Department of Statistics
The Wharton School
University of Pennsylvania

Statistics 621 Fall 2003

Module 2

The Simple Regression Model

The Signal-Noise Decomposition

One of the ideal setups in Statistics occurs when data

$$y_1,\ldots, y_n$$

can be treated as sequence of independent draws from a normal distribution with mean μ_y and standard deviation σ_y.

This statistical *model* for such data is denoted by

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

where *iid* stands for "independent and identically distributed".

Example

In Stat 603 we saw that the *iid* normal model was reasonable for the 1992-1993 daily returns on GM (see GM92.jmp).

There, based on n = 507 observations, we estimated

$$\mu_y \text{ by } \bar{y} = .00158, \quad \text{and} \quad \sigma_y \text{ by } s_y = .0202$$

Somewhat more suggestively, this model can also be written as

$$y_i = \mu_y + \varepsilon_i, \quad i = 1,\ldots, n$$
$$\varepsilon_1,\ldots,\varepsilon_n \ \ iid \sim N(0, \sigma_y^2)$$

Notice how the *data generating process* has two components[1]:

1) "signal": a fixed level μ_y

2) "noise": $\varepsilon_1,\ldots,\varepsilon_n$ iid mean 0 normal deviations

Although the *iid* normal model above is not always appropriate, it is a special case of a broadly applicable model formulation

$$y_i = signal_i + \varepsilon_i, \quad i = 1,\ldots, n$$
$$\varepsilon_1,\ldots,\varepsilon_n \ \ iid \sim N(0, \sigma_\varepsilon^2)$$

Again, the data generating process has two components:

1) the signal: $signal_i$
2) the noise: $\varepsilon_1,\ldots, \varepsilon_n$ iid mean 0 normal deviations

Note the three main properties of the noise $\varepsilon_1,\ldots, \varepsilon_n$

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

[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.

Although the decomposition

$$y_i = signal_i + \varepsilon_i$$

is not explicitly observed, a general strategy for finding such a model is based on finding a decomposition of the form

$$y_i = \hat{y}_i + e_i, \quad i = 1,\ldots, n$$

where $\hat{y}_i$ estimates $signal_i$

and $e_i = y_i - \hat{y}_i$ estimates noise $\varepsilon_1,\ldots,\varepsilon_n$.

Support for a particular form for $signal_i$ is obtained when $e_1,\ldots,e_n$ manifest *iid* normal behavior.

Jargon

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

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

As we shall now see, regression analysis falls exactly into this framework.

The Simple Regression Model (SRM)

Under this idealized statistical model, the data

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

are 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:

As a decomposition of the data into signal & noise, the signal here is

and the noise here is

People also sometimes refer to $\varepsilon_1,\ldots,\varepsilon_n$ as the "errors".[2]

[2] You can also think of these errors 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.

Think of the SRM as a hypothetical process that could have generated 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 $\sigma_\varepsilon = 1$

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

β_0, β_1 and σ_ε are the (usually) unknown parameters of the SRM. An objective of regression is to estimate them.

[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.

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.

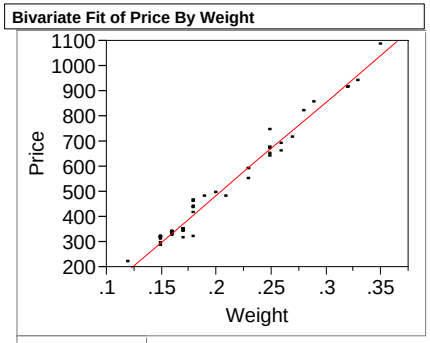
Weight	Price	Predicted Price	Residuals Price
0.17	355	372.95	-17.95
0.16	328	335.74	-7.74
0.17	350	372.95	-22.95
0.18	325	410.16	-85.16
0.25	642	670.63	-28.63
0.16	342	335.74	6.26
0.15	322	298.53	23.47
0.19	485	447.37	37.63
0.21	483	521.79	-38.79
0.15	323	298.53	24.47
0.18	462	410.16	51.84
0.28	823	782.26	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

Does the decomposition $y_i = \hat{y}_i + e_i$ hold here?

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

Looking at more of the output 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

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*. It is obtained by the formula

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

$RMSE^2$ 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[5].

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

RMSE measures the dispersion of the residuals 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 two *RMSE* of the LS line

[5] 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

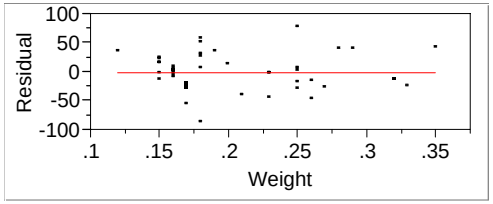
2. A scatterplot of the Residuals vs *x* should appear

3. A histogram of the residuals should appear
and a normal quantile plot of the residuals should appear

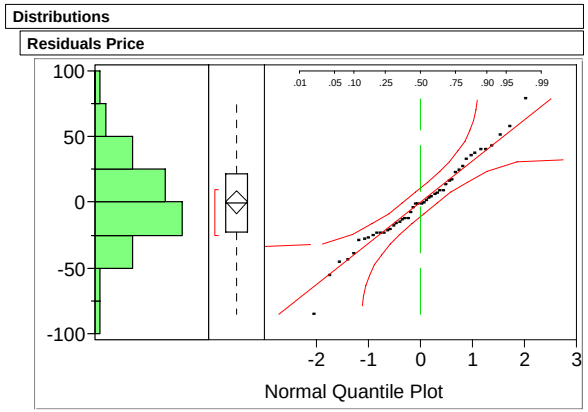
Example: Checking the Diamond Regression

The scatterplot of Price vs Weight (p 2-9) reveals

The scatterplot[6] of Residuals vs Weight reveals



The histogram and normal quantile plot of the residuals shows



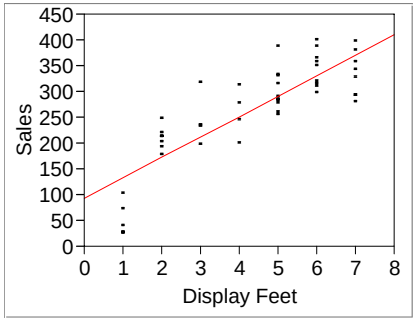
[6] After executing the Fit Line subcommand, right click on the triangle next to "—Linear Fit" and select Plot Residuals from the Pop-up menu to obtain this plot.

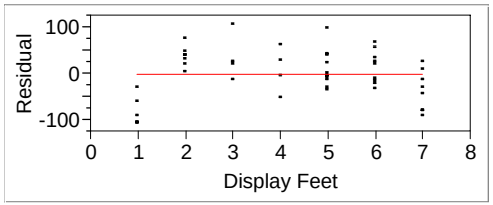
Anomalies to Look For

Nonlinearity

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

Recall the display.jmp data





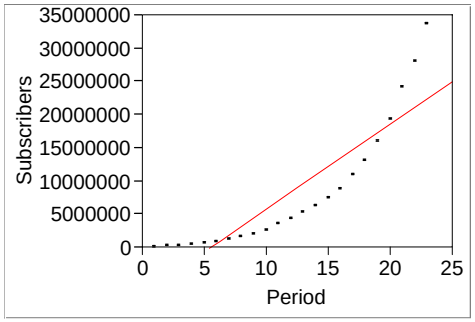
Remedy: Transform y and/or x.

Autocorrelated Residuals

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

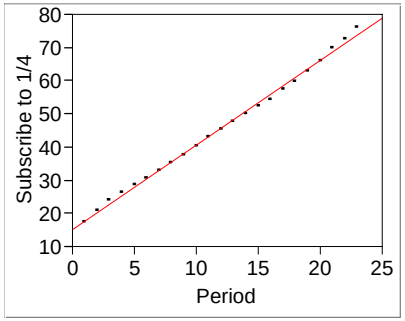
Thus one might consider fitting a trend model of the form

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

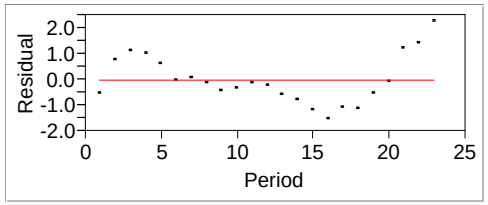
In this special case of the SRM, t plays the role of x.

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

At first glance the regression of $y^{1/4}$ on t appears to be wonderful.



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

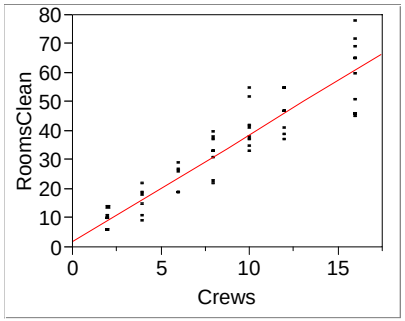


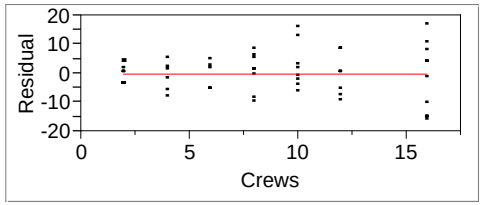
What SRM assumption has been violated?

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

Expanding Residuals

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.

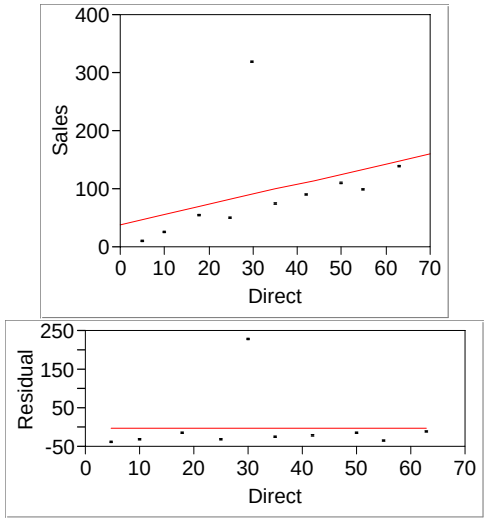
Remedy: Transform y or use weighted least squares (p 57-60) instead of least squares.

Outliers and Influential Points

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

The file direct.jmp contains the level of sales ($1000's) and the number of direct mail recipients (1000's) for 10 different mailings of a catalog.

Each catalog costs $1.50 and the company would like to assess its marginal profit increase per catalog.



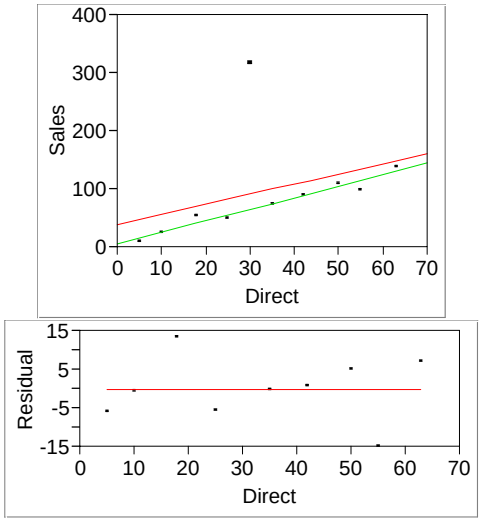
The LS line here is

Sales = 39.58 + 1.73 Direct

with RMSE = 85.7

Investigation of the unusual point above reveals that that mailing coincided with a large inventory sale.

Repeating the regression with the point excluded yields



The LS line here is

Sales = 5.78 + 1.98 Direct

with RMSE = 8.8

What changed?

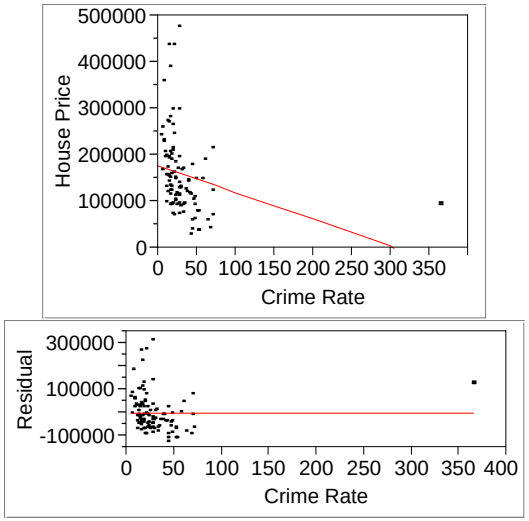
Should the outlier be excluded?

What is the company's estimated marginal profit per catalog?

Another Outlier Example

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

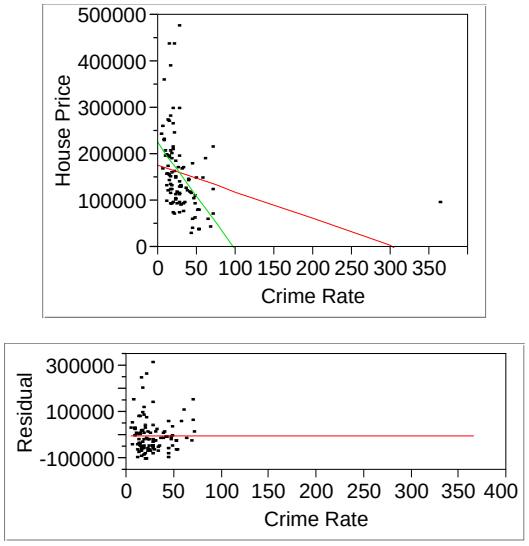
House Price = 176629 – 577 Crime Rate

with RMSE = 84325. Interpretation?

The unusual point is[8]

[8] 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.

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?

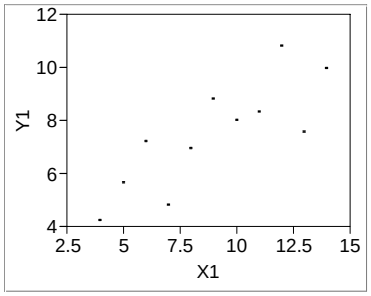
Note how one point can drastically influence a regression. Should this point be excluded?

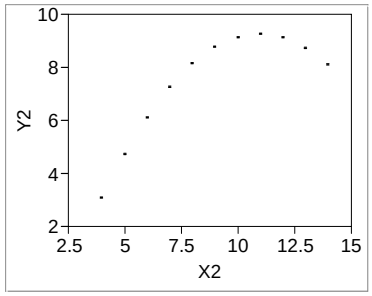
How does the unusual point affect the fit here, compared to the effect of the outlier in the previous direct mail regression?

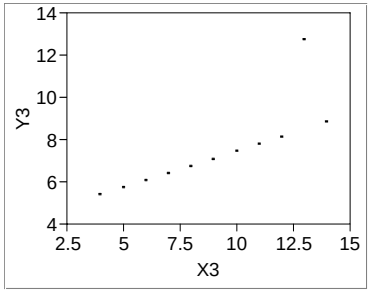
Don't forget to plot the data!

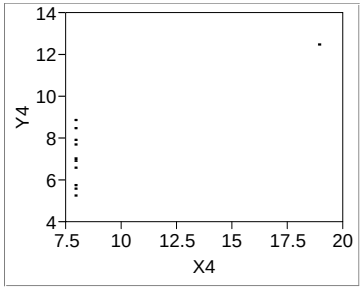
Before fitting a regression, it is crucial to first plot the data.

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



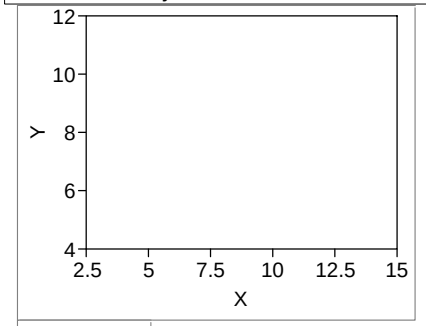






Which of the previous scatter plots yields the following regression output?

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.