slides originally by
Dr. Richard Burns,
modified by
Dr. Stephanie Schwartz

LINEAR REGRESSION

CSCI 452: Data Mining

Linear Regression

- *What*: for predicting a quantitative variable
- *Age*: "it's been around for a long time"
- *Complexity*: somewhat dull compared to more modern statistical learning techniques
- *Popularity*: still widely used

Fancier, modern, data mining approaches can be seen as generalization or extensions of Linear Regression.

Advertising Dataset

- Sales totals for a product in 200 different markets
  - Advertising budget in each market, broken down into TV, radio, newspaper

Advertising Dataset

```
> head(Advertising)
  X    TV Radio Newspaper Sales
1 1 230.1  37.8      69.2  22.1
2 2  44.5  39.3      45.1  10.4
3 3  17.2  45.9      69.3   9.3
4 4 151.5  41.3      58.5  18.5
5 5 180.8  10.8      58.4  12.9
6 6   8.7  48.9      75.0   7.2
```

Advertising Dataset

- *Goal:* What marketing plan for next year will result in high product sales?
- *Questions:*
  1. Is there a relationship between advertising budget and sales?
  2. How strong is the relationship between advertising budget and sales?
     - <u>Strong relationship:</u> given the advertising budget, we can predict sales with a high level of accuracy
     - <u>Weak relationship:</u> given the advertising budget, our prediction of sales is only slightly better than a random guess

Advertising Dataset

- *Goal:* What marketing plan for next year will result in high product sales?
- *Questions:*

3. Which media contribute to sales?
   - Need to separate the effects of each medium
4. How accurately can we estimate the effect of each medium on sales?
   - For every dollar spent on advertising in a particular medium, by what amount will sales increase? How accurately can we predict this increase?

Advertising Dataset

- *Goal:* What marketing plan for next year will result in high product sales?
- *Questions:*
  5. Is the relationship linear?
     - If the relationship between advertising budget and sales is a straight-line, then linear regression seems appropriate.
     - If not, all is not lost yet. (<u>Variable Transformation</u>)
  6. Is there any <u>interaction effect</u>? *(called "synergy" in business)*
     - *Example:* spending 50k on TV ads + 50k on radio ads results in more sales than spending 100k on only TV

Simple Linear Regression

- Predicting quantitative response *Y* based on a <u>single predictor variable</u> *X*
- Assumes linear relationship between *X* and *Y*

$$Y \approx B_0 + B_1 X$$

read $\approx$ as "is approximately modeled as"

"we are regressing *Y* onto X"

Simple Linear Regression

- Two unknown constants
  - Also called "model coefficients" or "parameters"

β_0 = intercept
β_1 = slope

$$Y \approx \beta_0 + \beta_1 X$$

- Use training data to produce estimates for the model coefficients:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

In practice, β_0 and β_1 are unknown.

Estimating the Coefficients

- Goal is to obtain coefficient estimates such that the linear model fits the available data well
  - To find an intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$ such that the resulting line is as close as possible to the data points
  - Q: How to determine "closeness"?
  - A: *Common approach:* <u>least squares</u>

Residual Sum of Squares (RSS)

- Prediction for *Y* based on the *i*th value of *X*

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- *i*th <u>residual</u>: difference between the *i*th observed response value and the *i*th predicted value

$$e_i = y_i - \hat{y}_i$$

- <u>Residual Sum of Squares (RSS):</u>

$$RSS = e_1^2 + e_2^2 + \ldots + e_n^2$$

Least Squares

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- <u>Residual Sum of Squares (RSS):</u>

$$RSS = e_1^2 + e_2^2 + \ldots + e_n^2$$

- <u>Least Squares:</u> chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS

Least Squares

- *Using some calculus to minimize the RSS, we get:*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

sample means:

$$\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$$

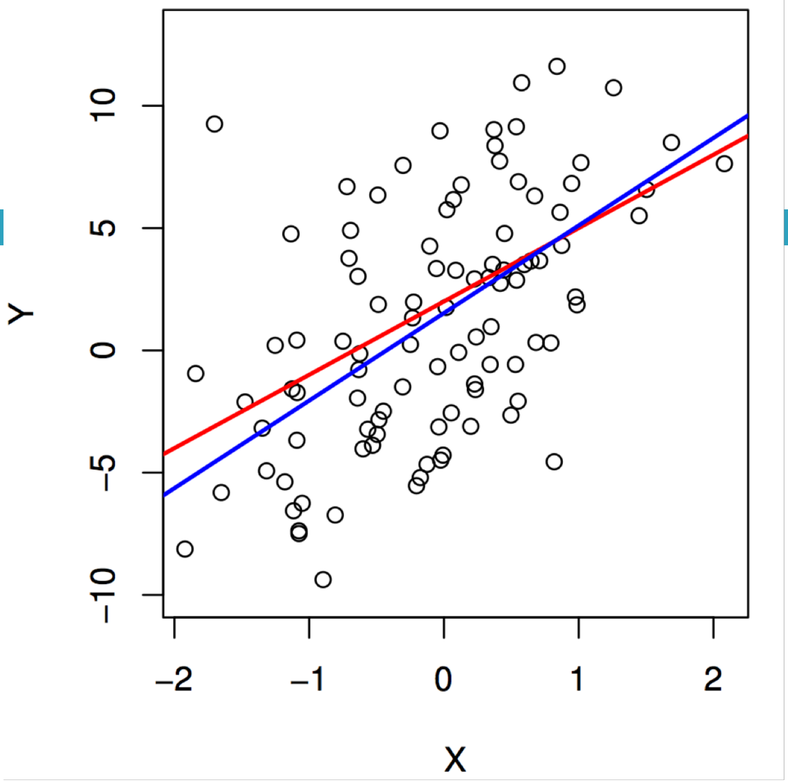
Simulated Example

- Population Regression Line:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Least Squares Line:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$



Red line: $Y=2+3X+\varepsilon$

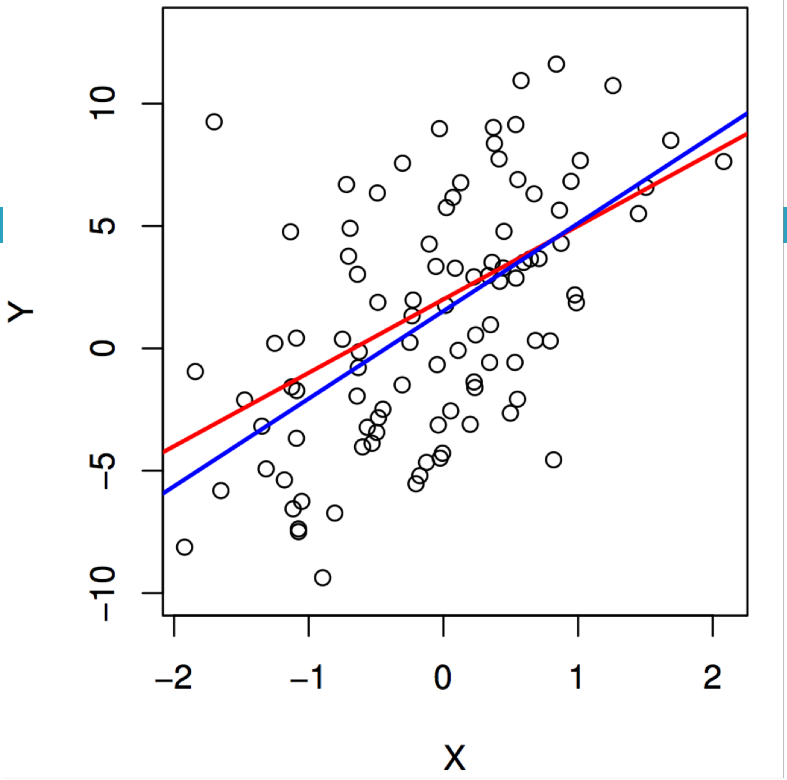
Blue line: least squares estimate based on observed data

Simulated data:

- 100 random *X*s
- 100 corresponding *Y*s from the model
- ε generated from normal distribution

Simulated Example

- True relationship "of the <u>population</u>" (red line) not usually known for real data
- Depending on the set of observations, "the <u>sample</u>", the estimated coefficients and model will change



Red line: $Y=2+3X+\varepsilon$

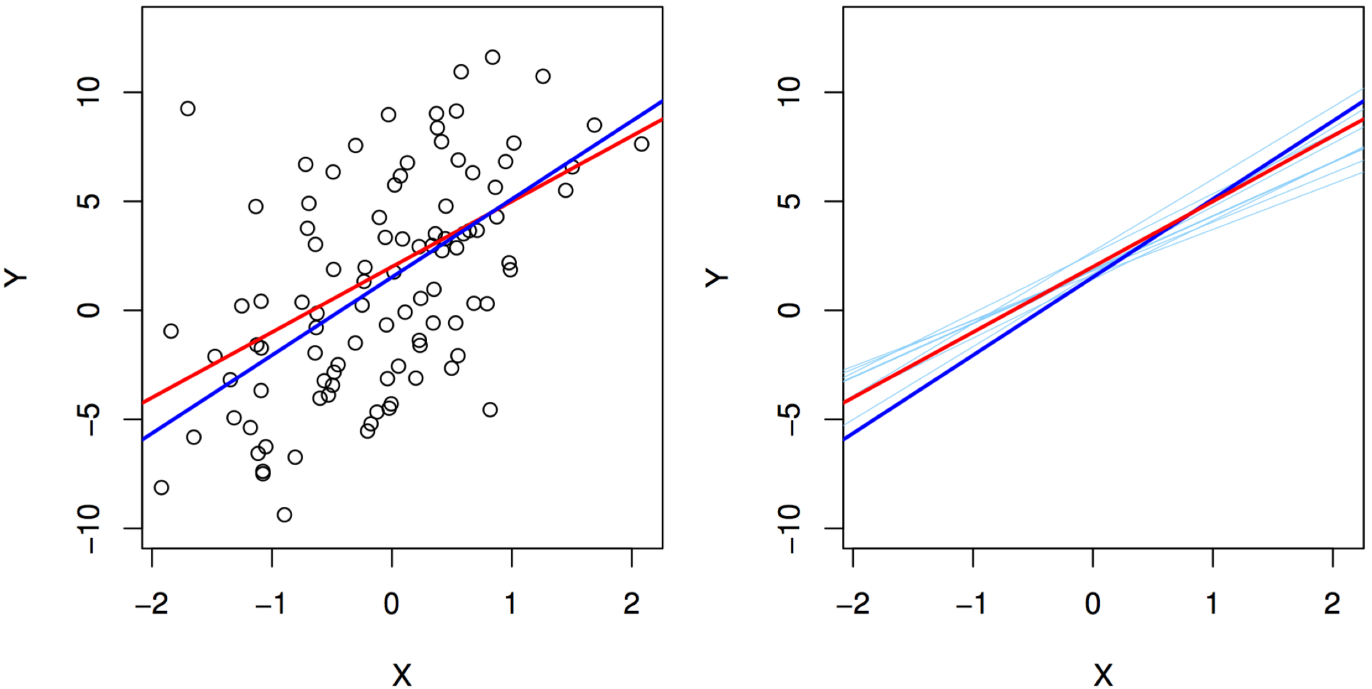
Blue line: least squares estimate based on observed data

Simulated data:

- 100 random *X*s
- 100 corresponding *Y*s from the model
- ε generated from normal distribution

Simulated Example

Red line is the "true relationship" in the population



Red line doesn't change

Right graph: ten least square lines (blueish), each for a different simulation of the red line.

Because of the error term, the "sample" data points are different for each simulation.

Advertising Dataset

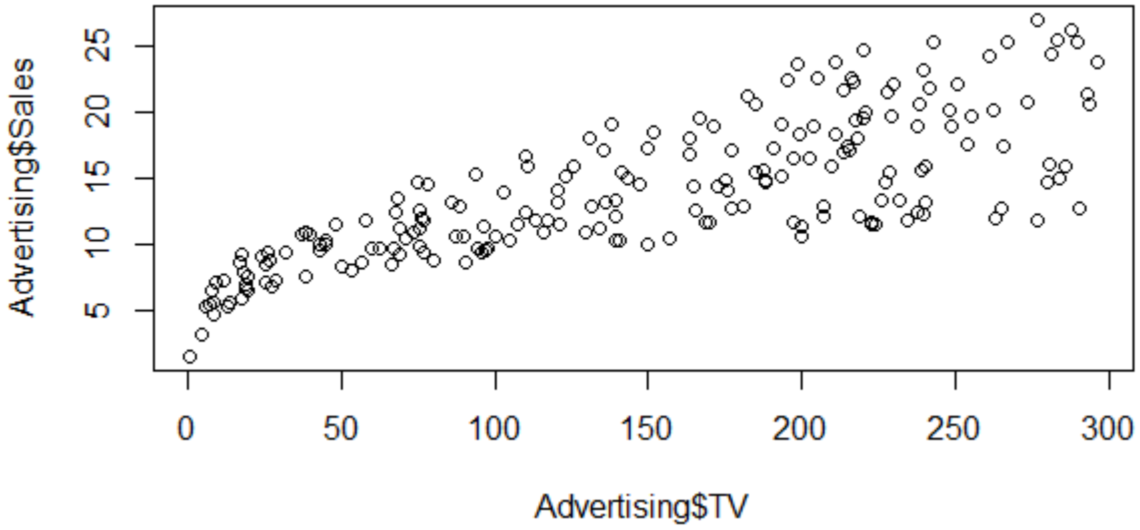
- What are some ways we can regress sales onto adverting using Simple Linear Regression?
- One model:

$$sales \approx \beta_0 + \beta_1 \times TV$$

$$Y \approx B_0 + B_1 X$$

Advertising Dataset

- Scatter plot visualization for TV and Sales.




```
> plot(Advertising$Sales ~ Advertising$TV)
```

Advertising Dataset

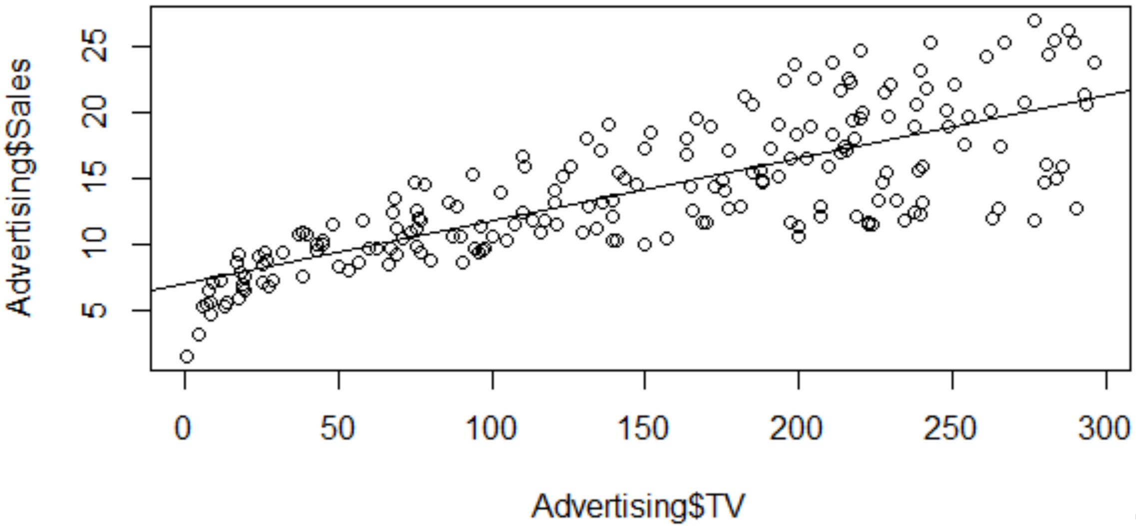
- Simple Linear Model in R:
  - *General form:* `lm(y~x, data)`
  - *Predictor: x*
  - *Response: y*

```
> lm(Advertising$Sales ~ Advertising$TV)

Call: lm(formula = Advertising$Sales ~ Advertising$TV)

Coefficients:
        (Intercept)        Advertising$TV
          7.03259                 0.04754
```

*Sales = 7.03259 + 0.04754 * TV*

Advertising Dataset

- Scatter plot visualization for TV and Sales with Linear Model.




```
> lm.fit=lm(Advertising$Sales ~ Advertising$TV)
> abline(lm.fit)
```

"a b line" – draw line of intercept *a* and slope *b*

Simple Linear Model

- Our assumption was that the relationship between X and Y took the form:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Expected value when $X=0$ is β_0
- Average increase in Y when there is a one-unit increase in X is β_1
- Error term: what model misses, measurement error, etc.

Assessing the Accuracy of the Model

- Trying to quantify the extent to which the model fits the data (so we can draw conclusions about population)
  - Typically assessed with:
    1. Residual standard error (RSE)
    2. R^2 statistic
- Different than measuring how good the model's predictions were on a <u>test set</u>
  - Root Mean Squared Error (RMSE)

Measuring the Quality of a Regression Model

Residual Standard Error

$$RSE = \sqrt{RSS/(n-2)}$$

- *(RSS "Residual Sum of Squares" sometimes called SSE "Sum of Squared Errors")*
- *(RSE "Residual Standard Error" sometimes called "Standard Error of the Estimate" or "**Residual Standard Deviation**" – it is the estimated standard deviation of the residuals)*
- *We use RSE because the standard deviation is unknown, so we can't calculate SE*

Cereal Dataset

- http://lib.stat.cmu.edu/DASL/Datafiles/Cereals.html
- From CMU Data and Story Library
- 77 cereals
- 15 Attributes: calories, sugar content, protein, etc.
- Target: Consumer Reports "Health Rating" (continuous)

Cereal - R

- Residual Sum of Squares
- Residual Standard Error
- Using Linear Model to Predict Value

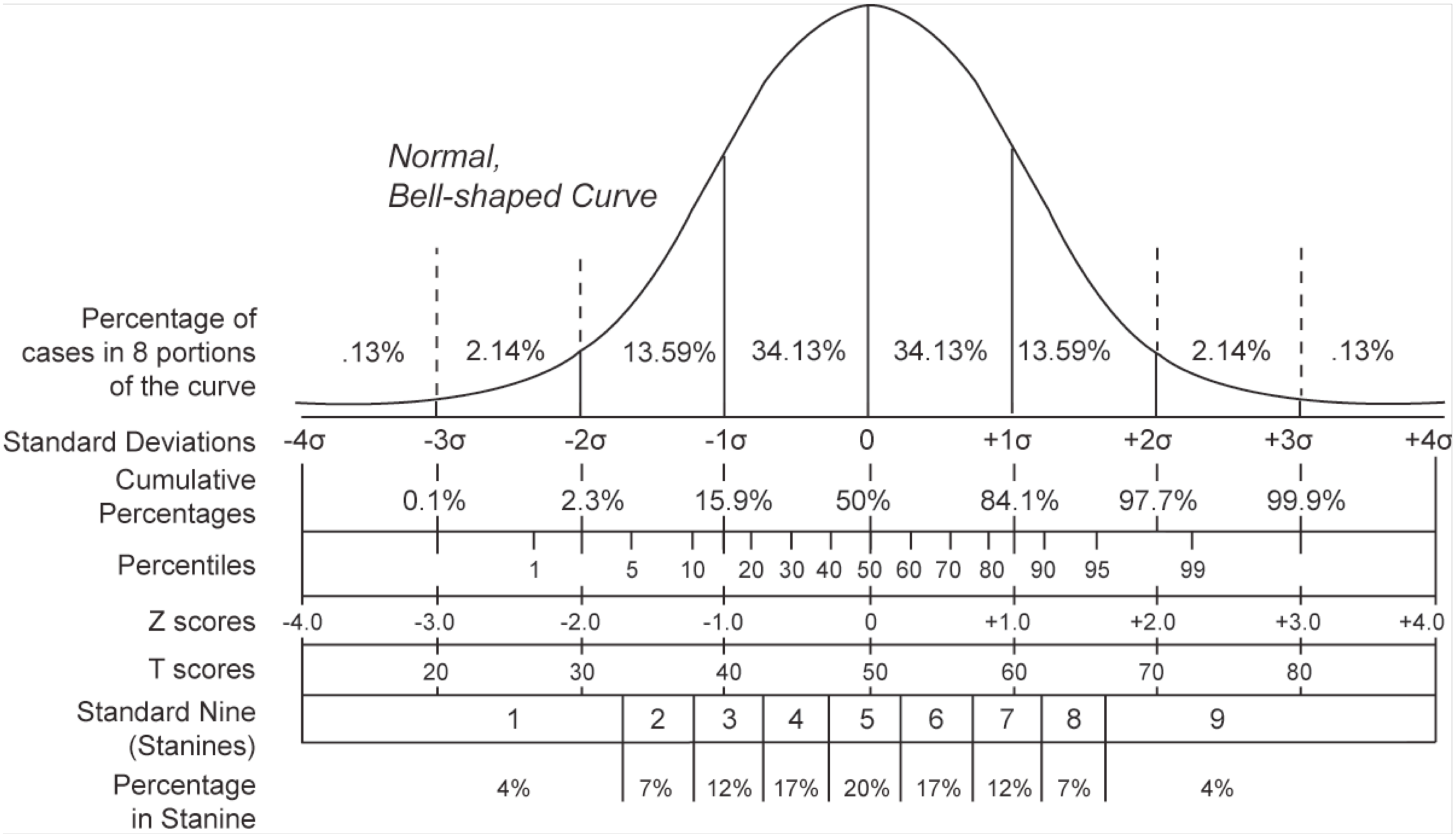
Example

- Cereal Dataset

$$RSE = \sqrt{RSS/(n-2)} = \sqrt{6342/75}$$

$RSE = 9.196 \approx 9.2$

  - *"Typical error in predicting nutritional rating will be about 9.2 points."*
  - *"Estimate of the new cereal's rating will be within 9.2 points about 68% of the time." (68% because it is one standard deviation)*


Normal,
Bell-shaped Curve
Percentage of cases in 8 portions of the curve
.13%
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
.13%
Standard Deviations
-4σ
-3σ
-2σ
-1σ
0
+1σ
+2σ
+3σ
+4σ
Cumulative Percentages
0.1%
2.3%
15.9%
50%
84.1%
97.7%
99.9%
Percentiles
1
5
10
20
30
40
50
60
70
80
90
95
99
Z scores
-4.0
-3.0
-2.0
-1.0
0
+1.0
+2.0
+3.0
+4.0
T scores
20
30
40
50
60
70
80
Standard Nine (Stanines)
1
2
3
4
5
6
7
8
9
Percentage in Stanine
4%
7%
12%
17%
20%
17%
12%
7%
4%

Confidence Intervals for Linear Regression

- Takes the form:

$$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$$

$$[\hat{\beta}_1 - 2 \cdot SE(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot SE(\hat{\beta}_1)]$$

- That is, there is a 95% chance that the true value is in the above range.
- Same form for β_0

Advertising Dataset

```
> Advertising <-
read.csv("C:/Users/75RBURNS/Dropbox/work/wcu/600-
DataMining/data/Advertising.csv")
> head(Advertising)
  X    TV Radio Newspaper Sales
1 1 230.1  37.8      69.2  22.1
2 2  44.5  39.3      45.1  10.4
3 3  17.2  45.9      69.3   9.3
4 4 151.5  41.3      58.5  18.5
5 5 180.8  10.8      58.4  12.9
6 6   8.7  48.9      75.0   7.2
```

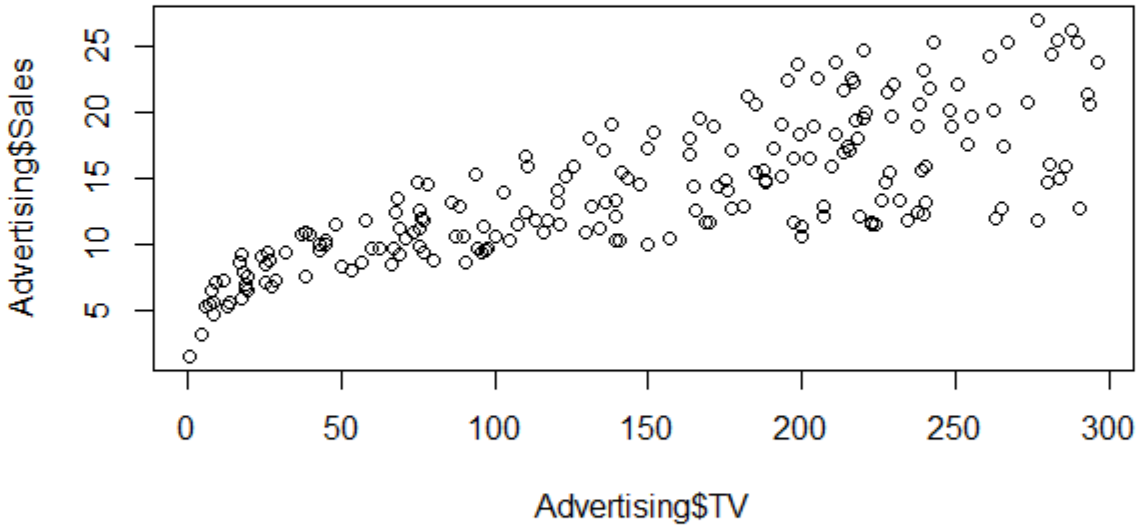
Simple Linear Regression Model for Advertising Dataset

$$sales \approx \beta_0 + \beta_1 \times TV$$

$$Y \approx B_0 + B_1 X$$

Advertising Dataset

- Scatter plot visualization for TV and Sales.




```
> plot(Advertising$Sales ~ Advertising$TV)
```

Advertising Dataset

- Simple Linear Model in R:
  - *General form:* `lm(y~x, data)`
  - *Predictor: x*
  - *Response: y*

```
> lm(Advertising$Sales ~ Advertising$TV)

Call: lm(formula = Advertising$Sales ~ Advertising$TV)

Coefficients:
      (Intercept)      Advertising$TV
        7.03259             0.04754
```

*Sales = 7.03259 + 0.04754 * TV*

Advertising:

*Sales = 7.03259 + 0.04754 * TV*

- 95% confidence interval for B_0 is [6.130, 7.935]
- 95% confidence interval for B_1 is [0.042, 0.053]
- *Prediction:* if TV = 30, then $Sales = 7.03259 + 0.04754(30) = 8.45879$

- In the absence of any advertising, sales will, on average, fall somewhere between 6,130 and 7,940 units.
- For each $1,000 increase in television advertising, there will be an average increase in sales between 42 and 53 units.

Advertising Dataset

- RSE = 3.26

Actual sales in each market deviate from the true regression line by approximately 3.26 units, on average.

- Is this error amount acceptable?
  - *Business answer:* depends on problem context
  - Worth noting the percentage error:

$$\text{Percentage Error} = \frac{RSE}{\text{mean } sales} = \frac{3.258656}{14.0225} = 0.23238 = 23.2\%$$

Assessing the Accuracy of the Model

- Trying to quantify the extent to which the model fits the data (so we can draw conclusions about population)
  - Typically assessed with:
    1. Residual standard error (RSE)
    2. R^2 statistic
- Different than measuring how good the model's predictions were on a <u>test set</u>
  - Root Mean Squared Error (RMSE)

R^2 Statistic

- <u>Proportion</u> of variance explained
  - Always a value between 0 and 1
  - Independent of the scale of *Y* (unlike RSE)

$$R^2 = \frac{TSS - RSS}{RSS} = 1 - \frac{RSS}{TSS}$$

$$TSS = \sum (y_i - \bar{y})^2$$

$$RSS = \sum (y_i - \hat{y}_i)^2$$

R^2 Statistic

$$R^2 = \frac{TSS - RSS}{RSS} = 1 - \frac{RSS}{TSS}$$

$$TSS = \sum (y_i - \bar{y})^2$$

$$RSS = \sum (y_i - \hat{y}_i)^2$$

- TSS: total variance in the response *Y*
  - Amount of variability inherent in the response, before the regression is performed
- RSS: amount of variability that is left unexplained after performing the regression
- TSS-RSS : the amount of variability that is explained

Advertising Dataset

- $R^2 = 0.61$

Just under two-thirds of the variability in *sales* is explained by a linear regression on *TV*.

Interpreting R^2 values

- R^2 is a measurement of the linear relationship between X and Y
- R^2 has an interpretational advantage over RSE in that it doesn't depend on the units of Y
- Q: What is a good R^2 value?
  A: Depends on the application, of course.
  - *Example*: problem from physics where it is known that a linear relationship exists, can expect a good R^2 value
  - *Example*: other domains where linear model is rough approximation…

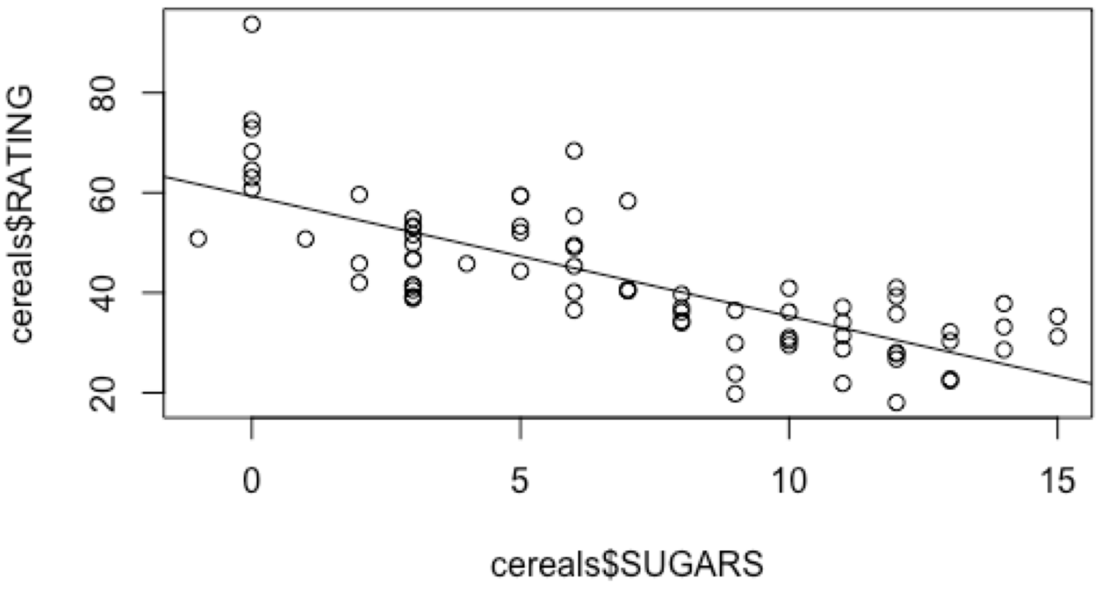
Assessing the Accuracy of the Model

- Trying to quantify the extent to which the model fits the data (so we can draw conclusions about population)
  - Typically assessed with:
    1. Residual standard error (RSE)
    2. R^2 statistic
- Different than measuring how good the model's predictions were on a <u>test set</u>
  - Root Mean Squared Error (RMSE)

Predicting Values for New Data

- Confidence Interval: What is the confidence interval for the *expected value* of y given x (the new data)
- Prediction Interval: What is the interval into which you would expect the individual data points to fall?

Confidence Interval vs Prediction Interval



Confidence Interval vs. Prediction Interval

- Analogy:
  - Trying to predict baseball batting average
  - “Team” batting averages (mean of the player batting averages on that team), for all 30 teams, have low variance
    - Should be easier to predict a team batting average
  - “Individual batting averages are quite varied
    - Estimate of team average will be more precise than an estimate of a randomly chosen player, for the same level of confidence

Cereal - R

- Computing Confidence Interval using LM
- Computing Prediction Interval using LM

Evaluating the LM using a Test Set

- Given a set of predictions for *m* new cases *for which we have results* (a test set), we can evaluate the model's predictions by:
  1. Mean Error (ME)
  2. Root Mean Square Error (RMSE)

Mean Error

- Mean error should be close to zero
- Mean errors different from zero indicate a bias in the model

$$ME = \left(\frac{1}{m}\right)\sum_{i=1}^{m}(y_i - \hat{y}_i)$$

Root Mean Square Error

- Root mean square error (vs mean square error) expresses the magnitude of the model's error in the units of the response variable

$$RMSE = \sqrt{\left(\frac{1}{m}\right)\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}$$

R Example

References

- *Data Mining and Business Analytics in R*, 1st edition, Ledolter
- *An Introduction to Statistical Learning*, 1st edition, James et al.
- *Discovering Knowledge in Data*, 2nd edition, Larose et al.