Public Policy 529
Linear Regression
Part 1

Jonathan Hanson

Gerald R. Ford School of Public Policy
University of Michigan

November 15, 2023

Outline

1. Preliminaries
2. Digging In
3. Example

Outline

1. Preliminaries

2. Digging In

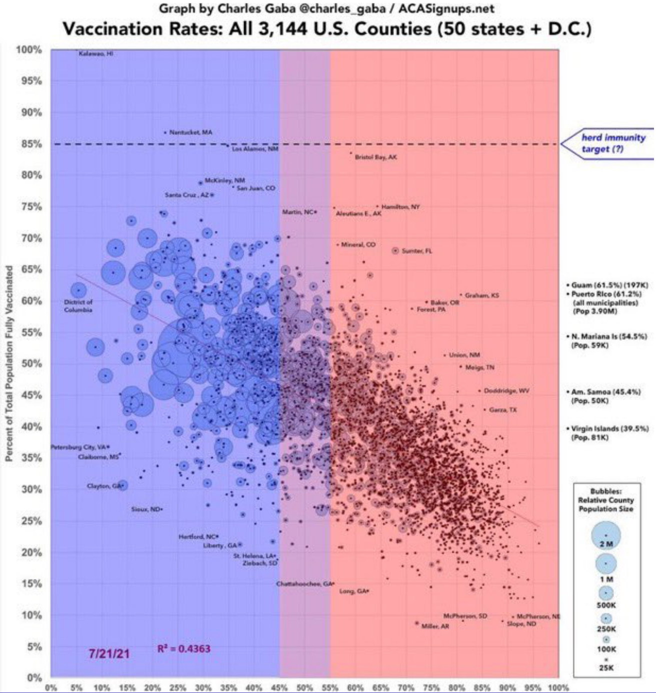
3. Example

Recap

- The Pearson's r correlation coefficient measures the strength (linearity) and direction of the relationship between two interval-level variables.
- The coefficient ranges from -1 to $+1$.
- It is "unit-free" in that its scale is not tied to either variable.
- It thus measures the degree of linearity of the relationship, but not the magnitude of the relationship.

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?
- Regression extends into multiple independent variables. We estimate the effect of one independent variable, controlling for the others.


Graph by Charles Gaba @charles_gaba / ACASignups.net
Vaccination Rates: All 3,144 U.S. Counties (50 states + D.C.)
Percent of Total Population Fully Vaccinated
herd immunity target (?)
Kalawao, HI
Nantucket, MA
Los Alamos, NM
Bristol Bay, AK
McKinley, NM
San Juan, CO
Santa Cruz, AZ
Martin, NC
Aleutians E., AK
Hamilton, NY
Mineral, CO
Sumter, FL
District of Columbia
Graham, KS
Baker, OR
Forest, PA
Union, NM
Meigs, TN
Doddridge, WV
Garza, TX
Petersburg City, VA
Claiborne, MS
Clayton, GA
Sioux, ND
Hertford, NC
Liberty, GA
St. Helena, LA
Ziebach, SD
Chattahoochee, GA
Long, GA
McPherson, SD
Miller, AR
McPherson, NE
Slope, ND
7/21/21
R² = 0.4363
• Guam (61.5%) (197K)
• Puerto Rico (61.2%) (all municipalities) (Pop 3.90M)
• N. Mariana Is (54.5%) (Pop. 59K)
• Am. Samoa (45.4%) (Pop. 50K)
• Virgin Islands (39.5%) (Pop. 81K)
Bubbles: Relative County Population Size
2 M
1 M
500K
250K
100K
25K
100%
95%
90%
85%
80%
75%
70%
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
0% 5% 10% 15% 20% 25% 30% 35% 40% 45% 50% 55% 60% 65% 70% 75% 80% 85% 90% 95% 100%

The Linear Functional Form

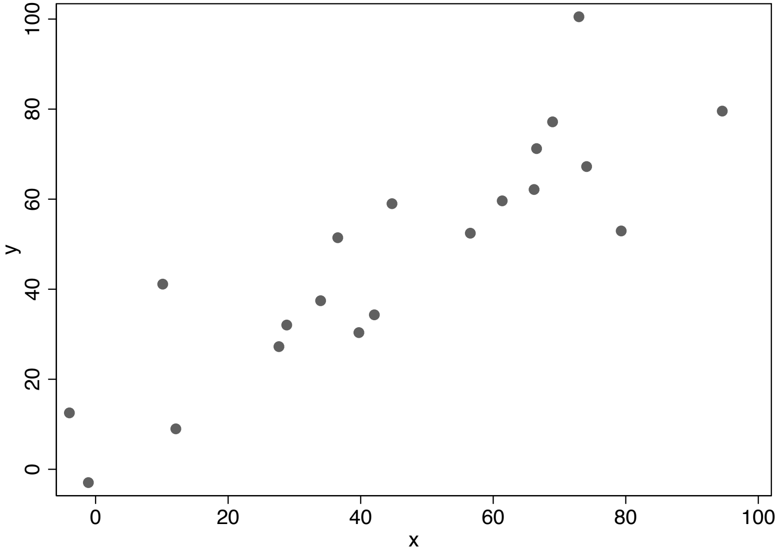
$$y = a + bx$$

- When we use linear regression, we make the assumption that the relationship between x and y is linear.
- Linear functions have two basic parameters:

  Intercept (a): The point at which the line drawn by the function crosses the y axis. It is the value of y when x=0.

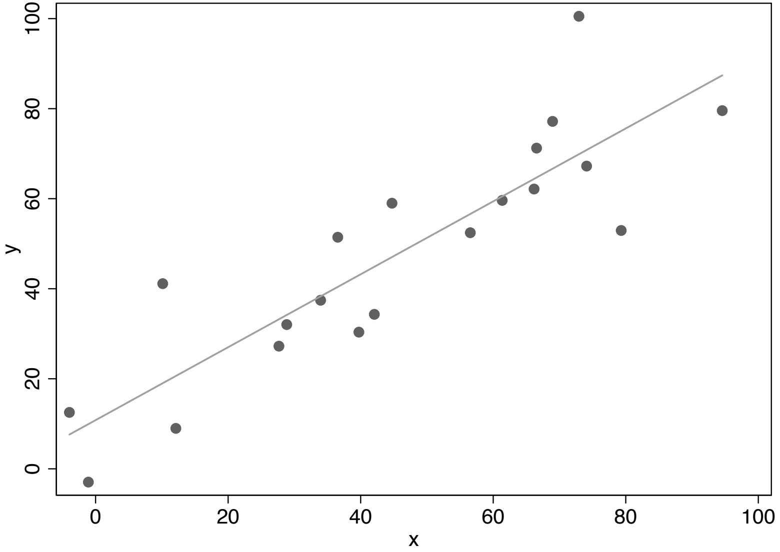
  Slope (b): The amount by which y changes for every one-unit increase in x. When x increases by 1, y changes by b.

Find the Best Linear Fit



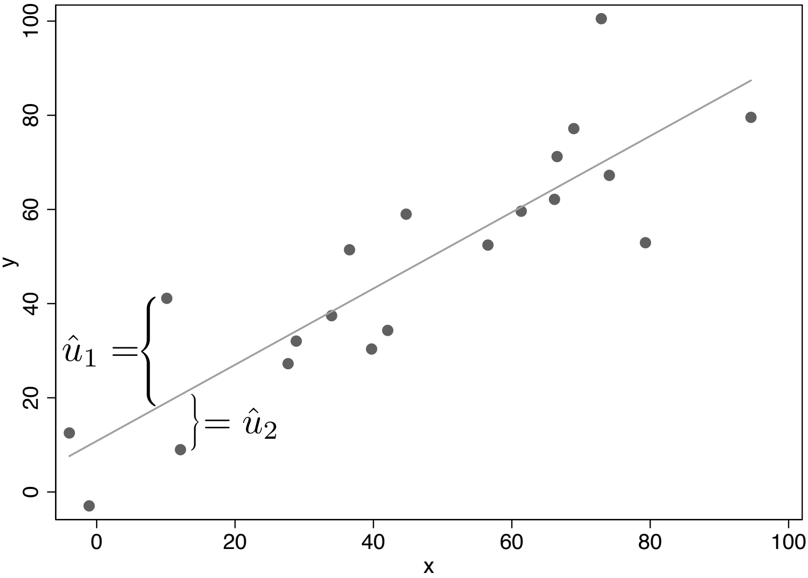
$$\hat{y}_i = 10.81 + .81x_i$$

Note: $\hat{y}_i$ is the predicted value of y_i given x_i. It is on the line.



$$y_i = 10.81 + .81x_i + \hat{u}_i$$

The difference between y_i and $\hat{y}_i$ is given by $\hat{u}_i$. It represents prediction error.



A Few Key Points

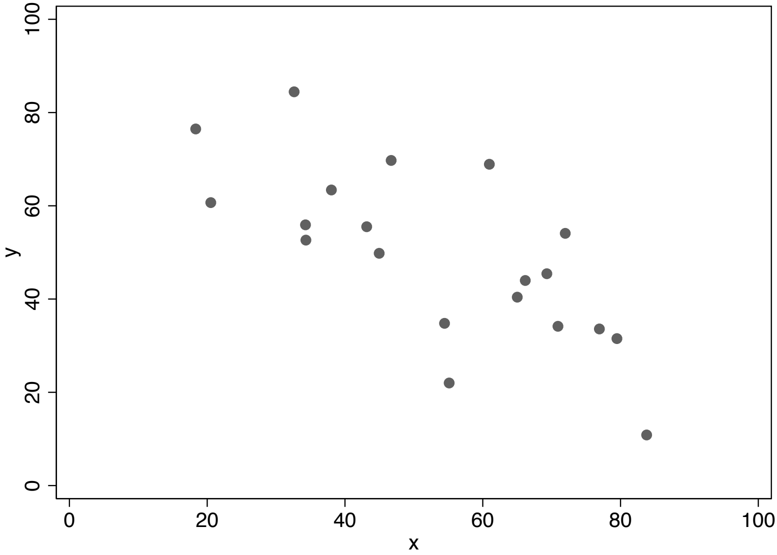
$$\hat{y}_i = 10.81 + .81x_i$$
$$y_i = 10.81 + .81x_i + \hat{u}_i$$

- The best-fit line is $\hat{y}$, the estimated or predicted value of y for a given level of x. We plug in a value for x and solve.
- When $x = 0$, y is predicted to be 10.81.
- For each one-unit increase in x, the estimated value of y rises by .81.
- The prediction error ($\hat{u}_i$) is also called a residual. It is the vertical distance between y and $\hat{y}$.

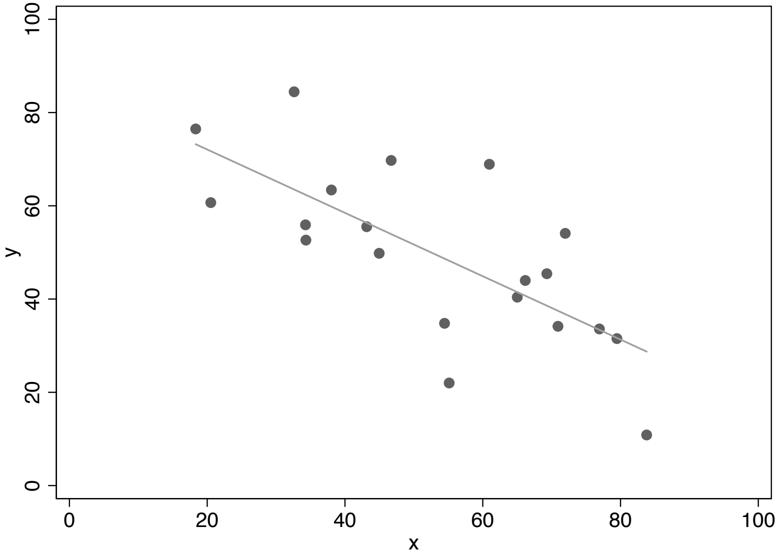
Some Additional Points

- When b is positive, y tends to increase as x increases.
- When b is negative, y tends to decrease as x increases.
- When b is 0, there is no linear relationship between x and y.
- b is measured in the same units as y. It provides the magnitude by which y is expected to change as x changes.

Find the Best Linear Fit



$$\hat{y}_i = 85.66 - .68x_i$$



Outline

General Approach of Linear Estimation

- From a broader standpoint, we can think of the dependent variable y as being produced by a combination of systematic and stochastic (i.e. random) factors.
- With bivariate linear regression, we model this systematic part of y as a linear function of x.
- This model may not be correct. The true relationship may not be linear, or x may not have any relationship with y.
- We estimate the regression to examine the fit of the model to the data.

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.

β_0 (beta naught) is the intercept of the linear function. It is called α some textbooks; just a different notation.

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

u_i is the part of y_i that is stochastic/random. It is unobserved and unknown.

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.

- $\hat{y}$ is determined mathematically by the $\hat{\beta}$'s and x. 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.

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

Estimating the β's

Regression involves finding the values of the β's that will minimize the sum of the squared deviations between y and $\hat{y}$.

- For each observation i out of n in our sample, the squared deviation between y_i and $\hat{y}_i$ is:

$$(y_i - \hat{y}_i)^2$$

- We can call these squared deviations "squared errors." The sum of the squared errors across all n items in the sample is thus:

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

- The best linear fit minimizes this sum.

Minimizing the Sum of Squared Errors

Since for any observation i:

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

It follows through substitution for $\hat{y}_i$ that:

$$(y_i - \hat{y}_i)^2 = (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Therefore, minimizing the sum of squared errors means finding the β's that minimize:

$$\mathsf{SSE} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

We do this with calculus! See Appendix 4.2 in Stock and Watson.

The OLS Estimator (Bivariate)

In bivariate regression, the formula for the slope ($\hat{\beta}_1$) is given by:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The formula for the intercept ($\hat{\beta}_0$) is:

$$\bar{y} - \hat{\beta}_1 \bar{x}$$

We won't use these formulas by hand, but there is value in understanding a bit of the math.

Interpreting the $\hat{\beta}_1$ Estimator

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The denominator is the variance of x:

- It is always positive, so the numerator determines the sign of $\hat{\beta}_1$.

The numerator is the covariance of x and y:

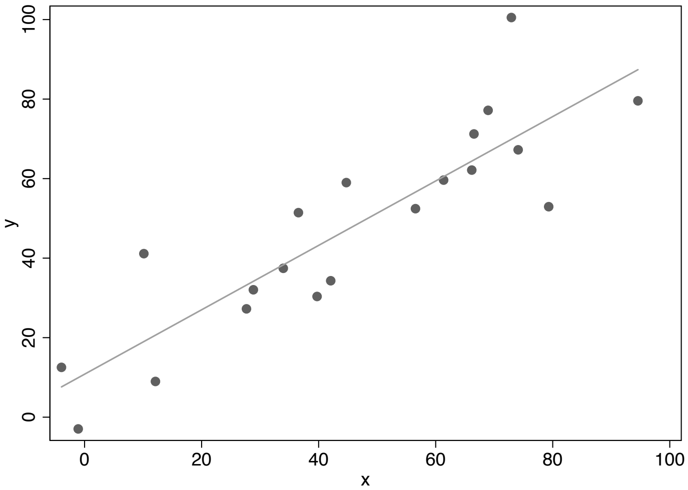
- If y tends to be above its mean when x is above its mean, then numerator is positive
- If y tends to be below its mean when x is above its mean, then the numerator is negative

Comparing Estimators for $\hat{\beta}_1$ and r

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y}_i)}{\sum(x_i - \bar{x})^2} \quad \text{vs.} \quad r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

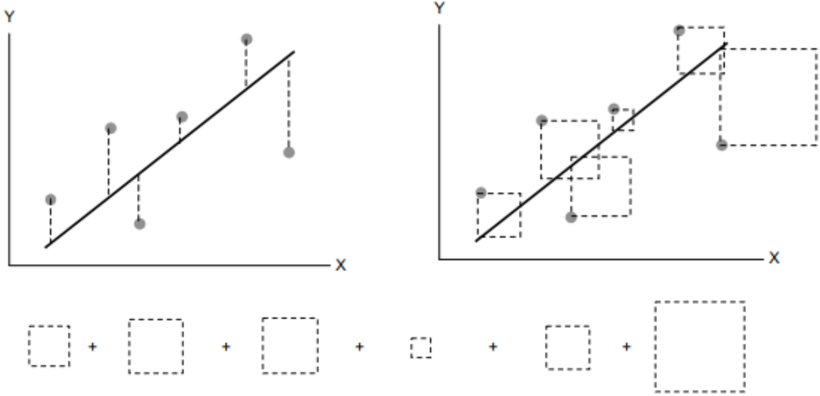
- The numerator is the same in both formulas.
- The correlation coefficient is unit-less and ranges from -1 to $+1$. It measures linearity and direction of relationship, but not slope.
- The regression coefficient has interpretable units: the change in y (in the units of y) for a one-unit change in x.

Visually



We find the intercept and slope that make the sum of the squared vertical distances between the points and the line as small as possible.

Sum of the Squared Errors



Visual representation of how much each point contributes to the sum of squared errors.

Outline

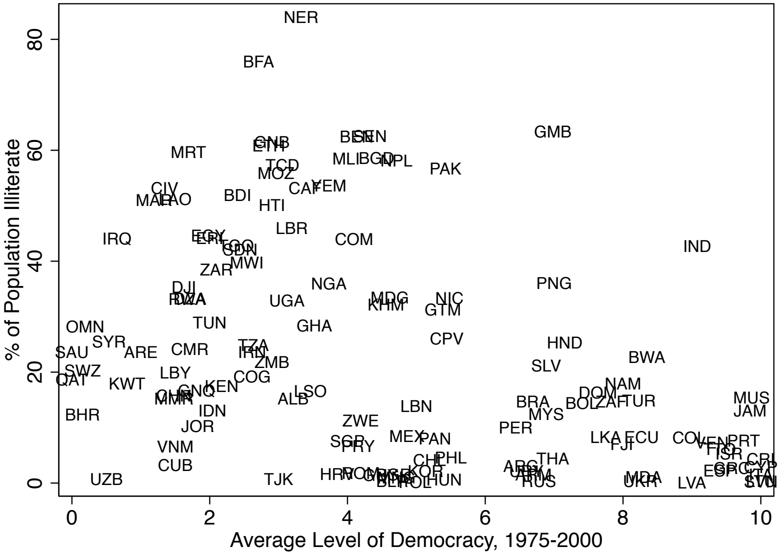
Working Example: Democracy and Illiteracy

- Across countries, there is variation in the rate of illiteracy.
- Some of this variation across cases is systematic; some of it is due to random (i.e. stochastic) factors.
- We hypothesize that a country's level of illiteracy is systematically related to its level of democracy.
- We could model the rate of illiteracy as a linear function of the level of democracy. Any remaining variation in illiteracy rates would be considered stochastic in this model.

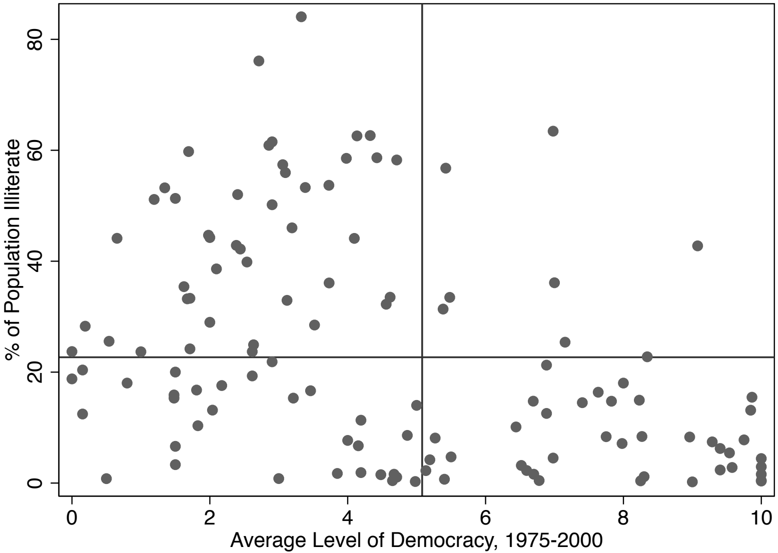
Measurement

- Illiteracy is measured in the year 2000 as the percentage of the population that cannot read or write. Data from the World Bank.
- Democracy is measured as mean score during the period 1975-2000 of the Polity index, rescaled to run from 0 - 10.
- Both variables can be treated as interval-level variables.
- We will regress `Illiteracy`, the dependent variable, on `Democracy`, the independent variable.

Democracy and Illiteracy

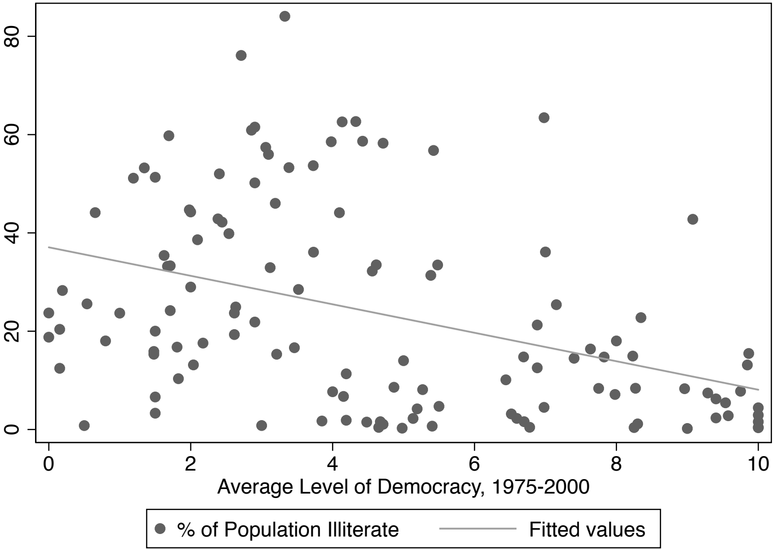


Pearson's $r = -.41$



Example: Estimated Regression Line ($\hat{y}$)

Illiteracy $=$ 37.1 - 2.9(Democracy)



Using Software

The Stata command for linear regression is regress, or just reg for short. This is followed by the name of the dependent variable and a list of independent variables.

```
reg depvar indvar

reg depvar indvar1 indvar2 indvar3
```

In R, we define a linear model (`lm`) and ask for a summary of the results:

```
mymodel <- lm(depvar ~ indvar1 + indvar2 +
indvar3, data=dataname)

summary(mymodel)
```

Example: Democracy and Illiteracy

$$\text{Illiteracy} = 37.1 - 2.9(\text{Democracy})$$

```
. 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
------------------------------------------------------------------------------
```

Note the Coef. column contains $\hat{\beta}_0$ (the intercept, labeled _cons) and $\hat{\beta}_1$ (the coefficient on Democracy).

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
```

Actual Data vs. Predictions

country	Democracy	Illiteracy	yhat	residual
Albania	3.211539	15.307	27.77394	-12.46694
Algeria	1.711538	33.301	32.1271	1.173901
Argentina	6.519231	3.167	18.17468	-15.00768
Armenia	6.7	1.585	17.65007	-16.06507
Bahrain	.1538462	12.443	36.64768	-24.20468
Bangladesh	4.423077	58.65	24.25793	34.39207
Belarus	4.65	.425	23.59938	-23.17438
Benin	4.134615	62.587	25.09508	37.49192
Bolivia	7.403846	14.489	15.60744	-1.118436
Botswana	8.346154	22.757	12.87276	9.884237
Brazil	6.692307	14.757	17.6724	-2.915395
Bulgaria	4.673077	1.584	23.53241	-21.94841
Burkina Faso	2.711539	76.102	29.22499	46.877
Burundi	2.403846	52.013	30.11795	21.89505
Cambodia	4.558824	32.235	23.86399	8.371017
Cameroon	1.711538	24.188	32.1271	-7.939098

Interpretation of Coefficients

$$\begin{aligned} \hat{y_i} &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\ \text{Illiteracy} &= 37.1 - 2.9(\text{Democracy}) \end{aligned}$$

- For each [one-unit] increase in `Democracy`, the predicted rate of `Illiteracy` is 2.9 [units] lower. Avoid the generic "unit."
- Since `Illiteracy` is measured as a percentage, that means it would decline by 2.9 percentage points.

  When `Democracy`=0, the predicted rate of `Illiteracy` is:

  $$37.1 - 2.9(0) = 37.1$$

  When `Democracy`=10, the predicted rate of `Illiteracy` is:

  $$37.1 - 2.9(10) = 8.1$$

Interpreting $\hat{\beta}_1$

- $\hat{\beta}_1$ is the average change in y associated with a 1-unit change in x.
- A one-point increase in the `Democracy` scale is associated on average with a 2.9 percentage point decrease in `Illiteracy`.
- On average, countries that are one point higher on `Democracy` have illiteracy that is 2.9 percentage points lower.

$$\frac{\Delta\text{Illiteracy}}{\Delta\text{Democracy}} = -2.9$$

Cautions on Interpreting $\hat{\beta}_1$

$\hat{\beta}_1$ is the average change in y associated with a 1-unit change in x.

- This statement is agnostic about causation: "associated with" is not "caused by."
- OLS is just a statistical calculation, just like calculating a mean.
- We still need to think through tools of causal inference:
  - research design
  - sample selection issues
  - identification strategy, etc.

Interpreting the Intercept $\hat{\beta}_0$

$$\begin{aligned} \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\ \text{Illiteracy} &= 37.1 - 2.9(\text{Democracy}) \end{aligned}$$

- Value of the y variable (Illiteracy) when the x variable (Democracy) is zero
- In our example: the OLS results tell us that for a district with Democracy of 0, we predict an illiteracy rate of 37.1 percentage points.
- Note: if a value of 0 in the x variable is not realistic, then this estimate is not directly meaningful.
- In most cases, $\hat{\beta}_0$ is not of primary interest

More Interpretation

$$\text{Illiteracy} = 37.1 - 2.9(\text{Democracy})$$

Suppose that Country A has a `Democracy` score of 3 and Country B has a `Democracy` score of 7. What is the difference in their predicted rates of illiteracy?

$$\begin{aligned} \text{Country A:} \quad 37.1 - 2.9(3) &= 28.4 \\ \text{Country B:} \quad 37.1 - 2.9(7) &= 16.8 \end{aligned}$$

This is a predicted difference of 11.6 percentage points.

More Interpretation

$$\text{Illiteracy} = 37.1 - 2.9(\text{Democracy})$$

Suppose `Democracy` were 3 points higher in a country. How much would its predicted rate of illiteracy change?

Key insight: with a linear prediction, it does not matter where you start. A 3-point increase has the same effect whether the initial level of `Democracy` is 0, 3, or 7.

Accordingly, if `Democracy` were 3 units higher, we would expect the predicted rate of illiteracy to change by $-2.9 \times 3 = -8.7$ points from its original value.

Marginal Effects

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

- A marginal effect is the instantaneous rate of change in $\hat{y}$ caused by changing one independent variable while holding the others constant.
- With bivariate regression, there are no other independent variables to hold constant, but the principle is the same.
- We can find it with calculus:

$$\frac{\partial \hat{y}}{\partial x} = \hat{\beta}_1$$

The marginal effect of x on $\hat{y}$ is $\hat{\beta}_1$. This makes sense!