Bivariate Relationships

September 19, 2022

Today

- Examining relationships among two variables
  - Remember—techniques differ for different types of variables
- Focus primarily on interval DVs:
  - Scatterplot (two interval variables)
  - Correlation (also two interval variables)
  - Ordinary least squares regression (interval DV)
- Brief look at two way tables (i.e. cross tabs)

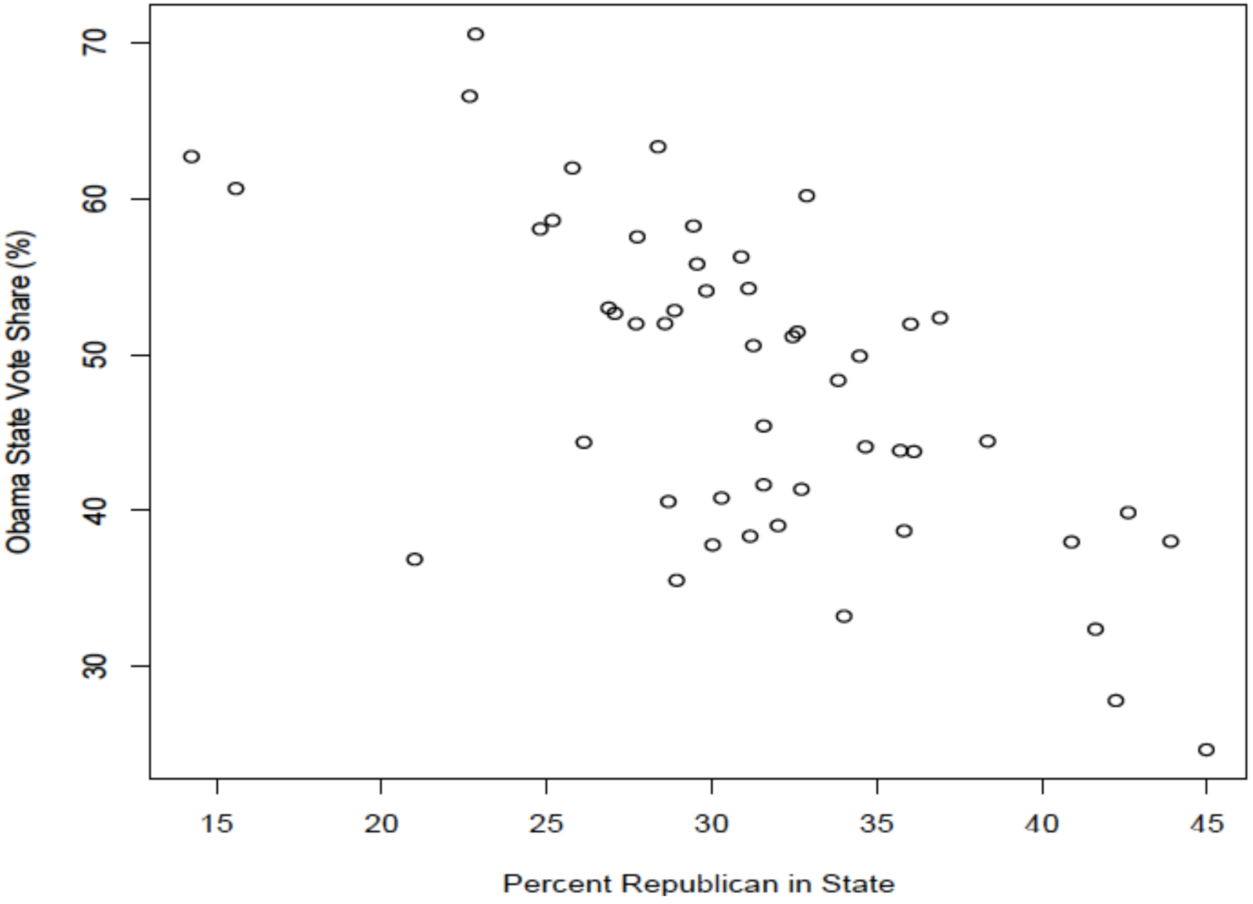
Basic Terminology

- "*Y*" variable-- Dependent Variable (DV), Explained Variable, Response Variable, Predicted Variable
- "*X*" variable-- Independent Variable, Explanatory Variable, Control Variable, Predictor Variable
- Variation is critical—a change in *Y* depends on a change in *X*
- Two possibilities:
  - Causal relationship—changes in *X* **cause** changes in *Y*
  - Correlation—changes in *X* **predict** changes in *Y*
    - Ex. SAT scores and performance in college
- Regression techniques show correlation, not causation
  - If testing causal arguments need to consider potential for sample selection, endogeneity, omitted/lurking variables, etc.
  - Very important element of research design

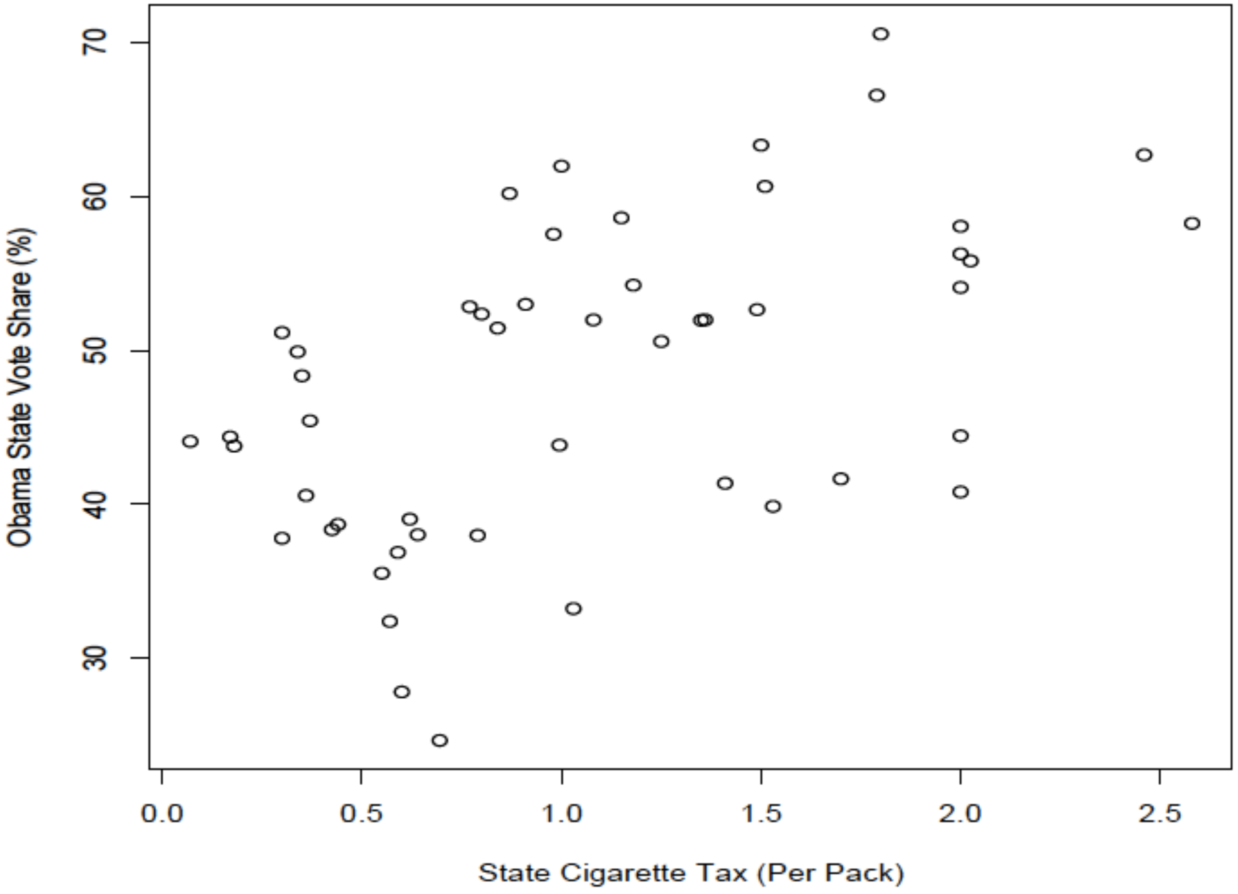
Scatterplots

- Basics:
  - Two interval variables
  - Plot values for IV, for each observation, on the X axis & values of DV, for each observation, on the Y axis
- What to look for:
  - Overall patterns
  - Deviations from overall patterns
  - Direction and strength of relationship
  - Outliers

Scatterplot Example # 1



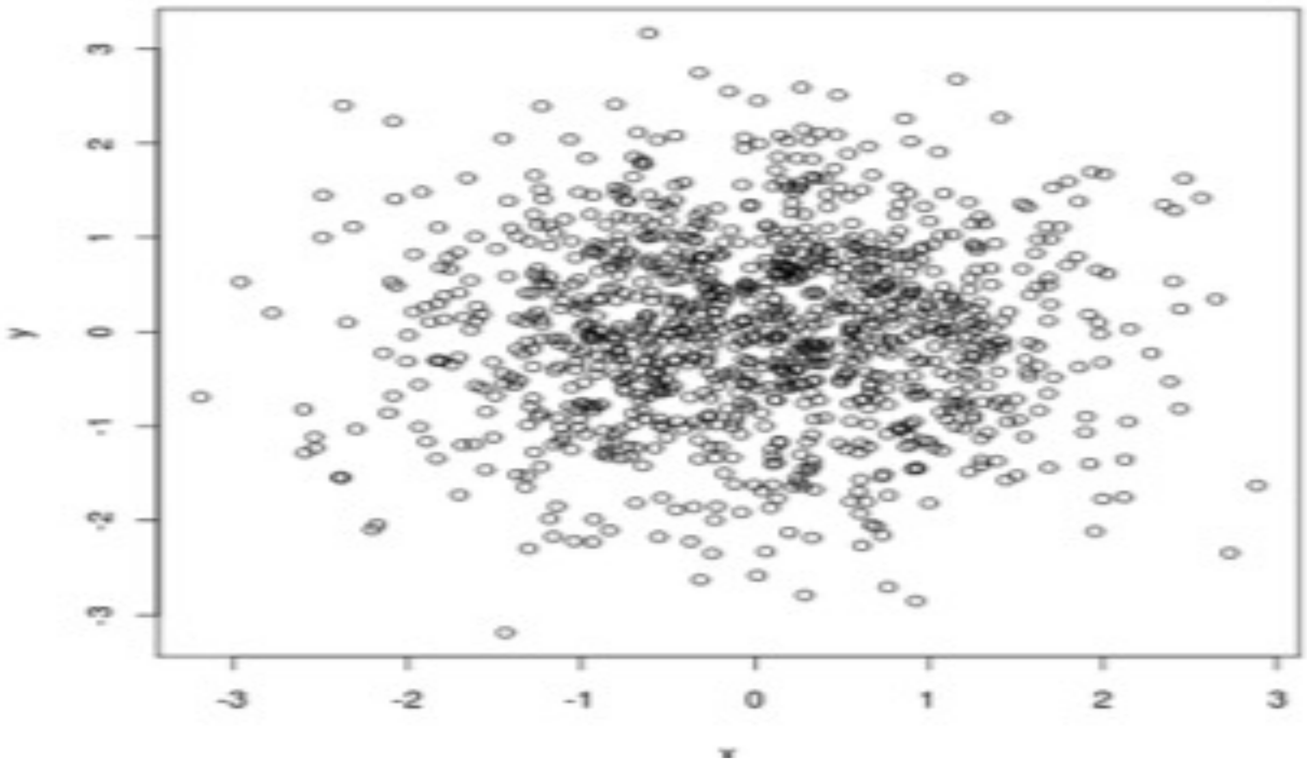
Scatterplot Example #2



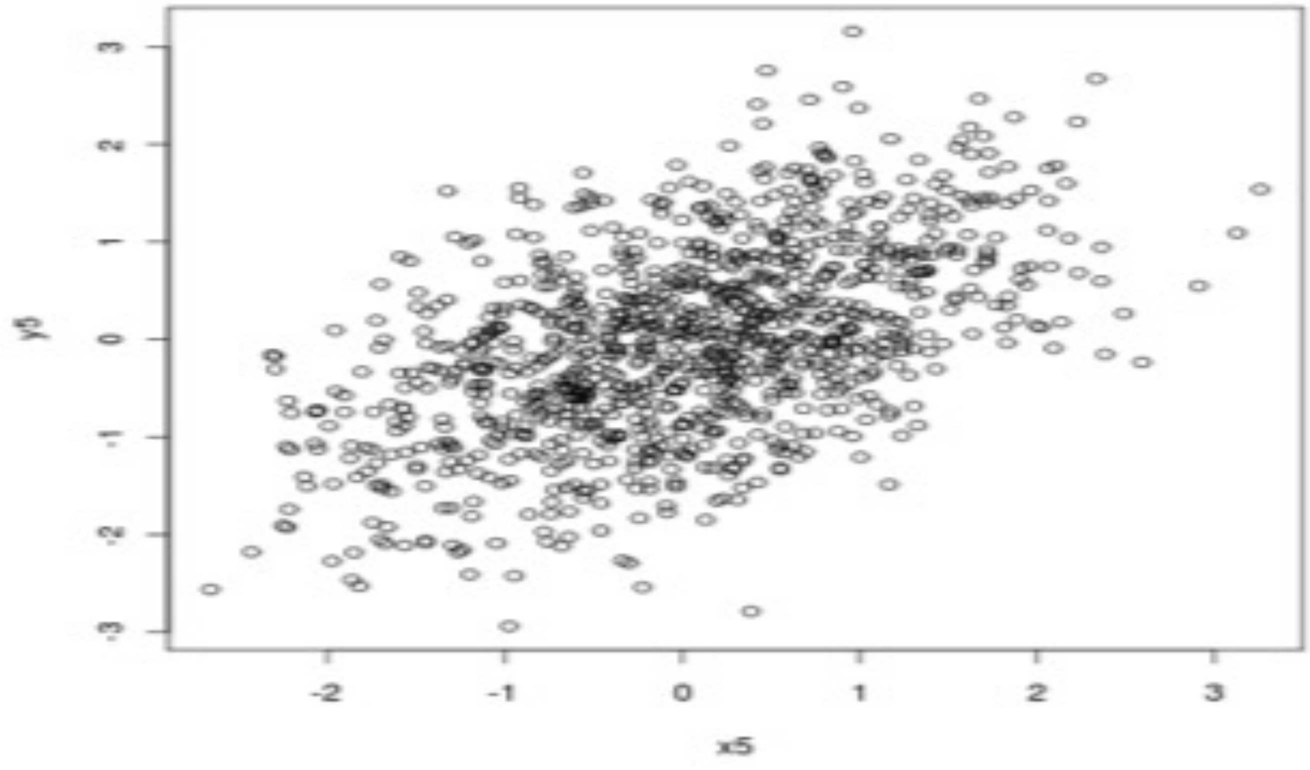
Correlation

- Scatterplots allow us to view the data & basic relationship
- Correlation tells us how closely variables relate to one another
  - Still only used with two interval/continuous variables
- Correlation coefficient indicates direction & strength
  - r= $\frac{\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)}{n-1}$
- Assumes a linear relationship, values bound between -1 and 1
- DV/IV distinction irrelevant with correlation coefficient

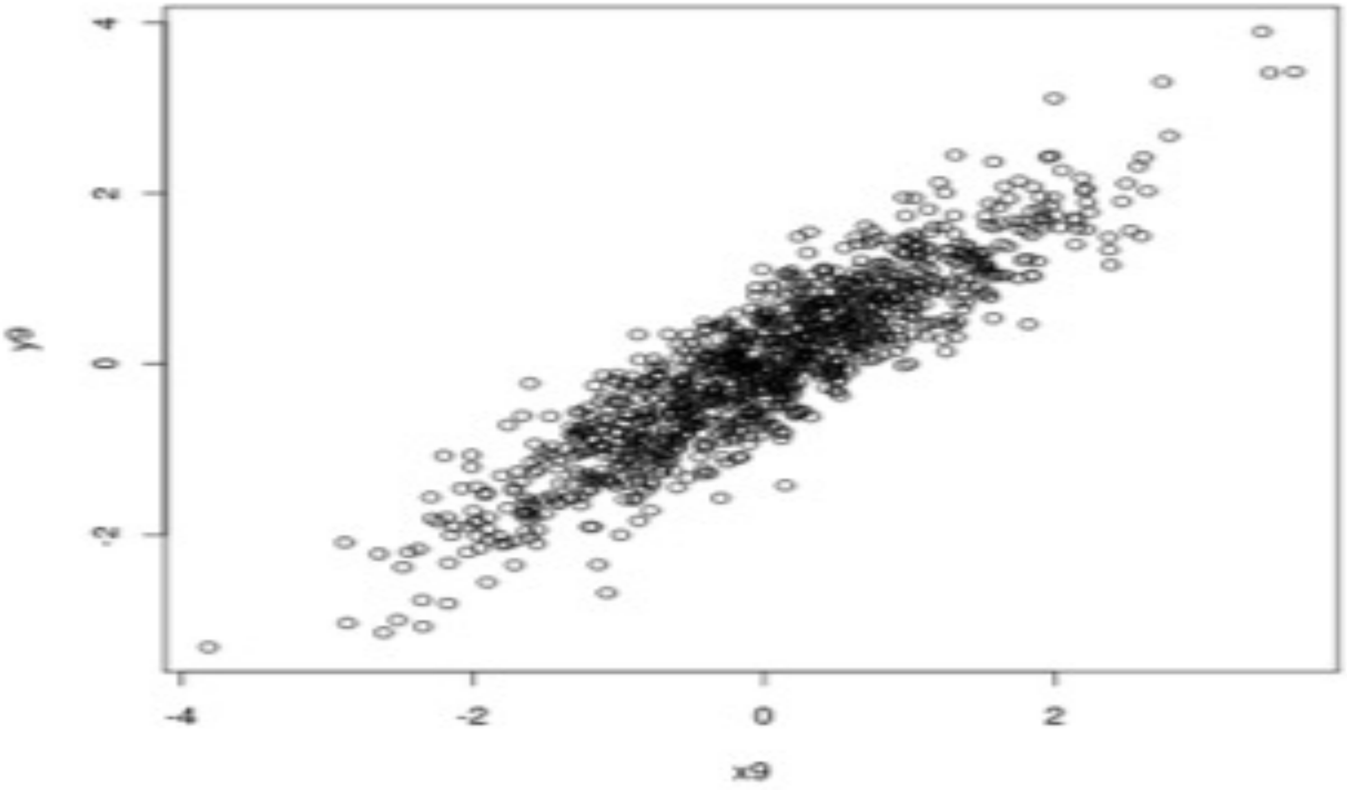
Zero Correlation



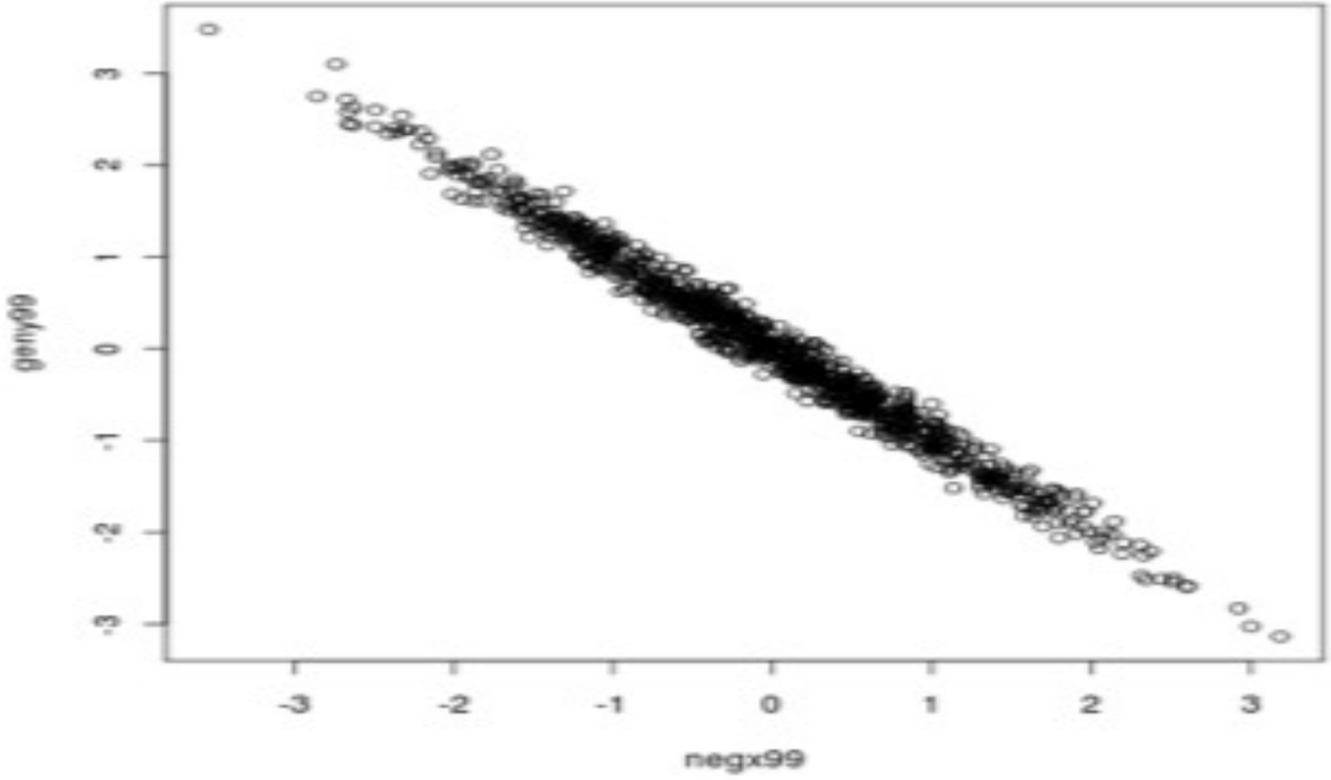
Positive Correlation



Strong Positive Correlation



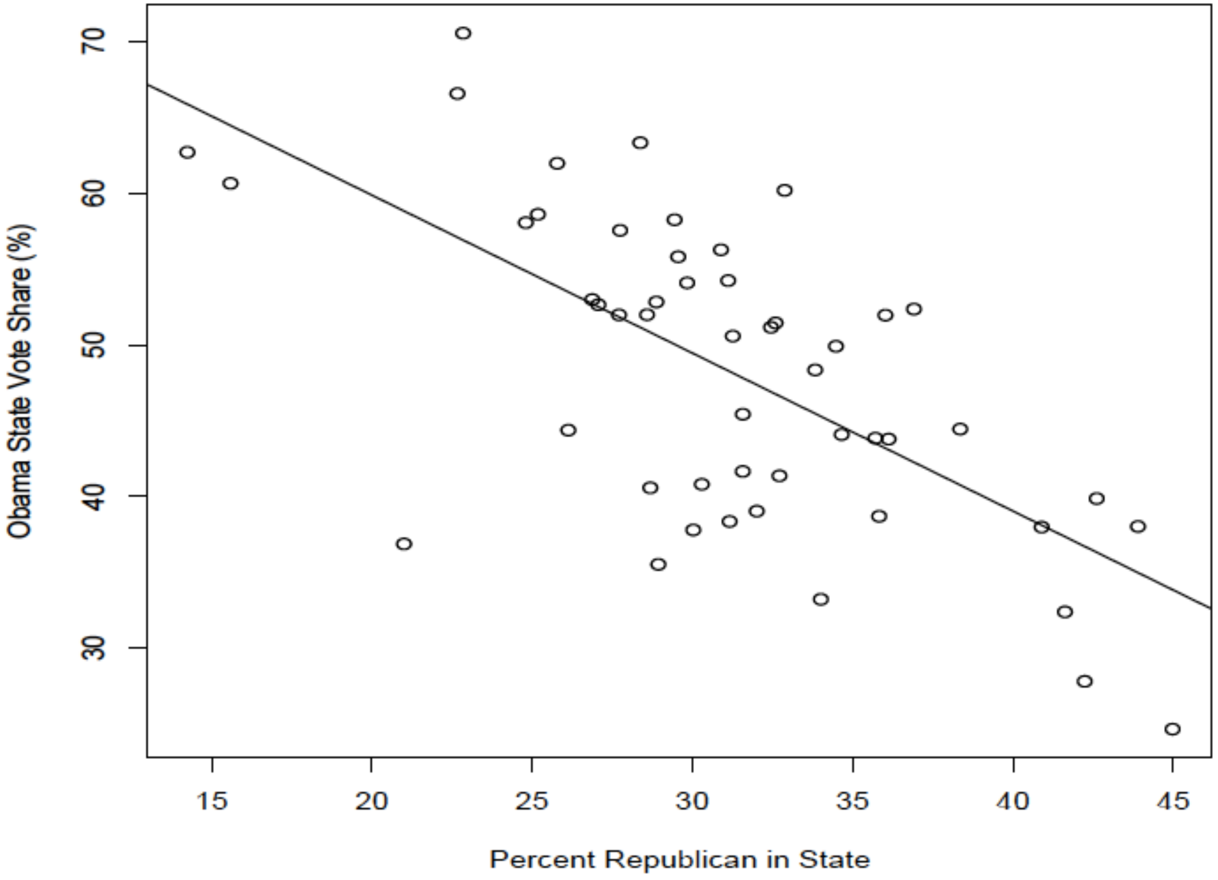
Stronger Negative Correlation



OLS Regression

- OLS Regression fits a "best-fit" line between points on a scatterplot
- The OLS (bivariate) equation: Y= $\beta_0 + \beta_1 X_1$
  - β_0=the intercept/constant estimate
    - The intersection point of the best-fit line with the y-axis (i.e. when x=0)
  - β_1=the slope coefficient
    - The magnitude of the impact of x on y, on average
    - Proper interpretation—**A one-unit change in x leads to an expected b_1-change in** y, **on average**
- Referred to as "ordinary least squares" because it minimizes the sum of squared "residuals" from the line
  - Residual—the difference between the predicted value (along the best-fit line) and the actual, observed, value

Scatterplot with Regression Line



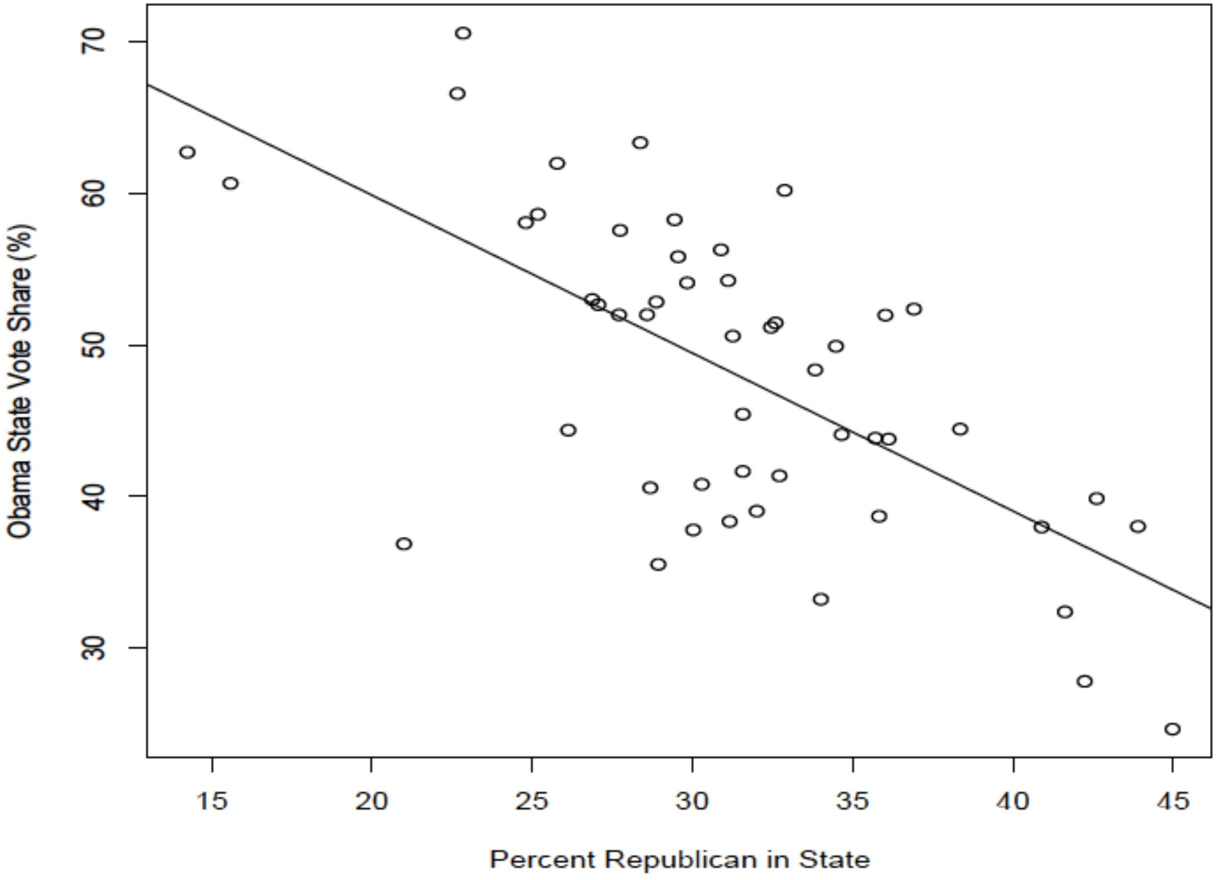
OLS Raw Output in R

```
Call:
lm(formula = Obama2012 ~ reppct_m, data = states)

Residuals:
     Min       1Q   Median       3Q      Max 
-21.9484  -5.1864   0.8801   5.0077  13.7267 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  80.7026     5.5661  14.499  < 2e-16 ***
reppct_m     -1.0415     0.1745  -5.968 2.81e-07 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.875 on 48 degrees of freedom
Multiple R-squared:  0.426,	Adjusted R-squared:  0.414 
F-statistic: 35.62 on 1 and 48 DF,  p-value: 2.805e-07
```

Predicted Values

- Can use OLS to generate predicted values
  - $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i}$
- Example
  - y=80.7026-1.0415(X_1)
  - Maryland has an observed value of 25.78% Republican identifiers in the data
  - What is the predicted value?
- Can compare these predicted values to observed values
  - Obama actually received 61.97% of the votes in Maryland in 2012

Scatterplot with Regression Line



OLS Error

- The **residual** is the difference between the predicted value and the observed value
  - Residual=$Y_i - \hat{Y}_i$
- What is the residual for Maryland?
- Since OLS regression generates the best-fit line, the residuals sum to zero
- Can measure the sum of squared residuals
  - Offers a measure of the total error in our OLS model
  - $RSS = \sum(y_i - \hat{y}_i)^2$

What does the regression equation come from?

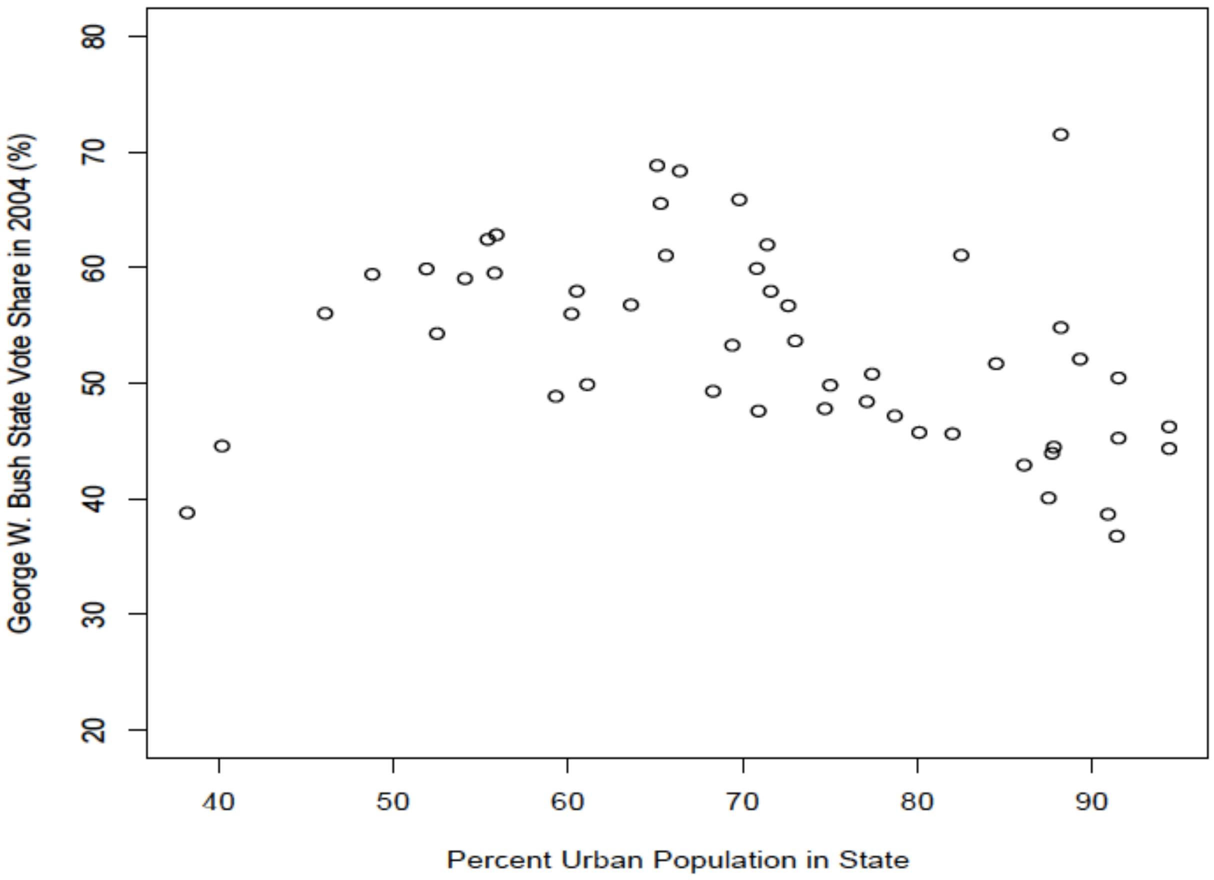
- The regression slope (β_1) is the correlation coefficient between X and Y, multiplied by the s.d. of Y divided by the s.d. of X
  - $\beta_1 = (r_{yx})(\frac{s_y}{s_x})$
  - $\beta_1 = (\frac{\sum(\frac{x_i-\overline{x}}{s_x})(\frac{y_i-\overline{y}}{s_y})}{n-1})(\frac{s_y}{s_x}) = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2}$
  - Intuition on the slope:
    - How much do the variables go together (i.e., covariance) divided by how much does the independent variable vary (i.e., variance)

- The intercept (β_0) is the mean of y minus β_1 times the mean of x
  - $\beta_0 = \overline{y} - \beta_1(\overline{x}_1)$

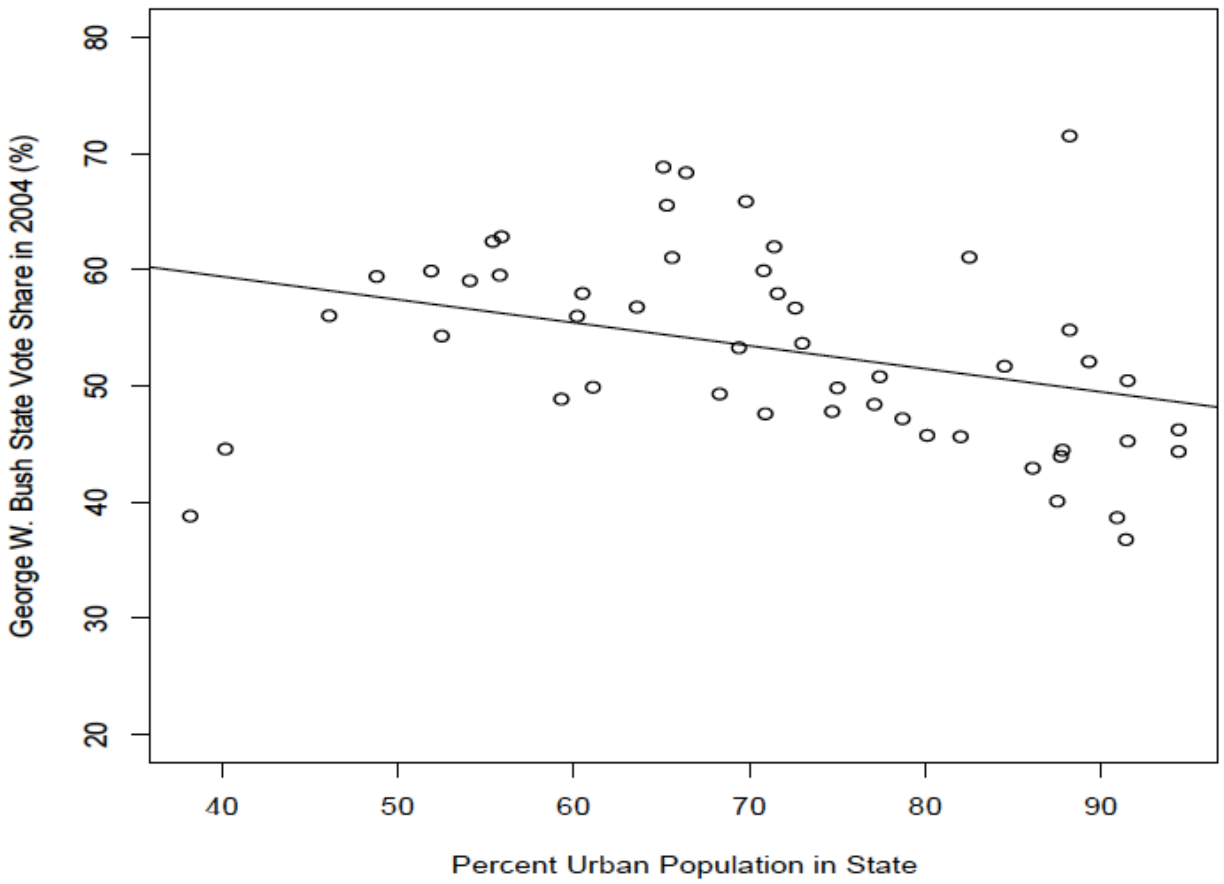
Things to Keep in Mind

- 1. OLS Regression is linear
  - Fits line that minimizes squared prediction error
  - The predicted magnitude of the impact of *X* on *Y* is constant across the range of x—i.e., one slope coefficient
  - True relationship between *X* and *Y* may not be linear
  - Very important to begin by plotting the relationship
    - Can often see on scatterplot whether the relationship appears linear in the data
- If relationship not linear, several possibilities:
  - Curvilinear—a sign/slope shift (or reversal), i.e., the effect of *X* on *Y* changes direction at different levels of *X*
  - Diminishing Returns—slope stays in same direction, but effect of one-unit change in *X* decreases (or increases) as values of *X* increase

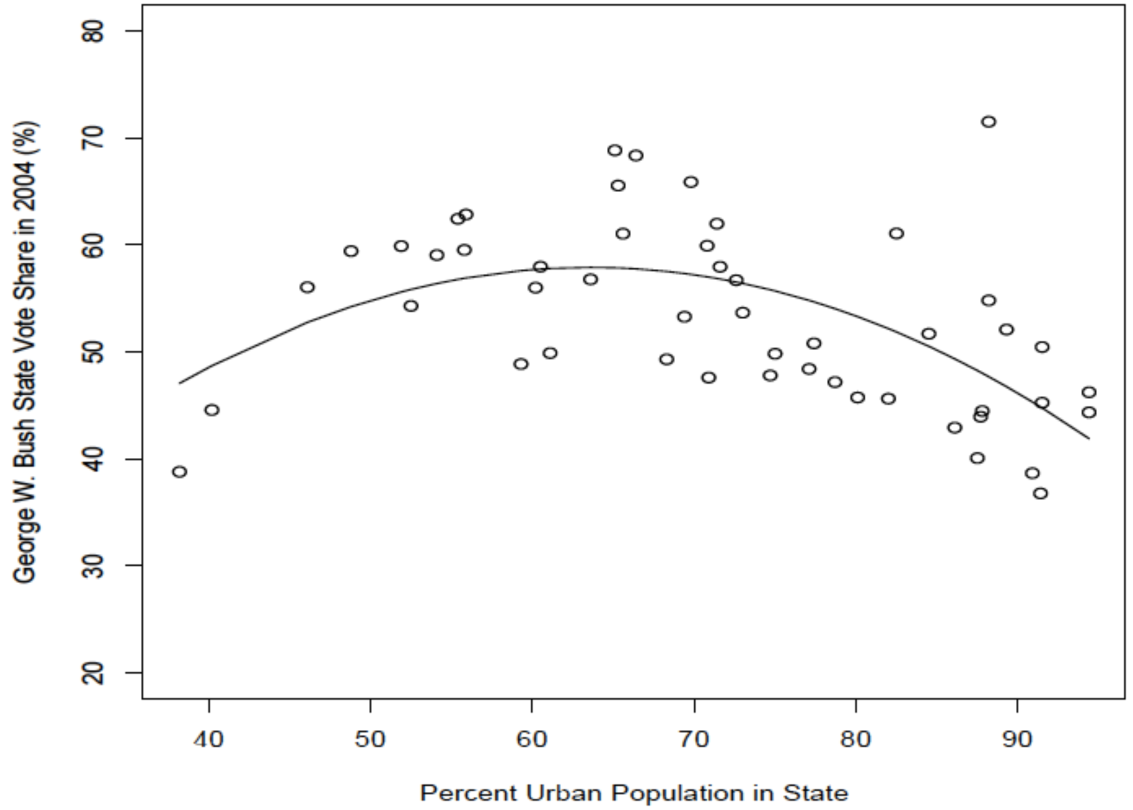
Linear Relationship?



Linear Relationship?



Curvilinear Relationship



Things to Keep in Mind

- 2. OLS always models a "best fit" line, need to judge how good of a fit
- Common measure of model fit: R^2
  - Percentage of variation in DV explained by the IV
  - "fraction of the variation in the values of Y that is explained by the least-squares regression of Y on X"
- Components:
  - Total Sum of Squares (TSS)—the sum of all of the squared differences of each observation (of Y) from the overall mean
  - Explained Sum of Squares (ESS)—the sum of the squares of the deviations of the predicted values from the mean value of Y
  - Residual Sum of Squares (RSS)—Difference between TSS and ESS
    - In essence, error not explained by model

Equation for R^2

- $R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \hat{y})^2}$
- $R^2 = \frac{ESS}{TSS} = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2}$

OLS Raw Output in R

```
Call:
lm(formula = Obama2012 ~ reppct_m, data = states)

Residuals:
     Min       1Q   Median       3Q      Max
-21.9484  -5.1864   0.8801   5.0077  13.7267

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  80.7026     5.5661  14.499  < 2e-16 ***
reppct_m     -1.0415     0.1745  -5.968 2.81e-07 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.875 on 48 degrees of freedom
Multiple R-squared:  0.426,     Adjusted R-squared:  0.414
F-statistic: 35.62 on 1 and 48 DF,  p-value: 2.805e-07
```

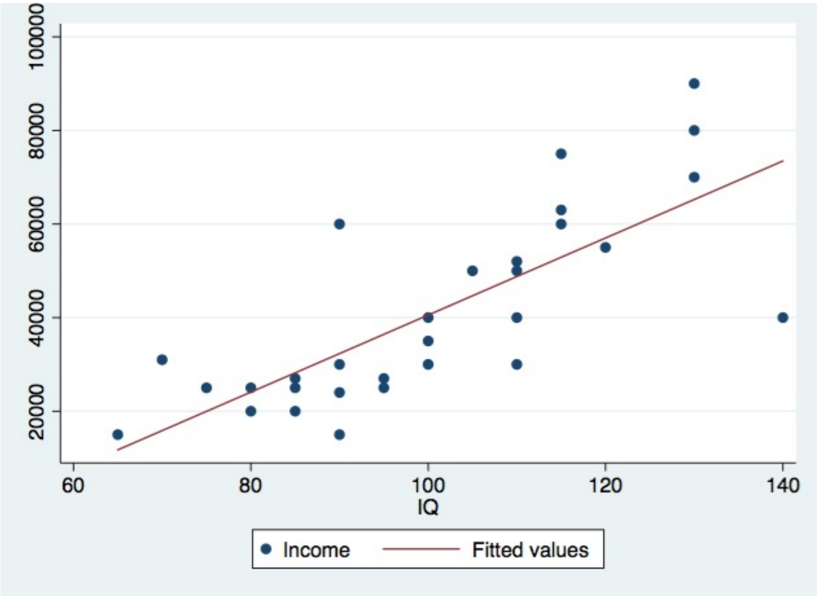
Sum of Squares in R

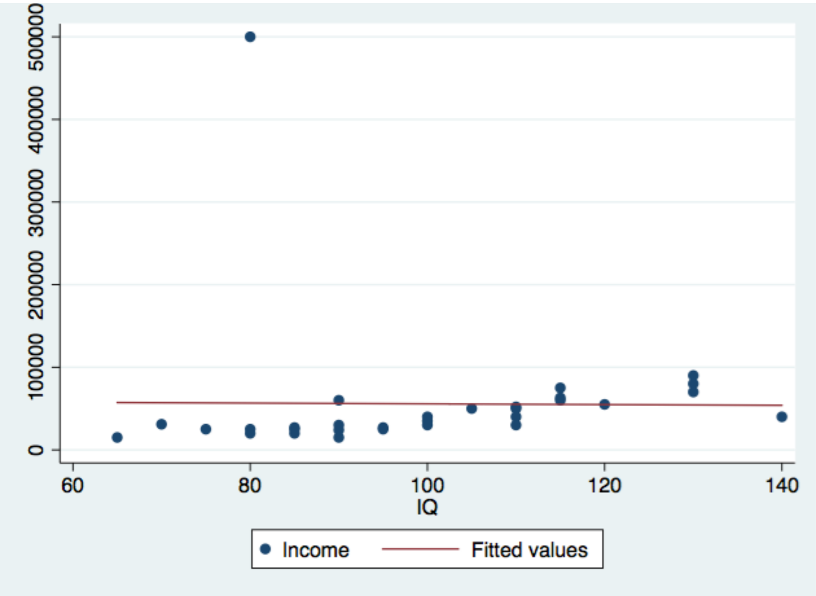
```
Analysis of Variance Table

Response: Obama2012
          Df Sum Sq Mean Sq F value    Pr(>F)
reppct_m   1 2209.1 2209.10  35.618 2.805e-07 ***
Residuals 48 2977.1   62.02
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Things to Keep in Mind

- 3. OLS can be heavily affected by outliers
  - Best fit line minimizes the distance from all points to the line
  - If one or more points are far out of the pattern, the slope of the line can change considerably

Outlier Example





Things to Keep in Mind

- 4. OLS allows for making unreasonable predictions:
  - Example—Prediction that a candidate will win 120% of vote
  - Or, prediction where there are no values of X
    - State with 70% of its population identifying as Republican
- We want to use OLS to generate reasonable predictions
- Evaluating predictions is key to assessing the relationship between variables and the strength of the model

Things to Keep in Mind

- 5. Correlation does not necessarily indicate causation
- OLS will assess the correlation between variables
- This correlation could be driven by lurking (unobserved) variable
  - Ex—Ice cream sales and drownings
- OLS is versatile—can include additional variables into model
  - Multiple Regression—will discuss this later in course (& GVPT 722)
- Need to think conceptually about "what else" could influence both the IV and the DV

Spurious Correlation

Ice Cream Consumption $\rightarrow$ Number of Drowning Deaths

Spurious Correlation

Ice Cream Consumption → Number of Drowning Deaths

Heat Index → Number of Drowning Deaths

Heat Index → Ice Cream

Things to Keep in Mind

- 6. Should (arguably) only use OLS Regression with an interval/continuous DV
  - Can use regression with other types of DVs, but relationship between *X* and *Y* then no longer linear
    - Ex-logit, probit, ordered logit/probit, multinomial logit/probit—see GVPT 729A (MLE) next Fall
- Can use OLS with dichotomous IV
  - Only two possible predicted values
- Cannot use OLS with multi-categorical (i.e. more., than two) IV (if it can't reasonably be treated as continuous)
  - Regression equation measures effect of 1 unit change in IV on DV
  - But, we can split categories into separate dichotomous variables

Regression Summary

- OLS very versatile
  - Models relationship between IV and DV
  - Allows for making predictions
  - Preview: Gauss-Markov & BLUE (GVPT 722)
- But, some limitations to keep in mind:
  - Gives best fit line, line may not be appropriate
  - Best fit line may not be good fit
  - OLS not resistant to outliers
  - Extrapolation likely to lead to incorrect predictions
  - Correlation potentially affected by lurking variables
  - Can only use with interval DVs & interval or dichotomous IVs

Multiple Regression

- Can incorporate additional variables into regression
  - Will discuss multiple regression more later
- Basic logic—fit a plane (or a hyperplane) through a 3- (or n-) dimensional scatterplot
- Interpreting coefficients similar:
  - Effect of a one-unit change in x on y, **holding constant other variables in the regression model**
- R^2 now indicates the percent of variation in DC explained by the entire model (i.e., all/multiple IVs)

Relationships among Categorical Variables

- Several techniques for evaluating relationships between categorical variables
- Graphical:
  - Side-by-side boxplots of different values of categorical variables
  - Line graphs for different sub-groups of data (ex—Democrats vs. Republicans)
- Tables:
  - Cross-tabulation—categorical IV/categorical DV
  - Mean-comparison table—categorical IV/interval DV

Raw Cross-Tabulation Output

When should abortion be permitted?	RECODE of partyid7 (Summary Party ID) Dem	Indep	Rep	Total
Never	52	64	36	152
	12.26	15.88	19.25	14.99
Some conds	115	110	65	290
	27.12	27.30	34.76	28.60
More conds	59	64	37	160
	13.92	15.88	19.79	15.78
Always	198	165	49	412
	46.70	40.94	26.20	40.63
Total	424	403	187	1,014
	100.00	100.00	100.00	100.00

Interpreting Cross-Tabs

- **Joint distribution—proportion of all observations in that cell**
  - Example—Proportion of respondents who are Democrats who say abortion should never be allowed
    - 52/1014=0.051
- **Marginal distribution—distribution of a single variable in a two-way table**
  - Example—Proportion of respondents who are Democrats
    - 424/1014=0.418
- **Conditional distribution—distribution of variable conditioned on value of one variable**
  - Example—Proportion of Independents saying abortion should never be permitted
    - 64/403=0.159

Mean-Comparison Table

RECODE of partyid7 (Summary Party ID)	Summary of Feeling thermometer: HILLARY CLINTON Mean	Std. Dev.	Freq.
Dem	78.695553	18.014381	877
Indep	61.442602	24.088741	784
Rep	41.24937	25.671601	397
Total	64.899417	26.069223	2058

Relationships among Categorical Variables

- Each of these share a common approach
  - Evaluate how values of DV differ across different categories of IV
  - For cross-tab—percentage of cases in different categories
  - For mean-comparison table—mean values of DV across categories of IV
- Also possible to incorporate control variables into cross-tab and mean-comparison table
  - Further divide into subcategories

Confidence in a Relationship

- One major omission here—confidence in the estimates of relationships
  - How confident should we be in the data and appearance of a relationship?
  - How confident should we be in our statistical results?
  - Related to R^2, but that doesn't tell us everything we need to know
- Interpretation of statistical significance in regression and other techniques will come later this semester
  - Again, the appropriate technique depends on the type of variable
  - Basic logic is the same
  - Knowledge of normal distribution and the empirical rule is key

Next Week

- Fundamentals of research design
- Design of study (data collection, etc.) influences the ability to evaluate hypotheses
- Basics:
  - Experiments vs. observational research
  - Sample vs. population