Fri Feb 4 Lab 2

How well my model preform compared with baseline?

I. Understand R^2: regression model v.s. base line

average of sample outcomes

- R^2: coefficient of determination

  measured on training data

sum of squares total

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i \in train} (y_i - \hat{y}_i^{train})^2}{\sum_{i \in train} (y_i - \bar{y})^2}$$

from linear regression, we have $\hat{\beta}$.

$\hat{y}^{train}$ is the prediction result using $\hat{\beta}$

$$\hat{y}_i^{train} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$$

prediction result using $\hat{\beta}$

X data in training set

- OSR^2 out-of-sample R^2

  measured on test set

$$OSR^2 = 1 - \frac{SSE\ (test\ set)}{SST\ (test\ set)}$$

Y data in test test

$$= 1 - \frac{\sum_{i \in test} (y_i - \hat{y}_i^{test})^2}{\sum_{i \in test} (y_i - \bar{y})^2}$$

$$\hat{y}_i^{test} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$$

X data in test set

$\bar{y}$ is the base line model calculated on the training set

$\bar{y}$ are same in R^2 and OSR^2:

$$\bar{y} = \frac{1}{n} \sum_{i \in train} y_i, \quad n = \text{train set size}$$

II. P-Value & Confidence Inteval

- Hypothesis Testing

X_j is useful in predicting the response ?

null hypothesis — $H_0: \beta_j = 0$; x_j is useless

alternative hypothesis — $H_a: \beta_j \neq 0$. x_j is useful

— Assumptions for statistical analysis

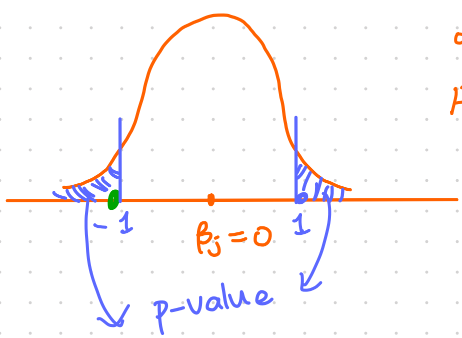
Linear model is true: $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

- p-value of feature j:

From our observed data $\{x_i, y_i\}$'s, we can find $\hat{\beta}_j$ using linear regression.

p-value is the probability that we observe such large $|\hat{\beta}_j|$ or even something deviates from 0 more if true $\beta_j = 0$.

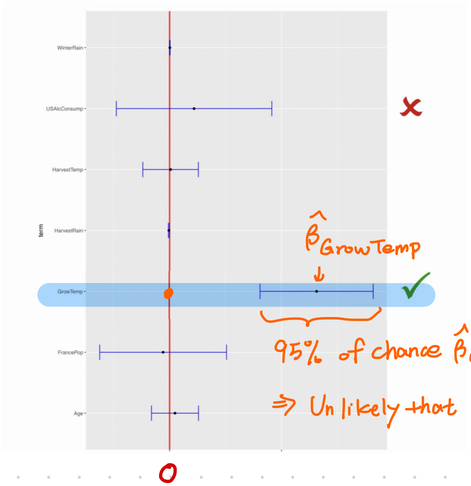


If true $\beta_j = 0$, then the distribution of $\hat{\beta}_j$ centered at 0.

After running linear regression, if it tells us $\hat{\beta}_j = 1$, p-value is the sum of two purple areas.

- Confidence Interval

$1-\alpha$ CI : there is $1-\alpha$ chance $\hat{\beta}_j$ falls in this range.



95% of chance $\hat{\beta}_{GrowTemp}$ in this range.

$\Rightarrow$ Unlikely that $\beta_{GrowTemp} = 0$!

- Relationship Between CI and p-value :

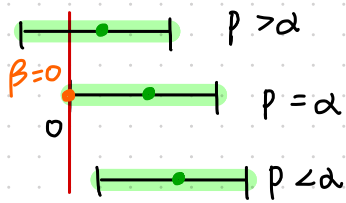
test whether $\beta_j = 0$

Hypothesis testing is equivalent to looking at CIs.

Reject null Hypothesis ($\beta_j = 0$) at significance level α

$\Updownarrow$

$(1-\alpha)$ CI does NOT contain 0



p-value $= \alpha$:	$(1-\alpha)$ CI exactly touches 0			
$> \alpha$:	$0 \in (1-\alpha)$ CI	$\beta_j = 0$	X_j not significant	
$< \alpha$:	$0 \notin (1-\alpha)$ CI	$\beta_j \neq 0$	X_j significant	

III. VIF

multicollinearity problem:

Occurs when 2 or more predictors are highly correlated.

can exist without large correlations in the correlation table!

→ Need to check VIF

VIF_j :

Consider regressing predictor variable X_j on all the other variables

$$X_j = a_0 + a_1 X_1 + \cdots + a_{j-1} X_{j-1} + a_{j+1} X_{j+1} + \cdots + a_p X_p \quad (*)$$

Let R_j^2 be the R^2 of the above linear regression problem.

If there is a perfect linear relationship between X_j and others, it means using other X variables to predict X_j can give an accurate prediction. Then, R_j^2 is close to 1.

$$\boxed{VIF_j := \frac{1}{1 - R_j^2}}$$

Then, $VIF_j \uparrow \infty$.

VIF_j very large $\Rightarrow$ X_j is a linear combination of others.

$\Rightarrow$ Remove X_j !