```
#PO 841
#11/19/2014
#Lab 9: Multivariate Regression
#============================

#To explore multiple regression, let's do an example with data on health
 outcomes across OECD countries:

library(foreign)
health<-read.dta(file.choose(),convert.factors=F)
names(health)
health$country

#The dependent variable we will be assessing is public-sector health
 expenditures (pubhth). Wealthier countries could be expected to spend
 more on health, on average. Let's start with a bivariate regression:

reg1<-lm(pubhth~gnppc, data=health)
summary(reg1)

#What do these summary results tell us? For every $1 more in GNP per
 capita, we can expect a .01353 percentage-point increase in the
 proportion of the national budget devoted to health. Translated into
 more relevant terms, we can say that for every 10,000 more of GNP per
 capita, we can expect an increase of 1.35 percentage points in the
 proportion of the budget devoted to health.

#Some governments might spend less on health because the health system is
 largely private, and individuals or insurance companies are spending the
 money instead (e.g., the USA). So let's do a multivariate regression and
 control for private sector health spending too. When we say we "control"
 for another variable, what we mean is that we are assessing the effect
 of each predictor variable in our model while holding the values of all
 the other variables constant; in other words, we are able to assess the
 effect of each predictor variable c.p., all else being equal. We add
 predictors to our model that are potential confounders to our bivariate
 regression model -- if we just did a bivariate regression and left out
 the private-sector health expenditures, and private-sector health
 expenditures WERE a significant factor in predicting overall public
 spending on health, then we would get biased results. We include
```

```
possible confounders in our model in order to estimate the effect of a
predictor variable after accounting for the presence and possible
effects of other predictor variables.

reg2<-lm(pubhth~gnppc + privhth, data=health) #When specifying a
multivariate model, you put a '+' between the X variables.
summary(reg2)

#What does the summary output tell us? The coefficient for "gnppc" barely
changed, and for "privhth" it is not significantly different from zero.
This suggests that private-sector health expenditures are NOT a
confounding variable. In other words, the effect of GNP per capita on
public health expenditures is essentially the same even after
controlling for other possible effects on public health spending (like
the amount of money that the private sector spends on health care).

#What are some other possible confounding variables that we haven't
included in our model and that might have an effect on public spending
on health? Countries might spend less on health because they spend more
on other things, like defense spending, say. So we can control for the
effect of the amount of defense spending on public health expenditures:

reg3<-lm(pubhth~gnppc + privhth + buddef, data=health)
summary(reg3)

#By examining the summary output, we can see that there is no evidence
that defense spending has any independent effect on health spending.
However, note that the effect of GNP per capita has become less
significant after controlling for defense spending than before
(significant at the .05 level but not at the .01 level). Also note the
increase in the multiple-r-squared statistic from regressions 1 to 2 to
3:

c(summary(reg1)$r.squared, summary(reg2)$r.squared,
summary(reg3)$r.squared)

#The adjusted-r-squared gives us a better picture of the explanatory
power of our model, though (accounting for the addition of variables):

c(summary(reg1)$adj.r.squared, summary(reg2)$adj.r.squared,
```

```
summary(reg3)$adj.r.squared)

#We can do all the same diagnostics with multiple regression, like
residual plots, but you have to choose one x variable to plot the
residuals against.

plot(reg3$residuals~gnppc, data=health, pch=19)
abline(h=0)

#What assumptions appear to be violated here, if any?

plot(reg3$residuals~privhth, data=health, pch=19)
abline(h=0)

#What assumptions appear to be violated here, if any?

plot(reg3$residuals~buddef, data=health, pch=19)
abline(h=0)

#What assumptions appear to be violated here, if any?

#Important note: Datasets may be missing data for some observations
(e.g., some countries didn't report the percentage of their budget spent
on education in X year, some people surveyed refused to answer a
particular question, etc). These missing values are typically stored as
NA in the data file (or you can recode them to equal NA). When you run a
regression, any observation that has missing values on Y or any X
variable is deleted and, as a result, you might end up with fewer
residuals than valid entries for a given X variable. If so, and you try
to generate a residuals plot using the approach above, you'll get an
error message saying that 'x' and 'y' lengths differ.

#One solution for this problem is to use the argument x=T in lm() to tell
R to save the version of each X variable actually used in the
regression, e.g., to strip out any rows containing missing variables.
Then you can plot the residuals against this version of the X variable.
For example:

reg3<-lm(pubhth~gnppc + privhth + buddef, data=health, x=T) #The only
difference here is that we're specifying x=T
```

```
plot(reg3$residuals~reg3$x[,'gnppc'], pch=19) #In the quotation marks you
 specify the X variable
abline(h=0)

plot(reg3$residuals~reg3$x[,'privhth'], pch=19)
abline(h=0)


# ----------------------------------------
# TESTING NULL HYPOTHESES OTHER THAN BETA=0
# ----------------------------------------

#As I briefly mentioned in last week's lab, when you examine regression
 results using summary(), the standard errors, t-statistics, and p-values
 reported are all those corresponding to a 2-tailed test of the null
 hypothesis that beta = 0. If you want to do a 1-tailed test, you can
 just divide the p-value by 2. But what if you want a different null
 hypothesis?

#Let's consider a different relationship: the effect of health spending
 on a health outcome such as life expenctancy. Let's do a multiple
 regression of life expectancy on public and private health expenditures,
 both of which should affect it:

reg4<-lm(lifexp~pubhth+privhth,data=health)
summary(reg4)

#What can we conclude about the effect of public and private health
 spending on life expectancy?

#Let's assume left-leaning health economists have long claimed that each
 percentage point increase in public health spending as a share of GDP
 buys a country an additional year of life expectancy.  Right-wing health
 economists say it's less. Whose claim do these data support? We need to
 test the null hypothesis that beta_1 = 1 against the alternative beta_1
 < 1. We have to do this "by hand":

coef(reg4)
coef(reg4)[2] #The coefficient for public health spending is the second
 one (beta_0, beta_1)
```

```
coef(reg4)['pubhth'] #Beta-hat for public health expenditures

#We'll go over variance-covariance matrices in class today, but
 essentially, this matrix provides us with the variance of beta-hat so
 that we can do a hypothesis test:

vcov(reg4) #Variance-covariance matrix of beta-hat
vcov(reg4)[2,2] #Row 2, column 2 is where the vcov matrix stores the
 information for the variance of beta_1 (the variance of the coefficient
 for "pubhth." The variance of a particular coefficient can be found
 where the same variables intersect (e.g., the variance of beta_2,
 "privhth", would be 0.10196436; the variance of the intercept beta_0
 would be 2.4239598). Another way of thinking about it is that the
 variances of the coefficients can be found by moving diagonally across
 the matrix from the upper-left element to the lower-right element.

vcov(reg4)['pubhth','pubhth'] #Gives us the same result

#Now let's calculate the t-statistic "by hand". Remember that our null
 hypothesis is the left-wing economists' claim that beta_1 = 1:

tstat<-(coef(reg4)['pubhth'] - 1)/sqrt(vcov(reg4)['pubhth','pubhth'])
tstat

#We also have to find the degrees of freedom in order to conduct a t-
 test:

reg4$df.residual

#Now we can conduct the t-test using pt():

pt(tstat,reg4$df.residual)

#The p-value for this hypothesis test is just above 0.05. Beta_1 may be
 significantly different from zero, but it is NOT significantly different
 from 1 at the .05 level. So there is not strong enough evidence to
 support to right-wing economists' claims.

#For another example, let's say that left-wing economists also claim that
 public health expenditures have a larger effect on life expectancy than
```

```
private health expenditures, whereas right-wing economists claim that
there is no difference. To see if the data support this, we need to test
the difference between beta_1 and beta_2. So our null hypothesis is that
beta_1 - beta_2 = 0 against an alternative that beta_1 - beta_2 > 0.

#Note: Remember that var(beta_1 - beta_2) = var(beta_1) + var(beta_2) - 2
* cov(beta_1, beta2). We'll need to compute that value, because we'll
need the variance not of beta_1 or beta_2 but of the difference between
the two coefficients.

#Let's take another look at the variance-covariance matrix:

vcov(reg4)

#Now let's hand-roll the variance of the difference between beta_1
("pubhth") and beta_2 ("privhth"):

var.diff<-vcov(reg4)['pubhth','pubhth'] + vcov(reg4)['privhth','privhth']
- (2 * vcov(reg4)['pubhth', 'privhth'])
var.diff

#Now let's conduct the hypothesis test. Remember that our null hypothesis
is that there is no difference between the effect of public health
expenditures on life expectancy and the effect of private health
expenditures on life expectancy:

tstat2<-((coef(reg4)['pubhth'] - coef(reg4)['privhth']) - 0)/
(sqrt(var.diff))
tstat2

#And now let's find the p-value:

1-pt(tstat2, reg4$df.residual)

#The p-value is > any conventionally-defined level of significance. We
cannot reject this null hypothesis: The estimated change in life
expectancy associated with an increase in public health spending is NOT
significantly different from the estimated change associated with an
increase in private health spending. In other words, we don't have
evidence to support the left-wing economists' claim.
```

```
# --------------------
# IDENTIFYING OUTLIERS
# --------------------

#Let's take another look at the bivariate relationship between public
 health expenditures and GNP per capita:

plot(pubhth~gnppc,data=health, pch=19)
abline(reg1)

#There are some outliers here, so let's point them out. We can use the
 labels=() command to tell R what to label our points. (I included this
 command as an afterthought in last week's lab after Mike's question in
 section, so now let's go over it in more detail!) We can use the
 labels=() command to label various things in R -- we can label the
 points on a scatterplot, the x and y axes, the values on the x and y
 axes, etc. Here is the example I used in last week's lab, about test
 scores and hours studied:

scores<-c(48,76,50,72,60,54,92,69,65,91,69,83,78,98)
hours<-c(2,4,3,4,3,4,6,6,5,7,5,5,6,8)
names<-c("Joe", "Mike", "Bill", "Claire", "Sarah", "Kate", "Emily",
 "Michelle", "Matt", "Sue", "Jane", "Bob", "John", "Jennifer")
data.frame(names, scores, hours)

reg1<-lm(scores~hours)
plot(scores~hours, pch=19)
abline(reg1)
text(hours, scores, labels=names, cex=.6, pos=3) #Label all the data
 points (cex denotes font size, pos denotes position of the text)

plot(scores~hours, pch=19) #You have to call up a new plot if you don't
 want the new text superimposed on the old plot
abline(reg1)
text(6, 78, labels="John", cex=.7, pos=3) #Label just one of the data
 points; specify the coordinates of your data point and what you want to
 label that point.
```

```
#Back to the health outcomes example: We can specify labels=country,
 which tells R to label the points with the data set's country names; if
 we leave the labels=() argument out, R will label the plot with
 identification (or case) numbers.

plot(health$gnppc, health$pubhth, pch=19)
abline(reg1)

#The identify() command lets us manually identify points according to
 what we've specified our labels will be -- in other words, to identify
 outliers. This is one of R's interactive graphic features, so when
 you're done identifying the points, hit ESC. (Note: Don't close out the
 plot before you run the identify() command.)

identify(health$gnppc, health$pubhth, labels=health$country)

#Because we see a few outliers that represent different regions of the
 world (e.g., Sweden, Turkey), it seems that region may matter for this
 relationship. So let's create some new categorical variables that
 represent region. First, we'll create a new variable specifying whether
 or not the country is in Europe or not. This will return a logical
 vector that will specify whether or not a country is in Europe (FALSE or
 TRUE).

health$country

europe<-health$country %in% c("Portugal", "Greece", "Spain", "Italy",
 "Ireland", "United Kingdom", "Netherlands", "Belgium", "Austria",
 "France", "Germany", "Denmark", "Finland", "Norway", "Sweden",
 "Switzerland") #We're telling R here, essentially, for all values of
 health$country, is the country one of the following (Portugal, Greece,
 Spain, etc.), true or false?

europe
summary(europe)

# ------------------------------------------------
# REGRESSION WITH CATEGORICAL OR DUMMY VARIABLES
# ------------------------------------------------
```

```
#So far we have only done regression with continuous variables, but it's
 easy to add a dummy variable to a multiple regression model.

reg4<-lm(pubhth ~ gnppc + europe, data=health) #The coefficient will give
 you the effect on Y of "europe"'s being TRUE.
summary(reg4)

#How do we interpret this output? The coefficient for "gnppc" tells us
 the effect of increasing GNP per capita on public health spending while
 holding "europe" constant -- that is, it tells us the effect of GNP per
 capita even after controlling for region (for a country's being in
 Europe or not).

#The coefficient for "europe" tells us what happens when that variable
 changes from 0 to 1, i.e., from false to true. This effect captures the
 idea that health spending is generally higher in Europe compared with
 other countries, above and beyond what can be accounted for by wealth.

#What if we wanted to graphically compare the relationship between GDP
 per capita and public spending in European versus non-European
 countries?

#First, let's subset our data according to whether or not a country is in
 Europe or not:

eur<-subset(health, europe=="TRUE")
eur
noneur<-subset(health, europe=="FALSE")
noneur

#Now let's run separate regressions for each subset:
regeur<-lm(eur$pubhth~eur$gnppc, data=eur)
summary(regeur)
regnoneur<-lm(noneur$pubhth~noneur$gnppc, data=noneur)
summary(regnoneur)

#Now let's plot all the data, for ALL countries:
plot(pubhth~gnppc, main="Public Health Spending by Country Wealth",
 ylab="Percent of GDP on Health", xlab="GNP per capita", pch=19,
 data=health)
```

```
#And finally, let's add in our regression lines -- black for the
 relationship between GNPPC and public health spending in European
 countries, and red for the relationship between GNPPC and public health
 spending in non-European countries:
abline(regeur, col="black", lty=1, lwd=2)
abline(regnoneur, col="red", lty=1, lwd=2)

#We can also superimpose the original bivariate regression line
 estimating the relationship between public health and GNPPC across all
 OECD countries:

abline(reg1, col="green", lty=1, lwd=2)

#We can also add a legend:
legend("bottomright", legend=c("Europe","Other OECD Countries", "All OECD
 Countries"), lty=1, lwd=2,  col=c("black", "red", "green"))

#What if we assumed that the slope of the relationship between GNPPC and
 public health spending was the same for both European and non-European
 countries, but that just the intercepts were different? In other words,
 let's say we conducted a hypothesis test that told us that the beta_1
 coefficient (for GDPPC) for our European regression was NOT
 significantly different from our beta_1 coefficient (for GDPPC) for our
 non-European regression? How would we go about graphing this?

values.europe<-data.frame(gnppc=range(health$gnppc), europe=T) #Tells R
 to create a new data frame that includes the range of values for GNP per
 capita -- the minimum and maximum values, produced by range(health
 $cgnppc) -- along with a variable that indicates that the country is in
 Europe.
values.europe

values.noteurope<-data.frame(gnppc=range(health$gnppc), europe=F) #Tells
 R to create a new data frame that includes the range of values for GNP
 per capita -- the minimum and maximum values, produced by range(health
 $cgnppc) -- along with a variable that indicates that the country is not
 in Europe.
values.noteurope
```

```
#We need to create these new data frames because when we plot the
 different regression lines on the scatterplot, we're going to tell R to
 predict the relationship for European and non-European countries (which
 will have the same slope) based on the asumption that the full range of
 values of GNP per capita are in Europe AND on the assumption that the
 full range of values of GNP per capita are NOT in Europe.

#Then we can plot the relationship of GNP per capita and health spending,
 for both European and non-European countries. First, let's call up the
 plot of the relationship between public spending and GDP per capita:

plot(pubhth~gnppc, main="Public Health Spending by Country Wealth",
 ylab="Percent of GDP on Health", xlab="GNP per capita", pch=19,
 data=health)

lines(range(health$gnppc), predict(reg4,values.europe), lty=1, lwd=2,
 col="black")
lines(range(health$gnppc), predict(reg4,values.noteurope), lty=2, lwd=2,
 col="red")

#Let's add a legend:

legend("bottomright", legend=c("Europe","Other OECD Countries"),
 lty=c(1,2), lwd=2, col=c("black", "red"))
```