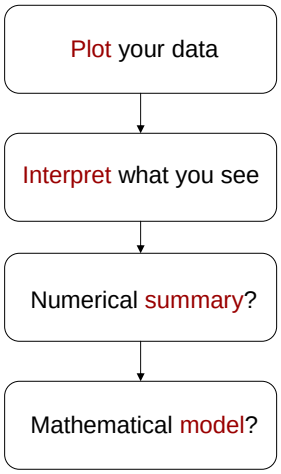
Correlation and Regression

Scatterplots
Correlation
Explanatory and response variables
Simple linear regression

General Principles of Data Analysis

- First plot the data, then add numerical summaries
- Look for overall patterns and deviations from those patterns
- When overall pattern is regular, use a compact mathematical model to describe it

General Principles of Data Analysis



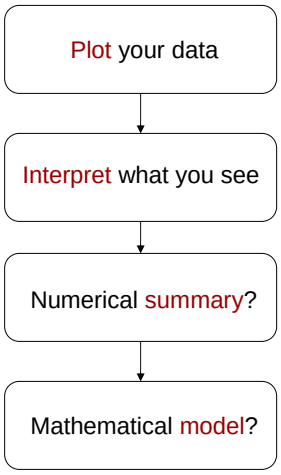
To understand the data, always start with a series of graphs

Look for overall pattern and deviations from that pattern

Choose an appropriate measure to describe the pattern and deviation

If the pattern is regular, summarize the data in a compact mathematical model

Univariate Data Analysis



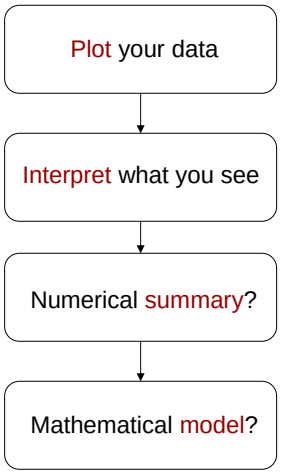
For continuous variable, use stemplot, histogram
For categorical variable, use bar chart, dot plot

Describe distribution in terms of shape, center, spread, outliers

Select measure of central tendency and spread: mean, standard deviation, five-number summary

If appropriate, use Normal distribution as model for overall pattern

Bivariate Data Analysis



For two quantitative variables, use a scatterplot

Describe the direction, form, and strength of the relationship

If pattern is roughly linear, summarize with correlation, means, and standard deviations

Regression gives a compact model of overall pattern, if relationship is roughly linear

Explanatory and Response Variables

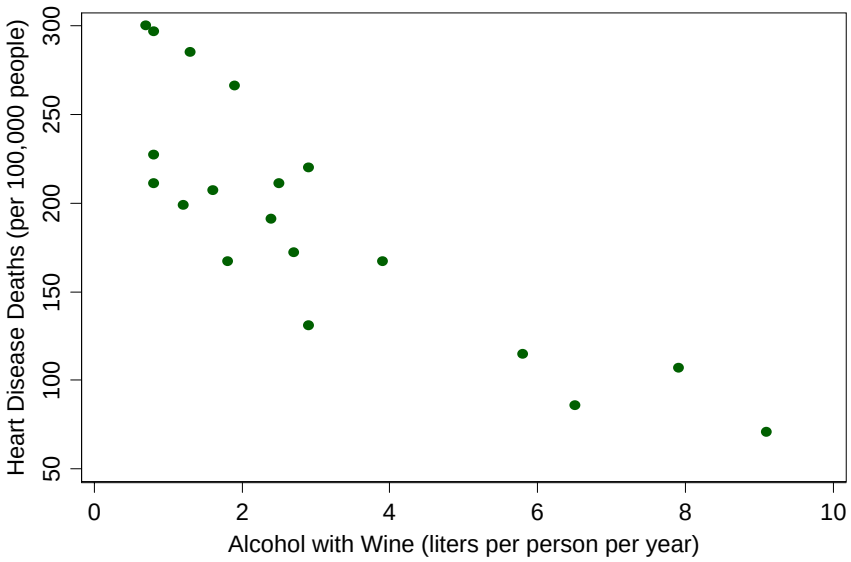
- Response variable measures outcome of a study
- Explanatory variable explains or influences change in response variable
- Response variables often called dependent variables, explanatory independent
- “Independent” and “dependent” have other meanings in statistics, so prefer to avoid
- But remember that calling one variable explanatory and other response doesn’t necessarily imply cause

Scatterplot

- Scatterplot shows relationship between two quantitative variables measured on same cases
- By convention, explanatory variable on horizontal axis, response on vertical axis
- Each case appears as point fixed by values of both variables

Wine Consumption and Heart Attacks

- Some evidence that drinking moderate amounts of wine my help prevent heart attacks
- We have the following data from 19 countries
  - Yearly wine consumption (liters of alcohol from wine, per person)
  - Yearly deaths (per 100,000 people) from heart disease
- In Stata, use `-twoway scatter-`
- Interpret a scatterplot
  - Overall pattern and striking deviations
  - Form, direction, and strength
  - Look for values that fall outside overall pattern of relationship (outlier)



Measuring Linear Association: Correlation

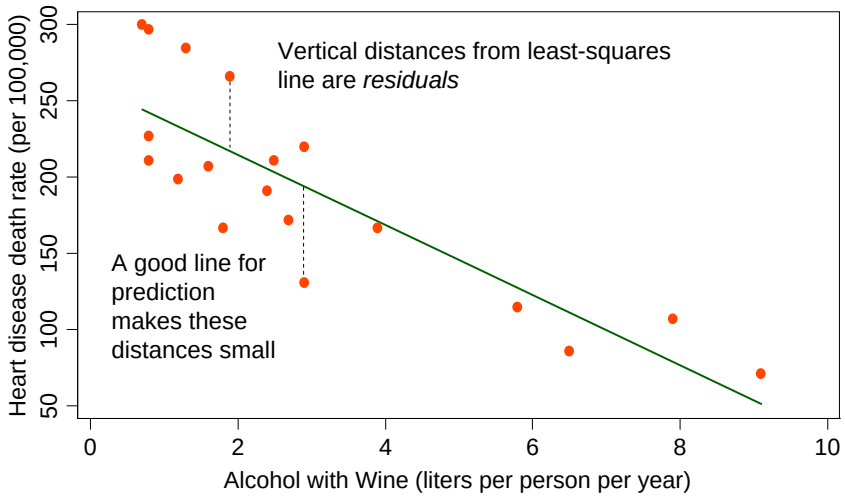
- Correlation measures direction and strength of linear relationship between two quantitative variables
- Usually written as *r*
- In Stata, use `-correlate-` or `-pwcorr-`
- Find and interpret correlation for wine and heart disease example

Using and Interpreting Correlation

- Ranges from -1 to 1; values closer to |1| indicate stronger linear relationship
- Positive values indicate positive association
- Does not distinguish between explanatory and response variables
- Requires that both variables be quantitative
- It has no unit of measurement – because it uses standardized values, it is scale free
- Measures strength only of linear relationships
- Like mean and standard deviation, strongly affected by outlying observations

Least-Squares Regression Line

- Correlation measures direction and strength of linear relationships
- A regression line summarizes relationship between explanatory, *x*, and response variable, *y*
- We can use regression line to predict value of *y* for a given value of *x*
- These predictions have error, called residuals
- The least-squares regression line of *y* is the line that minimizes residuals



In Stata, obtain this graph with `twoway (lfit heartdis alcohol) (scatter heartdis alcohol)`

Regression in Stata

```
regress heartdis alcohol

 Source |       SS       df       MS              Number of obs =      19
-------------+------------------------------           F(  1,    17) =   41.69
       Model |  59813.5718     1  59813.5718           Prob > F      =  0.0000
    Residual |  24391.3756    17   1434.7868           R-squared     =  0.7103
-------------+------------------------------           Adj R-squared =  0.6933
       Total |  84204.9474    18  4678.05263           Root MSE      =  37.879

------------------------------------------------------------------------------
    heartdis |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     alcohol |  -22.96877    3.55739    -6.46   0.000     -30.4742   -15.46333
       _cons |   260.5634   13.83536    18.83   0.000     231.3733    289.7534
------------------------------------------------------------------------------
```

slope, *b*

y-intercept, *a*

$$\hat{y} = a + bx$$

Estimated heart disease death rate = 260.56 + (-22.97)(Per capita alcohol consumption)

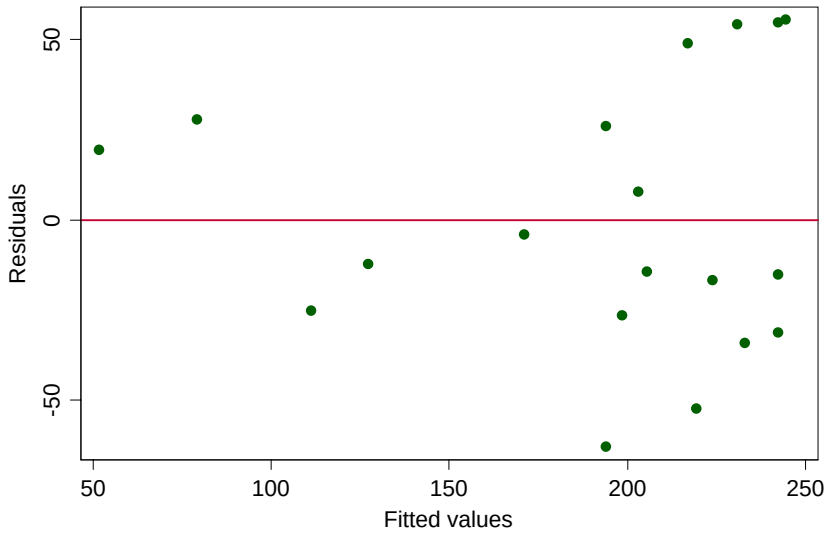
Using and Interpreting Regression

- Distinction between explanatory and response variables is essential in regression
  - Regression of *y* on *x* ≠ regression of *x* on *y*
- There is a close connection between correlation and slope of least-squares line
  - $b = r(s_y / s_x)$
  - Change of 1 sd in *x* corresponds to a change of *r* standard deviations in *y*
- Square of correlation, r^2, is fraction of variation in values of *y* explained by *x* (PRE)

Conditions for Inference

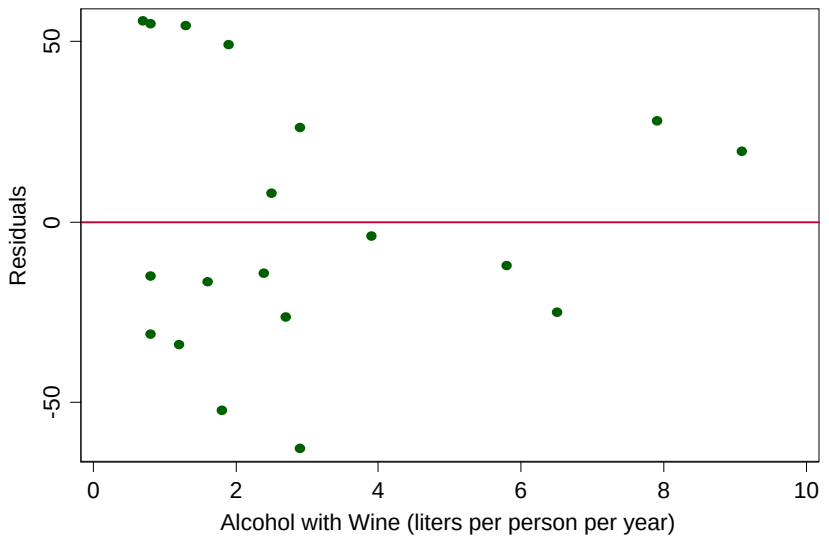
- Observations are independent
- True relationship is linear
- Standard deviation of response about the true line is same everywhere
- Response varies normally about true regression line
- Analysis of residuals is key to diagnosing violations of these conditions

Residuals-versus-Fitted Plot



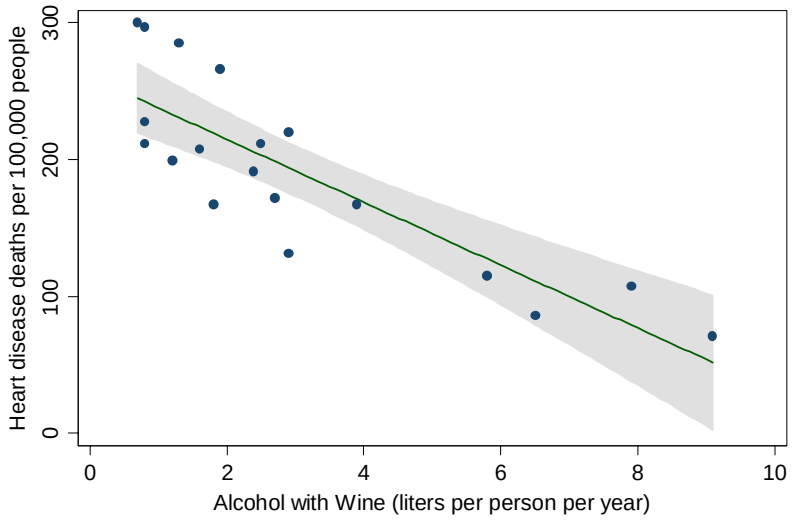
In Stata, obtain this plot after `regress` with `rvfplot, yline(0)`

Residuals-versus-Predictor Plot



In Stata, obtain this plot after `regress` with `rvpplot alcohol, yline(0)`

95% Confidence Interval for Least-Squares Line



Obtain this graph with `twoway (lfitci heartdis alcohol) (scatter heartdis alcohol)`