1 Chapter 1: Introduction

Sample

Subset of population which we have data on

$\Downarrow$

Descriptive Statistics

Summary Statistics of Sample Data

$\Downarrow$

Used to make *Inferences*

$\Downarrow$

Inferential Statistics

Parameters of Population

(Unknown)

$\Downarrow$

Population

Total set of subjects of interest

2 Chapter 2: Single Variable Analyses

Types of Variables

Nominal (Cate)

Meaningful Order $\Downarrow$

Ordinal (Cate)

Consistent Difference $\Downarrow$

Discrete (Quant)

(Uncountably) Infinite $\Downarrow$

Continuous (Quant)

Describing Data

Continuous Variables:

- Cluster/Gap Intervals
- Suspected Outliers [Small/Large] ($z-score > 3$)
- Modality
- Skewness
  1. Symmetric & Bell-Shaped: (Sensitive to Skew)
     - Mean
     - Var & Std Dev
  2. Highly Skewed: (Robust to Outliers)
     - Median
     - IQR (NOTE: Quantiles are non-unique)

$\Rightarrow$ Continuous can be categorised!

Categorical (or Discrete) Variables:

- % Proportion of Modal Category
- Special High/Low Categories
- (Ordinal) Apparent Trend in Proportions

Formulae and Results

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

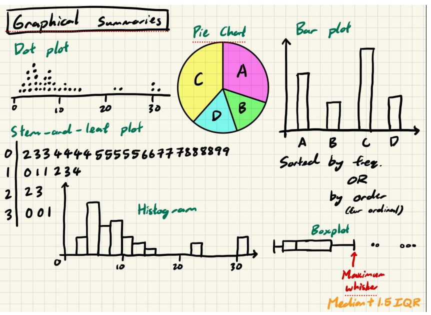
$Var(X) = \frac{1}{n-1}\sum_{i-1}^{n}(X_i - \overline{X})^2$

For $Y = aX + b$:

$\overline{Y} = a\overline{X} + b$

$Var(Y) = a^2 Var(X)$

$s_Y = |a|\, s_X$



3 Association between Variables

Categorical x Categorical

Association:

- Difference in Proportions
- Relative Risk $\frac{Occurrence\ in\ Exposed}{Occurrence\ in\ Unexposed}$

Graphical Summaries:

- Contingency Table
  Conditional Proportions on [Explanatory] for [Response]
  Columns: Response, Rows: Explanatory
  $\sum$ proportions across row $= 1$
  $Marginal\ Proportion = \frac{+\,Response}{Total\ Data}$
- Bar Charts (Stacked/Clustered)

Categorical x Quantitative

Association:

- Medians
- Skewness, Spread
- Outliers [Small/Large]

Graphical Summary: Side-by-side Boxplots

Quantitative x Quantitative

Association:

- Presence of Association [+/-]
- Type of Association [Linear/Non-Linear]
- Variance of Points
- Unusual Departure from Overall Trend
  Non-Constant Variance, Suspected Outliers
- Correlation Coefficient (Linear only)

$$R = \frac{1}{n-1}\sum\left(\frac{X_i - \overline{X}}{s_X}\right)\left(\frac{Y_i - \overline{Y}}{s_Y}\right)$$

Interpretation:

$\lvert R\rvert >= 0.8$	Very Strong
$0.5 < \lvert R\rvert < 0.8$	Strong
$\lvert R\rvert <= 0.5$	Not Strong

Correlation != Causation$\Downarrow$

Lurking and Confounding Variables

Lurking Variables: Unobserved variable that influences association between variables of interest. Has potential for confounding.

Confounding Variables: Explanatory variables which are associated with response, but also to each other. Condition by confounding variable

4 Study Design

Observational Studies

Case-Control Studies: Split by response $\Rightarrow$ What was done differently in the PAST? (Retrospective)

Sample Surveys: What does the population look like NOW? (Cross-Sectional)

Cohort Studies: Identify a group now $\Rightarrow$ Observe in the future. (Prospective)

Pros:

- Ethical and Easier to conduct
- Availability of Data
- Causality is not always required information

Cons:

- Not possible to establish cause and effect
- Lurking Variables can affect the results

Consistent concusion of observational studies $\Rightarrow$ Probable causal relationship but never definitely!

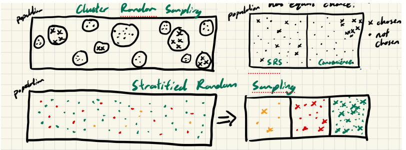
Conducting a Sample Survey

1. Sampling Frame: List of subjects for sampling
   $\Rightarrow$ Ideally = Population
2. Sampling Design: Method of subject selection
   $\Rightarrow$ Ideally sampled by chance
   - Simple Random Sampling
   - Cluster Random Sampling
     Use when:
     - Reliable frame unavailable
     - Cost of SRS too high

     BUT
     - Larger sample size required
     - Selecting small number of clusters might be more homogeneous than population
   - Stratified Random Sampling
     Use when:
     - Response differs typically across strata
     - To include enough subjects in each stratum

     BUT
     - Must know the stratum each subject belongs to
     - Need to define multiple sampling frames

   $\Rightarrow$ Non-Random Sampling sometimes required
   - Convenience Sample
     - Data can be obtained relatively cheaply
     - May poorly represent the population
     - Bias depends on the method of convenience sample
   - Volunteer Sample



3. Data Collection

F2F Interview	Phone Interview	Self-Admin Questionnaire
More likely to participate	Less patient, likely to hang up	Lower participation rates
High costs	Low costs	Low costs
May not answer sensitive questions on opinion and lifestyle		More willing to answer sensitive questions

Sampling Bias: Bias during sampling

- Non-Random Sampling
- Undercoverage: Non-representative sampling frame
  eg. Survey through landlines not reaching homeless people

Non-Sampling Bias: Bias during data collection

- Nonresponse Bias
  - Sampled subjects cannot be reached/refuse to participate
  - Missing data for certain questions
- Response Bias
  - Non-honest responses
  - Confusing/Leading questions
  - Answering wrongly

Experimental Studies

- More sure of a causal relationship as lurking variables' impacts more easily addressed.
- Random selection of treatments $\Rightarrow$ Reduced potential for lurking variables.

Conducting an Experiment

1. Obtaining experimental units
   Typically have to be a convenience sample
   $\Rightarrow$Representative?
2. Assigning to treatments
   The Control Group: Placebo/Existing treatments for ethical or comparison reasons
   Random Assignment:
   - Prevent bias with systematically different non-randomly assigned groups
   - Balance groups on lurking variables to prevent effects on association
3. Performing Treatment
   Blinding: Units unaware of treatment
   Double-Blinding: Those in contact with units unaware of treatment

5 Probability

Definitions:

- Probability: Proportion of an outcome in the long run
  (Converges due to law of large numbers)
- Sample Space: Set of ALL possible outcomes
- Event: A particular outcome, OR Set of possible outcomes (event $\subseteq$ Sample space)
  $P(Event) = \sum P(Outcome)$
- Disjoint Events $A, B \iff A \cap B = \varnothing$
- Independent Events A, B

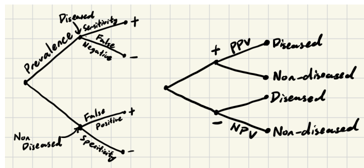
$$\begin{aligned} P(A \cap B) &= P(A) \times P(B) \\ P(A|B) &= P(A) \\ P(B|A) &= P(B) \end{aligned}$$

Probability Cheatsheet

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) \\ &\quad - P(A \cap B) \\ &= P(A) + P(B) \text{(Disjoint)} \\ P(A \cap B) &= P(A) \times P(B|A) \\ &= P(B) \times P(A|B) \\ &= P(A) \times P(B) \text{(Independent)} \\ P(A \cap B \cap C) &= P(A) + P(B) + P(C) \\ &\quad - P(A \cap B) - P(A \cap C) \\ &\quad - P(B \cap C) - \\ &\quad P(A \cap B \cap C) \\ P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= P(A) \times \frac{P(B|A)}{P(B)} \end{aligned}$$

For $B_i, B_2, ..., B_n$ partitioning S

$$\begin{aligned} P(A) &= \sum_{i=1}^{n} P(A \cap B_i) \\ &= \sum_{i=1}^{n}\{P(A|B_i)P(B_i)\} \end{aligned}$$



6 Distributions

Mean and Variance of Random Variable X

Discrete:

$E(X) = \mu_X = \sum xP(X = x)$

$Var(X) = \sum(x - \mu)^2 P(X = x)$

Continuous:

$E(X) = \mu_X = \int xP(X = x)dx$

$Var(X) = \int(x - \mu)^2 P(X = x)dx$

Expected Value $E(X)$:

Average in a long run of observations, NOT the expected value of a single observation!

Normal Distribution

$$Norm(\mu, \sigma^2) \qquad \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+$$

1. Symmetric
2. Bell-shaped
   0.68 in $(\mu - \sigma, \mu + \sigma)$
   0.95 in $(\mu - 2\sigma, \mu + 2\sigma)$
   0.997 in $(\mu - 3\sigma, \mu + 3\sigma)$
3. Approximates many discrete distributions (with large n)

$Z - Score = \frac{x-\mu}{\sigma}$ (No. of standard deviations a value falls from the mean) Standard Normal: $Norm(0, 1)$

Binomial Distribution

Let X = no. of successes in n trials.

$$X \sim Binom(n, p) \qquad n \in \mathbb{N}, p \in [0, 1]$$

1. Binary Outcomes
2. Independent n trials
   If sampling without replacement,
   $n < 10\%$ of Population
3. $P(success) = p$ is a constant
   Binomial is perfectly symmetric $\iff p = 0.5$

For $x \in \mathbb{N}_{\leq n}$,

$$P(X = x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

$$E(X) = np \qquad Var(X) = np(1-p)$$

Approximation by Normal Distribution

$np(1-p) \geq 5 \vee (np \geq 15 \wedge n(1-p) \geq 15)$

$\Rightarrow X \approx Norm(np, np(1-p))$

7 Sampling Distribution

Probability distribution that specifies probabilities for the possible values of the statistic

Population	Data	Sampling
Where a sample comes from	is THE sample	Describes how a sample is likely to look like

Categorical

Population: proportion p

$\Downarrow$

Sample: n observations with proportion $\hat{p}$

Sampling Distribution:

$successes \sim Binom(n, p) \quad \hat{p} = \frac{successes}{n}$

$\therefore \mu_{\hat{p}} = p \quad s_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

$np(1-p) \geq 5 \Rightarrow \hat{p} \approx Norm(p, \frac{p(1-p)}{n})$

Quantitative

Population:

Mean μ, Standard deviation: σ

$\Downarrow$

Sample $\approx$ Distribution of Population:

n observations with mean $\overline{X}$,

Standard deviation: s_X

Sampling Distribution:

$$\mu_{\overline{X}} = \mu \quad s_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

Central Limit Theorem

$n > 30 \vee pop \sim Norm(\mu, \sigma^2)$

$\Rightarrow \overline{X} \sim Norm(\mu, \frac{\sigma^2}{n})$

8 Confidence Intervals

Getting from Sample Dist. to Parameters: Estimations

Point Estimate

Ideal Properties:

- Unbiased (Centred at parameter)
- Small Standard Deviation ($\therefore$ Sample Mean over Median)

Confidence Interval

- Indicates precision
- Interval around the point estimate (Margin of Error)
- Associated with certain degree of confidence (≈ 0.95)
  The probability that it contains p

If we generated a $(1-\alpha)$ interval using the same method over many random samples, to estimate many population parameters, in the long run, $(1-\alpha)$ of those intervals will contain the population parameter.

Confidence Interval: Proportions

Assumptions:

- Data obtained by randomisation
- Distribution is $\approx$ Normal ($np(1-p) \geq 5$)
  For Binomial $\approx$ Normal, $s_{\hat{p}} \approx se$ approximation,

$(1-\alpha)$ Confidence Interval:

$$\hat{p} \pm \underbrace{q_{1-\frac{\alpha}{2}}(se)}_{\text{Margin of Error}} \qquad s_{\hat{p}} \approx se = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\text{Margin of Error} \leq W \iff n \geq \frac{q^2_{1-\frac{\alpha}{2}}}{W^2}\hat{p}(1-\hat{p})$$

$$n \Uparrow \Rightarrow MoE \Downarrow$$
$$(1-\alpha) \Uparrow \Rightarrow MoE \Uparrow$$
$$p \Uparrow \Rightarrow \hat{p} \Uparrow \Rightarrow MoE \Uparrow$$

n ultimately depends on costs and limitations. Rectify:
If $p \approx 0 \vee p \approx 1, +2 successes, +2 failures$

Confidence Interval: Mean

Assumptions:

- Data obtained by randomisation
- Population Distribution $\approx$ Normal
  Robust, but not to outliers:
  - Summary statistics $\overline{X}$ and s_X sensitive to outliers.
  - $\overline{X}$ no longer $\approx \mu_{\overline{X}} = \mu$

Robustness Confidence Interval is robust wrt. the normality assumption $\Rightarrow$ Performs adequately even when assumption is modestly violated

$(1-\alpha)$ Confidence Interval:

$$\overline{X} \pm \underbrace{t_{df=n-1,1-\frac{\alpha}{2}}(se)}_{\text{Margin of Error}} \qquad s_{\overline{X}} \approx se = \frac{s_X}{\sqrt{n}}$$

s_X is a point estimator of σ.

For $df \geq 30$ ($n > 30$), and $\mu \pm 3\sigma \approx Range(X)$

$$\text{Margin of Error} \leq W \iff n \geq \frac{\sigma^2 q_{1-\frac{\alpha}{2}}{}^2}{W^2}$$

$$n \Uparrow \Rightarrow MoE \Downarrow$$
$$(1-\alpha) \Uparrow \Rightarrow MoE \Uparrow$$
$$\sigma^2 \Uparrow \Rightarrow s^2 \Uparrow \Rightarrow MoE \Uparrow$$

t-Distribution

Distribution to allow generalisation for small sample sizes but assumes normal distribution

$$t_{df} \qquad df \in \mathbb{R}$$
$$lim_{df \to \infty} t_{df} = Norm(0,1)$$
$$t_{df=30} \approx Norm(0,1)$$

1. Bell-shaped
2. Slightly thicker tails than normal
3. Shows more variability than normal

9 Significance Tests

Assumptions

Certain conditions or assumptions that the test requires, or makes $\Downarrow$

Hypotheses

H_0: statement that the parameter takes a particular value (Usually no effect)

H_a: statement that the parameter falls in some alternative range of values. (Usually represents an effect)

- Assumed to be true until sufficient evidence against the hypothesis
- One/two-sided test ($>$ or $<$ or $\neq$)

$\Downarrow$

Test Statistic

How far the point estimate of the parameter falls from the H_0 value, usually in no. of se
Is a random variable, each sample is an observation
Distribution under H_0 is the null distribution

P-Value

Probability that the test statistic equals, or is more extreme, than the observed. Calculated by assuming H_0.
Smaller P-Value $\Rightarrow$ Stronger evidence against H_0

$\Downarrow$

Conclusion

Interpretation of the P-Value, and in context

Significance level α: the number such that we reject H_0 if the P-Value $\leq$

Statistically Significant: The results are statistically significant, if the data provides sufficient evidence to reject H_0 and support H_a

Types of Errors

	Decision	
Reality	Do not reject H_0	Reject H_0
H_0	Correct Conclusion	Type I Error
H_a	Type II Error	Correct Conclusion

P(Type I Error): Significance level α

P(Type II Error): Complex, but inversely related to P(Type I Error). For fixed α, prob. decreases:

- as parameter moves further into H_a, away from H_0
- as sample size increases

Plot: Probability against p_0 for fixed α, n

Power of a test$= 1 - P$(Type II Error)

Misinterpretations

- "Do not reject H_0"$\neq$"Accept H_0"
- Small P-Value does not imply confidence interval is far

Significance Test for p

Assumptions

- Categorical Variable
- Data obtained using randomisation
- Sample size is sufficiently large ($np(1-p) \geq 5$)
  Sample size is small $\Rightarrow$ two-sided test is robust.
  Otherwise null-dist$= Binom(n, p_0)$

Hypotheses:

$$H_0: p = p_0 \qquad H_a: p \neq p_0, p > P_0, p < p_0$$

Test Statistic:

$z - score_{\hat{p}}$ supposing H_0

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

P-Value:

Null Distribution: $Norm(0,1)$

$$P-Value = \begin{cases} P(Z < z) & left\text{-}sided \\ P(Z > z) & right\text{-}sided \\ 2 \times P(Z > z) & two\text{-}sided \end{cases}$$

Conclusion: If P-Value is $> \alpha$, strong evidence against H_0. Otherwise, we do not have strong evidence against H_0.

Significance test for $\overline{X}$

"One Sample t-Test"

Assumptions

- Quantitative Variable
- Data obtained using randomisation
- Population distribution $\approx$ Normal
  Two-sided test is robust (because CLT)
  Except when n is small and H_a is one-sided, sampling distribution is no longer t dist.

Hypotheses

$$H_0: \mu = \mu_0 \qquad H_a: \mu \neq \mu_0, \mu > \mu_0, \mu < \mu_0$$

Test Statistic $T - Score_{\overline{X}}$ supposing H_0

$$T = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

P-Value:

Null Distribution: $t_{df=n-1}$

$$P\text{-}Value = \begin{cases} P(t_{df} < T) & left\text{-}sided \\ P(t_{df} > T) & right\text{-}sided \\ 2 \times P(t_{df} > T) & two\text{-}sided \end{cases}$$

Conclusion If P-Value is $> \alpha$, strong evidence against H_0. Otherwise, we do not have strong evidence against H_0.

Two-sided Test vs. Confidence Interval

Two-sided test P-Value $\leq \alpha \iff$
$(1-\alpha)$ Conf-Int does not contain H_0 value

10 Bivariate Inference Methods

Sampling Distribution of $(p_1 - p_2)$

$\mu_{\hat{p_1} - \hat{p_2}} = p_1 - p_2$

$s_{\hat{p_1} - \hat{p_2}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Confidence Interval for $(p_1 - p_2)$

Assumptions

- Categorical Response variable observed
- Independent random samples for the two groups
- Large sample sizes, $np > 10 \wedge n(1-p) > 10$ for each group

$(1-\alpha)$ Confidence Interval:

$$\hat{p_1} - \hat{p_2} \pm q_{1-\frac{\alpha}{2}}(se)$$

$$se = \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$

If confidence interval contains 0, it is plausible that $(p_1 - p_2) = 0$, and the proportions might be equal.
Sign of values: $p_1 > p_2$ or $p_1 < p_2$
Magnitude of values: The size of the true difference in proportions

Significance Test for $(p_1 - p_2)$

Same assumptions as confidence interval

Hypotheses

$H_0: p_1 = p_2, H_a: p_1 \neq p_2, p_1 > p_2, p_1 < p_2$

Test Statistic, $\hat{p}$ is the pooled estimate

$$z = \frac{(\hat{p_1} - \hat{p_2}) - 0}{se_0}, se_0 = \sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}$$

P-Value

Null Distribution: $Norm(0,1)$

Conclusion Groups are (statistically) significantly different if P-Value is small

Sampling Distribution of $(\mu_1 - \mu_2)$

$$\mu_{\overline{X_1} - \overline{X_2}} = \mu_1 - \mu_2 \qquad s_{\overline{X_1} - \overline{X_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Null Distribution: t_{df}. df is complex.

Confidence Interval for $(\mu_1 - \mu_2)$

Assumptions

- Quantitative response variable observed
- Independent random samples for the two groups
- Approximately normal population dist. for each group
  Robust, except to outliers [Confidence Interval: Mean]

$(1-\alpha)$ Confidence Interval:

$$(\overline{X_1} - \overline{X_2}) \pm t_{df,1-\frac{\alpha}{2}}(se), se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}}$$

Significance Test for $(\mu_1 - \mu_2)$

Same assumptions as confidence interval

Hypotheses

$H_0: \mu_1 = \mu_2, H_a: \mu_1 \neq \mu_2 \mu_1 > \mu_2 \mu_1 < \mu_2$

Test Statistic, se same as confidence interval

$$t = \frac{(\overline{X_1} - \overline{X_2}) - 0}{se_0}$$

P-Value Null Distribution: t_{df}

Conclusion Groups are (statistically) significantly different if P-Value is small

Significance Test for $(\mu_1 - \mu_2), \sigma_1 = \sigma_2$

F test for comparing standard deviation:
P-Value$< \alpha = 0.05$
$\Uparrow$ NOT robust to population normality assumption

Assumptions, Test Statistic, P-Value, Conclusion are all the same as for non-equal std. dev, but with:

$$se = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{x_1 + n_2 - 2}}$$

$df = n_1 + n_2 - 2$

11 Linear Regression

$$Y = \overline{Y} - b(\overline{X}) + R(\frac{s_Y}{s_X})X + \epsilon$$

Y-intercept $\overline{Y} - b\overline{X}$

Predicted value of y when $x = 0$, Might have no interpretative value (if no observations near x=0).

Slope $R(\frac{s_Y}{s_X})$

Same sign as R. Amount that $\hat{y}$ changes with one unit increase of x R^2 is the % of variability in the response variable that can be explained by the linear relationship with the explanatory variable.

Error term ϵ

Assumptions

- Data obtained by randomisation
- Relationship between X and Y is linear
- Error term $\epsilon \sim Norm(0, \sigma^2)$ where σ is a constant

Implications of Assumptions

$\forall X(Y \sim Norm(\beta_0 + \beta_1 X, \sigma^2))$

Ordinary Least Squares Estimation

Best fit regression is the minimalisation of $\sum_{i=0}^{n-1} e_i^2$

Interpreting Info of Linear Regression Models

$\hat{Y} = \hat{B_0} - \hat{B_1}X$

Residuals

Quartiles of the residuals of each point with the model

Estimate $\hat{\beta_i}$

$\Rightarrow$ Point Estimates of each coefficient in the model

Std. Error

Standard error of each coefficient. Can be used to obtain a confidence interval

Residual Standard Error

The standard error of $\hat{\sigma}$ For each point $x_i, e_i = y_i - \hat{y_i}$

- Could be normalised, to get standard residuals $SR \sim Norm(0,1)$
- σ is the measure of how far the observations can deviate from best-fit line

σ^2 is the measure of how far the observations can deviate from the best-fit line

Multiple R-squared

Coefficient of Determination of linear model

Hypothesis Testing on Linear Models

t-Test

The significance of one regressor.

Assumptions Same as assumptions of model

Hypothesis

$H_0: \beta_i = 0$ OR Regressor i is NOT significant
$H_a: \beta_i \neq 0$ OR Regressor i is significant

Test Statistic

$$t = \frac{\hat{B_i}}{se}$$

P-Value

Null-Distribution: t Distribution, $df = n-$no. of coefficients

Conclusion

The coefficient is (not) significantly different from 0 at α-level

F-test

Assumptions Same as assumptions of model

Hypothesis

$H_0:$ model is NOT significant OR all the coefficients except β_0 are zero
$H_a:$ model is significant OR at least one of the coefficients except β_0 are non-zero

Test Statistic F-statistic from R output
$F = t^2$ for Simple Linear Regression

P-Value

Null-Distribution: F Distribution,
$df1 =$no. of coefficients-1
$df2 = n-$no. of coefficients
Find right-sided P-Value on F Distribution

Conclusion

The data provides (in)sufficient evidence that the built model is significant.
$P - Value < \alpha \Rightarrow$ ALL regressors used in the model are not significant, $Y = \beta_0$

Checking Assumptions of Linear Model

Before fitting model, scatterplot of Y against X:

1. Linearity (No curved bands)
2. Constant Variance (Funnel shape)

To verify the assumption $\epsilon \sim Norm(0, \sigma^2)$

1. SR against $\hat{Y_i}$
2. SR against X
   Points scatter randomly around 0, within $(-3, 3)$
   Funnel shaped observed $\Rightarrow$ Constant variance assumption violated
3. Histogram of SR
4. QQ Plot of SR
   SR has a normal distribution Skewed Distribution $\Rightarrow$ Normality assumption violated

Possible Fixes

- Add higher order terms
- Transform response into $\ln(Y), \sqrt{Y}, \frac{1}{Y}$
- Add more regressors
- Non-linear model required

$|SR| > 3 \Rightarrow$Potential outlier
Cook's Distance$> 1 \Rightarrow$Potential influential point
Avoid Extrapolation (Estimation using regressors outside domain)

Coefficient of Determination

Interpretation:

The proportion of total variation of the response (of sample mean $\overline{Y}$) that is explained by the model.

For simple model: $\sqrt{R^2} = |Cor(X, Y)|$

$R^2 = 1 \Rightarrow \forall i(\hat{Y_i} = Y_i)$ Adding regressors will always increase, or not change R^2. Use adjusted R^2

$$R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n - \text{no. of coefficients}}$$

Indicator Variables

Each indicator splits the model into two equations, on whether the indicator is 1 or 0
n categories require $n-1$ indicators
Identify the reference category when every indicator is 0.
Use anova P-Value for significance of categorical variables