3 Regression using Dummy Variables

In all the examples we have seen to date, we have had a quantitative response variable, necessarily due to the normality assumption, but our explanatory variables have been quantitative (continuous) measurements also. In this chapter we consider (multiple) linear regression, where one or more explanatory variables are binary (one which can take only two values, say 0 or 1).

For example, the `Credit` data set displayed in Figure 15 records `balance` (average credit card debt for a number of individuals) as well as several quantitative predictors: `age`, `cards` (number of credit cards), `education` (years of education), `income` (in thousands of dollars), `limit` (credit limit), and `rating` (credit rating). Each panel of Figure 15 is a scatterplot for a pair of variables whose identities are given by the corresponding row and column labels. For example, the scatterplot directly to the right of the word "Balance" depicts `balance` versus `age`, while the plot directly to the right of "Age" corresponds to `age` versus `cards`. In addition to these quantitative variables, we also have four qualitative variables: `gender`, `student` (student status), `status` (marital status), and `ethnicity` (Caucasian, African American or Asian).

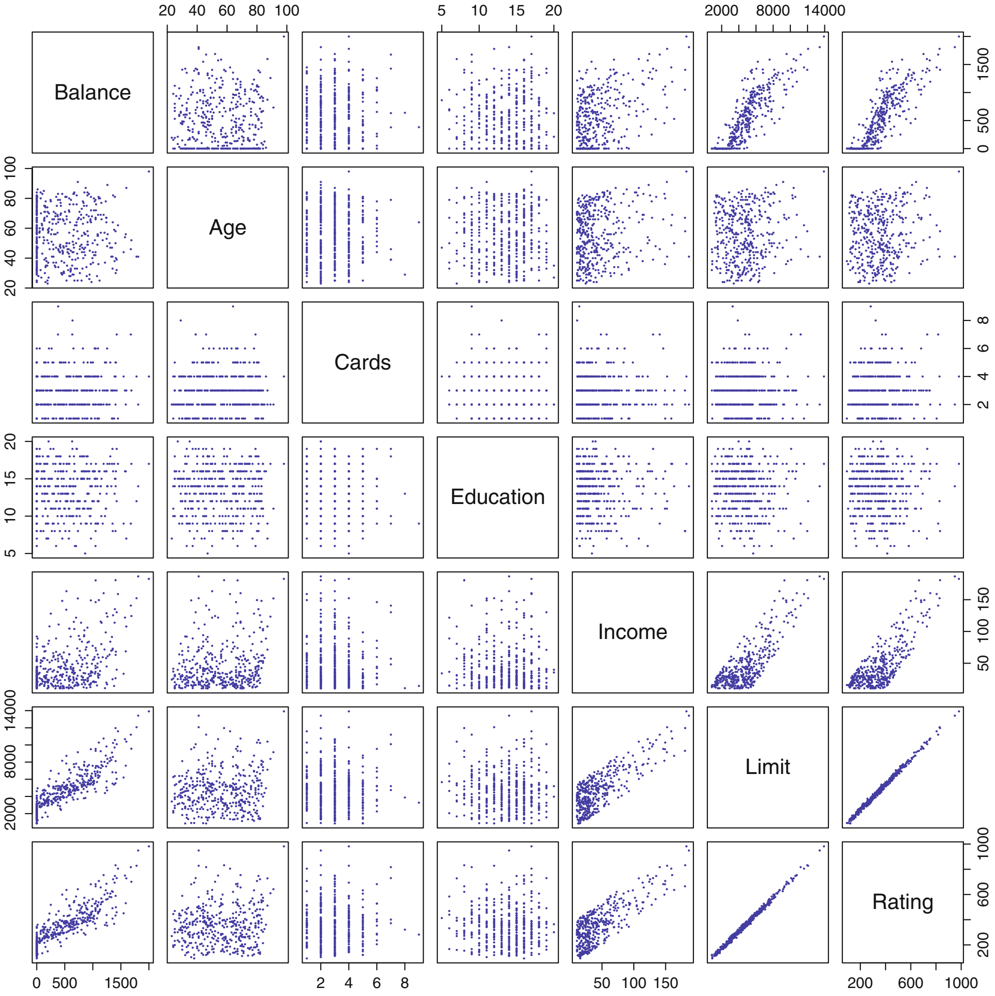

Figure 15: The `Credit` data set contains information about `balance`, `age`, `cards`, `education`, `income`, `limit`, and `rating` for a number of potential customers.

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`gender[Female]`	19.73	46.05	0.429	0.6690

Table 10: Least squares coefficient estimates associated with the regression of `balance` onto `gender` in the `Credit` data set.

3.1 Predictors with Only Two Levels

Suppose that we wish to investigate differences in credit card balance between males and females, ignoring the other variables for the moment. If a qualitative predictor (also known as a factor) only has two levels, or possible values, then incorporating it into a regression model is very simple. We simply create an indicator or dummy variable that takes on two possible numerical values. For example, based on the gender variable, we can create a variable that takes the form

$$x_i = \begin{cases} 1, & \text{unit } i \text{ is female} \\ 0, & \text{unit } i \text{ is male} \end{cases}$$

and use this variable as a predictor in the regression equation. This results in the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \text{unit } i \text{ is female} \\ \beta_0 + \epsilon_i, & \text{unit } i \text{ is male.} \end{cases}$$

Now β_0 can be interpreted as the average credit card balance among males, $\beta_0 + \beta_1$ as the average credit card balance among females, and β_1 as the average difference in credit card balance between females and males.

Table 10 displays the coefficient estimates and other information associated with the two models above. The average credit card debt for males is estimated to be \$509.80, whereas females are estimated to carry \$19.73 in additional debt for a total of \$509.80 + \$19.73 = \$529.53. However, we notice that the p-value for the dummy variable is very high. This indicates that there is no statistical evidence of a difference in average credit card balance between the genders.

The decision to code females as 1 and males as 0 is arbitrary, and has no effect on the regression fit, but does alter the interpretation of the coefficients. If we had coded males as 1 and females as 0, then the estimates for β_0 and β_1 would have been 529.53 and -19.73, respectively, leading once again to a prediction of credit card debt of \$529.53 - \$19.73 = \$509.80 for males and a prediction of \$529.53 for females.

3.2 Qualitative Predictors with More than Two Levels

When a qualitative predictor has more than two levels, a single dummy variable cannot represent all possible values. In this situation, we can create additional dummy variables. For example, for the `ethnicity` variable we create two dummy variables. The first could be

$$x_{i1} = \begin{cases} 1, & \text{unit } i \text{ is Asian} \\ 0, & \text{unit } i \text{ is not Asian,} \end{cases}$$

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	531.00	46.32	11.464	< 0.0001
`ethnicity[Asian]`	-18.69	65.02	-0.287	0.7740
`ethnicity[Caucasian]`	-12.50	56.68	-0.221	0.8260

Table 11: Least squares coefficient estimates associated with the regression of balance onto ethnicity in the `Credit` data set. `ethnicity` is encoded via two dummy variables.

and the second could be

$$x_{i2} = \begin{cases} 1, & \text{unit } i \text{ is Caucasian} \\ 0, & \text{unit } i \text{ is not Caucasian.} \end{cases}$$

Then both of these variables can be used in the regression equation, in order to obtain the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i + \epsilon_i, & \text{unit } i \text{ is Asian} \\ \beta_0 + \beta_2 + \epsilon_i, & \text{unit } i \text{ is Caucasian} \\ \beta_0 + \epsilon_i, & \text{unit } i \text{ is African American.} \end{cases}$$

Now β_0 can be interpreted as the average credit card balance for African Americans, β_1 can be interpreted as the difference in the average balance between the Asian and African American categories, and β_2 can be interpreted as the difference in the average balance between the Caucasian and African American categories. There will always be one fewer dummy variable than the number of levels. The level with no dummy variable - African American in this example - is known as the baseline.

From Table 11, we see that the estimated balance for the baseline, African American, is \$531.00. It is estimated that the Asian category will have \$18.69 less debt than the African American category, and that the Caucasian category will have \$12.50 less debt than the African American category. However, the p-values associated with the coefficient estimates for the two dummy variables are very large, suggesting no statistical evidence of a real difference in credit card balance between the ethnicities. Once again, the level selected as the baseline category is arbitrary, and the final predictions for each group will be the same regardless of this choice. However, the coefficients and their p-values do depend on the choice of dummy variable coding. Rather than rely on the individual coefficients, we can use an F-test to test $H_0 : \beta_1 = \beta_2 = 0$; this does not depend on the coding. This F-test has a p-value of 0.96, indicating that we cannot reject the null hypothesis that there is no relationship between `balance` and `ethnicity`. Using this dummy variable approach presents no difficulties when incorporating both quantitative and qualitative predictors. For example, to regress `balance` on both a quantitative variable such as `income` and a qualitative variable such as `student`, we must simply create a dummy variable for student and then fit a multiple regression model using income and the dummy variable as predictors for credit card balance.

There are many different ways of coding qualitative variables besides the dummy variable approach taken here. All of these approaches lead to equivalent model fits, but the coefficients are different and have different interpretations, and are designed to measure particular contrasts. This topic is beyond the scope of the book, and so we will not pursue it further.

3.3 Quantitative Predictor and Qualitative Predictor with Two Levels

Consider the following model:

$$y_i = \beta_0 + \beta x_i + \gamma z_i + \epsilon_i, \tag{23}$$

Note that (x_i, y_i) are the usual observations of the ith individual on the explanatory and response variables x and y respectively. The variable z_i is a dummy variable, which takes the following values

$$z_i = \begin{cases} 0, & \text{unit } i \text{ belongs to group A} \\ 1, & \text{unit } i \text{ belongs to group B} \end{cases}$$

The model (23) can hence be separated into two models, according to the group membership (or category) of i. For group/category A ($z_i = 0$), we have

$$y_i = \beta_0 + \beta x_i + \epsilon_i$$

and, for group/category B ($z_i = 1$)

$$y_i = (\beta_0 + \gamma) + \beta x_i + \epsilon_i$$

Hence, the parameter γ represents the difference in intercept value between groups/categories A and B. That is, model (23) has separate intercepts (but common slope) for the two groups, and this is achieved by adding the qualitative explanatory variable (or factor) z_i to the model, which has two categories relating to the group membership, A ($z_i = 0$) or B ($z_i = 1$). If the term γ is found to be significant in the model, then the need for a separate intercept for each group is evidenced.

Model (23) above allows for separate intercepts between groups, but an obvious extension allows us to include separate (i.e. *non-parallel*) slopes for each group.

$$y_i = \beta_0 + \beta x_i + \gamma z_i + \delta x_i z_i + \epsilon_i \tag{24}$$

where again z_i is a 0/1 indicator variable representing group membership. In this setting we have, for group A ($z_i = 0$)

$$y_i = \beta_0 + \beta x_i + \epsilon_i$$

and for group B ($z_i = 1$)

$$y_i = (\beta_0 + \gamma) + (\beta + \delta) x_i + \epsilon_i$$

i.e. we have simple linear regressions of y on x for each of the two groups. In this case the regressor $x_i z_i$ represents the *interaction* term between x_i and z_i. We say that two variables *interact* if the partial effect of one depends on the value of the other. In this case the parameter δ represents the difference in slopes between group A (control, $z_i = 0$) and group B in the combined regression model.

3.4 Quantitative Predictor and Qualitative Predictor with more than Two Levels

The process described above of using an indicator (or dummy) variable to represent group membership (or factor level) is easily extended to situations where a factor has more than two levels. Let us first of all consider the case of *three* levels. The separate intercepts regression model may now be written

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{i1} + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \gamma_3 z_{3i} + \epsilon_i \quad i = 1, \ldots, n, \\
&= (\beta_0 + \gamma_j) + \beta_1 x_{i1} + \epsilon_i \quad , \; j = 1, 2, 3. \qquad (25)
\end{aligned}
$$

where z_{ji}, $j = 1, 2, 3$ is the dummy regressor indicating whether an observation belongs to the jth category or factor level.

$$
z_{ji} = \begin{cases} 1 & \text{if unit } i \text{ belongs to category } j \\ 0 & \text{otherwise} \end{cases}
$$

One immediate problem is that this model is over-parameterized, with four parameters (β_0, γ_1, γ_2 and γ_3) to represent three group intercepts, and we would not be able to find unique estimates for these four parameters. This is because the γ_j parameters are associated with the dummy variables z_1, z_2 and z_3 which are perfectly collinear, e.g. $z_3 = 1 - z_1 - z_2$. To accommodate this, we rely on only two dummy variables and select one of the categories to be considered as a baseline.

In general, for a polytomous factor with m categories, we code $m - 1$ dummy regressors, selecting the *first* category as the baseline and code $z_{ij} = 1$ when observation i falls into category j, and 0 otherwise.

Category	z_2	z_3	z_4	$\ldots$	z_m
1	0	0	0	$\ldots$	0
2	1	0	0	$\ldots$	0
3	0	1	0	$\ldots$	0
4	0	0	1	$\ldots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
m	0	0	0	$\ldots$	1

In this way a polytomous factor can be entered into a (multiple) regression by simply coding a set of 0/1 dummy variables, one fewer than the number of categories for that factor. The 'omitted' category, coded 0 for all the dummy regressors in the set, serves as a baseline to which all other categories are compared. Such a model represents parallel regression lines (surfaces), one for each category of factor.

Interactions can be incorporated into the model by simply forming the product terms between the relevant explanatory variable and each of the $m - 1$ dummy variables representing the m-level factor. (The interaction between two factors of k levels and ℓ levels respectively will thus be modelled by the entering of $(k - 1)(\ell - 1)$ terms in the regression model).

3.4.1 The Principle of Marginality

An important consideration, when dealing with models containing interactions is that of *marginality.* In general, we do not test or interpret the main effects of explanatory variables that interact. As a corollary to this principle, it does not generally make sense to specify and fit models that include interaction terms (regressors), but omit main effects terms that are marginal to them. That is, the *principle of marginality* specifies that a model including a higher order term (such as an interaction) should normally also include the lower order relatives (i.e. the main effects that compose the interaction).

3.4.2 Inference for Main Effects and Interactions

We use the principle of marginality as a guide to constructing incremental F-tests for the terms in a model that includes interactions.

For each of the examples we have considered (i.e. the house value and the iris data), the following *four* models can be compared, where $j = 1, \ldots, m-1$ represents the factor category:

Model		Description
	$y_i = \beta_0 + \epsilon_i$	constant
(A)	$y_i = \beta_0 + \beta x_i + \epsilon_i$	single line
(B)	$y_i = (\beta_0 + \gamma_j) + \beta x_i + \epsilon_i$	parallel lines
(C)	$y_i = (\beta_0 + \gamma_j) + (\beta + \delta_j) x_i + \epsilon_i$	separate lines

A *sequential* ANOVA table can be constructed to test between such models hierarchically. Such a table can be produced directly using the function `anova()` in `R`.

3.5 Lab: Regression Using Dummy Variables

3.5.1 Regression Using Dummy Variables I

Consider the following dataset:

```
   Value Size Fireplace
1  234.4 2.00       Yes
2  227.4 1.71        No
3  225.7 1.45        No
4  235.9 1.76       Yes
5  229.1 1.93        No
6  220.4 1.20       Yes
7  225.8 1.55       Yes
8  235.9 1.93       Yes
9  228.5 1.59       Yes
10 229.2 1.50       Yes
11 236.7 1.90       Yes
12 229.3 1.39       Yes
13 224.5 1.54        No
14 233.8 1.89       Yes
15 226.8 1.59        No
```

which contains three variables; the response variable house value and two regressors size and fireplace. The data can be plotted as follows:

```
house <-read.table("house.txt", header = TRUE)
house

plot(house$Size, house$Value, type = "n", main = "House value vs Size",
     xlab = "Size", ylab = "Value")
points(house$Size[house$Fireplace=="No"], house$Value[house$Fireplace=="No"],
       pch=1)
points(house$Size[house$Fireplace=="Yes"], house$Value[house$Fireplace=="Yes"],
       pch=8)

legend(1.2, 237, c("Yes", "No"), pch =c(8,1))
```

Notice from the plot that the observations can be split into two, depending on whether the house has (‘Yes’) or has not (‘No’) a fireplace. There does appear to be a clear (possibly linear) relationship between house values and size. Notice also that the house value is generally higher for houses with a fireplace than for houses with no fireplace.

Model A: Fitting a simple linear regression

```
house.lmA <- lm(Value ~ Size, data = house)
summary(house.lmA)
```

Plots showing the fitted line and the standardized residuals against fitted values can be produced.

```
par(mfrow = c(1, 2))
plot(house$Size, house$Value, type = "n", main = "House value vs Size: Model A",
     xlab = "Size", ylab = "Value")
points(house$Size[house$Fireplace == "No"], house$Value[house$Fireplace == "No"], pch = 1)
points(house$Size[house$Fireplace == "Yes"], house$Value[house$Fireplace == "Yes"], pch = 8)
legend(1.2, 237, c("Yes", "No"), pch =c(8,1), bty="n")
abline(house.lmA)

sresA <- rstandard(house.lmA)
fitsA <- fitted(house.lmA)

plot(fitsA, sresA, type = "n", main = "Stand. Residuals vs Fitted Values",
     xlab = "Fitted values", ylab = "Standardized residuals")
points(fitsA[house$Fireplace == "No"], sresA[house$Fireplace == "No"], pch = 1)
points(fitsA[house$Fireplace == "Yes"], sresA[house$Fireplace == "Yes"], pch = 8)
```

Notice from the residual plot that the model does not appear to adequately fit both groups of house (with and without fireplace). Since houses with fireplacee are likely to have greater value, so their residuals are more likely to be positive than those of houses without fireplace. That is, on average, this simple linear regression model appears to underestimate the value for houses with fireplace, whilst overestimating that for houses without fireplace.

A natural question to ask is whether there are actually *separate* regression relationships (between `Value` and `Size`) for houses with and without fireplace. If so, is there a means by which we can test for the need for separate intercept and slope parameters.

We can achieve this by including the 'dummy variable' `Fireplace` in the model which takes the value 1 if the house has a fireplace (`Fireplace="Yes"`) and 0 if the house does not have a fireplace (`Fireplace="No"`).

Models B and C represent models with separate intercepts (model (23)) and non-parallel slopes (model (24)) respectively.

Model B

```
house.lmB <- lm(Value ~ Size + Fireplace, data = house)
summary(house.lmB)
```

Again the fitted regression lines and (standardized) residual plots can be assessed:

```
par(mfrow = c(1, 2))
plot(house$Size, house$Value, type = "n", main = "House value vs Size: Model B",
     xlab = "Size", ylab = "Value")
points(house$Size[house$Fireplace == "No"], house$Value[house$Fireplace == "No"], pch = 1)
points(house$Size[house$Fireplace == "Yes"], house$Value[house$Fireplace == "Yes"], pch = 8)
```

```
legend(1.2, 237, c("Yes", "No"), pch =c(8,1), bty="n")

ld <- seq(min(house$Size), max(house$Size), 0.1)
lines(ld, predict(house.lmB, data.frame(Size = ld, Fireplace = rep("Yes", length(ld))),
                    type = "response"))
lines(ld, predict(house.lmB, data.frame(Size = ld, Fireplace = rep("No", length(ld))),
                    type = "response"))

sresB <- rstandard(house.lmB)
fitsB <- fitted(house.lmB)
plot(fitsB, sresB, type = "n", main = "Stand. Residuals vs Fitted Values")
points(fitsB[house$Fireplace == "No"], sresB[house$Fireplace == "No"], pch = 1)
points(fitsB[house$Fireplace == "Yes"], sresB[house$Fireplace == "Yes"], pch = 8)
```

Here the model appears to be more appropriate, with house values predicted to be higher for houses with fireplace, for a similar level of size. The residual plot is much more acceptable with standardized residuals for houses with and without fireplace evenly spread about zero. The better fit of the model is confirmed by the R^2 values from the output. Model B gives a multiple R^2 value of 0.7796 which is better than the value of 0.6331 returned by model A, which also has a higher value of model standard deviation (2.919 versus 2.263). The coefficient of Fireplace in Model B is also highly significant, indicating that the intercepts between the two groups of house are different.

Finally, we consider whether there is a case for separate slope parameters too (non-parallel regression lines). Note that the interaction term is specified in R by using Size*Fireplace to the model formula.

Model C

```
house.lmC <- lm(Value ~ Size * Fireplace, data = house)
summary(house.lmC)
```

The output shows that there is no evidence for non-parallel lines, since the coefficient corresponding to the interaction term Size:FireplaceYes is not significant (P-vale=0.166), so that Model B is preferred.

A sequential ANOVA table can be constructed to test between such models hierarchically. Such a table can be produced directly using the function anova() in R as follows:

```
anova(house.lmC)
```

Recall that this tests the change in the regression sums of squares, sequentially, for the order in which terms are added to the model. We select a parsimonious model by stopping at the first stage beyond which there are no significant F-ratios, confirming our earlier results.

3.5.2 Regression Using Dummy Variables II

We will now examine the `Carseats` data, which is part of the `ISLR` library. We will attempt to predict `Sales` (child car seat sales) in 400 locations based on a number of predictors.

```
library(ISLR)
fix(Carseats)
names(Carseats)
```

The `Carseats` data includes qualitative predictors such as `Shelveloc`, an indicator of the quality of the shelving location - that is, the space within a store in which the car seat is displayed - at each location. The predictor `Shelveloc` takes on three possible values, Bad, Medium, and Good. Given a qualitative variable such as `Shelveloc`, `R` generates dummy variables automatically. Below we fit a multiple regression model.

```
lm.fit <- lm(Sales ~., data = Carseats)
summary(lm.fit)
```

Use `?contrasts` to learn about other contrasts, and how to set them. `R` has created a `ShelveLocGood` dummy variable that takes on a value of 1 if the shelving location is good, and 0 otherwise. It has also created a `ShelveLocMedium` dummy variable that equals 1 if the shelving location is medium, and 0 otherwise. A bad shelving location corresponds to a zero for each of the two dummy variables. The fact that the coefficient for `ShelveLocGood` in the regression output is positive indicates that a good shelving location is associated with high sales (relative to a bad location). And `ShelveLocMedium` has a smaller positive coefficient, indicating that a medium shelving location leads to higher sales than a bad shelving location but lower sales than a good shelving location.