Linear Regression - Part 1

Anders Stockmarr
Course developers: Anders Stockmarr, Elisabeth Wreford Andersen

DTU Department of Applied Mathematics and Computer Science
Section for Statistics and Data Analysis
Technical University of Denmark
anst@dtu.dk

January 7th, 2025

Plan for this week

Monday	Statistical inference, and the t-test
Tuesday	Simple and Multiple regression
Wednesday	ANOVA, ANCOVA, and linear models
Thursday	Categorical data, statistical report writing, logistic Regression
Friday	Introduction to repeated measures , Principal Component Analysis

Learning objectives

After this session you should be able to:

1. Understand what a linear regression model is and be able to fit it to data;

Learning objectives

After this session you should be able to:

1. Understand what a linear regression model is and be able to fit it to data;
2. Understand the assumptions of linear regression models and perform model diagnostics;

Learning objectives

After this session you should be able to:

1. Understand what a linear regression model is and be able to fit it to data;
2. Understand the assumptions of linear regression models and perform model diagnostics;
3. Try polynomials or logarithms if the assumptions are not fulfilled;

Learning objectives

After this session you should be able to:

1. Understand what a linear regression model is and be able to fit it to data;
2. Understand the assumptions of linear regression models and perform model diagnostics;
3. Try polynomials or logarithms if the assumptions are not fulfilled;
4. Illustrate the fitted model on the data.

Overview

03_Simple regression

Simple Linear Regression

03_Simple regression

- The association between two continuous variables:
- Y response / outcome / dependent variable;
- X explanatory / covariate / independent variable.

Data

03_Simple regression

- Observations of pairs (x_i, y_i) for all $i = 1, \ldots, n$ individuals or units.
- Note that regression is not symmetrical in X and Y.
- In some studies it is possible to chose X beforehand, this gives more precise results.

Example

03_Simple regression

Is higher education associated with higher murder rates?

- Crime data from 2003 for the US.
- Murder rate: The annual number of murders per 100,000 people in the population.
- Poverty: Percentage of residents with income below the poverty level.
- High school: Percentage of the adult residents who have at least a high school education.
- College: Percentage of the adult residents who have a college education.
- Single parent: Percentage of families headed by a single parent.

The outcome Y is the murder rate and the explanatory variable X college.
Always start by plotting the data!

Read data into R

03_Simple regression

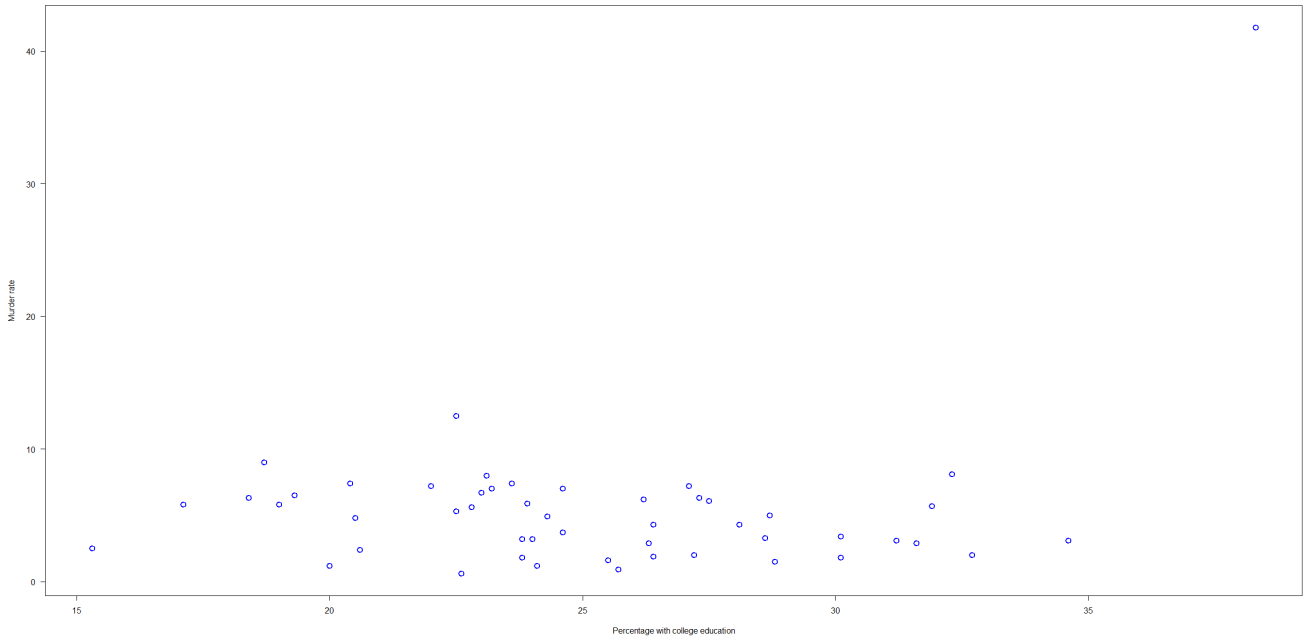
```
crime <- read.delim("us_statewide_crime.txt")
head(crime)
```

```
##          State violent.crime.rate murder.rate poverty high.school college
## 1      Alabama                486         7.4    14.7        77.5    20.4
## 2       Alaska                567         4.3     8.4        90.4    28.1
## 3      Arizona                532         7.0    13.5        85.1    24.6
## 4     Arkansas                445         6.3    15.8        81.7    18.4
## 5   California                622         6.1    14.0        81.2    27.5
## 6     Colorado                334         3.1     8.5        89.7    34.6
##   single.parent unemployed metropolitan
## 1          26.0        4.6         70.2
## 2          23.2        6.6         41.6
## 3          23.5        3.9         87.9
## 4          24.7        4.4         49.0
## 5          21.8        4.9         96.7
## 6          20.8        2.7         84.0
```

Scatter plot

03_Simple regression

```
plot(crime$college, crime$murder.rate, las=1, cex=1.5,col ="blue",
     lwd=2,ylab="Murder rate",
     xlab="Percentage with college education")
```

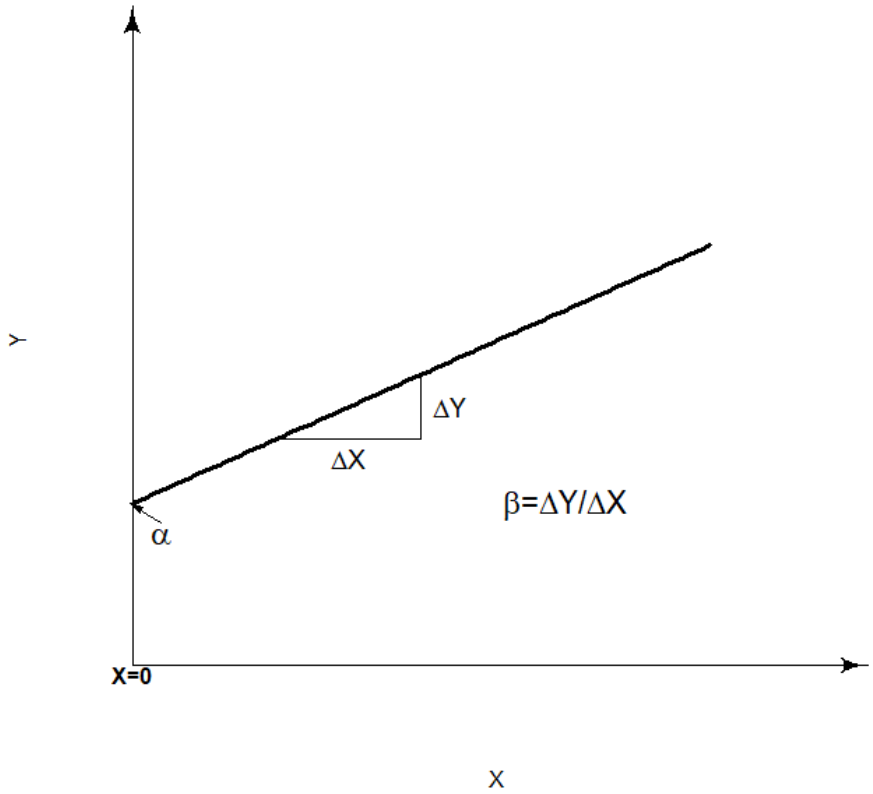


Mathematical Model

03_Simple regression

The expression for a straight line:

$$Y = \alpha + \beta X$$



The Parameters

03_Simple regression

- α: The intercept. The murder rate in a state with 0 percentage college education. Often a meaningless extrapolation! The intercept has the same units as the outcome.

The Parameters

03_Simple regression

- α: The intercept. The murder rate in a state with 0 percentage college education. Often a meaningless extrapolation! The intercept has the same units as the outcome.
- β: The slope, the regression coefficient. The difference in murder rate between two states with a difference of 1 percentage in college education. Often the main parameter of interest.

Statistical Model

03_Simple regression

$$Y_i = \alpha + \beta x_i + \varepsilon_i \qquad i = 1, \ldots, n \qquad \varepsilon_i \sim N(0, \sigma^2)$$

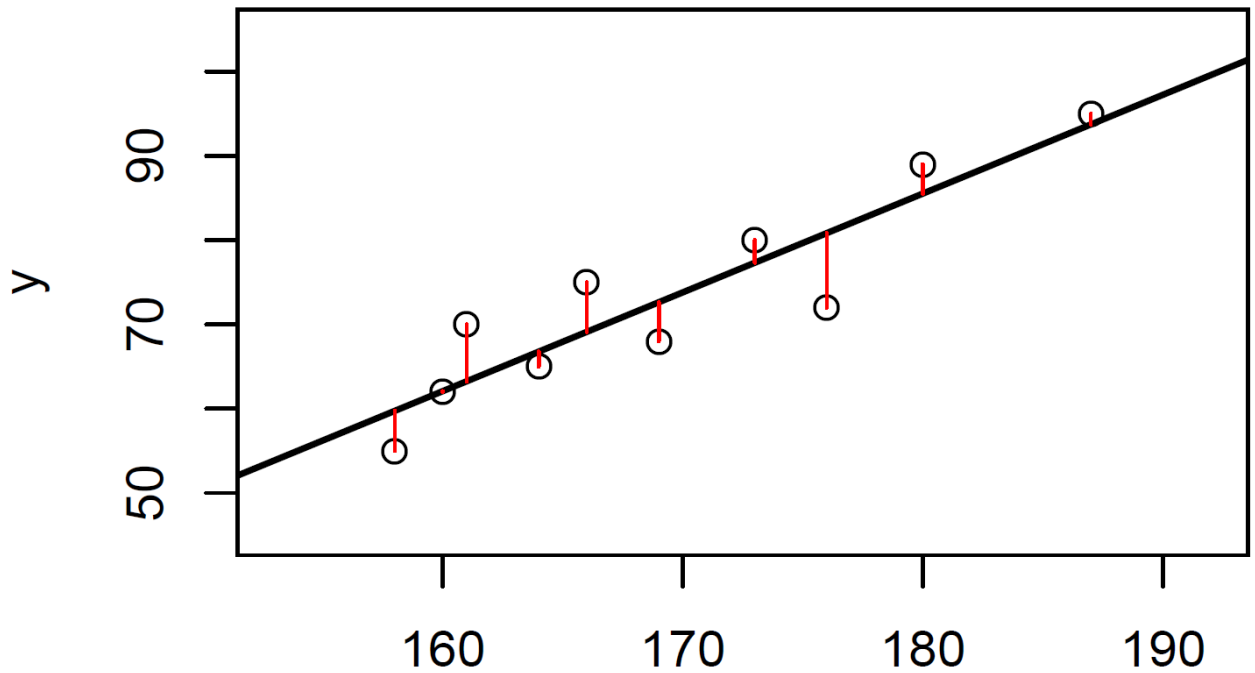
where

- Y_i is the response/dependent variable (random)
- α is the intercept (fixed, unknown)
- x_i is the covariate/independent variable (fixed)
- β is the slope (fixed, unknown)
- ε_i is the unobserved random error
- n is the number of observations (known)
- σ^2 is the residual variance (fixed, unknown)

Residuals

03_Simple regression

The ε_i are the difference between the observed Y_i and the expectation. The best fitting line is found by minimizing the squared residuals (the distances to the line - the estimated random errors). The observations are fixed, while the line varies.



Overview

03_Simple regression

Estimation

03_Simple regression

Find α and β so the squared distances to the line becomes as small as possible. I.e. minimize:

$$\sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

The resulting estimates:

$$\hat{\beta} = \frac{SP_{xy}}{SSD_x} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \ \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

where:
$\bar{x}$ is the average of x_i and $\bar{y}$ is the average of y_i.

Regression in R

03_Simple regression

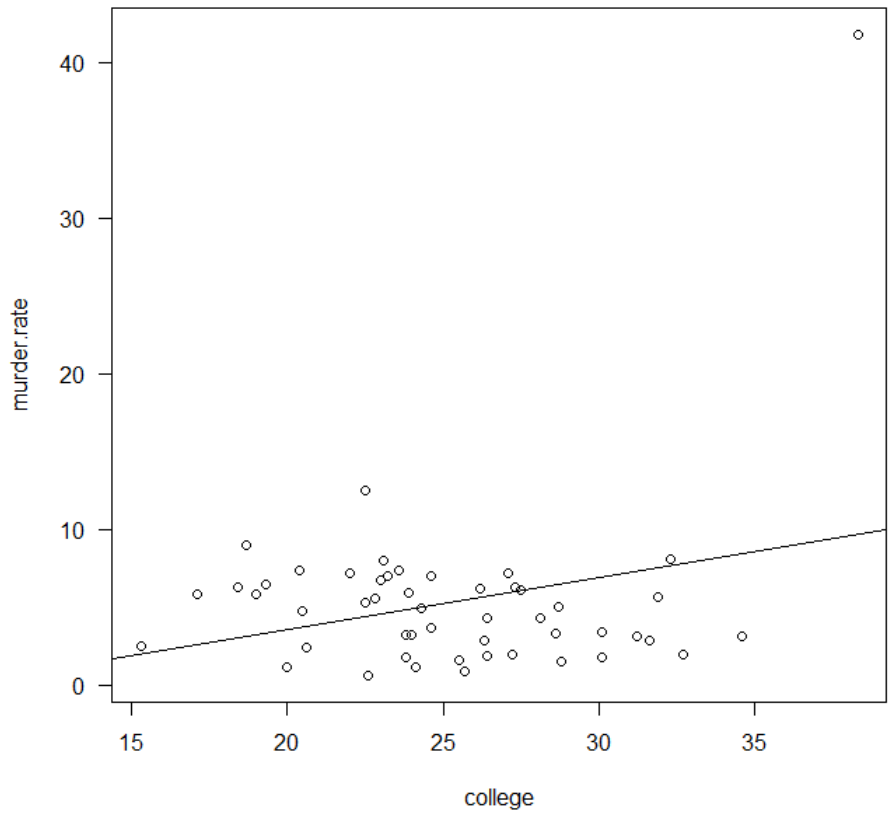
```
reg1 <- lm(murder.rate ~  college, data = crime)
summary(reg1)

##
## Call:
## lm(formula = murder.rate ~ college, data = crime)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.058      4.363   -0.70    0.487
## college        0.333      0.170    1.96    0.056 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.61 on 49 degrees of freedom
## Multiple R-squared:  0.0724,Adjusted R-squared:  0.0534
## F-statistic: 3.82 on 1 and 49 DF,  p-value: 0.0563
```

Estimated regression line

03_Simple regression

Estimated line: $\hat{y} = \hat{\alpha} + \hat{\beta}x$, murder.rate= -3.06 + 0.33 college

```
plot(murder.rate ~ college, data = crime, las=1)
abline(reg1)
```



3 Estimates from the regression analysis

03_Simple regression

- 2 estimates from the line (**intercept and slope**)
- The **variation** in y around the regression line (σ^2). Here the variation in murder rate for states with the same college percentage.

The estimate for σ^2 is:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

Nearly the average squared distance but n (the number of observations) is replaced by $n-2$ (the degrees of freedom). An estimator for σ is $s = \sqrt{s^2}$ which is denoted **residual standard error** in the output ($s = 5.61$).

Trust, but confirm - estimating parameters

03_Simple regression

```
# BY HAND
x <- crime$college
y <- crime$murder.rate
(beta <- sum((x-mean(x))*(y-mean(y)))/sum((x-mean(x))^2))

## [1] 0.33307

(alpha <- mean(y)- beta*mean(x))

## [1] -3.0581

Fitted <- alpha + beta*x
Resid <- y - Fitted
sigma2 <- sum(Resid^2)/(length(y)-2)
sqrt(sigma2)

## [1] 5.6146
```

Uncertainty of the Estimates

03_Simple regression

- How good are our estimates of the unknown parameters α and β?
- If we repeated the experiment how different would the estimates be?

Uncertainty of the Estimates

03_Simple regression

- How good are our estimates of the unknown parameters α and β?
- If we repeated the experiment how different would the estimates be?
- It may been shown that $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{SSD_x})$.
- The estimate of β does not vary so much if:

1. If σ^2 is small, i.e. the points are close to the line.
2. If SSD_x is large, i.e. the x values are spread out.

Confidence interval

The estimated uncertainty of $\hat{\beta}$ is called the standard error of $\hat{\beta}$:

$$SE(\hat{\beta}) = \frac{s}{\sqrt{SSD_x}}$$

We use the standard error to construct the 95% confidence interval

$$\hat{\beta} \pm t_{0.975}(n-2) \times SE(\hat{\beta})$$

Where $t_{0.975}(n-2)$ is the 97.5% percentile in the t-distribution with $n-2$ degrees of freedom.

Confidence interval

03_Simple regression

The estimated uncertainty of $\hat{\beta}$ is called the standard error of $\hat{\beta}$:

$$SE(\hat{\beta}) = \frac{s}{\sqrt{SSD_x}}$$

We use the standard error to construct the 95% confidence interval

$$\hat{\beta} \pm t_{0.975}(n-2) \times SE(\hat{\beta})$$

Where $t_{0.975}(n-2)$ is the 97.5% percentile in the t-distribution with $n-2$ degrees of freedom.

In our example: $0.3331 \pm 2.0096 \times 0.1703 = (-0.009; 0.675)$.

Test for $\beta = 0$

We can try to simplify the model, most often by:

$$H_0 : \beta = 0$$

I.e. no effect of the covariate (here college education). We use a t-test:

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})} \sim t(n-2)$$

Here $t = \frac{0.3331}{0.1703} = 1.956$ with a p-value:

$$2 \cdot P(T > |t|) = 2 \cdot (1 - P(T < |t|))$$

.

```
2*(1-pt(1.956, 49))
```

```
## [1] 0.056179
```

Tables of results in R

03_Simple regression

```
confint(reg1)

##                   2.5 %  97.5 %
## (Intercept) -11.8259006 5.70976
## college      -0.0092443 0.67538

# Nice table
tab <- cbind(coef(summary(reg1))[ , 1:2], "Lower" = confint(reg1)[ , 1],
             "Upper" = confint(reg1)[ , 2])
tab

##             Estimate Std. Error       Lower   Upper
## (Intercept) -3.05807    4.36303 -11.8259006 5.70976
## college      0.33307    0.17034  -0.0092443 0.67538

# Nice table with p-values
data.frame(round(tab, 2),
           "p-value" = format.pval(coef(summary(reg1))[ , 4], digits = 3, eps = 1e-3))

##             Estimate Std..Error  Lower Upper p.value
## (Intercept)    -3.06       4.36 -11.83  5.71  0.4867
## college         0.33       0.17  -0.01  0.68  0.0563
```

Overview

03_Simple regression

Model Check

03_Simple regression

- Is this a good model?

Model Check

03_Simple regression

- Is this a good model?
- We can only trust our conclusions if the model is appropriate.

Model Check

03_Simple regression

- Is this a good model?
- We can only trust our conclusions if the model is appropriate.
- We need to check:
  - Are the assumptions behind the model fulfilled?
  - Are there any influential observations that we need to check up on?

Model Check

03_Simple regression

- Is this a good model?
- We can only trust our conclusions if the model is appropriate.
- We need to check:
  - Are the assumptions behind the model fulfilled?
  - Are there any influential observations that we need to check up on?
- One would like to check earlier, but we need the estimates to be able to check.

Residual Analysis

03_Simple regression

The statistical model was:

$$Y_i = \alpha + \beta X_i + \epsilon_i, \ \epsilon_i \sim N(0, \sigma^2) \text{ independent}$$

We have to check:

- Normal residuals (observed - fitted)
- Variance homogeneity (one σ^2)
- Linear effect of X.

Notice that we have no assumption about normal distribution for X.

Residual Plots

03_Simple regression

We will do graphical model checks where the residuals are plotted:

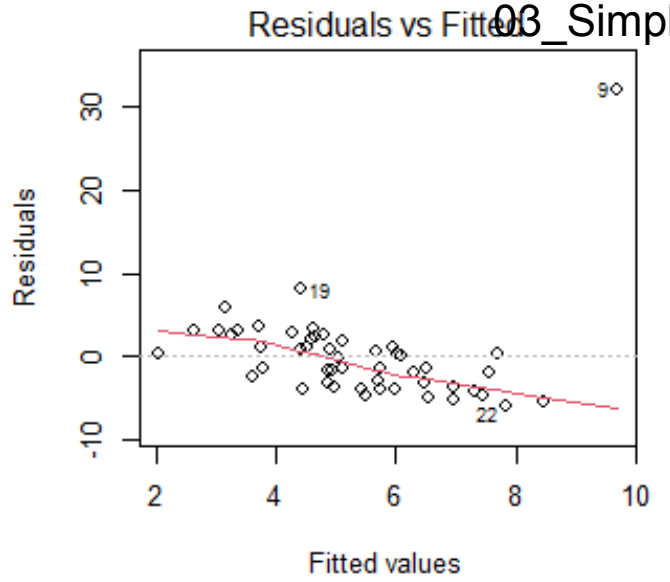
- Residuals plotted in qq-plots for normality (are the points close to straight line).
- Residuals plotted vs. fitted values ($\hat{y}_i$) to check variance homogeneity (Look for trumpet shape and other irregularities).
- Residuals plotted vs. explanatory variable (x_i) to check for linearity. (Look for curves or S-shape)

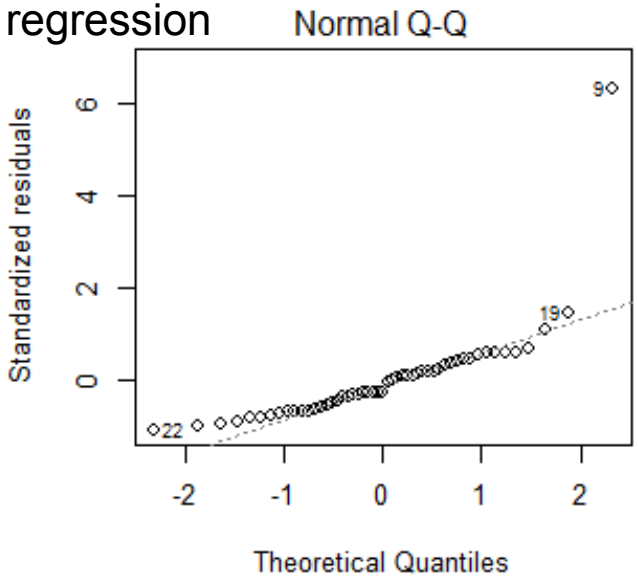
Influential Observations
03_Simple regression

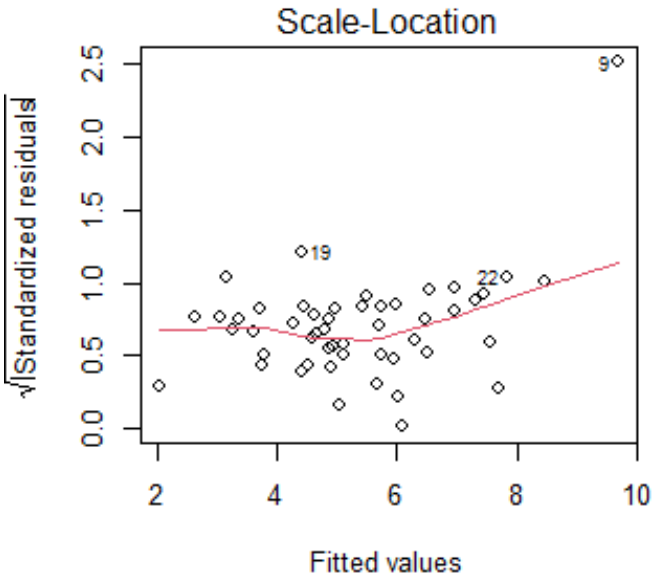
- May the results change if a few observations are left out, or are they insensitive to minor alterations of the data set?
- Influential observations are not necessarily problematic (althought the may be). They impact on the strength of the evidence. If the conclusion rests on for example the presence of a single observation, reservations need to be taken.
- Remove one observation at a time and see how results change.
- Calculate Cook's distance for each observation i. A measure of how much the results change if observation i is left out.
- Plot in **R**

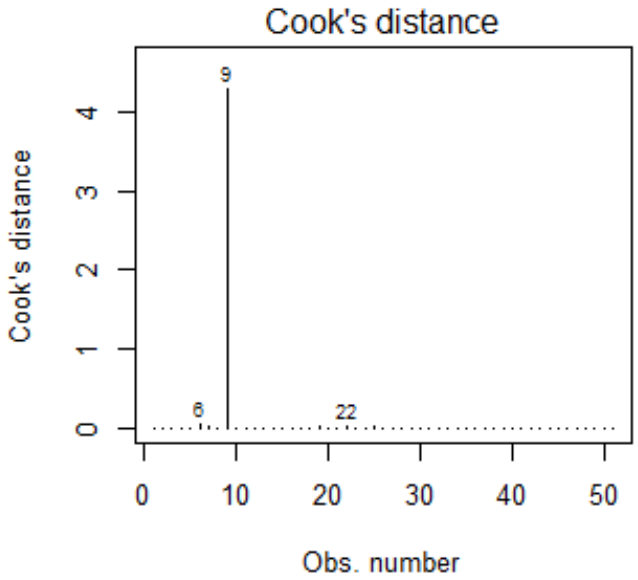
```
par(mfrow = c(2, 2))
plot(reg1, which = 1:4)
```

03_Simple regression









Model check in example 03_Simple regression

- Something is different about observation 9!
- It is the observation from Washington DC.
- Washington DC is the capital area of the USA; not a state like the rest.
- I will decide to leave it out, because it is structurally different from the rest, and I will remember that conclusions are not valid for DC.

New Data

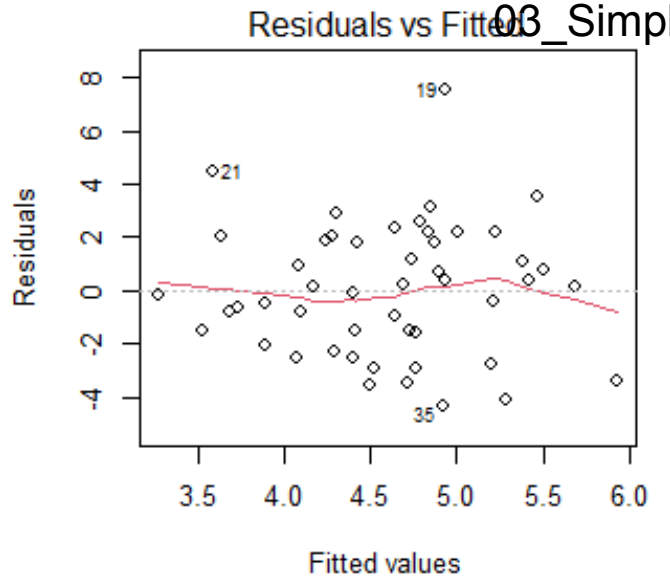
03_Simple regression

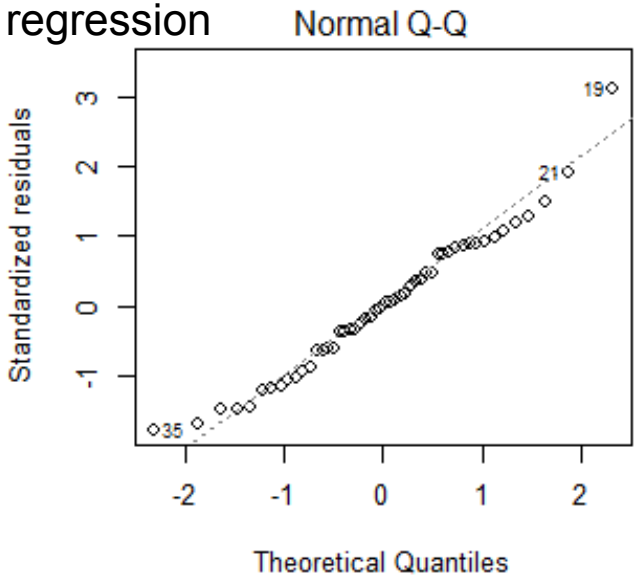
```
crime50 <- crime[-9, ]
head(crime50)
```

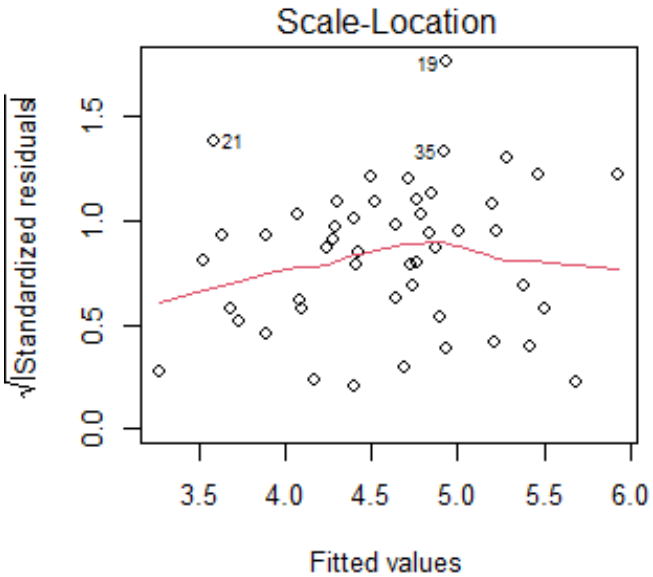
```
##        State violent.crime.rate murder.rate poverty high.school college
## 1    Alabama                486         7.4    14.7        77.5    20.4
## 2     Alaska                567         4.3     8.4        90.4    28.1
## 3    Arizona                532         7.0    13.5        85.1    24.6
## 4   Arkansas                445         6.3    15.8        81.7    18.4
## 5 California                622         6.1    14.0        81.2    27.5
## 6   Colorado                334         3.1     8.5        89.7    34.6
##   single.parent unemployed metropolitan
## 1          26.0        4.6         70.2
## 2          23.2        6.6         41.6
## 3          23.5        3.9         87.9
## 4          24.7        4.4         49.0
## 5          21.8        4.9         96.7
## 6          20.8        2.7         84.0
```

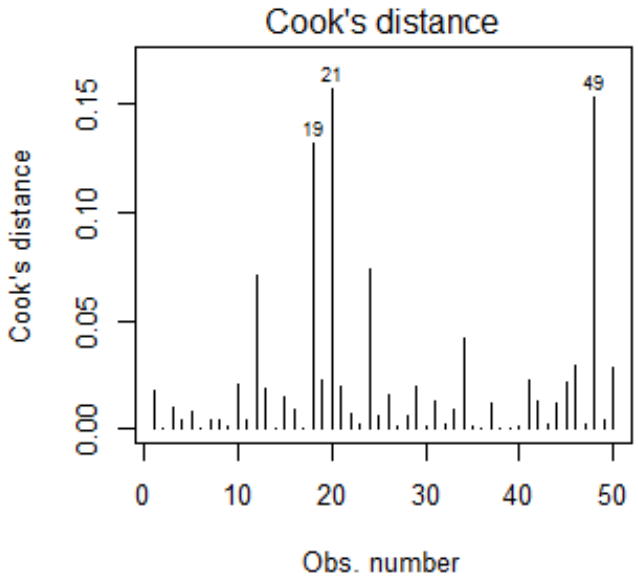
```
reg2 <- lm(murder.rate ~  college, data = crime50)
```

03_Simple regression





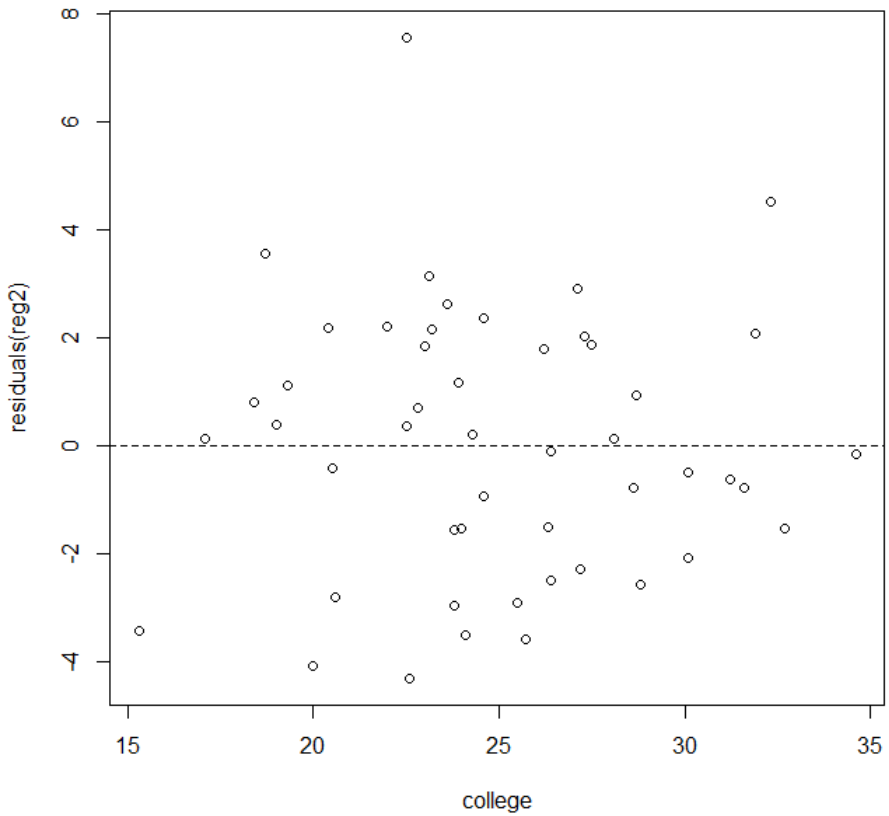




Check linearity

03_Simple regression

```
plot(residuals(reg2) ~ college, data = crime50,
     ylab = "Residuals")
abline(h = 0)
```



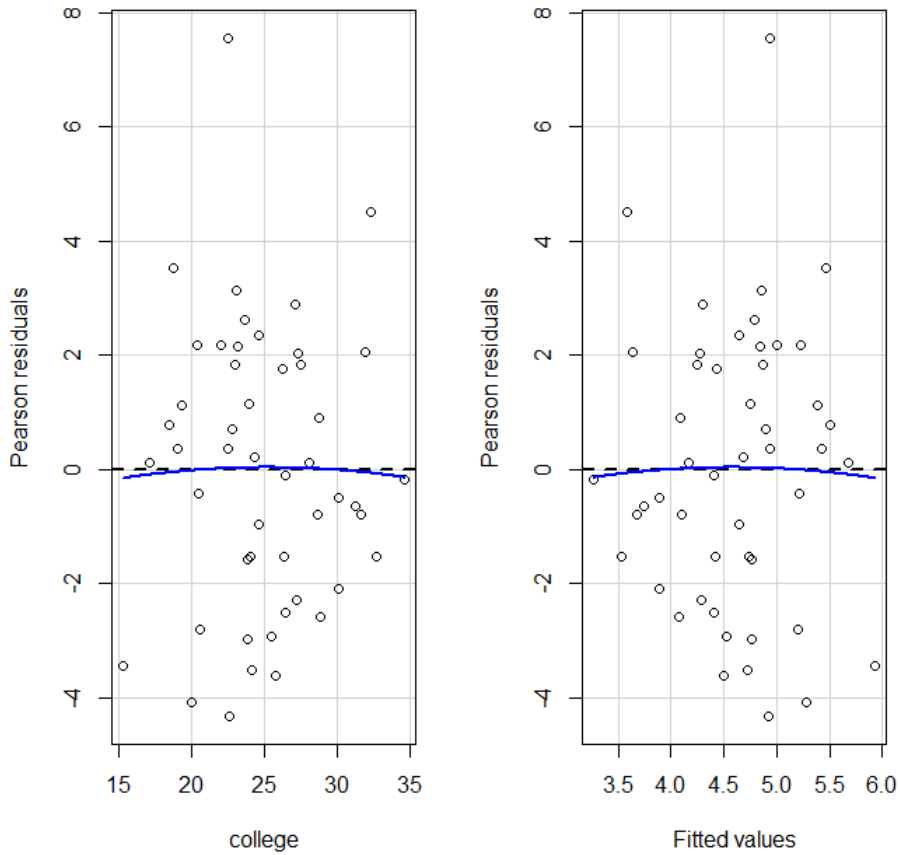
Check linearity, alternative using the car library

- Instead of making the plot ourselves we can use the package car (companion to applied regression).

```
library(car)
residualPlots(reg2)
```

Check linearity, alternative using the car library


```
##                Test stat Pr(>|t|)
## college           -0.128    0.899
## Tukey test        -0.128    0.898
```



Estimates for data without DC

03_Simple regression

```
summary(reg2)

##
## Call:
## lm(formula = murder.rate ~ college, data = crime50)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.0416     2.0648    3.89   0.0003 ***
## college      -0.1379     0.0816   -1.69   0.0977 .
##
## Residual standard error: 2.46 on 48 degrees of freedom
## Multiple R-squared:  0.0561,Adjusted R-squared:  0.0364
## F-statistic: 2.85 on 1 and 48 DF,  p-value: 0.0977
```

Estimates for data without DC in a table

03_Simple regression

```
##                Estimate Std. Error Lower Upper p.value
## (Intercept)        8.04       2.06  3.89 12.19  <0.001
## college           -0.14       0.08 -0.30  0.03  0.0977
```

- The fitted model has the form:

$$murder.rate_i = 8.04 - 0.14 \cdot college_i + \epsilon_i, \;\; N(0, 2.46^2)$$

Estimates for data without DC in a table

03_Simple regression

```
##                Estimate Std. Error Lower Upper p.value
## (Intercept)        8.04       2.06  3.89 12.19  <0.001
## college           -0.14       0.08 -0.30  0.03  0.0977
```

- The fitted model has the form:

$$murder.rate_i = 8.04 - 0.14 \cdot college_i + \epsilon_i, \ N(0, 2.46^2)$$

- **For every percentage increase in college education the murder rate decreases with 0.14** with 95% CI -0.03 to 0.30.

Estimates for data without DC in a table

03_Simple regression

```
##                Estimate Std. Error Lower Upper p.value
## (Intercept)        8.04       2.06  3.89 12.19  <0.001
## college           -0.14       0.08 -0.30  0.03  0.0977
```

- The fitted model has the form:

$$murder.rate_i = 8.04 - 0.14 \cdot college_i + \epsilon_i, \ N(0, 2.46^2)$$

- For every percentage increase in college education the murder rate decreases with 0.14 with 95% CI -0.03 to 0.30.
- College education is not statistically significant for the murder rate as 0 is in the confidence interval, and $p = 0.1$.

Estimates for data without DC in a table

03_Simple regression

```
##                Estimate Std. Error Lower Upper p.value
## (Intercept)        8.04       2.06  3.89 12.19  <0.001
## college           -0.14       0.08 -0.30  0.03  0.0977
```

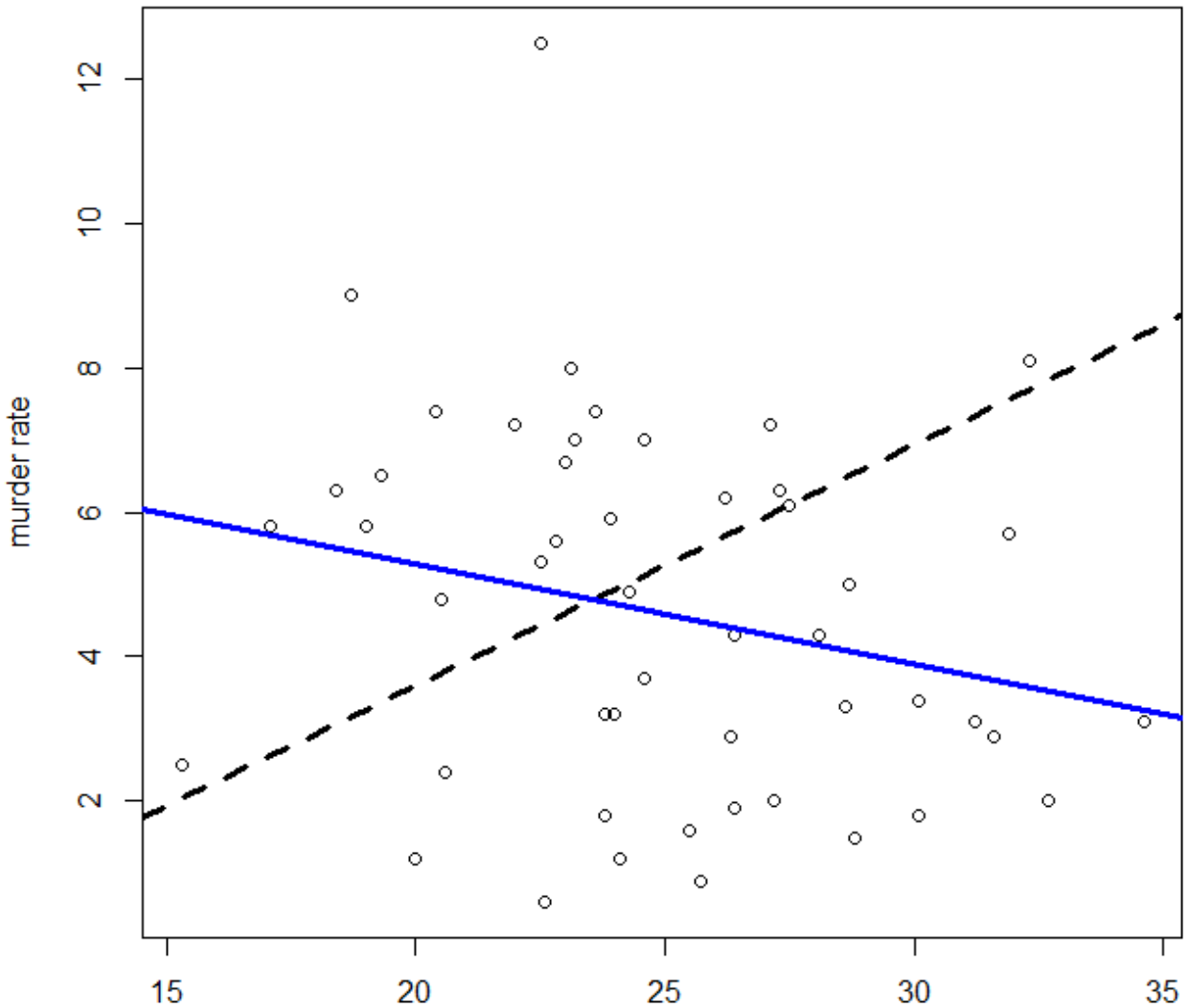
- The fitted model has the form:

$$murder.rate_i = 8.04 - 0.14 \cdot college_i + \epsilon_i, \ N(0, 2.46^2)$$

- For every percentage increase in college education the murder rate decreases with 0.14 with 95% CI -0.03 to 0.30.
- College education is not statistically significant for the murder rate as 0 is in the confidence interval, and $p = 0.1$.
- Washington DC was a very influential observation and changed the negative association to positive. However, DC is not a state and may behave in another fashion than the rest. This observation has therefore been left out of further analyses.

The Regression Lines

03_Simple regression



When is Regression Wrong?

03_Simple regression

Find a partner and explain to each other what is wrong with the use of regression in these examples.

- Winning times at the Boston marathon have followed a straight line decreasing trend from 160 minutes in 1927 to 130 minutes in 2004. After fitting a regression line you use the equation to predict that the winning time in 2023 will be about 123 minutes.
- A regression between x=years of education and y=annual income for 100 people shows a modest positive trend; person number 101 dropped out of school after the 10th grade and is now a multimillionaire. Since it is wrong to leave out data, we should report all results including this point resulting in a negative slope.

Overview

03_Simple regression

Confidence limits for regression line

03_Simple regression

The confidence interval for the fitted line at x_0:

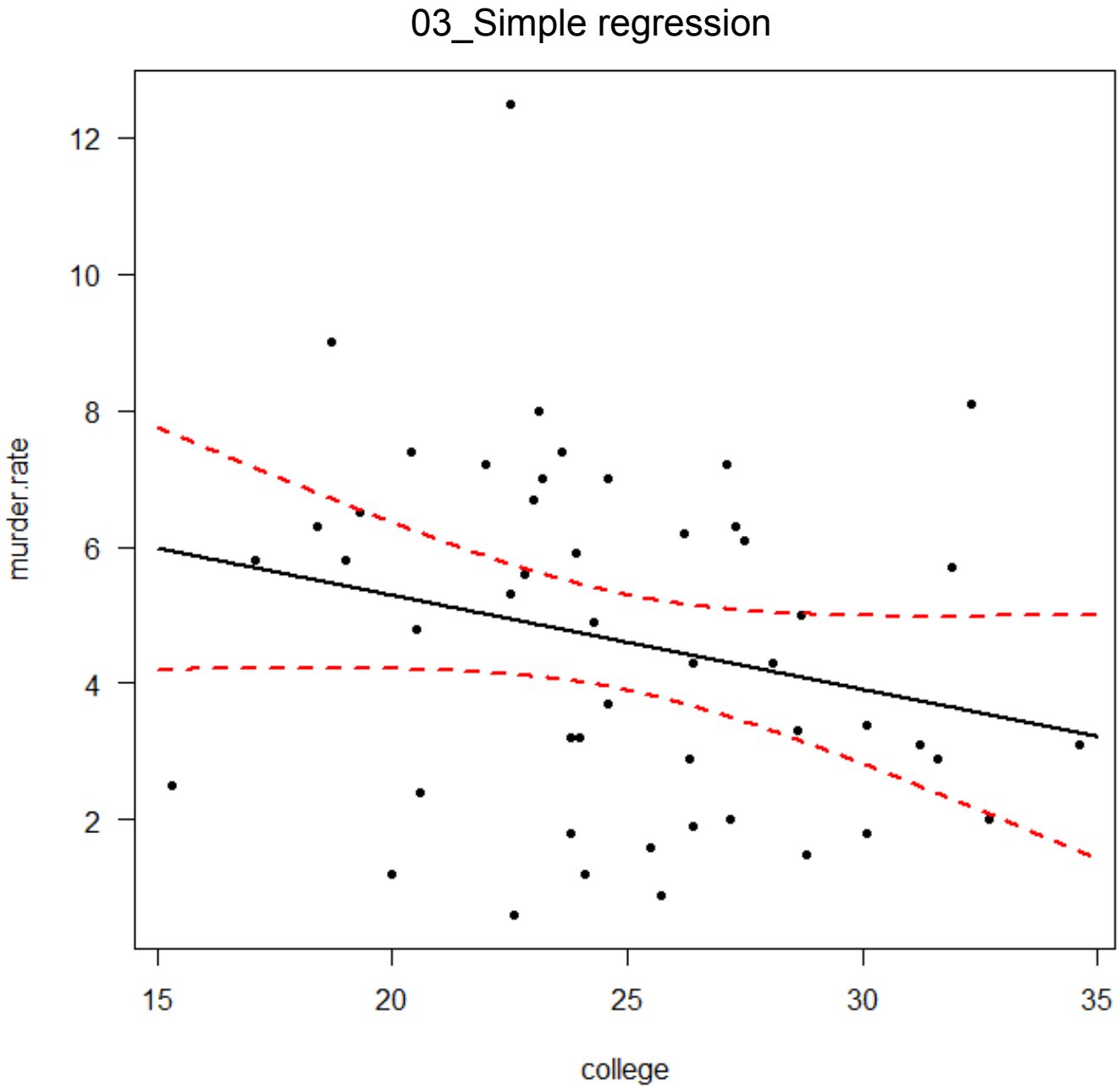
$$\hat{\alpha} + x_0\hat{\beta} \pm t_{\alpha/2}(n-2) \cdot s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SSD_x}}$$

- Expresses uncertainty about the fitted line/model.
- These limits become narrower when the number of observations is increased.
- The limits are narrowest close to the mean of x.

Confidence intervals with R

03_Simple regression

```
xval <- seq(from = 15, to = 35, length.out = 500)
newData <- data.frame(college = xval)
Pred.ci <- predict(reg2, newdata = newData,
                   interval = "confidence",
                   level = .95)

## Plot data, model and intervals:
plot(murder.rate ~ college, data = crime50, pch = 20, las = 1)
lines(xval, Pred.ci[, "fit"], lwd = 2)  ## or use: abline(reg2)
lines(xval, Pred.ci[, "lwr"], lty = 2, col = "red", lwd = 2)
lines(xval, Pred.ci[, "upr"], lty = 2, col = "red", lwd = 2)
```


03_Simple regression
murder.rate
college
12
10
8
6
4
2
15
20
25
30
35

Prediction limits for regression line

03_Simple regression

The prediction interval for the fitted line at x_0:

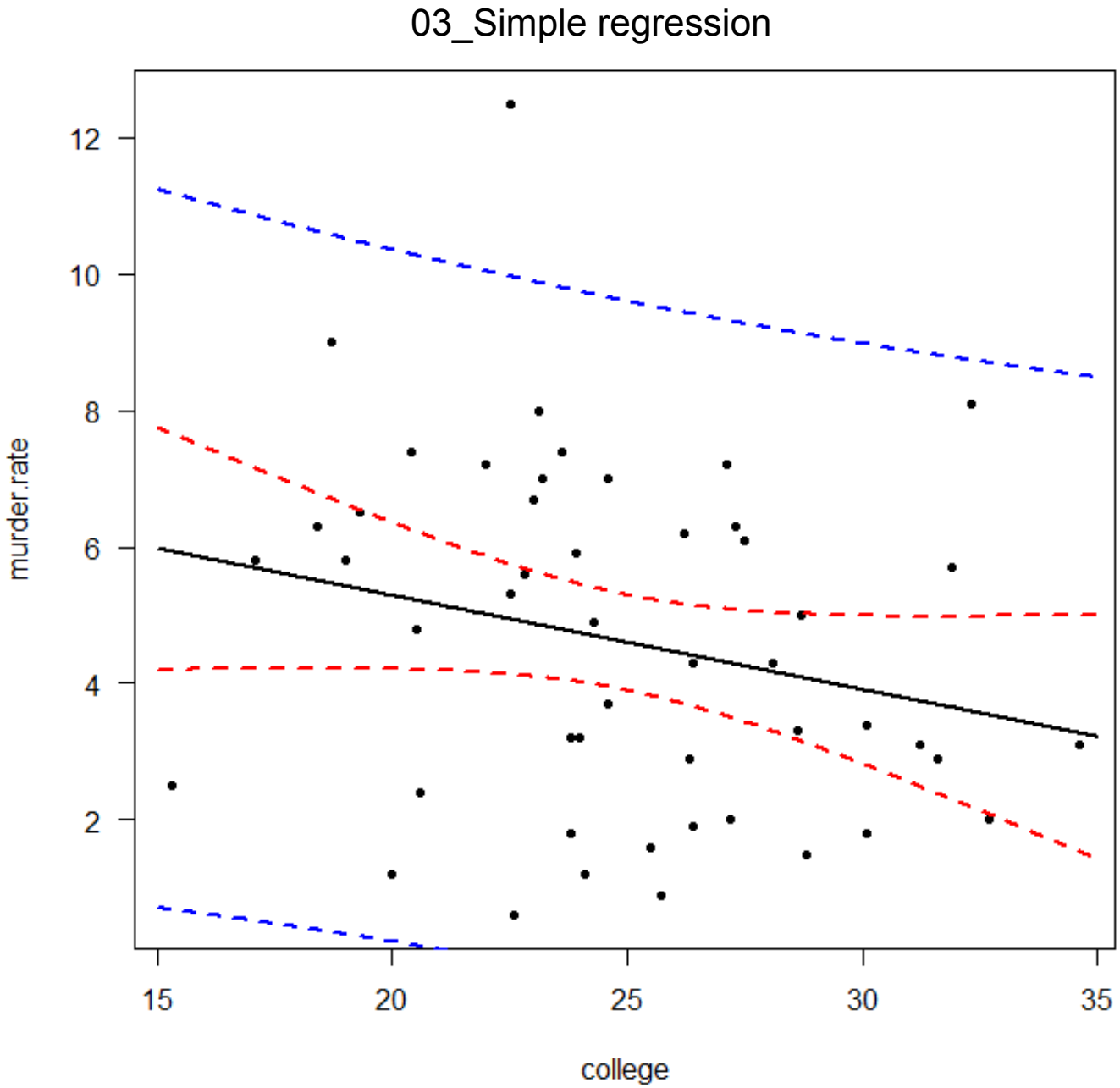
$$\hat{\alpha} + x_0\hat{\beta} \pm t_{\alpha/2}(n-2) \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}}$$

- Expresses uncertainty about the a single observation (x_0, y_0) when x_0 is known.
- Is used to decide whether a new observation is atypical, as the limits will include approximately 95% of future observations.
- The region where we would expect to see 95% of future observations.

Computing prediction intervals with R

03_Simple regression

```
## Prediction interval for a new observation:
Pred.pi <- predict(reg2, newdata = newData,
                   interval = "prediction")
## Add prediction intervals to plot:
lines(xval, Pred.pi[, "lwr"], lty=2,
      col="blue", lwd=2)
lines(xval, Pred.pi[, "upr"], lty=2,
      col="blue", lwd=2)
## could add legend here.
```


03_Simple regression
murder.rate
college
12
10
8
6
4
2
15
20
25
30
35

Prediction and Confidence Limits

03_Simple regression

- The narrow limits are confidence limits:
  - Corresponds to standard error.
  - Is used to assess the uncertainty of the estimate.
  - Depends a lot on the value of x_0.

Prediction and Confidence Limits

03_Simple regression

- The narrow limits are confidence limits:
  - Corresponds to standard error.
  - Is used to assess the uncertainty of the estimate.
  - Depends a lot on the value of x_0.
- The broad limits are prediction limits:
  - Corresponds to standard deviation.
  - Also called reference range.
  - Is used to assess individual observations.
  - Approximately calculated as ± 2 residual standard error.

Overview

03_Simple regression

What do we do if the model does not fit?

03_Simple regression

- In our example we had an influential observation, but without this observation the model had a good fit.
- We were checking:
  - Normality of residuals.
  - Variance homogeneity.
  - Linear effect of covariate.

Example about timber hardness

03_Simple regression

- We have 36 observations of timber hardness and density from some trees in Australia.
- The aim of the study was to estimate parameters that define the relationship between timber hardness and density, so we could predict (unknown) timber hardness in future samples based on the measured density.
- Our first model could be a simple regression:

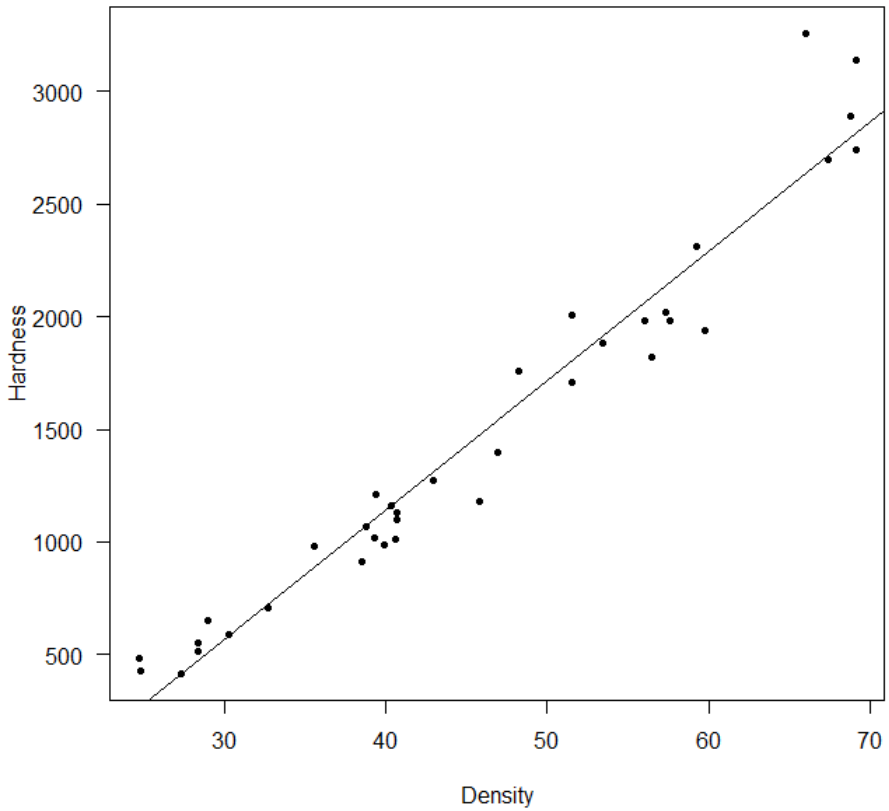
$$Hardness_i = \alpha + \beta Density_i + \epsilon_i, \ \epsilon_i \sim N(0, \sigma^2)$$

- But does this model fit?

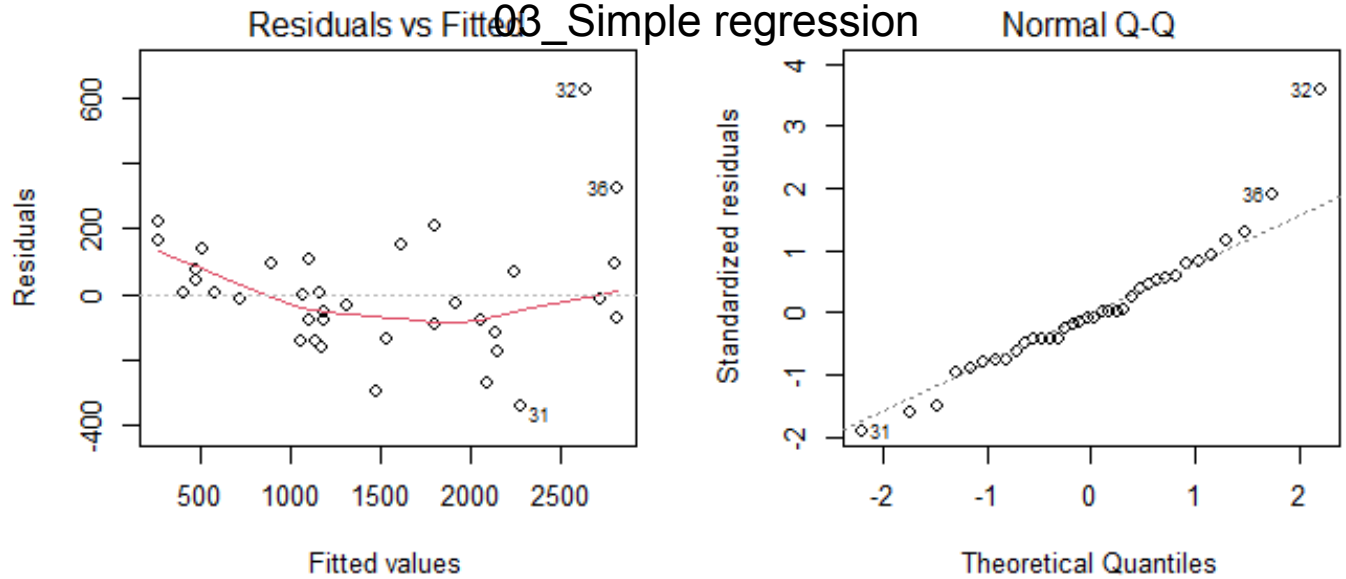
The Janka Example

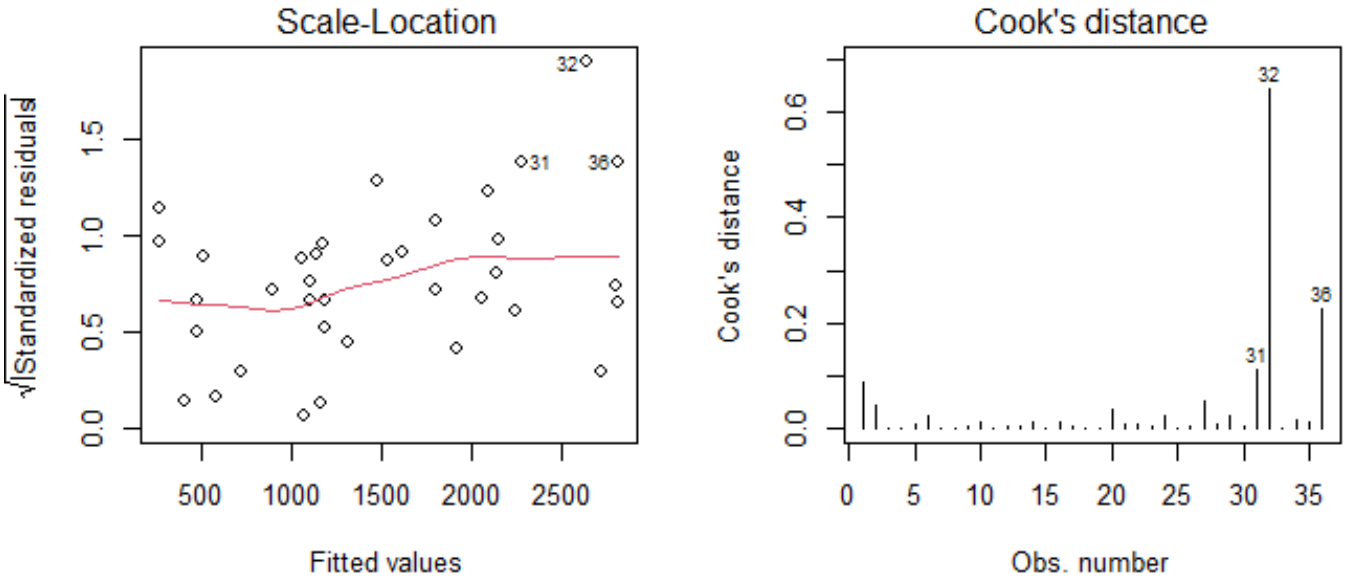
03_Simple regression

```
reg3 <- lm(Hardness ~ Density, data = janka)
plot(Hardness ~ Density, data = janka, pch = 20, las = 1)
abline(reg3)
```



03_Simple regression

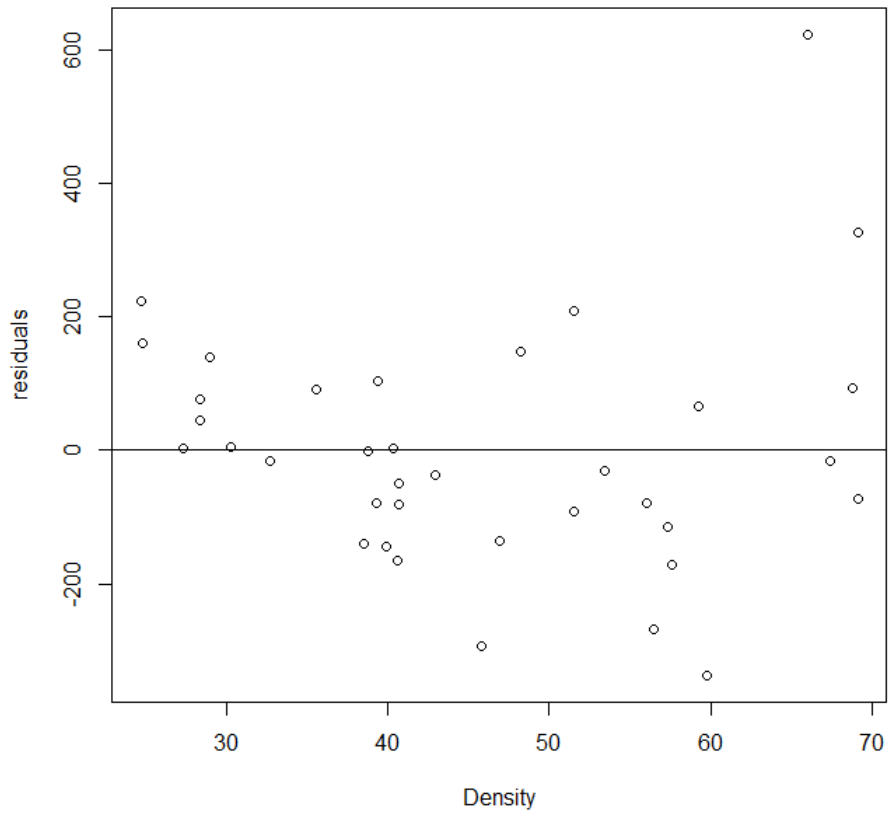




Check linearity

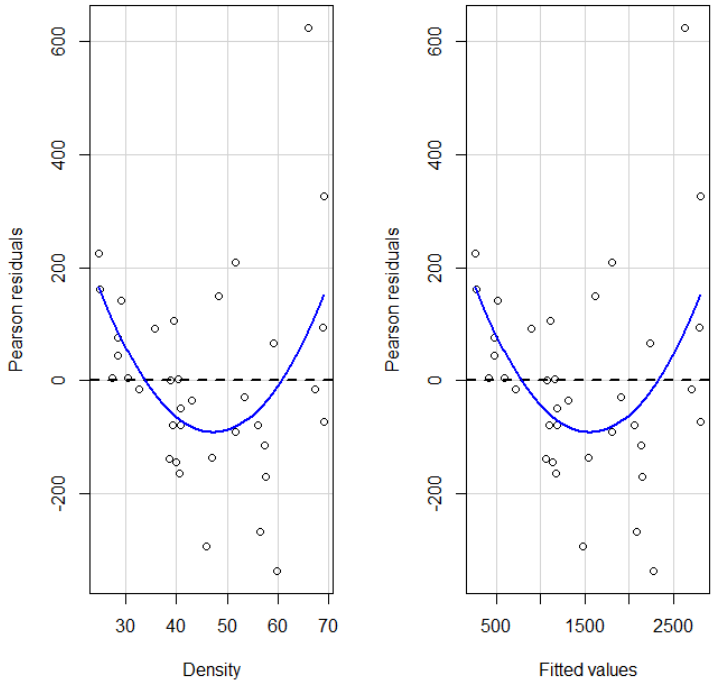
03_Simple regression

```
plot(residuals(reg3) ~ Density, data = janka,
     ylab = "residuals")
abline(h = 0)
```



Check linearity, using the car library


```
library(car)
residualPlots(reg3)

##               Test stat Pr(>|t|)
## Density           3.248    0.003
## Tukey test        3.248    0.001
```



Model check plots, what are we looking for?

03_Simple regression

1. Plot of Residuals vs Fitted Variance homogeneity. Problem if we see a trumpet shape, which we do here. We can also look for non-linear shapes - apart from the trumpet shape a curve is present.
2. Normal Q-Q Residuals have to follow a normal distribution. The observations should follow a straight line. They do here (sort of); this kind of model abberation cannot be detected by a qq-plot.
3. Scale-Location Variance homogeneity. The line should be horizontal (supplements plot 1). We see a slight increase.
4. Cook's distance Shows potentially influential observations.
5. Residual plots for the linear effect This should look like a random scatter of dots. Looking for shapes e.g. polynomial, here we see a curved association.

Problems with normality 03_Simple regression

Normally distrubuted residuals

If the normal distribution is not valid then we lose power and the prediction limits are not valid.

What can we do?

- If we have a heavy tail to the right, then try transfoming the response with the logarithm. We will get back to alternative transformations in an exercise on Thursday.
- Non-parametric methods, simulation

Problems with Variance Homogeneity

03_Simple regression

Variance Homogeneity

If the the variance is not constant then we lose power and the prediction limits are not valid.

What can we do?

- If we have a trumpet shape in the residual plot, then try transfoming the response with the logarithm.

Problems with Variance Homogeneity

Variance Homogeneity

If the the variance is not constant then we lose power and the prediction limits are not valid.

What can we do?

- If we have a trumpet shape in the residual plot, then try transfoming the response with the logarithm.

Trumpet shape

- Trumpet shapes are often seen when measuring small positive values, e.g. concentrations.
- Can be regarded as a constant relative uncertainty, constant coefficient of variation.

$$\text{Coefficient of variation} = \frac{\text{standard deviation}}{\text{mean}}$$

Normality and Variance Homogeneity

03_Simple regression

- The assumption about normality (and variance homogeneity) are not so crucial for the estimates (NB: It is very crucial for statistical inference in general, when the models are more complicated).
- Here we are using the t-distribution for $\hat{\beta}$ and this needs normality but the Central Limit Theorem helps.
- Central Limit Theorem: sums of more and more observations approaches normality.
- The prediction limits cannot be used (concern single observations).

Problems with Linearity 03_Simple regression

Linearity

If we have assumed linearity without it being approximately ok then we cannot interpret the model. What can we do?

- If we see a curve in the residual plots we can add our covariate squared, cubed...
- Transform variables with log, square root, inverse...

Polynomial regression models

03_Simple regression

What if the relation between y and x is not a straight line?

We can use a pth order polynomial to fit curves:

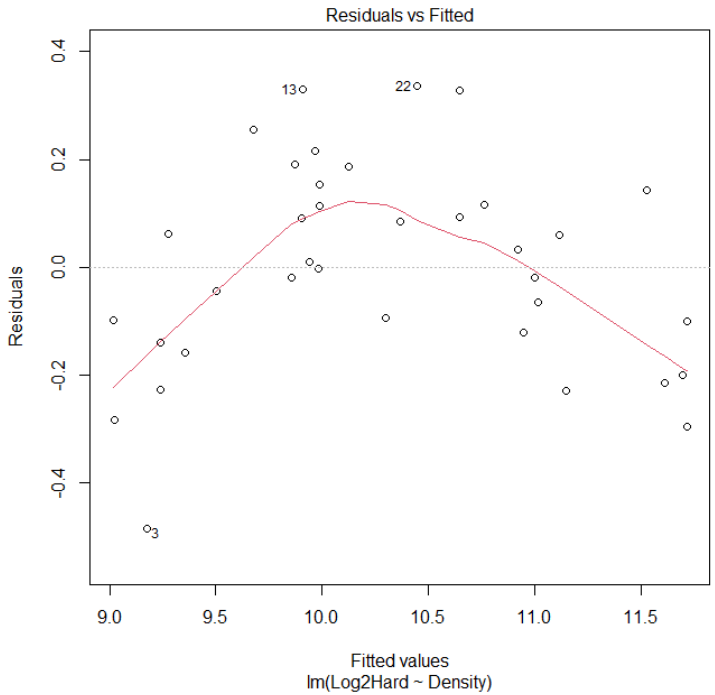
$$Y_i = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + \ldots + x_i^p\beta_p + \varepsilon_i$$

This is still a linear model since it is linear the parameters, β.

The Janka Example, a better model

03_Simple regression

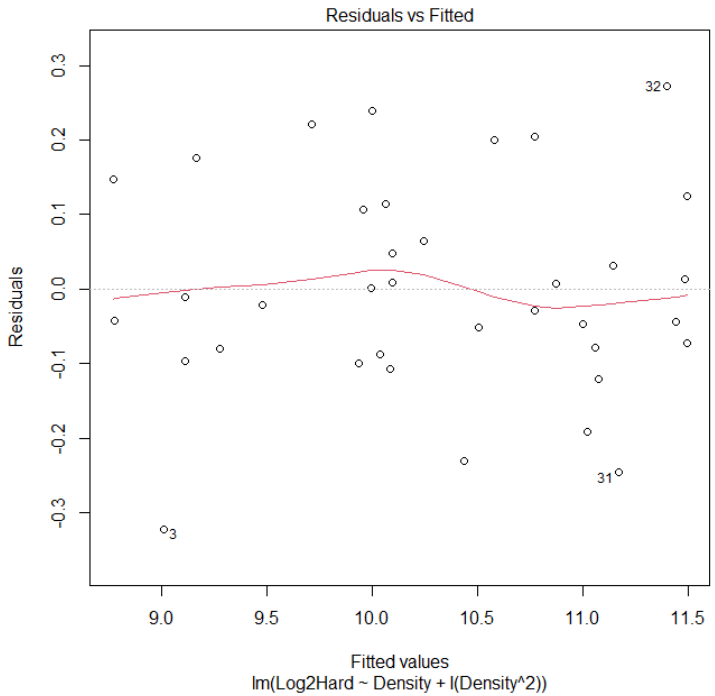
```
janka$Log2Hard <- log2(janka$Hardness)
reg4 <- lm(Log2Hard ~ Density, data = janka)
plot(reg4, which = 1)
```



The Janka Example, Adding a Squared Term

03_Simple regression

```
#TRANSFORM THE OUTCOME WITH LOG2 AND ADD SQUARED DENSITY
reg5 <- lm(Log2Hard ~ Density + I(Density^2), data = janka)
plot(reg5, which=1)
```



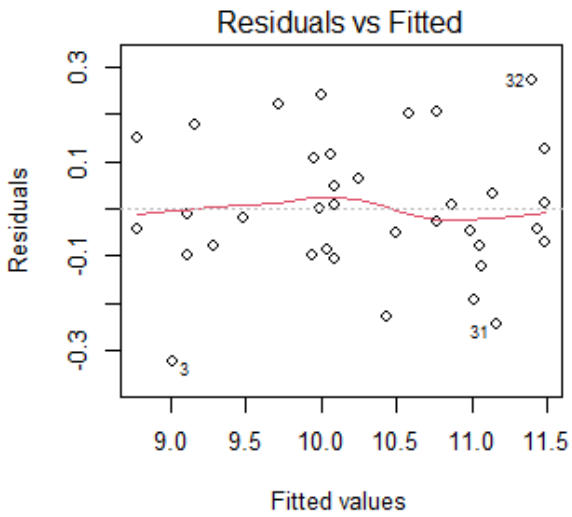
The Janka Example, Final Model

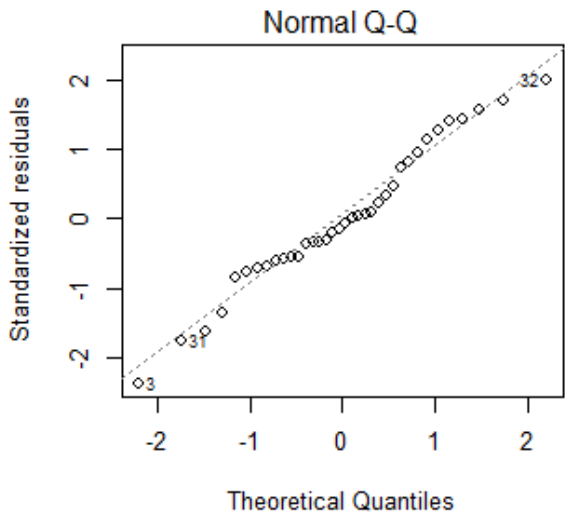
03_Simple regression

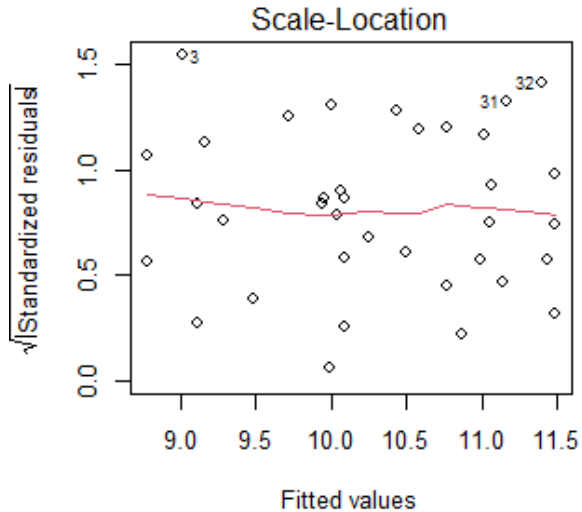
```
# LOOK AT THE RESULTS
summary(reg5)

##
## Call:
## lm(formula = Log2Hard ~ Density + I(Density^2), data = janka)
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.969895   0.301078   19.83  < 2e-16 ***
## Density        0.132029   0.013425    9.83  2.5e-11 ***
## I(Density^2) -0.000754   0.000141   -5.35  6.5e-06 ***
##
## Residual standard error: 0.145 on 33 degrees of freedom
## Multiple R-squared:  0.972,Adjusted R-squared:  0.971
## F-statistic:  579 on 2 and 33 DF,  p-value: <2e-16
```

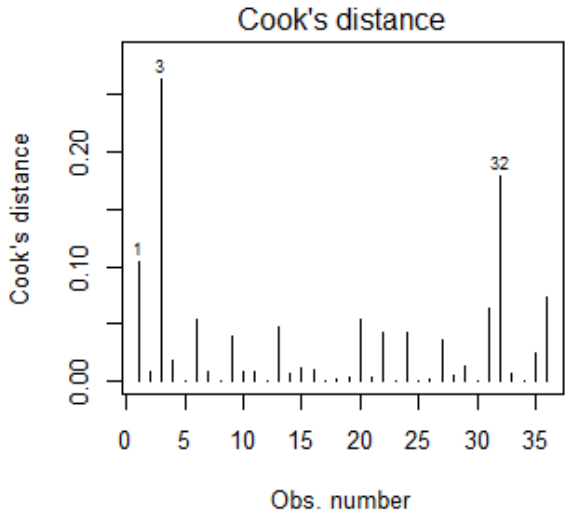
The Janka Example, Final Model

03_Simple regression









The Janka Example, Final Model

03_Simple regression

We have estimated

$$Log2Hardness_i = 5.97 + 0.13Density - 0.001Density^2 + \varepsilon_i$$

But this is on the log2 scale. We will want to see the results on the original scale.

The Janka Example, Final Model

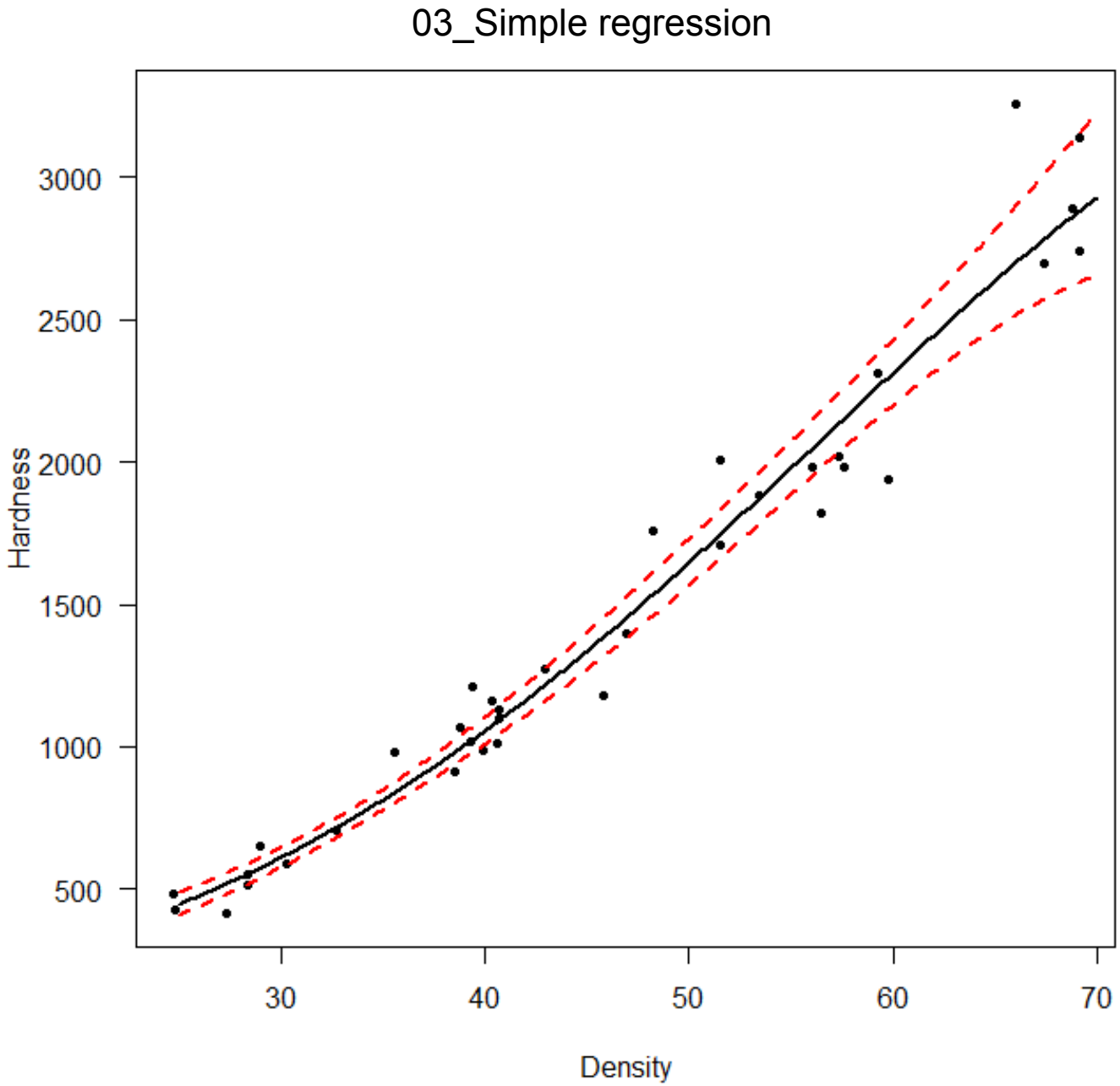
03_Simple regression

We have estimated

$$Log2Hardness_i = 5.97 + 0.13Density - 0.001Density^2 + \varepsilon_i$$

But this is on the log2 scale. We will want to see the results on the original scale.

```
xval <- seq(from = 25, to = 70, length.out = 500)
newData <- data.frame(Density = xval)
Pred.ci <- predict(reg5, newdata = newData,
                   interval = "confidence", level = .95)

# ON THE ORIGINAL SCALE, WE USED log2
plot(Hardness ~ Density, data = janka, pch = 20, las = 1)
lines(xval, 2^Pred.ci[, "fit"], lwd = 2)
lines(xval, 2^Pred.ci[, "lwr"], lty = 2, col = "red", lwd = 2)
lines(xval, 2^Pred.ci[, "upr"], lty = 2, col = "red", lwd = 2)
```


03_Simple regression
Hardness
Density
3000
2500
2000
1500
1000
500
30
40
50
60
70

Dos and don'ts with polynomial regression

03_Simple regression

- It is often a good idea to center x: $x' = (x - \bar{x})$ before fitting the models
- Retain the simplest order polynomial that fits the data (reasonably) well
- But: retain β_{p-1} in the model if β_p is present

Overview

03_Simple regression

Exercises for "simple" linear regression

03_Simple regression

1. PCB in trouts
   ⇒ Transformations, multiplicative models, residuals, linear regression models in R
2. Brain weight (If more time)
   ⇒ Transformations outliers predictions