Refresher Course in Calculus, Probability, and Statistics

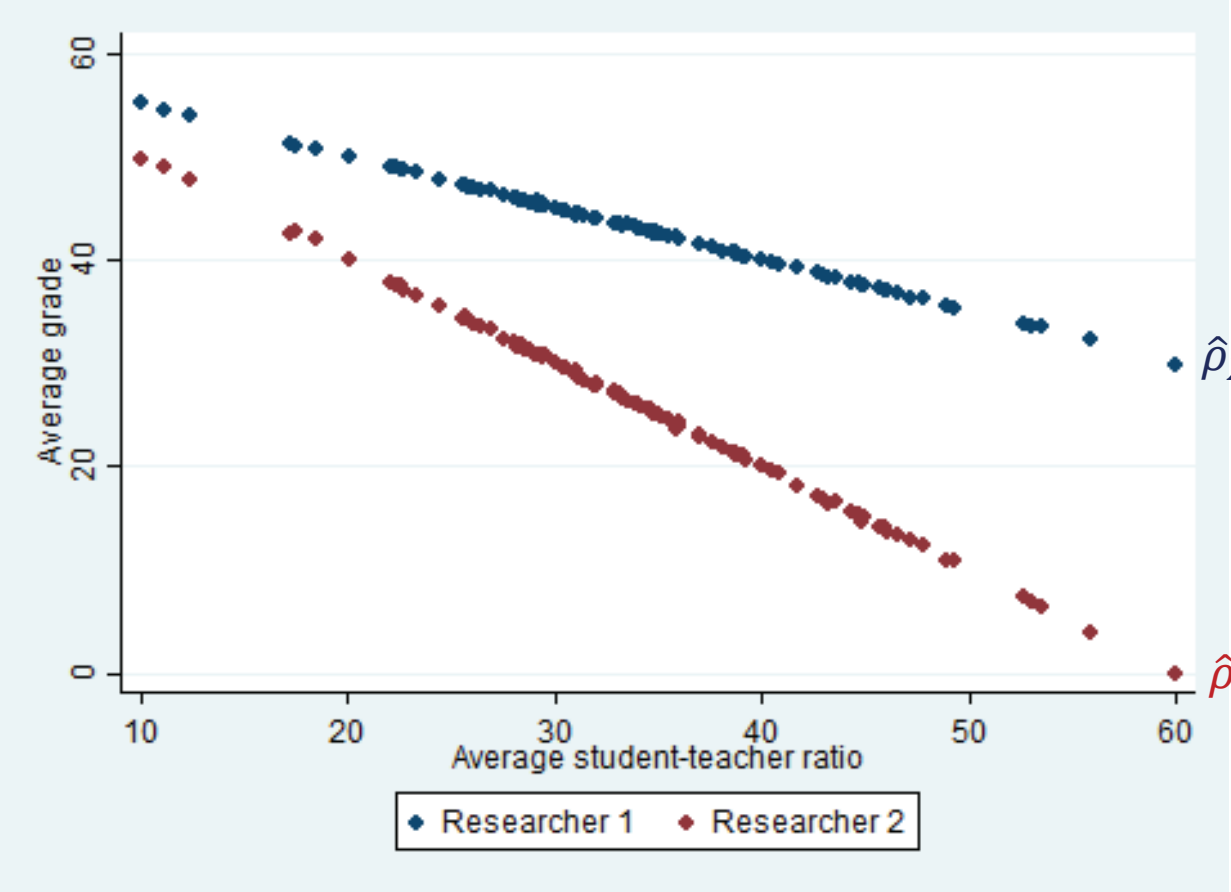
Day 5a: Linear Regression

Scatterplot, sample variance and sample correlation

- What is the relationship between variable X and Y?
- Different ways to summarize relationship between variables
  - Scatterplot: plot of n observations on X_i and Y_i, each observation is represented by a point (X_i, Y_i)
    - Good idea to begin an analysis by drawing one.
  - Sample covariance:
    - $\hat{\sigma}_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}[(X_i - \bar{X})(Y_i - \bar{Y})]$
  - Sample correlation:
    - $\hat{\rho}_{xy} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x \hat{\sigma}_y}$
    - Measure of strength of linear association between X and Y in sample

Sample Covariance and Sample Correlation

➲ Hypothetical example: Grades and student teacher ratio



$\hat{\rho}_{XY} = -0.9998$

$\hat{\rho}_{XY} = -0.9998$

Linear Regression

Univariate Model

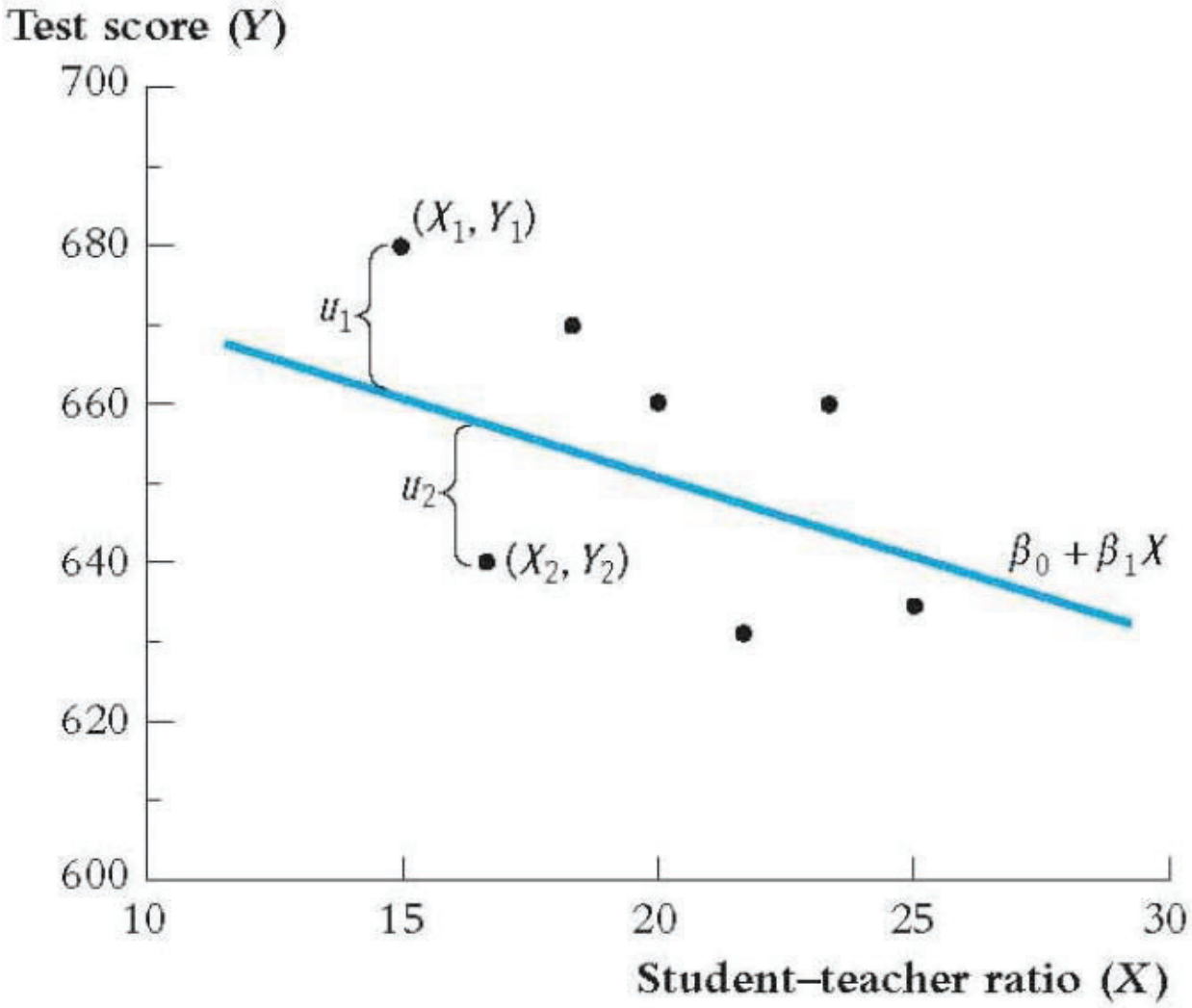
➲ By how much does Y change if X changes by one unit?

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- Y_i: *dependent variable*, *regressand* or *left-hand variable*
- X_i: *independent variable*, *predictor*, *regressor* or *right-hand variable*
- β_1: Slope (of the *population regression line*)
- β_0: Intercept (of the *population regression line*)
- u_i: *error term*

Linear Regression

Univariate Model (continued)



Stock & Watson (2012)

Linear Regression

Univariate Model (continued)

- How are coefficients of the linear regression model estimated?
  - In practice: β_0 and β_1 are unknown
  - → we must use data to estimate them

- Ordinary Least Squares Estimator (OLS)
  - $\hat{\beta} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2}$
  - $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
  - Predicted values: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
  - Residuals: $\hat{u}_i = Y_i - \hat{Y}_i$

- Practical Interpretation of coefficients
  - $\hat{\beta}_1$: average change in Y associated with a change of one unit in X
  - $\hat{\beta}_0$: expected value of Y if X is zero

Linear Regression

Univariate Model (continued)

- Measures of fit
  - The R^2
    - $R^2 \in [0,1]$
    - Interpretation: Fraction of variance in the dependent that is explained by the variance in the independents
    - Alternative measurein the multivariatecase is adjust R square to control the number of independent variables

Linear Regression

Univariate Model (continued)

```
. reg testscr str

      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     22.58
       Model |  7794.11004         1  7794.11004   Prob > F        =    0.0000
    Residual |  144315.484       418  345.252353   R-squared       =    0.0512
-------------+----------------------------------   Adj R-squared   =    0.0490
       Total |  152109.594       419  363.030056   Root MSE        =    18.581

------------------------------------------------------------------------------
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .4798256    -4.75   0.000     -3.22298   -1.336637
       _cons |    698.933   9.467491    73.82   0.000     680.3231    717.5428
------------------------------------------------------------------------------
```

Population model: $Test\ Score_i = \beta_0 + \beta_1 STR_i + u_i$

Estimated model: $\widehat{Test\ Score}_i = \hat{\beta}_0 + \hat{\beta}_1 STR_i = 698.93 - 2.28 * STR_i$

(We discuss all other numbers in following slides)

Linear Regression

Univariate Model (continued)

- Hypothesis testing concerning $\hat{\beta}_1$
  - Set up hypothesis/hypotheses

	Two-sided test	**One-sided test**
H_0	$\beta_1 = c$	$\beta_1 \geq c\ (\beta_1 \leq c)$
H_1	$\beta_1 \neq c$	$\beta_1 < c\ (\beta_1 > c\ )$

  - Calculate the empirical t-value

$$t_{emp} = \frac{estimator - value\ under\ the\ Null}{standard\ error\ of\ the\ estimator} = \frac{\hat{\beta}_1 - c}{\hat{\sigma}_{\hat{\beta}_1}} \left(= \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \right)$$

  - Compare this value to the critical values of a t-distribution with $n - k - 1$ df
    - If $\left|t_{emp}\right| > |t_{crit}^{\alpha}|$ you can reject the Null at a level of significance of α (= with a Type I error probability of $< \alpha$)
  - OR/AND: Compute the p-value
    - $p - value = \Pr_{H_0}(|t| > \left|t_{emp}\right|)$
    - If n is large → $t_{n-k-1} \sim N(0,1)$
      - Two-sided: $p - value = Pr_{H_0}(|Z| > \left|t_{emp}\right|) = 2\Phi(-\left|t_{emp}\right|)$
      - One-sided: $p - value = Pr_{H_0}(Z < t_{emp}) = \Phi(t_{emp})$

Z	p
1.645	10%
1.96	5%
2.58	1%

Linear Regression

Univariate Model (continued)

- Confidence Intervals for $\hat{\beta}_1$
  - Definition of a $1-\alpha\%$ Confidence Interval:
    - Set of values that cannot be rejected using a two-sided hypothesis at a $\alpha\%$ significance level
    - Interval that contains the true value of β_1 with $1-\alpha\%$ (read: in $1-\alpha\%$ of all samples)

- Calculating the $1-\alpha\%$ Confidence Interval
  - Two-sided: $CI_{1-\alpha} = \left[\hat{\beta}_1 - t_{n-k-1,\alpha} \cdot \hat{\sigma}_{\hat{\beta}_1}, \hat{\beta}_1 + t_{n-k-1,\alpha} \cdot \hat{\sigma}_{\hat{\beta}_1}\right]$
  - One-sided ($H_1: \beta_1 > c$): $CI_{1-\alpha} = \left[\hat{\beta}_1 - t_{n-k-1,\alpha} \cdot \hat{\sigma}_{\hat{\beta}_1}, \infty\right)$
  - One-sided ($H_1: \beta_1 < c$): $CI_{1-\alpha} = \left(-\infty, \hat{\beta}_1 + t_{n-k-1,\alpha} \cdot \hat{\sigma}_{\hat{\beta}_1}\right]$
    - with $t_{n-k-1,\alpha}$: critical value of the t-distribution with $n-k-1$ d.f. and $\alpha\%$ significance
    - If n is large these values will approach the critical values of the standard normal distribution (10%: 1.645; 5%: 1.96; 1%: 2.58)

Linear Regression

Univariate Model (continued)

```
. reg testscr str, robust

Linear regression                               Number of obs     =        420
                                                F(1, 418)         =      19.26
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0512
                                                Root MSE          =     18.581

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
       _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

Population model: $Test\ Score_i = \beta_0 + \beta_1 STR_i + u_i$

Estimated model: $\widehat{Test\ Score}_i = \hat{\beta}_0 + \hat{\beta}_1 STR_i = \underset{(10.36)}{698.93^{***}} - \underset{(0.52)}{2.28^{***}} * STR_i$

Linear Regression

Multivariate Model

➲ More general:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, \qquad i = 1, \ldots, n, \qquad k + 1 \leq n$$

- Predicted values: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}$
- Residuals: $\hat{u}_i = Y_i - \hat{Y}_i$