9. Linear regression, correlation and intro to general linear models

Regression and correlation

Both regression and correlation refer to associations between two quantitative variables. One variable, the predictor, is considered independent in regression, and its values are considered not to be random. The other variable, the response, is dependent on the values of the predictor with a certain level of error variability, i.e. it is a random variable. In the case of correlation, both variables are considered random. Regression and correlation are thus quite different – theoretically. In practice, however, they are numerically identical concerning both the measure of association and p-values (type I error probabilities) associated with rejecting the null hypothesis on independence between the two variables.

Linear regression

Linear association between two quantitative variables X and Y, of which Y is a random variable, can be described by the equation:

$\mathrm{Y} = a + b\mathrm{X} + \varepsilon$

where *a* and *b* are intercept and slope of a linear function, respectively. These represent the systematic (deterministic) component of the regression model, while *ε* is the error (residual) variation representing the stochastic component. *ε* is assumed to follow the normal distribution with mean = 0. The goal of regression model fitting is to estimate the population slope and intercept from sample data of Y and X. *a* and *b* are thus estimates of population parameters. There are multiple approaches to conduct such estimates. Maximum-likelihood estimation is most common, which provides numerically identical results to least-square estimation in ordinary regression. We shall discuss the least square estimation here, as it is fairly intuitive and will help us to understand the relationship with ANOVA. The least-square estimation aims at minimizing the sum of error squares ($\mathrm{SS}_{\mathrm{error}}$), i.e. the squares of the differences between fitted and observed values of the response variable (Fig. 9.1). Note that this mechanism is notably similar to that of analysis of variance. In parallel with ANOVA, we can also define the total sum of squares ($\mathrm{SS}_{\mathrm{total}}$) and the regression sum of squares ($\mathrm{SS}_{\mathrm{regr}}$). Subsequently, we can calculate mean squares (MS) by dividing SS by corresponding DF, with $\mathrm{DF}_{\mathrm{total}} = n - 1$, $\mathrm{DF}_{\mathrm{regr}} = 1$, and
$\mathrm{DF}_{\mathrm{error}} = \mathrm{DF}_{\mathrm{total}} - \mathrm{DF}_{\mathrm{effect}} = n - 2$, where *n* is total number of observations. Hence, we get:

$\mathrm{MS}_{\mathrm{regr}} = \mathrm{SS}_{\mathrm{regr}}/\mathrm{DF}_{\mathrm{regr}}$

$\mathrm{MS}_{\mathrm{error}} = \mathrm{SS}_{\mathrm{error}}/\mathrm{DF}_{\mathrm{error}}$

As in ANOVA, the ratio between MS can be used in an F-test of a null hypothesis that there is no linear relationship between the two variables:

$\mathrm{F}_{\mathrm{DFregr,DFerror}} = \mathrm{MS}_{\mathrm{regr}}/\ \mathrm{MS}_{\mathrm{error}}$

Rejecting the null hypothesis means that the two variables are linearly related. Note, however, that a non-significant result may also be produced in cases when the relationship exists but is not linear (e.g. when it is quadratic).

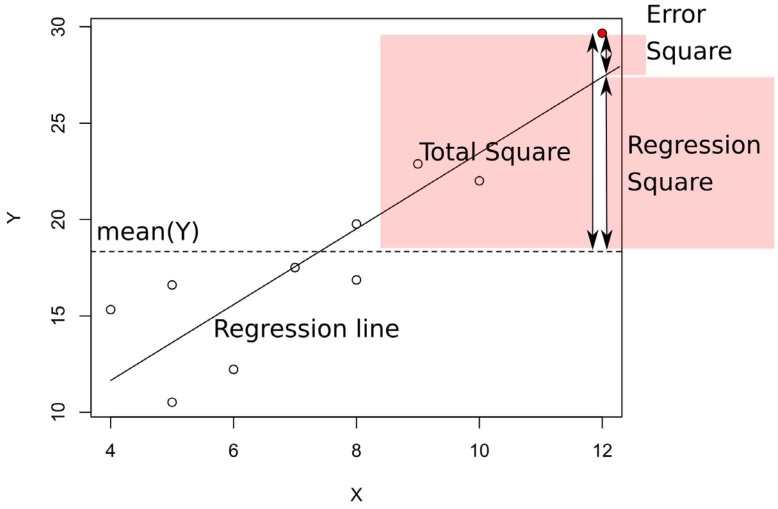

Fig. 9.1 Mechanism of least square estimation in regression: definition of squares exemplified with the red data point.

In regression, we are usually interested in statistical significance, and the strength of the association, i.e. the proportion of variability in Y explained by X. That is measured by the coefficient of determination (R^2):

$R^2 = SS_{regr}/SS_{total}$

which can range from 0 (no association) to 1 (deterministic linear relationship). Alternatively, so-called adjusted-R^2 may be used (and is reported by R), which accounts for the fact that the association is computed from samples and not from populations:

adjusted-$R^2 = 1 – MS_{error}/MS_{total}$

Coming back to the regression coefficients – the fact that these are estimates means that associated errors of such estimates may be computed. Their significance (i.e. significant difference from zero) may thus be tested by a single sample *t*-test. The p-value of such a test for the slope (*b*) is identical to that of the F-test in simple regression with a single predictor. Note that the test of the intercept (reported by R or other statistical software) is irrelevant for the significance of the regression itself. Significant intercept only indicates that mean(Y) is significantly different from zero.

Regression diagnostics

We have discussed the systematic component of the regression equation. However, the stochastic component is also important. This is because its properties can provide crucial information on the validity of regression assumptions and thus the validity of the whole model. The stochastic component of the model, called model **residuals,** can be computed using the equation:

$\varepsilon = Y – a – bX = Y – fitted(Y)$

Residuals form a vector of values for each of the data points. As such, they can be analyzed by descriptive statistics. They may also be standardized by division of their standard deviation. The basic assumptions concerning the residuals are:

1. Residuals should follow the normal distribution
2. The size of their absolute value should be independent of the fitted value.
3. There should be no obvious trend in residuals associated with fitted values, which would indicate the non-linearity of the relationship between X and Y.

These assumptions are best evaluated on a regression-diagnostics plot (Fig 9.2). In addition, it may be worth checking that the regression result is not driven by a single extreme observation (or a few of these), which is provided on the bottom-right plot in Fig 9.2.

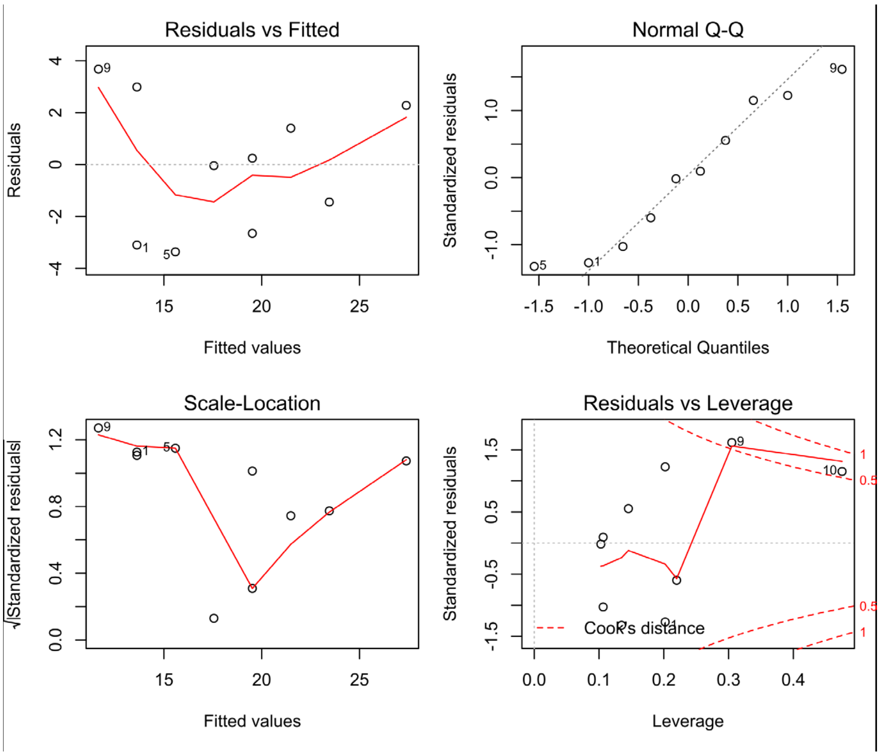

Fig 9.2. Regression diagnostics plots. 1. Residuals vs. fitted values indicate potential non-linearity of the relationship (smoothed trend displayed by the red line). 2. Normal Q-Q plot displays agreement between normal distribution and distribution of residuals (dashed line). 3. Square root of the absolute value of residuals indicate a potential correlation between the size of residuals and fitted values. 4. Residuals vs. leverage (https://en.wikipedia.org/wiki/Leverage_(statistics)) plot detect points, which have a strong influence on the regression parameter estimates (these points have high Cook distance; https://en.wikipedia.org/wiki/Cook%27s_distance).

See also the detailed explanation of regression diagnostics here: https://arc.lib.montana.edu/book/statistics-with-r-textbook/item/57

Correlation

Correlation is a symmetric measure of the association between two random variables, of which neither can be considered a predictor or a response. Correlation is most commonly measured by the Pearson correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Its values can range from -1 (absolute negative correlation) to +1 {absolute positive correlation), with *r* = 0 corresponding to no correlation. r^2 then refers to the amount of shared variability. Numerically, Pearson r^2 and regression R^2 have identical values for given data and have basically the same meaning. Pearson *r* is also an estimate of the population parameter; its significance (i.e. significant difference from zero) can thus be tested by a single sample *t*-test with *n* – 2 degrees of freedom.

On correlation and causality

Note that a significant result of a regression of observational data may only be interpreted as correlation (or coincidence) despite there is a variable called the predictor and the response. Causal explanations imply that a change of predictor value causes a directional change in the response. Causality may, therefore, only be tested in manipulative experiments, where the predictor is manipulated. the See more details on this in Chapter 6.

How to do in R

1. **Regression (or a linear model)**

start with function **lm** to fit the model and save the lm output into an object:

model.1<-lm(response~predictor)

or model.2<-lm(response~predictor1+predictor2+…)

anova(model.1) performs analysis of variance of the model (i.e. tests its significance by an F test). Models may also be compared by **anova(model.1, model.2)**

summary(model.1) displays a summary of the model, including the t-tests of individual coefficients.

resid(model.1) extracts model residuals

predict(model.1) returns predicted values

plot(model.1) plots regression diagnostic plots of the model

2. **Pearson correlation coefficient**

cor(Var1~Var2) computes just the coefficient value

cor.test(Var1~Var2) computes the coefficient value together with significance test

3. Plotting

Plotting a scatterplot with regression line is straightforward in ggplot

geom_point() is used to produce the scatterplot and then

geom_smooth(method="lm”) can be used to add the regression line with confidence intervals. The line color can be adjusted by parameter color in the geom_smooth function.

For instance, the full script to plot e.g. the plot of task #1 of the practicals is:

```
ggplot(snow, aes(x=temp, y=diam))+
geom_point()+geom_smooth(method="lm", color=1)+
theme_classic()+labs(x="Temperature", y="Snowflake diameter")
```