UCLA STAT 13

Introduction to Statistical Methods for the Life and Health Sciences

Instructor: **Ivo Dinov,**

Asst. Prof. of Statistics and Neurology

Teaching Assistants:

Brandi Shanata & Tiffany Head

University of California, Los Angeles, Fall 2007

http://www.stat.ucla.edu/~dinov/courses_students.html

Chapter 12

Randomized Block ANOVA & Intro to Regression

Two Way ANOVA

- There are many types of designs for Analysis of Variance
- Two way ANOVA incorporates analyses when there are two factors of interest
- Your book includes information on:
  - randomized block designs
  - factorial ANOVA

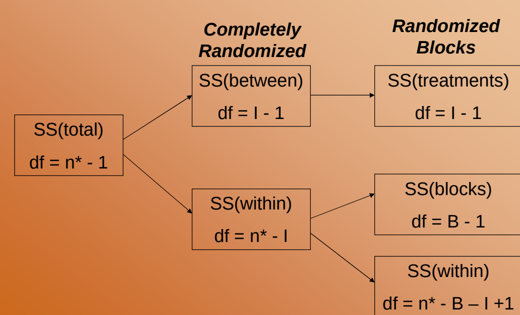
Randomized Block Design

- Recall that in statistics, blocking is the idea of grouping relatively similar units together into matched sets called blocks
  - The idea is that the inherent variability of the units will be reduced with the blocking
- In certain circumstances rather than use a completely randomized design, we can use a block design to control for extraneous variability
  - similar idea to pairing, but doesn't necessarily have to be just two observations per block

Randomized Block Design

- The idea in randomized block designs is to split the total variability into three parts:
  - variability between, same as before
  - variability within
  - variability between the blocks
- Note: the old variability within is subdivided into blocks and within
- Typically we are not interested in a formal hypothesis test for the blocks, we just use this describe the blocking effect on the response variable

Randomized Block Design



Randomized Block Design

In a randomized block design:

SS(total) = SS(within) + SS(treatments) + SS(blocks)

This means that we are adding a new row to our ANOVA table

Randomized Block Design

Example: A study was conducted to investigate whether plants can reduce stress in humans. Two weeks prior to final exams, 10 randomly selected students at a local university took part in an experiment to determine what effect the presence of a live plant, a photo of a plant, or absence of a plant has on the student's ability to relax while isolated in a dimly lit room. Each student participated in 3 sessions – one with a live plant, one with a photo, and one with no plant. During each session finger temperature was measure as an indication of relaxation (higher temperature = more relaxed).

Randomized Block Design

Does the data suggest that there is a difference in mean finger temperature (ie. relaxation) among the three treatment groups? Test using $\alpha = 0.05$.

Two-way ANOVA: Temp versus Plant, Student

Source	DF	SS	MS	F	P
Plant	2	2.942	1.47100	6.69	0.007
Student	9	15.232	1.69244	7.70	0.000
Error	18	3.958	0.21989		
Total	29	22.132			

S = 0.4689 R-Sq = 82.12% R-Sq(adj)= 71.19%

Next slide for Individual CI's

Randomized Block Design

```
Individual 95% CIs For Mean Based on  Pooled StDev
Plant   Mean   --+---------+---------+---------+-------
Live   95.85                         (--------*-------)
None   95.09   (--------*--------)
Photo  95.38           (--------*--------)
               --+---------+---------+---------+-------
              94.85     95.20     95.55     95.90
                   Individual 95% CIs For Mean Based on Pooled StDev
Student     Mean   ----+---------+---------+---------+-----
 1       94.1667   (-----*----)
 2       96.2333                     (----*-----)
 3       96.0000                  (-----*-----)
 4       96.4000                      (-----*-----)
 5       95.4000            (-----*-----)
 6       95.5333              (----*-----)
 7       94.3667     (-----*----)
 8       95.2000          (-----*-----)
 9       95.1333          (----*-----)
10       95.9667                  (-----*----)
                   ----+---------+---------+---------+-----
                    94.0      95.0      96.0      97.0
```


Randomized Block Design

H_o: $\mu_1 = \mu_2 = \mu_3$

H_a: at least two of the μ_k's are different

where 1=live plant, 2 = photo, 3 = no plant

F = 6.69, p = 0.0007

Reject H_o

Conclusion: These data provide evidence to suggest that at least 2 of the true mean finger temperatures are different among the three groups (live plant, photo, and no plant), even after blocking by student to control for extraneous variability.

Randomized Block Design

One-way ANOVA: Temp versus Plant

Source	DF	SS	MS	F	P
Plant	2	2.942	1.471	2.07	0.146
Error	27	19.190	0.711		
Total	29	22.132			

```
S = 0.8431   R-Sq = 13.29%   R-Sq(adj) = 6.87%
         Individual 95% CIs For Mean Based on Pooled StDev

Level   N    Mean  StDev  ---------+---------+---------+---------+
Live   10  95.850  1.042               (----------*----------)
None   10  95.090  0.734  (----------*----------)
Photo  10  95.380  0.713        (----------*----------)
                          ---------+---------+---------+---------+
                               95.00     95.50     96.00     96.50

Pooled StDev = 0.843
```

Linear Relationships

- Analyze the relationship, if any, between variables x and y by fitting a straight line to the data
  - If a relationship exists we can use our analysis to make predictions
- Data for regression consists of (x,y) pairs for each observation
  - For example: the height and weight of individuals

Linear Relationships

Example: The data below are airfares ($) and distance (miles) to various US cities from Baltimore, Maryland.

Destination	Distance	Airfare	Destination	Distance	Airfare
Atlanta	*576*	*178*	Miami	946	198
Boston	370	138	New Orleans	998	188
Chicago	612	94	New York	189	98
Dallas	1216	278	Orlando	787	179
Detroit	409	158	Pittsburgh	210	138
Denver	1502	258	St. Louis	737	98

Linear Relationships

Destination	Distance	Airfare
Atlanta	*576*	*178*
Boston	370	138
Chicago	612	94
Dallas	1216	278
Detroit	409	158
Denver	1502	258
Miami	946	198
New Orleans	998	188
New York	189	98
Orlando	787	179
Pittsburgh	210	138
St. Louis	737	98

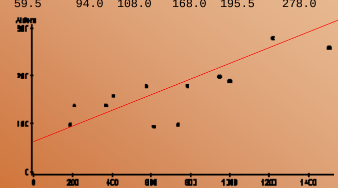
Linear Relationships

- Until now we have described data using statistics such as the sample mean

Descriptive Statistics: Distance, Airfare

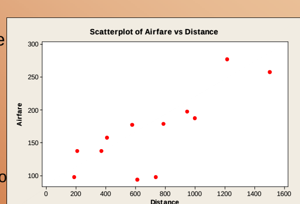
Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Distance	12	0	713	116	403	189	380	675	985	1502
Airfare	12	0	166.9	17.2	59.5	94.0	108.0	168.0	195.5	278.0

- What seems to be missing from this one sample view of the data?

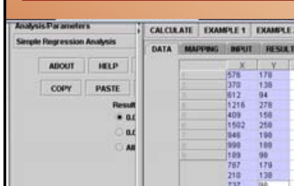


Linear Relationships

- This scatterplot gives us a view of how the dependent variable airfare (y) changes with the independent variable distance (x)
- From this data there appears to be a linear trend, but the data do not fall in an exact straight line
  - Still may be reasonable to fit a line to this data



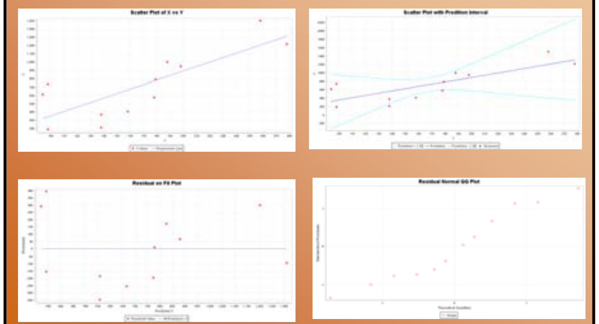
Linear Relationships
SOCR SLR: socr.ucla.edu/htmls/SOCR_Analyses.html



Sample Size = 12
Dependent Variable = X
Independent Variable = Y
Simple Linear Regression Results:
Mean of Y = 166.917
Mean of X = 712.667
Regression Line:
X = -186.090 + 5.384464042692271 Y
Correlation(Y, X) = .795
R-Square = .632
Intercept:
- Parameter Estimate: -186.090
- Standard Error: 229.137
- T-Statistics: -.812
- P-Value: .436

Slope:
- Parameter Estimate: 5.384
- Standard Error: 1.299
- T-Statistics: 4.144
- P-Value: .002

Linear Relationships
SOCR SLR: socr.ucla.edu/htmls/SOCR_Analyses.html



Linear Relationships

- Two Contexts for regression:

  1. Y is an observed variable and X is specified by the researcher
     - Ex. Y is hair growth after 2 months, for individuals at certain dose levels of hair growth cream

  2. X and Yy are observed variables
     - Ex. Height and weight for 20 randomly selected individuals

The Fitted Regression Line

- Suppose we have n pairs (x,y)
- If a scatterplot of the data suggests a general linear trend, it would be reasonable to fit a line to the data
- The question is which is the best line?

Example Airfare (cont')
  - We can see from the scatterplot that greater distance is associated with higher airfare
  - In other words airports that tend to be further from Baltimore than tend to be more expensive airfare

- To decide on the best fitting line, we use the least-squares method to fit the least squares (regression) line

Equation of the Regression Line

- RECALL: y = mx + b
- In statistics we call this $Y = b_0 + b_1X$

  where Y is the dependent variable
  X is the independent variable

  b_0 is the y-intercept $\bar{y} - b_1\bar{x}$

  b_1 is the slope of the line $\dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

Equation of the Regression Line

- **Example**: Airfare (cont')

```
Regression Analysis: Airfare versus Distance

The regression equation is
Airfare = 83.3 + 0.117 Distance
Predictor      Coef  SE Coef      T      P
Constant      83.27    22.95   3.63  0.005
Distance    0.11738  0.02832   4.14  0.002
S = 37.8270   R-Sq = 63.2%    R-Sq(adj) = 59.5%

Analysis of Variance
Source          DF     SS     MS      F      P
Regression       1  24574  24574  17.17  0.002
Residual Error  10  14309   1431
Total           11  38883
```


Equation of the Regression Line

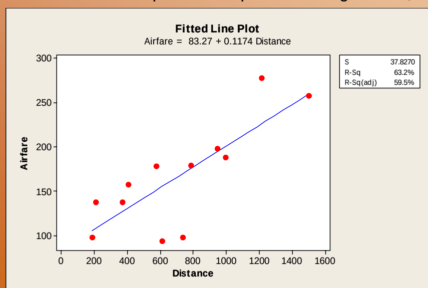
- When we write the least squares regression equation we use the following notation:

$$\hat{y} = 83.27 + 0.117x$$

  - b_1 expresses the rate of change of y with respect to x
    - For every one mile increase in distance, airfare will go up by an additional 0.117 dollars.
    - We could actually describe this as for a 100 mile increase in distance airfare rises by $11.70
  - b_0 expresses where the regression line will hit the y axis
    - It may or may not be interpretable, depends on the context
    - In this case does an airfare of $83.27 when distance traveled is 0 miles make sense?

Equation of the Regression Line

- NOTE: The least squares line passes through $(\bar{x}, \bar{y})$



Equation of the Regression Line

- Predict the airfare for a city that is 576 miles away. If you look at the original data set (first page), *Atlanta's* distance was 576 miles and the airfare was $178

$$\hat{y} = b_0 + b_1 x$$

= 83.27 + 0.11738(576)

= $150.88 (watch units!)

- Calculate the corresponding residual
  - HOLD that thought

  Residual = 178 – 150.88 = $27.12

Equation of the Regression Line

- It is important to only make predictions for values that are within our sampled range of x data
- Extrapolation beyond the scope of our sampled data is dangerous because we do not know what happens to the relationship between x (distance) and y (airfare) outside this range
- In other words, this line may not continue on with the same slope forever

Equation of the Regression Line

Predict the airfare for a city that is 2842 miles away from Baltimore. Does this seem like a legitimate prediction? Explain.

= 83.27 + 0.11738(2842) = $416.86

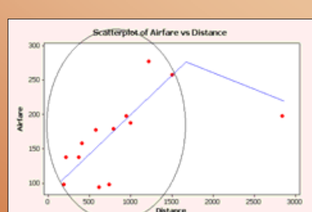
- This does not seem like a legitimate prediction because our sample range of data goes from 189 to 1502 miles
- No making predictions outside our sampled range of data!
- This city (San Francisco) falls outside of this range
- NOTE: The actual airfare for this city was $198

Equation of the Regression Line

We can predict Y for X that are "reasonable" (within the range of modeled X values)

Once we have fit the data with a regression line, if we have done a good job it is natural to use the line to make predictions about Y at certain values of X

We should not predict Y for X values that are "not reasonable" (outside the range of modeled X values)



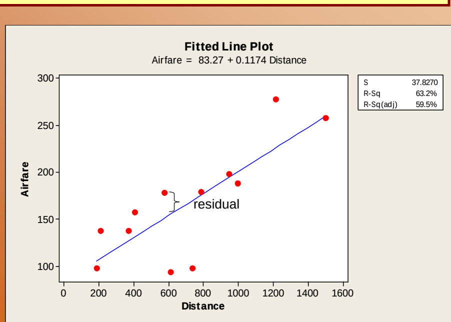
Residuals

- For each observed x value (x_i) there is a predicted y value ($\hat{y}$) based on the regression equation

$$\hat{y} = b_0 + b_1 x$$

- Also associated with each (x_i, y_i) there is a residual
  - the vertical distance between each predicted y ($\hat{y}$) and observed y
  - Residual = $y_i - \hat{y}_i$
- When we add up all the residuals they sum to 0

Residuals



Residuals

- Which city has the largest (in absolute value) residual? Quantify this value.
  - *HINT: look at the scatter plot.* How can you tell?
  - St. Louis because it lies the furthest (vertically) from the regression line

$\hat{y}$ = 83.27 + 0.11738(737) = \$169.78

Residual = 98 – 169.78 = -\$71.78

Residuals

- Which city has the largest predicted value ($\hat{y}$)? Quantify this value.
  - *HINT: look at the scatter plot.* How can you tell?

  Denver because it is the observation with the largest distance and therefore predicted value

$\hat{y}$ = 83.27 + 0.11738(1502) = \$259.57

NOTE: If the slope was negative the largest predicted value would be the observation with the smallest x.

Residuals

Regression Analysis: Airfare versus Distance

…Portion of output omitted…

Obs	Distance	Airfare	Fit	SE Fit	Residual	St Resid
1	576	178.0	150.9	11.6	27.1	0.75
2	370	138.0	126.7	14.6	11.3	0.32
3	612	94.0	155.1	11.3	-61.1	-1.69
4	1216	278.0	226.0	18.0	52.0	1.56
5	409	158.0	131.3	13.9	26.7	0.76
6	1502	258.0	259.6	24.9	-1.6	-0.05
7	946	198.0	194.3	12.8	3.7	0.10
8	998	188.0	200.4	13.6	-12.4	-0.35
9	189	98.0	105.5	18.4	-7.5	-0.23
10	787	179.0	175.6	11.1	3.4	0.09
11	210	138.0	107.9	17.9	30.1	0.90
12	737	98.0	169.8	10.9	-71.8	-1.98

The Residual Sums of Squares

- What we want to measure is how close each observed y_i is to it's predicted value ($\hat{y}$) based on the regression equation
- A summary measure of all the residual distances is called the residual sum of squares

SS(resid) = $\sum (y_i - \hat{y})^2$

Will be small if the observed values lie close to the regression line

The Residual Sums of Squares

Example: Airfare (cont')

Regression Analysis: Airfare versus Distance

The regression equation is
Airfare = 83.3 + 0.117 Distanc

Predictor	Coef	SE Coef	T	P
Constant	83.27	22.95	3.63	0.005
Distance	0.11738	0.02832	4.14	0.002

S = 37.8270 R-Sq = 63.2% R-Sq(adj) = 59.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	24574	24574	17.17	0.002
Residual Error	10	14309	1431		
Total	11	38883			

Residual Standard Deviation

- The 'best" straight line is the one that minimizes the residual sums of squares
- The residual standard deviation can be used as our description of the closeness of the data points to the regression line

$$s_{Y|X} = \sqrt{\frac{SS(resid)}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n-2}}$$

  - how far off predictions tend to be that are made using the regression model
  - Similar idea to s (measures variability around $\bar{y}$)
    $s_{Y|X}$ (measures variability about the regression line)

Residual Standard Deviation

- Similar interpretation to ch 2.
  68% of our data falls within $\pm$ 1 $s_{Y|X}$ from the line
  95% of our data falls within $\pm$ 2 $s_{Y|X}$ from the line
- We expect most of our data to fall within $2s_{Y|X}$ from the regression line

Example: Airfare (cont') $s_{Y|X} = \sqrt{\frac{SS(resid)}{n-2}} = 37.83$

  - Predictions tend to be off by $37.83
  - Most of our observed values will fall within $\pm$ 2(37.83) = $75.66 from their predicted values.

Residual Standard Deviation

Example: Airfare (cont')

```
Regression Analysis: Airfare versus Distance

The regression equation is
Airfare = 83.3 + 0.117 Distance
Predictor      Coef   SE Coef      T      P
Constant      83.27     22.95   3.63  0.005
Distance    0.11738   0.02832   4.14  0.002
S = 37.8270     R-Sq = 63.2%    R-Sq(adj) = 59.5%

Analysis of Variance

Source          DF      SS      MS       F       P
Regression       1   24574   24574   17.17   0.002
Residual Error  10   14309    1431
Total           11   38883
```


The Linear Model

- When we conduct linear regression think of Y as having a distribution that depends on X
- The conditional population of Y is associated with a fixed X
  $\mu_{Y|X}$ is the population mean Y for a fixed X.
  $\sigma_{Y|X}$ is the population standard deviation of Y for a fixed X.
  - In the airfare example: these are the mean and standard deviation of airfare in the subpopulation whose distance is X miles
  - There is a different subpopulation for each X
- Using this we will learn how to infer from the data to make generalizations about the population

The Linear Model

- For linear regression to be valid we must meet two conditions:

1. Linearity:
   - Y is the average at some X + error

$$Y = \mu_{Y|X} + \varepsilon = \beta_o + \beta_1 X + \varepsilon$$

2. Consistency of standard deviations:
   $\sigma_{Y|X}$ does not depend on x
   $\sigma_{Y|X}$ for each x is the same.

See figure 12.9, page 543 in text

The Linear Model

- Random subsampling model: for each (x,y) pair, we regard the value of Y as having been sampled at random from the conditional population of Y values associated with a fixed X
- The quantities we have estimated so far are:
  b_0 is an estimate of β_0
  b_1 is an estimate of β_1
  $s_{Y|X}$ is an estimate of $\sigma_{Y|X}$
  $b_0 + b_1x_i$ is an estimate of $\mu_{Y|X}$

The Linear Model

Example: Airfare (cont')

83.27 is an estimate of β_0

0.117 is an estimate of β_1

37.83 is an estimate of $\sigma_{Y|X}$

83.27 + 0.117x_i is an estimate of $\mu_{Y|X}$

Suppose we wanted to estimate the average airfare for a city that is 250 miles from Baltimore

$$\hat{y} = 83.27 + 0.117(250) = \$112.52$$

Suppose we wanted to estimate the standard deviation for a city that is 250 miles from Baltimore

$s_{Y|X}$ = \$37.82