STA 1013 : Statistics through Examples

Lecture 32: Linear Regression Analysis 2, and Final review

Hwiyoung Lee

December 4, 2019

Department of Statistics, Florida State University

Linear Regression analysis 2

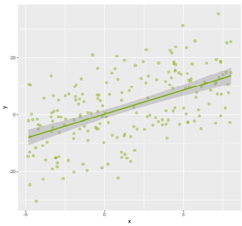
Coefficient of determination R^2

- A statistical measure of how close the data are to the fitted regression line

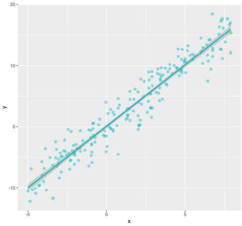
$$R^2 = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2} = \frac{\text{Variation explained by the model}}{\text{Total variation}}$$

- The proportion of the variation in a variable that is accounted for by the best-fit line
- $R^2 = \text{correlation}^2$
- R-squared is always between 0 and 1

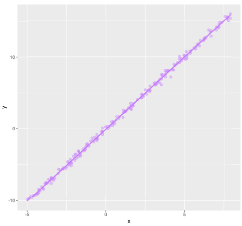
Coefficient of determination R^2



$R^2 = 0.28$



$R^2 = 0.9$



$R^2 \approx 1$

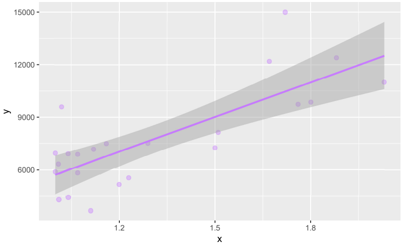
- 0% indicates that the model explains none of the variability of the response data
- 1 (or 100%) indicates that the model explains all the variability of the response data
- The higher the R-squared, the better the model fits your data

Example : Coefficient of determination R^2

TABLE 1	Prices and Characteristics of a Sample of 23 Diamonds from Gem Dealers					
Diamond	Price	Weight (carats)	Depth	Table	Color	Clarity
1	$6,958	1.00	60.5	65	3	4
2	$5,885	1.00	59.2	65	5	4
3	$6,333	1.01	62.3	55	4	4
4	$4,299	1.01	64.4	62	5	5
5	$9,589	1.02	63.9	58	2	3
6	$6,921	1.04	60.0	61	4	4
7	$4,426	1.04	62.0	62	5	5
8	$6,885	1.07	63.6	61	4	3
9	$5,826	1.07	61.6	62	5	5
10	$3,670	1.11	60.4	60	9	4
11	$7,176	1.12	60.2	65	2	3
12	$7,497	1.16	59.5	60	5	3
13	$5,170	1.20	62.6	61	6	4
14	$5,547	1.23	59.2	65	7	4
15	$7,521	1.29	59.6	59	6	2
16	$7,260	1.50	61.1	65	6	4
17	$8,139	1.51	63.0	60	6	4
18	$12,196	1.67	58.7	64	3	5
19	$14,998	1.72	58.5	61	4	3
20	$9,736	1.76	57.9	62	8	2
21	$9,859	1.80	59.6	63	5	5
22	$12,398	1.88	62.9	62	6	2
23	$11,008	2.03	62.0	63	8	3

$$\hat{y} = b_0 + b_1 \times x_{\text{weight}}$$

Example : Coefficient of determination R^2



- Estimated regression line : $\hat{y} = -873.1 + 6593.2 \cdot x_{\text{weight}}$
- $R^2 = 0.604$ which we can interpret as follows: About 0.6, or 60%, of the variation in the diamond prices is accounted for by the best-fit line relating weight and price.
- That leaves 40% of the variation in price that must be due to other factors, presumably such things as depth, table, color, and clarity

Note : Multiple Regression

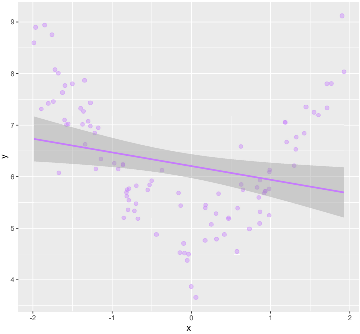
- Multiple regression is an extension of simple linear regression
- It is used when we want to predict the response variable based on the value of two or more other explanatory variables
- Model : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$
- For example :

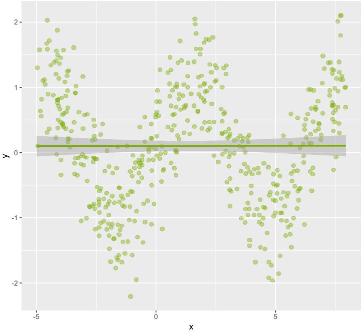
$$\underbrace{y}_{\text{Price of Diamond}} = \beta_0 + \beta_1 x_{\text{weight}} + \beta_2 x_{\text{Depth}} + \beta_3 x_{\text{table}} + \beta_4 x_{\text{Color}} + \beta_5 x_{\text{Clarity}}$$

- More accurate result

Nonlinear pattern

DO NOT use the least-squares line when the relationship between x and y is not linear





- $\hat{y} = 6.2047 - 0.2639x$
- $R^2 = 0.0630$

- $\hat{y} = 0.1029 + 0.0004x$
- $R^2 = 4.201e - 06 \approx 0$

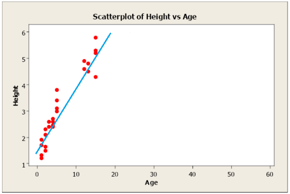
Look at the scatterplot before fitting the data.

Extrapolation

DO NOT extrapolate the fitted line outside the range of the data.

- Extrapolation is a type of estimation, beyond the original observation range
- The linear relationship may not hold there.

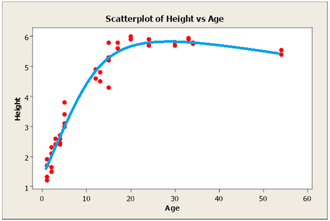
Example : Age vs Height



- We only observed data (Age $0 \sim 20$)
- If we extrapolate the best-fit lines as drawn, the average height will greater than 8.

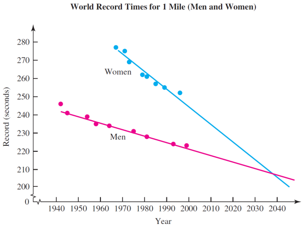
Example : Height and Age

- The linear relationship is not hold



Example : Will Women Be Faster Than Men?

The Figure shows data and best-fit lines for both men's and women's world record times in the 1-mile race.



- If we accept the best-fit lines as drawn, the women's world record will equal the men's world record by about 2040.
- However, this is **not** a valid prediction because it is based on extending the best-fit lines beyond the range of the actual data

Review and Practice problems

Statistical Hypothesis for μ

- Two-tailed hypothesis test :

$$H_0 : \mu = \mu_0$$
$$H_a : \mu \neq \mu_0$$

- Left-tailed hypothesis test :

$$H_0 : \mu = \mu_0$$
$$H_a : \mu < \mu_0$$

- Right-tailed hypothesis test :

$$H_0 : \mu = \mu_0$$
$$H_a : \mu > \mu_0$$

Z-test for μ (when σ is known)

Test Statistic

- **Test Statistic under H_0 (when σ known):**

$$z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

- We will reject H_0 if our test statistic (z_0) is **too "extreme."**
- "Too extreme" means the test statistic is located in the **rejection region**

Rejection Regions

1. Left-Tailed test :

$$\textbf{Rejection Region} < -z_\alpha$$

2. Right-Tailed test :

$$\textbf{Rejection Region} > z_\alpha$$

3. Two-Tailed test :

$$\textbf{Rejection Region} > z_{\alpha/2} \text{ or } \textbf{Rejection Region} < -z_{\alpha/2}$$

Commonly used critical values

- $z_{0.1} = 1.28$
- $z_{0.05} = 1.645$
- $z_{0.025} = 1.96$
- $z_{0.01} = 2.326$
- $z_{0.005} = 2.576$

P-value

1. Left-Tailed test

$$\textbf{P-value} \;=\; P(Z \leqslant z_0)$$

2. Right-Tailed test

$$\textbf{P-value} \;=\; P(Z \geqslant z_0)$$

3. Two-Tailed test

$$\begin{aligned}\textbf{P-value} \;&=\; P(Z \leqslant -|z_0| \text{ or } Z \geqslant |z_0|)\\ &= 2 * P(Z \geqslant |z_0|)\end{aligned}$$

Calculator (Z-Test)

Z-Test from statistics

1. Press the $\boxed{\text{stat}}$ and highlight **TESTS**
2. Scroll down to 1: **Z-Test**
3. Inpt : Highlight **Stats**
4. Enter values for
    - μ_0 : Claimed value in the null hypothesis (H_0)
    - σ : Population standard deviation
    - $\bar{X}$: Sample mean
    - n : sample size
    - μ : Select the test type ($\underbrace{\neq \mu_0}_{Two-Tailed}$, $\underbrace{< \mu_0}_{Left-Tail}$, $\underbrace{> \mu_0}_{Right-Tail}$)

Calculator (Z-Test)

Z-Test from Data

1. Press the $\boxed{\text{stat}}$ and highlight **TESTS**
2. Scroll down to 1: **Z-Test**
3. Inpt : Highlight **Data**
4. Enter values for
   - μ_0 : Claimed value in the null hypothesis (H_0)
   - σ : Population standard deviation
   - List : Data (ex : L_1)
   - μ : Select the test type ($\underbrace{\neq \mu_0}_{Two-Tailed}$, $\underbrace{< \mu_0}_{Left-Tail}$, $\underbrace{> \mu_0}_{Right-Tail}$)

Z-test : Exerciese 1

Alternative Hypothesis supported? ($\alpha = 0.05$)

1. $H_a : \mu < 75, n = 100, \bar{X} = 70, \sigma = 15$
2. $H_a : \mu < 75, n = 36, \bar{X} = 72, \sigma = 15$
3. $H_a : \mu > 12, n = 64, \bar{X} = 14, \sigma = 2$
4. $H_a : \mu > 1007, n = 225, \bar{X} = 1021, \sigma = 35$
5. $H_a : \mu \neq 2.55, n = 100, \bar{X} = 2.58, \sigma = 0.29$
6. $H_a : \mu \neq 156.2, n = 225, \bar{X} = 155.5, \sigma = 29$

Z-test : Exerciese 2

1. When $\sigma = 5, n = 100, z_0 = 4, \mu_0 = 10$, Find $\bar{x}$?

2. When $\sigma = 16, n = 64, z_0 = 4, \bar{x} = 9$, Find μ_0 ?

Z-test : Exerciese 3

1. Suppose we perform two-tailed test with $\alpha = 0.05$
   - $H_0 = 50$
   - $H_0 \neq 50$

   and $\bar{x} = 51, \sigma = 5$ are fixed. How many sample do we need to reject the H_0 ?

Z-test : Exerciese 3

2. Suppose we perform one-tailed test with $\alpha = 0.05$
    - $H_0 = 10$
    - $H_0 > 10$

   and $\bar{x} = 10.2, \sigma = 2$ are fixed. How many sample do we need to reject the H_0 ?

Z-test : Exerciese 3

3. Suppose we perform one-tailed test with $\alpha = 0.1$
   - $H_0 = 5$
   - $H_0 < 5$

   and $\bar{x} = 4, \sigma = 10$ are fixed. How many sample do we need to reject the H_0 ?

Z-test : Exerciese 4

(Weights of Bears) The health of the bear population in Yellowstone National Park is monitored by periodic measurements taken from anesthetized bears. A sample of 54 bears has a mean weight of 182.9 lb. Assuming that σ is known to be 121.8 lb, test the claim that the population mean of all such bear weights is greater than 150 lb ($\alpha = 0.05$)

1. State the hypothesis

2. Find the p-value

3. Conclusion

Z-test : Exerciese 5

Coke

Randomly selected cans of Coke are measured for the amount of cola, in ounces. The sample values listed below.

12.3	12.1	12.2	12.3	12.2	12.3	12.0	12.1	12.2
12.1	12.3	12.3	11.8	12.3	12.1	12.1	12.0	12.2
12.2	12.2	12.2	12.2	12.2	12.4	12.2	12.2	12.3
12.2	12.2	12.3	12.2	12.2	12.1	12.4	12.2	12.2

Assume that we want to use a 0.05 significance level to test the **claim that cans of Coke have a mean amount of cola greater than 12 ounces.** Assume that the population has a standard deviation of $\sigma = 0.115$ ounce.

1. **State the Hypothesis**
2. **Perform Z-test**

t-test for μ (when σ is unknown)

Test Statistic

When we don't know the population standard deviation σ,

- Use t distirbution
- **Test statistic for t-test :**

$$t_0 = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

Rejection Region Approach

1. Left-Tail Test

$$\textbf{Rejection Region} < -t_{\alpha,df}$$

2. Right-Tail Test

$$\textbf{Rejection Region} > t_{\alpha,df}$$

3. Two-Tailed Test

$$\textbf{Rejection Region} < -t_{\alpha/2,df} \text{ or } \textbf{Rejection Region} > t_{\alpha/2,df}$$

- Rejection region approach depends on t distribution
- df = n-1

P-value approach

1. Left-Tailed test

$$\textbf{P-value} = P(T_{df} \leqslant t_0)$$

2. Right-Tailed test

$$\textbf{P-value} = P(T_{df} \geqslant t_0)$$

3. Two-Tailed test

$$\textbf{P-value} = 2 * P(T_{df} \geqslant |t_0|)$$

- Reject H_0 when : P-value $\leqslant \alpha$
- Can't reject H_0 when : P-value $> \alpha$

T-Test from Statistics

1. Press the $\boxed{\text{stat}}$ and highlight **TESTS**
2. Scroll down to 2: **T-Test**
3. Inpt : Highlight **Stats**
4. Enter values for
   - μ_0 : Claimed value in the null hypothesis (H_0)
   - $\bar{X}$: Sample mean
   - S_x : Sample standard deviation
   - n : sample size
   - μ : Select the test type ($\underbrace{\neq \mu_0}_{Two-Tailed}$, $\underbrace{< \mu_0}_{Left-Tail}$, $\underbrace{> \mu_0}_{Right-Tail}$)

T-Test from Data

1. Press the $\boxed{\text{stat}}$ and highlight **TESTS**
2. Scroll down to 2: **T-Test**
3. Inpt : Highlight **Data**
4. Enter values for
   - μ_0 : Claimed value in the null hypothesis (H_0)
   - List : Data (ex : L_1)
   - μ : Select the test type ($\underbrace{\neq \mu_0}_{Two-Tailed}$, $\underbrace{< \mu_0}_{Left-Tail}$, $\underbrace{> \mu_0}_{Right-Tail}$)

T-test : Exercise 1

(Brain Volume) Listed below are brain volumes (cm^3) of unrelated subjects used in a study. Use a 0.1 significance level to test the **claim that the population of brain volumes has a mean equal to** $1100.0cm^3$**.** Data : 963, 1027, 1272, 1079, 1070, 1173, 1067, 1347, 1100, 1204

(Multiple choice). Find the critical value (boundary of the rejection region) ?

1. $t_{0.1,9}$
2. $t_{0.1,10}$
3. $t_{0.05,9}$
4. $t_{10.05,10}$

Note : $t_{\text{tail probability,df}}$

T-test : Exercise 2

ages of Race Car drivers Listed below are the ages (years) of randomly selected race car drivers (based on data reported in USa Today). Use a 0.05 significance level to test the **claim that the mean age of all race car drivers is greater than 30 years.**

Data : 32, 32, 33, 33, 41, 29, 38, 32, 33, 23, 27, 45, 52, 29, 25

1. State the hypothesis
2. Find the p-value
3. Conclusion

Hypothesis Tests For Population Proportions

Z-test for p

Test Statistic

Under the H_0, our z test statistic is given by,

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0,1)$$

Hypothesis Tests For Population Proportions

- We now consider hypothesis testing with proportions

$$H_0 : p = p_0$$
$$H_a : p \neq p_0 \text{ Two-Tailed test}$$
$$H_a : p < p_0 \text{ Left-Tailed test}$$
$$H_a : p > p_0 \text{ Right-Tailed test}$$

- All the ideas from previous tests apply
- p (or π) : denotes the population proportion
- $\hat{p}$: denotes the sample proportion

Calculator (1-PropZTest)

Z-Test for proportion

1. Press the $\boxed{\text{stat}}$ and highlight **TESTS**
2. Scroll down to 5: **1-PropZTest**
3. Enter values for
   - p_0 : Claimed proportion in the null hypothesis (H_0)
   - x : Number of success
   - n : Sample size
   - prop : Select the test type

     $\neq p_0$: Two-Tailed
     $< p_0$: Left-Tail
     $> p_0$: Right-Tail

Proportion Z test : Exerciese 1

Under the H_0

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0,1)$$

What is the distribution of $\hat{p}$?

Proportion Z test : Exerciese 2

(Voter Poll) In a pre-election poll, a candidate for district attorney receives 250 of 400 votes. Assuming that the people polled represent a random sample of the voting population, test the claim that a majority of voters support the candidate.

1. State the hypothesis

2. Find the test statistic

3. Find the p-value

4. Conclusion

Proportion Z test : Exerciese 3

(Poverty) According to recent estimates, 12.6% of the 4,342 people in Custer County, Idaho, live in poverty. Assume that the people in this county represent a random sample of all people in Idaho. Based on this sample, test the claim that the poverty rate in Idaho is less than the national rate of 13.3%. ($\alpha = 0.1$)

1. State the hypothesis

2. Find the test statistic

3. Find the rejection region

4. Conclusion

Linear Regression

Linear Regression : Exercise 1

Does it make sense ? (**Yes** or **No**)

1. Suppose $s_x = s_y$, then $b_1 > 1$

2. Suppose $r < 0$, then $b_1 > 0$

Linear Regression : Exercise 2

1. Suppose $b_1 = 3$, $s_x = 1$, and $s_y = 6$, Find r

2. Suppose $r = 0.9$, $s_x = 3$, and $s_y = 5$, Find b_1

3. Suppose $r = -0.5$, $b_1 = -3$, and $s_y = 6$, Find s_x

4. Suppose $\hat{y} = 5 - 0.4x$, and $R^2 = 0.81$, Find r

Linear Regression : Exercise 3

1. Suppose $b_0 = 3$, $\bar{x} = 2$, $\bar{y} = 5$, find b_1

2. Suppose $b_0 = 5$, $b_1 = 2$, $\bar{y} = -3$, find $\bar{x}$

3. Suppose $b_0 = 10, \bar{y} = 4, \bar{x} = 2, s_x = 3, s_y = 10$, find r

Thank you