Lesson 03 Intro to Linear Regression Models

Lusine Zilfimian

February 24 (Monday), 2020

Contents

Contents

Contents

Contents

- Did you see the HW?

- Did you see the HW?
- Are you shocked?

- Did you see the HW?
- Are you shocked?
- Well, we will have OH this Wednesday (instead of Shiny?)?

Last Lecture ReCap

- Bring an example of the power of the Visualization.

Last Lecture ReCap

- Bring an example of the power of the Visualization.
- Is the Correlation Coefficient a better measurement of the relationship between variables than Covariation?

Last Lecture ReCap

- Bring an example of the power of the Visualization.
- Is the Correlation Coefficient a better measurement of the relationship between variables than Covariation?
- Give some structural differences between the Barplot and Histogram.

Intro to Simple Linear Regression

Suggested materials to read (master) regression

- G. James, D. Witten, et al., An Introduction to Statistical Learning, Chapter 3, 7

Intro to Simple Linear Regression

Suggested materials to read (master) regression

- G. James, D. Witten, et al., An Introduction to Statistical Learning, Chapter 3, 7
- T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Chapter 3, 5

Intro to Simple Linear Regression

Suggested materials to read (master) regression

- G. James, D. Witten, et al., An Introduction to Statistical Learning, Chapter 3, 7
- T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Chapter 3, 5
- Magnus J., et al., Introduction to Econometrics, Chapter 2, 3, 4

The Phenomenon of Regression

- Do you believe that your talent and beauty must be inherited by your children?

The Phenomenon of Regression

- Do you believe that your talent and beauty must be inherited by your children?
- **Sir Francis Galton** was interested in the heredity of human characteristics.

The Phenomenon of Regression

- Do you believe that your talent and beauty must be inherited by your children?
- **Sir Francis Galton** was interested in the heredity of human characteristics.
- Galton observed that the characteristics of parents are not fully transmitted to their children.

The Phenomenon of Regression

- Do you believe that your talent and beauty must be inherited by your children?
- **Sir Francis Galton** was interested in the heredity of human characteristics.
- Galton observed that the characteristics of parents are not fully transmitted to their children.
- Galton found that children of short parents tended to be shorter than average, while children of tall parents tended to be taller than average, "however, the transmission of height across generations was imperfect."

The Phenomenon of Regression

- Do you believe that your talent and beauty must be inherited by your children?
- **Sir Francis Galton** was interested in the heredity of human characteristics.
- Galton observed that the characteristics of parents are not fully transmitted to their children.
- Galton found that children of short parents tended to be shorter than average, while children of tall parents tended to be taller than average, "however, the transmission of height across generations was imperfect."
- Original data from Galton's notebook http://www.medicine.mcgill.ca/epidemiology/hanley/galton/notebook/index.html lists 963 children in 205 families ranging from 1-15 adult children.

The Phenomenon of Regression

- Do you believe that your talent and beauty must be inherited by your children?
- **Sir Francis Galton** was interested in the heredity of human characteristics.
- Galton observed that the characteristics of parents are not fully transmitted to their children.
- Galton found that children of short parents tended to be shorter than average, while children of tall parents tended to be taller than average, “however, the transmission of height across generations was imperfect.”
- Original data from Galton’s notebook http://www.medicine.mcgill.ca/epidemiology/hanley/galton/notebook/index.html lists 963 children in 205 families ranging from 1-15 adult children.
- You can find Galton dataset `GaltonFamilies` from the package `HistData`.

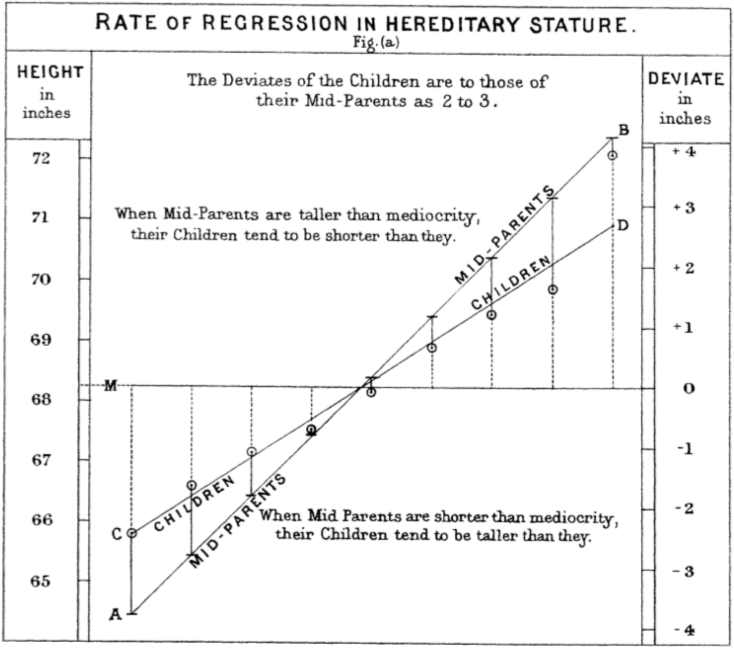
Source: Galton's classic paper *"Regression Towards Mediocrity in Hereditary Stature"*:
http://www.stat.ucla.edu/~nchristo/statistics100C/history_regression.pdf

TABLE I.

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.
(All Female heights have been multiplied by 1·08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.


RATE OF REGRESSION IN HEREDITARY STATURE.
Fig. (a)
HEIGHT in inches
DEVIATE in inches
The Deviates of the Children are to those of their Mid-Parents as 2 to 3.
When Mid-Parents are taller than mediocrity, their Children tend to be shorter than they.
When Mid Parents are shorter than mediocrity, their Children tend to be taller than they.
MID-PARENTS
CHILDREN
72
71
70
69
68
67
66
65
+4
+3
+2
+1
0
-1
-2
-3
-4
A
B
C
D
M

Simple (Modern) Linear Regression

- Relationship between **one continuous dependent** variable and **one or more (any) explanatory variables**

Simple (Modern) Linear Regression

- Relationship between **one continuous dependent** variable and **one or more (any) explanatory variables**
- Simple LR - one explanatory variable

Simple (Modern) Linear Regression

- Relationship between **one continuous dependent** variable and **one or more (any) explanatory variables**
- Simple LR - one explanatory variable
- Multiple LR - two or more explanatory variables

Simple (Modern) Linear Regression

- Relationship between **one continuous dependent** variable and **one or more (any) explanatory variables**
- Simple LR - one explanatory variable
- Multiple LR - two or more explanatory variables
- Used for prediction and estimation

Simple linear regression

True relationship between X and Y:

- $$Y = f(X) + \varepsilon,$$

Simple linear regression

True relationship between X and Y:

- $$Y = f(X) + \varepsilon,$$
- ε-mean-zero random error term

Simple linear regression

True relationship between X and Y:

- $$Y = f(X) + \varepsilon,$$
- ε-mean-zero random error term
- Approximately linear relationship between X and Y.

Simple linear regression

True relationship between X and Y:

- $$Y = f(X) + \varepsilon,$$
- ε-mean-zero random error term
- Approximately linear relationship between X and Y.
- $$Y \approx \beta_0 + \beta_1 X + \varepsilon$$

Simple linear regression

True relationship between X and Y:

- $$Y = f(X) + \varepsilon,$$
- ε-mean-zero random error term
- Approximately linear relationship between X and Y.
- $$Y \approx \beta_0 + \beta_1 X + \varepsilon$$
- **Unknown** constants: β_0-intercept, β_1-slope.

Reasons for having ε:

- The true relationship is probably not linear

Reasons for having ε:

- The true relationship is probably not linear
- Other variables that cause variation in Y

Reasons for having ε:

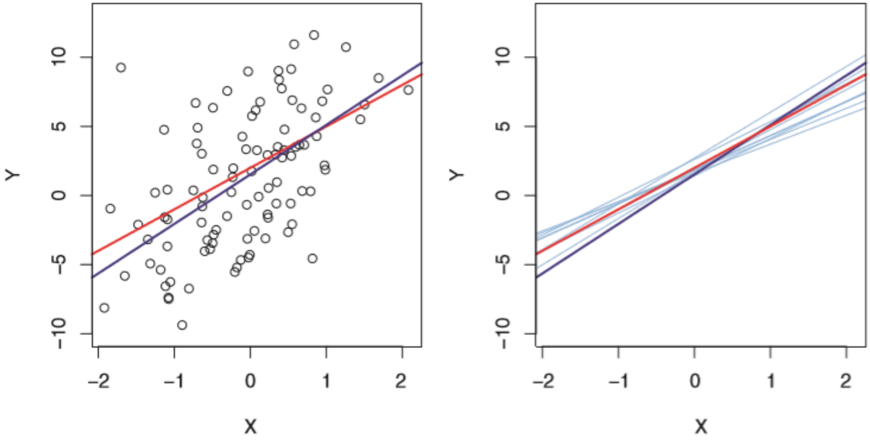
- The true relationship is probably not linear
- Other variables that cause variation in Y
- Estimated model:

Reasons for having ε:

- The true relationship is probably not linear
- Other variables that cause variation in Y
- Estimated model:
- $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\varepsilon}$$

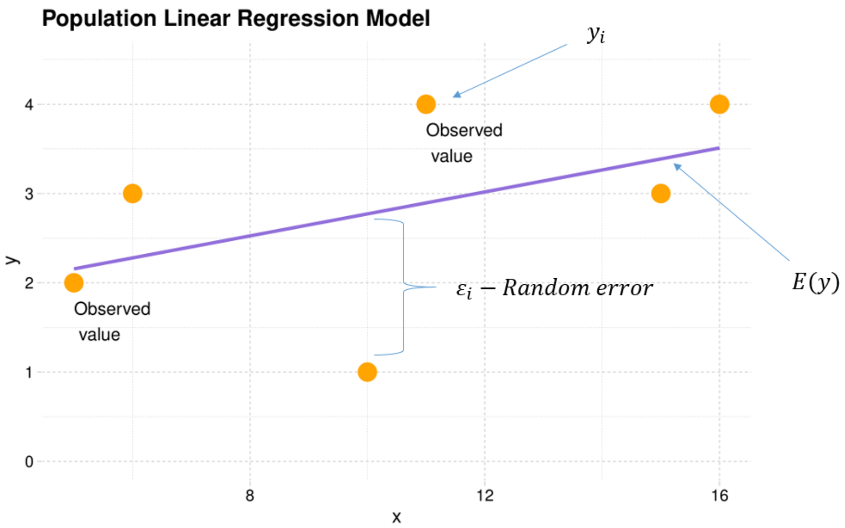
Population regression line vs least squares line

Source: G. James, D. Witten, et al., *An Introduction to Statistical Learning*

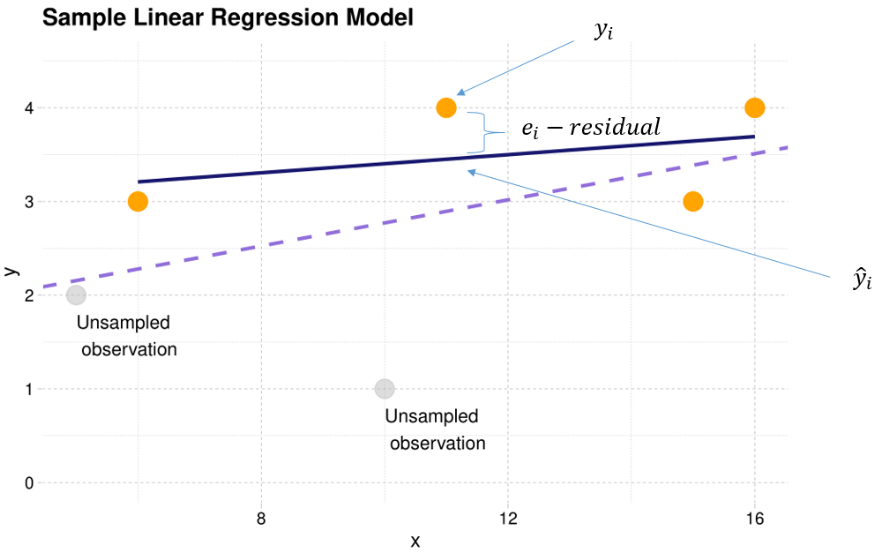


- Least squares line is computed using set of observations, however, the population regression line is unobserved.

Population Linear Regression Model



Sample Linear Regression Model



Estimating the Coefficients

- The most common approach – minimizing the least squares criterion.

Estimating the Coefficients

- The most common approach – minimizing the least squares criterion.
- Residual for i-th observation: $e_i = y_i - \hat{y}_i$

Estimating the Coefficients

- The most common approach – minimizing the least squares criterion.
- Residual for i-th observation: $e_i = y_i - \hat{y}_i$
- Residual sum of squares (RSS):

Estimating the Coefficients

- The most common approach – minimizing the least squares criterion.
- Residual for i-th observation: $e_i = y_i - \hat{y}_i$
- Residual sum of squares (RSS):
- $$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)^2 \rightarrow min$$

Estimating the Coefficients

- The most common approach – minimizing the least squares criterion.
- Residual for i-th observation: $e_i = y_i - \hat{y}_i$
- Residual sum of squares (RSS):
- $$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)^2 \rightarrow min$$
- The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS.

Estimating the Coefficients

- The most common approach – minimizing the least squares criterion.
- Residual for i-th observation: $e_i = y_i - \hat{y}_i$
- Residual sum of squares (RSS):
- $$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)^2 \rightarrow min$$
- The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS.
- $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{cov(x, y)}{var(x)}$$

Estimating the Coefficients

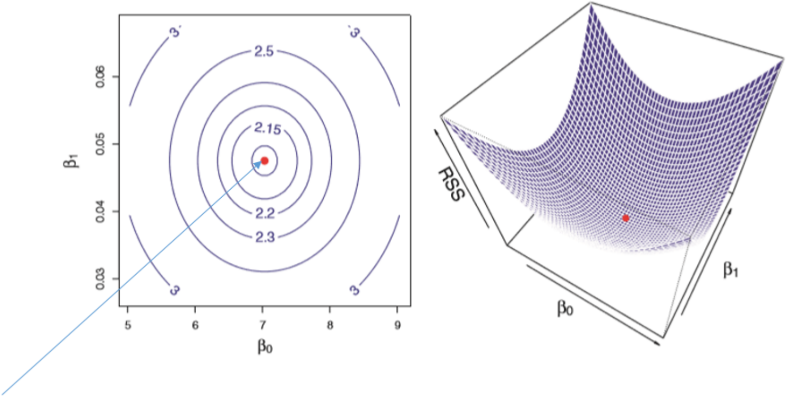
- The most common approach – minimizing the least squares criterion.
- Residual for i-th observation: $e_i = y_i - \hat{y}_i$
- Residual sum of squares (RSS):
- $$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)^2 \rightarrow min$$
- The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS.
- $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{cov(x,y)}{var(x)}$$
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Estimating the Coefficients

- The most common approach – minimizing the least squares criterion.
- Residual for i-th observation: $e_i = y_i - \hat{y}_i$
- Residual sum of squares (RSS):
- $$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)^2 \rightarrow min$$
- The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS.
- $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{cov(x, y)}{var(x)}$$
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- **Prove that:** $corr(x, y) = \hat{\beta}_1 \sqrt{\frac{var(x)}{vary}}$

Graphical representation

Source: G. James, D. Witten, et al., *An Introduction to Statistical Learning*



- The pair of least squares estimates $(\hat{\beta}_0, \hat{\beta}_1)$

How accurate are the coefficients?

- The average amount that $\hat{\beta}$ differs from the actual value β:

How accurate are the coefficients?

- The average amount that $\hat{\beta}$ differs from the actual value β:
- $$var(\hat{\beta}_0) = \sigma^2\{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\}$$

How accurate are the coefficients?

- The average amount that $\hat{\beta}$ differs from the actual value β:
- $$var(\hat{\beta}_0) = \sigma^2\{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\}$$
- $$var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

How accurate are the coefficients?

- The average amount that $\hat{\beta}$ differs from the actual value β:
- $$var(\hat{\beta}_0) = \sigma^2\{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\}$$
- $$var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
- $\sigma^2 = var(\varepsilon)$

How accurate are the coefficients?

- The average amount that $\hat{\beta}$ differs from the actual value β:
- $$var(\hat{\beta}_0) = \sigma^2\{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\}$$
- $$var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
- $\sigma^2 = var(\varepsilon)$
- $$var(x) \uparrow \Rightarrow \ var(\hat{\beta}) \downarrow$$

How accurate are the coefficients?

- The average amount that $\hat{\beta}$ differs from the actual value β:
- $$var(\hat{\beta}_0) = \sigma^2\{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\}$$
- $$var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
- $\sigma^2 = var(\varepsilon)$
- $$var(x) \uparrow \Rightarrow \; var(\hat{\beta}) \downarrow$$
- **Task**: Derive all formulas by yourself.

Hypothesis tests on the coefficients

- $$H_0 : \beta_1 = 0 \; H_\alpha : \beta_1 \neq 0$$

Hypothesis tests on the coefficients

- $$H_0 : \beta_1 = 0 \; H_\alpha : \beta_1 \neq 0$$
- If $\beta_1 = 0 \Rightarrow Y \approx \beta_0 + \varepsilon \Rightarrow$ there is no relationship between X and Y.

Hypothesis tests on the coefficients

- $$H_0 : \beta_1 = 0 \ H_\alpha : \beta_1 \neq 0$$
- If $\beta_1 = 0 \ \Rightarrow Y \approx \beta_0 + \varepsilon \Rightarrow$ there is no relationship between X and Y.
- $$\mathbb{P}(Reject\ H_0 | H_0\ is\ true) = \mathbb{P}(|t| \geq c)$$

Hypothesis tests on the coefficients

- $$H_0 : \beta_1 = 0 \; H_\alpha : \beta_1 \neq 0$$
- If $\beta_1 = 0 \;\Rightarrow Y \approx \beta_0 + \varepsilon \Rightarrow$ there is no relationship between X and Y.
- $$\mathbb{P}(Reject\ H_0 | H_0\ is\ true) = \mathbb{P}(|t| \geq c)$$
- $c = t_{n-(k+1),\frac{\alpha}{2}}$

Hypothesis tests on the coefficients

- $$H_0 : \beta_1 = 0 \ H_\alpha : \beta_1 \neq 0$$
- If $\beta_1 = 0 \ \Rightarrow Y \approx \beta_0 + \varepsilon \Rightarrow$ there is no relationship between X and Y.
- $$\mathbb{P}(Reject\ H_0 | H_0\ is\ true) = \mathbb{P}(|t| \geq c)$$
- $c = t_{n-(k+1), \frac{\alpha}{2}}$
- Test statistics: $t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$

Hypothesis tests on the coefficients

- $$H_0 : \beta_1 = 0 \; H_\alpha : \beta_1 \neq 0$$
- If $\beta_1 = 0 \;\Rightarrow Y \approx \beta_0 + \varepsilon \Rightarrow$ there is no relationship between X and Y.
- $$\mathbb{P}(Reject\ H_0 | H_0\ is\ true) = \mathbb{P}(|t| \geq c)$$
- $c = t_{n-(k+1), \frac{\alpha}{2}}$
- Test statistics: $t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$
- If the Null hypothesis is true:

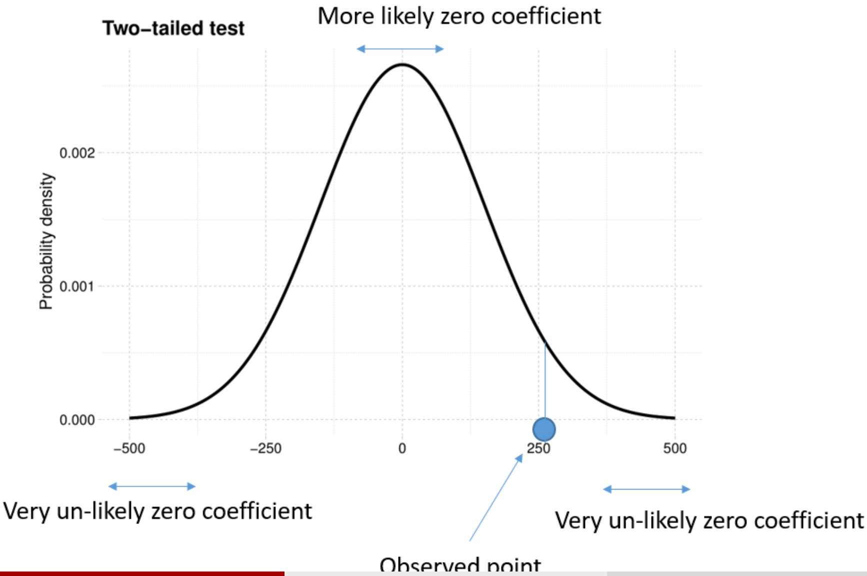
Hypothesis tests on the coefficients

- $$H_0 : \beta_1 = 0 \; H_\alpha : \beta_1 \neq 0$$
- If $\beta_1 = 0 \; \Rightarrow Y \approx \beta_0 + \varepsilon \Rightarrow$ there is no relationship between X and Y.
- $$\mathbb{P}(\mathit{Reject}\ H_0 | H_0\ \mathit{is\ true}) = \mathbb{P}(|t| \geq c)$$
- $c = t_{n-(k+1), \frac{\alpha}{2}}$
- Test statistics: $t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$
- If the Null hypothesis is true:
- $$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

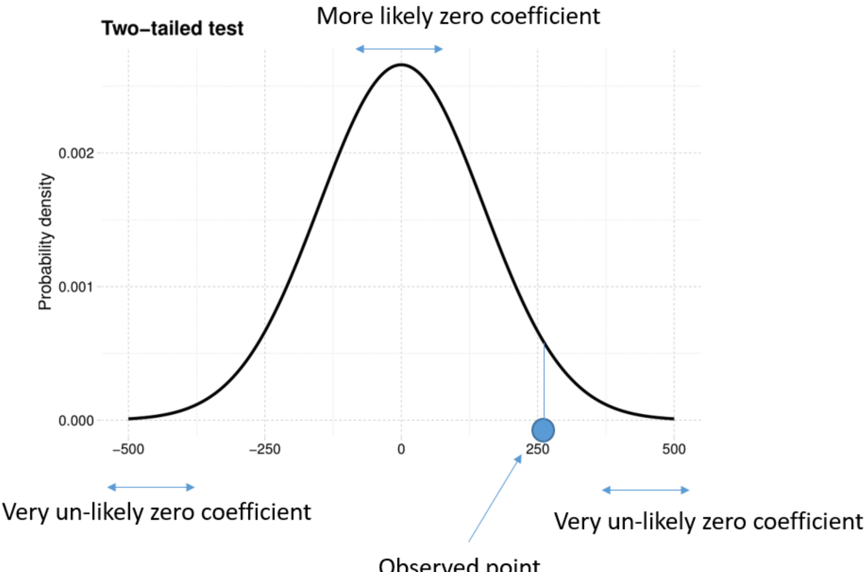
Hypothesis tests on the coefficients

- $$H_0 : \beta_1 = 0 \; H_\alpha : \beta_1 \neq 0$$
- If $\beta_1 = 0 \;\Rightarrow Y \approx \beta_0 + \varepsilon \Rightarrow$ there is no relationship between X and Y.
- $$\mathbb{P}(Reject\ H_0 | H_0\ is\ true) = \mathbb{P}(|t| \geq c)$$
- $c = t_{n-(k+1), \frac{\alpha}{2}}$
- Test statistics: $t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$
- If the Null hypothesis is true:
- $$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$
- $$|t| = \frac{|\hat{\beta}_1|}{SE(\hat{\beta}_1)} > t_{n-(k+1), \frac{\alpha}{2}}$$

- p-value - the probability of observing any value equal to $|t|$ or larger ($\mathbb{P} > |t|$)



- p-value - the probability of observing any value equal to $|t|$ or larger ($\mathbb{P} > |t|$)



Accuracy of the model

1. Residual standard error

Accuracy of the model

1. Residual standard error

- $$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Accuracy of the model

1. Residual standard error

- $$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$
- High RSE $\Rightarrow$ the model does not fit the data well

Accuracy of the model

1. Residual standard error
   - $$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$
   - High RSE $\Rightarrow$ the model does not fit the data well
2. R^2

Accuracy of the model

1. Residual standard error

- $$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- High RSE $\Rightarrow$ the model does not fit the data well

2. R^2

- High R^2 $\Rightarrow >$ the model fits the data well

Accuracy of the model

1. Residual standard error
   - $$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$
   - High RSE $\Rightarrow$ the model does not fit the data well
2. R^2
   - High R^2 $\Rightarrow >$ the model fits the data well
3. F test

R-squared

R^2 - the proportion of variability in Y that can be explained using X.

- The total variance of the response variable:

R-squared

R^2 - the proportion of variability in Y that can be explained using X.

- The total variance of the response variable:
- $$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

R-squared

R^2 - the proportion of variability in Y that can be explained using X.

- The total variance of the response variable:
- $$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$
- The estimated variance of the response variable:

R-squared

R^2 - the proportion of variability in Y that can be explained using X.

- The total variance of the response variable:
- $$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$
- The estimated variance of the response variable:
- $$ESS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

R-squared

R^2 - the proportion of variability in Y that can be explained using X.

- The total variance of the response variable:
- $$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$
- The estimated variance of the response variable:
- $$ESS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$
- $$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS}$$

R-squared

R^2 - the proportion of variability in Y that can be explained using X.

- The total variance of the response variable:
- $$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$
- The estimated variance of the response variable:
- $$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$
- $$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS}$$
- By definition $0 \leq R^2 \leq 1$

Ideas for project

- Piecewise Polynomials, Regression Splines

Ideas for project

- Piecewise Polynomials, Regression Splines
- Relaxing the assumtions of G-M theorem