IST772 Chapter Notes Template: After Completing Please Submit as a PDF.
Originality Assertion: By submitting this file you affirm that this writing is your own.

Name: Wanyue Xiao
Date: October 12, 2020
Chapter Number: #9
Title of Chapter: Linear Multiple Regression

Linear Regression

General Linear Model is a powerful framework which compares how one or more variables affect different continuous variables. The model includes the ANOVA, Pearson correlation, and regression analysis.

Some basic concepts:

1. **Linearity**: a relationship that could be graphically drawn as a straight line.
2. **Bivariate Normality**: two random variables are said to be bivariate normal of aX + bY has a normal distribution.
3. **Influential Outliers**: linear multiple regression model is sensitive to definite outliers, especially for those points that are at the end of the bivariate distribution.
4. **Error of prediction**: if one wants to find the line that fits shape of the scatter plot, one needs to minimize the sum of error of prediction, which is the sum of distances between each point and the best-fitting line. The prediction errors of a best-fitting line should be normally distributed since the majority of errors need to be 0 or pretty small. Generally, the smaller the sum of prediction errors are, the better the prediction about the best-fitting line.

Sum of squared errors of prediction: to avoid the negative prediction errors cancel out with the positive errors, one can use the sum of squared errors of prediction to estimate the best-fitting line. If there is no relationship between dependent variables and independent variables, the sum of the squared errors reaches the maximum. If we minimize the sum of squared errors of prediction, it is possible to find the best-fitting line in a bivariate relationship.

Matrix Algebra: a method that could be used to find the slope and intercept that follows the least-squares criterion. One can use lm() command to create a linear regression model. The outcome of such a model consists of three sections, including residuals (equal to error of predictions), coefficient section that shows the estimate of intercept and the estimates of coefficient of each independent variable, and the summary statistics.

- One can check the residuals via rediduals() command. If these residuals are notably abnormal, the underlying relationship between independent variables and dependent variable is not linear.
- Specifically, coefficient section also contains information of std error, t-value, and P(>|t|). The t-value shows the **t-test of the null hypothesis** which indicates each estimates of coefficient is equal to zero. The standard errors could be used to show the estimated spread of the sampling distribution around those estimated points.
- The effect size of regression model is the **R-squared value** which is also could be interpreted as the proportion of variability in the dependent variables that is explained by an independent variable or variables in a regression model.
- The summary statistics contains Residual Standard Error, df value, multiple R-squared, Adjusted R-squared, F-statistics, and p-value. The **formula of df calculation** should be

df = TotalObservation – (numbers_independent_variables) – 1. Here, n df is lost for calculating the slope of variables while another 1 df is lost for calculating the Y-intercept.

- The **F-statistics of Residual**, testing the null hypothesis that R-squared is equal to 0.

Multicollinearity: at least one independent variable is/are highly correlated with another one, leading to an inaccurate model result.

Making Sense of Adjusted R-Squared

Formula: $$R^2_{\text{adj}} = 1 - \left(\frac{SS_{\text{res}} / (n-p-1)}{SS_{\text{tot}} / (n-1)} \right)$$

Plain old R-squared: could be used to interpretate the proportion of variance accounted for in the dependent variable. Since the old R-squared used biased variance estimators, the result of R-squared will overestimated than the true population over the long run.
To calculate the adjusted R-squared, one should instead use unbiased estimators for the calculation of those two variances. The result equals to 1 minus the ratio of two variances.

- Denominator: the total sum of squares in dependent variable equals to the sum of squares divided by (sample size - 1) df.
- Numerator: the sum of squares for the residual equals to the sum of squares divided by (sample size - the number of predictors -1) df.
- Residual variance also known as error variance.
- If the sample size is large enough, a relatively small number of predictors won cause significant reduction when it comes to the calculation of df. Therefore, the adjusted *R*-squared will be virtually identical to the plain old *R*-squared.

Exercise Review

2.

```
> cor(myCars)
            mpg        cyl       disp         hp       drat         wt
mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594
cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.6999381  0.7824958
disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.7102139  0.8879799
hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.4487591  0.6587479
drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.0000000 -0.7124406
wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.7124406  1.0000000
```

Only the correlation between mpg and drat is positive while the rests are negative. wt has the lowest correlation value compared with the rests. Therefore, one can assume that wt might be the best predictor.

3.

IST772 Chapter Notes Template: After Completing Please Submit as a PDF.
Originality Assertion: By submitting this file you affirm that this writing is your own.

```
Call:
lm(formula = mpg ~ wt + hp, data = myCars)

Residuals:
   Min     1Q Median     3Q    Max
-3.941 -1.600 -0.182  1.050  5.854

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
wt          -3.87783    0.63273  -6.129 1.12e-06 ***
hp          -0.03177    0.00903  -3.519  0.00145 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.593 on 29 degrees of freedom
Multiple R-squared:  0.8268,    Adjusted R-squared:  0.8148
F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```

The values of F-statistics and p-value could be used to tell whether the overall R-squared was significant nor not. Given that the p-value is far lower than 0.05, one can reject the null hypothesis and claim the R-squared is significant. The result is 0.8268, which is quite satisfactory since nearly 83% of the variance of mpg could be explained by wt and hp.
The estimate of intercept, wt, and hp is 37.23, -3.88, -0.03 respectably. The Pr(>|t|) value of intercept, wt, and hp is 2e-16, 1.12e-06, and 0.00145. All those values are significantly lower than 0.05, indicating that the null hypothesis of each will be rejected.

4.
37.22727 + (-3.87783*3) + (-0.03177 * 110) = 22.09908

5.

```
Bayes factor analysis
--------------
[1] wt + hp : 788547604 ±0%

Against denominator:
  Intercept only
---
Bayes factor type: BFlinearModel, JZS
```

The Bayes factor of 788547604, very strong positive evidence in favor of the alternative hypothesis. In this case the alternative hypothesis is that the coefficients on hp and tw are nonzero. The result strength my conclusion.

6.

IST772 Chapter Notes Template: After Completing Please Submit as a PDF.
Originality Assertion: By submitting this file you affirm that this writing is your own.

```
Iterations = 1:10000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variabl
   plus standard error of the mean:

         Mean        SD  Naive SE Time-series SE
mu   20.08913  0.485482 4.855e-03      4.945e-03
wt   -3.79407  0.659473 6.595e-03      6.595e-03
hp   -0.03085  0.009474 9.474e-05      9.672e-05
sig2  7.52406  2.427966 2.428e-02      2.909e-02
g     3.76416 10.016231 1.002e-01      1.002e-01

2. Quantiles for each variable:

         2.5%      25%      50%      75%    97.5%
mu   19.11771 19.77036 20.09029 20.40793 21.04130
wt   -5.09691 -4.22206 -3.79818 -3.35769 -2.46817
hp   -0.04961 -0.03713 -0.03086 -0.02483 -0.01194
sig2  4.37407  5.96600  7.16244  8.63747 12.69463
g     0.35501  0.93856  1.67722  3.38008 19.88332
```

The coefficient of wt and hp is similar with the coefficient obtained from the q3. However, the intercept changed significantly (37.22 vs 20.08). The 95% HDI for the coefficient of hp and wt does not overlaps with 0, providing evidence that the population value of that coefficient of each predictor does credibly differ from 0.

7.

Variance Inflation Factors

Description

Calculates variance-inflation and generalized variance-inflation factors for linear, generalized linear, and other models.

Usage

```
vif(mod, ...)

## Default S3 method:
vif(mod, ...)

## S3 method for class 'merMod'
vif(mod, ...)
```

Arguments

mod	for the default method, an object that responds to coef, vcov, and model.matrix, such as an lm or glm object.
...	not used.

The vif will take output of lm() command and return value of each predictor. A VIF larger than 5 or 10 is large enough to indicate that the multicollinearity is strongly suggested. However, this in general does not degrade the quality of predictions. According to the Rdocument of vif() command, “If the VIF is larger than 1/(1-R2), where R2 is the Multiple R-squared of the regression, then that predictor is more related to the other predictors than it is to the response”.

8.

IST772 Chapter Notes Template: After Completing Please Submit as a PDF.
Originality Assertion: By submitting this file you affirm that this writing is your own.

```
> vif(lm(mpg ~ wt + hp,data=myCars))
       wt       hp
1.766625 1.766625
```

1.766625 means that the variance of wt (or hp) is 76% bigger than what we would expect if there was no multicollinearity.

R Code Fragment and Explanation

Variance inflation factor, aka vif(), could be used to evaluate or detect the multicollinearity of multiple regression model. Ranging from 1 upward, vif represent the percentage of variance is inflated for each coefficient.

Rule of thumb for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

General rule: However, this in general does not degrade the quality of predictions. According to the Rdocument of vif() command, "If the VIF is larger than 1/(1-R2), where R2 is the Multiple R-squared of the regression, then that predictor is more related to the other predictors than it is to the response" (Rdocument, 2020). Besides, "Taking the square root of the VIF tells you how much larger the standard error of the estimated coefficient is respect to the case when that predictor is independent of the other predictors (Rdocument, 2020)".
For more information, please check (https://www.rdocumentation.org/packages/regclass/versions/1.6/topics/VIF).

Example:

```
> vif(lm(mpg ~ wt + hp,data=myCars))
       wt       hp
1.766625 1.766625
```

Rdocument 2020, *'VIF'*, R documentation, https://www.rdocumentation.org/packages/regclass/versions/1.6/topics/VIF

Question for Class

Is there any other method that could be used to detect multicollinearity?