Class #14 Wednesday 2 March 2011

What did we cover last time?

- Hypothesis Testing Types
  - Student's t-test - practical equations
  - Effective degrees of freedom
- Parametric Tests
  - Chi squared test
  - Kolmogorov-Smirnov
- Nonparametric Tests
  - Classical
    - Wilcoxen-Mann-Whitney
    - Kolmogorov-Smirnov
    - Wilcoxen signed-rank test for paired samples
- Field Significance Testing

4.3 Nonparametric Tests cont...

❋(B) Resampling tests

❋Construct synthetic data, resample data (aka Randomization or Monte Carlo tests) in manner consistent with Null Hypothesis

❋Construct artificial values of test statistic from the original data

❋ Computer does resampling for you

❋(B1) Permutation Tests

❋ Natural generalization of Wilcoxen-Mann-Whitney test (Recall lightning example, pool data, Rank, sum for each group)

❋ Here you, pool two groups and randomly sample to get one U statistic, then calculate significance, repeat this 10,000 times, get distribution of U's

4.3 Nonparametric Tests cont...

❋(B1) Permutation Tests cont...

❋Example 5.9 Lightning data example

❋Apply L-scale statistic:

$$\lambda_2 = \frac{(n-2)!}{n!} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left| x_i - x_j \right| \quad 5.27$$

TABLE 5.5 Counts of cloud-to-ground lightning for experimentally seeded and nonseeded storms. From Baughman *et al.* (1976).

Seeded		Unseeded	
Date	Lightning strikes	Date	Lightning strikes
7/20/65	49	7/2/65	61
7/21/65	4	7/4/65	33
7/29/65	18	7/4/65	62
8/27/65	26	7/8/65	45
7/6/66	29	8/19/65	0
7/14/66	9	8/19/65	30
7/14/66	16	7/12/66	82
7/14/66	12	8/4/66	10
7/15/66	2	9/7/66	20
7/15/66	22	9/12/66	358
8/29/66	10	7/3/67	63
8/29/66	34		

❋ 5.27 - half the average difference (absolute value) between all possible pairs in a sample size n.

❋ Compare seeded & unseeded by λ diff or ratio.

❋ Choose λ(seeded)/λ(unseeded), 1 same distribution

❋ Choice of L-scale stat is arbitrary and for illustrative purposes.

❋ Null hypothesis, the samples have the same L-scale

❋ Sample 10,000 of 1,352,078 possible. (23!)/[12! 11!]

❋ L-scale 0.188 in observed case

4.3 Nonparametric Tests cont...

❋(B1) Permutation Tests cont...

❋Example 5.9 Lightning data example cont...

❋ 0.188 is smaller than all but 49 of the 10,000 cases, so the null hypothesis is rejected.

❋ 49 out of 10,000 then p=0.0049 (1-tail) or 0.0098 (2 tailed)

❋ Bimodal distribution

❋ the 358 outlier is cause

❋

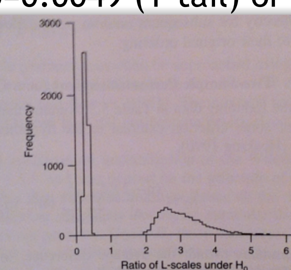


FIGURE 5.7 Histogram for the null distribution of the ratio of the *L*-scales for lightning counts of seeded vs. unseeded storms in Table 5.5. The observed ratio of 0.188 is smaller than all but 49 of the 10,000 permutation realizations of the ratio, which provides very strong evidence that the lightning production by seeded storms was less variable than by unseeded storms. This null distribution is bimodal because the one outlier (353 strikes on 9/12/66) produces a very large *L*-scale in whichever of the two partitions it has been randomly assigned.

4.3 Nonparametric Tests cont...

✷(B2) Bootstrap Test

✷ Used when you have one-sample setting

✷ Key difference: Resampling with replacement (slips of paper in hat analogy)

✳ Example 5.10: Use on Ithaca January precip (1933-82) with the statistic stdev of log x: s_{lnx}, arbitrary, with n=50 is 0.537.

4.3 Nonparametric Tests cont...

✷(B2) Bootstrap Test cont...

✳ Example continued... s_{lnx} is 0.537 for observed

✳ 10,000 bootstrap estimates, histogram below

✳ (1-α)% confidence interval, is n_B*α/2, values are 0.41 and 0.65 on histogram.

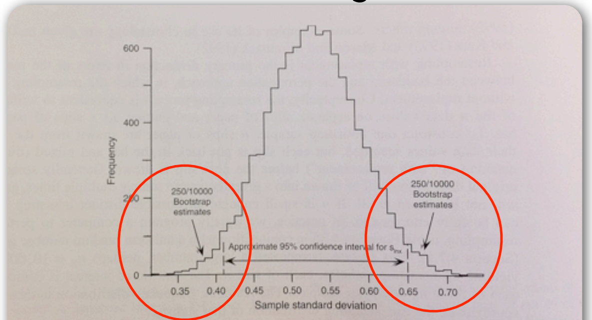


FIGURE 5.8 Histogram of $n_B = 10,000$ bootstrap estimates of the standard deviation of the logarithms of the 1933–1982 Ithaca January precipitation data. The sample standard deviation computed directly from the data is 0.537. The 95% confidence interval for the statistic, as estimated using the percentile method, is also shown.

4.3 Nonparametric Tests cont...

✷(B2) Bootstrap Test cont...

✳ Bootstrap can be used for two or more samples but kept separate and resampled, for example: Eq 5.5 for difference of means test.

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - E[\bar{x}_1 - \bar{x}_2]}{[V\hat{a}r(\bar{x}_1 - \bar{x}_2)]^{1/2}} \quad (5.5)$$

✳ Estimate the sampling distribution of the test statistic directly Example 5.11: Ratio of L-scales for lightning strikes for seeded and unseeded.

✳ Null hypothesis- All aspects of the lightning between the seeded and unseeded are the same.

4.3 Nonparametric Tests cont...

✷(B2) Bootstrap Test cont...

✳Example 5.11 continued: Bootstrap n_1=12 (seeded) and n_2=11 (unseeded). n_B=10,000, Figure 5.9 shows the distribution of λ(seeded)/λ(unseeded).

✳Grey arrow, 95% confidence limits from 0.08 to 0.75
- Getting 1 has a low probability.
- H_0 would be rejected at the 95% level.
- greater than 1 for 33 cases so can reject at 1% levels also

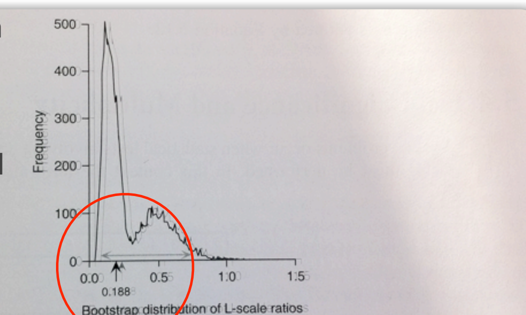


FIGURE 5.9 Bootstrap distribution for the ratio of L-scales for lightning strikes in seeded and unseeded storms, Table 5.5. The ratio is greater than 1 for only 33 of 10,000 bootstrap samples, indicating that a null hypothesis of equal L-scales would be rejected. Also shown (grey arrows) is the 95% confidence interval for the ratio, which ranges from 0.08–0.75.

Discussion Exercise: Problem 5.10 Wilks

5.10. A published article contains a statistical analysis of historical summer precipitation data in relation to summer temperatures using individual *t* tests for 121 locations at the 10% level. The study investigates the null hypothesis of no difference in total precipitation between the 10 warmest summers in the period 1900–1969 and the remaining 60 summers, reports that 15 of the 121 tests exhibit significant results, and claims that the overall pattern is therefore significant. Evaluate this claim.

Chapter 5 Statistical Forecasting

❋5.1 Background

❋ Weather and climate forecasting has a statistical basis as there are many things not in our models...

❋ Classical Methods - no input from NWP (Numerical Weather Prediction) models or fluid dynamical models. Prominent before NWP.

❋ Other Method - Use in conjunction with NWP, enhance numerical forecasts with statistics. Important to get forecast for a city as opposed to a grid point.

❋ Much of statistical weather forecasting is based on linear regression based on the least squares principle.

Chapter 5 Statistical Forecasting continued

❋ 5.2 Linear Regression (A) Calculate Regressions

❋ Linear relationship between 2 variables, x independent variable(predictor) and y dependent variable (predictand).

❋ Criteria 1: Minimize squared error, easy to compute but vulnerable to outliers.

❋ Other criteria: Least absolute deviation (LAD), does not have a theoretical basis so it is an iterative process.

$$\hat{y} = a + bx \qquad (6.2)$$

$$e_i = y_i - \hat{y}(x_i) \qquad (6.3)$$

❋ Each data pair will have a residual e_i.

$$y_i = \hat{y}(x_i) + e_i = a + bx + \hat{y}(x_i) \qquad (6.4)$$

5.2 Linear Regression continued ..

❋ Minimize the function and get b and a.

$$b = \frac{\sum_{i=1}^{n}\left[(x_i - \bar{x})(y_i - \bar{y})\right]}{\left[\sum_{i=1}^{n}\left[(x_i - \bar{x})\right]^2\right]} = \frac{n\sum_{i=1}^{n}\left[x_i y_i\right] - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n}\left[x_i\right]^2 - \left[\sum_{i=1}^{n} x_i\right]^2} \qquad (6.7a)$$

$$a = \bar{y} - b\bar{x} \qquad (6.7b)$$

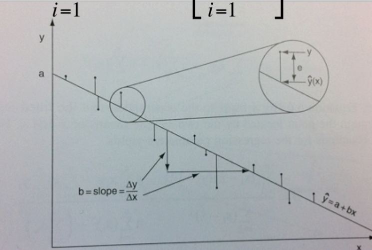


FIGURE 6.1 Schematic illustration of simple linear regression. The regression line, $\hat{y} = a + bx$, is chosen as the one minimizing some measure of the vertical differences (the residuals) between the points and the line. In least-squares regression that measure is the sum of the squared vertical distances. Inset shows the residual, e, as the difference between the data point and the regression line.

Chapter 5 Statistical Forecasting

- 5.2 Linear Regression continued..
- (B) Distribution of Residuals
- Typically assume that the errors are independent random variables with zero mean

$$\sum_{i=1}^{n} e_i = 0 \quad (6.8)$$

- Look at error variance, they scatter about some mean value (0 in this case). Conditional (less spread with known x) and Unconditional in Fig. 6.2.

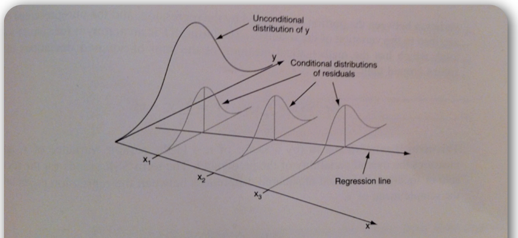


FIGURE 6.2 Schematic illustration of distributions of residuals around the regression line, conditional on values of the predictor variable, x. The actual residuals are regarded as having been drawn from these distributions.

(B) Distribution of Residuals continued ...

- Residual Variance, divide by n-2 since 2 parameters

$$s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = 0 \quad (6.9) \qquad s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left[y_i - \hat{y}(x_i)\right]^2 = 0 \quad (6.10)$$

- More usual way to calculate residual variance

$$SST = SSR + SSE \quad (6.11)$$

- **SST** (sum of squares) - Variation in predictand (y) is partitioned into variation due to regression and that due to residuals.

$$SST = \sum_{i=1}^{n}\left[y_i - \bar{y}\right]^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \quad (6.12)$$

(B) Distribution of Residuals continued ...

- SSR- Regression sum of squares in 2 forms, regression line with small slope with contribute little to SSR.

$$SSR = \sum_{i=1}^{n}\left[\hat{y}(x_i) - \bar{y}\right]^2 \quad (6.13a)$$

$$SSR = b^2\sum_{i=1}^{n}\left[x_i - \bar{x}\right]^2 = b^2\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \quad (6.13b)$$

- **SSE -** Sum of squared differences between residual and it's mean (which is 0).

$$SSE = \sum_{i=1}^{n} e_i^2 \quad (6.14)$$

$$s_e^2 = \frac{1}{n-2}\left[SST - SSR\right] = \frac{1}{n-2}\left\{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - b^2\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right]\right\} \quad (6.15)$$

(C) Analysis of Variance (ANOVA) (Section 6.2.3)

- Get this information with regression analysis and comes in this typical table form.

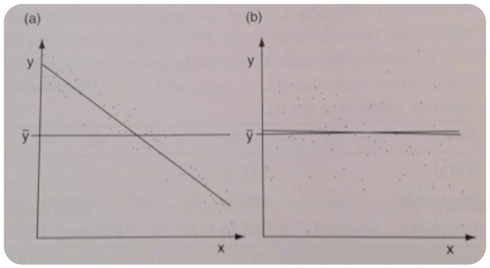
TABLE 6.1 Generic Analysis of Variance (ANOVA) table for simple linear regression. The column headings df, SS, and MS stand for degrees of freedom, sum of squares, and mean square, respectively. Regression df = 1 is particular to simple linear regression (i.e., a single predictor x). Parenthetical references are to equation numbers in the text.

Source	df	SS	MS	
Total	$n-1$	SST (6.12)		
Regression	1	SSR (6.13)	MSR = SSR/1	(F = MSR/MSE)
Residual	$n-2$	SSE (6.14)	MSE = s_e^2	

- Total is somewhat redundant and often omitted.
- MS - mean squared column, Total is just SST/(n-1) or sample variance of predictand.

Chapter 5 Statistical Forecasting

- 5.2 Linear Regression continued
- (D) Goodness of fit measures (Section 6.2.4)
- Components of ANOVA table can be used to calculate goodness of fit.
- MSE (Mean squared error) important as it tells us accuracy of forecast results. Fig. 6.3a, SSR and SST nearly the same ==> good fit. Fig6.3b SSR=0.



(D) Goodness of fit measures (Section 6.2.4) cont..

- Second measure of regression fit is R^2 coefficient of determination.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad (6.16)$$

- Proportion of the predictand variance (proportional to SST) that is described by the regression (SSR).

Note: R is the correlation between x and y!

- Third measure is F ratio MSR/MSE, increases with strength of regression. Can use F test as a measure of significance for regressions.

(E) Sampling distribution of Regression Coefficients (Section 6.2.5)

Residual variance gives estimates of sampling distribution of regression coefficients.

- Assume sampling distributions are gaussian for a and b.
- For Intercept $\mu_a = a \quad (6.17a)$

$$\sigma_a = s_e \left[\frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} [(x_i - \bar{x})]^2} \right]^{1/2} \qquad (6.17b)$$

- For slope

$$\mu_b = b \qquad (6.18a)$$

$$\sigma_b = \frac{s_e}{\left[\sum_{i=1}^{n} [(x_i - \bar{x})]^2 \right]^{1/2}} \qquad (6.18b)$$

(E) Sampling distribution of Regression Coefficients (Section 6.2.5) cont...

- Estimated slope and intercept are not independent and have the correlation:

$$r_{a,b} = \frac{-\bar{x}}{\left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 \right)^{1/2}} \qquad (6.19b)$$

- Valid for linear regression only. Regression programs also often output t and p values.
- Example 6.1 is a simple worth-looking-at example of how the ANOVA table results help evaluate a regression coefficient.

(F) Examining Residuals (Section 6.2.6)

- Examine Residuals for consistency, don't just take the computer's word for it! Not easy when done for gridded data.
- case of heteroscedasticity - non-constant residual variance. Remedy by ln transformation, which reduced bigger values more. If RHS more fanned then use y^2 to increase bigger values.

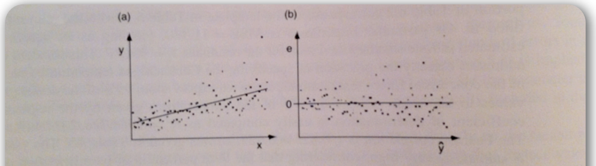
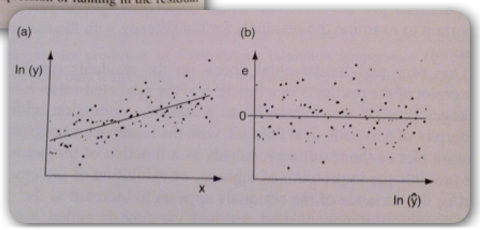


FIGURE 6.4 Hypothetical linear regression (a), and plot of the resulting residuals against the predicted values (b), for a case where the variance of the residuals is not constant. The scatter around the regression line in (a) increases for larger values of *x* and *y*, producing a visual impression of fanning in the residual plot (b). A transformation of the predictand is indicated.

Fig. 6.4 not transformed

Fig. 6.5 is ln transformed



- Scatter plot residual vs predictor, x Fig6.6.
- Q-Q plots helpful to see if residuals are Gaussian Fig 6.7. MORE STUFF ...

(G) Prediction Intervals (Section 6.2.7)

Want confidence intervals around forecast values!

❋ 95% limit is $\hat{y} = \pm 2 s_e$ good for a Gaussian distribution with a large number of n

❋ But real will be a bit higher since predictand mean and regression slope impacted by sampling variations. Also, future values used in equation will increase the variance.

$$s_{\hat{y}}^2 = s_e^2 \left[1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \qquad (6.22)$$

❋term 2- uncertainty from using sample size of n to estimate predictand mean, decreases as n increases.

❋term 3- uncertainty from slope estimation, predictions far from center of data less certain

❋ Pages 195-6 of section discusses how to put confidence intervals on the regressions.