Hypothesis Tests to Compare Models

Data Analytics and Visualization

Instructor: Debopriya Ghosh

Test of all predictors

Are any of the predictors useful in predicting the response?

- Full model (Ω) : $y = X\beta + \epsilon$, where X is a full-rank $n \times p$ matrix.
- Reduced model (ω) : $y = \mu + \epsilon$, predict y by the mean

We could write the null hypothesis in this case as $H_0 : \beta_1 = ... = \beta_p = 0$, all of them are not useful

Now, $RSS_\Omega = (y - X\hat{\beta})^T(y - X\hat{\beta}) = \hat{\epsilon}^T\hat{\epsilon} =$ RSS

$RSS_\omega = (y - \bar{y})^T(y - \bar{y}) =$ SST

Therefore, $F = \frac{(SST-RSS)/(p)}{RSS/(n-p-1)}$

We'd now refer to $F_{p-1,n-p}$ for a critical value or a p-value. Large values of F would indicate rejection of the null.

Even if we fail to reject the null hypothesis we must still investigate the possibility of non-linear transformations of the variables and of outliers which may obscure the relationship. We may have insufficient data to demonstrate a real effect and so we say "fail to reject" the null rather than "accept" the null. It would be a mistake to conclude that no real relationship exists.

When the null is rejected, it does not imply that the alternative model is the best model. We don't know whether all the predictors are required to predict the response or just some of them. Other predictors might also be added - for example quadratic terms in the existing predictors. Either way, the overall F-test is just the beginning of an analysis and not the end.

DATASET DESCRIPTION: This is an old economic dataset on 50 different countries. These data are averages over 1960-1970. dpi is per-capita disposable income in U.S. dollars; ddpi is the percent rate of change in per capita disposable income; sr is aggregate personal saving divided by disposable income. The percentage population under 15 (pop15) and over 75 (pop75) are also recorded.

```
library(faraway)
data(savings)
savings
```

```
##                   sr pop15 pop75     dpi  ddpi
## Australia      11.43 29.35  2.87 2329.68  2.87
## Austria        12.07 23.32  4.41 1507.99  3.93
## Belgium        13.17 23.80  4.43 2108.47  3.82
## Bolivia         5.75 41.89  1.67  189.13  0.22
## Brazil         12.88 42.19  0.83  728.47  4.56
## Canada          8.79 31.72  2.85 2982.88  2.43
## Chile           0.60 39.74  1.34  662.86  2.67
## China          11.90 44.75  0.67  289.52  6.51
## Colombia        4.98 46.64  1.06  276.65  3.08
## Costa Rica     10.78 47.64  1.14  471.24  2.80
## Denmark        16.85 24.42  3.93 2496.53  3.99
## Ecuador         3.59 46.31  1.19  287.77  2.19
## Finland        11.24 27.84  2.37 1681.25  4.32
## France         12.64 25.06  4.70 2213.82  4.52
## Germany        12.55 23.31  3.35 2457.12  3.44
## Greece         10.67 25.62  3.10  870.85  6.28
## Guatamala       3.01 46.05  0.87  289.71  1.48
## Honduras        7.70 47.32  0.58  232.44  3.19
## Iceland         1.27 34.03  3.08 1900.10  1.12
## India           9.00 41.31  0.96   88.94  1.54
## Ireland        11.34 31.16  4.19 1139.95  2.99
## Italy          14.28 24.52  3.48 1390.00  3.54
## Japan          21.10 27.01  1.91 1257.28  8.21
## Korea           3.98 41.74  0.91  207.68  5.81
## Luxembourg     10.35 21.80  3.73 2449.39  1.57
## Malta          15.48 32.54  2.47  601.05  8.12
## Norway         10.25 25.95  3.67 2231.03  3.62
## Netherlands    14.65 24.71  3.25 1740.70  7.66
## New Zealand    10.67 32.61  3.17 1487.52  1.76
## Nicaragua       7.30 45.04  1.21  325.54  2.48
## Panama          4.44 43.56  1.20  568.56  3.61
## Paraguay        2.02 41.18  1.05  220.56  1.03
## Peru           12.70 44.19  1.28  400.06  0.67
## Philippines    12.78 46.26  1.12  152.01  2.00
## Portugal       12.49 28.96  2.85  579.51  7.48
## South Africa   11.14 31.94  2.28  651.11  2.19
## South Rhodesia 13.30 31.92  1.52  250.96  2.00
## Spain          11.77 27.74  2.87  768.79  4.35
## Sweden          6.86 21.44  4.54 3299.49  3.01
## Switzerland    14.13 23.49  3.73 2630.96  2.70
## Turkey          5.13 43.42  1.08  389.66  2.96
## Tunisia         2.81 46.12  1.21  249.87  1.13
## United Kingdom  7.81 23.27  4.46 1813.93  2.01
## United States   7.56 29.81  3.43 4001.89  2.45
## Venezuela       9.22 46.40  0.90  813.39  0.53
## Zambia         18.56 45.25  0.56  138.33  5.14
## Jamaica         7.72 41.12  1.73  380.47 10.23
## Uruguay         9.24 28.13  2.72  766.54  1.88
## Libya           8.89 43.69  2.07  123.58 16.71
## Malaysia        4.71 47.20  0.66  242.69  5.08
```

First consider a model with all the predictors:

```
lm.fit = lm(sr ~ pop15 + pop75 + dpi + ddpi, data = savings)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2422 -2.6857 -0.2488  2.4280  9.7509
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
## pop15       -0.4611931  0.1446422  -3.189 0.002603 **
## pop75       -1.6914977  1.0835989  -1.561 0.125530
## dpi         -0.0003369  0.0009311  -0.362 0.719173
## ddpi         0.4096949  0.1961971   2.088 0.042471 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.803 on 45 degrees of freedom
## Multiple R-squared:  0.3385, Adjusted R-squared:  0.2797
## F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

We can see directly the result of the test of whether any of the predictors have significance in the model. since, the p-value is extremely small, the null hypotheses $H_0 : \beta_1 = ... = \beta_p = 0$ is rejected.

We can also do it directly using the F-testing formula:

```
 SST = sum((savings$sr-mean(savings$sr))^2)
 RSS = sum(lm.fit$residuals^2)
 F = ((SST-RSS)/(4))/(RSS/(50-4-1))
 F
```

```
## [1] 5.755681
```

Check for the p-value

```
1-pf( F,1,45)
```

```
## [1] 0.02063914
```

Testing just one predictor

Can one particular predictor be dropped from the model?

Null hypothesis would be $H_0 : \beta_i = 0$

RSS_Ω is the RSS for the model with all the predictors of interest (p parameters).

RSS_ω is the RSS for the model with all the above predictors except predictor i.

The F-statistic may be computed using the formula from above.

```
lm.fit1 = lm(sr ~ pop75 + dpi + ddpi, data = savings)
RSS1 = sum(lm.fit1$residuals^2)

F = (RSS1-RSS)/(RSS/(45))
F
```

```
## [1] 10.16659
```

The p-value is computed as:

```
1-pf(10.167,1,45)
```

```
## [1] 0.002602553
```

An alternative approach is to use a t-statistic for testing the hypothesis: $t_i = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)}$

Check for significance using a t distribution with $n - p$ degrees of freedom.

```
 t = sqrt(F)
 2*(1-pt(t,45))
```

```
## [1] 0.002603019
```

Convenient way to compare two nested models is–

```
anova(lm.fit1,lm.fit)
```

```
## Analysis of Variance Table
##
## Model 1: sr ~ pop75 + dpi + ddpi
## Model 2: sr ~ pop15 + pop75 + dpi + ddpi
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     46 797.72
## 2     45 650.71  1    147.01 10.167 0.002603 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing a subspace

Suppose that y is the miles-per-gallon for a make of car and X_j is the weight of the engine and X_k is the weight of the rest of the car. There are also some other predictors. Do we need two weight variables? Can they be replaced by the total weight, $X_j + X_k$?

So if the original model was: $y = \beta_0 + \beta_1 X_1 + ... + \beta_j X_j + \beta_k X_k +,,,+\epsilon$

$y = \beta_0 + \beta_1 X_1 + ... + \beta_j (X_j + X_k) +,,,+\epsilon$, given $\beta_j = \beta_k$

The null hypotheis is: $H_0 : \beta_j = \beta_k$ This defines a linear subspace to which the general F-testing procedure applies.

In the above example, we hypothesize that the effect of young and old people on the savings rate was the same. $H_0 : \beta_{pop15} = \beta_{pop75}$

The null model is: $y = \beta_0 + \beta_1 X_1 + ... + \beta_{pop15}(X_{pop15} + X_{pop75}) +,,,+\epsilon$.We can then compare this to the full model as follows:

```
lm.full =   lm(sr ~ .,savings)
lm.null = lm(sr ~ I(pop15+pop75)+dpi+ddpi, savings)
anova(lm.null,lm.full)
```

```
## Analysis of Variance Table
##
## Model 1: sr ~ I(pop15 + pop75) + dpi + ddpi
## Model 2: sr ~ pop15 + pop75 + dpi + ddpi
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     46 673.63
## 2     45 650.71  1    22.915 1.5847 0.2146
```

The p-value of 0.21 indicates that the null cannot be rejected here meaning that there is not evidence here that young and old people need to be treated separately in the context of this particular model.

Suppose we want to test whether one of the coefficients can be set to a particular value. For example, $H_0 : \beta_{ddpi} = 1$ Here the null model would take the form: $y = \beta_0 + \beta_{pop15} pop15 + \beta_{pop75} pop75 + \beta_{dpi} ddpi + ddpi + \epsilon$

A fixed term in the regression equation is called an offset. We fit this model and compare it to the full:

```
lm.null <- lm(sr ~ pop15+pop75+dpi+offset(ddpi),savings)
anova(lm.null,lm.full)
```

```
## Analysis of Variance Table
##
```

```
## Model 1: sr ~ pop15 + pop75 + dpi + offset(ddpi)
## Model 2: sr ~ pop15 + pop75 + dpi + ddpi
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     46 781.61
## 2     45 650.71  1     130.9 9.0525 0.004286 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that the p-value is small and the null hypothesis here is soundly rejected.

Can we test a hypothesis such as – $H_0 : \beta_j\beta_k = 1$?

No. This hypothesis is not linear in the parameters so we can't use our general method. We'd need to fit a non-linear model.

Permutation testing

For the Galapagos dataset, suppose that the number of species had no relation to the five geographic variables, then the observed response values would be randomly distributed between the islands without relation to the predictors. We then consider what the chance would be under this assumption that an F-statistic would be observed as large or larger than one we actually observed. We compute this exactly by computing the F-statistic for all possible permutations of the response variable and see what proportion exceed the observed Fstatistic. This is a permutation test. If the observed proportion is small, then we must reject the contention that the response is unrelated to the predictors.

Lets apply the permutation test to the savings data. We chose a model with just pop75 and dpi so as to get a p-value for the F-statistic that is not too small.

```r
lm.fit3 = lm(sr ~ pop75+dpi,data = savings)
summary(lm.fit3)
```

```
##
## Call:
## lm(formula = sr ~ pop75 + dpi, data = savings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1571 -3.1835 -0.0844  2.2877 11.9802
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.0566189  1.2904352   5.468  1.7e-06 ***
## pop75        1.3049653  0.7775328   1.678   0.0999 .
## dpi         -0.0003415  0.0010129  -0.337   0.7375
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.334 on 47 degrees of freedom
## Multiple R-squared:  0.1024, Adjusted R-squared:  0.06416
## F-statistic:  2.68 on 2 and 47 DF,  p-value: 0.07906
```

We extract the F-statistic as–

```r
summary(lm.fit3)$fstat
```

```
##     value     numdf     dendf
##  2.679647  2.000000 47.000000
```

We compute the F-statistic for 1000 randomly selected permutations and see what proportion exceed the the F-statistic for the original data:

```
fstats = numeric(1000)
for(i in 1:1000){
   model = lm(sample(sr) ~ pop75+dpi,data = savings)
   fstats[i] = summary(model)$fstat[1]
   }
length(fstats[fstats > 2.6796])/1000
```

```
## [1] 0.079
```

So our estimated p-value using the permutation test is 0.092 which is close to the normal theory based value of 0.0791. Thus it is possible to give some meaning to the p-value when the sample is the population or for samples of convenience although one has to be clear that one's conclusion apply only the particular sample.