Multivariate data analysis: overview of key concepts

James G. Scott

Models for grouping variables

Example data sets and scripts: rxntime

Notation. We use the letter y to denote a response variable, and x to denote a predictor. We usually have more than one predictor variable (x_1, x_2, and so forth), but at least in this course, only one response. The subscript i indexes cases or observations, and j will index variables. Thus y_i is the response for the ith case; x_{ij} is the value of the jth predictor for the ith case.

Residuals and fitted values. One important purpose of a statistical model is to partition variation into predictable and unpredictable components. In a simple group-wise model, we write each observation as "individual case = group mean + deviation of that case," or

$$y_i = \hat{y}_i + e_i = \text{Group mean} + \text{Residual}\,.$$

More generally, $\hat{y}_i$ is the predicted or fitted value from the model. The residual is often called the "error," but it need not be an error in the sense of observational noise. More often it is just the sum of all the effects we've chosen to leave out of the model. Residuals should have a mean of zero. If not, we could improve the model by moving the group means up or down.

Dummy variables. We usually express group-wise models in terms of *indicator* or *dummy* variables, rather than the actual means of the groups. Take the simple case of a single grouping variable x with two levels: "on" ($x = 1$) and "off" ($x = 0$). We can write this model in "baseline/offset" form:

$$y_i = \beta_0 + \beta_1 \mathbf{1}_{\{x_i=1\}} + e_i\,.$$

The quantity $\mathbf{1}_{\{x_i=1\}}$ is called a dummy variable; it takes the value 1 when $x_i = 1$, and the value 0 otherwise. We call β_0 and β_1 the *coefficients* of the model. This way of expressing the model implies the following.

$$\begin{aligned} \text{Group mean for case where } x \text{ is off} &= \beta_0 \\ \text{Group mean for case where } x \text{ is on} &= \beta_0 + \beta_1\,. \end{aligned}$$

Therefore, we can think of β_0 as the baseline (or *intercept*), and β_1 as the offset.

We estimate the values of β_0 and β_1 using the least-squares criterion: that is, make the sum of squared errors, $\sum_{i=1}^n e_i^2$, as small as possible. It turns out that this is mathematically equivalent to computing the group-wise means separately. In light of this, you might wonder: why bother with the baseline/offset form? One reason is simple: we are often interested not in the means themselves, but in the *differences* between the means (in this case, the offset β_1).

Variance decomposition and R^2. The variance decomposition of a linear statistical model is

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 ,$$

or

$$TV = PV + UV .$$

Variance of the data = variance of the fitted values + variance of the residuals. PV is the variation of the predictable part of y; UV is the variation of the unpredictable part. This additive decomposition doesn't work for sums of absolute values, only sums of squares. This is not just a metaphor. It turns out to be a important consequence of the Pythagorean theorem in a high-dimensional Euclidean space. It's a big reason we use sums of squares to describe variability in statistical models.

We define R^2 as the ratio of predictable variation to total variation: $R^2 = PV/TV = 1 - UV/TV$. This quantifies the preciseness of the fit, and therefore the information content of the predictor. Some people abbreviate the variance decomposition as TSS = ESS + RSS, but this can be ambiguous. Do the letters mean Total = Explained + Residual? Or Total = Error + Regression?

The individual terms in the variance decomposition are perfectly well defined in nonlinear statistical models. But the three terms will not, in general, add together. In this case people often just quote the mean squared error, or MSE, as:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 ,$$

where n is the sample size.

More than two levels. If the predictor x has more than two levels, we must expand it in terms of more than one dummy variable. Suppose that x can take four levels, labeled arbitrarily as `0` through 3. Then our model is

$$y_i = \beta_0 + \beta_1^{(1)} \mathbf{1}_{\{x_i=1\}} + \beta_1^{(2)} \mathbf{1}_{\{x_i=2\}} + \beta_1^{(3)} \mathbf{1}_{\{x_i=3\}} + e_i .$$

More generally, $\beta_j^{(k)}$ is the coefficient associated with the kth level of the jth variable. Notice that there is no dummy variable for the case $x = 0$: this is the baseline case, whose group mean

is described by the intercept β_0. In general, for a categorical variable with K levels, we will have $K-1$ dummy variables.

More than one grouping variable. Take the case of two grouping variables x_1 and x_2, each of which can take the value 0 ("off") or 1 ("on"). One approach to modeling the effect of x_1 and x_2 is to slice and dice. That is: take subsets of the data for each of the four combinations of x_1 and x_2, and compute the mean within each subset.

This approach is intuitively reasonable, but combinatorially explosive. For example, with 10 grouping variables, there will be $2^{10} = 1024$ possible subsets, and thus 1024 group-wise means to estimate. If you want to do this, you will need a lot of data—not merely overall, but for each combination separately.

A second strategy is to treat the effect of x_1 and x_2 as if they are separable:

$$y_i = \hat{y}_i + e_i = \text{Baseline} + (\text{Effect if } x_1 \text{ on}) + (\text{Effect if } x_2 \text{ on}) + \text{Residual}.$$

This notation gets cumbersome. We can write it more concisely as

$$y_i = \beta_0 + \beta_1 \mathbf{1}_{\{x_1\}} + \beta_2 \mathbf{1}_{\{x_2\}} + e_i.$$

Notice, for example, that if $x_2 = 0$, then the $\beta_2 \mathbf{1}_{\{x_2\}}$ term falls away, and we're left with the baseline, plus the effect of x_1 being on, plus the residual. We refer to β_1 and β_2 as the *main effects*.

Interactions. What if the effects of x_1 and x_2 aren't separable? That is, we believe

$$y_i = \text{Baseline} + (\text{Effect if } x_1 \text{ on}) + (\text{Effect if } x_2 \text{ on}) + (\text{Extra effect if both } x_1 \text{ and } x_2 \text{ on}) + \text{Residual}.$$

We can create such a model by multiplying dummy variables together:

$$y_i = \beta_0 + \beta_1 \mathbf{1}_{\{x_1=1\}} + \beta_2 \mathbf{1}_{\{x_2=1\}} + \beta_{12} \mathbf{1}_{\{x_1\}} \mathbf{1}_{\{x_2\}} + e_i.$$

We call β_{12} an *interaction term*. This one is a two-way interaction. We may also have multi-way interactions involving arbitrary numbers of predictors. The caveat is: the more multi-way interaction terms we add, the closer we come to the pure slice-and-dice approach.

Quantifying uncertainty

Sampling distributions. As we have seen, one purpose of a statistical model is to partition observed variation. Another purpose is to quantify uncertainty about any trends we see in the data. This pre-supposes that we don't know the whole story—in other words, that the data are an imperfect reflection of some underlying reality. There are three common "creation myths" that play a central role in statistical analysis.

1. The data set comprises the entire relevant population.

2. The data are a random sample from a wider population. (Archetypal examples: political polls, surveys, animals in a lab experiment.)

3. The data are one realization of a random process. (Archetypal examples: earthquakes, hurricanes, nucleotide sequences in extant organisms, photons from a distant star.)

In the first case, there is no uncertainty, and thus no need for statistical thinking. But in the other two cases, we reason as follows. Our model parameters are estimated from the data. But in a parallel universe, our data would have been different merely by random chance. Therefore our estimated model parameters might have been different, too.

How different? The answer to this question is the classical notion of an estimator's *sampling distribution*: that is, the distribution of model estimates we would get in all those parallel universes invoked by the relevant creation myth.

The standard deviation of an estimator's sampling distribution is referred to as the *standard error*. In quoting the standard error of an estimator's sampling distribution, you are saying: "If I were to take repeated samples from the population and use this estimator for every sample, my estimate is typically off from the truth by about this much." Notice that this is a claim about a procedure, not a particular estimate. The bigger the standard error, the less stable the estimator across different samples, and the less you can trust that estimator for any particular sample. This is the core idea of frequentist statistics: *uncertainty equals instability across different samples.*

If you really could take repeated samples from the population, life would be easy. You could simply peer into all of those alternate universes, tap each version of yourself on the shoulder, and ask, "What estimate you get for *your* sample?" By tallying up these estimates and seeing how much they differed from one another, you could discover precisely how much confidence you should place in your own estimates of β_0 and β_1, and report appropriate error bars. Let's ignore the obvious fact that, if you had access to all those alternate universes, you'd also have more data. The presence of sample-to-sample variability is the thing to focus on.

In reality, we're stuck with one sample. Thus we're stuck with one of two imperfect approaches for characterizing the sampling distribution: the bootstrap, or the parametric probability model. For many data sets, there is little practical difference between the two approaches, in that they give similar standard errors. Nonetheless, there is a conceptual distinction worth preserving.

Bootstrapping. In most cases we can't repeatedly take samples of size n from the population. But we can repeatedly take samples of size n *from the sample itself*, and compute our estimator afresh for each notional sample. The idea is that the variability of the estimates across all these notional samples can be used to approximate the sampling distribution of the corresponding estimator. Each block of n resampled data points is called a bootstrapped sample. Modern software makes a non-issue of the calculational tedium involved.

You might be puzzled by something here. If there are n data points in the original sample, and we resample n data points from this "pseudo-population," won't each bootstrapped sample be precisely equal to the original sample? It turns out that the answer is no—as long as the resampling is done *with replacement* from the original sample. Sampling with replacement means that each bootstrapped sample will have duplicates and omissions from the original sample. These duplicates and omissions induce variation from one bootstrapped sample to the next. Ths variation mimics the variation you'd expect to see across the real repeated samples you're unable to take.

Resampling won't yield the true sampling distribution of an estimator. Bu it is often good enough for approximating the standard error. The quality of the approximation depends almost entirely on one thing: how closely the original sample resembles the wider population. Alas, this often isn't under your control, and is almost always the limiting factor in the accuracy of the bootstrap. You can't magic your way to sensible error bars by bootstrapping a biased, woefully small, or otherwise poor sample.[1]

Normality assumptions. Another typical approach is to assume that the residuals in your model follow a normal distribution: $\epsilon_i \sim N(0, \sigma^2)$. Implicitly, this assumes that the residuals are independent of one another and have constant variance σ^2 that does not depend upon the predictors in the model. These assumptions can be used to derive (via probability theory) explicit formulas for the standard errors of estimators like the sample mean, the least-squares estimator, and so forth. These assumptions are typically baked in to most statistical software.

Confidence intervals. We use standard errors to construct *confidence intervals*, or error bars. If θ is a parameter, $\hat{\theta}$ is an estimate of that parameter, and $\text{se}(\hat{\theta})$ is the standard error of the estimate, then we can quote a confidence interval of the form

$$\hat{\theta} \pm z_\alpha \cdot \text{se}(\hat{\theta}).$$

The factor z_α is a constant that expresses your tolerance for error, denoted by α and expressed as a number between 0 and 1. A typical confidence level is 0.95, meaning that you'll allow your confidence interval to miss the answer 5% of the time ($\alpha = 0.05$) of the time. It's important to keep in mind that a confidence interval is a claim about the long-run properties of a statistical procedure—in how many parallel universes will the intervals so generated cover the true value? It is not a probabilistic claim about a specific data set. A procedure used to construct confidence intervals satisfies the frequentist coverage property if the confidence intervals so generated cover the true value the stated percentage (e.g. 95%) of the time.

In an introductory statistics course, you will likely have picked z_α by laborious calculations involving the t distribution. None of this is wrong. But we'll skip this tedium and use some simple rules of thumb that statisticians have discovered to reasonably accurate: For a 68% confidence

[1]The approximation also depends on how many bootstrapped samples you take from the original sample. More bootstrapped samples help—up to a point. But taking more bootstrapped samples is never a substitute for having more actual samples in the real data set.

interval, choose $z_\alpha = 1$. For a 95% confidence interval, choose $z_\alpha = 2$. For a 99.5% confidence interval, choose $z_\alpha = 3$.

Two practical guidelines. First, always plot your data. This will often give you a good sense of whether your modeling assumptions are sensible, or whether you're churning through a hapless exercise in "garbage in, garbage out." Second, try never to report a guess without an error bar. The corollary of this second point is: don't be afraid to quote an estimate with weak information! Just make sure the error bars are appropriately wide.

Linear regression

Example data sets and scripts: kidney, gala, ut2000, profs

One predictor. In a simple one-variable regression model, we relate the response y to the predictor x using a linear equation:

$$y_i = \hat{y}_i + e_i = \beta_0 + \beta_1 x_i + e_i \,.$$

As before, we fit the model parameters by least squares. The same variance decomposition (TV = PV = UV) holds here, as does the same definition of R^2.

There are several common goals of regression analysis:

(1) Predicting a future value of y at a given x. For example, we could regress a patient's score on a clinical test for kidney function (y) on his or her age (x). When a new 55-year-old patient walks in the door, we would estimate his score as $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 55$.

(2) Summarizing the trend. The intercept β_0 is the expected value of y when $x = 0$. The slope β_1 describes the expected change in y for every one-unit change in x.

(3) Statistical adjustment, or taking the "x"-ness out of y. The response variable y is systematically associated with x. If we fit a model

$$y_i = \hat{y}_i + e_i = \beta_0 + \beta_1 x_i + e_i \,,$$

then the fitted value $\hat{y}_i = \beta_0 + \beta_1 x_i$ captures this systematic component of variation. Thus the residual e_i can be interpreted as the y variable, having "adjusted for" or "partialled out" x.

(4) Quantifying the reduction in uncertainty from knowing a new piece of information (i.e. a predictor).

More than one predictor. In a multiple-regression model, we just add up the linear effects of each predictor individually,

$$y_i = \hat{y}_i + e_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i \,,$$

once again choosing the coefficients $\hat{\beta}_j$ by the principle of least squares. The interpretation of each coefficient is now slightly more involved than in the one-variable case. Each is an estimated *partial slope*: that is, the change in y associated with a one-unit change in x_j, holding all other variables constant. One way to see this is to isolate a single factor on the right-hand side, say the first:

$$\hat{y}_i - (\beta_2 x_{i2} + \cdots + \beta_p x_{ip}) = \beta_0 + \beta_1 x_{i1}\,.$$

On the right hand side, we see a usual one-variable regression model. On the left, we see the predicted value of y_i, adjusted for all the other factors $x_2, \ldots, x_p$.

Here's another way to see this: imagine two hypothetical people who are identical in all predictors x_j except the first: case i has $x_1 = x^\star$, and case j has $x_1 = x^\star + 1$. Then

$$\hat{y}_j - \hat{y}_i = \{\beta_0 + \beta_1(x^\star + 1) + \beta_2 x_2 + \cdots + \beta_p x_p\} - \{\beta_0 + \beta_1 x^\star + \beta_2 x_2 + \cdots + \beta_p x_p\}\,.$$

Because x_2 through x_p are held constant, all terms but those involving β_1 cancel. We are left with

$$\hat{y}_j - \hat{y}_i = \beta_1(x^\star + 1 - x^\star) = \beta_1\,.$$

In other words, the difference between the two predicted values is precisely β_1.

Continuous and grouping variables together. You will often encounter situations with both continuous and categorical predictors. To handle this we simply incorporate the dummy variables associated with the categorical predictor directly into the multiple-regression equation. These dummy variables systematically shift the intercept up or down, depending on their sign. To see this, consider a case with one continuous predictor x_1, and one grouping variable x_2 that takes three levels, arbitrarily labeled 0–2. The regression equation is then

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2^{(1)} \mathbf{1}_{\{x_{i2}=1\}} + \beta_2^{(2)} \mathbf{1}_{\{x_{i2}=2\}} + e_i\,.$$

As before, there is no dummy variable for the reference category, $x_2 = 0$, as that case is handled by the intercept β_0. We can interpret this as three separate regression equations, with three different intercepts and a common slope:

$$\begin{aligned}
\text{Model when } x_2 = 0\text{:} \quad y_i &= \beta_0 + \beta_1 x_{i1} + e_i \\
\text{Model when } x_2 = 1\text{:} \quad y_i &= \{\beta_0 + \beta_2^{(1)}\} + \beta_1 x_{i1} + e_i \\
\text{Model when } x_2 = 2\text{:} \quad y_i &= \{\beta_0 + \beta_2^{(2)}\} + \beta_1 x_{i1} + e_i\,.
\end{aligned}$$

It is quite common ask questions of the form: "How much larger or smaller are the $x_2 = 2$ cases than the $x_2 = 0$ cases, adjusting for x_1?" The estimate and error bar for $\beta_2^{(2)}$ provide the answer.

We can also have interactions between dummy variables and continuous predictors. The no-

tation for this is straightforward, if a bit cumbersome:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2^{(1)} \mathbf{1}_{\{x_{i2}=1\}} + \beta_2^{(2)} \mathbf{1}_{\{x_{i2}=2\}} + \gamma_1^{(1)} x_{i1} \mathbf{1}_{\{x_{i2}=1\}} + \gamma_1^{(2)} x_{i1} \mathbf{1}_{\{x_{i2}=2\}} + e_i \,.$$

We may again interpret this as three separate regression equations, each with a distinct slope and intercept:

$$\begin{aligned}
\text{Model when } x_2 = 0\text{:} \quad y_i &= \beta_0 + \beta_1 x_{i1} + e_i \\
\text{Model when } x_2 = 1\text{:} \quad y_i &= \{\beta_0 + \beta_2^{(1)}\} + \{\beta_1 + \gamma_1^{(1)}\} x_{i1} + e_i \\
\text{Model when } x_2 = 2\text{:} \quad y_i &= \{\beta_0 + \beta_2^{(2)}\} + \{\beta_1 + \gamma_1^{(2)}\} x_{i1} + e_i \,.
\end{aligned}$$

We may be interested in a question of the form: "How much faster or slower does y grow with x among the cases where $x_2 = 2$ than the cases where $x_2 = 0$?" The estimate and error bar for $\gamma_1^{(2)}$ provide the answer.

Testing a model. Often you'll face the problem of comparing models with different numbers of predictors. The model with more variables will always fit the data better, but it might not be a better model, because it might end up overfitting noise in the data. For these kinds of cross-dimensional comparisons, R^2 is useless. In fact, in many ways it is worse than useless: R^2 will always go up when we add new predictors, even if those predictors have nothing to do with the response. So if you chase the best R^2, you will overfit.

To test whether a variable should be included in a the model, we often use a hypothesis test. This involves four steps:

1. Choose a null hypothesis H_0, the hypothesis of "no effect."
2. Choose a test statistic T that is sensitive to departures from the null hypothesis.
3. Simulate or calculate $P(T \mid H_0)$, the sampling distribution of the test statistic T under the assumption that H_0 is true.
4. Check whether the observed test statistic for your data, t, is consistent with $P(T \mid H_0)$.

Step 3 is often carried out via a permutation test, in which the values of the variable in question are shuffled randomly among the cases, and the model re-fit to each "shuffled" data set. To quantify surprise in step 4, we often quote a p value: the probability under $P(T \mid H_0)$ of having seen a test statistic at least as surprising as t.