veverage - dist ($\hat{Y}_0$, mean) → outhers

heteroscadasticity — ~~e~~ error not const in predicted vs residuals

Question 1

Linear Regression :

1) $Y \longrightarrow$ output $(n \times 1)$

$X \longrightarrow$ input $(n \times p)$

$\beta \longrightarrow$ coefficients $(p \times 1)$

$\varepsilon \longrightarrow$ errors $(n \times 1)$

$$Y = X\beta + \varepsilon$$

Assuming that the output is a linear model of the input & errors have mean zero, & constant variance

2) OLS — sum of squares of (actual - predicted)

$$\Sigma (y_i - \hat{y}_i)^2$$

MSE — OLS / number of outputs

$$MSE = \frac{1}{n} \| \underset{\text{actual}}{y} - X\beta \|^2$$

$$= \frac{1}{n} (y - X\beta)^T (y - X\beta)$$

3) ~~$\frac{\partial (OLS)}{\partial \beta} =$~~

3) $\frac{\partial}{\partial \beta} (OLS/MSE) =$

$$OLS = (y^T - \beta^T X^T)(y - X\beta)$$

Key step – recognizing the dimension of $y^TX\beta \rightarrow (1\times1)$

$\therefore (y^TX\beta)^T = y^TX\beta$

$$y^Ty - y^TX\beta - \beta^TX^Ty + \beta^TX^TX\beta$$

$$y^Ty - 2\beta^TX^Ty + \beta^TX^TX\beta$$

$$\frac{d}{d\beta}\left(y^Ty - 2\beta^TX^Ty + \beta^TX^TX\beta\right)$$

$$0 = 0 - 2X^Ty + X^TX\hat\beta + (X^TX)^T\hat\beta$$

$$\downarrow$$

$$\frac{d(\beta^T\alpha)}{d\beta} = \alpha$$

$$\cancel{2}X^Ty = \cancel{2}X^TX\hat\beta$$

$$\hat\beta = (X^TX)^{-1}X^Ty$$

4) $\det(X^TX) \neq 0$ & X^TX should be a square matrix. X^TX will always be square

If we have multicollinearity occurs then we could have transformed the matrix to make a column zero and thus $|\det|$ would be zero

5) $\hat{y} = X\hat\beta + \cancel{\varepsilon}$

$y = X\beta + \varepsilon$

$\boxed{y - \hat{y} = \varepsilon}$

To prove : $X^T(y - \hat{y}) = 0$

$= X^T(y - X\hat\beta)$

$$= x^T y - x^T x \hat{\beta}$$

$$= 0 \quad \text{(Above derivation of } \hat{\beta} \text{)}$$

6) $$J(\beta) = \frac{1}{2n} \| x\beta - y \|^2$$

$$\nabla_\beta \, (x\beta - y)^T (x\beta - y)$$

$$\frac{\partial}{\partial\beta} \quad (\beta^T x^T - y^T)(x\beta - y)$$

~~$\beta^T x^T - y^T x\beta$~~

$$\beta^T x^T x\beta - y^T x\beta - \beta^T x^T y + y^T y$$

$$\boxed{\frac{\partial}{\partial\beta} = \frac{-2x^T y + 2x^T x\beta}{2n}}$$

$$\frac{\partial}{\partial\beta} = \frac{x^T x\beta - x^T y}{n}$$

By substituting this value we will get the result.

Batch gradient rule

$$\beta^{(t+1)} = \beta^{(t)} - \alpha \nabla_\beta J(\beta^{(t)})$$

10) Heavy tails — extreme value exist in a large number thus causing difficulty to predict the extreme outliers & error will be misleading

V shaped — upward — right-skewed & positive outliers

downward — left skewed & negative outliers

__/__/____

12) $E((y-\hat{f}(x))^2) = Bias^2 + Var + \sigma^2$

$y = f(n) + \varepsilon$

$E(\varepsilon) = 0$

$Var(\varepsilon) = E(\varepsilon^2) = \sigma^2$

$$E\left(\left((f(x) - \hat{f}(n)) + \varepsilon\right)^2\right)$$

$$= E\left((f(n) - \hat{f}(n))^2 + \underbrace{\cancel{2(f(x) - \hat{f}(n))\varepsilon}}_{0} + \underbrace{\varepsilon^2}_{\sigma^2}\right)$$

~~$E(f^2(n) - 2f(n)(f(n) + \hat{f}(n))^2 + \sigma^2)$~~

$$E\left(\left(f(n) - E(\hat{f}(n)) - (\hat{f}(n) - E\hat{f}(n))\right)^2\right)$$

$$E\left((f(n) - E(\hat{f}(n)))^2 - 2(f(n) - \cancel{E(\hat{f}(n)))(\hat{f}(n) - E}(\hat{f}(n))^{0}\right.$$

$$\left. + (\hat{f}(n) - E(\hat{f}(n)))^2\right) + \sigma^2$$

$E(f(n) - E(\hat{f}(n)))^2$ $\downarrow$ $Bias^2$ $+$ $E((\hat{f}(n) - E(\hat{f}(n)))^2)$ $\downarrow$ var

$Bias^2 + Var + \sigma^2$

(3) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

$y = \alpha_0 + \alpha_1 x_1 + u$

a) $$E(\alpha_1) = \frac{\text{cov}(x_1, y)}{\text{Var}(x_1)}$$

$$= \frac{\text{cov}(x_1, \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon)}{\text{var}(x_1)}$$

$$= \overset{\to 0}{\cancel{\text{cov}(x_1, \beta_0)}} + \text{cov}(x_1, \beta_1 x_1) + \text{cov}(x_1, \beta_2 x_2) + \cancel{\text{cov}(x_1, \varepsilon)} \to 0$$

$$= \frac{\beta_1 \text{Var}(x_1) + \beta_2 \text{cov}(x_1, x_2)}{\text{var}(x_1)}$$

$$= \beta_1 + \beta_2 \frac{\text{cov}(x_1, x_2)}{\text{Var}(x_1)}$$

b) $(E(\alpha_1) - \beta_1)$ shows us the bias of α_1 towards x_2

c) for removing the the bias eithe β_2 or $\text{cov}(x_1, x_2)$ should be zero .

_/_/_

14) ~~a) We can check for the ratio of coefficients of each column if the vaw~~

a) We can check for the ratio of max & min eigen value
if the ratio is close to one
the no multicollinearity otherwise if values are large the multicollinearity occurs.

b)

$$vr(\hat{\beta}_j) = \frac{\sigma^2}{(n-1)\, vr(x_j)} \cdot \boxed{\frac{1}{1-\underbrace{\boxed{R_j^2}}_{\text{correlatio}}}} \rightarrow \text{Varionce inflation factor}$$

So $var(\hat{\beta}_j) \uparrow$ as ~~corretion~~ correlation $\uparrow$

c) Instability is because ~~the~~ error of the loss functions are flattened
As long as the matrix in invertible
bias will be singular

Question 2

2) for education level to handle the missing data mode will be the most suitable as mean & median of a category cannot be calculated. Same thing for city

Salaries median would be suitable as this property is non-linear and thus it would be able to manage the outliers well. The same logic could work for performance score

3) Encode categorical features

assigning educational levels numbers as there is a clear hierarchy in the levels where PhD is the highest and high school is lowest

for remote worker either 0 or 1

one hot encoding for the rest of the features because assigning numbers will be a problem as by assigning numbers it will generate an ordinality between the categories.

4) for stratification the variable used is salary. Salary helps us bifurcate the data so that extreme high outliers do not affect the low salaries as would have happened if the datasets were randomly chosen.