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")

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

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

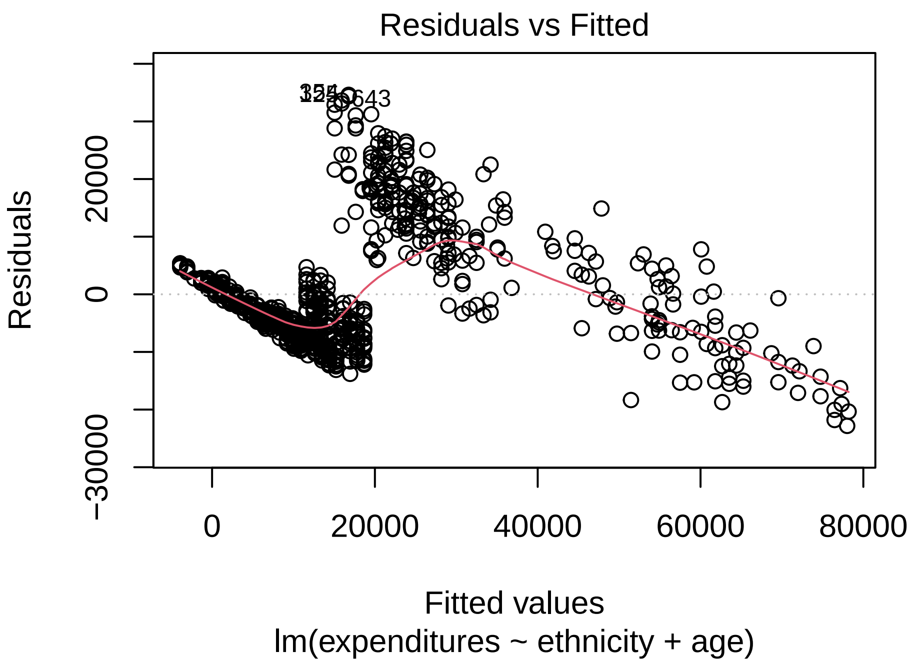
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.1     v purrr   0.3.4
## v tibble  3.2.1     v dplyr   1.1.1
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.3     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

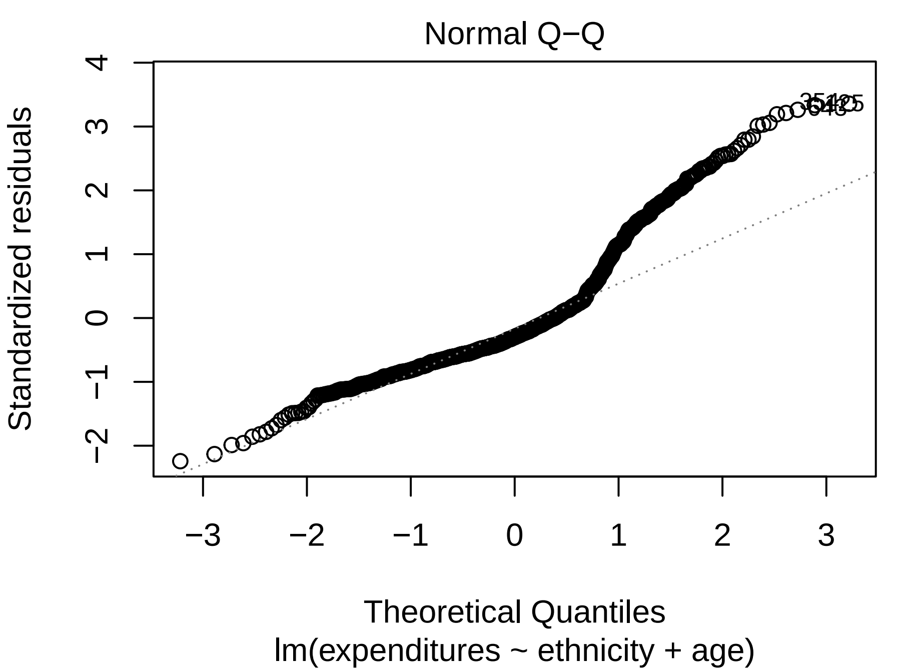
a) Fit a multiple regression model predicting expenditures from ethnicity and age. Interpret the ethnicity coefficient and investigate the model assumptions with residual plots.

```
mod1 <- lm(expenditures~ethnicity+age,dds.subset)
summary(mod1)
```

```
##
## 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
```

```r
#look at residuals versus fitted values
plot(mod1,which=c(1,2))
```





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.

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?

```
mod2 <- lm(expenditures~ethnicity+age,dds.subset[dds.subset$age<18,])
summary(mod2)$coef
```

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

```
mod3 <- lm(expenditures~ethnicity+age,dds.subset[dds.subset$age>=18 & dds.subset$age<=21,])
summary(mod3)$coef
```

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

```
mod4 <- lm(expenditures~ethnicity+age,dds.subset[dds.subset$age>21,])
summary(mod4)$coef
```

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

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

```
#summary output
```

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
dds.subset <- dds.subset %>%
  mutate(age.grp = case_when(age < 18 ~ "under 18",
                             age >=18 & age <= 21 ~ "18 to 21",
                             age > 21 ~ "over 21"))
```

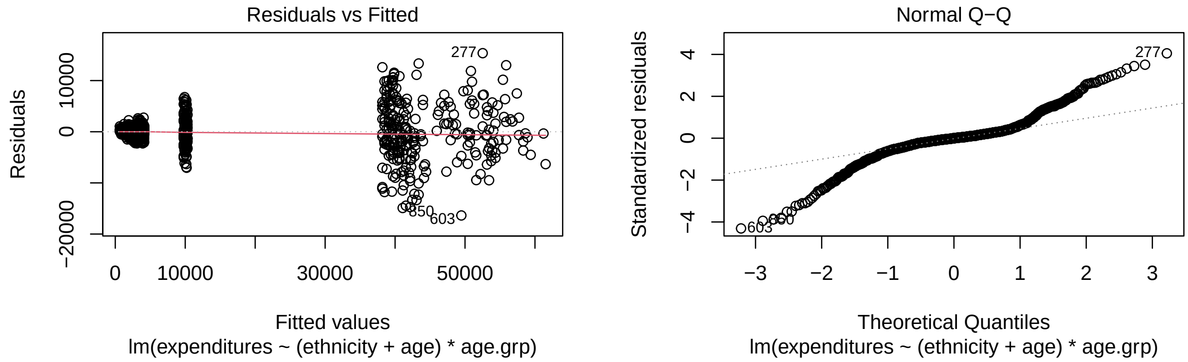
b) Fit a model predicting expenditures from ethnicity, age, and age group. Interpret the model coefficients.

```r
mod5 <- lm(expenditures~(ethnicity+age)*age.grp, dds.subset)
summary(mod5)
```

```
##
## 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
## (Intercept)                               12327.8     5634.0   2.188
## ethnicityWhite not Hispanic                 169.8      630.2   0.270
## age                                        -121.0      287.1  -0.422
## age.grpover 21                            20210.7     5676.1   3.561
## age.grpunder 18                          -11607.8     5656.6  -2.052
## ethnicityWhite not Hispanic:age.grpover 21  -1234.7      844.2  -1.463
## ethnicityWhite not Hispanic:age.grpunder 18  -240.2      753.6  -0.319
## age:age.grpover 21                          426.1      287.4   1.483
## age:age.grpunder 18                         320.9      290.1   1.106
##                                          Pr(>|t|)
## (Intercept)                              0.028961 *
## ethnicityWhite not Hispanic              0.787615
## age                                      0.673478
## age.grpover 21                           0.000393 ***
## age.grpunder 18                          0.040500 *
## ethnicityWhite not Hispanic:age.grpover 21  0.143993
## ethnicityWhite not Hispanic:age.grpunder 18 0.750011
## age:age.grpover 21                       0.138517
## age:age.grpunder 18                      0.269039
## ---
## 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
```

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

```
plot(mod5, which=c(1,2))
```



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.

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

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.

```r
# create model.main
model.main <- lm(expenditures~age.cohort + age + gender + ethnicity, dds.train)
summary(model.main)
```

```
##
## Call:
## lm(formula = expenditures ~ age.cohort + age + gender + ethnicity,
##     data = dds.train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17452.5  -1369.1    -12.3   1300.8  15673.8
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1573.17     612.59   2.568 0.010472 *
## age.cohort13-17                603.04     771.62   0.782 0.434806
## age.cohort18-21               6143.94     878.84   6.991 7.40e-12 ***
## age.cohort22-50              34876.54    1111.04  31.391  < 2e-16 ***
## age.cohort51+                40865.64    2344.15  17.433  < 2e-16 ***
## age.cohort6-12                -212.14     708.55  -0.299 0.764743
## age                            164.77      33.66   4.895 1.27e-06 ***
## genderMale                   -1147.99     315.56  -3.638 0.000299 ***
## ethnicityWhite not Hispanic  -449.63     339.28  -1.325 0.185600
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3857 on 591 degrees of freedom
## Multiple R-squared:  0.962,  Adjusted R-squared:  0.9615
## F-statistic:  1871 on 8 and 591 DF,  p-value: < 2.2e-16
```

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

```r
# create model.interact
model.interact <- lm(expenditures~(age.cohort + age + gender + ethnicity)^2, dds.train)
summary(model.interact)
```

```
##
## Call:
## lm(formula = expenditures ~ (age.cohort + age + gender + ethnicity)^2,
##     data = dds.train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13175.6  -1160.5     19.4   1163.2  17330.8
```

```
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                   1491.34    1472.41   1.013
## age.cohort13-17                               5777.28    3900.55   1.481
## age.cohort18-21                              16151.83    6755.85   2.391
## age.cohort22-50                              47880.80    2976.54  16.086
## age.cohort51+                                52417.60    5708.67   9.182
## age.cohort6-12                                1904.10    2244.44   0.848
## age                                            -27.29     332.56  -0.082
## genderMale                                    -687.89    1148.34  -0.599
## ethnicityWhite not Hispanic                   -711.72    1262.49  -0.564
## age.cohort13-17:age                           -183.00     401.28  -0.456
## age.cohort18-21:age                           -344.49     469.12  -0.734
## age.cohort22-50:age                           -213.74     336.47  -0.635
## age.cohort51+:age                               78.70     333.76   0.236
## age.cohort6-12:age                             -80.74     375.12  -0.215
## age.cohort13-17:genderMale                   -2544.55    1513.47  -1.681
## age.cohort18-21:genderMale                   -3822.19    1727.43  -2.213
## age.cohort22-50:genderMale                   -8754.65    2196.48  -3.986
## age.cohort51+:genderMale                    -16667.55    4684.11  -3.558
## age.cohort6-12:genderMale                    -1516.94    1381.32  -1.098
## age.cohort13-17:ethnicityWhite not Hispanic  -2266.86    1636.03  -1.386
## age.cohort18-21:ethnicityWhite not Hispanic  -2802.88    1861.32  -1.506
## age.cohort22-50:ethnicityWhite not Hispanic  -4736.92    2376.64  -1.993
## age.cohort51+:ethnicityWhite not Hispanic   -15089.85    5228.54  -2.886
## age.cohort6-12:ethnicityWhite not Hispanic   -1230.41    1497.09  -0.822
## age:genderMale                                 200.64      67.32   2.980
## age:ethnicityWhite not Hispanic                174.71      73.78   2.368
## genderMale:ethnicityWhite not Hispanic         354.71     668.37   0.531
##                                             Pr(>|t|)
## (Intercept)                                 0.311557
## age.cohort13-17                             0.139117
## age.cohort18-21                             0.017134 *
## age.cohort22-50                              < 2e-16 ***
## age.cohort51+                                < 2e-16 ***
## age.cohort6-12                              0.396590
## age                                         0.934628
## genderMale                                  0.549390
## ethnicityWhite not Hispanic                 0.573149
## age.cohort13-17:age                         0.648534
## age.cohort18-21:age                         0.463052
## age.cohort22-50:age                         0.525523
## age.cohort51+:age                           0.813666
## age.cohort6-12:age                          0.829649
## age.cohort13-17:genderMale                  0.093257 .
## age.cohort18-21:genderMale                  0.027316 *
## age.cohort22-50:genderMale                   7.6e-05 ***
```

```
## age.cohort51+:genderMale                       0.000404 ***
## age.cohort6-12:genderMale                      0.272586
## age.cohort13-17:ethnicityWhite not Hispanic 0.166413
## age.cohort18-21:ethnicityWhite not Hispanic 0.132655
## age.cohort22-50:ethnicityWhite not Hispanic 0.046722 *
## age.cohort51+:ethnicityWhite not Hispanic   0.004048 **
## age.cohort6-12:ethnicityWhite not Hispanic  0.411494
## age:genderMale                                 0.003002 **
## age:ethnicityWhite not Hispanic                0.018212 *
## genderMale:ethnicityWhite not Hispanic      0.595830
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3733 on 573 degrees of freedom
## Multiple R-squared:  0.9655, Adjusted R-squared:  0.9639
## F-statistic: 616.8 on 26 and 573 DF,  p-value: < 2.2e-16
```

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`.

```r
# 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),
                direction="both",
                trace=2,
                k=log(n)) # BIC criterion
```

```
## Start:  AIC=9960.02
## expenditures ~ age.cohort + age + gender + ethnicity
##
##              Df  Sum of Sq        RSS     AIC
## - ethnicity   1 2.6128e+07 8.8183e+09  9955.1
## <none>                     8.7922e+09  9960.0
## - gender      1 1.9688e+08 8.9891e+09  9966.7
## - age         1 3.5639e+08 9.1486e+09  9977.2
## - age.cohort  5 5.5760e+10 6.4552e+10 11122.9
##
## Step:  AIC=9955.14
## expenditures ~ age.cohort + age + gender
##
##              Df  Sum of Sq        RSS     AIC
## <none>                     8.8183e+09  9955.1
## - gender      1 1.9413e+08 9.0125e+09  9961.6
## - age         1 3.5729e+08 9.1756e+09  9972.3
## - age.cohort  5 5.8563e+10 6.7381e+10 11142.0
```

```r
formula(model.step)
```

```
## expenditures ~ age.cohort + age + gender
```

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?

```r
# 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)
}
```

'model.step' is favored by the results on the test set and the F test. 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?

```r
# 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)
}
```

f) 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?

```r
dim(dds)
```

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

```r
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
```