16:42 Saturday 16th January, 2016

Chapter 2

The Truth about Linear Regression

We need to say some more about how linear regression, and especially about how it really works and how it can fail. Linear regression is important because

1. it's a fairly straightforward technique which sometimes works tolerably for prediction;
2. it's a simple foundation for some more sophisticated techniques;
3. it's a standard method so people use it to communicate; and
4. it's a standard method so people have come to confuse it with prediction and even with causal inference as such.

We need to go over (1)–(3), and provide prophylaxis against (4).

2.1 Optimal Linear Prediction: Multiple Variables

We have a response variable Y and a p-dimensional vector of predictor variables or features $\vec{X}$. We would like to predict Y using $\vec{X}$. We saw last time that the best predictor we could use, at least in a mean-squared sense, is the conditional expectation,

$$\mu(\vec{x}) = \mathbb{E}\left[Y|\vec{X} = \vec{x}\right] \tag{2.1}$$

Instead of using the optimal predictor $\mu(\vec{x})$, let's try to predict as well as possible while using only a linear function of $\vec{x}$, say $\beta_0 + \beta \cdot \vec{x}$. This is not an assumption about the world, but rather a decision on our part; a choice, not a hypothesis. This decision can be good — $\beta_0 + \vec{x} \cdot \beta$ could be a tolerable approximation to $\mu(\vec{x})$ — even if the linear hypothesis is strictly wrong. It might also be that no linear approximation to μ is much good mathematically, but we might be forced to make it for practical reasons, e.g., speed of computation.

Perhaps the best reason to hope the choice to use a linear model isn't crazy is that we may hope μ is a smooth function. If it is, then we can Taylor expand it about our favorite point, say $\vec{u}$:

$$\mu(\vec{x}) = \mu(\vec{u}) + \sum_{i=1}^{p} \left(\left. \frac{\partial r}{\partial x_i} \right|_{\vec{u}} \right) (x_i - u_i) + O(||\vec{x} - \vec{u}||^2) \tag{2.2}$$

or, in the more compact vector-calculus notation,

$$\mu(\vec{x}) = \mu(\vec{u}) + (\vec{x} - \vec{u}) \cdot \nabla \mu(\vec{u}) + O(||\vec{x} - \vec{u}||^2) \tag{2.3}$$

If we only look at points $\vec{x}$ which are close to $\vec{u}$, then the remainder terms $O(||\vec{x} - \vec{u}||^2)$ are small, and a linear approximation is a good one[1]. Here, "close to $\vec{u}$" really means "so close that all the non-linear terms in the Taylor series are comparatively negligible".

Of course there are lots of linear functions so we need to pick one, and we may as well do that by minimizing mean-squared error again:

$$MSE(\beta) = \mathbb{E}\left[\left(Y - \beta_0 - \vec{X} \cdot \beta \right)^2 \right] \tag{2.4}$$

Going through the optimization is parallel to the one-dimensional case (see last chapter), with the conclusion that the optimal β is

$$\beta = \mathbf{v}^{-1} \text{Cov}\left[\vec{X}, Y\right] \tag{2.5}$$

where $\mathbf{v}$ is the covariance matrix of $\vec{X}$, i.e., $v_{ij} = \text{Cov}\left[X_i, X_j\right]$, and $\text{Cov}\left[\vec{X}, Y\right]$ is the vector of covariances between the predictor variables and Y, i.e. $\text{Cov}\left[\vec{X}, Y\right]_i = \text{Cov}\left[X_i, Y\right]$. We also get

$$\beta_0 = \mathbb{E}[Y] - \beta \cdot \mathbb{E}\left[\vec{X}\right] \tag{2.6}$$

just as in the one-dimensional case (Exercise 1).

Multiple regression would be a lot simpler if we could just do a simple regression for each predictor variable, and add them up; but really, this is what multiple regression *does*, just in a disguised form. If the input variables are uncorrelated, $\mathbf{v}$ is diagonal ($v_{ij} = 0$ unless $i = j$), and so is $\mathbf{v}^{-1}$. Then doing multiple regression breaks up into a sum of separate simple regressions across each input variable. When the input variables are correlated and $\mathbf{v}$ is not diagonal, we can think of the multiplication by $\mathbf{v}^{-1}$ as **de-correlating** $\vec{X}$ — applying a linear transformation to come up with a new set of inputs which are uncorrelated with each other.[2]

[1] If you are not familiar with the big-O notation like $O(||\vec{x} - \vec{u}||^2)$, now would be a good time to read Appendix C.

[2] If $\vec{Z}$ is a random vector with covariance matrix I, then $\mathbf{w}\vec{Z}$ is a random vector with covariance matrix $\mathbf{w}^T\mathbf{w}$. Conversely, if we start with a random vector $\vec{X}$ with covariance matrix $\mathbf{v}$, the latter has a "square root" $\mathbf{v}^{1/2}$ (i.e., $\mathbf{v}^{1/2}\mathbf{v}^{1/2} = \mathbf{v}$), and $\mathbf{v}^{-1/2}\vec{X}$ will be a random vector with covariance matrix $\mathbf{I}$. When we write our predictions as $\vec{X}\mathbf{v}^{-1}\text{Cov}\left[\vec{X}, Y\right]$, we should think of this as $\left(\vec{X}\mathbf{v}^{-1/2}\right)\left(\mathbf{v}^{-1/2}\text{Cov}\left[\vec{X}, Y\right]\right)$. We use one power of $\mathbf{v}^{-1/2}$ to transform the input features into uncorrelated variables before taking their correlations with the response, and the other power to decorrelate $\vec{X}$.

Notice: β depends on the marginal distribution of $\vec{X}$ (through the covariance matrix $\mathbf{v}$). If that shifts, the optimal coefficients β will shift, *unless* the real regression function is linear.

2.1.1 Collinearity

The formula $\beta = \mathbf{v}^{-1}\text{Cov}\left[\vec{X}, Y\right]$ makes no sense if $\mathbf{v}$ has no inverse. This will happen if, and only if, the predictor variables are linearly dependent on each other — if one of the predictors is really a linear combination of the others. Then (as we learned in linear algebra) the covariance matrix is of less than "full rank" (i.e., "rank deficient") and it doesn't have an inverse. Equivalently, $\mathbf{v}$ has at least one eigenvalue which is exactly zero.

So much for the algebra; what does that mean statistically? Let's take an easy case where one of the predictors is just a multiple of the others — say you've included people's weight in pounds (X_1) and mass in kilograms (X_2), so $X_1 = 2.2X_2$. Then if we try to predict Y, we'd have

$$\begin{aligned}
\widehat{\mu}(\vec{X}) &= \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p && (2.7)\\
&= 0X_1 + (2.2\beta_1 + \beta_2)X_2 + \sum_{i=3}^{p} \beta_i X_i && (2.8)\\
&= (\beta_1 + \beta_2/2.2)X_1 + 0X_2 + \sum_{i=3}^{p} \beta_i X_i && (2.9)\\
&= -2200X_1 + (1000 + \beta_1 + \beta_2)X_2 + \sum_{i=3}^{p} \beta_i X_i && (2.10)
\end{aligned}$$

In other words, because there's a linear relationship between X_1 and X_2, we make the coefficient for X_1 whatever we like, provided we make a corresponding adjustment to the coefficient for X_2, and it has no effect at all on our prediction. So rather than having one optimal linear predictor, we have infinitely many of them.[3]

There are three ways of dealing with collinearity. One is to get a different data set where the predictor variables are no longer collinear. A second is to identify one of the collinear variables (it usually doesn't matter which) and drop it from the data set. This can get complicated; principal components analysis (Chapter 16) can help here. Thirdly, since the issue is that there are infinitely many different coefficient vectors which all minimize the MSE, we could appeal to some extra principle, beyond prediction accuracy, to select just one of them. We might, for instance, prefer smaller coefficient vectors (all else being equal), or ones where more of the coefficients were exactly zero. Using some quality other than the squared error to pick out a unique solution is called "regularizing" the optimization problem, and a lot of attention has been given to regularized regression, especially in the "high dimensional" setting where the number of coefficients is comparable to, or even greater than, the number of data points. See Appendix H.5.5, and exercise 2 in Chapter 8.

[[Add chapter on ridge and lasso?]]

[3] Algebraically, there is a linear combination of two (or more) of the predictor variables which is con-

2.1.2 The Prediction and Its Error

Once we have coefficients β, we can use them to make predictions for the expected value of Y at *arbitrary* values of $\vec{X}$, whether we've an observation there before or not. How good are these?

If we have the optimal coefficients, then the prediction error will be uncorrelated with the predictor variables:

$$\begin{aligned}
\text{Cov}\left[Y - \vec{X}\cdot\beta, \vec{X}\right] &= \text{Cov}\left[Y, \vec{X}\right] - \text{Cov}\left[\vec{X}\cdot(\mathbf{v}^{-1}\text{Cov}\left[\vec{X}, Y\right]), \vec{X}\right] && (2.11)\\
&= \text{Cov}\left[Y, \vec{X}\right] - \mathbf{v}\mathbf{v}^{-1}\text{Cov}\left[Y, \vec{X}\right] && (2.12)\\
&= 0 && (2.13)
\end{aligned}$$

Moreover, the expected prediction error, averaged over all $\vec{X}$, will be zero (Exercise 2). In general, however, the conditional expectation of the error is not zero,

$$\mathbb{E}\left[Y - \vec{X}\cdot\beta \,|\, \vec{X} = \vec{x}\right] \neq 0 \tag{2.14}$$

and the conditional variance is not constant in $\vec{x}$,

$$\mathbb{V}\left[Y - \vec{X}\cdot\beta \,|\, \vec{X} = \vec{x}_1\right] \neq \mathbb{V}\left[Y - \vec{X}\cdot\beta \,|\, \vec{X} = \vec{x}_2\right] \tag{2.15}$$

The optimal linear predictor can be arbitrarily bad, and it can make arbitrarily big systematic mistakes. It is generally very biased[4].

2.1.3 Estimating the Optimal Linear Predictor

To actually estimate β from data, we need to make some probabilistic assumptions about where the data comes from. A fairly weak but often sufficient assumption is that observations $(\vec{X}_i, Y_i)$ are independent for different values of i, with unchanging covariances. Then if we look at the sample covariances, they will, by the law of large numbers, converge on the true covariances:

$$\begin{aligned}
\frac{1}{n}\mathbf{X}^T\mathbf{Y} &\to \text{Cov}\left[\vec{X}, Y\right] && (2.16)\\
\frac{1}{n}\mathbf{X}^T\mathbf{X} &\to \mathbf{v} && (2.17)
\end{aligned}$$

where as before $\mathbf{X}$ is the data-frame matrix with one row for each data point and one column for each feature, and similarly for $\mathbf{Y}$.

So, by continuity,

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \to \beta \tag{2.18}$$

and we have a consistent estimator.

stant. The coefficients of this linear combination are given by one of the zerp eigenvectors of $\mathbf{v}$.

[4] You were taught in your linear models course that linear regression makes unbiased predictions. This presumed that the linear model was true.

On the other hand, we could start with the residual sum of squares

$$\text{RSS}(\beta) \equiv \sum_{i=1}^{n} (y_i - \vec{x}_i \cdot \beta)^2 \tag{2.19}$$

and try to minimize it. The minimizer is the same $\widehat{\beta}$ we got by plugging in the sample covariances. No probabilistic assumption is needed to minimize the RSS, but it doesn't let us say anything about the *convergence* of $\widehat{\beta}$. For that, we do need some assumptions about $\vec{X}$ and Y coming from distributions with unchanging covariances.

(One can also show that the least-squares estimate is the linear predictor with the minimax prediction risk. That is, its worst-case performance, when everything goes wrong and the data are horrible, will be better than any other linear method. This is some comfort, especially if you have a gloomy and pessimistic view of data, but other methods of estimation may work better in less-than-worst-case scenarios.)

2.1.3.1 Unbiasedness and Variance of Ordinary Least Squares Estimates

The very weak assumptions we have made are still strong enough to let us say a little bit more about the properties of the ordinary least squares estimate $\widehat{\beta}$. To do so, we need to think about why $\widehat{\beta}$ fluctuates. For the moment, let's fix $\mathbf{X}$ at a particular value $\mathbf{x}$, but allow $\mathbf{Y}$ to vary randomly (what's called "fixed design" regression).

The key fact is that $\widehat{\beta}$ is linear in the observed responses $\mathbf{Y}$. We can use this by writing, as you're used to from your linear regression class,

$$Y = \vec{X} \cdot \beta + \epsilon \tag{2.20}$$

Here ϵ is the noise around the optimal linear predictor; we have to remember that while $\mathbb{E}[\epsilon] = 0$ and $\text{Cov}\left[\epsilon, \vec{X}\right] = 0$, it is not generally true that $\mathbb{E}\left[\epsilon | \vec{X} = \vec{x}\right] = 0$ or that $\mathbb{V}\left[\epsilon | \vec{X} = \vec{x}\right]$ is constant. Even with these limitations, we can still say that

$$\begin{aligned}
\widehat{\beta} &= (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y} && (2.21)\\
&= (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T(\mathbf{x}\beta + \epsilon) && (2.22)\\
&= \beta + (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\epsilon && (2.23)
\end{aligned}$$

This directly tells us that $\widehat{\beta}$ is an unbiased estimate of β:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\beta} | \mathbf{X} = \mathbf{x}\right] &= \beta + (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbb{E}[\epsilon] && (2.24)\\
&= \beta + 0 = \beta && (2.25)
\end{aligned}$$

We can also get the variance matrix of $\widehat{\beta}$:

$$\begin{aligned}
\mathbb{V}\left[\widehat{\beta} | \mathbf{X} = \mathbf{x}\right] &= \mathbb{V}\left[\beta + (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\epsilon \,|\, \mathbf{x}\right] && (2.26)\\
&= \mathbb{V}\left[(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\epsilon \,|\, \mathbf{X} = \mathbf{x}\right] && (2.27)\\
&= (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbb{V}[\epsilon \,|\, \mathbf{X} = \mathbf{x}]\,\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1} && (2.28)
\end{aligned}$$

Let's write $\mathbb{V}[\epsilon|\mathbf{X}=\mathbf{x}]$ as a single matrix $\mathbf{\Sigma}(\mathbf{x})$. If the linear-prediction errors are uncorrelated with each other, then $\mathbf{\Sigma}$ will be diagonal. If they're also of equal variance, then $\mathbf{\Sigma}=\sigma^2\mathbf{I}$, and we have

$$\mathbb{V}\left[\widehat{\beta}|\mathbf{X}=\mathbf{x}\right] = \sigma^2(\mathbf{x}^T\mathbf{x})^{-1} = \frac{\sigma^2}{n}\left(\frac{1}{n}\mathbf{x}^T\mathbf{x}\right)^{-1} \tag{2.29}$$

Said in words, this means that the variance of our estimates of the linear-regression coefficient will (i) go down with the sample size n, (ii) go up as the linear regression gets worse (σ^2 grows), and (iii) go down as the predictor variables, the components of $\vec{X}$, have more sample variance themselves, and are more nearly uncorrelated with each other.

If we allow $\mathbf{X}$ to vary, then by the law of total variance,

$$\mathbb{V}\left[\widehat{\beta}\right] = \mathbb{E}\left[\mathbb{V}\left[\widehat{\beta}|X\right]\right] + \mathbb{V}\left[\mathbb{E}\left[\widehat{\beta}|X\right]\right] = \frac{\sigma^2}{n}\mathbb{E}\left[\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)^{-1}\right] \tag{2.30}$$

As $n\to\infty$, the sample variance matrix $n^{-1}\mathbf{X}^T\mathbf{X}\to\mathbf{v}$. Since matrix inversion is continuous, $\mathbb{V}\left[\widehat{\beta}\right]\to n^{-1}\sigma^2\mathbf{v}^{-1}$, and points (i)–(iii) still hold.

2.2 Shifting Distributions, Omitted Variables, and Transformations

2.2.1 Changing Slopes

I said earlier that the best β in linear regression will depend on the distribution of the predictor variables, unless the conditional mean is exactly linear. Here is an illustration. For simplicity, let's say that $p=1$, so there's only one predictor variable. I generated data from $Y=\sqrt{X}+\epsilon$, with $\epsilon\sim\mathcal{N}(0,0.05^2)$ (i.e. the standard deviation of the noise was 0.05).

Figure 2.1 shows the regression lines inferred from samples with three different distributions of X: the black points are $X\sim\text{Unif}(0,1)$, the blue are $X\sim\mathcal{N}(0.5,0.01)$ and the red $X\sim\text{Unif}(2,3)$. The regression lines are shown as colored solid lines; those from the blue and the black data are quite similar — and similarly wrong. The dashed black line is the regression line fitted to the complete data set. Finally, the light grey curve is the true regression function, $\mu(x)=\sqrt{x}$.

2.2.1.1 R^2: Distraction or Nuisance?

This little set-up, by the way, illustrates that R^2 is not a stable property of the distribution either. For the black points, $R^2=0.92$; for the blue, $R^2=0.70$; and for the red, $R^2=0.77$; and for the complete data, 0.96. Other sets of x_i values would give other values for R^2. Note that while the global linear fit isn't even a good approximation anywhere in particular, it has the highest R^2.

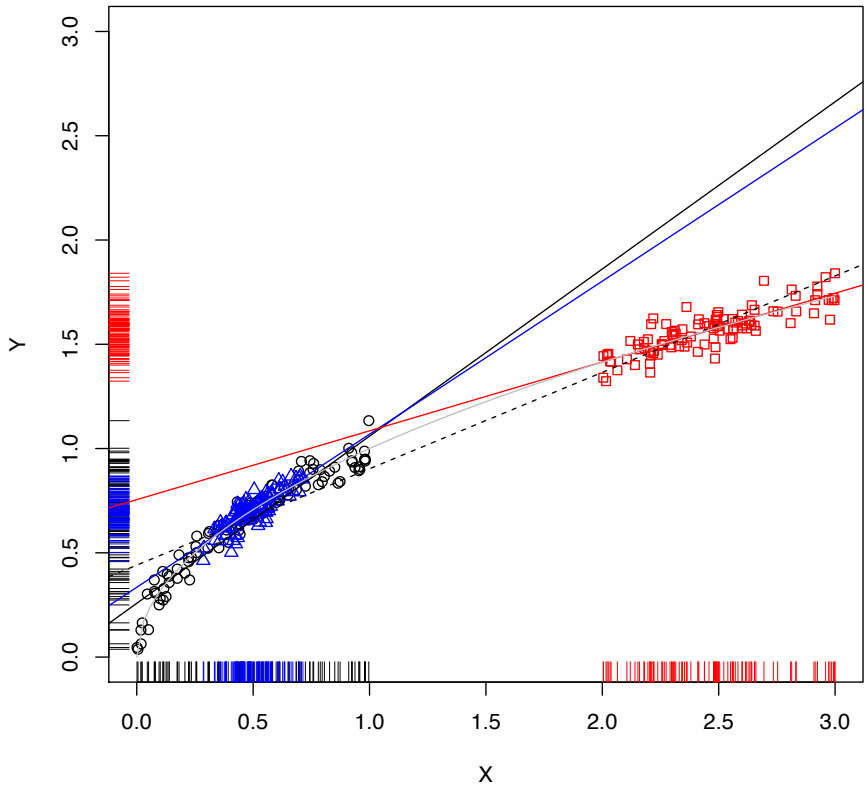


FIGURE 2.1: *Behavior of the conditional distribution* $Y|X \sim \mathcal{N}(\sqrt{X}, 0.05^2)$ *with different distributions of* X. *Black circles:* X *is uniformly distributed in the unit interval. Blue triangles: Gaussian with mean* 0.5 *and standard deviation* 0.1. *Red squares: uniform between 2 and 3. Axis tick-marks show the location of the actual sample points. Solid colored lines show the three regression lines obtained by fitting to the three different data sets; the dashed line is from fitting to all three. The grey curve is the true regression function. (See accompanying R file for commands used to make this figure.)*

This kind of perversity can happen even in a completely linear set-up. Suppose now that $Y = aX + \epsilon$, and we happen to know a exactly. The variance of Y will be $a^2\mathbb{V}[X] + \mathbb{V}[\epsilon]$. The amount of variance our regression "explains" — really, the variance of our predictions — will be $a^2\mathbb{V}[X]$. So $R^2 = \frac{a^2\mathbb{V}[X]}{a^2\mathbb{V}[X]+\mathbb{V}[\epsilon]}$. This goes to zero as $\mathbb{V}[X] \to 0$ and it goes to 1 as $\mathbb{V}[X] \to \infty$. It thus has little to do with the quality of the fit, and a lot to do with how spread out the predictor variable is.

Notice also how easy it is to get a very high R^2 even when the true model is not linear!

2.2.2 Omitted Variables and Shifting Distributions

That the optimal regression coefficients can change with the distribution of the predictor features is annoying, but one could after all *notice* that the distribution has shifted, and so be cautious about relying on the old regression. More subtle is that the regression coefficients can depend on variables which you do not measure, and those can shift without your noticing anything.

Mathematically, the issue is that

$$\mathbb{E}\left[Y|\vec{X}\right] = \mathbb{E}\left[\mathbb{E}\left[Y|Z,\vec{X}\right]|\vec{X}\right] \tag{2.31}$$

Now, if Y is independent of Z given $\vec{X}$, then the extra conditioning in the inner expectation does nothing and changing Z doesn't alter our predictions. But in general there will be plenty of variables Z which we don't measure (so they're not included in $\vec{X}$) but which have some non-redundant information about the response (so that Y depends on Z even conditional on $\vec{X}$). If the distribution of $\vec{X}$ given Z changes, then the optimal regression of Y on $\vec{X}$ should change too.

Here's an example. X and Z are both $\mathcal{N}(0,1)$, but with a positive correlation of 0.1. In reality, $Y \sim \mathcal{N}(X+Z, 0.01)$. Figure 2.2 shows a scatterplot of all three variables together ($n = 100$).

Now I change the correlation between X and Z to -0.1. This leaves both marginal distributions alone, and is barely detectable by eye (Figure 2.3).

Figure 2.4 shows just the X and Y values from the two data sets, in black for the points with a positive correlation between X and Z, and in blue when the correlation is negative. Looking by eye at the points and at the axis tick-marks, one sees that, as promised, there is very little change in the *marginal* distribution of either variable. Furthermore, the correlation between X and Y doesn't change much, going only from 0.74 to 0.63. On the other hand, the regression lines are noticeably different. When $\mathrm{Cov}[X,Z] = 0.1$, the slope of the regression line is 0.96 — high values for X tend to indicate high values for Z, which also increases Y. When $\mathrm{Cov}[X,Z] = -0.1$, the slope of the regression line is 0.84, because now extreme values of X are signs that Z is at the opposite extreme, bringing Y closer back to its mean. But, to repeat, the difference here is due to a change in the correlation between X and Z, not how those variables themselves relate to Y. If I regress Y on X and Z, I get $\widehat{\beta} = 1, 1$ in the first case and $\widehat{\beta} = 1, 1$ in the second.

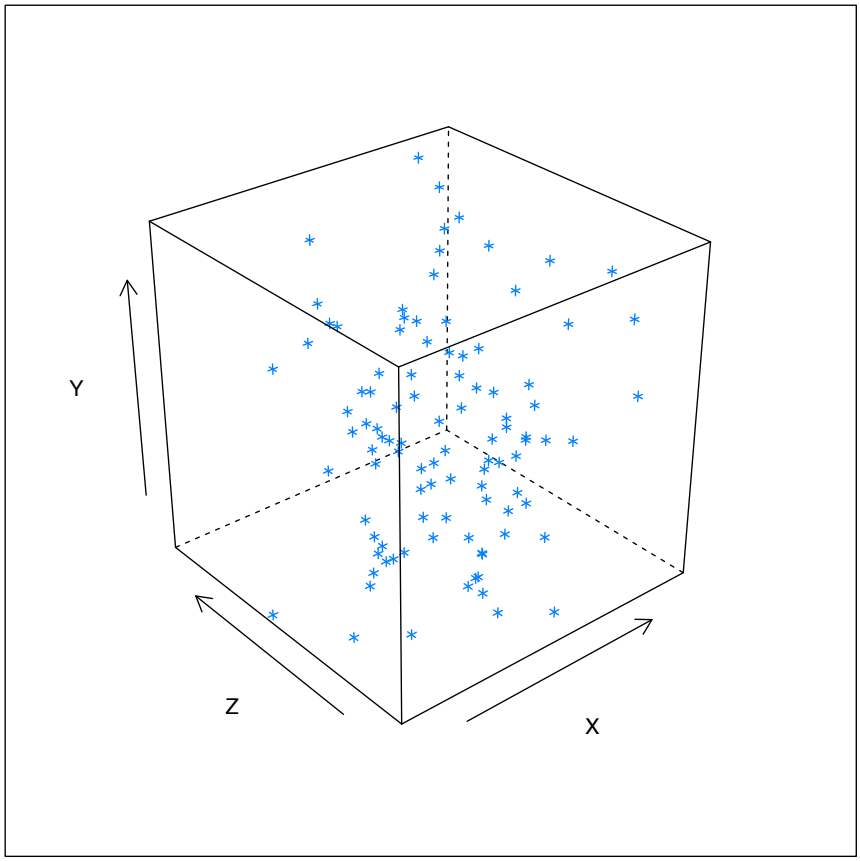


FIGURE 2.2: *Scatter-plot of response variable Y (vertical axis) and two variables which influence it (horizontal axes): X, which is included in the regression, and Z, which is omitted. X and Z have a correlation of $+0.1$. (Figure created using the* `cloud` *command in the package* `lattice`*; see accompanying R file.)*

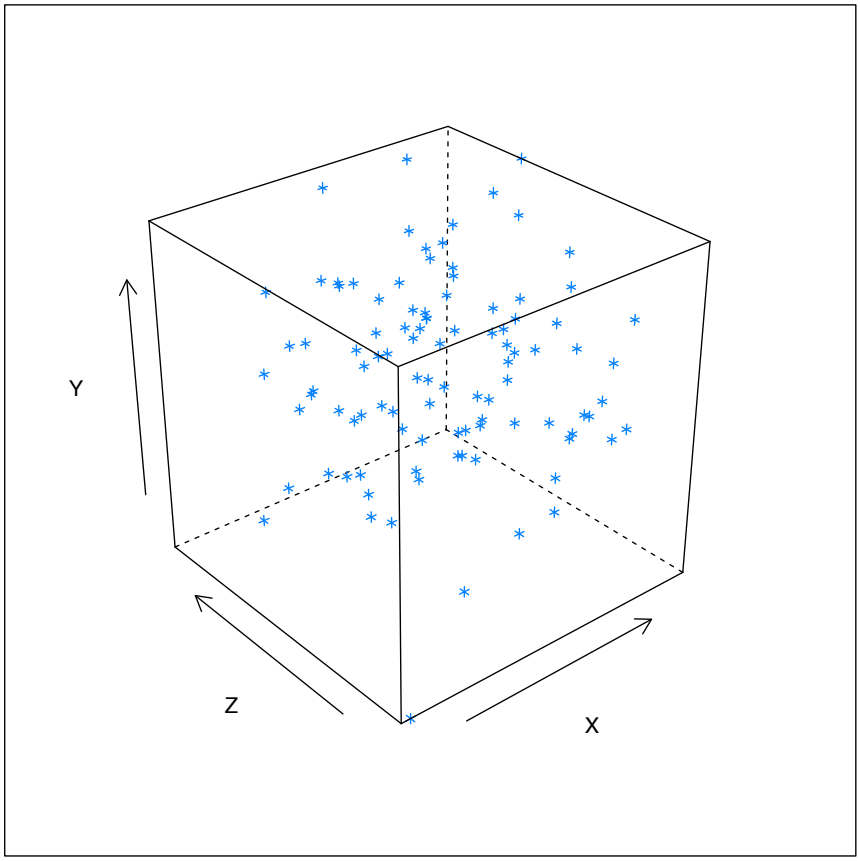


FIGURE 2.3: *As in Figure 2.2, but shifting so that the correlation between X and Z is now -0.1, though the marginal distributions, and the distribution of Y given X and Z, are unchanged. (See accompanying R file for commands used to make this figure.)*

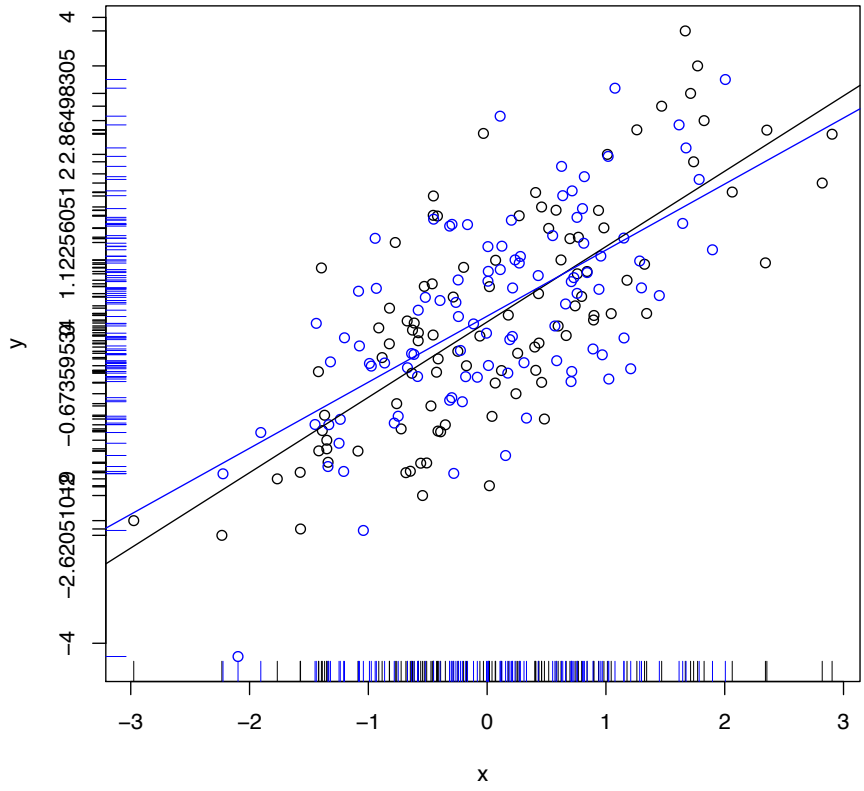


FIGURE 2.4: *Joint distribution of X and Y from Figure 2.2 (black, with a positive correlation between X and Z) and from Figure 2.3 (blue, with a negative correlation between X and Z). Tick-marks on the axes show the marginal distributions, which are manifestly little-changed. (See accompanying R file for commands.) [[TODO: fix labels on y axis]]*

We'll return to this issue of omitted variables when we look at causal inference in Part IV.

2.2.3 Errors in Variables

It is often the case that the input features we can actually measure, $\vec{X}$, are distorted versions of some other variables $\vec{U}$ we wish we could measure, but can't:

$$\vec{X} = \vec{U} + \vec{\eta} \tag{2.32}$$

with $\vec{\eta}$ being some sort of noise. Regressing Y on $\vec{X}$ then gives us what's called an **errors-in-variables** problem.

In one sense, the errors-in-variables problem is huge. We are often much more interested in the connections between *actual* variables in the *real* world, than with our imperfect, noisy measurements of them. Endless ink has been spilled, for instance, on what determines students' examination scores. One thing commonly thrown into the regression — a feature included in $\vec{X}$ — is the income of children's families. But this is typically *not* measured with absolute precision[5], so what we are really interested in — the relationship between actual income and school performance — is not what we are estimating in our regression. Typically, adding noise to the input features makes them less predictive of the response — in linear regression, it tends to push $\widehat{\beta}$ closer to zero than it would be if we could regress Y on $\vec{U}$.

On account of the error-in-variables problem, some people get very upset when they see imprecisely-measured features as inputs to a regression. Some of them, in fact, demand that the input variables be measured *exactly*, with no noise whatsoever.

This position, however, is crazy, and indeed there's a sense in which errors-in-variables isn't a problem at all. Our earlier reasoning about how to find the optimal linear predictor of Y from $\vec{X}$ remains valid whether something like Eq. 2.32 is true or not. Similarly, the reasoning in Ch. 1 about the actual regression function being the over-all optimal predictor, etc., is unaffected. If in the future we will continue to have $\vec{X}$ rather than $\vec{U}$ available to us for prediction, then Eq. 2.32 is irrelevant *for prediction*. Without better data, the relationship of Y to $\vec{U}$ is just one of the unanswerable questions the world is full of, as much as "what song the sirens sang, or what name Achilles took when he hid among the women".

Now, if you are willing to assume that $\vec{\eta}$ is a *very* nicely behaved Gaussian and you know its variance, then there are standard solutions to the error-in-variables problem for linear regression — ways of estimating the coefficients you'd get if you could regress Y on $\vec{U}$. I'm not going to go over them, partly because they're in standard textbooks, but mostly because the assumptions are hopelessly demanding.[6]

[5] One common proxy is to ask *the child* what they think their family income is. (I didn't believe that either when I first heard about it.)

[6] Non-parametric error-in-variable methods are an active topic of research (Carroll *et al.*, 2009).

2.2.4 Transformation

Let's look at a simple non-linear example, $Y|X \sim \mathcal{N}(\log X, 1)$. The problem with smoothing data from this source on to a straight line is that the true regression curve isn't very straight, $\mathbb{E}[Y|X=x] = \log x$. (Figure 2.5.) This suggests replacing the variables we have with ones where the relationship *is* linear, and then undoing the transformation to get back to what we actually measure and care about.

We have two choices: we can transform the response Y, or the predictor X. Here transforming the response would mean regressing $\exp Y$ on X, and transforming the predictor would mean regressing Y on $\log X$. Both kinds of transformations can be worth trying, but transforming the predictors is, in my experience, often a better bet, for three reasons.

1. Mathematically, $\mathbb{E}[f(Y)] \neq f(\mathbb{E}[Y])$. A mean-squared optimal prediction of $f(Y)$ is not necessarily close to the transformation of an optimal prediction of Y. And Y is, presumably, what we really want to predict. (Here, however, it works out.)

2. Imagine that $Y = \sqrt{X} + \log Z$. There's not going to be any particularly nice transformation of Y that makes everything linear; though there will be transformations of the features.

3. This generalizes to more complicated models with features built from multiple covariates.

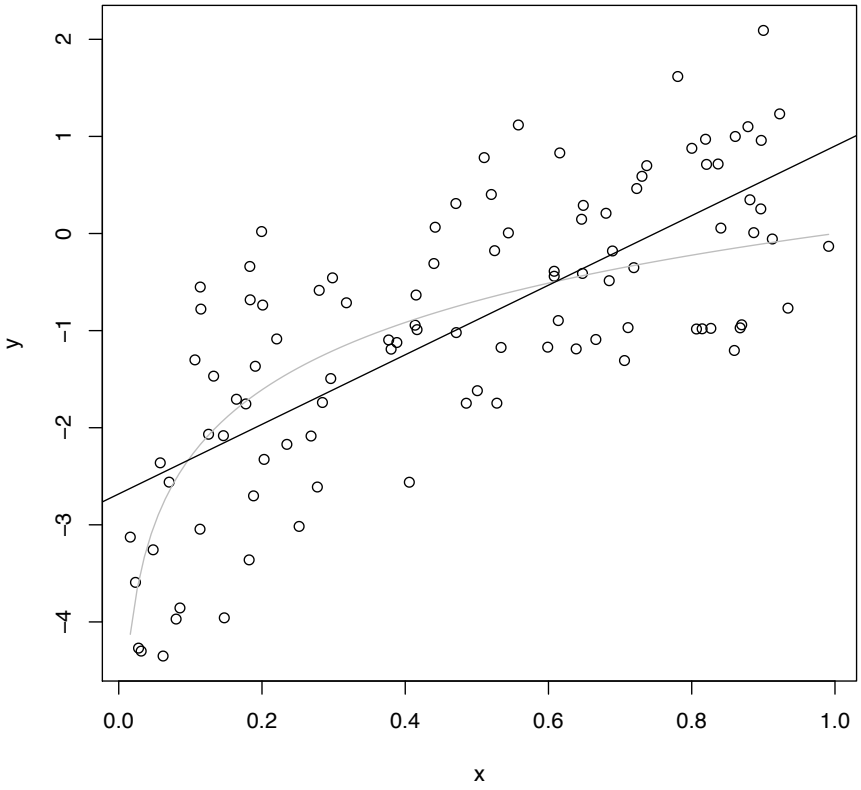
4. Suppose that we are in luck and $Y = \mu(X) + \epsilon$, with ϵ independent of X, and Gaussian, so all the usual default calculations about statistical inference apply. Then it will generally *not* be the case that $f(Y) = s(X) + \eta$, with η a Gaussian random variable independent of X. In other words, transforming Y completely messes up the noise model. (Consider the simple case where we take the logarithm of Y. Gaussian noise after the transformation implies log-normal noise before the transformation. Conversely, Gaussian noise before the transformation implies a very weird, nameless noise distribution after the transformation.)

Figure 2.6 shows the effect of these transformations. Here transforming the predictor does, indeed, work out more nicely; but of course I chose the example so that it does so.

To expand on that last point, imagine a model like so:

$$\mu(\vec{x}) = \sum_{j=1}^{q} c_j f_j(\vec{x}) \tag{2.33}$$

If we know the functions f_j, we can estimate the optimal values of the coefficients c_j by least squares — this is a regression of the response on new features, which happen to be defined in terms of the old ones. Because the parameters are outside the functions, that part of the estimation works just like linear regression. Models embraced under the heading of Eq. 2.33 include linear regressions with **interactions**




```
x <- runif(100)
y <- rnorm(100,mean=log(x),sd=1)
plot(y~x)
curve(log(x),add=TRUE,col="grey")
abline(lm(y~x))
```

FIGURE 2.5: *Sample of data for* $Y|X \sim \mathcal{N}(\log X, 1)$. *(Here* $X \sim \text{Unif}(0,1)$*, and all logs are natural logs.) The true, logarithmic regression curve is shown in grey (because it's not really observable), and the linear regression fit is shown in black.*

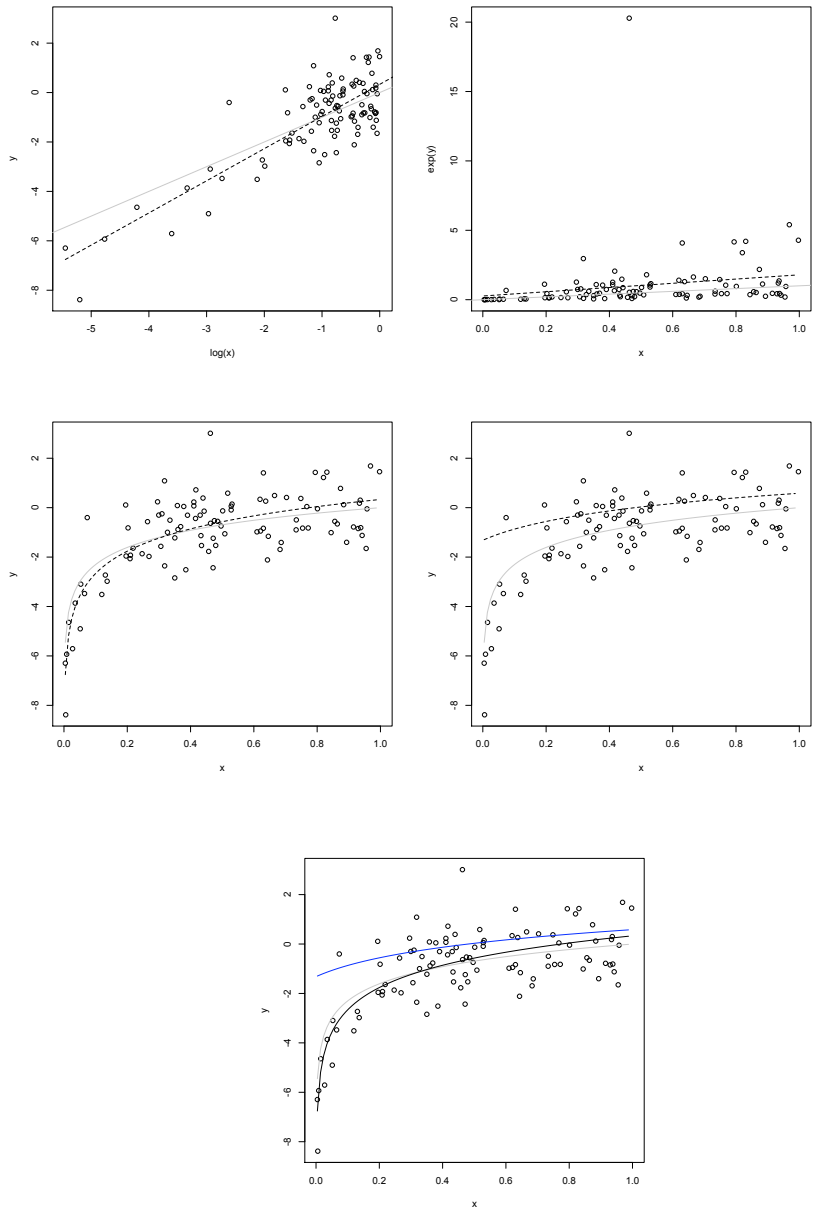


FIGURE 2.6: *Transforming the predictor (left column) and the response (right) in the data from Figure 2.5, shown in both the transformed coordinates (top) and the original coordinates (middle). The bottom figure super-imposes the two estimated curves (transformed X in black, transformed Y in blue). The true regression curve is always in grey. (R code deliberately omitted; can you reproduce this?) [[TODO: Re-create from code, but omit from purled code output]]*

between the predictor variables (set $f_j = x_i x_k$, for various combinations of i and k), and **polynomial regression**. There is however nothing magical about using products and powers of the predictor variables; we could regress Y on $\sin x$, $\sin 2x$, $\sin 3x$, etc.

To apply models like Eq. 2.33, we can either (a) fix the functions f_j in advance, based on guesses about what should be good features for this problem; (b) fix the functions in advance by always using some "library" of mathematically convenient functions, like polynomials or trigonometric functions; or (c) try to *find* good functions from the data. Option (c) takes us beyond the realm of linear regression as such, into things like **splines** (Chapter 8) and **additive models** (Chapter 9). It is also possible to search for transformations of *both* sides of a regression model; see Breiman and Friedman (1985) and, for an R implementation, Spector *et al.* (2013).

2.3 Adding Probabilistic Assumptions

The usual treatment of linear regression adds many more probabilistic assumptions, namely that

$$Y|\vec{X} \sim \mathcal{N}(\vec{X} \cdot \beta, \sigma^2) \tag{2.34}$$

and that Y values are independent conditional on their $\vec{X}$ values. So now we are *assuming* that the regression function is exactly linear; we are *assuming* that at each $\vec{X}$ the scatter of Y around the regression function is Gaussian; we are *assuming* that the variance of this scatter is constant; and we are *assuming* that there is no dependence between this scatter and anything else.

None of these assumptions was needed in deriving the optimal linear predictor. None of them is so mild that it should go without comment or without at least some attempt at testing.

Leaving that aside just for the moment, why make those assumptions? As you know from your earlier classes, they let us write down the likelihood of the observed responses $y_1, y_2, \ldots y_n$ (conditional on the covariates $\vec{x}_1, \ldots \vec{x}_n$), and then estimate β and σ^2 by maximizing this likelihood. As you also know, the maximum likelihood estimate of β is exactly the same as the β obtained by minimizing the residual sum of squares. This coincidence would not hold in other models, with non-Gaussian noise.

We saw earlier that $\widehat{\beta}$ is consistent under comparatively weak assumptions — that it converges to the optimal coefficients. But then there might, possibly, still be other estimators are also consistent, but which converge faster. If we make the extra statistical assumptions, so that $\widehat{\beta}$ is also the maximum likelihood estimate, we can lay that worry to rest. The MLE is generically (and certainly here!) **asymptotically efficient**, meaning that it converges as fast as any other consistent estimator, at least in the long run. So we are not, so to speak, wasting any of our data by using the MLE.

A further advantage of the MLE is that, as $n \to \infty$, its sampling distribution is itself a Gaussian, centered around the true parameter values. This lets us calculate standard errors and confidence intervals quite easily. Here, with the Gaussian assumptions, much more exact statements can be made about the distribution of $\widehat{\beta}$

around β. You can find the formulas in any textbook on regression, so I won't get into that.

We can also use a general property of MLEs for model testing. Suppose we have two classes of models, Ω and ω. Ω is the general case, with p parameters, and ω is a special case, where some of those parameters are constrained, but $q < p$ of them are left free to be estimated from the data. The constrained model class ω is then **nested** within Ω. Say that the MLEs with and without the constraints are, respectively, $\widehat{\Theta}$ and $\widehat{\theta}$, so the maximum log-likelihoods are $L(\widehat{\Theta})$ and $L(\widehat{\theta})$. Because it's a maximum over a larger parameter space, $L(\widehat{\Theta}) \geq L(\widehat{\theta})$. On the other hand, if the true model really is in ω, we'd expect the constrained and unconstrained estimates to be converging. It turns out that the difference in log-likelihoods has an asymptotic distribution which doesn't depend on any of the model details, namely

$$2\left[L(\widehat{\Theta}) - L(\widehat{\theta})\right] \rightsquigarrow \chi^2_{p-q} \tag{2.35}$$

That is, a χ^2 distribution with one degree of freedom for each extra parameter in Ω (that's why they're called "degrees of freedom").[7]

This approach can be used to test particular restrictions on the model, and so it is sometimes used to assess whether certain variables influence the response. This, however, gets us into the concerns of the next section.

2.3.1 Examine the Residuals

By construction, the residuals of a fitted linear regression have mean zero and are uncorrelated with the predictor variables. If the usual probabilistic assumptions hold, however, they have many other properties as well.

1. The residuals have a Gaussian distribution at each $\vec{x}$.
2. The residuals have the *same* Gaussian distribution at each $\vec{x}$, i.e., they are *independent* of the predictor variables. In particular, they must have the same variance (i.e., they must be homoskedastic).
3. The residuals are *independent of each other*. In particular, they must be *uncorrelated* with each other.

These properties — Gaussianity, homoskedasticity, lack of correlation — are all *testable* properties. When they all hold, we say that the residuals are **white noise.** One would never expect them to hold exactly in any finite sample, but if you do test for them and find them strongly violated, you should be extremely suspicious of your model. These tests are much more important than checking whether the coefficients are significantly different from zero.

Every time someone uses linear regression with the standard assumptions for inference and does *not* test whether the residuals are white noise, an angel loses its wings.

[7] If you assume the noise is Gaussian, the left-hand side of Eq. 2.35 can be written in terms of various residual sums of squares. However, the equation itself remains valid under other noise distributions, which just change the form of the likelihood function. See Appendix I.

2.3.2 On Significant Coefficients

If all the usual distributional assumptions hold, then t-tests can be used to decide whether particular coefficients are statistically-significantly different from zero. Pretty much any piece of statistical software, R very much included, reports the results of these tests automatically. It is far too common to seriously over-interpret those results, for a variety of reasons.

Begin with what hypothesis, exactly, is being tested when R (or whatever) runs those t-tests. Say, without loss of generality, that there are p predictor variables, $\vec{X} = (X_1, \ldots X_p)$, and that we are testing the coefficient on X_p. Then the null hypothesis is not just "$\beta_p = 0$", but "$\beta_p = 0$ in a linear model which also includes $X_1, \ldots X_{p-1}$, and nothing else". The alternative hypothesis is not just "$\beta_p \neq 0$", but "$\beta_p \neq 0$ in a model which also includes $X_1, \ldots X_{p-1}$, but nothing else". The optimal linear coefficient on X_p will depend on not just on the relationship between X_p and the response Y, but also on which other variables are included in the model. The t-test checks whether adding X_p really improves predictions more than would be expected, under all these assumptions, if one is already using all the other variables, and only those other variables. It does not and cannot test whether X_p is important in any absolute sense.

Even if you are willing to say "Yes, all I really want to know about this variable is whether adding it to the model really helps me predict in a linear approximation", remember that the question which a t-test answers is whether adding that variable will help *at all*. Of course, as you know from your regression class, and as we'll see in more detail in Chapter 3, expanding the model never hurts its performance on the *training* data. The point of the t-test is to gauge whether the improvement in prediction is small enough to be due to chance, or so large, *compared to what noise could produce*, that one could confidently say the variable adds *some* predictive ability. This has several implications which are insufficiently appreciated among users.

In the first place, tests on individual coefficients can seem to contradict tests on groups of coefficients. Adding multiple variables to the model could significantly improve the fit (as checked by, say, a partial F test), even if *none* of the coefficients is significant on its own. In fact, every single coefficient in the model could be insignificant, while the model as a whole is highly significant (i.e., better than a flat line).

In the second place, it's worth thinking about which variables will show up as statistically significant. Remember that the t-statistic is $\widehat{\beta}_i / \mathrm{se}(\widehat{\beta}_i)$, the ratio of the estimated coefficient to its standard error. We saw above that $\mathbb{V}\left[\widehat{\beta} | \mathbf{X} = \mathbf{x}\right] = \frac{\sigma^2}{n}\left(n^{-1}\mathbf{x}^T\mathbf{x}\right)^{-1} \to n^{-1}\sigma^2\mathbf{v}^{-1}$. This means that the standard errors will shrink as the sample size grows, so more and more variables will become significant as we get more data — but how much data we collect is irrelevant to how the process we're studying actually works. Moreover, at a fixed sample size, the coefficients with smaller standard errors will tend to be the ones whose variables have more variance, and whose variables are less correlated with the other predictors. High input variance and low correlation help us *estimate* the coefficient precisely, but, again, they have nothing to

do with whether the input variable actually *influences* the response a lot.

To sum up, it is *never* the case that statistical significance is the same as scientific, real-world significance. The most important variables are *not* those with the largest-magnitude t statistics or smallest p-values. Statistical significance is always about what "signals" can be picked out clearly from background noise[8]. In the case of linear regression coefficients, statistical significance runs together the size of the coefficients, how bad the linear regression model is, the sample size, the variance in the input variable, and the correlation of that variable with all the others.

Of course, even the limited "does it help linear predictions enough to bother with?" utility of the usual t-test (and F-test) calculations goes away if the standard distributional assumptions do not hold, so that the calculated p-values are just wrong. One can sometimes get away with using bootstrapping (Chapter 6) to get accurate p-values for standard tests under non-standard conditions.

2.4 Linear Regression Is Not the Philosopher's Stone

The philosopher's stone, remember, was supposed to be able to transmute base metals (e.g., lead) into the perfect metal, gold (Eliade, 1971). Many people treat linear regression as though it had a similar ability to transmute a correlation matrix into a scientific theory. In particular, people often argue that:

1. because a variable has a significant regression coefficient, it must influence the response;
2. because a variable has an insignificant regression coefficient, it must not influence the response;
3. if the input variables change, we can predict how much the response will change by plugging in to the regression.

All of this is wrong, or at best right only under very particular circumstances.

We have already seen examples where influential variables have regression coefficients of zero. We have also seen examples of situations where a variable with no influence has a non-zero coefficient (e.g., because it is correlated with an omitted variable which does have influence). *If* there are no nonlinearities and *if* there are no omitted influential variables and *if* the noise terms are always independent of the predictor variables, are we good?

No. Remember from Equation 2.5 that the optimal regression coefficients depend on both the marginal distribution of the predictors and the joint distribution (covariances) of the response and the predictors. There is no reason whatsoever to suppose that if we *change* the system, this will leave the conditional distribution of the response alone.

A simple example may drive the point home. Suppose we surveyed all the cars in Pittsburgh, recording the maximum speed they reach over a week, and how often they are waxed and polished. I don't think anyone doubts that there will be a

[8] In retrospect, it might have been clearer to say "statistically *detectable*" rather than "statistically *significant*".

positive correlation here, and in fact that there will be a positive regression coefficient, even if we add in many other variables as predictors. Let us even postulate that the relationship is linear (perhaps after a suitable transformation). Would anyone believe that polishing cars will make them go faster? Manifestly not. But this is exactly how people interpret regressions in all kinds of applied fields — instead of saying polishing makes cars go faster, it might be saying that receiving targeted ads makes customers buy more, or that consuming dairy foods makes diabetes progress faster, or .... Those claims might be *true*, but the regressions could easily come out the same way were the claims false. Hence, the regression results provide little or no *evidence* for the claims.

Similar remarks apply to the idea of using regression to "control for" extra variables. If we are interested in the relationship between one predictor, or a few predictors, and the response, it is common to add a bunch of other variables to the regression, to check both whether the apparent relationship might be due to correlations with something else, and to "control for" those other variables. The regression coefficient is interpreted as how much the response would change, on average, if the predictor variable were increased by one unit, "holding everything else constant". There is a very particular sense in which this is true: it's a prediction about the difference in expected responses (conditional on the given values for the other predictors), assuming that the form of the regression model is right, *and* that observations are randomly drawn from the same population we used to fit the regression.

In a word, what regression does is *probabilistic* prediction. It says what will happen if we keep drawing from the same population, but *select* a sub-set of the observations, namely those with given values of the predictor variables. A **causal** or **counter-factual** prediction would say what would happen if we (or Someone) *made* those variables take on those values. There may be no difference between selection and intervention, in which case regression can work as a tool for causal inference[9]; but in general there is. Probabilistic prediction is a worthwhile endeavor, but it's important to be clear that this is what regression does. There are techniques for doing actually causal prediction, which we will explore in Part IV.

Every time someone thoughtlessly uses regression for causal inference, an angel not only loses its wings, but is cast out of Heaven and falls in extremest agony into the everlasting fire.

2.5 Further Reading

There are many excellent textbooks on linear regression. Among them, I would mention Weisberg (1985) for general statistical good sense, along with Faraway (2004) for R practicalities, and Hastie *et al.* (2009) for emphasizing connections to more advanced methods. Berk (2004) omits the details those books cover, but is superb on the big picture, and especially on what must be assumed in order to do certain things with linear regression and what cannot be done under any assumption.

[9] In particular, if our model was estimated from data where Someone *assigned* values of the predictor variables in a way which breaks possible dependencies with omitted variables and noise — either by randomization or by experimental control — then regression can, in fact, work for causal inference.

2.6 Exercises

1. (a) Write the expected squared error of a linear predictor with slopes $\vec{b}$ and intercept b_0 as a function of those coefficients.
   (b) Find the derivatives of the expected squared error with respect to all the coefficients.
   (c) Show that when we set all the derivatives to zero, the solutions are Eq. 2.5 and 2.6.
2. Show that the expected error of the optimal linear predictor, $\mathbb{E}\left[Y - \vec{X} \cdot \beta\right]$, is zero.
3. Convince yourself that if the real regression function is linear, β does not depend on the marginal distribution of X. You may want to start with the case of one predictor variable.
4. Run the code from Figure 2.5. Then replicate the plots in Figure 2.6.
5. Which kind of transformation is superior for the model where $Y|X \sim \mathcal{N}(\sqrt{X}, 1)$?