L16

Basic Statistics for Economists

Spring 2020

Department of Statistics

Stockholm University

Today

Sections **12.1 – 12.4** + parts of **12.5** – **12.6 NCT**

- Several explanatory variables: $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$, …
- Model specification, how it is done
- Parameter estimate
- Coefficient of determination
- Inference:
  - Confidence interval and test for individual coefficients
  - test of the whole model

Multiple variable regression model

- The dependent or the endogenous variable: Y
- K exogenous or explanatory variables: $X_1, X_2, X_3, \ldots, X_K$
- A natural extension of the simple linear model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon \quad \text{where } \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

Interpretation of the coefficients:

- Intercept β_0 = expected value of Y given $X_1 = X_2 = \cdots = X_K = 0$
- Coefficient β_j = expected increase of Y when X_j increases by 1 and ***everything else is held constant***
  - The marginal change that is related to X_j

Example 2

Household consumption = *f*(disposable income)

$\boldsymbol{Y}$ = household consumption, $\boldsymbol{X_1}$ = disposable income

$$\mu_{Y|x} = \beta_0 + \beta_1 x_1$$

- If disposable income increases by one unit the average consumption increases by β_1 units:

$$\textbf{Diff} = \left(\mu_{Y|x_1+1} - \mu_{Y|x_1}\right) = \beta_0 + \beta_1(x_1 + 1) - \beta_0 - \beta_1 x_1 = \boldsymbol{\beta_1}$$

By it turns out (as expected) that household consumption also depends on the size of the household.

Example 1, cont.

H.hold consump. = *f*(disposable income, household size)

Add: $\boldsymbol{X_2}$ = household size

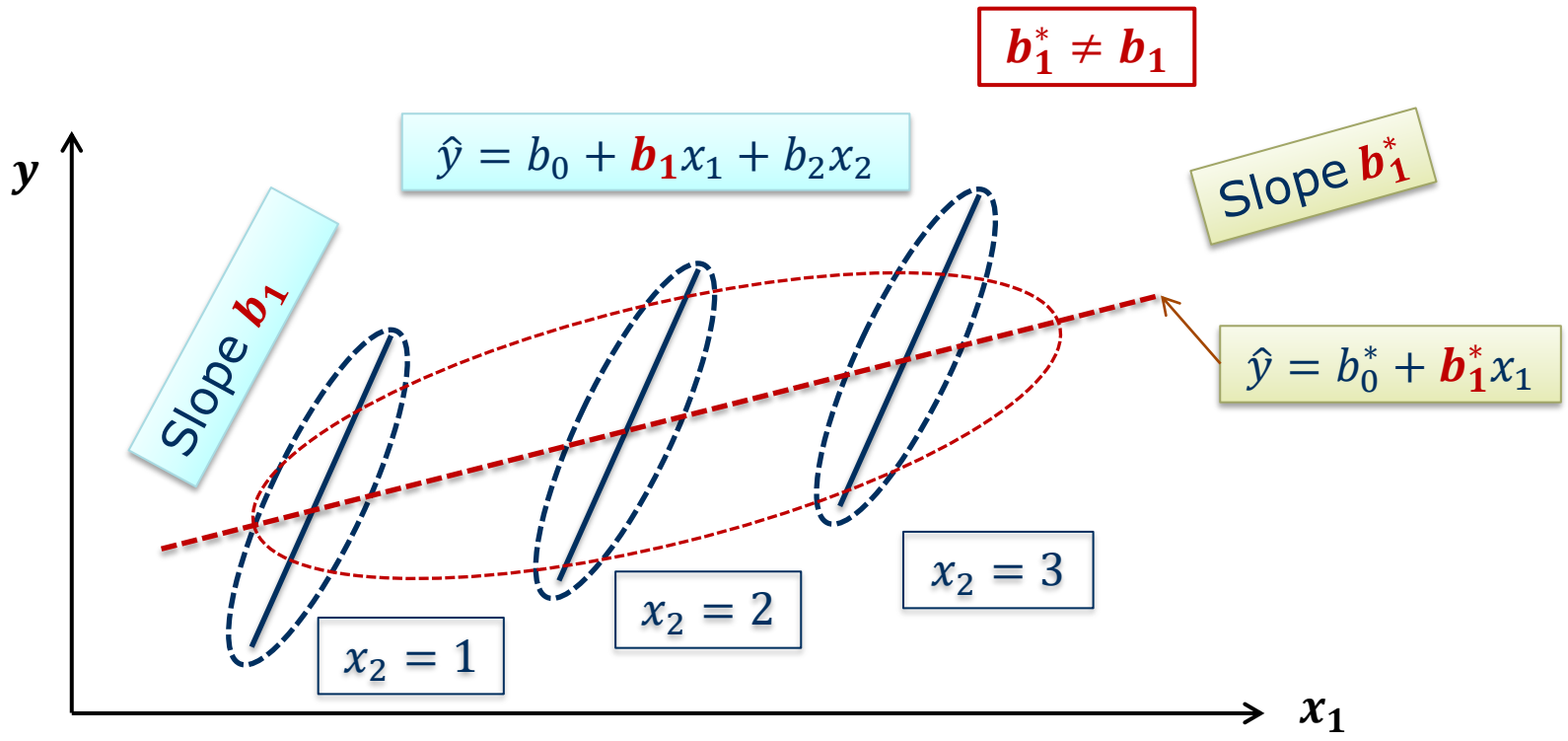
$$\mu_{Y|x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- If disposable income x_1 increases by one unit the average consumption changes by β_1 units, ***given that the household size x_2 stays the same***, i.e. if ***size is held constant***:

$$\textbf{Diff} = \left(\mu_{Y|x_1+1,x_2} - \mu_{Y|x_1,x_2}\right) =$$

$$= [\beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2] - [\beta_0 + \beta_1 x_1 + \beta_2 x_2] = \boldsymbol{\beta_1}$$

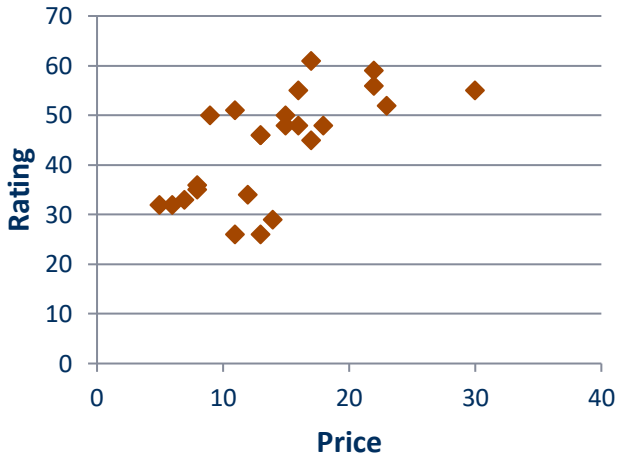
Example 1, cont.



Stockholm University

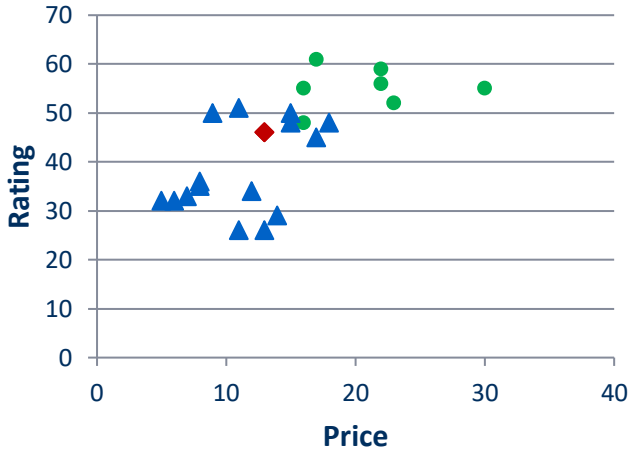
Example 2

ID	Type	Rating	PricePerLoad
1	Liquid	61	17
2	Liquid	59	22
3	Liquid	56	22
4	Liquid	55	16
5	Liquid	55	30
6	Liquid	52	23
7	Powder	51	11
8	Powder	50	15
9	Powder	50	9
10	Liquid	48	16
11	Powder	48	15
12	Powder	48	18
13	Gel	46	13
14	Gel	46	13
15	Powder	45	17
16	Powder	36	8
17	Powder	35	8
18	Powder	34	12
19	Powder	33	7
20	Powder	32	6
21	Powder	32	5
22	Powder	29	14
23	Powder	26	11
24	Powder	26	13



$s_{xy} = 43,2$

$r_{xy} = 0,67$



$r_{xy} = 0,42$

$r_{xy} = 0,07$

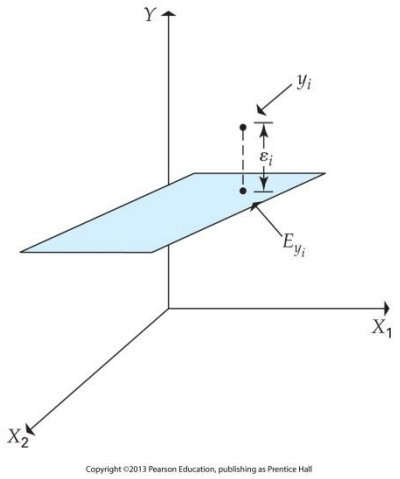
$r_{xy} =$ not def.

Stockholm University

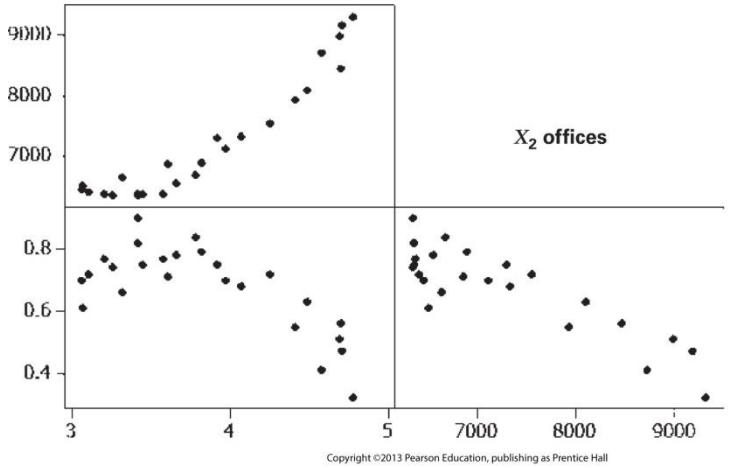
Graphically - diagram

- Combine colors (see previous p.)

- 3 dimensions, maybe …
  - E.g. Figure 12.2, s. 479 NCT



- Common to present a matrix consisting of several scatter plots where relationships between all pairs of $\boldsymbol{Y}$, $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, …, $\boldsymbol{X_K}$ are shown
  - E.g. Figure 12.6 s. 487 NCT



Stockholm University

Assumptions of regression model

1. Y is a ***linear*** function of x plus a random error term ε:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

2. x_{ki} values, $k = 1, \ldots, K$, are assumed to either be constant or realized values of the r.v. $X_1, \ldots, X_K$ all of which are ***independent*** of the error term ε_i. $X_1, \ldots, X_K$ may be pairwise dependent…

3. The error terms ε_i are ***normally distributed*** r.v. with ***expectation*** $E(\varepsilon_i) = 0$ and ***constant variance*** σ_ε^2.
   - Homoskedasticity

4. The error terms are pairwise ***independent***.

Assumptions of regression model, cont.

5. Predictor variables $\boldsymbol{X_1, \ldots, X_K}$ cannot be dependent of each other in such a way that there exists a combination of constants c_1, c_2, …, c_K where every $c_j \neq 0$ such that

$$\boldsymbol{c_1 x_{1i} + \cdots + c_K x_{Ki} = 0}$$

- E.g. $\boldsymbol{X_1, \ldots, X_K}$ may not be deterministically linearly dependent of each other
- This is called **multicolinearity**
- May feel a little abstract at the moment, but we will return to this during L17

Estimation of the model parameters

- Method of Least-Squares, OLS
- When we have more than one $\boldsymbol{K = 1}$ independent variable, this becomes more complicated to solve
- ***We will not need to compute coeffients! Use computer!***
- The principle is the same: minimize

$$\sum_{i=1}^{n}(y_i - b_0 - b_1 x_{1i} - \cdots - b_K x_{Ki})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{SSE}$$

Solved mathematically by taking the derviatives of SSE with respect to b_0, b_1, …, b_K, set the $K + 1$ derivatives equal to zero and solve for b_0, b_1, …, b_K.

Predictions & Residuals

- Comment: instead of pairs (x_i, y_i), we have ***vectors*** $(x_{1i}, \ldots, x_{Ki}, y_i)$

- When the parameters have been estimated, we calculate the ***predicted values*** $\hat{y}_i$ as before $i = 1, \ldots, n$

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_K x_{Ki}$$

- Then the **residuals** e_i are calculated for $i = 1, \ldots, n$ as

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

- ***Let a computer do it!***

ANOVA and degrees of freedom

- **SST** = $(n-1)\cdot \boldsymbol{s_y^2} = \sum_{i=1}^{n}(y_i - \boldsymbol{\bar{y}})^2$

  $\bar{y}$ is calculated before s_y^2 can be calculated ⇒ $\boldsymbol{n-1}$ d.f.

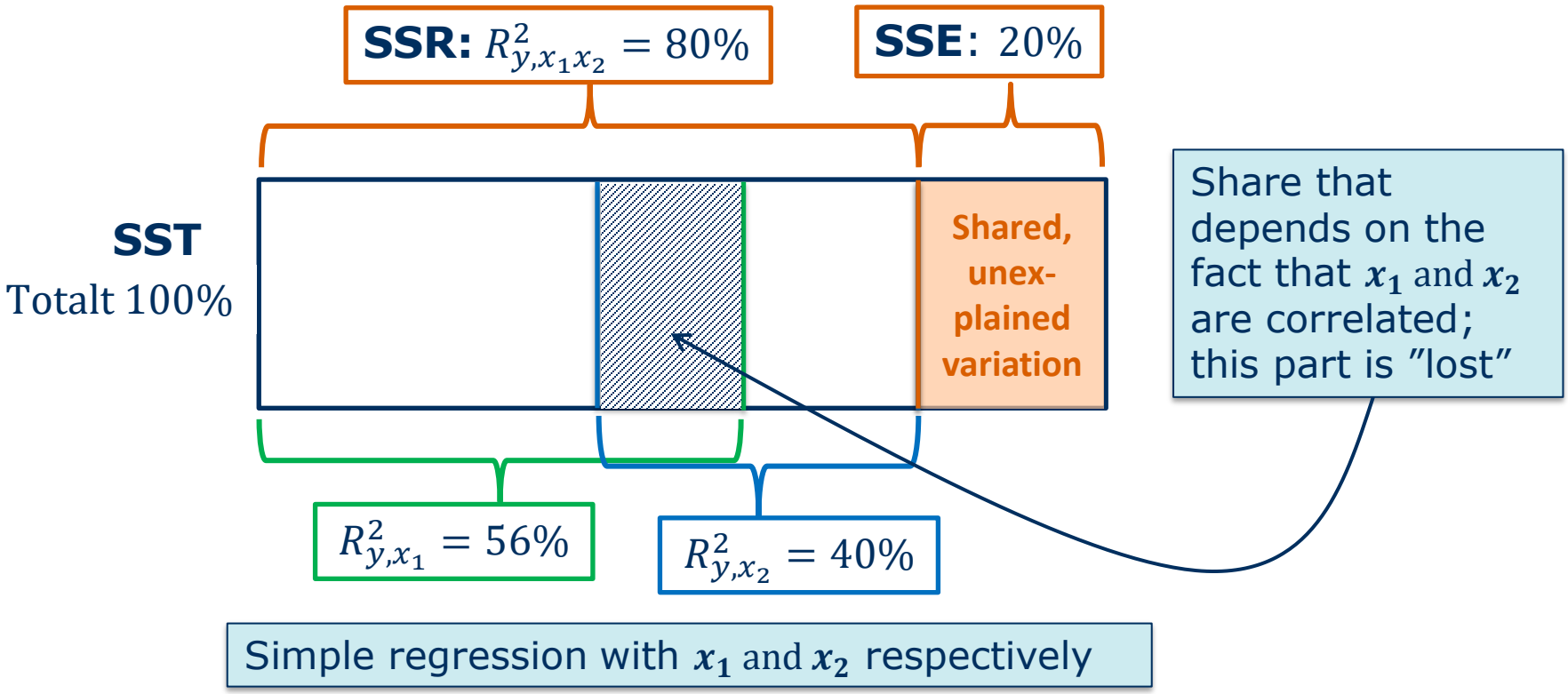
- **SSE** = $(n-K-1)\cdot \boldsymbol{s_e^2} = \sum_{i=1}^{n}(y_i - \boldsymbol{b_0} - \boldsymbol{b_1}x_{1i} - \cdots - \boldsymbol{b_K}x_{Ki})^2$

  b_0 and b_1,…, b_K are calculated before s_e^2 can be calculated
  ⇒ $\boldsymbol{n-K-1}$ d.f.

- **SSR** = $1\cdot\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ = SST – SSE

  $\mathrm{fg}_{\mathrm{regression}} = \mathrm{fg}_{\mathrm{total}} - \mathrm{fg}_{\mathrm{error}} = \boldsymbol{K}$ degrees of freeom left
  K explanatory variables ⇒ $\boldsymbol{K}$ degrees of freedom for SSR

Comparison to simple regression



Stockholm University

See Formulary

Coefficient of determination R^2

- Proportion of the total variation in $\boldsymbol{Y}$ that can be explained by the model:

$$\boldsymbol{R^2} = \frac{\boldsymbol{SSR}}{\boldsymbol{SST}} = \frac{\boldsymbol{SST - SSE}}{\boldsymbol{SST}} = \boldsymbol{1 - \frac{SSE}{SST}}$$

  $\boldsymbol{R^2}$ always increased as more explanatory variables are added!

- $\boldsymbol{R^2}$ adjusted for the number of **degrees of freedom**:

$$\bar{\boldsymbol{R}}^{\boldsymbol{2}} = \boldsymbol{R^2_{adj}} = 1 - \frac{MSE}{MST} = 1 - \frac{s_e^2}{s_y^2} = \boldsymbol{1 - \frac{SSE/(n-K-1)}{SST/(n-1)}} \qquad \boxed{\bar{\boldsymbol{R}}^{\boldsymbol{2}} < \boldsymbol{R^2}}$$

- $\boldsymbol{R^2_{adj}}$ can actually decrease when you add explanatory variables! If the variable that you add already is strongly correlated with a variable that is already in the model

Example - Excel: solution 12.82 in NCT

- **Correlation matrix**, shows corr. of every pair of variables
- Can be accessed in the Data Analysis tool in Excel
- Which variables should we used in the model?

	X1	X2	X3	Y
X1	1			
X2	-0,00406	1		
X3	-0,02163	-0,00029	1	
Y	0,29074	0,48716	-0,27071	1

Example - Excel

Check that you can calculate $\boldsymbol{R^2}$ and $\boldsymbol{\bar{R}^2}$ and $\boldsymbol{MSx = SSx/f.g.}$ for $x = T$, R and E

UTDATASAMMANFATTNING						
Regressionsstatistik						
Multipel-R	0,62678376					
R-kvadrat	0,392857881					
Justerad R-kvadrat	0,351461828					
Standardfel	3,624308448					
Observationer	48					
ANOVA						
	fg	KvS	MKv	F	p-värde för F	
Regression	3	373,9797506	124,6599	9,490225	5,9484E-05	
Residual	44	577,9669161	13,13561			
Totalt	47	951,9466667				
				ärde	Nedre 95%	Övre 95%
				31748	0,19664944	4,10425715
X1	0,493358413	0,20197985	2,442612	0,018666	0,08629477	0,900422054
X2	0,269912608	0,064939272	4,156385	0,000147	0,1390361	0,400789111
X3	-0,117090608	0,052041463	-2,24995	0,029504	-0,2219733	-0,01220793

$\boldsymbol{R^2}$ and $\boldsymbol{\bar{R}^2 = R^2_{adj}}$

ANOVA: deg. of freedom, SS and MS

Stockholm University

Properties for b_0 and b_j

- Estimates $\boldsymbol{b_0}$ and $\boldsymbol{b_j}$ are linear combinations of iid normally distributed random variables (will not be shown!). Means that

$$\boldsymbol{b_0 \sim N(\beta_0, \sigma^2_{b_0})} \qquad \boldsymbol{b_j \sim N\left(\beta_j, \sigma^2_{b_j}\right)} \quad \boldsymbol{j = 1, \ldots, K}$$

- Estimates $\boldsymbol{b_0}$ and $\boldsymbol{b_j}$ are ***unbiased***!
- Variances $\boldsymbol{\sigma^2_{b_0}}$ and $\boldsymbol{\sigma^2_{b_1}}$,…, $\boldsymbol{\sigma^2_{b_K}}$ are unknown; must be estimated:

Too complex to cover! Use computer software!

Example of K=2: see p. 495

- Use software computed estimates: $\boldsymbol{s^2_{b_0}}$ and $\boldsymbol{s^2_{b_1}}$,…, $\boldsymbol{s^2_{b_K}}$
- We will use the *t*-distribution.

Confidence interval for β_0 and β_j

- A $100(1-\alpha)$ % confidence interval for $\boldsymbol{\beta_1}$ is given by

$$\boldsymbol{b_j \pm t_{n-K-1\,;\,\alpha/2} \cdot s_{b_j}} \qquad j = 0, 1, \ldots, K$$

Same as K=1, see L15
Keep track of the degrees of freedom, f.g.!

  where $\boldsymbol{t_{n-K-1\,;\,\alpha/2}}$ is the table value from the *t*-dist. table

- Relies on the assumptions of the linear model

- REMEMBER! If we want to test $\boldsymbol{H_0: \beta_j = \beta_j^*}$ vs. $\boldsymbol{H_1: \beta_j \neq \beta_j^*}$ where $\boldsymbol{\beta_j^*}$ is a value of our choice, it should be enough to check that $\boldsymbol{\beta_j^*}$ lies inside the conficence interval.
  - E.g. for the hypotesis $\boldsymbol{H_0: \beta_j = 0}$, check whether 0 belongs to the interval

Hypothesis test for β_j

- Hypothesis test to find out if a coefficient is significant (adds to the model) ***given that everything else is in the model***:

  $$H_0: \beta_j = 0 \quad \text{vs.} \quad H_1: \beta_j \neq 0$$

- Assume: normally distributed iid error terms ε, independent of the x-values
- Test variable: $t_{n-K-1} = \dfrac{b_j - 0}{s_{b_j}}$ t-distributed with $n - K - 1$ d.f.
- Critical values and decision rule:
  - Reject H_0 if $|t_{obs}| > t_{n-K-1\,;\,\alpha/2}$

Same as the case K=1, L15
Just be aware of the d.f.!

Example - Excel

You can check that t_{obs} and the confidence intervals are derived from the estimates and standard errors!

UTDATASAMMANFATTNING						
Regressionsstatistik						
Multipel-R	0,62678376					
R-kvadrat	0,392857881					
Justerad R-kvadrat	0,351461828					
Standardfel	3,624308448					
Observationer	48					
	fg	KvS	MKv	F		
Regression	3	373,9797506	124,6599	9,490225	5,9484E-05	
Residual	44	577,9669161	13,13561			
Totalt	47	951,9466667				
	Koefficienter	Standardfel	t-kvot	p-värde	Nedre 95%	Övre 95%
Konstant	2,150453297	0,969452857	2,218213	0,031748	0,19664944	4,10425715
X1	0,493358413	0,20197985	2,442612	0,018666	0,08629477	0,900422054
X2	0,269912608	0,064939272	4,156385	0,000147	0,1390361	0,400789111
X3	-0,117090608	0,052041463	-2,24995	0,029504	-0,2219733	-0,01220793

Hypotesis test for every coefficient with t_{obs} and p values

Estimates, standard errors

95 % CI for every coefficient

Inference – to test the entire model

If $\boldsymbol{\beta_1 = \beta_2 = \cdots = \beta_K = 0}$ then the model useless. None of the explanatory variables help to explain the variation in $\boldsymbol{Y}$. Test!

- Hypotheses:

$$\boldsymbol{H_0}: \boldsymbol{\beta_1 = \beta_2 = \cdots = \beta_K = 0}$$

$$\boldsymbol{H_1}: \text{at least one } \boldsymbol{\beta_j \neq 0}, j = 1, \ldots, K$$

- Test variable:

$$\boldsymbol{F = \frac{SSR/K}{SSE/(n-K-1)} = \frac{MSR}{MSE} = \frac{MSR}{s_e^2} \sim F-\text{distribution}}$$

- We do not have to learn the *F*-distribution! We will determine by looking at the ***p*-value**! Next page!

Example - Excel

UTDATASAMMANFATTNING						
Regressionsstatistik						
Multipel-R	0,62678376					
R-kvadrat	0,392857881					
Justerad R-kvadrat	0,351461828					
Standardfel	3,624308448					
Observationer	48					
ANOVA						
	fg	KvS	MKv	F	p-värde för F	
Regression	3	373,9797506	124,6599	9,490225	5,9484E-05	
Residual	44	577,9669161	13,13561			
Totalt	47	951,9466667				
				-värde	Nedre 95%	Övre 95%
				,031748	0,19664944	4,10425715
X1	0,493358413	0,20197985	2,442612	0,018666	0,08629477	0,900422054
X2	0,269912608	0,064939272	4,156385	0,000147	0,1390361	0,400789111
X3	-0,117090608	0,052041463	-2,24995	0,029504	-0,2219733	-0,01220793

F-test for the entire model. Use p-value!

ANOVA: d.f., SS and MS

Stockholm University

Inference for conditional expectation

- Remember that the expectation of $\boldsymbol{Y}$ depends on the value of $\boldsymbol{x}$, we condition on the event $\boldsymbol{X} = \boldsymbol{x}$ and on $\hat{\boldsymbol{\mu}}_{\boldsymbol{Y}|\boldsymbol{x}_{n+1}} = \boldsymbol{b}_0 + \boldsymbol{b}_1\boldsymbol{x}_{1,n+1} + \cdots + \boldsymbol{b}_K\boldsymbol{x}_{K,n+1}$
  - NCT denotes this "new" x-value by $\boldsymbol{x}_{n+1}$

- A $100(1-a)\,\%$ CI for $\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{x}_{n+1}}$ is given by

A complicated formula that you don't need to know

- Unfortunately, Excel cannot calculate it for us!
- BUT REMEMBER: $\hat{\boldsymbol{\mu}}_{\boldsymbol{Y}|\boldsymbol{x}_{n+1}} = \boldsymbol{b}_0 + \boldsymbol{b}_1\boldsymbol{x}_{1,n+1} + \cdots + \boldsymbol{b}_K\boldsymbol{x}_{K,n+1}$
- So we can always compute a ***point estimate*** (by hand)

Predict future Y using regression

- A ***prediction interval*** for $\boldsymbol{y_{n+1}}$ given $\boldsymbol{x_{n+1}}$ is given by

A complicated formula that you don't need to know

- Unfortunately, Excel cannot calculate it for us!
- BUT REMEMBER: $\boldsymbol{\hat{\mu}_{Y|x_{n+1}} = b_0 + b_1 x_{1,n+1} + \cdots + b_K x_{K,n+1}}$
- So we can always compute a ***point estimate*** (by hand)

Next time

Selected topics from sections **12.8-9** + **13.1** + **13.4-6 NCT**

- Dummy variables
- Application multiple regression – how it is done
- Model building
- Specification bias, multicollinearity, heteroscedasticity