L15

Basic Statistic for Economists

Spring 2020

Department of Statistics

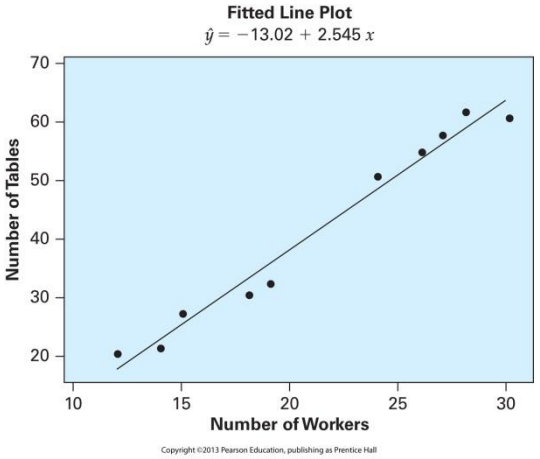
Stockholm University

Summary of last time

- Start with two r.v $\boldsymbol{X}$ and $\boldsymbol{Y}$ which covary ***linearly***

$$\boldsymbol{Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i} \quad \text{where} \quad \boldsymbol{\varepsilon_i \sim N(0, \sigma_\varepsilon^2)}$$

- Estimate according to OLS using formulas in the formulary
- We estimate the parameters $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ with $\boldsymbol{b_0}$ and $\boldsymbol{b_1}$
- Predictions: $\boldsymbol{\hat{y}_i = b_0 - b_1 x_i}$
- Residuals: $\boldsymbol{e_i = y_i - \hat{y}_i}$
- Residual variance: $\boldsymbol{s_e^2 = \frac{\sum_{i-1}^{n} e_i^2}{n-2}}$



Stockholm University

Assumptions of the regression model

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

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

2. The x_i values are assumed to be constant (set in advanced) or realizations of a r.v. X that are ***independent*** of the error terms ε_i.

3. The error terms ε_i are ***normally distributed*** r.v. with ***expectation*** $E(\varepsilon_i) = 0$ and ***constant variance*** σ_ε^2.
   - E.g. the variance is not different depending on x.

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

Stockholm University

Terms & definitions

- Y is called ***dependent***, ***endogenous***, ***predicted variable***, or ***regressor***
  - depends on X
  - The observations should pairwise independent (iid)
- X is called ***independent***, ***exogenuos***, ***explanatory-*** or ***predictor variable*** or ***regressor***
- The **residuals** e_i are not the same as the **error terms** ε_i
  - The error terms cannot be observed, they are hidden
  - The residuals are used as proxy for ε_i
  - But sometimes people mix up the terminology and call residuals "the errors"

Stockholm University

Terms & definitions, cont.

Parameters

- β_0 is called the ***intercept***, ***y-intercept*** or sometimes ***the constant***
- β_1 is called ***regression coefficient***, ***slope***, ***the slope coefficient***
- β_0 & β_1 are often together called the ***coefficients of the model***
- The regression line, the regression model, the model, … are often synonymous

Stockholm University

Today

- ANOVA and coefficient of determination $\boldsymbol{R^2}$ and adjusted R^2
- Inference for $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$
- Inference for $\boldsymbol{\mu_{Y|x}}$ conditional on a value $\boldsymbol{x}$
- Prediction of $\boldsymbol{y_{n+1}}$ given some value $\boldsymbol{x}$
- Brief discussion of regression and correlation
- A little more on graphic representation, what to look for
  - Model assumptions

Stockholm University

See formula sheet

ANOVA – breaking the variance into parts

- ANOVA = Analysis of Variances
- In the output of (all) statistical software
- Indirectly a method of comparing the variances σ_Y^2 and $\sigma_{Y|X=x}^2 = \sigma_\varepsilon^2$ or rather the estimates s_y^2 and $s_{y|x}^2 = s_e^2$
- **SST** = **S**um-of-**S**quares-**T**otal $= (n-1) \cdot s_y^2$
- **SSE** = **S**um-of-**S**quare-**E**rror $= (n-2) \cdot s_e^2$

here: Error = Residual

- **SSR** = **S**um-of-**S**quares-**R**egression = **SST** – **SSE**

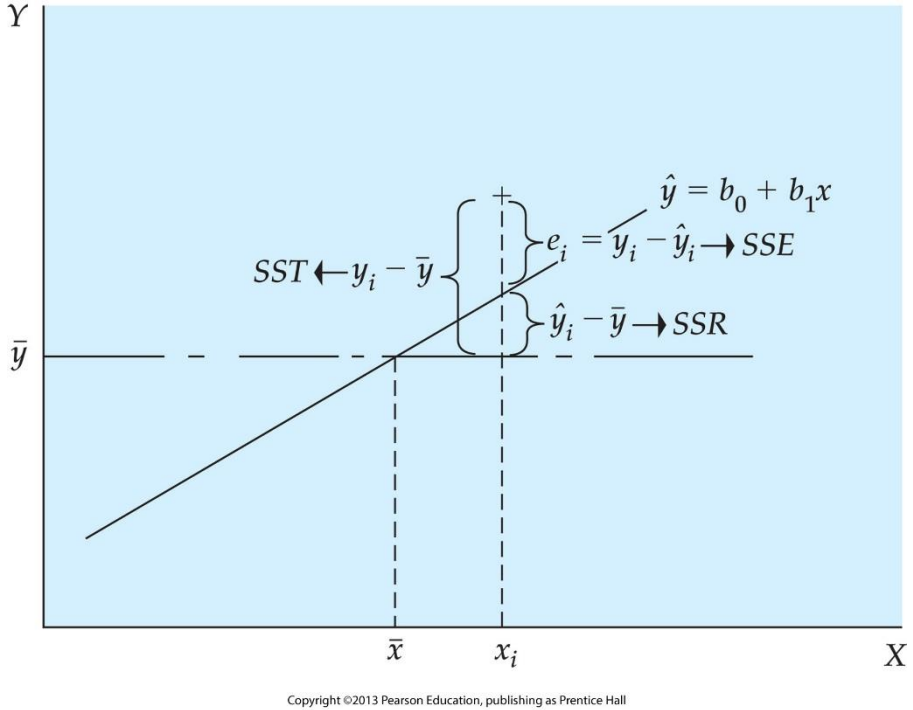
SST = SSR + SSE

Stockholm University

ANOVA, cont.

SST = SSR + SSE

- SST = total variation in *Y*
- SSR = the part of SST that the model explains
- SSE = rest, residual, the variation <u>not</u> explained by the model



Copyright ©2013 Pearson Education, publishing as Prentice Hall

Stockholm University

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 calc. ⇒ $\boldsymbol{n-1}$ deg. of freedom

- **SSE** = $(n-2)\cdot \boldsymbol{s_e^2} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \boldsymbol{b_0} - \boldsymbol{b_1}x_i)^2$

  b_1 and b_0 calculated before s_e^2 can be calc. ⇒ $\boldsymbol{n-2}$ d.f.

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

  f.g.$_{\text{Regression}}$ =f.g.$_{\text{Total}}$ – f.g.Er$_{\text{Error}}$ = $\boldsymbol{n-1-n+2=1}$ d.f.

  $K = 1$ explanatory variable ⇒ $\boldsymbol{1}$ degree of freedom for SSR

Stockholms universitet

ANOVA, cont.

Mean square sums – adjust for the degrees of freedom

- **MSE** = Mean Square Error = $SSE/(n-2) = \boldsymbol{s_e^2}$
- **MSR** = Mean Square Regression = $SSR/1$
- **MST** = Mean Square Total = $SST/(n-1) = \boldsymbol{s_y^2}$
  - Usually not found on software output

Stockholm University

See formula sheet

Coefficient of determination R^2

- Proportion of the total variation of $\boldsymbol{Y}$ that is explained by the model, i.e. the variation explained by differences in $\boldsymbol{x}$:

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

Sometimes reported as %

- **MST** = Mean Square Total = $SST/(n-1) = \boldsymbol{s_y^2}$
- **MSE** = Mean Square Error = $SSE/(n-2) = \boldsymbol{s_e^2}$

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

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

$\boldsymbol{\bar{R}^2 < R^2}$

Stockholm University

Exercise, cont. ANOVA and R^2 in Excel

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0,737902433					
R Square	0,5445					
Adjusted R Square	0,430625					
Standard Error	1,948289848					
Observations	6					
ANOVA						
		df	SS	MS		
Regression		1	18,15	18,15		
Residual		4	15,18333333	3,795833		
Total		5	33,33333333			
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	2,916666667	1,488030951	1,960085	0,121542838	-1,214769584	7,048102917
X Variable 1	0,55	0,251523138	2,186678	0,094040288	-0,148340185	1,248340185

$R^2 = 54,45\ \%$

$R^2_{adj} = 43,06\ \%$

df = degrees of freedom:

Total = Error + Regr

Regression = SSR

Residual = SSE

Total = SST

Stockholm University

Coefficient of determination R^2, cont.

- $\boldsymbol{R^2}$ is a measure of how well $\boldsymbol{x}$ "explains" $\boldsymbol{y}$
  - i.e. how the variation in $\boldsymbol{y}$ can be explained by the variation of $\boldsymbol{x}$
- Possible values: $\mathbf{0\ \%} \leq \boldsymbol{R^2} \leq \mathbf{100\ \%}$
  - NOTE! Adjusted value $\bar{\boldsymbol{R}}^2 = \boldsymbol{R^2_{adj}}$ can take values outside this intervall!
- $R^2 = 0\ \%$ means that the model is pointless $\boldsymbol{\sigma_\varepsilon^2 = \sigma_Y^2}$
- $R^2 = 100\ \%$ means that all y values are on the line $\boldsymbol{\sigma_\varepsilon^2 = 0}$
- $50\ \% \leq R^2 \leq 80\ \%$ is often regarded as great, above $80\ \%$ as even better. It all depends on the context
- You should also consider $\boldsymbol{s_e^2 = MSE = SSE/(n-2)}$ as a measure of the fit when you different data (see p. 435 NCT)

Stockholm University

Linear regression and correlation

- The relationship between the linear regression model and correlation coefficient is apparent in the formulas:

$$b_1 = \frac{Cov(x,y)}{s_x^2} = r_{xy} \cdot \frac{s_y}{s_x} \qquad r_{xy} = \frac{Cov(x,y)}{s_x \cdot s_y} = b_1 \cdot \frac{s_x}{s_y}$$

- The relationship between correlation coefficient and R^2:

$$R^2 = \left(r_{xy}\right)^2$$

- The sign (plus or minus) of the correlation is given by the sign of the regression coefficient b_1. Should be same sign!

Stockholm University

Properties of b_1 and b_0

- The estimates $\boldsymbol{b_0}$ and $\boldsymbol{b_1}$ are linear combinations of iid normally distributed random variables. It can be shown that:

$$\boldsymbol{b_1 \sim N(\beta_1, \sigma^2_{b_1})} \qquad \boldsymbol{b_0 \sim N(\beta_0, \sigma^2_{b_0})}$$

- So the esimates $\boldsymbol{b_0}$ and $\boldsymbol{b_1}$ are ***unbiased***!

- Variances $\boldsymbol{\sigma^2_{b_0}}$ and $\boldsymbol{\sigma^2_{b_1}}$ are unknown and must be estimated:

$$\boldsymbol{s^2_{b_1} = \frac{s^2_e}{(n-1)s^2_x}} \qquad s^2_{b_0} = s^2_e\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s^2_x}\right)$$

- We can already guess that we will use the *t*-distribution. Why?

Stockholm University

Inference – to test the model

W typically want to know if $\boldsymbol{\beta_1 = 0}$. Why?

- If $\boldsymbol{\beta_1 = 0}$ the model can be simplified to

$$\boldsymbol{Y = \beta_0 + 0 \cdot x + \varepsilon = \beta_0 + \varepsilon} \qquad \text{where} \quad \boldsymbol{\varepsilon \sim N(0, \sigma_\varepsilon^2)}$$

  Which is the same as saying that $\boldsymbol{\beta_0 = \mu_Y}$ i.e. $\boldsymbol{Y = \mu_Y + \varepsilon}$

If $\boldsymbol{\beta_1 = 0}$ it follows that **(se F14 s. 20)**

- The slope of the line is zero; line is parallel to the x-axis
- The correlation and coefficient of determination: $\boldsymbol{r_{xy} = R^2 = 0}$
- We gain nothing by conditioning on $\boldsymbol{X = x}$

Stockholm University

Hypothesis test for β_1

- Typical hypoteses to test a simpple linear model:

$$H_0: \beta_1 = 0 \quad \text{vs.} \qquad H_1: \beta_1 \neq 0$$

- Assumptions:
  - Normally distributed iid random errors ε, independent of the x-values
- Test variable: $t_{n-2} = \dfrac{b_1 - 0}{s_{b_1}}$ that is t-distributed with $n-2$ d.g.

- Critical values and decision rule:
  - reject H_0 if $|t_{obs}| > t_{n-2;\alpha/2}$

- You can test for values different from 0, in the formulary it says β_1^* which denotes any choice of value. You can use any value in its place.

Stockholm University

Confidence interval for β_0 and β_1

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

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

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

- Needs the same assumptions as on last page.

- NOTE! If we want to test $\boldsymbol{H_0: \beta_1 = \beta_j^*}$ vs. $\boldsymbol{H_1: \beta_1 \neq \beta_j^*}$ where $\boldsymbol{\beta_j^*}$ is some number, it is enough to check whether $\boldsymbol{\beta_j^*}$ lies within the confidence interval.
  - E.g. for hypotesis $\boldsymbol{H_0: \beta_1 = 0}$, check whether the value 0 is in the interval

Stockholm University

Hypotesis testing for the intercept β_0

- Not as common.
- If you were to test $H_0: \beta_0 = 0$ mot $H_1: \beta_0 \neq 0$ and a non-significant result, i.e. that b_0 cannot be assumed to be different from 0
- Maybe make $Y = \beta_1 x + \varepsilon$; estimate a new model with no intercept?
- The problem is that β_1 no longer unbiased!
- Solution 1: let β_0 stay in the model and accept that it is not significantly different from zero
- Solution 2: us a more complicated model

Stockholm University

Inference for conditional expectation $\mu_{Y|x_{n+1}}$

- Remember that the expected value of Y depends on the value of x; we condition on the event $X = x$ using $\hat{\mu}_{Y|x_{n+1}} = b_0 + b_1 x_{n+1}$
  - NCT dentoed this new x-value by x_{n+1}
- A $100(1-a)\,\%$ CI for $\mu_{Y|x_{n+1}}$ is given by

$$\hat{\mu}_{Y|x_{n+1}} \pm t_{n-2,\alpha/2} \cdot s_e \sqrt{\left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{(n-1)s_x^2}\right)}$$

Looks a little differnt in the formulary but it is the same formula.

- Compare to the regular CI for μ_Y:

$$\bar{y} \pm t_{n-1,\alpha/2} \cdot s_y \sqrt{\frac{1}{n}}$$

We gain accuracy if $s_e < s_y$ and if x_{n+1} is not too far away from $\bar{x}$.

Stockholm University

Predicting future Y without regression

- Create a confidence interval around the expected value of μ_Y:

$$\mu_Y \pm 1,96 \cdot \sqrt{\sigma_Y^2}$$ 95 % probability to get a value in the interval

- If μ_Y is unknown, we can estimate the value with $\bar{Y} \sim N(\mu_Y, \sigma_Y^2/n)$
  - Carries extra uncertainty:

$$\bar{y} \pm 1,96 \cdot \sqrt{\sigma_Y^2 + \frac{\sigma_Y^2}{n}}$$ extra uncertainty because of the estimate of $\bar{y}$

- If σ_Y^2 also is unknown, it is estimated with s_y^2 and we use t

$$\bar{y} \pm t_{n-1;0,025} \sqrt{s_y^2 + \frac{s_y^2}{n}}$$ extra uncertainty b/c of the estimate s_y^2

Stockholm University

Predicting future Y without regression

- Corresponding ***prediction interval*** for y_{n+1} given x_{n+1}

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} \cdot \sqrt{s_e^2\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{(n-1)s_x^2}\right)}$$

t-distribution since we estimate σ_ε^2 with s_e^2

More uncertain the further away from $\bar{x}$ we are

Since we estimate $\mu_{Y|x_{n+1}}$

- REMEMBER: $\hat{y}_{n+1} = b_0 + b_1 x_{n+1}$

Natural variation in Y give $X = x_{n+1}$

Stockholm University

Exercise 1, same as last time

- Suppose we have a sample $n = 6$ paired values from random variables X and Y:

i	x_i	y_i
1	1	2
2	2	3
3	10	8
4	8	6
5	5	7
6	4	8
Summa	30	34

$\bar{x} = 5$ $\quad \bar{y} = 5,6667$

$s_x^2 = 12$ $\quad s_y^2 = 6,6667$

$s_x = 3,4610$ $\quad s_y = 2,5820$

$s_{xy} = 6,6$ $\quad r_{xy} = 0,73790$

$b_1 = 0,55$ $\quad b_0 = 2,91667$

$s_e^2 = 3,7958$ $\quad s_e = 1,9483$

Stockholm University

Exercise

1. Test $\boldsymbol{H_0}: \boldsymbol{\beta_1} = \mathbf{0}$ vs. $\boldsymbol{H_1}: \boldsymbol{\beta_1} \neq \mathbf{0}$ at the 5 % significance level
   - Verify the results against the output on the next slide
2. Calculate a 95 % CI for $\boldsymbol{\beta_1}$
   - Verify the result against the output
3. Calculate a 95 % CI for $\boldsymbol{\beta_0}$ using only the output
4. Calculate $\boldsymbol{R^2}$ and $\boldsymbol{\bar{R}^2}$ (adjusted $\boldsymbol{R^2}$)
   - verify the result against the output
5. Calculate a 95 % KI for $\boldsymbol{\mu_{Y|X=6}}$
6. Calculate a 95 % PI for $\boldsymbol{y_{n+1}}$ when $\boldsymbol{x_{n+1}} = \mathbf{6}$

Stockholm University

Exercise using Excel – interpret output

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0,737902433					
R Square	0,5445					
Adjusted R Square	0,430625					
Standard Error	1,948289848					
Observations	6					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	18,15	18,15	4,781558727	0,094040288	
Residual	4	15,18333333	3,795833			
Total	5	33,33333333				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	2,916666667	1,488030951	1,960085	0,121542838	-1,214769584	7,048102917
X Variable 1	0,55	0,251523138	2,186678	0,094040288	-0,148340185	1,248340185

Stockholm University

Next time

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

Stockholm University