Model Formulas and Coefficients

Sean Hellingman ©

Regression for Applied Data Science (ADSC2020)

shellingman@tru.ca

Winter 2025

THOMPSON RIVERS UNIVERSITY

Topics

Introduction

- Often it is not practical to express statistical models visually.
- Models often have many explanatory variables and sometimes have complicated relationships.
- Using formulas will help us quantify these relationships.
- We will cover presenting your models using formulas and coefficients.

Equation of a Line

- Linear regression is based on estimating the linear relationship between the dependent and independent variable(s).
- Recall the equation of a line:

$$y = mx + b. \tag{1}$$

  - m is the slope
  - b is the y-intercept

Formula

- In R we would express a simple linear model:

  ```
  Science_Score ~ 1 + Study_Hours
  ```

- R generates an estimate for the intercept (b_0) and coefficient/slope (b_1) of *Study_Hours* (X) on *Science_Score* (Y).

- To express this relationship as a *model formula:*

  $$Science\_Score = b_0 + b_1 Study\_Hours$$

Example 1

- Using the simulated *Scores* data, estimate the linear regression model:

  $$Science\_Score = b_0 + b_1 Study\_Hours$$

- Express the resulting model as a model formula.

Model Formula

- In design language: $Science\_Score = 1 + Study\_Hours$
- The *model formula* takes each term and multiplies it by a number.
  - These numbers are called **model coefficients** (not *slope*).
- The coefficients are estimated through **fitting the model to the data**.
- The coefficient results depend on the fitting process chosen and the data used.

Example 2

- Using the simulated *Scores* data, do the following:

1. Set your seed to 123
2. Use R to generate two subsets of the *Scores* simulated data ($n_1 = 70$, $n_2 = 80$).
3. Estimate the regression model from Example 1 using each of the subsets.
4. Using the model formulas, are the estimates the same?

Linear Models with Multiple Terms

- As we have already encountered, multiple variables may explain the variability in our response variable.

- **Multiple Linear Regression Model** (Expected value of Y):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon. \tag{2}$$

  - Y is the dependent variable.
  - β_0 is the intercept.
  - X_1, X_2, ..., X_k are the independent variables.
  - β_1, β_2, ..., β_k are the regression coefficients for the independent variables.
  - ϵ is the random error term.
    - Follows an assumed distribution with $E[\epsilon] = 0$ and constant variance σ_ϵ^2

Formula

- In R we would express a linear model:

  `Science_Score ~ 1 + Study_Hours + Entry_Exam`

- R generates an estimate for the intercept (b_0) and coefficients (b_1, b_2) of *Study_Hours* (X_1) and *Entry_Exam* (X_2) on *Science_Score* (Y).

- To express this relationship as a *model formula:*

  $$Science\_Score = b_0 + b_1 Study\_Hours + b_2 Entry\_Exam$$

Model Formula

- Now we will have multiple **model coefficients** that multiply to each term.
- Again, these coefficients are estimated through fitting the model to the data.
- To include multiple explanatory variables, simply add terms to the formula.

Example 3

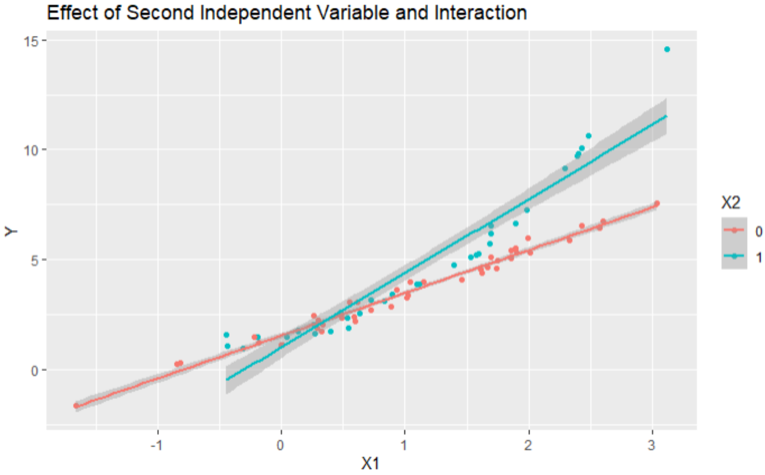
- Using the simulated *Scores* data, estimate the linear regression model:

$$Science\_Score = b_0 + b_1 Study\_Hours + b_2 Entry\_Exam$$

- Express the resulting model as a model formula.

Interaction Terms

- **Interaction terms** occur when one explanatory variable modulates the effect of another on the response variable.
- It does **NOT** refer to a relationship between two variables.
- You just have to remember to multiply the coefficient by the product of all the variables in the term.


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

Formula

- In R we would express a linear model with interaction terms:

  `Science_Score ~ 1 + Study_Hours + Entry_Exam + Study_Hours:Entry_Exam`

- R generates an estimate for the intercept (b_0) and coefficients (b_1, b_2, b_3) of *Study_Hours* (X_1) and *Entry_Exam* (X_2) and their interaction ($X_1 * X_2$) on *Science_Score* (Y).

- To express this relationship as a *model formula:*

  $$Science\_Score = b_0 + b_1 Study\_Hours + b_2 Entry\_Exam + b_3 Study\_Hours * Entry\_Exam$$

Example 4

- Using the simulated *Scores* data, estimate the linear regression model:

$$Science\_Score = b_0 + b_1 Study\_Hours + b_2 Entry\_Exam + b_3 Study\_Hours{*}Entry\_Exam$$

- Express the resulting model as a model formula.

Interpreting Interaction Terms

- Again, these results quantify how one variable impacts the effects of another variable on the dependent variable.
- From Example 4:
  - *The positive impact of studying longer is greater when the entry exam score is higher.*
- *This interpretation is under the assumption that we have properly estimated the model.*

- Quantitative (numeric) variables are naturally reflected in a model formula.
    - Multiply the value of the model term by the coefficient on that term.
- We use indicator variables to model the impacts of categorical variables.
- The coefficients express a change in dependent variable compared to the reference category.
    - The reference category is omitted from the model to prevent multicollinearity.

Formula

- In R we would express a linear model with interaction and categorical terms (assume *Province* has 3 levels):

  `Science_Score ~ 1 + Study_Hours + Entry_Exam + Study_Hours:Entry_Exam + Province`

- R generates an estimate for the intercept (b_0) and coefficients (b_1, b_2, b_3, b_4, b_5) of *Study_Hours* (X_1) and *Entry_Exam* (X_2), their interaction (X_1*X_2), and the categories *BC* & *Other* (X_3) on *Science_Score* (Y).

- To express this relationship as a *model formula:*

$$Science\_Score = b_0 + b_1 Study\_Hours + b_2 Entry\_Exam + b_3 Study\_Hours{*}Entry\_Exam + b_4 BC + b_5 Other$$

Example 5

- Using the simulated *Scores* data, estimate the linear regression model:

$$Science\_Score = b_0 + b_1 Study\_Hours + b_2 Entry\_Exam + b_3 Study\_Hours*Entry\_Exam + Province$$

- Express the resulting model as a model formula.

Effect Size

- An important step in statistical inference is to study the implied relationships found in your data.
- The **effect size** is the measurement of the size of a relationship is based on comparing changes.
- How does a one unit change in X_j change the value of Y.
- From Example 1:
  - *For every hour of studying completed, the science score increases by 5.9866*.

Effect Size of Categorical Variables

- For categorical variables, the coefficient on each level represents how much difference there is in the model value compared to the reference category.

- From Example 5:
  - *The science scores of students from BC are on average 3.5986 higher than those from Alberta.*
  - *The science scores of students from Other provinces are on average 3.7285 lower than those from Alberta (-3.7285).*

Residuals

- Unless the relationship is deterministic, the model values (fitted values) will not be exact match with the actual response variable in your data.
- The **residuals** show how far each observation is from its *model value*.
  - Residuals are always measured: *actual value minus fitted value*. OR
  - $y_i = \hat{y}_i + e_i$
- The residuals are likely to change every time we make adjustments to the model.

Explanatory Example I

- Import the *Wages.csv* dataset into R.

1. Use a linear model to determine the average *Salary*.
2. Next, include the categorical variable of the provinces and interpret the meaning of the coefficient estimates.
3. Suppress the inclusion of the intercept by using -1. Interpret the meaning of these coefficient estimates.

Explanatory Example II

- Using the *Wages.csv* dataset in R.

4. Create a simple linear regression model: $Salary \sim Experience$
   - Comment on the resulting intercept and slope.

5. Next, add the categorical variable of the provinces back into the model. Interpret the results.

6. Finally, include an interaction term between the *Experience* and *Province* variables.

Coefficients

- It is important to keep in mind that the coefficients have units.
- Generally, the units are not included when presenting the model formulas.
- Ignoring the units can be extremely misleading.
- The units of a slope: units of the response variable divided by the units of the explanatory variable.

Example 6

- Using the *Wages.csv* data estimate the following linear model:
  Salary $\sim 1 +$ GPA $+$ Experience
- Ignoring significance, which one of the variables has a larger impact on the Salary?

Correlation in Explanatory Variables

- Using formulas to describe models allows for the inclusion of multiple explanatory variables.
- Although it is theoretically better for explanatory variables to be completely independent, this is rarely the case.
- An effect attributed to one variable might equally well be assigned to some other variable.
- The way the *tangling* shows up is in the way the coefficient on a variable will change when another variable is added to the model or taken away from the model.
- **Very important to be aware of this, and to run appropriate diagnostics.**

Example 7

- Using the *Wages.csv* data estimate a linear with *Salary* as the dependent variable and all other variables as the explanatory variables.
- Are there any counter-intuitive coefficient estimates?

Simpson's Paradox

- **Simpson's Paradox**: *the coefficient on an explanatory variable can depend on what other explanatory variables have been included in the model.*
- As we can see in Example 7 there are some results that just do not make sense and this is due to the inclusion of all of the variables.
- *You can't look at explanatory variables in isolation; you have to interpret them in context.*

Why Linear Models?

- It can feel a bit unnatural to model complex relationships using a linear model.
- Linear models are very powerful tools and are often the chosen tool for many tasks.
  - Able to capture general linear relationships between multiple variables.
  - The results are easy to interpret and explain.
  - Often, the linear relationships are difficult to see when just plotting two variables at a time.
  - Start with main effects and add terms to improve the model.
- **There are situations where linear models will not work.**

Example 8

- Examine the *Results* section of the examples to see how linear models can be useful when examining multiple variables.

Exercise 1

- What exactly do the coefficients tell us?
- What happens when we add categorical variables to our linear model? (*Hint: Think about the intercept)*
- What happens when we add an interaction term to our model?

Exercise 2

- Using the *Wages.csv* dataset:
  - Estimate a linear model for *Salary* that contains at least one numeric and one categorical explanatory variable.
    - Express your results using a model formula and describe the meaning of all of the coefficients.
  - Estimate a linear model for *Salary* that contains at least one numeric, one categorical, and one interaction term as explanatory variables.
    - Express your results using a model formula and describe the meaning of all of the coefficients.

References & Resources

1. Kaplan, Daniel T. (2017). *Statistical Modelling: A Fresh Approach. (Second Edition).* Retrieved from https://dtkaplan.github.io/SM2-bookdown/

2. Fox, J. (2015). *Applied regression analysis and generalized linear models (Third Edition).* Sage Publications.

- https://cran.r-project.org/web/packages/interactions/vignettes/interactions.html