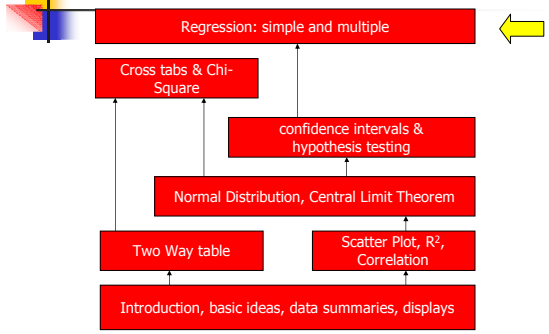
PUB – POS 316
Week 14a

Simple linear regression

Navid Ghaffarzadegan
navidg@gmail.com
Last updated – Jan 1, 10

Course Road Map



Agenda

- Introduction
  - Association
  - Scatter plots
- The linear regression model
- Tests for significance and CI
- ANOVA
- F-test

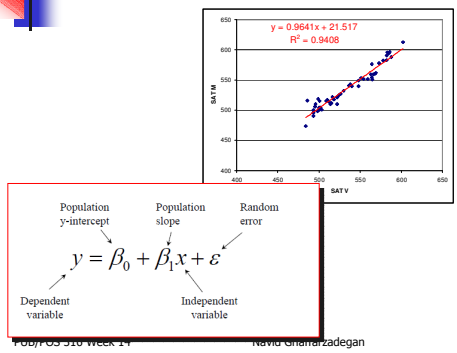
Introduction

- Review from last class
- Association between variables:

Two variables are associated if knowing the value of one of them tells you something about the other one.

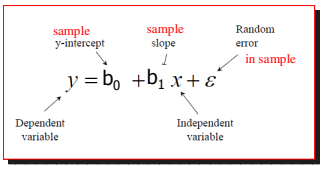
- Examples:
  - *Effort and grade*
    - *Positive association*
  - *Price and demand*
    - *Negative association*

Least Square Regression



Least Square Regression

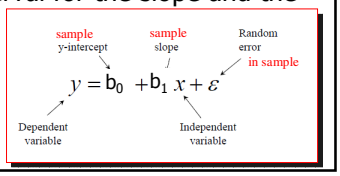
- What if we do not have the complete information about our population?



What does estimation of slope and intercept mean? (b estimation of β)

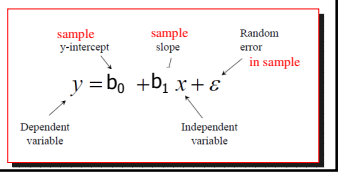
Tests for significance and CI

- So, if we are estimating the slope and the intercept of the line,…
- WE CAN BE WRONG
- → We need to report confidence intervals!
- Confidence interval for the slope and the intercept



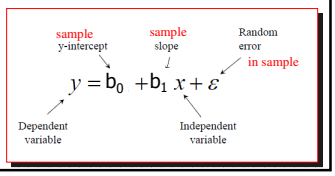
Tests for significance and CI

- Remember:
- Margin of error = z. (proper standard deviation)
- And if you do not have the st dev in population, you use t.
- The same here: (And the good thing is that excel gives you the proper standard deviation (standard error)
- Margin of error = $t^*_{0.025}$. (SE) (df=n-2)
- excel



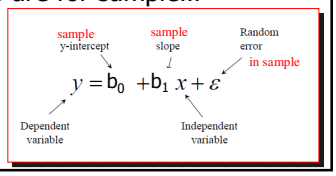
Tests for significance and CI

- What will happen for the slope and intercept if we conduct the study many times?
- The important question: Are you confident enough that the slope is not zero? ($\beta_1 \neq 0$)



Tests for significance and CI

- Hypotheses: H_0: $\beta_1 = 0$
- H_a: $\beta_1 \neq 0$
- Don't forget: β's are related to the population – b's are for sample…
- Very simple:
- $t = b_1 / SE_{b1}$



Tests for significance and CI

- So, 1. we should report confidence intervals for β's., or 2. We should test hypothesis that β's are different from zero.
- Back to excel.
- For your own work learning one of these two methods is enough.

Analysis of Variance (ANOVA)

- ANOVA:

Analysis **o**f **Va**riance

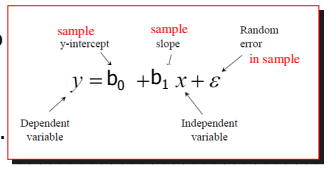
- As you have seen in this class, we are very interested to learned about variance (or standard deviation) in a data set. Remember?
- How can we explain why there is a variation in a data set?

Analysis of Variance (ANOVA)

- Example:
  - Why some students perform better?
  - Why some countries have a better health status?
  - What can explain variation in the divorce rate?
- Isn't that the whole purpose of social science?!!

Analysis of Variance (ANOVA)

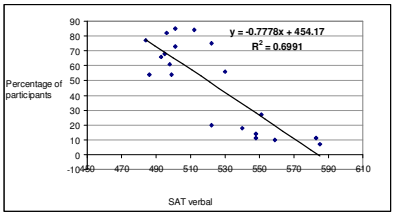
A regression analysis helps us to understand the reasons of the variance in our dependent variable.



- If you can show that $\beta_1 \neq 0$ (i.e., reject the null hypothesis), you are saying that some portion of variance in your dependent variable (y) is explained by your independent variable (x).

F-Test

- In general:
  - DATA = FIT + RESIDUAL



F-Test

- In general:
  - DATA = FIT + RESIDUAL

$$y_i = \hat{y}_i + (y_i - \hat{y}_i)$$

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$ Sum of Squares

SST = SSM + SSE

F-Test

- DATA = FIT + RESIDUAL

SST = SSM + SSE

R^2 = SSM / SST

MSE (Mean Square Error) = SSE/(n-2)

$$F = \frac{SSM / df_m}{SSE / df_e}$$

Analysis of Variance (ANOVA)

- What you need to remember:
  - F shows if your regression shows anything at all. (or it is just a random pattern between your x and y).
  - Excel reports F, compares it with F-table, reports p-value. Just we should be able to read it and know what it is about.

Analysis of Variance (ANOVA)

- Back to excel. Read F-test.

Summary

- What we need to know:
  - 1. When to conduct a regression.
  - 2. To use excel to conduct regression.
  - 3. To interpret the results.
  - 4. To know how to get t-value and test significance of coefficients and confidence intervals (if t or p or both are not given)

Example

- Example:
- We have data from a sample of computer science students. We would like to test to see if there is any association between their high school math grades and their SAT math.
- What should we do? How?

Summary

- What we need to know:
  - 1. When to conduct a regression.
  - 2. To use excel to conduct regression.
  - 3. To interpret the results.
  - 4. To know how to get t-value and test significance of coefficients and confidence intervals (if t or p or both are not given)

Example

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.382981264
R Square	0.146674648
Adjusted R Square	0.12851879
Standard Error	92.39064342
Observations	49

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	68959.52299	68959.52299	8.078640185	0.006606171
Residual	47	401193.4566	8536.030992		
Total	48	470152.9796			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	403.2044621	67.38424363	5.983660873	2.84931E-07	267.6448515	538.7640727
X Variable 1	22.72341076	7.99474077	2.84229488	0.006606171	6.640067123	38.80675439

Example

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.382981264
R Square	0.146674648
Adjusted R Square	0.12851879
Standard Error	92.39064342
Observations	49

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	68959.52299	68959.52299	8.078640185	0.006606171
Residual	47	401193.4566	8536.030992		
Total	48	470152.9796			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	403.2044621	67.38424363	5.983660873	2.84931E-07	267.6448515	538.7640727
X Variable 1	22.72341076	7.99474077	2.84229488	0.006606171	6.640067123	38.80675439

Check if the slope is significantly different from zero..
That's **the most important** thing

Summary

- What we need to know:
  - 1. When to conduct regression.
  - 2. To use excel to conduct regression.
  - 3. To interpret the results.
  - 4. To know how to get t-value and test significance of coefficients and confidence intervals (if t or p or both are not given)

Example

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.382981264
R Square	0.146674648
Adjusted R Square	0.12851879
Standard Error	92.39064342
Observations	49

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	68959.52299	68959.52299		
Residual	47	401193.4566	8536.030992		
Total	48	470152.9796			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	403.2044621	67.38424363				
X Variable 1	22.72341076	7.99474077				

Example

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.382981264
R Square	0.146674648
Adjusted R Square	0.12851879
Standard Error	92.39064342
Observations	49

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	68959.52299	68959.52299		
Residual	47	401193.4566	8536.030992		
Total	48	470152.9796			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	403.2044621	67.38424363				
X Variable 1	22.72341076	7.99474077				

$F=(SSM/df_M)/(SSE/df_E)$

table

t=Coefficient/Standard Error

Find it from t-table, using t and df (n-2)

=coefficient ± margin of error (margin of error=$t_{0.025}$*SE)

Example

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.382981264
R Square	0.146674648
Adjusted R Square	0.12851879
Standard Error	92.39064342
Observations	49

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	68959.52299	68959.52299	8.078640185	0.006606171
Residual	47	401193.4566	8536.030992		
Total	48	470152.9796			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	403.2044621	67.38424363	5.983660873	2.84931E-07	267.6448515	538.7640727
X Variable 1	22.72341076	7.99474077	2.84229488	0.006606171	6.640067123	38.80675439

Summary

- What we need to know:
  - 1. When to conduct a regression.
  - 2. To use excel to conduct regression.
  - 3. To interpret the results.
  - 4. To know how to get t-value and test significance of coefficients and confidence intervals (if t or p or both are not given)
- That's almost every thing that we need to know about regression!..