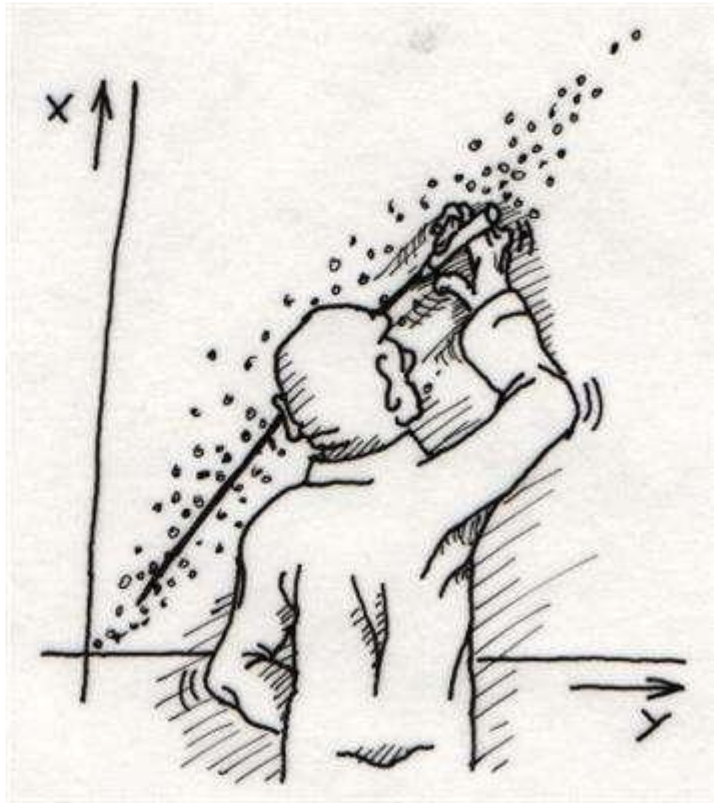


Multiple Linear Regression

강필성
고려대학교 산업경영공학부
pilsung_kang@korea.ac.kr

AGENDA

01	Multiple Linear Regression
02	Evaluating Regression Models
03	R Exercise

Multiple Linear Regression

- Regression Example: Predict the selling price of Toyota Corolla

Dependent variable (target)

Independent variables (attributes, features)

Variable	Description
Price	Offer Price in EUROs
Age_08_04	Age in months as in August 2004
KM	Accumulated Kilometers on odometer
Fuel_Type	Fuel Type (Petrol, Diesel, CNG)
HP	Horse Power
Met_Color	Metallic Color? (Yes=1, No=0)
Automatic	Automatic ((Yes=1, No=0)
CC	Cylinder Volume in cubic centimeters
Doors	Number of doors
Quarterly_Tax	Quarterly road tax in EUROs
Weight	Weight in Kilograms

Multiple Linear Regression

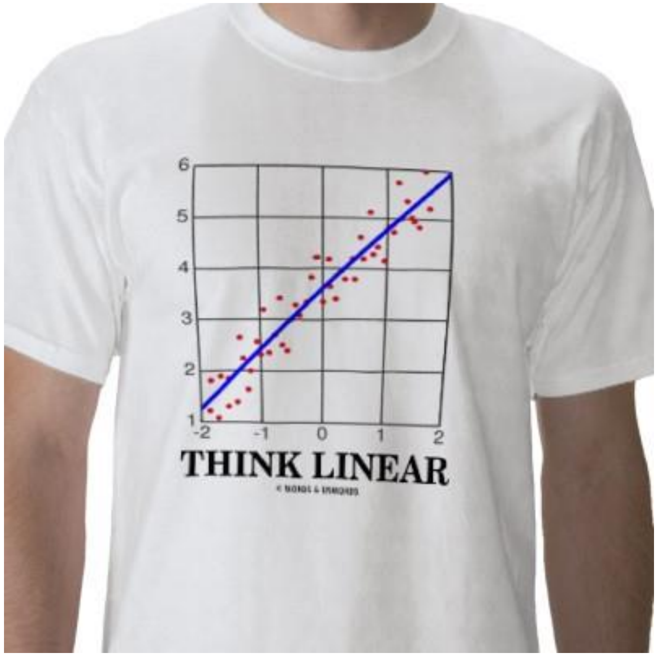
- Goal
  - ✓ Fit a linear relationship between a quantitative dependent variable Y and a set of predictors $X_1, X_2, \ldots, X_p$.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_d x_d + \epsilon$$

unexplained

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d$$

coefficients



Multiple Linear Regression

- Explanatory vs. Predictive

Explanatory Regression

- Explain relationship between predictors (explanatory variables) and target.
- Familiar use of regression in data analysis.
- Model Goal: Fit the data well and understand the contribution of explanatory variables to the model.
- “goodness-of-fit”: R^2, residual analysis, p-values.

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

Predictive Regression

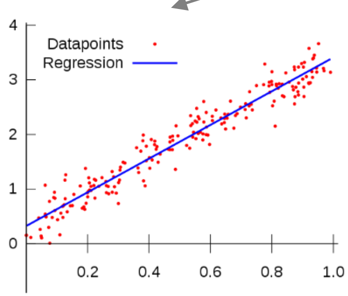
- Predict target values in other data where we have predictor values, but not target values.
- Classic data mining context
- Model Goal: Optimize predictive accuracy
- Train model on training data
- Assess performance on validation (hold-out) data
- Explaining role of predictors is not primary purpose (but useful)

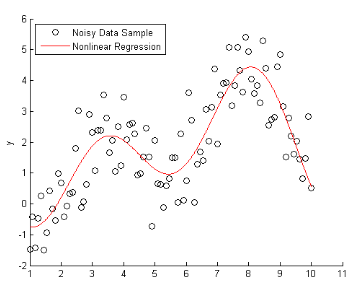
$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

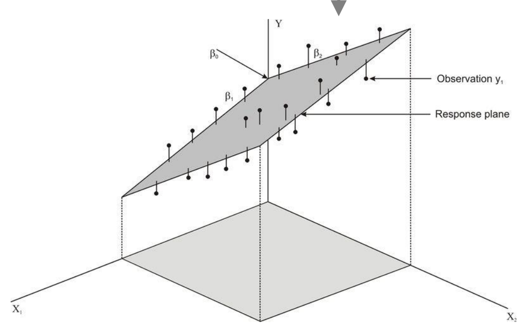
Multiple Linear Regression

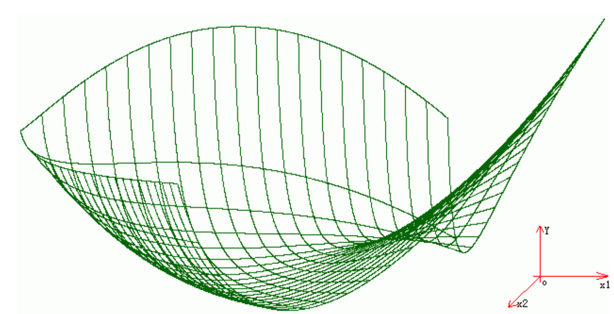
- Type of Regression

Regression Models
- Simple
  - Linear
  - Non-linear
- Multiple
  - Linear
  - Non-linear





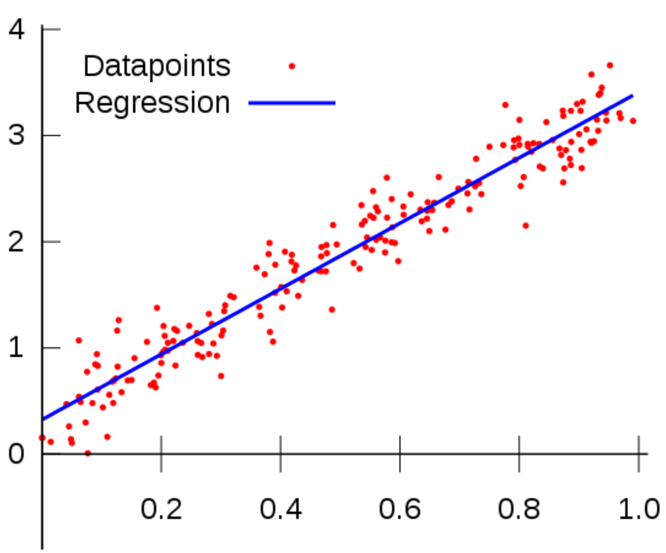


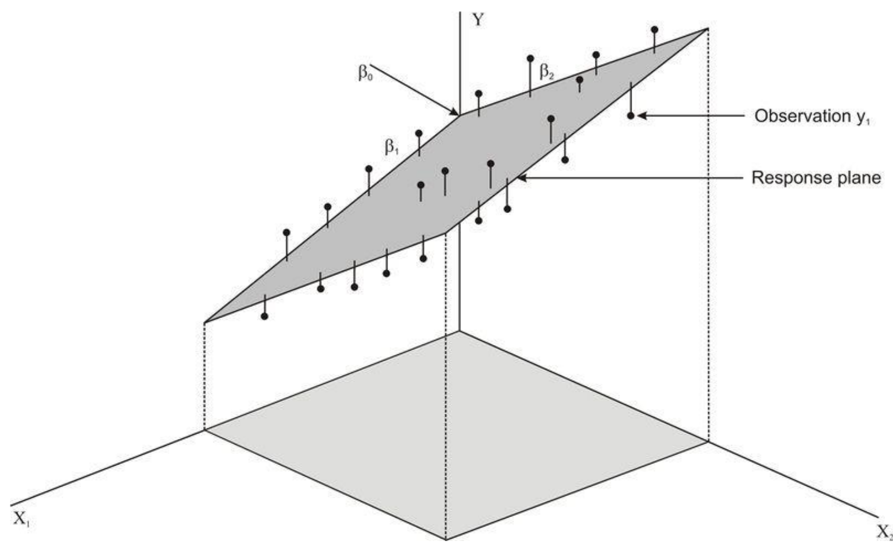


Multiple Linear Regression

- Linear Regression
  - ✓ Assume that the relationship between the input variable and the target variable is always linear.

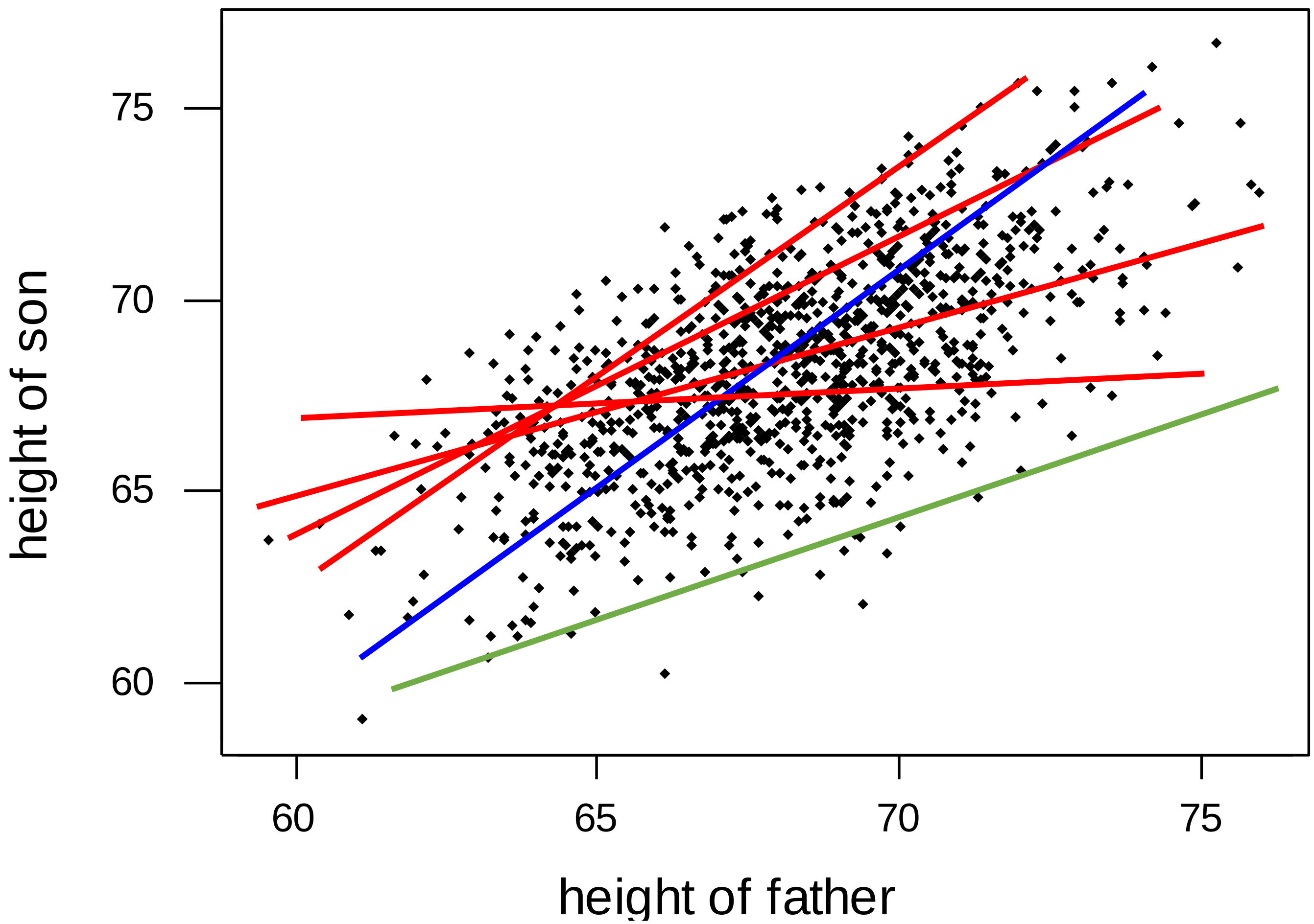
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d$$





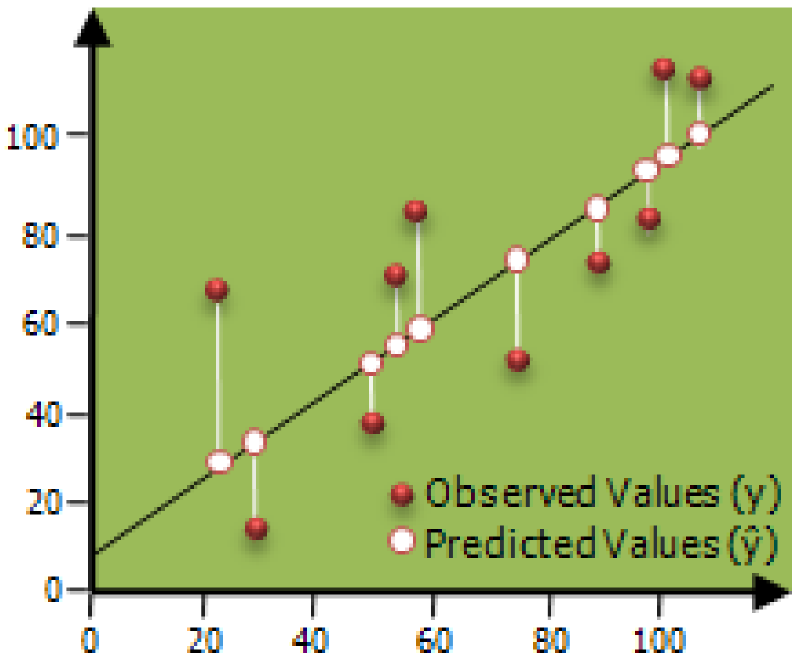
Multiple Linear Regression

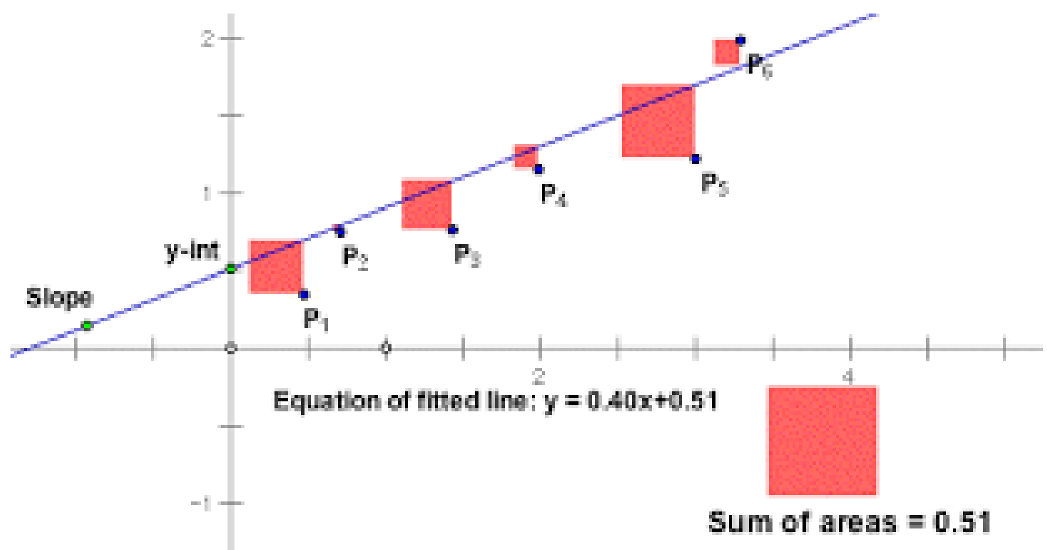
- Which line is optimal?



Multiple Linear Regression

- Estimating the coefficients
  - ✓ Ordinary least square (OLS): Minimize the squared difference between the actual target value and the estimated value by the regression model





Multiple Linear Regression

- Estimating the coefficients
  - ✓ Ordinary least square (OLS)
    - Actual target: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_d x_d + \epsilon$
    - Predicted target: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d$
    - Goal: minimize the difference between the actual and predicted target.

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

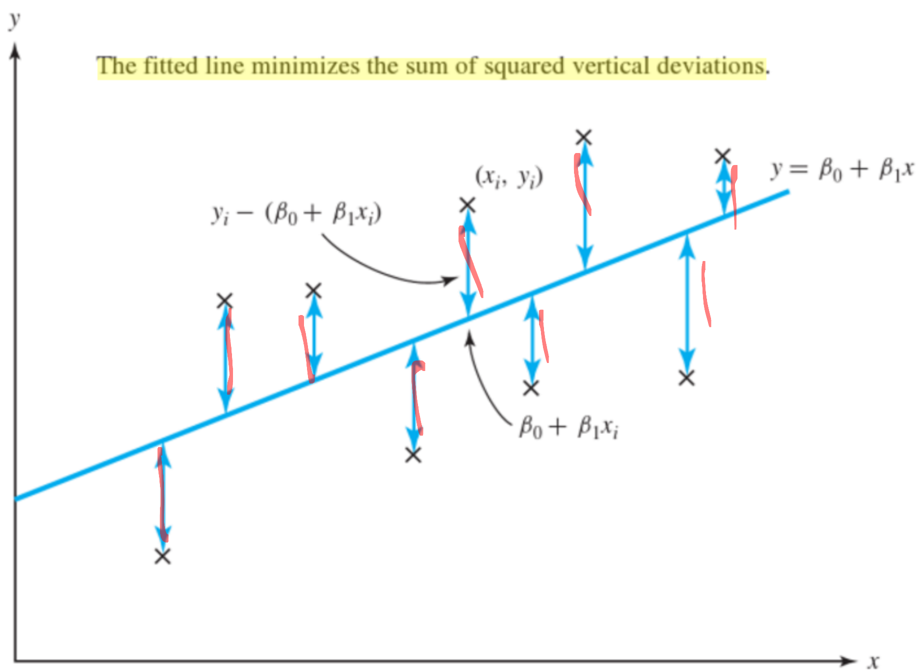
$$= \frac{1}{2} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} \cdots + \hat{\beta}_d x_{id})^2$$

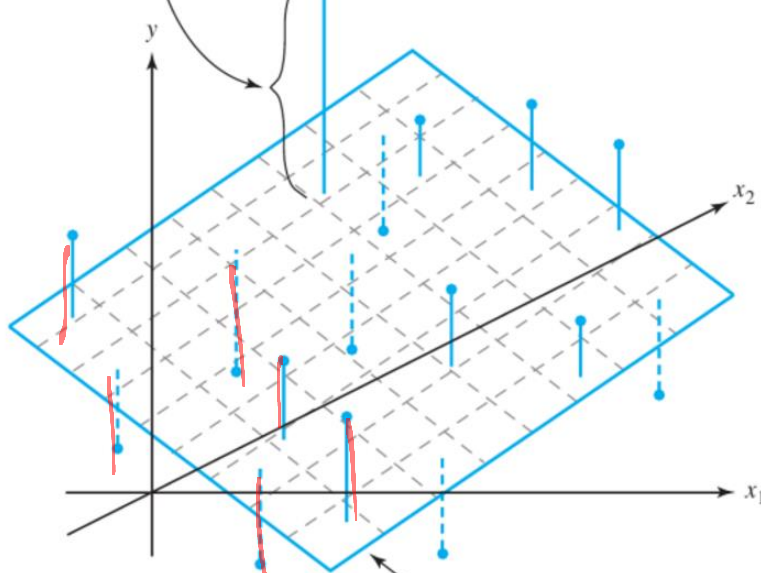
고려대학교 KOREA UNIVERSITY

DSBA Data Science & Business Analytics

Multiple Linear Regression

- Estimating the coefficients
  - ✓ Ordinary least square (OLS)



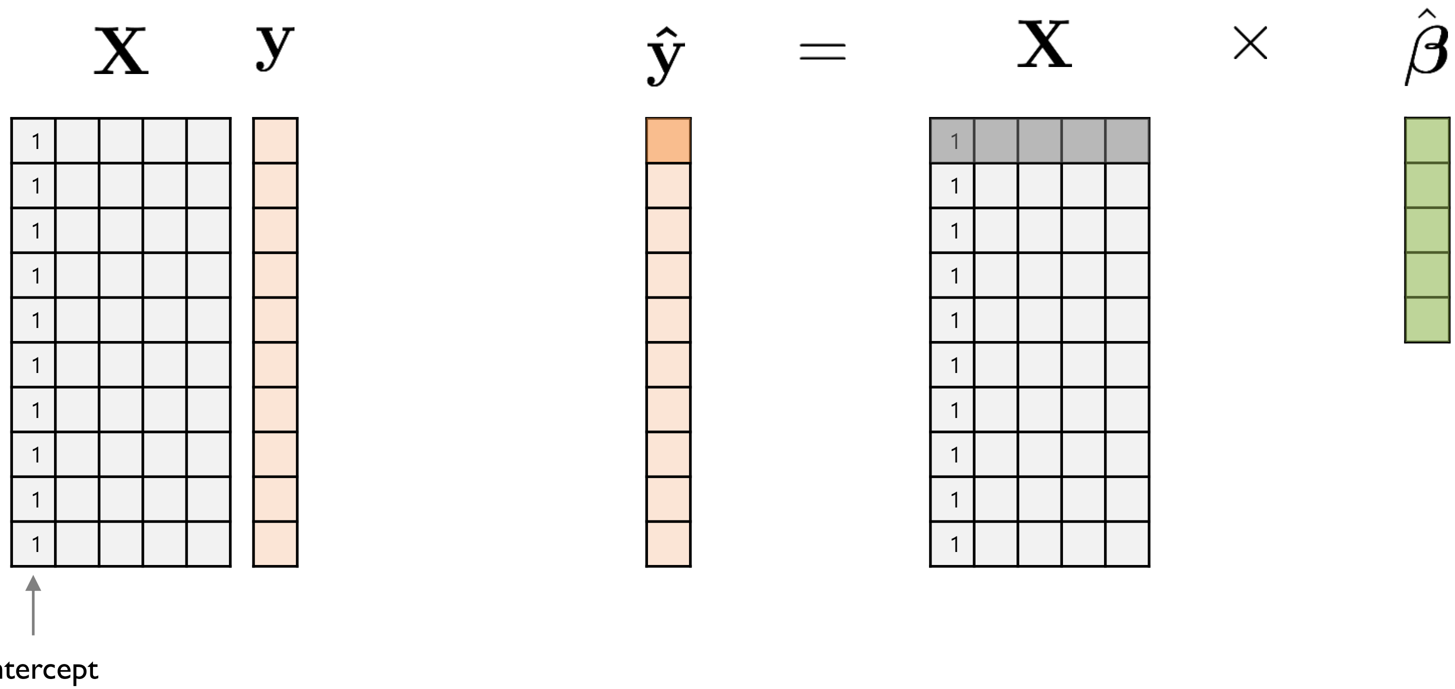


Multiple Linear Regression

- Ordinary least square: Matrix solution

$$\mathbf{X}: \; n \times (d+1) \; matrix, \;\; \mathbf{y}: \; n \times 1 \; vector$$

$$\hat{\beta}: (d+1) \times 1 \; vector$$



Multiple Linear Regression

- Ordinary least square: Matrix solution

$$\mathbf{X}: \; n \times (d+1) \; matrix, \; \mathbf{y}: \; n \times 1 \; vector$$

$$\hat{\boldsymbol{\beta}}: (d+1) \times 1 \; vector$$

$$\min E(\mathbf{X}) = \frac{1}{2}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^T \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$$

$$\Rightarrow \frac{\partial E(\mathbf{X})}{\partial \hat{\boldsymbol{\beta}}} = -\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) = 0$$

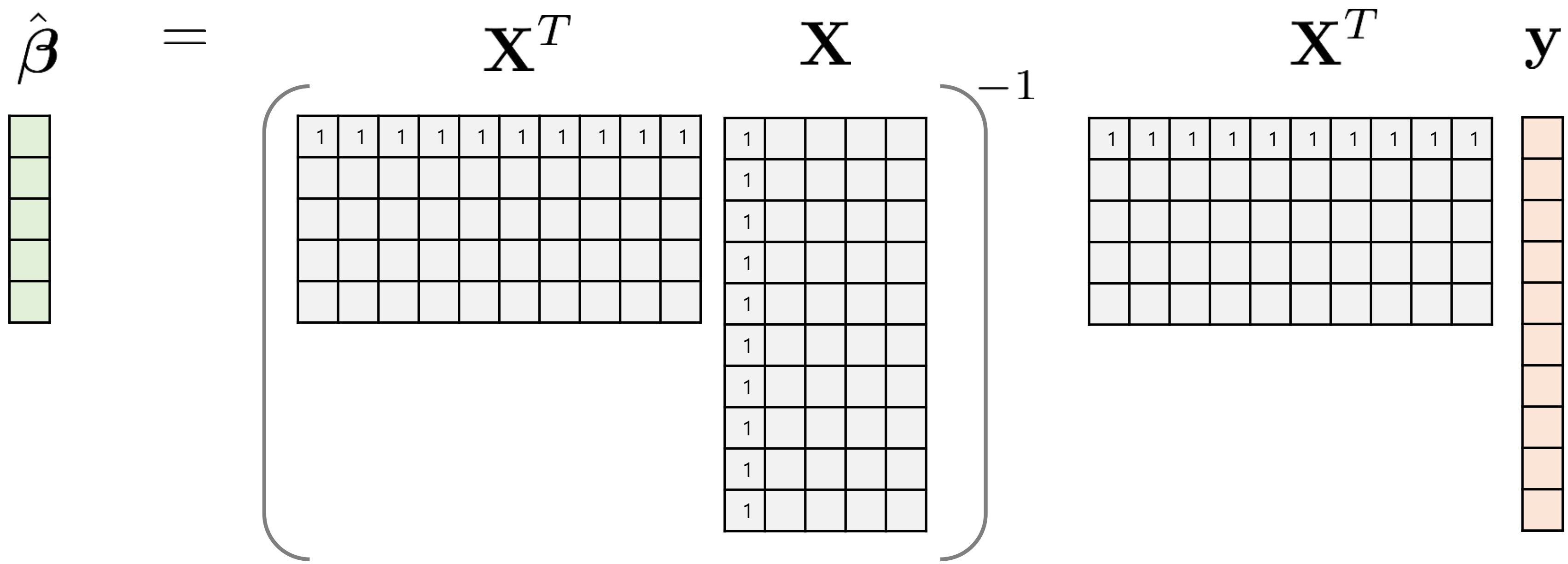
$$\Rightarrow \mathbf{X}^T\mathbf{y} + \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = 0$$

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \longrightarrow$$ Unique and explicit solution exists!

Multiple Linear Regression

- Ordinary least square: Matrix solution

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



Closed form solution for the regression coefficient

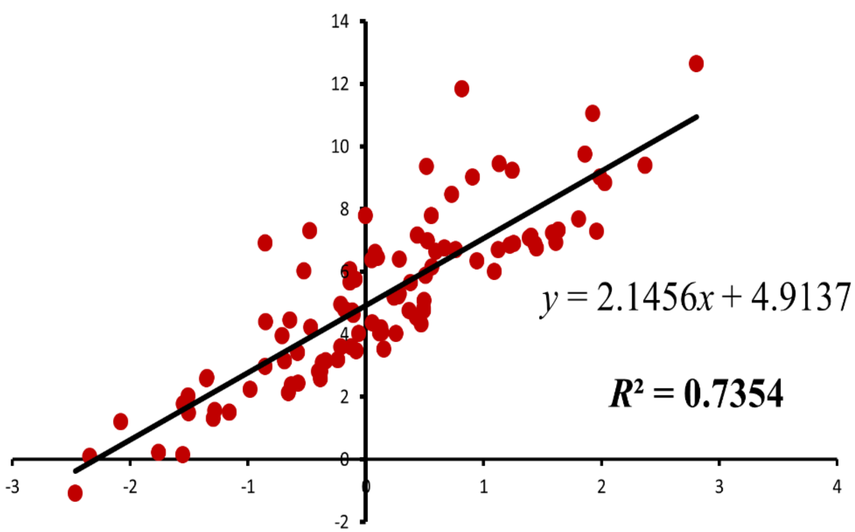
Multiple Linear Regression

- Ordinary least square
  - ✓ Finds the best estimates **β** when the following conditions are satisfied:
    - The noise ε follows a normal distribution.
    - The linear relationship is correct.
    - The cases are independent of each other.
    - The variability in Y values for a given set of predictors is the same regardless of the values of the predictors (<u>homoskedasticity</u>).

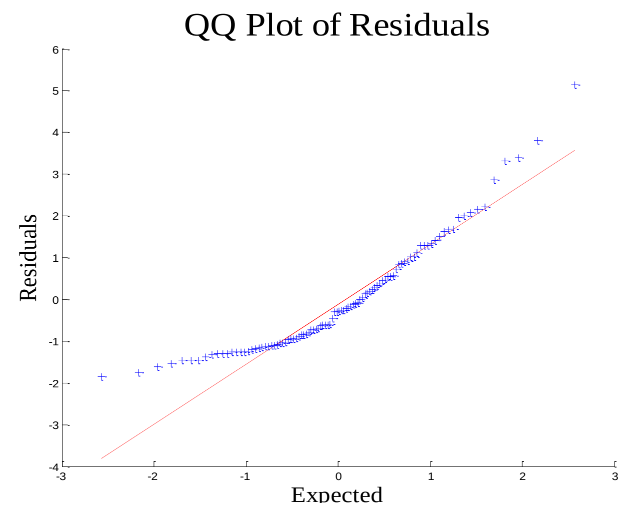
Multiple Linear Regression

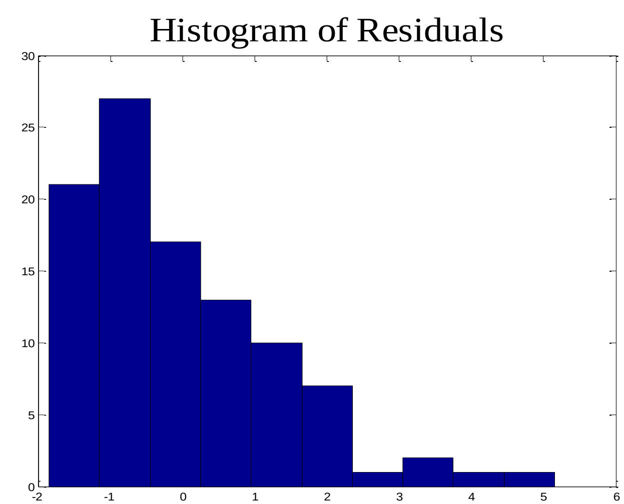
- Model checking

$$y = 2x + \varepsilon, \qquad \varepsilon \sim Gamma(2,1)$$



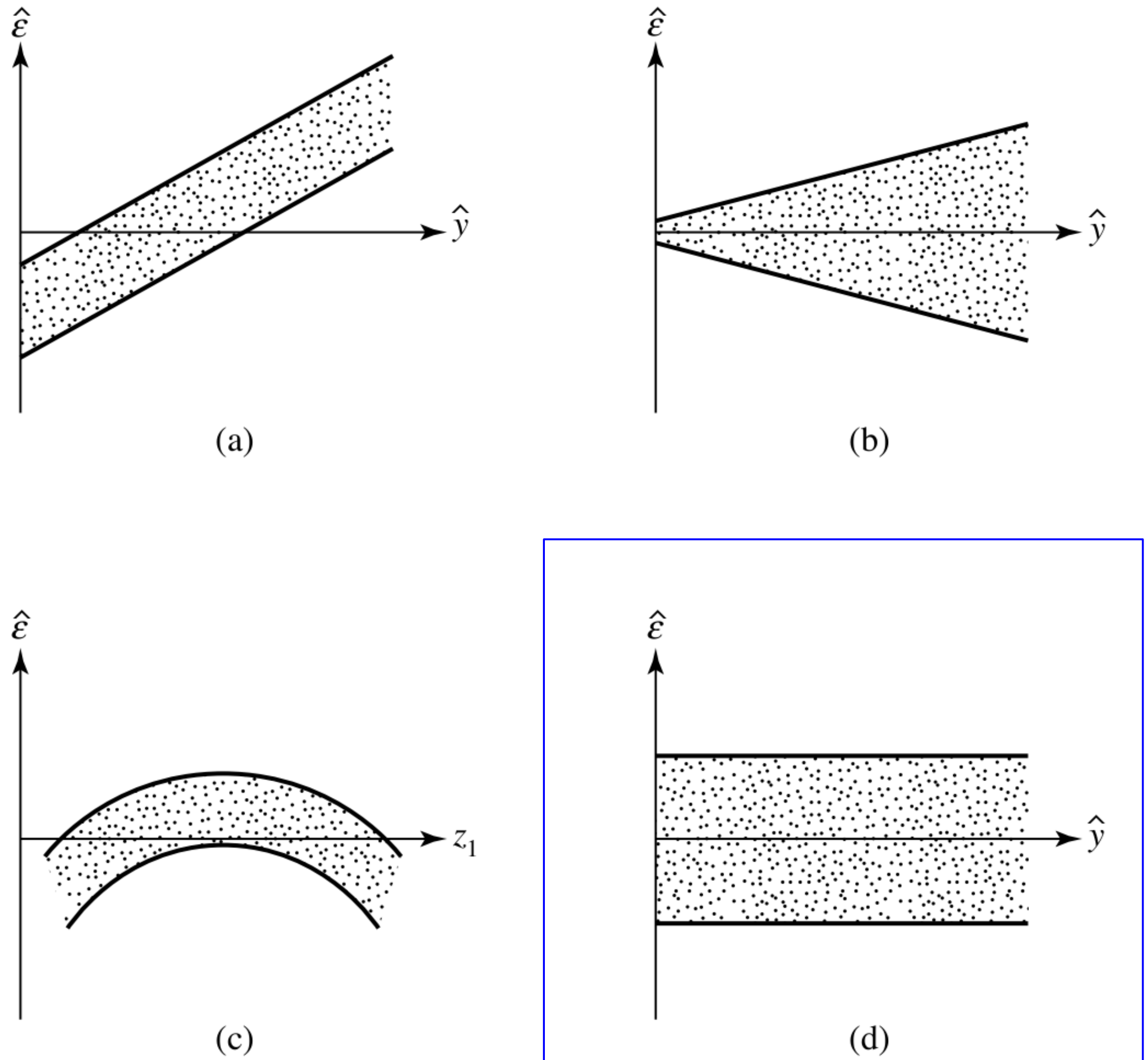
Regression model





Multiple Linear Regression

- Residual plots

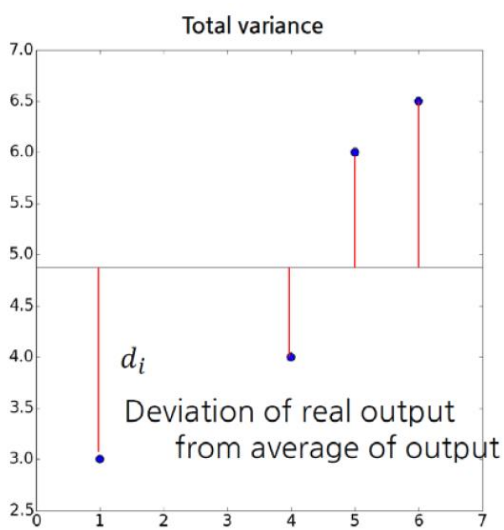


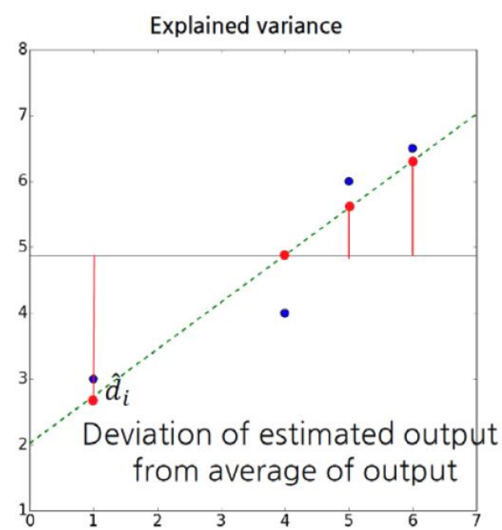
Multiple Linear Regression

- Goodness of fit
  - ✓ Sum-of-Squares Decomposition

$$\underbrace{\sum_{j=1}^{n}\left(y_j-\bar{y}\right)^2}_{\begin{pmatrix}\text{total sum of squares}\\ \text{about mean}\end{pmatrix}} = \underbrace{\sum_{j=1}^{n}\left(\hat{y}_j-\bar{y}\right)^2}_{\begin{pmatrix}\text{regression}\\ \text{sum of squares}\end{pmatrix}} + \underbrace{\sum_{j=1}^{n}\hat{\varepsilon}_j^2}_{\begin{pmatrix}\text{residual (error)}\\ \text{sum of squares}\end{pmatrix}}.$$

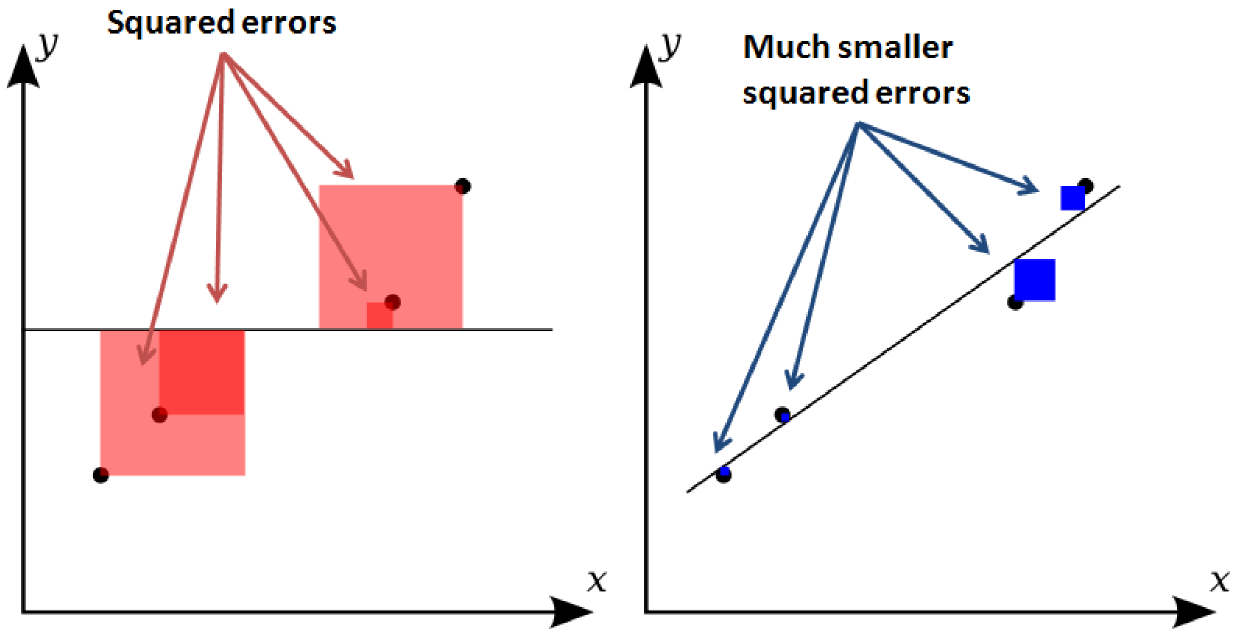
SST SSR SSE





Multiple Linear Regression

- Goodness-of-fit: (Adjusted) R^2
  - ✓ Graphical interpretation



Computationally:

$$\text{R-squared} = 1 - \left[\frac{SS_{error}}{SS_{total}}\right]$$

Conceptually:
Force x and y to be independent, calculate **the squared error.**
Allow for a relationship between x and y, does this reduce your **error?**

http://compare2what.blogspot.kr/2013_01_01_archive.html

Multiple Linear Regression

- Goodness-of-fit: (Adjusted) R^2

$$R^2 = 1 - \frac{\sum_{j=1}^{n} \hat{\varepsilon}_j^2}{\sum_{j=1}^{n} (y_j - \bar{y})^2} = \frac{\sum_{j=1}^{n} (\hat{y}_j - \bar{y})^2}{\sum_{j=1}^{n} (y_j - \bar{y})^2}$$

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

✓ Gives the proportion of the total variation in the y_i's explained by the predictor variables

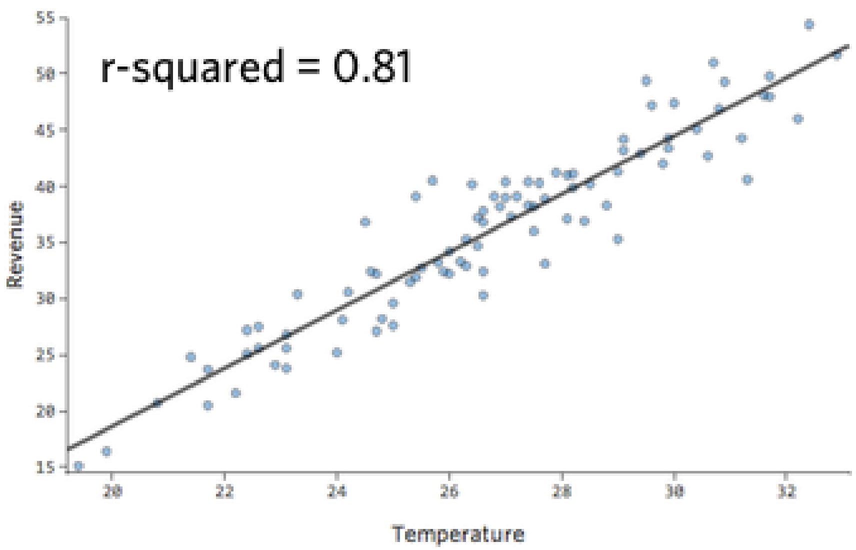
✓ $0 \leq R^2 \leq 1$

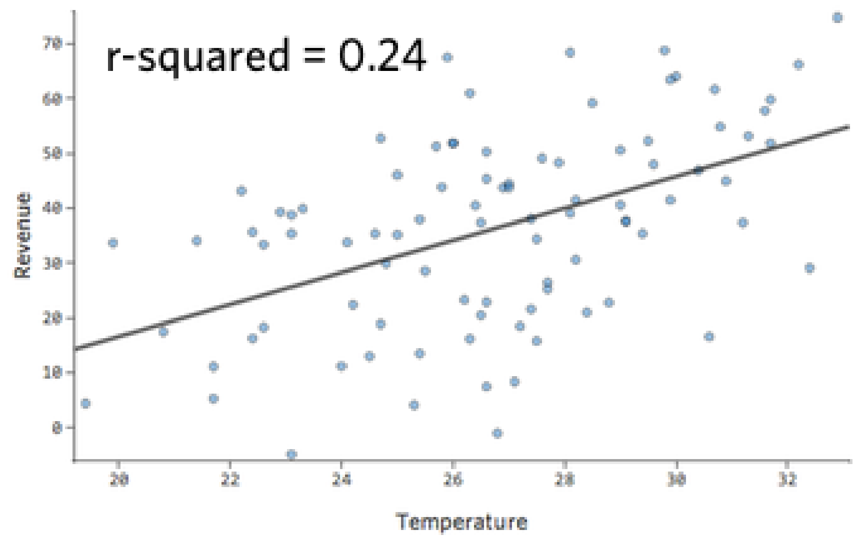
✓ $R^2 = 1$ → The fitted equation passes through all the data points

✓ $R^2 = 0$ → There is no linear relationship between the predictor variables and the target variable

Multiple Linear Regression

- Goodness-of-fit: (Adjusted) R^2
  - ✓ The proportionate reduction of total variation associated with the use of the predictor variable Z.





  - R^2 of your model is very high
    - I did a good job! (No!)
    - This dataset has a strong linear relationship between the X and y
    - Because everyone can have the same solution

Multiple Linear Regression

- Goodness-of-fit: (Adjusted) R^2
  - ✓ Adjusted R^2

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

  - ✓ R^2 increases monotonically when a (possibly not significant) new variable is added
  - ✓ Adjusted R^2 fix this problem
  - ✓ If an insignificant variable is added, the adjusted R^2 does not increase
- Model verification
  - ✓ Check whether the model satisfies the following assumptions
    - Residuals are independent
    - Residuals have zero mean and a constant variance

Multiple Linear Regression: Example

- Example: predict the selling price of Toyota corolla

Y X

Price	Age_08_04	KM	Fuel_Type	HP	Met_Color	Automatic	cc	Doors	Quarterly_Tax	Weight
13500	23	46986	Diesel	90	1	0	2000	3	210	1165
13750	23	72937	Diesel	90	1	0	2000	3	210	1165
13950	24	41711	Diesel	90	1	0	2000	3	210	1165
14950	26	48000	Diesel	90	0	0	2000	3	210	1165
13750	30	38500	Diesel	90	0	0	2000	3	210	1170
12950	32	61000	Diesel	90	0	0	2000	3	210	1170
16900	27	94612	Diesel	90	1	0	2000	3	210	1245
18600	30	75889	Diesel	90	1	0	2000	3	210	1245
21500	27	19700	Petrol	192	0	0	1800	3	100	1185
12950	23	71138	Diesel	69	0	0	1900	3	185	1105
20950	25	31461	Petrol	192	0	0	1800	3	100	1185
19950	22	43610	Petrol	192	0	0	1800	3	100	1185
19600	25	32189	Petrol	192	0	0	1800	3	100	1185
21500	31	23000	Petrol	192	1	0	1800	3	100	1185
22500	32	34131	Petrol	192	1	0	1800	3	100	1185
22000	28	18739	Petrol	192	0	0	1800	3	100	1185
22750	30	34000	Petrol	192	1	0	1800	3	100	1185
17950	24	21716	Petrol	110	1	0	1600	3	85	1105
16750	24	25563	Petrol	110	0	0	1600	3	19	1065

Multiple Linear Regression: Example

- Data preprocessing
  - ✓ Create dummy variables for fuel types

	Fuel_type = Disel	Fuel_type = Petrol	Fuel_type = CNG
Diesel	1	0	0
Petrol	0	1	0
CNG	0	0	1

- Data partitioning
  - ✓ 60% training data / 40% validation data

Id	Model	Price	Age_08_04	Mfg_Month	Mfg_Year	KM	Fuel_Type_Diesel	Fuel_Type_Petrol
1	RRA 2/3-Doors	13500	23	10	2002	46986	1	0
4	RRA 2/3-Doors	14950	26	7	2002	48000	1	0
5	SOL 2/3-Doors	13750	30	3	2002	38500	1	0
6	SOL 2/3-Doors	12950	32	1	2002	61000	1	0
9	/VT I 2/3-Doors	21500	27	6	2002	19700	0	1
10	RRA 2/3-Doors	12950	23	10	2002	71138	1	0
12	BNS 2/3-Doors	19950	22	11	2002	43610	0	1
17	ORT 2/3-Doors	22750	30	3	2002	34000	0	1

Multiple Linear Regression: Example

- Fitted linear regression model

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	-3608.418457	1458.620728	0.0137	97276410000
Age_08_04	-123.8319168	3.367589	0	8033339000
KM	-0.017482	0.00175105	0	251574500
Fuel_Type_Diesel	210.9862518	474.9978333	0.6571036	6212673
Fuel_Type_Petrol	2522.066895	463.6594238	0.00000008	4594.9375
HP	20.71352959	4.67398977	0.00001152	330138600
Met_Color	-50.48505402	97.85591125	0.60614568	596053.75
Automatic	178.1519013	212.0528565	0.40124047	19223190
cc	0.01385481	0.09319961	0.88188446	1272449
Doors	20.02487946	51.0899086	0.69526076	39265060
Quarterly_Tax	16.7742424	2.09381151	0	160667200
Weight	15.41666317	1.40446579	0	214696000

β

Significance Probability

Multiple Linear Regression: Example

- Interpret the result
  - ✓ Regression coefficient
    - Beta value for the corresponding predictor variable
    - The amount of change when the predictor variable increases by 1
    - If it is positive/negative, then the predictor variable and the target variable are positively/negatively correlated

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	-3608.418457	1458.620728	0.0137	97276410000
Age_08_04	-123.8319168	3.367589	0	8033339000
KM	-0.017482	0.00175105	0	251574500
Fuel_Type_Diesel	210.9862518	474.9978333	0.6571036	6212673
Fuel_Type_Petrol	2522.066895	463.6594238	0.00000008	4594.9375
HP	20.71352959	4.67398977	0.00001152	330138600
Met_Color	-50.48505402	97.85591125	0.60614568	596053.75
Automatic	178.1519013	212.0528565	0.40124047	19223190
cc	0.01385481	0.09319961	0.88188446	1272449
Doors	20.02487946	51.0899086	0.69526076	39265060
Quarterly_Tax	16.7742424	2.09381151	0	160667200
Weight	15.41666317	1.40446579	0	214696000

Multiple Linear Regression: Example

- Interpret the result
  - ✓ p-value
    - Indicate whether the regression coefficient is statistically significant or not
    - A predictor variable is important for modeling when its p-value is close to 0
    - Can be used to select significant variables (e.g., use the variables with p-value less than 0.05)

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	-3608.418457	1458.620728	0.0137	97276410000
Age_08_04	-123.8319168	3.367589	0	8033339000
KM	-0.017482	0.00175105	0	251574500
Fuel_Type_Diesel	210.9862518	474.9978333	0.6571036	6212673
Fuel_Type_Petrol	2522.066895	463.6594238	0.00000008	4594.9375
HP	20.71352959	4.67398977	0.00001152	330138600
Met_Color	-50.48505402	97.85591125	0.60614568	596053.75
Automatic	178.1519013	212.0528565	0.40124047	19223190
cc	0.01385481	0.09319961	0.88188446	1272449
Doors	20.02487946	51.0899086	0.69526076	39265060
Quarterly_Tax	16.7742424	2.09381151	0	160667200
Weight	15.41666317	1.40446579	0	214696000


ANY
questions?