Testing a Submodel

In general, we can construct an F-Test to compare a model that contains a subset of a collection of regressors to the model that contains all of the regressors.

\( H_0 \): Submodel is adequate.
\( H_1 \): Full model explains more of the variability in the response.

- Let \( X_F \) and \( X_S \) represent the model matrices for the full model and the submodel respectively. Then the projection matrices for these models are:

\[
P_F = X_F (X_F^t X_F)^{-1} X_F^t
\]

\[
P_S = X_S (X_S^t X_S)^{-1} X_S^t
\]
The Added Variable F-Test

To construct the F-test we need the projection of the response vector \( y \) on to the orthogonal compliment of \( \text{colsp}(X_S) \) in \( \text{colsp}(X_F) \): The projection matrix is given by \( P_F - P_S \).

- If none of the extra variables in the full model, provide additional information for predicting the response then projecting \( y \) onto this subspace is equivalent to just projecting \( \epsilon \).

- However if some of them do provide additional information then the squared length of the image should tend to be larger.
The Added Variable F-Test

Suppose the full model contains $k$ regressors and the submodel contains $p$ of these regressors ($p < k$):

\[
F\text{-stat} = \frac{\| (P_F - P_S) y \|^2 / (k - p)}{\| (I - P_F) y \|^2 / (n - k - 1)}
= \frac{y^t (P_F - P_S) y / (k - p)}{y^t (I - P_F) y / (n - k - 1)}
\]

\[
p\text{-value} = \Pr(F_{k-p, n-k-1} \geq \text{F-stat})
\]
Example from Catheter Data

For the catheter data we could test adding weight to the model that already contains height.

```r
> Xs<-cbind(1,ht)
> Xf<-cbind(1,ht,wt)
> Ps<-Xs%*%solve(t(Xs)%*%Xs)%*%t(Xs)
> Pf<-Xf%*%solve(t(Xf)%*%Xf)%*%t(Xf)
> num<-(t(y)%*%(Pf-Ps)%*%y)/1
> denom<-(t(y)%*%(diag(12)-Pf)%*%y)/9
> fstat<-num/denom
> fstat
   [,1]
[1,] 1.457787
> pval<-1-pf(fstat,1,9)
> pval
   [,1]
[1,] 0.2580548
```
Both forward selection and backward elimination involve doing a sequence of added variable tests.

- Each step involves doing an added variable test for the addition or deletion of one variable.

- The t-tests from the `summary(name.lm)` give the same p-values as the added variable F-test that looks at eliminating one variable at a time.
The **anova** Command Output

The output from anova applied to an `lm` object, produces results for a sequence of F-tests.

```r
> anova(catheter.lm)
Response: ca

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ht</td>
<td>1</td>
<td>586.38</td>
<td>586.38</td>
<td>41.0760</td>
</tr>
<tr>
<td>wt</td>
<td>1</td>
<td>20.81</td>
<td>20.81</td>
<td>1.4578</td>
</tr>
<tr>
<td>Residuals</td>
<td>9</td>
<td>128.48</td>
<td>14.28</td>
<td></td>
</tr>
</tbody>
</table>
```

- The `ht` line tests adding `ht` to the null model.
- The `wt` line tests adding `wt` to the model that contains `ht`. 
The `anova` Command Output II

There is a smallish wrinkle: the `anova` command always uses the projection on to the error space for the model that contains all of the variables as the basis for its denominator. Note that this is not the same as using the “full model” error space.

- Suppose $P_1$ is the projection matrix for the model that just contains $ht$ and $P_2$ is the projection matrix for the model that contains both $ht$ and $wt$. Consider testing whether $ht$ should be added to the null model (projection matrix $P_0$).

\[
\text{anova test: } F\text{-stat} = \frac{\| (P_1 - P_0)y \|^2 / 1}{\| (I - P_2)y \|^2 / 9} \quad \text{p-value} = P (F_{1,9} \geq F\text{-stat})
\]

\[
\text{added variable: } F\text{-stat} = \frac{\| (P_1 - P_0)y \|^2 / 1}{\| (I - P_1)y \|^2 / 10} \quad \text{p-value} = P (F_{1,10} \geq F\text{-stat})
\]
The anova Command Output

As variables are added sequentially, the order of variables in the model statement are important unless the vectors for the explanatory variables are orthogonal to each other – this will usually only occur for a designed experiment.

```r
> catheterA.lm <- lm(ca~wt+ht, data=catheter.df)
> anova(catheterA.lm)
Response: ca

              Df Sum Sq Mean Sq   F value Pr(>F) 
wt             1 601.88 601.88 42.16234  0.0001123 ***
ht             1  5.31  5.31  0.37207  0.5570028 
Residuals      9 128.48 14.28
```

9 / 32
Categorical Explanatory Variables

If an explanatory variable is categorical (a factor) rather than numerical then:

- We use indicator ("dummy") variables in the linear model.
- For each factor, the number of additional indicator variables is equal to the number of levels minus one.
- As far our “vector space approach” to fitting the linear model, these dummy variable vectors are treated the same as our other explanatory variable vectors.
Categorical Explanatory Variables Example

One theory regarding memory is that verbal material is remembered as a function of the degree to which it was processed when it was initially presented. Eysenck (1974) randomly assigned 50 younger subjects and 50 older (between 55 and 65 years old) to one of five learning groups who were each given a list of 27 words which they were later asked to recall:

- Counting group: counted the number of letters in each word.
- Rhyming group: thought of a word that rhymed with each word.
- Adjective group: thought of an adjective for each word.
- Imagery group: formed vivid mental images of each word.
- Intentional group: was asked to memorize the list for later recall.
Two categorical explanatory variables:

**Age**: younger or older.

**Method**: counting, rhyming, adjective, imagery or intentional.

- we want the mean response to depend on the levels of **Age** and **Method**.
- to accomplish this we need to incorporate a set of indicator variables in the regression model.
- indicator variables correspond to vectors that indicate the levels of the factors - often (but not necessarily) 0/1 binary vectors.
Dummy Vectors

There are lots of options for setting up the “dummy” vectors for orthogonal factors. The default in “R” is to use the baseline model.

▶ One level is designated as the baseline – the mean for this level is captured by the coefficient for the intercept.
▶ Each of the other levels is compared to the baseline level using a binary dummy vector that has a 1 when the specified level occurs and a 0 for all other levels.
## Memory Dataframe in R


ger eys.df

<table>
<thead>
<tr>
<th>Age</th>
<th>Method</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Younger</td>
<td>Counting</td>
<td>8</td>
</tr>
<tr>
<td>Younger</td>
<td>Counting</td>
<td>6</td>
</tr>
<tr>
<td>Younger</td>
<td>Counting</td>
<td>4</td>
</tr>
<tr>
<td>Younger</td>
<td>Counting</td>
<td>6</td>
</tr>
<tr>
<td>Younger</td>
<td>Counting</td>
<td>7</td>
</tr>
<tr>
<td>Older</td>
<td>Intentional</td>
<td>11</td>
</tr>
<tr>
<td>Older</td>
<td>Intentional</td>
<td>14</td>
</tr>
<tr>
<td>Older</td>
<td>Intentional</td>
<td>15</td>
</tr>
<tr>
<td>Older</td>
<td>Intentional</td>
<td>11</td>
</tr>
<tr>
<td>Older</td>
<td>Intentional</td>
<td>11</td>
</tr>
</tbody>
</table>
Baseline Model Indicator Variables for Method

The contrasts command can be used to see the default indicator variables used by R.

> contrasts(eyes.df$Method)

<table>
<thead>
<tr>
<th></th>
<th>Counting</th>
<th>Imagery</th>
<th>Intentional</th>
<th>Rhyming</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjective</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Counting</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Imagery</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Intentional</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Rhyming</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Baseline Model Indicator Variables for Method

Suppose that we fit the regression model that just uses Method as an explanatory variable. Then the model matrix has 5 columns:

- A intercept column (all 1’s).
- Four indicator columns as defined on the previous slide.
- The intercept represents the mean response for the baseline level (Adjective).
- Each of the other coefficients represents the difference between one of the remaining levels and the baseline level.
### Fitted Baseline Model

```r
eys.lm <- lm(Words ~ Method, data=eys.df)
> summary(eys.lm)

Coefficients:

|               | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------|----------|------------|---------|----------|
| (Intercept)   | 12.900   | 0.779      | 16.561  | < 2e-16  *** |
| MethodCounting| -6.150   | 1.102      | -5.583  | 2.24e-07 *** |
| MethodImagery | 2.600    | 1.102      | 2.360   | 0.0203 * |
| MethodIntentional | 2.750   | 1.102      | 2.496   | 0.0143 * |
| MethodRhyming | -5.650   | 1.102      | -5.129  | 1.54e-06 *** |

Residual standard error: 3.484 on 95 degrees of freedom
Multiple R-squared: 0.5679, Adjusted R-squared: 0.5497
F-statistic: 31.21 on 4 and 95 DF,  p-value: < 2.2e-16
```
Output from \texttt{anova} Function in \textit{R}

The \texttt{anova} function in R simultaneously tests the inclusion of all of the indicator variables for Method:

\begin{verbatim}
> anova(eys.lm)
Analysis of Variance Table

  Response: Words
         Df Sum Sq  Mean Sq   F value Pr(>F) 
Method     4 1514.9  378.744  31.2093  <2e-16 ***
Residuals 95 1152.8   12.140

---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1
\end{verbatim}
Equivalent Models

Any two models that define the same model space will produce the same set of fitted values.

- We can take linear combinations of the explanatory variable vectors to create a new basis that spans the same model space.

- For our memory example, any set of four vectors for Method that (along with the intercept vector) span the same vector space as those give
Suppose we define our own indicator variables in $R$ as follows:

```r
> c1<-c(0,1,1,1,1)
> c2<-c(0,0,1,1,1)
> c3<-c(0,0,0,1,1)
> c4<-c(0,0,0,0,1)
> cmat<-cbind(c1,c2,c3,c4)
> eys.df2<-eys.df
> eys.df2$Method<-C(eys.df2$Method,cmat)
```
We end up with columns in the model matrix which are linear combinations of those used for the baseline model:

```r
> contrasts(eys.df2$Method)
   c1  c2  c3  c4
Adjective  0  0  0  0
Counting   1  0  0  0
Imagery    1  1  0  0
Intentional 1  1  1  1
Rhyming    1  1  1  1
```
The fitted coefficients are different and so care must be taken to interpret them appropriately:

```r
> summary(eyes.lm2)
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    12.900     0.779  16.561  < 2e-16 ***
Methodc1       -6.150     1.102  -5.583  2.24e-07 ***
Methodc2        8.750     1.102   7.943  3.97e-12 ***
Methodc3        0.150     1.102   0.136    0.892
Methodc4       -8.400     1.102  -7.625  1.84e-11 ***
```

Residual standard error: 3.484 on 95 degrees of freedom
Multiple R-squared: 0.5679, Adjusted R-squared: 0.5497
F-statistic: 31.21 on 4 and 95 DF,   p-value: < 2.2e-16
Output for `anova`

However the ANOVA table is exactly the same (why?).

```r
> anova(eys.lm2)
Analysis of Variance Table

Response: Words

    Df Sum Sq Mean Sq F value Pr(>F)
Method   4 1514.9 378.74  31.209  < 2.2e-16 ***
Residuals 95 1152.8 12.14

---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
```
Consider fitting an alternative linear model which has model matrix as $W = XA$ where $A$ is a nonsingular $(k + 1) \times (k + 1)$ matrix.

- The vector of fitted values for this alternative model will be identical to that for the original model since the projection matrices for the two models will be identical.

\[
W (W^t W)^{-1} W^t = X (X^t X)^{-1} X^t
\]

- The estimated coefficients are related as:

\[
\hat{\beta}_W = A^{-1} \hat{\beta}_X
\]
Multi-way ANOVA

For the memory data there were two categorical variables: Method and Age. Putting these both into our regression model means that the mean response is a function of both of these factors. This can be done in two ways: the variables interact or the variables do not interact.

- If the interaction term is not included then the impact that changing the level of Method has on the mean response is the same for all levels of Age (and vice versa).
- If the interaction term is included then the impact of Method depends on the level of Age (and vice versa).
Baseline Model Column

If we use the baseline model then we need one column for Age defined as:

\[
> \text{contrasts(eyes.df$Age)} \\
| \text{Older} | 0 \\
| \text{Younger} | 1 \\
\]

26 / 32
Interaction Indicator Columns

The indicator columns for the interaction are defined as follows for the base line model:

<table>
<thead>
<tr>
<th></th>
<th>Yo:Co</th>
<th>Yo:Im</th>
<th>Yo:In</th>
<th>Yo:Rh</th>
</tr>
</thead>
<tbody>
<tr>
<td>Older:Adjective</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Older:Counting</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Older:Imagery</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Older:Intentional</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Older:Rhyming</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Younger:Adjective</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Younger:Counting</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Younger:Imagery</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Younger:Intentional</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Younger:Rhyming</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Multi-way ANOVA

The output from `anova` indicates the interaction model is appropriate:

```r
> eys.lm3 <- lm(Words ~ Age * Method, data = eys.df)
> anova(eys.lm3)
```

Analysis of Variance Table

Response: Words

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>1</td>
<td>240.25</td>
<td>240.25</td>
<td>29.94</td>
<td>3.981e-07 ***</td>
</tr>
<tr>
<td>Method</td>
<td>4</td>
<td>1514.94</td>
<td>378.74</td>
<td>47.19</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>Age:Method</td>
<td>4</td>
<td>190.30</td>
<td>47.58</td>
<td>5.93</td>
<td>0.0002793 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>90</td>
<td>722.30</td>
<td>8.03</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Estimated Coefficients

| Coefficient                  | Estimate | Std. Error | t value | Pr(>|t|)   |
|------------------------------|----------|------------|---------|-----------|
| (Intercept)                  | 11.0000  | 0.8959     | 12.279  | < 2e-16   *** |
| AgeYounger                   | 3.8000   | 1.2669     | 2.999   | 0.00350   ** |
| MethodCounting               | -4.0000  | 1.2669     | -3.157  | 0.00217   ** |
| MethodImagery                | 2.4000   | 1.2669     | 1.894   | 0.06139   . |
| MethodIntentional            | 1.0000   | 1.2669     | 0.789   | 0.43201   |
| MethodRhyming                | -4.1000  | 1.2669     | -3.236  | 0.00170   ** |
| AgeYounger:MethodCounting    | -4.3000  | 1.7917     | -2.400  | 0.01846   * |
| AgeYounger:MethodImagery     | 0.4000   | 1.7917     | 0.223   | 0.82385   |
| AgeYounger:MethodIntentional | 3.5000   | 1.7917     | 1.953   | 0.05387   . |
| AgeYounger:MethodRhyming     | -3.1000  | 1.7917     | -1.730  | 0.08702   . |
Estimated Means

So how to we go from these estimated coefficients to the estimated mean response for each combination of Age and Method?
Wrap-Up

The additional lectures for STATS 762 have presented a geometric interpretation of the linear model.

- Think of the response and the explanatory variables as being vectors in $R^n$.
- Different models define different subspaces of $R^n$.
- Model fitting (using least squares) is equivalent to orthogonally projecting the response vector onto a subspace defined by a model.
- Degrees of freedom are equivalent to the dimensions of subspaces.
Good Luck

Good luck with the final exam.

- For the final exam, there will one question on this material.