STAT1006 Week 6 Cheat Sheet

Lisa Luff

10/3/2020

Contents

Linear Models

- The key idea is that of an additive model:
  - Response - $y =$ Explanatory - $g(x_0, x_1, x_2, ..., \beta_0, \beta_1, \beta_2, ...) +$ Error - ϵ
  - With assumptions about for form of $g(\cdot)$ and ϵ

Fitting a Linear Relationship

- We speculate that the relationship in the population follows the following model:
  - Response - $y_i =$ Y Intersect - $\beta_0 +$ Slope - $\beta_1 x_i +$ Error - ϵ_i, where $\epsilon_i \sim N(0, \sigma^2)$
  - The parameters β_0 and β_1 are fixed constants that we want to estimate the values of using the observed data

Uses for Linear Models

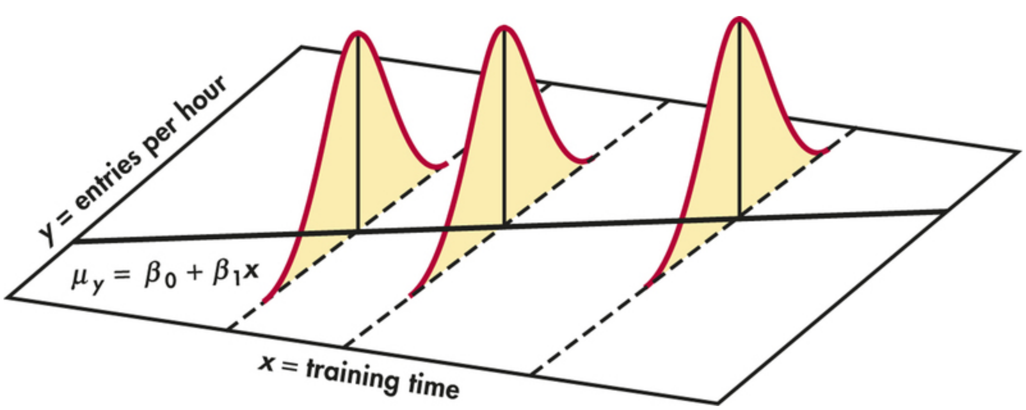
- In some cases the underlying relationship is approximately linear
- A simple model might be "good enough" for the purposes
- Might be a good approximation to a non-linear model (eg. over a narrow region)
- Makes sense to try it before trying more complex models
- **All models are wrong, but some are useful**

Simple Linear Regression (SLR)

- Used for a scatterplot that shows a relatively linear relationship between a numerical explanatory variable x and a numerical response variable y
  - Use least-squares line fitted to the data to predict y for a given value of x
  - The pattern of variation in the slope is described by its' *sampling distribution*
    - Linear regression assumes **equal variance of y** (σ is the same for all values of x)
- When used on data from a random sample of a larger population, you can use statistical inference to answer questions about the relationship between x and y

SLR Population Equations

- In the population, the linear regression equation is:
  - $\mu_y = \beta_0 + \beta_1 x$
    - Where μ_y is the mean of y
- Sample data is used to estimate:
  - Data = Fit + Error
  - $Y_i = (\beta_0 + \beta_1 X_i) + (\epsilon_i)$
    - Where ϵ_i are independent and Normally distributed $N(0, \sigma)$



Estimating the Parameters

- β_0, β_1 and σ of y are the unknown parameters of the regression model
- We use the **random sample** data to provide an unbiased estimate of the parameters
  - With the least-squares regression line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, where
    * $\hat{y}$ estimates mean response μ_y
    * $\hat{\beta}_0$ estimates y intercept β_0
    * $\hat{\beta}_1$ estimates slope of β_1
- **Unbiased** means that $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables, and are subject to variation in different samples
  - As such if you take lots of samples, then take the average of the estimate for $\hat{\beta}_0$ and $\hat{\beta}_1$, this will be equal to the true population values β_0 and β_1
    * $E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$

Conditions for SLR Inference

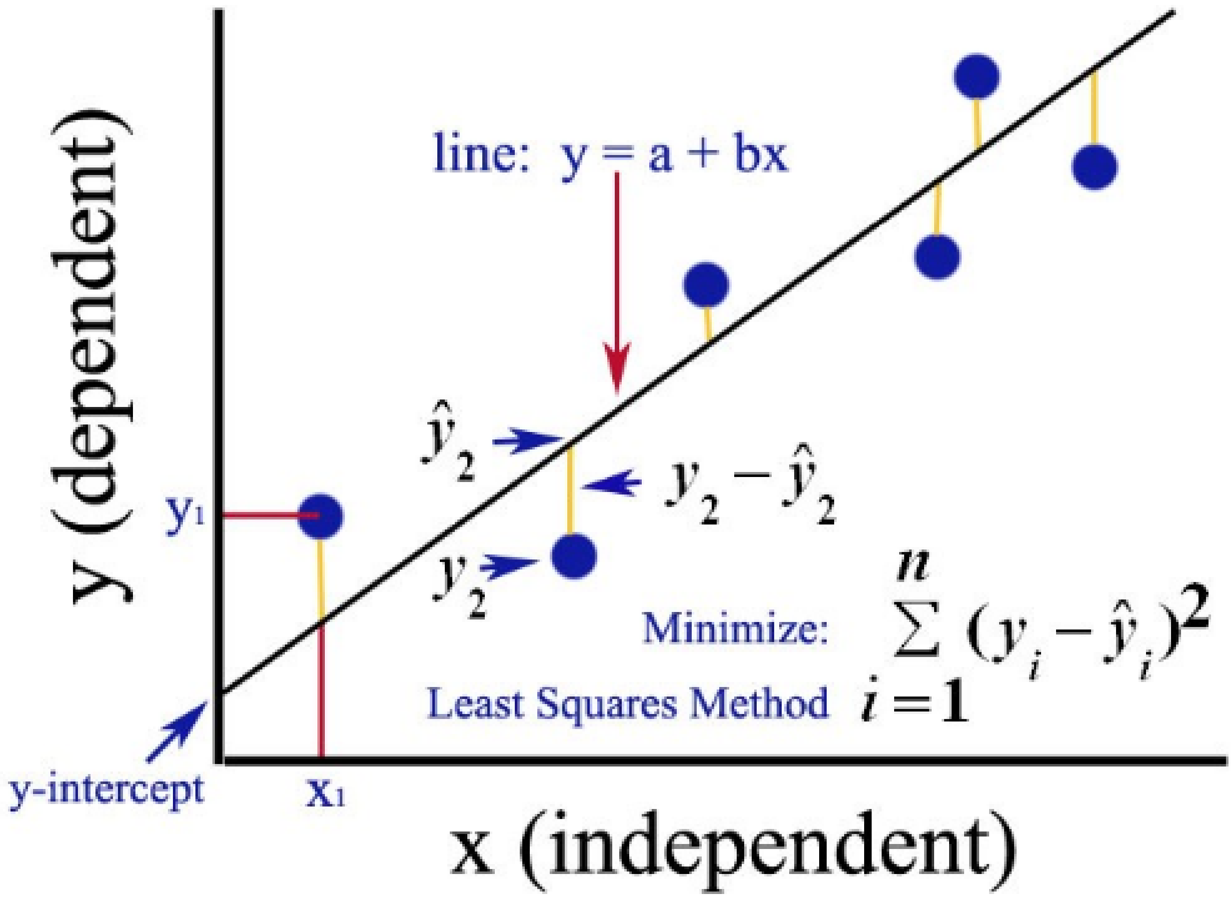
- Based on that the slope and intercept of the least-squares line are statistics from the sample data and will be different for every sample we take.
- So to do this inference we need the following conditions:
  - For any fixed value of x, the response y varies according to a Normal distribution
  - Repeated responses are independent of each other
  - The mean response μ_y has a straight-line relationship with x given by a population regression line as seen above
  - The slope and intercept are unknown parameters
  - The standard deviation of y (σ) is the same for all values of x, and σ is unknown

SLR Assumptions

- To allow completion of the model specifications we assume:
  1. $E(\epsilon_i) = 0, \text{ for all } i$
  2. $var(\epsilon_i) = \sigma^2, \text{ for all } i$
  3. ϵ_i and ϵ_j are independent for all $i \neq j$
  4. $\epsilon_i \sim N(0, \sigma^2)$ if we wish to make inferences about the regression model
- These assumptions imply that:
  - $E(Y|X = x) = \beta_0 + \beta_1 x$ and
  - $var(Y|X = x) = \sigma^2$
- Checking these assumptions is an important part of model-checking

The Least Squares Method

- The least squares regression line is found using the least squares method
  - A line is drawn on the scatterplot and the aim is to have the vertical distances of the observations from the drawn line to be as small as possible
  - The least squares regression line is the unique line using this method such that the sum of the squared vertical differences $(y - \hat{y})$ (to even out positive and negative) between the observed data y and the redicted value $\hat{y}$ of the line is the smallest



- This method therefore aims to minimise the Sum of Squares of Error (SSE)
  - $SSE = \sum(Y_i - \beta_0 - \beta_1 X_i)^2$
  - To do this the 1^{st} partial derivatives are set to 0, and where rearranged end up as:
    - $\sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i$
    - $\sum_{i=1}^{n} X_i Y_i = \beta_0 \sum_{i=1}^{n} X_i + \beta_i \sum_{i=1}^{n} X_i^2$
    - There are called the *normal equations*, and are what you need to solve to find $\hat{\beta}_0$ and $\hat{\beta}_1$
- We can solve these equations using data points Y and X with
  - $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$
  - $\hat{\beta}_1 = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2} \sim \frac{Cov(X,Y)}{Var(X)} = \frac{S_{xy}}{S_{xx}}$
  - Use R to do this!
- The regression line always passes through the mean of x and y

Interpretation

- $\hat{\beta}_0$ = y intercept = y value at $x = 0$
  - This is only interpretable if $x = 0$ is of practical value or interest
- $\hat{\beta}_1$ = slope = change in y for every 1 unit of increase of x
  - Always interpretable

Variance

- $Var(\hat{\beta}_0) = \sigma^2 \left(\dfrac{1}{n} + \dfrac{\overline{X}^2}{\sum(X_i - \overline{X})^2} \right)$
- $Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum(X_i - \overline{X})^2}$
  - We can also use c_i to express this as $Var(\beta_1) = Var(\sum c_i(Y_i - \overline{Y})) = Var(\sum c_i Y_i)$
  - Or we can use matrices
  - Or just use R!

Variance of the Error

- $\sigma^2 = Var(\epsilon_i) = Var(y_i - \beta_0 - \beta_1 x_i)$
  - But we can only estimate the coefficients
- An unbiased estimate of σ^2 is
  - $s^2 = \frac{SSE}{n-2} = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\epsilon}_i^2$, where
    * s^2 is the variance of the sample error
    * SSE is the Sum of Squares of the Error
    * $\hat{\epsilon}_i^2$ is the estimate of the error
    * The divisor is $n - 2$ because we have estimated two parameters (degrees of freedom)

Inference About the Slope

- Sampling distribution of $\hat{\beta}_1$
  - If the assumptions are satisfied then;
  - $\hat{\beta}_1 | X \sim N(\beta_1, \frac{\sigma^2}{\sum(X_i - \overline{X})^2})$, where
    * Unknown σ^2 is estimates by s^2
- Test statistic T
  - Because we are estimating σ^2 with s^2, we use
  - $T = \frac{\hat{\beta}_1 - \beta_1^0}{\frac{s}{\sqrt{\sum(X_i - \overline{X})^2}}} = \frac{\hat{\beta}_1 - \beta_1^0}{se(\hat{\beta}_1)} \sim t_{n-2}$, where
    * se is standard error (standard deviation)
    * t_{n-2} is the test distribution

Confidence Intervals

- If we assume that $\epsilon_i \sim N(0, \sigma^2)$
  - Then that means $Y_i | X_i \sim N(\beta_0 + \beta_i X_i, \sigma^2)$
    * Because each parameter estimate is a linear function of the data Y, each estimate $\hat{\beta}$ is also normally distributed
  - We can place confidence intervals on the regression parameter estimates using central limit theorum, and the variance terms from above
- Using the arguments from the slope inference, we can calculate a $100(1 - \alpha)\%$ confidence interval for β_1 with
  - $\hat{\beta}_1 \pm t_{\frac{\alpha}{2}, n-2}$ x $se(\hat{\beta}_1)$

Significance Test for Regression Slope

- To test the hypothesis $H_0 : \beta_1 =$ hypothesised value, compute the test statisitc
  - $t = \frac{\hat{\beta}_1 - \text{ hypothesised value}}{SE_{\hat{\beta}_1}}$
- Get the p-value by calculating the probability of getting a t statistic as large or larger in the direction specifed by H_A for t distribution with $df = n - 2$

Testing the Hypothesis of No Relationship

- If we hypothesis that there is a relationship between x and y, then we test for β_1
  - We rarely do a test of hypothesis for β_0 as it often has no practical interpretation

1. Hypotheses

- $H_0 : \beta_1 = 0$
  - Equivalent to testing the hypothesis of *no correlation*
- $H_A : \beta_1 \neq 0$

2. Test Statistic

- $T = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)}$

3. Sampling distribution

- $T \sim t_{n-2}$

4. P-Value

- Determined by H_A
- P-value $= P(|t_{n-2}| > t)$ for two sided

5. and 6. Decision and Conslusion

Inference About the Intercept

- As demonstated before
  - $\hat{\beta}_0 | X \sim N(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(X_i - \overline{X})^2}\right))$, where
    - σ^2 is estimated by s^2
- To test the hypothesis $H_0 : \beta_0 = \beta_0^0$, R tests $H_0 : \beta_0 = 0$ by default
  - We use $T = \frac{\hat{\beta}_0 - \beta_0^0}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(X_i - \overline{X})^2}}} = \frac{\hat{\beta}_0 - \beta_0^0}{se(\hat{\beta}_0^0)}$
- A $100(1-\alpha)\%$ confidence interval for β_0 is give by
  - $\hat{\beta}_0 \pm t_{\frac{\alpha}{2}, n-2}$ x $se(\hat{\beta}_0)$

Linear Modelling in R

- To find the linear model use the lm function
  - *variablelm* = lm(*reponse*~*explanatory*, data = *table/dataframe*)
  - R output of lm function shown with the summary function
    * Residuals: Min, 1Q, Median, 3Q, Max
    * Coefficients: Intercept and Explanatory
      · For each: Estimate (gives intercept for intercept value and slope for explanatory value), Standard Error, t value, Pr(>|t|)(p-value for β_1) (twosided, halve for 1 sided, or use pt function with lower.tail = FALSE)
    * Residual standard error and degrees of freedom
    * Multiple R-squared and Adjusted R-squared
    * F-statistic, degrees of freedom and p-value
  - There are different variables created by lm within the dataframe *variablelm*, including things like fitted which is $\hat{y}_i$

- To find the confidence interval use the confint function
  - confint(*variablelm*, level = *percent as decimal*) (default 95% CI)
  - R output of confint function
    * Intercept and Explanatory
      · Point Estimate, lower and upper confidence interval values for each

- To plot the least squares regression line using the abline function
  - First plot the data
  - Then add abline(*variablelm*, col = "colour", lwd = *line width*)
- To plot confidence intervals use matlines(sort(*explanatory*), *civariable*[order(*explanatory*), 2:3], lwd = *line width*, lty = 1)

Reading CSV Files in R

- You can click on the file in your directory and choose import
- OR use the read_csv function
  - library(readr)
  - *variable* <- read_csv("*name.csv*")