Partial $F$ Test, Multiple testing, Out of Sample Prediction

Max H. Farrell
The University of Chicago Booth School of Business
Modeling Building

How do we know which $X$ variables to include?

▶ Are any important to our study?
▶ What variables does the subject-area knowledge demand?
▶ Can the data help us decide?

Next two classes address these questions.

Today we start with a simple approach: $F$-testing.

▶ How does regression 1 compare to regression 2?
▶ Limitations make for important lessons.
  ▶ Multiple testing
  ▶ Always need human input!
Partial $F$ Test

Pick up where we left off: how employee ratings of their supervisor relate to performance metrics.

The Data:

- **Y**: Overall rating of supervisor
- **X1**: Handles employee complaints
- **X2**: Opportunity to learn new things
- **X3**: Does not allow special privileges
- **X4**: Raises based on performance
- **X5**: Overly critical of performance
- **X6**: Rate of advancing to better jobs
> attach(supervisor)
> bosslm <- lm(Y ~ X1 + X2 + X3 + X4 + X5 + X6)
> summary(bosslm)

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) | (Intercept) | 10.78708 | 11.58926 | 0.931 | 0.361634 |
|----------|------------|---------|----------|-------------|----------|----------|-------|---------|
| X1 | 0.61319 | 0.16098 | 3.809 | 0.000903 | *** |
| X2 | 0.32033 | 0.16852 | 1.901 | 0.069925 | . |
| X3 | -0.07305 | 0.13572 | -0.538 | 0.595594 |
| X4 | 0.08173 | 0.22148 | 0.369 | 0.715480 |
| X5 | 0.03838 | 0.14700 | 0.261 | 0.796334 |
| X6 | -0.21706 | 0.17821 | -1.218 | 0.235577 |

Residual standard error: 7.068 on 23 degrees of freedom
Multiple R-squared:  0.7326,  Adjusted R-squared:  0.6628
F-statistic:  10.5 on 6 and 23 DF,  p-value:  1.24e-05
The $F$ test says that the regression as a whole is worthwhile. But it looks (from the $t$-statistics and $p$-values) as though only $X_1$ and $X_2$ have a significant effect on $Y$.

What about a reduced model with only these two $X$’s?

```r
> summary(bosslm2 <- lm(Y ~ X1 + X2))
```

Coefficients: Abbreviated output:

|                  | Estimate | Std. Error | t value | Pr(>|t|)  |
|------------------|----------|------------|---------|-----------|
| (Intercept)      | 9.8709   | 7.0612     | 1.398   | 0.174     |
| X1               | 0.6435   | 0.1185     | 5.432   | 9.57e-06 *** |
| X2               | 0.2112   | 0.1344     | 1.571   | 0.128     |

Residual standard error: 6.817 on 27 degrees of freedom
Multiple R-squared: 0.708, Adjusted R-squared: 0.6864
F-statistic: 32.74 on 2 and 27 DF, p-value: 6.058e-08
The full model (6 covariates) has $R^2_{\text{full}} = 0.733$, while the second model (2 covariates) has $R^2_{\text{base}} = 0.708$.

Is this difference worth 4 extra covariates?

The $R^2$ will always increase as more variables are added

- If you have more $b$'s to tune, you can get a smaller SSE.
- Least squares is content fit “noise” in the data.
- This is known as overfitting.

More parameters will always result in a “better fit” to the sample data, but will not necessarily lead to better predictions.

... And remember the coefficient interpretation changes.
Partial $F$-test

At first, we were asking:

“Is this regression worthwhile?”

Now, we’re asking:

“Is it useful to add extra covariates to the regression?”

You **always** want to use the simplest model possible.

- Only add covariates if they are truly informative.
- I.e., only if the extra complexity is useful.
Consider the regression model

\[ Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{d_{\text{base}}} X_{d_{\text{base}}} \\
+ \beta_{d_{\text{base}}+1} X_{d_{\text{base}}+1} + \cdots + \beta_{d_{\text{full}}} X_{d_{\text{full}}} + \varepsilon \]

where

- \( d_{\text{base}} \) is the \# of covariates in the base (small) model, and
- \( d_{\text{full}} > d_{\text{base}} \) is the \# in the full (larger) model.

The partial \( F \)-test is concerned with the hypotheses

\[ H_0 : \beta_{d_{\text{base}}+1} = \beta_{d_{\text{base}}+2} = \cdots = \beta_{d_{\text{full}}} = 0 \]
\[ H_1 : \text{at least one } \beta_j \neq 0 \text{ for } j > d_{\text{base}}. \]
New test statistic:

$$f_{\text{Partial}} = \frac{(R_{\text{full}}^2 - R_{\text{base}}^2)/(d_{\text{full}} - d_{\text{base}})}{(1 - R_{\text{full}}^2)/(n - d_{\text{full}} - 1)}$$

- Big $f$ means that $R_{\text{full}}^2 - R_{\text{base}}^2$ is statistically significant.
- Big $f$ means that at least one of the added $X$’s is useful.
As always, this is super easy to do in R!

```r
> anova(bosslm2, bosslm)
Analysis of Variance Table

Model 1: Y ~ X1 + X2
Model 2: Y ~ X1 + X2 + X3 + X4 + X5 + X6

<table>
<thead>
<tr>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1254.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1149.0</td>
<td>4</td>
<td>105.65</td>
<td>0.5287</td>
<td>0.7158</td>
</tr>
</tbody>
</table>

A *p*-value of 0.71 is not significant, so we stick with the null hypothesis and assume the base (2 covariate) model.

Partial-$F$ is a fine way to compare two different regressions. But what if we have more?
Case study in interaction

Use census data to explore the relationship between log wage rate \((\log(\text{income}/\text{hours}))\) and age—a proxy for experience.

We look at people earning \(>$5000\), working \(>500\) hrs, and \(<60\) years old.
A discrepancy between mean $\log(WR)$ for men and women.

- Female wages flatten at about 30, while men’s keep rising.

```r
> men <- sex=='M'
> malemean <- tapply(log.WR[men], age[men], mean)
> femalemean <- tapply(log.WR[!men], age[!men], mean)
```
The most simple model has

$$E[\log(WR)] = 2 + 0.016 \cdot \text{age}. $$

```r
> wagereg1 <- lm(log.WR ~ age)
```

You get one line for both men and women.
Add a sex effect with
\[E[\log(\text{WR})] = 1.9 + 0.016 \cdot \text{age} + 0.2 \cdot 1_{[\text{sex}=\text{M}]}.\]

> wagereg2 <- lm(log.WR ~ age + sex)

\[\begin{array}{cc}
20 & 30 & 40 & 50 & 60 \\
2.0 & 2.2 & 2.4 & 2.6 & 2.8 & 3.0
\end{array}\]

The male wage line is shifted up from the female line.
With interactions

\[ \mathbb{E}[\log(\text{WR})] = 2.1 + 0.011 \cdot \text{age} + (-0.13 + 0.009 \cdot \text{age}) \mathbb{1}_{[\text{sex}=M]} \cdot \]

\[
> \text{wagereg3 <- lm(log.WR ~ age*sex)}
\]

▶ The interaction term gives us different slopes for each sex.
& quadratics ...

\[ E[\log(WR)] = 0.9 + 0.077 \cdot \text{age} - 0.0008 \cdot \text{age}^2 + (-0.13 + 0.009 \cdot \text{age})\mathbb{1}_{[\text{sex}=M]} . \]

\[ \text{wagereg4} \leftarrow \text{lm}(\log.WR \sim \text{age*sex + age2}) \]

\[ \text{age}^2 \text{ allows us to capture a nonlinear wage curve.} \]
Finally, add an interaction term on the curvature ($age^2$)

$$E[\log(WR)] = 1 + .07 \cdot age - .0008 \cdot age^2 + (.02 \cdot age - .00015 \cdot age^2 - .34) \mathbb{1}_{[sex=M]}.$$  

\[ \text{wagereg5} <- \text{lm(} \log.WR \sim age*sex + age2*sex \text{)} \]

This model provides a generally decent looking fit.
We could also consider a model that has an interaction between age and edu.

▶ `reg <- lm(log.WR ~ edu*age)`

Maybe we don’t need the age main effect?

▶ `reg <- lm(log.WR ~ edu*age - age)`

Or perhaps all of the extra edu effects are unnecessary?

▶ `reg <- lm(log.WR ~ edu*age - edu)`

Which of these is the best?
Model Selection

Our job this lecture and next is to decide which $X$ variables should be in our model.

- A good model summarizes the data but does not overfit.
- A good model answers the question at issue.
  - Better predictions don’t matter if the model doesn’t answer the question.

Last week: a good regression model meets the assumptions.

- Especially important when the goal is inference/relationships.

Next week we will also discuss when a regression is causal.

- A causal model is only good when it meets even more assumptions.
What is the **goal**?

1. Relationship-type questions and inference?
   - Are women paid differently than men on average?
     > \texttt{lm(log.WR \sim sex)}
   - Does age/experience differently affect men and women?
     > \texttt{lm(log.WR \sim age*sex - sex)}
   - No other models matter

2. Data summarization?
   - Is matching the trends enough?
     - In the census data, we matched the blue/pink curves well with a simple(?) model.

3. Prediction?
   - Need an **objective** criterion
We need a method for selecting a final regression specification.

Why not include all variables and be done with it?

- Bad forecasts
- Impossible to interpret
- Bad decision making

Over-fit ⇒ less general model
Overfitting

We have already seen overfitting twice:

1. **Week 3**: $R^2 \uparrow$ as more variables went into MLR
   
   ```r
   > c(summary(trucklm1)$r.square, summary(trucklm3)$r.square, 
     +     summary(trucklm6)$r.square)
   [1] 0.021 0.511 0.693
   ```

2. **Week 4**: Classification error ↓ as more variables into logit

   full history     empty
   0.214 0.283 0.300

Fitting the data at hand better and better

...but getting worse at predicting the next observation.

How can we use the data to pick the model without relying on the data too much?
We need a method that is **disciplined** before we see the data.

**First solution:** Partial $F$ test

- Objective criteria
- Rigorously grounded
- Significance level pre-set

But we’re going to see some big **downsides**.

- These shortfalls have important lessons.
- General messages to carry with you.
Example: Back to the census wage data.

Matching the curves gave us

$$E[\log(WR) \mid \text{age, sex}] =$$

$$1 + 0.07 \text{age} - 0.0008 \text{age}^2 + (0.02 \text{age} - 0.00015 \text{age}^2 - 0.34) \mathbf{1}_{[\text{sex=M}]}.$$

But there were other possible variables:

- Education: 9 levels from none to PhD.
- Marital status: married, divorced, separated, or single.
- Race: white, black, Asian, other.

Should we consider other main effects / interactions?
> summary(wagereg5)

Call:
lm(formula = log.WR ~ age * sex + age2 * sex)

Residuals:
     Min       1Q   Median       3Q      Max
-2.3907 -0.3747 -0.0040  0.3480  3.3820

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.019e+00  7.078e-02 14.395  < 2e-16 ***
age          6.975e-02  3.814e-03 18.290  < 2e-16 ***
sexM        -3.391e-01  9.464e-02  -3.584  0.00034 ***
age:sexM     2.078e-02  5.101e-03   4.074 4.63e-05 ***
sexM:age2    -1.548e-04  6.526e-05  -2.373  0.01767 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5868 on 25397 degrees of freedom
Multiple R-squared:  0.1315,  Adjusted R-squared:  0.1313
F-statistic: 769.1 on 5 and 25397 DF,  p-value: < 2.2e-16
> summary(wagereg6 <- lm(log.WR ~ age*sex + age2*sex + ., data=Wages))

Coefficients: ## output abbreviated

Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.196e+00 6.744e-02 17.737 < 2e-16 ***
age 4.657e-02 3.549e-03 13.123 < 2e-16 ***
sexM -2.133e-01 8.594e-02 -2.482 0.01306 *
age2 -4.832e-04 4.510e-05 -10.715 < 2e-16 ***
raceAsian 1.397e-02 1.860e-02 0.751 0.45267
raceBlack -3.165e-02 1.134e-02 -2.791 0.00525 **
raceNativeAmerican -7.479e-02 3.824e-02 -1.956 0.05048 .
raceOther -8.112e-02 1.338e-02 -6.063 1.36e-09 ***
maritalDivorced -6.981e-02 1.066e-02 -6.549 5.91e-11 ***
maritalSeparated -1.381e-01 1.612e-02 -8.563 < 2e-16 ***
maritalSingle -1.065e-01 9.413e-03 -11.316 < 2e-16 ***
maritalWidow -1.502e-01 3.213e-02 -4.674 2.98e-06 ***
hsTRUE 1.499e-01 1.157e-02 12.947 < 2e-16 ***
assocTRUE 3.111e-01 1.146e-02 27.157 < 2e-16 ***
collTRUE 6.082e-01 1.278e-02 47.602 < 2e-16 ***
gradTRUE 7.970e-01 1.498e-02 53.203 < 2e-16 ***
age:sexM 1.876e-02 4.631e-03 4.051 5.12e-05 ***
sexM:age2 -1.721e-04 5.927e-05 -2.903 0.00369 **
Is it worthwhile to add all the main effects?

```r
> anova(wagereg5, wagereg6)
Analysis of Variance Table

Model 1: log.WR ~ age * sex + age2 * sex
Model 2: log.WR ~ age * sex + age2 * sex
   + (age + age2 + sex + race +
      marital + hs + assoc + coll + grad)

Res.Df   RSS  Df Sum of Sq     F  Pr(>F)
1 25397 8744.8
2 25385 7187.4 12 1557.4 458.37 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

⇒ The new variables are significant!
```
Bring interactions of age with race and education:

```r
> wagereg7 <- lm(log.WR ~ age*sex + age2*sex + marital +
+ (hs+assoc+coll+grad)*age + race*age , data=Wages)
> anova(wagereg6, wagereg7)
Analysis of Variance Table

Model 1: log.WR ~ age * sex + age2 * sex + (age
  + age2 + sex + race + marital + hs + assoc + coll + grad)
Model 2: log.WR ~ age * sex + age2 * sex + marital
  + (hs + assoc + coll + grad) * age + race * age

Res.Df RSS Df Sum of Sq F Pr(>F)
1 25385 7187.4
2 25377 7163.7 8 23.656 10.475 8.891e-15 ***

⇒ The new variables are significant too!
```
Three way interaction!

> wagereg8 <- lm(log.WR ~ race*age*sex + age2*sex + marital +
+ (hs+assoc+coll+grad)*age, data=Wages)
> anova(wagereg7, wagereg8)

Analysis of Variance Table

Model 1: log.WR ~ age * sex + age2 * sex + marital
   + (hs + assoc + coll + grad) * age + race * age
Model 2: log.WR ~ race * age * sex + age2 * sex + marital
   + (hs + assoc + coll + grad) * age

<table>
<thead>
<tr>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25377</td>
<td>7163.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>25369</td>
<td>7145.8</td>
<td>8</td>
<td>17.957</td>
<td>7.9688</td>
</tr>
</tbody>
</table>

⇒ These additions appear to be useful too!
Do we get away without race main effects? \((-\text{race})\)

\begin{verbatim}
> wagereg9 <- lm(log.WR ~ race*age*sex - race + age2*sex +
+     marital + (hs+assoc+coll+grad)*age, data=YX)
> anova(wagereg8, wagereg9)

Model 1: log.WR ~ race * age * sex + age2 * sex + marital +
   (hs + assoc + coll + grad) * age
Model 2: log.WR ~ race * age * sex - race + age2 * sex +
   marital + (hs + assoc + coll + grad) * age
Res.Df   RSS  Df Sum of Sq    F Pr(>F)
1  25369 7145.8
2  25373 7146.0 -4   -0.20565 0.1825 0.9476

⇒ Reduced model is best.
\end{verbatim}
Testing is a difficult and imperfect way to compare models.

- You need a good prior sense of what model you want.
- $H_0$ vs $H_1$ is not designed to for model search.
- What “direction” do you search?
- A $p$-value doesn’t measure how much better a model is, only a yes/no answer.
- **Multiple Testing:** If you use $\alpha = 0.05 = 1/20$ to judge significance, then you expect to reject a true null about once every 20 tests.
Multiple testing

A big problem with using tests ($t$ or $F$) for comparing models is the false discovery rate associated with multiple testing:

- If you do 20 tests of true $H_0$, with $\alpha = 0.05$, you expect to see 1 false positive (i.e. you expect to reject a true null).

Suppose you have 100 predictors, but only 10 are useful

- You find all 10 of them significant . . . but what else?
- Reject $H_0$ for 5% of the useless 90 variables
  $\Rightarrow 0.05 \times 90 = 4.5$ false positives!
- Final model has $10 + 4.5 = 14.5$ variables
  $\Rightarrow 4.5/14.5 \approx 1/3$ are junk
- What happens if you set $\alpha = 0.01$?

In some online marketing data, $<1\%$ of variables are useful.
Data-Driven Model Selection

We need to find a trade-off between data fit and simplicity.

The partial $F$ test did this, but had problems:

1. You need a good prior sense of what model you want.
2. $H_0$ vs $H_1$ is not designed to for model search.
3. What “direction” do you search?
4. A $p$-value doesn’t measure how much better a model is.
5. Multiple Testing

Data-driven variable selection solves all 5 issues.

Use your head! Nothing is automatic.

Two steps:

1. Select the “universe of variables”.
2. Choose the best model.
The universe of variables is **HUGE**!

- includes all possible covariates that you think might have a linear effect on the response
- . . . and all squared terms . . . and all interactions . . .

**You** decide on this universe through your experience and discipline-based knowledge (and data availability).

- Consult subject matter research and experts.
- Consider carefully what variables have explanatory power, and how they should be transformed.
- If you can avoid it, don’t just throw everything in.

**This step is very important!** And also difficult.

. . . and sadly, not much we can do today.
Data Mining

“Data Mining” refers to tools that seek to uncover a small number of influential variables within large, high-dimensional datasets.

Unfortunately, it is more like the guy on the left.
As mentioned before, this is a very hard problem:
▶ Since very few variables are influential, testing is useless.
▶ You cannot consider all transformations and interactions.
▶ It is easy to overfit, which leads to bad predictions.

For industrial mining, more powerful tools are needed.

There are two full classes in this area: 41201 & 41204.
Out-of-sample prediction

How do we evaluate a forecasting model?

▶ Make predictions!

Basic Idea: We want to use the model to forecast outcomes for observations we have not seen before.

▶ Use the data to create a prediction problem. *(coming up)*
▶ See how our candidate models perform.

We’ll use most of the data for training the model, and the left over part for validating/testing it.
In a validation scheme, you

▸ fit a bunch of models to most of the data (training set)
▸ choose the one performing best on the rest (testing set).

For each model:

▸ Obtain $b_0, \ldots, b_d$ via least squares on the training data.
▸ Use the model to obtain fitted $\hat{Y}_j = x'_j b$ values for all of the $n_{\text{test}}$ testing data points.
▸ Calculate the Mean Square Error for these predictions.

\[
MSE = \frac{1}{n_{\text{test}}} \sum_{j=1}^{n_{\text{test}}} (Y_j - \hat{Y}_j)^2
\]
Example: Back to census data.

We aim to predict $\log(\text{wage rate})$ using demographics.

We tried many regressions:

```r
> wagereg6 <- lm(log.WR ~ age*sex + age2*sex + ., data=Wages)
> wagereg7 <- lm(log.WR ~ age*sex + age2*sex + marital +
+       (hs+assoc+coll+grad)*age + race*age , data=Wages)
> wagereg8 <- lm(log.WR ~ race*age*sex + age2*sex + marital +
+       (hs+assoc+coll+grad)*age, data=Wages)
> wagereg9 <- lm(log.WR ~ race*age*sex - race + age2*sex +
+       marital + (hs+assoc+coll+grad)*age, data=Wages)
```

$F$-tests showed `wagereg9` was the best.
Out of sample validation steps:

1) Split the data into testing/training samples.
   > training.samples <- sample.int(nrow(Wages), 0.75*nrow(Wages))
   > train <- Wages[trainingsamples,]
   > test <- Wages[-trainingsamples,]

2) Fit models on the training data
   > wagereg6 <- lm(log.WR ~ age*sex + age2*sex + ., data=train)

3) Predict on the test data
   > error6 <- predict(wagereg6, newdata=test) - test$log.WR

4) Compute MSE
   > c(error6=mean(error6^2), error7=mean(error7^2), 
     +  error8=mean(error8^2), error9=mean(error9^2))
     error6   error7   error8   error9
     0.2982959 0.2972645 0.2975347 0.2974996
wagereg7 is the winner, but...

- Only by a small amount. (MSE is directly interpretable!)
- Remember, test/train sampling is random.
  
  *Cross validation can help. What’s that?*

- $F$-testing favored wagereg9. So which is better?

Why did we only compare these models?

- We followed our nose down a very particular $F$ testing path. Any reason for that?
- Lots of other variables/interactions/etc.

We want to use the training sample to help select models for later comparison.
Coming Up Next

Next class: Use the training data to select models.
- Data will build models from scratch.
- We will go over a few methods, classical and modern, all useful

Also next class: Causal inference
- What $X$ variables do we need in order to claim “$X$ causes $Y$”?  
- Totally different goal, requires a different approach

Last lecture: Advanced GLMS

Then final and projects!
Glossary and Equations

**F-test**

- \( H_0 : \beta_{d_{\text{base}}+1} = \beta_{d_{\text{base}}+2} = \ldots = \beta_{d_{\text{full}}} = 0 \).
- \( H_1 : \) at least one \( \beta_j \neq 0 \) for \( j > 0 \).

**Null hypothesis distributions**

- **Total:** \( f = \frac{(R^2)/(p-1)}{(1-R^2)/(n-p)} \sim F_{p-1,n-p} \)
- **Partial:** \( f = \frac{(R^2_{\text{full}} - R^2_{\text{base}})/(d_{\text{full}}-d_{\text{base}})}{(1-R^2_{\text{full}})/(n-d_{\text{full}}-1)} \sim F_{d_{\text{full}}-d_{\text{base}},n-d_{\text{full}}-1} \)

**The partial \( F \)-test vs \( t \)-test**

- You have covariates \( X_1, \ldots, X_d \), and want to add only \( X_{d+1} \).
- Both tests have the same null and alternative:
  \( H_0 : \beta_{d+1} = 0 \) vs \( H_1 : \beta_{d+1} \neq 0 \)
- Different test statistics, same p-value