Lab 8: Multiple regression

Google's Transparency Report

> "Transparency is a core value at Google. As a company we feel it is our responsibility to ensure that we maximize transparency around the flow of information related to our tools and services. We believe that more information means more choice, more freedom and ultimately more power for the individual."

So begins Google's recently released Transparency Report. As a company, they have access to prodigious amounts of information about their users, information that is often of keen interest to governments. The report contains information related to the number of government inquiries for user data as well as the number of requests to remove content from Google's services. In this lab we'll consider a handful of countries and various statistics on each to help us get a sense of what defines countries with various levels of internet censorship. Our tool of choice here will be multiple linear regression, and while it might not be the method best suited to this particular situation, it can still provide us with useful insights.

The data

The data consist of the number of requests Google received for user account information as part of criminal investigations in the first half of 2011 and the rate of compliance as well as some other indicators on the countries.

Let's load up the Google data:

```
goog <- read.csv("http://www.openintro.org/stat/data/goog.csv")
```

Below is a list of the variables and descriptions.

`country` name of country.

`complied` percentage of requests Google complied with.

`requests` number of requests Google received for user account information as part of criminal investigations.

`pop` population of country, in thousands.

`hdi` human development index, a composite measure of life expectancy, literacy, education, and standard of living on a scale of 0 (least developed) to 1 (most developed). For more information, click here.

`dem` democracy index, categorized into `full` democracies, `flawed` democracies, and `hybrid` regimes. For more information, click here.

`internet` percentage of internet users. For more information, click here.

`freepress` free press index, scored on a scale from 1 (most free) to 100 (least free). For more information, click here.

We'll first focus on modeling Google's compliance rate (`complied`) using the rest of the explanatory variables.

Pairwise plots

We start our analysis by examining pairwise plots of the data. Since the first column is just the country name, we leave that column out.

```
plot(goog[, -1])
```

Note that you could also get the same plot with

```
plot(goog[, 2:8])
```

but it's a bit quicker to use the earlier command, saying that you want all columns except for the first one.

> **Exercise 1** Examine the pairwise scatterplots. What can you say about the trends in relationships between `complied` and the other variables? Are there any observations that stand out from the rest?

Drop outliers

One of the observations that stands out as unusual is the country with the highest population. Let's see which country this is:

```
which.max(goog$pop)
```

The result of the above function indicates that the observation in the 10th row of the data set is the country with the highest population, and this country is India. Next, we subset the data to exclude this observation.

```
goog_sub <- subset(goog, country != "india")
```

Let's take another look at the pairwise plots.

```
plot(goog_sub[, -1])
```

> **Exercise 2** Do any other countries appear unusual in this plot? If so, determine which one(s).

Regardless of your answer, we'll proceed with the analysis with the 25 countries remaining in the data frame `goog_sub`.

Defining the reference level

Currently there is one categorical variable in the data, `dem`. Let's make a frequency table for this variable.

```
table(goog_sub$dem)
```

The level that shows up first, `free`, is by default the reference level. R usually chooses the reference level as the level that comes up first alphabetically. We can change the reference level to `hybrid` using the following command:

```r
goog_sub$dem <- relevel(goog_sub$dem, ref = "hybrid")
```

Exercise 3 Confirm that the reference level for `dem` has been changed to `hybrid`.

Multiple linear regression

Let's fit a multiple linear regression model to predict compliance rate using the rest of the explanatory variables.

```r
m_full = lm(complied ~ requests + pop + hdi + dem + internet + freepress,
            data = goog_sub)
summary(m_full)
```

The format here is similar to what we used for simple regression, only now we have many predictors, each separated by `+`.

Exercise 4 Interpret the coefficient for `pop` (population) and the coefficients associated with the democracy index.

Model selection

p-value method, backwards-elimination

Exercise 5 Which variables in the full model are significant predictors?

Let's try to find a model that better predicts Google's compliance rate than the full model eliminating variables that are not found to be significant predictors of the response variable. When using a backwards-elimination approach, we do this sequentially, dropping the variable with the highest p-value, and then refitting the model.

The variable with the highest p-value in the full model is `internet`, therefore it is eliminated from the model first.

```r
m_step1 <- lm(complied ~ requests + pop + hdi + dem + freepress, data = goog_sub)
summary(m_step1)
```

Exercise 6 Are there any variables in this new model (`m_step1`) that are not significant. If so, which variable should be eliminated in the next step? Refit a new model called `m_step2` excluding this variable.

Exercise 7 Continue the backwards-elimination process untill all variables left in the model are significant. Which variables are included in the final model?

Adjusted R^2 method, backwards selection

Alternatively we can use the adjusted R^2 method where variables are dropped in order to increase the adjusted R^2 of the model.

Exercise 8 Record the adjusted R^2 of the full model in the table below. Then, drop one variable at a time from the model, and record the new adjusted R^2. Which variable should be dropped from the model in the first step of this approach?

Model	adjusted R^2
Full: `complied ~ requests + pop + hdi + dem + internet + freepress`	
`complied ~ pop + hdi + dem + internet + freepress`	
`complied ~ requests + hdi + dem + internet + freepress`	
`complied ~ requests + pop + dem + internet + freepress`	
`complied ~ requests + pop + hdi + internet + freepress`	
`complied ~ requests + pop + hdi + dem + freepress`	
`complied ~ requests + pop + hdi + dem + internet`	

The next step involves refitting the model many times again, dropping one variable at a time, and determining which dropped variable yields the highest increase in adjusted R^2. As you can imagine, refitting these models many times can get cumbersome.

A shortcut

The built in `step` function in R does stepwise model selection with one command. By default this function does backwards-elimination.

```
step(m_full)
```

The second to last argument in the output shows that the formula for the best model is `lm(formula = complied ~ pop + hdi, data = goog_sub)`. So let's try that model and see what the output looks like.

```
m <- lm(complied ~ pop + hdi, data = goog_sub)
summary(m)
```

On Your Own

You should work with your teammates on the lab, as well as the homework from the book, but you must turn in your own work. The answers to questions must be in your own words.

1. Do the countries in this data set represent a population or a sample? If you answered "sample", then what is the population of interest? Do these countries represent a random sample from this population? Is it reasonable to assume this sample is representative of the population? Why or why not?
2. Copy and paste the R output of the final model and interpret the slope estimates for each variable included in the model. Also comment on whether or not the p-values for these coefficients are meaningful in this context. (Hint: What is the null hypothesis for testing for significance of a slope coefficient in the context of multiple linear regression? What does a p-value mean in this context?)
3. Check model diagnostics on the final model. You may want to refer to code from the previous lab to do this as well as the Section 8.3 from the book on graphical model diagnostics for multiple linear regression.
4. Fit another model predicting number of `requests` from all other explanatory variables except for compliance rate. Using the `step` function find the "best" model and include only the output for this model.

Note that you might find that your model violates some assumptions. Multiple linear regression, while a powerful method, is still considered basic. Many real data sets will require more advanced techniques for proper analysis. If you are interested in what else is out there we encourage you to take a higher level statistics course.