BLUE group NHANES Report

Caleb Ki, Stephany Flores-Ramos, Jordan Browning

November 9, 2016

NHANES Use the `NHANES` training dataset to fit and interpret a linear regression model of BMI (body mass index) as a function of being physically active, using alcohol, age, gender, and poverty status.

Be sure to report RMSE for the training set and for the test set.

Your report should provide background on these data, describe the analytic sample, fit and interpret the model, and undertake model assessment. You should include one figure that summarizes key findings.

SOLUTION:

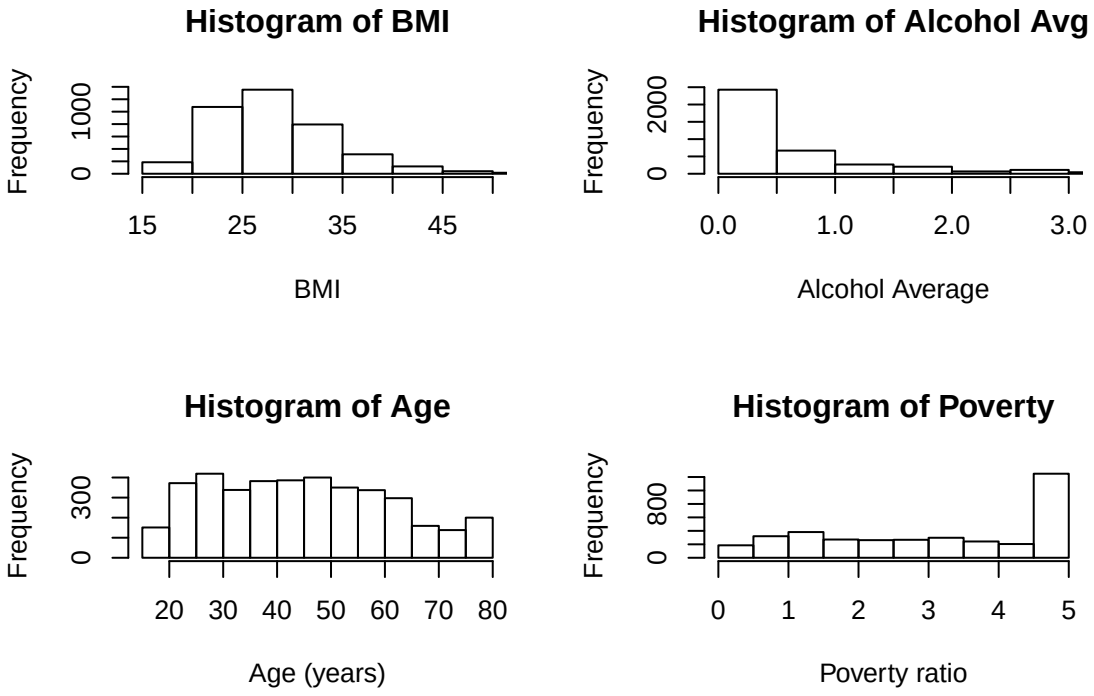
The dataset

The NHANES dataset is survey data from 2009-2012 gathered by the US National Center for Health Statistics (NCHS), who started collecting health and nutrition survey data in the 1960s. This data is a collection of survey data on health and nutrition topics along with vital health information collected via a health examination in one of NCH's mobile examination centers (MEC). Since the target population is "non-institutionalized civilian resident population of the United States", the sample is designed to oversample certain subpopulations. This is important to any analysis performed on the dataset as they could have huge implications on any conclusions drawn.

Sample Description

To fit the model a random sample of 3929 was taken from the NHANES dataset, which will be called our training dataset. This training dataset is 4/5 of the original NHANES, where there were 4912 observations after filtering for those above 18 and removing observations where there were missing values. We have chosen to remove those below 18 because alcohol consumption was not recorded for those below 18. We couldn't just code these values to 0, since the fact that they were not recorded does not necessarily mean that the amount of alcohol consumed for those below 18 was 0.

The following is a list of variables that we are concerned with and a look at them in our training dataset:



BMI:
Body mass index calculated by $\frac{weight}{height^2}$ in $\frac{kg}{m^2}$. The distribution of BMI is nearly normal, though it is slightly right-skewed. Most of the values seem to lie within the 20-30 range. There are 50 missing data points for BMI.

PhysActive:
A dichotomous variable. Inactive or active depending on the answer to the question, does the participant do moderate or vigorous-intensity sports or recreational activities? (Inactive for no and active for yes).

There are slightly more physically active people in this sample than not, with 52.97% being physically active and 47.02% being not. There are no missing data points for physically active.

alcoholavg:
The average number of drinks consumed per day over the past year. This was found by multiplying the AlcoholDay variable which is the average number of drinks consumed on days that participant drank alcoholic beverages and the AlcoholYear variable which is the estimated number of days over the past year that participant drank alcoholic beverages then we divided by 365 to get the average. This was done because the original variables concerning alcohol were not a comprehensive in their description of alcoholic behavior. This new coded variable tries to account for both frequency and quantity of drinking.

Nearly all have less than one drink per day. There are some high outliers with a max of 23 drinks per day on average during the year, so the distribution is right skewed. The observation with 23 drinks per day on average was removed as upon further inspection, it was reported that he drank 82 drinks on average for every occasion he drank. There were 2045 missing data points for alcohol average.

Age:
The age in years of the study participant at the time of the screening. For the age range 20-60, the proportion of people at each age seems to be relatively the same. However, there are fewer people in the dataset who are under 20 and over 60. There are no missing data points for age.

Gender:
Gender of participant coded as male or female. The number of males and females are roughly equal, with 49.21% of the sample being male and 50.79% being female. There are no missing data points for gender.

Poverty:
Ratio of family income to poverty guidelines. Kept the ratio instead of using an indicator variable. This way

the model gets more information about their income rather than just that they are above/below the poverty line.

There sample is bimodal. However, the first mode is the much less frequent value. There is a peak around the poverty value 1 which shows that those who are poorer generally have family incomes close to the poverty level. The other peak is around the value 5. This shows there is a significant proportion of people who have family incomes greater than the poverty line by a factor of 5+ . There are 461 missing data points here.

Our model

```
mod <- lm(BMI ~ PhysActive + Age + Gender + Poverty + alcoholavg, data=train, na.action = na.exclude)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	27.6852	0.3610	76.68	0.0000
PhysActiveInactive	1.5536	0.2173	7.15	0.0000
Age	0.0157	0.0066	2.39	0.0171
Gendermale	0.8066	0.2099	3.84	0.0001
Poverty	-0.2354	0.0660	-3.57	0.0004
alcoholavg	-0.4656	0.0843	-5.52	0.0000

In this model, we see that accounting for the effects of other variables, the mean BMI for a population who are physically inactive is 1.56 more than that of a physically active population. Then, if a person is male, accounting for other variables, we see that the mean BMI for this population is 0.81 higher as compare to the mean BMI of a female population. In addition, accounting for other factors, with each unit increase in the average number of drinks consumed per day over the past year average, the mean BMI decreases by 0.47. Furthermore, this model tells us that after accounting for the effects otf other variables the mean BMI of a population increases by 0.02 for each additional year in someone's age and this mean BMI actually decreases by 0.24 with each 1 unit increase in the ratio of family income to poverty guidelines.

Something else we can note from this model is that it also tells us that all of these factors are statistically significant at the 0.05 α level. This is an indicator that these variables perhaps do not add much predictive value. Furthermore, the R^2 value for this model was only 0.03 which means that a lot of the variability of the dataset is unaccounted by the model. If we were to recreate the model, we would heavily consider removing the aformentioned variables since they do not exhibit statistical significance or improve R^2.

Cross-Validation

In addition to checking the LINE assumptions, we also used a cross validation method to test the robustness of our model against a test set, a dataframe with the other observations from the `NHANES` dataset not included in the training set.

```
test$predictVal <- predict(mod, test)
test2 <- test %>% filter(!is.na(predictVal)) %>% filter(!is.na(BMI)) %>% mutate(sqerror = (BMI -
    predictVal)^2)
train$predictVal <- predict(mod, train)
train2 <- train %>% filter(!is.na(predictVal)) %>% filter(!is.na(BMI)) %>% mutate(sqerror = (BMI -
    predictVal)^2)

mseTest <- sum(test2$sqerror)/nrow(test2)
rmseTest <- sqrt(mseTest)  #rmse for test set
rmseTest
```

```
## [1] 6.73121
```

```
mseTrain <- sum(train2$sqerror)/nrow(train2)
rmseTrain <- sqrt(mseTrain)  #rmse for training set
rmseTrain
```

```
## [1] 6.206139
```

```
mean(NHANES$BMI, na.rm = TRUE)  #mean of BMI values
```

```
## [1] 28.38323
```

```
median(NHANES$BMI, na.rm = TRUE)  # median of BMI values
```

```
## [1] 27.4
```

```
rsquared <- 1 - sum(test2$sqerror)/sum((test2$BMI - mean(test2$BMI))^2)
```

The root mean squared error is 6.7312103 for the test set and 6.206139 for the training set. This shows that the model performs similarly for both the data it was fitted for and the data it was tested against, although the rmse is slightly higher for the training set. While the performance is consistent across the two datasets, the rmse values are very high relative to the values that the BMI takes. The average value of BMI is around 26 which is only around 4 times as large as the rmse. Additionally, the R^2 value of the model on the test set is 0.0251321, which shows that only 2.513% of the variability in BMI is captured by the model. A Overall, these measures show that the model is not a great fit for the data and that the actual predictive value of the model is not very good. I looked at the scatterplots of the different predictors against BMI but there seems to be a lot of noise. There doesn't seem to be an obvious relationship, so I believe that not even a transformation would help.

Model Assesment

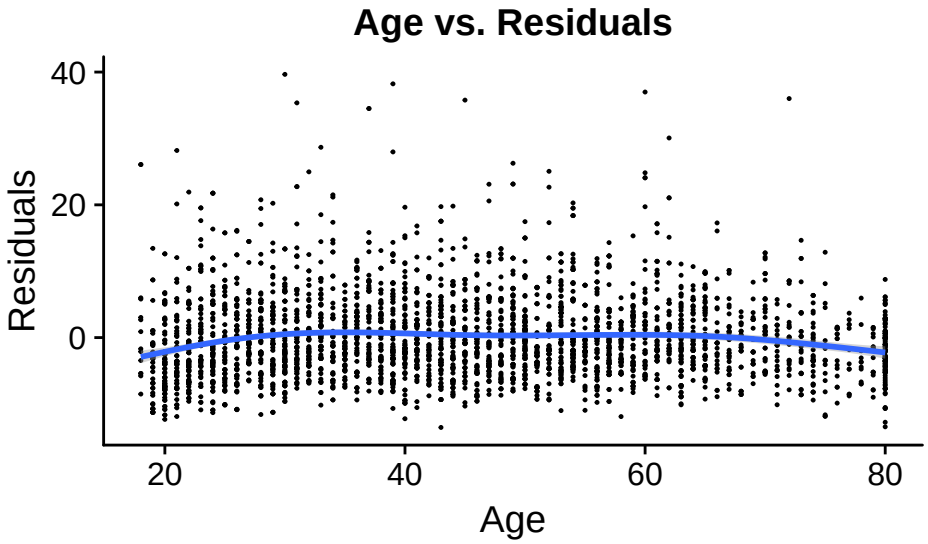
LINE

To check the appropriateness of our model we checked the following assumptions:

Linearity

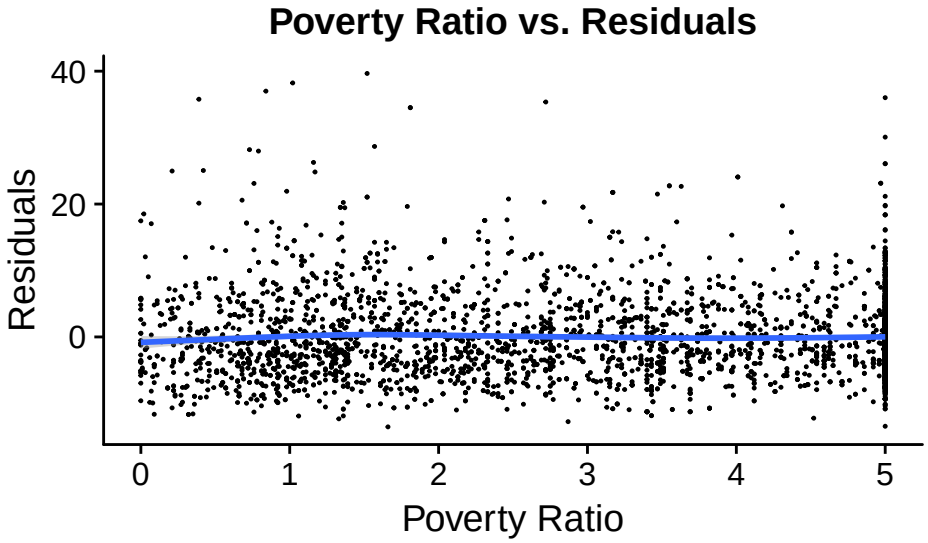
To test the linearity assumption we look at scatterplots where a quantitative variable is on the x-axis and the residuals from the model are on the y-axis. In this case, there are three quantitative predictors so we look at three different scatterplots. If the linearity assumption is satisfied than the residuals would be symmetrically distributed across a horizontal line with roughly constant variance. If the linearity assumption is not satisfied there is generally some pattern or curvature that can be found in the data. For categorical variables, we will use boxplots.

```
## `geom_smooth()` using method = 'gam'
```

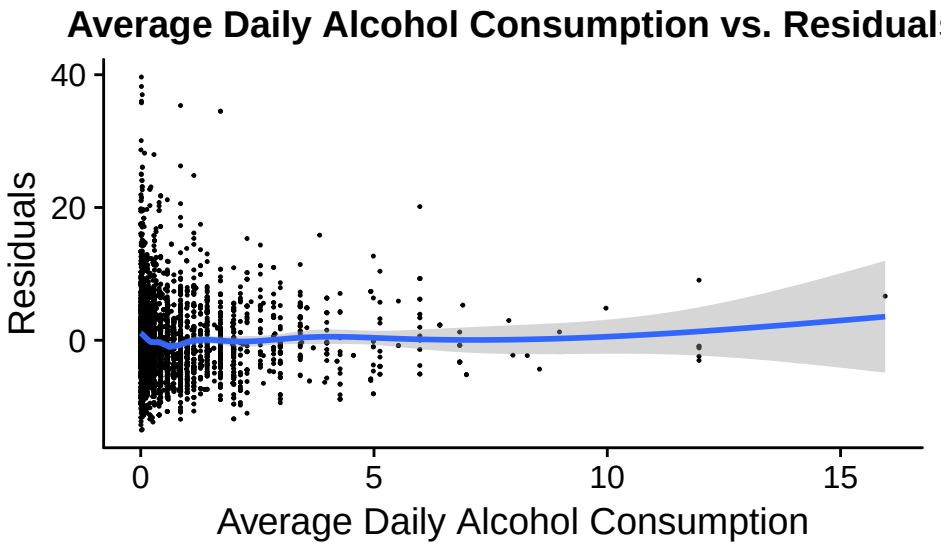


From this scatterplot, the residuals do seem to have roughly constant variance for all values of age, and that they are relatively symmetric across a horizontal line (although there are many more values with extreme positive residuals than extreme negative residuals). We say that the linearity condition is satisfied in this case.

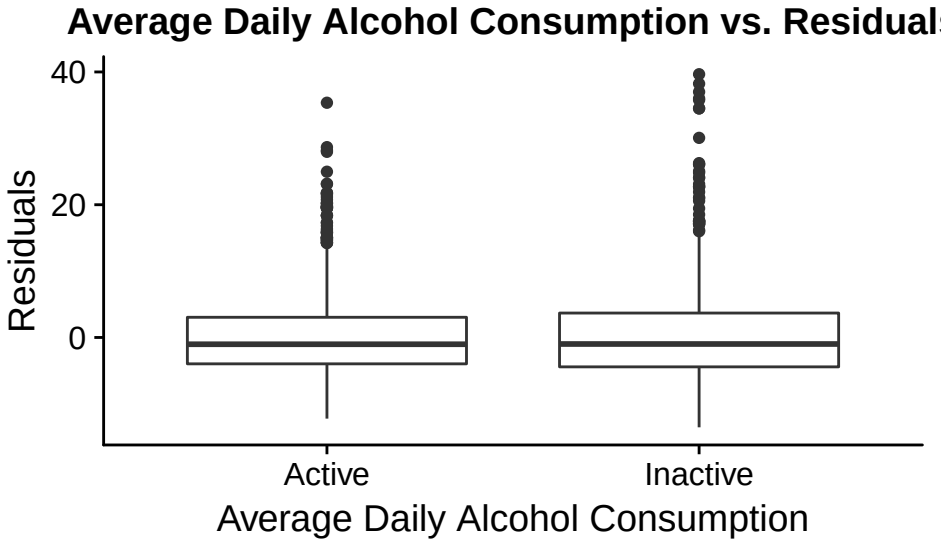
```
## `geom_smooth()` using method = 'gam'
```

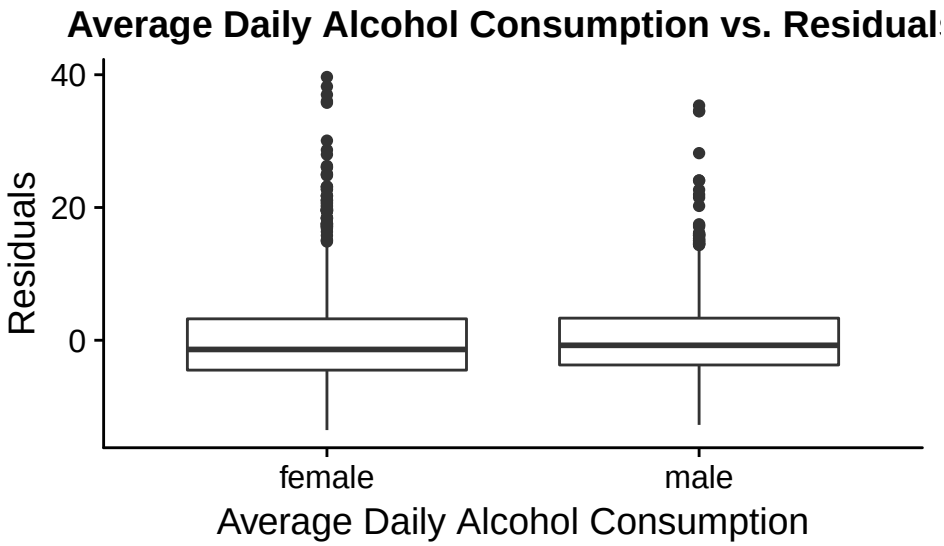



```
## `geom_smooth()` using method = 'gam'
```



The linearity condition is not satisfied for the alcoholavg variable. Clearly there is a curvature in the plot which is an indicator that there may not be a linear relationship between BMI and alcoholavg. In addition, there is a point with high leverage. While it would be outside the scope of this analysis, it may be prudent to see what happens to the model if we remove this outlier.

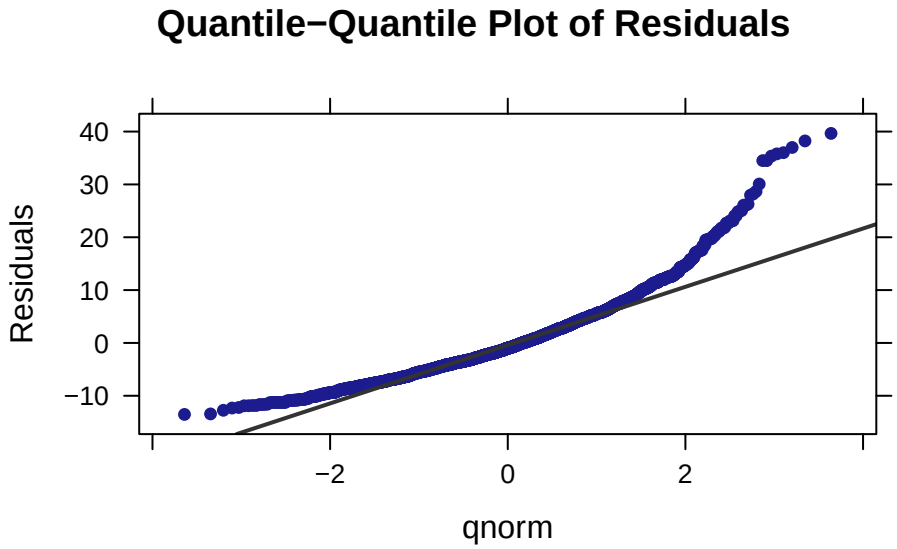




For both categorical variables, the box plots for each group within the variables looks relatively the same. The distribution

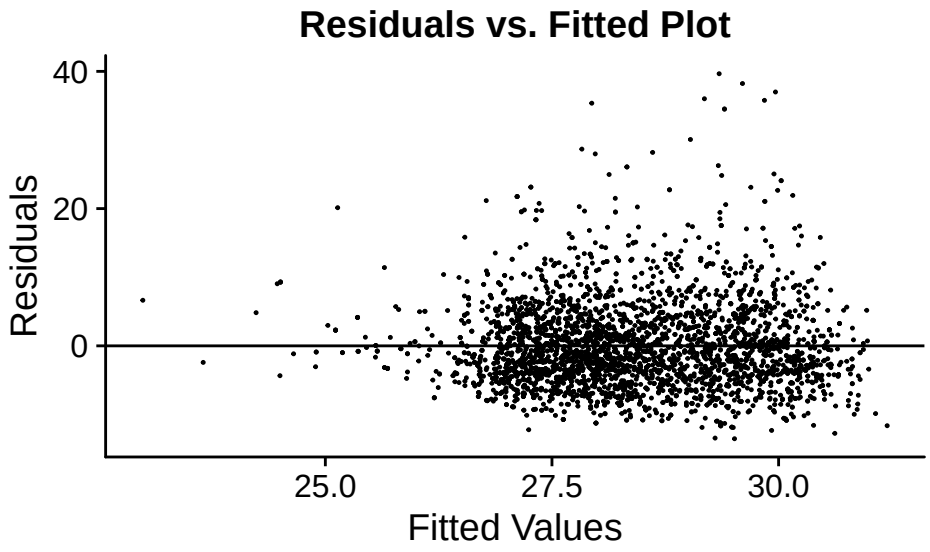
Normality

```
qqmath(~residuals(mod), ylab = "Residuals", main = "Quantile-Quantile Plot of Residuals")
ladd(panel.qqmathline(residuals(mod)))
```



Based on the qqplot above we can see that our model does not meet the normality assumption because we can see that both tails of the residual distribution are highly skewed.

Equal Variance



The constant variance condition of errors does not hold. Although there are less points with smaller fitted values, the residuals are less variable for the smaller fitted values. As the fitted values become larger, the residuals become slightly more variable. However, there is a point where as the fitted values increase, the variance of the residuals decreases.

Collinearity

```
corTrain <- cbind(train$Age, train$alcoholavg, train$Poverty)
cor(corTrain, use = "complete.obs")
```

```
##              [,1]         [,2]        [,3]
## [1,]  1.000000000 -0.003166747  0.22456550
## [2,] -0.003166747  1.000000000 -0.05724508
## [3,]  0.224565496 -0.057245085  1.00000000
```

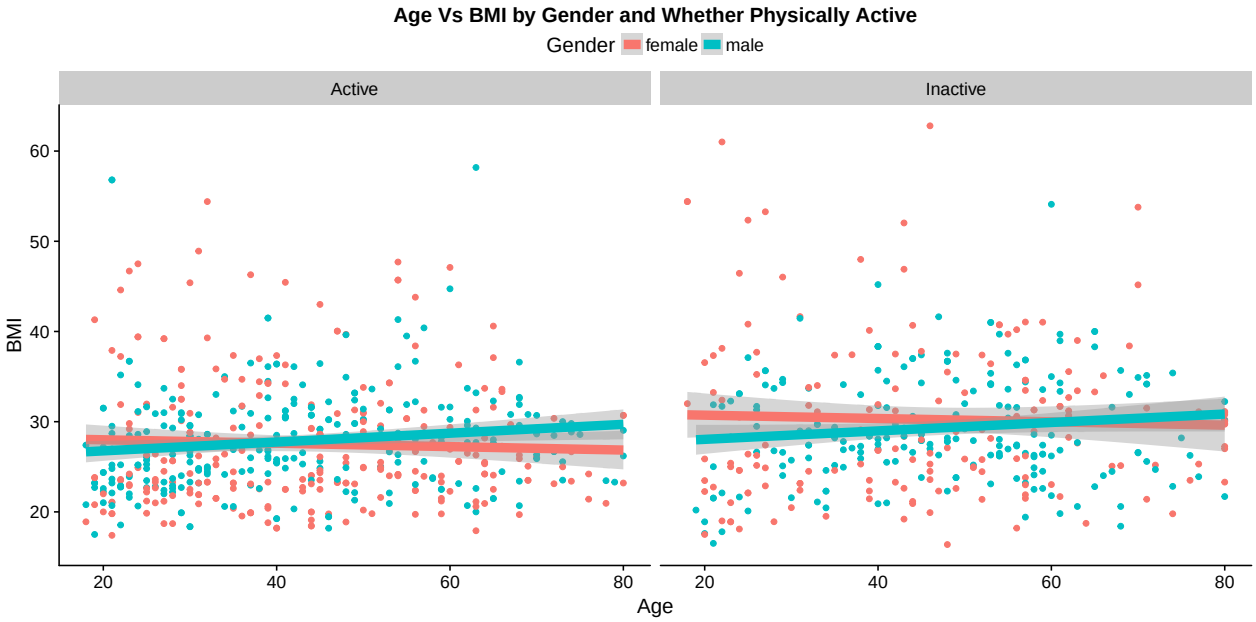
From the correlation matrix (ignoring the diagonals since we always expect one variable to have perfect correlation with itself), none of the three quantitative variables show a concerning level of correlation with each other. The correlation values are all below .25 so we can assume that there is no multicollinearity or only a little bit.

Independence

We assume independence because NHANES goes through a 4 stage sampling process in order to get their survey sample. First they sample counties, then segments within those counties, then households in those segments, and finally people within those households. Thus we can assume each observation in the sample is independent.

Visual Accounting for Noise

```
ggplot(data=test2, aes(x=Age, y=BMI)) +
  geom_point()  + aes(colour=Gender) +
  facet_wrap(~PhysActive, ncol=2)  + stat_smooth(method=lm, size=3) +
  theme(legend.position="top") +
  labs(title="Age Vs BMI by Gender and Whether Physically Active") +
  theme(axis.text = element_text(size = 12))
```



The purpose of this plot is to demonstrate how noisy the variables actually are. An initial look at the graph seems to show that there is almost no difference in BMI for different ages, for different genders, or for different levels of physical activity. However, as the model indicates, there is some slight differences. It is with the help of the lines in the graph that we can ascertain some of the small differences. The lines in the graph show that those who are physically active seem to have a lower BMI than those who are not. And those who are very young seem to have generally on average seem to have a lower BMI. This plot confirms the findings from our model: while there may be differences in BMI based on the different predictors, the differences themselves are very very small, and it's difficult to perceive them as the data is very noisy. In this case, since the model is still able to capture around 3.247% of the variability in such a noisy dataset, it may not be as bad as the diagnosis above suggests. Another thing to note from the plot is that for those who are inactive, age seems to be slightly negatively correlated with BMI for women and postively correlated for men. This could be something to explore in further analysis.

Technical Appendix

```
set.seed(1994)
NHANES <- NHANES %>% mutate(alcoholavg = AlcoholDay * AlcoholYear / 365) %>%
  filter(Age >= 18) %>%
  mutate(PhysActive = ifelse(PhysActive == "Yes", "Active", "Inactive")) %>%
  filter(alcoholavg < 23)
rows <- sample(1:nrow(NHANES), 4*(4912)/5)
train <- NHANES[rows,]
test <- NHANES[-rows,]
```