14
Multiple regression: the basics

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 14.1 on page 144, 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. (Note: in general it's hard to interpret 3D graphics and they're not a good idea. We've been very careful here to try to show a 3D graphic just this once, specifically so you can get some intuition for the idea of a regression plane; in general, this is not how we recommend visualizing data.)

Fitting a linear equation for y versus x_1 and x_2 results in a re-

[1] These are the same SUVs shown in the second-from-right panel in FIgure 13.4, when we discussed ANOVA for models involving correlated predictors.

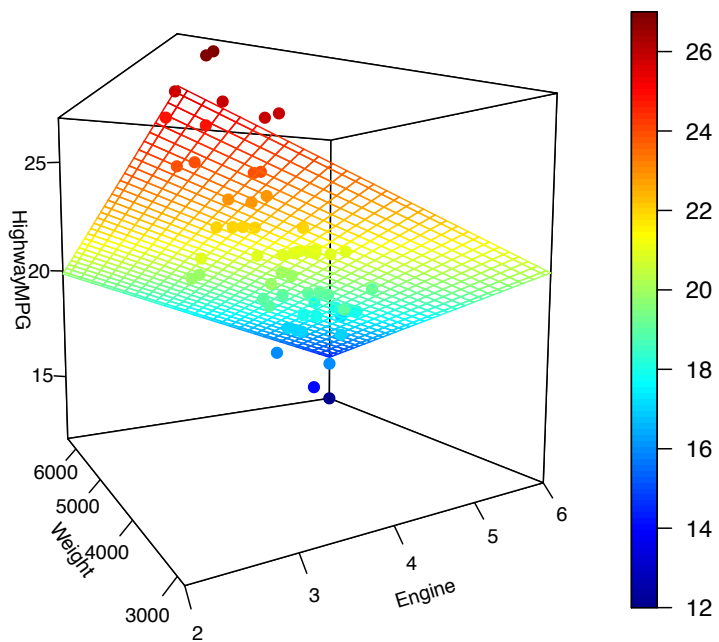**Mileage versus weight and engine power**
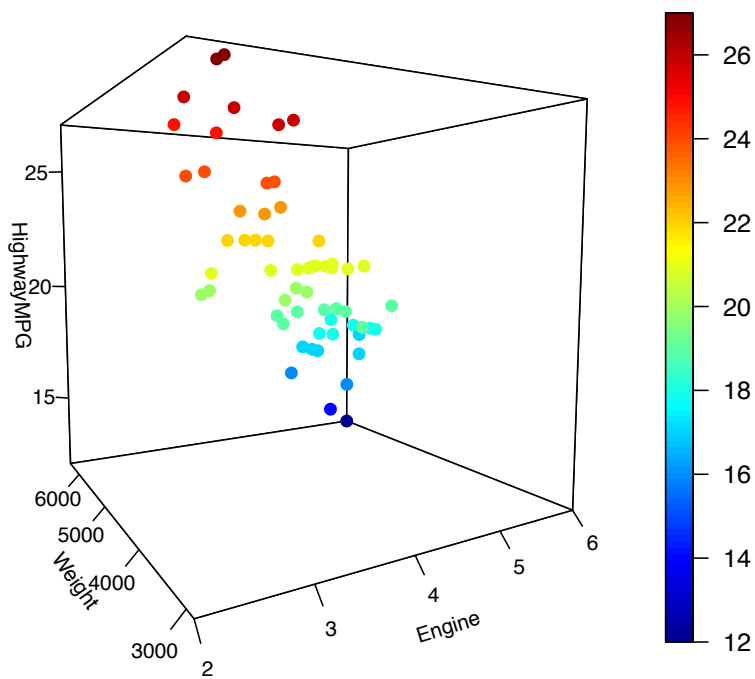
With fitted plane

Figure 14.1: Highway gas mileage versus weight and engine displacement for 59 SUVs, with the least-squares fit shown in the bottom panel.

gression 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 14.1, this fitted equation is

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

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 14.1. 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).

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 $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,p})$:

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

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.

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

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 $(x_{i,1}, x_{i,2}, \ldots, x_{i,p})$ every time. When writing things out by hand, a little arrow can be used instead, since you obviously can't write things in bold: $\vec{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,p})$. By the same logic, we also write $\vec{\beta}$ for the vector $(\beta_0, \beta_1, \ldots, \beta_p)$.

residual sum of squares in the same way as before, as the sum of squared differences between fitted and observed values:

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

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:

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

The only difference is that $\hat{y}_i$ is now a function of more than just an intercept and a single slope.

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 $\hat{\beta}_k$ 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 14.1:

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

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 14.2. 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.

Focus on the left panel of Figure 14.2 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.

Overall relationship of mileage with engine displacement

$$\hat{y}_i = 30.3 - 2.5 \cdot x_{i1}$$

Overall relationship of mileage with vehicle weight

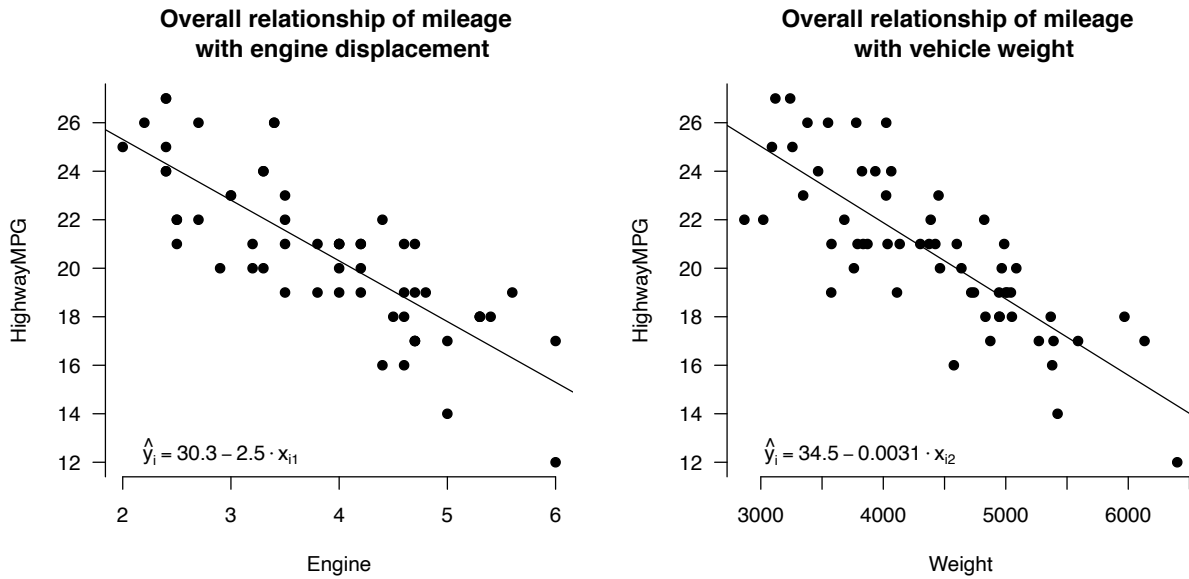$$\hat{y}_i = 34.5 - 0.0031 \cdot x_{i2}$$

Figure 14.2: Overall relationships for highway gas mileage versus weight and engine displacement individually.
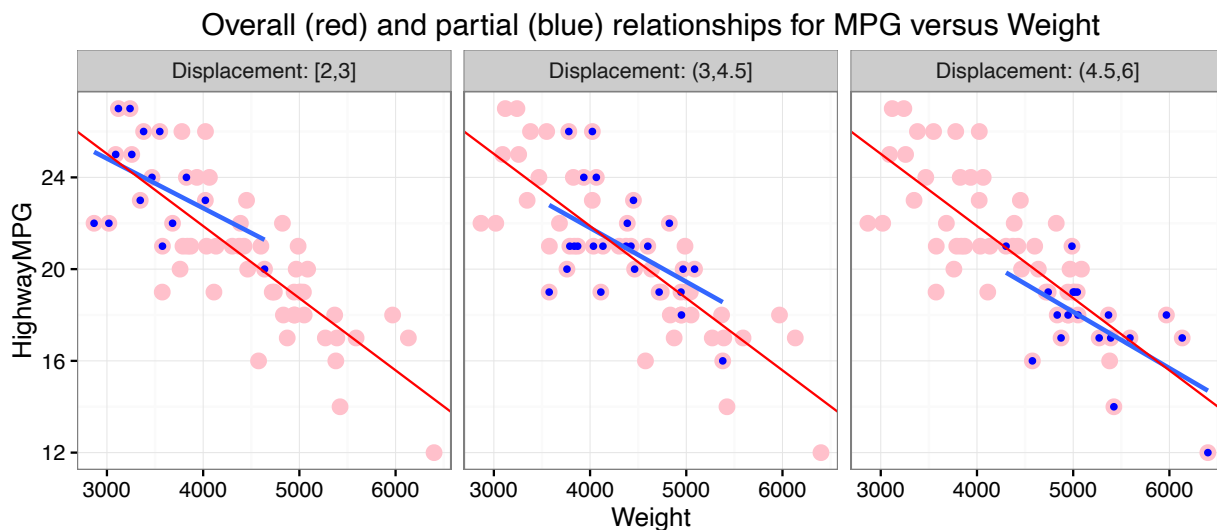
That's because this model doesn't attempt to adjust for the effects of vehicle weight. Because weight is correlated with engine displacement, we get a steeper estimate for the overall relationship than for the partial relationship: for cars where engine displacement is larger, weight also tends to be 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 14.3 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

Overall (red) and partial (blue) relationships for MPG versus Weight

Figure 14.3: 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 14.2. 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.

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 14.2.[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 14.3, 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.

[2] This is a very general property of regression: if x_1 and x_2 are two correlated (collinear) predictors, then adding x_2 to the model will change the coefficient on x_1, compared to a model with x_1 alone.

## Using multiple regression to address real-world questions

While there are many possible uses of multiple regression, most applications will fall into one of two categories:

(1) Isolating a partial relationship between the response and a predictor of interest, adjusting for possible confounders.

(2) Building a predictive model for forecasting the response, using all available sources of information.

In the rest of this chapter, we'l focus on the first use of regression (isolating a partial relationship), deferring the second (building a predictive model) for later. As a case study, we'll use a running example on house prices from Saratoga County, New York, distributed as part of the mosaic R package. We'll show how, together with multiple regression, this data set can be used 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.

How much is a fireplace worth?

Our first question is: how much does a fireplace improve the value of a house for sale? Figure 14.4 would seem to say: by about $66,700 per fireplace. This dot plot shows the sale price of houses
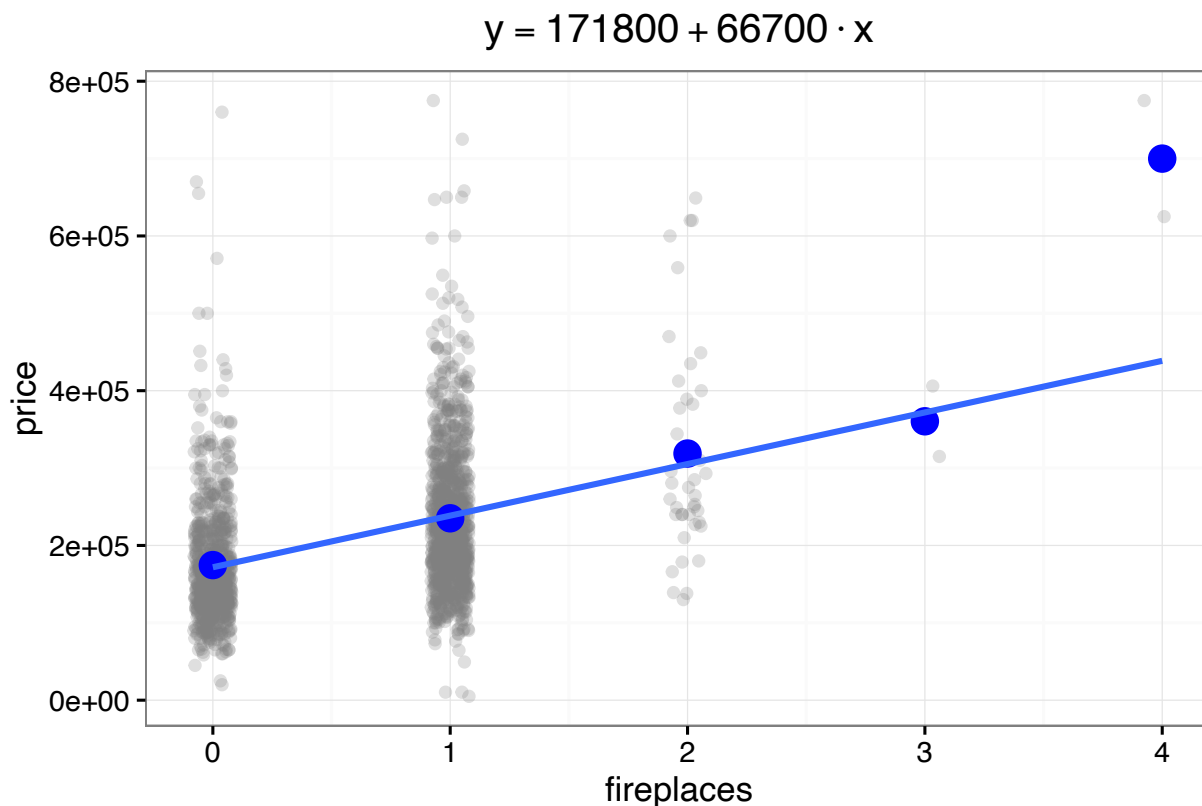
$$y = 171800 + 66700 \cdot x$$

Figure 14.4: The relationship between the price of a house and the number of fireplaces it has.

in Saratoga County, NY that were on the market in 2006.[3] We also see a linear regression model 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 14.4. 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 bricklayer so that you might cash in on your new fireplace, consult Figure 14.5 on page 153. 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
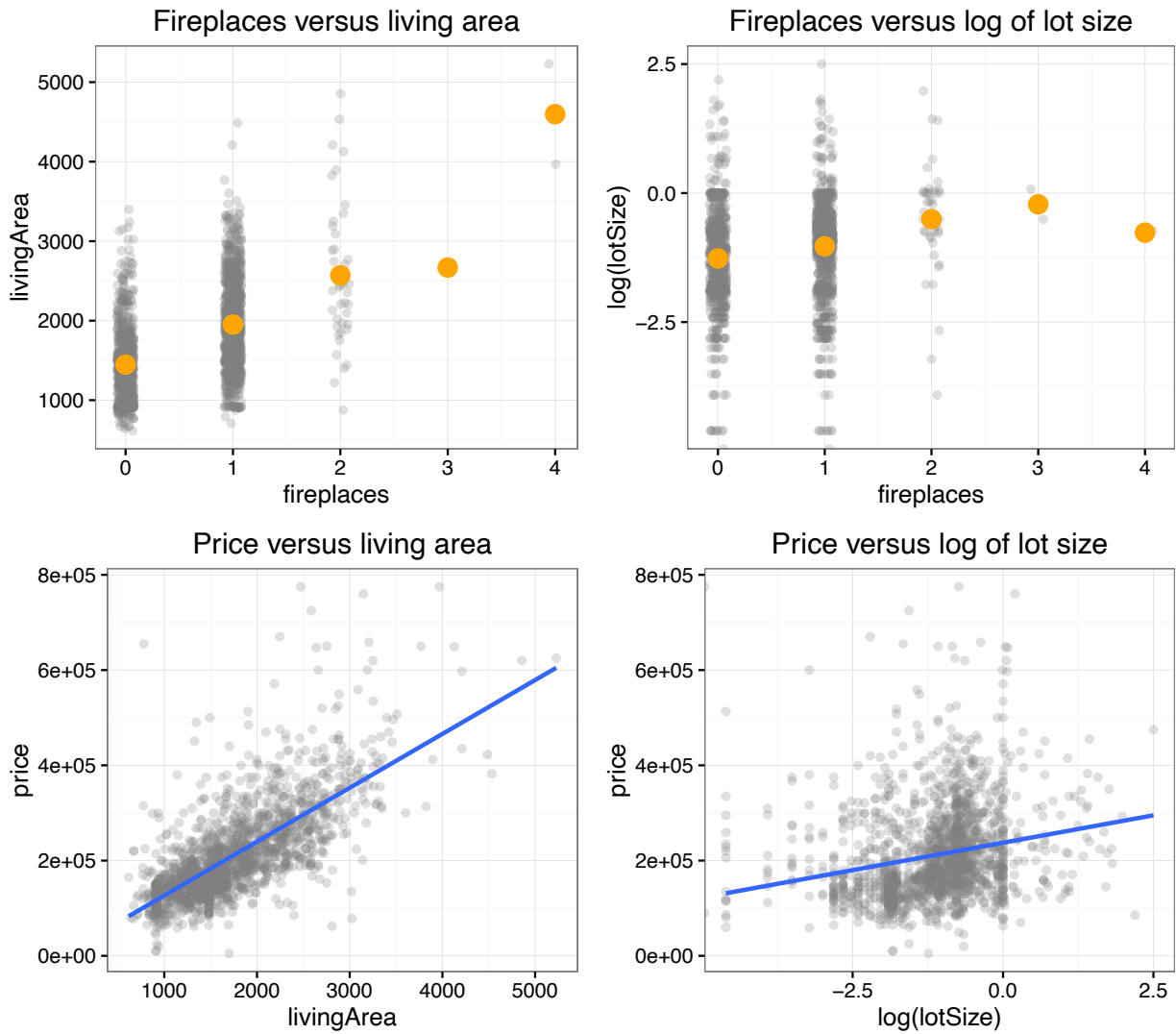
Figure 14.5: The relationship of house price with living area (bottom left) and with the logarithm of lot size in acres (bottom right). Both of these variables are potential confounders for the relationship between fireplaces and price, because they are also correlated with the number of fireplaces (top row).

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.

Thus we have two possible explanations for the relationship we see in Figure 14.4. This correlation may happen because fireplaces are so valuable. On the other hand, it may instead (or also) happen because fireplaces happen to occur more frequently in houses that are desireable for other reasons (i.e. they are bigger). This is confounding again: when some third variable is correlated with both the response and the predictor of interest.

Disentangling these two possibilities requires estimating the partial relationship between fireplaces and prices, rather than the overall relationship shown in Figure 14.4. After all, when someone like a realtor or the county tax assessor asks how much a fireplace is worth, what they really 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:

$$\text{Price} = \$17787 + 108.3 \cdot \text{SqFt} + 1257 \cdot \log(\text{Acres}) + 8783 \cdot \text{Fireplaces} + \text{Residual} \, .$$
$$(14.1)$$

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 14.4.

The example emphasizes the use of multiple regression to adjust statistically for the effect of confounders, by estimating a partial relationship between the response and the predictor of interest. This is one of the most useful real-world applications of regression modeling, and we'll see many similar examples. In general, the advice is: if you want to estimate a partial relationship, make sure you include the potential confounders in the model.

Bootstrapped sampling distribution
for fireplace coefficient

Bootstrapped sampling distribution
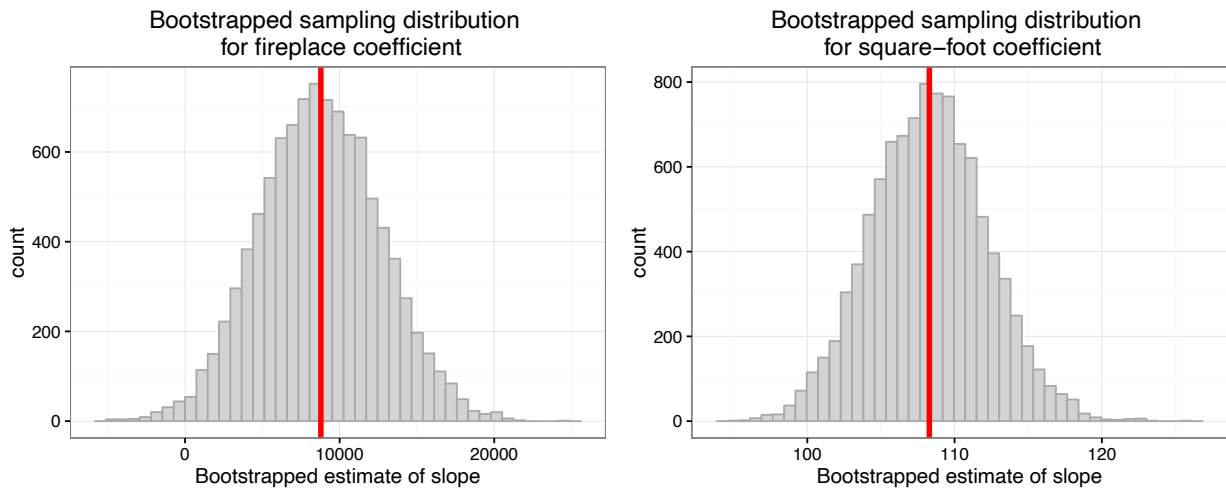for square−foot coefficient

Figure 14.6: Bootstrapped estimates
for the sampling distributions of the
partial slopes for number of fireplaces
(left) and square footage (right) from
the model in Equation 14.1 on page 154.
The least-squares estimates are shown
as vertical red lines.

Uncertainty quantification

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

The left panel of Figure 14.6 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 14.4.
If the county tax assessor wanted to value your new fireplace at
\$66,700 for property-tax purposes, Figure 14.6 would make a good
argument in your appeal.[4]

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

The right-hand side of Figure 14.6 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.

Model checking

However, before we put too much faith in the conclusions of your
fitted model, it's important to check whether the assumption of a
linear regression model is appropriate in the first place. We call
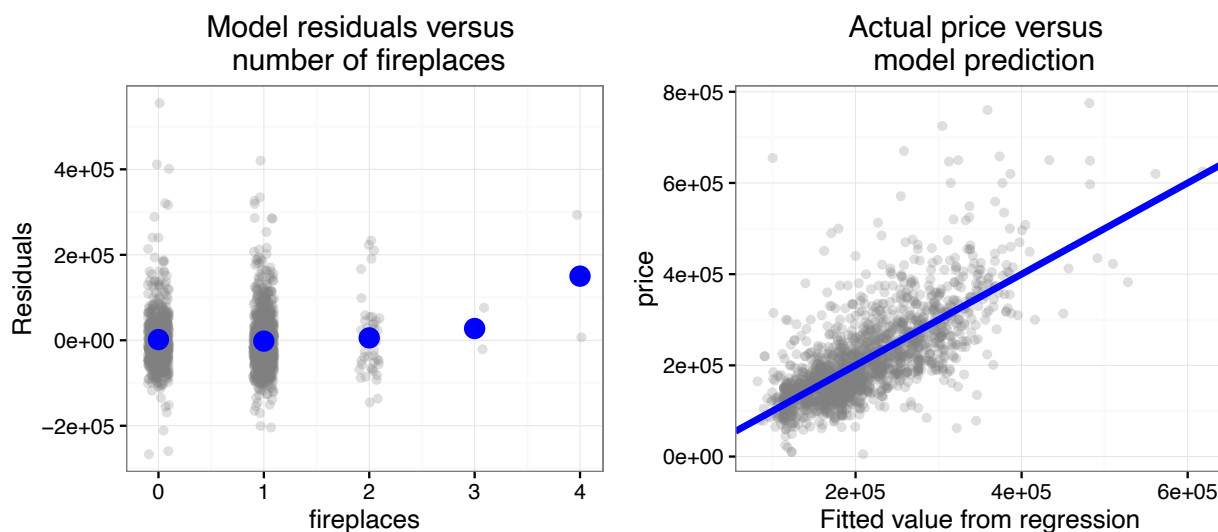this step *model checking*. We'll learn a lot more about model check-

Figure 14.7: Left: model residuals versus number of fireplaces. In a model where the assumption of linearity is met, these residuals should look randomly distributed around zero at every value of the predictor (here, fireplaces). Right: observed house prices versus fitted house prices from the multiple regression model. In a model where the assumption of linearity is reasonable, these points should fall randomly around the straight "y=x" line, with slope 1 and intercept 0.

ing later, but for now we'll cover the most basic step: validating that the response varies linearly with the predictors.

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. 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 14.7 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: there are no obvious departures from linearity. The one caveat is that the predictions for houses with four fireplaces may be too low, which we can see from the fact that 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
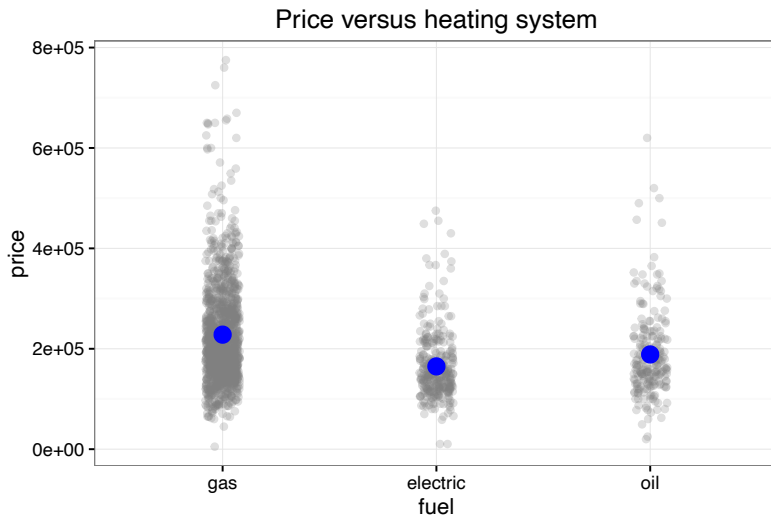
Figure 14.8: Prices of houses with gas, electric and fuel-oil heating systems.

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 it would involve extrapolation: we don't have a lot of data that can inform us about these houses.

The right panel of Figure 14.7 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 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.

How much is gas heating worth? Grouping variables in multiple regression

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 14.8 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). One possible reason is that gas heaters are cheaper to run and maintain.

Variable	Estimate	Std. Error	2.5%	97.5%
Intercept	29868	6743	16644	43093
livingArea	105	3	99	112
log(lotSize)	2705	1913	-1047	6457
fireplaces	7547	3348	980	14113
fuel=electric	-14010	4471	-22778	-5242
fuel=oil	-15879	5295	-26265	-5494

Table 14.1: Coefficients, standard errors, and 95% confidence intervals for the multiple regression model for house price (y) versus living area, log of lot size, number of fireplaces, and heating system type.

But this figure shows an overall relationship. What does the story look like 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, or maybe fireplaces are used more in homes with expensive-to-use heating systems.

Remember: if you want to isolate a partial relationship, include potential confounders in the model. We'll do this here by including two sets of terms: (1) dummy variables for heating-system type, to model the partial relationship of interest; and (2) all the possible confounding variables that we had in our previous regression equation (on page 154), which includes living area, lot size, and number of fireplaces. Fitting this model by least squares yields the following equation:

$$\text{Price} = \$29868 + 105.3 \cdot \text{SqFt} + 2705 \cdot \log(\text{lotSize}) + 7546 \cdot \text{Fireplaces}$$
$$- 14010 \cdot \mathbf{1}_{\{\text{fuel = electric}\}} - 15879 \cdot \mathbf{1}_{\{\text{fuel = oil}\}} + \text{Residual}.$$

The full table of coefficients, standard errors, and 95% confidence intervals is in Table 14.1. The baseline 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 ± 4500 over electric heating (estimate, plus-or-minus one standard error), and about $16,000 ± 5300 over fuel-oil heating. Because these are terms in a multiple regression model, these numbers represent partial relationships, adjusting for size, lot acreage, and number of fireplaces.

Assessing statistical significance

A question that often comes up in multiple regression is whether a particular term in the model is "statistically significant" at some specified level (e.g. 5%). All this means is whether zero is a plausible value for that partial slope in the model. Remember, a coefficient of zero means that there is no partial relationship between the response and the corresponding predictor, adjusting for the other terms in the model. So when we say that a predictor is statistically significant, all we mean is that it we think it has a nonzero (partial) relationship with the response.

Here are a few quick observations and guidelines about the topic of "statistical significance."

First, by convention, people express the statistical significance level as the opposite of the confidence level. So a confidence level of 95% means a significance level of 5%; a confidence level of 99% means a significance level of 1%; and so forth. This is confusing at first, but you'll get used to it. Just remember: the *lower* the significance level, the stronger the evidence that some variable has a nonzero relationship with the response.

Second, in regression models we can often assess statistical significance just by looking at whether zero is included in the confidence interval. That's because "statistically significant" just means "zero is not a plausible value," and a confidence interval gives us a range of plausible values. For example, let's take the 95% confidence intervals for two terms in Table 14.1:

- The 95% confidence interval for the partial slope on fireplaces is $(980, 14113)$. We can rule out zero as a plausible value at a 95% confidence level, and so we can say that the lot size variable is statistically significant at the 5% level.

- The 95% confidence interval for the partial slope on lot size is $(-1047, 6457)$. We cannot rule out zero as a plausible value with 95% confidence, and so the lot size variable is not statistically significant at the 5% level.

Third, the fact that some variable is "statistically significant" does not mean that this variable is important in practical terms. A "significant" variable does not necessarily have a large effect on the response, nor is it automatically important for generating good predictions. Statistical significance means that we think the corresponding coefficient isn't zero. But it could still be very small.

This is why, in most cases, it is better to focus on a variable's confidence interval, rather than on whether a variable is significant. The confidence interval carries a lot more information than a simplistic distinction between "significant" and "insignificant," because it gives you a range of plausible values for the coefficient.

Finally, the fact that some variable is *not* statistically significant does not imply that this variable has no relationship with the response, or that it should automatically be dropped from the model. A lack of statistical significance could just mean a lot of uncertainty—in other words, that the data aren't capable of precisely estimating the numerical magnitude of some variable's partial relationship with the response. There's an important but subtle distinction here: an insignificant coefficient means that we have an *absence of compelling evidence* for a nonzero effect. It does not mean that we have found *compelling evidence that the effect is absent.*

For example, the confidence interval for the log(acres) term in Table 14.1 is $(-1047, 6457)$. We therefore cannot rule out zero as a plausible value. But there are lot of large values, like 5000 or 6000, that we cannot rule out, either! There's a lot of uncertainty here. One symptom of this is a wide confidence interval; another symptom is a lack of statistical significance at the 5% level. But it does not follow that lot size is irrelevant for predicting house price.[5]

[5] In this the large standard error is almost surely due to collinearity between lot size and other predictors.