Hypothesis testing in regression

To finish off this chapter, we will show how the permutation-testing framework can be used to answer questions about partial relationships in multiple regression modeling.

In a previous chapter, we asked the following question about houses in Saratoga, NY: what is the partial relationship between heating system type (gas, electric, or fuel oil) and sale price, once we adjust for the effect of living area, lot size, and the number of fireplaces? We fit a multiple regression model with these four predictors, which led to the following equation:

$$\begin{aligned}\text{Price} &= \$29868 + 105.3 \cdot \text{SqFt} + 2705 \cdot \log(\text{Acres}) + 7546 \cdot \text{Fireplaces} \\ &\quad - 14010 \cdot \mathbf{1}_{\{\text{fuel = electric}\}} - 15879 \cdot \mathbf{1}_{\{\text{fuel = oil}\}} + \text{Residual}\,.\end{aligned}$$

Remember that the baseline case here is gas heating, since it has no dummy variable. Our model estimated the premium associated with gas heating to be about $14,000 over electric heating, and about $16,000 over fuel-oil heating.

But are these differences due to heating-system type statistically significant, or could they be explained due to chance?

To answer this question, you could look at the confidence intervals for every coefficient associated with the heating-system variable, just as we learned to do in the chapter on multiple regression. The main difference is that before, we had one coefficient to look at, whereas now we have two: one dummy variable for fuel = electric, and one for fuel = oil. Two coefficients means two confidence intervals to look at.

Sometimes this strategy—that is, looking at the confidence intervals for all coefficients associated with a single variable—works just fine. For example, when the confidence intervals for all coefficients associated with a single variable are very far from zero, it's pretty obvious that the categorical variable in question is statistically significant.

But at other times, this strategy can lead to ambiguous results. In the context of the heating-system type variable, what if the 95% confidence interval for one dummy-variable coefficient contains zero, but the other doesn't? Or what if both confidence intervals contain zero, but just barely? Should we say that heating-system type is significant or not? This potential for ambiguous confidence intervals gets even worse when your categorical variable has more than just a few levels, because then there will be many more confidence intervals to look at.

The core of the difficulty here is that we want to assess the significance of the heating-system variable itself, not the significance of any individual *level* of that variable. To assess the significance of the whole variable, with all of its levels, we'll use a permutation test. Specifically, we will compare two models:

- The *full model*, which contains variables for square footage, lot size, number of fireplaces, and heating system.
- The *reduced model*, which contains variables for square footage, lot size, and number of fireplaces, but not for heating system. We say that the reduced model is *nested* within the full model, since it contains a subset of the variables in the full model, but no additional variables.

As always, we must start by specifying H_0. Loosely speaking, our null hypothesis is that the reduced model provides an adequate description of house prices, and that the full model is needlessly complex. To be a bit more precise: the null hypothesis is that *there is no partial relationship* between heating system and house prices, once we adjust for square footage, lot size, and number of fireplaces. This implies that all of the *true* dummy variable coefficients for heating-system type are zero.

Next, we must pick a test statistic. A natural way to assess the evidence against the null hypothesis is to use improvement in R^2 under the full model, compared to the reduced model. This is the same quantity we look at when assessing the importance of a variable in an ANOVA table. The idea is simple: if we see a big jump in R^2 when moving from the reduced to the full model, then the variable we added (here, heating system) is important for predicting the outcome, and the null hypothesis of no partial relationship is probably wrong.

You might wonder here: why not use the coefficients on the dummy variables for heating-system type as test statistics? The reason is that there are two such coefficients (or in general, $K-1$ coefficients for a categorical variable with K levels). But we need a single number to use as our test statistic in a permutation test. Therefore we use R^2: it is a single number that summarizes the predictive improvement of the full model over the reduced model.

Of course, even if we were to add a useless predictor to the reduced model, we would expect R^2 to go up, at least by a little bit, since the model would have more degrees of freedom (i.e. param-

Remember the four basic steps in a permutation test:

(1) Choose a null hypothesis H_0.

(2) Choose a test statistic Δ that is sensitive to departures from the null hypothesis.

(3) Repeatedly shuffle the predictor of interest and recalculate the test statistic after each shuffle, to approximate $P(\Delta \mid H_0)$, the sampling distribution of the test statistic T under the assumption that H_0 is true.

(4) Check whether the observed test statistic for your data, δ, is consistent with $P(\Delta \mid H_0)$.

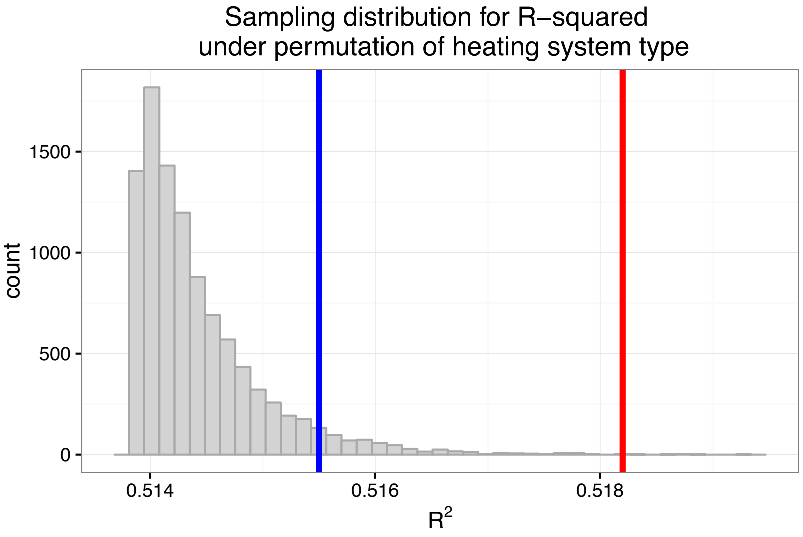

Figure 7.7: Sampling distribution of R^2 under the null hypothesis that there is no partial relationship between heating system and price after adjusting for effects due to square footage, lot size, and number of fireplaces. The blue vertical line marks the 95th percentile of the sampling distribution (and so corresponds to a rejection region at the 5% level). The red line marks the actual value of $R^2 = 0.518$ when we fit the full model by adding heating system to a model already containing the other three variables.

eters) that it can use to predict the observed outcome. Therefore, a more precise way of stating our null hypothesis is that, when we add heating system to a model already containing variables for square footage, lot size, and number of fireplaces, the improvement we see in R^2 could plausibly be explained by chance, even if this variable had no partial relationship with price.

To carry out a hypothesis test, we need to approximate the sampling distribution of R^2 under the null hypothesis. We will do so by repeatedly shuffling the heating system for every house (keeping all other variables the same), and re-fitting our model to each permuted data set. This breaks any partial relationship between heating system and price that may be present in our data. It tells us how big an improvement in R^2 we'd expect to see when fitting the full model, even if the null hypothesis were true.

This sampling distribution is shown in Figure 7.7, which was generating by fitting the model to 10,000 data sets in which the heating-system variable had been randomly shuffled, but where the response and the variables in the reduced model have been left alone. As expected, R^2 of the full model under permutation is always bigger than than the value of $R^2 = 0.513$ from the reduced model—but rarely by much. The blue line at $R^2 = 0.5155$ shows the 95th percentile of the sampling distribution (i.e. the critical value for a rejection region at the 5% level). The red line shows the actual value of $R^2 = 0.518$ from the full model fit the original data set (i.e. with no shuffling). This test statistic falls far beyond the 5% rejection region. We therefore reject the null hypothesis and conclude that there is statistically significant evidence for an effect on price due to heating-system type.

One key point here is that we shuffled *only* heating-system type—or in general, whatever variable is being tested. We don't shuffle the response or any of the other variables. That's because we are interested in a partial relationship between heating-system type and price. Partial relationships are always defined with respect to a specific context of other control variables, and we have to leave these control variables as they are in order to provide the correct context for that partial relationship to be measured.

To summarize: we can compare any two nested models using a permutation test based on R^2, regardless of whether the variable in question is categorical or numerical. To do so, we repeatedly shuffle the extra variable in the full model—without shuffling either the response or the control variables (i.e. those that also appear in the reduced model). We fit the full model to each shuffled data set, and we track the sampling distribution of R^2. We then compare this distribution with the R^2 we get when fitting the full model to the *actual* data set. If the actual R^2 is a lot bigger than what we'd expect under the sampling distribution for R^2 that we get under the permutation test, then we conclude that the extra variable in the full model is statistically significant.

F tests and the normal linear regression model. Most statistical software will produce an ANOVA table with an associated p-value for all variables. These p-values are approximations to the p-values that you'd get if you ran sequential permutation tests, adding and testing one variable at a time as you construct the ANOVA table. To be a bit more specific, they correspond to something called an F test under the normal linear regression model that we met awhile back:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + e_i \, , \quad e_i \sim N(0, \sigma^2) \, .$$

You might want to revisit the discussion of the normal linear regression model starting on page 120. But the upshot is that an F test is conceptually similar to a permutation test based on R^2—and if you're happy with the assumption of normally distributed residuals, you can treat the p-values from these two tests as virtually interchangeable.[8]

[8] If you're not happy with this assumption, then you're better off with the permutation test.