Revised Homework 6

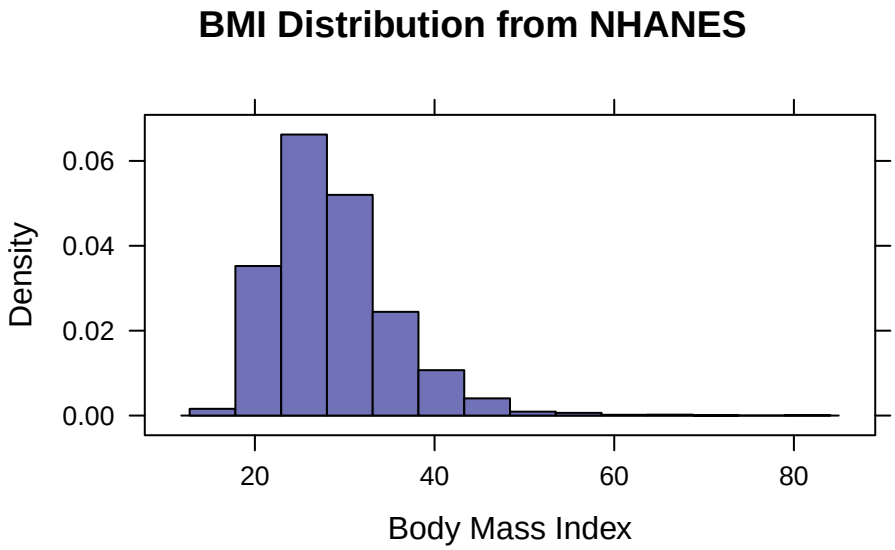
Orange Group - Tam Tran The, Coco Kusiak, Connor Haley

December 5, 2016

Background

The NHANES data set includes information on the National Health and Nutrition Examination Survey from 1999 to 2004. The set includes information on the race of the participant, their weight, if they use hard drugs, and if they have diabetes, as well as the variables used in the following model.

```
histogram(~BMI, data=NHANES, xlab = "Body Mass Index", ylab = "Density",
          main= "BMI Distribution from NHANES")
```



BMI

BMI has a unimodal distribution which is right skewed.

Our Predictors

- **PhysActiveDays:** physically active days per week
- **AlcoholDay:** days of consumption per year
- **Age:** in years
- **Gender:** female or male
- **Poverty Status:** ratio of income to poverty line, low values indicate lower wealth, capped at 5

```
pa <- favstats(~PhysActiveDays, data= NHANES)[c("min", "median", "mean", "max", "n")]
ad <- favstats(~AlcoholDay, data= NHANES)[c("min", "median", "mean", "max", "n")]
age <- favstats(~Age, data= NHANES)[c("min", "median", "mean", "max", "n")]
pov <- favstats(~Poverty, data= NHANES)[c("min", "median", "mean", "max", "n")]
```

```
preds <- rbind(pa, ad, age, pov)
rownames(preds) <- c("PhysActiveDays", "AlcoholDay", "Age", "Poverty")
xtable(preds)
```

% latex table generated in R 3.3.1 by xtable 1.8-2 package % Mon Dec 5 15:58:23 2016

	min	median	mean	max	n
PhysActiveDays	0.00	0.00	1.81	7.00	7481
AlcoholDay	0.00	1.00	1.91	82.00	7481
Age	18.00	45.00	46.23	80.00	7481
Poverty	0.00	2.88	2.92	5.00	6910

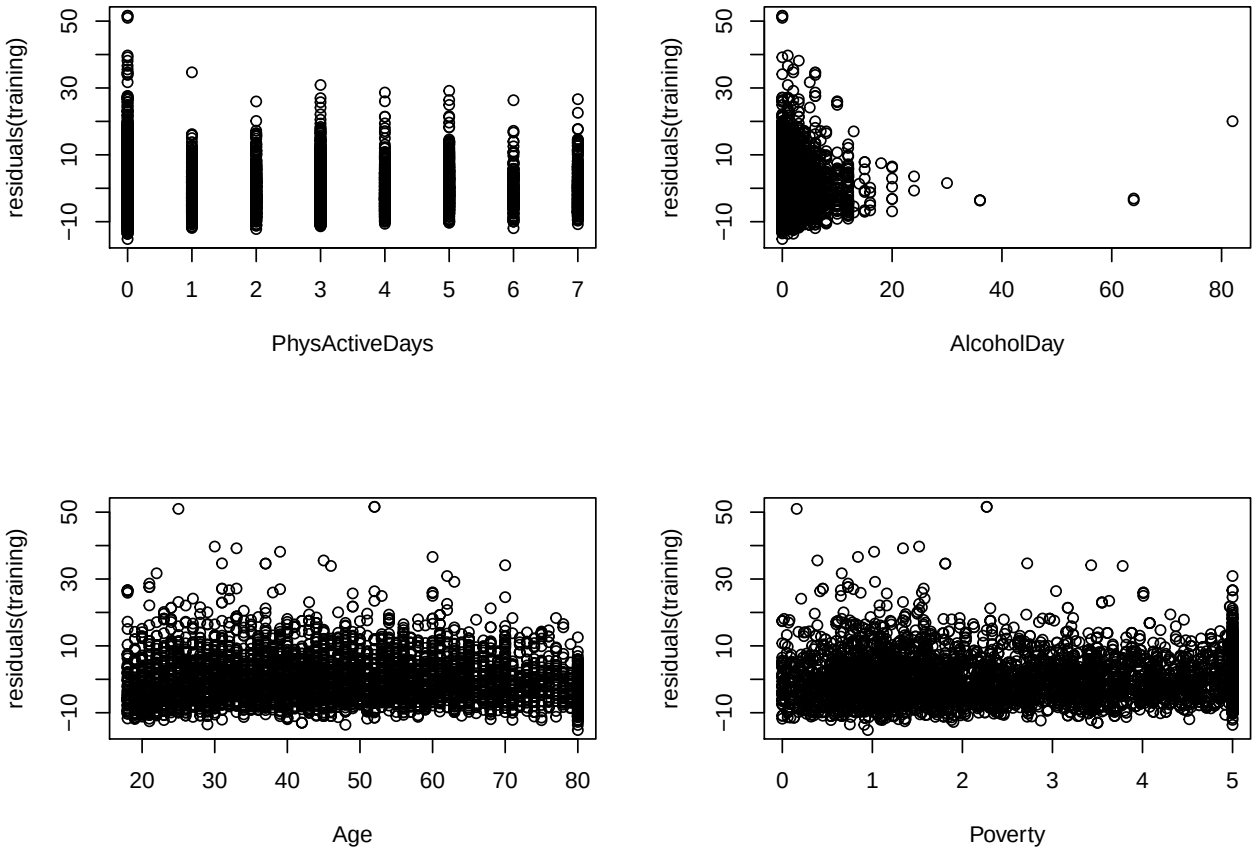
Our initial exploration of these variables shows that AlcoholDay has a strong right skew, with a max of 82 drinks per day. In addition, we are missing 571 values for Poverty, and all missing values for PhysActiveDays and AlcoholDay have been recoded to be 0.

There is an approximately even split between male and female participants in this study, with 3795 females and 3686 males.

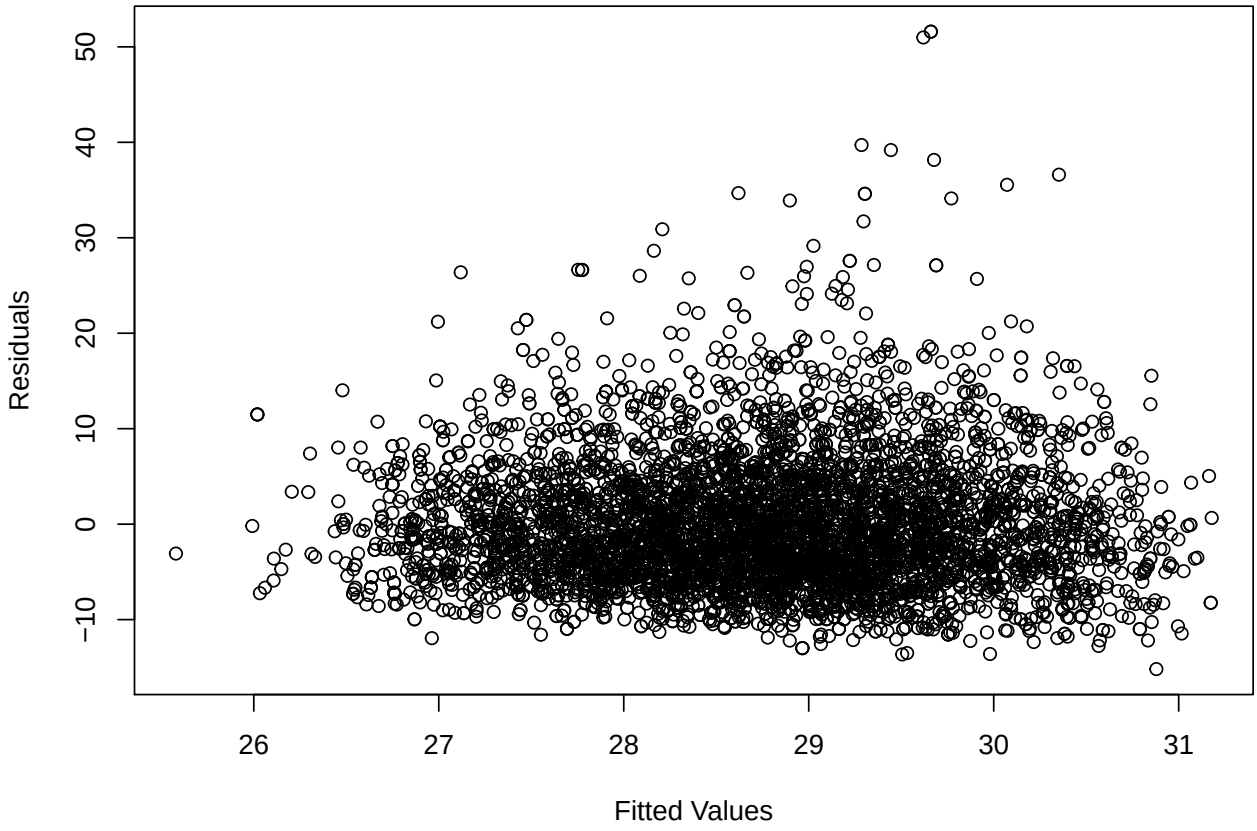
Analysis

The Assumptions

Linearity and Equal Variance:



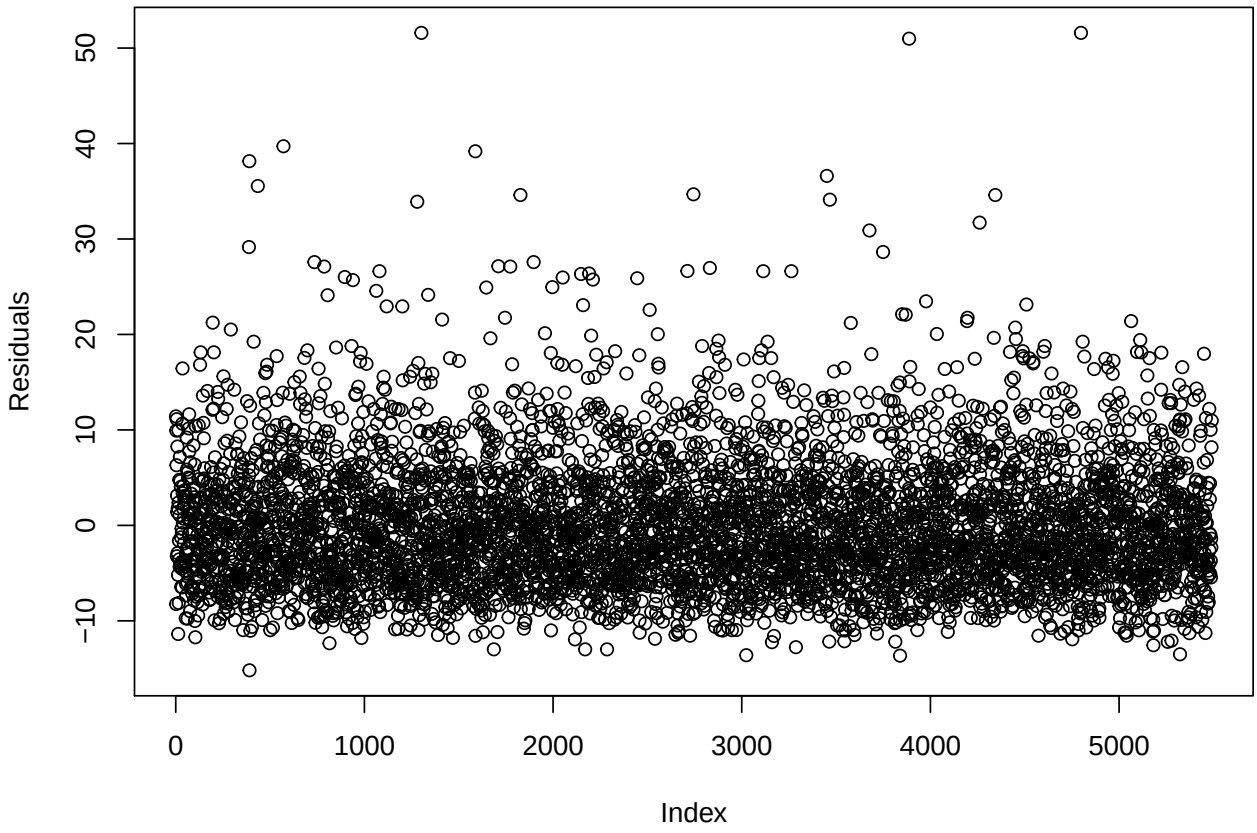
Based on the plots above, we do not see any patterns that indicate non-linear patterns in the data. Therefore, the linearity assumption is met.



Using the plot above, we see approximately equal variance of the residuals across all fitted values. Therefore, the equal variance assumption is met.

```
plot(residuals(training), ylab = "Residuals")
```

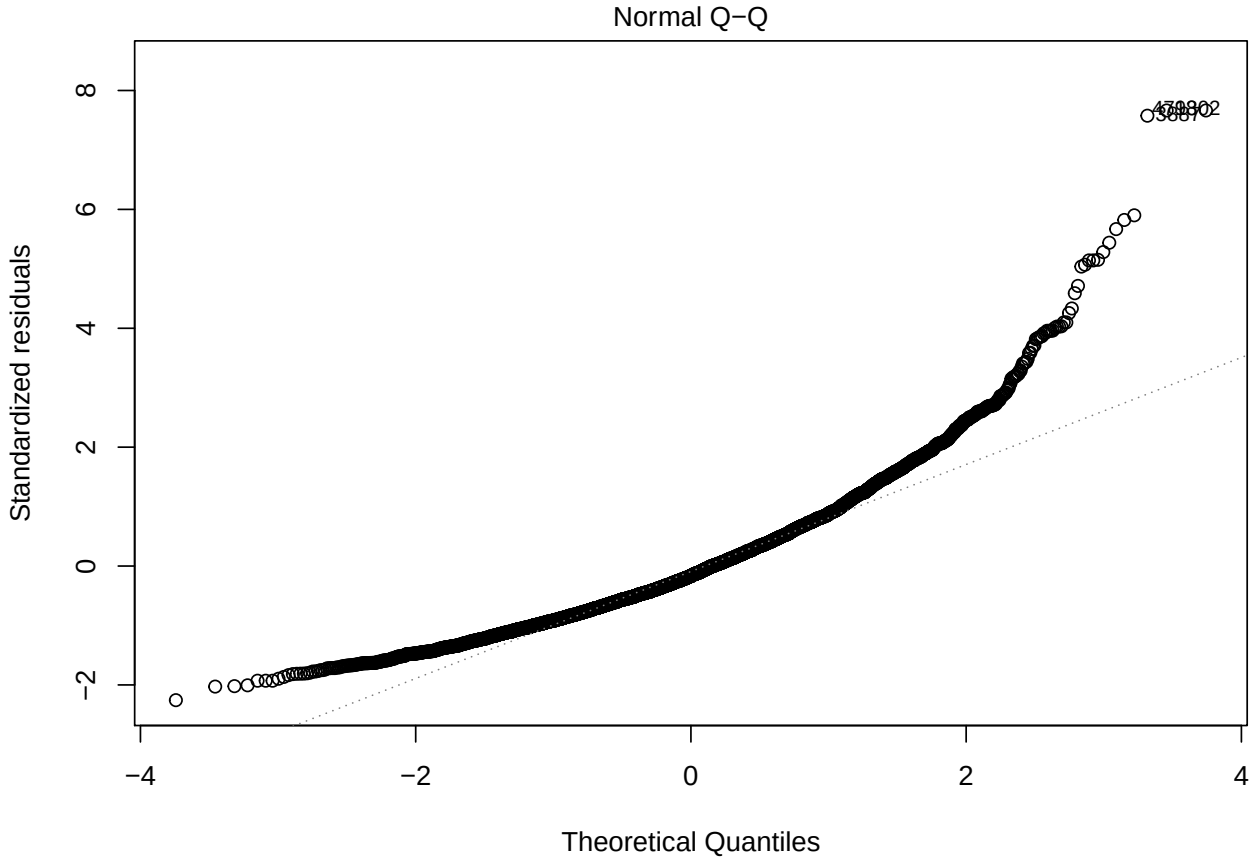
Independence:



We do not see any structure in the plot above, suggesting that the residuals are independent.

```
plot(training, which=2, sub = "")
```

Normality:



From this plot, we are also worried about the normality assumption because the distribution of our residuals is not consistent with a normal distribution, since our plotted residuals do not follow the line representing a normal distribution.

Although we are not completely satisfied that we have met the conditions for linear regression, we will proceed with the analysis below.

```
xtable(summary(training))
```

Modeling % latex table generated in R 3.3.1 by xtable 1.8-2 package % Mon Dec 5 15:58:28 2016

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	28.9759	0.3246	89.27	0.0000
PhysActiveDays	-0.2648	0.0404	-6.55	0.0000
AlcoholDay	-0.0141	0.0307	-0.46	0.6453
Age	0.0279	0.0054	5.13	0.0000
Gendermale	-0.0087	0.1857	-0.05	0.9627
Poverty	-0.3379	0.0553	-6.11	0.0000

This model was created using 6000 of the total 7481 observations, this is our `train` set. From this model, we can conclude that at an α level of 0.05, the variables PhysActive, Poverty, and Age are

significant in predicting a participant's Body Mass Index. For a one day per week increase in physical activity, there is expected to be a 0.265 unit decrease in BMI. A one year increase in a participant's age predicts a 0.028 unit increase in BMI. A one unit increase in a participant's poverty level predicts a 0.338 unit decrease in BMI. Intuitively, these results make sense. A poorer person may have less body mass, as well as a person who is physically active. An older person may have more body mass.

We next tested this linear model on the `test` data which includes the remaining 1481 observations from the NHANES data set. The RMSEs are reported below.

```
rmseTrain <- sqrt(mean((training$residuals)^2))
simsTest <- predict(training, test)
realsTest <- test$BMI
rmseTest <- sqrt(mean((simsTest - realsTest)^2, na.rm=TRUE))
FrameH <- data.frame(Set = c("Training", "Test"), RMSE = c(rmseTrain, rmseTest),
                     Variance = c(var(train$BMI, na.rm=TRUE), var(test$BMI, na.rm=TRUE)))
xtable(FrameH)
```

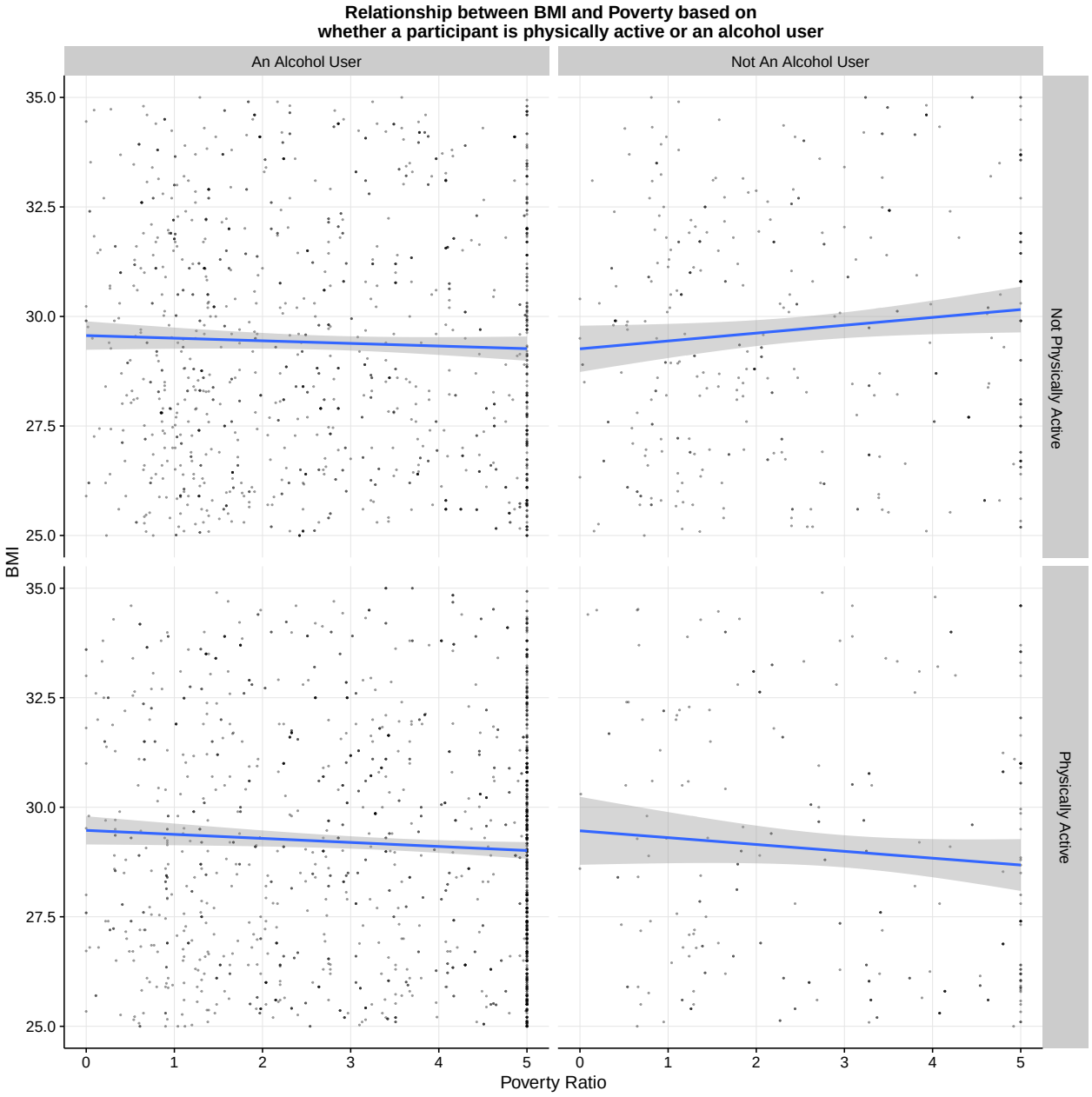
% latex table generated in R 3.3.1 by xtable 1.8-2 package % Mon Dec 5 15:58:28 2016

	Set	RMSE	Variance
1	Training	6.73	46.20
2	Test	6.38	40.83

Contrarily to what might be expected, the RMSE for the training set is greater than the RMSE for the test set. This however may be because the variance of the training set is greater than that of the test set. These two values are low and pretty similar between two data sets, which suggests that this model is a good fit.

Visualization

```
train1 <- NHANES[!is.na(NHANES$PhysActive),]
train1 <- train1[!is.na(train1$Alcohol12PlusYr),]
train1 <- train1 %>%
  mutate(PhysActive = ifelse(PhysActive == "No", "
                             Not Physically Active", "Physically Active")) %>%
  mutate(Alcohol12PlusYr = ifelse(Alcohol12PlusYr == "No",
                                  "Not An Alcohol User", "An Alcohol User"))
ggplot(data = train1, aes(x=Poverty, y=BMI)) +
  geom_point(size=0.3, alpha=0.4)  +
  facet_grid(PhysActive~Alcohol12PlusYr) +
  stat_smooth(method=lm) +
  theme(legend.position="right") + ylim(min = 25, max=35) +
  labs(list(x = "Poverty Ratio", title="Relationship between BMI and Poverty based on
       whether a participant is physically active or an alcohol user")) +
  background_grid(major="xy", minor="none")
```



As shown in this plot, we can conclude that after controlling for physical activity and alcohol use, BMI does not vary much based on poverty level.

This study found a lot more alcohol users than non-alcohol users shown by the density of points in left plots. The study however, found a similiar number of phsycially active and not physically active participants, as shown by the densities in the plots from top to bottom.

Technical Appendix

```
require(mosaic)   # Load additional packages here
require(xtable)
```

```
require(cowplot)
options(comment.xtable=FALSE)
# Some customization.  You can alter or delete as desired (if you know what you are doing).
trellis.par.set(theme=theme.mosaic()) # change default color scheme for lattice
knitr::opts_chunk$set(
  tidy=FALSE,     # display code as typed
  size="small")   # slightly smaller font for code
```

```
set.seed(1994)
library(NHANES)
#glimpse(NHANES)
data("NHANES")

NHANES <- filter(NHANES, Age >= 18) %>%
  mutate(PhysActiveDays = ifelse(is.na(PhysActiveDays), 0, PhysActiveDays),
         AlcoholDay = ifelse(is.na(AlcoholDay), 0, AlcoholDay))

rows <- sample(1:nrow(NHANES), 6000)
train <- NHANES[rows,]
dim(train)
test <- NHANES[-rows,]
dim(test)
```

```
train <- select(train, BMI, PhysActiveDays, AlcoholDay, Age, Gender, Poverty)
train <- train[complete.cases(train),]
training <- lm(BMI ~ PhysActiveDays + AlcoholDay + Age + Gender + Poverty, data = train)

par(mfrow=c(2,2))
plot(train$PhysActiveDays, residuals(training), xlab = "PhysActiveDays")
plot(train$AlcoholDay, residuals(training), xlab = "AlcoholDay")
plot(train$Age, residuals(training), xlab = "Age")
plot(train$Poverty, residuals(training), xlab = "Poverty")
```

```
plot(fitted(training), residuals(training), xlab = "Fitted Values", ylab = "Residuals")
```

```
train1 <- NHANES[!is.na(NHANES$PhysActive),]
train1 <- train1[!is.na(train1$Alcohol12PlusYr),]
train1 <- train1 %>%
  mutate(PhysActive = ifelse(PhysActive == "No", "
                             Not Physically Active", "Physically Active")) %>%
  mutate(Alcohol12PlusYr = ifelse(Alcohol12PlusYr == "No",
                                  "Not An Alcohol User", "An Alcohol User"))
ggplot(data = train1, aes(x=Poverty, y=BMI)) +
  geom_point(size=0.3, alpha=0.4)  +
  facet_grid(PhysActive~Alcohol12PlusYr) +
  stat_smooth(method=lm) +
  theme(legend.position="right") + ylim(min = 25, max=35) +
  labs(list(x = "Poverty Ratio", title="Relationship between BMI and Poverty based on
       whether a participant is physically active or an alcohol user")) +
  background_grid(major="xy", minor="none")
```