Jacob Dineen

UNEXECUTED:

**#ADS IST 687**

**#Jacob Dineen**

**#Homework 8**

**#Due 9/10/2017**

**#################All Calls to Clear Environment and Fetch Packages**

**#CLEAR ENVIRONMENT AND INSTALL INITIAL PACKAGES**

**rm(list = ls(all = TRUE))#Clear Enviroment**

**################ START OF HOMEWORK ASSIGNMENT 8. IST687 – Making Predictions**

**#Read in Dataset**

**URLtoRead <- "http://college.cengage.com/mathematics/brase/understandable\_statistics/7e/students/datasets/mlr/excel/mlr01.xls"**

**regression <- read.xls(URLtoRead)**

**str(regression)**

**View(regression)**

**#Update Column Names**

**colnames(regression) <- c("Fawn", "Antelope", "Precipitation", "WinterRating")**

**str(regression)**

**#Bivariate Plots**

**##Fawn V Antelope**

**ggplot(regression,aes(x=regression$Antelope,y=regression$Fawn))+ geom\_point(color="red", size=2)+ labs(x = "Antelope", y= "Fawn", title = "Antelope V Fawn")**

**##Fawn V Precipitation**

**ggplot(regression,aes(x=regression$Precipitation,y=regression$Fawn))+ geom\_point(color="red", size=2)+ labs(x = "Precipitation", y= "Fawn", title = "Precipitation V Fawn")**

**##Fawn V Severity of Winter**

**ggplot(regression, aes(x=regression$WinterRating, y=regression$Fawn))+ geom\_point(color="red", size=2)+ labs(x = "WinterRating", y= "Fawn", title = "WinterRating V Fawn")**

**#Fawn is the Dependent Variable, so it goes on the Y axis of each of these plots.**

**#linear models**

**##Fawn V Severity of Winter**

**model1 <- lm(formula= Fawn ~ WinterRating, data=regression)**

**summary(model1)**

**plot(regression$WinterRating, regression$Fawn)**

**abline(model1)**

**#Predict#**

**range(regression$WinterRating)**

**newdata <- data.frame (WinterRating=3)**

**predict(model1, newdata, type="response")**

**#Analysis**

**# Adjusted R-squared: 0.4702, so the variance in Fawns is not heavily reliant on the variance in the severity of the winter.**

**#The predictor 'winterRating' was not statistally significant in this model - pvalue= 0.036263**

**##Fawn v Precipation & Severity of Winter**

**model2 <- lm(formula=Fawn ~ WinterRating+Precipitation, data=regression)**

**summary(model2)**

**range(regression$WinterRating)**

**range (regression$Precipitation)**

**#Predict#**

**newdata1 <- data.frame(Precipitation= 12.5, WinterRating=4)**

**predict(model2, newdata1, type="response")**

**#Analysis**

**# Adjusted R-squared: 0.86, so this model was a pretty solid fit for our data.**

**#We can see that precipitation, with a p value of .008 was a very statistically significant predictor variable for our model.**

**##Fawn V All Variables**

**model3 <- lm(formula= Fawn~ ., data=regression)**

**summary(model3)**

**range(regression$WinterRating)**

**range (regression$Precipitation)**

**range (regression$Antelope)**

**#Predict#**

**newdata2 <- data.frame(Precipitation= 14, WinterRating=2, Antelope=8.5)**

**predict(model3, newdata2, type= "response")**

**#Analysis**

**#Adjusted R-squared: 0.955, so this model, with 3 variables used, was the best fit for our data.**

**#All of the variables were statistically significant in the model, with everything registering a pvalue < .05.**

**#Look at step function in week8synchcode file**

**#Parasimonious Model**

**step(model3,data=regression,direction='backward')**

**#if we decide to use the AIC function to best determine the best model, with the least amount of variables, we see that the result is a model that**

**#includes all relevant variables. In this case, we see a model that includes all three of the initial variables.**

**#With all 3 variables wee see an adjusted r2 of .955 and see that all variables are statistically relevant to the outcome (y)**

**summary(lm(formula = Fawn ~ Antelope + Precipitation + WinterRating,**

**data = regression))**

**#Plotting**

**model1 <- lm(formula= Fawn ~ WinterRating, data=regression)**

**summary(model1)**

**g <- ggplot(regression, aes(x=WinterRating, y= Fawn)) + geom\_point()**

**g + geom\_smooth(method = "lm")**

**model4 <- lm(formula= Fawn ~ Antelope, data=regression)**

**summary(model4)**

**g1 <- ggplot(regression, aes(x=Antelope, y= Fawn)) + geom\_point()**

**g1 + geom\_smooth(method = "lm")**

**model5 <- lm(formula= Fawn ~ Precipitation, data=regression)**

**summary(model5)**

**g2 <- ggplot(regression, aes(x=Precipitation, y= Fawn)) + geom\_point()**

**g2 + geom\_smooth(method = "lm")**

**model6 <- lm(formula= Fawn ~ Precipitation+Antelope, data=regression)**

**summary(model6)**

|  |
| --- |
| > ################ START OF HOMEWORK ASSIGNMENT 8. IST687 – Making Predictions  >  > #Read in Dataset  > URLtoRead <- "http://college.cengage.com/mathematics/brase/understandable\_statistics/7e/students/datasets/mlr/excel/mlr01.xls"  > regression <- read.xls(URLtoRead)  trying URL 'http://college.cengage.com/mathematics/brase/understandable\_statistics/7e/students/datasets/mlr/excel/mlr01.xls'  Content type 'application/vnd.ms-excel' length 5632 bytes  downloaded 5632 bytes  >  > str(regression)  'data.frame': 8 obs. of 4 variables:  $ X1: num 2.9 2.4 2 2.3 3.2 ...  $ X2: num 9.2 8.7 7.2 8.5 9.6 ...  $ X3: num 13.2 11.5 10.8 12.3 12.6 ...  $ X4: int 2 3 4 2 3 5 1 3  > View(regression)  >  > #Update Column Names  > colnames(regression) <- c("Fawn", "Antelope", "Precipitation", "WinterRating")  > str(regression)  'data.frame': 8 obs. of 4 variables:  $ Fawn : num 2.9 2.4 2 2.3 3.2 ...  $ Antelope : num 9.2 8.7 7.2 8.5 9.6 ...  $ Precipitation: num 13.2 11.5 10.8 12.3 12.6 ...  $ WinterRating : int 2 3 4 2 3 5 1 3  >  > #Bivariate Plots  >  > ##Fawn V Antelope  > ggplot(regression,aes(x=regression$Antelope,y=regression$Fawn))+ geom\_point(color="red", size=2)+ labs(x = "Antelope", y= "Fawn", title = "Antelope V Fawn")  >  > ##Fawn V Precipitation  > ggplot(regression,aes(x=regression$Precipitation,y=regression$Fawn))+ geom\_point(color="red", size=2)+ labs(x = "Precipitation", y= "Fawn", title = "Precipitation V Fawn")  >  > ##Fawn V Severity of Winter  > ggplot(regression, aes(x=regression$WinterRating, y=regression$Fawn))+ geom\_point(color="red", size=2)+ labs(x = "WinterRating", y= "Fawn", title = "WinterRating V Fawn")  >  > #Fawn is the Dependent Variable, so it goes on the Y axis of each of these plots.  >  > #linear models  >  > ##Fawn V Severity of Winter  > model1 <- lm(formula= Fawn ~ WinterRating, data=regression)  > summary(model1)  Call:  lm(formula = Fawn ~ WinterRating, data = regression)  Residuals:  Min 1Q Median 3Q Max  -0.52069 -0.20431 -0.00172 0.13017 0.71724  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 3.4966 0.3904 8.957 0.000108 \*\*\*  WinterRating -0.3379 0.1258 -2.686 0.036263 \*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.415 on 6 degrees of freedom  Multiple R-squared: 0.5459, Adjusted R-squared: 0.4702  F-statistic: 7.213 on 1 and 6 DF, p-value: 0.03626  > plot(regression$WinterRating, regression$Fawn)  > abline(model1)  >  > #Predict#  > range(regression$WinterRating)  [1] 1 5  > newdata <- data.frame (WinterRating=3)  > predict(model1, newdata, type="response")  1  2.482759  >  > #Analysis  > # Adjusted R-squared: 0.4702, so the variance in Fawns is not heavily reliant on the variance in the severity of the winter.  > #The predictor 'winterRating' was not statistally significant in this model - pvalue= 0.036263  >  >  >  > ##Fawn v Precipation & Severity of Winter  > model2 <- lm(formula=Fawn ~ WinterRating+Precipitation, data=regression)  > summary(model2)  Call:  lm(formula = Fawn ~ WinterRating + Precipitation, data = regression)  Residuals:  1 2 3 4 5 6 7 8  -0.165458 0.188313 0.006417 -0.193358 0.289080 -0.193312 -0.010695 0.079013  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -5.7791 2.2139 -2.610 0.04765 \*  WinterRating 0.2269 0.1490 1.522 0.18842  Precipitation 0.6357 0.1511 4.207 0.00843 \*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.2133 on 5 degrees of freedom  Multiple R-squared: 0.9, Adjusted R-squared: 0.86  F-statistic: 22.49 on 2 and 5 DF, p-value: 0.003164  > range(regression$WinterRating)  [1] 1 5  > range (regression$Precipitation)  [1] 10.6 14.1  >  > #Predict#  > newdata1 <- data.frame(Precipitation= 12.5, WinterRating=4)  > predict(model2, newdata1, type="response")  1  3.074216  >  > #Analysis  > # Adjusted R-squared: 0.86, so this model was a pretty solid fit for our data.  > #We can see that precipitation, with a p value of .008 was a very statistically significant predictor variable for our model.  >  >  > ##Fawn V All Variables  > model3 <- lm(formula= Fawn~ ., data=regression)  > summary(model3)  Call:  lm(formula = Fawn ~ ., data = regression)  Residuals:  1 2 3 4 5 6 7 8  -0.11533 -0.02661 0.09882 -0.11723 0.02734 -0.04854 0.11715 0.06441  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -5.92201 1.25562 -4.716 0.0092 \*\*  Antelope 0.33822 0.09947 3.400 0.0273 \*  Precipitation 0.40150 0.10990 3.653 0.0217 \*  WinterRating 0.26295 0.08514 3.089 0.0366 \*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.1209 on 4 degrees of freedom  Multiple R-squared: 0.9743, Adjusted R-squared: 0.955  F-statistic: 50.52 on 3 and 4 DF, p-value: 0.001229  > range(regression$WinterRating)  [1] 1 5  > range (regression$Precipitation)  [1] 10.6 14.1  > range (regression$Antelope)  [1] 6.8 9.7  >  > #Predict#  > newdata2 <- data.frame(Precipitation= 14, WinterRating=2, Antelope=8.5)  > predict(model3, newdata2, type= "response")  1  3.099784  >  > #Analysis  > #Adjusted R-squared: 0.955, so this model, with 3 variables used, was the best fit for our data.  > #All of the variables were statistically significant in the model, with everything registering a pvalue < .05.  >  >  > #Look at step function in week8synchcode file  >  >  > #Parasimonious Model  >  > step(model3,data=regression,direction='backward')  Start: AIC=-31.35  Fawn ~ Antelope + Precipitation + WinterRating  Df Sum of Sq RSS AIC  <none> 0.058494 -31.346  - WinterRating 1 0.13950 0.197989 -23.592  - Antelope 1 0.16907 0.227561 -22.478  - Precipitation 1 0.19518 0.253673 -21.609  Call:  lm(formula = Fawn ~ Antelope + Precipitation + WinterRating,  data = regression)  Coefficients:  (Intercept) Antelope Precipitation WinterRating  -5.9220 0.3382 0.4015 0.2629  >  >  > #if we decide to use the AIC function to best determine the best model, with the least amount of variables, we see that the result is a model that  > #includes all relevant variables. In this case, we see a model that includes all three of the initial variables.  > #With all 3 variables wee see an adjusted r2 of .955 and see that all variables are statistically relevant to the outcome (y)  >  > summary(lm(formula = Fawn ~ Antelope + Precipitation + WinterRating,  + data = regression))  Call:  lm(formula = Fawn ~ Antelope + Precipitation + WinterRating,  data = regression)  Residuals:  1 2 3 4 5 6 7 8  -0.11533 -0.02661 0.09882 -0.11723 0.02734 -0.04854 0.11715 0.06441  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -5.92201 1.25562 -4.716 0.0092 \*\*  Antelope 0.33822 0.09947 3.400 0.0273 \*  Precipitation 0.40150 0.10990 3.653 0.0217 \*  WinterRating 0.26295 0.08514 3.089 0.0366 \*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.1209 on 4 degrees of freedom  Multiple R-squared: 0.9743, Adjusted R-squared: 0.955  F-statistic: 50.52 on 3 and 4 DF, p-value: 0.001229  >  >  >  > #Plotting  >  > model1 <- lm(formula= Fawn ~ WinterRating, data=regression)  > summary(model1)  Call:  lm(formula = Fawn ~ WinterRating, data = regression)  Residuals:  Min 1Q Median 3Q Max  -0.52069 -0.20431 -0.00172 0.13017 0.71724  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 3.4966 0.3904 8.957 0.000108 \*\*\*  WinterRating -0.3379 0.1258 -2.686 0.036263 \*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.415 on 6 degrees of freedom  Multiple R-squared: 0.5459, Adjusted R-squared: 0.4702  F-statistic: 7.213 on 1 and 6 DF, p-value: 0.03626  > g <- ggplot(regression, aes(x=WinterRating, y= Fawn)) + geom\_point()  > g + geom\_smooth(method = "lm")  >  >  > model4 <- lm(formula= Fawn ~ Antelope, data=regression)  > summary(model4)  Call:  lm(formula = Fawn ~ Antelope, data = regression)  Residuals:  Min 1Q Median 3Q Max  -0.24988 -0.17586 0.04938 0.12611 0.25309  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -1.67914 0.63422 -2.648 0.038152 \*  Antelope 0.49753 0.07453 6.676 0.000547 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.2121 on 6 degrees of freedom  Multiple R-squared: 0.8813, Adjusted R-squared: 0.8616  F-statistic: 44.56 on 1 and 6 DF, p-value: 0.0005471  > g1 <- ggplot(regression, aes(x=Antelope, y= Fawn)) + geom\_point()  > g1 + geom\_smooth(method = "lm")  >  > model5 <- lm(formula= Fawn ~ Precipitation, data=regression)  > summary(model5)  Call:  lm(formula = Fawn ~ Precipitation, data = regression)  Residuals:  Min 1Q Median 3Q Max  -0.33747 -0.08040 -0.00889 0.03023 0.43399  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -2.63251 0.87591 -3.005 0.02384 \*  Precipitation 0.42845 0.07244 5.915 0.00104 \*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.2356 on 6 degrees of freedom  Multiple R-squared: 0.8536, Adjusted R-squared: 0.8292  F-statistic: 34.99 on 1 and 6 DF, p-value: 0.001039  > g2 <- ggplot(regression, aes(x=Precipitation, y= Fawn)) + geom\_point()  > g2 + geom\_smooth(method = "lm")  >  > model6 <- lm(formula= Fawn ~ Precipitation+Antelope, data=regression)  > summary(model6)  Call:  lm(formula = Fawn ~ Precipitation + Antelope, data = regression)  Residuals:  1 2 3 4 5 6 7 8  -0.07265 -0.09701 0.08698 -0.29029 0.22233 0.14526 0.10497 -0.09960  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -2.3155 0.7595 -3.049 0.0285 \*  Precipitation 0.1916 0.1421 1.348 0.2355  Antelope 0.2999 0.1624 1.847 0.1241  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.199 on 5 degrees of freedom  Multiple R-squared: 0.913, Adjusted R-squared: 0.8782  F-statistic: 26.23 on 2 and 5 DF, p-value: 0.002234 |
|  |
| |  | | --- | | > | |