Categorical Predictors

Statistics & Methodology Lecture 7

Kyle M. Lang

Department of Methodology & Statistics
Tilburg University

Outline

1. Adding categorical predictors into MLR models

Categorical Predictors

Most of the predictors we've considered thus far have been *quantitative*.

- Continuous variables that can take any real value in their range
- Interval or Ratio scaling
- If we use ordinal items as predictors, we assume interval scaling.

We often want to include grouping factors as predictors.

- These variables are *qualitative*.
  - Their values are simply labels.
  - There is no ordering of the categories.
  - Nominal scaling

How to Model Categorical Predictors

We need to be careful when we include categorical predictors into a regression model.

- The variables need to be coded before entering the model.

Consider the following indicator of major:
$X_{maj} = \{1 = Law, 2 = Economics, 3 = Data\ Science\}$

- What would happen if we naïvely used this variable to predict program satisfaction?

How to Model Categorical Predictors

```r
mDat <- readRDS("../data/major_data.rds")
mDat[seq(25, 150, 25), ]

##     sat majF majN
## 25  1.9  law    1
## 50  1.4  law    1
## 75  4.3 econ    2
## 100 4.1 econ    2
## 125 5.7   ds    3
## 150 5.1   ds    3

out <- lm(sat ~ majN, data = mDat)
```

How to Model Categorical Predictors

```
partSummary(out, -c(1, 2))

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.33200    0.12060  -2.753  0.00664
## majN         2.04500    0.05582  36.632  < 2e-16
##
## Residual standard error: 0.5582 on 148 degrees of freedom
## Multiple R-squared:  0.9007,Adjusted R-squared:    0.9
## F-statistic:  1342 on 1 and 148 DF,  p-value: < 2.2e-16
```

Dummy Coding

The most common way to code categorical predictors is *dummy coding*.

- A G-level factor (i.e., one that represents G groups) will be transformed into a set of $G - 1$ dummy codes.
- Each code is a variable on the dataset that equals 1 for observations corresponding to the code's group and equals 0, otherwise.
- The group without a code is called the *reference group*.

Example Dummy Code

Let's look at the simple example of coding biological sex:

	sex	male
1	female	0
2	male	1
3	male	1
4	female	0
5	male	1
6	female	0
7	female	0
8	male	1
9	female	0
10	female	0

Example Dummy Codes

Now, a slightly more complex example:

	drink	juice	tea
1	juice	1	0
2	coffee	0	0
3	tea	0	1
4	tea	0	1
5	tea	0	1
6	tea	0	1
7	juice	1	0
8	tea	0	1
9	coffee	0	0
10	juice	1	0

Using Dummy Codes

To use the dummy codes, we simply include the $G - 1$ codes as $G - 1$ predictor variables in our regression model.

$$Y = \beta_0 + \beta_1 X_{male} + \varepsilon$$

$$Y = \beta_0 + \beta_1 X_{juice} + \beta_2 X_{tea} + \varepsilon$$

- The intercept corresponds to the mean of Y in the reference group.
- Each slope represents the difference between the mean of Y in the coded group and the mean of Y in the reference group.

Example

First, an example with a single, binary dummy-coded variable:

```r
## Read in some data:
cDat <- readRDS("../data/cars_data.rds")

## Fit and summarize the model:
out1 <- lm(price ~ mtOpt, data = cDat)
```

Example

```
partSummary(out1, -c(1, 2))

## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    23.841      1.623  14.691   <2e-16
## mtOpt          -6.603      2.004  -3.295   0.0014
##
## Residual standard error: 9.18 on 91 degrees of freedom
## Multiple R-squared:  0.1066,Adjusted R-squared:  0.09679
## F-statistic: 10.86 on 1 and 91 DF,  p-value: 0.001403
```

Example

Fit a more complex model:

```
out2 <- lm(price ~ front + rear, data = cDat)
partSummary(out2, -c(1, 2))

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.63000    2.76119   6.385 7.33e-09
## front       -0.09418    2.96008  -0.032  0.97469
## rear        11.32000    3.51984   3.216  0.00181
##
## Residual standard error: 8.732 on 90 degrees of freedom
## Multiple R-squared:  0.2006,Adjusted R-squared:  0.1829
## F-statistic: 11.29 on 2 and 90 DF,  p-value: 4.202e-05
```

Example

Include two sets of dummy codes:

```
out3 <- lm(price ~ mtOpt + front + rear, data = cDat)
partSummary(out3, -c(1, 2))

## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   21.7187     2.9222   7.432 6.25e-11
## mtOpt         -5.8410     1.8223  -3.205  0.00187
## front         -0.2598     2.8189  -0.092  0.92677
## rear          10.5169     3.3608   3.129  0.00237
##
## Residual standard error: 8.314 on 89 degrees of freedom
## Multiple R-squared:  0.2834,Adjusted R-squared:  0.2592
## F-statistic: 11.73 on 3 and 89 DF,  p-value: 1.51e-06
```

Cell-Means Coding

If we include all G dummy codes, we get a *cell-means* coded model.

- Cell-means coding estimates the so-called *normal means model*:

$$Y = \mu_g + \varepsilon$$

- We directly estimate each group-specific mean.
- We cannot estimate an intercept when using cell-means coded predictors.

Example Cell-Means Code

Let's look at the cell-means coding of biological sex:

	sex	female	male
1	female	1	0
2	male	0	1
3	male	0	1
4	female	1	0
5	male	0	1
6	female	1	0
7	female	1	0
8	male	0	1
9	female	1	0
10	female	1	0

Example Cell-Means Codes

Now, cell-means for the drinks example:

	drink	coffee	juice	tea
1	juice	0	1	0
2	coffee	1	0	0
3	tea	0	0	1
4	tea	0	0	1
5	tea	0	0	1
6	tea	0	0	1
7	juice	0	1	0
8	tea	0	0	1
9	coffee	1	0	0
10	juice	0	1	0

Using Cell-Means Codes

When using cell-means codes, we include all G codes into our model, so we must not estimate an intercept:

$$Y = \beta_1 X_{female} + \beta_2 X_{male} + \varepsilon$$

$$Y = \beta_1 X_{coffee} + \beta_2 X_{juice} + \beta_3 X_{tea} + \varepsilon$$

- Each "slope" is an estimate of the group-specific mean of Y in the coded group.
- The significance tests for the "slopes" are testing if the group-specific means are different from zero.

Example

First, an example with a two-level cell-means coded variable:

```
## Read in some data:
cDat <- readRDS("../data/cars_data.rds")

## Fit and summarize the model:
out4 <- lm(price ~ atOnly + mtOpt - 1, data = cDat)

## HACK: Add a new class attribute to dispatch
##       summary.cellMeans() in place of summary.lm():
class(out4) <- c("cellMeans", class(out4))
```

Example

```
partSummary(out4, -c(1, 2))

## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## atOnly   23.841      1.623   14.69   <2e-16
## mtOpt    17.238      1.175   14.67   <2e-16
##
## Residual standard error: 9.18 on 91 degrees of freedom
## Multiple R-squared:  0.1066,Adjusted R-squared:  0.09679
## F-statistic: 10.86 on 1 and 91 DF,  p-value: 0.001403
```

Example

Fit a model with a three-level factor:

```
out5 <- lm(price ~ four + front + rear - 1, data = cDat)
class(out5) <- c("cellMeans", class(out5))
partSummary(out5, -c(1, 2))

## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## four    17.630      2.761   6.385 7.33e-09
## front   17.536      1.067  16.439  < 2e-16
## rear    28.950      2.183  13.262  < 2e-16
##
## Residual standard error: 8.732 on 90 degrees of freedom
## Multiple R-squared:  0.2006,Adjusted R-squared:  0.1829
## F-statistic: 11.29 on 2 and 90 DF,  p-value: 4.202e-05
```

Effects Coding

Another useful form of categorical variable coding is *effects coding*.

- Effects codes can be *weighted* or *unweighted*.

Effects Coding

Another useful form of categorical variable coding is *effects coding*.

- Effects codes can be *weighted* or *unweighted*.

We'll first discuss *unweighted* effects codes.

- Unweighted effects codes are identical to dummy codes except that "reference group" rows get values of -1 on all codes.
- The intercept is interpreted as the unweighted mean of the group-specific means of Y.
- The slope associated with each code represents the difference between the coded group's mean of Y and the mean of the group-specific means of Y.

Example Unweighted Effects Codes

	sex	male.ec
1	female	-1
2	male	1
3	male	1
4	female	-1
5	male	1
6	female	-1
7	female	-1
8	male	1
9	female	-1
10	female	-1

	drink	juice.ec	tea.ec
1	juice	1	0
2	coffee	-1	-1
3	tea	0	1
4	tea	0	1
5	tea	0	1
6	tea	0	1
7	juice	1	0
8	tea	0	1
9	coffee	-1	-1
10	juice	1	0

Using Unweighted Effects Codes

We use the unweighted effects codes as we would use dummy codes.

- We include the $G - 1$ effects codes as $G - 1$ predictor variables in our regression model.

$$Y = \beta_0 + \beta_1 X_{male.ec} + \varepsilon$$

$$Y = \beta_0 + \beta_1 X_{juice.ec} + \beta_2 X_{tea.ec} + \varepsilon$$

- The intercept corresponds to the unweighted mean of the group-specific means of Y.
- Each slope represents the difference between the mean of Y in the coded group and the mean of the group-specific means of Y.

Example

```r
## Model with single effects code:
out6 <- lm(price ~ mtOpt.ec, data = cDat)
partSummary(out6, -c(1, 2))

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   20.539      1.002  20.501   <2e-16
## mtOpt.ec      -3.301      1.002  -3.295   0.0014
##
## Residual standard error: 9.18 on 91 degrees of freedom
## Multiple R-squared:  0.1066,Adjusted R-squared:  0.09679
## F-statistic: 10.86 on 1 and 91 DF,  p-value: 0.001403
```

Example

```
## Model with two effects codes (for a variable with G = 3):
out7 <- lm(price ~ front.ec + rear.ec, data = cDat)
partSummary(out7, -c(1, 2))

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   21.372      1.226  17.433  < 2e-16
## front.ec      -3.836      1.372  -2.796  0.00632
## rear.ec        7.578      1.758   4.310 4.16e-05
##
## Residual standard error: 8.732 on 90 degrees of freedom
## Multiple R-squared:  0.2006,Adjusted R-squared:  0.1829
## F-statistic: 11.29 on 2 and 90 DF,  p-value: 4.202e-05
```

Why is $\hat{\beta}_0$ the Unweighted Mean of Y?

First, define the group-specific means:

$$\begin{aligned}
\hat{\mu}_1 &= \hat{\beta}_0 + \hat{\beta}_1(1) \quad + \hat{\beta}_2(0) \quad = \hat{\beta}_0 + \hat{\beta}_1 \\
\hat{\mu}_2 &= \hat{\beta}_0 + \hat{\beta}_1(0) \quad + \hat{\beta}_2(1) \quad = \hat{\beta}_0 + \hat{\beta}_2 \\
\hat{\mu}_3 &= \hat{\beta}_0 + \hat{\beta}_1(-1) + \hat{\beta}_2(-1) = \hat{\beta}_0 - \hat{\beta}_1 - \hat{\beta}_2
\end{aligned}$$

Why is $\hat{\beta}_0$ the Unweighted Mean of Y?

First, define the group-specific means:

$$\begin{aligned}
\hat{\mu}_1 &= \hat{\beta}_0 + \hat{\beta}_1(1) \quad + \hat{\beta}_2(0) \quad = \hat{\beta}_0 + \hat{\beta}_1 \\
\hat{\mu}_2 &= \hat{\beta}_0 + \hat{\beta}_1(0) \quad + \hat{\beta}_2(1) \quad = \hat{\beta}_0 + \hat{\beta}_2 \\
\hat{\mu}_3 &= \hat{\beta}_0 + \hat{\beta}_1(-1) + \hat{\beta}_2(-1) = \hat{\beta}_0 - \hat{\beta}_1 - \hat{\beta}_2
\end{aligned}$$

Next, solve for $\hat{\beta}_1$ and $\hat{\beta}_2$:

$$\begin{aligned}
\hat{\beta}_1 &= \hat{\mu}_1 - \hat{\beta}_0 \\
\hat{\beta}_2 &= \hat{\mu}_2 - \hat{\beta}_0
\end{aligned}$$

Why is $\hat{\beta}_0$ the Unweighted Mean of Y?

First, define the group-specific means:

$$\begin{aligned}
\hat{\mu}_1 &= \hat{\beta}_0 + \hat{\beta}_1(1) \quad + \hat{\beta}_2(0) \quad = \hat{\beta}_0 + \hat{\beta}_1 \\
\hat{\mu}_2 &= \hat{\beta}_0 + \hat{\beta}_1(0) \quad + \hat{\beta}_2(1) \quad = \hat{\beta}_0 + \hat{\beta}_2 \\
\hat{\mu}_3 &= \hat{\beta}_0 + \hat{\beta}_1(-1) + \hat{\beta}_2(-1) = \hat{\beta}_0 - \hat{\beta}_1 - \hat{\beta}_2
\end{aligned}$$

Next, solve for $\hat{\beta}_1$ and $\hat{\beta}_2$:

$$\begin{aligned}
\hat{\beta}_1 &= \hat{\mu}_1 - \hat{\beta}_0 \\
\hat{\beta}_2 &= \hat{\mu}_2 - \hat{\beta}_0
\end{aligned}$$

Finally, substitute and solve for $\hat{\beta}_0$:

$$\begin{aligned}
\hat{\mu}_3 &= \hat{\beta}_0 - (\hat{\mu}_1 - \hat{\beta}_0) - (\hat{\mu}_2 - \hat{\beta}_0) \\
\hat{\mu}_3 &= 3\hat{\beta}_0 - \hat{\mu}_1 - \hat{\mu}_2 \\
\hat{\beta}_0 &= \frac{\hat{\mu}_1 + \hat{\mu}_2 + \hat{\mu}_3}{3}
\end{aligned}$$

Weighted Effects Coding

Weighted effects codes differ from the unweighted version only in how they code the "reference group" rows.

- In weighted effects codes the "reference group" rows get negative fractional values on all codes.
  - Let $g = 1, 2, \ldots, G$ index groups.
  - Take the first group as the "reference group."
  - Then, the gth code's reference group rows will take values of $-N_g/N_1$.
- The intercept is interpreted as the weighted mean of the group-specific outcome means.
  - The arithmetic mean of Y.
- Each slope represents the difference from that group's mean outcome and the overall mean of Y.

Example Weighted Effects Codes

	sex	male.wec
1	female	$-N_{male}/N_{female}$
2	male	1
3	male	1
4	female	$-N_{male}/N_{female}$
5	male	1
6	female	$-N_{male}/N_{female}$
7	female	$-N_{male}/N_{female}$
8	male	1
9	female	$-N_{male}/N_{female}$
10	female	$-N_{male}/N_{female}$

	drink	juice.wec	tea.wec
1	juice	1	0
2	coffee	$-N_{juice}/N_{coffee}$	$-N_{tea}/N_{coffee}$
3	tea	0	1
4	tea	0	1
5	tea	0	1
6	tea	0	1
7	juice	1	0
8	tea	0	1
9	coffee	$-N_{juice}/N_{coffee}$	$-N_{tea}/N_{coffee}$
10	juice	1	0

Using Weighted Effects Codes

Weighted effects codes work the same way as all of our other codes.

- As before, we include the $G - 1$ effects codes as $G - 1$ predictor variables in our regression model.

$$Y = \beta_0 + \beta_1 X_{male.wec} + \varepsilon$$

$$Y = \beta_0 + \beta_1 X_{juice.wec} + \beta_2 X_{tea.wec} + \varepsilon$$

- The intercept corresponds to the weighted mean of the group-specific means of Y (i.e., the arithmetic average of Y).
- Each slope represents the difference between the coded group's mean of Y and the overall mean of Y.

Example

```r
## Model with single effects code:
out8 <- lm(price ~ mtOpt.wec, data = cDat)
partSummary(out8, -c(1, 2))

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.5097     0.9519  20.495   <2e-16
## mtOpt.wec    -2.2720     0.6895  -3.295   0.0014
##
## Residual standard error: 9.18 on 91 degrees of freedom
## Multiple R-squared:  0.1066,Adjusted R-squared:  0.09679
## F-statistic: 10.86 on 1 and 91 DF,  p-value: 0.001403
```

Example

```
## Model with two effects codes (for a variable with G = 3):
out9 <- lm(price ~ front.wec + rear.wec, data = cDat)
partSummary(out9, -c(1, 2))

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.5097     0.9054  21.547  < 2e-16
## front.wec    -1.9739     0.5640  -3.500 0.000727
## rear.wec      9.4403     1.9863   4.753 7.57e-06
##
## Residual standard error: 8.732 on 90 degrees of freedom
## Multiple R-squared:  0.2006,Adjusted R-squared:  0.1829
## F-statistic: 11.29 on 2 and 90 DF,  p-value: 4.202e-05
```

Why is $\hat{\beta}_0$ the Weighted Mean of Y?

Define the group-specific means:

$$\hat{\mu}_1 = \hat{\beta}_0 + \hat{\beta}_1(1) \qquad + \hat{\beta}_2(0)$$

$$\hat{\mu}_2 = \hat{\beta}_0 + \hat{\beta}_1(0) \qquad + \hat{\beta}_2(1)$$

$$\hat{\mu}_3 = \hat{\beta}_0 + \hat{\beta}_1\left(\frac{-N_1}{N_3}\right) + \hat{\beta}_2\left(\frac{-N_2}{N_3}\right)$$

Why is $\hat{\beta}_0$ the Weighted Mean of Y?

Define the group-specific means:

$$\hat{\mu}_1 = \hat{\beta}_0 + \hat{\beta}_1(1) \qquad + \hat{\beta}_2(0)$$

$$\hat{\mu}_2 = \hat{\beta}_0 + \hat{\beta}_1(0) \qquad + \hat{\beta}_2(1)$$

$$\hat{\mu}_3 = \hat{\beta}_0 + \hat{\beta}_1\left(\frac{-N_1}{N_3}\right) + \hat{\beta}_2\left(\frac{-N_2}{N_3}\right)$$

Solve for $\hat{\beta}_1$ and $\hat{\beta}_2$:

$$\hat{\beta}_1 = \hat{\mu}_1 - \hat{\beta}_0$$
$$\hat{\beta}_2 = \hat{\mu}_2 - \hat{\beta}_0$$

Why is $\hat{\beta}_0$ the Weighted Mean of Y?

Define the group-specific means:

$$\hat{\mu}_1 = \hat{\beta}_0 + \hat{\beta}_1(1) \qquad + \hat{\beta}_2(0)$$

$$\hat{\mu}_2 = \hat{\beta}_0 + \hat{\beta}_1(0) \qquad + \hat{\beta}_2(1)$$

$$\hat{\mu}_3 = \hat{\beta}_0 + \hat{\beta}_1\left(\frac{-N_1}{N_3}\right) + \hat{\beta}_2\left(\frac{-N_2}{N_3}\right)$$

Solve for $\hat{\beta}_1$ and $\hat{\beta}_2$:

$$\hat{\beta}_1 = \hat{\mu}_1 - \hat{\beta}_0$$
$$\hat{\beta}_2 = \hat{\mu}_2 - \hat{\beta}_0$$

Substitute and solve for $\hat{\beta}_0$:

$$\hat{\mu}_3 = \hat{\beta}_0 + \left(\frac{-N_1}{N_3}\right)(\hat{\mu}_1 - \hat{\beta}_0) + \left(\frac{-N_2}{N_3}\right)(\hat{\mu}_2 - \hat{\beta}_0)$$

$$\hat{\mu}_3 = \frac{N_3}{N_3}\hat{\beta}_0 - \frac{N_1}{N_3}\hat{\mu}_1 + \frac{N_1}{N_3}\hat{\beta}_0 - \frac{N_2}{N_3}\hat{\mu}_2 + \frac{N_2}{N_3}\hat{\beta}_0$$

$$\hat{\mu}_3 = \frac{N_1 + N_2 + N_3}{N_3}\hat{\beta}_0 - \frac{N_1}{N_3}\hat{\mu}_1 - \frac{N_2}{N_3}\hat{\mu}_2$$

$$N_3\hat{\mu}_3 = (N_1 + N_2 + N_3)\hat{\beta}_0 - N_1\hat{\mu}_1 - N_2\hat{\mu}_2$$

$$\hat{\beta}_0 = \frac{N_1\hat{\mu}_1 + N_2\hat{\mu}_2 + N_3\hat{\mu}_3}{N_1 + N_2 + N_3}$$

Significance Testing for Categorical Variables

For variables with only two levels, we can test the overall factor's significance by evaluating the significance of its single code.

- This won't work when we're using cell-means coding.
- Cell-means coding will always produce two or more codes.

```
partSummary(out1, -c(1, 2))

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   23.841      1.623  14.691   <2e-16
## mtOpt         -6.603      2.004  -3.295   0.0014
##
## Residual standard error: 9.18 on 91 degrees of freedom
## Multiple R-squared:  0.1066,Adjusted R-squared:  0.09679
## F-statistic: 10.86 on 1 and 91 DF,  p-value: 0.001403
```

Significance Testing for Categorical Variables

For variables with more than two levels (or whenever using cell-means codes), we need to simultaneously evaluate the significance of all of the variable's codes.

```
partSummary(out3, -c(1, 2))

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.7187     2.9222   7.432 6.25e-11
## mtOpt        -5.8410     1.8223  -3.205  0.00187
## front        -0.2598     2.8189  -0.092  0.92677
## rear         10.5169     3.3608   3.129  0.00237
##
## Residual standard error: 8.314 on 89 degrees of freedom
## Multiple R-squared:  0.2834,Adjusted R-squared:  0.2592
## F-statistic: 11.73 on 3 and 89 DF,  p-value: 1.51e-06
```

Significance Testing for Categorical Variables

```
summary(out3)$r.squared - summary(out1)$r.squared

## [1] 0.1767569

anova(out1, out3)

## Analysis of Variance Table
##
## Model 1: price ~ mtOpt
## Model 2: price ~ mtOpt + front + rear
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     91 7668.9
## 2     89 6151.6  2    1517.3 10.976 5.488e-05 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Significance Testing for Categorical Variables

What about models where a single nominal factor is the only predictor?

```
partSummary(out2, -c(1, 2))

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.63000    2.76119   6.385 7.33e-09
## front       -0.09418    2.96008  -0.032  0.97469
## rear        11.32000    3.51984   3.216  0.00181
##
## Residual standard error: 8.732 on 90 degrees of freedom
## Multiple R-squared:  0.2006,Adjusted R-squared:  0.1829
## F-statistic: 11.29 on 2 and 90 DF,  p-value: 4.202e-05
```

Significance Testing for Categorical Variables

We can compare back to an "intercept-only" model.

```
out0 <- lm(price ~ 1, data = cDat)
partSummary(out0, -1)

## Residuals:
##    Min     1Q Median     3Q    Max
## -12.11  -7.31  -1.81   3.79  42.39
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   19.510      1.002   19.48   <2e-16
##
## Residual standard error: 9.659 on 92 degrees of freedom
```

Significance Testing for Categorical Variables

```
r2Diff <- summary(out2)$r.squared - summary(out0)$r.squared
r2Diff

## [1] 0.2006386

anova(out0, out2)

## Analysis of Variance Table
##
## Model 1: price ~ 1
## Model 2: price ~ front + rear
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     92 8584.0
## 2     90 6861.7  2    1722.3 11.295 4.202e-05 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Significance Testing for Categorical Variables

We don't actually need to do the explicit model comparison, though.

```
r2Diff

## [1] 0.2006386

summary(out2)$r.squared

## [1] 0.2006386

anova(out0, out2)[2, "F"]

## [1] 11.29494

summary(out2)$fstatistic[1]

##    value
## 11.29494
```

Compare Codings

Let's dig into some numerical properties of the three coding schemes.

```
## Fit models using all four codings:
dcOut  <- lm(price ~ front + rear,                data = cDat)
cmOut  <- lm(price ~ four + front + rear - 1, data = cDat)
ecOut  <- lm(price ~ front.ec + rear.ec,         data = cDat)
wecOut <- lm(price ~ front.wec + rear.wec,     data = cDat)

## Compute group-specific means of 'price':
grpMeans <- tapply(cDat$price, cDat$dr, mean)
```

Compare the parameter estimates to their theoretical equivalents:

```
coef(dcOut)[1] - grpMeans["4WD"]

##   (Intercept)
## -1.421085e-14

coef(cmOut) - grpMeans

##           4WD         Front          Rear
##  0.000000e+00  7.105427e-15 -3.552714e-15

coef(ecOut)[1] - mean(grpMeans)

##   (Intercept)
## -3.552714e-15

coef(wecOut)[1] - mean(cDat$price)

##   (Intercept)
## -1.065814e-14
```

Compare the R^2 values from each coding scheme:

```
summary(dcOut)$r.squared

## [1] 0.2006386

summary.cellMeans(cmOut)$r.squared

## [1] 0.2006386

summary(ecOut)$r.squared

## [1] 0.2006386

summary(wecOut)$r.squared

## [1] 0.2006386
```

Compare the F-statistics:

```
summary(dcOut)$fstatistic

##    value    numdf    dendf
## 11.29494  2.00000 90.00000

summary.cellMeans(cmOut)$fstatistic

##    value    numdf    dendf
## 11.29494  2.00000 90.00000

summary(ecOut)$fstatistic

##    value    numdf    dendf
## 11.29494  2.00000 90.00000

summary(wecOut)$fstatistic

##    value    numdf    dendf
## 11.29494  2.00000 90.00000
```

Compare the residual standard errors:

```
summary(dcOut)$sigma

## [1] 8.731638

summary.cellMeans(cmOut)$sigma

## [1] 8.731638

summary(ecOut)$sigma

## [1] 8.731638

summary(wecOut)$sigma

## [1] 8.731638
```

Choosing a Coding Scheme

Any valid coding scheme will represent the information in the categorical variable equally well.

- All valid coding schemes produce equivalent models.

We choose a particular coding scheme based on the interpretations that we want.

- Dummy coding is useful with a meaningful reference group.
    - Control group in an experiment
    - An "industry standard" or benchmark implementation of some feature
- Dummy coding is also preferred if we don't care about interpretation.
    - Dummy codes are the simplest to construct.

Choosing a Coding Scheme

- Cell-means coding is useful when you want to directly test for non-zero means within each group.
  - The interpretation of cell-means effects is probably the most intuitive of any coding scheme, as well.
- Weighted effects codes are good when you believe your sample is representative of the population.
  - Larger groups should be weighted more heavily in the model.
  - Parameter estimates will correctly generalize to the population.

Choosing a Coding Scheme

- Unweighted effects codes are good when the group sizes in your sample do not generalize to the population.
  - Convenience samples, for example, are usually not representative.
  - When your sample is not representative, larger groups should not be weighted more heavily.
  - Unweighted effects codes are “agnostic” to differing group sizes.
    - We need to be careful with very small groups.
- Weighted effects codes with known weights are another option.

Conclusion

When we use categorical predictors, they must be coded before entering the model.

- We discussed four of the most popular coding schemes:
  1. Dummy coding
  2. Cell-means coding
  3. Unweighted effects coding
  4. Weighted effects coding

- Apart from cell-means, these coding schemes differ primarily in how the "reference" group is defined.

All valid coding schemes produce equivalent models.

- We choose a particular scheme for interpretational convenience.