MA50260 Statistical Modelling

Lecture 6: Interactions and Covariate Selection in Linear Regression

Ilaria Bussoli

February 23, 2024

Interactions

Suppose that we have two explanatory variables $x_{i,1}$ and $x_{i,2}$.

We consider two models:

1. Model the main effects only,

$$\mathbb{E}(Y_i) = \beta_1 + \beta_2 x_{i,1} + \beta_3 x_{i,2}.$$

2. Include an **interaction** as well,

$$\mathbb{E}(Y_i) = \beta_1 + \beta_2 x_{i,1} + \beta_3 x_{i,2} + \beta_4 x_{i,1} x_{i,2}.$$

We already considered such a model in the `Gas consumption` example.

Illustration

Suppose we conduct a clinical trial with two drugs (A and B) and that we have two possible doses for each drug.

The response variable is the increase in the red blood cell count.

Then, we can imagine three possible interaction effects:

1. A and B affect the response independently $\rightarrow$ No interaction.
2. High doses for both A and B lead to a smaller increase than if only one of the doses is high $\rightarrow$ Negative interaction.
3. The combined effect of the high doses is greater than the individual effects $\rightarrow$ Positive interaction.

Example - Gas Consumption (I)

Recall the model

$$\mathbb{E}(Y_i) = \beta_1 + \beta_2 x_{i,1} + \beta_3 x_{i,2} + \beta_4 x_{i,1} x_{i,2}$$

where

- β_2 describes the **main effect** of outside temperature on gas consumption.
- β_4 is the **interaction** effect between temperature and whether or not insulation is installed.

Do we require the interaction effect?

Example - Gas Consumption (II)

We wish to test whether or not there is an interaction,

$$H_0 : \beta_4 = 0 \qquad \text{vs.} \qquad H_1 : \beta_4 \neq 0.$$

The test statistic is

$$t = \frac{\hat{\beta}_4 - 0}{\text{se}(\hat{\beta}_4)} = 3.22$$

We have to compare the test statistic to $t_{40}(0.975) = 2.02$.

Since $3.22 > 2.021$, we reject H_0 at the 5% level.

$\Rightarrow$ There was a significant change in the relationship between outside temperature and gas consumption following insulation.

Transformations of the Response / Explanatory Variables

In some cases, the linear model fit may be improved by transforming the response or explanatory variables, e.g.:

- Logarithmic or square root transformation
- Box-Cox transformation

$$y^* = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0, \\ \log(y) & \text{for } \lambda = 0. \end{cases}$$

  - $\lambda = 1$ leaves the response unchanged
  - $\lambda = 0$ corresponds to the log transformation
  - $\lambda = \frac{1}{2}$ gives the square root transformation

Covariate Selection - Motivation

Suppose we have p explanatory variables: do we use all of them in the model? If not, which we could use?

Which explanatory variables should we include?

Covariate Selection - Motivation

Suppose we have p explanatory variables: do we use all of them in the model? If not, which we could use?

Which explanatory variables should we include?

There are two options:

- Fit the model with all p variables and drop the ones that are not significant, or
- Use the F-test, which we cover today.

The F-test is generally applied to compare **nested models**.

Nested Models

Consider two models:

1. Model 1 with p_1 explanatory variables;
2. Model 2 with p_2 explanatory variables.

If $p_2 > p_1$, we call Model 1 the **simpler model**.

If Model 2 contains all explanatory variables of Model 1, we say that Model 1 is **nested** in Model 2.

Definition (Nested Model): Consider the models $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\underline{\beta}$ and $\mathbb{E}(\mathbf{Y}) = \tilde{\mathbf{X}}\underline{\beta}$, where is an $n \times p_1$ matrix and $\tilde{\mathbf{X}}$ is an n $\times p_2$ matrix, with $p_1 < p_2$. Then **Model 1 is nested in Model 2** if $\mathbf{X}$ is a (strict) subspace of $\tilde{\mathbf{X}}$.

Example

Consider the models

1. $\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_{i,1}$,
2. $\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2}$,
3. $\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3}$.

Example

Consider the models

1. $\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_{i,1}$,
2. $\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2}$,
3. $\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3}$.

Then

- Model 1 is simpler than Model 2
- Model 1 is nested in Model 2
- Model 2 is nested in Model 3

Example - Brain Weights of Mammals (I)

Let Y_i be the brain weight, $x_{i,1}$ denote body weight and $x_{i,2}$ denote number of hours asleep per day.

Which of the following models L1-L3 fits the data best?

$$\begin{aligned}
\mathbb{E}(\log Y_i) &= \beta_1 + \beta_2 \log x_{i,1}, \\
\mathbb{E}(\log Y_i) &= \beta_1 + \beta_2 x_{i,2}, \\
\mathbb{E}(\log Y_i) &= \beta_1 + \beta_2 \log x_{i,1} + \beta_3 x_{i,2}.
\end{aligned}$$

The estimates and standard errors are

Model	β_1	β_2	β_3
L1	2.15 (0.0991)	0.759 (0.0303)	NA
L2	6.17 (0.675)	-0.299 (0.0588)	NA
L3	2.60 (0.288)	0.728 (0.0352)	-0.0386 (0.0237)

Example - Brain Weights of Mammals (II)

Let's check which variables are significant, i.e., we test

$$H_0 : \beta_j = 0 \qquad \text{vs} \qquad H_1 : \beta_j \neq 0.$$

- For Model L1, the test statistic is $t = 25.09$. As $t_{56}(0.975) = 2.0$, $\log x_{i,1}$ is significant at the 5% level.
- For Model L2, the test statistic is $t = -5.092$. Again, $|-5.092| > t_{56}(0.975)$. So $x_{i,2}$ is significant at the 5% level.
- For Model L3, for β_2 we have $t = 20.67$. The test statistics for β_3 is $t = -1.632$. So $\log x_{i,1}$ is significant but $x_{i,2}$ is not.

The F-test (I)

The F-test gives a formal statistical test to choose between

- Model 1: $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\underline{\beta}$,
- Model 2: $\mathbb{E}(\mathbf{Y}) = \tilde{\mathbf{X}}\underline{\gamma}$,

with Model 1 nested in Model 2.

Consider the sum of squares

$$\begin{aligned} SS_1 &= (\mathbf{y} - \mathbf{X}\underline{\hat{\beta}})^T(\mathbf{y} - \mathbf{X}\underline{\hat{\beta}}), \\ SS_2 &= (\mathbf{y} - \tilde{\mathbf{X}}\underline{\hat{\gamma}})^T(\mathbf{y} - \tilde{\mathbf{X}}\underline{\hat{\gamma}}). \end{aligned}$$

Then

$$SS_2 \leq SS_1.$$

The F-test (II)

Consider the hypothesis test

$$H_0 : \text{Model 1 is the best fit} \quad \text{vs.} \quad H_1 : \text{Model 2 is the best fit.}$$

To conduct this test, we perform

1. Calculate the test statistic

$$F = \frac{(SS_1 - SS_2)/(p_2 - p_1)}{SS_2/(n - p_2)}.$$

2. Compare the test statistic to the $F_{p_2-p_1,n-p_2}$ distribution, and reject H_0 if the test statistic exceeds the critical value.

Example - Brain Weights of Mammals (III)

We considered the three models

$$\begin{aligned}
\mathbb{E}(\log Y_i) &= \beta_1 + \beta_2 \log x_{i,1}, \\
\mathbb{E}(\log Y_i) &= \beta_1 + \beta_2 x_{i,2}, \\
\mathbb{E}(\log Y_i) &= \beta_1 + \beta_2 \log x_{i,1} + \beta_3 x_{i,2}.
\end{aligned}$$

The F-test enables us to choose between Model 1 and Model 3, and between Model 2 and Model 3.

Let's test

H_0 : Model 1 is the best fit vs. H_1 : Model 3 is the best fit.

Example - Brain Weights of Mammals (IV)

We calculate $SS_1 = 28.0$ and $SS_3 = 26.7$.

Since $n = 58$, $p_1 = 2$ and $p_3 = 3$, the test statistic is

$$\begin{aligned} F &= \frac{[SS(L1) - SS(L3)]/(p_2 - p_1)}{SS(L3)/(n - p_2)} \\ &= \frac{(28.0 - 26.7)/(3 - 2)}{26.7/(58 - 3)} \\ &= 2.67. \end{aligned}$$

To test at the 5% level, we need the 95% quantile of the F-distribution with $(p_2 - p_1, n - p_2) = (1, 55)$ degrees of freedom, which is about 4.02.

Since $2.67 < 4.02$, there is not enough evidence to reject H_0.

Origins of the F-test (I)

We know that

$$\mathbb{E}\left[\frac{SS}{(n-p)}\right] = \sigma^2,$$

which also implies that $\mathbb{E}(SS) = (n-p)\sigma^2$.

Origins of the F-test (I)

We know that

$$\mathbb{E}\left[\frac{SS}{(n-p)}\right] = \sigma^2,$$

which also implies that $\mathbb{E}(SS) = (n-p)\sigma^2$.

We then get

$$\mathbb{E}(SS_1 - SS_2) = (n-p_1)\sigma^2 - (n-p_2)\sigma^2 = (p_2-p_1)\sigma^2.$$

So, $(SS_1 - SS_2)/(p_2 - p_1)$ is an unbiased estimator of σ^2.

Origins of the F-test (I)

We know that

$$\mathbb{E}\left[\frac{SS}{(n-p)}\right] = \sigma^2,$$

which also implies that $\mathbb{E}(SS) = (n-p)\sigma^2$.

We then get

$$\mathbb{E}(SS_1 - SS_2) = (n-p_1)\sigma^2 - (n-p_2)\sigma^2 = (p_2-p_1)\sigma^2.$$

So, $(SS_1 - SS_2)/(p_2 - p_1)$ is an unbiased estimator of σ^2.

But if Model 1 is not a sufficiently good model for the data,

$$\mathbb{E}\left[\frac{SS_1 - SS_2}{p_2 - p_1}\right] > \sigma^2.$$

Origins of the F-test (II)

Consequently, the F-statistic

$$F = \frac{(SS_1 - SS_2)/(p_2 - p_1)}{SS_2/(n - p_2)}$$

is the ratio of two estimates of σ^2.

If Model 1 is a sufficient fit, this ratio will be close to 1, otherwise it will be greater than 1.

To see how far the F-statistic must be from 1 for the result not to have occurred by chance, we need its sampling distribution, which is the F-distribution.