



Tutorial Business Analytics

Tutorial 3: Linear Regression
Decision Sciences & Systems (DSS)
Department of Informatics
TU München

Tutorial Business Analytics

Today's agenda

1. Theory:

- Model relationship of numerical variables using linear regression model
- Estimators
 - Find the best estimators
 - Interpreting estimators
 - Significance of estimators
- Evaluation of model

2. Practice: **Exercise 3.1**

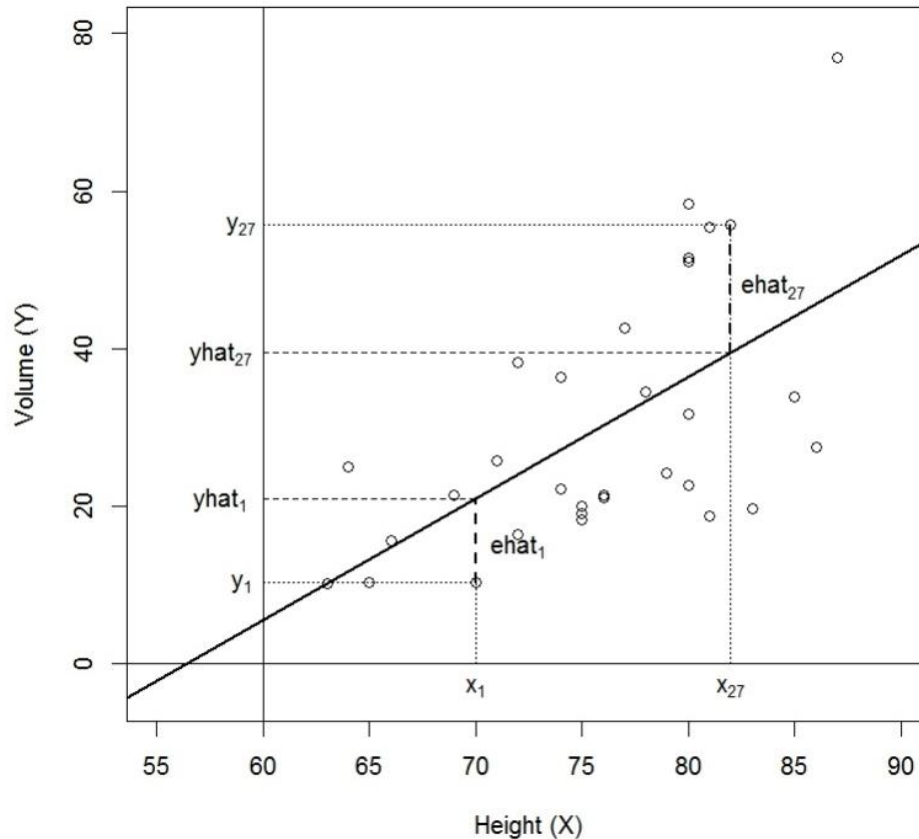
3. Practice: **Exercise 3.2 (R)**

Homework: **Exercise 3.3** and **Exercise 3.4**

Tutorial Business Analytics

Regression line

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



- X is **predictor variable**
- Y is **respond variable**
- Residual** \hat{e}_i , is difference between the observed y_i and predicted \hat{y}_i :

$$\hat{e}_i = y_i - \hat{y}_i$$

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

Tutorial Business Analytics

Finding the estimators

- Squared error of a point (residual): $\hat{e}_i^2 = (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$
- Residual Sum Squares: $RSS = \sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$


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

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

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

Tutorial Business Analytics

Interpreting the estimators of linear regression model

- $\hat{\beta}_0$: 

The output of linear regression model when all predictor variables (x_j) are set to 0.

Also called the intercept on y .

- $\hat{\beta}_j$: 

The change in y , for each unit increase in x_j , while keeping other predictor variables constant.

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



Tutorial Business Analytics

Testing the significance of estimated values, $\hat{\beta}_1$

- 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

t-distribution

t Table											
cum. prob	<i>t</i> _{.50}	<i>t</i> _{.75}	<i>t</i> _{.90}	<i>t</i> _{.95}	<i>t</i> _{.99}	<i>t</i> _{.95}	<i>t</i> _{.975}	<i>t</i> _{.99}	<i>t</i> _{.995}	<i>t</i> _{.999}	<i>t</i> _{.9995}
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
Z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Tutorial Business Analytics

When to **reject H_0** ?




H_1	using p-value	using test statistic
$\hat{\beta}_j \neq 0$	$p < \alpha$	$ t_0 \geq t_{1-\frac{\alpha}{2};df}^c $
$\hat{\beta}_j > 0$	$p < \alpha$	$t_0 \geq t_{1-\alpha;df}^c$
$\hat{\beta}_j < 0$	$p < \alpha$	$t_0 \leq t_{\alpha;df}^c$

Tutorial Business Analytics

Evaluation of model

Measure the difference between true observations and the regression line

- Residual Sum of Squares (RSS):


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

- Mean Squared Error (MSE):

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

- Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{MSE}$$

Tutorial Business Analytics

Exercise 3.1

The following table displays the per capita gross national product (X in 1000\$) and the percentage of literate people among the population (Y).

Country	X	Y
Nepal	0.5	5
Uganda	0.6	28
Thailand	1.0	68
South Korea	1.4	77
Peru	1.8	48
Lebanon	3.6	48
Ireland	5.7	98
France	6.4	96
New Zealand	13.0	99

Note:

$$\sum x_i = 34, \sum x_i^2 = 262.22,$$

$$\sum y_i = 567, \sum x_i y_i = 2914.3,$$

$$\bar{x} = 3.78, \bar{y} = 63$$

- a) Calculate the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ of the simple linear regression model using the ordinary least squares. Find the regression line using the formulas below:

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

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

Tutorial Business Analytics

b) Interpret the coefficients calculated in exercise a).

Tutorial Business Analytics

c) Test the null hypothesis $H_0: \hat{\beta}_1 \leq 0$ with significance level $\alpha = 0.05$

Use the following t-test with:

- $RSS = 4411.4$
- $\sum_{i=1}^n (x_i - \bar{x})^2 = 133.77$

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

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

t Table											
cum. prob	$t_{.50}$	$t_{.75}$	$t_{.90}$	$t_{.95}$	$t_{.99}$	$t_{.995}$	$t_{.9975}$	$t_{.999}$	$t_{.9995}$	$t_{.99975}$	$t_{.9999}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Tutorial Business Analytics

- d) Now the above linear regression model will be used to estimate the percentage of literates among a country with known gross national product.

Which problems might occur? Briefly explain your concerns using an example.

Tutorial Business Analytics

e) Repeat c) using R (Exercise 3.1_R-Script.R).

Tutorial Business Analytics

Exercise 3.2

t	Demand
0	28.20
1	37.65
2	47.28
3	59.76
4	73.44
5	86.19
6	100.31
7	112.58
8	121.63
9	
10	

Note: You may use R to solve this exercise (exercise 3.2_R-Script.R).

a) For the time series above, calculate the forecasted demand value for $t = 10$ using the simple linear regression and the formula below:

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 \cdot t$$

Tutorial Business Analytics

b) Calculate the RMSE and explain its meaning.

Tutorial Business Analytics

c) For the time series above, calculate the forecasted demand value for $t = 10$ assuming a biannual seasonal component of the following form:

Starting from the first period $t = 0$, suppose after every second period a new year begins. Make use of the formula below:

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 \cdot t + \hat{\beta}_2 \cdot Q_1$$

Tutorial Business Analytics

d) Does the data reflect biannual data?