Discussion Regarding Homework 5 and R code interpretation

Hypothesis Testing using the F- and T-tests: Writing Code and interpreting output

Before we begin: when testing hypotheses, you always need to specify both the null hypothesis and the alternative hypothesis. You need to know both the smaller and the bigger model: the answer you get, and the correct interpretation of that answer, depends on this!

These notes are not meant to discuss the theory of F- and T-testing, nor to be a comprehensive guide to using them in practice. The goal here is to provide a clear account of basic hypothesis testing using commands in R.

The notes will be based on the last problem in homework 5, using Ericksen's data, but the discussion to follow is not meant to be a solution to the problem exactly.

The first step is obviously to load the data; I will also transform crime, highschool, and poverty to the log scales.

```
ericksen = read.table("http://socserv.socsci.mcmaster.ca/
jfox/Books/Applied-Regression-2E/datasets/Ericksen.txt", header = T)

attach(ericksen)
crime = log(crime)
highschool = log(highschool)
poverty = log(poverty)
```

Then, we need to create a linear model. I will start with the model involving all three covariates, plus an intercept term.

```
mod.3 = lm(crime~highschool+poverty+minority)
summary(mod.3)
```

The output is

```
Call:
lm(formula = crime ~ highschool + poverty + minority)

Residuals:
     Min       1Q   Median       3Q      Max
-0.59213 -0.12056 -0.01524  0.11288  0.71592

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.306927   0.455679  13.841  < 2e-16 ***
highschool  -0.685577   0.161240  -4.252 7.27e-05 ***
poverty     -0.087566   0.152785  -0.573    0.569
minority     0.019617   0.002372   8.269 1.35e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '. ' 0.1 ' ' 1

Residual standard error: 0.2354 on 62 degrees of freedom
Multiple R-squared: 0.6031,Adjusted R-squared: 0.5839
F-statistic:  31.4 on 3 and 62 DF,  p-value: 1.808e-12
```

From an inferential point of view, the interesting parts of the model ar the t statistics in the table above, and the F-statistic listed in the last line.

The t statisic in the second line of the table, for poverty, tests the hypothesis that poverty has no predictive power in the linear model for crime. In full detail, this test makes the assumption (which may or may not be true) that the distribution of variables for each observation satisfies

$$(crime) = \alpha + \beta_1 \cdot (highschool) + \beta_2 \cdot (poverty) + \beta_3 \cdot (minority) + \epsilon,$$

where ϵ is random normal, and independent for each observation. Then, the hypothesis asks whether we have reason to beleive that β_2 in the model above is equal to zero. We are making no restrictions ont he other three coefficients α, β_1 and β_3. If we accept the null hypothesis (which we do here), it says that we do not have enough evidence to claim that β_2 must not be equal to zero. If we reject, it means there is enough evidence that β_2 is nonzero.

A word about causation: the hypothesis above does not say that (*crime*) was *generated* from the equation above, only that its *distribution* behaves in that way. So testing whether $\beta_2 = 0$ is *not* testing whether (*poverty*) affects (*crime*). Statistics alone from observational data can never make such a decision; one has to either do a properly designed experiment, or make causal assumptions which are stronger than the statistical assumptions made.

Aside from the t statistics, which test for the predictive power of each variable in the presence of all the others, the other test in the output above is the F-test at the bottom. This tests the null hypothesis that *all* of the β_j's are equal to zero against the alternative that allows them to take any values. If we reject this null hypothesis (which we do because the p-value is small), then this is the same as saying there is enough evidence to conclude that *at least one* of the covariates has predictive power in our linear model, i.e. that using a regression is predictively 'better' than just guessing the average.

Next, let us consider testing the hypothesis that two of the three covariates above, say (*highschool*) and (*minority*), matter, while one does not. This can be done via an F statistic. We can obtain the F statistic in several different ways. Both of the following blocks of code explicitly create the F-statistic (the first requires the `car` package)

```
linear.hypothesis(mod.3, c('poverty = 0'))

mod.2.nopov = lm(crime~highschool+minority)
anova(mod.3,mod.2.nopov)
```

These two commands can always be used to test nested models. If we want to test a model that has more than one variable set to zero, we just add more terms, in the string syntax used above, to the `c(...)` vector of hypotheses. If we want to test two nested models using the `anova` function, we just have to run the code to make the two models, and then call `anova` as above. Note that we used `anova` with a lowercase 'a;' the same call to `Anova` gives different results. To be honest, I'm not sure what `Anova` is doing in this case, but at any rate it is of no interest to us. Make sure to use `anova` when trying to test hypotheses like this.

The output of the first is

```
Hypothesis:
poverty = 0

Model 1: crime ~ highschool + poverty + minority
Model 2: restricted model

  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     62  3.4355
2     63  3.4537 -1   -0.0182 0.3285 0.5686
```

and the output of the second is essentially the same (the preamble looks different, but the table is identical). First, note that the p-value above is exactly the same as that in the row for `poverty` of the t-table in the summary. This is because the two hypotheses, which were phrased differently, are mathematically identical. The reader should check that they understand this. Hence, in reality there was no reason to do the F-test above. But the syntax is still useful.

Also, the `linear.hypothesis` function above is nice because we can also test hypotheses other than $\beta_2 = 0$. For instance, if we want to know whether we have reason to beleive $\beta_2 \neq 1$, we can call

```
linear.hypothesis(mod.3, c(’poverty = 1’))
```

The output of this hypothesis test is not easy to get from the `anova` function, and is also not in the T-statistic table we got from `summary(mod.3)`. If we reject, it means that the data provide enough evidence to conclude that $\beta_2 \neq 1$.

Another reason the functions above are useful is because it is not possible to test the null hypothesis that $\beta_1 = \beta_2 = 0$ using a t-statistic. But we can test this using the F-test: either call `linear.hypothesis`, or construct the nested `lm` objects corresponding to the full and restricted model and then use the same `anova` call as above:

```
linear.hypothesis(mod.3, c(’poverty = 0’, ’highschool=0’))

mod.1.crime = lm(crime~minority)
anova(mod.3,mod.1.crime)
```

The output of the `anova` call is

```
Analysis of Variance Table

Model 1: crime ~ highschool + poverty + minority
Model 2: crime ~ minority
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     62  3.4355
2     64  5.1059 -2   -1.6704 15.072 4.629e-06 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

and again essentially the same as the `linear.hypothesis` output. Again, the `linear.hypothesis` is a little more general because it can also test hypotheses in which certain coefficients are set different from 0, which `anova` cannot do.

The syntax above is actually all we need to do any nested hypothesis tests we want: figure out the appropreate hypothesis and call `linear.hypothesis`, or else make the two models and call `anova`.

The other calls available, of the form `anova(model)` and `Anova(model)`, allow us to test lots of hypothesis in a single call, which has advantages, but we could just as easily test the hypotheses individually using the code above. The only reason to use the calls above is to save time, so *DO NOT DO IT* unless you understand what the output will be. The remainder of these notes discusses the output of these calls; you can skip it if you are willing to avoid the calls.

By default, `anova` does Type-I tests. Inputting the code `anova(mod.3)` leads to the following output:

```
Analysis of Variance Table

Response: crime
           Df Sum Sq Mean Sq F value    Pr(>F)
highschool  1 0.0159  0.0159  0.2878    0.5936
poverty     1 1.4156  1.4156 25.5478 4.081e-06 ***
minority    1 3.7887  3.7887 68.3730 1.354e-11 ***
Residuals  62 3.4355  0.0554
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that there are multiply p-values above. Each one corresponds to a *different* hypothesis test. The hypothesis here are nested: the first row tests the model with just an intercept versus the model with an intercept and highschool. The second tests the model with an intercept and highschool against that with everything except minority, and the last tests the model with everything except minority against the full model. Note that the p-values here are all different from what we've seen before, except the very last one, which is the same as the p-value from the t-test for minority at the very biginning. This is because the hypotheses being tested above are all different from what we've already done except the last one. In the last row, we are testing $H_0 : \beta_3 = 0$, which is the same as what the t-test did.

On the other hand if we call `Anova(mod.3)` we get:

```
Anova Table (Type II tests)

Response: crime
           Sum Sq Df F value    Pr(>F)    
highschool 1.0018  1 18.0787 7.269e-05 ***
poverty    0.0182  1  0.3285    0.5686    
minority   3.7887  1 68.3730 1.354e-11 ***
Residuals  3.4355 62                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By default, `Anova` does type-II tests. This means that in each row, we are testing the hypothesis that the variable in that row does not matter, versus the full model. Note that the p-values are the same as in the t-table before. This is because the hypotheses being tested are the same.

A word of caution: I wrote these notes discussing regression, but we are now starting analysis of variance, where there are factors instead of/ as well as numerical covariates. Here the F tests become more important, and the analysis of variance tables will *not* be equivalent to t-tests. The reasons for this are explained in the book, and I will not try to delve into detail here. When working with factors, one should usually ignore t-tests and focus just on F-tests.

Also, the call to `Anova` above is not equivalent to the t-tests if our model includes interactions, unless we change it to `Anova(mod.3,type = 'III')`, which is usually a bad idea. This is because, when testing whether a variable matters, we should usually remove any higher-order interactions involving that variable during the test. This is explained in the book, and may be discussed in more detail later. It is in the sections where he discusses the principle of marginality.