Take Away: Dummy Variable

(Jing Li, Miami University)

1. A young couple wants to know how gender and parents' heights affect their kids' heights. The first two observations in Height data are shown below

```
> ad = "https://www.fsb.miamioh.edu/lij14/311r_height.txt"
> data = read.table(url(ad), header=T)
> data[1:2,]
  family father mother midparentHeight children childNum gender childHeight
1    001   78.5     67           75.43        4        1   male        73.2
2    001   78.5     67           75.43        4        2 female        69.2
> attach(data)
> d = as.numeric(gender=="male")
> mean(d)
[1] 0.5149893
```

   Notice that gender is categorical. So we convert it to a dummy variable d that equals one for male (base group is female). About 51.49 percent kids are male.

2. The DVR I reports the two-sample t test

```
> summary(lm(childHeight~d))$coef
              Estimate Std. Error   t value      Pr(>|t|)
(Intercept) 64.103974  0.1173385 546.31651  0.000000e+00
d            5.130122  0.1635089  31.37518 4.681854e-148
```

   The intercept 64.103974 is average height for girls; the slope 5.130122 is the average heights of boys minus average heights of girls. The gap between two genders is statistically significant according to the t value 31.37518

3. With DVR II, we can show how father's height affects kids' height

```
> inte = d*father
> summary(lm(childHeight~d+father+inte))$coef
               Estimate Std. Error    t value     Pr(>|t|)
(Intercept) 37.69497433 2.81095074 13.4100444 1.290768e-37
d            0.66760670 4.20331655  0.1588286 8.738384e-01
```

```
father        0.38130087 0.04055579  9.4018844 4.045563e-20
inte          0.06522174 0.06070986  1.0743186 2.829585e-01
```

4. More interpretable results are based on group-wise regressions. For instance, the intercept and coefficient of father in previous DVR II can be duplicated in the following regression that uses girls only

```
> summary(lm(childHeight[d==0]~father[d==0]))$coef
                 Estimate Std. Error  t value     Pr(>|t|)
(Intercept)    37.6949743 2.62458412 14.36227 8.691097e-39
father[d == 0]  0.3813009 0.03786693 10.06949 1.197041e-21
```

5. With DVR II, we can test the hypothesis that the relationship between kids' height and father's height does not depend on gender

```
> dvr2 = lm(childHeight~d+father+inte)
> library("car")
> linearHypothesis(dvr2, c("d=0", "inte=0"))
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
2    930 4843.7  2      6263 601.25 < 2.2e-16 ***
```

The Chow or F test is 601.25. We reject the hypothesis since p value is less than 0.05

6. This data illustrate an important phenomenon called Regression towards Mean (RTM). Consider father-and-son. The slope coefficient shown below, 0.4465226, is less than one, implying that, on average, if a father is one inch above (or below) the average, the son is only 0.4465226 inch above (or below) average. In other words, sons tend to be closer to average (mean) relative to their dads. Similar results are found for mother-and-daughter as well.

```
> summary(lm(childHeight[d==1]~father[d==1]))$coef
                 Estimate Std. Error   t value     Pr(>|t|)
(Intercept)    38.3625810 3.30837361 11.595601 1.425209e-27
father[d == 1]  0.4465226 0.04782546  9.336504 3.737109e-19
```

Read `http://www.fsb.miamioh.edu/lij14/311r_or_rtm.pdf` for more discussion about RTM. If your parents are very successful, do not feel bad if you are not that successful! On the other hand, you have a better chance to become top dog if your parents are underdogs.