Class 1. Simple linear regression

Concise regression review notes are available from Stat608 1997 homepage.[1]

1 Today

Review of simple linear regression.

New idea. Heavy tailed residual distributions and what they may indicate.

Extending categorical variables to "broken stick" models.

Illustrations: 30 year returns on gold.

2 The regression set-up

The model for the mean relationship:

$$Av(Y|x) = \beta_0 + \beta_1 x.$$

The model for the raw data:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

This is the straight line or *linear* model.
Assumptions are on the ϵ_i.

Independent

Constant variance. Mean zero.

Approximately normally distributed.

Biggest problems. Dependence, skewness and non-constant variance.

Call the ϵ_i the "true error terms".

[1] http://stattemp.wharton.upenn.edu/ waterman/Teaching/608f97/

Distance from point to the true line. $y_i - (\beta_0 + \beta_1 x_i)$.

We don't know them as we don't know the regression line.

Substitute with the "residuals", estimated error terms.

Distance from point to estimated regression line. $y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{y}_i$.

ALWAYS check assumptions on the residuals.
Why so important?

Standard least squares regression is sensitive to individual data points.

A single point can dominate the regression.

Everything you say and conclude may be driven by a single data point.

Residual plots are one of the tools available to help identify these points.

Even if you keep it, it is important to know that it is there.

Inference, p-values, CI's etc only has validity if assumptions hold.

Key diagnostics.
1. Residual plot. Good plots lack structure.
2. Normal scores plot of the residuals.

3 More depth – the normal scores plot

Most model diagnostics (here the model is the normal distribution for the error terms) compare reality (what we observe) to theory (what we expect). In general OBSERVED versus EXPECTED.

This is how the normal scores plot is constructed.

On the X-axis is what we expect.

On the Y axis is what we observe.

The idea is simple: say there were 100 observations (n = 100) and therefore 100 residuals.

The model says that the residuals come from an approximate normal distribution.

Now order the residuals from lowest to highest.

Where would you expect the smallest of 100 observations from a normal distribution to lie?

Plot where you expect it to be against where it actually is.

Repeat for the other 99 points.

If the model is correct than theory and reality should coincide, ie observed equals expected and the points should roughly (because there's inherent variability) lie along a line.

Extension.

There is no reason why for the X-axis we have to use the normal distribution, perhaps the data has a gamma distribution (useful for life length data). You just calculate where you EXPECT the data to be if a gamma distribution is true. These more general plots are called Quantile-Quantile[2] or Q-Q plots.

4 Heavy tailed residuals

The ends of the normal scores plot have greater slopes than the reference line because the observations in the tails are spreading out more than the normal theory predicts.

One reason for heavy tails.[3] The residuals come from TWO groups with different variances. Always leads to heavy tails. Interpretation: two different volatility regimes, low and high.

The volatility regime model, in which trades take place at different rates over the course of a day, relies on knowing the key facts that

The sum of normal random variables is again normal

The variance of a sum is the sum of the variances when the observations are independent

[2] http://www-stat.wharton.upenn.edu/ waterman/Teaching/701f97/quantile.html
[3] http://www-stat.wharton.upenn.edu/ waterman/Teaching/701f97/goldvol.html

Table 1: Comparing hourly returns and the daily return.

HOUR	0	1	2	3	4	5	6	7
Value	100.0000	99.4819	99.6610	100.6495	101.5622	102.7503	100.9752	101.29
Hourly return	***	-0.0052	0.0018	0.0099	0.0091	0.0117	-0.0173	0.003

The return over a fixed period (say a day) is approximately equal to the sum of returns over contained shorter periods (say hours)

Example:

Daily return = 0.0129.
Sum of hourly returns = 0.0131

The sum of the hourly returns approximates the daily return.

If the hourly returns are approximately normal, then so is their sum, but this is just the daily return. The daily return is a sum, and the variance of a sum is the sum of the variances (provided the returns are independent), so if you buy into the model then daily returns have more variance than the hourly returns.

Take home:graphical observation generates sensible questions.

5 Categorical variable regression

Enables comparisons between groups while accounting for other related variables – Amazonian Indian Stress Study.

5.1 Parallel lines

First case: a single dichotomous variable. Our example: Pre 1980 vs Post 1980.

The way JMP does it (contrast = sum in S-Plus): model

$$Av(Y|x) = \beta_0 + \beta_1 x + \beta_2 z$$

where z = 1 if observation is in the first group and -1 if observation is in the second group.

Check to understand the model: plug in z = 1 and -1.

Group 1 model

$$Av(Y|x, z = 1) = \beta_0 + \beta_1 x + \beta_2 \times 1.$$

$$Av(Y|x, z = 1) = \beta_0 + \beta_1 x + \beta_2.$$

$$Av(Y|x, z = 1) = (\beta_0 + \beta_2) + \beta_1 x.$$

Group 2 model

$$Av(Y|x, z = -1) = \beta_0 + \beta_1 x + \beta_2 \times -1.$$

$$Av(Y|x, z = -1) = \beta_0 + \beta_1 x - \beta_2.$$

$$Av(Y|x, z = -1) = (\beta_0 - \beta_2) + \beta_1 x.$$

Compare Group 1 and Group 2.

$Av(Y|x, z = 1) - Av(Y|x, z = -1)$ is the difference in height between the two regression lines.

Notes.

Both groups have the same slopes (β_1) – parallel lines.

The difference in heights is $2\beta_2$.

Note that β_1 represents a comparison against the "norm".

5.2 Interaction terms in categorical variables

Interaction: a three variable concept. One Y and two X's. X1 and X2.

The impact of X1 on Y depends on the level of X2.

In the gold example; the impact of SP500 return on gold return depends on the date (Pre 1980 vs. Post 1980).

In formulae, denote th categorical variable by z, and let z = 1 for group 1 and z = -1 for group 2. Then for group 1:

$$Av(Y|x, z) = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 x \times z.$$

$$Av(Y|x, z = 1) = \beta_0 + \beta_1 x + \beta_2 \times 1 + \beta_3 x \times 1$$
$$Av(Y|x, z) = \beta_0 + \beta_2 + (\beta_1 + \beta_3)x,$$

and group 2

$$Av(Y|x, z = -1) = \beta_0 + \beta_1 x + \beta_2 \times -1 + \beta_3 x \times -1$$
$$Av(Y|x, z) = \beta_0 - \beta_2 + (\beta_1 - \beta_3)x.$$

Hence $2\beta_2$ is the difference in intercepts and $2\beta_3$ is the difference in slopes.

When doing categorical variable regression always check the residuals for each group.

Comparison boxplots are good for this.

There are many different types of coding schemes for categorical variables. We will investigate them in more detail.

Example from the Gold data set.

Consider Q-Q plot of one set of residuals against the other.

5.3 Broken stick regression

Useful for systems that may suffer a "shock".

Another application of categorical variables.

Model:

$$Av(Y|x, z) = \beta_0 + \beta_1 x + \beta_2 \times z \times (x - T),$$

where z = 0 if $x < T$ and z = 1 if $x \geq T$, and T is the "breakpoint".

Case 1, $x < T$, plug in to get

$$Av(Y|x, z) = \beta_0 + \beta_1 x + \beta_2 \times 0 \times (x - T),$$
$$Av(Y|x, z) = \beta_0 + \beta_1 x,$$

Case 2, $x \geq T$, plug in to get

$$Av(Y|x, z) = \beta_0 + \beta_1 x + \beta_2 \times 1 \times (x - T),$$
$$Av(Y|x, z) = \beta_0 + \beta_1 x + \beta_2 (x - T),$$
$$Av(Y|x, z) = \beta_0 - \beta_2 T + (\beta_1 + \beta_2)x,$$

Slope before T is β_1, slope after T is $\beta_1 + \beta_2$.

Therefore β_2 measures the change in slope after time T.

Implementation in S-Plus.

1. Create the indicator variable column z.

2. Create a new column by multiplying column z by column x.

3. Run the regression with column x and the "product column".

Technical name: Piecewise linear regression.
Issues: searching for the breakpoint.

6 What you need to know

The basic graphics for simple regression

The interpretation of the regression coefficients

The interpretation of the number summaries, R^2, RMSE.

Inference, confidence intervals for the slope, p-values

7 Next time

Start multiple regression.

We will use this to review transformation, prediction and categorical variables with more than two groups.