# NRES 798 — Statistical Methods for Ecologists
# Chapter 9: Regression

Oscar García

March 20, 2013

## Contents

**1 Regression models, least-squares** **1**

**2 Linear regression** **3**

**3 Hypothesis testing** **5**
3.1 $t$-tests . . . . . 5
3.2 $F$-ratio tests, variance components . . . . . 6
3.3 ANOVA for linear regression . . . . . 7

**4 Confidence intervals for $y$ in simple linear regression** **8**

**5 Design** **9**

**6 Model selection** **10**

**7 Transformations** **10**

## 1 Regression models, least-squares

Regression models relationships between a random variable $Y$ (dependent variable, response) as a function of other variables $x_i$ (independent variables, predictors). The general form is:

$$Y = f(\boldsymbol{x}, \boldsymbol{\beta}) + \varepsilon \,,$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$ is a vector of predictors, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ is a vector of unknown parameters, and $\varepsilon$ is a random variable with mean 0 and variance $\sigma^2$. In other words, $f(\boldsymbol{x}, \boldsymbol{\beta})$ is the mean of $Y$, and depends on the values of $\boldsymbol{x}$.

An example studied in detail in the lab, is estimating tree volume ($Y =$ *Volume*) using variables that are easier to measure, the diameter at breast height ($x_1 =$ *Dbh*) and the tree height ($x_2 =$ *Height*). A possible $f$ might be $Y = \beta_1 x_1^{\beta_2} x_2^{\beta_3}$. We want to find a $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ that produces a good fit.

The data is assumed to be obtained as a simple random sample, so that the observations are independent, and they all have the same distribution as the population (they are independent and identically distributed, *iid*):

$$Y_i = f(\boldsymbol{x}_i, \boldsymbol{\beta}) + \varepsilon_i \ , \tag{1}$$

with $i = 1, 2, \ldots, n$. The actual values obtained may be written as

$$y_i = f(\boldsymbol{x}_i, \boldsymbol{\beta}) + e_i \ ,$$

where $y_i$ and $e_i$ are simple real numbers, not RVs. The *residuals* $e_i = y_i - f(\boldsymbol{x}_i, \boldsymbol{\beta}) = y_1 - \hat{y}_i$ are the differences between observed and predicted values.

The parameters are usually estimated by the *method of least-squares*. It consists of estimating $\boldsymbol{\beta}$ by the value $\hat{\boldsymbol{\beta}}$ that minimizes the residual sum of squares

$$\mathrm{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - f(\boldsymbol{x}_i, \boldsymbol{\beta})]^2 \ . \tag{2}$$

For general $f$, this is done with iterative procedures, like the function `nls` in *R*. That is called *nonlinear regression*. If $f$ is a linear function of the parameters (*linear regression*), then there are explicit formulas for $\hat{\boldsymbol{\beta}}$.

One reason for minimizing RSS is that we want the deviations (residuals) to be small, and the function to go through the "middle" of the data. Positive deviations should balance negative deviations. Squaring the deviations a square loss function is minimized, resulting in the estimated $y$ approaching the mean for a given $\boldsymbol{x}$, subject to the constraints of the function form. Instead of squares, absolute values could be used, tending to the median instead of the mean; that is used in *quantile regression*. However, least-squares is by far more commonly used and mathematically more convenient.

If the $\varepsilon$ are assumed to be normally distributed, another justification is that then the least-squaresestimates are the maximum-likelihood (ML) estimates[1]. ML has desirable statistical properties, mainly asymptotically for large samples.

It will be seen that in linear regression the Gauss Markov Theorem gives yet another justification for least-squares, that does not depend on assuming normality.

The variance $\sigma$ is estimated by

$$\hat{\sigma^2} = \frac{\text{RSS}}{n-p} \; . \tag{3}$$

The square root is called the *standard error of regression* or the *residual standard error* (SE or RSE) [2].

## 2 Linear regression

Linear regression is a special case of the general equation (1), where $f$ is a linear function of the parameters. Usually, but not always, the first variable is taken as the constant 1, with a corresponding parameter $\beta_0$, representing the *intercept* (the value of the function when all the $x_j$ are 0):

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon \; ,$$

or

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + e_i \; .$$

For $p > 2$, this is *multiple linear regression.* The special case $p = 2$ is *simple linear regression* (SLR):

$$y_i = \beta_0 + \beta_1 x + e_i \; .$$

Explicit formulas for the parameter estimators can be obtained equating to 0 the partial derivatives of $\sum e_i^2$ with respect to the $\beta$'s, and solving

[1] If $\varepsilon_i \sim N(0, \sigma)$, the PDF of $Y_i$ is $\exp[-\varepsilon_i^2/(2\sigma^2)]/\sqrt{2\pi\sigma^2}$. Because the observations are independent, the PDF for the sample is the product of these. The likelihood function is the sample PDF evaluated at the observed data, looked at as a function of the parameters. It is therefore proportional to $\prod_i \exp[-e_i^2/(2\sigma^2)] = \exp[-\sum_i e_i^2/(2\sigma^2)]$ The maximum occurs where $\sum_i e_i^2$ is minimum.

[2] Maximizing the likelihood gives the ML estimate $\hat{\sigma^2} = \text{RSS}/n$. This is biased, and the unbiased estimate (3) is usually preferred although it is somewhat less precise.

these $p$ equations for the $\beta$'s (exercise!). Statistical software arranges the calculations somewhat differently, to reduce the effects of rounding error in difficult situations. Anyhow, this is more reliable than nonlinear regression, with no failures of convergence or the danger of converging to spurious local optima.

In *R* all the calculations are done by the function `lm` (linear model). Use `summary` on the result to get more outputs, including $p$-values, etc.

Note that the regression function is linear *in the parameters*, not necessarily in the variables of interest. Both $y$ and the $x$'s can be variable transformations, so that the regression can represent curves (or curved surfaces, etc.) on the original variables. E.g., $\log y = \beta_0 + \beta_1 \log x$, $y = \beta_0 + \beta_1 x + \beta_2 x^2$, $1/z = \beta_1/x + \beta_2/y$ [3].

The RSS and RSE are as before.

The *correlation coefficient* $r$, and its square the *coefficient of determination* or *R-square* $r^2$, are often calculated as an index of fit. For SLR,

$$r = \frac{s_{xy}}{s_x s_y} ,$$

where $s_{xy} = \frac{1}{n-1}\sum(x_i - \overline{x})(y_i - \overline{y})$ is the sample covariance, and $s_x^2 = \frac{1}{n-1}\sum(x_i - \overline{x})^2$and $s_y^2$ are the sample variances of $x$ and $y$. Another way of calculating $r^2$ is shown in the next section. This is often misused and misinterpreted. As a measure of association $r$ is only valid when $x = X$ is random, with the pairs $(X, Y)$ being a random sample from a bivariate normal distribution. Otherwise, it is seen from the formula above that $r$, and the R-square, change depending on how spread out the $x_i$ are ($s_x$). At best, R-square can be used to compare alternative models based on the same data, but then the RSE may well be more meaningful.

We saw a very general heuristic justification for the least-squares method, and also that it gives the ML estimates if the distribution is normal. It would be nice if the estimates were always the best even without normality, for instance in the sense of having the smallest sample variance and no bias. Of course, that may be too much to ask. But the Gauss-Markov Theorem comes close, saying that regardless of distribution, in linear regression the least-squaresparameter estimators are unbiased with minimum variance among all linear estimators, that is, those of the form $\sum w_i y_i$, for some set of weights $w_i$.

[3] An "inverse polynomial", usually better fitted as $xy/z = \beta_1 y + \beta_2 x$

They are *best linear unbiased* (BLUE). The limitation to linear estimators might seem a little arbitrary, but perhaps not too unreasonable.

# 3 Hypothesis testing

One is often interested in testing a null hypothesis that one or more of the $\beta$'s are zero, or that they equal some given constant. This can be done with $t$-values, test statistics having a $t$ distribution, or with $F$-ratios, that have an $F$ distribution.

## 3.1 $t$-tests

Take as an example a SLR $\hat{y} = \beta_0 + \beta_1 x$. One might want to test $H_0$: $\beta_0 = 0$, the hypothesis that the line goes through the origin. Or $H_0$: $\beta_1 = 0$, that there is no relationship between $x$ and $y$. Or $H_0$: $\beta_1 = 1$, that the slope is 1.

In general, for any $k$, the null hypothesis $\beta_k = b$ can be tested calculating the test statistic

$$\frac{\hat{\beta}_k - b}{\mathrm{SE}_{\beta_k}},$$

where the parameter estimate $\hat{\beta}_k$ and its estimated standard error $\mathrm{SE}_{\beta_k}$ can be obtained from the computer regression output. This test statistic, sometimes called a $t$-value, has a $t$ distribution with $n-p$ degrees of freedom.

The hypothesis test goes like always, let's go over it one more time (draw a picture; the $t$ PDF is bell-shaped, symmetric around 0). The observed test statistic, call it $t$, can be positive or negative, and under $H_0$ it should not be too far from 0. If it is negative, the probability of observing a $t$ less that what we got is the area under the PDF to the left of $t$. It can be calculated as the value of the CDF at $t$ (function `pt` in $R$). As this is a two-tailed test, we have to add the similar area at the other end, so that the $p$ value is that number multiplied by 2. If $t$ was positive the procedure is similar, work it through. If $p$ is smaller than the standard significance level $\alpha$, e.g., 0.05, then $H_0$ is rejected, it seems unlikely to get a $t$ like this by pure chance if $H_0$ is true. Otherwise, $H_0$ is "accepted".

With the more traditional method, the critical values are set in advance. The acceptance region is the interval around 0 that has probability (area

under the PDF) equal to $1-\alpha$. Approximately $[-2,2]$ for $\alpha = 0.05$; it can be calculated with the $t$ quantile function (`qt`). If the observed $t$ falls outside that, $H_0$ is rejected, etc.

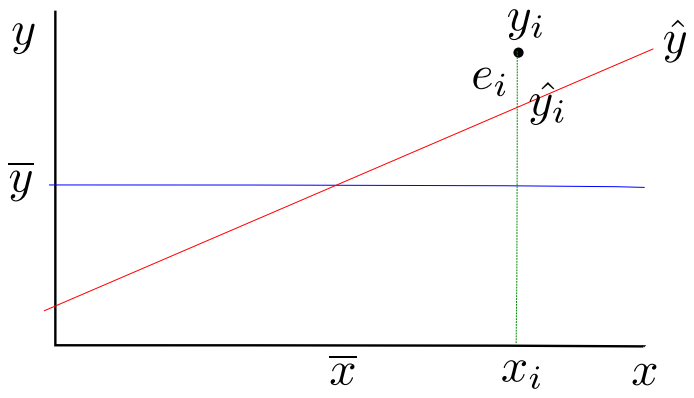
This is closely related to confidence intervals. For $t$-tests in general, the limits are of the form estimate$\pm t_\alpha$SE, where $t_\alpha$ is the $t$-distribution quantile for the confidence level $\alpha$, around 2 for $\alpha = 0.05$. The confidence interval for $\beta_k$ is the rejection region centered around $\hat{\beta}_k$. You can see that $H_0$ is rejected if the confidence interval does not contain $b$.

The `summary` of the `lm` fit in *R* gives the $t$- and $p$-values for the common case $b = 0$. The confidence limits are obtained with `confint`.

All this works the same for nonlinear regression, but there the results are only approximate because the test statistic does not have exactly a $t$-distribution.

### 3.2 $F$-ratio tests, variance components

This is less intuitive than the $t$ tests for most people that are not ANOVA experts. It is based on splitting the variability into components, and forming ratios among them. We explain the process in detail in a SLR, because the ideas form the basis for the classical analysis of variance (ANOVA).




The picture shows an SLR line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ [4]. The deviations from the mean can be partitioned into a a part "due to" the regression, and the residual

[4] It can be shown that it can also be written as $\hat{y} = \overline{y} + \hat{\beta}_1(x - \overline{x})$, which is sometimes useful.

deviation from the regression line:

$$y_i - \overline{y} = (\hat{y}_i - \overline{y}) + e_i \ .$$

Squaring and adding over all the observations:

$$\begin{aligned}(y_i - \overline{y})^2 &= (\hat{y}_i - \overline{y})^2 + e_i^2 + 2(\hat{y}_i - \overline{y})e_i \\ \sum (y_i - \overline{y})^2 &= \sum (\hat{y}_i - \overline{y})^2 + 0 \\ \mathrm{SS}_y &= \mathrm{SS}_{\mathrm{reg}} + RSS\end{aligned}$$

The sum of products turns out to be 0 because the components are "orthogonal", we omit the proof. The result is that the sum of squares around the mean can be partitioned into a sum of squares due to regression and a residual sum of squares.

If there were no relationship, $\beta_1 = 0$, the line would be close to horizontal and $\mathrm{SS}_{\mathrm{reg}}$ should be around 0. If the relationship were perfect, all the points would be on the line, RSS would be 0, and $\mathrm{SS}_{\mathrm{reg}}$ would equal the total sum of squares $\mathrm{SS}_y$. Therefore $\mathrm{SS}_{\mathrm{reg}}$, or some function of it could be used as a test statistic for the null hypothesis $H_0$: $\beta_1 = 0$. The ratio of mean squares ($F$-ratio)

$$\frac{\mathrm{SS}_{\mathrm{reg}}/1}{\mathrm{RSS}/(n-2)}$$

has an $F$ distribution with 1 and $n-2$ degrees of freedom, and can be used to test $H_0$. The null hypothesis would seem unlikely if the $F$-ratio is large, the test is one-tailed.

This $F$-ratio test is equivalent to the corresponding $t$-test, giving the same $p$-value. In fact, the square of a $t$ variable has an $F$ distribution.

Incidentally, the ratio

$$r^2 = \frac{\mathrm{SS}_{\mathrm{reg}}}{\mathrm{SS}_y}$$

is another way of computing the R-square from the previous section. It shows it as the proportion of the variation in $y$ that is "explained by" the regression.

### 3.3 ANOVA for linear regression

It is conventional to present the $F$-ratio test above in the form of a table, based on how the calculations were done by hand in the old days. It is

perhaps curious that this historical artifact has persisted, but we discuss it in detail because its use is standard in what is known as the analysis of variance. The first column in the table shows the "sources" of variation, with the sums of squares in the second column (frequently the degrees of freedom are shown in second place, but I will leave them for later):

| Source | SS | df | MS | $F$-ratio | $p$-value |
|---|---|---|---|---|---|
| Regression | $\mathrm{SS}_{\mathrm{reg}}$ | 1 | $\mathrm{SS}_{\mathrm{reg}}$ | $\mathrm{SS}_{\mathrm{reg}}/[\mathrm{RSS}/(n-2)]$ | |
| Residual | RSS | $n-2$ | RSS / $(n-2)$ | | |
| Total | $\mathrm{SS}_y$ | $n-1$ | | | |

$\mathrm{SS}_y$ and RSS are easy to calculate, $\mathrm{SS}_{\mathrm{reg}}$ not so much. So the $\mathrm{SS}_{\mathrm{reg}}$ entry used to be filled from the difference $\mathrm{SS}_y - \mathrm{RSS}$. The degrees of freedom (df) for the total are $n-1$, because one estimates the mean. For the residual it is $n-2$, from the estimation of the two $\beta$'s, and the regression df can be filled-in as a difference. Then the mean squares (MS) are calculated dividing the sums of squares by the degrees of freedom. Finally, the ratio of the regression MS to the residual MS is the $F$-ratio. The corresponding $p$-value may be included, or the significance can be shown with stars or by saying "$< 0.05$" or "$< 0.01$", or both.

Nowadays the last row is commonly omitted in publications. More complex ANOVAs include more sources, but the principles are the same.

# 4 Confidence intervals for $y$ in simple linear regression

Writing $\hat{y} = \overline{y} + \hat{\beta}_1(x - \overline{x})$, the usual properties of the variance and the fact that $\overline{y}$ and $\hat{\beta}_1$are independent give $V[\hat{y}] = V[\overline{y}] + V[\hat{\beta}_1](x-\overline{x})^2$. Using the estimated sample variances $V[\overline{y}] = \hat{\sigma^2}/n$ and $V[\hat{\beta}_1] = \hat{\sigma^2}/\mathrm{SS}_x$, the variance of the regression estimate is

$$V[\hat{y}] = \hat{\sigma^2}\left[\frac{1}{n} + \frac{(x-\overline{x})^2}{\mathrm{SS}_x}\right] .$$

The SE is the square root of this. As might be expected, it coincides with the SE of the mean when $x = \overline{x}$, but increases as $x$ gets away from $\overline{x}$. This can be used to draw confidence limits around the regression line as $\hat{y} \pm t_\alpha \mathrm{SE}_{\hat{y}}$.

Also of interest might be the confidence limits for the $y$ in a future $(x, y)$ pair. The new $y$ would be the $\hat{y}$ at $x$ plus a residual $\varepsilon$. Therefore, the estimated variance would be the one above plus $\widehat{V[\varepsilon]} = \hat{\sigma}^2$.

# 5 Design

Very often one can influence the predictors (the $x$'s) through the sampling strategy in observational studies, or they can be freely chosen in experimental work. No statistical assumptions are violated by doing that, provided that one is careful not to influence the $y$ for a given $x$. What would be the best way of choosing $x$?

In simple linear regression, it was seen in Section 4 that both the SE of $\hat{\beta}_1$ and the regression and prediction errors have $\mathrm{SS}_x$ in a denominator. Therefore, precision improves if the spread (SS, variance, SD) of $x$ is wider. Clearly, the best would be to choose half of the observations with an $x$ as large as we can get, and the other half with $x$ as small as possible. That assumes that we trust the model; it is usually a good idea to get some observations around the middle to check the linearity.

The same is true for each predictor in multiple regression. An additional issue in this case is *multicolinearity*. It is common to find relationships between some of the predictors, for instance between diameter and height in the example of estimating tree volumes. If there is a close linear relationship between two predictors, then it is difficult to disentangle the corresponding parameters. Confidence intervals for the parameters become large, and the power of some hypothesis tests are low. The consequences for $y$ predictions are less serious, the main problem being the use of predictors that become redundant. In some instances of observational studies there is little that can be done about this, in other situations one may strive to spread out the observations away from the correlation line.

In general, in regression getting samples that are "representative" or evenly distributed is not a good idea. It is better to try covering the extremes, with some points elsewhere to check the appropriateness of the models.

# 6 Model selection

See Lab 9.

# 7 Transformations

Important assumptions in linear regression are that the relationship is linear on the parameters, and that the variance of the residual RV $\varepsilon$ is independent of the predictors (homoscedasticity). A third assumption, that $\varepsilon$ has a normal distribution, is important for hypothesis testing but not so much for response estimation and prediction.

Transformations of the predictors can help with linearity. For instance, remember the use of logarithms, `Dbh^2` and `Dbh^2 * Height` for tree volume estimation in Lab 8.

Transformations of the response can help with linearity, homoscedasticity, and normality. It may be optimistic to expect that one transformation will produce a model satisfying the three assumptions, but in some instances it can get reasonably close.

It is common for the standard deviation of a measurement error. $\varepsilon$ to increase roughly proportionally to the size of the measurement $Y$. Then, $\log Y$ is approximately homoscedastic. Typically $Y$ is a non-negative variable, and the predictors and error are multiplicative. In that case, the logarithm produces a linear relation with the logs of the predictors and the error term, and also can make the distribution more symmetric. For these reasons, logarithmic transformations, changing $Y \to \log Y$, are commonly used.

Small counts usually tend to be approximately Poisson. Apart from $Y$ then being asymmetric, the variance iincreases with the mean (in the Poisson both equal the rate parameter $\lambda$), failing the homoscedasticity assumption. A square root transformation, $Y \to \sqrt{Y}$, improves the situation.

When $Y$ is a proportion, $Yn$ tends to be close to a binomial distribution, where again the variance is related to the mean. In addition, it is desirable to constrain the predicted $Y$ to be between 0 and 1. Three different transformations are commonly used in this case: the arc sine transformation $Y \to sin^{-1}\sqrt{Y}$, the *logit* $Y \to \log[Y/(1-Y)]$, and the *probit*, which is a

normal quantile function. A simple linear regression using the logit is called *logistic regression*.

An more sophisticated alternative to transformation in these cases is to use *generalized linear models* (GLM). These are a special form of nonlinear model, and are implemented in *R* in the `glm` package.

It should be recognized that many (most?) variables used in practice cannot have exactly a normal distribution, because they cannot be negative (weights, lengths, pH, etc.). What is usually required is a rough approximation.

Care may be needed when values are close to zero. In the tree volumes example $V \approx \beta_0 + \beta_1 D^2 H$, if $V$ is a utilizable or merchantable volume then $\beta_0$ is usually negative, there is no volume in small trees, for $D^2 H$ smaller than some positive value. Negative volume estimates do not make sense, and the model is actually $\max\{\beta_0 + \beta_1 D^2 H, 0\}$. Data to the left of the zero-crossing point should be excluded, because it does not correspond to the linear model and would bias the estimates. In fact, points close to that point on the right cannot have a near-normal $V$, and should be excluded too.