Multiple linear regression

With multiple predictors

Prof. Dr. Jan Kirenz
HdM Stuttgart

Indicator and categorical predictors

First six rows of the loans dataset.

interest_rate	verified_income	debt_to_income	credit_util	bankruptcy	term	credit
14.07	Verified	18.01	0.548	0	60	
12.61	Not Verified	5.04	0.150	1	36	
17.09	Source Verified	21.15	0.661	0	36	
6.72	Not Verified	10.16	0.197	0	36	
14.07	Verified	57.96	0.755	0	36	
6.72	Not Verified	6.46	0.093	0	36	

Variables and their descriptions for the loans dataset.

Variable	Description
`interest_rate`	Interest rate on the loan, in an annual percentage.
`verified_income`	Categorical variable describing whether the borrower's income source and amount have been verified, with levels `Verified`, `Source Verified`, and `Not Verified`.
`debt_to_income`	Debt-to-income ratio, which is the percentage of total debt of the borrower divided by their total income.
`credit_util`	Of all the credit available to the borrower, what fraction are they utilizing. For example, the credit utilization on a credit card would be the card's balance divided by the card's credit limit.
`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*.
`term`	The length of the loan, in months.
`issue_month`	The month and year the loan was issued, which for these loans is always during the first quarter of 2018.
`credit_checks`	Number of credit checks in the last 12 months. For example, when filing an application for a credit card, it is common for the company receiving the application to run a credit check.

Summary of a linear model for predicting interest rate based on whether the borrower has a bankruptcy in their record.

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

Interpret the coefficient for the past bankruptcy variable in the model.

term	estimate	std.error	statistic	p.value
`(Intercept)`	12.34	0.05	231.49	<0.0001
`bankruptcy1`	0.74	0.15	4.82	<0.0001

The variable takes one of two values: 1 when the borrower has a bankruptcy in their history and 0 otherwise. A slope of 0.74 means that the model predicts a 0.74% higher interest rate for those borrowers with a bankruptcy in their record.

Categorical predictor with three levels

term	estimate	std.error	statistic	p.value
`(Intercept)`	11.10	0.08	137.2	<0.0001
`verified_incomeSource Verified`	1.42	0.11	12.8	<0.0001
`verified_incomeVerified`	3.25	0.13	25.1	<0.0001

The “missing level” is called the reference level and it represents the default level that other levels are measured against.

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

Example

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

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

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

Categorical predictors with multiple levels

- Categorical variable that has k levels where $k>2$
- Software will provide a coefficient for $k-1$ of those levels.
- For the last level that does not receive a coefficient, this is the **reference level**, and the coefficients listed for the other levels are all considered relative to this reference level.

Many predictors in a model

Multiple regression

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

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

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

Output for the regression model

term	estimate	std.error	statistic	p.value
(Intercept)	1.89	0.21	9.01	<0.0001
verified_incomeSource Verified	1.00	0.10	10.06	<0.0001
verified_incomeVerified	2.56	0.12	21.87	<0.0001
debt_to_income	0.02	0.00	7.43	<0.0001
credit_util	4.90	0.16	30.25	<0.0001
bankruptcy1	0.39	0.13	2.96	0.0031
term	0.15	0.00	38.89	<0.0001
credit_checks	0.23	0.02	12.52	<0.0001
issue_monthJan-2018	0.05	0.11	0.42	0.6736
issue_monthMar-2018	-0.04	0.11	-0.39	0.696

Multiple regression model

$$\hat{y} = \boldsymbol{b}_0 + \boldsymbol{b}_1 \boldsymbol{x}_1 + \boldsymbol{b}_2 \boldsymbol{x}_2 + \cdots + \boldsymbol{b}_k \boldsymbol{x}_k$$

Adjusted R-squared

R-squared

$$R^2 = 1 - \frac{\text{variability in residuals}}{\text{variability in the outcome}} = 1 - \frac{Var(e_i)}{Var(y_i)}$$

- **Var** = variance (s^2)
- $\mathbf{e_i}$ = residuals of the model for observation i
- $\mathbf{y_i}$ = outcome for observation i

Problem: regular R^2 is a biased estimate of the amount of variability explained by the model when applied to model with more than one predictor.

Adjusted R-squared as a tool for model assessment.

$$R^2_{adj} = 1 - \frac{s^2_{\text{residuals}}/(n-k-1)}{s^2_{\text{outcome}}/(n-1)}$$
$$= 1 - \frac{s^2_{\text{residuals}}}{s^2_{\text{outcome}}} \times \frac{n-1}{n-k-1}$$

n: number of observations used to fit the model

k: number of predictor variables in the model.

Model selection

Common issue in multiple regression

- **Correlation** among predictor variables is not good.
- Two predictor variables are **collinear** (pronounced as co-linear) when they are correlated
- This "**multicollinearity**" complicates model estimation.

Full model vs parsimonious model

- **Full model:** model that includes all available predictors
- Often not desirable

- **Parsimonious model**
- A model that achieves a desired level of goodness of fit (R^2) using as few explanatory variables as possible

Stepwise selection

Backward elimination

- Starts with model that includes all potential predictor variables.
- Variables are eliminated one-at-a-time from the model until we cannot improve the model any further.

Forward selection

- We add variables one-at-a-time
- Until we cannot find any variables that improve the model any further.

Terms you should know

adjusted R-squared

backward elimination

degrees of freedom

forward selection

full model

multicollinearity

multiple regression

parsimonious

reference level

stepwise selection