Categorical Predictors

We have dealt so far with explanatory variables that are quantitative. If \( y \) is a response variable, we use the method of least squares to fit models of form

\[
y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon
\]

We now turn to the case where the explanatory variable is categorical. (Such explanatory variables are often called factors and the possible values of a factor are called levels.) Categorical explanatory variables often appear in the case of randomized comparative experiments.

For example, consider the chickwts data that are an internal \( \text{R} \) dataset. In this experiment, baby chicks are divided at random into six groups and each group fed a different feed. After six weeks, the chicks are weighed. We want to see whether some feeds perform better than others.

![Weight graph](image)

Obviously, there is considerable variation in each group of chicks. That variation is due to the normal differences between individual chicks and to variables that the experimenter was unable to control. This variation within groups makes it more difficult to see if there is a systematic difference between groups – say between the meatmeal and soybean group.

Our model in this situation is that each of the six treatments has a "true" or "population" mean and that sampled chicks will have weights varying around that mean. Our goal is to decide whether these means are different and how different they are. For example, we want to test the hypothesis that

\[
H_0 : \mu_{\text{casein}} = \mu_{\text{horsebean}} = \mu_{\text{linseed}} = \mu_{\text{meatmeal}} = \mu_{\text{soybean}} = \mu_{\text{sunflower}} = 0
\]

We formulate this model as follows. We have \( I \) groups and so we have means \( \mu_1, \mu_2, \ldots, \mu_I \). We write \( y_{ij} \) for the \( j^{\text{th}} \) observation in the \( i^{\text{th}} \) group and we write \( n_i \) for the number of observations in that group. We still use \( n \) for the total number of observations. Our model is

\[
y = \mu_i + \varepsilon \quad \text{for observations } y \text{ in the } i^{\text{th}} \text{ group}
\]

We have \( I \) parameters to fit and we fit these by minimizing the sums of the squares of the residuals. It turns out that \( \hat{\mu}_i \) is just the mean of the observations in the \( i^{\text{th}} \) group.

We can use \texttt{lm} to do the work here. When we do, the model looks slightly different. We first consider the case of \( I = 2 \).
One factor with two categories

The case of a categorical explanatory variable with just two levels arises in three different natural scenarios.

1. We perform an experiment in which we assign subjects to two different experimental treatments at random. 
   An example of this was the helicopter project. You flew two different models of helicopters and you wanted to know whether the average flight time of the two models was different.

2. We have two different populations and we randomly sample from each to determine if the means of the two populations are different.
   For example, we have bolts produced by two different machines and we want to see whether their average lengths are different.

3. We randomly sample from one population but divide the population into two groups based on a categorical variable.
   For example, we randomly sample Calvin students and want to determine the average height difference between males and females.

The model for each of these situations is that $y = \mu_i + \varepsilon$. We are going to revise that model slightly to make it look even more like the linear models we have considered before and to make it more suitable for \texttt{lm}.

In order to do this we need the notion of an \textit{indicator} variable. An indicator variable $x$ is a variable that takes on just two values, 0 or 1. We can use an indicator variable to code a categorical variable – we use $x = 1$ to code one group and $x = 0$ the other.

We can now write the linear model as follows:

$$ y = \beta_0 + \beta_1 x + \varepsilon $$

Here we see that $\beta_0$ is the mean of $y$ when $x$ is 0 (the first group), and $\beta_0 + \beta_1$ is the mean of $y$ when $x$ is 1 (the second group). Therefore $\beta_1$ is the difference in the means of the true groups. This means that $\beta_1$ is the parameter that we are usually interested in. (We are usually interested in whether there is a real difference in the means of the two groups.)

Hot rubberbands

Refer to the dataframe \texttt{rubberbands} in the M241 package. 21 rubberbands were randomly assigned to two groups. One group (A) was left at the ambient temperature. The other group (H) was placed in hot water for four minutes. After a ten minute wait, the amounts of stretch under a 1.35 kg weight were recorded.

\begin{verbatim}
stretch    quantitative, amount of stretch
 treat     categorical, treatment A or H
\end{verbatim}

A linear model for this data is: $\text{stretch} = \beta_0 + \beta_1 \text{treatH} + \varepsilon$

Here $\text{treatH}$ is a variable that has value 1 if the treatment is H and 0 if the treatment is not H (i.e., is A). (The choice of the indicator variable is arbitrary. We could just as well have used a variable that is 1 if the treatment is A.)

We fit the model using \texttt{lm} just as before.
\begin{verbatim}
> lrb=lm(stretch~treat,data=rubberbands)
> summary(lrb)
Call:
  lm(formula = stretch ~ treat, data = rubberbands)

Residuals:
       Min       1Q   Median       3Q      Max
-27.090  -4.500    0.500    7.910   15.500

Coefficients:
             Estimate Std. Error     t value  Pr(>|t|)
(Intercept) 244.08876   3.28752   74.16870 < 2e-16 ***
treatH       9.41143   4.76930    1.96963    0.063 .

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.9 on 19 degrees of freedom
Multiple R-squared: 0.17,   Adjusted R-squared: 0.126
F-statistic: 3.89 on 1 and 19 DF,  p-value: 0.0632
\end{verbatim}

Note that R decides to use treatH for the treatment variable. That means that R has chosen that this variable is 1 only if the treatment is H. R chooses the indicator variable based on the order of the levels in the vector treat. In fitting the model, we simply name the categorical variable.

So the fitted model is

\[
\hat{\text{stretch}} = 244.09 \quad \text{if treatment is A} \\
\hat{\text{stretch}} = 244.09 + 9.41 \quad \text{if treatment is H}
\]

Therefore the estimate of the difference in means between the two treatments is 9.41. There isn’t strong evidence to believe that this difference isn’t 0 ($P = 0.063$). Another way of seeing that there isn’t strong evidence for a treatment effect is to do an analysis of variance. The function \texttt{anova} with just one argument compares the given model to the model that doesn’t have the explanatory variable in it.

\begin{verbatim}
> lr=lm(stretch~1,data=rubberbands)
> anova(lr,lrb)
Analysis of Variance Table

  Model 1: stretch ~ 1
  Model 2: stretch ~ treat
             Res.Df RSS Df Sum of Sq F Pr(>F)
Model 2:  1   20 2727
Model 1:  2   19 2263 1  464 3.89 0.063 .

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
\end{verbatim}

\begin{verbatim}
> anova(lrb)
Analysis of Variance Table

Response: stretch
             Df Sum Sq Mean Sq F value  Pr(>F)
Model 1:  1   464  464 3.89  0.063 .
Model 2: 19  2263 119

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
\end{verbatim}