14.2: Model Selection

The best model is not always the most complicated. Sometimes including variables that are not evidently important can actually reduce the accuracy of predictions. In this section we discuss model selection strategies, which will help us eliminate from the model variables that are less important. In this section, and in practice, the model that includes all available explanatory variables is often referred to as the **full model**. Our goal is to assess whether the full model is the best model. If it isn't, we want to identify a smaller model that is preferable.

Identifying Variables in the Model that may not be Helpful

Table 8.6 provides a summary of the regression output for the full model for the auction data. The last column of the table lists p-values that can be used to assess hypotheses of the following form:

- H_0: $\beta_i = 0$ when the other explanatory variables are included in the model.
- H_A: $\beta_i \neq 0$ when the other explanatory variables are included in the model.

Table 8.6: The fit for the full regression model, including the adjusted R^2.

	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

Example 14.2.1

The coefficient of cond new has a t test statistic of T = 4.88 and a p-value for its corresponding hypotheses ($H_0 : \beta_1 = 0, H_A : \beta_1 \neq 0$) of about zero. How can this be interpreted?

Solution

If we keep all the other variables in the model and add no others, then there is strong evidence that a game's condition (new or used) has a real relationship with the total auction price.

Example 14.2.2

Is there strong evidence that using a stock photo is related to the total auction price?

Solution

The t test statistic for stock photo is T = 1.02 and the p-value is about 0.31. After accounting for the other predictors, there is not strong evidence that using a stock photo in an auction is related to the total price of the auction. We might consider removing the stock photo variable from the model.

Exercise 14.2.1

Identify the p-values for both the duration and wheels variables in the model. Is there strong evidence supporting the connection of these variables with the total price in the model?

Answer

The p-value for the auction duration is 0.8882, which indicates that there is not statistically significant evidence that the duration is related to the total auction price when accounting for the other variables. The p-value for the Wii wheels variable is about zero, indicating that this variable is associated with the total auction price.

There is not statistically significant evidence that either the stock photo or duration variables contribute meaningfully to the model. Next we consider common strategies for pruning such variables from a model.

TIP: Using adjusted R^2 instead of p-values for model selection

The adjusted R^2 may be used as an alternative to p-values for model selection, where a higher adjusted R^2 represents a better model t. For instance, we could compare two models using their adjusted R^2, and the model with the higher adjusted R^2 would be preferred. This approach tends to include more variables in the final model when compared to the p-value approach.

Two model selection strategies

Two common strategies for adding or removing variables in a multiple regression model are called backward-selection and forward-selection. These techniques are often referred to as **stepwise** model selection strategies, because they add or delete one variable at a time as they "step" through the candidate predictors. We will discuss these strategies in the context of the p-value approach. Alternatively, we could have employed an R^2_{adj} approach.

The **backward-elimination** strategy starts with the model that includes all potential predictor variables. Variables are eliminated one-at-a-time from the model until only variables with statistically significant p-values remain. The strategy within each elimination step is to drop the variable with the largest p-value, re t the model, and reassess the inclusion of all variables.

Example 14.2.3

Results corresponding to the full model for the mario kart data are shown in Table 8.6. How should we proceed under the backward-elimination strategy?

Solution

There are two variables with coefficients that are not statistically different from zero: stock_photo and duration. We first drop the duration variable since it has a larger corresponding p-value, then we re t the model. A regression summary for the new model is shown in Table 8.7.

In the new model, there is not strong evidence that the coefficient for stock photo is different from zero, even though the p-value decreased slightly, and the other p-values remain very small. Next, we again eliminate the variable with the largest non-significant p-value, stock photo, and re t the model. The updated regression summary is shown in Table 8.8.

In the latest model, we see that the two remaining predictors have statistically significant coefficients with p-values of about zero. Since there are no variables remaining that could be eliminated from the model, we stop. The final model includes only the cond_new and wheels variables in predicting the total auction price:

$$\hat{y} = b_0 + b_1 x_1 + b_4 x_4 \quad (14.2.1)$$

$$= 36.78 + 5.58 x_1 + 7.23 x_4 \quad (14.2.2)$$

where x_1 represents cond new and x4 represents wheels.

An alternative to using p-values in model selection is to use the adjusted R^2. At each elimination step, we refit the model without each of the variables up for potential elimination. For example, in the first step, we would fit four models, where each would be missing a different predictor. If one of these smaller models has a higher adjusted R^2 than our current model, we pick the smaller model with the largest adjusted R^2. We continue in this way until removing variables does not increase R^2_{adj}. Had we used the adjusted R^2 criteria, we would have kept the stock photo variable along with the cond new and wheels variables.

Notice that the p-value for stock photo changed a little from the full model (0.309) to the model that did not include the duration variable (0.275). It is common for p-values of one variable to change, due to collinearity, after eliminating a different variable. This fluctuation emphasizes the importance of refitting a model after each variable elimination step. The p-values tend to change dramatically when the eliminated variable is highly correlated with another variable in the model.

The **forward-selection** strategy is the reverse of the backward-elimination technique. Instead of eliminating variables one-at-a-time, we add variables one-at-a-time until we cannot nd any variables that present strong evidence of their importance in the model.

Table 8.7: The output for the regression model where price is the outcome and the duration variable has been eliminated from the model.

	Estimate	Std. Error	t value	Pr(>|t|)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.0483	0.9745	36.99	0.0000
cond_new	5.1763	0.9961	5.20	0.0000
stock_photo	1.1177	1.0192	1.10	0.2747
wheels	7.2984	0.5448	13.40	0.0000

Table 8.8: The output for the regression model where price is the outcome and the duration and stock photo variables have been eliminated from the model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.7849	0.7066	52.06	0.0000
cond_new	5.5848	0.9245	6.04	0.0000
wheels	7.2328	0.5419	13.35	0.0000

Example 14.2.4: forward selection strategy

Construct a model for the mario kart data set using the forward selection strategy.

Solution

We start with the model that includes no variables. Then we t each of the possible models with just one variable. That is, we fit the model including just the cond new predictor, then the model including just the stock photo variable, then a model with just duration, and a model with just wheels. Each of the four models (yes, we fit four models!) provides a p-value for the coefficient of the predictor variable. Out of these four variables, the wheels variable had the smallest p-value. Since its p-value is less than 0.05 (the p-value was smaller than 2e-16), we add the Wii wheels variable to the model. Once a variable is added in forward-selection, it will be included in all models considered as well as the nal model.

Since we successfully found a first variable to add, we consider adding another. We fit three new models: (1) the model including just the cond_new and wheels variables (output in Table 8.8), (2) the model including just the stock photo and wheels variables, and (3) the model including only the duration and wheels variables. Of these models, the first had the lowest p-value for its new variable (the p-value corresponding to cond new was 1.4e-08). Because this p-value is below 0.05, we add the cond_new variable to the model. Now the final model is guaranteed to include both the condition and wheels variables.

We must then repeat the process a third time, fitting two new models: (1) the model including the stock photo, cond_new, and wheels variables (output in Table 8.7) and (2) the model including the duration, cond new, and wheels variables. The p-value corresponding to stock photo in the first model (0.275) was smaller than the p-value corresponding to duration in the second model (0.682). However, since this smaller p-value was not below 0.05, there was not strong evidence that it should be included in the model. Therefore, neither variable is added and we are finished.

The final model is the same as that arrived at using the backward-selection strategy.

Example 14.2.5: backward-selection strategy

As before, we could have used the R^2_{adj} criteria instead of examining p-values in selecting variables for the model. Rather than look for variables with the smallest p-value, we look for the model with the largest R^2_{adj}. What would the result of forward-selection be using the adjusted R^2 approach?

Solution

Using the forward-selection strategy, we start with the model with no predictors. Next we look at each model with a single predictor. If one of these models has a larger R^2_{adj} than the model with no variables, we use this new model. We repeat this procedure, adding one variable at a time, until we cannot nd a model with a larger R^2_{adj}. If we had done the forward-selection strategy using R^2_{adj}, we would have arrived at the model including cond new, stock photo, and wheels, which is a slightly larger model than we arrived at using the p-value approach and the same model we arrived at using the adjusted R^2 and backwards-elimination.

Model selection strategies

The backward-elimination strategy begins with the largest model and eliminates variables one-by-one until we are satis ed that all remaining variables are important to the model. The forward-selection strategy starts with no variables included in the model, then it adds in variables according to their importance until no other important variables are found.

There is no guarantee that the backward-elimination and forward-selection strategies will arrive at the same nal model using the p-value or adjusted R^2 methods. If the backwards-elimination and forward-selection strategies are both tried and they arrive at different models, choose the model with the larger R^2_{adj} as a tie-breaker; other tie-break options exist but are beyond the scope of this book.

It is generally acceptable to use just one strategy, usually backward-elimination with either the p-value or adjusted R^2 criteria. However, before reporting the model results, we must verify the model conditions are reasonable.
