Final Exam schedule is up: May 12, 5:40 - 7:40pm.
(check Banner.)

Midterm 1: Monday March 2.

Special homework for students registered for 531 is posted. If you are registered for 455 do NOT do this homework.

Two ideas from chapter 3: the bootstrap and the jackknife.

The bootstrap: suppose we have a data set with n rows. We can sample this data set with replacement to get many data sets out of one.

Let's say we have an estimator $\hat{\beta}_1$.
One way of getting a variance or s.d. for $\hat{\beta}_1$ is to use the theory of linear regression (or whatever theory we have of our estimator)

This depends on assumptions.

With the bootstrap, we can get a variance without making such assumptions. How?

We sample our data set with replacement, creating $data_1, data_2, \ldots, data_N$.

We compute $\hat{\beta}_1(data_1), \ldots, \hat{\beta}_1(data_N)$.

These numbers have a sample variance.

This is a different type of estimate from the one in the linear regression output.

Standard fact about the bootstrap: on average, the new data sets will leave out about $1/3$ of the rows in our original data set.

Why $1/3$? Actually it is about $1/e$.

What is the probability that a single row, say row 1, is left out? We sample n times.

The chance we miss row 1 is $\left(1 - \frac{1}{n}\right)$

(n rows in original data set.)

After n samples, the chance of missing row 1 is $\prod_{i=1}^{n}\left(1-\frac{1}{n}\right) \approx e^{-1}$ by Calc 2.

This applies to any row, not just row 1.

So, on average $1/e$ of the rows are left out.

This leftover data has a use: the estimates we made ($\hat{\beta}_1(\text{data}_i)$) never "saw" the rows left out of data_i. So we can look at the errors of our model on these "unseen" rows.

This is a better error estimate than the error we see on the data used to create the estimate.

See also: cross-validation.

Basic idea: divide data into bins (say 5) train model on any 4 bins: test model on 1 bin, which was not used for training.

When is this bootstrap/cross-validation/etc.

NOT a good idea?

A: When the data has extra structure which is NOT preserved by sampling / binning / etc.

Example: airpass data set, Faraway Ch. 4.

Note that we have courses in "time-series analysis" in which special techniques are used to deal with this type of data.

What is the jackknife?

The jackknife: leave out one row of data.

estimate $\hat{\beta}_1$ computed from data$_{(i)}$ = data with i^{th} row deleted.

Then we can look at the error on the i^{th} row.

This (as in the bootstrap) is error on "unseen" data.

Advantage over bootstrap: if we leave out <u>exactly</u> one point, we have simple formulas and don't need to do so much computing.

Formulas for jackknife residuals Faraway p. 87.

In the standard regression output, we see p-values.

These are associated to a test of $H_0 : \beta_i = 0$

vs. H_a $\beta_i \neq 0$.

What if we want to test $H_0 : \beta_i = 0.5$

vs. $H_a : \beta_i \neq 0.5$?

What we do: compute $\dfrac{\hat{\beta}_i - 0.5}{\text{s.e.}(\hat{\beta}_i)}$.

this has the t-dist. (with $n-p$ d.o.f.)

We can look up the p-value with R.

Another test we can do: Note that e.g. $\hat{\beta}_0, \hat{\beta}_1$ not only have variances but also a covariance.

We can ~~test~~ create (instead of a confidence interval for β_i) a joint confidence ellipse for β_0, β_1. (See Faraway Ch 3.5)

Remember that there is a close relationship between hypothesis testing and confidence intervals.

The most basic confidence interval:

$$\left[\hat{\beta}_i - t_{n-p}^{(\alpha/2)} \text{ s.e.}(\hat{\beta}_i),\ \hat{\beta}_i + t_{n-p}^{(\alpha/2)} \text{ s.e.}(\hat{\beta}_i)\right]$$

this is the confidence interval for β_i.

Super-simple form (95% CI with n large)

$$\left[\hat{\beta}_i - 2 \text{ s.e.}(\hat{\beta}_i),\ \hat{\beta}_i + 2 \text{ s.e.}(\hat{\beta}_i)\right]$$