STAT1006 Week 10 Cheat Sheet

Lisa Luff

11/2/2020

Contents

More MLR **2**
ANOVA for MLR . . . 2
F Testing in Model Comparision . . . 2
Diagnostics . . . 3
Leverage . . . 3
Influential Observations . . . 3

Polynomial Regression **3**

Categorical/Indicator Variables **4**

In R: **5**

More MLR

ANOVA for MLR

- Hypotheses:
  - $H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$
  - H_A : at least one of the $\beta_i \neq 0$
    - To test this, we must have a comparison, so we use a linear model of y against 1
- ANOVA table:

Source of variation	Degrees of freedom (df)	Sum of Squares (SS)	Mean Squares (MS)	F
Regression	p (DFR)	SSReg (SSR)	$\frac{SSReg}{p}$ (MSR)	$\frac{MSR}{MSE}$
Residual	$n - p - 1$ (DFE)	SSE	$S^2 = \frac{SSE}{n-p-1}$ (MSE)	
Total	$n - 1$ (DFT)	SST		

- Use the ANOVA table the same as with SLR

F Testing in Model Comparision

- Competing models:
  - **Model 1 -** $H_0 : \boldsymbol{Y} = \boldsymbol{X_1 \beta_1} + \boldsymbol{\varepsilon}$
  - **Model 2 -** $H_A : \boldsymbol{Y} = \boldsymbol{X_1 \beta_1} + \boldsymbol{X_2 \beta_2} + \boldsymbol{\varepsilon}$
    - Where $\boldsymbol{X_2 \beta_2}$ can include more than one explanatory variable
- We are testing if the additional variables in Model 2 are necessary
- If Model 2 adds c parameters:
  - The F-statistic compares the Mean Squares of 1 & 2, which is the Sum of the Squares of 1 & 2 / c
  - $F = \frac{\left(\frac{SSE_1 - SSE_2}{c}\right)}{MSE_2} = \frac{\frac{SS_{1,2}}{c}}{\hat{\sigma}_2^2} \sim F_{c,n-q-c}$
- Adding more variables will **always** decrease SSE, for need to find if it's a significant difference
  - Test this is a **partial F-test**
    - Like all others, it involves ratios of SS
- Partial F-Test:
  - Compares a "small" model and a "big" model
    - The small model is model 1 from before
    - The big model is model 2 from before
  - Comparing c>1 number of differences
    - $H_0 : \beta_c 1 = \beta_c 2 = ... = 0$

    - $H_A : \beta_c 1, \beta_c 2, ...$ are not all 0
      - Where $\beta_c n$ is the n^{th} additional variable in the big model
    - If rejecting H_0, interpret as we need to add at least 1 of the c variables
  - Comparing $c = 1$ number of differences
    - $H_0 : \beta_{testing} = 0$, given estimates of all other coefficients

    - $H_A : \beta_{testing} \neq 0$, given estimate of all other coefficients
    - If rejecting H_0, interpret as evidence to keep the testing variable in the model

Diagnostics

- Use:
  - Histogram and QQ plot of standardised results
  - Plot of standardised residuals against **each** of the explantory variables
  - Plot of standardised residuals against fitted values
  - Plot of standardised residuals against actual values
- Standardised residuals are used as they have an actual mean of 0 with equal variance
- Best way to standardise residuals to have constant variance:
  - $r_i = \frac{\hat{e}_i}{s\sqrt{1-h_{ii}}}$, where
    * r_i - Is the standardised residual for $\boldsymbol{X}_i$
    * $\hat{e}_i$ - Is the error for $\boldsymbol{X}_i$
    * s is the standard deviation
    * h_{ii} is the i^{th} diagonal element of the hat matrix $\mathbf{H}$

Leverage

- Use the i^{th} diagonal element of the hat matrix $\mathbf{H}$, h_{ii} to measure leverage
  - Rule of thumb: When a point has a $h_{ii} > \frac{2(p+1)}{n}$, it is a high leverage point

Influential Observations

- Easiest way to see if a data point is influential is to remove one at a time
- Help identify possible data points for removal with Cooks distance
  - Measure of influence by reflecting when a point is a large residual AND a large leverage
  - $D_i = \frac{r_i^2}{2}\frac{h_{ii}}{1-h_{ii}}$
    * Rule of thumb: $2(p+1)/(n-2)$
- Once identified, remove and compare Model 1 with all observations, vs Model 2 without influential observations

Polynomial Regression

- A polynomial regression takes the form:
  - $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$
- It is still a linear model, because the parameters are linear
  - If you assign z as x^2, then
    * $Y = \beta_0, +\beta_1 x + \beta_2 z + \varepsilon$
  - It is a MLR, where the point fall in curve, but on a straight plane in 3D space
- Everything else is the same as a regular MLR

Categorical/Indicator Variables

- When you have a single variable, but that variable can be broken up into different categories, turn it into an MLR
- To turn it into an MLR (for two categories):
  - Combine into a single MLR: $y_i = \beta_0 + \beta_i x_i + \beta_2 z_i + \beta_3 x_i z_i + \varepsilon_i$, where
    - z_i is the indicator variable
    - The indicator variable **must** be binary (0 or 1)
  - For category1, $z_i = 0$, so
    - $E(Y|X) = \beta_0 + \beta_1 x$
  - For category2, $z_i = 1$, so
    - $E(Y|X) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x$
  - If p-values for both are large, no need to separate them, can keep as single model
- Or you can look it as:
  - Model will be: $y_i = \beta_0 + \beta_i x_i + \beta_2 z_{1i} + \beta_3 z_{2i} + \varepsilon_i$, where
    - $z_1 = 1$ for category1, $z_2 = 0$ otherwise
    - $z_2 = 1$ for category2, $z_1 = 0$ otherwise
- T-tests:
  - T-test for linear model
    - Tells you if there is a significant difference between the values of y in each category compared to those of the other categories
  - T-test for the linear regression
    - With equality of variances, is equivalent to fitting $y_i = \beta_0 + \beta_1 z_i + \varepsilon_i$
    - Where z is a binary value associated with a categorical variable
    - Tells you the same as the t-test for linear model
- Comparing more than 2 categorical variables:
  - For two variables:
    - Model will be: $y_i = \beta_0 + \beta_i x_i + \beta_2 z_{1i} + \beta_3 z_{2i} + \varepsilon_i$, where
      - $z_1 = 1$ for category1, $z_2 = 0$ otherwise
      - $z_2 = 1$ for category2, $z_1 = 0$ otherwise
  - There will be 1 less $\beta_n z_{mi}$ than the total number of variables
  - If complex, might be:
    - $y_i = \beta_0 + \beta_i x_i + \beta_2 z_{1i} + \beta_3 z_{2i} + \beta_4 x_i z_{1i} + \beta_5 x_i z_{2i} + \varepsilon_i$
  - R does this automatically

In R:

- **Notes**:
  - ~. is shorthand for all other explanatory variables when creating MLR's with lm()
  - subset = -c(n) is a subset excluding n in lm()
- ANOVA -
  - $H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$
  - H_A : at least one of the $\beta_i \neq 0$
    - Create lm of each combo
      - This might just be lm(y ~ 1, data = *data*)
    - anova(*lm1*, *lm2*)
    - OR use package rms: anova(rms::ols(y ~ $x_1, x_2, ..., x_i$, data = *data*)
- F-test for comparing models
  - **Model 1 -** $H_0 : \boldsymbol{Y} = \boldsymbol{X_1\beta_1} + \boldsymbol{\varepsilon}$
  - **Model 2 -** $H_A : \boldsymbol{Y} = \boldsymbol{X_1\beta_1} + \boldsymbol{X_2\beta_2} + \boldsymbol{\varepsilon}$
    - Use ANOVA or summary
- Partial F-test to compare models with variables removed
  - $H_0 : \beta_c 1 = \beta_c 2 = ... = 0$

  - $H_A : \beta_c 1, \beta_c 2, ...$ are not all 0
    - Use ANOVA or summary
- Leverage -
  - Plotting:
    - For plotting with actual data:
      - plot(1:n, hatvalues(*lm*), xlab = x_i, ylab = leverage)
      - abline(h = $\frac{2(p+1)}{n}$)
    - For plotting with standarised variables:
      - plot(hatvalues(*lm*), stdres(*lm*), xlab = leverage, ylab = standardised residuals)
      - abline(v = $\frac{2(p+1)}{n}$)
      - abline(h = 0)
  - Get hat values with:
    - For actual data:
      - identify(1:n, hatvalues(*lm*), labels = rownames(*data*))
    - For standardised data
      - identify(hatvalues(*lm*), rstandard(*lm*), labels = rownames(*data*))
- Influential Observations (Cooks Distance) -
  - Plotting:
    - Cooks distance against actual data:
      - plot(cooks.distance(*lm*), xlab = x_i, ylab = cooks distance)
    - Cooks distance against leverage:
      - plot(hatvalues(*lm*), cooks.distance(*lm*), xlab = leverage, ylab = cooks distance)
      - abline(v = $\frac{2(p+1)}{n}$, h = $2(p+1)/(n-2)$)
  - Get Cooks distance values:
    - identify(1:n, cooks.distance(*lm*), label = rownames(*lm*))
      - **Note**: Must be run in the console

- Polynomial Regression -
  - Linear model:
    - lm($y \sim x$ + Ix^2, data = *data*)
  - Fitted model:
    - *lm*$fitted
  - Summary:
    - summary(*lm*), where
      - Intercept = β_0
      - $x = \beta_1$
      - I(x^2) = β_2
- Categorical/indicator variable -
  - Linear model:
    - lm(formula = $y \sim x$ + *category* + x*category, *data* = data*) OR
    - lm($y \sim x$ + *category* + x:*category*, data = *data*), where
  - To only get intercept:
    - lm($y \sim x$ + *category*, data = *data*)
  - Summary:
    - summary(*lm*), where
      - Intercept = β_0 - Category 1 intercept
      - $x = \beta_1$ - Category 1 slope
      - *categoryn* = $\beta_{n+1} z_{ni}$ - Category n intercept
      - x:*categoryn* = $\beta_m x_i z_{ni}$ = category n slope
  - T-test of the linear model
    - t.test($y \sim$ *category*, var.equal = TRUE, data = *data*), where
      - p-value is if there is significant difference between observations between variables
  - T-test of the linear regression
    - lm($y \sim$ *category*, data = *data*)
    - summary(*lm*), where
      - p-value will be exact same as t-test of the linear model