

Mathematics 231

Lecture 9
Liam O'Brien

1

Announcements

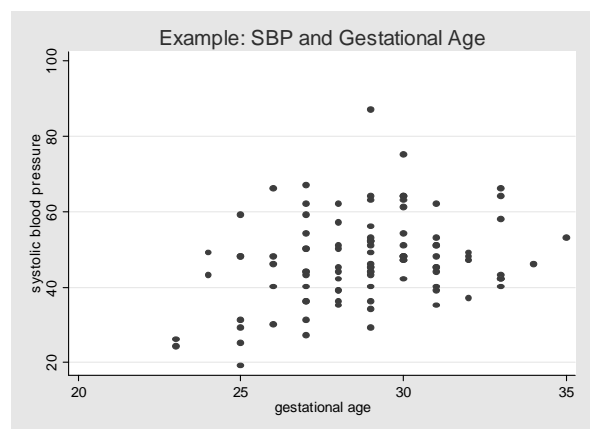
- Reading
 - Today M&M 2.3 117-119
 - M&M 2.4 125-132
 - Next class M&M 2.5 142-151

2

Conditional Standard Deviation

- Conditional Standard Deviation
- Conditional SD in Regression
- Regression Assumptions
- Predicting Y from X versus Predicting X from Y

3



4

Conditional Mean

- The mean SBP of infants with a gestational age of 25 weeks is approximately 42 mm Hg.
- This is the **conditional** mean given a gestational age of 25 weeks, since it is based only on infants who satisfy some condition (25 weeks gestation).
- The **marginal** mean SBP is the mean SBP of all infants (47 mm Hg), regardless of gestational age.

5

Conditional Distributions – One more time

- In general, we can consider the distribution of Y variables (e.g., height) for observations that satisfy some condition $X = x$ (e.g., age equals 25 weeks).
- This is called the **conditional distribution of Y given $X = x$** .
- In a scatter plot, the conditional distribution of Y given $X = x$ is the distribution of points in the vertical strip above a given value of x .

6

Linear Regression

- Linear regression fits a straight line to the conditional mean of Y , given X .
- How might we determine the conditional SD at any given x =value?
- For example, what is the conditional SD of SBP for infants with a gestational age of 25 weeks?

7

Conditional Standard Deviation

- Conditional SD of Y given $X = x$:
 - In a scatterplot, the conditional standard deviation of Y given $X = x$ is the spread of points in the vertical strip above a given value of x .
- The spread is determined relative to the center (mean) of the distribution of points in the vertical strip.

8

Conditional Standard Deviation

- **Conditional SD of Y given X = x:**
- The spread can be determined by the residuals:

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

- In calculating the SD, should we consider the spread of points only in the vertical strip above the particular value of x (e.g., 25 weeks)?

9

Recall: Regression Assumptions

- The regression line estimates the conditional mean of Y given $X=x$ for any point x if the following assumptions are met.
 1. Conditional mean of Y is a linear function of X .
 2. Conditional SD of Y is constant for all X .
- We often make an additional assumption:
 3. The conditional distribution of Y is a normal distribution for any value of x .

10

Conditional Standard Deviation

- **Conditional SD of Y given X = x:**

$$s_{y|x} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

- Measures the degree of scatter of the points about the regression line (in any give vertical strip).
- In Stata, this is denoted by **Root Mean Square Error (MSE)**.
- This is the variation NOT explained by the linear regression model.

11

Example: Height and Age

TABLE 2.7 Mean height of Kalama children

Age x in months	Height y in centimeters
18	76.1
19	77.0
20	78.1
21	78.2
22	78.8
23	79.7
24	79.9
25	81.1
26	81.2
27	81.8
28	82.8
29	83.5

12

Example: Height and Age

```
. regress height age
```

Source	SS	df	MS	Number of obs = 12	
Model	57.6548678	1	57.6548678	F(1, 10) =	880.00
Residual	.655171562	10	.065517156	Prob > F =	0.0000
Total	58.3100394	11	5.30091267	R-squared =	0.9888
				Adj R-squared =	0.9876
				Root MSE =	.25596

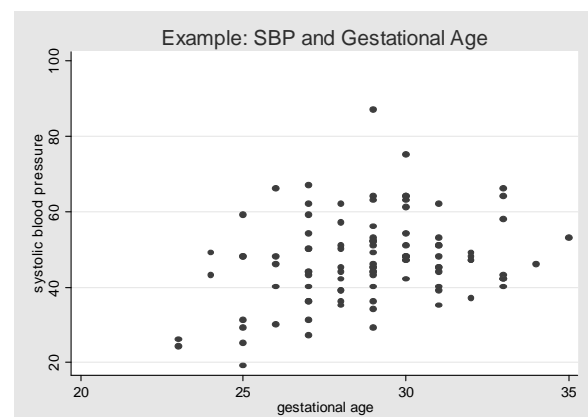
height	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.6349653	.0214047	29.66	0.000	.5872726	.682658
_cons	64.92832	.508409	127.71	0.000	63.79551	66.06112

13

Example: SBP and Gestational Age

- Data: SBP and gestational age for 100 infants.
- Mean Gestational age = 28.9 weeks, SD = 2.53 weeks.
- Mean SBP = 47.1 mm Hg, SD = 11.4 mm Hg
- Correlation between gestational age and SBP, $r=0.28$.
- Suppose we are interested in predicting SBP from gestational age.

14



15

Example: SBP and Gestational Age

- Regression line:

$$SBP = 10.6 + 1.26 (\text{gestational age})$$

Conditional SD = 11

- Of infants 25 weeks in gestation, what proportion have a SBP between 31 and 53 mm Hg.

16

Example: SBP and Gestational Age

- If we make assumption (3), then the SBP of 25-week old infants have a normal distribution with mean = $10.6 + 1.26$ (gestational age).
- What's the SD of this conditional distribution?
11 mm Hg.
- Of 25-week old infants, what proportion have an SBP between 31 and 53 mm Hg?

17

Recall: The Empirical Rule

- All normal distributions have the following property:
- 68% of the area under the curve lies with σ of the mean.
- 95% of the area of the curve lies within 2σ of the mean.
- 99.7% of the area of the curve lies within 3σ of the mean.

18

Example: SBP and Gestational Age

- For 25-week old infants, SBP's between 31 and 53 mm Hg are 1 SD above and below the mean (42 mm Hg for 25-week old infants).
- So, 68% of 25-week old infants have SBP's between 31 and 53 mm Hg.
- **Previously:** Calculating the proportion of **all** infants with SBP's between 31 and 53.
- **Now:** Can calculate for infants of a given age only.

19

Predicting X from Y

- Regression line:
 $SBP = 10.6 + 1.26$ (gestational age)
- For an infant 25 weeks into gestation, our prediction for its SBP is
 $SBP = 10.6 + 1.26(25) = 42$ mm Hg
- Now consider an infant with an SBP of 42 mm Hg, what is our prediction of its gestational age?

20

Predicting X from Y

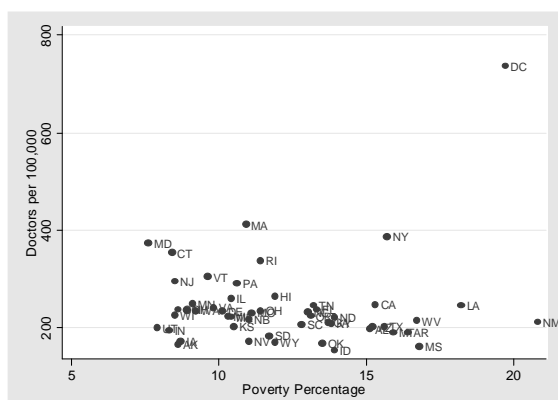
- Consider the scatterplot of SBP (Y) versus gestational age (X).
- Then the conditional mean gestation age of infants with an SBP of 42 mm Hg is the mean of the points within the horizontal strip at Y = 42.
- In general, predicting Y from X is NOT the same as predicting X from Y (although the data in this example provide similar regression results).

21

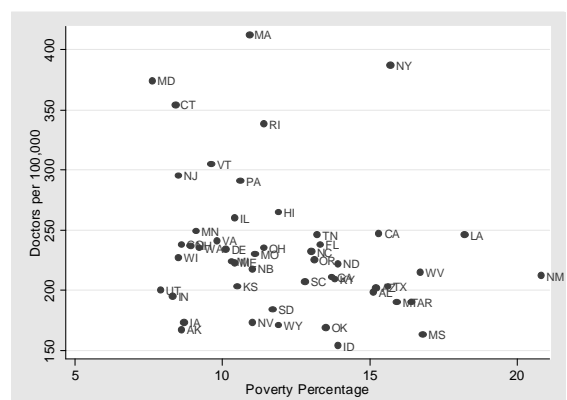
Example: Poverty and Doctors

- Between 1997 and 1999, data were collected on poverty rates and the number of doctors in each of the 50 states and DC.
- Of interest is how strongly poverty and the number of doctors is related.
- How do we expect these two to relate?

22



23



24

Example: Poverty and Doctors

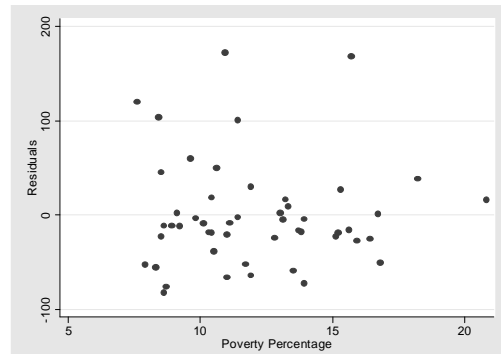
```
. regress doctors poverty if state != "DC"
```

Source	SS	df	MS			
Model	8670.18752	1	8670.18752	Number of obs =	50	
Residual	154011.032	48	3208.56318	F(1, 48) =	2.70	
Total	162681.22	49	3320.0249	Prob > F =	0.1067	
				R-squared =	0.0533	
				Adj R-squared =	0.0336	
				Root MSE =	56.644	

doctors	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
poverty	-4.375241	2.661601	-1.64	0.107	-9.726749	.9762674
_cons	287.0354	33.04209	8.69	0.000	220.5998	353.471

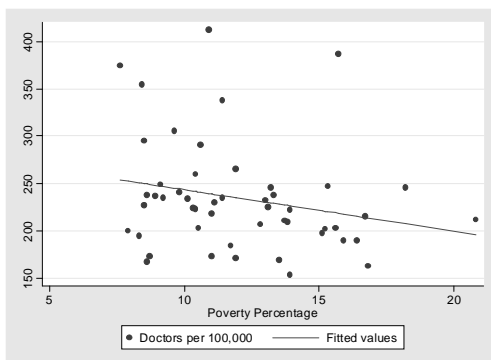
25

Residual Plot



26

Example: Poverty and Doctors



27

Reverse the Prediction

```
. regress poverty doctors if state!="DC"
```

Source	SS	df	MS			
Model	24.1387972	1	24.1387972	Number of obs =	50	
Residual	428.784391	48	8.93300815	F(1, 48) =	2.70	
Total	452.923188	49	9.24333037	Prob > F =	0.1067	
				R-squared =	0.0533	
				Adj R-squared =	0.0336	
				Root MSE =	2.9888	

poverty	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
doctors	-.0121812	.0074102	-1.64	0.107	-.0270804	.002718
_cons	14.89854	1.787209	8.34	0.000	11.30511	18.49196

- Is this the same as the line we'd get by rearranging terms from the previous regression? NO

28