

Week 2b, Thursday
Models for Nominal Covariates

Dummy Variables: An Introduction
(Vittinghoff et al., 2005 pp. 76-77)

By the end of this lecture you should be able to fit and interpret linear regression models with a single independent variable which is either:

1. binary (e.g. smoker vs. non-smoker); or
2. nominal having three or more groups (e.g. mother's race classified as white, black, other).

- Dummy variables are used to model nominal scale variables (e.g. smoking status, race).
- Dummy variables are also known as indicator or design variables.
- Simple linear regression of birth weight in grams (Y_i) on smoking status (X_i):

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_i$$

where $X_i = 1$ for mother's who smoked during pregnancy and $X_i = 0$ for mother's who did not smoke during pregnancy.

**Dummy Variables:
Interpreting Regression Coefficients**

- Mean birth weight (g): non-smokers ($X_i = 0$)

$$\begin{aligned} E(Y_i|X_i = 0) &= \beta_0 + (\beta_1 \times 0) \\ &= \beta_0. \end{aligned}$$

- Mean birth weight (g): smokers ($X_i = 1$)

$$\begin{aligned} E(Y_i|X_i = 1) &= \beta_0 + (\beta_1 \times 1) \\ &= \beta_0 + \beta_1. \end{aligned}$$

- Consequently the regression coefficient

$$\beta_1 = E(Y_i|X_i = 1) - E(Y_i|X_i = 0)$$

denotes the difference in mean birth weight (grams) of babies born to smokers vs. those born to non-smokers.

```
proc means;var bwt;class smoke;
```

	N	Mean	Std Dev	
SMK=0:	115	3054.96	752.41	Non-smokers
SMK=1:	74	2772.30	659.81	Smokers

- Based on the specified model...

$$\hat{\beta}_0 = 3054.96 \text{ grams}$$

$$\hat{\beta}_1 = 2772.30 - 3054.96 = -282.66 \text{ grams.}$$

- Saturated models and number of covariate patterns (Vittinghoff et al., 2005 pp. 78-79).

```
proc reg;model bwt = smoke / clb;
```

```
Number of Observations Read      189
Number of Observations Used      189
```

```

      Analysis of Variance
      Sum of      Mean      F
Source  DF  Squares  Square  Value  Pr>F
Model   1  3597444  3597444  6.98  0.0089
Error  187  96317854  515069
Corr Tot 188  99915299
```

```
Root MSE      717.68290  R-Square  0.0360
Dependent Mean 2944.28571
```

```

      Parameter Std      t
Variable  DF  Estimate  Error  Value  Pr>|t|
Intercept 1  3054.957  66.924  45.65  <.0001
smoke     1  -282.659  106.954  -2.64  0.0089
```

```

      95% Confidence Limits
Intercept  2922.93293  3186.98012
smoke     -493.65152  -71.66693
```

```
proc ttest;class smoke;
      var bwt;
run;
```

The TTEST Procedure

```

      Lower      Upper
      CL      CL
      smoke  N  Mean  Mean  Mean
Non-smokers  0  115  2916  3055  3193.9
Smokers      1   74  2619  2772  2925.2

Diff (0-1)      72  283  494
```

T-Tests

```

      t
Variable Method  Variances  DF  Value  Pr>|t|
bwt      Pooled  Equal    187  2.64  0.0089
```

Is it a co-incidence that identical p-values are obtained using linear regression and a two-sample t-test?

**Dummy Variables:
Selecting a Reference Group**

- The reference group should be selected to simplify interpretation.
- Typically non-exposed subjects (e.g. non-smokers) are chosen as the reference group.
- For a binary covariate the choice of reference group determines the...
 - interpretation of the intercept, and the
 - sign of the point estimate and confidence limits, but
 - has no effect on tests of $H_o : \beta_1 = 0$.

```
smoke=1-smoke;
proc reg;model bwt = smoke / clb;
```

```
Number of Observations Read      189
Number of Observations Used      189
```

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	1	3597444	3597444	6.98	0.0089
Error	187	96317854	515069		
Corr. Total	188	99915299			

```
Root MSE 717.68290    R-Square 0.0360
```

Variable	DF	Parameter Estimate	Std Error	t Value	Pr> t
Intercept	1	2772.297	83.429	33.23	<.0001
smoke	1	282.659	106.954	2.64	0.0089

95% Confidence Limits		
Intercept	2607.71443	2936.88016
smoke	71.66693	493.65152

Modeling Nominal Covariates: Race

- Mothers in the birth weight study were classified as white (RACE=1), black (RACE=2) or other (RACE=3).
- Interpretation of the association between race and birth weight is complicated since race effects reflect
 - genetics,
 - culture (e.g. diet, exercise),
 - acculturation,
 - socio-economic status, and
 - societal biases.
- Comstock et al. (2004) provide research recommendations for measuring and modelling effects of race and ethnicity.

Coding Dummy Variables: Three or More Levels

Vittinghoff et al., (2005, Pages 77-79)

- A variable with k levels is modelled using k-1 dummy variables.
- Race is modelled using two dummy variables (i.e. women classified as white, black or other).
- Dummy variables are typically constructed so that regression coefficients compare responses of subjects from the j'th group to that of subjects from a reference category.
- Consider the model...

$$E(Y_i|X_{i1}, X_{i2}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$$

$$\text{where } X_{i1} = \begin{cases} 1 & \text{for white mothers} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{and } X_{i2} = \begin{cases} 1 & \text{for other mothers} \\ 0 & \text{otherwise} \end{cases}$$

- Reference group?

We can summarize the specification of the dummy variables using the table... Interpreting Regression Coefficients

Mother's Race	Dummy Variables	
	X_{i1}	X_{i2}
Black	0	0
White	1	0
Other	0	1

$$E(Y_i|X_{i1}, X_{i2}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$$

$$\text{where } X_{i1} = \begin{cases} 1 & \text{for white mothers} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{and } X_{i2} = \begin{cases} 1 & \text{for other mothers} \\ 0 & \text{otherwise} \end{cases}$$

- For black mothers...

$$E(Y_i|X_{i1} = 0, X_{i2} = 0) = \beta_0$$

- For white mothers...

$$E(Y_i|X_{i1} = 1, X_{i2} = 0) = \beta_0 + \beta_1 \text{ so that}$$

$$\beta_1 = E(Y_i|X_{i1} = 1, X_{i2} = 0) - E(Y_i|X_{i1} = 0, X_{i2} = 0),$$

denotes the difference in mean birth weight (grams) comparing babies born to white vs. black mothers.

- For other mothers...

$$E(Y_i|X_{i1} = 0, X_{i2} = 1) = \beta_0 + \beta_2 \text{ so that}$$

$$\beta_2 = E(Y_i|X_{i1} = 0, X_{i2} = 1) - E(Y_i|X_{i1} = 0, X_{i2} = 0),$$

denotes the difference in mean birth weight (grams) comparing babies born to other vs. black mothers.

```

/*
race=1 for white mothers
race=2 for black mothers
race=3 for other mothers.
*/

*Black moms are the reference group;
race_white=0;race_other=0;
if race=1 then race_white=1;
if race=3 then race_other=1;

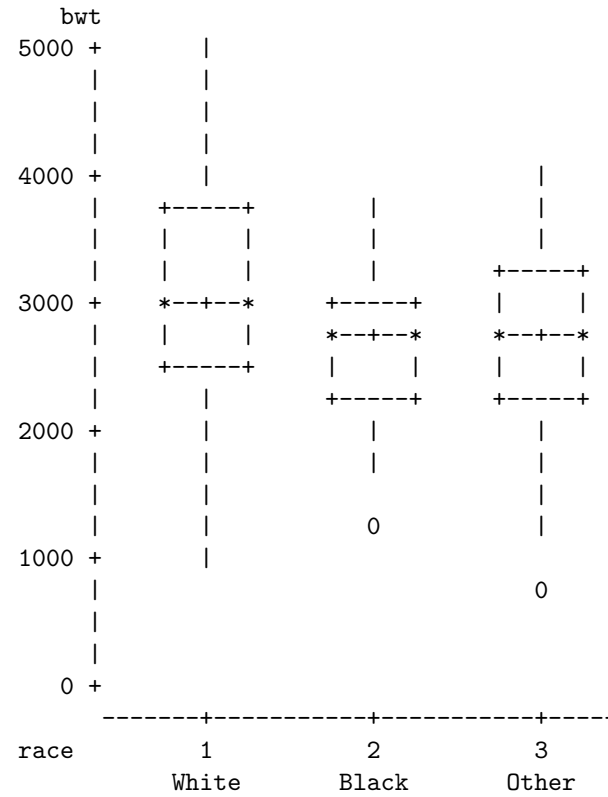
proc means;var bwt;class race;

proc sort;by race;
proc univariate plots;var bwt;by race;

proc reg;
model bwt = race_white race_other/clb;

```

Race	N	Mean	Std Dev
1 - White	96	3103.01	727.87
2 - Black	26	2719.69	638.68
3 - Other	67	2804.01	721.30



Black Moms are the Reference Group

Number of Observations Read 189

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	2	5048361	2524181	4.95	0.0081
Error	186	94866938	510037		
Corr Total	188	99915299			

Root MSE 714.16896 R-Square 0.0505

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr> t
Intercept	1	2719.692	140.061	19.42	<.0001
race_white	1	383.318	157.891	2.43	0.0161
race_other	1	84.323	165.013	0.51	0.6100

Variable	95% Confidence Limits	
Intercept	2443.382	2996.003
race_white	71.830	694.807
race_other	-241.215	409.860

Why 186 degrees of freedom for the error term?

Estimated Regression Coefficients

- Is this a saturated model?

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2}$$

$$\text{where } X_{i1} = \begin{cases} 1 & \text{for white mothers} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{and } X_{i2} = \begin{cases} 1 & \text{for other mothers} \\ 0 & \text{otherwise} \end{cases}$$

- Estimated difference in mean birth weight of babies born to white vs. black mothers

$$\hat{\beta}_1 = 3103.01 - 2719.69 = 383.32 \text{ grams.}$$

- Estimated difference in mean birth weight of babies born to other vs. black mothers

$$\hat{\beta}_2 = 2804.01 - 2719.69 = 84.32 \text{ grams.}$$

Race	N	Mean
1 - White	96	3103.01
2 - Black	26	2719.69
3 - Other	67	2804.01

Statistical Inferences

- Statistical inferences about individual regression coefficients are constructed using t-tests and associated confidence intervals.
- For example, $H_o : \beta_1 = 0$ may be tested using a t-test with $189 - 3 = 186$ degrees of freedom.
- Statistical inferences about the variable RACE must be constructed using F-tests since RACE is being modelled using three categories.
- An F-test with two and 186 degrees of freedom may be used to test the effects of RACE, i.e.

$$H_o : \beta_1 = \beta_2 = 0$$

Vs.

$$H_a : \text{At least one of } \beta_1 \neq 0 \text{ or } \beta_2 \neq 0.$$

- Why two and 186 degrees of freedom?

Choosing a Reference Group

- The reference group should be selected primarily based on study objectives.
- The choice of reference group does not affect F tests of

$$H_o : \beta_1 = \beta_2 = 0$$

Vs.

$$H_a : \text{At least one of } \beta_1 \neq 0 \text{ or } \beta_2 \neq 0.$$

- Interpretation of individual regression coefficients and associated hypothesis tests and confidence intervals are a consequence of the selected reference group.

Selecting White Mothers as the Reference Group

- Consider the model...

$$E(Y_i|X_{i1}, X_{i2}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$$

$$\text{where } X_{i1} = \begin{cases} 1 & \text{for black mothers} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{and } X_{i2} = \begin{cases} 1 & \text{for other mothers} \\ 0 & \text{otherwise} \end{cases}$$

Mother's Race	Dummy Variables	
	X_{i1}	X_{i2}
White	0	0
Black	1	0
Other	0	1

Interpreting Regression Coefficients

- For white mothers...

$$E(Y_i|X_{i1} = 0, X_{i2} = 0) = \beta_0$$

- For black mothers...

$$E(Y_i|X_{i1} = 1, X_{i2} = 0) = \beta_0 + \beta_1 \text{ so that}$$

$$\beta_1 = E(Y_i|X_{i1} = 1, X_{i2} = 0) - E(Y_i|X_{i1} = 0, X_{i2} = 0)$$

denotes the difference in mean birth weight (grams) of babies born to black vs. white mothers.

- For other mothers...

$$E(Y_i|X_{i1} = 0, X_{i2} = 1) = \beta_0 + \beta_2 \text{ so that}$$

$$\beta_2 = E(Y_i|X_{i1} = 0, X_{i2} = 1) - E(Y_i|X_{i1} = 0, X_{i2} = 0)$$

denotes the difference in mean birth weight (grams) of babies born to other vs. white mothers.

Reference Group = White Mothers

```

/*
race=1 for white mothers
race=2 for black mothers
race=3 for other mothers.
*/

race_black=0;race_other=0;
if race=2 then race_black=1;
if race=3 then race_other=1;

proc reg;
model bwt = race_black race_other / clb;
run;

```

$$E(Y_i|X_{i1}, X_{i2}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$$

$$\text{where } X_{i1} = \begin{cases} 1 & \text{for black mothers} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{and } X_{i2} = \begin{cases} 1 & \text{for other mothers} \\ 0 & \text{otherwise} \end{cases}$$

Reference Group = White Mothers

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	2	5048361	2524181	4.95	0.0081
Error	186	94866938	510037		
Corr Total	188	99915299			
Root MSE	714.16896	R-Square	0.0505		
Parameter Estimates					
Variable	DF	Estimate	Std Error	t Value	Pr> t
Intercept	1	3103.010	72.890	42.57	<.0001
race_black	1	-383.318	157.891	-2.43	0.0161
race_other	1	-298.995	113.690	-2.63	0.0093
95% Confidence Limits					
Intercept		2959.21388		3246.80696	
race_black		-694.80637		-71.82985	
race_other		-523.28287		-74.70811	

Estimating Birth Weight (Grams)

- Consider the model...

$$E(Y_i|X_{i1}, X_{i2}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$$

$$\text{where } X_{i1} = \begin{cases} 1 & \text{for black mothers} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{and } X_{i2} = \begin{cases} 1 & \text{for other mothers} \\ 0 & \text{otherwise} \end{cases}$$

- Model predicted birth weight (grams):

$$\text{White Moms: } \hat{Y}_i =$$

$$\text{Black Moms: } \hat{Y}_i =$$

$$\text{Other Moms: } \hat{Y}_i =$$

- Implications for choice of reference group?

Analysis of Variance (ANOVA) and Linear Regression

- One-way ANOVA or linear regression may be used to test H_0 : mean population birth weight is the same for babies born to white, black or other mothers.
- Statistical inferences will be identical.
- Either approach can be extended to account for continuous covariates as confounders or sources of effect measure modification.
- Linear regression focuses attention on regression coefficients and associated confidence intervals rather than on hypothesis tests.

References

1. Vittinghoff E, Glidden DV, Shiboski SC, McCulloch CE. Regression Methods in Biostatistics. Linear, Logistic, Survival, and Repeated Measures Models. New York: Springer 2005.