Department of Statistics
The Wharton School
University of Pennsylvania

STAT 621 Fall 2005

Business Analysis Using Regression
Syllabus

Instructors:

Edward George	edgeorge@wharton	446 JMHH	898 8229
Robert Stine	stine@wharton	444 JMHH	898 3114
Jonathan Stroud	stroud@wharton	465 JMHH	573 3637
Richard Waterman	waterman@wharton	443 JMHH	898 8227

Source Material

Required

- Class Notes. A full copy of these notes (two sided copies, 3 hole punched) can be purchased from Wharton Reprographics. These can also be downloaded directly from the 621 Webcafe e-room.
- Foster, Stine, and Waterman, *Business Analysis Using Regression: A Casebook*, Springer-Verlag, Revised Printing (Note: This is a different casebook from the one used in 603. Be sure that your copy is distinguished by the phrase "Revised Printing" that appears in white lettering on the cover).
- SAS Institute, *JMP IN*, Release 5.1, Windows or Macintosh version, Duxbury Press (software purchase includes Sall, Creighton and Lehman (SCL), *JMP Start Statistics*, 3rd Edition).

Optional

- Freedman, Pisani and Purves, *Statistics*, 3rd edition, Norton
- Hildebrand, Ott and Gray, 2nd edition, *Basic Statistical Ideas for Managers*, Duxbury Press.

The fundamental material for the class is contained in the Class Notes which will be discussed and elaborated in the class lectures. A good deal (but not all) of the Class Notes is also elaborated in the *Business Analysis Using Regression* (BAR). For those who would like to go beyond the material in the Class Notes, also suggested below are

optional supplemental readings in Freedman, Pisani, and Purves (FPP) and Hildebrand, Ott and Gray (HOG). FPP is a highly verbal and conceptual book, and is an excellent introduction both for "poets" who are unfamiliar with technical readings and for "quant jocks" who would like a better sense of the reasoning process of statistics. HOG is the traditional "reference manual" and explains the details of statistical procedures more fully than can be done in class.

JMP is the computer package we'll use to for statistical calculations and graphics. Those who took Stat 603 in pre-term will be familiar with the package. It will be employed considerably in Stat 621. In particular, an essential component of 621 will entail project work that will require substantial use of JMP. Although JMP is merely a tool and not the central point of the course, it is sufficiently useful that you need it. The JMP manual SCL provides a detailed introduction to the many powerful facets of JMP.

Course Overview

In this course, you will learn the fundamental statistical methods of regression analysis. These methods and their application will reappear in many other MBA classes and are part of the basic "tool kit" expected of all MBAs in their careers.

The Class Notes are organized into modules which will be covered in order.

Module 0 – Getting Started
Module 1 – Fitting Equations to Data
Module 2 – The Simple Regression Model (SRM)
Module 3 – Inference in Simple Regression
Module 4 – The Multiple Regression Model (MRM)
Module 5 – More on Multiple Regression
Module 6 – Categorical Predictors
Module 7 – More on Categorical Predictors
Module 8 – Model Building
Module 9 – Time Series Modeling

Before each class, you should review the material from the previous class and you should skim the Class Notes that will be covered. This is a course that builds on itself and it is crucial to not fall behind. The classes will focus on critical interpretation of results and analysis of assumptions. We will use JMP to carry out the computations, although the software itself is not the main focus of the course.

Students enrolled in this course are expected to be familiar with the key ideas covered in Statistics 603. These foundations include data displays (boxplots, histograms, quantile plots, and scatterplots), summary statistics (such as the mean, median, standard deviation, and correlation), and basic features of statistical estimation and testing (including sampling distributions, standard error, confidence intervals, t statistics, p-values). If you need to refresh your knowledge of this material, you can find all the Stat 603 materials in

the Stat 621 WebCafe e-room. In particular, you should work through Assignment 3 of Stat 603 before 621 begins.

Assignments, Quizzes and Exam

There will be five weekly assignments. These are posted in the Course Materials folder on Webcafe. Although these assignments will not be collected, they are essential for the learning process and you should treat them as a requirement. Solutions will be posted for you to check your work.

There will be four short in-class quizzes throughout the course. These will take place on Sept 19, 26, Oct 3, 10. (See the WebCafe calendar). Assignment 1 should be completed before Quiz 1, Assignment 2 should be completed before Quiz 2, etc.

There will be a two hour final exam from 6-8PM on Monday, October 24.

Learning Team Project

A statistics project will be assigned to each learning team early in the course. This project will involve the analysis of a substantial data set which will be reported in three group installments. The first installment is due no later than 5PM on Thursday, September 22; the second installment is due no later than 5PM on Friday, October 7; and the third installment is due no later than 5PM on Tuesday, October 25. (See the WebCafe calendar). You are encouraged to turn in these installments earlier than these due dates. These installments must only reflect the work of your learning team. You are strictly forbidden from discussing this project with anyone outside your learning team.

Grading

Grades for the course will be based on the final examination (50%), quizzes (20%), statistics project (30%). You must pass the final exam to pass the course.

Teaching Assistants (TAs)

Four TAs for Stat 621 will hold office hours and regular JMP IN help sessions throughout the course. Times and locations will be posted in the 621 Webcafe e-room.

Classroom Expectations - Concert Rules

- Class starts on time
- Sit according to the seating chart
- Late entry or reentry is severely discouraged
- Name tents displayed
- All phones and electronic devices turned off

Department of Statistics
The Wharton School
University of Pennsylvania

Statistics 621 Fall 2005

Module 0

Getting Started

Administrative Issues

Webcafe - Stat 621

- Syllabus, data, assignments, calendar, discussion
- These lecture notes

Syllabus

- Office Hours
- Assignments: Not handed in but essential
- Quizzes: 4 in-class, short quizzes
- Learning Team Project: 3 installments
- Grading: Final 50% Final, Quizzes 20%, Project 30%
- Passing final is necessary to pass course
- Materials: BAR Casebook and JMP-IN software
- Other books are optional
- Teaching Assistants
- Concert Rules!
  - Class starts on time
  - Sit according to the seating chart
  - Late entry or reentry is severely discouraged
  - Name tents displayed
  - All phones and electronic devices turned off

What You Should Already Know (Stat 603)

Graphical tools

Histogram, normal quantile plot, boxplot, comparison boxplots, and scatterplot.

Expected value, variance and correlation

Expected value is an average, weighted by probabilities.
Variance is the expected squared deviation from mean.
Correlation measures the strength of linear association.

Normal distribution

95% of the distribution lies in the range $\mu \pm 2\sigma$
Normal quantile plot as a diagnostic

Standard error

Sample-to-sample variability of a statistic

$SE(\bar{x}) = s/\sqrt{n}$

Confidence interval

Estimate ± 2 SE(Estimate)
Contains "truth" for 95% of samples

Hypothesis test

t-statistic/t-ratio counts the SE's from conjectured value
p-value measures "plausibility" of H_0
p-value < 0.05 <=> reject H_0 at the .05 significance level
<=> hypothesized value lies outside 95% CI

Software

JMP – to the extent it was used in Stat 603

Course Overview

Regression Analysis

Decision making in the presence of one or more factors that affect the outcome of interest. We will rely on *regression models*, first of a simple form and then more elaborate:

Simple regression: one predictor to model a response

Multiple regression: Using several predictors, of various types to model a response

Module 0 – Getting Started
Module 1 – Fitting Equations to Data
Module 2 – The Simple Regression Model (SRM)
Module 3 – Inference in Simple Regression
Module 4 – The Multiple Regression Model (MRM)
Module 5 – More on Multiple Regression
Module 6 – Categorical Predictors
Module 7 – More on Categorical Predictors
Module 8 – Model Building
Module 9 – Time Series Modeling

A Supply Problem

This problem introduces some of the complexities of real decision making that regression can simplify. Regression provides you with a tool to deal with several things at once.

Suppose you manage a company that supplies food at professional sports games.

Question: The food perishes if you prepare too much, but you miss out on possible sales if you do not have enough. So, how much should you have ready for the next game?

Data: From 20 recent games, you have data like this…[1]

Game	Time of Day	Temperature	Crowd Size	Sales
1	Night	94	46703	2114
2	Night	78	48646	1821
3	Night	94	45116	2035
4	Day	67	24230	705
5	Night	59	46215	1183
6	Night	65	42781	1399
7	Night	94	49765	2407
8	Night	77	57220	1963
9	Day	90	29138	1292
10	Day	84	25399	1091
11	Night	77	52301	1981
12	Night	84	43482	1796
13	Night	89	48626	1959
14	Day	86	31985	1265
15	Night	82	51043	2153
16	Night	56	45652	1158
17	Night	89	42427	1956
18	Day	59	23730	682
19	Day	73	22129	825
20	Night	63	59528	1619

[1] This data is in the file vendor.xls.

How much should you have on hand for these upcoming games?

Day / Night	Expected Temp	Sold Tickets
Night	64	46413
Night	62	41495
Night	77	41797
Day	73	43624
Night	58	51844
Night	65	46625
Night	70	53925
Night	56	47268
Day	66	29253
Day	64	43197
Night	94	59790
Night	66	41431
Night	82	57174
Day	85	28301
Night	84	47003
Night	62	47718
Night	64	58335
Day	85	46227
Day	59	30951
Night	63	58330

We'll see that not only is there an easy way to generate the needed estimates, but the approach also gives you a sense for how accurate its predictions will be.

Department of Statistics
The Wharton School
University of Pennsylvania

Statistics 621 Fall 2005

Module 1

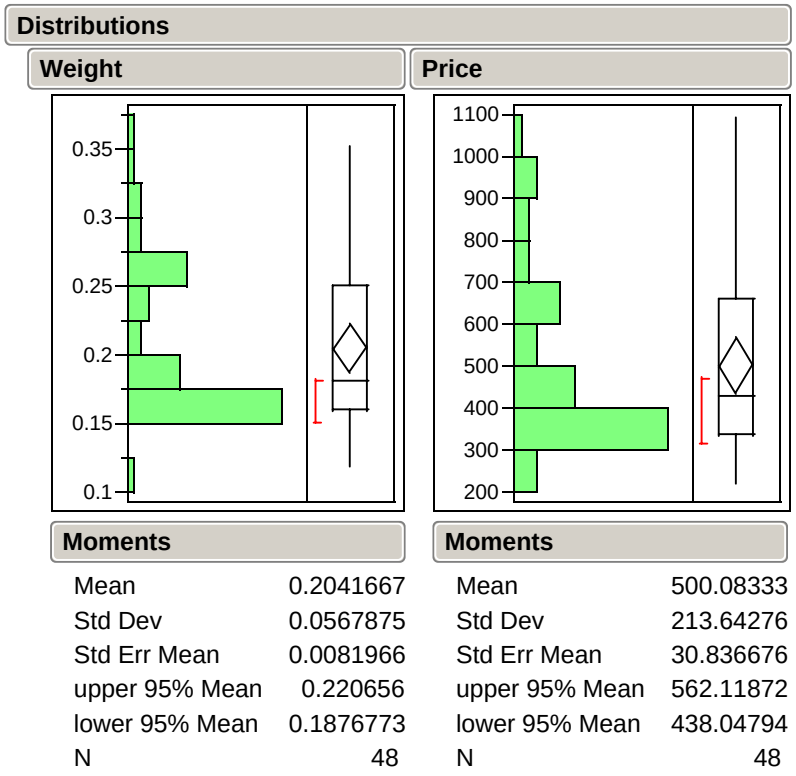
Fitting Equations to Data

Relationships in Bivariate Data

How does the size of a diamond influence the selling price of a diamond ring?

The file diamond.jmp contains the price (in Singapore $) and weight (in carats) of 48 diamond rings.[1]

JMP summaries for these two variables are obtained as[2]

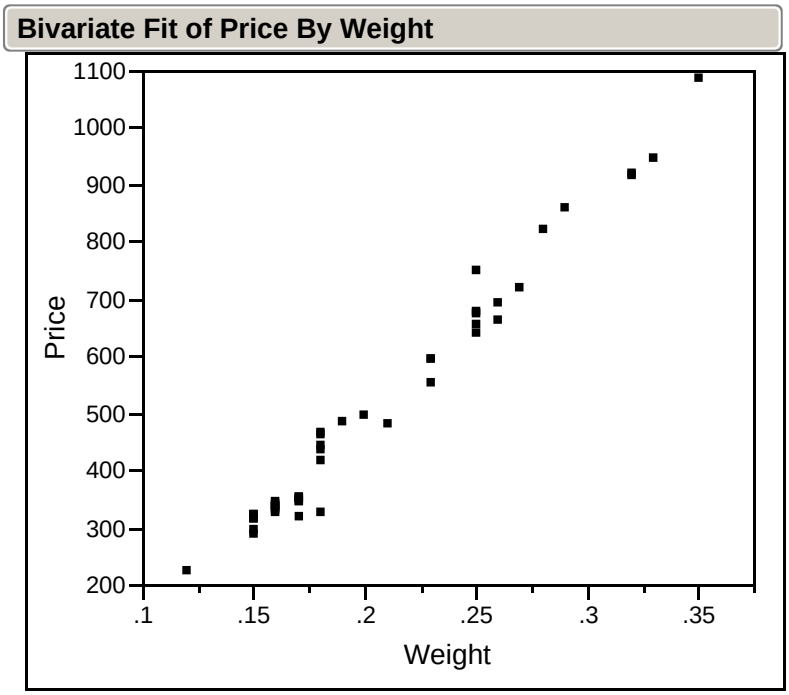


Moments		Moments	
Mean	0.2041667	Mean	500.08333
Std Dev	0.0567875	Std Dev	213.64276
Std Err Mean	0.0081966	Std Err Mean	30.836676
upper 95% Mean	0.220656	upper 95% Mean	562.11872
lower 95% Mean	0.1876773	lower 95% Mean	438.04794
N	48	N	48

[1] The source of this data was a February 29, 1992 advertisement in the *Straights Times*, a Singapore newspaper.
[2] Use Analyze/Distribution with quantiles deselected.

What do these summaries tell you about the relationship between weight and price?

The relationship between weight and price is revealed by a *scatterplot*.[3] Which goes on the *x*-axis and which goes on the y-axis?

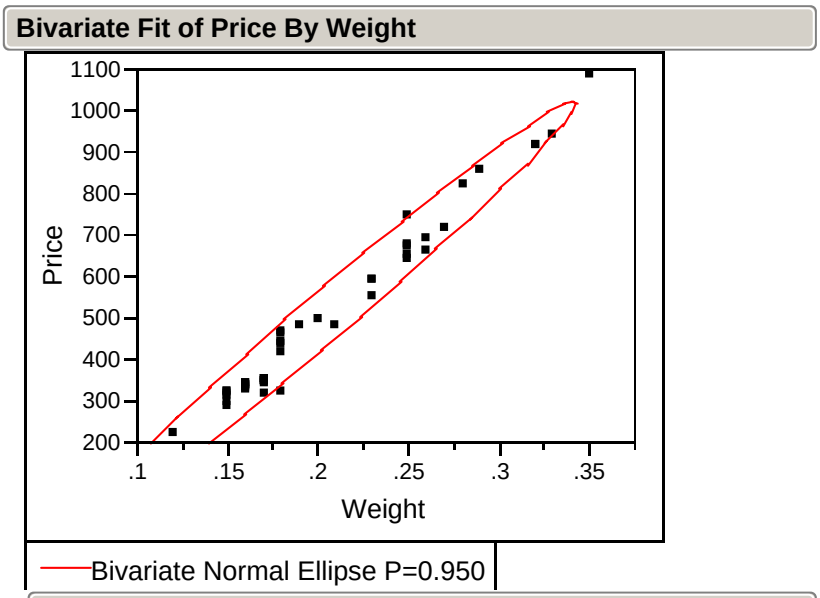


How would you describe the relationship between the weight and price of these diamonds?

[3] Apply Fit Y by X selecting Price as Y and Weight as X

A statistic that measures linear association is the (sample) correlation coefficient which we denote by r.[4]

For the diamond data, JMP computes[5] $r = .989$



Correlation

Variable	Mean	Std Dev	Correlation	Signif. Prob	Number
Weight	0.204167	0.056788	0.989071	0.0000	48
Price	500.0833	213.6428			

r lies between -1 and 1 and is an estimate of the (population) correlation ρ_{XY}

Although correlation is a useful summary measure, we can say much more with a regression analysis. For example, regression can be used to predict price from weight.

---

[4] If you would like to know, $r = \frac{\frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y})}{s_X \, s_Y}$. measuring the degree of linear association.

[5] Right click on the title bar to obtain the Pop-up menu and then choose the Density ellipse option to obtain the correlation output. JMP will also add an ellipse to the initial plot suggesting where the population density is largest.

The Simple Regression Setup

Several characteristics of the data used in a simple regression

Observe n independent pairs $(x_1, y_1), \ldots, (x_n, y_n)$

x explains or predicts y

x is called the independent[6] variable, explanatory variable or predictor

y is called the dependent variable or response

Examples

x	y
weight of diamond	price
protein in diet	weight gain
past DJIA	present DJIA
market return	stock return
interest rate	savings rate

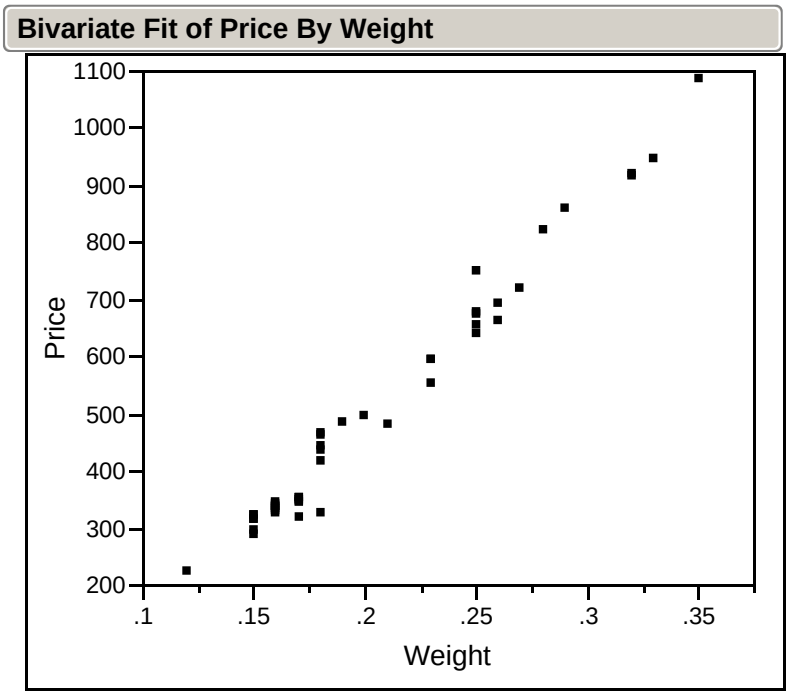
Jargon

Regression with <u>one x</u> is called <u>simple regression</u>
Regression with <u>several x's</u> is called <u>multiple regression</u>

[6] This terminology for the predictor in a regression can be a bit confusing but is nonetheless common. To say that "x is the independent variable" does not imply that the observations are independent or that x is independent of y. The terminology owes to the asymmetry of regression: we think of predicting y from x, not vice versa.

The Least Squares (LS) Regression Line

In scatter plots for regression, *y* is traditionally shown on the vertical axis and *x* is shown on the horizontal axis



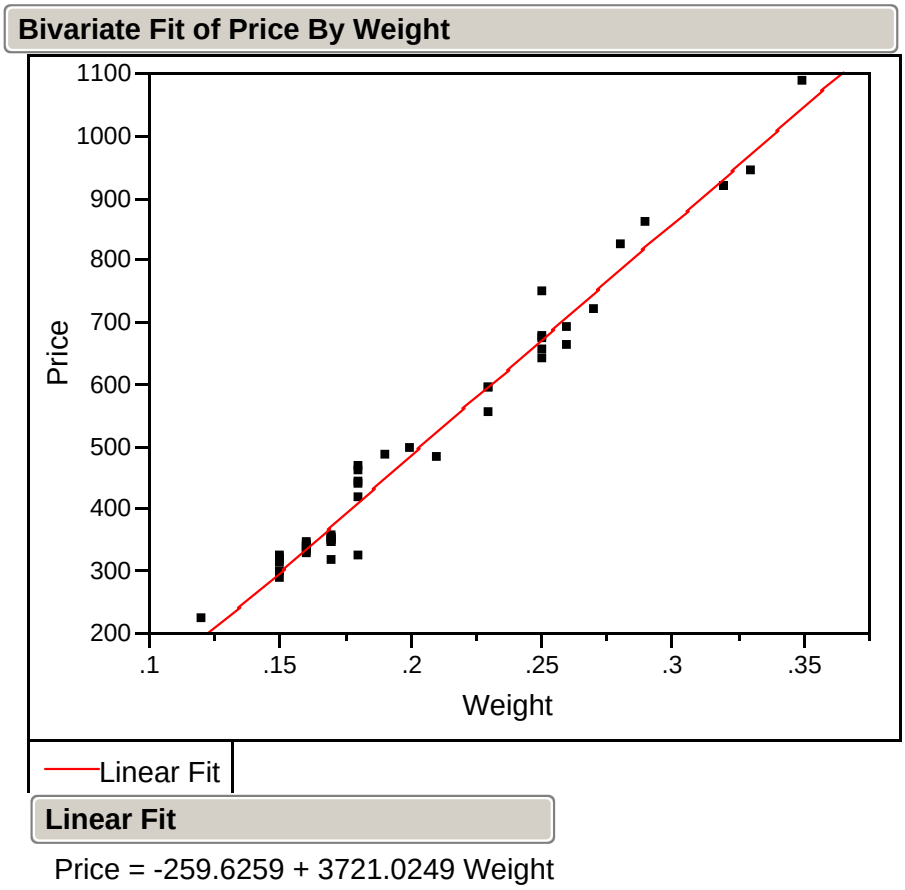
A regression analysis computes the line that minimizes the sum of squared vertical[7] distances from the line to the data.

This line is called the *least squares (LS) regression line.*
To identify this LS line, we will use the equation

$$y = b_0 + b_1 x$$

[7] Why vertical distances and not the sort of distance that you learned in geometry? Again, it's the asymmetry of how we treat the predictor and the response. The vertical distance is the prediction error.

JMP computes this equation for the LS regression line, and then adds the line to the scatterplot.[8]



The LS regression line for this regression is (approximately)

---

[8] Use Fit Y by X as before. Right click on the title bar to obtain the Pop-up menu and then choose the subcommand Fit Line. (For the moment we focus only on the output above). If you would like to know, the formulas to obtain b_0 and b_1 from the data are

$$b_0 = \bar{y} - b_1\bar{x} \quad \text{and} \quad b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} .$$

Interpreting the LS Regression Line

As will become clear in Module 2, the LS regression line estimates the "population" mean of the response y given the value of the predictor x.

Based on this, in the diamond regression what are the interpretations of these values?

$$b_0 + b_1(.3) \approx -260 + 3721(.3) = 856.30$$

$$b_1 \approx 3721$$

$$b_0 \approx -260$$

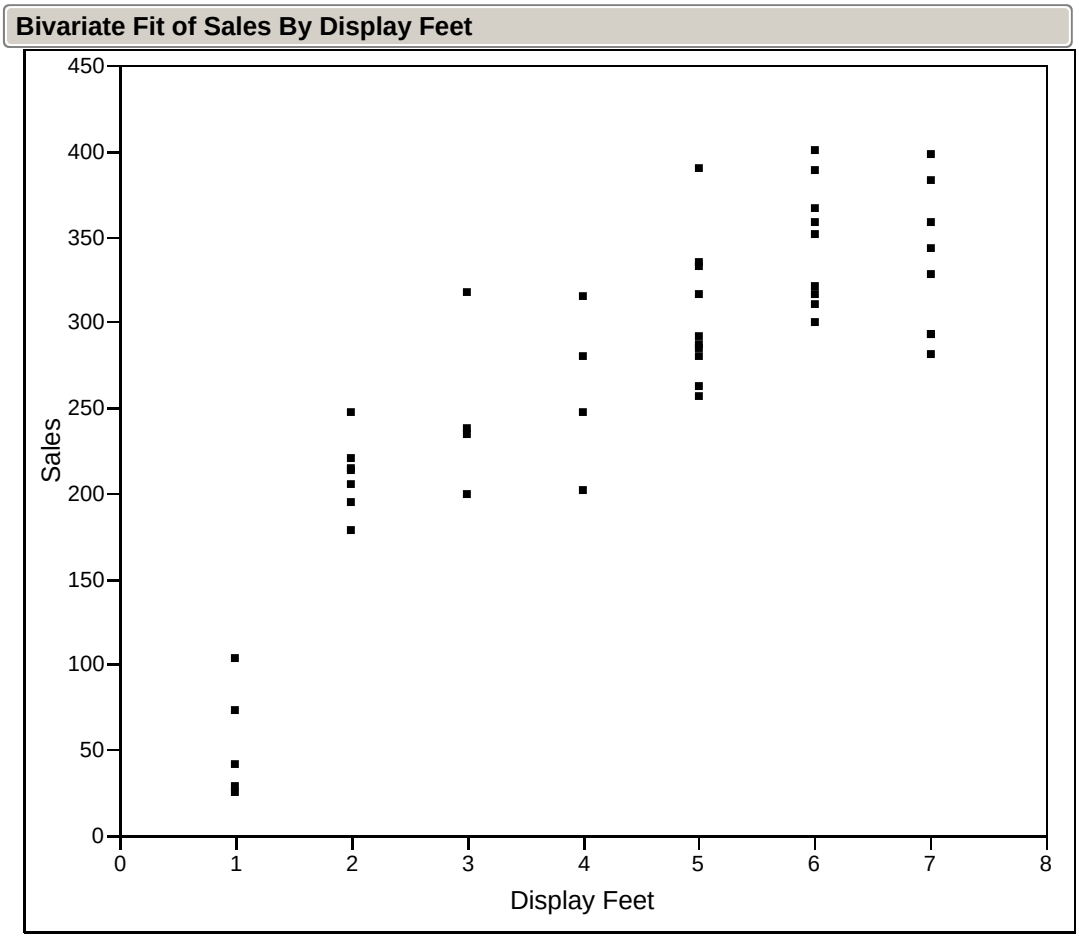
Although one could simply fit a line of the form $y = b_1x$, which goes exactly through the origin (0,0), it is usually not done.

Interesting fact: If all the data $(x_1, y_1), \ldots, (x_n, y_n)$ had the same x value, the LS estimate would be the point $(x, \bar{y})$, but with no variation in x, we could not fit a unique line.

A Nonlinear Simple Regression (BAR, *p* 12)

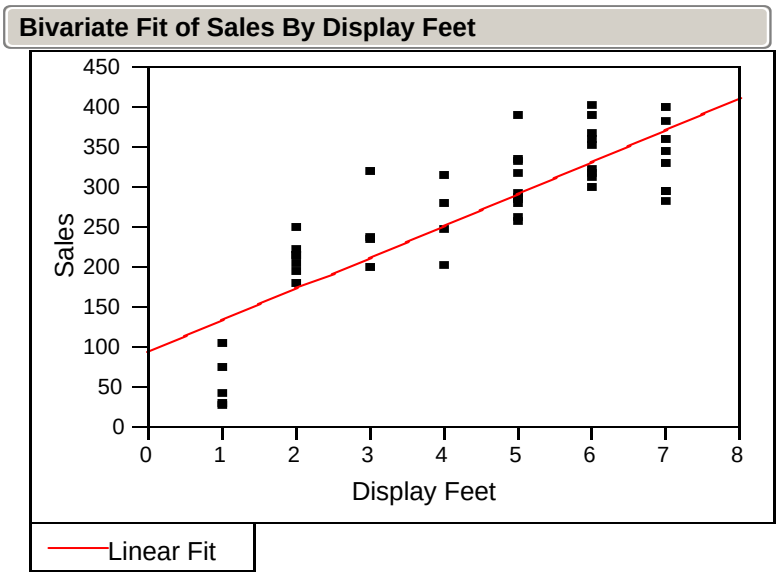
A large chain of liquor stores would like to know how much display space in its stores to devote to a new wine.

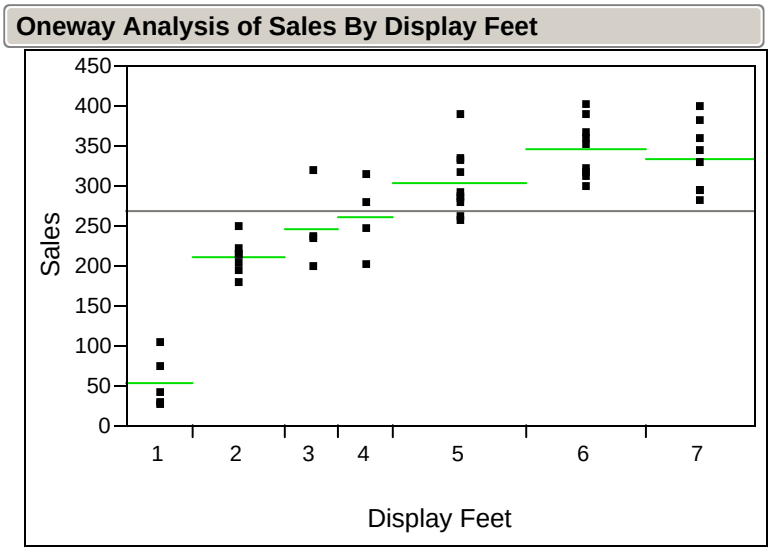
The file display.jmp contains sales ($) and linear shelf-feet per month collected from 47 stores of the chain.



Using a line to estimate the average of sales for a given amount of promotion (*y* given *x*) seems silly here. Why?

Note how the LS regression line severely misses some of the different group means[9]





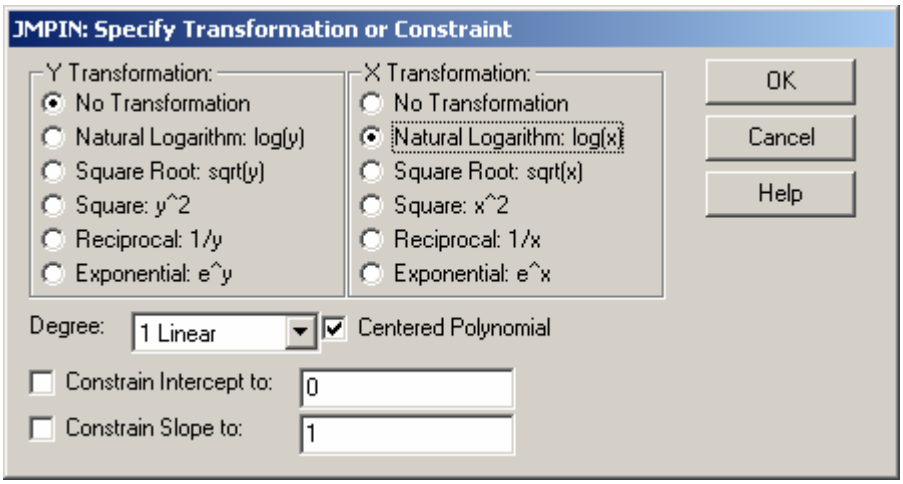
To get a feel for the shape of the relationship between sales and display feet, one might simply sketch a smooth curve that is closer to the center of each group. (BAR, *p* 13)

[9] To obtain the second plot in JMP, use Fit Y by X, treating Display Feet as nominal and use the subcommand Display Options > Mean Lines.

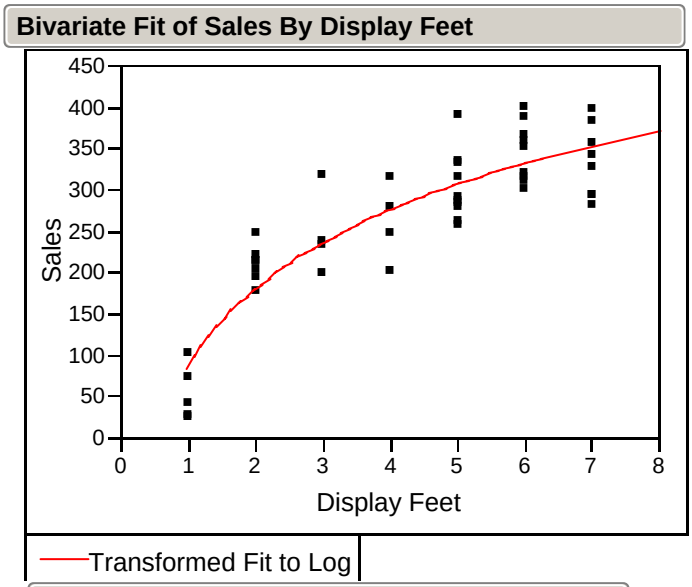
This shape of such a curve is similar to the shape of $y = \log x$, and so we might consider fitting a curve of the form

$$y = b_0 + b_1 \log x$$

This can be done using the Fit Y by X subcommand Fit Special and selecting Natural Logarithm for the X transformation



to obtain



The fitted function

$$y = 83.56 + 138.62 \log x$$

is a least squares fit in the sense that of all functions of the form $f(x) = b_0 + b_1 \log x$, this one minimizes the sum of squared vertical distances from the line to the data.

Visually, this fitted function appears to describe well the relationship between average sales and display feet. Does it make sense?

What is the interpretation of $b_1 \approx 138.62$?

If x increases by 1%, then y will increase by

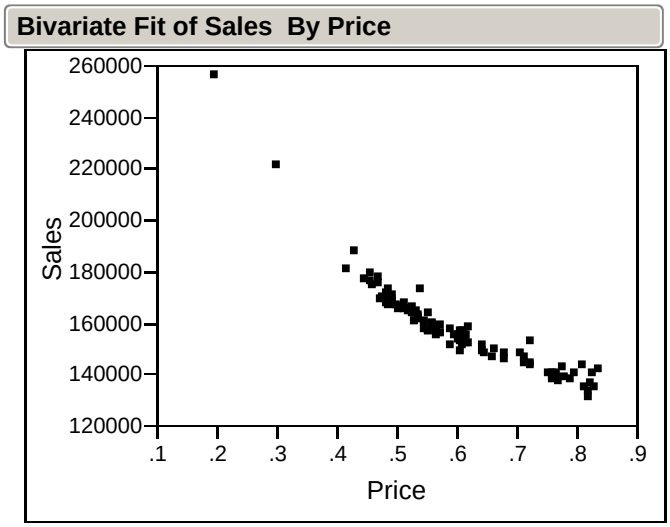
What is the interpretation of $b_0 \approx 83.56$?

$y = 83.56$ when $x =$

Another Nonlinear Example: The Constant Elasticity Model

A large manufacturer of cat food wishes to determine how demand for their 375 gram wet cat food changes as a function of price (Cat food.jmp).

The relationship between sales volume (y) and excess price per can[10] (x) does not appear to be linear



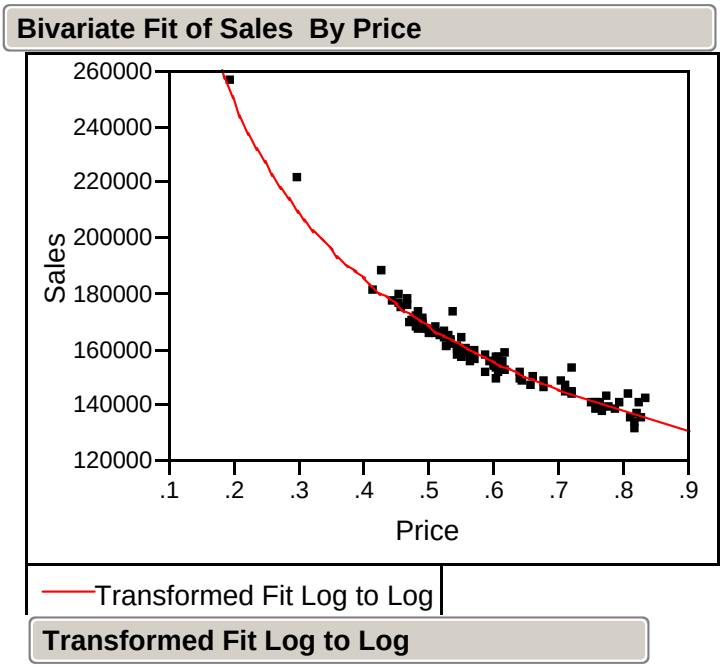
For this data, we might consider fitting a curve of the form

$$\log y = b_0 + b_1 \log x$$

which is a simple linear relationship between $\log y$ and $\log x$

[10] Excess Price = (Actual Price - \$1.50) here. This adjustment yielded a better fit of the log-log model below.

We can use regression to fit this equation[11]



An interpretation of b_1 for this curve is:

If x increases by 1%, then y will increase by

Here, b_1 is called the *elasticity* of demand with respect to price, and the relationship $\log y = b_0 + b_1 \log x$ is often expressed as

$$y = e^{b_0} x^{b_1}$$

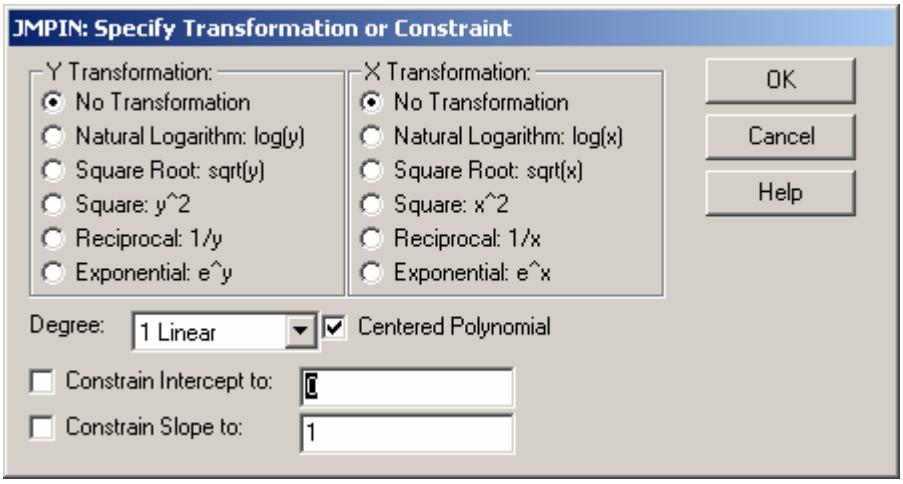
[11] Again use the Fit Y by X subcommand Fit Special and select the Natural Logarithm transformation for both y and x.

Other Transformations

The choice of the transformations for the nonlinear regressions above was guided by the shape of the relationships and the sensibility of its interpretation.

However, other choices might also be reasonable, and one typically proceeds by trial and error.[12]

The subcommand Fit Special offers a variety of such choices for transforming y and/or x.



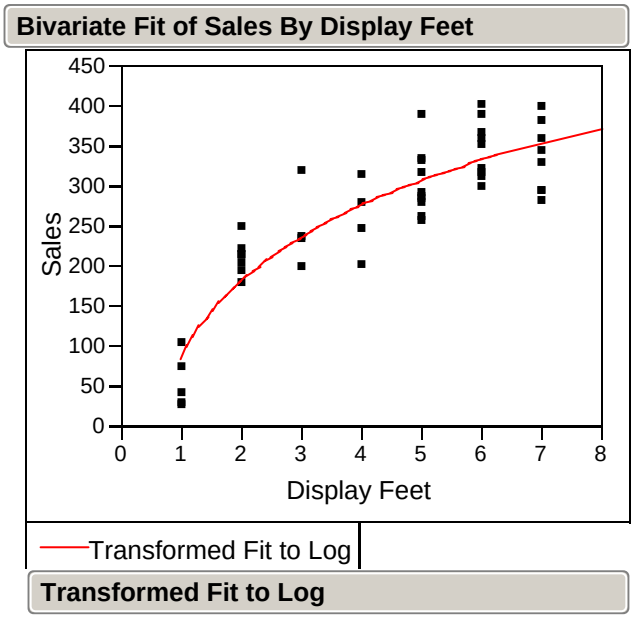
Letting y^* denote the (possibly) transformed y, and letting x^* denote the (possibly) transformed x, JMP simply fits a least squares regression line of the form

$$y^* = b_0 + b_1 x^*$$

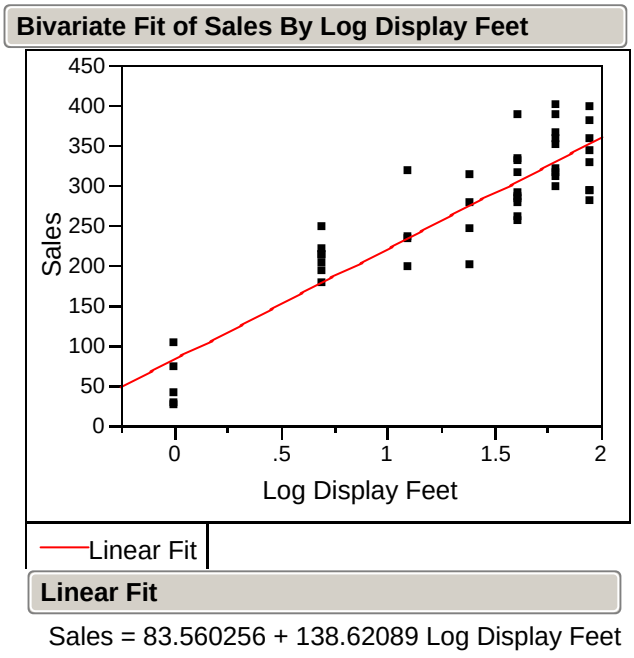
This is just a simple linear regression of y^* on x^* !!

[12] The "bulging rule" heuristic shown on page 15 of BAR offers some suggestions for which transformations might work given your impression of the curvature of the data.

For example, the previous fit of $y = b_0 + b_1 \log x$ on the display data



is really just a fit of $y = b_0 + b_1 x^*$ where $x^* = \log x$



Strategy: A successful transformation will yield a scatterplot that shows linear association between y^* and x^*.

Further BAR Examples

1) Managing Benefit Costs - Insure.jmp (p 23)

"Will paying insurance costs entice labor into a contract?"

Plot of cost of life insurance on age of beneficiary
Quadratic pattern (up, then down cannot be captured by transformations such as logs)
Linear suggests continuing up, missing structure of data
Better equation suggests otherwise

2) Predicting Cellular Phone Use - Cellular.jmp (p 29)

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

Time series
Remarkable pattern (p 32) that initially looks great, but misses a lot under closer scrutiny (p 55) and would mistakenly predict a drop in rates if naively used.

Take-Away Review

A regression line offers a summary of the relationship between a predictor (called *x*) and a response (called *y*).

A regression line is an estimate of the average of *y* at each value of *x*

Transformations to new coordinates (such as through logs) allow regression to capture nonlinear patterns as well.

Next Module

Where are the statistics and p-values? Inference?

These ideas require a *model* for the underlying sampling process and populations.

Department of Statistics
The Wharton School
University of Pennsylvania

Statistics 621 Fall 2005

Module 2

The Simple Regression Model

The Simple Regression Model (SRM) is an idealized statistical model, under which the data

$$(x_1, y_1), \ldots, (x_n, y_n)$$

are treated as a realization of

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

$$\varepsilon_1, \ldots, \varepsilon_n \text{ iid} \sim N(0, \sigma_\varepsilon^2)$$

Pictorially:

β_0 , β_1 and σ_ε are the (usually) unknown parameters of the SRM. An objective of regression analysis is to estimate them. Once we know these three parameters, we have a complete description of an idealized process that produced the data.

It may be useful to think of SRM data as the sum of two components[1]:

signal:

noise[2]:

Think of the SRM as a hypothetical "data generating process" that could have produced the data.

Example

To get a feel for how the SRM generates data, the file Utopia.jmp contains a simulation of pairs

$$(x_1, y_1), \ldots, (x_n, y_n)$$

from a SRM with[3] $\beta_0 = 7$, $\beta_1 = .5$ and $\sigma_\varepsilon = 1$

---

[1] The terminology "signal" and "noise" originated in electrical engineering. The methods we are studying can also be used to improve the reception of a TV or radio station. The goal of engineers is to transmit a clear signal from the station, one free of noise. For us, signal is an underlying structure that we seek to separate from random noise.

[2] People sometimes refer to $\varepsilon_1,\ldots, \varepsilon_n$ as the "errors". You can also think of $\varepsilon_1,\ldots, \varepsilon_n$ as coming from all of the other factors that influence the response aside from the one that we have chosen to highlight in the simple regression.

[3] The simulation is determined by the formulas that define the y and error columns. Note that it is necessary to Unhide the error column to see its formula.

What are the interpretations of $\beta_0 + \beta_1 x$, β_0 , β_1 and σ_ε in the SRM?

The familiar iid normal model that we saw in Stat 603

$$y_1, \ldots, y_n \ \ iid \ \sim N(\mu_y, \sigma_y^2)$$

is a special case of the SRM when $\beta_0 =$, $\beta_1 =$ and $\sigma_\varepsilon =$

Note the three main properties of $\varepsilon_1, \ldots, \varepsilon_n$ are precisely those that identify an *iid* sample of normal observations:

a) independence
b) equal variance σ_ε^2
c) normally distributed

Typical Regression Situation

The general course of a regression analysis includes these steps:

Figure out for your problem if it makes sense to think of one variable as a predictor, and one as a response.

Observe pairs of data, $(x_1, y_1), \ldots, (x_n, y_n)$

Plot the data!

If necessary, transform the data to obtain linear association

Suspect (or hope) SRM assumptions are justified

Estimate the "true" regression line

$$y = \beta_0 + \beta_1 x$$

by the LS regression line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ to denote b_0 and b_1 from Module 1.

WARNING! The true regression line and the LS regression line are different. DON'T CONFUSE THEM!

Pictorially

Jargon: $\hat{\beta}_0$ and $\hat{\beta}_1$ are often referred to as the *least squares (LS) estimates* of β_0 and β_1.

The Fitted Values and the Residuals

The LS regression line decomposes the data into two parts

$$y_i = \hat{y}_i + e_i$$

where

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad \text{and} \quad e_i = y_i - \hat{y}_i$$

Pictorially

Jargon (again)

$\hat{y}_1, \ldots, \hat{y}_n$ are called the *fitted* or *predicted values*

$e_1, \ldots, e_n$ are called the *residuals*

The following page shows the fitted values and the residuals for the Module 1 diamond regression[4].

[4] After executing the Fit Line subcommand, JMP will store the fitted values and residuals in the data table by right clicking next to "—Linear Fit" and selecting Save Predicteds and Save Residuals form the Pop-up menu.

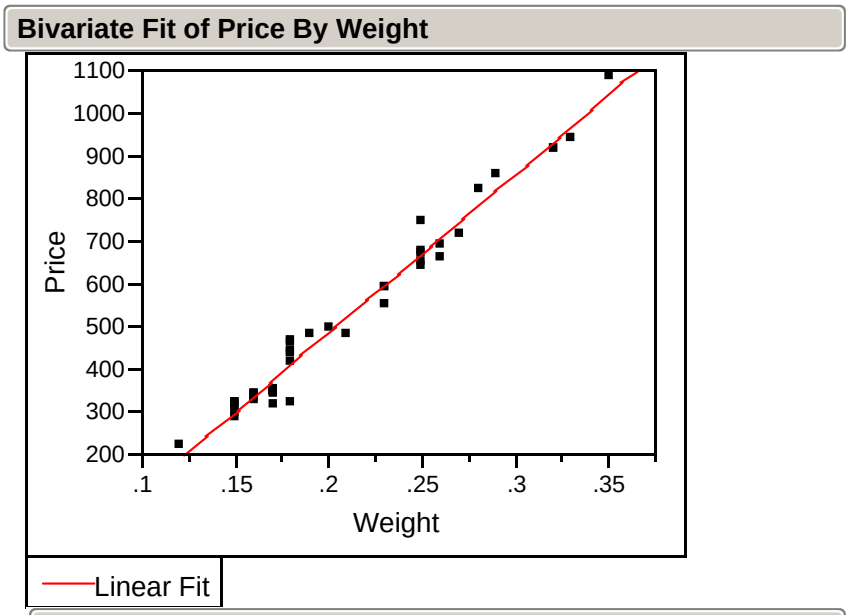
Weight	Price	Predicted Price	Residuals Price
0.17	355	372.95	-17.95
0.16	328	335.74	-7.74
0.17	350	372.95	-22.95
0.18	325	410.16	-85.16
0.25	642	670.63	-28.63
0.16	342	335.74	6.26
0.15	322	298.53	23.47
0.19	485	447.37	37.63
0.21	483	521.79	-38.79
0.15	323	298.53	24.47
0.18	462	410.16	51.84
0.28	823	782.26	40.74
0.16	336	335.74	0.26
0.2	498	484.58	13.42
0.23	595	596.21	-1.21
0.29	860	819.47	40.53
0.12	223	186.90	36.10
0.26	663	707.84	-44.84
0.25	750	670.63	79.37
0.27	720	745.05	-25.05
0.18	468	410.16	57.84
0.16	345	335.74	9.26
0.17	352	372.95	-20.95
0.16	332	335.74	-3.74
0.17	353	372.95	-19.95
0.18	438	410.16	27.84
0.17	318	372.95	-54.95
0.18	419	410.16	8.84
0.17	346	372.95	-26.95
0.15	315	298.53	16.47
0.17	350	372.95	-22.95
0.32	918	931.10	-13.10
0.32	919	931.10	-12.10
0.15	298	298.53	-0.53
0.16	339	335.74	3.26
0.16	338	335.74	2.26
0.23	595	596.21	-1.21
0.23	553	596.21	-43.21
0.17	345	372.95	-27.95
0.33	945	968.31	-23.31
0.25	655	670.63	-15.63
0.35	1086	1042.73	43.27
0.18	443	410.16	32.84
0.25	678	670.63	7.37
0.25	675	670.63	4.37
0.15	287	298.53	-11.53
0.26	693	707.84	-14.84
0.15	316	298.53	17.47

Note that the decomposition $y_i = \hat{y}_i + e_i$ holds.

Root Mean Squared Error (*RMSE*) – An Estimate of σ_ε

Looking at more of the output[5] from the diamond regression, a key quantity of interest is the

Root Mean Square Error (*RMSE*) = 31.84



Linear Fit

Price = -259.6259 + 3721.0249 Weight

Summary of Fit

R-square is the portion of variation attributed to the model, between 0 and 1. Root Mean Squared Error "RMSE" estimates the standard deviation of the residual.

RSquare	0.978261
RSquare Adj	0.977788
Root Mean Square Error	31.84052
Mean of Response	500.0833
Observations (or Sum Wgts)	48

RMSE estimates σ_ε, and is often called the *standard deviation of the residuals*.

[5] When the "Show Explanations" box is checked under File> Preferences > General, the JMP will display explanations such the two sentences in Summary of Fit table.

The formula for *RMSE* is

$$RMSE = \sqrt{\frac{1}{n-2}\sum (y_i - \hat{y}_i)^2}$$

How does this formula compare to the formula for s_y, the sample standard deviation of y?

$RMSE^2$ is the "average" squared deviation between the data and the LS regression line (i.e. the variance of the residuals).

We divide by $(n - 2)$ instead of n to compensate for the fact that the LS line obtains smaller sum of squared deviations than the true regression line[6].

RMSE measures the dispersion of the data around the LS regression line. Why is this value important in the regression?

If the SRM holds, then approximately

of the data will lie within *one RMSE* of the LS line

of the data will lie within $2 \times RMSE$ of the LS line

[6] The quantity $(n - 2)$ here is sometimes called the degrees of freedom (df) and is often used in regression calculations.

Model Checking

Any conclusions drawn from a regression analysis depend on the assumption that the SRM is appropriate.

Good statistical practice entails using the data to make sure there are no gross violations of the SRM.

What to look for:

1) Is the relationship between x and y linear?

2) Are there outliers or influential values that distort the model fit?

3) Do the residuals manifest *iid* normal behavior? (i.e., independent, constant variance, normal)

Three crucial model checks:

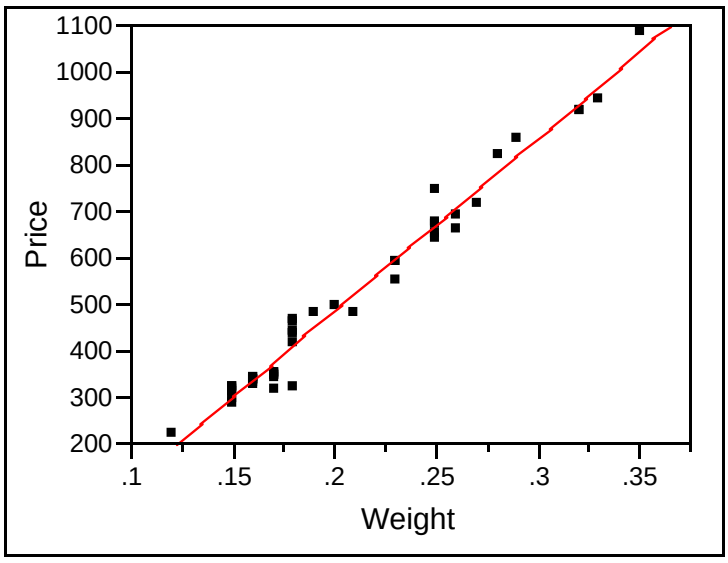
1. A scatterplot of y vs x should reveal

2. A scatterplot of the Residuals vs x should appear

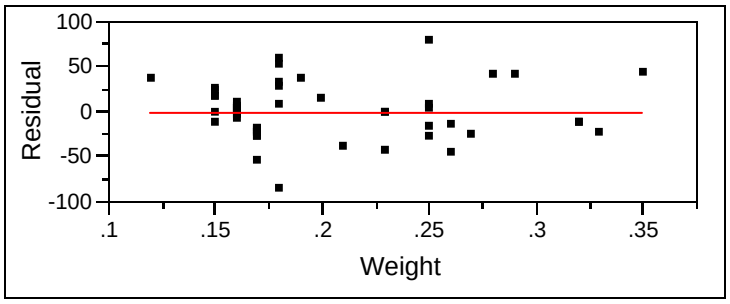
3. A histogram and normal quantile plot of the residuals should be consistent with the assumption of normality of the errors.

Example: Checking the Diamond Regression

The scatterplot of Price vs Weight

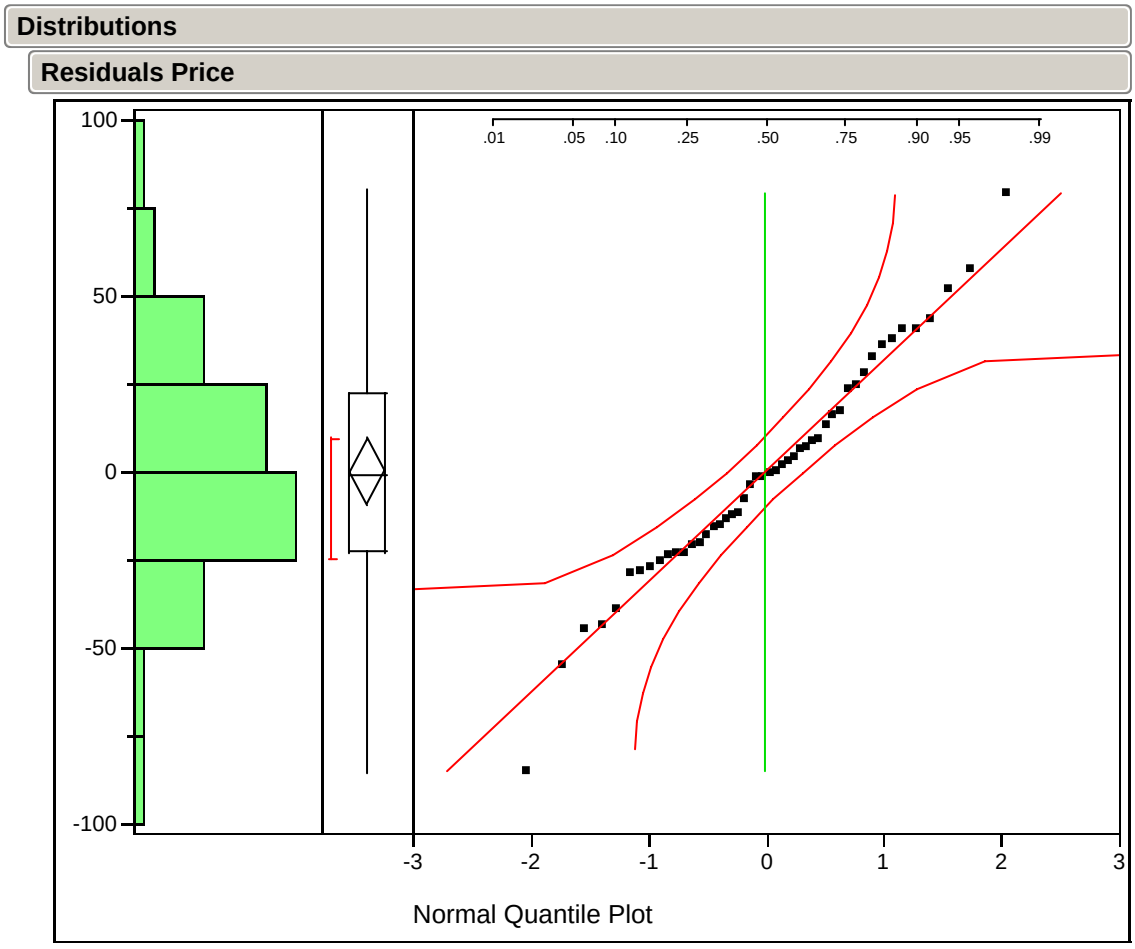


The scatterplot[7] of Residuals vs Weight



[7] After executing the Fit Line subcommand, right click on the triangle next to "—Linear Fit" and select Plot Residuals from the Pop-up menu to obtain this plot.

The histogram and normal quantile plot of the residuals

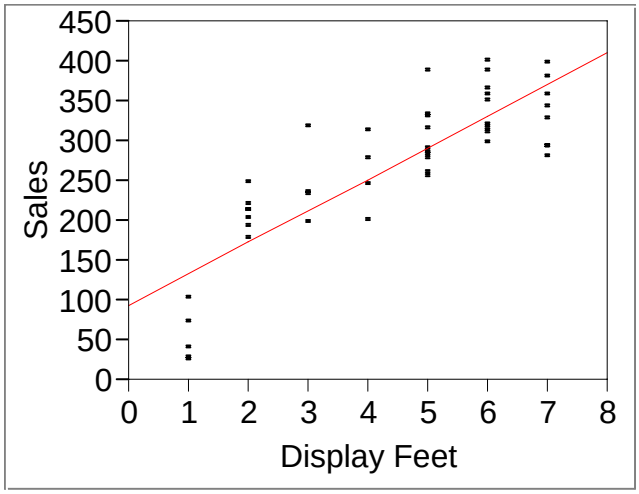


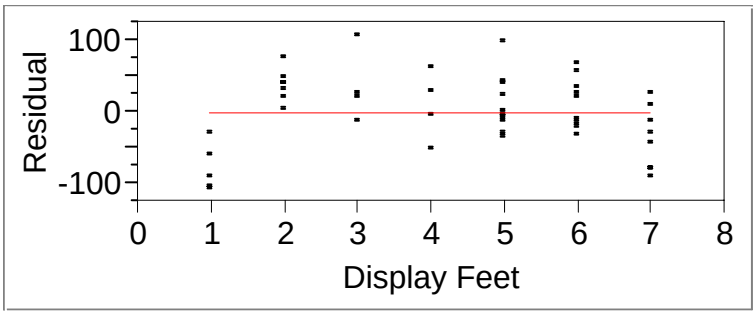
Anomalies to Look For

Nonlinearity

Can be revealed by the *y* vs *x* scatterplot or by the Residuals vs *x* scatterplot

Recall the display.jmp data





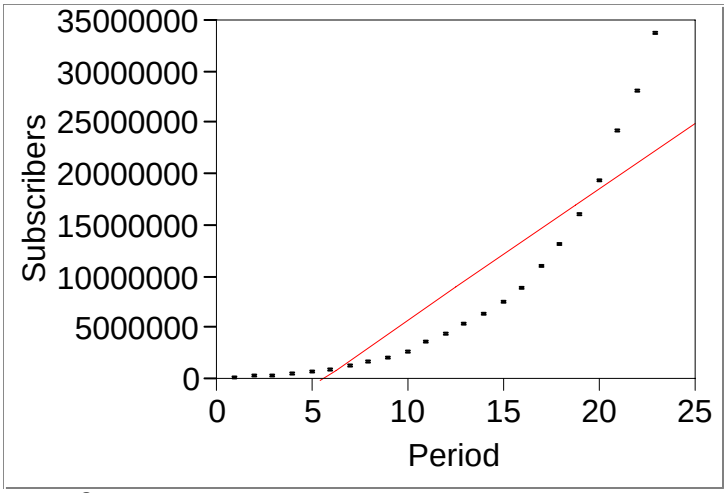
Remedy: Transform y and/or x.

Autocorrelated Residuals (BAR, *p* 29)

The file cellular.jmp contains the number of subscribers to cellphone service in the US every six months from the end of 1984 to the end of 1995.

The data is a time series $y_1,\ldots,y_n$ where y_t is the number of subscribers at time period t.

A scatterplot of y vs t (i.e. a time series plot of y) shows nonlinear growth in the number of subscribers.



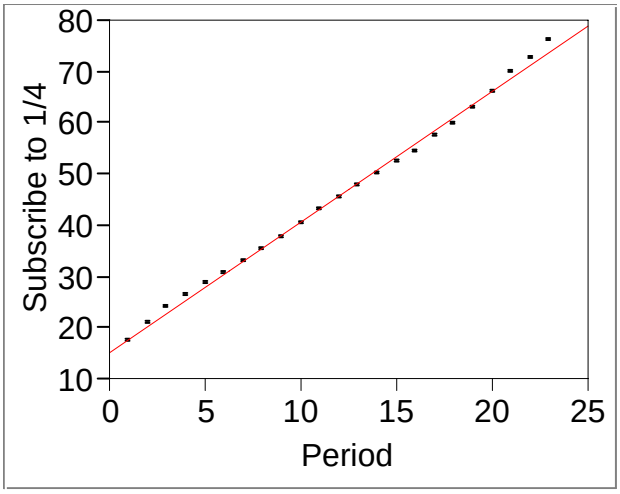
By trial and error[8], one discovers that the transformation $y^* = y^{1/4}$ yields what appears to be an ideal linear relationship.

Thus one might consider fitting a trend model of the form

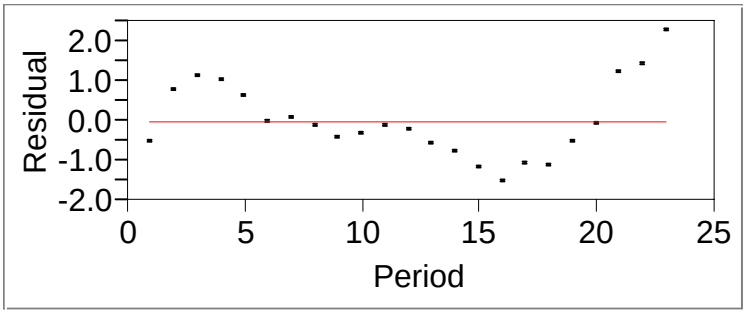
$$y_t^{1/4} = \beta_0 + \beta_1 t + \varepsilon_t , \quad t = 1,\ldots, n$$

In this special case of the SRM, t plays the role of x.
At first glance the regression of $y^{1/4}$ on t appears to be wonderful.

[8] As described in Lecture 1 of BAR, p 29-38.



However, the scatterplot of residuals vs *t* reveals a serious problem.

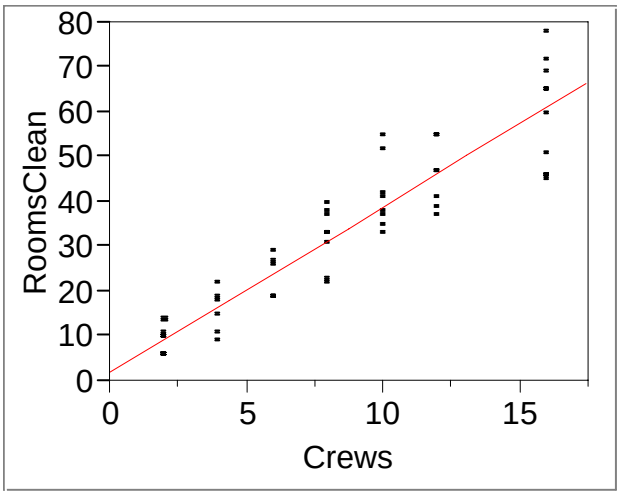


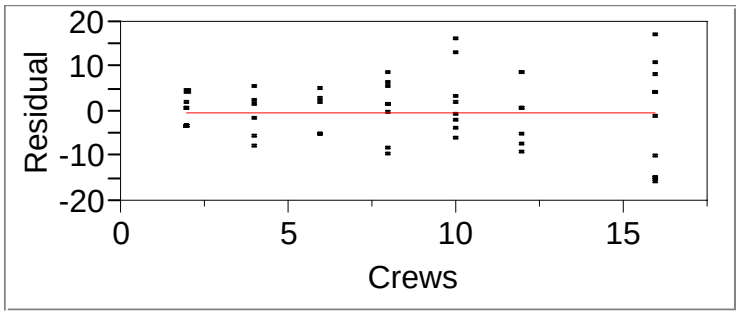
What SRM assumption has been violated?

Such meandering residuals are often called autocorrelated because $e_{t\text{-}1}$ and e_t appear correlated.

Expanding Residuals (BAR, p7)

The file cleaning1.jmp contains the number of crews (Crews) and the number of rooms cleaned (RoomsClean) for 53 teams of building maintenance workers.





Which assumption of the SRM is violated here?

This violation has only a minor effect on the estimation of β_0 and β_1. However, it does affect the prediction statements to be discussed in Module 3.

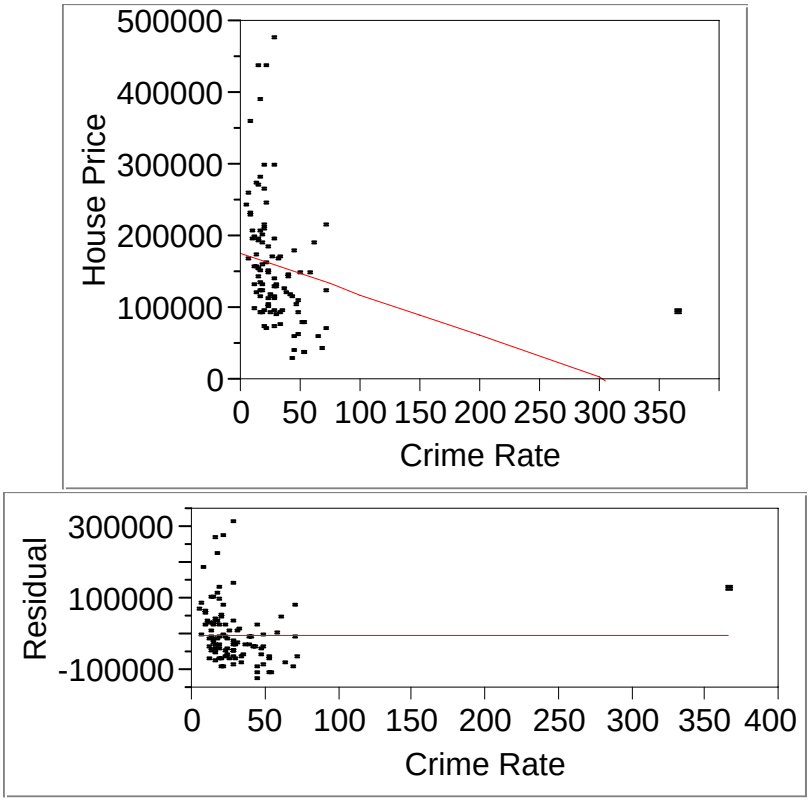
Remedy: Transform *y* or use weighted least squares (BAR, p 57-60) instead of least squares.

Outliers and Influential Points (BAR, *p* 62)

Main idea: outliers are unusual points. They should always be investigated. If warranted, they should be excluded.

The file phila.jmp contains the average prices of houses sold in the prior year and crime rates for 110 Pennsylvania communities in and near Philadelphia in April 1996.

To gauge the relationship between house prices and crime rates, one might consider the following regression

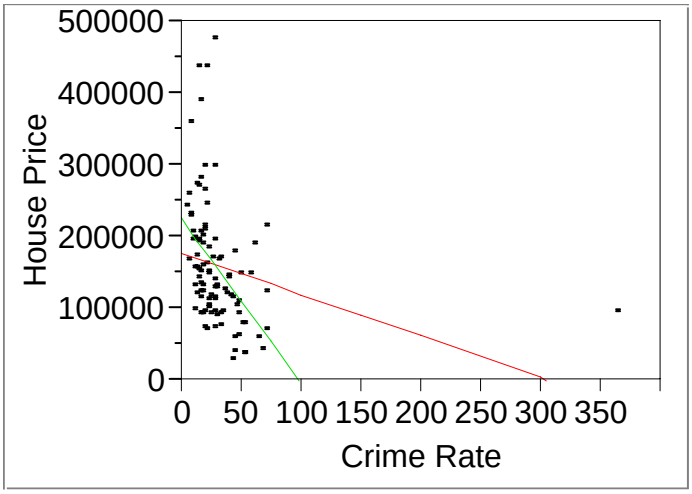


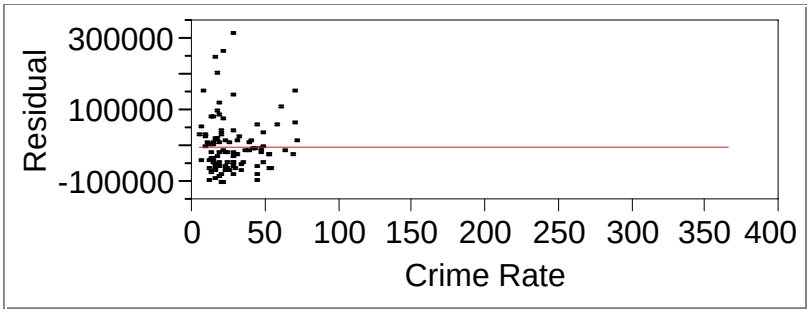
The LS line here is *House Price = 176629 – 577 Crime Rate*

with RMSE = 84325. Interpretation?

The unusual point is[9]

Repeating the regression with the unusual point excluded yields





The LS line here is

House Price = 225234 – 2289 Crime Rate

with RMSE = 78861. How does this fit change the implications of the previous model?

Should this point be excluded?

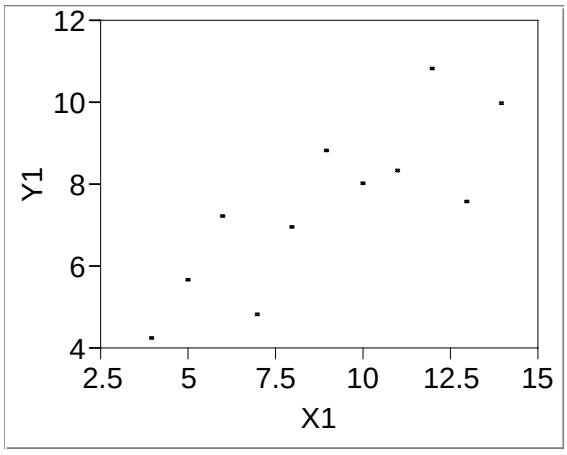
Remark: Here, one point drastically influenced the regression. However, this is not always the case. See the direct marketing example in BAR p 72-77.

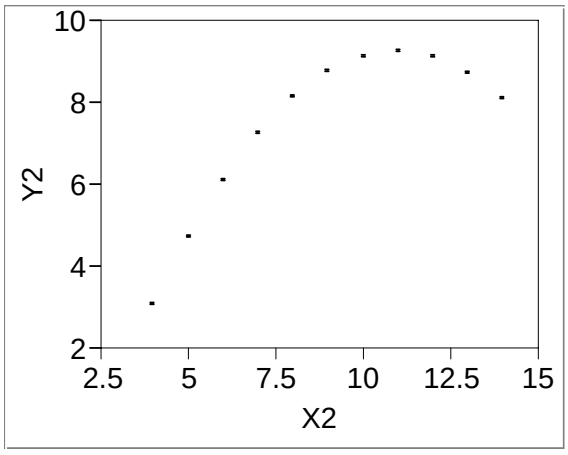
[9] Point labels are very helpful when it comes to identifying outliers. The default point label is the row number in the JMP data set. You can assign a variable to be the label as well.

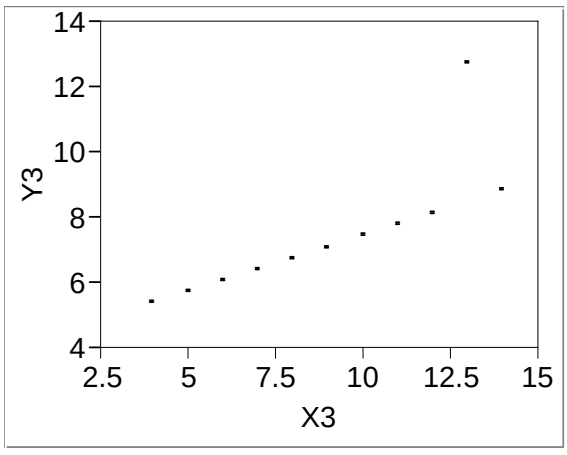
Don't forget to plot the data!

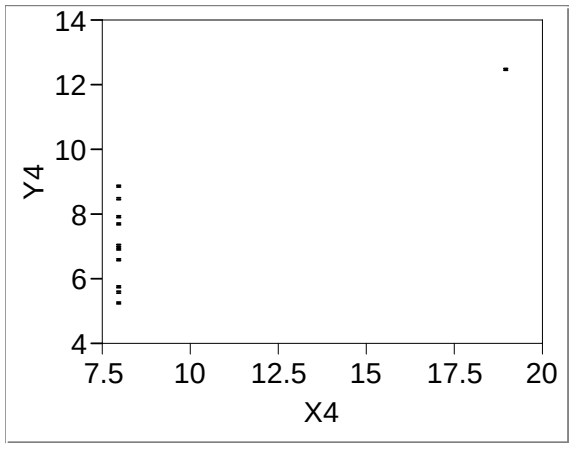
Before fitting a regression, it is crucial to first plot the data.

Example: Which of the following four data sets seems compatible with the SRM assumptions?



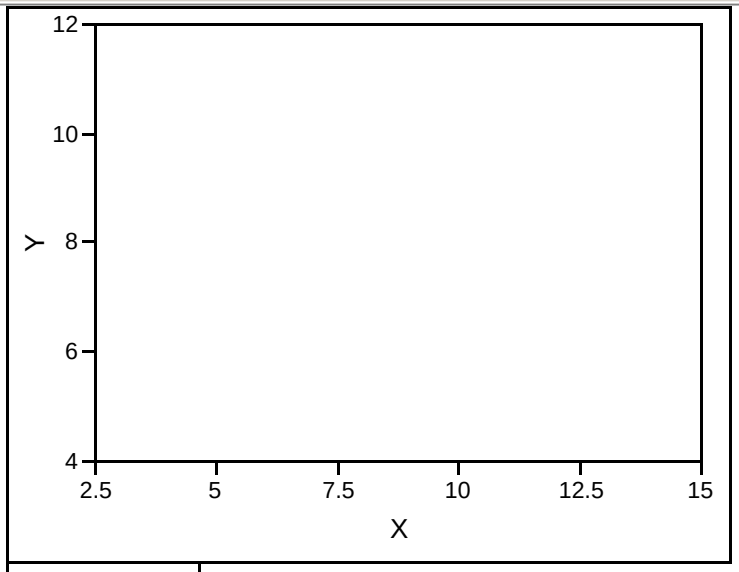






Which of the previous scatter plots yields the following regression output?

Bivariate Fit of Y By X



Linear Fit

Linear Fit

Y = 3.0000909 + 0.5000909 X

Summary of Fit

RSquare	0.666542
RSquare Adj	0.629492
Root Mean Square Error	1.236603
Mean of Response	7.500909
Observations (or Sum Wgts)	11

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	27.510001	27.5100	17.9899
Error	9	13.762690	1.5292	Prob > F
C. Total	10	41.272691		0.0022

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.0000909	1.124747	2.67	0.0257
X	0.5000909	0.117906	4.24	0.0022

Take-Away Summary

The simple regression model (SRM) is the basis for inference from regression with one predictor. In this model, the observed data

$$(x_1, y_1), \ldots, (x_n, y_n)$$

are assumed to be a realization of a "signal + noise" data generating process that ideally has the form

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

$$\varepsilon_1, \ldots, \varepsilon_n \text{ iid} \sim N(0, \sigma_\varepsilon^2)$$

Important diagnostics to keep in mind are plots that check for

Outliers
Linearity
Independence when the data are ordered (particularly when the data are a time series)
Equal variance
Normality

Next Module

The SRM is the basis for confidence intervals, prediction intervals, and hypothesis tests.

Department of Statistics
The Wharton School
University of Pennsylvania

Statistics 621 Fall 2005

Module 3

Inference about the SRM

Mini-Review: Inference for a Mean

An ideal setup for inference about a mean assumes normality,

$$y_1,\ldots,y_n \text{ iid } \sim N(\mu_y,\sigma_y^2)$$

and $\bar{y}$ and s_y are used to estimate μ_y and σ_y.

Fact: The sampling distribution of $\bar{y}$ is normal with mean μ_y

Fact: The standard error of $\bar{y}$ is $SE(\bar{y}) = s_y/\sqrt{n}$

Based on these facts we saw that

$\bar{y} \pm 2\,SE(\bar{y})$ are approx 95% CI limits for μ_y

For testing H_0: $\mu_y = c$ vs H_1: $\mu_y \neq c$, t ratio = $(\bar{y} - c)/SE(\bar{y})$

if $|t \text{ ratio}| > 2$ *or* p-value < .05 *or* 95% CI does not contain c, reject H_0 at the .05 level of significance.

Sampling Distributions and Standard Errors in Regression

For data $(x_1, y_1),\ldots, (x_n, y_n)$ generated with the SRM,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1,\ldots, n$$
$$\varepsilon_1,\ldots,\varepsilon_n \text{ iid} \sim N(0, \sigma_\varepsilon^2)$$

$\hat{\beta}_0, \hat{\beta}_1$ and $RMSE$ are used to estimate β_0, β_1 and σ_ε.

Fact: The sampling distributions[1] of $\hat{\beta}_0$ and $\hat{\beta}_1$ are normal[2] with means β_0 and β_1

> How could you use the simulation in utopia.jmp to generate these sampling distributions?[3]

Fact: The standard error of $\hat{\beta}_1$ is

$$SE(\hat{\beta}_1) \approx \frac{RMSE}{\sqrt{n}} \times \frac{1}{s_x}$$

$SE(\hat{\beta}_0)$, the standard error of $\hat{\beta}_0$, is given by a similar formula[4].

Note that $SE(\hat{\beta}_1)$ decreases as s_x increases. Does this make sense?

---

[1] More precisely, this fact refers to the sampling distribution when $y_1,\ldots,y_n$ vary, but the values of $x_1,\ldots,x_n$ are treated as fixed constants that are the same for every sample. Things work out similarly even if the x_i are random.

[2] These sampling distributions are approximately normal even if the errors $\varepsilon_1,\ldots,\varepsilon_n$ are not normally distributed.

[3] Think back to the M&M's from Stat 603!

[4] These standard errors are routinely computed by statistical software such as JMP.

Further Inference about β_0 and β_1

Confidence Intervals

Approximate 95% CI's for β_0 and β_1 are given by

$$\hat{\beta}_0 \pm 2SE(\hat{\beta}_0) \quad \text{and} \quad \hat{\beta}_1 \pm 2SE(\hat{\beta}_1)$$

Hypothesis Tests

For H_0: $\beta_0 = c$ vs H_1: $\beta_0 \neq c$,, t ratio = $\dfrac{\hat{\beta}_0 - c}{SE(\hat{\beta}_0)}$

For H_0: $\beta_1 = c$ vs H_1: $\beta_1 \neq c$,, t ratio = $\dfrac{\hat{\beta}_1 - c}{SE(\hat{\beta}_1)}$

If $|t \text{ ratio}| > 2$ *or* p-value < .05 *or* 95% CI does not contain c, reject H_0 at the .05 level of significance.

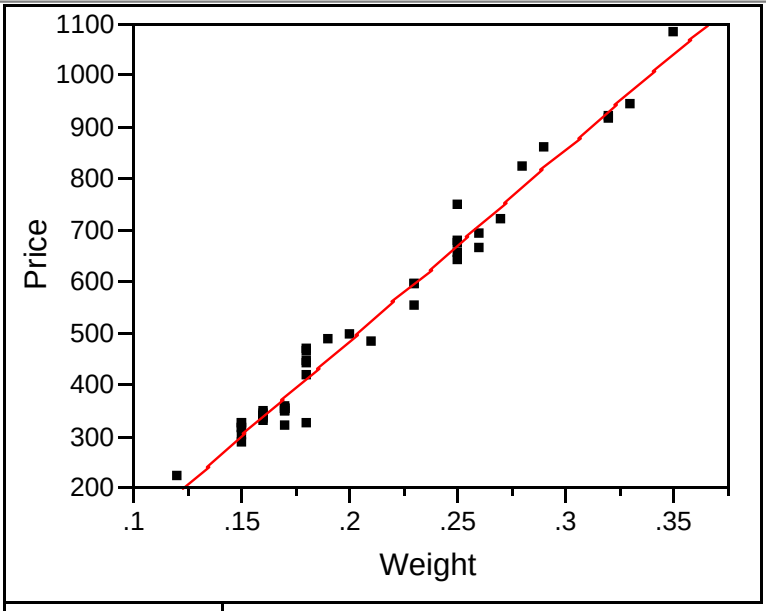
H_0: $\beta_1 = 0$ is the usual null hypothesis of interest. What does it say about the relationship between predictor and response if this null hypothesis is true?

JMP provides the details: estimates, SE's, *t* statistics, and p-values for inference about β_0 and β_1 arranged in a table, with one line for the intercept and one for the slope.

Example

Consider inference for β_0 and β_1 in the diamond regression.

Bivariate Fit of Price By Weight



Linear Fit

Linear Fit

Price = -259.6259 + 3721.0249 Weight

Summary of Fit

RSquare	0.978261
RSquare Adj	0.977788
Root Mean Square Error	31.84052
Mean of Response	500.0833
Observations (or Sum Wgts)	48

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-259.6259	17.31886	-14.99	<.0001
Weight	3721.0249	81.78588	45.50	<.0001

Here, β_0 and β_1 are estimated by

$$\hat{\beta}_0 \approx -259.6 \qquad \text{and} \qquad \hat{\beta}_1 \approx 3721.0$$

The standard errors of these estimates are

$$SE(\hat{\beta}_0) \approx 17.3 \qquad \text{and} \qquad SE(\hat{\beta}_1) \approx 81.8$$

Approximate 95% confidence interval limits for β_0 and β_1 are

$$-259.6 \pm 2\ (17.3) \qquad \text{and} \qquad 3721.0 \pm 2\ (81.8)$$

Should the hypothesis H_0: $\beta_1 = 0$ be rejected?[5]

Yes, because $t = 45.5 > 2$ or because p-value < .0001

Should the hypothesis H_0: $\beta_1 = 3800$ be rejected?

No, because $|t| = |3721.0 - 3800|/81.8 = .96 < 2$

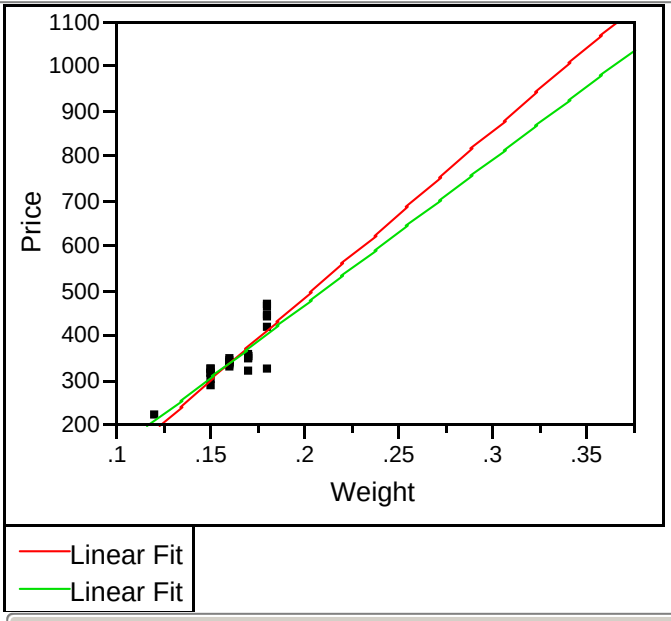
Why is it interesting that H_0: $\beta_0 = 0$ can be rejected?

[5] You can also answer these questions by using a confidence interval for the slope. For example, this question asks whether 0 lies inside the 95% confidence interval for the slope.

Suppose that instead of having the complete diamond.jmp data set, we only had the observations for which Weight ≤ .18.

The LS regression with these 28 observations yields[6]

Bivariate Fit of Price By Weight



Linear Fit

Price = -182.2635 + 3247.0255 Weight

Summary of Fit

RSquare	0.687211
RSquare Adj	0.67518
Root Mean Square Error	30.50833
Mean of Response	348.8571
Observations (or Sum Wgts)	28

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-182.2635	70.50896	-2.58	0.0157
Weight	3247.0255	429.6156	7.56	<.0001

How has the regression output changed?

[6] To exclude the observations for which Weight > .18, use the command: Row Selection > Select Where..., and then Exclude the selected rows.

Confidence Bands for the True Regression Line

"Where does the true population regression line lie?"

> "What is the *average* price for *all* diamonds of a chosen weight?"

After running a regression based on (x_1, y_1), …, (x_n, y_n), the point $(x, \hat{y}_x)$ on the LS regression line

$$\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 x$$

is an estimate of the corresponding point $(x, \mu_{y|x})$ on the true regression line

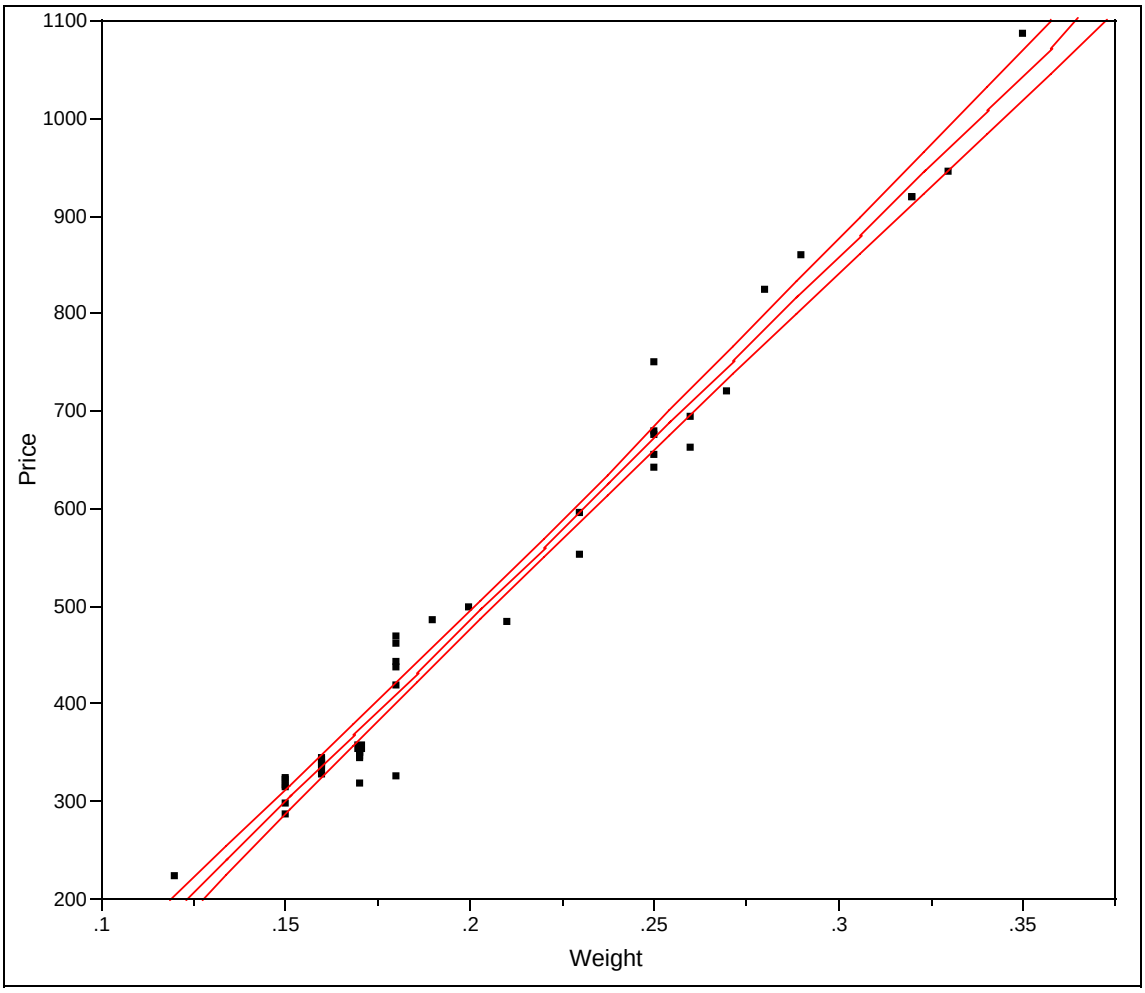
$$\mu_{y|x} = \beta_0 + \beta_1 x$$

Further inference about the true regression line is obtained using 95% Confidence Bands. Pictorially:

For each x value, the interval within the confidence band is a 95% Confidence Interval[7] for the (unknown) value of $\mu_{y|x}$.

[7] The estimate $\hat{y}_x$ is a statistic with a sampling distribution and a standard error SE($\hat{y}_x$) given by a complicated formula. This interval is approximately $\hat{y}_x \pm 2$ SE($\hat{y}_x$).

JMP provides a graph[8] of the 95% Confidence Bands for the true regression line. For example, for the diamond data regression, these are seen to be

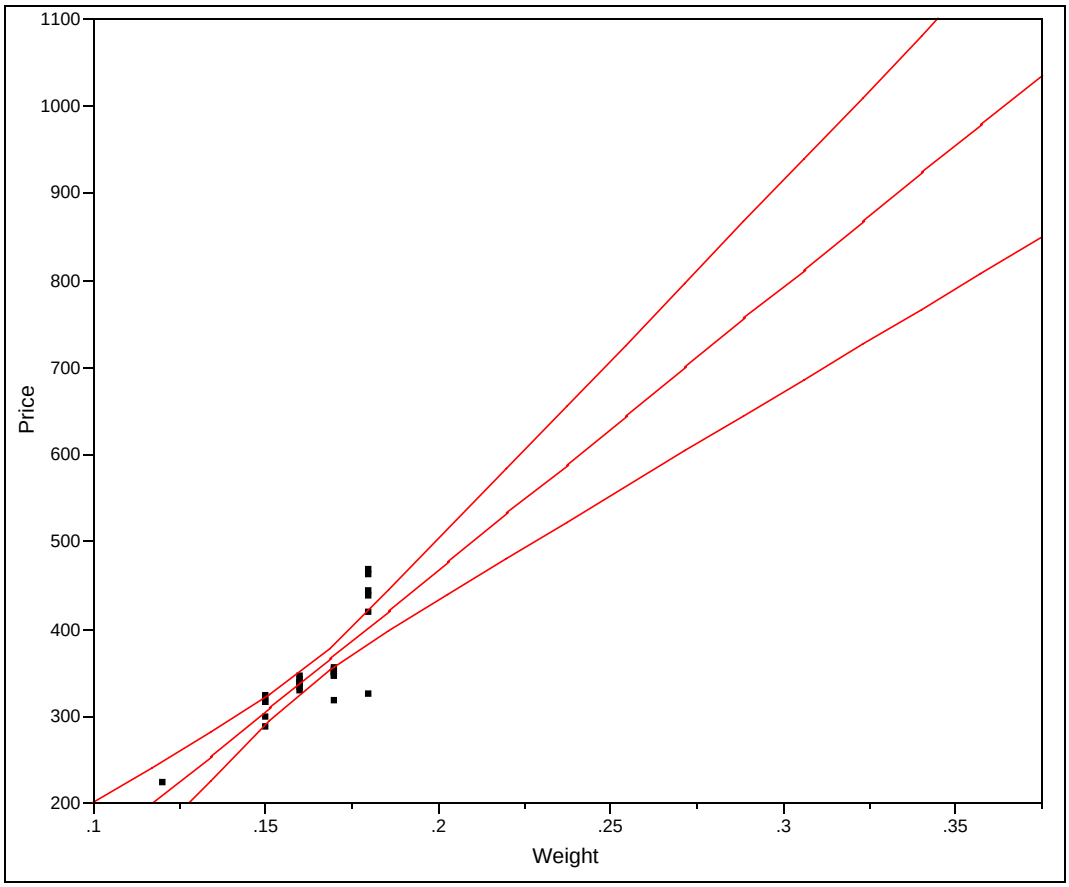


When Weight = .3, the interval between the confidence bands is (842, 875).[9] What is the interpretation of this interval?

[8] After executing the Fit Line subcommand, right click next to "—Linear Fit" and select Confid Curves Fit from the Pop-up menu to obtain this plot.

[9] A feature in JMP that you may wish to use to read off values from graphs (such as values of these confidence intervals), is obtained by changing the cursor to crosshair + in the Tools menu. Using this feature you can simply click on a point in a graph and read off its x and y coordinates.

If we had only the data for the smaller diamonds (Weight ≤ .18), the 95% Confidence Bands would be



What happens to the 95% Confidence Bands for the true regression line as x gets farther away from $\bar{x}$?[10]

This phenomenon can be thought of as a "Statistical Extrapolation Penalty".[11]

---

[10] For extrapolation beyond the range of the data graph, you may find it useful to first add a row with extreme values to increase the axes of a plot and then select and exclude it before doing the analysis.

[11] Even though the intervals widen as we extrapolate, as the output shows, this penalty is rather optimistic because it assumes that the model we have fit is correct for all values of x. If you price a big diamond with this model, you'll see that the interval is not nearly wide enough!

Prediction Bands for Future Values

"Where will a future value of the response *y* lie?"

"How much might I pay for a specific 1/4 carat diamond?"

After running a regression based on $(x_1, y_1), \ldots, (x_n, y_n)$, the point $(x, \hat{y}_x)$ on the LS regression line

$$\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 x$$

is a prediction of the future point (x, y_x) generated by the SRM

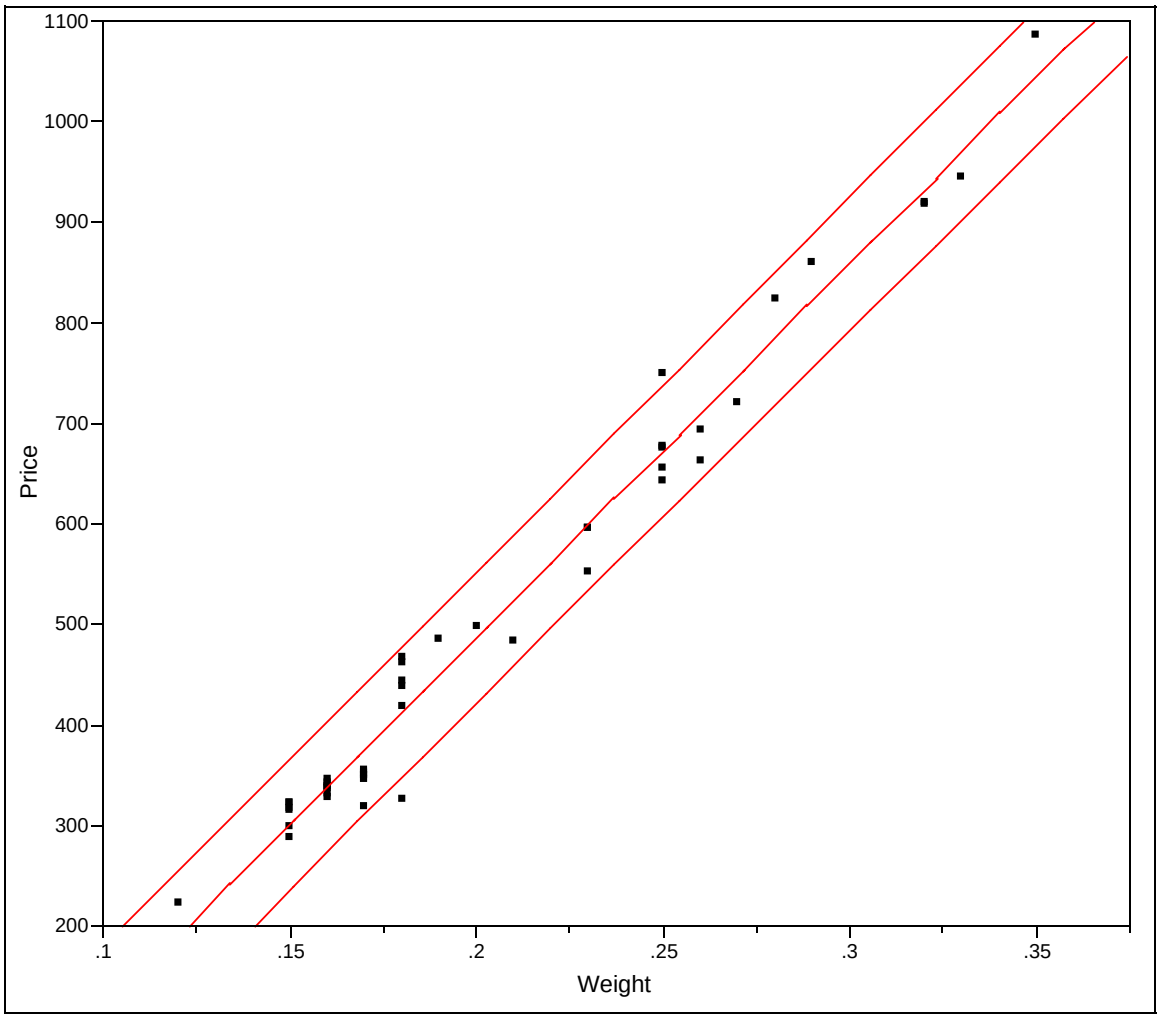
$$y_x = \beta_0 + \beta_1 x + \varepsilon_x$$

Further inference about future values is obtained using 95% Prediction Bands. Pictorially:

For each *x* value, the interval within the Prediction Band is a 95% Prediction Interval[12] for the (unknown) value of y_x.

[12]The variation of y_x is due of the variation of $\hat{y}_x$ and the variation of ε_x. These components are independent, and so the 95% prediction interval (PI) for y_x is approximately given by $\hat{y}_x \pm 2\sqrt{SE(\hat{y}_x)^2 + RMSE^2}$. When x is close to $\bar{x}$, an even simpler rule of thumb approximation $\hat{y}_x \pm 2RMSE$ works reasonably well.

JMP provides[13] a graph of the exact 95% Prediction Bands for y_x over the whole line. For the regression on the smaller diamond data, these prediction bands are seen to be

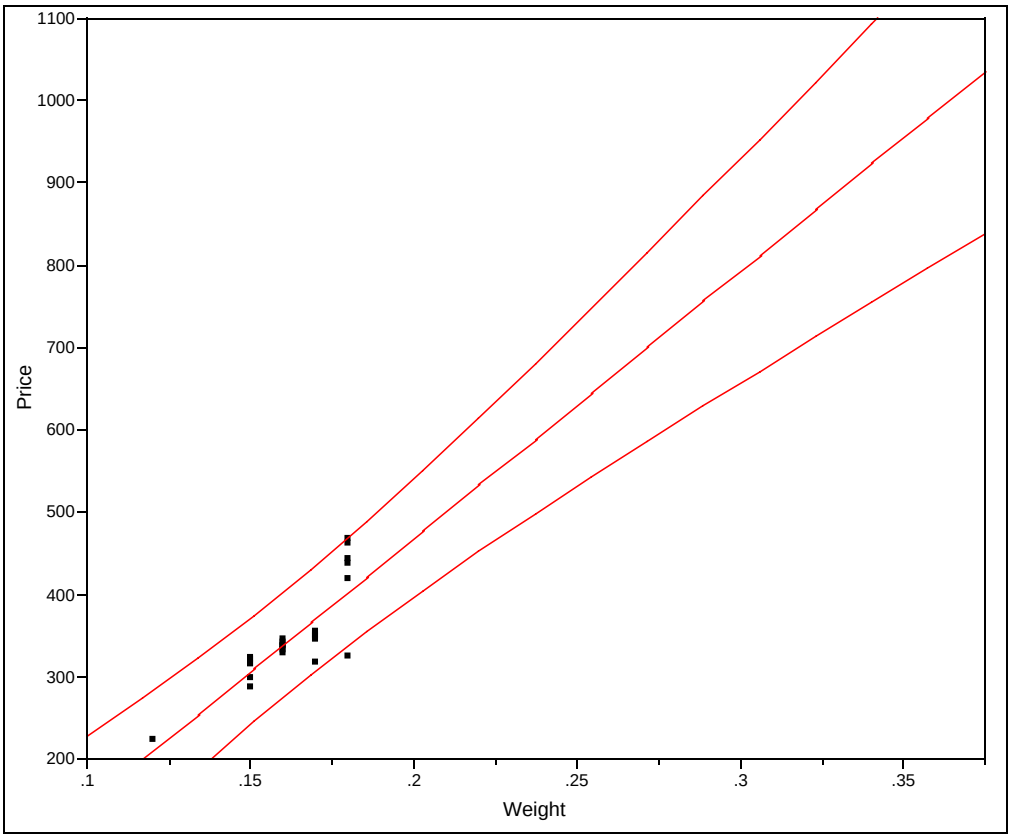


When Weight = .3, the interval between the prediction bands is (789, 923).[14] What is the interpretation of this interval?

[13] After executing the Fit Line subcommand, right click next to "—Linear Fit" and select Confid Curves Indiv from the Pop-up menu to obtain this plot.

[14] Again changing the cursor to Crosshairs +, you can simply click on a point in a graph and read off its x and y coordinates.

If we had only the data for the smaller diamonds (Weight ≤ .18), the 95% Prediction Bands would be



Note that the prediction bands are wider than the confidence bands on pgs 3-8 and 3-9. Why is this reasonable?

WARNING: Extrapolate with caution!

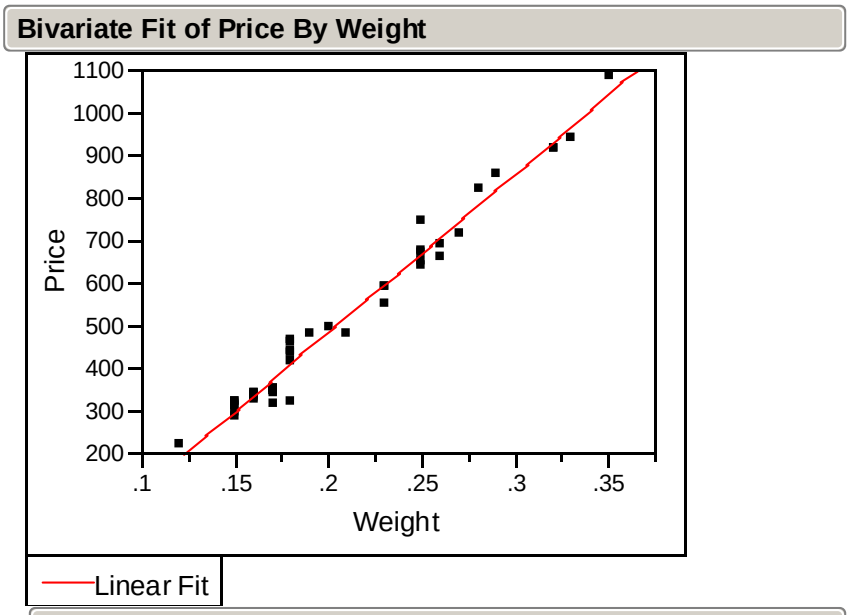
If x is not in the range of the data, predicting y_x is especially dangerous because the linear model may fail. Consider pricing the Hope Diamond (at 45.5 carats) with this model.

Another example: Average systolic blood pressure in people is approximated by $y \approx 118 + .0043\,x^2$ for $20 \le x \le 60$ where y = blood pressure and x = age. But when $x = 1000$, $y =$

R^2 Index of Performance

The next piece of output that we'll consider from the full diamond regression, is

RSquare = .978



Linear Fit

Price = -259.6259 + 3721.0249 Weight

Summary of Fit

R-square is the portion of variation attributed to the model, between 0 and 1. Root Mean Squared Error "RMSE" estimates the standard deviation of the residual.

RSquare	0.978261
RSquare Adj	0.977788
Root Mean Square Error	31.84052
Mean of Response	500.0833
Observations (or Sum Wgts)	48

Fact: In simple regression $R^2 = r^2$, where as in Stat 603, r is the sample correlation.

Beyond this fact, R^2 is widely interpreted as

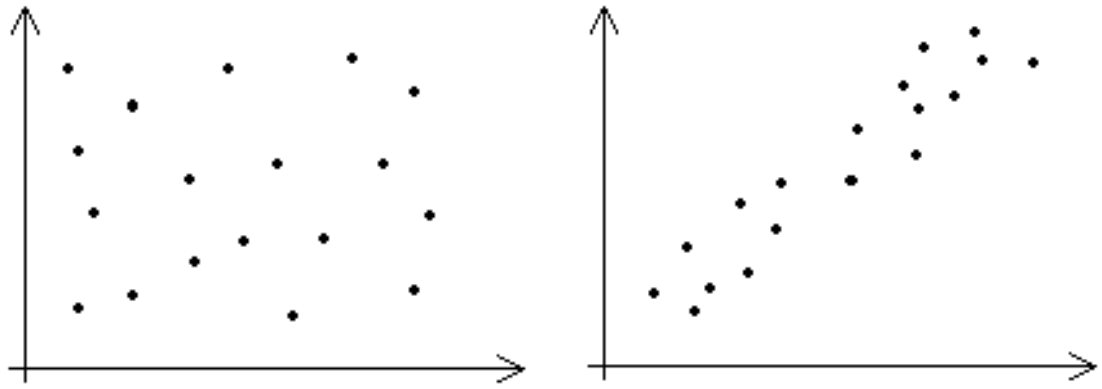
"the proportion of variation explained by the regression"

where variation refers to the variance of y, namely s_y^2

This interpretation is based on the fact that

$$R^2 \approx \frac{s_y^2 - RMSE^2}{s_y^2}$$

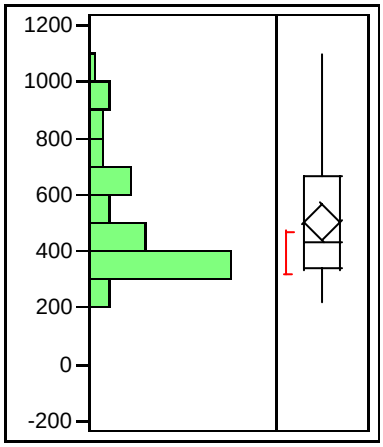
To see how this works, compare s_y^2, $RMSE^2$ and R^2 in the following two plots:



Here's another way to think about R^2: Notice how much less variation remains in the residuals after fitting the regression.

Distributions

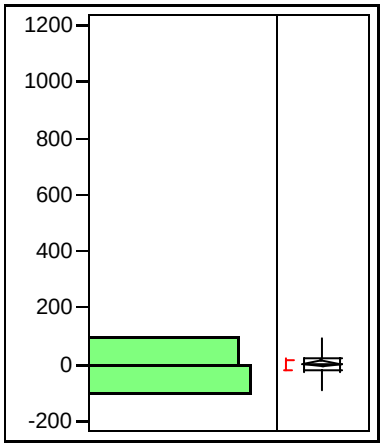
Price



Moments

Mean	500.08333
Std Dev	213.64276
Std Err Mean	30.836676
upper 95% Mean	562.11872
lower 95% Mean	438.04794
N	48
Sum Wgt	48
Sum	24004
Variance	45643.227
Skewness	1.0102859
Kurtosis	0.0706385
CV	42.721431
N Missing	0

Residuals Price



Moments

Mean	1.184e-15
Std Dev	31.499972
Std Err Mean	4.5466293
upper 95% Mean	9.1466385
lower 95% Mean	-9.146638
N	48
Sum Wgt	48
Sum	5.684e-14
Variance	992.24824
Skewness	0.0577218
Kurtosis	0.4471929
CV	2.6599e18
N Missing	0

R^2 captures the usefulness of using x to predict y. (p 92)

R^2 is often used as a measure of the "effectiveness" of a regression.

Advice: Resist the temptation to think of R^2 in absolute terms. Regression is a statistical tool for extracting information from data. Its value depends on the value of the information provided.

RMSE and R^2

From the fact on pg 3-14:

$$RMSE \approx s_y \sqrt{1 - R^2}$$

Does *RMSE* provide the same information as R^2?

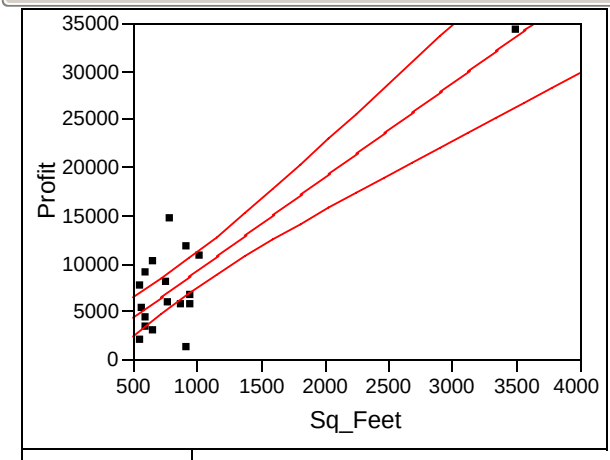
The Impact of Outliers: Another Example

Outliers can impact inferences from regression in a dramatic fashion.

Value of Housing Construction (BAR, p 89)

The data set cottage.jmp gives the profits obtained by a construction firm and the square footage of the properties.

The scatterplot shows that the firm has built one rather large "cottage". This is an "outlier" in the sense that it is very different from the rest of the points. (p 90)



Bivariate Fit of Profit By Sq_Feet

Linear Fit

Profit = -416.8594 + 9.7505469 Sq_Feet

Summary of Fit

RSquare	0.779667
RSquare Adj	0.765896
Root Mean Square Error	3570.379
Mean of Response	8347.799
Observations (or Sum Wgts)	18

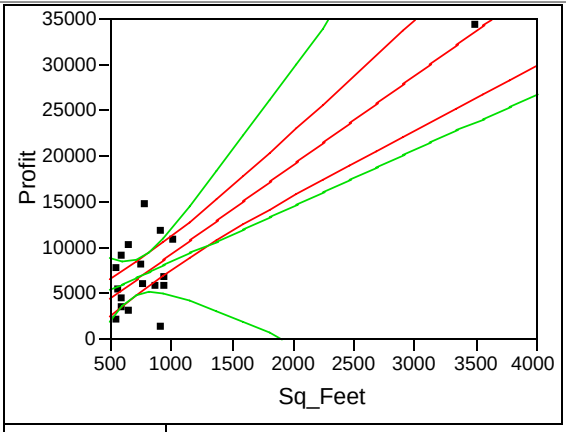
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-416.8594	1437.015	-0.29	0.7755
Sq_Feet	9.7505469	1.295848	7.52	<.0001

Without the Large Property

How does the fitted model change when we set aside the large cottage and refit the model without this one? (p 94)

Bivariate Fit of Profit By Sq_Feet



Linear Fit

Profit = 2245.4005 + 6.1370246 Sq_Feet

Summary of Fit

RSquare	0.075205
RSquare Adj	0.013552
Root Mean Square Error	3633.591
Mean of Response	6822.899
Observations (or Sum Wgts)	17

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2245.4005	4237.249	0.53	0.6039
Sq_Feet	6.1370246	5.556627	1.10	0.2868

What has happened to R^2? To *RMSE*? To the Confidence Bands for the slope?

Which version of this model should the firm use to estimate profits on the next large cottage it is considering building?

What additional information about these construction projects would you like to have in order to make a decision?

Leverage and Outliers

The dramatic effects of removing the outlying large "cottage" in this last example illustrates the impact of a *leveraged* outlier.

Leverage: points that are far from the rest of the data along the x-axis are said to be leveraged. BAR gives the formula for leverage on page 63.

Heuristic: Moving away from the center of the predictor impacts the possible effect of a single observation in regression much like moving your weight out to the end of a see-saw. As your weight moves farther from the fulcrum, you can lift more weight on the other side.

Leverage is a property of the values of the predictor, not the response.

Leveraged points are not necessarily bad and in fact improve the accuracy of your estimate of the slope.[15] Just recognize that you are giving some observations a bigger role than others.

[15] In particular, leveraged observations are those that contribute the most to the variation in the predictor. Since these points spread out the values of the predictor, they make it easier to estimate the slope of the regression.

Take-Away Review

Inference for regression benefits from the same sort of sampling distribution facts that made inference for means possible in Stat 603.

In particular, the sampling distribution of the slope estimate is normal with mean β_1 and standard error

$$SE(\hat{\beta}_1) \approx \frac{RMSE}{\sqrt{n}} \times \frac{1}{s_x}$$

So, we can form confidence intervals as before, forming the intervals as before, namely as [estimate ± 2 SE(estimate)].

We can use these same ideas to construct confidence intervals for the *average* of the response for any value of the predictor as well as for a specific response.

The R^2 summary measures the proportion of the variation of the response "explained" by the model; the *RMSE* measures the SD of the noise that remains.

Next Time

Getting more data gives you better estimates of the slope and intercept, but has little impact on the accuracy of prediction.

The only way to improve R^2 and reduce *RMSE* is to add more predictors. This is the domain of multiple regression.

Department of Statistics
The Wharton School
University of Pennsylvania

Statistics 621 Fall 2005

Module 4

The Multiple Regression Model

Example: Explaining and predicting fuel efficiency

The file car89.jmp contains many characteristics of various makes and models of cars. Variables include:[1]

> MPG City, Make/Model, Weight, Cargo, Seating, Horsepower, Displacement, Number of cylinders, Length, Headroom, Legroom, Price…

Questions of interest

> "What is the predicted mileage for a 4000 lb. new design, and which characteristics of the design are crucial?"[2]
>
> "How much does my 200 pound brother owe me for 3000 miles of riding with me?"

To get started, let's consider using simple regression to model the effect of Weight (lb) on MPG City

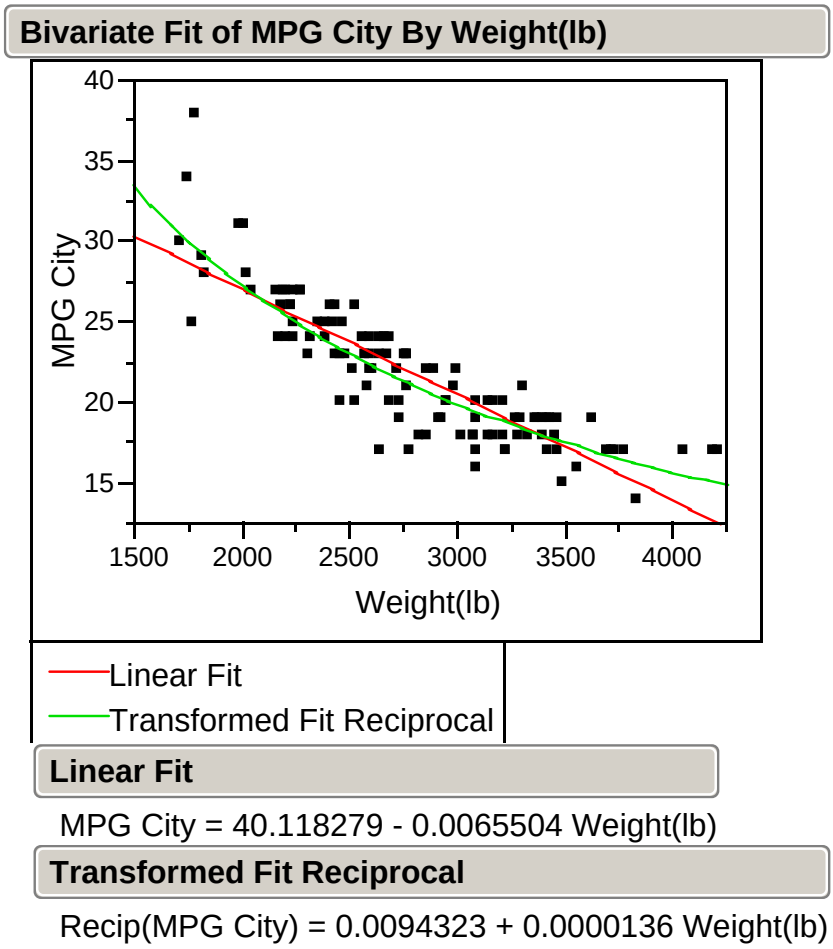
---

[1] See BAR, page 109, for more information on this dataset.

[2] Such questions of mileage are important to manufacturers that sell cars in the US. The so-called CAFE standards set requirements for the average fuel efficiency of the fleet of cars produced by a manufacturer.

Applying Fit Y by X, we consider the regression of MPG City on Weight(lb) (using Fit Line) and the regression of (p 110)

(1/MPG City) on Weight(lb)[3]



Which of these regressions seems more reasonable?

Do the signs of the slope coefficients make sense here?

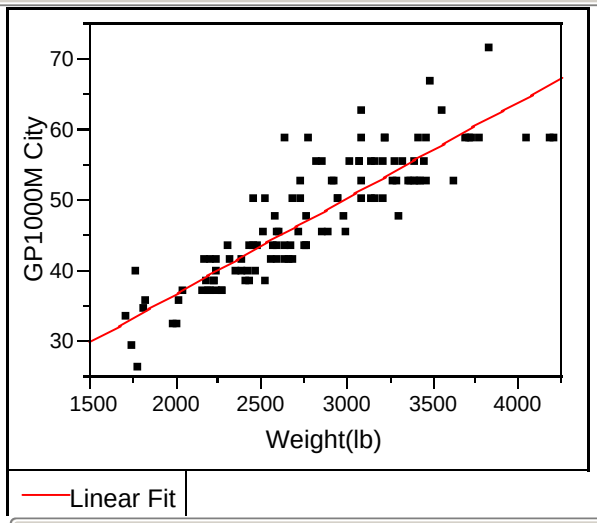
[3] Use the fit special dialog to get the reciprocal 1/Y of the response.

Based on the previous regressions we created a “new”, rescaled dependent variable[4]

$$GP1000M = 1000/MPG$$

The regression of GP1000M on Weight(lb) yields (p 111)

Bivariate Fit of GP1000M City By Weight(lb)



Linear Fit

GP1000M City = 9.4323391 + 0.0136159 Weight(lb)

Summary of Fit

RSquare	0.765268
RSquare Adj	0.763134
Root Mean Square Error	4.233195
Mean of Response	47.59468
Observations (or Sum Wgts)	112

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	9.4323391	2.054514	4.59	<.0001
Weight(lb)	0.0136159	0.000719	18.94	<.0001

What is the interpretation of the LS regression slope here?

Is the only difference between the BMW 735i and the Suzuki Swift just the 2000 pounds in weight?

[4] Multiplying 1/MPG by 1000 serves only to multiply the intercept and slope estimates by 1000 resulting in "friendlier" (and more impressive) regression output. What other easily motivated change of scales would make the slope 2000 times larger still?

Other factors obviously contribute as well. Let's use the Multivariate command[5] to explore the pairwise relationships between some of these. (p 115-116)

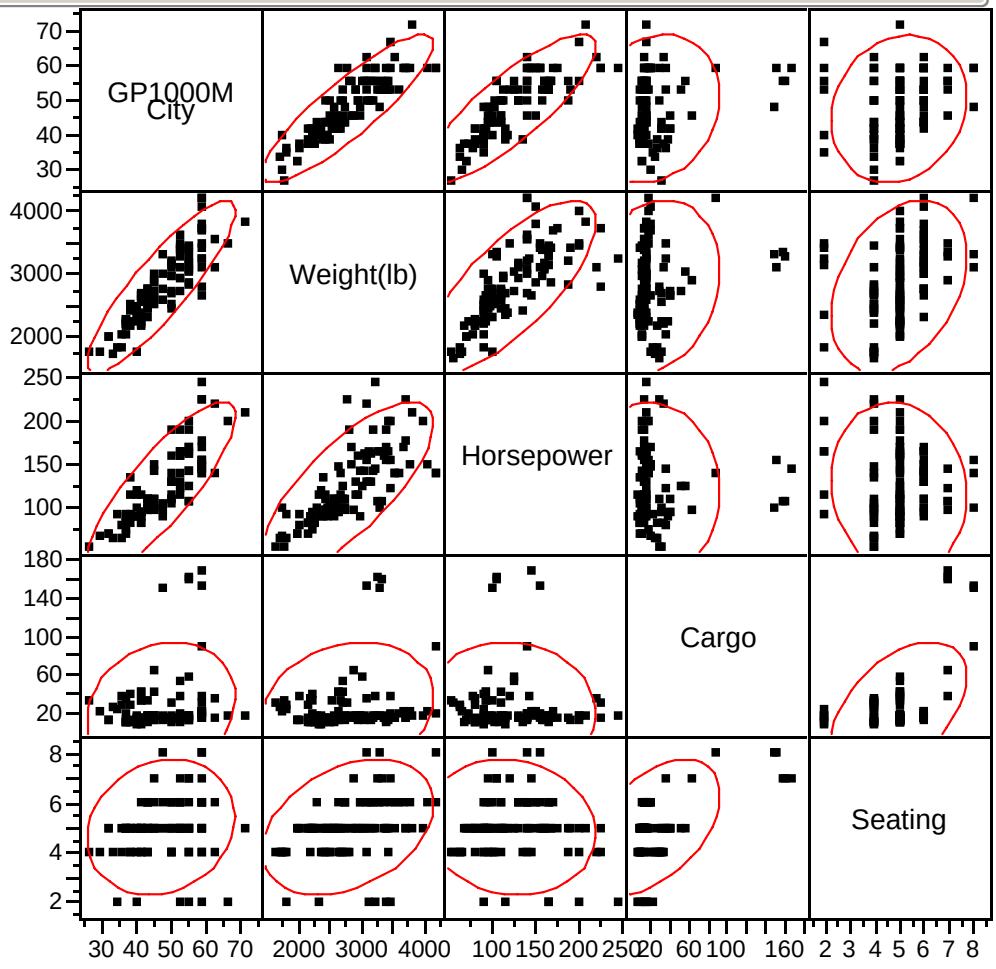
Multivariate

Correlations

	GP1000M City	Weight(lb)	Horsepower	Cargo	Seating
GP1000M City	1.0000	0.8798	0.8334	0.1672	0.1620
Weight(lb)	0.8798	1.0000	0.7509	0.1816	0.3499
Horsepower	0.8334	0.7509	1.0000	-0.0548	-0.0914
Cargo	0.1672	0.1816	-0.0548	1.0000	0.4894
Seating	0.1620	0.3499	-0.0914	0.4894	1.0000

7 rows not used due to missing values.

Scatterplot Matrix



[5] Find it under Analyze > Multivariate Methods > Multivariate

The multivariate command provides a correlation matrix and scatterplot matrix[6] for all pairwise relationships between the five variables GP1000M, Weight, Horsepower, Cargo and Seating.

> Besides Weight, which variable appears most strongly associated with GP1000M?

To consider the *joint* effect of Weight and Horsepower on GP1000M, we apply the Fit Model command[7] to obtain the multiple regression output (p 118)

Response GP1000M City

Summary of Fit

RSquare	0.841022
RSquare Adj	0.838105
Root Mean Square Error	3.499726
Mean of Response	47.59468
Observations (or Sum Wgts)	112

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	11.684254	1.727038	6.77	<.0001
Weight(lb)	0.0089183	0.000882	10.11	<.0001
Horsepower	0.0883837	0.012264	7.21	<.0001

To interpret this output, let's first describe the underlying multiple regression model.

---

[6] The density ellipses in each of these plots are estimates of the highest density population regions under the assumption of joint normality. Note how these ellipses guide your eye towards the strongest linear associations.

[7] Fit Y by X in JMP only performs simple regressions. To fit a multiple regression, use Fit Model. Here we select GP1000M as Y and add Weight and Horsepower to the Model Effects box in the dialog used to specify the multiple regression.

The Multiple Regression Model (MRM)

A model for the relationship between

y - a dependent variable or response, and

$x_1,\ldots, x_K$ - a set of independent variables, explanatory variables or predictors

Denote the n observations of the K+1 terms y, x_1, …, x_K by

$$y_i, x_{1i},\ldots, x_{Ki}, \qquad i = 1,...,n$$

Under the MRM, the data is assumed to be a realization of

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \varepsilon_i, \qquad i = 1,...,n$$

$$\varepsilon_1,\ldots,\varepsilon_n \text{ iid} \sim N(0,\sigma_\varepsilon^2)$$

Pictorially[8]

$K = 1$

$K = 2$

$K \geq 3$, hyperplane

[8] The case $K = 2$ can be visualized with JMP command Factor Profiler > Surface Profiling after right clicking on the Fit Model output title bar.

Remark: Even for $K > 1$, $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K$ is usually just called the regression line.

Some key interpretations:

$\beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki}$

β_0

β_k for $k = 1,\ldots, K$ (Careful!)

σ_ε

$\beta_0, \beta_1, \ldots, \beta_K$ and σ_ε are the (usually) unknown parameters of the MRM. An objective of regression is to estimate them.

The Least Squares (LS) Regression

In order to estimate the *"true" regression*

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K$$

we use the *least squares (LS) regression*

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_K x_K$$

which has the property of minimizing the sum of squared vertical distances from the plane to the data

The values of $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_K$ are calculated by computer programs such as JMP which insert the data into formulas, (*which if you would like to know, we'll tell you during office hours*).

The values $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_K$ are called the *least squares (LS) estimates* of $\beta_0, \beta_1, \ldots, \beta_K$

Partial versus Marginal Regression Coefficients

Returning to the previous regressions, let

y = GP1000M, x_1 = Weight and x_2 = Horsepower

From the output on p 4-5, we can see that the LS regression

GP1000M = 11.68 + 0.00891 Weight(lb) + 0.0883 Horsepower

estimates the "true" regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

In this model, β_1 is called a *partial regression coefficient*.

The interpretation of $\hat{\beta}_1 = 0.00891$ here is

In contrast, the LS regression line on p 4-3,

GP1000M = 9.43 + 0.01362 Weight(lb)

is an estimate of the "true" regression line

$$y = \beta_0 + \beta_1 x_1$$

In this model, β_1 is called a *marginal regression coefficient*.

The interpretation of $\hat{\beta}_1 = 0.01362$ here is

What is the essential difference between partial and marginal regression coefficients?

To get some insight into what's going on, we note that a simple regression of Horsepower on Weight yields (p 120)

$$Horsepower = -26.10 + 0.0533\ Weight$$

Substituting this expression for Horsepower into the multiple regression yields

$$GP1000M = 11.68 + 0.00891\ Weight + 0.0883\ Horsepower$$

$$= 11.68 + 0.00891\ Weight + 0.0883\ (-26.10 + 0.0533\ Weight)$$

$$= 9.43 + 0.01362\ Weight$$

which is just the previous simple regression!

A path analysis diagram provides a convenient representation of what's going on.

"How much does my 200 pound brother owe me for 3000 miles of riding with me?"[9]

[9] Is there ever a context in which you would rather have the marginal coefficient? Yes. Suppose you only know the weight of a car. Which slope would help you estimate its fuel consumption?

Inference about $\beta_0, \beta_1, \ldots, \beta_K$

Tests and confidence intervals used in simple regression generalize naturally to multiple regression.

Fact:

> Under the MRM, the sampling distributions of $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_K$ are normal with means $\beta_0, \beta_1, \ldots, \beta_K$

Along with the estimates $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_K$, programs such as JMP report their standard errors SE($\hat{\beta}_0$), SE($\hat{\beta}_1$),..., SE($\hat{\beta}_K$)

Confidence Intervals for β_k

Approximate 95% CI's for $\beta_0, \beta_1, \ldots, \beta_K$ are given by

Hypothesis Tests for β_k

For testing the null hypothesis H_0: [$\beta_k = c$ in the fitted model] vs H_1: [$\beta_k \neq c$ in the fitted model], t ratio = $\dfrac{\hat{\beta}_k - c}{SE(\hat{\beta}_k)}$

> Hypotheses of the form H_0: [$\beta_k = 0$ in the fitted model] are usually of most interest. Why?

If $|t \text{ ratio}| > 2$ *or* p-value < .05 *or* 95% CI does not contain c, reject H_0 at the .05 level of significance.

Example

JMP provides *t* ratios and p-values for testing

Suppose we consider adding the variables Cargo and Seating to the car89 regression.

Response GP1000M City

Summary of Fit

RSquare	0.852239
RSquare Adj	0.846556
Root Mean Square Error	3.411697
Mean of Response	47.67511
Observations (or Sum Wgts)	109

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	12.930547	2.020835	6.40	<.0001
Weight(lb)	0.0091318	0.001159	7.88	<.0001
Horsepower	0.0857712	0.01509	5.68	<.0001
Cargo	0.0346363	0.013277	2.61	0.0104
Seating	-0.476467	0.412437	-1.16	0.2506

What would you conclude about the effect of either addition from this output? (p 125)

Note that if Seating is removed here, the other *t* ratios and p-values will change.

WARNING! Used properly, the *t* ratios justify removing *at most one variable at a time*. Regression must then be rerun to get a new set of *t* ratios.

The Fitted Values and the Residuals

As in simple regression, the LS regression line again serves to decompose the data into the fitted values and the residuals

$$y_i = \hat{y}_i + e_i$$

where

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_K x_{Ki} \quad \text{and} \quad e_i = y_i - \hat{y}_i$$

Root Mean Squared Error (*RMSE*) – An Estimate of σ_ε

When the MRM holds, σ_ε is estimated by *RMSE*.

For example, in the car89 regression output on p 4-12, *RMSE* is given by Root Mean Square Error = 3.41.

As in a simple regression, *RMSE* is also called the *standard deviation of the residuals* and measures the dispersion of the residuals about the LS regression line. It again measures the predictive accuracy of the model used to forecast values for new cases.

The formula for the *RMSE* is[10]

$$RMSE = \sqrt{\frac{1}{n-K-1}\sum (y_i - \hat{y}_i)^2}$$

[10] $RMSE^2$ is the "average" sum of squared deviations from the regression line. We divide by $(n - K - 1)$ instead of n to compensate for the fact that the LS line always obtains a smaller sum of squared deviations than the true regression line.

R-square and Adjusted R-square

Just as in simple regression, R^2 in multiple regression is widely interpreted as

"the proportion of variation explained by the regression"

where variation refers to the variance of y, namely s_y^2

Just as before

$$R^2 \approx \frac{s_y^2 - RMSE^2}{s_y^2}$$

so that

$$RMSE \approx s_y \sqrt{1 - R^2}$$

In the simple regression of GP1000M on Weight, R^2 = 76.5%. When Horsepower is added, R^2 increases to 84.1%.

> Fact: R^2 can never decrease when another independent variable x is added to a regression.

To avoid this limitation, people sometimes use adjusted R^2 which is essentially R^2 penalized by the number of x's in the regression.

In the previous two regressions, adjusted R^2 goes from 76.3% to 83.8% when Horsepower is added.

Prediction of a Future Observation

"Where will a future value of the response *y* lie?"

"What GP1000M will I get with a car of a given weight and horsepower?"

After running a multiple regression, each point on the LS regression

$$\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_K x_K$$

is a prediction of the corresponding future point generated by the MRM

$$y_x = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + \varepsilon_x$$

For example, suppose we wanted to predict GP1000M for a new design where Weight = 4000 and Horsepower = 200.

Using the regression of GP1000M on Weight and Horsepower, the predicted value of GP1000M for this new design is (p125)

GP1000M = 11.68 + 0.00891 Weight(lb) + 0.0883 Horsepower

= 11.6 + 0.00891 (4000) + 0.0883 (200) = 65.03

JMP provides this calculation for you.[11]

[11] At least, it will if you figure out how to ask it. The trick is to add an extra row to the data. The last row of car89.jump contains the *x* values for the new design. After running Fit Model to obtain the regression output, right-click on one of the title bars , and select Save Columns > Predicted Values from the Pop-up menu. The predicted values for all of the rows, including the new one, will be placed in a column to the right of the data. This can also be done by selecting Save Columns > Prediction Formula which also includes the prediction formula in the calculator window.

JMP also provides[12] [57.9, 72.1] as a 95% prediction interval (PI) for y_x when Weight = 4000 and Horsepower = 200

What is the interpretation of this interval?

These results can also easily be used to predict MPG for the new design. By using the transformation 1000/GP1000M, the prediction of MPG is (1000/65.03) = 15.8 and the 95% PI is [1000/72.1, 1000/57.9] = [13.9, 17.27]. (p 131)

The prediction of GP1000M for the simple regression on Weight is

$$GP1000M = 9.43 + 0.01362\ Weight(lb)$$

$$= 9.43 + 0.01362\ (4000) = 63.9$$

and the 95% PI is [55.3, 72.5]. How do these compare with the above?

As in simple regression, extrapolate with caution! If $x_1,\ldots, x_K$ are not in the range of the data, predicting y_x is dangerous and the PIs are unreliable.

Often the intercept in a regression is an extrapolation itself. The intercept *is* the prediction when all of the predictors are set to zero. For data like these cars, we don't see any cases like that, and so the intercept is quite far from the data.

[12]Follow the steps in the previous footnote and select Save Columns > Indiv Confidence Interval.

Some New Graphical Model Diagnostics

In addition to the model checking methods we saw for simple regression, a variety of graphical methods are especially useful for multiple regression.

Plots of the Raw Data: Although it is not possible to plot all the data when $K > 3$, it may be useful to look at scatterplot matrices (pg 4-4) or 3-D spinning plots[13].

The following plots: Actual by Predicted, Residual by Predicted, and Leverage Plots were produced by the Fit Model platform[14] for the regression of GP1000M on Weight, Horsepower, Cargo and Seating.[15]

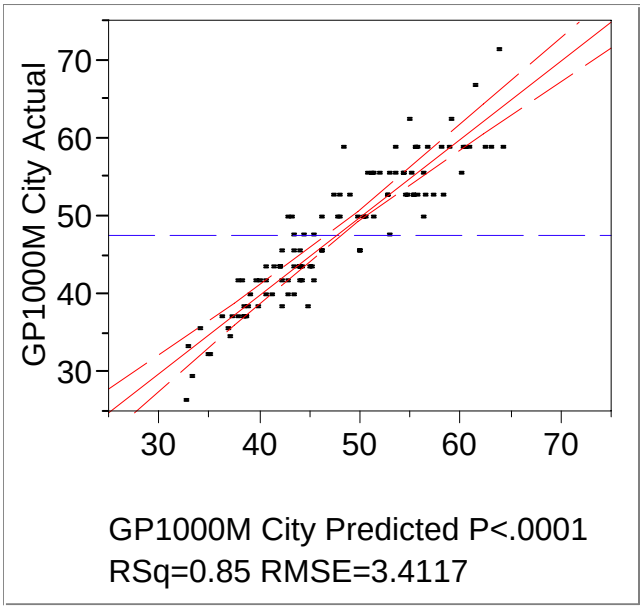
How should each of these be used? The key feature shared by all of these is that each offers you a "simple regression view" of the multiple regression.

Simple regression is simple because you can easily plot the data and see what is happening. The diagnostic plots shown by JMP with a multiple regression present various scatterplot views of a multiple regression.

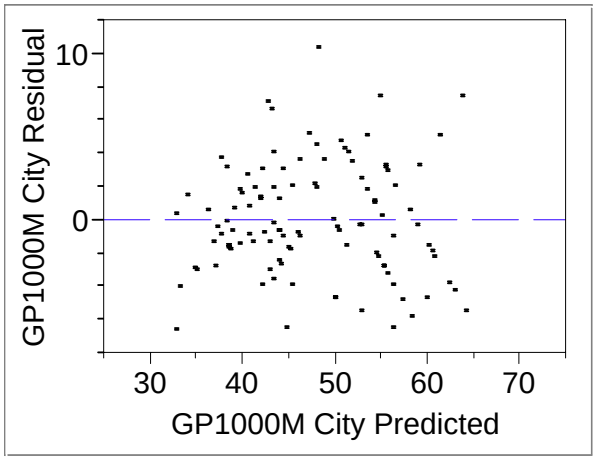
[13] Obtained with Graph > Spinning Plot In JMP.
[14] When Emphasis Select Leverage (the default) is selected.

The first two plots resemble the Fit Y by X plots of the data and residuals. For a simple regression, the one predictor supplied the x-axis. For multiple regression, these use a mixture of the predictors for the x-axis – namely the predicted values.

Actual By Predicted Plot [16]



Residual by Predicted Plot



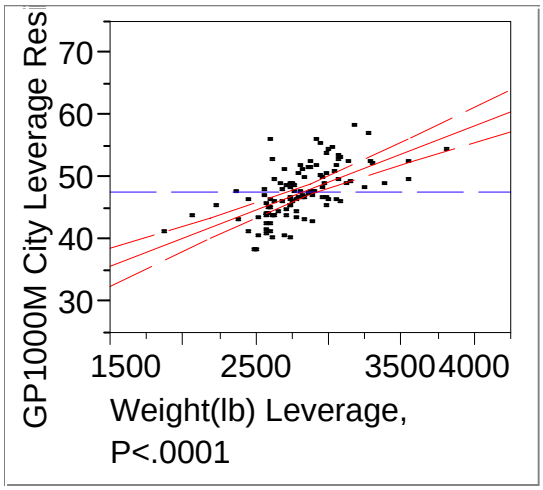
[16] JMP tries to show you a plot for every summary statistic. This time, its showing a plot that goes with the R^2 summary. The higher the R^2, the more the points in this plot cluster along the diagonal.

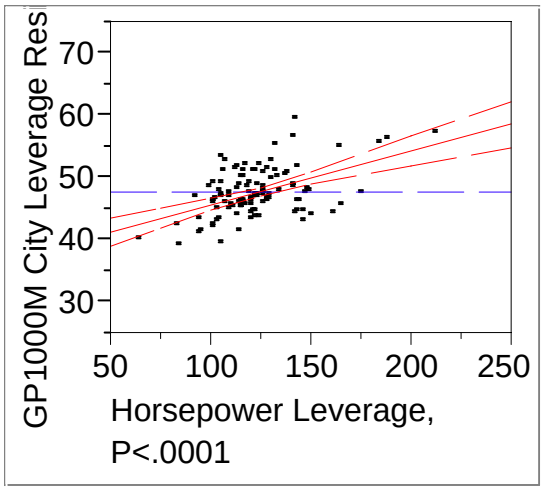
Leverage Plots

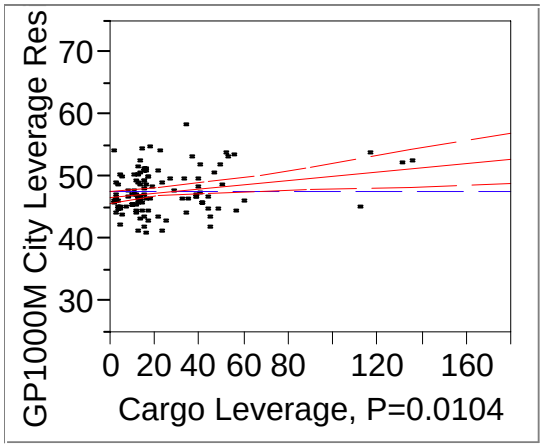
Leverage plots show you a simple regression view of the partial regression coefficient

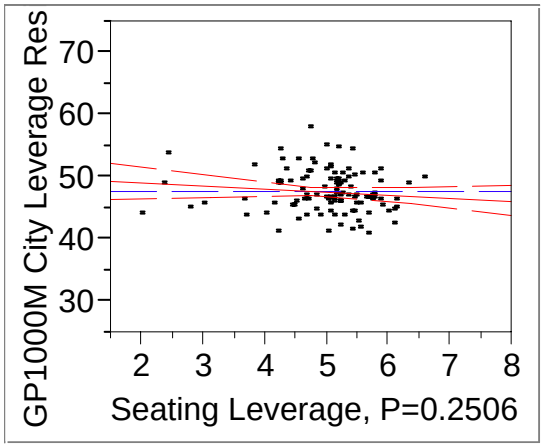
Scatterplot: marginal coefficient
Leverage plot: partial coefficient

The slope of the fitted line in each leverage plot is the slope for the indicated predictor in the multiple regression.[17]









[17] So why are these called leverage plots? They excel at revealing leverage points in the multiple regression that are hard to spot in the marginal views of the fit. BAR (p 63) describes the calculation of leverage in a simple regression. To make a version of these plots by hand is straightforward, but tedious. To make the leverage plot for Weight, regress fuel consumption on the other predictors (HP, cargo, seating); save the residuals. Now regress Weight on these three other predictors; save these residuals. Finally (whew), plot the residuals of fuel consumption on the residuals of Weight. Though tedious, you can see how the leverage plot removes the effects of the other predictors. It uses regression!

Take-Away Review

Multiple regression extends the ideas of simple regression, allowing one to use several predictors to model simultaneously the variation in the response.

The addition of other variables changes the interpretation of the slope: the slope in a multiple regression is a "partial" effect, adjusted for the other predictors.

The underlying MRM is a natural extension of the SRM, allowing for more predictors. Under these assumptions, we can again use standard errors to form confidence intervals and test hypotheses.

To assess the assumptions of the MRM, new diagnostic plots include plots of fitted value on actual values of the response, residuals on fitted values, and leverage plots.

Next Module

More on multiple regression, with an emphasis on the effects of correlation among the predictors (i.e., collinearity).

Department of Statistics
The Wharton School
University of Pennsylvania

Statistics 621 Fall 2005

Module 5

Further Aspects of Multiple Regression

The ANOVA Table

All statistics programs including JMP provide an ANOVA (Analysis of Variance) table. The p-value on this table is used for testing

$$H_0: \beta_1 = 0, \ldots, \beta_K = 0$$

What does this hypothesis imply about the relationship between y and $x_1,\ldots, x_K$?

This hypothesis is tested with the familiar p-value strategy: If the p-value < .05, then $H_0: \beta_1 = 0, \ldots, \beta_K = 0$ can be rejected at the .05 level of significance.

This test is based on F ratio statistic

$$F = \frac{R^2 / K}{(1 - R^2)/(n - K - 1)}$$

Larger values of F correspond to smaller p-values. A rule of thumb[1] is to reject H_0 at the .05 level if $F > 4$.

[1] Use this rule if you do not have a p-value handy. This rule is "conservative": any time the F > 4, the p-value < 0.05. However, there are cases in which F < 4 but the p-value p < 0.05 even though F < 4.

Example (car89.jmp)

The multiple regression of GP1000M on Weight, Horsepower, Cargo and Seating yields

Response GP1000M City

Analysis of Variance

The test that the whole model fits better than a simple mean, i.e. testing that all the parameters are zero except the intercept

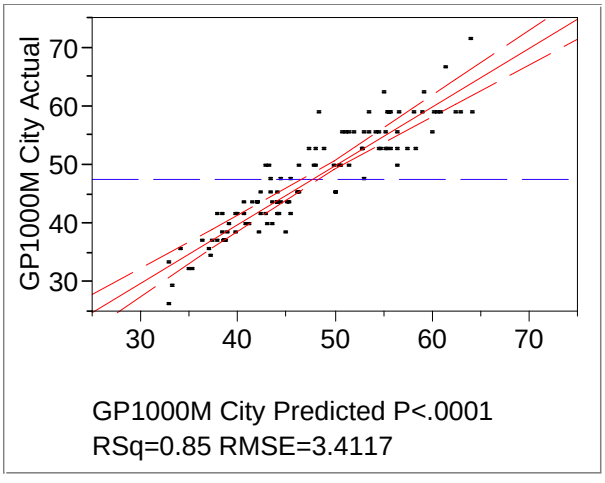
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	6981.9348	1745.48	149.9598
Error	104	1210.5264	11.64	Prob > F
C. Total	108	8192.4611		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	12.930547	2.020835	6.40	<.0001
Weight(lb)	0.0091318	0.001159	7.88	<.0001
Horsepower	0.0857712	0.01509	5.68	<.0001
Cargo	0.0346363	0.013277	2.61	0.0104
Seating	-0.476467	0.412437	-1.16	0.2506

What should we conclude from the ANOVA table?[2]

The y vs $\hat{y}$ plot confirms this conclusion.



[2] Note that we have turned on "Show explanations" for this output.

The F Test and Correlated Predictors

The ANOVA test comes in handy when, as often happens, the predictors in a regression are correlated. The following example illustrates an extreme case.

Example: A Market Model [3]

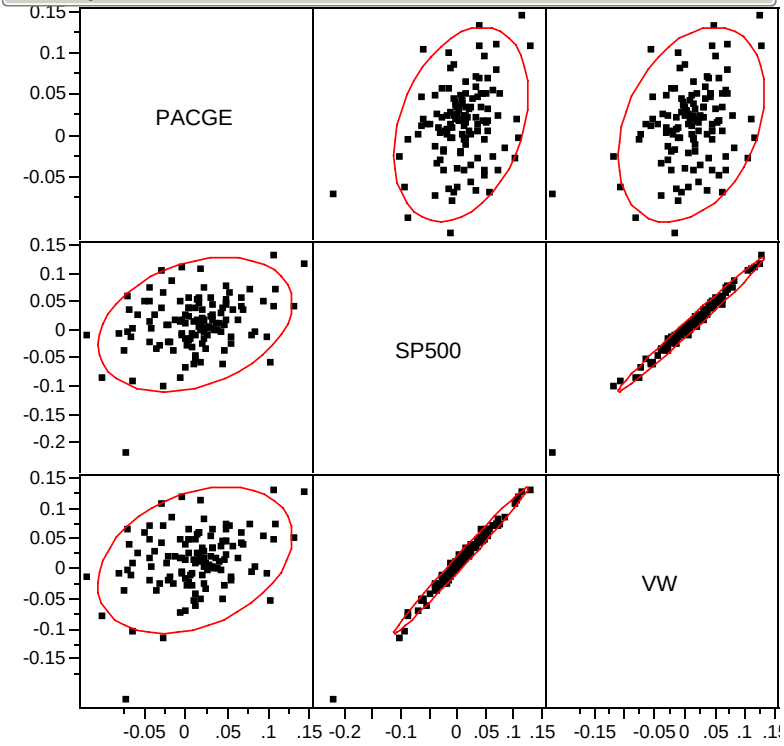
The file stocks.jmp contains monthly returns from 2/78 to 12/87 of VW, SP500, IBM, PACGE and Walmart. Let's focus on the relationship between PACGE, SP500 and VW.

Multivariate

Correlations

	PACGE	SP500	VW
PACGE	1.0000	0.3116	0.3249
SP500	0.3116	1.0000	0.9932
VW	0.3249	0.9932	1.0000

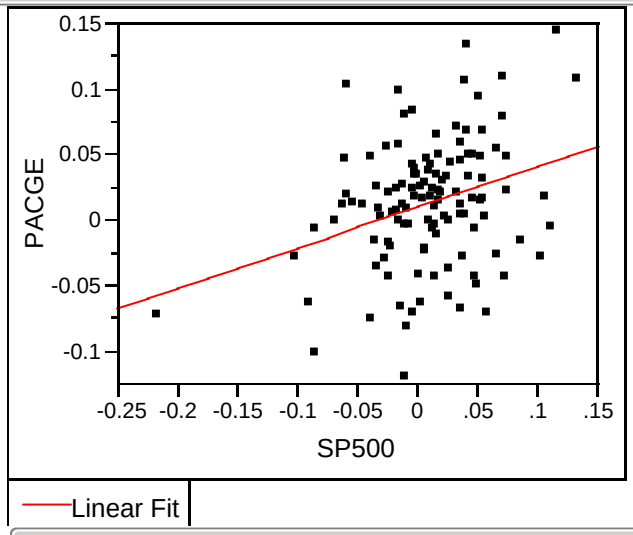
Scatterplot Matrix



[3] The BAR casebook example that uses this data (p 138) focuses instead on the relationship of these indices with the returns on Walmart stock. The results are similar and issues of collinearity arise there as well.

A simple regression of PACGE on SP500 yields

Bivariate Fit of PACGE By SP500



Linear Fit

PACGE = 0.0096823 + 0.3102945 SP500

Summary of Fit

RSquare	0.097073
RSquare Adj	0.089356
Root Mean Square Error	0.046161
Mean of Response	0.012718
Observations (or Sum Wgts)	119

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.02680342	0.026803	12.5786
Error	117	0.24931178	0.002131	Prob > F
C. Total	118	0.27611520		0.0006

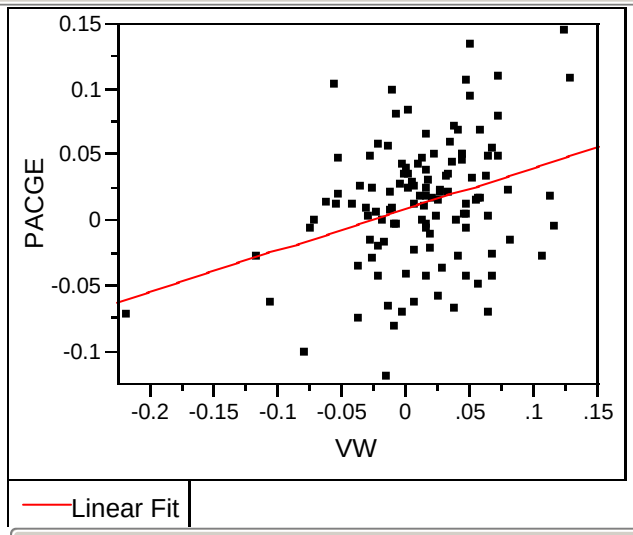
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0096823	0.004317	2.24	0.0268
SP500	0.3102945	0.08749	3.55	0.0006

What is the interpretation of $\hat{\beta}_1$ here?

A simple regression of PACGE on VW yields

Bivariate Fit of PACGE By VW



Linear Fit

PACGE = 0.0083714 + 0.3156957 VW

Summary of Fit

RSquare	0.105532
RSquare Adj	0.097887
Root Mean Square Error	0.045945
Mean of Response	0.012718
Observations (or Sum Wgts)	119

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.02913890	0.029139	13.8040
Error	117	0.24697630	0.002111	Prob > F
C. Total	118	0.27611520		0.0003

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0083714	0.004371	1.92	0.0579
VW	0.3156957	0.08497	3.72	0.0003

What is the interpretation of $\hat{\beta}_1$ here?

Consider now what happens when *both* SP500 and VW are used together in a multiple regression

Response PACGE

Summary of Fit

RSquare	0.114681
RSquare Adj	0.099417
Root Mean Square Error	0.045906
Mean of Response	0.012718
Observations (or Sum Wgts)	119

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	0.03166507	0.015833	7.5131
Error	116	0.24445013	0.002107	Prob > F
C. Total	118	0.27611520		0.0009

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0054478	0.005119	1.06	0.2895
SP500	-0.821098	0.749946	-1.09	0.2758
VW	1.1114984	0.731784	1.52	0.1315

What has happened?[4]

[4] If the response is returns on Walmart, the regression shown in the casebook on page 143 finds a significant effect for the value-weighted index. Thus, in that case, VW significantly improves a regression with SP500 alone, but not vice versa. Adding SP500 to a model that already has VW does not improve the fit, agreeing with underlying finance.

Collinearity

In a multiple regression of y on $x_1,\ldots, x_K$, linear redundancy – or correlation – among $x_1,\ldots, x_K$, is called *collinearity*.

Effects of collinearity:

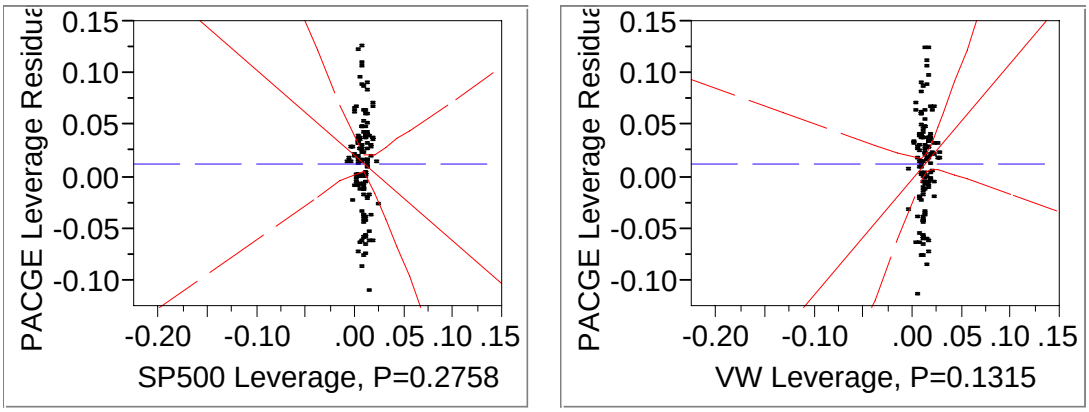
> Coefficient standard errors increase
> t-ratios decrease (and so p-values increase)
> Difficulty interpreting coefficients
> Coefficients change as others come and go.

These effects can be serious when collinearity is severe.

Why these effects happen:

> Key fact: In a multiple regression, $\hat{\beta}_k$ is the effect of adding x_k last. (As shown in the leverage plots)
>
> Variation of x_k with the other x's fixed is limited (BAR, p 121). This manifests itself as
>
> $$\mathrm{SE}(\hat{\beta}_k) \approx \frac{RMSE}{\sqrt{n}} \times \frac{1}{SD(adjusted\ x_k)}$$
>
> where adjusted x_k is the residual from a multiple regression of x_k on all the other x's
>
> The increase in $\mathrm{SE}(\hat{\beta}_k)$ leads to smaller t-ratios.

The following leverage plots for the multiple regression of PACGE on SP500 and VW illustrate this phenomenon.



What to do if you have severe collinearity (BAR, p. 147)

- Suffer[5]

- Remove natural proxies

- Transform or combine some of your predictors

[5] Collinearity does not violate an assumption of the MRM. Rather, it causes problems in interpretation: the coefficients may not make much sense. If you only need to predict cases like the ones you have seen, it's not a problem. If you want to explain your predictions, it is.

Example: Market Segmentation

A marketing project identified a list of affluent customers for its new PDA. Should it focus on the younger or older members of this list?

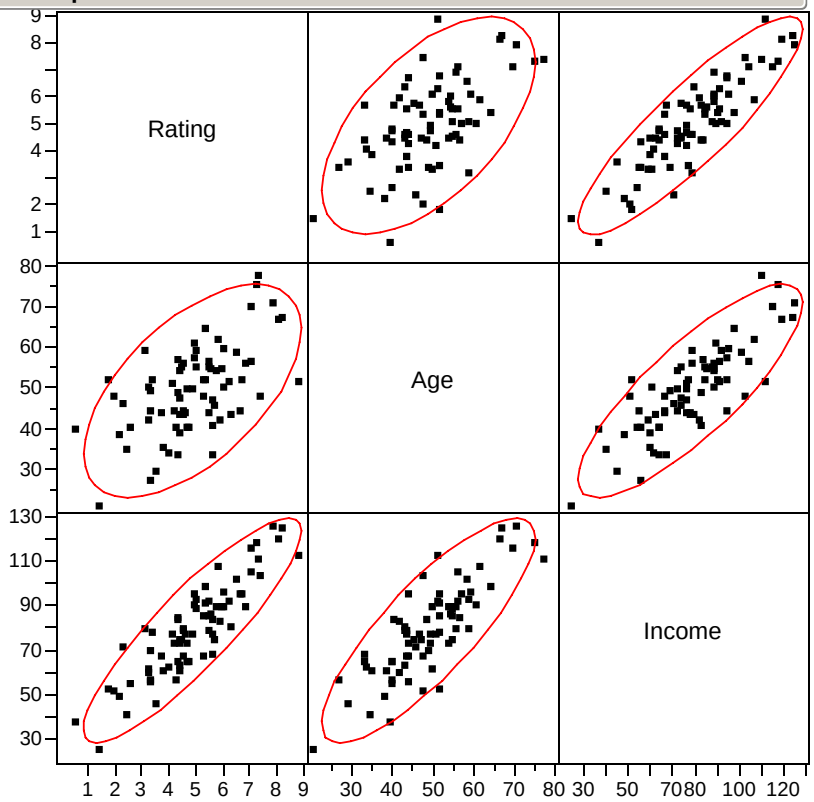
To answer this question, the marketing firm obtained a sample of 75 consumers and asked them to rate their "likelihood of purchase" on a scale of 1 to10. The data is in PDA.jmp Age and Income of consumers were also recorded.

Multivariate

Correlations

	Rating	Age	Income
Rating	1.0000	0.5867	0.8845
Age	0.5867	1.0000	0.8286
Income	0.8845	0.8286	1.0000

Scatterplot Matrix



The two simple regressions and multiple regression of *Rating* on *Age* and *Income* yields the following:

Regression of Rating on Age

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.4923901	0.733017	0.67	0.5039
Age	0.0900056	0.014541	6.19	<.0001

Regression of Rating on Income

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.595877	0.352356	-1.69	0.0951
Income	0.0700332	0.004322	16.20	<.0001

Multiple Regression Estimates

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.517709	0.352139	1.47	0.1459
Age	-0.071628	0.012479	-5.74	<.0001
Income	0.1006638	0.00644	15.63	<.0001

What's going on?

Based on these results, how should the marketing firm direct their marketing efforts?

Another Example

Just because we are doing multiple regression does not mean we should ignore transformations. Logs, in particular, can be very important in economic models.

Models with logs of Y and X lead to slope interpretations as *elasticities*. The BAR casebook gives an example (p 148).

Take-Away Review

The **F-test** allows for you to look at the importance of several factors simultaneously. When predictors are *collinear*, the F-test reveals their net effect rather than trying to separate their effects as a t-ratio does.

A **leverage** plot shows the contribution of each predictor to the regression, giving you a picture of what that variable adds to a model that contains *all* of the others.

Collinearity does not violate any assumption of the MRM, but it does make regression harder to interpret. In the presence of collinearity, slopes become less precise and the effect of one predictor depends on the others that happen to be in the model.

Next Module

Not all predictors are numerical. Some of the most important predictors of a response label an attribute of the observation, such as the sex or specialty of a doctor.

JMP allows you to easily include such categorical predictors in a regression, but leaves you with the burden of figuring out how to interpret the results. We'll start with that next time.

Department of Statistics
The Wharton School
University of Pennsylvania

Statistics 621 Fall 2005

Module 6

Categorical Predictors in Regression

Categorical variables

Represent group membership (e.g. type of car, sex, race)

JMP denotes such columns as "nominal".

A dummy variable is a 0-1 variable indicating the presence or absence of a certain characteristic,

$$x = \begin{cases} 1 & \textit{if characteristic present} \\ 0 & \textit{if characteristic absent} \end{cases}$$

Dummy variables in regression incorporate categorical variables

Modeling the differences between 2 groups

Do cars from different regions differ in gasoline consumption?

Are women systematically underpaid at a firm?

Does a company discriminate in hiring of minorities?

Example: Employee Performance Study (BAR, p 161)

"Which of two prospective job candidates should we hire, the internal manager or the externally recruited manager?"

Data set Manager.jmp

150 managers: 88 internal and 62 external

MngrRating is an evaluation score of the employee in the job they do, indicating the "value" of the employee to the firm.

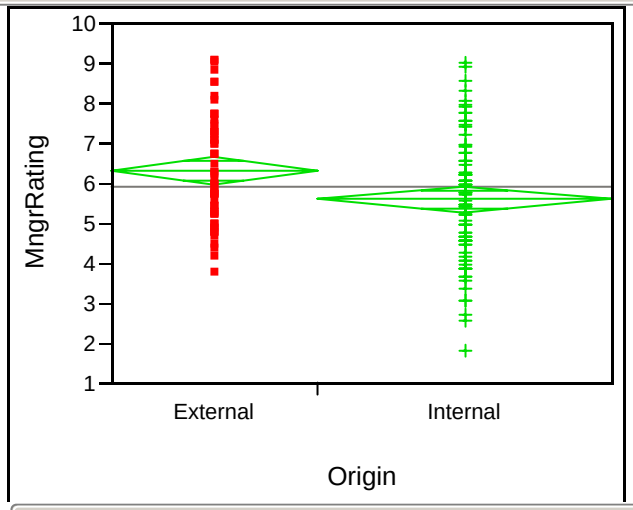
Origin is a categorical variable which identifies the managers as either Externally or Internally hired

Salary is the starting salary of the employee when hired and indicates what sort of job the person was hired to do

Let's begin with a simple two-sample comparison obtained using Fit Y by X with *MngrRating* as Y and *Origin*, a categorical variable[1], selected as X.

[1] Make sure that the columns you want JMP to treat as categorical are marked as *nominal*. Check the status of each column by examining the left panel of the spreadsheet window: each column name is preceded by a symbol denoting how JMP treats that column: *c* for "continuous", *n* for "nominal" and *o* for "ordinal". Change the "modeling type" by clicking on the status symbol itself. Ordinal columns represent ordered categories, such a low, medium, and high on some attribute. We won't consider such special categorizations.

Oneway Analysis of MngrRating By Origin



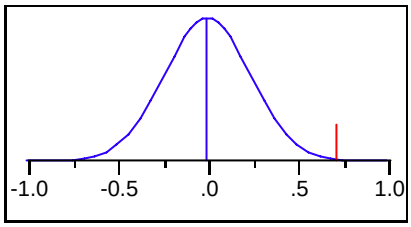
Oneway Anova

t Test

External-Internal

Assuming equal variances

Difference	0.71642	t Ratio	2.983912
Std Err Dif	0.24010	DF	148
Upper CL Dif	1.19088	Prob > |t|	0.0033
Lower CL Dif	0.24197	Prob > t	0.0017
Confidence	0.95	Prob < t	0.9983



Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
External	62	6.32097	0.18390	5.9576	6.6844
Internal	88	5.60455	0.15436	5.2995	5.9096

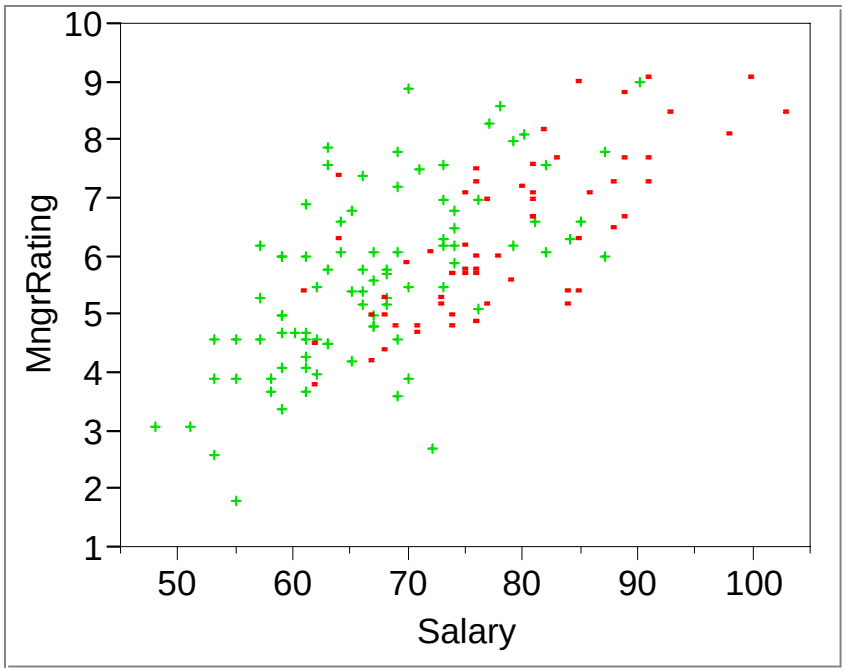
Std Error uses a pooled estimate of error variance

This output shows that the average performance rating for external managers, 6.32, is significantly higher than that for internal managers, 5.60, since (p 161)

$$(6.32 - 5.60) = 0.72 \text{ with t ratio} = 2.98$$

However, before we jump to the conclusion that the external candidate should be hired, let's explore the relationship between *MngrRating* and *Salary*.

Consider the following scatterplot of *MngrRating* vs *Salary* where green + denotes internal and red · denotes external[2]



What does this plot reveal? (p 166)

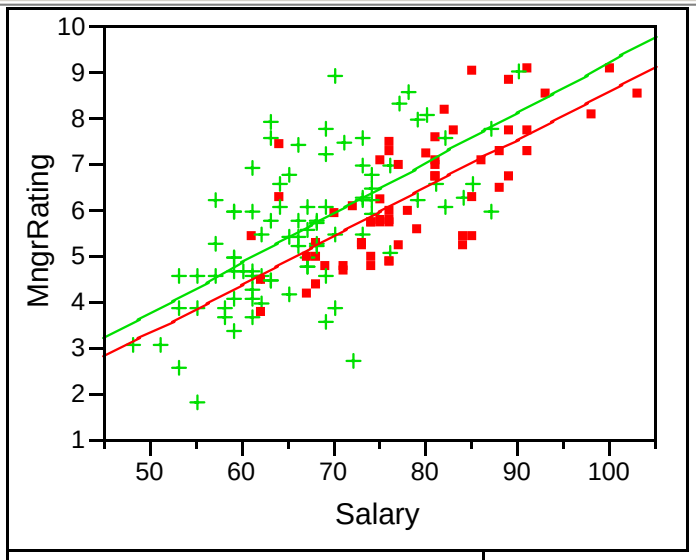
We've discovered collinearity in a new form: *Salary* is related to the one predictor used to explain the rating of a manager, namely the internal/external status of a manager.[3]

[2] Use the JMP row command Color/Mark by column, and pick Origin as the column to label the points. Plots of data that mix observations from different groups are usually more informative if you color/mark the points by their group membership.

[3] Collinearity in this context is often called "confounding." We cannot tell if the difference in average ratings of managers is due to interval/external labels or some other factor.

Let's consider separate regressions of *MngrRating* on *Salary*[4]

Bivariate Fit of MngrRating By Salary



Linear Fit Origin=="External"

MngrRating = -1.936941 + 0.1053912 Salary

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1.936941	0.986231	-1.96	0.0542
Salary	0.1053912	0.012499	8.43	<.0001

Linear Fit Origin=="Internal"

MngrRating = -1.693524 + 0.1090929 Salary

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1.693524	0.949254	-1.78	0.0779
Salary	0.1090929	0.014066	7.76	<.0001

At each given salary, internal managers now appear to be more highly rated! (p 166-7)

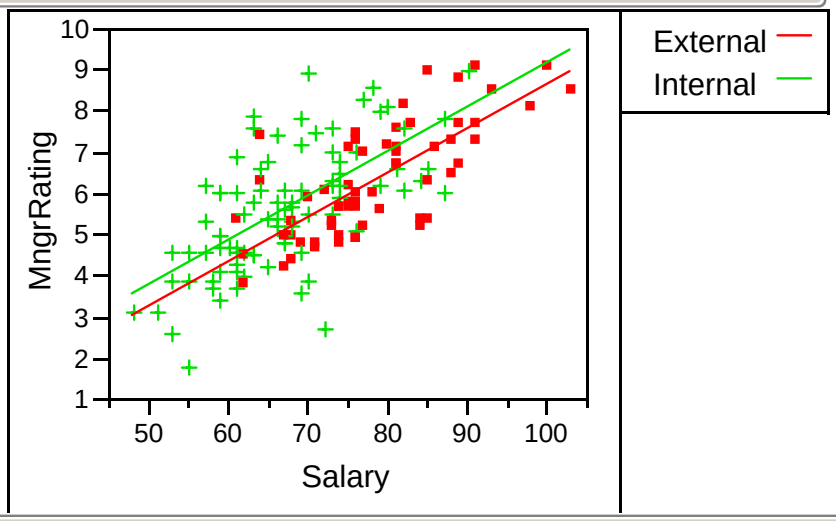
[4] This is conveniently done with Fit Y by X by selecting the "Group by" option using Origin *before* selecting Fit Line on the pop-up menu at the red triangle. If you fit the overall line first, the coloring will be confusing.

The slopes are so similar in the previous two regressions, that it seems reasonable to consider a model that forces them to be the same.

The Fit Model[5] command yields the following output:

Response MngrRating

Regression Plot



Expanded Estimates

Nominal factors expanded to all levels

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1.842976	0.706344	-2.61	0.0100
Salary	0.1074781	0.009649	11.14	<.0001
Origin[External]	-0.257483	0.104515	-2.46	0.0149
Origin[Internal]	0.2574829	0.104515	2.46	0.0149

[5] Select MngrRating as Y, and add Salary and Origin as predictors. Right-click on a title bar of the output and select Estimates > Expanded Estimates from the pop-up menu. Coefficients for all groups are revealed only in the "expanded" estimates output. Note that the coefficients for Origin are redundant: they add to zero. If you know one, you also know the other.

To incorporate *Origin* in the regression, JMP has included the two *dummy variables*, one for each group,

$$\begin{aligned} \text{Origin[External]} &= 1 \quad \text{if Origin = External} \\ &= 0 \quad \text{otherwise} \\ \text{Origin[Internal]} &= 1 \quad \text{if Origin = Internal} \\ &= 0 \quad \text{otherwise} \end{aligned}$$

With this in mind, the JMP output can be interpreted as two *parallel* regression lines

$$\text{MngrRating} = -1.84 + 0.107\,\text{Salary} - 0.257 \quad \text{if Origin = External}$$

$$\text{MngrRating} = -1.84 + 0.107\,\text{Salary} + 0.257 \quad \text{if Origin = Internal}$$

The difference between the intercepts is (p 169)

$$(-1.84 + .257) - (-1.84 - .257) = 2(.257)$$

and can be interpreted as

This difference is significantly different from 0 since .0149, the p-value for Origin[External], is less than .05.[6]

Thus, if we *assume* slopes are equal (i.e. no evident interaction), a model using a categorical predictor implies that controlling for initial salary, internal managers rate significantly higher.[7]

---

[6] Equivalently, we could have used the p-value for *Origin*[Internal] which is also .0149. With just two categories, the size of the two coefficient estimates and their p-values will always be identical.

[7] Use of a regression model that contains both categorical and continuous predictors with a focus or emphasis on the difference in the intercepts among the groups is sometimes called the Analysis of Covariance

How can we check the assumption that the slopes are parallel?

Rather than only look at the plot, we can fit a model that allows the slopes to differ. This model gives an estimate of the difference between the slopes.

This estimate is known as an *interaction*. It measures the difference between the slopes.[8]

Expanded Estimates

Nominal factors expanded to all levels

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1.815233	0.723615	-2.51	0.0132
Salary	0.1072421	0.00976	10.99	<.0001
Origin[External]	-0.121709	0.723615	-0.17	0.8667
Origin[Internal]	0.1217087	0.723615	0.17	0.8667
Salary*Origin[External]	-0.001851	0.00976	-0.19	0.8499
Salary*Origin[Internal]	0.0018509	0.00976	0.19	0.8499

[8] It is added to the model in JMP's Fit Model dialog by highlighting *Salary* and *Origin*, and then using the "cross" button. You should right-click on the red indicator at the top left of the output and turn off "Center Polynomials".

To interpret this output, extract the estimated regression line for the external group

$$\text{MngrRating} = -1.815 + 0.107\ \text{Salary} - 0.122 - 0.0018\ \text{Salary}$$

$$= (-1.815 - 0.122) + (0.107 - 0.0018)\ \text{Salary}$$

Similarly, for the internal group

$$\text{MngrRating} = (-1.815 + 0.122) + (0.107 + 0.0018)\ \text{Salary}$$

In this example, the interaction terms are not significant. How can you tell?

Note: The assumption of equal error variance should also be checked (p 174)

Conclusion

Should an internal or external candidate be hired?

Review Example: Wage discrimination (p 180)

"Are men paid more than women in management positions?"

Data set Salary.jmp:

220 managers, 145 men and 75 women.

Confounding

Marginally, women are paid *less* ($t = -2.06$) than men (p 181)
Confounding: men have more experience and work in positions of higher responsibility (or position) (p 182-3)

Regression with categorical predictor

Adjusting for confounding yields a reversal (p 184).
When adjusted for differences in experience and position, women are paid *more* than men.

Interactions

The casebook includes discussion of various types of interaction, such as that between two continuous variables.

Practical comment

JMP's centering of interactions reduces the collinearity often associated with these terms. Just the same, simplicity warrants removing extraneous interaction terms. Unless an interaction is significant, remove it.

Conclusion

Comparison of groups with statistically adjusted backgrounds suggest that men are paid less. However, why do the women have less background?

Take-Away Review

Categorical variables allow us to model the differences between two groups using regression.

> In a model with a categorical variable, the coefficients of the categorical terms indicate differences between parallel lines.
>
> In a model that includes an interaction, the coefficients of the interaction measure the differences in the slopes for the two groups.

Interaction, in general, measures how the slope of one predictor depends upon levels of others. Does the effect of one predictor (e.g., Salary) depend upon another (Internal/External).

Questions to address when using categorical variables

> Are the groups really so similar that it makes sense to combine them into one regression.
>
> If so, are the slopes different for the distinct groups? (i.e., Is interaction present?)
>
> Are the error variances comparable? Heteroscedasticity can be a problem: Since the groups have different slopes, why should the variances be the same?

Next Module

Categorical variables can define more than two groups. The ideas are the same, only with more slopes. The proliferation of coefficients in the resulting fitted model requires new type of statistical test.

Department of Statistics
The Wharton School
University of Pennsylvania

Statistics 621 Fall 2005

Module 7

More on Categorical Predictors in Regression

Common Business Questions

Did advertising affect sales differently in several marketing regions?

How can I compare the performance of several production managers?

To answer both questions requires regression with categorical predictors that may have *more than two* categories.

Example: Timing production runs (BAR, p 189)

"Are three production managers doing equally well?"

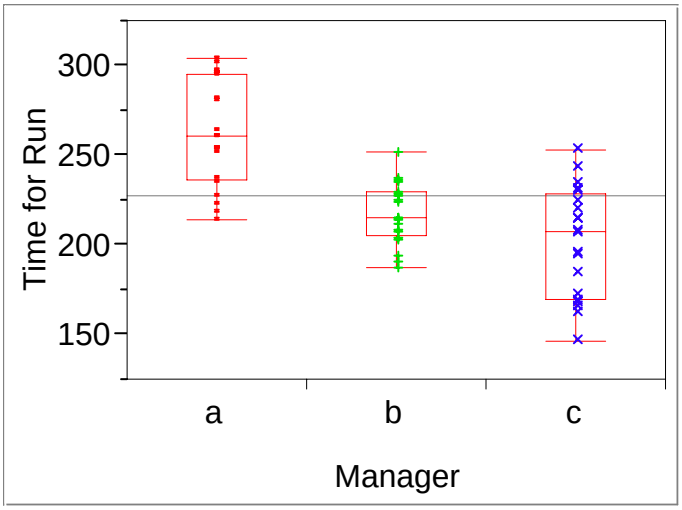
Data set Prodtime.jmp

20 production runs supervised by each of three managers. Each observation relates the time (in minutes) to complete the task to the number of units produced and the manager.

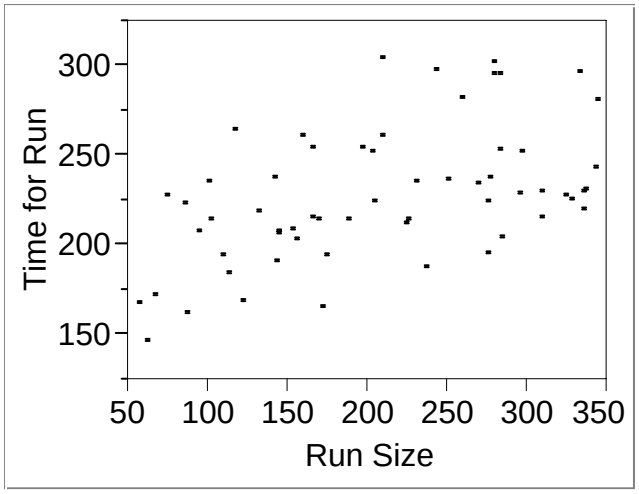
Three variables: *Time for Run* is the response with *Manager* and *Run Size* as predictors

A marginal comparison of the run times of the three managers (shown on the next page) may not be appropriate since, at least initially, we do not know if the jobs are comparable.

Nonetheless, production times do vary among the managers, with Manager c looking the best from this perspective.



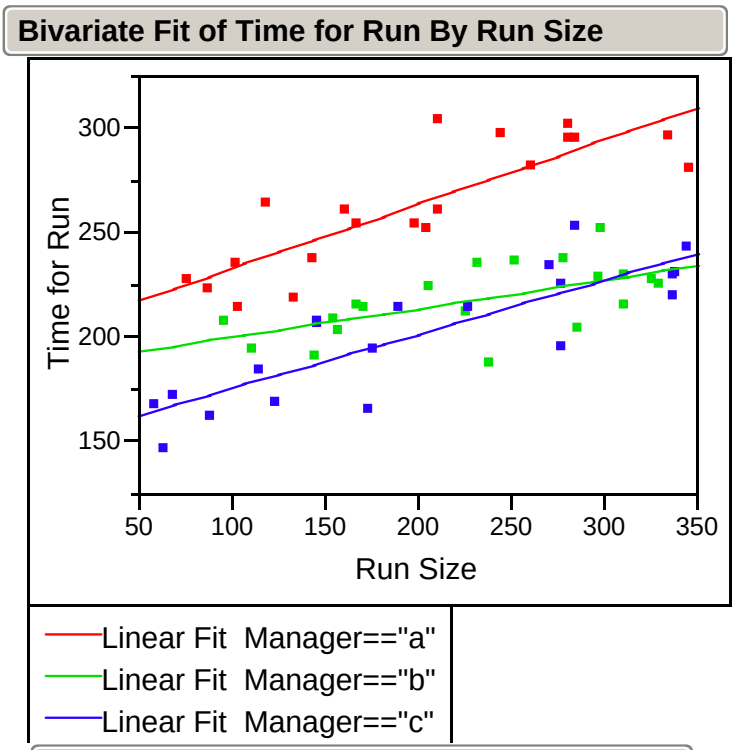
But run time and run size are related



How might this affect the marginal comparison of the managers?[1]

[1] Suppose that Run Size, one predictor, is related to the other predictor, here the Manager. In that case, the two predictors are collinear, even though one is continuous and the other is categorical. What plot could you use to see this effect, known as *confounding*?

Fitting three separate regressions with color-coded points reveals a more complete comparison of the features of the three managers, especially the presence of interaction (differing slopes) (p 190-1)



Linear Fit Manager=="a"

Time for Run = 202.53415 + 0.307265 Run Size

Linear Fit Manager=="b"

Time for Run = 186.49386 + 0.1367771 Run Size

Linear Fit Manager=="c"

Time for Run = 149.7477 + 0.2592431 Run Size

Questions

Which manager would you pick to supervise a large job?

To supervise a small job?

How do we tell if these apparent differences are meaningful (i.e. not just due to randomness)?

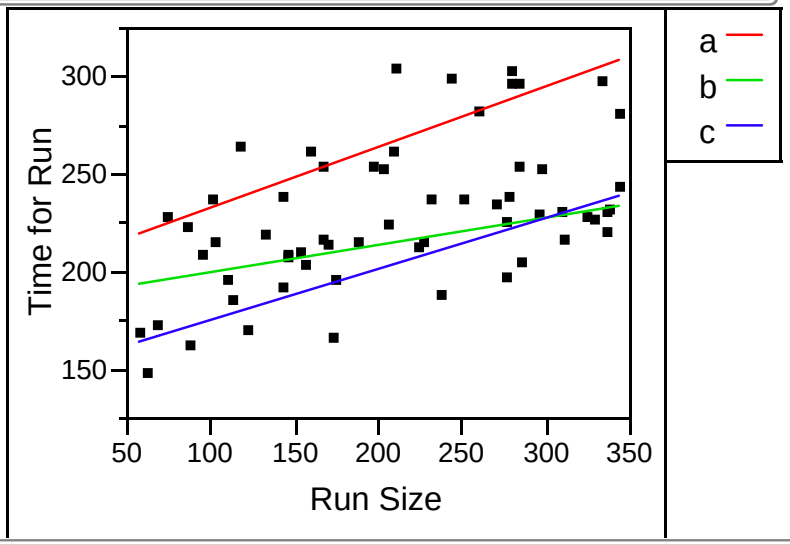
Interactions

Let's fit a model with interactions to this data (p. 195)[2]

Response Time for Run

Whole Model

Regression Plot



Expanded Estimates

Nominal factors expanded to all levels

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	179.59191	5.619643	31.96	<.0001
Run Size	0.2344284	0.024708	9.49	<.0001
Manager[a]	22.942248	7.76019	2.96	0.0046
Manager[b]	6.9019547	8.730525	0.79	0.4327
Manager[c]	-29.8442	7.282836	-4.10	0.0001
Manager[a]*Run Size	0.0728366	0.035263	2.07	0.0437
Manager[b]*Run Size	-0.097651	0.037178	-2.63	0.0112
Manager[c]*Run Size	0.0248147	0.032207	0.77	0.4444

[2] The interaction term is added to the model in JMP's Fit Model dialog by highlighting *RunSize* and *Manager*, and then using the "cross" button. You should right-click on the red indicator at the top left of the output and turn off "Center Polynomials".

The fitted model from this output can be interpreted as three distinct regression lines.

For example, the fitted line for Manager a is obtained from the output, as

Time for Run = baseline model + effects for Manager a

= 179.59 + 0.234 Run Size + 22.94 + 0.073 Run Size

= (179.59 + 22.94) + (0.23+0.073) Run Size

= 202.50 + 0.307 Run Size

Notes

- This *is* the previous simple regression for Manager a.[3]
- The Manager terms sum to zero.
- The interaction terms sum to zero.
- Each set is redundant and thus adds only two new predictors to the model.

[3] It's reasonable to ask at this point "Why bother?" since we got the same fit using a simple regression. The reason to bother is that we now can easily test the size of the difference.

Effect Tests and Categorical Variables

As with other predictors, the standard errors in the shown Expanded Estimates can be used to gauge the precision of each of the coefficient estimates. (e.g., form confidence intervals)

The *t* ratios can be used to test hypotheses of the form

$$H_0: [\beta_k = 0 \text{ in the fitted model}]$$

For example, what can we conclude about the baseline coefficient of Run Size?

However, the several *t*-ratios associated with the categorical variable cannot be used to test other hypotheses of interest such as

$$H_0: \text{All Manager coefficients} = 0$$

$$H_0: \text{All Manager-Run Size interactions} = 0$$

These are instead tested using the F ratios that appear in the Effect Tests[4] summary (p. 195)

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Run Size	1	1	22070.614	90.0192	<.0001
Manager	2	2	43981.452	89.6934	<.0001
Manager*Run Size	2	2	1778.661	3.6273	0.0333

Conclusions:

[4] An effect test is JMP's name for a partial F-test. As we'll see on p 7-7, a partial F-test is a test for a subset of the slopes in a regression model. In fact, the usual t-test as well as the overall F-test in the Anova table are special cases of the partial F-test. See p 7 following and the casebook.

Further Comparison of the Managers

The following output[5] provides

- adjusted mean run times for the managers[6]
- confidence intervals for differences between these means
- a grouping of means that are statistically equivalent

Manager

Least Squares Means Table

Level	Least Sq Mean	Std Error	Mean
a	266.84984	3.5424931	263.050
b	215.12358	3.6302438	217.850
c	204.01161	3.5117154	202.050

LSMeans Differences Tukey HSD

Alpha= 0.050 Q= 2.40999

LSMean[i] \ LSMean[j]	a	b	c
Mean[i]-Mean[j]			
Std Err Dif			
Lower CL Dif			
Upper CL Dif			
a	0	51.7263	62.8382
	0	5.07227	4.98813
	0	39.5022	50.8169
	0	63.9504	74.8596
b	-51.726	0	11.112
	5.07227	0	5.05082
	-63.95	0	-1.0605
	-39.502	0	23.2844
c	-62.838	-11.112	0
	4.98813	5.05082	0
	-74.86	-23.284	0
	-50.817	1.06045	0

Level		Least Sq Mean
a	A	266.84984
b	B	215.12358
c	B	204.01161

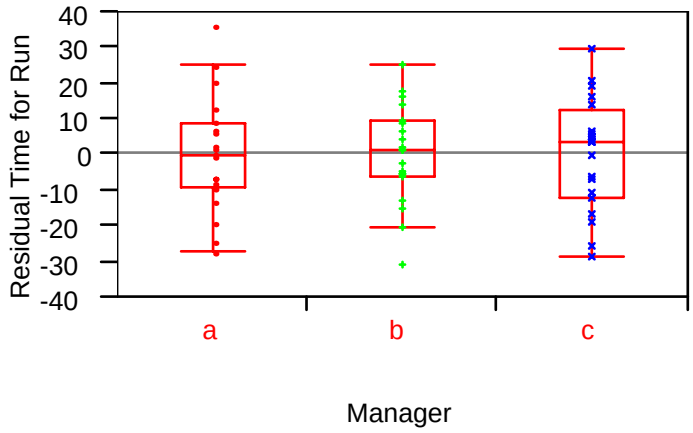
Levels not connected by same letter are significantly different

[5] The first table above appears in the Fit model output under the Manager Leverage Plot. To get the second table, right click on the Manager title bar and select LSMeans Tukey HSD.

[6] These adjusted means are the predicted values when runtime is set equal to its average value.

Model Checking

Don’t forget to check assumptions, particularly for equal variation in the residuals across the three managers (p. 196).



Conclusions for the Manager Analysis

Substantial differences exist among managers, both in setup times (intercepts) and in how they handle the impact of increased size of the production run (slopes).

Manager c gets the job started the quickest, but manager b does well for larger jobs.

Further questions for management include…

- What does b do to make large jobs run more quickly?
- How does c get the process started so quickly?
- How can we help a?

The Partial F-Test

In a multiple regression of y on $x_1, \ldots, x_K$, it is sometimes of interest to consider the *simultaneous* contribution of a subset of the x_k's to the model. This is equivalent to testing

H_0: Corresponding Subset of β_k's are all 0

This is tested with the partial F-ratio[7]

$$F = \frac{(R^2_{complete} - R^2_{reduced})/\#removed}{(1 - R^2_{complete})/(n - K - 1)}$$

where

$R^2_{complete}$ = the R^2 for the complete model with $x_1, \ldots, x_K$

$R^2_{reduced}$ = the R^2 for the reduced model where the subset of x_k's has been removed

$\#\ removed$ = number of effective coefficients removed

The F-ratios in the Effect Tests summary are of this form[8]

For the p-value associated with these or any other F-ratio, use the familiar strategy: If the p-value < .05, then H_0 above can be rejected at the .05 level of significance.[9]

---

[7] Examples of the use of the partial F test in multiple regression *without* categorical terms appear in the BAR casebook on pages 127 and 152.

[8] The familiar t ratios are also of the form of the partial F test. Note that the square of t = 9.74 on p 7-4 is equal to F = 90.02 on p 7-5.

[9] If you do not have a p-value handy, a useful rule of thumb is to reject H_0 at the .05 level whenever $F > 4$. This rule is "conservative": any time the F > 4, the p-value < 0.05. However, there are some cases in which the F < 4 but the p-value is still significant (p < 0.05).

Key Take-Away Points

Categorical predictors in regression

Allow regression to estimate and test for differences among the intercepts and slope of models fit to many groups, not just two.

The baseline model serves as an overall point of reference, with the categorical terms representing how the fits for the separate groups differ from this common baseline.

The benefit in collecting models for the separate groups into one is that we can test for the significance of the observed differences.

Effect tests

Partial F test for adding a categorical term (or the associated interaction) measures the statistical significance of the observed differences. When there are just two categories, the effect test is redundant with the t-test.

Significant categorical variable -> different intercepts[10]

Significant interaction -> different slopes

Next Module

How does one build a regression model from scratch? We'll see that it's a process of iterative refinement, combining what we learn from the data with our knowledge of the context.

[10] Beware testing for the categorical term itself in a model that has a significant interaction. In the presence of one or more interactions, the interpretation of the categorical term is no longer a simple shift in the intercept.

Department of Statistics
The Wharton School
University of Pennsylvania

Statistics 621 Fall 2005

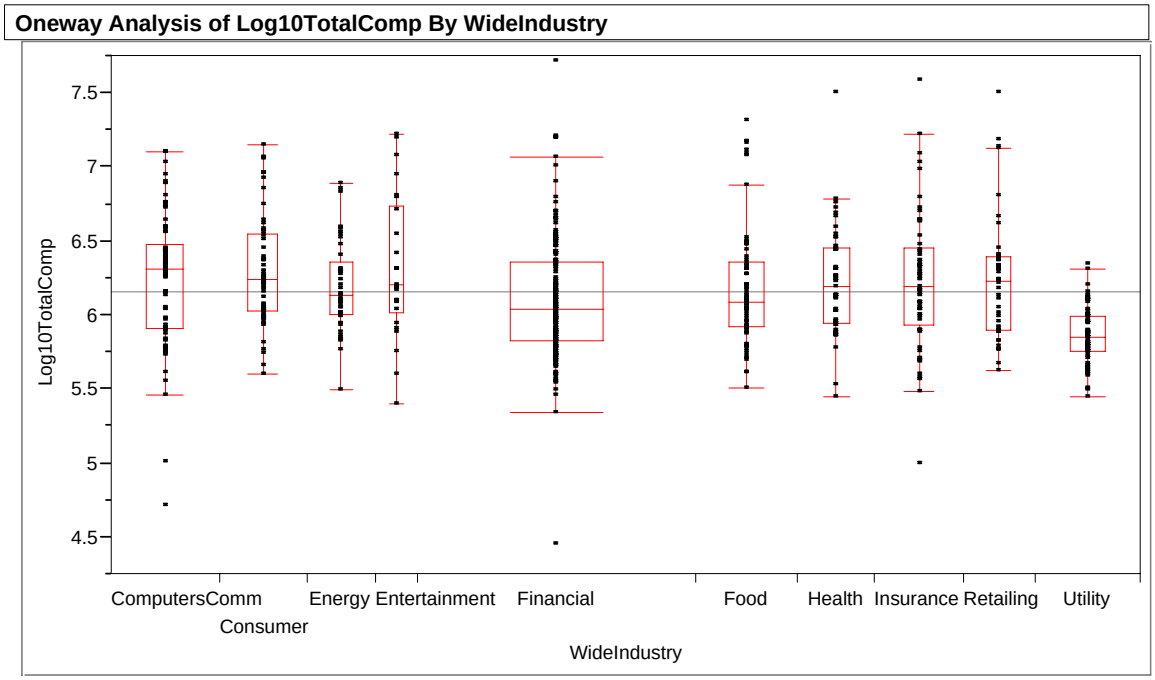
Module 8

Building a Regression Model

Example: Another Look at the Forbes94 Data (BAR, *p* 202)

In Stat 603, we studied the variation of annual CEO compensation of 790 executives as reported in the May 23, 1994 issue of Forbes. This data is in forbes94.jmp.

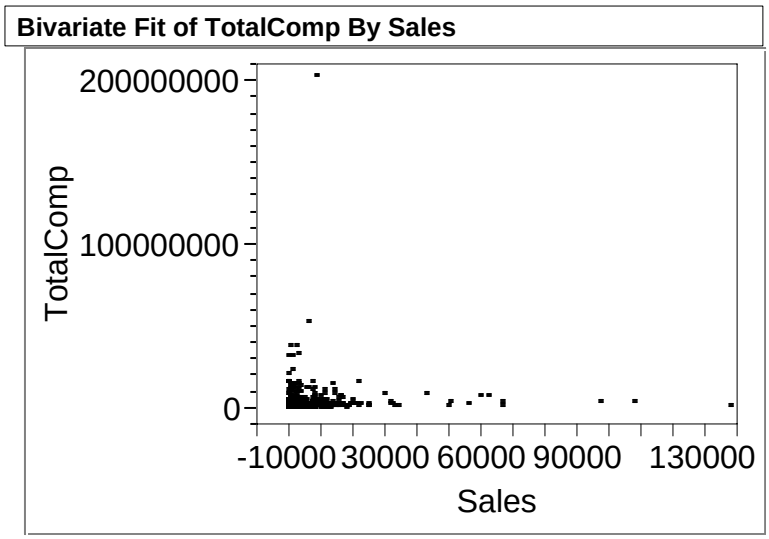
In Module 3 of Stat 603, we looked at comparison boxplots to compare total compensation across industries.



Such comparisons were facilitated by looking at Total Compensation on the $\log_{10}$ scale[1]. Why?

Let's now try to "explain" the variation in compensation among executives by using a regression model constructed from other variables recorded in the data. Which of these other variables might be expected to explain variation in compensation?

Let's start with the annual sales of the company. A scatterplot of Total Comp by Sales yields (BAR, p 203):



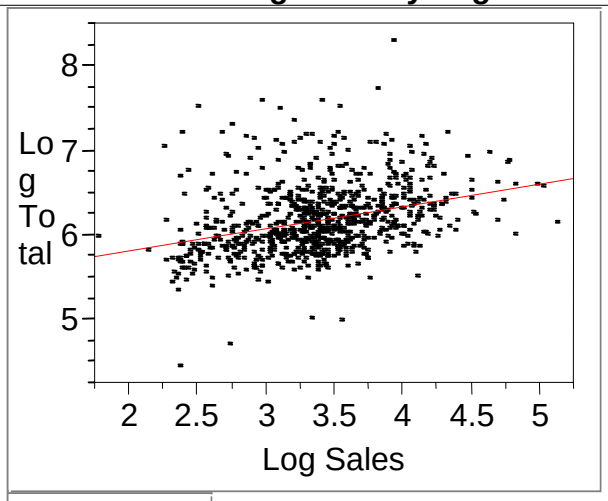
Why is it difficult to see what's going on from this plot?

[1] Recall that we used log base 10 to reveal more of the variation in the data that was otherwise dominated by extreme outliers. Using the base 10 lets us interpret these values as the number of digits in the original scale (minus one).

The plot suggests that it may be easier to work with transformation of both of these variables on the log scale.

The following plot of Log_{10} Total versus Log_{10} Sales reveals much more and also shows a rather linear relationship (p 204), albeit with asymmetry in the dispersion about the line.

Bivariate Fit of Log Total By Log



Linear Fit

Linear

Log Total = 5.2816554 + 0.265047 Log Sales

Summary of

RSquare	0.108212
RSquare Adj	0.10708
Root Mean Square Error	0.389444
Mean of Response	6.177785
Observations (or Sum Wgts)	790

Parameter

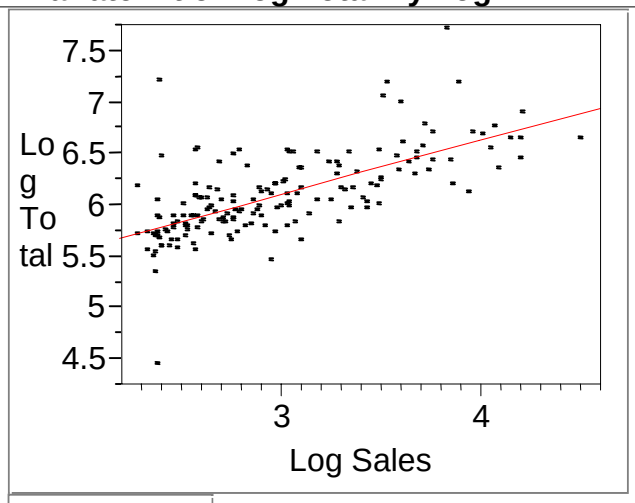
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	5.2816554	0.092685	56.99	<.0001
Log Sales	0.265047	0.027105	9.78	<.0001

What is the interpretation of the slope estimate here?[2]

[2] What does a p-value mean here? Since these are the top executives, in what sense is this a sample? The concept of a p-value here also harkens back to Stat 603 and the idea of using random variables to think about "what might have happened."

To get a more homogeneous subset, let's restrict attention to the financial industry. The regression of Log_{10} Total on Log_{10} Sales for the 168 CEOs in this subset of the data shows (p 207)

Bivariate Fit of Log Total By Log



Linear Fit

Linear

Log Total = 4.5180312 + 0.528997 Log Sales

Summary of

RSquare	0.449868
RSquare Adj	0.446554
Root Mean Square Error	0.30301
Mean of Response	6.100385
Observations (or Sum Wgts)	168

Parameter

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4.5180312	0.13781	32.78	<.0001
Log Sales	0.528997	0.045404	11.65	<.0001

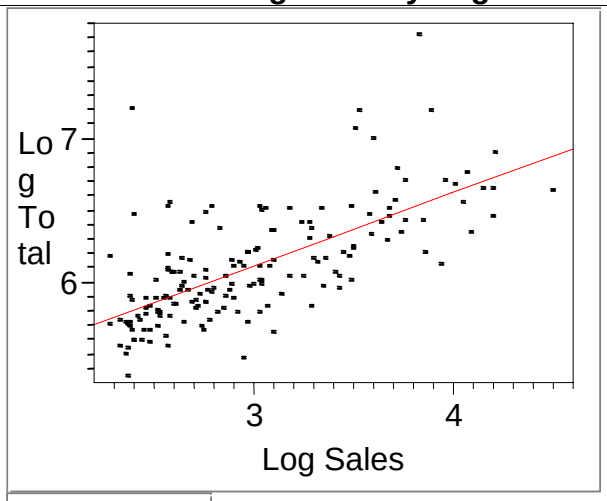
How has the fitted model changed?

Who is the outlying CEO in the lower left corner?

How would you *anticipate* the regression model to change if we exclude this outlier from the analysis?

Removing the outlying CEO yields a more revealing plot and the following simple regression.

Bivariate Fit of Log Total By Log



Linear

Log Total = 4.5806015 + 0.5107426 Log Sales

Summary of

RSquare	0.460916
RSquare Adj	0.457649
Root Mean Square Error	0.285777
Mean of Response	6.11021
Observations (or Sum Wgts)	167

Parameter

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4.5806015	0.130667	35.06	<.0001
Log Sales	0.5107426	0.043001	11.88	<.0001

Of course, other variables may help to further explain the variation of Log_{10} Total. If we add the age of the CEO and return on the company's stock over the last 5 years, we obtain

Response Log

Summary of

RSquare	0.5131
RSquare Adj	0.503913
Root Mean Square Error	0.272723
Mean of Response	6.112593
Observations (or Sum Wgts)	163

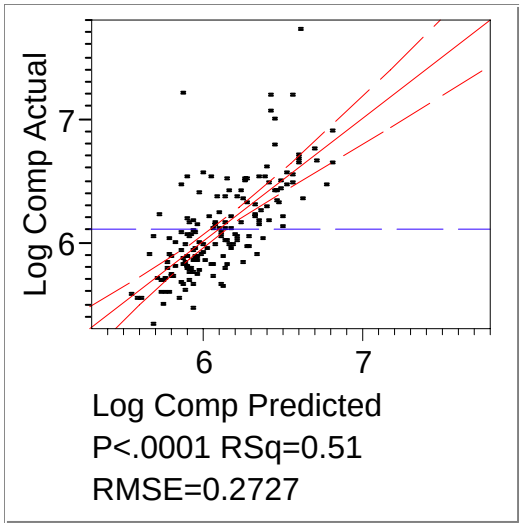
Analysis of

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	12.462441	4.15415	55.8520
Error	159	11.826075	0.07438	Prob > F
C. Total	162	24.288515		<.0001

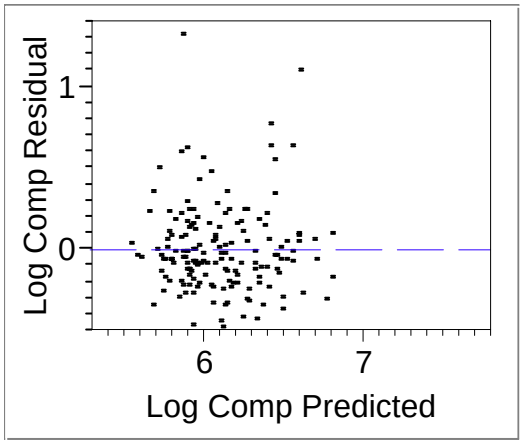
Parameter

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4.0248079	0.211547	19.03	<.0001
Log Sales	0.4938189	0.04225	11.69	<.0001
Age	0.0096423	0.003348	2.88	0.0045
ReturnOver5Yrs	0.0042525	0.001167	3.64	0.0004

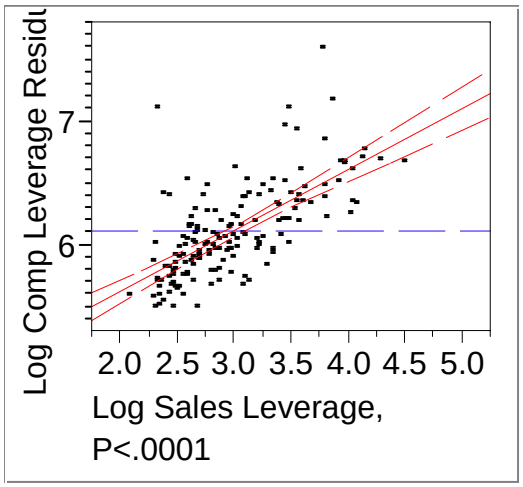
Actual vs Predicted Plot

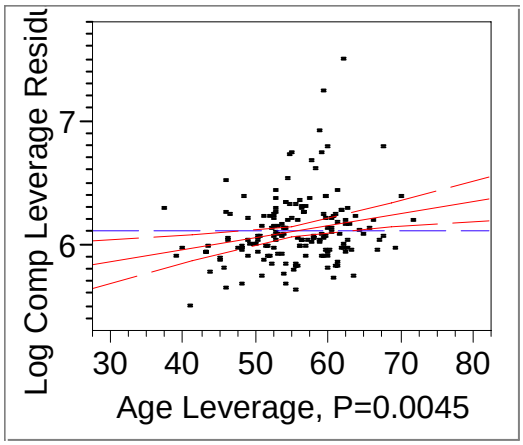


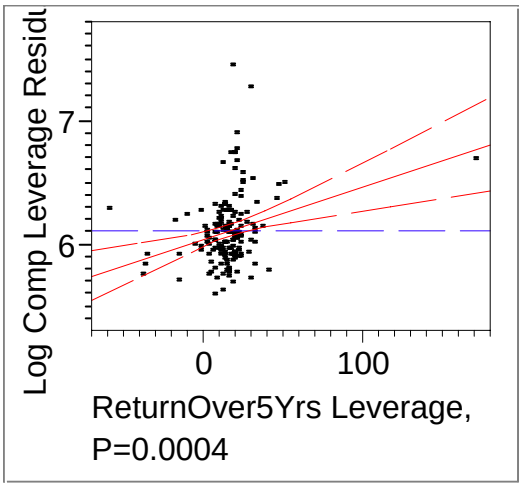
Residual vs Predicted Plot



Leverage Plots







How is this model different, statistically? Substantively?

What about the leverage point for *ReturnOver5Yrs*? Excluding this point and rerunning the regression has very little effect on the results, so should we keep it or remove it?

More importantly, how did we know to add the predictors *Age* and *ReturnOver5Yrs* to the model? Should we consider adding other variables to the model as well?

Automatic Variable Selection by Stepwise Methods

When building a regression model, we often have many potential *x* variables. Some predictors are obviously relevant (e.g., sales) but having used these, we may have no clear sense of how to proceed.

Solution A: Try every model…

To examine *every* possible regression may require too much effort. With *K=10* potential *x*'s, for example, how many models are possible?

Solution B: Consider a sequence of "interesting" models…

A convenient strategy is to use automatic methods that incrementally build potential models by using "greedy algorithms". Examples of such methods are the *stepwise regression* algorithms described below.

CAVEATS!!

Such automatic methods are not a substitute for thinking!

They merely select a manageable subset of models for further consideration.

They do not eliminate the need for model checking, removal of outliers, transformations, etc.

When feasible, it is usually best to build a model manually.

The Mechanics of Stepwise Regression

Stepwise regression proceeds by simply adding or removing *x* variables, one at a time, based on their p-values.

Add step: Add the *x* yielding the biggest improvement in R^2 so long as its p-value < p_{ENTER}

Remove step: Remove the *x* yielding the smallest decrease in R^2 so long as its p-value > p_{LEAVE}

The thresholds p_{ENTER} and p_{LEAVE} are picked by the user.

Three Variations[3]

Mixed Stepwise - First perform an *add* step, then a *delete* step, then an *add* step, etc. until no more *x*'s can be added or deleted. (i.e., none meet the p-value conditions)

Forward Selection - Start with an empty model and only add *x*'s, gradually building up a model.

Backward Elimination - Start with the full model and only remove *x*'s, "cleaning up" a model by eliminating predictors.

A hybrid approach combines substantive modeling with stepwise. Begin with a model motivated by the context (e.g., promotion as a predictor of sales) and use forward stepwise to explore additions (e.g., sift through possible interactions).

[3] JMP's Fit Model platform offers all three of these variations.

Back to the Forbes94.jmp Analysis

With Log_{10} Total as the response, reasonable candidate predictors include Log_{10}Sales, Age, ReturnOver5Yrs, AgeofUnder, MBA?, MasterPhD?, StockOwned, and Profits.

Rather than put all of these into one large multiple regression, let's use *forward stepwise* selection to see what happens[4].

Stepwise Fit

Response: Log Comp

Stepwise Regression Control

Prob to Enter 0.250
Prob to Leave 0.100
Direction: Forward

654 rows not used due to missing values.

Current Estimates

SSE	DFE	MSE	RSquare	RSquare Adj	Cp	AIC
9.6196681	141	0.068225	0.5308	0.5175	4.19044	-387.09

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	3.95358251	1	0	0.000	1.0000
☐	☑	Log Sales	0.47608288	1	8.294888	121.582	0.0000
☐	☑	Age	0.01044455	1	0.587633	8.613	0.0039
☐	☑	ReturnOver5Yrs	0.00645341	1	0.866502	12.701	0.0005
☐	☐	AgeOfUnder	.	1	0.064967	0.952	0.3309
☐	☑	MBA?	0.09734615	1	0.304663	4.466	0.0363
☐	☐	MasterPhd?	.	1	0.020711	0.302	0.5835
☐	☐	StockOwned	.	1	0.038468	0.562	0.4547
☐	☐	Profits	.	1	0.088932	1.306	0.2550

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p
1	Log Sales	Entered	0.0000	9.228567	0.4501	22.295	2
2	ReturnOver5Yrs	Entered	0.0006	0.884936	0.4933	11.399	3
3	Age	Entered	0.0110	0.464441	0.5159	6.6304	4
4	MBA?	Entered	0.0363	0.304663	0.5308	4.1904	5

What happened? What model has the procedure constructed?[5]

[4] This is done by selecting the Stepwise "personality" in the Fit Model window, clicking the "Run Model" button and selecting "Go" in the Stepwise window

[5] Stepwise search is also happy to look over the possible interaction terms in these models as well. Use the "response surface" macro as described later in these notes in the analysis of the VW stock index.

Now use the "Make Model" option to construct a standard regression dialog, with the selected variables already present. Regression on the selected variables yields[6]

Response Log

Summary of

RSquare	0.520008
RSquare Adj	0.507856
Root Mean Square Error	0.271637
Mean of Response	6.112593
Observations (or Sum Wgts)	163

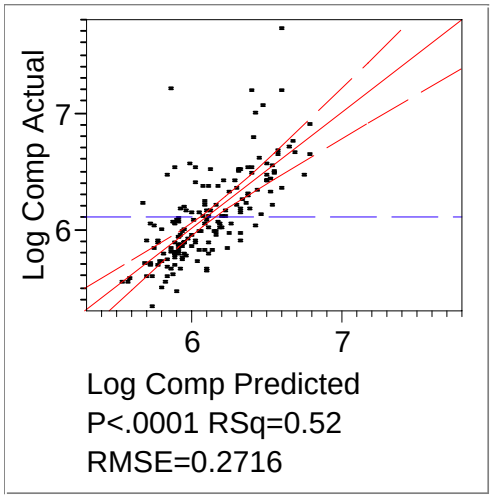
Analysis of

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	12.630215	3.15755	42.7930
Error	158	11.658301	0.07379	Prob > F
C. Total	162	24.288515		<.0001

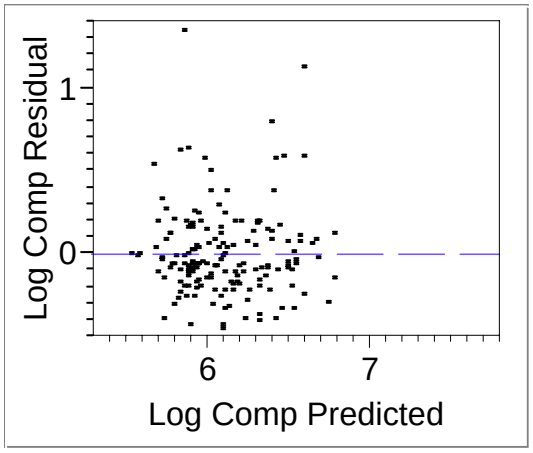
Parameter

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.9487475	0.216658	18.23	<.0001
Log Sales	0.4903204	0.042146	11.63	<.0001
ReturnOver5Yrs	0.0043727	0.001165	3.75	0.0002
Age	0.0107493	0.003415	3.15	0.0020
MBA?	0.0702732	0.046603	1.51	0.1336

Actual vs Predicted Plot

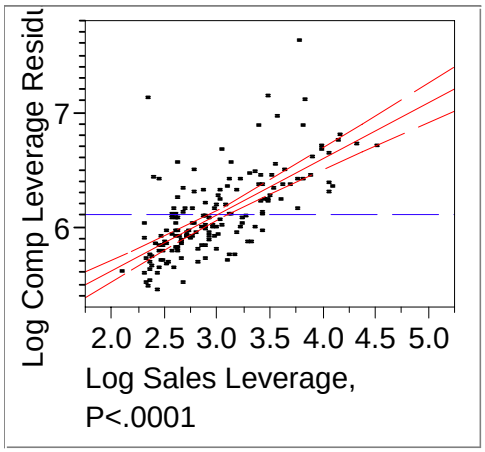


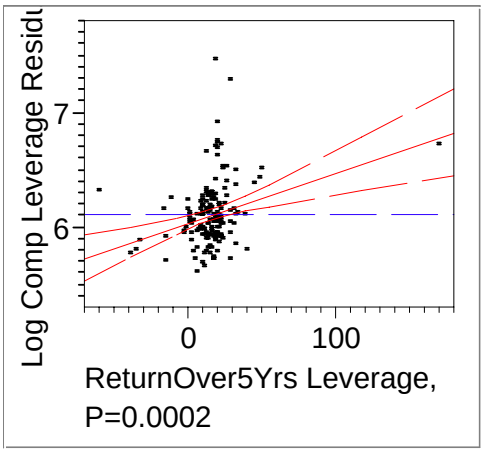
Residual vs Predicted Plot

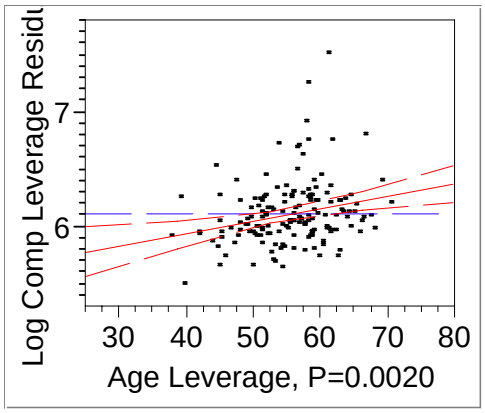


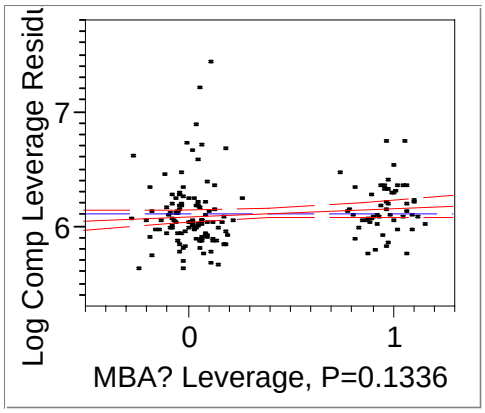
[6] The casebook offers some discussion of stepwise in the context of this data as well. See page 216.

Leverage Plots









Conclusions[7]

[7] The casebook (p 202-219) offers a different analysis of this data, one that is more focused on the role (and value) of the MBA categorical variable.

Building a Regression Model: An Outline

Before You Gather And Look At The Data

Identify the question of interest, the goals of the analysis.

Prediction	In/out of sample? Extrapolation?
	Allowable margin for error?
	What sort of *RMSE* is useful?
Interpretation	Does the estimate "make sense"?
	Is there collinearity? How much?
	Need marginal or partial slope?

Anticipate Important Features Of The Model

Which variables do you expect to find important?
Do you anticipate nonlinear patterns or interactions?
What do you expect the coefficients to be?

Evaluate The Data

Is there enough? (role of preliminary or "pilot" study)
Is the data representative? (sampling biases, coding errors)
Is there a hidden "clumping" factor? A lurking variable?

Assess Univariate Features And Marginal Relationships

Identify scales, ranges, distributions of the various factors.
Are data normal or skewed? Outliers present?
Look at scatterplots, time series plots (if appropriate).
Nonlinear (curvature)? Outliers, leverage points?
Marginal associations with response?
Correlation among predictors? (suggesting collinearity)
Differences among categories? (color coding)

Fit An Initial Model

Modeling is an iterative process. No one gets it right the first time.

> If possible, fit the model suggested by your understanding of the problem, in form that makes the most sense given the context. Sometimes the needs of a "client" can help.
>
> In the absence of a clear way to proceed, stepwise methods can be used to build a preliminary model.

Does model explain much variation in data? (*RMSE*, *F* and R^2)

> Are estimates significant? What is the length of CIs? of PIs?

Evaluate your model graphically.

> Do leverage plots indicate problems? Unusual points?
> Do leverage points, outliers affect the fit?

Are residuals reasonable (i.e., constant variance, normal)? (Don't dwell on these until you get a decent model).

Assess the parameters (slopes, intercept) of the fitted model, focusing on a mixture of statistics and substance.

Can you interpret the slopes, using appropriate units?

> How do the partial and marginal slopes differ?
> What is the impact of collinearity? Can you ignore it?

Revise The Fitted Model As Necessary

Procedure depends on the use of the model

For interpretation, collinearity may be an issue, since it obscures the effects of predictors.

For prediction, don't use factors that are not contributing significantly to the model – they only add error to the prediction. Check these with the t-statistics, effect tests.

Identify other omitted factors

- Are variables appropriately transformed?
- What factors explain the unexplained residual variation?

Use a cautious, one-at-a-time strategy.

- Removing several is dangerous if collinearity is present.
- Check for missed nonlinearity.

Is overfitting a problem. How does the model perform on a left out subset of the data.

Continue revising until satisfied and then

Make sure that you can interpret the end result.
Make sure that you can answer the question of interest.
Run a careful check of residuals

- Does anything in the analysis suggest dependence?
- Do different groups have comparable variance?
- Are they normal (quantile plot from saved residuals)?

Finally - Report your Results

Determine how to communicate results to others. Know your audience.

Do they know statistics?
Do they appreciate subtleties of analysis, such as plots?

What common beliefs does your analysis support? Contradict?

Focus on things that would make analysis simpler, better.

What data are missing?
Which predictors are missing?
Would more data help?

Take-Away Summary

Regression modeling is an *iterative process* that combines substantive criteria with automatic methods like stepwise selection to build models.

Diagnostics like leverage plots and residual plots remain useful for checking for flaws in the model.

Next Module

Regression can be a powerful method for predicting time series. The profitable use of regression requires only a couple of additional ideas.

Department of Statistics
The Wharton School
University of Pennsylvania

Statistics 621 Fall 2005

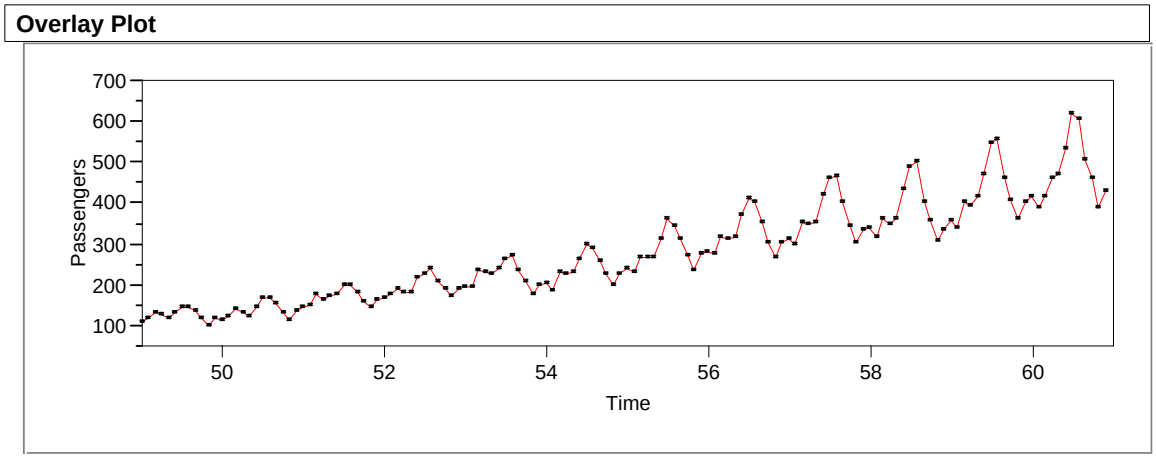
Module 9

Time Series Modeling

When data are a sequence of observations over time, they are called a *time series*. Looking back over previous examples, which of the data sets that we have discussed so far have been time series?

Predicting Airline Passenger Demand Revisited

The file IntlAir-reg.jmp contains monthly passenger data (in units of 1000 passengers) from 1949 to 1960, a period of rapid growth.

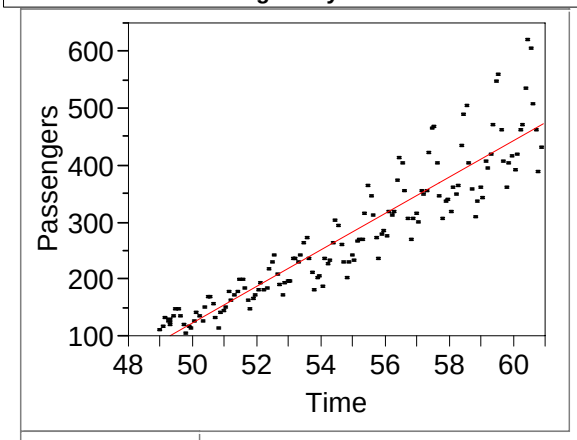


In Stat 603, we looked at relative changes, estimated monthly effects, and came up with a forecast of about 443,000 for January 1961. Multiple regression provides a natural approach for modeling this kind of data.[1]

[1] Box and Jenkins, in their classic 1976 book *Tme Series Analysis*, use this data to motivate another type of model for seasonal data. The form of their model is rather different from the regression model used here and requires methods specialized to the analysis of time series.

We begin with a simple regression of *Passengers* on *Time*, a simple linear model for the trend that captures much of the variation in *Passengers*.

Bivariate Fit of Passengers By Time



Linear Fit

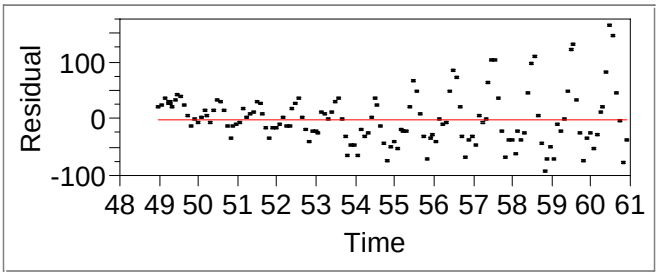
Passengers = -1472.114 + 31.886207 Time

Summary of Fit

RSquare	0.853638
RSquare Adj	0.852607
Root Mean Square Error	46.05715
Mean of Response	280.2986
Observations (or Sum Wgts)	144

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1472.114	61.0141	-24.13	<.0001
Time	31.886207	1.10799	28.78	<.0001

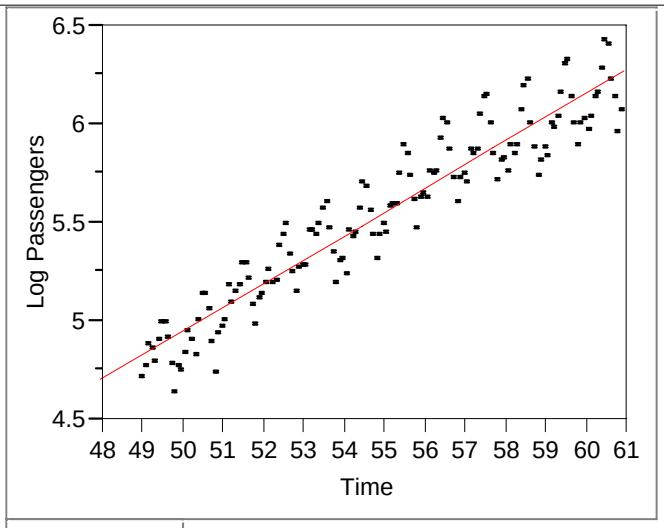


The residuals reveal curvature and heteroscedasticity!

Clearly, structure remains that this simple model ignores.

To eliminate the curvature and heteroscedasticity, let's instead consider a regression[2] of Log *Passengers* on *Time*.

Bivariate Fit of Log Passengers By Time



Linear Fit

Linear Fit

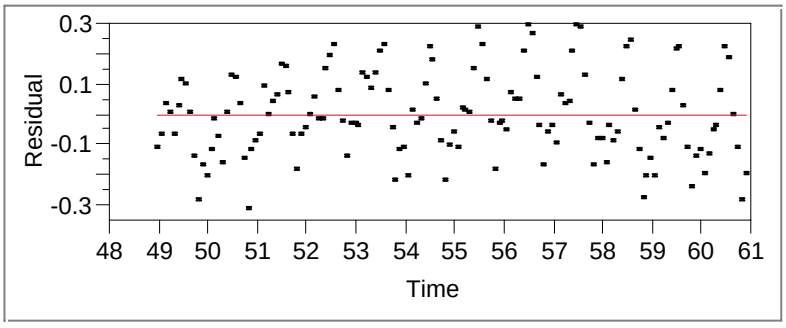
Log Passengers = -1.084732 + 0.1205806 Time

Summary of Fit

RSquare	0.9015
RSquare Adj	0.900807
Root Mean Square Error	0.139037
Mean of Response	5.542176
Observations (or Sum Wgts)	144

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1.084732	0.184189	-5.89	<.0001
Time	0.1205806	0.003345	36.05	<.0001

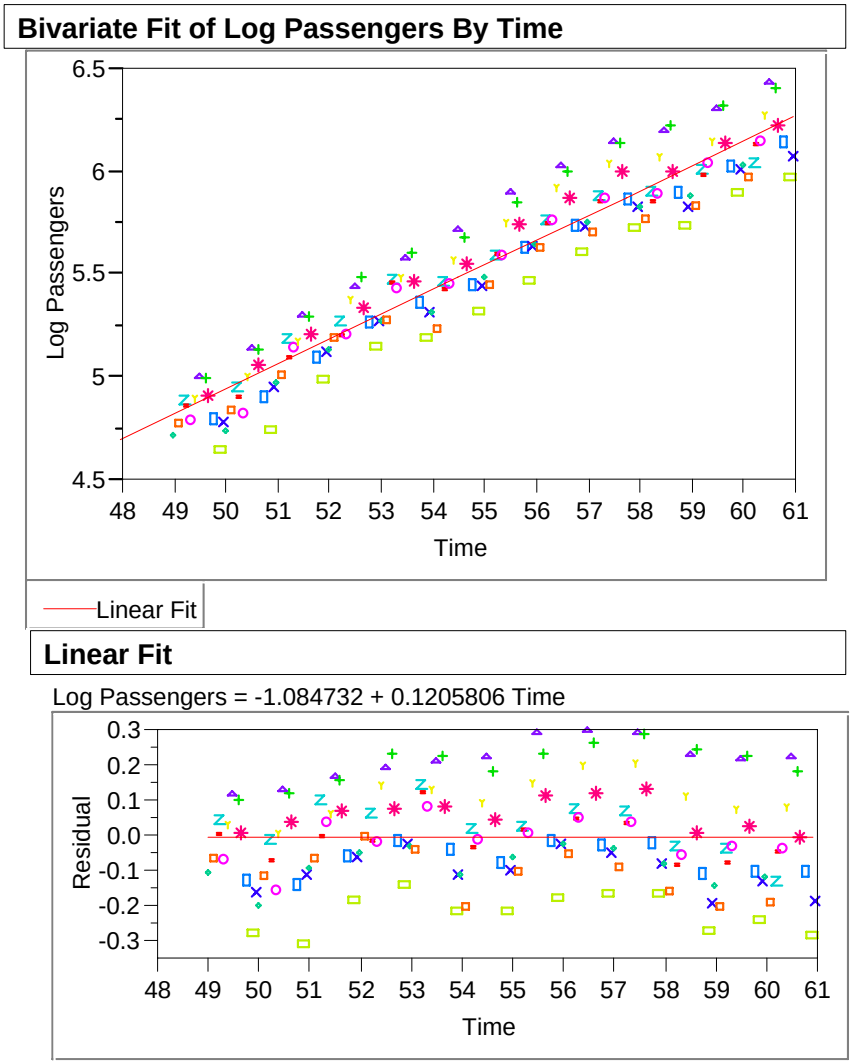


This looks much better, but…

[2] We used the natural log here, but any base will do. For example, the natural log of any value x is 2.3 (= $\log_e$ 10) times the log base 10 of *x*, ln *x* = 2.3 $\log_{10}$ x

Seasonality

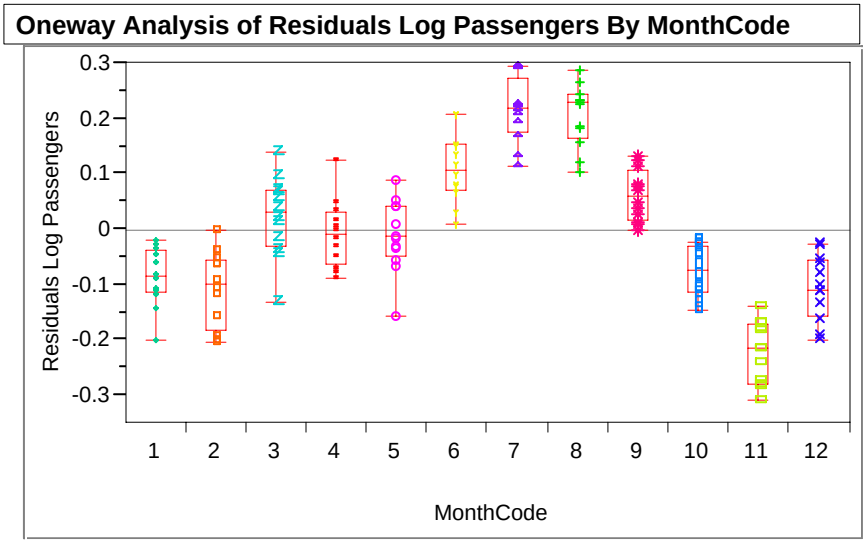
Look what happens when we color code the data by month



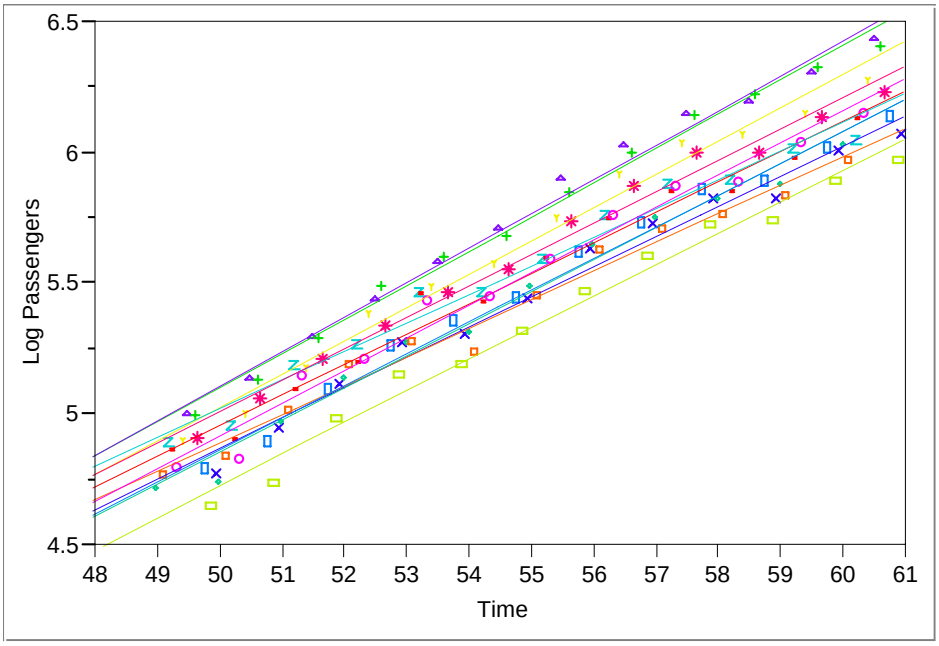
It appears that some months are systematically above or below the regress line.

What assumption of the MRM has been violated?

Another perspective on this residual phenomenon is obtained by plotting the residuals by month[3]



Look what happens when we fit a fit a separate line for each month[4] Do you think we'll need interaction terms?



[3] Use the Fit Y by X command using MonthCode as X to keep the months in order.
[4] Just select the group by option that we have been using when working with categorical variables.

Separate, nearly parallel lines capture variation in the values for each month. Such a systematic monthly effect is called *seasonality*.

To incorporate seasonality in our model, simply add the categorical variable *Month* to the prior simple regression

Response Log Passengers

Summary of Fit

RSquare	0.983468
RSquare Adj	0.981954
Root Mean Square Error	0.059304
Mean of Response	5.542176
Observations (or Sum Wgts)	144

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Time	1	1	25.051608	7123.183	<.0001
Month	11	11	2.284315	59.0475	<.0001

Expanded Estimates

Nominal factors expanded to all levels

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1.098201	0.078833	-13.93	<.0001
Time	0.1208257	0.001432	84.40	<.0001
Month[Apr]	-0.008504	0.016393	-0.52	0.6048
Month[Aug]	0.2059172	0.016392	12.56	<.0001
Month[Dec]	-0.106728	0.016404	-6.51	<.0001
Month[Feb]	-0.107462	0.016399	-6.55	<.0001
Month[Jan]	-0.085407	0.016404	-5.21	<.0001
Month[Jul]	0.2152121	0.016391	13.13	<.0001
Month[Jun]	0.1112698	0.016391	6.79	<.0001
Month[Mar]	0.0227651	0.016396	1.39	0.1674
Month[May]	-0.010876	0.016392	-0.66	0.5082
Month[Nov]	-0.220593	0.016399	-13.45	<.0001
Month[Oct]	-0.076876	0.016396	-4.69	<.0001
Month[Sep]	0.0612826	0.016393	3.74	0.0003

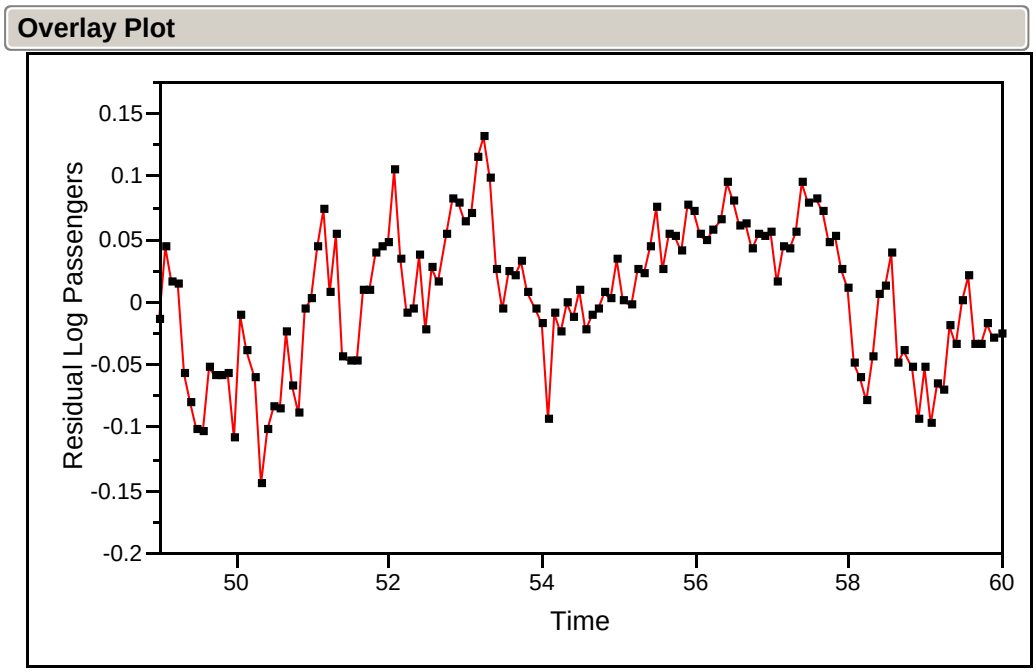
Durbin-Watson

Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
0.4251836	144	0.7788	0.0000

Good news: R^2 has increased from 90.2% to 98.3% and *RMSE* has decreased from to .060 to .026. Both *Time* and *Month* are strongly statistically significant.

Bad news: The Durbin-Watson table[5] above reveals a problem…

This problem can also be seen from the meandering behavior of a time series plot of the residuals



[5] This table is obtained by right-clicking on any title bar and selecting Row Diagnostics > Durbin Watson Test from the pop-up menu. Right click on the red triangle in this table to obtain the p-value.

Residual Autocorrelation

Whenever a regression is performed with time series data, the possibility of sequential dependence between adjacent residuals e_t and e_{t-1} should be checked.

Such dependence between residuals suggests that the underlying true error terms may violate the assumption of independence. We need a way of judging whether the size of this dependence is large (i.e., statistically significant).

A Clever Idea: $e_1,\ldots,e_n$ actually contains many repeated observations of the relationship between adjacent residuals e_t and $e_{t\text{-}1}$, namely

$$(e_1, e_2), (e_2, e_3),\ldots,(e_{n-1}, e_n)$$

This construction allows us to use a scatter plot to study the relationship between e_t and e_{t-1}.

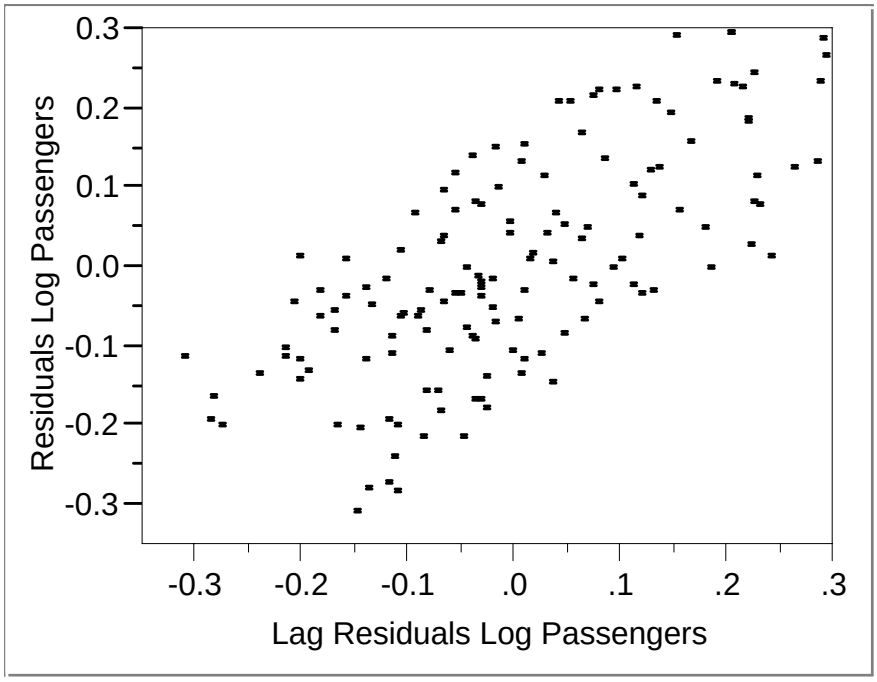
To construct $(e_1, e_2), (e_2, e_3),\ldots,(e_{n-1}, e_n)$ from $e_1,\ldots,e_n$, we create a new variable representing e_{t-1}, which is called a *lagged variable*.

For example, JMP easily creates[6] the variable *Lag Residuals* from the *Residuals*.

The first few rows of those variables look like these; the column with the lag is the original column shifted down one row.

Residuals	Lag Residuals
-0.01409	.
0.04440	-0.01409
0.01622	0.04440
0.01443	0.01622
-0.05728	0.01443

A scatter plot of *Residuals* vs *Lag Residuals* reveals clear linear association between e_t and e_{t-1}.[7]



[6] Use the function Row > Lag in the formula window. Notice that the lag function just "pushes" the column of residuals down by one row. Other lags are also possible.

[7] The casebook has similar analysis, applied to cellular phone growth. See page 306-307.

The correlation between e_t and e_{t-1} is called the *residual autocorrelation.*[8] It is estimated as $r = 0.7788$ shown in the Durbin-Watson table on page 9-6.[9]

Does an estimated residual autocorrelation of $r = 0.7788$ imply that the model errors violate the assumption of independence for the MRM?

To answer this you can use the p-value given in the Durbin-Watson output which tests the hypothesis

$$H_0: \varepsilon_1,\ldots,\varepsilon_n \text{ are independent}$$

From the Durbin-Watson table (p 9-6), *p-value* = . 00 < .05, so we reject the assumption of independent errors.

[8] People also sometimes consider correlations using larger lags. In that context, this is then called the lag 1 residual autocorrelation.

[9] A slightly different formula than the one on p 1-3 in Module 1 is used to compute this autocorrelation.

To reduce autocorrelation, we can simply add the lagged of residuals to the regression model. This is ok since we are using the past to predict the future.

Response Log Passengers

Summary of Fit

RSquare	0.993531
RSquare Adj	0.992879
Root Mean Square Error	0.036922
Mean of Response	5.547936
Observations (or Sum Wgts)	143

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Time	1	1	24.336755	17852.31	<.0001
Month	11	11	1.202899	80.2174	<.0001
Lag Residuals Log Passengers	1	1	0.284485	208.6847	<.0001

Expanded Estimates

Nominal factors expanded to all levels

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1.070768	0.049623	-21.58	<.0001
Time	0.1203201	0.000901	133.61	<.0001
Month[Apr]	-0.026253	0.010283	-2.55	0.0118
Month[Aug]	0.0356479	0.015586	2.29	0.0238
Month[Dec]	0.068731	0.015876	4.33	<.0001
Month[Feb]	-0.039474	0.011252	-3.51	0.0006
Month[Jan]	-0.004981	0.011943	-0.42	0.6773
Month[Jul]	0.1273505	0.011878	10.72	<.0001
Month[Jun]	0.120253	0.010228	11.76	<.0001
Month[Mar]	0.1082698	0.011809	9.17	<.0001
Month[May]	-0.003801	0.010222	-0.37	0.7106
Month[Nov]	-0.159138	0.011066	-14.38	<.0001
Month[Oct]	-0.125016	0.010738	-11.64	<.0001
Month[Sep]	-0.101589	0.015203	-6.68	<.0001
Lag Residuals Log Passengers	0.7930716	0.054899	14.45	<.0001

Durbin-Watson

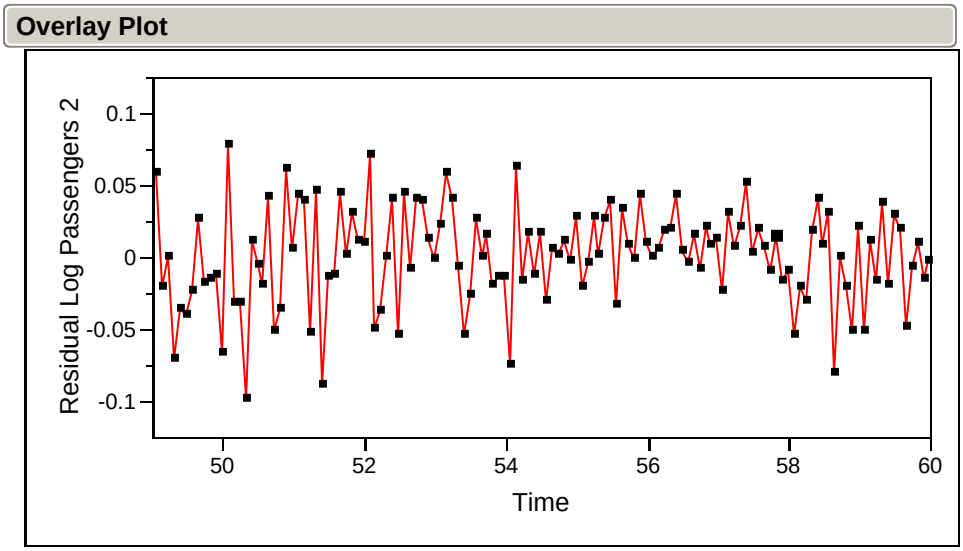
Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
2.191914	143	-0.1089	0.8478

The Durbin-Watson output shows a much smaller degree of dependence in the residuals. This looks much better!

From the effects table, all three predictors *Time*, *Month* and *Lag Residual Log Passengers* are significant.

The residual autocorrelation, now $r = -0.109$, is not significant since p-value = .85 > .05. Thus, the independent error assumption of the MRM is now reasonable.[10]

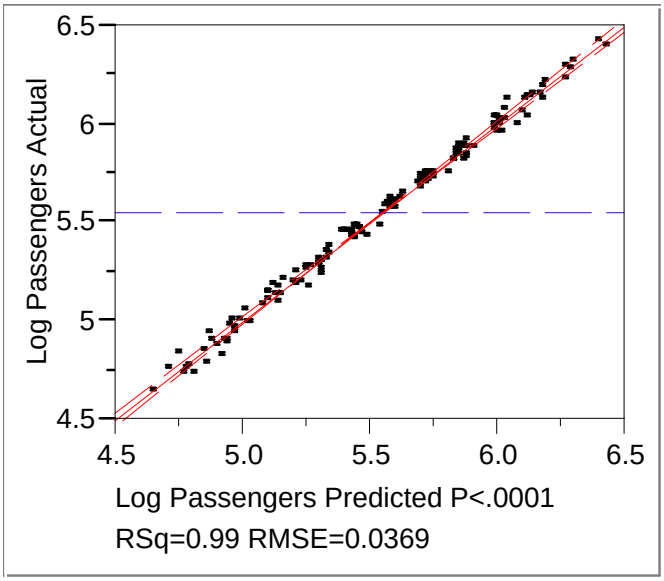
This is further confirmed by the time series plot of the residuals which is no longer meandering

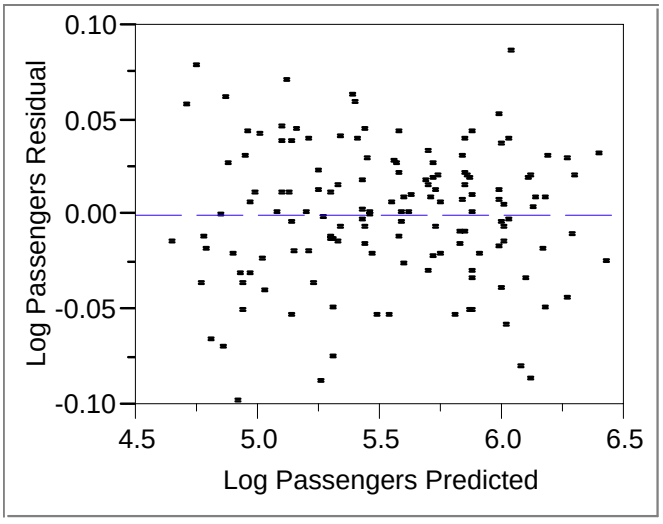


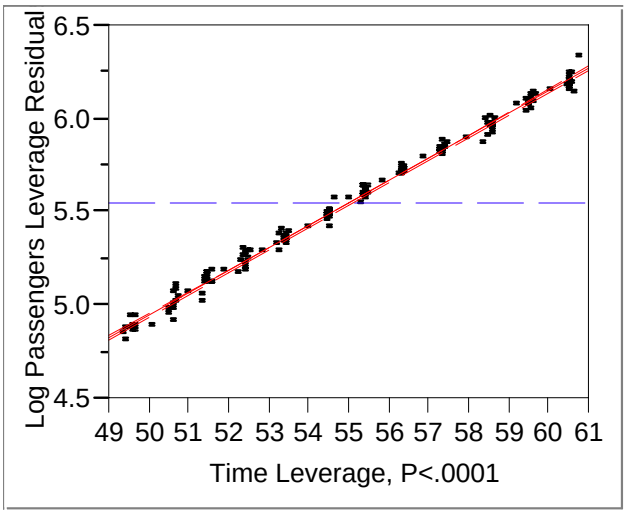
[10] And it needs to be! If this model has correlated errors, the least squares estimates becomes rather unreliable. But this is the subject of another course, such as Stat 701.

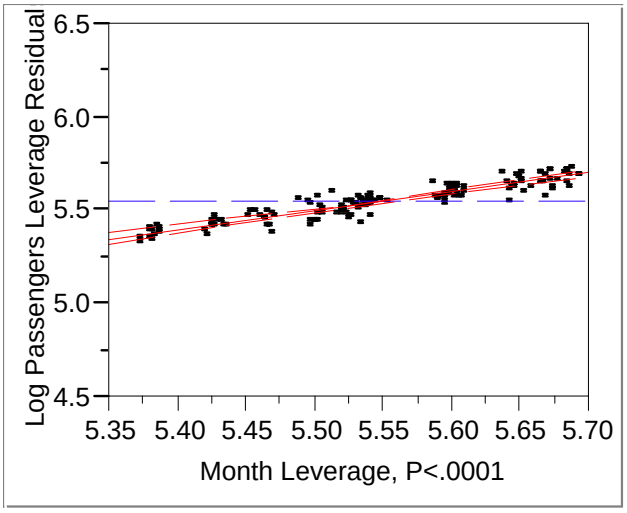
Of course, the usual model checks must be carried out to make sure the other aspects of the MRM model are reasonable.

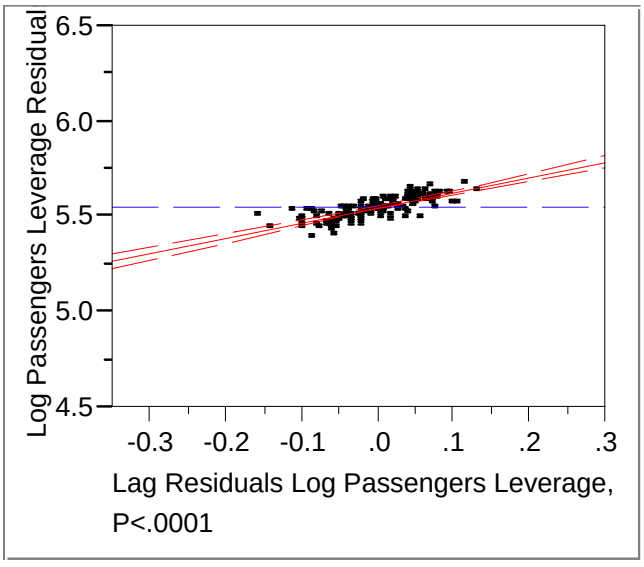
The diagnostic plots reveal no disturbing anomalies.



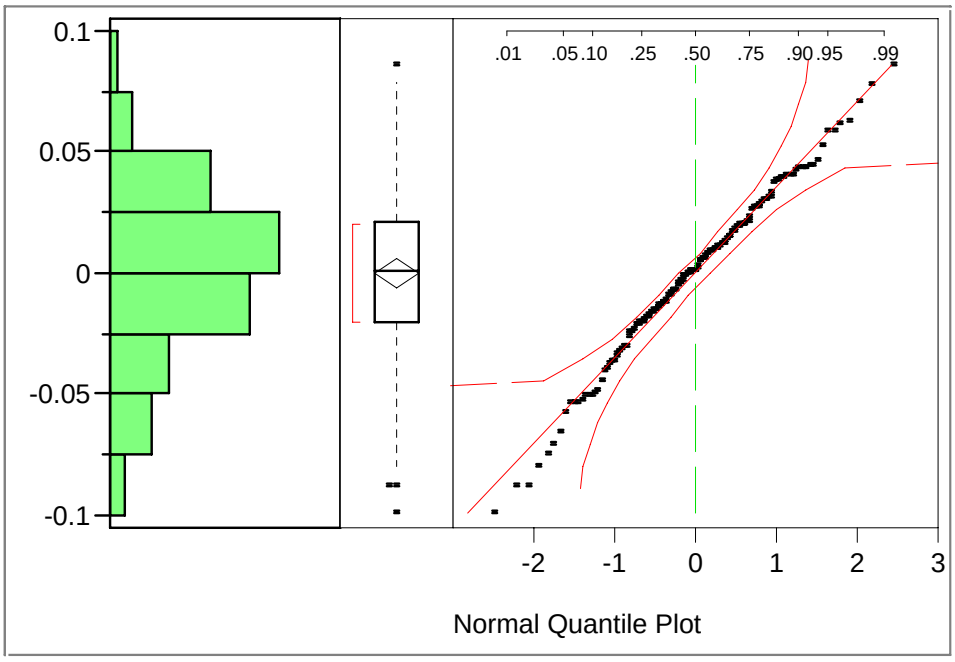



Log Passengers Leverage Residual
6.5
6.0
5.5
5.0
4.5
49 50 51 52 53 54 55 56 57 58 59 60 61
Time Leverage, P<.0001


Log Passengers Leverage Residual
6.5
6.0
5.5
5.0
4.5
5.35 5.40 5.45 5.50 5.55 5.60 5.65 5.70
Month Leverage, P<.0001


Log Passengers Leverage Residual
6.5
6.0
5.5
5.0
4.5
-0.3 -0.2 -0.1 .0 .1 .2 .3
Lag Residuals Log Passengers Leverage, P<.0001

And the normal quantile plot does not contradict the normality assumption of the MRM.



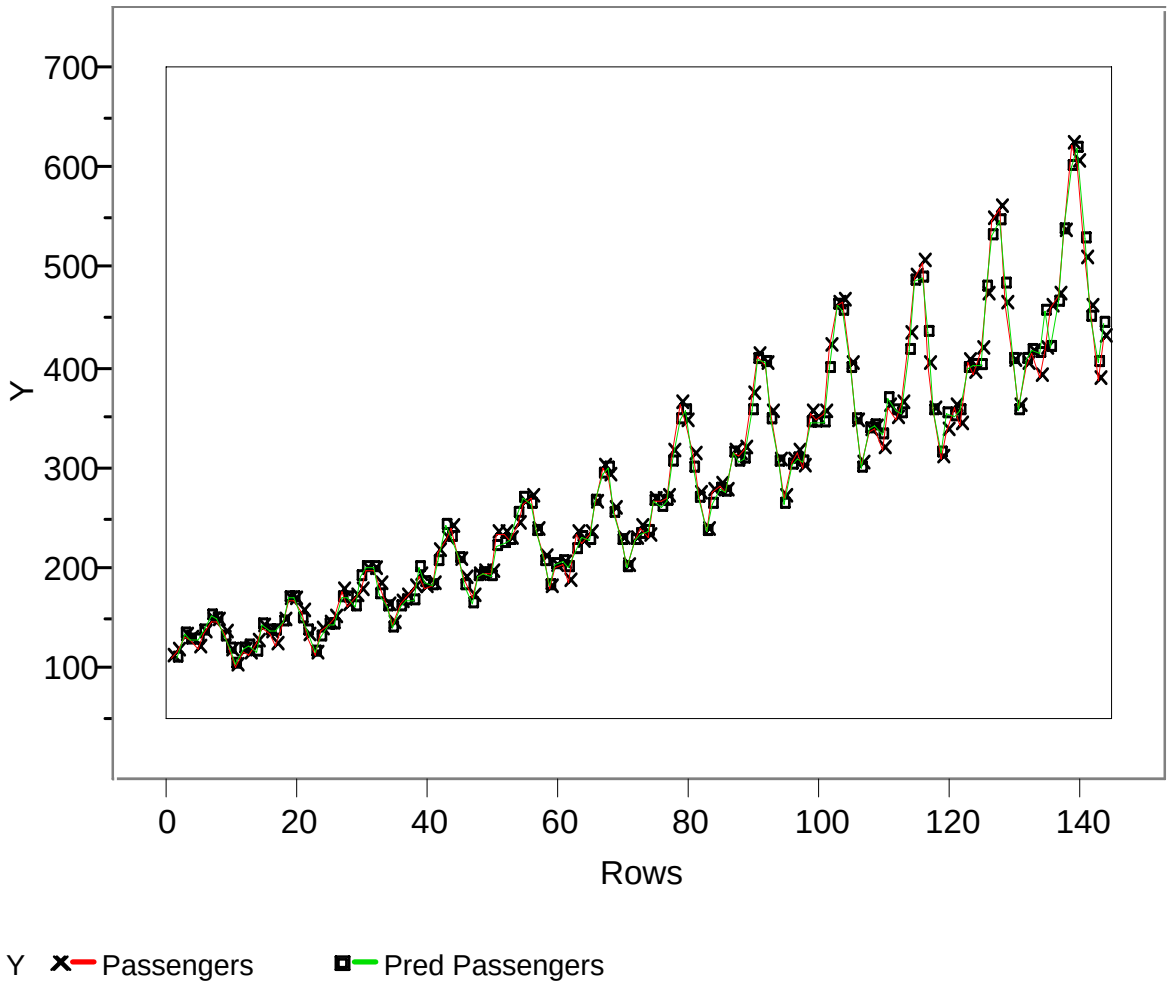
Forecasting

We are ready to use the model for forecasting.

Letting $\hat{y}_t$ stand for the predicted value at time *t* of Log Passengers from the fitted model on p 8-10, the predicted value of Passengers is obtained as

$$e^{\hat{y}_t}$$

An overlay of time series plots of *Passengers* vs *Time* and *Predicted Passengers* vs *Time* shows remarkable fit.



Finally, to forecast *Log Passengers* for January 1961, we simply fill in the appropriate *x* values (see the last row of IntlAir-reg.jmp) and obtain[11]

$$\hat{y}_{\text{Jan 1961}} = 6.11$$

The forecast of *Passengers* for January 1961 is then obtained as

$$e^{6.12} \approx 451$$

which is slightly larger than our previous forecast of 443 from Stat 603.

Better yet, we can add a prediction interval. To get the prediction interval, we'll use JMP to compute the width of the prediction interval for the log response.[12] Once we have the interval on the log scale, we can "unlog" the endpoints of the interval and get a prediction interval for the number of passengers itself.

Starting from (6.034, 6.182), the 95% prediction interval for *Log Passengers*, the 95% prediction interval for *Passengers* is approximately (418, 487).

[11] The prediction appears in the column obtained by Save Columns > Prediction Formula from the pop-up menu.

[12] To compute this for a predicted point in JMP, it is necessary to manually insert the prediction to back into the response column and then rerun Fit Model. Alternatively, the approximation $\hat{y}_{\text{Jan 1961}} \pm 2$ RMSE works just fine here.

Going Further

To see some other examples of time series methods, take a look in Class 12 of the casebook. It gives more examples (such as for the cellular telephone data) of these methods and related ideas.

To learn a lot more about modeling time series data, consider the course Stat 711 which focuses on models designed for prediction.