


Simple Linear Regression: John McGready



Section E

Measuring the Strength of A Linear Association

Strength of Association

- The slope of a regression line estimates the magnitude and direction of the relationship between y and x; it encapsulates how much y differs on average with differences in x
- The slope estimate and standard error can be used to address the uncertainty in this estimate with regards to the true magnitude and direction of the association in the population from which the sample was taken from
- Slopes do not impart any information about how well the regression line fits the data in the sample; the slope gives no indication of how close the points get to the estimated regression line

2

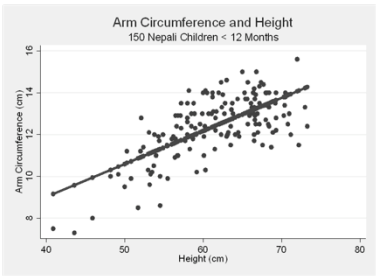
Strength of Association

- Another quantity that can be estimated via linear regression is the coefficient of determination, R^2
 - This is a number that ranges from 0 to 1, with larger values indicate “closer fits” of the data points and regression line
- R^2 measures strength of association by comparing variability of points around the regression line to variability in y-values ignoring x

3

Example: Arm Circumference and Height

- How close do the points get to the line: can we quantify?



4

Example: Arm Circumference and Height

- (SR1 flashback) the sample standard deviation of the y-values ignoring the corresponding potential information in x is

$$s = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}}$$

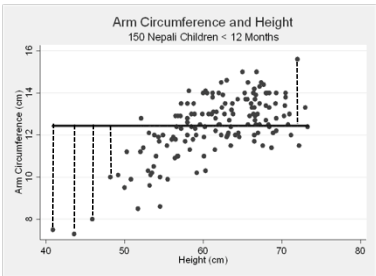
- This measures how far on average each of the sample y-values falls from the overall mean all y-values
- In this example $s = 1.48$ cm

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|----------|-----|----------|-----------|-----|------|
| armcirc | 150 | 12.42733 | 1.479875 | 7.3 | 15.6 |

5

Example: Arm Circumference and Height

- “Visualization” on the scatterplot



6

Simple Linear Regression: John McGready

Example: Arm Circumference and Height

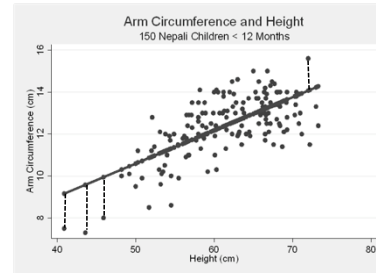
- Standard deviation of regression, referred to as root mean square error is "average" distance of points from the line: how far on average each y falls from its mean predicted by its corresponding x -value

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

7

Example: Arm Circumference and Height

- Each distance is $y - \hat{y} = y - (\hat{\beta}_0 + \hat{\beta}_1 x)$: this is computed for each data point in the sample



8

Using Stata: Arm Circumference and Height

- regress command in Stata gives $s_{y|x}$

```
. regress armcirc height
-----+-----
Source |      SS      df       MS              Number of obs =   150
-----+-----
Model | 148.874597      1 148.874597          F( 1, 148) = 124.30
Residual | 177.263335     148 1.19772523          Prob > F      = 0.0000
-----+-----
Total | 326.137932     149 2.18884518          R-squared     = 0.4565
                                          Adj R-squared = 0.4528
                                          Root MSE    = 1.0944

-----+-----
armcirc |      Coef.   Std. Err.      t    P>|t|   [95% Conf. Interval]
-----+-----
height |  .1579469   .0141671    11.15  0.000   .1299511   .1859428
 _cons |  2.695906   .8774225     3.07  0.003   .9620116   4.4299
```

9

Example: Arm Circumference and Height

- If $s = s_{y|x}$, then knowing x does not yield a better guess for the mean of y than using the overall mean \bar{y} (flat regression line)
- The smaller $s_{y|x}$ is relative to s , the closer the points are to the regression line
- R^2 functionally measures how much smaller $s_{y|x}$ is than s : as such it is an estimate of the amount of variability in y explained by taking x into account

10

Using Stata: Arm Circumference and Height

- regress command in Stata gives R^2 : child's height explains (an estimated) 46% of the variation in arm circumferences

```
. regress armcirc height
-----+-----
Source |      SS      df       MS              Number of obs =   150
-----+-----
Model | 148.874597      1 148.874597          F( 1, 148) = 124.30
Residual | 177.263335     148 1.19772523          Prob > F      = 0.0000
-----+-----
Total | 326.137932     149 2.18884518          R-squared     = 0.4565
                                          Adj R-squared = 0.4528
                                          Root MSE    = 1.0944

-----+-----
armcirc |      Coef.   Std. Err.      t    P>|t|   [95% Conf. Interval]
-----+-----
height |  .1579469   .0141671    11.15  0.000   .1299511   .1859428
 _cons |  2.695906   .8774225     3.07  0.003   .9620116   4.4299
```

11

Example: Arm Circumference and Height

- R^2 and r
- $r =$ the properly signed square root of R^2 ; the proper sign is the same sign as the slope in the regression
- r is called the correlation coefficient (not to be confused with the "regression coefficients"—great names, huh)
- Allowable values
 - $0 \leq R^2 \leq 1$
 - If relationship between y and x is positive $0 \leq r \leq 1$
 - If relationship between y and x is negative $-1 \leq r \leq 0$
- In this example, $r = +\sqrt{R^2} = +\sqrt{0.46} = 0.68$

12

Simple Linear Regression: John McGready

Example: Arm Circumference and Height

- So from the example: child height explains (an estimated) 46% of the variation in arm circumferences
 - This is just an estimate based on the sample; a 95% CI can be computed but its not easy to do (and not given readily by the computer); also the procedure for estimating the 95% CI is not so good
- So this means an estimated 54% of the variability in arm circumference is not explained by child's height
 - Some if this unexplained variability may be explained by factors other than height
 - Multiple linear regression will allow us to estimate the relationship between arm circumference, height, and other child characteristics in one analysis

13

Example 2: Hemoglobin and "Packed Cell Volume"

- regress command in Stata gives R^2 : PCV explains (an estimated) 51% of the variation in hemoglobin levels

```
. regress Hb PCV
```

| Source | SS | df | MS | Number of obs = | 21 |
|----------|------------|----|------------|-----------------|--------|
| Model | 53.7803079 | 1 | 53.7803079 | F(1, 19) = | 19.81 |
| Residual | 51.5711174 | 19 | 2.71426934 | Prob > F = | 0.0003 |
| | | | | R-squared = | 0.5101 |
| | | | | Adj R-squared = | 0.4847 |
| | | | | Root MSE = | 1.6475 |

| Hb | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] |
|-------|----------|-----------|------|-------|----------------------|
| PCV | .2033502 | .0456835 | 4.45 | 0.000 | .1077325 .2989668 |
| _cons | 5.77645 | 1.913624 | 3.02 | 0.007 | 1.771188 9.781712 |

14

Example: Hemoglobin and PCV

- regress command in Stata gives R^2 of 0.51; the slope is positive, so $r = +\sqrt{R^2} = +\sqrt{0.51} = 0.71$

15

Example 3: Wages and Years of Education

- regress command in Stata gives R^2 : years of education explains (an estimated) 15% of the variation in hourly wages

```
. regress wage edlevel
```

| Source | SS | df | MS | Number of obs = | 534 |
|----------|------------|-----|------------|-----------------|--------|
| Model | 2053.29014 | 1 | 2053.29014 | F(1, 532) = | 90.85 |
| Residual | 12023.4084 | 532 | 22.6003917 | Prob > F = | 0.0000 |
| | | | | R-squared = | 0.1450 |
| | | | | Adj R-squared = | 0.1445 |
| | | | | Root MSE = | 4.754 |

| wage | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] |
|---------|-----------|-----------|-------|-------|----------------------|
| edlevel | .7504607 | .0787337 | 9.53 | 0.000 | .5957936 .9051279 |
| _cons | -.7459796 | 1.045454 | -0.71 | 0.476 | -2.799704 1.307745 |

- Here $r = +\sqrt{R^2} = +\sqrt{0.15} = 0.39$

16

Example 4: Arm Circumference and Child Sex

- regress command in Stata gives R^2 : sex (female = 1) explains (an estimated) 0.2% of the variation in arm circumference

```
. regress armcirc sex
```

| Source | SS | df | MS | Number of obs = | 150 |
|----------|------------|-----|------------|-----------------|---------|
| Model | .608719585 | 1 | .608719585 | F(1, 148) = | 0.28 |
| Residual | 325.529212 | 148 | 2.1995217 | Prob > F = | 0.5956 |
| | | | | R-squared = | 0.0019 |
| | | | | Adj R-squared = | -0.0049 |
| | | | | Root MSE = | 1.4931 |

| armcirc | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] |
|---------|-----------|-----------|-------|-------|----------------------|
| sex | -.1274182 | .2422072 | -0.53 | 0.600 | -.6066494 .3512129 |
| _cons | 12.49189 | .1724044 | 72.46 | 0.000 | 12.1512 12.83250 |

- Here $r = -\sqrt{R^2} = -\sqrt{0.002} = -0.045$; in this sample of data female sex is negatively correlated with arm circumference

17

What Is a "Good" R^2 ?

- There are some important things to keep in mind about R^2 and r
 - These quantities are both estimates based on the sample of data frequently reported without some recognition of sampling variability (for example, a 95% confidence interval)
 - Low R^2 and r is not necessarily "bad"
 - Many outcomes can not/will not be fully or close to fully explained, in terms of variability, by any one single predictor

18

Simple Linear Regression: John McGready

What Is a "Good" R^2 ?

- The higher the R^2 values, the better the x predicts y for individuals in a sample/population, as individual y-values vary less about their estimated means based on x
- However, there may be important overall associations between the mean of y and x even though there is a lot of individual variability in y-values about their means estimated by x
 - In the wages example, years of education explained an estimated 15% of the variability in hourly wages
 - The association was statistically significant showing that average wages were greater for persons with more years of education
 - However, for any single education level (year), still there is a lot of variation in wages for individual workers

19

Slope versus R^2

- Slope estimates the magnitude and direction of the relationship between y and x
 - Estimates a mean difference in y for two groups who differ by one-unit in x
 - The slope will change if the units change for y and/or for x
 - *Larger slopes are not indicative of stronger linear association: smaller slopes are not indicative of weaker linear association*
- R^2 measures strength of linear association; r measures strength and direction
 - Neither R^2 or r measures magnitude
 - Neither R^2 or r changes with changes in units

20

Using Stat: Arm Circumference and Height

- Regression of arm circumference (cm) on height in centimeters

```
. regress armcirc height
-----+-----
Source |      SS      df       MS              Number of obs =   150
-----+-----
Model | 148.874597    1 148.874597          F( 1, 148) = 124.30
Residual | 177.263335   148 1.19772523          Prob > F      = 0.0000
-----+-----
Total | 326.137932   149 2.18884518          R-squared     = 0.4565
                                          Adj R-squared  = 0.4528
                                          Root MSE     = 1.0944

-----+-----
armcirc |      Coef.   Std. Err.      t    P>|t|   [95% Conf. Interval]
-----+-----
height | .1579469   .0141671    11.15  0.000   .1299511   .1859428
_cone | 2.695906   .8774225     3.07  0.003   .9620116   4.4298
```

$$\hat{y} = 2.7 + 0.16x$$

- $R^2 = 0.46$ or 46%; $\hat{\beta}_1 = 0.16$

21

Using Stat: Arm Circumference and Height

- Regression of arm circumference on height in inches

```
. regress armcirc height_inch
-----+-----
Source |      SS      df       MS              Number of obs =   150
-----+-----
Model | 148.874589    1 148.874589          F( 1, 148) = 124.30
Residual | 177.263343   148 1.19772529          Prob > F      = 0.0000
-----+-----
Total | 326.137932   149 2.18884518          R-squared     = 0.4565
                                          Adj R-squared  = 0.4528
                                          Root MSE     = 1.0944

-----+-----
armcirc |      Coef.   Std. Err.      t    P>|t|   [95% Conf. Interval]
-----+-----
height_inch | .4008806   .035857    11.15  0.000   .3298251   .471936
_cone | 2.695906   .8774225     3.07  0.003   .9620119   4.429801
```

$$\hat{y} = 2.7 + 0.40x$$

- $R^2 = 0.46$ or 46%; $\hat{\beta}_1 = 0.40$

22