HUDM 5123 - Linear Models and Experimental Design
Notes 03 - Categorical Predictors

1 Dichotomous Predictors

A *dichotomous* predictor is one that has only two categories. Dichotomous predictors in multiple regression are typically coded with 1s and 0s, which makes them a special case of a dummy coding, which will be defined in more detail below. In any case, when we regress a continuous outcome on a dichotomous, 0/1 predictor, the result is pretty straightforward. As a motivating example, consider the relationship between special education status (SPED) and 5th grade math score (MATH5). In the ECLS data uploaded for class today (called "ecls2.Rdata"), variable SPED is 1 if the student had any involvement in special education between kindergarten and fifth grade, and 0 otherwise. In this case, there are 429 cases that were exposed to special education (i.e., SPED = 1) and 6933 cases that were not exposed to special education. The linear model is specified as follows.

$$MATH5_i = \beta_0 + \beta_1 SPED_i + \epsilon_i,$$

and the prediction model with estimated coefficients is as follows:

$$MATH5_i = 128.2 - 19.2 SPED_i.$$

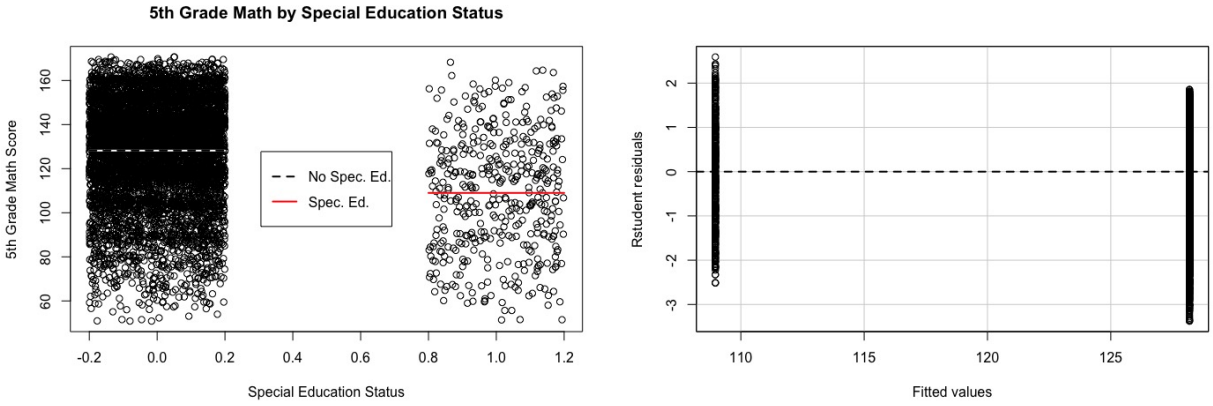

Figure 1: Left pane: plot of 5th grade math score on special education status; right pane: studentized residuals plotted against fitted values

1.1 Checking Regression Assumptions with Categorical Predictors

Recall that all three empirically testable assumptions (linearity, normality, and constant variance) are specified conditional on the predictor values. That is, the residuals must be everywhere (i.e., over all covariate values and combinations) normally distributed, have mean zero, and have identical variance. With only a single dichotomous predictor, there are only

two possible predictor values to condition on: 0 and 1. Thus, we need to check (a) that the mean of the residuals is 0 in both groups (0 and 1), (b) that the residuals (studentized residuals) are normally (students t) distributed, and (c) that the residual variances are not too different across groups.

(a) Linearity is most difficult to check with categorical predictors because there are not enough predictor values to get a sense as to whether the functional form is correct or not. The residuals will average out to (nearly) zero in both groups.

(b) Normality may be checked visually by making histograms of values by group, or by plotting boxplots to check for skew or other attributes such as outliers that would make you question the normality assumption.

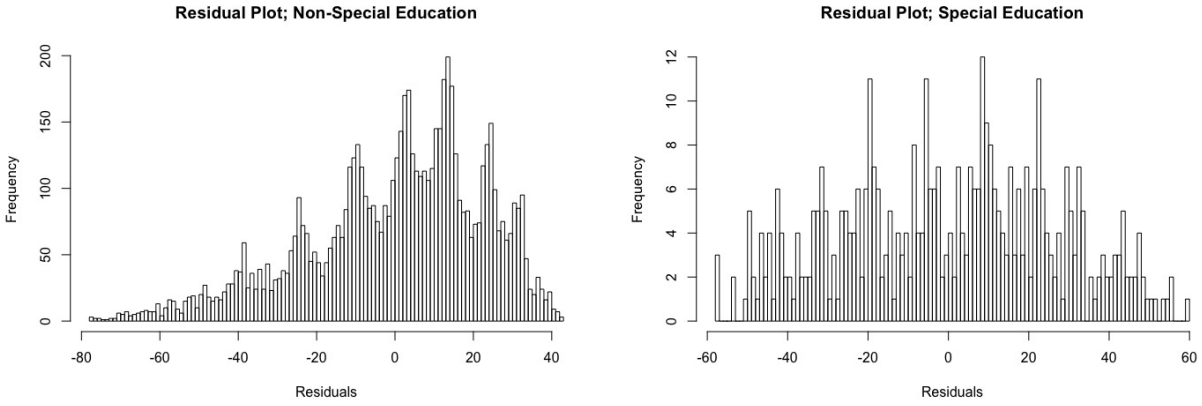

Figure 2: Residual plots for non-special ed (left) and special ed (right) students

(c) Constant variance may be assessed by simply taking effect size estimates of the spread in each group. The sample variance is a useful measure that is often used for this purpose. Here, the sample variances are 513.9 and 718.4 for non-special ed and special ed, respectively. There are plenty of rules of thumb for how for a variance ratio should be before it is of concern. Nevertheless, none are perfect and all are arbitrary. A ratio of variances of larger than 3 is probably too large to safely conclude constant variance is satisfied. Remember that the standard deviation is on the square root scale relative to the variance, so that if the variances differ by a factor of 3, the standard deviations only differ by a factor of $\sqrt{3} \approx 1.7$.

The model estimates that, on average, receipt of special education services (i.e., SPED = 1) is associated with a 19.2 point decrease in 5th grade math score relative to the group of students who did not receive special education services (i.e., SPED = 0). But how should the coefficients be interpreted? In general, it can be instructive to take expected values (i.e., averages). By conditioning on the levels of special education exposure, we arrive at two expressions, one that describes the linear model for students who were not exposed to

special education, and one that describes the model for students who were exposed.

$$\begin{aligned} E[MATH5_i|SPED_i] &= \beta_0 + \beta_1 SPED_i \\ E[MATH5_i|SPED_i = 0] &= \beta_0 \\ E[MATH5_i|SPED_i = 1] &= \beta_0 + \beta 1 \end{aligned}$$

2 Polytomous Factors

Since there is no multi-category variable in the ecls data, one will be created by binning the four quartiles of the socioeconomic status variable, SES. The categorical version is called 'SEScat' in the 'ecls2' data file. 'SEScat' is a categorical variable that takes on four values: 1, 2, 3, and 4, each of which correspond to an SES quartile. It doesn't make sense to simply include the 'SEScat' variable as a predictor in the model, however, because the four categories are not numeric. Instead, they represent group membership in a quartile. Thus, in order to include the predictor, its levels must be coded somehow.

2.1 Dummy-Coding

A categorical factor is said to be *dummy-coded* if each level of the factor is coded as a separate dichotomous (0/1) variable equal to 1 when the factor attains that level and 0 otherwise. For the four-category SES variable, four dummy variables may be constructed as follows.

Table 1: Dummy-coding scheme for a four-category SES variable; category 4 is reference

SEScat	E1	E2	E3
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

In multiple regression, we cannot include all four dummies as predictors in a linear regression because they are perfectly multicollinear. Thus, we must leave out one of the four dummy variables as a *reference group*. Suppose the fourth dummy variable, D4, is held out as a reference group. The model takes on the following form:

$$MATH5_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + \epsilon_i$$

Let $\mu_1, \mu_2, \mu_3, \mu_4$ be the mean 5th grade math scores for groups 1 through 4, respectively. As above, we can learn about how to interpret the coefficients by taking conditional expectations.

Table 2: Dummy-coding example for a four-category SES variable; category 4 is reference

Participant	SEScat	D1	D2	D3	D4
1	3	0	0	1	0
2	3	0	0	1	0
3	2	0	1	0	0
4	1	1	0	0	0
5	4	0	0	0	1
6	1	1	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮
$n-1$	4	0	0	1	1
n	1	1	0	1	0

$$
\begin{aligned}
E[MATH5_i|D_{1i}, D_{2i}, D_{3i}] =& \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} \\
\mu_1 = E[MATH5_i|D_{1i} = 1, D_{2i} = 0, D_{3i} = 0] =& (\beta_0 + \beta_1) \\
\mu_2 = E[MATH5_i|D_{1i} = 0, D_{2i} = 1, D_{3i} = 0] =& (\beta_0 + \beta_2) \\
\mu_3 = E[MATH5_i|D_{1i} = 0, D_{2i} = 0, D_{3i} = 1] =& (\beta_0 + \beta_3) \\
\mu_4 = E[MATH5_i|D_{1i} = 0, D_{2i} = 0, D_{3i} = 0] =& \beta_0
\end{aligned}
$$

Note that β_0 represents the average 5th grade math score for a student in the highest (4th) SES quartile; $\beta_0 + \beta_1$ represents the average 5th grade math score for a student in the lowest (1st) SES quartile; $\beta_0 + \beta_2$ represents the average 5th grade math score for a student in the 2nd SES quartile; and $\beta_0 + \beta_3$ represents the average 5th grade math score for a student in the 3rd SES quartile.

2.2 Deviation Coding

An *effect coding* scheme is similar to a dummy coding scheme except that the reference level is always coded -1. For example, suppose we choose level 4 of the four category factor to be the reference category. Then the effect coding scheme would look as follows.

Table 3: Deviation-coding scheme for a four-category SES variable; category 4 is reference

SEScat	E1	E2	E3
1	1	0	0
2	0	1	0
3	0	0	1
4	-1	-1	-1

The model may be specified as follows.

Table 4: Deviation-coding example for a four-category SES variable; category 4 is reference

Participant	SEScat	E1	E2	E3
1	3	0	0	1
2	3	0	0	1
3	2	0	1	0
4	1	1	0	0
5	4	-1	-1	-1
6	1	1	0	0
⋮	⋮	⋮	⋮	⋮
$n-1$	4	-1	-1	-1
n	1	1	0	1

$$MATH5_i = \beta_0 + \beta_1 E_{1i} + \beta_2 E_{2i} + \beta_3 E_{3i} + \epsilon_i$$

Let $\mu_1, \mu_2, \mu_3, \mu_4$ be the mean 5th grade math scores for groups 1 through 4, respectively. Taking conditional expectations yields the following.

$$\begin{aligned}
E[MATH5_i|E_{1i}, E_{2i}, E_{3i}] =& \beta_0 + \beta_1 E_{1i} + \beta_2 E_{2i} + \beta_3 E_{3i} \\
\mu_1 =& E[MATH5_i|E_{1i} = 1, E_{2i} = 0, E_{3i} = 0] = && \beta_0 + \beta_1 \\
\mu_2 =& E[MATH5_i|E_{1i} = 0, E_{2i} = 1, E_{3i} = 0] = && \beta_0 + \beta_2 \\
\mu_3 =& E[MATH5_i|E_{1i} = 0, E_{2i} = 0, E_{3i} = 1] = && \beta_0 + \beta_3 \\
\mu_4 =& E[MATH5_i|E_{1i} = -1, E_{2i} = -1, E_{3i} = -1] = && \beta_0 - \beta_1 - \beta_2 - \beta_3
\end{aligned}$$

Note that β_0 represents the *grand mean*; that is, the mean of the group means. And each of β_1 through β_3 represent the respective group means minus the grand mean (i.e., each group's *deviation* from the grand mean). This is why this coding scheme is referred to as deviation coding.

3 The Principle of Marginality

The *principle of marginality* asserts that higher-order terms like interactions should only be included in a model if all the associated lower-order terms are also included. Otherwise, the models are needlessly constrained, which leads to parameters that are difficult to interpret. Consider, as an example, the full interaction model for the dichotomous special education predictor we investigated above:

$$MATH5_i = \beta_0 + \beta_1 MATHK_i + \beta_2 SPED_i + \beta_3 MATHK_i \times SPED_i + \epsilon_i$$

Consider two models that violate the principle of marginality:

$$\text{M1: } MATH5_i = \beta_0 + \beta_1 MATHK_i + \beta_3 MATHK_i \times SPED_i + \epsilon_i$$
$$\text{M2: } MATH5_i = \beta_0 + \beta_2 SPED_i + \beta_3 MATHK_i \times SPED_i + \epsilon_i$$

The models may still be fit by least squares, and coefficients estimated, but the interpretation is constrained. For model M1, when SPED $= 0$, the slope is constrained to be 0. For model M2, the intercept is constrained to be the same for both groups. Neither of these constraints make sense for the given situation.

4 ANOVA

So far we have discussed two coding frameworks for expression categorical predictors in multiple regression: dummy coding and deviation coding. As we saw above, a four-category predictor with the last category as reference may be modeled via dummy codes as follows:

$$MATH5_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + \epsilon_i$$

and via deviation codes as follows:

$$MATH5_i = \beta_0 + \beta_1 E_{1i} + \beta_2 E_{2i} + \beta_3 E_{3i} + \epsilon_i$$

You learned in your last regression course that you can use t tests to test the individual regression coefficients in simple and multiple linear regression so that, for example, it is possible to examine regression output and see whether the intercept or any of the particular slope coefficients are significantly different from 0. Consider the output from the multiple regression of 5th grade math score on the dummy coded four-category SES variable.

```
> lm5 <- lm(MATH5 ~ SEScat, data = ecls2)
> summary(lm5)

Call:
lm(formula = MATH5 ~ SEScat, data = ecls2)

Residuals:
    Min      1Q  Median      3Q     Max
-78.877 -12.235   2.847  15.653  55.220
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 139.5758     0.4990  279.69   <2e-16 ***
SEScat1     -25.2759     0.7111  -35.54   <2e-16 ***
SEScat2     -15.9383     0.7047  -22.62   <2e-16 ***
SEScat3      -9.1590     0.7059  -12.97   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.48 on 7358 degrees of freedom
Multiple R-squared:  0.1556,Adjusted R-squared:  0.1553
F-statistic:   452 on 3 and 7358 DF,  p-value: < 2.2e-16
```

Clearly the intercept and each of the slope coefficients have been tested ($H_0 : \beta = 0$), and p-values have been produced via the t-distribution (see Fox, ch. 6-7 for a refresher). However, each of the dummy variables represents only a single category of the larger categorical variable, SEScat. What if we just want to know if the categorical variable itself, SEScat, is a significant predictor of the outcome? So far, we don't have the tools to do that.

Analysis of Variance (ANOVA, for short) is the name given to regression-based tests of categorical predictors in situations like this; there's more to ANOVA that we will discuss in future lectures, but this is the basic idea. ANOVA tests can be built up on incremental tests of nested regression models. These tests, which are called *incremental F tests*, use the F distribution rather than the t distribution.

Two linear regression models are said to be *nested* if one can be made identical to the other by fixing some of its coefficients (βs) to constant values such as 0. Back to the categorical SES example. To devise an overall test of the SEScat variable, we need to simultaneously test all three coefficients. That is, we need a test of the null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ against the alternative that at least one of the parameters differs from 0. To devise this test we fit two models. The first is the *full* model which is the fully complex model that contains the coefficients we wish to test. Here, the full model is simply the dummy variable model

$$\text{Full Model:} \qquad MATH5_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + \epsilon_i$$

We then also need a *reduced* model for comparison. The reduced model must be nested within the full model and the only difference between the two models is that the slope parameters we wish to test have been constrained in the reduced model based on the null hypothesis of interest. Here, then, we must set $\beta_1 = \beta_2 = \beta_3 = 0$ so that the reduced model is

$$\text{Reduced Model:} \qquad MATH5_i = \beta_0 + \epsilon_i$$

A full model will always fit the data at least as well as a model nested within it. That is, $\text{RSS}_R \geq \text{RSS}_F$, where R is for "reduced" and F is for "full." The F-statistic for testing the

omnibus null hypothesis is

$$F_0 = \frac{(RSS_R - RSS_F)/(df_R - df_F)}{RSS_F/df_F}, \text{ or}$$

$$F_0 = \frac{(RegSS_F - RegSS_R)/(df_R - df_F)}{RSS_F/df_F}.$$

These formulations are identical because both models have the same TSS, and TSS = RegSS + RSS. The `anova()` function in R is used to run incremental F tests for nested models. You use the function by running two linear regression models, one full and one reduced (i.e., nested in the full model). Then you pass the models to the anova function as shown below, with the reduced model as the first argument and the full model as the second.

```
> lm1 <- lm(formula = MATH5 ~ SEScat, data = ecls2)
> summary(lm1)

Call:
lm(formula = MATH5 ~ SEScat, data = ecls2)

Residuals:
    Min      1Q  Median      3Q     Max
-78.877 -12.235   2.847  15.653  55.220

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 139.5758     0.4990  279.69   <2e-16 ***
SEScat1     -25.2759     0.7111  -35.54   <2e-16 ***
SEScat2     -15.9383     0.7047  -22.62   <2e-16 ***
SEScat3      -9.1590     0.7059  -12.97   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.48 on 7358 degrees of freedom
Multiple R-squared:  0.1556,Adjusted R-squared:  0.1553
F-statistic:   452 on 3 and 7358 DF,  p-value: < 2.2e-16

> lm2 <- lm(formula = MATH5 ~ 1, data = ecls2)
> summary(lm2)

Call:
lm(formula = MATH5 ~ 1, data = ecls2)

Residuals:
    Min      1Q  Median      3Q     Max
-76.211 -13.346   3.859  16.399  43.589

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 127.0713     0.2723   466.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.37 on 7361 degrees of freedom

> anova(lm2, lm1)
Analysis of Variance Table

Model 1: MATH5 ~ 1
Model 2: MATH5 ~ SEScat
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1   7361 4019063
2   7358 3393701  3    625362 451.96 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We conclude, therefore, that SEScat is a significant predictor of MATH5 ($F(3, 7358) = 451.96$, $p < .001$).

Output based on incremental F tests are often organized in ANOVA tables. ANOVA tables are used to summarize sums of squares, degrees of freedom, mean squares, and F_0 value. Here is the ANOVA table for this with general notation:

Table 5: ANOVA table for the incremental F test (notation)

Source	Sum of Squares	df	Mean Square	F
Regression	RegSS	$df_R - df_F$	$\frac{RegSS}{df_R - df_F}$	$\frac{RegMS}{RMS}$
Residuals	RSS	df_F	$\frac{RSS}{df_F}$	