

Unit 16: Regression

Regression analysis

- Regression Analysis: Fitting the “best” line through the scatter
- Very closely related to correlation, but treats one variable as **dependent** and the other(s) as **explanatory**, while correlation is asymmetric

Some geometry: equation of a line

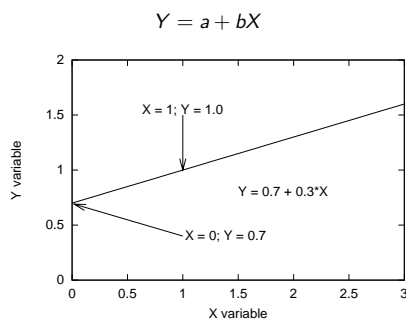


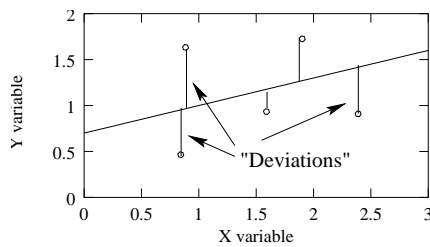
Figure 9: The equation of a line.

Predictive in intent

- Asymmetric: use X to predict Y
- PRE: implied causality (Proportional Reduction in Error: if we know X, we can guess Y better)
- Find the best a and b to summarise the data scatter:
 - 'Best' is defined as minimising the squared deviations between the observed data-points and the fitted line, hence often called 'least-squares' regression
 - Deviations are the vertical distance between the line and the observed data points.
 - Very similar logic to the mean (minimise variance).

Deviations again

Figure 10: Deviations from the line



Predicted values

- The line gives a **predicted** value of Y for each value of X:

$$\hat{Y} = a + bX$$

$$Y = \hat{Y} + e$$

$$Y = a + bX + e$$

e is the 'residual' or deviation.

- That is, knowing X we “predict” or guess Y as $a + bX$
- In general this is more accurate than guessing Y as \bar{Y} , the mean: “Proportionate Reduction in Error”

Regression equation

- Regression equation: the estimate of Y, called \hat{Y} , depends on X:

$$\hat{Y} = a + bX$$

- The regression slope b depends on SXY and SXX, the intercept a is calculated from b and the mean values of Y and X:

$$b = \frac{SXY}{SXX}$$

$$a = \bar{Y} - b\bar{X}$$

$$SXX = \sum (X_i - \bar{X})^2$$

$$SXY = \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

Pitfalls

- Like correlation, non-linear relationships may be missed
- Spurious relationships will fit just as well as real ones (e.g., if A affects B and A affects C, B and C will seem to be related and a regression line might fit well)
- Predicting outside the range of the data: the relationship we see only holds for the data we use, and it may well not hold for higher (or lower) values of X and Y

Fit

- How well does it "fit"? We use R^2 to tell:
 - ranges from 0: no relationship at all
 - to 1: perfect relationship, all Y 's are exactly equal to $a + bX$
 - values from 0.7 up indicate quite a good relationship
 - smaller values may indicate an interesting relationship
- In the case of bivariate regression (one independent variable), R^2 is the same as $r \times r$ (squared correlation coefficient).

266

Regression in Stata:

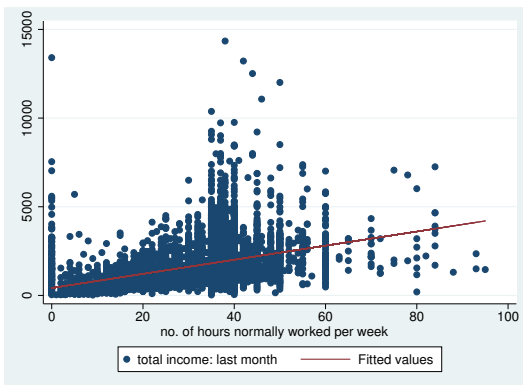
```
. reg ofimm ojbhrs
```

Source	SS	df	MS			
Model	1.7116e+09	1	1.7116e+09	Number of obs =	7894	
Residual	9.6048e+09	7892	1217033.86	F(1, 7892) =	1406.39	
Total	1.1316e+10	7893	1433733.83	Prob > F =	0.0000	
				R-squared =	0.1513	
				Adj R-squared =	0.1511	
				Root MSE =	1103.2	

	ofimm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
ojbhrs		39.77896	1.060718	37.50	0.000	37.69967 41.85825
_cons		420.2139	37.02457	11.35	0.000	347.6359 492.7919

267

Predicted regression line



268

Multiple explanatory variables

- Regression analysis can be extended to the case where there is more than one explanatory variable – multivariate regression
- This allows us to estimate the net simultaneous effect of many variables, and thus to begin to disentangle more complex relationships
- Interpretation is relatively easy: each variable gets its own slope coefficient, standard error and significance
- The slope coefficient is the effect on the dependent variable of a 1 unit change in the explanatory variable, *while taking account of the other variables*

269

Example

- Example: domestic work time may be affected by gender, and also by paid work time: competing explanations – one or the other, or both could have effects
- We can fit bivariate regressions:

$$DWT = a + b \times PaidWork$$

or

$$DWT = a + b \times Female$$

- We can also fit a single multivariate regression

$$DWT = a + b \times PaidWork + c \times Female$$

270

Dichotomous variables

- We deal with gender in a special way: this is a *binary* or *dichotomous* variable – has two values
- We turn it into a yes/no or 0/1 variable – e.g., female or not
- If we put this in as an explanatory variable a *one unit change in the explanatory variable* is the difference between being male and female
- Thus the c coefficient we get in the $DWT = a + b \times PaidWork + c \times Female$ regression is the net change in predicted domestic work time for females, once you take account of paid work time.
- The b coefficient is then the net effect of a unit change in paid work time, once you take gender into account.

271

SO5041, Autumn 2016

Unit 17: Review

272

- Retrospective Course Outline:
 - Representing information as numbers
 - Probability distributions and CIs
 - Hypothesis testing
 - Questionnaires and sampling
 - Correlation and regression

273

- Representing information as numbers
 - levels of measurement
 - univariate descriptive statistics and graphics
 - bivariate descriptive statistics and graphics
 - Agresti/Finlay ch 2.1, ch 3, assignment 1
- Probability distributions and CIs
 - The normal distribution
 - The t-distribution
 - The χ^2 (chi-sq) distribution
 - Confidence intervals
 - Agresti/Finlay chs 4, 5, assignment 2

- Hypothesis testing
 - Extension of CI framework
 - Test the "null hypothesis", if rejected \Rightarrow support for initial hypothesis
 - Agresti/Finlay ch 6
- Questionnaires and sampling
 - Questionnaire design
 - The principles of sampling
 - Different sampling strategies
 - Sampling theory: Agresti/Finlay ch 2.2–2.4
 - Methodological overview: de Vaus, chs 6–8

- Correlation and regression
 - Correlation as bivariate summary of interval/ratio variables
 - Regression: $\hat{Y} = a + bX$; directional
 - Agresti ch 9.1–9.4