Anthony Cunningham

STAT 3200

Due April 26

**Homework 9**

#1. > attach(Bfox)

> time=as.numeric(rownames(Bfox))

> Bfox.new=cbind(Bfox, time)

> names(Bfox.new)

[1] "partic" "tfr" "menwage" "womwage" "debt" "parttime" "time"

A. > lm.out=lm(partic~., data=Bfox.new)

> summary(lm.out)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.043e+01 1.913e+02 0.264 0.79443

tfr 2.618e-04 4.093e-04 0.640 0.52874

menwage -4.150e-02 1.478e-01 -0.281 0.78133

womwage 5.724e-02 1.758e-01 0.325 0.74776

debt 6.689e-02 1.558e-02 4.294 0.00027 \*\*\*

parttime 6.865e-01 8.185e-02 8.387 1.89e-08 \*\*\*

time -1.806e-02 9.944e-02 -0.182 0.85745

Residual standard error: 0.5333 on 23 degrees of freedom

Multiple R-squared: 0.9936, Adjusted R-squared: 0.992

F-statistic: 597.8 on 6 and 23 DF, p-value: < 2.2e-16

The effect of fertility differs from the expectation of the researchers. They were expecting a negative effect on women’s labor force participation rate, while the estimated regression coefficient is positive.

Insignificant partial regression coefficients: tfr, menwage, womwage, time

B. > plot(Bfox.new, pch=16)

> round(cor(Bfox.new), 4)

partic tfr menwage womwage debt parttime time

partic 1.0000 -0.8784 0.9595 0.9674 0.9819 0.9504 0.9531

tfr -0.8784 1.0000 -0.7942 -0.8537 -0.8437 -0.8904 -0.7612

menwage 0.9595 -0.7942 1.0000 0.9830 0.9861 0.8533 0.9891

womwage 0.9674 -0.8537 0.9830 1.0000 0.9868 0.8715 0.9637

debt 0.9819 -0.8437 0.9861 0.9868 1.0000 0.8875 0.9805

parttime 0.9504 -0.8904 0.8533 0.8715 0.8875 1.0000 0.8459

time 0.9531 -0.7612 0.9891 0.9637 0.9805 0.8459 1.0000

\*Yes, there is a lot of correlation between some predictors. In fact, I counted 6 different pairs of predictors with an absolute correlation greater than 0.9, while the lowest absolute correlation value between a pair of predictors is 0.76.

C.1) > vif(lm.out) \* most of the vif’s are large (greater than 10), which is concerning

tfr menwage womwage debt parttime time

9.085772 112.194563 69.637039 95.032322 7.649806 78.136864

2) > summary(lm(menwage~tfr + womwage + debt + parttime + time, data =Bfox.new))$r.squared

[1] 0.9910869

VIF = 1/(1 – R2) = 1/(1 – 0.9910869) = 112.19

D. > model.selection = step(lm.out)

Predictors removed using AIC: time, menwage, womwage, tfr

Predictors included in model: debt, parttime

E. > model.selection.BIC = step(lm.out, k=log(nrow(Bfox.new)))

Predictors removed using BIC: time, menwage, womwage, tfr

Predictors included in model: parttime, debt

\*Yes, BIC’s model includes the same predictors as AIC’s model.

F.1) > summary(model.selection)

Coefficients:

Estimate Std. Error t value Pr(>|t|) \* Yes, the partial regression

(Intercept) 16.325014 0.402170 40.59 < 2e-16 \*\*\* coefficients are significant.

debt 0.062573 0.003233 19.35 < 2e-16 \*\*\*

parttime 0.661329 0.059883 11.04 1.62e-11 \*\*\*

Residual standard error: 0.4973 on 27 degrees of freedom

Multiple R-squared: 0.9935, Adjusted R-squared: 0.993

F-statistic: 2062 on 2 and 27 DF, p-value: < 2.2e-16

2) > AIC(lm.out)

[1] 55.44862 \*AIC’s model is better than the original according

> AIC(model.selection) to AIC.

[1] 48.06513

3) > BIC(lm.out) \*BIC prefers the final model to the original model.

[1] 66.6582

> BIC(model.selection)

[1] 53.66992

4) The final model has a slightly larger Adjusted R2 value than the original model (0.993 to 0.992).

5) > vif(model.selection)

debt parttime \*No, there isn’t great concern for multicollinearity as the

4.708948 4.708948 VIF’s aren’t very big.

6) > plot(hatvalues(model.selection), rstudent(model.selection), type="n")

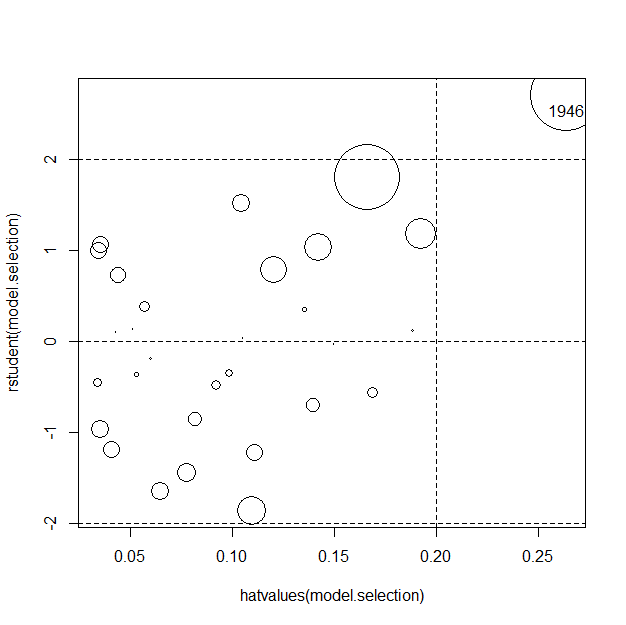
> cook=sqrt(cooks.distance(lm.out))

> points(hatvalues(model.selection), rstudent(model.selection), cex=10\*cook/max(cook))

> abline(h=c(-2,0,2), lty=2)

> abline(v=c(2,3)\*3/30, lty=2)

> identify(hatvalues(model.selection), rstudent(model.selection), row.names(Bfox))



1946’s data appears to be most influential on our model, mainly due to a large studentized residual, although it has a decent-sized hat-value as well.

G. > model.1 = lm(partic ~ tfr + debt + parttime)

> model.2 = lm(partic ~ menwage + debt + time)

> AIC(model.1)

[1] 49.89694

> AIC(model.2)

[1] 98.61183 \* Both AIC and BIC prefer the first model.

> BIC(model.1)

[1] 56.90292

> BIC(model.2)

[1] 105.6178

H.1)> fit1 = lm(partic ~ debt + parttime + debt:parttime)

> vif(fit1)

debt parttime debt:parttime

42.22116 10.93810 71.87413

\*Yes, there is a large concern of multicollinearity since every VIF is large.

2) > c.debt = debt - mean(debt)

> c.parttime = parttime - mean(parttime)

> c.int = (debt - mean(debt))\*(parttime - mean(parttime))

> model.1.centered = lm(partic~c.debt + c.parttime + c.int)

> vif(model.1.centered) \*There is no longer a concern for multicollinearity.

c.debt c.parttime c.int

5.661267 4.722036 1.767241

#2. A. > myfit = lm(intensity ~ commerce + tradition + commerce:tradition + midpeasant + inequality)

> vif(myfit)

commerce tradition midpeasant inequality

475.075987 6.369546 1.956909 2.092182

commerce:tradition

506.379119

\*The VIF’s for commerce and the interaction between commerce and tradition are extremely large, meaning that multicollinearity is a significant issue for the predictors.

B. > model.selection = step(myfit)

Remaining predictors in model: inequality, commerce, tradition, commerce:tradition

C. > model.selection.BIC = step(myfit, k=log(nrow(Chirot)))

Remaining predictors in model: commerce, tradition, commerce:tradition

\* BIC’s model is smaller

D. > summary(myfit)

Multiple R-squared: 0.6721, Adjusted R-squared: 0.609

F-statistic: 10.66 on 5 and 26 DF, p-value: 1.191e-05

> summary(model.selection)

Multiple R-squared: 0.672, Adjusted R-squared: 0.6235

F-statistic: 13.83 on 4 and 27 DF, p-value: 2.928e-06

> summary(model.selection.BIC)

Multiple R-squared: 0.6504, Adjusted R-squared: 0.6129

F-statistic: 17.36 on 3 and 28 DF, p-value: 1.45e-06

\*AIC’s model is the best according to Adjusted R2

E. 1) > vif(model.selection)

commerce tradition inequality commerce:tradition

472.869985 6.368064 1.159980 504.162192

Both commerce and the interaction between commerce and tradition still have large VIF’s, which is concerning since multicollinearity is still an issue in the model.

2) > c.commerce = commerce - mean(commerce)

> c.tradition = tradition - mean(tradition)

> model.selection.centered = lm(intensity ~ inequality + c.commerce + c.tradition + commerce:tradition)

> vif(model.selection.centered)

inequality c.commerce c.tradition commerce:tradition

1.159980 472.869985 6.368064 504.162192

\*Yes, there is still multicollinearity present in the model since the VIF’s are still large. We need to center the highest-ordered term (i.e. the interaction term commerce:tradition).