LINEAR REGRESSION

Week04

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

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

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

  - Decision rule

$$\text{If } |t_j| \leq t\left(1 - \frac{\alpha}{2}; n - p - 1\right), \text{conclude } H_0$$
$$\text{If } |t_j| > t\left(1 - \frac{\alpha}{2}; n - p - 1\right), \text{conclude } H_1$$

Test Concerning Regression Coefficients

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

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

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

※ Variance and Covariance

- Variance
  - Variance is the expectation of the squared deviation of a random variable from its mean

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

    - sample variance

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

    - $\bar{x}$: sample mean
- Covariance
  - Covariance is a measure of the joint variability of two random variables

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

  - sample covariance

$$q_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

※ Covariance Matrix

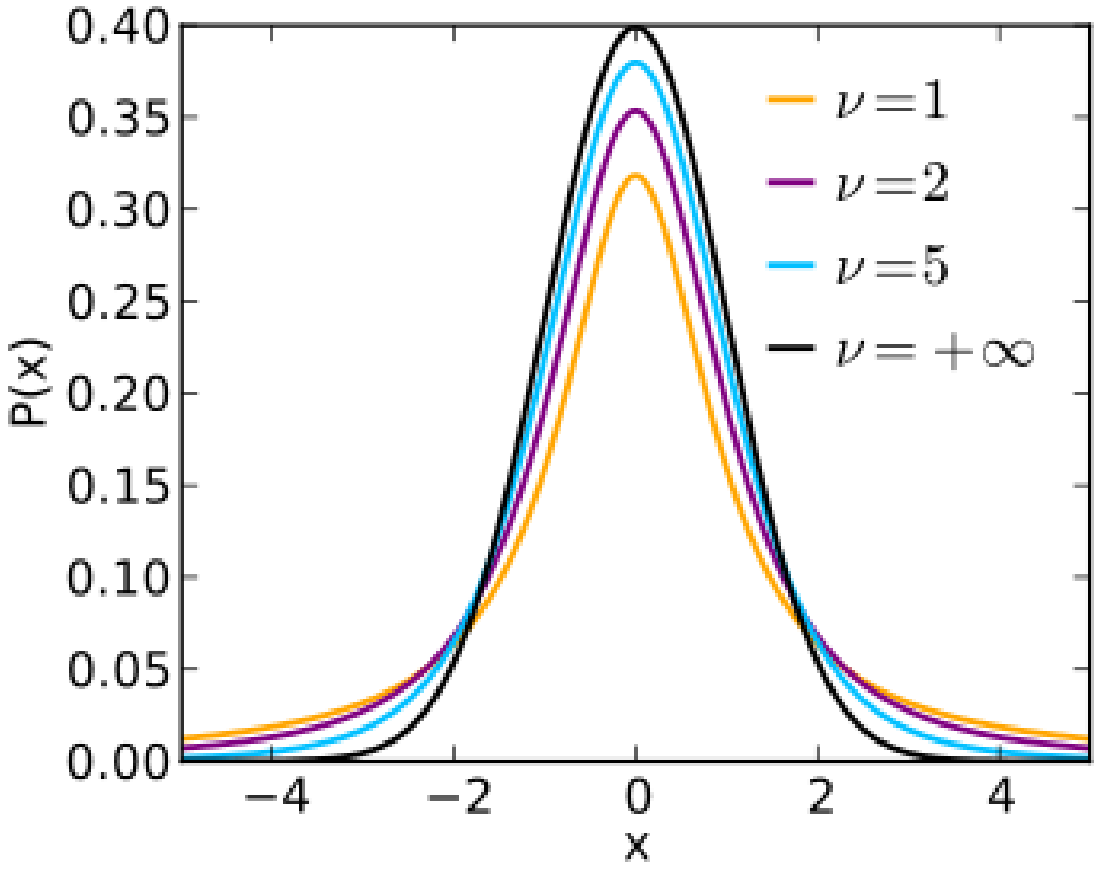
- Covariance matrix
  - A covariance matrix is a square matrix giving the covariance between each pair of elements of a given random vector

$$\begin{bmatrix} cov[X_1, X_1] & cov[X_1, X_2] & \cdots & cov[X_1, X_p] \\ cov[X_2, X_1] & cov[X_2, X_2] & \cdots & cov[X_2, X_p] \\ \vdots & \vdots & & \vdots \\ cov[X_p, X_1] & cov[X_p, X_2] & \cdots & cov[X_p, X_p] \end{bmatrix}$$

$$cov[X_i, X_i] = var[X_i]$$

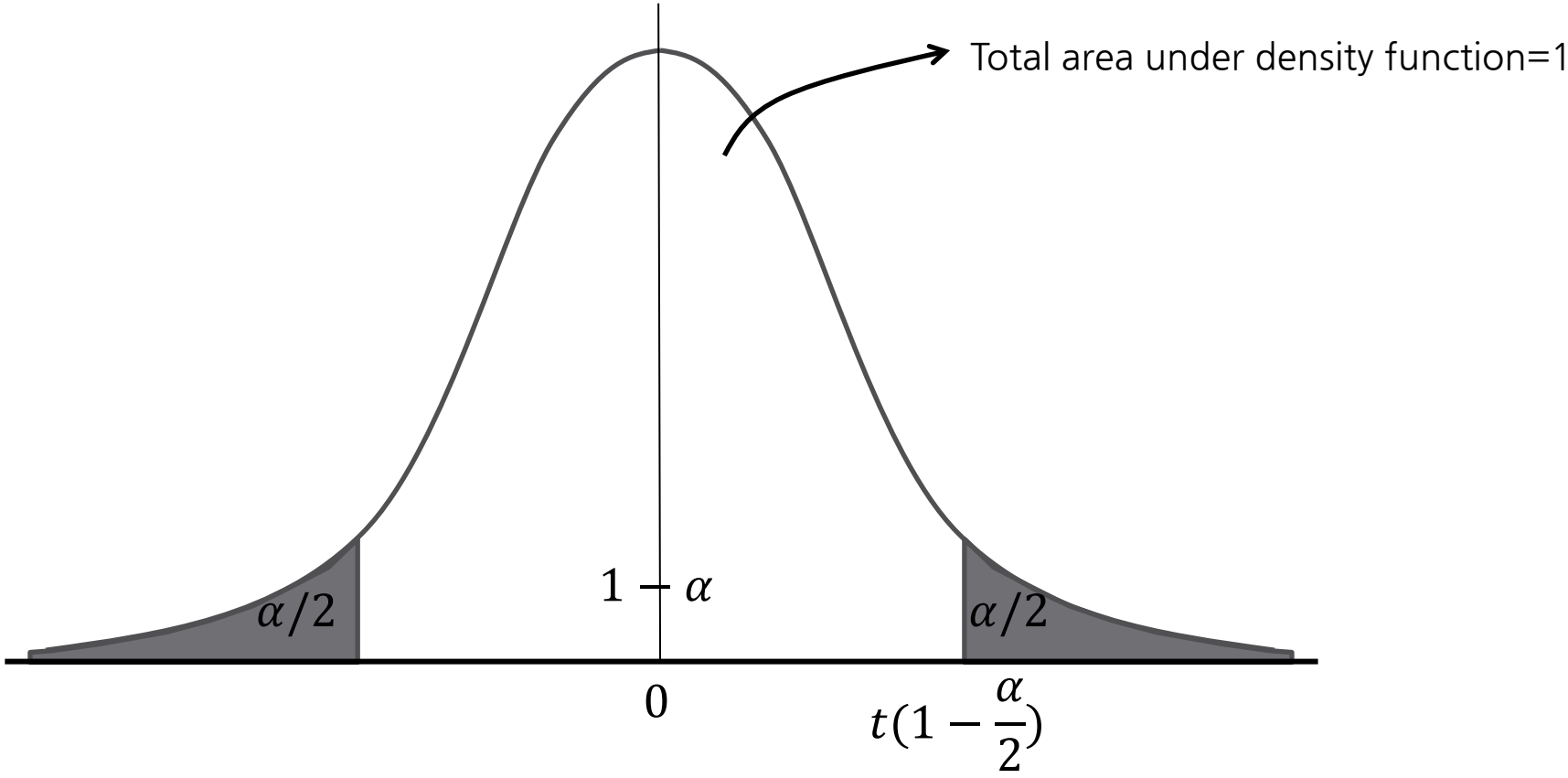
Test Concerning Regression Coefficients

- 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)



Test Concerning Regression Coefficients

- If (area under density function from $|t|$ to ∞) $< \frac{\alpha}{2}$
  $\rightarrow$ Reject null hypothesis $\rightarrow \beta_i$ 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

- 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	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

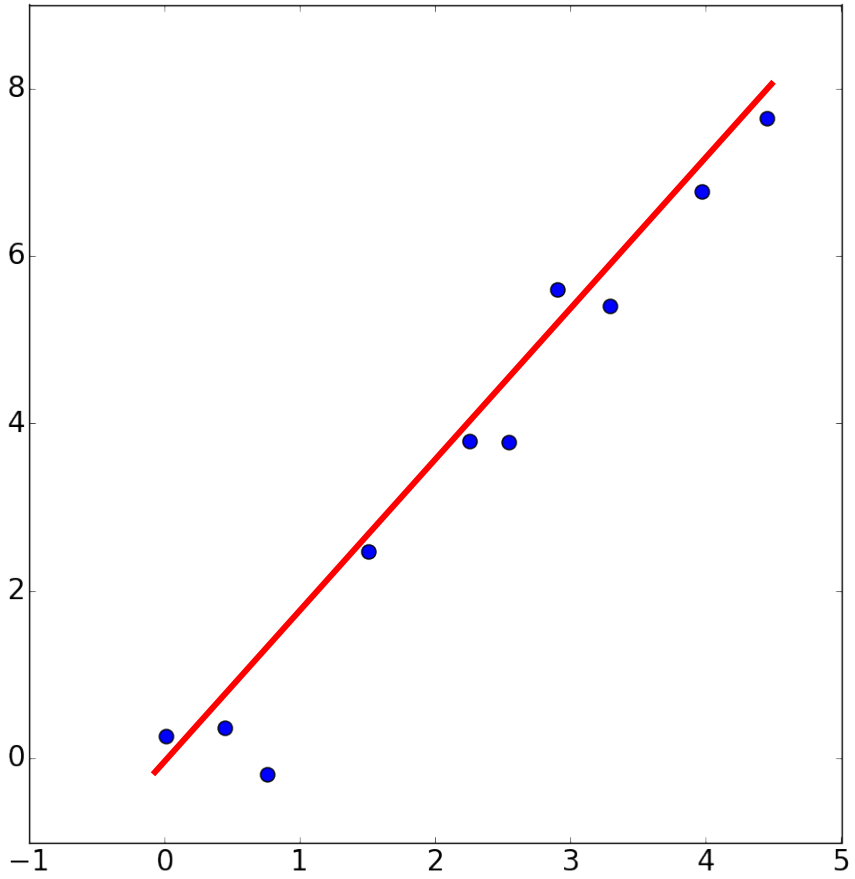
Table for Distributions: F-distribution

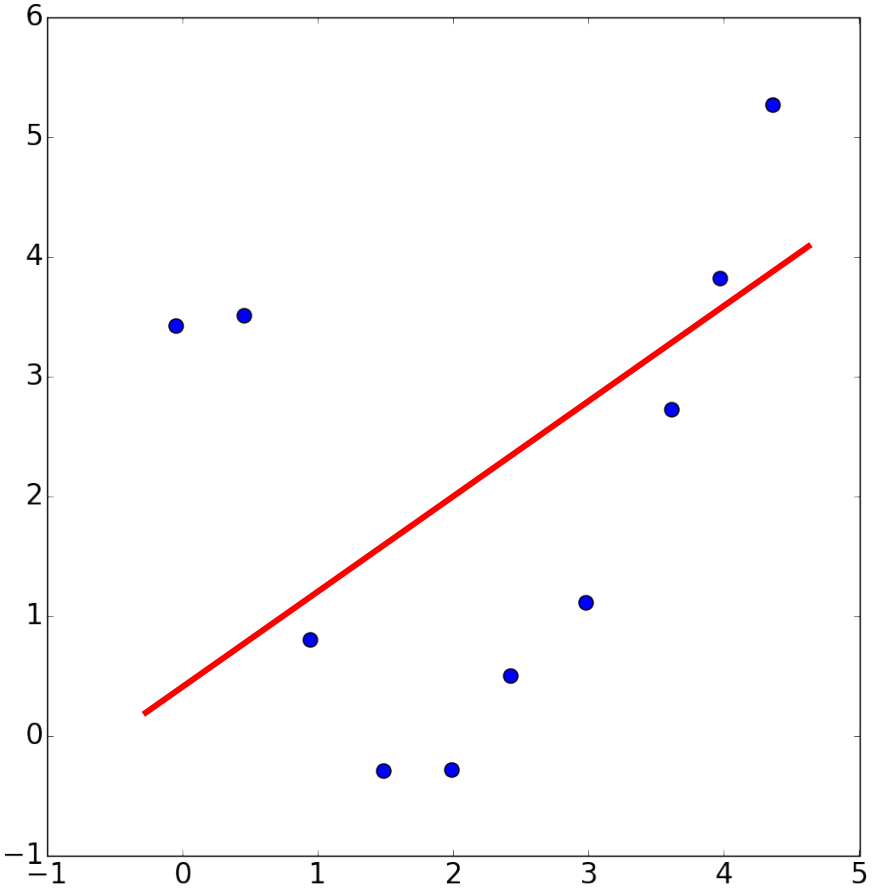
F-Distribution, Continued
Upper 0.01 Critical Points

$F_{0.01}(r_1, r_2)$									
	r_1								
r_2	10	15	20	25	30	40	60	120	∞
1	6055.9	6157.3	6208.7	6239.8	6260.7	6286.8	6313.0	6339.4	6365.9
2	99.40	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	27.23	26.87	26.69	26.58	26.50	26.41	26.32	26.22	26.13
4	14.55	14.20	14.02	13.91	13.84	13.75	13.65	13.56	13.46
5	10.05	9.72	9.55	9.45	9.38	9.29	9.20	9.11	9.02
6	7.87	7.56	7.40	7.30	7.23	7.14	7.06	6.97	6.88
7	6.62	6.31	6.16	6.06	5.99	5.91	5.82	5.74	5.65
8	5.81	5.52	5.36	5.26	5.20	5.12	5.03	4.95	4.86
9	5.26	4.96	4.81	4.71	4.65	4.57	4.48	4.40	4.31
10	4.85	4.56	4.41	4.31	4.25	4.17	4.08	4.00	3.91
11	4.54	4.25	4.10	4.01	3.94	3.86	3.78	3.69	3.60
12	4.30	4.01	3.86	3.76	3.70	3.62	3.54	3.45	3.36
13	4.10	3.82	3.66	3.57	3.51	3.43	3.34	3.25	3.17
14	3.94	3.66	3.51	3.41	3.35	3.27	3.18	3.09	3.00
15	3.80	3.52	3.37	3.28	3.21	3.13	3.05	2.96	2.87
16	3.69	3.41	3.26	3.16	3.10	3.02	2.93	2.84	2.75
17	3.59	3.31	3.16	3.07	3.00	2.92	2.83	2.75	2.65
18	3.51	3.23	3.08	2.98	2.92	2.84	2.75	2.66	2.57
19	3.43	3.15	3.00	2.91	2.84	2.76	2.67	2.58	2.49
20	3.37	3.09	2.94	2.84	2.78	2.69	2.61	2.52	2.42
21	3.31	3.03	2.88	2.79	2.72	2.64	2.55	2.46	2.36
22	3.26	2.98	2.83	2.73	2.67	2.58	2.50	2.40	2.31
23	3.21	2.93	2.78	2.69	2.62	2.54	2.45	2.35	2.26
24	3.17	2.89	2.74	2.64	2.58	2.49	2.40	2.31	2.21
25	3.13	2.85	2.70	2.60	2.54	2.45	2.36	2.27	2.17
26	3.09	2.81	2.66	2.57	2.50	2.42	2.33	2.23	2.13
27	3.06	2.78	2.63	2.54	2.47	2.38	2.29	2.20	2.10
28	3.03	2.75	2.60	2.51	2.44	2.35	2.26	2.17	2.06
29	3.00	2.73	2.57	2.48	2.41	2.33	2.23	2.14	2.03
30	2.98	2.70	2.55	2.45	2.39	2.30	2.21	2.11	2.01
40	2.80	2.52	2.37	2.27	2.20	2.11	2.02	1.92	1.80
60	2.63	2.35	2.20	2.10	2.03	1.94	1.84	1.73	1.60
120	2.47	2.19	2.03	1.93	1.86	1.76	1.66	1.53	1.38
∞	2.32	2.04	1.88	1.77	1.70	1.59	1.47	1.32	1.00

Goodness-of-fit

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





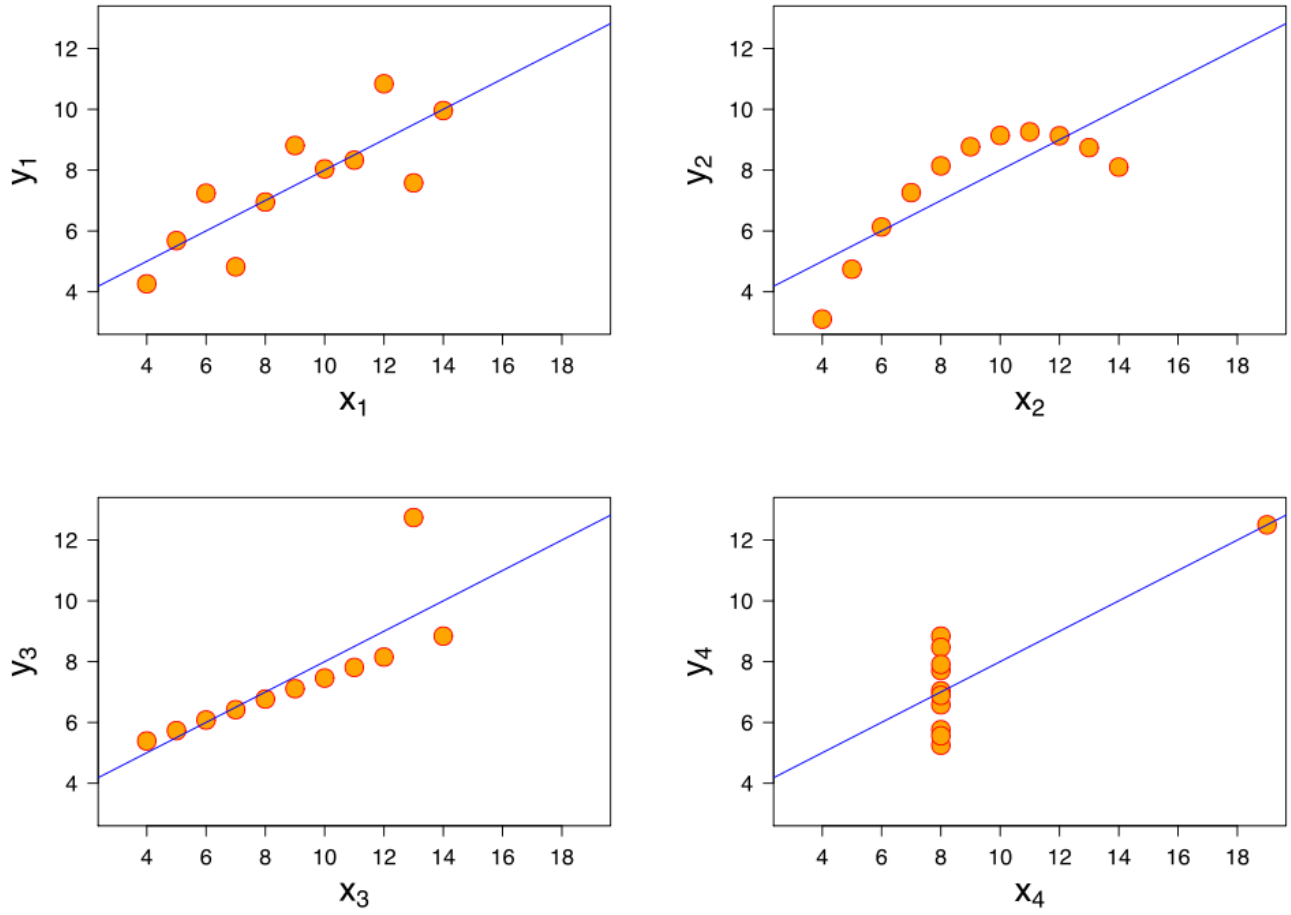
R^2

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

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

R^2 is NOT All-around Player

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



〈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

- Functions to measure regression performance
  - Mean squared error

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

  - Mean absolute error

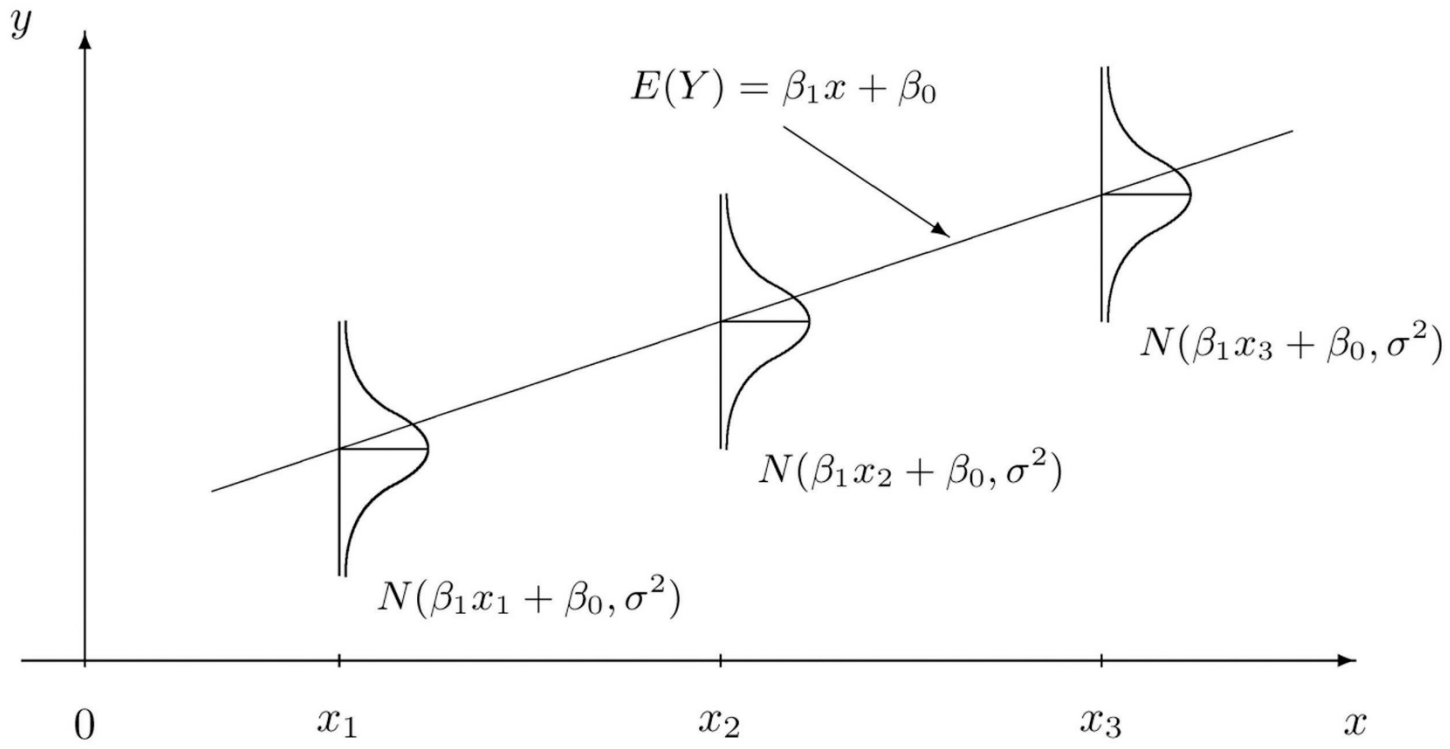
$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n}$$

  - Median absolute error
    - robust to outliers

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

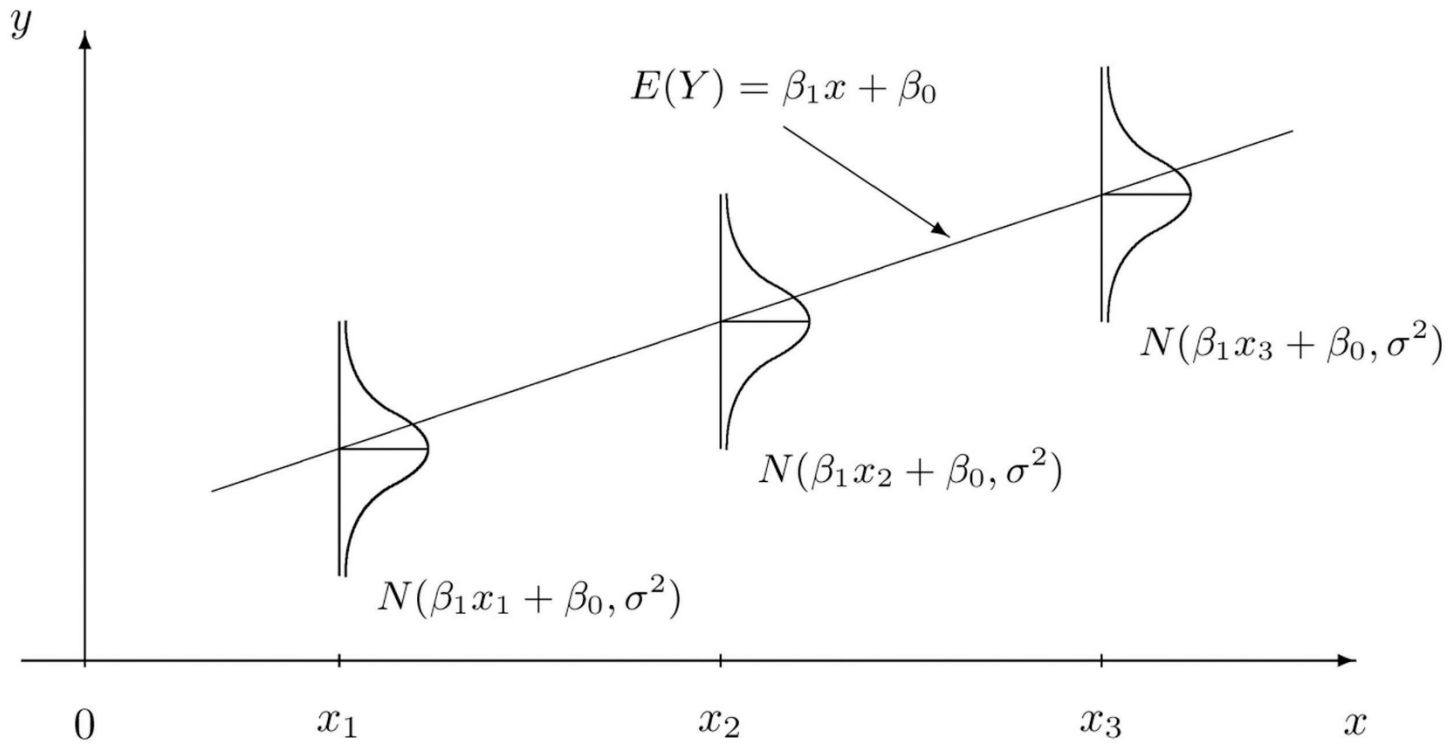
Check Appropriateness of Linear Regression

- Do you remember main assumptions of linear regression?



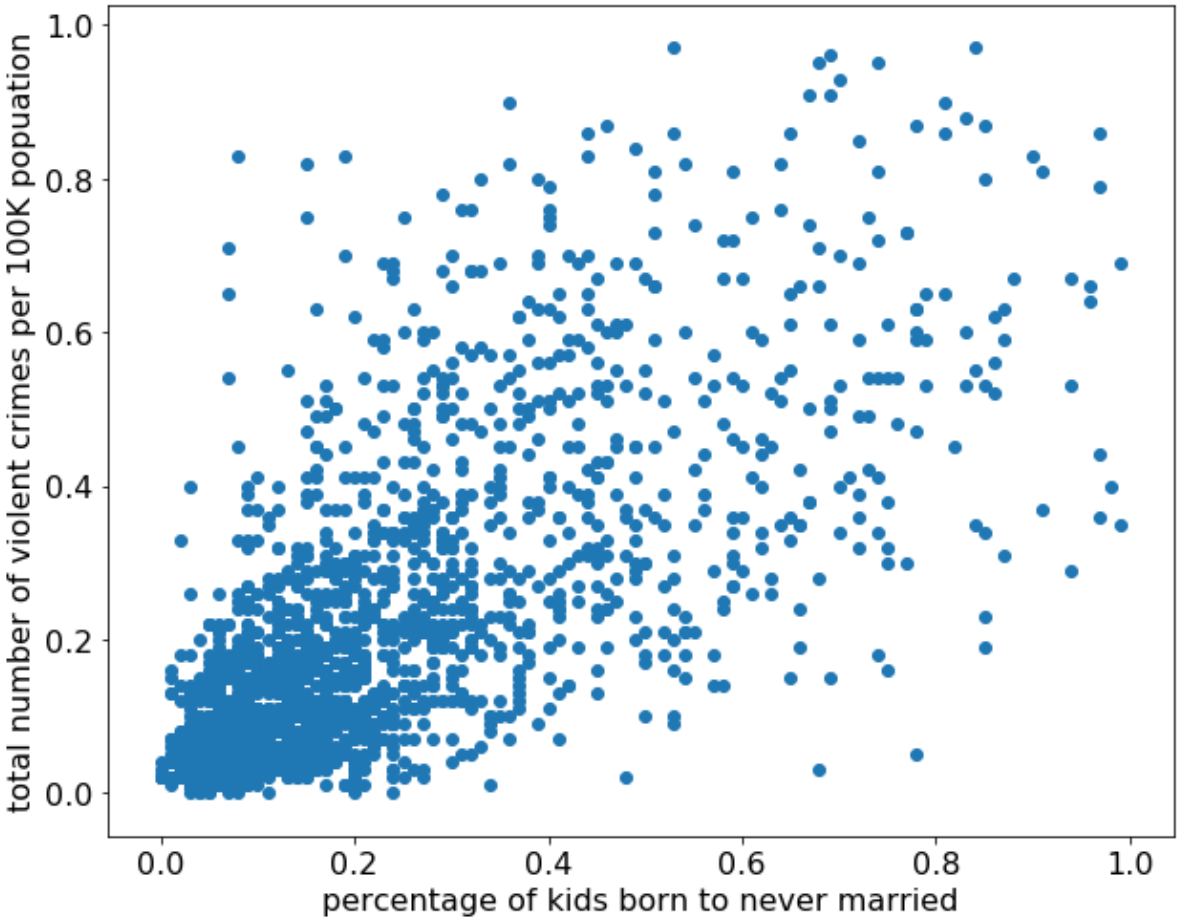
Main Assumption of Linear Regression

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



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

	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

$$corr = \begin{bmatrix} 1.00 & 0.98 & 0.17 \\ 0.98 & 1.00 & 0.11 \\ 0.17 & 0.11 & 1.00 \end{bmatrix}$$

  - x_1 and x_2 are highly correlated

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: $x = 1, y = 1$

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

Variance Inflation Factor

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

> **[Multicollinearity]**
> 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 $j-\text{th}$ input variable

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

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

- VIF

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

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