ACE 261
Fall 2002
Prof. Katchova

Lecture 14

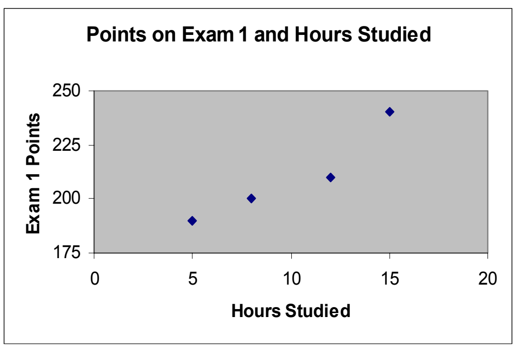
Simple Linear Regression

Question

- What is the relationship between the time a student studied for the exam and his/her points on exam 1?

Hours studied	Points on exam 1
5	190
8	200
12	210
15	240

2



3

Answer

- We already know from chapters 2 and 3:
- We can look at the scatter diagram and the correlation coefficient.
- Conclusion: correlation coefficient = 95%, therefore there is _____________ relationship.

4

New questions

- Can I predict if a student studied 5 hours, how much he/she would score?
- For each hour of study, a student's grade increases by how many points?
- If the increase is 4 points for each hour studied, is that a significant increase?

5

Definitions

- Dependent variable is the variable being predicted or explained. Usually denoted by y.
  - Example: y=exam points
- Independent variables are the variables being used to predict or explain the dependent variable. Usually denoted by x_1, x_2, etc.
  - Example: x=hours studied
- Regression analysis (model) is used to predict the value of the dependent variable based on the values of the independent variables.
  - Given that someone studied 5 hours, how much would he/she score?

6

More definitions

- Be careful! Regression analysis does not establish a cause-and-effect relationship, just that there is a relationship.
  - For example: students who study more have higher score, but it is also true that students who are taller also have higher score.
  - The cause-and-effect must be established with a theoretical or logical reason.
- Simple linear regression model is a regression model where the relationship between the dependent and *one* independent variable is approximated by *a straight line*.
- Multiple linear regression model is a regression model where the relationship between the dependent and *two or more* independent variables is approximated by *a straight line*.

7

A potential answer

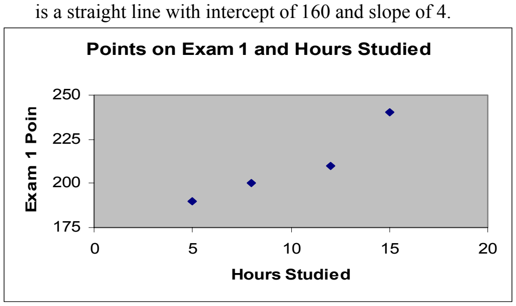
- Suppose I tell you that if a student studies x hours, he will score y points based on the following equation.
- Predicted points = 160 + 4*(Hours studied)

Hours studied	Observed points	Predicted points	
5	190		
8	200		
12	210		
15	240		

8

Graphical representation

- The equation Predicted points = 160 + 4*(Hours studied) is a straight line with intercept of 160 and slope of 4.



9

Is this a good model?

- Does this model fit perfectly? No, it has an error.
- Error = observed points – predicted points or
- Observed points = 160 + 4*(hours studied) + error
- We'll have a good model if the errors are as small as possible. Can (how do) we make the errors as small as possible?
  - Least squares method: minimize sum (observed points – predicted points)^2

10

Regression model and estimated equation

Simple linear regression model: an equation that describes how the dependent variable y is related to the independent variable x and an error.

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

Observed points = $\beta_0 + \beta_1$ (hours studied) + error

- Unfortunately, we don't know β_0 and β_1 but we can estimate them by using sample data, so we get b_0 and b_1.

Estimated simple linear equation:

yhat = $b_0 + b_1$ x

Predicted points = 160 + 4*(Hours studied)

11

The least squares method

- The least squares method uses sample data to find the estimated regression equation.
- It provides values of b_0 and b_1 that minimize the sum of squared errors (SSE).
- Min SSE $= \Sigma (y - \text{yhat})^2$ or

  Minimize sum (observed points – predicted points)^2

12

Least squares estimates

- The slope of the regression line is
- b_1 = cov(x,y)/var(x) = Σ (x xbar)(y ybar)/Σ(x xbar)^2
- The intercept of the regression line is
- b_0 = ybar – b_1 xbar

- Find b_1 and b_0 using the least squares method.

13

Calculations

Hours studied (x_i)	Points on exam 1 (y_i)	x_i-x	y_i-y	(x_i-x)* (y_i-y)	(x_i-x)^2
5	190				
8	200				
12	210				
15	240				

14

Least squares estimates

- b_1 = Σ (x-xbar)(y-ybar)/Σ(x-xbar)^2=

- b_0 = ybar – b_1 xbar =

- The estimated regression equations is:
- yhat = 163.45 + 4.66 x
- Interpretation: we predict that a student who studied 0 hours will score 163.45, and 4.66 points more for each additional hour of study.
- Given x=5, the predicted score = yhat =

15

Determining goodness of fit

- How well does the model fit the data?
- SST=SSR+SSE
- sum of squares total = sum of squares regression + sum of squares error.
  Σ (y ybar)^2 =Σ (yhat ybar)^2 +Σ (y yhat)^2
- Total variation = explained variation by the regression + unexplained variation associated with error

16

Coefficient of determination

- <u>Coefficient of determination</u> (R^2) provides a measure of the goodness of fit for the estimated regression equation.
- R^2 = SSR/SST = 1 – SSE/SST
- Values of R^2 close to 1 indicate perfect fit, values close to zero indicate poor fit. R^2 of more than 0.25 is considered good in the ag economics field.

- If SSE =143.1034 and SST =1400, R^2=
- This means that 89.78% of the variation is explained by the regression and the rest of the variation is due to error.

17

Correlation coefficient

- r_{xy} = (sign of b_1) *sqrt (R^2) =
- Goodness of fit is a better measure than the correlation coefficient because it can be applied when:
  - there are more independent variables
  - the relationship between the dependent and independent variables is not linear.

18

Testing for significance

- Is the increase of b_1=4.66 points for each hour studied, significant or not?
- In other words, is the slope β_1= zero? If the slope is zero then y and x are not related (y does not depend on x).
- H_0: $\beta_1 = 0$ and H_a: $\beta_1 \neq 0$
- Two tests
  - t-test for a coefficient significance (β_1=0 or not)
  - F-test for an overall significance (are y and x related? Are all coefficients jointly equal to zero?)
  - If one independent variable, these two tests have the same results, with more independent variables, the tests have different results.

19

F-test for overall significance of all coefficients

- H_0: $\beta_1 = 0$ and H_a: $\beta_1 \neq 0$
- SST=SSR+SSE
  Total variation = explained variation by the regression + unexplained variation associated with error
- F = (SSR/p)/[SSE/(n-p-1)] = MSR/MSE
  p is the number of independent variables, n is the number of observations
- If the error explains a lot of the variation in score and the regression doesn't, then the regression is not significant, i.e., $\beta_1 = 0$

20

ANOVA table

Source	Sum of Squares	Degrees of Freedom	Mean Square	F
Regression	SSR= $\sum(\text{yhat-ybar})^2$	p = number of independent variables	MSR=SSR/p	F=MSR/MSE
Error	SSE= $\sum(\text{y-yhat})^2$	n-p-1	MSE= SSE/(n-p-1)	
Total	SST= $\sum(\text{y-ybar})^2$	n-1		

- SST=SSR+SSE
- Total variation = explained variation by the regression + unexplained variation associated with error

21

ANOVA table

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1256.897	1256.897	17.56627	0.052486
Residual (Error)	2	143.1034	71.55172		
Total	3	1400			

- Since p-value=0.0525 > 0.05, the relationship between hours studied and score is not significant.

22

t-test for a coefficient significance

- H_0: $\beta_1 = 0$ and H_a: $\beta_1 \neq 0$
- $t = (b_1-\beta_1) / s_{b1}$ with d.f.= n-p-1
- s_{b1} = sqrt(MSE)/ sqrt [Σ(x-xbar)^2]
- Since the p-value for time studied is 0.0525 > 0.05 we accept the null hypothesis, there is no relation between time studied and points scored.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	163.4483	11.88499	13.7525	0.005246
Time studied	4.655172	1.110698	4.191213	0.052486

23

Confidence interval for β_1

- The confidence interval for β_1 is $b_1 \pm t_{\alpha/2}\, s_{b1}$ =
- Interpretation: I'm 95% confident that the value for β_1 is between -0.12378 and 9.434124.

	Lower 95.0%	*Upper 95.0%*
Intercept	112.3113	214.5853
Time studied	-0.12378	9.434124

24

Model Assumptions

- The error e is a random variable with mean of zero.
- The variance of e , denoted by σ^2, is the same for all values of the independent variable.
- The values of e are independent.
- The error e is a normally distributed random variable.

25

Detecting outliers

- An outlier is an observation that has unusually large or small values. Solutions?
  - Maybe a mistake was made – correct it
  - Maybe the model doesn't fit well
  - May just happened by chance?

26

Detecting influential observations

- Influential observation is an observation with extreme values for the independent variable. An influential observation has a high leverage.
- The leverage is determined by how far the values of the independent variables are from their means.
- Solutions?
  - Run the regression without the influential observation and see if the results change.

27

Regression of price of stock on number of stocks sold

Regression Statistics	
Multiple R	0.862428
R Square	0.743781
Adjusted R Square	0.711754
Standard Error	1.419338
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	46.78384	46.78384	23.22333	0.001323
Residual	8	16.11616	2.01452		
Total	9	62.9			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	9.264947	1.099136	8.429297	2.99E-05	6.730332	11.79956
Shares	0.710515	0.147438	4.819059	0.001323	0.370521	1.050509

28