I. Multivariate Methods: Correlation and Regression

1. Same idea: Draw new observations from old multivariate distribution.
   a. Calculate test statistic whose distribution is desired for each new sample,
   b. Use methods from earlier to get confidence interval for expectation of this statistic
      i. Percentile
      ii. BCa
      iii. etc.
   c. Use for correlation, or regression $Y_j = \beta_0 + \beta_1 X_j + \epsilon_j, \epsilon_j \text{ iid.}$
   d. More generally, $Y = X\beta + \epsilon$, with $X = \begin{pmatrix} 1 & X_1 \\ & \vdots \\ & \vdots \\ & 1 & X_n \end{pmatrix}$. [Mark A R]

2. New idea: Keep old X’s, but draw new residuals.
   a. Why? Maybe covariates were fixed by design
   b. How?
      i. Fit regression line
ii. Calculate residuals and fitted values

iii. Resample from residuals

iv. Remember from 960:563 that residuals are not IID?
   • Sum to 0
   • Residuals associated with extreme values of explanatory variables are more variable than others.

v. Partial Solution:
   • Calculate $h_j = \text{diagonal elements of the matrix } X(X^\top X)^{-1}X^\top$.
   • Divide residual by $\sqrt{1 - h_j}$ for $h_j$ the “hat value”,

vi. Create new response variable by adding resampled residuals to old fitted values

vii. Fit regression line to new response and old explanatory variable

viii. Keep as the statistic value the regression parameter associated with the explanatory variable

c. Use same methods as above for inference.

3. Either the method of resampling new observation vectors, or resampling new residuals, works for multiple regression.

a. Method can be used to get Cl for
Lecture 10

i. one or more regression coefficients,

ii. $R^2$,

iii. etc.

4. In order to use the bootstrap-t confidence interval method, retain not only the parameter estimates, but the standard errors as well.

5. We can use simulation to assess how well our procedure does.

J. Analysis of Variance Models

1. Special Case: Two Sample Testing with Equal Variance
   a. Want inference on a shift parameter
   b. Assume distributions are the same, up to a shift
      i. Same assumption as for MWW test.
   c. Write as regression model.
      i. Let $x_j = \begin{cases} 1 & \text{if observation is in group 2} \\ 0 & \text{if observation is in group 1} \end{cases}$
      ii. Then $\beta_1$ is shift parameter
      iii. Fitted values are group means.
      iv. Residuals are differences from group means.
v. Estimate of shift parameter is difference in group averages.

vi. Standard error of shift estimate is the usual pooled estimate of standard deviation, times the square root of the sum of the sample sizes.

d. Writing as $t$-test as below is generally faster.

2. Special NonCase: Two Sample Testing with Unequal Error Distribution

a. Not a special case of the regression model, since variances are unequal

b. Bootstrap approach is pretty general: don’t assume that any aspect of distribution in two groups is the same.

c. Proceed as above, except resampling of residuals is done within group.

i. So, if group sizes are $m$ and $n$, the total number of possible bootstrap samples is $m^m n^n$ rather than $(m + n)^{m+n}$.

ii. You can resample Ys instead of residuals, since items resampled in each group have the same mean. [Mark D R]

3. $K$-sample testing, for $K > 2$

a. Same as earlier:
i. calculate residuals

ii. resample new residuals from original residuals

iii. Add residuals to fitted values to get new data

iv. Proceed with analysis.

b. Sample tasks:

i. Can do inference on separate pre-specified group differences
   • Advantage over approach retaining only these groups is likely limited.

ii. Can do more exotic things like confidence interval for extreme values of means.

4. Parametric Bootstrap:

a. Replace sampling from sample by sampling from a parametric distribution whose mean and variance match the sample. [Mark E-R]

ST: 1.3

K. Related technique: Jackknife

1. Suppose that bias of estimator \( T \) of \( \theta \) based on \( n \) observations is

\[
a/n + b/n^2 + O(1/n^2).
\]

a. Ex., \( X_1, \ldots, X_n \) independent and identically distributed
Lecture 11

\[ N(\mu, \sigma^2), \]

i. MLE for \( \sigma^2 \) is \( \hat{\sigma}^2 = \frac{\sum_{j=1}^{n} (X_j - \bar{X})^2}{n} \)

ii. \( n\hat{\sigma}^2/(n - 1) \) is unbiased estimator of \( \sigma^2 \).

iii. \( E_{\sigma^2} [\hat{\sigma}^2] = (n - 1)\sigma^2/n = \sigma^2 - \sigma^2/n \)

b. Ex., \( W \) is an unbiased estimate of \( \omega \), \( T = g(W) \), \( \theta = g(\omega) \), \( \text{Var} [W] < \infty \).

c. \( g(W) \approx g(\omega) + g'(\omega)(W - \omega) + g''(\omega)(W - \omega)^2/2 + g'''(\omega)(W - \omega)^3/6 + g''''(\omega)(W - \omega)^4/24 \)

d. \( E[g(W)] \approx g(\omega) + g''(\omega)\text{Var}_W [/] 2 + g'''(\omega)E[(W - \omega)^3]/6 + g''''(\omega)E[(W - \omega)^4]/24 \)

e. Get above expansion if skewness is 0.

2. Let \( T_{n-1,i}^* \) be the estimator based on the sample of size \( n - 1 \) with observation \( i \) omitted.

3. Let \( \bar{T}_n^* = \sum_{i=1}^{n} T_{n-1,i}^*/n \).

a. Then bias of \( \bar{T}_n^* \) is \( \approx a/(n - 1) \).

b. Let \( B = (n - 1)(\bar{T}_n^* - T_n) \) be an estimate of the bias.

i. Why? \( E[B] \approx (n - 1)(\theta + a/(n - 1) - \theta - a/n) = a(1 - a(n - 1)/n) = a/n. \)

ii. Then bias of \( T_n - B \) is \( O(1/n^2) \).