This week: Results of Cross Validation
Bootstrap
Monte Carlo tests (Panini)

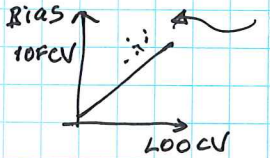
- Recall $Y = f(X) + \varepsilon$ Goal: estimate θ (expected test MSE) of an estimator $\hat{f}$ of f using cross-validation

  ex. $\hat{\theta}_{LOOCV}$, $\hat{\theta}_{10FCV}$. * (fit on 90% of data, so it is more biased, overestimate)

  - LOOCV: average the test error of each datum.
    → more $\hat{f}$ but less correlated
  - 10FCV: get the test error over the (whole) dataset (ie. not individual data)
    → fewer $\hat{f}$ but more correlated
  - Select / change X, ε, params

(Reviewing X Valid'n results of assignment surveys - ref R code)

↳ assess / compare: test MSE vs. mean of 10FCV, LOOCV (results obtained via CV)
- Variance of 10FCV, LOOCV



we expect positive bias for 10FCV (overestimates)
10FCV has also higher Variance.
(MSE is dominated by Variance, not so much by bias.

- Bootstrap - see lecture slides. Additional remarks:
  - P^* is the conditional distribution based on
  - estimate bias using boot package in R
    - note that boot. is asymptotic
  - $Var(\hat{\theta}_n)$ and $Var^*(\hat{\theta}_n^*)$ are similar for large sample sizes.

$$\hat{\theta}_n^* - \hat{\theta}_n \approx \hat{\theta}_n - \theta$$

  - Consider the slightly weaker statement $\dfrac{\hat{\theta}_n^* - \hat{\theta}_n}{sd(\hat{\theta}_n^*)} \approx \dfrac{\hat{\theta}_n - \theta}{sd(\hat{\theta}_n)}$ (how to obtain these?)

- Given $Z_1, \ldots, Z_n$,

so $Z_1^{*1}, \ldots, Z_n^{*1} \to \hat{\theta}_n^{*1}$
$Z_1^{*2}, \ldots, Z_n^{*2} \to \hat{\theta}_n^{*2}$
$\vdots$
$Z_1^{*B}, \ldots, Z_n^{*B} \to \hat{\theta}_n^{*B}$

B bootstraps → get the sample variance = $\hat{sd}(\hat{\theta}_n)$

- How do we get $\hat{sd}(\hat{\theta}_n^*)$?
  Do another level of bootstrap:
  ex. for $\hat{\theta}_n^{*1}$,

$Z_1^{**}, \ldots, Z_n^{**} \to \hat{\theta}_n^{**1}$
$Z_1^{**}, \ldots, Z_n^{**} \to \hat{\theta}_n^{**2}$
$\vdots$
$Z_1^{**}, \ldots, Z_n^{**} \to \hat{\theta}_n^{**M}$

M bootstraps → sample variance = $\hat{sd}(\hat{\theta}_n^{*1})$

 Computational Statistics Lecture 12 Thu, 12 April 18

- Bootstrap CIs: normal: 1st order asymptotically correct
  bootstrap T: 2nd order asymptotically correct
  → good for location params but not for ratios.
  - Is the ref. size (n) enough? Assess via QQ plot: if it is roughly Gaussian distributed, then that's good (n is probably big enough) – can be confident about bootstrap results. If not, probably shouldn't trust. $\hat{P}_n$ is empirical.
  - If the parametric model is bad, use nonparametric perhaps?

$Z_1, \dots, Z_n \to \hat{\theta}$. To get $\hat{sd}(\hat{\theta}_n)$, we generate B bootstrapped datasets and we take the sample variance of $\hat{\theta}_n^{*1}, \dots, \hat{\theta}_n^{*B}$. Next, we need the quantiles of $\frac{\hat{\theta}_n^* - \hat{\theta}_n}{sd(\hat{\theta}_n^*)}$

$Z_1^{*1}, \dots, Z_n^{*1} \to \hat{\theta}_n^{*1}$

$\vdots$

$Z_1^{*B}, \dots, Z_n^{*B} \to \hat{\theta}_n^{*B}$

We approximate this by the empirical quantiles of

$\frac{\hat{\theta}_n^{*1} - \hat{\theta}_n}{\hat{sd}(\hat{\theta}_n^{*1})}, \dots, \frac{\hat{\theta}_n^{*B} - \hat{\theta}_n}{\hat{sd}(\hat{\theta}_n^{*B})}$

Hesterberg: check that bias, variance is consistent; else $\hat{\theta}_n - \theta$ not so accurate? $\frac{\hat{\theta}_n - \theta}{sd(\hat{\theta}_n)}$

- ref R code (ex. airconditioning: aircond (from boot lib.)
  - Parametric bootstrap: sample from param. data rather than from $\hat{P}_n$
  - rexp(~): samples from exponential distribution using some params. arg air.rg function corresponds to params.

- Bootstrap for regression $E[Y|X=x] = \mu(x)$
  - Check that the bootstrapped samples (sets) look like the data.
  - Why can't we fix the x's and just resample the y's? We lose any trends in the data

ex. 

If, however, there are heteroskedastic errors, ie. errors are not iid, then we lose trends in sampling the residuals.

ex.