I. INTRODUCTION

A. DEFINITIONS

1. **Regression** of \( Y \) on \( X_1 \) and \( X_2 \): \( \mu(Y|X_1,X_2) \)
   
   \[
   \mu(Y|X_1,X_2) = \text{“the mean of } Y \text{ as a function of } X_1 \text{ and } X_2\text{”}
   \]

2. **Regression model**: a formula to approximate \( \mu(Y|X_1,X_2) \)
   
   Example: \( \mu(Y|X_1,X_2) = \beta_0 + \beta_1X_1 + \beta_2X_2 \)

   \[\text{Response Variable} \quad \text{Explanatory Variables} \quad \text{Regression coefficients (unknown parameters)}\]

3. **Linear regression model**: a regression model linear in \( \beta \)s

4. **Regression analysis**: tools for answering questions via regression models
B. Example: sex discrimination in salaries

1. Data set:

<table>
<thead>
<tr>
<th>salary</th>
<th>yrs.exper</th>
<th>sex</th>
<th>male.ind</th>
</tr>
</thead>
<tbody>
<tr>
<td>20000</td>
<td>5</td>
<td>male</td>
<td>1</td>
</tr>
<tr>
<td>18500</td>
<td>3</td>
<td>female</td>
<td>0</td>
</tr>
<tr>
<td>24500</td>
<td>10</td>
<td>female</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

2. Possible model: \( \mu(\text{salary}| \text{male.ind}, \text{yrs.exper}) = \beta_0 + \beta_1 \text{male.ind} + \beta_2 \text{yrs.exper} \)

At any particular level of experience, the mean salary for males is $\beta_1$ more than the mean for females.
C. Some purposes of regression analysis

1. Examine a relationship after accounting for other variables (as in example B)

2. Prediction of future Y’s at some values of X (example: predict sales price of house from sq. footage, etc.)

3. Get adjusted values (example: compute statewide SAT scores, adjusted for percent of seniors who take exam)

4. Test a theory (example: mean recession velocity of galaxies = $H \times \text{distance} + qH^2\text{distance}^2$)

5. Find “important” X’s for predicting Y (use with caution)

6. Find a value of X to maximize the mean of Y
D. ST 412/512 Topics

1. Regression modeling
   a. Indicators variables for categorical variables (like sex)
   b. Product terms for interaction
   c. Polynomial terms for curvature ($\beta_0 + \beta_1 X + \beta_2 X^2$)
   d. Variable selection (choosing a useful set of X’s)
   e. Model checking and remedies for various problems

2. Inferential tools (various tests and conf. intervals)

3. Analysis of variance (all categorical X’s)

4. Simple time series, multivariate responses, and repeated measures
E. Review of Simple Linear Regression

1. Model with constant variance:
\[ \mu(Y|X) = \beta_0 + \beta_1 X; \quad \text{var}(Y|X) = \sigma^2 \]

2. *Least Squares*: choose estimators \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) to minimize the
sum of squared residuals,
\[ \sum_{i=1}^{n} \left( Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right)^2 \]

3. Solution (calculus):
\[ \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \]
4. Properties of least squares estimators (Display 7.7)

5. Summary of robustness and resistance of least squares:
   
a. The “constant variance” assumption is important

b. Normality is not too important for confidence intervals and p-values, but is important for prediction intervals

c. Long-tailed distributions and/or outliers can heavily influence the results

d. Non-independence problems: serial correlation (Ch. 15) and cluster effects (we deal with this in Ch. 9-14)

6. Strategy for dealing with these potential problems

   a. Plots; residual plots; Consider outliers (more in Ch. 11)

   b. Transform (especially log) (see Display 8.6)
7. Gas chromatography example

5 samples of material at each of 4 known amounts

Goal: use chromatograph on future samples with unknown amounts of the substance

LS line \( R^2 = .9995 \)
(1) Curvature!
(2) Increasing variance with increasing mean!

=> try LOG(Y)
This didn’t straighten the relationship; try taking log of X?
$R^2 = .998$
RESIDUAL PLOT

Fitted Values

Problems?
8. Questions for discussion

a. How do we decide which is Y and which is X here?

b. How can $R^2$ be so large for such a poor fitting model?

c. Why are the residual plots such a big help here?

d. Will a lack-of-fit test (Section 8.5) help resolve the adequacy of the fit on p. 10 and 11 above?

e. What are possible explanations for the problems?

f. What other information might be useful?

g. What inferential tool would we like to use (on a good fitting model)?

h. Any ideas for next steps?
9. **SAS output (for fit shown on page 10 above)**

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>68.82878</td>
<td>68.82878</td>
<td>10816.7</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>18</td>
<td>0.11454</td>
<td>0.00636</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>19</td>
<td>68.94332</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE: 0.07977
R-Square: 0.9983
Dependent Mean: 4.37741
Adj R-Sq: 0.9982
Coeff Var: 1.82231

Parameter Estimates

| Parameter         | DF  | Estimate | Standard Error | t Value | Pr > |t| | Standardized Estimate |
|-------------------|-----|----------|----------------|---------|-------|----------|-----------------------|
| Intercept         | 1   | 3.47291  | 0.01984        | 175.01  | <.0001| 0        | 0                     |
| amount_log        | 1   | 1.12399  | 0.01081        | 104.00  | <.0001| 0.99917  |                       |

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>3.43122    3.51460</td>
</tr>
<tr>
<td>amount_log</td>
<td>1</td>
<td>1.10128    1.14669</td>
</tr>
</tbody>
</table>

See Section 8.5.2

\( \sigma^2 \)
10. Lack-of-fit F-test (Section 8.5.3)

a. Idea: compare size of residuals (reg and 1-way ANOVA)

Reduced: simple reg.  
Full: 4 separate means

SS(res) = .11454 (18 df)  
SS(res) = .05422 (16 df)

b. Details for extra-sum-of-squares F-test

\[ F - stat = \frac{(0.11454 - 0.05422)/2}{0.05422/16} = 8.9 \]
11. Output for fit to a quadratic model: $\mu(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2$

Model: MODEL1
Dependent Variable: response_log

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>2</td>
<td>68.83125</td>
<td>34.41563</td>
<td>5220.79</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>17</td>
<td>0.11206</td>
<td>0.00659</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>19</td>
<td>68.94332</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE: 0.08119  R-Square: 0.9984
Dependent Mean: 4.37741  Adj R-Sq: 0.9982
Coeff Var: 1.85478

Parameter Estimates

| Variable        | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|-----------------|----|--------------------|----------------|---------|-------|---|
| Intercept       | 1  | 3.46179            | 0.02716        | 127.47  | <.0001|
| amount_log      | 1  | 1.11537            | 0.01786        | 62.45   | <.0001|
| x2              | 1  | 0.00536            | 0.00874        | 0.61    | 0.5483|
12. Output for fit to cubic model: $\mu(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$

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>3</td>
<td>68.88909</td>
<td>22.96303</td>
<td>6776.06</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>16</td>
<td>0.05422</td>
<td>0.00339</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>19</td>
<td>68.94332</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 0.05821
R-Square 0.9992
Dependent Mean 4.37741
Adj R-Sq 0.9991
Coeff Var 1.32987

Parameter Estimates

| Variable | DF | Estimate | Standard Error | t Value | Pr > |t| | Standardized Estimate |
|----------|----|----------|----------------|---------|-------|---|-----------------------|
| Intercept| 1  | 3.39040  | 0.02603        | 130.23  | <.0001| 0 | 0                     |
| x        | 1  | 1.12008  | 0.01286        | 87.12   | <.0001| 0.99569   | 0.11256               |
| x2       | 1  | 0.06198  | 0.01507        | 4.11    | 0.0008| 0.11004   | -0.11004              |
| x3       | 1  | -0.00066010 | 0.00015978    | -4.13   | 0.0008| -0.11004  | -0.11004              |

$H_0: \beta_3 = 0$

Small p-value => poor fit of reduced model

a. The *sample correlation coefficient* between two variables $X$ and $Y$ is

$$r_{XY} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})/(n-1)}{s_{X}s_{Y}}$$

b. $r_{XY}$ summarizes the degree of *linear* association of $X,Y$

c. It doesn’t depend on one variable being a “response”

d. $-1 \leq r_{XY} \leq 1$

e. Fact: in simple linear reg., $R^2 = r_{XY}^2$

f. Typically: $r_{XY}$ is a useful summary; regression is better for answering questions of interest
II. Multiple Regression Models

A. Data

1. Y  X₁  X₂  X₃

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>34</td>
<td>15</td>
<td>-37</td>
<td>3.331</td>
</tr>
<tr>
<td>24</td>
<td>18</td>
<td>59</td>
<td>1.111</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

B. Linear regression models (p.231)

1. Model with 2 X’s: \( \mu(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \)

2. Ex: Y: 1st year GPA, X₁: Math SAT, X₂: Verbal SAT

3. Ex: Y = log(tree volume), X₁: log(height), X₂: log(diameter)
C. Important notes about interpretation of $\beta$’s

1. Causal conclusions can be made from randomized experiments but not from observational studies!!!!

2. Geometrically, $\beta_0 + \beta_1 X_1 + \beta_2 X_2$ describes a plane. For a fixed value of $X_1$ the mean of $Y$ changes by $\beta_2$ for each one-unit increase in $X_2$.

3. The meaning of a coefficient depends on what explanatory variables are included!!!! $\beta_2$ in $\mu(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ is not the same as $\beta_2$ in $\mu(Y|X_1, X_2, X_3) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$.

4. Constant variance assumption. If $\text{var}(Y|X_1, X_2) = \sigma^2$, a focus on the mean, $\mu(Y|X_1, X_2)$, is better justified.
D. Special constructed explanatory variables

1. **Polynomial terms**, eg $X^2$, for curvature (see p. 15, 16 above)

2. **Indicator variables** to model effects of categorical variables
   
   a. One indicator variable ($X=0,1$) to distinguish 2 groups; 
      Ex: $X = 1$ for females, 0 for males (see p. 2 above)
   
   b. $(K-1)$ indicator variables to distinguish $K$ groups; Ex: 
      $X_2 = 1$ if fertilizer B was used, 0 if A or C was used
      $X_3 = 1$ if fertilizer C was used, 0 if A or B was used

3. **Product terms** for interaction
   
   $\mu(Y|X_1,X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2)$
   
   So $\mu(Y|X_1,X_2=7) = (\beta_0 + \beta_2 7) + (\beta_1 + \beta_3 7)X_1$
   
   $\mu(Y|X_1,X_2=9) = (\beta_0 + \beta_2 9) + (\beta_1 + \beta_3 9)X_1$

   The effect of $X_1$ on $\mu(Y)$ depends on the value of $X_2$
E. Example with several indicator variables (fake)

1. Data: \( Y = \) Post-training score, \( X = \) Pre-training score, Method = code for method of training (A, B, or C)

<table>
<thead>
<tr>
<th>Y</th>
<th>X</th>
<th>Method</th>
<th>( I_A )</th>
<th>( I_B )</th>
<th>( I_C )</th>
</tr>
</thead>
<tbody>
<tr>
<td>53</td>
<td>49</td>
<td>A</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>52</td>
<td>47</td>
<td>B</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>71</td>
<td>63</td>
<td>C</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>59</td>
<td>61</td>
<td>A</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Note: The Final Exam is on WEDNESDAY June 7 at 9:30. The syllabus incorrectly listed the day as Tuesday.
2. Coded scatterplot of Y versus X:

Post Training Test Score Vs. Pre-training test score
for 30 subjects trained with methods A, B, or C
3. Model: \( \mu(Y|X, I_A, I_B, I_C) = \beta_0 + \beta_1 X + \beta_2 I_B + \beta_3 I_C \)

So:

\[
\begin{align*}
\mu(Y|X, I_A=1, I_B=0, I_C=0) &= \beta_0 + \beta_1 X \\
\mu(Y|X, I_A=0, I_B=1, I_C=0) &= \beta_0 + \beta_2 + \beta_1 X \\
\mu(Y|X, I_A=0, I_B=0, I_C=1) &= \beta_0 + \beta_3 + \beta_1 X
\end{align*}
\]
4. Estimates of coefficients

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Value</th>
<th>Std. Error</th>
<th>t value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>4.7808</td>
<td>3.7141</td>
<td>1.2872</td>
<td>0.2094</td>
</tr>
<tr>
<td>x</td>
<td>0.9243</td>
<td>0.0686</td>
<td>13.4677</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ib</td>
<td>3.7269</td>
<td>1.6242</td>
<td>2.2947</td>
<td>0.0301</td>
</tr>
<tr>
<td>ic</td>
<td>9.0430</td>
<td>1.6436</td>
<td>5.5021</td>
<td>&lt;0.0000</td>
</tr>
</tbody>
</table>

5. Results

a. For individuals of similar initial ability, the mean test score for those trained with method B is an estimated 3.7 points higher than that for method A (95% confidence interval 0.5 to 7.0).

b. Those with training method C have a mean that is an estimated 9.0 points higher those taught with method A, after accounting for initial test score (95% confidence interval to 5.8 to 12.3)
c. The adjusted C mean is an estimated 5.3 points higher than the adjusted B mean.

6. Different parameterization for the same model

\[ \mu(Y|X,I_A,I_B,I_C) = \beta_0 + \beta_1 X + \beta_2 I_A + \beta_3 I_C \]

Same fit.
Same residuals.
Same est. of \( \sigma \).
Just different names for the 3 intercepts.
The estimate of $\beta_3$ is 5.3 with SE 1.6. So a 95% confidence interval for the amount by which the mean for method C exceeds that for B is 2.8 to 8.5.

7. Another possible parameterization (drop intercept):
   \[
   \mu(Y|X,I_A,I_B,I_C) = \beta_1 X + \beta_2 I_A + \beta_3 I_B + \beta_4 I_C
   \]

F. Notes about indicator variables

1. A t-test for $H_0: \beta_1=0$ in the regression of $Y$ on a single indicator variable $I_B$, $\mu(Y|I_B) = \beta_0 + \beta_2 I_B$ is the 2-sample t-test

2. Regression when all explanatory variables are categorical is “analysis of variance”

3. Regression with categorical variables and one numerical $X$ is often called “analysis of covariance”
G. Short-hand notation for models

1. Write $\mu(Y|x_1, \text{CAT}) = x_1 + \text{CAT}$

   - Write the variable name in lower case when it is to be treated as a numerical variable
   - Upper case refers to a variable that is to be treated as categorical

   i.e. $\mu(Y|x_1, \text{CAT}) = \beta_0 + \beta_1 x_1 + (\beta_2 I_B + \beta_3 I_C)$

   - Indicator variables to distinguish the levels of the categorical variable
2. Ex: $\mu(\text{flowers}|\text{light, TIME}) = \text{light} + \text{TIME}$  
stands for 
$\mu(\text{flowers}|\text{light,TIME}) = \beta_0 + \beta_1 \text{flowers} + (\beta_2 I_{\text{time}=24})$

3. To include interaction terms,
write $\mu(Y|x_1,x_2) = x_1 + x_2 + x_1 \times x_2$ to represent  
$\mu(Y|x_1,x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$

or $\mu(Y|x_1,\text{CAT}) = x_1 + \text{CAT} + x_1 \times \text{CAT}$ to represent  
$\beta_0 + \beta_1 x_1 + (\beta_2 I_B + \beta_3 I_C) + (\beta_4 x_1 I_B + \beta_5 x_1 I_C)$
III. Strategies and Graphical Tools for Model Exploration

A. Strategy for data analysis with models (p. 241)

B. Graphical tools for exploration and communication

1. Matrix of scatterplots (9.5.1)
   
   a. SAS ANALYST -- can’t be done
   
   b. SAS INSIGHT
      -> Solutions -> Analysis -> Interactive Data Analysis
      Select Library, Select Data Set, Click on Open
      -> Analyze -> Multivariate (YX)
      For each variable, click on the variable name, then on Y
(and enter the response variable last)
Click on “Output”, put a check next to “Matrix of Scatterplots” (and uncheck everything else). Click OK.

c. S+
Import the data set of interest (From -> File -> Import)
Highlight the variables of interest; click on the matrix of scatterplots icon (make sure View -> toolbars -> 2d plots is selected)

2. Coded scatterplot (9.5.2)

a. SAS ANALYST
Graphs -> Scatter Plot. Select “Two-Dimensional”
X-Axis: INTENS, Y-AXIS: FLOWERS, class: TIME
(i.e. choose for the class variable the column that contains the codes)
b. SAS INSIGHT
   -> Edit -> Observations -> Label in Plots
   Select the coded variable to use as a label. Then click and drag so all possible codes are highlighted. Click OK.
   -> Analyze -> Scatter Plot (YX). Select the Y-axis and X-axis variables and selected the coded variable for “label.” Click OK. Click “Apply.”

c. S+
   Click on “Text as Symbols” and follow directions, or click on “Scatter/line plot” include a z variable with codes for symbols, click on the tab “Vary Symbols by” and follow directions
C. Aside: Study on 300 children born prematurely

1. \[ Y = \text{IQ at age 8} \]
   \[ X_1 = \text{mother’s social class (1,2,...,5)} \]
   \[ X_2 = \text{mother’s education (1,2,...,5)} \]
   \[ X_3 = 1 \text{ if female, 0 if male} \]
   \[ X_4 = \# \text{ days of ventilation} \]
   \[ X_5 = 1 \text{ if given breast milk as infant, 0 if not} \]

2. Results (p. 249)

<table>
<thead>
<tr>
<th></th>
<th>Est. coef.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>social class</td>
<td>-3.5</td>
<td>.0004</td>
</tr>
<tr>
<td>mother’s ed</td>
<td>2.0</td>
<td>.01</td>
</tr>
<tr>
<td>female ind.</td>
<td>4.2</td>
<td>.01</td>
</tr>
<tr>
<td>days ventilation</td>
<td>-2.6</td>
<td>.02</td>
</tr>
<tr>
<td>breast milk ind.</td>
<td>8.3</td>
<td>&lt; .0001</td>
</tr>
</tbody>
</table>
IV. Inferential Tools

A. Topics

1. Least squares estimators of multiple regression coefficients
2. t-tests and confidence intervals for individual coefficients
3. t-tests and CIs for linear combinations of $\beta$’s
4. Prediction intervals
5. Extra Sum of Squares F-test [Main topic of Ch. 10]

Ex:

Full model: $\beta_0 + \beta_1 x_1 + \beta_2 I_B + \beta_3 I_C$

Reduced model: $\beta_0 + \beta_1 x_1$

Note on homework problem 10.23. If you get confused, retain the model with different intercepts and different slopes in order to answer the questions of interest, even though the coefficients may not be significantly different from zero.
B. Introduction to Bat Echolocation Data (p. 258)

1. Q: Do echolocating bats expend more energy than non-echolocating bats and birds, after accounting for size?

2. Strategy
   
   a. Explore (resolve need for transformation)
   
   b. Test for interaction
   
   c. If no interaction, answer question with the three parallel lines model
3. Review: 4 parameterizations for the 3 parallel lines model
\[ \mu(y|x,\text{TYPE}) = x + \text{TYPE} \]

- **a.** \[ \beta_0 + \beta_1 x_1 + \beta_2 I_{\text{bird}} + \beta_3 I_{\text{ebat}} \] (p. 260)
- **b.** \[ \beta_0 + \beta_1 x_1 + \beta_2 I_{\text{bat}} + \beta_3 I_{\text{bird}} \] (SAS default)
- **c.** \[ \beta_0 + \beta_1 x_1 + \beta_2 I_{\text{bat}} + \beta_3 I_{\text{ebat}} \]
- **d.** \[ \beta_1 x_1 + \beta_2 I_{\text{bat}} + \beta_3 I_{\text{bird}} + \beta_4 I_{\text{ebat}} \] (drop intercept)
C. Least Squares

1. \( \mu(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2; \) \( \text{var}(Y|x_1, x_2) = \sigma^2 \)

   **Unknown parameters:**
   - **Regression coefficients**
   - **Variance (about regression)**

2. **Fitted (or predicted) values, \( \hat{y}_i \):**
   \[ \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} \quad \text{for } i = 1, 2, ..., n \]

3. **Residuals, \( \text{res}_i = y_i - \hat{y}_i \)**

4. **Least squares estimators, \( (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) \), chosen to minimize the sum of squared residuals (matrix algebra formula)**

5. \( \sigma^2 = \text{(sum of squared residuals)}/(n-p) \quad [p=\text{number of } \beta\text{'s}] \)
D. t-tests and confidence intervals for individual $\beta$’s

1. Note: a matrix algebra formula for $SE(\hat{\beta_j})$ is also available

2. If distribution of $Y$ given $X$’s is normal, then

$$t - ratio = \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)}$$

has a $t$-distribution on $n-p$ degrees of freedom

3. For testing the hypothesis $H_0: \beta_2 = 7$; compare

$$t - stat = \frac{\hat{\beta}_2 - 7}{SE(\hat{\beta}_2)}$$

to a $t$-distribution on $n-p$ degrees of freedom. Reasonable?
4. The p-value for the test of $H_0: \beta_j = 0$ is standard output

5. It’s often useful to think of $H_0: \beta_2 = 0$ (for example) as
   
   Full model: $\mu(y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
   
   Reduced model: $\beta_0 + \beta_1 x_1 + \beta_3 x_3$

   (Is the $\beta_2 x_2$ term needed in a model with the other x’s?)

6. 95% confidence interval for $\beta_j$:

   $$\hat{\beta}_j \pm t_{n-p}(.975) \times SE(\hat{\beta}_j)$$

7. The meaning of a coefficient (and its significance) depends on what other X’s are in the model (p. 262).

8. The t-based inference works well even without normality.
E. t-tests and CIs for Bat Data

1. From the output in Display 10.6:
   The data are consistent with the hypothesis of no energy differences between echolocating and non-echolocating bats, after accounting for body size (2-sided p-value = .7)

2. But that doesn’t prove that there is no difference. A “large” p-value means either:
   
   (i) there is no difference ($H_0$ is true) or

   (ii) there is a difference and this study is not powerful enough to detect it.
3. So: report a confidence interval in addition to the p-value.  
95% CI for \( \beta_3 \): \( 0.0787 \pm 2.12 \times 0.2027 = (-0.35, 0.51) \).

Interpretation (back-transform, \( e^{0.0787} = 1.08 \), \( e^{-0.35} = 0.70 \) and \( e^{0.51} = 1.67 \)):

It is estimated that the median energy expenditure for echolocating bats is 1.08 times the median for non-echolocating bats of the same body weight (95% confidence interval: 0.70 to 1.67 times).

4. What about a test and CI for \( \beta_3 - \beta_2 \)? Do either:

a. Refit with parameterization (c) (p.35 above) [easiest], or

b. Compute the SE as in Display 10.15.
F. Review

1. Model: \( \mu(y|x_1,x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2; \) \( \text{var}(y|x_1,x_2)=\sigma^2 \)

2. Special terms: (a) (K-1) indicator variables for K levels of a categorical factor, (b) polynomial terms for curvature, (c) cross-product terms for interaction

3. Least squares: \( (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) \)

4. Fitted (predicted) values: \( \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} \)

5. Residuals: \( res_i = y_i - \hat{y}_i \)

6. Estimator of \( \sigma^2 = \) mean square of residuals (d.f. = n-p)
G. Review of "variance"

1. \( \text{var}(y) = \text{Mean}\{(y-\mu)^2\} \) (population variance)

2. \( \text{SD}(y) = \{\text{var}(y)\}^{1/2} \)

3. \( \text{var}(y|x) = \text{Mean}\{[y-\mu(y|x)]^2\} \) in subpopulation of y’s at x

4. \( \text{var}(\hat{\beta}) = \text{Mean}\{(\hat{\beta} - \beta)^2\} \) (sampling variance)

5. \( \text{SD}(\hat{\beta}) = \{\text{var}(\hat{\beta})\}^{1/2} \)

6. \( \text{SE}(\hat{\beta}) = \text{Estimate of} \ \text{SD}(\hat{\beta}), \text{usually obtained by using} \ \hat{\sigma} \) in place of the unknown \( \sigma \) (with associated d.f.)
H. The main set of inferential tools

1. t-tests and confidence intervals for individual $\beta$’s

2. t-tests and confidence intervals for linear combinations of $\beta$’s (ex: $\beta_1 - \beta_2$)

3. Confidence intervals for $\mu(y|x_1=x_1^*,x_2=x_2^*)$—giving a range of values of $\mu(y|x_1=x_1^*,x_2=x_2^*)$ consistent with the data

4. Prediction intervals—giving a range of plausible values of $y$ when $x_1=x_1^*$ and $x_2=x_2^*$

5. Extra SS F-test for testing whether several $\beta$’s are all 0
I. Reminder about interpolation and extrapolation

We can safely make statements about the distribution of $y$ given $x$-values within the sampled range.

Extrapolation involves additional speculation.

1. We can safely make statements about the distribution of $y$ given $x$-values within the sampled range.

2. Extrapolation involves additional speculation.
J. Inference about linear combinations of $\beta$’s

1. Sometimes we need a test or CI for $C_0\beta_0 + C_1\beta_1 + C_2\beta_2$ for constants $C_0, C_1, C_2$.

2. Examples:
   a. $\beta_1 - \beta_2$
   b. $\beta_0 + \beta_1 x_1^* + \beta_2 x_2^*$ (the mean of $y$ at $x_1=x_1^*$ and $x_2=x_2^*$)

3. A formula for $\text{var}(C_0\hat{\beta}_0 + C_1\hat{\beta}_1 + C_2\hat{\beta}_2)$ is on p. 278

4. Computer trick is sometimes useful: redefine reference level
   a. Change the set of indicator variables (for factor levels)
   b. Re-center the $x$’s (for 2b above)
5. Trick for getting SE, test, and CI for $\beta_0 + \beta_1 x_1^* + \beta_2 x_2^*$

a. Transform to get a new set of x’s by subtracting off the value where inference is desired
   
   $\text{newx}_1 = x_1 - x_1^*$
   $\text{newx}_2 = x_2 - x_2^*$

b. Note: $\mu(y|\text{newx}_1, \text{newx}_2) = \beta_0^* + \beta_1^* \text{newx}_1 + \beta_2^* \text{newx}_2$
   where $\beta_0^*$ represents $\mu(y|\text{newx}_1=0, \text{newx}_2=0)$
   i.e. $\beta_0^*$ represents $\mu(y| x_1 - x_1^* = 0, x_2 - x_2^* = 0)$
   i.e. $\beta_0^*$ represents $\mu(y| x_1 = x_1^*, x_2 = x_2^*)$

c. So, fit the regression of y on $\text{newx}_1$ and $\text{newx}_2$ and carry out inference about the intercept
K. Prediction and prediction intervals

1. \( \text{var}_{\text{pred}}(y|x_1=x_1^*, x_2=x_2^*) = \text{var}\{\hat{\mu}(y | x_1 = x_1^*, x_2 = x_2^*)\} + \sigma^2 \)

2. \( \text{SE}(\text{pred}) = \{\text{var}_{\text{pred}}(y|x_1=x_1^*, x_2=x_2^*)\}^{1/2} \)

3. 95% Prediction interval:
   \[ \hat{\mu}(y | x_1 = x_1^*, x_2 = x_2^*) \pm t_{n-p}(.975)SE(\text{pred}) \]

4. Typically: Supply program (SAS?) with \( x^* \)s and prediction level (95%) and it will produce intervals
L. Extra SS F-test

1. Full model: \( \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \)
   Reduced model: \( \beta_0 + \beta_1 x_1 \)

   i.e. \( H_0: \beta_2 = \beta_3 = 0 \)

2. Extra sum of squares = (Sum of squared residuals from reduced model) - (Sum of squared residuals from full model)

3. F-statistic = \([\text{Extra SS}/\text{Extra # of } \beta\text{'s}] / \hat{\sigma}_{full}^2\)
M. Special Cases of Extra SS F-test

1. F-test for overall significance of regression

   Full model: $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

   Reduced model: $\beta_0$

   Asks: are *any* of the x’s useful predictors of y?
   The calculations are laid out in the ANOVA table

2. F-test for a single coefficient

   If full model has one more term than the reduced model, the F-test is equivalent to the (2-sided) t-test

3. For $\mu(y|\text{CAT}) = \text{CAT}$ (i.e. a single categorical factor), the F-test for overall significance is the one-way ANOVA F-test
N. Miscellaneous notes from Chapter 10

1. $R^2$ can help, somewhat, with practical significance (10.4.1)
   
   a. $R^2$ from model with $x$, $x^2$, and $x^3$: .9994
   
   b. $R^2$ from model with $x$, $x^2$: .9903
   
   c. So $x^3$, although significant, explains only .91% more of the variation in $y$

2. $R^2$ can always be made 100% by adding enough terms
3. Formula for variance of linear combinations (10.4.3)

   a. Example: linear comb. of 2 coefficients

   \[ SE(C_1 \hat{\beta}_1 + C_2 \hat{\beta}_2) = \left[ C_1^2 SE(\hat{\beta}_1)^2 + C_2^2 SE(\hat{\beta}_2)^2 + 2C_1C_2 \hat{\text{cov}}(\hat{\beta}_1, \hat{\beta}_2) \right]^{1/2} \]

   b. Note—covariance: \( \text{cov}(Y_1, Y_2) = \text{mean}\{(Y_1-\mu_1)(Y_2-\mu_2)\} \)

   c. The estimated covariance matrix of \( \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots \)
      is available upon request in a regression program
      (See Display 10.15 for an example)

4. Principle of Occam’s Razor (10.4.5): prefer simple models

5. Informal tests in model exploration phase (10.4.6)

   Be aware that p-values are sample-size dependent
6. Estimating the $x$ that maximizes or minimizes the mean of $y$, when $\mu(y|x) = \beta_0 + \beta_1 x + \beta_2 x^2$ (from calculus)

\[ x_{\text{max}} = -\frac{\beta_1}{2\beta_2} \]

a. This is a maximum if $\beta_2 < 0$ and a minimum if $\beta_2 > 0$

b. Approximately (for future reference):

\[
SE[-\hat{\beta}_1/(2\hat{\beta}_2)] = \frac{[SE(\hat{\beta}_1)]^2}{4\hat{\beta}_2^2} + \frac{\hat{\beta}_1^2[SE(\hat{\beta}_2)]^2}{4\hat{\beta}_2^4} - \frac{\hat{\beta}_1[\text{cov}(\hat{\beta}_1, \hat{\beta}_2)]}{2\hat{\beta}_2^3}
\]
O. Example: predicting college enrollment

1. roll=enrollment at University of New Mexico

[Scatter plots showing year, unemployment rate, highschool graduates, and college enrollment.]
2. SAS output

Summary of Fit
Mean of Response        12707.0345  R-Square        0.9702
Root MSE                  594.2032  Adj R-Sq        0.9667

Analysis of Variance
Source     DF  Sum of Squares  Mean Square      F Stat  Pr > F
Model       3    287665541.9707  95888513.99      271.58  <.0001
Error      25    8826934.9949      353077.3998         .     .
C Total    28   296492476.9655        .             .     .

Parameter Estimates
Pr
Variable     DF    Estimate   Std Error      t Stat    >|t|        Tolerance  Var Inflation
Intercept     1  -2380.3968   1096.1569       -2.17    0.0396       .                 0.0000
year          1    191.9126     19.0115       10.09     <.0001      0.4812         2.0781
unem          1    308.5925    108.6785        2.84      0.0088      0.8463         1.1816
hgrad         1      0.5946          0.0520       11.43     <.0001      0.5439         1.8385

3. Predict enrollment at year = 31 (the “next year”) if unemployment is 10% and number of HS grads is 17000
a. Could use re-centering trick on page 46 above to get

\[ SE\{\mu(y|x_1=31,x_2=10,x_3=17000)\} \]

i.e. transform to get

newx1 = \(x_1 - 31\)

newx2 = \(x_2 - 10\)

newx3 = \(x_3 - 17000\)

---

### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Stat</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>287665541.9707</td>
<td>95888513.99</td>
<td>271.58</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>25</td>
<td>8826934.9949</td>
<td>353077.3998</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>C Total</td>
<td>28</td>
<td>296492476.9655</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

### Parameter Estimates

| Variable   | DF | Estimate   | Std Error | t Stat  | >|t| | Tolerance | Var Inflation |
|------------|----|------------|-----------|---------|----|-------------------|----------------|
| Intercept  | 1  | 16762.6363 | 323.5834  | 51.80   | <.0001 | .                 | 0.0000         |
| newx1      | 1  | 191.9126   | 19.0115   | 10.09   | <.0001 | 0.4812            | 2.0781         |
| newx2      | 1  | 308.5925   | 108.6785  | 2.84    | 0.0088 | 0.8463            | 1.1816         |
| newx3      | 1  | 0.5946     | 0.0520    | 11.43   | <.0001 | 0.5439            | 1.8385         |

b. and use formula on p. 47 to get \(SE_{\text{pred}}\)
c. Or, easier (in ANALYST) just ask for predictions
(1) Create a new data set with variable names year, unem, hgrad; and list the values of these variables at which you want to predict. (2) Fit the regression of roll on year, unem, hgrad; select “PREDICTIONS”, specify that you want to create predictions for a separate data file.

<table>
<thead>
<tr>
<th>Obs</th>
<th>year</th>
<th>roll</th>
<th>unem</th>
<th>hgrad</th>
<th>inc</th>
<th>Predicted roll</th>
<th>Lower prediction limit of roll</th>
<th>Upper prediction limit of roll</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31</td>
<td>.</td>
<td>10</td>
<td>17000</td>
<td>.</td>
<td>16762.64</td>
<td>16096.20</td>
<td>17429.07</td>
</tr>
</tbody>
</table>

4. Problems with this analysis

a. Extrapolation to future years (requires speculation)

b. Residuals correlated with time (Need Ch. 15 tools)
IV. Model Checking and Refinement

A. Introduction

1. The problem: least squares is NOT resistant; one or several observations can have undue influence

Example:
A quadratic-in-x term is significant here, but not when largest x is removed.

2. Why is this a problem?
3. Tools
   a. Scatterplots of data and of residuals
   b. Tentative fits of models with one or more cases set aside
   c. Tools to help detect outliers and influential cases (11.4)
   d. Strategy for dealing with influential observations (11.3)

4. Difficulties to overcome:
   a. Detection of outliers and influential observations depends on having the right scale for y (a good transformation) first, but assessment of the need for a transformation may be influenced by outliers and influential observations
b. Detection of outliers and influential observations can depend on having the appropriate x’s in the model, but assessment of appropriate x’s can be affected by influential observations

5. General strategy:
   a. Start with a fairly rich model; i.e. tend to include x’s even if you’re not sure they will be in the final model (but be careful about this with small sample sizes)
   
   b. Resolve influence and transformation simultaneously, early in the data analysis
   
   c. In complicated problems, be prepared for deadends
B. Influence

1. By *influential observation(s)* we mean one or several observations whose removal causes a different conclusion or course of action (subjective assessment)

2. Two strategies

   a. Use an estimating procedure that is more resistant than least squares (and don’t worry about the problem)

   b. Use least squares with the strategy in Display 11.8
C. Case Influence Statistics

1. Introduction

   a. These help identify influential observations and help to clarify the course of action in Display 11.8
   b. Use them when (a) you suspect influence problems and when (b) graphical displays may not be adequate
   c. One useful set:

   \[ D_i \] (Cook’s Distance) for measuring influence
   \[ h_i \] (leverage) for measuring “unusualness” of x’s
   \[ r_i \] (Studentized residual) for measuring “outlierness”
   \[ i = 1,2,..., n \]
2. Leverage: $h_i$ (also called: diagonal element of the hat matrix)

   a. It measures the multivariate distance between the x’s for case $i$ and the average x’s, accounting for the correlation structure (see Display 11.10)

   b. If there is only one x, $h_i = \frac{1}{(n-1)} \left( \frac{x_i - \bar{x}}{s_x} \right)^2 + \frac{1}{n}$

   c. For several x’s, $h_i$ has a matrix expression

   ![Scatter plot with comment: This case has a relatively large leverage]
3. Studentized residual for detecting outliers (in y direction)

a. Formula: \( \text{studres}_i = \frac{\text{res}_i}{\text{SE}(\text{res}_i)} \)

b. Fact: \( \text{SE}(\text{res}_i) = \hat{\sigma} \sqrt{1 - h_i} \)

i.e. different residuals have different variances, and since \( 0 < h_i < 1 \) those with largest \( h_i \) (unusual x’s) have the smallest \( \text{SE}(\text{res}_i) \)

c. For outlier detection use this type of residual
   (but use ordinary residuals in the standard residual plots)
4. $D_i$: Cook’s Distance for identifying influential cases

a. One formula: 
$$D_i = \sum_{j=1}^{n} \frac{(\hat{y}_{j(i)} - \hat{y}_j)^2}{p\hat{\sigma}^2}$$

where $\hat{y}_{j(i)}$ is the estimated mean of $y$ at observation $j$, based on the reduced data set with observation $i$ deleted.

b. Equivalent formula (admittedly mysteriously here):
$$D_i = \frac{1}{p} (\text{studres}_i)^2 \left( \frac{h_i}{1-h_i} \right)$$

This term is big if case $i$ is unusual in the $y$-direction.

This term is big if case $i$ is unusual in the $x$-direction.
5. The upshot

a. Get the triplet \((D_i, h_i, \text{studres}_i)\) for each \(i\) from 1 to \(n\)

b. Look to see whether any \(D_i\)’s are “large”

Large \(D_i\)’s indicate influential observations (Note: you ARE allowed to investigate these more closely by manual case deletion)

c. \(h_i\) and \(\text{studres}_i\) help explain the reason for influence (unusual x-value or outlier or both); which helps in deciding the course of action in the strategy of Display 11.8
6. ROUGH guidelines for “large”

   a. $D_i$ values near or larger than 1 are good indications of influential cases; sometimes a $D_i$ much larger than the others in the data set is worth looking at

   b. The average of $h_i$ is always $p/n$; some people suggest using $h_i > 2p/n$ as “large”

   c. Based on normality, $|\text{studres}| > 2$ is considered “large”

   **KEY WORD: ROUGH**
7. Sample situations with a single $x$

- **Large $h_i$**
  - Moderate studres$_i$
  - Large $D_i$

- **Moderate $h_i$**
  - Large studres$_i$
  - Large $D_i$

- **Small $h_i$**
  - Large studres$_i$
  - Small $D_i$
8. Alternative case influence statistics

   a. Alternative to $D_i$: $\text{dffits}_i$ (and others)

   b. Alternative to $\text{studres}_i$: externally-studentized residual

   c. Suggestion: use whatever is convenient with the statistical computer package you’re using

9. Note: $D_i$ only detects influence of single-cases; influential pairs may go undetected
D. Partial Residual Plots

1. We would like to visually explore the function $f(x_2)$ in
   
   \[ \mu(y \mid x_1, x_2) = \beta_0 + \beta_1 x_1 + f(x_2) \]

2. That is, we’d like to plot $y$ versus $x_2$ but with the effect of $x_1$ subtracted out; i.e. plot $y - \beta_0 - \beta_1 x_1$ versus $x_2$

3. To approximate this, get the partial residual for $x_2$:

   a. Get $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ in \( \mu(y \mid x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \)

   b. Compute the partial residual as $\text{pres} = y - \hat{\beta}_0 - \hat{\beta}_1 x_1$

4. This is also called a component plus residual; if $\text{res}$ is the residual from 3a, $\text{pres} = \text{res} + \hat{\beta}_2 x_2$ (easier computation)
E. Weighted regression for certain types of non-constant variance

1. Suppose \( \mu(y | x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \), \( \text{var}(y | x_1, x_2) = \sigma^2 / w_i \)

   and the \( w_i \)'s are known

2. *Weighted least squares* is the appropriate tool for this model; it minimizes the weighted sum of squared residuals

\[
\sum_{i=1}^{n} w_i (y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2
\]

3. In statistical computer programs: use linear regression in the usual way, specify the column \( w \) as a weight, read the output in the usual way
4. Important special cases where this is useful

a. $y_i$ is an average based on a sample of size $m_i$
   In this case, the weights are $w_i = 1/m_i$

b. the variance is proportional to $x$;
   so $w_i = 1/x_i$
F. Measurement errors in x’s

1. Fact: least squares estimates are biased and inferences about $\mu(y \mid x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ can be misleading if the available data for estimating the regression are observations $y, x_1, x_2^*$, where $x_2^*$ is an imprecise measurement of $x_2$ (even though it may be an unbiased measurement).

2. This is an important problem to be aware of; general purpose solutions do not exist in standard statistical programs.

3. Exception: if the purpose of the regression is to predict future $y$’s from future values of $x_1$ and $x_2^*$ then there is no need to worry that $x_2^*$ is an imprecise measurement.
V. Variable Selection

A. Introduction

1. The problem: we may be faced with a fairly large number of potential explanatory variables (5, 10, 20, 50)

2. Two good reasons for seeking a subset:
   a. General principle: smaller is better (Occam’s razor)
   b. Unnecessary terms add imprecision to inferences

3. Computer assisted tools
   a. Comparison of (Cp, AIC, or BIC criteria) in all subsets
   b. Stepwise regression (search along favorable directions)

4. But don’t expect a single best model!
B. Objectives when there are many X’s (12.2.1)

1. Adjustment
   
a. Ex: Do males receive higher salaries than females, after accounting for ...

   b. Strategy: first find a good set of X’s to account for; then see if the sex indicator is significant when added in

2. Fishing for association; i.e. what are the important X’s?
   
a. The trouble with this: we can find several subsets of X’s that are useful predictors of Y; but that doesn’t imply importance or causation

   b. Best attitude: use this for hypothesis generation

3. Prediction (this is a straightforward objective)
C. Loss of precision due to multicollinearity


\[
\frac{\sigma^2}{(n-1)s^2_x}
\]

Variance about the regression

Sample variance of x

2. Fact: variance of L.S. estimator of coef. of \( X_j \) in mult. reg. =

\[
\frac{\sigma^2}{(n-1)s^2_j(1 - R^2_j)}
\]

Sample variance of \( x_j \)

\( R^2 \) in the regression of \( x_j \) on the other x’s in model

3. So variance of an estimated coef. will tend to be larger if there are other X’s in the model that can predict \( x_j \)
4. This is also true for other inferences; for example: the variance of prediction will tend to be larger if there are unnecessary or redundant x’s in the model.

5. **Multicollinearity**: the situation in which \( s_j^2(1 - R_j^2) \) is small for one or more j’s.

6. **Variance inflation factor** for \( x_j \): \( \frac{1}{(1 - R_j^2)} \) is sometimes used to identify multicollinearity problems (see SAS output).

7. But, this doesn’t help too much; e.g. two highly correlated x’s may still both be needed in the model.

8. “Good” subsets of x’s: (a) lead to a small \( \hat{\sigma}^2 \) (b) with as few x’s as possible (Criteria \( C_p \), AIC, and BIC formalize this).
D. Strategy for dealing with many x’s

1. Identify objectives; identify relevant set of x’s

2. Exploration: matrix of scatterplots; correlation matrix; residual plots after fitting tentative models

3. Transformation? Influence?

4. Computer-assisted variable selection

5. Proceed with analysis

E. Sequential variable selection

1. Forward selection

   a. Start with no x’s “in” the model
b. Find the “most significant” additional x

c. If its p-value is less than some cutoff (like .05) add it to the model (and re-fit the model with the new set of x’s)

d. Repeat (b) and (c) until no further x’s can be added

2. Backward elimination

a. Start with all x’s “in” the model

b. Find the “least significant” of the x’s currently in the model

c. If it’s p-value is greater than some cutoff (like .05) drop it from the model (and re-fit with the remaining x’s)

d. Repeat until no further x’s can be dropped

3. Stepwise regression
a. Start with no x’s “in”
b. Do one step of forward selection
c. Do one step of backward elimination
d. Repeat (b) and (c) until no further x’s can be added or dropped

4. Notes

a. Add and drop factor indicator variables as a group

b. Don’t take p-values and CI’s for selected variables seriously—because of serious data snooping (not a problem for objectives 1 and 3)

c. A drawback: the product is a single model. This is deceptive. Think not: “here is the best model.” Think instead: “here is one, possibly useful model.”
F. Criteria for comparing models

Criterion to minimize

\[ f(\hat{\sigma}^2) + g(p) \]

Cp

\[ \frac{(n - p)(\hat{\sigma}^2 - \hat{\sigma}_{full}^2)}{\hat{\sigma}_{full}^2} + p \]

BIC

\[ n \log(\hat{\sigma}^2) + p \log(n) \]

AIC

\[ n \log(\hat{\sigma}^2) + 2p \]

Idea: favor models with small mean square of residuals but penalize for too many x’s
1. The proposed criteria: Mallow’s Cp Statistic, Schwarz’s Bayesian Information Criterion (BIC, or SBC in SAS), and Akaike’s Information Criterion (AIC)

2. The idea behind these is the same, but the theory for arriving at the tradeoff between small $\hat{\sigma}^2$ and small $p$ differs

3. My opinion: there’s no way to truly say that one of these criteria is better than the others

4. Computer programs: Fit all possible models; report the best 10 or so according to the selected criteria

5. Note: one other criteria: $\hat{\sigma}^2 + 0$ (sometimes used; but isn’t as good). An equivalent criterion is $-R^2_{\text{adjusted}}$
G. Cross Validation (12.6.4)

1. If tests, CIs, or prediction intervals are needed after variable selection and if n is large, maybe try:

2. Cross validation
   a. Randomly divide the data into 75% for model construction and 25% for inference
   b. Perform variable selection with the 75%
   c. Refit the same model (don’t drop or add anything) on the remaining 25% and proceed with inference
H. Review


2. Fitted values, residuals, least squares method of estimation

3. Properties of least squares; tests and confidence intervals for individual coefficients; prediction intervals; extra SS F-tests (full and reduced models)

4. Model building and refinement: transformation, indicator variables, $x^2$, interaction, variable selection

5. Influence and case-influence statistics
6. A note on the difference between “confounding variable” and “interaction”

a. Is there an association between gestation and mean brain weight after accounting for body weight?

\[ \mu\{\text{brain}\} = \beta_0 + \beta_1 \text{body} + \beta_2 \text{gest} \]

(\(\beta_2\) represents the association of gestation with mean brain weight after accounting for body weight.)

b. Is the association between gestation and brain weight different for animals of different body sizes?

\[ \mu\{\text{brain}\} = \beta_0 + \beta_1 \text{body} + \beta_2 \text{gest} + \beta_3 \text{body} \times \text{gest} \]

(There is an interactive association of body and gest)
I. Example: sex discrimination data

1. Matrix of scatterplots; response= beginning salary (BSAL)
2. Residual plots; BSAL regressed on all 5 X’s

\[
y = \text{Beginning salary}
\]

\[
y = \text{Log beginning salary}
\]

3. Correlation matrix

<table>
<thead>
<tr>
<th></th>
<th>FSEX</th>
<th>SENIOR</th>
<th>AGE</th>
<th>EDUC</th>
<th>EXPER</th>
<th>LN.BSAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSEX</td>
<td>1.00</td>
<td>-0.10</td>
<td>0.26</td>
<td>-0.33</td>
<td>-0.02</td>
<td>-0.54</td>
</tr>
<tr>
<td>SENIOR</td>
<td>-0.10</td>
<td>1.00</td>
<td>-0.18</td>
<td>0.06</td>
<td>-0.07</td>
<td>-0.29</td>
</tr>
<tr>
<td>AGE</td>
<td>0.26</td>
<td>-0.18</td>
<td>1.00</td>
<td>-0.23</td>
<td>0.80</td>
<td>0.06</td>
</tr>
<tr>
<td>EDUC</td>
<td>-0.33</td>
<td>0.06</td>
<td>-0.23</td>
<td>1.00</td>
<td>-0.10</td>
<td>0.41</td>
</tr>
<tr>
<td>EXPER</td>
<td>-0.02</td>
<td>-0.07</td>
<td>0.80</td>
<td>-0.10</td>
<td>1.00</td>
<td>0.19</td>
</tr>
<tr>
<td>LN.BSAL</td>
<td>-0.54</td>
<td>-0.29</td>
<td>0.06</td>
<td>0.41</td>
<td>0.19</td>
<td>1.00</td>
</tr>
</tbody>
</table>
4. Coded scatterplots (triangles represent females)

5. Various approaches to model fitting and testing of sex effect
   a. Straightforward: do manual backward elimination, withholding the sex indicator until the end
STEP 1:

|               | Value   | Std. Error | t value | Pr(|t|) |
|---------------|---------|------------|---------|--------|
| (Intercept)   | 8.6580  | 0.1389     | 62.3380 | 0.0000 |
| SENIOR        | -0.0041 | 0.0011     | -3.6315 | 0.0005 |
| AGE           | -0.0002 | 0.0001     | -1.1735 | 0.2438 |
| EDUC          | 0.0239  | 0.0051     | 4.6415  | 0.0000 |
| EXPER         | 0.0005  | 0.0002     | 2.3651  | 0.0202 |

STEP 2 (drop AGE):

|               | Value   | Std. Error | t value | Pr(|t|) |
|---------------|---------|------------|---------|--------|
| (Intercept)   | 8.5596  | 0.1109     | 77.1498 | 0.0000 |
| SENIOR        | -0.0038 | 0.0011     | -3.4589 | 0.0008 |
| EDUC          | 0.0253  | 0.0050     | 5.0495  | 0.0000 |
| EXPER         | 0.0003  | 0.0001     | 2.3784  | 0.0195 |

STOP BACKWARD ELIMINATION; NOW ADD IN SEX INDICATOR VARIABLE

|               | Value     | Std. Error | t value | Pr(|t|) | Note: exp(-.1296) = .88 |
|---------------|-----------|------------|---------|--------|--------------------------|
| (Intercept)   | 8.8006    | 0.1018     | 86.4431 | 0.0000 |
| SENIOR        | -0.0043   | 0.0009     | -4.6057 | 0.0000 |
| EDUC          | 0.0164    | 0.0045     | 3.6623  | 0.0004 |
| EXPER         | 0.0003    | 0.0001     | 2.4510  | 0.0162 |
| FSEX          | -0.1296   | 0.0214     | -6.0600 | 0.0000 |

b. Try stepwise regression, with FSEX forced into the model. (For these data, this leads to same model as a.)
c. Select model (without FSEX) using Cp (or AIC or BIC) (For these data, this leads to same model as a.)

d. Note: a key point is that even when all these methods do not lead to the same model, they do lead to approximately the same coefficient for FSEX when added into the model.

e. Section 12.5 does these analyses but with quadratic and interaction terms; but result is nearly identical

6. Bayesian Model Averaging (advanced topic; currently in vogue in some disciplines)

a. Idea: Get a statistic of interest (like the coeff of FSEX) from lots of models (with different other X’s) and take a weighted average of these
b. The Bayesian weighting using BIC:

\[ statistic_{averaged} = \sum_{k=1}^{K} w_k \times statistic_k \]

where \( k \) represents the particular model out of \( K \) different ones considered; and \( w_k \) is the weight attached to that model, given by:

\[ w_k = \exp(-BIC_k) / \sum_{k=1}^{K} \exp(-BIC_k) \]

i.e. small BIC (good model) implies large weight.

c. Note: this results from the formula on p. 345 when there is no reason before hand to prefer any particular model; i.e. when the subjective probability associated with model \( k \) is \( \Pr(M_k) = 1/K \)

d. Answer for coef. of FSEX: -.1206; note: \( \exp(-.1206) = .89 \)
J. Aside on Bayesian Statistics

1. Philosophical distinction:
   a. Frequentist statistics: parameters (like $\beta$’s) are unknown constants
   b. Bayesian statistics: parameters are random variables, meaning we have probability distributions to express our subjective belief about their values

2. Bayesian terminology
   a. Prior distribution: a probability distribution describing belief in the parameter, prior to seeing the data; $f_{\text{prior}}(\beta)$
   b. Posterior distribution: probability distribution describing the parameter after seeing the data $f_{\text{post}}(\beta|\text{data})$
3. Bayes law: \( f_{\text{posterior}}(\beta|\text{data}) \propto f_{\text{prior}}(\beta)f_{\text{likelihood}}(\text{data}|\beta) \), tells how to update knowledge about the parameter

4. This instructor’s opinion about Bayesian statistics
   a. Philosophical interpretation is nice; but not a big deal in practice
   b. The difficulty in specifying a prior is the big drawback
   c. Bayesian analysis is THE way to go if you truly have prior information
   d. Because of computational advances (called Markov Chain Monte Carlo) there has been a resurgence of interest in Bayesian analysis
   e. Some scientists mistakenly believe more information can be extracted from data with Bayesian methods.
K. Additional review

1. What about all those F-tests?

   a. **All** F-tests we’ve considered are special cases of the extra sum of squares F-test (Sect. 10.3)

   b. **F-test for overall significance of regression**
      Full: a model of interest
      Reduced: model with $\beta_0$ only

   c. **F-test for lack-of fit**
      Full: one-way anova (separate means for each distinct combination of x’s)
      Reduced: a model of interest

   d. **Partial F-test** is an F-test for a single $\beta$
e. *One-way ANOVA F-test*
   Full: model with a separate mean for each group
   i.e. $\beta_0$ and $k-1$ indicators to distinguish $k$ groups
   Reduced: $\beta_0$ only (single mean model)

f. SAS *Type III F-tests*
   Full: model that has been specified
   Reduced: model without a particular term

g. SAS and S+ *Sequential F-tests*
   i. Full: intercept and $x_1$
      Reduced: intercept
   ii. Full: intercept, $x_1$, and $x_2$
      Reduced: intercept and $x_1$
   iii. Full: intercept, $x_1$, $x_2$, and $x_3$
      Reduced: intercept, $x_1$, and $x_2
2. In “linear regression,” what does “linear in β’s” mean?
   a. \( \beta_0 \times \text{something} + \beta_1 \times \text{something} + \beta_2 \times \text{something} + \ldots \)
   b. Ex. of nonlinear regression: \( \mu(y|x) = \beta_0 x^{\beta_1} \)

3. A note about “mean response.” It is useful to explicitly write \( \mu(y|x_1, x_2, x_3) \) to talk about the mean of y as a function of \( x_1, x_2, \) and \( x_3 \). Sometimes we abbreviate this to “the mean of the response” if it’s clear what x’s we’re talking about.

4. Partial residuals
   a. You may find a plot of partial residuals vs. \( x_1 \) to be useful when it is desired to study the relationship between y and \( x_1 \), after getting the effects of \( x_2, x_3, \) etc. out of the
way, especially if the effect of $x_1$ is relatively small (in which case the plot of $y$ versus $x_1$ does not reveal much).

b. For example: How is mammal brain weight related to litter size, after accounting for body weight?

c. Suppose $\mu(y|x_1,x_2) = \beta_0 + \beta_1x_1 + \beta_2x_2$. A plot of $y$ versus $x_1$ won’t show a linear relationship whose slope is $\beta_1$ if $x_1$ and $x_2$ are correlated. However, a plot of $y - (\beta_0 + \beta_2x_2)$ versus $x_1$ will show a pattern whose slope is $\beta_1$.

d. So, the partial residuals are $y_i - (\hat{\beta}_0 + \hat{\beta}_2x_{2i})$, where the $\beta$’s are the estimates from the regression of $y$ on $x_1$ and $x_2$. 
L. Review of course topics

1. Ch. 9: Regression model, special terms
2. Ch. 10: Inferential tools, extra SS F-test
3. Ch. 11: Case influence statistics
4. Ch. 12: Variable selection
5. Ch. 13: 2-way ANOVA (with replicates)
6. Ch. 14: 2-way and higher-way ANOVA, without replicates
7. Ch. 15: Time series
8. Ch. 16: Repeated measures
VI. Two-way classifications

A. Introduction

1. Regression model example (short-hand notation): $\mu(y) = x_1 + \text{FACTOR}_a + x_2 + \text{FACTOR}_b + x_1 \text{FACTOR}_a + x_1 x_2$

2. *Analysis of variance*: regression analysis with only factors

3. Notes:
   a. ANOVA / regression distinction isn’t too important
   b. In ANOVA, F-tests and ANOVA tables are central (why?)
   c. *2-way ANOVA* has 2 factors, 3-way ANOVA has ..
d. **Additive** 2-way model: $\mu(y) = \text{FACTOR}_a + \text{FACTOR}_b$

**Non-additive**: $\text{FACTOR}_a + \text{FACTOR}_b + \text{FACTOR}_a \times \text{FACTOR}_b$
4. The *two-way classification* or *two-way layout*:

<table>
<thead>
<tr>
<th>ROW FACTOR</th>
<th>COLUMN FACTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>C1</td>
</tr>
<tr>
<td>R2</td>
<td>Y1</td>
</tr>
<tr>
<td>R3</td>
<td>Y2</td>
</tr>
<tr>
<td>R2</td>
<td>Y9</td>
</tr>
<tr>
<td>R3</td>
<td>Y10</td>
</tr>
<tr>
<td>R3</td>
<td>Y17</td>
</tr>
<tr>
<td>R3</td>
<td>Y18</td>
</tr>
</tbody>
</table>
5. Important experimental design terms

a. *Experimental unit*: the object to which the treatment is applied

b. *Treatment*: the condition or manipulation that is applied to the units

c. *Block*: a collection of similar experimental units
d. **Randomized block design**: randomizations of treatments to exper. units are conducted separately for each block

6. Data in a two-way classification may arise from

   a. observational study (row=sex, col=yrs educ, y=salary)
   b. randomized experiment with 2 treatments
   c. randomized block exper. (row = block, col = treatment)
7. Replication

2-way table with replicates

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2-way table without replicates (1 obs. per cell)

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. **Balance**: same number of observations per treatment combination (i.e. per cell of the table)

b. One *cannot* estimate or *test for* interaction in a 2-way table without replicates (Ch.14)
B. Strategy for 2-way tables, with replicates

1. Note: question(s) of interest may be about: (a) interaction, (b) all levels of one or both factors, (c) a few levels of one or both factors, or (d) certain linear combinations

2. Explore: plot, if possible; fit 2-way non-additive model then plot residuals versus fitted values; transform if necessary

3. Test for interaction with an F-test

4. If question is primarily about main effects, drop the interaction terms if they are not significant (Note: some controversy here due to sample size/significance relationship)
5. Answer specific questions of interest with
   a. linear combinations of $\beta$’s
   b. linear combinations of row or column averages
   c. OR, coefficients in wisely chose parameterizations (my preference)

6. Use multiple comparison techniques, if necessary, for
   (i) multiple comparisons or (ii) data snooping

   e.g.
   (i) Which column factor levels differ from which others?

   (ii) Test column levels 3 and 7 since they turned out to be
        the ones with the smallest and largest effects.
7. A note about a classical parameterization and its estimates

a. Classical parameterization for additive 2-way model (13.5.6):

\[ \mu_{ij} = \mu + \alpha_i + \beta_j \]

Mean in ith row and jth column

overall mean

ith row effect

jth column effect

where \( \alpha_1 + \alpha_2 + \ldots + \alpha_J = 0 \) and \( \beta_1 + \beta_2 + \ldots + \beta_J = 0 \).

b. With BALANCED data, the least squares estimators are

\[ \hat{\mu} = \bar{y} \text{ grand ave.} \]

\[ \hat{\alpha}_i = \bar{y}_i - \bar{y} \text{ ave. of ith row minus grand ave.} \]

\[ \hat{\beta}_j = \bar{y}_j - \bar{y} \text{ ave. of jth column minus grand ave.} \]
## C. A note about interaction

### LARGE FISH

<table>
<thead>
<tr>
<th></th>
<th>Not present</th>
<th>Present</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMALL FISH</td>
<td>µ</td>
<td>µ_\text{fF}</td>
</tr>
<tr>
<td>Present</td>
<td>µ_\text{f}</td>
<td>µ_\text{F}</td>
</tr>
</tbody>
</table>

(a) Does the mean depend on whether small fish are present?
Row 2 average minus row 1 average
(b) Does the mean depend on whether large fish are present?
Column 2 average minus column 1 average
(c) Does the mean depend on the presence of small fish differently when large fish are also present than when large fish are not present?
\[
(\mu_{\text{fF}} - \mu_{\text{F}}) - (\mu_{\text{f}} - \mu)
\]

Notice that estimates of all four means are necessary to answer (c); so it is impossible to estimate interaction without data in all four combinations.
D. Related Issues

1. Paired-t Analysis (Ch. 2) is a special case of 2-way ANOVA

2. In a randomized block experiment, retain block effects even if they are not significant, because
   a. The p-values (from F and t-tests) closely approximate randomization test p-values (as in Sect. 13.4.4)
   b. Conceptually, therefore, the p-values are tied directly to the chance mechanism involved in randomization

3. A *Factorial treatment arrangement* (Sect. 9.6.2) describes the formation of experimental treatments from all combinations of the levels of 2 or more distinct, individual treatments. (e.g. 4 treatment levels formed from 2 levels of Limpit presence × 2 levels of fish presence)
4. Should you use ANOVA or regression routines for 2-way tables?

a. If data are balanced, the F-tests are conveniently laid out in ANOVA routines. The subsequent investigation into specific questions of interest could proceed either with row and column averages or using regression with appropriately chosen indicator variables.

b. If data are unbalanced, there is potential ambiguity in the ANOVA F-tests, but not in the regression extra SS F-tests (because the identification of full and reduced models is more clear).
E. 2-Way tables with 1 obs. per cell (Sect. 14.2)

1. The difficulty: the model ROW + COL + ROW×COL has 0 degrees of freedom for estimating $\sigma^2$

ex: LARGE FISH

<table>
<thead>
<tr>
<th></th>
<th>Not present</th>
<th>Present</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMALL</td>
<td>16</td>
<td>24</td>
</tr>
<tr>
<td>FISH</td>
<td>20</td>
<td>30</td>
</tr>
</tbody>
</table>

Each observation estimates its cell mean, so all residuals are 0.

2. Strategies

   a. **If** ROW (or COL) can be treated as *numerical*, try $x + COL + x\times COL$ (i.e. linear effect of row factor)

   b. or proceed (with caution) supposing no interaction
F. Aside: example of 2x2 factorial, with reps.

1. FOOD TASTING PROBLEM: In food taste tests, General Foods employs a 7-point scale from -3 (terrible) to +3 (excellent). In one experiment, 16 groups of 50 people were randomly assigned to one of 4 treatment groups, formed as all possible combinations of 2 levels of SCREEN (course or fine) and 2 levels of liquid concentration (low or high). “SCORE” is the sum of taste scores for 50 subjects.

<table>
<thead>
<tr>
<th>SCORE</th>
<th>SCR</th>
<th>LIQ</th>
<th>SCORE</th>
<th>SCR</th>
<th>LIQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>0</td>
<td>0</td>
<td>24</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>39</td>
<td>0</td>
<td>0</td>
<td>21</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>77</td>
<td>0</td>
<td>0</td>
<td>39</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>0</td>
<td>60</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>104</td>
<td>1</td>
<td>0</td>
<td>65</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>129</td>
<td>1</td>
<td>0</td>
<td>94</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>97</td>
<td>1</td>
<td>0</td>
<td>86</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>84</td>
<td>1</td>
<td>0</td>
<td>86</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
2.

SUMS OF FOOD TASTE SCORES FOR 16 GROUPS OF 50 PEOPLE
3. F-TEST for interaction: p-value = .32

4. Additive model $\mu = \beta_0 + \beta_1 I_{\text{highSCR}} + \beta_2 I_{\text{highLIQ}}$

   SCREEN effect = $\mu(\text{SCORE}|I_{\text{highSCR}}=1,I_{\text{highLIQ}})$ - $\mu(\text{SCORE}|I_{\text{highSCR}}=0,I_{\text{highLIQ}})$ =
   
   $(\beta_0 + \beta_1 + \beta_2 I_{\text{highLIQ}}) - (\beta_0 + \beta_2 I_{\text{highLIQ}}) = \beta_1$

   Similarly, LIQUID effect = $\beta_2$

5. |
<table>
<thead>
<tr>
<th>Model</th>
<th>df</th>
<th>SS (Res)</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIQ</td>
<td>14</td>
<td>15698</td>
<td>7%</td>
</tr>
<tr>
<td>SCR</td>
<td>14</td>
<td>6113</td>
<td>63%</td>
</tr>
<tr>
<td>LIQ + SCR</td>
<td>13</td>
<td>5089</td>
<td>70%</td>
</tr>
</tbody>
</table>

   So taste scores depend on SCR much more than LIQ
6. Additive model fit: \(46.9 + 51.5 I_{\text{high SCR}} - 16.0 I_{\text{high LIQ}}\)

7. Suppose interest is in \(\beta_1 - \beta_2\) (how much bigger is SCR effect than LIQ effect?)

   a. \(\hat{\beta}_1 - \hat{\beta}_2 = 51.5 - (-16.0) = 67.5\)

   b. \(\text{var}(\hat{\beta}_1 - \hat{\beta}_2) = \text{var}(\hat{\beta}_1) + \text{var}(\hat{\beta}_2) - 2 \text{cov}(\hat{\beta}_1, \hat{\beta}_2)\) (Sect. 10.4.3). From regression output: Estimated \textit{Variance-covariance matrix} of estimated coefficients

<table>
<thead>
<tr>
<th>(Intercept)</th>
<th>SCR</th>
<th>LIQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>73.40</td>
<td>-48.93</td>
</tr>
<tr>
<td>SCR</td>
<td>-48.93</td>
<td>97.86</td>
</tr>
<tr>
<td>LIQ</td>
<td>-48.93</td>
<td>0.00</td>
</tr>
</tbody>
</table>

   \(\text{Est. var}(\hat{\beta}_1 - \hat{\beta}_2) = 97.86 + 97.86 - 2 \times 0 = 195.72; \text{SE}=14.0\)
G. Review of 2-way Tables

1. Data: Response variable, factor A code, factor B code

2. Types of studies leading to 2-way tables:
   a. Observational
   b. Randomized block exper. (factorA = block, B=treat.)
   c. Randomized exper. with 2 treatments

3. Are there replicates (i.e. multiple Y’s for some combinations of factor A code and factor B code)?

4. Strategy when there are replicates
   a. Coded scatterplots if possible
   b. Residual plots; transform if necessary
c. Test for interaction (extra SS F-test)

d. If appropriate, fit additive model (without interaction); use extra SS F-tests to test for Factor A and Factor B effects

e. If factor levels are numerical, explore simpler models (like linear effect of factor levels)

f. Investigate particular questions of interest with: linear combinations of coefficients (or coefficients of indicator variables in carefully chosen reparameterizations), with attention to multiple comparisons if appropriate

5. Strategy when there is a single observation per cell: same as 4 above but without (c) and with special attention to (e)
H. Related Issues

1. Introduction to random effects (14.5.1)
   
a. For chimp data suppose \( \mu(l\text{time}|\text{CHIMPS, SIGNS}) = \alpha_1 C_1 + ... + \alpha_4 C_4 + \beta_1 S_1 + ... + \beta_{10} S_{10} \)
   and \( \text{var}(l\text{time}|\text{CHIMPS, SIGNS}) = \sigma^2 \)

b. If (i) there is no interest in the 4 particular chimps and (ii) those 4 can be considered a random sample, then...

c. treat \( \alpha_1, \alpha_2, \alpha_3, \) and \( \alpha_4 \) as a random sample with mean 0
   and variance \( \sigma_C^2 \)

d. Then \( \mu(l\text{time}|\text{CHIMPS, SIGNS}) = \beta_1 S_1 + ... + \beta_{10} S_{10} \)
   and \( \text{var}(l\text{time}|\text{CHIMPS, SIGNS}) = \sigma^2 + \sigma_C^2 \)
2. Introduction to *nested effects* (14.5.2)

a. Example: $Y = \%$ damage to a leaf

<table>
<thead>
<tr>
<th>Plant</th>
<th>Leaf</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>$Y_1$</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>$Y_2$</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>$Y_3$</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>$Y_4$</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b. Leaf 1 on plant 1 does not correspond to leaf 1 on plant 2; the factor “leaf” is *nested* within the factor “plant”

c. $\mu(Y|\text{PLANT, LEAF}) = \text{PLANT} + \text{PLANT/LEAF}$

means include main effect of PLANT and interaction of PLANT and LEAF (with NO LEAF main effect)
VII. Adjustments for serial correlation

A. Review of independence assumption

1. Model: \( y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \) \( (i = 1, 2, ..., n) \)
   \( \varepsilon_i \) is independent of \( \varepsilon_j \) for all distinct indices \( i, j \)

2. Consequences of non-independence: (a) SE’s, tests, and CIs will be incorrect; (b) LS isn’t the best way to estimate \( \beta \)’s
3. Main violations and remedies

a. Cluster effects (ex: mice litter mates)

Remedy: include cluster as a factor in the model

Plotting code for cluster

Remedy: include cluster as a factor in the model
b. Serial effects (may be present for data collected over time or space)

i) Easy remedy that sometimes works: adjust for serial correlation (Ch. 15)

ii) Or: more advanced time series analysis

Plotting code for time order of data collection
B. Preview

1. Introduction of AR(1) “Auto-regressive model of lag 1” and the first serial correlation coefficient

2. How to estimate the first serial correlation coefficient

3. How to compare two samples when AR(1) serial correlation is present

4. How to fit regression models when AR(1) serial correlation is present (by using a special transformation: filtering)

5. How to examine residuals from a usual fit to see whether
   a. serial correlation is present
   b. serial correlation can be modeled as AR(1)
C. Autoregressive model of lag 1: AR(1)

1. Suppose $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ are random error terms from measurements at equally-spaced time points, and $\mu(\varepsilon_t) = 0$

2. $\mu(\varepsilon_t | \varepsilon_1, \ldots, \varepsilon_{t-1}) = \alpha \varepsilon_{t-1}$

   Regression of $t$th error term on all previous error terms

   autoregression coefficient (a parameter)

3. Notice that the $\varepsilon$’s are not independent, but...

4. the dependence is only through the previous error term

5. Since $\text{corr}(\varepsilon_t, \varepsilon_{t-1}) = \alpha$, it is called the 1st serial correlation coefficient
D. Estimating the first serial correlation coefficient from residuals of a single series

1. Let $e_1, e_2, ..., e_n$ be the residuals from the series

2. Let $c_1 = \sum_{t=2}^{n} e_t e_{t-1}$ and $c_0 = \sum_{t=1}^{n} e_t^2$

3. The estimate of the first serial correlation coefficient ($\alpha$) is $r_1 = c_1/c_0$

4. Note: this is (almost) the sample correlation of residuals $e_2, e_3, ..., e_n$ with the “lag 1” residuals $e_1, e_2, ..., e_{n-1}$
E. Comparing averages of two AR(1) series

1. Series 1: $Y_1, \ldots, Y_{n_1}$; where $Y_t = \mu_1 + \varepsilon_t$

Series 2: $Y^*_1, \ldots, Y^*_{n_2}$; where $Y^*_t = \mu_2 + \varepsilon^*_t$

Suppose both series have AR(1) dependence with common but unknown 1st serial correlation coefficient $\alpha$

Interest is in $\mu_1 - \mu_2$

2. Pooled estimate of first serial correlation coefficient, $\alpha$

   a. Get pooled estimates $c_1$ and $c_0$ (see Display 15.6)

   b. $r_1 = c_1/c_0$
3. Fact from statistical theory: the standard error for the difference in averages, adjusted for serial correlation of lag 1, is

$$SE(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{1 + r_1}{1 - r_1}} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

4. So test \(H_0: \mu_1 - \mu_2 = 0\) or get a CI for \(\mu_1 - \mu_2\) in the usual way, but use this SE. (Note: there is no t-theory in this case; rather the normal distribution is appropriate, if the n’s are large.)
F. Regression in the AR(1) Model

1. Introduction: Steps in global warming analysis

   a. Fit the usual regression of TEMP ($Y_t$) on YEAR ($X_t$).
   b. Estimate the 1st ser. corr. coeff, $r_1$, from the residuals.
   c. Is there serial correlation present? (Sect. 15.4)
   d. Is the serial correlation of the AR(1) type? (Sect. 15.5)
   e. If yes, use the filtering transformation (Sect. 15.3.2):

   $$V_t = Y_t - r_1 Y_{t-1}$$
   $$U_t = X_t - r_1 X_{t-1}$$

   f. Regress $V_t$ on $U_t$ to get AR(1)-adjusted reg. estimates.
2. Filtering, and why it works

a. Simple regression model with AR(1) error structure:

\[ Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t ; \quad \mu(\varepsilon_t | \varepsilon_1, \ldots, \varepsilon_{t-1}) = \alpha \varepsilon_{t-1} \]

b. Ideal “filtering transformations:"

\[ V_t = Y_t - \alpha Y_{t-1} \quad \text{and} \quad U_t = X_t - \alpha X_{t-1} \]

c. Algebra showing the induced regression of \( V_t \) on \( U_t \)

\[
V_t = Y_t - \alpha Y_{t-1} = (\beta_0 + \beta_1 X_t + \varepsilon_t) - \alpha (\beta_0 + \beta_1 X_{t-1} + \varepsilon_{t-1}) \\
= (\beta_0 - \alpha \beta_0) + \beta_1 (X_t - \alpha X_{t-1}) + (\varepsilon_t - \alpha \varepsilon_{t-1}) \\
= \gamma_0 + \beta_1 U_t + \varepsilon_t^* 
\]

Reg of \( V \) on \( U \) has same slope as \( Y \) on \( X \)

There is no serial corr. present in the \( \varepsilon^* \)’s
d. The AR(1) serial correlation has been filtered out by this transformation; i.e.

\[ \mu(\varepsilon_t^* | \varepsilon_1^*, \ldots, \varepsilon_{t-1}^*) = \mu(\varepsilon_t - \alpha \varepsilon_{t-1} | \varepsilon_1, \ldots, \varepsilon_{t-1} - \alpha \varepsilon_{t-2}) = 0 \]

since \( \varepsilon_t - \alpha \varepsilon_{t-1} \) is independent of all previous residuals

e. So, least squares inference about \( \beta_1 \) in the regression of \( V \) on \( U \) is correct

f. Since \( \alpha \) is unknown, use its estimate, \( r_1 \), instead
3. Estimate of warming trend from the global warming data

   a. Fit the simple reg. of TEMP on YEAR  
      Estimated slope: .00449 (SE = .00035)

   b. Get the estimate of 1st serial correlation coefficient from  
      the residuals: r_1 = .452

   c. Compute new columns of data:  
       \[ V_t = \text{TEMP}_t - r_1 \text{TEMP}_{t-1} \]  
       and  
       \[ U_t = \text{YEAR}_t - r_1 \text{YEAR}_{t-1} \]

   d. Fit the simple reg. of U on V  
      Estimated slope: .00460 (SE = .00058)

   e. Use (d) in the summary of statistical findings
f. Filtering in multiple regression

\[ V_t = Y_t - \alpha Y_{t-1} \]

\[ U_{1t} = X_{1t} - \alpha X_{1(t-1)} \]

\[ U_{2t} = X_{2t} - \alpha X_{2(t-1)} \]

etc.

Fit the least squares regression of \( V \) on \( U_1, U_2, \) etc. 
(using \( r_1 \) as an estimate of \( \alpha \))
G. AR(2) Models, etc.

1. AR(2): \( \mu(\varepsilon_t | \varepsilon_{t-1}, \ldots, \varepsilon_{t-2}) = \alpha_1 \varepsilon_{t-1} + \alpha_2 \varepsilon_{t-2} \)

2. In this model the deviations from the regression depend on the previous two deviations

3. Estimate of \( \alpha_1 = r_1 \) as before (estimate of 1st serial correlation coefficient is also the first partial autocorrelation)

4. Estimate of \( \alpha_2 = r_2 \) comes from the multiple regression of residuals on the lag 1 and lag 2 residuals
5. Autocorrelation and partial autocorrelation coefficients

a. *Estimated autocorrelation coefficients of lag k* are (essentially) the correlation coefficients between the residuals and the lag k residuals

b. *Estimated partial autocorrelation coefficients of lag k* are (essentially) the correlation coefficients between the residuals and the lag k residuals, after accounting for the lag 1,...,lag (k-1) residuals

c. In checking to see what order of AR model is necessary, it is (b), not (a) that must be used.
H. Checking for Serial Correlation

1. Introduction: Section 15.4 describes 3 ways to check for serial correlation: (a) Large sample test of $H: \alpha = 0$, (b) runs test, and (c) Durbin-Watson Test.

2. We’ll skip (b) and (c) and use the following:
   If $r_k$ is the $k$th partial autocorrelation coefficient, then
   \[ SE(r_k) \approx \frac{1}{\sqrt{n}} \]
   So, informally, we reject $H_0: \alpha_k = 0$ if $|r_k| > 2/\sqrt{n}$

3. By examining partial autocorrelation coefficients for lags $k = 1, 2, \ldots$, (eg in a PACF plot like Display 15.10) we can ask
   a. Are all $\alpha_k = 0$? (If so, ignore serial dep.)
b. Are all $\alpha_k$ except $\alpha_1$ 0? [If so, use AR(1)]

I. Brief Big Review

1. Regression, least squares, special x terms, inference (t-, F-tests, etc.), transformation

2. Tools and strategies for

   a. data analysis with regression models (Display 9.9)
   b. dealing with outliers and influential observations
   c. dealing with many possible x’s
   d. dealing with multifactor studies
   e. dealing with data collected over time or space
f. (next:) dealing with repeated measurements (Ch. 16)

J. Adjustment for serial correlation—review

1. One of the assumptions justifying the good properties of least squares: independence of y’s given x; i.e. indep. of ε’s

2. One particular type of dependence: serial correlation

3. Autoregressive models for serial correlation
   a. AR(1): \( \mu(\varepsilon_t | \varepsilon_1, \ldots, \varepsilon_{t-1}) = \alpha_1 \varepsilon_{t-1} \)
   b. AR(2): \( \mu(\varepsilon_t | \varepsilon_1, \ldots, \varepsilon_{t-1}) = \alpha_1 \varepsilon_{t-1} + \alpha_2 \varepsilon_{t-2} \)

4. Data exploration if serial dependence is suspected
a. Is there serial correlation? Get $r_1$ from ordinary residuals, $\text{SE}(r_1) = 1/\sqrt{n}$, $Z\text{-stat} = r_1/\text{SE}(r_1)$; $|Z\text{-stat}| > 2$?

b. Is there AR(2) or higher serial correlation? Get $r_2$ (2nd partial autocorrelation, which has only be vaguely defined in this class), $\text{SE}(r_2) = 1/\sqrt{n}$, $Z\text{-stat} = r_2/\text{SE}(r_2)$; $|Z\text{-stat}| > 2$?

5. Adjustment to account for AR(1) serial correlation

a. Adjust SEs in 1-sample and 2-sample problems (15.2)

b. Filtering: regress $V_t = Y_t - r_1 Y_{t-1}$ on $U_t = X_t - r_1 X_{t-1}$ and then perform inference on coefficients as usual
6. Regression via filtering in the AR(2) model:
   \[ \text{regress } V_t = Y_t - r_1 Y_{t-1} - r_2 Y_{t-2} \text{ on } U_t = X_t - r_1 X_{t-1} - r_2 X_{t-2} \]
VIII. Strategies for repeated measures
A. Some definitions

1. *Multivariate response*: a response variable with multiple components [ex: \( y = (\text{length, weight}) \), \( y = (\text{max temp, min temp, ave temp}) \)]

2. *Repeated measure*: a type of multivariate response in which the same variable is measured repeatedly on the sampling or experimental unit [ex: \( y = (\text{blood pressure 1 week after treatment, 2 weeks after, 3 weeks after, ...}) \)]

3. *Longitudinal study*: a randomized experiment or observational study in which experimental or sampling units are followed over time
4. **Cross-over experiment**: an experiment in which each experimental unit receives more than one treatment [ex: half of subjects receive control diet first, half receive special diet; measure cholesterol after 8 weeks then switch diets and measure cholesterol again after 8 more weeks]

5. **Split-plot experiment over time**: experimental units are randomly assigned to one of the levels of one treatment factor (the *whole plot* treatment) and then, via cross-over, receive all combinations of a second treatment factor (the *split plot* treatment) (Two levels of randomization: (1) assignment to whole plot level and (2) assignment to an order for receiving split plot levels.

6. **Split-plot experiment over space**: As in (5) with the split plot levels applied to various plots within a field
B. Strategies for analyzing repeated measures

1. Key: you need not analyze responses in the form they came!

2. Three approaches

   a. Univariate analyses on one or more well-chosen one-dimensional summary of the response (average, maximum, final value, final - first, slope)

   b. Multivariate analysis on several summaries of the response

   c. Repeated measures ANOVA (subject is a factor; there may be dependence among responses from the same subject; subject effect may be modeled as random)
3. How to decide between 2(a) and 2(b):

a. Goal in both cases: one summary response for each research question or each component of a multi-component research question

b. Use 2(a) if there are several distinct research questions (one analysis for each research question)

c. Use 2(b) if there is one research question with several components (Ex: it is hypothesized that treatment will effect the size of fish, and size is measured by (weight, length)
C. Aside: other Stat classes at OSU

1. St 513: (regression for categorical responses, logistic and log-linear regression, experimental design; exploratory multivariate analysis)

2. St 515: (experimental design and analysis of variance associated with advanced designs)

3. St 421/521, 422/522: introduction to mathematical statistics

4. St 435/535: quantitative ecology (requires 412/512)

5. St 557: multivariate analysis (requires 412/512 and math)

6. St 431/531: sampling methods

7. St 565: time series (requires 412/512 and 422/522)
D. Review

1. Multivariate data structure of interest:

| Response | | Explanatory |
|----------|---------------|
| Y_1 Y_2 Y_3 Y_4 Y_5 Y_6 | X_1 X_2 X_3 |
| 13 18 21 33 44 59 | 1 0 22 |
| 3 4 7 22 31 41 | 0 1 5 |

... ... ...

2. Think about choosing numerical summaries of the Y’s (like average, max, slope) as responses for a regression on X’s.

3. How should you use the right summary?

a. Best approach: match summary to question of interest

b. Also available: tools for suggesting summaries (Ch. 17)
E. Two-sample analysis; multivariate response

1. Monkey memory example: Compare $\mu_T$ to $\mu_C$

   $\mu_T = (\mu_{\text{short-term}}, \mu_{\text{long-term}})_T$; $\mu_C = (\mu_{\text{short-term}}, \mu_{\text{long-term}})_C$

2. Hotelling’s $T^2$ statistic for comparing 2 bivariate means

   $$T^2 \text{stat} = \frac{t_1^2 + t_2^2 - 2R t_1 t_2}{1 - R^2}$$

   where

   $t_1 = \text{ordinary t-statistic for comparing component 1}$

   $t_2 = \ldots \text{ component 2}$

   $R = \text{pooled estimate of correlation between first and second component (eg short-term and long-term memory score)}$
3. Computation of R:

a. Recall sample correlation of $y_1$ and $y_2$ is $r = c/(s_1s_2)$ where $c$ is the sample covariance of $y_1$ and $y_2$

$$c = \frac{1}{(n-1)} \sum_{i=1}^{n} (y_{1i} - \bar{y}_1)(y_{2i} - \bar{y}_2)$$

and where $s_1$ and $s_2$ are the sample standard deviations of $y_1$ and $y_2$

b. Pooled estimate of correlation: $R = C/(S_1S_2)$ where $C$, $S_1$, and $S_2$ are pooled estimates; for example:

$$C = [(n_1 - 1)c_1 + (n_2 - 1)c_2]/(n_1 + n_2 - 2)$$

4. To get p-value, convert $T^2$ to an F-statistic this way:

$$F - \text{stat} = \frac{n_1 + n_2 - 3}{2(n_1 + n_2 - 2)} T^2$$
5. Confidence region for $\mu_T - \mu_C$

a. Formally: a 95% confidence ellipse is the set of all $(\delta_1, \delta_2)$ such that Hotelling’s $T^2$ test for $H_0: \mu_{T1} - \mu_{C1} = \delta_1$ and $\mu_{T2} - \mu_{C2} = \delta_2$ has p-value > .05

b. But calculations and communication are too difficult

c. So, instead, we get a conservative confidence rectangle as the intersection of the two confidence intervals

$$(\bar{y}_{T1} - \bar{y}_{C1}) \pm Multiplier \times SE(\bar{y}_{T1} - \bar{y}_{C1})$$

$$(\bar{y}_{T2} - \bar{y}_{C2}) \pm Multiplier \times SE(\bar{y}_{T2} - \bar{y}_{C2})$$

where

$$Multiplier = \sqrt{\frac{2(n_1 + n_2 - 2)}{n_1 + n_2 - 3} F_{2, n_1 + n_2 - 3} (.95)}$$
F. One-sample analysis with bivariate responses

1. Oat bran study, bivariate response for each subject:
   a. Component 1: cholesterol after high fiber diet minus baseline cholesterol
   b. Component 2: cholesterol after low fiber diet minus baseline cholesterol

2. Hypothesis: means of both components are zero

3. Hotelling’s
   \[ T^2_{stat} = \frac{t_1^2 + t_2^2 - 2rt_1t_2}{1 - r^2} \]
   where \( t_1 \) and \( t_2 \) are the univariate t-statistics for testing, individually, the means of component 1 and 2 are zero; and \( r \) is the sample correlation of components 1 and 2.
G. Main Points of Ch. 16

1. Definition of multivariate response and repeated measure

2. Strategies for repeated measures

3. Hotelling’s $T^2$ test for comparing 2 samples with multivariate responses (and confidence intervals)

4. Hotelling’s $T^2$ test for testing the mean of a single sample of multivariate responses (and confidence intervals)
H. Refined suggestion for multivariate response

1. Identify appropriate univariate summaries; do the appropriate *univariate* analysis for each individually

2. Examine the correlation(s) of the residuals from the separate univariate analyses

3. If the correlation(s) are

   a. < .2 in magnitude, don’t worry about multivariate analysis (rationale: the questions are roughly independent)

   b. < .6, use the Hotelling’s $T^2$ test and confidence intervals with the F-multiplier (the questions and answers are not independent; adjust to account for the “overlap”)

   c. > .6, try to pick new summaries that are less dependent