## 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 the case of regression and its values are considered not to be random. The other variable, the response, is dependent on the values of the predictor with certain level of error variability, i.e. it is a random variable. In 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:

$Y = a + bX + \varepsilon$

where $a$ and $b$ are intercept and slope of a linear function, which represents the systematic (deterministic) component of the regression model and $\varepsilon$ is the error (residual) variation representing the stochastic component. $\varepsilon$ is assumed to follow 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 as 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 ($SS_{error}$), 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 ($SS_{total}$) and regression sum of squares ($SS_{regr}$). Subsequently, we can calculate mean squares (MS) by dividing SS by corresponding DF, with $DF_{total} = n - 1$, $DF_{regr} = 1$, and $DF_{error} = DF_{total} - DF_{effect} = n - 2$, where $n$ is total number of observations. Hence we get:

$MS_{regr} = SS_{regr}/DF_{regr}$

$MS_{error} = SS_{error}/DF_{error}$

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:

$F_{DFregr,DFerror} = MS_{regr}/ MS_{error}$

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

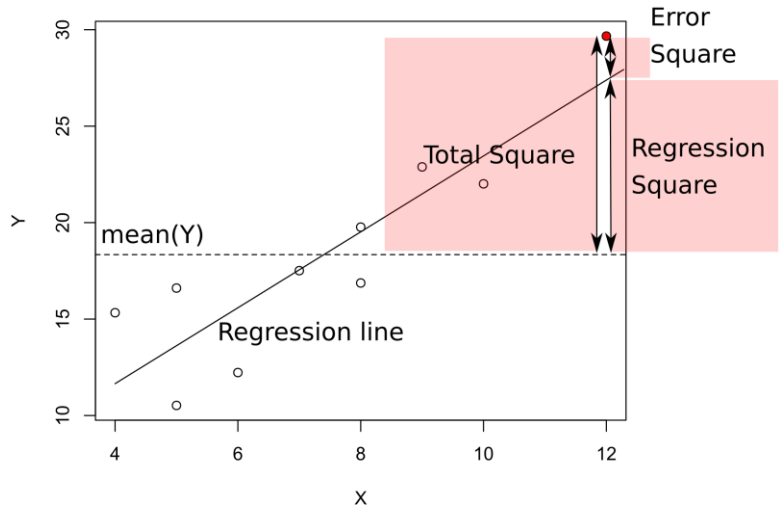


**Fig. 9.1** Mechanism of least square estimation in regression: definition of squares exemplified with the red data point.

In regression, we are usually interested not only in statistical significance but also in 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-$R^2 = 1 - MS_{error}/MS_{total}$

Returning back to the regression coefficients – their estimate nature means that errors of the estimate 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 test for the slope (*b*) is identical to that of the F-test in simple regression with single predictor. Note, that the test of the intercept (reported by R or other statistical software) is irrelevant for 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 validity of regression assumptions and thus validity of the whole model. The stochastic component of the model, called model **residuals,** can be computed using equation:

$\varepsilon = Y - a - bX = Y - \text{fitted}(Y)$

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. Size of their absolute value should be independent of fitted value.
3. There should be no obvious trend in residuals associated with fitted values, which would indicate 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 to check that the regression result is not driven by a single extreme observation (or few of these), which is provided on the bottom-right plot on Fig 9.2.

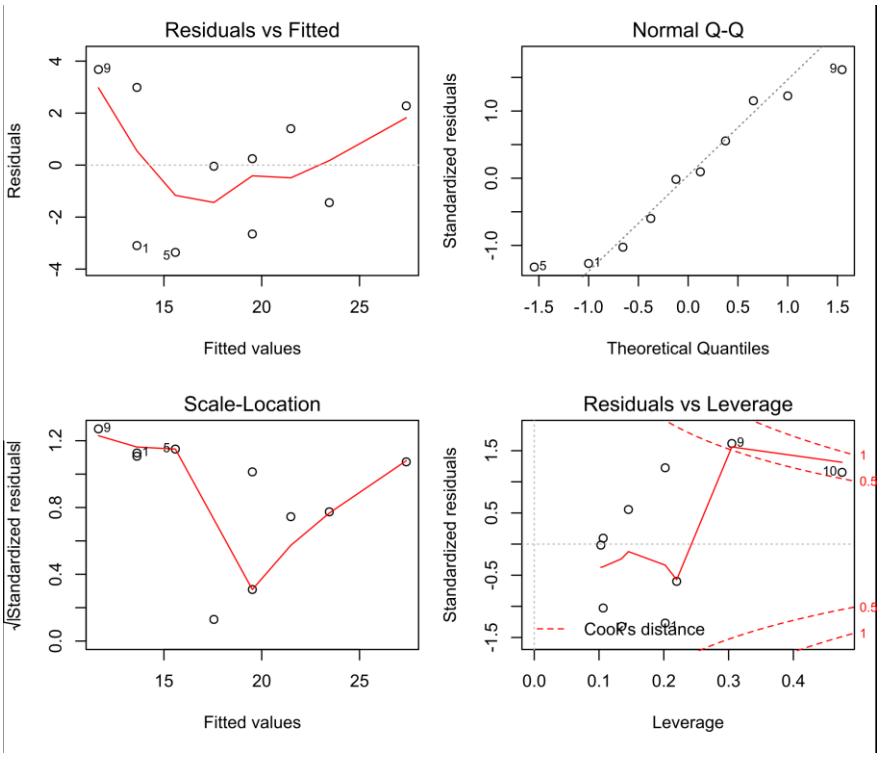


**Fig 9.2.** Regression diagnostics plots. 1. Residuals vs. fitted values indicate potential non-linearity of the relationship (smoothed trend displayed by red line). 2. Normal Q-Q plot displays agreement between normal distribution and distribution of residuals (dashed line). 3. Square root of absolute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

*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 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 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 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. See more details on this in Chapter 6.

*Brief introduction to multiple regression and general linear models*

In regression, multiple predictors may be used in a model:

$Y = a + b_1X + b_2X + \ldots + b_nX + \varepsilon$

provided that number of predictors is lower than number of observations – 1. While statistical testing of such models is easy and largely follows the same principles as in simple regression, it is more difficult to decide which predictors to include in the model and which not (i.e. how to find the best model) as some may have significant effects while others not. This is done by a model selection procedure, details of which are beyond the scope of the present text.

It is important, that categorical predictors of ANOVA may be decomposed into a series of $k$ - 1 quantitative variables (e.g. binary variables containing 0s and 1s), where $k$ is the number of categories.

This means that ANOVA and regression are fundamentally identical. In addition, both quantitative and categorical variables may be used as predictors in a single analysis. Statistical models containing a single response variable and possibly multiple quantitative and categorical predictors and assuming normal distribution of the residuals are called general linear models (or simply linear models).

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 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