STA255 Week-11

Shahriar Shams

23/03/2020

Review of week-10

- Testing population proportion, $p = p_0$
- Idea of p-value.
- Z test and confidence interval for difference between two population means
- Two sample t-test and confidence interval
- Analysis of paired data

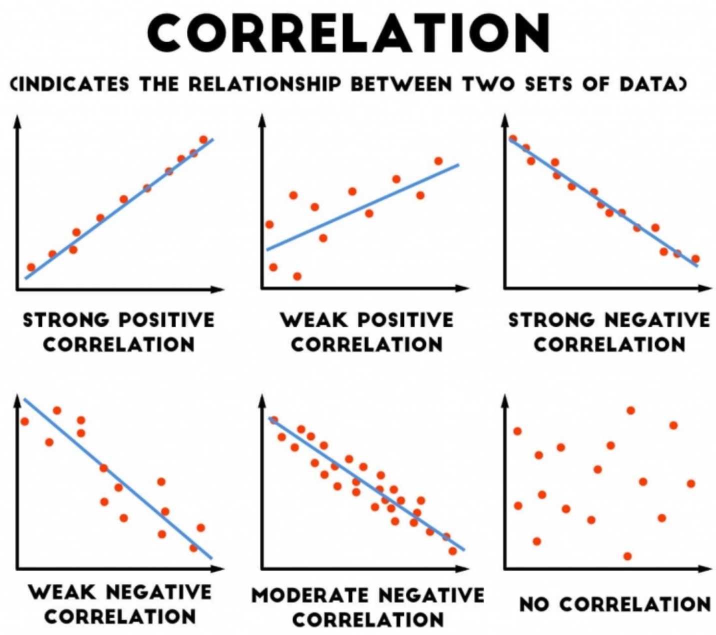
Learning goals

- Relationship among quantitative variables
- Pearson correlation coefficient
- Least square regression
- Regression under Normal distribution
  - Parameter estimation
  - Interpretation of regression parameters
  - Properties of estimators of regression parameters
  - Confidence interval/t-test for the slope
  - Sum of squares decomposition

Relationship among quantitative variables

- Suppose we have two quantitative variables X (say represents income) and Y (say represents expense).
- 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.
  - x_1 is the income of the first individual and y_1 is his/her expense.
- 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 $i = 1, 2, ..., n$
- **Recall:** Scatter plot suggests the direction and magnitude of **correlation** between X and Y

Pearson correlation coefficient



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

- 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

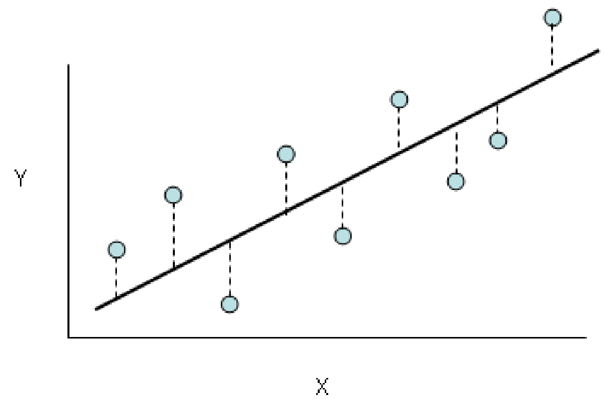
- Correlation coefficient, r measures the **linear** relation ship between two variables.
- It's a unit free number which ranges from -1 to 1.
- $r = -1 \implies$ Perfect Negative Correlation (All the points are exactly on a downward line)
- $r = 1 \implies$ Perfect Positive Correlation (All the points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

- Correlation coefficient calculated from a sample,

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

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.



Source: https://medium.com/statistical-guess/least-square-regression-34aaef3f76ec

- 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:

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
- $b_1 = 3.39 - 2.059062 * 1 = 1.330938 \approx 1.331$

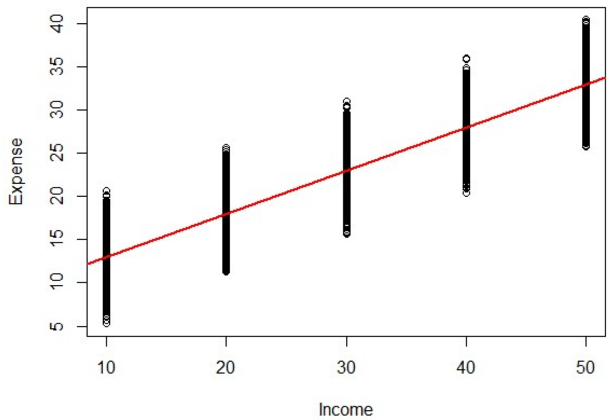
The least square regression line is: $y = 1.331 + 2.059x$

Some comments:

- 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. 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.

Linear Regression under Normal distribution

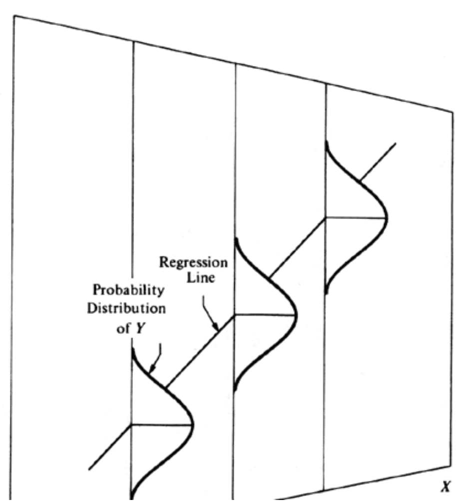
- Let's take a look at a hypothetical example (may look a bit unrealistic)



- X represents income category (10K, 20K etc.)
  - For the sake of this example, say income can only be \$10K or \$20K or ... and nothing in between.
- For each category of X, we have expenses of 10000 different individuals.
- In total we have 50,000 individuals in our population.

Some assumptions:

- $(Y|X = x) \sim N(\beta_1 + \beta_2 x, \sigma^2)$
- The conditional mean of Y is a linear function of X
- The conditional variance of Y, (σ^2) is constant
- y_i's are independent



Parameter estimation

- 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.

- Hence, the optimization becomes same as the least square regression (which does not involve any Normality assumption)
- Therefore,

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

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

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

Using our example:

- β_1 is the average expense when income = 0.
- β_2 is the change in the average expense, when income increases by 1-unit (the unit here is $1000).

Properties of estimators of regression parameters

- 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 estimator(random variable) and $\bar{x}$ is the value from our sample ie. estimate.

- 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 here and x is treated as fixed constant.
- B_2 can be expressed as

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

- 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

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

- We can write,

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

Confidence interval/t-test for the slope

- An unbiased estimator of σ^2 is

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

- 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 $100 * (1-\alpha)\%$ level confidence interval for β_2

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

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

- For the numeric example on page 4, $b_2 = 2.06$, $SE(B_2) = 0.1023$, $t_{0.975,(8)} = 2.306$
- 95% CI for β_2

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

```
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)

y_pred = 1.331 + 2.059*x

S2 = sum((y-y_pred)^2)/8

SE_B2 = sqrt(S2/sum((x-mean(x))^2))
SE_B2
```

```
## [1] 0.1022622
```

- Using $T = \frac{B_2 - \beta_2}{\sqrt{\frac{S^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}$ as a test statistic which follows a $t_{(df=n-2)}$ distribution we can conduct any test of hypothesis like

$$H_0 : \beta_2 = 0$$

(or some other value)

```r
m=lm(y~x)
summary(m)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6727 -0.3960  0.1155  0.6541  1.3931
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.3309     0.3408   3.905  0.00451 **
## x             2.0591     0.1023  20.135 3.86e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.028 on 8 degrees of freedom
## Multiple R-squared:  0.9806, Adjusted R-squared:  0.9782
## F-statistic: 405.4 on 1 and 8 DF,  p-value: 3.864e-08
```

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:
  - Error/Residual sum of square (ESS) $= \sum_{i=1}^{n}(y_i - b_1 - b_2 x_i)^2$
  - Regression sum of square (RSS) $= b_2^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$
- It can be shown that

$$TSS = RSS + ESS$$

Coefficient of determination (R^2)

- 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 our numeric example,

```
TSS = sum((y-mean(y))^2)
TSS
```

```
## [1] 437.009
```

```
ESS = sum((y-y_pred)^2)
ESS
```

```
## [1] 8.456399
```

```
RSS = TSS - ESS

R2= RSS/TSS
R2
```

```
## [1] 0.9806494
```

- $R^2 = 0.9806 \implies$ 98.06% variation in Y can be explained by the model/by the variation in X.
- $r = \sqrt{R^2} = \sqrt{0.9806} = 0.9903$ (why should "r" be +ve)
- So, there a is strong +ve relationship between X and Y

Homework

Chapter 12.2

13, 16, 18

Chapter 12.3

34, 36(a,b,c)