Continuing . . .

Kathleen M. Saul

April 28, 2011

Simple Linear Regression

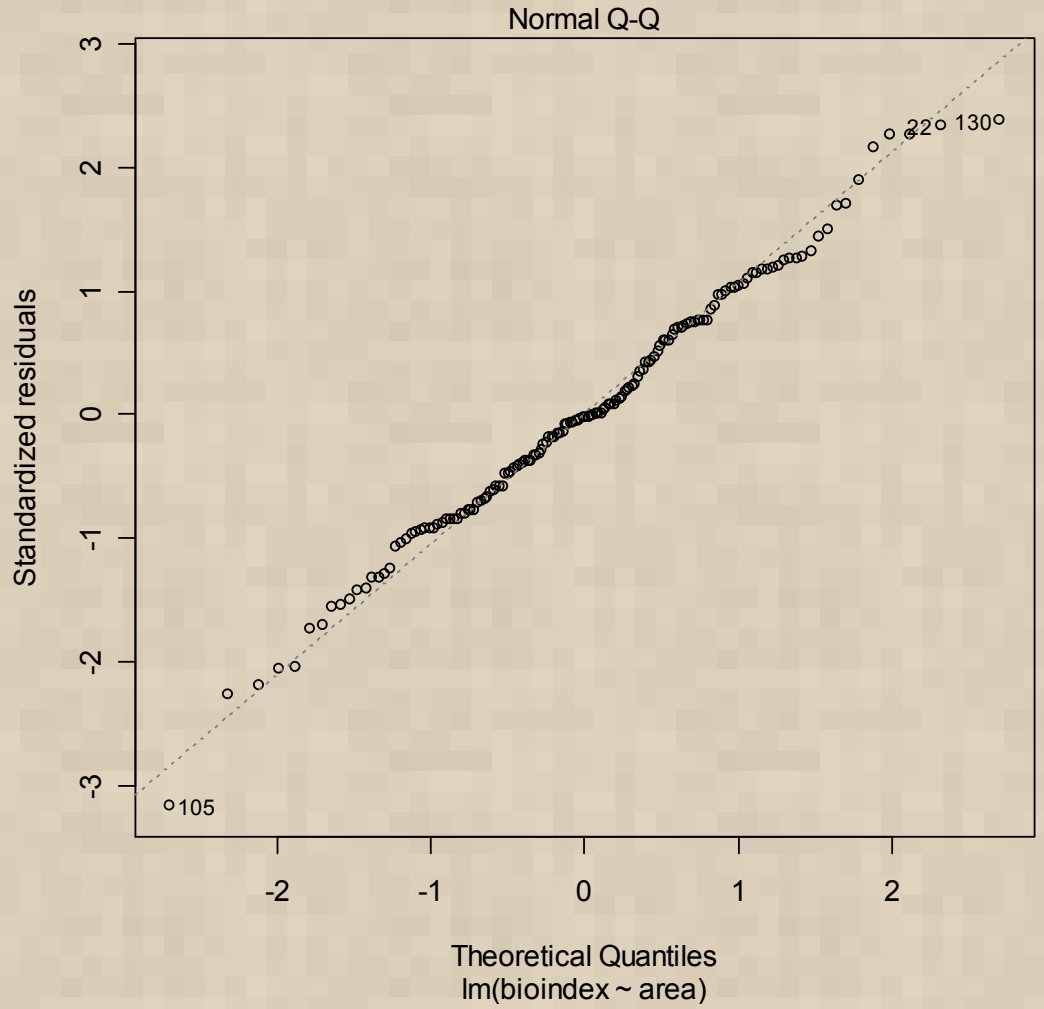
- Assumes a cause and effect relationship
- Analyzes the linear relationship between variables: $y = a + bx + error$
- Minimizes the squared vertical distance between each point and the regression line (“least squares”)
- Can be used to predict a value for the dependent variable (y), given a value for the independent variable (x)

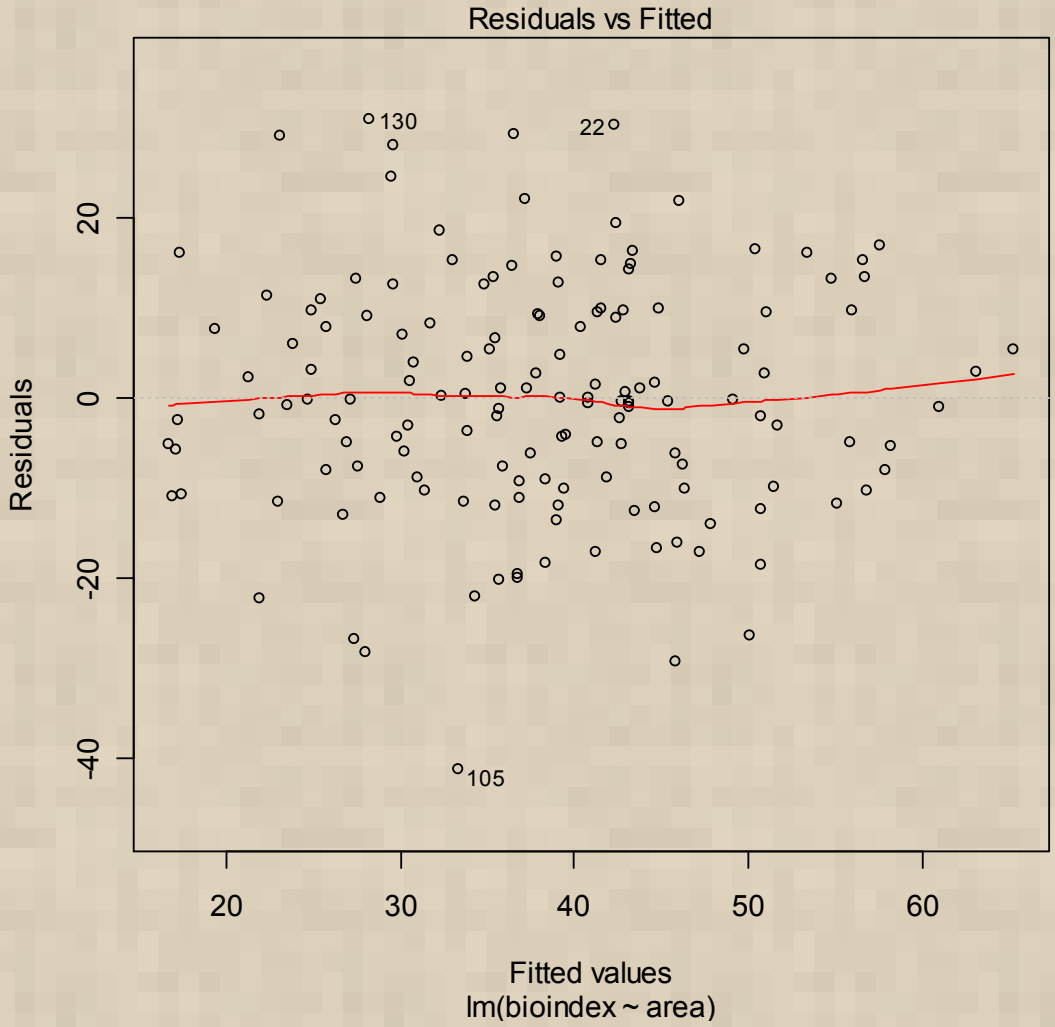
- Gives an r^2 value that tells the proportion of the variation in y explained by the variation in x
  - r^2 is just the square of the Pearson's r (correlation)

- Residuals:
- Min 1Q Median 3Q Max
- -41.1073 -9.0200 -0.3371 9.3601 30.8269
- Coefficients:
- Estimate Std. Error t value Pr(>|t|)
- (Intercept) 16.64706 2.40695 6.916 1.30e-0 ***
  - **This Pr(>|t|) tells us that there is statistically significant evidence to reject the hypothesis that that the value of the intercept is equal to zero**
- area 0.95050 0.09573 9.929 < 2e-16 ***
  - **Here, P(>|t|) tells us there is significant evidence to reject the hypothesis that the coefficient of area is equal to zero**
- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
- Residual standard error: 13 on 148 degrees of freedom
- Multiple R-squared: 0.3998, Adjusted R-squared: 0.3957
  - **The r2 value tells us that 39.98% of the variation in bioindex is explained by the variation in area**
- F-statistic: 98.58 on 1 and 148 DF, p-value: < 2.2e-16
  - **The small p-value indicates that the model is indeed a valid one.**

- For regression to be valid, the error terms must be normally distributed and independent of each other.
- How can we tell???
  - If the Residual = Observed value of y minus the expected value of y (the line)
    - Standardized Residual = Residual / standard error
  - Draw a histogram or normal quartile plot of the residuals
  - Plot the predicted value on the x axis and residuals on the y
    - There should be no pattern in the resulting plot

```
> model<-lm(bioindex~area)
> plot(model)
```




Residuals vs Fitted
130
22
105
Residuals
20
0
-20
-40
20
30
40
50
60
Fitted values
lm(bioindex ~ area)

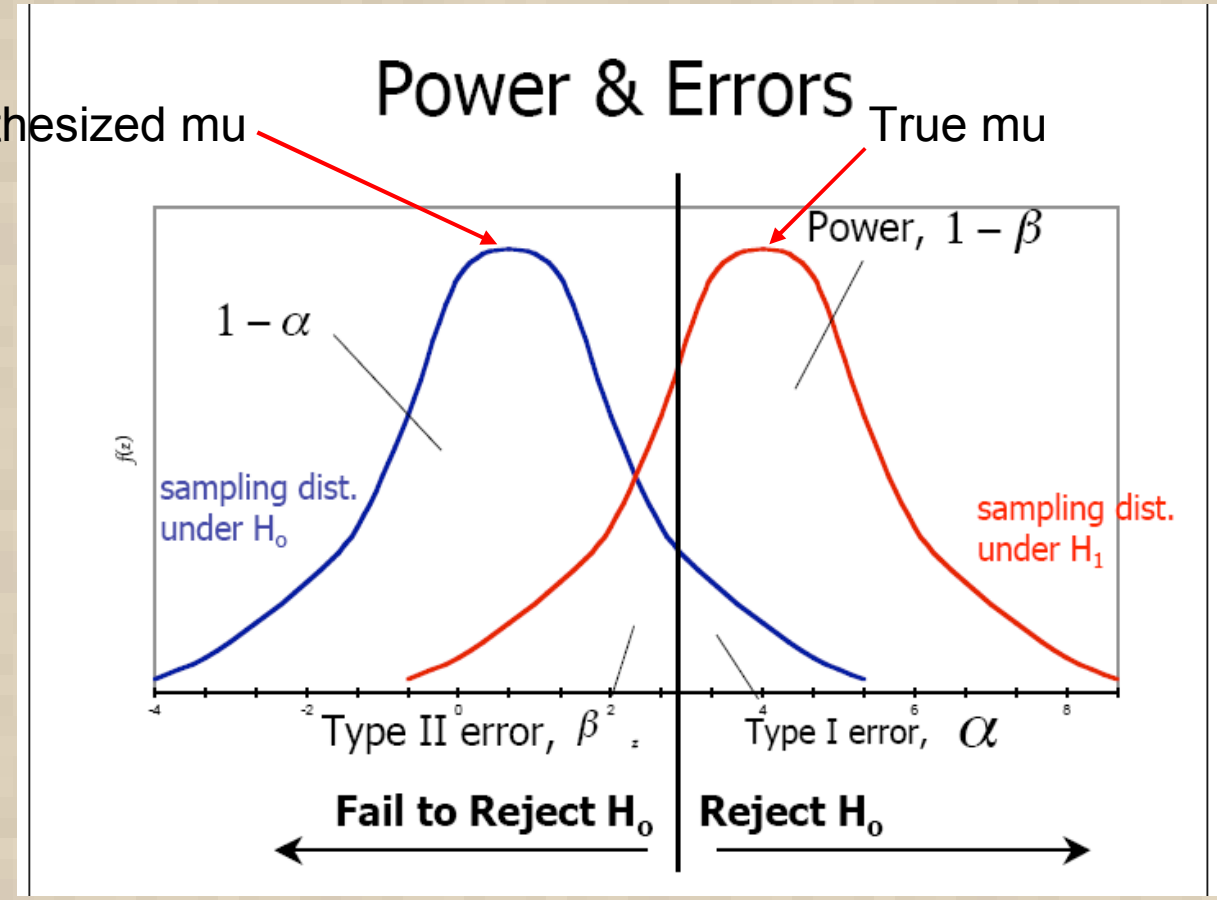
- Time series data present often result in correlated error terms and regression should not be used.
  - Ex: Water flow and sediment levels on day 1, day 5, day 10, day 15, . . .
    - The variables of interest are x = water flow and y = sediment level.
- However, if your explanatory variable is a measure of time, you can use regression with x = time.

http://andrewsforest.oregonstate.edu/

Power of a Test

Truth About a Population

Decision Based on Sample	Ho True	Ha True
Reject Ho	**Type I error** **alpha**	**Okay** **power = 1 - beta**
Do not Reject Ho	**Okay**	**Type II error** **beta**


Power & Errors
Hypothesized mu
True mu
Power, $1-\beta$
$1-\alpha$
sampling dist. under H_o
sampling dist. under H_1
Type II error, β
Type I error, α
Fail to Reject H_o
Reject H_o

- Higher power implies a test is correct more frequently.