**Multiple linear regression models**

There can be more than one independent (explanatory) variable in a regression model! For example, a model may look like:

E(Y) = β0 + β1x1 + β2x2

where x1 and x2 are two separate independent variables. The interpretation of x1 becomes:

E(Y) will increase by β1 for every one unit increase in x1, holding x2 constant.

A similar interpretation for x2 can be made.

To include additional independent variables in the lm() function, the variable names are separated by “+”. For example, the formula argument would look like

formula = y ~ x1 + x2

where x1 represents x1 and x2 represents x2 in a data frame.

There are various transformations of independent variables that are sometimes useful to include in a model. For example, suppose there is a quadratic relationship between x and Y. You may be interested in the model

E(Y) = β0 + β1x + β2x2

To include the x2 term in the lm() function, the formula argument would be:

formula = y ~ x + I(x^2)

Finally, there may be situations where an interaction between two independent variables (i.e., x1x2) is of interest to include in a model. For example, the model of interest may be

E(Y) = β0 + β1x1 + β2x2 + β3x1x2

And the corresponding formula argument for the lm() function could be one of the following three:

formula = y ~ x1 + x2 + x1:x2

formula = y ~ x1\*x2

formula = y ~ (x1 + x2)^2

Examples are provided in the gpa\_regression.R program. A whole course on regression analysis would provide much more detail.

Qualitative independent variables

In this chapter, we have only examined situations where there are quantitative (numerical) variables. What if there is a qualitative variable like gender, political party affiliation, or some of the type of classification? These variables can be included in a regression model too!

Example: Gender as an independent variable

Consider using gender to estimate some variable Y, where we use a binary definition for it. We can simply let one gender level be equal to 0 and one gender level be equal to 1:

x = 1 if female (F)

0 if male (M)

This 0-1 variable is referred to as an “indicator” variable. Notice there is one indicator variable for 2 levels (male or female) of the qualitative variable; however, each level has a unique coding.

Suppose the model is E(Y) = β0 + β1x. We have one of two possible values for E(Y):

If x = 0 (male), then E(Y) = μM = β0.

If x = 1 (female), then E(Y) = μF = β0 + β1.

where μM and μF are the mean of the response variable for male and female, respectively. Thus, μF - μM = β0 + β1 - β0 = β1. This means that the hypothesis test of Ho:β1 = 0 vs. Ha:β1 ≠ 0 is equivalent toa hypothesis test for   
Ho:μF - μM = 0 vs. Ho:μF - μM ≠ 0 (hypothesis test for differences between means of independent samples).

Furthermore, the model above is simply an ANOVA model for one factor! We used the expression below previously in the notes:

E(Y) = μ + αi for i = 1, 2

where μ is the grand mean and αi represents the “effect” of the ith treatment level. In this setting, one often will let α1 = 0. Then E(Y) = μ for i = 1 and E(Y) = μ + α2 for i = 2. Notice how this matches up with E(Y) = β0 and E(Y) = β0 + β1 shown previously. Thus, a regression model with one qualitative variable is equivalent to a one-factor ANOVA model!

Question: Will anything change if x = 1 for male and 0 for female? The estimates for β0 and β1 will change, but the overall conclusions from a hypothesis test of Ho:β1=0 vs. Ha:β1≠0 will not. Also, your interpretation of the model will not change. This is similar to doing the hypothesis test for Ho:μ1 - μ2 = 0 vs. Ho:μ1 - μ2 ≠ 0 where you flip the meaning of “1” and “2”.

Example: Political party affiliation (Republican, Democrat, and Independent)

Consider using political party affiliation to estimate some variable Y. We can create two indicator variables to represent this three-level qualitative variable:

x1 =1 if Republican, 0 otherwise

x2 =1 if Democrat, 0 otherwise

Each level of political party affiliation has a unique coding:

| **Party** | **x1** | **x2** |
| --- | --- | --- |
| Republican | 1 | 0 |
| Democrat | 0 | 1 |
| Independent | 0 | 0 |

Suppose the model is E(Y) = β0 + β1x1 + β2x2. The model can be represented as

Republican: E(Y) = β0 + β1

Democrat: E(Y) = β0 + β2

Independent: E(Y) = β0

This again is an ANOVA model for one factor! The model is

E(Y) = μ + αi for i = 1, 2, 3

where μ is the grand mean and αi represents the “effect” of the ith treatment level. To test

Ho:μR = μD = μI

Ha: At least two means are unequal

we would construct an ANOVA table and use Fobs = MST/MSE to perform the test. For a regression setting, we can do the exact same thing by testing

Ho: β1 = β2 = 0

Ha: At one βi ≠ 0

An ANOVA table can also be constructed for a regression model that has the same form as in previously in the notes, and the same test statistic can be used. This will be illustrated in the next example.

In general, if there are c different levels for a qualitative variable, then c – 1 indicator variables are needed. No matter what you choose as the coding for the indicator variables (Republican, Democrat, or Independent is the “all 0 level”), you will get the same  and hypothesis tests involving all of the indicator variables will be the same.

Example: Wheaties Cereal (wheaties\_regression.R, wheaties.csv)

Below is the R code from using the aov() function:

> wheaties <- read.csv(file = "wheaties.csv")

> wheaties

Design Store Response

1 1 1 12

2 1 2 18

3 2 1 14

4 2 2 12

5 2 3 13

6 3 1 19

7 3 2 17

8 3 3 21

9 4 1 24

10 4 2 30

> mod.fit <- aov(formula = Response ~ factor(Design), data

= wheaties)

> summary(object = mod.fit)

Df Sum Sq Mean Sq F value Pr(>F)

factor(Design) 3 258 86.000 11.217 0.007135 \*\*

Residuals 6 46 7.667

---

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

We saw previously how the lm() function could be used to estimate the same model and help to produce the ANOVA table:

> mod.fit.lm <- lm(formula = Response ~ factor(Design),

data = wheaties)

> mod.fit.lm$coefficients

(Intercept) factor(Design)2 factor(Design)3 factor(Design)4

15 -2 4 12

> summary(object = mod.fit.lm)

Call:

lm(formula = Response ~ factor(Design), data = wheaties)

Residuals:

Min 1Q Median 3Q Max

-3.00 -1.75 0.00 1.75 3.00

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 15.000 1.958 7.661 0.000258 \*\*\*

factor(Design)2 -2.000 2.528 -0.791 0.458922

factor(Design)3 4.000 2.528 1.583 0.164620

factor(Design)4 12.000 2.769 4.334 0.004908 \*\*

---

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

Residual standard error: 2.769 on 6 degrees of freedom

Multiple R-squared: 0.8487, Adjusted R-squared: 0.773

F-statistic: 11.22 on 3 and 6 DF, p-value: 0.007135

> anova(object = mod.fit.lm)

Analysis of Variance Table

Response: Response

Df Sum Sq Mean Sq F value Pr(>F)

factor(Design) 3 258 86.000 11.217 0.007135 \*\*

Residuals 6 46 7.667

---

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

R automatically creates the indicator variables for the regression model. The contrasts() function shows how the indicator variables are constructed:

> contrasts(x = factor(wheaties$Design))

2 3 4

1 0 0 0

2 1 0 0

3 0 1 0

4 0 0 1

Thus, x1 = 1 for level 2 of Design and 0 otherwise. The sample regression model is



where x1 represents level 2, x2 represents level 3, and x3 represents level 4.

To help see how the lm() function works here, below is my code used to estimate the model again, but now I create the indicator variables myself first:

> x1 <- c(0, 0, 1, 1, 1, 0, 0, 0, 0, 0)

> x2 <- c(0, 0, 0, 0, 0, 1, 1, 1, 0, 0)

> x3 <- c(0, 0, 0, 0, 0, 0, 0, 0, 1, 1)

> wheaties2 <- data.frame(wheaties, x1, x2, x3)

> mod.fit.lm2 <- lm(formula = Response ~ x1 + x2 + x3, data

= wheaties2)

> mod.fit.lm2$coefficients

(Intercept) x1 x2 x3

15 -2 4 12

> summary(object = mod.fit.lm2)

Call:

lm(formula = Response ~ x1 + x2 + x3, data = wheaties2)

Residuals:

Min 1Q Median 3Q Max

-3.00 -1.75 0.00 1.75 3.00

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 15.000 1.958 7.661 0.000258 \*\*\*

x1 -2.000 2.528 -0.791 0.458922

x2 4.000 2.528 1.583 0.164620

x3 12.000 2.769 4.334 0.004908 \*\*

---

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

Residual standard error: 2.769 on 6 degrees of freedom

Multiple R-squared: 0.8487, Adjusted R-squared: 0.773

F-statistic: 11.22 on 3 and 6 DF, p-value: 0.007135

> anova(object = mod.fit.lm2) #Does not give exactly what

we want

Analysis of Variance Table

Response: Response

Df Sum Sq Mean Sq F value Pr(>F)

x1 1 107.143 107.143 13.9752 0.009641 \*\*

x2 1 6.857 6.857 0.8944 0.380797

x3 1 144.000 144.000 18.7826 0.004908 \*\*

Residuals 6 46.000 7.667

---

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

> #No independent variable model

> mod.fit.lm3 <- lm(formula = Response ~ 1, data =

wheaties2)

> summary(mod.fit.lm3)

Call:

lm(formula = Response ~ 1, data = wheaties2)

Residuals:

Min 1Q Median 3Q Max

-6.00 -4.75 -0.50 2.50 12.00

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 18.000 1.838 9.794 4.25e-06 \*\*\*

---

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

Residual standard error: 5.812 on 9 degrees of freedom

> mean(wheaties2$Response)

[1] 18

> anova(mod.fit.lm3, mod.fit.lm2, test = "F")

Analysis of Variance Table

Model 1: Response ~ 1

Model 2: Response ~ x1 + x2 + x3

Res.Df RSS Df Sum of Sq F Pr(>F)

1 9 304

2 6 46 3 258 11.217 0.007135 \*\*

---

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

The final use of anova() gives the same F test statistic and p-value as we obtained earlier.