Linear Regression III

Sean Hellingman ©

Introduction to Statistical Data Analysis (ADSC1000)

shellingman@tru.ca

Fall 2024

THOMPSON RIVERS UNIVERSITY

Topics

Introduction

- We have covered regression models with a single independent and multiple independent (explanatory) variables.
- We can also include categorical (factor) variables into our regression models.
- Binary factors or factors with multiple levels may be included in regression models.

Dummy Variables

- We can include *Yes* or *No* variables as 1 and 0 respectively in regression models.
- Such variables are often called **dummy variables**.
- Essentially, our coefficient estimate (b_j) will calculate indicate the change in the expected value of Y when going from *No* to *Yes*.

Example Model I

- Assume we would like to create a model with the dependent variable as a salary and the explanatory variables as the age and the respondents' sex:
- Theoretical model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

  where

  Y = salary

  X_1 = age

  X_2 = sex indicator (1 = Female, 0 = Male)

Example 1

- Load the *SLID* data into R and estimate the following linear regression model:

$$\hat{Y}_{salary} = b_0 + b_1 X_{age} + b_2 X_{sexIND}$$

- Interpret your results.
- *We will ignore the missing observations for now.*

Categorical Variables

- When the categorical variables have only two levels we can code the levels as 0 and 1 (Example 1).
- When we have more than two levels ($k > 2$) we need to take a different approach.
- To avoid any multicollinearity issues, we will add $k-1$ variables to the model.
- The factor level not included is called the **reference category**.

Reference Category

- The category that is left out of the regression equation is called the reference category.
- All of the parameters of the *dummy* variables represent the difference or change from this reference category.
- We are able to choose the reference category in our regression models.

Example Model II

- Assume we would like to create a model with the dependent variable as a salary and the explanatory variables as the age and the respondents' first language (English, French, or Other):

- We will need $k - 1 = 2$ *dummy* variables corresponding to two levels of our categorical variable. The level not included will be our reference category.

- Theoretical model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$
  where
  Y = salary
  X_1 = age
  $X_2 = 1$ if the first language is French
  $X_3 = 1$ if the first language is Other
- Note: *When $X_2 = X_3 = 0$ then by default, the first language is English*.

Selecting Reference Category in R

- You permanently change the reference category in your variable, or you can select a reference category in your regression model.

- Permanently change:

  ```
  data.frame$variable.name <-
  relevel(data.frame$variable.name, ref =
  "reference.name")
  ```

- Set in model:

  ```
  lm1 <- lm(variable.y ~ relevel(variable.name,
  "reference.name"), data = data.frame )
  ```

- *Must be a factor.*

Example 2

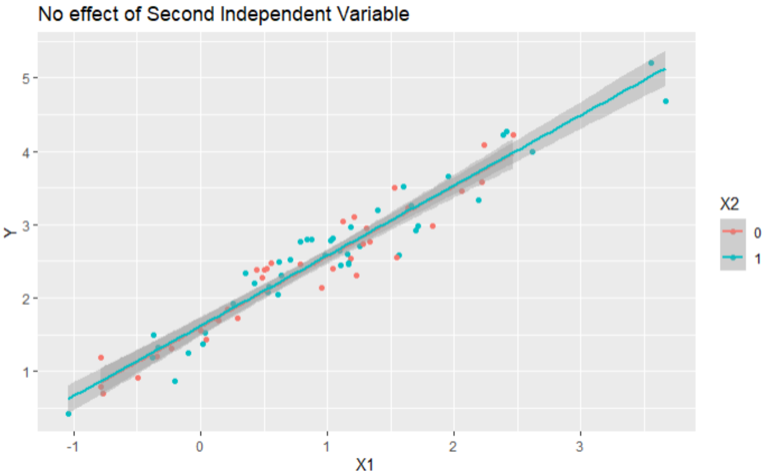
- Load the *SLID* data into R and estimate the following linear regression model:

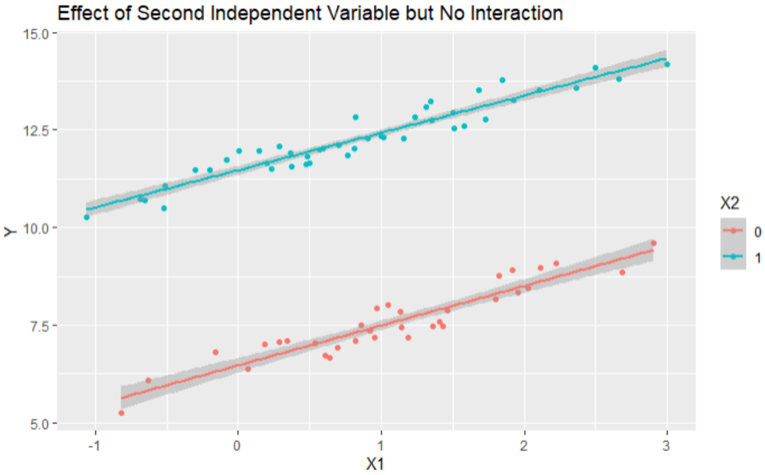
$$\hat{Y}_{salary} = b_0 + b_1 X_{age} + b_2 X_{English} + b_3 X_{Other}$$

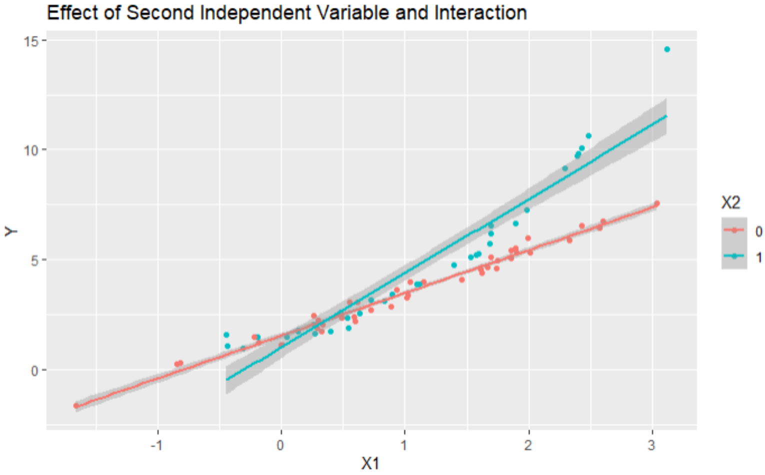
- *Make sure that French is the reference category.*
- Interpret your results.
- *We will ignore the missing observations for now.*

Interactions

- An **interaction** occurs when an independent variable has a different effect on the outcome depending on the values of another independent variable.
- In other words, the slope changes depending on which category (categorical variables) we are in.
- We can use visualizations to help determine if interactions may exist.


No effect of Second Independent Variable
Y
X1
X2
0
1
-1
0
1
2
3
1
2
3
4
5


Effect of Second Independent Variable but No Interaction
15.0
12.5
10.0
7.5
5.0
Y
-1
0
1
2
3
X1
X2
0
1


Effect of Second Independent Variable and Interaction
15
10
5
0
Y
-1
0
1
2
3
X1
X2
0
1

Interaction Terms in R

- We can include interaction terms in our models in R by using the `*` symbol *or* the `:` symbol.

- `lm1 <- lm(Y  X1 + X2 + X1*X2, data = data.frame)`

  OR

- `lm1 <- lm(Y  X1 + X2 + X1:X2, data = data.frame)`

Example 3

- Use the *SLID* data to visualize any potential interactions between *education* and *sex* on the wages the respondents earn.
- Create a linear regression model to validate your findings.
- **Remember, for inferences we need to conduct regression diagnostics on every model.**

Comparing models

- In practice, we should always select a simpler model (fewer independent variables) when two models are comparable.
- Selecting simpler models can save us from *overfitting*.
- If we decide to go with a more complex model, it **must** provide a much better fit to the data.
- There are some ways we can compare models:
  1. ANOVA
  2. Akaike information criterion (AIC)
  3. Bayesian information criterion (BIC)

`anova()`

- The ANOVA tests whether the more complex model is significantly better at capturing variability in the data than the simpler model.
- We can use the `anova(lm1,lm2)` function in R
- A small *p*-value (< 0.05) indicates that the complex model is significantly better at capturing the variability.
- A large *p*-value (> 0.05) indicates that there is very little difference and we should select the simpler model.

Akaike information criterion (AIC)

- The **Akaike information criterion (AIC)** estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model.

- In other words, models with a lower AIC are said to be better based on this criterion.

$$AIC = 2k - 2\ln(\hat{L}).$$

k is the number of estimated parameters.
$\hat{L}$ is the maximized value of the likelihood function (estimation method).

- Therefore $2k$ is a penalization term for adding more parameters to the model.

Bayesian information criterion (BIC)

- The **Bayesian information criterion (BIC)** also estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model.

- In other words, models with a lower BIC are said to be better based on this criterion.

$$BIC = k\ln(n) - 2\ln(\hat{L}).$$

  k is the number of estimated parameters and n is the number of observations.
  $\hat{L}$ is the maximized value of the likelihood function (estimation method).

- Therefore $k\ln(n)$ is a **larger** penalization term for adding more parameters to the model.

AIC and BIC in R

- AIC: `AIC(lm1)`
- BIC: `BIC(lm1)`
- Remember, models with the smallest values are considered better.
- These are two different methods and they may result in different preferences when we are comparing models.

Example 4

- Import the *diamonds* dataset from the *ggplot2* package and do the following:

  1. Estimate the following linear models:
     - `value ~ carat`
     - `value ~ carat + clarity` (be sure to identify the reference category)
     - `value ~ carat + clarity + color` (be sure to identify the reference category)
     - `value ~ carat + clarity + color + carat:clarity`
  2. Use `anova()`, `AIC()`, and `BIC()` to select your preferred model.
  3. Why did you select the model that you did?

Exercise 1

- Using the *SILD* data estimate increasingly complex models including interaction terms. Use the techniques we covered to select the model that you think is the best. Why did you select this model?
- Perform the necessary diagnostic tests on the model that you have selected. Does it pass?

Exercise 2

- We have covered most of the basic concepts of **linear** regression.
- Use the regression techniques that we have learned on your project data.
- Did you uncover anything interesting?

References & Resources

1. Evans, J. R., Olson, D. L., & Olson, D. L. (2007). *Statistics, data analysis, and decision modeling.* Upper Saddle River, NJ: Pearson/Prentice Hall.

2. Devore, J. L., Berk, K. N., & Carlton, M. A. (2012). *Modern mathematical statistics with applications (Second Edition).* New York: Springer.

- https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/lm
- https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/anova
- https://cran.r-project.org/web/packages/car/index.html
- https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/AIC