Lecture 1: Class Logistics/SAS
Outline

- Overview
- Class Information and Policies
- SAS Software/Example
- Background Reading
Overview

- We will cover:
  - Simple linear regression (SLR) – Chapters 1-5
  - Multiple linear regression (MLR) – Chapters 6-11
  - Analysis of variance (ANOVA) – Chapters 16-25
Overview

- The emphasis will be placed on selected practical tools using SAS rather than on the mathematical manipulations.

- Want to understand the theory so that we can apply it appropriately.
Class Website

http://www.stat.purdue.edu/~ghobbs/STAT_512/stat512.htm

- Course syllabus / schedule
- Lecture notes
- Homework assignments
- Sample SAS programs
- Data sets
- Announcements
- Link to Blackboard Vista
Class Policies

Class Time
- 8:30-9:20 am, MWF, REC 121
- Please arrive on time and stay the duration of the lecture.

Texts:
- Required - *Applied Linear Statistical Models, 5th edition*, by Kutner, Nachtsheim, Neter, and Li
Class Policies

- Communication
  - Office hours: (MATH 510)
    - Mon 1:30 - 2:30
    - Wed 1:30 - 2:30
  - Email: ghobbs@purdue.edu – put STAT 512 in subject line
  - Announcements posted on web page
  - Note: It may be difficult to talk with me right before or after class, so please primarily use the above mechanisms for communication.
Lecture Notes
- Available on website – please print yourself
- Usually (hopefully) prepared a week in advance
- Not comprehensive (Be prepared to take notes).
- One/two chapters per week
- Ask questions if you’re confused
Class Policies

Grades

- 25% Exam I (Tentatively week of February 21)
- 25% Exam II (Tentatively week of March 28)
- 20% Final Exam (Set by University)
- 30% Homework

Exam I and II will be evening exams. Two classes will be cancelled to compensate.

Once the exam dates are set, please notify me at least a week in advance if there is a conflict.
Class Policies

- Homework Assignments
  - Generally one per week – assigned on Fri
  - Will be due *beginning* of class the following Fri
  - Assignments posted on the website
  - Can discuss with others, but solutions must be your own
  - Guidelines for homework/re-grades in syllabus
  - No late homework accepted
  - Drop lowest two homeworks
  - 30% of grade
SAS Software

SAS (Statistical Analysis System) is the program we will use to perform data analysis for this class. Learning to use SAS will be a large part of the course.

- Available on all ITaP computers
- Installation on personal computers FREE for Purdue faculty, staff, and students
  - STEW G65 (Contracts and Licensing Office)
  - Take your Purdue ID

- ITaP Software Remote:
SAS Software

- Getting Help with SAS – try yourself
  - SAS Help Files and Online Doc.
  - World Wide Web (look up the syntax in your favorite search engine)
  - Introductory Tutorials (on class website)
  - The Recommended Text
  - SAS Files on class website
SAS Software

Getting Help with SAS – ask for help

- Statistical Consulting Service - Software Help Desk
  - Math G175 Hours 10-4 M through F
  - [http://www.stat.purdue.edu/scs/](http://www.stat.purdue.edu/scs/)

- Wednesday Evening Help Sessions
  - Help with SAS for multiple Stat courses
  - Staffed with graduate student TA
  - More info will be given as it becomes available

- Your Instructor
I will often give examples from SAS in class. The programs used in lecture (and any other programs you should need) will be available for you to download from the website.

I will usually have to edit the output somewhat to get it to fit in the notes. You should run the SAS programs yourself to see the real output and experiment with changing the commands to learn how they work.

I will tell you the names of all SAS files I use in these notes. If the notes differ from the SAS file, take the SAS file to be correct, since there may be cut-and-paste errors.
There is a tutorial in SAS to help you get started.
  - Help/Getting Started with SAS Software

You should spend some time before next week getting comfortable with SAS.

For today, don’t worry about the detailed syntax of the commands. Just try to get a sense of what is going on.
**SAS Example:**
Price Analysis for Diamond Rings in Singapore

**Variables**
- *Response variable* – price in Singapore dollars ($Y$)
- *Explanatory variable* – weight of diamond in carats ($X$)

**Goals**
- Create a scatterplot
- Fit a regression line
- Predict the price of a sale for a 0.43 carat diamond ring
SAS Data Step

- File diamond.sas on website.
- One way to input data in SAS is to type or paste it in. In this case, we have a sequence of ordered pairs (weight, price).

```sas
DATA diamonds;
  input weight price @@;
datalines;
.17 355 .16 328 .17 350 .18 325 .25 642 .16 342 .15 322 .19 485
.21 483 .15 323 .18 462 .28 823 .16 336 .20 498 .23 595 .29 860
.12 223 .26 663 .25 750 .27 720 .18 468 .16 345 .17 352 .16 332
.17 353 .18 438 .17 318 .18 419 .17 346 .15 315 .17 350 .32 918
.32 919 .15 298 .16 339 .16 338 .23 595 .23 553 .17 345 .33 945
.25 655 .35 1086 .18 443 .25 678 .25 675 .15 287 .26 693 .15 316
.43 .
;
data diamonds1;
set diamonds;
if price ne .; 

**Syntax Notes**

- Each line must end with a semi-colon.
- There is no output from this statement, but information does appear in the log window.
- Often you will obtain data from an existing SAS file or import it from another file, such as a spreadsheet. Examples showing how to do this will come later.
Now we want to see what the data look like.

```sas
proc print data=diamonds;
run;
```

<table>
<thead>
<tr>
<th>Obs</th>
<th>weight</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.17</td>
<td>355</td>
</tr>
<tr>
<td>2</td>
<td>0.16</td>
<td>328</td>
</tr>
<tr>
<td>3</td>
<td>0.17</td>
<td>350</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>47</td>
<td>0.26</td>
<td>693</td>
</tr>
<tr>
<td>48</td>
<td>0.15</td>
<td>316</td>
</tr>
<tr>
<td>49</td>
<td>0.43</td>
<td></td>
</tr>
</tbody>
</table>
We want to plot the data as a scatterplot, using circles to represent data points and adding a smoothing curve to see if it looks linear.

The symbol statement “v=circle” (v stands for “value”) lets us do this.

The symbol statement “i=sm70” will add a smooth line using splines (interpolation = smooth). These are options which stay on until you turn them off.

In order for the smoothing to work properly we need to sort the data by the X variable.
**SAS Proc Gplot**

```sas
proc sort data=diamonds1;
by weight;

symbol1 v=circle i=sm70;
title1 'Diamond Ring Price Study';
title2 'Scatter plot of Price vs. Weight with Smoothing Curve';
axis1 label=('Weight (Carats)');
axis2 label=(angle=90 'Price (Singapore $$)');
proc gplot data=diamonds1;
plot price*weight / haxis=axis1 vaxis=axis2;
run;
```
Diamond Ring Price Study

Scatter plot of Price vs. Weight with Smoothing Curve
Now we want to use simple linear regression to fit a line through the data. We use the symbol option “i=rl” meaning “interpolation = regression line” (that’s an L not a one).

symbol1 v=circle i=rl;

title2 'Scatter plot of Price vs. Weight with Regression Line';

proc gplot data=diamonds1;
plot price*weight / haxis=axis1 vaxis=axis2;
run;
Diamond Ring Price Study

Scatter plot of Price vs. Weight with Regression Line
We use proc reg(regression) to estimate a regression line and calculate predictors and residuals from the straight line.

We tell it what the data are, what the model is, and what options we want.

```sas
proc reg data=diamonds;
  model price=weight/clb p r;
  output out=diag p=pred r=resid;
  id weight;
run;
```
## Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>2098596</td>
<td>2098596</td>
<td>2069.99</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>46</td>
<td>46636</td>
<td>1013.81886</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>47</td>
<td>2145232</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Root MSE
31.84052

### R-Square
0.9783

### Dependent Mean
500.08333

### Adj R-Sq
0.9778

### Coeff Var
6.36704

## Parameter Estimates

| Variable | DF | Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|----------|----------------|---------|------|---|
| Intercept| 1  | -259.62591| 17.31886       | -14.99  | <.0001 |
| weight   | 1  | 3721.02485| 81.78588       | 45.50   | <.0001 |
**proc print data=diag;**

**run;**

<table>
<thead>
<tr>
<th>Obs</th>
<th>weight</th>
<th>price</th>
<th>pred</th>
<th>resid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.17</td>
<td>355</td>
<td>372.95</td>
<td>-17.9483</td>
</tr>
<tr>
<td>2</td>
<td>0.16</td>
<td>328</td>
<td>335.74</td>
<td>-7.7381</td>
</tr>
<tr>
<td>3</td>
<td>0.17</td>
<td>350</td>
<td>372.95</td>
<td>-22.9483</td>
</tr>
<tr>
<td>4</td>
<td>0.18</td>
<td>325</td>
<td>410.16</td>
<td>-85.1586</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>46</td>
<td>0.15</td>
<td>287</td>
<td>298.53</td>
<td>-11.5278</td>
</tr>
<tr>
<td>47</td>
<td>0.26</td>
<td>693</td>
<td>707.84</td>
<td>-14.8406</td>
</tr>
<tr>
<td>48</td>
<td>0.15</td>
<td>316</td>
<td>298.53</td>
<td>17.4722</td>
</tr>
<tr>
<td>49</td>
<td>0.43</td>
<td>.</td>
<td>1340.41</td>
<td>.</td>
</tr>
</tbody>
</table>
Background Reading

- Start Reading Chapter 1