HW01 - Statistical Modeling

Kyle Barisone

```r
knitr::opts_chunk$set(warning=FALSE, message=FALSE)
# load appropriate libraries and import data here

# For this assignment, to get you started, I have included code to directly import
# data sets from my website.
# After you create your own data management files, this approach should not be used.

FEV <- read.delim("https://norcalbiostat.netlify.com/data/Lung_081217.txt",
                  sep="\t", header=TRUE)
hiv <- read.delim("https://norcalbiostat.netlify.com/data/PARHIV_081217.txt",
                  sep="\t", stringsAsFactors = FALSE, header=TRUE)
depress <- read.delim("https://norcalbiostat.netlify.com/data/depress_081217.txt",
                  sep="\t", header=TRUE)

library(ggplot2)
library(dplyr)
library(leaps)
library(pander)
```

Part I: Statistical Modeling

1. Fit a linear regression model

```r
mv_model <- lm(cesd ~ income + age, data=depress)
summary(mv_model)
```

```
##
## Call:
## lm(formula = cesd ~ income + age, data = depress)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.860  -5.891  -2.438   3.620  36.456
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.59442    1.61773   9.640  < 2e-16 ***
## income      -0.11353    0.03334  -3.405 0.000755 ***
## age         -0.09848    0.02819  -3.494 0.000551 ***
## ---
```

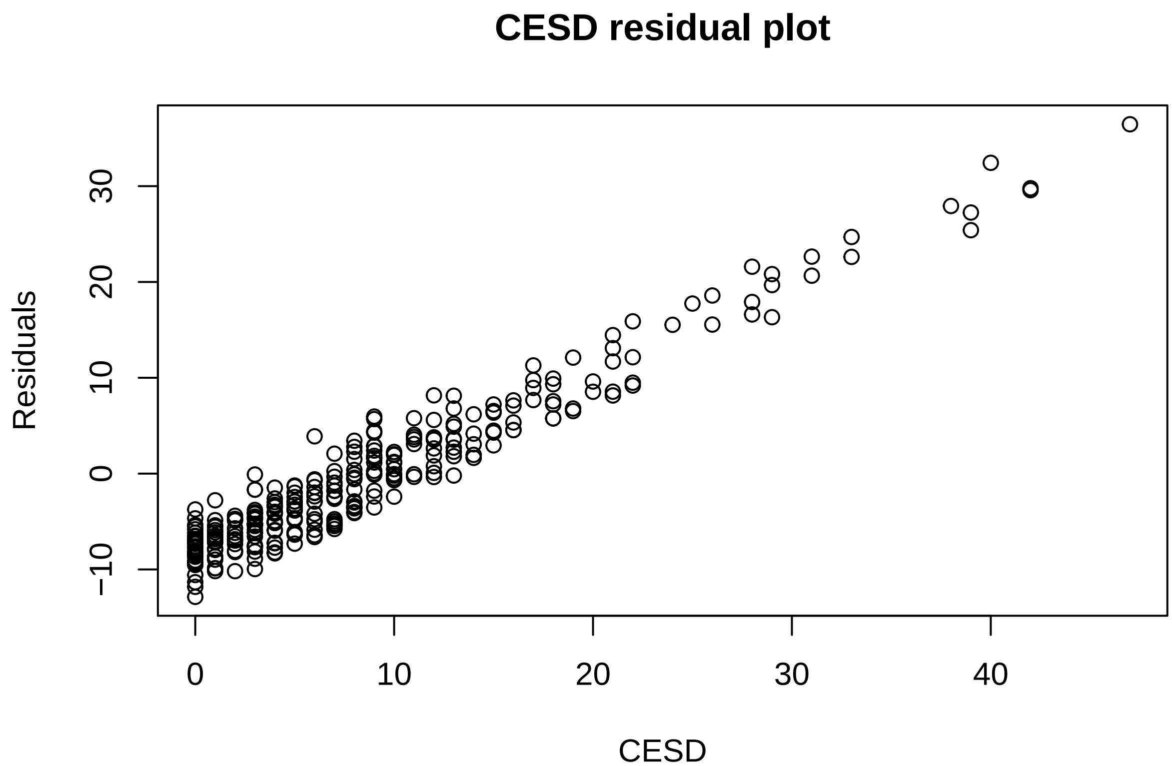
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.565 on 291 degrees of freedom
## Multiple R-squared:  0.06423,    Adjusted R-squared:  0.0578
## F-statistic: 9.986 on 2 and 291 DF,  p-value: 6.387e-05
```

a. Analyze the residuals

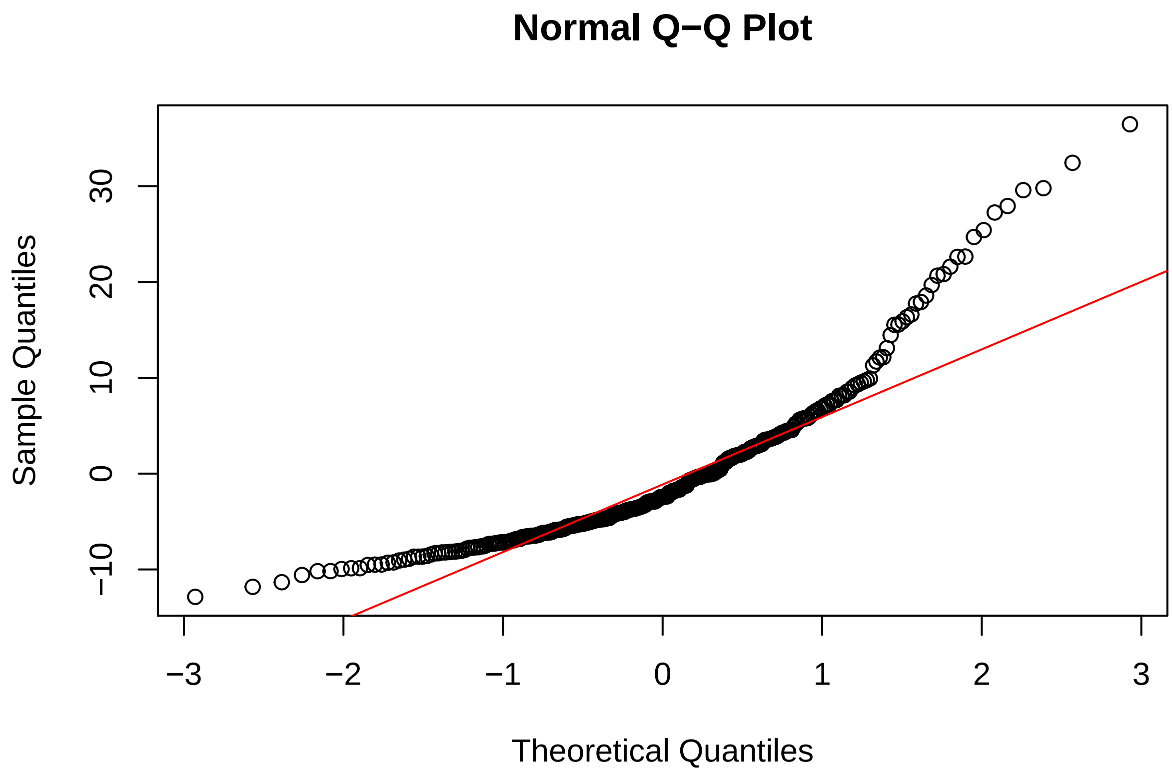
```
mv_model <- lm(cesd ~ income + age, data=depress)
summary(mv_model)
```

```
##
## Call:
## lm(formula = cesd ~ income + age, data = depress)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.860  -5.891  -2.438   3.620  36.456
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.59442    1.61773   9.640  < 2e-16 ***
## income      -0.11353    0.03334  -3.405 0.000755 ***
## age         -0.09848    0.02819  -3.494 0.000551 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.565 on 291 degrees of freedom
## Multiple R-squared:  0.06423,    Adjusted R-squared:  0.0578
## F-statistic: 9.986 on 2 and 291 DF,  p-value: 6.387e-05
```

```
model.resid <- resid(mv_model)
plot(depress$cesd, model.resid, ylab="Residuals", xlab="CESD",
         main="CESD residual plot")
```




```r
qqnorm(mv_model$residuals)
qqline(mv_model$residuals, col='red')
```



The qqplot indicates that the data is fairly normal since slight deviations in the tails are expected. However since it seems that it curves slightly more as the x-axis increases it could indicate data that is slightly right skewed. ### b. Interpret each coefficient

```
confint(mv_model)
```

```
##                   2.5 %      97.5 %
## (Intercept) 12.4104917 18.77834390
## income      -0.1791574 -0.04790915
## age         -0.1539645 -0.04300222
```

B_0: someone who makes no money per year, and is 0 years old is expected to have a cesd score of 15.59. 95% CI: (12.36, 18.71) with p-value < 0.001.

B_1: After controlling for age, someone who earns $1000 per year greater is expected have 0.11 lower cesd. 95% CI: (-0.18, -0.05) p-value < 0.0001.

B_2: After controlling for income, for every year older someone gets, they are expected to have a cesd score that is 0.10 lower. 95% CI: (-0.15, -0.04) p-value < 0.0001. ## 2. Test gender as a moderator using a) using a stratified model

```
depress$SEX1 <- ifelse(depress$sex=="0", "Male", "Female")

mv_model <- depress %>% select(cesd, age, income, SEX1)
male <- mv_model %>% select(cesd, age, income, SEX1) %>% filter(SEX1 == "Male")
female <- mv_model %>% select(cesd, age, income, SEX1) %>% filter(SEX1 == "Female")
model_male <- lm(cesd ~ age + income, data = male)
```

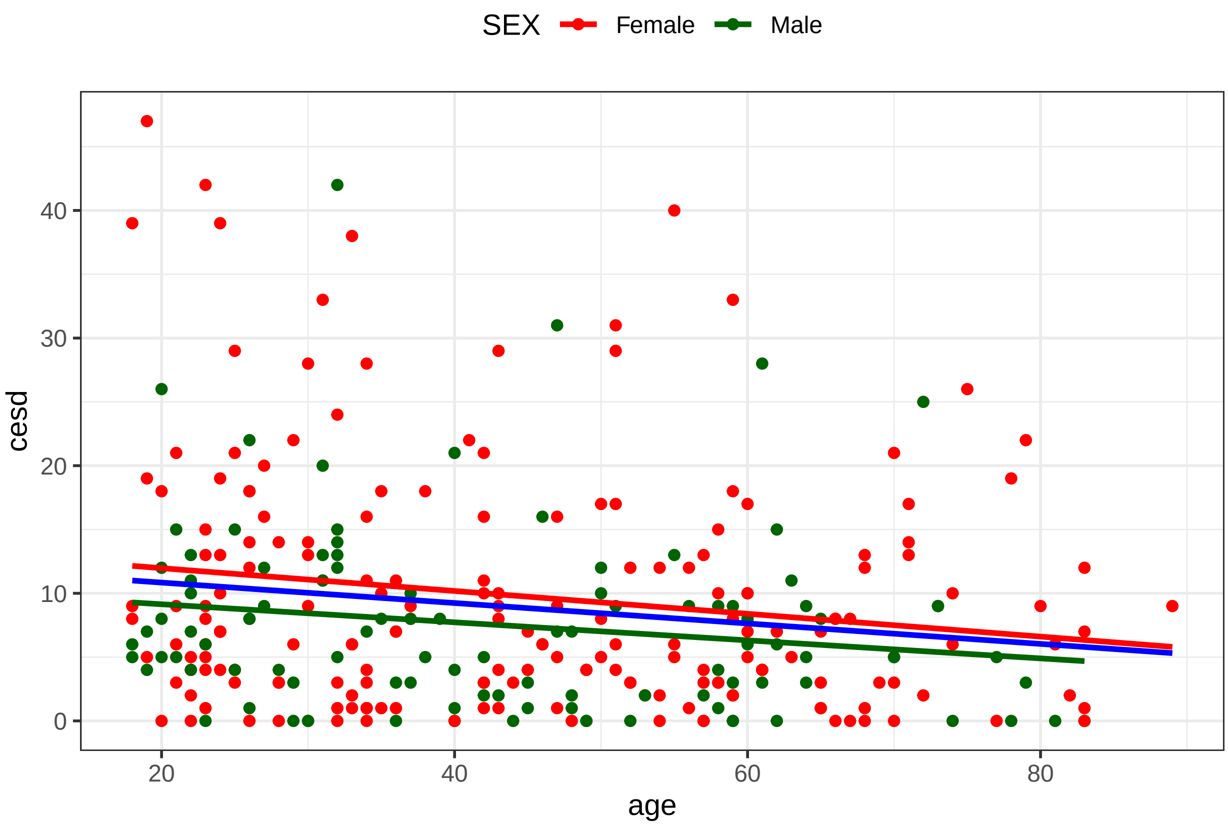
```r
model_female <- lm(cesd ~ age + income, data = female)

summary(model_male)
```

```
##
## Call:
## lm(formula = cesd ~ age + income, data = male)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.323 -4.553 -1.805  2.972 31.054
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.98752    2.19210   6.381  4.5e-09 ***
## age         -0.08810    0.03865  -2.279  0.02461 *
## income      -0.11105    0.04136  -2.685  0.00839 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.965 on 108 degrees of freedom
## Multiple R-squared:  0.0898, Adjusted R-squared:  0.07294
## F-statistic: 5.328 on 2 and 108 DF,  p-value: 0.006214
```

```r
summary(model_female)
```

```
##
## Call:
## lm(formula = cesd ~ age + income, data = female)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.440  -7.004  -2.532   4.528  35.477
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.22231    2.22641   7.286 9.65e-12 ***
## age         -0.10445    0.03843  -2.718  0.00721 **
## income      -0.09696    0.04978  -1.948  0.05300 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.391 on 180 degrees of freedom
## Multiple R-squared:  0.04979,    Adjusted R-squared:  0.03923
## F-statistic: 4.716 on 2 and 180 DF,  p-value: 0.01008
```

```r
ggplot(depress, aes(x=age, y=cesd, col=as.factor(SEX1))) +
  geom_point() + theme_bw() + theme(legend.position="top") +
  scale_color_manual(name="SEX", values=c("red", "darkgreen")) +
  geom_smooth(se=FALSE, method="lm") +
  geom_smooth(aes(x=age, y=cesd), col="blue", se=FALSE, method='lm')
```



b) using an interaction model.

```
summary(lm(cesd ~ age + sex + income + age*sex, data=depress))
```

```
##
## Call:
## lm(formula = cesd ~ age + sex + income + age * sex, data = depress)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.546  -5.814  -2.136   3.677  35.512
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.74519    2.42275   5.673  3.4e-08 ***
## age         -0.08687    0.04706  -1.846  0.06595 .
## sex          2.63601    2.75547   0.957  0.33955
## income      -0.10321    0.03380  -3.053  0.00247 **
## age:sex     -0.01855    0.05797  -0.320  0.74924
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.548 on 289 degrees of freedom
## Multiple R-squared:  0.07426,    Adjusted R-squared:  0.06145
```

```
## F-statistic: 5.796 on 4 and 289 DF,  p-value: 0.0001686
```

3. Which of the two models in question 2 assumes that the affect of income on depression is constant (does not change) between males and females?

The interaction model assumes the affects are constant while the stratified model fits them separately based on gender.

4. Determine whether the regression plane can be improved by also including `weight`. Use two measures of model fit to justify your answer to this question

```r
summary(lm(FFEV1 ~ FAGE + FHEIGHT, data=FEV))
```

```
##
## Call:
## lm(formula = FFEV1 ~ FAGE + FHEIGHT, data = FEV)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.34708 -0.34142  0.00917  0.37174  1.41853
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.760747   1.137746  -2.427   0.0165 *
## FAGE        -0.026639   0.006369  -4.183 4.93e-05 ***
## FHEIGHT      0.114397   0.015789   7.245 2.25e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5348 on 147 degrees of freedom
## Multiple R-squared:  0.3337, Adjusted R-squared:  0.3247
## F-statistic: 36.81 on 2 and 147 DF,  p-value: 1.094e-13
```

```r
summary(lm(FFEV1 ~ FAGE + FHEIGHT + FWEIGHT, data=FEV))
```

```
##
## Call:
## lm(formula = FFEV1 ~ FAGE + FHEIGHT + FWEIGHT, data = FEV)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3927 -0.3316  0.0223  0.3871  1.6223
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.383883   1.155410  -2.929  0.00395 **
## FAGE        -0.026520   0.006282  -4.222 4.24e-05 ***
## FHEIGHT      0.135901   0.018242   7.450 7.54e-12 ***
## FWEIGHT     -0.004783   0.002114  -2.263  0.02510 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5274 on 146 degrees of freedom
## Multiple R-squared:  0.3563, Adjusted R-squared:  0.3431
## F-statistic: 26.94 on 3 and 146 DF,  p-value: 6.331e-14
```

Weight improves the model but the change is not very significant. Multiple R-squared changes from .334 to .356 and adjusted R-squared changes from .325 to .343 when the variable weight is added, Standard error also decreases slightly (by less than .01)

5. Does weight *confound* the relationship between age or height and FEV1?

```r
summary(lm(FFEV1 ~ FHEIGHT, data=FEV))
```

```
##
## Call:
## lm(formula = FFEV1 ~ FHEIGHT, data = FEV)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.56688 -0.35290  0.04365  0.34149  1.42555
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.08670    1.15198  -3.548 0.000521 ***
## FHEIGHT      0.11811    0.01662   7.106 4.68e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5638 on 148 degrees of freedom
## Multiple R-squared:  0.2544, Adjusted R-squared:  0.2494
## F-statistic:  50.5 on 1 and 148 DF,  p-value: 4.677e-11
```

```r
summary(lm(FFEV1 ~ FHEIGHT + FWEIGHT, data=FEV))
```

```
##
## Call:
## lm(formula = FFEV1 ~ FHEIGHT + FWEIGHT, data = FEV)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.42790 -0.38764  0.04791  0.29479  1.62625
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.713564   1.173551  -4.016 9.39e-05 ***
## FHEIGHT      0.139929   0.019232   7.276 1.91e-11 ***
## FWEIGHT     -0.004858   0.002231  -2.177    0.031 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.5568 on 147 degrees of freedom
## Multiple R-squared:  0.2777, Adjusted R-squared:  0.2679
## F-statistic: 28.26 on 2 and 147 DF,  p-value: 4.123e-11
```

```r
summary(lm(FFEV1 ~ FAGE, data=FEV))
```

```
##
## Call:
## lm(formula = FFEV1 ~ FAGE, data = FEV)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73332 -0.46620 -0.01332  0.42572  1.89899
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.266374   0.300590  17.520  < 2e-16 ***
## FAGE        -0.029230   0.007382  -3.959 0.000116 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6209 on 148 degrees of freedom
## Multiple R-squared:  0.09578,    Adjusted R-squared:  0.08967
## F-statistic: 15.68 on 1 and 148 DF,  p-value: 0.0001163
```

```r
summary(lm(FFEV1 ~ FAGE + FWEIGHT, data=FEV))
```

```
##
## Call:
## lm(formula = FFEV1 ~ FAGE + FWEIGHT, data = FEV)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.68723 -0.44163 -0.06107  0.43963  1.88433
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.633388   0.492329   9.411  < 2e-16 ***
## FAGE        -0.028967   0.007344  -3.944 0.000124 ***
## FWEIGHT      0.003418   0.002112   1.618 0.107763
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6175 on 147 degrees of freedom
## Multiple R-squared:  0.1116, Adjusted R-squared:  0.09952
## F-statistic: 9.233 on 2 and 147 DF,  p-value: 0.0001669
```

Since both models are significant $p<.0001$ before and after weight is added, it seems that weight is not a confounder of age or height when measuring FEV1. ## 6. Fit a model of income using age, sex, educational level and religion as predictors.

```r
new.dep <- depress %>% select(income, age, sex, educat, relig)
View(new.dep)

new.dep$relig <- factor(new.dep$relig, labels = c("Protestant", "Catholic", "Jewish", "No Religion"))
new.dep$sex <- factor(new.dep$sex, labels = c("Male", "Female"))
new.dep$educat <- factor(new.dep$educat, labels = c("less than highschool", "some highschool", "finishe

dep.model2 <- lm(income ~ age + sex + educat + relig, data = new.dep)

summary(dep.model2)
```

```
##
## Call:
## lm(formula = income ~ age + sex + educat + relig, data = new.dep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.043  -8.464  -2.338   7.824  46.443
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              13.13246    7.51487   1.748 0.081642 .
## age                      -0.09113    0.04654  -1.958 0.051226 .
## sexFemale                -4.41282    1.69690  -2.601 0.009803 **
## educatsome highschool    18.27960    7.31087   2.500 0.012980 *
## educatfinished highschool 11.98268    7.07084   1.695 0.091252 .
## educatsome college       29.13932    7.80152   3.735 0.000227 ***
## educatbachelors          29.67750    8.25078   3.597 0.000381 ***
## educatmasters            20.76515    7.25661   2.862 0.004534 **
## educatdoctorate           7.64372    7.13261   1.072 0.284796
## religCatholic            -2.92897    2.22518  -1.316 0.189155
## religJewish               3.32137    2.75808   1.204 0.229516
## religNo Religion         -0.90837    2.20940  -0.411 0.681286
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.56 on 280 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.2463, Adjusted R-squared:  0.2167
## F-statistic:  8.32 on 11 and 280 DF,  p-value: 1.222e-12
```

a. Use a general F test to determine whether religion has an effect on income.

```r
new.dep2 <- depress %>% select(income, age, sex, educat, relig)
new.dep2$relig <- factor(new.dep2$relig, labels = c("Protestant", "Catholic", "Jewish", "No Religion"))
new.dep2$sex <- factor(new.dep2$sex, labels = c("Male", "Female"))
new.dep2$educat <- factor(new.dep2$educat, labels = c("less than highschool", "some highschool", "finis
full_model <- lm(income ~ age + sex + educat +relig, data=new.dep2)
pander(summary(full_model))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	13.13	7.515	1.748	0.08164
age	-0.09113	0.04654	-1.958	0.05123
sexFemale	-4.413	1.697	-2.601	0.009803
educatsome highschool	18.28	7.311	2.5	0.01298
educatfinished highschool	11.98	7.071	1.695	0.09125
educatsome college	29.14	7.802	3.735	0.0002274
educatbachelors	29.68	8.251	3.597	0.0003805
educatmasters	20.77	7.257	2.862	0.004534
educatdoctorate	7.644	7.133	1.072	0.2848
religCatholic	-2.929	2.225	-1.316	0.1892
religJewish	3.321	2.758	1.204	0.2295
religNo Religion	-0.9084	2.209	-0.4111	0.6813

Table 2: Fitting linear model: income ~ age + sex + educat + relig

Observations	Residual Std. Error	R^2	Adjusted R^2
292	13.56	0.2463	0.2167

```
survey::regTermTest(full_model, "relig")
```

```
## Wald test for relig
##  in lm(formula = income ~ age + sex + educat + relig, data = new.dep2)
## F =  1.389428  on  3  and  280  df: p= 0.24623
```

The P - Value is .246 so we can conclude that religion is not a good predictor of income level for an individual.

b. State the reference categories for both religion and educational level.

The reference category for religion is protestant and the reference category for education level is less than highschool.

c. Interpret the coefficient for each level of educational level

```
confint(full_model)
```

```
##                                2.5 %        97.5 %
## (Intercept)               -1.6603635 27.9252820194
## age                       -0.1827384  0.0004884122
## sexFemale                 -7.7531142 -1.0725188333
## educatsome highschool      3.8883610 32.6708413542
## educatfinished highschool -1.9360797 25.9014397629
## educatsome college        13.7822495 44.4963965050
## educatbachelors           13.4360575 45.9189339916
## educatmasters              6.4807204 35.0495819187
## educatdoctorate           -6.3966187 21.6840645751
## religCatholic             -7.3091744  1.4512352488
```

```
## religJewish              -2.1078259  8.7505734537
## religNo Religion         -5.2575041  3.4407723588
```

B_3: After controlling for sex, age, and religion, individuals who have completed some high school are expected to have an income of 18.28 thousand dollars higher than people who did not complete any high school. 95% CI: (3.89, 32.67) (p-value = 0.012)

B_4: After controlling for sex, age, and religion, individuals who have finished high school are expected to have an income of 11.98 thousand dollars higher than people who did not complete any high school. 95% CI: (-1.93, 25.90) (p-value = .0913)

B_5: After controlling for sex, age, and religion, individuals who have completed some college are expected to have an income of 29.14 thousand dollars higher than people who did not complete any high school. 95% CI: (13.78, 44.50) (p-value = .0002)

B_6: After controlling for sex, age, and religion, individuals who have completed their bachelors are expected to have an income 29.68 thousand dollars higher than people who did not complete any high school. 95% CI: (13.44, 45.92) (p-value = .0004)

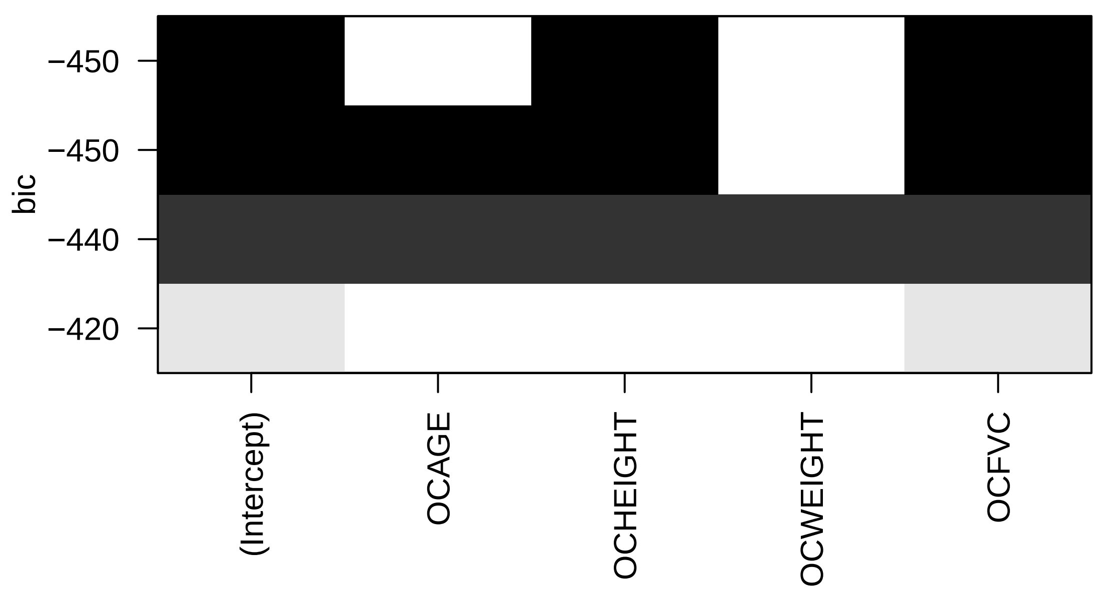
B_7: After controlling for sex, age, and religion, individuals who have completed their masters are expected to have an income 20.77 thousand dollars higher than people who did not complete any high school. 95% CI: (6.48, 35.05) (p-value = .0045)

B_8: After controlling for sex, age, and religion, individuals who have completed their doctorate are expected to have an income 7.644 thousand dollars higher than people who did not complete any high school. 95% CI: (-6.40, 21.68) (p-value = .2848)

Part II: Variable Selection

1. PMA6 9.9

```
OC.data <- FEV %>% select(OCFEV1, OCAGE, OCHEIGHT, OCWEIGHT, OCFVC)

subset_result <- regsubsets(OCFEV1~.,data=OC.data, nvmax = 5)
plot(subset_result, scale="bic")
```




```
summary(subset_result)
```

```
## Subset selection object
## Call: regsubsets.formula(OCFEV1 ~ ., data = OC.data, nvmax = 5)
## 4 Variables  (and intercept)
##          Forced in Forced out
## OCAGE        FALSE      FALSE
## OCHEIGHT     FALSE      FALSE
## OCWEIGHT     FALSE      FALSE
## OCFVC        FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: exhaustive
##          OCAGE OCHEIGHT OCWEIGHT OCFVC
## 1  ( 1 ) " "   " "      " "      "*"
```

```
## 2  ( 1 ) " "   "*"      " "      "*"
## 3  ( 1 ) "*"   "*"      " "      "*"
## 4  ( 1 ) "*"   "*"      "*"      "*"
```

```r
nullmodel=lm(OCFEV1~1, data=OC.data)
fullmodel=lm(OCFEV1~., data=OC.data)

model_step_b <- step(fullmodel,direction='backward')
```

```
## Start:  AIC=-439.51
## OCFEV1 ~ OCAGE + OCHEIGHT + OCWEIGHT + OCFVC
##
##            Df Sum of Sq     RSS     AIC
## - OCWEIGHT  1    0.0195  7.5119 -441.12
## <none>                   7.4924 -439.51
## - OCAGE     1    0.1051  7.5975 -439.42
## - OCHEIGHT  1    0.4127  7.9051 -433.47
## - OCFVC     1   13.9042 21.3966 -284.11
##
## Step:  AIC=-441.12
## OCFEV1 ~ OCAGE + OCHEIGHT + OCFVC
##
##            Df Sum of Sq     RSS     AIC
## <none>                   7.5119 -441.12
## - OCAGE     1    0.1240  7.6359 -440.67
## - OCHEIGHT  1    0.5058  8.0176 -433.35
## - OCFVC     1   17.1876 24.6994 -264.58
```

```r
model_step_f <- step(nullmodel, scope=list(lower=nullmodel, upper=fullmodel), direction='forward')
```

```
## Start:  AIC=18.8
## OCFEV1 ~ 1
##
##            Df Sum of Sq     RSS     AIC
## + OCFVC     1    158.27   9.505 -409.82
## + OCHEIGHT  1    143.07  24.703 -266.56
## + OCWEIGHT  1    133.89  33.886 -219.14
## + OCAGE     1    119.33  48.445 -165.53
## <none>                  167.777   18.80
##
## Step:  AIC=-409.82
## OCFEV1 ~ OCFVC
##
##            Df Sum of Sq    RSS     AIC
## + OCHEIGHT  1   1.86902 7.6359 -440.67
## + OCAGE     1   1.48730 8.0176 -433.35
## + OCWEIGHT  1   0.73429 8.7706 -419.88
## <none>                  9.5049 -409.82
##
## Step:  AIC=-440.67
## OCFEV1 ~ OCFVC + OCHEIGHT
##
##            Df Sum of Sq    RSS     AIC
```

```
## + OCAGE     1  0.124029 7.5119 -441.12
## <none>                  7.6359 -440.67
## + OCWEIGHT  1  0.038397 7.5975 -439.42
##
## Step:  AIC=-441.12
## OCFEV1 ~ OCFVC + OCHEIGHT + OCAGE
##
##            Df Sum of Sq    RSS     AIC
## <none>                  7.5119 -441.12
## + OCWEIGHT  1  0.019471 7.4924 -439.51
```

```r
summary(model_step_b)
```

```
##
## Call:
## lm(formula = OCFEV1 ~ OCAGE + OCHEIGHT + OCFVC, data = OC.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7877 -0.1039  0.0295  0.1407  0.7271
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.2635095  0.3472279  -3.639 0.000379 ***
## OCAGE        0.0220124  0.0141776   1.553 0.122680
## OCHEIGHT     0.0280260  0.0089390   3.135 0.002076 **
## OCFVC        0.0063140  0.0003455  18.277  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2268 on 146 degrees of freedom
## Multiple R-squared:  0.9552, Adjusted R-squared:  0.9543
## F-statistic:  1038 on 3 and 146 DF,  p-value: < 2.2e-16
```

```r
summary(model_step_f)
```

```
##
## Call:
## lm(formula = OCFEV1 ~ OCFVC + OCHEIGHT + OCAGE, data = OC.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7877 -0.1039  0.0295  0.1407  0.7271
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.2635095  0.3472279  -3.639 0.000379 ***
## OCFVC        0.0063140  0.0003455  18.277  < 2e-16 ***
## OCHEIGHT     0.0280260  0.0089390   3.135 0.002076 **
## OCAGE        0.0220124  0.0141776   1.553 0.122680
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.2268 on 146 degrees of freedom
## Multiple R-squared:  0.9552, Adjusted R-squared:  0.9543
## F-statistic:  1038 on 3 and 146 DF,  p-value: < 2.2e-16
```

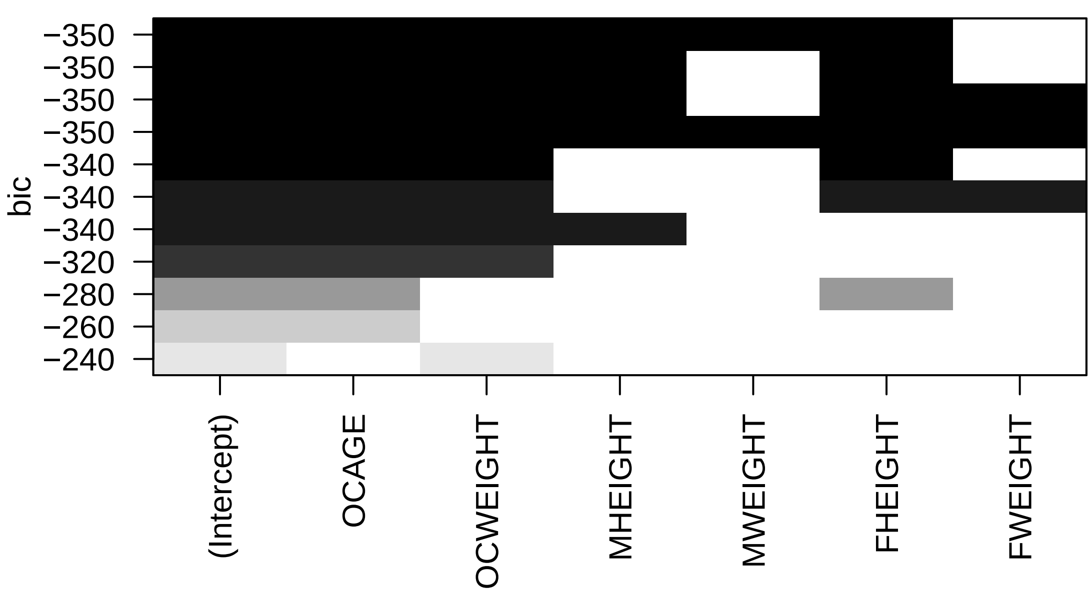
```r
models <- summary(subset_result)
data.frame(AdjustedR2 = models$adjr2)
```

```
##   AdjustedR2
## 1  0.9429650
## 2  0.9538685
## 3  0.9543070
## 4  0.9541111
```

Using the models above, and looking at the r-squared values for each of predictors, the variables that best predict fev1 in the oldest child in order from strongest to least strong are FVC, Height, AGE, then weight.

2. PMA6 9.11

```r
subset_result <- regsubsets(OCHEIGHT ~ OCAGE + OCWEIGHT + MHEIGHT + MWEIGHT + FHEIGHT + FWEIGHT,data=FEV
plot(subset_result, scale="bic")
```



```
summary(subset_result)
```

```
## Subset selection object
## Call: regsubsets.formula(OCHEIGHT ~ OCAGE + OCWEIGHT + MHEIGHT + MWEIGHT +
##     FHEIGHT + FWEIGHT, data = FEV, nbest = 2, nvmax = 14)
## 6 Variables  (and intercept)
##          Forced in Forced out
## OCAGE        FALSE      FALSE
## OCWEIGHT     FALSE      FALSE
## MHEIGHT      FALSE      FALSE
## MWEIGHT      FALSE      FALSE
## FHEIGHT      FALSE      FALSE
## FWEIGHT      FALSE      FALSE
## 2 subsets of each size up to 6
## Selection Algorithm: exhaustive
##          OCAGE OCWEIGHT MHEIGHT MWEIGHT FHEIGHT FWEIGHT
## 1  ( 1 ) "*"   " "      " "     " "     " "     " "
## 1  ( 2 ) " "   "*"      " "     " "     " "     " "
## 2  ( 1 ) "*"   "*"      " "     " "     " "     " "
## 2  ( 2 ) "*"   " "      " "     " "     "*"     " "
## 3  ( 1 ) "*"   "*"      " "     " "     "*"     " "
## 3  ( 2 ) "*"   "*"      "*"     " "     " "     " "
## 4  ( 1 ) "*"   "*"      "*"     " "     "*"     " "
## 4  ( 2 ) "*"   "*"      " "     " "     "*"     "*"
## 5  ( 1 ) "*"   "*"      "*"     "*"     "*"     " "
## 5  ( 2 ) "*"   "*"      "*"     " "     "*"     "*"
## 6  ( 1 ) "*"   "*"      "*"     "*"     "*"     "*"
```

Using subset regression, the variables that best predict height in the oldest child are their age, their weight, and the height of the father.

3. PMA6 9.12

```
model_12 <- FEV %>% select(OCHEIGHT, OCSEX, OCAGE, OCWEIGHT, MHEIGHT, MWEIGHT, FHEIGHT, FWEIGHT)
model_12$OCSEX <- ifelse(model_12$OCSEX == 1, "Male", "Female")

model_12female <- model_12 %>% select(OCHEIGHT, OCSEX, OCAGE, OCWEIGHT, MHEIGHT, MWEIGHT, FHEIGHT, FWEI
  filter(OCSEX == "Female")
model_12male <- model_12 %>% select(OCHEIGHT, OCSEX, OCAGE, OCWEIGHT, MHEIGHT, MWEIGHT, FHEIGHT, FWEIGHT
  filter(OCSEX == "Male")

model_12female <- model_12female %>% select(OCHEIGHT, OCAGE, OCWEIGHT, MHEIGHT, MWEIGHT, FHEIGHT, FWEIG
model_12male <- model_12male %>% select(OCHEIGHT, OCAGE, OCWEIGHT, MHEIGHT, MWEIGHT, FHEIGHT, FWEIGHT)
View(model_12)
summary(regsubsets(OCHEIGHT ~., data = model_12female, nvmax = 3))
```

```
## Subset selection object
## Call: regsubsets.formula(OCHEIGHT ~ ., data = model_12female, nvmax = 3)
## 6 Variables  (and intercept)
##          Forced in Forced out
```

```
## OCAGE        FALSE      FALSE
## OCWEIGHT     FALSE      FALSE
## MHEIGHT      FALSE      FALSE
## MWEIGHT      FALSE      FALSE
## FHEIGHT      FALSE      FALSE
## FWEIGHT      FALSE      FALSE
## 1 subsets of each size up to 3
## Selection Algorithm: exhaustive
##          OCAGE OCWEIGHT MHEIGHT MWEIGHT FHEIGHT FWEIGHT
## 1  ( 1 ) "*"   " "      " "     " "     " "     " "
## 2  ( 1 ) "*"   " "      " "     " "     "*"     " "
## 3  ( 1 ) "*"   "*"      "*"     " "     " "     " "
```

```
summary(regsubsets(OCHEIGHT ~., data = model_12male, nvmax = 3))
```

```
## Subset selection object
## Call: regsubsets.formula(OCHEIGHT ~ ., data = model_12male, nvmax = 3)
## 6 Variables  (and intercept)
##          Forced in Forced out
## OCAGE        FALSE      FALSE
## OCWEIGHT     FALSE      FALSE
## MHEIGHT      FALSE      FALSE
## MWEIGHT      FALSE      FALSE
## FHEIGHT      FALSE      FALSE
## FWEIGHT      FALSE      FALSE
## 1 subsets of each size up to 3
## Selection Algorithm: exhaustive
##          OCAGE OCWEIGHT MHEIGHT MWEIGHT FHEIGHT FWEIGHT
## 1  ( 1 ) "*"   " "      " "     " "     " "     " "
## 2  ( 1 ) "*"   "*"      " "     " "     " "     " "
## 3  ( 1 ) "*"   "*"      " "     " "     "*"     " "
```

For girls who are the oldest child the best variables to predict height are their age, weight and height of the middle child. For boys who are the oldest child, the best predictors are age, weight, and fathers height.

4. PMA6 9.13

Some potential confounding variables could be race or ethnicity since different cultures drink at different ages. A precision variable could be the amount of parents they live with or their friends/siblings.

```
hiv.model <- hiv %>% select(AGEALC, ETHN, LIVWITH, SIBLINGS, FRNDS) %>% na.omit()
hiv.model$ETHN <- factor(hiv.model$ETHN,labels = c("Hispanic","Black","Other"))

hiv.model$LIVWITH <- factor(hiv.model$LIVWITH,labels = c("Both_parents","One_parent","Other"))

hiv_lm <- lm(AGEALC ~  ETHN + LIVWITH + LIVWITH*ETHN, data = hiv.model)
summary(hiv_lm)
```

```
##
## Call:
## lm(formula = AGEALC ~ ETHN + LIVWITH + LIVWITH * ETHN, data = hiv.model)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.250  -6.125  -2.750   6.875  10.875
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    6.0400     1.3334   4.530 9.33e-06 ***
## ETHNBlack                     -3.2900     2.3414  -1.405    0.161
## ETHNOther                     -2.3257     2.8510  -0.816    0.415
## LIVWITHOne_parent              0.3308     1.5091   0.219    0.827
## LIVWITHOther                   4.2100     2.7082   1.555    0.121
## ETHNBlack:LIVWITHOne_parent    3.0442     2.5689   1.185    0.237
## ETHNOther:LIVWITHOne_parent    2.5105     3.3312   0.754    0.452
## ETHNBlack:LIVWITHOther         1.8733     3.8396   0.488    0.626
## ETHNOther:LIVWITHOther         1.8757     4.7513   0.395    0.693
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.667 on 239 degrees of freedom
## Multiple R-squared:  0.04204,    Adjusted R-squared:  0.009971
## F-statistic: 1.311 on 8 and 239 DF,  p-value: 0.2386
```

None of the variables had a significant p level. I might need to fit more variables with the model.

4. PMA6 9.14

```r
drink.model <- regsubsets(AGEALC ~ GENDER + LIVWITH + SIBLINGS + JOBMO +
EDUMO + HOWREL + ATTSERV + NGHB1 + NGHB2 + NGHB3 + NGHB4 + NGHB5 + NGHB6 +
NGHB7 + NGHB8 + NGHB9 + NGHB10 + NGHB11 + FINSIT + ETHN + AGESMOKE + AGEMAR +
LIKESCH + HOOKEY + NHOOKEY, data=filter(hiv, AGEALC!="0"))
summary(drink.model)
```

```
## Subset selection object
## Call: regsubsets.formula(AGEALC ~ GENDER + LIVWITH + SIBLINGS + JOBMO +
##     EDUMO + HOWREL + ATTSERV + NGHB1 + NGHB2 + NGHB3 + NGHB4 +
##     NGHB5 + NGHB6 + NGHB7 + NGHB8 + NGHB9 + NGHB10 + NGHB11 +
##     FINSIT + ETHN + AGESMOKE + AGEMAR + LIKESCH + HOOKEY + NHOOKEY,
##     data = filter(hiv, AGEALC != "0"))
## 25 Variables  (and intercept)
##            Forced in Forced out
## GENDERMale     FALSE      FALSE
## LIVWITH        FALSE      FALSE
## SIBLINGS       FALSE      FALSE
## JOBMO          FALSE      FALSE
## EDUMO          FALSE      FALSE
## HOWREL         FALSE      FALSE
## ATTSERV        FALSE      FALSE
## NGHB1          FALSE      FALSE
## NGHB2          FALSE      FALSE
## NGHB3          FALSE      FALSE
## NGHB4          FALSE      FALSE
```

```
## NGHB5          FALSE      FALSE
## NGHB6          FALSE      FALSE
## NGHB7          FALSE      FALSE
## NGHB8          FALSE      FALSE
## NGHB9          FALSE      FALSE
## NGHB10         FALSE      FALSE
## NGHB11         FALSE      FALSE
## FINSIT         FALSE      FALSE
## ETHN           FALSE      FALSE
## AGESMOKE       FALSE      FALSE
## AGEMAR         FALSE      FALSE
## LIKESCH        FALSE      FALSE
## HOOKEY         FALSE      FALSE
## NHOOKEY        FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          GENDERMale LIVWITH SIBLINGS JOBMO EDUMO HOWREL ATTSERV NGHB1 NGHB2
## 1  ( 1 ) " "        " "     " "      " "   " "   " "    " "     " "   " "
## 2  ( 1 ) " "        " "     " "      " "   " "   " "    " "     " "   " "
## 3  ( 1 ) " "        " "     " "      " "   " "   " "    " "     " "   " "
## 4  ( 1 ) "*"        " "     " "      " "   " "   " "    " "     " "   " "
## 5  ( 1 ) "*"        " "     " "      " "   " "   " "    " "     " "   " "
## 6  ( 1 ) "*"        " "     "*"      " "   "*"   " "    " "     " "   " "
## 7  ( 1 ) "*"        " "     "*"      " "   "*"   " "    " "     " "   " "
## 8  ( 1 ) "*"        " "     "*"      " "   "*"   " "    " "     " "   " "
##          NGHB3 NGHB4 NGHB5 NGHB6 NGHB7 NGHB8 NGHB9 NGHB10 NGHB11 FINSIT ETHN
## 1  ( 1 ) " "   " "   " "   " "   " "   " "   " "   " "    " "    " "    " "
## 2  ( 1 ) " "   " "   " "   " "   " "   " "   "*"   " "    " "    " "    " "
## 3  ( 1 ) " "   " "   " "   " "   " "   " "   "*"   " "    " "    " "    " "
## 4  ( 1 ) " "   " "   " "   " "   " "   " "   "*"   " "    " "    " "    " "
## 5  ( 1 ) " "   " "   " "   " "   " "   " "   "*"   " "    "*"    " "    " "
## 6  ( 1 ) " "   " "   " "   " "   " "   " "   "*"   " "    " "    " "    " "
## 7  ( 1 ) " "   " "   " "   " "   " "   " "   "*"   " "    "*"    " "    " "
## 8  ( 1 ) " "   " "   " "   " "   " "   "*"   "*"   " "    "*"    " "    " "
##          AGESMOKE AGEMAR LIKESCH HOOKEY NHOOKEY
## 1  ( 1 ) "*"      " "    " "     " "    " "
## 2  ( 1 ) "*"      " "    " "     " "    " "
## 3  ( 1 ) "*"      " "    " "     " "    "*"
## 4  ( 1 ) "*"      " "    " "     " "    "*"
## 5  ( 1 ) "*"      " "    " "     " "    "*"
## 6  ( 1 ) "*"      " "    " "     " "    "*"
## 7  ( 1 ) "*"      " "    " "     " "    "*"
## 8  ( 1 ) "*"      " "    " "     " "    "*"
```

```r
drink.lm <- lm(AGEALC ~ NGHB9 + AGESMOKE + NHOOKEY, data=filter(hiv, AGEALC!="0"))
summary(drink.lm)
```

```
##
## Call:
## lm(formula = AGEALC ~ NGHB9 + AGESMOKE + NHOOKEY, data = filter(hiv,
##     AGEALC != "0"))
##
## Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -8.8044 -0.6072  0.4447  1.4086  5.3881
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.51485    1.26714   6.720 1.55e-09 ***
## NGHB9       -0.35708    0.21577  -1.655 0.101383
## AGESMOKE     0.37546    0.09672   3.882 0.000196 ***
## NHOOKEY      0.13043    0.07808   1.670 0.098266 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.442 on 91 degrees of freedom
##   (31 observations deleted due to missingness)
## Multiple R-squared:  0.1821, Adjusted R-squared:  0.1552
## F-statistic: 6.755 on 3 and 91 DF,  p-value: 0.0003646
```

The best variables to predict the age at which adolescents started drinking are NGHB9(homelessness in the community),AGESMOKE(the age that they started smoking), and NHOOKEY(how often they skipped school). When a linear model is fit using these 3 variables as predictors, we get P-values which are all $<.01$. The age at which they started smoking seems to be the best predictor with a p-value of .0002 when homelessness and hookey are held constant.