STAT1006 Week 8 Cheat Sheet

Lisa Luff

10/11/2020

Contents

Assumptions of Error

- Remembering the 4 assumptions of error for SLR's
  1. $E(\epsilon_i) = 0$, for all i - Mean error is 0
  2. $var(\epsilon_i) = \sigma^2$, for all i - Constant variance
  3. ϵ_i and ϵ_j are independent for all $i \neq j$ - Independence and randomness
  4. $\epsilon_i \sim N(0, \sigma^2)$ to be able to make inferences from the regression model - Normality of error
- If these assumptions are met, linearity is implied
- Test these assumptions with residual plots

Critically Assessing the Regression Model

1. Generate a scatterplot

- Is the data linear?
  - Form
  - Direction
  - Strength

2. Linear relationship statistically significant?

- Hypothesis testing
  - $H_0 : \beta_1 = 0$

3. Does the model explain the variance of y?

- Coefficient of determination - R^2

4. Check the residuals

- Diagnostic checking

Diagnostic Testing

- Check the assumptions of error with graphs of standardised residuals
- Simple standardisation:
  - $\frac{\hat{\epsilon}_i}{s}$, where
    - $\hat{\epsilon}_i$ - is the sample residuals
    - s - is the sample standard deviation
- Standard decision:
  - $z = \frac{\hat{\epsilon}_i - E(\hat{\epsilon}_i)}{s}$
  - *Look at this later*
- **In R**
  - To use the residuals, create the lm, and then create a variable that holds *variable*lm$residuals
  - To create standardised residuals, use rstandard(*variable*.lm)

Checking Model Validity

1. Determine if proposed regression is a valid model

- Use plots of standardised residuals

2. Visually assess if the assumptions are being violated

- If so, what can we do to overcome these violations →

3. Find any outliers
4. Find if any outliers are bad leverage points

- Assess its influence on the model

5. Is the assumption of constant variance reasonable

- If not can we use transformations to overcome this

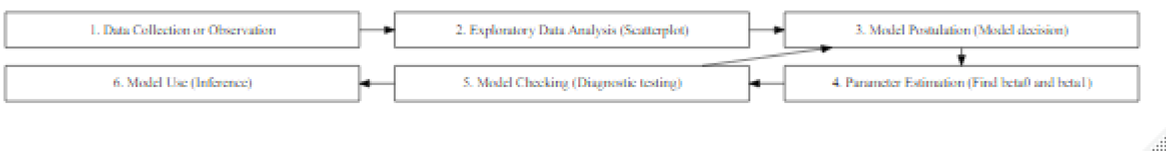
6. If the data is collected over time, is it correlated over time

- Or do we need to use a time series

7. If the sample is small, or we want prediction intervals

- Test assumption that errors are Normally distributed

Role of Diagnostic Testing



Residual Analysis

- Checks the appropriateness of the Least Squares Line (model)
- 1. Find the predicted value, $\hat{y}$, for each case (x, y) in the data set
  2. Find the residuals: $residual = y - \hat{y}$
  3. Plot residuals against the x values

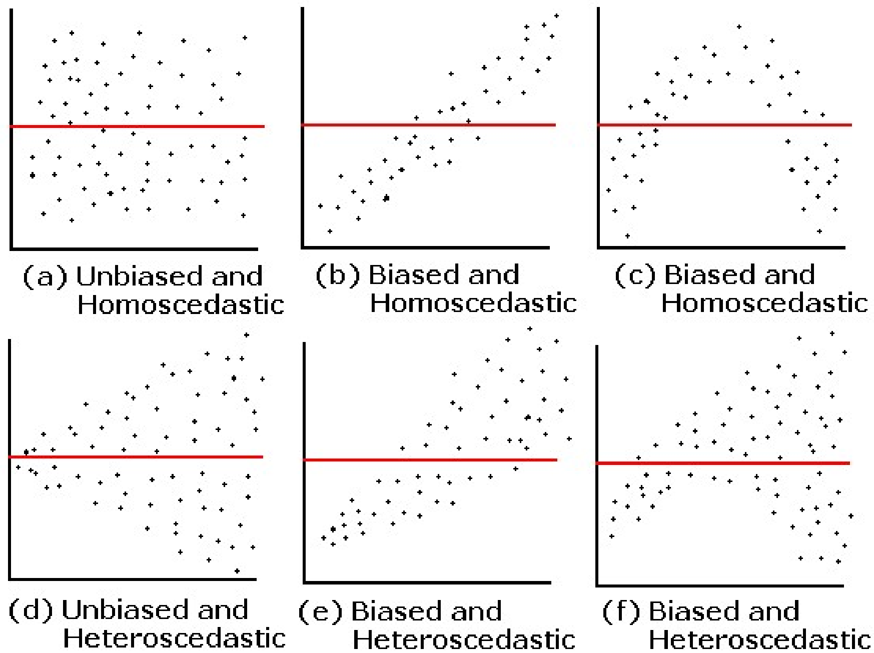


Figure 1: ‘Residual Plots’

- The linear model is appropriate if:

1. Pattern-less residuals:

- **Unbiased in fig1**
- Residuals should be randomly distributed around the horizontal line through zero
- Shows data is random and independent
- If there is any pattern, the model is not a good summary of the data
- Look at the actual methodology of how the data was collected
- Are the residuals correlated when plotted against time? * We can check this by plotting the residuals against time, or the values of x * Any pattern means that a linear model was not appropriate for this data

2. Constant variable:

- **Homoscedastic in fig1**
- If the residuals are more spread out at one end than the other
- If this occurs, the model is not a good summary of the data
- Look at the scatterplot and see if the spread of residuals is consistent, or of there is clustering

3. Normally distributed:

- Residuals should follow an *approximately* Normal distribution
- It is approximately, because if the sample is small, it can be hard to assess
- Create a histogram of the residuals and compare
- Use a QQ plot
  - Points on a QQ plot:
    - $(F^{-1}\left(\frac{i}{n+1}\right), z_i)$, where $i = 1, 2, ..., n$
- **In R**
  - Use hist(*variable*.lm) to create the histogram, and use that for residual analysis
  - Use qqplot(*variable*.lm) and qqline(*variable*.lm) for the QQ plot
- **In R**
  - Use plot(*variable*.lm) to create 4 plots with residuals/fitted, standardised residuals/fitted, QQ plot and residuals and leverage (see below)

Outliers and Leverage Points

- Outliers are observations far from the rest of the data
  - Shouldn't delete unless it is an error, can be important data!

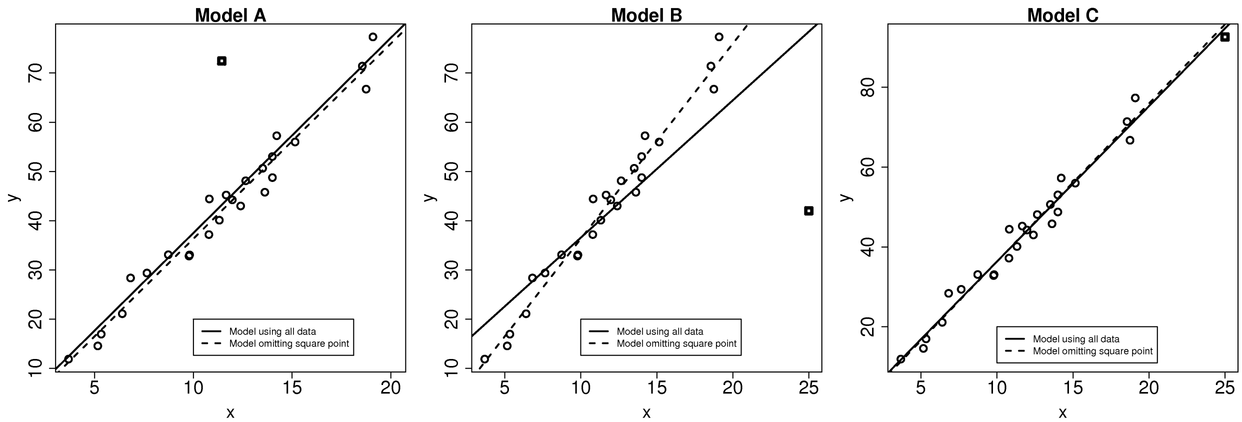


Figure 2: 'Leverage Points'

- Leverage points are outliers that have the ability to influence the fitted line
  - **Model A** Leverage points might only mildly affect the fitted line
  - **Model B** Leverage points that do significantly affect the fitted line are "bad leverage" points, and can become a candidate for removal
  - **Model C** A leverage point that is a long way away but along the fitted line is considered a "good leverage" point
- How to see if an outlier is a bad leverage point:
  1. The hat matrix elements h_i (This is for SLR)
  2. Cook's distance statistic D_i (Also for SLR)
  3. The Studentized deleted residuals T_i^* (Later down the track)
  - Only when all three criteria provide consistent results, should an observation be removed
- The hat matrix element h_i:
  - $h_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$
  - Rule of thumb: if $h_i > \frac{4}{n}$, then X_i is an outlier
    * As such it *may* be considered for removal
  - **In R**
    * hatvalues(*variable*.lm)
    * Or plot(*variable*.lm)
- Cook's distance statistic D_i:
  - $SR_i = \frac{e_i}{S_{YX}\sqrt{1-h_i}} \rightarrow$
  - $D_i = \frac{SR_i^2 h_i}{2(1-h_i)}$
  - Rule of thumb: if $D_i > \frac{4}{n-2}$, then X_i is an influential point (leverage point)
  - **In R**
    * Use cooks.distance(*variable*.lm)
    * Or plot(*variable*.lm)

Handling Outliers and Leverage Points

- Outliers and leverage points can point out important problems with the model, or a problem not considered previously
  - So you don't want to routinely delete them
- Can be worth considering an alternative model
- Adding one or more dummy variables can be helpful (will learn later)
- Or can use transformations

Transformations

- Used to:
  - Overcome problems due to non-constant variance
    - Mostly this
  - Estimate percentage effects
  - Overcome problems due to non-linearity (not in this unit)
- Visualise the distributions of responses in vertical strips if:
  - Mean ~ straight line
  - StD ~ constant
- Transform X if:
  - Mean ~ curved
  - StD ~ constant
- Transform X^2 if:
  - Mean significantly curved
  - StD ~ constant
- Transform Y if:
  - Mean ~ curved
  - StD increasing
- Report skewness (do not remedy, only report skewness) if:
  - Mean ~ straight line
  - StD ~ constant
  - Shape is skewed
- Weighted Regression (not in this unit) if:
  - Mean ~ straight line
  - StD increasing
  - Mean being dragged by leverage point/s
- **Note**: If X and Y use the same values (both counts, both seconds, both meters, etc.), then you may want to transform both variables