STA 563 Project 2

OCCUPATIONAL PRESTIGE: PREDICTIVE MODELING

Submitted by:

Alisha Rajbhandari

Erika Rasnick

December 11, 2017

Introduction

Occupational prestige can be defined as the admiration and respect that a particular occupation holds in society.[1] In the mid-1960s, the occupations in Canada were scored on a Pineo-Porter prestige scale based on results from a social survey. In 1971, these prestige scores for each occupation were matched up with information from the Canadian Census.[2] The occupational prestige dataset consists of observations for 102 occupations and the following variables:

	Variables	Type	Description
Response (Y)	prestige	numeric	Pineo-Porter prestige score on a scale from 0 to 100
Predictor (X1)	education	numeric	Average education of incumbents in 1971 (years)
Predictor (X2)	income	numeric	Average income of incumbents in 1971 (dollars)
Predictor (X3)	women	numeric	Percentage of incumbents who are women
Predictor (X4)	type	categorical	Type of occupation with three levels: • "prof" - refers to professional, managerial, and technical occupations • "wc" - refers to white collar occupations • "bc" - refers to blue collar occupations

Table 1: Variables and their description

We would like to see how well can we predict the prestige points for different occupations. Our goal is to develop a predictive model by using the best subsets method and 10-fold cross validation to select the best predictors from these variables as well as any important 2-factor interactions between these variables. We will also use this model to predict the prestige scores for four occupations that are not used to develop the model: athletes, newsboys, babysitters, and farmers.

Exploratory Data Analysis

The scatter plot matrix of the variables (Figure 1) reveals that there appears to be a linear relationship between prestige and education, and possibly a logarithmic relationship between prestige and income. There does not seem to be an obvious relationship between the percentage of women in the occupation and prestige, but there does seem to be a correlation between the type of occupation and prestige. It also appears that there is a logarithmic relationship between education and income, and a possible non-linear relationship between income and percentage of women. Type of occupation looks like it may be correlated with all the other predictors except percentage of women.

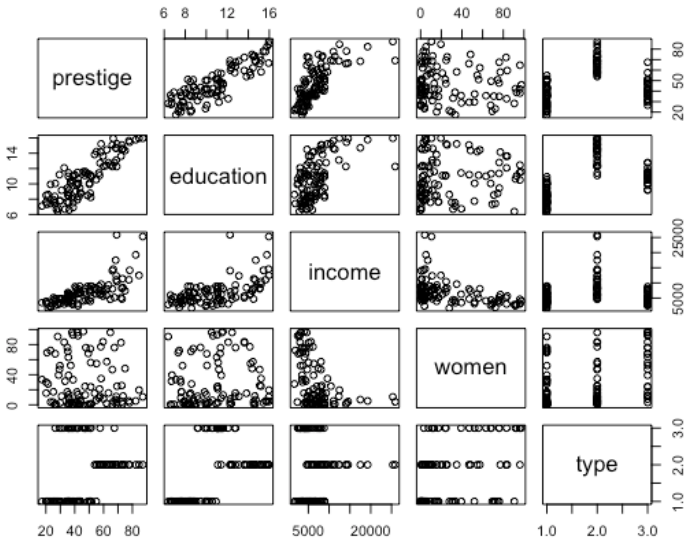

Figure 1. Scatter plot matrix of variables

Looking at interaction plots, there does not seem to be an obvious interaction between occupation type and education (Figure 3) or occupation type and percentage of women (Figure 4). However, there appears to be a distinctly different relationship between income and prestige depending on the type of occupation (Figure 2). This leads us to believe that the interaction between income and occupation type should be included in the model.

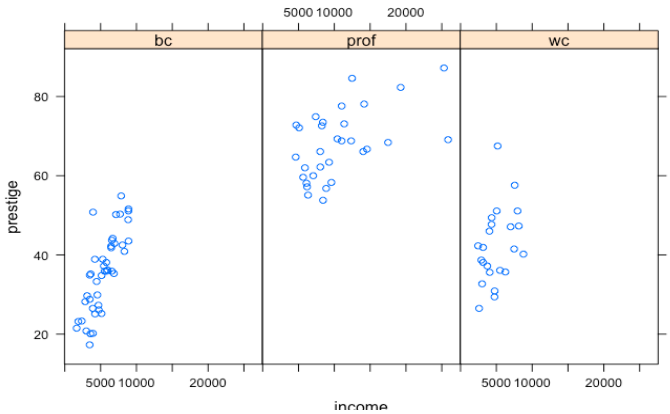

Figure 2. Interaction plot for Income and Occupation Type

Analysis and Model Selection

Since we have five predictors (education, income, women, type-prof, and type-wc) and 9 possible 2-factor interactions, we will have 2^14 = 16,384 possible models. Evaluating all the predictors might not be a best approach. Hence, we will use a variable selection procedure to identify a small group of regression models that performs well according to some specified criterion. For this project, we will consider the R-squared, the adjusted R-squared, Mallow's Cp, and BIC. We are looking for higher R-squared and adjusted R-squared and lower Mallow's Cp and BIC.

There are several automatic search procedures for model selection that will simplify our task of selecting the best set of predictors. This paper will discuss the best subset method for variable selection and the k-fold cross validation method for model validation. The best subset method is often helpful because it identifies the best subset in terms of R-squared for each possible number of predictors in the model, which makes the task of selecting the final subset of predictors easier.

We first fit a full model with all the interaction terms to find a best subset of predictors for prestige. Based on Cp and BIC, we found that the models with 7 predictors and 5 predictors are the best models to predict prestige (Appendix, Figure 5). The predictors for the 7-predictor model were income, typeprof, education:income, education:women, income:typeprof, income:typewc, and women:typeprof . The predictors for the 5-predictor model were income, education:women, education:typeprof, income:typeprof, and women:typeprof (Appendix, Figure 6). To check the validity of the models, we chose a 10-fold cross validation method and found the root mean squared error (RMSE) for both the models to be around 6, which seems relatively low Appendix, Figure 11). We also used these multiple linear regression models with 5 predictors and 7 predictors to predict the mean prestige for 4 new observations from the out-of-sample data, namely, athletes, babysitters, newsboys, and farmers. The predictions were relatively accurate compared to the known values of prestige for these occupations. However, we wanted to see if we could find a simpler model that could better explain prestige.

After looking at the interaction plots between the predictors, we found that there is an interaction between income and type but there is no significant interaction with other predictors. For the reduced model, we kept the main effects and the interaction for education:income, education:women, income:women, and income:type. We selected the best subset for the reduced model and ran the analysis again. We found that the 5-predictor model is favorable according to BIC and 6-predictor model is favorable according to Mallow's Cp (Appendix, Figure 7). The predictors for the 6-predictor model are education, income, women, typeprof, income:typeprof, and income:typewc. The predictors for the 5-predictor model are education, income, women, typeprof, and income:typeprof (Appendix, Figure 8). The 5-predictor model seems relatively simpler than the

6-predictor model. We think this could be the best model. To further support our idea, we perform a 10-fold cross validation again and found that there is no significant difference in the RMSE, hence we choose the 5-predictor model.

We also ran this analysis for the model with only the main effects terms to serve as a comparison to our other models. Mallow's Cp and BIC both favor a 3-predictor model with education, income, and typeprof (Appendix, Figures 9 and 10). The RMSE was higher for the main effects only model, so we rejected this model.

We came to a conclusion that the model with the predictors education, income, women, typeprof, and income:typeprof is the best model because of its relatively simple structure and high predictive power. Now, we perform a multiple linear analysis on our model which showed us that all the predictors are now significant. Our final model is

$$\widehat{\text{Prestige}} = 2.4856\ \text{education} + 0.0036\ \text{income} + 0.0744\ \text{women} + 28.1102\ \text{typeprof} - 0.0027\ \text{income} * \text{typeprof}$$

We used our final model to predict prestige for four occupations that were not used to build the models because they had missing values for occupation type. To make the predictions, we used our best judgement to assign a type for these occupations. We decided that athletes, newsboys, and babysitters were not professionals, but since farmers were most likely considered very important to society in the 1970s, we called them professionals.

Occupation	**TypeProf**	**Prestige Actual Value**	**Prestige Predicted (final)**	**Prestige Predicted (main effects)**
Athletes	FALSE	54.1	52.05	56.88
Newsboys	FALSE	14.8	21.42	34.29
Babysitters	FALSE	25.9	26.59	33.45
Farmers	TRUE	44.1	42.24	37.31

Table 2: Predicted values of prestige for final model and main effects model

From the table above, we can see that our final model predicted prestige for these occupations better than the main effects model. The main effects model predicted too high for athletes, newsboys, and babysitters, and predicted too low for farmers. The values predicted by our final model are within 6 prestige points of the actual values, which agrees with our value for RMSE for this model. This supports the idea that the model with only three predictors is too simple and our final chosen model is better for prediction. Further, the predicted values vs. prestige plot shows a strong linear trend which means that the prestige is well predicted by our model (Appendix, Figure 12(a)). The residual plot also shows no major outliers since all the residuals are within 3 standard deviations (Appendix, Figure 12(b)).

Conclusions

Our best model tells us that prestige is best predicted by the amount of education required for the occupation (years), the average income of the occupation (dollars), the percentage of women in the occupation, whether the occupation is considered professional or not, and the interaction between the average income and whether the occupation is professional or not. The R-squared value for this model is relatively high at 0.8706 which means that around 87.06% of the variability in prestige can be explained by its linear relationship with our chosen predictors. This model also makes relatively accurate predictions for the four out-of-sample occupations. However, we do not believe this model should be used to predict prestige for occupations in the present day, because not only has income become inflated, but the public view of the prestige of a given occupation can change with current and historical events and fluctuation in cultural views. Therefore, different variables may be more important for predicting prestige today.

References

1. Occupational prestige. (2017, January 12). In *Wikipedia, The Free Encyclopedia*. Retrieved 19:58, December 8, 2017, from https://en.wikipedia.org/w/index.php?title=Occupational_prestige&oldid=75966334

2. Canada (1971) Census of Canada. Vol. 3, Part 6. Statistics Canada [pp. 19-1–19-21]. Personal communication from B. Blishen, W. Carroll, and C. Moore, Departments of Sociology, York University and University of Victoria.

Appendix

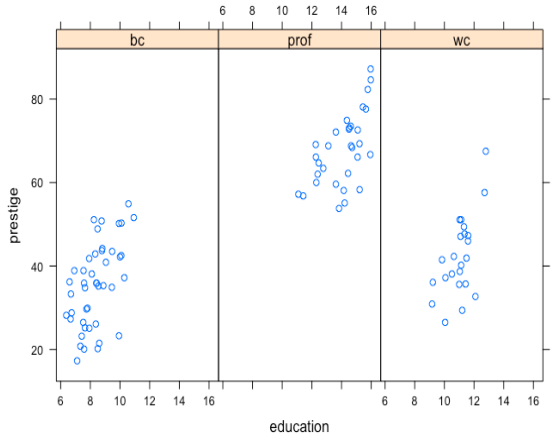

Figure 3. Interaction plot for Education and Occupation Type

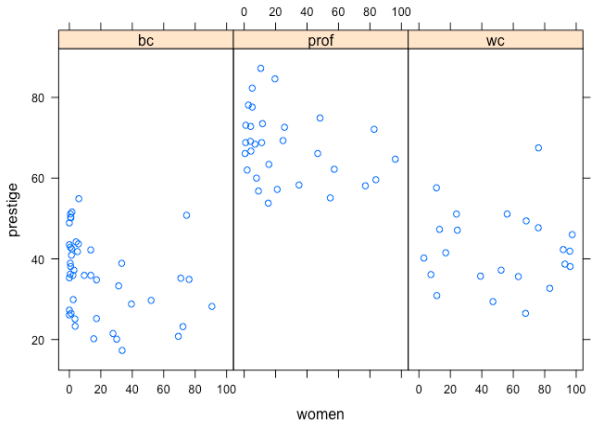

Figure 4. Interaction plot for % Women and Occupation Type

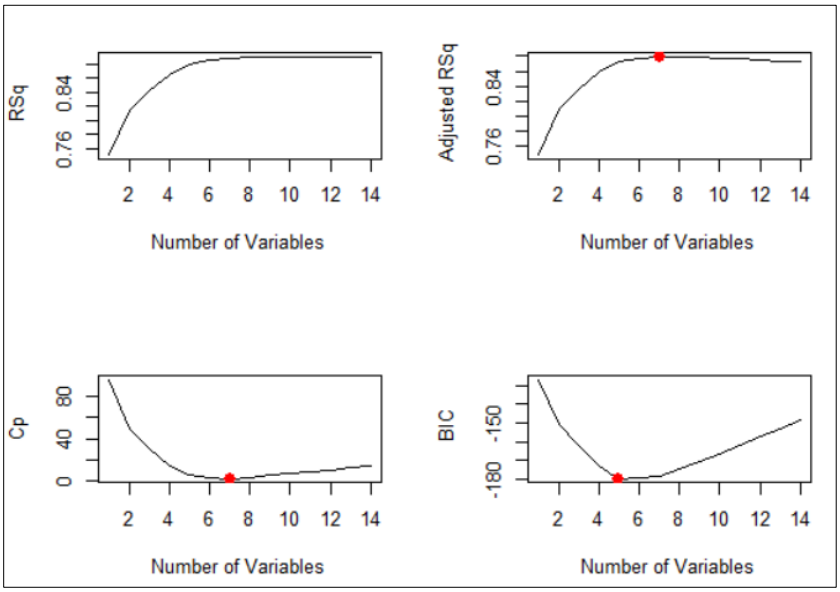

Figure 5. Model Selection Plots for Full Model

```
> coef(prestige.full,w.max.2full)
     (Intercept)           income         typeprof education:income
    6.4835610051     0.0029302436    55.0846743806     0.0002261590
 education:women  income:typeprof    income:typewc   women:typeprof
    0.0193146836    -0.0055222600    -0.0006145087    -0.2873657371
> coef(prestige.full,w.max.3full)
       (Intercept)            income   education:women education:typeprof
       5.404667623       0.005063570       0.019452740       4.020789781
   income:typeprof     women:typeprof
      -0.004507172      -0.280126731
```

Figure 6. Coefficients for Full Model

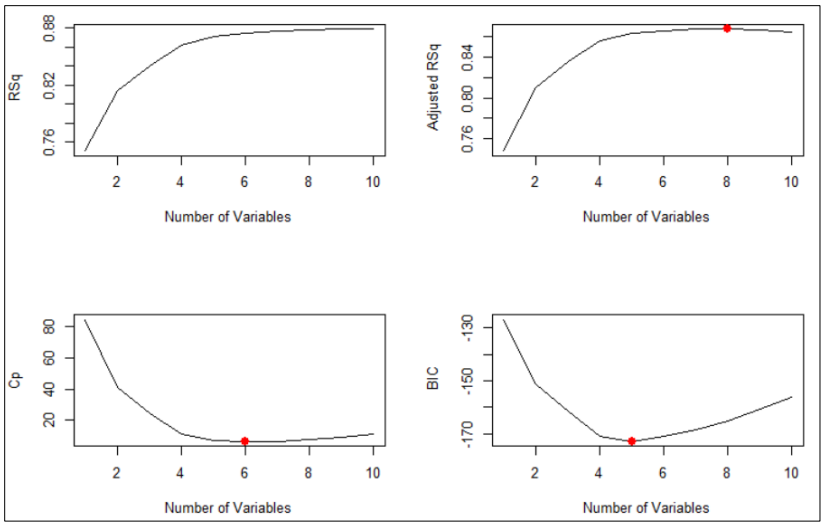

Figure 7. Model Selection Plots for Reduced Model

```
> coef(prestige.reduced,w.max.2red)
    (Intercept)       education          income           women        typeprof
  -1.050577e+01    2.945023e+00    3.714884e-03    8.209505e-02    2.579325e+01
income:typeprof   income:typewc
  -2.863678e-03   -5.978179e-04
> coef(prestige.reduced,w.max.3red)
    (Intercept)       education          income           women        typeprof
   -6.288937518     2.485553791     0.003571092     0.074380217    28.110187585
income:typeprof
   -0.002707169
```

Figure 8. Coefficients for Reduced Model

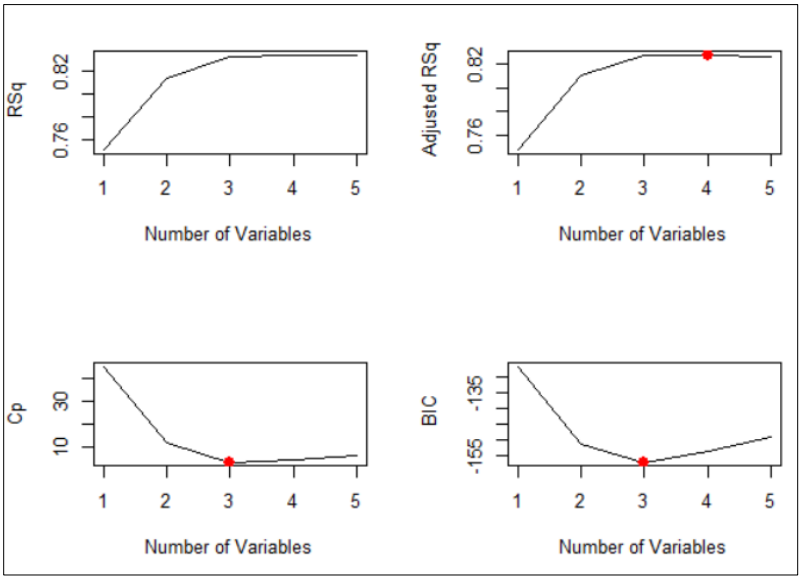

Figure 9. Model Selection Plots for Main Effects Model

```
> coef(prestige.me,w.max.2me)
(Intercept)   education      income    typeprof
2.581473970 3.194059988 0.001069967 8.982891659
> coef(prestige.me,w.max.3me)
(Intercept)   education      income    typeprof
2.581473970 3.194059988 0.001069967 8.982891659
```

Figure 10. Coefficients for Main Effects Model

```
> sqrt(full7.err$delta[1])
[1] 6.191222
> sqrt(full5.err$delta[1])
[1] 6.363635
> sqrt(red6.err$delta[1])
[1] 6.388648
> sqrt(red5.err$delta[1])
[1] 6.609954
> sqrt(me3.err$delta[1])
[1] 7.325393
```

Figure 11. RMSE for different sets of predictors model

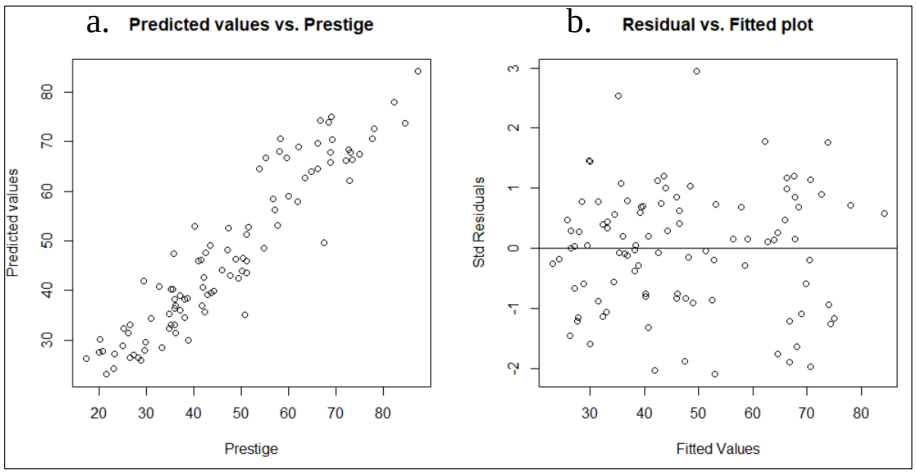

Figure 12. (a) Predicted vs. Prestige, (b) Residual Plot

Relevant Codes

```
library(GGally)
library(lattice)
library(leaps)
library(car)
library(boot)

PrestigeData <-
read.table("/Users/erikarasnick/Documents/STA/STA563/prestige.txt",
header=TRUE)
PrestigeData <- na.omit(PrestigeData)

######### EDA
# scatter plot matrix
pairs(~prestige+education+income+women+type, data=PrestigeData)

# interaction plots
xyplot(prestige~education|type,data=PrestigeData)
xyplot(prestige~income|type,data=PrestigeData)
xyplot(prestige~women|type,data=PrestigeData)


######### Model Selction by All Subsets

prestige.full <-regsubsets(prestige~education+income+women+type
                           +education:income+education:women
                           +education:type+income:women+income:type
                           +women:type, data=PrestigeData, nvmax=20)
full.summary <- summary(prestige.full)

prestige.reduced <-regsubsets(prestige~education+income+women+type
                           +education:income+education:women+income:women
                           +income:type, data=PrestigeData, nvmax=20)
reduced.summary <- summary(prestige.reduced)

prestige.me <-regsubsets(prestige~education+income+women+type,
                         data=PrestigeData, nvmax=20)
me.summary <- summary(prestige.me)

# evaluation of models

# full model
par(mfrow=c(2,2))
plot(full.summary$rsq ,xlab=" Number of Variables ",ylab=" RSq",
     type="l")
plot(full.summary$adjr2 ,xlab =" Number of Variables ",
     ylab=" Adjusted RSq",type="l")
w.max.1full=which.max(full.summary$adjr2)
points(w.max.1full, full.summary$adjr2[w.max.1full], col ="red",cex=2,pch=20)
plot(full.summary$cp,xlab =" Number of Variables ",ylab="Cp",type="l")
w.max.2full=which.min(full.summary$cp)
points(w.max.2full,full.summary$cp[w.max.2full],col="red",cex=2,pch=20)
w.max.3full=which.min (full.summary$bic)
plot(full.summary$bic,xlab=" Number of Variables ",ylab=" BIC",
     type="l")
points(w.max.3full,full.summary$bic[w.max.3full],col="red",cex=2,pch=20)

coef(prestige.full,w.max.2full)
coef(prestige.full,w.max.3full)

# reduced model
par(mfrow=c(2,2))
plot(reduced.summary$rsq ,xlab=" Number of Variables ",ylab=" RSq",
```

```
     type="l")
plot(reduced.summary$adjr2 ,xlab =" Number of Variables ",
     ylab=" Adjusted RSq",type="l")
w.max.1red=which.max(reduced.summary$adjr2)
points(w.max.1red, reduced.summary$adjr2[w.max.1red], col
="red",cex=2,pch=20)
plot(reduced.summary$cp,xlab =" Number of Variables ",ylab="Cp",type="l")
w.max.2red=which.min(reduced.summary$cp)
points(w.max.2red,reduced.summary$cp[w.max.2red],col="red",cex=2,pch=20)
w.max.3red=which.min (reduced.summary$bic)
plot(reduced.summary$bic,xlab=" Number of Variables ",ylab=" BIC",
     type="l")
points(w.max.3red,reduced.summary$bic[w.max.3red],col="red",cex=2,pch=20)

coef(prestige.reduced,w.max.2red)
coef(prestige.reduced,w.max.3red)

# main effects only model
par(mfrow=c(2,2))
plot(me.summary$rsq ,xlab=" Number of Variables ",ylab=" RSq",
     type="l")
plot(me.summary$adjr2 ,xlab =" Number of Variables ",
     ylab=" Adjusted RSq",type="l")
w.max.1me=which.max(me.summary$adjr2)
points(w.max.1me, me.summary$adjr2[w.max.1me], col ="red",cex=2,pch=20)
plot(me.summary$cp,xlab =" Number of Variables ",ylab="Cp",type="l")
w.max.2me=which.min(me.summary$cp)
points(w.max.2me,me.summary$cp[w.max.2me],col="red",cex=2,pch=20)
w.max.3me=which.min (me.summary$bic)
plot(me.summary$bic,xlab=" Number of Variables ",ylab=" BIC",
     type="l")
points(w.max.3me,me.summary$bic[w.max.3me],col="red",cex=2,pch=20)

coef(prestige.me,w.max.2me)
coef(prestige.me,w.max.3me)

################ cross validation

PrestigeData$typeprof <- PrestigeData$type=="prof"
PrestigeData$typewc <- PrestigeData$type=="wc"

# full model -> Cp -> 7 predictors

prestige.full7 <-glm(prestige~income+typeprof+education:income
                     +education:women+income:typeprof
                     +income:typewc+women:typeprof,data=PrestigeData)
full7.err<-cv.glm(PrestigeData,prestige.full7, K=10)

# full model -> BIC -> 5 predictors

prestige.full5 <-glm(prestige~income+education:women+education:typeprof
                     +income:typeprof+women:typeprof,data=PrestigeData)
full5.err<-cv.glm(PrestigeData,prestige.full5, K=10)

# reduced model -> Cp -> 6 predictors

prestige.red6 <-glm(prestige~education+income+women+typeprof
                     +income:typeprof+income:typewc,data=PrestigeData)
red6.err<-cv.glm(PrestigeData,prestige.red6, K=10)

# reduced model -> BIC -> 5 predictors

prestige.red5 <-glm(prestige~education+income+women+typeprof
                     +income:typeprof,data=PrestigeData)
```

```
red5.err<-cv.glm(PrestigeData,prestige.red5, K=10)

# main effects only -> 3 predictors

prestige.me3 <-glm(prestige~education+income+typeprof,data=PrestigeData)
me3.err<-cv.glm(PrestigeData,prestige.me3, K=10)

sqrt(full7.err$delta[1])
sqrt(full5.err$delta[1])
sqrt(red6.err$delta[1])
sqrt(red5.err$delta[1])
sqrt(me3.err$delta[1])


################ predictions

# full model -> Cp -> 7 predictors
mr.full7 <- lm(prestige~income+typeprof+education:income
          +education:women+income:typeprof
          +income:typewc+women:typeprof,data=PrestigeData)
summary(mr.full7)
# athletes
new = data.frame(education = 11.44, income=8206, women=8.13, typeprof=FALSE,
typewc=FALSE)
predict(mr.full7, new, interval="confidence")
predict(mr.full7, new, interval="prediction")
# newsboys
new1 = data.frame(education = 9.62, income=918, women=7, typeprof=FALSE,
typewc=FALSE)
predict(mr.full7, new1, interval="confidence")
predict(mr.full7, new1, interval="prediction")
# babysitters
new2 = data.frame(education = 9.46, income=611, women=96.53, typeprof=FALSE,
typewc=FALSE)
predict(mr.full7, new2, interval="confidence")
predict(mr.full7, new2, interval="prediction")
# farmers
new3 = data.frame(education = 6.84, income=3643, women=3.6, typeprof=FALSE,
typewc=FALSE)
predict(mr.full7, new3, interval="confidence")
predict(mr.full7, new3, interval="prediction")


# full model -> BIC -> 5 predictors
mr.full5 <- lm(prestige~income+education:women+education:typeprof
          +income:typeprof+women:typeprof,data=PrestigeData)
summary(mr.full5)
# athletes
new = data.frame(education = 11.44, income=8206, women=8.13, typeprof=FALSE)
predict(mr.full5, new, interval="confidence")
predict(mr.full5, new, interval="prediction")
# newsboys
new1 = data.frame(education = 9.62, income=918, women=7, typeprof=FALSE)
predict(mr.full5, new1, interval="confidence")
predict(mr.full5, new1, interval="prediction")
# babysitters
new2 = data.frame(education = 9.46, income=611, women=96.53, typeprof=FALSE)
predict(mr.full5, new2, interval="confidence")
predict(mr.full5, new2, interval="prediction")
# farmers
new3 = data.frame(education = 6.84, income=3643, women=3.6, typeprof=FALSE)
predict(mr.full5, new3, interval="confidence")
predict(mr.full5, new3, interval="prediction")
```

```
# reduced model -> Cp -> 6 predictors
mr.red6 <- lm(prestige~education+income+women+typeprof
          +income:typeprof+income:typewc,data=PrestigeData)
summary(mr.red6)
# athletes
new = data.frame(education = 11.44, income=8206, women=8.13, typeprof=FALSE,
typewc=FALSE)
predict(mr.red6, new, interval="confidence")
predict(mr.red6, new, interval="prediction")
# newsboys
new1 = data.frame(education = 9.62, income=918, women=7, typeprof=FALSE,
typewc=FALSE)
predict(mr.red6, new1, interval="confidence")
predict(mr.red6, new1, interval="prediction")
# babysitters
new2 = data.frame(education = 9.46, income=611, women=96.53, typeprof=FALSE,
typewc=FALSE)
predict(mr.red6, new2, interval="confidence")
predict(mr.red6, new2, interval="prediction")
# farmers
new3 = data.frame(education = 6.84, income=3643, women=3.6, typeprof=FALSE,
typewc=FALSE)
predict(mr.red6, new3, interval="confidence")
predict(mr.red6, new3, interval="prediction")


# reduced model -> BIC -> 5 predictors
mr.red5 <- lm(prestige~education+income+women+typeprof
          +income:typeprof,data=PrestigeData)
summary(mr.red5)
# athletes
new = data.frame(education = 11.44, income=8206, women=8.13, typeprof=FALSE)
predict(mr.red5, new, interval="confidence")
predict(mr.red5, new, interval="prediction")
# newsboys
new1 = data.frame(education = 9.62, income=918, women=7, typeprof=FALSE)
predict(mr.red5, new1, interval="confidence")
predict(mr.red5, new1, interval="prediction")
# babysitters
new2 = data.frame(education = 9.46, income=611, women=96.53, typeprof=FALSE)
predict(mr.red5, new2, interval="confidence")
predict(mr.red5, new2, interval="prediction")
# farmers
new3 = data.frame(education = 6.84, income=3643, women=3.6, typeprof=FALSE)
predict(mr.red5, new3, interval="confidence")
predict(mr.red5, new3, interval="prediction")


# main effects only model -> 3 predictors
mr.me3 <- lm(prestige~education+income+typeprof,data=PrestigeData)
summary(mr.me3)
# athletes
new = data.frame(education = 11.44, income=8206, typeprof=FALSE)
predict(mr.me3, new, interval="confidence")
predict(mr.me3, new, interval="prediction")
# newsboys
new1 = data.frame(education = 9.62, income=918, typeprof=FALSE)
predict(mr.me3, new1, interval="confidence")
predict(mr.me3, new1, interval="prediction")
# babysitters
new2 = data.frame(education = 9.46, income=611, typeprof=FALSE)
predict(mr.me3, new2, interval="confidence")
predict(mr.me3, new2, interval="prediction")
```

```
# farmers
new3 = data.frame(education = 6.84, income=3643, typeprof=FALSE)
predict(mr.me3, new3, interval="confidence")
predict(mr.me3, new3, interval="prediction")

# Best model summary output and plots
mr.red5 <- lm(prestige~education+income+women+typeprof
              +income:typeprof,data=PrestigeData)
summary(mr.red5)

par(mfrow=c(1,2))
plot(PrestigeData$prestige, mr.red5$fitted.values, main="Predicted values vs.
Prestige",
     xlab="Prestige", ylab="Predicted values")
plot(mr.red5$fitted.values, rstandard(mr.red5), main = "Residual vs. Fitted
plot",
     xlab="Fitted Values", ylab="Std Residuals")
abline(h=0)
```