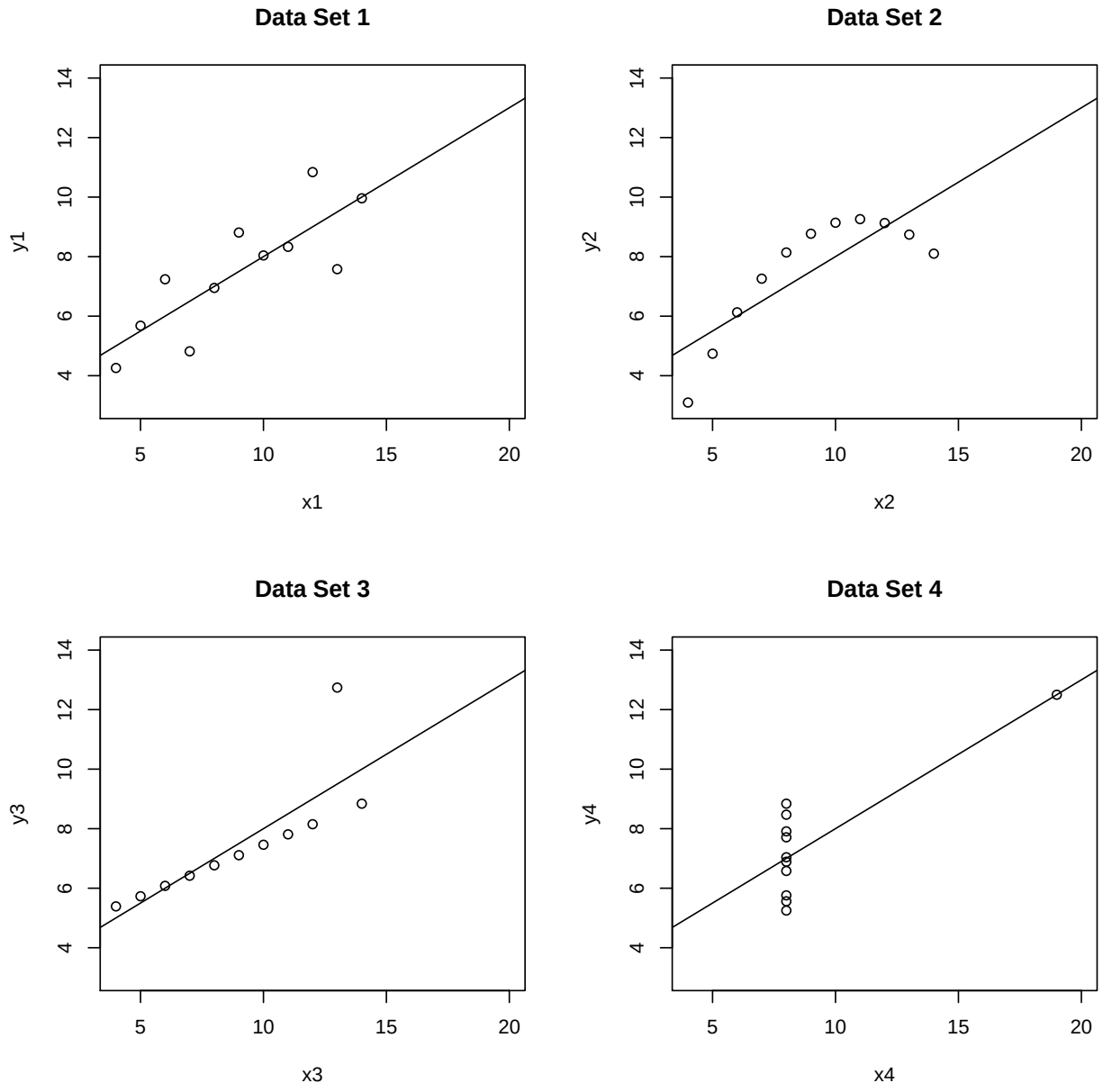
Lecture 4:
Diagnostics for Simple Linear Regression
STAT 632, Spring 2020

Anscombe's Four Data Sets



Anscombe's Four Data Sets

- All four data sets have the same least squares regression line, $\hat{y} = 3.0 + 0.5x$, and the same $R^2 = 0.67$.
- For which data set(s) is fitting a simple linear regression model reasonable?

Regression Diagnostics

- Assumptions of SLR model
  1. Linearity:
  2. Independence:
  3. Constant Variance:
  4. Normality:
- Regression diagnostics are a set of techniques (both numerical and graphical) that can be used to check the validity of these assumptions.
- Regression diagnostics often suggest improvements, which means model building is an iterative and interactive process.

Regression Diagnostics

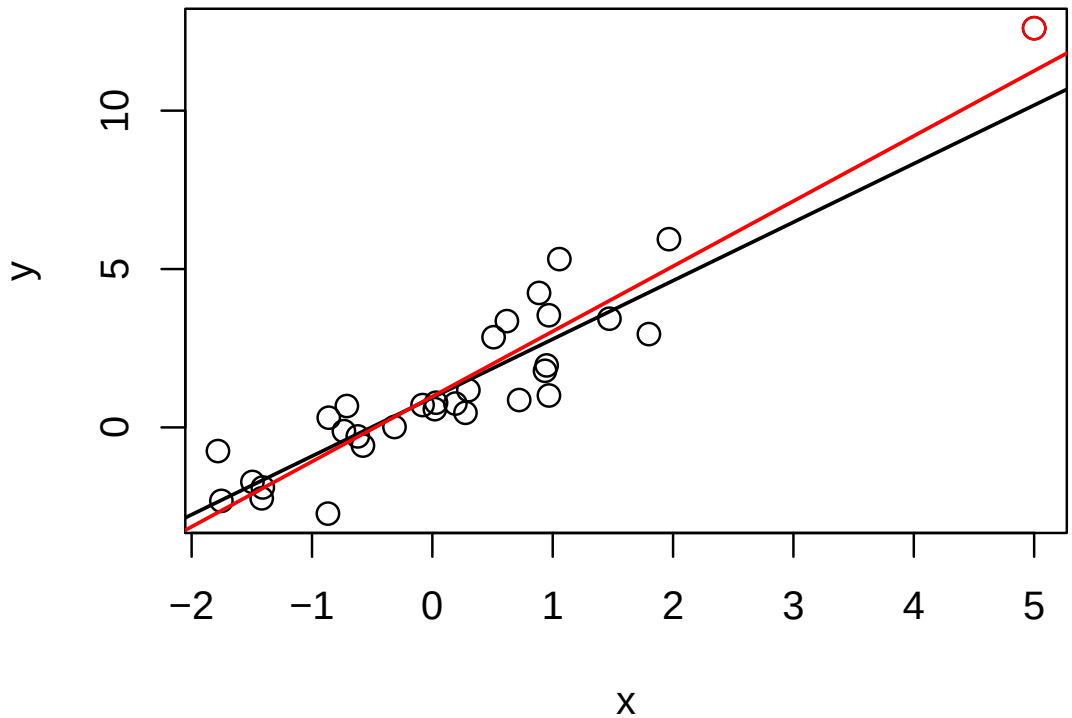
When fitting a regression model to a data set, we can use diagnostics for the following important tasks:

- Determine if any points are unusual, and deviate from the bulk of the data in someway. Such points are called outliers or high leverage points.
- Examine whether the assumptions of linearity and constant variance seem reasonable. Residual plots are useful for this. Consider transformations to fix any problems.
- Examine whether the residuals are normally distributed.
- If the data are collected over time or space, examine whether the residuals are autocorrelated.

A **leverage point** is a point whose x-values are distant from other x-values.

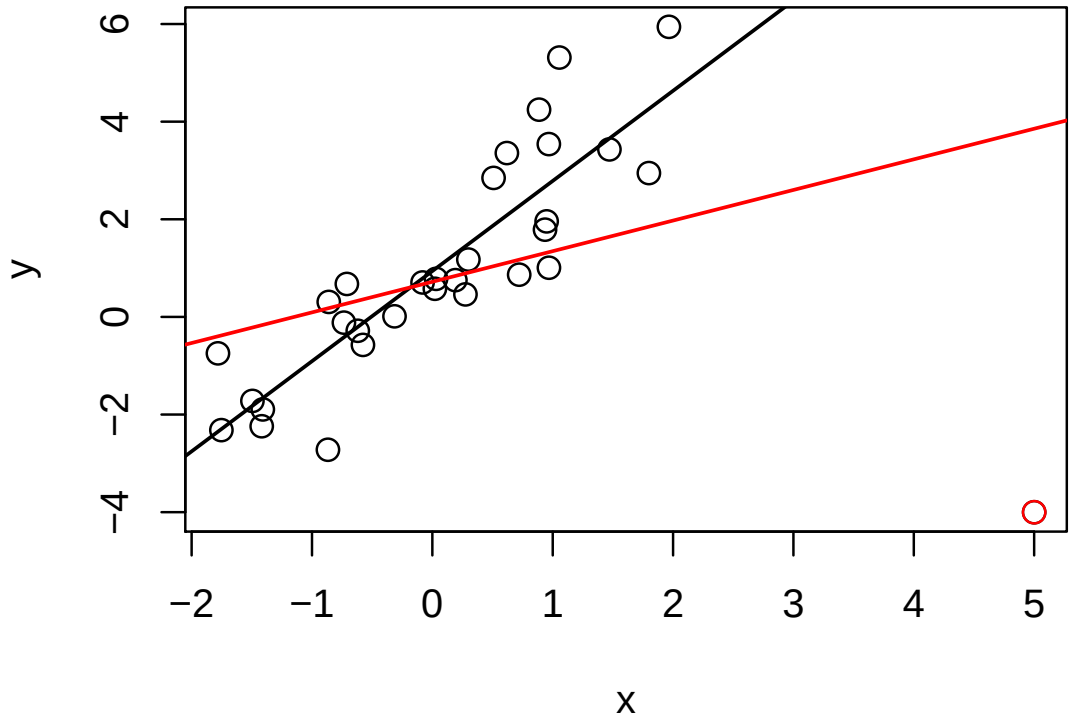
An **outlier** is a point that does not follow the pattern set by the bulk of the data, when one takes into account the given model. That is, outliers have y-values that do not follow the pattern set by the bulk of the data.

Example of a "good" leverage point. A "good" leverage point is a leverage point that is not an outlier.



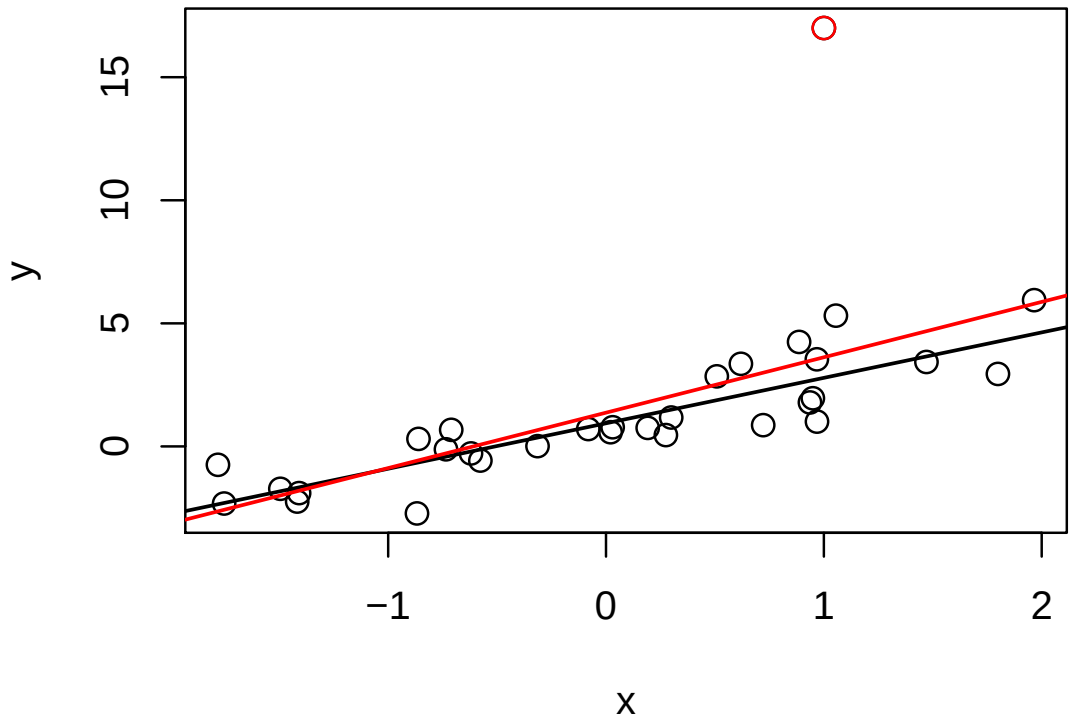
The red line is the least squares line including the high leverage point (red point). The black line is the least squares line without the high leverage point.

Example of a "bad" leverage point. A "bad" leverage point is a leverage point that is also an outlier.



The red line is the least squares line including the high leverage point (red point). The black line is the least squares line without the high leverage point. In this case, removing the high leverage point substantially changes the estimate of the least squares line.

Example of an outlier that is not a leverage point.



The red line is the least squares line including the outlier (red point).
The black line is the least squares line without the outlier.

Quantifying Leverage

The **leverage** of the i^{th} data point is quantified as

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

- h_i increases the further x_i is from $\bar{x}$
- h_i is between $1/n$ and 1
- For simple linear regression, average$(h_i) = 2/n$

See Sheather, Section 3.2.1, pp. 55-56 for a derivation.

Identifying Leverage Points

A popular rule is to classify x_i as a point of high leverage in a simple linear regression model if

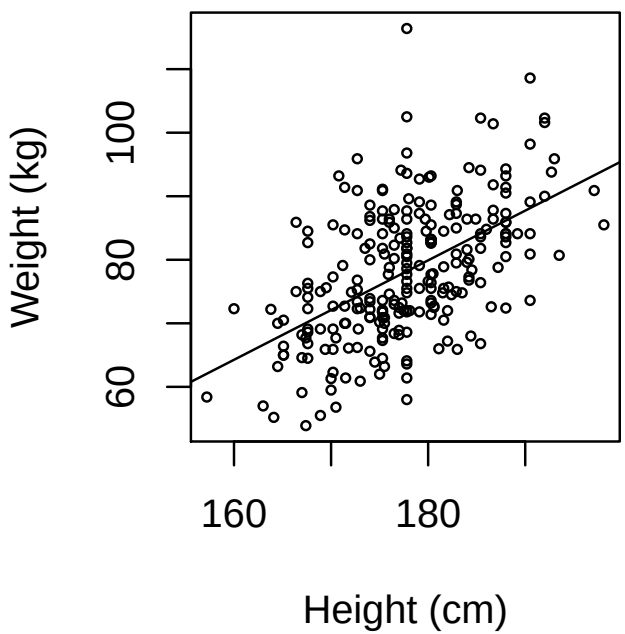
$$h_i > 2 \times \text{average}(h_i) = 2 \times \frac{2}{n} = 4/n$$

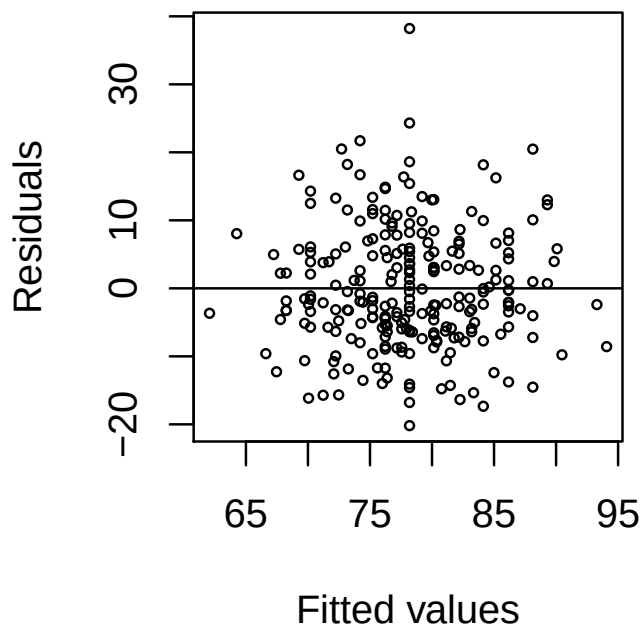
Residual Plots

- One of the most useful diagnostics is a plot of the residuals $\hat{e}_i = y_i - \hat{y}_i$ versus the fitted values $\hat{y}_i$, for $i = 1, \cdots, n$. It is also common to plot the reisduals $\hat{e}_i$ versus the predictor x_i.
- One purpose of residual plots is to identify characteristics or patterns still apparent in the data after fitting the model.
- Residual plots are especially useful for assessing linearity and constant variance.
- Ideally, the residual plot should show no obvious pattern, and the points are randomly scattered around 0.

Residual Plots

For the simple linear regression model of male weight versus height, the points in the residual plot look randomly scattered and show no obvious patterns, indicating that the assumptions are reasonably satisfied. Although, there is one potential outlier.





Here is the code used to create the last plot.

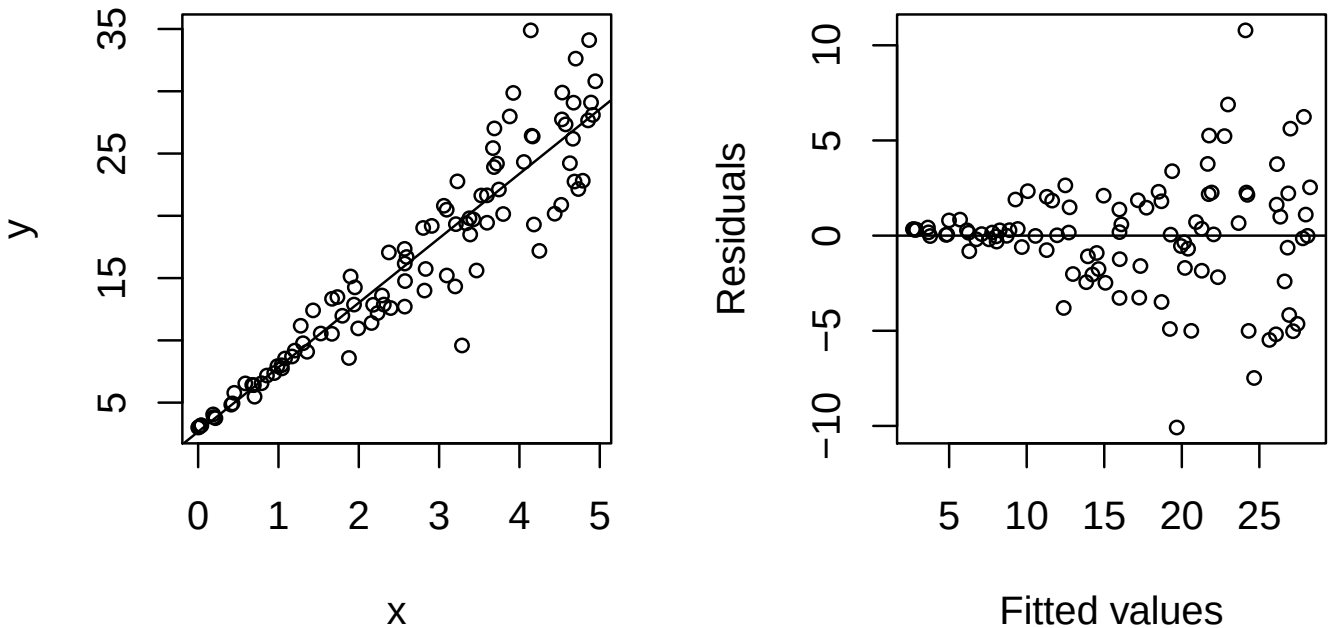
```
> library(openintro)
> bdims_males <- subset(bdims, sex == 1)
> lm1 <- lm(wgt ~ hgt, data=bdims_males)

> par(mfrow=c(1,2)) # split plot into 2 panes
# scatter plot with least squares line
> plot(wgt ~ hgt, data=bdims_males,
       xlab = 'Height (cm)' , ylab = 'Weight (kg)')
> abline(lm1)

# residual plot
> plot(predict(lm1), resid(lm1),
       xlab='Fitted values', ylab='Residuals')
> abline(h=0)
```

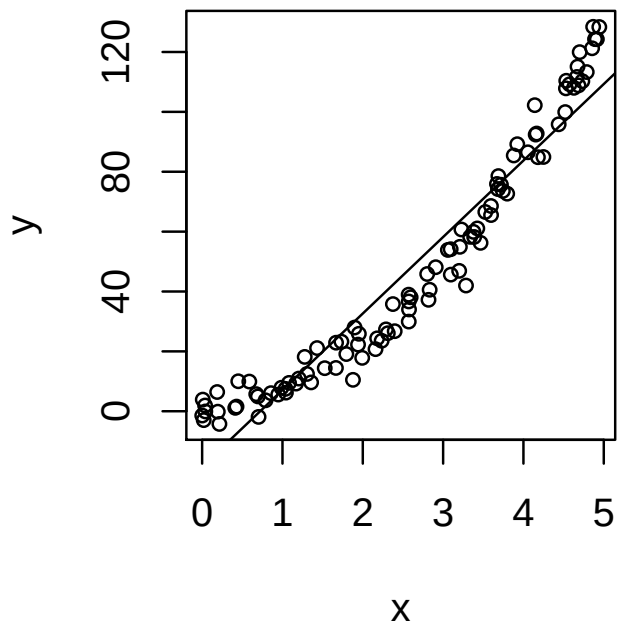
Residual Plots

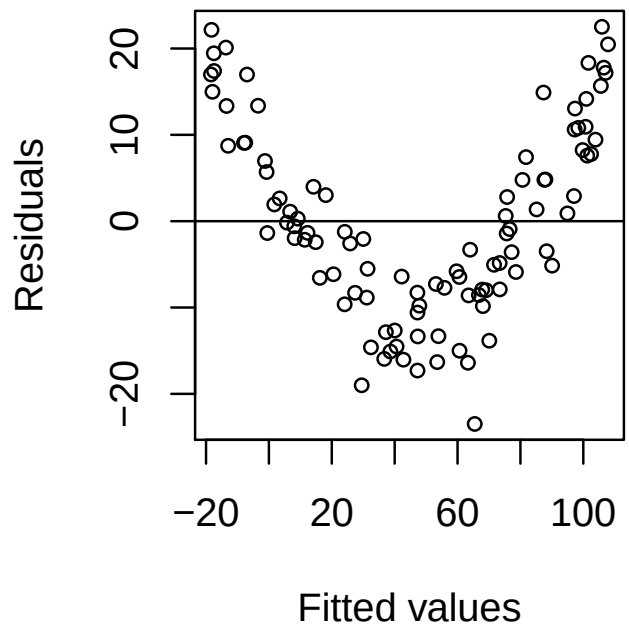
An example of nonconstant variance, also called **heteroscedasticity**. The residual plot below shows a fan pattern.



Residual Plots

An example of nonlinearity.





Question

What is $\hat{e}_i$ and how is it different than e_i?

Standardized Residuals

It can be shown that the variance of the i^{th} residual is given by

$$Var(\hat{e}_i) = \sigma^2[1 - h_i],$$

where h_i is the leverage of the i^{th} point defined earlier. (See Sheather, Section 3.2.2, pp. 60-61, for a derivation.)

This implies residuals do not have equal variance, even though we make the assumption that the errors have constant variance $Var(e_i) = \sigma^2$.

Standardized Residuals

The problem of the residuals having different variances can be overcome by standardizing each residual by its standard error. The i^{th} standardized residual, r_i is given by

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_i}},$$

where $\hat{\sigma} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{e}_i^2}$ is the residual standard error (i.e., the estimate of σ)

Standardized Residuals

- A plot of the standardized residuals versus the fitted values is recommended when there are points of high leverage in a data set.
- When there are no points of high leverage, there is generally little difference between the plot of the raw residuals and the standardized residuals.

Identifying Outliers

- One advantage of standardized residuals is that they inform us how many estimated standard errors any point is away from the fitted regression line.
  - For example, suppose that a point has a standardized residual of 4.3, then this point is 4.3 standard errors away from the fitted regression line, and would therefore be considered an outlier that should be investigated.
- Sheather suggests labeling a point as an outlier if its standardized residual falls outside the interval from **-2 to 2**. For large data sets, he suggests changing this rule to **-4 to 4** (otherwise, too many points would be flagged).

To summarize the rules for identifying outliers and leverage points:

- An **outlier** is a point whose standardized residual falls outside the interval from **-2 to 2**
- A **high leverage point** is a point that has $h_i > 4/n$ (note, this rule is only for simple linear regression)
- A **"bad" leverage point** is leverage point that is also an outlier
- A **"good" leverage point** is a leverage point that is not an outlier

Example: 2000 US Presidental Elections

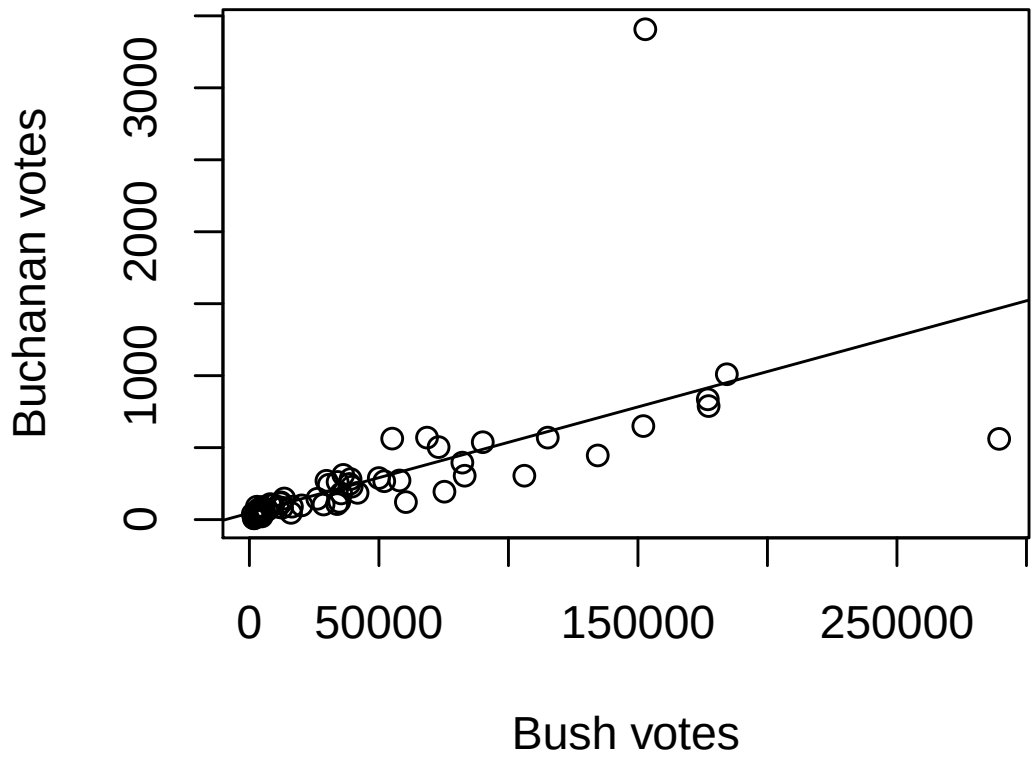
Data set with number of votes for George W. Bush and Pat Buchanan in Florida counties for the 2000 US presidential election.

```
> library(Stat2Data)
> data("PalmBeach")
> head(PalmBeach)
    County Buchanan   Bush
1  ALACHUA      262  34062
2    BAKER       73   5610
3      BAY      248  38637
4 BRADFORD       65   5413
5  BREVARD      570 115185
6  BROWARD      789 177279
```

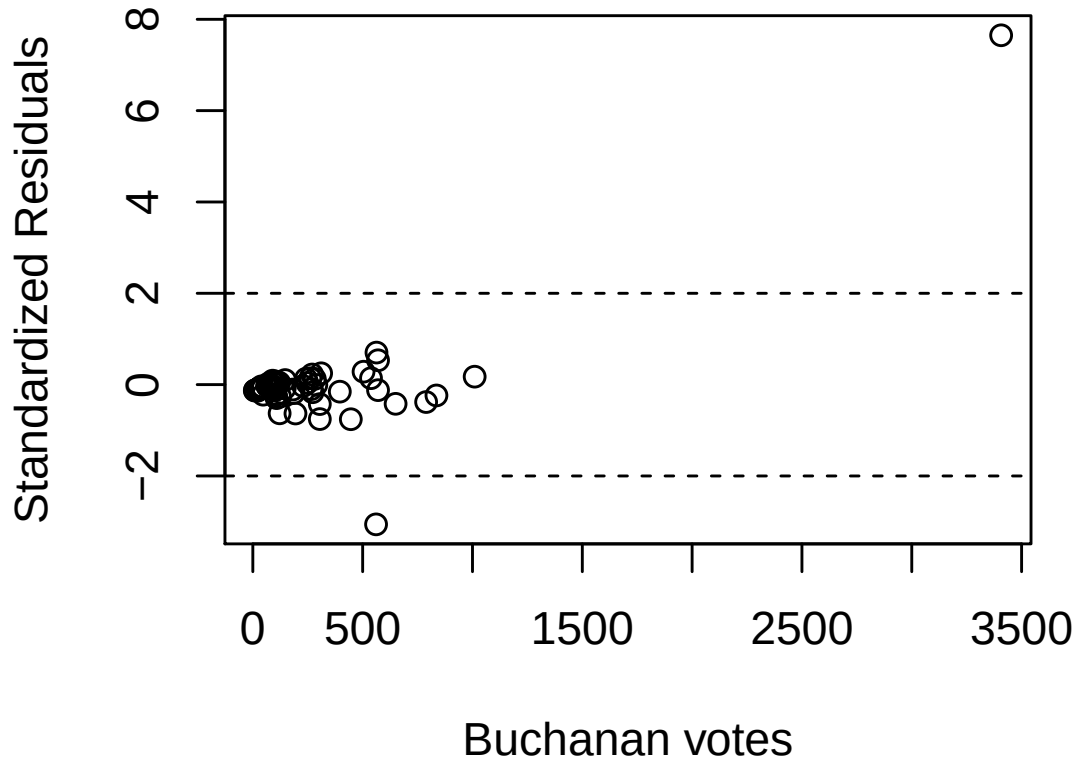
- The 2000 presidential election between George W. Bush and Al Gore was very close, with the electoral college votes from Florida determining the outcome.
- After a period of manual recounting, the Florida vote was ultimately settled in Bush's favor by a margin of 532 votes.[1]
- About 2% of the votes cast in Florida were awarded to other candidates, including the Reform Party candidate Pat Buchanan.

[1] https://en.wikipedia.org/wiki/2000_United_States_presidential_election_recount_in_Florida

```
> lm1 <- lm(Buchanan ~ Bush, data=PalmBeach)
> plot(Buchanan ~ Bush, data=PalmBeach,
    ylab = "Buchanan votes", xlab = "Bush votes")
> abline(lm1)
```



```
> plot(PalmBeach$Buchanan, rstandard(lm1),
    xlab = "Buchanan votes", ylab = "Standardized Residuals")
> abline(h=c(-2,2), lty=2)
```



```
# identify outliers
> ind <- which(abs(rstandard(lm1)) > 2)
> PalmBeach[ind, ]
       County Buchanan   Bush
13       DADE      561 289456
50 PALM BEACH     3407 152846
```

- Palm Beach county used a unique "butterfly ballot," which had a layout that was confusing for many voters.
- Some voters that intedended to vote for Al Gore mistakenly marked their ballots for Pat Buchanan.

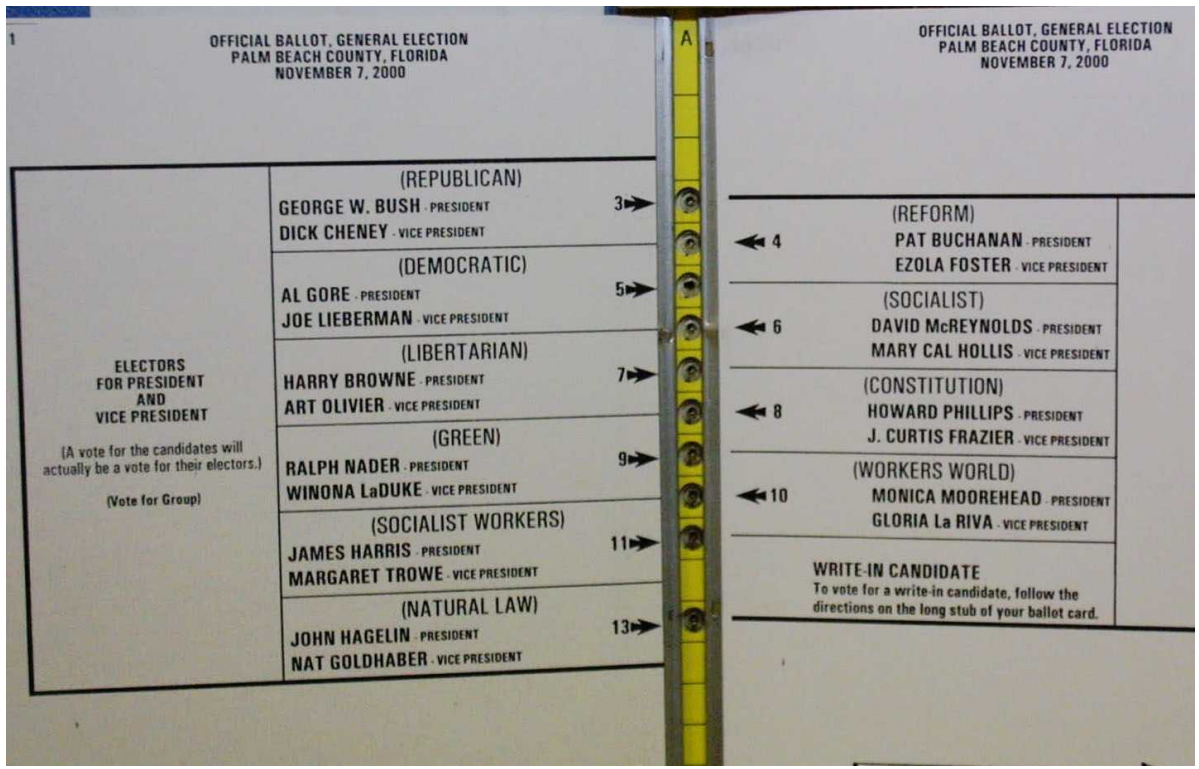
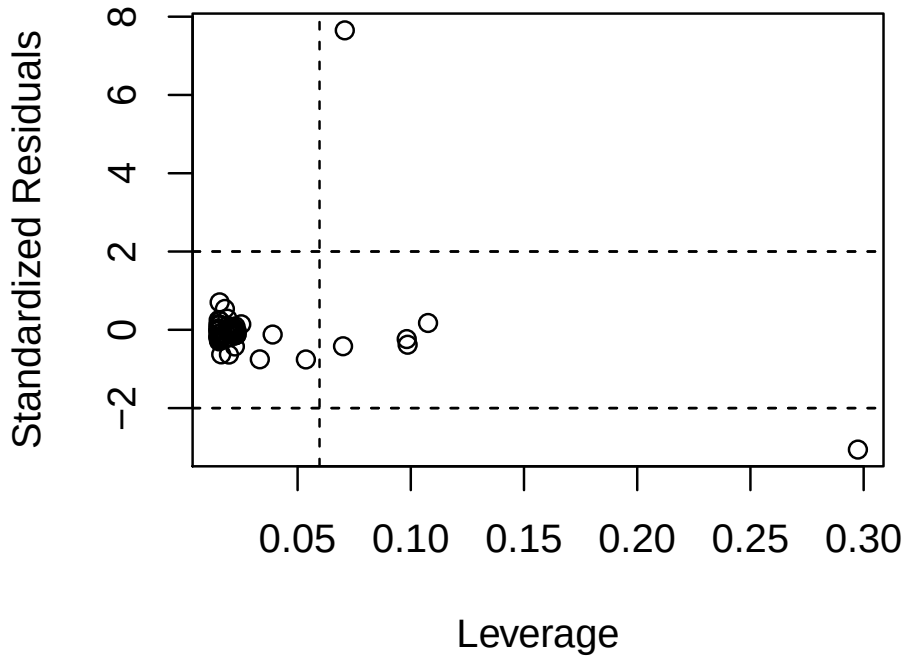


Image Source: https://commons.wikimedia.org/wiki/File:Butterfly_Ballot,_Florida_2000_(large).jpg

```
> plot(hatvalues(lm1), rstandard(lm1),
    xlab = "Leverage", ylab = "Standardized Residuals")
> n <- nrow(PalmBeach)
> abline(v=4/n, lty=2) # threshold for high leverage
> abline(h=c(-2,2), lty=2) # threshold for outliers
```



Question

Write R code that identifies those counties (rows) that correspond to high leverage points?

Write R code that identifies those counties that correspond to "bad" leverage points?

Recommendations for handling outliers and leverage points:

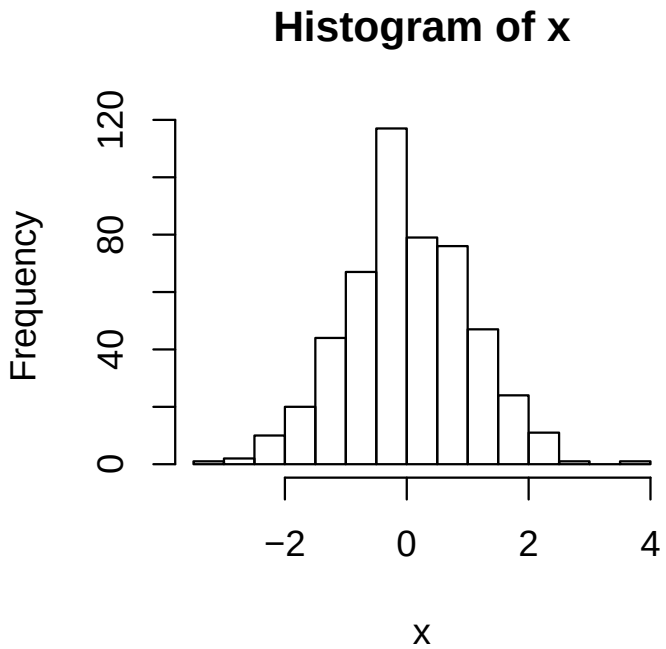
- Points should not be routinely deleted from an analysis because they do not fit the model. Outliers and high leverage points should be carefully inspected. There might be a data entry problem, or the points may be different than the rest of the data in some way (e.g., the "butterfly ballots" in Palm Beach county).
- Outliers and high leverage points may suggest an **alternative model**, in which the points are no longer unusual. Consider including transformations, polynomial terms (e.g., x, x^2), or indicator variables. These methods will be discussed in future lectures.

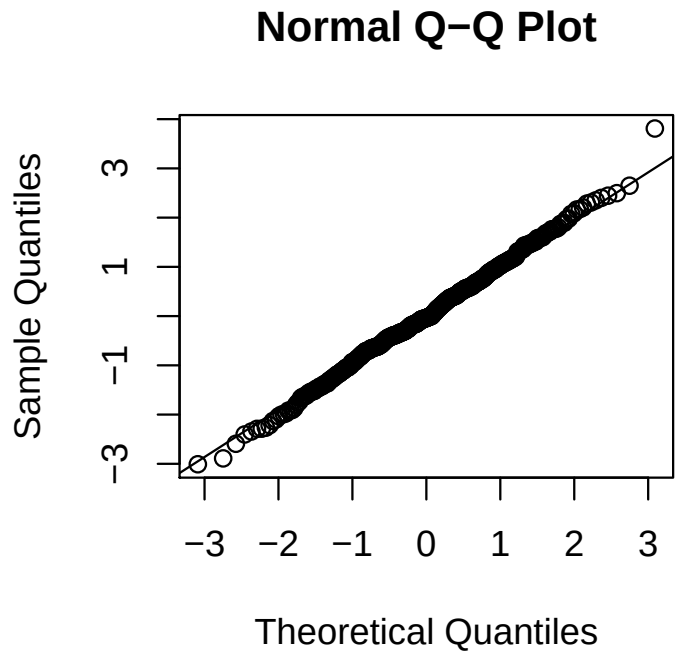
Assessing Normality (review)

- A **normal probability plot**, or **normal QQ plot**, is a useful graphical technique for determining whether a data set follows an approximate normal distribution.
- Technically, a QQ plot is a plot of the sample quantiles (sorted data) on the y-axis, and the corresponding theoretical quantiles from the standard normal distribution on the x-axis.
- If the points follow a straight line then the data are approximately normally distributed. Any deviations form the straight line indicate deviations in the data from the normal distribution.
- Histograms are also useful, but depend on arbitrary bin widths.

For example, if we generate some data from a normal distribution using `rnorm()` then the points follow a straight line in the QQ plot.

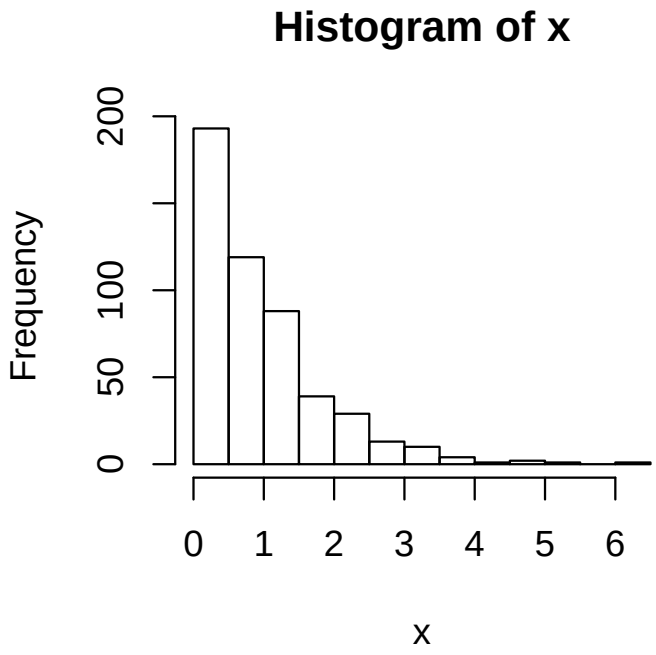
```
> set.seed(1)
> x <- rnorm(500)
> hist(x)
> qqnorm(x)
> qqline(x)
```

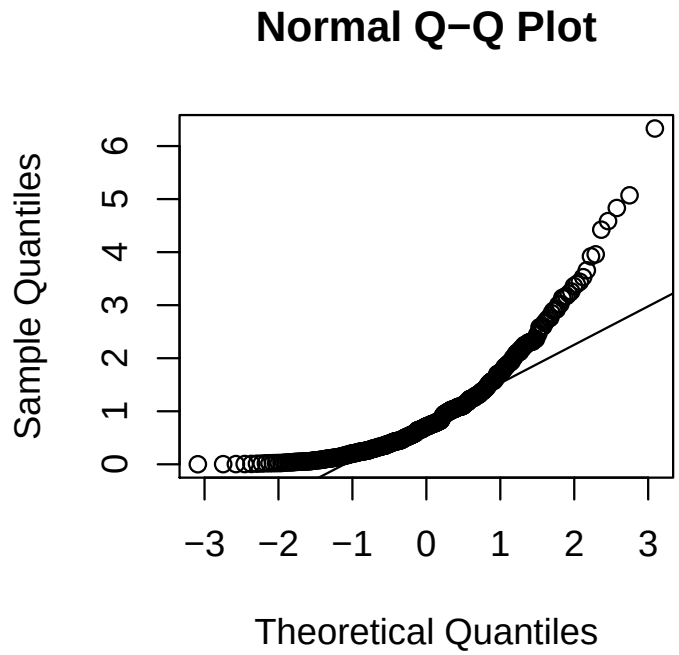




If the data do not follow a normal distribution, then there are deviations from the straight line.

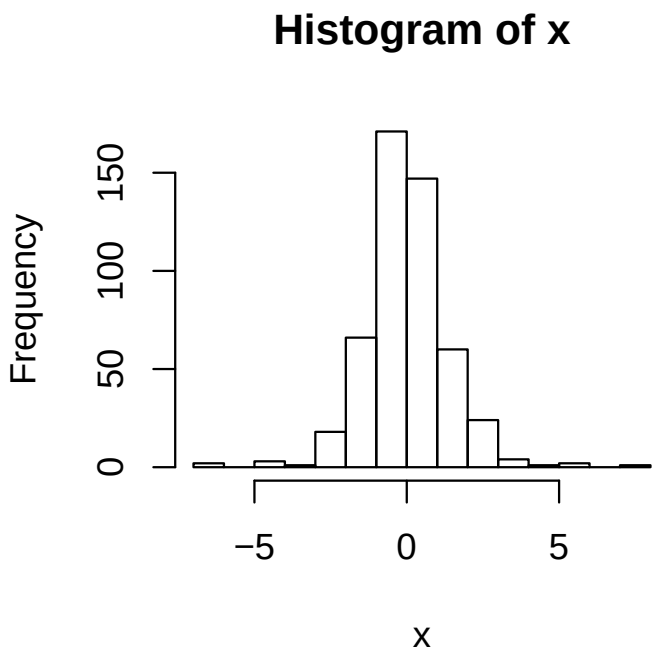
```
> set.seed(1)
> x <- rexp(500) # random numbers from an exponential distribution
> hist(x)
> qqnorm(x)
> qqline(x)
```

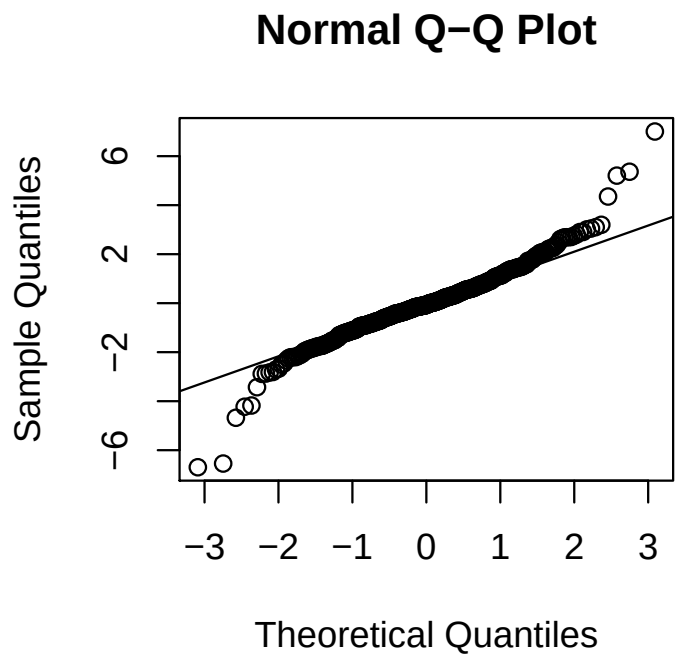




It may not always be obvious from the histogram that the data are not normally distributed.

```
> set.seed(1)
> x <- rt(500, df=5) # random numbers from an t-distribution
> hist(x)
> qqnorm(x)
> qqline(x)
```





- Recall that one of the assumptions for SLR is that the errors are normally distributed (i.e., $e_i \sim N(0, \sigma^2)$)
- QQ plots and histograms are commonly used to check whether the residuals are normally distributed.

For example, below is a QQ plot of the residuals for an SLR model of male weight versus height. The points follow a straight line in the QQ plot, indicating that the residuals are approximately normal. One observation is a potential outlier.

```
> lm1 <- lm(wgt ~ hgt, data=bdims_males)
> hist(resid(lm1))
> qqnorm(resid(lm1))
> qqline(resid(lm1))
```