Poli 5D Social Science Data Analytics

Regression in Stata

Shane Xinyang Xuan
ShaneXuan.com

February 10, 2017

Contact Information

Shane Xinyang Xuan

xxuan@ucsd.edu

The teaching staff is a team!

Professor Roberts	M	1600-1800 (SSB 299)
Jason Bigenho	Th	1000-1200 (Econ 116)
Shane Xuan	M	1100-1150 (SSB 332)
	Th	1200-1250 (SSB 332)

Supplemental Materials

UCLA STATA starter kit

http://www.ats.ucla.edu/stat/stata/sk/

Princeton data analysis

http://dss.princeton.edu/training/

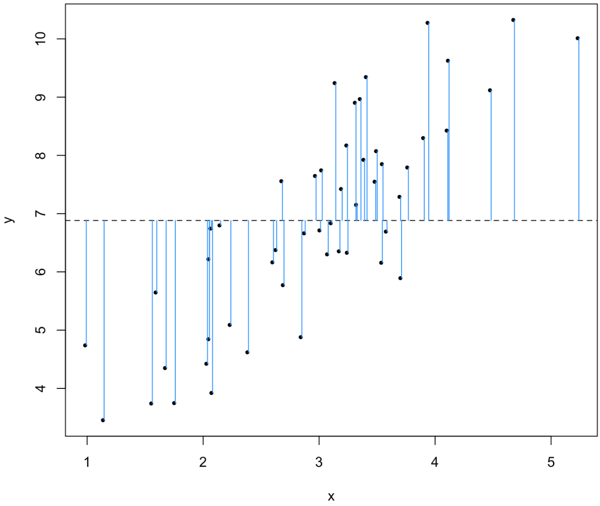
Road map

Some quick notes before we start today's section:

- Make sure that you pass around the attendance sheet
- Open a .do file
- Import your data ("h1_fams_data.xlsx")
- I will be using my slides, and you will need to type the code in your .do file

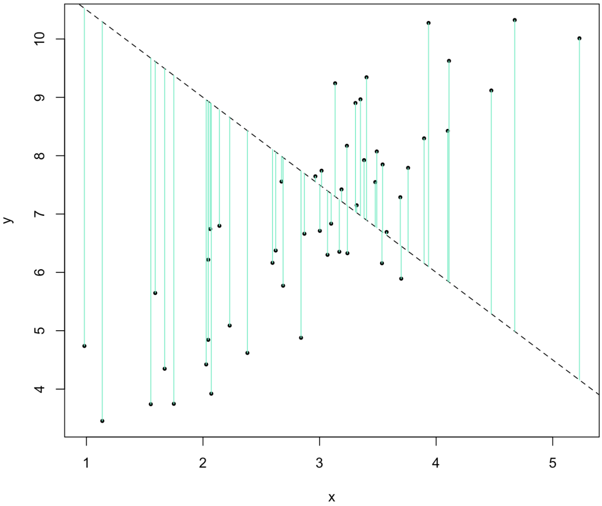
Regression: Examples!

Figure: Data points



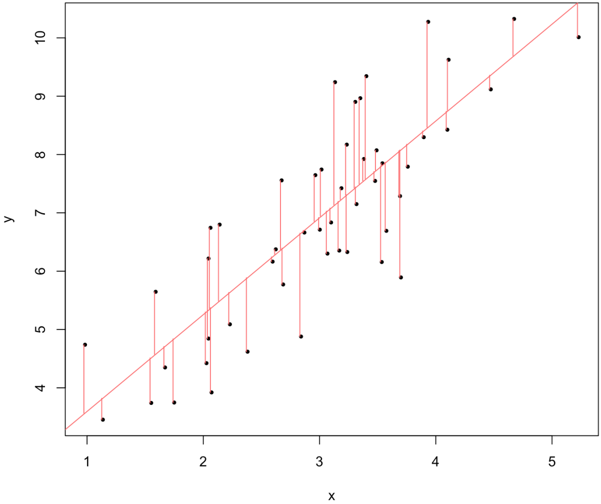
Regression: Examples!

Figure: Bad fit



Regression: Examples!

Figure: Good fit



Model

- Population

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

Model

- Population

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

- Estimation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{e}_i$$

Model

- Population

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

- Estimation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{e}_i$$

- (You don't need to memorize this) Regression Coefficient is calculated by

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}$$

Interpretation of regression coefficient

Suppose we have the model

$$y = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_0 + \hat{e}$$

Interpretation of regression coefficient

Suppose we have the model

$$y = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_0 + \hat{e}$$

- A 1-unit change in x_1 is associated with a β_1-unit change in y, all else equal.

Interpretation of regression coefficient

Suppose we have the model

$$y = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_0 + \hat{e}$$

- A 1-unit change in x_1 is associated with a β_1-unit change in y, all else equal.
- A 1-unit change in x_2 is associated with a β_2-unit change in y, all else equal.

Application

- Suppose consumption ($cons$) is a function of family income (inc):

$$cons = \beta_0 + \beta_1 inc + u$$

where u contains other factors affecting consumption. What change do you expect to see in $cons$ with a two-unit increase in inc?

Application

- Suppose consumption ($cons$) is a function of family income (inc):

$$cons = \beta_0 + \beta_1 inc + u$$

  where u contains other factors affecting consumption. What change do you expect to see in $cons$ with a two-unit increase in inc?
- With a two-unit increase in inc,

Application

- Suppose consumption ($cons$) is a function of family income (inc):

$$cons = \beta_0 + \beta_1 inc + u$$

  where u contains other factors affecting consumption. What change do you expect to see in $cons$ with a two-unit increase in inc?

- With a two-unit increase in inc,

$$\begin{aligned} cons &= \beta_0 + \beta_1(inc + 2) + u \\ &= \beta_0 + (\beta_1 inc + 2\beta_1) + u \\ &= (\beta_0 + \beta_1 inc + u) + 2\beta_1 \end{aligned}$$

Application

- Suppose consumption ($cons$) is a function of family income (inc):

$$cons = \beta_0 + \beta_1 inc + u$$

  where u contains other factors affecting consumption. What change do you expect to see in $cons$ with a two-unit increase in inc?

- With a two-unit increase in inc,

$$\begin{aligned} cons &= \beta_0 + \beta_1(inc + 2) + u \\ &= \beta_0 + (\beta_1 inc + 2\beta_1) + u \\ &= (\beta_0 + \beta_1 inc + u) + 2\beta_1 \end{aligned}$$

  Thus, we see a $2\beta_1$ increase in $cons$ with a 2-unit increase in inc!

Code

- Scatter plot: twoway (scatter povertyratio_mom age_mom, mlabsize(tiny) msize(tiny))

Code

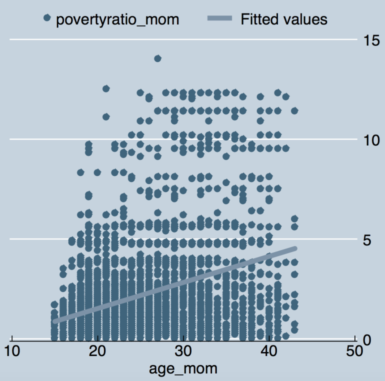
- Scatter plot: twoway (scatter povertyratio_mom age_mom, mlabsize(tiny) msize(tiny))
- Regression: regress povertyratio_mom age_mom

Code

- Scatter plot: twoway (scatter povertyratio_mom age_mom, mlabsize(tiny) msize(tiny))
- Regression: regress povertyratio_mom age_mom
- Visualization: twoway (scatter povertyratio_mom age_mom, mlabsize(tiny) msize(tiny)) (lfit povertyratio_mom age_mom)

Code

- Scatter plot: twoway (scatter povertyratio_mom age_mom, mlabsize(tiny) msize(tiny))
- Regression: regress povertyratio_mom age_mom
- Visualization: twoway (scatter povertyratio_mom age_mom, mlabsize(tiny) msize(tiny)) (lfit povertyratio_mom age_mom)



Residuals

- Fitted values

Residuals

- Fitted values
  - Manually: gen fitted = -1.091357 + .1305531 * age_mom

Residuals

- Fitted values
  - Manually: gen fitted = -1.091357 + .1305531 * age_mom
  - Stata command: predict fv

Residuals

- Fitted values
  - Manually: gen fitted = -1.091357 + .1305531 * age_mom
  - Stata command: predict fv
- Residuals

Residuals

- Fitted values
  - Manually: gen fitted = -1.091357 + .1305531 * age_mom
  - Stata command: predict fv
- Residuals
  - Manually: gen resid = povertyratio_mom - fv

Residuals

- Fitted values
  - Manually: gen fitted = -1.091357 + .1305531 * age_mom
  - Stata command: predict fv
- Residuals
  - Manually: gen resid = povertyratio_mom - fv
  - Stata command: predict e, residual

Residuals

- Fitted values
  - Manually: gen fitted = -1.091357 + .1305531 * age_mom
  - Stata command: predict fv
- Residuals
  - Manually: gen resid = povertyratio_mom - fv
  - Stata command: predict e, residual

fitted	fv	resid	e
.9974926	.9974928	-.6974928	-.6974928
1.519705	1.519705	-.4197054	-.4197053
1.258599	1.258599	.5414009	.5414008
1.911364	1.911365	-1.711365	-1.711365
2.955789	2.95579	-2.75579	-2.75579

Figure: Similar results for fitted values, and residuals

What else can you do using regressions?

- Suppose you run a regression of y on x_1, and get an error term $\hat{e}$. You can then do a scatterplot of error term ($\hat{e}$) and a different variable (x_2) to see how much of the difference can be explained by this variable:

What else can you do using regressions?

- Suppose you run a regression of y on x_1, and get an error term $\hat{e}$. You can then do a scatterplot of error term ($\hat{e}$) and a different variable (x_2) to see how much of the difference can be explained by this variable:
  - twoway scatter e x_2

What else can you do using regressions?

- Suppose you run a regression of y on x_1, and get an error term $\hat{e}$. You can then do a scatterplot of error term ($\hat{e}$) and a different variable (x_2) to see how much of the difference can be explained by this variable:
  - twoway scatter e x_2
- You can do a multiple regression

What else can you do using regressions?

- Suppose you run a regression of y on x_1, and get an error term $\hat{e}$. You can then do a scatterplot of error term ($\hat{e}$) and a different variable (x_2) to see how much of the difference can be explained by this variable:
  - twoway scatter e x_2
- You can do a multiple regression
  - regress y_1 x_1 x_2 ...