Regression
Simple Linear regression

Morten Frydenberg ©
Department of Biostatisics, Aarhus Univ, Denmark

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

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 — Random part

$$y = f(x;\theta) \text{ "+" } e(\sigma)$$

Unknown Parameters — Unknown Parameters

Regression in general

$$y = f(x;\theta) \text{ "+" } e(\sigma)$$

Some examples:

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

$pefr = \beta_0 + \beta_1 \cdot height + \beta_2 \cdot height^2 + E$ and $E \sim N(0,\sigma^2)$

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

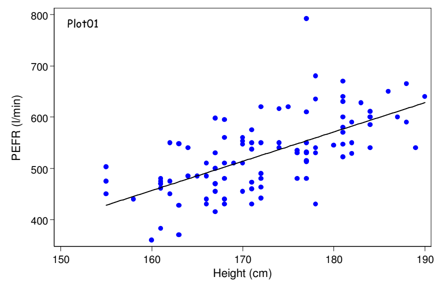
$conc(t) = dose \cdot V \cdot [\exp(-\lambda_{abs} \cdot t) - \exp(-\lambda_{eli} \cdot t)] + E$

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

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

Simple linear regression

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



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

Simple linear regression: The model

Let $PEFR_i$ and $height_i$ be the data for the ith 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**.

Simple linear regression: The parameters

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

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 meaningful estimate of the general level of $PEFR$!

Simple linear regression: The parameters

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

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 $\pm 1.96\sigma$.

Simple linear regression: The estimates (by hand)

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

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.

$$\hat{\beta}_1 = \frac{\sum\left(y_i - \overline{y}\right)\left(x_i - \overline{x}\right)}{\sum\left(x_i - \overline{x}\right)^2} \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \cdot \overline{x}$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_i\right)^2 = \frac{1}{n-2}\sum r_i^2$$

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
-------------+------------------------------           F(  1,    99) =   69.90
       Model |  226303.854     1  226303.854           Prob > F      =  0.0000
    Residual |  320519.473    99  3237.57044           R-squared     =  0.4139
-------------+------------------------------           Adj R-squared =  0.4079
       Total |  546823.327   100  5468.23327           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
------------------------------------------------------------------------------
```

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

Standard errors

95% confidence intervals

Simple linear regression: The distribution of the estimates

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2 \frac{1}{\sum\left(x_i - \overline{x}\right)^2}\right) \qquad \operatorname{se}\left(\hat{\beta}_1\right) = \hat{\sigma}\Big/\sqrt{\sum\left(x_i - \overline{x}\right)^2}$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{\sum\left(x_i - \overline{x}\right)^2}\right]\right) \qquad \operatorname{se}\left(\hat{\beta}_0\right) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum\left(x_i - \overline{x}\right)^2}}$$

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n-2}\chi^2\left(n-2\right)$$

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

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

$$95\% \text{ CI for } \beta_1 : \hat{\beta}_1 \pm t_{n-2}^{0.975} \cdot \operatorname{se}\left(\hat{\beta}_1\right)$$

$$95\% \text{ CI for } \beta_0 : \hat{\beta}_0 \pm t_{n-2}^{0.975} \cdot \operatorname{se}\left(\hat{\beta}_0\right)$$

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:**

$$\text{Approx. } 95\% \text{ CI for } \beta_1 : \hat{\beta}_1 \pm 1.96 \cdot \operatorname{se}\left(\hat{\beta}_1\right)$$

$$\text{Approx. } 95\% \text{ CI for } \beta_0 : \hat{\beta}_0 \pm 1.96 \cdot \operatorname{se}\left(\hat{\beta}_0\right)$$

Simple linear regression: Confidence intervals

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

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

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
```

Changing the reference value and scale for x

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

In this model the parameter β_0 does not make sense.

But if we consider the **equivalent** model:

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

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:

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

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

Simple linear regression: The intercept

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

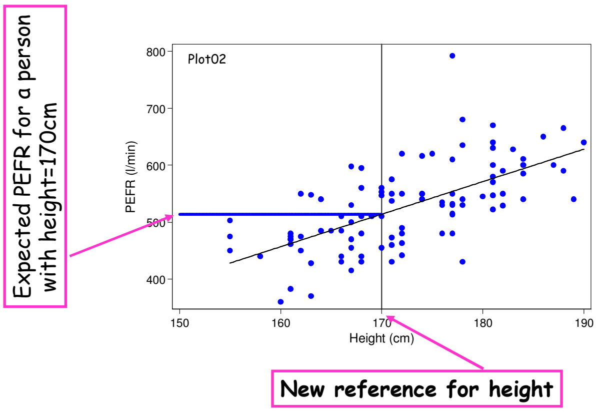
```
generate height170=height-170

regress PEFR height170
```

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



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:

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

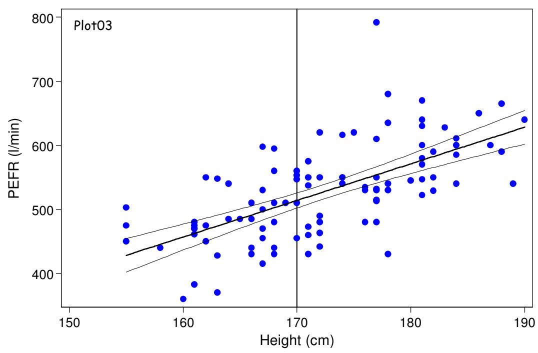
with the 95% (pointwise) confidence interval

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

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

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

Confidence interval for the estimated line




```
twoway ///
 (scatter PEFR height, mco(blue) msym(O))           ///
 (lfitci PEFR height,clpat(1) cip(rline) )          ///
  ,legend(off) ytit("PEFR (l/min)") xline(170 )
```


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: $z = \dfrac{\hat{\beta}_1 - \beta_{1H}}{\text{se}\left(\hat{\beta}_1\right)}$ P-value: $2 \cdot P\left(t_{n-2} < -\left|z\right|\right)$

An example: Hypothesis $\beta_1 = 5$

$$z = \frac{5.712 - 5}{0.6832} = 1.04$$

P-value 30%

In Stata : ***lincom* height170-5**

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 se 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.

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.