BUS41100 Applied Regression Analysis

Week 5: More Topics in MLR

Causal inference, Targeting
Model Building I: $F$-testing, multiple testing, multicollinearity

Max H. Farrell
The University of Chicago Booth School of Business
Causality

When does correlation ⇒ causation?

- We have been careful to never say that $X$ causes $Y$ . . .
- . . . but we’ve really wanted to.
- We want to find a “real” underlying mechanism:

  What’s the change in $Y$ as $T$ moves
  independent of all other influences?

But how can we do this in regression?
- First we’ll look at the Gold Standard: experiments
- Watch out for multiple testing
- Then see how this works in regression
Randomized Experiments

We want to know the effect of treatment $T$ on outcome $Y$.

What’s the problem with “regular” data? Selection.

- People choose their treatments
  - Eg: (i) Firm investment & tax laws; (ii) people & training/education; (iii) . . . .

Experiments are the best way to find a true causal effect.

Why? The key is randomization:

- No systematic relationship between units and treatments
  - $T$ moves independently by design.
- $T$ is discrete, usually binary.
  - Classic: drug vs. placebo
  - Newer: Website experience (A/B testing)
- Experiments are important (& common) in their own right
The fundamental question: Is $Y$ better on average with $T$?

$$\mathbb{E}[Y \mid T = 1] > \mathbb{E}[Y \mid T = 0]$$

We need a model for $\mathbb{E}[Y \mid T]$

- $T$ is just a special $X$ variable:

$$\mathbb{E}[Y \mid T] = \beta_0 + \beta_T T$$

- $\beta_T$ is the Average Treatment Effect (ATE)
- This is not a prediction problem, . . .
- . . . it’s an inference problem, about a single coefficient.

Estimation:

$$b_T = \hat{\beta}_T = \bar{Y}_{T=1} - \bar{Y}_{T=0}$$

Can’t usually do better than this. (Be wary of any claims.)
Why do we care about the average $Y$?

First, we might care about $Y$ directly, for an individual unit:

- Does $Y = \text{earnings}$ increase after $T = \text{training}$?
  - e.g. does getting an MBA increase earnings?
- Do firms benefit from consulting?
- Do people live longer with a medication/procedure?
- Do people stay longer on my website with the new design?

Or, we might care about aggregate measures:

- $Y = \text{purchase yes/no}$, then profit is $P = \text{price} \times Y$
  - Average profit per customer: $\mathbb{E}[P \times Y]$
  - Total profit: (No. customers)$\times\mathbb{E}[P \times Y]$
- Higher price means fewer customers, but perhaps more profit overall?  
  *(Ignore Giffen goods)*
Profit Maximization

Data from an online recruiting service

- Customers are firms looking to hire
- Fixed price is charged for access
  - Post job openings, find candidates, etc

Question is: what price to charge?

\[
\text{Profit at price } P = \text{Quantity}(P) \times (P - \text{Cost})
\]

Arriving customers are shown a random price \( P \)

- \( P \) is our treatment variable \( T \)
- How to randomize matters:
  - Why not do: \( P_1 \) in June, \( P_2 \) in July, \ldots? What's wrong?

Data set includes

- \( P = \text{price} - \text{price they were shown}, \$99 \text{ or } \$249 \)
- \( Y = \text{buy} - \text{did this firm sign up for service: yes/no} \)
Let’s see the data

```r
> price.data <- read.csv("priceExperiment.csv")
> summary(price.data)
> head(price.data)
```

Note that \( Y = \text{buy} \) is binary. That’s okay!

\[
E[Y] = P[Y = 1]
\]

Computing the ATE and Profit:

```r
> purchases <- by(price.data$buy, price.data$price, mean)
-0.1291639
4.311221
```

4.31 what? For whom? How many?
Regression version: computing ATE

> summary(lm(price.data$buy ~ price.data$price))

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|)  |
|----------------|----------|------------|---------|-----------|
| (Intercept)    | 0.3284017| 0.0195456  | 16.802  | <2e-16 ***|
| price.data$price | -0.0008611| 0.0001039  | -8.287  | <2e-16 ***|

Careful with how you code the variables!

> summary(lm(price.data$buy ~ (price.data$price==249)))

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|)  |
|----------------|----------|------------|---------|-----------|
| (Intercept)    | 0.24315  | 0.01091    | 22.285  | <2e-16 ***|
| price.data$price == 249TRUE | -0.12916| 0.01559    | -8.287  | <2e-16 ***|

What’s so special about $T = 0/1$?
Regression version: computing profit

> profit <- price.data$buy*price.data$price
> summary(lm(profit ~ (price.data$price==249)))

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 24.072   | 1.820      | 13.226  | <2e-16   *** |
| price.data$price == 249TRUE | 4.311    | 2.600      | 1.658   | 0.0974   .  |

---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 63.18 on 2361 degrees of freedom
Multiple R-squared: 0.001163, Adjusted R-squared: 0.0007402
F-statistic: 2.75 on 1 and 2361 DF, p-value: 0.09741

- Same profit estimate, thanks to transformed $Y$ variable
- Tiny $R^2$! Why?
- What’s 24.072?
Example: Job Training Program & Income

Un(der)employed men were randomized: 185 received job training, 260 didn’t

```r
> nsw <- read.csv("nsw.csv")
> nsw.outcomes <- by(nsw$income.after, nsw$treat, mean)
1794.342
```

- Outliers?
- Bunching?
- Income = 0?

Today, we’ll ignore these problems.
What do we learn from the regression output?

```r
> summary(nsw.reg <- lm(nsw$income.after ~ nsw$treat))
```

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) 4554.8 | 408.0 | 11.162 | < 2e-16 *** |
| nsw$treat 1794.3 | 632.9 | 2.835 | 0.00479 ** |

Residual standard error: 6580 on 443 degrees of freedom
Multiple R-squared: 0.01782, Adjusted R-squared: 0.01561
F-statistic: 8.039 on 1 and 443 DF, p-value: 0.004788

- Training increases earnings
- Tiny $R^2$! Why?
How does the TE vary over $X$? Useful for targeting

$$E[Y | T = 1, \text{black}] - E[Y | T = 0, \text{black}] = ?$$

> summary(lm(income.after ~ treat, data=nsw[nsw$black==1,]))

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 4107.7   | 457.9      | 8.971   | <2e-16   *** |
| treat          | 2028.7   | 706.1      | 2.873   | 0.0043   **  |

<table>
<thead>
<tr>
<th>Subpopulation</th>
<th>n</th>
<th>ATE</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>371</td>
<td>2029</td>
<td>0.004</td>
</tr>
<tr>
<td>Not Black</td>
<td>74</td>
<td>803</td>
<td>0.549</td>
</tr>
<tr>
<td>Hispanic</td>
<td>39</td>
<td>793</td>
<td>0.708</td>
</tr>
<tr>
<td>Not Hispanic</td>
<td>406</td>
<td>1960</td>
<td>0.003</td>
</tr>
<tr>
<td>Married</td>
<td>75</td>
<td>3709</td>
<td>0.017</td>
</tr>
<tr>
<td>Unmarried</td>
<td>370</td>
<td>1373</td>
<td>0.049</td>
</tr>
<tr>
<td>HS Degree</td>
<td>97</td>
<td>3192</td>
<td>0.038</td>
</tr>
<tr>
<td>No High School</td>
<td>348</td>
<td>1154</td>
<td>0.098</td>
</tr>
<tr>
<td>Black + Unmarried</td>
<td>307</td>
<td>1548</td>
<td>0.046</td>
</tr>
<tr>
<td>Black + No HS</td>
<td>293</td>
<td>1129</td>
<td>0.139</td>
</tr>
<tr>
<td>Unmarried + No HS</td>
<td>292</td>
<td>795</td>
<td>0.308</td>
</tr>
<tr>
<td>Black, Unmarried, No HS</td>
<td>244</td>
<td>644</td>
<td>0.448</td>
</tr>
</tbody>
</table>

Watch out for multiple testing and multicollinearity
Using *regression*

$\mathbb{E}[Y \mid T, \text{black}]$ is just $\mathbb{E}[Y \mid T, X]$: just a MLR

First try: $\mathbb{E}[Y \mid T, \text{black}] = \beta_0 + \beta_T T + \beta_1\text{black}$

- $\mathbb{E}[Y \mid T = 1, \text{black}] - \mathbb{E}[Y \mid T = 0, \text{black}] = \beta_T$

  ```r
  > summary(lm(income.after ~ treat + black, data=nsw))
  
  Estimate  Std. Error   t value   Pr(>|t|)
  (Intercept)    6289.2     799.3     7.869    2.79e-14  ***
  treat          1828.6     629.2     2.906     0.00384   **
  black         -2097.4     832.9    -2.518     0.01214   *
  
  Not the same! Why?

- Interpret conditional on the model
- **Same** $\beta_T$ for Not black, why?

Recall dummy variables: different intercepts, same slope.
A better model includes interactions:

\[ E[Y \mid T, \text{black}] = \beta_0 + \beta_T T + \beta_1 \text{black} + \beta_2 (T \times \text{black}) \]

\[ \Rightarrow E[Y \mid T = 1, \text{black}] - E[Y \mid T = 0, \text{black}] = \beta_T + \beta_2 \]

> summary(lm(income.after ~ treat*black, data=nsw))

| Estimate  | Std. Error | t value   | Pr(>|t|)   |
|-----------|------------|-----------|------------|
| (Intercept) | 6691.2     | 975.5     | 6.859 2.35e-11 *** |
| treat      | 802.8      | 1558.3    | 0.515 0.6067    |
| black      | -2583.5    | 1072.7    | -2.408 0.0164 * |
| treat:black| 1225.9     | 1703.5    | 0.720 0.4721    |

\[ \beta_T + \beta_2 = 2029 \]

\[ F\text{-test: 0.002045} \]

Continuous variables are fine too:

> summary(lm(income.after ~ treat*education, data=nsw))

| Estimate  | Std. Error | t value   | Pr(>|t|)   |
|-----------|------------|-----------|------------|
| (Intercept) | 3803.01    | 2568.91   | 1.480 0.1395 |
| treat      | -4585.18   | 3601.53   | -1.273 0.2036 |
| education  | 74.52      | 251.45    | 0.296 0.7671 |
| treat:education | 614.77 | 347.28    | 1.770 0.0774 . |
Causality Without Randomization

We want to find:

*The change in $Y$ caused by $T$ moving independently of all other influences.*

Our MLR interpretation of $\mathbb{E}[Y \mid T, X]$:

*The change in $Y$ associated with $T$, holding fixed all $X$ variables.*

⇒ We need $T$ to be randomly assigned given $X$

- $X$ must include enough variables so $T$ is random.
  - Requires a lot of knowledge!
- No systematic relationship between units and treatments, conditional on $X$.
  - It’s OK if $X$ is predictive of $Y$. 
The model is the same as always:

\[ \mathbb{E}[Y \mid T, X] = \beta_0 + \beta_T T + \beta_1 X_1 + \cdots \beta_d X_d. \]

But the assumptions change:

▶ This is a \textit{structural} model: it says something true about the real world.
▶ Need \( X \) to control for \textbf{all} sources of non-randomness.
  ▶ Even possible?

Then the interpretation changes:

\( \beta_T \) \textit{is the average treatment effect}

▶ Continuous “treatments” are easy.
▶ \textbf{Not} a “conditional average treatment effect”
  ▶ What happens to \( \beta_T \) as the variables change? To \( b_T \)?
▶ No \( T \times X \) interactions, why? What would these mean?
Example: Bike Sharing & Weather: does a change in humidity cause a change in bike rentals?

From Capital Bikeshare (D.C.’s Divvy) we have daily bike rentals & weather info.

- $Y_1 = \text{registered} - \#\,\text{rentals by registered users}$
- $Y_2 = \text{casual} - \#\,\text{rentals by non-registered users}$
- $T = \text{humidity} - \text{relative humidity (continuous!)}$

Possible controls/confounders:

- season
- holiday – Is the day a holiday?
- workingday – Is it a work day (not holiday, not weekend)?
- weather – coded 1=nice, 2=OK, 3=bad
- temp – degrees Celsius
- feels.like – “feels like” in Celsius
- windspeed
Is humidity randomly assigned to days?

humidity $\uparrow \Rightarrow$ rentals $\downarrow$!
Or is this because of something else?
The “randomized experiment” coefficient

\[
\begin{align*}
&> \text{summary(casual.reg <- lm(casual ~ humidity, data=bike))} \\
&\quad \text{Estimate Std. Error t value Pr(>|t|)} \\
&(\text{Intercept}) 1092.719 \quad 114.116 \quad 9.576 \quad < 2\text{e-16} \quad *** \\
&\text{humidity} \quad -5.652 \quad 1.912 \quad -2.957 \quad 0.00327 \quad ** \\
\end{align*}
\]

...is pretty similar to the effect with controls. So what?

\[
\begin{align*}
&> \text{summary(casual.reg.main <- lm(casual ~ humidity + season + holiday + workingday + weather + temp + windspeed, data=bike))} \\
&\quad \text{Estimate Std. Error t value Pr(>|t|)} \\
&(\text{Intercept}) \quad 716.964 \quad 203.273 \quad 3.527 \quad 0.000464 \quad *** \\
&\text{humidity} \quad -6.845 \quad 1.496 \quad -4.574 \quad 6.2\text{e-06} \quad *** \\
&\text{seasonspring} \quad -94.041 \quad 82.189 \quad -1.144 \quad 0.253152 \\
&\text{seasonsummer} \quad 182.964 \quad 53.249 \quad 3.436 \quad 0.000646 \quad *** \\
&\text{seasonwinter} \quad 57.194 \quad 68.849 \quad 0.831 \quad 0.406578 \\
&\text{holiday} \quad -285.327 \quad 103.757 \quad -2.750 \quad 0.006203 \quad ** \\
&\text{workingday} \quad -796.933 \quad 37.381 \quad -21.319 \quad < 2\text{e-16} \quad *** \\
&\text{weathernice} \quad 308.495 \quad 100.633 \quad 3.066 \quad 0.002305 \quad ** \\
&\text{weatherok} \quad 264.843 \quad 92.695 \quad 2.857 \quad 0.004475 \quad ** \\
&\text{temp} \quad 39.430 \quad 4.045 \quad 9.747 \quad < 2\text{e-16} \quad *** \\
&\text{windspeed} \quad -10.912 \quad 3.071 \quad -3.554 \quad 0.000420 \quad *** \\
\end{align*}
\]
The bottom line:

You only get causal effects with **strong** assumptions.

- Real-world concerns take precedence over statistics.
- Is there an economic/business/etc justification for your choice of $X$?

Diagnostics **do not help** with any of this!

Causal inference from observational data may be the hardest problem in statistics.

- We are just scratching the surface in terms of ideas, methods, applications, . . . .
- Still an active area of research in econometrics, statistics, & machine learning.
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.

- Does this regression have information?
- How does regression 1 compare to regression 2?

Next lecture we will look at more modern methods.

Again: none of this will help establish causality!
The *F*-test

The *F*-test tries to formalize the idea of a **big** $R^2$.

The test statistic is

$$f = \frac{SSR/(p - 1)}{SSE/(n - p)} = \frac{R^2/(p - 1)}{(1 - R^2)/(n - p)}$$

If $f$ is big, then the regression is “worthwhile”:

- Big SSR relative to SSE?
- $R^2$ close to one?
What we are really testing:

\[ H_0 : \beta_1 = \beta_2 = \cdots = \beta_d = 0 \]

\[ H_1 : \text{at least one } \beta_j \neq 0. \]

Hypothesis testing only gives a yes/no answer.

- Which \( \beta_j \neq 0 \)?
- How many?

The test is contained in the R `summary` for any MLR fit.

```r
> summary(lm(income.after ~ treat, data=nsw))
```

Multiple R-squared: 0.01782, Adjusted R-squared: 0.01561
F-statistic: 8.039 on 1 and 443 DF, p-value: 0.004788
Example: 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
```r
> attach(supervisor)
> bosslm <- lm(Y ~ X1 + X2 + X3 + X4 + X5 + X6)
> summary(bosslm)

Coefficients: # abbreviated output

|                  | 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

▶ $F$-value of 10.5 is very significant ($p$-val = $1.24 \times 10^{-5}$).
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 \textit{p-value} of 0.71 is not at all significant, so we stick with the null hypothesis and assume the base (2 covariate) model.
**Example:** Recall the wage rate curves from week 3.

We decided that

$$E[\log(WR)] = 1 + 0.07age - 0.0008age^2 + (0.02age - 0.00015age^2 - 0.34)1_M.$$  

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.

We also need to consider possible interactions.
```r
> summary(reg1 <- lm(log.WR ~ age*sex + age2*sex + ., data=YX))

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 ** |
```
Bring interactions of age with **race and education**:

```r
> reg2 <- lm(log.WR ~ age*sex + age2*sex + marital +
  +        (hs+assoc+coll+grad)*age + race*age , data=YX)
> anova(reg1, reg2)
```

Analysis of Variance Table

**Model 1:** \( \text{log.WR} \sim \text{age} \times \text{sex} + \text{age2} \times \text{sex} + \)

\[
(\text{age} + \text{age2} + \text{sex} + \text{race} + \text{marital} + \text{hs} + \text{assoc} + \text{coll} + \text{grad})
\]

**Model 2:** \( \text{log.WR} \sim \text{age} \times \text{sex} + \text{age2} \times \text{sex} + \text{marital} +

\[
(\text{hs} + \text{assoc} + \text{coll} + \text{grad}) \times \text{age} + \text{race} \times \text{age}
\]

<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>25385</td>
<td>7187.4</td>
<td>1</td>
<td>23.656</td>
<td>10.475</td>
</tr>
<tr>
<td>2</td>
<td>25377</td>
<td>7163.7</td>
<td>8</td>
<td>23.656</td>
<td>10.475</td>
</tr>
</tbody>
</table>

⇒ The new variables are significant!
Three way interaction!

```r
> reg3 <- lm(log.WR ~ race*age*sex + age2*sex + marital +
+ (hs+assoc+coll+grad)*age, data=YX)
> anova(reg2, reg3)

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? (-race)

```r
> reg4 <- lm(log.WR ~ race*age*sex - race + age2*sex +
+            marital + (hs+assoc+coll+grad)*age, data=YX)
> anova(reg3, reg4)

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.
```
**The $F$-test vs $t$-test**

You have $d$ covariates, $X_1, \ldots, X_d$, and want to add $X_{d+1}$ in.

The $t$-test is

$$H_0 : \beta_{d+1} = 0$$

$$H_1 : \beta_{d+1} \neq 0,$$

with

test stat $z = b_{d+1}/s_{b_{d+1}}$ and $p$-value $P(Z_{n-d-2} > |z|)$.

Partial $F$-test is

$$H_0 : \beta_{d+1} = 0$$

$$H_1 : \beta_{d+1} \neq 0,$$

with

test stat $f = (SSE_{\text{full}} - SSE_{\text{base}})/MSE_{\text{full}}$ and $p$-value $P(F_{1,n-d-2} > f)$.

Different stats, but same hypotheses!
Turns out that the tests also lead to the same \( p \)-values.

Consider testing \( X_1, X_2 \) supervisor regression vs. just \( X_1 \).

\( t \)-test asks: is \( b_2 \) far enough from zero to be significant?

```r
> summary(bosslm2) ## severely abbreviated

|         | Estimate | Std. Error | t value | Pr(>|t|) |
|---------|----------|------------|---------|----------|
| X2      | 0.2112   | 0.1344     | 1.571   | 0.128    |
```

\( F \)-test asks: is the increase in \( R^2 \) significantly big?

```r
> bosslm1 <- lm(Y ~ X1)
> anova(bosslm1, bosslm2)

<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>2</td>
<td>27</td>
<td>1</td>
<td>1254.7</td>
<td>114.73</td>
<td>2.4691</td>
</tr>
</tbody>
</table>
```
Limitations

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.
Limitations

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.

Multicollinearity: If the $X$’s are highly correlated with each other, then

- $s_{b_j}$’s will be very the big (since you don’t know which $X_j$ to regress onto), and
- therefore you may fail to reject $\beta_j = 0$ for all of the $X_j$’s even if they do have a strong effect on $Y$. 


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 $\alpha = .05$, 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
  \[ 0.05 \times 90 = 4.5 \text{ false positives!} \]
- Final model has $10 + 4.5 = 14.5$ variables
  \[ 4.5/14.5 \approx 1/3 \text{ are junk} \]
- What happens if you set $\alpha = 0.01$?

In some online marketing data, $<1\%$ of variables are useful.
Multicollinearity refers to strong linear dependence between some of the covariates in a multiple regression model.

The usual marginal effect interpretation is lost:
- change in one $X$ variable leads to change in others.

Coefficient standard errors will be large, such that multicollinearity leads to large uncertainty about the $b_j$'s.
Suppose that you regress $Y$ onto $X_1$ and $X_2 = 10 \times X_1$.

Then

$$E[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 = \beta_0 + \beta_1 X_1 + \beta_2 (10X_1)$$

and the marginal effect of $X_1$ on $Y$ is

$$\frac{\partial E[Y|X_1, X_2]}{\partial X_1} = \beta_1 + 10\beta_2$$

$\triangleright$ $X_1$ and $X_2$ do not act independently!
Consider 3 of our supervisor rating covariates:

**X2:** Opportunity to learn new things

**X3:** Does not allow special privileges

**X4:** Raises based on performance

⇒ A boss good at one aspect is usually good at the others.

Sure enough, they are all correlated with each other.
In the 3 covariate regression, none of the effects are significant:

```
> summary(lm(Y ~ X2 + X3 + X4))
```

Coefficients: # Abbreviated output

|            | Estimate | Std. Error | t value | Pr(>|t|) |
|------------|----------|------------|---------|----------|
| (Intercept)| 14.1672  | 11.5195    | 1.230   | 0.2298   |
| X2         | 0.3936   | 0.2044     | 1.926   | 0.0651 . |
| X3         | 0.1046   | 0.1682     | 0.622   | 0.5396   |
| X4         | 0.3516   | 0.2242     | 1.568   | 0.1289   |

Residual standard error: 9.458 on 26 degrees of freedom
Multiple R-squared: 0.4587, Adjusted R-squared: 0.3963
F-statistic: 7.345 on 3 and 26 DF, p-value: 0.00101

- But the $f$-stat is significant with $p$-value $= 0.001$!
If you look at individual regression effects, all 3 are significant:

> summary(lm(Y ~ X2)) ## severely abbreviated

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| 0.6468   | 0.1532     | 4.222   | 0.000231 *** |

> summary(lm(Y ~ X3)) ## severely abbreviated

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| 0.4239   | 0.1701     | 2.492   | 0.0189 *  |

> summary(lm(Y ~ X4)) ## severely abbreviated

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| 0.6909   | 0.1786     | 3.868   | 0.000598 *** |
Multicollinearity is not a big problem in and of itself, you just need to know that it is there.

If you recognize multicollinearity:

- Understand that the $\beta_j$ are not true marginal effects.
- Consider dropping variables to get a more simple model (use the partial $F$-test!).
- Expect to see big standard errors on your coefficients (i.e., your coefficient estimates are unstable).
Causality is hard!

- But do we care? Is correlation enough?

Model selection is more tractable. $F$-testing has big limitations. Next lecture we look at modern model selection methods that:

- Get around multiple testing
- Can’t do anything about multicollinearity
- Always need human input!
  - What does this model mean?
  - Does it answer our question?
Glossary and Equations

Causal Inference

- Treatment $T$.
- Mode $\mathbb{E}[Y \mid T] = \beta_0 + \beta_T T$.
- $\beta_T$ is the average treatment effect.
- In a regression: $\mathbb{E}[Y \mid T, X] = \beta_0 + \beta_T T + \beta_1 X + \beta_2 T X$.

F-test

- $H_0 : \beta_{d_{base}+1} = \beta_{d_{base}+2} = \ldots = \beta_{d_{full}} = 0$.
- $H_1 : \text{at least one } \beta_j \neq 0 \text{ for } j > 0$.
- Null hypothesis distributions
  - Total: $f = \frac{(R^2)/(p-1)}{(1-R^2)/(n-p)} \sim F_{p-1,n-p}$
  - Partial: $f = \frac{(R^2_{full} - R^2_{base})/(d_{full} - d_{base})}{(1-R^2_{full})/(n-d_{full} - 1)} \sim F_{d_{full} - d_{base}, n-d_{full} - 1}$