

Activity 15.10

[S2008/S2008-whole8]

The Null Hypothesis is that the explanatory model vectors are unrelated to the response variable. You can create a simulation of the Null Hypothesis by generating random explanatory vectors using the `rand` operator in the model expression.

To start, read in some data frame from which a response variable will be drawn. It can be anything. Here the `net` variable from the `ten-mile-race` data is being used:

```
> running = ISMdata("ten-mile-race.csv")
```

Pick a sample size n that's of interest to you, say $n = 50$:

```
> mydata = shuffle(running, 50)
```

Now fit a model of order m using $m - 1$ random vectors and the intercept. Find the R^2 :

```
> r.squared( lm( net ~ rand(9), data=mysamp))
```

and a model order m that's of interest to you. Here $m = 10$, so 9 random vectors are being used.

```
> r.squared( lm( net ~ rand(9), data=mysamp))
```

```
[1] 0.2295271
```

```
> r.squared( lm( net ~ rand(9), data=mysamp))
```

```
[1] 0.1269188
```

Use a statement like the ones above along with the `do` operator to create a sample of 1000 trials. In each trial, compute the R^2 of a random model with n and m as specified. Use your trials to answer the following questions (you will have to make a different run for each question):

1. When $n = 50$ and $m = 10$, what is a typical value of R^2 ?

Pick the closest: 0.001 0.012 0.05 0.20 0.50 0.75 0.95 Activity 15.10-1

2. When $n = 200$ and $m = 10$, what is a typical value of R^2 ?

Pick the closest: 0.001 0.012 0.05 0.20 0.50 0.75 0.95 Activity 15.10-2

3. When $n = 800$ and $m = 10$, what is a typical value of R^2 ?

Pick the closest: 0.001 0.012 0.05 0.20 0.50 0.75 0.95 Activity 15.10-3

4. Based on your answers to the above, what appears to be the general relationship between R^2 and n ?

- A R^2 is independent of n .
- B R^2 increases with n .
- C R^2 is proportional to $1/n$.
- D R^2 increases with \sqrt{n} .
- E R^2 is proportional to $1/\sqrt{n}$.

Activity 15.10-4

For a whole model, the F statistic is related to R^2 according to the formula F calculate the F statistic directly from your simulated R^2 using an expression like this:

```
> myF = (myR2 / 9) / ((1-myR2)/(50-10))
```

In this statement, $n = 50$, $m = 10$, and myR2 is a set of simulated R^2 values that you have generated. (You will have to create myR2 before the statement will work.)

Generate such simulated values of F under the Null Hypothesis in the same way you did for R^2 . Based on your simulations, answer the following questions:

1. What is a typical value of F for $n = 50$ and $m = 10$?

0.01 0.1 1.0 2.0 5.0 10.0 Activity 15.10-5

2. What is a typical value of F for $n = 200$ and $m = 10$?

0.01 0.1 1.0 2.0 5.0 10.0 Activity 15.10-6

3. What is a typical value of F for $n = 800$ and $m = 10$?

0.01 0.1 1.0 2.0 5.0 10.0 Activity 15.10-7

Pick one of your simulations and plot out a histogram of the simulated F values. Then figure out the degrees of freedom of the numerator and the degrees of freedom of the denominator based on your n and m .

Make a plot that compares the distribution of your simulated version of F (under the Null Hypothesis) with the tabulated F distribution given by the R operator df. You can use these commands, substituting your own simulated set of myF and inserting the appropriate degrees of freedom in place of (9,40):

```
> densityplot( myF, plot.points=FALSE, lwd=3)
> trellis.focus()
> panel.mathdensity( df, args=list(9,40),col='red')
> trellis.unfocus()
```