Introduction to Regression

Administrative Items

Getting help

– See me Monday 3-5:30, Wednesday from 4-5:30, or make an appointment.
– Send an e-mail to stine@wharton.
– Visit the StatLab/TAs, particularly for help using the computer.

Midterm Exam #2

Tuesday April 4, from 6-8 p.m. in Annenberg 110.

Assignment #4

Due Wednesday, March 29.

Questions about Anova

How do you find Tukey intervals in a two-way anova?

It depends upon *which* means you are making a comparison between. For example in the gummy bear experiment, are you comparing (a) the six means based on the 2 positions and 3 elevations, (b) the three means for elevation overall, or (c) the 2 means for positions?

Here's the approach for each. Always keep the "conceptual version" of the formula for the intervals in mind.

$$(\textit{difference in means}) \pm q_{\alpha}(\textit{number of means compared}, \textit{error df}) \sqrt{\hat{\sigma}^2 / \textit{observations for mean}}$$

First off, no matter which situation,the estimate for "sigma-hat squared" is always the mean squared error term from the anova table.

(a) Use $q_{\alpha}(6,MSE)$ with 4 observations per mean
(b) Use $q_{\alpha}(3,MSE)$ with 8 observations per mean
(c) You don't need multiple comparisons in this case since you only are comparing two means – this is the usual t-test.

The example on the next page illustrates the calculations with output from an experiment done in class.

Challenge Question

Why do you think that the value labeled q* in JMP's Tukey-Kramer output differs by the square root of 2 from that found in the table in the textbook?

Example: Gummy Bear Anova

Purpose of the experiment

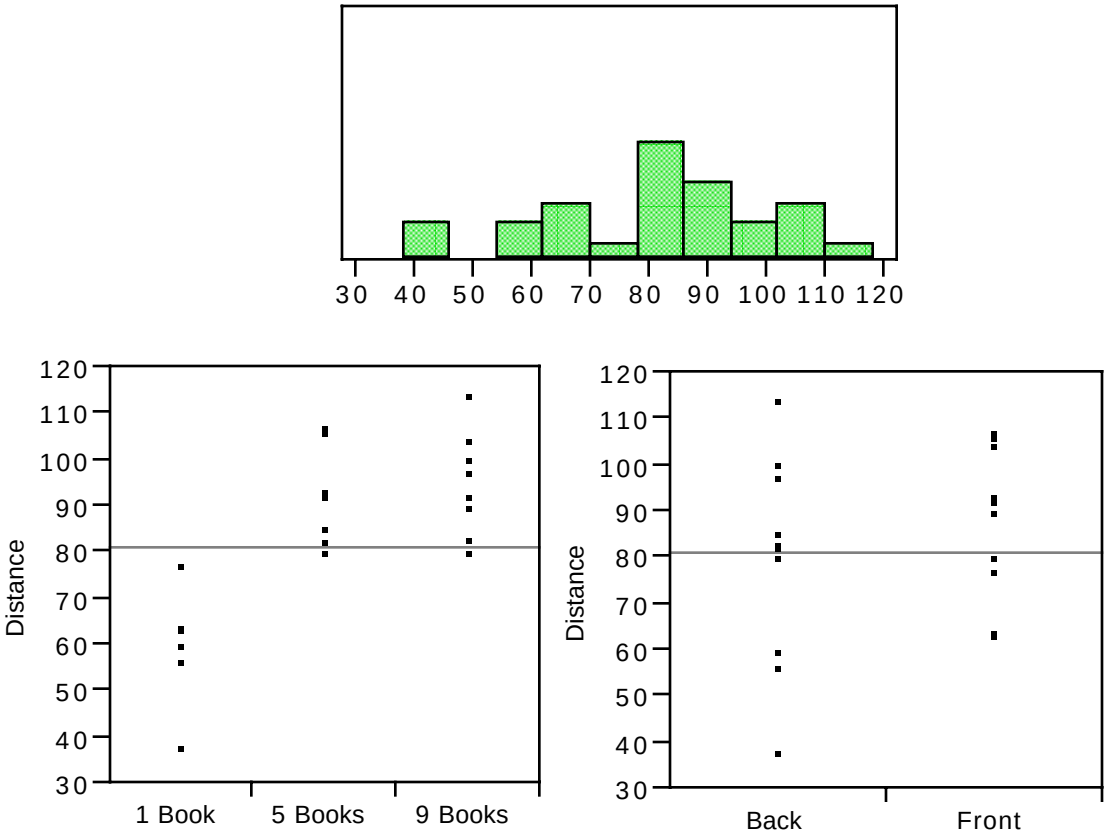
– What's the right combination of elevation and position for launching gummy bears the farthest?

– Learn that the error terms in an anova measure the effect of other factors (often weird things) that happen when you try to repeat the experiment under the same conditions.

Initial thoughts on data quality

Was the experiment done in a way that makes the assumptions of anova reasonable?

Initial data analysis



Should the marginal (overall) distribution of distance be normal?

Does the plot of distance by front/back indicate a lack of constant variance?

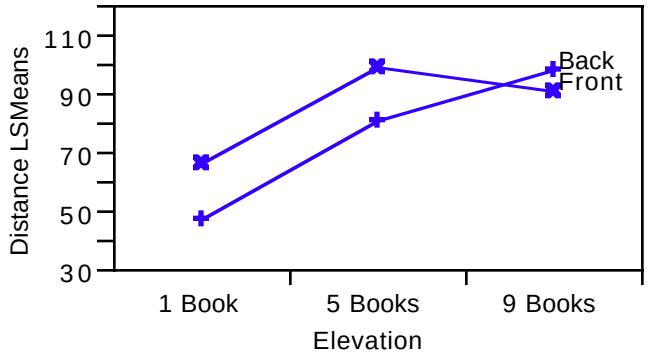
Anova output

Effect Test

Source	Nparm	DF	Sum of Squares	F Ratio	Prob>F
Position	1	1	590.0	6.9	0.0169
Elevation	2	2	6727.1	39.5	<.0001
Position*Elevation	2	2	860.1	5.1	0.0181

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	8177.2	1635.44	19.2185
Error	18	1531.8	85.10	**Prob>F**
C Total	23	9709.0		<.0001



Is interaction present? What would it mean if it is (is not)?

The F test indicates significant interaction, implying that the differences between front to back, for example, changes as the elevation changes. You can see this effect most clearly in the profile plot.

Is the combination that went farthest significantly better than the others?

This requires Hsu's comparison from JMP. Front with 5 books beats the three lowest combinations, but not those at 9 books or back with 5 books.

Mean[i]-Mean[j]-LSD	Front/5	Back/9 B	Front/9	Back/5 B	Front/1	Back/1 B
Front/5	-15.7	-14.7	-7.7	2.0	16.8	35.8
Back/9 B	-16.7	-15.7	-8.7	1.0	15.8	34.8
Front/9	-23.7	-22.7	-15.7	-6.0	8.8	27.8
Back/5 B	-33.5	-32.5	-25.5	-15.7	-1.0	18.0
Front/1	-48.2	-47.2	-40.2	-30.5	-15.7	3.3
Back/1 B	-67.2	-66.2	-59.2	-49.5	-34.7	-15.7

If a column has any positive values, the mean is significantly less than the max.

Are the means for front/back different for each elevation?

Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean
Back/1 Book	4	48.00	11.66	5.83
Back/5 Books	4	81.75	2.36	1.18
Back/9 Books	4	98.50	12.71	6.36
Front/1 Book	4	67.00	6.68	3.34
Front/5 Book	4	99.50	8.10	4.05
Front/9 Book	4	91.50	9.85	4.92

For the comparing back to front at 1 book, the Tukey interval is

$$(67-48) \pm q_{.05}(6, 24-6)\sqrt{85.1/4} = 19 \pm 4.49(4.61) = 19 \pm 20.7$$

Notice that the width factor in this interval (20.7) is bigger than the width factor used in Hsu's comparisons (15.7, found on the diagonal on Hsu's output).

How would you compare elevation (ignoring interaction)?

Level	Number	Mean	Std Dev	Std Err Mean
1 Book	8	57.50	13.44	4.75
5 Books	8	90.62	10.98	3.88
9 Books	8	95.00	11.17	3.95

In this case the Tukey interval for comparing, e.g. 1 to 5 books, is

$$(90.6-57.5) \pm q_{.05}(3,18)\sqrt{85.1/8} = 33.1 \pm 3.61(3.26) = 33.1 \pm 11.8$$

and the difference is significant. Notice that the MSE from the overall anova is used, but that the constant q and divisor under the square root change.

Overall conclusion?

Modeling Relationships

Some "interesting" questions...

(1) How much should you expect to pay for a diamond?

(2) What level of promotion maximizes profit?

(3) Where are levels of cellular telephone use headed?

Pricing diamonds

– What factors are important in determining the price of a diamond?

– How are these factors related to the price of the diamond? (Graphically)

– Can you use this relationship to price diamonds?

Diamond data from web page

Prices (in Singapore $'s) for diamonds sold retail in 1990.

How would you describe the relationship between diamond size (measured in carats) and price?

Would you feel comfortable extrapolating this relationship to very small (e.g. industrial) diamonds or exceptionally large stones (e.g. Hope diamond)?

Linear equations

The diamond relationship is well characterized by a linear relationship. We will write these in "slope-intercept" form as

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

where the constant β_0 is the intercept (point where the line intersects the Y axis) and β_1 is the slope (change in Y/change in X).

JMP analysis

With two continuous variables, JMP's Fit Y by X shows a scatterplot, and we can add a summary line.

Modeling Nonlinear Relationships

Curvature

Linear is the most convenient relationship, but is not guaranteed to hold in all problems. Other types of relations are generically termed "nonlinear" or "curved".

Can you think of some that are non-linear? (e.g. in economics, finance)

Examples of nonlinear relationships

(a) Promotion effects

(b) Cellular telephone subscription rates

Effect of display space on liquor sales (page 12 of casebook)

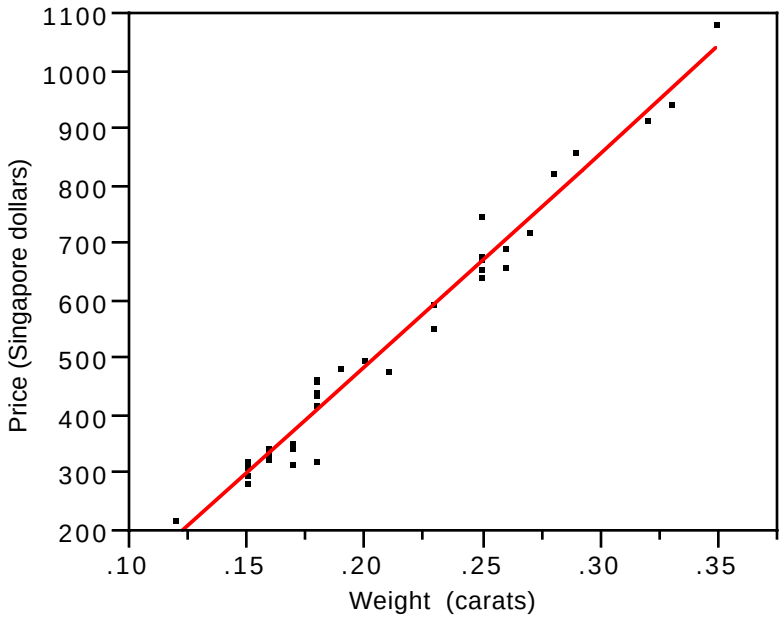
"How much shelf space is needed for a new product?"

– Plot of sales ($) on number of shelf-feet used in display
– Nonlinear (decreasing returns to scale)
– Smoothing splines help you visually detect the presence of curvature.
– Tukey's bulging rule is a graphical device for recognizing which transformations of X and Y help(p 15 of regression casebook)
– Calculus reveals optimal shelf space (baseline of $50/foot); the optimal amount of display space is *not* to use the most display space because of substitution effects.

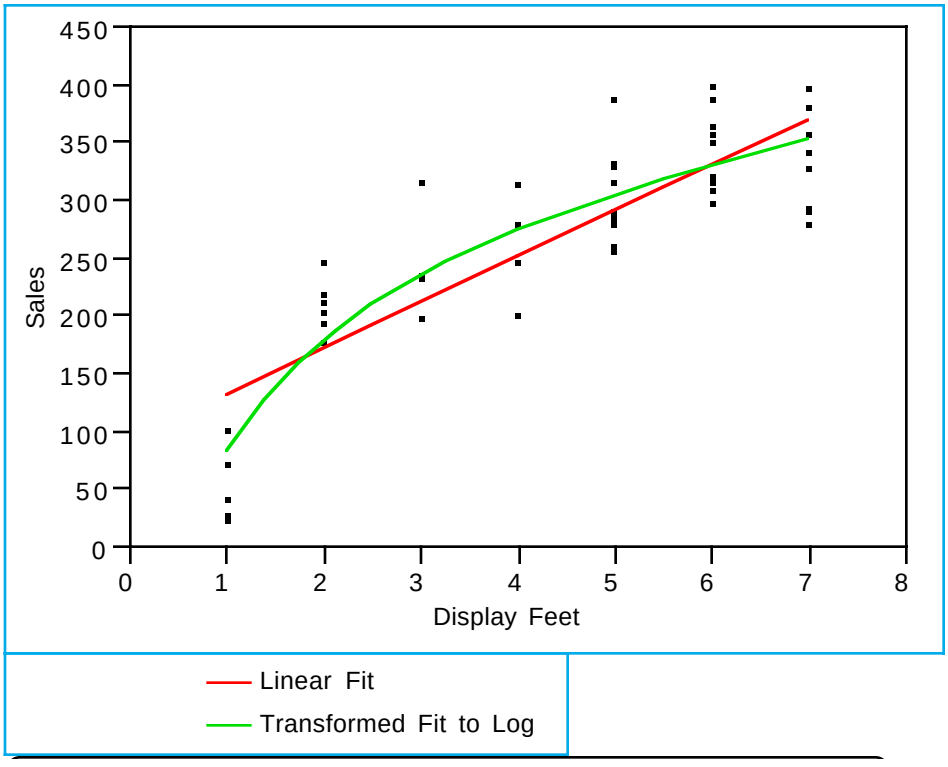
Predicting cellular phone use (page 29 of casebook)

"How many subscribers are expected by the end of this year?"

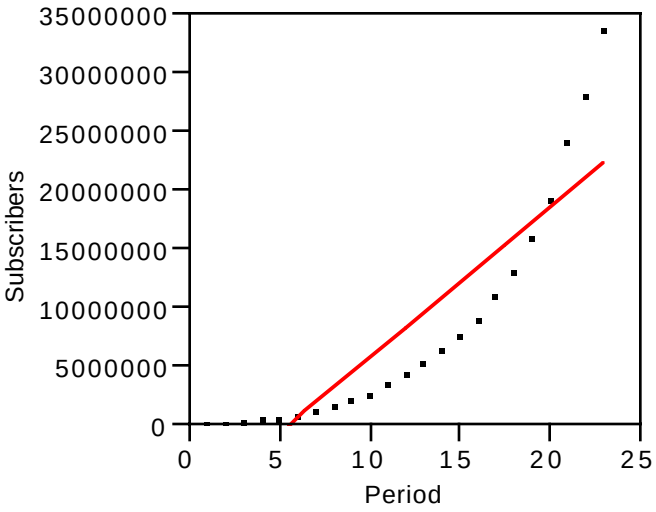
– Time series
– Remarkable pattern misses a lot
– Lesson to learn...
Plot should be linear on the transformed scale.

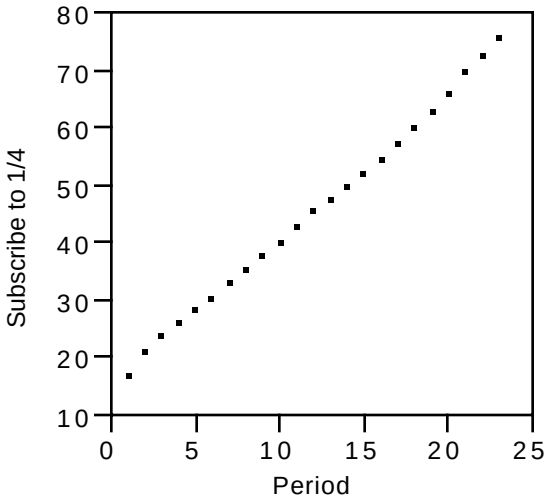


Price (Singapore dollars) = -259.63 + 3721.02 Weight (carats)



Sales = 83.5603 + 138.621 Log(Display Feet)


35000000
30000000
25000000
20000000
15000000
10000000
5000000
0
Subscribers
0
5
10
15
20
25
Period


80
70
60
50
40
30
20
10
Subscribe to 1/4
0
5
10
15
20
25
Period