6
Multiple regression

From lines to planes

Linear regression, as we've learned, is a powerful tool for finding patterns in data. So far, we've only considered models that involve a single numerical predictor, together with as many grouping variables as we want. These grouping variables were allowed to modulate the intercept, or both the slope and intercept, of the underlying relationship between the numerical predictor (like SAT score) and the response (like GPA). This allowed us to fit different lines to different groups, all within the context of a single regression equation.

In this chapter, we learn how to build more complex models that incorporate two or more numerical predictors. For example, consider the data in Figure 6.1 on page 136, which shows the highway gas mileage versus engine displacement (in liters) and weight (in pounds) for 59 different sport-utility vehicles.[1] The data points in the first panel are arranged in a three-dimensional point cloud, where the three coordinates (x_{i1}, x_{i2}, y_i) for vehicle i are:

- x_{i1}, engine displacement, increasing from left to right.
- x_{i2}, weight, increasing from foreground to background.
- y_i, highway gas mileage, increasing from bottom to top.

Since it can be hard to show a 3D cloud of points on a 2D page, a color scale has been added to encode the height of each point in the y direction.

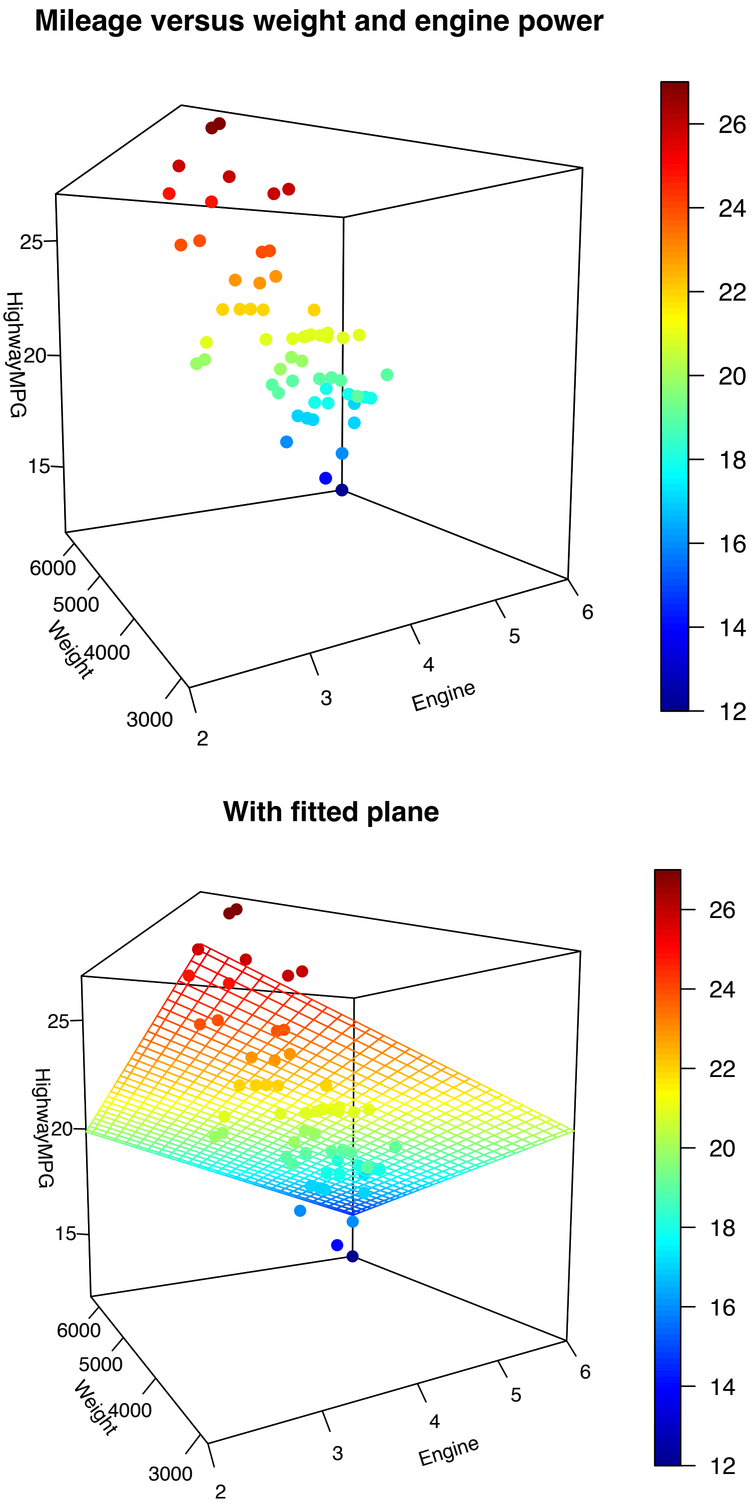
Fitting a linear equation for y versus x_1 and x_2 results in a regression model of the following form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i \, .$$

Just as before, we call the β's the coefficients of the model and the e_i's the residuals. In Figure 6.1, this fitted equation is

$$\text{MPG} = 33 - 1.35 \cdot \text{Displacement} - 0.00164 \cdot \text{Weight} + \text{Residual} \, .$$

[1] These are the same SUVs shown in the second-from-right panel in FIgure 4.9, when we discussed ANOVA for models involving correlated predictors.



Mileage versus weight and engine power

With fitted plane

Figure 6.1: Highway gas mileage versus weight and engine displacement for 59 SUVs, with the least-squares fit shown in the bottom panel.

Both coefficients are negative, showing that gas mileage gets worse with increasing weight and engine displacement.

This equation is called a *multiple regression model*. In geometric terms, it describes a plane passing through a three-dimensional cloud of points, which we can see slicing roughly through the middle of the points in the bottom panel in Figure 6.1. This plane has a similar interpretation as the line did in a simple one-dimensional linear regression. If you read off the height of the plane along the y axis, then you know where the response variable is expected to be, on average, for a particular pair of values (x_1, x_2). We use the term *predictor space* to mean all possible combinations of the predictor variables.

In more than two dimensions. In principle, there's no reason to stop at two predictors. We can easily generalize this idea to fit regression equations using p different predictors $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,p})$:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} = \beta_0 + \sum_{k=1}^{p} \beta_k x_{i,k} \, .$$

This is the equation of a p-dimensional plane embedded in $(p+1)$-dimensional space. This plane is nearly impossible to visualize beyond $p = 2$, but straightforward to describe mathematically.

We use a bolded $\mathbf{x}_i$ as shorthand to denote the whole vector of predictor values for observation i. That way we don't have to write out $(x_{i,1}, x_{i,2}, \dots, x_{i,p})$ every time. When writing things out by hand, a little arrow can be used instead, since you obviously can't write things in bold: $\vec{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,p})$. By the same logic, we also write $\vec{\beta}$ for the vector $(\beta_0, \beta_1, \dots, \beta_p)$.

From simple to multiple regression: what stays the same. In this jump from the familiar (straight lines in two dimensions) to the foreign (planes in arbitrary dimensions), it helps to start out by cataloguing several important features that don't change.

First, we still fit parameters of the model using the principle of least squares. As before, we will denote our estimates by $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$, and so on. For a given choice of these coefficients, and a given point in predictor space, the fitted value of y is

$$\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i,1} + \widehat{\beta}_2 x_{i,2} + \cdots + \widehat{\beta}_p x_{i,p} \, .$$

This is a scalar quantity, even though the regression parameters describe a p-dimensional hyperplane. Therefore, we can define the residual sum of squares in the same way as before, as the sum of squared differences between fitted and observed values:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left\{ y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_{i,1} + \widehat{\beta}_2 x_{i,2} + \cdots + \widehat{\beta}_p x_{i,p}) \right\}^2 \, .$$

The principle of least squares prescribes that we should choose the estimates so as to make the residual sum of squares as small as possible, thereby distributing the "misses" among the observations in a roughly equal fashion. Just as before, the little e_i is the amount by which the fitted plane misses the actual observation y_i.

Second, these residuals still have the same interpretation as before: as the part of y that is unexplained by the predictors. For a least-squares fit, the residuals will be uncorrelated with each of the original predictors. Thus we can interpret $e_i = y_i - \hat{y}_i$ as a statistically adjusted quantity: the y variable, adjusted for the systematic relationship between y and all of the x's in the regression equation. Here, as before, statistical adjustment just means subtraction.

Third, we still summarize preciseness of fit using R^2, which has the same definition as before:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{UV}{TV} = \frac{PV}{TV}\,.$$

The only difference is that $\hat{y}_i$ is now a function of more than just an intercept and a single slope. Also, just as before, it will still be the case R^2 is the square of the correlation coefficient between y_i and $\hat{y}_i$. It will not, however, be expressible as the correlation between y and any of the original predictors, since we now have more than one predictor to account for. (Indeed, R^2 is a natural generalization of Pearson's r for measuring correlation between one response and a whole basket of predictors.)

Fourth, it remains important to respect the distinction between the true model parameters (β_0, β_1, and so forth) and the estimated parameters ($\widehat{\beta}_0$, $\widehat{\beta}_1$ and so forth). When using the multiple regression model, we imagine that there is some true hyperplane described by β_0 through β_p, and some true residual standard deviation σ_e, that gave rise to our data. We can infer what those parameters are likely to be on the basis of observed data, but we can never know their values exactly.

Multiple regression and partial relationships

Not everything about our inferential process stays the same when we move from lines to planes. We will focus more on some of the differences later, but for now, we'll mention a major one: the interpretation of each β coefficient is no longer quite so simple as the interpretation of the slope in one-variable linear regression.

The best way to think of $\widehat{\beta}_k$ is as an estimated *partial slope*: that is, the change in y associated with a one-unit change in x_k, holding all other variables constant. This is a subtle interpretation that is worth considering at length. To understand it, it helps to isolate the contribution of x_k on the right-hand side of the regression equation. For example, suppose we have two numerical predictors, and we want to interpret the coefficient associated with x_2. Our equation is

$$\underbrace{y_i}_{\text{Response}} = \beta_0 + \underbrace{\beta_1 x_{i1}}_{\text{Effect of } x_1} + \underbrace{\beta_2 x_{i2}}_{\text{Effect of } x_2} + \underbrace{e_i}_{\text{Residual}} \, .$$

To interpret the effect of the x_2 variable, we isolate that part of the equation on the right-hand side, by subtracting the contribution of x_1 from both sides:

$$\underbrace{y_i - \beta_1 x_{i1}}_{\text{Response, adjusted for } x_1} = \underbrace{\beta_0 + \beta_2 x_{i2}}_{\text{Regression on } x_2} + \underbrace{e_i}_{\text{Residual}} \, .$$

On the left-hand side, we have something familiar from one-variable linear regression: the y variable, adjusted for the effect of x_1. If it weren't for the x_2 variable, this would just be the residual in a one-variable regression model. Thus we might call this term a *partial residual*.

On the right-hand side we also have something familiar: an ordinary one-dimensional regression equation with x_2 as a predictor. We know how to interpret this as well: the slope of a linear regression quantifies the change of the left-hand side that we expect to see with a one-unit change in the predictor (here, x_2). But here the left-hand side isn't y; it is y, adjusted for x_1. We therefore conclude that β_2 is the change in y, *once we adjust for the changes in y due to x_1*, that we expect to see with a one-unit change in the x_2 variable.

This same line of reasoning can allow us to interpret β_1 as well:

$$\underbrace{y_i - \beta_2 x_{i2}}_{\text{Response, adjusted for } x_2} = \underbrace{\beta_0 + \beta_1 x_{i1}}_{\text{Regression on } x_1} + \underbrace{e_i}_{\text{Residual}} \, .$$

Thus β_1 is the change in y, *once we adjust for the changes in y due to x_2*, that we expect to see with a one-unit change in the x_1 variable.

We can make the same argument in any multiple regression model involving two or more predictors, which we recall takes the form

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{i,k} + e_i \, .$$

To interpret the coefficient on the jth predictor, we isolate it on the right-hand side:

$$\underbrace{y_i - \sum_{k \neq j} \beta_k x_{i,k}}_{\text{Response adjusted for all other } x\text{'s}} = \underbrace{\beta_0 + \beta_j x_{ij}}_{\text{Regression on } x_j} + \underbrace{e_i}_{\text{Residual}} .$$

Thus β_j represents the rate of change in y associated with one-unit change in x_j, after adjusting for all the changes in y that can be predicted by the other predictor variables.

Partial versus overall relationships. A multiple regression equation isolates a set of *partial relationships* between y and each of the predictor variables. By a partial relationship, we mean the relationship between y and a single variable x, holding other variables constant. The partial relationship between y and x is very different than the *overall relationship* between y and x, because the latter ignores the effects of the other variables. When the two predictor variables are correlated, this difference matters a great deal.

To compare these two types of relationships, let's take the multiple regression model we fit to the data on SUVs in Figure 6.1:

$$\text{MPG} = 33 - 1.35 \cdot \text{Displacement} - 0.00164 \cdot \text{Weight} + \text{Residual} .$$

This model isolates two partial relationships:

- We expect highway gas mileage to decrease by 1.35 MPG for every 1-liter increase in engine displacement, after adjusting for the simultaneous effect of vehicle weight on mileage. That is, if we held weight constant and increased the engine size by 1 liter, we'd expect mileage to go down by 1.35 MPG.

- We expect highway gas mileage to decrease by 1.64 MPG for every additional 1,000 pounds of vehicle weight, after adjusting for the simultaneous effect of engine displacement on gas mileage. That is, if we held engine displacement constant and added 1,000 pounds of weight to an SUV, we'd expect mileage to go down by 1.64 MPG.

Let's compare these partial relationships with the overall relationships depicted in Figure 6.2. Here we've fit two separate one-variable regression models: mileage versus engine displacement on the left, and mileage versus vehicle weight on the right.

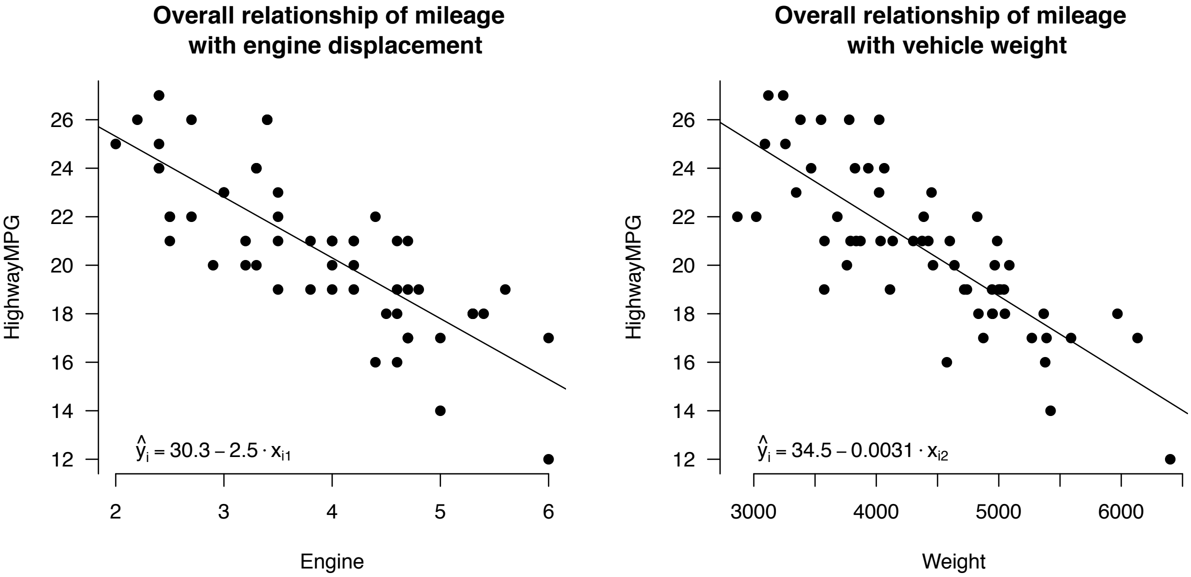


Figure 6.2: Overall relationships for highway gas mileage versus weight and engine displacement individually.

Focus on the left panel of Figure 6.2 first. The least-squares fit to the data is

$$\text{MPG} = 30.3 - 2.5 \cdot \text{Displacement} + \text{Residual}\,.$$

Thus when displacement goes up by 1 liter, we expect mileage to go down by 2.5 MPG. This overall slope is quite different from the partial slope of -1.35 isolated by the multiple regression equation. That's because this model doesn't attempt to adjust for the effects of vehicle weight. Because weight is correlated with engine displacement, we get a steeper estimate for the overall relationship than for the partial relationship: for cars where engine displacement is larger, weight also tends to be larger, and the corresponding effect on the y variable isn't controlled for in the left panel.

Similarly, the overall relationship between mileage and weight is

$$\text{MPG} = 34.5 - 0.0031 \cdot \text{Weight} + \text{Residual}\,.$$

The overall slope of -0.0031 is nearly twice as steep the partial slope of -0.00164. The one-variable regression model hasn't successfully isolated the marginal effect of increased weight from that of increased engine displacement. But the multiple regression model has—and once we hold engine displacement constant, the marginal effect of increased weight on mileage looks smaller.

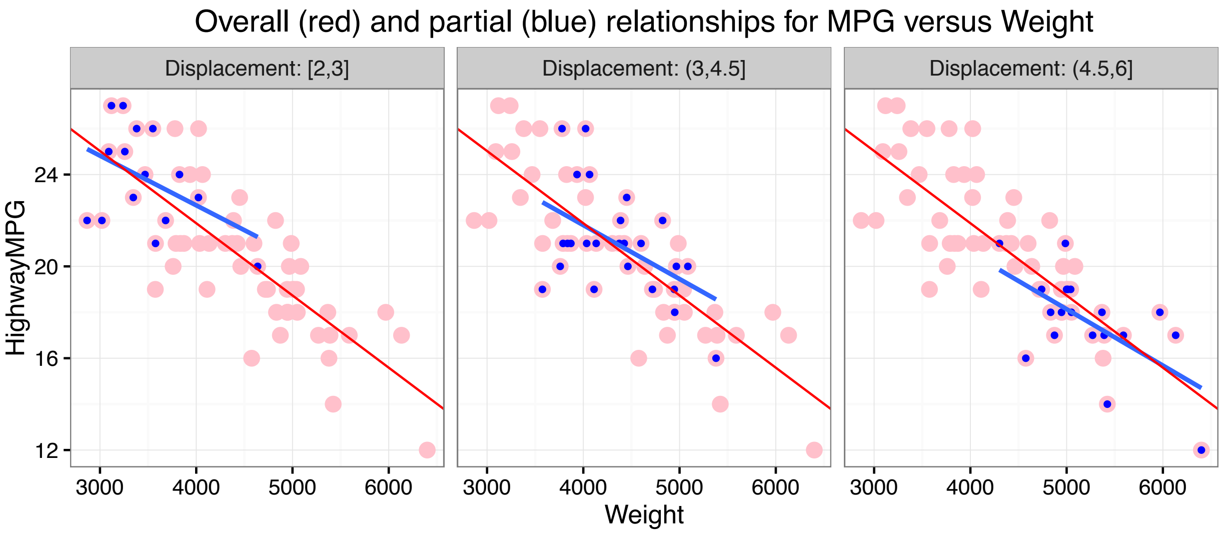


Figure 6.3: A lattice plot of mileage versus weight, stratified by engine displacement. The blue points within each panel show only the SUVs within a specific range of engine displacements: ≤ 3 liters on the left, 3–4.5 liters in the middle, and > 4.5 liters on the right. The blue line shows the least-squares fit to the blue points alone within each panel. For reference, the entire data set is also shown in each panel (pink dots), together with the overall fit (red line) from the right-hand side of Figure 6.2. The blue lines are shallower than the red line, suggesting that once we hold engine displacement approximately (thought not perfectly) constant, we estimate a different (less steep) relationship between mileage and weight.

Figure 6.3 provides some intuition here about the difference between an overall and a partial relationship. The figure shows a lattice plot where the panels correspond to different strata of engine displacement: 2–3 liters, 3–4.5 liters, and 4.5–6 liters. Within each stratum, engine displacement doesn't vary by much—that is, it is approximately held constant. Each panel in the figure shows a straight line fit that is specific to the SUVs in each stratum (blue dots and line), together with the overall linear fit to the whole data set (red dots and line).

The two important things to notice here are the following.

(1) The SUVs within each stratum of engine displacement are in systematically different parts of the x–y plane. For the most part, the smaller engines are in the upper left, the middle-size engines are in the middle, and the bigger engines are in the bottom right. When weight varies, displacement also varies, and each of these variables have an effect on mileage. Another way of saying this is that engine displacement is a *confounding variable* for the relationship between mileage and weight. A confounder is something that is correlated with both the predictor and response.

(2) In each panel, the blue line has a shallower slope than the red line. That is, when we compare SUVs that are similar in engine displacement, the mileage–weight relationship is not as steep

as it is when we compare SUVs with very different engine displacements.

This second point—that when we hold displacement roughly constant, we get a shallower slope for mileage versus weight—explains why the partial relationship estimated by the multiple regression model is different than the overall relationship from the left panel of Figure 6.2. The slope of -1.64×10^{-3} MPG per pound from the multiple regression model addresses the question: how fast should we expect mileage to change when we compare SUVs with different weights, but with the same engine displacement? This is similar to the question answered by the blue lines in Figure 6.3, but different than the question answer by the red line.

It is important to keep in mind that this "isolation" or "adjustment" is statistical in nature, rather than experimental. Most real-world systems simply don't have isolated variables. Confounding tends to be the rule, rather than the exception. The only real way to isolate a single factor is to run an experiment that actively manipulates the value of one predictor, holding the others constant, and to see how these changes affect y. Still, using a multiple-regression model to perform a statistical adjustment is often the best we can do when facing questions about partial relationships that, for whatever reason, aren't amenable to experimentation.

Further issues in multiple regression

In this section, we will address four important issues that arise frequently in multiple regression, all of which involve extensions of ideas we've already covered:

- Checking the assumption of linearity.
- Using the bootstrap to quantify uncertainty.
- Incorporating grouping variables in multiple regression.
- Using permutation tests to determine whether adding a variable produces a statistically significant improvement in the overall fit of the model.

Throughout this section, we'll use a running example of a data set on house prices from Saratoga County, New York, distributed as part of the `mosaic` R package. We'll use this data set to address a few interesting questions of the kind that might be relevant to anyone buying, selling, or assessing the taxable value of a house.

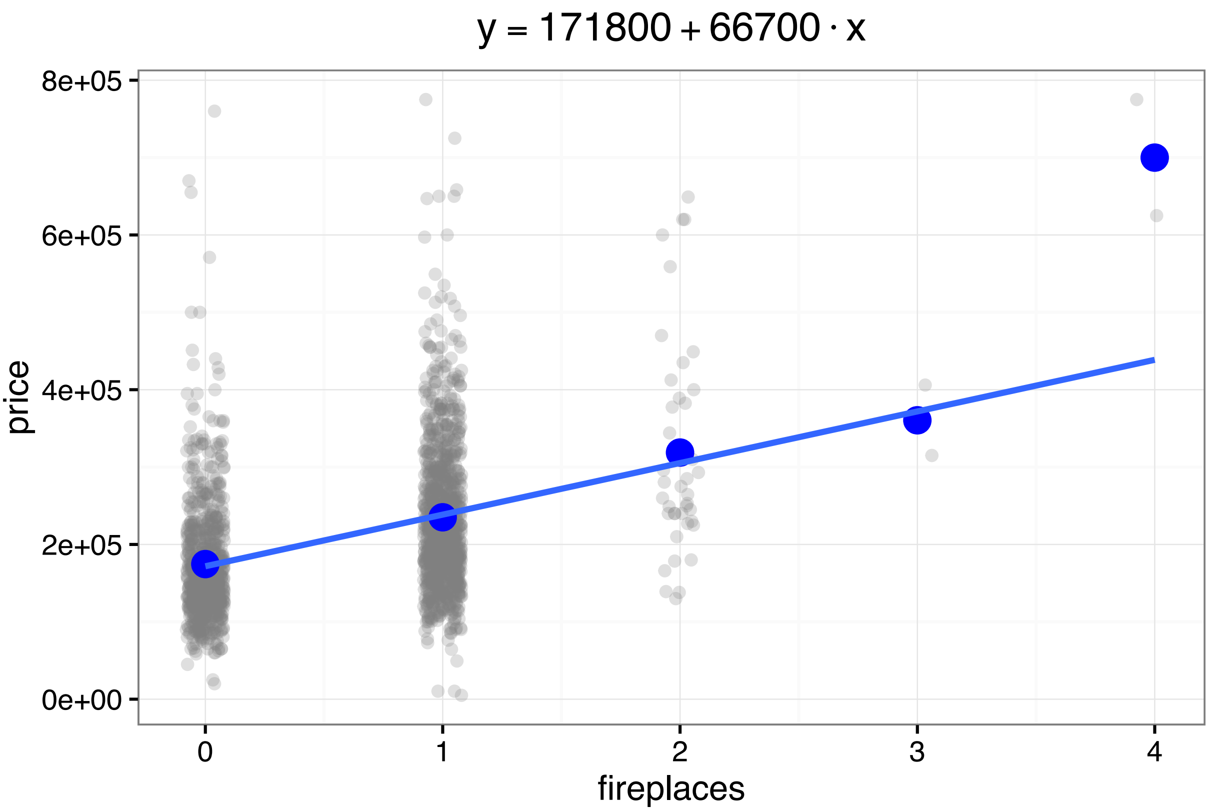


Figure 6.4: The relationship between the price of a house and the number of fireplaces it has.

How much is a fireplace worth?

How much does a fireplace improve the value of a house for sale? Figure 6.4 would seem to say: by about \$66,700 per fireplace. This dot plot shows the sale price of houses in Saratoga County, NY that were on the market in 2006.[2] We have fit a linear regression of for house price versus number of fireplaces, leading to the equation

$$\text{Price} = \$171800 + 66,700 \cdot \text{Fireplaces} + \text{Residual}\,,$$

This fitted equation is shown as a blue line in Figure 6.4. The means of the individual groups (1 fireplace, 2 fireplaces, etc) are also shown as blue dots. This helps us to verify that the assumption of linearity is reasonable here: the line passes almost right through the group means, except the one for houses with four fireplaces (which corresponds to just two houses).

But before you go knocking a hole in your ceiling and hiring a

[2] Data from "House Price Capitalization of Education by Part Year Residents," by Candice Corvetti. Williams College honors thesis, 2007, available here, and in the mosaic R package.

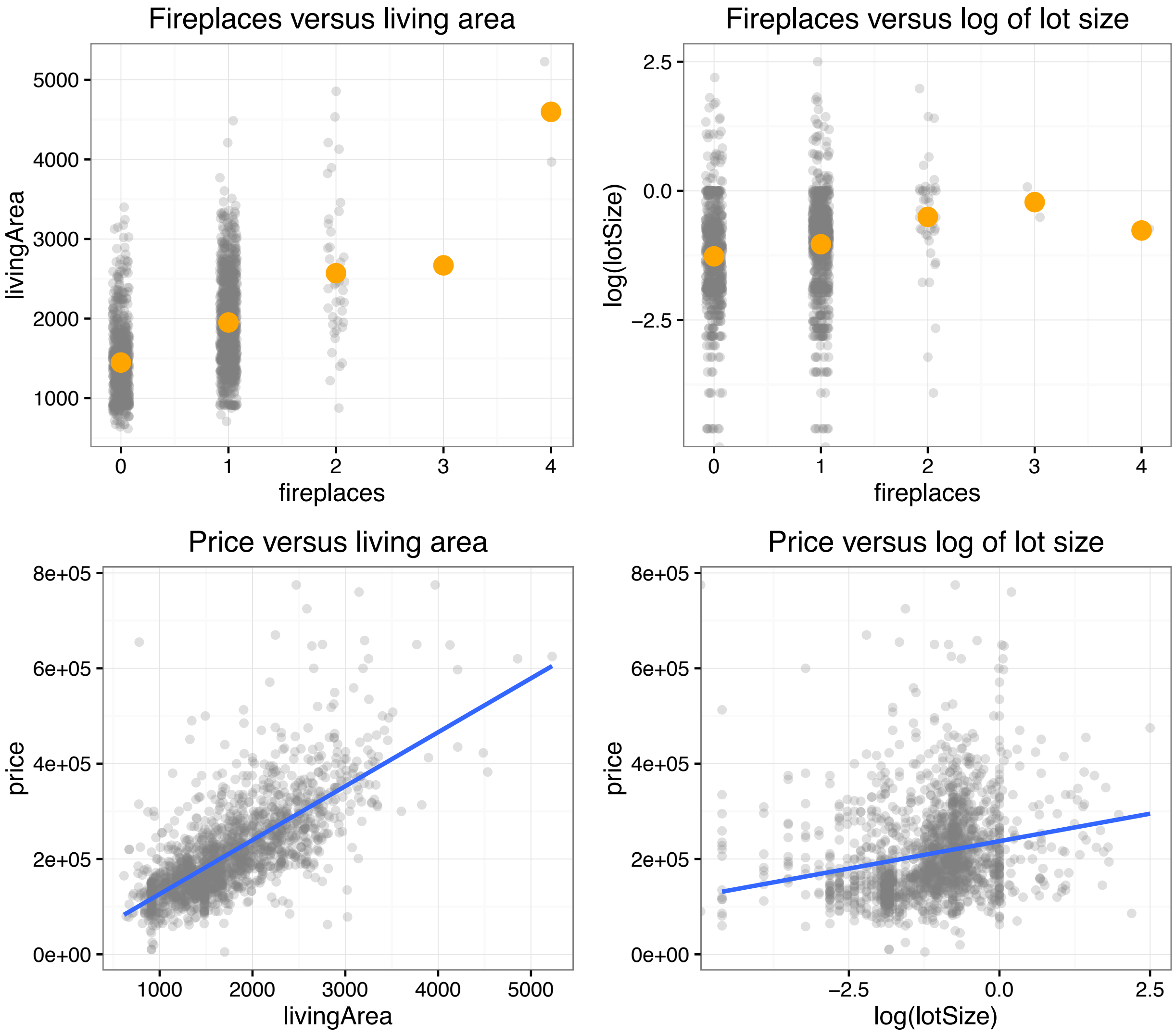


Figure 6.5: The relationship of house price with living area (bottom left) and with the logarithm of lot size in acres (bottom right). Both of these variables are potential confounders for the relationship between fireplaces and price, because they are also correlated with the number of fireplaces (top row).

bricklayer so that you might cash in on your new fireplace, consult Figure 6.5 on page 145. This figure shows that we should be careful in interpreting the figure of $66,700 per fireplace arising from the simple one-variable model. Specifically, it shows that houses with more fireplaces also tend to be bigger (top left panel) and to sit on lots that have more land area (top right). These factors are also correlated with the price of a house. In light of this potential for confounding, we have two possible explanations for the relationship we see in Figure 6.4. This correlation may happen because fireplaces are so valuable. On the other hand, it may also happen because fireplaces happen to occur more frequently in houses that are desireable for other reasons (i.e. they are bigger).

Disentangling these two possibilities requires estimating the partial relationship between fireplaces and prices, rather than the overall relationship shown in Figure 6.4. After all, when someone like a realtor or the county tax assessor wants to know how much a fireplace is worth, what they probably want to know is: how much is a fireplace worth, holding other relevant features of the house constant?

To address this question, we can fit a multiple regression model for price versus living area, lot size, and number of fireplaces. This will allow us to estimate the partial relationship between fireplaces and price, holding square footage and lot size constant. Such a model can tell us how much more we should expect a house with a fireplace to be worth, compared to a house that is identical in size and acreage but without a fireplace.

Fitting such a model to the data from Saratoga County yields the following equation:

$$\text{Price} = \$17787 + 108.3 \cdot \text{SqFt} + 1257 \cdot \log(\text{Acres}) + 8783 \cdot \text{Fireplaces} + \text{Residual}\,. \tag{6.1}$$

According to this model, the value of one extra fireplace is about $8,783, holding square footage and lot size constant. This is a much lower figure than the $66,700 fireplace premium that we would naïvely estimate from the overall relationship in Figure 6.4.

Model checking. Is the assumption of a linear regression model appropriate? In one-variable regression models, we addressed this question using a plot of the residuals e_i versus the original predictor x_i. This allowed us to check whether there was still a pattern in the residuals that suggested a nonlinear relationship between the predictor and response. (Recall Figure 2.6 on page 45.)

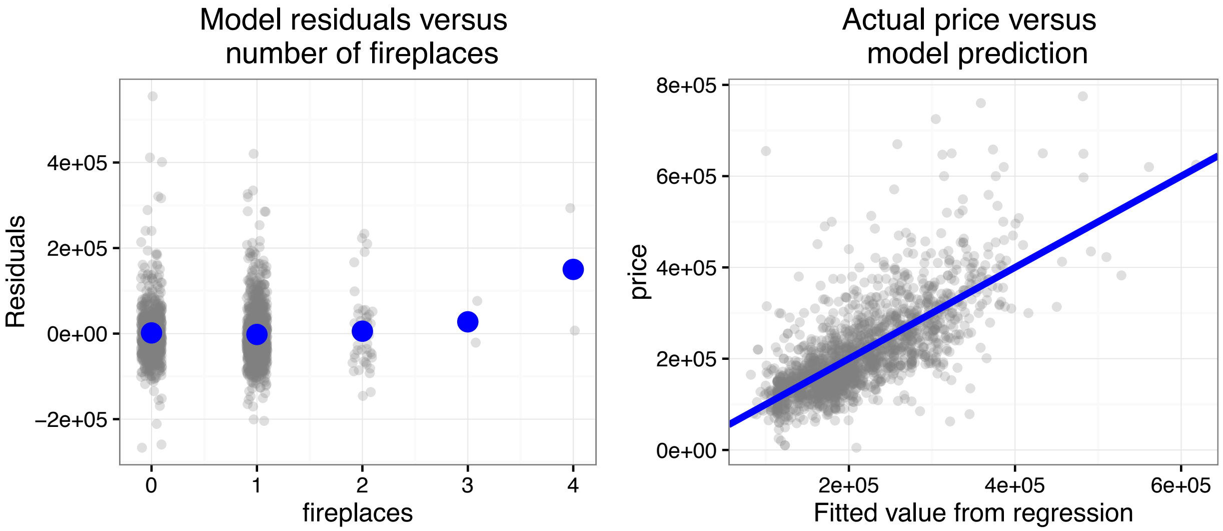


Figure 6.6: Left: model residuals versus number of fireplaces. Right: observed house prices versus fitted house prices from the multiple regression model.

There are two ways to extend the idea of a residual plot to multiple regression models:

- plotting the residuals versus each of the predictors x_{ij} individually. This allows us to check whether the response changes linearly as a function of the jth predictor.
- plotting the actual values y_i versus the fitted values $\hat{y}_i$ and looking for nonlinearities. This allows us to check whether the responses depart in a systematically nonlinear way from the model predictions.

Figure 6.6 shows an example of each plot. The left panel shows each the residual for each house versus the number of fireplaces it contains. Overall, this plot looks healthy.[3] The one caveat is that the predictions for houses with four fireplaces may be too low, which we know because the mean residual for four-fireplace houses is positive. Then again, there are only two such houses, making it difficult to draw a firm conclusion here. We probably shouldn't change our model just to chase a better fit for two (very unusual) houses out of 1,726. But we should also recognize that our model might not be great at predicting the price for a house with four fireplaces, simply because we don't have a lot of data that would allow us to do so.

The right panel of Figure 6.6 shows a plot of y_i versus $\hat{y}_i$. This also looks like a nice linear relationship, giving us further confidence that our model isn't severely distorting the true relationship

[3] What would an unhealthy residual plot look like? To give a hypothetical example, suppose we saw that the residuals for houses with no fireplace were systematically above zero, while the residuals for houses with one fireplace were systematically below zero. This would suggest a nonlinear effect that our model hasn't captured.

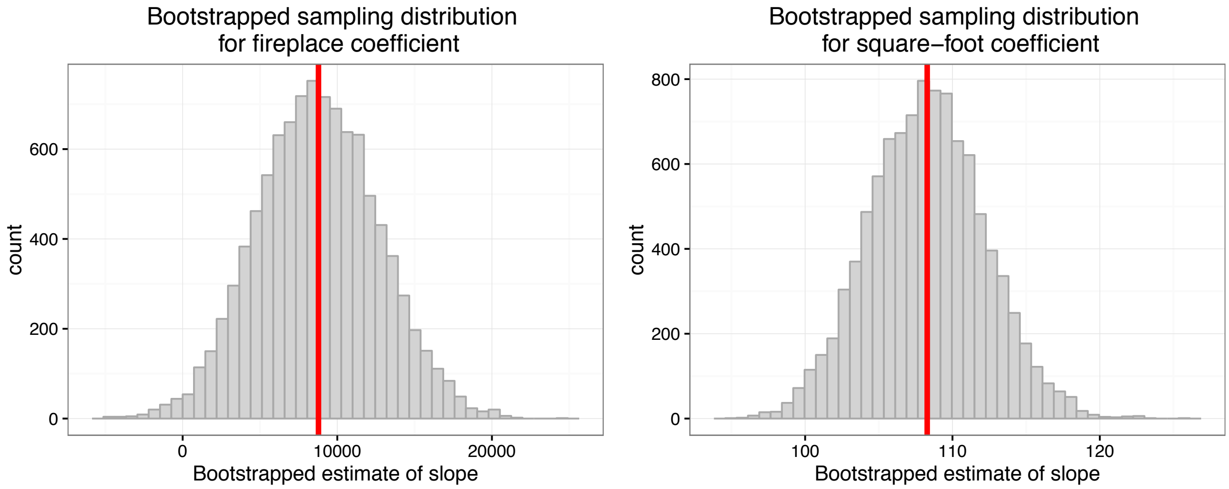


Figure 6.7: Bootstrapped estimates for the sampling distributions of the partial slopes for number of fireplaces (left) and square footage (right) from the model in Equation 6.1 on page 146. The least-squares estimates are shown as vertical red lines.

between predictors and response. In a large multiple regression model with many predictors, it may be tedious to look at e_i versus each of those predictors individually. In such cases, a plot of y_i versus $\hat{y}_i$ should be the first thing you examine to check for nonlinearities in the overall fit.

Quantifying uncertainty. We can get confidence intervals for partial relationships in a multiple regression model via bootstrapping, just as we do in a one-variable regression model.

The left panel of Figure 6.7 shows the bootstrapped estimate of the sampling distribution for the fireplace coefficient in our multiple regression model. The 95% confidence interval here is $(1095, 16380)$. Thus while we do have some uncertainty we have about the value of a fireplace, we can definitively rule out the number estimated using the overall relationship from Figure 6.4. If the county tax assessor wanted to value your new fireplace at $66,700 for property-tax purposes, Figure 6.7 would make a good argument in your appeal.[4]

The right-hand side of Figure 6.7 shows the bootstrapped sampling distribution for the square-foot coefficient. While this wasn't the focus of our analysis here, it's interesting to know that an additional square foot improves the value of a property by about $108, plus or minus about $8.

[4] At a 2% property tax rate, this might save you over $1000 a year in taxes.

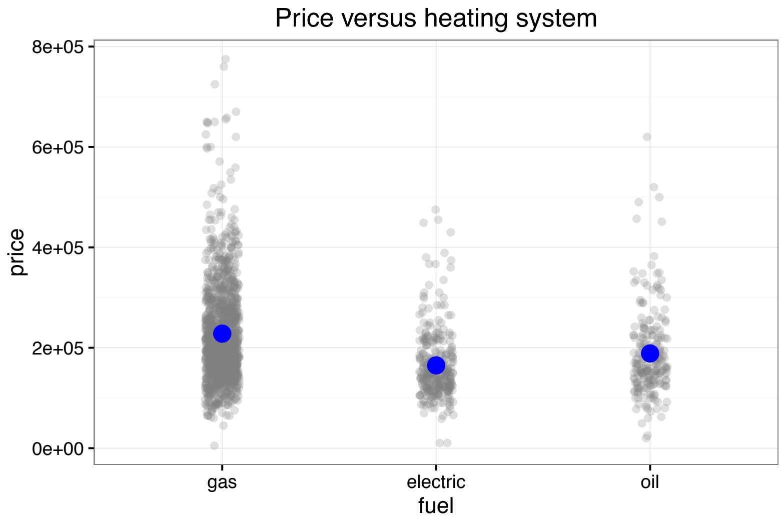


Figure 6.8: Prices of houses with gas, electric and fuel-oil heating systems.

How much is gas heating worth?

Saratoga, NY is cold in the winter: the average January day has a low of 13° F and a high of 31° F. As you might imagine, residents spend a fair amount of money heating their homes, and are sensitive to the cost differences between gas, electric, and fuel-oil heaters. Figure 6.8 suggests that the Saratoga real-estate market puts a big premium for houses with gas heaters (mean price of $228,000) versus those with electric or fuel-oil heaters (mean prices of $165,000 and $189,000, respectively).

But this is an overall relationship. Do these differences persist when we adjust for the effect of living area, lot size, and the number of fireplaces? There could be a confounding effect here. For example, maybe the bigger houses tend to have gas heaters more frequently than the small houses. Moreover, accounting for this effect might change our assessment of the value of a fireplace, since fireplaces might be used more frequently in homes with expensive-to-use heating systems.

We can investigate these issues by adding dummy variables for heating-system type to the regression equation we fit previously (on page 146). Fitting this model by least squares yields the following equation:

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

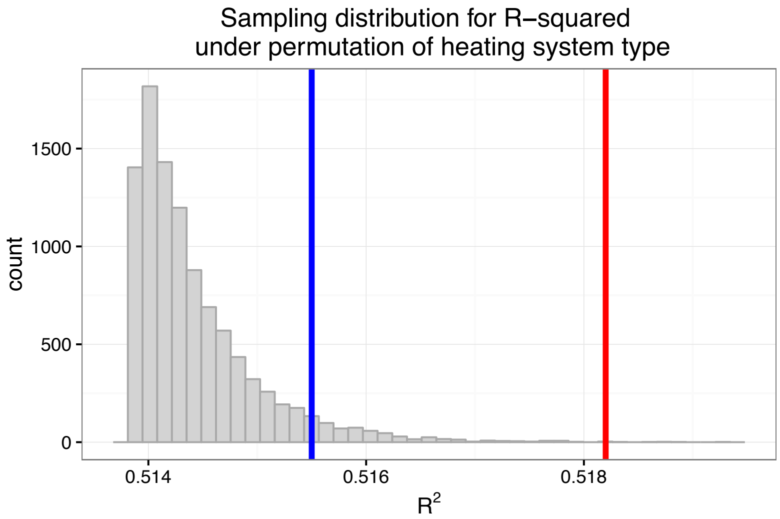


Figure 6.9: Sampling distribution of R^2 under the null hypothesis that there is no partial relationship between heating system and price after adjusting for effects due to square footage, lot size, and number of fireplaces. The blue vertical line marks the critical value for the rejection region at the $\alpha = 0.05$ 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. The red line falls in the rejection region, implying that there is statistically significant evidence for an effect on price due to heating system at the $\alpha = 0.05$ level.

The baseline case here is gas heating, since it has no dummy variable. Notice how the coefficients on the dummy variables for the other two types of heating systems shift the entire regression equation up or down. This model estimates the premium associated with gas heating to be about $14,000 over electric heating, and about $16,000 over fuel-oil heating.

Assessing statistical significance. Are these differences statistically significant, or could they be explained due to chance? To assess this, we'll use a permutation test to 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.

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. A natural way to assess the evidence against the null hypothesis is to use improvement in R^2 as a test statistic. (This is the same quantity we look at when assessing the importance of a variable in an ANOVA table.) If we see a big jump in R^2 when moving from the reduced to the full model, it stands to reason that the variable we added (here, heating system) was important, and that the null hypothesis is wrong.

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. parameters) 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 Neyman–Pearson 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 has the effect of breaking any partial relationship between heating system and price that might be present in our original data. It tells us how big an improvement in R^2 we'd expect to see when fitting the full model, even the null hypothesis were true.

This sampling distribution is shown in Figure 6.9, 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 critical value for the rejection region at the $\alpha = 0.05$ 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 in the 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 at the $\alpha = 0.05$ level.

In general, we can compare any two nested models using a permutation test based on R^2. To do so, we repeatedly shuffle the extra variables in the full model, without shuffling either the response or the variables 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.