Appendix

Regression Models in Systems Biology with R

Uwe Menzel 2014

www.matstat.org

Finding the best regression line

$$SS_{res} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - (\alpha^* + \beta^* x_i)\right)^2$$ to be minimized

$$\left.\begin{aligned} \frac{\partial SS_{res}}{\partial \alpha^*} &= -2 \sum_{i=1}^{n} (y_i - \alpha^* - \beta^* x_i) = 0 \\ \frac{\partial SS_{res}}{\partial \beta^*} &= -2 \sum_{i=1}^{n} x_i \, (y_i - \alpha^* - \beta^* x_i) = 0 \end{aligned}\right\}$$ derivatives must be zero

$$\Rightarrow \left.\begin{aligned} \sum_{i=1}^{n} y_i - n \cdot \alpha^* - \beta^* \sum_{i=1}^{n} x_i &= 0 \\ \sum_{i=1}^{n} y_i \cdot x_i - \alpha^* \cdot \sum_{i=1}^{n} x_i - \beta^* \sum_{i=1}^{n} x_i^2 &= 0 \end{aligned}\right\}$$ two linear equations for α^* and β^* (everything else known from sample)

Finding the best regression line

$$\sum_{i=1}^{n} y_i - n \cdot \alpha^* - \beta^* \sum_{i=1}^{n} x_i = 0 \qquad (1)$$

$$\sum_{i=1}^{n} y_i \cdot x_i - \alpha^* \cdot \sum_{i=1}^{n} x_i - \beta^* \sum_{i=1}^{n} x_i^2 = 0 \qquad (2)$$

$$n \cdot \bar{y} - n \cdot \alpha^* - \beta^* \cdot n \cdot \bar{x} = 0 \qquad (1')$$

$$\sum_{i=1}^{n} y_i \cdot x_i - \alpha^* \cdot n \cdot \bar{x} - \beta^* \sum_{i=1}^{n} x_i^2 = 0 \qquad (2')$$

$$n \cdot \alpha^* = n \cdot \bar{y} - \beta^* \cdot n \cdot \bar{x} \qquad (1'') \text{ in } (2')$$

$$\sum_{i=1}^{n} y_i \cdot x_i - n \cdot \bar{y} \cdot \bar{x} - \beta^* \cdot n \cdot \bar{x}^2 - \beta^* \sum_{i=1}^{n} x_i^2 = 0 \qquad (2'')$$

Finding the best regression line

$$\sum_{i=1}^{n} y_i \cdot x_i - n \cdot \bar{y} \cdot \bar{x} - \beta^* \cdot n \cdot \bar{x}^2 - \beta^* \sum_{i=1}^{n} x_i^2 = 0 \qquad (2'')$$

$$\beta^* \cdot \left[\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2 \right] = \sum_{i=1}^{n} y_i \cdot x_i - n \cdot \bar{y} \cdot \bar{x} \qquad (2''')$$

$$\beta^* \cdot S_{xx} = S_{xy}$$

$$\boxed{\beta^* = \frac{S_{xy}}{S_{xx}}}$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2$$ Definition

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}) = \sum_{i=1}^{n} x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}$$ Definition

Finding the best regression line

$$n \cdot \bar{y} - n \cdot \alpha^* - \beta^* \cdot n \cdot \bar{x} = 0 \qquad (1')$$

$$\Longrightarrow \quad \bar{y} = \alpha^* + \beta^* \cdot \bar{x}$$

That means that the point $(\bar{x}, \bar{y})$ is located on the regression line.

$$\boxed{\alpha^* = \bar{y} - \beta^* \cdot \bar{x}}$$

The slope estimator β^* is a linear combination of the y_i

$$\begin{aligned}
\beta^* &= \frac{S_{xy}}{S_{xx}} = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \\
&= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot y_i \\
&= \sum_{i=1}^{n} c_i \cdot y_i \quad \text{with} \quad c_i = \frac{x_i - \bar{x}}{S_{xx}}
\end{aligned}$$

c_i is not a random variable because x is not random !

This term disappears:

$$\sum_{i=1}^{n} (x_i - \bar{x}) \cdot \bar{y} = \bar{y} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) = \bar{y} \cdot \left[\sum_{i=1}^{n} x_i - n\bar{x} \right] = 0$$

The intercept estimator α^* is a linear combination of the y_i

$$\begin{aligned}
\alpha^* &= \bar{y} - \beta^* \bar{x} = \frac{1}{n}\sum_{i=1}^{n} y_i - \bar{x}\sum_{i=1}^{n} c_i \cdot y_i \\
&= \sum_{i=1}^{n}\left(\frac{1}{n} - c_i \bar{x}\right) y_i \\
&= \sum_{i=1}^{n} d_i \cdot y_i \quad \text{with} \quad d_i = \frac{1}{n} - c_i \bar{x}
\end{aligned}$$

d_i is not a random variable!

The slope estimator β^* is unbiased

$$
\begin{aligned}
E\left(\beta^*\right) &= E\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i E(y_i) = \sum_{i=1}^{n} c_i \left(\alpha + \beta x_i\right) \\
&= \alpha \sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_{xx}} + \beta \sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_{xx}} x_i \quad \text{cause} \quad c_i = \frac{x_i - \bar{x}}{S_{xx}} \\
&= \frac{\alpha}{S_{xx}} \cdot \sum_{i=1}^{n} \left(x_i - \bar{x}\right) + \frac{\beta}{S_{xx}} \sum_{i=1}^{n} \left(x_i - \bar{x}\right) x_i \\
&= \frac{\alpha}{S_{xx}} \cdot 0 + \frac{\beta}{S_{xx}} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(x_i - \bar{x}\right) \quad \text{cause} \quad \sum_{i=1}^{n} \left(x_i - \bar{x}\right) \bar{x} = 0 \\
&= 0 + \frac{\beta}{S_{xx}} S_{xx} \\
&= \beta \qquad \text{unbiased!}
\end{aligned}
$$

The slope estimator β^* is consistent

$$
\begin{aligned}
V(\beta^*) &= V\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} V(c_i y_i) \\
&= \sum_{i=1}^{n} c_i^2 V(y_i) = \sum_{i=1}^{n} c_i^2 \sigma^2 \\
&= \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_{xx}}\right)^2 \sigma^2 \\
&= \frac{\sigma^2}{S_{xx}^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \text{using} \quad S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= \frac{\sigma^2}{S_{xx}}
\end{aligned}
$$

because the noise is independent!

S_{xx} grows when $n \rightarrow \infty$

Extractor functions for “`lm`”

https://www.zoology.ubc.ca/~schluter/R/fit-model

```
summary(z)    # parameter estimates and overall model fit
plot(z)       # plots of residuals, q-q, leverage
coef(z)       # model coefficients (means, slopes, intercepts)
confint(z)    # confidence intervals for parameters

resid(z)      # residuals
fitted(z)     # predicted values
abline(z)     # adds simple linear regression line to scatter plot

predict(z, newdata = mynewdata) # predicted values for new observations
                                # contained in your data frame "mynewdata".
                                # The variable must have the same name
                                # in mynewdata as in original data frame.

anova(z1, z2)  # compare fits of 2 models, "full" vs "reduced"
anova(z)       # ANOVA table (** terms tested sequentially **)
```

Regression ⇄ ANOVA

http://www.strath.ac.uk/aer/materials/4dataanalysisineducationalresearch/unit6/anovaandregression/

- Regression is more flexible than ANOVA
- ANOVA can only have a limited number of categorical predictors
- Regression can include all types of variables (continuous, categorical)
- Regression: different error distributions possible
  - binomial, Poisson, negative binomial ...
- ANOVA is a special case of the GLM. Both consider the observations to be the sum of a model (fit) and a residual (error) to be minimized.

see Regression_and_ANOVA.R

Confounding Variables
- Simpsons Paradox -

https://de.wikipedia.org/wiki/Simpson-Paradoxon

	Treatment A	Treatment B
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

treatment A is better

???

treatment B is better !

Just adding the numbers for "Small Stones" and "Large Stones" fooles you into thinking that treatment B is more succesful!