Assignment2

Cameron Todd (5611230)

7 April 2017

Question 1

library("VGAMdata")

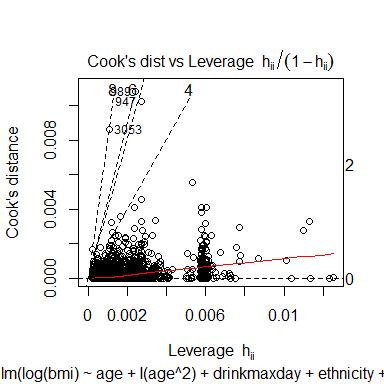
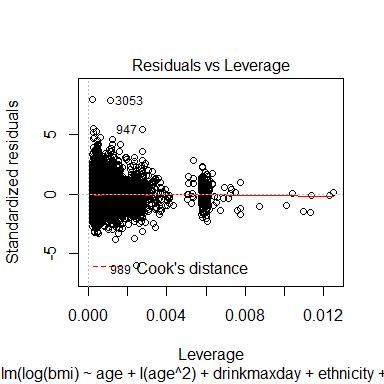
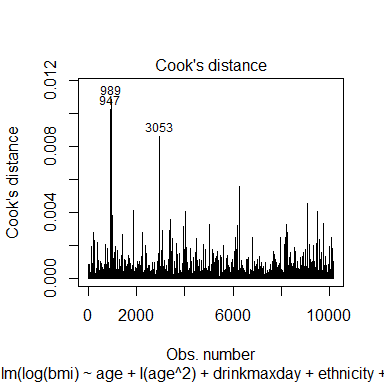
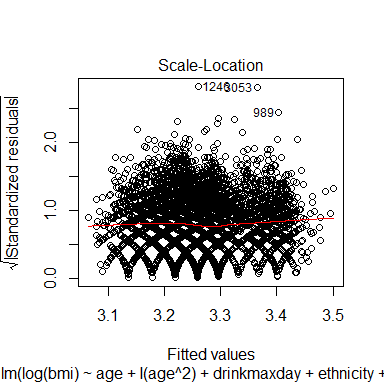
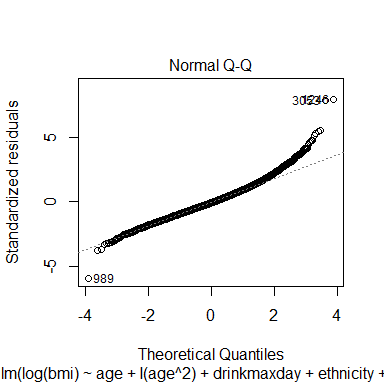
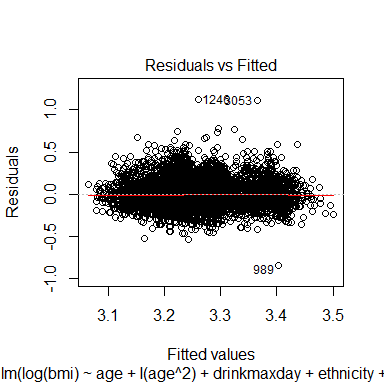
## Warning: package 'VGAMdata' was built under R version 3.3.3

data(xs.nz)  
educat.df = xs.nz[,c(3,4,9,10,30,50)]  
educat.df <- transform(educat.df, bmi = round(weight/height^2))  
final.df = na.omit(educat.df)  
  
summary(final.df)

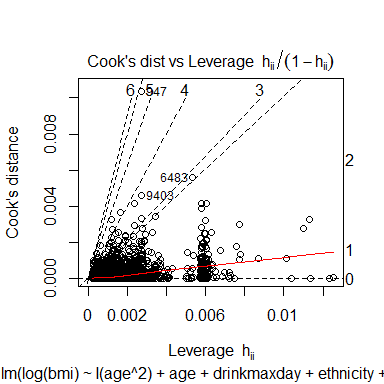
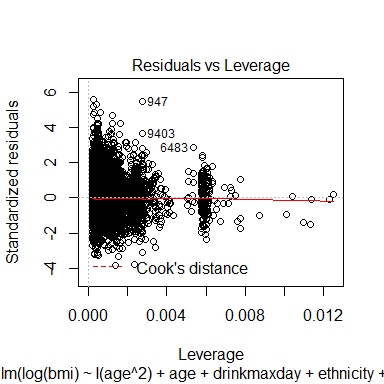
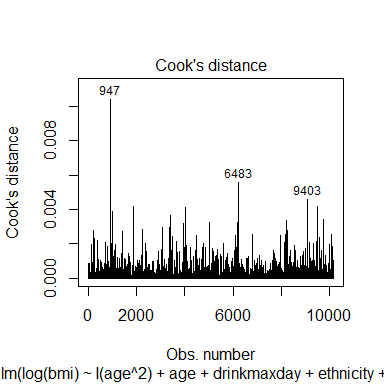
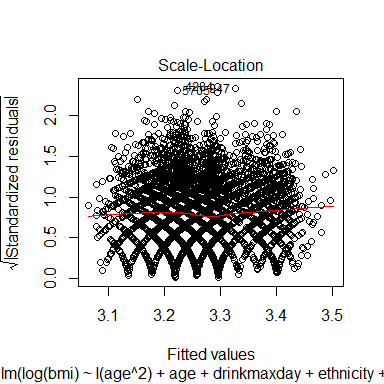
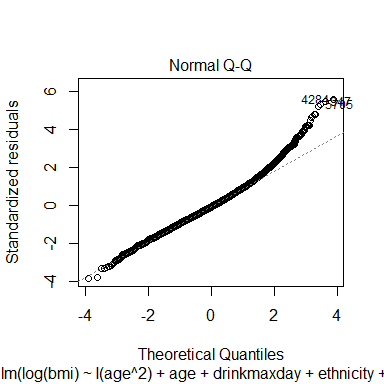
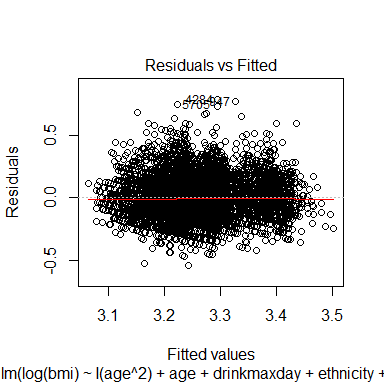
## age sex height weight   
## Min. :16.00 F:2877 Min. :1.060 Min. : 33.60   
## 1st Qu.:32.00 M:7311 1st Qu.:1.670 1st Qu.: 68.00   
## Median :42.00 Median :1.730 Median : 78.00   
## Mean :43.69 Mean :1.723 Mean : 78.55   
## 3rd Qu.:52.00 3rd Qu.:1.790 3rd Qu.: 88.00   
## Max. :88.00 Max. :2.050 Max. :199.00   
## drinkmaxday ethnicity bmi   
## Min. : 0.000 European :8553 Min. :13.00   
## 1st Qu.: 2.000 Maori :1018 1st Qu.:24.00   
## Median : 6.000 Polynesian: 442 Median :26.00   
## Mean : 7.325 Other : 175 Mean :26.37   
## 3rd Qu.:10.000 3rd Qu.:29.00   
## Max. :90.000 Max. :88.00

Comment - In the dataset we obviously have 7 variables, 5 of which are continuous variables and 2 variables (ethnicity and sex) are categorical variables. The data appears to be consistent as we have dealth with the missing values and after viewing and summary, the data appears to have no special values or unusual observations or obvious errors and therefore the data is clean to fit a model.Also their appears to be no skewage in the data with the continuous variables means and medians being almost the same.

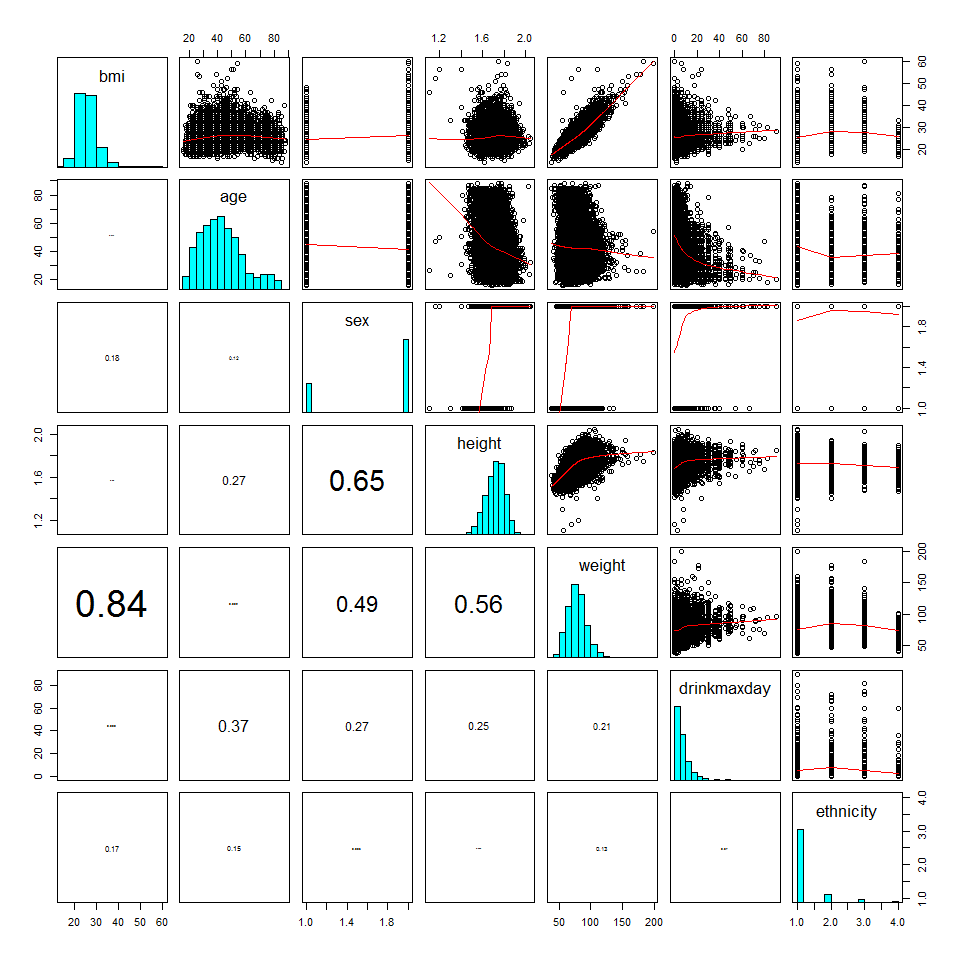
BMI.lm = lm(log(bmi) ~ age + I(age^2) + drinkmaxday + ethnicity + sex, data = final.df )  
  
hat.extreme <- 3\*length(coef(BMI.lm))/nrow(final.df)  
plot(BMI.lm, which = 1:6)

 We can clearly see in the residuals vs leverage plot that we have observation 989 3053 and 1246 as fairly low-leverage large outliers, all three are the only real influential points and therefore will be removed from the model.

library(s20x)  
final.df1 = educat.df[-c(989,3053,1246),]  
final.df1 = na.omit(final.df1)  
BMI.lm2 <- lm(log(bmi) ~ I(age^2) + age +  
 drinkmaxday + ethnicity + sex, data = final.df1)  
plot(BMI.lm2, which = 1:6)



pairs20x(bmi ~ ., data = final.df1)

 After taking out the 3 outliers and refitting the model, it is safe to say their are no more outliers that are needed to be removed. The cooks plot shows observations 947, 6483 and 9403 and potential outliers but are all below 0.4. These observations also show up in the residuals vs leverage plot but as fairly low leverage observations and therefore won't be influential in the model. I also plotted a pairs plot to check for any multicollinearity. After reviewing the plot it appears their are no issues and with the highest correlation between the explanatory variables of 0.65 between sex and height, this is expected, no further actions required.

summary(BMI.lm2)

##   
## Call:  
## lm(formula = log(bmi) ~ I(age^2) + age + drinkmaxday + ethnicity +   
## sex, data = final.df1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.53493 -0.09325 -0.00924 0.08302 0.78233   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.913e+00 1.186e-02 245.571 < 2e-16 \*\*\*  
## I(age^2) -9.852e-05 4.983e-06 -19.771 < 2e-16 \*\*\*  
## age 1.129e-02 4.961e-04 22.765 < 2e-16 \*\*\*  
## drinkmaxday 1.755e-03 1.991e-04 8.813 < 2e-16 \*\*\*  
## ethnicityMaori 1.109e-01 4.752e-03 23.344 < 2e-16 \*\*\*  
## ethnicityPolynesian 1.374e-01 6.876e-03 19.985 < 2e-16 \*\*\*  
## ethnicityOther -6.153e-02 1.074e-02 -5.730 1.04e-08 \*\*\*  
## sexM 4.908e-02 3.246e-03 15.118 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1401 on 10177 degrees of freedom  
## Multiple R-squared: 0.1656, Adjusted R-squared: 0.165   
## F-statistic: 288.5 on 7 and 10177 DF, p-value: < 2.2e-16

print( 100 \* (exp(0.001755) - 1))

## [1] 0.1756541

print(100 \* (exp(0.109) - 1))

## [1] 11.51624

print(100 \* (exp(0.1374) - 1))

## [1] 14.7287

print(100 \* (exp(-0.06153) - 1))

## [1] -5.967526

print(100 \* (exp(0.04908) - 1))

## [1] 5.030437

I had to exponentiate and times by 100 to make the summary coeffficients appropriate to interpret I am also not going to interpret age and quadratic of age as requested in question. The summary of the model shows that the drinkmaxday variable, ethnicity and sex are all significant variables to the median BMI. drinkmaxday shows that every extra one unit drink increase it can have roughly a .18 % increase on the median BMI.. Ethnicity - The model shows that if the person is Maori, they generally are 11.5% larger then the median BMI population. A Polynesian person will generally be 14.7% larger then the median BMI population and other ethnicities will generally be have a BMI of 6% less then the median population. Sex - It appears that SexM is linearly posititve to BMI, if you are male you generally have a BMI that is 5% larger then the median population .

1. I am reading this question that is asking for what age is giving rise to the larest BMI values and therefore we need to find the maximum value by getting the first derivative of the model above with respect to age and then set the gradient to 0 and solve for age. I did this with the values above and deriving the model by age which =
2. . I then solved for age which gave me roughly 57 years old tends to give the highest BMI's

xmat = model.matrix(BMI.lm2)[,-1]  
round(diag(solve(cor(xmat))),2)

## I(age^2) age drinkmaxday   
## 28.56 29.02 1.26   
## ethnicityMaori ethnicityPolynesian ethnicityOther   
## 1.05 1.02 1.01   
## sexM   
## 1.11

Appears to be no evidence of multicollinearity for this data

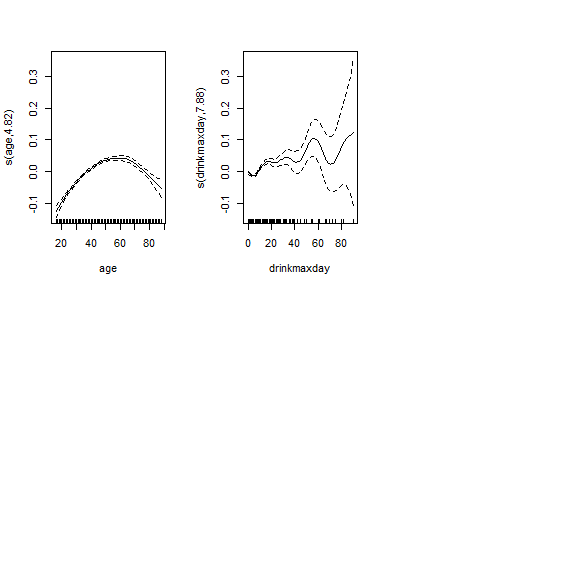
library('mgcv')

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

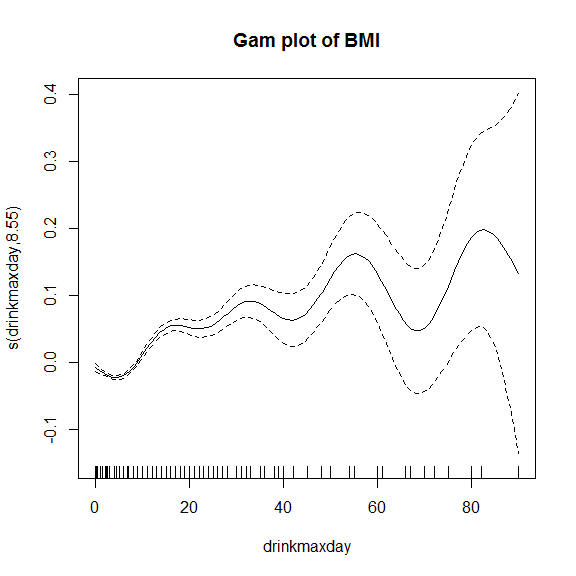
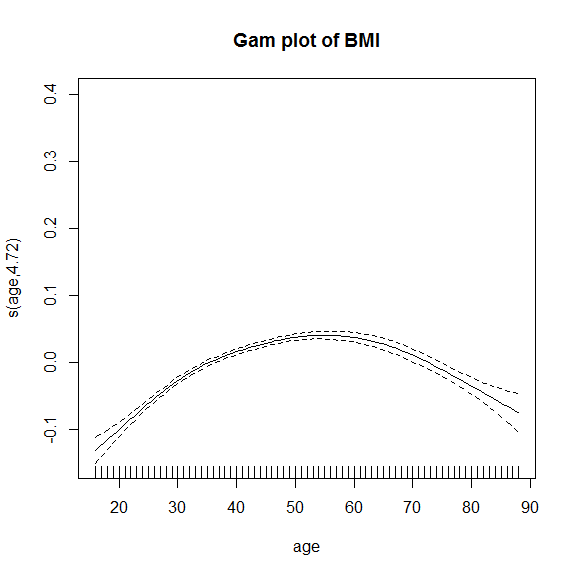
## Loading required package: nlme

## This is mgcv 1.8-17. For overview type 'help("mgcv-package")'.

bmi.gam1 <- gam(log(bmi) ~ s(age)+s(drinkmaxday)+ethnicity+sex,  
 data = final.df1)  
par(mfrow = c(2, 3))  
plot(bmi.gam1)  
par(mfrow = c(1, 1))

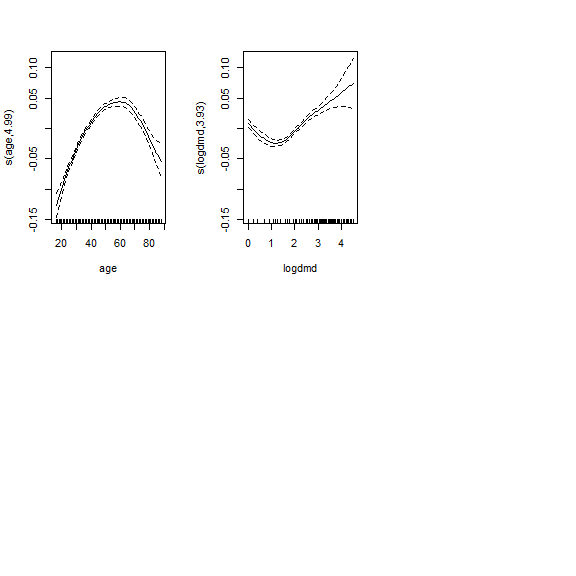


plot(gam(log(bmi)~ s(age) + s(drinkmaxday), data = final.df1), lwd = 1, col = "black", pch=12, main = "Gam plot of BMI")

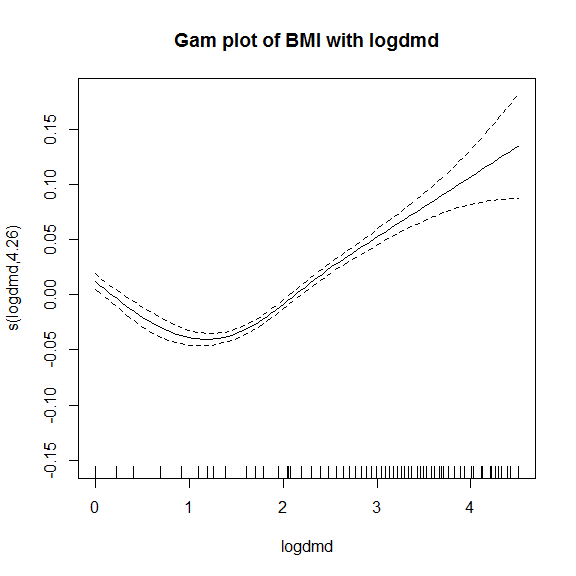
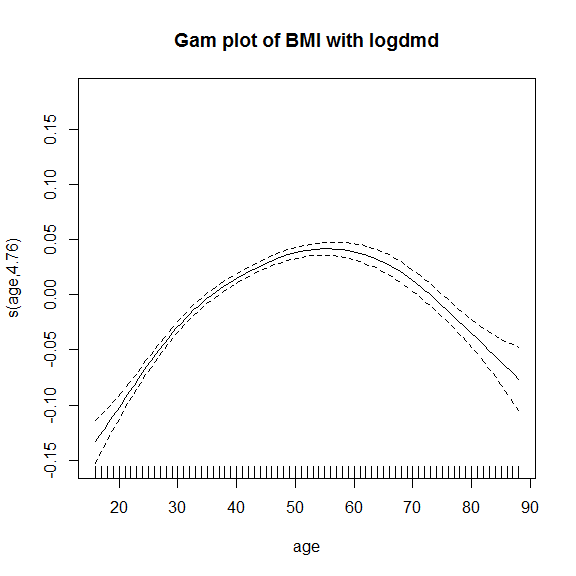


After observing the additive model above, It looks like a transformation of Age and drinkmaxday could be useful. We would try a cubic polynomial for Age, however the gam model also suggests a polynomial function to the power of 7 for the variable drinkmaxday, this polynomial would make the model very complex and hard to logically make sense of, io think it would be best to log the drinkmaxday variable to help bring down the recommended polynomical factor.

final.df1 <- transform(final.df1, logdmd = log1p(drinkmaxday))  
bmi.gam2 <- gam(log(bmi) ~ s(age)+s(logdmd)+ethnicity+sex,  
 data = final.df1)  
par(mfrow = c(2, 3))  
plot(bmi.gam2)  
par(mfrow = c(1, 1))



plot(gam(log(bmi)~ s(age) + s(logdmd), data = final.df1), lwd = 1, col = "black", pch=12, main = "Gam plot of BMI with logdmd")

 After adding a new varialbe logdmd which is computed to log(1+drinkmaxday), we do this because we know some of the observations would have max drink per day values of 0, these obeservations if when logged would errors log(0) = - infinity and therefore log1p just creates the log(0) observations to be 0. The plots above have shown a better result then before as it suggests we can transform log(drinkmaxday)into only a quadratic which is much simpler to interpret then before with 7 powers. Age has relatively stayed the same with the component function suggesting also a quadratic as before

bmi.lm3= lm(log(bmi) ~ age + I(age^2) + logdmd + I(logdmd^2) + ethnicity + sex, data = final.df1)  
summary(bmi.lm3)

##   
## Call:  
## lm(formula = log(bmi) ~ age + I(age^2) + logdmd + I(logdmd^2) +   
## ethnicity + sex, data = final.df1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.52365 -0.09179 -0.00937 0.08191 0.78728   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.920e+00 1.195e-02 244.320 < 2e-16 \*\*\*  
## age 1.190e-02 5.013e-04 23.743 < 2e-16 \*\*\*  
## I(age^2) -1.050e-04 5.059e-06 -20.746 < 2e-16 \*\*\*  
## logdmd -3.531e-02 4.456e-03 -7.924 2.54e-15 \*\*\*  
## I(logdmd^2) 1.435e-02 1.356e-03 10.585 < 2e-16 \*\*\*  
## ethnicityMaori 1.069e-01 4.772e-03 22.397 < 2e-16 \*\*\*  
## ethnicityPolynesian 1.304e-01 6.958e-03 18.734 < 2e-16 \*\*\*  
## ethnicityOther -6.652e-02 1.078e-02 -6.174 6.93e-10 \*\*\*  
## sexM 4.815e-02 3.272e-03 14.714 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1397 on 10176 degrees of freedom  
## Multiple R-squared: 0.1704, Adjusted R-squared: 0.1698   
## F-statistic: 261.3 on 8 and 10176 DF, p-value: < 2.2e-16

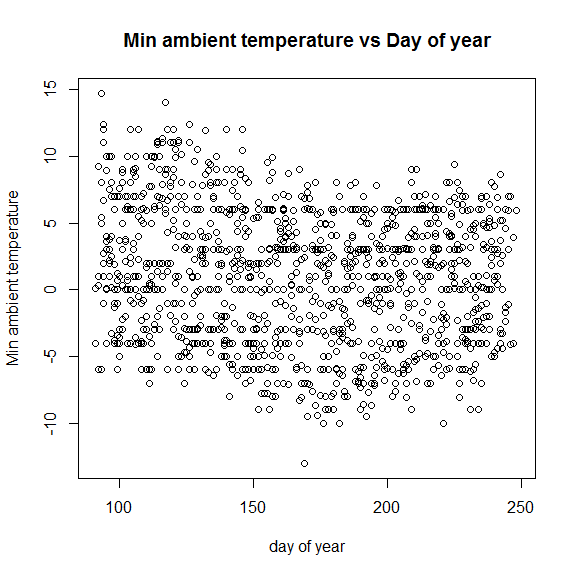
The model using log(bmi) as a response with the age variable as a quadratic and the logged drink max per day variable also as a quadratic with the other explanatory variable of ethnicity and sex does not fit very well at all. The R-squared of this model of 0.17 meaning the model only explains 17% of all variability in the data, therefore the model should definitely not be relied upon for any sort of prediction. However it is better then the model we previous developed with their R-squared of only .16. As you can see from the summary all the variables do have a significant relationship with the median BMI. Mathmatical Formula =

where

and ethm, ethp, etho = 1, if person has ethnicity fo maori, polynesian or other ethnicity respectively and sexM = 1, if person is male

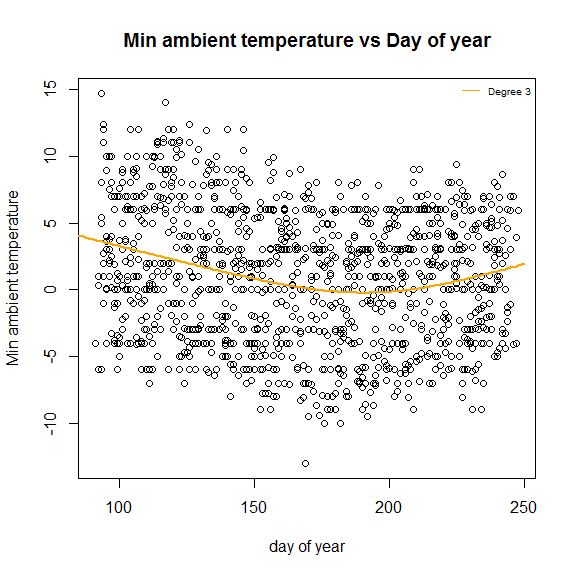
Q2 (A)

data("trapO", package = "VGAMdata")  
plot(trapO$doy,trapO$MinAT, main = "Min ambient temperature vs Day of year", ylab = "Min ambient temperature", xlab = "day of year")



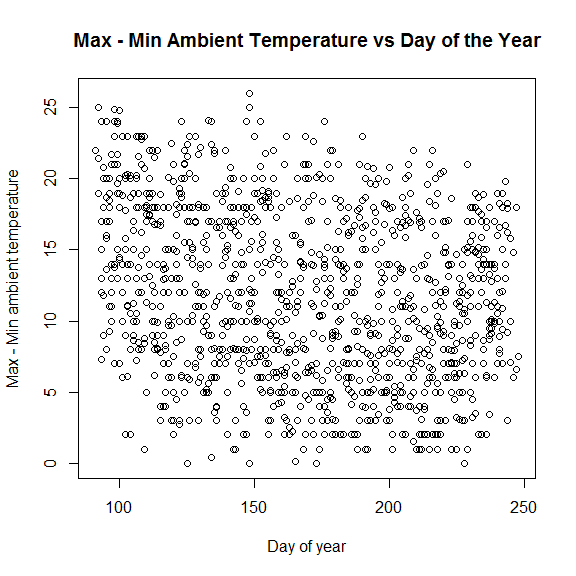
The plot appears to have fairly constant scatter and their appears to be slight curvatures, meaning their will be a higher power variable. Their appears to be no outliers (obviously you can see their was a very cold day, well below -10 at some stage). The plots show curvature which is what we expected for some place in the world where it is seasonal with Summer and Winters. Their appears to be no skewage in the data.

library("splines"); n = 50; set.seed(123); knots = 1:10  
  
mycol <- c("Purple", "green4", "orange", "blue")  
myCex <- 1  
my.cex <- 1  
mylty <- c("solid", "solid", "solid", "dashed")  
mylwd <- 2  
mymgp <- c(1.6, 0.6, 0)  
dd <- seq(0, 250, length.out = 250)  
  
trap.df = na.omit(trapO)  
mysp1 <- lm(MinAT ~ ns(doy, df = 3), data = trap.df)  
plot(trap.df$doy,trap.df$MinAT, main = "Min ambient temperature vs Day of year", ylab = "Min ambient temperature", xlab = "day of year"  
 )  
  
  
lines(dd, predict(mysp1, data.frame(doy = dd)),  
 col = mycol[3], lty = mylty[3], lwd = mylwd)  
  
legend("topright", bty = "n",  
 legend = paste("Degree 3"),  
 col = "orange",cex = 0.6,  
 lty = c("solid"), lwd = 1.1)

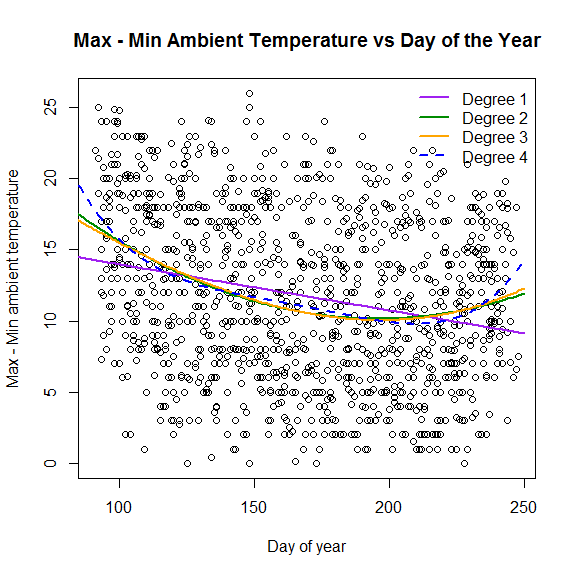


I liked the B-spline with a quadratic spline which has a continuous derivative up to order M - 2 = 1 at the knots- this is aka a parabolic spline, it is also linear beyond the boundaries, it appears to be not undersmoothed or oversmoothed, their appears to be no outliers that affect the ends of the line

plot(MaxAT-MinAT ~doy, data = trap.df, main = "Max - Min Ambient Temperature vs Day of the Year", xlab = "Day of year", ylab = "Max - Min ambient temperature", col = "black", cex = 1)

 The above plot shows very large levels of scatter, it appears to be negatively correlated between Day of year and Max-Min temperatures. Their appears to be no outliers but very hard to spot, we may see higher leverage points when plotting a large degree smoother against it, as their appears to be larger scatter in the early days of the years which might create the corners of the smoothers to spike up.

plot(MaxAT-MinAT ~doy, data = trap.df, main = "Max - Min Ambient Temperature vs Day of the Year", xlab = "Day of year", ylab = "Max - Min ambient temperature", col = "black", cex = 1)  
#doy.spl = with(trap.df, smooth.spline(MaxAT - MinAT, doy))  
  
  
for (Degree in 1:4) {  
 fm <- lm(MaxAT-MinAT ~ poly(doy, Degree), data = trap.df)  
 assign(paste("Ambient temperature", Degree, sep = "."), fm)  
 lines(dd, predict(fm, data.frame(doy = dd)),  
 col = mycol[Degree], lty = mylty[Degree], lwd = mylwd)  
}  
  
  
  
  
legend("topright", bty = "n",  
 legend = paste("Degree", as.character(1:4)),  
 cex = myCex, col = mycol,  
 lty = mylty, lwd = mylwd)



I wanted to plot all 4 of the different poly smoothers up to degree of 4. I think after reviewing all the smoothers a poly of degrees 2 appears to be fine, this is a smaller degree then the above smoother using the response variable of Min temperature, therefore fitting a squared polynomial of Day of Year to the max min temperatures might be best. Obviously if you can have a smaller polynomial of 2 instead of 3 or 4 the models is alot less complex