STA 1013 : Statistics through Examples

Lecture 31: Linear Regression Analysis

Hwiyoung Lee

December 2, 2019

Department of Statistics, Florida State University

Final Exam

Exam Day :

- Section 5 : December 10th (Tuesday), 10:00 - 12:00 noon.
- Section 15 : December 12th (Thursday), 7:30 – 9:30 AM.

Check website : `https://registrar.fsu.edu/registration_guide/fall/exam_schedule/`

Topics :

1. Hypothesis test
   - Z-test for mean
   - t-test for mean
   - Z-test for proportion
2. Correlation
3. Linear Regression

Note : Open book exam

Review : Correlation Coefficient

Correlation Coefficient

Statisticians measure the strength of a **linear correlation** with a number called the correlation coefficient.

Correlation coefficient

$$r = \frac{\sum_{i=1}^{n}\left(\frac{(x_i-\bar{x})(y_i-\bar{y})}{s_x s_y}\right)}{n-1}$$
$$= \frac{n(\sum_{i=1}^{n} x_i y_i) - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}}$$

Use Calculator : LinRegTTest

Properties of a Correlation

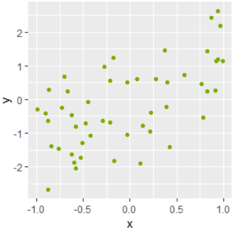
- The correlation coefficient, r, is a measure of the strength of a correlation. Its value can range only from -1 to 1.

$$-1 \leqslant r \leqslant 1$$

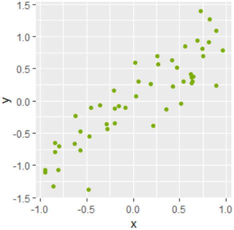
- If there is no correlation, the value of r is close to 0.
- If there is a positive correlation, the correlation coefficient is positive $(0 < r \leqslant 1)$. Values of r close to 1 indicate a strong positive correlation and positive values closer to 0 indicate a weak positive correlation.
- If there is a negative correlation, the correlation coefficient is negative$(-1 \leqslant r < 0)$: Values of r close to -1 indicate a strong negative correlation and negative values closer to 0 indicate a weak negative correlation.

Scatter plot and Correlation

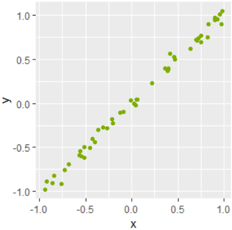
- Positive correlation



$r = 0.2$

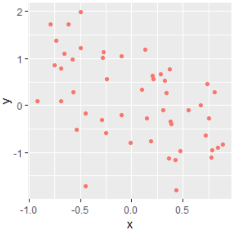


$r = 0.6$

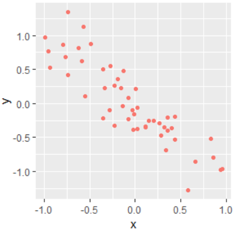


$r = 0.9$

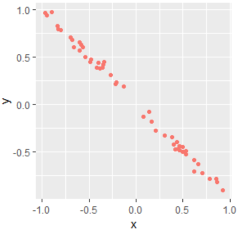
- Negative correlation



$r = -0.2$



$r = -0.6$



$r = -0.9$

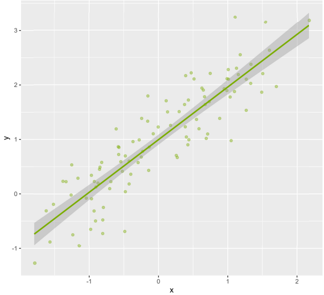
Linear Regression

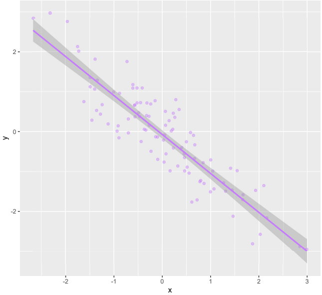
Linear Regression

- When two variables are related, one can be used to predict the value of the other
- Examples :
    1. Knowing the amount of advertising expenses, one can predict the amount of sales.
    2. Knowing the daily temperature, one can predict the amount of water consumption.
    3. Knowing Fathers heights, one can predict Sons heights

Best-fit line (or regression line)

- We could draw any number of lines, but we need to have a criteria for determining which one is "best."





- The best-fit line (or regression line) on a scatterplot is a line that lies closer to the data points than any other possible line (according to a standard statistical measure of closeness).

Linear Regression Analysis

- The relationship between the two variables is approximated by a straight line.
- A statistical procedure called regression analysis can be used to develop an equation showing how the variables are related.
- The simple linear regression model is:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Terminology : Linear Regression model

The simple linear regression model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

1. **Dependent variable (y)**
   - The variable being predicted
   - Response variable
2. **Independent variable (x)**
   - The variable being used to predict the value of the dependent variable.
   - Predictor or Explanatory variable
3. β_0 : Intercept of the regression line
4. β_1 : Slope of the regression line

Note : β_0, β_1 are called parameters of the model

Estimation

- The estimated simple linear regression equation

$$\hat{y} = b_0 + b_1 x$$

	Population Parameter	Sample Statistic
y-Intercept	β_0	b_0
Slope	β_1	b_1
Equation	$y = \beta_0 + \beta_1 x$	$\hat{y} = b_0 + b_1 x$

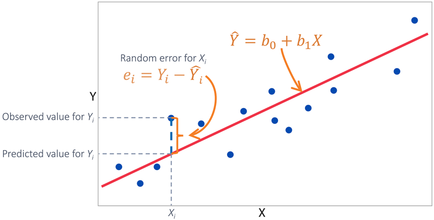
- How can we estimate b_0, b_1 ? **Least Squares methods**

Least Squares Methods

Residual

$$r_i(\text{ or } e_i) = \underbrace{y_i}_{\text{Observed value}} - \underbrace{\hat{y}_i}_{\text{Predicted value}}$$

- For a pair of sample x and y values, the residual is the difference between the observed sample value of y and the $\hat{y}$ value that is predicted by using the regression equation.



Least Squares Methods

$$\textbf{Minimize} \sum_{i=1}^{n} r_i^2 = \textbf{Minimize} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$= \textbf{Minimize} \sum_{i=1}^{n} (y_i - b_0 - b_1 x)^2$$

- A straight line satisfies the least-squares property if the sum of the squares of the residuals is the smallest sum possible.

Least Squares Methods (Formulas for b_0 and b_1)

1. **Slope :** b_1

$$\begin{aligned} b_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{n(\sum_{i=1}^{n} x_i y_i) - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \\ &= r\frac{s_y}{s_x} \end{aligned}$$

,where r is the correlation coefficient, s_x, s_y denote the standard deviation of x and y, respectively

2. y-**Intercept :** b_0

$$\begin{aligned} b_0 &= \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \\ &= \bar{y} - b_1\bar{x} \end{aligned}$$

Example : Least Squares Methods

- Data

x	1	2	4	5
y	4	24	8	32

- Calculate

x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	4				
2	24				
4	8				
5	32				

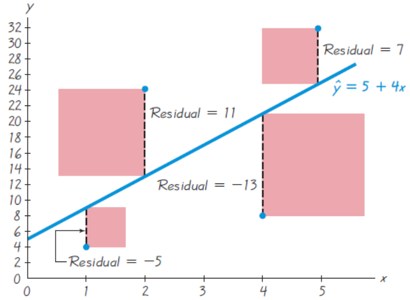
- b_1 :
- b_0 :

Example : Least Squares Methods

- The best linear line (Estimated Regression line)

$$\hat{y} = 5 + 4x$$

- Residuals and the Estimated Regression line



Calculating the Estimated Regression line

LinReg(a+bx)

1. Press the $\boxed{\text{stat}}$ and highlight **CALC**
2. Scroll down to 8: **LinReg(a+bx)**
3. Enter values
    - Xlist : L_1
    - Ylist : L_2
    - Freq :
    - Store RegEQ : Y_1

Example : Calculating the Estimated Regression line

Police sometimes use footprint evidence to estimate the height of a suspect, and the height is included in a description that becomes part of a BOLO ("be on the lookout").

Shoe Print (cm)	29.7	29.7	31.4	31.8	27.6
Height (cm)	175.3	177.8	185.4	175.3	172.7

- Find the best linear line ?

Example :Altitude and Temperature

Altitude (thousand feet)	3	10	14	22	28	31	33
Temperature	57	37	24	-5	-30	-41	-54

1. Find the regression line.

2. At 6327 ft (or 6.327 thousand feet), find the best predicted temperature.

Interpreting the Regression Equation

1. Meaning of β_1

   change in the response variable for one unit of change in the independent variable

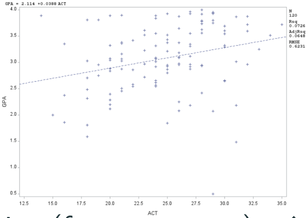
   - $\beta_1 < 0$: the change is a decrease
   - $\beta_1 > 0$: the change is an increase

2. Meaning of β_0

   - If $x = 0$ makes sense, then it is the value of y when x $= 0$.
   - If $x = 0$ does not make sense, it is just an intercept used to fit a more flexible model.

Example : ACT score and GPA

The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (y) be predicted by the ACT score (x).



- Estimated equation (from computer) : $\hat{y} = 2.114 + .0388x$
- Interpretation of β_1 :
- Predict the GPA of a student with ACT of 20 :
- Predict the GPA of a student with ACT of 30 :

Example : Blaze Pizza

Blaze Pizza

Blaze Pizza's most successful locations are near college campuses. The managers believe that quarterly sales for these restaurants are related positively to the size of the student population; that is, restaurants near campuses with a large student population tend to generate more sales than those located near campuses with a small student population.

- Dependent variable : Quarterly Sales
- Independent variable : Student population

Suppose data were collected from a sample of 10 Blaze Pizza restaurants located near college campuses

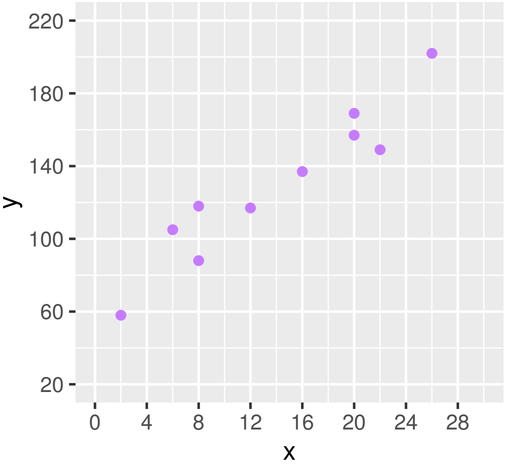
Example : Blaze Pizza

- For the ith observation or restaurant in the sample,
  - x_i is the size of the student population (in thousands)
  - y_i is the quarterly sales (in thousands of dollars).

Restaurant	Student Population (1,000)	Quarterly Sales($ 1,000)
i	x_i	y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

Example : Blaze Pizza

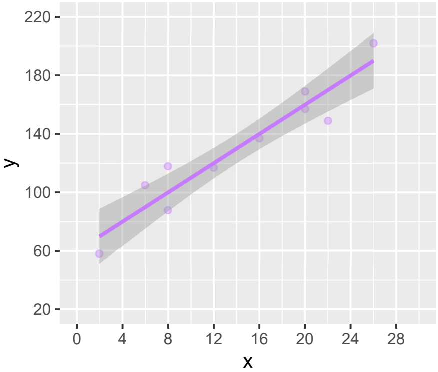
- Scatterplot



- Find the regression coefficients

Example : Blaze Pizza

- Estimated Regression line : $\hat{y} = 60 + 5x$



- The slope of the estimated regression equation $(b_1 = 5)$ is positive, implying that as student population increases, sales increase.

Example : Blaze Pizza

- In fact, we can conclude (based on sales measured in \$1,000s and student population in 1,000s) that an increase in the student population of 1,000 is associated with an increase of \$5,000 in expected sales
- that is, quarterly sales are expected to increase by \$5 per student
- If we wanted to predict quarterly sales for a restaurant to be located near a campus with 16,000 students, we would compute

$$\hat{y} =$$

- Hence, we would predict quarterly sales of \$__________ for this restaurant.