Revised Homework 6

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

```r
set.seed(1994)
library(NHANES)
#glimpse(NHANES)
rows <- sample(1:nrow(NHANES), 8000)
train <- NHANES[rows,]
dim(train)
```

```
## [1] 8000   76
```

```r
test <- NHANES[-rows,]
dim(test)
```

```
## [1] 2000   76
```

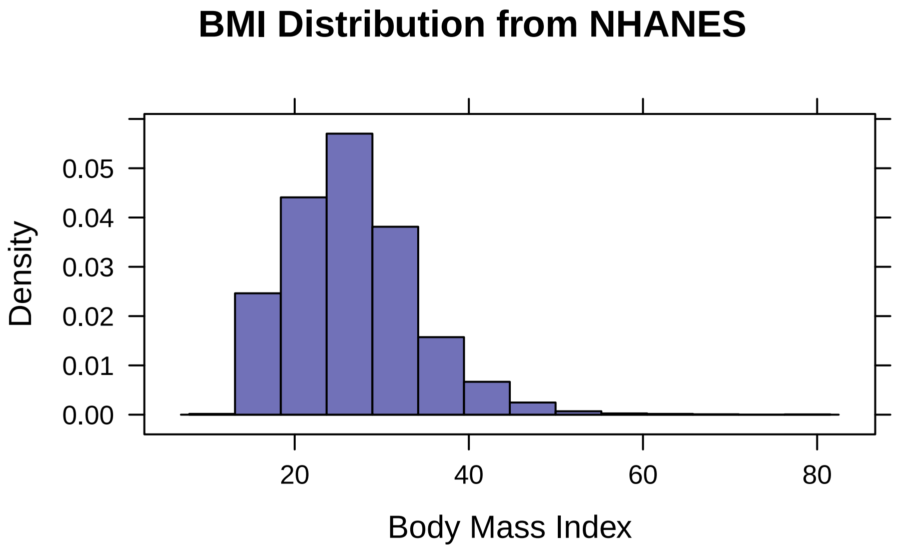
Your report should provide background on these data, describe the analytic sample, fit and interpret the model, and undertake model assessment. You should include one figure that summarizes key findings.

SOLUTION:

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.

```r
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

```r
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)
```

	min	median	mean	max	n
PhysActiveDays	1.00	3.00	3.74	7.00	4663
AlcoholDay	1.00	2.00	2.91	82.00	4914
Age	0.00	36.00	36.74	80.00	10000
Poverty	0.00	2.70	2.80	5.00	9274

Our initial exploration of these variables shows that AlcoholDay has a strong right skew, with a max of 82 drinks per day. In addition, for both PhysActiveDays and AlcoholDay, were are missing apporixmately half of our initial observations with n equal to 4663 and 4914, respectively.

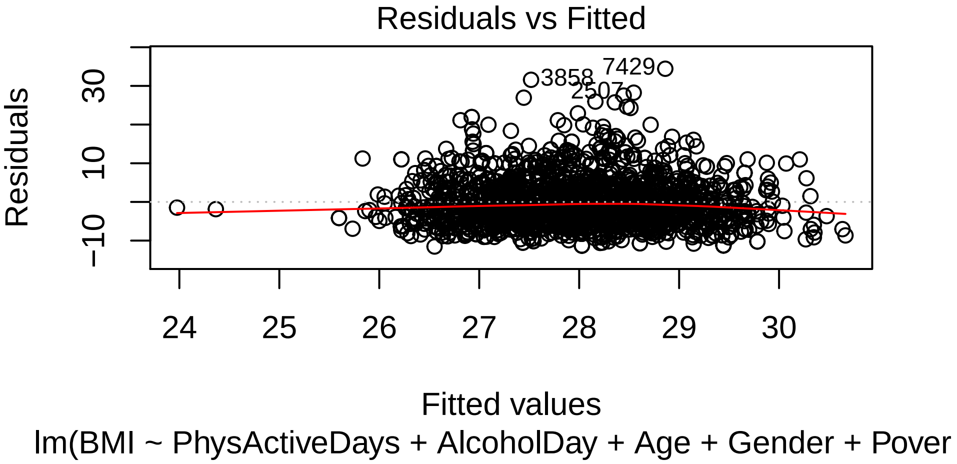
There is an approximately even split between male and female participants in this study, with 5020 females and 4980 males.

Analysis

The Assumptions

Linearity and Equal Variance:

```r
training <- lm(BMI ~ PhysActiveDays + AlcoholDay + Age + Gender + Poverty, data = train)
plot(training, which=1)
```



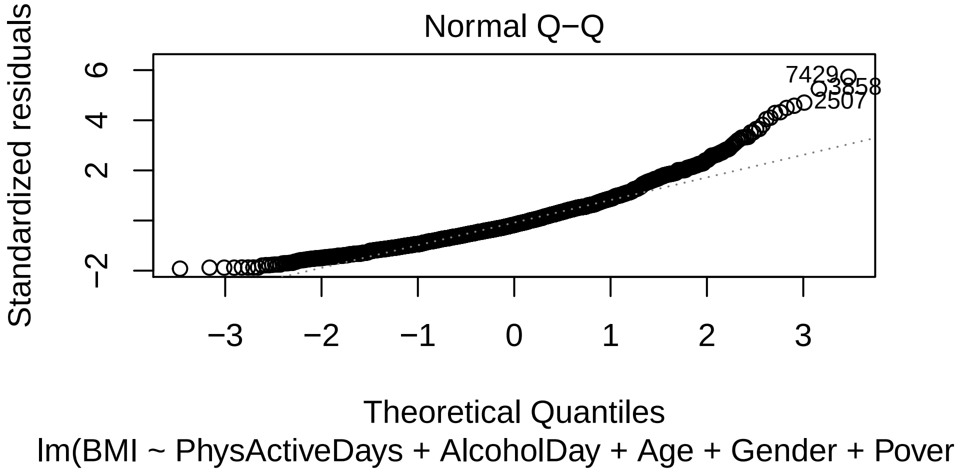
Based on this plot, we have no evidence that the relationship is not linear. There is no clear pattern in the residuals. Also from this plot, we worry about the equal variance assumption due to there not being an even distribution of points through the fitted values.

Independence:

We assume that the BMIs of participants in the study are not dependent on eachother.

Normality:

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



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))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	28.3764	0.5762	49.25	0.0000
PhysActiveDays	-0.1943	0.0762	-2.55	0.0108
AlcoholDay	-0.0482	0.0434	-1.11	0.2664
Age	0.0252	0.0092	2.75	0.0060
Gendermale	0.9679	0.2810	3.44	0.0006
Poverty	-0.3579	0.0876	-4.09	0.0000

Modeling

This model was created using 8,000 of the total 10,000 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 0.194 unit decrease in BMI. A one year increase in a participant's age predicts a 0.025 unit increase in BMI. A one unit increase in a participant's poverty level predicts a 0.358 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.

While our model finds statistically significant predictors, it is interesting to note that the r^2 is only 0.020. This means our model only accounts for 2% of the variation in BMI.

We next tested this linear model on the `test` data which includes the remaining 2,000 observations from the original NHANES data set. The RMSEs are reported below.

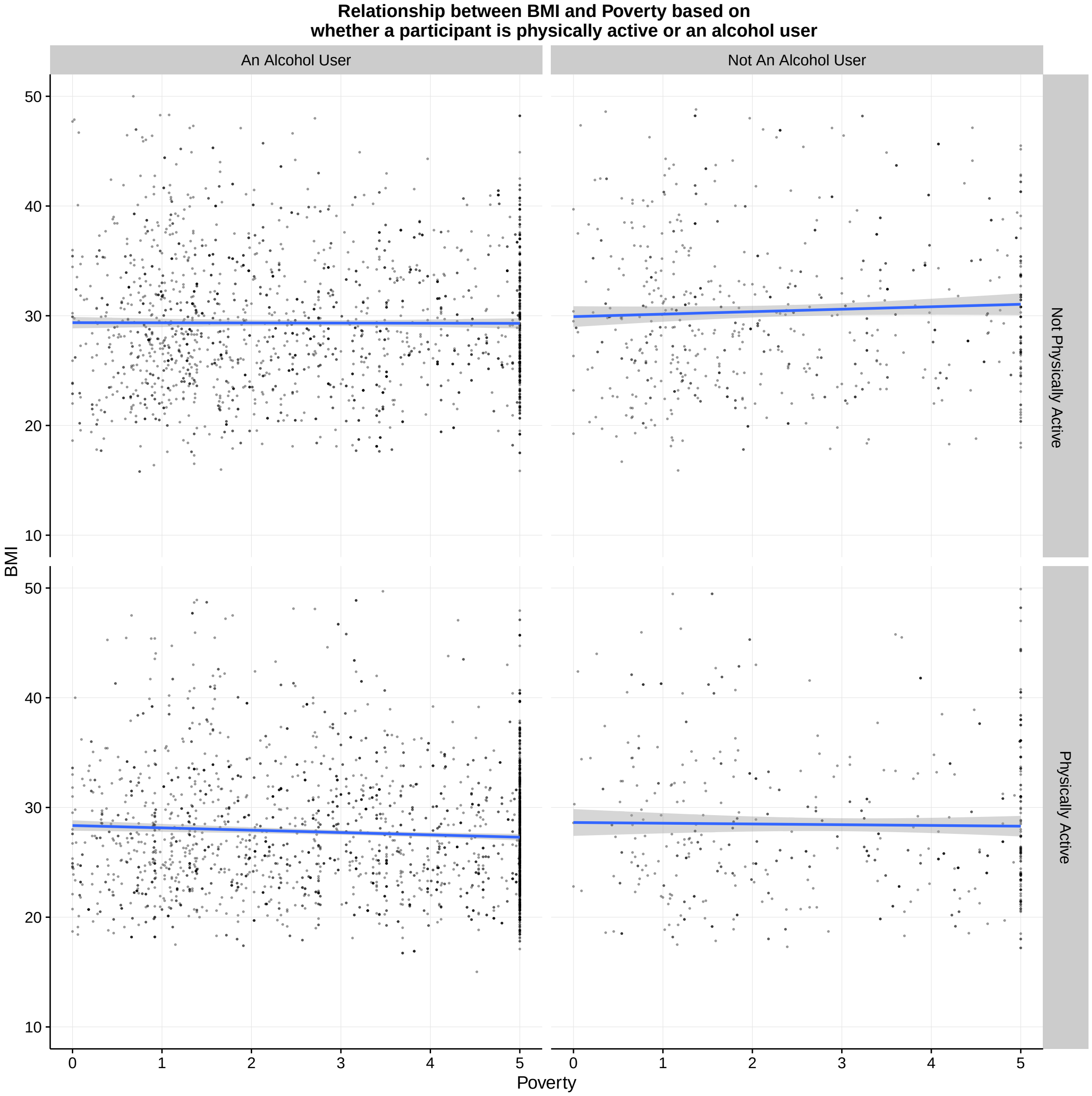
```
rmseTrain <- sqrt(mean((training$residuals)^2))
simsTest <- predict(training, test)
realsTest <- test$BMI
rmseTest <- sqrt(mean((simsTest - realsTest)^2, na.rm=TRUE))
FrameH <- data.frame(Set = c("Training", "Test"), RMSE = c(rmseTrain, rmseTest),
                     Variance = c(var(train$BMI,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

	Set	RMSE	Variance
1	Training	6.01	55.22
2	Test	6.02	51.20

Contrarily to what might be expected, the RMSE for the test set is greater than the RMSE for the training 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 = 10, max=50) +
  labs(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.