correlation between residual and the fitted value is zero.

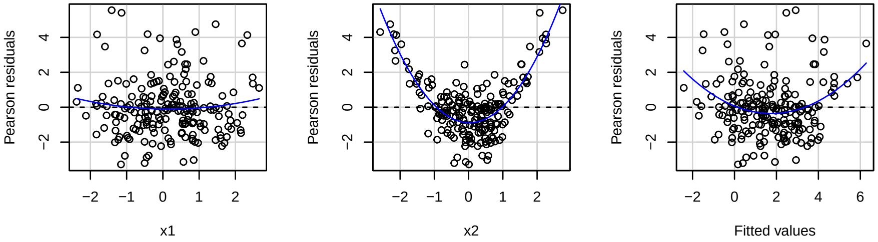
However, if the model is misspecified, we may still observe patterns. We can plot residuals to see whether something has gone wrong.

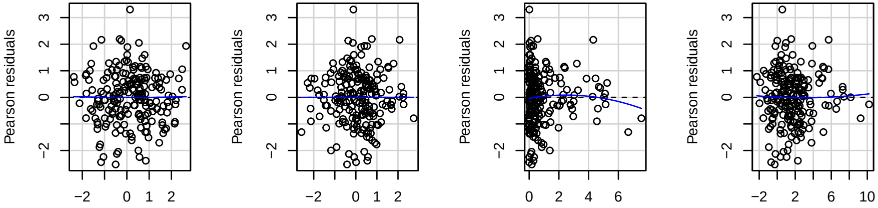
Detect Lack of Fit Using Residuals (1)

The residuals should be plotted against each of the explanatory variables of the model, and against the fitted value.

- If the model fits the data well, no systematic patterns can be observed.
- A systematic pattern suggests another model form or additional terms.

Detect Lack of Fit Using Residuals (1)





Normality Assumption

We often assume $\boldsymbol{e}$ is normally distributed.

1. If $\boldsymbol{e} \mid \boldsymbol{X}$ is normally distributed, then

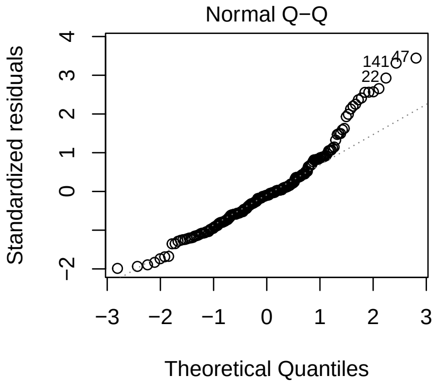
$$\hat{\boldsymbol{e}} \mid \boldsymbol{X} \quad \sim \quad N\left(\mathbf{0},\ \sigma^2\left(\boldsymbol{I} - \boldsymbol{H}\right)\right).$$

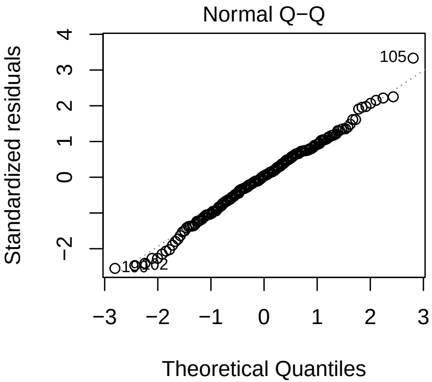
   We can test whether the standardized residuals

Detect Lack of Fit Using Residuals (2)

QQ plot: plot the sample quantiles of the standardized residuals against the expected quantiles from a assumed distribution.

- Departures from the straight line indicate departures from the assumed distribution.





Studentized Residual

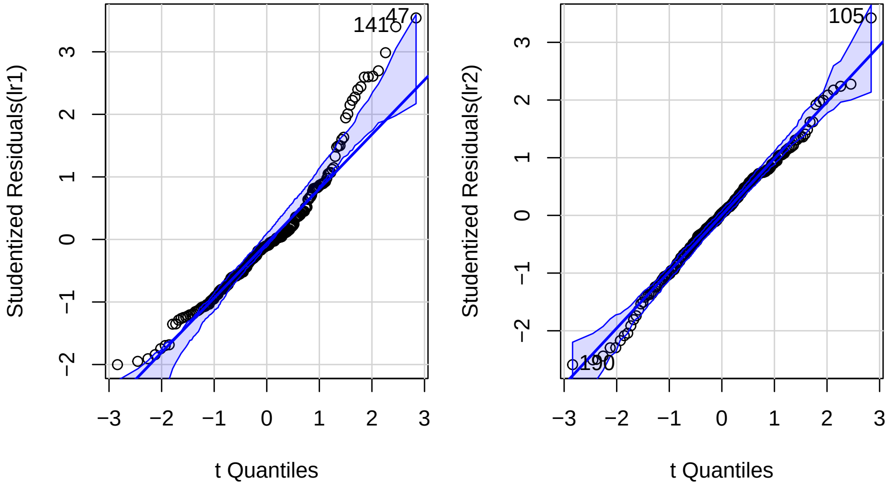
Let $\hat{\boldsymbol{\beta}}$ be the WLS estimator based on the entire data set. A studentized residual is

$$\frac{\sqrt{w_i}\left(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}\right)}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}},$$

where $\sqrt{w_i}\left(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}\right)$ is the Pearson residual, and h_{ii} is the (i,i)th entry of $\boldsymbol{H}$.

- Here, $\hat{\sigma}^2_{(i)}$ is an estimator of σ^2, but not from the model based on the entire data set.
- Instead, we delete the ith observation and refit the model. $\hat{\sigma}^2_{(i)}$ is the estimator of σ^2 from the refitted model.

More QQ Plot



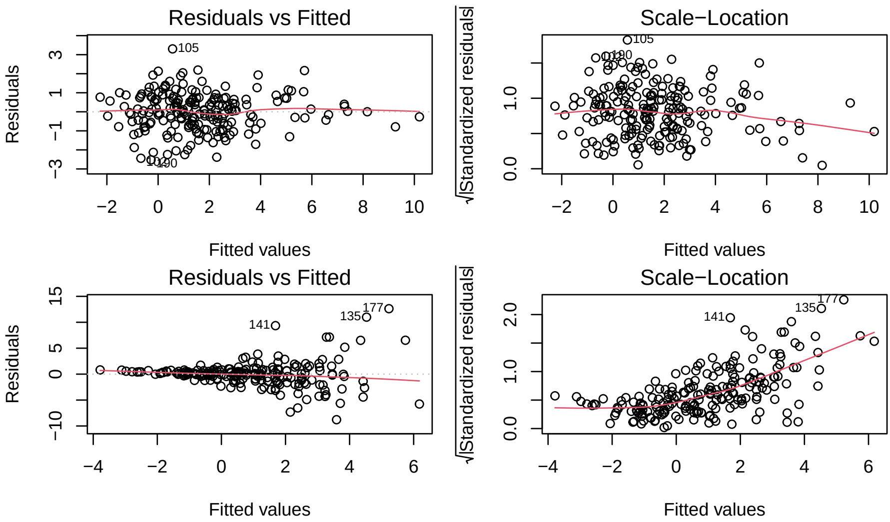
Studentized Residuals(lr1)
t Quantiles
47
141
Studentized Residuals(lr2)
t Quantiles
105
190

Detect Lack of Fit Using Residuals (3)

The residuals should also be plotted against the fitted values to detect changes in variance.

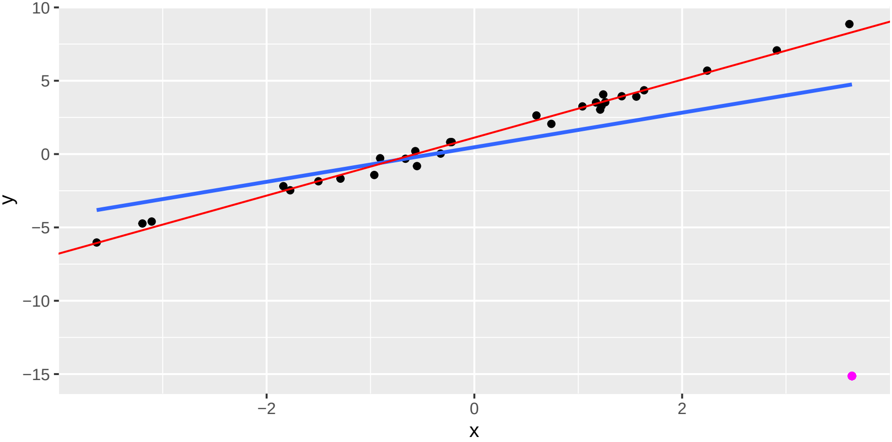
- If a systematic pattern is observed, the variance of standardized residuals is a function of fitted values, not homoskedasticity.
- Some plots do not show systematic deviation from zero but the variation depends on $\hat{y}$.

Detect Lack of Fit Using Residuals (3)



Influential Point or Outlier

An influential point or an outlier is a point that differs greatly from any other points.



Influence Analysis

The general idea of influence analysis is to study changes in a specific part of the analysis when the data are perturbed.

1. The studentized residual is an example of influence analysis.
2. The Cook's distance is defined as

$$D_i = \frac{\left(\boldsymbol{X}\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^T \left(\boldsymbol{X}\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)}{p\hat{\sigma}^2},$$

   where $\hat{\boldsymbol{\beta}}$ is the estimator with the whole data set and $\hat{\boldsymbol{\beta}}_{(i)}$ is the estimator after deleting the ith observation.

   - If the ith observation has a substantial influence, then we expect $\hat{\boldsymbol{\beta}}$ differ much from $\hat{\boldsymbol{\beta}}_{(i)}$.
   - A rule-of-thumb is that if $D_i > 1$, then it deserves some consideration.

Cook's Distance and Leverage

In fact, we do not need to refit the model n times. It can be shown that

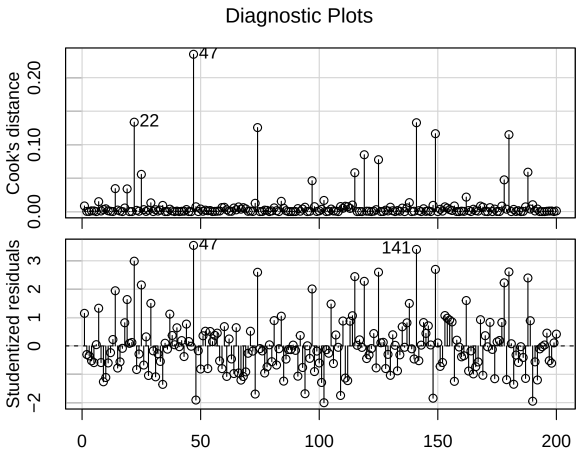
$$D_i = \frac{1}{p} r_i^2 \frac{h_{ii}}{1 - h_{ii}},$$

where r_i is the standardized residual, and h_{ii} is the leverage.

- If the leverage is close to 1, we probably have a large Cook's distance.

Detect Lack of Fit Using Residuals (4)

Use the Cook's distance and other tools to detect influential observations/outliers, and other plots.



DFBETAS

Another quantity to measure the influence of an observation is

$$\text{DFBETAS}_j = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\hat{\sigma}_{(i)} \sqrt{\left[\left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \right)^{-1} \right]_{jj}}},$$

where $\hat{\beta}_{j(i)}$ is the estimator of β_j without the ith observation, and $\hat{\sigma}^2_{(i)}$ is an estimator of σ^2 using the data set without ith observation.

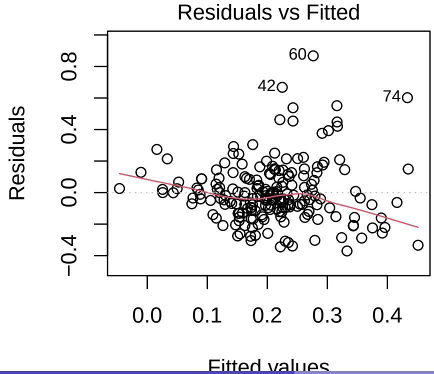
It measures how much the coefficient chnages when the ith observation is deleted.

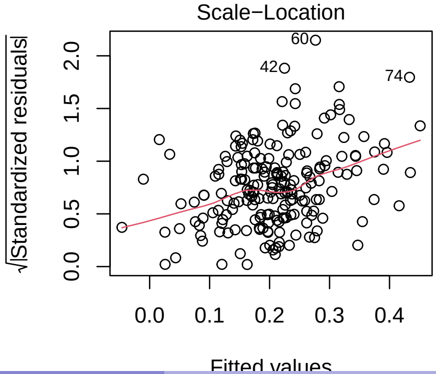
OBS!

But keep in mind that the residual plots only suggest that something is wrong but never tell what is definitely wrong. We generate data from

$$\mathrm{E}\left(Y \mid \boldsymbol{x}\right) = \frac{|x_1|}{2 + (1.5 + x_2)^2}$$

but fit a linear model $\mathrm{E}\left(Y \mid \boldsymbol{x}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.





Test Outlier: T-Test

Suppose that the ith observation is an outlier.

- The mean function for all other points are $\mathrm{E}\left(Y_j \mid \boldsymbol{x}_j\right) = \boldsymbol{x}_j^T \boldsymbol{\beta}$, $j \neq i$.
- The mean function for the outlier is $\mathrm{E}\left(Y_i \mid \boldsymbol{x}_i\right) = \boldsymbol{x}_i^T \boldsymbol{\beta} + \delta$ for some δ.

Now we suspect the ith observation to be an outlier.

1. Define a variable U that has 1 for the ith observation and 0 for all other elements.
2. We can test H_0: $\delta = 0$ versus H_1: $\delta \neq 0$ using a t-test if we assume $\mathrm{Var}\left(Y \mid \boldsymbol{x}\right) = \sigma^2$ and data are normally distributed.

Test Outlier: Leave-One-Out

Another approach works as follows.

1. Delete the ith observation from the data. Denote the remaining design matrix by $\boldsymbol{X}_{(i)}$.
2. Estimate $\boldsymbol{\beta}$ and σ^2 using the remaining data. Denote the estimator by $\hat{\boldsymbol{\beta}}_{(i)}$ and $\hat{\sigma}^2_{(i)}$.
3. The predicted value of y_i is $\hat{y}_{i(i)} = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(i)}$, which is independent of y_i if we assume observations are mutually independent.
4. Under the assumptions that $\operatorname{Var}(Y \mid \boldsymbol{x}) = \sigma^2$, we can obtain

$$\operatorname{Var}\left(y_i - \hat{y}_{i(i)} \mid \boldsymbol{X}\right) = \sigma^2 + \sigma^2 \boldsymbol{x}_i^T \left(\boldsymbol{X}_{(i)}^T \boldsymbol{X}_{(i)}\right)^{-1} \boldsymbol{x}_i.$$

5. Test $\operatorname{E}\left(y_i - \hat{y}_{i(i)} \mid \boldsymbol{X}\right)$ using a t-test under the normality assumption.

It turns out that this procedure is equivalent to the t-test using regression.

Added Variable Plot

Suppose that we have fitted a model $\hat{y} = \boldsymbol{x}^T \hat{\boldsymbol{\beta}}$ using the variables in $\boldsymbol{x}$, but we have an extra variable Z. An added variable plot or partial regression plot can reveal whether we can use Z to improve the model.

- Let $\hat{\boldsymbol{e}}\left(Y \mid \boldsymbol{x}\right)$ be the residual vector when regress Y on $\boldsymbol{x}$ (part of Y not explained by $\boldsymbol{x}$).
- Let $\hat{\boldsymbol{e}}\left(Z \mid \boldsymbol{x}\right)$ be the residual vector when regress Z on $\boldsymbol{x}$ (part of Z not explained by $\boldsymbol{x}$).
- We regress $\hat{\boldsymbol{e}}\left(Y \mid \boldsymbol{x}\right)$ on $\hat{\boldsymbol{e}}\left(Z \mid \boldsymbol{x}\right)$ and check whether part of Y not explained by $\boldsymbol{x}$ can be explained by part of Z not explained by $\boldsymbol{x}$.

Illustration 1

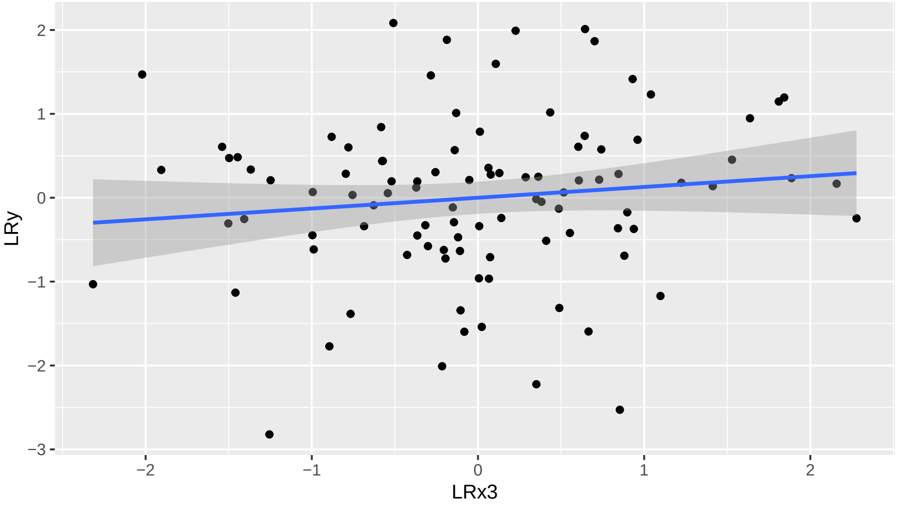
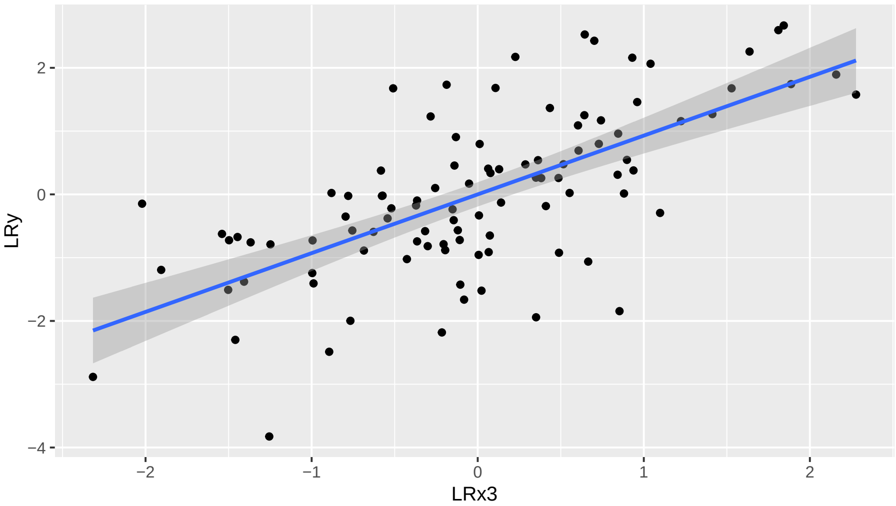
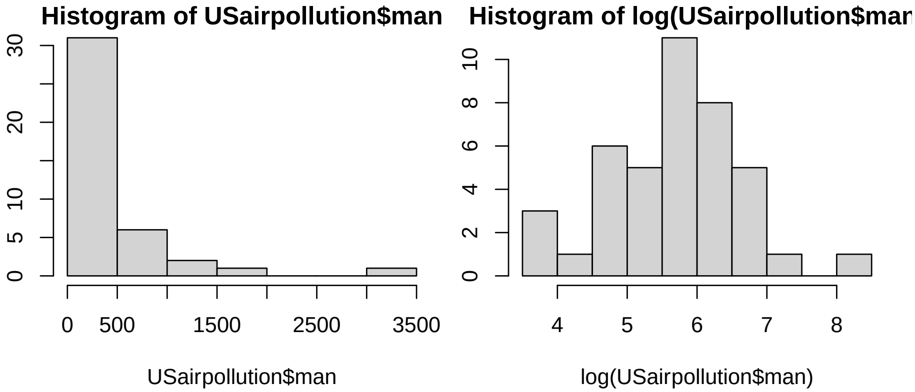


Illustration 2



Purpose of Transformation

The purpose of transformation (of the response variable and the regressors) is to make our assumptions more plausible.



Box-Cox Transformation

The Box-Cox transformation is often used to transform the response variable so that the normal assumption is more plausible:

$$\psi\left(y,\lambda\right) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0, \\ \log\left(y\right), & \text{if } \lambda = 0. \end{cases}$$

In case y is not positive, we have two-parameter Box-Cox transformation.

$$\psi\left(y,\lambda,\varepsilon\right) = \begin{cases} \frac{\left(y+\varepsilon\right)^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0, \\ \log\left(y+\varepsilon\right), & \text{if } \lambda = 0. \end{cases}$$

Which λ to Choose

The idea is to find the value of λ such that the residual from our regression model is most normal.

- The regression model is

$$\psi\left(y, \lambda\right) = \boldsymbol{x}^T \boldsymbol{\beta} + e.$$

- With normal error, $\psi\left(y, \lambda\right) - \boldsymbol{x}^T \boldsymbol{\beta}$ should be normal.
- Find the λ value that maximizes the normal likelihood.

Bootstrap

Another approach when the assumptions are violated is the bootstrap that can be used to handle

1. residuals are not normally distributed,
2. homoscedasticity is violated.

method = "case" eller method = "residual"?

Boot {car} R Documentation

Bootstrapping for regression models

Description

This function provides a simple front-end to the `boot` function in the **boot** package that is tailored to bootstrapping based on regression models. Whereas `boot` is very general and therefore has many arguments, the `Boot` function has very few arguments.

Usage

```
Boot(object, f=coef, labels=names(f(object)), R=999,
  method=c("case", "residual"), ncores=1, ...)
```

method = "case" eller method = “residual”?

method = “residual” requires

- homoscedasticity
- Y is random, but $\boldsymbol{x}$ is not random
  - For example, we want to compare the caffeine content of Lindvalls Mörkrost och Lindvalls Brygg.

method = "case"

- does not require homoscedasticity
- requires both Y and $\boldsymbol{x}$ are random.