Coding Categorical Variables

BIOS 6611

CU Anschutz

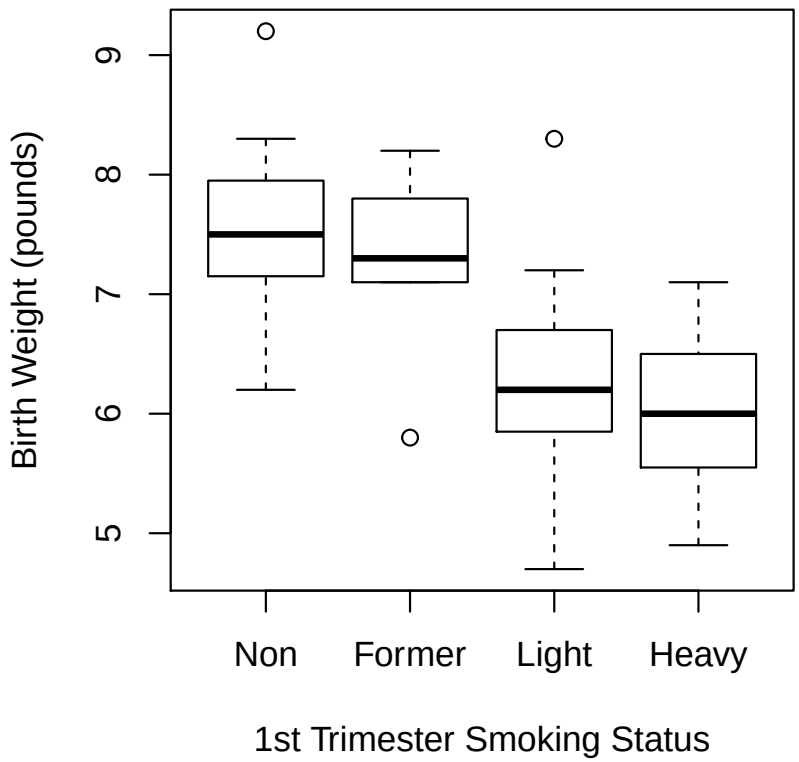
Week 11

Categorical Explanatory Variables: More than 2 Categories

Categorical Explanatory Variables: More than 2 Categories

Motivating Example: interested in relationship between infant birthweight (pounds) and mother's smoking status during first trimester. Sample pregnant women from a larger study from each of **4 smoking categories**.

	Smoking Status			
i	**non**	**Former**	**Light**	**Heavy**
1	7.5	5.8	5.9	6.2
2	6.2	7.3	6.2	6.8
3	6.9	8.2	5.8	5.7
4	7.4	7.1	4.7	4.9
5	9.2	7.8	8.3	6.2
6	8.3		7.2	7.1
7	7.6		6.2	5.8
8				5.4
$\bar{Y}_j$	7.59	7.24	6.33	6.01
s_j^2	0.93	0.83	1.3	0.52



Motivating Example (cont.)

Potential Scientific Questions:

- Is there an association between smoking status and birthweight?
- Is there a difference in birthweight between non and former smokers?

Indicator ("dummy") variables

To address these questions in a regression model, can create indicator or "dummy" variables for each category:

$$I_{non} = \begin{cases} 1 & \text{if smoke=non} \\ 0 & \text{if otherwise} \end{cases} \qquad I_{former} = \begin{cases} 1 & \text{if smoke=former} \\ 0 & \text{if otherwise} \end{cases}$$

$$I_{light} = \begin{cases} 1 & \text{if smoke=light} \\ 0 & \text{if otherwise} \end{cases} \qquad I_{heavy} = \begin{cases} 1 & \text{if smoke=heavy} \\ 0 & \text{if otherwise} \end{cases}$$

In general, I_{heavy} is used to represent indicator (a.k.a. "dummy") variables.

May also see $I(heavy = 1)$ or $I_{heavy=1}$.

Coding indicator variables in R

```
BWT <- read.csv("birthweight_smoking_dataset.csv", header=T)

# Create indicator variables
BWT$Xf <- BWT$momsmoke=='Former'
BWT$Xl <- BWT$momsmoke=='Light'
BWT$Xh <- BWT$momsmoke=='Heavy'
BWT$Xn <- BWT$momsmoke=='Non'

BWT[c(6,7,9,10,14,15,20,21),]
```

```
##    birthwt momsmoke    Xf    Xl    Xh    Xn
## 6      8.3      Non FALSE FALSE FALSE  TRUE
## 7      7.6      Non FALSE FALSE FALSE  TRUE
## 9      7.3   Former  TRUE FALSE FALSE FALSE
## 10     8.2   Former  TRUE FALSE FALSE FALSE
## 14     6.2    Light FALSE  TRUE FALSE FALSE
## 15     5.8    Light FALSE  TRUE FALSE FALSE
## 20     6.2    Heavy FALSE FALSE  TRUE FALSE
## 21     6.8    Heavy FALSE FALSE  TRUE FALSE
```

Reference Cell Models

Reference Cell Models

Reference Cell Coding: Leave one indicator variable out of model, which is represented by the intercept. This is called a *reference cell model.*

For example, making non-smoker the reference category, the reference cell model is:

$$E[\text{birthweight}] = \beta_0 + \beta_{former} I_{former} + \beta_{light} I_{light} + \beta_{heavy} I_{heavy}$$

The estimated mean of each group can be obtained by setting the corresponding indicators equal to 0 or 1:

$$\begin{aligned} E[\text{birthweight}|\text{non}] &= \beta_0 = \mu_{non} \\ E[\text{birthweight}|\text{former}] &= \beta_0 + \beta_{former} = \mu_{former} \\ E[\text{birthweight}|\text{light}] &= \beta_0 + \beta_{light} = \mu_{light} \\ E[\text{birthweight}|\text{heavy}] &= \beta_0 + \beta_{heavy} = \mu_{heavy} \end{aligned}$$

Reference Cell: Estimating Means between Groups

The regression parameters (β's) can be used to estimate the difference between the mean of any two groups.

The expected difference in birthweight between a mother who lightly smoked and a mother who was a non-smoker is:

$$\begin{aligned} E[\text{birthweight}|\text{light}] - E[\text{birthweight}|\text{non}] &= (\beta_0 + \beta_{light}) - \beta_0 \\ &= \beta_{light} \end{aligned}$$

The expected difference in birthweight between a mother who lightly smoked and a mother who heavily smoked is:

$$\begin{aligned} E[\text{birthweight}|\text{light}] - E[\text{birthweight}|\text{heavy}] &= (\beta_0 + \beta_{light}) - (\beta_0 + \beta_{heavy}) \\ &= \beta_{light} - \beta_{heavy} \end{aligned}$$

In general, for reference cell coding, the intercept is the expected mean for the reference group, and every other coefficient is the expected change in the mean from the reference group to the corresponding coefficient's group.

Testing a Category's Coefficient with Indicator Variables

The test of one category's coefficient is *conceptually* equivalent to a t-test of that category against the reference category. However, it is **not** mathematically identical.

- Using a "pooled variance" from **all four** groups, not just the two groups being compared

Note: the parameter estimate and some of the p-values will change if the reference category is changed. (Why?)

The overall F-test or partial F-test (if more than one categorical predictor) can used to test the overall significance of the categorical variable (i.e., does mothers smoking status effect birthweight). Will not change if reference category is changed.

Association between Smoking and Birthweight

To test if there is an association between smoking and birthweight, we can test the null hypothesis:

$$
\begin{aligned}
H_0: &\quad \beta_{former} = \beta_{light} = \beta_{heavy} = 0 \\
\iff & \\
H_0: &\quad \beta_{former} + \beta_0 = \beta_{light} + \beta_0 = \beta_{heavy} + \beta_0 = +\beta_0 \\
\iff & \\
H_0: &\quad \mu_{former} = \mu_{light} = \mu_{heavy} = \mu_{non}
\end{aligned}
$$

Our hypothesis can be evaluated using the Overall F-test!

Association between Smoking and Birthweight

```
fullmod <- glm(birthwt~Xf+Xl+Xh,data=BWT)
nullmod <- glm(birthwt~1,data=BWT)
anova(fullmod,nullmod,test='F')
```

```
## Analysis of Deviance Table
##
## Model 1: birthwt ~ Xf + Xl + Xh
## Model 2: birthwt ~ 1
##   Resid. Df Resid. Dev Df Deviance      F  Pr(>F)
## 1        23     20.304
## 2        26     31.976 -3  -11.673 4.4076 0.01371 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion: there is evidence that mother's smoking status (the *entire* set of indicator variables) contributes significantly to the prediction of birthweight.

To know *which* smoking levels are significant, need to do *post-hoc* testing (future lecture).

Change Reference Category

```
mod_nonref <- lm(birthwt~Xf+Xl+Xh,data=BWT) # non is reference group
summary(mod_nonref)
```

```
##
## Call:
## lm(formula = birthwt ~ Xf + Xl + Xh, data = BWT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6286 -0.4786 -0.1286  0.6371  1.9714
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.5857     0.3551  21.361  < 2e-16 ***
## XfTRUE       -0.3457     0.5501  -0.628  0.53593
## XlTRUE       -1.2571     0.5022  -2.503  0.01985 *
## XhTRUE       -1.5732     0.4863  -3.235  0.00366 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9396 on 23 degrees of freedom
## Multiple R-squared:  0.365,  Adjusted R-squared:  0.2822
## F-statistic: 4.408 on 3 and 23 DF,  p-value: 0.01371
```

Change Reference Category (cont.)

```
mod_heavyref <- lm(birthwt~Xf+Xl+Xn,data=BWT) # heavy is reference group
summary(mod_heavyref)
```

```
##
## Call:
## lm(formula = birthwt ~ Xf + Xl + Xn, data = BWT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6286 -0.4786 -0.1286  0.6371  1.9714
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.0125     0.3322  18.100 4.21e-15 ***
## XfTRUE        1.2275     0.5356   2.292  0.03141 *
## XlTRUE        0.3161     0.4863   0.650  0.52213
## XnTRUE        1.5732     0.4863   3.235  0.00366 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9396 on 23 degrees of freedom
## Multiple R-squared:  0.365,  Adjusted R-squared:  0.2822
## F-statistic: 4.408 on 3 and 23 DF,  p-value: 0.01371
```

Change Reference Category (cont.)

We see the overall F-test does not depend on choice of reference group used in the model. However, the parameter estimate table *does*. This is because we are now comparing each category to a different reference group.

Reference Cell Interpretation

Let's interpret the model with non-smoker as the reference category:

$$\hat{Y} = 7.59 + (-0.35) \times I_{former} + (-1.26) \times I_{light} + (-1.57) \times I_{heavy}$$

What is the interpretation of the intercept?
The expected mean birthweight for the non-smokers (the reference group) is 7.59.

What is the expected birthweight for heavy smokers?
$\hat{Y} = 7.59 + (-0.35) \times 0 + (-1.26) \times 0 + (-1.57) \times 1 = 6.02$
The expected birthweight for baby's born from mothers who heavily smoked is 6.02 pounds.

Reference Cell Interpretation

Let's interpret the model with non-smoker as the reference category:

$$\hat{Y} = 7.59 + (-0.35) \times I_{former} + (-1.26) \times I_{light} + (-1.57) \times I_{heavy}$$

What is the expected difference in birthweight between heavy smokers and non-smokers? Is this difference significant?

The expected difference between heavy smokers and non-smokers is $\hat{\beta}_{heavy} = -1.57$. The corresponding parameter p-value is 0.004, indicating this difference is significant. (Again, this is mathematically *different* from a t-test because we are using the pooled variance from all four categories.)

Reference Cell Interpretation (cont.)

$$\hat{Y} = 7.59 + (-0.35) \times I_{former} + (-1.26) \times I_{light} + (-1.57) \times I_{heavy}$$

What is the estimated difference in average birthweight between heavy smokers and light smokers?

$$\begin{aligned} E[Y|heavy] - E[Y|light] &= (\hat{\beta}_0 + \hat{\beta}_{heavy}) - (\hat{\beta}_0 + \hat{\beta}_{light}) = \hat{\beta}_{heavy} - \hat{\beta}_{light} \\ &= -1.57 - (-1.26) = -0.31 \end{aligned}$$

Is the estimated difference between light and heavy smokers significantly different from zero?

...

Is the estimated difference between light and heavy smokers significantly different from zero?

$$
\begin{aligned}
t &= \frac{\hat{\beta}_{heavy} - \hat{\beta}_{light}}{SE(\hat{\beta}_{heavy} - \hat{\beta}_{light})} \\
&= \frac{\hat{\beta}_{heavy} - \hat{\beta}_{light}}{\sqrt{Var(\hat{\beta}_{heavy}) + Var(\hat{\beta}_{light}) - 2Cov(\hat{\beta}_{heavy}, \hat{\beta}_{light})}} \\
&= \frac{-1.57 - (-1.26)}{\sqrt{0.2522 + 0.2365 - 2(0.126)}} = -0.65 \sim t_{23}; p = 0.522
\end{aligned}
$$

Alternatively, we could have fit a model with either heavy or light as the reference group, and looked at the corresponding parameter estimate and p-value. If we look at the output from when we fit heavy as the reference group, the results agree!

Note: variance and covariance estimates were obtained using `vcov(mod_nonref)`.

Other Coding Schemes and Analysis Strategies for Categorical Variables

Other Coding Schemes and Analysis Strategies for Categorical Variables

In addition to the reference cell model, there are other approaches to model categorical variables.

Effect Coding: uses -1,0,1 for classification

Cell Means: includes *all* dummy variables, excludes intercept. For a model with only one categorical predictor (with ≥ 2 levels), is extremely similar to one-way ANOVA that we will discuss in another lecture.

Continuous: could treat categories as single continuous predictor. For example, non-smoker=0, former=1, light=2, heavy=3. Uses less degrees of freedom, but we are assuming linearity between levels (i.e., the difference between levels 0 and 1 is equivalent to between 1 and 2, and 2 and 3, which is often unlikely.)

Effect Coding

Effect Coding: uses 1, 0, -1 to classify each category. Intercept represents the grand mean, $\frac{\mu_1+...+\mu_k}{k}$, and β coefficients represent deviation in category mean from the grand mean. Here, non-smokers is the -1 "reference" category, and e1-e3 represent other categories:

Group	*e1*	*e2*	*e3*
Non-smokers	-1	-1	-1
Former smokers	1	0	0
Light Smokers	0	1	0
Heavy Smokers	0	0	1

If have an interaction of 2 categorical variables, then effect coding directly provides estimates of the main effects and interaction. In contrast, with reference cell coding, we obtain simple effects, i.e., the effect of one variable at a particular level of the other variable.

Example with Cell Means Model

```r
cellmeans_mod <- lm(birthwt~-1+momsmoke,data=BWT) # cell means model
summary(cellmeans_mod)
```

```
##
## Call:
## lm(formula = birthwt ~ -1 + momsmoke, data = BWT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6286 -0.4786 -0.1286  0.6371  1.9714
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## momsmokeFormer   7.2400     0.4202   17.23 1.21e-14 ***
## momsmokeHeavy    6.0125     0.3322   18.10 4.21e-15 ***
## momsmokeLight    6.3286     0.3551   17.82 5.88e-15 ***
## momsmokeNon      7.5857     0.3551   21.36  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9396 on 23 degrees of freedom
## Multiple R-squared:  0.9838, Adjusted R-squared:  0.981
## F-statistic: 349.6 on 4 and 23 DF,  p-value: < 2.2e-16
```

Cell Means Model Interpretation

From previous slide, our cell means model is:

$$\begin{aligned} E[Y] &= \beta_{non} I_{non} + \beta_{former} I_{former} + \beta_{light} I_{light} + \beta_{heavy} I_{heavy} \\ \hat{Y} &= 7.59 I_{non} + 7.24 I_{former} + 6.33 I_{light} + 6.01 I_{heavy} \end{aligned}$$

Coefficient interpretation:

$$\begin{aligned} E[birthweight|non] &= \beta_{non} = \mu_{non} \\ E[birthweight|former] &= \beta_{former} = \mu_{former} \\ E[birthweight|light] &= \beta_{light} = \mu_{light} \\ E[birthweight|heavy] &= \beta_{heavy} = \mu_{heavy} \end{aligned}$$

Cell Means model interpretation (cont.)

What is the expected change in mean birthweight between former smokers and non-smokers?

$$E[birthweight|former] - E[birthweight|non] = \beta_{former} - \beta_{non}$$

What is the overall F-test testing?

$$\begin{aligned} H_0: &\quad \beta_{non} = \beta_{former} = \beta_{light} = \beta_{heavy} = 0 \\ \iff & \\ H_0: &\quad \mu_{non} = \mu_{former} = \mu_{light} = \mu_{heavy} = 0 \end{aligned}$$

Careful of what is being testing!

Continuous Variable (no dummy codes)

```
# Create continuous variable
BWT$contsmoke<-ifelse(BWT$Xn==TRUE,0,ifelse(BWT$Xf==TRUE,1,
                  ifelse(BWT$Xl==TRUE,2,ifelse(BWT$Xh==TRUE,3,NA))))
cont_mod<-lm(birthwt~contsmoke,data=BWT) # Fit continuous model
summary(cont_mod)
```

```
##
## Call:
## lm(formula = birthwt ~ contsmoke, data = BWT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8034 -0.5759 -0.1138  0.6914  1.7966
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.6138     0.2971  25.628  < 2e-16 ***
## contsmoke    -0.5552     0.1507  -3.685  0.00111 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9104 on 25 degrees of freedom
## Multiple R-squared:  0.352,  Adjusted R-squared:  0.3261
## F-statistic: 13.58 on 1 and 25 DF,  p-value: 0.001107
```

Continuous Variable (no dummy codes) (cont.)

$$\hat{Y} = 7.61 - 0.56 \times Group$$

Interpretation of intercept? Expected birthweight for a non-smoking mother is 7.61 pounds. (Group=0)

Interpretation of $\hat{\beta}_{group}$? On average, birthweight decreases by 0.56 pounds for every increase in category in smoking status (assumed to be the same increase between all adjacent categories.)