Quantitative Analysis in Ecology and Evolution

The Linear Model II: Analysis of Variance

Øystein H. Opedal

11 Oct 2022

6. The linear model IV: multiple regression

Linear models are easily extendable to multiple predictor variables. If there are several continuous predictors, the analysis is called a multiple-regression analysis. Multiple regression has some very useful properties. For example, the parameter estimates represent the *marginal effect* of each predcitor, that is the effect of the predictor when all other variables in the model are held constant at their mean. This allows us to evaluate the independent effects of several, potentially correlated, variables (asking for example which have the stronger effect on the response variable), or to 'control for' some nuisance variables (say, sampling effort).

```
set.seed(187)
x1 = rnorm(200, 10, 2)
x2 = 0.5*x1 + rnorm(200, 0, 4)
y = 0.7*x1 + 2.2*x2 + rnorm(200, 0, 4)

m = lm(y~x1+x2)

summary(m)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4276 -2.7240 -0.0065  2.7041  9.7580
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.48722    1.34745   0.362    0.718
## x1           0.64178    0.13246   4.845 2.56e-06 ***
## x2           2.18446    0.06422  34.017  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.618 on 197 degrees of freedom
## Multiple R-squared:  0.8683, Adjusted R-squared:  0.8669
## F-statistic: 649.3 on 2 and 197 DF,  p-value: < 2.2e-16
```

```
coefs = summary(m)$coef
```

First, note that the r^2 of the model is 0.868, which means that 86.8% of the variance in y is explained. As before, we can see why this is the case by computing the variance in the predicted values $\hat{y}$,

$$V(\hat{y}) = X\beta$$

, and then divide this by the total variance in the response varible $V(y)$.

```
y_hat = coefs[1,1] + coefs[2,1]*x1 + coefs[3,1]*x2
var(y_hat)
```

```
## [1] 85.4221
```

```
var(y_hat)/var(y)
```

```
## [1] 0.8682827
```

This is the total variance explained by the model. Now what about the variance explained by each of the predictors x_1 and x_2? To compute the predicted values associated only with x_1, we keep x_2 constant at its mean, and *vice versa* for the variance associated with x_2.

```
y_hat1 = coefs[1,1] + coefs[2,1]*x1 + coefs[3,1]*mean(x2)
var(y_hat1)
```

```
## [1] 1.608668
```

```
var(y_hat1)/var(y)
```

```
## [1] 0.01635149
```

```
y_hat2 = coefs[1,1] + coefs[2,1]*mean(x1) + coefs[3,1]*x2
var(y_hat2)
```

```
## [1] 79.29333
```

```
var(y_hat2)/var(y)
```

```
## [1] 0.8059861
```

```
var(y_hat)
```

```
## [1] 85.4221
```

```
var(y_hat1) + var(y_hat2)
```

```
## [1] 80.902
```

So, what happened to the last few percent of the variance? Recall that

$Var(x + y) = Var(x) + Var(y) + 2Cov(x, y)$.

```
var(y_hat1) + var(y_hat2) + 2*cov(y_hat1, y_hat2)
```

```
## [1] 85.4221
```

As before, we can also do this by computing $V(x) = \beta_x^2 \sigma_x^2$.

```
coefs[2,1]^2*var(x1)
```

```
## [1] 1.608668
```

To include the covariance between the predictors, we can do this in matrix notation $V(\hat{y}) = \hat{\beta}^{\mathbf{T}} \mathbf{S} \hat{\beta}$. Recall the matrix multiplication operator `%*%`.

```
t(coefs[2:3,1]) %*% cov(cbind(x1,x2)) %*% coefs[2:3,1]
```

```
##          [,1]
## [1,] 85.4221
```

The latter approach is the most general, as it extends to any number of predictors in the model. Note that we could, for example, compute the variance explained by a subset of the predictors by specifying the correct vector of β coefficients and their corresponding variance-covariance matrix. This is useful if we had, say, 3 variables related to climate and 3 other variables related to local land-use, and wanted to know how these sets each explain variance in some variable (say, the size of pine trees).

This procedure also hints at a method for obtaining parameter estimates that directly reflect the strength of the effects of each predictor. If all variables had the same variance, then the variance explained would be directly proportional to the regression slope. The most common way to standardize predictor variables is to scale them to zero mean and unit variance, a so-called z-transform

$z = \frac{x - \bar{x}}{\sigma(x)}$

The resulting variable will have a mean of zero and a standard deviation (and variance) of one.

```
x1_z = (x1 - mean(x1))/sd(x1)
x2_z = (x2 - mean(x2))/sd(x2)

m = lm(y~x1_z + x2_z)
summary(m)
```

```
##
## Call:
## lm(formula = y ~ x1_z + x2_z)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4276 -2.7240 -0.0065  2.7041  9.7580
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.4090     0.2558  75.866  < 2e-16 ***
## x1_z          1.2683     0.2618   4.845 2.56e-06 ***
## x2_z          8.9047     0.2618  34.017  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.618 on 197 degrees of freedom
## Multiple R-squared:  0.8683, Adjusted R-squared:  0.8669
## F-statistic: 649.3 on 2 and 197 DF,  p-value: < 2.2e-16
```

Note that the model fit (e.g. the r^2) has not changed, but the parameter estimates have. First, the intercept can now be interpreted as the mean of y, because it represents the value of y when both predictors have a value of 0 (i.e. their mean after the z-transform). This effect can be obtained also by mean-centering the variables without scaling them to a standard deviation of 1.

Second, the slopes now have units of standard deviations, i.e. they describe the change in y per standard deviation change in each predictor. This shown directly that the predictor x_2 explains more variance in y than does x_1.

Another useful transformation could be a natural log-transform, or similarly mean-scaling, which would give the slopes units of means, and allow interpreting the change in y per percent change in x.

```
x1_m = (x1 - mean(x1))/mean(x1)
x2_m = (x2 - mean(x2))/mean(x2)

summary(lm(y~x1_m + x2_m))
```

```
##
## Call:
## lm(formula = y ~ x1_m + x2_m)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4276 -2.7240 -0.0065  2.7041  9.7580
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.4090     0.2558  75.866  < 2e-16 ***
## x1_m          6.5254     1.3468   4.845 2.56e-06 ***
## x2_m         12.3964     0.3644  34.017  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.618 on 197 degrees of freedom
## Multiple R-squared:  0.8683, Adjusted R-squared:  0.8669
## F-statistic: 649.3 on 2 and 197 DF,  p-value: < 2.2e-16
```

Multicollinearity

When we have several predictors that are strongly correlated with each other, it becomes difficult to estimate their independent effects. A rule of thumb is that such *multicollinearity* becomes a potential problem when the correlation between the predictors is greater than 0.6 or 0.7. One way of assessing the degree of multicollinearity is to compute *variance inflation factors*, defined as

$VIF_i = \frac{1}{1-r_i^2}$

where the r^2 is from a regression of covariate i on the other covariates included in the model. For our example model, the variance inflation factor for covariate x_1 is thus

```r
m1 = lm(x1~x2)
r2 = summary(m1)$r.squared
1/(1-r2)
```

```
## [1] 1.041714
```

This is very low, because the two predictors are not strongly correlated. Rules of thumb for what constitutes severe variance inflation range from $VIF > 3$ to $VIF > 10$. When this occurs, the parameter estimates becomes associated with excessive variance and are thus less reliable. In these cases it may be good to simplify the model by removing some of the correlated predictors, especially if there are several predictors that essentially represent the same property (e.g. multiple measures of body size). If the effects of the correlated predictors are of specific interest, it can also make sense to fit alternative models including each of the candidate predictor, and compare estimates. If the 'best model' is desired, the choice among the predictor may be based on model selection techniques.

Data exercise: multiple regression and variable selection

A common problem in biology is that we have a large number of variables, many of which could potentially predict variation in a response variable. Including a lot of predictors in the same model can lead to problems with multicollinearity, with 'overfitting' (resulting in a model that explains a lot of variance but fails to predicts independent test data), and difficulties in interpretation.

We will return in a later section to a more complete treatment of the problem of model selection, but for now let's consider two main approaches to choosing among a large set of potential variables. A 'statistical' approach to the problem is to look for the simplest model that does a decent job in explaining variation in the data. This can be done e.g. by so-called backward selection of variables, which means that we start from a full model including all potential predictors, and then sequantially drop non-significant terms until all terms are statistically significant.

The problem with this approach is that is focusses on hypothesis testing over interpretation of effects. As we have discussed previously, a 'significant' hypothesis test does not necessarily mean that the effect is biologically important. One strategy for avoiding this fallacy is to start from a well defined biological hypothesis that can be formulated as a statistical model. The focus is then moved from statistical hypothesis testing to a 'simpler' task of parameter estimation and interpretation.

In the following exercise, we will try both approaches for the same dataset, and see if we end up with the same final model. The following dataset includes data on the local abundance of two alpine plant species, measured as the number of times the species was hit in a so-called pinpoint analysis, where 25 metal pins were passed vertically through the vegetation within a 25*25 cm plot. The data also include a number of environmental variables. Start by exploring the data. Then, think about some possible hypotheses (models) that could explain the distribution of the plant species. Fit the models and interpret the results.

Then, do a backward selection in which you start from a saturated (full) model and sequentially drop non-significant terms.

```r
plants = read.csv(file="datasets/alpineplants.csv")
```