Untitled

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.

```
set.seed(1994)
NHANES <- NHANES %>% mutate(alcoholavg = AlcoholDay * AlcoholYear / 365) %>%
  mutate(alcoholavg = ifelse(is.na(alcoholavg), 0, alcoholavg)) %>%
  mutate(PhysActive = ifelse(PhysActive == "Yes", "Active", "Inactive"))
rows <- sample(1:nrow(NHANES), 8000)
train <- NHANES[rows,] %>% filter(Age >= 18)
test <- NHANES[-rows,] %>% filter(Age >= 18)
```

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 NHANES dataset is survey data gathered by the US National Center for Health Statistics (NCHS) beginning in the early 1960s. The survey is focused on health and nutrition. 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.

The following is a list of variables that we are concerned with:

BMI - Body mass index calculated by $\frac{weight}{height^2}$ in $\frac{kg}{m^2}$

PhysActive - A dichotomous variable. Yes or no depending on the answer to the question, does the participant do moderate or vigorous-intensity sports?

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. Anyone with NA values for AlcoholDay and AlcoholYear were coded to have 0 for their alcoholavg.

Age - the age in years of the study participant at the time of the screening

Gender - gender of participant coded as male or female

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.

Included in the linear model is an interaction term between alcoholavg and age. Originally, when I fit the model without the interaction term, the coefficient for alcoholavg was negative. It didn't make sense that as people drank more that there body mass index would go down. Since alcohol consumption generally occurs after a certain range, there is likely a difference between how alcohol effects your body at different ages.

To fit the model a random sample of 8000 was taken from the NHANES dataset.

```
mod <- lm(BMI ~ PhysActive + Age + Gender + Poverty + alcoholavg + Age*alcoholavg, data=train)
summary(mod)
```

```
##
## Call:
## lm(formula = BMI ~ PhysActive + Age + Gender + Poverty + alcoholavg +
##     Age * alcoholavg, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.447  -4.576  -1.164   3.480  51.219
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        27.908365   0.334366  83.466  < 2e-16 ***
## PhysActiveInactive  1.734300   0.190551   9.102  < 2e-16 ***
## Age                 0.017244   0.005880   2.933  0.00337 **
## Gendermale          0.306546   0.184796   1.659  0.09721 .
## Poverty            -0.224106   0.057007  -3.931 8.56e-05 ***
## alcoholavg          0.041012   0.257354   0.159  0.87339
## Age:alcoholavg     -0.009737   0.005527  -1.762  0.07817 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.701 on 5463 degrees of freedom
##   (514 observations deleted due to missingness)
## Multiple R-squared:  0.02964,    Adjusted R-squared:  0.02858
## F-statistic: 27.81 on 6 and 5463 DF,  p-value: < 2.2e-16
```

The coefficients of the model for PhysActiveYes, Age, Gendermale, Poverty, alcoholavg, and age:alcoholavg are -1.705, .063, .156, -.258, .978, -.027 respectively. This means that those who are physically active have a BMI that is lower by 1.705 on average than those who are not physically active. For a one year age increase, after accounting for other variables, the BMI would be expected to increase by .063. On average, males have a .156 high BMI than females. For a one unit increase in poverty ratio, after accounting for other variables, we would expect BMI to decrease by .258. As you drink more alcohol we would expect your BMI to increase, but the older you are, the effect that alcohol has on your BMI decreases.

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

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

```
## [1] 6.292538
```

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

```
## [1] 6.696863
```

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

```
## [1] 26.66014
```

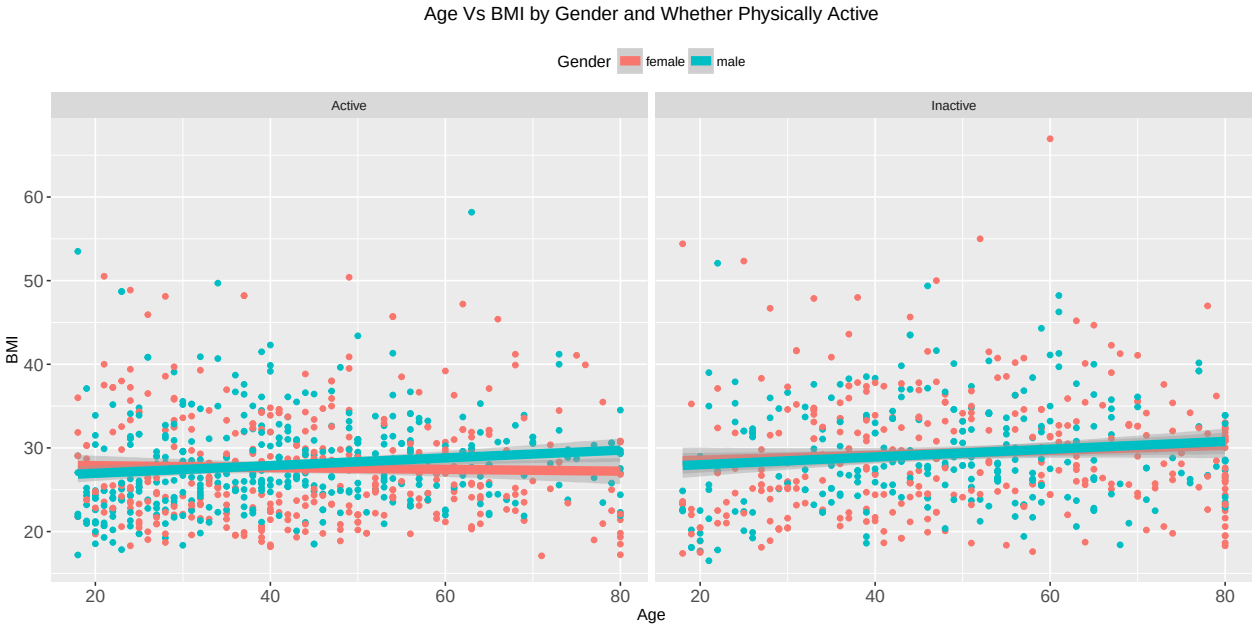
```r
median(NHANES$BMI, na.rm = TRUE) # median of BMI values
```

```
## [1] 25.98
```

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

The root mean squared error is 6.69 for the test set and 6.59 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. 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 is 0.0535, which shows that only slightly more than 5% of the variability in BMI is captured by the model. 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.

```r
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 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. 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 5.3% of the variability in such a noisy dataset, it may not be as bad as the diagnosis above suggests.