23:38 Wednesday 17th February, 2016

Chapter 3

Evaluating Statistical Models: Error and Inference

3.1 What Are Statistical Models For? Summaries, Forecasts, Simulators

There are (at least) three ways we can use statistical models in data analysis: as summaries of the data, as predictors, and as simulators.

The least demanding use of a model is to summarize the data — to use it for **data reduction**, or **compression**. Just as the sample mean or sample quantiles can be descriptive statistics, recording some features of the data and saying nothing about a population or a generative process, we could use estimates of a model's parameters as descriptive summaries. Rather than remembering all the points on a scatter-plot, say, we'd just remember what the OLS regression surface was.

It's hard to be wrong about a summary, unless we just make a mistake. (It may not be *helpful* for us later, but that's different.) When we say "the slope which minimized the sum of squares was 4.02", we make no claims about anything *but* the training data. That statement relies on no assumptions, beyond our calculating correctly. But it also asserts nothing about the rest of the world. As soon as we try to connect our training data to anything else, we start relying on assumptions, and we run the risk of being wrong.

Probably the most common connection to want to make is to say what *other* data will look like — to make predictions. In a statistical model, with random variables, we do not anticipate that our predictions will ever be *exactly* right, but we also anticipate that our mistakes will show stable probabilistic patterns. We can evaluate predictions based on those patterns of error — how big is our typical mistake? are we biased in a particular direction? do we make a lot of little errors or a few huge ones?

Statistical inference about model parameters — estimation and hypothesis testing — can be seen as a kind of prediction, extrapolating from what we saw in a small piece of data to what we would see in the whole population, or whole process. When

we *estimate* the regression coefficient $\hat{b} = 4.02$, that involves predicting new values of the dependent variable, but also predicting that if we repeated the experiment and re-estimated $\hat{b}$, we'd get a value close to 4.02.

Using a model to summarize old data, or to predict new data, doesn't commit us to assuming that the model describes the process which generates the data. But we often want to do that, because we want to interpret parts of the model as aspects of the real world. We think that in neighborhoods where people have more money, they spend more on houses — perhaps each extra $1000 in income translates into an extra $4020 in house prices. Used this way, statistical models become stories about how the data were generated. If they are accurate, we should be able to use them to *simulate* that process, to step through it and produce something that looks, probabilistically, just like the actual data. This is often what people have in mind when they talk about *scientific* models, rather than just statistical ones.

An example: if you want to predict where in the night sky the planets will be, you can actually do very well with a model where the Earth is at the center of the universe, and the Sun and everything else revolve around it. You can even estimate, from data, how fast Mars (for example) goes around the Earth, or where, in this model, it should be tonight. But, since the Earth is *not* at the center of the solar system, those parameters don't actually refer to anything in reality. They are just mathematical fictions. On the other hand, we can also predict where the planets will appear in the sky using models where all the planets orbit the Sun, and the parameters of the orbit of Mars in that model *do* refer to reality.[1]

This chapter focuses on evaluating predictions, for three reasons. First, often we just want prediction. Second, if a model can't even predict well, it's hard to see how it could be right scientifically. Third, often the best way of checking a scientific model is to turn some of its implications into statistical predictions.

3.2 Errors, In and Out of Sample

With any predictive model, we can gauge how well it works by looking at its errors. We want these to be small; if they can't be small all the time we'd like them to be small on average. We may also want them to be patternless or unsystematic (because if there was a pattern to them, why not adjust for that, and make smaller mistakes). We'll come back to patterns in errors later, when we look at specification testing (Chapter 10). For now, we'll concentrate on the size of the errors.

To be a little more mathematical, we have a data set with points $\mathbf{z}_n = z_1, z_2, \ldots z_n$. (For regression problems, think of each data point as the pair of input and output values, so $z_i = (x_i, y_i)$, with x_i possibly a vector.) We also have various possible models, each with different parameter settings, conventionally written θ. For regression, θ tells us which regression function to use, so $m_\theta(x)$ or $m(x;\theta)$ is the prediction we make at point x with parameters set to θ. Finally, we have a **loss function** L which tells us how big the error is when we use a certain θ on a certain data point, $L(z, \theta)$.

[1] We can be pretty sure of this, because we use our parameter estimates to send our robots to Mars, and they get there.

For mean-squared error, this would just be

$$L(z, \theta) = (y - m_\theta(x))^2 \tag{3.1}$$

But we could also use the mean absolute error

$$L(z, \theta) = |y - m_\theta(x)| \tag{3.2}$$

or many other loss functions. Sometimes we will actually be able to measure how costly our mistakes are, in dollars or harm to patients. If we had a model which gave us a distribution for the data, then $p_\theta(z)$ would a probability density at z, and a typical loss function would be the negative log-likelihood, $-\log m_\theta(z)$. No matter what the loss function is, I'll abbreviate the sample average of the loss over the whole data set by $\overline{L}(\mathbf{z}_n, \theta)$.

What we would like, ideally, is a predictive model which has zero error on future data. We basically never achieve this:

- The world just really is a noisy and stochastic place, and this means even the true, ideal model has non-zero error.[2] This corresponds to the first, σ_x^2, term in the bias-variance decomposition, Eq. 1.27 from Chapter 1.
- Our models are usually more or less **mis-specified**, or, in plain words, wrong. We hardly ever get the functional form of the regression, the distribution of the noise, the form of the causal dependence between two factors, etc., *exactly* right.[3] This is the origin of the bias term in the bias-variance decomposition. Of course we can get any of the details in the model specification *more or less* wrong, and we'd prefer to be less wrong.
- Our models are never perfectly estimated. Even if our data come from a perfect IID source, we only ever have a finite sample, and so our parameter estimates are (almost!) never quite the true, infinite-limit values. This is the origin of the variance term in the bias-variance decomposition. But as we get more and more data, the sample should become more and more representative of the whole process, and estimates should converge too.

So, because our models are flawed, we have limited data and the world is stochastic, we cannot expect even the best model to have zero error. Instead, we would like to minimize the **expected error**, or **risk**, or **generalization error**, on new data.

What we would like to do is to minimize the risk or expected loss

$$\mathbb{E}[L(Z, \theta)] = \int L(z, \theta) p(z) dz \tag{3.3}$$

To do this, however, we'd have to be able to calculate that expectation. Doing that would mean knowing the distribution of Z — the joint distribution of X and Y, for

[2] This is so even if you believe in some kind of ultimate determinism, because the variables we plug in to our predictive models are not complete descriptions of the physical state of the universe, but rather immensely coarser, and this coarseness shows up as randomness.

[3] Except maybe in fundamental physics, and even there our predictions are about our fundamental theories *in the context of experimental set-ups*, which we never model in complete detail.

the regression problem. Since we don't know the true joint distribution, we need to approximate it somehow.

A natural approximation is to use our training data $\mathbf{z}_n$. For each possible model θ, we can could calculate the sample mean of the error on the data, $\overline{L}(\mathbf{z}_n, \theta)$, called the **in-sample loss** or the **empirical risk**. The simplest strategy for estimation is then to pick the model, the value of θ, which minimizes the in-sample loss. This strategy is imaginatively called **empirical risk minimization**. Formally,

$$\widehat{\theta_n} \equiv \operatorname*{argmin}_{\theta \in \Theta} \overline{L}(\mathbf{z}_n, \theta) \tag{3.4}$$

This means picking the regression which minimizes the sum of squared errors, or the density with the highest likelihood[4]. This is what you've usually done in statistics courses so far, and it's very natural, but it does have some issues, notably optimism and over-fitting.

The problem of optimism comes from the fact that our training data isn't perfectly representative. The in-sample loss is a sample average. By the law of large numbers, then, we anticipate that, for each θ,

$$\overline{L}(\mathbf{z}_n, \theta) \to \mathbb{E}[L(Z, \theta)] \tag{3.5}$$

as $n \to \infty$. This means that, with enough data, the in-sample error is a good approximation to the generalization error of any given model θ. (Big samples are representative of the underlying population or process.) But this does *not* mean that the in-sample performance of $\hat{\theta}$ tells us how well it will generalize, because we purposely picked it to match the training data $\mathbf{z}_n$. To see this, notice that the in-sample loss equals the risk plus sampling noise:

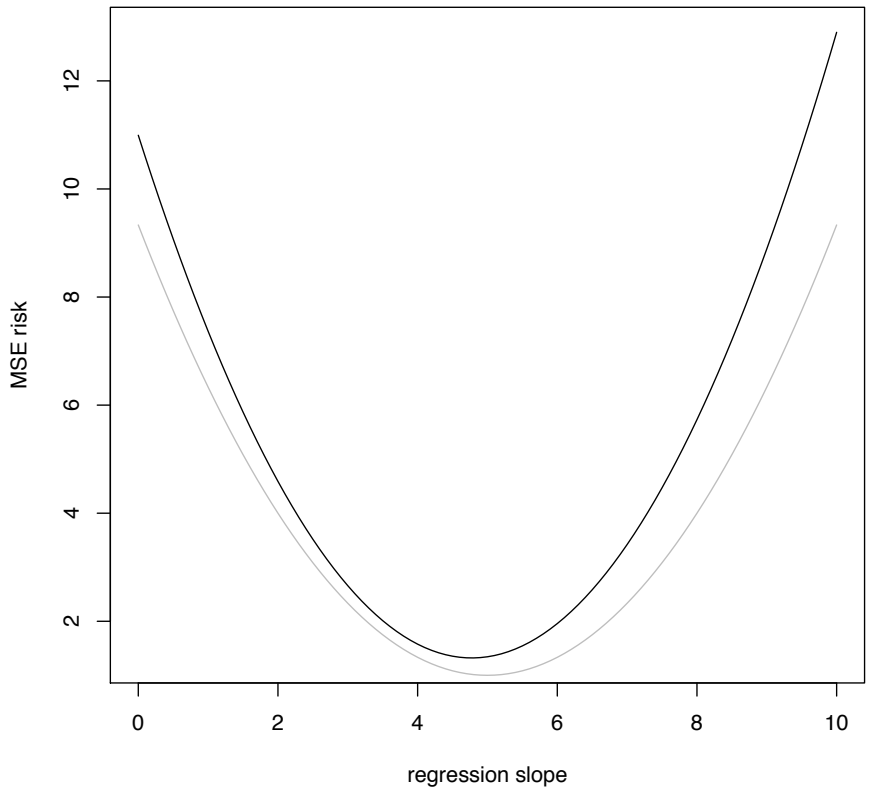
$$\overline{L}(\mathbf{z}_n, \theta) = \mathbb{E}[L(\mathbf{Z}, \theta)] + \eta_n(\theta) \tag{3.6}$$

Here $\eta_n(\theta)$ is a random term which has mean zero, and represents the effects of having only a finite quantity of data, of size n, rather than the complete probability distribution. (I write it $\eta_n(\theta)$ as a reminder that different values of θ are going to be affected differently by the same sampling fluctuations.) The problem, then, is that the model which minimizes the in-sample loss could be one with good generalization performance ($\mathbb{E}[L(\mathbf{Z}, \theta)]$ is small), or it could be one which got very lucky ($\eta_n(\theta)$ was large and negative):

$$\widehat{\theta_n} = \operatorname*{argmin}_{\theta \in \Theta} \left(\mathbb{E}[L(Z, \theta)] + \eta_n(\theta) \right) \tag{3.7}$$

We only want to minimize $\mathbb{E}[L(Z, \theta)]$, but we can't separate it from $\eta_n(\theta)$, so we're almost surely going to end up picking a $\widehat{\theta_n}$ which was more or less lucky ($\eta_n < 0$) as well as good ($\mathbb{E}[L(Z, \theta)]$ small). This is the reason why picking the model which best fits the data tends to exaggerate how well it will do in the future (Figure 3.1).

[4]Remember, maximizing the likelihood is the same as maximizing the log-likelihood, because log is an increasing function. Therefore maximizing the likelihood is the same as *minimizing* the *negative* log-likelihood.




```
n <- 20
theta <- 5
x <- runif(n)
y <- x * theta + rnorm(n)
empirical.risk <- function(b) {
    mean((y - b * x)^2)
}
true.risk <- function(b) {
    1 + (theta - b)^2 * (0.5^2 + 1/12)
}
curve(Vectorize(empirical.risk)(x), from = 0, to = 2 * theta, xlab = "regression slope",
    ylab = "MSE risk")
curve(true.risk, add = TRUE, col = "grey")
```

FIGURE 3.1: *Empirical and generalization risk for regression through the origin,* $Y = \theta X + \epsilon$, $\epsilon \sim \mathcal{N}(0,1)$, *with true* $\theta = 5$, *and* $X \sim \text{Unif}(0,1)$. *Black: MSE on a particular sample (*$n = 20$*) as a function of slope, minimized at* $\hat{\theta} = 4.78$. *Grey: true or generalization risk (Exercise 2). The gap between the curves is the text's* $\eta_n(\theta)$. *(Code comments online.)*

Again, by the law of large numbers $\eta_n(\theta) \to 0$ for each θ, but now we need to worry about how fast it's going to zero, and whether that rate depends on θ. Suppose we knew that $\min_\theta \eta_n(\theta) \to 0$, or $\max_\theta |\eta_n(\theta)| \to 0$. Then it would follow that $\eta_n(\widehat{\theta_n}) \to 0$, and the over-optimism in using the in-sample error to approximate the generalization error would at least be shrinking. If we knew how fast $\max_\theta |\eta_n(\theta)|$ was going to zero, we could even say something about how much bigger the true risk was likely to be. A lot of more advanced statistics and machine learning theory is thus about uniform laws of large numbers (showing $\max_\theta |\eta_n(\theta)| \to 0$) and rates of convergence.

Learning theory is a beautiful, deep, and practically important subject, but also a subtle and involved one. (See §3.6 for references.) To stick closer to analyzing real data, and to not turn this into an advanced probability class, I will only talk about some more-or-less heuristic methods, which are good enough for many purposes.

3.3 Over-Fitting and Model Selection

The big problem with using the in-sample error is related to over-optimism, but at once trickier to grasp and more important. This is the problem of **over-fitting**. To illustrate it, let's start with Figure 3.2. This has the twenty X values from a Gaussian distribution, and $Y = 7X^2 - 0.5X + \epsilon$, $\epsilon \sim \mathcal{N}(0,1)$. That is, the true regression curve is a parabola, with additive and independent Gaussian noise. Let's try fitting this — but pretend that we didn't know that the curve was a parabola. We'll try fitting polynomials of different degrees in x — degree 0 (a flat line), degree 1 (a linear regression), degree 2 (quadratic regression), up through degree 9. Figure 3.3 shows the data with the polynomial curves, and Figure 3.4 shows the in-sample mean squared error as a function of the degree of the polynomial.

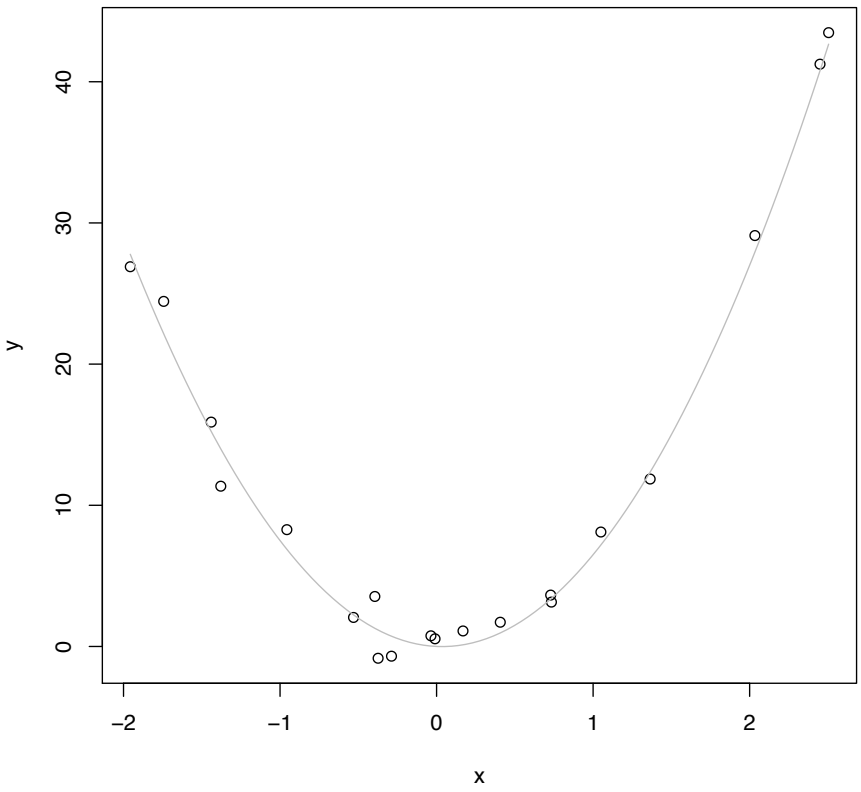
Notice that the in-sample error goes down as the degree of the polynomial increases; it has to. Every polynomial of degree p can also be written as a polynomial of degree $p+1$ (with a zero coefficient for x^{p+1}), so going to a higher-degree model can only reduce the in-sample error. Quite generally, in fact, as one uses more and more complex and flexible models, the in-sample error will get smaller and smaller.[5]

Things are quite different if we turn to the generalization error. In principle, I could calculate that for any of the models, since I know the true distribution, but it would involve calculating things like $\mathbb{E}\left[X^{18}\right]$, which won't be very illuminating. Instead, I will just draw a lot more data from the same source, twenty thousand data points in fact, and use the error of the old models on the new data as their generalization error[6]. The results are in Figure 3.5.

What is happening here is that the higher-degree polynomials — beyond degree 2 — are not just a *little* optimistic about how well they fit, they are *wildly* over-optimistic. The models which seemed to do notably better than a quadratic actu-

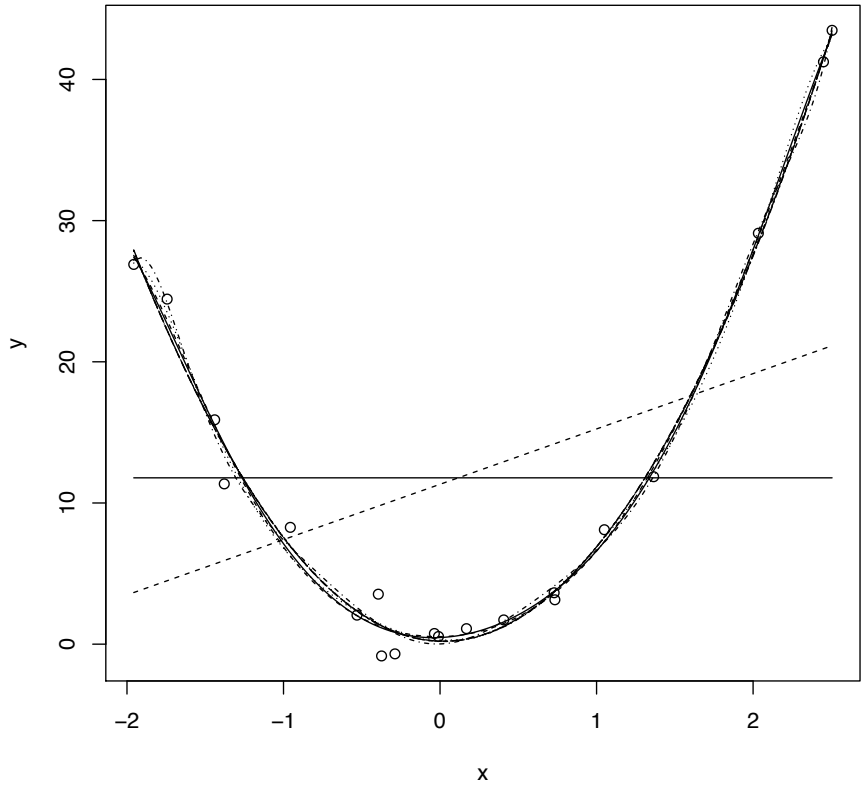
[5] In fact, since there are only 20 data points, they could all be fit exactly if the degree of the polynomials went up to 19. (Remember that any two points define a line, any three points a parabola, etc. — $p+1$ points define a polynomial of degree p which passes through them.)

[6] This works, yet again, because of the law of large numbers. In Chapters 5 and especially 6, we will see much more about replacing complicated probabilistic calculations with simple simulations.



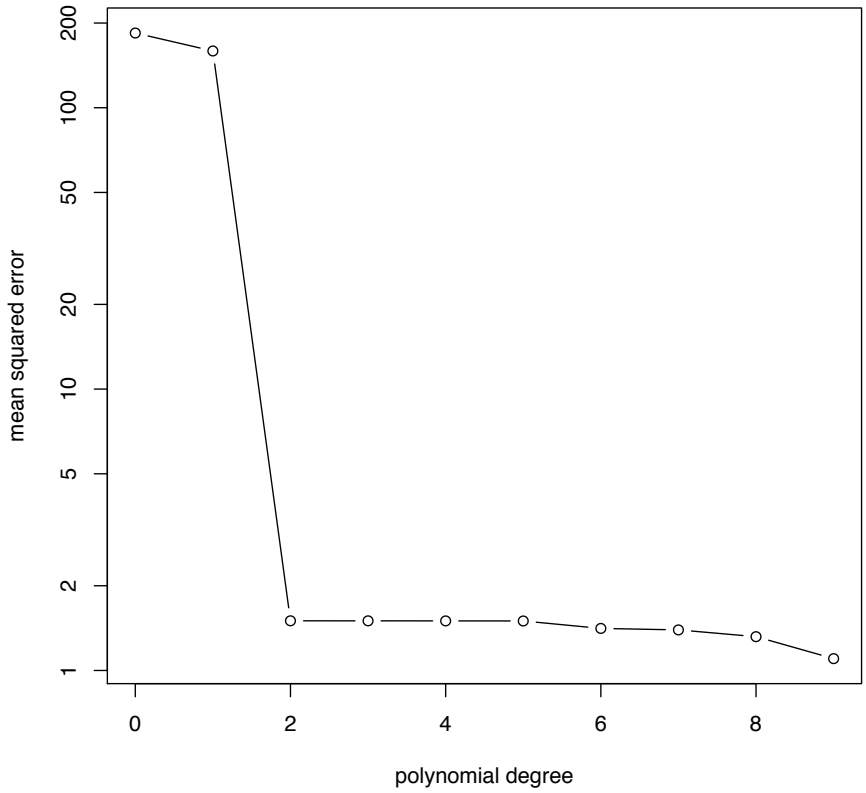

```
x = rnorm(20)
y = 7 * x^2 - 0.5 * x + rnorm(20)
plot(x, y)
curve(7 * x^2 - 0.5 * x, col = "grey", add = TRUE)
```

FIGURE 3.2: *Scatter-plot showing sample data and the true, quadratic regression curve (grey parabola).*



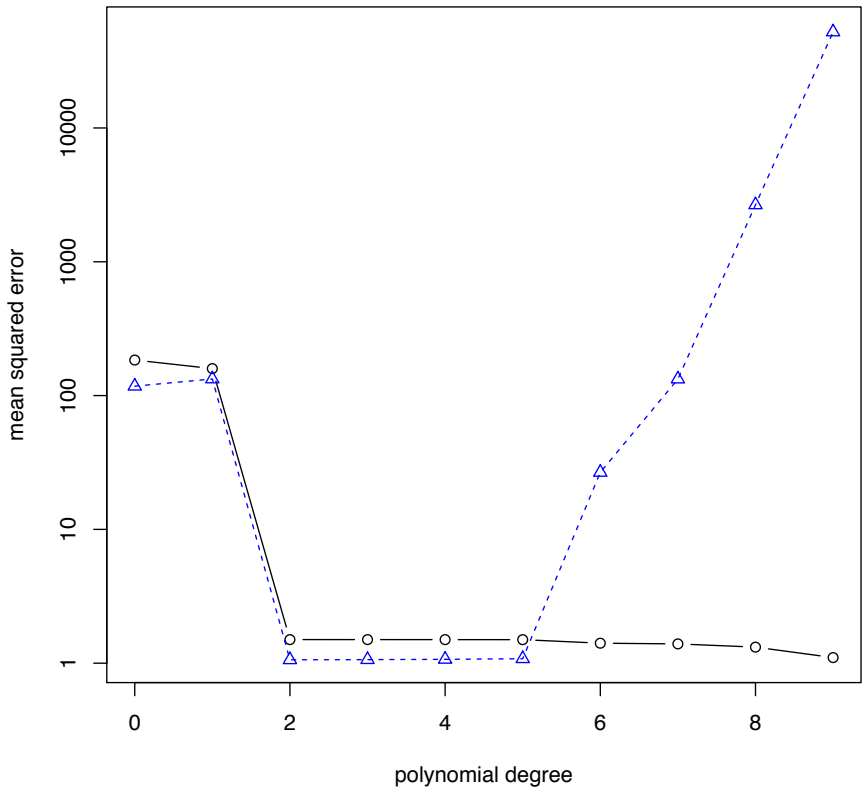

```
plot(x, y)
poly.formulae <- c("y~1", paste("y ~ poly(x,", 1:9, ")", sep = ""))
poly.formulae <- sapply(poly.formulae, as.formula)
df.plot <- data.frame(x = seq(min(x), max(x), length.out = 200))
fitted.models <- list(length = length(poly.formulae))
for (model_index in 1:length(poly.formulae)) {
    fm <- lm(formula = poly.formulae[[model_index]])
    lines(df.plot$x, predict(fm, newdata = df.plot), lty = model_index)
    fitted.models[[model_index]] <- fm
}
```

FIGURE 3.3: *Twenty training data points (dots), and ten different fitted regression lines (polynomials of degree 0 to 9, indicated by different line types).* R NOTES: *The* `poly` *command constructs orthogonal (uncorrelated) polynomials of the specified degree from its first argument; regressing on them is conceptually equivalent to regressing on* $1, x, x^2, \ldots x^{\text{degree}}$*, but more numerically stable. (See* `?poly`*.) This builds a vector of model formulae and then fits each one in turn, storing the fitted models in a new list.*




```
mse.q <- sapply(fitted.models, function(mdl) {
    mean(residuals(mdl)^2)
})
plot(0:9, mse.q, type = "b", xlab = "polynomial degree", ylab = "mean squared error",
    log = "y")
```

FIGURE 3.4: *Empirical MSE vs. degree of polynomial for the data from the previous figure. Note the logarithmic scale for the vertical axis.*




```
x.new = rnorm(20000)
y.new = 7 * x.new^2 - 0.5 * x.new + rnorm(20000)
gmse <- function(mdl) {
    mean((y.new - predict(mdl, data.frame(x = x.new)))^2)
}
gmse.q <- sapply(fitted.models, gmse)
plot(0:9, mse.q, type = "b", xlab = "polynomial degree", ylab = "mean squared error",
    log = "y", ylim = c(min(mse.q), max(gmse.q)))
lines(0:9, gmse.q, lty = 2, col = "blue")
points(0:9, gmse.q, pch = 24, col = "blue")
```

FIGURE 3.5: *In-sample error (black dots) compared to generalization error (blue triangles). Note the logarithmic scale for the vertical axis.*

ally do much, much worse. If we picked a polynomial regression model based on in-sample fit, we'd chose the highest-degree polynomial available, and suffer for it.

In this example, the more complicated models — the higher-degree polynomials, with more terms and parameters — were not actually fitting the *generalizable* features of the data. Instead, they were fitting the sampling noise, the accidents which don't repeat. That is, the more complicated models **over-fit** the data. In terms of our earlier notation, η is bigger for the more flexible models. The model which does best here is the quadratic, because the true regression function happens to be of that form. The more powerful, more flexible, higher-degree polynomials were able to get closer to the training data, but that just meant matching the noise better. In terms of the bias-variance decomposition, the bias shrinks with the model degree, but the variance of estimation grows.

Notice that the models of degrees 0 and 1 also do worse than the quadratic model — their problem is not over-fitting but *under*-fitting; they would do better if they were more flexible. Plots of generalization error like this usually have a minimum. If we have a choice of models — if we need to do **model selection** — we would like to find the minimum. Even if we do not have a *choice* of models, we might like to know how big the gap between our in-sample error and our generalization error is likely to be.

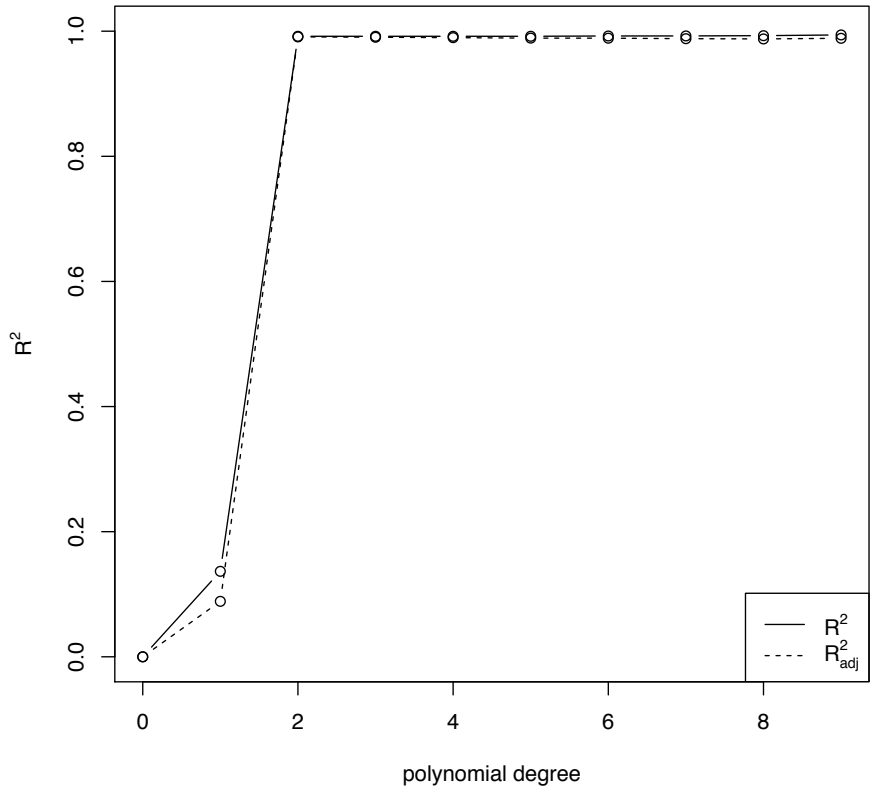
There is nothing special about polynomials here. All of the same lessons apply to variable selection in linear regression, to k-nearest neighbors (where we need to choose k), to kernel regression (where we need to choose the bandwidth), and to other methods we'll see later. In every case, there is going to be a minimum for the generalization error curve, which we'd like to find.

(A minimum with respect to what, though? In Figure 3.5, the horizontal axis is the model degree, which here is the number of parameters [minus one for the intercept]. More generally, however, what we care about is some measure of how complex the model space is, which is not necessarily the same thing as the number of parameters. What's more relevant is how flexible the class of models is, how many different functions it can approximate. Linear polynomials can approximate a smaller set of functions than quadratics can, so the latter are more complex, or have higher **capacity**. More advanced learning theory has a number of ways of quantifying this, but the details get pretty arcane, and we will just use the concept of complexity or capacity informally.)

3.4 Cross-Validation

The most straightforward way to find the generalization error would be to do what I did above, and to use fresh, independent data from the same source — a **testing** or **validation** data-set. Call this $\mathbf{z}'_m$, as opposed to our training data $\mathbf{z}_n$. We fit our model to $\mathbf{z}_n$, and get $\widehat{\theta_n}$. The loss of this on the validation data is

$$\mathbb{E}\left[L(Z, \widehat{\theta_n})\right] + \eta'_m(\widehat{\theta_n}) \tag{3.8}$$




```
extract.rsqd <- function(mdl) {
    c(summary(mdl)$r.squared, summary(mdl)$adj.r.squared)
}
rsqd.q <- sapply(fitted.models, extract.rsqd)
plot(0:9, rsqd.q[1, ], type = "b", xlab = "polynomial degree", ylab = expression(R^2),
    ylim = c(0, 1))
lines(0:9, rsqd.q[2, ], type = "b", lty = "dashed")
legend("bottomright", legend = c(expression(R^2), expression(R[adj]^2)), lty = c("solid",
    "dashed"))
```

FIGURE 3.6: *R^2 and adjusted R^2 for the polynomial fits, to reinforce §2.2.1.1's point that neither is really a useful measure of how well a model fits, or a good way to select among models.*

```
CAPA <- na.omit(read.csv("http://www.stat.cmu.edu/~cshalizi/uADA/13/hw/01/calif_penn_2011.csv"))
half_A <- sample(1:nrow(CAPA), size = nrow(CAPA)/2, replace = FALSE)
half_B <- setdiff(1:nrow(CAPA), half_A)
small_formula = "Median_house_value ~ Median_household_income"
large_formula = "Median_house_value ~ Median_household_income + Median_rooms"
small_formula <- as.formula(small_formula)
large_formula <- as.formula(large_formula)
msmall <- lm(small_formula, data = CAPA, subset = half_A)
mlarge <- lm(large_formula, data = CAPA, subset = half_A)
in.sample.mse <- function(model) {
    mean(residuals(model)^2)
}
new.sample.mse <- function(model, half) {
    test <- CAPA[half, ]
    predictions <- predict(model, newdata = test)
    return(mean((test$Median_house_value - predictions)^2))
}
```

CODE EXAMPLE 1: *Code used to generate the numbers in Figure 3.7. See online for comments.*

where now the sampling noise on the *validation* set, η'_m, is independent of $\widehat{\theta_m}$. So this gives us an unbiased estimate of the generalization error, and, if m is large, a precise one. If we need to select one model from among many, we can pick the one which does best on the validation data, with confidence that we are not just over-fitting.

The problem with this approach is that we absolutely, positively, cannot use any of the validation data in estimating the model. Since collecting data is expensive — it takes time, effort, and usually money, organization, effort and skill — this means getting a validation data set is expensive, and we often won't have that luxury.

3.4.1 Data Splitting

The next logical step, however, is to realize that we don't strictly need a separate validation set. We can just take our data and *split* it ourselves into training and testing sets. If we divide the data into two parts at random, we ensure that they have (as much as possible) the same distribution, and that they are independent of each other. Then we can act just as though we had a real validation set. Fitting to one part of the data, and evaluating on the other, gives us an unbiased estimate of generalization error. Of course it doesn't matter which half of the data is used to train and which half is used to test.

Figure 3.7 illustrates the idea with a bit of the data and linear models from §A.12, and Code Example 1 shows the code used to make Figure 3.7.

3.4.2 k-Fold Cross-Validation (CV)

The problem with data splitting is that, while it's an unbiased estimate of the risk, it is often a very noisy one. If we split the data evenly, then the test set has $n/2$

	Median_house_value	Median_household_income	Median_rooms
2	909600	111667	6.0
3	748700	66094	4.6
4	773600	87306	5.0
5	579200	62386	4.5
11274	209500	56667	6.0
11275	253400	71638	6.6

	Median_house_value	Median_household_income	Median_rooms
3	748700	66094	4.6
4	773600	87306	5.0
5	579200	62386	4.5

	Median_house_value	Median_household_income	Median_rooms
2	909600	111667	6.0
11274	209500	56667	6.0
11275	253400	71638	6.6

	RMSE($A \to A$)	RMSE($A \to B$)
Income only	1.6301137×10^5	1.6003643×10^5
Income + Rooms	1.2771308×10^5	1.2647686×10^5

FIGURE 3.7: *Example of data splitting. The top table shows three columns and seven rows of the housing-price data used in §A.12. I then randomly split this into two equally-sized parts (next two tables). I estimate a linear model which predicts house value from income alone, and another model which predicts from income and the median number of rooms, on the first half. The third table fourth row shows the performance of each estimated model both on the first half of the data (left column) and on the second (right column). The latter is a valid estimate of generalization error. The larger model* always *has a lower in-sample error, whether or not it is really better, so the in-sample MSEs provide little* evidence *that we should use the larger model. Having a lower score under data splitting, however,* is *evidence that the larger model generalizes better. (For R commands used to get these numbers, see Code Example 1.)*

```
cv.lm <- function(data, formulae, nfolds = 5) {
    data <- na.omit(data)
    formulae <- sapply(formulae, as.formula)
    n <- nrow(data)
    fold.labels <- sample(rep(1:nfolds, length.out = n))
    mses <- matrix(NA, nrow = nfolds, ncol = length(formulae))
    colnames <- as.character(formulae)
    for (fold in 1:nfolds) {
        test.rows <- which(fold.labels == fold)
        train <- data[-test.rows, ]
        test <- data[test.rows, ]
        for (form in 1:length(formulae)) {
            current.model <- lm(formula = formulae[[form]], data = train)
            predictions <- predict(current.model, newdata = test)
            test.responses <- eval(formulae[[form]][[2]], envir = test)
            test.errors <- test.responses - predictions
            mses[fold, form] <- mean(test.errors^2)
        }
    }
    return(colMeans(mses))
}
```

CODE EXAMPLE 2: *Function to do k-fold cross-validation on linear models, given as a vector (or list) of model formulae. Note that this only returns the CV MSE, not the parameter estimates on each fold.*

data points — we've cut in half the number of sample points we're averaging over. It would be nice if we could reduce that noise somewhat, especially if we are going to use this for model selection.

One solution to this, which is pretty much the industry standard, is what's called **k-fold cross-validation**. Pick a small integer k, usually 5 or 10, and divide the data at random into k equally-sized subsets. (The subsets are often called "folds".) Take the first subset and make it the test set; fit the models to the rest of the data, and evaluate their predictions on the test set. Now make the second subset the test set and the rest of the training sets. Repeat until each subset has been the test set. At the end, average the performance across test sets. (This is the same as data-set splitting if $k = 2$.) This is the cross-validated estimate of generalization error for each model. Model selection then picks the model with the smallest estimated risk.[7] Code Example 2 performs k-fold cross-validation for linear models specified by formulae.

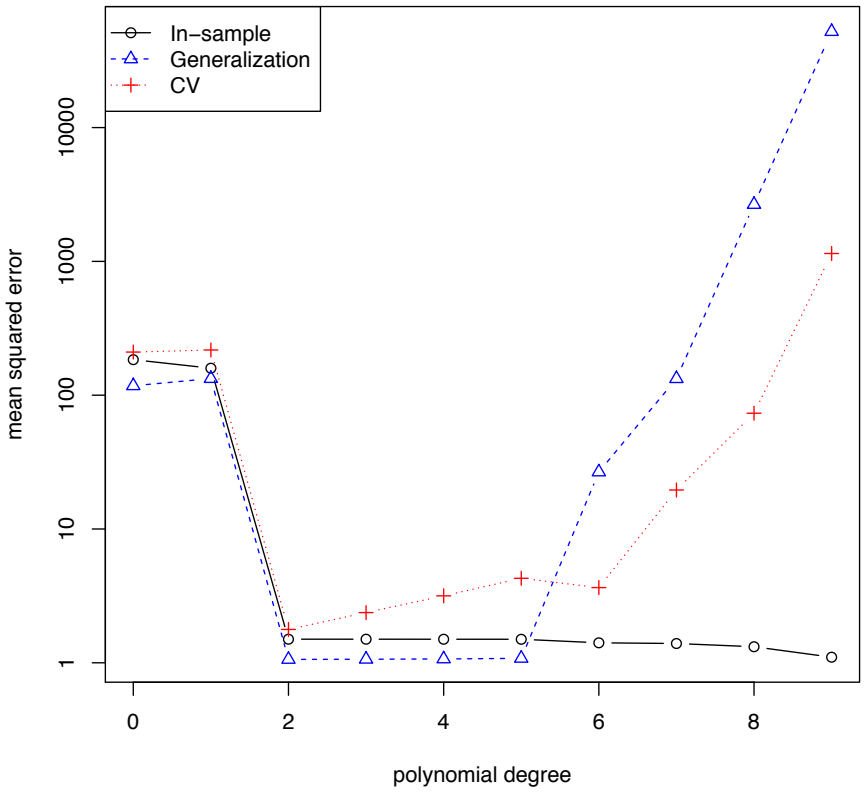
The reason cross-validation works is that it uses the existing data to simulate the process of generalizing to new data. If the full sample is large, then even the smaller portion of it in the testing data is, with high probability, fairly representative of the

[7] A closely related procedure, sometimes also called "k-fold CV", is to pick $1/k$ of the data points at random to be the test set (using the rest as a training set), and then pick an *independent* $1/k$ of the data points as the test set, etc., repeating k times and averaging. The differences are subtle, but what's described in the main text makes sure that each point is used in the test set just once.

data-generating process. *Randomly* dividing the data into training and test sets makes it very unlikely that the division is rigged to favor any one model class, over and above what it would do on real new data. Of course the original data set is never *perfectly* representative of the full data, and a smaller testing set is even less representative, so this isn't ideal, but the approximation is often quite good. It is especially good at getting the *relative* order of different models right, that is, at controlling over-fitting.[8] Figure 3.8 demonstrates these points for the polynomial fits we considered earlier (in Figures 3.3–3.5).

Cross-validation is probably the most widely-used method for model selection, and for picking control settings, in modern statistics. There are circumstances where it can fail — especially if you give it *too many* models to pick among — but it's the first thought of seasoned practitioners, and it should be your first thought, too. The assignments to come will make you *very* familiar with it.

[8]The cross-validation score for the selected model still tends to be somewhat over-optimistic, because it's still picking the luckiest model — though the influence of luck is much attenuated. Tibshirani and Tibshirani (2009) provides a simple correction.




```
little.df <- data.frame(x = x, y = y)
cv.q <- cv.lm(little.df, poly.formulae)
plot(0:9, mse.q, type = "b", xlab = "polynomial degree", ylab = "mean squared error",
    log = "y", ylim = c(min(mse.q), max(gmse.q)))
lines(0:9, gmse.q, lty = 2, col = "blue", type = "b", pch = 2)
lines(0:9, cv.q, lty = 3, col = "red", type = "b", pch = 3)
legend("topleft", legend = c("In-sample", "Generalization", "CV"), col = c("black",
    "blue", "red"), lty = 1:3, pch = 1:3)
```

FIGURE 3.8: *In-sample, generalization, and cross-validated MSE for the polynomial fits of Figures 3.3, 3.4 and 3.5. Note that the cross-validation is done entirely within the initial set of only 20 data points.*

3.4.3 Leave-one-out Cross-Validation

Suppose we did k-fold cross-validation, but with $k = n$. Our testing sets would then consist of single points, and each point would be used in testing once. This is called **leave-one-out cross-validation**. It actually came before k-fold cross-validation, and has two advantages. First, it doesn't require any random number generation, or keeping track of which data point is in which subset. Second, and more importantly, because we are only testing on *one* data point, it's often possible to find what the prediction on the left-out point would be by doing calculations on a model fit to the *whole* data. (See below.) This means that we only have to fit each model once, rather than k times, which can be a big savings of computing time.

The drawback to leave-one-out CV is subtle but often decisive. Since each training set has $n-1$ points, any two training sets must share $n-2$ points. The models fit to those training sets tend to be strongly correlated with each other. Even though we are averaging n out-of-sample forecasts, those are correlated forecasts, so we are not really averaging away all that much noise. With k-fold CV, on the other hand, the fraction of data shared between any two training sets is just $\frac{k-2}{k-1}$, not $\frac{n-2}{n-1}$, so even though the number of terms being averaged is smaller, they are less correlated.

There are situations where this issue doesn't really matter, or where it's overwhelmed by leave-one-out's advantages in speed and simplicity, so there is certainly still a place for it, but one subordinate to k-fold CV.[9]

[[ATTN: AIC appendix?]]

A Short-cut for Linear Smoothers Suppose the model m is a linear smoother (§1.5). For each of the data points i, then, the predicted value is a linear combination of the observed values of y, $m(x_i) = \sum_j \hat{w}(x_i, x_j) y_j$ (Eq. 1.48). As in §1.5.3, define the "influence", "smoothing" or "hat" matrix $\hat{\mathbf{w}}$ by $\hat{w}_{ij} = \hat{w}(x_i, x_j)$. What happens when we hold back data point i, and then make a prediction at x_i? Well, the observed response at i can't contribute to the prediction, but otherwise the linear smoother should work as before, so

$$m^{(-i)}(x_i) = \frac{(\hat{\mathbf{w}}\mathbf{y})_i - \hat{w}_{ii} y_i}{1 - \hat{w}_{ii}}$$

The numerator just removes the contribution to $m(x_i)$ that came from y_i, and the denominator just re-normalizes the weights in the smoother. Now a little algebra says that

$$y_i - m^{(-i)}(x_i) = \frac{y_i - m(x_i)}{1 - \hat{w}_{ii}}$$

The quantity on the left of that equation is what we want to square and average to get the leave-one-out CV score, but everything on the right can be calculated from

[9] At this point, it may be appropriate to say a few words about the Akaike information criterion, or AIC. AIC also tries to estimate how well a model will generalize to new data. One can show that, under standard assumptions, as the sample size gets large, leave-one-out CV actually gives the same estimate as AIC (Claeskens and Hjort, 2008, §2.9). However, there do not seem to be any situations where AIC works where leave-one-out CV does not work at least as well. So AIC should really be understood as a very fast, but often very crude, approximation to the more accurate cross-validation.

the fit we did to the whole data. The leave-one-out CV score is therefore

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - m(x_i)}{1-\hat{w}_{ii}}\right)^2 \tag{3.9}$$

Thus, if we restrict ourselves to leave-one-out and to linear smoothers, we can calculate the CV score with just one estimation on the whole data, rather than n re-estimates.

An even faster approximation that this is what's called "generalized" cross-validation, which is just the in-sample MSE divided by $(1-n^{-1}\operatorname{tr}\hat{\mathbf{w}})^2$. That is, rather than dividing each term in Eq. 3.9 by a unique factor that depends on its own diagonal entry in the hat matrix, we use the average of all the diagonal entries, $n^{-1}\operatorname{tr}\hat{\mathbf{w}}$. (Recall from §1.5.3.2 that $\operatorname{tr}\hat{\mathbf{w}}$ is the number of effective degrees of freedom for a linear smoother.) In addition to speed, this tends to reduce the influence of points with high values of $\hat{w}_{ii}$, which may or may not be desirable.

3.5 Warnings

Some caveats are in order.

1. All of these model selection methods aim at getting models which will generalize well to new data, *if it follows the same distribution* as old data. Generalizing well even when distributions change is a much harder and much less well-understood problem (Quiñonero-Candela *et al.*, 2009). It is particularly troublesome for a lot of applications involving large numbers of human beings, because society keeps changing all the time — variables vary by definition, but the *relationships* between variables also change. (That's history.)

   [[ATTN: Only mentioned CV in this version!]]

2. All of the standard theory of statistical inference you have learned so far presumes that you have a model which was fixed in advance of seeing the data. If you use the data to select the model, that theory becomes invalid, and it will no longer give you correct p-values for hypothesis tests, confidence sets for parameters, etc., etc. Typically, using the same data both to select a model and to do inference leads to too much confidence that the model is correct, significant, and estimated precisely.

3. All the model selection methods we have discussed aim at getting models which *predict well*. This is not necessarily the same as getting the *true theory of the world*. Presumably the true theory will also predict well, but the converse does not necessarily follow. We will see examples later where false but low-capacity models, because they have such low variance of estimation, actually out-predict correctly specified models.

   [[TODO: cross-refs]]

The last two items — combining selection with inference, and parameter interpretation — deserve elaboration.

3.5.1 Inference after Selection

You have, by this point, learned a lot of inferential statistics — how to test various hypotheses, calculate p-values, find confidence regions, etc. Most likely, you have been taught procedures or calculations which all presume that the model you are working with is fixed in advance of seeing the data. But, of course, if you do model selection, the model you do inference within is *not* fixed in advance, but is actually a function of the data. What happens then?

This depends on whether you do inference with the same data used to select the model, or with another, independent data set. If it's the same data, then all of the inferential statistics become invalid — none of the calculations of probabilities on which they rest are right any more. Typically, if you select a model so that it fits the data well, what happens is that confidence regions become too small[10], as do p-values for testing hypotheses about parameters. Nothing can be trusted as it stands.

The essential difficulty is this: Your data are random variables. Since you're doing model selection, making your model a function of the data, that means your model is random too. That means there is some extra randomness in your estimated parameters (and everything else), which isn't accounted for by formulas which assume a fixed model (Exercise 4). This is not just a problem with formal model-selection devices like cross-validation. If you do an initial, exploratory data analysis before deciding which model to use — and that's generally a good idea — you are, yourself, acting as a noisy, complicated model-selection device.

There are three main ways of dealing with this issue of **post-selection inference**.

1. *Ignore it.* This can actually make sense if you don't really care about doing inference within your selected model, you just care about what model is selected. Otherwise, I can't recommend it.

2. *Beat it with more statistical theory.* There is, currently, a lot of interest among statisticians in working out exactly what happens to sampling distributions under various combinations of models, model-selection methods, and assumptions about the true, data-generating process. Since this is an active area of research in statistical theory, I will pass it by, with some references in §3.6[11].

3. *Evade it with an independent data set.* Remember that if the events A and B are probabilistically independent, then $\Pr(A|B) = \Pr(A)$. Now set $A =$ "the confidence set we calculated from this new data covers the truth" and $B =$ "the model selected from this old data was such-and-such". So long as the old and the new data are independent, it doesn't matter that the model was selected using data, rather than being fixed in advance.

The last approach is of course our old friend data splitting (§3.4.1). We divide the data into two parts, and we use one of them to select the model. We then re-estimate the selected model on the other part of the data, and only use that second part in calculating our inferential statistics. Experimentally, using part of the data to

[10] Or, if you prefer, the same confidence region really has a lower confidence level, a lower probability of containing or covering the truth, than you think it does.

[11] But I am sure that Prof. G'Sell or Prof. Tibshirani would be happy to tell you all about it.

do selection, and then all of the data to do inference, does not work as well as a strict split (Faraway, 2016). Using equal amounts of data for selection and for inference is somewhat arbitrary, but, again it's not clear that there's a much better division.

Of course, if you only use a portion of your data to calculate confidence regions, they will typically be larger than if you used all of the data. (Or, if you're running hypothesis tests, fewer coefficients will be significantly different from zero, etc.) This drawback is more apparent than real, since using all of your data to select a model and do inference gives you apparently-precise confidence regions which aren't actually valid.

The simple data-splitting approach to combining model selection and inference only works if the individual data points were independent to begin with. When we deal with dependent data, in Part III, other approaches will be necessary.

3.5.2 Parameter Interpretation

In many situations, it is very natural to want to attach some substantive, real-world meaning to the parameters of our statistical model, or at least to some of them. I have mentioned examples above like astronomy, and it is easy to come up with many others from the natural sciences. This is also extremely common in the social sciences. It is fair to say that this is much less carefully attended to than it should be.

To take just one example, consider the paper "Luther and Suleyman" by Prof. Murat Iyigun (Iyigun, 2008). The major idea of the paper is to try to help explain why the Protestant Reformation was not wiped out during the European wars of religion (or alternately, why the Protestants did not crush all the Catholic powers), leading western Europe to have a mixture of religions, with profound consequences. Iyigun's contention is that all of the Christians were so busy fighting the Ottoman Turks, or perhaps so afraid of what might happen if they did not, that conflicts among the European Christians were suppressed. To quote his abstract:

> at the turn of the sixteenth century, Ottoman conquests lowered the number of all newly initiated conflicts among the Europeans roughly by 25 percent, while they dampened all longer-running feuds by more than 15 percent. The Ottomans' military activities influenced the length of intra-European feuds too, with each Ottoman-European military engagement shortening the duration of intra-European conflicts by more than 50 percent.

To back this up, and provide those quantitative figures, Prof. Iyigun estimates linear regression models, of the form[12]

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 Z_t + \beta_3 U_t + \epsilon_t \tag{3.10}$$

where Y_t is "the number of violent conflicts initiated among or within continental European countries at time t"[13], X_t is "the number of conflicts in which the Ottoman Empire confronted European powers at time t", Z_t is "the count at time t of the

[12] His Eq. 1 on pp. 1473; I have modified the notation to match mine.

[13] In one part of the paper; he uses other dependent variables elsewhere.

newly initiated number of Ottoman conflicts with others and its own domestic civil discords", U_t is control variables reflecting things like the availability of harvests to feed armies, and ϵ_t is Gaussian noise.

The qualitative idea here, about the influence of the Ottoman Empire on the European wars of religion, has been suggested by quite a few historians before[14]. The point of this paper is to support this rigorously, and make it precise. That support and precision requires Eq. 3.10 to be an accurate depiction of at least part of the process which led European powers to fight wars of religion. Prof. Iyigun, after all, wants to be able to interpret a negative estimate of β_1 as saying that fighting off the Ottomans *kept* Christians from fighting each other. If Eq. 3.10 is inaccurate, if the model is badly mis-specified, however, β_1 becomes the best approximation to the truth within a systematically wrong model, and the support for claims like "Ottoman conquests lowered the number of all newly initiated conflicts among the Europeans roughly by 25 percent" drains away.

To back up the use of Eq. 3.10, Prof. Iyigun looks at a range of slightly different linear-model specifications (e.g., regress the number of intra-Christian conflicts in year t on the number of Ottoman attacks in year $t-1$), and slightly different methods of estimating the parameters. What he does *not* do is look at the other implications of the model: that residuals should be (at least approximately) Gaussian, that they should be unpredictable from the regressor variables. He does not look at whether the relationships he thinks are linear really are linear (see Chapters 4, 9, and 10). He does not try to simulate his model and look at whether the patterns of European wars it produces resemble actual history (see Chapter 5). He does not try to check whether he has a model which really supports causal inference, though he has a causal question (see Part IV).

I do not say any of this to denigrate Prof. Iyigun. His paper is actually *much better* than most quantitative work in the social sciences. This is reflected by the fact that it was published in the *Quarterly Journal of Economics*, one of the most prestigious, and rigorously-reviewed, journals in the field. The point is that by the end of this course, you will have the tools to do *better*.

3.6 Further Reading

[[TODO: Link to stuff on penalties in splines and optimization appendix]]

[[TODO: Link to stuff on sieves and capacity control, perhaps in optimization appendix]]

Data splitting and cross-validation go back in statistical practice for many decades, though often as a very informal tool. One of the first important papers on the subject was Stone (1974), which goes over the earlier history. Arlot and Celisse (2010) is a good recent review of cross-validation. Faraway (1992, 2016) reviews computational evidence that data splitting reduces the over-confidence that results from model selection even if one only wants to do prediction. Györfi *et al.* (2002, chs. 7–8) has important results on data splitting and cross-validation, though the proofs are rather more advanced than this book.

Some comparatively easy starting points on statistical learning theory are Kearns and Vazirani (1994), Cristianini and Shawe-Taylor (2000) and Mohri *et al.* (2012). At a more advanced level, look at the tutorial papers by Bousquet *et al.* (2004); von

[14] See §1–2 of Iyigun (2008), and MacCulloch (2004, *passim*).

Luxburg and Schölkopf (2008), or the textbooks by Vidyasagar (2003) and by Anthony and Bartlett (1999) (the latter is much more general than its title suggests), or read the book by Vapnik (2000) (one of the founders). Hastie *et al.* (2009), while invaluable, is much more oriented towards models and practical methods than towards learning *theory*.

On model selection in general, the best recent summary is the book by Claeskens and Hjort (2008); it is more theoretically demanding than this book, but includes many real-data examples.

The literature on doing statistical inference after model selection by accounting for selection effects, rather than simple data splitting, is already large and rapidly growing. Taylor and Tibshirani (2015) is a comparatively readable introduction to the "selective inference" approach associated with those authors and their collaborators. Tibshirani *et al.* (2015) draws connections between this approach and the bootstrap (ch. 6). Berk *et al.* (2013) provides yet another approach to post-selection inference; nor is this an exhaustive list.

White (1994) is a thorough treatment of parameter estimation in models which may be mis-specified, and some general tests for mis-specification. It also briefly discusses the interpretation of parameters in mis-specified models. That very important topic deserves a more in-depth treatment, but I don't know of one.

3.7 Exercises

1. Suppose that one of our model classes contains the true and correct model, but we also consider more complicated and flexible model classes. Does the bias-variance trade-off mean that we will over-shoot the true model, and always go for something more flexible, when we have enough data? (This would mean there was such a thing as *too much* data to be reliable.)

2. Derive the formula for the generalization risk in the situation depicted in Figure 3.1, as given by the `true.risk` function in the code for that figure. In particular, explain to yourself where the constants 0.5^2 and $1/12$ come from.

3. *"Optimism" and degrees of freedom* Suppose that we observe data of the form $Y_i = \mu(x_i) + \epsilon_i$, where the noise terms ϵ_i have mean zero, are uncorrelated, and all have variance σ^2. We use a linear smoother (§1.5) to get estimates $\widehat{\mu}$ from n such data points. The "optimism" of the estimate is

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i' - \widehat{\mu}(x_i))^2\right] - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{\mu}(x_i))^2\right] \tag{3.11}$$

   where Y_i' is an independent copy of Y_i. That is, the optimism is the difference between the in-sample MSE, and how well the model would predict on new data taken at exactly the same x_i values.

   (a) Find a formula for the optimism in terms of n, σ^2, and the number of effective degrees of freedom (in the sense of §1.5.3).

(b) When (and why) does $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i' - \widehat{\mu}(x_i))^2\right]$ differ from the risk?

4. *The perils of post-selection inference, and data splitting to the rescue*[15] Generate a 1000×101 array, where all the entries are IID standard Gaussian variables. We'll call the first column the response variable Y, and the others the predictors $X_1, \ldots X_{100}$. By design, there is no true relationship between the response and the predictors (but all the usual linear-Gaussian-modeling assumptions hold).

   (a) Estimate the model $Y = \beta_0 + \beta_1 X_1 + \beta_{50} X_{50} + \epsilon$. Extract the p-value for the F test of the whole model. Repeat the simulation, estimation and testing 100 times, and plot the histogram of the p-values. What does it look like? What should it look like?

   (b) Use the `step` function to select a linear model by forward stepwise selection. Extract the p-value for the F-test of the selected model. Repeat 100 times and plot the histogram of p-values. Explain what's going on.

   (c) Again use `step` to select a model based on one random 1000×101 array. Now re-estimate the selected model on a new 1000×101 array, and extract the new p-value. Repeat 100 times, with new selection and inference sets each time, and plot the histogram of p-values.

[15] Inspired by Freedman (1983).