STATS 412

Twenty-first Class Note

In Son Zeng

03 December, 2018

My Office Hour:

The office hours this week are on **16:30 - 18:00 Tuesday** and **13:30 - 15:00 Friday**, at **USB 2165**. You may check the campus map to get to my office. I am prepared for your questions, so please feel free to come to my office hours.

Exam 3 Updates:

I will be updating the revision for the final exam day by day, so check closely to the update for the final exam.

• The revision contents are various: concepts, test convention, sample questions and solutions. Hope you find the materials helpful for performing better in the final exam.

• If our class can improve the handwriting for the rest of the homeworks, I can have more time to write the revision day by day to improve the general quality. Also, when I have more time, I can be more efficiently respond questions on Piazza and provide better designed questions and solutions (with video).

Homework Grading Policy:

Please include the final answer for each homework question. If the final answer is not included, you will risk 0.5 points for each missing part.

Homework 11 Reminder:

• **Check Class Note 20**.

Key Concepts During Lecture 21:

Correlation: As suggested by the lecture notes, correlation measures the strength and direction of the linear relationship between two quantitative variables. We can express the correlation coefficient r as the following ways:

$$\boxed{r = \frac{cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\cdot\left(\frac{y_i-\bar{y}}{s_y}\right)} \quad (1)$$

Strength and Direction of Correlation: The range of r is $-1 \leq r \leq 1$. If $r = 1$, we say perfectly positive correlation. If $r = -1$, we say perfectly negative correlation. Generally, $r \geq 0.8$ or $r \leq -0.8$ might indicate a strong correlation in engineering field, while $r \geq 0.6$ or $r \leq -0.6$ might indicate a strong correlation in psychology or medical research.

Plot: Given a plot of the independent variable as x-axis and dependent variable as y-axis, we can draw the least square line, which is given by $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Then, we can draw a horizontal line indicating the value $\bar{y}$. Based on these two lines, we can see that the total variability in y for each observation is the quantity $y_i - \bar{y}$. In addition, the variability explained by x is $\hat{y}_i - \bar{y}$, while the unexplained variability is $e_i = (y_i - \bar{y}) - (\hat{y}_i - \bar{y}) = y_i - \hat{y}_i$.

JMP outputs in Lecture note: It is crucial to understand what the JMP outputs represent. Firstly, the Linear Fit shows the least square line. Secondly, Parameter Estimates gives the regression coefficients β_0 for the intercept term and β_1 for the regression effect (slope).
The Summary of Fit gives the RSquare, which represents the proportion of variability in the model captured by the model. It also gives the RMSE (square root os the Mean Square for Error in Analysis of Variance) and number of observations.

JMP outputs, ANOVA: Finally, JMP gives the Analysis of Variance table. First, JMP gives the degree of freedoms: total degree of freedom, degree of freedom of error and degree of freedom of model. It also gives Total: the total sum of squares (SST), Model: regression sum of squares (SSR) and Error: error sum of squares (SSE), and as a matter of fact, $SST = SSR + SSE$. Then, JMP gives the mean square for regression and error, which is defined by sum of squares divided by the corresponding degree of freedom.
Finally, the F-ratio is given by mean square for regression divided by mean square error, and the p-value gives the result of the following hypothesis (if we have multiple predictors): $H_0 : \beta_1 = \beta_2 = ... = \beta_{p-1} = 0$ vs

$H_1 : \beta_j \neq 0$ for at least one $j = 1, 2, ..., p-1$, where p is the number of parameters in the model, including the intercept and $p-1$ predictors. Similar to what we have learnt in chapter 6, we specify the significance level α, then if the p-value is lower than the significance level, we reject the null hypothesis. We have sufficient evidence to conclude that not all regression effects $\beta_j, j = 1, 2, ...p-1$ are 0.

JMP outputs for Exam 3: If the R^2 is not given during the exam, you can find by looking at the Analysis of Variance. Specifically, $R^2 = \frac{SSR}{SST} = \frac{Sum\ of\ Square\ Model}{Sum\ of\ Square\ C.Total}$

For more information about the output, please take time to refer to: http://facweb.cs.depaul.edu/sjost/csc423/documents/f-test-reg.htm https://www.jmp.com/en_us/statistics-knowledge-portal/what-is-regression/interpreting-regression-results.html

Coefficient of Determination: Coefficient of Determination is the R^2 value. As mentioned above, the R^2 value is given as follows:

$$\boxed{R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}} \tag{2}$$

Interpretation: If $R^2 = 0.6$, we claim that 60% of the observed variation in the model can be explained by the simple linear regression relationship between **(the independent variable)** and **(the dependent variable)**.

Relationship to correlation: We have a nice formula between r and R^2. We have $R^2 = r^2 \rightarrow r = \sqrt{R^2}$. However, we need to be careful here. If the regression line demonstrates positive correlation (reflected by slope), then we take the positive square root, whereas if the regression line shows negative correlation, we take the negative square root. Remember, both R^2 and r only work for evaluating linear relationships.

Simple Linear Model: The simple linear model (true model for least-squares regression) is given by:

$$\boxed{y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad or \quad Y = \beta_0 + \beta_1 X + \epsilon} \tag{3}$$

, where y_i is the observed value of the response variable, β_0 is the least-squares y-intercept, β_1 is the least-squares slope for the model. The ϵ_i is the error term with the ith measurement.

Normal Assumption for Linear Model:

• 1) We need the observations are independent.

• 2) We need the errors $\epsilon_1, \epsilon_2, ..., \epsilon_n$ are random and independent. In particular, the magnitude of any error term does not influence the value of the other error terms. Again, if the error terms appear to be monotone increasing or decreasing, U-shape or inverted-U-shape, the assumption for linear models is violated.

• 3) We need the error terms to be normally distributed, with mean 0 and with same variance (homoscedasticity). Notationally speaking, we need $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$

If the normal assumption is fulfilled, we can further have the response variable y_i follows normal distribution. Notationally speaking, $y_i \overset{\text{i.i.d.}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$, for $i = 1, 2, ..., n$.

Last Comment:

Please inform me to fix the typos and grammatical mistakes if they exist. It is a great practice of writing and I appreciate your help!