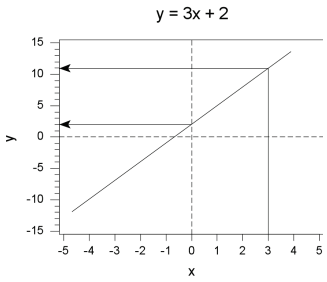
Chapters 8 and 9
Linear Regression and Regression Wisdom

Review - The equation of a straight line



- $y = mx + b$
- m is the slope — the change in y over the change in x — or rise over run.
- b is the $y-$intercept — the value where the line cuts the y axis.

Finding the Equation of a Line Using Two Points

- Finding the slope:

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

- Finding the y-intercept:

$$b = y - mx$$

Example: $(x, y) = (1, 3)$ and $(4, -3)$ are points on a line. Find the Equation of this line.

Example

If we find a linear association between two quantitative variables (perhaps by using a scatterplot), we can use this knowledge to better summarize the relationship.

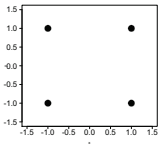
We can use a straight line, or regression line, to

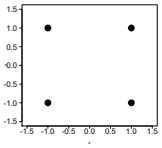
Notation:

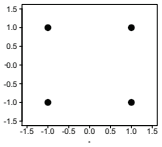
There are lots of straight lines that go through the data, however, not all choices provide a *good fit* to the data. We wish to choose the **line of best fit** (i.e. **least squares line**) for the data.

- We need to **minimize** the sum of squared residuals, i.e. we must minimize

$$\sum e_i^2 = \sum (y - \hat{y})^2$$







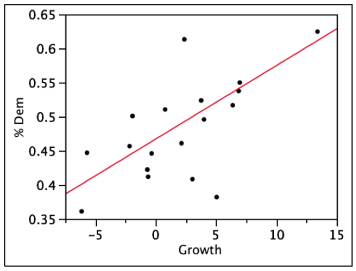
Regression line equation:

$$\hat{y} = b_0 + b_1 x$$

where

$$b_1 = r\frac{s_y}{s_x} \qquad b_0 = \bar{y} - b_1\bar{x}$$

We have information on the U.S. economy, which we wish to use to predict the proportion of the vote obtained by the Democratic candidate in presidential elections. More specifically, we have information on the growth rate of the U.S. economy (in %) and the Democratic share of the two-party vote (ignoring third parties). The data contain information from 1916 to 1988, however, we will exclude 1932 and 1972 as explainable anomalies (the depression and Watergate).



Fitting a linear regression model to this data produces the following formula

$$\hat{y} = 0.467 + 0.0107\ x$$

Provide interpretations of the regression coefficients for the example.

$$\hat{y} = \ b_0 \ + \ b_1\ x$$
$$\hat{y} = 0.467 + 0.0107\ x$$

- b_0 — y-intercept
  INTERPRETATION:

- b_1 — slope
  INTERPRETATION:

Given the least squares regression line, we can predict values of the response variable for specified values of the explanatory variable.

- What proportion of the two-party vote are the Democrats expected to receive in 1992, based on a economic growth rate of 3.4%?

- What proportion of the two-party vote are the Democrats expected to receive in 2008, based on a economic growth rate of 0.4%?

Residuals

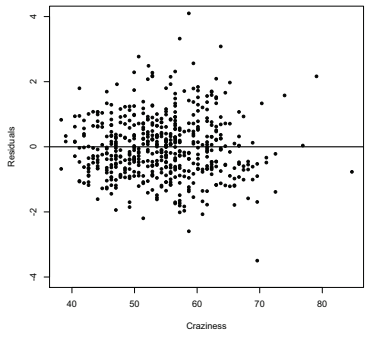
A residual is defined as the difference between an observed value and its predicted value, $\hat{y}$, based on the least squares regression line.

- Calculate the residual for the Democratic proportion of the two-party vote in 1992, if Clinton actually received 0.534 of the two-party vote.

- Calculate the residual for the Democratic proportion of the two-party vote in 2008, if Obama actually received 0.537 of the two-party vote.

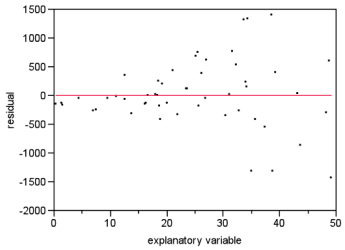
Residual Plot

A residual plot is a special scatterplot where the explanatory variable, x, is displayed on the horizontal axis and the residuals, e, are displayed on the vertical axis. Additionally, a horizontal line is drawn at $e = 0$.



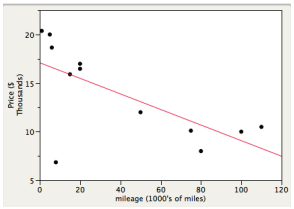
Examples of "suspect" residual plots

"Megaphone" Pattern: data are linear, but there is more variation in y for either larger or smaller values of x (i.e. there is increasing/decreasing spread of the response variable).

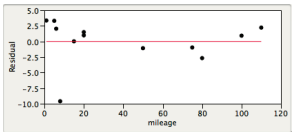


Curved Pattern:

Regressing the Price of a Camero on it's Mileage



Scatterplot



Residual Plot

Outliers:

Model Assessment - R^2

The goal of regression is to describe the variation in the response variable using the explanatory variable. We are able to determine *how much* of the variation is explained by the linear model and how much is unexplained.

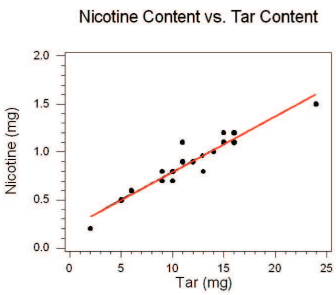
- r^2

- 1 - r^2

Formula: $\frac{s_{\hat{y}}}{s_y}$

Properties of R^2

Example: When we regressed the "Nicotine contents" of cigarettes on the "Tar contents" of the cigarettes, How well did the model do explaining the variation in the data?
$r = 0.956$



We cannot simply use least squares regression without making sure that certain assumptions and conditions are met.

- **Linearity Assumption:**

- **Equal Variance Assumption:**

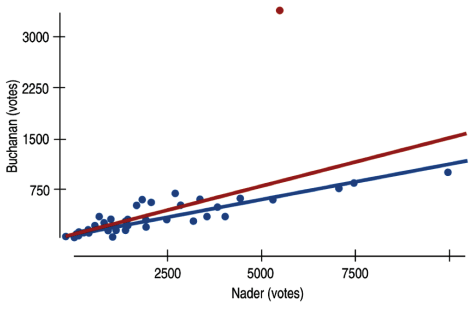
- **Outlier Condition:**

Regression Wisdom

Extrapolation: We fit a linear model to the data in hand, however, **we cannot assume that a linear relationship in the data exists beyond the range of the observed data**.

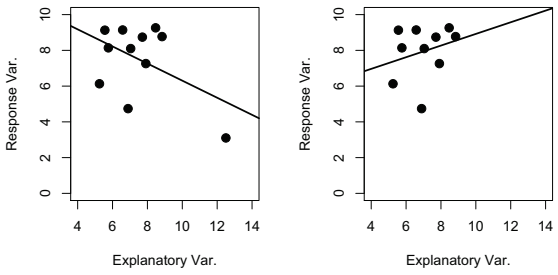
Extrapolation:

Outliers:
Outlying points can strongly influence a regression. They can greatly influence effect the intercept and slope of the least squares line. That is, omitting an outlier from the analysis can sometimes result in a very different model.

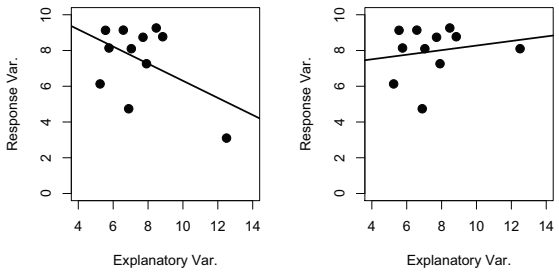


Large Residual:

Influencial Observations:



High Leverage Data Values:



Correlation $\neq$ Causation

WARNING: There is no way to conclude from regression alone that one variable *causes* the other.

- There is always the possibility that some third variable is driving both observed variables (i.e. there is a **lurking variable**).