23-24. Categorical Predictors and Testing General Linear Hypotheses

Readings: Kleinbaum, Kupper, Nizam, and Rosenberg (KKNR): Ch. 12

SAS: PROC REG

Homework: Homework 9 due by 11:59 pm on November 28
Final Project due by 11:59 on December 6

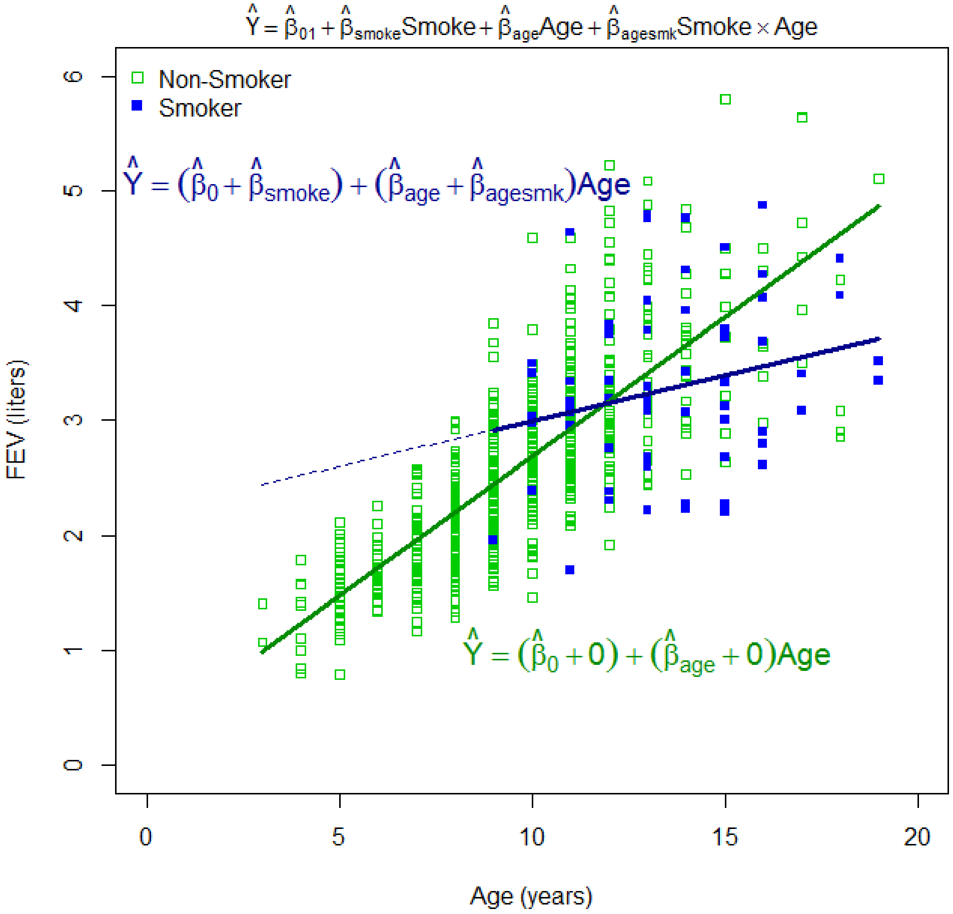
Overview

A) Re/Preview of Topics
B) Categorical Predictors with >2 Categories (Reference Cell Model)
C) Other Coding Strategies for Categorical Variables (Cell Means, Effect Coding, Continuous)
D) Tests of General Linear Hypotheses
E) Contrasts
F) ANOVA Table and Degrees of Freedom Summary

A. Review (Lecture 22)/Current (Lecture 23-24)/ Preview (Lecture 25)

Lecture 22:

- Effect Modification (Interactions):
  - $E[FEV_i] = \beta_0 + \beta_{age} Age_i + \beta_{smoke} Smoke_i + \beta_{agesmk} Age_i \times Smoke_i$
  - Allows for different slopes (FEV vs. age) for smokers and non-smokers
- MLE vs LSE (same β's, different variance)



Lectures 23-24:

- Categorical Predictors
  - Indicator variables
- Test of general linear hypothesis
- Contrasts

Lecture 25:

- Polynomial Regression: quadratic, cubic, quartic
- Other remedies for non-linearity

B. Categorical Explanatory Variables: More Than 2 Categories

Motivating Example: An investigator is interested in studying the relationship between infant birthweight (pounds) and smoking status of the mother during the first trimester. The investigator chose **five pregnant women from each of four smoking categories** (X) (*never*, *former*, *light*, and *heavy* smokers, with X coded 0, 1, 2, 3) from a larger study. The table below provides the birthweight (Y) of each baby and the average birthweight for each smoking category.

	Never Smokers (X=0)	*Former Smokers (X=1)*	*Light Smokers (X=2)*	*Heavy Smokers (X=3)*
	7.50	5.80	5.90	6.20
	6.20	7.30	6.20	6.80
	6.90	8.20	5.80	5.70
	7.40	7.10	4.70	4.90
	9.20	7.80	8.30	6.20
$\overline{Y}\|X_i$	7.44	7.24	6.18	5.96
${S^2}_{Y\|X_i}$	1.233	0.833	1.727	0.503

```
proc import
datafile="~/birthweight_smoking_5per
group_dataset.csv"
     out=bwt5 /* name for data set
for SAS to reference */
     dbms=csv /* identify file as
csv */
     replace; /* overwrite BWT if
already present */
     getnames=yes; /* take first row
as column names from data */
run;
```

Potential Scientific Questions:

- Is there an association between smoking status and birthweight?
- Is there a difference in birthweight between never smokers and former smokers?
- Is there a difference in birthweight between non-smokers and current smokers?
- Is there an association between smoking and birthweight adjusting for weight of the mother?

To address these scientific questions in a regression model, you can create a different indicator variable or “dummy variable” for each of the categories:

$$\text{never}=\begin{cases}1 \text{ if } \text{smoke}=0.\\ 0 \text{ if } \text{smoke}=1,2,3.\end{cases} \qquad \text{former}=\begin{cases}1 \text{ if } \text{smoke}=1.\\ 0 \text{ if } \text{smoke}=0,2,3.\end{cases}$$

$$\text{light}=\begin{cases}1 \text{ if } \text{smoke}=2.\\ 0 \text{ if } \text{smoke}=0,1,3.\end{cases} \qquad \text{heavy}=\begin{cases}1 \text{ if } \text{smoke}=3.\\ 0 \text{ if } \text{smoke}=0,1,2.\end{cases}$$

Any *three* of these indicator variables can be used in the model if an intercept is included.

```
/* create dummy variables */
DATA bwt5;
    set bwt5;

    *** Create dummy variables ****;
    IF momsmoke = 'Never' THEN Never = 1; ELSE Never = 0;
    IF momsmoke = 'Former' THEN Former = 1; ELSE Former = 0;
    IF momsmoke = 'Light' THEN Light = 1; ELSE Light = 0;
    IF momsmoke = 'Heavy' THEN Heavy = 1; ELSE Heavy = 0;

    *** Create variable for two groups with current status ****;
    IF momsmoke = 'Never' THEN group = 0;
    IF momsmoke = 'Former' THEN group = 1;
    IF momsmoke = 'Light' THEN group = 2;
    IF momsmoke = 'Heavy' THEN group = 3;

    non = (group = 0 or group = 1);
    smoke = (group = 2 or group = 3);

RUN;
```

Notes on Using Indicator Variables: Reference Cell Models

The reference category is the category associated with the indicator variable left out of the model (if specifying a model with an intercept). This is called a *reference cell model.*

Using never smoker as the reference category:

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

From this regression equation we can still determine the estimated mean for each group:

$$\begin{aligned}
E[\text{birthweight}|\text{never}] &= \beta_0 &&= \mu_{never} \\
E[\text{birthweight}|\text{former}] &= \beta_0 + \beta_{former} &&= \mu_{former} \\
E[\text{birthweight}|\text{light}] &= \beta_0 + \beta_{light} &&= \mu_{light} \\
E[\text{birthweight}|\text{heavy}] &= \beta_0 + \beta_{heavy} &&= \mu_{heavy}
\end{aligned}$$

The β’s can be used to estimate the difference between the mean of any two groups:

$$\begin{aligned}
E[\text{birthweight}|\text{light}] - E[\text{birthweight}|\text{never}] &= (\beta_0 + \beta_{light}) - \beta_0 &&= \beta_{light} \\
E[\text{birthweight}|\text{heavy}] - E[\text{birthweight}|\text{never}] &= (\beta_0 + \beta_{heavy}) - \beta_0 &&= \beta_{heavy} \\
E[\text{birthweight}|\text{light}] - E[\text{birthweight}|\text{heavy}] &= (\beta_0 + \beta_{light}) - (\beta_0 + \beta_{heavy}) &&= \beta_{light} - \beta_{heavy}
\end{aligned}$$

Notes on Using Indicator Variables (cont.)

From our regression equation, we can conduct the **Overall F-test** and make the direct connection to the one-way ANOVA:

$H_0: \beta_{former} = \beta_{light} = \beta_{heavy} = 0$ (Step 1: add β_0 to the H_0)

$H_0: \beta_{former} + \beta_0 = \beta_{light} + \beta_0 = \beta_{heavy} + \beta_0 = \beta_0$ (Step 2: substitute in definition for μ_x)

$H_0: \mu_{former} = \mu_{light} = \mu_{heavy} = \mu_{never}$

The intercept represents the level of the outcome in the reference category:

$$E[\text{birthweight}] = \beta_0 + \beta_{former} I_{former} + \beta_{light} I_{light} + \beta_{heavy} I_{heavy} \Rightarrow E[\text{birthweight} \,|\, \text{never}] = \beta_0$$

You can choose a different reference category by selecting which indicator variables are included in the model. For example, if we made heavy smoking mothers our reference category:

$$E[\text{birthweight}] = \beta_0^* + \beta_{never}^* I_{never} + \beta_{former}^* I_{former} + \beta_{light}^* I_{light} \Rightarrow E[\text{birthweight} \,|\, \text{heavy}] = \beta_0^*$$

Notes on Using Indicator Variables – Testing A Category's Coefficient

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

- Because we are using a "pooled variance" from all four categories/groups
- Not just the two groups we are comparing

The parameter estimates and some of the p-values for the parameter estimates will change if the reference category is changed.

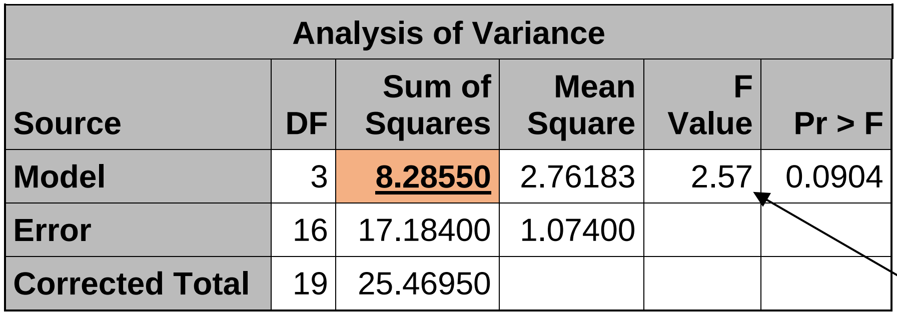
The *F* test or partial *F* test can be used to test the overall significance of the categorical variable (this does not depend on the reference category).

- The *F* test and partial *F* test will **not** change if the reference category is changed.

Association between smoking and birthweight (reference group: never smokers)

```
/* REFERENCE CELL MODEL (reference group: never smokers) */
PROC REG DATA=bwt5;
   MODEL birthwt = former light heavy;
RUN;
```

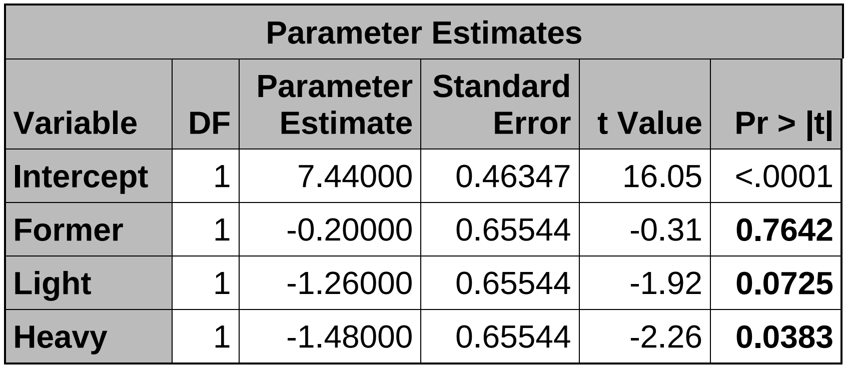
$$E[\text{birthweight}] = \beta_0 + \beta_{former} I_{former} + \beta_{light} I_{light} + \beta_{heavy} I_{heavy}$$



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	**8.28550**	2.76183	2.57	0.0904
Error	16	17.18400	1.07400		
Corrected Total	19	25.46950			

SS explained by smoking status.



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.44000	0.46347	16.05	<.0001
Former	1	-0.20000	0.65544	-0.31	**0.7642**
Light	1	-1.26000	0.65544	-1.92	**0.0725**
Heavy	1	-1.48000	0.65544	-2.26	**0.0383**

H_0: $\beta_{former}=\beta_{light}=\beta_{heavy}=0$

or

H_0: $\beta_{former}+\beta_0=\beta_{light}+\beta_0=\beta_{heavy}+\beta_0 = \beta_0$

or

H_0: $\mu_{former}=\mu_{light}=\mu_{heavy}=\mu_{never}$

Global Hypotheses vs Multiple Comparisons

If you **reject** the null hypothesis H_0: $\mu_{former}=\mu_{light}=\mu_{heavy}=\mu_{never}$ (which we did **not**, p=0.0904) **AND** you want to perform 6 additional tests, *then*:

Correct the alpha level, ***if*** you perform the additional tests (**Review Lectures 13-14**)

1. H_0: $\mu_{former} = \mu_{never}$
2. H_0: $\mu_{light} = \mu_{never}$
3. H_0: $\mu_{heavy} = \mu_{never}$
4. H_0: $\mu_{former} = \mu_{light}$
5. H_0: $\mu_{former} = \mu_{heavy}$
6. H_0: $\mu_{light} = \mu_{heavy}$

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	
Intercept	1	7.44000	0.46347	16.05	<.0001	
Former	1	-0.20000	0.65544	-0.31	**0.7642**	$H_0: \mu_{former} = \mu_{never}$
Light	1	-1.26000	0.65544	-1.92	**0.0725**	$H_0: \mu_{light} = \mu_{never}$
Heavy	1	-1.48000	0.65544	-2.26	**0.0383**	$H_0: \mu_{heavy} = \mu_{never}$

This is the explanation as to why H_0: $\mu_{heavy} = \mu_{never}$ is rejected at alpha=0.05, *but* the overall F-test for H_0: $\mu_{former}=\mu_{light}=\mu_{heavy}=\mu_{never}$ is not significant.

Note: no multiple comparison correction is needed for the null hypothesis of all means are equal (H_0: $\mu_{former}=\mu_{light}=\mu_{heavy}=\mu_{never}$), because it is only 1 test.

Association between smoking and birthweight (reference group: never smokers)

```
PROC REG DATA=bwt5;
    MODEL birthwt = former light heavy / covb;
RUN;
```

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

Analysis of Variance

Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	**8.28550**	2.76183	2.57	0.0904
Error	16	17.18400	1.07400		
Corrected Total	19	25.46950			

SS explained by smoking status.

H_0: $\beta_{former}=\beta_{light}=\beta_{heavy}=0$

or

H_0: $\beta_{former}+\beta_0=\beta_{light}+\beta_0=\beta_{heavy}+\beta_0 = \beta_0$

or

H_0: $\mu_{former}=\mu_{light}=\mu_{heavy}=\mu_{never}$

Parameter Estimates

Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	7.44000	0.46347	16.05	<.0001
Former	1	-0.20000	0.65544	-0.31	**0.7642**
Light	1	-1.26000	0.65544	-1.92	**0.0725**
Heavy	1	-1.48000	0.65544	-2.26	**0.0383**

Overall test: Does smoking status (the *entire set* of indicator variables) contribute significantly to the prediction of birthweight?

H_0: $\beta_{former} = \beta_{light} = \beta_{heavy} = 0$; No, because F=2.57, p=0.0904.

REDUCED MODEL for Partial *F*

```
PROC REG DATA=bwt5;
    MODEL birthwt = ;
RUN;
```

$$E[\text{birthweight}] = \beta_0 = \bar{Y}$$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	0	0	.	.	.
Error	19	25.46950	1.34050		
Corrected Total	19	25.46950			

Root MSE	1.15780	R-Square	0.0000
Dependent Mean	6.70500	Adj R-Sq	0.0000
Coeff Var	17.26771		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.70500	0.25889	25.90	<.0001

Overall *F* test (using Partial *F* test):

$$\frac{[SS_{model}(full) - SS_{model}(reduced)]/k}{MS_{error}(full)} = \frac{[SS_{model}(full) - 0]/k}{MS_{error}(full)} = \frac{MS_{model}(full)}{MS_{error}(full)} = \frac{2.76183}{1.07400} = \mathbf{2.57}$$

Association between smoking and birthweight (reference group: heavy smokers)

```
PROC REG DATA=bwt5;
   MODEL birthwt = never former light;
RUN;
```

$$E[\text{birthweight}] = \beta_0 + \beta_{never}I_{never} + \beta_{former}I_{former} + \beta_{light}I_{light}$$

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	8.28550	2.76183	2.57	0.0904
Error	16	17.18400	1.07400		
Corrected Total	19	25.46950			

H_0: $\beta_{never}=\beta_{former}=\beta_{light}=0$
or
H_0: $\mu_{former}=\mu_{light}=\mu_{heavy}=\mu_{never}$

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.96000	0.46347	12.86	<.0001
Never	1	1.48000	0.65544	2.26	0.0383
Former	1	1.28000	0.65544	1.95	0.0686
Light	1	0.22000	0.65544	0.34	0.7415

The overall F test does not depend on the choice of reference group used in the model. The parameter estimates table *does* depend on the choice of indicator variables. Note that the ANOVA table is identical to the results on slide 10, but the parameter estimates table has changed.

Tests of Individual Coefficients (reference group: never smokers)

```
PROC REG DATA=bwt5;
   MODEL birthwt = former light heavy / covb;
RUN;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	8.28550	2.76183	2.57	0.0904
Error	16	17.18400	1.07400		
Corrected Total	19	25.46950			

H_0: $\beta_{former}=\beta_{light}=\beta_{heavy}=0$

or

H_0: $\mu_{former}=\mu_{light}=\mu_{heavy}=\mu_{never}$

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.44000	0.46347	16.05	<.0001
Former	1	-0.20000	0.65544	-0.31	0.7642
Light	1	-1.26000	0.65544	-1.92	0.0725
Heavy	1	-1.48000	0.65544	-2.26	0.0383

Covariance of Estimates

Variable	Intercept	Former	Light	Heavy
Intercept	0.2148	-0.2148	-0.2148	-0.2148
Former	-0.2148	0.4296	0.2148	0.2148
Light	-0.2148	0.2148	0.4296	0.2148
Heavy	-0.2148	0.2148	0.2148	0.4296

$$\boldsymbol{\Sigma} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}^2_{Y|X}$$

$$\hat{Y} = 7.44 + (-0.20) \times \text{former} + (-1.26) \times \text{light} + (-1.48) \times \text{heavy}$$

$$\hat{Y} = 7.44 + (-0.20) \times \text{former} + (-1.26) \times \text{light} + (-1.48) \times \text{heavy}$$

What is the interpretation of the intercept?

This is the expected mean birthweight for the reference group (non-smokers) or expected birthweight for an individual baby born to a non-smoking mother.

$\hat{Y}$ = 7.44 + (-0.20)×0 +(-1.26)×0 + (-1.48)×0 = 7.44 lbs

What is the expected birthweight for former smokers?

$\hat{Y}$ = 7.440 + (-0.20)×**1** +(-1.26)×0 + (-1.48)×0 = 7.24 lbs

What is the expected birthweight for heavy smokers?

$\hat{Y}$ = 7.44 + (-0.20)×0 +(-1.26)×0 + (-1.48)×**1** = 5.96 lbs

What is the difference in expected birthweight between heavy smokers and never smokers?

E[birthweight|heavy]-E[birthweight|never] = $\left(\beta_0 + \beta_{heavy}\right) - \beta_0 = \beta_{heavy}$
t = $\hat{\beta}_{heavy}/SE\left(\hat{\beta}_{heavy}\right)$ = -1.48/0.65544 = -2.26, p = 0.0383

Why isn't this mathematically the same as an independent samples t-test?

Because we are using a "pooled variance" from all four smoking categories/groups, not just the two groups we are comparing

Form of the Variance Covariance Matrix

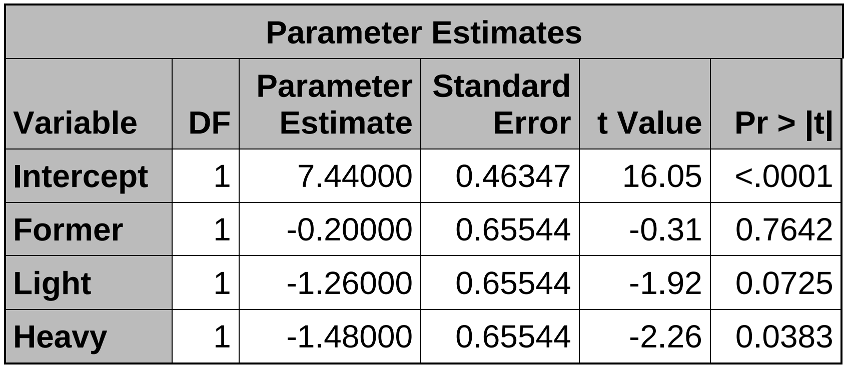
$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 &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

$$Var(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2_{Y|X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

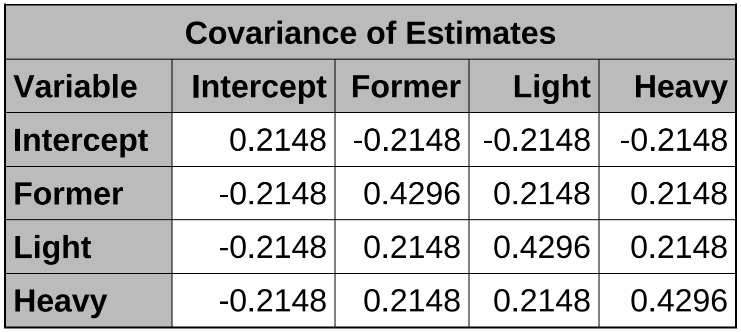
$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} 20 & 5 & 5 & 5 \\ 5 & 5 & 0 & 0 \\ 5 & 0 & 5 & 0 \\ 5 & 0 & 0 & 5 \end{bmatrix}$$

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix} 0.2 & -0.2 & -0.2 & -0.2 \\ -0.2 & 0.4 & 0.2 & 0.2 \\ -0.2 & 0.2 & 0.4 & 0.2 \\ -0.2 & 0.2 & 0.2 & 0.4 \end{bmatrix}$$

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



Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.44000	0.46347	16.05	<.0001
Former	1	-0.20000	0.65544	-0.31	0.7642
Light	1	-1.26000	0.65544	-1.92	0.0725
Heavy	1	-1.48000	0.65544	-2.26	0.0383



Covariance of Estimates				
Variable	Intercept	Former	Light	Heavy
Intercept	0.2148	-0.2148	-0.2148	-0.2148
Former	-0.2148	0.4296	0.2148	0.2148
Light	-0.2148	0.2148	0.4296	0.2148
Heavy	-0.2148	0.2148	0.2148	0.4296

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

E[birthweight|heavy]-E[birthweight|light] = $\left(\beta_0 + \beta_{\text{heavy}}\right) - \left(\beta_0 + \beta_{\text{light}}\right) = \beta_{\text{heavy}} - \beta_{\text{light}}$

Then $\hat{\beta}_{\text{heavy}}$-$\hat{\beta}_{\text{light}}$= -1.48 – (-1.26) = -0.22

Is this difference significantly different from zero?

$$t = \frac{\hat{\beta}_{heavy} - \hat{\beta}_{light}}{SE\left(\hat{\beta}_{heavy} - \hat{\beta}_{light}\right)} = \frac{\hat{\beta}_{heavy} - \hat{\beta}_{light}}{\sqrt{Var\left(\hat{\beta}_{heavy}\right) + Var\left(\hat{\beta}_{light}\right) - 2Cov\left(\hat{\beta}_{heavy}, \hat{\beta}_{light}\right)}}$$

$$= \frac{-1.48 - (-1.26)}{\sqrt{0.4296 + 0.4296 - 2 * 0.2148}} = \frac{-0.22}{\sqrt{0.4296}} = 0.336 \sim t_{16}; p = 0.742$$

C. Other Coding Schemes and Analysis Strategies for Categorical Variables

In addition to the reference cell model, there are many other schemes and approaches one can use to model categorical variables and their relationship with the outcome.

Cell Means: this approach fits the reference cell model with *all* dummy variables included and the intercept excluded. For a model with only one categorical predictor (with $\geq$2 levels), this approach is extremely similar to the one-way ANOVA model discussed in Lecture 13.

Effect Coding: this approach uses 1, 0, and -1 as values to classify each category (some sources note any coding can be used as long as they sum to 0). In this model, the intercept term represents the grand mean, $\frac{\mu_1+\cdots+\mu_k}{k}$, of the data, and the beta coefficients represent the deviation in the category mean from the given grand mean. Below is an example where non-smokers is the -1 “reference” category, and e1-e3 represent estimates for the other smoking 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

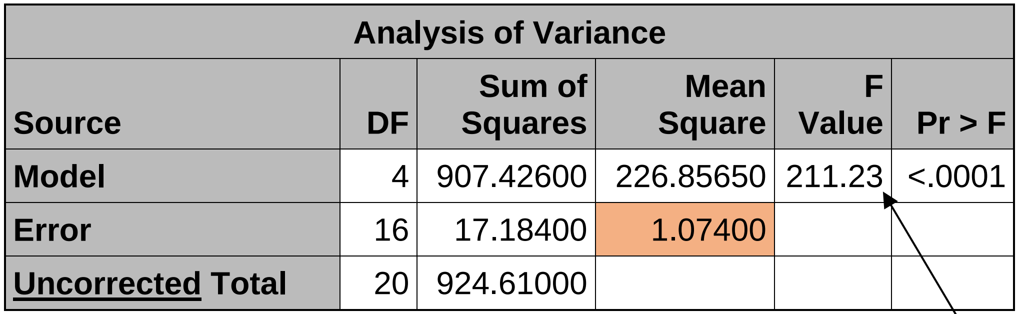
Continuous: alternatively, we could treat the categories as a single continuous predictor. For example, non-smoker=0, former smoker=1, light smoker=2, and heavy smoker=3. While this approach uses less degrees of freedom, we are assuming we have ratio data (i.e., the move from 0 to 1 is equivalent to the move to 1 to 2, which is an unlikely assumption in our smoking example).

Cell Means Model Example: Mother's Smoking Status and Birthweight

```
PROC REG DATA=bwt5;
    MODEL birthwt = never former light heavy / noint;
RUN;
```

NOTE: No intercept in model. R-Square is redefined. It uses the uncorrected sum of squares and is not meaningful to compare to the R^2 from models which include an intercept.

$$E[birthweight] = \beta_{never} I_{never} + \beta_{former} I_{former} + \beta_{light} I_{light} + \beta_{heavy} I_{heavy}$$



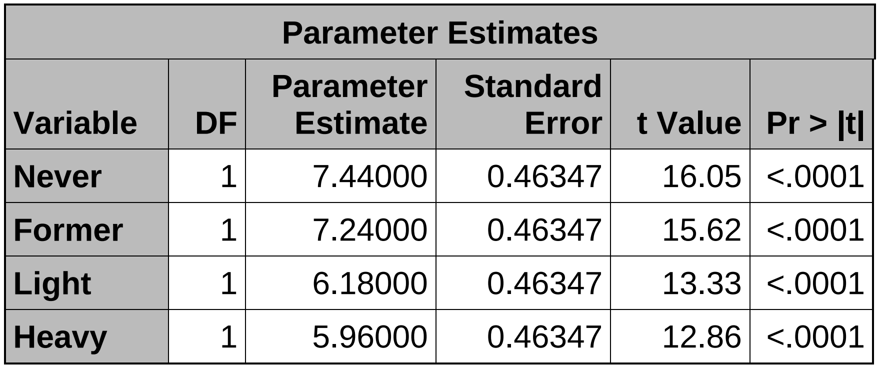
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	907.42600	226.85650	211.23	<.0001
Error	16	17.18400	1.07400		
Uncorrected Total	20	924.61000			

H_0: $\beta_{never}=\beta_{former}=\beta_{light}=\beta_{heavy}=0$

or

H_0: $\mu_{never}=\mu_{former}=\mu_{light}=\mu_{heavy}=0$



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Never	1	7.44000	0.46347	16.05	<.0001
Former	1	7.24000	0.46347	15.62	<.0001
Light	1	6.18000	0.46347	13.33	<.0001
Heavy	1	5.96000	0.46347	12.86	<.0001

Group Means

Notes on Using Indicator Variables – No Intercept (i.e., Cell Means Model)

Model **without** an intercept:

$$E[\text{birthweight}] = \beta_{never} I_{never} + \beta_{former} I_{former} + \beta_{light} I_{light} + \beta_{heavy} I_{heavy}$$

$$E[\text{birthweight} \,|\, \text{never}] = \beta_{never} = \mu_{never}$$
$$E[\text{birthweight} \,|\, \text{former}] = \beta_{former} = \mu_{former}$$
$$E[\text{birthweight} \,|\, \text{light}] = \beta_{light} = \mu_{light}$$
$$E[\text{birthweight} \,|\, \text{heavy}] = \beta_{heavy} = \mu_{heavy}$$

$$E[\text{birthweight} \,|\, \text{former}] - E[\text{birthweight} \,|\, \text{never}] = \beta_{former} - \beta_{never}$$
$$E[\text{birthweight} \,|\, \text{light}] - E[\text{birthweight} \,|\, \text{never}] = \beta_{light} - \beta_{never}$$
$$E[\text{birthweight} \,|\, \text{heavy}] - E[\text{birthweight} \,|\, \text{never}] = \beta_{heavy} - \beta_{never}$$

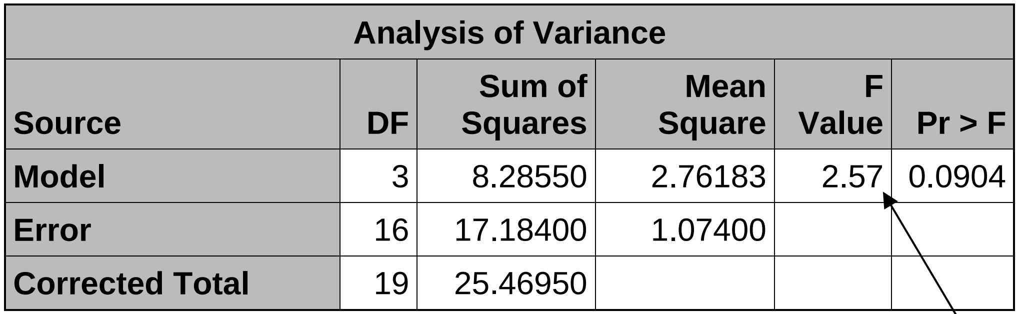
Overall F-test for the model without an intercept:

$$H_0: \beta_{never} = \beta_{former} = \beta_{light} = \beta_{heavy} = 0 \Rightarrow H_0: \mu_{former} = \mu_{light} = \mu_{heavy} = \mu_{never} = 0$$

Effect Coding Example: Mother's Smoking Status and Birthweight

```
PROC REG DATA=bwt5;
    MODEL birthwt = e1 e2 e3;
RUN;
```

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

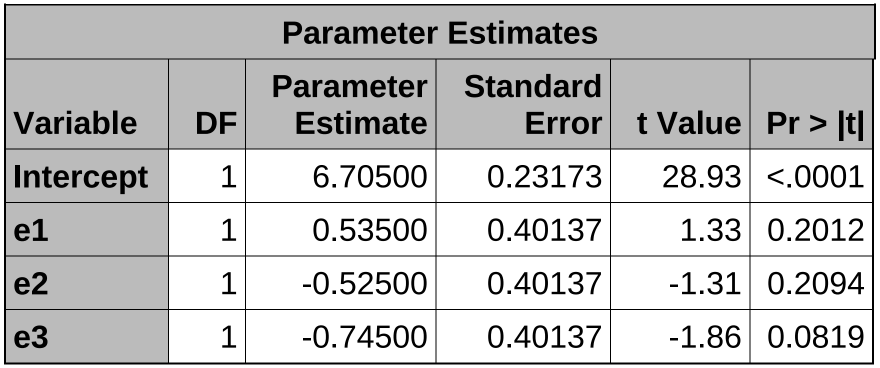


Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	8.28550	2.76183	2.57	0.0904
Error	16	17.18400	1.07400		
Corrected Total	19	25.46950			

H_0: $\beta_{former}=\beta_{light}=\beta_{heavy}=0$

or

H_0: $\mu_{never}=\mu_{former}=\mu_{light}=\mu_{heavy}=0$



Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.70500	0.23173	28.93	<.0001
e1	1	0.53500	0.40137	1.33	0.2012
e2	1	-0.52500	0.40137	-1.31	0.2094
e3	1	-0.74500	0.40137	-1.86	0.0819

Group Means

Note, the ANOVA table here matches our reference cell model on Slide 13.

Notes on Using Indicator Variables – Effect Coding

Based on our fitted regression model on the previous slide:

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

$$E[\text{birthweight} \,|\, \text{never}] = \beta_0 - (\beta_{former} + \beta_{light} + \beta_{heavy}) = \mu_{never}$$
$$E[\text{birthweight} \,|\, \text{former}] = \beta_0 + \beta_{former} = \mu_{former}$$
$$E[\text{birthweight} \,|\, \text{light}] = \beta_0 + \beta_{light} = \mu_{light}$$
$$E[\text{birthweight} \,|\, \text{heavy}] = \beta_0 + \beta_{heavy} = \mu_{heavy}$$

$$E[\text{birthweight} \,|\, \text{former}] - E[\text{birthweight} \,|\, \text{never}] = 2 \times \beta_{former} + \beta_{light} + \beta_{heavy}$$
$$E[\text{birthweight} \,|\, \text{light}] - E[\text{birthweight} \,|\, \text{never}] = \beta_{former} + 2 \times \beta_{light} + \beta_{heavy}$$
$$E[\text{birthweight} \,|\, \text{heavy}] - E[\text{birthweight} \,|\, \text{never}] = \beta_{former} + \beta_{light} + 2 \times \beta_{heavy}$$

Overall F-test for the model without an intercept:

$$H_0: \beta_{former} = \beta_{light} = \beta_{heavy} = 0 \Rightarrow H_0: \mu_{former} = \mu_{light} = \mu_{heavy} = \mu_{never} = 0$$

Model treating Smoking Status as a *continuous* variable (no dummy codes):

```
PROC REG DATA=bwt5;
    MODEL birthwt = group;
RUN;
```

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	7.56250	7.56250	7.60	0.0130
Error	18	17.90700	0.99483		
Corrected Total	19	25.46950			

H_0: $\beta_{smkgroup} = 0$

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	7.53000	0.37320	20.18	<.0001
group	1	-0.55000	0.19948	-2.76	0.0130

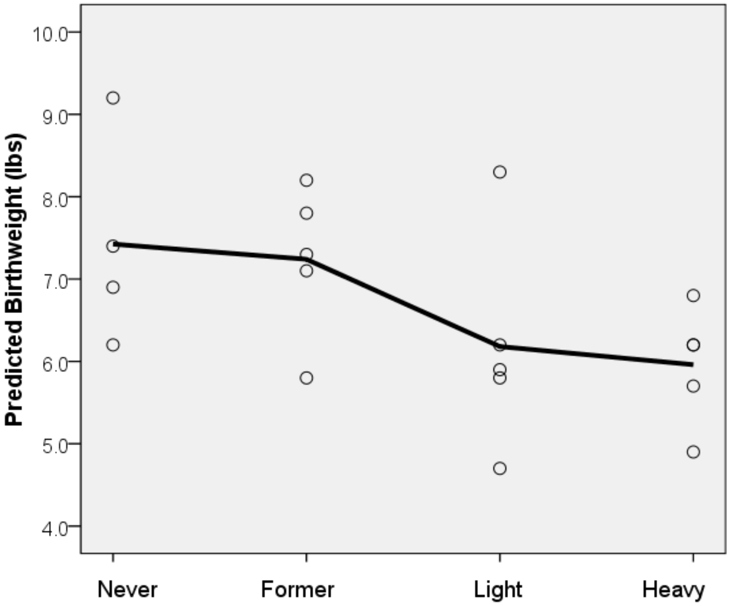
$$\hat{Y} = 7.53 + (-0.55) \times Group$$

$$\hat{Y} = 7.53 + (-0.55) \times Group$$

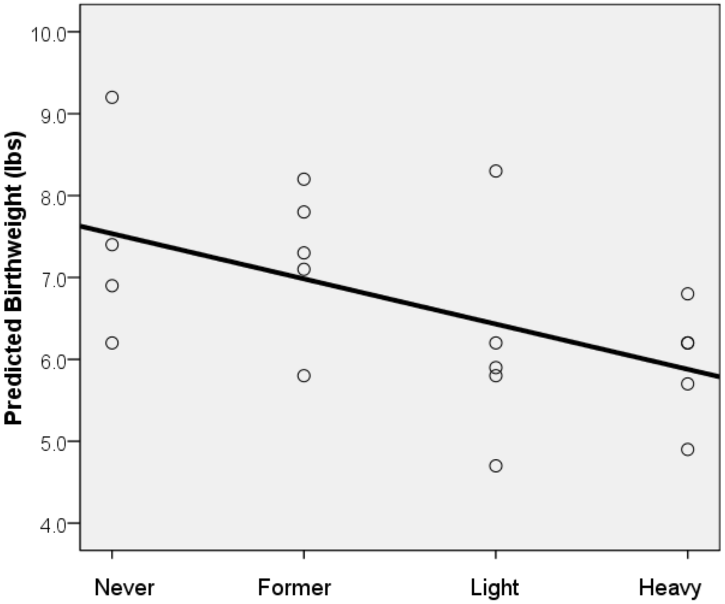
Interpretation of the intercept? Expected birthweight for a non-smoking mother is 7.53 lbs. $E[\text{birthweight}] = \beta_0 + \beta_{group}Group \Rightarrow E[\text{birthweight} \,|\, \text{never}] = \beta_0$

Interpretation of $\hat{\beta}_1$? On average, birthweight decreases by 0.55 pounds for every category increase in smoking status (assumed to be the same increase between all adjacent categories).

Predicted Model using Dummy Coding (using 4 degrees of freedom)



Predicted Model using Continuous Variable (using 2 degrees of freedom)



D. Tests of General Linear Hypotheses

Tests on individual regression parameters and on subsets of parameters can be put into a more general framework that allows much more flexibility by the use of the general linear hypothesis:

$$H_0: \mathbf{c}\boldsymbol{\beta} = \mathbf{d}$$
$$H_1: \mathbf{c}\boldsymbol{\beta} \neq \mathbf{d}$$

The matrix **c** is an $r \times p^*$ matrix that is of rank r and $r \leq p^*$, where

- $p^* = p$ when an intercept is included in the model
- $p^* = p$-1 for a no intercept model.

In other words, we can postulate $r \leq p^*$ non-redundant and non-contradictory statements about the parameters.

We can use this framework to test a single parameter:

$$H_0: \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = 0 \Rightarrow H_0: \beta_2 = 0$$

We can use this framework to compare two or more parameters:

$$H_0: \begin{pmatrix} 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = 0 \Rightarrow H_0: \beta_1 - \beta_2 = 0 \text{ or } H_0: \beta_1 = \beta_2$$

We can use this framework for simultaneous hypothesis tests:

$$H_0: \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow H_0: \begin{matrix} \beta_2 = 0 \\ \beta_1 = 0 \end{matrix}$$

$$H_0: \begin{pmatrix} 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow H_0: \begin{matrix} \beta_1 - \beta_2 = 0 \\ \beta_1 - \beta_3 = 0 \end{matrix} \text{ or } H_0: \begin{matrix} \beta_1 = \beta_2 \\ \beta_1 = \beta_3 \end{matrix}$$

The *F*-test can be used to test our general linear hypotheses:

$$F = \frac{[(\boldsymbol{c}\widehat{\boldsymbol{\beta}} - \boldsymbol{d})^T (\boldsymbol{c}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{c}^T)^{-1} (\boldsymbol{c}\widehat{\boldsymbol{\beta}} - \boldsymbol{d})/r]}{\hat{\sigma}_{Y|X}^2} \sim F_{r,n-p-1}$$

where *r* is the number of linear combinations of parameters we wish to test (which is equal to the number of rows in **c**).

This reduces to our Partial *F* test for testing a group of variables, since:

$$(\boldsymbol{c}\widehat{\boldsymbol{\beta}})^T (\boldsymbol{c}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{c}^T)^{-1} (\boldsymbol{c}\widehat{\boldsymbol{\beta}}) = SS_{model}(full) - SS_{model}(reduced).$$

And reduces to our *t* test for a single parameter (or test of a single linear hypothesis):

$$t = \frac{\boldsymbol{c}\widehat{\boldsymbol{\beta}}}{\sqrt{\boldsymbol{c}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{c}^T \hat{\sigma}_{Y|X}^2}}.$$

Recall:
$(X^T X)^{-1} \hat{\sigma}_{Y|X}^2 = \Sigma$

In terms of the covariance matrix of $\boldsymbol{\beta}$, **Σ**, the *F* and *t* tests become:

$$F = (\boldsymbol{c}\widehat{\boldsymbol{\beta}} - \boldsymbol{d})^T (\boldsymbol{c}\sum \boldsymbol{c}^T)^{-1} (\boldsymbol{c}\widehat{\boldsymbol{\beta}} - \boldsymbol{d})/r \sim F_{r,n-1-p}$$

$$t = (\boldsymbol{c}\widehat{\boldsymbol{\beta}})\left(\sqrt{\boldsymbol{c}\sum \boldsymbol{c}^T}\right)^{-1} \sim t_{n-1-p}$$

Tests of General Linear Hypotheses: Example

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

We want to test the hypothesis: H$_0$: $\beta_{\text{heavy}} = \beta_{\text{light}}$, or equivalently H$_0$: $\beta_{\text{heavy}} - \beta_{\text{light}} = 0$

Which can also be written as: $H_0: \begin{pmatrix} 0 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_{former} \\ \beta_{light} \\ \beta_{heavy} \end{pmatrix} = 0$

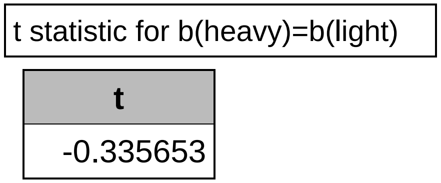
$$\boldsymbol{b} = \hat{\boldsymbol{\beta}} = \begin{pmatrix} 7.440 \\ \text{-}0.200 \\ \text{-}1.260 \\ \text{-}1.480 \end{pmatrix} \quad \boldsymbol{\Sigma} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\hat{\sigma}^2_{Y|X} = \begin{pmatrix} 0.2148 & \text{-}0.2148 & \text{-}0.2148 & \text{-}0.2148 \\ \text{-}0.2148 & 0.4296 & 0.2148 & 0.2148 \\ \text{-}0.2148 & 0.2148 & 0.4296 & 0.2148 \\ \text{-}0.2148 & 0.2148 & 0.2148 & 0.4296 \end{pmatrix}$$

$$t = (\boldsymbol{cb})\left(\sqrt{\boldsymbol{c\Sigma c}'}\right)^{-1} \quad t = \begin{pmatrix} 0 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} 7.440 \\ \text{-}0.200 \\ \text{-}1.260 \\ \text{-}1.480 \end{pmatrix} \left(\sqrt{ \begin{pmatrix} 0 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} 0.2148 & \text{-}0.2148 & \text{-}0.2148 & \text{-}0.2148 \\ \text{-}0.2148 & 0.4296 & 0.2148 & 0.2148 \\ \text{-}0.2148 & 0.2148 & 0.4296 & 0.2148 \\ \text{-}0.2148 & 0.2148 & 0.2148 & 0.4296 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ -1 \\ 1 \end{pmatrix} } \right)^{-1}$$

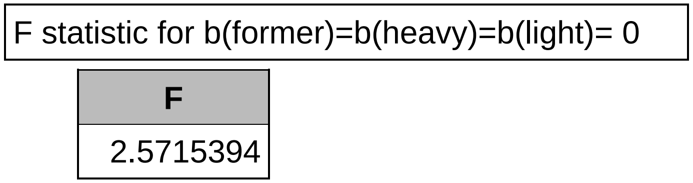
$$t = \left(-1.260 - (-1.480)\right)\left(\sqrt{(0.4296\text{+}0.4296\text{-}2 \times 0.2148)}\right)^{-1} = \frac{0.22}{0.6554} = 0.336 \sim t_{16}$$

Testing Generalized Linear Hypotheses in SAS with Matrices using PROC IML

```
PROC IML;
 beta = {7.44000, -0.20000,  -1.26000,  -1.48000};

 sigma ={0.2148         -0.2148         -0.2148         -0.2148,
       -0.2148          0.4296          0.2148          0.2148,
       -0.2148          0.2148          0.4296          0.2148,
       -0.2148          0.2148          0.2148          0.4296};

PRINT "t statistic for b(heavy)=b(light)";
c = {0 0 -1 1};
t = (c*beta)*INV(SQRT(c*sigma*c`));
PRINT t;

PRINT "F statistic for b(former)=b(heavy)=b(light)= 0";
c = {0 1 0 0, 0 0 1 0, 0 0 0 1};
F = (c*beta)`*INV(c*sigma*c`)*(c*beta)/NROW(c);
PRINT F;
```



```
t statistic for b(heavy)=b(light)

        t
-0.335653
```



```
F statistic for b(former)=b(heavy)=b(light)= 0

        F
2.5715394
```

Testing General Linear Hypotheses Directly in SAS

```
PROC REG DATA=bwt5;
    MODEL birthwt = former light heavy;

    /* these 3 statement request the equivalent test */
    TEST light = heavy;
    TEST light-heavy;
    TEST light-heavy=0;

    /* these 3 statement request the equivalent test */
    TEST former=light=heavy=0;
    TEST former,light,heavy;
    test former=0, light=0, heavy=0;
RUN;
```

Test 1 Results for Dependent Variable birthwt

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.12100	0.11	0.7415
Denominator	16	1.07400		

Test 4 Results for Dependent Variable birthwt

Source	DF	Mean Square	F Value	Pr > F
Numerator	3	2.76183	2.57	0.0904
Denominator	16	1.07400		

Note: Tests 2 and 3 results are identical to Test 1 and are not shown here. Similarly, Tests 5 and 6 are identical to Test 4 and are not shown here.

Compare Test 4’s F- and p-value to ANOVA table on Slides 8/10.

E. Contrasts

Contrasts are most often used to test linear combinations of group means in a Cell Means Model, but can also be utilized in other modeling approaches as well.

A linear contrast (L) is any linear combination of the parameters such that the linear coefficients add up to 0. Specifically,

$$L = \sum_{i=1}^{k} c_i \mu_i \quad \text{where} \quad \sum_{i=1}^{k} c_i = 0$$

Orthogonal contrasts are a set of contrasts such that for all pairs of contrasts, the cross-products of the coefficients is zero (assuming equal sample sizes):

$$\sum_{i=1}^{k} \frac{c_{Ai} c_{Bi}}{n_i} = 0 \qquad \text{or} \qquad \sum_{i=1}^{k} c_{Ai} c_{Bi} = 0 \text{ (when the } n_i\text{'s are equal.)}$$

Orthogonality is a desirable property because the Model Sums of Squares is partitioned into statistically independent sums of squares.

Orthogonal polynomial contrasts are orthogonal contrasts that test for polynomial patterns in the data by removing the inherent multicollinearity associated with polynomial terms.

These are described in more detail in the Lecture S1 (supplemental) slides.

F. ANOVA Table and Degrees of Freedom Summary

Now that we have two approaches for regression modeling (the reference cell/group model *with* an intercept **versus** the cell means model *without* an intercept) we can summarize the similarities and differences between the two approaches.

Let N be the overall sample size and assume that we do not necessarily have the same X's between models (X versus X^*). For the given regression equation, the ANOVA tables will be:

Reference Cell Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$ (note there are **p+1** beta coefficients)

SAS ANOVA Table for Reference Cell Model (Includes Intercept)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model (Explained Variability)	p	SS_{Model}	SS_{Model} / p	MS_{Model}/MS_{Error}	Compare to $F_{p,N\text{-}p\text{-}1}$ distribution
Error (Unexplained Variability)	N-p-1	SS_{Error}	SS_{Error} / (N-p-1)		
Corrected Total	N-1	SS_{Total}			

Cell Means Model: $\hat{Y}^* = \hat{\beta}_1^* X_1^* + \cdots + \hat{\beta}_k^* X_k^*$ (note there are **k** beta coefficients)

SAS ANOVA Table for Cell Means Model (Does NOT Include Intercept)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model (Explained Variability)	k	SS_{Model}	SS_{Model} / k	MS_{Model}/MS_{Error}	Compare to $F_{k,N\text{-}k}$ distribution
Error (Unexplained Variability)	N-k	SS_{Error}	SS_{Error} / (N-k)		
Uncorrected Total	N	SS_{Total}			

Note: If we fit a reference cell model with 3 dummy variables for a 4-category variable, the corresponding cell means model has 4 dummy variables. In this context p=3 for the reference cell model and k=p+1=4 for our cell means model.

Why the different degrees of freedom between the two modeling approaches?

First, let us revisit the type of variability described by our ANOVA table sources:

- ***Model:*** the variation from our observed outcome explained by the regression model (i.e., the variation explained by the X's we included in our model)
- ***Error:*** the variation from our observed outcome that is not explained by the regression model (i.e., it is unlikely that we have all relevant X's in our model and there will be some variation from our outcome that is explained by variables that are not included in the regression model)
- ***Total:*** the variation of our outcome (i.e., the difference between each observed outcome and the mean of the outcome)

The error degrees of freedom for the reference cell model is **N-p-1**, which accounts for the estimation of our intercept term plus all p slope beta coefficients.

However, when we do not estimate an intercept we are essentially fixing this value at 0 (i.e., $\beta_0 = 0$). By not estimating this term, we have gained 1 degree of freedom since we have one fewer parameter to estimate for the mean. However, it does not change the model degrees of freedom because there is no variable (X) associated with the intercept. The intercept is a constant that applies to every observation equally in our linear regression models.

How do we determine the degrees of freedom to use for the different approaches to statistical inference for regression that we have covered so far? (Lecture 19)

Overall F Test for the Entire Set of Independent Variables

If we are interested in testing the null hypothesis that the entire set of beta coefficients for our independent variables is equal to 0 (i.e., $H_0\colon \beta_1 = \cdots = \beta_p = 0$), the degrees of freedom for the numerator of our F distribution is p and for the denominator is:

- N-p-1 for the reference cell model ($F_{p,N-p-1}$)
- N-k for the cell means model ($F_{k,N-k}$)

Testing Addition of a Single Variable

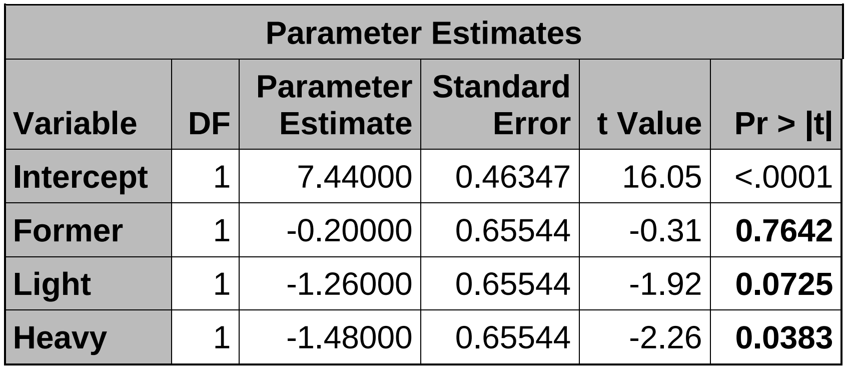
If we are interested in testing if *one* particular independent variable adds significantly to the prediction of the outcome over and above that achieved by the other independent variables already present in the model, then we use:

- N-p-1 degrees of freedom for the reference cell model (t_{N-p-1})
- N-k for the cell means model (t_{N-k})

The reason we use N-p-1 or N-k degrees of freedom is because we need to account for the estimation of the other beta coefficients (including the intercept for the reference cell model).

But wait, why does the Parameter Estimates table provide a degrees of freedom column with the “wrong” degrees of freedom!?

Indeed, this is one of the potentially misleading aspects of the PROC REG output. Consider the parameter estimates table from Slide 8 (never smoker used as the reference category):



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.44000	0.46347	16.05	<.0001
Former	1	-0.20000	0.65544	-0.31	**0.7642**
Light	1	-1.26000	0.65544	-1.92	**0.0725**
Heavy	1	-1.48000	0.65544	-2.26	**0.0383**

Here the DF column does not correspond to the degrees of freedom used in the “Pr > |t|” column, but it represents the degrees of freedom contributed by that specific variable. Unless the model is not full rank (e.g., if we included an intercept *and* all 4 smoker categories), the DF column will always list 1.

Given that we used the dummy variable representation so that the included Former, Light, and Heavy all contribute 1 degree of freedom, we can also note that the overall categorical variable of Smoking Status contributes the number of groups minus 1 degree of freedom, or simply the sum of the DF in our table: 4-1 = 1+1+1 = 3.