Introduction to Statistics

Linear Regression

Alexandra Posekany
alexandra.posekany@wu.ac.at

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)

Testing for Correlation

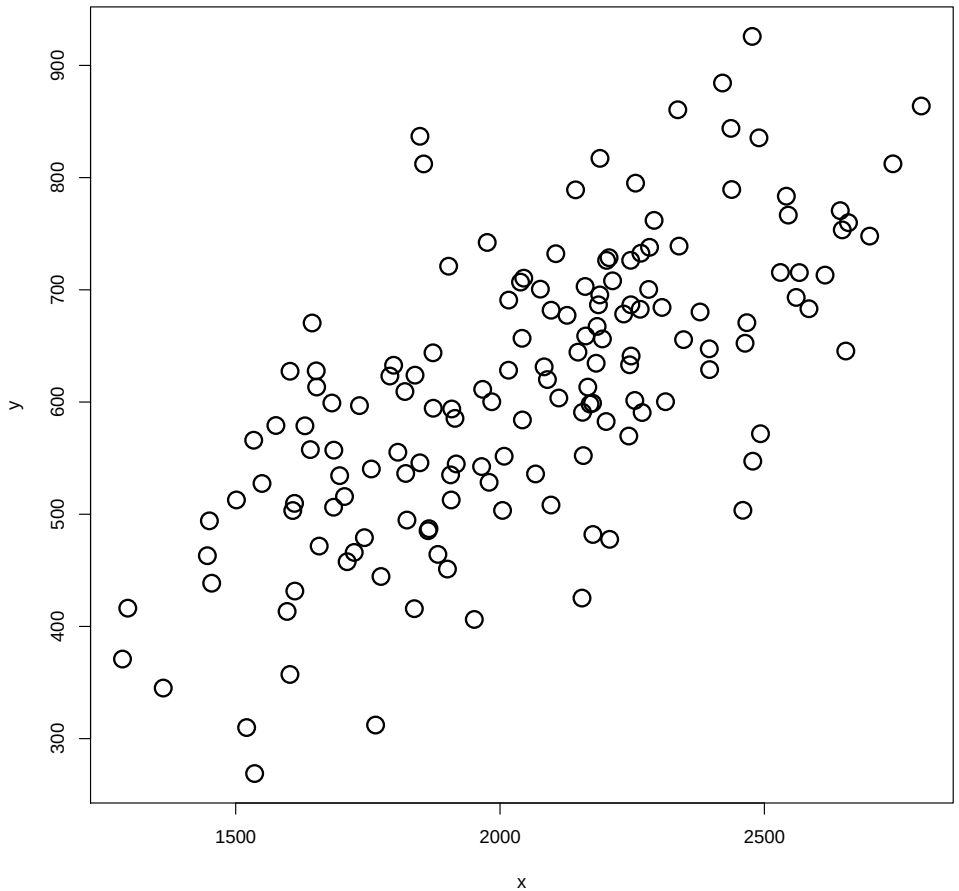
If a justification for 'non-zero' correlation is required, one can test for correlation with a t-test for the two-sided hypothesis

$$H_0: \quad r_{xy} = 0$$
$$H_A: \quad r_{xy} \neq 0$$

R

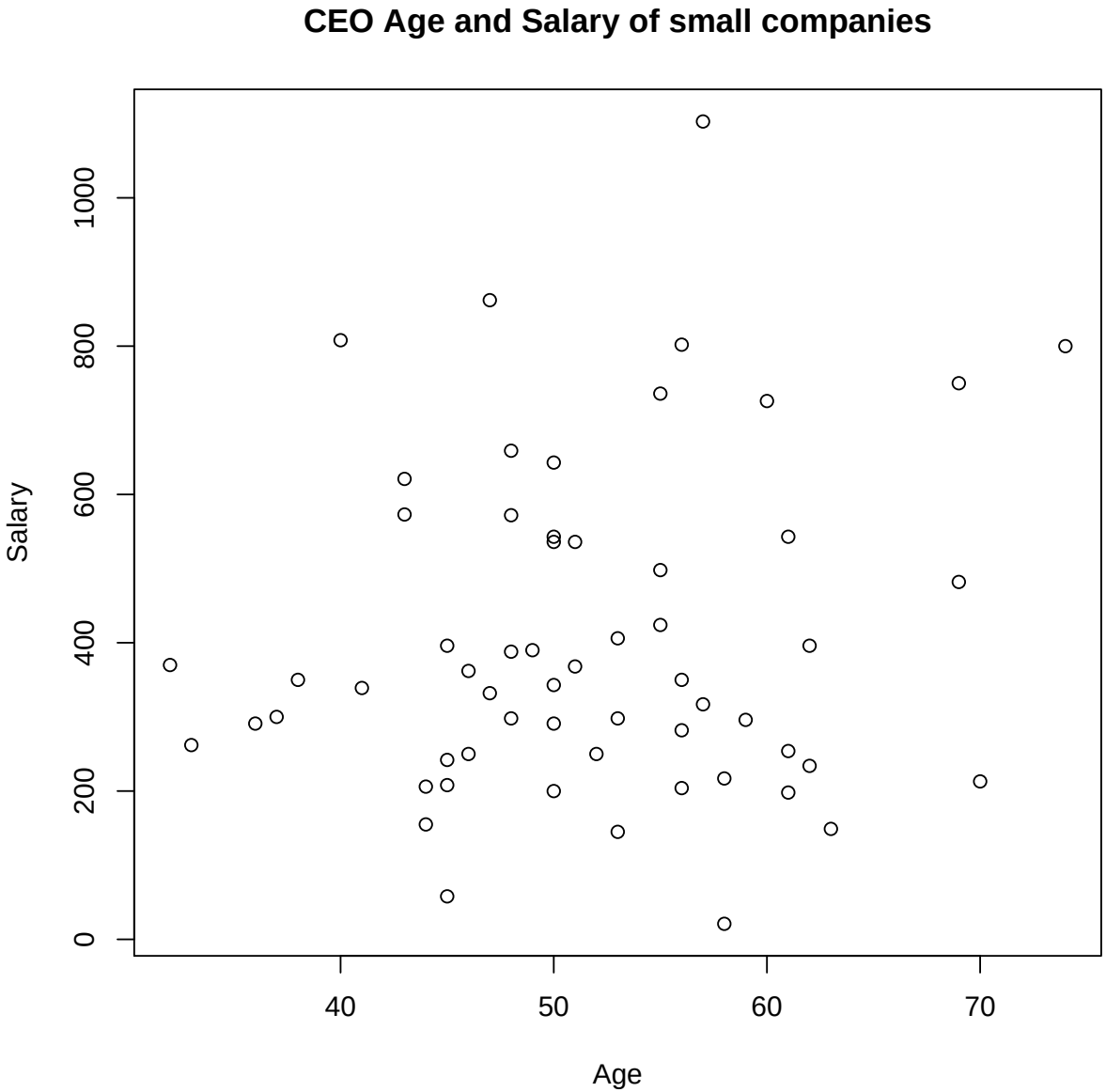
```
cor.test(x, y,
         alternative = c(two.sided, less, greater),
         method = c(pearson, kendall, spearman),
         conf.level = 0.95,, ...)
```

Regression-Motivation



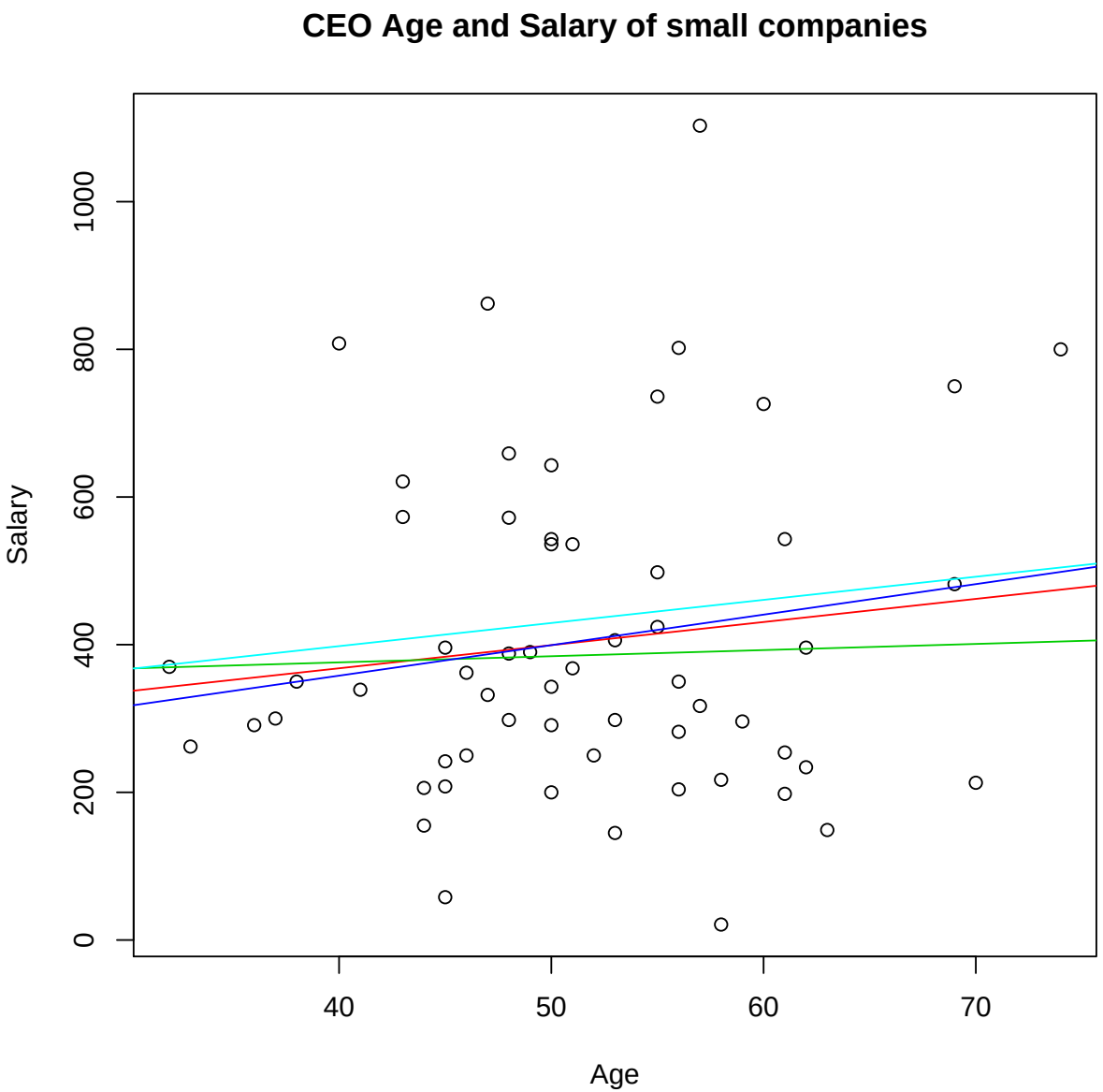
Correlation
$r = 0.68$

Regression-Motivation



Correlation
$r = 0.13$

Regression- Which line is the “right” line?



Regression model vs. model equation

The regression model

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

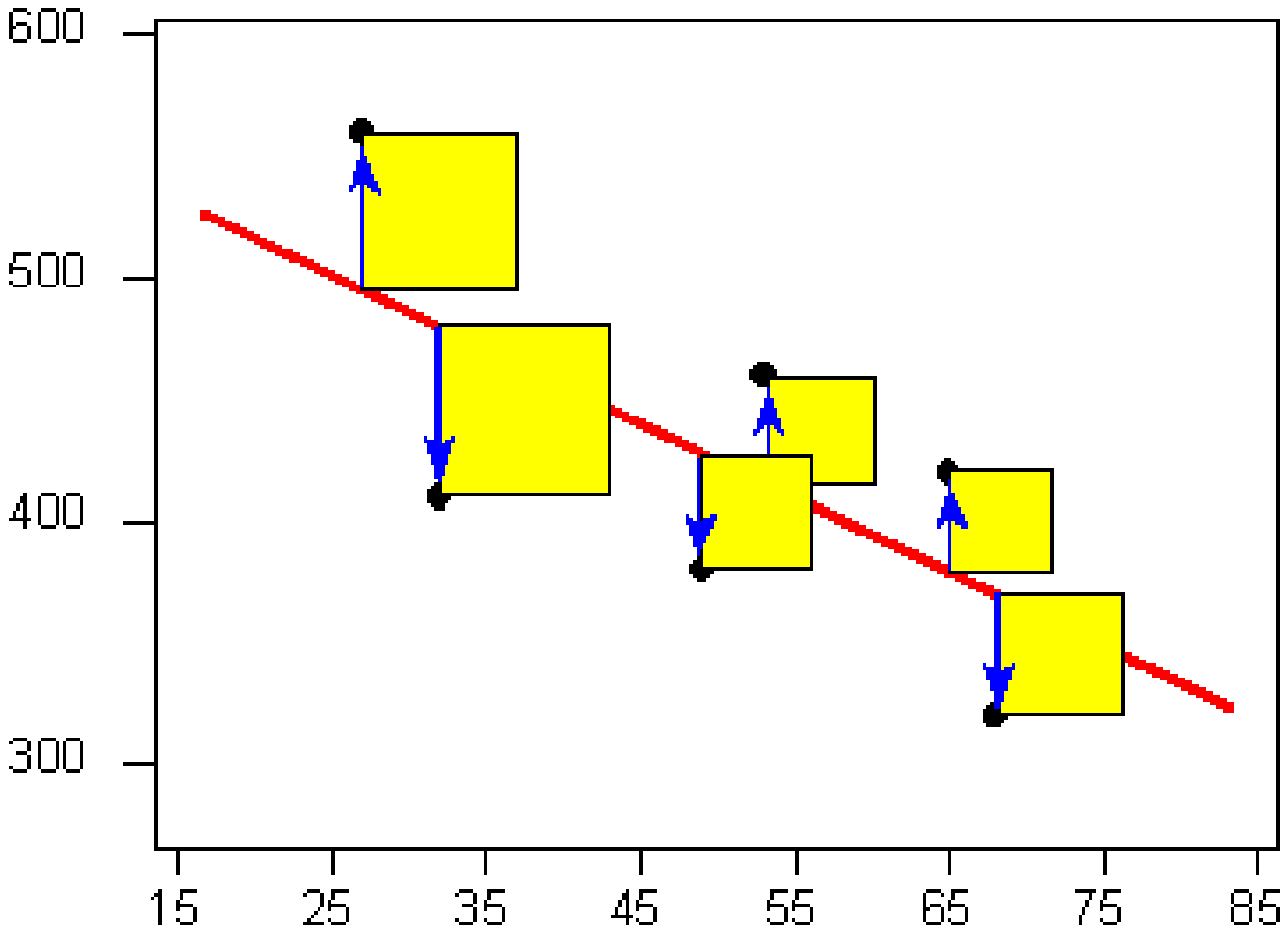
explains the observed values y_i.
Once a specific line is selected, we obtain an actual model equation of the "solution line"

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

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.

Least Squares Estimates for Regression

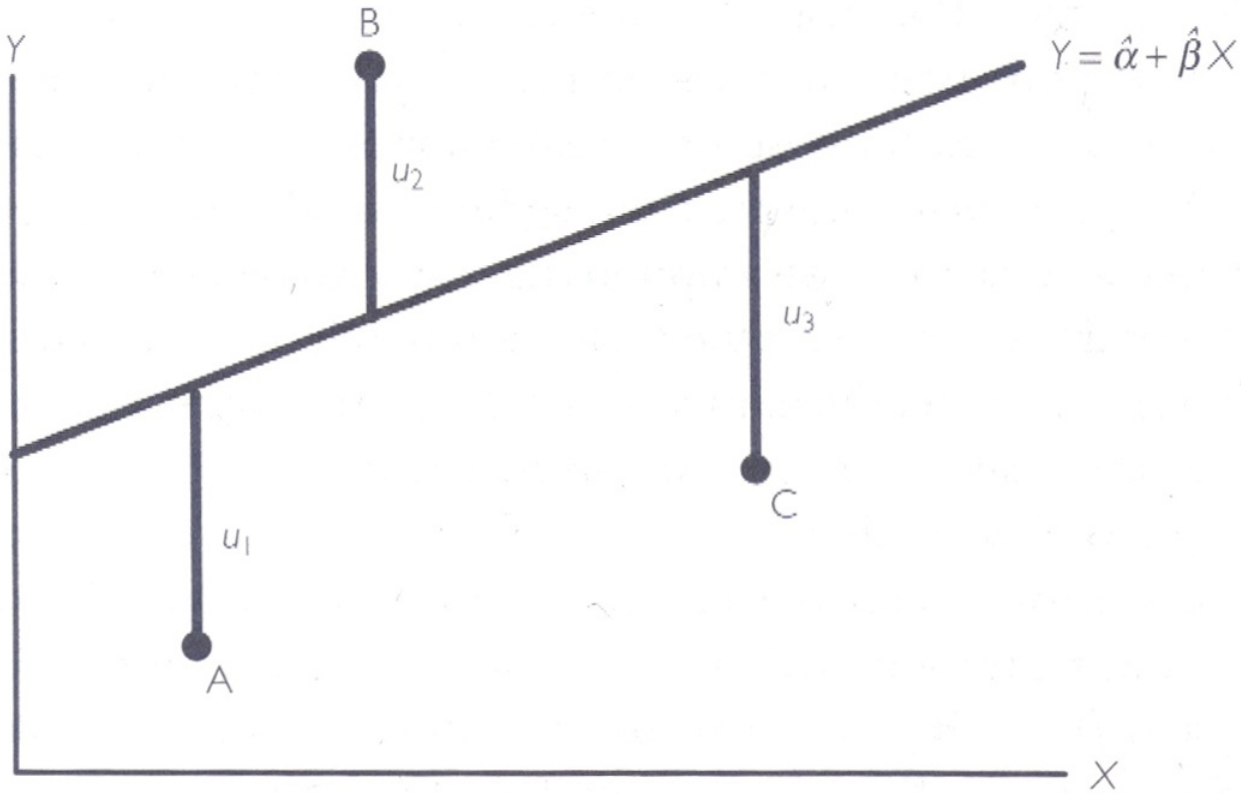
residual sum of squares = sum of areas of squares



Least Squares Estimates for Regression

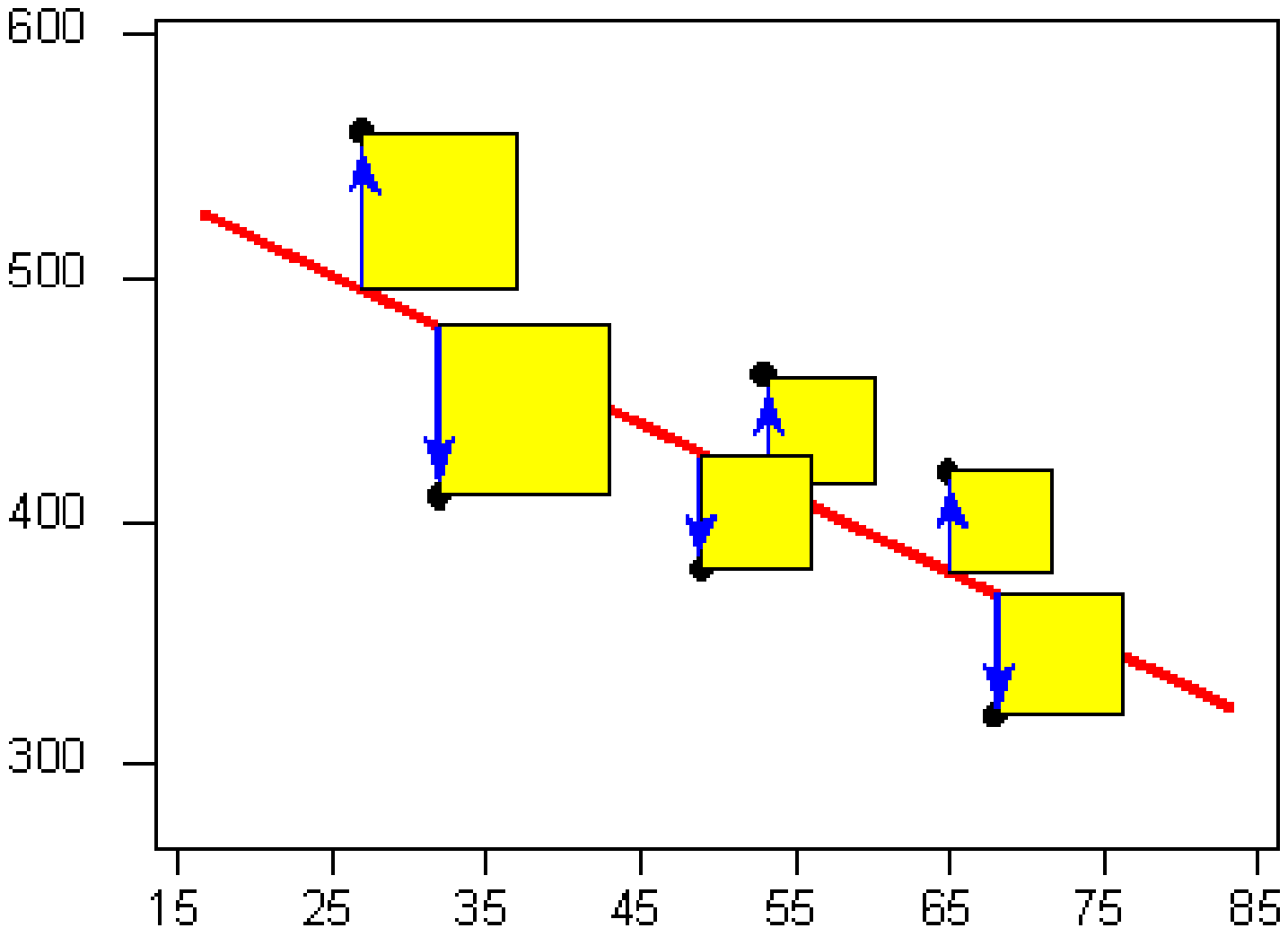
The optimal model is determined by minimising the sum of squared residuals $e_i = u_i^2$ defined by

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



Least Squares Estimates for Regression

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

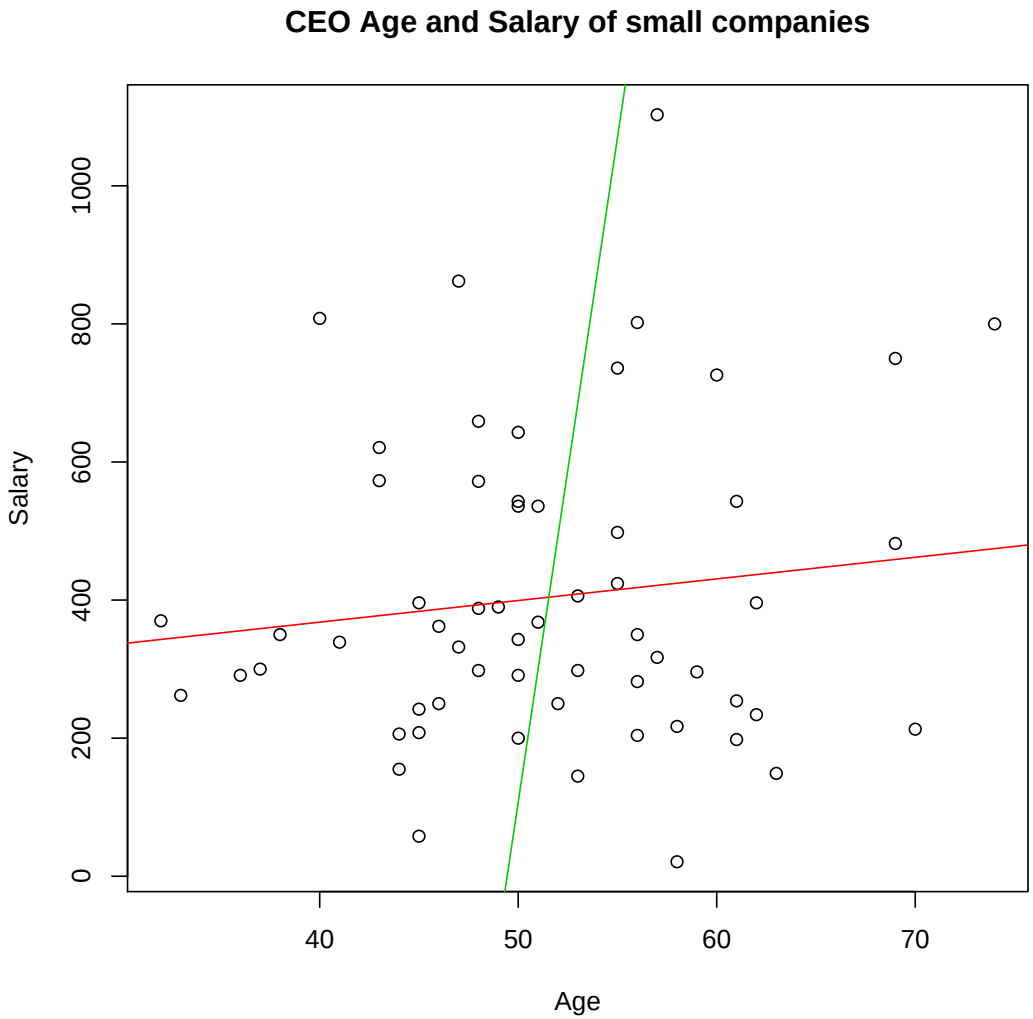
$$\begin{aligned} \hat{\beta}_{XY} &= \frac{\widehat{\text{cov}}(X,Y)}{s_X^2}, \\ r_{XY} &= \frac{\widehat{\text{cov}}(X,Y)}{s_X s_Y}. \end{aligned}$$

But

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

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

We assume that the model can be written in the form

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

where the error terms are

- **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. necessary assumption for testing and estimating confidence bounds for parameters and the regression line itself!

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

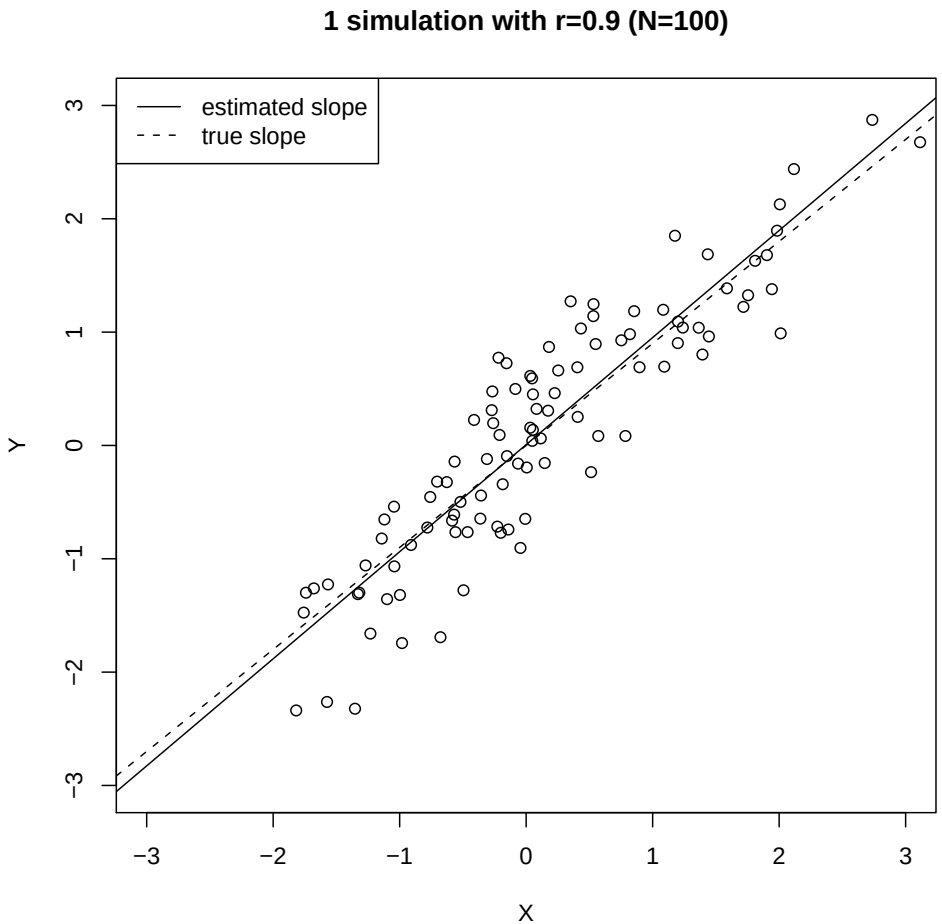
$$s_b = \sqrt{\frac{\text{SSR}}{(N-2)\sum(X_i - \bar{X})^2}}.$$

Standard errors

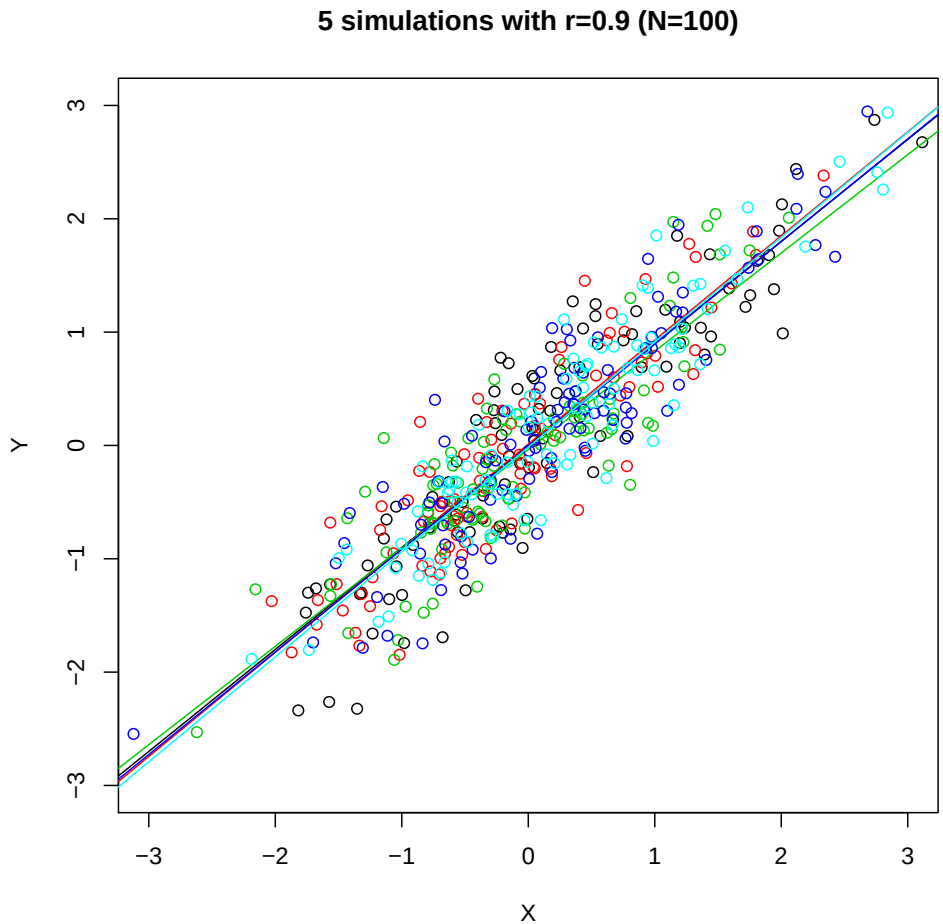
The slope standard deviation formula is consistent with the three factors that influence the precision of $\hat{\beta}$:

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).

About the precision of $\hat{\beta}$

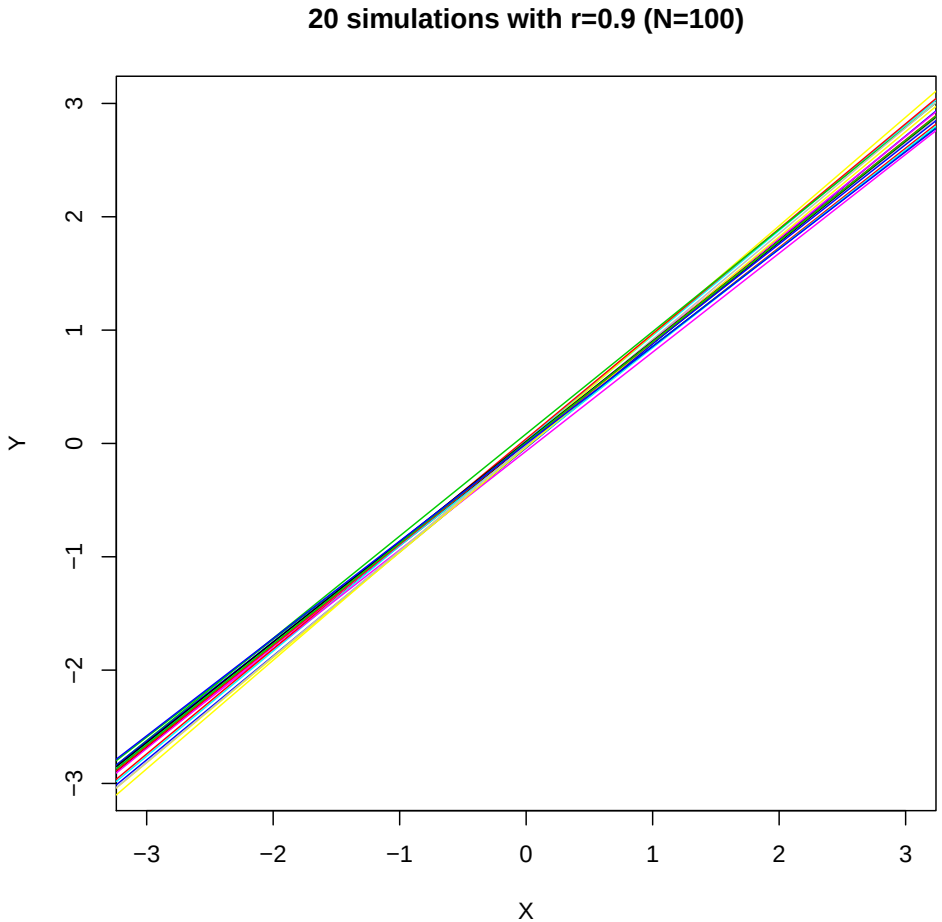


$$0.9448 \leq \hat{\beta} \leq 0.9448$$

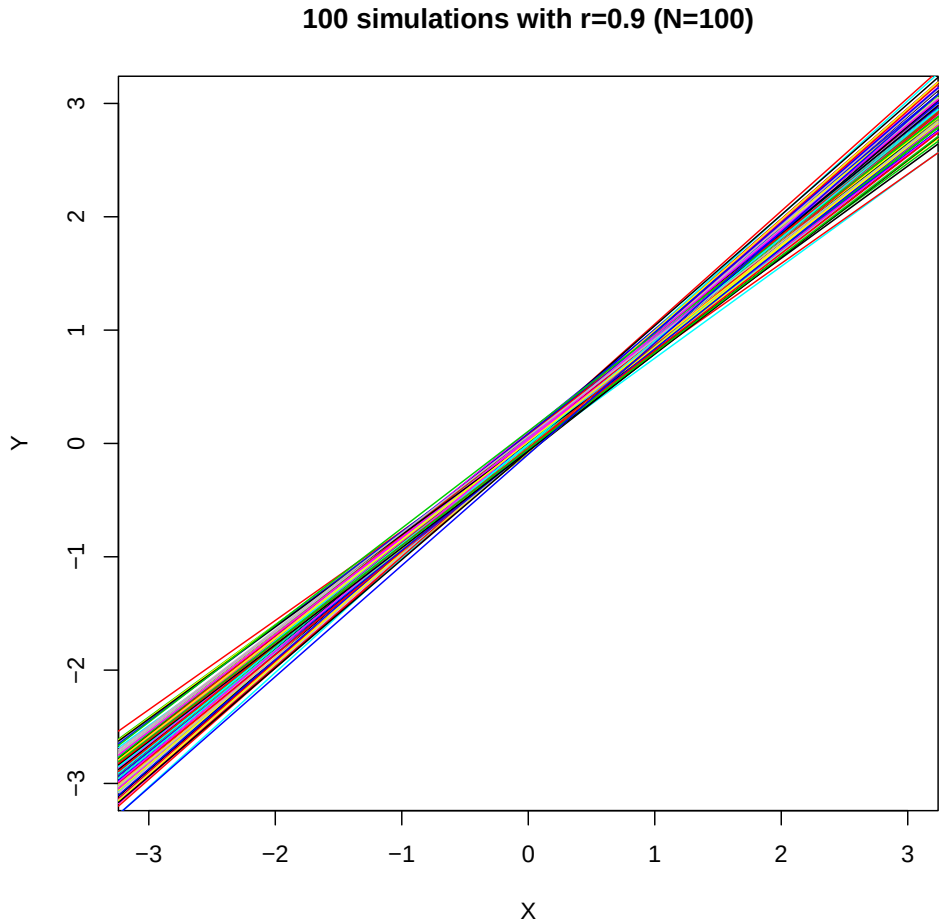


$$0.8677 \leq \hat{\beta} \leq 0.9448$$

About the precision of $\hat{\beta}$

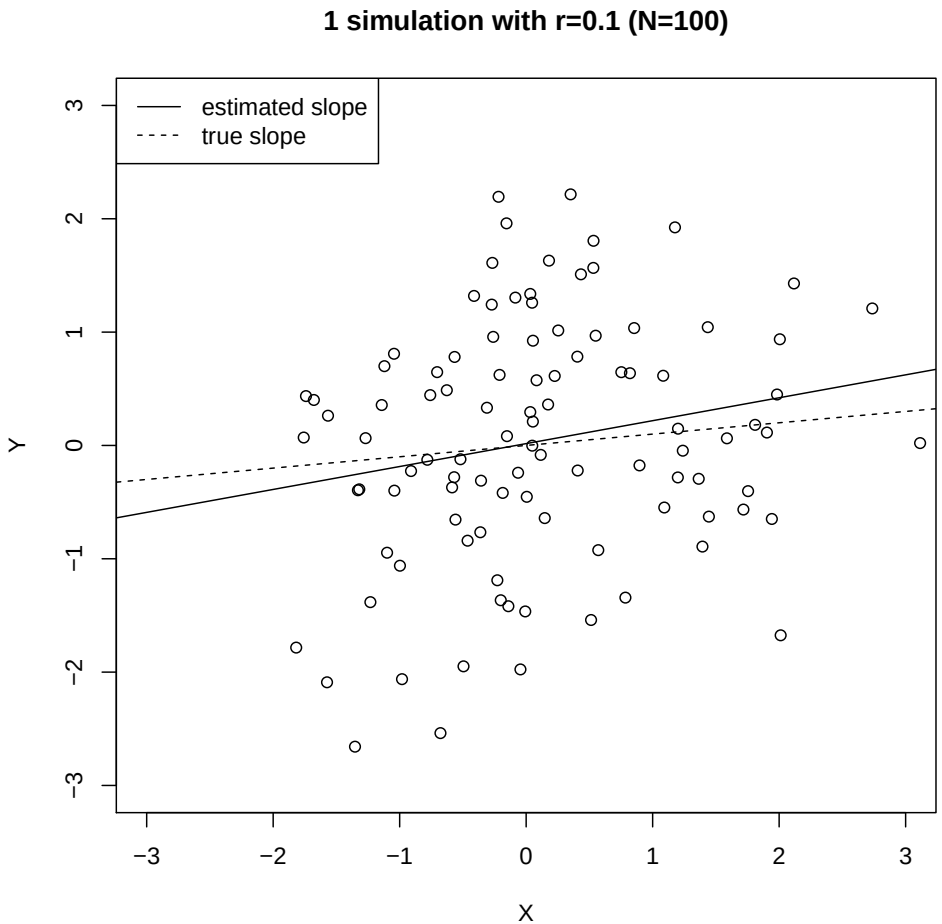


$$0.8588 \leq \hat{\beta} \leq 0.9582$$

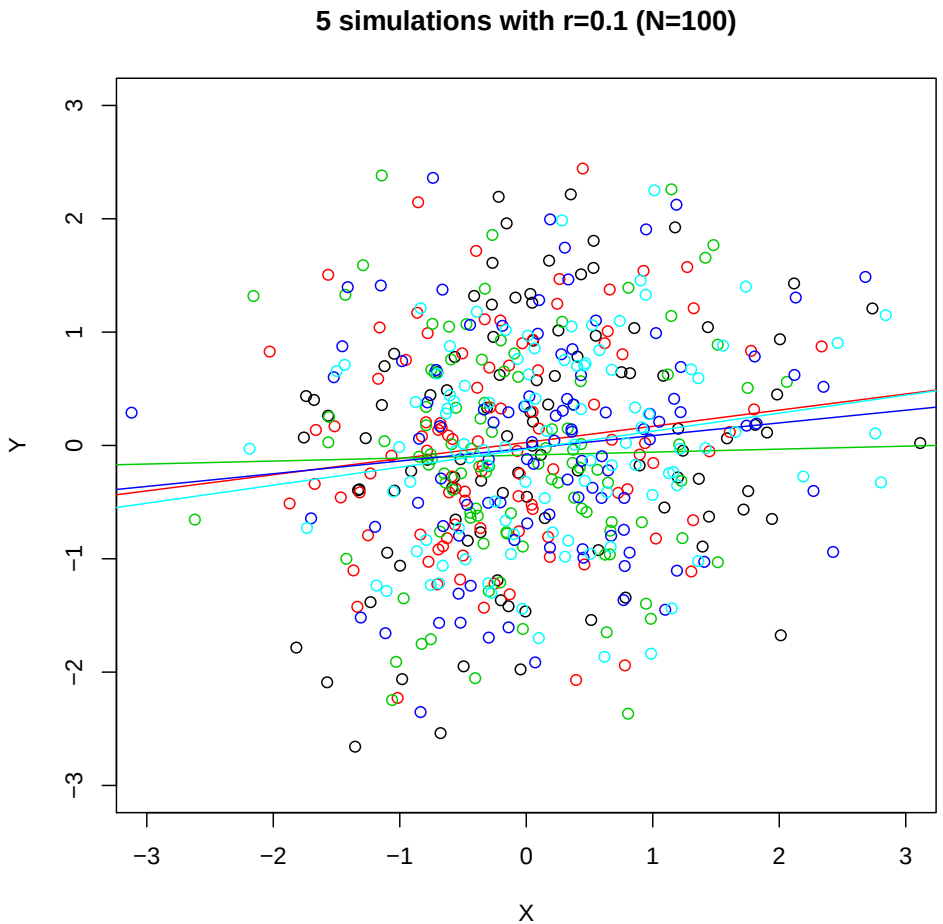


$$0.7874 \leq \hat{\beta} \leq 1.0089$$

About the precision of $\hat{\beta}$

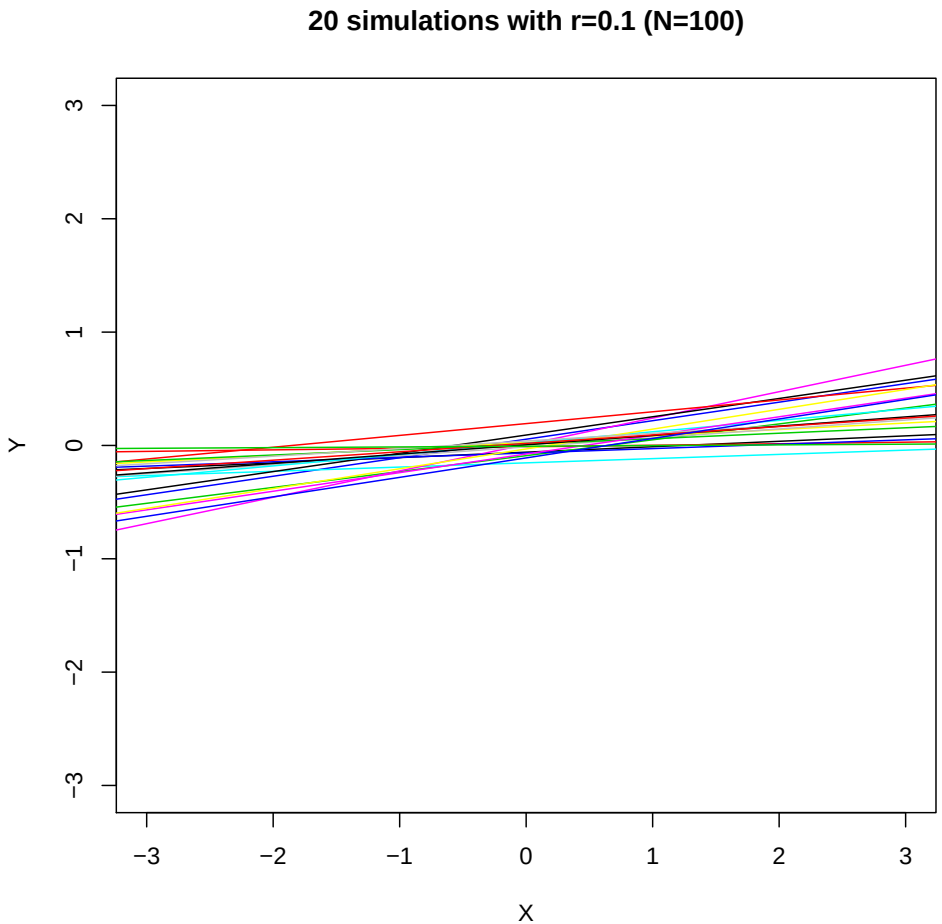


$$0.2022 \leq \hat{\beta} \leq 0.2022$$

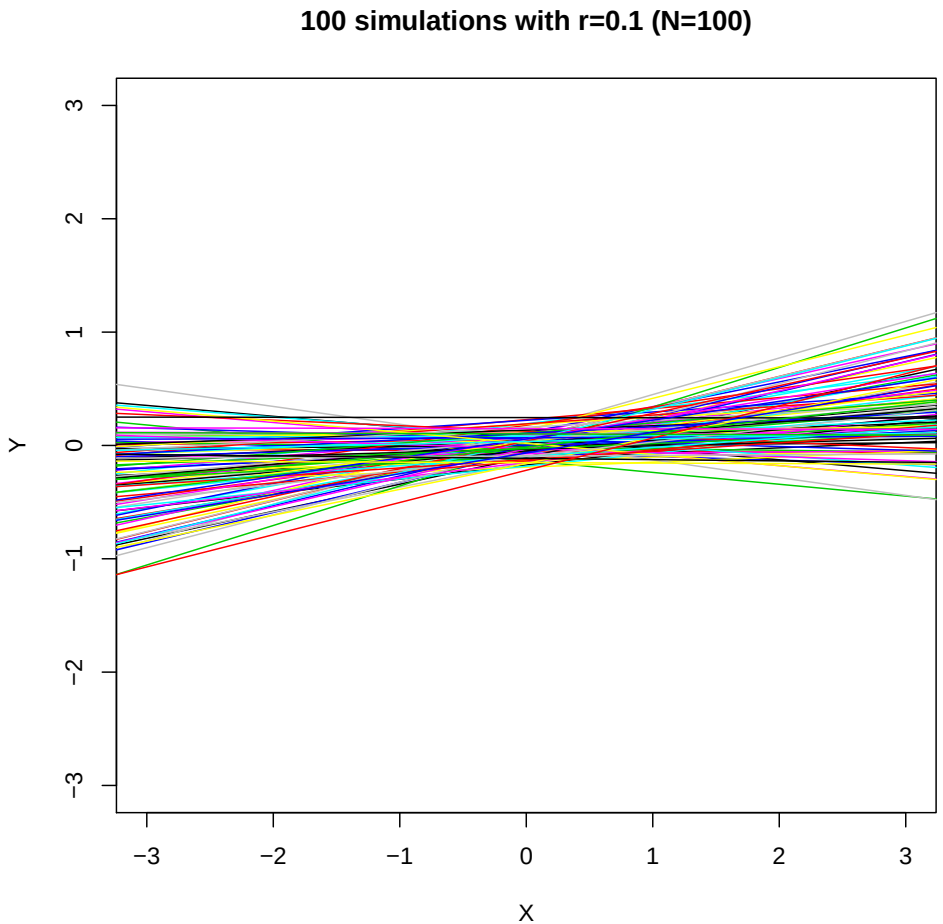


$$0.0263 \leq \hat{\beta} \leq 0.2022$$

About the precision of $\hat{\beta}$



$$0.0061 \leq \hat{\beta} \leq 0.2329$$



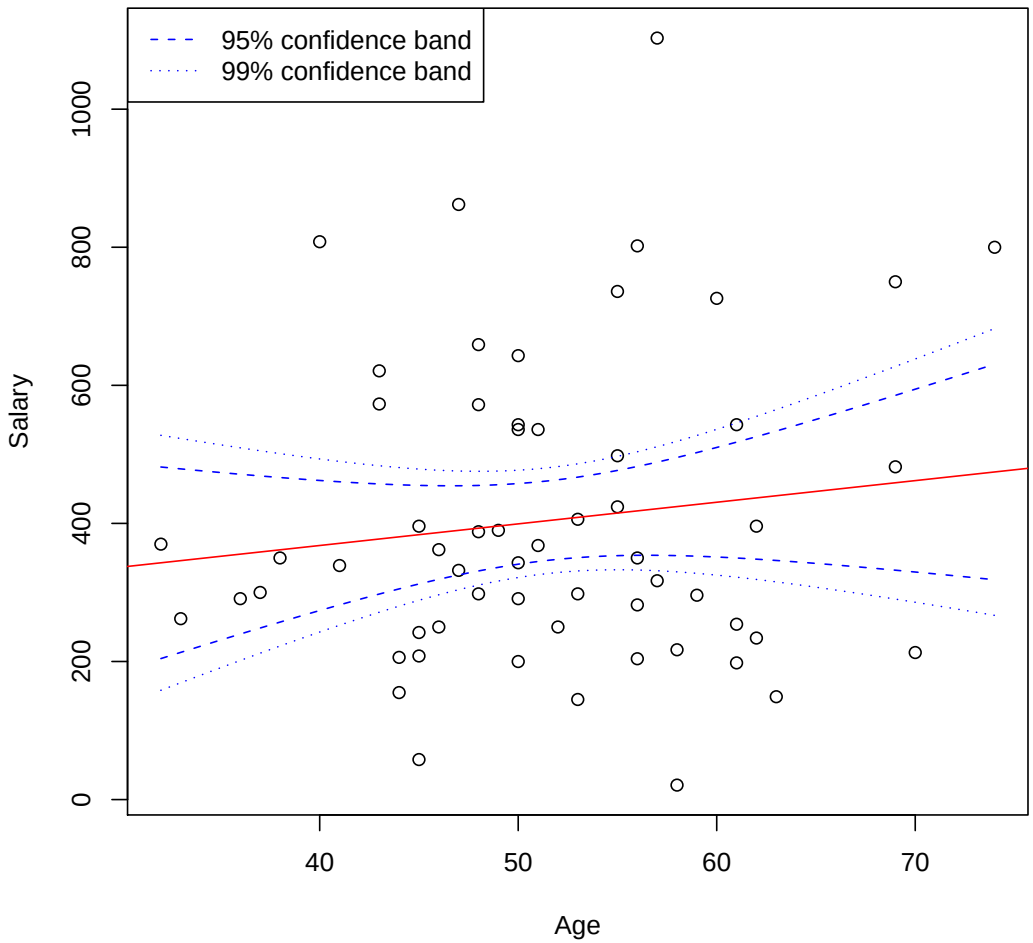
$$-0.1569 \leq \hat{\beta} \leq 0.3486$$

Confidence and Prediction Bands

Confidence and prediction bands for the regression line arise

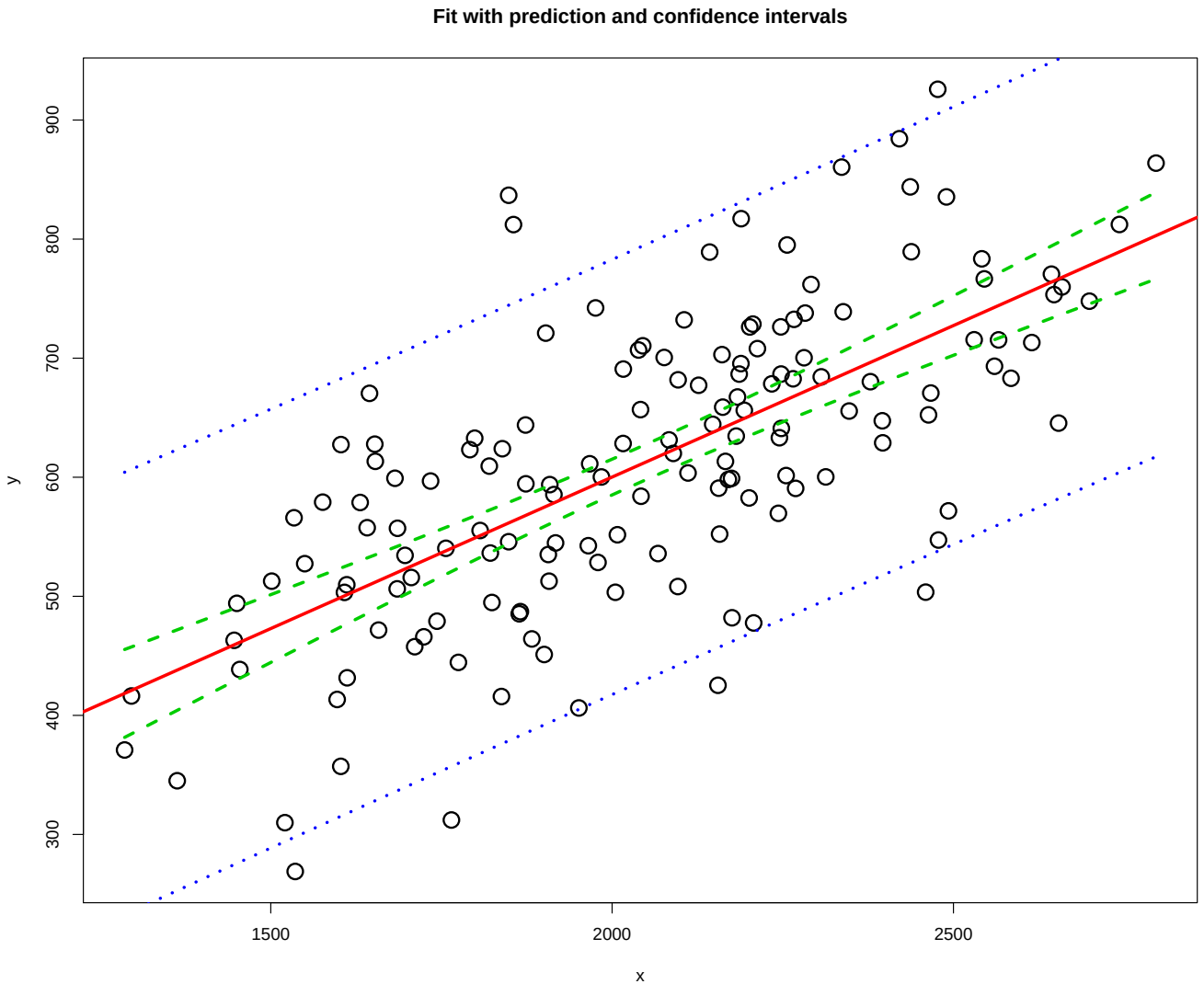
- **pointwise**, i. e. at each value of $\boldsymbol{x}_i$ a confidence interval for a fixed confidence level α is calculated
  Caution: although for each point the *confidence level* α is kept, the *coverage probability* α need not be kept for the regression line as a whole
- **simultaneous**, i. e. 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



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

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. e. influential.

$$H_0: \quad \alpha \text{ or } \beta \quad = 0$$

$$H_A: \quad \alpha \text{ or } \beta \quad \neq 0$$

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.

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

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

Summary

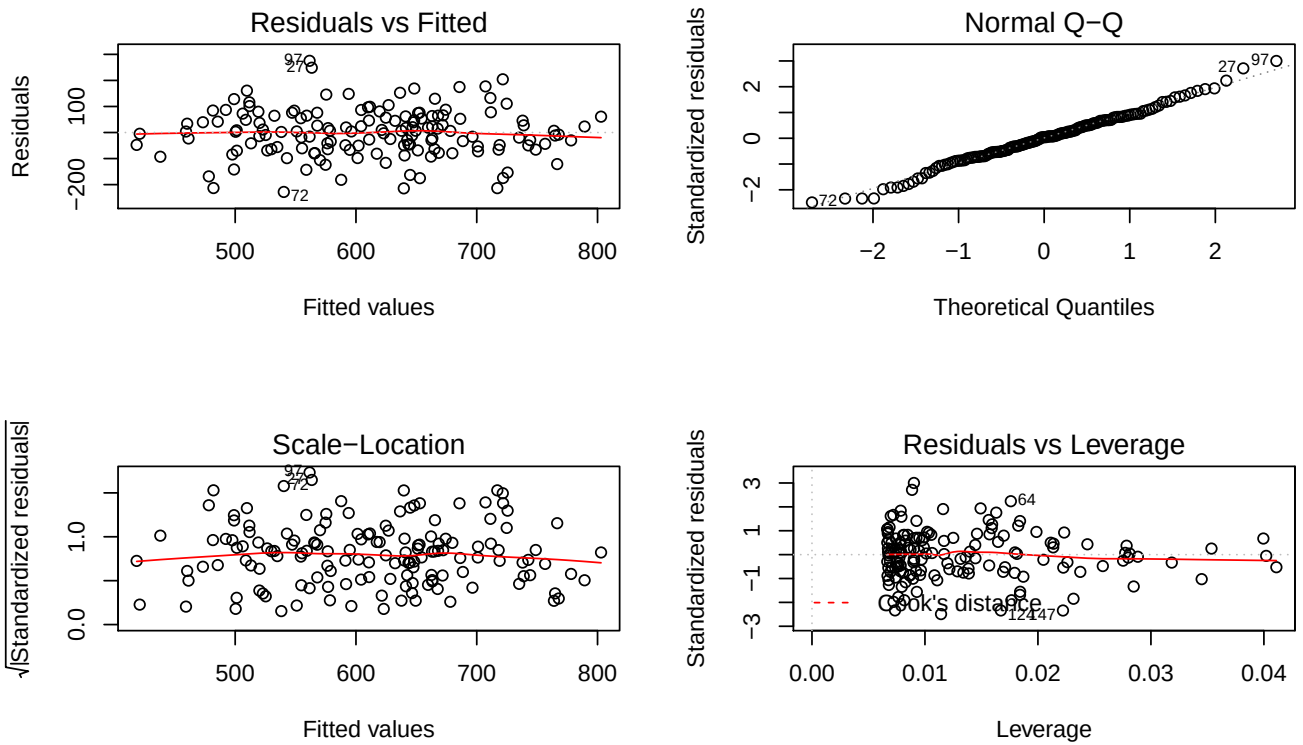
Given that the model assumptions are fulfilled, regression techniques provide the following information about β:

- $\hat{\beta}$, 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 $\beta = 0$.

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-228.101  -59.325    3.949   61.650  275.267

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 91.17348   46.56991   1.958   0.0521 .
x            0.25449    0.02244  11.339   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 92.14 on 148 degrees of freedom
Multiple R-squared:  0.4649, Adjusted R-squared:  0.4613
F-statistic: 128.6 on 1 and 148 DF,  p-value: < 2.2e-16
```

Residual plots

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



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 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 $x_{k,.}$ have to be stochastically independent

OLS estimates for multiple regression

The *ordinary least squares (OLS)* estimates:

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

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 $\hat{\beta}$ is the **best linear unbiased estimator** (BLUE).

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**

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

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**

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

where n is the number of observations.

R

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

Anscombe data example

The historical data collected by Anscombe et al. in 1970 contain the expenditures of U. S. States on public schools for the U. S. states and Washington, D. C. separately.
The following variables are observed:

education	Per-capita education expenditures, dollars.
income	Per-capita income, dollars.
young	Proportion under 18, per 1000.
urban	Proportion urban, per 1000.

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.

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.

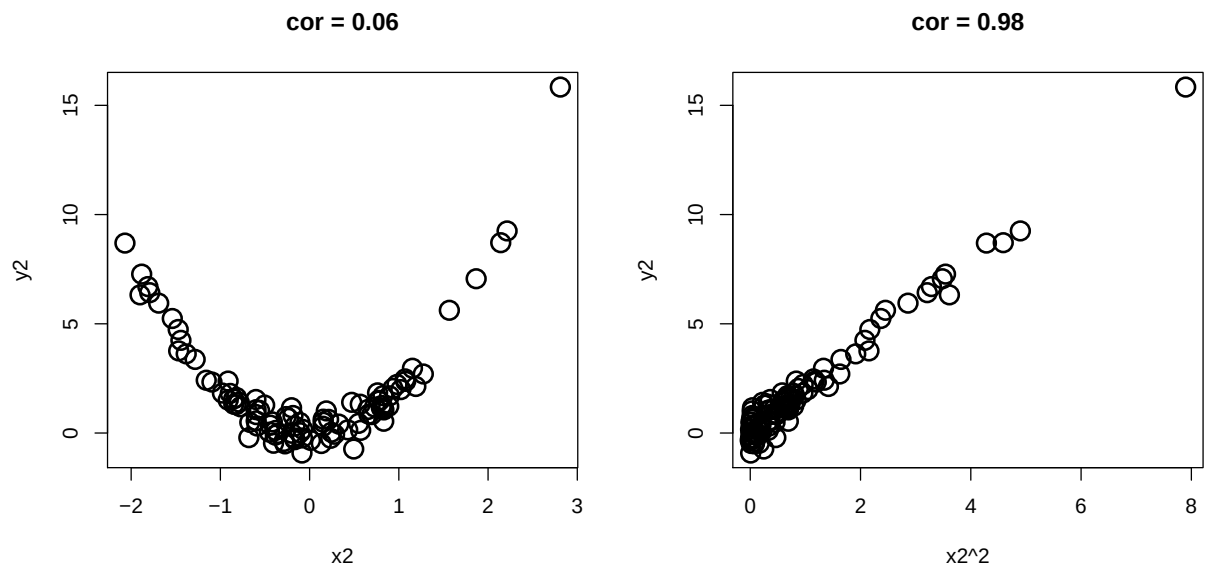
Non-linear Transformations

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

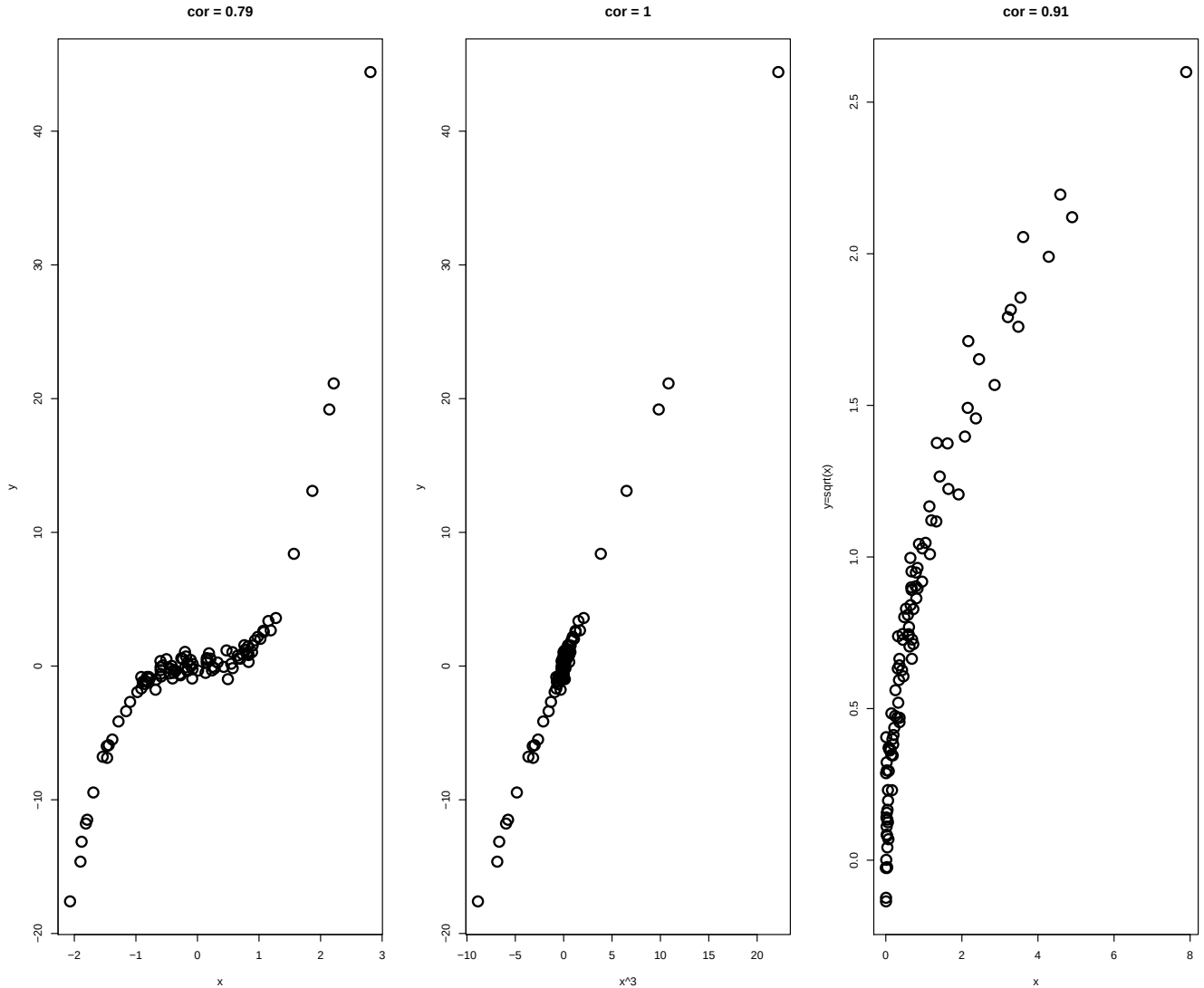
$$P_i = X_i^k$$

most importantly, the quadratic transformation is

$$Q_i = X_i^2$$



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

$$L_i = log(X_i)$$

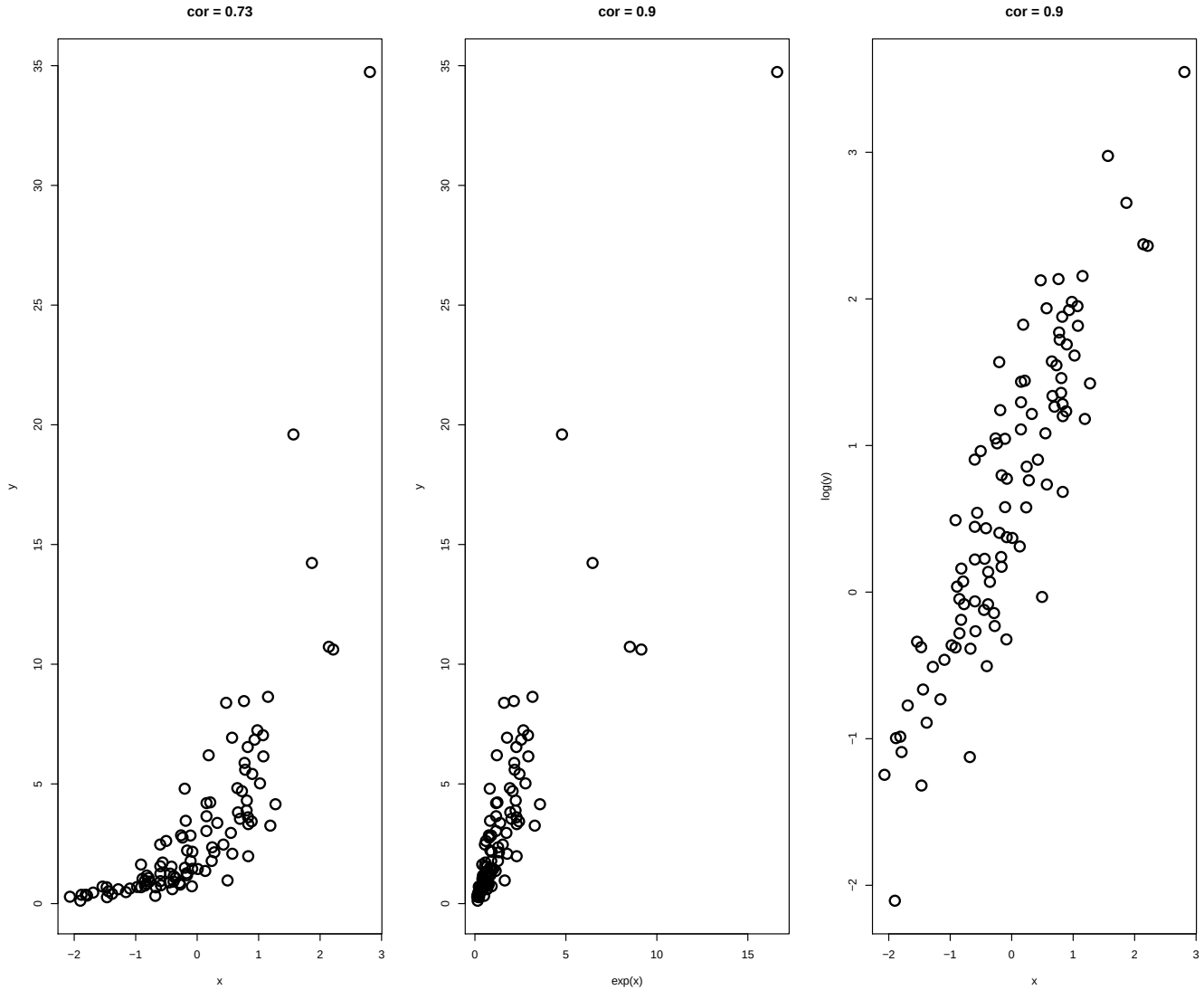
These two transformations form the bridge between the class of exponential models

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

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: