Multiple linear regression

With multiple predictors

Prof. Dr. Jan Kirenz

Some columns and the first 5 rows of the loans dataset.

	interest_rate	verified_income	debt_to_income	credit_util	bankruptcy	term	issue_month
0	14.07	Verified	18.01	38767	0	60	Mar-2018
1	12.61	Not Verified	5.04	4321	1	36	Feb-2018
2	17.09	Source Verified	21.15	16000	0	36	Feb-2018
3	6.72	Not Verified	10.16	4997	0	36	Jan-2018
4	14.07	Verified	57.96	52722	0	36	Mar-2018

Linear model for predicting interest rate based on whether the borrower has a **bankruptcy** in their record.

benkruptcy
- 0
- 1

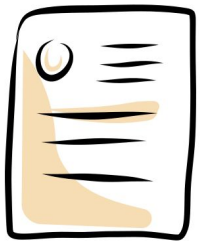
`bankruptcy`	An indicator variable for whether the borrower has a past bankruptcy in their record. This variable takes a value of `1` if the answer is *yes* and `0` if the answer is *no*.

$$\widehat{\texttt{interest\_rate}} = 12.34 + 0.74 \times \texttt{bankruptcy}$$

term	estimate
(Intercept)	12.34
bankruptcy1	0.74

Source: Çetinkaya-Rundel & Hardin (2021)

Lab: Jupyter Notebook 08a-1



Build the model

Question 8a-1:

Open quiz

$$\widehat{\texttt{interest\_rate}} = 12.34 + 0.74 \times \texttt{bankruptcy}$$

term	estimate
(Intercept)	12.34
bankruptcy1	0.74

Interpret the coefficients

Categorical predictors with **multiple levels**

- Categorical variable that has k levels where **k greater or equal 2**
- You only need coefficients for **k−1** of those levels (which one doesn't matter)
- For the last level that does not receive a coefficient, this is the **reference level**
- The coefficients listed for the other levels are all considered **relative to this reference level**.

Source: Çetinkaya-Rundel & Hardin (2021)

Categorical predictor with **three levels**

verified_income
- Not Verified
- Source Verified
- Verified

`verified_income`	Categorical variable describing whether the borrower's income source and amount have been verified, with levels `Verified`, `Source Verified`, and `Not Verified`.

term	estimate
`(Intercept)`	11.10
`verified_incomeSource Verified`	1.42
`verified_incomeVerified`	3.25

The "missing level" is called the **reference level** and it represents the default level (intercept) that other levels are measured against.

Example

verified_income
- Not Verified
- Source Verified
- Verified

Our linear regression model

$$\widehat{\texttt{interest\_rate}} = 11.10 + 1.42 \times \texttt{verified\_income}_{\texttt{Source Verified}} + 3.25 \times \texttt{verified\_income}_{\texttt{Verified}}$$

verified_income
- **Not Verified**
- Source Verified
- Verified

Model for a person without verified income (not verified)

$$\widehat{\texttt{interest\_rate}} = 11.10 + 1.42 \times 0 + 3.25 \times 0 = 11.10$$

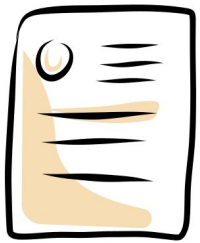
verified_income
- Not Verified
- **Source Verified**
- Verified

Model for a person with source verified

$$\widehat{\texttt{interest\_rate}} = 11.10 + 1.42 \times 1 + 3.25 \times 0 = 12.52$$

Source: Çetinkaya-Rundel & Hardin (2021)

Lab: Jupyter Notebook 08a-2



Build the model

Many predictors in a model

Multiple regression

$$\widehat{\texttt{interest\_rate}} = b_0 + b_1 \times \texttt{verified\_income}_{\texttt{Source Verified}} + b_2 \times \texttt{verified\_income}_{\texttt{Verified}} + b_3 \times \texttt{debt\_to\_income} + b_4 \times \texttt{credit\_util} + b_5 \times \texttt{bankruptcy} + b_6 \times \texttt{term} + b_9 \times \texttt{credit\_checks} + b_7 \times \texttt{issue\_month}_{\texttt{Jan-2018}} + b_8 \times \texttt{issue\_month}_{\texttt{Mar-2018}}$$

Source: Çetinkaya-Rundel & Hardin (2021)

We select values for b_0, b_1, … b_9 that minimize the sum of the squared residuals

$$SSE = e_1^2 + e_2^2 + \cdots + e_{10000}^2 = \sum_{i=1}^{10000} e_i^2 = \sum_{i=1}^{10000} (y_i - \hat{y}_i)^2$$

Source: Çetinkaya-Rundel & Hardin (2021)

Output for the regression model

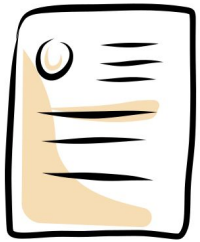
term	estimate
(Intercept)	1.89
verified_incomeSource Verified	1.00
verified_incomeVerified	2.56
debt_to_income	0.02
credit_util	4.90
bankruptcy1	0.39
term	0.15
credit_checks	0.23
issue_monthJan-2018	0.05
issue_monthMar-2018	-0.04

Source: Çetinkaya-Rundel & Hardin (2021)

Multiple regression model

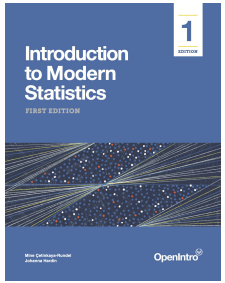
$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

Lab: Jupyter Notebook 08a-3



Build the model. Note that we don't use the predictor "credit_checks" in our model.

Resources



The content of this lecture is mainly based on the excellent book (can be accessed for free)

Introduction to Modern Statistics" by Mine Çetinkaya-Rundel and Johanna Hardin (2024)

https://openintro-ims.netlify.app/index.html