Crash Course in Statistics for Neuroscience Center Zurich University of Zurich

Dr. C.J. Luchsinger

8 Regression (linear model)

Further readings: Chapter 13 in Stahel or Chapter 11 in Cartoon Guide

Aim of this chapter: you know the simplest, non trivial statistical model (linear model). With this linear model you can describe a (linear) relationship between two variables. Classical example is number of points in Maths and Physics. Theoretical model:

$$Y_i = \beta_1 + \beta_2 x_i + \epsilon_i \tag{8.1}$$

We will assume the x's (Predictor, in German: erklärende Variable) to be fix and try to predict or explain the Y's (Response variable, in German: abhängige Variable) as good as possible. This would be simple if there were not this error term ϵ (in German: Störterm, er stört eben wirklich!). We will then use data points $(x_i, y_i), 1 \leq i \leq n$, to estimate (remember Chapter 6) β_1 and β_2 and test (remember Chapter 7) whether $\beta_1 = 0$ or, usually more important, $\beta_2 = 0$ or not.

Reasons for regression analysis are:

* predict Y with x (interpolation ($x \in [x_{(1)}, x_{(n)}]$) and extrapolation ($x \notin [x_{(1)}, x_{(n)}]$))

* explain relationship between x and y with current data

8.1 Introductory example - Pitfalls

Let us look at a data analysis that shows fallout of radioactivity after the nuclear disaster of Tschernobyl.

A first plot of data (radioactivity vs distance) is not very informative. What is often done in statistics (and often criticized and even more often misunderstood) is a data transformation. More to data transformation can be found in Stahel 2.6. For experienced data analysts, it is clear, that a first data transformation can be to take the logarithm of both axis and look at the plot again.

8.1.1 Try to explain (or predict) radioactivity with distance alone

Using some statistical package (R for example), we get the following estimate

$$\log(bq) = 9.803 - 1.091 * \log(dist) + \epsilon$$

Now: what does this "-1.091" mean if you have to give an advice to the authorities?

8.1.2 Try to explain (or predict) radioactivity with rain alone

Again, using some statistical package (R for example), we get the following estimate

$$\log(bq) = 3.7784 - 0.8247 * \text{rain} + \epsilon$$

Oooups? What happened?

8.1.3 Try to explain (or predict) radioactivity with distance *and* rain

Again, using some statistical package (R for example), we get the following estimate

$$\log(bq) = 10.522 - 1.360 * \log(dist) + 2.723 * \text{rain} + \epsilon$$

How do we interpret this equation?

In 8.1.1 and 8.1.2 we used simple regression. 8.1.3 was multiple regression, which is more complicated to explain in detail. But almost all features of multiple regression can be explained with simple regression. That's why we focus on simple regression and treat multiple regression only briefly in 8.5.

A general remark: our Tschernobyl-data is problematic, because we only have data with rain far away from Tschernobyl. This lead to these strange results in 8.1.2, which were clarified in 8.1.3.

8.1.4 Which is the true model? Which one should we choose?

First question is a typical question asked by people who forget to distinguish between model and reality. I don't know, which is the true model. I am not God and don't even know if God lets the world evolve according to equations like (8.1) with random effects. Only case where this question makes sense is simulated data given to students with them having to estimate the true parameters.

Second question makes much more sense. Some remarks:

* The more you know about Statistics, the more you can choose a good model!

* Scientist must understand model used!

* A good explanation of data with few explaining predictors is a good thing!

8.1.5 Assumptions we are going to make

Data $(x_i, y_i), 1 \leq i \leq n$, comes from RV's such that:

$$Y_i = \beta_1 + \beta_2 x_i + \epsilon_i. \tag{8.1}$$

β_1, β_2 are parameters to be estimated, $x_1, \ldots, x_n$ are fix and $\epsilon_1, \ldots, \epsilon_n$ are iid $\mathcal{N}(0, \sigma^2)$-RV's. σ^2 must be estimated too. We include small effects like

* measurement errors
* random variation
* effects with small influence that we don't want to include in model to keep it simple

in the ϵ.

8.2 Estimation of β_1, β_2 and σ^2 with OLS [Remember Chapter 6?]

Let us denote $(\hat{\beta}_1, \hat{\beta}_2)$ estimators of (β_1, β_2), then

$$\hat{y}_i := \hat{\beta}_1 + \hat{\beta}_2 x_i, 1 \leq i \leq n,$$

is the estimated line. We now want a line trough data points $(x_i, y_i)_{i=1}^n$ such that Sum of Squared Errors

$$SSE := \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 := \sum_{i=1}^{n} (y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_i))^2$$

is minimal (OLS=Ordinary Least Squares):

Results are (if you like maths: use calculus to prove it!)

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \tag{8.2}$$

and

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}. \tag{8.3}$$

Solve exercises 8.1 and 8.2.

We still need an estimate for σ^2: one possibility is to use the so called observed residuals $e_i := y_i - \hat{y}_i, 1 \leq i \leq n$.

$$\hat{\sigma}^2 := \frac{1}{n-2} \sum_{i=1}^{n} e_i^2. \tag{8.4}$$

Why "$(n-2)$"? Because we estimated two parameters from the data!

Without proving it, we state that all 3 estimators are unbiased and consistent and have a more complicated super feature: they are BLUE=best linear unbiased estimator. The "best" is complicated to explain - it has something to do with the variance of the estimator, we omit this point.

8.3 Tests for $\beta_2 = 0$ (and $\beta_1 = 0$) [Remember Chapter 7?]

Test for $\beta_1 = 0$ is done with some not too complicated formula which can be found in statistics text books. The usual test statistic has a t-distribution - we therefore make a t-test. It is (usually) two sided, rejecting the hypothesis that $\beta_1 = 0$ if we have a value too far away from 0. Computers usually do the computations for us.

Let us first talk about why **the question** $\beta_2 = 0$ **or not** is an intelligent question to ask:

We are going to give the test statistic for the general case of $\mathcal{H}_0 : \beta_2 = \beta_*$, for some given β_*, including the usual case of $\beta_2 = 0$ or not (choose $\beta_* = 0$). The test statistic used is

$$T_{n-2} := \frac{\hat{\beta}_2 - \beta_*}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}} = \frac{\hat{\beta}_2 - \beta_*}{\sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}} \quad .$$

It has a t-distribution with $(n-2)$ degrees of freedom. Default in statistical packages is usually such that it tests for $\beta_2 = 0$ or not.

We look at our Tschernobyl-Data again (in the full model!) and see that both explanatory variables (distance and rain) are highly significant. The β_1 (intercept) is significant too, but not of much interest for our data analysis. Small mathematical remark: that data analysis there is with multiple regression. Significance would be slightly different in simple regression, but also very significant.

8.4 Problems & Diagnostic Checking

Only brief overview:

* Outliers

* Variance not constant

* remaining structure in the observed residuals e_i's

* ϵ_i's not independent

* ϵ_i's not normally distributed

8.5 Some remarks about multiple regression

* We used it in Tschernobyl data, as soon as we took both explanatory variables

* Estimators and tests become complicated (need's linear algebra, matrices)

* can not be presented graphically as well as simple regression (2 dimensions!)

* many possible pitfalls for non-experienced data analysts

8.6 Why is linear regression using OLS so well known and accepted?

* can be understood easily by non-quants

* not too complicated mathematics

* simple regression is generalized to multiple regression strait forward

* historically the only method implemented in statistical packages - not relevant argument anymore

why linear (linear is something like "$y = ax + b$")

* human beings can only grasp linear dependence well; examples: "I drive at 100 km/h for 10 hours. How far do I get (linear)?" vs "1.- at 5 % interest rate on bank. How much will I have in 10 years?" Oehae, good question, tya, and the answer is... ?

* non-linear dependencies can be transformed into linear dependencies

why OLS

* other methods are computing intensive - not relevant argument anymore

* OLS is BLUE (Best Linear Unbiased Estimator); stays true for multiple regression

8.7 Exercises

8.1 What happens to equations (8.2) and (8.3) if $\overline{x} = \overline{y} = 0$?

8.2 What happens to (8.2) if estimated variances of x and y are equal, consult 6.1.3?

8.3 We have (for simplicity of computation only) 7 Datapoints (x_i, y_i) from some experiment. These values are $(2, 3.3), (2.1, 3.6), (2.2, 3.5), (2.3, 3.3), (2.4, 3.8), (2.5, 3.9), (2.6, 4.2)$. You now want to make a linear regression. Compute

a) OLS-Estimate of β_1

b) OLS-Estimate of β_2

c) estimated variance of error-terms.

d) Test at 0.05, whether $\beta_2 = 0$ or not.