

Regression

Simple Linear regression

Morten Frydenberg ©
Institut for Biostatistik

Regression in general

Simple linear regression.

The model.

The assumptions.

The parameters.

Estimation.

The distribution of the estimates

Confidence intervals

Changing the reference value and scale for x

Tests

The example: Summarising

Morten FrydenbergLinear and Logistic regression - Note 1.11

Regression in general

A regression model can be many things!

In general it **models** the relationship between:

y : **dependent**/response

and a set of

x 's: **independent**/explanatory variables.

The dependent variable is **modelled** as a function of the independent variable plus some unexplained random variation:

Systematic part

$y = f(x; \theta)$

Unknown Parameters

Random part

$+ e(\sigma)$

Unknown Parameters

Morten FrydenbergLinear and Logistic regression - Note 1.12

Regression in general

$y = f(x; \theta) + e(\sigma)$

Some examples:

$pefr = \beta_0 + \beta_1 \cdot height + E$

$pefr = \beta_0 + \beta_1 \cdot height + \beta_2 \cdot height^2 + E$

$gfr = \exp(\beta_0 + \beta_1 \cdot \ln[Cr]) + E$

$conc(t) = dose \cdot V \cdot [\exp(-\lambda_{abs} \cdot t) - \exp(-\lambda_{eli} \cdot t)] + E$

The first two are **linear** regressions, the last two **non-linear**.

In this course we will **focus** on the **linear** regressions.

Morten FrydenbergLinear and Logistic regression - Note 1.13

Simple linear regression

The relationship between measured $PEFR$ and $height$ in 101 medical students.

A model : $PEFR = \text{line} + \text{some random variation}$ seems to be valid.

Morten FrydenbergLinear and Logistic regression - Note 1.14

Simple linear regression: The model

Let $PEFR_i$ and $height_i$ be the data for the i th person.

$PEFR_i = \beta_0 + \beta_1 \cdot height_i + E_i \quad E_i \sim N(0, \sigma^2)$

This model is based on the **assumptions**:

1. The **expected** value of $PEFR$ is a **linear function** of $height$.

2. The **unexplained** random deviations are **independent**.

3. The unexplained random deviations have the **same distributions**.

4. This distribution is **normal**.

Morten FrydenbergLinear and Logistic regression - Note 1.15

Simple linear regression: The parameters

$PEFR_i = \beta_0 + \beta_1 \cdot height_i + E_i \quad E_i \sim N(0, \sigma^2)$

The model have **three** unknown **parameters**:

1. The **intercept** β_0

2. The **slope** (or **regression coefficient**) β_1

3. The **residual variance** σ^2 or **residual standard deviation** σ .

The **interpretation** of the parameters:

β_0 is expected $PEFR$ of a person with $height=0$.

Obviously, this does not make sense.

We will later look at how one can get a meaning full estimate of the general level of $PEFR$!

Morten FrydenbergLinear and Logistic regression - Note 1.16

Simple linear regression: The parameters

$$PEFR_i = \beta_0 + \beta_1 \cdot height_i + E_i \quad E_i \sim N(0, \sigma^2)$$

β_1 is the **expected difference** in **PEFR** for two persons who differ with **one unit** (here cm) in **height**.

If a person is **6 cm** higher than another, then we will expected that his **PEFR** is **$6\beta_1$** higher than the other.

σ is best understood by the fact that a **95%-prediction interval** around the line is given by **$\pm 1.96\sigma$** .

Morten Frydenberg Linear and Logistic regression - Note 1.1 7

Simple linear regression: The estimates (by hand)

$$PEFR_i = \beta_0 + \beta_1 \cdot height_i + E_i \quad E_i \sim N(0, \sigma^2)$$

The estimates of the parameters are found by the method of **least square**, which, for this model, is equivalent to the **maximum likelihood** method.

The estimates can be calculated in hand, but they are of course found much easier by using a computer program.

$$\hat{\beta}_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}$$
$$\hat{\sigma}^2 = \frac{1}{n-2} \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_i)^2 = \frac{1}{n-2} \sum r_i^2$$

Morten Frydenberg Linear and Logistic regression - Note 1.1 8

Simple linear regression: The estimates (by computer)

In STATA we fit the model by the command

`regress PEFR height`

n: Always check this

Source	SS	df	MS	Number of obs = 101			
Model	226303.854	1	226303.854	F(1, 99)	= 69.90		
Residual	320519.473	99	3237.57044	Prob > F	= 0.0000		
Total	546823.327	100	5468.23327	R-squared	= 0.4139		
				Adj R-squared	= 0.4079		
				Root MSE	= 56.9		

PEFR	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
height	5.711578	.6831558	8.36	0.000	4.356049 7.067107
_cons	-456.9205	117.9567	-3.87	0.000	-690.9721 -222.869

$\hat{\beta}_1$ $\hat{\beta}_0$ σ^2 $\hat{\sigma}$

Standard errors 95% confidence intervals

Morten Frydenberg Linear and Logistic regression - Note 1.1 9

Simple linear regression: The distribution of the estimates

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2 \frac{1}{\sum (x_i - \bar{x})^2}\right) \quad se(\hat{\beta}_1) = \hat{\sigma} / \sqrt{\sum (x_i - \bar{x})^2}$$
$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right]\right) \quad se(\hat{\beta}_0) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$
$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n-2} \chi^2(n-2)$$

Some comments:

The precision of the estimates of β_1 and β_0 depends on the size of the variation around the line.

The precision of the estimate of β_1 increases with the variation of x 's

Morten Frydenberg Linear and Logistic regression - Note 1.1 10

Simple linear regression: Confidence intervals

Exact 95% confidence intervals, CI's, for β_0 and β_1 are found from the estimates and standard errors

95% CI for β_1 : $\hat{\beta}_1 \pm t_{n-2}^{0.975} \cdot se(\hat{\beta}_1)$

95% CI for β_0 : $\hat{\beta}_0 \pm t_{n-2}^{0.975} \cdot se(\hat{\beta}_0)$

Where $t_{n-2}^{0.975}$ is the upper 97.5 percentile in the t-distribution $n-2$ degrees of freedom.

These confidence intervals are found in the output.

Note that if n is large then this percentile is close to **1.96** and one can use the **approximate confidence intervals**:

Approx. 95% CI for β_1 : $\hat{\beta}_1 \pm 1.96 \cdot se(\hat{\beta}_1)$

Approx. 95% CI for β_0 : $\hat{\beta}_0 \pm 1.96 \cdot se(\hat{\beta}_0)$

Morten Frydenberg Linear and Logistic regression - Note 1.1 11

Simple linear regression: Confidence intervals

Exact 95% confidence intervals, CI's, for σ using the χ^2 distribution with $n-2$ degrees of freedom.

95% CI for σ : $\hat{\sigma} \cdot \sqrt{\frac{n-2}{\chi_{n-2}^2(0.975)}} \leq \sigma \leq \hat{\sigma} \cdot \sqrt{\frac{n-2}{\chi_{n-2}^2(0.025)}}$

Where $\chi_{n-2}^2(0.975)$ is the **upper** 97.5 percentile and $\chi_{n-2}^2(0.025)$ the **lower** 2.5 percentile in the χ^2 -distribution $n-2$ degrees of freedom.

This confidence interval is **rarely** given in the output !

Using STATA we find:

```
display 56.9*sqrt(99/invchi2(99,0.975))
49.95859
display 56.9*sqrt(99/invchi2(99,0.025))
66.099322
```

Morten Frydenberg Linear and Logistic regression - Note 1.1 12

Changing the reference value and scale for x

$$PEFR_i = \beta_0 + \beta_1 \cdot height_i + E_i \quad E_i \sim N(0, \sigma^2)$$

In this model the parameter β_0 does not make sense.

But if we consider the equivalent model:

$$PEFR_i = \alpha_0 + \alpha_1 \cdot (height_i - 170cm) + E_i \quad E_i \sim N(0, \tau^2)$$

then α_0 is the expected $PEFR$ of a person with height 170cm.

The two other parameters are unchanged, i.e. $\beta_1 = \alpha_1$ and $\sigma = \tau$

If $HEIGHT$ denote the height in m, i.e. $HEIGHT = height/100$ and we consider the equivalent model:

$$PEFR_i = \gamma_0 + \gamma_1 \cdot HEIGHT_i + E_i \quad E_i \sim N(0, \omega^2)$$

then $\gamma_1 = 100 \cdot \beta_1$, $\gamma_0 = \beta_0$ and $\omega = \sigma$

Morten Frydenberg Linear and Logistic regression - Note 1.1 13

Simple linear regression: The intercept

Let us fit the model with a meaningful intercept/constant:

generate height170=height-170

regress PEFR height170

Source	SS	df	MS	Number of obs =	101
Model	226303.854	1	226303.854	F(1, 99) =	69.90
Residual	320519.473	99	3237.57044	Prob > F =	0.0000
Total	546823.327	100	5468.23327	R-squared =	0.4139
				Adj R-squared =	0.4079
				Root MSE =	56.9

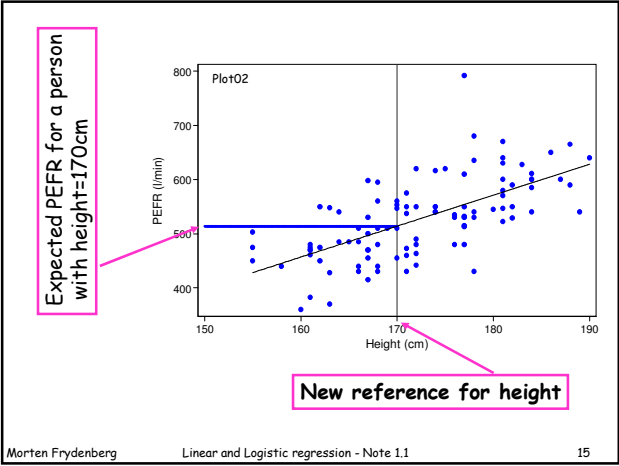
PEFR	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
height170	5.7115	.6831558	8.36	0.000	4.356 7.0671
_cons	514.0477	5.906923	87.02	0.000	502.32 525.76

Nothing is changed except this

The expected $PEFR$ for a person with height=170cm is:

514 (502;526) l/min

Morten Frydenberg Linear and Logistic regression - Note 1.1 14



Confidence interval for the estimated line

The true line is given as : $y = \beta_0 + \beta_1 \cdot x$

and estimated by plugging in the estimates $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$

The standard error of this estimate is given by:

$$se(\hat{\beta}_0 + \hat{\beta}_1 \cdot x) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

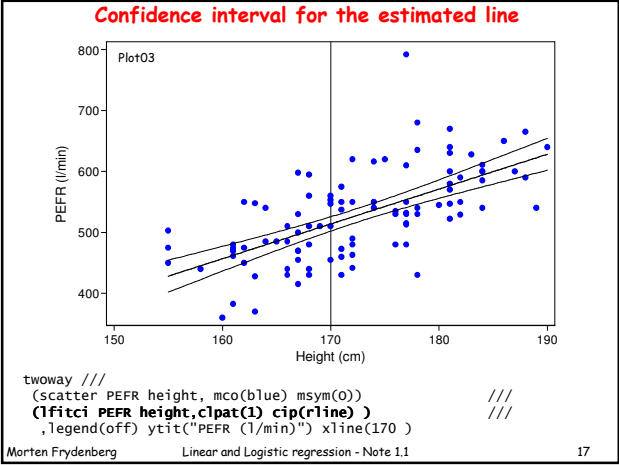
with the 95% (pointwise) confidence interval

$$\hat{\beta}_0 + \hat{\beta}_1 \cdot x \pm t_{n-2}^{0.975} \cdot se(\hat{\beta}_0 + \hat{\beta}_1 \cdot x)$$

Many programs can make a plot with the fitted line and its confidence limits.

In STATA its done by the *lfitci* graph command.

Morten Frydenberg Linear and Logistic regression - Note 1.1 16



Simple linear regression: Tests

Statistical test concerning β_0 and β_1 can be calculated in the standard way based on estimates, standard errors and the t-distribution:

Hypothesis: $\beta_1 = \beta_{1H}$

Test statistics: $z = \frac{\hat{\beta}_1 - \beta_{1H}}{se(\hat{\beta}_1)}$ P-value: $2 \cdot P(t_{n-2} < -|z|)$

An example: Hypothesis $\beta_1 = 5$

$$z = \frac{5.771 - 5}{0.6832} = 0.7116$$
 P-value 30%

In STATA : *lfitcom* height170-5

Morten Frydenberg Linear and Logistic regression - Note 1.1 18

Simple linear regression: Tests/confidence intervals

The p-values found in the **regression output** corresponds to the hypothesis that the given parameter is **zero**, e.g. $\beta_1 = 0$.

In the example we find that β_1 is highly significant ($p < 0.001$) different from 0.

That is, there is a **statistical significant association** between *PEFR* and *Height*.

The estimate with **confidence interval** does of course contain much more information than the p-value:

95% CI for β_1 : 5.71 (4.36;7.07) l/min/cm

From this we can see that the difference in **mean PEFR** between two persons, who differ one cm in height, is in interval from **4.36** to **7.07** l/min.

Morten FrydenbergLinear and Logistic regression - Note 1.119

The example: Summarising

$$PEFR_i = \beta_0 + \beta_1 \cdot (height_i - 170) + E_i \quad E_i \sim N(0, \sigma^2)$$

The estimates:

β_1 : **5.71 (4.36;7.07) l/min/cm**

β_0 : **514 (502;526) l/min**

σ : **56.9 (50.0;66.1) l/min**

The difference in **mean PEFR** between two persons who **differ one cm** in height is in interval from **4.36** to **7.07** l/min - the best guess is **5.71** l/min.

The mean PEFR for a person who is 170 cm is in the interval **502** to **526** l/min - the best guess is **514** l/min.

A 95% prediction interval is given as **±112** l/min.

Morten FrydenbergLinear and Logistic regression - Note 1.120