Section 9.1: Introduction to Multiple Regression

Diez, D. M., Çetinkaya-Rundel, M., Barr, C. D. (2019). OpenIntro Statistics (4th ed.). OpenIntro. https://www.openintro.org/book/os/ CC BY-SA 3.0

STAT 1201
Introduction to Probability and Statistics

Chapter 9

Multiple and logistic regression

The principles of simple linear regression lay the foundation for more sophisticated regression models used in a wide range of challenging settings. In Chapter 9, we explore multiple regression, which introduces the possibility of more than one predictor in a linear model, and logistic regression, a technique for predicting categorical outcomes with two levels.

For videos, slides, and other resources, please visit
www.openintro.org/os

9.1 Introduction to multiple regression

Multiple regression extends simple two-variable regression to the case that still has one response but many predictors (denoted x_1, x_2, x_3, ...). The method is motivated by scenarios where many variables may be simultaneously connected to an output.

We will consider data about loans from the peer-to-peer lender, Lending Club, which is a data set we first encountered in Chapters 1 and 2. The loan data includes terms of the loan as well as information about the borrower. The outcome variable we would like to better understand is the interest rate assigned to the loan. For instance, all other characteristics held constant, does it matter how much debt someone already has? Does it matter if their income has been verified? Multiple regression will help us answer these and other questions.

The data set `loans` includes results from 10,000 loans, and we'll be looking at a subset of the available variables, some of which will be new from those we saw in earlier chapters. The first six observations in the data set are shown in Figure 9.1, and descriptions for each variable are shown in Figure 9.2. Notice that the past bankruptcy variable (`bankruptcy`) is an indicator variable, where it takes the value 1 if the borrower had a past bankruptcy in their record and 0 if not. Using an indicator variable in place of a category name allows for these variables to be directly used in regression. Two of the other variables are categorical (`income_ver` and `issued`), each of which can take one of a few different non-numerical values; we'll discuss how these are handled in the model in Section 9.1.1.

	interest_rate	income_ver	debt_to_income	credit_util	bankruptcy	term	issued	credit_checks
1	14.07	verified	18.01	0.55	0	60	Mar2018	6
2	12.61	not	5.04	0.15	1	36	Feb2018	1
3	17.09	source_only	21.15	0.66	0	36	Feb2018	4
4	6.72	not	10.16	0.20	0	36	Jan2018	0
5	14.07	verified	57.96	0.75	0	36	Mar2018	7
6	6.72	not	6.46	0.09	0	36	Jan2018	6
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 9.1: First six rows from the `loans` data set.

variable	description
`interest_rate`	Interest rate for the loan.
`income_ver`	Categorical variable describing whether the borrower's income source and amount have been verified, with levels `verified`, `source_only`, and `not`.
`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 her 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.
`issued`	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.

Figure 9.2: Variables and their descriptions for the `loans` data set.

9.1.1 Indicator and categorical variables as predictors

Let's start by fitting a linear regression model for interest rate with a single predictor indicating whether or not a person has a bankruptcy in their record:

$$\widehat{rate} = 12.33 + 0.74 \times bankruptcy$$

Results of this model are shown in Figure 9.3.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.3380	0.0533	231.49	<0.0001
bankruptcy	0.7368	0.1529	4.82	<0.0001
				$df = 9998$

Figure 9.3: Summary of a linear model for predicting interest rate based on whether the borrower has a bankruptcy in their record.

EXAMPLE 9.1

Interpret the coefficient for the past bankruptcy variable in the model. Is this coefficient significantly different from 0?

(E) The `bankruptcy` 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. (See Section 8.2.8 for a review of the interpretation for two-level categorical predictor variables.) Examining the regression output in Figure 9.3, we can see that the p-value for `bankruptcy` is very close to zero, indicating there is strong evidence the coefficient is different from zero when using this simple one-predictor model.

Suppose we had fit a model using a 3-level categorical variable, such as `income_ver`. The output from software is shown in Figure 9.4. This regression output provides multiple rows for the `income_ver` variable. Each row represents the relative difference for each level of `income_ver`. However, we are missing one of the levels: `not` (for *not verified*). The missing level is called the **reference level**, and it represents the default level that other levels are measured against.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.0995	0.0809	137.18	<0.0001
income_ver: *source_only*	1.4160	0.1107	12.79	<0.0001
income_ver: *verified*	3.2543	0.1297	25.09	<0.0001
				$df = 9998$

Figure 9.4: Summary of a linear model for predicting interest rate based on whether the borrower's income source and amount has been verified. This predictor has three levels, which results in 2 rows in the regression output.

EXAMPLE 9.2

How would we write an equation for this regression model?

(E) The equation for the regression model may be written as a model with two predictors:

$$\widehat{rate} = 11.10 + 1.42 \times \texttt{income\_ver}_{\texttt{source\_only}} + 3.25 \times \texttt{income\_ver}_{\texttt{verified}}$$

We use the notation $\texttt{variable}_{\texttt{level}}$ to represent indicator variables for when the categorical variable takes a particular value. For example, $\texttt{income\_ver}_{\texttt{source\_only}}$ would take a value of 1 if `income_ver` was `source_only` for a loan, and it would take a value of 0 otherwise. Likewise, $\texttt{income\_ver}_{\texttt{verified}}$ would take a value of 1 if `income_ver` took a value of `verified` and 0 if it took any other value.

The notation used in Example 9.2 may feel a bit confusing. Let's figure out how to use the equation for each level of the `income_ver` variable.

EXAMPLE 9.3

Using the model from Example 9.2, compute the average interest rate for borrowers whose income source and amount are both unverified.

When `income_ver` takes a value of `not`, then both indicator functions in the equation from Example 9.2 are set to zero:

E

$$\begin{aligned}\widehat{rate} &= 11.10 + 1.42 \times 0 + 3.25 \times 0 \\ &= 11.10\end{aligned}$$

The average interest rate for these borrowers is 11.1%. Because the `not` level does not have its own coefficient and it is the reference value, the indicators for the other levels for this variable all drop out.

EXAMPLE 9.4

Using the model from Example 9.2, compute the average interest rate for borrowers whose income source is verified but the amount is not.

When `income_ver` takes a value of `source_only`, then the corresponding variable takes a value of 1 while the other ($\texttt{income\_ver}_{\texttt{verified}}$) is 0:

E

$$\begin{aligned}\widehat{rate} &= 11.10 + 1.42 \times 1 + 3.25 \times 0 \\ &= 12.52\end{aligned}$$

The average interest rate for these borrowers is 12.52%.

GUIDED PRACTICE 9.5

G

Compute the average interest rate for borrowers whose income source and amount are both verified.[1]

PREDICTORS WITH SEVERAL CATEGORIES

When fitting a regression model with a 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.

[1] When `income_ver` takes a value of `verified`, then the corresponding variable takes a value of 1 while the other ($\texttt{income\_ver}_{\texttt{source\_only}}$) is 0:

$$\begin{aligned}\widehat{rate} &= 11.10 + 1.42 \times 0 + 3.25 \times 1 \\ &= 14.35\end{aligned}$$

The average interest rate for these borrowers is 14.35%.

GUIDED PRACTICE 9.6

Interpret the coefficients in the `income_ver` model.[2]

The higher interest rate for borrowers who have verified their income source or amount is surprising. Intuitively, we'd think that a loan would look *less* risky if the borrower's income has been verified. However, note that the situation may be more complex, and there may be confounding variables that we didn't account for. For example, perhaps lender require borrowers with poor credit to verify their income. That is, verifying income in our data set might be a signal of some concerns about the borrower rather than a reassurance that the borrower will pay back the loan. For this reason, the borrower could be deemed higher risk, resulting in a higher interest rate. (What other confounding variables might explain this counter-intuitive relationship suggested by the model?)

GUIDED PRACTICE 9.7

How much larger of an interest rate would we expect for a borrower who has verified their income source and amount vs a borrower whose income source has only been verified?[3]

9.1.2 Including and assessing many variables in a model

The world is complex, and it can be helpful to consider many factors at once in statistical modeling. For example, we might like to use the full context of borrower to predict the interest rate they receive rather than using a single variable. This is the strategy used in **multiple regression**. While we remain cautious about making any causal interpretations using multiple regression on observational data, such models are a common first step in gaining insights or providing some evidence of a causal connection.

We want to construct a model that accounts for not only for any past bankruptcy or whether the borrower had their income source or amount verified, but simultaneously accounts for all the variables in the data set: `income_ver`, `debt_to_income`, `credit_util`, `bankruptcy`, `term`, `issued`, and `credit_checks`.

$$\begin{aligned}\widehat{\texttt{rate}} = \beta_0 &+ \beta_1 \times \texttt{income\_ver}_{\texttt{source\_only}} + \beta_2 \times \texttt{income\_ver}_{\texttt{verified}} + \beta_3 \times \texttt{debt\_to\_income} \\ &+ \beta_4 \times \texttt{credit\_util} + \beta_5 \times \texttt{bankruptcy} + \beta_6 \times \texttt{term} \\ &+ \beta_7 \times \texttt{issued}_{\texttt{Jan2018}} + \beta_8 \times \texttt{issued}_{\texttt{Mar2018}} + \beta_9 \times \texttt{credit\_checks}\end{aligned}$$

This equation represents a holistic approach for modeling all of the variables simultaneously. Notice that there are two coefficients for `income_ver` and also two coefficients for `issued`, since both are 3-level categorical variables.

We estimate the parameters β_0, β_1, β_2, ..., β_9 in the same way as we did in the case of a single predictor. We select b_0, b_1, b_2, ..., b_9 that minimize the sum of the squared residuals:

$$SSE = e_1^2 + e_2^2 + \cdots + e_{10000}^2 = \sum_{i=1}^{10000} e_i^2 = \sum_{i=1}^{10000} (y_i - \hat{y}_i)^2 \tag{9.8}$$

where y_i and $\hat{y}_i$ represent the observed interest rates and their estimated values according to the model, respectively. 10,000 residuals are calculated, one for each observation. We typically use a computer to minimize the sum of squares and compute point estimates, as shown in the sample output in Figure 9.5. Using this output, we identify the point estimates b_i of each β_i, just as we did in the one-predictor case.

[2] Each of the coefficients gives the incremental interest rate for the corresponding level relative to the `not` level, which is the reference level. For example, for a borrower whose income source and amount have been verified, the model predicts that they will have a 3.25% higher interest rate than a borrower who has not had their income source or amount verified.

[3] Relative to the `not` category, the `verified` category has an interest rate of 3.25% higher, while the `source_only` category is only 1.42% higher. Thus, `verified` borrowers will tend to get an interest rate about 3.25%−1.42% = 1.83% higher than `source_only` borrowers.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.9251	0.2102	9.16	<0.0001
income_ver: *source_only*	0.9750	0.0991	9.83	<0.0001
income_ver: *verified*	2.5374	0.1172	21.65	<0.0001
debt_to_income	0.0211	0.0029	7.18	<0.0001
credit_util	4.8959	0.1619	30.24	<0.0001
bankruptcy	0.3864	0.1324	2.92	0.0035
term	0.1537	0.0039	38.96	<0.0001
issued: *Jan2018*	0.0276	0.1081	0.26	0.7981
issued: *Mar2018*	-0.0397	0.1065	-0.37	0.7093
credit_checks	0.2282	0.0182	12.51	<0.0001
				$df = 9990$

Figure 9.5: Output for the regression model, where `interest_rate` is the outcome and the variables listed are the predictors.

MULTIPLE REGRESSION MODEL

A multiple regression model is a linear model with many predictors. In general, we write the model as

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

when there are k predictors. We always estimate the β_i parameters using statistical software.

EXAMPLE 9.9

Write out the regression model using the point estimates from Figure 9.5. How many predictors are there in this model?

The fitted model for the interest rate is given by:

$$\widehat{\texttt{rate}} = 1.925 + 0.975 \times \texttt{income\_ver}_{\texttt{source\_only}} + 2.537 \times \texttt{income\_ver}_{\texttt{verified}} + 0.021 \times \texttt{debt\_to\_income} + 4.896 \times \texttt{credit\_util} + 0.386 \times \texttt{bankruptcy} + 0.154 \times \texttt{term} + 0.028 \times \texttt{issued}_{\texttt{Jan2018}} - 0.040 \times \texttt{issued}_{\texttt{Mar2018}} + 0.228 \times \texttt{credit\_checks}$$

If we count up the number of predictor coefficients, we get the *effective* number of predictors in the model: $k = 9$. Notice that the `issued` categorical predictor counts as two, once for the two levels shown in the model. In general, a categorical predictor with p different levels will be represented by $p - 1$ terms in a multiple regression model.

GUIDED PRACTICE 9.10

What does β_4, the coefficient of variable `credit_util`, represent? What is the point estimate of β_4?[4]

[4] β_4 represents the change in interest rate we would expect if someone's credit utilization was 0 and went to 1, all other factors held even. The point estimate is $b_4 = 4.90\%$.

EXAMPLE 9.11

Compute the residual of the first observation in Figure 9.1 on page 343 using the equation identified in Guided Practice 9.9.

(E) To compute the residual, we first need the predicted value, which we compute by plugging values into the equation from Example 9.9. For example, $\texttt{income\_ver}_{\texttt{source\_only}}$ takes a value of 0, $\texttt{income\_ver}_{\texttt{verified}}$ takes a value of 1 (since the borrower's income source and amount were verified), `debt_to_income` was 18.01, and so on. This leads to a prediction of $\widehat{rate}_1 = 18.09$. The observed interest rate was 14.07%, which leads to a residual of $e_1 = 14.07 - 18.09 = -4.02$.

EXAMPLE 9.12

We estimated a coefficient for `bankruptcy` in Section 9.1.1 of $b_4 = 0.74$ with a standard error of $SE_{b_1} = 0.15$ when using simple linear regression. Why is there a difference between that estimate and the estimated coefficient of 0.39 in the multiple regression setting?

(E) If we examined the data carefully, we would see that some predictors are correlated. For instance, when we estimated the connection of the outcome `interest_rate` and predictor `bankruptcy` using simple linear regression, we were unable to control for other variables like whether the borrower had her income verified, the borrower's debt-to-income ratio, and other variables. That original model was constructed in a vacuum and did not consider the full context. When we include all of the variables, underlying and unintentional bias that was missed by these other variables is reduced or eliminated. Of course, bias can still exist from other confounding variables.

Example 9.12 describes a common issue in multiple regression: correlation among predictor variables. We say the two predictor variables are **collinear** (pronounced as *co-linear*) when they are correlated, and this collinearity complicates model estimation. While it is impossible to prevent collinearity from arising in observational data, experiments are usually designed to prevent predictors from being collinear.

GUIDED PRACTICE 9.13

(G) The estimated value of the intercept is 1.925, and one might be tempted to make some interpretation of this coefficient, such as, it is the model's predicted price when each of the variables take value zero: income source is not verified, the borrower has no debt (debt-to-income and credit utilization are zero), and so on. Is this reasonable? Is there any value gained by making this interpretation?[5]

[5] Many of the variables do take a value 0 for at least one data point, and for those variables, it is reasonable. However, one variable never takes a value of zero: `term`, which describes the length of the loan, in months. If `term` is set to zero, then the loan must be paid back immediately; the borrower must give the money back as soon as she receives it, which means it is not a real loan. Ultimately, the interpretation of the intercept in this setting is not insightful.

9.1.3 Adjusted R^2 as a better tool for multiple regression

We first used R^2 in Section 8.2 to determine the amount of variability in the response that was explained by the model:

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

where e_i represents the residuals of the model and y_i the outcomes. This equation remains valid in the multiple regression framework, but a small enhancement can make it even more informative when comparing models.

GUIDED PRACTICE 9.14

G The variance of the residuals for the model given in Guided Practice 9.9 is 18.53, and the variance of the interest rate in all the loans is 25.01. Calculate R^2 for this model.[6]

This strategy for estimating R^2 is acceptable when there is just a single variable. However, it becomes less helpful when there are many variables. The regular R^2 is a biased estimate of the amount of variability explained by the model when applied to a new sample of data. To get a better estimate, we use the adjusted R^2.

ADJUSTED R^2 AS A TOOL FOR MODEL ASSESSMENT

The **adjusted R^2** is computed as

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

where n is the number of cases used to fit the model and k is the number of predictor variables in the model. Remember that a categorical predictor with p levels will contribute $p-1$ to the number of variables in the model.

Because k is never negative, the adjusted R^2 will be smaller – often times just a little smaller – than the unadjusted R^2. The reasoning behind the adjusted R^2 lies in the **degrees of freedom** associated with each variance, which is equal to $n-k-1$ for the multiple regression context. If we were to make predictions for *new data* using our current model, we would find that the unadjusted R^2 would tend to be slightly overly optimistic, while the adjusted R^2 formula helps correct this bias.

GUIDED PRACTICE 9.15

G There were $n = 10000$ loans in the `loans` data set and $k = 9$ predictor variables in the model. Use n, k, and the variances from Guided Practice 9.14 to calculate R^2_{adj} for the interest rate model.[7]

GUIDED PRACTICE 9.16

G Suppose you added another predictor to the model, but the variance of the errors $Var(e_i)$ didn't go down. What would happen to the R^2? What would happen to the adjusted R^2?[8]

Adjusted R^2 could have been used in Chapter 8. However, when there is only $k = 1$ predictors, adjusted R^2 is very close to regular R^2, so this nuance isn't typically important when the model has only one predictor.

[6] $R^2 = 1 - \frac{18.53}{25.01} = 0.2591$.

[7] $R^2_{adj} = 1 - \frac{18.53}{25.01} \times \frac{10000-1}{1000-9-1} = 0.2584$. While the difference is very small, it will be important when we fine tune the model in the next section.

[8] The unadjusted R^2 would stay the same and the adjusted R^2 would go down.