Stat 102

Introduction to Business Statistics
Class 10

Richard P. Waterman

Wharton

Table of contents I

1. Today's module

2. The SRM defined
   - The mean function
   - The error terms
   - The regression story

3. The utopian situation

4. Standard error and inference

5. Summary

6. Next time

Today's module

Topics to be covered in this module:

- Last time
- The SRM structure and assumptions
- Utopia – the best situation possible
- Checking the assumptions and the modeling process
- Standard error for the slope
- Confidence intervals and test statistics
- Summary
- Next time

Last time

- Defining the log transform
- Prospecting for a log relationship
- Fitting log models
- Optimizing with log models
- Four model possibilities and the interpretation of the slope in each one

Big picture

- We want to do *inference* in regression.
- In particular, is the regression line flat (does knowing X tell you anything about Y)?
- Therefore we are going to need some assumptions to do the inference.
- Today we are going to describe those assumptions.

The structure of the mean

We will follow the paradigm:

$$Data = Signal + Noise$$

and discuss each part in turn.
For the signal (the sytematic part of the model) we claim:

$$E(Y|X) = \beta_0 + \beta_1 X.$$

Note the β's rather than b's. This is the *population/sample* paradigm again.
In English: the **mean** of Y as a function of X, lies along a straight line.
You can use transformed versions of X and Y for more exotic relationships, but the key point is that in a regression, the **mean** of the outcome is modeled as a linear function of the predictor.

The noise in the system

We believe that the deviations (ϵ_i) about the true regression line have the following properties:

1. Independent
2. Constant variance
3. Normally distributed

These deviations constitute the *noise* in the system.
Where do they come from in practice? Possible explanations:

1. Measurement error
2. Unobserved/missing predictors

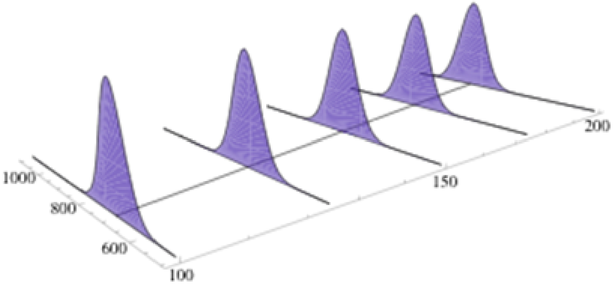
Putting it together

The Simple Regression Model (SRM) states that:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where ϵ_i are *iid* and

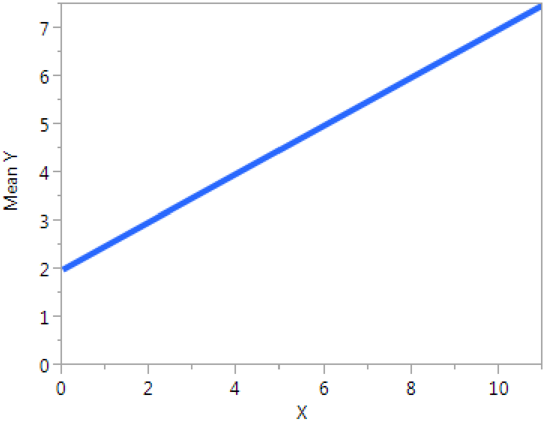
$$\epsilon_i \sim N(0, \sigma_\epsilon^2).$$



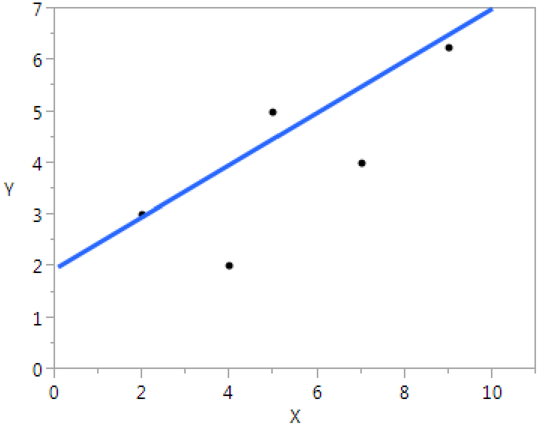
The regression story

There is a truth, but we don't know it, that is the true regression line: $\beta_0 + \beta_1 x$. The *true blue line*.

$$E(Y|X) = 2 + 0.5x.$$

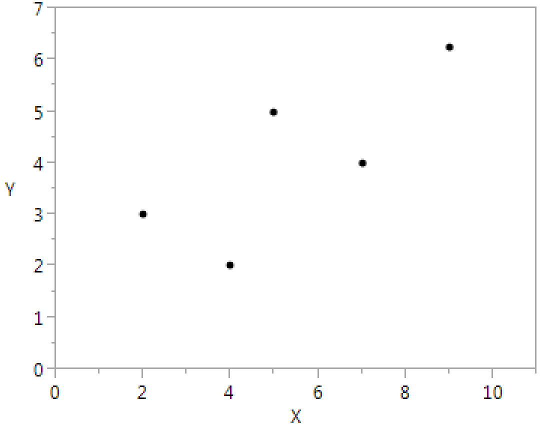


Data is generated according to this model



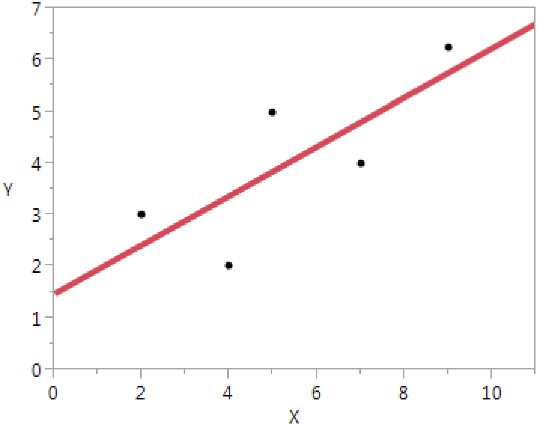
The points have a normal distribution about this true line (the errors, ϵ_i).

The analyst's problem



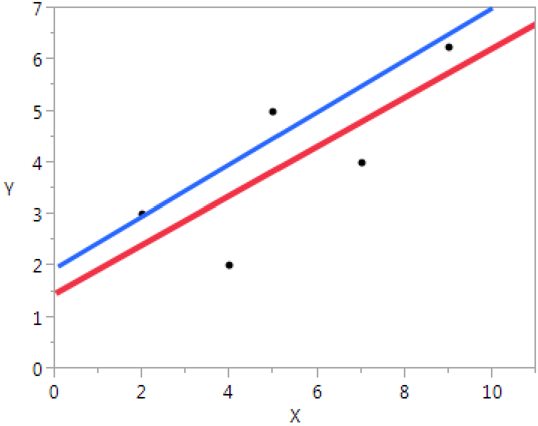
The analyst only has the data and tries to reconstruct the *true blue line* using the method of least squares.

Overlaying the least squares line on the data



The least squares line is the red line. As humans, this is the best we can do: $\hat{y} = 1.48 + 0.48x$.

Hopefully least squares is close to the truth



We use the red line (least squares) as an approximation to the true blue line.

The errors and the residuals

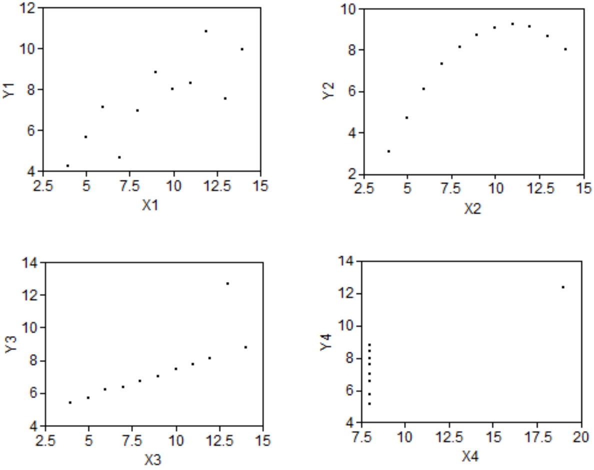
- The errors (ϵ_i) are the vertical distances from the points to the true regression line (the blue line).
- We just made assumptions about the errors terms, but we never see them, because we never see the true regression line. This seems like a problem.
- But, if the least squares line is close to the true regression line (red close to blue) then the distance from the point to the red line must be close to the distance from the point to the blue line.
- The distance from the point to the red line is the *residual* (e_i).
- Thus, we can use the residuals as proxies for the ϵ's.
- That is:

$$e_i \approx \epsilon_i.$$

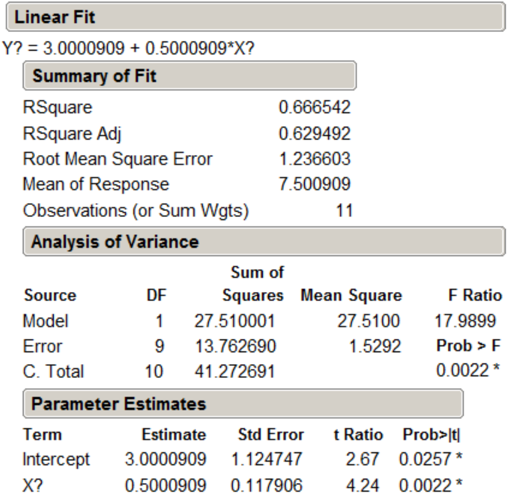
- We make assumptions on the error terms but check them using the residuals from the least squares regression.

Don’t forget to plot the data

Four qualitatively different relationships between X and Y:



What do you learn from the numerical output?



```
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
                          Sum of
  Source        DF       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 *
```

Good modeling practice

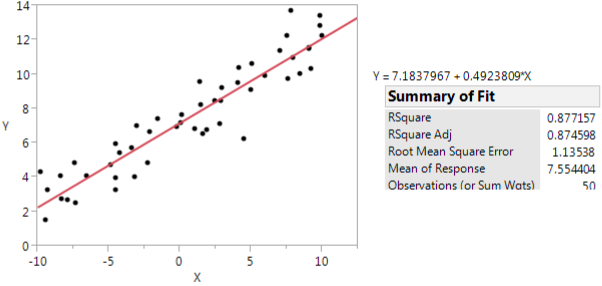
1. Create a scatterplot of Y against X. Assess linearity and transform if necessary.
2. Fit the regression and obtain residuals (e_i).
3. Plot residuals looking for potential problems:
    - Residuals against X
    - Residuals against time (if a time-series)
    - NQP of the residuals
4. Inspect plots for outliers and explain if possible.

The Utopia example

The best it can get. Truth:

$$E(Y|X) = 7 + 0.5X.$$

$$\epsilon_i \sim N(0,1).$$



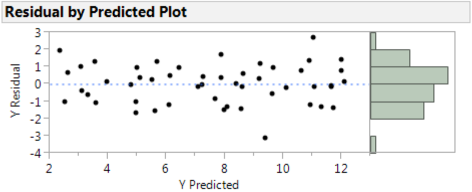
Y = 7.1837967 + 0.4923809*X

Summary of Fit

RSquare	0.877157
RSquare Adj	0.874598
Root Mean Square Error	1.13538
Mean of Response	7.554404
Observations (or Sum Wgts)	50

Note that the Least Squares (LS) line is a good approximation to the true line and RMSE is attempting to estimate the standard deviation of the error terms, that is 1.

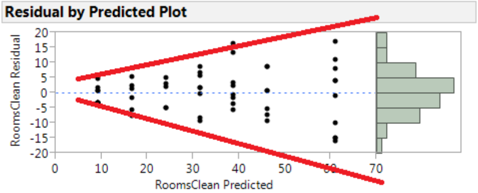
The good residual plot from Utopia



Note the lack of structure. Not exactly surprising as we faked the data exactly from the SRM.

A bad residual plot from Cleaning

Note the increasing variance with X. This breaks the *constant variance* assumption.



Standard error and inference

To do inference you need a standard error (the measure of the sample-to-sample variation in the estimate).

In regression we have

$$SE(b_1) = \frac{\sigma_\epsilon}{\sqrt{n-1}} \times \frac{1}{s_x}.$$

σ_ϵ is the standard deviation of the error terms, and s_x is the standard deviation of the x-variable.

As we don't know σ_ϵ, we use its estimate RMSE (s_e in Stine's notation).

The approximation you need to know is:

$$se(b_1) \approx \frac{RMSE}{\sqrt{n}} \times \frac{1}{s_x}.$$

Drivers of the standard error of the slope

Three drivers:

1. The inherent noise is the system as measured by RMSE.
2. The sample size, n.
3. The spread of the Xs.

Inference for b_1

This is the key regression hypothesis: does knowing X tell us anything about Y?
The standard hypothesis test:

$$H_0 : \beta_1 = 0 \quad v. \quad H_1 : \beta_1 \neq 0.$$

Test statistic:

$$t = \frac{b_1 - 0}{std.err(b_1)}.$$

[1]Approximate 95% CI for β_1:

$$\{b_1 - 2 \times std.err(b_1), b_1 + 2 \times std.err(b_1)\}.$$

[1] Technically, the 2's should be quantiles of the t-distribution on n -2 degrees of freedom, $t_{0.025,n-2}$.

Inference for b_0

The standard hypothesis test:

$$H_0 : \beta_0 = c \quad v. \quad H_1 : \beta_0 \neq c.$$

Test statistic:

$$t = \frac{b_0 - c}{std.err(b_0)}.$$

[2]Approximate 95% CI for β_0:

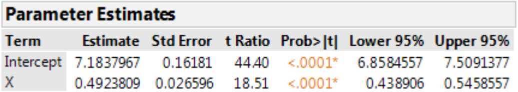
$$\{b_0 - 2 \times std.err(b_0), b_0 + 2 \times std.err(b_0)\}.$$

You don't need to know the standard error formula for b_0.

[2] Technically, the 2's should be quantiles of the t-distribution on n -2 degrees of freedom, $t_{0.025,n-2}$.

Inference output

From Utopia:



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	7.1837967	0.16181	44.40	<.0001*	6.8584557	7.5091377
X	0.4923809	0.026596	18.51	<.0001*	0.438906	0.5458557

- Be very careful: JMP by default tests null hypotheses of β_1 and β_0 equal to zero. This may or may not be what you want in any particular situation.
- Obtain the CI in JMP by right-clicking in the parameter estimates table and using the “columns dialog”.

Module summary

1. The SRM structure and assumptions
2. Utopia – the best situation possible
3. Checking the assumptions and the modeling process
4. Standard errors for the slope
5. Confidence intervals and test statistics

Next time

- Confidence and prediction intervals for the regression