PartA\_Tatum.R

Jennifer

Sun Mar 26 12:45:43 2017

getwd()

## [1] "C:/Users/Jennifer/Downloads"

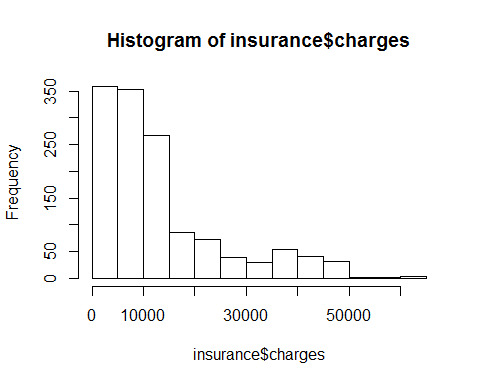
setwd("C:/Users/Jennifer/Documents/ADM/HW 1")  
insurance<-read.csv(file="insurance.csv")  
  
str(insurance)

## 'data.frame': 1338 obs. of 7 variables:  
## $ age : int 19 18 28 33 32 31 46 37 37 60 ...  
## $ sex : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...  
## $ bmi : num 27.9 33.8 33 22.7 28.9 ...  
## $ children: int 0 1 3 0 0 0 1 3 2 0 ...  
## $ smoker : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...  
## $ region : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...  
## $ charges : num 16885 1726 4449 21984 3867 ...

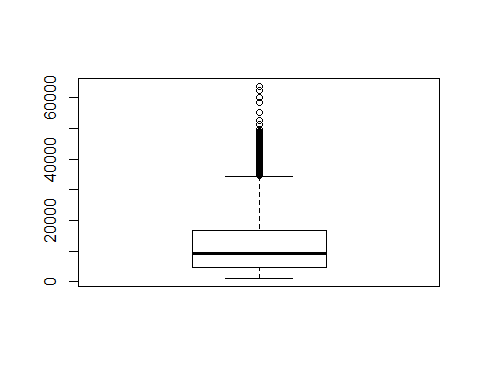
summary(insurance)

## age sex bmi children smoker   
## Min. :18.00 female:662 Min. :15.96 Min. :0.000 no :1064   
## 1st Qu.:27.00 male :676 1st Qu.:26.30 1st Qu.:0.000 yes: 274   
## Median :39.00 Median :30.40 Median :1.000   
## Mean :39.21 Mean :30.66 Mean :1.095   
## 3rd Qu.:51.00 3rd Qu.:34.69 3rd Qu.:2.000   
## Max. :64.00 Max. :53.13 Max. :5.000   
## region charges   
## northeast:324 Min. : 1122   
## northwest:325 1st Qu.: 4740   
## southeast:364 Median : 9382   
## southwest:325 Mean :13270   
## 3rd Qu.:16640   
## Max. :63770

hist(insurance$charges)



boxplot(insurance$charges)



count\_region<-table(insurance$region)  
count\_region

##   
## northeast northwest southeast southwest   
## 324 325 364 325

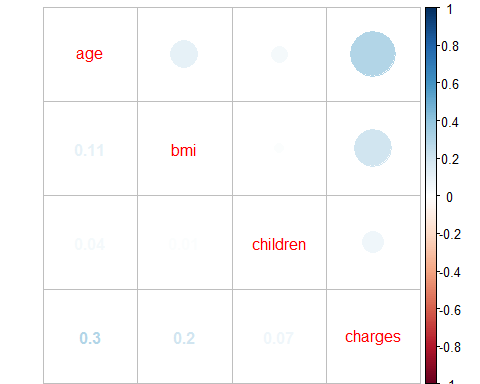
Numeric\_Vars <-insurance[,c(1,3, 4,7)]  
Correlation\_Matrix <- cor(Numeric\_Vars)  
Correlation\_Matrix

## age bmi children charges  
## age 1.0000000 0.1092719 0.04246900 0.29900819  
## bmi 0.1092719 1.0000000 0.01275890 0.19834097  
## children 0.0424690 0.0127589 1.00000000 0.06799823  
## charges 0.2990082 0.1983410 0.06799823 1.00000000

library(corrplot)

## Warning: package 'corrplot' was built under R version 3.3.3

corrplot.mixed(Correlation\_Matrix)



plot(insurance$age,insurance$charges)  
  
MODELAGE<-lm(insurance$charges~insurance$age)  
summary(MODELAGE)

##   
## Call:  
## lm(formula = insurance$charges ~ insurance$age)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8059 -6671 -5939 5440 47829   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3165.9 937.1 3.378 0.000751 \*\*\*  
## insurance$age 257.7 22.5 11.453 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11560 on 1336 degrees of freedom  
## Multiple R-squared: 0.08941, Adjusted R-squared: 0.08872   
## F-statistic: 131.2 on 1 and 1336 DF, p-value: < 2.2e-16

MODELSEX<-lm(insurance$charges~insurance$sex)  
summary(MODELSEX)

##   
## Call:  
## lm(formula = insurance$charges ~ insurance$sex)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12835 -8435 -3980 3476 51201   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12569.6 470.1 26.740 <2e-16 \*\*\*  
## insurance$sexmale 1387.2 661.3 2.098 0.0361 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12090 on 1336 degrees of freedom  
## Multiple R-squared: 0.003282, Adjusted R-squared: 0.002536   
## F-statistic: 4.4 on 1 and 1336 DF, p-value: 0.03613

MODELBMI<-lm(insurance$charges~insurance$bmi)  
summary(MODELBMI)

##   
## Call:  
## lm(formula = insurance$charges ~ insurance$bmi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20956 -8118 -3757 4722 49442   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1192.94 1664.80 0.717 0.474   
## insurance$bmi 393.87 53.25 7.397 2.46e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11870 on 1336 degrees of freedom  
## Multiple R-squared: 0.03934, Adjusted R-squared: 0.03862   
## F-statistic: 54.71 on 1 and 1336 DF, p-value: 2.459e-13

MODELCHILDREN<-lm(insurance$charges~insurance$children)  
summary(MODELCHILDREN)

##   
## Call:  
## lm(formula = insurance$charges ~ insurance$children)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11585 -8759 -4071 3468 51248   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12522.5 446.5 28.049 <2e-16 \*\*\*  
## insurance$children 683.1 274.2 2.491 0.0129 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12090 on 1336 degrees of freedom  
## Multiple R-squared: 0.004624, Adjusted R-squared: 0.003879   
## F-statistic: 6.206 on 1 and 1336 DF, p-value: 0.01285

MODELSMOKER<-lm(insurance$charges~insurance$smoker)  
summary(MODELSMOKER)

##   
## Call:  
## lm(formula = insurance$charges ~ insurance$smoker)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19221 -5042 -919 3705 31720   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8434.3 229.0 36.83 <2e-16 \*\*\*  
## insurance$smokeryes 23616.0 506.1 46.66 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7470 on 1336 degrees of freedom  
## Multiple R-squared: 0.6198, Adjusted R-squared: 0.6195   
## F-statistic: 2178 on 1 and 1336 DF, p-value: < 2.2e-16

insurance$region<-relevel(insurance$region, ref = "northeast")  
MODELREGION<-lm(insurance$charges~insurance$region)  
summary(MODELREGION)

##   
## Call:  
## lm(formula = insurance$charges ~ insurance$region)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13614 -8463 -3793 3385 49035   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13406.4 671.3 19.971 <2e-16 \*\*\*  
## insurance$regionnorthwest -988.8 948.6 -1.042 0.297   
## insurance$regionsoutheast 1329.0 922.9 1.440 0.150   
## insurance$regionsouthwest -1059.4 948.6 -1.117 0.264   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12080 on 1334 degrees of freedom  
## Multiple R-squared: 0.006634, Adjusted R-squared: 0.0044   
## F-statistic: 2.97 on 3 and 1334 DF, p-value: 0.03089

insurance$region<-relevel(insurance$region, ref = "northwest")  
MODELREGION2<-lm(insurance$charges~insurance$region)  
summary(MODELREGION2)

##   
## Call:  
## lm(formula = insurance$charges ~ insurance$region)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13614 -8463 -3793 3385 49035   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12417.58 670.26 18.526 <2e-16 \*\*\*  
## insurance$regionnortheast 988.81 948.63 1.042 0.2974   
## insurance$regionsoutheast 2317.84 922.16 2.513 0.0121 \*   
## insurance$regionsouthwest -70.64 947.90 -0.075 0.9406   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12080 on 1334 degrees of freedom  
## Multiple R-squared: 0.006634, Adjusted R-squared: 0.0044   
## F-statistic: 2.97 on 3 and 1334 DF, p-value: 0.03089

insurance$sex<-relevel(insurance$sex, ref = "female")  
MODEL1<-lm(insurance$charges~insurance$age+insurance$sex+insurance$bmi+insurance$children+insurance$smoker+insurance$region)  
summary(MODEL1)

##   
## Call:  
## lm(formula = insurance$charges ~ insurance$age + insurance$sex +   
## insurance$bmi + insurance$children + insurance$smoker + insurance$region)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11304.9 -2848.1 -982.1 1393.9 29992.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -12291.5 988.2 -12.438 < 2e-16 \*\*\*  
## insurance$age 256.9 11.9 21.587 < 2e-16 \*\*\*  
## insurance$sexmale -131.3 332.9 -0.394 0.693348   
## insurance$bmi 339.2 28.6 11.860 < 2e-16 \*\*\*  
## insurance$children 475.5 137.8 3.451 0.000577 \*\*\*  
## insurance$smokeryes 23848.5 413.1 57.723 < 2e-16 \*\*\*  
## insurance$regionnortheast 353.0 476.3 0.741 0.458769   
## insurance$regionsoutheast -682.1 479.0 -1.424 0.154669   
## insurance$regionsouthwest -607.1 477.2 -1.272 0.203533   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6062 on 1329 degrees of freedom  
## Multiple R-squared: 0.7509, Adjusted R-squared: 0.7494   
## F-statistic: 500.8 on 8 and 1329 DF, p-value: < 2.2e-16

MODEL2<-lm(insurance$charges~insurance$age+insurance$bmi+insurance$children+insurance$smoker)  
summary(MODEL2)

##   
## Call:  
## lm(formula = insurance$charges ~ insurance$age + insurance$bmi +   
## insurance$children + insurance$smoker)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11897.9 -2920.8 -986.6 1392.2 29509.6   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -12102.77 941.98 -12.848 < 2e-16 \*\*\*  
## insurance$age 257.85 11.90 21.675 < 2e-16 \*\*\*  
## insurance$bmi 321.85 27.38 11.756 < 2e-16 \*\*\*  
## insurance$children 473.50 137.79 3.436 0.000608 \*\*\*  
## insurance$smokeryes 23811.40 411.22 57.904 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6068 on 1333 degrees of freedom  
## Multiple R-squared: 0.7497, Adjusted R-squared: 0.7489   
## F-statistic: 998.1 on 4 and 1333 DF, p-value: < 2.2e-16

OBESITY<-insurance$bmi>=30  
MODEL3<-lm(insurance$charges~insurance$age+insurance$bmi+insurance$children+insurance$smoker+OBESITY)  
summary(MODEL3)

##   
## Call:  
## lm(formula = insurance$charges ~ insurance$age + insurance$bmi +   
## insurance$children + insurance$smoker + OBESITY)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12530 -3503 -232 1494 28075   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7816.28 1233.73 -6.335 3.23e-10 \*\*\*  
## insurance$age 258.03 11.78 21.909 < 2e-16 \*\*\*  
## insurance$bmi 131.71 44.93 2.932 0.003428 \*\*   
## insurance$children 473.87 136.41 3.474 0.000529 \*\*\*  
## insurance$smokeryes 23819.41 407.10 58.511 < 2e-16 \*\*\*  
## OBESITYTRUE 2904.47 547.33 5.307 1.31e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6007 on 1332 degrees of freedom  
## Multiple R-squared: 0.7549, Adjusted R-squared: 0.754   
## F-statistic: 820.4 on 5 and 1332 DF, p-value: < 2.2e-16

MODEL4<-lm(insurance$charges~insurance$age+insurance$bmi+insurance$children+insurance$smoker+OBESITY+OBESITY:insurance$smoker)  
summary(MODEL4)

##   
## Call:  
## lm(formula = insurance$charges ~ insurance$age + insurance$bmi +   
## insurance$children + insurance$smoker + OBESITY + OBESITY:insurance$smoker)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19074.0 -1865.8 -1254.9 -440.4 24385.3   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5097.881 929.887 -5.482 5.02e-08 \*\*\*  
## insurance$age 264.226 8.842 29.883 < 2e-16 \*\*\*  
## insurance$bmi 97.698 33.738 2.896 0.00384 \*\*   
## insurance$children 512.424 102.395 5.004 6.35e-07 \*\*\*  
## insurance$smokeryes 13412.395 445.186 30.128 < 2e-16 \*\*\*  
## OBESITYTRUE -809.791 426.763 -1.898 0.05798 .   
## insurance$smokeryes:OBESITYTRUE 19684.869 612.394 32.144 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4509 on 1331 degrees of freedom  
## Multiple R-squared: 0.862, Adjusted R-squared: 0.8614   
## F-statistic: 1386 on 6 and 1331 DF, p-value: < 2.2e-16

anova(MODEL1, MODEL2)

## Analysis of Variance Table  
##   
## Model 1: insurance$charges ~ insurance$age + insurance$sex + insurance$bmi +   
## insurance$children + insurance$smoker + insurance$region  
## Model 2: insurance$charges ~ insurance$age + insurance$bmi + insurance$children +   
## insurance$smoker  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 1329 4.8840e+10   
## 2 1333 4.9078e+10 -4 -238917273 1.6253 0.1654

anova(MODEL2, MODEL3)

## Analysis of Variance Table  
##   
## Model 1: insurance$charges ~ insurance$age + insurance$bmi + insurance$children +   
## insurance$smoker  
## Model 2: insurance$charges ~ insurance$age + insurance$bmi + insurance$children +   
## insurance$smoker + OBESITY  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 1333 4.9078e+10   
## 2 1332 4.8062e+10 1 1016105153 28.16 1.307e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(MODEL3, MODEL4)

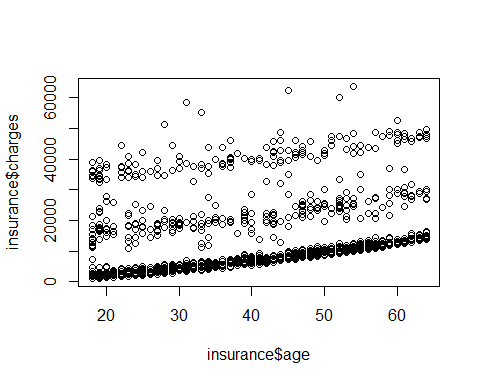
## Analysis of Variance Table  
##   
## Model 1: insurance$charges ~ insurance$age + insurance$bmi + insurance$children +   
## insurance$smoker + OBESITY  
## Model 2: insurance$charges ~ insurance$age + insurance$bmi + insurance$children +   
## insurance$smoker + OBESITY + OBESITY:insurance$smoker  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 1332 4.8062e+10   
## 2 1331 2.7058e+10 1 2.1005e+10 1033.2 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(MODEL2, MODEL4)

## Analysis of Variance Table  
##   
## Model 1: insurance$charges ~ insurance$age + insurance$bmi + insurance$children +   
## insurance$smoker  
## Model 2: insurance$charges ~ insurance$age + insurance$bmi + insurance$children +   
## insurance$smoker + OBESITY + OBESITY:insurance$smoker  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 1333 4.9078e+10   
## 2 1331 2.7058e+10 2 2.2021e+10 541.61 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

library(car)

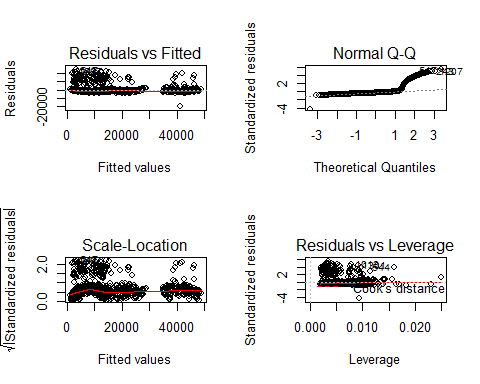
## Warning: package 'car' was built under R version 3.3.3



vif(MODEL4)

## insurance$age insurance$bmi insurance$children   
## 1.014989 2.783933 1.002087   
## insurance$smoker OBESITY insurance$smoker:OBESITY   
## 2.124258 2.987120 2.385075

par(mfrow=c(2,2))  
plot(MODEL4)



durbinWatsonTest(MODEL4)

## lag Autocorrelation D-W Statistic p-value  
## 1 -0.02961429 2.059074 0.242  
## Alternative hypothesis: rho != 0

ABC<-read.csv(file="CompanyABC.csv")  
str(ABC)

## 'data.frame': 50 obs. of 6 variables:  
## $ region : Factor w/ 4 levels "northeast","northwest",..: 1 4 4 1 1 3 4 2 4 2 ...  
## $ sex : Factor w/ 2 levels "female","male": 2 2 1 2 1 1 1 2 1 2 ...  
## $ bmi : num 19.9 19.3 31.6 25.5 30.1 ...  
## $ children: int 0 0 0 0 0 3 1 1 1 2 ...  
## $ smoker : Factor w/ 2 levels "no","yes": 2 2 1 1 1 2 1 2 1 1 ...  
## $ age : int 56 38 41 30 18 61 34 20 19 26 ...

summary(ABC)

## region sex bmi children smoker   
## northeast:13 female:25 Min. :17.77 Min. :0.00 no :38   
## northwest:13 male :25 1st Qu.:27.25 1st Qu.:0.00 yes:12   
## southeast:11 Median :29.34 Median :0.50   
## southwest:13 Mean :29.88 Mean :0.98   
## 3rd Qu.:33.34 3rd Qu.:2.00   
## Max. :49.06 Max. :3.00   
## age   
## Min. :18.00   
## 1st Qu.:23.00   
## Median :34.00   
## Mean :37.14   
## 3rd Qu.:51.75   
## Max. :63.00

ABC$estimated\_charges<-NA