Day 10 - Categorical Variables and Nested F Tests

SDS 291

February 26, 2020

Birthweight and parental smoking

Using data from the mosaic package. Birth weight, date, and gestational period collected as part of the Child Health and Development Studies from Oakland, CA in 1961 and 1962; we're working with a sample of 1,263 babies and their parents. The study is still ongoing, now following its 3rd generation.

- Response variable
  - `wt`: birth weight (in ounces)
- Explanatory Variable(s)
  - `smoke`: smoke does mother smoke? 0=never, 1=smokes now, 2=until current pregnancy, 3=once did, not now
  - `age`: mother's age in years at termination of pregnancy

Bring in the data

```
require(mosaic)
require(tidyverse)
require(magrittr)
data("Gestation")
Gestation<-Gestation %>% filter(!is.na(smoke), !is.na(wt))
```

Now fit the model:

$$bithweight = \beta_0 + \beta_1 smoke + \epsilon$$

```
m_quant<-lm(wt~smoke, data=Gestation)
summary(m_quant)
```

```
##
## Call:
## lm(formula = wt ~ smoke, data = Gestation)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.855 -10.855   0.274  11.145  56.145
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 119.8553     0.6949  172.48   <2e-16 ***
## smoke        -0.4198     0.5749   -0.73    0.465
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.21 on 1224 degrees of freedom
## Multiple R-squared:  0.0004353,  Adjusted R-squared:  -0.0003814
```

```
## F-statistic: 0.533 on 1 and 1224 DF,  p-value: 0.4655
```

####How do you interpret the coefficient for `smoke`?

Answer: Like the example with gender above, this is treating `smoke` as a quantitative variable. So it can be interpreted as a 1-unit increase in smoking status is associated with a 0.42lb lighter child at birth, on average in the population.

What if smoking status were a categorical variable?

###R uses `factor` variables to indicate categorical variables.

You could make a new variable `smoke_factor` with the same values as `smoke` but where `R` knows the variable should be formatted as a factor.

```
Gestation<-Gestation %>%
  mutate(smoke_factor=as.factor(smoke))
m_cat1<-lm(wt~smoke_factor, data=Gestation)
summary(m_cat1)
```

```
##
## Call:
## lm(formula = wt ~ smoke_factor, data = Gestation)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -67.78 -11.11   0.89  11.22  53.22
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   122.7776     0.7583 161.904  < 2e-16 ***
## smoke_factor1  -8.6681     1.1052  -7.843 9.53e-15 ***
## smoke_factor2   0.3066     1.9668   0.156    0.876
## smoke_factor3   1.6593     1.9006   0.873    0.383
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.69 on 1222 degrees of freedom
## Multiple R-squared:  0.05823,    Adjusted R-squared:  0.05592
## F-statistic: 25.19 on 3 and 1222 DF,  p-value: 8.196e-16
```

What is the average expected birthweight of a child whose mother never smoked?

Answer: The average expected birthweight of a child whose mother never smoked is 122.78 ounces.

Interpret the coefficient for `smoke_factor1`.

Answer: The coefficient for `smoke_factor1` reflects the difference in the average expected birthweight of a child whose mother is a current smoker to a child whose mother never smoked. Specifically, a child born to a mother who is a current smoker will weigh 8.67 ounces lighter at birth than a child born to a mother who never smoked, on average in this population. This -8.67 ounce difference in birthweight between current and never smoker mothers is statisticially significant from 0 since the t-statistic (-7.8) is below the critical value of 1.96, and the p-value (9.53e-15) is <0.05.

Interpret the coefficient for `smoke_factor2`.

Answer: The coefficient for `smoke_factor2` reflects the difference in the average expected birthweight of a child whose mother who smoked until pregnancy to a mother who never smoked. A child of a woman who smoked prior to pregnancy weighed, on average, .3066 ounces more than a child of a never smoker mother in this population. This difference was not statistically signigicantly different from 0, since the t-statistic (0.156) was <1.96 and the p-value (0.876) was >0.05.

It's hard to keep the values of these levels straight. Especially if you had multiple factor variables. Instead, you might make the factor levels be something more conceptually understandable.

```
Gestation<-Gestation %>%
  mutate(smoke_cat=as.factor(if_else(smoke==0,"never smoker",
                             if_else(smoke==1,"current smoker",
                             if_else(smoke==2,"pre-pregnancy smoker",
                             if_else(smoke==3,"other former smoker","NA"))))
                    )
        )
tally(~smoke_cat, data=Gestation)
```

```
## smoke_cat
##       current smoker         never smoker  other former smoker
##                  484                  544                  103
## pre-pregnancy smoker
##                   95
```

```
smoke_factor_labels<-lm(wt~I(smoke_cat), data=Gestation)
summary(smoke_factor_labels)
```

```
##
## Call:
## lm(formula = wt ~ I(smoke_cat), data = Gestation)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -67.78 -11.11   0.89  11.22  53.22
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       114.109      0.804 141.933  < 2e-16 ***
## I(smoke_cat)never smoker            8.668      1.105   7.843 9.53e-15 ***
## I(smoke_cat)other former smoker    10.327      1.919   5.381 8.88e-08 ***
## I(smoke_cat)pre-pregnancy smoker    8.975      1.985   4.522 6.73e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.69 on 1222 degrees of freedom
## Multiple R-squared:  0.05823,    Adjusted R-squared:  0.05592
## F-statistic: 25.19 on 3 and 1222 DF,  p-value: 8.196e-16
```

What is the average expected birthweight of a child whose mother never smoked?

Answer: The average expected birthweight of a child whose mother never smoked was 122.78 ounces (114.109+8.668).

What is the average expected birthweight of a child whose mother was a current smoker?

Answer: The average expected birthweight of a child whose mother was a current smoker was 114.11 ounces.

How to get the right reference group?

You can tell `R` what you want the reference group to be with `relevel()` function.

```
Gestation$smoke_cat<-relevel(Gestation$smoke_cat, ref = "never smoker")
```

But *the better option* is to make dummy/indicator variables for each of the categories.

```
Gestation<-Gestation %>%
  mutate(
    smoke_nev=if_else(smoke==0,1,0),
    smoke_cur=if_else(smoke==1,1,0),
    smoke_pre=if_else(smoke==2,1,0),
    smoke_fmr=if_else(smoke==3,1,0)
  )
tally(c("smoke_nev", "smoke_cur", "smoke_pre", "smoke_fmr"), data=Gestation)
```

```
## X
## smoke_cur smoke_fmr smoke_nev smoke_pre
##         1         1         1         1
```

```
smoke_indicators<-lm(wt~smoke_nev+smoke_cur+smoke_pre+smoke_fmr, data=Gestation)
summary(smoke_indicators)
```

```
##
## Call:
## lm(formula = wt ~ smoke_nev + smoke_cur + smoke_pre + smoke_fmr,
##     data = Gestation)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -67.78 -11.11   0.89  11.22  53.22
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  124.437      1.743  71.401  < 2e-16 ***
## smoke_nev     -1.659      1.901  -0.873    0.383
## smoke_cur    -10.327      1.919  -5.381 8.88e-08 ***
## smoke_pre     -1.353      2.516  -0.538    0.591
## smoke_fmr         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.69 on 1222 degrees of freedom
## Multiple R-squared:  0.05823,    Adjusted R-squared:  0.05592
## F-statistic: 25.19 on 3 and 1222 DF,  p-value: 8.196e-16
```

What does the ANOVA table tell us?

```
anova(smoke_indicators)
```

```
## Analysis of Variance Table
##
## Response: wt
##             Df Sum Sq Mean Sq F value    Pr(>F)
## smoke_nev    1  10385 10385.4  33.197 1.053e-08 ***
## smoke_cur    1  13162 13162.3  42.074 1.274e-10 ***
## smoke_pre    1     90    90.4   0.289    0.5909
## Residuals 1222 382290   312.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What's the MSModel?

The MSModel is 7879.333 ((23399.5+.1+238.4)/3), which has an F-statistic of 25.19 on 3 and 1222 DF. This model explains a statistically signficant more of the variance in birthweight than a constant model with no variables, as evidenced by the small p-value (8.196e-16) which is <0.05.

What about a more parsimonious model?

Webster's Dictionary defines parsimony as "the quality of being careful with money or resources." Essentially, we use parsimonious as an adjective to describe a simpler model; we're being careful with our degrees of freedom and number of coefficients in the model, and would prefer to "spend" fewer of them. A simpler model is generally better – it's easier to explain.

```
Gestation<-Gestation %>%
  mutate(smoke_cur_d=as.factor(if_else(smoke==1,"Smoker","NonSmoker")))
smoke_cur_d<-lm(wt~smoke_cur_d, data=Gestation)
summary(smoke_cur_d)
```

```
##
## Call:
## lm(formula = wt ~ smoke_cur_d, data = Gestation)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -68.05 -11.05   0.89  10.95  52.95
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        123.047      0.649 189.597   <2e-16 ***
## smoke_cur_dSmoker   -8.938      1.033  -8.653   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.68 on 1224 degrees of freedom
## Multiple R-squared:  0.05764,    Adjusted R-squared:  0.05687
## F-statistic: 74.87 on 1 and 1224 DF,  p-value: < 2.2e-16
```

How does this change our interpretation of `smoke_cur_d`?

Answer: The reference group is now not-current-smokers, a combination of never smokers and former smokers (both those pre-pregnancy and who had quit longer before pregnancy). A child of a mother who currently smokes will be, on average, 8.94 ounces lighter than a mother who does not currently smoke in this population.

Do we need the 4 category variable or is the dichotomous/binary variable enough?

```
anova(smoke_cur_d,smoke_indicators)
```

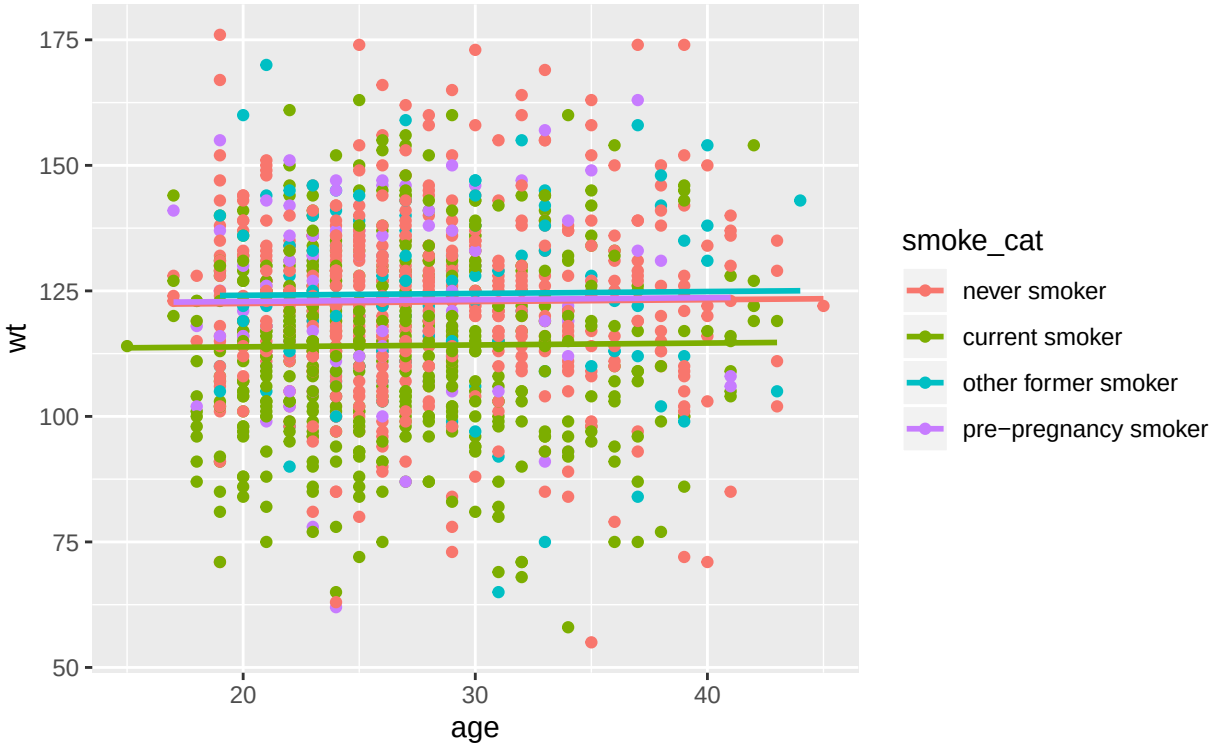
```
## Analysis of Variance Table
##
## Model 1: wt ~ smoke_cur_d
## Model 2: wt ~ smoke_nev + smoke_cur + smoke_pre + smoke_fmr
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1224 382529
## 2   1222 382290  2     238.6 0.3813  0.683
```

Answer: We fail to reject the null hypothesis that the coefficients for smoke_pre and smoke_fmr both equal 0. We can conclude that, since these extra coefficients in the model were together not different from 0, that the nested, parsimonious model of a binary variable (current vs. non-current smokers) is sufficient)

What we want to do is to test whether the model with four categories of smoking is *better* than the a binary definition of smoking (current vs. not current, which includes never, former, and smokers until pregnancy).

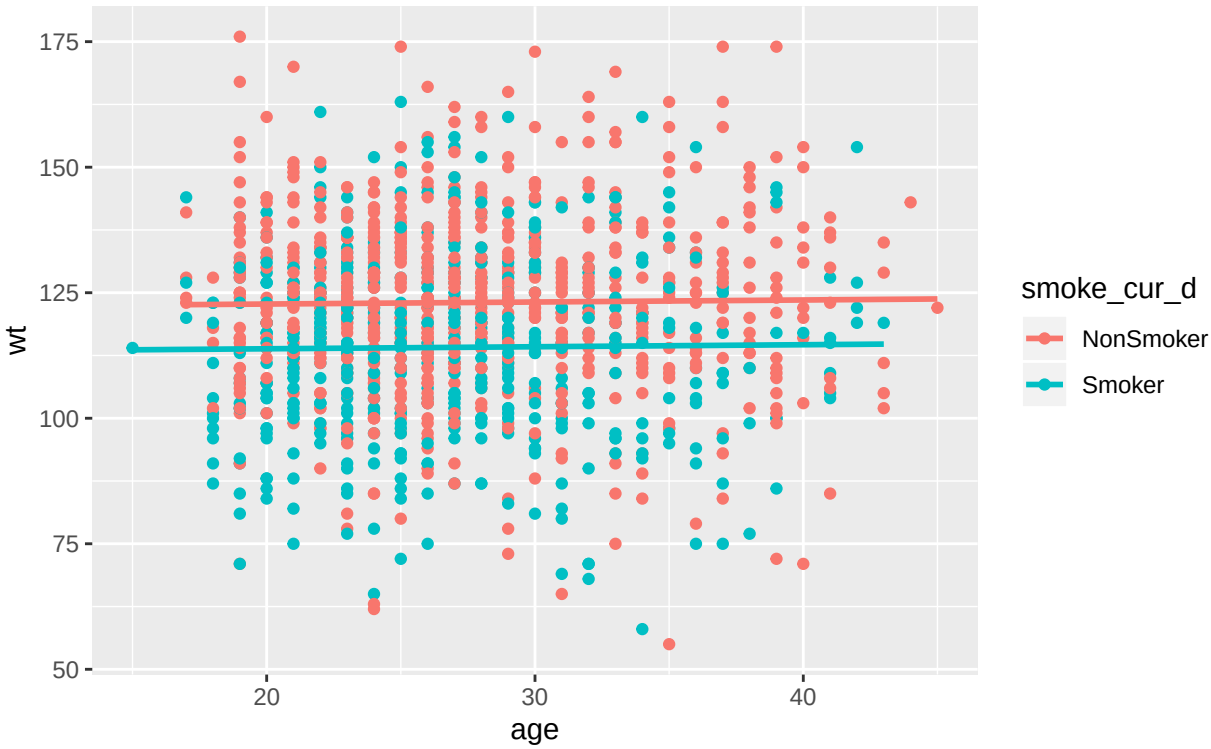
Essentially, is the right plot with a binary explanatory variable better than the left of a model with four categories of `smoke`:

```
library(moderndive)
qplot(y=wt, x=age, color=smoke_cat, data=Gestation)+geom_parallel_slopes(se=FALSE)
```




```
smoke_cur_dage<-lm(wt~smoke_cur_d+age, data=Gestation)
```

```
qplot(y=wt, x=age, color=smoke_cur_d, data=Gestation)+geom_parallel_slopes(se=FALSE)
```




```
wt4cat_temp<-lm(wt~smoke_cat+age, data=Gestation)
```

We are essentially testing the hypothesis that former smokers and quit at pregnancy are the same as never smokers vs. the alternative that one of them is different.

$$H_0 : \beta_{\text{former}} = \beta_{\text{quit-at-pregancy}} = 0$$
$$H_A : \beta_i \neq 0$$

Nested Model

```
anova(smoke_cur_dage)
```

```
## Analysis of Variance Table
##
## Response: wt
##               Df Sum Sq Mean Sq F value Pr(>F)
## smoke_cur_d    1  23393 23392.7 74.6822 <2e-16 ***
## age            1     67    66.6  0.2126 0.6448
## Residuals   1221 382454   313.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Full Model

```
anova(wt4cat_temp)
```

```
## Analysis of Variance Table
##
## Response: wt
##             Df Sum Sq Mean Sq F value    Pr(>F)
## smoke_cat    3  23632  7877.2 25.1221 8.991e-16 ***
## age          1     58    57.9  0.1848    0.6674
## Residuals 1219 382224   313.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nested F Test

```
anova(smoke_cur_dage, wt4cat_temp)
```

```
## Analysis of Variance Table
##
## Model 1: wt ~ smoke_cur_d + age
## Model 2: wt ~ smoke_cat + age
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1   1221 382454
## 2   1219 382224  2    230.14 0.367 0.6929
```

The nested-F test is

$$NestedF = \frac{\frac{SSM_{\text{full}} - SSM_{\text{nested}}}{\text{Number of predictors}}}{\frac{SSE_{\text{Full}}}{n-k-1}}$$

In this case, we have $NestedF = \frac{\frac{23632-23393}{2}}{\frac{382224}{1219}} = \frac{\frac{230.14}{2}}{313.5} = \frac{115.07}{313.5} = 0.367$

We fail to reject the null hypothesis (the F statistic is very small, <1, and the p-value is very large and above our standard threshold $0.69>0.05$) and conclude that we don't have evidence that former and quit-at-pregnancy aren't the same as each other. Thus, a model with just non-smokers vs. smokers does just as well as a model with all four categories (current vs. never, former, quit-at-pregnancy).