Comparing Nested Models

Two regression models are called *nested* if one contains all the predictors of the other, and some additional predictors.

For example, the first-order model in two independent variables,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

is nested within the complete second-order model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon.$$

How to choose between them?

If the models are being considered for making *predictions* about the mean response or about future observations, you could just use PRESS or P^2.

But you may be interested in whether the simpler model is adequate as a description of the relationship, and not necessarily in whether it gives better predictions.

A single added predictor

If the larger model has just one more predictor than the smaller model, you could just test the significance of the one additional coefficient, using the t-statistic.

Multiple added predictors

When the models differ by $r > 1$ added predictors, you cannot compare them using t-statistics.

The conventional test is based on comparing the regression sums of squares for the two models: the *general regression test*, or the *extra sum of squares* test.

Write $SS_{R,\text{reduced}}$ and $SS_{R,\text{full}}$ for the regression sums of squares of the two models, where the "reduced" model is nested within the "full" model.

The *extra sum of squares* is

$$SS_{R,\text{extra}} = SS_{R,\text{full}} - SS_{R,\text{reduced}}$$

and if this is large, the r additional predictors have explained a substantial additional amount of variability.

We test the null hypothesis that the added predictors all have zero coefficients using the F-statistic

$$F_{\text{obs}} = \frac{SS_{R,\text{extra}}/r}{MS_{E,\text{full}}}.$$

In R

The R function `anova()` (not to be confused with `aov()`) implements the extra sum of squares test:

```
wireBondLm2 <- lm(Strength ~ Length  + I(Length^2) + Height,
                  wireBond)
wireBondLm3 <- lm(Strength ~ Length  + I(Length^2) + Height +
                  I(Height^2) + I(Length * Height), wireBond)
anova(wireBondLm1, wireBondLm3)
```

It can also compare a sequence of more than two nested models:

```
anova(wireBondLm1, wireBondLm2, wireBondLm3)
```

Note

Because

$$SS_R = SS_T - SS_E$$

and SS_T is the same for all models, the extra sum of squares can also be written

$$SS_{R,\text{extra}} = SS_{E,\text{reduced}} - SS_{E,\text{full}}$$

That is, the extra sum of squares is also the amount by which the *residual* sum of squares is *reduced* by the additional predictors.

Note

The nested model F-test can also be used when $r = 1$, and is equvalent to the $|t|$-test for the added coefficient, because $F = t^2$.

Indicator Variables

Recall that an *indicator variable* is a variable that takes only the values 0 and 1.

A single indicator variable divides the data into two groups, and is a quantitative representation of a categorical factor with two levels.

To represent a factor with $a > 2$ levels, you need $a - 1$ indicator variables.

Recall the paper strength example: the factor is Hardwood Concentration, with levels 5%, 10%, 15%, and 20%.

Define indicator variables

$$x_1 = \begin{cases} 1 & \text{for 5\% hardwood} \\ 0 & \text{otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{for 10\% hardwood} \\ 0 & \text{otherwise} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{for 15\% hardwood} \\ 0 & \text{otherwise} \end{cases}$$

$$x_4 = \begin{cases} 1 & \text{for 20\% hardwood} \\ 0 & \text{otherwise} \end{cases}$$

Consider the regression model

$$Y_i = \beta_0 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \epsilon_i.$$

The interpretation of β_0 is, as always, the mean response when $x_2 = x_3 = x_4 = 0$; in this case, that is for the remaining (baseline) category, 5% hardwood.

For 10% hardwood, $x_2 = 1$ and $x_3 = x_4 = 0$, so the mean response is $\beta_0 + \beta_2$; the interpretation of β_2 is the *difference* between the mean responses for 10% hardwood and the baseline category.

Similarly β_3 is the difference between 15% hardwood and the baseline category, and β_4 is the difference between 20% hardwood and the baseline category.

So the interpretations of β_0, β_2, β_3, and β_4 are exactly the same as the interpretations of μ, τ_2, τ_3, and τ_4 in the one-factor model

$$Y_{i,j} = \mu + \tau_i + \epsilon_{i,j}.$$

The factorial model may be viewed as a special form of regression model with these indicator variables as constructed predictors.

Modern statistical software fits factorial models using regression with indicator functions.

Combining Categorical and Quantitative Predictors

Example: surface finish

The response Y is a measure of the roughness of the surface of a metal part finished on a lathe.

Factors

- RPM;
- Type of cutting tool (2 types, 302 and 416).

Begin with the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where x_1 is RPM and x_2 is the indicator for type 416:

```
parts <- read.csv("Data/Table-12-11.csv")
pairs(parts)
parts$Type <- as.factor(parts$Type)
summary(lm(Finish ~ RPM + Type, parts))
```

```
Call:
lm(formula = Finish ~ RPM + Type, data = finish)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9546 -0.5039 -0.1804  0.4893  1.5188

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.276196   2.091214   6.827 2.94e-06 ***
RPM          0.141150   0.008833  15.979 1.13e-11 ***
Type416    -13.280195   0.302879 -43.847  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.6771 on 17 degrees of freedom
Multiple R-squared:  0.9924,Adjusted R-squared:  0.9915
F-statistic:  1104 on 2 and 17 DF,  p-value: < 2.2e-16
```

For tool type 302, $x_2 = 0$, so the fitted equation is

$$\hat{y} = 14.276 + 0.141 \times \text{RPM}$$

while for tool type 416, $x_2 = 1$, and the fitted equation is

$$\begin{aligned}\hat{y} &= 14.276 - 13.280 + 0.141 \times \text{RPM} \\ &= 0.996 + 0.141 \times \text{RPM}\end{aligned}$$

We are essentially fitting *parallel* straight lines against RPM for the two tool types: the same slope, but different intercepts.

We could also allow the slopes to be different:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{1,2} x_1 x_2 + \epsilon.$$

In this model, the slopes versus RPM are β_1 for type 302 and $\beta_1 + \beta_{1,2}$ for type 416.

```
summary(lm(Finish ~ RPM * Type, parts))
```

```
Call:
lm(formula = Finish ~ RPM * Type, data = finish)

Residuals:
     Min       1Q   Median       3Q      Max
-0.68655 -0.44881 -0.07609  0.30171  1.76690

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.50294    2.50430   4.593   0.0003 ***
RPM          0.15293    0.01060  14.428 1.37e-10 ***
Type416     -6.09423    4.02457  -1.514   0.1495
RPM:Type416 -0.03057    0.01708  -1.790   0.0924 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.6371 on 16 degrees of freedom
Multiple R-squared:  0.9936,Adjusted R-squared:  0.9924
F-statistic: 832.3 on 3 and 16 DF,  p-value: < 2.2e-16
```

We do not reject the null hypothesis that the interaction term has a zero coefficient, so the fitted lines are not significantly different from parallel.