14.1: Introduction to Multiple Regression

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

We will consider Ebay auctions of a video game called Mario Kart for the Nintendo Wii. The outcome variable of interest is the total price of an auction, which is the highest bid plus the shipping cost. We will try to determine how total price is related to each characteristic in an auction while simultaneously controlling for other variables. For instance, all other characteristics held constant, are longer auctions associated with higher or lower prices? And, on average, how much more do buyers tend to pay for additional Wii wheels(plastic steering wheels that attach to the Wii controller) in auctions? Multiple regression will help us answer these and other questions.

The data set mario kart includes results from 141 auctions.[1] Four observations from this data set are shown in Table 14.1.1, and descriptions for each variable are shown in Table 14.1.2. Notice that the condition and stock photo variables are indicator variables. For instance, the cond new variable takes value 1 if the game up for auction is new and 0 if it is used. Using indicator variables in place of category names allows for these variables to be directly used in regression. See Section 7.2.7 for additional details. Multiple regression also allows for categorical variables with many levels, though we do not have any such variables in this analysis, and we save these details for a second or third course.

[1]*Diez DM, Barr CD, and Cetinkaya-Rundel M. 2012. openintro: OpenIntro data sets and supplemental functions. cran.r-project.org/web/packages/openintro.*

Table 14.1.1: Four observations from the mario kart data set.

	price	cond new	stock photo	duration	wheels
1	51.55	1	1	3	1
2	37.04	0	1	3	1
⋮	⋮	⋮	⋮	⋮	⋮
140	38.76	0	0	7	0
141	54.51	1	1	1	2

Table 14.1.2*: Variables and their descriptions for the mario kart data set.*

variable	description
price	final auction price plus shipping costs, in US dollars
cond_new	a coded two-level categorical variable, which takes value 1 when the game is new and 0 if the game is used
stock_photo	a coded two-level categorical variable, which takes value 1 if the primary photo used in the auction was a stock photo and 0 if the photo was unique to that auction
duration	the length of the auction, in days, taking values from 1 to 10
wheels	the number of Wii wheels included with the auction (a Wii wheel is a plastic racing wheel that holds the Wii controller and is an optional but helpful accessory for playing Mario Kart)

A Single-Variable Model for the Mario Kart Data

Let's fit a linear regression model with the game's condition as a predictor of auction price. The model may be written as

$$\hat{price} = 42.87 + 10.90 \times \text{cond\_ new} \tag{14.1.1}$$

Results of this model are shown in Table 14.1.3 and a scatterplot for price versus game condition is shown in Figure 14.1.4

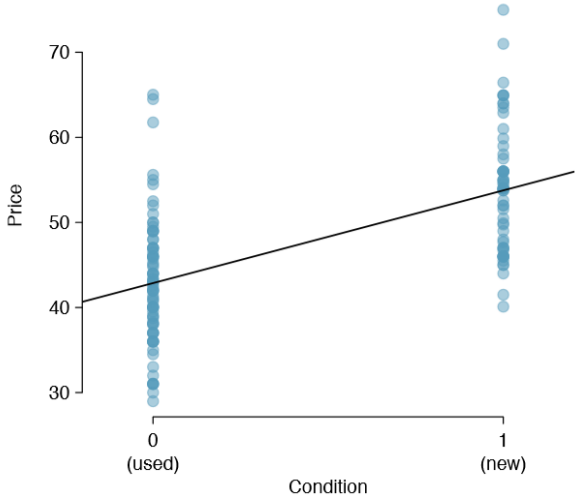

Figure 14.1.4: Scatterplot of the total auction price against the game's condition. The least squares line is also shown.

Table 14.1.3: Summary of a linear model for predicting auction price based on game condition.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	42.8711	0.8140	52.67	0.0000
cond_new	10.8996	1.2583	8.66	0.0000

Exercise 14.1.1

Figure 14.1.4 Does the linear model seem reasonable?

Answer

Yes. Constant variability, nearly normal residuals, and linearity all appear reasonable.

Exercise 14.1.2

Interpret the coefficient for the game's condition in the model. Is this coefficient significantly different from 0?

Note that cond new is a two-level categorical variable that takes value 1 when the game is new and value 0 when the game is used. So 10.90 means that the model predicts an extra $10.90 for those games that are new versus those that are used. (See Section 7.2.7 for a review of the interpretation for two-level categorical predictor variables.) Examining the regression output in Table 14.1.3, we can see that the p-value for cond new is very close to zero, indicating there is strong evidence that the coefficient is different from zero when using this simple one-variable model.

Including and Assessing Many Variables in a Model

Sometimes there are underlying structures or relationships between predictor variables. For instance, new games sold on Ebay tend to come with more Wii wheels, which may have led to higher prices for those auctions. We would like to fit a model that includes all potentially important variables simultaneously. This would help us evaluate the relationship between a predictor variable and the outcome while controlling for the potential influence of other variables. This is the strategy used in **multiple regression**. While we remain cautious about making any causal interpretations using multiple regression, such models are a common first step in providing evidence of a causal connection.

We want to construct a model that accounts for not only the game condition, as in Section 8.1.1, but simultaneously accounts for three other variables: stock photo, duration, and wheels.

$$\hat{price} = \beta_0 + \beta_1 \times \text{cond\_ new} + \beta_2 \times \text{stock\_ photo} + \beta_3 \times \text{duration} + \beta_4 \times \text{wheels} \tag{14.1.2}$$

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \tag{14.1.3}$$

In this equation, y represents the total price, x_1 indicates whether the game is new, x_2 indicates whether a stock photo was used, x_3 is the duration of the auction, and x_4 is the number of Wii wheels included with the game. Just as with the single predictor case, a multiple regression model may be missing important components or it might not precisely represent the relationship between the outcome and the available explanatory variables.

While no model is perfect, we wish to explore the possibility that this one may fit the data reasonably well.

We estimate the parameters $\beta_0, \beta_1, \ldots, \beta_4$ in the same way as we did in the case of a single predictor. We select $b_0, b_1, \ldots, b_4$ that minimize the sum of the squared residuals:

$$SSE = e_1^2 + e_2^2 + \cdots + e_{141}^2 = \sum_{i=1}^{141} e_i^2 = \sum_{i=1}^{141} (y_i - \hat{y}_i)^2 \tag{14.1.4}$$

Here there are 141 residuals, one for each observation. We typically use a computer to minimize the sum in Equation (8.4) and compute point estimates, as shown in the sample output in Table 14.1.5. Using this output, we identify the point estimates b_i of each β_i, just as we did in the one-predictor case.

Table 14.1.5: Output for the regression model where price is the outcome and cond_new, stock_photo, duration, and wheels are the predictors.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.2110	1.5140	23.92	0.0000
cond_new	5.1306	1.0511	4.88	0.0000
stock_photo	1.0803	1.0568	1.02	0.3085
duration	-0.0268	0.1904	-0.14	0.8882
wheels	7.2852	0.5547	13.13	0.0000

Multiple regression model

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

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{14.1.5}$$

when there are k predictors. We often estimate the β_i parameters using a computer.

Exercise 14.1.3

Write out the model in Equation (8.3) using the point estimates from Table 14.1.5. How many predictors are there in this model?[3]

Answer

$\hat{y} = 36.21 + 5.13x_1 + 1.08x_2 - 0.03x_3 + 7.29x_4$, and there are k = 4 predictor variables.

Exercise 14.1.4

What does β_4, the coefficient of variable x_4 (Wii wheels), represent? What is the point estimate of β_4?

Answer

It is the average difference in auction price for each additional Wii wheel included when holding the other variables constant. The point estimate is $b_4 = 7.29$.

Exercise 14.1.5

Compute the residual of the first observation in Table 14.1.1 on page 355 using the equation identified in Exercise 8.5.

Answer

$e_i = y_i - \hat{y}_i = 51.55 - 49.62 = 1.93$, where 49.62 was computed using the variables values from the observation and the equation identified in Exercise 14.1.3.

Example 14.1.1

We estimated a coefficient for cond new in Section 8.1.1 of $b_1 = 10.90$ with a standard error of $SE_{b1} = 1.26$ when using simple linear regression. Why might there be a difference between that estimate and the one in the multiple regression setting?

If we examined the data carefully, we would see that some predictors are correlated. For instance, when we estimated the connection of the outcome price and predictor cond new using simple linear regression, we were unable to control for other variables like the number of Wii wheels included in the auction. That model was biased by the confounding variable wheels. When we use both variables, this particular underlying and unintentional bias is reduced or eliminated (though bias from other confounding variables may still remain).

Example 14.1.1 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.

Exercise 14.1.6

The estimated value of the intercept is 36.21, 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: the game is used, the primary image is not a stock photo, the auction duration is zero days, and there are no wheels included. Is there any value gained by making this interpretation?

Solution

Three of the variables (cond new, stock photo, and wheels) do take value 0, but the auction duration is always one or more days. If the auction is not up for any days, then no one can bid on it! That means the total auction price would always be zero for such an auction; the interpretation of the intercept in this setting is not insightful.

Adjusted R^2 as a better estimate of explained variance

We first used R^2 in Section 7.2 to determine the amount of variability in the response that was explained by the model:

$$R^2 = 1 - \frac{\text{variability in residuals}}{\text{variability in the outcome}} = 1 - \frac{Var(e_i)}{Var(y_i)} \qquad (14.1.6)$$

where e_i represents the residuals of the model and yi the outcomes. This equation remains valid in the multiple regression framework, but a small enhancement can often be even more informative.

Exercise 14.1.7

The variance of the residuals for the model given in Exercise 8.7 is 23.34, and the variance of the total price in all the auctions is 83.06. Calculate R^2 for this model.

Solution

$$R^2 = 1 - \frac{23.34}{83.06} = 0.719\,.$$

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 actually a biased estimate of the amount of variability explained by the model. 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

$$R^2_{adj} = 1 - \frac{\frac{Var(e_i)}{(n-k-1)}}{\frac{Var(y_i)}{(n-1)}} = 1 - \frac{Var(e_i)}{Var(y_i)} \times \frac{n-1}{n-k-1} \quad (14.1.7)$$

where n is the number of cases used to fit the model and k is the number of predictor 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.

In multiple regression, the degrees of freedom associated with the variance of the estimate of the residuals is n - k - 1, not n - 1. For instance, if we were to make predictions for new data using our current model, we would nd that the unadjusted R^2 is an overly optimistic estimate of the reduction in variance in the response, and using the degrees of freedom in the adjusted R^2 formula helps correct this bias.

Exercise 14.1.8

There were n = 141 auctions in the mario_kart data set and k = 4 predictor variables in the model. Use n, k, and the variances from Exercise 8.10 to calculate R^2_{adj} for the Mario Kart model.[9]

Solution

$R^2_{adj} = 1 - \frac{23.34}{83.06} \times \frac{141-1}{141-4-1} = 0.711$.

Exercise 14.1.9

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?

Solution

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

This page titled 14.1: Introduction to Multiple Regression is shared under a CC BY-SA 3.0 license and was authored, remixed, and/or curated by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel via source content that was edited to the style and standards of the LibreTexts platform.

- **8.1: Introduction to Multiple Regression** by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel is licensed CC BY-SA 3.0. Original source: https://www.openintro.org/book/os.