Joint Test for Significance, Regression with Dummy Variables, Interactions, Outliers, Etc.

D. Alex Hughes

November 17, 2014
This Week
  Dummy Variables
  Interactions
  ANOVA

Presenting Results
  Tables and More
  Predicted Values
This Week

Dummy Variables

Interactions

ANOVA

Presenting Results

Tables and More

Predicted Values
Dummy Variables

- Definition: A binary categorical variable with one category coded as “0” and the other as “1”.
- South/Not South
- Democracy/not democracy
- Might see “Woman” or $I_{\text{Woman}}$
- Avoid “Gender”.
- More examples?
Dummy Variables

Given an indicator or dummy variable $x$ and another dummy or indicator variable $z$:

- $\bar{x} = \frac{\sum x_i}{n} = p_x$
- $1 - x = x'$
- $\text{Median}(x) = \text{Mode}(x)$
- The product of two dummy variables $x$ and $z$ produces a new dummy coded 1 when both $x$ and $z$ are 1, and 0 when either or both are not 1. Or, $z \times x = I_{x \cap z}$
  
This is particularly useful if you believe in necessary and sufficient conditions.
Simple bivariate regression with a dummy variable:

\[ \text{Income} = a + b \times \text{Female} + \text{Error} \]
Bivariate Regression With Dummy

Histogram of Income

Histogram of Female
Bivariate Regression With Dummy

Histogram of Income[Female == 1]  
Histogram of Income[Female == 0]
Bivariate Regression With Dummy
Bivariate Regression With Dummy

![Graph showing income vs. female status]

- Income on the y-axis, ranging from 10000 to 35000
- Female status on the x-axis, ranging from -2 to 2
- The graph includes a regression line and data points for different female statuses.
## Dummy variables

## Example 1

### Income = a + b*Woman + e

n <- 2000
a <- 25000
b <- 4000
sderror <- 1000

Female <- rbinom(n,1,.5)
Income <- a + b*Female + rnorm(n,sd=sderror)

par(mfrow=c(1,2))
  hist(Income,nclass=40)
  hist(Female)

plot(Female,Income,type="n",xlim=c(-2,2),ylim=c(10000,35000))

plot(Female,Income,type="n",xlim=c(-2,2),ylim=c(10000,35000))
  abline(a,b)
plot(Female,Income,type="n",xlim=c(-2,2),ylim=c(10000,35000))
  abline(a,b)
  points(Female,Income)

par(mfrow=c(1,2))
  hist(Income[Female==1],xlim=c(min(Income),max(Income)))
  hist(Income[Female==0],xlim=c(min(Income),max(Income)))
```r
## Example 2
sderror <- 4
n <- 30
Left <- rbinom(n,1,.2)
a <- 75
b <- 15
Cohesion <- a + b*Left + rnorm(n,sd=sderror)

hist(Left)
hist(Cohesion)

plot(Left,Cohesion,type="n",xlim=c(-1,2.5),ylim=c(50,130))
abline(a,b)

plot(Left,Cohesion)
abline(a,b)

hist(Cohesion[Left==0])
hist(Cohesion[Left==1])
```
## Example 3: Inference for Dummy Variables

```r
umsims <- 1000
results <- array(NA,dim=c(numsims,2))

par(mfrow=c(1,2))
for(i in 1:numsims){
  Cohesion <- a + b*Left + rnorm(n,sd=sderror)
  plot(Left,Cohesion)#,xlim=c(0,1.1),ylim=c(0,101),main=i)
  coefs<-lm(Cohesion~Left)$coefficients
  abline(coefs)
  results[i,]<-coefs
  plot(results[1:i,],main=i,xlab="Slope",ylab="Intercept")
}
```
Bivariate Regression With Dummy

\[ Y = a + b \times X + \epsilon \]

How is this different from a difference of means test?

\[ H_0 : \mu_1 = \mu_2 \]
\[ H_A : \mu_1 \neq \mu_2 \]

Would this be different from a difference of means test?

\[ Y = 0 + b \times X + \epsilon \]
\[ H_0 = \mu_Y = \mu_Z \]
\[ H_A = \mu_Y \neq \mu_Z \]
Multivariate Regression With Dummy

\[ Y = a + b_1X_1 + b_2X_2 + \text{Error} \]

```
require(rgl)
n<-100
nspin=250
x<-rnorm(n)
y<-rnorm(n)
z<-rbinom(n=n,size=1,p=.5)
a<-1
b1<-2
b2<-10
y<-a+b1*x+b2*z+3*rnorm(n)
minx<-min(x)
maxx<-max(x)
plot3d(x,y,z,mouseMode=c("zAxis","fov","yAxis"),pch=19,size=6,col=2+2*z)
for(i in seq(from=1,to=360,length.out=nspin)){
  rgl.viewpoint(theta=i,phi=15,fov=100,zoom=.75)
}
```
rgl.lines()
x0 <- c(minx, maxx)
z0 <- c(0, 0)
z1 <- c(1, 1)
y0 <- a+b1*x0+b2*z0
y1 <- a+b1*x0+b2*z1
rgl.lines(x=x0, y=y0, z=z0, col="red", lwd=3)
rgl.lines(x=x0, y=y1, z=z1, col="red", lwd=3)

for(i in seq(from=1, to=360, length.out=nspin)){
  rgl.viewpoint(theta=i, phi=15, fov=100, zoom=.5)
}

rgl.lines()
x1<-c(minx,minx,maxx,maxx)
z1<-c(0,1,1,0)
y1<-a+b1*x1+b2*z1
rgl.quads(x=x1,y=y1,z=z1,add=TRUE)

for(i in seq(from=1,to=360,length.out=nspin)){
  rgl.viewpoint(theta=i,phi=15,fov=100,zoom=.5)
}

yfun<-function(x,z){
  return(1+2*x+2*z)
}
yhat<-yfun(x,z)
for(i in 1:n){
  rgl.lines(c(x[i],x[i]),c(y[i],yhat[i]),c(z[i],z[i]),col="red")
}
for(i in seq(from=1,to=360,length.out=nspin)){
  rgl.viewpoint(theta=i,phi=15,fov=100,zoom=.5)
}

plot(x,y,col=2+2*z,pch=19)
abline(a,b1)
abline(a+b2,b1)
• Same as any other multivariate regression
• All else equal, being "CATEGORY" increases Y by B.
• Easy to show trivariate regression in a two-dimensional plot as an "intercept shift"
Trivariate Regression With Dummy
Three Variable Regression With Dummy

\texttt{lm(formula = y \sim x1 + x2)}

Coefficients:
\begin{tabular}{ccc}
(Intercept) & x1 & x2 \\
0.8676 & 2.1429 & 2.3285
\end{tabular}
Trivariate Regression With Dummy
• Race
  • White
  • African American
  • Asian/Pacific Islander
  • Native American

• Include multiple dummy variables \((I_W, I_{AA}, I_{API}, I_{NAmer})\)

• If complete and mutually exclusive, exclude one to act as the “baseline”. Does race fit this condition?

• Including all categories will include redundant information.
  • Matrix example
  • Think about the excluded category with all included. What does \(\alpha\) mean?
Why do we leave out one category?

Consider a nominal variable with 2 categories, coded into dummies (gender? party or vote choice in the US?). In this case, your $X$ matrix looks like this:

$$
X = \begin{bmatrix}
1 & 1 & 0 \\
1 & 1 & 0 \\
1 & 1 & 0 \\
1 & 0 & 1 \\
1 & 0 & 1 \\
\end{bmatrix}
$$

Notice that the sum of columns 2 and 3 are equal to column one. That means that... the matrix contains some duplicate information, and we won’t be able to invert $X'X$. 
Why do we leave out one category?

Another way to think about it - what’s the intercept in this case? The expected value when belonging to none of the two categories.... but EVERYONE belongs to one of the categories!
The bottom line: when using dummy variables capturing mutually exclusive categories, you must leave one baseline category out of the regression.
Categorical Variables and Dummies

- Include all but one category.
- Coefficients are the impact of being \( X \) versus the excluded category, all else equal.
- Bivariate: South
- Multinomial: South, Northeast, Midwest, West
• Watch out for categories that are not mutually exclusive
• Which category to exclude?
• P-values in regression output test narrow hypotheses
• Use the F-Test for more general hypotheses
An Example

Dependent variable $Y$, independent ratio variable $X_1$, and independent nominal variable $X_2$, with categories $A$, $B$, and $C$. There are lots of $A$’s and $C$’s, not many $B$’s.
An Example

```r
X1 <- rnorm(100)
X2 <- c(rep("A", 45), rep("B", 10), rep("C", 45))

X2Alt <- c(rep(1, 45), rep(2, 10), rep(3, 45))

Y <- 1 + 2 * X1 + rnorm(100) + .3 * c(rep(-1, 45), rep(0, 10), rep(1, 45))
plot(X1, Y)
plot(X1, Y, col = X2Alt)
X2A <- X2 == "A"
X2B <- X2 == "B"
X2C <- X2 == "C"

X <- cbind(1, X1, X2A, X2B, X2C)
solve(t(X) %*% X) %*% t(X) %*% Y

m0 <- lm(Y ~ X1)
m1 <- lm(Y ~ X1 + X2A + X2C)
m2 <- lm(Y ~ X1 + X2B + X2C)
m3 <- lm(Y ~ X1 + X2A + X2B)
m4 <- lm(Y ~ X1 + X2A + X2B + X2C)
```
summary(m0)
summary(m1)
summary(m2)
summary(m3)
summary(m4)

anova(m0,m1)
anova(m0,m2)
anova(m0,m3)
anova(m0,m4)
• t-tests for dummies in regression just test whether category i’s mean is significantly different than the baseline (excluded) category mean.

• Shortcomings:
  • Might not be the comparison of interest.
  • Your theory might be much broader.
  • What if there are just a handful of observations in one or the other category?

• To test a broad hypothesis that your nominal variable “matters” (Race matters, religion matters, regime type matters), use an F-test. Compare the fit of the model including and excluding the dummy variables that represent the nominal categories.

• Q: will SSTotal be higher or lower when you drop the dummy variables?
Sidebar: Fixed-Effects Regression

Consider studying data where we have lots of data from lots of different countries - say 100 individuals (people? years? elections? parties?) from each of 100 countries, for a total dataset of size 10,000. Analysts will often include a dummy variable for each country, to “soak up” the country effects, if what they are really interested in are the differences between individuals. Estimated coefficients for each country are usually not even reported as they are secondary to the questions at hand. This approach is called a “fixed-effects regression”. There are many extensions of this approach, including a random-effects regression, and hierarchical linear models.
This Week
- Dummy Variables
- Interactions
- ANOVA

Presenting Results
- Tables and More
- Predicted Values
We’re “limited” to using a linear functional form. This means that we need to assume the $Y_i$ are just some linear combination of all the $X$s.

Is the world that simple?

How then, can we build more complexity into our thinking? How can we test conditional hypotheses?

Interactions!
The impact of X1 on Y varies with X2
For participation, education has a bigger impact for women than for men in Latin America.
For Income, education matters more in white collar jobs than blue collar jobs.
For party discipline, party labels matter more in developed areas than in underdeveloped areas.
For peace, trade matters more for manufactured goods than commodities
Interactions where one X is a dummy variable

- Income = a + b1*Education + b2*White Collar + b3*Education*White Collar + Error
- See R Code
# Dummy Interaction

```r
require(rgl)
n<-250
nspin=1000
x<-rnorm(n)
y<-rnorm(n)
z<-rbinom(n=n,size=1,p=.5)
a<-1
b1<-2
b2<-10
b3<-4
y<-a+b1*x+b2*z+b3*x*z+3*rnorm(n)
minx<-min(x)
maxx<-max(x)
plot3d(x,y,z,mouseMode=c("zAxis","fov","yAxis"),pch=19,size=6,col=2+2*z)
for(i in seq(from=1,to=360,length.out=nspin)){
  rgl.viewpoint(theta=i,phi=15,fov=100,zoom=.75)
}
```
rgl.lines()

x0 <- c(minx,maxx)
z0 <- c(0,0)
z1 <- c(1,1)
y0 <- a+b1*x0+b2*z0+b3*x0*z0
y1 <- a+b1*x0+b2*z1+b3*x0*z1
rgl.lines(x=x0,y=y0,z=z0,col="red",lwd=3)
rgl.lines(x=x0,y=y1,z=z1,col="blue",lwd=3)

for(i in seq(from=1,to=360,length.out=nspin)){
  rgl.viewpoint(theta=i,phi=15,fov=100,zoom=.5)
}

rgl.lines()
npts<-10
xs<-seq(from=minx,to=maxx,length.out=npts)
zs<-seq(from=0,to=1,length.out=npts)
Y<-array(NA,dim=c(npts,npts))
for(i in 1:npts){
  for(j in 1:npts){
    Y[i,j]<-a+b1*xs[i]+b2*zs[j]+b3*xs[i]*zs[j]
  }
}
rgl.surface(x=xs,y=Y,z=zs)

for(i in seq(from=1,to=360,length.out=nspin)){
  rgl.viewpoint(theta=i,phi=15,fov=100,zoom=.5)
}

yfun<-function(x,z){
  return(a+b1*x+b2*z+b3*x*z)
} yhat<-yfun(x,z)
for(i in 1:n){
  rgl.lines(c(x[i],x[i]),c(y[i],yhat[i]),c(z[i],z[i]),col="red")
} for(i in seq(from=1,to=360,length.out=nspin)){
  rgl.viewpoint(theta=i,phi=15,fov=100,zoom=.5)
}

plot(x,y,col=2+2*z,pch=19)
abline(a,b1,col="red",lwd=3)
abline(a+b2,b1+b3,col="blue",lwd=3)
Interactions where one $X$ is a dummy variable

- $\text{Income} = a + b_1 \times \text{Education} + b_2 \times \text{White Collar} + b_3 \times \text{Education} \times \text{White Collar} + \text{Error}$
- This can be seen as both an intercept and a slope shift
Interactions where one X is a dummy variable
Interactions where one X is a dummy variable

x1 <- runif(100)
x2 <- rbinom(100,1, prob = .5)
y <- 1 + 4*x1 + 2*x2 + 2*x1*x2 + rnorm(100, sd = .5)
summary(lm(y ~ x1 + x2))
summary(lm(y ~ x1 + x2 + x1*x2))

plot(x1,y)
plot(x2,y)
Interactions where one X is a dummy variable
Interactions where one $X$ is a dummy variable

\[
Participation = \alpha + \beta_1 \ast \text{Age} + \beta_2 \ast \text{Woman} \\
+ \beta_3 \ast \text{Age} \ast \text{Woman} + \epsilon
\]

- Predicted Participation for Unemployed Men is $\alpha$
- Predicted Participation for Employed Men is $\alpha + \beta_1$
- Predicted Participation for Unemployed Women is $\alpha + \beta_2$
- Predicted Participation for Employed Women is $\alpha + \beta_1 + \beta_2 + \beta_3$

**Note!!**

- Can’t speak of “The Impact of Gender” or “The Impact of Employment Status”.
- Instead, we have to speak of “The impact of employment on women”, for example.
Interactions where one X is continuous

\[ Participation = \alpha + \beta_1 \times \text{Age} + \beta_2 \times \text{Woman} \]
\[ + \beta_3 \times \text{Age} \times \text{Woman} + \epsilon \]

1. What is the impact of gender conditional on age?
   \[ \beta_2 + \beta_3 \times \text{Age} \]

2. Or, what’s the impact of age conditional on gender?
   \[ \beta_1 + \beta_3 \times \text{Woman} \]
Inference with Interactions

What has been our test statistic for a single coefficient?

\[ t_{n-k-1} = \frac{\hat{\beta} - \beta_0}{\hat{SE}(\hat{\beta}_j)} \]

But what is wrong (or potentially wrong) with this setup for testing \( \beta_3(X_1 \times X_2) \)? We’re not testing \( \beta_3 \) in insolation – we know that it is related to \( \beta_2 \)!
Inference with Interactions

Standard error for the impact of $X_2$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \ast X_2$$

$$\text{Var}(\beta_2 + \beta_3 \ast x_1) = ?$$

$$\text{Var}(\beta_2) + \text{Var}(\beta_3 \ast x_1) + 2\text{Cov}(\beta_2, \beta_3 \ast x_1)$$

$$\text{Var}(\beta_2) + x_1^2 \text{Var}(\beta_3) + x_1 \ast 2\text{Cov}(\beta_2, \beta_3)$$

$$\sigma_{\beta_2 + \beta_3 x_1} = \sqrt{\text{Var}(\beta_2) + x_1^2 \text{Var}(\beta_3) + x_1 \ast 2\text{Cov}(\beta_2, \beta_3)}$$
General Advice for Interactions

- Use predicted values to show impact and develop intuition.
- Also use graphs of coefficients to show how the impact of $x_1$ varies with $x_2$.
- Don’t forget that these quickly get complicated, and that it is difficult to eyeball coefficients.
An Example

Given the following:

\[ \text{Income} = \alpha + \beta_1 \text{Education} + \beta_2 \text{Gender} + \beta_3 \text{Gender} \times \text{Education} \]

- Try making a picture that represents this in two lines and interpreting each of the coefficients.
- Let’s think about three questions:
  - Does gender have any impact on income, interactive or not? F-test on \( \beta_2 \) and \( \beta_3 \)
  - What is the impact of gender on income? \( \beta_2 + \beta_3 \times \text{Education} \)
  - Is the impact of education on income different for men than for women? \( \beta_3 \)
A different example

Now try for:

\[
Income = \alpha + \beta_1 Education + \beta_2 Age + \beta_3 Age \ast Education
\]

- Does age have any impact on income, interactive or not?
- What is the impact of age on income?
- Is the impact of education on income different across different ages?
An Example Using Predicted Values

Given the following:

\[ \text{Income} = \alpha + \beta_1 \text{Education} + \beta_2 \text{Gender} + \beta_3 \text{Gender} \times \text{Education} \]

A second way to proceed: predicted values: drawing one or more regression lines.

\[ \text{Income} = \alpha + \beta_1 \text{Education} + \beta_2 \text{Age} + \beta_3 \text{Age} \times \text{Education} \]

- Include Interaction Terms whenever you have a conditional hypothesis.
- Include all constitutive terms.
- Constitutive terms are not Unconditional Marginal Effects
- SE’s Require Covariance; Marginal Effects require examples
- Difficult to eyeball coefficients

1. Fit a model that does not include either term you’re going to interact
2. Fit a model with the “main-effects” (and all other RHS vars) but no interaction
3. Fit a model with all the terms
4. Test for model improvement from the first to the second to the third (F-test)
5. If you’re not doing any better, *even if* interaction is significant, little justification for conditional hypothesis.
Do Prestige Example
This Week

- Dummy Variables
- Interactions
- ANOVA

Presenting Results

- Tables and More
- Predicted Values
ANOVA: Analysis of Variance

- Used When: continuous dependent variable, one or more unordered, categorical independent variables.
- Examples:
  - Mobility as a function of region
  - Affect for politician as a function of ad exposure
  - Income as a function of race and gender
- Most common in political psychology. Frequently used in experimental literature
One-way ANOVA

- $Y_{ij} = \mu + \alpha_j + e_{ij}$, where $i$ refers to the individual and $j$ refers to the group.
- The model is just estimating the mean $Y$ for each group $j$, and the standard error of the error $\epsilon \sim N(0, \sigma^2)$.
- Unordered categorical variables.

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\mu + \alpha_1$</td>
<td>$Y_{11}, Y_{21}, Y_{31}, Y_{41}, \ldots Y_{n1}$</td>
</tr>
<tr>
<td>2</td>
<td>$\mu + \alpha_2$</td>
<td>$Y_{12}, Y_{22}, Y_{32}, Y_{42}, \ldots Y_{n2}$</td>
</tr>
<tr>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
</tr>
<tr>
<td>$m$</td>
<td>$\mu + \alpha_m$</td>
<td>$Y_{1m}, Y_{2m}, Y_{3m}, Y_{4m}, \ldots Y_{n2m}$</td>
</tr>
</tbody>
</table>
ANOVA really is just regression adjusted to make it more interpretable...

- Regression: exclude a base category (like White).
  - Other estimates are interpreted as difference from White.
  - To get a group mean, need Intercept + Estimate.
What’s the Difference

ANOVA really is just regression adjusted to make it more interpretable...

- **ANOVA**: Instead of restricting one category to be zero, and picking it up with the intercept, we use the following:

  \[
  \mu_j = \mu + \alpha_j
  \]

  \[
  \sum \alpha_j = 0
  \]

- What does this do? It just gives the coefficients an easy interpretation: it is each group’s average *difference* from the overall mean. So you can tell at a glance if a category is above or below the overall mean.
Inference for ANOVA

- Does knowing group means $\mu_j$ explain much of the variance in $Y$?
- Is there mostly within or between group variance?
- $H_0$: The $\alpha_j$ are all zero. This means that none of the groups deviate from the overall mean - knowing that an observation is in group $j$ tells us nothing about $Y_{ij}$.
- $H_A$: The $\alpha_j$ are not all zero. Some of the groups deviate from the overall mean. Knowing that an observation is in group $j$ does tells us something about $Y_{ij}$.
• Individual $\alpha$’s usually aren’t meaningful.
• Instead, we use an F-Test, identical to regression, to ask if the explanatory variable soaks up a significant part of the variance of $Y$. 
Inference for ANOVA

\[ RegSS = \sum_{j=1}^{m} \sum_{1}^{n_j} (\hat{Y}_{ij} - \bar{Y})^2 \]

\[ = \sum_{j=1}^{m} n_j (\bar{Y}_j - \bar{Y})^2 \]

\[ RSS = \sum_{j=1}^{m} \sum_{n_j=1}^{n_j} (Y_{ij} - \hat{Y}_{ij})^2 \]

\[ = \sum_{j=1}^{m} \sum_{1}^{n_j} (Y_{ij} - \bar{Y}_i)^2 \]

\[ F = \frac{RegSS/(m - 1)}{RSS/(n - m)} \]
Graphically, what does this look like?

groupvar<-c(rep(1,25),rep(2,25),rep(3,25))
t1<-rnorm(3,sd=3)
truemeans<-sort(rep(t1,25))
Yij <- truemeans+rnorm(75)
plot(groupvar,Yij)
abline(mean(Yij),0)
for(i in 1:3){
  abline(mean(t1[i]),0,lty=2)
}
lm(Yij~groupvar)
lm(Yij~as.factor(groupvar))
anova(lm(Yij~as.factor(groupvar)))
Inference for ANOVA

<table>
<thead>
<tr>
<th>Group</th>
<th>Data: $\mu = .66$</th>
<th>Group Mean</th>
<th>$n$</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pos Pol</td>
<td>0.97, 0.51, 3.18, 0.59</td>
<td>1.31</td>
<td>4</td>
<td>.85</td>
</tr>
<tr>
<td>Pos Trait</td>
<td>-0.10, -2.19, -0.95, 0.22</td>
<td>-0.65</td>
<td>4</td>
<td>.01</td>
</tr>
<tr>
<td>Neg Pol</td>
<td>0.78, -0.45, -0.45, 1.03</td>
<td>.23</td>
<td>4</td>
<td>-.43</td>
</tr>
<tr>
<td>Neg Trait</td>
<td>3.10, 0.70, 2.10, 1.50</td>
<td>1.84</td>
<td>4</td>
<td>1.18</td>
</tr>
</tbody>
</table>

\[
\text{RegSS} = 4 \times (1.31 - .66)^2 + \cdots + 4 \times (1.84 - .66)^2
\]
\[
= 16.12
\]
\[
\text{RSS} = (.97 - 1.31)^2 + (.51 - 1.31)^2 + \cdots + (1.50 - 1.84)^2
\]
\[
= 35.98
\]
\[
F = \frac{16.12/(4 - 1)}{35.98/(16 - 4)} = 1.792823734
\]
\[
P(F) \sim 0.20
\]
Multivariate ANOVA

- Called “two-way”, “three-way”, or higher order ANOVA.
- RXC table
- Core results similar
- The Model
  \[ Y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \epsilon_{ijk} \]
- This is still a regression, albeit one with lots of interactions:
The parameters of the model are:

- $\mu$, the grand mean.
- $\alpha_j$, the deviation of row $j$ from the grand mean.
- $\beta_k$, the deviation of column $k$ from the grand mean.
- $\gamma_{jk}$, the deviation of cell $jk$ from $\mu$ mean plus $\alpha_j$ plus $\beta_k$.

Again, we are just transforming the regression coefficients from “deviation from baseline category” to “deviation from expectation”.
Multivariate ANOVA

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laborer</td>
<td>$\mu_{11}$</td>
<td>$\mu_{12}$</td>
<td>$\mu_{13}$</td>
</tr>
<tr>
<td>Manager</td>
<td>$\mu_{21}$</td>
<td>$\mu_{22}$</td>
<td>$\mu_{23}$</td>
</tr>
<tr>
<td>Craftsman</td>
<td>$\mu_{31}$</td>
<td>$\mu_{32}$</td>
<td>$\mu_{33}$</td>
</tr>
<tr>
<td>Other</td>
<td>$\mu_{41}$</td>
<td>$\mu_{42}$</td>
<td>$\mu_{43}$</td>
</tr>
</tbody>
</table>

Mean for cell $jk$ is:

$$\mu_{jk} = \mu_\cdot + \alpha_j + \beta_k + \gamma_{jk}$$
Two-way ANOVA

- A common question is about interactions - does the impact of $X_1$ vary with $X_2$, both categorical variables.
- The easiest way to show this is with line graphs.
- If the lines are parallel, there is no interaction.
- If the lines are not parallel, there *may* be an interaction.
- Interactions refer to the $\gamma_{jk}$ terms.
Final Thoughts

- Higher order ANOVA is an option - Income = f(race, gender, hair color, college type, employment category)
- As the number of categories grows, the number of cells and parameters grows even faster...
- At some point, the number of cells will exceed the number of datapoints.
This Week
  Dummy Variables
  Interactions
  ANOVA

Presenting Results
  Tables and More
  Predicted Values
Standardized Regression Coefficients

“Betas”

\[
\hat{\theta} = \hat{\beta} \frac{S_X}{S_Y}
\]

- Just *standardized regression coefficients*
- Interpretation: A one standard deviation change in X produces a Beta standard deviation change in Y
- Advantage: places impact in the context of the data. Gain in comparability.
- Disadvantage: Removes the “real world” element of the data.
- Less common today, depending on who you ask.
Typical Regression Tables

- Models in columns
- Variables are rows
- Report estimated coefficient, standard error, p-value/sig test.
- $n$, $R^2$, F-Test or $LL$ (Fit statistics)
### TABLE 2. Association Between City Liberalism and Policy Outcomes

<table>
<thead>
<tr>
<th></th>
<th>Scaled Policy</th>
<th>Per Capita Expend.</th>
<th>Per Capita Taxes</th>
<th>Sales Tax Share</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.13**</td>
<td>0.18</td>
<td>1,838.85**</td>
<td>1,780.02**</td>
</tr>
<tr>
<td></td>
<td>(0.07)</td>
<td>(0.15)</td>
<td>(207.31)</td>
<td>(217.62)</td>
</tr>
<tr>
<td>Policy Conservatism</td>
<td>1.19**</td>
<td>1.04**</td>
<td>-760.75**</td>
<td>-347.42**</td>
</tr>
<tr>
<td></td>
<td>(0.18)</td>
<td>(0.19)</td>
<td>(97.42)</td>
<td>(116.80)</td>
</tr>
<tr>
<td>Median Income</td>
<td>0.29</td>
<td></td>
<td>-720.01**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.35)</td>
<td></td>
<td>(204.90)</td>
<td></td>
</tr>
<tr>
<td>City Population</td>
<td>-0.01</td>
<td></td>
<td></td>
<td>54.41**</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td></td>
<td></td>
<td>(7.77)</td>
</tr>
<tr>
<td>Percent Black</td>
<td>0.15</td>
<td></td>
<td>330.05**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.39)</td>
<td></td>
<td>(177.82)</td>
<td></td>
</tr>
<tr>
<td>Med. Housing Value</td>
<td>-0.15**</td>
<td></td>
<td>238.69**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td></td>
<td>(39.24)</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>437</td>
<td>436</td>
<td>1,619</td>
<td>1,618</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-546.35</td>
<td>-547.24</td>
<td>-13,218.37</td>
<td>-13,146.86</td>
</tr>
<tr>
<td>Akaike Inf. Crit.</td>
<td>1,100.71</td>
<td>1,110.48</td>
<td>26,444.74</td>
<td>26,309.71</td>
</tr>
<tr>
<td>Bayesian Inf. Crit.</td>
<td>1,117.03</td>
<td>1,143.10</td>
<td>26,466.30</td>
<td>26,352.62</td>
</tr>
</tbody>
</table>

* **p < 0.05; *p < 0.1.
• Keep it simple
• Don’t overwhelm with tons of models
• Consider zeroing in on the core hypothesis test and putting the rest in the appendix
• Write out estimating equation?
Graphs Instead of Tables?

D. Alex Hughes
Omnibus, F, and Dummy Tests
November 17, 2014
This Week
  Dummy Variables
  Interactions
  ANOVA

Presenting Results
  Tables and More
  Predicted Values
**Goal:** Convey the substantive import of your findings, or the magnitude of the impact of your explanatory variables.

- Instead of Betas, use predicted values
- Compare how predicted values of Y change with X
- Include standard errors

\[
\hat{y}_i = \hat{\alpha} + \hat{\beta}x_0
\]

\[
\hat{y} = x_0\hat{\beta}
\]
We know that our predicted value will be off. Our uncertainty reflects:

- Uncertainty about the intercept
- Uncertainty about the slope
- Uncertainty about the standard error of the regression $\sigma$

We want a confidence interval for a predicted value of $Y$. 
Inference for Predicted Values

Key distinction:

- *Mean* prediction; or,
- *Individual* prediction?

What is our prediction of the average at $x_0$, or what is our prediction for a new individual 0’s $y$ value?

In both cases, the prediction is the same:

$$\hat{y}_0 = x_0 \hat{\beta}$$

But our uncertainty is different.
Inference for \textit{Mean} Predicted Value

- We want a confidence interval around what the true value of the line is at $x_0$.
- A confidence interval should reflect our uncertainty about what the true slope is and what the true intercept is.

\textbf{Interpretation:} What is our prediction of the average individual at $x_0$? Or: Where’s the true line at $x_0$?

- Prediction uncertainty minimized when?
- Prediction uncertainty increases when?
$\delta = \hat{y}_0 - E[y_0]$

$E[\delta] = E[x_0b] - E[x_0^t\beta + \epsilon]$

$= E[x_0]E[b] - E[x_0]E[\beta] + E[\epsilon]$

$= 0$

$Var[\delta] = Var(\hat{y}_0 - E[y_0])$

$= Var(x_0\beta - 0)$

$= x_0^tVar(\beta)x_0$

$= \sigma_\epsilon^2[x_0'(X'X)^{-1}x_0]$
CI for Individual Predicted Value

\[ D \equiv \hat{Y}_0 - Y_0 = x'_0 b - (x'_0 \beta + \epsilon_0) \]
\[ = x'_0 (b - \beta) - \epsilon_0 \]
\[ E[D] = E[x'_0 b] - E[x'_0 \beta + \epsilon_0] \]
\[ = E[x'_0] E[b] - E[x'_0 \beta] + E[\epsilon_0] \]
\[ = E[x'_0] \beta - E[x'_0] \beta \]
\[ = 0 \]
\[ V[D] = V[x'_0(b - \beta) - \epsilon_0] \]
\[ = V[x'_0(b - \beta)] + V[\epsilon_0] \]
\[ = x'_0 V[(b - \beta)]x_0 + \sigma^2_\epsilon \]
\[ = x'_0 [V[b] - V[\beta]]x_0 + \sigma^2_\epsilon \]
\[ = x'_0 (x'x)^{-1}\sigma^2_\epsilon x_0 + \sigma^2_\epsilon \]
\[ = \sigma^2_\epsilon \left[ 1 + x'_0(x'x)^{-1}x_0 \right] \]
Comparing the Two

\[ \hat{y}_0 \pm t_{\lambda/2} \sqrt{\frac{\sigma^2}{\epsilon} \left[ x_0'(x'x)^{-1}x_0 \right]} \]

\[ \hat{y}_0 \pm t_{\lambda/2} \sqrt{\frac{\sigma^2}{\epsilon} \left[ 1 + x_0'(x'x)^{-1}x_0 \right]} \]

Show Code -- `predictSim`
Predicted Values - Additional Considerations

- Choose range of $X_i$ to show
- Choose values of $X_{j\neq i}$ to set as fixed values - typically mean or median (why not set all to $+100,000,000$?)
- Calculate $\hat{y}$, $\hat{y} \pm 1.96 \times se(\hat{y})$
- Graph or report in a table
- With interactions, can be complex, depending on what you want to show
using data of this type. As reported in Table 1, weighting did not affect the substance of our conclusions.

RESULTS

The estimates indicate there is a strong positive correlation between the flow of external resources to local NGOs and the Left’s share of the vote. The results, presented in Table 1, are substantively and statistically significant. The regression model estimates the correlation between the funds distributed to NGOs by the Plantaffo project and the change in the left’s vote share from 1994 to 1998. These results were robust to multiple diagnostics including DFFIT’s, Cook’s D, and Leverages as well as visually inspecting the residuals plotted against the predicted values. In addition to the variables reported in Table 1, we varied the model using measures of population density, different levels of educational attainment, the percentage of a municipality’s housing that represent nonpermanent structures (domestic/fish improvements), and the number of residents per school. Whether these variables were excluded or included had no significant impact on the results.

Roughly 85% of the variance is explained by the weighted model. Most of the socioeconomic variables show little correlation with changes in the Left’s share of the vote once the linear restrictions of the other variables have been removed. Only education has a statistically significant relationship with votes for the Left. To give an illustration of the substantive significance of the
results, consider the municipalities of Corumbiara, Cacaúlandia, and Monte Negro. Each of the three towns received no money through the Planalto program during the period in question. The model predicts that in such instances, the Left’s share of the vote should decrease by 7 percentage points (±2; standard error of the prediction) from 1994 to 1998. In well-funded municipalities (Theobroma, Candias do Jamari, and Rio Crespo), the model predicts an increase of 7 percentage points for the Left (±4, standard error of the prediction). Although the center-Right candidates easily won most of Rondónia’s vote in 1994 and 1998, the differences between their gains and losses in poorly funded and well-funded municipalities was approximately 14 percentage points. Figure 1 provides predicted values for the change in the Left’s share of the vote from 1994 to 1998 in addition to the 95% confidence intervals. The predicted values were generated by the weighted model.

DIAGNOSTICS ON THE MODEL

We validated our findings by testing additional models that controlled for other dramatic political events, especially for important developments in the landless peasant movement (the MST). During the period we analyzed, an important political shock occurred. Military police killed nine landless peasants at the Santa Elina Estate, Corumbiara, Rondonia State on August 9, 1995.
The Easy Way To Show Predicted Values?


- Fit Model
- Set $x_0$
- Simulate values of $\beta$ from the asymptotic normal distribution
- Calculate predicted values
- Repeat many times
The General Way To Show Predicted Values?

- Fit your model
- Create predData dataset
- Predict, and plot
- See code