STA 302/1001

Note Title 9/15/2011

Tuesdays 11:00 - 12:00

Tues. Sept 20 - demonstrating using SAS via remote access to Quest

Oct 11 - Andriy away

Background necessary for STA 302/1001:

Math: Matrix Algebra
- see Announcements on Blackboard

- idempotent matrices $A^2 = A$
  - see Facts on Idempotent Matrices on course web site

Second year calculus

Probability
- r.v. X, Y
- expected value (mean) and variance of r.v.
- conditional expectation and variance
  $E(Y|X)$, $\text{Var}(Y|X)$
- standard Normal distribution
- Central Limit Theorem

Statistics - estimate / estimator - unbiased
- t-test - one- and 2-sample t-tests
& Confidence Intervals

From Preface to textbook:

"It makes sense to base inferences or conclusions only on valid models."

Example: Effect of Montreal Protocol on Atmospheric Concentration of ChloroFluoroCarbons (CFCs)

Ozone layer - protects us from UV radiation

1985 - first discovery of hole in ozone layer

CFCs - destroy the ozone layer.

Montreal Protocol - schedule for phasing out manufacture of CFCs

Was it effective?

Data - atmospheric concentrations of CFCs
on Mauna Loa
units: # of parts per trillion

Define: Before Montreal Protocal (Pre-MP) before 1990
After Montreal Protocal (Post-MP) after 1994

Test: $H_0: \mu_1 = \mu_2$ where μ_1 is mean of
CFC concentration
distribution pre-MP
& μ_2 " " " " " "
post-MP

$H_a: \mu_1 \neq \mu_2$

Two-sample t-test using SAS ttest procedure
(whether I assume two groups have same variance or not)

↳ But should not assume it; calc'd s.d's are 5 and 36

$p < .0001$

Strong evidence that mean CFC concentration is not the same pre & post MP.

But mean concentration is higher post than pre-MP. Maybe CFC concentration was growing in pre-MP period!

Fatal error in this analysis: We didn't look at the data.

Looking at plot of CFC concentration versus time, interested in rate of change of CFC concentration

→ slope of line

Need a statistical model

Statistical Modelling

2 components — systematic

— random

Model how — Observed value of Y (CFC concentration) = fitted value which is function of X (time) + random error "residual"

Goal: Want to find an appropriate model (appropriate function of X) and understand error

Particular model we're going to use

Simple Linear Regression (SLR)

"simple" – one X

"linear" - in parameters (β's)
"beta"

SLR model: Y - dependent variable
response
- modelled as random

X - independent
predictor variable
explanatory
- sometimes random, sometimes not
(CFC example, X pre-planned
dates, not random)

Model: Straight line Model is an example

$$Y = \beta_0 + \beta_1 x + e$$

Y: random variable

e: random variable

"All models are wrong, some are useful."

Parameters - constants in statistical model which we usually don't know but will use data to estimate.

In straight line model, parameters:

β_0 - intercept

β_1 - slope

Estimate from data: b_0, b_1

Estimators for β_0, β_1: $\hat{\beta}_0, \hat{\beta}_1$

e - random noise

- variations in data we can't account for
- assume random with $E(e) = 0$

$$Y = \beta_0 + \beta_1 x + e$$

$$E(Y \mid X = x) = \beta_0 + \beta_1 x$$

"systematic" part of relationship

between x and y