Chapter 11

Basic Multivariate Regression

This chapter extends the bivariate model to a multivariate model, i.e., the case with more than one independent variable. There is also a YouTube Video associated with this section:

- Multivariate Regression with R - Video

The following topics associated with multivariate regression models are covered in this chapter:

- Dummy variables
- Natural logarithm
- Functional forms
- Interaction Terms
- Multicollinearity

11.1 Introduction

Recall the bivariate regression model with one independent and one dependent variable:

$$y = \beta_0 + \beta_1 \cdot x_1 + \epsilon$$

The multivariate linear regression model includes more than one independent variable and is simply an extension of the bivariate regression model:

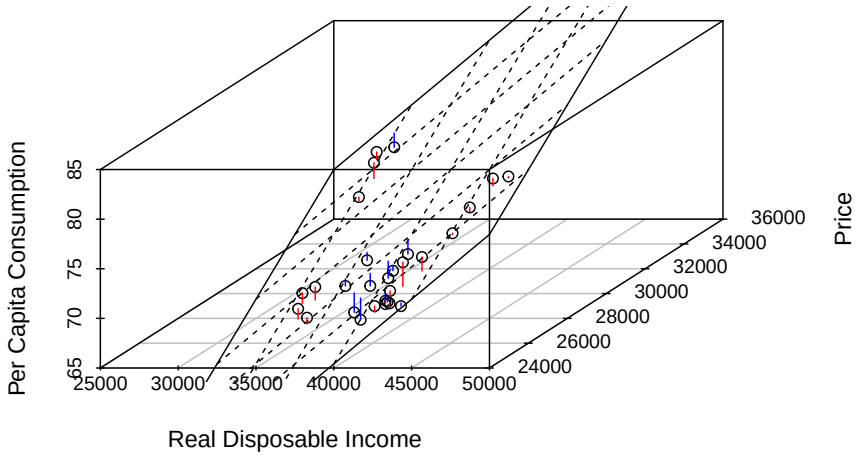
$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k + \epsilon$$

Whether we consider the bivariate or multivariate model, the objective is always to minimize the sum of squared errors which has led to the name ordinary least

square (OLS) model. The equation of a line can be determined using slope (β_0) and the intercept ($\beta 1$), i.e.:

$$E(y|x_1) = \beta_0 + \beta_1 \cdot x_1$$

The case of a regression model with two independent variable can still be represented in a 3-dimensional graph as depicted below



The purpose of the multivariate regression model is to measure the effect of independent variables on the dependent variable. It is crucial to control for everything else that could influence the dependent variable. For example, measuring the weekly grocery bill as a function of years of education might give you a statistically significant effect for education but if income is included, the effect for education might (most likely) disappear.

The first example involves estimating home values based on square footage and number of garage spots of a house in the 46268 ZIP code in Indianapolis. The data is contained in `indyhomes`.

```
indyhomes46268 = subset(indyhomes,zip==46268)
bhat = lm(price~sqft+garage,data=indyhomes46268)
summary(bhat)
```

```
##
## Call:
## lm(formula = price ~ sqft + garage, data = indyhomes46268)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -58780  -7817   1582   7886  51803
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 81733.141  15896.004   5.142 5.20e-06 ***
## sqft           40.897      4.383   9.331 2.85e-12 ***
## garage      16580.964   7136.866   2.323   0.0245 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20710 on 47 degrees of freedom
## Multiple R-squared:  0.675,  Adjusted R-squared:  0.6611
## F-statistic:  48.8 on 2 and 47 DF,  p-value: 3.388e-12
```

Depending on the nature of the variables, it might be necessary to scale your variables for ease of interpretation. This might be necessary if coefficients are very large or very small. A rescaling, e.g., dividing income by 1000, does affect the coefficients and the standard errors but has no effect on the t-statistics.

11.2 Dummy Variables

So far, independent variables were quantitative such as price, income, square footage, miles, and so on. But very often, a qualitative characteristic such as religion or gender must be modeled. For this purpose, dummy variables that can be either 0 or 1 are used. Dummy variables represent a single qualitative characteristic. For example, consider the price (y_i) of a car depending on miles (x_i) and whether the car has all-wheel drive (AWD) or rear-wheel drive (RWD). This characteristic can be modeled using a dummy variable (d_i). If $d_i = 1$, the car has AWD and if $d_1 = 0$, the car has RWD. The regression equation can be written as follows:

$$y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot d_i + \epsilon_i$$

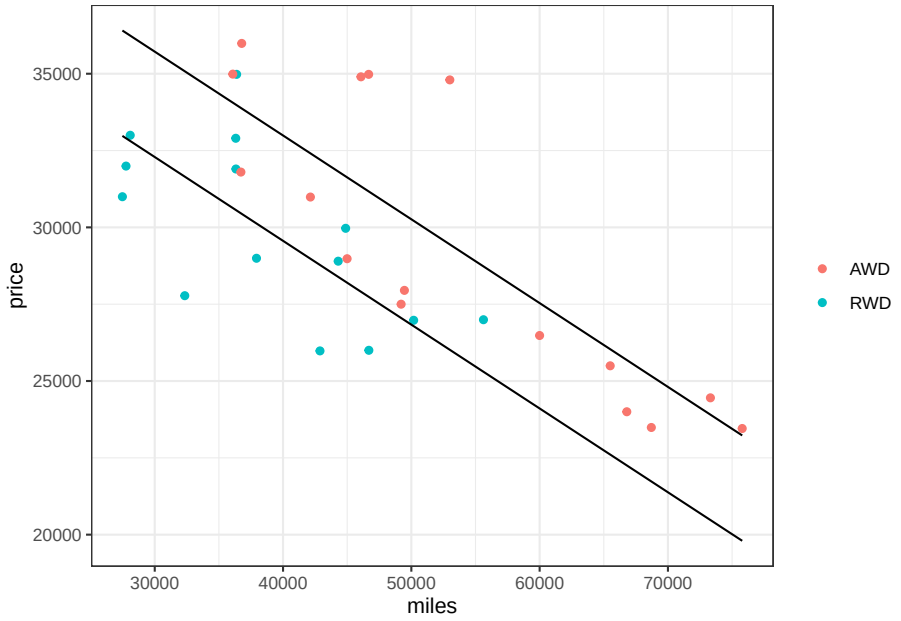
his regression can theoretically be separated into two single equations:

- RWD: $y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$
- AWD: $y_i = (\beta_0 + \beta_2) + \beta_1 \cdot x_i + \epsilon_i$

To interpret the dummy variables, it is necessary to know how it is coded. In the above case, if the coefficient β_2 is positive, then the dummay variable adds to the price. That is, the coefficient β_2 represents the value of all-wheel drive.

```
##
## Call:
## lm(formula = price ~ miles + allwheeldrive, data = bmw)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3874.1 -1724.0  -176.5  1604.5  5355.0
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.047e+04  1.711e+03  23.660  < 2e-16 ***
## miles         -2.728e-01  4.044e-02  -6.745 3.05e-07 ***
## allwheeldrive  3.429e+03  1.063e+03   3.227  0.00327 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2449 on 27 degrees of freedom
## Multiple R-squared:  0.6287, Adjusted R-squared:  0.6012
## F-statistic: 22.86 on 2 and 27 DF,  p-value: 1.553e-06
```



11.3 Natural Logarithm

Transforming the dependent and/or independent variables using the natural logarithm has some important and useful interpretations. Consider the following simple consumption equation in which both variables are in logarithmic form:

$$ln(consumption) = \beta_0 + \beta_1 \cdot \ln(income) + \epsilon$$

In this case, β_1 is the elasticity of consumption with respect to income, i.e., a 1% increase in income leads to a $\beta_1 \cdot 1\%$ increase in consumption. For example,

if $\beta_1 = 0.4$, then a 1% increase in income will rise consumption by 0.4%. Note that the percentage increase is only an approximation for small changes.

Dep. Var.	Indep. Var	Interpretation
y	x	1 dollar change in x changes y by $\hat{\beta}$ dollars
$\ln(y)$	x	1 dollar change in x changes y by 100 $\times\hat{\beta}$ percent
$\ln(y)$	$\ln(x)$	1 percent change in x changes y by $\hat{\beta}$ percent
y	$\ln(x)$	1 percent change in x changes y by $\hat{\beta}/100$ dollars

```
bhat = lm(log(total)~yards+att+exp+draft1+veteran+changeteam+pbowlever,data=nfl)
summary(bhat)
```

```
##
## Call:
## lm(formula = log(total) ~ yards + att + exp + draft1 + veteran +
##     changeteam + pbowlever, data = nfl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4899 -0.4998 -0.0801  0.4554  3.1959
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.9289322  0.0767846 -12.098  < 2e-16 ***
## yards        0.0003566  0.0001411   2.527 0.011783 *
## att          0.0003927  0.0009657   0.407 0.684408
## exp          0.0108812  0.0160213   0.679 0.497312
## draft1       0.8876564  0.1132374   7.839 2.30e-14 ***
## veteran      0.6735244  0.1144567   5.885 6.88e-09 ***
## changeteam  -0.3095919  0.0893125  -3.466 0.000568 ***
## pbowlever    0.4093324  0.0936078   4.373 1.46e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7909 on 560 degrees of freedom
##   (441 observations deleted due to missingness)
## Multiple R-squared:   0.55,  Adjusted R-squared:  0.5444
## F-statistic:  97.8 on 7 and 560 DF,  p-value: < 2.2e-16
```

11.4 Functional Form

To model non-linear relationships, an independent variable can be transformed by squaring it. For example, consider the relationship between income and food

expenditure. The regular OLS assumes a linear relationship in the sense that an increase in income always leads to a proportional increase in food expenditure. In realty, there is likely a flattening out of food expenditure for high incomes because only so much money can be spent on food.

```
bhat = lm(log(total)~yards+att+exp+exp2+draft1+veteran+changeteam+pbowlever,data=nfl)
summary(bhat)
```

```
##
## Call:
## lm(formula = log(total) ~ yards + att + exp + exp2 + draft1 +
##     veteran + changeteam + pbowlever, data = nfl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4477 -0.5010 -0.0807  0.4452  3.1638
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.0270821  0.1228294  -8.362 4.93e-16 ***
## yards        0.0003353  0.0001426   2.351 0.019087 *
## att          0.0005137  0.0009728   0.528 0.597691
## exp          0.0702541  0.0601679   1.168 0.243452
## exp2        -0.0037576  0.0036704  -1.024 0.306398
## draft1       0.8910570  0.1132812   7.866 1.90e-14 ***
## veteran      0.5752784  0.1493617   3.852 0.000131 ***
## changeteam  -0.3221519  0.0901474  -3.574 0.000383 ***
## pbowlever    0.4158535  0.0938203   4.432 1.12e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7909 on 559 degrees of freedom
##   (441 observations deleted due to missingness)
## Multiple R-squared:  0.5509, Adjusted R-squared:  0.5445
## F-statistic: 85.71 on 8 and 559 DF,  p-value: < 2.2e-16
```

11.5 Interaction Effects

Interaction effects are used when the influence of one independent variable depends on the level of another independent variable. Suppose that you want to measure time spent volunteering (y) and you think that it depends on the marital status (x_1), the number of children (x_2), and some other independent variables (X). So you could have the following regression equation

$$y = \beta_1 \cdot x_1 \cdot x_2 + \beta \cdot X + \epsilon$$