[Lecture 17]

0.1 Robust Inference via Partial Specification

In the previous section we saw how using a non-parametric inference method–the bootstrap–allowed us to avoid the pitfalls of mis-specified parametric models. Next we will explore a different idea, called *partial specification* or *robust standard errors*. Here we stay within a parametric model, but we derive algebraic formulas that hold even when the particular parametric model is wrong, as long as certain "orthogonality assumptions" are true.

Specifically, we will consider linear regression, deriving standard error estimates via typical parametric confidence regions as with GLMs. We will see that these are brittle, but that they are primarily brittle because they implicitly assume that certain equations hold. If we instead explicitly subtitute the right-hand side of those equations, we get better confidence intervals that hold under fewer assumptions. As a bonus, we'll be able to study how linear regression performs under distribution shift.

Starting point: linear response with Gaussian errors. In the simplest setting, suppose that we completely believe our model:

$$Y = \langle \beta, X \rangle + Z, \text{ where } Z \sim \mathcal{N}(0, \sigma^2 I). \tag{1}$$

We observe samples $(x_1, y_1), \dots, (x_n, y_n) \sim p$. Suppose that we estimate β using the ordinary least squares estimator:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle \beta, x_i \rangle)^2 = (\sum_{i=1}^{n} x_i x_i^\top)^{-1} \sum_{i=1}^{n} x_i y_i. \tag{2}$$

Define $S = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top$. Then since $y_i = x_i^\top \beta + z_i$, we can further write

$$\begin{aligned}
\hat{\beta} &= (\sum_{i=1}^{n} x_i x_i^\top)^{-1} (\sum_{i=1}^{n} x_i x_i^\top \beta + x_i z_i) && (3)\\
&= (nS)^{-1} (nS\beta + \sum_{i=1}^{n} x_i z_i) && (4)\\
&= \beta + \frac{1}{n} S^{-1} \sum_{i=1}^{n} x_i z_i. && (5)
\end{aligned}$$

From this we see that, conditional on the x_i, $\hat{\beta} - \beta$ is a zero-mean Gaussian distribution. Its covariance matrix is given by

$$\frac{1}{n^2} S^{-1} \sum_{i=1}^{n} \mathbb{E}[z_i^2 \mid x_i] x_i x_i^\top S^{-1} = \frac{\sigma^2}{n} S^{-1}. \tag{6}$$

Confidence regions. The above calculation shows that the error $\hat{\beta} - \beta$ is *exactly* Gaussian with covariance matrix $\frac{\sigma^2}{n} S^{-1}$ (at least assuming the errors z_i are i.i.d. Gaussian). Thus the (parametric) confidence region for $\hat{\beta} - \beta$ would be an ellipsoid with shape S^{-1} and radius depending on σ, n, and the significance level α of the test. As a specific consequence, the standard error for β_i is $\sigma \sqrt{(S^{-1})_{ii}/n}$. This is the standard error estimate returned by default in most software packages.

Of course, this all so far rests on the assumption of Gaussian error. Can we do better?

Calculation from moment assumptions. It turns out that our calculation above relied only on conditional moments of the errors, rather than Gaussianity. We will show this explicitly by doing the calculations more carefully. Re-using steps above, we have that

$$\hat{\beta} - \beta = \frac{1}{n} S^{-1} \sum_{i=1}^{n} x_i z_i. \tag{7}$$

In particular, assuming that the (x_i, y_i) are i.i.d., we have

$$\mathbb{E}[\hat{\beta} - \beta \mid x_1, \ldots, x_n] = \frac{1}{n} S^{-1} \sum_{i=1}^{n} x_i \mathbb{E}[z_i \mid x_i] = S^{-1} b, \tag{8}$$

where $b \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} x_i \mathbb{E}[z_i \mid x_i]$.

In particular, as long as $\mathbb{E}[Z \mid X] = 0$ for all X, $\hat{\beta}$ is an unbiased estimator for X. In fact, since this only needs to hold on average, as long as $\mathbb{E}[ZX] = 0$ (covariates uncorrelated with noise) then $\mathbb{E}[\hat{\beta} - \beta] = 0$, and $\mathbb{E}[\hat{\beta} - \beta \mid x_{1:n}]$ converges to zero as $n \to \infty$. This yields an insight that is important more generally:

> *Orindary least squares yields an unbiased estimate of β whenever the covariates X and noise Z are uncorrelated.*

This partly explains the success of OLS compared to other alternatives (e.g. penalizing the absolute error or fourth power of the error). While OLS might initially look like the maximum likelihood estimator under Gaussian errors, it yields consistent estimates of β under much weaker assumptions. Minimizing the fourth power of the error requires stronger assumptions for consistency, while minimizing the absolute error would yield a different condition in terms of medians rather than expectations.

Next we turn to the covariance of $\hat{\beta}$. Assuming again that the (x_i, y_i) are independent across i, we have

$$\begin{aligned}
\mathsf{Cov}[\hat{\beta} \mid x_{1:n}] &= \mathsf{Cov}[\frac{1}{n} S^{-1} \sum_{i=1}^{n} x_i z_i \mid x_{1:n}] && (9)\\
&= \frac{1}{n^2} S^{-1} \sum_{i,j=1}^{n} x_i \mathsf{Cov}[z_i, z_j \mid x_i, x_j] x_j^\top S^{-1} && (10)\\
&= \frac{1}{n^2} S^{-1} \sum_{i=1}^{n} x_i \mathsf{Var}[z_i \mid x_i] x_i^\top S^{-1}, && (11)
\end{aligned}$$

where the final line is because z_i, z_j are independent for $i \neq j$. If we define $\Omega = \frac{1}{n} \sum_{i=1}^{n} x_i \mathsf{Var}[z_i \mid x_i] x_i^\top$, then the final term becomes $\frac{1}{n} S^{-1} \Omega S^{-1}$.

This quantity can be upper-bounded under much weaker assumptions than Gaussianity. If we, for instance, merely assume that $\mathsf{Var}[z_i \mid x_i] \leq \sigma^2$ for all i, then we have that $\Omega \preceq \sigma^2 S$ and hence $\mathsf{Cov}[\hat{\beta} \mid x_{1:n}] \preceq \frac{\sigma^2}{n} S^{-1}$.

Even better, this quantity can be estimated from data. Let $u_i^2 = (y_i - \hat{\beta}^\top x_i)^2$. This is a downward-biased, but asymptotically unbiased, estimate for $\mathsf{Var}[z_i \mid x_i]$ (it would be unbiased if we used β instead of $\hat{\beta}$). Therefore, form the matrix

$$\hat{\Omega}_n = \frac{1}{n} \sum_{i=1}^{n} x_i u_i^2 x_i^\top. \tag{12}$$

Then $\frac{1}{n} S^{-1} \hat{\Omega}_n S^{-1}$ can be used to generate confidence regions and standard errors. In particular, the standard error estimate for β_i is $\sqrt{(S^{-1} \hat{\Omega}_n S^{-1})_{ii} / n}$. This is called the *robust standard error* or *heteroskedacity-consistent standard error.*

There are a couple of simple improvements on this estimate. The first is a "degrees of freedom" correction: we know that u_i^2 is downward-biased, and it is more downward-biased the larger the dimension d (because then $\hat{\beta}$ can more easily overfit). We often instead use $\frac{1}{n-d} S^{-1} \hat{\Omega}_n S^{-1}$, which corrects for this.

A fancier correction, based on the jacknnife, first corrects the errors u_i, via

$$u_i' = u_i / (1 - \kappa_i), \text{ with } \kappa_i = \frac{1}{n} x_i^\top S^{-1} x_i.$$

We obtain a corresponding $\Omega_n' = \frac{1}{n} \sum_{i=1}^{n} x_i (u_i')^2 x_i^\top$, and the matrix for the standard errors is

$$\frac{1}{n} S^{-1} (\Omega_n' - \zeta \zeta^\top) S^{-1}, \text{ where } \zeta = \frac{1}{n} \sum_{i=1}^{n} x_i u_i'.$$

The main difference is that each u_i gets a different correction factor $\frac{1}{1-\kappa_i}$ (which is however roughly equal to $\frac{n}{n-d}$) and also that we subtract off the mean ζ. There is some evidence that this more complicated estimator works better when the sample size is small, see for instance **?**.

Out-of-distribution error. Now suppose that we wish to estimate the error on test samples $\bar{x}_{1:m}$ drawn from a distribution $\bar{p} \neq p$. Start again with the Gaussian assumption that $y = \beta^\top x + z$.

The expected error on sample $\bar{x}_i$ (over test noise $\bar{z}_i$) is $\sigma^2 + \langle \hat{\beta} - \beta, \bar{x}_i \rangle^2$. If we let $\bar{S} = \frac{1}{m}\sum_{i=1}^m \bar{x}_i \bar{x}_i^\top$, and let $\mathbb{E}_Z$ denote the expectation with respect to the training noise $z_1, \ldots, z_n$, then the overall average expected error (conditional on $x_{1:n}, \bar{x}_{1:m}$) is

$$\sigma^2 + \mathbb{E}_Z[\frac{1}{m}\sum_{i=1}^m (\bar{x}_i^\top(\beta - \hat{\beta}))^2] = \sigma^2 + \langle \frac{1}{m}\sum_{i=1}^m \bar{x}_i \bar{x}_i^\top, \mathbb{E}_Z[(\beta - \hat{\beta})(\beta - \hat{\beta})^\top] \rangle \tag{13}$$

$$= \sigma^2 + \langle \bar{S}, \frac{\sigma^2}{n} S^{-1} \rangle \tag{14}$$

$$+ \sigma^2 \Big(1 + \frac{1}{n} \langle \bar{S}, S^{-1} \rangle \Big). \tag{15}$$

This shows that the error depends on the divergence between the second moment matrices of $p(x)$ and $\bar{p}(x)$:

- When $p(x) = \bar{p}(x)$, then $\langle \bar{S}, S^{-1} \rangle = \mathrm{tr}(\bar{S}S^{-1}) \approx \mathrm{tr}(I) = d$, so the error decays as $\frac{d}{n}$.
- If S is low-rank and is missing any directions that appear in $\bar{S}$, then the error is infinite. This makes sense, as we have no way of estimating β along the missing directions, and we need to be able to estimate β in those directions to get good error under $\bar{p}$. We can get non-infinite bounds if we further assume some norm bound on β; e.g. if $\|\beta\|_2$ is bounded then the missing directions only contribute some finite error.
- On the other hand, if S is full-rank but $\bar{S}$ is low-rank, then we still achieve finite error. For instance, suppose that $S = I$ is the identity, and $\bar{S} = \frac{d}{k}P$ is a projection matrix onto a k-dimensional subspace, scaled to have trace d. Then we get a sample complexity of $\frac{d}{n}$, although if we had observed samples with second moment matrix $\bar{S}$ at training time, we would have gotten a better sample complexity of $\frac{k}{n}$.
- In general it is always better for S to be bigger. This is partially an artefact of the noise σ^2 being the same for all X, so we would always rather have X be as far out as possible since it pins down β more effectively. If the noise was proportional to $\|X\|_F$ (for instance) then the answer would be different.

Robust OOD error estimate. We can also estimate the OOD error even when the Gaussian assumption doesn't hold, using the same idea as for robust standard errors. Letting $\bar{z}_i$ be the noise for $\bar{x}_i$, the squared error is then $\frac{1}{m}\sum_{j=1}^m (\langle \beta - \hat{\beta}, \bar{x}_i \rangle + \bar{z}_i)^2$, and computing the expectation given $x_{1:n}, \bar{x}_{1:m}$ yields

$$\mathbb{E}[\frac{1}{m}\sum_{j=1}^m (\langle \beta - \hat{\beta}, \bar{x}_i \rangle + \bar{z}_i)^2 \mid x_{1:n}, \bar{x}_{1:m}] \tag{16}$$

$$= \frac{1}{m}\sum_{i=1}^m \bar{x}_i^\top \mathbb{E}[(\beta - \hat{\beta})(\beta - \hat{\beta})^\top \mid x_{1:n}] \bar{x}_i + 2\bar{x}_i^\top \mathbb{E}[\beta - \hat{\beta} \mid x_{1:n}] \mathbb{E}[\bar{z}_i \mid x_i] + \mathbb{E}[\bar{z}_i^2 \mid \bar{x}_i] \tag{17}$$

$$= \Big\langle \bar{S}, S^{-1} \Big(\frac{1}{n}\Omega + bb^\top \Big) S^{-1} \Big\rangle + 2 \Big\langle \bar{b}, S^{-1} b \Big\rangle + \frac{1}{m}\sum_{j=1}^m \mathbb{E}[\bar{z}_i^2 \mid \bar{x}_i]. \tag{18}$$

To interpret this expression, first assume that the true model is "actually linear", meaning that $b = \bar{b} = 0$. Then the expression reduces to $\frac{1}{n}\langle \bar{S}, S^{-1}\Omega S^{-1} \rangle + \frac{1}{m}\sum_{j=1}^m \mathbb{E}[\bar{z}_i^2 \mid x_i]$. The second term is the intrinsic variance in the data, while the first term is similar to the $\frac{1}{n}\langle \bar{S}, S^{-1} \rangle$ term from before, but accounts for correlation between X and the variation in Z.

If the model is not actually linear, then we need to decide how to define β (since the optimal β is then no longer independent of the distribution). In that case a natural choice is to let β be the minimizer under the training distribution, in which case $b \to 0$ as $n \to \infty$ and thus the $\langle \bar{b}, S^{-1} b \rangle$ term conveniently becomes asymptotically negligible. The twist is that $\mathbb{E}[\bar{z}_i^2 \mid \bar{x}_i]$ now measures not just the intrinsic variance but also the departure from linearity, and could be quite large if the linear extrapolation away from the training points ends up being poor.

Partial specification. In general, we see that we can actually form good estimates of the mean-squared error on $\bar{p}$ making only certain moment assumptions (e.g. $b = \bar{b} = 0$) rather than needing to assume the Gaussian model is correct. This idea is called *partial specification*, where rather than making assumptions that are stringent enough to specify a parametric family, we make weaker assumptions that are typically insufficient to even yield a likelihood, but show that our estimates are still valid under those weaker assumptions. The weaker the assumptions, the more happy we are. Of course $b = \bar{b} = 0$ is still fairly strong, but much better than Gaussianity. The goal of partial specification aligns with our earlier desire to design estimators for the entire family of resilient distributions, rather than specific parametric classes. We will study other variants of partial specification later in the course, in the context of clustering algorithms.