Public Policy 529
Linear Regression
Part 2

Jonathan Hanson

Gerald R. Ford School of Public Policy
University of Michigan

November 20, 2023

Outline

1. Recap

2. Statistical Significance

3. Explaining Variance in y

Outline

1. Recap

2. Statistical Significance

3. Explaining Variance in y

Linear Regression

- Fits the dependent variable (y) as a linear function of the independent variable (x).
- Like correlation analysis, it captures linearity of the relationship between x and y.
- Unlike correlation analysis, it estimates the magnitude of the relationship. How much does y change for a given change in x?
- The dependent variable is interval-level. The independent variables are interval-level or dichotomous.

The Bivariate Linear Regression Model

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

The relationship between x and y is specified by the "true" population parameters β_0, β_1, and u_i.

β_0 (beta naught) is the intercept of the linear function.

β_1 (beta 1) is the slope, or the coefficient on x.

u_i is the part of each y_i that is stochastic. We model it as drawn from a normal distribution with mean 0 and variance σ_u^2.

What we Estimate

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

With linear regression, we estimate the β's. These produce the intercept and slope of the best-fit line.

- $\hat{y}_i$ is determined mathematically by the $\hat{\beta}$'s and x_i. The regression line depicts $\hat{y}$ across the values of x.
- The difference between y_i and $\hat{y}_i$, which is prediction error, represents our best guess at u_i for each observation.

$$y_i - \hat{y}_i = \hat{u}_i$$

- We use these prediction errors to estimate the variance of the distribution of u, in other words $\hat{\sigma}_u^2$.

Example: Water Sources and Infant Mortality

Let's test the idea that, across countries, access to an improved water source (well or plumbing system) is associated with lower infant mortality.

InfMort (dependent variable): the number of babies, out of every 1,000 born, that die before age 1. Measured in the year 2000.

Water (independent variable): the percentage of the country's population that has access to an improved water source. Measured in the year 2000.

$$\text{InfMort}_i = \beta_0 + \beta_1 \text{Water}_i + u_i$$

Stata Regression Output

```
. reg InfMort Water

      Source |       SS           df       MS      Number of obs   =       172
-------------+----------------------------------   F(1, 170)       =    430.79
       Model |  152188.787         1  152188.787   Prob > F        =    0.0000
    Residual |   60057.948       170  353.282047   R-squared       =    0.7170
-------------+----------------------------------   Adj R-squared   =    0.7154
       Total |  212246.735       171  1241.20897   Root MSE        =    18.796

------------------------------------------------------------------------------
     InfMort |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Water |  -1.537692   .0740864   -20.76   0.000     -1.68394   -1.391444
       _cons |    167.306   6.222146    26.89   0.000     155.0234    179.5886
------------------------------------------------------------------------------
```

R Regression Output

```
> model <- lm(InfMort ~ Water, data = infmort_data)
> summary(model)

Call:
lm(formula = InfMort ~ Water, data = infmort_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-59.871  -9.143  -3.956   8.461  57.449 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 167.30601    6.22215   26.89   <2e-16 ***
Water        -1.53769    0.07409  -20.75   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.8 on 170 degrees of freedom
  (23 observations deleted due to missingness)
Multiple R-squared:  0.717,	Adjusted R-squared:  0.7154 
F-statistic: 430.8 on 1 and 170 DF,  p-value: < 2.2e-16
```

Interpretation

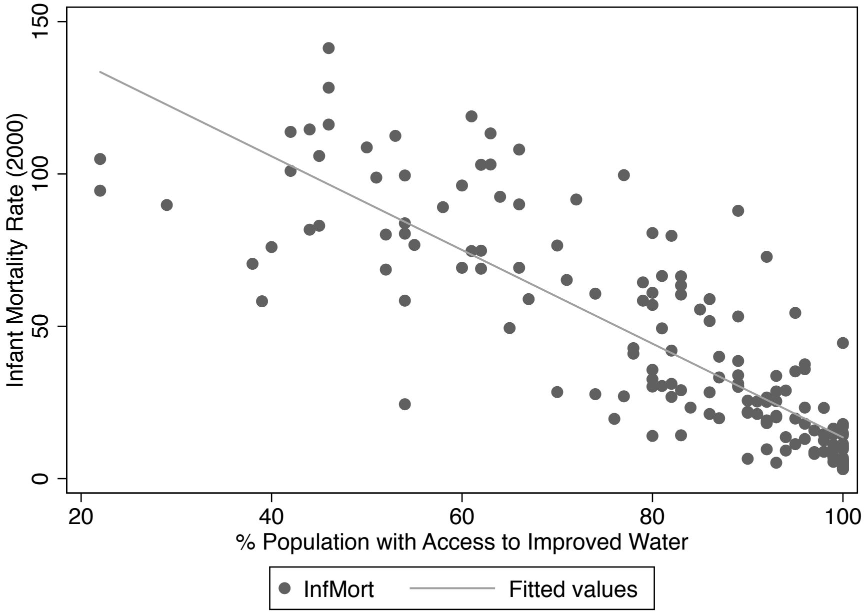
$$\widehat{\text{InfMort}} = 167.3 - 1.5(\text{Water})$$

- Generically: for each [one-unit] increase in Water, InfMort is predicted to be 1.5 [units] lower. But don't be generic!
- Substantively: for each percentage point increase in the population with access to improved water, we predict 1.5 fewer infant deaths out of every 1,000 births.
- In a country where 20% of the population has access to an improved water source, the infant mortality rate is predicted to be 167.3 - 1.5(20) = 137.

Interpretation

$$\widehat{\text{InfMort}} = 167.3 - 1.5(\text{Water})$$

- The predicted rate of infant mortality in a country with no access to improved water sources would be 167.3
- If access to water is 10 percentage points higher in Country A than Country B, the rate of infant mortality is predicted to be -1.5 $\times$ 10 = -15 deaths lower.
- If 100% of a country's residents have access to improved water sources, then 167.3 - 1.5(100) = 17.3 is the predicted infant mortality rate.



Infant Mortality Rate (2000)
0
50
100
150
20
40
60
80
100
% Population with Access to Improved Water
InfMort
Fitted values

Outline

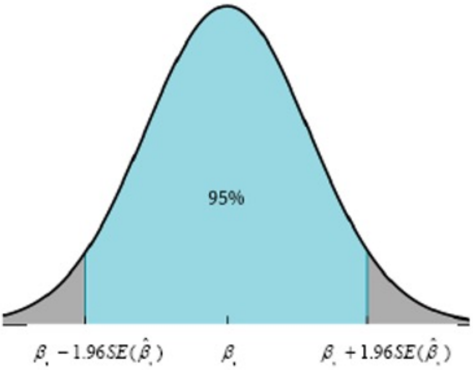
Statistical Significance of Regression Coefficients

- Regression tables report some combination of standard errors, t-statistics, p-values, and 95% confidence intervals.
- The standard error of $\hat{\beta}$ is interpreted the same as usual: it is the typical deviation of $\hat{\beta}$ from the "true" β in repeated trials.
- The formula for the standard error in bivariate regression requires some explanation.
- This formula gets more complicated when we add more independent variables (beyond our scope).

The Sampling Distribution for $\hat{\beta}_1$

- Since the OLS slope estimator (i.e. the formula for $\hat{\beta}_1$) is calculated from a sample, it is subject to random sampling error.
- Across repeated samples, the estimated coefficients will vary. They have a sampling distribution.
- We want to use statistical tools to:
  - Quantify the sampling uncertainty associated with $\hat{\beta}_1$ (i.e. estimate its standard errror).
  - Use $\hat{\beta}_1$ to test hypotheses such as $\beta_1 = 0$.
  - Construct a confidence interval for $\hat{\beta}_1$.

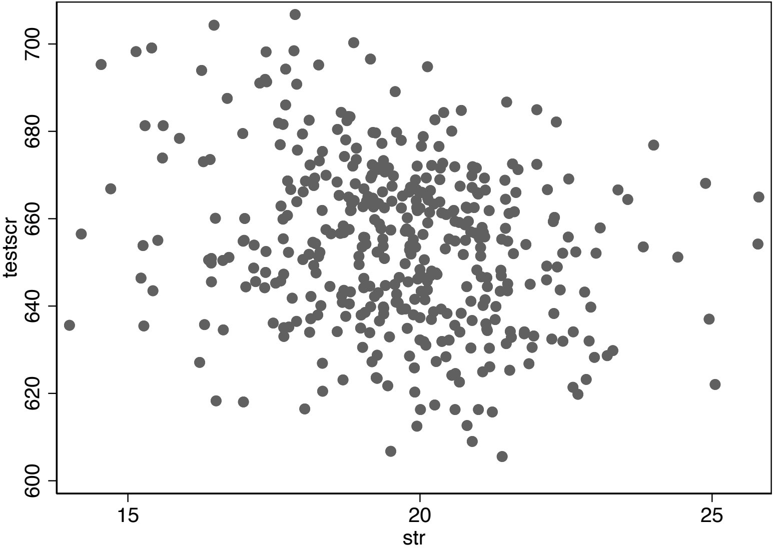
The Sampling Distribution for $\hat{\beta}_1$



- The sampling distribution of $\hat{\beta}_1$ is normal and centered upon the true population regression slope β_1.
- The standard deviation of this distribution is the standard error of $\hat{\beta}_1$, which is $\sigma_{\hat{\beta}_1}$. We need to estimate this standard error.

Visually

Imagine how different samples would lead to different estimates of β_1.



The Standard Error of $\hat{\beta}_1$

- If key assumptions hold, the true standard error for $\hat{\beta}_1$ can be expressed as:

$$\text{se}(\hat{\beta}) \; = \sigma_{\hat{\beta}} = \sqrt{\frac{\sigma_u^2}{\sum(x_i - \bar{x})^2}}$$

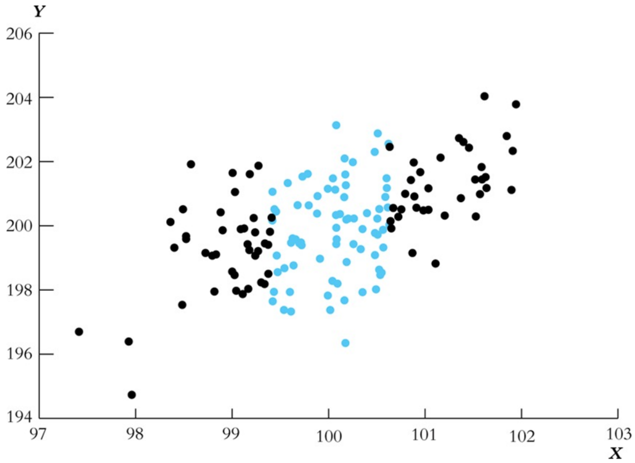
- Just like the standard error for the sample mean, however, we must estimate the standard error for $\hat{\beta}$:

$$\text{est. se}(\hat{\beta}) \; = \hat{\sigma}_{\hat{\beta}} = \sqrt{\frac{\frac{1}{n-2}\sum \hat{u}_i^2}{\sum(x_i - \bar{x})^2}}$$

- We use the squared residuals to estimate the variance of u.

The Variance of x and se($\hat{\beta}$)

Would using the black dots or the blue dots produce a $\hat{\beta}$ with a lower standard error? Why? (hint: look at the formula)



Interpreting the Standard Error of $\hat{\beta}_1$

- The standard error of $\hat{\beta}_1$ is the standard deviation of its sampling distribution.
- It is the amount by which, in repeated samples, our estimated $\hat{\beta}_1$ typically deviates from the "true" β_1.
- It thus measures our level of precision and provides the basis for constructing a confidence interval.
- Degrees of freedom matter for estimating se($\hat{\beta}$). We thus use the t-distribution to represent the sampling distribution.

Testing Hypotheses in OLS

- We can test hypotheses about coefficients just like we do with means or differences of means.
- e.g. the null hypothesis states that the true coefficient is zero
- These tests follow the usual form:

$$t = \frac{\text{estimate} - \text{expected value under } H_0}{\text{se of estimate}}$$

- The critical value of t leaves an area of $\alpha/2$ in each tail of the t-distribution with $n-k$ degrees of freedom.
  - In bivariate regression, $k = 2$.

Testing Hypotheses in OLS

$$H_0: \quad \beta_1 = 0$$
$$H_A: \quad \beta_1 \neq 0$$

- With the usual null hypothesis that $\beta_1 = 0$, the formula for t becomes:
$$t = \frac{\hat{\beta}_1 - H_0}{\text{se}(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)}$$
- p-values for each coefficient come from their t statistics with $n - k$ degrees of freedom

  k is the number of β's that must be estimated. In a bivariate regression, $k = 2$.

Confidence Intervals for OLS Coefficients

$$\hat{\beta}_1 \pm t \cdot \mathsf{se}(\hat{\beta}_1)$$

- In the above, t is chosen to create an interval with the desired level of confidence.
- Again, the degrees of freedom are $n - k$, where k is the number of coefficients being estimated.
- Interpretation is the same as any other confidence interval.

Example: Democracy and Illiteracy

Note the location of all statistics associated with statistical significance.

```
. reg Illiteracy Democracy

      Source |       SS           df       MS      Number of obs   =       123
-------------+----------------------------------   F(1, 121)       =     23.70
       Model |  8501.65111         1  8501.65111   Prob > F        =    0.0000
    Residual |  43398.0246       121  358.661361   R-squared       =    0.1638
-------------+----------------------------------   Adj R-squared   =    0.1569
       Total |  51899.6757       122  425.407178   Root MSE        =    18.938

------------------------------------------------------------------------------
  Illiteracy | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
   Democracy |  -2.902103   .5960784    -4.87   0.000    -4.082197   -1.722008
       _cons |   37.09416   3.248717    11.42   0.000     30.66247    43.52585
------------------------------------------------------------------------------
```

Example: Democracy and Illiteracy

```
> ols_mod <- lm(Illiteracy ~ Democracy, data = democ_wealth)
> summary(ols_mod)

Call:
lm(formula = Illiteracy ~ Democracy, data = democ_wealth)

Residuals:
    Min      1Q  Median      3Q     Max 
-34.851 -14.041  -4.569  10.010  56.626 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  37.0942     3.2487  11.418  < 2e-16 ***
Democracy    -2.9021     0.5961  -4.869 3.43e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.94 on 121 degrees of freedom
  (72 observations deleted due to missingness)
Multiple R-squared:  0.1638,	Adjusted R-squared:  0.1569 
F-statistic:  23.7 on 1 and 121 DF,  p-value: 3.431e-06
```

Example 2: Water Quality and Infant Mortality

```
. reg InfMort Water

      Source |       SS       df       MS              Number of obs =     172
-------------+------------------------------           F(  1,   170) =  430.79
       Model |  152188.787     1  152188.787           Prob > F      =  0.0000
    Residual |   60057.948   170  353.282047           R-squared     =  0.7170
-------------+------------------------------           Adj R-squared =  0.7154
       Total |  212246.735   171  1241.20897           Root MSE      =  18.796

------------------------------------------------------------------------------
     InfMort |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Water |  -1.537692   .0740864   -20.76   0.000     -1.68394   -1.391444
       _cons |    167.306   6.222146    26.89   0.000     155.0234    179.5886
------------------------------------------------------------------------------
```

Outline

Regression and “Explained” Variance

- In linear regression we estimate the line – i.e. the intercept and slope – that minimizes the sum of the squared (vertical) deviations between y and the line ($\hat{y}$).

- Since this line is a function of x, we may say that x “explains” some of the variance of y.

- We would like to know how much of the variance of y is explained by x.

- Be careful: regression by itself does not tell us whether the relationship is causal. That comes from research design and theory.

Breaking Down the Variance of y

- Total Sum of Squares (TSS): the total variation of y

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

- Explained Sum of Squares (ESS): total explained variation of y.

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

- Sum of Squared Errors (SSE): the unexplained variation of y.

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Measuring Explained Variance: R^2

$$TSS = ESS + SSE$$

- The total variation in y can be partitioned into the explained variation (ESS) and unexplained variation (SSE).
- This gives us a way to measure the proportion of variation in y explained by our regression model:

$$R^2 = \frac{TSS - SSE}{TSS} = \frac{ESS}{TSS}$$

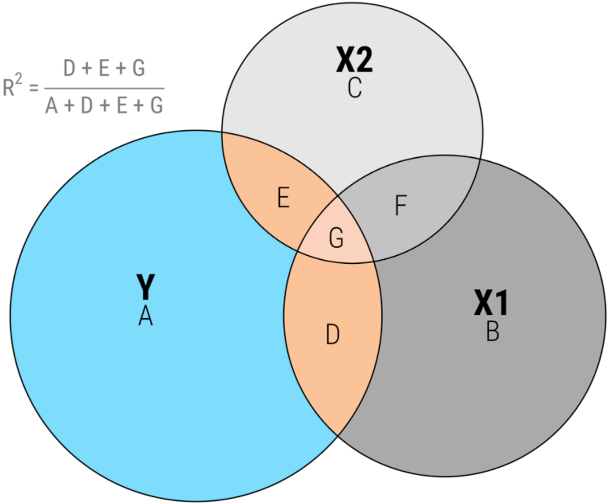
- As a proportion, R^2 is bounded in the range 0 to 1.

Interpreting R^2

$$R^2 = \frac{TSS - SSE}{TSS} = \frac{ESS}{TSS}$$

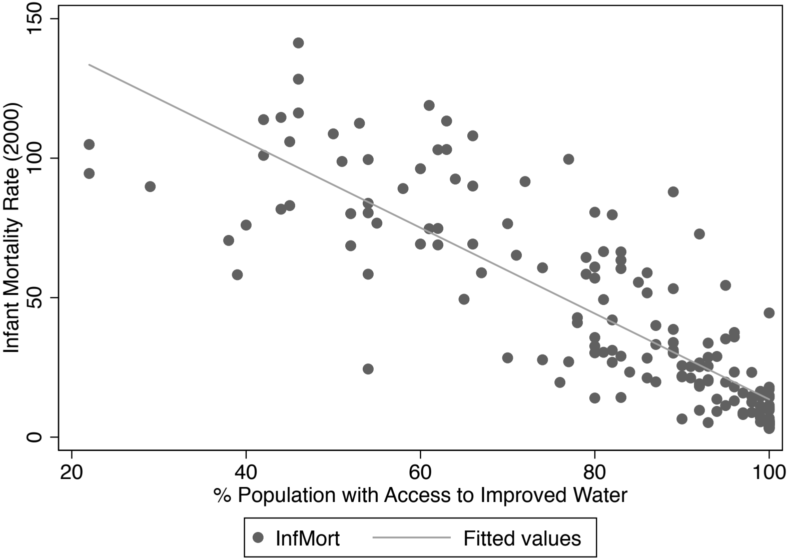
- If our model perfectly explains y, then SSE=0 (in other words, ESS=TSS). This means $R^2 = 1$.
- If our model explains no variation in y, then SSE=TSS (in other words, ESS=0). This means $R^2 = 0$.
- The greater is R^2, the more variation in y is explained by our model.
- Like the correlation coefficient (r), R^2 measures linear association.

Orange area (D + E + G) shows the total variance in outcome Y that is jointly explained by X1 and X2



https://www.andrewheiss.com/blog/2021/08/21/r2-euler/

e.g. Water and Infant Morality: $R^2 = .72$



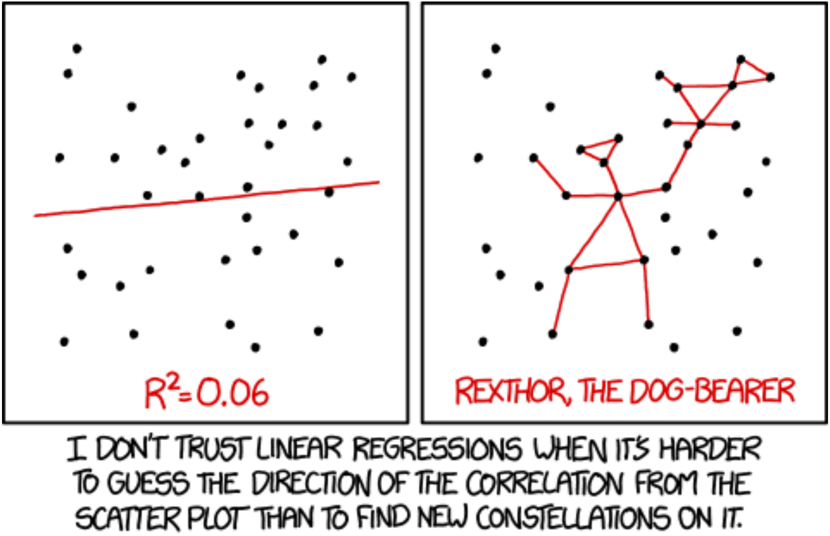
Some Cautions about R^2

- R^2 can be a useful diagnostic tool to assess fit of a model, but maximizing R^2 is not the goal.
  - Our goal is to determine whether particular independent variables have a statistically significant, and substantively relevant, relationship with y.
- The formula is imperfect: R^2 increases every time we add an an independent variable, even if that variable does nothing to explain variation in y.
  - The "adjusted R^2" measure corrects for this phenomenon.
- In general, too much is made of R^2.

What is a "Large" R^2

- Depends on the context.
- R^2 measures the importance of the explanatory variable we model *relative to the importance of other factors*.
- If R^2 is low, then there are other important factors influencing our outcome variable.
- Models of human behavior tend to have low R^2 because there are many things people do that we cannot explain well.

The "Best-Fit"



R²=0.06
REXTHOR, THE DOG-BEARER
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Explained/Unexplained Variance and Stata Output

- Stata reports the R^2 and adjusted R^2 on upper right of the regression output. See next slide.
- In the Source section of the output, Stata reports the TSS, ESS, and SSE.
- In the same section, Stata reports the mean sums of squares, which divide TSS, ESS, and SSE by their degrees of freedom to get "typical" squared deviations.

Location of Variance Statistics

ESS

SSE

TSS

. reg InfMort Water

Source	SS	df	MS
Model	**152188.787**	**1**	**152188.787**
Residual	**60057.948**	**170**	**353.282047**
Total	**212246.735**	**171**	**1241.20897**

Number of obs	=	**172**
F(1, 170)	=	**430.79**
Prob > F	=	**0.0000**
R-squared	=	**0.7170**
Adj R-squared	=	**0.7154**
Root MSE	=	**18.796**

InfMort	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Water	**-1.537692**	**.0740864**	**-20.76**	**0.000**	**-1.68394**	**-1.391444**
_cons	**167.306**	**6.222146**	**26.89**	**0.000**	**155.0234**	**179.5886**

Standard Error of the Estimate ($\hat{y}$)

```
. reg InfMort Water

      Source |       SS       df       MS              Number of obs =     172
-------------+------------------------------           F(  1,   170) =  430.79
       Model |  152188.787     1  152188.787           Prob > F      =  0.0000
    Residual |   60057.948   170  353.282047           R-squared     =  0.7170
-------------+------------------------------           Adj R-squared =  0.7154
       Total |  212246.735   171  1241.20897           Root MSE      =  18.796

------------------------------------------------------------------------------
     InfMort |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Water |  -1.537692   .0740864   -20.76   0.000     -1.68394   -1.391444
       _cons |    167.306   6.222146    26.89   0.000     155.0234    179.5886
------------------------------------------------------------------------------
```

square root

Root MSE is also called the Standard Error of the Estimate. It is the typical deviation of $\hat{y}$ from y and thus represents typical prediction error. It is equal to the square root of the Residual MS.

Similarities with ANOVA

- ANOVA and bivariate regression with a dichotomous independent variable are very similar.
- The "between-group variance" from the ANOVA will match the ESS from the linear regression.
- The "within-group variance" from the ANOVA will match the SSE from the linear regression.
- The R^2 from both will be identical.

The F-test

- Stata always reports an F test as part of the regression output.
- It represents a test of whether the model as a whole is statistically significant for explaining variance in y.
- The t statistics are for individual coefficients; the F test is for all of them working together.
- Along with the F statistic, Stata reports the associated p-value.

Location of F-test Information

```
. reg InfMort Water

      Source |       SS           df       MS      Number of obs   =       172
-------------+----------------------------------   F(1, 170)       =    430.79
       Model |  152188.787         1  152188.787   Prob > F        =    0.0000
    Residual |   60057.948       170  353.282047   R-squared       =    0.7170
-------------+----------------------------------   Adj R-squared   =    0.7154
       Total |  212246.735       171  1241.20897   Root MSE        =    18.796

------------------------------------------------------------------------------
     InfMort |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Water |  -1.537692   .0740864   -20.76   0.000     -1.68394   -1.391444
       _cons |    167.306   6.222146    26.89   0.000     155.0234    179.5886
------------------------------------------------------------------------------
```

The F-test is a ratio of variances with numerator and denominator degrees of freedom. The p-value associated with our F-statistic of 430.79 is .0000.

Diagnostic Statistics in R

```
> model <- lm(InfMort ~ Water, data = infmort_data)
> summary(model)

Call:
lm(formula = InfMort ~ Water, data = infmort_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-59.871  -9.143  -3.956   8.461  57.449 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 167.30601    6.22215   26.89   <2e-16 ***
Water        -1.53769    0.07409  -20.75   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.8 on 170 degrees of freedom
  (23 observations deleted due to missingness)
Multiple R-squared:  0.717,	Adjusted R-squared:  0.7154 
F-statistic: 430.8 on 1 and 170 DF,  p-value: < 2.2e-16
```