STAT489HW2

Nicholas Thompson

9/14/2021

#Question 1  
  
hw2.data=read.csv("insurance.csv", header = TRUE)  
  
str(hw2.data)

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

#7 variables: age(int), sex(chr), body mass index (num), # of children (int), smoking status (chr), residential region (chr), and medical charges, or costs (num)  
  
model1=lm(charges~age, data = hw2.data)  
model1

##   
## Call:  
## lm(formula = charges ~ age, data = hw2.data)  
##   
## Coefficients:  
## (Intercept) age   
## 3165.9 257.7

#below is the prediction equation

#Interpret Slope: For every 1 year increase in age, we expect estimated individual medical costs to increase by $257.7  
summary(model1)

##   
## Call:  
## lm(formula = charges ~ age, data = hw2.data)  
##   
## 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 \*\*\*  
## 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

#p-value is less than 2 to the -16th power from both the t test and F test from the above summary command. This value is substantially smaller than any useful value of alpha. Therefore, we declare that age is significantly related to medical costs.  
  
#R-squared value is 0.08941, which means that about 8.9% of the variation in medical costs can be explained by a person's age. This low of an R-squared value demonstrates that the simple linear regression model is not enough to predict individual medical costs. Multiple regression analysis is necessary.

#Question 2  
  
model2=lm(charges~age+bmi+children+sex+smoker+region,data = hw2.data)  
  
summary(model2)

##   
## Call:  
## lm(formula = charges ~ age + bmi + children + sex + smoker +   
## region, data = hw2.data)  
##   
## 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) -11938.5 987.8 -12.086 < 2e-16 \*\*\*  
## age 256.9 11.9 21.587 < 2e-16 \*\*\*  
## bmi 339.2 28.6 11.860 < 2e-16 \*\*\*  
## children 475.5 137.8 3.451 0.000577 \*\*\*  
## sexmale -131.3 332.9 -0.394 0.693348   
## smokeryes 23848.5 413.1 57.723 < 2e-16 \*\*\*  
## regionnorthwest -353.0 476.3 -0.741 0.458769   
## regionsoutheast -1035.0 478.7 -2.162 0.030782 \*   
## regionsouthwest -960.0 477.9 -2.009 0.044765 \*   
## ---  
## 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

anova(model2)

## Analysis of Variance Table  
##   
## Response: charges  
## Df Sum Sq Mean Sq F value Pr(>F)   
## age 1 1.7530e+10 1.7530e+10 477.0239 < 2.2e-16 \*\*\*  
## bmi 1 5.4464e+09 5.4464e+09 148.2064 < 2.2e-16 \*\*\*  
## children 1 5.7152e+08 5.7152e+08 15.5519 8.446e-05 \*\*\*  
## sex 1 5.8245e+08 5.8245e+08 15.8494 7.230e-05 \*\*\*  
## smoker 1 1.2287e+11 1.2287e+11 3343.5022 < 2.2e-16 \*\*\*  
## region 3 2.3343e+08 7.7810e+07 2.1173 0.09622 .   
## Residuals 1329 4.8840e+10 3.6749e+07   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#In terms of the overall F-test: the F statistic is reasonably high at 500.8 and p-value is substantially low at being less than 2 to the -16th power. Adjusted R-squared value demonstrates that all predictors explain 74% of the variation in medical costs. A significant improvement over the 8.9% R-squared of the simple model.  
  
model2.red=lm(charges~age+bmi+children+sex+smoker, data = hw2.data)  
summary(model2.red)

##   
## Call:  
## lm(formula = charges ~ age + bmi + children + sex + smoker, data = hw2.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11837.2 -2916.7 -994.2 1375.3 29565.5   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -12052.46 951.26 -12.670 < 2e-16 \*\*\*  
## age 257.73 11.90 21.651 < 2e-16 \*\*\*  
## bmi 322.36 27.42 11.757 < 2e-16 \*\*\*  
## children 474.41 137.86 3.441 0.000597 \*\*\*  
## sexmale -128.64 333.36 -0.386 0.699641   
## smokeryes 23823.39 412.52 57.750 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6070 on 1332 degrees of freedom  
## Multiple R-squared: 0.7497, Adjusted R-squared: 0.7488   
## F-statistic: 798 on 5 and 1332 DF, p-value: < 2.2e-16

anova(model2.red,model2)

## Analysis of Variance Table  
##   
## Model 1: charges ~ age + bmi + children + sex + smoker  
## Model 2: charges ~ age + bmi + children + sex + smoker + region  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 1332 4.9073e+10   
## 2 1329 4.8840e+10 3 233431209 2.1173 0.09622 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Based on the F-test, the p-value is 0.09622. Testing at alpha = 0.05, we find that p is greater than alpha. Therefore, we declare that region has no statistical significance on medical costs.   
  
coef(summary(model2))

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -11938.5386 987.81918 -12.0857530 5.579044e-32  
## age 256.8564 11.89885 21.5866552 7.783217e-89  
## bmi 339.1935 28.59947 11.8601306 6.498194e-31  
## children 475.5005 137.80409 3.4505546 5.769682e-04  
## sexmale -131.3144 332.94544 -0.3944020 6.933475e-01  
## smokeryes 23848.5345 413.15335 57.7232020 0.000000e+00  
## regionnorthwest -352.9639 476.27579 -0.7410914 4.587689e-01  
## regionsoutheast -1035.0220 478.69221 -2.1621870 3.078174e-02  
## regionsouthwest -960.0510 477.93302 -2.0087563 4.476493e-02

#Based on the t values from the summary function and using the alpha = 0.05 test, we can remove the sex predictor since its corresponding p-value (0.69) is greater than 0.05.   
  
refined.model2=lm(charges~age+bmi+children+smoker, data = hw2.data)  
summary(refined.model2)

##   
## Call:  
## lm(formula = charges ~ age + bmi + children + smoker, data = hw2.data)  
##   
## 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 \*\*\*  
## age 257.85 11.90 21.675 < 2e-16 \*\*\*  
## bmi 321.85 27.38 11.756 < 2e-16 \*\*\*  
## children 473.50 137.79 3.436 0.000608 \*\*\*  
## 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

#Question 3  
attach(hw2.data)  
age.sq=(hw2.data$age)^2  
bmi.sq=(hw2.data$bmi)^2  
child.sq=(hw2.data$children)^2  
  
  
model3=lm(charges~age+bmi+children+smoker+age.sq+bmi.sq+child.sq, data = hw2.data)  
summary(model3)

##   
## Call:  
## lm(formula = charges ~ age + bmi + children + smoker + age.sq +   
## bmi.sq + child.sq, data = hw2.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10551 -3114 -1196 1702 30359   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -13518.329 3498.607 -3.864 0.000117 \*\*\*  
## age -87.357 82.479 -1.059 0.289726   
## bmi 792.804 206.940 3.831 0.000134 \*\*\*  
## children 1272.677 371.985 3.421 0.000642 \*\*\*  
## smokeryes 23813.533 408.529 58.291 < 2e-16 \*\*\*  
## age.sq 4.322 1.028 4.204 2.8e-05 \*\*\*  
## bmi.sq -7.542 3.251 -2.320 0.020496 \*   
## child.sq -185.366 100.799 -1.839 0.066142 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6021 on 1330 degrees of freedom  
## Multiple R-squared: 0.7541, Adjusted R-squared: 0.7528   
## F-statistic: 582.7 on 7 and 1330 DF, p-value: < 2.2e-16

#the square of the children variable is not statistically significant by t-test.  
model4=lm(charges~age+bmi+children+smoker+age.sq+bmi.sq+bmi\*smoker)  
summary(model4)

##   
## Call:  
## lm(formula = charges ~ age + bmi + children + smoker + age.sq +   
## bmi.sq + bmi \* smoker)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11500.7 -1759.9 -1351.3 -597.5 30367.1   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5.857e+03 2.810e+03 -2.084 0.037330 \*   
## age -3.580e+01 6.468e+01 -0.553 0.580041   
## bmi 5.499e+02 1.657e+02 3.318 0.000931 \*\*\*  
## children 6.689e+02 1.146e+02 5.838 6.62e-09 \*\*\*  
## smokeryes -2.019e+04 1.636e+03 -12.339 < 2e-16 \*\*\*  
## age.sq 3.769e+00 8.068e-01 4.672 3.29e-06 \*\*\*  
## bmi.sq -8.684e+00 2.600e+00 -3.340 0.000861 \*\*\*  
## bmi:smokeryes 1.435e+03 5.223e+01 27.467 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4815 on 1330 degrees of freedom  
## Multiple R-squared: 0.8427, Adjusted R-squared: 0.8419   
## F-statistic: 1018 on 7 and 1330 DF, p-value: < 2.2e-16

#According to individual t-test, the interaction term is statistically significant  
#Since we are keeping age.sq, we will also keep age as a predictor  
#RMSE=4815, adj R-squared=0.8419  
  
#Compare RMSE and adj R-squared to the multiple regression model  
summary(refined.model2)

##   
## Call:  
## lm(formula = charges ~ age + bmi + children + smoker, data = hw2.data)  
##   
## 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 \*\*\*  
## age 257.85 11.90 21.675 < 2e-16 \*\*\*  
## bmi 321.85 27.38 11.756 < 2e-16 \*\*\*  
## children 473.50 137.79 3.436 0.000608 \*\*\*  
## 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

#RMSE=6068, adj R-squared= 0.7489  
#This model with the quadratic and interaction terms (model4) has a smaller residual standard error and higher adjusted R-square value compared to the finalized multiple regression model (model2). Model 4's performance is better than Model 2's.

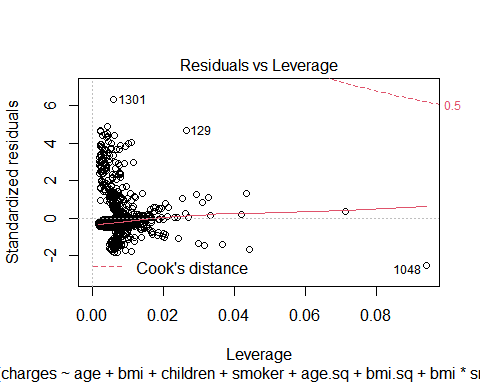
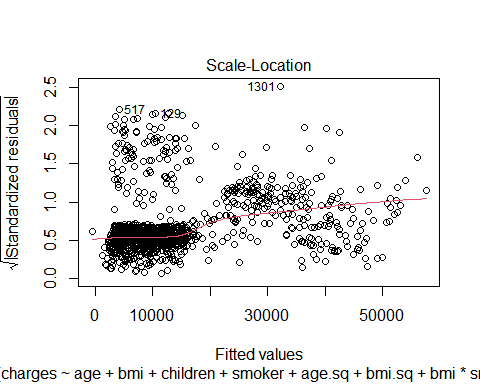
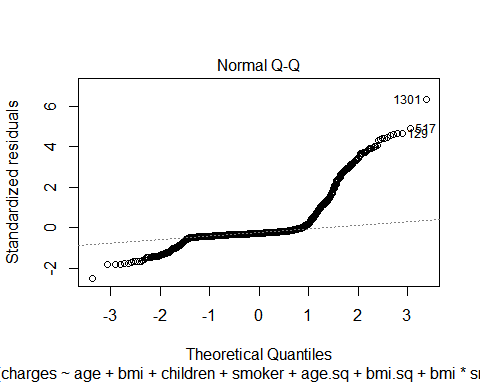
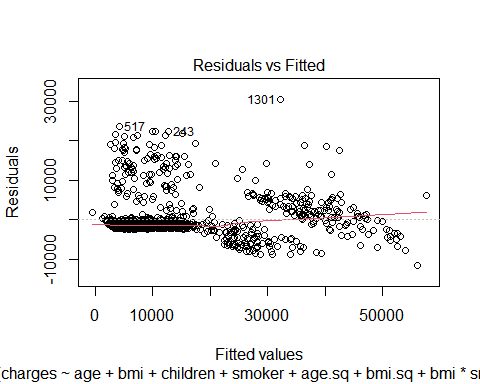
#Question 4  
  
testhw2data=hw2.data[,c(1,3,4)]  
cor(testhw2data)

## age bmi children  
## age 1.0000000 0.1092719 0.0424690  
## bmi 0.1092719 1.0000000 0.0127589  
## children 0.0424690 0.0127589 1.0000000

vif(lm(charges~age+bmi+children))

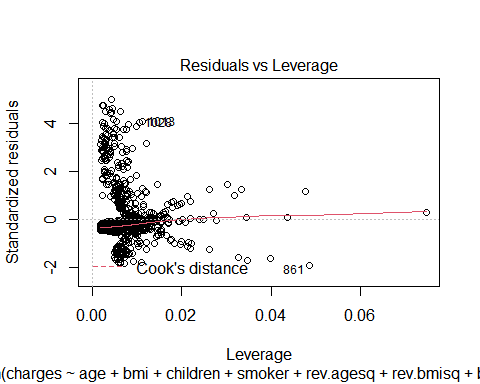
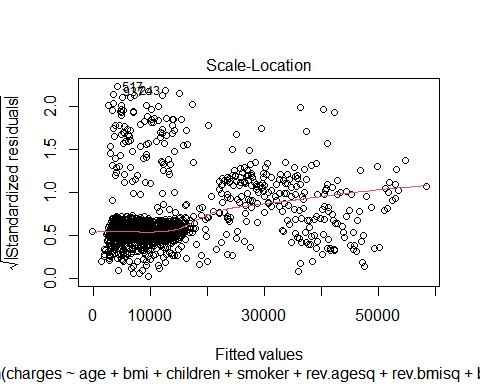
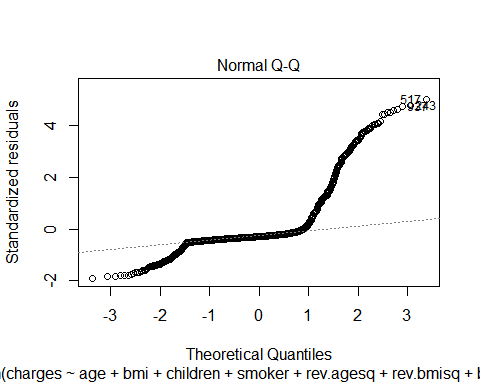
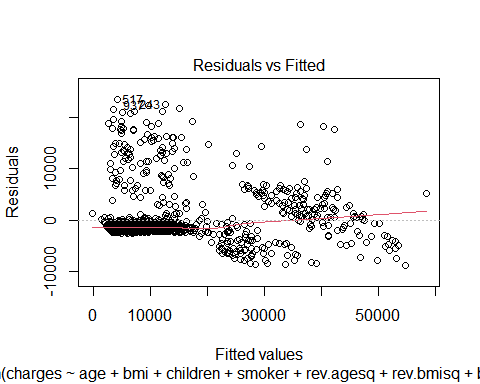
## age bmi children   
## 1.013816 1.012152 1.001874

#no multicollinearity, VIF < 5  
plot(model4)



#A residuals plot shows if the residuals have any non-linear patterns. Looking at Model 4's residual plot, we see that the residuals are not equally spread and in no distinct pattern. This could be a potential problem. Along with this, we have the leverage plot which determines if there are influential outliers. We can see three different outliers pointed out by the plot. The problem here is that these outliers can skew the fit line and over-influence it.

#Question 5  
revised.data=hw2.data[-c(1301,129,1048),]  
rev.agesq=(revised.data$age)^2  
rev.bmisq=(revised.data$bmi)^2  
  
  
model5=lm(charges~age+bmi+children+smoker+rev.agesq+rev.bmisq+bmi\*smoker, data = revised.data)  
plot(model5)



summary(model5)

##   
## Call:  
## lm(formula = charges ~ age + bmi + children + smoker + rev.agesq +   
## rev.bmisq + bmi \* smoker, data = revised.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8704.6 -1757.3 -1327.8 -619.9 23480.1   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.723e+03 2.804e+03 -1.685 0.09230 .   
## age -5.774e+01 6.313e+01 -0.915 0.36053   
## bmi 5.004e+02 1.660e+02 3.015 0.00262 \*\*   
## children 6.876e+02 1.118e+02 6.151 1.02e-09 \*\*\*  
## smokeryes -2.198e+04 1.631e+03 -13.476 < 2e-16 \*\*\*  
## rev.agesq 4.031e+00 7.873e-01 5.121 3.50e-07 \*\*\*  
## rev.bmisq -7.897e+00 2.606e+00 -3.030 0.00249 \*\*   
## bmi:smokeryes 1.488e+03 5.215e+01 28.536 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4695 on 1327 degrees of freedom  
## Multiple R-squared: 0.8479, Adjusted R-squared: 0.8471   
## F-statistic: 1056 on 7 and 1327 DF, p-value: < 2.2e-16

#Compared to model 4, RMSE is smaller (4695) and adjusted R-square is slightly higher (84.71% of variability in the medical costs can be explained by model 5's predictors)  
model5$coefficients

## (Intercept) age bmi children smokeryes   
## -4723.312383 -57.738250 500.376358 687.614282 -21981.144719   
## rev.agesq rev.bmisq bmi:smokeryes   
## 4.031368 -7.897179 1488.232927

#Below is the final model