Eco311 Optional Reading: NBA Data and Categorical Variable

(Jing Li, Miami University)

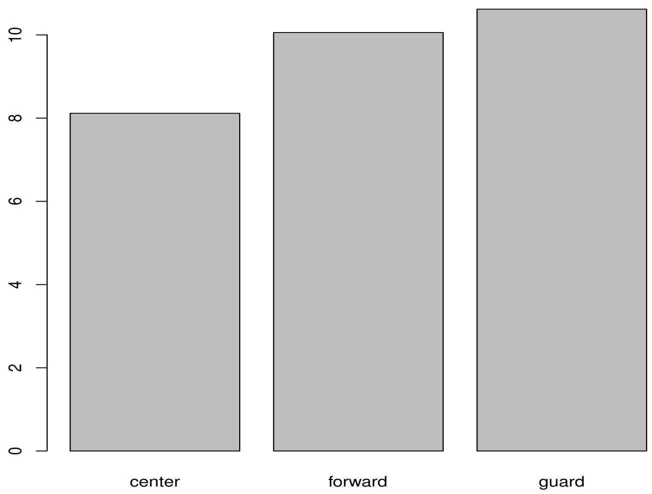
1. We use NBA data to illustrate how to analyze a categorical (nominal, factor) variable that takes more than two values (levels). We are interested in how position of a NBA player affects points. We need to be careful because position is string and categorical. Unlike a dummy variable, position can take three values (levels)—center, forward, guard. In terms of statistics position follows multinomial distribution

```
> library("readxl")
> setwd("I:/311")
> data = read_excel("311_nba.xls")
> attach(data)
> position[1:5]
[1] "guard"   "guard"   "center"  "guard"   "forward"
> prop.table(xtabs(~position))
   center   forward     guard
0.1740614 0.3959044 0.4300341
```

   About 17 percent of NBA players play center position.

2. The bar graph below compares average points across three positions

```
> barplot(tapply(points,position,mean,na.rm=T))
> mean(points[position=="center"],na.rm=T)
[1] 8.115686
> mean(points[position=="forward"],na.rm=T)
[1] 10.05652
> mean(points[position=="guard"],na.rm=T)
[1] 10.61818
```



3. The statistical significance of the difference or gap between bars cannot be seen from the graph. To obtain the significance based on a regression, we need to generate a set of dummy variables, one dummy for each position:

   ```
   > center = as.integer(position=="center")
   > forward = as.integer(position=="forward")
   > guard = as.integer(position=="guard")
   ```

   We use **as.integer** function to convert Boolean values TRUE, FALSE to 1, 0.

4. The fact that the sum of guard, center and forward is one, a constant, implies that we cannot include in the regression all three dummy variables along with the constant term. Otherwise we would run into dummy variable trap, a situation in which perfect multicollinearity arises.

5. Intuitively, because there are only three positions, we know a person must be center if the player is not forward or guard. In other words, the center dummy is redundant once forward and guard dummies are included in the regression. Another example is, once female dummy variable is included in the regression, there is no need to include male dummy variable.

6. So, to avoid dummy variable trap, we try using only two dummy variables along with the constant term

```
> summary(lm(points~forward+guard))$coef
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 8.115686  0.8142268 9.967353 2.916266e-20
forward     1.940836  0.9782515 1.983984 4.822014e-02
guard       2.502496  0.9707714 2.577842 1.044614e-02
```

$\hat{\beta}_0 = 8.115686$ is the average points for center (the base group, for which both forward and guard equal zero, or for which we drop the corresponding dummy variable). $\hat{\beta}_1 = 1.940836$ is the difference of average points between forward and center; $\hat{\beta}_2 = 2.502496$ is the difference of average points between guard and center. In short, all comparison is made relative to the base group, and in this case, the base group is center.

7. We can test the null hypothesis that position does not matter for points, i.e., no difference between forward and center, and no difference between guard and center:

```
> m = lm(points~forward+guard)
> library("car")
> linearHypothesis(m, c("forward=0", "guard=0"))
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    286 9829.3
2    284 9602.4  2    226.93 3.3558 0.03627 *
```

The p-value 0.03627 of F test is less than 0.05, so we reject the hypothesis of no difference across positions. In other words, position matters for points.

8. In fields like biology, people would say position is treatment, and F test is commonly called analysis of variance (ANOVA). Simply put, we can carry out ANOVA by regressing a variable onto a set of dummy variables and conduct the F test that all coefficients of dummy variables equal zero.

9. We can include all three dummy variables in regression, but then we have to drop the constant (intercept) term

```
> summary(lm(points~center+forward+guard-1))$coef
         Estimate Std. Error   t value     Pr(>|t|)
center   8.115686  0.8142268  9.967353 2.916266e-20
forward 10.056522  0.5422276 18.546682 7.367468e-51
guard   10.618182  0.5286130 20.086873 1.860402e-56
```

The advantage of this regression without intercept is that we can get average points for each position directly. Nevertheless, the disadvantage is that it is hard to test the difference across position

10. In practice we can skip generating the set of dummy variables. Instead, R function **factor** can be used as a shortcut

```
> summary(lm(points~factor(position)))$coef
                        Estimate Std. Error  t value     Pr(>|t|)
(Intercept)             8.115686  0.8142268 9.967353 2.916266e-20
factor(position)forward 1.940836  0.9782515 1.983984 4.822014e-02
factor(position)guard   2.502496  0.9707714 2.577842 1.044614e-02
```

The result is the same as the regression that uses forward and guard dummy variables as regressors.