DSS
Decision Sciences & Systems

Technische Universität München

Tutorial Business Analytics

Tutorial 3: Linear Regression

Decision Sciences & Systems (DSS)
Department of Informatics
TU München

Tutorial Business Analytics

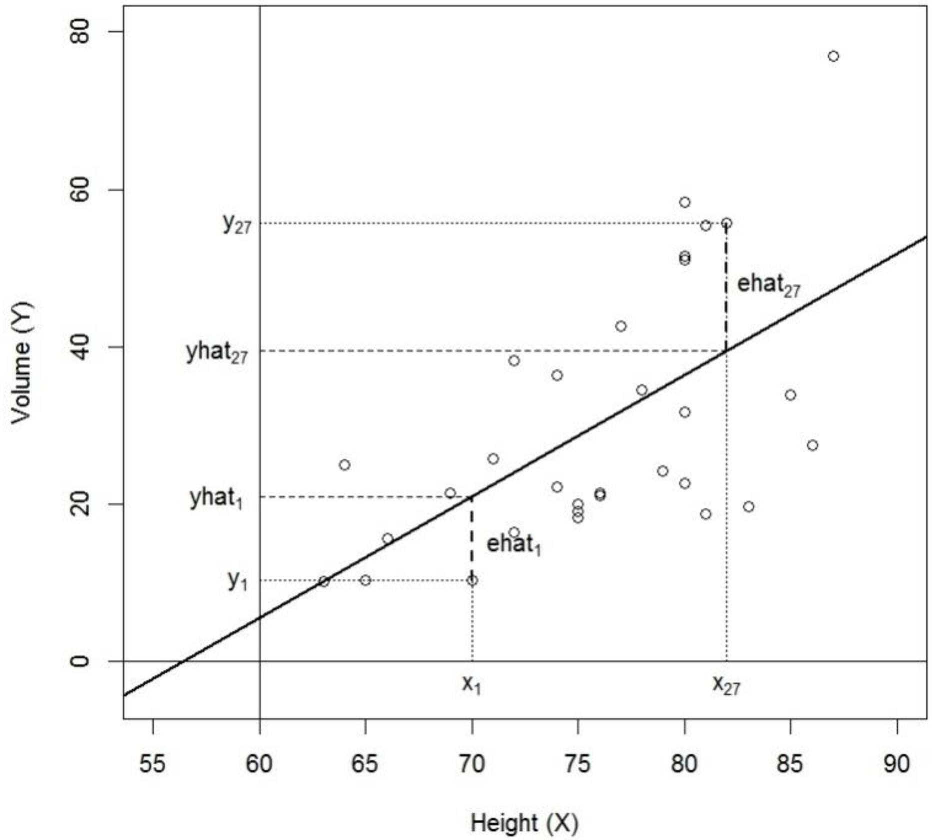
Agenda

1. Simple Linear Regression
2. Multiple Linear Regression
3. Significance Tests of Estimators
4. Model Evaluation
5. Gauss-Markov Theorem
6. Panel Regression

Tutorial Business Analytics

Simple Linear Regression

- Fitting a linear function through the data: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$



- X: **predictor variable**
- Y: **response variable**
- **Residual** e_i is the difference between the observed y_i and predicted $\hat{y}_i$:

$$e_i = y_i - \hat{y}_i$$
$$= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Tutorial Business Analytics

Finding the estimators

- Squared error of a point (residual): $e_i^2 = \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2$

- Residual Sum Squares: $\text{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2$

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \left\{RSS = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2\right\}$$

… (set partial derivatives equal to zero)

$$\Rightarrow \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\Rightarrow \quad \hat{\beta}_1 = \frac{Cov(x,y)}{Var(x)} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2} = \frac{\frac{1}{n}\sum_i^n x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum_i^n x_i^2 - \bar{x}^2} = \frac{S_{xy}}{S_{xx}}$$

Tutorial Business Analytics

Interpreting the estimators of a simple linear regression model

- $\boldsymbol{\hat{\beta}_0}$:

The output of the linear regression model when the predictor variable (x_i) is set to 0.
Also called the intercept on y.

- $\boldsymbol{\hat{\beta}_1}$:

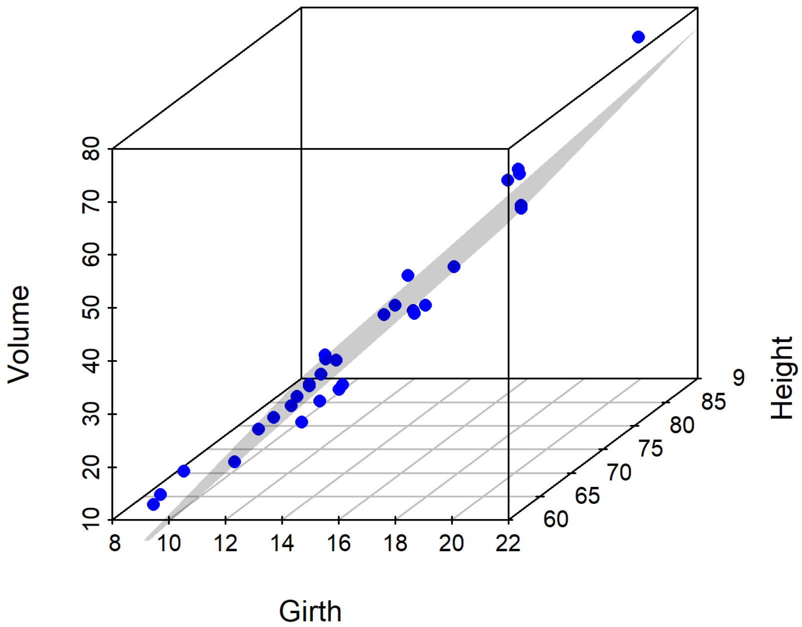
The change in $\hat{y}_i$, for each unit increase in x_i. Also called the slope on y.

Note: If the variables are transformed, they have to be interpreted differently!
E.g. $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{log}(\boldsymbol{x_i})$: If x_i increases by **1%**, then $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \log(1.01 * x_i) = \hat{\beta}_0 + \hat{\beta}_1 \log(x_i) + \hat{\beta}_1 \log(1.01)$. So the change in $\hat{y}_i$ equals $\boldsymbol{\hat{\beta}_1 log}(\mathbf{1.01})$.

Tutorial Business Analytics

Multiple Linear Regression

- Fitting a linear function through the data: $\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij} \Leftrightarrow y = \mathbf{X}\beta + \varepsilon$



- X: **predictor variables**
- Y: **response variable**
- **Residual** e_i is the difference between the observed y_i and predicted $\hat{y}_i$:

$$e = y - \mathbf{X}\hat{\beta}$$

Tutorial Business Analytics

Finding the estimators

- Squared error of a point (residual): $e_i^2 = \left(y_i - (\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij})\right)^2$

- Residual Sum Squares: $\text{RSS} = e^T e = (y - \mathbf{X}\hat{\beta})^T (y - \mathbf{X}\hat{\beta})$

$$\min_{\hat{\beta}} \left\{ RSS = (y - \mathbf{X}\hat{\beta})^T (y - \mathbf{X}\hat{\beta}) \right\}$$

… (take derivative and use FOC and SOC)

$$\Rightarrow \qquad \hat{\beta} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T y$$

Tutorial Business Analytics

Interpreting the estimators of a multiple linear regression model

- $\widehat{\boldsymbol{\beta}}_{\mathbf{0}}$:

The output of the linear regression model when all predictor variables (x_{ij}) are set to 0.
Also called the intercept on y.

- $\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}}$:

The change in y_i, for each unit increase in x_{ij}, while keeping the other predictor variables constant.
Also called the partial slope on y.

Note: If the variables are transformed, they have to be interpreted differently!

Tutorial Business Analytics

Testing the significance of regression coefficients

- Follow “test manual ” from Tutorial 2 to do the Hypothesis testing

- The **test statistic** is calculated as follows:

$$t_0 = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

$$SE(\hat{\beta}_1) = \sqrt{\frac{RSS}{\sum_{i=1}^{n}(x_i - \bar{x})^2} * \frac{1}{n-2}}$$

Tutorial Business Analytics

When to **reject H_0**?

H_1	using p-value	using test statistic
$\hat{\beta}_j \neq 0$	p < α	$\lvert t_0 \rvert \geq \left\lvert t^c_{1-\frac{\alpha}{2};df} \right\rvert$
$\hat{\beta}_j > 0$	p < α	$t_0 \geq t^c_{1-\alpha;df}$
$\hat{\beta}_j < 0$	p < α	$t_0 \leq t^c_{\alpha;df}$

Tutorial Business Analytics

Evaluation of model

Measure the difference between true observations and the regression line

- Residual Sum of Squares (RSS):

$$\text{RSS} = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2$$

- Mean Squared Error (MSE):

$$\text{MSE} = \frac{RSS}{n}$$

- Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{MSE}$$

- Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS}$$

Tutorial Business Analytics

Exemplary R Output

```
> myModel = lm(trees$Volume ~ trees$Girth + trees$Height)
> summary(myModel)

Call:
lm(formula = trees$Volume ~ trees$Girth + trees$Height)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.4065 -2.6493 -0.2876  2.2003  8.4847 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -57.9877     8.6382  -6.713 2.75e-07 ***
trees$Girth    4.7082     0.2643  17.816  < 2e-16 ***
trees$Height   0.3393     0.1302   2.607   0.0145 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared:  0.948,     Adjusted R-squared:  0.9442 
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

Tutorial Business Analytics

Gauss-Markov Theorem

Property	What does it mean?	Why do we need that?	How can we test that?
Linearity	Regression linear in the coefficients β	Core assumption of **linear** regression	Do not transform β, only the covariates
No Multicollinearity	• $rank(\mathbf{X}) = p$ • No high correlation between covariates	• Impossible to estimate coefficients • Non-significant coefficients	Variance Inflation Factor
Homoskedasticity	$Var(\varepsilon_i \| \boldsymbol{X}) = \sigma^2 \ \forall i$	• Some observations have more „weight" • Biased standard errors	• White Test • Breusch-Pagan Test
No Autocorrelation	$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0 \ \ \forall i, j$	• Omitted variables • Functional misfit • Measurement errors	Durbin-Watson Statistic
Exogeneity	$\mathrm{E}(\varepsilon_i \| \boldsymbol{X}) = 0 \ \ \forall i$	• Omitted variables • Measurement errors	Instrument Variables

Under these assumptions, the OLS estimator is BLUE

Tutorial Business Analytics

Panel regression

- **Fixed Effects Model:**

$$y_{it} = (\beta_0 + \lambda_i) + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_p x_{pit} + \varepsilon_{it}$$

- **Random Effects Model:**

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_p x_{pit} + \lambda_i + u_{it}$$

- **Lagrange Multiplier Test:** Test of individual effects for panel models

H_0: No individual effects

- **Hausman Test:** Test of fixed effects vs. random effects

H_0: Random effects estimator is consistent and efficient

Tutorial Business Analytics

Agenda

1. Simple Linear Regression
2. Multiple Linear Regression
3. Significance Tests of Estimators
4. Model Evaluation
5. Gauss-Markov Theorem
6. Panel Regression