Multiple Linear Regression Modeling Estimation

K Boswell

2023-07-17

In this methodology, we use estimation and hypothesis testing in multiple linear regression.

Step 1: Overall Test

We are interested in asking the question whether BMI or Age is useful in explaining the variability of LDL cholesterol. LDLi = β0 + β1BMIi + β2Agei + εi H0 : β1 = β2 = 0

What is alternative hypothesis H1? Ha: β1 OR β2 ≠0

Perform the test in R and interpret the results:

```
# Load haven to read sas data
library(haven)
hersdata <- read_sas("hersdata.sas7bdat")

# Load tidy to eliminate na values from selected column
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.3.1
```

```
herstidy <- hersdata %>% drop_na(LDL, BMI, AGE)
selected_columns <- c("LDL", "BMI", "AGE", "SMOKING", "DRINKANY", "NONWHITE")

# Verify column names and na values
selected_data <- summary(herstidy[, selected_columns])
print(selected_data)
```

```
##       LDL             BMI             AGE           SMOKING      
##  Min.   : 36.8   Min.   :15.21   Min.   :44.00   Min.   :0.0000  
##  1st Qu.:119.6   1st Qu.:24.62   1st Qu.:62.00   1st Qu.:0.0000  
##  Median :141.0   Median :27.74   Median :67.00   Median :0.0000  
##  Mean   :145.1   Mean   :28.57   Mean   :66.65   Mean   :0.1289  
##  3rd Qu.:166.0   3rd Qu.:31.73   3rd Qu.:72.00   3rd Qu.:0.0000  
##  Max.   :393.4   Max.   :54.13   Max.   :79.00   Max.   :1.0000  
##                                                                  
##     DRINKANY         NONWHITE     
##  Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :0.0000   Median :0.0000  
##  Mean   :0.3913   Mean   :0.1129  
##  3rd Qu.:1.0000   3rd Qu.:0.0000  
##  Max.   :1.0000   Max.   :1.0000  
##  NA's   :2
```

```
# Fit the full multiple linear regression model
full_model <- lm(LDL ~ BMI + AGE, data = herstidy)

# Perform Overall F-test
overall_ftest <- anova(full_model)
print(overall_ftest)
```

```
## Analysis of Variance Table
## 
## Response: LDL
##             Df  Sum Sq Mean Sq F value   Pr(>F)   
## BMI          1   14446 14446.0 10.1553 0.001455 **
## AGE          1    7567  7567.2  5.3196 0.021161 * 
## Residuals 2744 3903361  1422.5                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results indicate that at least one of the variables is significant at the .05 alpha level (p = .0015 for BMI, p = .02 for AGE). Therefore, both variables are showing significance, not accounting for interactions. As a result, I initially reject H0 (that none of the variables are significant), and will continue to evaluate this model.

Step 2: Tests of Individual Regression Coefficients

Given the model E(LDLi) = β0 + β1BMIi + β2Agei we want to test H0 : β1 = 0 vs H1 : β1 /= 0, given that Agei is in the model. We will do it in different ways.

Method 1: T-Test

```
# T-test results for the individual coefficients
t_test_result <- summary(full_model)
print(t_test_result)
```

```
## 
## Call:
## lm(formula = LDL ~ BMI + AGE, data = herstidy)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -110.388  -25.335   -3.691   20.997  246.236 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 151.4427     8.7744  17.260  < 2e-16 ***
## BMI           0.3665     0.1320   2.778  0.00551 ** 
## AGE          -0.2530     0.1097  -2.306  0.02116 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 37.72 on 2744 degrees of freedom
## Multiple R-squared:  0.005608,   Adjusted R-squared:  0.004883 
## F-statistic: 7.737 on 2 and 2744 DF,  p-value: 0.0004458
```

The p-value is .0055, so we reject the null hypothesis that H0: β_1=0 and conclude that BMI is significantly associated with LDL, controlling for Age.

Method 2, Part 1: Partial F-Test by comparing SSR from two models: linear regression models with and without BMI.

```
# STEP TWO *********************************************************************
# Tests of individual regression coefficients
# Questions:
# 1. Is there an increase in the SSR, and is it enough to warrant
# an additional predictor in the model?
# 2. Are we adding an unimportant predictor that increases the residual mean square
# and therefore reduces the usefulness of the model?

# Two-Model non-direct comparison:

# Fitting the linear regression model with LDL regressed solely on Age
age_model <- lm(LDL ~ AGE, data = herstidy)

# Extracting the Sum of Squares Regression (SSR) for the Age model
age_ssr <- sum((predict(age_model) - mean(herstidy$LDL))^2)

# Extracting the Sum of Squares Regression (SSR) for the full model
full_model_ssr <- sum((predict(full_model) - mean(herstidy$LDL))^2)

# Creating a data frame to display the SSR values for both models
ssr_comparison <- data.frame(
  Model = c("Age Model", "Full Model (Age + BMI)"),
  SSR = c(age_ssr, full_model_ssr)
)

# Printing the SSR for the Age model
print(ssr_comparison)
```

```
##                     Model      SSR
## 1               Age Model 11038.85
## 2 Full Model (Age + BMI) 22013.22
```

Our SSR in the full model is increased quite a bit, which is a good sign. The SSR helps us see how much variance in the response variable is explained with the model. Because it increases, is suggests that the predictors improve the model’s performance.We now need to extract the F-statistic.

Method 2, Part 2:

```
# Calculate the MSE, which represents the variance of the residuals in the full model.
# It is calculated as the SSE/DF
sse <- sum((residuals(full_model))^2)
df_error <- nrow(herstidy) - length(coef(full_model)) - 1
mse <- sse / df_error

# Take the SSR values for both models
ssr_bmi_age <- full_model_ssr
ssr_age <- age_ssr

# Calculate the F-statistic: Length(coef(full_model)) is the number of coefficients in the full
model,
# including the intercept, and Length(coef(age_model) gives the number of coefficients in the ag
e-only model,
# including the itnercept
f_statistic <- ((ssr_bmi_age - ssr_age) / (length(coef(full_model)) - length(coef(age_model))))
/ mse
print(f_statistic)
```

```
## [1] 7.71199
```

Because the F-statistic is large, it suggests that the predictors significantly improve the model’s performance. We now verify with a partial F-test and extract the p-value.

Method 3: Partial F-Test

```
# Use the partial f-test to show that the full model compared to only one variable
# is significant.

reduced_model <- lm(LDL ~ 1, data = herstidy)
partial_f_result <- anova(reduced_model, full_model)
print(partial_f_result)
```

```
## Analysis of Variance Table
##
## Model 1: LDL ~ 1
## Model 2: LDL ~ BMI + AGE
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   2746 3925375
## 2   2744 3903361  2     22013 7.7375 0.0004458 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the p-value of these results, we have .00045. That is statistically significant. Given the increase in SSR, the significant F-statistic, and the p-value with significance at the .001 alpha level, we can reject the null hypothesis.

Tests for Groups of Predictors

Often, it is of interest to determine if, collectively, a group of predictors significantly contribute to the variability in Y given another group of predictors are in the model.

Given the model

E(LDLi) = β0 + β1statinsi + β2BMIi +β3statinsi × BMIi + β4Agei + β5Smokingi +β6Drinkanyi + β7Nonwhitei

where there are *two* terms associated with BMI, we would like to know if BMI is significantly associated with LDL levels, given the model that this association differs by statin use?

In other words, we want to test H0: β2 = β3 = 0 vs H1: at least one of β2, β3 /= 0, given other predictors are in the model. We will use several different methods.

Method for Groups of Predictors 1: SSR and F-Statistic Comparison between Models

This is similar to the steps we took above, where we extracted SSR. Here, we are testing SSR extraction from the interaction model and the reduced model without any BMI.

```
# Step One:
# Fitting the full multiple linear regression model with all predictors of interest
# Making sure to multiply BMI with statins to account for interaction
interaction_model_bmi <- lm(LDL ~ BMI + STATINS + BMI:STATINS + AGE + SMOKING +
                   DRINKANY + NONWHITE, data = herstidy)
# Fitting the reduced model without any BMI
model_no_bmi <- lm(LDL ~ STATINS + AGE + SMOKING + DRINKANY + NONWHITE, data = herstidy)



# Extracting the Sum of Squares Regression (SSR) for the Interaction model
interaction_ssr_bmi <- sum((predict(interaction_model_bmi) - mean(herstidy$LDL))^2)

# Extracting the Sum of Squares Regression (SSR) for the Non-BMI model
no_bmi_ssr <- sum((predict(model_no_bmi) - mean(herstidy$LDL))^2)

# Creating a data frame to display the SSR values for both models
group_ssr_comparison <- data.frame(
  Model = c("Interaction Model with BMI", "No BMI Model"),
  SSR = c(interaction_ssr_bmi, no_bmi_ssr)
)
print(group_ssr_comparison)
```

```
##                        Model      SSR
## 1 Interaction Model with BMI 216682.3
## 2               No BMI Model 198227.9
```

We now extract the f-statistic to further test for significance.

We get an increase in SSR in the interaction model. Now we need to look at the F-statistic and the p-value.

```
# Performing the partial F-test to compare the two models
group_f_result <- anova(model_no_bmi, interaction_model_bmi)
print(group_f_result)
```

```
## Analysis of Variance Table
##
## Model 1: LDL ~ STATINS + AGE + SMOKING + DRINKANY + NONWHITE
## Model 2: LDL ~ BMI + STATINS + BMI:STATINS + AGE + SMOKING + DRINKANY +
##     NONWHITE
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1   2739 3725955
## 2   2737 3707501  2     18454 6.8118 0.001119 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is .001119 and the F-statistic is 6.8118. Therefore, using this method, we reject the null hypothesis, as we have strong evidence that BMI is associated with LDL levels.

Method for Groups of Predictors 2: Simultaneous Regression

What if we want to test whether BMI is significantly associated with LDL levels for those people receiving statins?

In other words, we want to test H0 : β2 + β3 = 0 vs H1: β2 + β3 /= 0, given that other predictors are in the model. Notice the differences in H0 from the above.

Step 1: Partial F Comparison in R:

```
# Testing BMI significance associated with LDL for those people
# receiving statins, given other predictors are in the model

herstidy$BMIMINUSINTERACTION <- herstidy$BMI - herstidy$BMI*herstidy$STATINS

interaction_model_withbmistatin <- lm(LDL ~ STATINS + AGE + SMOKING + DRINKANY + NONWHITE + STAT
INS:BMI +
                               BMIMINUSINTERACTION, data = herstidy)

reduced_model_no_bmistatin <- lm(LDL ~ STATINS + AGE + SMOKING + DRINKANY + NONWHITE + BMIMINUSI
NTERACTION, data = herstidy)

summary(interaction_model_withbmistatin)
```

```
##
## Call:
## lm(formula = LDL ~ STATINS + AGE + SMOKING + DRINKANY + NONWHITE +
##     STATINS:BMI + BMIMINUSINTERACTION, data = herstidy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -105.649  -24.061   -3.601   19.862  238.167
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          145.7684     9.4701  15.393  < 2e-16 ***
## STATINS                3.8081     7.8249   0.487 0.626536
## AGE                   -0.1729     0.1106  -1.563 0.118099
## SMOKING                3.1098     2.1670   1.435 0.151386
## DRINKANY              -2.0753     1.4666  -1.415 0.157168
## NONWHITE               4.0728     2.2751   1.790 0.073544 .
## BMIMINUSINTERACTION    0.5821     0.1601   3.636 0.000282 ***
## STATINS:BMI           -0.1198     0.2207  -0.543 0.587206
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.8 on 2737 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.05522,    Adjusted R-squared:  0.0528
## F-statistic: 22.85 on 7 and 2737 DF,  p-value: < 2.2e-16
```

```
anova(interaction_model_withbmistatin, reduced_model_no_bmistatin)
```

```
## Analysis of Variance Table
##
## Model 1: LDL ~ STATINS + AGE + SMOKING + DRINKANY + NONWHITE + STATINS:BMI +
##     BMIMINUSINTERACTION
## Model 2: LDL ~ STATINS + AGE + SMOKING + DRINKANY + NONWHITE + BMIMINUSINTERACTION
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   2737 3707501
## 2   2738 3707900 -1   -399.33 0.2948 0.5872
```

We are interested in the p-value from the simultaneous test for general linear hypothesis. The p-value is .5872, indicating no evidence that BMI is associated with LDL levels for people taking statins. I do not reject the null hypothesis.