

Influential variables

You have used a t-statistic to judge whether an independent variable is *significantly related* to the dependent variable. Here we look at a related question: of all the significant independent variables, which exerts the *greatest influence* on the dependent variable? There are three main ways to answer this question.

The first is the use of standardized coefficients (often called beta coefficients). The standardized coefficients show by how many standard deviations the dependent variable will change, for a one standard deviation change in the independent variable. This is equivalent to the estimated coefficient times the standard deviation of the independent variable divided by the standard deviation of the dependent variable.

The second is the calculation of elasticities. An elasticity is the percentage change in a dependent variable caused by a one percent change in the independent variable (e.g., the percentage change in quantity demanded caused by a one percent change in price). In a model in which all variables are converted to their natural logs (such as the Cobb-Douglas production function), the coefficient estimates can be directly interpreted as elasticities. In a linear regression, one multiplies the estimated coefficient by the mean of the independent variable divided by the mean of the dependent variable.

The third way to determine the relative influences of the independent variables is to decompose the model R^2 into the portion attributable to each independent variable. The R^2 is the percent of the variation in the dependent variable that can be explained by the model. If the independent variables are perfectly orthogonal (that is, not correlated with each other), then each independent variable will explain a unique portion of the variation in the dependent variable. But independent variables are almost never orthogonal—they will share some variation, so that two or more independent variables will account for a portion of the variation in the dependent variables. Decomposition of R^2 into the portion accounted for by each independent variable is thus not straightforward. R contains several algorithms to perform this.

Influential observations

OLS coefficients are derived so as to minimize the sum of squared errors. Outlier observations can have a big effect, since errors are squared. It is often useful, particularly with small datasets, to see how the presence of a particular observation affects ones results. The seminal source for influence diagnostics is:

- Belsley, D. A., E. Kuh, and R. E. Welsch. 2004. *Regression diagnostics: Identifying influential data and sources of collinearity*. Wiley-IEEE.

R will perform a number of different tests recommended by Belsley and Kuh. The measure *dfbets* gives an overall measure of the influence of a particular observation. I've found their *dfbeta* measure especially useful—it gives some sense of how a particular regression coefficient has changed because of the inclusion of a particular observation.

Comments on Presentation of Results

A model starts from *theory*; independent variables should encompass the forces which *cause* the dependent variable. You must justify the variables you include in your model in an *ex ante* discussion, explaining how changes in these variables should lead to changes in the dependent variable.

Results should be explained. Regression output is the raw material for an explanation, not the explanation itself. The *ex post* discussion interprets the coefficients in the final model, explaining relationships, noting when they are in agreement and when they conflict with *ex ante* assumptions. In cases where the relationships are surprising, try to present a plausible reason why your results might disagree with your expectations.

In addition to appropriate specialized tests (which we will learn during the remainder of the semester), the following items should always be reported:

- descriptive statistics for the variables included in the unrestricted model, with *ex ante* justification
- the F-statistic used to justify the restricted model
- the results of the restricted model, including coefficient estimates, t-statistics and their p-values, the model R-squared, the number of observations.
- the *ex post* story interpreting the restricted model results.

Homework (due next Thursday, before noon)

This week's homework introduces you to the Standard Cross-Cultural Sample (SCCS). Each of the 186 observations is a society, and there are over 2,000 variables covering all areas of social life. A description of the SCCS, with useful links, can be found on Wikipedia:

http://en.wikipedia.org/wiki/Standard_cross-cultural_sample

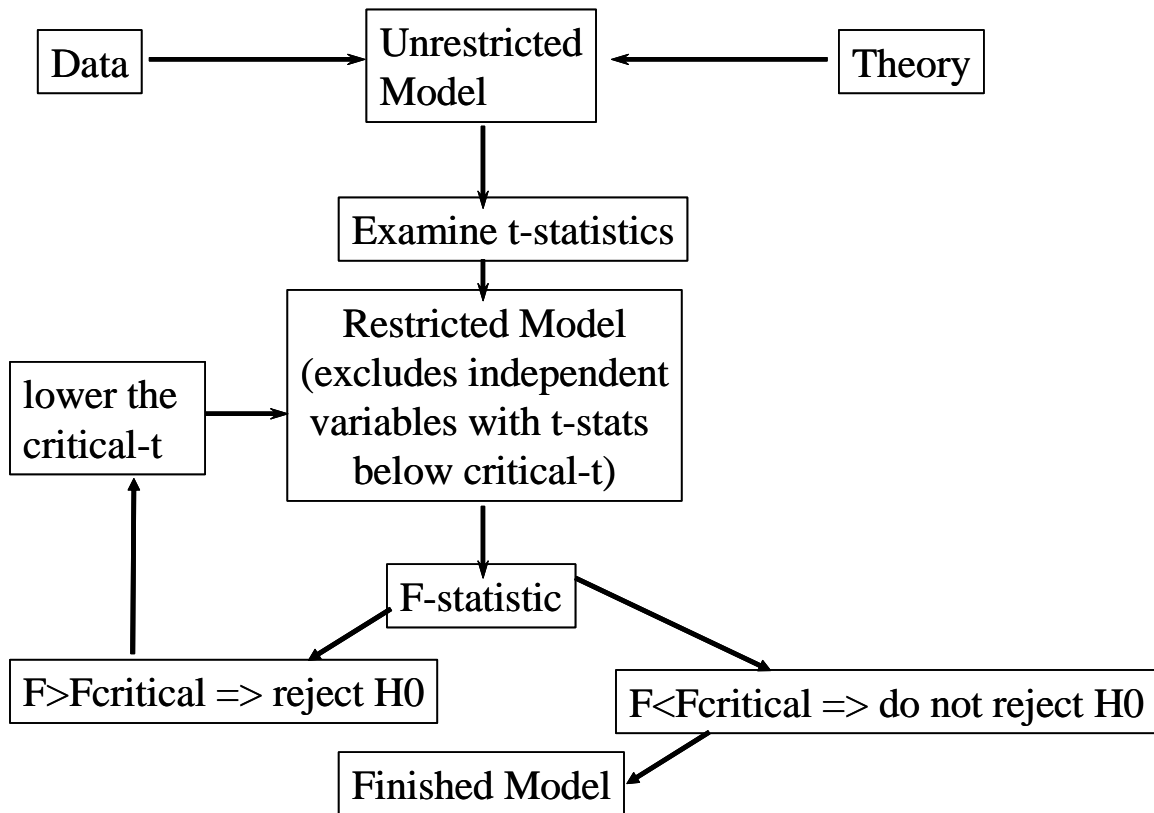
Our course webpage contains a link to the codebook.

Your assignment is to develop a model explaining the degree to which a society values children. The R program `S:\TEFF\662\R\r03.R` provides some code that you can cannibalize for this work. Follow your usual model-building strategy. Spend some time finding plausible independent variables (using the codebook).

Produce two nicely formatted tables: the first with descriptive statistics for the variables used; the second with regression results for the restricted model. Write at least a page or two of *ex ante* and *ex post* discussion. Include in your *ex post* discussion some analysis of the

relative importance of the independent variables in your final model. Additionally, if any of your results are different from what you had expected, briefly discuss the observations that were most influential.

Schema of General Model-Building Strategy



R program: S:\TEFF\662\R\r03.R

```
#SCCS--model for value of children--
rm(list=ls(all=TRUE))
#--Set path to your directory with data and program--
setwd("d:/class/662/R")
options(echo=TRUE)
#--need these packages for estimation and diagnostics--
library(foreign)
library(car)

#--Read in data over web-
load(url("http://frank.mtsu.edu/~eaeff/downloads/sccs200909.Rdata"), .GlobalEnv)
ls();dim(sccs200909)
#--rename dataframe--shorter name is easier--
SCCS<-sccs200909

#--extract variables to be used, put in dataframe fx--
fx<-data.frame(
  socname=SCCS$socname,socID=SCCS$ord,
  valchild=(SCCS$v473+SCCS$v474+SCCS$v475+SCCS$v476),
  roots=(SCCS$v233==5)*1,gath=SCCS$v203,hunt=SCCS$v204,
  milk=(SCCS$v245>1)*1,bovines=(SCCS$v244==7)*1,tree=(SCCS$v233==4)*1,
  foodscarc=SCCS$v1685,popdens=SCCS$v156,
  exogamy=SCCS$v72,setType=SCCS$v234,
  localjh=(SCCS$v236-1),moralgods=SCCS$v238,
  war=SCCS$v1648,himilexp=(SCCS$v899==1)*1,money=SCCS$v155)

row.names(fx)<-fx$socname

#--check to see number of missing values,
#--whether variables are numeric,
#--and number of discrete values for each variable---
vvn<-names(fx)
pp<-NULL
for (i in 1:length(vvn)){
  nmis<-length(which(is.na(fx[,vvn[i]])))
  numeric<-is.numeric(fx[,vvn[i]])
  numDiscrVals<-length(table(fx[,vvn[i]]))
  pp<-rbind(pp,cbind(data.frame(numeric),nmis,numDiscrVals))
}
row.names(pp)<-vvn
pp

#--number of observations left after listwise deletion--
nobs<-length(which(!is.na(rowSums(fx[,-1]))))
nobs

xUR<-lm(valchild~roots+gath+
  hunt+milk+bovines+tree+foodscarc+
  popdens+exogamy+
  setType+localjh+moralgods+
  war+himilexp+money
  ,data=fx)

#--standardized coefficients---
z<-which(!is.na(rowSums(fx[,-1])))
cc<-coef(xUR)[-1]
sdX<-sd(fx[z,names(cc)])
stdcoef<-cc*sdX/sd(fx[z,"valchild"])
stdcoef[order(stdcoef)]

#--elasticities---
z<-which(!is.na(rowSums(fx[,-1])))
cc<-coef(xUR)[-1]
mnx<-mean(fx[z,names(cc)])
elast<-cc*mnx/mean(fx[z,"valchild"])
elast[order(elast)]
```

```

#--partitioning R2--this is slow with more than 10 or so variables--
library(realmipo)
om<-calc.relimp(xUR)$lmg
om[order(om)]

#--dffits---
o<-as.matrix(dffits(xUR))
dft<-o[order(o[,1]),]
cbind(dft)

#--dfbetas: just look at the most influential---
o<-data.frame(dfbetas(xUR))[, -1]
pp<-names(o)
mx<-NULL
for (i in 1:length(pp)){
z<-which(abs(o[,pp[i]])>2/nobs^.5)
pz<-cbind(o[z,pp[i]],pp[i],row.names(o)[z])
pz<-pz[order(pz[,1]),]
mx<-rbind(mx,data.frame(pz))
}
names(mx)<-c("dfbeta","variable","observation")
mx

#--select independent variables to drop, do F-test---
z<-which(summary(xUR)$coefficients[-1,4]>.1)
coefs<-names(coef(xUR))[-1]
coefs<-coefs[z]
linear.hypothesis(xUR,coefs)

#--write results to csv file for perusal in spreadsheet--
write.csv("==OLS model for value of child==",file="OLS results.csv",append=FALSE)
write.csv(summary(xUR)$coefficients,file="OLS results.csv",append=TRUE)
write.csv(nobs,file="OLS results.csv",append=TRUE)
write.csv(summary(xUR)$r.squared,file="OLS results.csv",append=TRUE)
write.csv(linear.hypothesis(xUR,coefs),file="OLS results.csv",append=TRUE)
write.csv("==Influential variables==",file="OLS results.csv",append=TRUE)
write.csv(stdcoef[order(stdcoef)],file="OLS results.csv",append=TRUE)
write.csv(elast[order(elast)],file="OLS results.csv",append=TRUE)
write.csv(om[order(om)],file="OLS results.csv",append=TRUE)
write.csv("==Influential observations==",file="OLS results.csv",append=TRUE)
write.csv(dft,file="OLS results.csv",append=TRUE)
write.csv(mx,file="OLS results.csv",append=TRUE)

```

Individual Homework Questions; the answers are due next week, in class. Define briefly, in your own words, each of the following. Please try to give an intuitive, descriptive definition. Please write neatly.

a) **Error Sum of Squares**

b) **P-Value**

c) **R-squared**

d) **Chow test**

e) **J-test**

f) **Degrees of Freedom**

g) **Null Hypothesis**