ASSESSING The accuracy of the ~~model~~ coefficient estimates.

We assume that true relationship b/w X & Y takes the form

$$Y = f(X) + \epsilon$$

function f & ϵ ⇒ mean zero random error term.

IF Y is to be estimated by a linear function then

$$Y \approx \beta_0 + \beta_1 X$$

$$\text{or } Y = \beta_0 + \beta_1 X + \epsilon \Leftarrow *$$

Defn:- Population Regression lines are slightly different to the OLS lines but are close enough as OLS is estimating.

on a real data, it is just unobserved.

The difference is subtle & can be understood by understanding the difference b/w sample mean & population mean.

In fact $\boxed{\hat{\mu} = \bar{y}}$ is widely used.

An intuitive question :- We have established that one estimate might over or underestimate a given variable like $\hat{\mu}$ or regression lines, but how can we judge how far off is that particular estimate?

Ans: Standard Error $SE(\hat{\mu})$

$$\Rightarrow Var(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{(\sigma)^2}{n}$$

SE tells us the avg. amount that this est. $\hat{\mu}$ differs from the actual value of μ.

It also tells us that this value would shrink with "n" increase i.e. more obs. less error.

Similarly $SE(\hat{\beta}_0)$ & $SE(\hat{\beta}_1)$

$$SE(\hat{\beta}_0)^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{n=1}^{n}(x_i - \bar{x})^2}\right]$$

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{n=1}^{n}(x_i - \bar{x})^2}$$

{ where $\sigma^2 = Var(\epsilon)$

$\boxed{\text{Assumption}}$:- errors ϵ_i for each obs. are uncorrelated w/ common variance.

Notice :- ① $SE(\hat{\beta}_1)^2$ is smaller if obs. are more spread out, i.e. we have more <u>leverage</u> to estimate the slope.

② $SE(\hat{\beta}_0)^2$ would be $= SE(\hat{\mu})$ if $\boxed{\bar{x} = 0}$

i.e $\hat{\beta}_0 = \bar{y}$

In general σ^2 is not known, thus can be estimate from the data & is known as RSE (Residual Standard Error)

$$RSE = \sqrt{\frac{RSS}{(n-2)}}$$

Standard errors can also be used to compute "Confidence Intervals"

Defn :- 95% CI means that with 95% prob., the range would contain the true unknown value.

For LR, 95% CI for β_1

$$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$$

i.e. 95% chance that the range.

$$[\hat{\beta}_1 - 2 \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot SE(\hat{\beta}_1)]$$

Similarly for β_0

$$\hat{\beta}_0 \pm 2 \cdot SE(\hat{\beta}_0)$$

Standard errors can also be used to create hypothesis tests. Eg: Null Hypothesis

Eg. $H_0 \Rightarrow$ No relation b/w X & Y

$H_1 \Rightarrow$ Some relation.

Mathematically $H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Since if $\beta_1 = 0$, the model reduces to $\Rightarrow \boxed{y = \beta_0 + \epsilon}$

$\therefore$ we basically need to test that β_1 should be greater than 0, Ques? How much greater? Need "t - statistic" for that.

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)},$$

measures the no. of standard deviations that $\hat{\beta}_1$ is away from zero.

The t- distribution is quite similar to bell curve. with $n > 30$,

Consequently, it is a simple matter to compute the probability of observing any value equal to $|t|$ or larger, assuming $\beta_1 = 0$ $\Rightarrow$ p- value.

Generally speaking $\leftrightarrow$ The interpretation of p-value.
a small p-value indicates that it is unlikely to observe such a substantial association b/w the predictor and the response due to chance, in the absence of any real association b/w the predictor & the response.

ASSESSING THE ACCURACY OF THE MODEL

"Once we have rejected the null hypothesis in favour of the alt., it is natural to want to quantify the extent to which the model fits the data

The quality of the linear regression fit is typically assessed using two related quantities :- ① $\boxed{RSE}$ & ② R^2

RSE : ① Estimate of the SD of ϵ
[what is a good RSE?] ② Avg. Amt. that the response will deviate the true regression line.

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

③ Considered as a measure of lack of fit

R^2 : ① Alternative measure of fit that removes the problem of what is a good RSE?

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

([0, 1])

$$\begin{cases} TSS = \sum (y_i - \bar{y})^2 \\ RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \end{cases}$$

→ Can be thought of as the amount of variability in the response variable Y before regression is performed.

→ Can be thought as the variability that is left after the regression.

∴ R^2 is thus the variability explained through regression

R^2 statistic is a linear relationship b/w X & Y. Recalling correlation, which is also defined similarly =

$$Cor(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

This suggests that we might be able to use $r = Cor(X, Y)$ instead of R^2 to assess the fit of the model.

In fact, for the simple linear regression with one variable, we can show that $R^2 = r^2$,

in other words the squared correlation & R^2 statistic are ~~similar~~ identical.

BUT in multiple linear regression, we can't use correlation 'cuz by defn. it only quantifies the relationship b/w a pair of variables & thus R^2 fills that role perfectly.

MULTIPLE LINEAR REGRESSION

Extending simple linear regression by giving each ~~coefficient~~ predictor a separate ~~term~~ slope coefficient

$$\Rightarrow Y = \beta_0 + \beta_1 X + \beta_2 X_2 + \dots \beta_p X_p + \epsilon$$

β_j is the average effect on Y of one unit increase in X_j, holding all other predictors fixed.

Eg sales $= \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times newspaper + \epsilon$

ESTIMATING THE REGRESSION COEFFICIENTS

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots \hat{\beta}_p x_p$$

Eg :- In a 3D with two predictors X_1 & X_2 we have to find the plane (regression line becomes regression plane), & we min. the sum of squared vertical distance between each obs. and the plane.

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

$$= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots \hat{\beta}_p x_{ip})^2$$

The values $(\hat{\beta}_0, \hat{\beta}_1, \dots \hat{\beta}_p)$ that minimize the eqn are the multiple least squares regression coefficient estimates.

FUNFACT : Running a regression of shark attacks versus ice cream sales for data collected at a given beach community over a period of time would show a positive relationship

In reality higher temp causes people to come to beaches in higher nos. & have ice creams as well.

Some Imp Ques.

1. Is atleast one of the $X_1, X_2 \cdots X_p$ useful?
2. Do all fit or only a subset?
3. How well does it fit the data.
4. Given a set of predictor values, what response should we predict, and how accurate is the prediction?

Ans 1 $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$.

H_a: at least one $\beta_j \neq 0$

This hypothesis test is performed using the F-statistic:

$$F = \frac{\frac{(TSS - RSS)}{p}}{(RSS/(n-p-1)}$$

{ if the linear model assumptions are true, then $\frac{RSS}{(n-p-1)} = \sigma^2$

and if H_0 is true, then $\frac{(TSS - RSS)}{p} = \sigma^2$

Hence we could conclude that if F is close to one, then no relationship b/w predictors but if far greater than 1 then [H_a is true]

Ans ②.

Variable Selection :- The task of determining which predictors are associated with the response, in order to fit a single model involving only those predictors is known as variable selection.

Which model is the best? AIC, BIC, Mallow's Cp. and adjusted R^2

Unfortunately there will be 2^p models.
Eg: if 30 predictors then 2^{30} models to test :-

Three of the better approaches are listed below:-

① Forward Selection :- We start with the null model and add a predictor that results in the lowest RSS. We then select a variable/predictor that results in the lowest RSS for the new 2 variable model. Continue until some threshold reached.

② Backward Selection :- We start with the full model and start removing predictors with highest p value until all variables have a p-value below some threshold.

③ Mixed Selection :- We start with null model, keep adding variables that best fit but we dont add variables that have p value above a certain threshold.

Ans ③

MODEL FIT: RSE & R^2 } fraction of Variance explained.

In simple linear regression,

$$R^2 = Cor(X, Y)^2 =$$

in MLR, $R^2 = Cor(Y, \hat{Y})^2$
explains the property that fitted linear model maximises the correlation among all possible linear models.

It turns out that R^2 would always increase with more predictor variables, even if there is a weak relationship.

Eg. sales ~ TV, newspaper, Radio

$\Downarrow$

$R^2 = 0.8972$

sales ~ TV, Radio

$\hookrightarrow R^2 = 0.8719$

cuz newspaper & radio are correlated & its p value is not significant.

If we dont remove it, we might lead the model to overfit.

$$RSE = \sqrt{\frac{RSS}{(n-p-1)}}$$

Ans 4

PREDICTIONS & ACCURACY of such pred.

3 uncertainties :-

1. CONFIDENCE intervals for removing the uncertainty behind closeness of y & $f(x)$.
2. Model Bias, ignore.
3. Irreducible error, the error term, we use prediction intervals for that $\Rightarrow$ will be wider than CI as they incorporate uncertainty behind $f(x)$ & individual point difference.

Qualitative predictors with more than two levels.

$$x_{i1} = \begin{cases} 1 & \text{if Asian} \\ 0 & \text{if not asian} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if Caucasian.} \\ 0 & \text{if not.} \end{cases}$$

& Baseline African American.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{Caucasian} \\ \beta_0 + \epsilon_i & \text{African American} \end{cases}$$

There will always be one fewer dummy variables than the no. of levels.

Imagine if p-values are large then there wont be any real difference b/w ethnicities for our use case {eg credit card debt balance}

EXTENTIONS OF LINEAR MODEL

2 Assumptions:

① ADDITIVE Assumption :- This states that the effect of change of X_j on Y is independent of any change of other predictors.

Solve Relaxing this assumption
⇒ Introduce an interaction term :-

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 * X_2)$$

$$= \beta_0 + \tilde{\beta}(X_1) + \beta_2 X_2$$

$$\{ \tilde{\beta} = (\beta_1 + \beta_3 X_2)$$

This means change in X_1 is no longer independent.

Eg :- sales ~ $\beta_0 + \beta_1 TV + \beta_2 radio + \beta_3 (TV \times radio)$

② NON/LINEAR RELATIONSHIP

We can directly extend linear ~~relationship~~ regression by including transformed versions of predictor variables.

Eg : mpg ~ $\beta_0 + \beta_1 \times horsepower + \beta_2 \times HP^2 + \epsilon$

This method is also known as polynomial regression

POTENTIAL PROBLEMS.

① **Non Linearity of the data** :-
Residual plots i.e $e_i = y_i - \hat{y}_i$ versus the predictor x_i can be used to identify non linearity. {Solve: use polyregression}

② **Correlation of Error terms** {We assume uncorrelation}
The standard errors (SE) that are computed for the estimated regression coefficients or the fitted values are based on the assumption of uncorrelated error terms.

③ **Non Constant variance of error terms** :-
Another imp. assumption of LR is constant variance of error terms $Var(\epsilon_i) = \sigma^2$. CI, SE, & hypothesis tests rely on this assumption.

Solve :- Transform "Y" response variable through concave functions like $\log(Y)$ or $\sqrt{Y}$.

Identify :- It is known as ~~heteroscedasity~~ heteroscedasity i.e. non constant variance, by plotting Residuals. If a funnel shape appears then apply such transformations

④ **Outliers**
In practice, it can be difficult to detect outliers, since how far is far enough?

Solve :- STUDENTIZED RESIDUALS ~~computed~~ (SR)

$$= \frac{e_i}{SE(e_i)}$$

If $||SR|| > 3$ then the point can be removed.

5 High Leverage points :-

SOLVE: LEVERAGE STATISTIC

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$ {For Simple LR

It is clear that h_i increases with x_i's distance from its sample mean ($\bar{x}$) thus h_i (high) can be removed from the observations.

For M.L.R, the formula can be checked on wiki,

but $\frac{1}{n} \leq h_i \leq 1$

and average $h_i = \frac{(p+1)}{n}$

Thus if h_i greatly exceeds $\frac{(p+1)}{n}$ then we can think of removing those observations.

6 COLLINEARITY :-

When two or more predictors are more closely related to one another.

Correlation matrix ⇒ to detect a pair of correlated variables. But sometimes it is tough to find correlation b/w 3 or more when no pair in particular has any correlation. This is also known as multicollinearity.

SOLVE: Variance Inflation Factor (VIF)

$$VIF(\hat{B}_j) = \frac{1}{1 - R^2_{(X_j | X_{-j})}}$$

{$R^2_{X_j | X_{-j}}$ is R^2 from a reg. of X_j onto all other predictors

If $R^2_{x_j|x_{-j}}$ is close to one, & VIF will be large then colinarity is present.

⇒ a VIF of greater than 5 or 10 can be problematic.

Solutions to colinear variables :-

1. Drop out.
2. Avg. Standardized versions
   Eg. combine "limit" and "rating" to create "credit worthness"

The Marketing plan.

① Is there a relationship b/w advertising sales & budget?

- MLR : Sales ~ TV, radio, newspaper.
- H_0 : $\beta_{TV} = \beta_{radio} = \beta_{newspaper} = 0$.
- F Statistic & p values.

② How strong is the relationship.
Model Accuracy
- RSE : est. the (σ) of the response from the reg. line.
- R^2 : % variability explained by the predictors.

③ Which media contributes to sales?
- p values of different predictors' coefficients.

④ How large is the effect of each medium on sales.
- $SE(\hat{\beta}_j) \Rightarrow$ can be used to construct CI for β_j
- colinearity & (VIF).

⑤ How accurately can we predict sales?
- indiv. respons.

$$Y = f(X) + \epsilon.$$

$\Downarrow$ prediction interval. (wider as they include irreducible error).

- $f(X)$ avg. response.

$\Downarrow$ Confidence interval.

⑥ Is the relationship linear?
- transformations of predictors
- " of response.

⑦ Synergy? $\Rightarrow$ interaction terms.