Chapter 12: Regression Analysis

Yu Yang

School of Statistics
University of Minnesota

November 28, 2022

Example 12.1

How well does the production cost of a movie predict how well it will do at the box office? Data was collected from a random sample of 10 Hollywood movies.

Box	Prod	Promo	Book
1	85.1	8.5	5.1	4.7
2	106.3	12.9	5.8	8.8
.	.	.	.	.
.	.	.	.	.

where (all in millions),

- Box = 1st year Box office receipts
- Prod = production costs
- Promo = promotional costs
- Book = book sales

Full data can be found at
`http://www.stat.umn.edu/~wuxxx725/data/movies.txt`

Goal

Goal: Explore the association between 2 quantitative variables.
Example: production cost and box office sales

The Approach:

1. Get to know the sample data using both graphical and numerical summaries

2. Use the sample data to make inferences about the true population relationship

Graphical Summaries - The Scatterplot

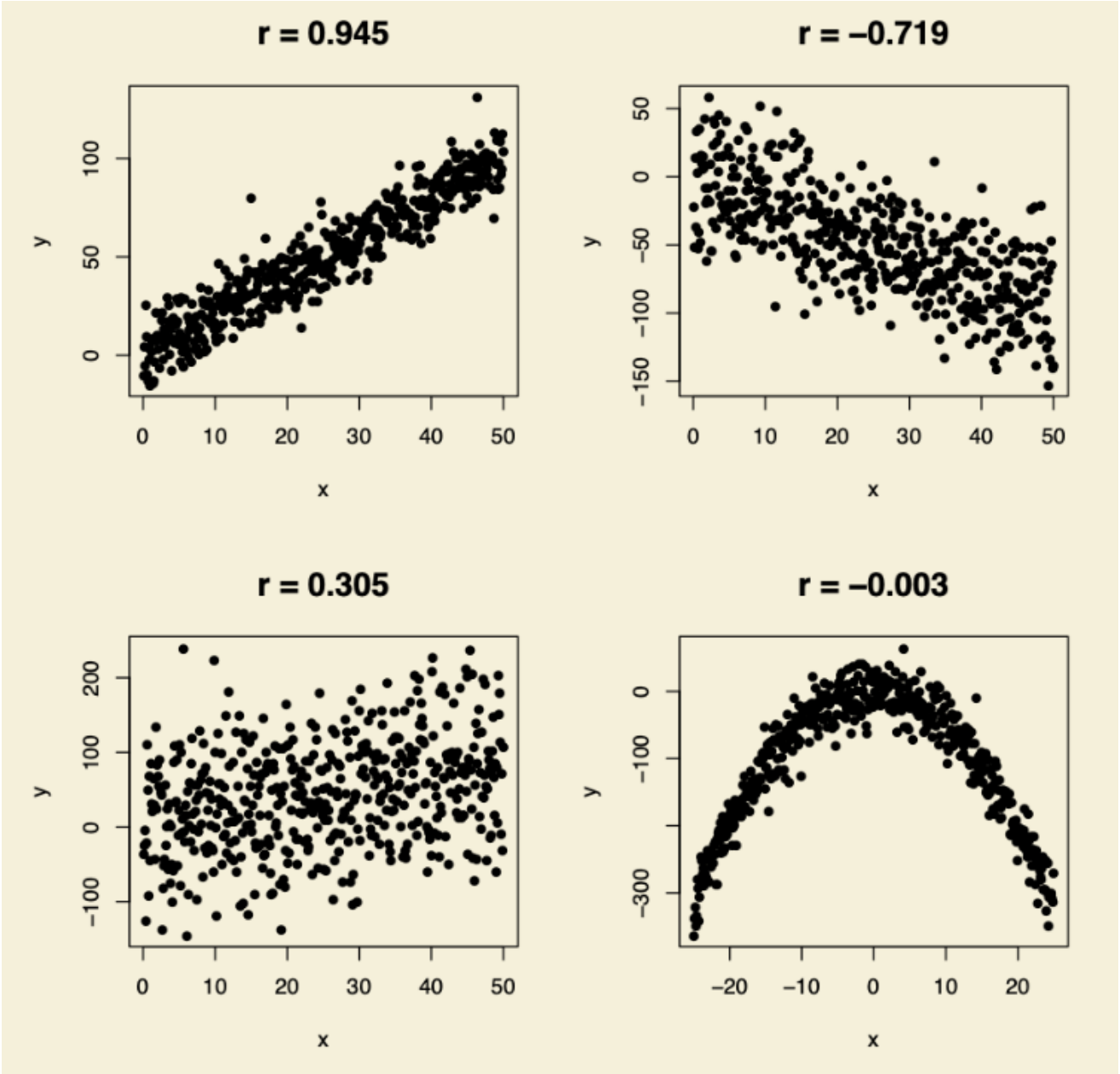
The **scatterplot** is a graphical tool used to display the relationship between two quantitative variables.

Steps to constructing a scatterplot:

1. Label the x-axis (horizontal) with the explanatory variable
   Label the y-axis (vertical) with the response variable

2. Represent each observation with a point in the graph at its (x, y) coordinate.

Example Scatterplots



What to Look For in a Scatterplot

1. **overall pattern** - ex: linear, curved, etc

2. **strength** - How closely do the points follow a pattern?
   ex: weak, moderate, strong

3. **direction** - Do the variables have a positive or negative association?
   Definition: positive association
   when high values of one of the variables tend to occur with high values of the other.

   Definition: negative association
   when high values of one of the variables tend to occur with low values of the other.

4. **outliers** - Are there any unusual points that fall outside the cloud of data points?

Example 12.2

Is there an association between Olympic winning times for the 200 meter dash and the year in which the Olympics took place? The following is a partial data set of the Olympic winning times (in s) starting in 1908 and ending in 1996, where

Year = number of years after 1900
Time = Olympic winning times for the 200m dash.

Data: `http://www.stat.umn.edu/~wuxxx725/data/dash.txt`

```
  Year Time
1    8 22.6
2   12 21.7
3   20 22.0
.    .    .
```

Which is the explanatory variable?

Example 12.2

Is there an association between Olympic winning times for the 200 meter dash and the year in which the Olympics took place? The following is a partial data set of the Olympic winning times (in s) starting in 1908 and ending in 1996, where

Year = number of years after 1900
Time = Olympic winning times for the 200m dash.

Data: `http://www.stat.umn.edu/~wuxxx725/data/dash.txt`

	Year	Time
1	8	22.6
2	12	21.7
3	20	22.0
.	.	.

Which is the explanatory variable? Year
Which is the response variable?

Example 12.2

Is there an association between Olympic winning times for the 200 meter dash and the year in which the Olympics took place? The following is a partial data set of the Olympic winning times (in s) starting in 1908 and ending in 1996, where

Year = number of years after 1900
Time = Olympic winning times for the 200m dash.

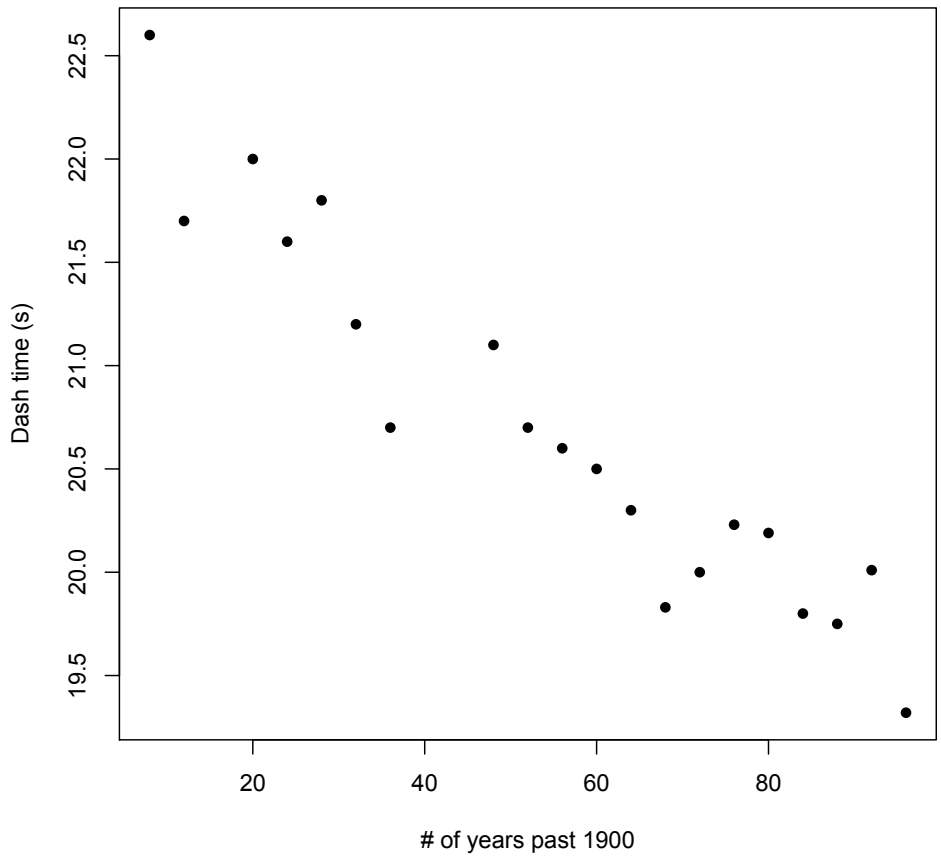
Data: `http://www.stat.umn.edu/~wuxxx725/data/dash.txt`

	Year	Time
1	8	22.6
2	12	21.7
3	20	22.0
.	.	.

Which is the explanatory variable? Year
Which is the response variable? Time

Example 12.2: Scatterplot

```
> dash <- read.table("http://www.stat.umn.edu/~wuxxx725/data/dash.txt",
+                    header = TRUE)
> plot(x = dash$Year, y = dash$Time, xlab = "# of years past 1900",
+      ylab = "Dash time (s)", pch = 16)
```



Scatterplot

- In the `plot()` function, always put the x-axis variable first and the y-axis variable second
- `xlab =` , `ylab =` , and `main =` are used to label the x-axis and y-axis and give the plot a title
- `pch = 16` is used for making the dots solid (purely an aesthetic option)

Describe the relationship between the Olympic winning time of the 200m dash and the year in which the Olympics took place (in terms of overall pattern, strength, direction, and outliers):

Scatterplot

- In the `plot()` function, always put the x-axis variable first and the y-axis variable second
- `xlab =` , `ylab =` , and `main =` are used to label the x-axis and y-axis and give the plot a title
- `pch = 16` is used for making the dots solid (purely an aesthetic option)

Describe the relationship between the Olympic winning time of the 200m dash and the year in which the Olympics took place (in terms of overall pattern, strength, direction, and outliers):

linear, strong, negative relationship
200 m dash times have been improving over time.

Numerical Summaries - Correlation

A scatterplot only allows us to eyeball the strength of the linear relationship. We also want to quantify the relationship

Notation: Let $x = \{x_1, x_2, \ldots, x_n\}$ and $y = \{y_1, y_2, \ldots y_n\}$ denote two paired samples of size n corresponding two two different quantitative variables. Also, let

$\bar{x}$ = sample mean of the x data
$\bar{y}$ = sample mean of the y data
s_x = sample standard deviation of the x data
s_y = sample standard deviation of the y data

Correlation (r)

Correlation is a measure of the strength and direction of the **linear** relationship between two quantitative variables.

Given the random sample $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, the sample correlation between X and Y is

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Properties of r

1. The value of r does not change if we reverse the roles of x and y.
2. The value of r does not change if we change the units of the variable.
   ex: converting pounds to kilograms won't affect r
3. r only measures the strength and direction of the **linear** (not curved) relationship.
4. r is not resistant to outliers.
   the formula for r has $\bar{x}$ and $\bar{y}$, which are not resistant to outliers.

Properties of r (continued)

5. $-1 \leq r \leq 1$

   Direction:

   $r > 0 \Rightarrow$ positive association

   $r < 0 \Rightarrow$ negative association

   $r = 0 \Rightarrow$ uncorrelated (no linear relationship)

   Strength:

   $r \approx 0 \Rightarrow$ weak linear relationship

   $r \approx \pm 1 \Rightarrow$ strong linear relationship

   $r = 1$ or -1 only if all points fall exactly on a straight line.

General rules: If r is in

$[-1, -.8) =$ strong negative

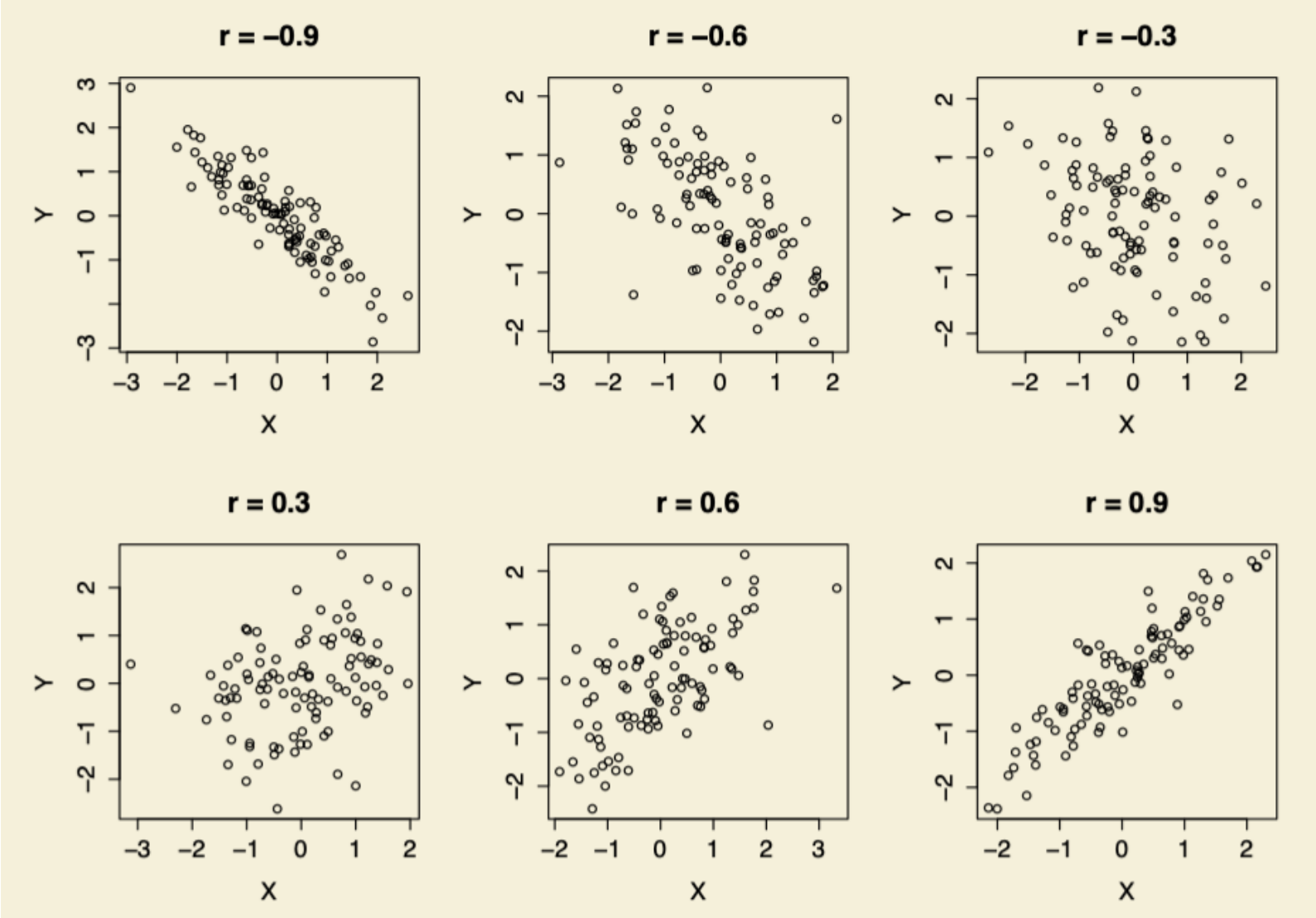
$[-.8, -.5] =$ moderate negative

$[-.5, .5] =$ weak

$(.5, .8] =$ moderate positive

$(.8, 1] =$ strong positive

Examples



Example 12.2 Continued

Use R to calculate the correlation between the year in which the Olympics took place and the winning time for the 200m dash.

```
> cor(dash$Year, dash$Time)
[1] -0.9496326
> cor(dash$Time, dash$Year)
[1] -0.9496326
```

r = -0.950. Therefore, there is a strong , negative linear association between Time and Year. This also agrees with what we saw in the scatterplot.

Nov 28 Lecture Stopped Here

Numerical Summaries - Least Squares Regression

Regression Line

A regression line formulates the linear relationship between a response variable (y) and explanatory variable (x) and can be used to predict the value of y for a given value of x.

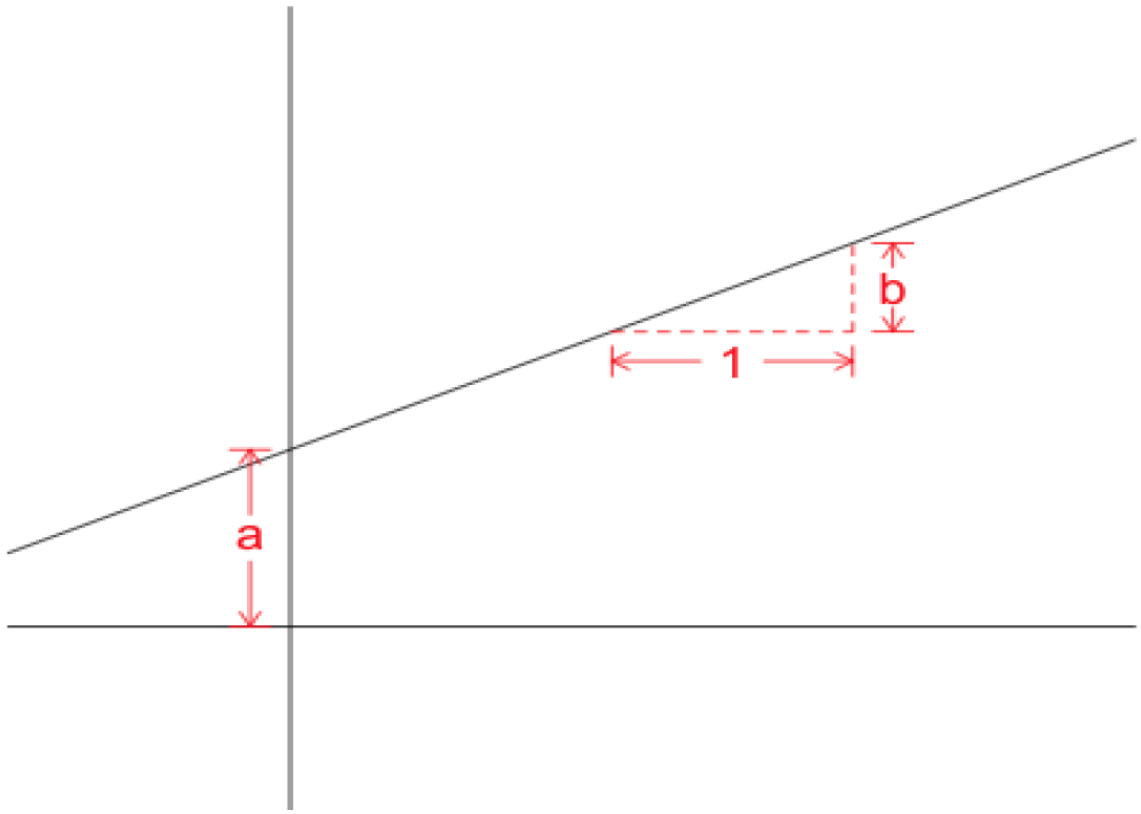
Sample Regression Line

Suppose we have explanatory variable x and a response variable y with observed data pairs (x_1, y_1), $(x_2, y_2), \ldots, (x_n, y_n)$. Then the equation of the sample regression line is

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

- $\hat{y}$ = the predicted value of y at x
- a = intercept (value of y when $x = 0$)
  this may not have any interpretive value if no observation have x values near 0
- b = slope of the line = expected change in y per unit change in x

Sample Regression Line (continued)



Relationship Between r and b

$b > 0 \Rightarrow$ positive slope (positive association)

Therefore $b > 0 \iff r > 0$ and $b < 0 \iff r < 0$.

Using the Regression Line for Prediction

We can predict the value of the response variable at some value of x by plugging x into the equation of the regression line.

Example: Suppose we have a regression line with $a = 2$ and $b = -1/4$.

(a) Write down the formula and draw a picture of the regression line.

$$\hat{y} = a + bx = 2 - \frac{1}{4}x$$

(b) Predict y for $x = 100$.

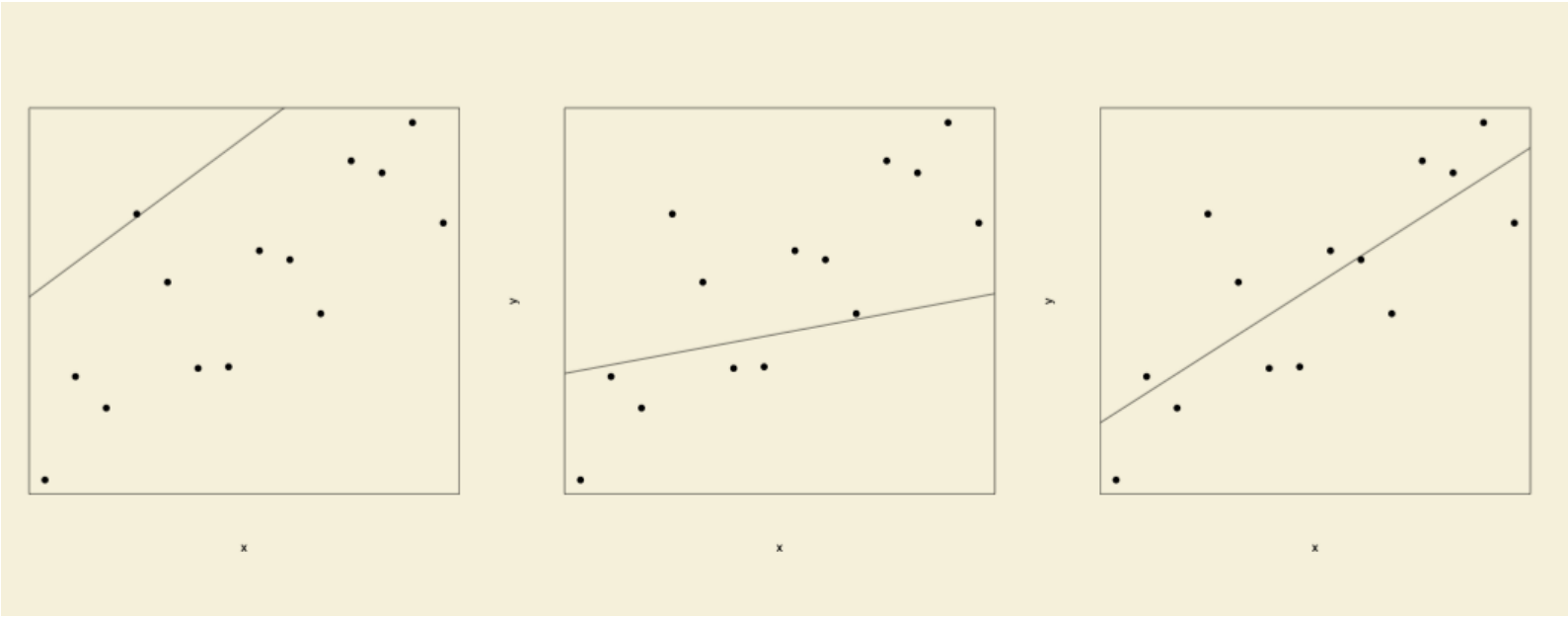
$$\hat{y} = 2 - \frac{1}{4}100 = -23$$

(c) Predict y for $x = 4$.

$$\hat{y} = 2 - \frac{1}{4}4 = 1$$

Find a and b

It is unlikely that the observations will fall exactly on a straight line.

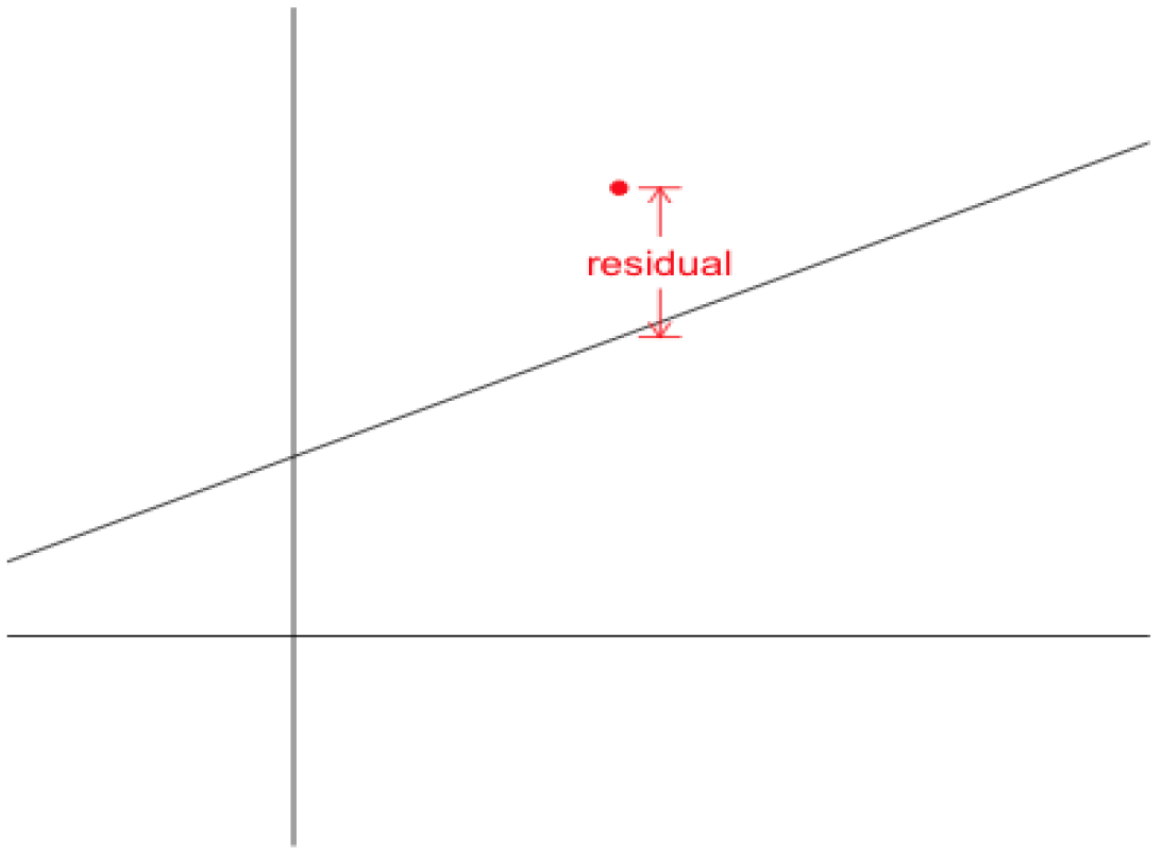


So how do we pick a and b for the regression line?

Residual

A *residual* is the prediction error for an observation. Specifically, it is the difference between an observed value of the response variable y_i and the value predicted by the regression line, $\hat{y}_i = a + bx_i$:

$$\text{residual}_i = y_i - \hat{y}_i$$



Residual (continued)

Notes:

1. Cannot switch the order of y_i and $\hat{y}_i$ in the formula for the residual

2. What does it mean when residual$_i = 0$?
   The predicted value is spot on.

3. What does it mean when the residual $< 0, > 0$?
   residual$_i < 0 \Rightarrow y_i < \hat{y}_i \Rightarrow$ predicted value is too high.
   residual$_i > 0 \Rightarrow y_i > \hat{y}_i \Rightarrow$ predicted value is too low.

Goal: Choose a and b so that

- residuals are small
- not systematically over or under predicting the true value.

Least Squares Regression

Residual Sum of Squares

The *residual sum of squares* (RSS) is one measure of how well a regression line predicts values of the response variable. It is literally the sum of the squared residuals.

$$RSS = \sum_{i=1}^{n} (\text{residual}_i^2) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Least Squares Criterion

Choose a and b to minimize RSS.

Least Squares Line (LSL)

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

where

$$b = r\frac{s_y}{s_x}$$

$$a = \bar{y} - b\bar{x}$$

Nov 30 Lecture Stopped Here

Properties of the Least Squares Regression Line (LSL)

1. Interchanging x and y will give a different regression line.
2. Some residuals are positive and some residuals are negative, but it is **always** true that

$$\sum_{i=1}^{n} \text{residual}_i = \sum (y_i - \hat{y}_i) = 0$$

   Interpretation: low predictions are balanced out by high predictions
3. RSS for the LSL is smaller than for any other line.
4. The LSL **always** passes through the point $(\bar{x}, \bar{y})$, i.e. $\bar{y} = a + b\bar{x}$

Properties of LSL (continued)

5. r^2 provides a measure of how well the LSL describes the relationship between x and y (where r is the correlation)
   - r^2 = proportion of variation in y that is explained by its linear relationship with x
   - $0 \leq r^2 \leq 1$
   - The closer r^2 is closer to 1, the better x can be used to predict y
   - $r^2 = 1 \Rightarrow$ all variability in y is explained by its linear relationship with x

Example 12.1 Continued

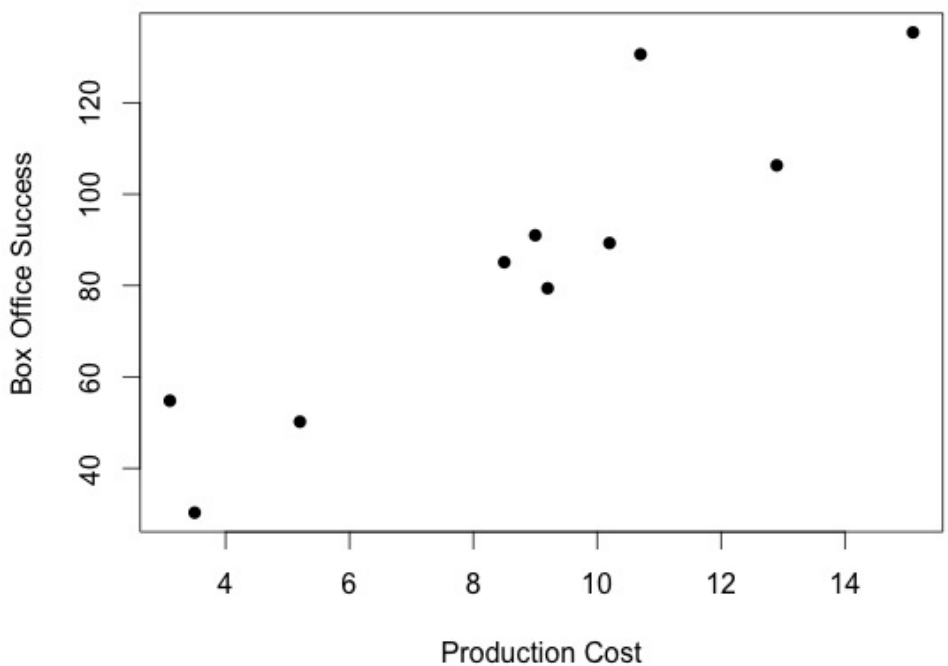
Recall: We are interested in exploring the relationship between box office sales and the cost of production. Which is the response and which is the explanatory variable?

$y =$ **Response: box office sales (in millions)**
$x =$ **Explanatory variable: production cost (in millions)**

Example 12.1 Continued (Graphical Summaries)

Use R to draw a scatterplot and estimate the true linear relationship between box office success and the cost of production:

```
dat <- read.table("http://www.stat.umn.edu/~wuxxx725/data/movies.txt",
                  header = TRUE)
plot(x = dat$Prod, y = dat$Box, xlab = "Production Cost",
     ylab = "Box Office Sales", pch = 16)
```



Example 12.1 Continued (Numerical Summaries)

```
> mod <- lm(Box ~ Prod, data = dat) # lm(response ~ explanatory)
> summary(mod)

Call:
lm(formula = Box ~ Prod, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-13.136  -9.029  -3.689   3.208  29.723

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   15.513     11.603   1.337 0.217989
Prod           7.978      1.223   6.522 0.000184 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.26 on 8 degrees of freedom
Multiple R-squared:  0.8417,Adjusted R-squared:  0.8219
F-statistic: 42.54 on 1 and 8 DF,  p-value: 0.0001838
```

Example 12.1 Continued (Numerical Summaries)

Notes:

- Function "`lm`" stands for "**l**inear **m**odel".
  (That is, "`lm`" is the *letter* l and the letter m. Don't make a typo!)
- The form of the "`lm`" function is always "`lm(response ~ explanatory)`".
- "mod" (or whatever other name you assign your results) has more information than just coefficient estimates and r^2. Let's take a look:

```
> names(mod)  #shows what mod holds
 [1] "coefficients"  "residuals"     "effects"
 [4] "rank"          "fitted.values" "assign"
 [7] "qr"            "df.residual"   "xlevels"
[10] "call"          "terms"         "model"
> mod$fitted.values  #predicted values of Time based on the model
       1         2         3         4         5         6
 83.32530 118.42820  56.99811 100.87675  40.24445  43.43563
       7         8         9        10
 88.90985  87.31426 135.97966  96.88778
> mod$residuals  #residuals
         1          2          3          4          5
 1.7747041 -12.1282047  -6.7981143  29.7232497  14.5555467
         6          7          8          9         10
```

Example 12.1 Continued (Numerical Summaries)

(a) What is the y-intercept of the least squares regression line? (a)

Example 12.1 Continued (Numerical Summaries)

(a) What is the y-intercept of the least squares regression line? (a)
$a = 15.513$

(b) What is the slope of the least squares regression line? (b). Interpret the slope in the context.

Example 12.1 Continued (Numerical Summaries)

(a) What is the y-intercept of the least squares regression line? (a)
$a = 15.513$

(b) What is the slope of the least squares regression line? (b). Interpret the slope in the context.
$b = 7.978$
we estimate box office sales to increase by almost \$8 million (on average) for every one million spent on the movie's production

(c) Write down the equation of the least squares regression line.

Example 12.1 Continued (Numerical Summaries)

(a) What is the y-intercept of the least squares regression line? (a)
$a = 15.513$

(b) What is the slope of the least squares regression line? (b). Interpret the slope in the context.
$b = 7.978$
we estimate box office sales to increase by almost \$8 million (on average) for every one million spent on the movie's production

(c) Write down the equation of the least squares regression line.
$\widehat{\text{Box}} = \text{a} + \text{b(Prod)} = 15.513 + 7.978\ \text{(Prod)}$ or
$\hat{y} = 11.513 + 7.978x$

(d) Predict the box office sales for a movie with production costs of \$7 million.

Example 12.1 Continued (Numerical Summaries)

(a) What is the y-intercept of the least squares regression line? (a)
$a = 15.513$

(b) What is the slope of the least squares regression line? (b). Interpret the slope in the context.
$b = 7.978$
we estimate box office sales to increase by almost \$8 million (on average) for every one million spent on the movie's production

(c) Write down the equation of the least squares regression line.
$\widehat{\text{Box}} = \text{a} + \text{b(Prod)} = 15.513 + 7.978\ \text{(Prod)}$ or
$\hat{y} = 11.513 + 7.978x$

(d) Predict the box office sales for a movie with production costs of \$7 million.
$\widehat{\text{Box}} = \text{a} + \text{b(Prod)} = 15.513 + 7.978\ (7) = \$71.4$

Example 12.1 Continued (Numerical Summaries)

(e) State and interpret r^2.

Example 12.1 Continued (Numerical Summaries)

(e) State and interpret r^2.
$r^2 = 0.8417$
Therefore, about 84% of the variability in Box office sales is accounted for by its linear relationship with production cost.
The other 16% of the variation in Box office sales is accounted for by other factors (ex: rating, genre, number of screens, run time, etc)

(f) Calculate the sample correlation r.

Example 12.1 Continued (Numerical Summaries)

(e) State and interpret r^2.
$r^2 = 0.8417$
Therefore, about 84% of the variability in Box office sales is accounted for by its linear relationship with production cost.
The other 16% of the variation in Box office sales is accounted for by other factors (ex: rating, genre, number of screens, run time, etc)

(f) Calculate the sample correlation r.
$r =$ sign of (b)$\sqrt{r^2} = (+)(.8417) = 0.917$

Example 12.1 Continued (Numerical Summaries)

(g) *Finding Nemo*, a computer animated film produced by Pixar in 2003, had its production budget 94 million USD and box office sale 940.3 million USD. Use the regression equation from part (c) and find the residual of this movie.

$$\hat{y} = 15.513 + 7.978(94) = 765.445$$

$$\text{residual} = y - \hat{y} = 940.3 - 765.445 = 174.855$$

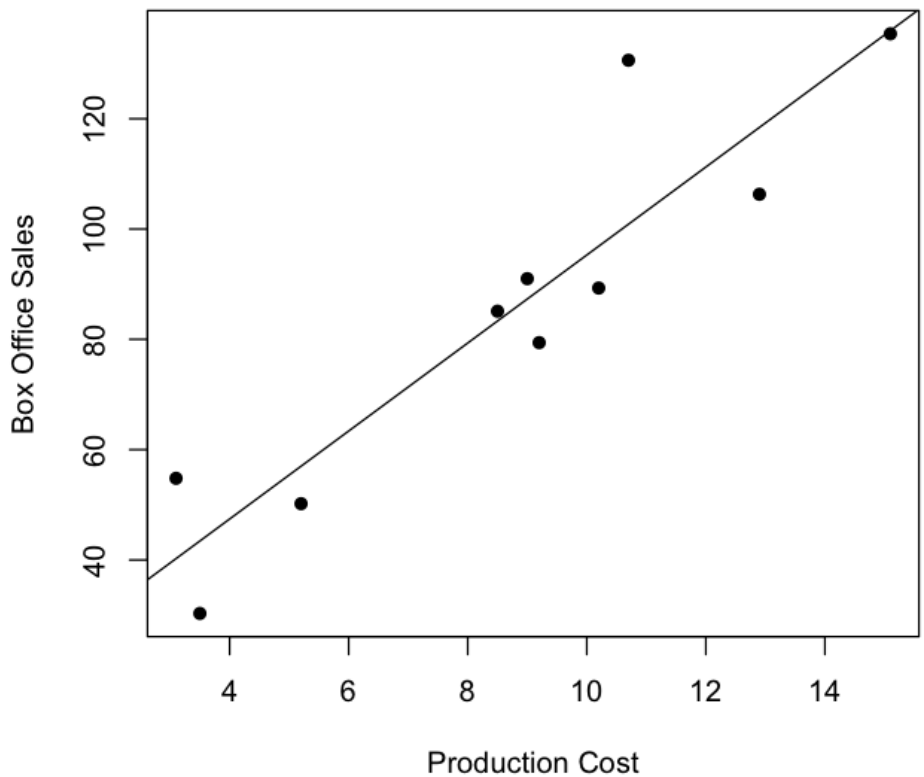
The prediction provided by the regression line for *Finding Nemo* was not good, even though the relationship in the data looks quite linear and the r^2 value is high.

Why did the prediction fail? (There could be multiple reasons.)
Extrapolation: The value of x in the new observation is too far outside of the range of the observed values of x.
Non-linearity: While the r^2 value of this model appears to be high, there might exist other non-linear models that work even better.

Example 12.1 Continued (Plot the Least Squares Line)

After using R to find the least squares line, we can draw a graph that includes the data points along with the fitted line.

```
> plot(x = dat$Prod, y = dat$Box, xlab = "Production Cost",
+      ylab = "Box Office Sales", pch = 16)
> abline(mod)
```



Example 12.2 Continued

Recall: We are interested in quantifying the linear relationship between the Olympic winning time of the 200m dash and the year in which the Olympics took place. Let

- x = Year = number of years after 1900
- y = Time = Olympic winning times for the 200m dash (in s)

Step 1: Graphical Summary
We already looked!

Example 12.2 Continued (Numerical Summaries)

```
> mod2 <- lm(Time ~ Year, data = dash)
> summary(mod2)

Call:
lm(formula = Time ~ Year, data = dash)

Residuals:
     Min       1Q   Median       3Q      Max
-0.56550 -0.14113 -0.03391  0.21083  0.48704

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.355087   0.143763  155.50  < 2e-16 ***
Year        -0.030266   0.002354  -12.86 1.65e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2838 on 18 degrees of freedom
Multiple R-squared:  0.9018,Adjusted R-squared:  0.8963
F-statistic: 165.3 on 1 and 18 DF,  p-value: 1.649e-10
```

Example 12.2 Continued (Numerical Summaries)

(a) Use the R output to find the equation of the least squares regression line

Example 12.2 Continued (Numerical Summaries)

(a) Use the R output to find the equation of the least squares regression line $\hat{y} = 22.36 - .0303x$

(b) Interpret the value of the slope

Example 12.2 Continued (Numerical Summaries)

(a) Use the R output to find the equation of the least squares regression line
$\hat{y} = 22.36 - .0303x$

(b) Interpret the value of the slope
On average, 200m dash times improve by .0303 seconds every year

(c) By how much would we expect the 200m dash time to improve from one Olympics to the next (4 years apart)?

Example 12.2 Continued (Numerical Summaries)

(a) Use the R output to find the equation of the least squares regression line
$\hat{y} = 22.36 - .0303x$

(b) Interpret the value of the slope
On average, 200m dash times improve by .0303 seconds every year

(c) By how much would we expect the 200m dash time to improve from one Olympics to the next (4 years apart)?
-.0303(4) = -.1212

(d) Estimate what the winning 200m dash time might have been had the Olympics been held in 1950.

Example 12.2 Continued (Numerical Summaries)

(a) Use the R output to find the equation of the least squares regression line
$\hat{y} = 22.36 - .0303x$

(b) Interpret the value of the slope
On average, 200m dash times improve by .0303 seconds every year

(c) By how much would we expect the 200m dash time to improve from one Olympics to the next (4 years apart)?
-.0303(4) = -.1212

(d) Estimate what the winning 200m dash time might have been had the Olympics been held in 1950.
$\hat{y} = 22.36 - 0.0303(50) = 20.84$

(e) State and interpret r^2.

Example 12.2 Continued (Numerical Summaries)

(a) Use the R output to find the equation of the least squares regression line
$\hat{y} = 22.36 - .0303x$

(b) Interpret the value of the slope
On average, 200m dash times improve by .0303 seconds every year

(c) By how much would we expect the 200m dash time to improve from one Olympics to the next (4 years apart)?
-.0303(4) = -.1212

(d) Estimate what the winning 200m dash time might have been had the Olympics been held in 1950.
$\hat{y} = 22.36 - 0.0303(50) = 20.84$

(e) State and interpret r^2.
$r^2 = .902$
About 90% of the variability in winning times is accounted by its linear relationship with year. The other 10% of the variation is accounted by other factors that are not in this study (e.g. temp, running surface, etc.)

Example 12.2 Continued (Numerical Summaries)

(f) Calculate the sample correlation r.

Example 12.2 Continued (Numerical Summaries)

(f) Calculate the sample correlation r.
$r = sign(b)\sqrt{r^2} = (-1)(.950) = -.950$

(g) The 2012 Olympics 200m dash was won by Usain Bolt of Jamaica in 19.32 sec. Find the residual for this observation.

Example 12.2 Continued (Numerical Summaries)

(f) Calculate the sample correlation r.
$r = sign(b)\sqrt{r^2} = (-1)(.950) = -.950$

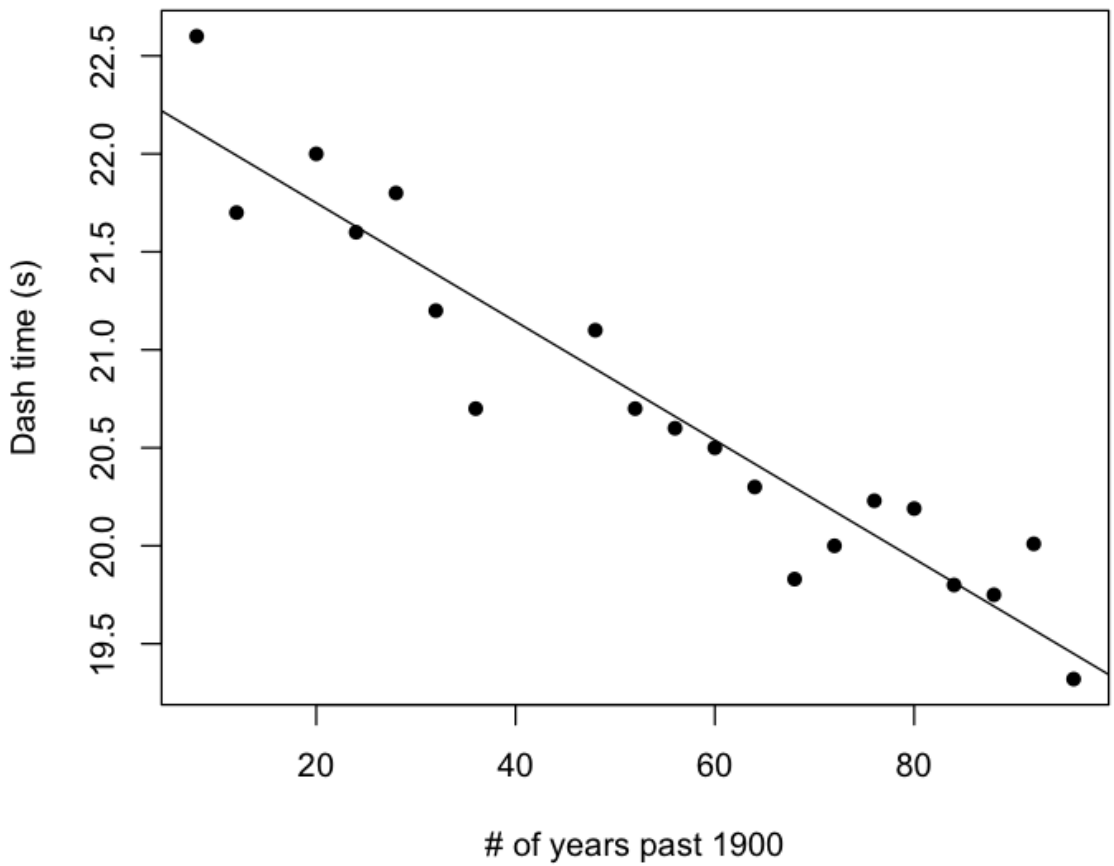
(g) The 2012 Olympics 200m dash was won by Usain Bolt of Jamaica in 19.32 sec. Find the residual for this observation.
$\hat{y} = 22.355 - .0303(112) = 18.97$
residual $= y - \hat{y} = 19.32 - 18.97 = 0.35$

The prediction for the 2012 Olympics was not very good, even though relationship is linear and r^2 is high. Why did the prediction fail?

- Extrapolation
- Non-linearity

Example 12.2 Continued (Plot the Least Squares Line)

```
> plot(x = dash$Year, y = dash$Time, xlab = "# of years past 1900",
+      ylab = "Dash time (s)", pch = 16)
> abline(mod2)
```



Dec 2 Lecture Stopped Here

Regression Analysis

We will now learn to make inferences about the true relationship of the two quantitative variables.

Population Regression Model
Let

- $y =$ value of the response variable
- $x =$ value of the explanatory variable

The trend of the linear relationship between x and y is described by

$$\mu_x = \alpha + \beta x$$

What is the interpretation of the trend?
μ_x is the expected value of y given a particular value of x.

Model Assumptions

When using such a model to describe the linear relationship between x and y (quantitative), we make 3 assumptions.

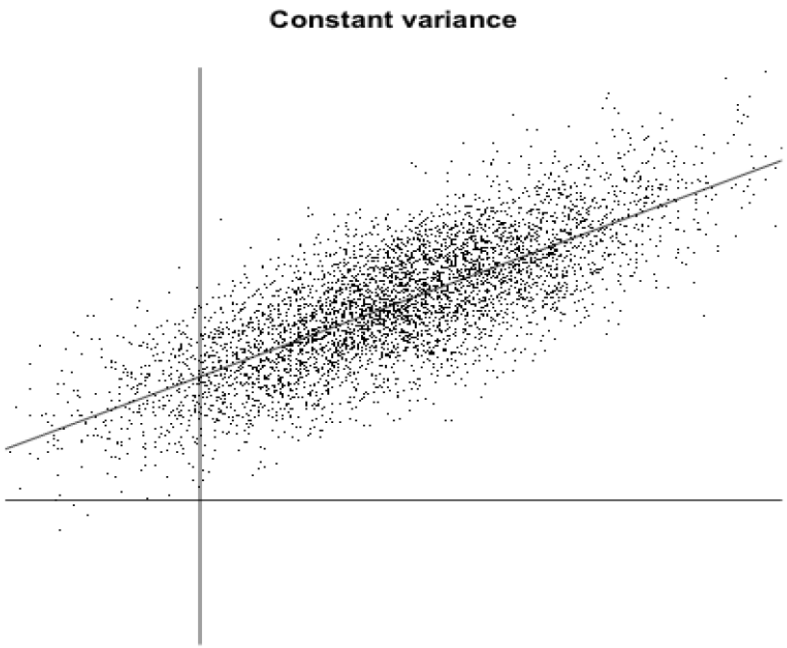
1. A linear relationship between x and y is appropriate

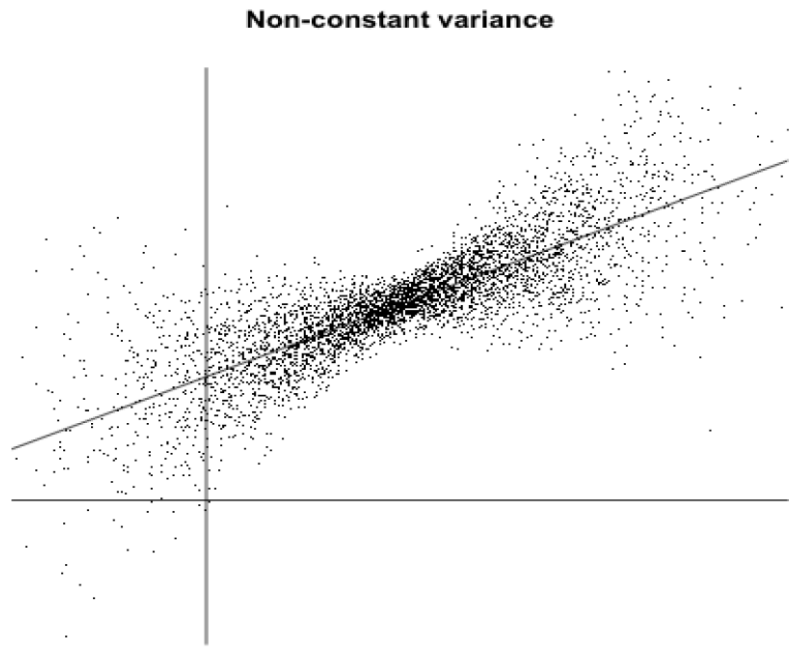
2. Normality: the distribution of y values at a given value of x is

$$y \sim N(\mu_x, \sigma) = N(\alpha + \beta x, \sigma)$$

3. Constant variance: the standard deviation of y, σ, is the same for all values of x.

Non-Constant Variance





Inference

We can make inference about the unknown true population regression model through

1. point and interval estimation

2. hypothesis testing

Estimating α and β

Recall that the population regression model can be written as

$$\mu_x = \alpha + \beta x$$

However, the *parameters* α and β are unknown. That is, the true linear relationship between x and y is unknown.

Goal: Choose estimates α and β that will result in small prediction errors.

The Bottom Line: The least squares line

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

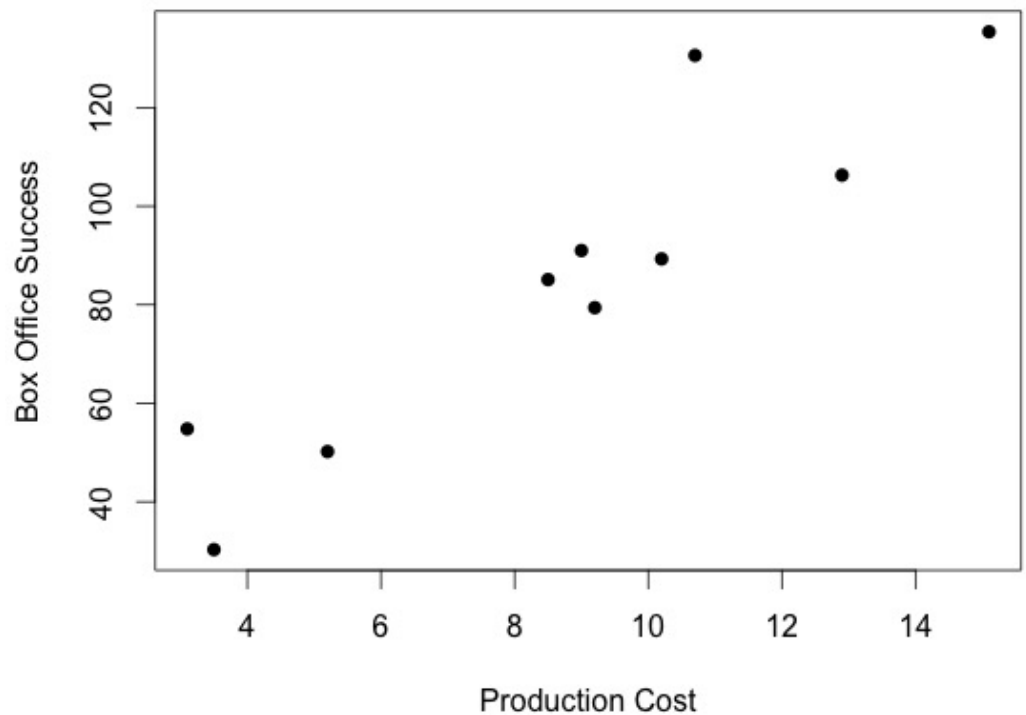
serves as a **point estimate** of the true population regression model:

- a is an unbiased estimator of α.
- b is an unbiased estimator of β.

Example 12.1 Continued

Relationship between box office success and the cost of production.

```
> plot(Prod, Box, xlab = "Production Cost",
       ylab = "Box Office Success", pch = 16)
```



Example 12.1 Continued

```
> mod <- lm(Box ~ Prod, data = dat) # lm(response ~ explanatory)
> summary(mod)

Call:
lm(formula = Box ~ Prod, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-13.136  -9.029  -3.689   3.208  29.723

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   15.513     11.603   1.337 0.217989
Prod           7.978      1.223   6.522 0.000184 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.26 on 8 degrees of freedom
Multiple R-squared:  0.8417,Adjusted R-squared:  0.8219
F-statistic: 42.54 on 1 and 8 DF,  p-value: 0.0001838
```

Example 12.1 Continued

(a) What is the estimate of α?

Example 12.1 Continued

(a) What is the estimate of α?
a = 15.513

(b) State and interpret the estimate of β?

Example 12.1 Continued

(a) What is the estimate of α?
a = 15.513

(b) State and interpret the estimate of β?
b = 7.978.
We expect box office sales to increase by almost $8 million for every one million spent on the movie's production

Hypothesis Tests and Confidence Intervals for β

Both α and β are population parameters, about which we want to gain information. We have a point estimate (the regression line), now we want to make **confidence intervals** and perform **hypothesis tests**!

We rely on the following information
b = point estimate of β

standard error of $b = se(b)$
We will calculate this in R since the formula is complicated

Sampling distribution of b:

$$\frac{b-\beta}{se(b)} \sim t_{n-2}$$

where n = sample size

Confidence Intervals for β

Assumptions:

1. Random sample of n pairs of data
2. Model Assumptions:
   (a) Linear relationship is appropriate
   (b) Normality
   (c) Constant variance

$1 - \alpha$ confidence level CI for β:

$$\text{point est} \pm \text{margin of error}$$

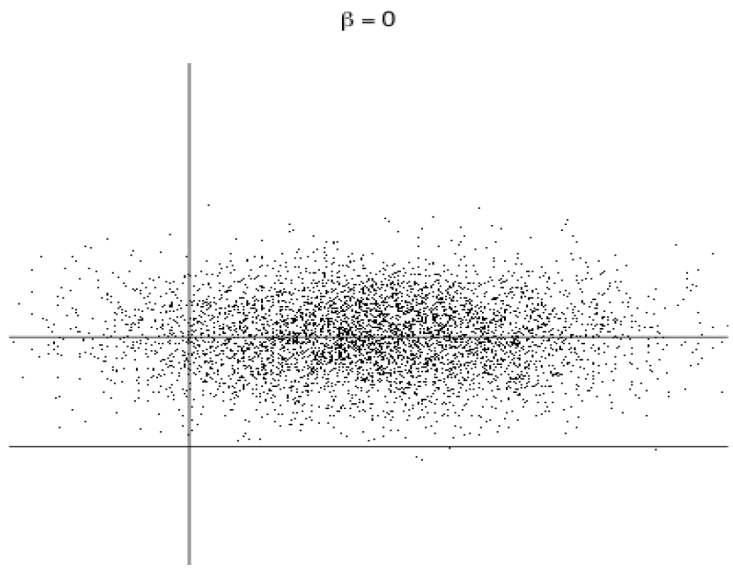
$$b \pm t_{\alpha/2, n-2} \cdot se(b)$$

Hypothesis Tests for β

Remember, we are hypothesizing that x is unrelated to y.

Assumptions: Same as for the confidence intervals

Hypotheses:
$H_0 : \beta = 0$
$H_a : \beta \neq 0$
Interpretation: When $\beta = 0$, y does not depend on x.



Hypothesis Tests for β (continued)

The following are equivalent:

$$H_0 : \beta = 0 \Leftrightarrow H_0 : x \text{ and } y \text{ are independent}$$
$$\Leftrightarrow H_0 : \text{we wouldn't want to use } x \text{ to predict } y$$

Test Statistic:

$$t = \frac{\text{point est - hyp value}}{se(\text{point est})} = \frac{b}{se(b)} \sim t_{n-2} \text{ when } H_0 \text{ is true}$$

p-value: $p\text{-value} = 2P(t_{n-2} > |t|)$

Conclusion:
If p-value $< \alpha$, reject H_0
If p-value $\geq \alpha$, fail to reject H_0.

Measuring the Strength of the Linear Relationship

Rejecting $H_0 : \beta = 0$ only tells us that a significant linear association exists. It does not give us an idea of how *strong* this association is.

1. Correlation R = a measure of the strength and direction of the linear association between two quantitative variables

   R is unknown. We estimate R using r, the sample correlation.

2. R^2 = proportion of variation in y that is explained by its linear relationship with x.

   R^2 is unknown. Estimate R^2 using r^2.

Example 12.1 Continued

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   15.513     11.603   1.337 0.217989
Prod           7.978      1.223   6.522 0.000184 ***
```

(a) At the .05 level, use the R output to test $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$ where β is the slope of the regression line between box office sales and production costs.

Assumptions: scatterplot indicates linear relationship between box office sales and production cost is a valid assumption. Assuming normality, constant variance and random samples.

Test Statistic: $t = \dfrac{b}{se(b)} =$

Example 12.1 Continued

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   15.513     11.603   1.337 0.217989
Prod           7.978      1.223   6.522 0.000184 ***
```

(a) At the .05 level, use the R output to test $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$ where β is the slope of the regression line between box office sales and production costs.

Assumptions: scatterplot indicates linear relationship between box office sales and production cost is a valid assumption. Assuming normality, constant variance and random samples.

Test Statistic: $t = \dfrac{b}{se(b)} = \dfrac{7.978}{1.223} = 6.522$

p-value: $2P(t_{n-2} > |t|) = 2P(t_8 > 6.522) = .000184$

Conclusion: p-value $< .05$. Reject H_0. We have strong evidence that a linear association exists between box office sales and production costs.

Example 12.1 Continued

(b) Calculate and interpret the 95% confidence interval for β.

$b \pm t_{\alpha/2,n-2} se(b) =$

Example 12.1 Continued

(b) Calculate and interpret the 95% confidence interval for β.

$$b \pm t_{\alpha/2,n-2} se(b) = 7.978 \pm 2.306(1.223) = 7.978 \pm 2.820 = (5.518, 10.798)$$

We are 95% confident that Box office sales increase somewhere between \$5.158 and \$10.798 million for \$1 million increase in production cost.

Example 12.1 Continued

```
Residual standard error: 14.26 on 8 degrees of freedom
Multiple R-squared:  0.8417,Adjusted R-squared:  0.8219
F-statistic: 42.54 on 1 and 8 DF,  p-value: 0.0001838
```

(c) What percentage of the variation in box office sales is accounted for by its linear relationship with production costs?

Example 12.1 Continued

```
Residual standard error: 14.26 on 8 degrees of freedom
Multiple R-squared:  0.8417,Adjusted R-squared:  0.8219
F-statistic: 42.54 on 1 and 8 DF,  p-value: 0.0001838
```

(c) What percentage of the variation in box office sales is accounted for by its linear relationship with production costs?

84.17%

Checking Assumptions

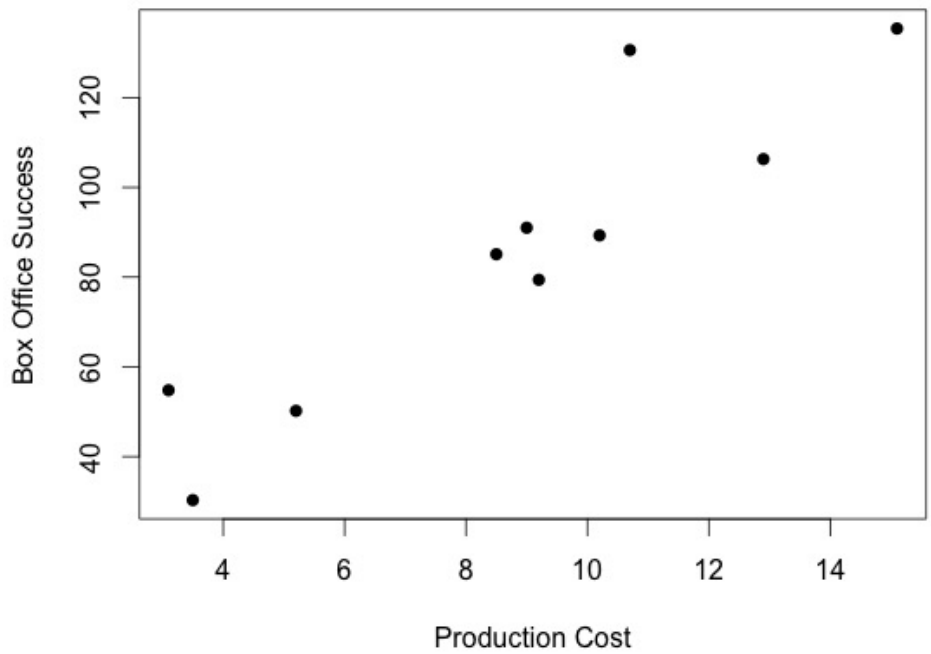
1. **Linearity:** There is a linear relationship between y and x.
   Check: the scatterplot to see if relationship looks linear

2. **Normality:** The distribution of the y values at values of x is $N(\alpha + \beta x, \sigma)$.
   Check: A Q-Q plot of the residuals

3. **Constant Variance:** the standard deviation of y, σ, is the same for every value of x
   Check: plot of residual versus fitted values.

Example 12.1: Checking Assumptions

Linearity: Scatterplot

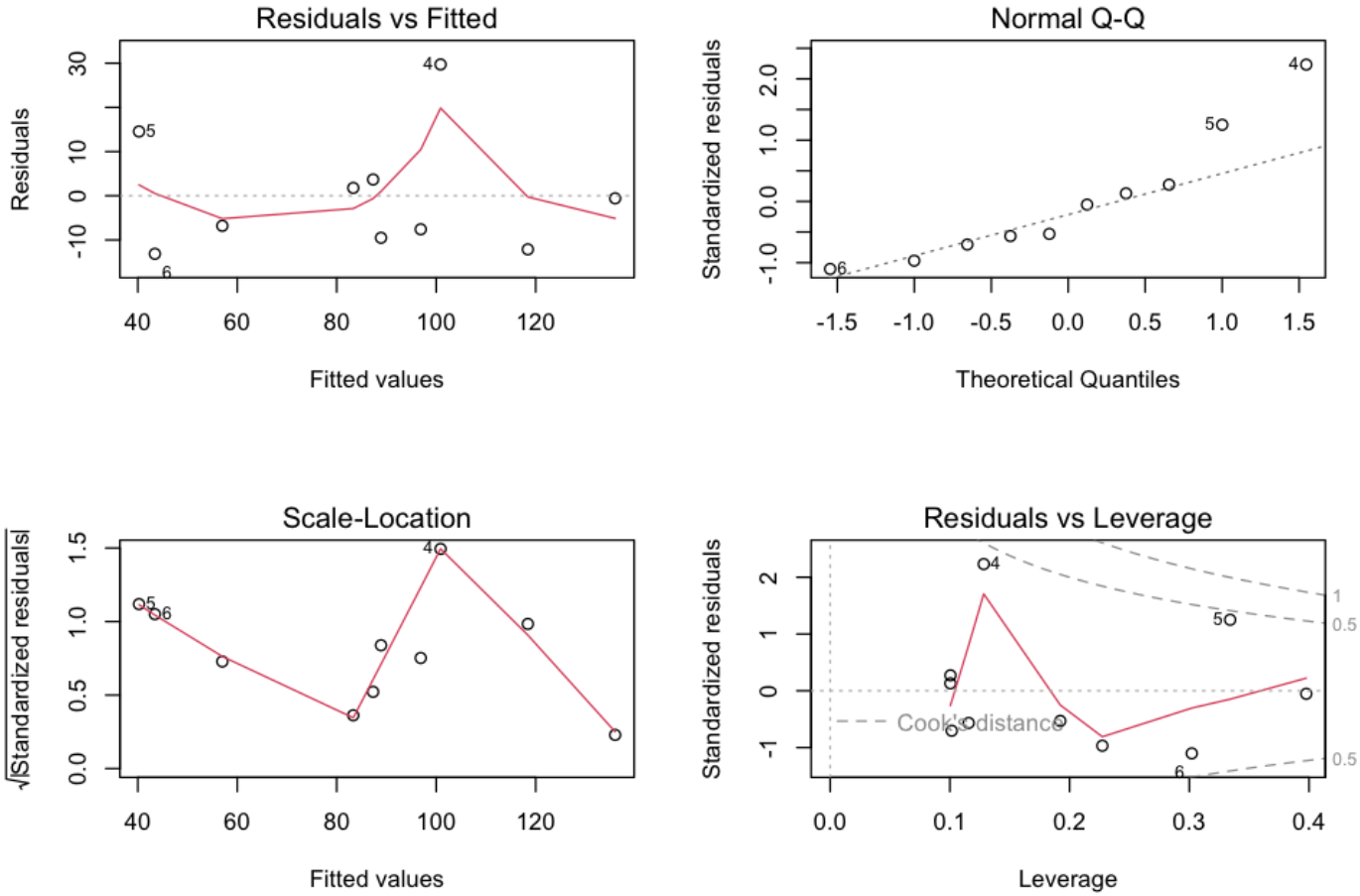


It looks like a straight line would fit this scatterplot well enough.

Example 12.1: Checking Assumptions

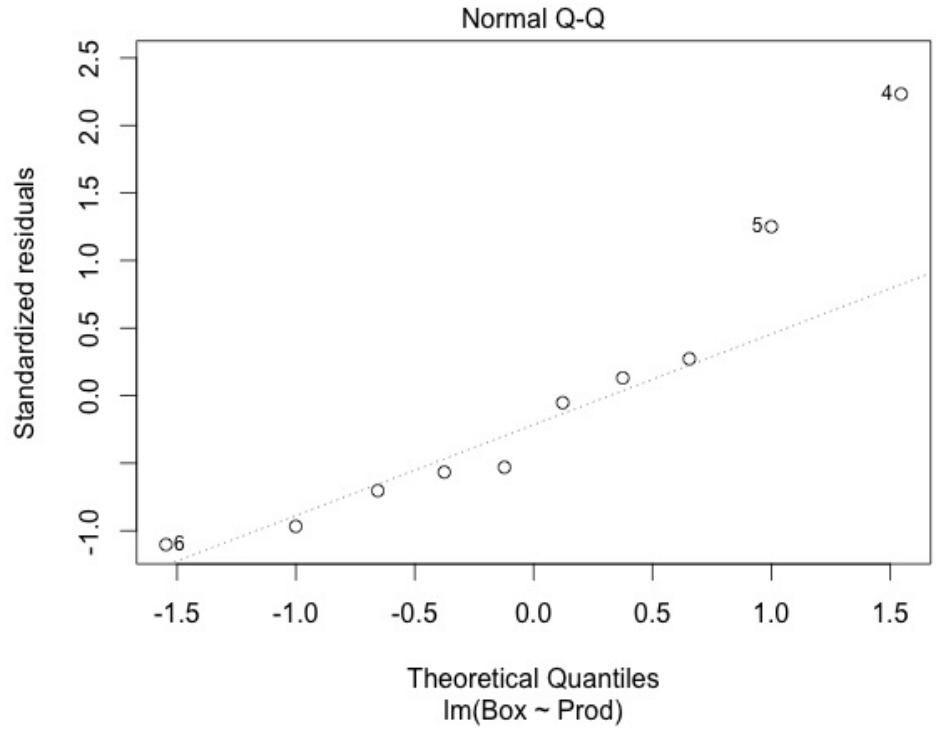
Pass your model output, in this case `mod`, to the `plot()` function to generate the plots needed to check your model assumptions.

```
> par(mfrow = c(2, 2))
> plot(mod)
> par(mfrow = c(1, 1))
```



Example 12.1: Checking Assumptions

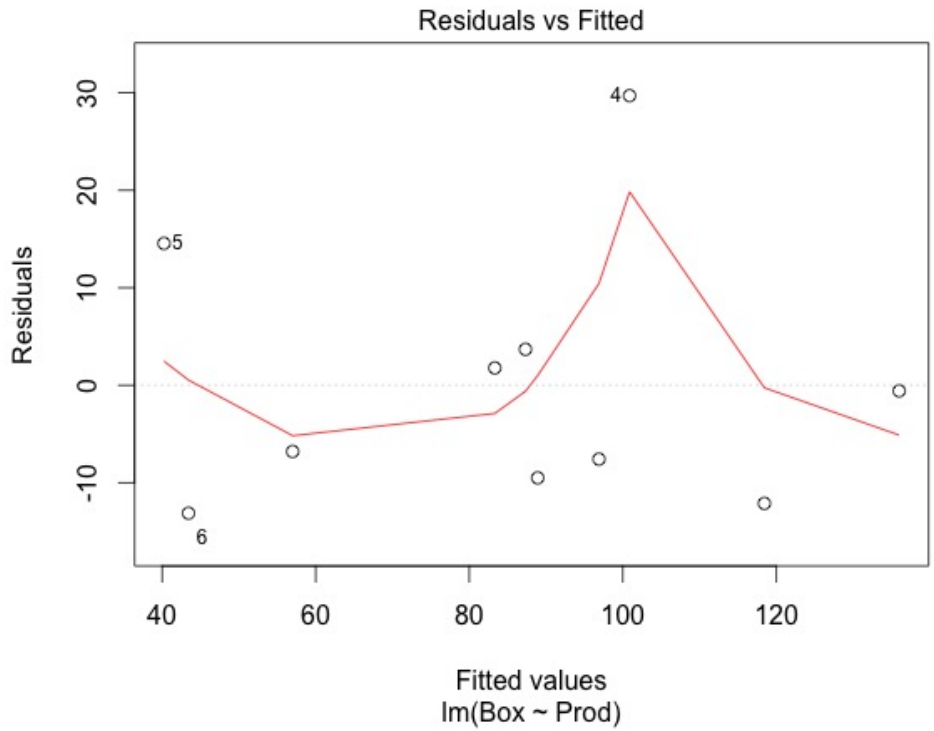
Normality: Check Q-Q plot of residuals



Remember Q-Q plots should have points on the line. The assumption of normality is only approximately satisfied since there are 2 points at the top which deviate from the line.

Example 12.1: Checking Assumptions

Constant Variance: Check the scatterplot of residuals versus fitted values



We want to see points equally far apart (ignore the red line). The constant variance assumption is approximately satisfied since the points seem equally far apart.

Correlation and Regression: Caution

Keep in mind the following when reading results

1. Correlation and regression only describe the **linear** relationships

2. Correlation and regression are **not** resistant to outliers

3. Regression lines should not be extrapolated far outside the range of observed data.

   Extrapolation is the use of a regression line to predict values of the response variable for values of the explanatory variable that are far outside the observed range of the data.

   This is a problem since we don't know if the linear trend continues beyond the observed range of data.

Example 12.2 Continued

We previously fit the following least squares regression line that describes the linear relationship between Olympic winning time of the 200m dash (Time) and the year in which the Olympics took place (Year):

$$\widehat{\text{Time}} = 22.355 - 0.0303 \text{ Year}$$

To convince yourself that extrapolating is the wrong thing to do, use the fitted line to predict the winning dash time in the 2016 and 2640 Olympics.

For the 2016 Olympics:

$$\widehat{\text{Time}} = 22.355 - 0.0303 \text{ Year} = 22.355 - 0.0303(2016 - 1900) = 18.84$$

For the 2640 Olympics:

$$\widehat{\text{Time}} = 22.355 - 0.0303 \text{ Year} = 22.355 - 0.0303(2640 - 1900) = -0.07$$

The winning time cannot be negative!

Correlation and Regression: Caution

4. Correlation does not necessarily imply causation

   Suppose r is the correlation between two quantitative variables x and y. A value close to -1 or 1 means they are associated, does not mean one causes the other! There is often another variable, that can explain this association.

   Lurking Variable: a variable which is not included in the study but strongly affects the relationship among the variables of primary interest. It may either falsely suggest a strong relationship or hide an important existing relationship.

Example 12.3

Based on United Nations data, it can be shown that there is a strong, negative correlation between a nation's infant mortality rate (IMR) and the per capita television ownership. How can this be explained?

Nobody actually thinks that watching more TV causes an improvement in infant health!

Lurking variable: National GDP

People in wealthier countries tend to have healthier babies and own more TVs. Thus, association does not imply causation!

How can we establish causation?

1. Set up a carefully designed experiment that controls for potential lurking variables
2. Evaluate criteria for establishing causation:
   (a) association is strong
   (b) association is consistent across many studies
   (c) alleged cause is plausible
   (d) alleged cause preceded the effect (in time)
   (e) higher doses evoke stronger responses.