IST772 Summary Template: Chapter 8 – Linear Multiple Regression
Originality Assertion: By submitting this file you affirm that this writing is your own.

Name: Hendi Kushta
Date: 03/13/2023

****Important**** **Copying and/or pasting anything from the textbook will not be acceptable for your chapter notes submissions. You must write your notes in your own words and generate your own code, results, and graphs in R. This is what forces your brain to process the material that you read.**

INTRODUCTION

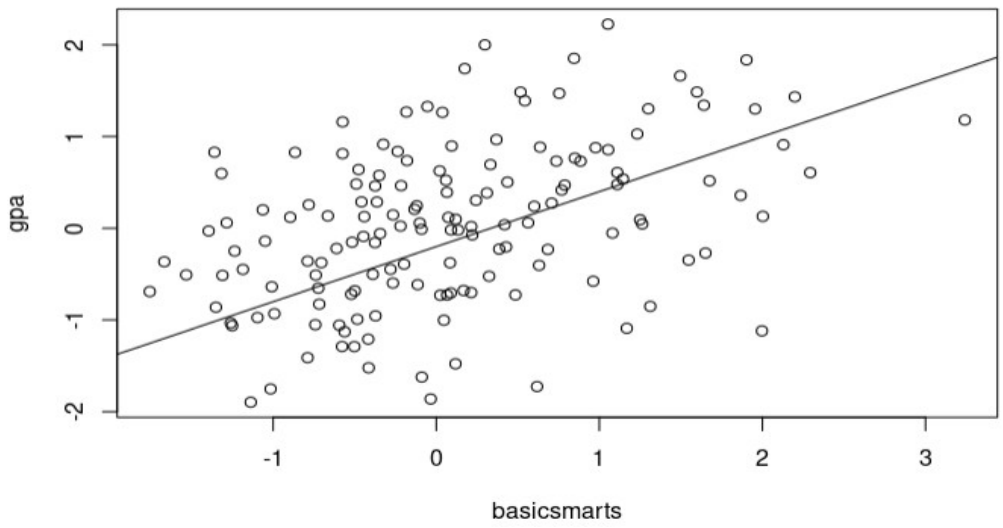
The general linear model is a family of analysis techniques that models one dependent variable as a function of one or more independent variables. ANOVA and Pearson correlation are part of this family. The model uses coefficients to represent the importance of each independent variable and can be used for forecasting, such as predicting a student's GPA based on their hard work, basic smarts, and curiosity. This can be useful in developing an educational early warning system to help students improve their grades.
The B coefficients in the general linear model can be used to understand the relative importance of predictors and select metrics in both research and practice. Regression analysis can have value beyond creating a forecasting equation. Understanding regression is like discovering a best-fitting line between two metric variables, and we can explore it through creating correlated variables and multivariate regression on random variables.
In this example, a sample of 150 observations on three independent variables (hard work, basic smarts, and curiosity) is created using random normal distribution. A dependent variable (GPA) is synthesized by adding these variables together along with a random noise component. The standard deviation of GPA is confirmed to be approximately one. The example then focuses on exploring the relationship between the dependent variable and one of the predictor variables, hard work.

```{r}
set.seed(123)
hardwork <- rnorm(150)
basicsmarts <- rnorm(150)
curiosity <- rnorm(150)

# add a random noise
randomnoise <- rnorm(150)
# We divide each input variable by 2 because it is the square root of 4
# and we have four inputs into creating our fake dependent variable.
gpa <- hardwork/2 + basicsmarts/2 + curiosity/2 + randomnoise/2
sd(gpa)
```

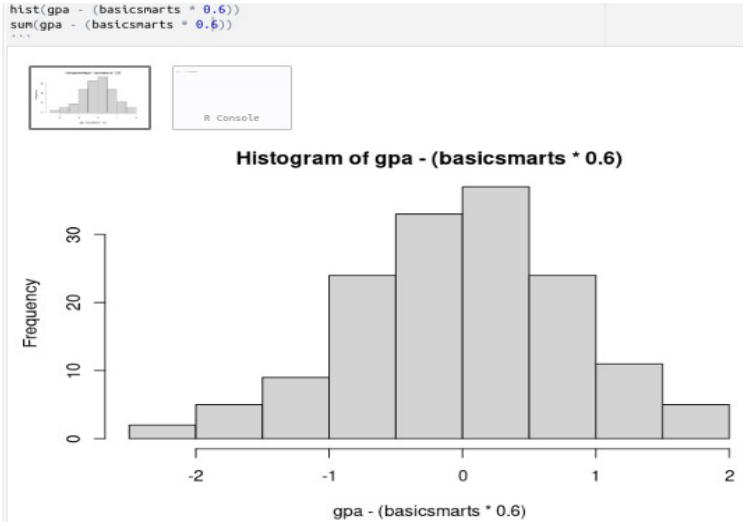
```
[1] 0.8751566
```

IST772 Summary Template: Chapter 8 – Linear Multiple Regression
Originality Assertion: By submitting this file you affirm that this writing is your own.

Scatterplots arein analyzing bivariate relationships between variables. There are three important characteristics to look for in a scatterplot: linearity, bivariate normality, and influential outliers. Below is presented a scatterplot of the variables basicsmarsts and GPA and notes that it exhibits linearity, bivariate normality, and the absence of outliers. Finally, I have added a line to the scatterplot based on the guess for the slope and intercept, which is 0.6 and -0.2, respectively.



It is calculated errors of prediction for the made-up data points and the "guesstimated" best-fitting line using the expression gpa - (basicsmarsts * 0.6). The resulting histogram of prediction errors shows a normal distribution centered around 0, which is what one would expect for a best-fitting line. The sum of the errors is a little lower than -1.8, which is not bad considering there are 150 points in the dataset.



IST772 Summary Template: Chapter 8 – Linear Multiple Regression
Originality Assertion: By submitting this file you affirm that this writing is your own.

The model specification syntax used in lm() is similar to that of aov(). The dependent variable is listed first, followed by independent variables separated by the ~ character, and the model is constructed using + or * depending on the type of model. The article also includes sample code for constructing a data frame and using lm() with the model syntax.

```r
# Put everything in a data frame first
educdata <- data.frame(gpa, hardwork, basicsmarts, curiosity)
regOut <- lm(gpa ~ hardwork, data=educdata)    # Predict gpa with hardwork
summary(regOut)                       # Show regression results
```

```
Call:
lm(formula = gpa ~ hardwork, data = educdata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.75814 -0.60489  0.01609  0.51340  1.97768

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.05143    0.06716   0.766    0.445
hardwork     0.32325    0.07093   4.557 1.08e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8223 on 148 degrees of freedom
Multiple R-squared:  0.1231,    Adjusted R-squared:  0.1171
F-statistic: 20.77 on 1 and 148 DF,  p-value: 1.077e-05
```

The output above is the result of running a linear regression model using the lm() function in R. The model is specified with the formula gpa ~ hardwork, where gpa is the dependent variable and hardwork is the independent variable.

The first section, "Call", shows the syntax of the function call used to create the model.

The "Residuals" section shows the minimum, 1st quartile, median, 3rd quartile, and maximum values of the residuals (the difference between the observed values and predicted values) of the model.

The "Coefficients" section shows the estimated coefficients of the model, which include the intercept and the slope of the hardwork variable. The "Std. Error" column shows the standard errors of these coefficients, while the "t value" column shows the t-statistics, which is calculated by dividing the coefficient estimate by its standard error. The "Pr(>|t|)" column shows the p-values of the t-tests for each coefficient.

IST772 Summary Template: Chapter 8 – Linear Multiple Regression
Originality Assertion: By submitting this file you affirm that this writing is your own.

The "Residual standard error" shows the residual standard error, which is an estimate of the standard deviation of the error term in the model. The "Multiple R-squared" and "Adjusted R-squared" show the proportion of the variance in the dependent variable that is explained by the independent variable. Finally, the "F-statistic" and "p-value" show the overall significance of the model, which tests whether the independent variable is significantly related to the dependent variable.

BOX ON P.169: MAKING SENSE OF ADJUSTED R-SQUARED

Degrees of freedom is a statistical concept that represents the number of independent pieces of information available in a sample. It plays a crucial role in estimating population parameters from a sample, as it helps to account for the variability that arises due to chance. When we calculate sample-based estimates, such as the sample variance, we have to take into account the degrees of freedom in the calculation to ensure that our estimate is not biased. Using n - 1 instead of n in the denominator of the sample variance calculation is a common way to adjust for the degrees of freedom, as it ensures that our estimate is not too low over the long run.

Adjusted R-squared is a modified version of the R-squared statistic, which is commonly used to measure the proportion of variance in the dependent variable that is explained by the independent variables in a regression model. While R-squared uses biased variance estimators, adjusted R-squared uses unbiased estimators to account for the degrees of freedom in the sample. The proper degrees of freedom for the adjusted R-squared calculation is n - p - 1, where n is the sample size and p is the number of predictors. This penalty for having more predictors in the model helps to account for the increased chance of capitalizing on chance with each additional predictor.

Adjusted R-squared is a statistical measure that is calculated using the formula: Adjusted R-squared = 1 - (residual sum of squares / total sum of squares). It is used to determine the accuracy of a regression model by taking into account the number of predictors and sample size. The degrees of freedom for the residual variance decrease as the number of predictors increases, resulting in a lower adjusted R-squared value. Adjusted R-squared is important to report in situations where the sample size is small relative to the number of predictors, but it is not necessary to report it in situations where the sample size is large. Reporting adjusted R-squared provides a more realistic picture of the model's performance, regardless of sample size or number of predictors.

IST772 Summary Template: Chapter 8 – Linear Multiple Regression
Originality Assertion: By submitting this file you affirm that this writing is your own.

THE BAYESIAN APPROACH TO LINEAR REGRESSION

The Bayesian approach to inference provides direct and detailed probability information about each parameter of a model. It starts with assumptions about priors, modifies them with data, and ends up with posterior probability distributions for each coefficient. The Cauchy distribution is a commonly used prior distribution for modeling standardized coefficients in regression analysis. It has a standard deviation of 1, which is weakly informative and represents diffuse knowledge about the likelihood of various beta values. The lmBF() function in the BayesFactor package can be used to conduct linear multiple regression analysis with posterior sampling using the Markov chain Monte Carlo technique. The output includes means and 95% highest density intervals (HDIs) for each coefficient's posterior distribution.

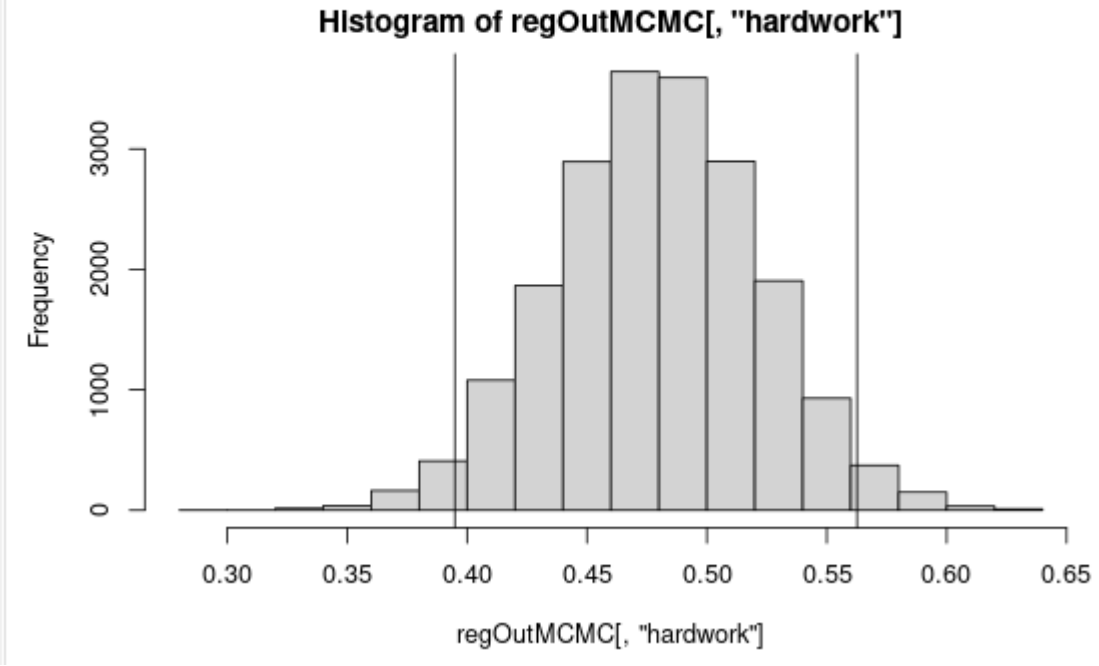
```r
regOutMCMC <- lmBF(gpa ~ hardwork + basicsmarts + curiosity,
data=educdata, posterior=TRUE, iterations=20000)
summary(regOutMCMC)
```

```
0% 10   20   30   40   50   60   70   80   90   100%
|----|----|----|----|----|----|----|----|----|----|
**************************************************|

Iterations = 1:20000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 20000

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

               Mean      SD  Naive SE Time-series SE
mu          0.04312 0.03954 0.0002796      0.0002796
hardwork    0.47925 0.04271 0.0003020      0.0003020
basicsmarts 0.50276 0.04291 0.0003034      0.0003072
curiosity   0.49846 0.03879 0.0002743      0.0002743
sig2        0.23167 0.02770 0.0001958      0.0002062
g           1.20527 1.96513 0.0138956      0.0141678

2. Quantiles for each variable:

                2.5%     25%     50%     75%  97.5%
mu          -0.03515 0.01705 0.04302 0.06945 0.1210
hardwork     0.39490 0.45086 0.47945 0.50810 0.5626
basicsmarts  0.41933 0.47375 0.50254 0.53185 0.5875
curiosity    0.42226 0.47249 0.49868 0.52461 0.5739
sig2         0.18351 0.21211 0.22977 0.24881 0.2926
g            0.21000 0.44452 0.72261 1.27486 5.0107
```

The output shows the results of a Bayesian linear regression analysis, which was conducted using the lmBF function from the BayesFactor package. The output includes two sections: 1) empirical mean, standard deviation, naive standard

error, and time-series standard error for each variable, and 2) quantiles for each variable.
The first section provides summary statistics for each of the model parameters, including the intercept (mu), as well as the regression coefficients for hardwork, basicsmarts, curiosity, and sig2 (the residual variance). The "g" parameter represents the standard deviation of the Cauchy prior, which was set to 1. The mean, standard deviation, naive standard error, and time-series standard error are provided for each parameter.
The second section provides quantiles for each variable, including the 2.5th, 25th, 50th (median), 75th, and 97.5th percentiles. These quantiles can be used to estimate the range of plausible values for each parameter, as well as the level of uncertainty associated with each estimate. For example, the median value of hardwork is 0.479, but there is some uncertainty around this estimate, as indicated by the range from the 25th to 75th percentiles (0.451 to 0.508). Similarly, the residual variance sig2 has a median value of 0.230, with a range from 0.212 to 0.249.

```{r}
hist(regOutMCMC[,"hardwork"])
abline(v=quantile(regOutMCMC[,"hardwork"],c(0.025)), col="black")
abline(v=quantile(regOutMCMC[,"hardwork"],c(0.975)), col="black")
```

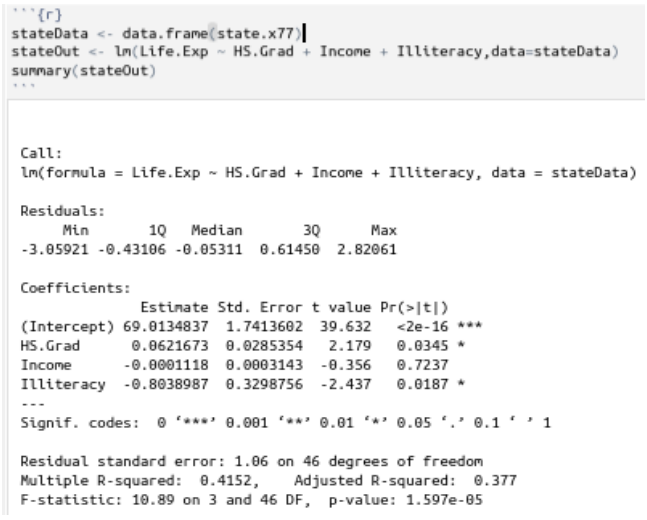


The code is analyzing the output of a Bayesian regression model for the variable "hardwork". The histogram shows the distribution of the estimates and is centered on 0.56. Two vertical lines are added to show the extent of the 95% highest density interval (HDI). The "sig2" estimates summarize the error in the

model, and smaller values indicate better quality of prediction. With the sig2 estimates, the code calculates the R-squared value for each model in the posterior distribution, which is equal to 1 minus sig2 divided by the variance of the dependent variable. The code then calculates a list of these R-squared values and displays the mean value.
The mean value of the distribution of R-squared values is calculated to be 0.75, which is slightly lower than the R-squared obtained from the traditional lm() model but almost equal to the adjusted R-squared obtained from that model. The likely range of possibilities for the predictive strength of the model is estimated to be between 0.67 and 0.81, with the most likely values surrounding 0.75. The distribution of R-squared values is asymmetric with a skew to the left due to a ceiling effect that squashes the right-hand tail as R-squared approaches 1. The importance of plotting the 95% highest density interval (HDI) lines to identify the most likely values of R-squared is emphasized. Finally, the Bayes factor for the model is obtained, which is expected to be very strong given all the evidence gathered so far.

A LINEAR REGRESSION MODEL WITH REAL DATA

The chapter discusses analyzing and interpreting regression results from a real dataset using the built-in data sets in R called state.x77. The dataset contains eight different statistics about the 50 US states, and the variable representing life expectancy is used as the dependent variable. The percentage of high school graduates, per capita income, and the percentage of the population that is illiterate are used as predictors. The state.x77 data object is converted to a data frame named stateData to make it easier to work with.



```r
stateData <- data.frame(state.x77)
stateOut <- lm(Life.Exp ~ HS.Grad + Income + Illiteracy,data=stateData)
summary(stateOut)
```

```
Call:
lm(formula = Life.Exp ~ HS.Grad + Income + Illiteracy, data = stateData)

Residuals:
     Min       1Q   Median       3Q      Max
-3.05921 -0.43106 -0.05311  0.61450  2.82061

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 69.0134837  1.7413602  39.632   <2e-16 ***
HS.Grad      0.0621673  0.0285354   2.179   0.0345 *
Income      -0.0001118  0.0003143  -0.356   0.7237
Illiteracy  -0.8038987  0.3298756  -2.437   0.0187 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.06 on 46 degrees of freedom
Multiple R-squared:  0.4152,    Adjusted R-squared:  0.377
F-statistic: 10.89 on 3 and 46 DF,  p-value: 1.597e-05
```

This is a summary of a linear regression model in R, with the following characteristics:

IST772 Summary Template: Chapter 8 – Linear Multiple Regression
Originality Assertion: By submitting this file you affirm that this writing is your own.

Dependent variable: Life.Exp (life expectancy)
Independent variables (predictors): HS.Grad (percentage of high school graduates), Income (per capita income), Illiteracy (percentage of illiterate population)
Model equation: Life.Exp = 69.013 + 0.062HS.Grad - 0.0001Income - 0.804*Illiteracy
Residuals (errors) have a minimum of -3.06 and a maximum of 2.82, with a mean of 0.
Coefficients for the intercept, HS.Grad, and Illiteracy are statistically significant ($p < 0.05$), while Income is not ($p > 0.05$).
Adjusted R-squared (a measure of how well the model fits the data) is 0.377.
F-statistic (a measure of how well the overall model fits the data) is 10.89, with a p-value of 1.597e-05 (very low, indicating that the model is a good fit for the data).

```{r}
stateOutMCMC <- lmBF(Life.Exp ~ HS.Grad + Income + Illiteracy,
data=stateData, posterior=TRUE, iterations=10000)
summary(stateOutMCMC)
```

```
0% 10   20   30   40   50   60   70   80   90   100%
|----|----|----|----|----|----|----|----|----|----|
**************************************************|

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

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

                Mean        SD  Naive SE Time-series SE
mu          7.088e+01 0.1604241 1.604e-03      1.604e-03
HS.Grad     5.545e-02 0.0298731 2.987e-04      3.180e-04
Income     -9.862e-05 0.0003181 3.181e-06      3.181e-06
Illiteracy -7.118e-01 0.3307936 3.308e-03      3.534e-03
sig2        1.196e+00 1.0078291 1.008e-02      1.037e-02
g           3.540e-01 0.8296901 8.297e-03      9.304e-03

2. Quantiles for each variable:

                2.5%        25%        50%        75%     97.5%
mu         70.5756371 70.7714828 70.8792125 70.9815291 71.180869
HS.Grad     0.0011997  0.0363874  0.0552827  0.0739043  0.110434
Income     -0.0007026 -0.0003037 -0.0001006  0.0001037  0.000503
Illiteracy -1.3605213 -0.9255835 -0.7117736 -0.4949562 -0.069081
sig2        0.7867567  1.0018064  1.1505873  1.3257918  1.808627
g           0.0462955  0.1141233  0.1987309  0.3613314  1.574610
```

The output shows the results of a Bayesian Markov Chain Monte Carlo (MCMC) analysis with 10,000 iterations, where the chain was thinned to only retain one

sample per iteration. The analysis was conducted on a single chain, and each chain had a sample size of 10,000.
The output provides summary statistics for six variables: mu, HS.Grad, Income, Illiteracy, sig2, and g. For each variable, the output reports the empirical mean, standard deviation, and standard error of the mean, as well as the time-series standard error.
The quantiles for each variable are also reported, including the 2.5%, 25%, 50%, 75%, and 97.5% quantiles. These quantiles provide a measure of the distribution of the posterior samples for each variable.

```{r}
stateOutBF <- lmBF(Life.Exp ~ HS.Grad + Income + Illiteracy,data=stateData)
stateOutBF
```

```
Bayes factor analysis
--------------
[1] HS.Grad + Income + Illiteracy : 1467.725 ±0%

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

This output is from a Bayesian linear regression analysis using the Bayes factor approach. The analysis compares two models: one with three predictors (HS.Grad, Income, and Illiteracy) and an intercept, and the other with only an intercept (Intercept only). The Bayes factor is used to compare the evidence for these two models.
The output shows the Bayes factor in favor of the three-predictor model relative to the intercept-only model. In this case, the Bayes factor is 1467.725, which means that the evidence for the three-predictor model is much stronger than the evidence for the intercept-only model.
The output also shows the precision of the Bayes factor estimate, indicated by the ±0% after the value. This indicates that the estimate has essentially zero uncertainty.