STA261: Probability and Statistics II

Shahriar Shams

Week 11(Correlation and Regression)

Winter 2020

Recap of Week 10

- Part-1 of lecture:
  - Idea of Bayesian Inference
  - Prior, Likelihood and Posterior
  - Some examples of calculating posterior dist
- Part-2 of lecture:
  - Inference using posterior dist
    - Summary of posterior dist
    - Credible Region
  - Different types of prior

Learning goal for this lecture

- Part-1 of lecture: correlation and Least square regression
  - Relationship among quantitative variables
  - Pearson correlation coefficient
  - Least square regression

- Part-2 of lecture: Regression under Normal distribution
  - Properties of estimators of regression parameters
  - Confidence interval/t-test for β_2
  - Sum of squares decomposition/ANOVA test

These are selected topics from E&R chapter 10.3

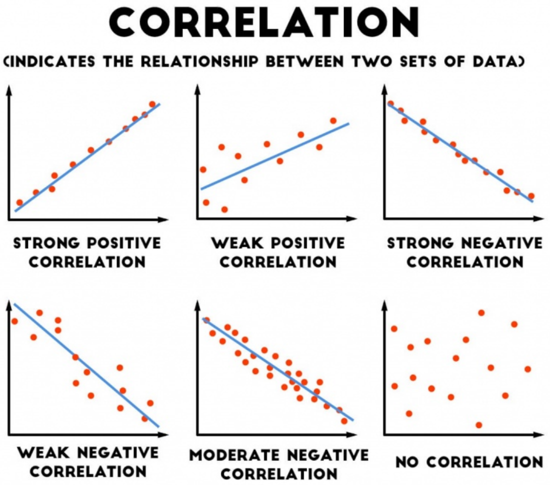
Section 1

Correlation and least square regression

Relationship among quantitative variables

- Suppose we have quantitative variables X and Y.
- We want to check whether there is any relationship between them or not.
- Let $(x_1, x_2, ..., x_n)$ and $(y_1, y_2, ..., y_n)$ are the two corresponding data vectors.
- A visual display of these two vectors can be done by drawing a **Scatter Plot**.
- Plotting y_i's against x_i's will give us the scatter plot where $i = 1, 2, ..., n$
- Scatter plot suggests the direction and magnitude of **correlation** between X and Y

Interpretation of Scatter plot



Source:`http://www.pythagorasandthat.co.uk/scatter-graphs`

Interpretation of Scatter plot (cont...)

- Think of a hypothetical line that goes through the points.
- Direction of the line:
  - the line is going upward $\implies$ the correlation is positive.
  - the line is going downward $\implies$ then the correlation is negative.
- Closeness of the points to the line suggests the strength of the correlation
  - points are closely clustered around the line $\implies$ strong correlation
  - points are not so close to the line $\implies$ moderate/weak correlation
- If the points look totally random $\implies$ No relationship between X and Y

Pearson Correlation Coefficient (r)

- Correlation coefficient, r measures the **linear** relation ship between two variables.
- It's a unit free number which ranges from -1 to 1.
- $r = -1 \implies$ Perfect Negative Correlation (All the points are exactly on a downward line)
- $r = 1 \implies$ Perfect Positive Correlation (All the points are exactly on a upward line)
- $r = 0 \implies$ Zero correlation.
- Geometric definition of r:

$$r = cos(\theta)$$

where, θ is the angle between n dimensional vector $X - \bar{X}$ and vector $Y - \bar{Y}$

Least Square Regression

- Let $y = b_1 + b_2 x$ is the equation of the hypothetical line that we thought is going throw the points.
- $(y_i - b_1 - b_2 x_i)$ is the deviation of y_i from the line.
- Least square regression is the technique of finding the line (in other words, finding b_1 and b_2) that **minimizes** sum of the squared deviations,

$$\sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)^2$$

- Differentiating this expression with respect to b_1 and b_2 and equating to zero gives us:

$$b_1 = \bar{y} - b_2 \bar{x}$$

$$b_2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Example 10.3.3

x	y	$(x-\bar{x})$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
3.9	8.9	2.9	8.41	5.51	30.360	15.979
2.6	7.1	1.6	2.56	3.71	13.764	5.936
2.4	4.6	1.4	1.96	1.21	1.464	1.694
4.1	10.7	3.1	9.61	7.31	53.436	22.661
-0.2	1.0	-1.2	1.44	-2.39	5.712	2.868
5.4	12.6	4.4	19.36	9.21	84.824	40.524
0.6	3.3	-0.4	0.16	-0.09	0.008	0.036
-5.6	-10.4	-6.6	43.56	-13.79	190.164	91.014
-1.1	-2.3	-2.1	4.41	-5.69	32.376	11.949
-2.1	-1.6	-3.1	9.61	-4.99	24.900	15.469
$\bar{x}=1$	$\bar{y}=3.39$	-	$sum = 101.08$	-	$sum = 437.009$	$sum = 208.13$

Therefore,

- $b_2 = \frac{208.13}{101.08} = 2.059062 \approx 2.059$ and
- $b_1 = 3.39 - 2.059062 * 1 = 1.330938 \approx 1.331$

The least square regression line is: $y = 1.331 + 2.059x$

Some notes on least square regression

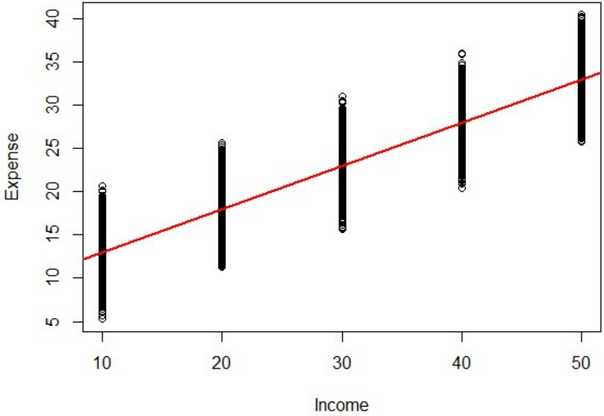
- Least square regression doesn't require any distributional assumption.
- It is more like how to fit a linear regression line if we have all have population level data.
- I found this page online which explains the concept of least square interactively (I think it's really cool!!!) `https://setosa.io/ev/ordinary-least-squares-regression/` One the second graph of this page, try changing the intercept or slope value and see what happens graphically.

Section 2

Classical linear regression under Normal dist.

Idea of regression under Normal dist

This is a hypothetical example:



- X represents income category
- For each category of X, we have 10000 different individuals.
- In total we have 50,000 individuals in our population.

Simple Linear Regression

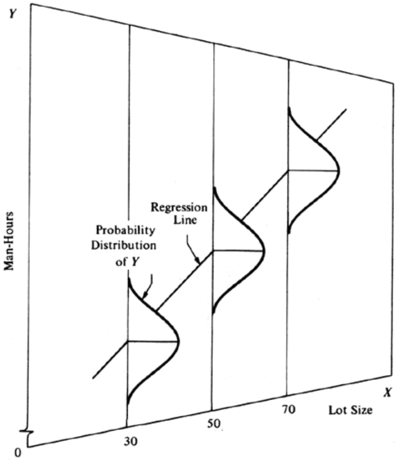


FIGURE 2.4 Pictorial representation of linear regression model

source: Neter, Wasserman & Kutner (1983) *Applied Linear Regression Models*

Assumptions:

- $(Y|X = x) \sim N(\beta_1 + \beta_2 x, \sigma^2)$
- The mean of Y is a linear function of X
- The variance (σ^2) is constant
- $(y_1, y_2, ..., y_n)$ are observed values of Y
- $(x_1, x_2, ..., x_n)$ are observed values of X
- y_i's are independent

Likelihood func of Simple Linear Regression

- The conditional distribution of Y is assumed to be Normal.
- $E[Y_i|X_i = x_i] = \beta_1 + \beta_2 x_i$
- $var[Y_i|X_i = x_i] = \sigma^2$
- The likelihood function of $(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n)$ will be a function of $(x_1, x_2, ..., x_n)$, β_1, β_2 and $\sigma^2 \implies$

$$L(\beta_1, \beta_2, \sigma^2|data) = (2\pi\sigma^2)^{-n/2} exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_1 - \beta_2 x_i)^2]$$

- For any given σ^2, this likelihood will be maximized when $\sum_{i=1}^{n}(y_i - \beta_1 - \beta_2 x_i)^2$ will be minimized.

Maximizing the likelihood func.

- Maximizing the likelihood function written in the previous slide is same as minimizing the sum of squared differences between y_i's and $b_1 + b_2 x_i$'s .
- Hence, the optimization becomes same as the least square regression (which does not involve any Normality assumption)
- Therefore,

$$\hat{\beta}_1 = b_1 = \bar{y} - b_2\bar{x}$$

$$\hat{\beta}_2 = b_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Example 10.3.3

x	y	$(x-\bar{x})$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
3.9	8.9	2.9	8.41	5.51	30.360	15.979
2.6	7.1	1.6	2.56	3.71	13.764	5.936
2.4	4.6	1.4	1.96	1.21	1.464	1.694
4.1	10.7	3.1	9.61	7.31	53.436	22.661
-0.2	1.0	-1.2	1.44	-2.39	5.712	2.868
5.4	12.6	4.4	19.36	9.21	84.824	40.524
0.6	3.3	-0.4	0.16	-0.09	0.008	0.036
-5.6	-10.4	-6.6	43.56	-13.79	190.164	91.014
-1.1	-2.3	-2.1	4.41	-5.69	32.376	11.949
-2.1	-1.6	-3.1	9.61	-4.99	24.900	15.469
$\bar{x}=1$	$\bar{y}=3.39$	-	$sum = 101.08$	-	$sum = 437.009$	$sum = 208.13$

Therefore,

- $b_2 = \frac{208.13}{101.08} = 2.059062 \approx 2.059$ and
- $b_1 = 3.39 - 2.059062 * 1 = 1.330938 \approx 1.331$

The least square regression line is: $y = 1.331 + 2.059x$

Interpretation of Regression parameters

- β_1 represents the expected value of Y when $X = 0$
- β_2 represents the change in expected value of Y for 1-unit increase in X

Different parameterization of the same model

- In a lot of text books, linear regression is written slightly differently though it represents the same model.
- the model is writen as

$$Y_i = \beta_1 + \beta_2 x_i + \epsilon_i$$

- with the assumption: $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$

- This is equivalent of saying

$$Y_i \stackrel{iid}{\sim} N(\beta_1 + \beta_2 x_i, \sigma^2)$$

Subsection 1

Properties of estimators of regression parameters

Parameters, Estimators, Estimates...

- If we had the population level data, we would have been able to calculate the "true" intercept and slope
  - Population parameters: β_1 and β_2

- Instead we observe a sample and calculate estimates of those parameters.
  - Estimates: b_1 and b_2

- If we keep taking random samples and keep calculating the intercept and the slope we will get different values(likely)
  - Estimators: B_1 and $B_2 \leftarrow$ these two are random variables.

Recall: μ is the parameter, $\bar{X}$ is the variable and $\bar{x}$ is the value from our sample.

Properties of Estimators

- We can re-write the equations of the estimators

$$B_1 = \bar{Y} - B_2\bar{x}$$

$$B_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Technical Note: Since we are dealing with bunch of conditional distributions, Y is the random variable and x is treated as fixed constant.

- B_2 can be expressed as

$$B_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Properties of Estimators (cont...)

- B_2 is a linear combinations of Y_i's (which are bunch of Normal variables). So is B_1.
- Then both B_1 and B_2 follows Normal distribution.
- B_1 and B_2 are unbiased estimators of β_1 and β_2
    - $E[B_1] = \beta_1$
    - $E[B_2] = \beta_2$
- $var[B_1]$ and $var[B_2]$ can be calculated and will be a function of σ^2 (Theorem 10.3.3)

$$var[B_2] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- We can write,

$$B_2 \sim N\left(\beta_2\,,\ \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

Subsection 2

Confidence interval/t-test for β_2

Confidence Interval of β_2

- An unbiased estimator of σ^2 is

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - b_1 - b_2 x_i)^2$$

- It can be proved that

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{(n-2)}$$

- Then using the definition of t-distribution,

$$\frac{B_2 - \beta_2}{\sqrt{\frac{S^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \sim t_{(n-2)}$$

- Then γ level confidence interval for β_2

$$B_2 \pm t_{(1+\gamma)/2(df=n-2)} * SE(B_2)$$

where $SE(B_2) = \sqrt{\frac{S^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

Testing $H_0 : \beta_2 = 0$

- $\beta_2 = 0 \implies$ There is no relationship between X and Y
- We can either calculate the confidence interval of β_2 using the formula given in previous slide and check whether zero is inside or not.
- Or we can use the following test statistic to calculate the p-value.

$$T = \frac{B_2}{SE(B_2)} \sim t_{(n-2)}$$

- For example 10.3.3
  $b_2 = 2.06$, $SE(B_2) = 0.1023$, $t_{0.975(8)} = 2.306$
- 95% CI

$$2.06 \pm 2.306 * 0.1023 = (1.824, 2.296)$$

- Zero is not inside the interval, so there is evidence of relationship between X and Y

Subsection 3

Sum of Square decomposition and ANOVA test

Sum of Squares decomposition

- Total sum of square (TSS) $= \sum_{i=1}^{n}(y_i - \bar{y})^2$
- TSS can be written as the sum of two terms:
  - Regression sum of square (RSS) $= b_2^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$
  - Error/Residual sum of square (ESS) $= \sum_{i=1}^{n}(y_i - b_1 - b_2 x_i)^2$
  - Therefore,

$$TSS = RSS + ESS$$

ANOVA table (Another way of testing $H_0 : \beta_2 = 0$)

Source	df	Sum of Square (SS)	Mean SS = SS/df
X	1	$b_2^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$	$b_2^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$
Error	$n-2$	$\sum_{i=1}^{n}(y_i - b_1 - b_2 x_i)^2$	s^2
Total	$n-1$	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	-

For example 10.3.3

Source	df	Sum of Square (SS)	Mean SS = SS/df
X	1	$(2.059)^2 * 101.08 = 428.527$	428.527
Error	8	$437.009 - 428.527 = 8.482$	$8.482/8 = 1.06$
Total	9	437.009	-

Finally, $F = \frac{\text{Mean SS for X}}{\text{Mean SS for error}} = \frac{428.527}{1.06} = 404.27 \sim F_{(1,8)}$

where, $F_{(1,8)}$ represents F distribution with df 1 and 8

$$p - value = 1 - pf(404.27, df1 = 1, df2 = 8) = 0.000$$

Coefficient of determination and Correlation coefficient

- Coefficient of determination (R^2) is defined as

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

- R^2 represents the proportion of variation in Y that can be explained by the model.
- For simple linear regression (only one X variable),

$$r^2 = R^2 \implies r = \sqrt{R^2}$$

For example 10.3.3,

- $R^2 = \frac{428.527}{437.009} = 0.9805908 \implies$ 98.05% variation in Y can be explained by the model/by the variation in X.
- $r = \sqrt{R^2} = \sqrt{0.985908} = 0.9902478$ (why should "r" be +ve)
- So, there is strong +ve relationship between X and Y

Little R code for Example 10.3.3

In R, linear regression is fitted using the command called "lm( )" where "lm" stands for linear model.

```
x=c(3.9,2.6,2.4,4.1,-0.2,5.4,0.6,-5.6,-1.1, -2.1)
y=c(8.9,7.1,4.6,10.7,1.0,12.6,3.3,-10.4,-2.3,-1.6)
m=lm(y ~ x)
summary(m)
anova(m)
```

Assignment (Non-credit)

Evans and Rosenthal

Exercise: 10.3.4(a,b,f,g,h), 10.3.7(a,b,e,f,g)