MET3220C
Computational Statistics

Linear Regression

(Chapter 6 of Wilk's book)

Key Points:

1) Determining best fit parameters

2) Determining uncertainty in best fit parameter

3) Working with uncertain observations

http://campus.fsu.edu/
bourassa@met.fsu.edu

Fitting Parameters for a Line

- The calculations for the y-intercept and slope are:

$$\text{y-intercept} = \frac{\left(\sum_{i}^{n} x_i^2\right)\left(\sum_{i}^{n} y_i\right) - \left(\sum_{i}^{n} x_i\right)\left(\sum_{i}^{n} x_i y_i\right)}{\Delta}$$

$$\text{slope} = \frac{n\left(\sum_{i}^{n} x_i y_i\right) - \left(\sum_{i}^{n} x_i\right)\left(\sum_{i}^{n} y_i\right)}{\Delta}$$

$$\Delta = n\left(\sum_{i}^{n} x_i^2\right) - \left(\sum_{i}^{n} x_i\right)^2$$

Uncertainty Calculations

- The 'uncertainty in *y*' (Ω_y) is

$$\Omega_y^2 = \frac{1}{n-2}\sum_{i}^{n}\left(y_i - mx_i - b\right)^2$$

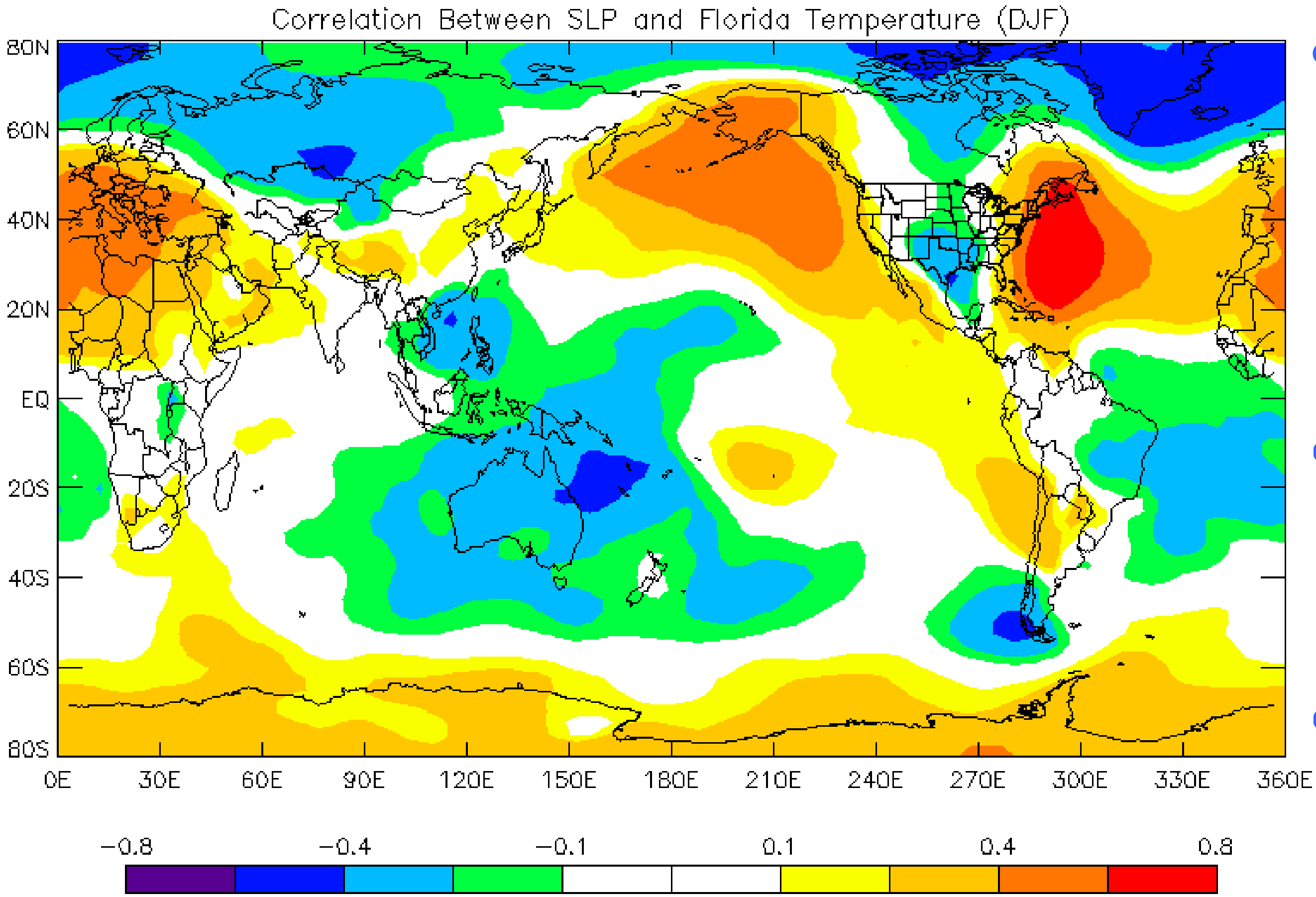
  - This can be used as error bars about your best fit line.

- The uncertainty in the slope and y-intercept are

$$\Omega_{slope}^2 = n\Omega_y^2 / \Delta$$

$$\Omega_{y-\mathrm{int}}^2 = \Omega_y^2\left(\sum_{i=1}^{n} x_i^2\right) / \Delta$$

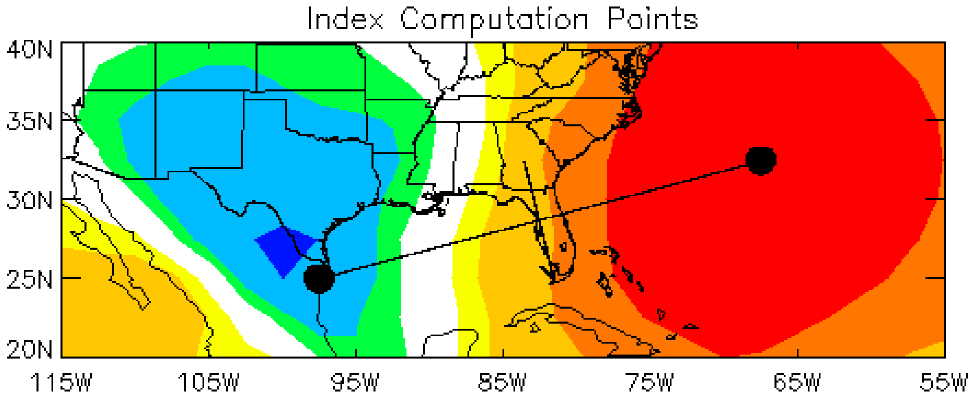
Example Correlation Between Pressure and Florida Winter Temperatures

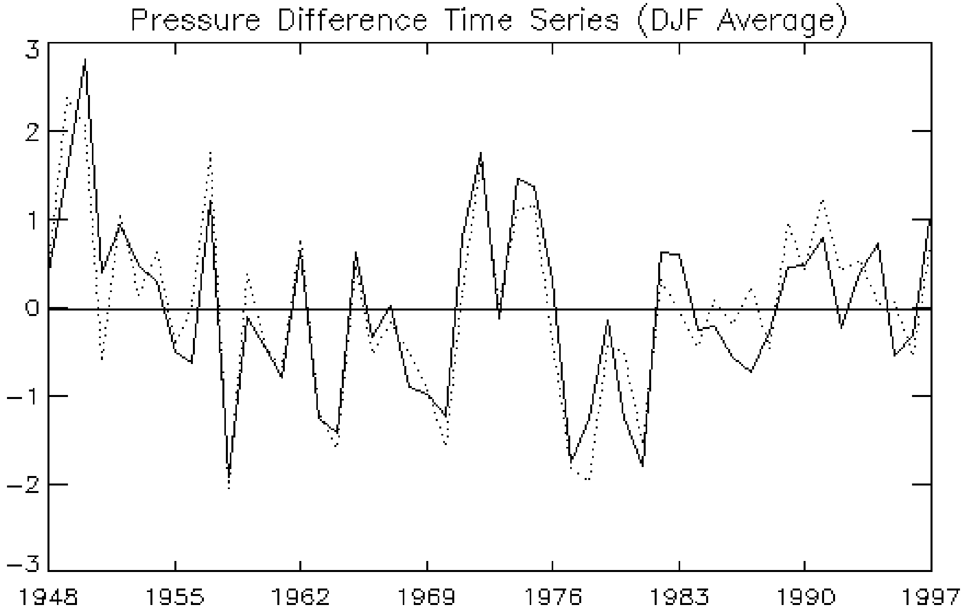


Correlation Coefficient

- A time series of average (of sorts) Winter temperature in Florida can be determined from station data.
- That time series can be correlated with modeled (analysis) pressure fields.
- The areas with high (positive or negative) correlation indicate teleconnections.

Selecting a Predictor



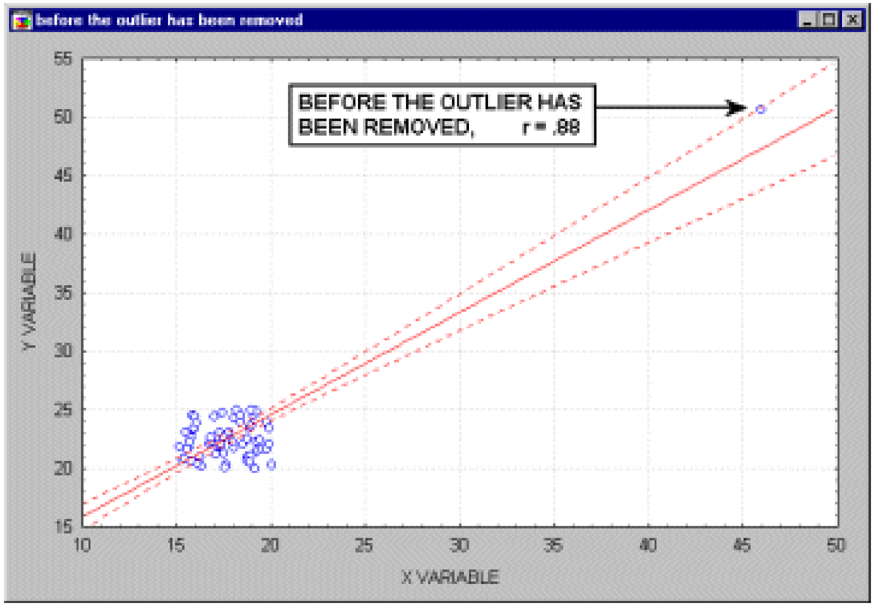


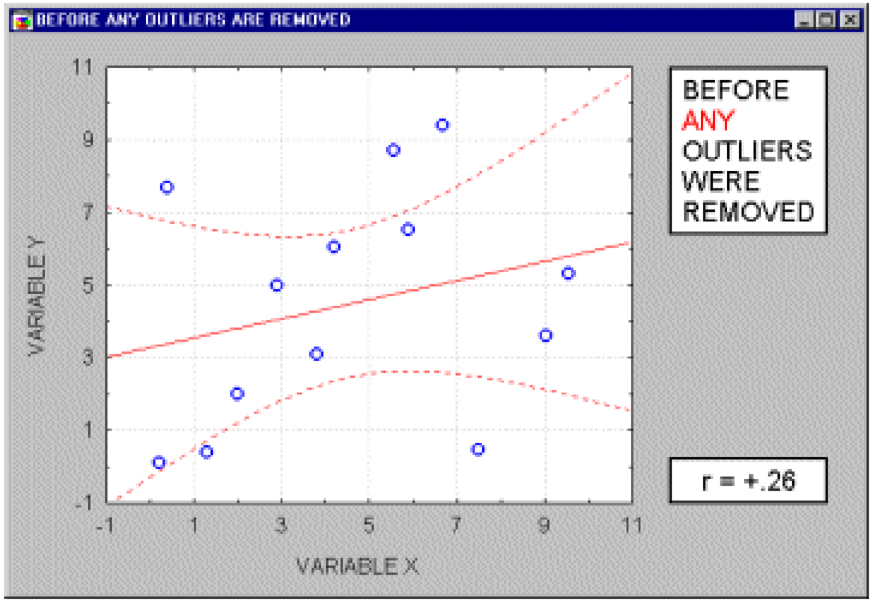
- It is somewhat safer to take predictors that are close to the location of interest.
- Less likely predictors can also have a good match. That is predictors with high correlations, but with a less obvious physical connection.
- The less likely predictors are more likely to fail in forecasts.
- Here it is assumed that a local pressure gradient is a good predictor.
- The pressure gradient explains >98% of variability in monthly temperatures.

Hypothesis Testing With Slopes

- Assume that we are interested in global warming. Why?
  - The consequences of rapid global warming could be dire.
  - The cost of attempting to prevent global warming could be huge.
    - If the cost is huge, it will come at the expense of other activities.
    - If the cost is huge, and the data does not support the existence of the problem, then much better things could be done with the money!
- We can calculate a rate of change of temperature with time (a slope).
  - How do we tell if the slope is statistically significant?
- We can assume (reasonably) that the null distribution (of values for slope) has a Gaussian distribution.
  - A z-value can be calculated by dividing the slope by the uncertainty in the slope.

Cool Examples



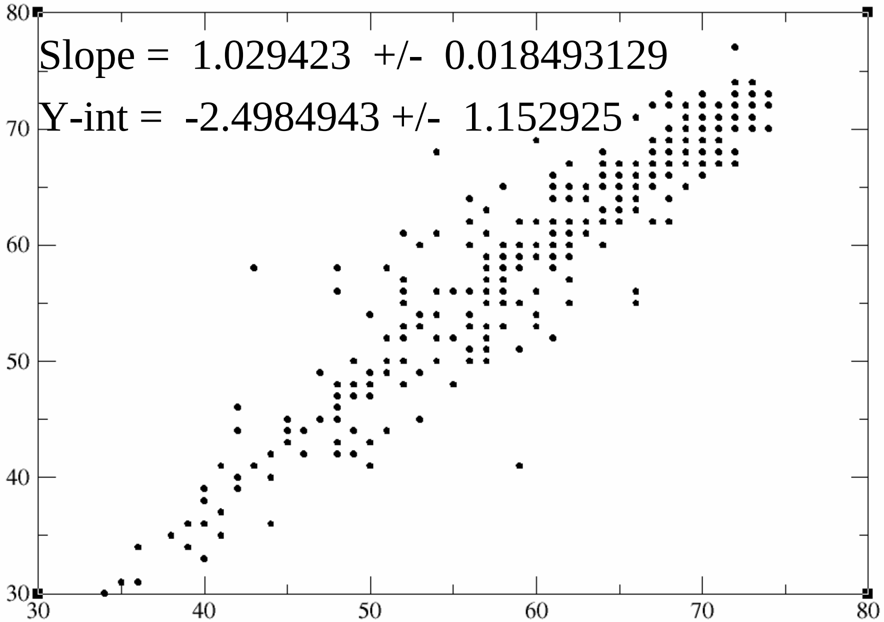


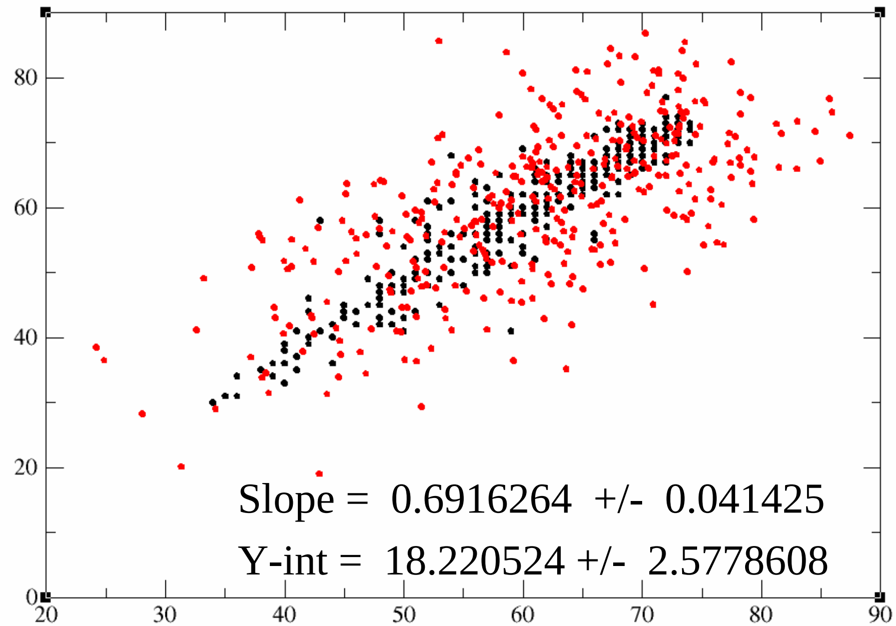
- How Big a problem are outliers?

Cool graphics from http://www.statsoft.com/textbook/stbasic.html#Correlationsd

A Problem:
Uncertainty In Paired Observations

- Linear regression assumes that uncertainty in the observations can be ignored.
  - This assumption is often not valid.
  - Results in (potentially large) errors in a slope and y-intercept!
    - Example: Uncorrupted data (left), and
    - Added random error equal to a standard deviation (right).





Why Did Purely Random Noise Change The Slope?

- When we add noise we tend to add outliers, but we also distribute the data in a more random fashion.
- The distribution looks more like a sphere, and the slope tends to be closer to zero.
- Interestingly, if we exchange the axis, the slope also is decreased, which would seem to conflict with the previously determined slope.

One Solution To the Problem of Uncertainty In Paired Observations

- In many cases there is little bias in either of the paired observations, and the gain (proportionality) is equal to 1 ($x = y$).
- The uncertainty in both sets of data can be estimated by randomly adding Gaussianly distributed noise to a perfect fit, and modifying the standard deviations of the noise with the goal of match the observed curves of $x(y)$ and $y(x)$.
- Ideally a similar data distribution should be used in this approach. Fortunately, any ball park distribution will do!