# Data Science: Correlation and Regression

Won Kim

2022

# Roadmap

- Correlation
- Regression
  - Linear Regression
  - Polynomial Regression
  - Multiple Regression

# Acknowledgments

- http://www.pitt.edu/~super4/33011-34001/33851.ppt
- https://www.fil.ion.ucl.ac.uk/mfd_archive/2005/Corr-and-Regress.ppt

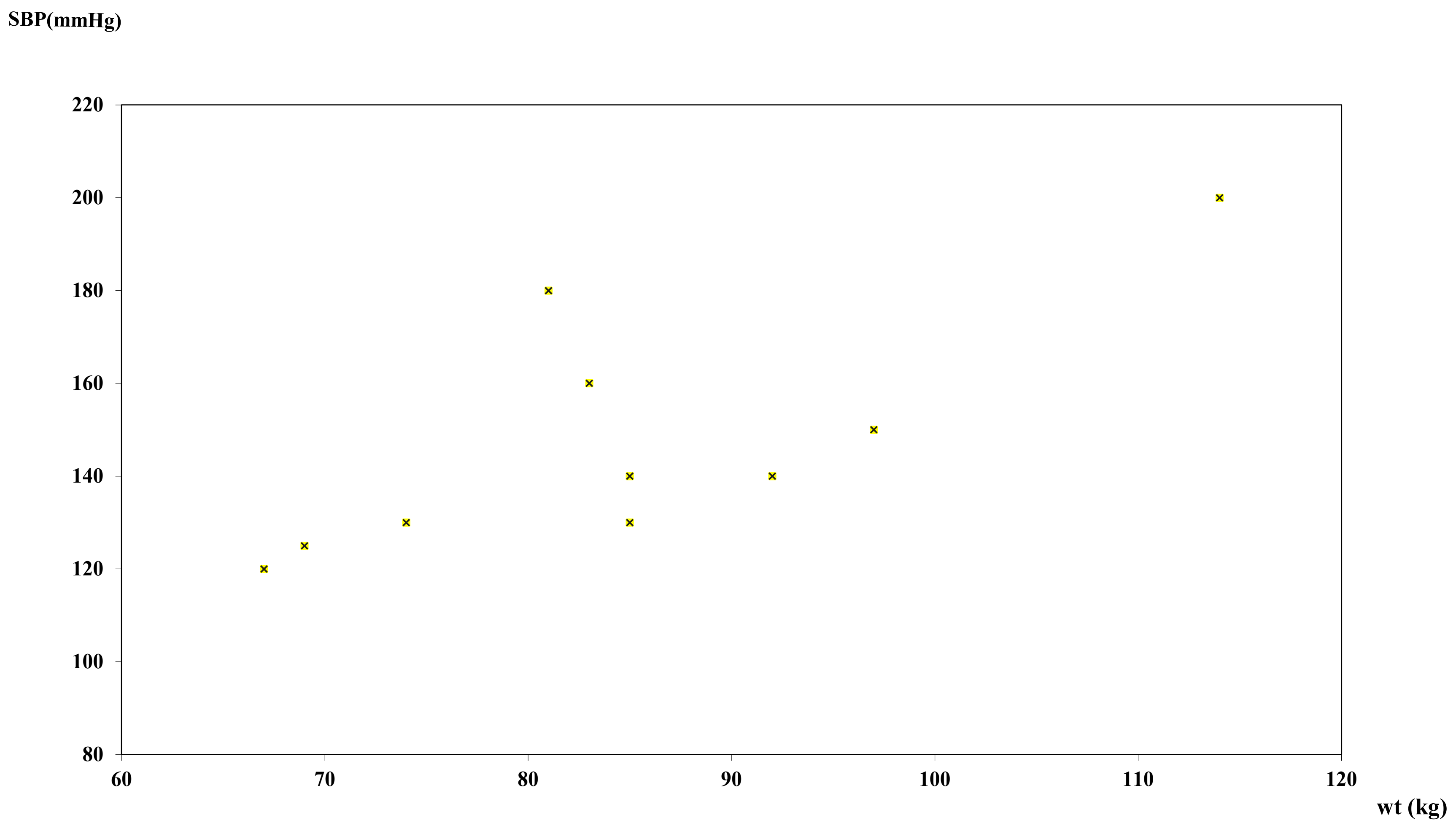
# Correlation

- Correlation is a statistical technique used to determine the degree to which two variables are related
- Finding the relationship between two quantitative variables without being able to infer causal relationships

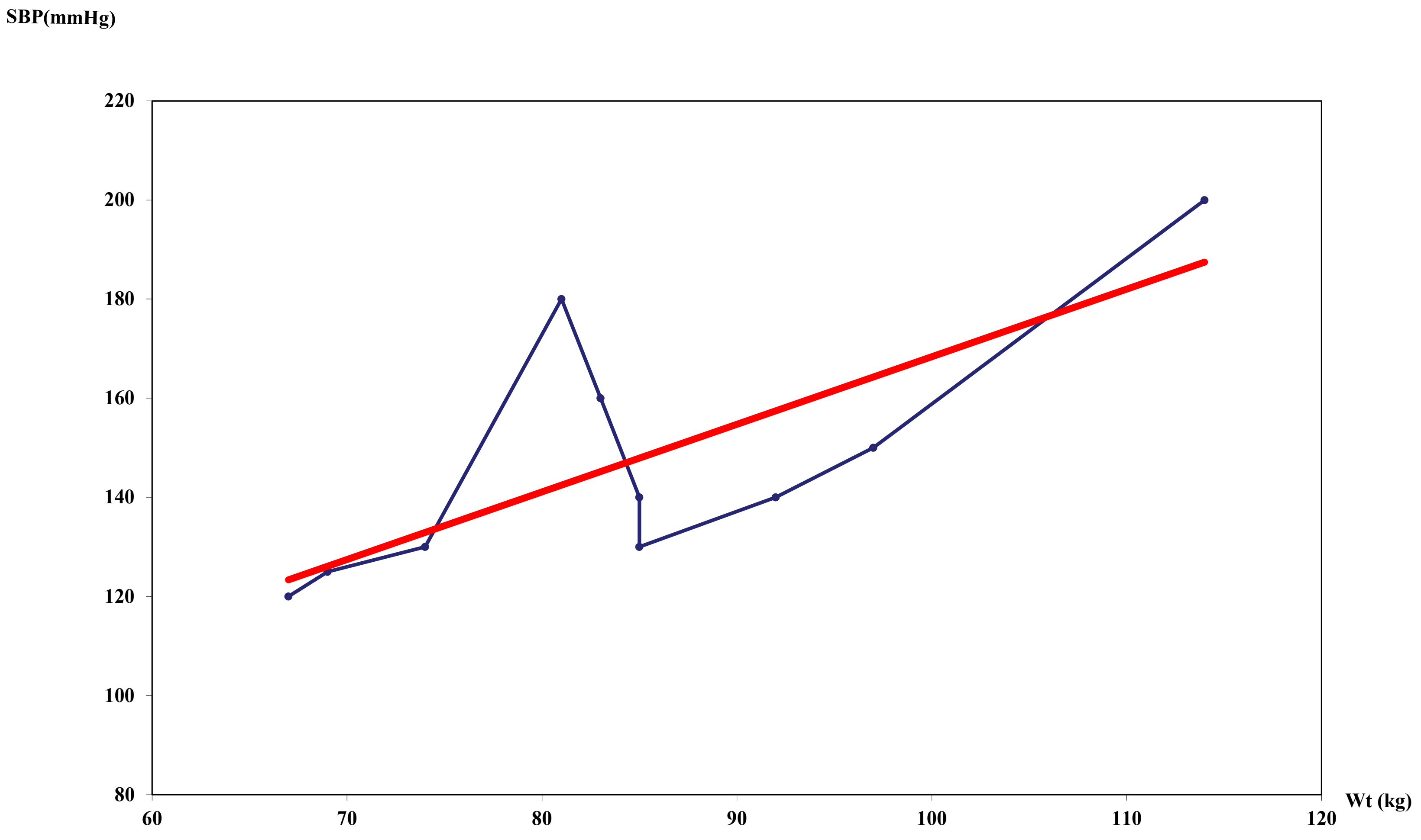
# Example: Weight vs. Systolic Blood Pressure

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Wt. (kg) | 67 | 69 | 85 | 83 | 74 | 81 | 97 | 92 | 114 | 85 |
| SBP mHg) | 120 | 125 | 140 | 160 | 130 | 180 | 150 | 140 | 200 | 130 |

# Scatter Plot of Weight and Systolic Blood Pressure



# Linear Regression Line

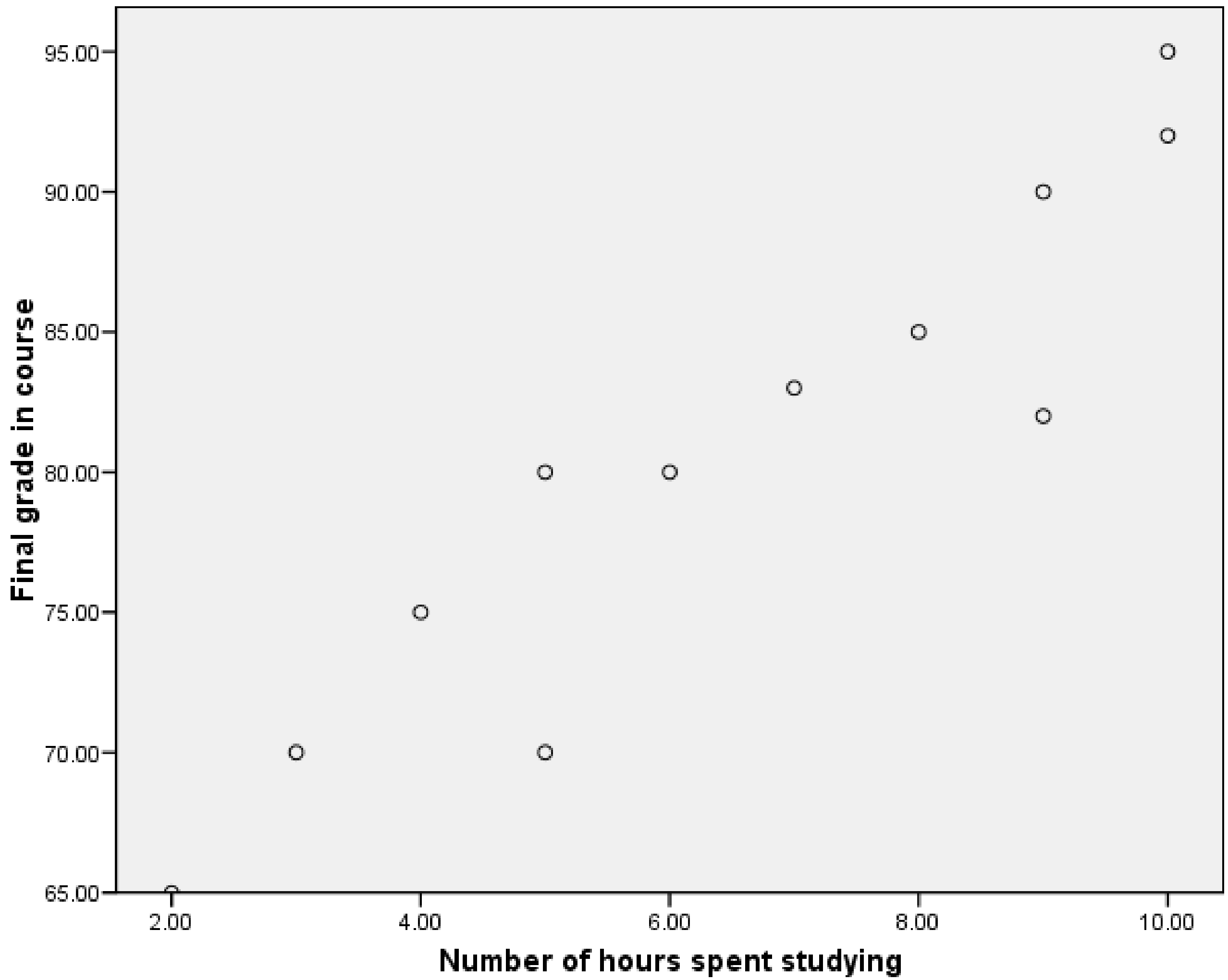


# Scatter Plots

- The pattern of data is indicative of the type of relationship between your two variables
  - positive relationship
  - negative relationship
  - no relationship

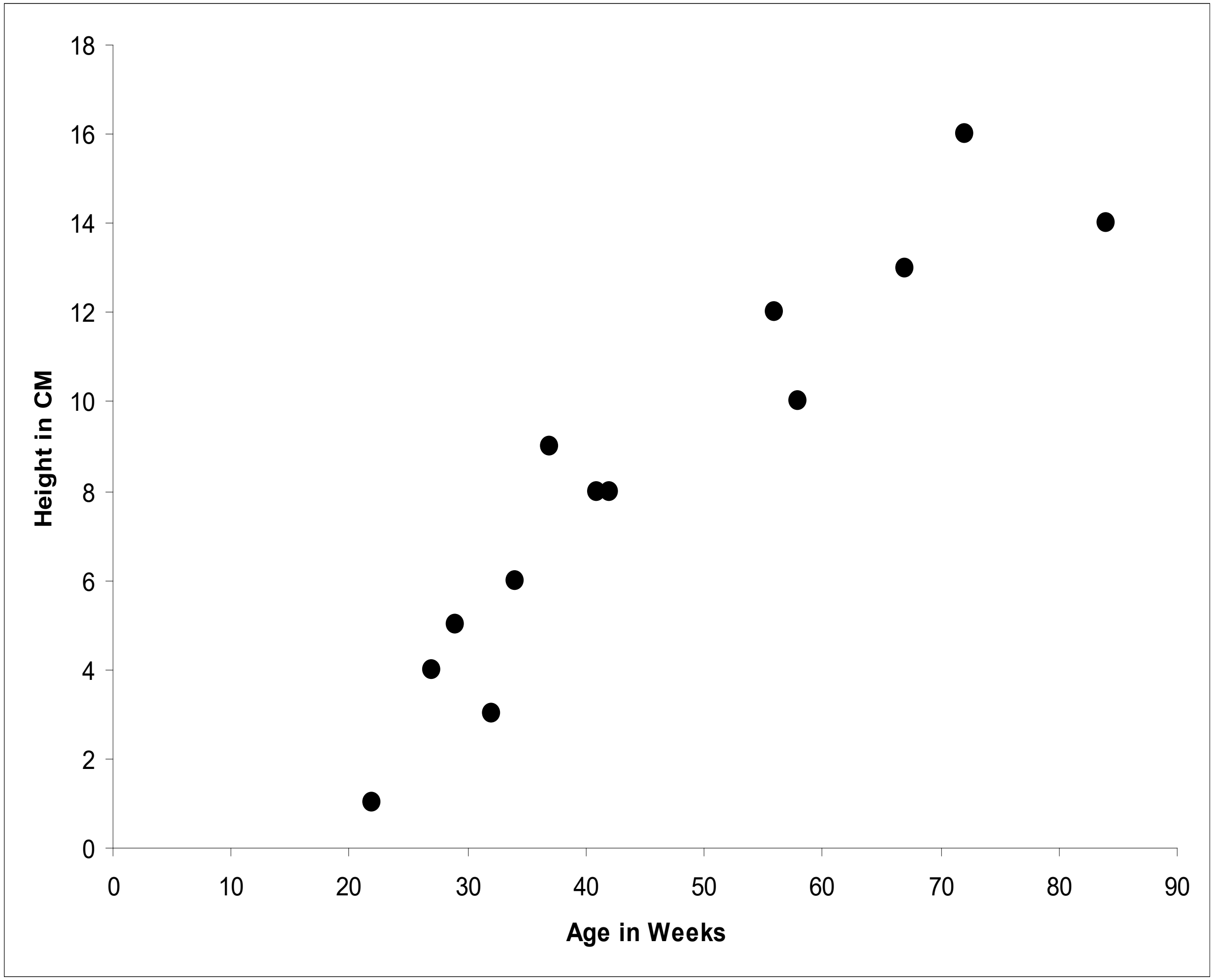
# Positive Relationship

- # of hours studying vs. grade



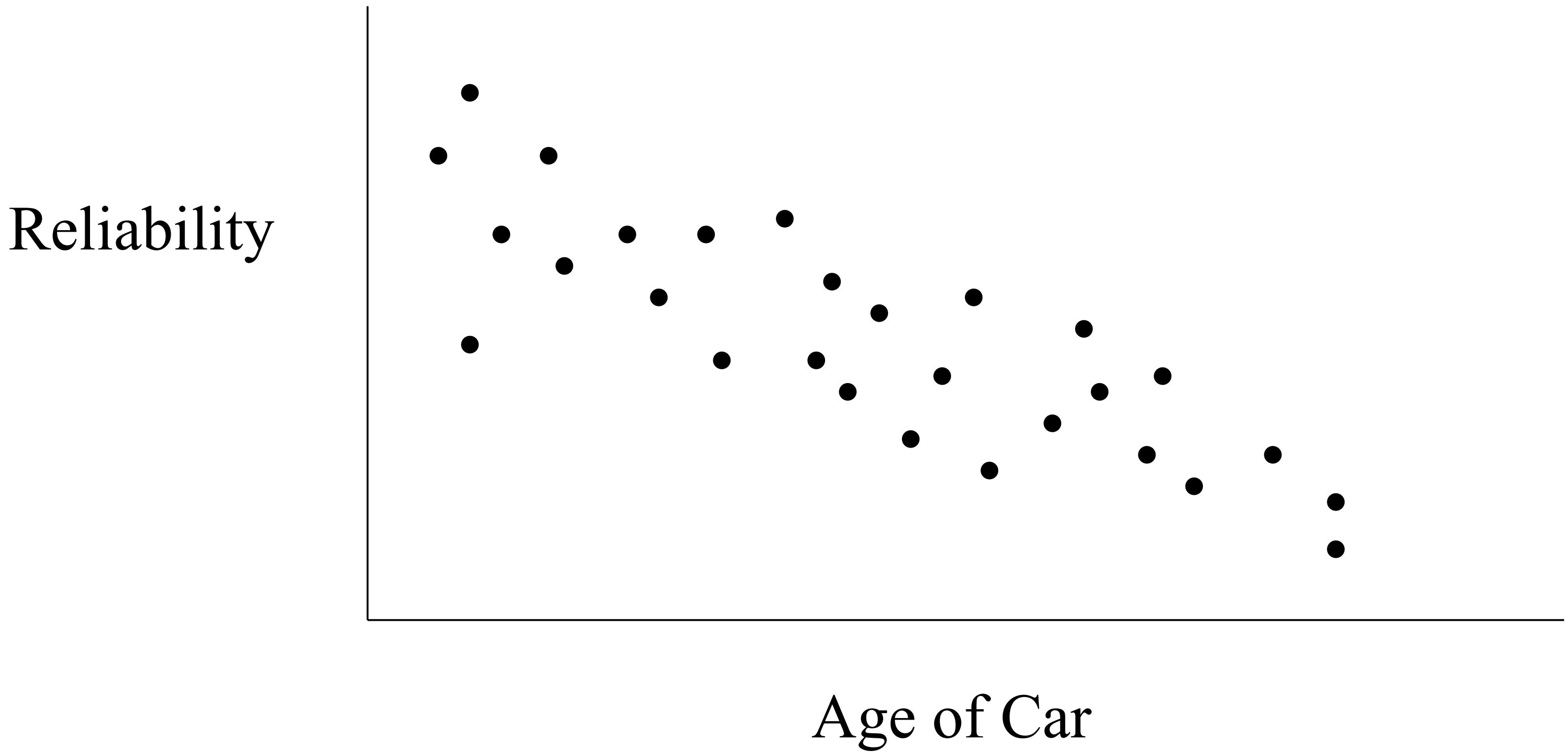
# Positive Relationship

- baby age in weeks vs. height



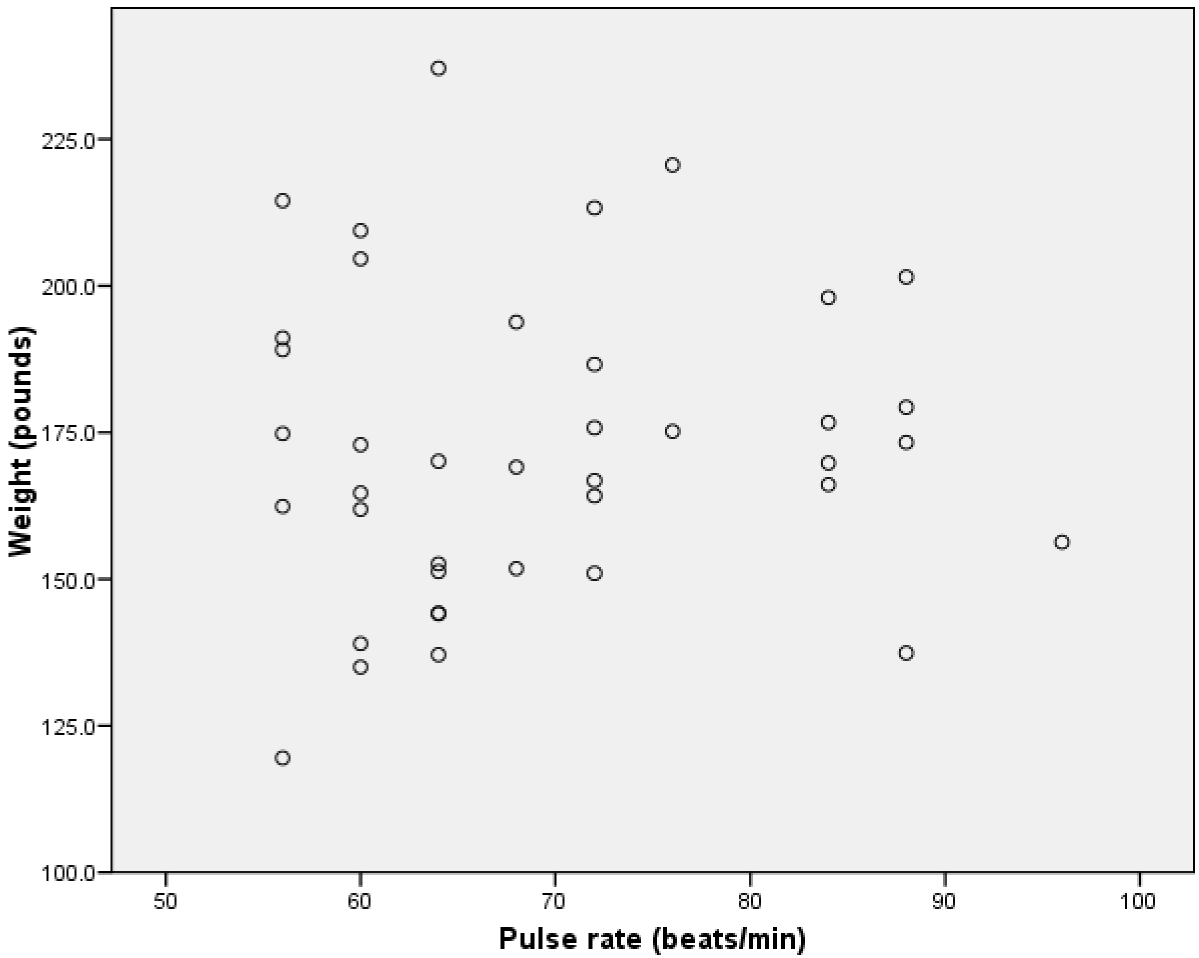
# Negative Relationship

- age of car vs. reliability



# No Relationship

- pulse rate vs. weight



# Variance and Covariance

- Variance
  - variability of a single variable

$$S_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- Covariance
  - degree to which two variables vary together

$$\mathrm{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Note:
(n-1) for sample
(n) for population

# Variance and Covariance

- Covariance is similar to variance
  - The equation simply multiplies x's error scores by y's error scores (not squaring x's error scores)

- When X⇧ and Y⇧ : cov (x,y) = pos.
- When X⇧ and Y⇩ : cov (x,y) = neg.
- When no constant relationship: cov (x,y) = 0

# Problem with Covariance

- The covariance value depends on the size of the data's standard deviations
- If large, the value will be greater than if small, even if the relationship between x and y is the same.

# How Covariance Depends on Variance

| | High Variance Data | | | | Low Variance Data | | |
|---|---|---|---|---|---|---|---|
| **Subject** | **x** | **y** | **x error * y error** | | **x** | **y** | **x error * y error** |
| 1 | 101 | 100 | 2500 | | 54 | 53 | 9 |
| 2 | 81 | 80 | 900 | | 53 | 52 | 4 |
| 3 | 61 | 60 | 100 | | 52 | 51 | 1 |
| 4 | 51 | 50 | 0 | | 51 | 50 | 0 |
| 5 | 41 | 40 | 100 | | 50 | 49 | 1 |
| 6 | 21 | 20 | 900 | | 49 | 48 | 4 |
| 7 | 1 | 0 | 2500 | | 48 | 47 | 9 |
| **Mean** | 51 | 50 | | | 51 | 50 | |
| **Sum of x error * y error :** | | | 7000 | | **Sum of x error * y error :** | | 28 |
| **Covariance:** | | | **1166.67** | | **Covariance:** | | **4.67** |

# Solutions: Correlation Coefficients

- Statistic showing the degree of relation between two variables
- Pearson's Coefficient  r
- Spearman's Coefficient  rho

# Pearson's r

- Pearson's r standardizes the covariance value.
- Divides the covariance by the multiplied standard deviations of X and Y.

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{s_x s_y}$$

# Pearson's r Formula

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right) \cdot \left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

| | | |
|---|---|---|
| $N$ | = | number of pairs of scores |
| $\Sigma xy$ | = | sum of the products of paired scores |
| $\Sigma x$ | = | sum of x scores |
| $\Sigma y$ | = | sum of y scores |
| $\Sigma x^2$ | = | sum of squared x scores |
| $\Sigma y^2$ | = | sum of squared y scores |

# Pearson's r (cont'd)

- The value of r denotes the strength of association.
  - The value of r ranges between ( -1) and ( +1)
- The sign of r denotes the nature of association
  - If the sign is +, the relationship is direct
  - If the sign is -, the relationship is inverse or indirect

# Strength of Relationship

- If $r = 0$, no association or correlation between the two variables
- If $0 < r < 0.25$, weak correlation
- If $0.25 \le r < 0.75$, intermediate correlation
- If $0.75 \le r < 1$, strong correlation
- If $r = 1$, perfect correlation

# Example 1

- A sample of 6 children was selected.
- Data about their age in years and weight in kilograms were recorded as shown below .
- We want to find the correlation between age and weight.

| serial No | Age (years) | Weight (Kg) |
|---|---|---|
| 1 | 7 | 12 |
| 2 | 6 | 8 |
| 3 | 8 | 12 |
| 4 | 5 | 10 |
| 5 | 6 | 11 |
| 6 | 9 | 13 |

# Computing Steps (1/2)

| Serial n. | Age (years) (x) | Weight (Kg) (y) | xy | $X^2$ | $Y^2$ |
|---|---|---|---|---|---|
| 1 | 7 | 12 | 84 | 49 | 144 |
| 2 | 6 | 8 | 48 | 36 | 64 |
| 3 | 8 | 12 | 96 | 64 | 144 |
| 4 | 5 | 10 | 50 | 25 | 100 |
| 5 | 6 | 11 | 66 | 36 | 121 |
| 6 | 9 | 13 | 117 | 81 | 169 |
| Total | $\sum x=$ 41 | $\sum y=$ 66 | $\sum xy=$ 461 | $\sum x^2=$ 291 | $\sum y^2=$ 742 |

# Computing Steps (2/2)

$$r = \frac{461 - \dfrac{41 \times 66}{6}}{\sqrt{\left[291 - \dfrac{(41)^2}{6}\right] . \left[742 - \dfrac{(66)^2}{6}\right]}}$$

r = 0.759

strong positive (direct) correlation

# Example

- Relationship between anxiety and test scores

| Anxiety (X) | Test score (Y) | $X^2$ | $Y^2$ | $XY$ |
|---|---|---|---|---|
| 10 | 2 | 100 | 4 | 20 |
| 8 | 3 | 64 | 9 | 24 |
| 2 | 9 | 4 | 81 | 18 |
| 1 | 7 | 1 | 49 | 7 |
| 5 | 6 | 25 | 36 | 30 |
| 6 | 5 | 36 | 25 | 30 |
| $\sum X = 32$ | $\sum Y = 32$ | $\sum X^2 = 230$ | $\sum Y^2 = 204$ | $\sum XY = 129$ |

# Result

$$r = \frac{(6)(129)-(32)(32)}{\sqrt{\left(6(230)-32^2\right)\left(6(204)-32^2\right)}} = \frac{774-1024}{\sqrt{(356)(200)}} = -.94$$

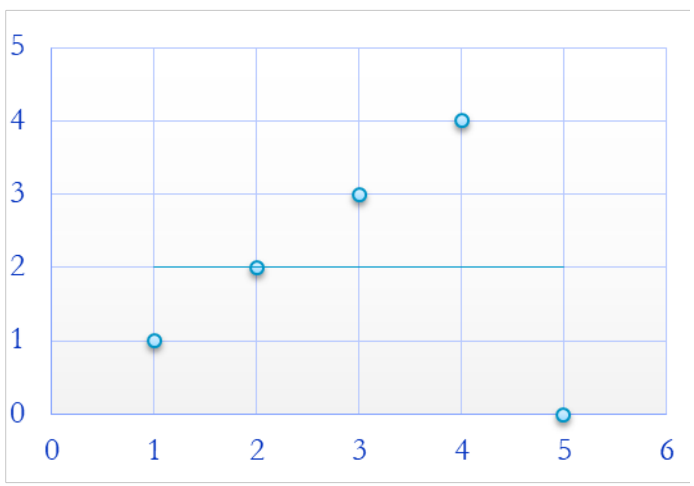
r = - 0.94

strong negative (indirect) correlation

# Limitations of r

- When r = 1 or r = -1
  - We can predict y from x with certainty

  All data points are on a straight line: y = ax + b
- r is actually $\hat{r}$
  - r = true r of whole population
  - $\hat{r}$ = estimate of r based on data
  - r is very sensitive to extreme values



# Spearman's Correlation Coefficient $(r_s)$

- May be used in the following cases:
  - Both variables are quantitative.
  - Both variables are qualitative ordinal.
  - One variable is quantitative and the other is qualitative ordinal.

# Computing $r_s$

- Rank the values of X from 1 to n where n is the number of pairs of values of X and Y in the sample.
- Rank the values of Y from 1 to n.
- Compute the value of $d_i$ for each pair of observation by subtracting the rank of $Y_i$ from the rank of $X_i$
- Square each $d_i$ and compute $\Sigma d_i^2$ (which is the sum of the squared values).
- Apply the formula $$r_s = 1 - \frac{6\sum (di)^2}{n(n^2 - 1)}$$
- The value of $r_s$ denotes the magnitude and nature of association giving the same interpretation as simple r.

# Example 2

- Find the relationship between education level and income from the following data.

| sample numbers | Education Level (X) | Income (Y) |
|---|---|---|
| A | preparatory | 25 |
| B | primary | 10 |
| C | university | 8 |
| D | secondary | 10 |
| E | secondary | 15 |
| F | illiterate | 50 |
| G | university | 60 |

# Ranking (X)

| (X) | Rank X | Adjusted Rank X |
|---|---|---|
| university | 1 | 1.5 |
| university | 2 | 1.5 |
| secondary | 3 | 3.5 |
| secondary | 4 | 3.5 |
| preparatory | 5 | 5 |
| primary | 6 | 6 |
| illiterate | 7 | 7 |

# Ranking (Y)

| (Y) | Rank Y | Adjusted Rank Y |
|---|---|---|
| 60 | 1 | 1 |
| 50 | 2 | 2 |
| 25 | 3 | 3 |
| 15 | 4 | 4 |
| 10 | 5 | 5.5 |
| 10 | 6 | 5.5 |
| 8 | 7 | 7 |

# Solution

| | (X) | (Y) | Rank X | Rank Y | $d_i$ | $d_i^2$ |
|---|---|---|---|---|---|---|
| A | preparatory | 25 | 5 | 3 | 2 | 4 |
| B | primary | 10 | 6 | 5.5 | 0.5 | 0.25 |
| C | university | 8 | 1.5 | 7 | -5.5 | 30.25 |
| D | secondary | 10 | 3.5 | 5.5 | -2 | 4 |
| E | secondary | 15 | 3.5 | 4 | -0.5 | 0.25 |
| F | illiterate | 50 | 7 | 2 | 5 | 25 |
| G | university | 60 | 1.5 | 1 | 0.5 | 0.25 |

$\sum d_i^2 = 64$

$$r_s = 1 - \frac{6 \times 64}{7(48)} = -0.1$$

indirect weak correlation between level of education and income

# Covariance Matrix

- https://stattrek.com/matrix-algebra/covariance-matrix.aspx

# Covariance Matrix

- Actually, "Variance and Covariance Matrix"

$$V = \begin{bmatrix} \Sigma\, x_1^2 / N & \Sigma\, x_1\, x_2 / N & \ldots & \Sigma\, x_1\, x_n / N \\ \Sigma\, x_2\, x_1 / N & \Sigma\, x_2^2 / N & \ldots & \Sigma\, x_2\, x_n / N \\ \ldots & \ldots \quad \ldots & \ldots & \\ \Sigma\, x_n\, x_1 / N & \Sigma\, x_n\, x_2 / N & \ldots & \Sigma\, x_n^2 / N \end{bmatrix}$$

V : n x n variance-covariance matrix

N : the number of scores in each of the n data sets

$X_i$ : a deviation score from the $i^{th}$ data set

$\Sigma\, x_i^2 / N$ : the variance of elements from the $i^{th}$ data set

$\Sigma\, x_i\, x_j / N$ : the covariance for elements from the $i^{th}$ and $j^{th}$ data sets

Variances along the diagonal

Covariances along the off-diagonal

# Covariance Matrix Example

- Raw Data

| Student | Math | English | Art |
|---|---|---|---|
| 1 | 90 | 60 | 90 |
| 2 | 90 | 90 | 30 |
| 3 | 60 | 60 | 60 |
| 4 | 60 | 60 | 90 |
| 5 | 30 | 30 | 30 |

# Result

$$V = \begin{bmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{bmatrix}$$

- The diagonal shows the variances (or eigenvectors or principal components)
- Off-diagonal shows the covariances

How do we get the V matrix?

# Deriving a Covariance Matrix from an nxk Matrix

- Represent the raw data as a matrix
- (our example)

A 5 x 3 table : 5 rows, 3 features

| Student | Math | English | Art |
|---|---|---|---|
| 1 | 90 | 60 | 90 |
| 2 | 90 | 90 | 30 |
| 3 | 60 | 60 | 60 |
| 4 | 60 | 60 | 90 |
| 5 | 30 | 30 | 30 |

# Transformation

- Represent the data from the table as matrix M, where each column in the matrix shows scores on a test and each row shows scores for a student.

$$M = \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix}$$

- Compute the variance of each test (feature) and the covariance between each pair of tests.

# Solution: Step 1 of 3

- Transform the raw data from *M* (an n x k matrix) into matrix *d* of variance (deviation) scores, using the formula

$$d = M - \mathbf{1}\,\mathbf{1}'\,M\,(1/n)$$

where

$\mathbf{1}$ is an $n$ x $1$ column vector of one(1)s
$\mathbf{1}'$ is the transpose of $\mathbf{1}$
$d$ is an $n$ x $k$ matrix of *variance* scores:
$d_{11}, d_{12}, \ldots, d_{nk}$
$M$ is an $n$ x $k$ matrix of *raw* scores:
$M_{11}, M_{12}, \ldots, M_{nk}$

# Reminder: Matrix Dot Product

$$\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \qquad \mathbf{1}' = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$$

$$\mathbf{1}\,\mathbf{1}' = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

# Illustration

$$d = \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix} - \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix} (1/5)$$

$$d = \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix} - \underbrace{\begin{bmatrix} 66 & 60 & 60 \\ 66 & 60 & 60 \\ 66 & 60 & 60 \\ 66 & 60 & 60 \\ 66 & 60 & 60 \end{bmatrix}}_{\text{means}} = \underbrace{\begin{bmatrix} 24 & 0 & 30 \\ 24 & 30 & -30 \\ -6 & 0 & 0 \\ -6 & 0 & 30 \\ -36 & -30 & -30 \end{bmatrix}}_{\text{deviations}}$$

# Transformation (2/2)

- (2) Compute d'd, the *k* x *k* deviation sums of squares and cross products matrix for *d*.
- (3) Divide each term in the deviation sums of squares and cross product matrix by *n* to create the variance-covariance matrix.

That is, V = d'd ( 1 / n )

where

V is a *k* x *k* variance-covariance matrix

d'd is the deviation sums of squares and cross

product matrix n is the number of scores in

each column of the original matrix M

# Solution: Step 2 of 3

- Then we compute d'd, to find the deviation score sums of squares matrix.

$$d'd = \begin{bmatrix} 24 & 24 & -6 & -6 & -36 \\ 0 & 30 & 0 & 0 & -30 \\ 30 & -30 & 0 & 30 & -30 \end{bmatrix} \begin{bmatrix} 24 & 0 & 30 \\ 24 & 30 & -30 \\ -6 & 0 & 0 \\ -6 & 0 & 30 \\ -36 & -30 & -30 \end{bmatrix}$$

$$= \begin{bmatrix} 2520 & 1800 & 900 \\ 1800 & 1800 & 0 \\ 900 & 0 & 3600 \end{bmatrix}$$

# Solution: Step 3 of 3

- Next we divide each element in the deviation sum of squares matrix by *n*.

$$V = d'd \,/\, n$$

- We now have the variance-covariance matrix V.

$$V = \begin{bmatrix} 2520/5 & 1800/5 & 900/5 \\ 1800/5 & 1800/5 & 0/5 \\ 900/5 & 0/5 & 3600/5 \end{bmatrix}$$

$$= \begin{bmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{bmatrix}$$

# Interpreting the Covariance Matrix

- 3 features: Math English Art test scores

$$\begin{array}{cc} & \begin{array}{ccc} \text{Math} & \text{English} & \text{Art} \end{array} \\ \begin{array}{c} \text{Math} \\ \text{English} \\ \text{Art} \end{array} & \begin{bmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{bmatrix} \end{array}$$

- The Art test has the biggest variance; English the smallest.
- The covariance between Math and English (and Math and Art) is positive; as scores on Math go up, scores on English and Art tend to go up; and vice versa.
- The covariance between English and Art is zero. There is no predictable relationship between these two.

# Exercise 1

- Using hand calculation, derive and interpret a covariance matrix for the following dataset.

| Person | Age | Income | Yrs worked | Vacation |
|---|---|---|---|---|
| 1 | 30 | 200 | 10 | 4 |
| 2 | 40 | 300 | 20 | 4 |
| 3 | 50 | 800 | 20 | 1 |
| 4 | 60 | 600 | 20 | 2 |
| 5 | 40 | 300 | 20 | 5 |

# Using Numpy (1/3): Creating a Population Covariance Matrix

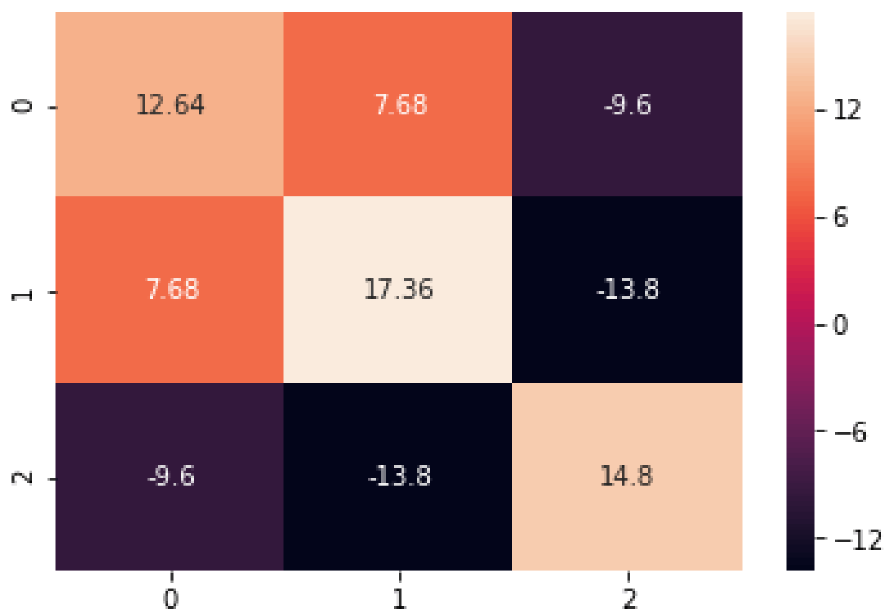
- https://datatofish.com/covariance-matrix-python/

```
import numpy as np
# input data
A = [45,37,42,35,39]
B = [38,31,26,28,33]
C = [10,15,17,21,12]

data = np.array([A,B,C])
# population covariance matrix (N)
covMatrix = np.cov(data,bias=True)
print (covMatrix)
```

```
[[ 12.64   7.68  -9.6 ]
 [  7.68  17.36 -13.8 ]
 [ -9.6  -13.8   14.8 ]]
```

# Using Numpy and Seaborn (2/3): Visualizing a Covariance Matrix

```
import seaborn as sn
import matplotlib.pyplot as plt

sn.heatmap(covMatrix, annot=True, fmt='g')
plt.show()
```



# Using Numpy (3/3): Creating a Sample Covariance Matrix

```
# sample covariance matrix (N-1)
covMatrix = np.cov(data,bias=False)
print (covMatrix)
```

```
[[ 15.8    9.6  -12.  ]
 [  9.6   21.7  -17.25]
 [-12.   -17.25  18.5 ]]
```

# Using Pandas (1/2): Creating a Sample Covariance Matrix

```
import pandas as pd

data = {'A': [45,37,42,35,39],
      'B': [38,31,26,28,33],
      'C': [10,15,17,21,12]
      }

df = pd.DataFrame(data,columns=['A','B','C'])
# sample covariance matrix
covMatrix = pd.DataFrame.cov(df)
print (covMatrix)
```

# Using Pandas and Seaborn (2/2): Visualizing a Covariance Matrix

```
import seaborn as sn
import matplotlib.pyplot as plt

sn.heatmap(covMatrix, annot=True, fmt='g')
plt.show()
```

# Exercise 2

- As shown previously, using NumPy and Pandas (and Seaborn), create a covariance matrix and visualize it. For this exercise, use the dataset used for Exercise 1.
  - A population covariance matrix
  - A sample covariance matrix

# Roadmap

- Regression
  - Linear Regression
  - Polynomial Regression
  - Multiple Regression

# Regression Analysis

- Regression is a technique for predicting some variables given values of other variables.
- The process of predicting some outcome variable (y) using an input variable (x)
- Tells you how values in y change as a function of changes in values of x.

# Correlation vs. Regression

- Correlation describes the strength of a linear relationship between two variables
  - “linear” means “straight line”
- Regression tells us how to draw the straight line described by the correlation
  - Calculates the “best-fit” line for a given set of data (training dataset)

# Regression Terminology

- Linear Regression (linear (i.e., single line) function)
  - $Y = \alpha + \beta X + \varepsilon$
- Polynomial Regression (single independent variable)
  - $Y_i = \alpha + \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k + \varepsilon_i$
- Multiple Regression (multiple independent variables)
  - $Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon_i$
- Non-Linear Regression (nonlinear function)
  - exponential function, log function, power function,…
- Multivariate Regression (multiple dependent variables)

# Variety of Terms for X and Y Axis

| X axis | Y axis |
|---|---|
| predictor | predicted |
| | target, response |
| explanatory | class, label |
| input | output |
| independent | dependent |

# Roadmap: Regression

- Linear Regression (review of errors)
- Polynomial Regression
- Multiple Regression

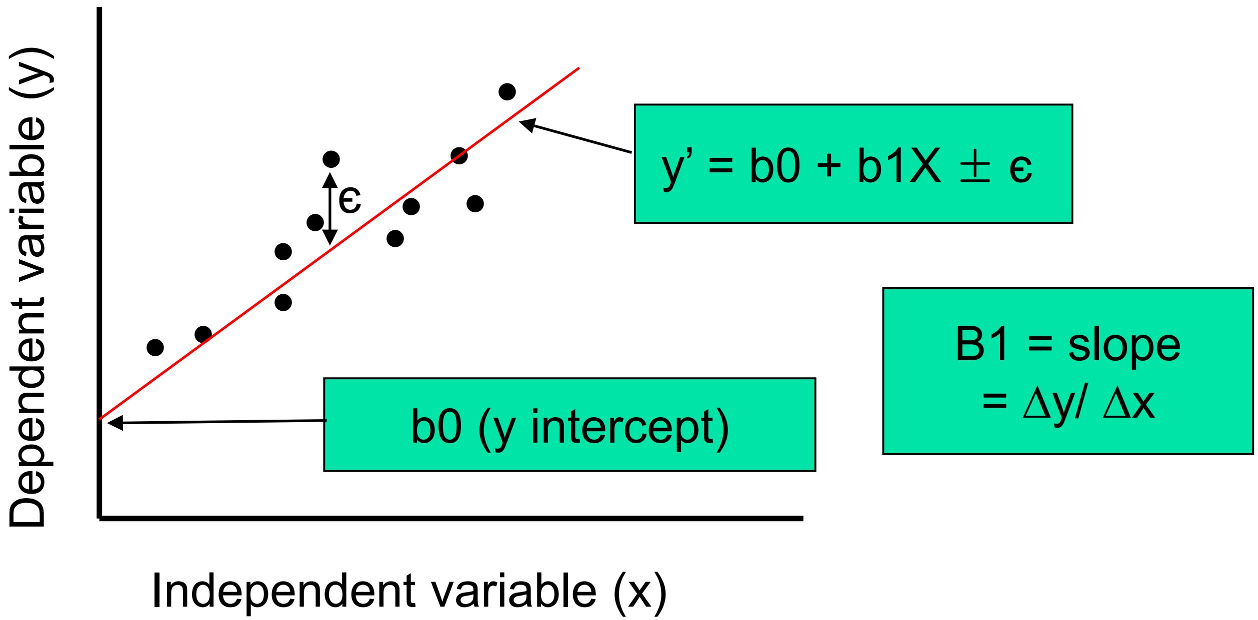
# Acknowledgments

- http://www2.gsu.edu/~dscaas/pptdsc/regression.ppt

# Linear Regression

- The least squared method
  - procedure that minimizes the vertical deviations of plotted points surrounding a straight line
- By using the least squares method we are able to construct a best fitting straight line to the scatter diagram points and then formulate a regression equation.
- The regression line makes the sum of the squares of the residuals (prediction errors) smaller than for any other line.

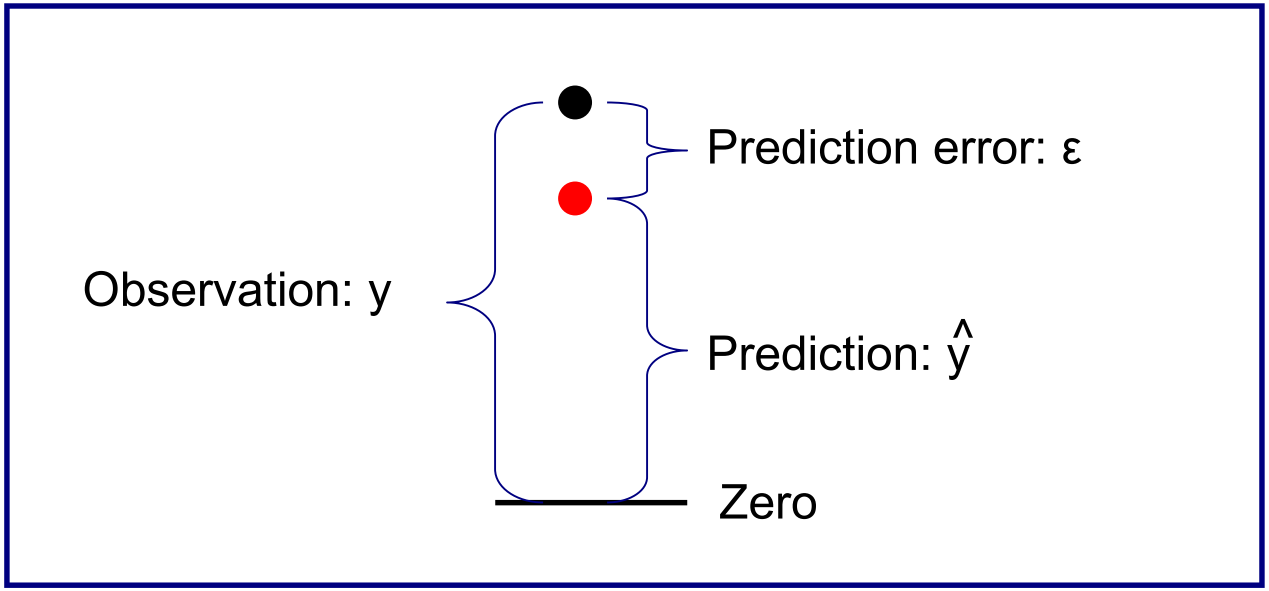
# Linear Regression (with Error Term)




The output of a regression is a function that predicts the dependent variable based upon values of the independent variables.

Simple regression fits a straight line to the data.

# Linear Regression (with Error)



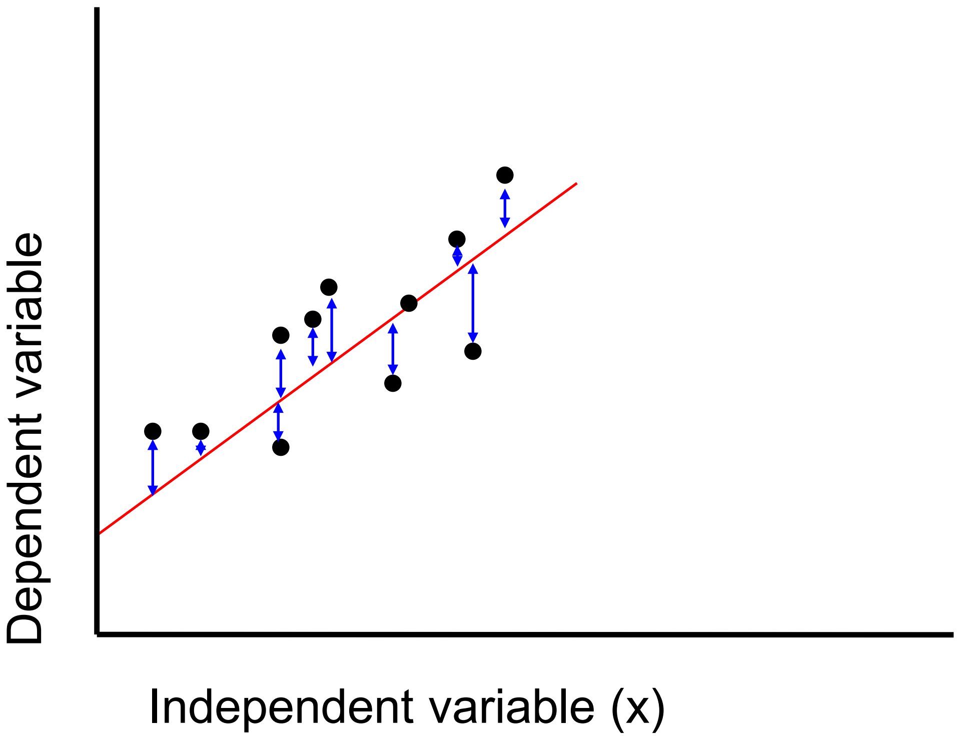

- For each observation, the variation can be described as

$$y = \hat{y} + \varepsilon$$

Actual = Predicted + Error

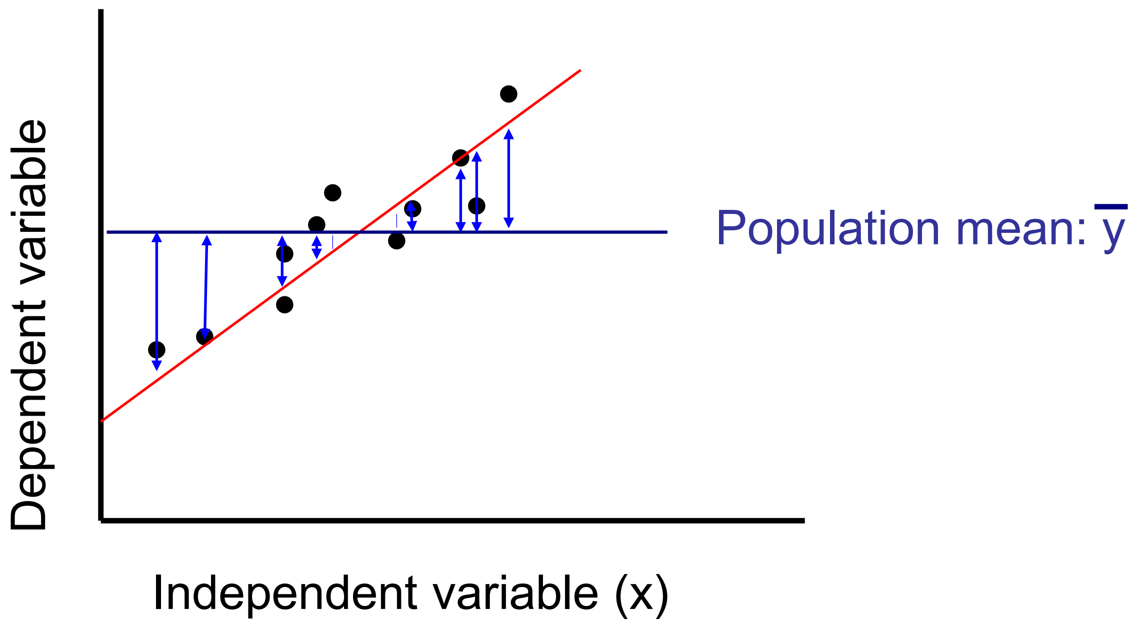
# Sum of Squares of Error

- A least squares regression selects the line with the lowest sum of squared prediction errors (or residual errors).
- This value is called the Sum of Squares of Error (SSE).



# Calculating the Sum of Squares of Regression

- The Sum of Squares of Regression (SSR) is the sum of the squared differences between the prediction for each observation and the population mean.



# Regression Formulas

- Total Sum of Squares (TSS or SST) = SSR + SSE.

SSR = $\Sigma\,(\hat{y} - \bar{y})^2$ (measure of explained variation)

SSE = $\Sigma\,(y - \hat{y})^2$ (measure of unexplained variation)

SST = SSR + SSE = $\Sigma\,(y - \bar{y})^2$ (measure of total variation in y)

# Linear Regression Formula

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

a: intercept
b: slope

$$\hat{y} = \bar{y} + b(x - \bar{x})$$

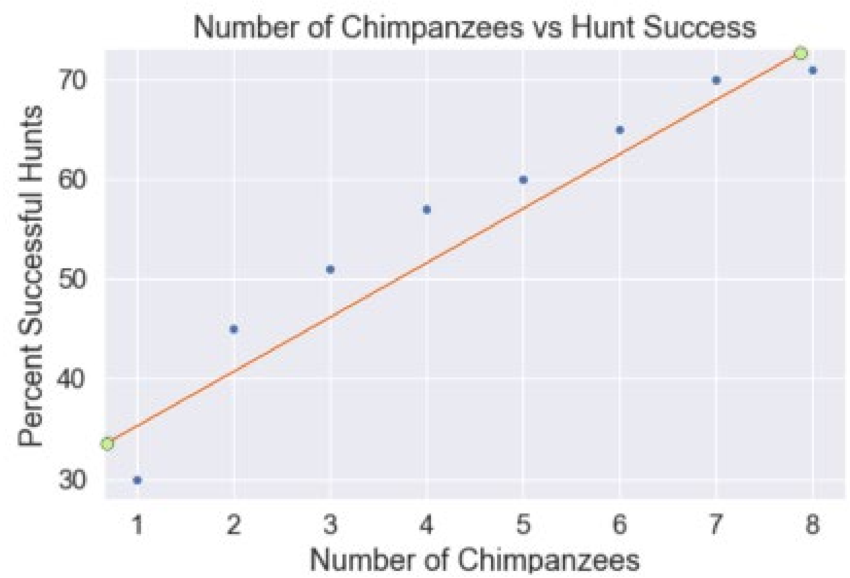
$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

# Walkthrough Example

- https://towardsdatascience.com/linear-regression-by-hand-ee7fe5a751bf
- Dataset: #of chimpanzes and hunting success

| | Number of Chimpanzees | Percent Successful Hunts |
|---|---|---|
| 0 | 1 | 30 |
| 1 | 2 | 45 |
| 2 | 3 | 51 |
| 3 | 4 | 57 |
| 4 | 5 | 60 |
| 5 | 6 | 65 |
| 6 | 7 | 70 |
| 7 | 8 | 71 |



# First, Calculate All the Terms

| Number of Chimpanzees (x) | Percent Successful Hunts (y) | xy | $x^2$ | $y^2$ |
|---|---|---|---|---|
| 1 | 30 | 30 | 1 | 900 |
| 2 | 45 | 90 | 4 | 2025 |
| 3 | 51 | 153 | 9 | 2601 |
| 4 | 57 | 228 | 16 | 3249 |
| 5 | 60 | 300 | 25 | 3600 |
| 6 | 65 | 390 | 36 | 4225 |
| 7 | 70 | 490 | 49 | 4900 |
| 8 | 71 | 568 | 64 | 5041 |
| | | | | |
| $\sum x$ | $\sum y$ | $\sum xy$ | $\sum x^2$ | $\sum y^2$ |
| 36 | 449 | 2249 | 204 | 26541 |

# Next, Plug the Values into the Formulas

$$m = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} \qquad b = \frac{\Sigma y - m(\Sigma x)}{n}$$

$$m = \frac{8(2249) - (36)(449)}{8(204) - (36)^2} \qquad b = \frac{449 - 5.4405(36)}{8}$$

$m = 5.4405$ $\qquad$ $b = 31.6429$

y = mx + b

y = 5.4405x + 31.6429

# Homework

- The following dataset is the amount a person spends on recreation and the person's income.
- 1. Using the following dataset, hand calculate the least squares regression line. Then predict the income of two new persons who spend 3500 and 5300.
- 2. Using scikit-learn and seaborn library, find the regression line and also draw the line and a scatter plot of the dataset.

| **spends** | **income** |
|---|---|
| 2400 | 41200 |
| 2650 | 50100 |
| 2350 | 52000 |
| 4950 | 66000 |
| 3100 | 44500 |
| 2500 | 37700 |
| 5106 | 73500 |
| 3100 | 37500 |
| 2900 | 56700 |
| 1750 | 35600 |

# Coefficient of Determination ($R^2$)

- The proportion of total variation (SST) that is explained by the regression (SSR)
- It is often referred to as $R^2$

  $R^2$ = SSR /SST = 1 – (SSE / SST)

- The value of $R^2$ can range between 0 and 1
- The higher its value, the more accurate the regression model is.

# $R^2$ and Adjusted $R^2$

- A drawback of $R^2$
  - If new predictors are added, $R^2$ increases or remains constant, but never decreases.
  - (We cannot judge that by increasing the complexity of our model, whether we are making it more accurate.)
- Adjusted $R^2$ adjusts $R^2$ for the number of predictors (i.e., degree of freedom) in the model.

  $$R^2_{adjusted} = 1 - (1 - R^2)(N-1) / (N - p - 1)$$

  where p is the number of predictors, and

  N is the total sample size
- The adjusted $R^2$ increases only if new predictors improve the model accuracy.

# Standard Error of Regression

- The Standard Error of a regression is a measure of its variability.
- It can be used in a similar manner to standard deviation, allowing for prediction intervals.
- y ± 2 standard errors will provide approximately 95% accuracy, and 3 standard errors will provide a 99% confidence interval.
- Standard Error is calculated by taking the square root of the average prediction error.

$$\text{Standard Error} = \sqrt{\frac{\text{SSE}}{\text{n-k}}}$$

where n is the number of observations and

k is the total number of variables in the model

# Linear Regression
# Using Pandas and Scikit-Learn

# Build and Evaluate the Model

```
X = pd.DataFrame(df['OAT (F)'])
y = pd.DataFrame(df['Power (kW)'])
model = LinearRegression()

scores = []
kfold = KFold(n_splits=3, shuffle=True,
random_state=42)
for i, (train, test) in enumerate(kfold.split(X, y)):
   model.fit(X.iloc[train,:], y.iloc[train,:])
   score = model.score(X.iloc[test,:], y.iloc[test,:])
   scores.append(score)
print(scores)
```

```
[0.38433441420926338, 0.3938593327006433, 0.4015006377550042]
```

# Explanations for the Code

model = LinearRegression() creates a linear regression model.

The *for* loop divides the dataset into three folds (by shuffling its indexes) # We will learn about this later.

Inside the loop, we fit the data (train the model).

Then we assess the model's performance by appending its score to a list.

scikit-learn returns the $R^2$ scores (for the 3 folds) -- the coefficient of determination ($R^2$ closer to 1 for better linear regression)

# Roadmap: Regression

- Linear Regression
- Polynomial Regression
- Multiple Regression

# Acknowledgments

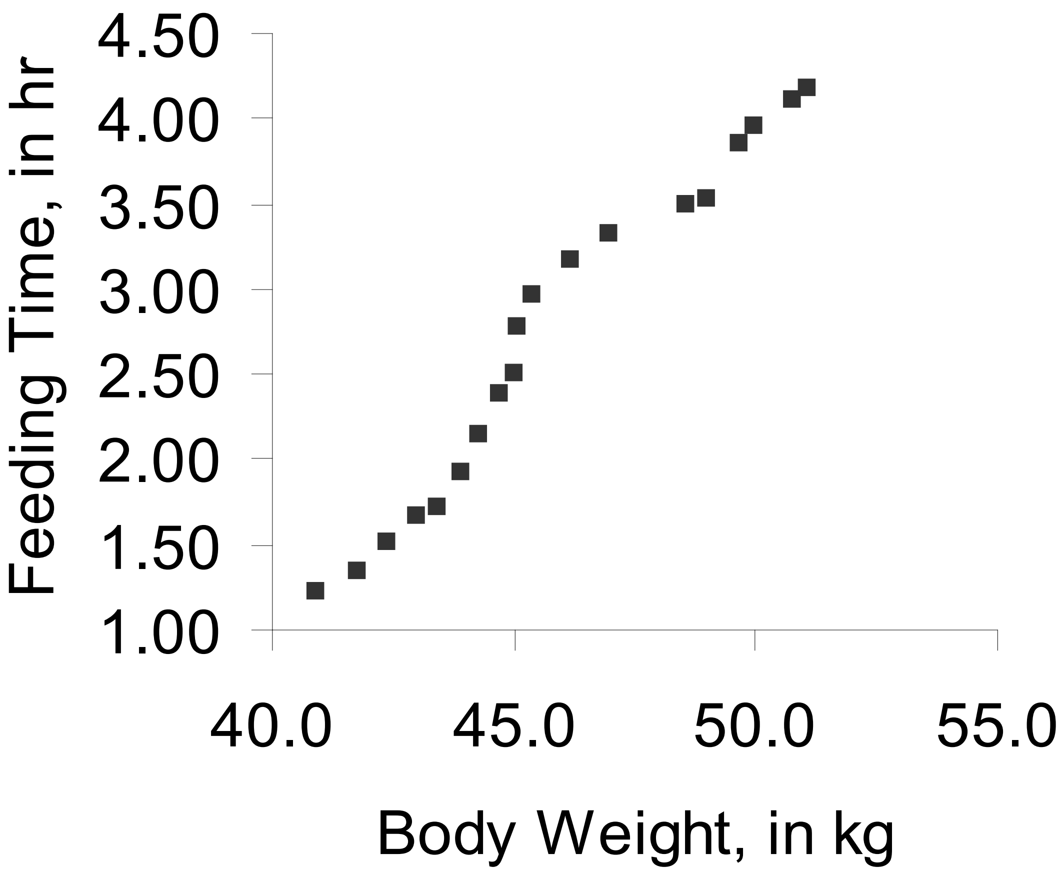
- http://www.fkm.utm.my/~mohsin/mmj1113/03.mohsin.stuff/05.curve.fitting.interpolation.Polynomial.ppt

# Motivating Example (1/2)

- A biologist is interested in the relationship between feeding time and body weight in the males of a mammalian species.
- The data he recorded are shown in the table. The objectives are to
  - construct an equation relating Time to Wt,
  - understand the model selection criteria, and
  - estimate the mean Time for a given Wt with 95% CLM (classical linear model).

| Time (hr) | Wt (kg) |
|---|---|
| 1.22 | 40.9 |
| 2.14 | 44.3 |
| 2.39 | 44.7 |
| 3.50 | 48.6 |
| 1.66 | 43.0 |
| 2.97 | 45.4 |
| 3.95 | 50.0 |
| 1.34 | 41.8 |
| 2.51 | 45.0 |
| 3.53 | 49.0 |
| 1.72 | 43.4 |
| 3.17 | 46.2 |
| 4.11 | 50.8 |
| 1.51 | 42.4 |
| 2.78 | 45.1 |
| 3.85 | 49.7 |
| 1.93 | 43.9 |
| 3.32 | 47.0 |
| 4.18 | 51.1 |

# Motivating Example (2/2)
# Nonlinear Relationship




$Y = a + b\,X$ ?

$Y = a\,e^{X}$ ?

$Y = a\,X^{b}$ ?

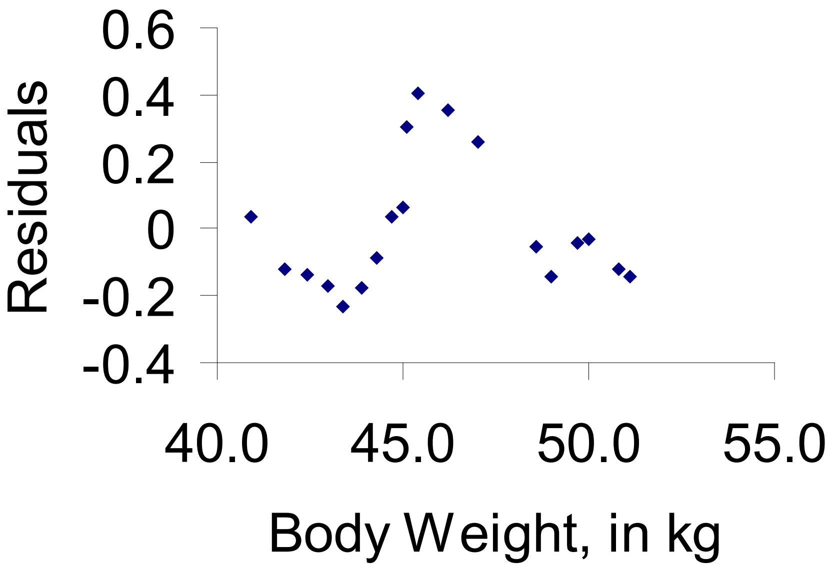
# Polynomial Regression (1/2)

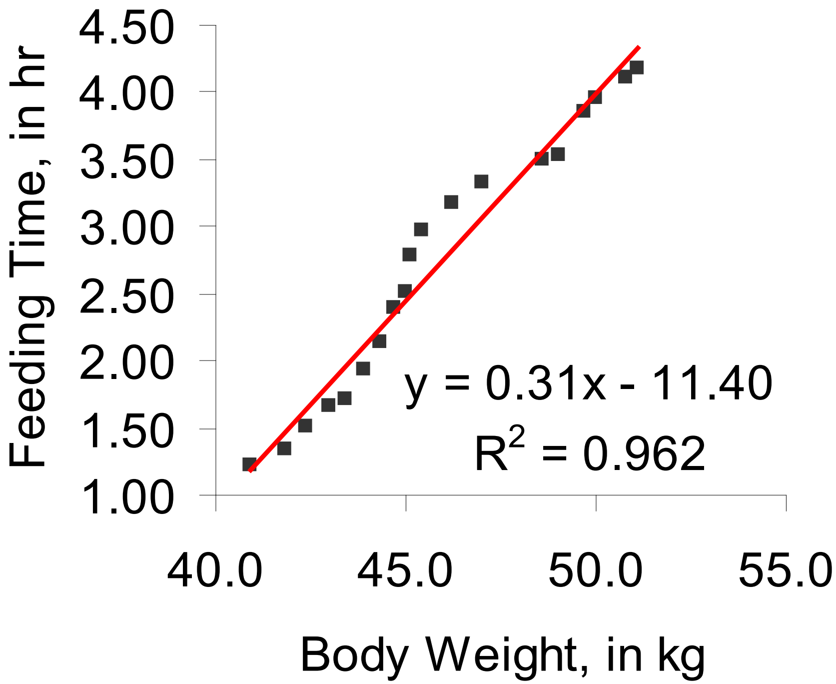
- Polynomial regression is a special type of multiple regression whose independent variables are powers of a single variable X.
- It is used to approximate a curve with unknown functional form.
  $Y_i = \alpha + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k + \varepsilon_i$
- Model selection is done by successively testing highest order terms and discarding insignificant highest-order terms.
  - Tests should use a liberal level of significance, such as $\alpha = 0.25$. The starting order should usually be $k \leq N/10$, where N is the number of observations.

# Polynomial Regression (2/2)

- Higher degree terms are successively discarded, because they are more prone to random error in X (i.e., the random error is multiplied several times in higher order terms).
- Suppose the true value for X is 2 but, because of measurement error, we obtain a value of 3.
  - $X^2$ is then 9.
  - If we had measured the X value accurately, the $X^2$ value would have been 4.
  - So the value of 9 obtained is 4 + 5 units of error.
  - $X^3 = 27 = 8 + 19$ units of error.
- Thus, if an order-4 regression is not significantly better than an order-3 regression, then the $X^4$ term is dropped.

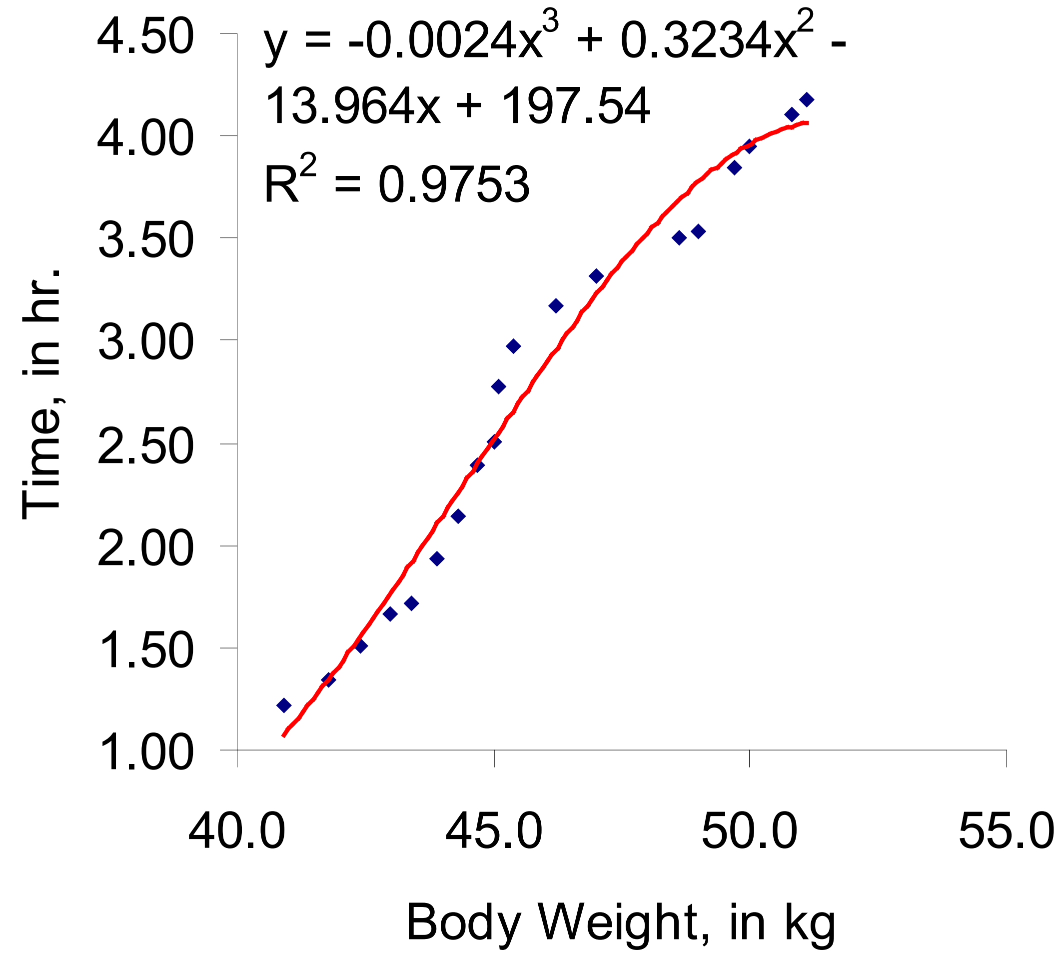
# Linear Regression



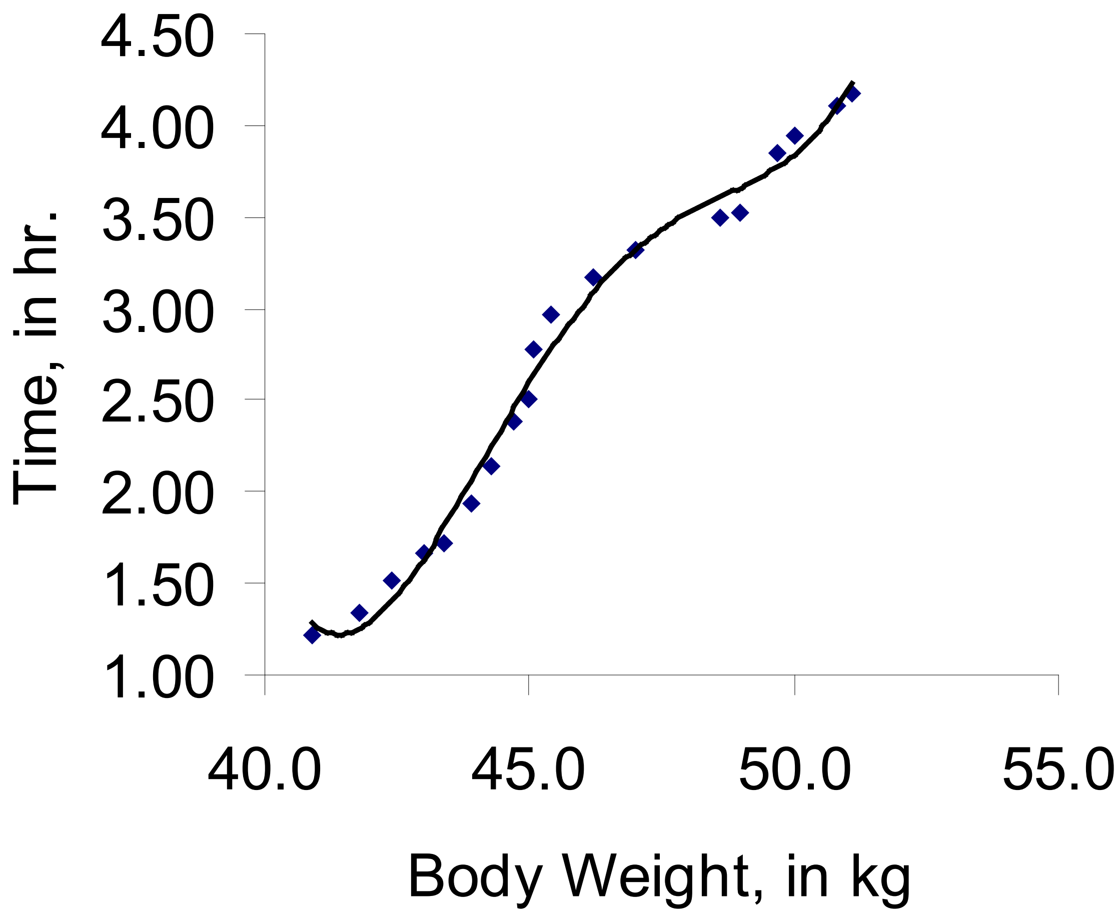



# Polynomial Regression (order 3)

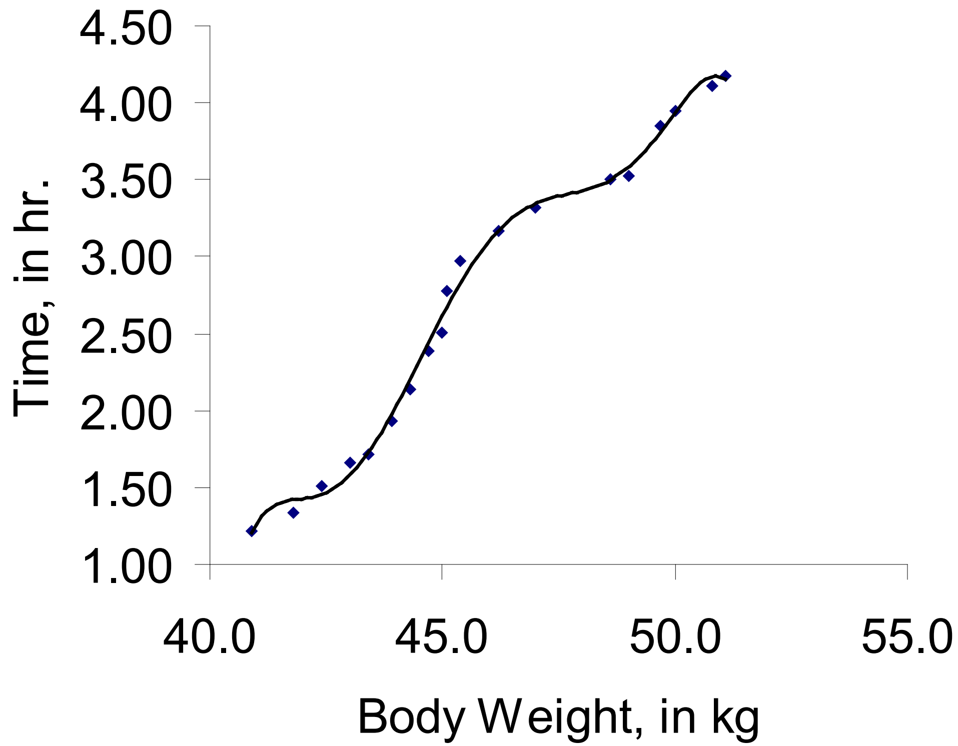
| | | | |
|---|---|---|---|
| 2.14 | 44.3 | 1962.5 | 86938.3 |
| 2.39 | 44.7 | 1998.1 | 89314.6 |
| 3.50 | 48.6 | 2362.0 | 114791.3 |
| 1.66 | 43.0 | 1849.0 | 79507.0 |
| 2.97 | 45.4 | 2061.2 | 93576.7 |
| 3.95 | 50.0 | 2500.0 | 125000.0 |
| 1.34 | 41.8 | 1747.2 | 73034.6 |
| 2.51 | 45.0 | 2025.0 | 91125.0 |
| 3.53 | 49.0 | 2401.0 | 117649.0 |
| 1.72 | 43.4 | 1883.6 | 81746.5 |
| 3.17 | 46.2 | 2134.4 | 98611.1 |
| 4.11 | 50.8 | 2580.6 | 131096.5 |
| 1.51 | 42.4 | 1797.8 | 76225.0 |
| 2.78 | 45.1 | 2034.0 | 91733.9 |
| 3.85 | 49.7 | 2470.1 | 122763.5 |
| 1.93 | 43.9 | 1927.2 | 84604.5 |
| 3.32 | 47.0 | 2209.0 | 103823.0 |
| 4.18 | 51.1 | 2611.2 | 133432.8 |
| | | | |
| | | | |



# Polynomial Regression (order 4)



# Polynomial Regression (order 6)




If we keep increasing the number of polynomial terms in the equation, eventually we will have perfect fit. Is that what we want? (may overfit)

# Criteria for Model Selection

$$R_a^2 = 1 - \frac{n-1}{n-m-1}(1-R^2)$$

| | | | | |
|---|---|---|---|---|
| n | 19 | 19 | 19 | 19 |
| m | 1 | 2 | 3 | 4 |
| $R^2$ | 0.9619 | 0.972 | 0.9753 | 0.9755 |
| $R^2_{adj}$ | 0.9597 | 0.9685 | 0.9704 | 0.9685 |

select

n: number of observations
m: highest order of regression term

# Roadmap: Regression

- Linear Regression
- Polynomial Regression
- Multiple Regression

# Acknowledgments

- http://www.sjsu.edu/faculty/gerstman/biostat-text/Gerstman_PP15.ppt
- https://stat.duke.edu/~gp42/sta101/notes/MultipleRegression.ppt

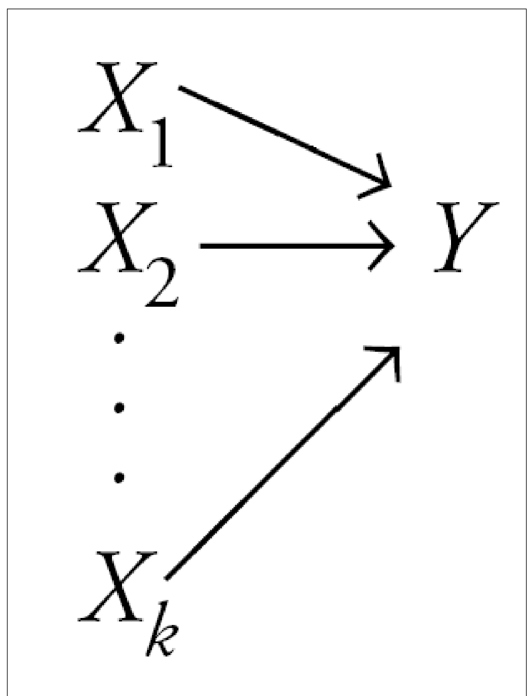
# Intuitive Introduction (1/3)

- Simple regression considers the relation between a single explanatory variable and response variable.

$$X \rightarrow Y$$

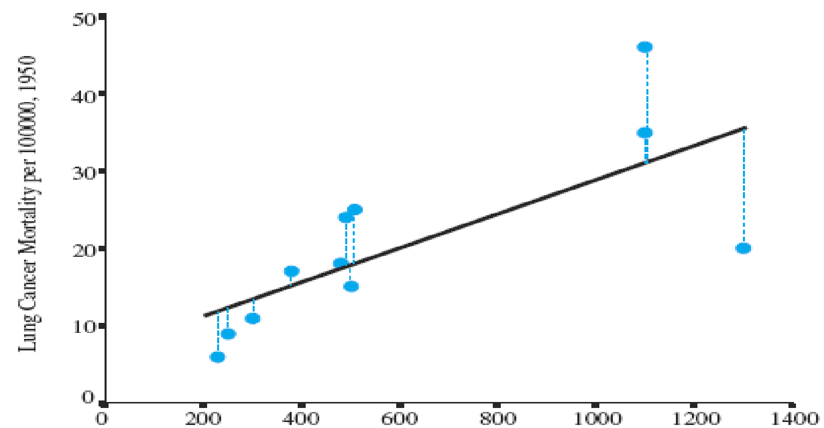
# Intuitive Introduction (2/3)

- Multiple regression simultaneously considers the influence of multiple explanatory variables on a response variable Y.
- The intent is to look at the independent effect of each variable while "adjusting out" the influence of potential confounders.



# Intuitive Introduction (3/3)

- A simple regression model (one independent variable) fits a regression *line* in 2-dimensional space




- A multiple regression model with two explanatory variables fits a regression *plane* in *3*-dimensional space

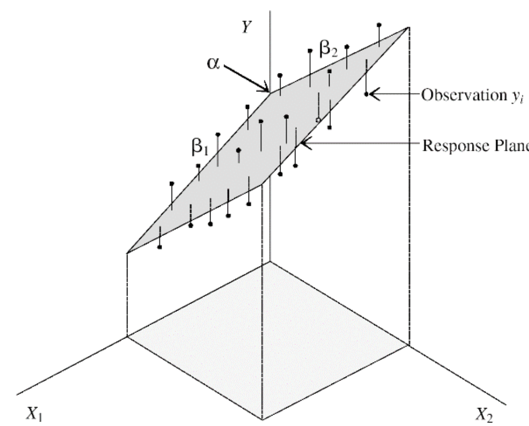



FIGURE 15.1 Three-dimensional response plane.

# Multiple Regression

- Again, estimates for the multiple slope coefficients are derived by minimizing Σresiduals$^2$ to derive this multiple regression model:

$$\hat{y} = a + b_1 x_1 + b_2 x_2$$

- Again, the standard error of the regression is based on the Σresiduals$^2$:

$$S_{Y|x} = \sqrt{\sum \text{residuals}^2 / df_{\text{res}}}$$

- The df(residual) (df = degree of freedom) is the sample size minus (one less than) the number of parameters being estimated.    df(residual) = n - k - 1

# Multiple Regression Model

- Intercept α predicts where the regression *plane* crosses the Y axis
- Slope for variable $X_1$ ($\beta_1$) predicts the change in Y per unit $X_1$ holding $X_2$ constant
- The slope for variable $X_2$ ($\beta_2$) predicts the change in Y per unit $X_2$ holding $X_1$ constant

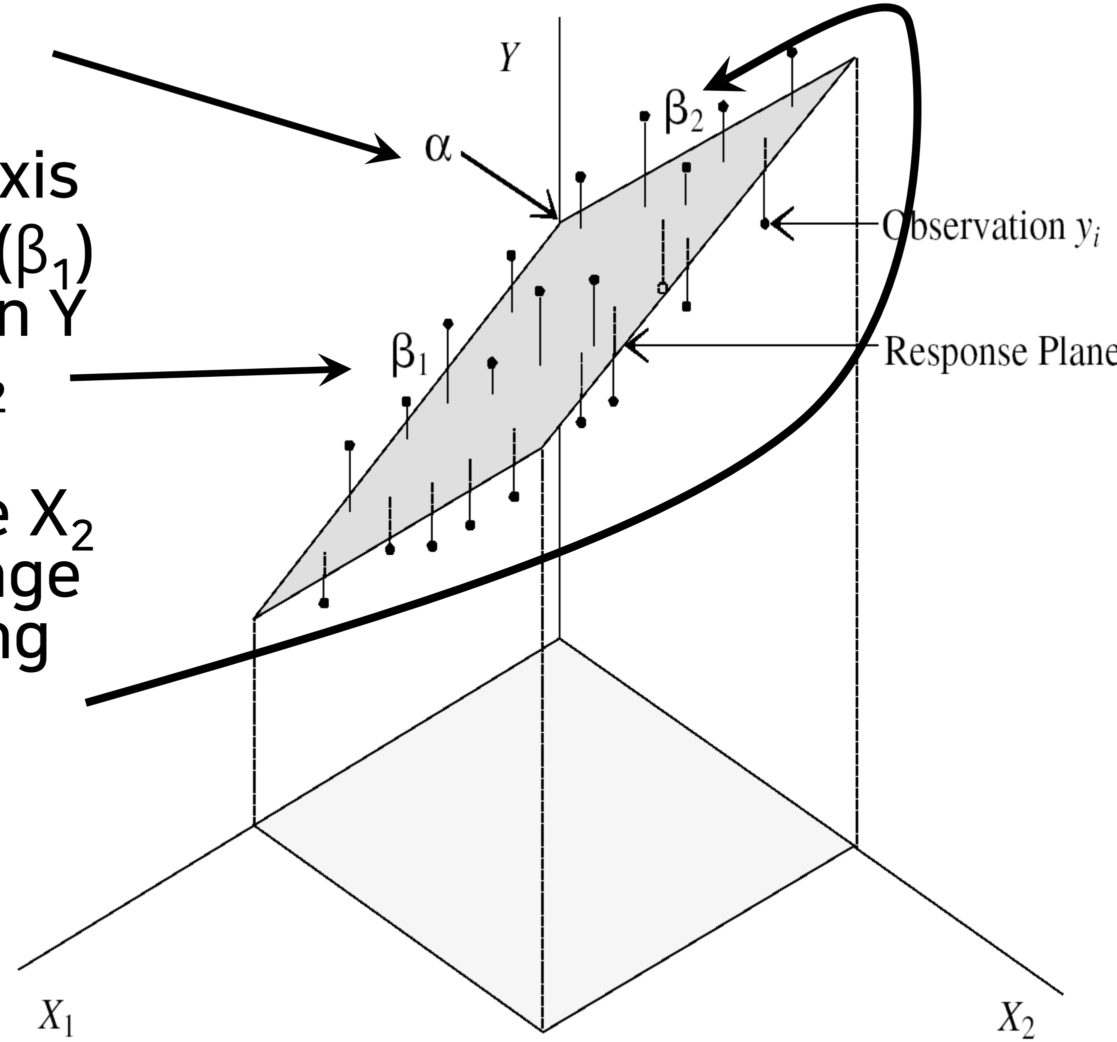



FIGURE 15.1 Three-dimensional response plane.

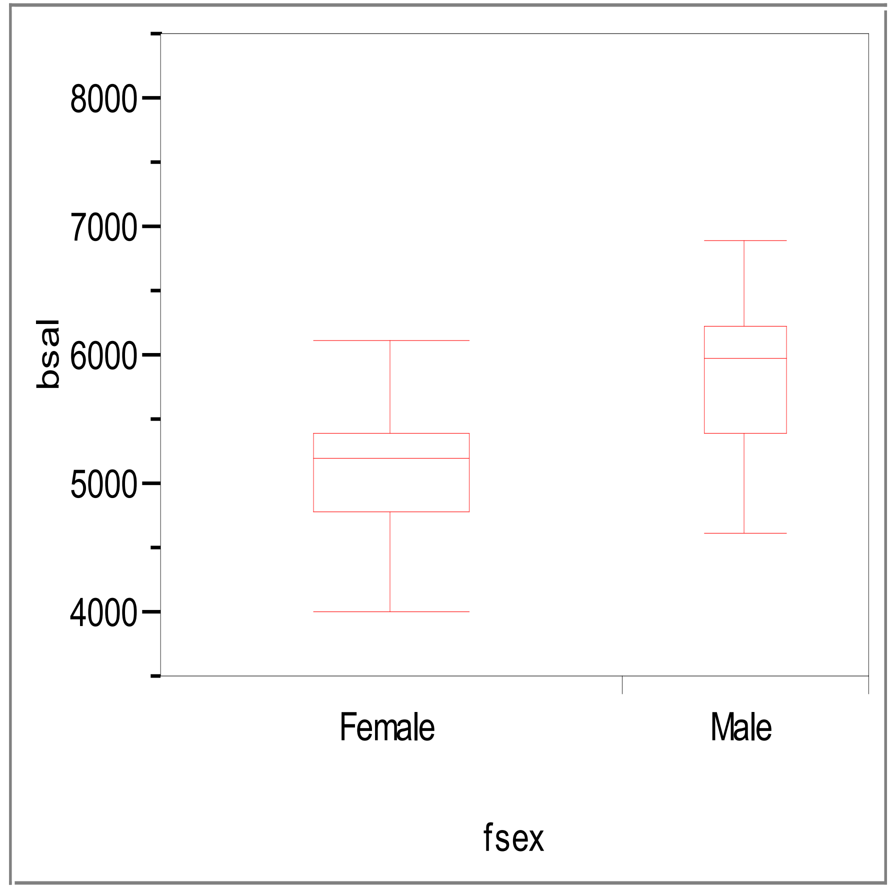
# Example

- Lawsuit for gender discrimination in salaries in a US bank in the 1970s.
- 93 employees on data file (61 female, 32 male).
  - bsal: Annual salary at time of hire.
  - sal77 : Annual salary in 1977.
  - educ: years of education.
  - exper: months previous work prior to hire at bank.
  - fsex: 1 if female, 0 if male
  - senior: months worked at bank since hired
  - age: months
- There are six *x*'s and and one *y* (bsal). However, in what follows we won't use sal77.

# Comparison for Male and Female

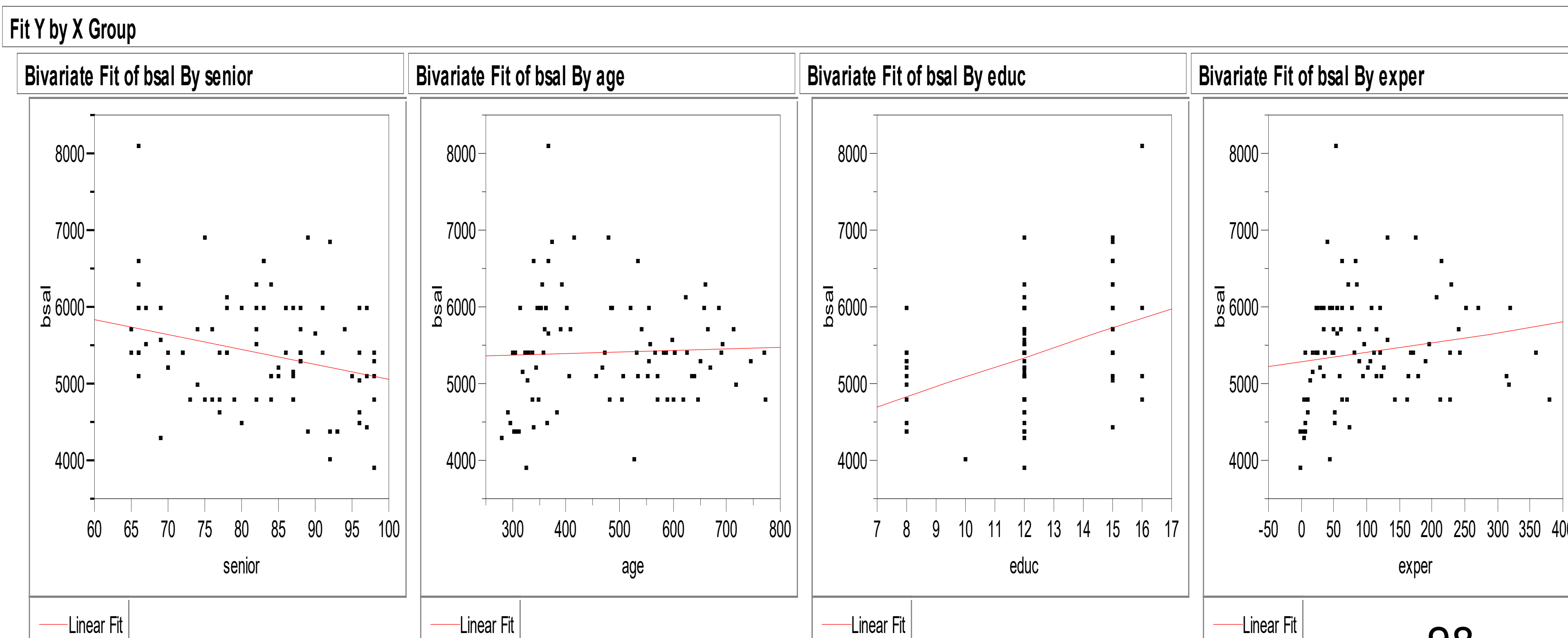
- This shows men started at higher salaries than women.
- But “fsex” (gender) doesn’t control other characteristics.

Oneway Analysis of bsal By fsex



# Relationships of bsal with Predictor Variables

- "senior" and "education" predict bsal well.  We want to control them when judging gender effect.



# Multiple Regression Model

- For any combination of values of the predictor variables, the average value of the response (bsal) lies on a straight line:

$$\text{bsal}_i = \alpha + \beta_1 \text{fsex}_i + \beta_2 \text{senior}_i + \beta_3 \text{age}_i + \beta_4 \text{educ}_i + \beta_5 \text{exper}_i + \varepsilon_i$$

- Just like in simple regression, assume that ε follows a normal curve within any combination of predictors.

# Output from Regression

## Summary of Fit

| | |
|---|---|
| RSquare | 0.515156 |
| RSquare Adj | 0.487291 |
| Root Mean Square Error | 508.0906 |
| Mean of Response | 5420.323 |
| Oservations (or Sum Wgts) | 93 |

# Parameter Estimates
## (fsex = 1 for females,   0 for males)

| Term | Estimate | Std Error | t Ratio | Prob>\|t\| |
|---|---|---|---|---|
| Intcept | 6277.9 | 652 | 9.62 | <.0001 |
| Fsex | -767.9 | 128.9 | -5.95 | <.0001 |
| Senior | -22.6 | 5.3 | -4.26 | <.0001 |
| Age | 0.63 | .72 | .88 | .3837 |
| Educ | 92.3 | 24.8 | 3.71 | .0004 |
| Exper | 0.50 | 1.05 | .47 | .6364 |

*   t ratio, used in t statistics, is similar to z score

# Example Predictions

- Prediction of beginning wages for a woman with 10 months seniority, 25 years (300 months) old, with 12 years of education, and 2 years (24 months) of experience:

$$\text{bsal}_i = \alpha + \beta_1 \text{fsex}_i + \beta_2 \text{senior}_i + \beta_3 \text{age}_i + \beta_4 \text{educ}_i + \beta_5 \text{exper}_i + \varepsilon_i$$

- (fsex = 1 for females, 0 for males)
- Predicted bsal = 6277.9 - 767.9*1 - 22.6*10

  + .63*300 + 92.3*12 + .50*24

  = 6592.6

# Interpretation of Coefficients in Multiple Regression

- Each estimated coefficient is the amount Y is expected to increase by when the value of its corresponding predictor is increased by one, holding constant the values of all other predictors.

- Example: estimated coefficient of education = 92.3.

  For each additional year of education of employee, we expect salary to increase by about 92 dollars, holding all other variables constant.

- Estimated coefficient of fsex = -767.

  For employees who started at the same time, had the same education and experience, and were the same age, women earned $767 less on average than men.

# End of Class