Simple linear regression

Tron Anders Moger
3.10.2007

Example 6: Population proportions One sample

- Assume $X \sim Bin(n,P)$, so that $\hat{P} = \frac{X}{n}$ is a frequency.
- Then $\frac{\hat{P}-P}{\sqrt{P(1-P)/n}} \sim N(0,1)$ (approximately, for large n)
- Thus $\frac{\hat{P}-P}{\sqrt{\hat{P}(1-\hat{P})/n}} \sim N(0,1)$ (approximately, for large n)
- Thus $P\left(\hat{P}-Z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} < P < \hat{P}+Z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right) = \alpha$
- Confidence interval for P

$$\left(\hat{P}-Z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}, \hat{P}+Z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right)$$

Example 6 (Hypothesis testing)

- Hypotheses: H_0:$P=P_0$ H_1:$P \neq P_0$
- Test statistic $\frac{\hat{P}-P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} \sim N(0,1)$

under H_0, for large n

- Reject H_0 if $\frac{\hat{P}-P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} < -Z_{\alpha/2}$,or if $\frac{\hat{P}-P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} > Z_{\alpha/2}$

Example 7: Differences between population proportions-two samples

- Assume $X_1 \sim Bin(n_1, P_1)$ and $X_2 \sim Bin(n_2, P_2)$ so that $\hat{P}_1 = \frac{X_1}{n_1}$ and $\hat{P}_2 = \frac{X_2}{n_2}$ are frequencies
- Then $\frac{\hat{P}_1-\hat{P}_2-(P_1-P_2)}{\sqrt{\frac{P_1(1-P_1)}{n_1}+\frac{P_2(1-P_2)}{n_2}}} \sim N(0,1)$ (approximately)
- Confidence interval for P_1-P_2

$$\left(\hat{P}_1-\hat{P}_2 \pm Z_{\alpha/2}\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1}+\frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}\right)$$

Example 7 (Hypothesis testing)

- Hypotheses: $H_0: P_1 = P_2$ $H_1: P_1 \neq P_2$
- Test statistic $\dfrac{\hat{P}_1 - \hat{P}_2}{\sqrt{\dfrac{\hat{P}_0(1-\hat{P}_0)}{n_1} + \dfrac{\hat{P}_0(1-\hat{P}_0)}{n_2}}} \sim N(0,1)$

  where $\hat{P}_0 = \dfrac{n_1 \hat{P}_1 + n_2 \hat{P}_2}{n_1 + n_2}$
- Reject H_0 if $\left| \dfrac{\hat{P}_1 - \hat{P}_2}{\sqrt{\dfrac{\hat{P}_0(1-\hat{P}_0)}{n_1} + \dfrac{\hat{P}_0(1-\hat{P}_0)}{n_2}}} \right| > Z_{\alpha/2}$

- Spontanous abortions among surgical nurses and other nurses
- Want to test if there is difference between the proportions of abortions in the two groups
- H_0: $P_{op.nurses} = P_{others}$ H_1: $P_{op.nurses} \neq P_{others}$

	Surgical nurses	Other nurses
No. interviewed	67	92
No. pregnancies	36	34
No. abortions	10	3
Percent abortions	27.8	8.8

Calculation:

- $P_1=0.278$ $P_2=0.088$ $n_1=36$ $n_2=34$

$$\bar{p} = \frac{\text{Total no. abortions}}{\text{Total no. pregnancies}} = \frac{10+3}{36+34} = 0.186$$

$$z = \frac{0.278 - 0.088}{\sqrt{(\frac{1}{36} + \frac{1}{34})0.186(1-0.186)}} = 2.04$$

- P-value 0.0414=4.1%, reject H_0 on 5%-sig.level (can't do this in SPSS)
- 95% confidence interval for P_1-P_2:

$$(\hat{P}_1 - \hat{P}_2) \pm 1.96 * \sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}} = (0.015, 0.190)$$

Repetition:

- Testing:
  - Identify data; continuous->t-tests; proportions->Normal approx. to binomial dist.
  - If continous: one-sample, matched pairs, two independent samples?
  - Assumptions: Are data normally distributed? If two ind. samples, equal variances in both groups?
  - Formulate H_0 and H_1 (H_0 is always no difference, no effect of treatment etc.), choose sig. level (α=5%)
  - Calculate test statistic

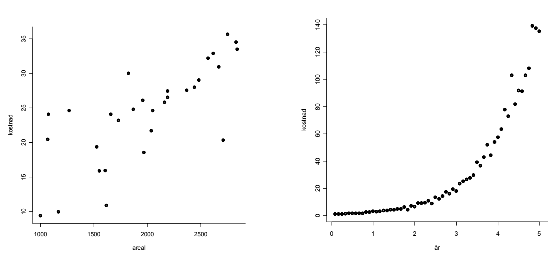
Inference:

- Test statistic usually standardized; (estimator-expected value of estimator under H_0)/(estimated standard error)
- Gives you a location on the x-axis in a distribution
- Compare this value to the value at the 2.5%-percentile and 97.5%-percentile of the distribution
- If smaller than the 2.5%-percentile or larger than the 97.5%-percentile, reject H_0
- P-value: Area in the tails of the distribution below value of test statistic+area above value of test-statistic (two-sided testing)
- If smaller than 0.05, reject H_0
- If confidence interval for mean or mean difference (depends on test what you use) does not include H_0 value from, reject H_0

Last week:

- Looked at continuous, normally distributed variables
- Used t-tests to see if there was significant difference between means in two groups
- How strong is the relationship between two such variables? Correlation
- What if one wants to study the relationship between several such variables? Linear regression

Connection between variables

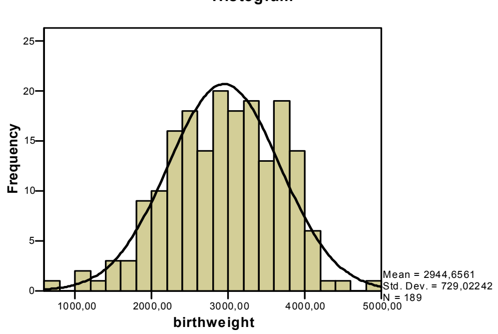


We would like to study connection between x and y!

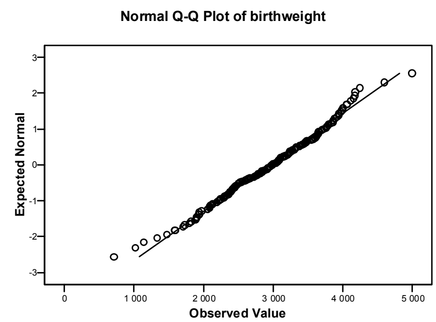
Data from the first obligatory assignment:

- Birth weight and smoking
- Children of 189 women
- Low birth weight is a medical risk factor
- Does mother's smoking status have any influence on the birth weight?
- Also interested in relationship with other variables: Mother's age, mother's weight, high blood pressure, ethincity etc.

Is birth weight normally distributed?



Q-Q plot (check *Normality plots with tests* under *plots*):



Tests for normality:

Tests of Normality

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
birthweight	,043	189	,200(*)	,992	189	,438

* This is a lower bound of the true significance.
a Liljefors Significance Correction

The null hypothesis is that the data are normal. Large p-value indicates normal distribution. For large samples, the p-value tends to be low. The graphical methods are more important

Pearsons correlation coefficient r

- Measures the linear relationship between variables
- r=1: All data lie on an increasing straight line
- r=-1: All data lie on a decreasing straight line
- r=0: No linear relationship
- In linear regression, often use R^2 (r^2) as a meansure of the explanatory power of the model
- R^2 close to 1 means that the observations are close to the line, r^2 close to 0 means that there is no linear relationship between the observations

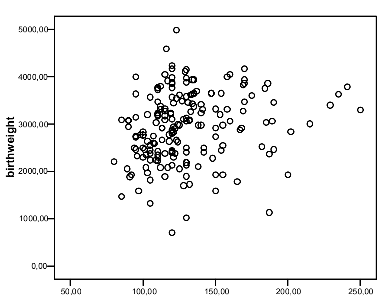
Testing for correlation

- It is also possible to test whether a sample correlation *r* is large enough to indicate a nonzero population correlation
- Test statistic: $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$
- Note: The test only works for normal distributions and linear correlations: Always also investigate scatter plot!

Pearsons correlation coefficient in SPSS:

- Analyze->Correlate->bivariate
  Check Pearson
- Tests if r is significantly different from 0
- Null hypothesis is that r=0
- The variables have to be normally distributed
- Independence between observations

Example:



Correlation from SPSS:

Correlations

		birthweight	weight in pounds
birthweight	Pearson Correlation	1	,186*
	Sig. (2-tailed)		,010
	N	189	189
weight in pounds	Pearson Correlation	,186*	1
	Sig. (2-tailed)	,010	
	N	189	189

*. Correlation is significant at the 0.05 level (2-tailed).

If the data are not normally distributed: Spearmans rank correlation, r_s

- Measures all monotonous relationships, not only linear ones
- No distribution assumptions
- r_s is between -1 and 1, similar to Pearsons correlation coefficient
- In SPSS: Analyze->Correlate->bivariate
  Check Spearman
- Also provides a test on whether r_s is different from 0

Spearman correlation:

Correlations

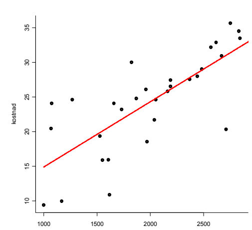
			birthweight	weight in pounds
Spearman's rho	birthweight	Correlation Coefficient	1,000	,248**
		Sig. (2-tailed)	.	,001
		N	189	189
	weight in pounds	Correlation Coefficient	,248**	1,000
		Sig. (2-tailed)	,001	.
		N	189	189

**. Correlation is significant at the 0.01 level (2-tailed).

Linear regression

- Wish to fit a line as close to the observed data (two normally distributed varaibles) as possible
- Example: Birth weight=*a*+*b**mother's weight
- In SPSS: Analyze->Regression->Linear
- Click Statistics and check Confidence interval for B
- Choose one variable as dependent (Birth weight) as dependent, and one variable (mother's weight) as independent
- Important to know which variable is your dependent variable!

Connection between variables



Fit a line!

The standard simple regression model

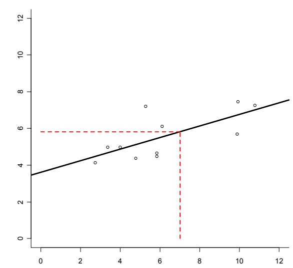
- We define a *model*

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

  where ε_i are independent, normally distributed, with equal variance σ^2
- We can then use data to *estimate* the model *parameters*, and to make statements about their uncertainty

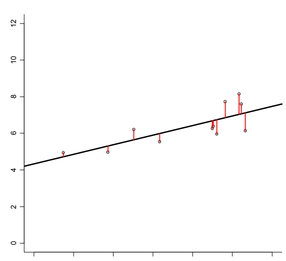
What can you do with a fitted line?

- Interpolation
- Extrapolation (sometimes dangerous!)
- Interpret the parameters of the line



How to define the line that ”fits best”?

$$\boxed{\text{The sum of the squares of the ”errors” minimized} = \text{Least squares method!}}$$



- Note: Many other ways to fit the line can be imagined

How to compute the line fit with the least squares method?

- Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ denote the points in the plane.
- Find a and b so that y=a+bx fit the points by minimizing

$$S = (a+bx_1-y_1)^2 + (a+bx_2-y_2)^2 + \cdots + (a+bx_n-y_n)^2 = \sum_{i=1}^{n}(a+bx_i-y_i)^2$$

- Solution:

$$b = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}_i^2}$$

$$a = \frac{\sum y_i - b\sum x_i}{n} = \bar{y} - b\bar{x}$$

where $\bar{x} = \frac{1}{n}\sum x_i, \bar{y} = \frac{1}{n}\sum y_i$ and all sums are done for i=1,...,n.

How do you get this answer?

- Differentiate S with respect to a og b, and set the result to 0

$$\frac{\partial S}{\partial a} = \sum_{i=1}^{n} 2(a + bx_i - y_i) = 0$$
$$\frac{\partial S}{\partial b} = \sum_{i=1}^{n} 2(a + bx_i - y_i)x_i = 0$$

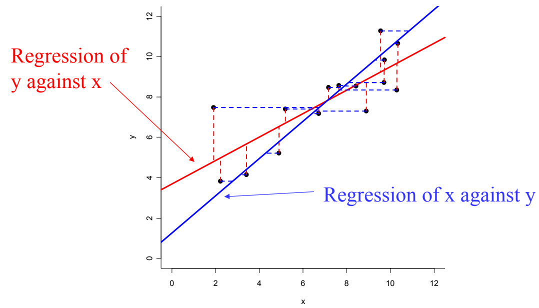
We get:

$$a \cdot n + b\left(\sum x_i\right) - \sum y_i = 0$$
$$a\left(\sum x_i\right) + b\left(\sum x_i^2\right) - \sum x_i y_i = 0$$

This is two equations with two unknowns, and the solution of these give the answer.

y against x ≠ x against y

- Linear regression of y against x does not give the same result as the opposite.

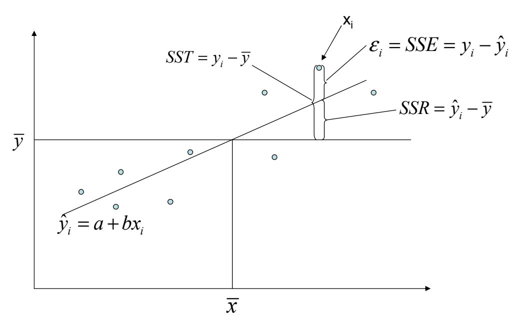


Anaylzing the variance

- Define
  - SSE: Error sum of squares $\sum (a + bx_i - y_i)^2$
  - SSR: Regression sum of squares $\sum (a + bx_i - \bar{y})^2$
  - SST: Total sum of squares $\sum (y_i - \bar{y})^2$
- We can show that

  SST = SSR + SSE
- Define $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = corr(x, y)^2$
- R^2 is the "coefficient of determination"

What is the logic behind R^2?



Assumptions

- Usually check that the dependent variable is normally distributed
- More formally, the residuals, i.e. the distance from each observation to the line, should be normally distributed
- In SPSS:
  - In linear regression, click Statistics. Under residuals check casewise diagnostics, and you will get ”outliers” larger than 3 or less than -3 in a separate table.
  - In linear regression, also click Plots. Under standardized residuals plots, check Histogram and Normal probability plot. Choose *Zresid as y-variable and *Zpred as x-variable

Example: Regression of birth weight with mother’s weight as independent variable

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,186[a]	,035	,029	718,24270

a. Predictors: (Constant), weight in pounds

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3448881	1	3448881,301	6,686	,010[a]
	Residual	96468171	187	515872,574		
	Total	99917053	188			

a. Predictors: (Constant), weight in pounds

b. Dependent Variable: birthweight

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	2369,672	228,431		10,374	,000	1919,040	2820,304
	weight in pounds	4,429	1,713	,186	2,586	,010	1,050	7,809

a. Dependent Variable: birthweight

Residuals:

Casewise Diagnostics[a]

Case Number	Std. Residual	birthweight	Predicted Value	Residual
1	-3,052	709,00	2901,1837	-2192,18

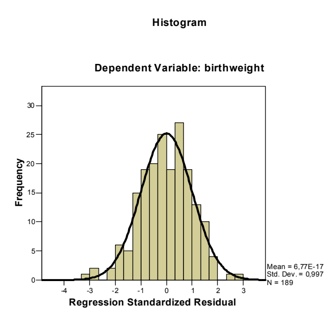
a. Dependent Variable: birthweight

Residuals Statistics[a]

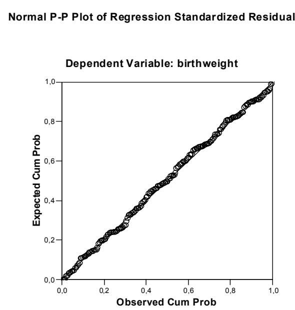
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	2724,0132	3476,9880	2944,6561	135,44413	189
Residual	-2192,18	2075,529	,00000	716,32993	189
Std. Predicted Value	-1,629	3,930	,000	1,000	189
Std. Residual	-3,052	2,890	,000	,997	189

a. Dependent Variable: birthweight

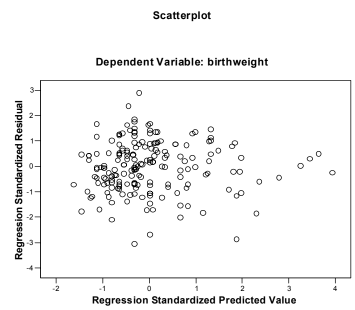
Check of assumptions:



Check of assumptions cont'd:



Check of assumptions cont'd:



Interpretation:

- Have fitted the line

  Birth weight=2369.672+4.429*mother's weight
- If mother's weight increases by 20 pounds, what is the predicted impact on infant's birth weight?

  4.429*20=89 grams
- What's the predicted birth weight of an infant with a 150 pound mother?

  2369.672+4.429*150=3034 grams

Influence of extreme observations

- NOTE: The result of a regression analysis is very much influenced by points with extreme values, in either the x or the y direction.
- Always investigate visually, and determine if outliers are actually erroneous observations

But how to answer questions like:

- Given that a positive slope (b) has been estimated: Does it give a reproducible indication that there is a positive trend, or is it a result of random variation?
- What is a confidence interval for the estimated slope?
- What is the prediction, with uncertainty, at a new x value?

Confidence intervals for simple regression

- In a simple regression model,
  - a estimates β_0
  - b estimates β_1
  - $\hat{\sigma}^2 = SSE/(n-2)$ estimates σ^2
- Also, $(b-\beta_1)/S_b \sim t_{n-2}$ where $S_b^2 = \dfrac{\hat{\sigma}^2}{(n-1)s_x^2}$ estimates variance of b
- So a confidence interval for β_1 is given by $b \pm t_{n-2,\alpha/2} S_b$

Hypothesis testing for simple regression

- Choose hypotheses: $H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$
- Test statistic: $b/S_b \sim t_{n-2}$
- Reject H_0 if $b/S_b < -t_{n-2,\alpha/2}$ or $b/S_b > t_{n-2,\alpha/2}$