Multiple Linear Regression Modeling Methods

K Boswell

2023-07-17

Multiple Linear Regression Model Formulation with Binary Indicators

Effects of Air Pollution on Vasoconstriction

Researchers in toxicology study were interested in evaluating the effects of air pollution on vasoconstriction of small pulmonary arteries in rats. In addition, they wanted to know if these health effects differed according to whether the rats had preexisting pulmonary disease. Thus, chronic bronchitis was induced in some of the rats by exposing them to SO2 for 6 weeks prior to pollution exposure. All rats were randomized to one of four groups:

- filtered air, not exposed to SO2 (INDICATOR 0)
- concentrated air particles (CAPs), not exposed to SO2 (INDICATOR 0)
- filtered air, exposed to SO2 (INDICATOR 1)
- concentrated air particles (CAPs), exposed to SO2 (INDICATOR 1)

and the amount of pulmonary inflammation, as measured by neutrophil numerical den- sity (Nn) in each animal was measured after three successive days of air pollution exposure. A large value of Nn denotes pulmonary inflammation.

The first step is to consider the main effect model and the interaction model (which contains both main effects and interactions) that simultaneously assess two categorical factors (SO2 and CAPs).

- **Nn = response variable**
- **caps, so2 = predictor variables**

Main effect model: Nn = β_0 + β_1(so2) + β_2(caps) + ε, where:

- **Nn is the numerical density of neutrophils**
- **Caps is the categorical variable for concentrated air particles, where 0 = not exposed and 1 = exposed**
- **β_0 is the intercept term. This will represent the average Nn when caps and so2 are at 0.**
- **β_1 is the coefficient for the effect of s02 on Nn. This will represent the change in Nn associated with being exposed to so2 when caps are held constant at 0.**
- **β_2 is the coefficient for the effect of caps on Nn. This will represent the change in Nn associated with being exposed to caps when so2 is held constant at 0.**
- **ε is the error term.**

Interaction model: Nn = β_0 + β_1(so2) + β_2(caps) + β3 (so2 * caps) + ε, where

- **Nn, caps, so2, β_0, β_1, β_2, ε all have the same meaning as above.**
- **β_3 is the coefficient for the interaction between caps and so2. This will represent the additional change in Nn due to the combined effect of being exposed to both so2 and caps. We are looking for statistical significance of β_3 to tell us if one pollutant is modified by the other pollutant.**

The next step is to consider a test of whether there is a difference in the health effects of air pollution inhalation for healthy animals and that for chronic bronchitic animals (i.e. those who received SO2) under the interaction model.

The null hypothesis is: H0: β3 = 0

The alternative hypothesis is: Ha: β3 ≠ 0

Since we cannot test the null hypothesis under the main model, I would perform a partial F-test of the models to see the p-value and test it at the .05 alpha level. We could also compare the Sum of Squares Regression between the interaction model and the main model to test if there is an increase. An increase in SSR from the main model suggests that we can reject the null hypothesis.

```
# Fit the interaction model
interaction_model_pollution <- lm(nn ~ so2 + caps + so2:caps, data = pollution)
main_model_pollution <- lm(nn ~ so2 + caps, data = pollution)

# Extracting SSR for the Interaction Model
interaction_ssr_pollution <- sum((predict(interaction_model_pollution) - mean(pollution$nn))^2)

# Extracting SSR for the Caps Model
reduced_ssr <- sum((predict(main_model_pollution) - mean(pollution$nn))^2)

# Creating a data frame to dsiplay the SSR values for both models:
ssr_compare_pollution <- data.frame(
  Model = c("Interaction Model", "Main Model"),
  SSR = c(interaction_ssr_pollution, reduced_ssr)
)
print(ssr_compare_pollution)
```

```
##               Model      SSR
## 1 Interaction Model 7.779052
## 2        Main Model 7.506952
```

```
# partial f-test for model comparison
group_f_caps_so2 <- anova(main_model_pollution, interaction_model_pollution)
print(group_f_caps_so2)
```

```
## Analysis of Variance Table
##
## Model 1: nn ~ so2 + caps
## Model 2: nn ~ so2 + caps + so2:caps
##   Res.Df     RSS Df Sum of Sq     F    Pr(>F)
## 1    272 1.03568
## 2    271 0.76358  1    0.2721 96.57 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is significant at the .001 level, so I reject the null hypothesis.

In our scenario, the investigator neglected to mention that half of the animals received SO2, so we fit a regression model for Nn using CAPs exposure only. Without all the information, we ignored whether the animal is chronic bronchitic or not. The corresponding regression model is therefore reduced. However, this is useful because We can fit both this model and the main effect model that includes CAPs and SO2 to check for confounding.

Caps model: $Nn = \beta_0 + \beta_1(caps) + \varepsilon$

Main effect model: $Nn = \beta_0 + \beta_1(so2) + \beta_2(caps) + \varepsilon$

```
# Fit a model with caps alone
caps_model <- lm(nn ~ caps, data = pollution)
summary(caps_model)
```

```
##
## Call:
## lm(formula = nn ~ caps, data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25407 -0.07357 -0.00360  0.07866  0.24093
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.292600   0.008534   34.29   <2e-16 ***
## caps        0.289474   0.012180   23.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.101 on 273 degrees of freedom
## Multiple R-squared:  0.6742, Adjusted R-squared:  0.673
## F-statistic: 564.8 on 1 and 273 DF,  p-value: < 2.2e-16
```

```
# Fit the main effect model
main_model_pollution <- lm(nn ~ so2 + caps, data = pollution)
summary(main_model_pollution)
```

```
##
## Call:
## lm(formula = nn ~ so2 + caps, data = pollution)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.189497 -0.039997  0.003503  0.039662  0.171593
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.203407   0.006673   30.48   <2e-16 ***
## so2         0.160090   0.007472   21.43   <2e-16 ***
## caps        0.302773   0.007469   40.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06171 on 272 degrees of freedom
## Multiple R-squared:  0.8788, Adjusted R-squared:  0.8779
## F-statistic: 985.8 on 2 and 272 DF,  p-value: < 2.2e-16
```

The parameter estimate for the caps coefficient in the caps model by itself is .28974. In the model with so2 added in, it is .30277. This is an increase of 4.6%, so it is an increase, but it does not reach the 10% level, so it is unclear whether so2 confounds the association and we might need to look at another method to check for confounding.

Another way to justify whether SO2 confounds the association between CAPs and Nn is by looking at the pairwise association among these three variables directly. We can use Pearson coefficient testing and build a matrix in R to test the correlations.

```
# compare the correlations directly using Pearson
# Calculate the correlation matrix for the variables
cor_matrix <- cor(pollution[c("caps", "so2", "nn")])

# Print the correlation matrix
print(cor_matrix)
```

```
##             caps         so2        nn
## caps  1.00000000 -0.08309957 0.8210664
## so2  -0.08309957  1.00000000 0.3825468
## nn    0.82106644  0.38254677 1.0000000
```

Based on these results, I don't believe that so2 confounds the association between caps and nn. The correlation between caps and nn is very strong (.82) and the correlation between so2 and nn is only moderate at .38. I believe that caps and so2 are independently associated with an increase in nn, but I would be cautious with this interpretation.

Now we begin the step of looking for interaction. Under the main effect model and the interaction model, we test whether there is a health effect of air pollution inhalation for healthy animals, and also test whether there is a health effect of air pollution inhalation for chronic bronchitic animals (i.e. those who received SO2).

Step 1: Use the main effect model to test the null hypothesis that H0: β1=β2=0

```
# Run F-test on the main effect model
main_ftest_pollution <- anova(main_model_pollution)
print(main_ftest_pollution)
```

```
## Analysis of Variance Table
##
## Response: nn
##            Df Sum Sq Mean Sq F value    Pr(>F)
## so2         1 1.2501  1.2501  328.32 < 2.2e-16 ***
## caps        1 6.2568  6.2568 1643.22 < 2.2e-16 ***
## Residuals 272 1.0357  0.0038
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

These results indicate that caps is significant at the .05 alpha level.

Step 2: test the coefficients with a t-test. The null hypothesis is at least one of the variables is not significant at the .05 level. We already did this above, but I will re-run the code for ease of reading.

```
summary(main_model_pollution)
```

```
##
## Call:
## lm(formula = nn ~ so2 + caps, data = pollution)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.189497 -0.039997  0.003503  0.039662  0.171593
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.203407   0.006673   30.48   <2e-16 ***
## so2         0.160090   0.007472   21.43   <2e-16 ***
## caps        0.302773   0.007469   40.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06171 on 272 degrees of freedom
## Multiple R-squared:  0.8788, Adjusted R-squared:  0.8779
## F-statistic: 985.8 on 2 and 272 DF,  p-value: < 2.2e-16
```

Each of the coefficients is significant at the .001 level, not accounting for interactions, so I reject the null hypothesis.

Step 3: look at the SSR between models.

```
# Extracting the Sum of Squares Regression (SSR) for the Caps Model
caps_ssr <- sum((predict(caps_model) - mean(pollution$nn))^2)

# Extracting the Sum of Squares Regression (SSR) for the main model
main_model_ssr <- sum((predict(main_model_pollution) - mean(pollution$nn))^2)

# Creating a data frame to siplay the SSR values for both models
ssr_comparison_pollution <- data.frame(
  Model = c("Caps Model", "Main Model (CAPS + SO2)"),
  SSR = c(caps_ssr, main_model_ssr)
)

# Printing the SSR for the main model
print(ssr_comparison_pollution)
```

```
##                     Model      SSR
## 1              Caps Model 5.759018
## 2 Main Model (CAPS + SO2) 7.506952
```

The SSR is increased in the model with so2, indicating that the variable is useful to the model.

Step 4: Run a partial F-test to show the main model compared to only one variable

```
# use the partial f-test to show the full model compared to only variable

reduced_model_pollution <- lm(nn ~ 1, data = pollution)
partial_f_two_model_pollution <- anova(reduced_model_pollution, main_model_pollution)
print(partial_f_two_model_pollution)
```

```
## Analysis of Variance Table
##
## Model 1: nn ~ 1
## Model 2: nn ~ so2 + caps
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    274 8.5426
## 2    272 1.0357  2     7.507 985.77 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is very significant and the residuals are reduced, indicating that the main model captures a significant relationship between the two variables and nn, as opposed to one variable alone.

Step 5: look at the interaction model. Extract the SSR for the interaction model and the main model and compare.

```
# Extracting SSR for the Interaction Model
interaction_ssr_pollution <- sum((predict(interaction_model_pollution) - mean(pollution$nn))^2)

# Extracting SSR for the Caps Model
caps_ssr <- sum((predict(caps_model) - mean(pollution$nn))^2)

# Creating a data frame to dsiplay the SSR values for both models:
group_ssr_compare_pollution <- data.frame(
  Model = c("Interaction Model with SO2", "No SO2 Model"),
  SSR = c(interaction_ssr_pollution, caps_ssr)
)
print(group_ssr_compare_pollution)
```

```
##                        Model      SSR
## 1 Interaction Model with SO2 7.779052
## 2               No SO2 Model 5.759018
```

The SSR increases in the interaction model, suggesting it is useful in the model.

Step 6: run a partial f-test for model comparison.

```
# partial f-test for model comparison
group_f_caps_so2 <- anova(main_model_pollution, interaction_model_pollution)
print(group_f_caps_so2)
```

```
## Analysis of Variance Table
##
## Model 1: nn ~ so2 + caps
## Model 2: nn ~ so2 + caps + so2:caps
##   Res.Df     RSS Df Sum of Sq     F    Pr(>F)
## 1    272 1.03568
## 2    271 0.76358  1    0.2721 96.57 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Once again, this is a very significant p-value.

Multiple Linear Regression Modeling with Continuous and Categorical Variables

Patient Satisfaction in Relation to Age, Severity of Illness, and Anxiety Level

A hospital administrator wished to study the relation between patient satisfaction (Y) and patient’s age (X1, in years), severity of illness (X2, an index) and anxiety level (X3, an index). The administrator randomly selected 23 patients and collected the data in patsat.sas7bdat, where larger values of Y, X2, and X3 are, respectively, associated with more satisfaction, increased severity of illness and more anxiety.

The regression model is $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$, where:

- **β_1 is the coefficient for the effect of age on patient satisfaction.**
- **β_2 is the coefficient for the effect of severity of illness on patient satisfaction.**
- **β_3 is the coefficient for the effect of anxiety level on patient satisfaction.**

```
library(haven)
patsat <- read_sas("patsat.sas7bdat")
patsat_main_model <- lm(Y ~ X1 + X2 + X3, data = patsat)
```

After fitting the model, we then test whether there is a regression relationship here; that is, if the regression as a whole explains variability in the response. Our hypotheses are:

- **Null Hypothesis: H_0: β_1 OR β_2 OR β_3 =0**
- **Alternative Hypothesis: H_a:β_1,β_2,$\beta_3 \neq 0$**

The next step is to run an overall f-test, Using significance level $\alpha = 0.05$ to test the significane of the variables.

```
overall_ftest_patsat <- anova(patsat_main_model)
print(overall_ftest_patsat)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X1         1 3678.4  3678.4 34.7439 1.124e-05 ***
## X2         1  402.8   402.8  3.8044   0.06603 .
## X3         1   52.4    52.4  0.4951   0.49021
## Residuals 19 2011.6   105.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion: The test implies that at least one of β_1, β_2, and β_3 is not equal to zero.

In fact, at least one of the variables is significant (β_1), so I reject the null hypothesis.

Since β_1 is significant, I run a t-test of the individual coefficient at the .05 alpha-level.

```
t_test_patsat <- summary(patsat_main_model)
print(t_test_patsat)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = patsat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.954  -7.154   1.550   6.599  14.888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 162.8759    25.7757   6.319 4.59e-06 ***
## X1           -1.2103     0.3015  -4.015  0.00074 ***
## X2           -0.6659     0.8210  -0.811  0.42736
## X3           -8.6130    12.2413  -0.704  0.49021
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.29 on 19 degrees of freedom
## Multiple R-squared:  0.6727, Adjusted R-squared:  0.621
## F-statistic: 13.01 on 3 and 19 DF,  p-value: 7.482e-05
```

The t-test indicates that β1 is not equal to zero, with a .001 significance level. Therefore, I reject the null hypothesis.

I then ran a 95% confidence interval estimates of β1, β2, and β3.

```
conf_intervals_patsat <- confint(patsat_main_model)
conf_intervals_patsat
```

```
##                  2.5 %      97.5 %
## (Intercept) 108.926839 216.8249581
## X1           -1.841264  -0.5793727
## X2           -2.384272   1.0524608
## X3          -34.234265  17.0082018
```

The confidence intervals give a range of values for each coefficient at the 95% confidence level. They show a 95% confidence that the true value of each coefficient is between the upper and lower boundary displayed. For example, for Age (X1), the lower boundary is -1.841 and the upper boundary is -0.579. We are 95% confident that the true value of X1 lies between those two numbers. And although X2(Illness Severity) and X3(Anxiety Level) are not shown to be statistically significant, the confidence interval still holds.

The hospital administrator would like to know a 95% confidence interval for someone age 35, with a severity illness level of 45 and an anxiety level of 2.2

```r
# Create a new data frame with the specific values of X1, X2, and X3
new_data <- data.frame(X1 = 35, X2 = 45, X3 = 2.2)

# Use the 'predict()' function to calculate the predicted mean satisfaction
predicted_satisfaction <- predict(patsat_main_model, newdata = new_data, interval = "confidenc
e", level = 0.95)
print(predicted_satisfaction)
```

```
##        fit      lwr      upr
## 1 71.60034 62.30057 80.90012
```

Based on these results, we can be 95% confident that the true mean satisfaction for someone age 35, with an Illness Severity level of 45, and a Anxiety Level of 2.2 will fall between 62.30057 and 80.900012. The point estimate of the mean satisfaction for this combination of variables is 71.60034.