Modeling Considerations

Lab 8 Handout Solutions

Statistics 139

Topics

- Inferential Modeling
- Predictive Modeling
  - Sequential Variable Selection
  - Comparing Models (with and without CV)

Background Information

This handout will step through a case study examining evidence for ethnic discrimination in the amount of financial support offered by the State of California to individuals with developmental disabilities. Although an initial look at the data suggested an association between expenditures and ethnicity (specifically between Hispanics and White non-Hispanics), further exploratory analysis suggested that age is a confounding variable for the relationship.

The data in `dds.discr` represent a random sample of 1,000 individuals who receive financial support from the California Department of Developmental Services (out of a total population of 250,000). The following variables are included in the dataset.

- `ID`: consumer ID number
  - `Age.Cohort`: age group, where `1` refers to 0 - 5 years, `2` refers to 51+ years, `3` refers to 13 - 17 years, `4` refers to 18 - 21 years, `5` refers to 22 - 50 years, and `6` refers to 6 - 12 years.
  - `Age`: age in years
  - `Gender`: gender, recorded as `1` for female and `2` for male
  - `Expenditures`: annual expenditure in dollars
  - `Ethnicity`: ethnicity, recorded as either `1` for American Indian, `2` for Asian, `3` for Black, `4` for Hispanic, `5` for Multi Race, `6` for Native Hawaiian, `7` for Other, and `8` for White not Hispanic.

In this handout, we return to the data with the tools of inference and regression modeling to conduct a formal analysis:

After adjusting for age as a confounder, is there evidence that the mean amount of financial support differs between Hispanics and White non-Hispanics?

Problem 1: Initial Model Fitting Run the code below to read in the data set and create a subset of the data to include only observations from Hispanic and White non-Hispanic consumers. Use this for all future analyses.

```
#load the data
dds = read.csv("data/dds_discr.csv")
```

```r
#subset the data
dds.subset = dds[dds$ethnicity == "Hispanic" |
                 dds$ethnicity == "White not Hispanic", ]

#how about with tidyverse?
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
dds.subset2 <- dds %>% filter(ethnicity %in% c("Hispanic", "White not Hispanic"))
```

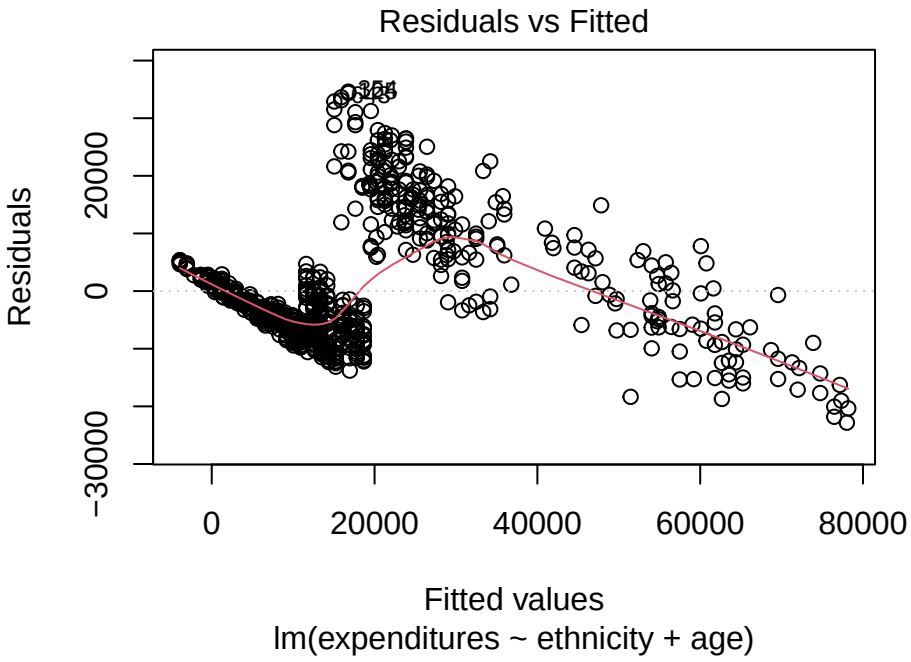
a) Fit a multiple regression model predicting expenditures from ethnicity and age. Interpret the ethnicity coefficient and investigate the model assumptions with residual plots.

```r
model1 <- lm(expenditures ~ ethnicity + age, data = dds.subset)
summary(model1)
```

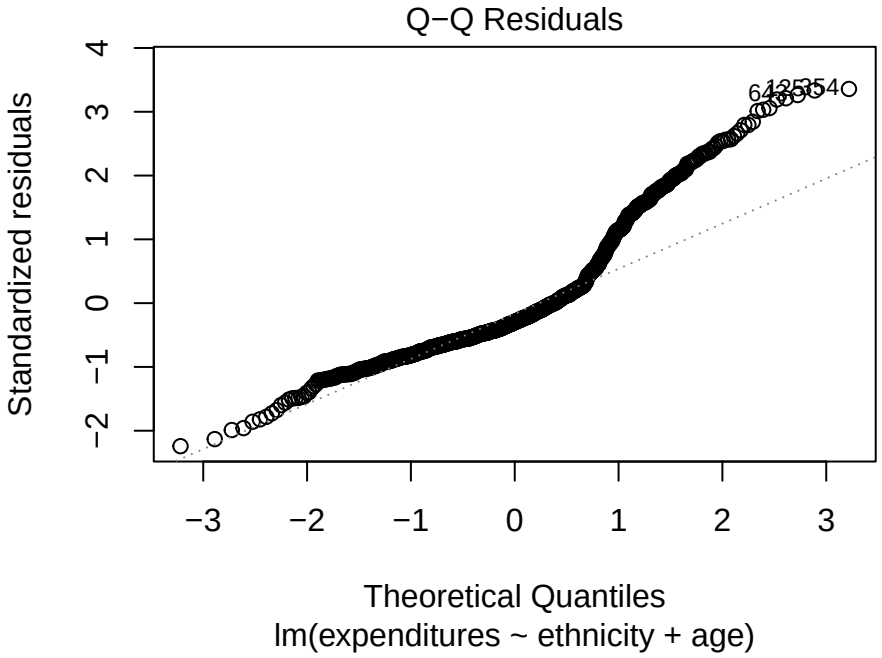
```
##
## Call:
## lm(formula = expenditures ~ ethnicity + age, data = dds.subset)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -22829  -6633  -3083   3168  34612
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -3920.06     645.89  -6.069 2.01e-09 ***
## ethnicityWhite not Hispanic   4489.61     773.93   5.801 9.60e-09 ***
## age                            862.48      21.05  40.977  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10320 on 774 degrees of freedom
## Multiple R-squared:  0.7227, Adjusted R-squared:  0.722
## F-statistic:  1009 on 2 and 774 DF,  p-value: < 2.2e-16
```

On average, White non-Hispanics have about $4490 higher annual expenditures, conditional on age. This is highly significant.

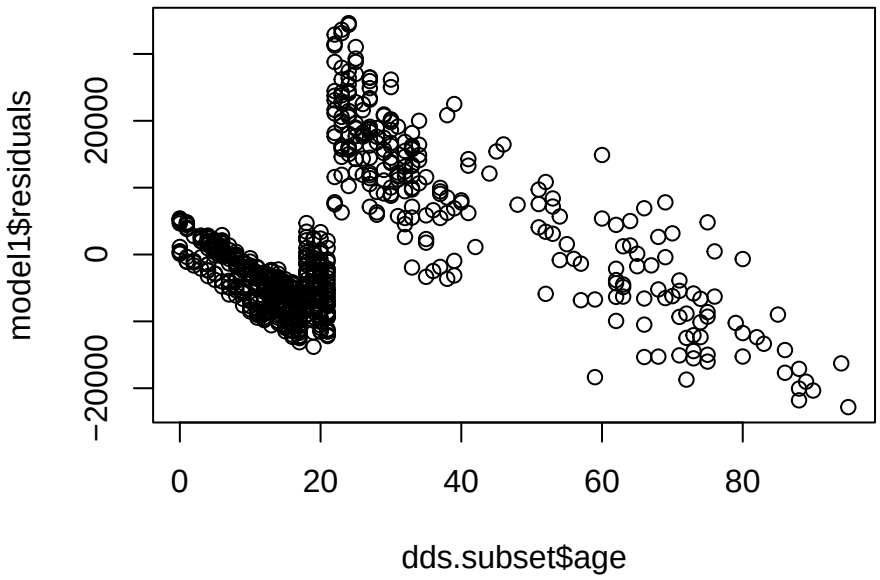
```r
#look at residuals versus fitted values
plot(model1, which=1)
```




```
plot(model1, which=2)
```




```
plot(y=model1$residuals, x=dds.subset$age)
```



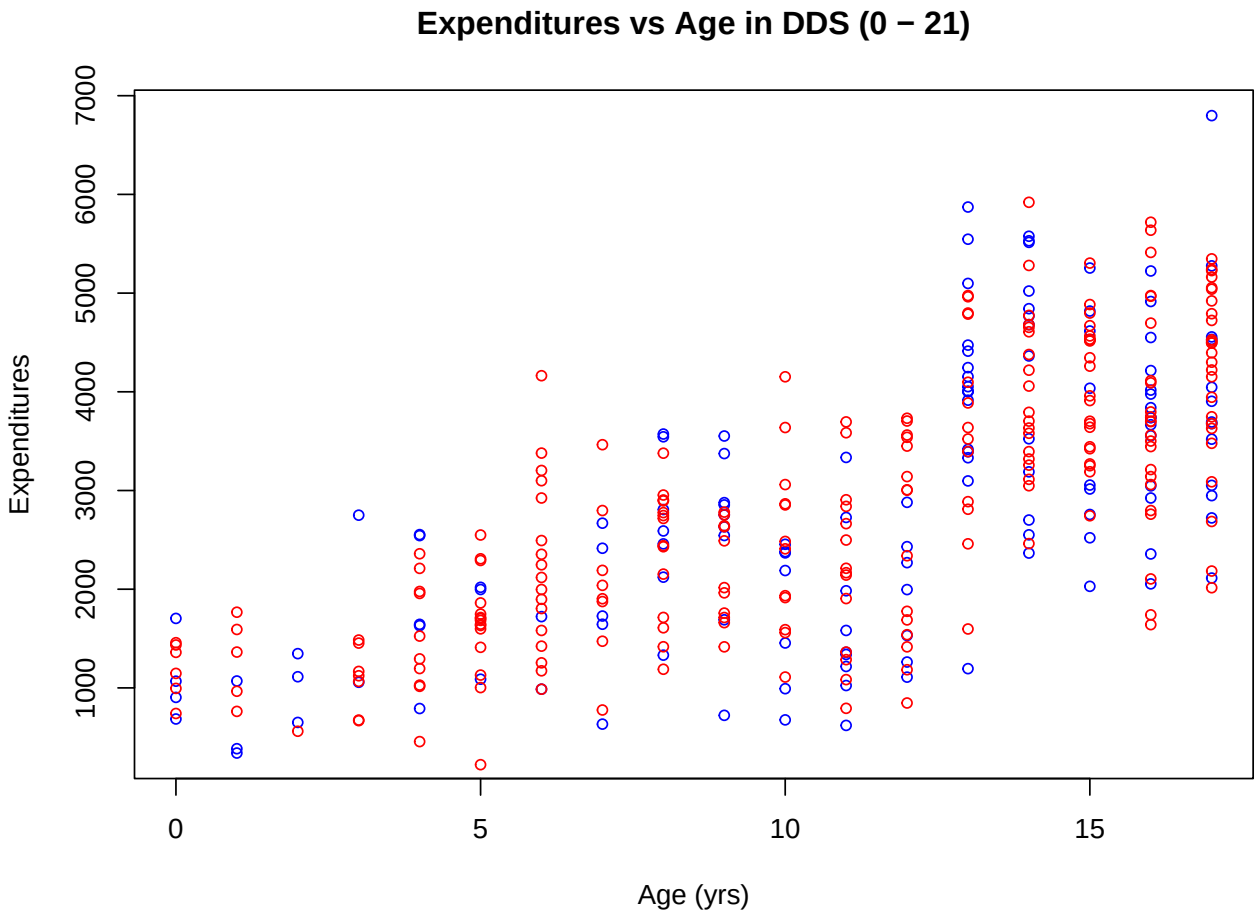
The residuals clearly show a pattern, rather than random scatter about the $y = 0$ line. There are lots of violations, but clearly the biggest is the non-linearity with respect to age. The residuals show marked departures from normality, particularly in the upper tail. There are many more large residuals than expected if the residuals were normally distributed. As-is, this model is not appropriate for modeling the relationship between expenditures, ethnicity, and age.

The residuals suggest that the coverage policy determining the amount of financial support (expenditure) differs drastically as individuals transition from adolescence to legal majority age.

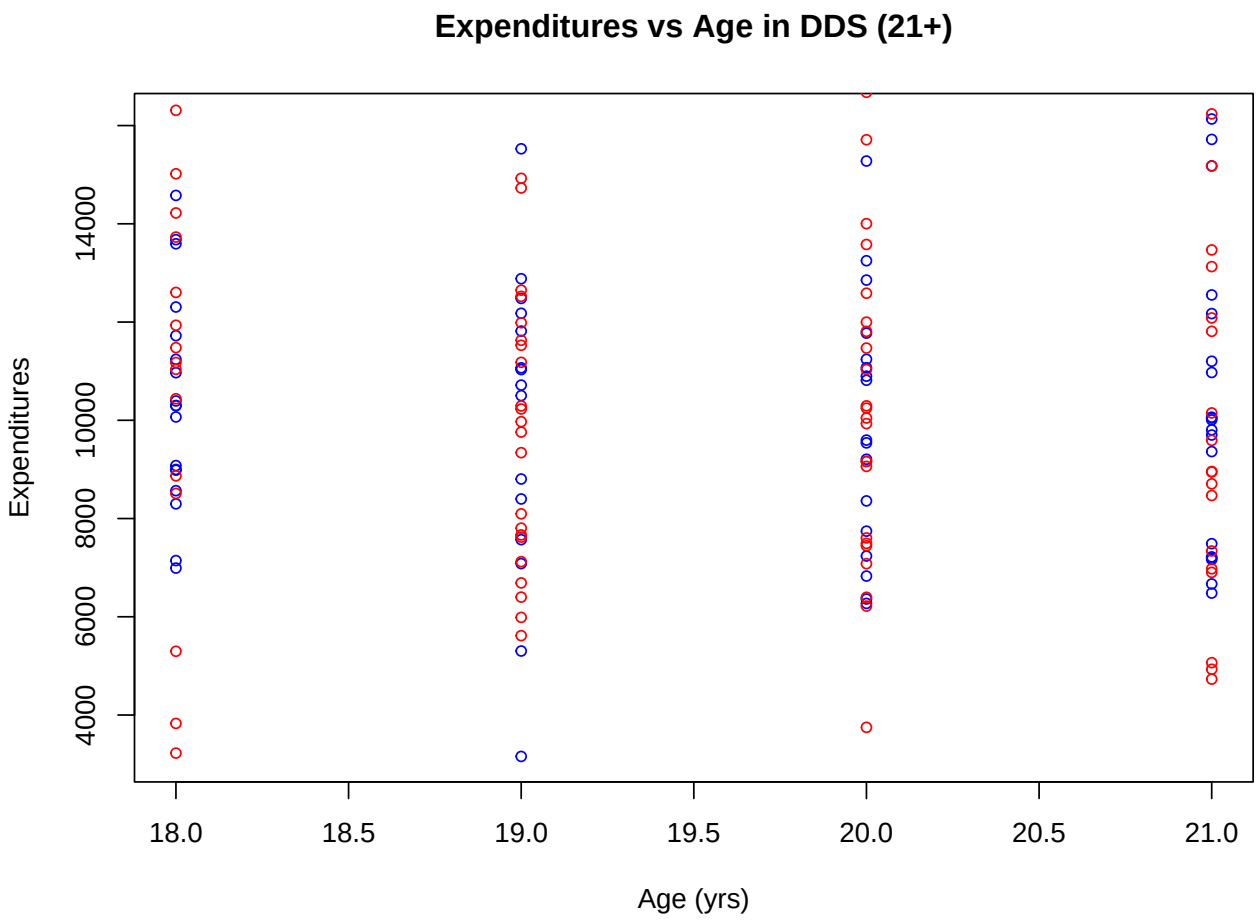
c) Investigate the association of expenditures and age for three separate age groups with scatter plots: under 18 years, between 18 and 21 years (inclusive), and above 21 years. Use color to differentiate between Hispanics and White non-Hispanics and explain what you see.

```
#create hispanic and white.not.hisp logicals
hispanic = (dds.subset$ethnicity == "Hispanic")
white.not.hisp = (dds.subset$ethnicity == "White not Hispanic")

## expenditures vs age for ages >= 21 , ages < 21
youngest <- (dds.subset$age < 18)
middle = (dds.subset$age <= 21 & dds.subset$age >=18)
oldest = (dds.subset$age > 21)



#Plot in youngest group
#plot blue points, white not hispanic
plot(expenditures[white.not.hisp & youngest] ~ age[white.not.hisp & youngest],
     data = dds.subset, pch = 21, col =  "blue", cex = 0.8,
     xlab = "Age (yrs)", ylab = "Expenditures",
     main = "Expenditures vs Age in DDS (0 - 21)")

#plot red points, hispanic
points(expenditures[hispanic & youngest] ~ age[hispanic & youngest],
```

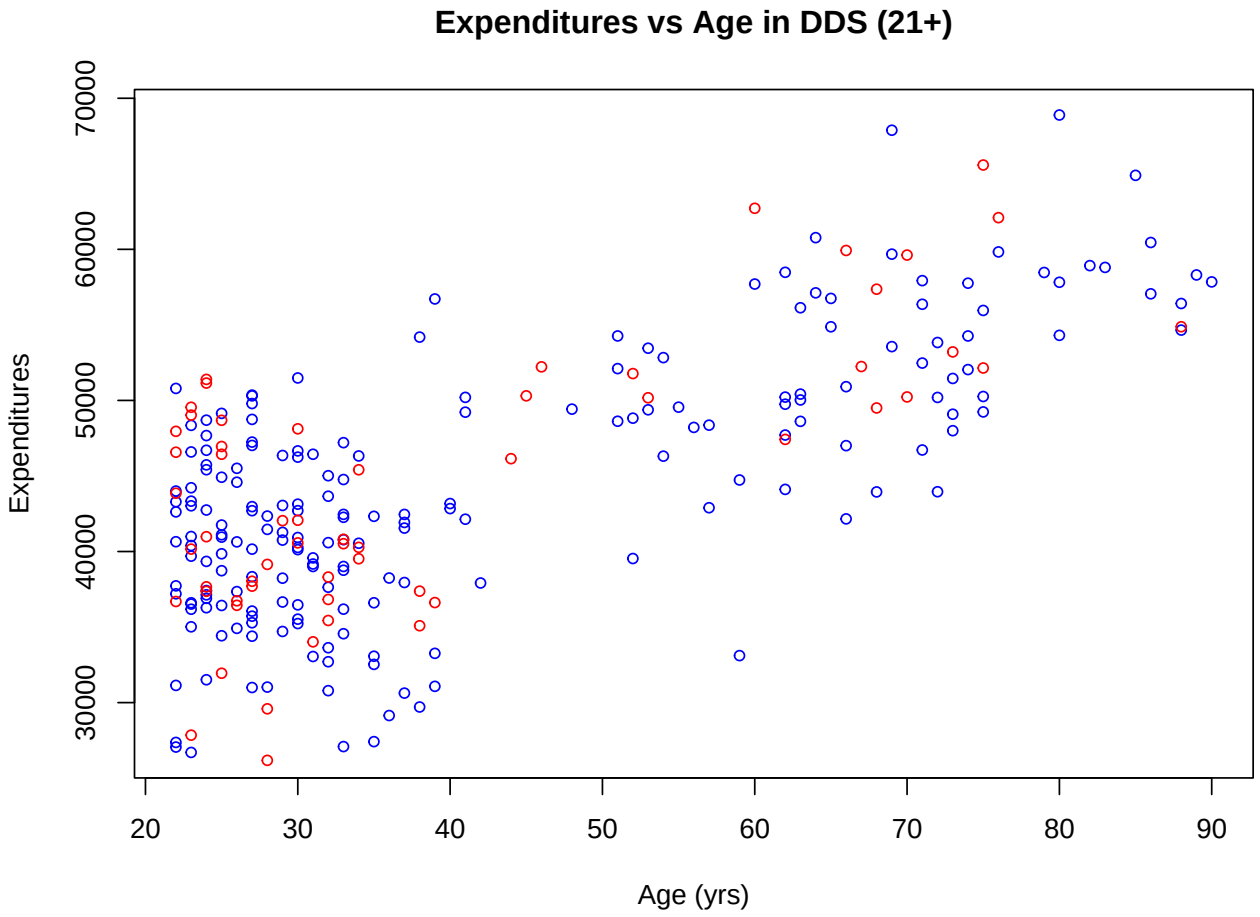
```
       data = dds.subset, pch = 21, col =  "red", cex = 0.8)
```




```
#Plot in oldest group
#plot blue points, white not hispanic
plot(expenditures[white.not.hisp & middle] ~ age[white.not.hisp & middle],
     data = dds.subset, pch = 21, col = "blue", cex = 0.8,
     xlab = "Age (yrs)", ylab = "Expenditures",
     main = "Expenditures vs Age in DDS (21+)")

#plot red points, hispanic
points(expenditures[hispanic & middle] ~ age[hispanic & middle],
       data = dds.subset, pch = 21, col =  "red", cex = 0.8)
```



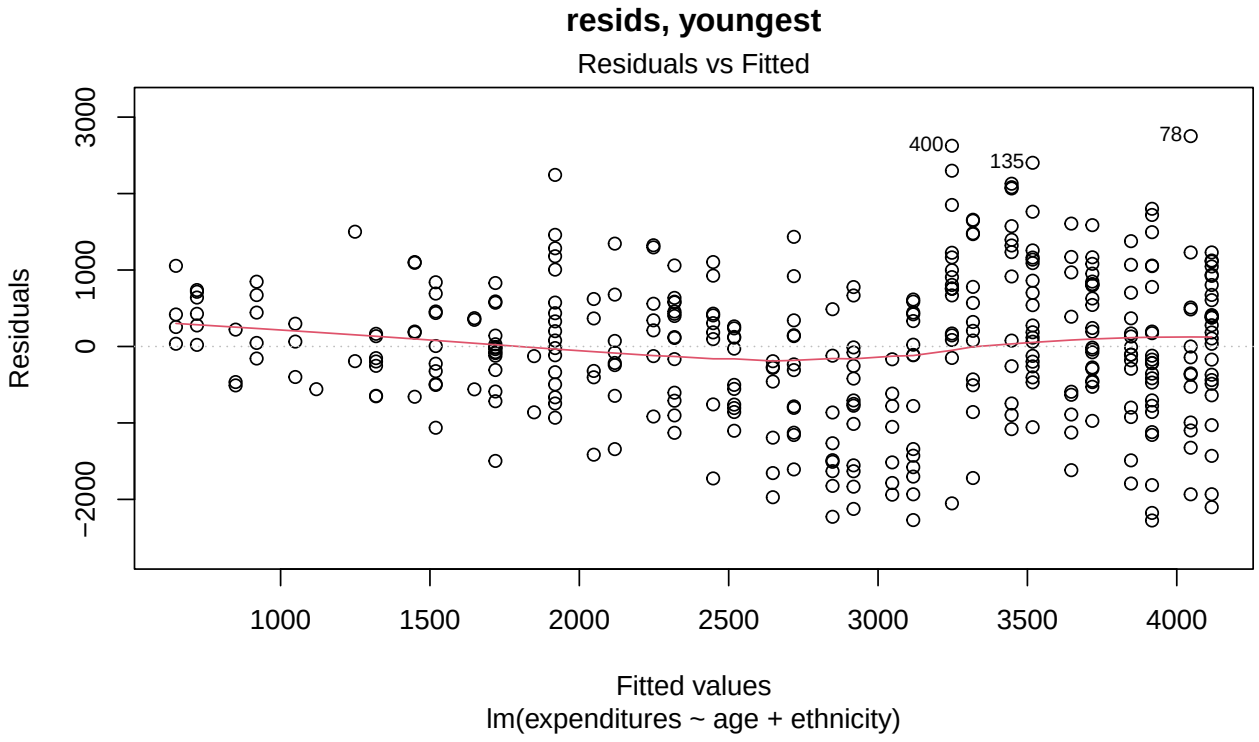

```
#Plot in oldest group
#plot blue points, white not hispanic
plot(expenditures[white.not.hisp & oldest] ~ age[white.not.hisp & oldest],
     data = dds.subset, pch = 21, col = "blue", cex = 0.8,
     xlab = "Age (yrs)", ylab = "Expenditures",
     main = "Expenditures vs Age in DDS (21+)")

#plot red points, hispanic
points(expenditures[hispanic & oldest] ~ age[hispanic & oldest],
       data = dds.subset, pch = 21, col =  "red", cex = 0.8)
```



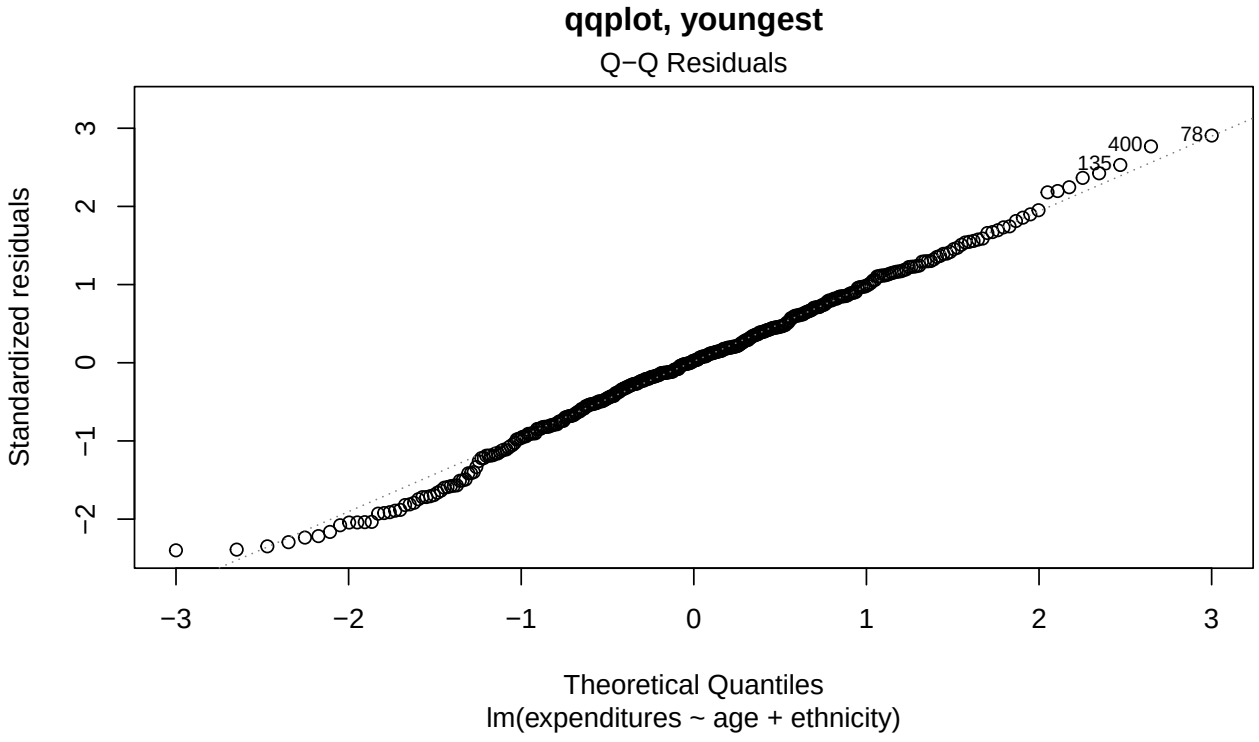
There does not seem to be a systematic difference in the amount of financial support granted to Hispanics versus White non-Hispanics, when adjusting for age. At any given age in the plots, there seems to be an even spread of blue points (white non-Hispanics) and red points (Hispanics) across the range of expenditures at that age.

d) Now fit three separate linear regression models predicting expenditures from age and ethnicity, considering only the individuals in a particular age group at a time: under 18 years, between 18 and 21 years (inclusive), and above 21 years. Comment on these models based on the diagnostics?

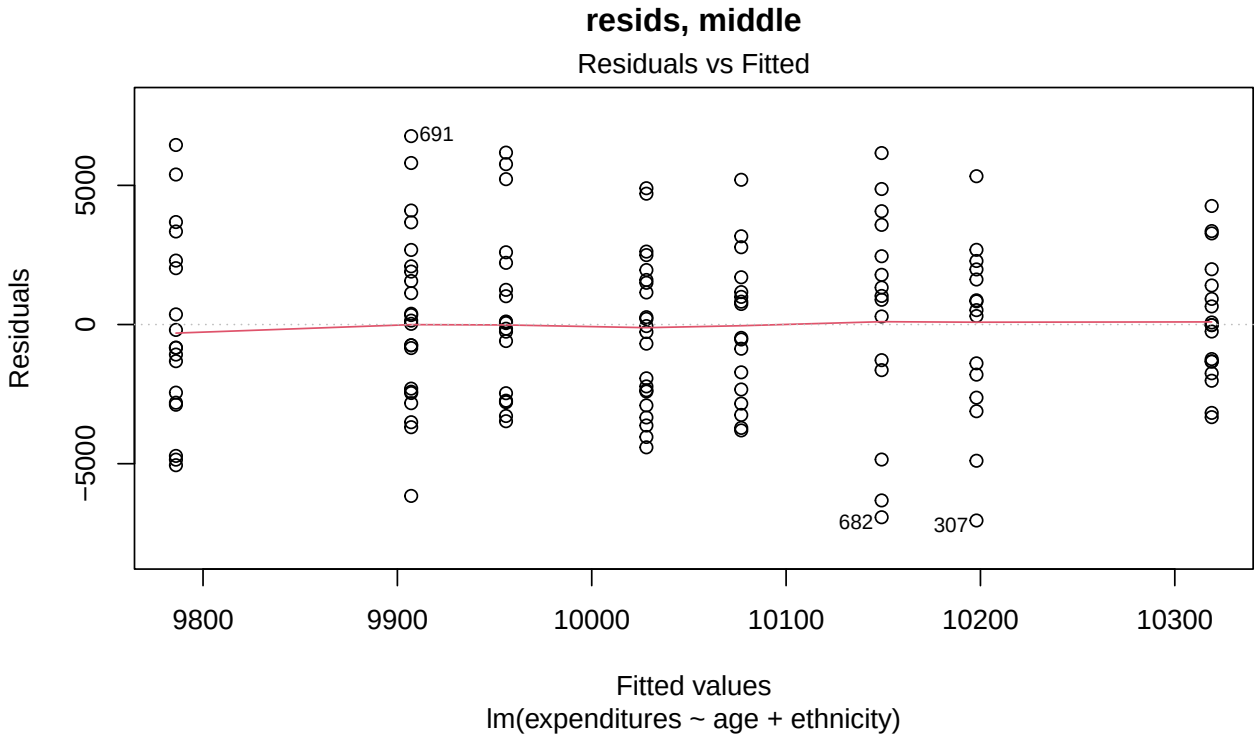
```
model.youngest = lm(expenditures[youngest] ~ age[youngest] + ethnicity[youngest],
                    data = dds.subset)

#or a bit simpler, like this:
model.youngest <- lm(expenditures ~ age + ethnicity, data=dds.subset, subset=youngest)

plot(model.youngest, which=1, main="resids, youngest") #residuals by fitted
```



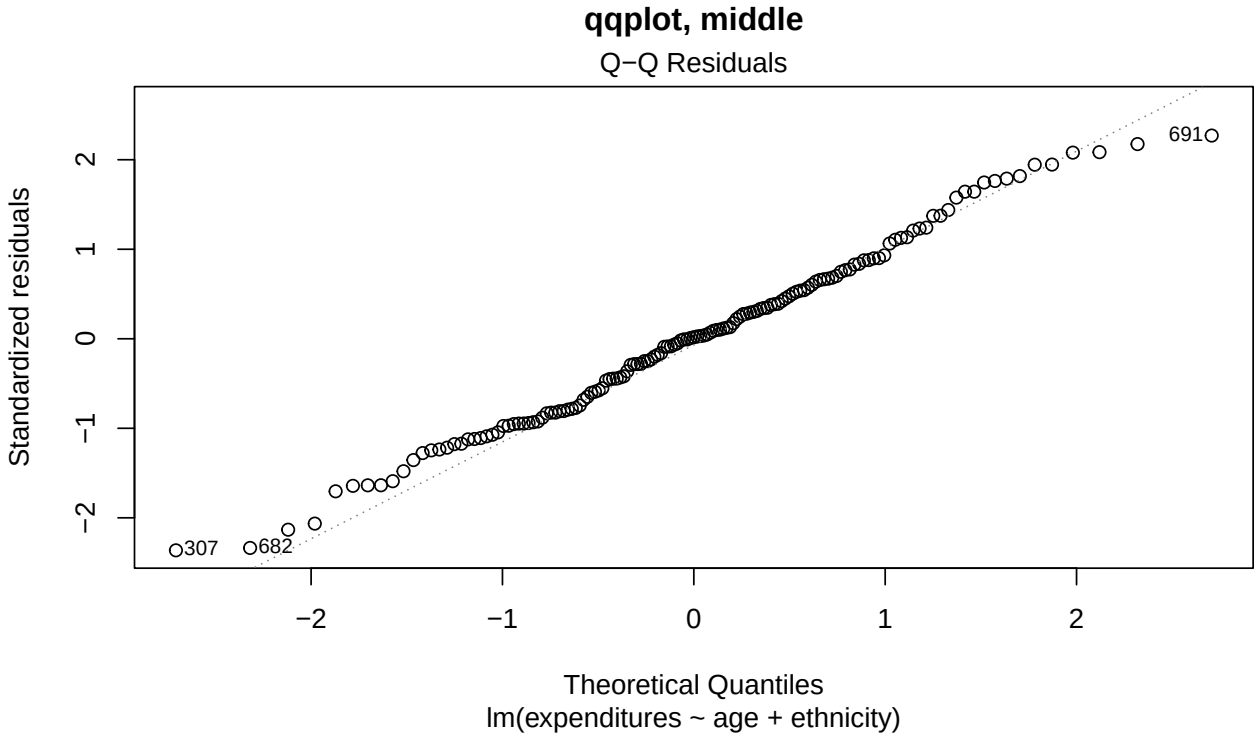

```
plot(model.youngest, which=2, main="qqplot, youngest")
```



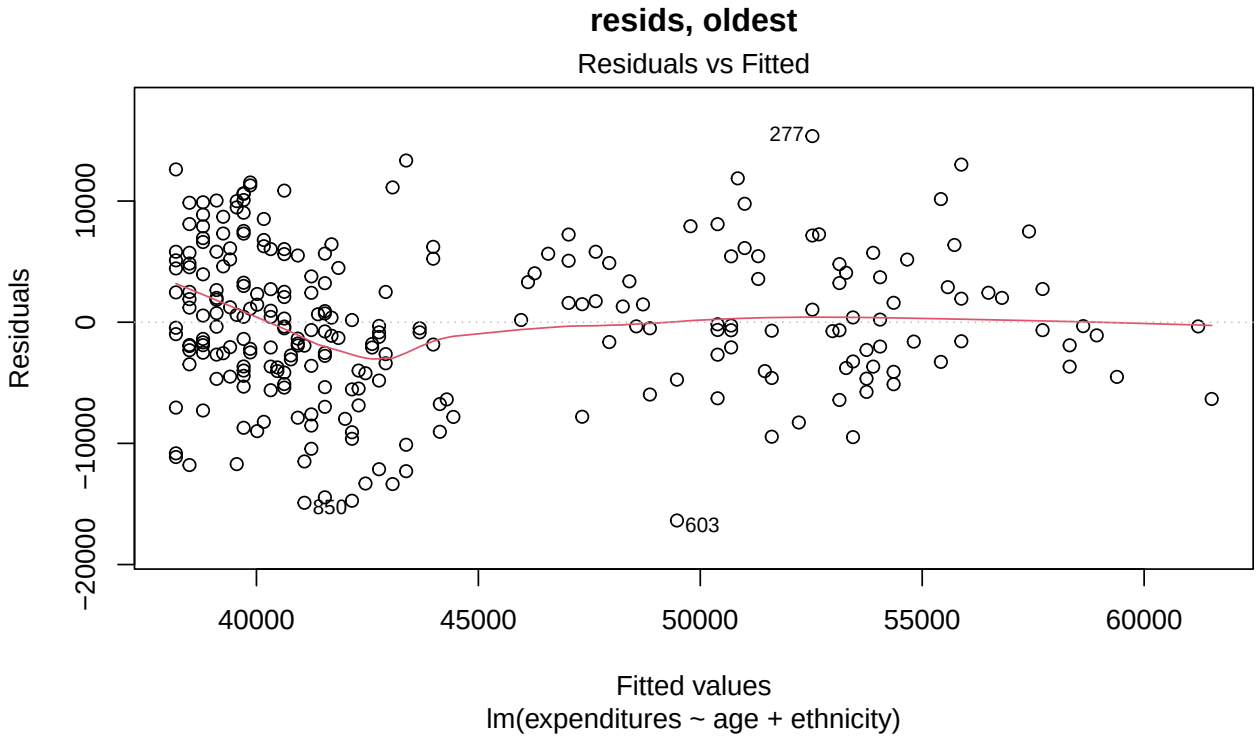

```
model.middle <- lm(expenditures ~ age + ethnicity, data=dds.subset, subset=middle)

#Do you know why there are exactly four groups here?
plot(model.middle, which=1, main="resids, middle") #residuals by fitted
```



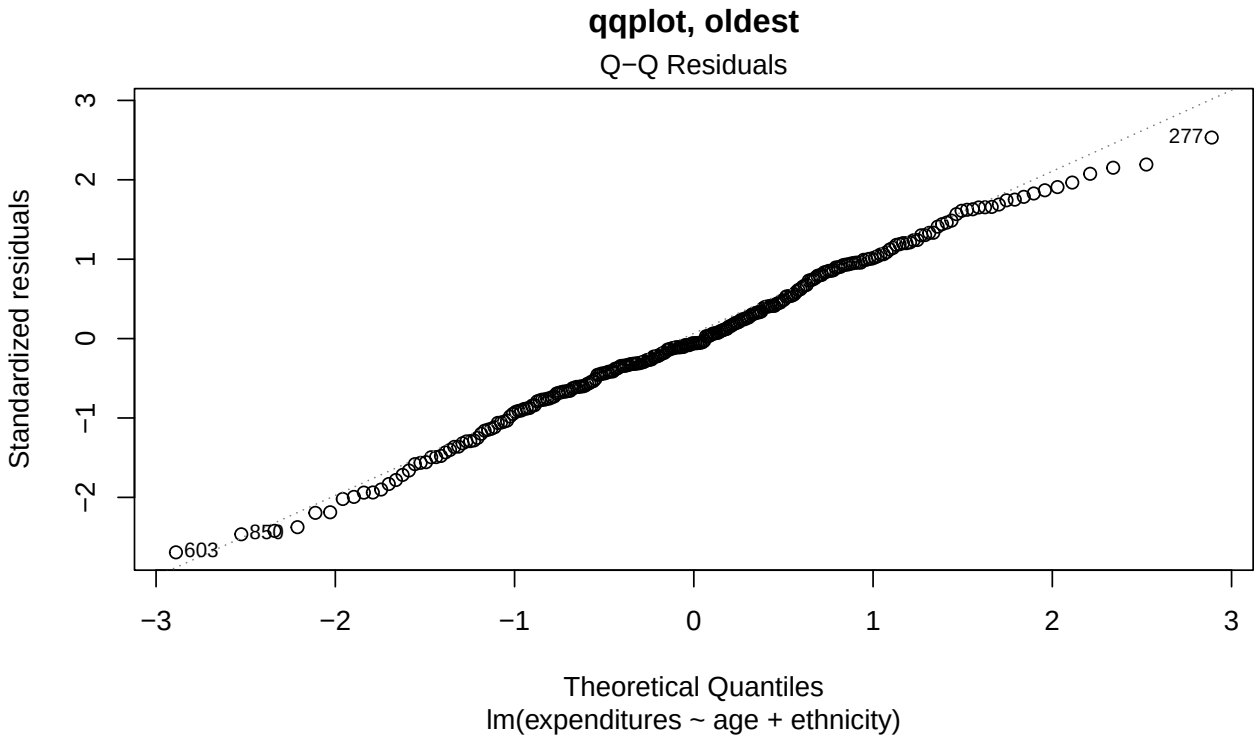

```
plot(model.middle, which=2, main="qqplot, middle")
```




```
model.oldest <- lm(expenditures ~ age + ethnicity, data=dds.subset, subset=oldest)

#Do you know why there are exactly four groups here?
plot(model.oldest, which=1, main="resids, oldest") #residuals by fitted
```




```
plot(model.oldest, which=2, main="qqplot, oldest")
```



The residual plots indicate that when applied to specific age cohorts at a time, modeling assumptions seem reasonably satisfied. Variance is roughly constant in all three groups, and the residuals follow a normal distribution well.}

e) Discuss the inference from these models - is ethnicity associated with expenditure?

```r
#summary output

summary(model.youngest)$coef
```

```
##                                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)                  719.98970  126.28973  5.7010948 2.445122e-08
## age                          199.82151   10.31961 19.3632871 4.235636e-58
## ethnicityWhite not Hispanic -70.37136  103.22065 -0.6817566 4.958217e-01
```

```r
summary(model.middle)$coef
```

```
##                                 Estimate Std. Error    t value    Pr(>|t|)
## (Intercept)                  12327.8091  4438.0504  2.7777533 0.006203036
## age                           -121.0361   226.1796 -0.5351327 0.593383527
## ethnicityWhite not Hispanic    169.8385   496.4201  0.3421264 0.732754555
```

```r
summary(model.oldest)$coef
```

```
##                                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)                  32538.5127 1105.15037 29.442611 3.679267e-84
## age                            305.0978   18.85863 16.178157 4.932641e-41
## ethnicityWhite not Hispanic -1064.8352  898.93546 -1.184551 2.372935e-01
```

Inference based on these models supports the idea that after adjusting for age, there is not evidence of an association between expenditures and ethnicity; the p-value associated with ethnicity in all three models is non-significant.

Problem 2: Refining the Model

a) One strategy for improving the model is to explicitly include a predictor that contains information about which age group an observation belongs to, since the relationship between expenditures and age is distinctly different between age groups. To this end, create a categorical variable called `age.grp` that has levels under 18 years of age, 18 - 21 years of age (inclusive), and over 21 years of age.

```r
age.grp <- rep(1, nrow(dds.subset))
age.grp[dds.subset$age < 18] <- 0
age.grp[dds.subset$age > 21] <- 2

dds.subset$age.grp <- factor(age.grp, levels = 0:2,
                             labels = c("Under 18", "18 - 21", "Over 21"))
```

b) Fit a model predicting expenditures from ethnicity, age, and age group. Interpret the model coefficients.

```r
model2 = lm(expenditures ~ (ethnicity + age)* age.grp, data = dds.subset)
summary(model2)
```

```
##
## Call:
## lm(formula = expenditures ~ (ethnicity + age) * age.grp, data = dds.subset)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16364.4  -1334.0     -7.8   1162.5  15358.6
##
## Coefficients:
##                                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                                    719.99     505.66   1.424   0.1549
## ethnicityWhite not Hispanic                    -70.37     413.29  -0.170   0.8648
## age                                            199.82      41.32   4.836  1.6e-06
## age.grp18 - 21                               11607.82    5656.61   2.052   0.0405
## age.grpOver 21                               31818.52     855.89  37.176  < 2e-16
## ethnicityWhite not Hispanic:age.grp18 - 21     240.21     753.62   0.319   0.7500
## ethnicityWhite not Hispanic:age.grpOver 21    -994.46     697.36  -1.426   0.1543
## age:age.grp18 - 21                            -320.86     290.09  -1.106   0.2690
## age:age.grpOver 21                             105.28      42.97   2.450   0.0145
##
## (Intercept)
## ethnicityWhite not Hispanic
## age                                          ***
## age.grp18 - 21                               *
## age.grpOver 21                               ***
## ethnicityWhite not Hispanic:age.grp18 - 21
## ethnicityWhite not Hispanic:age.grpOver 21
## age:age.grp18 - 21
## age:age.grpOver 21                           *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3813 on 768 degrees of freedom
## Multiple R-squared:  0.9625, Adjusted R-squared:  0.9621
## F-statistic:  2462 on 8 and 768 DF,  p-value: < 2.2e-16
```

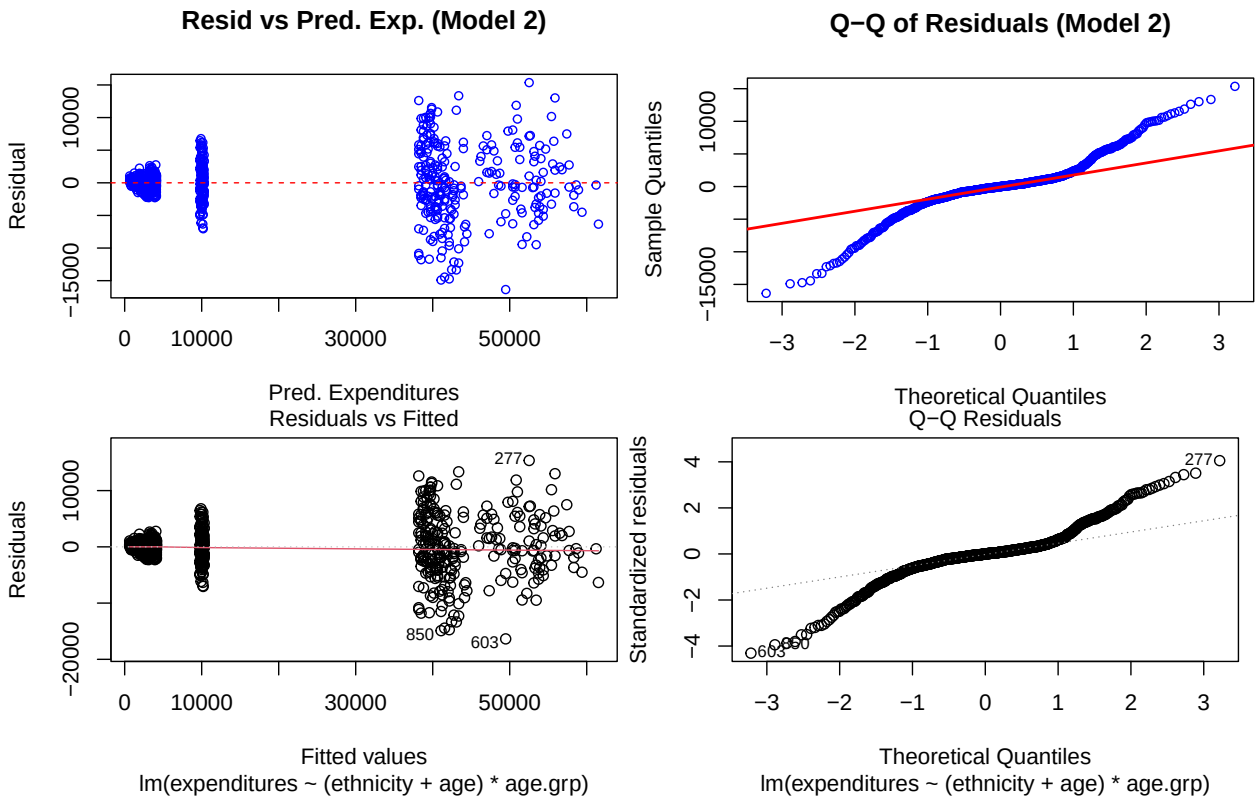
On average, an individual who is White non-Hispanic receives $70 less in expenditures than a Hispanic individual in the youngest age group, assuming age is held constant. This estimate becomes -70+240 = 170 (positive) in the middle age group and -70.4-994.5 = -1065 in the older age group.

The coefficient for age (199.8) represents the mean increase in expenditures per one year increase in age in the youngest cohort, assuming ethnicity is held constant. This estimate becomes 199.8-320.86 = -121 in the middle age group, and 199.8+105.3 = 305.1 in the older age group.

The coefficient for each age group represents the difference in the intercept of the line of age's relationship with expenditures for individuals in that age group compared to the youngest age group, assuming that age and ethnicity is held constant, but is not interpretable for the two age groups since an age of zero is not in those age groups.

c) Check the associated residual plots. What are some potential issues with the model fit in the previous subpart?

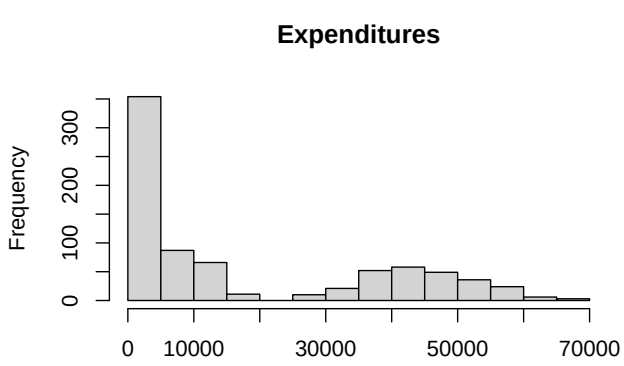
```r
plot(resid(model2) ~ fitted(model2),
     main = "Resid vs Pred. Exp. (Model 2)",
     pch = 21, col = "blue",
     cex = 0.8,
     xlab = "Pred. Expenditures",
     ylab = "Residual")
abline(h = 0, col = "red", lty = 2)

qqnorm(resid(model2),
       pch = 21, col = "blue", cex = 0.8,
       main = "Q-Q of Residuals (Model 2)")
qqline(resid(model2),
       col = "red", lwd = 2)



#I prefer to just make them like this:
plot(model2, which=1)
plot(model2, which=2)
```

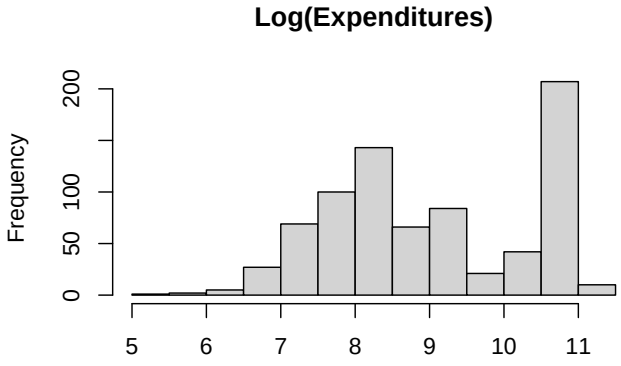


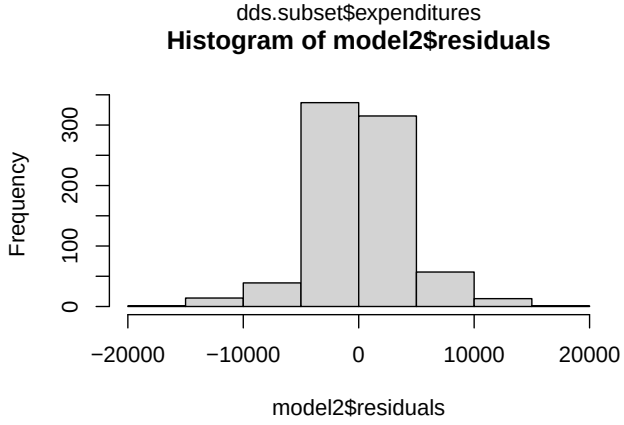
The most serious issue with the model fit is the non-constant variance between groups; it is highest in the individuals with the highest predicted expenditures (the over 21 years group) and lowest in the individuals with the lowest predicted expenditures (the under 18 years group). The residuals show departures from normality in both tails.

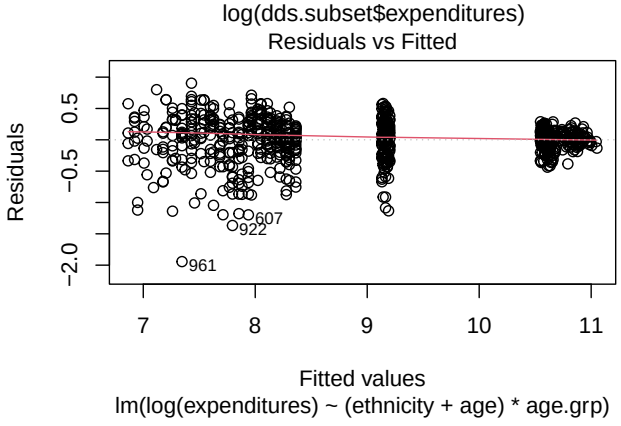
d) Do you think applying a log transformation to expenditures might address the observed issues from the model in the previous part? Try it and look at the residual plots. Does it seem preferable to continue with this model or return to the previous model? Explain your answer.

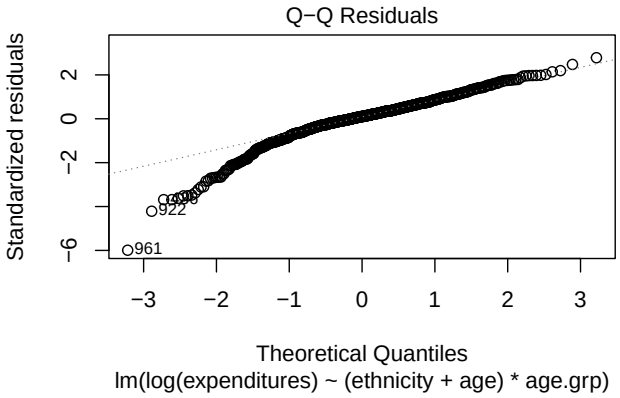
```
#some people like to look at the histograms of the dependent variable:
hist(dds.subset$expenditures, main = "Expenditures")
hist(log(dds.subset$expenditures), main = "Log(Expenditures)")

#I prefer to look at the model residuals (technically, you did this already in the qqplot abou
hist(model2$residuals)


model3 <- lm(log(expenditures) ~ (ethnicity + age)* age.grp, data=dds.subset)

plot(model3, which=1)
plot(model3, which=2)
```











The residuals are symmetric (though non-normal) so I wouldn't think a log() transformation would do any good. Sure enough, when we fit the model, the residuals are no longer symmetric. The previous model is better (it's also more interpretable since the dependent variable is untransformed)

e) Formally test whether ethnicity is an important predictor in `model`.

```
model.noEthnicity <- lm(expenditures ~ age* age.grp, data=dds.subset)
anova(model.noEthnicity, model2)
```

```
## Analysis of Variance Table
##
## Model 1: expenditures ~ age * age.grp
## Model 2: expenditures ~ (ethnicity + age) * age.grp
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1    771 1.1219e+10
## 2    768 1.1165e+10  3  53724980 1.2318 0.2971
```

The ESS F-test is not significant; controlling for age, ethnicity is not a significant predictor of expenditures. But we should probably pursue a bootstrap approach to account for the non-constant variance.

Problem 3: Building a Best Model Now that out inferential modeling is done, let's see if we can improve predictions based on the complete set of predictors.

Below we split into into train and test for you before performing prediction modeling techniques (`n.train = 600`).

```
set.seed(139); n = nrow(dds.subset); n.train = 600
rows.train = sample(1:n,n.train,replace=F)
dds.train = dds.subset[rows.train,]
dds.test = dds.subset[-rows.train,]
dim(dds.train); dim(dds.test)
```

```
## [1] 600   7
```

```
## [1] 177   7
```

a) For this problem, let's first build a model including the variables `age.cohort + age + gender + ethnicity` as a main effects only model.

```
# create model.main

model.main <- lm(expenditures ~ age.cohort + age + gender + ethnicity, data=dds.train)
```

b) Now create a model including the variables `age.cohort + age + gender + ethnicity` as main effect and the interactions between them all.

```
# create model.interact
model.interact = lm(expenditures ~ (age.cohort + age + gender + ethnicity)^2, data=dds.train)
```

c) Now build a stepwise (combined directions) sequential model starting from the `model.main` and considering a lower bound of the intercept only model and the upper bound of `model.interact`.

```
# create model.step, the next line should get you started, plus in for ___
model.step = step(model.main,scope=c(lower=formula(expenditures~1),
                                     upper=model.interact),trace=0)
```

d) Compare the 3 models above using 5 different metrics: R^2 in train, R^2 in test, adjusted-R^2, AIC, and the ESS F-test (only R^2 should be considered in the test set). Which model wins in each case?

```
# calculate the 5 metrics above for each of the 3 models
# (ESS F-test should only be calculated twice)

r.sq = function(y,yhat){
  SST = sum((y-mean(y))^2)
  SSE = sum((y-yhat)^2)
  return(1-SSE/SST)
}


rsq.train = c(r.sq(dds.train$expenditures,predict(model.main,new=dds.train)),
        r.sq(dds.train$expenditures,predict(model.interact,new=dds.train)),
        r.sq(dds.train$expenditures,predict(model.step,new=dds.train)))
```

```
rsq.test = c(r.sq(dds.test$expenditures,predict(model.main,new=dds.test)),
        r.sq(dds.test$expenditures,predict(model.interact,new=dds.test)),
        r.sq(dds.test$expenditures,predict(model.step,new=dds.test)))

adj.rsq = c(summary(model.main)$adj.r.squared,
            summary(model.interact)$adj.r.squared,
            summary(model.step)$adj.r.squared)

aic = AIC(model.main,model.interact,model.step)$AIC


data.frame(rsq.train=rsq.train,
           rsq.test=rsq.test,
           adj.rsq=adj.rsq,
           aic=aic,
           row.names=c("model.main","model.interact","model.step"))
```

```
##                rsq.train  rsq.test   adj.rsq      aic
## model.main     0.9620075 0.9703681 0.9614932 11622.85
## model.interact 0.9655020 0.9678946 0.9639366 11600.95
## model.step     0.9618946 0.9704609 0.9614440 11622.63
```

```
#ANOVA comparing step to main (step significantly better):
anova(model.step,model.main)["Pr(>F)"][[1]][2]
```

```
## [1] 0.1856003
```

```
#ANOVA comparing interact to step (interact is no better)
anova(model.interact,model.step)["Pr(>F)"][[1]][2]
```

```
## [1] 8.508738e-06
```

'model.step' is favored by the results on the test set, while `model.interact` is favored by the F-test and, adjusted R^2 and AIC. R^2 on the train set should be ignored (it will always be higher for larger models) since there are better measures considered here.

e) Perform a 'leave p out' cross-validation (keeping 500 in each "train" set) to compare the 3 models in this problem. Which model wins out using MSE as the error metric in the validation sets? Was this expected based on the previous part?

```
# this is a helper function for you.
# be careful using this in the presence of missingness (there is not here)
MSE = function(model,newdata,y){
   yhat=predict(model,newdata=newdata)
   MSE = sum((y-yhat)^2)/nrow(newdata)
   return(MSE)
}

set.seed(13939); nsims=100; n.train=nrow(dds.train);
mse.main=mse.interact=mse.step=rep(NA,nsims)
```

```
for(i in 1:nsims){
  # sample data into train-val splits
  ids = sample(1:n.train,500,replace=F)
  train = dds.train[ids,]
  val =  dds.train[-ids,]

  # fit the 3 models: 1 is given below
  fit.main=lm(formula(model.main),data=train)
  fit.interact=lm(formula(model.interact),data=train)
  fit.step=lm(formula(model.step),data=train)

  # evaluate the 3 models: the MSE function above is probably helpful
  mse.main[i] = MSE(fit.main,val,val$expenditures)
  mse.interact[i] = MSE(fit.interact,val,val$expenditures)
  mse.step[i] = MSE(fit.step,val,val$expenditures)
}

# compare results
mean(mse.main)
```

```
## [1] 15630629
```

```
mean(mse.interact)
```

```
## [1] 15205044
```

```
mean(mse.step)
```

```
## [1] 15613203
```

The mean squared error is smaller using the interaction model than the other two, which have similar performance. In this example, the step approach did not actually find a better model than using the full interaction model, which is likely just due to never adding some of the main effects involved that were necessary to consider the interaction terms

e) What challenges/issues may arise if cross-validation was used on the entire dds data set? How could these be handled in order to not "throw away" data?

Adding in the other ethnic groups would add an extra 223 observations. The issue is that many of the groups have just a few individuals (only American Indians, 3 Native Hawaiians, and 2 Other ethnicities) which could pose problems if they show up in the validation sets and not in the training sets. A simple solution would be to "collapse" these smaller groups into an "other" group, or stratify sample based on ethnicity, or throw away a train-validation split (and re-split) if one of the ethnic groups is missing from the train set.

```
dim(dds)
```

```
## [1] 1000    6
```

```
dim(dds.subset)
```

```
## [1] 777   7
```

```r
table(dds$ethnicity)
```

```
##
##    American Indian              Asian              Black           Hispanic
##                  4                129                 59                376
##         Multi Race    Native Hawaiian              Other White not Hispanic
##                 26                  3                  2                401
```