| | |
|---|---|
| **Started on** | Sunday, January 27, 2019, 6:06 PM |
| **State** | Finished |
| **Completed on** | Thursday, January 31, 2019, 9:22 PM |
| **Time taken** | 4 days 3 hours |
| **Grade** | **30.00** out of 30.00 (**100**%) |

Question 1

Correct

2.00 points out of 2.00

The output of the linear regression model is depicted below. Write down the equation for the Linear Regression Model using only statistically significant predictors.

To allow for automatic grading,

- do NOT use white-spaces
- use * for multiplication
- use the coefficient before its predictor
- round the coefficients to the fourth digit after the period
- use full names for the response and predictors
- list predictors in the decreasing order of importance

```
Call:
lm(formula = Murder ~ Population + Illiteracy + Income + Frost,
    data = dtrain)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6803 -1.9564  0.8795  2.0245  3.6822

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.5562389  5.3083560  -0.105  0.91759
Population   0.0003505  0.0001413   2.481  0.02209 *
Illiteracy   5.1626172  1.4240915   3.625  0.00169 **
Income      -0.0001088  0.0008357  -0.130  0.89774
Frost        0.0129793  0.0168122   0.772  0.44913
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.619 on 20 degrees of freedom
Multiple R-squared:  0.6088, Adjusted R-squared:  0.5305
F-statistic:  7.78 on 4 and 20 DF,  p-value: 0.0005955
```

Answer: Murder=5.1626*Illiteracy+0.0004*Population ✔

The correct answer is: Murder=5.1626 Illiteracy+0.0004 Population

Question 2

Correct

3.00 points out of 3.00

The output of the linear regression model is depicted below.

Holding all the other predictors constant, how much increase in one unit of **Illitreacy** contributes to the increase or decrease in the Murder rate?

To allow for automatic grading, do NOT use white-space characters and report using the following format rounded to the whole percentages (no decimals):

- increase:10
- decrease:7

```
Call:
lm(formula = Murder ~ Population + Illiteracy + Income + Frost,
    data = dtrain)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6803 -1.9564  0.8795  2.0245  3.6822

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.5562389  5.3083560  -0.105  0.91759
Population   0.0003505  0.0001413   2.481  0.02209 *
Illiteracy   5.1626172  1.4240915   3.625  0.00169 **
Income      -0.0001088  0.0008357  -0.130  0.89774
Frost        0.0129793  0.0168122   0.772  0.44913
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.619 on 20 degrees of freedom
Multiple R-squared:  0.6088, Adjusted R-squared:  0.5305
F-statistic:  7.78 on 4 and 20 DF,  p-value: 0.0005955
```

Answer: increase:5 ✔

The correct answer is: increase:5

Question 3

Complete

20.00 points out of 20.00

Suppose that you are aiming for building a linear regression model using the following code snippet below. Justify whether all the assumptions of the linear regression model are satisfied for building such a model. Include the corresponding R code with your answer.

The attached figure lists the assumptions you are expected to check.

- **Independence:** The observations (rows) must be independent of each other
- **Linearity**: Relationship between the Response (Y) and the predictors (X's) must be linear in terms of the model parameters (β's):
  - Use **crPlots()** in the car pkg for systematic departures from linear model
  - Think how to transform X's to achieve this
- **Normality**: The Response (Y) in a linear regression model must be from the normal (Gaussian distribution):
  - Test for normality using **qqPlot()** in the **car** pkg
  - Check **gvlma** package for global test for linear model assumptions
- **Error/Noise**: The Residuals (predicted – actual) must come from the normal distribution $\mathcal{N}(0,1)$ and should not be autocorrelated (**durbinWatsonTest()**)
- **Homoscedasticity**: The errors (residuals) must be structured:
  - The **car** package provides **ncvTest()** to the hypothesis of constant error variance against the alternative that the error variance changes with the level of fitted values: significant result suggest heteroscedasticity
- **Multicollinearity**: Test for absence of multicollinearity (**vif()** in the **car** pkg)
- **Sensitivity to outliers**: Sensitivity to outliers may affect model performance:
  - Use **outlierTest()** in the car pkg to identify ***high-leverage observations*** and ***influential observations***
- **Model complexity**: Feature selection (forward, backward, stepwise regression) and significant coefficient may guide towards models with reduced complexity

```
states <- as.data.frame(
  state.x77[,c("Murder","Population",
               "Illiteracy", "Income", "Frost")])
dim(states)
t(states[1,])
dtrain <- states[1:25,]
dtest <- states[26:50,]
murderModel <- lm (Murder ~ Population + Illiteracy
             + Income + Frost, data=dtrain)
summary (murderModel)
```

########################  Shashank Shekhar: Code Begins

########################  Discussed with: Rahul Aettapu

par("mar")

```
par(mar=c(1,1,1,1))
### To plot 2*2 = 4 graphs
par(mfrow=c(2,2))
#-----------------------------------------------------------------
# To check for Independence, plot residuals against any time variable
plot(1:length(murderModel$residuals), murderModel$residuals)
### Conclusion: Since there is no definitive pattern in the residuals wrt. time,
### we can conclude that the observations are independent of each other
#-----------------------------------------------------------------
# To check for Linearity, use crPlots which gives the plots of all independent
# variables against dependent variable thus telling about the Linearity of the
# model
crPlots(murderModel)
### Conclusion: From the model summary, it is clear that Illiteracy is the
### attribute that has the highest significance. Looking at the plot of Illiteracy
### and the Residuals, it shows a positive increasing trend in the data values,
### hence the assumption of Linearity is satisfied.
### For the Population feature carrying 2nd highest significance, transformation
### of the predictor variable is required to fit the data properly, otherwise the
### data [population attribute] is underfit.
#-----------------------------------------------------------------
# To check for Normality of the Response variable, using the qqPlot
qqPlot(murderModel)
### Conclusion: The values closely follow the dark blue line (45 degree) and are
### well within the confidence interval bounds indicating that the values of the
### Response variable holds Normality. Although a slight helix shaped pattern is
### observed, it is not significant enough to reject normality.
#-----------------------------------------------------------------
# To check for Error/Noise normality, plot residuals against fitted values
dtrain$prediction <- predict(murderModel,newdata=dtrain)
plot(dtrain$prediction, murderModel$residuals)
### Conclusion: There is no trend in the plotted data indicating that there is no
### correlation among the errors.
# ~~~~~~ # Alternatively, we can use the Durbin Watson test
durbinWatsonTest(murderModel)
```

```
### Conclusion: p value > 0.05 and Statistic value > 2 indicates that we cannot
### reject the null hypothesis and hence there is no correllation among residuals.
#-----------------------------------------------------------------------
# To check for Homoscedasticity
ncvTest(murderModel)
### Conclusion: p-value > 0.05 for Heteroscedasticity from the nvcTest indicates that
### the null hypothesis cannot the rejected, hence the model has Homoscedasticity
-----------------------------------------------------------------------
# To test for Multicollinearity
vif(murderModel)
### Since the values of the variance inflation factor < 4, there is no
### multicollinearity among the independent variables
#-----------------------------------------------------------------------
# To test for the sensitivity of the outliers
outlierTest(murderModel)
### Conclusion: p-value > 0.05 indicates that we can not reject the null hypothesis.
### Hence, the model is not sensitive to outliers.
#-----------------------------------------------------------------------
# To test for Model Complexity,
AIC(murderModel)
### Conclusion: The current value of AIC obtained is high which suggests that the
### model is complex. Also from the summary it is clear that Illiteracy and
### Population are significant features. Therefore, constructing the model with
### Illiteracy and Population.
murderModel2 <- lm(Murder ~ Population + Illiteracy, data=dtrain)
AIC(murderModel2)
### This model has only significant features and is also less complex than the
### original model. AIC value for this model is lower than the original model
### suggesting that it is a better model.
#-----------------------------------------------------------------------
```

Quiz_Code [Shashank Shekhar].R

```
states <- as.data.frame( state.x77[,c("Murder","Population", "Illiteracy", "Income", "Frost")])
dim(states)
```

```
t(states[1,])
dtrain <- states[1:25,]
dtest <- states[26:50,]
murderModel <- lm (Murder ~ Population + Illiteracy + Income + Frost, data=dtrain)
summary (murderModel)

#-------------
#Independence:
#-------------
# It is a necessary assumption that these data are independent.

library(car)

#-------------
# Linearity:
#-------------
help(crPlots)
crPlots(murderModel)
plot(dtrain$Murder,residuals(murderModel))

#The component and residual plots show that
# there might be some linearity issue with the population IV.
# It appears to exhibit the desired linearity up to
# population values of approximately 50,000.
# Above 50,000 population, there is limited data
# from which to draw any conclusions.
# All other DVs seem okay.

#-------------
# Normality:
#-------------
# A normal probability plot is a scatterplot of
# the data vs. the expected quantiles.
# If the data indeed come from a normal distribution,
# then the scatterplot should deviate in
```

```
# a random fashion from the reference line.
# Note that the 45 degree line serves
# as a convenient reference line for detecting
# a systematic departure from normality
qqPlot(murderModel)

# The studentized residuals look okay when
# plotted against t-quantiles.

# also use gvlma for a general overall assessment
# of how well the model fits the assumptions
install.packages("gvlma")
library(gvlma)
modelAssess <- gvlma(murderModel)
summary(modelAssess)

#----------------
# Error/Noise:
#----------------
durbinWatsonTest(murderModel)
# The high p value and high Statistic value > 2
# indicates that we cannot reject the null hypothesis
# that there is no correllation among residuals.



library(gvlma)
gvmodel<-gvlma(murderModel)
#finds that all assumptions are acceptable.
gvmodel.del<-deletion.gvlma(gvmodel)
summary(gvmodel.del)
plot(gvmodel.del)
#leave one out tests indicate that observations
# 1, 7, and 18 may have undue influence over the model results.
# Furthermore it appears this is due to the values of
# income and frost for those observations.
```

```
# So if we leave those out in the final model,
# that should avoid any problems.

#-------------
# Homoscedasticity:
#-------------
ncvTest(murderModel)
# The p value 0f 0.935 is sufficiently large enough for us reason
# to fail to reject the null hypothesis of
# Homoscedasticity of model residuals

plot(murderModel)

#-------------
# Multicollinearity:
#-------------
# If sqrt() if vifresults > 2.0,
# then multicollinearity is present in the data
vifstats <- vif(murderModel)
sqrt(vifstats) > 2.0
# No Multicolinearity . . .

#-------------------
# Sensitivity to outliers:
#-------------------
outlierTest(murderModel)
# NA: p-value > 1
# Hence, we can not reject the null hypothesis:
# Model is not sensitive to outliers

#-------------------
# Model Complexity:
#-------------------
library(leaps)
help(regsubsets)
```

```
leap<-regsubsets(Murder~Population + Illiteracy + Income + Frost,
                 data=dtrain ,nvmax=20,nbest=10)
#force.in=1,
summary(leap)
plot(leap,scale="adjr2")
plot(leap,scale="bic",)
leapsum<-summary(leap)
minbic<-min(leapsum$bic)
evidence<-leapsum$bic<=minbic+4
orderbic<-order(leapsum$bic[evidence],decreasing=F)
orderadjr2<-order(leapsum$adjr2[evidence],decreasing=T)
leapsum$outmat[evidence,][orderbic,]

# look at all models without positive
# BIC evidence against them compared
# to best BIC model
leapsum$outmat[evidence,][orderadjr2,]
leapsum$adjr2[evidence][orderadjr2]
leapsum$bic[evidence][orderbic]

#This procedure suggests that the best model
# is Murder~Population+Illiteracy.
#The next best model it suggests is Population+Illiteracy+Frost,

#Evaluating the models by adjusted r^2 yields similar conclusions.
```

Comment:

Question 4

Correct

5.00 points out of 5.00

Both the output of the logistic regression model summary and the R code used to produce this output are depicted below.

Holding all the other predictors constant, the **infidelity** (ynaffair) is increased or decreased by what factor for a unit increase in the **rating** predictor?

To allow for automatic grading, do NOT use white-space characters and report using the following format rounded to the two decimal points:

- increase:1.05
- decrease:0.86

```
data(Affairs, package="AER")
Affairs$ynaffair[Affairs$affairs > 0] <- 1
Affairs$ynaffair[Affairs$affairs == 0] <- 0
Affairs$ynaffair <- factor(Affairs$ynaffair,
                           levels = c(0,1),
                           labels = c("No","Yes"))
```

```
fit.full <- glm (ynaffair ~
                   gender +
                   age +
                   yearsmarried +
                   children +
                   religiousness +
                   education +
                   occupation +
                   rating,
                 data = Affairs,
                 family = binomial(link="logit")
             )
summary(fit.full)
```

```
glm(formula = ynaffair ~ gender + age + yearsmarried + children +
    religiousness + education + occupation + rating, family = binomial(link = "log
it"),
    data = Affairs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5713  -0.7499  -0.5690  -0.2539   2.5191

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.37726    0.88776   1.551 0.120807
gendermale    0.28029    0.23909   1.172 0.241083
age          -0.04426    0.01825  -2.425 0.015301 *
yearsmarried  0.09477    0.03221   2.942 0.003262 **
childrenyes   0.39767    0.29151   1.364 0.172508
religiousness -0.32472    0.08975  -3.618 0.000297 ***
education     0.02105    0.05051   0.417 0.676851
occupation    0.03092    0.07178   0.431 0.666630
rating       -0.46845    0.09091  -5.153 2.56e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 675.38  on 600  degrees of freedom
Residual deviance: 609.51  on 592  degrees of freedom
AIC: 627.51

Number of Fisher Scoring iterations: 4
```

Answer: decrease:0.63 ✔

The correct answer is: decrease:0.63

◄ (DUE: 01/23/2019): SUBMIT: QUIZ: Generalized Linear Models: Basics

Jump to...

(DUE: 01/30/2019): SUBMIT: QUIZ: Bayesian Inference ►