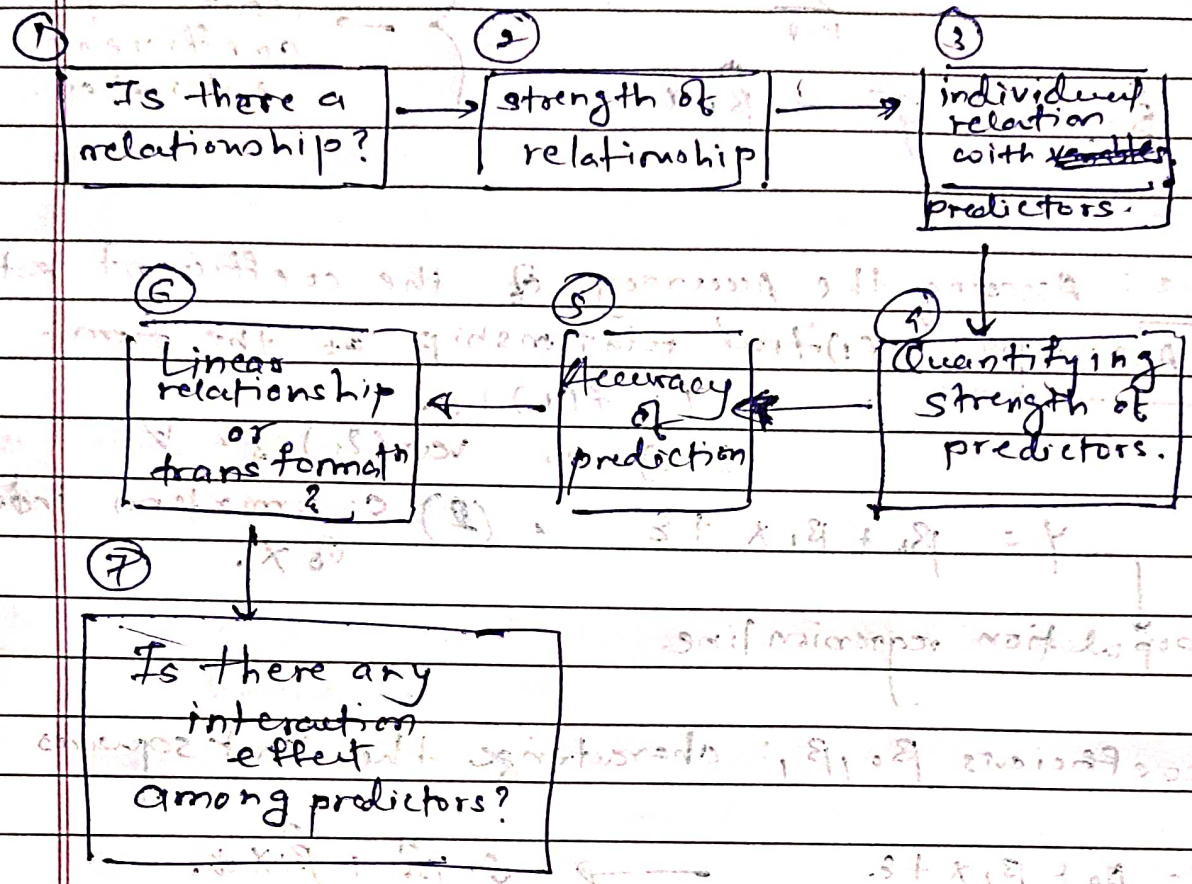
Chapter 3: Linear Regression.

- Tool for predicting quantitative response.
- very simple approach.



3.1. Simple Linear Regression.

- predicting a quantitative response Y on the basis of a single predictor variable X.

$$y \approx \beta_0 + \beta_1 X.$$

β_0 (intercept), β_1 (slope): model coefficients or parameters.

β_1 → average increase in y associated with one unit increase in X.

prediction: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$e_i = y_i - \hat{y}_i$: ith residual

Residual Sum of Squares:

$$RSS = \sum_{i=1}^{n} e_i^2$$

$$= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \longrightarrow \text{minimizing wrt } \beta_0, \beta_1$$

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

} → least square coefficient estimates.

3.1.2 : Assessing the Accuracy of the coefficient estimates.

Assumptn : (1) true relationship of the form :

$$Y = f(x) + \varepsilon.$$

$$E(\varepsilon) = 0. \quad \text{var}(\varepsilon_i) = \sigma^2 \ \forall i$$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$, (2) ε: error term independent of X.

↳ population regression line

coefficients β_0, β_1: characterize the least squares line.

$Y = \beta_0 + \beta_1 X + \varepsilon$. ⟶ $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

population regression line. estimates from least squares lines.

$\hat{\beta}_0$ & $\hat{\beta}_1$ unbiased for β_0 & β_1 respectively ⟶ close to true population regression line.

standard error of $\hat{T}$: the avg amount that this $\hat{T}$ will differ from T.

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

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

where $\sigma^2 = \text{Var}(\varepsilon)$

- Residual standard error:

$$RSE = \sqrt{\frac{RSS}{n-2}} = \hat{\sigma}$$

* Meaning of 95% confidence interval:
range of values s.t. with 95% probability, the range will contain the true unknown value of the parameter.
(using assumption $\varepsilon \sim N(0, \sigma^2)$)

* 95% CI for β_1 : $[\hat{\beta_1} - 2SE(\hat{\beta_1}), \hat{\beta_1} + 2SE(\hat{\beta_1})]$

95% CI for β_0 : $[\hat{\beta_0} - 2SE(\hat{\beta_0}), \hat{\beta_0} + 2SE(\hat{\beta_0})]$

H_0: There is no relationship i.e $\beta_1 = 0$
between X & Y

H_1: There is some relationship i.e $\beta_1 \neq 0$.
between X & Y

Under H_0:

$$t = \frac{\hat{\beta_1} - 0}{SE(\hat{\beta_1})} \sim t_{n-2}.$$

3.1.3: Assessing the accuracy of the model.

To quantify: The extent to which the model fits the data.

(1) Residual Standard Error.
- estimate of the standard deviation of ε.

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

$$RSE = \sqrt{\frac{RSS}{n-2}}$$ → measures the lack of fit. (absolute measure).

(2) R^2 statistic $\in [0,1]$.

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

For simple linear regression
$R^2 = \rho^2 = [\text{for}(X,Y)]^2$

TSS:= total sum of squares $\sum_{i=1}^{n} (y_i - \bar{y})^2$

R^2: proportion of variability in Y that can be explained using X.

3.2 : Multiple linear regression.

Doubt Shark attacks & ice cream sale example: Is temperature a latent variable? What if it is some unknown problem.

Assumption of functional form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

estimates:

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

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \xrightarrow{\text{minimizing wrt } \beta_j} \text{estimates } \hat{\beta}_j$$

Important questions:

(1) Is there a relationship between response & predictors?

H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$

H_1: at least one β_j is non zero.

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

where? $TSS = \sum (y_i - \bar{y})^2$

$$RSS = \sum (y_i - \hat{y}_i)^2 .$$

$$E\left[\frac{RSS}{n-p-1}\right] = \sigma^2 .$$

Under H_0: $$E\left[\frac{TSS - RSS}{p}\right] = \sigma^2 .$$

→ value of F statistics depends on n and p.

* To test whether particular of the coefficients (q) are zero.

H_0: $\beta_{p-q+1} = \beta_{p-q+2} = \dots = \beta_p = 0$.

$$\rightarrow F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$$

RSS: RSS for whole model

RSS_0: RSS for model which excludes q coefficients.

- F statistics test: p is relatively small.
- If $p > n$: forward selection or high dimensional models.

(2) Deciding on Important variables.

- classical approaches to variable selection.
- For p features: 2^p models to be considered.

Forward Selection

null model (only intercept)
↓
fit 'p' simple linear regressions.
↓
add the predictor with least RSS.
↓
two var model
↓
add the projector with least RSS
↓
untill some stopping criterion is satisfied.

→ can be always used
→ might include variables early that later become redundant. → mixed selection.

Backward Selection

All variables in model
↓
remove var with largest p value.
↓
fit for p-1
↓
remove least significant.
↓
untill some stopping criterion is reached.
→ if $p > n$ cannot be used.

*** Mixed Selection**

- Combination of forward & backward.
- start with null model
- go on adding variables, one by one which provide best fit.
- if p values become larger, remove that variable.
- untill all p values are below certain threshold.

Doubt What exactly is the difference between confidence interval & prediction interval?

(3) Model fit: How well does the model fit the data?

$R^2 = \{Cor(Y, \hat{Y})\}^2$.

- ~~For~~ MLR maximizes $\{Cor(Y, \hat{Y})\}^2$ among all possible linear models.

$$RSE = \sqrt{\frac{RSS}{n-p-1}} \quad (p=1 : \text{simple linear regression})$$

(4) Prediction: Given a set of predictor values, what response value should we predict and how accurate is our prediction?

Three uncertainties involved:

(1) $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots \hat{\beta}_p X_p$ only estimate for the true population regression plane:

$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.

- inaccuracy in estimates - related to reducible error
- compute CI for $\hat{Y}$.

(2) model bias.

(3) Using prediction intervals to find how much Y varies from $\hat{Y}$?

- they are wider than CI.
- prediction interval → reducible error: error in estimate of $f(X)$
- prediction interval → Irreducible error: how much an individual point will differ from the population regression plane.

3.3. Other considerations in the regression model

3.3.1 Qualitative Predictors.

(1) predictors with only two levels.

ex: Credit data set:

→ A factor with two levels → $x_i = \begin{cases} 1, & \text{if } i\text{th person owns a house} \\ 0, & \text{if } i\text{th person does not own a house} \end{cases}$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{— if } i^{th} \text{ person owns a house} \\ \beta_0 + \varepsilon_i & \text{— if } i^{th} \text{ person does not own a house.} \end{cases}$$

interpretation:

β_0: avg. credit card balance who do not own.

$\beta_0 + \beta_1$: avg credit card balance among those who do own their house

β_1: avg diff in credict card balance between owners & non owners.

→ interpretations changes accordingly with definition of dummy variable. (coded as -1 or 1, interpretation will change, but values will remain same).

(2) Qualitative predictors with more than two levels.

x → qualitative variable { : } m levels.

We consider m-1 dummy variables:

$$x_{i1} = \begin{cases} 1 & \text{→ } i^{th} \text{ obs. } \in 1^{st} \text{ level} \\ 0 & \not\in 1^{st} \text{ level} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & ; \ i^{th} \text{ obs } \in 2^{nd} \text{ level} \\ 0 & ; \ \not\in 2^{nd} \text{ level} \end{cases}$$

$$\vdots$$

$$x_{i,m-1} = \begin{cases} 1 & ; \ i^{th} \text{ obs } \in (m-1)^{th} \text{ level} \\ 0 & ; \ \not\in (m-1)^{th} \text{ level} \end{cases}$$

The level with no dummy variable is known as baseline.

For $m = 3$.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & , \in I^{st} \text{ level} \\ \beta_0 + \beta_2 + \varepsilon_i & ; \in II^{nd} \text{ level} \\ \beta_0 + \varepsilon_i & ; \in III^{rd} \text{ level.} \end{cases}$$

3.3.2 Extensions of the linear model:

- Two restrictive assumptions that are violated often:

(i) additivity:

Association betⁿ ~~predi~~ X_j & Y does not depend on the values of other predictors.

(ii) linearity:

Functional form of: population regression ~~line~~ is linear - change in Y associated with a one-unit change in X_j is constant, regardless of the value of X_j.

(1) Removing additive assumption:

Considering interaction term:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

$\beta_3 X_1 X_2$ → interact° term

$$\Rightarrow Y = \beta_0 + \beta_1 X_1 + \beta_3 X_1 X_2 + \beta_2 X_2 + \varepsilon$$

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

$$= \beta_0 + \tilde{\beta_1} X_1 + \beta_2 X_2 + \varepsilon$$

$\tilde{\beta_1} = \beta_1 + \beta_3 X_2$ ⇒ $\tilde{\beta_1}$ is function of X_2.

⇒ change in value of X_2 will change the association between X_1 and Y.

How to know which variables are interacting?

- The hierarchical principle:

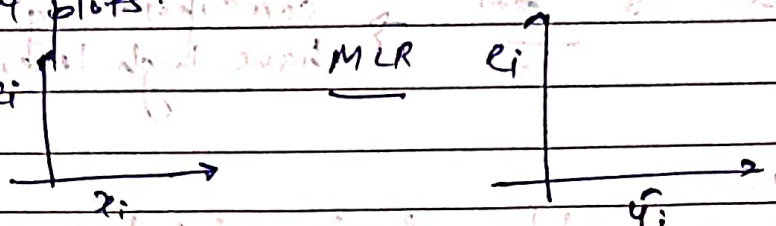
If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.

- for nonlinear relationships: polynomial regression.

3.3.3 : Potential problems.

(1) non-linearity of the response-predictor relationships
- residual plots:

SLR : e_i vs x_i ; MLR e_i vs $\hat{y}_i$



> ideally residual plot should not show any pattern, if present, indicates serious problem with some aspect of linear model.
→ take transformations such as $\log x$, $\sqrt{x}$, x^2.

(2) Correlation of error terms
- If error terms are correlated : CI will be narrower
- funnel shape : $\sqrt{y}$.
  U shape : $\log y$
  } (3) heteroscedasticity.
- weighted least squares.

~~(3) heteroscedasticity.~~

(4) Outliers.
- a point for which y_i is far from the value predicted by the model.

~~Ques~~ what is a studentized residual & why we need it?

$$\text{studentized residual} = \frac{e_i}{\sqrt{SE(e_i)}}$$

(5) High leverage points.
- points with unusual value of x_i
- difficult to identify in multiple linear regression.

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

$$\frac{1}{n} \leq h_i \leq 1, \quad \text{average} = \frac{p+1}{n}.$$

→ so if a given observation has a leverage statistic that greatly exceeds $\frac{p+1}{n}$ ⇒ corresponding point might have high leverage.

(6) Collinearity:

- two or more variables closely related to one another.
- ~~multiple~~ for multiple estimates, RSS is minimum.
- CI for $\hat{\beta}$ widens.
- t statistics significantly decreases ⇒ p value ↑
  ⇒ i.e. fail to reject H_0.
  ⇒ masks importance of some variable.

*How do we decide whether CI is 'too wide' or not?

- VIF : variance inflation factor. ← measure?

VIF = 1 ⇒ complete absence of collinearity.

* VIF > 5 or VIF > 10 ⇒ problematic amount of collinearity.

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

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

If $R^2_{X_j | X_{-j}} \approx 1$ ⇒ collinearity present, large VIF.

Solutions:

(1) Drop one of the problematic variable

(2) Combine the collinear variables together into a single predictor.

How to exactly combine?

*How to decide if we have to add interaction terms or not? Which ones to ~~add~~ consider?

3.5 Comparison of Linear Regression with K nearest neighbors.

- linear regression is example of parametric.
- if specified functional form is far from truth, prediction accuracy will be very less.

Do all non parametric methods donot have a loss function optimization problem inside them? (like KNN).

- non parametric methods : K Nearest Neighbors. (KNN regression)

Algorithm:

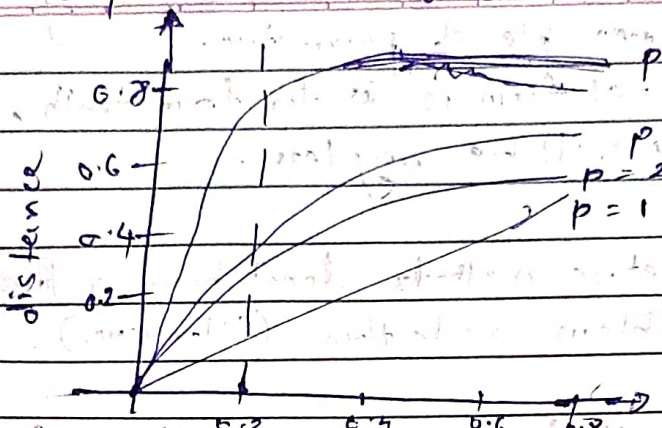
(1) K, prediction point x_0.
(2) identify K training observations that are closest to x_0 (represented by N_0)
(3) $\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in N_0} y_i$

- optimum value of k depends on bias-variance trade off.
small k :- low bias high variance.
large k :- smoother & less variable.
↳ smoothing may cause bias by masking some of the structure in f(x)
- The parametric approach will outperform the non-parametric approach if the parametric form that has been selected is close to the true form of f.

How to know for which values K alt is not overfitting? (for k=1, it is definitely overfit).

- performance of KNN decreases as no of features increase.
- Curse of dimensionality. K observations that are nearest to a given test observation x_0 may be very far away from x_0 in p dim space when p is large.
↳ poor prediction of $f(x_0)$ in KNN.

e.g (ESL) In 10 dimentions we need to cover 80 % of the range of each cordinate to capture 16 % of data.



(Figure 2.6: ESL).

- As a general rule : parametric methods will tend to out perform nonparametric approaches when there is small no. of observations per predictor.