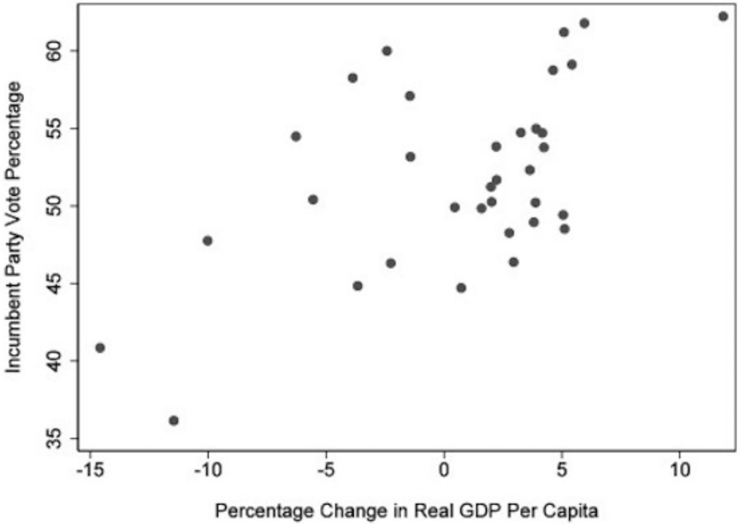
Simple Regression

POLI 205 Doing Research in Politics

Fall 2015

Scatterplot



Fitting a Line

- The basic idea of two-variable regression is that we are fitting the “best” line through a scatterplot of data
- This line, which is defined by its *slope* and *y-intercept*, serves as a statistical model of reality
  - $Y = mX + b$
- Where b is the y-intercept and m is the slope or “rise-over-run”
- For a one-unit increase (run) in X, m is the corresponding amount of rise in Y (or fall in Y, if m is negative)

Best Fitting Line

- The best fitting line *minimizes the sum of the squared residuals*
    - $\sum_{i=1}^{n} \epsilon_i^2$
- *Ordinary least-squares (OLS) regression*
- OLS is the best linear unbiased estimator (BLUE)

Regression Model

Population

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

- **Systematic component**: $\alpha + \beta X_i$
  - $\alpha =$ y-intercept parameter; constant
  - $\beta =$ slope parameter
- **Stochastic component**: ϵ_i
  - We do not expect all of our data points to line up perfectly on a straight line
  - Error term or residuals

Regression Model

Sample

$$Y = A + BX_i + E_i$$

- A represents the estimate of α
- B represents the estimate of β
- E represents the estimate of ϵ
    - Can also be written: $E_i = Y_i - \hat{Y}_i$

Estimating α and β

$$B = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$A = \bar{Y} - \hat{\beta}\bar{X}$$

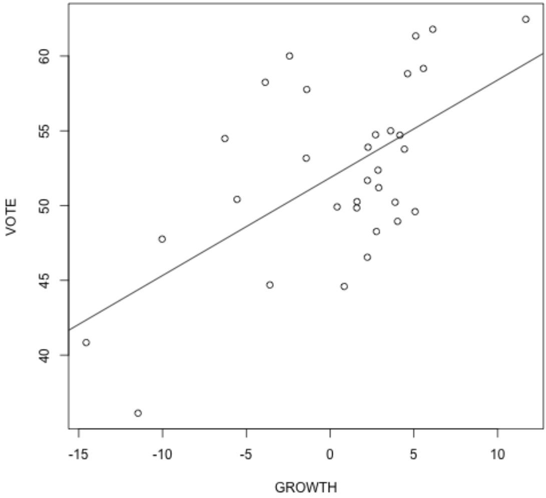
Example: GDP Growth and Presidential Vote

```
Coefficients:
Estimate Std.  Error t value Pr(>|t|)
(Intercept) 51.860 0.882 58.82 < 2e-16 ***
GROWTH 0.654 0.161 4.07 0.00032 ***
---
Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1
Residual standard error:  4.95 on 30 degrees of freedom
Multiple R-squared:  0.356, Adjusted R-squared:  0.334
F-statistic:  16.6 on 1 and 30 DF, p-value:  0.000316
```

Example: GDP Growth and Presidential Vote

- $Y = 51.86 + 0.654(\text{X})$
- Use to predict value of Y ($\hat{Y}$) for a given value of X
- With real GDP per capita growth of 2 (the rate in 2012) what would be the predicted presidential vote, $\hat{Y}$?
- $Y = 51.86 + 0.654(2) = 53.168$

Example: GDP Growth and Presidential Vote



Uncertainty

Goodness-of-Fit: Model

- How well does the regression model explain the variance of Y?
- Root Mean-Squared Error
  - The overall average "miss"
- `sqrt(deviance(ols1)/df.residual(ols1))`
- `## [1] 4.955`

Goodness-of-Fit: Model

- R^2: Ranges from 0 to 1 and indicates the *proportion of the variation in the dependent variable that is accounted by the model*
- `summary(ols1)$r.squared`
- `## [1] 0.3555`

Model Components: Standard Error

- σ^2: variance of the population stochastic component
  - The spread of observations around the regression
  - Estimated using the sum of squared E divided by $n - 2$
- Standard error of B
  - Square root of the variance of B
- Standard error of A
  - Square root of the variance of A

Hypothesis Testing with OLS

- We specify a null hypothesis and working hypothesis usually *about the slope parameter*
  - Null hypothesis is that the slope of $\beta = 0$
- Same logic as bivariate hypothesis testing
  - Observe a sample slope parameter, which is an estimate of the population slope
  - Evaluate how likely we are to observe the sample slope if the true (population) slope is *0*
  - If the probability is less than .05, then the estimate of β is said to be statistically significant
- Two-tailed vs. one-tailed test

Hypothesis Testing with OLS

- Null hypothesis $H_0 : \beta = 0$
- Working hypothesis $H_1 : \beta \neq 0$
- Directional hypothesis
  - $H_1 : \beta < 0$
  - $H_1 : \beta > 0$

t-test

- The statistical test for regression is the *t*-test
  - $t = \frac{B}{se(B)}$
- 0.654 / 0.161
- ```
  t <- coef(ols1)[2] / coef(summary(ols1))[2,2]
  t
  ```
- ```
   GROWTH
    4.068
  ```