

Soci708 – Statistics for Sociologists

Module 10 – Inference for Regression¹

François Nielsen

University of North Carolina
Chapel Hill

Fall 2009

¹Adapted from slides for the course Quantitative Methods in Sociology (Sociology 6Z3) taught at McMaster University by Robert Andersen (now at University of Toronto)

Goals of This Module

- ▶ Review of *least-squares regression* analysis
 - ▶ Simple and multiple regression
 - ▶ Slope, intercept and R^2
 - ▶ Standard error of the regression
- ▶ Inference for Regression
 - ▶ Confidence intervals and hypothesis tests for the slope
 - ▶ F-test for the entire regression model
- ▶ Assumptions of regression and how to check them

Review of Regression Analysis (1)

- ▶ Simple least squares regression fits a straight line to the relationship between *two linearly-related quantitative variables*
- ▶ The fitted *simple regression* line is represented by the equation:

$$\hat{y} = A + Bx$$

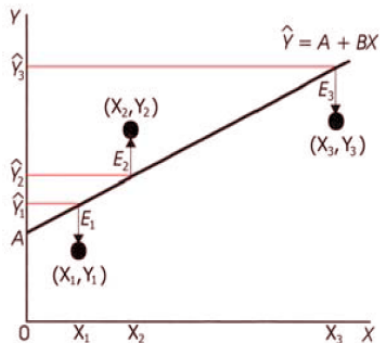
- ▶ A represents the *y-intercept* (the value of y for $x = 0$)
- ▶ The *slope of the line*, B , indicates how much y changes on average when x is increased by 1
 - ▶ If B is positive, y increases as x increases;
 - ▶ if B is negative, y decreases as x increases;
 - ▶ if $B = 0$, then the line is horizontal and thus remains constant as x changes

Review of Regression Analysis (2)

- ▶ Least-squares regression fits a straight line to linearly related data by minimizing the *residuals* (vertical distances between the predicted values and the observed values):

$$\begin{aligned} Y_i &= A + BX_i + E_i \\ &= \hat{Y}_i + E_i \\ E_i &= Y_i - \hat{Y}_i \end{aligned}$$

where $\hat{Y}_i = A + BX_i$ is the *fitted value* of Y_i



- ▶ Least squares regression chooses the values of A and B that minimize the sum of the *squared* residuals $\sum E_i^2$ or equivalently $\sum (Y_i - A - BX_i)^2$

Review of Regression Analysis (2b)

OPTIONAL – How to find the values of A and B that minimize $\sum E_i^2$?

- ▶ The sum of squared residuals can be viewed as a function Q of two *unknowns* A and B:

$$Q = \sum E_i^2 = \sum (Y_i - A - BX_i)^2$$

- ▶ To find the values A and B that minimize Q one could use:
 1. A “brute force” numerical search using a grid of values for A and B, or
 2. The calculus-based solution discovered by French mathematician Adrien-Marie Legendre in 1805 (next)
- ▶ From calculus, the values of A and B that minimize Q are found by equating the partial derivatives of Q to zero:

$$\frac{\delta Q}{\delta A} = -2 \sum (Y_i - A - BX_i) = 0 \quad (1)$$

$$\frac{\delta Q}{\delta B} = -2 \sum X_i(Y_i - A - BX_i) = 0 \quad (2)$$

Review of Regression Analysis (2c)

OPTIONAL – How to find the values of A and B that minimize $\sum E_i^2$? (cont'd)

- ▶ Which can be rearranged into the *normal equations*

$$\sum Y_i = nA + B \sum X_i \quad (3)$$

$$\sum X_i Y_i = A \sum X_i + B \sum X_i^2 \quad (4)$$

- ▶ Which can in turn be rearranged into the formulae

$$B = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \quad (5)$$

$$A = \bar{Y} - B\bar{X} \quad (6)$$

- ▶ These formulae for B and A can be shown to be equivalent to the ones shown on the next slide

Review of Regression Analysis (3)

- ▶ The *slope* B of the regression line is calculated as follows:

$$B = r \frac{s_y}{s_x}$$

Where r is the correlation between y and x ; s_y is the standard deviation of y ; and s_x is the standard deviation of x

- ▶ The *intercept* A is calculated as follows:

$$A = \bar{y} - B\bar{x}$$

where \bar{x} and \bar{y} are the means of the variables

Review of Regression Analysis (4)

- ▶ We can assess the fit of the regression line using the r^2 and the *standard error about the regression line* S_E
- ▶ *Squared correlation coefficient*, r^2
 - ▶ Varies between 0 (no linear relationship) and 1 (perfect linear relationship)
 - ▶ Represents the proportion of variation in y that is accounted for by its regression on x
- ▶ The *standard deviation of the residuals* (or *standard error of the regression*, S_E)
 - ▶ Measured in the *units of the dependent variable*
 - ▶ Represents the “typical” residual (standard error in this case has nothing to do with the sampling distribution):

$$S_E = \sqrt{\frac{\sum E_i^2}{n - 2}}$$

An Example of Simple Regression (1)

- ▶ Below we regress income (y) on education (x), resulting in the following:

$$\hat{y} = -7677 + 2419x$$
$$\widehat{\text{income}} = -7677 + 2419 \times \text{education}$$

- ▶ The *slope* tells us that as education goes up by one year, on average income increases by 2419 dollars

obs	x	y	\hat{y}	residual, E_i
1	12	20,000	21,354.8	-1,354.8
2	13	22,000	23,774.2	-1,774.2
3	12	23,000	21,354.8	1,645.2
4	14	25,000	26,193.5	-1,193.5
5	12	18,000	21,354.8	-3,354.8
6	15	30,000	28,612.9	1,387.1
7	12	26,000	21,354.8	4,645.2

An Example of Simple Regression (2)

- ▶ We can now calculate the *standard error about the regression line* (standard deviation of the residuals):

$$\begin{aligned}S_E &= \sqrt{\frac{\sum E_i^2}{n - 2}} \\ &= \sqrt{\frac{43,870,968}{5}} \\ &= 2962.1\end{aligned}$$

- ▶ We see here that the typical mistake in predicting income from this regression line is about 2962 dollars
- ▶ We have $n - 2$ degrees of freedom as a consequence of estimating the 2 parameters α and β with A and B

Inference for Regression

- ▶ So far we can only talk about our sample
 - ▶ That is, we have not discussed statistical inference for regression
- ▶ Usually we are interested in the *slope in the population*, β (beta) which we estimate using the sample slope B
- ▶ The now familiar methods for finding confidence intervals and hypothesis tests can be adapted to regression analysis
 - ▶ As with inference for means and proportions we must first estimate the *standard error of the least-squares slope*, B , in order to make inferences about β

Inference for Regression (2): Assumptions

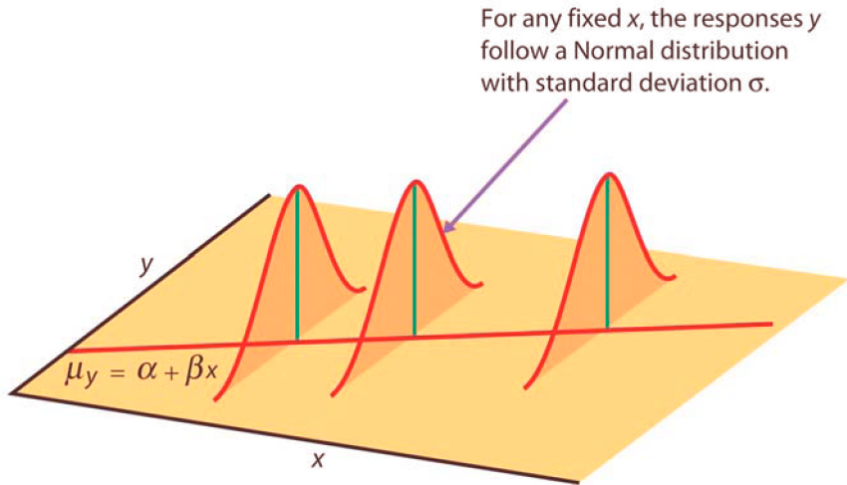
1. For any fixed value of x , the response variable y varies according to a normal distribution
2. The mean response μ_y has a linear relationship with x :

$$\mu_y = \alpha + \beta x$$

where α is the true population intercept and β is the true population slope

3. The standard deviation of y in the population (σ) is the same for all values of x (in other words, y has constant spread through the range of x)

Inference for Regression (3)



Confidence Intervals for the Slope, β (1)

- ▶ Inference for β uses the t distribution with $n - 2$ degrees of freedom
- ▶ The method is very similar to that used for inference about a population mean μ :

$$\text{estimate} \pm t^* SE_{\text{estimate}}$$
$$B \pm t^* SE_B$$

- ▶ where t^* is for a level C confidence interval with upper $(1 - C)/2$ critical value from the t distribution with $n - 2$ degrees of freedom
 - ▶ Unless n is very small, when $C = 95$ percent, the value of t^* (with probability .025 to the right) is approximately 2

Confidence Intervals for the Slope, β (2)

1. The *standard error of the least squares slope B* is:

$$SE_B = \frac{s}{\sqrt{\sum (x - \bar{x})^2}} = \frac{S_E}{\sqrt{\sum (x - \bar{x})^2}}$$

So, for the regression of income on education:

$$s = 2962.1 \text{ and } \sqrt{\sum (x - \bar{x})^2} = 8.86$$

Therefore:

$$SE_B = \frac{2962.1}{\sqrt{8.86}}$$

The degrees of freedom are $n - 2 = 7 - 2 = 5$

- ▶ From the *t*-table (Table D in Moore & McCabe 2006) we see that the critical value of *t* with 5 df for a 95% CI is 2.571

Confidence Intervals for the Slope, β (3)

2. We now have the following information:

$$B = 2419; \quad SE_B = 995.1; \quad t^* = 2.571$$

- ▶ The 95% confidence interval for β is then:

$$\begin{aligned} B \pm t^* SE_B &= 2419 \pm 2.571 \times 995.1 \\ &= 2419 \pm 2558.4 \\ &= -139.4 \text{ to } 4977.4 \end{aligned}$$

3. As usual, we can see this as equivalent to a non-directional (two-tail) hypothesis test at the $\alpha = 0.05$ level for $H_0 : \beta = 0$
- ▶ Since the interval *does* contain 0, we fail to reject the null hypothesis that education has no effect on income

Hypothesis Test for the Slope β (1)

- ▶ Significance tests for the population slope β are also similar to significance tests for the population mean μ
 1. As usual, we start with the null hypothesis. Usually we are interested in the null hypothesis that there is no relationship between x and y :

$$H_0 : \beta = 0$$

- ▶ H_0 specifies that *the regression line in the population is horizontal* (i.e., there is no correlation between the two variables)
- 2. The alternative hypothesis as usual can be one-sided if we expect a positive or a negative relationship in advance:

$$H_a : \beta > 0 \text{ or } H_a : \beta < 0$$

Or more commonly in regression, two-sided where we don't have a specific direction in mind:

$$H_a : \beta \neq 0$$

Hypothesis Test for the Slope β (2)

3. The t statistic for testing the null hypothesis is:

$$t = \frac{B}{SE_B} \text{ with } n - 2 \text{ degrees of freedom}$$

- ▶ For our example of income regressed on education, we specify a null hypothesis that there is no relationship between income and education:

$$H_0 : \beta = 0$$

- ▶ Since we expect a positive relationship, the alternative hypothesis is directional:

$$H_a : \beta > 0$$

- ▶ The t -statistic is:

$$t = \frac{B}{SE_B} = \frac{2419}{995.1} = 2.43$$

Hypothesis Test for the Slope β (3)

- ▶ We obtain the P-value for this right-tail test as between .025 and .05 (from Table D in Moore & McCabe 2006) or exactly .0297 (from software)
 - ▶ Since the P-value .0297 is less than $\alpha = .05$ we *can reject* the null hypothesis at the $\alpha = .05$ level
- ▶ Alternatively, we find that the critical value of t with 5 df for right tail P-value for $\alpha = .05$ is 2.015
 - ▶ Since the t -statistic (2.43) is higher than this value, again we *can reject* the null hypothesis at the $\alpha = .05$ level
- ▶ For a *two-sided* test the null become $H_0 : \beta \neq 0$. The corresponding two-sided P-value is calculated as $2 \times P(T > |t|) = 2 \times .0297 = .0593$
 - ▶ Since the two-sided P-value .0593 is greater than $\alpha = .05$, we *fail to reject* the null hypothesis at the $\alpha = .05$ level
 - ▶ We conclude that β is not significantly different from zero at the $\alpha = .05$ level in this two-sided test

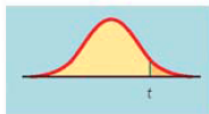
SIGNIFICANCE TEST FOR REGRESSION SLOPE

To test the hypothesis $H_0: \beta = 0$, compute the t statistic

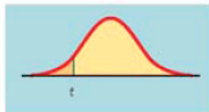
$$t = \frac{b}{SE_b}$$

In terms of a random variable T having the $t(n - 2)$ distribution, the P -value for a test of H_0 against

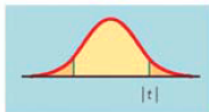
$$H_a: \beta > 0 \text{ is } P(T \geq t)$$



$$H_a: \beta < 0 \text{ is } P(T \leq t)$$



$$H_a: \beta \neq 0 \text{ is } 2P(T \geq |t|)$$



Simple Regression in R

Typical regression output showing the *two-sided* test of $H_0 : \beta = 0$

```
> # in R
> x<-c(12,13,12,14,12,15,12)
> y<-c(20000,22000,23000,25000,18000,30000,26000)
> # Creating variables "x" is education, "y" is income
> reg.model<-lm(y~x)
> summary(reg.model)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-7677.4	12845.7	-0.598	0.5761
x	2419.4	995.3	2.431	0.0593 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2962 on 5 degrees of freedom

Multiple R-Squared: 0.5416, Adjusted R-squared: 0.45

F-statistic: 5.909 on 1 and 5 DF, p-value: 0.05932