Topic 1: Introduction

STAT 525 - Fall 2013

Outline

- Class Website
- Class Policies / Schedule
- Overview of Course Material
- SAS Statistical Software
- Tower of Pisa Example

Class Website

www.stat.purdue.edu/∼bacraig/stat525.html

- Course syllabus / Announcements
- Lecture notes
- Sample SAS programs
- Homework assignments
- Exam and homework schedule
- Information about projects
- Data sets for class and homework

Class Policies

- Attendance
  - Not required but you are responsible for announcements and lecture material
  - If you have to leave early or arrive late, notify me in advance and sit near door
- Class participation encouraged
- Questions welcomed at all times

Exams

- There will be two evening exams and a final
  - Each worth 20% of your grade
  - Must notify me at least a week prior to exam if there is scheduling conflict....prefer you to take it earlier
  - Will need a calculator with $\sqrt{\ }$ function
  - Open book / open notes
  - Strongly encourage constructing a summary sheet

Homework

- Expect "weekly" homework assignments
  - Will be due Wed at end of class
  - Format guidelines in syllabus
  - Individual vs group effort
  - Worst grade will be dropped
  - Represents 25% of your grade
  - Answer key posted after due date

Project

- Group / Team project
  - Teams determined after week 3 or 4
  - Will find "real" problem to address
  - Represents 15% of your grade
  - Check web site for upcoming details

Communication

- Office Hours
  - Mon 3:30-5:00
  - Tue 12:45-1:45
  - By appt.
- Email - bacraig@stat.purdue.edu
- Will have a class email list
- Announcements made on Web page
- Will also use piazza.com

Statistical Software

- Class Software
  - Will be using SAS for Windows 9.3
  - Available on computer lab machines
  - Can get own copy (5th floor Young)
- Free to use any software for homeworks but you are then responsible for your own software support

Getting Started with SAS

- Will provide template programs to be "copied"
- SAS handout on Web page
- Syntax Help / Examples available
  - Click 'Help'
  - Click 'SAS Help and Documentation'
  - Click 'SAS Products'
  - Click 'SAS/STAT'
  - Click 'SAS/STAT 9.3 User's Guide'
- Software Consulting Service (MATH G175)

Overview

To *conceptually* understand the use of **multiple linear regression, ANOVA, logistic, and log-linear** models for inference. This will not be a "plug-and-chug" methods course. Nor will it be a mathematical statistics course. You are expected to understand the advantages and shortcomings of each model, how to estimate the parameters, and draw valid conclusions.

Much of the homework will focus on the analysis of "real" problems and interpreting the results. Emphasis will be on the ability to present (both written and oral) these conclusions in a concise and clear manner.

Schedule

- Simple linear regression (2 wks)
- Multiple linear regression (4 wks)
- ANOVA - fixed, random, mixed (4 wks)
- Analysis of Covariance (1 wk)
- Logistic Regression (1 week)
- Categorical Data Analysis (1 wk)
- Group projects / Review (1 wk)

Statistical Model I

- Attempts to describe how the "data were generated"
- Given inherent and/or systematic variability, cannot predict outcomes/data with certainty
- Utilizes mathematical equations and probability distributions to describe the "chance" of particular outcomes
- Simplification of reality but can still be used to learn about complex system
- "All models are wrong but some are useful" - G.E. Box

Statistical Model II

- We will focus on models that look at the relationship between an outcome (response) variable Y and a set of explanatory (predictor) variables $\mathbf{X}$
- Used to serve three major purposes
  - Description
  - Control
  - Prediction
- Be wary of observational versus experimental studies
- When can model results be used to imply causality?
- Also **always** need to consider the scope of the model

Example of Linear Regression Model

- Leaning Tower of Pisa
  - Construction began in 1173 and by 1178 (2nd floor), it began to sink
  - Construction resumed in 1272. To compensate for tilt, engineers built upper levels with one side taller
  - Seventh floor completed in 1319 with bell tower added in 1372
  - Tilt continued to grow over time and was monitored. Closed in 1990.
  - Stabilization completed in 2008 by removing ground from taller side

Leaning Tower of Pisa

The Data

- Prior to stabilization, annual measurements of its lean taken for monitoring
- We have observations from 1975 - 1987
- Lean (Y) measured in tenths of a mm $>$ 2.9 meters
- Year (X) is the explantory variable
- Goals:
  - To **characterize** lean over time
  - To **predict** future observations

The Data Set

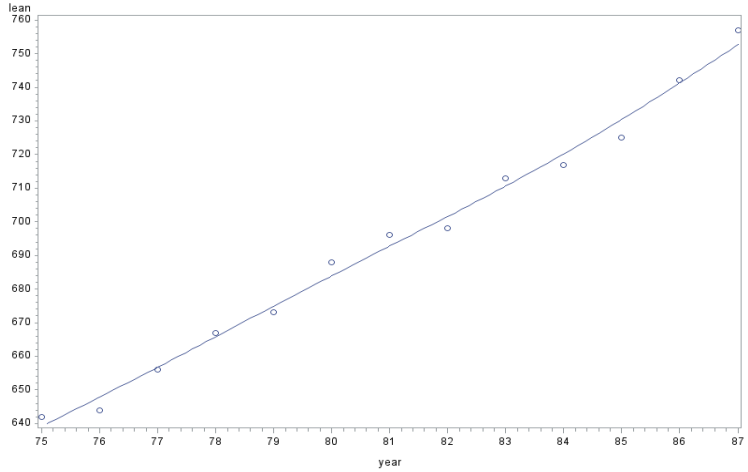
Obs	year	lean
1	75	642
2	76	644
3	77	656
4	78	667
5	79	673
6	80	688
7	81	696
8	82	698
9	83	713
10	84	717
11	85	725
12	86	742
13	87	757
14	113	.

Step 1: Study the relationship

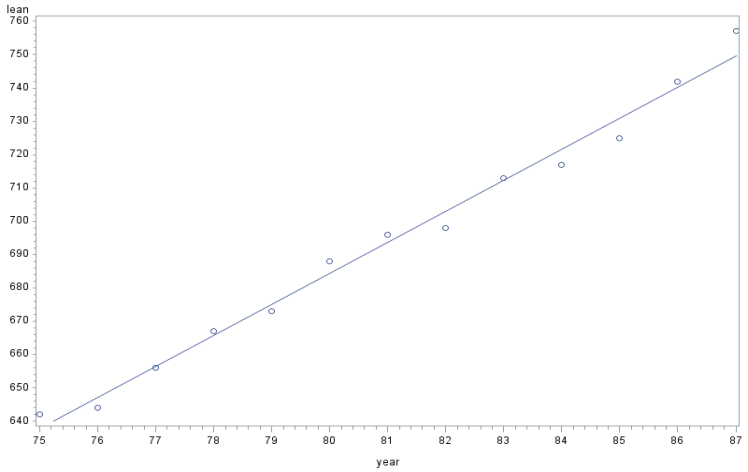
Should always plot first!!!!!

```
data a1; input year lean @@;
cards;
75 642 76 644 77 656 78 667 79 673 80 688
81 696 82 698 83 713 84 717 85 725 86 742
87 757 102 .
;
data a1p; set a1; if lean ne .;

symbol1 v=circle i=sm70;
proc gplot data=a1p; plot lean*year;

symbol1 v=circle i=rl;
proc gplot data=a1p; plot lean*year;
run;
```


What is the Trend?



Linear Trend?



Straight Line Equation

- Straight line describes smoothed curve well
- Formula for a straight line

$$Y = \beta_0 + \beta_1 X$$

β_0 is the intercept

β_1 in the slope

- Need to **estimate** β_0 and β_1
- Will use method of **least squares**

SAS Proc Reg

```
proc reg data=a1;
   model lean=year/clb p r;
   output out=a2  p=pred r=resid;
   id year;

proc gplot data=a2;
   plot resid*year/ vref=0;
   where lean ne .;
run;
```



```
The REG Procedure
                  Analysis of Variance
                               Sum of          Mean
Source             DF       Squares        Square  F Value  Pr > F
Model               1         15804         15804   904.12  <.0001
Error              11     192.28571      17.48052
Corrected Total    12         15997

Root MSE         4.18097    R-Square     0.9880
Dependent Mean 693.69231    Adj R-Sq     0.9869
Coeff Var        0.60271

                  Parameter     Standard
Variable   DF      Estimate        Error  t Value   Pr > |t|
Intercept   1     -61.12088     25.12982    -2.43     0.0333
year        1      9.31868      0.30991    30.07     <.0001

             Parameter Estimates
Variable     DF       95% Confidence Limits
Intercept     1     -116.43124       -5.81052
year          1        8.63656       10.00080
```

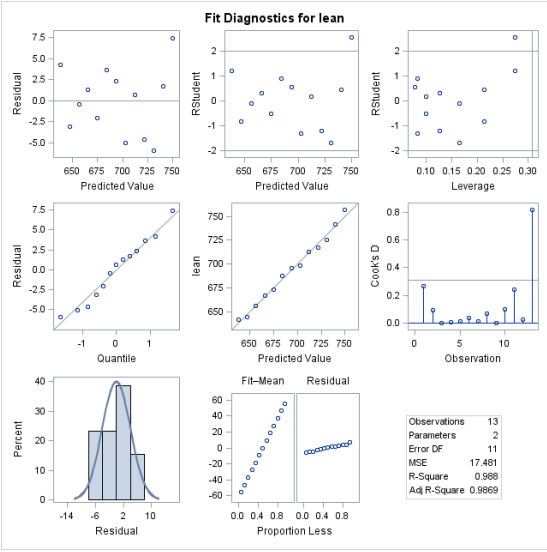
Output Statistics

Obs	year	Dep Var lean	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual
1	75	642.0000	637.7802	2.1914	4.2198	3.561
2	76	644.0000	647.0989	1.9354	-3.0989	3.706
3	77	656.0000	656.4176	1.6975	-0.4176	3.821
4	78	667.0000	665.7363	1.4863	1.2637	3.908
5	79	673.0000	675.0549	1.3149	-2.0549	3.969
6	80	688.0000	684.3736	1.2003	3.6264	4.005
7	81	696.0000	693.6923	1.1596	2.3077	4.017
8	82	698.0000	703.0110	1.2003	-5.0110	4.005
9	83	713.0000	712.3297	1.3149	0.6703	3.969
10	84	717.0000	721.6484	1.4863	-4.6484	3.908
11	85	725.0000	730.9670	1.6975	-5.9670	3.821
12	86	742.0000	740.2857	1.9354	1.7143	3.706
13	87	757.0000	749.6044	2.1914	7.3956	3.561
14	113	.	991.8901	9.9848	.	.

Obs	year	Student Residual	-2-1 0 1 2	Cook's D
1	75	1.185	|** |	0.266
2	76	-0.836	*| |	0.095
3	77	-0.109	| |	0.001
4	78	0.323	| |	0.008
5	79	-0.518	*| |	0.015
6	80	0.905	|* |	0.037
7	81	0.574	|* |	0.014
8	82	-1.251	**| |	0.070
9	83	0.169	| |	0.002
10	84	-1.189	**| |	0.102
11	85	-1.562	***| |	0.241
12	86	0.463	| |	0.029
13	87	2.077	|**** |	0.817
14	102	.		.

Sum of Residuals	0
Sum of Squared Residuals	192.28571
Predicted Residual SS (PRESS)	297.29196

Model Fit Diagnostics



Background Reading

- Appendix A : Review?
- KNNL Chapters 1 and 2
- SAS template file pisa.sas