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 $\sqrt{\text{function}}$
  - 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 $X$
- 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 explanatory variable
- Goals:
  - To **characterize** lean over time
  - To **predict** future observations

The Data Set

<table>
<thead>
<tr>
<th>Obs</th>
<th>Year</th>
<th>Lean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>75</td>
<td>642</td>
</tr>
<tr>
<td>2</td>
<td>76</td>
<td>644</td>
</tr>
<tr>
<td>3</td>
<td>77</td>
<td>656</td>
</tr>
<tr>
<td>4</td>
<td>78</td>
<td>667</td>
</tr>
<tr>
<td>5</td>
<td>79</td>
<td>673</td>
</tr>
<tr>
<td>6</td>
<td>80</td>
<td>688</td>
</tr>
<tr>
<td>7</td>
<td>81</td>
<td>696</td>
</tr>
<tr>
<td>8</td>
<td>82</td>
<td>698</td>
</tr>
<tr>
<td>9</td>
<td>83</td>
<td>713</td>
</tr>
<tr>
<td>10</td>
<td>84</td>
<td>717</td>
</tr>
<tr>
<td>11</td>
<td>85</td>
<td>725</td>
</tr>
<tr>
<td>12</td>
<td>86</td>
<td>742</td>
</tr>
<tr>
<td>13</td>
<td>87</td>
<td>757</td>
</tr>
<tr>
<td>14</td>
<td>113</td>
<td>.</td>
</tr>
</tbody>
</table>

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?

```
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;
```

Straight Line Equation

- Straight line describes smoothed curve well
- Formula for a straight line
  \[ Y = \beta_0 + \beta_1 X \]
  \( \beta_0 \) is the intercept
  \( \beta_1 \) in the slope
- Need to estimate \( \beta_0 \) and \( \beta_1 \)
- Will use method of least squares

The REG Procedure

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>15804</td>
<td>15804</td>
<td>904.12</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>11</td>
<td>192.28571</td>
<td>17.48052</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>12</td>
<td>15997</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

Parameter Estimates

| Variable | DF | Estimate  | Standard 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

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-116.43124 -5.81052</td>
</tr>
<tr>
<td>year</td>
<td>1</td>
<td>8.63656 10.00080</td>
</tr>
</tbody>
</table>
### Output Statistics

<table>
<thead>
<tr>
<th>Obs year</th>
<th>Dep Var</th>
<th>Predicted Value</th>
<th>Mean Predict</th>
<th>Residual Value</th>
<th>Mean Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>75</td>
<td>642.0000</td>
<td>637.7802</td>
<td>2.1914</td>
<td>4.2198</td>
</tr>
<tr>
<td>2</td>
<td>76</td>
<td>644.0000</td>
<td>647.9989</td>
<td>1.9354</td>
<td>-3.0989</td>
</tr>
<tr>
<td>3</td>
<td>77</td>
<td>656.0000</td>
<td>656.4176</td>
<td>1.6975</td>
<td>-0.4176</td>
</tr>
<tr>
<td>4</td>
<td>78</td>
<td>667.0000</td>
<td>665.7363</td>
<td>1.4863</td>
<td>1.2637</td>
</tr>
<tr>
<td>5</td>
<td>79</td>
<td>673.0000</td>
<td>675.0549</td>
<td>1.3149</td>
<td>-2.0549</td>
</tr>
<tr>
<td>6</td>
<td>80</td>
<td>688.0000</td>
<td>684.3736</td>
<td>1.2003</td>
<td>3.6264</td>
</tr>
<tr>
<td>7</td>
<td>81</td>
<td>696.0000</td>
<td>693.6923</td>
<td>1.1596</td>
<td>2.3077</td>
</tr>
<tr>
<td>8</td>
<td>82</td>
<td>698.0000</td>
<td>703.0110</td>
<td>1.2003</td>
<td>-5.0110</td>
</tr>
<tr>
<td>9</td>
<td>83</td>
<td>713.0000</td>
<td>712.3297</td>
<td>1.3149</td>
<td>0.6703</td>
</tr>
<tr>
<td>10</td>
<td>84</td>
<td>717.0000</td>
<td>721.6484</td>
<td>1.4863</td>
<td>-4.6484</td>
</tr>
<tr>
<td>11</td>
<td>85</td>
<td>725.0000</td>
<td>730.9670</td>
<td>1.6975</td>
<td>-5.9670</td>
</tr>
<tr>
<td>12</td>
<td>86</td>
<td>742.0000</td>
<td>740.2857</td>
<td>1.9354</td>
<td>1.7143</td>
</tr>
<tr>
<td>13</td>
<td>87</td>
<td>757.0000</td>
<td>749.6044</td>
<td>2.1914</td>
<td>7.3956</td>
</tr>
<tr>
<td>14</td>
<td>113</td>
<td>991.8901</td>
<td>9.8848</td>
<td>.</td>
<td></td>
</tr>
</tbody>
</table>

### Student Residuals

<table>
<thead>
<tr>
<th>Obs year</th>
<th>Residual</th>
<th>Student Cook's D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.185</td>
<td>**</td>
</tr>
<tr>
<td>2</td>
<td>-0.836</td>
<td>*</td>
</tr>
<tr>
<td>3</td>
<td>-0.109</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.323</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-0.518</td>
<td>*</td>
</tr>
<tr>
<td>6</td>
<td>0.905</td>
<td>*</td>
</tr>
<tr>
<td>7</td>
<td>0.574</td>
<td>*</td>
</tr>
<tr>
<td>8</td>
<td>-1.251</td>
<td>**</td>
</tr>
<tr>
<td>9</td>
<td>0.169</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>-1.189</td>
<td>**</td>
</tr>
<tr>
<td>11</td>
<td>-1.562</td>
<td>***</td>
</tr>
<tr>
<td>12</td>
<td>0.463</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>2.077</td>
<td>****</td>
</tr>
</tbody>
</table>

### Model Fit Diagnostics

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

### Background Reading

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