Introduction to Data Science CS61 June 12 - July 12, 2018

Dr. Ash Pahwa

Lesson 6: Regression

Lesson 6.1: Regression – Centering the Model

Outline

- Centered Model
  - Intercept = 0
- Matrix Solution
  - Non-Centered Model
  - Centered model
- Python Solution

Mechanics of Regression

- Closed Form Solution
  - Valid for only 2 variable regression
    - Mean of x, y, x*y, x^2
    - Correlation of x and y + standard dev of x, y
  - Valid for multi variable regression
    - Matrix approach
- Iterative Approach
  - Gradient Decent algorithm
  - Built in R and other software packages

Method#1
Mean of x, y, x*y, x^2

$$m = \frac{\sum y_i x_i - \frac{\sum y_i \sum x_i}{N}}{\sum x_i^2 - \frac{\sum x_i \sum x_i}{N}}$$

$$b = \left(\frac{\sum y_i}{N} - m\frac{\sum x_i}{N}\right)$$

	A	B	C	D	E	F	G
1							
2							
3		X	Y		X*Y		X^2
4		0	1		0		0
5		1	3		3		1
6		2	7		14		4
7		3	13		39		9
8		4	21		84		16
9							
10	SUM	10	45		140		30
11	AVERAGE	2	9		28		6
12	StdDev	1.58113883	8.124038405				
13	Correlation	0.97312368					
14							

- $$m = \frac{\sum y_i x_i - \frac{\sum y_i \sum x_i}{N}}{\sum x_i^2 - \frac{\sum x_i \sum x_i}{N}}$$
- Divide both numerator and denominator by N
- $$m = \frac{\frac{\sum y_i x_i}{N} - \frac{\sum y_i \sum x_i}{N.N}}{\frac{\sum x_i^2}{N} - \frac{\sum x_i \sum x_i}{N.N}} = \frac{Mean\ of\ X*Y - (Mean\ of\ X)*(Mean\ of\ Y)}{Mean\ of\ x^2 - (Mean\ of\ X)*(Mean\ of\ X)}$$

Regression Equation
$y = 5x - 1$

	A	B	C	D	E	F	G
22							
23		Slope : Using Average					
24		Numerator	10		(Mean of X * Y) - (Mean of X)*(Mean of Y)		
25		Denominator	2		(Mean of X^2) - (Mean of X)*(Mean of X)		
26		Slope	5				
27							
28							

Method#2

Correlation of x and y + standard dev of x, y

$$m = \frac{\sum y_i x_i - \frac{\sum y_i \sum x_i}{N}}{\sum x_i^2 - \frac{\sum x_i \sum x_i}{N}}$$

$$b = \left(\frac{\sum y_i}{N} - m\frac{\sum x_i}{N}\right)$$

- $m = \frac{\sum y_i x_i - \frac{\sum y_i \sum x_i}{N}}{\sum x_i^2 - \frac{\sum x_i \sum x_i}{N}}$
- $= r\frac{\sigma_y}{\sigma_x} = Correlation \frac{Std\ Dev\ of\ y}{Std\ Dev\ of\ x}$

	A	B	C	D	E	F	G
1							
2							
3		X	Y		X*Y		X^2
4		0	1		0		0
5		1	3		3		1
6		2	7		14		4
7		3	13		39		9
8		4	21		84		16
9							
10	SUM	10	45		140		30
11	AVERAGE	2	9		28		6
12	StdDev	1.58113883	8.124038405				
13	Correlation	0.97312368					
14							

Regression Equation

$y = 5x - 1$

C29 =B13*C12/B12

	A	B	C	D	E	F	G
27							
28							
29	Statistics	Slope	5		(Correlation * StdDev of Y) / StdDev of X		
30		Intercept	-1		(Mean of Y) - slope * (Mean of X)		
31							

Matrix Approach

- $RSS = (\boldsymbol{Y} - \boldsymbol{XA})^{\boldsymbol{T}}(\boldsymbol{Y} - \boldsymbol{XA})$
- $RSS = \boldsymbol{Y^TY} - \boldsymbol{2Y^TXA} + \boldsymbol{AX^TXA}$
- $Set\ \frac{\partial RSS}{\partial A} = 0,$
  - $to\ compute\ the\ value\ of\ A\ for\ minimum\ RSS$
- $\nabla RSS = \nabla[(\boldsymbol{Y} - \boldsymbol{XA})^{\boldsymbol{T}}(\boldsymbol{Y} - \boldsymbol{XA})]$
- $\nabla RSS = -2\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{XA}) = 0$
- $-2\boldsymbol{X}^T\boldsymbol{Y} + 2\boldsymbol{X}^T\boldsymbol{XA} = 0$
- $\boldsymbol{X}^T\boldsymbol{XA} = \boldsymbol{X}^T\boldsymbol{Y}$
- Multiply both sides with $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$
- $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{XA} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$
- $Since$: $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I} \quad \& \quad \boldsymbol{IA} = \boldsymbol{A}$
- $\boldsymbol{A} = (\boldsymbol{X^TX})^{-\boldsymbol{1}}\boldsymbol{X^TY}$

- By Analogy 1D case
- $\frac{d}{dA}(Y - XA)(Y - XA)$
- $= \frac{d}{dA}(Y - XA)^2$
- $= 2(Y - XA)(-X)$
- $= -2X(Y - XA)$

Using ‘lm’ command

Price	Demand
$49	124
$69	95
$89	71
$99	45
$109	18

```
> summary(lm(demand~price))

Call:
lm(formula = demand ~ price)

Residuals:
      1       2       3       4       5 
-4.2241  0.6724 10.5690  1.5172 -8.5345 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 211.2707    14.7215  14.351 0.000733 ***
price        -1.6948     0.1717  -9.872 0.002210 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.269 on 3 degrees of freedom
Multiple R-squared:  0.9701,      Adjusted R-squared:  0.9602 
F-statistic: 97.46 on 1 and 3 DF,  p-value: 0.00221
```

Regression Equation

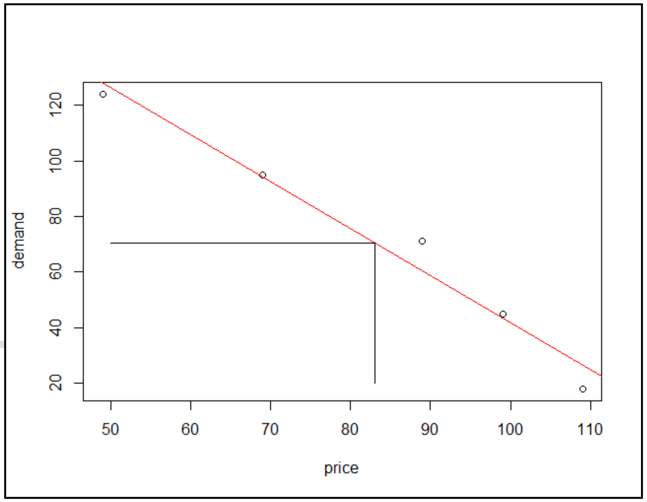
$y = -1.6948x + 211.2707$

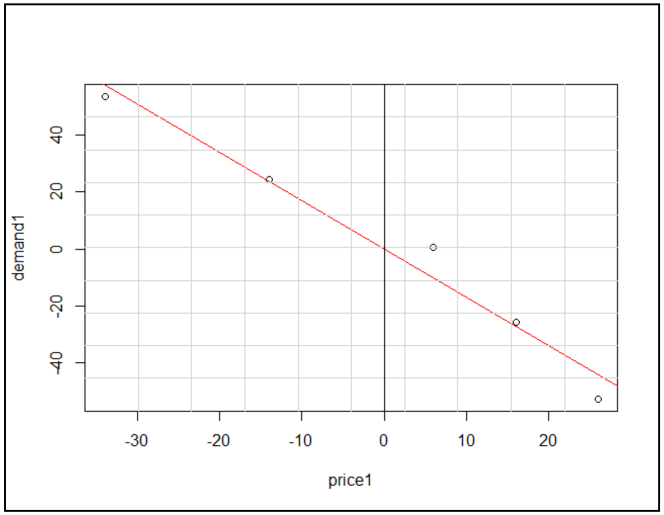
Centered Model

Intercept = 0

Centered Model

- Regression Model
  - $y_i = b_0 + m_1 x_i \quad (1)$
  - $\bar{y} = b_0 + m_1 \bar{x} \quad (2)$
- Where $\bar{x}$ *and* $\bar{y}$ are mean of x and y values
- Subtract equation #2 from equation #1
  - $y_i - \bar{y} = m_1(x_i - \bar{x})$
- This means $x = x_{original} - mean(x_{original})$
- This means y $= y_{original} - mean(y_{original})$
- This is new regression model in centered form
- The new model remains the same except intercept term b_0 is forced to zero
  - $b_0 = 0$

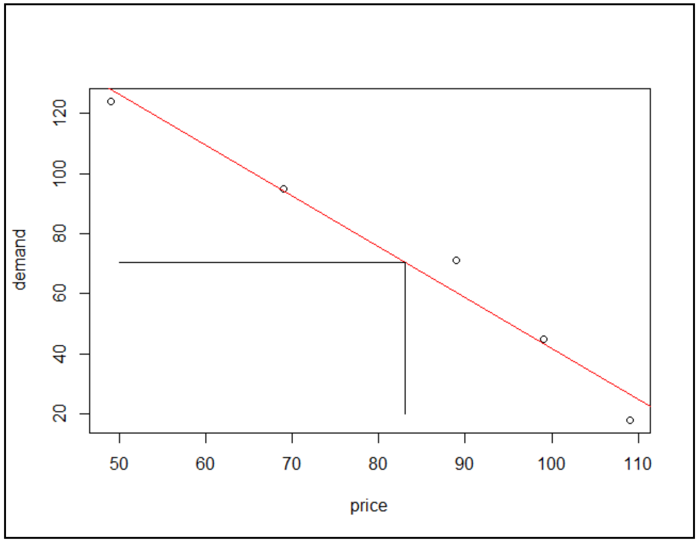




Standard Model

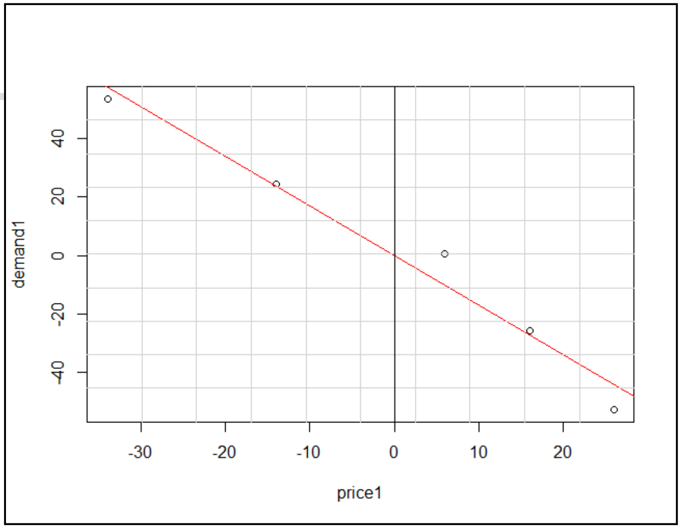
Price	Demand
$49	124
$69	95
$89	71
$99	45
$109	18

```
> price = c(49, 69, 89, 99, 109)
> demand = c(124, 95, 71, 45, 18)
> plot(price,demand)
> result <- lm(demand~price)
> summary(result)
Call:
lm(formula = demand ~ price)
Residuals:
      1        2       3       4       5
-4.2241  0.6724 10.5690  1.5172 -8.5345
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 211.2707    14.7215  14.351 0.000733
***
price        -1.6948     0.1717  -9.872 0.002210 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
Residual standard error: 8.269 on 3 degrees of
freedom
Multiple R-squared:  0.9701,  Adjusted R-squared:
0.9602
F-statistic: 97.46 on 1 and 3 DF,  p-value: 0.00221
> abline(result, col="red")
```



Centered Model : Intercept = 0

```
> (mp = mean(price))
[1] 83
> (md = mean(demand))
[1] 70.6
> lines(c(mp,mp),c(20,md),type="l")
> lines(c(50,mp),c(md,md),type="l")
> ###############################################
> price1 = price - mean(price)
> demand1 = demand - mean(demand)
> plot(price1,demand1)
> result <- (lm(demand1~price1))
> summary(result)
Call:
lm(formula = demand1 ~ price1)
Residuals:
      1        2        3        4        5
-4.2241  0.6724 10.5690  1.5172 -8.5345
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.0000     3.6981   0.000  1.00000
price1       -1.6948     0.1717  -9.872  0.00221 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
Residual standard error: 8.269 on 3 degrees of freedom
Multiple R-squared:  0.9701,  Adjusted R-squared:  0.9602
F-statistic: 97.46 on 1 and 3 DF,  p-value: 0.00221
> abline(result, col="red")
> grid(10,10,lty=1)
> lines(c(0,0),c(-60,60))
```



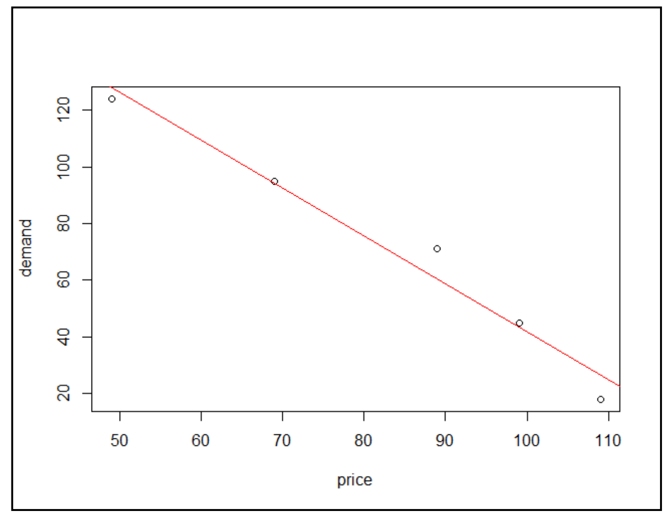
Centered Model

- Intercept can always be recovered later, if required
- What are advantages of Centered Model
  - No need to estimate the intercept, because it is zero
  - Centered data is more interpretable

Example Non-Centered

Price	Demand
$49	124
$69	95
$89	71
$99	45
$109	18

- $Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{bmatrix} \quad A = \begin{bmatrix} b \\ m \end{bmatrix} \quad E = \begin{bmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{bmatrix}$
- $Y = \begin{bmatrix} 124 \\ 95 \\ 71 \\ 45 \\ 18 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 49 \\ 1 & 69 \\ 1 & 89 \\ 1 & 99 \\ 1 & 109 \end{bmatrix}$
- $Y = XA + E$
- Solution for Least RSS = $A = (X^T X)^{-1} X^T Y$



Example Non-Centered

- $Y = \begin{bmatrix} 124 \\ 95 \\ 71 \\ 45 \\ 18 \end{bmatrix}$ $X = \begin{bmatrix} 1 & 49 \\ 1 & 69 \\ 1 & 89 \\ 1 & 99 \\ 1 & 109 \end{bmatrix}$

- $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$
- $A^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

- Solution for Least RSS = $A = (X^TX)^{-1}X^TY$
- $X^TX = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 49 & 69 & 89 & 99 & 109 \end{bmatrix}\begin{bmatrix} 1 & 49 \\ 1 & 69 \\ 1 & 89 \\ 1 & 99 \\ 1 & 109 \end{bmatrix} = \begin{bmatrix} 5 & 415 \\ 415 & 36765 \end{bmatrix}$
- $(X^TX)^{-1} = \frac{1}{11600}\begin{bmatrix} 36765 & -415 \\ -415 & 5 \end{bmatrix}$
- $X^TY = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 49 & 69 & 89 & 99 & 109 \end{bmatrix}\begin{bmatrix} 124 \\ 95 \\ 71 \\ 45 \\ 18 \end{bmatrix} = \begin{bmatrix} 353 \\ 25367 \end{bmatrix}$
- $A = (X^TX)^{-1}X^TY = \frac{1}{11600}\begin{bmatrix} 36765 & -415 \\ -415 & 5 \end{bmatrix}\begin{bmatrix} 353 \\ 25367 \end{bmatrix} = \begin{bmatrix} 211.2707 \\ -1.6948 \end{bmatrix}$

Example: Non-Centered Final Answer

- $A = (X^T X)^{-1} X^T Y = \frac{1}{11600}\begin{bmatrix} 36765 & -415 \\ -415 & 5 \end{bmatrix}\begin{bmatrix} 353 \\ 25367 \end{bmatrix} = \begin{bmatrix} 211.2707 \\ -1.6948 \end{bmatrix}$
- $f(x) = 211.27 - 1.6948x$

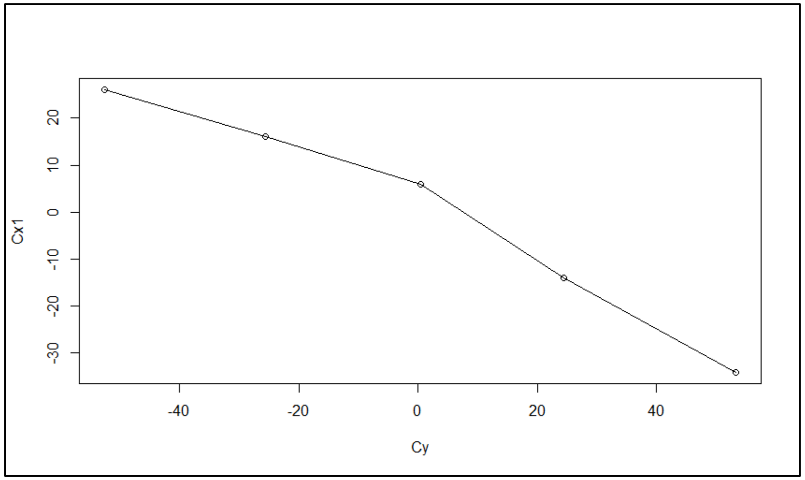
Regression Equation
$y = -1.6948x + 211.2707$

Matrix Solution Centered

Price	Demand
$49	124
$69	95
$89	71
$99	45
$109	18
Mean = $83	Mean = 70.6

Price - Mean	Demand - Mean
-$34	53.4
-$14	24.4
$6	0.4
$16	-25.6
$26	-52.6

- $Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}$ $X = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}$ $A = \begin{bmatrix} b \\ m \end{bmatrix}$ $E = \begin{bmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{bmatrix}$
- $Y = \begin{bmatrix} 53.4 \\ 24.4 \\ 0.4 \\ -25.6 \\ -52.6 \end{bmatrix}$ $X = \begin{bmatrix} -34 \\ -14 \\ 6 \\ 16 \\ 26 \end{bmatrix}$
- $Y = XA + E$
- Solution for Least RSS = $A = (X^T X)^{-1} X^T Y$



Matrix Solution Centered

- $Y = \begin{bmatrix} 53.4 \\ 24.4 \\ 0.4 \\ -25.6 \\ -52.6 \end{bmatrix}$ $X = \begin{bmatrix} -34 \\ -14 \\ 6 \\ 16 \\ 26 \end{bmatrix}$

- $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$
- $A^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

- Solution for Least RSS = $A = (X^T X)^{-1} X^T Y$
- $X^T X = \begin{bmatrix} -34 & -14 & 6 & 16 & 26 \end{bmatrix} \begin{bmatrix} -34 \\ -14 \\ 6 \\ 16 \\ 26 \end{bmatrix} = [2320]$
- $(X^T X)^{-1} = 0.000431$
- $X^T Y = \begin{bmatrix} -34 & -14 & 6 & 16 & 26 \end{bmatrix} \begin{bmatrix} 53.4 \\ 24.4 \\ 0.4 \\ -25.6 \\ -52.6 \end{bmatrix} = [-3932]$
- $A = (X^T X)^{-1} X^T Y = [0.000431][-3932] = [-1.6948]$

Matrix Solution: Final Answer Centered

- $A = (X^T X)^{-1} X^T Y = [0.000431][-3932] = [-1.6948]$
- $f(x) = -1.6948x$

Non-Centered Model

Regression Equation
$y = -1.6948x + 211.2707$

Centered Model

Regression Equation
$y = -1.6948x$

Centered Model in Python: Scikit-Learn

```
import numpy as np
import pandas as pd
from sklearn import linear_model
######################################
# 1. Read Data File
Price = [49, 69, 89, 99, 109]
Demand = [124, 95, 71, 45, 18]

#######################################
Xarray = np.array(Price)
Yarray = np.array(Demand)
meanX = np.mean(Xarray)
print(meanX)
83.0

meanY = np.mean(Yarray)
print(meanY)
70.6

Xarray_modified = Xarray - meanX
print(Xarray_modified)
[-34. -14.   6.  16.  26.]

Yarray_modified = Yarray - meanY
print(Yarray_modified)
[ 53.4  24.4   0.4 -25.6 -52.6]
```

Price - Mean	Demand - Mean
-$34	53.4
-$14	24.4
$6	0.4
$16	-25.6
$26	-52.6

Price	Demand
$49	124
$69	95
$89	71
$99	45
$109	18
Mean = $83	Mean = 70.6

Centered Model in Python: Scikit-Learn

```
####################################
# Regression Using Scikit-Learn
# Without Centering
#
df_x = pd.DataFrame(Xarray)
df_y = pd.DataFrame(Yarray)
reg = linear_model.LinearRegression()
reg.fit(df_x,df_y)
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)

print(reg.coef_)
[[-1.69482759]]

print(reg.intercept_)
[ 211.27068966]
```

$$y = -1.6948x + 211.2707$$

```
#########################################
df_x = pd.DataFrame(Xarray_modified)
df_y = pd.DataFrame(Yarray_modified)

reg = linear_model.LinearRegression()
reg.fit(df_x,df_y)
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)

print(reg.coef_)
[[-1.69482759]]

print(reg.intercept_)
[  5.68434189e-15]
```

$$y = -1.6948x + 0$$

Summary

- Centered Model
  - Intercept = 0
- Matrix Solution
  - Non-Centered Model
  - Centered model
- Python Solution