STA302 - Lecture 5

Cedric Beaulac

May 23, 2019

Introduction

Today's plan

- Today we conclude (almost) the simple linear regression model
  - Review of the model
  - Perform predictive inference
  - Diagnostic : checking the model assumptions

Review of the model

Review of the model

- Last week, we have established this basic model: $y_i = \beta_0 + \beta_1 x_i + e_i$, where e_i are independantly distributed $\sim N(0, \sigma^2)$.
- $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$ which implies $\mathbf{y}|\mathbf{X} \sim N(\mathbf{X}\beta, I\sigma^2)$.
- We estimate β with $\hat{\beta}_{MLE} = (\mathbf{X}^T\mathbf{X})^1\mathbf{X}^T\mathbf{y}$
- It allowed us to establish distributions for $\hat{\beta}$:

$$\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

Review of the model

$$\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

- It is interesting to use our statistical model to understand the relationship between **x** and **y** (β_1).
- It is also interesting to be able to assess our degree of certainty regarding the possible values of β_1.
- Could β_1 be 0 ? Could **x** provides no information on **y** ?
- It seems important being able to say things about the response as well.

Predictive inference

What about the response ?

- The response is the most important variable for us. (usually)
- In the test we have established the distribution for the fitted values $\hat{y} = \mathbf{X}\hat{\beta}$:

$$\hat{y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)$$

- That's good!

Prediction (predictive inference)

- Something important in data analysis is **prediction**.
- I would even say that it is central in supervised learning.
- Can we use the knowledge we have acquired from the data set to predict the response for a new, unobserved predictor x^* ?
- A simple prediction for the response could be : $y^* = \beta_0 + \beta_1 x^*$

Prediction (predictive inference)

- We don't know the exact values for β_0 nor β_1 but we have $\hat{\beta} = [\hat{\beta}_0, \hat{\beta}_1]^T$. Let's use it!
- A simple prediction : $\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$
- How accurate is this prediction ?
- Let us establish the distribution of $\hat{y}^*$ and build prediction intervals.
- Given a vector of new observation $\mathbf{x}^*$, we can create our matrix of predictors as usual by adding a comlumn of 1s to get $\mathbf{X}^*$.
- The vector of predictions is $\widehat{\mathbf{y}}^* = \mathbf{X}^* \hat{\beta}$

Prediction (predictive inference)

- $\hat{\beta}|\mathbf{X} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$ still holds.
- $\mathbf{y}^* \sim N(\mu, \Sigma)$, where

$$\mu = \mathbf{X}^*\beta$$
$$\Sigma = \sigma^2\mathbf{X}^*(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^{*T}$$

- Using this, we can build confidence interval for $\mathbf{E}(\widehat{\mathbf{y}}^*) = \mathbf{X}^*\beta$.

Prediction : Confidence interval

$$\frac{\widehat{\mathbf{y}}^* - \mathbf{X}^*\beta}{\sigma\sqrt{\mathbf{X}^*(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^{*T}}} \sim N(0, I)$$

$$\Rightarrow P(-1.96 < \frac{\hat{y}_i^* - \mathbf{X}^*\beta}{\sigma\sqrt{(\mathbf{X}^*(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^{*T})_{i,i}}} < 1.96) = 0.95$$

$$\Rightarrow CI : \hat{y}_i^* + / - 1.96 * \sigma\sqrt{(\mathbf{X}^*(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^{*T})_{i,i}}$$

- Notice that this is a confidence interval for $\mathbf{X}^*\beta$.

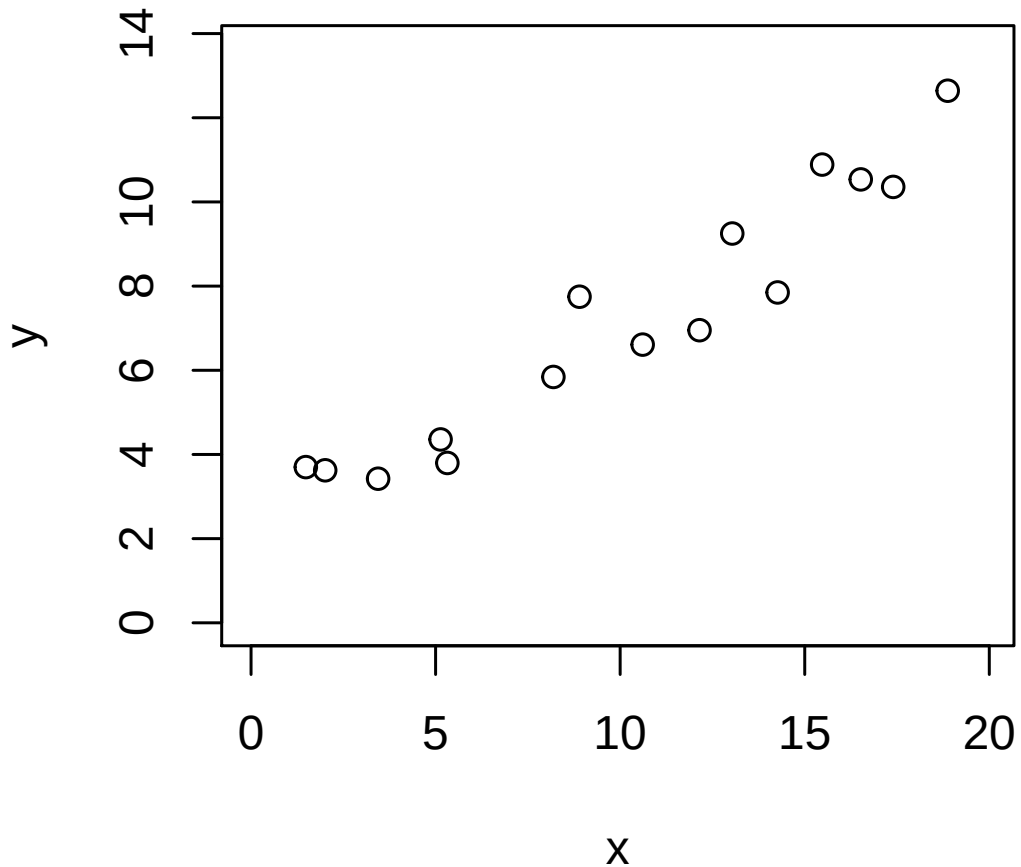
Prediction : Confidence interval

- If σ is unknown, we can estimate it (with s) and use a student distribution!

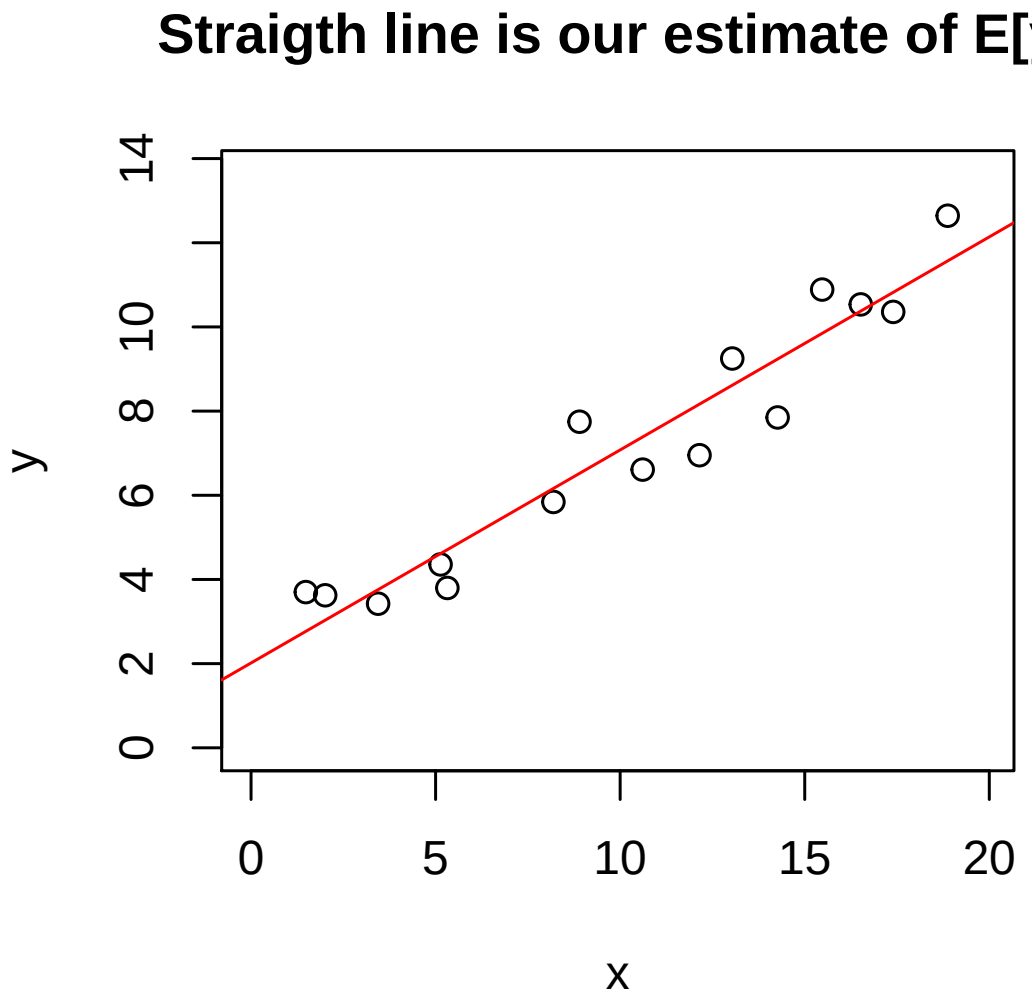
$$\Rightarrow CI : \hat{y}_i^* + / - t_{(n-2)}^{(0.025)} * \hat{\sigma} \sqrt{(\mathbf{X}^*(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^{*T})_{i,i}}$$

- Notice that this is a confidence interval for $\mathbf{X}^*\beta$.

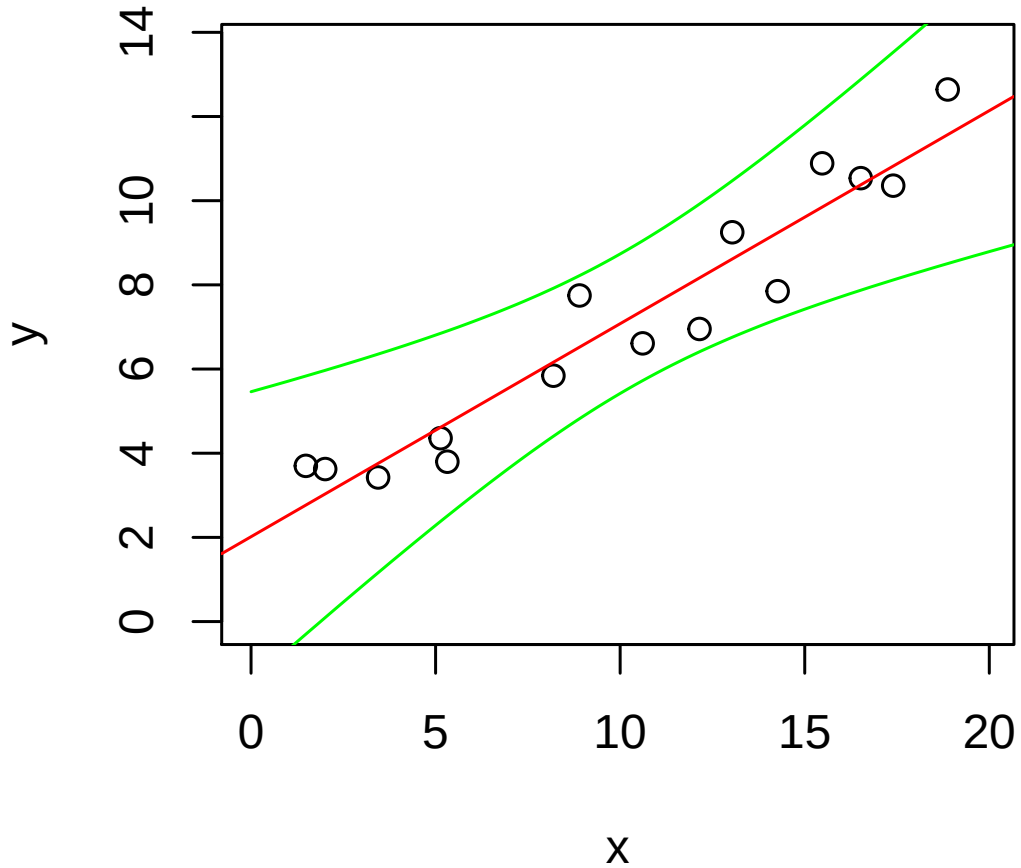
Prediction : Confidence interval



Prediction : Confidence interval



Prediction : Confidence interval



- This is in fact a confidence interval for $\mathrm{E}[\mathbf{y}^*]$.

Prediction : Prediction interval

- The confidence interval reflects our uncertainty about the population regression line (its parameters).
- We know that y itself is more variable as $\beta_0 + \beta_1 x$ is only its expectation.
- A *prediction interval* reflects the possible values for a new data points generated according to the model (considering ε)
- Let's construct an interval for our prediction error when we use $\hat{\beta}_0 + \hat{\beta}_1 x^*$.

Prediction : Prediction interval

- We know that y itself is more variable as $\beta_0 + \beta_1 x$ is only its expectation.
- A prediction interval reflects the possible values for a new data points generated according to the model (considering ε)
- Let's construct an interval for our prediction error when we use $\hat{\beta}_0 + \hat{\beta}_1 x^*$.

Prediction : Prediction interval

- $y^* = \beta_0 + \beta_1 x^* + \varepsilon^*$
- Our prediction :$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$
- The prediction error : $y^* - \hat{y}^* = \beta_0 + \beta_1 x^* + \varepsilon^* - \hat{\beta}_0 + \hat{\beta}_1 x^*$.
- The prediction error : $y^* - \hat{y}^* = \beta_0 + \beta_1 x^* + \varepsilon^* - \hat{\beta}_0 + \hat{\beta}_1 x^*$.
- Matrix notation :$\mathbf{y}^* - \hat{\mathbf{y}}^* = \mathbf{X}^* \beta + \varepsilon - \mathbf{X}^* \hat{\beta}$

Prediction : Prediction interval

$\mathbf{y}^* - \hat{\mathbf{y}}^* \sim N(\mu, \Sigma)$, where

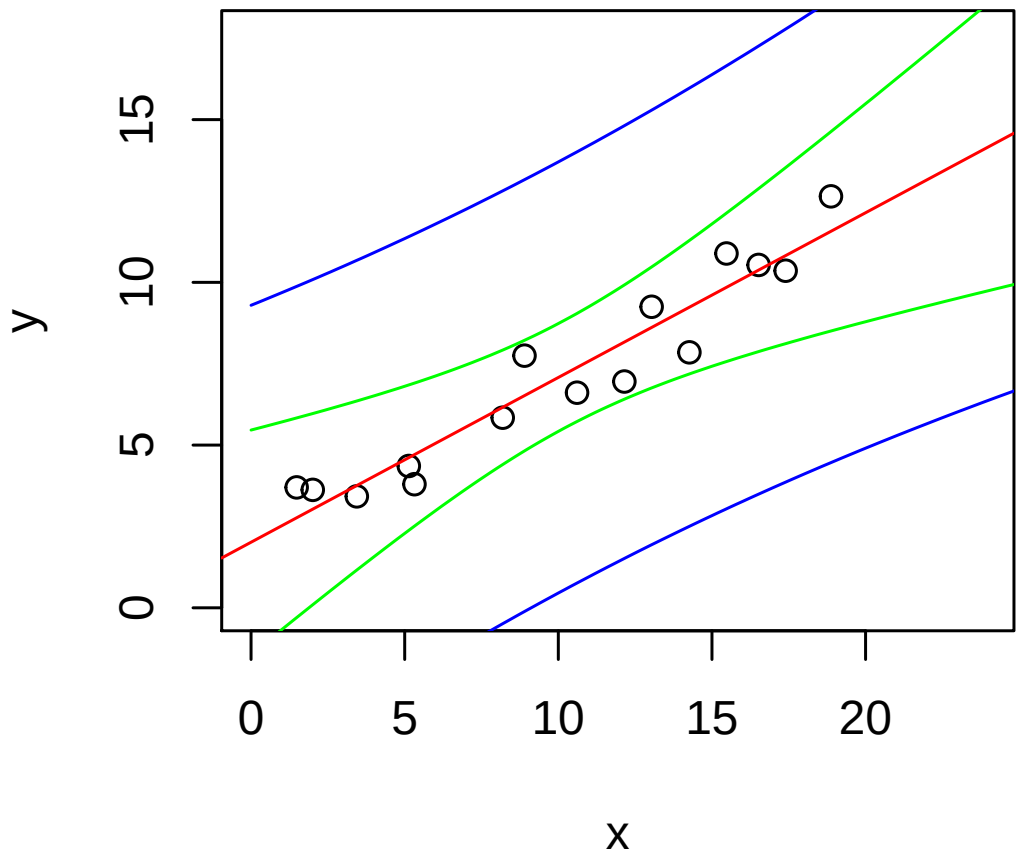
$$
\begin{aligned}
\mu &= E[\mathbf{X}^*\beta + \varepsilon - \mathbf{X}^*\hat{\beta}] \\
&= \mathbf{X}^*\beta + E[\varepsilon] - E[\mathbf{X}^*\hat{\beta}] \\
&= \mathbf{X}^*\beta + 0 - \mathbf{X}^*\beta = 0 \\
\Sigma &= \mathrm{Var}[\mathbf{X}^*\beta + \varepsilon - \mathbf{X}^*\hat{\beta}] \\
&= \mathrm{Var}[\varepsilon - \mathbf{X}^*\hat{\beta}] \\
&= \mathrm{Var}[\varepsilon] + \mathrm{Var}[\mathbf{X}^*\hat{\beta}] + 2\mathrm{cov}[\varepsilon, \mathbf{X}^*\hat{\beta}] \\
&= \sigma^2 I + \sigma^2 \mathbf{X}^*(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^{*T} + 0 \\
&= \sigma^2[I + \mathbf{X}^*(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^{*T}]
\end{aligned}
$$

Prediction : Prediction interval

- Since σ is unknown, we estimate it and use a student distribution!
- The prediction interval is :

$$\Rightarrow CI : \hat{y}_i^* + / - t_{(n-2)}^{(0.025)} * \hat{\sigma} \sqrt{[I + \mathbf{X}^*(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^{*T}]_{i,i}}$$

Prediction : Prediction interval



Prediction

- This conclude the inference we can do with our current model.
- Under a mild assumption $y = \beta_0 + \beta_1 x + \varepsilon$ with $\varepsilon \sim N(0, \sigma^2)$ we first estimated the paramaters.
- Because of our probabilistic modeling we could build confidence interval on our parameters and predicted values.
- Our results a more complete than simple point estimation, but came at a cost: our assumptions.

How's our model doing ?

- What is a *good* a model ?
- A good model is good at *explaining* the response. It is *usefull*
- R-squarred introduce in Lecture 3 is one way to check that.
- For a model to be a *valid* model, the assumptions must be respected.
- Even if the R^2 coefficient is low and the parameters are all non-significant we can have a *valid* model.

Checking the model assumption

Model assumptions

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

- Gauss-Markov assumptions (conditions) :
  - $\mathbf{E}(e_i) = 0$
  - $\text{Cor}(e_i, e_j) = 0 \; i \neq j$
  - $\text{Var}(e_i) = \sigma^2 < \inf \forall i$
- Our model assumptions : e_i are independantly distributed according to $N(0, \sigma^2)$.
- We also assume the relationship is linear : $\mathbf{E}[y_i] = \beta_0 + \beta_1 x_i$
- Finally we assume all the data points are generated from the same distribution (the first i in i.i.d).

Model checking

- We will divide model checking into three pieces :
  - Errors assumption (e_i i.i.d. $N(0, \sigma^2)$)
  - Identical distribution (checking for unexpected observations)
  - Model assumption (linearity)

Model checking

Diagnostic techniques can be graphical, which are more flexible but harder to definitively interpret, or numerical, which are narrower in scope, but require no intuition.

- In the slides to come we will focus on graphical techniques.

Model checking

- We will not learn to fix all the issues, but rather to identify if there is an issue.
- We only learn about simple linear models in STA302, good thing they can solve many problems.
- If it linear models are not appropriated, we must be able to identify it and mention it.
- Later (STA303 and more advance courses) you will learn more refined techniques to extend the linear model beyond some of it's assumption.

Check error assumptions

- To begin, let's note that it is technically impossible to check the errors e_i directly.
- We only have acess to the residuals (observed errors) $\hat{e}_i = y_i - \hat{y}_i$.
- Thus, diagnostics are often applied to the residuals in order to check the assumptions on the error.
- But, the residuals and the errors don't exactly have the same distributions.
- Remember $\hat{e}_i = y_i - \hat{y}_i \Rightarrow \widehat{\mathbf{e}} = (I - H)\mathbf{y}$

Check error assumptions : Constant variance

- The first assumption we will learn to check is the constant variance (remember we assumed $\text{Var}(e_i) = \sigma^2 \ \forall i$).
- We assume there is no structure in the error variance. One graphical way to check this is to plot the residual against the fitted values:

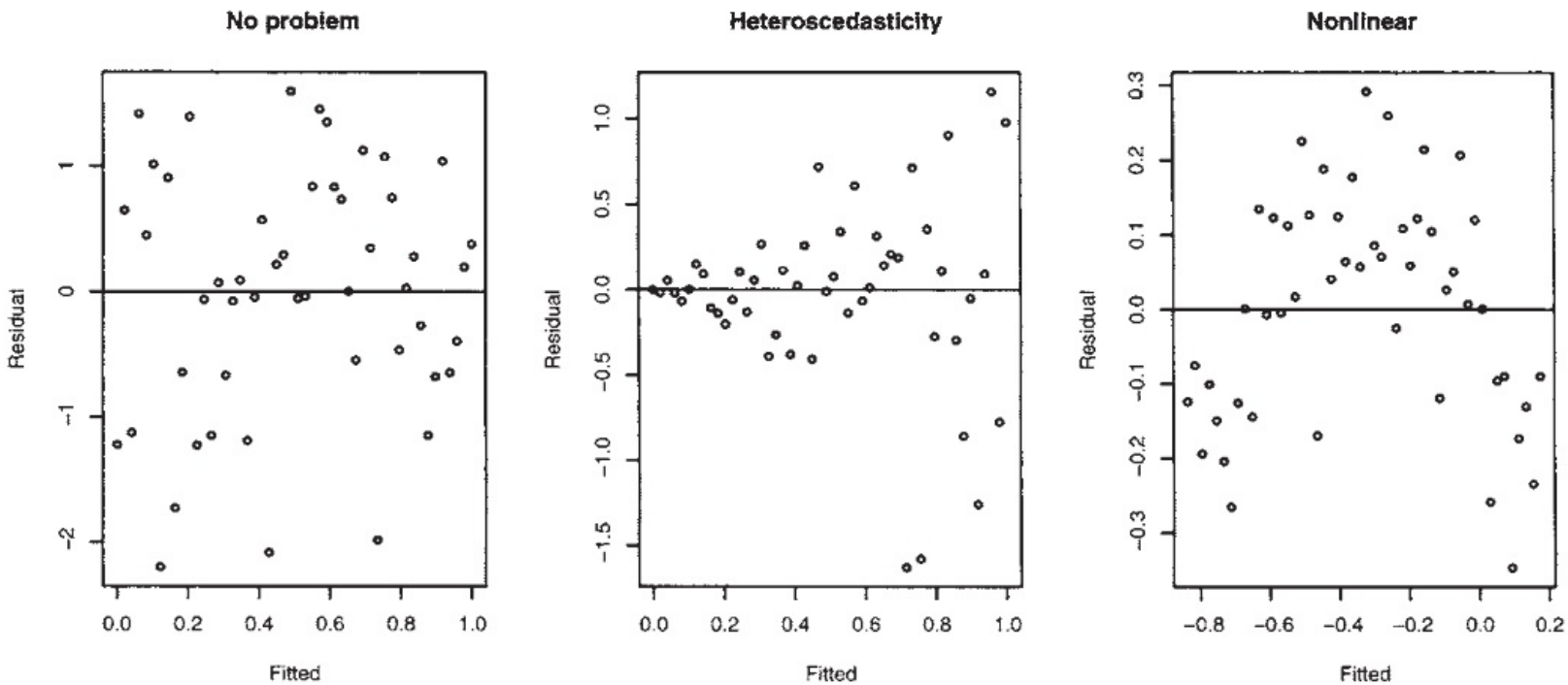


Figure 1: Residuals against fitted

Check error assumptions : Constant variance

- These plots are the bread and butter for checking constant variance (and uncorrelatedness).
- We expect well-dispersed residuals, with no clear pattern (like in the first plot).
- It takes some experience. (We are getting some experience right now!)

Check error assumptions : Constant variance

- It is also usefull to plot the residuals against the predictor.

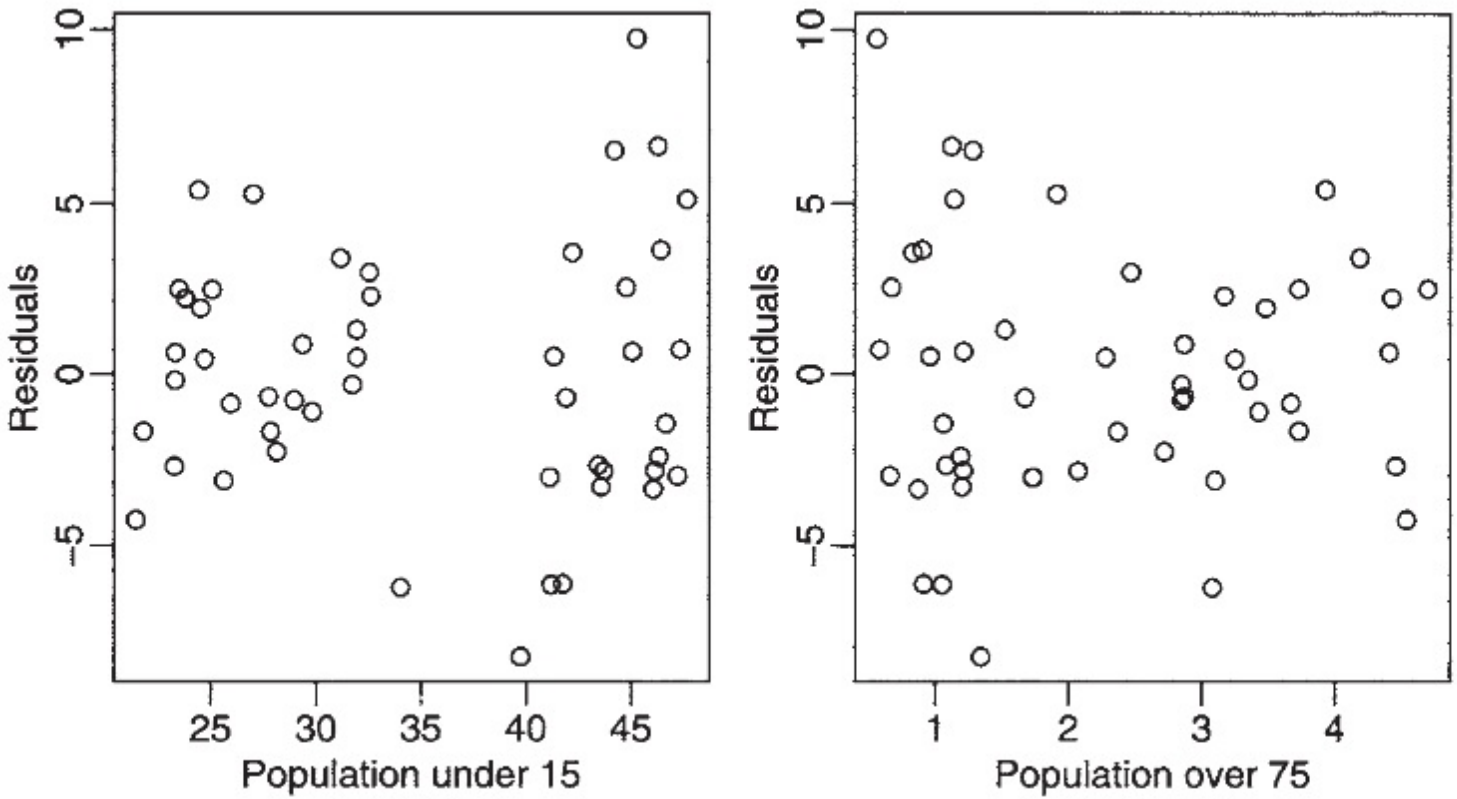


Figure 2: Residuals against predictor

Check error assumptions : Constant variance

- If the assumption looks to be roughly respected: we are happy!
- If the assumption looks violated : we are sad!
- Most important thing is to actually notice the assumption is violated, mention/document it.
- You should be more nuanced about the data.
- Transformation of the response can help with heteroscedasticity.
- A clear pattern in the residual plots indicates the need for a new model since it indicates a clear pattern has been missed by our model.

Check error assumptions : Constant variance

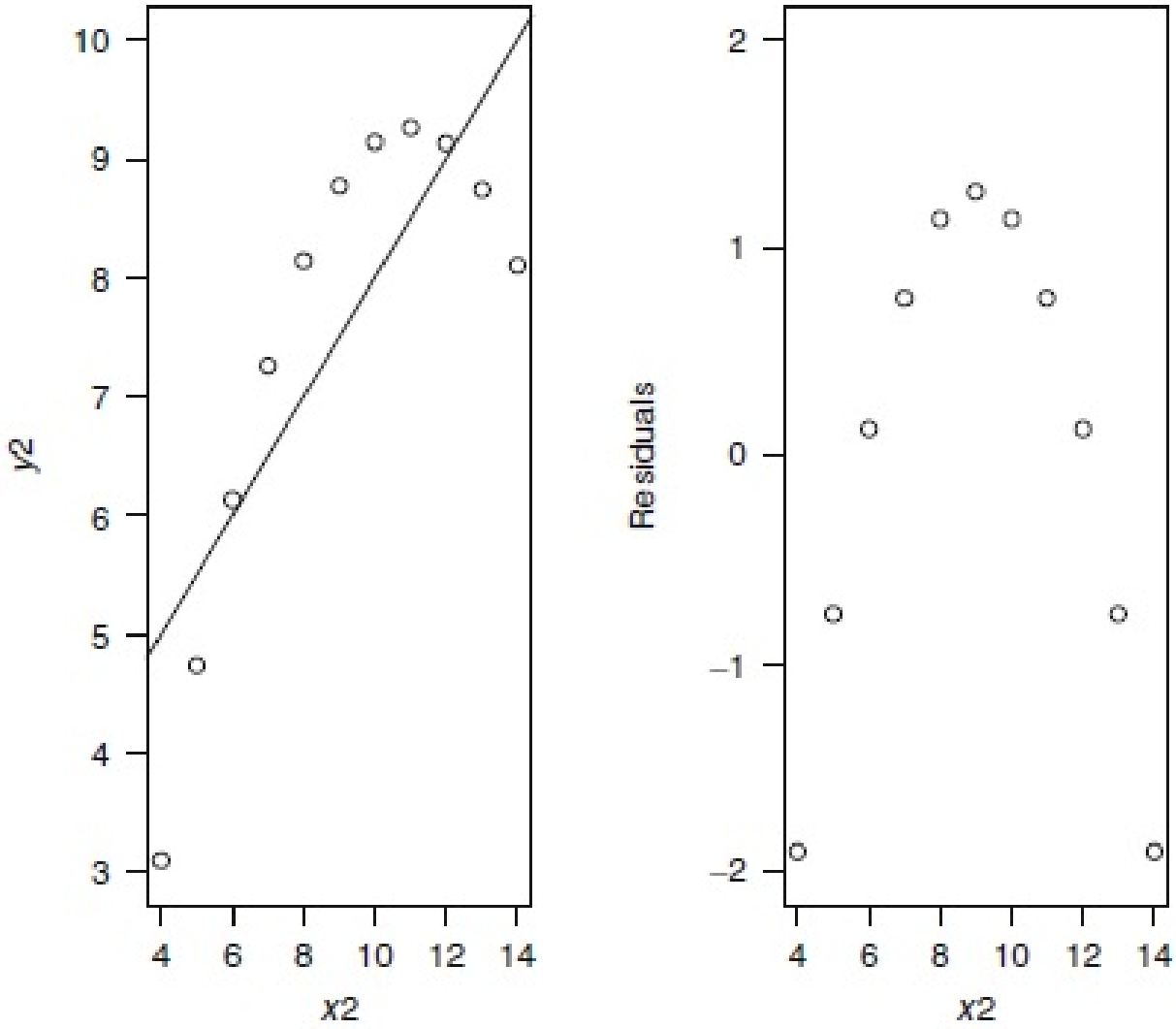


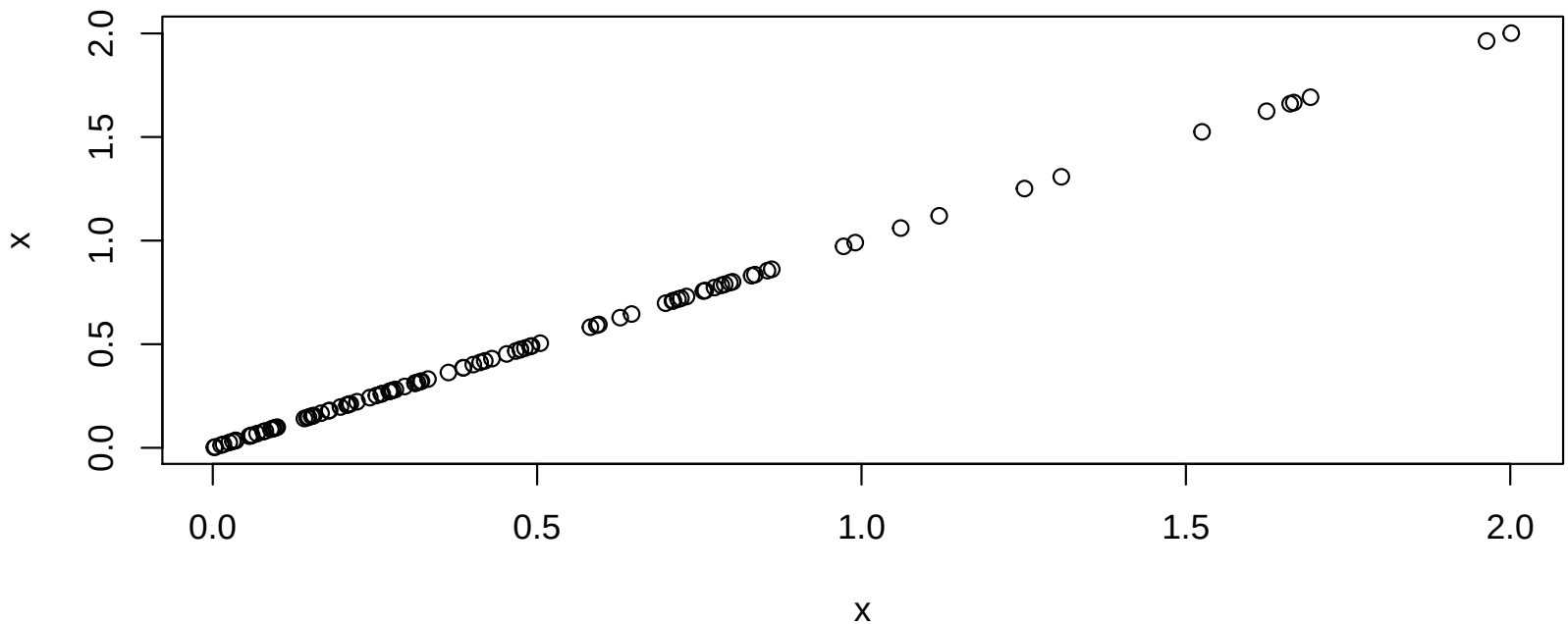
Figure 3: Residuals against fitted

Check error assumptions : Normality

- QQ-plot is our friend!
- QQ-plots stands for Quantile to Quantile plot.
- It plots the quantile of one distribution against another one.
- If the two distributions are the same with should expect a straitgh line

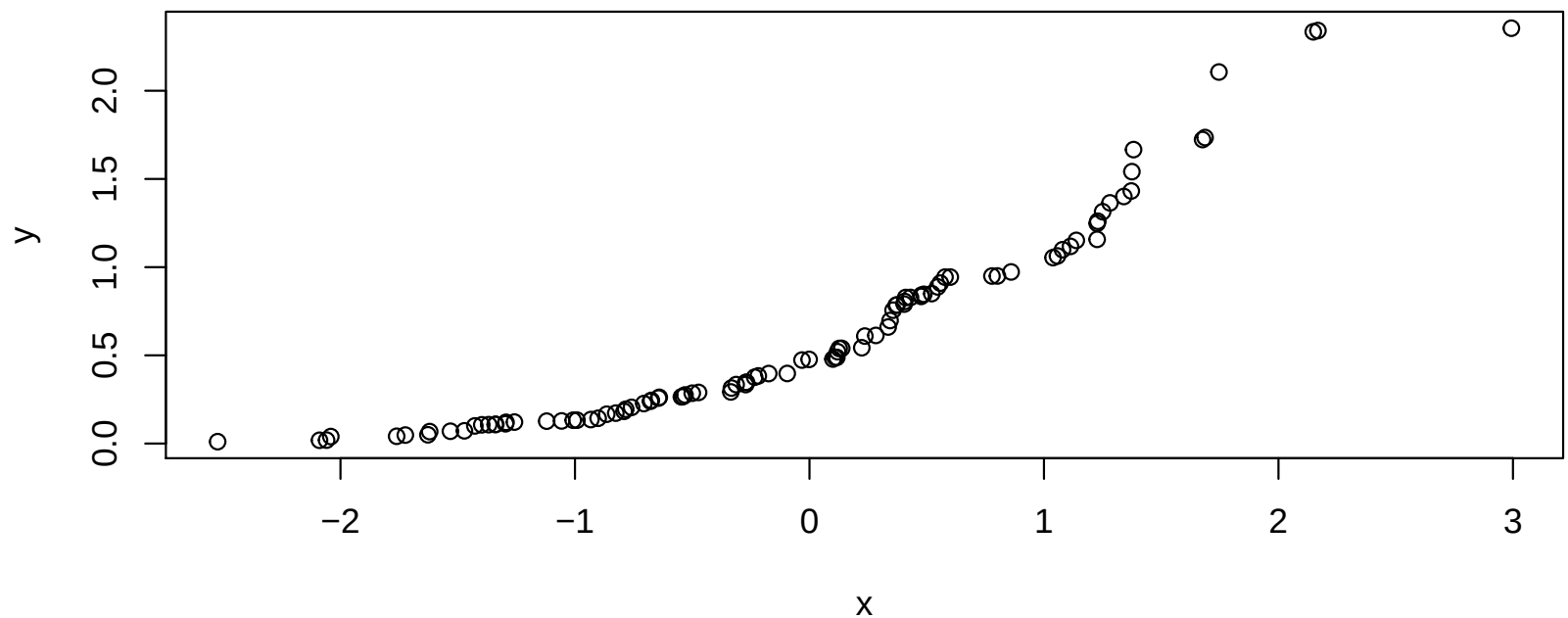
Check error assumptions : Normality

```
x <- rexp(n=100,rate=2)
qqplot(x,x)
```



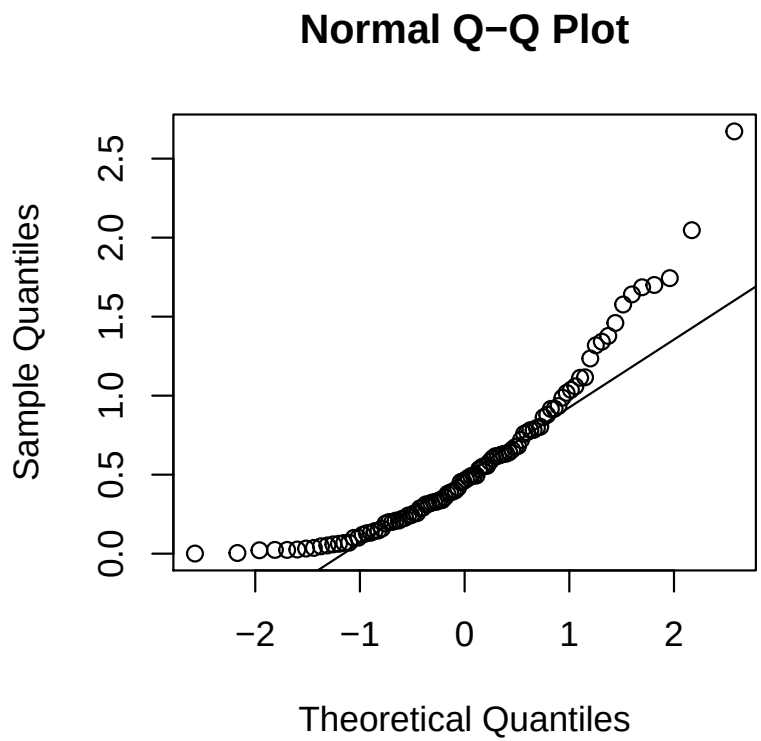
Check error assumptions : Normality

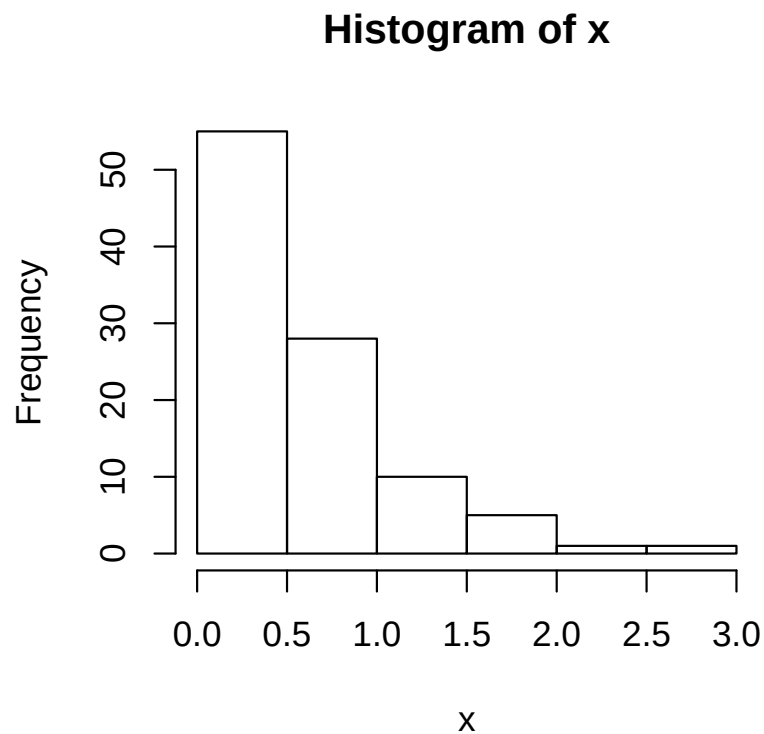
```
y <- rexp(n=100,rate=2)
x <- rnorm(n=100,0,1)
qqplot(x,y)
```



Check error assumptions : Normality

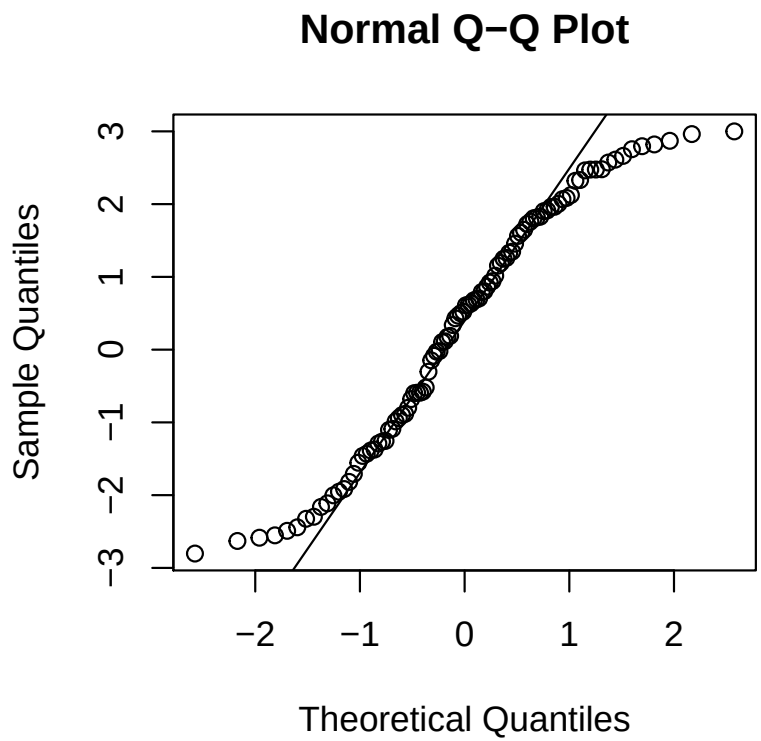
```
x <- rexp(n=100,rate=2)
par(mfrow=c(1,2))
qqnorm(x)
qqline(x)
hist(x)
```

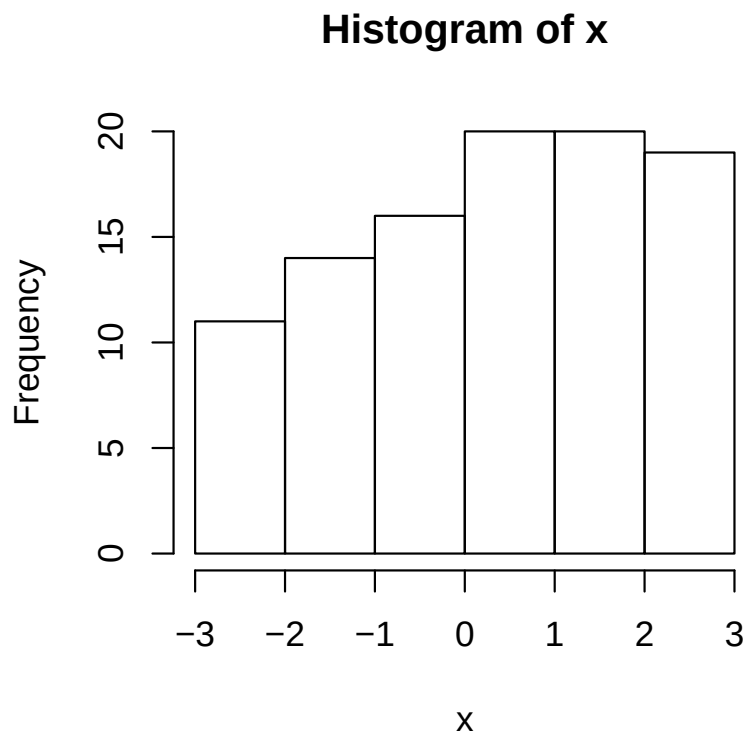




Check error assumptions : Normality

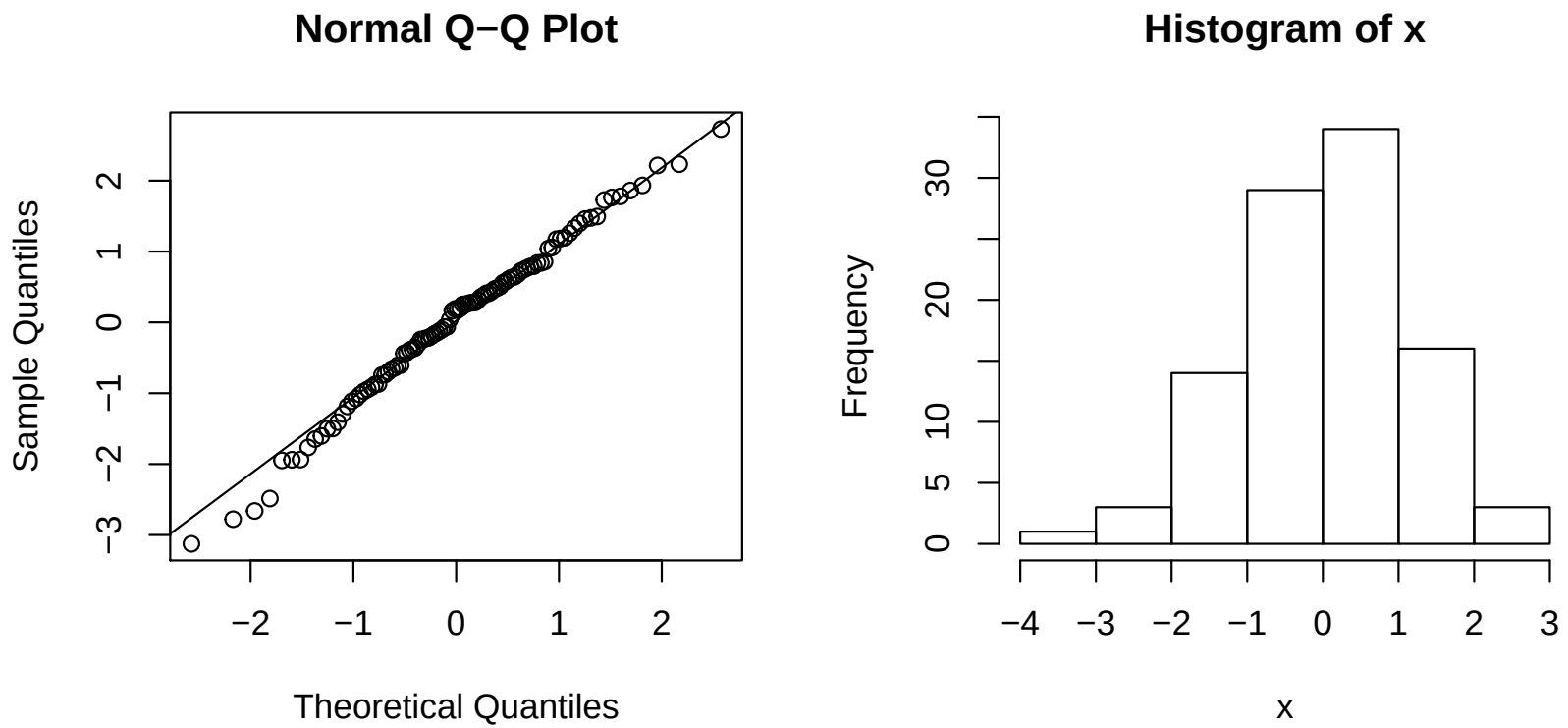
```
x <- runif(n=100,-3,3)
par(mfrow=c(1,2))
qqnorm(x)
qqline(x)
hist(x)
```





Check error assumptions : Normality

```
x <- rnorm(n=100,0,1)
par(mfrow=c(1,2))
qqnorm(x)
qqline(x)
hist(x)
```



Check error assumptions : Normality

- The closer the sample quantiles are close to the theoretical one, the more inclined to believe the assumption is satisfied.
- When the errors are not normal:
  - Our estimates may not be optimal.
  - The tests and confidences intervals are not exact.
  - But mild nonnormality can safely be ignored and the larger the sample size the less troublesome the nonnomrality.

Check error assumptions : Normality

- There exist formal statistical tests for normality (Shapiro-Wilk for example)
- They are not truly reliable. (Practice problems)
- The p-value is not a really good guide regarding the action to take to fix the issue. (Compared to QQ-plot that can be interpreted.)

Check error assumptions : Uncorrelatedness

- For temporally or spatially related data set is wise to check the uncorrelated assumption. (Daily measurements, repeated measurements using the same tool, spatial index. . .)
- We are asking ourselves do e_i depends on e_{i-1} ?
- A graphical check would be to plot $\hat{e}_i$ against $\hat{e}_{i-1}$.
- There also exist statistical test (Durbin-Watson for example).

Check error assumptions : Uncorrelatedness

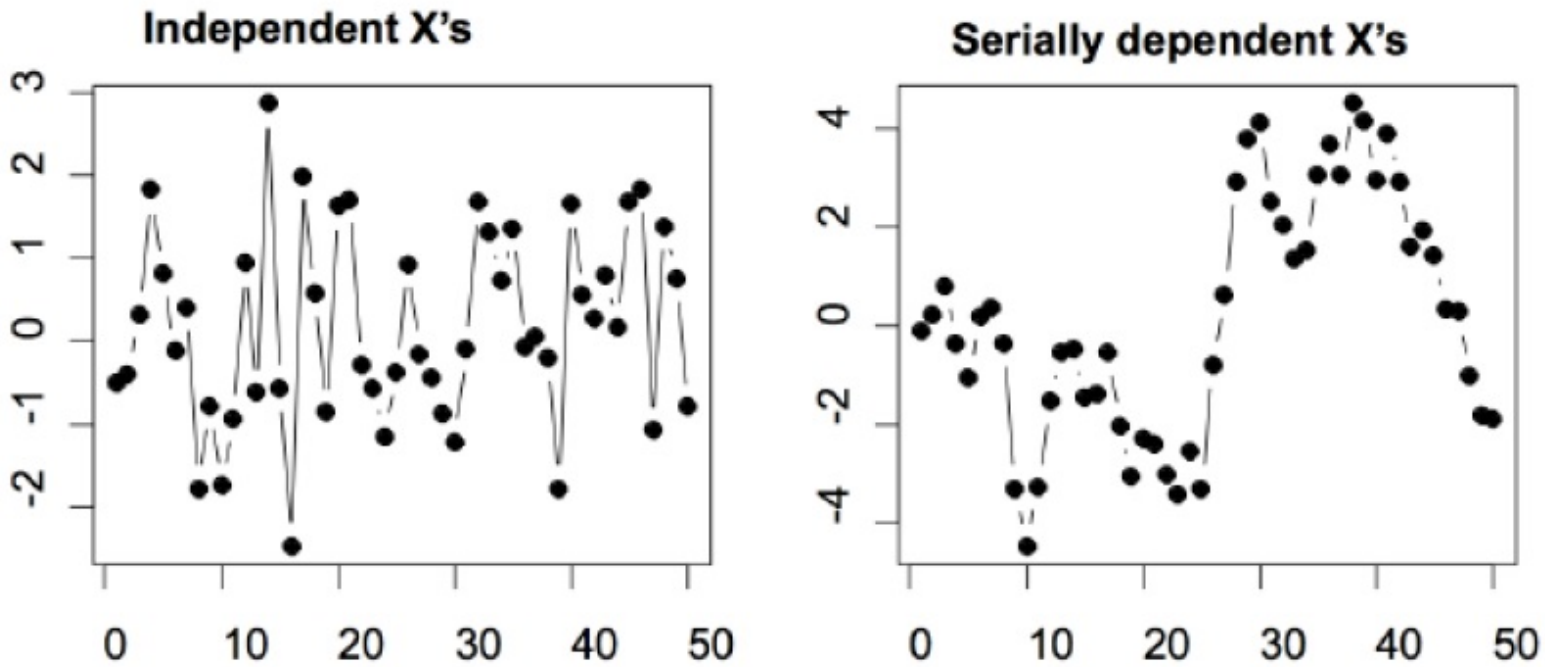


Figure 4: Serial correlation

Check error assumptions : Uncorrelatedness

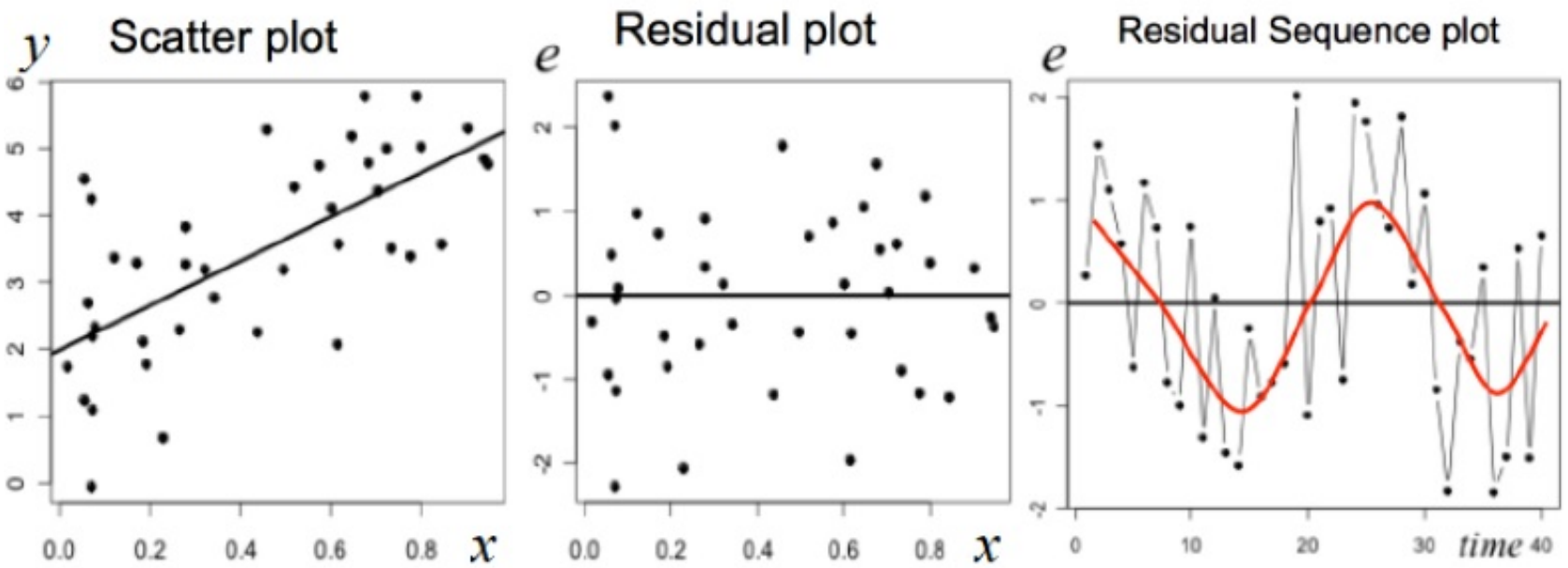


Figure 5: Serial correlation

Check error assumptions : Uncorrelatedness

- We rarely check for this assumption.
- Usually when we do we have a reason to believe there is.
- If there is these type of relationship, it must be included in the data somehow (Time series, spatial statistics,. . .)

Check error assumptions

- What do to when the assumptions are violated ?
- You can start by noticing it, it should raise a flag.
- Be more nuanced about your result (depending on the violated assumption)
- Apply a transformation (next week).

Check error assumptions

- Finally, since $\widehat{\mathbf{e}} = (I - H)\mathbf{y}$, $\mathbf{Var}(\hat{e}_i) = (I - H)_{i,i}\sigma^2$.
- Some recommend using *standardized* residuals instead: $r_i = \frac{\hat{e}_i}{s\sqrt{(I-H)_{i,i}}}$, where $s = \sqrt{SSE/n-2}$ the unbiased estimator for σ.
- Usually plots fo standardized residuals are similar to plots of residuals.
- BUT when the data set contains unusual observations, such as leverage points, the standardized residuals are more informative.

Unusual observations

- Somes observations are *special* observations within our data set.
- Some are unusual in the predictor space; they have a predictor value far from other points. They are called **leverage points**
- Some do not fit well within the model. They are called **outliers**.
- Finally some change the fit in a substantive manner. They are called **influential** observations. They could be leverage points, outliers but usually they are both.

Unusual observations

- We have to find and identify those unusual observations.
- Then we have to decide what to do with them.
- My personnal grudge about this part of the diagnostic process : we don't know what to do with them. It's hard to find a consensus among statisticians.
- Let us define those unusual observations.

Unusual observations : Leverage points

- A leverage point is a point whose x-value is distant fro the other x-values.
- Typically we define the $h_i = H_{i,i}$ as the leverages and are usefull diagnostics.
- $h_i = H_{i,i} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$
- Since $\sum_{i=1}^{n} h_i = 2$ (number of parameters) then the average value for h is $2/n$.
- Usually we say leverages larger than $4/n$ should be looked at more closely.

Unusual observations : Leverage points

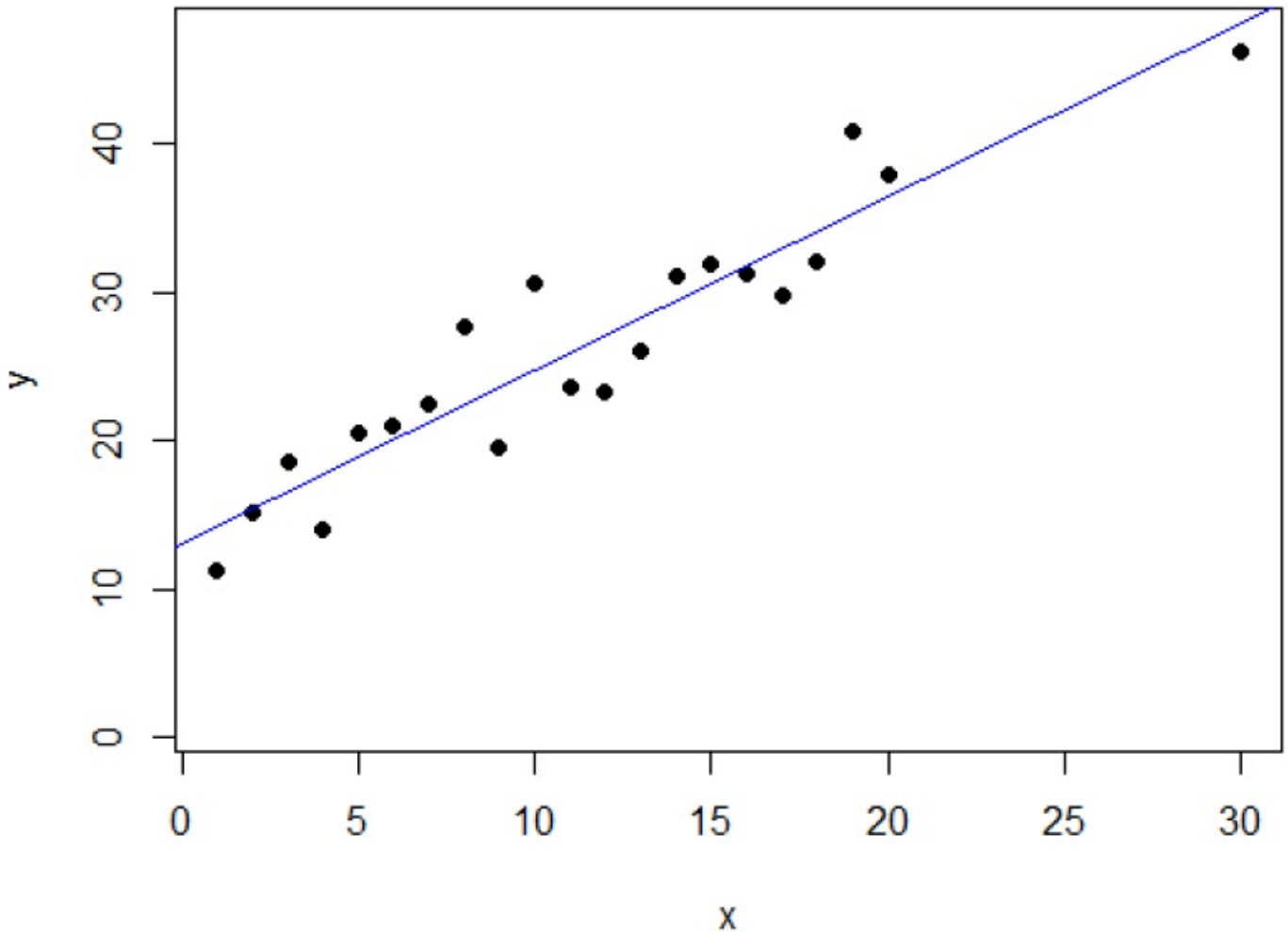


Figure 6: Leverage point

Unusual observations : Outliers

- Outliers have y-values distant from the other y-values. At least *different* from what you would expect.
- Usually large residual $\hat{y}_i - y_i$ might indicate outliers.

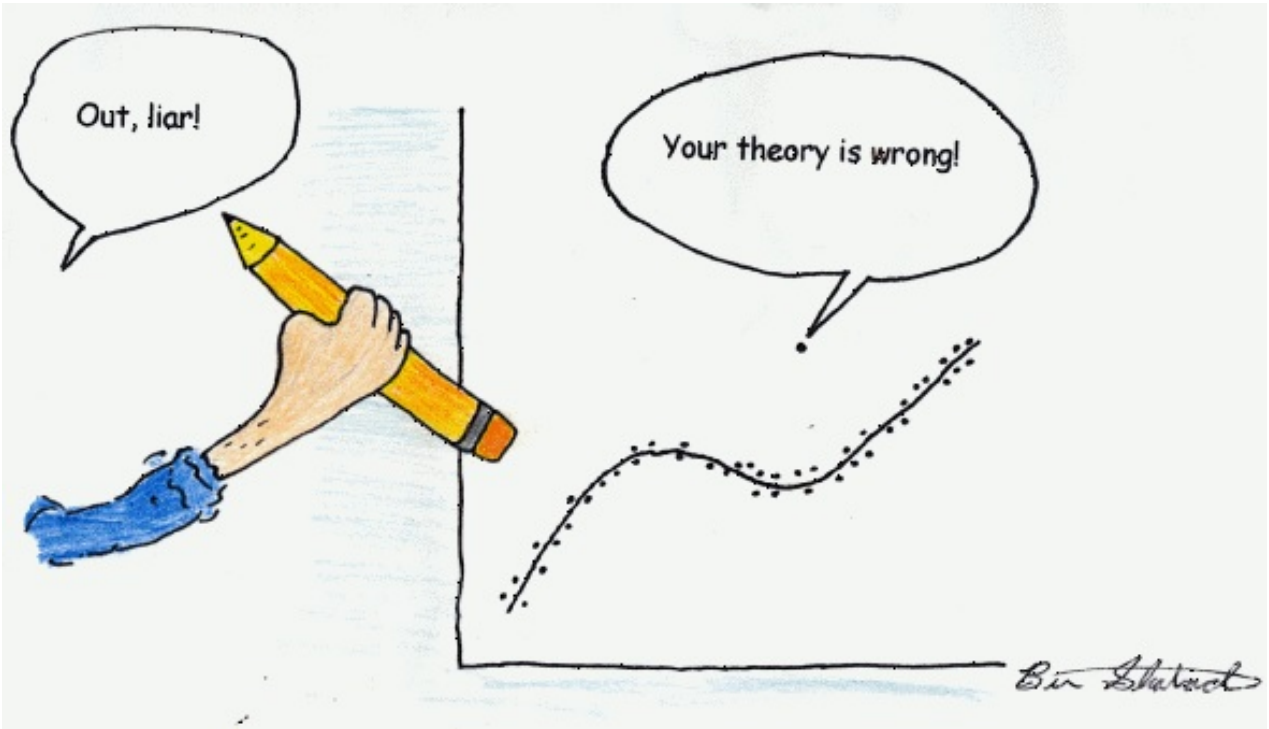


Figure 7: Outlier

Unusual observations : Influential observations

- An influential point is one whose removal from the dataset would cause a large change in the fit.
- It may or may not be an outlier. It may or may not be a leverage point but it will tend to be at least one of those.
- An outlier with a large leverage will definetly be an influential observation sometime named *bad leverage*.

Unusual observations : Influential observations

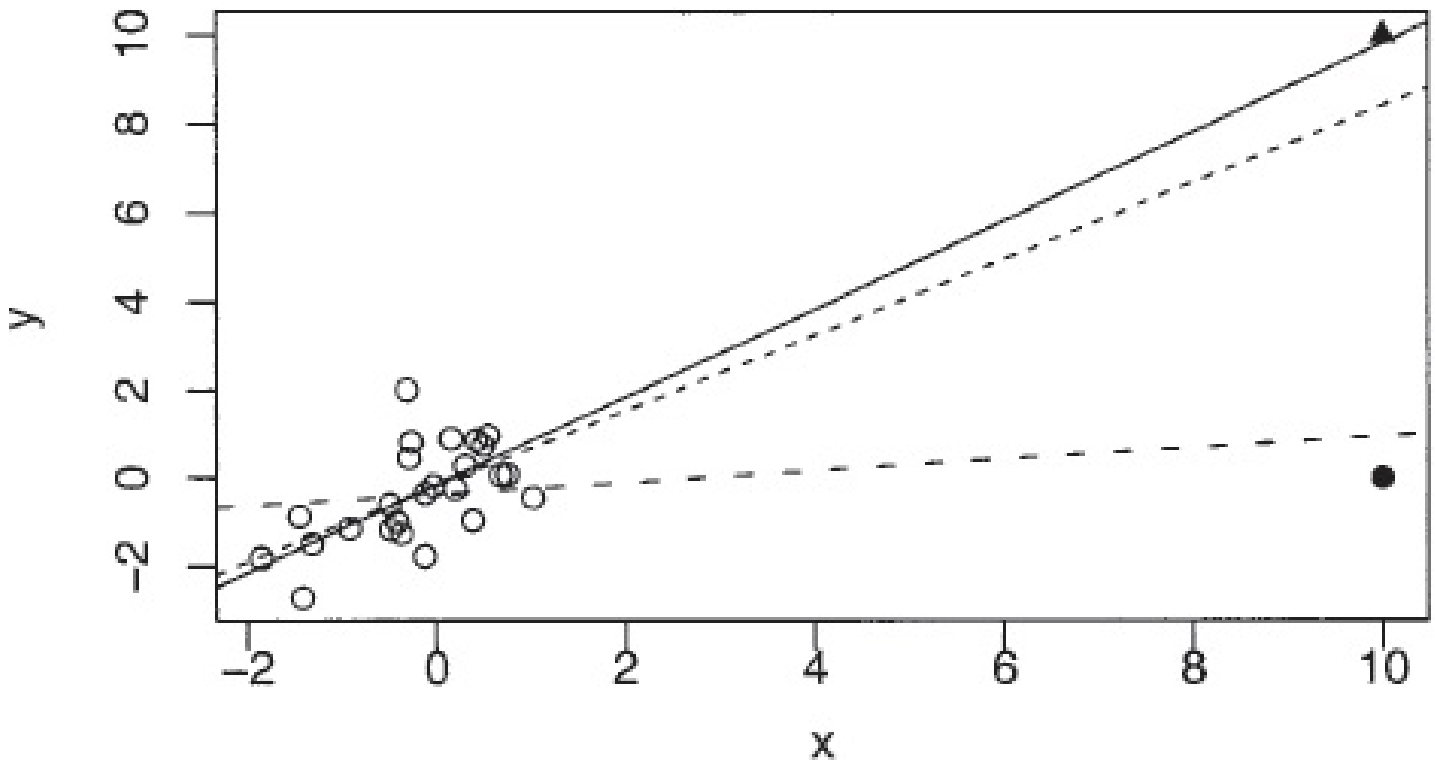
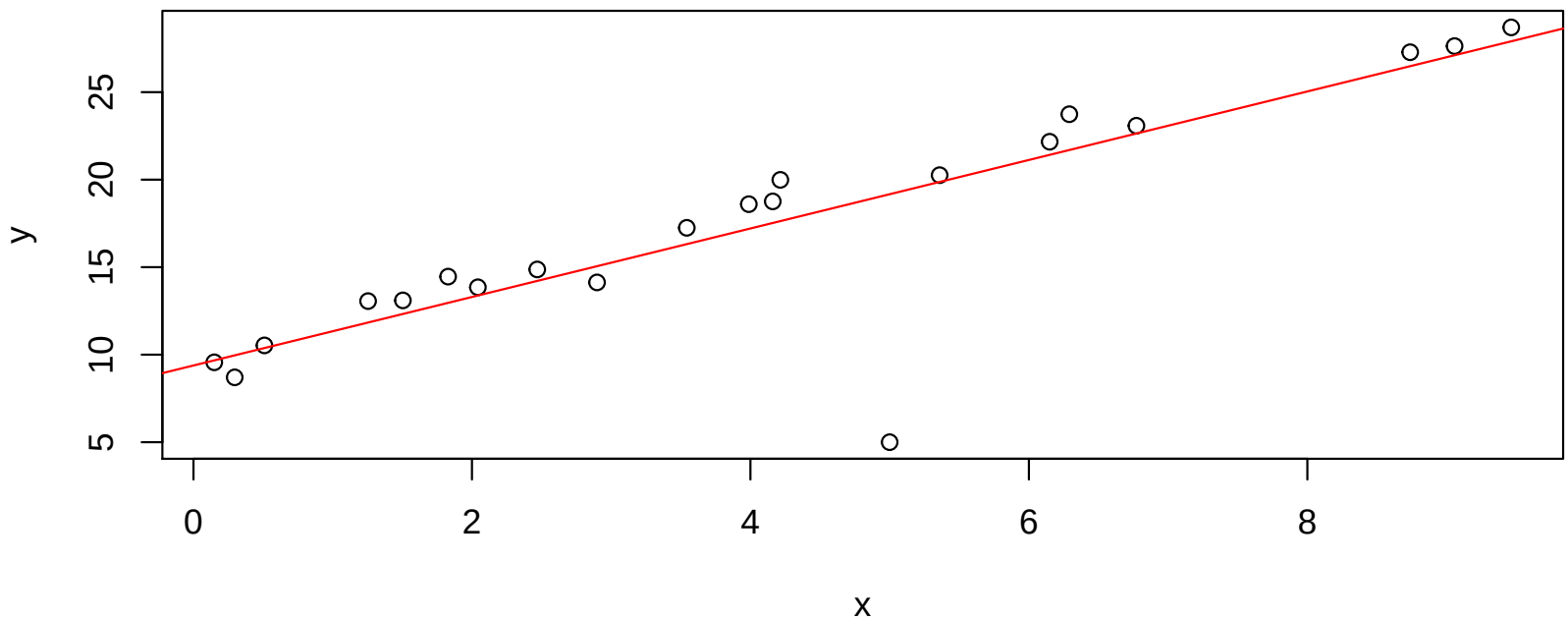


Figure 4.10 *Outliers can conceal themselves. The solid line is the fit including the ▲point but not the ● point. The dotted line is the fit without either additional point and the dashed line is the fit with the ● point but not the ▲ point.*

Figure 8:

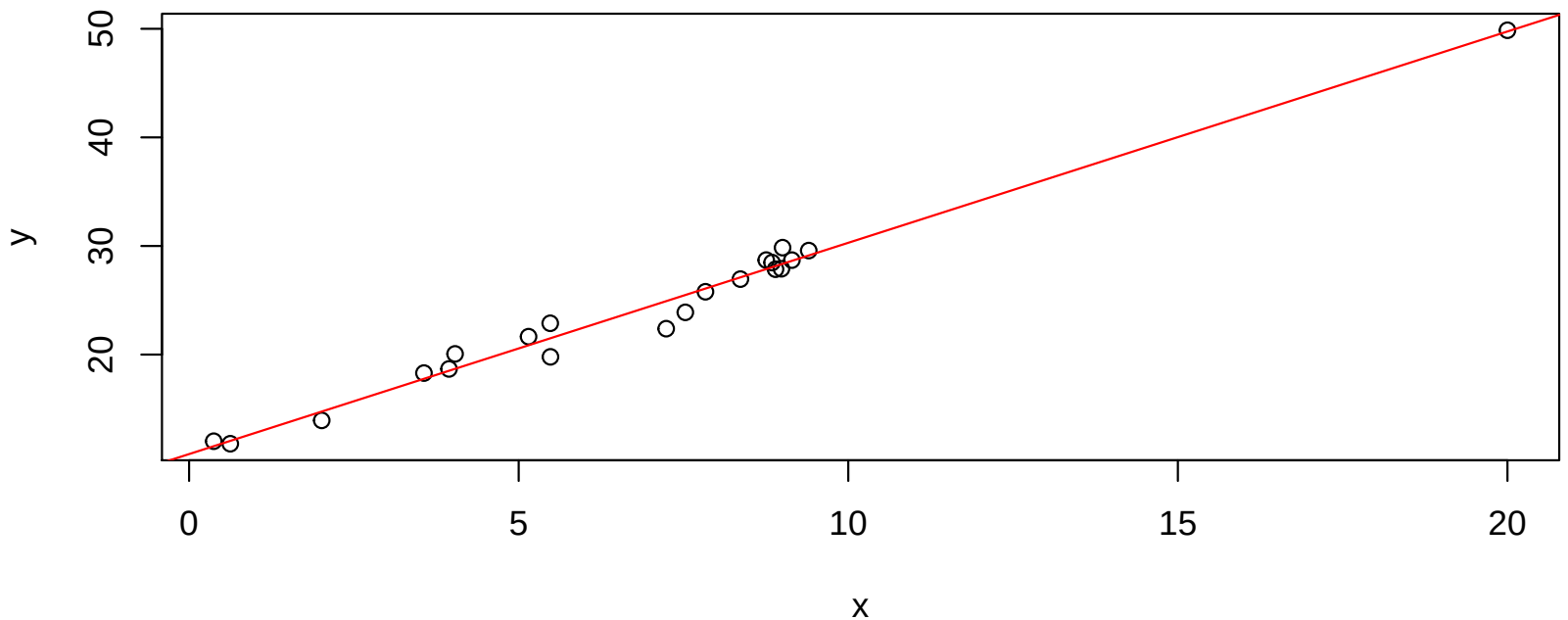
Unusual observations : Influential observations

- So if the observation x-value is close to others x-values (small leverage) but y-value is not (outlier), we have a small problem.



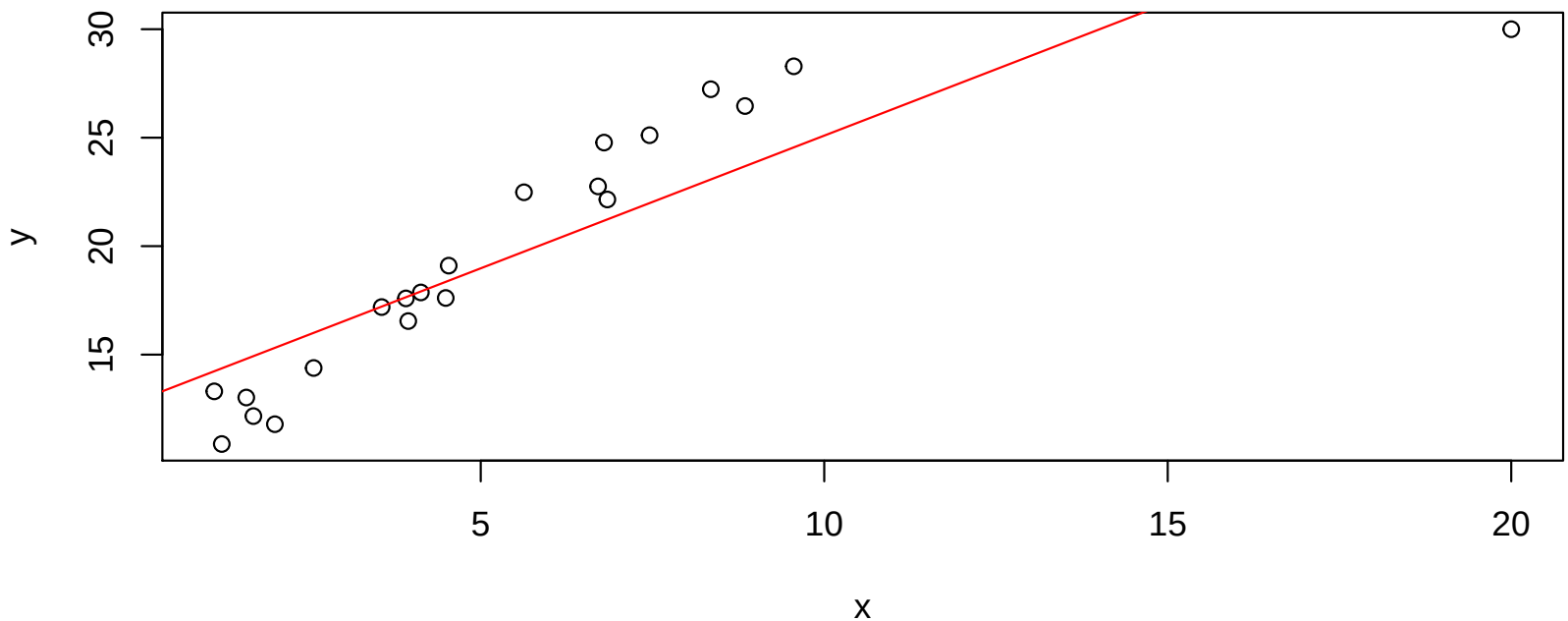
Unusual observations : Influential observations

- If the observation y-value is close to others y-values (no outlier) but x-value is not (large leverage), we have a small problem.



Unusual observations : Influential observations

- But if the x-value is not close to others x-values (large leverage) and y-value is not close to outher y-values (outlier), we have a big problem.



Unusual observations : Influential observations

- The Cook distance is inspire by the idea that the exist a *multiplicative effect* between leverages and outliers.
- For observation i the Cook's distance is $D_i = \frac{r_i^2}{2} \frac{h_i}{1-h_i}$, where r_i is the standardize residuals (accounting for outliers) and h_i is the leverage.
- Simple rules of thumb : There is a problem when
  - $D_i > 4/n$ on large datasets
  - $D_i > 1$ on small datasets
  - D_i is separated by a large gap from the other D_js

Unusual observations : Influential observations

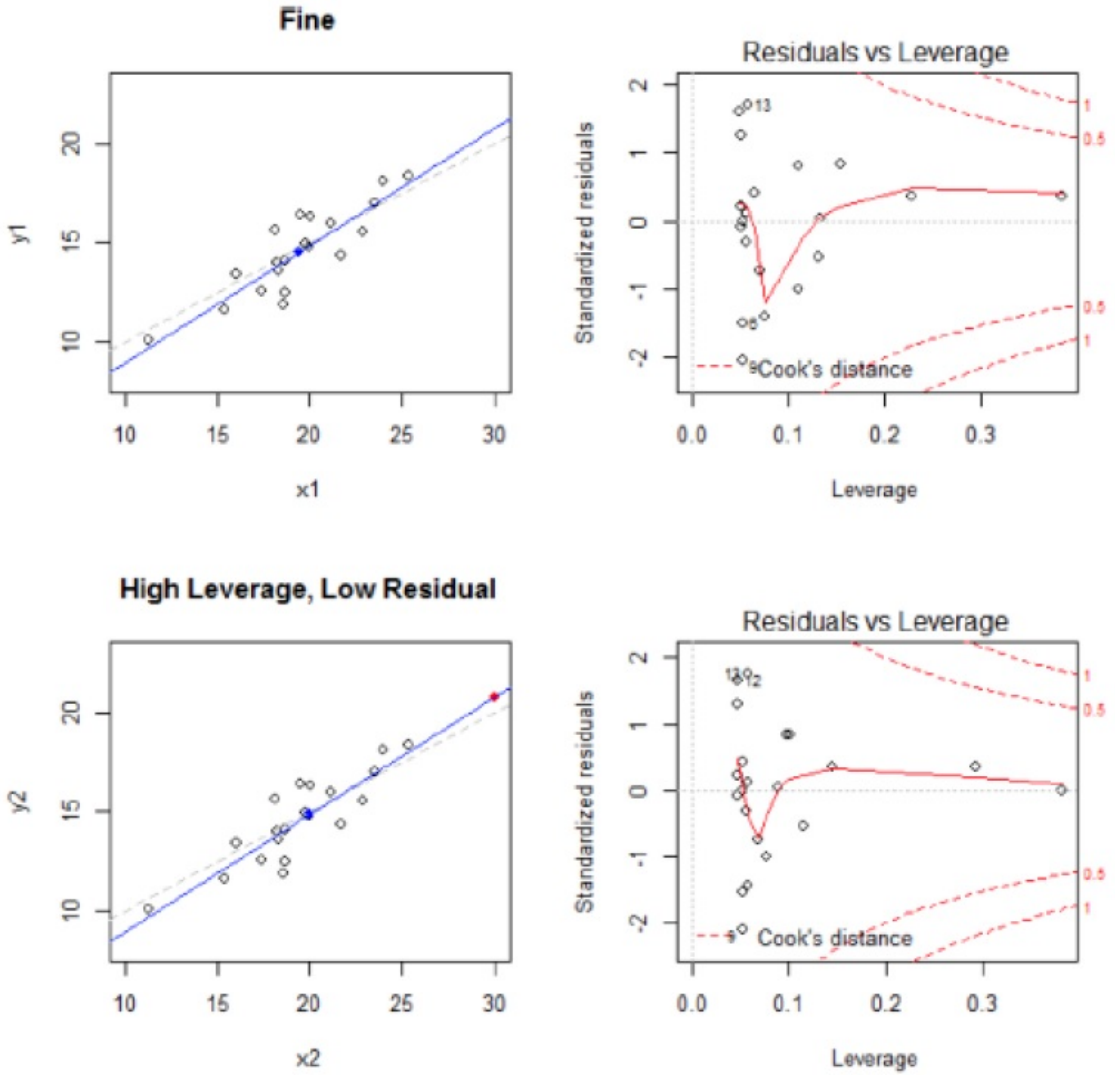


Figure 9: Thank you Professor Ebden

Unusual observations : Influential observations

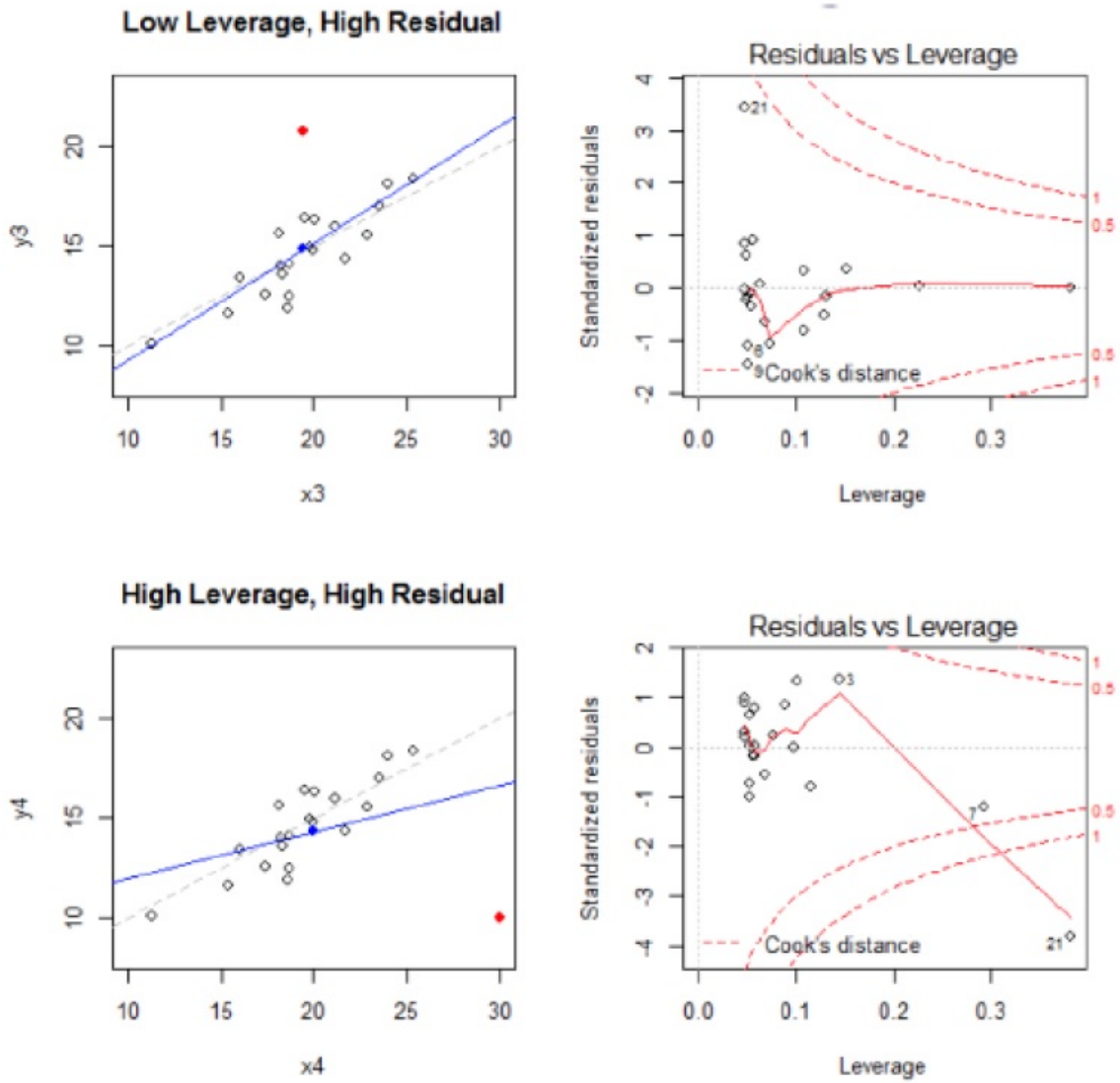


Figure 10: Thank you Professor Ebden

Unusual observations

- What to do with influential observations ?
- Try to figure out what hapenned :
  - Check for a data-entry error
  - Examine the physical context of the data
  - You might want to take it out in some cases (I don't like to do that)
- Check the different fit with different model (if you do that YOU HAVE to be transparent about it)
- Pick a different model ? Apply a transformation to the data (next week).

Checking linearity

- Usually the residuals plot will be used to check this assumption.

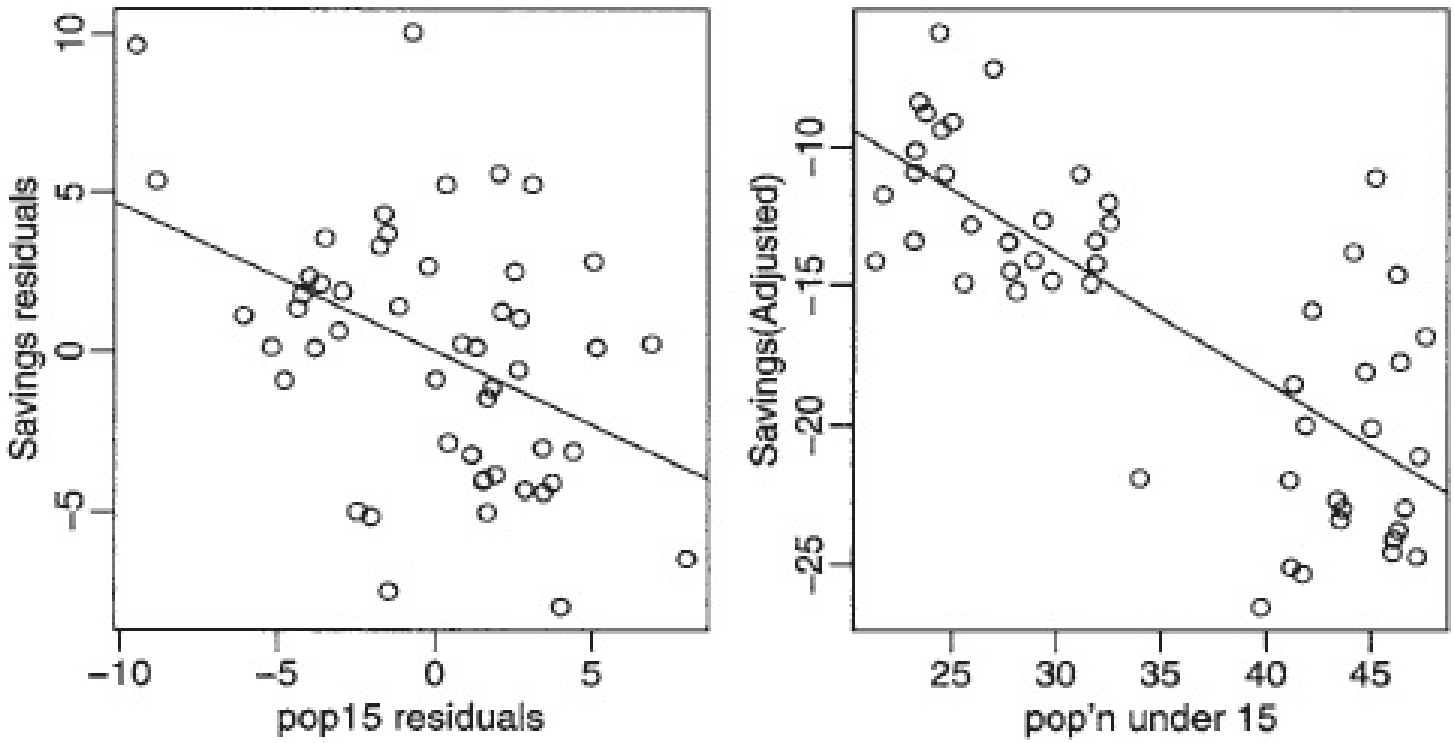


Figure 11: Maybe you need more predictors

Checking linearity

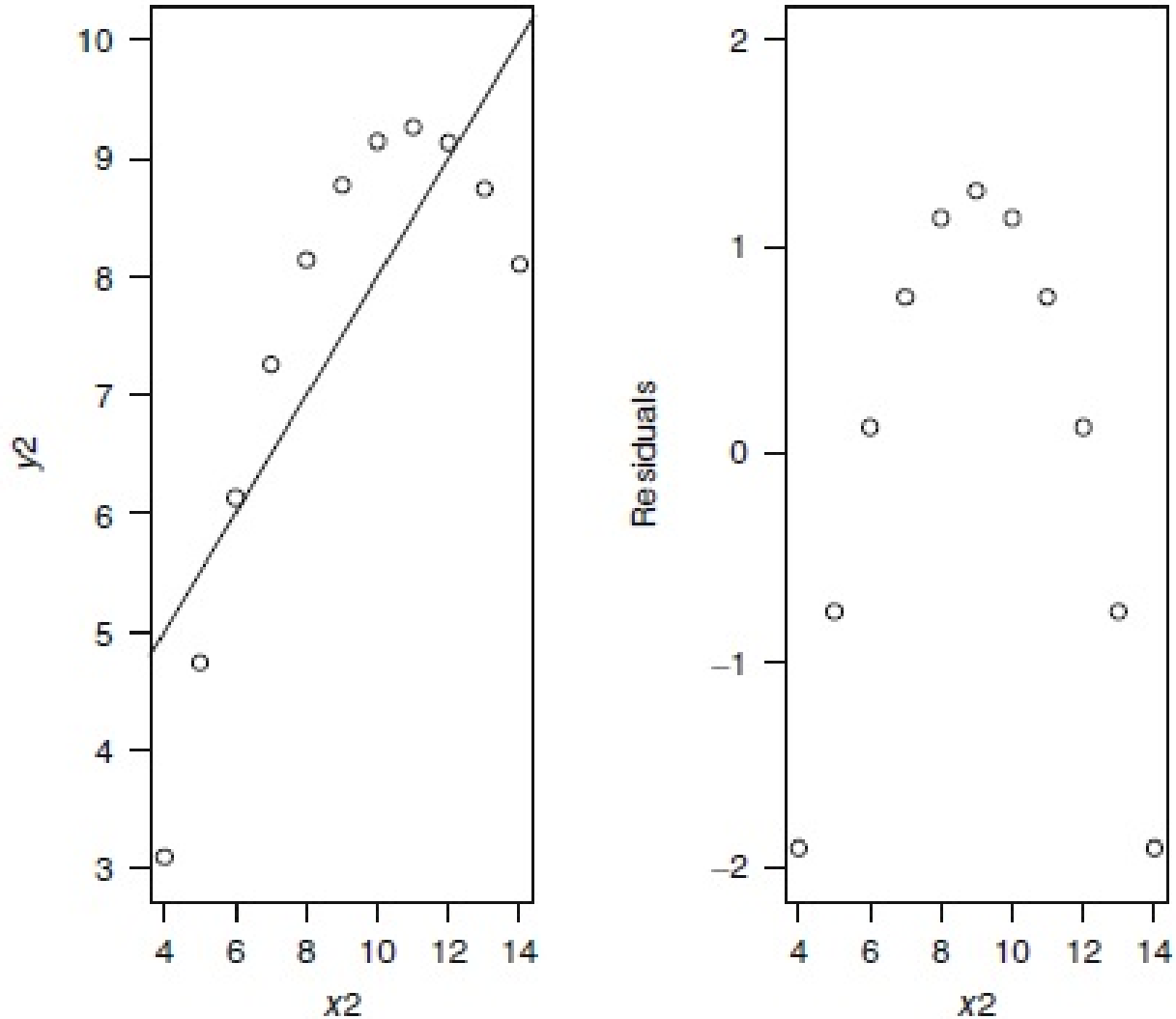


Figure 12: Maybe the relation ship is linear with x^2

Checking linearity

- This is the assumption of all linear models.
- If it is not respected, it is because the relationship is not linear.
- Adding predictors, transforming the predictor or the response or using more advanced model are all solutions.

Conclusion

- We must check the model's assumptions.
- The first step is at least to be able to identify if the assumptions are violated.
- Next week we will quickly review the diagnostic procedure and attempt to fix some of those issues with simple transformations.

Practice Problems

- A Modern Approach to Regression with R ch.2 : 1(c,d) (do with R)
- Alison Gibbs' additional chapter 3 practice problems : 1 (here)
- On R, run the simulated data test loop suggested in Linear Models with R ch.4 (p. 60 and 65)
- 8)h) From Craig Burkett's list of problems (Quercus)
- A Modern Approach to Regression with R ch.3 : 1,4(a) solutions (here)
- As usual, try to run today's Rlab, get some experience with diagnostic plots.

External Sources

- A Modern Approach to Regression with R ch.2
- A Modern Approach to Regression with R ch.3
- Linear Models with R ch.4
- Linear Models with R ch.3
- A Modern Approach to Regression with R ch.6