STAT1006 Week 7 Cheat Sheet

Lisa Luff

10/3/2020

Contents

Predicting the Mean Value of the Population Regression Line

- $\mu_y = E(Y|X)$ is the value of the regression line at $X = X_0$
- For any given value of X_0, we know
  - $E(Y|X) = \beta_0 + \beta_1 X$
  - $Var(Y|X) = \sigma^2$
- To predict the **average** value of Y for a given value of X_0
  - $E(Y_i|X_0) \sim \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_0$, where
    - $E(Y_i|X_0)$ is the point estimate
    - $\hat{Y}$ is the prediction of y

Confidence Intervals for the Population Regression Line

- We can calculate a confidence interval for the population mean μ_y of all responses y when x take the value x^* (within the range of tested data)
  - **ONE x for ALL y**
  - x^* is a specific value of x
- This is given by $\hat{\mu}_y \pm t^* SE_{\hat{\mu}}$, where
  - $\hat{\mu}_y$ is the estimate of the parameter
  - t^* is the critical value of the t_{n-2} distribution for the confidence interval
  - $SE_{\hat{\mu}}$ if the standard error
  - $t^* SE_{\hat{\mu}}$ together is the margin of error
- This can be found across all values of x and shown as a continuous curve on either side of $\hat{y}$

- To place a confidence interval around the prediction of the mean $E(Y|X)$ we need to estimate
  - $Var(E(Y|X_0)) = Var(\hat{Y}) = Var(\hat{\beta}_0 + \hat{\beta}_1 X_0)$
  - And because the random variable $\hat{\beta}_1 = \overline{Y} - \hat{\beta}_1 \overline{X}$
    - Then $\hat{Y} = \overline{Y} + \hat{\beta}_1 (X_0 - \overline{X})$
    - And so the variance will be the variance of those two parts
- So $SE(\hat{Y}|X_0) = \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum (X_i - \overline{X})^2} \right)}$
  - Broken into the two parts
    - $Var(\overline{Y}) = \frac{\sigma^2}{n}$ - Variance of Observations
    - $Var(\hat{\beta}_1 (X_0 - \overline{X}))$ - Variance of Estimated
  - Together
    - $\frac{\sigma^2 (X_0 - \overline{X})^2}{\sum (X_i - \overline{X})^2}$
- To create a confidence interval we must assume Normality (or some other distribution) for the residuals, so
  - $(\hat{\beta}_0 + \hat{\beta}_1 X_0) \pm t_{n-2, \frac{1-\alpha}{2}}$ x $SE(\hat{Y}|X_0)$
    - Preferrable to use σ^2 if that is know instead of SE
- This is a confidence interval for the regression line at any point X_0

Fitted (Predicted) Values

- **ALL x for ONE y**
- We can use the equation of the least squares line to predict y for each value of x (within the range of x)
  - $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
    - Just substititute x to find $\hat{y}$

Prediction Interval for the Least Squares Line

- We use a prediction interval to estimate the *individual* response of y for a given x
- For many samples for a given x there will be many values for y following $N(0, \sigma)$ around the mean response μ_y
- The prediction interval of μ_y is given by
  - $\hat{y} \pm t^* SE_{\hat{y}}$, where
    - t^* is the critical value of the t_{n-2} distribution for the confidence level
  - This is shown as a continuous curve on either side of $\hat{y}$
- These prediction intervals are wider than the corresponding confidence intervals for μ_y

- If you want a confidence interval for all future values of y, then we can obtain that with
  - $SE(Y|X_0) = \sqrt{\sigma^2 \left(1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum (X_i - \overline{X})^2}\right)}$
  - Again this has two parts
    - $Var(Y|X = X_0) = \sigma^2$ - Variance of Observations
    - $Var(\hat{\beta}_0 + \hat{\beta}_1 X_0) = Var(\hat{Y}|X = X_0)$ - Variance of Estimated
  - It is the extra part that means this interval is extra wide

- To make a confidence interval we must assume normality (or some other distribution) for the residuals
  - $(\hat{\beta}_0 + |hat\beta_1 X_0) \pm t_{n-2, \frac{1-\alpha}{2}}$ x $SE(Y|X_0)$

Intervals in R

- Confidence Intervals:
- Use the predict function
  - *variableconfmean* = predict(*variablelm*, interval = “confidence”, level = *percent as decimal*)
  - To predict as specific values add the argument
    - newdata = data.frame(*explanatory* = c(*values*))
- Add the lines to a plot with the matlines function
  - matlines(sort(*explanatory*), *variableconfmean*[order(*explanatory*), 2:3], lwd = *line width*, col = “colour”, lty = 1)

- Prediction Intervals:
- Use the predict function again
  - *variablepi* = predict(*variablelm*, interval = “prediction”, level = *percent as decimal*)
  - To predict as specific values add the argument
    - newdata = data.frame(*explanatory* = c(*values*)))
- Again use the matlines function to the plot the lines
  - matlines(sort(*Explanatory*), *variablePI*[order(*explanatory*), 2:3], lwd = *line width*, col = “colour”, lty = 1)

The ANOVA Table

Source	Sum of Squares (SS)	DF	Mean Squares (MS)	F	P-value
Regression	$SSReg = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	1	$MSR = \frac{SSReg}{DFR}$	$\frac{MSR}{DFR}$	Tail area above F
Error	$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-2$	$MSE = \frac{SSE}{DFE}$ $s^2 = MSE$		
Total	$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ SST = SSReg + SSE	$n-1$ DFT = DFR + DFE			

* Where s^2 is an unbiased estimate of the regression variance σ^2

Analysis of Variance (ANOVA) for Regression

- The regression model resembles an ANOVA, which also assumes equal variance, where
  - SST = SSR + SSE, and
  - DFT = DFR + DFE as seen above
    - These are related to the t distribution
- ANOVA is very important for regression
  - It can compare different but similar models and choose the best model for a set of data

ANOVA Hypotheses for Regression

- For a simple linear relationship, ANOVA tests the hypotheses:
  - $H_0 : \beta_1 = 0$ - There isn't a relationship, versus
  - $H_A : \beta_1 \neq 0$ - There is a relationship
  - We can look at these as two competing models
    - $(1) Y_i = \beta_0 + \epsilon_1$ - There isn't any relationship, versus
    - $(2) Y_i = \beta_0 + \beta_1 X_i + \epsilon_1$ - There is a relationship

ANOVA Analysis for Regression Models

- We can use these models to evaluate SST, SSE and therefore SSR for each model:
- Model 1
  - SSE -
    * The least squares line for the model minimises to $SSE = \sum(Y_i - \beta_0)^2 = \hat{\beta}_0 = \overline{Y}$
  - SST -
    * $SST = \sum(Y_i - \hat{\beta}_0)^2 = \sum(Y_i - \overline{Y})^2 = (n-1)s_y^2$, where s_y^2 is the sample variance of y
  - SSR -
    * There is no SSR, because the model is a mean line (not regression line)

- Model 2
  - SSE -
    * $SSE = \sum(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \sum \hat{\epsilon}_i^2 = (n-2)\hat{\sigma}^2$, where
    * $\sum \hat{\epsilon}_i^2$ is the sum of the squares of the residuals (Not SSR, SSR is for Regression)
  - SST -
    * $SST = \sum(Y_i - \overline{Y})^2 = (n-1)s_y^2$ as shown for Model 1
  - SSR (SSReg) -
    * $SSR = SST - SSE = \sum(Y_i - \overline{Y})^2 - \sum \hat{\epsilon}_i^2 = (n-1)s_y^2 - (n-2)\hat{\sigma}^2$
    * OR $SSR(SSReg) = SST - SSE = \sum(Y_i - \overline{Y})^2 - \sum(Y_i - \hat{Y})_i^2 = \sum(\hat{Y}_i - \overline{Y})^2$
    * This is called the Sum of the Squares DUE to Regression (SSReg), because it measures how much residual variation has decreased by fitting a linear model, rather than a constant mean to the data
- If Y and X are linearly related, then SSReg will be large
- If Y and X are not linearly related, then SSReg will be small

- Degrees of freedom $= SST = SSReg + SSE \rightarrow (n-1) = 1 + (n-2)$

Model Means

- If we divide a sum of squares by it's degrees of freedom, we get a Mean Square (MS)
  - Mean Square of the Error (MSE)
    * $MSE = \frac{SSE}{(n-2)} = \hat{\sigma}^2$
  - Mean Square Total (MST)
    * $MST = \frac{SST}{n-1}$
  - Mean Square of the Regression (MSR)
    * $MSR = \frac{SSReg}{1}$
- For the linear model (2) to fit better than the constant mean model (1), we want MSE to be much less than MST
  - Otherwise why bother?

F Distribution for Models

- The F statistic compares the MSR to the MSE
  - $F = \frac{(SST-SSE)}{MSE} = \frac{\frac{SSReg}{1}}{\hat{\sigma}^2}$
  - If SSReg is large, then F will be large
  - If SSReg is small, then F will be small
- F is a statistic and subject to random variation based on the parameters it estimates:
  - $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2, s^2, \overline{Y}$
  - As such the F distribution has two different degrees of freedom, one for the numerator and one for the denomenator
    * $F(\text{distribution}) = \frac{\frac{SSReg}{1}}{\frac{SSE}{(n-2)}}$
    * Or, a distribution of $F_{1,n-2}$

- We can use this distribution test our hypotheses
  - Comparing Mean Square Regression (MSR) to Mean Square Error (MSE)
  - Where $F = \frac{MSR}{MSE}$
  - When H_0 is true, F follows the $F_{1,n-2}$ distribution
  - And the p-value is $P(F \geq f)$
- The ANOVA test and the two sided t-test for $H_0 : \beta_1 = 0$ give the exact or very close to exact same p-value
  - Software output might provide t, F, or both alone with the p-value
  - So don't worry which test is used

Assumptions for ANOVA

- The statistic will have an F-distribution **only** if the residuals are Normally distributed
  - So when using the test, we need to examine the residuals
  - If they are highly non-Normal, the test is not valid
    * Can use a diagnostic to check this
  - Usually it is enough if the histogram is roughly mound shaped
    * This means the histogram is roughly symmetric with the highest density in the middle, and lowest density in the tails

Chi-Squared, F-Distribution and T-Distribution

- **This will NOT be tested, just FYI**
- Both SSE and SSR follow χ^2 distributions
  - SSE - χ^2_{n-2} distribution
  - SSR - χ^2_1 distribution
- A χ^2 variable describes the distribution of a sum of squares of normal data minue its mean
  - Eg, $SSE = \sum(Y_i - \hat{Y}_i)^2 \sim \chi^2_{n-2}$
- If Y has a Normal distribution with estimated mean $\hat{Y}$ and variance σ^2, then
  - $\frac{SST}{\sigma^2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{\sigma^2} \sim \chi^2_{df}$
  - So, $SSE = \sum(Y_i - \hat{Y}_i)^2 \sim \chi^2_{n-2}$
  - And $E(\chi^2_{df}) = df$
- The F distribution is a ratio of χ^2 with it's degree of freedom
- So if an F distribution is formed by two different chi-squared distributions
  - Say, $U \sim \chi^2_u$
  - and, $V \sim \chi^2_v$
  - Then $F = \frac{\frac{U}{u}}{\frac{V}{v}} \sim F_{u,v}$
    - And, usually $E(F) \sim 1$
  - Eg, $F = \frac{\frac{SSR}{1}}{\frac{SSE}{(n-2)}} = \frac{\frac{\chi^2_1}{1}}{\frac{\chi^2_{n-2}}{(n-2)}} \sim F_{1,n-2}$
- Both the χ^2 and F distributions tend towards normal as their degrees of freedom increase

- And the F distribution is the t distribution squared
  - $F = T^2$

T-Test or F-Test (ANOVA)

- Although we previously used a t-test
  - $T = \frac{\hat{\beta}_1 - \beta^0_0}{se(\hat{\beta}_1)} \sim t_{n-2}$
  - To test $H_0 : \beta_1 = \beta^0_0$ against $H_A : \beta_1 \neq \beta^0_0$
- We can also use the F-test as before
  - $F = \frac{\frac{SSReg}{1}}{\frac{SSE}{n-2}} \sim F_{1,n-2}$
  - When H_0 is true
- This is due to the relationship $F = T^2$

Hypothesis Testing for Regression in R

- T-test
- We use the lm function to create the linear model, and use the summary function to get details as before
  - If you want a one sided p-value use pt function with lower.tail = FALSE
  - Or if β_0^0 for H_0 isn't 0, find T using the equation and use it in the pt function

- F-test
- Use ANOVA as we have previously
  - Where the output will have:
  - Coefficients: Explanatory and Residuals
    - For each: Df, Sum Sq (SSR and MSR), Mean Sq (SSE and MSE), F Value and PR(>F) (P-value)
  - If you want a one sided p-value use pf function with lower.tail = FALSE

- Both will give the exact same, or very close to the exact same P-value

Examples of T-Test and F-Test Hypothesis Testing

Step	T-Test	F-Test (ANOVA)
1 Hypotheses	$H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$	$H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$
2 Test statistic	$T = 9.478$	$F = 84.824$
3 Sampling distribution	$T \sim t_{df=(n-2)} = t_8$	$F \sim F_{df=(1,n-2)} = F_{df=(1,8)}$
4 P-value	$P(\|t_8\| > 9.478) = 2^*pt(9.478, 8)$ $= 1.27e-05$	$P(F_{1,8} > 89.824) = 2^*pf(89.824, 1, 8)$ $= 1.265e-05$
5 Decision	As the p-value is very small, we reject H_0	As the p-value is very small, we reject H_0
6 Conclusion	We conclude there is a positive relationship	We conclude there is a positive relationship

Coefficient of Determination

- Describes the fraction of the variation in the values of y that is explained by the least squares regression line
- It is a number between 0 and 1, which represents the percentage between 0 and 100%
- The higher the value of the coefficient, the better the least squares regression line explains the variation in the data
- R^2 is the correlation coeffidient (r) squared
- In R:
  - This is in the summary of lm as Multiple R-Squared

Read TXT Files in R

- Need to manually call them in
  - *variable* = read.table("*name.txt*", header = T)