Chapter 6

Linear Models II

6.1 Key ideas

Consider a situation in which we take measurements of some attribute $Y$ on two distinct groups. We want to know whether the mean of group 1, $\mu_1$, is different from that of group 2, $\mu_2$. The null hypothesis is equality, that is, $H_0 : \mu_1 = \mu_2$ or $H_0 : \mu_1 - \mu_2 = 0$. Suppose that we sample $n_1$ individuals from group 1 and $n_2$ individuals from group 2. A model for these data is, for $i = 1, 2$ and $j = 1, \ldots, n_i$,

$$Y_{i,j} = \mu_i + \varepsilon_{i,j}, \quad \{\varepsilon_{i,j}\} \sim iN(0, \sigma^2_\varepsilon). \quad (6.1)$$

Our observations can be treated as coming from a single population. All of our response variable values can be stacked in a single vector $(Y_1, \ldots, Y_n)'$ where $n = n_1 + n_2$ and $Y_i = Y_{1,i}$ for $i = 1, \ldots, n_1$ and $Y_i = Y_{2,i-n_1}$ for $i = n_1 + 1, \ldots, n$. We also set up an additional variable $d$; this is an indicator for being in group 2, so $d_i = 0$ for $i = 1, \ldots, n_1$ and $d_i = 1$ for $i = n_1 + 1, \ldots, n$. Model (6.1) is then equivalent to

$$Y_i = \mu + \lambda d_i + \varepsilon_i, \quad \{\varepsilon_i\} \sim iN(0, \sigma^2_\varepsilon),$$

where $\mu = \mu_1$ and $\lambda = \mu_2 - \mu_1$. Testing $H_0 : \mu_1 = \mu_2$ is then equivalent to testing $H_0 : \lambda = 0$. We can generate a test statistic for this hypothesis by regressing $Y$ on the dummy variable $d$. Notice that we need one dummy variable and the constant term to completely characterise a problem with two groups.

The linear model framework described above is readily generalised to other analysis of variance problems. For example, suppose that we are interested in determining the effect on our response variable, $Y$, of a factor with $k$ levels. We could use a model

$$Y_i = \lambda_1 + \lambda_2 d_{2,i} + \ldots + \lambda_k d_{k,i} + \varepsilon_i, \quad \{\varepsilon_i\} \sim iN(0, \sigma^2_\varepsilon),$$

where $d_r$ is an indicator for the level $r$ of the factor; thus, $d_{r,i} = 1$ if the level of the factor is $r$ for observation $i$, and $d_{r,i} = 0$ otherwise. Notice that there is no $d_1$ variable. For a factor with $k$ levels we require a constant and $k - 1$ dummy variables to completely characterise the problem.
6.2 Factors

Factors are variables that have a number of different levels rather than numeric values (these levels may have numbers as labels but these are not interpreted as numeric values). For example, in the Cars93 data set the Type variable is a factor; vehicles in the study are of type "Compact", "Large", "Midsize", "Small", "Sporty" or "Van".

```r
> library(MASS)
> class(Cars93$Type)
> levels(Cars93$Type)
```

Note that Cylinders is also a factor. Some of the levels of cylinders are labeled by number but these are not interpreted as numeric values in model fitting.

```r
> class(Cars93$Cylinders)
> levels(Cars93$Cylinders)
```

Factors are used to identify groups within a data set. For example, in order to work out the mean of a variable for each type of vehicle, we would use the function `tapply(...)`. This function takes three arguments: the variable of interest, the factor (or list of factors) and the function that we would like to apply. A few examples follow.

```r
> tapply(MPG.highway, Type, mean)
> tapply(MPG.highway, Type, length)
> tapply(Price, list(Type, AirBags), mean)
```

In order to create a factor variable, we use the function `factor(...)`. This returns an object with class "factor". We can specify an order for this factor using the `levels` argument and `ordered=TRUE`. To illustrate, consider question 7 from the exercise in week 2. We have daily readings of log returns and we would like to group these by month and year (that is, Jan 2000, Feb 2000, ..., Sep 2004) and work out the standard deviation.

```r
> load("IntNas.Rdata")
> attach(intnas2)
> Monthf <- factor(Month, levels=c("Jan", "Feb", "Mar", "Apr", "May", "Jun",
> IntelLRsds <- tapply(IntelLR, list(Monthf, Year), sd, na.rm=TRUE)
> plot(as.vector(IntelLRsds),type="l")
```

6.3 Linear models using factors

The results from fitting linear models in which the explanatory variables are constructed from factors are often referred to as analysis of variance (ANOVA). Venables and Ripley, section 6.2, gives the details of the construction of explanatory variables from factors; this material is technical and is not required for the course. In most instances the R default for constructing the model matrix $X$ from factors is sensible. We will build up the complexity of our models
using examples.

### 6.3.1 Example A: One-way ANOVA

Consider the following question: does the provision of airbags affect the maximum price that people are willing to pay for a car? We can test this using a linear model with a single factor (one-way analysis of variance). The null hypothesis is that there is no effect. We start by plotting \texttt{Max.Price} against \texttt{AirBags} notice the type of plot that is generated. Try to interpret the output of the following commands.

\begin{verbatim}
> plot(AirBags,Max.Price)
> lmMaxP1 <- lm(Max.Price ~ AirBags, data=Cars93)
> summary(lmMaxP1)
> plot(lmMaxP1)
\end{verbatim}

Notice that R ensures that the columns of the model matrix are not linearly dependent by excluding one level from the linear model. The validity of analysis of variance results is dependent on constant variance within groups. We can see from the diagnostic plots that this is not entirely unreasonable for these data. We could check using the function \texttt{var.test(...)}

\begin{verbatim}
> MaxP0 <- Cars93$Max.Price[Cars93$AirBags=="None"]
> MaxP1 <- Cars93$Max.Price[Cars93$AirBags=="Driver only"]
> MaxP2 <- Cars93$Max.Price[Cars93$AirBags=="Driver & Passenger"]
> var.test(MaxP0, MaxP1)
> var.test(MaxP0, MaxP2)
> var.test(MaxP1, MaxP2)
\end{verbatim}

### 6.3.2 Example B: Two-way ANOVA

A linear model with two factors is referred to as two-way analysis of variance. We can use this technique to test the effect of two factors simultaneously. For example, we may ask whether, in addition to provision of airbags, the availability of manual transmission explains differences in maximum price.

\begin{verbatim}
> lmMaxP2 <- lm(Max.Price ~ AirBags+Man.trans.avail, data=Cars93)
> anova(lmMaxP2)
\end{verbatim}

As with multiple regression models, the sequential analysis of variance table is affected by the order in which variables are included.

\begin{verbatim}
> lmMaxP3 <- lm(Max.Price ~ Man.trans.avail+AirBags, data=Cars93)
> anova(lmMaxP3)
\end{verbatim}

We can readily include interaction terms in our model. A term of the form \texttt{a:b} is used to denote an interaction. We can include interactions and all lower order terms by the notation \texttt{a*b}. The following example illustrates.
> anova(lm(Max.Price ~ AirBags + Man.trans.avail + AirBags:Man.trans.avail, + data=Cars93))
> anova(lm(Max.Price ~ AirBags*Man.trans.avail, data=Cars93))

Note that the output of these commands is identical. What do you conclude about the interaction between provision of airbags and manual transmission in determining maximum price?

6.3.3 Example C: Factorial Design

Hines and Montgomery (1990, p.416) give the following results for an experiment in integrated circuit manufacture in which arsenic is deposited on silicon wafers. The factors are deposition time \((A)\) and arsenic flow rate \((B)\). Although both variables are quantitative, measurements are only taken at a high (labeled 1) and low level (labeled 0) of each. The purpose is to find out whether the factors have an effect and, if so, to estimate the sign and magnitude of the effect. The response is the thickness of the deposited layer.

<table>
<thead>
<tr>
<th>Treatment combination</th>
<th>(A)</th>
<th>(B)</th>
<th>Thickness</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>0</td>
<td>0</td>
<td>14.037, 14.165, 13.972, 13.907</td>
</tr>
<tr>
<td>(a)</td>
<td>1</td>
<td>0</td>
<td>14.821, 14.757, 14.843, 14.878</td>
</tr>
<tr>
<td>(b)</td>
<td>0</td>
<td>1</td>
<td>13.880, 13.860, 14.032, 13.914</td>
</tr>
<tr>
<td>(ab)</td>
<td>1</td>
<td>1</td>
<td>14.888, 14.921, 14.415, 14.932</td>
</tr>
</tbody>
</table>

The figures for thickness are stored in a single column row by row in the file \texttt{arsenic.dat}. We will create a data frame for these data and then generate variables to represent the levels of the two factors \(A\) and \(B\).

```r
> arsenic <- read.table("arsenic.dat", header=TRUE)
> A <- rep(0:1, each=4, times=2)
> B <- rep(0:1, each=8)
> A
> B
> arsenic$A <- A
> arsenic$B <- B
> rm(A,B)
> fix(arsenic)
> anova(lm(Thickness~ A*B, data=arsenic))
```

Having established from this that neither \(B\) (the flow rate) nor the interaction term are significant, we fit a model with just \(A\) the deposition time and generate some diagnostic plots.

```r
> lmArsenic <- lm(Thickness~ A, data=arsenic)
> summary(lmArsenic)
> par(mfrow=c(2,2))
> plot(lmArsenic)
> par(mfrow=c(1,1))
```
6.3.4 Example D: Analysis of covariance

In an analysis of covariance we are interested in the relationship between the response and covariates (explanatory variables in the usual sense) for different levels of a set of factors. For example, suppose we are interested in the behaviour of the relationship between fuel consumption in city driving and vehicle weight, comparing vehicles where manual transmission is available with those where it is not. We start with a plot of \texttt{MPG.city} against \texttt{Weight} with the level of \texttt{Man.trans.avail} identified for each point. The \texttt{legend(...)} function pauses while we choose a location for the legend on the plot.

\begin{verbatim}
> plot(Weight, MPG.city, pch=as.numeric(Man.trans.avail),
  + col=as.numeric(Man.trans.avail)+1)
> legend(locator(n=1), legend=c("Not available", "Available"),
  + pch=1:2, col=2:3)
\end{verbatim}

Initially we fit a model in which the slope is the same for both level of the factor.

\begin{verbatim}
> lmMPGadd <- lm(MPG.city ~ Weight+Man.trans.avail, data=Cars93)
> summary(lmMPGadd)
\end{verbatim}

This indicates that, if the same slope is fitted for each level of the factor then the difference intercepts is not significant. We can allow for an interaction between \texttt{Weight} and \texttt{Man.trans.avail}.

\begin{verbatim}
> lmMPGint <- lm(MPG.city ~ Weight*Man.trans.avail, data=Cars93)
> summary(lmMPGint)
\end{verbatim}

With the interaction term included, both the \texttt{Man.trans.avail} variable and the interaction between \texttt{Weight} and \texttt{Man.trans.avail} are significant. This indicates that both slope and intercept are significantly different when the levels of \texttt{Man.trans.avail} are taken into account.

\begin{verbatim}
> lmMPGmt0 <- lm(MPG.city ~ Weight, subset = (Man.trans.avail="No"))
> lmMPGmt1 <- lm(MPG.city~ Weight, subset = (Man.trans.avail="Yes"))
> summary(lmMPGmt0)
> summary(lmMPGmt1)
> abline(lmMPGmt0, col=2)
> abline(lmMPGmt1, col=3)
\end{verbatim}

Note that the difference between the slope estimates of these two models is precisely the interaction term from \texttt{lmMPGint}.

We can fit these models simultaneously using a model formula of the form \texttt{a/x -1} where \texttt{a} is a factor. This will fit the separate regression models. The -1 is to remove the overall intercept term that we do not want.

\begin{verbatim}
> lmMPGsep <- lm(MPG.city ~ Man.trans.avail/Weight-1, data=Cars93)
> summary(lmMPGsep)
\end{verbatim}
6.4 Exercise

1. This question illustrates the importance of including marginal terms. Fit quadratic regression models for \( \text{MPG.highway} \) with \texttt{Horsepower} as an explanatory variable. First of all just with the quadratic term and then with the marginal term included. The commands are:

\[
\text{lmMPG1} \leftarrow \text{lm}(\text{MPG.highway} \sim \text{I}(\text{Horsepower})^2, \text{data}=\text{Cars93})
\]

\[
\text{lmMPG2} \leftarrow \text{update}(\text{lmMPG1}, \sim \text{Horsepower} + .)
\]

Plot \( \text{MPG.highway} \) against \texttt{Horsepower} and add the fitted lines generated by the two regressions. Which do you think is more appropriate.

2. We wish to determine whether the log price is affected by the origin of the vehicle. Why might we consider performing a log transformation? Create a variable \texttt{logPrice} by taking logs of the \texttt{Price} variable.

(a) Carry out one-way analysis of variance with \texttt{Origin} as a factor. What conclusions do you draw?

(b) Now include \texttt{Type} in the model. Can you explain the resulting changes?

(c) Is there a significant interaction between price and origin? [You will need to fit another model to answer this question.]

(d) Draw a plot of \texttt{logPrice} against \texttt{as.numeric(Type)} with the colour of each point determined by its origin. [You will need to use \texttt{col = as.numeric(Origin)}.] Create a legend for this plot. Does this help to explain the results of our models?

3. Write a function that takes two variables and the data frame that contains these variables as arguments. The function should perform the quadratic regression of first variable on the second and generate a plot of the response against the explanatory variable with the fitted curve added to the plot.

4. † Consider the zero-mean auto-regressive model of order 1, AR(1):

\[
w_t = \phi w_{t-1} + \varepsilon_t, \quad \{\varepsilon_t\} \sim iN(0, 1).
\]

Write a function to simulate a series of \( n \) values of an AR(1). Now consider a simple regression model with time as the explanatory variable and autoregressive errors:

\[
Y_t = \alpha + \beta t + w_t, \quad \{w_t\} \sim AR(1).
\]

Write a function that simulates \( n \) values from this regression model and returns \( \hat{\beta} \). Write a function that simulates \( r \) instances of \( \hat{\beta} \) and use it to investigate the MSE of the associated estimator. How does this compare with (analytic) results for the case of uncorrelated errors.

6.5 Reading

6.5.1 Directly related reading

- Venables, W. N. \\ et. al. (2001) \textit{An Introduction to R}. [Chapter 4 for stuff on factors]

6.5.2 Supplementary reading†


• Dalgaard, P. (2002) *Introductory Statistics with R*, Springer. [Section 10.3 dummy variables, 10.5 interaction, 10.6 two-way ANOVA and 10.7 analysis of covariance - all R no theory]