Linear and generalised linear Regression

Organizational Issues

Alexandra Posekany

SS 2020

Linear Regression - simple univariate model

(linear) regression models the dependence between

- a **dependent** numeric variable, **regressand** Y, and
- one or more **independent** explanatory numeric variables, **regressor(s)** X, $\boldsymbol{X}$

Mathematically, the simple linear regression model is

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

- α and β are unknown parameters of the population
- ε_i are iid errors with mean 0 and a common unknown variance σ^2 (no heteroscedasticity).

Regression - Correlation

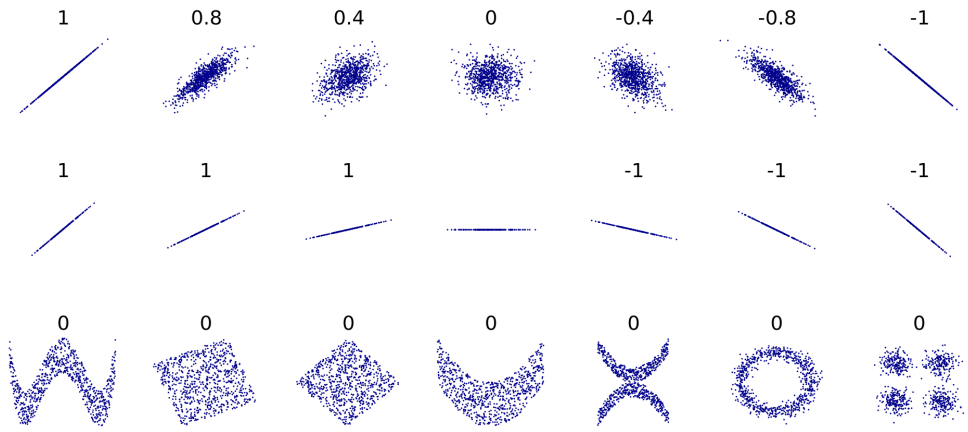
Covariance σ_{xy} and **correlation** ρ_{xy} measure linear dependence of simple linear regression, their multivariate analoga are the **covariance matrix** $Cov(\boldsymbol{X})$ and **correlation matrix** $Cor(\boldsymbol{X})$.

$$Cov(\boldsymbol{X}) = \begin{pmatrix} \sigma_1^2 & \sigma_{x_1x_2} & \dots & \sigma_{x_1x_n} \\ \sigma_{x_2x_1} & \sigma_2^2 & \dots & \sigma_{x_2x_n} \\ & & \vdots & \\ \sigma_{x_nx_1} & \sigma_{x_nx_2} & \dots & \sigma_n^2 \end{pmatrix},$$

$$Cor(\boldsymbol{X}) = \begin{pmatrix} 1 & \rho_{x_1x_2} & \dots & \rho_{x_1x_n} \\ \rho_{x_2x_1} & 1 & \dots & \rho_{x_2x_n} \\ & & \vdots & \\ \rho_{x_nx_1} & \rho_{x_nx_2} & \dots & 1 \end{pmatrix}$$

R cov(x), cor(x)

Visualising Correlation



Testing for Correlation

If a justification for 'non-zero' correlation is required, one can test for correlation with a t-test for the two-sided hypothesis

$$\begin{aligned} H_0: \quad r_{xy} &= 0 \\ H_A: \quad r_{xy} &\neq 0 \end{aligned}$$

```
cor.test(x, y,
         alternative = c("two.sided", "less", "greater"),
         method = c("pearson", "kendall", "spearman"),
         conf.level = 0.95)
```

Remember that this is tested in every regression model and the model is always more valuable than the test!

Data transformations

Often the dependent variable has no linear relation to the independent variable(s), but a more complex mathematical relation which can however be obtained by applying a single mathematical function to either. These **data transformations**

$$\begin{aligned} \tilde{Y} &= f(Y) \\ \tilde{\boldsymbol{X}} &= g(\boldsymbol{X}) \end{aligned}$$

of the regressand Y and/or the regressor(s) $\boldsymbol{X}$ can assure a linear relation, where f and g are suitable transformation functions.

Linear Transformations

Basic linear data transformations for $X_1, X_2, \ldots, X_n$.

- W_i are a **translation** of the X_i, if
$$W_i = X_i + b$$
- Y_i are a **scaling** of the X_i, if
$$Y_i = cX_i$$
- Z_i are a combination of a specific translation and scaling, called **normalisation** or **standardisation** of the X_i, if
$$Z_i = (X_i - \bar{X})/sd(X)$$
Note that Z_i have mean 0 and variance 1.

Non-linear Transformations

P_i are a **polynomial** transformation to the power of k of data X_i, if

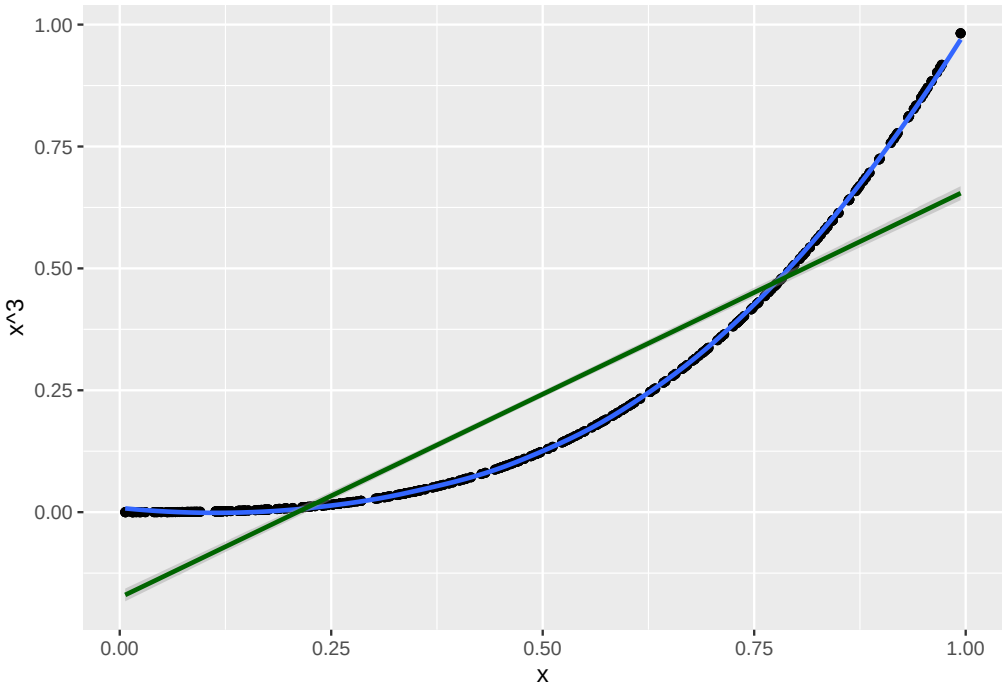
$$P_i = X_i^k$$

most importantly, the quadratic transformation is

$$Q_i = X_i^2$$

\begin{minipage}{10cm} \begin{center}

```
ggplot(dfpoly, aes(x = x, y = x^3)) + geom_point() + geom_smooth(method = "loess") +
    geom_smooth(method = "lm", col = "darkgreen")
```

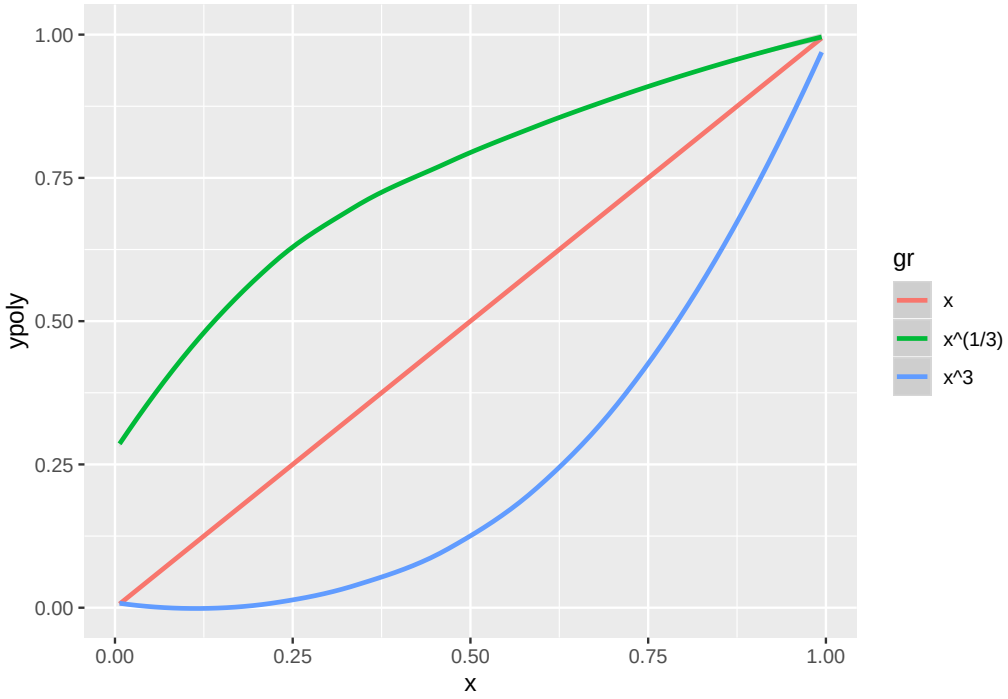


\end{center}

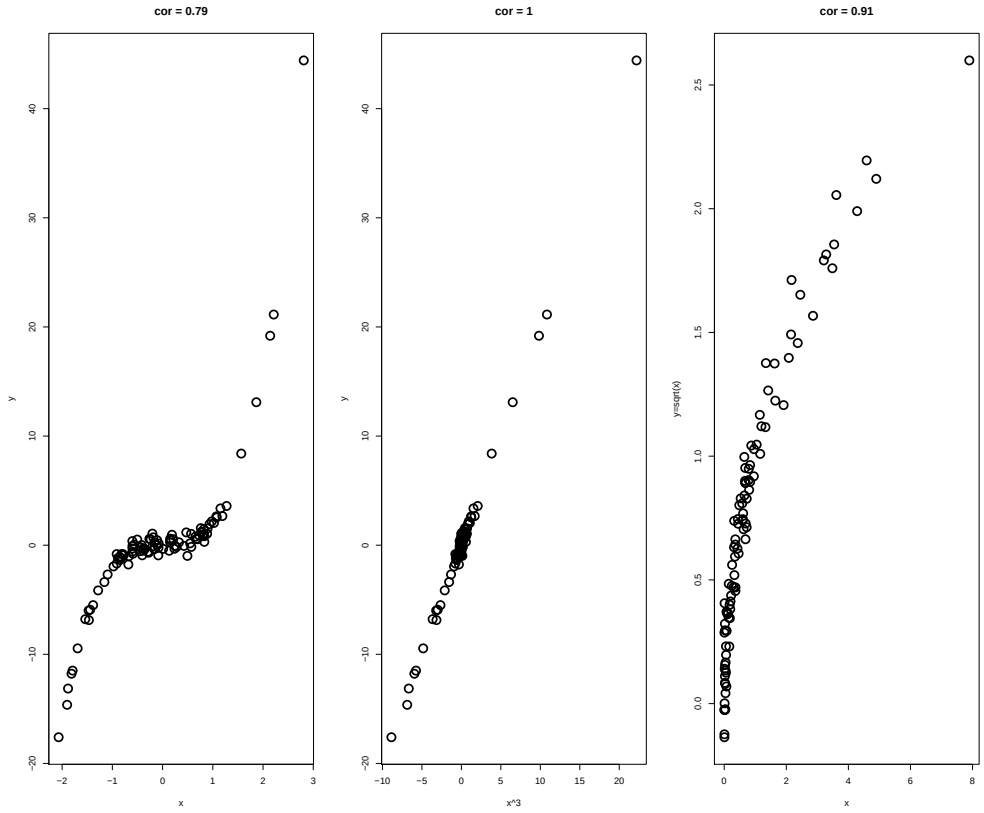
\end{minipage}

Root and Polynomials

```
ggplot(dfpoly, aes(x = x, y = ypoly, col = gr)) + geom_smooth(method = "loess")
```



Non-linear Transformations



Non-linear Transformations

- E_i are a **exponential** transformation of data X_i, if

$$E_i = exp(X_i)$$

- L_i are a **logarithmic** transformation of data X_i, if

$$L_i = log(X_i)$$

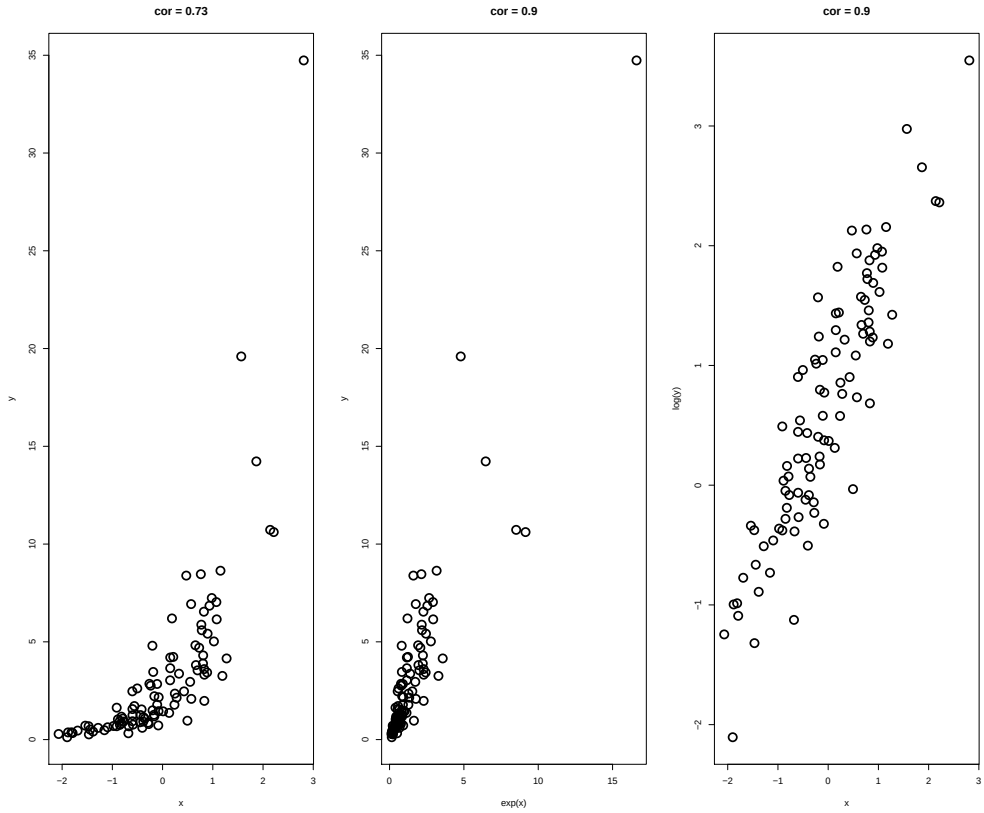
These two transformations form the bridge between the class of exponential models

$$Y_i = C \cdot exp(\beta \boldsymbol{X}_i) \cdot \epsilon_i$$

and linear models, as

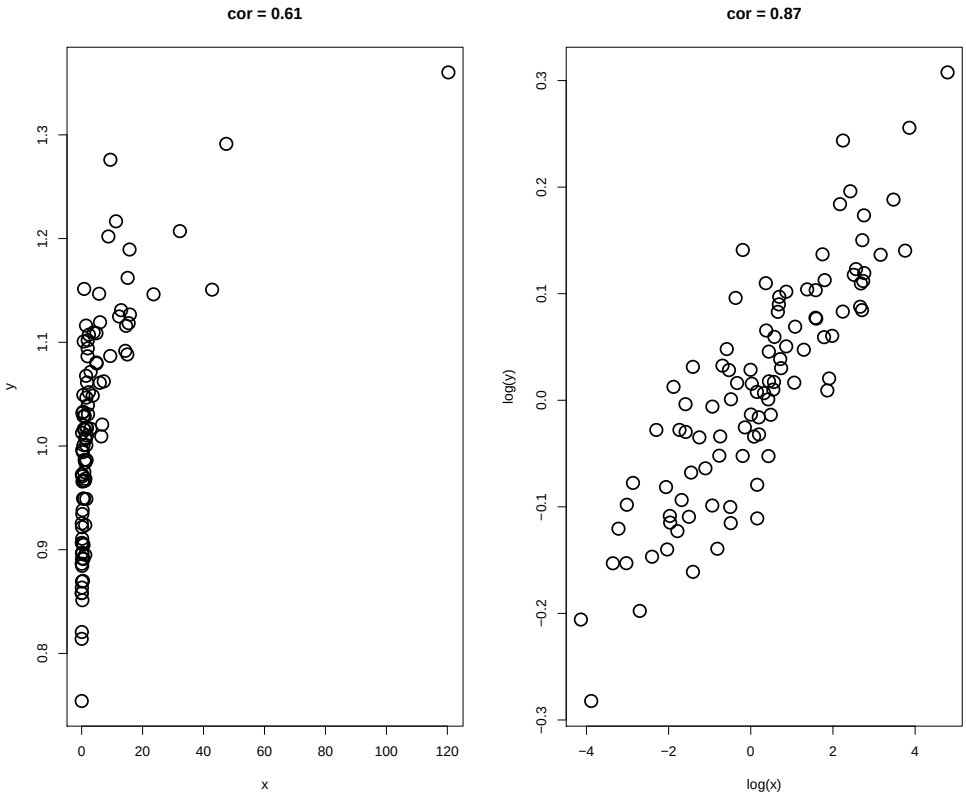
$$\begin{aligned} \log(Y_i) &= \log(C \cdot exp(\beta \boldsymbol{X}_i) \cdot \epsilon_i) \\ &= \log(C) + \log(exp(\beta \boldsymbol{X}_i)) + \log(\epsilon_i) \\ &= \tilde{\alpha} + \beta \boldsymbol{X}_i + \tilde{\varepsilon}_i. \end{aligned}$$

Non-linear Transformations



Non-linear Transformations

Logarithmically transforming both variables (a "log/log" plot) can reduce both heteroscedasticity and skewness:



Regression model vs. model equation

The regression model

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

explains the observed values y_i.

Once a specific line is selected, we obtain an actual model equation of the "solution line"

$$y = \hat{\alpha} + \hat{\beta} x$$

This line provides for each observed independent value x_i the corresponding estimated value on the regression line $\hat{y}_i$ which is the reason why the is no residual term left, as all $\hat{y}_i$ are located on the regression line.

Example

To fit the weight of persons given their height the **regression model**

$$weight_i = \alpha + \beta_{height} \cdot height_i + \varepsilon_i$$

Once a regression model has been fitted with estimated model parameters $\hat{\alpha}$ = - 100 and $\hat{\beta} = 1.01$ the **model equation**

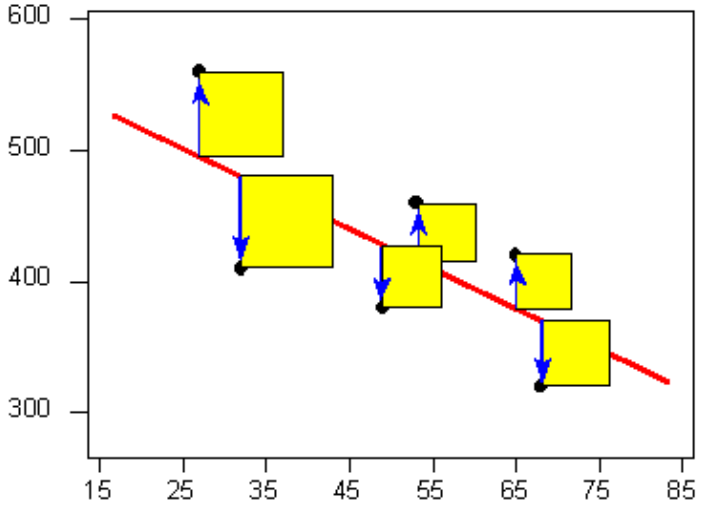
$$weight = -100 + 1.01 \cdot height$$

is determined.

For a person with height 178 cm the weight can be predicted as $-100 + 1.1 \cdot 178 = 79.78$ kg.

Least Squares Estimates for Regression visualised

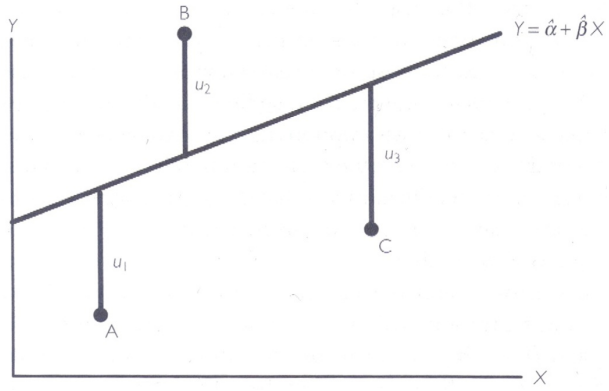
residual sum of squares = sum of areas of squares



Least Squares Estimates for Regression

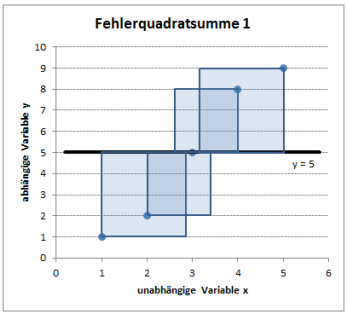
The optimal model is determined by minimising the sum of squared residuals $e_i = u_i^2$ defined by

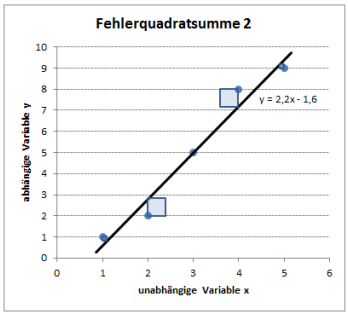
$$e_i = (y_i - \hat{\alpha} - \hat{\beta} y_i)^2, i = 1, 2, \ldots, N.$$



Least Squares Estimates for Regression visualised

residual sum of squares = sum of areas of squares





Linear Regression - OLS estimate

For the univariate one-way regression model

$$\begin{aligned} \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} \\ \hat{\beta} &= r_{xy}\frac{s_y}{s_x} = \frac{\frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum x_i^2 - \bar{x}^2} \end{aligned}$$

Thus, only in case of the one-way regression model $\hat{\beta}$ has the same sign as the correlation coefficient r_{xy}.

Correlation and the regression coefficient

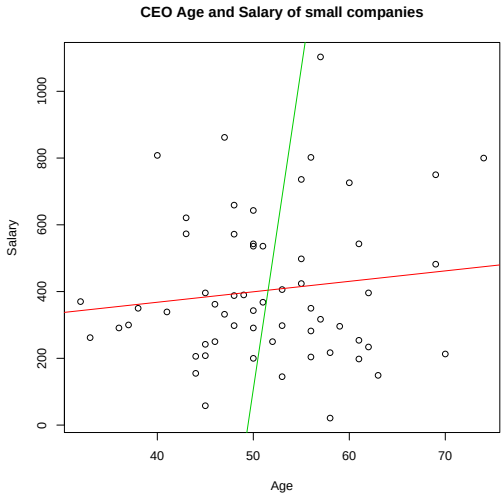
$$\begin{aligned} \hat{\beta}_{XY} &= \frac{\widehat{\mathrm{cov}}(X,Y)}{s_X^2}, \\ r_{XY} &= \frac{\widehat{\mathrm{cov}}(X,Y)}{s_X s_Y}. \end{aligned}$$

But

$$\hat{\beta}_{YX} = \frac{\widehat{\mathrm{cov}}(X,Y)}{s_Y^2},$$

which means that β is (unlike r) not symmetric. In other words: regression of Y onto X will generally not yield the same results as regression of X onto Y.

Which line is the "right" line?



Different Regression of Y onto X (red) and X onto Y (green).

Important properties of β

Correlation coefficient r and slope β are closely related:

- Positive values of β indicate a positive correlation between X and Y. Negative values indicate a negative correlation. $\beta \approx 0$ means that X and Y are (practically) uncorrelated.
- $\beta_{XX} = 1$
- $-\infty < \beta < \infty$
- Larger absolute values of β do not necessarily indicate stronger correlation.
- $\beta_{XY} \neq \beta_{YX}$ (in general)

Caveat!

- β is only a *linear* measure of dependence.
- $\beta \neq 0$ does not imply causality!

Model assumptions

Model assumptions We assume that the model can be written in the form

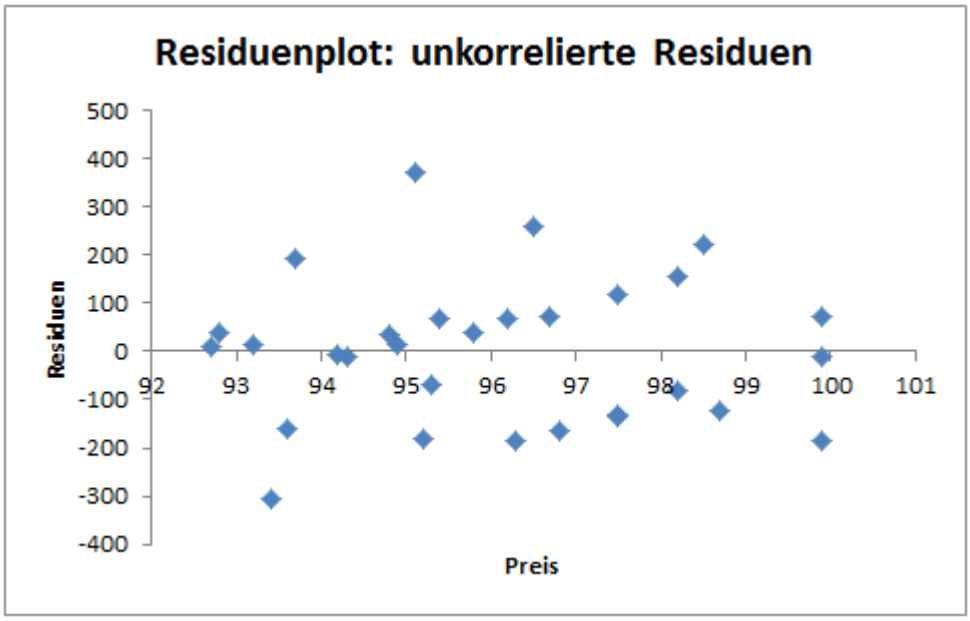
$$y_i = \alpha + \beta x_i + \varepsilon_i,$$

where the error terms are
* **unbiased**, i.e. their mean is equal to 0,
* **homoscedastic**, i.e. they have a constant variance,
* **uncorrelated**, i.e. they don't influence each other,
* **normally distributed**, i.e. they follow a Gaussian distribution.

This is only a necessary assumption for testing and estimating confidence bounds for parameters and the regression line itself!

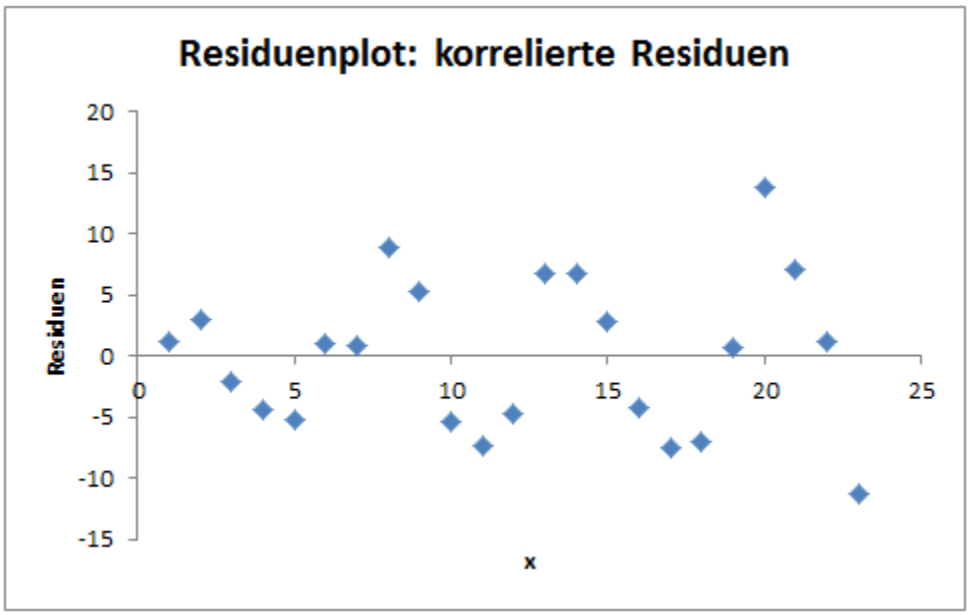
Visualisation of uncorrelated errors

This is what the residual vs. fitted plot should look like



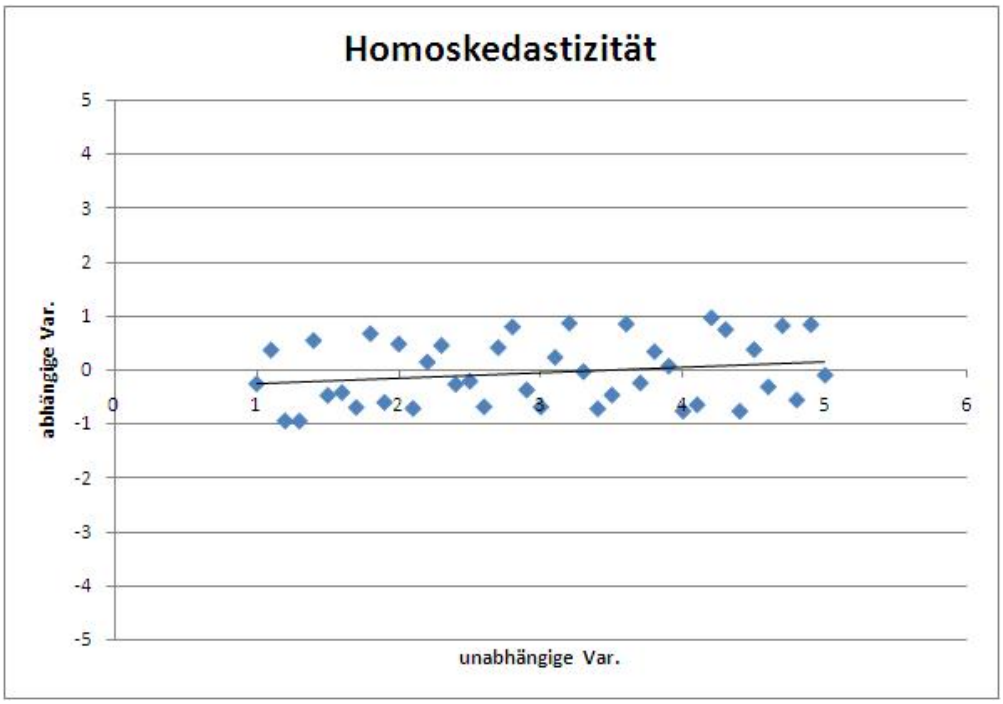
Visualisation of correlated errors

This is what the residual vs. fitted plot should NOT look like



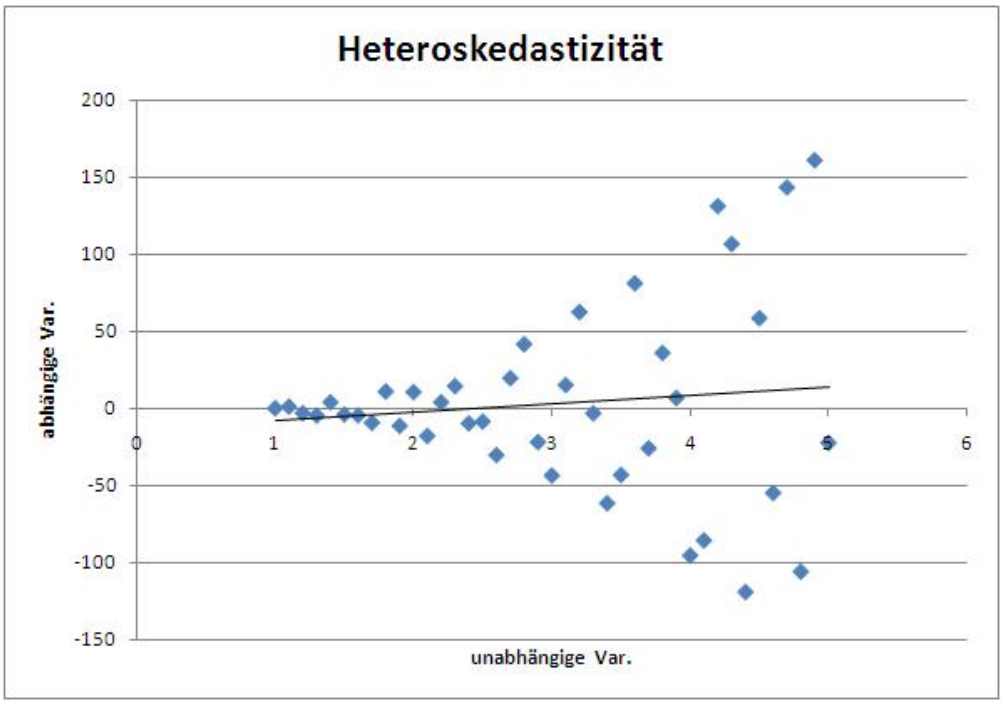
Visualisation of homoscedasticity (= equal variance)

This is what the residual vs. fitted plot should look like



Visualisation of heteroscedasticity (= unequal variance)

This is what the residual vs. fitted plot should NOT look like



About *Calculating* the precision of $\hat{\beta}$: s_b

We are looking for a (symmetrical) interval which covers β lies with a probability α (usually 0.95, 0.99 or even 0.999):

$$P(\hat{\beta} - q_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n-1}\hat{\sigma}_X} \leq \beta \leq \hat{\beta} + q_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n-1}\hat{\sigma}_X}) = 1 - \frac{\alpha}{2}$$

Under the assumption of *uncorrelated, homoscedastic, normally distributed errors* ε_i (see next slide), this interval can be calculated:

$$\hat{\beta} - t_b s_b \leq \beta \leq \hat{\beta} + t_b s_b,$$

where t_b denotes the proper quantile from the Student t distribution with $n-2$ degrees of freedom, and s_b denotes the standard deviation of $\hat{\beta}$ (often referred to as the *standard error*), given through

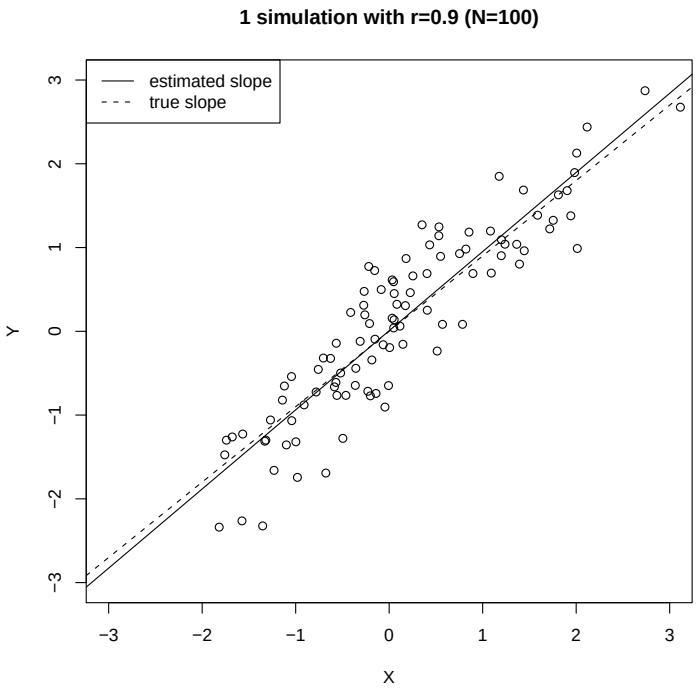
$$s_b = \sqrt{\frac{\text{SSR}}{(N-2)\sum(X_i - \bar{X})^2}}.$$

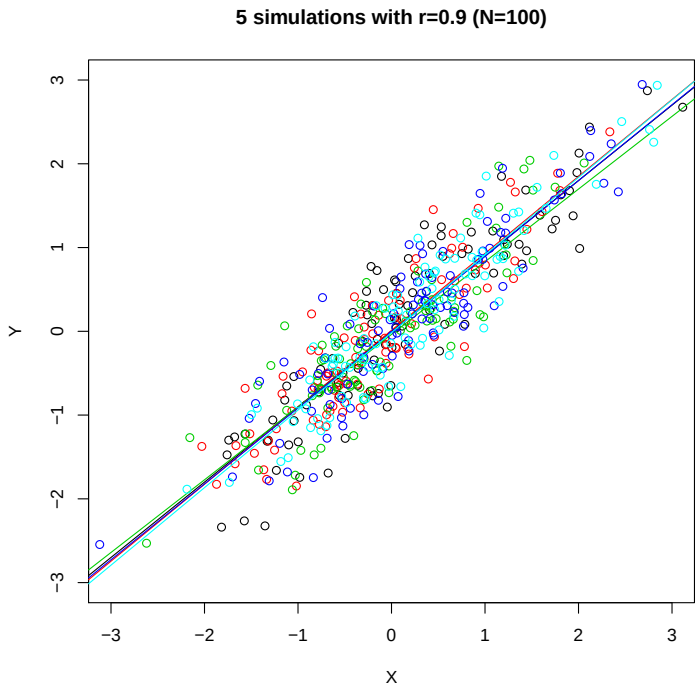
Standard errors

The slope standard deviation formula is consistent with the three factors that influence the precision of $\hat{\beta}$:

1. greater sample size reduces the standard deviation (resulting in a better correlation estimate)
2. greater σ^2 increases the standard deviation (resulting in smaller correlation)
3. greater X variability ($\hat{\sigma}_X$), i. e. a larger spread of X , reduces the standard deviation (resulting in larger correlation).

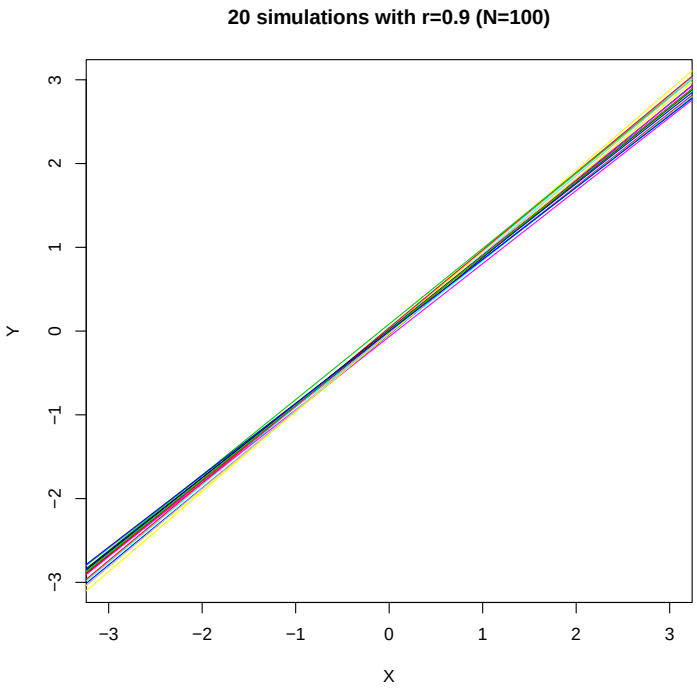
Let's talk about the precision of $\hat{\beta}$

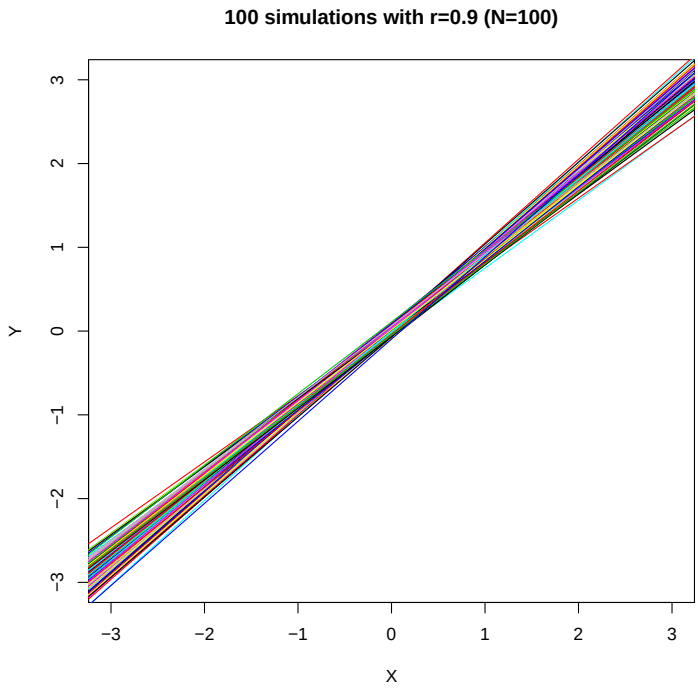




$0.9448 \leq \hat{\beta} \leq 0.9448$ $0.8677 \leq \hat{\beta} \leq 0.9448$

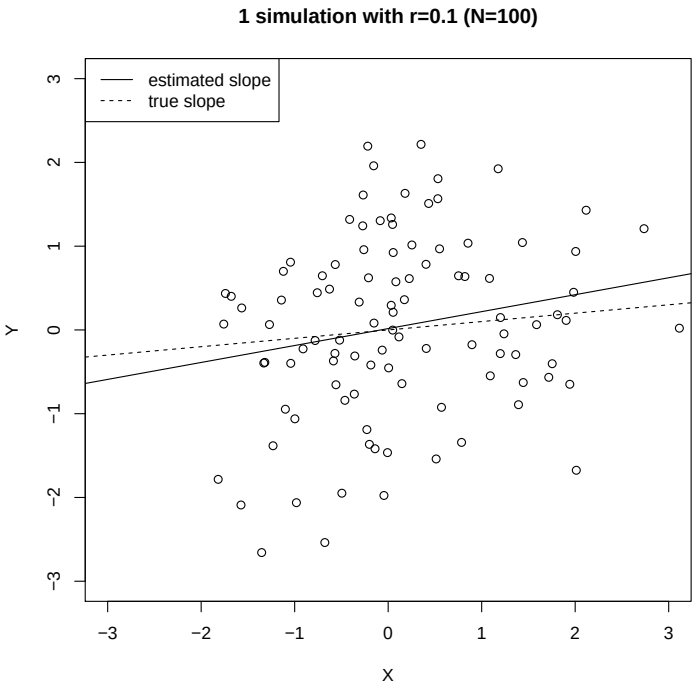
Let's talk about the precision of $\hat{\beta}$



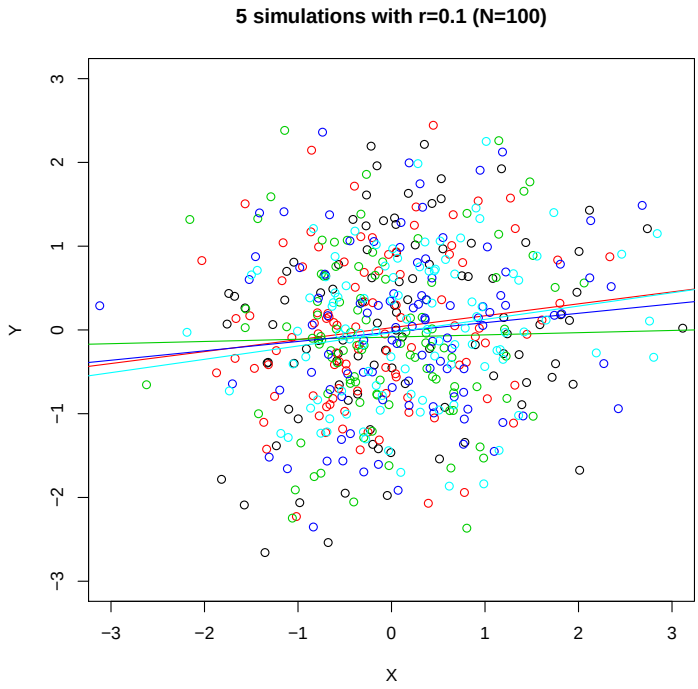


$0.8588 \leq \hat{\beta} \leq 0.9582$ $0.7874 \leq \hat{\beta} \leq 1.0089$

Let's talk about the precision of $\hat{\beta}$

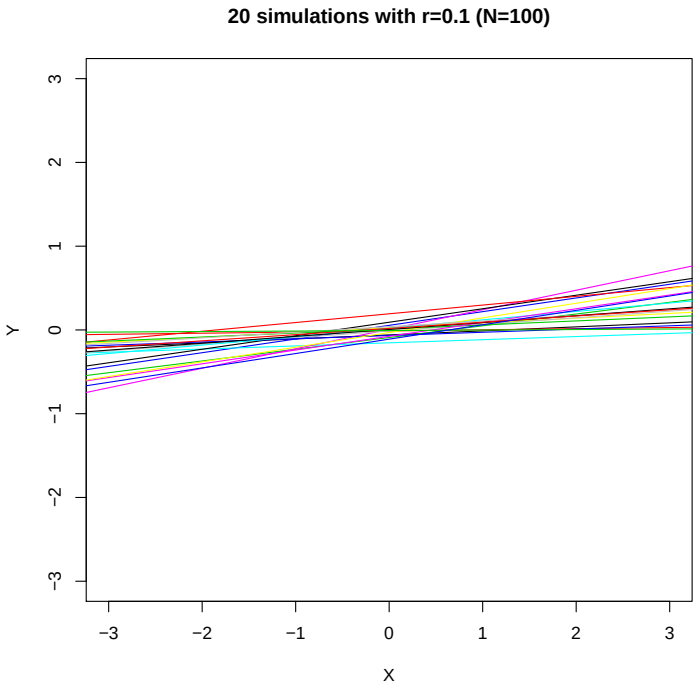


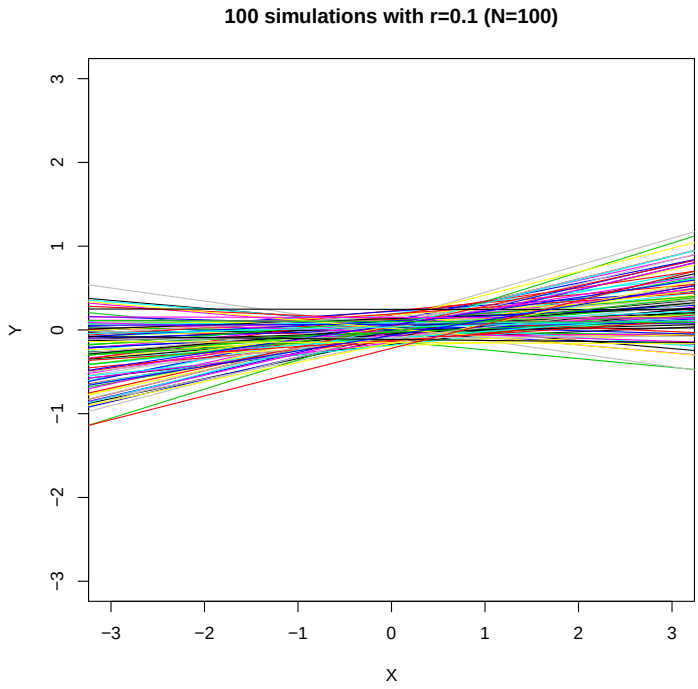
$0.2022 \leq \hat{\beta} \leq 0.2022$



$0.0263 \leq \hat{\beta} \leq 0.2022$

Let's talk about the precision of $\hat{\beta}$





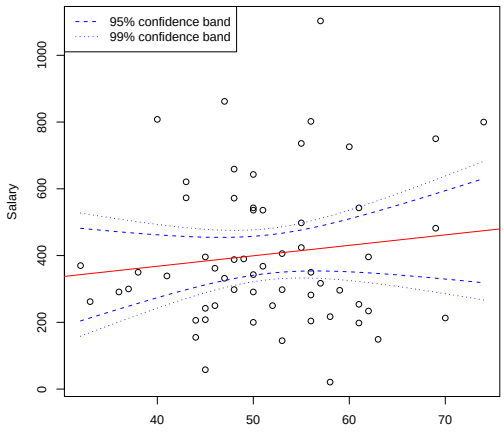
$0.0061 \leq \hat{\beta} \leq 0.2329$ $\quad -0.1569 \leq \hat{\beta} \leq 0.3486$

Confidence and Prediction Bands

Confidence and prediction bands for the regression line arise

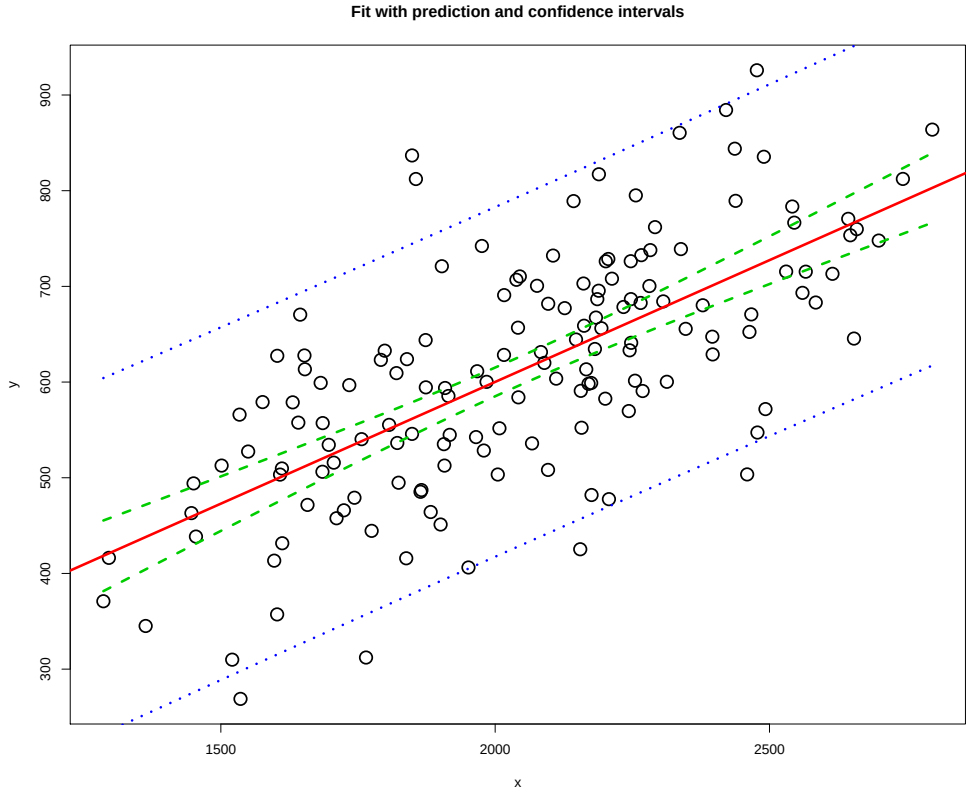
- **pointwise**, i. e. at each value of $\boldsymbol{x}_i$ a confidence interval for a fixed confidence level α is calculated
  Caution: although for each point the *confidence level* α is kept, the *coverage probability* α need not be kept for the regression line as a whole
- **simultaneous**, i. e. the **familiy wise error rate** is estimated such that the *coverage probability* α is kept for the whole regression line, but the *confidence level* α_i has to be changed at each value of $\boldsymbol{x}_i$
  methods for familywise error rate estimation: Bonferroni, Scheffè

CEO regression with confidence bands



$\hat{\alpha} = 242.702$ [168.760], $\hat{\beta} = 3.133$ [3.226].

Confidence and Prediction bands



Testing coefficients

Under the same assumptions as above (independently and identically distributed normal errors with mean 0 and homoscedastic variance σ^2), the confidence interval with specific variance calculation correspond to the

two-sided **t-test**.

$\Rightarrow$ simplest form of **model selection**:

testing whether each coefficient α, β is significantly different from 0, i.$_{\text{e.}}$influential.

$$\begin{aligned} H_0: &\quad \alpha \text{ or } \beta &&= 0 \\ H_A: &\quad \alpha \text{ or } \beta &&\neq 0 \end{aligned}$$

Testing coefficients

Given that H_0 is true (`under the Null`), the t-statistic

$$t = \frac{\hat{\beta}}{s_b},$$

follows a *Student's t distribution*. Thus, we expect values of t close to 0. Large values (in an absolute value sense) indicate that the assumption of H_0 might be wrong, and $\beta \neq 0$.

This test is similar to the t test for the mean of an approx. normally distributed variable. \[0.2cm]

R The summary() of a linear model object always includes t-test for each of the coefficients.

Coefficient of determination R^2

Coefficient of determination The **coefficient of determination** $R^2 = r^2_{\hat{y}y}$ is an indicator of how well data points fit the linear regression model.
As this increases automatically by increasing the number of independent variables, even if they have no explanatory effect, a corrected version of R^2 is included in regression outputs for measuring **goodness-of-fit**

Only in case of linear regression on a single independent variable

$$R^2 = r^2_{xy}.$$

Content of the R Summary

Given that the model assumptions are fulfilled, regression techniques provide the following information about β:

- $\hat{\beta}$, the OLS point estimate, or best guess, of what β is.
- A 95%/99%-confidence interval, where we are 95%/99% confident β will lie.
- The standard deviation (or standard error) of $\hat{\beta}$, s_b as a measure of how accurate $\hat{\beta}$ is. s_b is also a key component in the mathematical formula for the confidence interval and the test statistic for testing $\beta = 0$.
- The test statistic, t, for testing $\beta = 0$.
- The P-value for testing $\beta = 0$.

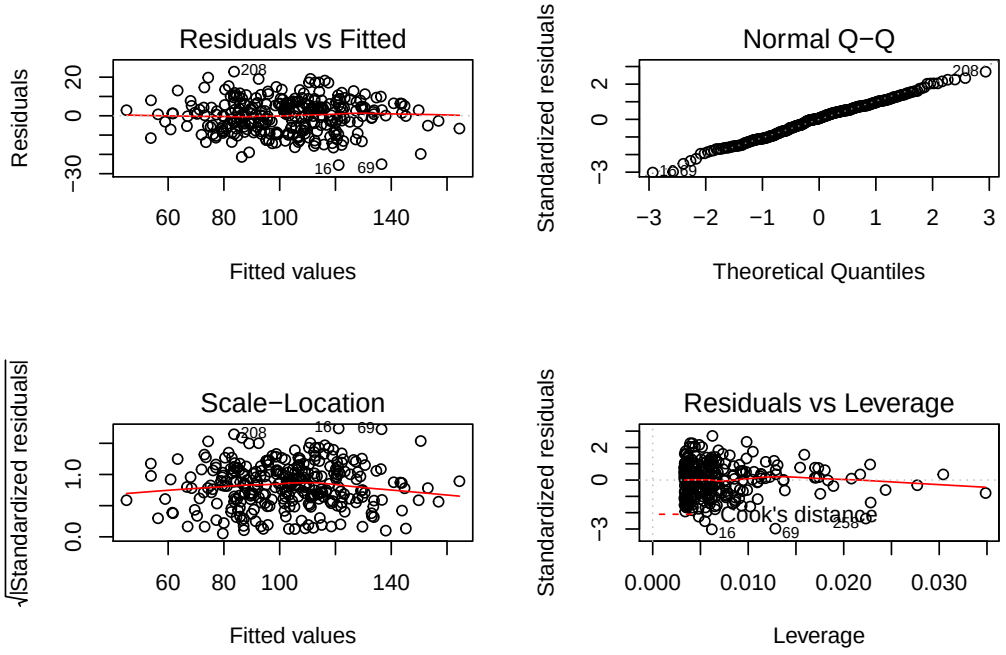
Summary of the linear regression model

```
summary(lm(y~x,data = df))
##
## Call:
## lm(formula = y ~ x, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.5353  -5.9668   0.5668   5.9204  22.8628
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -124.14513    5.52990  -22.45   <2e-16 ***
## x              1.29547    0.03148   41.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.479 on 298 degrees of freedom
## Multiple R-squared:  0.8503, Adjusted R-squared:  0.8498
## F-statistic:  1693 on 1 and 298 DF,  p-value: < 2.2e-16
```

Residual plots

Checking the assumptions for residuals: **residual plots**

```
par(mfrow=c(2,2))
plot(lm(y~x,data = df))
```



Residual plots

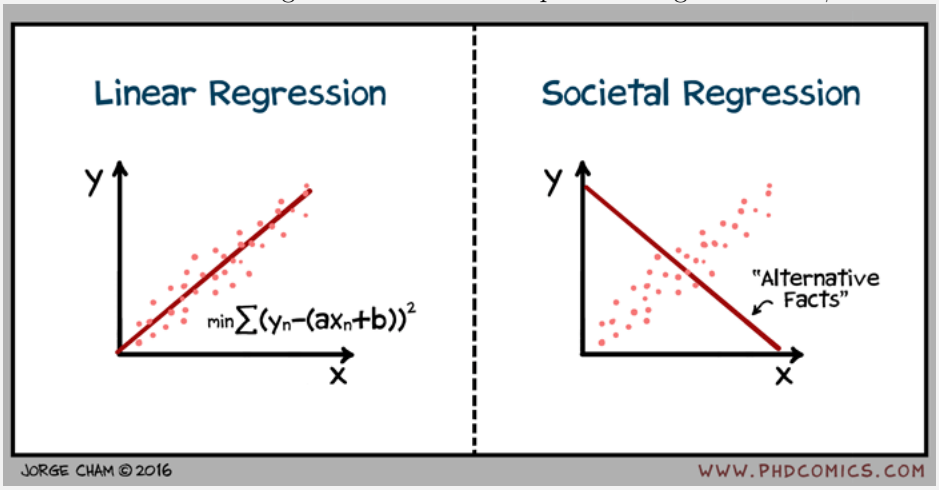
Cook's distance Cook's distance is a measure of influence of a single data point.

$$D([y_1, \boldsymbol{x}_i]) \quad = \quad \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j,-i}^{pred})^2}{p \cdot MSE}$$

where $\hat{y}_{j,-i}^{pred}$ is the estimate of y_j from a (refitted) regression model in which observation $[y_i, \boldsymbol{x}_i]$ was left out.

Leverage

Leverage
Leverage points are observations made at extreme or outlying values of the independent variables $\boldsymbol{X}$ which therefore have large influence on the slope of the regression line β.



Linear Regression - multiple model

Mathematically, the simple linear regression model is

$$y_i = \alpha + \beta_1 x_{1,i} + \ldots + \beta_k x_{k,i} + \varepsilon_i$$

in the notation of vectors and matrices this model corresponds to

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- μ and the α_i are unknown parameters of the population
- ϵ_{ij} are iid errors with mean 0 and a common unknown variance σ^2 (no heteroscedasticity).
- in case of multivariate X, the columns of $x_{k,\cdot}$ have to be stochastically independent

OLS estimates for multiple regression

The *ordinary least squares (OLS)* estimates:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

fulfills the Gauss-Markov theorem.

> Gauss-Markov Theorem For a linear regression model with errors having expectation zero and being uncorrelated and of equal variances (homoscedastic), the **ordinary least squares** (OLS) estimator of the coefficients $\hat{\boldsymbol{\beta}}$ is the **best linear unbiased estimator** (BLUE).

Linear Model with assumptions

`We assume that the model can be written in the form`

$$y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_p x_{p,i} + \varepsilon_i,$$

where the error terms are

- **centered around 0**, i.e. their mean is approximately 0,
- **homoscedastic**, i.e. they have a constant variance,
- **uncorrelated**, i.e. they don't influence each other, and are iid.
- **normally distributed**, i.e. they follow a Gaussian distribution.

  necessary assumption for testing and estimating confidence bounds for parameters and the regression line itself!
- **explanatory variables are uncorrelated**, i.e. x_i and x_j do not have linear dependences

Interpretation of coefficients

Interpretation of coefficients:

- The intercept β_0 is the average value of Y, when all X_i are equal to 0.
- β_i: The expected value of Y changes by β_i, if X_i is increased by 1 unit while all other X_j, $j \neq i$, are kept at the same value. "**marginal change**"

Model Selection

Several approaches towards model selection exist:

- **t-test** for each coefficient β_i
  simplest way of model selection comparing the model including the regressor against the simpler model excluding the regressor
- **ANOVA** comparison of nested models
  ANOVA can compare the residual sums of squares of any two nested models
- general model selection based on **"goodness-of-fit"** measures
  stepwise model selection based on AIC, BIC

Information Criteria

AIC **Akaike Information Criterion**

$$AIC \quad = \quad -2ln(L) + 2k$$

where k is the number of regressors, i. e. model parameters, and L is the value of the model likelihood function at its maximum.

BIC **Bayesian Information Criterion**

$$BIC \quad = \quad -2ln(L) + kln(n)$$

where n is the number of observations.

R AIC(x), BIC(x)
step(x) performs stepwise model selection based on information criteria

Regression on categorical variables

Categorical variables split the regression space up, such that a **separate regression** for each fixed category is fitted on the other regressors.

In the special case where all regressors are categorical variables, only a separate intercept is fitted for each combination of categories, which is exactly what **Analysis of Variance (ANOVA)** estimates.

Polynomial regression

A polynomial of degree n is a function formed by linear combinations of the powers of its argument up to n:

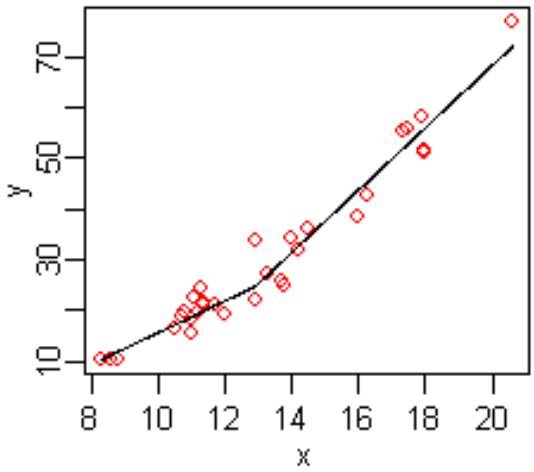
$$y = a_0 + a_1x + a_2x^2 + a_3x^3 + \ldots + a_nx^n$$

Specific Polynomials:

1. Linear $y = a_0 + a_1x + a_2x^2$
2. Quadratic $y = a_0 + a_1x + a_2x^2$
3. Cubic $y = a_0 + a_1x + a_2x^2 + a_3x^3$

Splines in regression

extension of linear models that automatically models nonlinearities and interactions between variables

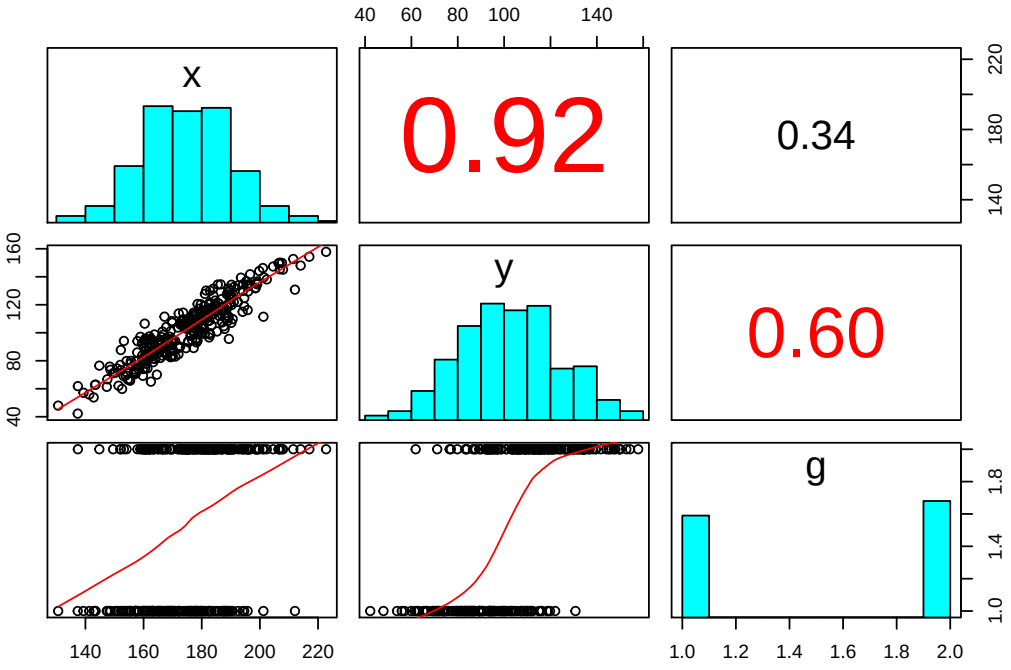


- Splines models partially linear (or quadratic, cubic etc.) models
- non-parametric regression technique
- moves "to the right" and approaches a normal distribution as $df \to \infty$

Linear Regression Models in R

First, we start with taking a look at the pairwise scatter plot matrix

```
pairs(df, lower.panel = panel.smooth,
      diag.panel = panel.hist, upper.panel = panel.cor)
```



ANOVA fit in R for additive model

linear models (lm) fit a dependent numerical variable y depending on independent variables x, g; at least one independent variables has to be numeric

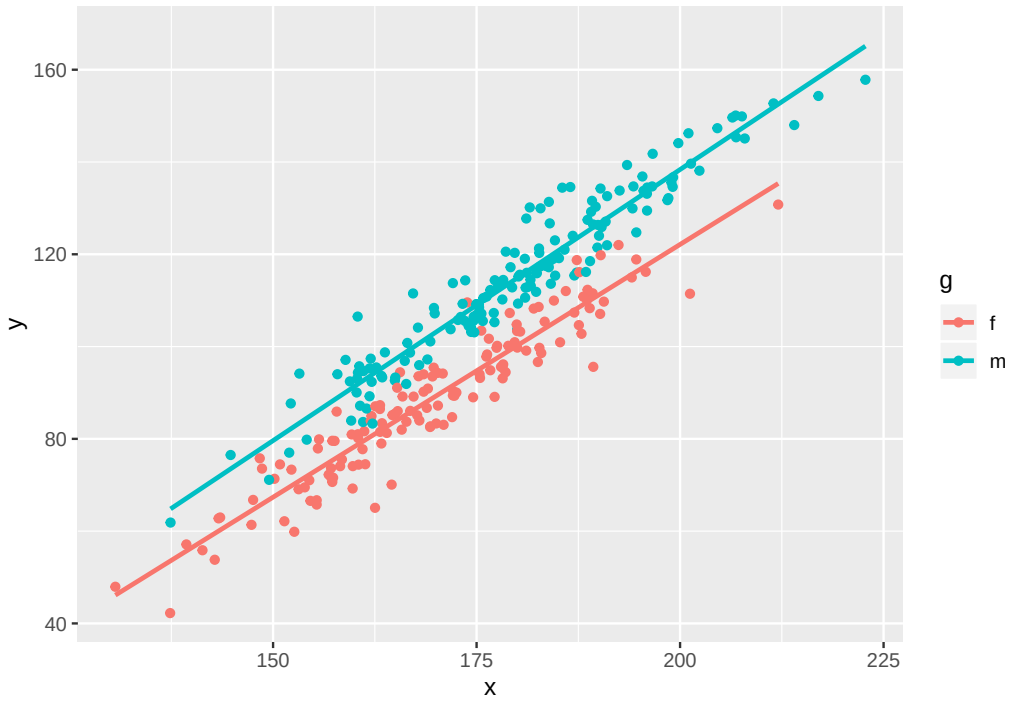
```r
# fitting additive regression model
linear_regression_model <- lm(y~x+g,data = df)
linear_regression_model
##
## Call:
## lm(formula = y ~ x + g, data = df)
##
## Coefficients:
## (Intercept)            x           gm
##     -104.45         1.14        14.17
```

Summary of the linear regression model

```r
summary(linear_regression_model)
##
## Call:
## lm(formula = y ~ x + g, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7383  -3.3550  -0.5116   3.5357  15.8740
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -104.44516    3.54351  -29.48   <2e-16 ***
## x              1.13968    0.02077   54.87   <2e-16 ***
## gm            14.16625    0.64739   21.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.255 on 297 degrees of freedom
## Multiple R-squared:  0.9427, Adjusted R-squared:  0.9423
## F-statistic:  2443 on 2 and 297 DF,  p-value: < 2.2e-16
```

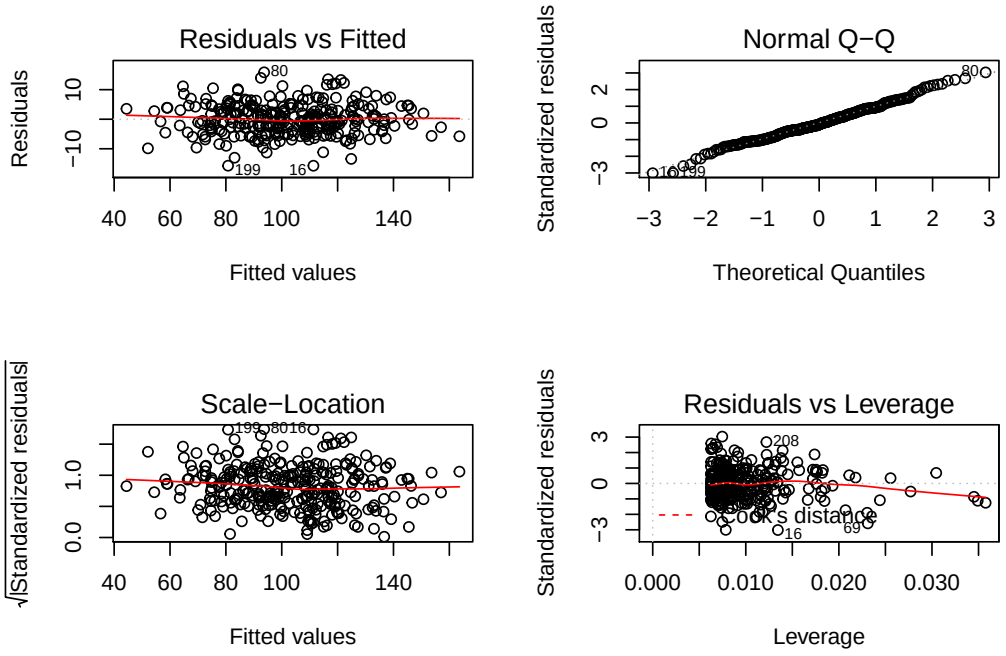
Visualisation of the linear regression model

```r
ggplot(df, aes(x = x, y = y, color = g)) +
    geom_point() + geom_smooth(method = "lm",
    fill = NA)
```



Model quality plot

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



Bayesian Regression model

Linear model

$$\mathbf{E}[y|X] = f(X'\beta)$$

$$y|X \sim N(X'\beta, \sigma^2)$$

with model matrix X.

Model matrix The model matrix contains the numerical values of x_i in column i for numeric variables and dummy coded columns for categorical variables.
model.matrix(formula,data)

Bayesian Regression model

Bayesian Linear model

$$y|X \sim N(X'\beta, \sigma^2)$$

with prior distributions

$$\beta \sim N(0, s_b^2)$$

$$\sigma^2 \sim IGamma(a, b)$$

Bayesian Regression model

The Bayesian linear model estimates the posterior distributions of the coefficients β_i of the explanatory variables x_i.

Similar to the classical linear regression model, these coefficients can be tested to be 'different from 0' with certain odds as opposed to being 0

this means that the variable influences y with with certain odds as opposed to having no influence.

This leads to a Bayesian hypothesis test.

Bayesian Hypothesis tests

Bayesian posterior distribution provides the probability of the parameter given the observed data

Bayes factors The Bayes Factor is the ratio of posterior probabilites of parameters under 2 hypotheses.

$$BF = \frac{\mathbb{P}[\theta \in \Theta_0|x]}{\mathbb{P}[\theta \in \Theta_1|x]}$$

This is symmetric w. r. t. the two hypotheses and on ratio scale, thus $BF = 2$ can be interpreted as H_0 is twice as likely as H_1. If you want asymmetry you can add weights (this means "loss" when making the wrong decision). The decision is then based on $\frac{k_0}{k_1} \cdot BF$

What are Generalised linear models ?

Question: How do we use the regression method not for (approximately) normal data (errors are approximately normal!), but for binary data instead?

Answer: Use the binomial distribution in the likelihood function.

$$f(y \mid p) \quad = \quad p^y \cdot (1-p)^{1-y}$$

with probability of success p and $y = 0$ in case of a failure and $y = 1$ in case of a success.

The expectation for this is $E(y) = p$.

Generalised linear models and a different problem formulation

We have to reformulate the model with expectation $E(y_i) = \mu_i = p_i$ in such a way that we can use the *linear predictor* of the original linear model $x_i^\top \beta$ again.

Problem: $x_i^\top \beta$ can result in any real number, but the probability of success $\mu_i = p_i$ has to lie within the interval $[0, 1]$.

Generalised linear models and the link function

Solution: Use a *Link-Funktion* g, which maps the interval $[0, 1]$ on the real numbers in such a way that

$$g(\mu_i) \quad = \quad x_i^\top \beta$$

Question: What is a good link-function? }

Revision of Contingency tables

When comparing 2 categorical variables we construct contigency tables. The visualisation of a contingency table is the mosaic plot.

Example: `screening` and `gender` in using cancer prescreening

gender vs. screening	no	yes
women	273	183
men	627	217

Contingency tables and relative frequencies

To compare the probability of men and women to use prescreening, we obtain row-wise relative frequencies.

gender vs. screening	no	yes	margin
women	0.599	0.401	1
men	0.743	0.257	1

Odds and Odds ratios

These relative frequencies can be tranformed into odds which are well-known in betting and gambling.

$$\text{Odds(screening)} \quad = \quad \frac{P(\text{screening})}{1 - P(\text{screening})}$$

To compare two odds, we calculate the *Odds Ratio*, which is

$$OR = \frac{\text{Odds(screening} \cup female)}{\text{Odds(screening} \cup male)} \quad = \quad \frac{\frac{P(\text{screening} \cup female)}{1-P(\text{screening} \cup female)}}{} \frac{P(\text{screening} \cup male)}{1 - P(\text{screening} \cup male)}$$

Interpreting Odds

For women the odds of using pre-screening are $0.401/0.599 = 0.67$, approximately 2:3.

For men the odds of using pre-screening are $0.257/0.743 = 0.346$, approximately 1:3.

To relate both odds, we caluclate the *Odds Ratio*. The odds ratio $0.346/0.67 = 0.516$ tells us that the odds of men to use pre-screening are half of women's odds for using prescreening.

Odds and Odds ratios

Based on the original table, we calculate odds ratios

gender vs. screening	no	yes
women	273	183
men	627	217

$$\text{OddsRatio} \quad = \quad \frac{273 \cdot 217}{627 \cdot 183} \; = \; 0.516$$

Odds ratio and relative risk

Relative risk is frequently used in medical terminology. It described the ratio of probabilities

$$RR = \frac{\pi_{women}}{\pi_{men}} = \frac{0.401}{0.257} = 1.560311$$

Women are therefore 1.56 times more likely to use pre-screening than men.

Exercise Odds, relative risk

A research study estimated that under a certain condition, the probability that a subject would be referred for heart catheterization was 0.906 for whites and 0.847 for blacks.

- Press release about the study stated that the odds of referral for cardiac catheterization for blacks are 60% of the odds for whites.

  Explain how they obtained 60% more accurately, 57% .

- An Associated Press story later described the study and said 'Doctors were only 60% as likely to order cardiac catheterization for blacks as for whites.'

  Explain what is wrong with this interpretation.

  Give the correct percentage for this interpretation.

  In stating results to the general public, it is better to use the relative risk than the odds ratio. It is simpler to understand and less likely to be misinterpreted.

  For details, see New Engl.J.Med.341:279-283,1999

Logistic regression

The Link-function for logistic regression is the *Logit-function*

$$g(\mu) \quad = \quad \log\left(\frac{\mu}{1-\mu}\right)$$

This function calculate the logarithms of odds (log-odds).

This leads us to the first generalised linear model:

$$\begin{aligned} E(y_i) &= \mu_i \\ g(\mu_i) &= x_i^\top \beta \quad g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) \end{aligned}$$

Interpretation of log-odds in Logistic regression

Interpretation: The logistic regression describes the log-odds

$$\log\left(\frac{P(y_i = 1 \,|\, x_i)}{P(y_i = 0 \,|\, x_i)}\right) \quad = \quad x_i^\top \beta$$

Example: Descrbing the relation between `screening` and `gender`.

The odds for men are

$$\frac{P(y=1\,|\,x=1)}{P(y=0\,|\,x=1)} \quad = \quad \exp(\beta_1 + \beta_2 \cdot 1)$$

Logistic regression and the odds ratio

and the odds for women:

$$\frac{P(y=1\,|\,x=0)}{P(y=0\,|\,x=0)} \quad = \quad \exp(\beta_1 + \beta_2 \cdot 0)$$

resulting in an odds ratio

$$\begin{aligned} \frac{\exp(\beta_1 + \beta_2 \cdot 1)}{\exp(\beta_1 + \beta_2 \cdot 0)} &= \frac{\exp(\beta_1 + \beta_2)}{\exp(\beta_1)} \\ &= \exp(\beta_1 + \beta_2 - \beta_1) \\ &= \exp(\beta_2) \end{aligned}$$

Interpreting Logistic regression coefficients

The relation between the coefficients of the logistic regression and the odds ratios is: The coefficient β_i describes the change of the log-odds, if x_i increases by 1.

Therefore, e^{β_i} describes the odds ratio, if x_i increases by 1.

In our pre-screening example the coefficents are

$$\beta \quad = \quad (-0.400, -0.661)^\top$$

The odds for pre-screening therefore are 48.4% less for men, as $\exp(-0.661) = 0.516$.

Exercise - Pima Indian Data

Use the Pima Indian data from package MASS.

- Visualise and explore the data appropriately.
- Fit a logistic regression model.
- Is there a significant dependence between bmi and glu?
- Explain the odds ratio interpretation and 'percentual' increase of the chance for diabetes for the significant explanatory variables.

Generalised linear Model

The generalised linear model (GLM) is described by

$$\begin{aligned} E(y_i) &= \mu_i \\ g(\mu_i) &= x_i^\top \beta \end{aligned}$$

- y_i — random variable of the exponential family (normal, binomial, negat. binomial, Poisson, Gamma, beta and some other distributions) with expectation μ_i.
- g — Link-function which maps from the scale of the expectation μ_i into the scale of the linear predictor (real numbers).
- x_i — vector of explanatory variables.
- β — vector of regression coefficients.

Exponential family

A variable y has an exponential family distribution, if its density function is a special cas eof the following function

$$f(y \,|\, \theta, \phi) \quad = \quad \exp\left(\frac{y \cdot \theta - a(\theta)}{\phi} \, + \, b(y, \phi) \right)$$

with known functions $a(\cdot)$ and $b(\cdot)$, scale parameter ϕ and shape parameter θ which regulate the shape and scale of the distribution.

The exponetial family includes:

- normal distribution
- binomial distribution, negative binomial
- Poisson distribution
- beta distribution
- gamma distribution, exponential distribution, χ^2 distribution

Iteratively Weighted Least Squares

When fixing the distribution of y_i, the likelihood function, the marginal probability of the data and the parameter estimates via maximum likelihood (ML) approach can be determined.

With the exception of few special cases, the maximum likelihood estimator cannot be calculated in closed from. Thus, it is determined by an iterative procedure: IWLS (iteratively weighted least squares).

Model selection

As the fitted GLMs are estimated with maximum likelihood method based on exponential family distributions, we can again use

- t- and F-tests or

- AIC and BIC

for model selection, all based on the log-likelihood.

Residuals

Residuals in the linear model:

$$y_i - \hat{\mu}_i$$

We had assumed that the variance of the residuals is the same independent of x (i.e. homoscedasticity).

In GLMs this is generally not the case!

Therefore, residuals are 'reweighted with their respective variances determined by the variance function $V(\mu)$ which depends on the distribution function.

Also other reasonable definitions of residuals exist, such as *Deviance-Residuals* which are R's default.

Pearson-Residuals

the residuals

$$\frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

are called *Pearson-Residuals*.

The variance function $V(\mu)$ depends on the distribution function:

- normal: $V(\mu) = 1$,
- binomial: $V(\mu) = \mu \cdot (1 - \mu)$,
- Poisson: $V(\mu) = \mu$.

Linear Regression as a special case

Linear regression is a special case of generalized linear regression with iid normally distributed data y_i with identical variance (homoscedasticity).

The Link-Function is the identity function $g(\mu) = \mu$.

This scenario is equivalent to the 5 assumption of the linear regression model with numerical y.

Binomial regression

For binary independent variable: binomial distribution model. Two general variants exist:

1. Logit Link-function models the log-odds

$$g(\mu) \quad = \quad \log\left(\frac{\mu}{1-\mu}\right)$$

2. Probit Link-function models based on the cumulative distribution function of the normal distribution $\Phi(\cdot)$

$$g(\mu) \quad = \quad \Phi^{-1}(\mu)$$

Probit model - Motivation

Motivation: We assume, a biological or biochemical response only applies, if a certain limit L for the stimulus is reached. If this is the case, a given gene is expressed or hormon is produced.

However, the limit L will not be identical for all tested individuals, but we expect this limit to be distributed according to a normal distribution with mean m and variance s^2

Probit model - Normal distribution link

Then, the probability p for expressing the gene or producing the hormon is

$$p \quad = \quad P(R \leq \text{stimulus}) \quad = \quad \Phi\left(\frac{\text{stimulus} - m}{s}\right)$$

resulting in

$$\Phi^{-1}(p) \quad = \quad -\frac{m}{s} + \frac{1}{s} \cdot \text{stimulus} \quad = \quad \beta_1 + \beta_2 \cdot \text{stimulus}$$

This is the Probit-link generalized linear model.

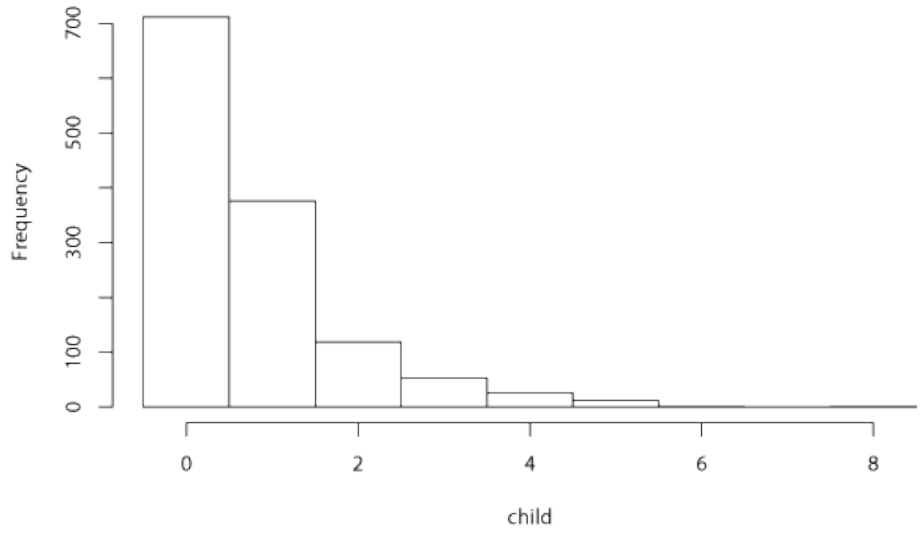
Poisson Regression

To model count data the Poisson-distribution is the most frequently used distribution.

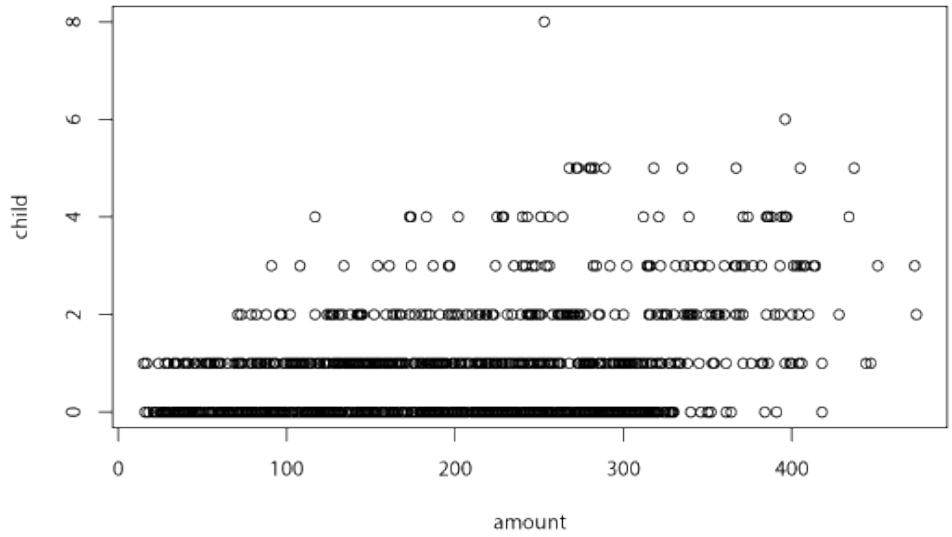
The link-function for Poisson-regression generally is the logarithm :

$g(\mu) = \log(\mu)$

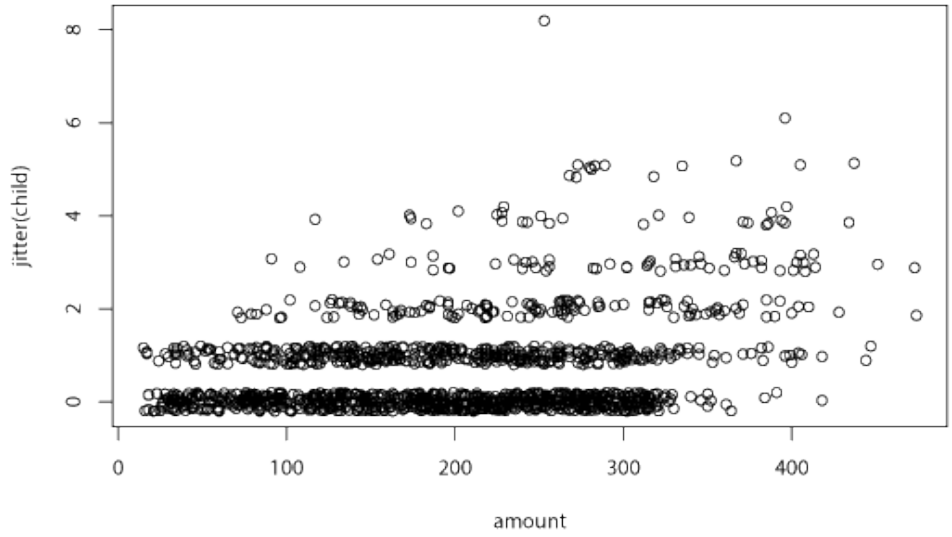
Poisson Regression



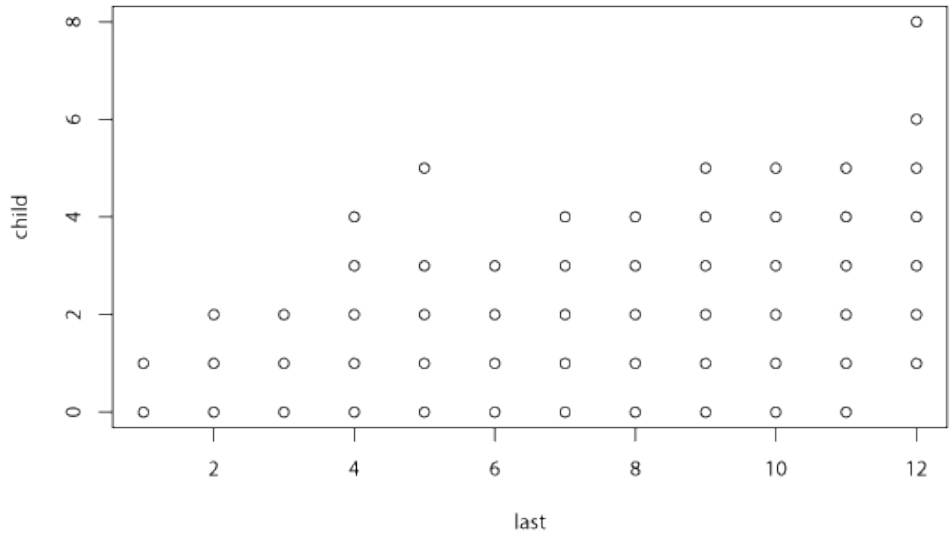
Poisson Regression



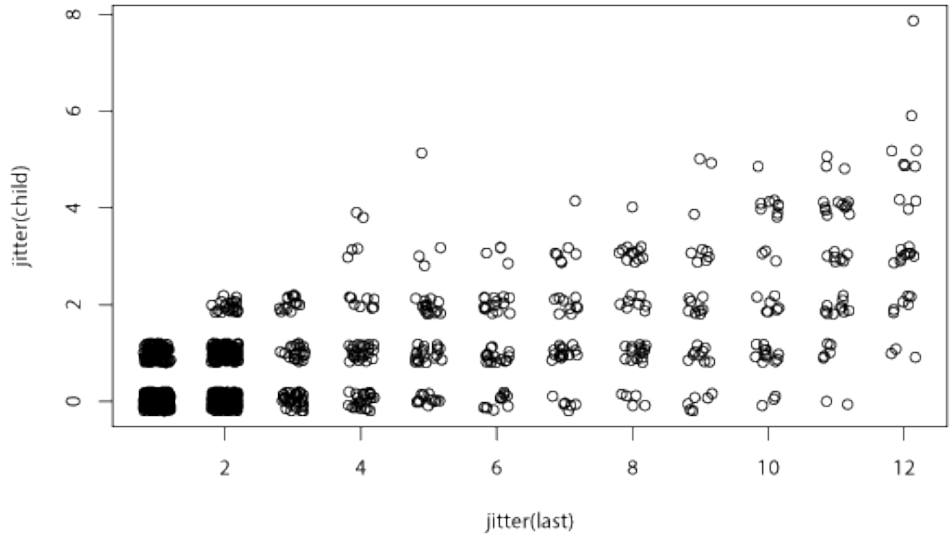
Poisson Regression



Poisson Regression



Poisson Regression



Choice Models

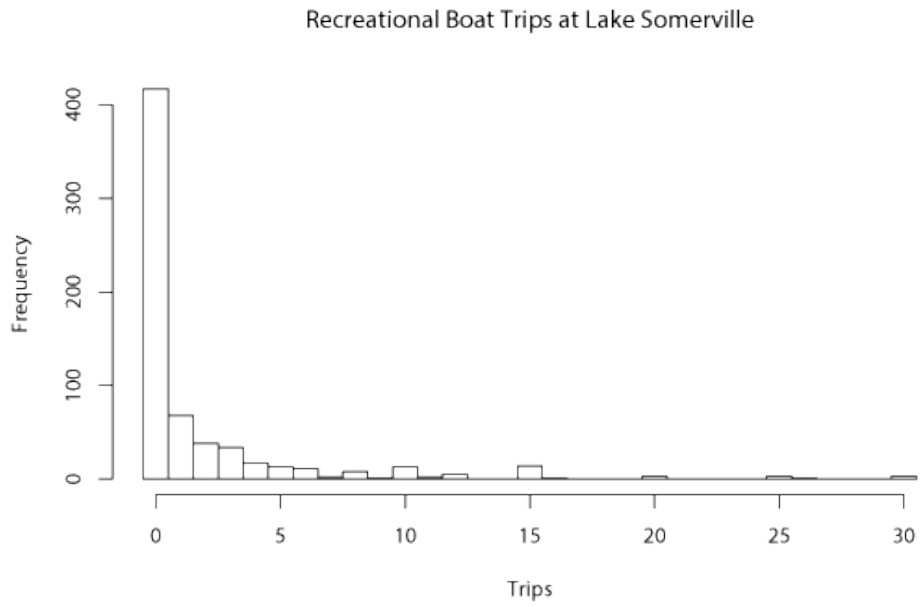
- **Multinomial model**: similar to the binomial model, but with $c > 2$ possible categories.
- **Ordinal logistic Regression = proportional odds model**: The dependent variable contains $c > 2$ ordered categories (e.g. Ratings, age categories, etc.).
- **Conjoint Analysis**: similar to the multinomial model, but not every entity can choose between all c possible categories.

Count data

For many economical and biological data sets count data contain more '0' observations than expected by a given probability distribution ("zero-inflation").

- Zero-inflated Poisson (ZIP) Model
- Zero-inflated Multinomial Model

Zero-inflated Poisson (ZIP) Model



Machine Learning

If the probabilistic model is irrelevant, yet an algorithmic `classification'' or`prognosis" is to be estimated, algorithms and methods from *machine learning* or *statistical learning* can be applied:

- Neural Networks,
- Support Vector Machines,
- tree-based methods: classifikation trees, Bagging, Random Forest,
- Boosting

Tree based method