LINEAR REGRESSION

Week05

Linear Regression

Is the Regression Model Significant?

- Modeling learning is not the end of the analysis
  - Check overall significance in regression models
    - Whether the regression model is overall significant for predicting a target
  - Check significance of regression coefficients
    - Whether the specific variable is significant for predicting a target
- In the case of simple linear regression, testing overall significance of the model is the same as testing significance of regression coefficients
  - Because only one explanatory variable is used

Test Concerning Regression Coefficients

- Test for $\beta_j (j = 0,1,2,\cdots,p)$
  - Hypothesis

$$H_0: \beta_j = 0$$
$$H_1: \beta_j \neq 0$$

  - Test statistic

$$t_j = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

standard error

    - $se^2(\hat{\boldsymbol{\beta}}) = MSE(\mathbf{X}^T\mathbf{X})^{-1} \rightarrow se^2(\hat{\beta}_j) = [MSE(\mathbf{X}^T\mathbf{X})^{-1}]_{j,j}$

  - Decision rule

If $|t_j| \leq t\left(1 - \frac{\alpha}{2}; n - p - 1\right)$, conclude H_0

If $|t_j| > t\left(1 - \frac{\alpha}{2}; n - p - 1\right)$, conclude H_1

Test Concerning Regression Coefficients

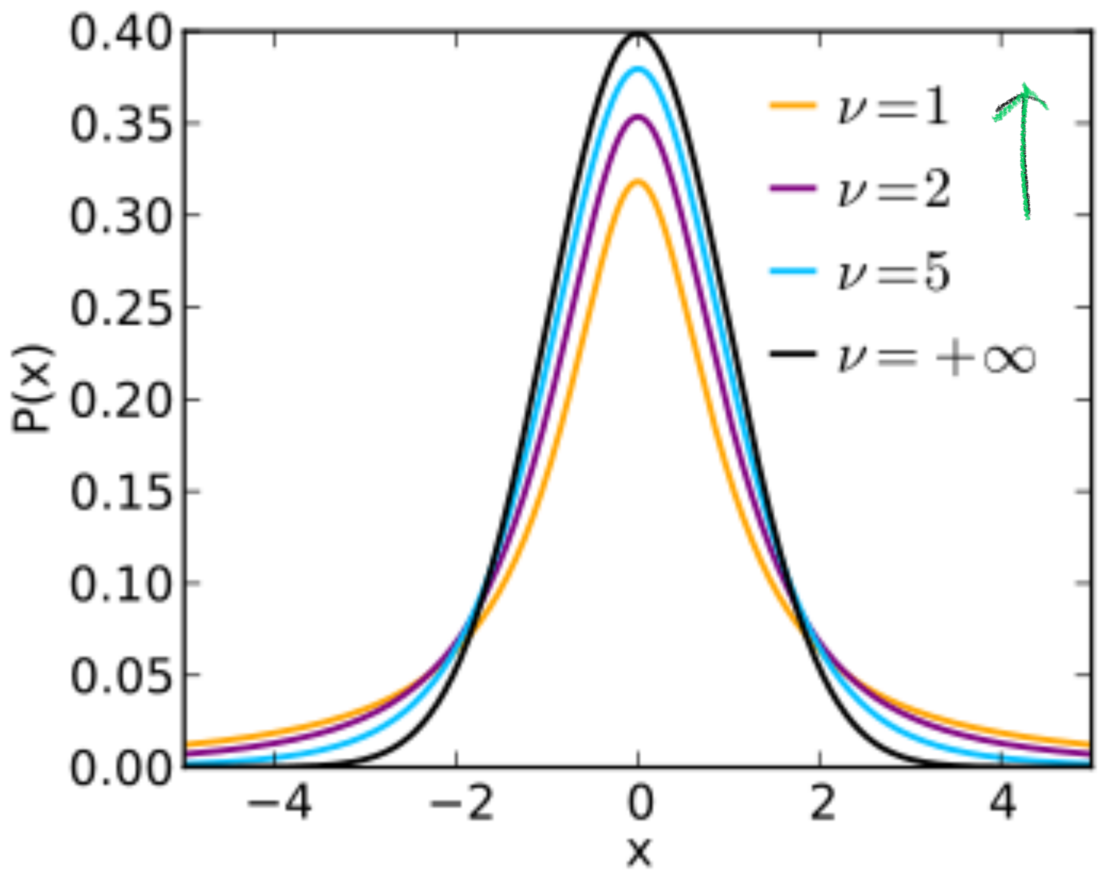
- $se^2(\hat{\beta}_i)$
  - Ex) two input variables

$$
\begin{array}{c}
\begin{array}{ccc} \beta_0 & \beta_1 & \beta_2 \end{array} \\
(X^T X)^{-1} = \begin{bmatrix} x_{00} & & \\ & x_{11} & \\ & & x_{22} \end{bmatrix}
\end{array}
$$

- $se^2(\hat{\beta}_0) = MSE \cdot x_{00} \rightarrow se(\hat{\beta}_0) = \sqrt{MSE \cdot x_{00}}$
- $se^2(\hat{\beta}_1) = MSE \cdot x_{11} \rightarrow se(\hat{\beta}_1) = \sqrt{MSE \cdot x_{11}}$
- $se^2(\hat{\beta}_2) = MSE \cdot x_{22} \rightarrow se(\hat{\beta}_2) = \sqrt{MSE \cdot x_{22}}$

Test Concerning Regression Coefficients

- Test statistics of t-test follows student's t distribution with $n - p - 1$ degree of freedom
  - Probability density function of student's t distribution with different parameters(degree of freedom)



작을수록 뾰족.

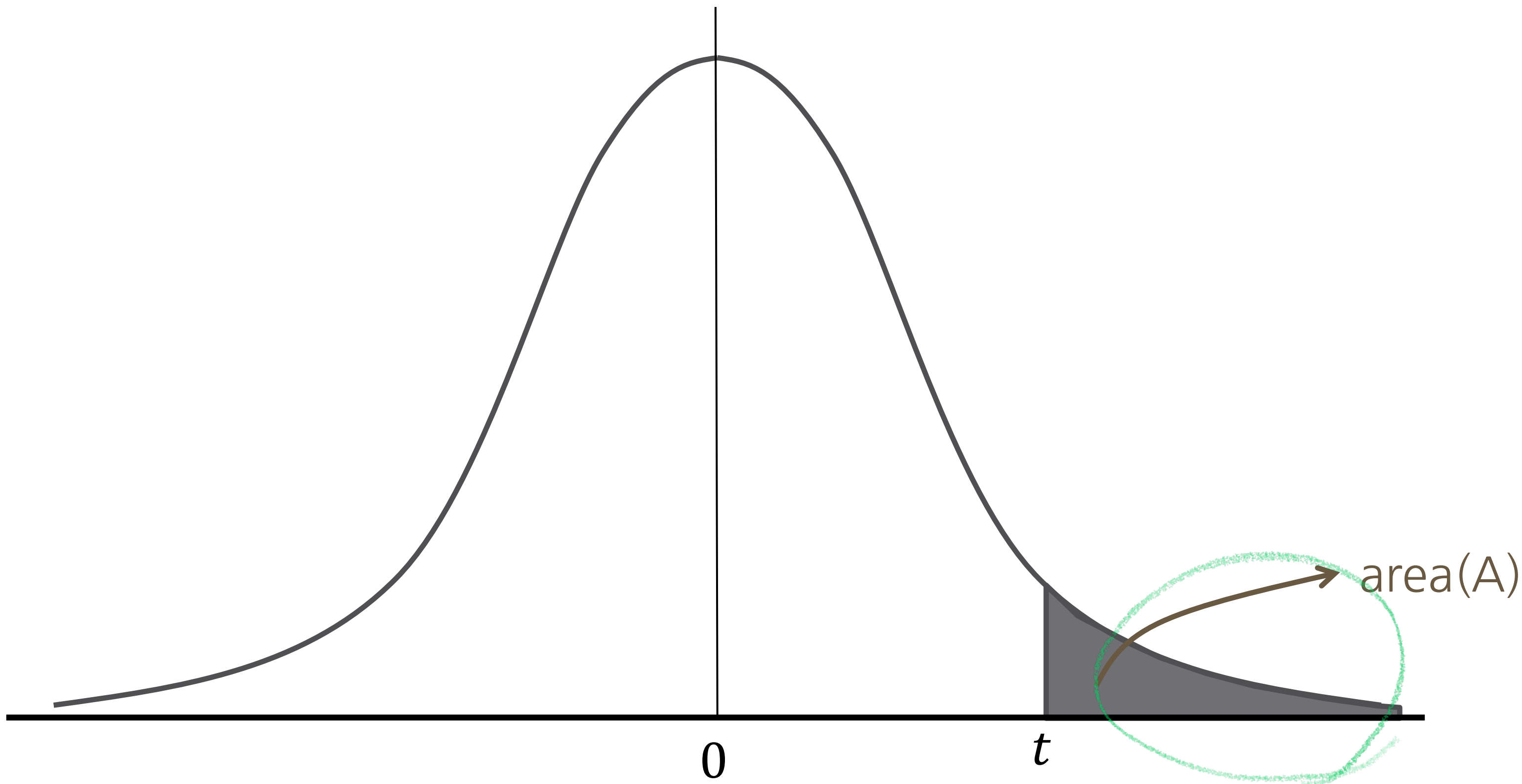
= 정규분포와 가까움.

sample ↓ 오차 ↑ 가능성

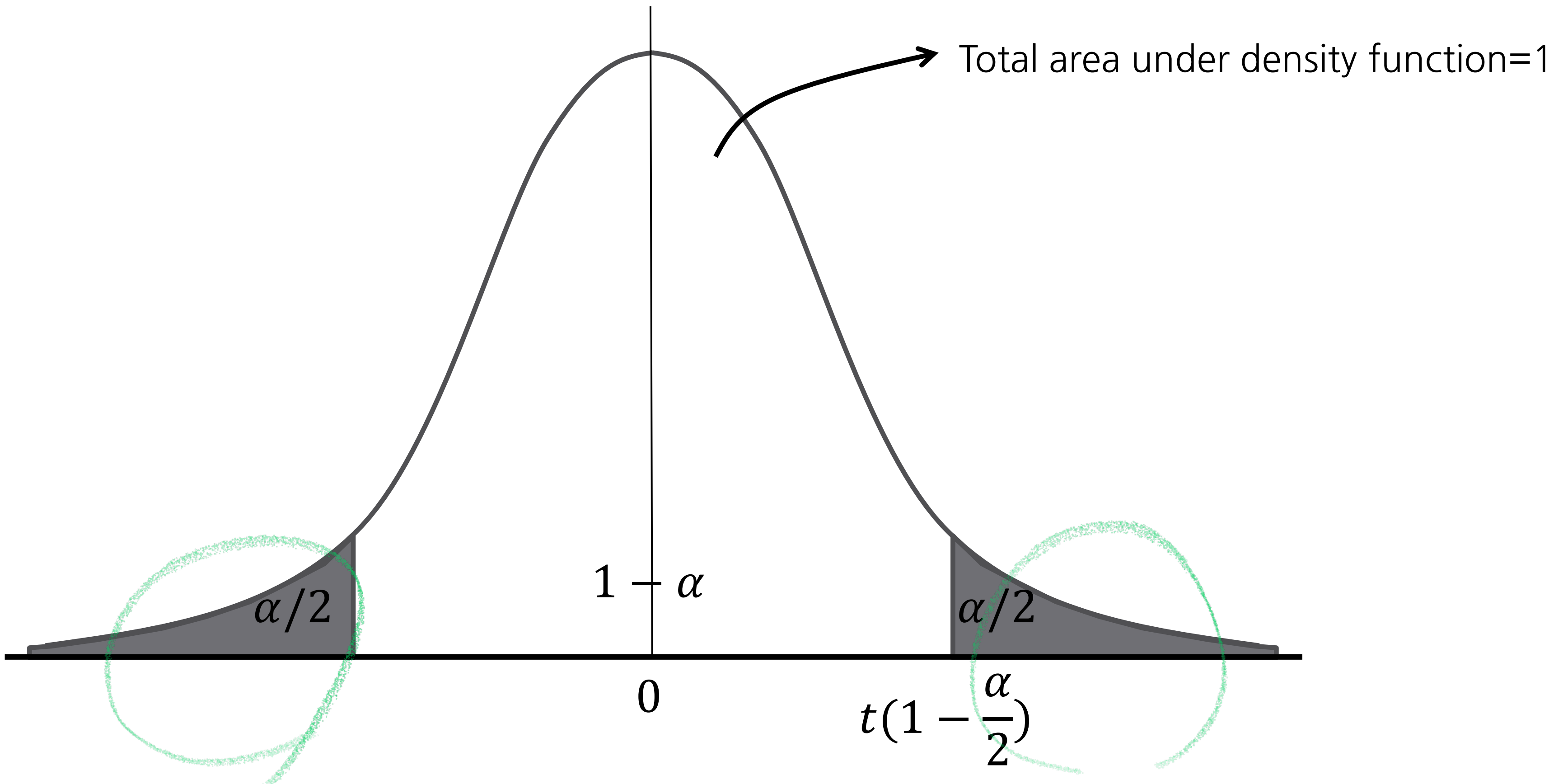
〃 ↑ 정확.

Test Concerning Regression Coefficients

- If (area under density function from $|t|$ to ∞) $< \frac{\alpha}{2}$
  $\rightarrow$ Reject null hypothesis $\rightarrow$ β_i is not zero
  - α is significance value
  - significance level is usually set to 0.1, 0.05
    - The higher significance level, the higher probability to reject null hypothesis



Test Concerning Regression Coefficients



Test Concerning Regression Coefficients

- How to calculate area?
  - Don't worry. There is pre-calculated table!

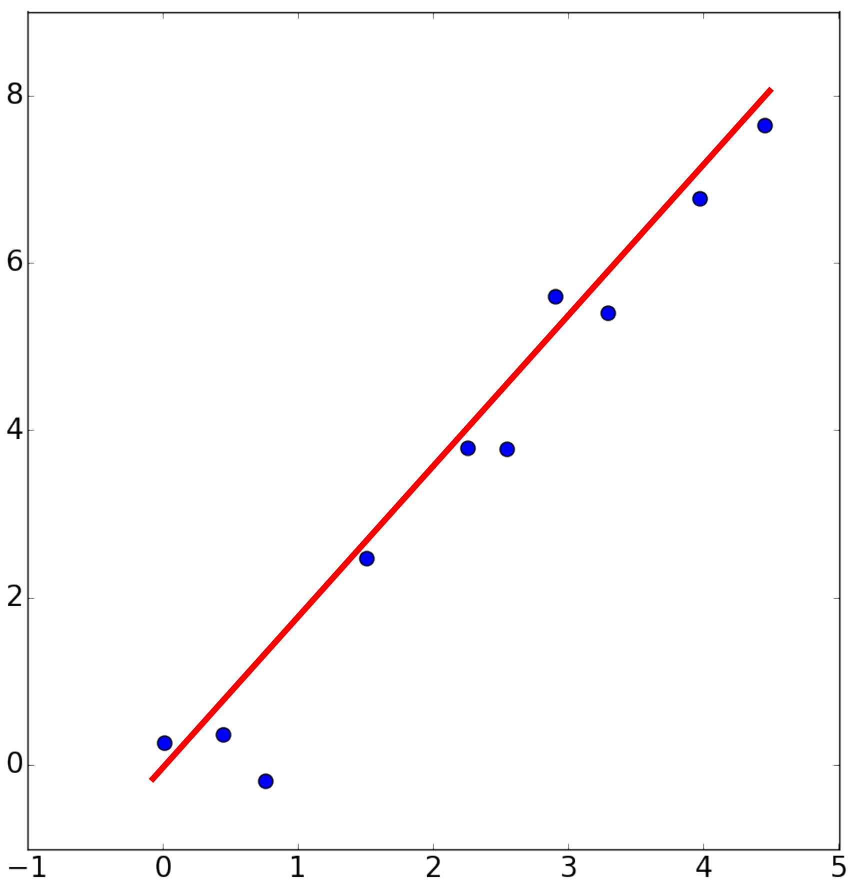
Student t-Table

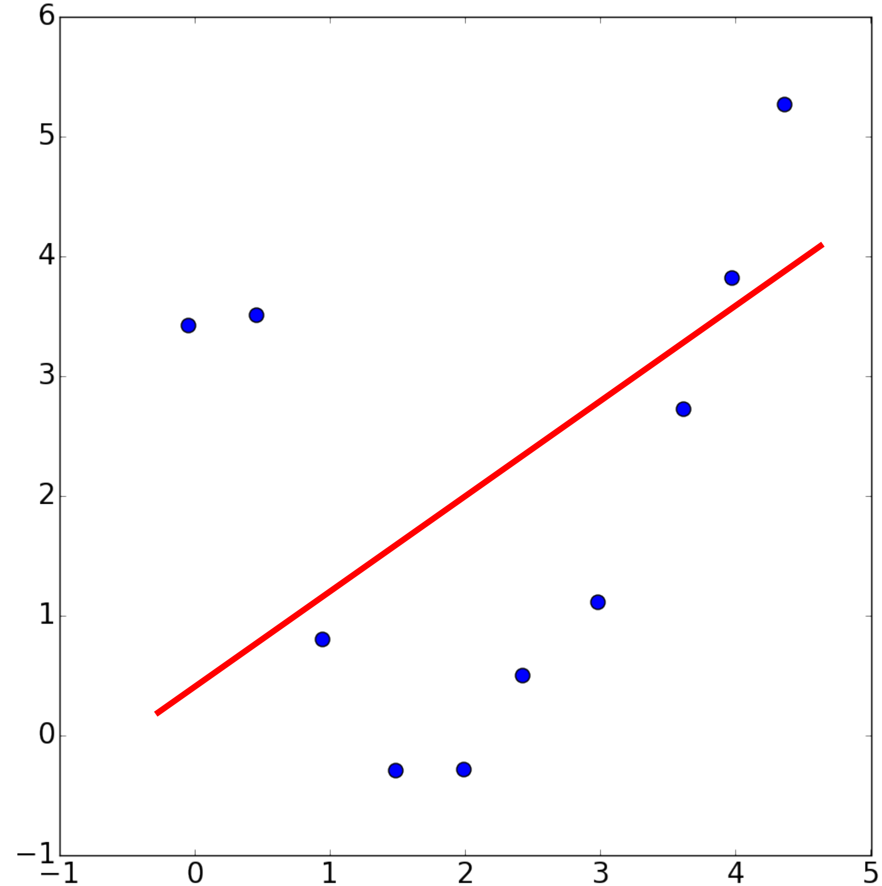
Alpha	0.250	0.200	0.150	0.100	0.050	0.025	0.010	0.005	0.0005
df									
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.656	636.578
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.600
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	$\boxed{0.687}$	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850

t value that area is 0.25 with 20 degree of freedom

Goodness-of-fit $\rightarrow R^2$ 모델이 유의미한지 test

- How to measure quantitatively performance of fitted models?
  - Calculate goodness-of-fit





R^2

- Statistical measures for goodness-of-fit
  - R^2 $(0 \le R^2 \le 1)$

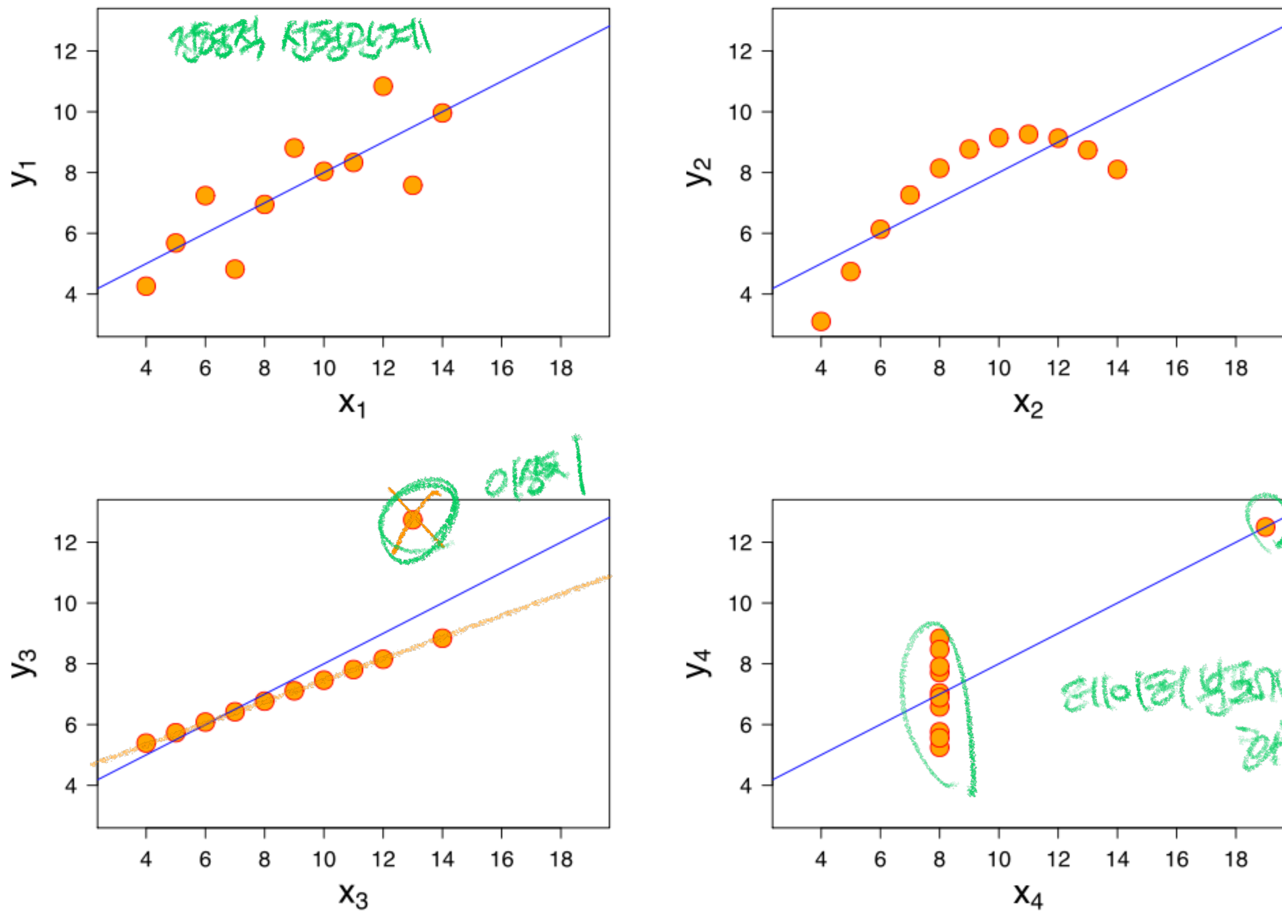
$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

내 모델이 설명을 많이 하고 있는거니까

$R^2 \uparrow$ 좋음

R^2 is NOT All-around Player → adjusted R^2

- Anscombe's quartet
  - The same linear regression line but are themselves very different.



곡선형태인데 억지로 직선. 잘못 적용

<https://en.wikipedia.org/wiki/Linear_regression>

Adjusted R^2

- Adding more input variables to the regression model increases R^2 and never reduce it
  - Tend to add more input variables to the model

Is always right to add more variables?

- Adjusted R^2

$$R^2{}_{adj} = 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}} = 1 - \left(\frac{n-1}{n-p-1}\right)(1-R^2)$$

Depend on the number of input variables

  - Penalty on the number of input variable by $n-p-1$
  - Adjusted R^2 may actually become smaller when another input variable is introduced into the model

Performance metrics

F^*, t검정, R^2, AR^2 외의 다른 방법들

- Functions to measure regression performance
  - Mean squared error

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

  - Mean absolute error

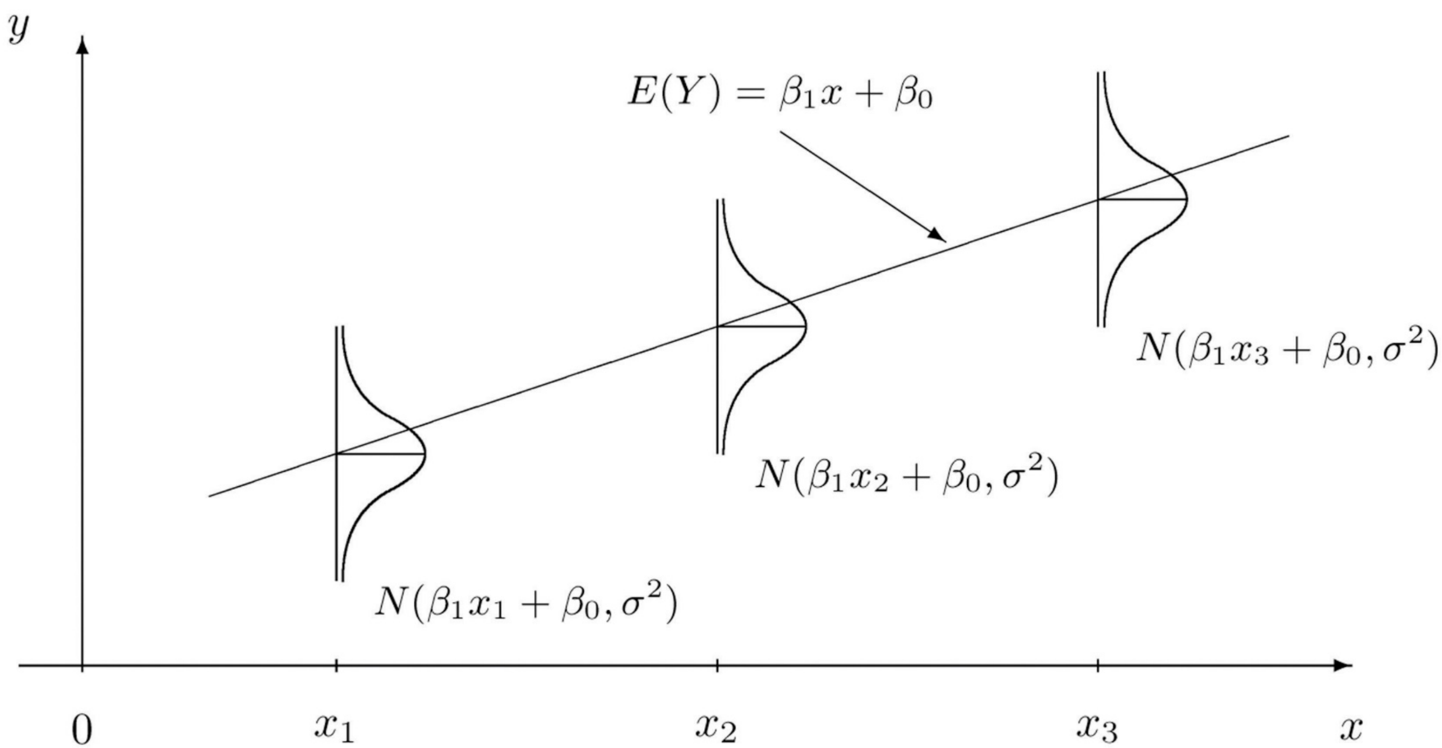
$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n}$$

  - Median absolute error
    - robust to outliers

$$MedAE = \text{median}(|y_1 - \hat{y}_1|, \ldots, |y_n - \hat{y}_n|)$$

Check Appropriateness of Linear Regression

- Do you remember main assumptions of linear regression?



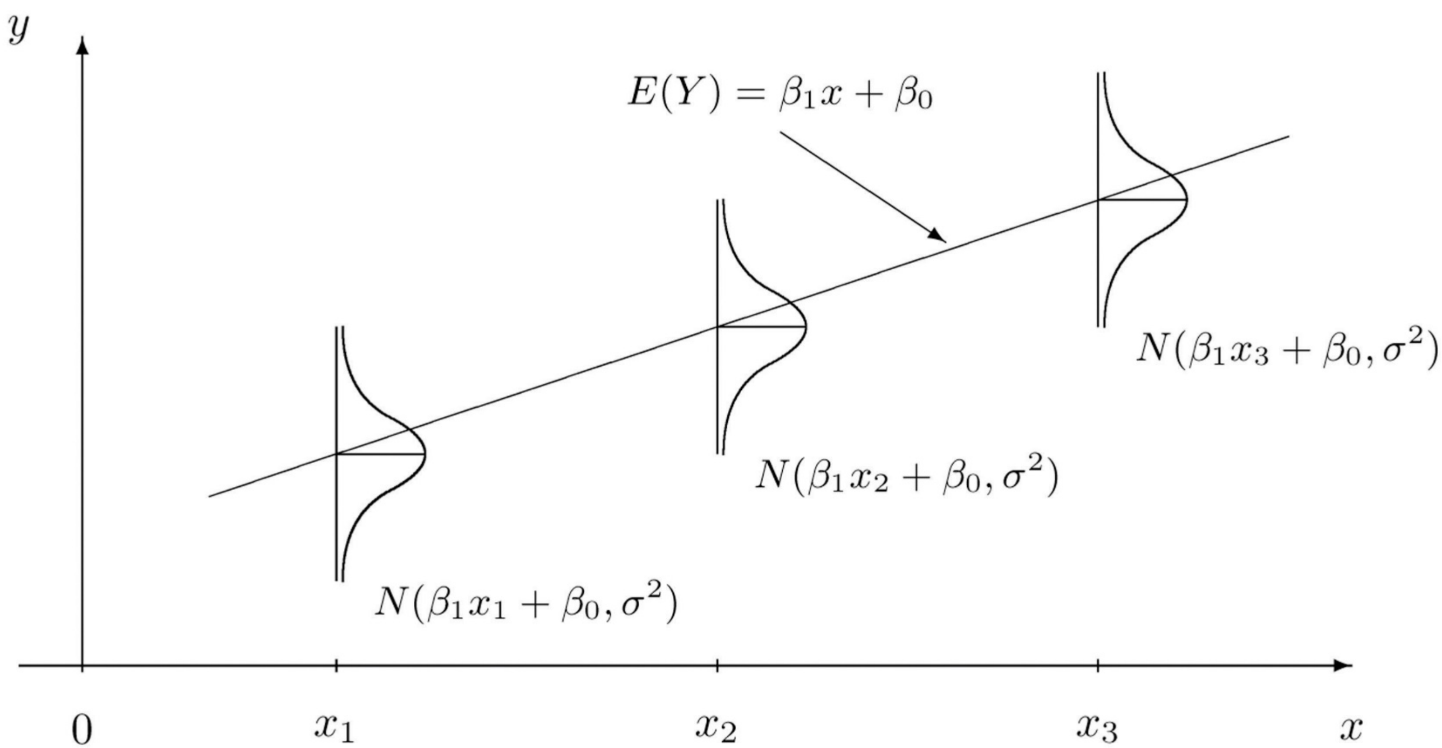
Main Assumption of Linear Regression

- Linear regression analysis makes several key assumptions
  - Linear relationship
  - Homoscedasticity
  - Normality
  - No or little multicollinearity

에러에 대한 가정.

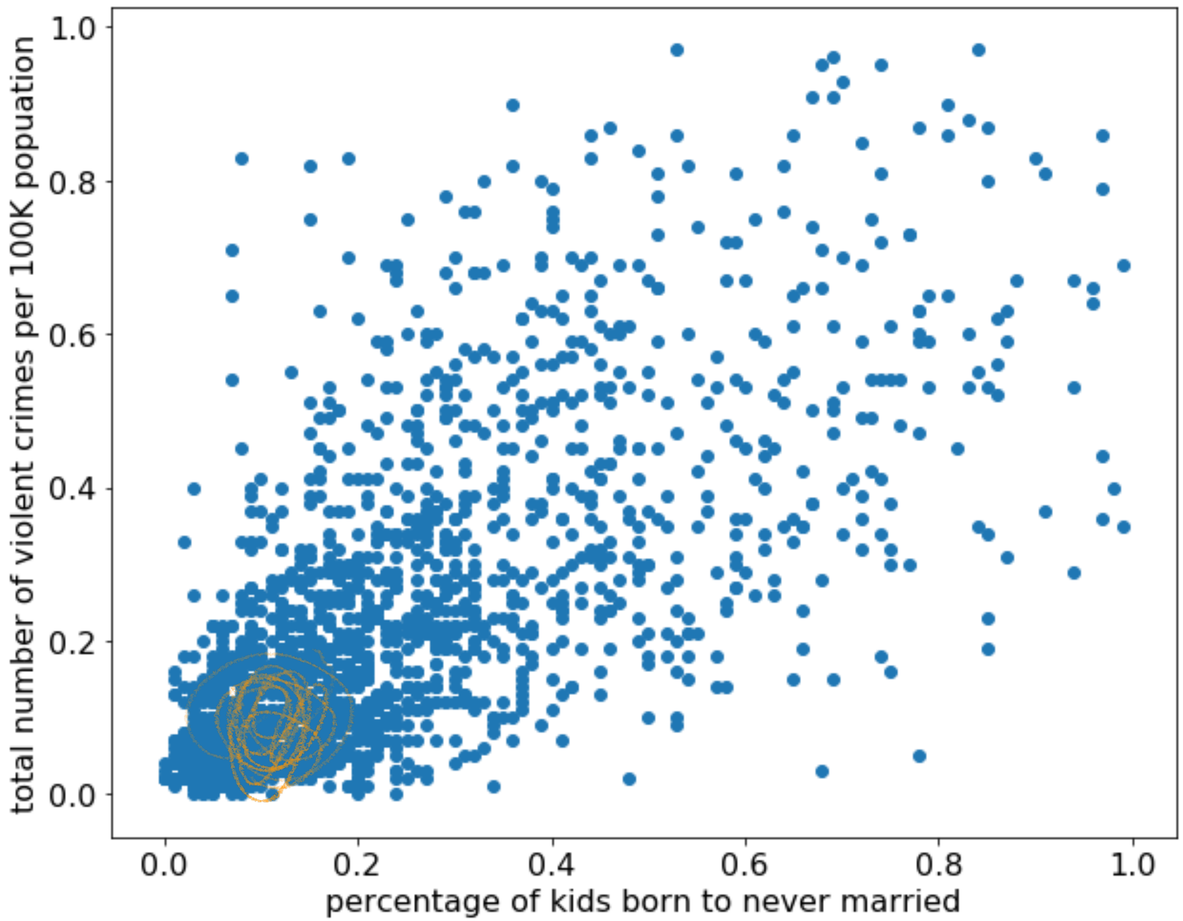
에러 정규분포를 따름

제일 중요한 가정



Check Appropriateness of Linear Regression

- Linear relationship
  - Check relationships between input variables and a responsive variable



Relationships between Input Variables

- If some of input variables are highly correlated, regression coefficients are unstable x에 대한 변동을 받는 것.

	1	2	3	4	5	6	7	8	9	10
x_1	98	120	140	195	181	128	107	106	88	77
x_2	24	35	36	51	45	30	29	24	22	19
x_3	21	11	31	42	57	82	67	13	55	36

- Correlation matrix

$$
\begin{array}{cc}
 & \begin{array}{ccc} x_1 & 2 & 3 \end{array} \\
corr = \begin{array}{c} 1 \\ 2 \\ 3 \end{array} & \begin{bmatrix} 1.00 & 0.98 & 0.17 \\ 0.98 & 1.00 & 0.11 \\ 0.17 & 0.11 & 1.00 \end{bmatrix}
\end{array}
$$

- x_1 and x_2 are highly correlated

※ Covariance

- Variance of a random variable X is the expected value of the squared deviation from the mean($\mu = \mathbb{E}[X]$)

$$Var(X) = \mathbb{E}[(X - \mu)^2]$$

  - Sample variance is calculated by

$$s_x^2 = \frac{\sum_{i=1}^{n}(x - \bar{x})^2}{(n-1)}$$

- Covariance is a measure of how much two random variables change together

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

  - Variance is the covariance of a random variable with itself

$$Var(X) = Cov(X, X)$$

  - Sample covariance is calculated by

$$q_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_j - \bar{y})}{n-1}$$

※ Correlation

- Any statistical relationship, whether causal or not, between two random variables or bivariate data
  - Pearson's correlation coefficient
    - The most popular correlation

$$\rho_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

    - Sample correlation coefficient

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

Relationships between Input Variables

- Two difference cases

	1	2	3	4	5	6	7	8	9	10
y_1	295	310	404	567	574	532	442	283	366	285
y_2	282	311	402	581	573	523	446	277	374	274

  - Output values of two cases are quite similar
  - Regression coefficient for y_1 and y_2

$$\text{Case 1: } [\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3] = [2.16 \quad 0.14 \quad 2.88]$$

$$\text{Case 2: } [\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3] = [1.73 \quad 2.18 \quad 2.97]$$

  - Because x_1 and x_2 are highly correlated, explained variance by x_2 is also explained by x_1 $\rightarrow$ Coefficient of x_2 is quite unstable

Why This Situation Happens

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y}$$

- To estimate regression coefficients, inverse matrix of $\mathbf{X}^{\mathbf{T}}\mathbf{X}$ should be calculated

- Ill-conditioned matrices
  - If a small change in the coefficient matrix results in a large change in the solution, the coefficient matrix is called ill-conditioned

$$\begin{cases} x + y = 2 \\ x + 1.001y = 2 \end{cases} \quad \text{and} \quad \begin{cases} x + y = 2 \\ x + 1.001y = 2.001 \end{cases}$$

    - Left: $x = 2, y = 0$
    - Right: $x = 1, y = 1$

$$\begin{bmatrix} 1 & 1 \\ 1 & 1.001 \end{bmatrix} \text{ is ill-conditioned}$$

Variance Inflation Factor

- Variance inflation factor(VIF) quantifies the severity of multicollinearity in a least square method

[Multicollinearity] 가 있나 없나를 따지기 위한

A Phenomenon in which two or more input variables in a multiple regression model are highly correlated
→ In this situation the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data

- Variance of estimated coefficients for $j-\mathrm{th}$ input variable

$$var(\hat{\beta}_j) = se^2(\hat{\beta}_j) = [MSE(\mathbf{X}^T\mathbf{X})^{-1}]_{j,j} = \frac{MSE}{(n-1)se^2(x_j)}\frac{1}{1-R_j^2}$$

  - R_j^2 is the R^2 for the regression of the x_j on the other input variables

나머지
x_j | x들

관계를 따지는것

R^2 y | x들

- VIF 는 feature 별로 하나씩.

$$VIF = \frac{1}{1-R_j^2}$$

$$\frac{1}{1-\frac{SSR}{SST}}$$

Variance Inflation Factor

- Calculate VIF
  - Step 1) Apply least square method to regression problem that i-th input variable is regressed by the remained input variables

$$x_i = \alpha_1 x_1 + \cdots + \alpha_{i-1} x_{i-1} + \alpha_{i+1} x_{i+1} + \cdots + \alpha_p x_p + \alpha_0 + \epsilon$$

  - Step 2) Calculate R^2 for above regression problem and set the value as R_i^2
  - Step 3) Calculate VIF from R_i^2

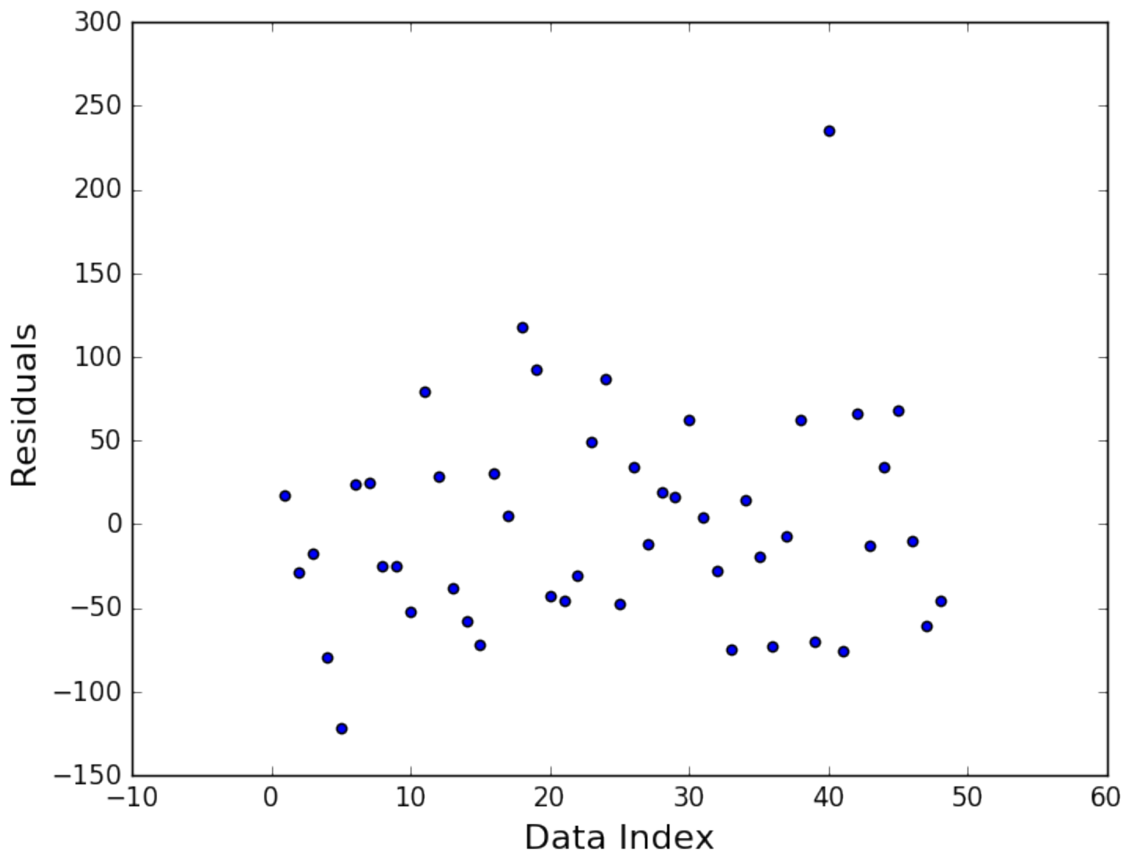
$$VIF = \frac{1}{1 - R_i^2}$$

- A rule of thumb is that if VIF($\widehat{\beta_i}$)>10 then multicollinearity is high
  - In this case, do not use x_i as explanatory variable to estimate output

Check Appropriateness of Linear Regression

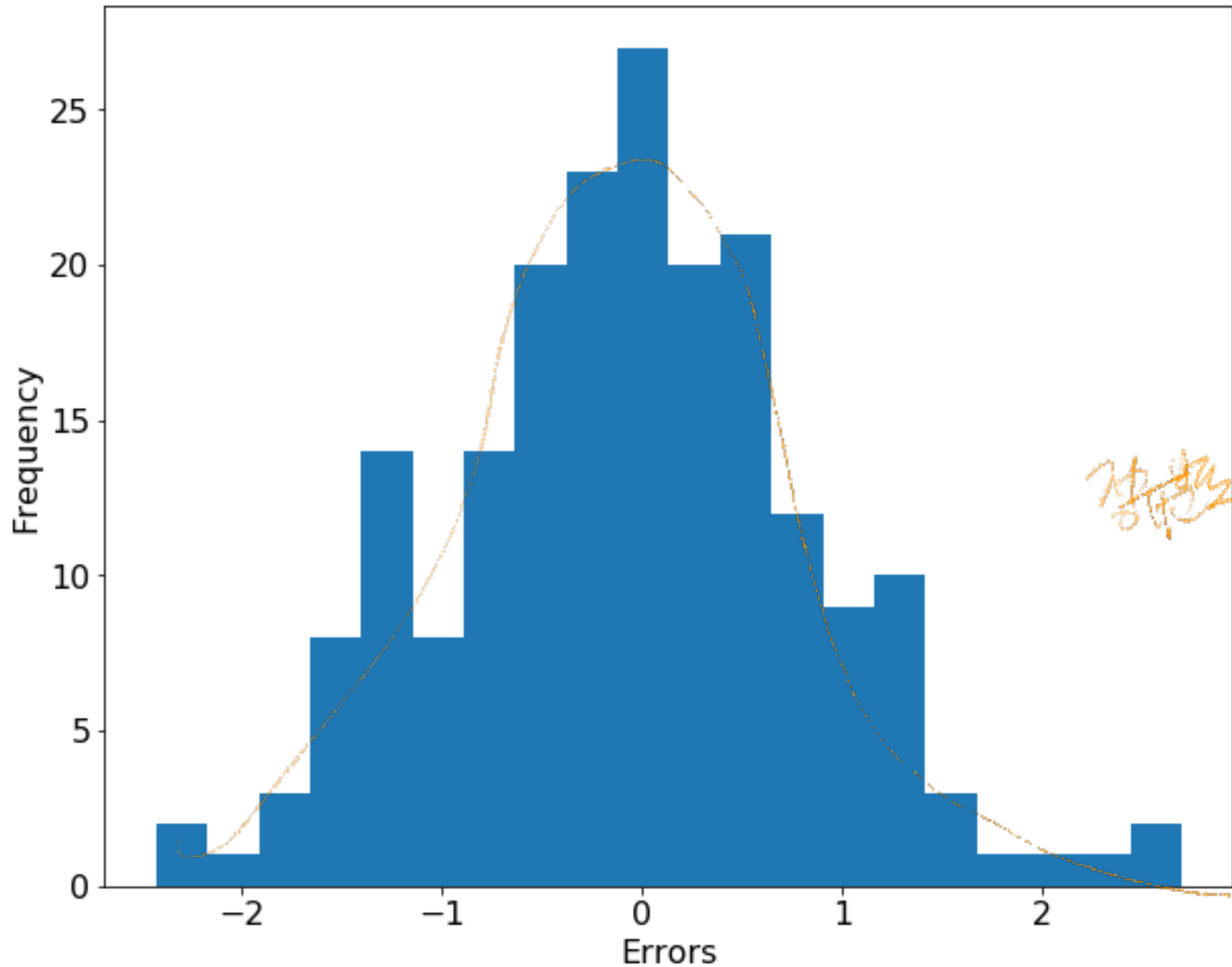
- Normality
  - Errors should follow normal distribution
  - Calculate errors (residuals) and check normality

$$e_i = y_i - \hat{y}_i$$



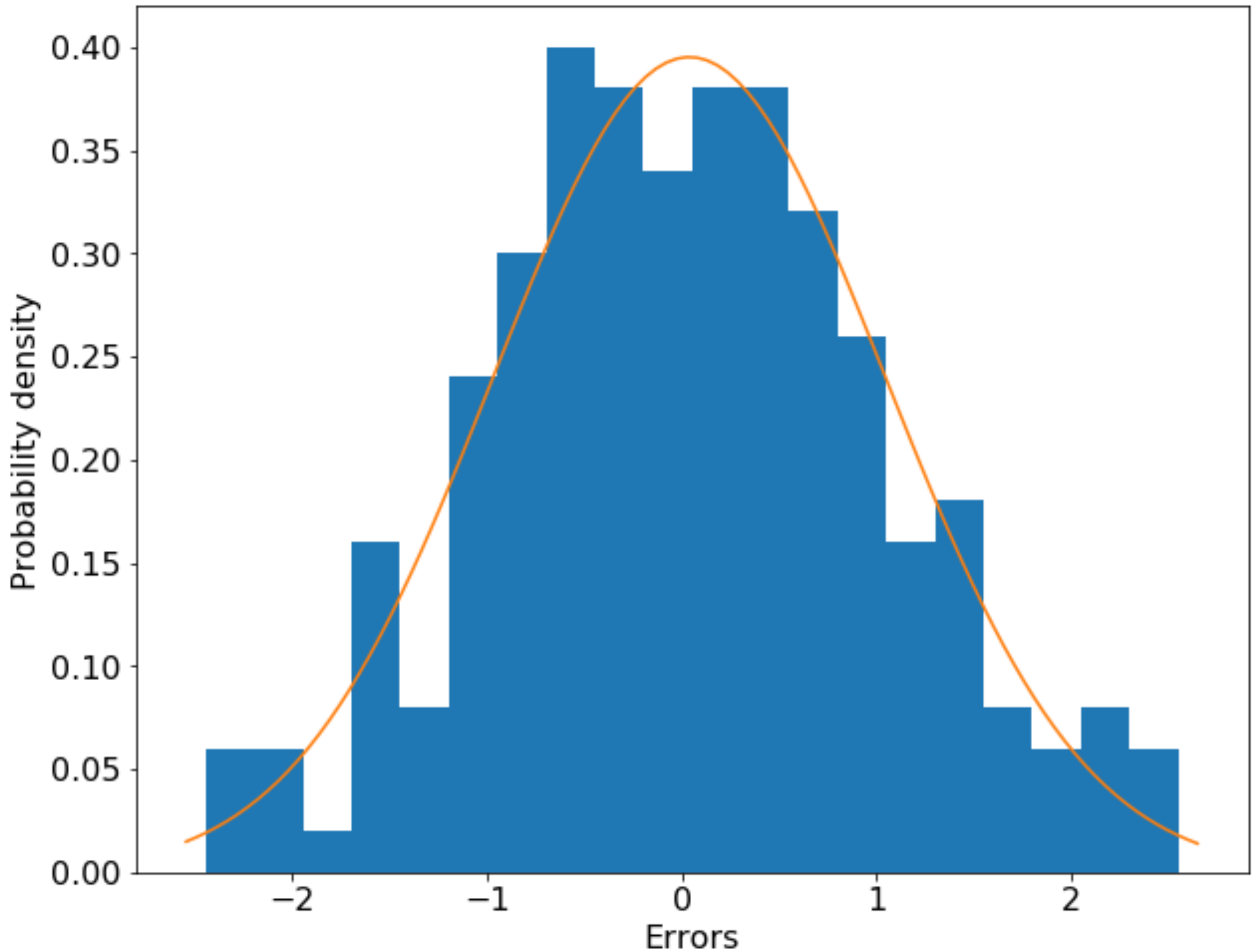
Check Appropriateness of Linear Regression

- Histogram



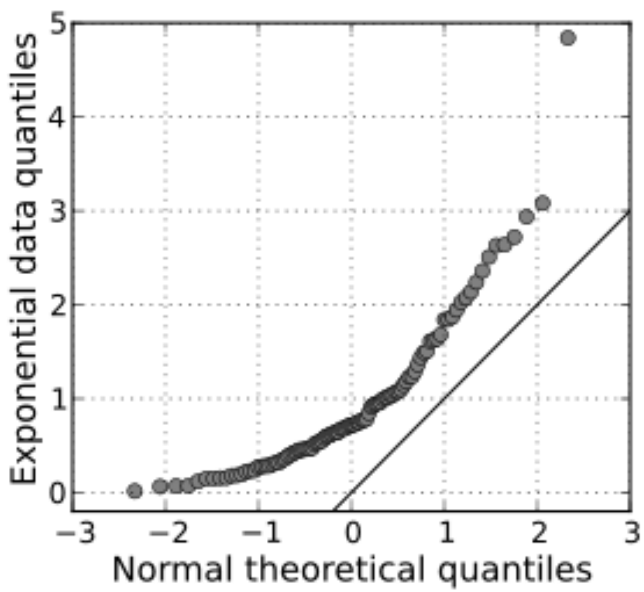
Check Appropriateness of Linear Regression

- Histogram

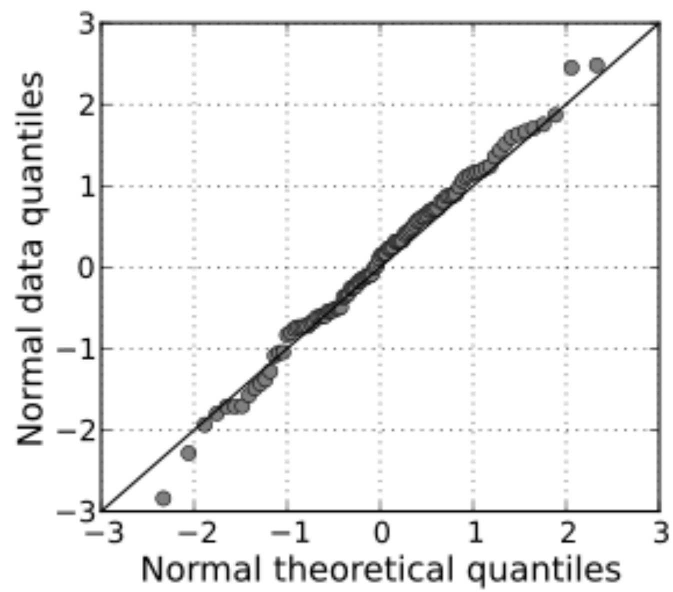


Check Appropriateness of Linear Regression

- Q-Q plot
  - A probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other
  - Quantiles are cutpoints dividing a set of observations into equal sized groups
    - q-Quantiles are values that partition a finite set of values into q subsets of (nearly) equal sizes
    - Median is 2-quartile, 0.5 quantile and 50 percentile

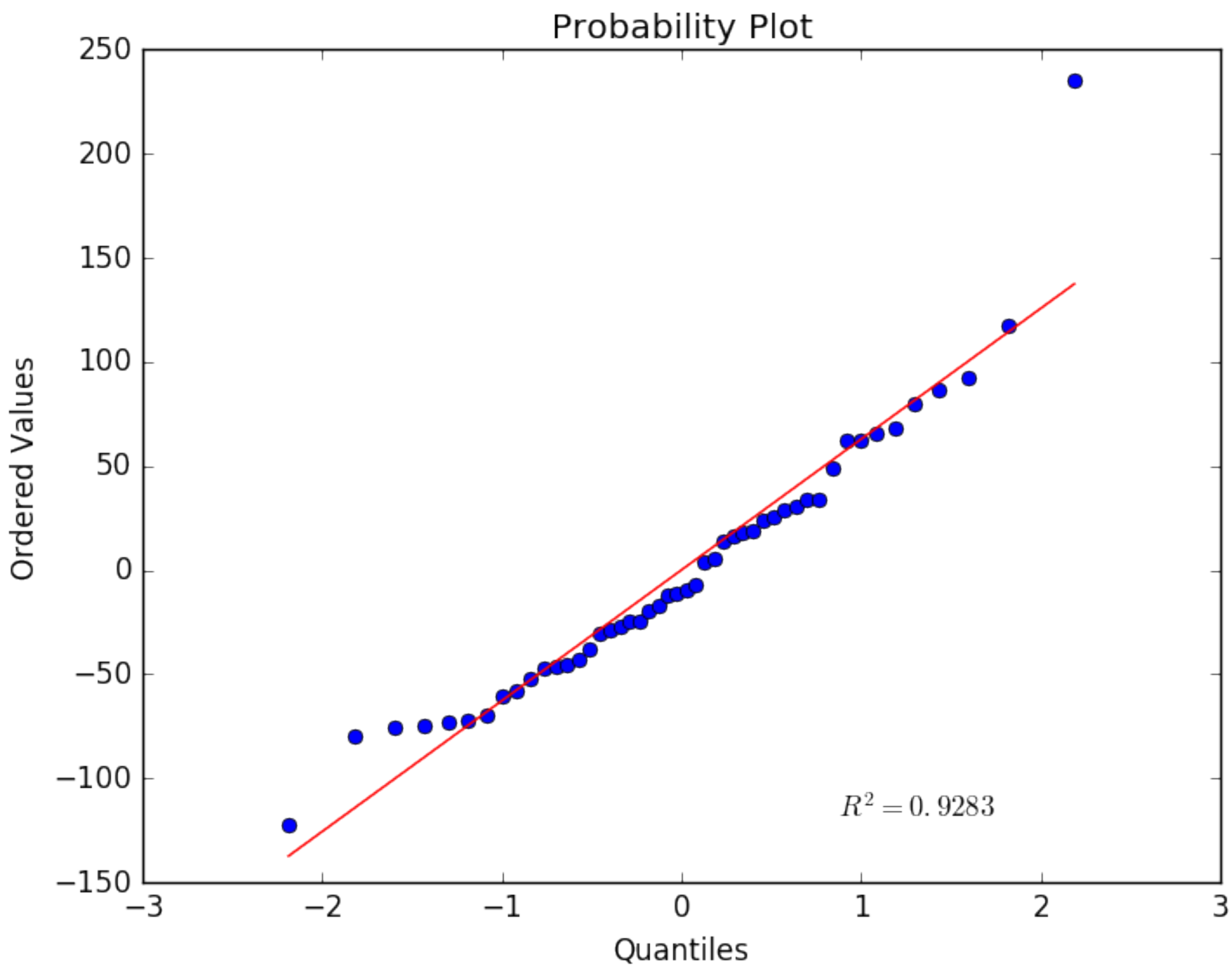


$X \sim Exp(1)$



$X \sim N(0,1)$

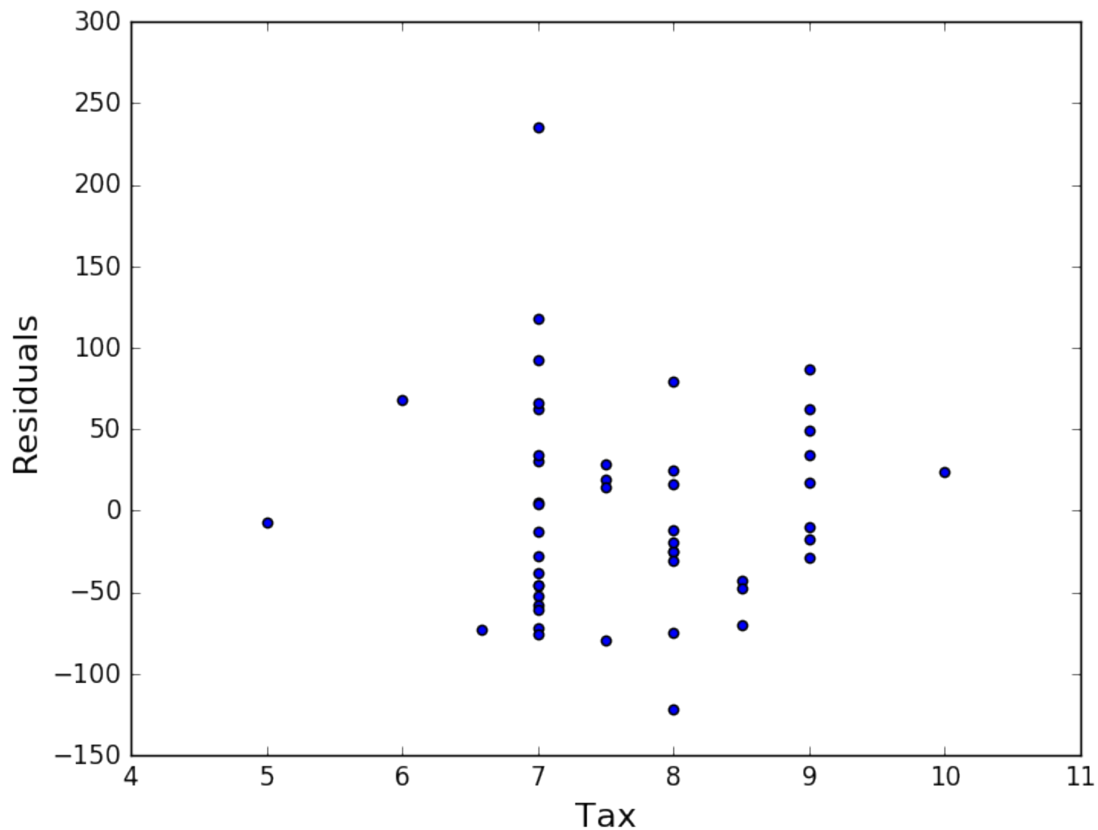
Q-Q Plot



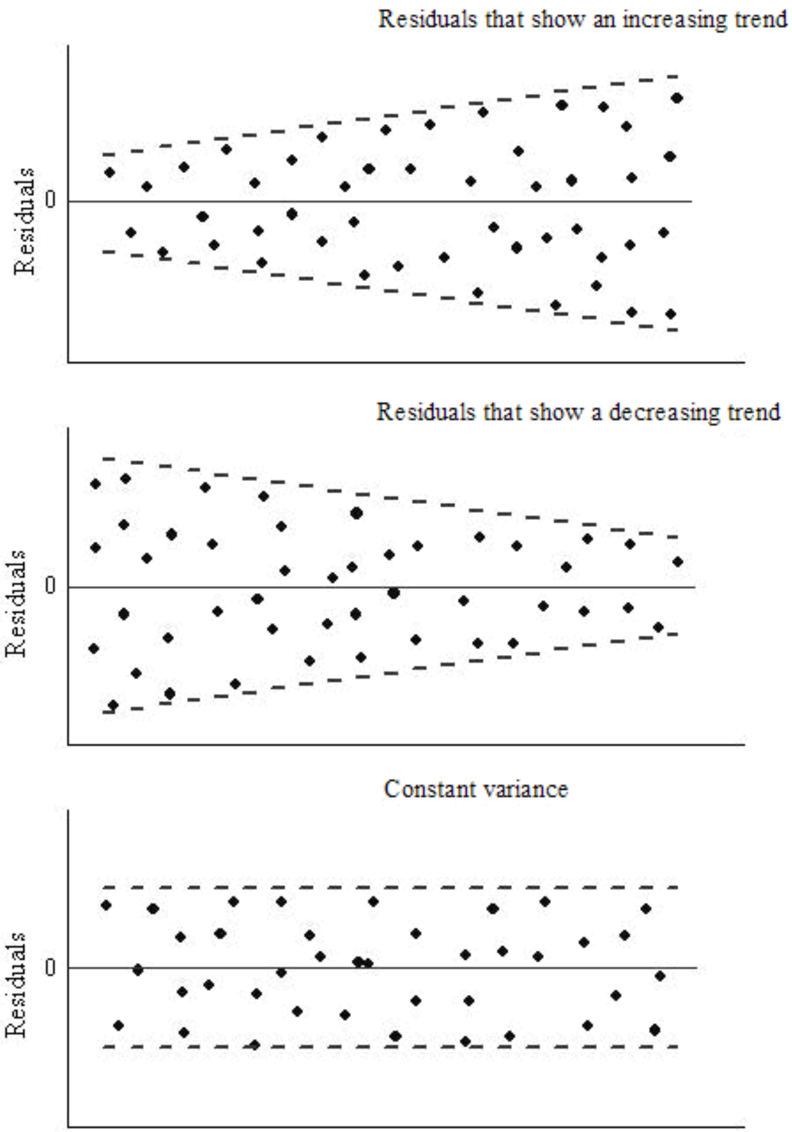
Probability Plot
Ordered Values
Quantiles
250
200
150
100
50
0
−50
−100
−150
−3
−2
−1
0
1
2
3
$R^2 = 0.9283$

Check Appropriateness of Linear Regression

- Homoscedasticity↔Heteoscedasticity
  - Check whether all random variables in the sequence or vector have the same finite variance

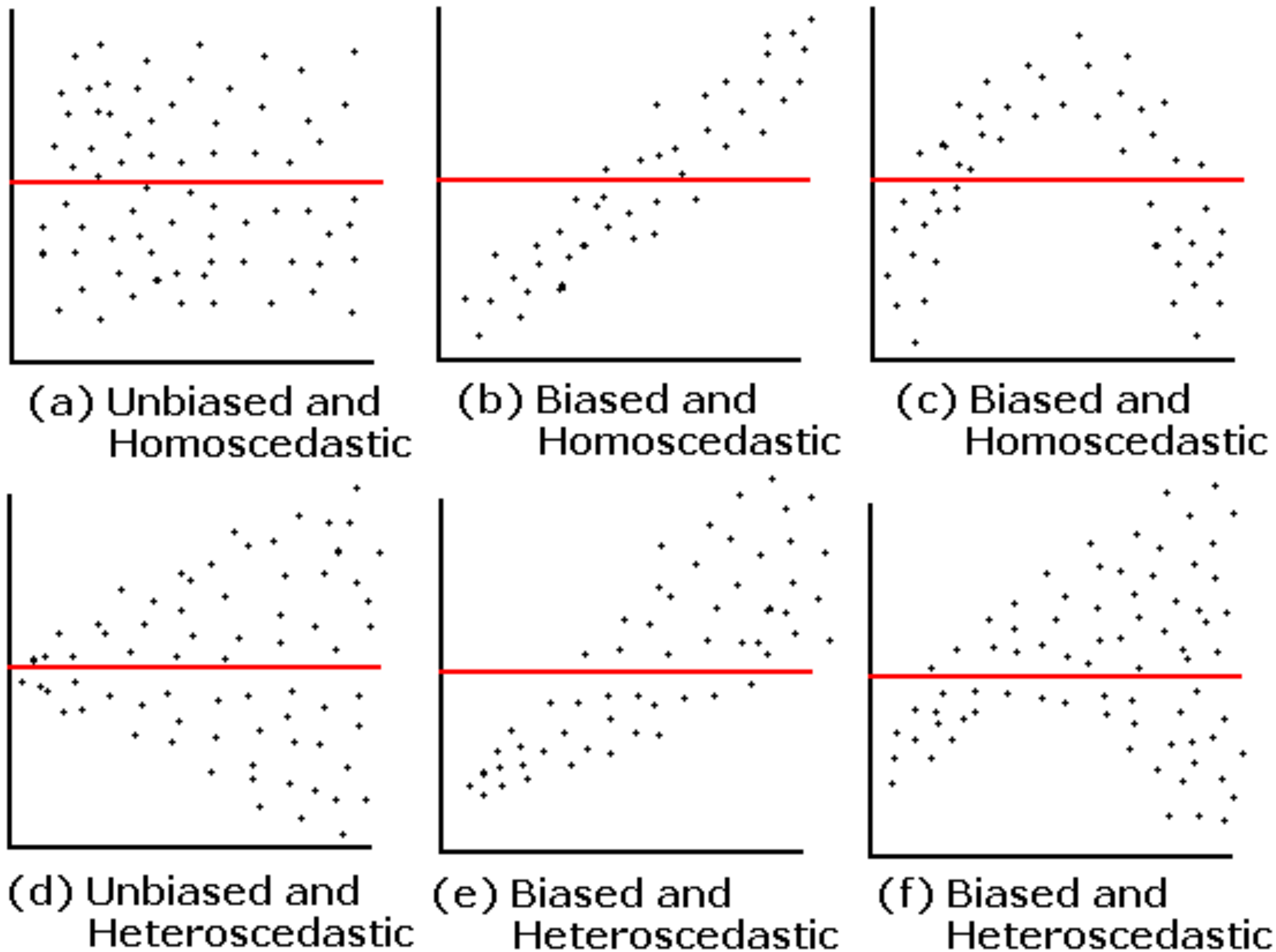


Check Appropriateness of Linear Regression



Check Appropriateness of Linear Regression

- Residual plot



Interpretation & Prediction

- If the fitted regression model is appropriate and significant you can use the model for future use
  - Linear regression models have strength in interpretation
    - Each coefficient explain relationship between each explanatory variable and the target variable
  - Based on the fitted model, predict the target on test samples

Overall Process for Linear Regression

Validity of Linear Regression Model

Before learning

Linearity EDA

Multicolinearity → 변수선택 or feature 제거

X끼리 확인

After learning

Normality

Homoscedasticity

[illegible]

Significance of model

Goodness-of-fit

[illegible]

얼마나 좋은지

Model Utilization

- Figure out significant input variables
- Figure out relations between input variables and target variable → Interpretation
- Predict target with respect to new observations

Feature Scaling

- Predict consumption of petrol
  - Linear model by least square method

$$y = -34.8x_1 - 0.0666x_2 - 0.002x_3 + 1336x_4 + 377.3$$

→ 무의미하지X
상관없음

Petrol Tax($)	Average Income ($)	Paved Highways (miles)	Proportion of population with driver's license	Consumption of petrol (M of gallons)
9	3571	1976	0.525	541
9	4092	1250	0.572	524
9	3865	1586	0.58	561
7.5	4870	2351	0.529	414
…	…	…	…	…

↔ scale이 다름

How about changing scale of variable?

Feature Scaling

- Change unit of paved highways from mile to cm

  1mile=160934.4cm

Petrol Tax($)	Average Income ($)	Paved Highways (cm)	Proportion of population with driver's license	Consumption of petrol (M of gallons)
9	3571	31683974.4	0.525	541
9	4092	20043000	0.572	524
9	3865	25430558.4	0.58	561
7.5	4870	37696874.4	0.529	414
…	…	…	…	…

- Linear regression on new data

$$y = -34.8x_1 - 0.0666x_2 - 1.5 \times 10^{-7}x_3 + 1336x_4 + 377.3$$

영향 X

Feature Scaling

- Scale change only affects on the changed variable
  - Coefficients of other variables are not changed
  - If variable x is replaced with ax, coefficient of x, β by linear regression is changed to β/a
  - If scale of certain variable is too large, coefficient of the variable might be too small
    → It is better to change scale

Variable Transformation

- Linear regression algorithm is quite simple, but it can be extended using transformation
  - $x \rightarrow x^2$
  - $x \rightarrow \log x$
  - $x \rightarrow \sqrt{x}$