HW #2

Matthew Fein

February 27, 2019

## Homework #2

1. A wine’s vintage is the year in which the grapes used were grown and harvested.
2. The response variable used in this paper is the log price of the vintage relative to 1961.

require(leaps)

## Loading required package: leaps

## Warning: package 'leaps' was built under R version 3.5.2

require(DAAG)

## Loading required package: DAAG

## Warning: package 'DAAG' was built under R version 3.5.2

## Loading required package: lattice

require(tidyverse)

## Loading required package: tidyverse

## -- Attaching packages ------------------------------------------------------------------------ tidyverse 1.2.1 --

## v ggplot2 3.0.0 v purrr 0.2.5  
## v tibble 1.4.2 v dplyr 0.7.6  
## v tidyr 0.8.1 v stringr 1.3.1  
## v readr 1.1.1 v forcats 0.3.0

## -- Conflicts --------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

require(gridExtra)

## Loading required package: gridExtra

## Warning: package 'gridExtra' was built under R version 3.5.2

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

require(GGally)

## Loading required package: GGally

## Warning: package 'GGally' was built under R version 3.5.2

##   
## Attaching package: 'GGally'

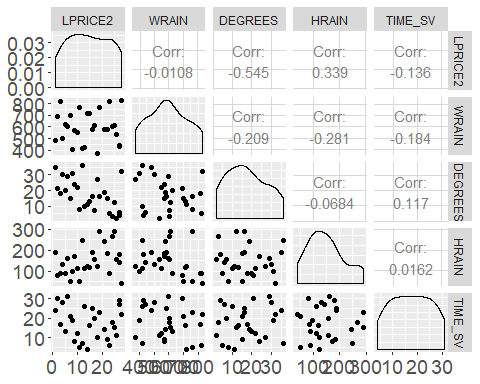
## The following object is masked from 'package:dplyr':  
##   
## nasa

theme.info <- theme(plot.title = element\_text(size=16, hjust=0.5),  
 axis.title = element\_text(size=14),  
 axis.text = element\_text(size=14))#theme for plots  
wine <- read.csv("wine.dat", header = TRUE) #import data  
wine$LPRICE2 <- as.integer(wine$LPRICE2)  
wine$DEGREES <- as.integer((wine$DEGREES))

1. The prices for 1954 and 1956 are missing because these vintages are so poor that they are no longer traded.

d,

wine[1:28,-1 ] %>% ggpairs(aes(color=), columns=c(1:5), cardinality\_threshold = 36) +  
 theme.info

 It looks like the rain in the months preceding growing season and the temperature during growing season are highly correlated, far more than any other combination.

lm.1 <- lm(LPRICE2 ~ TIME\_SV, data=wine)  
summary(lm.1)

##   
## Call:  
## lm(formula = LPRICE2 ~ TIME\_SV, data = wine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.912 -7.304 2.120 3.957 16.093   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 23.0034 1.9157 12.008 3.77e-14 \*\*\*  
## TIME\_SV -0.4171 0.1152 -3.621 0.000897 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.788 on 36 degrees of freedom  
## Multiple R-squared: 0.2669, Adjusted R-squared: 0.2466   
## F-statistic: 13.11 on 1 and 36 DF, p-value: 0.0008972

lm.2 <- lm(LPRICE2 ~ TIME\_SV + DEGREES + HRAIN + WRAIN, data=wine)  
summary(lm.2)

##   
## Call:  
## lm(formula = LPRICE2 ~ TIME\_SV + DEGREES + HRAIN + WRAIN, data = wine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.7509 -4.8452 0.2843 3.2106 20.2489   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 22.476094 7.472863 3.008 0.005007 \*\*   
## TIME\_SV -0.463944 0.113861 -4.075 0.000272 \*\*\*  
## DEGREES -0.234496 0.119255 -1.966 0.057718 .   
## HRAIN 0.029621 0.019466 1.522 0.137619   
## WRAIN 0.002241 0.009441 0.237 0.813828   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.431 on 33 degrees of freedom  
## Multiple R-squared: 0.3882, Adjusted R-squared: 0.3141   
## F-statistic: 5.235 on 4 and 33 DF, p-value: 0.002231

Looking at the R-squared values of each model it seems that the second model looking at multiple independent variables is the better of the two models. This is the same conclusion the authors reached as well. f. Both models have n=37. g.LPRICE2 = 22.48 -.46(TIME\_SV) -.23(DEGREES)+.03(HRAIN)+.002(WRAIN)+e The y-intercept represents the log price of wine ignoring all factors except error so I say it is not usable. h.

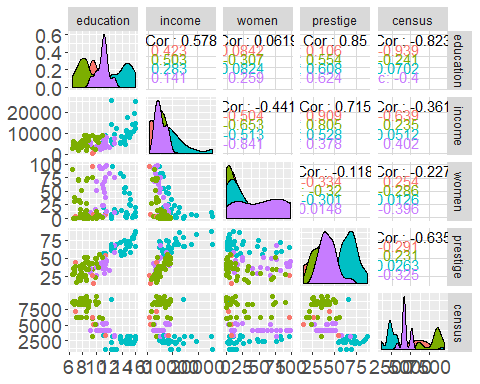
prediction.1 <- predict(lm.1)  
prediction.4 <- predict(lm.2)  
  
#SSE  
SSE <- c(sum(lm.1$resid^2), sum(lm.2$resid^2))  
#RMSE  
RMSE <- c(sqrt(sum((prediction.1-wine$LPRICE2)^2)/(nrow(wine) - 1- 1)), sqrt(sum((prediction.4-wine$LPRICE2)^2)/(nrow(wine) - 4- 1)))  
#PRESS stat  
PRESS <- c(press(lm.1), press(lm.2))  
# computing RMSE jackknife  
RMSEj <- c(sqrt(press(lm.1)/(nrow(wine)-1-1)), sqrt(press(lm.2)/(nrow(wine)-4-1)))  
  
tab <- rbind(SSE,RMSE,PRESS,RMSEj)  
colnames(tab) <- c("lm.1", "lm.2")  
show(tab)

## lm.1 lm.2  
## SSE 2183.297516 1822.10031  
## RMSE 7.787628 7.43069  
## PRESS 2428.611861 2526.01963  
## RMSEj 8.213491 8.74906

I would not change my previous decision, the model with more predictors has a lower RMSE. 1. If we were able to come up with all of the data for the relevant time up through 2005 in theory we could use this model. This assumes that nothing about the wine making process has changed, too.

2)a. The Pineo-Porter scale is a measure of an occupation’s “prestige” or more simply how an occupation rank’s in terms of benefits to oneself and society. It is primarily computed by measures of different socioeconomic statistics and location information (Goyder & Frank, 2007). This method began to come under criticism in the late 20th/early21st centuries in favor of newer measures which incorporate some information regarding the level of certain skills required. Considering what the newer scales take into account I do not think the Pineo-Porter rating is useful.

prestige <- read.csv("prestige.dat", header = TRUE) #imiport data  
  
#make scattermatrix (NOTE: colors are different professions rather than shapes)  
prestige %>% ggpairs(aes(color=type), columns=c(2:6), cardinality\_threshold = 36) +  
 theme.info

 c.

missing <- is.na(prestige$type)  
show(missing)

## [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [34] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [45] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [56] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [67] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [78] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [89] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [100] FALSE FALSE FALSE

Athletes, newsboys, babysitters and farmers do not have types. Since these professions vary widely in terms of income and are proobably less frequently seen they should be removed from the analysis. I don’t think they represent a meaningful sample of a particular population.

missing <- prestige[is.na(prestige$type), ]  
show(missing)

## [1] occupation.group education income women   
## [5] prestige census type   
## <0 rows> (or 0-length row.names)

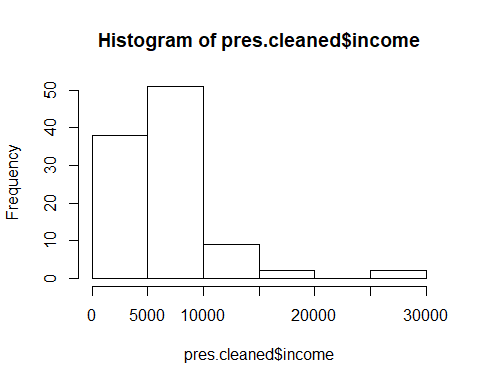
1. There seems to be a strong interaction between type and education and a moderate to weak interaction between type and income.

pres.cleaned <- prestige[!is.na(prestige$type), ]  
  
lm.pres <- lm(prestige~ income + education + type +income\*type + education\*type, data = pres.cleaned)  
summary(lm.pres)

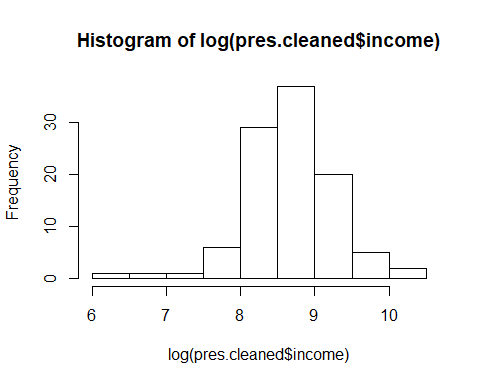
##   
## Call:  
## lm(formula = prestige ~ income + education + type + income \*   
## type + education \* type, data = pres.cleaned)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.462 -4.225 1.193 3.826 19.631   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 46.682990 18.928014 2.466 0.015545 \*   
## income 0.005294 0.001151 4.600 1.38e-05 \*\*\*  
## education -3.175963 2.137155 -1.486 0.140755   
## typebc -44.407237 20.212568 -2.197 0.030588 \*   
## typeprof -29.055341 22.314208 -1.302 0.196205   
## typewc -77.943889 24.953270 -3.124 0.002404 \*\*   
## income:typebc -0.001772 0.001279 -1.385 0.169579   
## income:typeprof -0.004674 0.001172 -3.987 0.000136 \*\*\*  
## income:typewc -0.003843 0.001349 -2.850 0.005422 \*\*   
## education:typebc 4.889237 2.343635 2.086 0.039792 \*   
## education:typeprof 6.277046 2.306275 2.722 0.007797 \*\*   
## education:typewc 9.180112 2.600022 3.531 0.000656 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.349 on 90 degrees of freedom  
## Multiple R-squared: 0.8787, Adjusted R-squared: 0.8638   
## F-statistic: 59.25 on 11 and 90 DF, p-value: < 2.2e-16

Income alone and the income x type “prof” are the most significant predictors of prestige, while incomextype“wc” and education x “typewc” are significant, too.

hist(pres.cleaned$income)



hist(log(pres.cleaned$income))

 The income alone histogram is positively skewed while the log income resembles a normal distribution.

lm.pres2 <- lm(prestige~ log(income) + education + type +log(income)\*type + education\*type, data = pres.cleaned)  
summary(lm.pres2)

##   
## Call:  
## lm(formula = prestige ~ log(income) + education + type + log(income) \*   
## type + education \* type, data = pres.cleaned)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.970 -4.124 1.186 3.829 18.059   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -65.0492 28.1379 -2.312 0.0231 \*   
## log(income) 13.6120 3.1497 4.322 3.98e-05 \*\*\*  
## education -0.4121 2.0067 -0.205 0.8378   
## typebc -54.9967 34.7700 -1.582 0.1172   
## typeprof 30.1634 37.0516 0.814 0.4177   
## typewc -24.7555 43.0760 -0.575 0.5669   
## log(income):typebc 2.3705 4.1101 0.577 0.5656   
## log(income):typeprof -7.0583 4.1930 -1.683 0.0958 .   
## log(income):typewc -5.7851 4.7805 -1.210 0.2294   
## education:typebc 2.7477 2.2160 1.240 0.2182   
## education:typeprof 3.4451 2.2023 1.564 0.1213   
## education:typewc 6.3878 2.5139 2.541 0.0128 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.494 on 90 degrees of freedom  
## Multiple R-squared: 0.873, Adjusted R-squared: 0.8575   
## F-statistic: 56.26 on 11 and 90 DF, p-value: < 2.2e-16

Simply put there are a couple more significant IV’s using the log income.

1. The model quality (R^2) decreases slightly when log income is used. A partial F test would not be appropriate because we have not changed the number of predictor variables, rather we are measuring some on a log scale.