A Regression is used to understand quality, cause - effect relationship.

8. Predictive Analytics

8.1 Linear Regression

* Simple Linear Regression

#what happens to sales of a brand of shampoo when discount of 15% is offered in particular week.

We expect sales to go up

Here

The Cause: reduction in price

effect: increase in sales

Also if I want to know by how much the sales has increased?

That is the quantification of the impact

* Regression analysis is a statistical technique used to infer the magnitude and direction of a possible ~~casual~~ causal relationship between an observed pattern and variables assumed to have an impact on the observed pattern.

Statistical -

A mathematical approach that assumes that the pattern of interest and the variables that impact the pattern are all random samples from underlying population.

Magnitude – Size of impact (2 times, 10 times)

Direction – Positive or Negative

Causal –

Magnitude of rainfall has impact on crop yield, but crop yield does not influence rainfall

Observed Pattern – Dependent variable, Distribution

* ~~Simple Linear Regression~~

eg. You work for a hospital and are looking to understand factors that may influence the birth weight of a baby

→ Maternal Diet

→ Gestation period

→ Maternal Health Issues

→ Ethnicity

→ Age of the mother

Sample Data

Birth Weight	Weeks of Gestation	Years of Educath	Race	Smoked During Pregancy

Sam

What are possible ways to accessing these relationships?

- Graphical Visualization
- Correlations
- Run Regression Model

A better way to identify ther relationship between these variables is to use a regression technique.

Why Regression?

- multiple factor impact on the effect
- Statistical Significance of the impact

* Simple Linear Regression

we are trying to find relationship between baby birth weight and gestation period

Birthweight = f (gestation weeks)

$$y = mx + c$$

m ⇒ slope
c ⇒ intercept

Slope ⇒ rate of change of y when x changes

Intercept ⇒ value of y when $x = 0$

$Y = \beta_0 + \beta_1 x + e$

where $\beta_0 \Rightarrow$ Intercept

$\beta_1 \Rightarrow$ Slope

$e \Rightarrow$ Error

Terminology

Dependent Variable / Predicted / Target Variable

Independent Variable / Predictor / X

Beta coefficient(s):

The estimate of magnitude of impact of changes in the predictors on the predicted variable.

Error:

The impact of the unobserved variables on the dependent variable, usually calculated as the difference between the predicted value of y given the estimated regression function and the actual value of y.

The ordinary least squares Regression (OLS) technique estimates coefficients on the variables hypothesized to have an impact on the variable of interest by identifying the line that minimizes the sum of squared differences between points on the estimated line and on the actual values of the independent variable.

- Coefficients : Betas

Minimizes : Least

Sum of Squared Differences : Square of Residuals

Estimated Line : Regression Line

Actual Values : values in dataset

OLS Estimates

The OLS Regression find that line by looking at the residuals (or the difference between the points on each line and ~~it~~ actual Y) and minimizing the sum of their squares.

$$\varphi = \sum_{i=1}^{N} (y_i - b_0 - b_1 x_i)^2$$

Using differential calculus we get

$$b_0 = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

$$b_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

How to measure Model fit? (Validating the model)

→ R^2

→ Fit Chart - Actual v/s Fitted values

→ MAPE - Mean Absolute Percentage Error

* Higher the R^2 better the model

Regression Assumptions

1. Model is linear in parameters
2. The data is a random sample of population
   - The errors are statistically independent from one another
3. The expected value of the errors is always zero.
4. The independent variables are not too strongly collinear
5. The independent variables are measured precisely

6. The residuals have constant variance

7. The errors are normally distributed

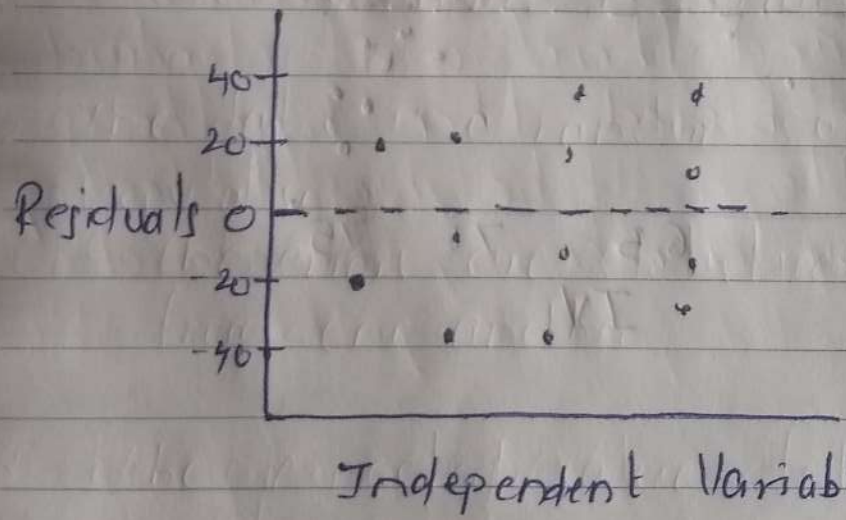
8. The model is correctly specified

~~The~~ If all the these conditions hold, then OLS estimates are BLUE - Best Linear Unbiased Estimators

Checking If Assumptions are Valid

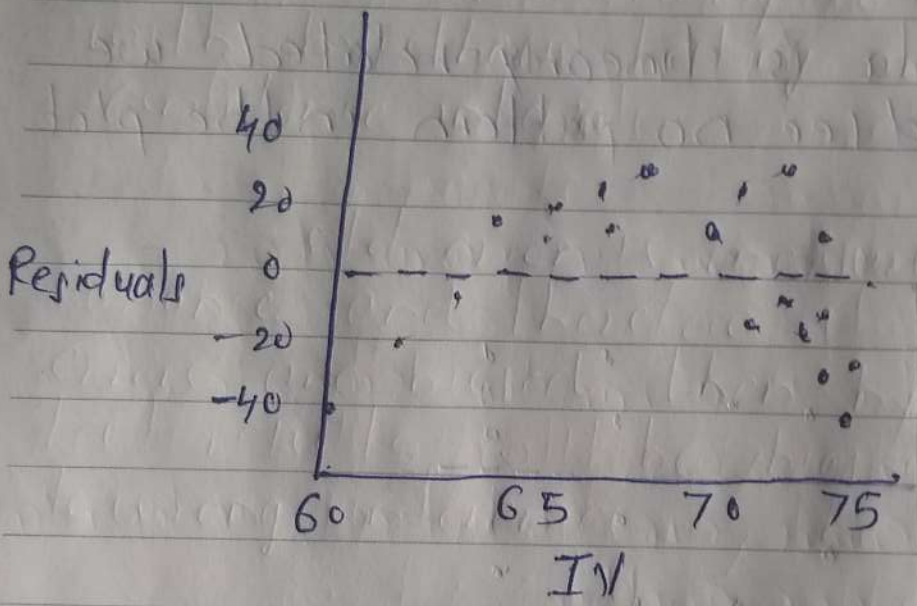
1) check for Linearity - Plot the residuals against each IV

- If data is linearly related, we should see no pattern in the plot



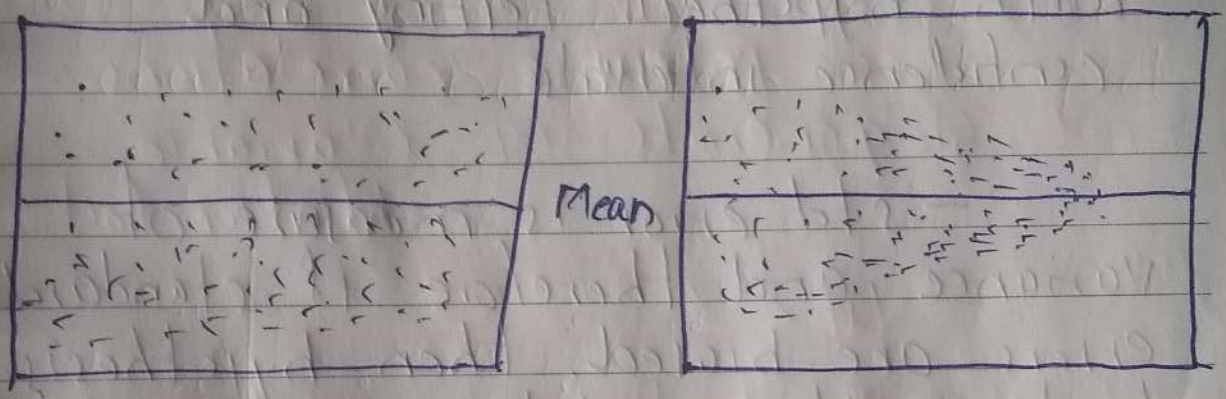
If the relationship is non-linear?

we will see some pattern in the plot



2. The residuals should have constant variance → homoscedasticity

Plot the residuals ~~against~~ against predicted Y



Homoscedasticity

No pattern in data
Variance is constant

Heteroscedasticity

we can see pattern in the data

→ The presense of heteroscedasticity does not imply bias in the estimates

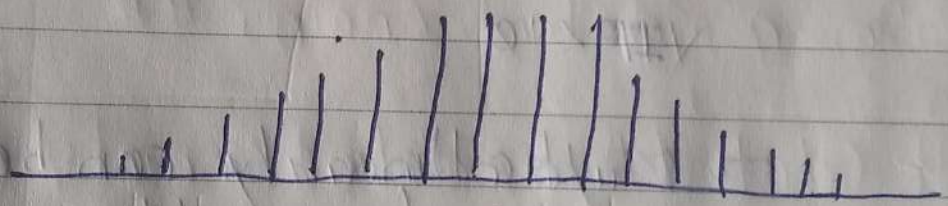
→ Heteroscedasticity leads to bias in the standard errors, leading to issues with hypothesis testing and confidence intervals

- std error is measure of variance, and therefore if standard errors are biased, then hypothesis test results will be biased leading to wrong inferences

3. The residuals are normally distributed

Histogram or probability plot for residuals

Distribution of errors



If the residuals are not normally distributed?

Hypothesis test outcomes may be invalid, though less of an issue with large samples

4. The IVs are not too correlated - multicollinearity

- The IVs should not be highly correlated to one another (<90% and >0)

- check pairwise correlations or generate VIF (Variance Inflation factor)

VIF < 10

- Multicollinearity can be reduced by Transforming the variable or by combining the variable

R Code

Age — Old / Middle / Young

Gender — Male / Female

Own Home — Own / Rent

Married — Single / Married

Location — Far / Close

Salary —

Children — 0 / 1 / 2 / 3

History — High / Low / NA / Medium
no. of purchase

Catalogs — 6 to 24

Amount Spent —

```
setwd (" ")

data <- read.csv ("DirectMarketing.csv")

library (dplyr)
library (ggplot2)
library ( car)

head (data)

## Do exploratory Analysis ##

plot (data$Age , data$Amountspent, col="red")

# middle & old Age have similar spent

# Combine middle & old levels together

data$Age1 <- ifelse (data$ Age == "Young",
                     "middle-old"; as.character (data$Age))

data$ Age 1 <- as.factor (data$Age1)

summary (data$Age1)
```

```
plot (data$Age1, data$AmountSpent)

# Gender

plot (data$Gender, data$Amountspent)

# own Home

Summary (data$ ownHome)

plot (data$ ownHome, data$Amount Spent,
      col = "red")

# Married

summary (data $ Married)

plot ( data$ Married, data$ Amountspent,
      col = "red")
```

```
# Location

summary (data$Location)

plot (data$Location, data$AmountSpent,
      col = "red")


# Salary

summary (data$Salary)

plot (data$Salary, data$AmountSpent)

# Might be heteroscedasticity


# children

summary (data$children)

data$children <- as.factor (data$children)

plot (data$children, data$Amount-spent,
      col = "red")
```

```
data$children1
<- ifelse(data$children==3 | data$children==2,
   "3-2", as.character(data$children))
data$children1 <- as.factor(data$children1)

summary(data$children1)

plot(Data$children1, Data$Amountspent,
   col = "red")

# History

summary(data$History)

# Impute missing values

tapply(data$Amountspent, data$History,
   mean)

ind <- which(is.na(data$History))

mean(data[ind, "Amountspent"])
```

```
data %>% filter(History == "Medium") %>%
select(AmountSpent) -> Amt_M

p <- ggplot(data[ind, ], aes(x = AmountSpent))

q <- ggplot(Amt_M, aes(x = AmountSpent))

p + geom_histogram()

q + geom_histogram()

# Create a category called missing

data$History1 <- ifelse(is.na(data$History),
                        "Missing", as.factor(data$History))
data$History1 <- as.factor(data$History1)

summary(data$History1)
```

```
data$History1 <- factor(data$History1,
                        labels = c("High", "Low", "Medium",
                                   "Missing"))

# catalogues
summary(data$Catalogs)
data1 <- data[, -c(1,7,8)]

mod1 <- lm(AmountSpent ~., data = data1)

summary(mod1)

mod2
<- lm(formula = AmountSpent ~ Gender + Location
      + Salary + Catalogs + Children1 + History1,
      data = data1)

summary(mod2)
```

```
summary (data1)

#Remove insignificent variables
#History Missing
#Gender male

#Create Dummy Variables

data$ Male_d <- ifelse (data1$Gender == "Male",
                        1, 0)

data$ Male_f <- ifelse(data1$Gender == "Female",
                       1, 0)

data1$Missing_d <- ifelse (data$History1 ==
                           "Missing", 1, 0)

data$ Missing_

data$ Low_d <- ifelse (data$History1 == "Low",
                       1, 0)

data$ Mid_d <- ifelse (data$History1 == "Medium",
                       1, 0)
```

```
data$High_d <- ifelse(data$History1 == "High", 1, 0)

mod3 <- lm(formula = AmountSpent ~ Male_d + Location + Salary + Catalogs + Children1 + Med_d + Low_d, data = data1)

summary(mod3)

mod4
<- lm(Formula = AmountSpent ~ Location + salary + Catalogs + Children1 + Med_d + ~~Med~~Low_d, data = data1)

summary(mod4)

# Signs
```

```
tapply (data$Amount Spent, data$History,
        mean)

data1 %>% filter(History1 != "Medium",
History1 != "Low") %>% summarize
(Mean = mean(AmountSpent)) # InLine

tapply (data1$AmountSpent, data1$Location,
        mean)      # Inline

# Assumption checks

hist (mod4$residuals)

qqplot (mod4$ residuals)

# Non normal behaviour observed

45° line denotes the behaviour if
data has been normal
```

Multicollinearity check

vif (mod4)

Constant variance check

plot4 (mod4$ fitted.values, mod4$ residuals) # funnel shape

~~Remedies~~ Remedies : Apply log transform to y variable

mod5
<- lm (formula = log (AmountSpent) ~ Location + Salary + catalogs + children + Med_d + Low_d, data = data1)

summary (mod5)

qqplot (mod5$ residuals) # looks okay

plot (mod5$ fitted values, mod5$ residuals)
still funnel

Apply square root transform

mod6
<-lm (formula = sqrt(Amount.spent) ~ Location + salary + catalogs + Children, 1. + med_d + Low_d, data = data1)

summary (mod 6)

qqplot (mod6 $ residuals)

plot (mod6 $ fitted values, mod 6 $ residuals)

seems okay

vif (mod6)

I can also try cube root transformation or Normal Logarithmic Transformation

```
predicted <- mod6$ fitted.values

actual <- sqrt (data1 $ Amount spent)

data <- data.frame (predicted, actual)

P <- ggplot (dat, aes (x = row (dat)[,2],
     y = predicted))

P + geom_line ((colour = "blue")) + geom_line (data = dat,
  aes (y = actual), colour = "black")
```