Multiple linear regression: Inference, Part I

Patrick Breheny

January 27

Introduction

- In our last lecture, we discussed how to estimate the regression coefficients
- Our goal today is to start addressing the question: how accurate are those estimates?
- In particular, we will be deriving the expectation and variance of our estimates, and some related concepts

Our assumptions for today

- The results we will derive today are based on the following central assumption: Suppose that

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\mathbf{X}$ is a fixed $n \times p$ matrix of full column rank and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of random errors $\{\epsilon_i\}$ which are identically and independently distributed with mean 0 and variance σ^2

- In other words,

$$\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$$
$$\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$$

- For the rest of this lecture, I will refer to the above set of assumptions by saying something along the lines of "Suppose that (1) holds"

Expectation and variance of linear and quadratic forms

- For today's derivations, we will need to calculate the expectation and variance of linear and quadratic forms
- Letting $\mathbf{A}$ denote a matrix of constants and $\mathbf{x}$ a random vector with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$,

$$\begin{aligned}
\mathrm{E}(\mathbf{A}^T\mathbf{x}) &= \mathbf{A}^T\boldsymbol{\mu} \\
\mathrm{Var}(\mathbf{A}^T\mathbf{x}) &= \mathbf{A}^T\boldsymbol{\Sigma}\mathbf{A} \\
\mathrm{E}(\mathbf{x}^T\mathbf{A}\mathbf{x}) &= \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} + \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma})
\end{aligned}$$

The trace

- The operator tr (defined for any square matrix) refers to the *trace* of a matrix, defined as the sum of its diagonal elements:

$$\mathrm{tr}(\mathbf{A}) = \sum_i A_{ii}$$

- Some basic facts about traces:

$$\begin{aligned} \mathrm{tr}(\mathbf{AB}) &= \mathrm{tr}(\mathbf{BA}) \\ \mathrm{tr}(\mathbf{A} + \mathbf{B}) &= \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B}) \\ \mathrm{tr}(c\mathbf{A}) &= c\,\mathrm{tr}(\mathbf{A}) \end{aligned}$$

- A further fact about traces that is not at all obvious but nonetheless useful is that if a matrix $\mathbf{A}$ is idempotent, then $\mathrm{tr}(\mathbf{A}) = \mathrm{rank}(\mathbf{A})$

$\hat{\beta}$ is unbiased

- With these facts in mind, we are ready to prove that
- **Theorem:** Suppose that (1) holds. Then

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

  i.e., estimating the regression coefficients by minimizing the residual sum of squares produces an unbiased estimator
- An important caveat here is this holds only **if the model is correct**; if the model is not correct (for example, it does not adjust for an important confounder), then estimates can be badly biased

The variance of $\hat{\boldsymbol{\beta}}$

- The other important component in assessing an estimator's accuracy is its variance
- **Theorem:** Suppose that (1) holds.

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

- Note that the result is a symmetric $p \times p$ matrix with $\text{Var}(\hat{\beta}_j)$ on the diagonals and $\text{Cov}(\hat{\beta}_j, \hat{\beta}_k)$ in the off-diagonal elements

Helpful facts

- Observe, however, that we can't actually calculate this variance (yet), because we don't know σ^2
- Before we go about deriving an unbiased estimator for σ^2, let's prove the following simple results which will help simplify our calculations:

$$\begin{aligned}
\mathbf{r} &= (\mathbf{I} - \mathbf{H})\mathbf{y} \\
&\mathbf{H} \text{ and } \mathbf{I} - \mathbf{H} \text{ are symmetric} \\
&\mathbf{H} \text{ and } \mathbf{I} - \mathbf{H} \text{ are idempotent} \\
\mathbf{H}\mathbf{X} &= \mathbf{X} \text{ and } \mathbf{X}^T\mathbf{H} = \mathbf{X}^T \\
\mathbf{X}^T\mathbf{r} &= \mathbf{0}
\end{aligned}$$

- I will also state the following without proof: $\mathrm{rank}(\mathbf{X}) = \mathrm{rank}(\mathbf{X}^T\mathbf{X}) = \mathrm{rank}(\mathbf{H})$; *i.e.*, if $\mathbf{X}$ is full rank, all of those matrices have rank p

Trying to estimate σ^2

- In principle, we could estimate σ^2 by

$$\frac{1}{n}\sum \epsilon_i^2$$

  but of course the $\{\epsilon_i\}$ are not observable

- We could use

$$\frac{1}{n}\sum r_i^2,$$

  but since our model was specifically chosen so as to reduce the residual sum of squares, this turns out to underestimate σ^2

$\hat{\sigma}^2$

- Consider instead the following estimator:

$$\hat{\sigma}^2 = \frac{RSS}{n - p}$$

- **Theorem:** Suppose that (1) holds. Then

$$\mathrm{E}(\hat{\sigma}^2) = \sigma^2$$

- Note that this estimator reduces to the usual unbiased estimators of variance and pooled variance in the one-sample and pooled two-sample cases, with $n - p$ as the *degrees of freedom*

Estimating the variance and standard error of $\hat{\beta}$

- A reasonable estimator for the variance of $\hat{\boldsymbol{\beta}}$ is therefore

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$$

- Furthermore, we can obtain standard errors by taking the square root of the diagonal elements

Decomposition of variance

- One final interesting result for today is that one can decompose the sample variance of y into two parts:

$$\widehat{\mathrm{Var}}(y) = \widehat{\mathrm{Var}}(\hat{\mu}) + \widehat{\mathrm{Var}}(r)$$

where $\widehat{\mathrm{Var}}$ means the usual sample variance ($\{y_i\}$, $\{\hat{\mu}_i\}$, and $\{r_i\}$ are all observable)

- Or equivalently,

$$TSS = MSS + RSS$$

where

  - TSS = Total sum of squares, $\sum(y_i - \bar{y})^2$
  - MSS = Model sum of squares, $\sum(\hat{\mu}_i - \bar{\mu})^2$
  - RSS = Residual sum of squares, $\sum r_i^2$

The coefficient of determination

- A useful way of summarizing how good our explanatory variables are at explaining y, then, is to look at the proportional reduction in variability that our model achieves
- This quantity is referred to as the *coefficient of determination* and is denoted R^2:

$$\begin{aligned} R^2 &= \frac{MSS}{TSS} \\ &= 1 - \frac{RSS}{TSS} \end{aligned}$$

- Remark: In the case of simple linear regression, R^2 is the square of r, the correlation coefficient

What if $\mathbf{X}$ is random?

- We've treating $\mathbf{X}$ as fixed for mathematical convenience
- When $\mathbf{X}$ is random (as it would be in an observational study), what changes (besides the fact that you'd have to add "given $\mathbf{X}$" to all the expectations and variances)
- It turns out that all of the results still hold, **if** each of the random variables that make up $\mathbf{X}$ are independent of the random error ϵ
- So once again, a confounder will cause problems, as it will introduce correlation between the explanatory variables and the error, and this could cause all manner of biases
- Remark: The random variables that make up $\mathbf{X}$ do not have to be independent of each other, just independent of the random error

What we don't need

- So we must keep in mind the major, crucial assumption we've made today: that the model we fit is actually true and that $\mathbf{X}$, if it is random, must be uncorrelated with the random error
- However, it's also worth pointing out a big assumption that we didn't make: we did not assume a distribution for Y or ϵ