Dummy Variables in Multiple Linear Regression

Basic Idea

We start by considering a simple linear regression with a single variable that is dichotomous rather than continuous. Recall that in simple linear regression, we had the basic equation:

$$Y = \alpha + \beta \times X$$

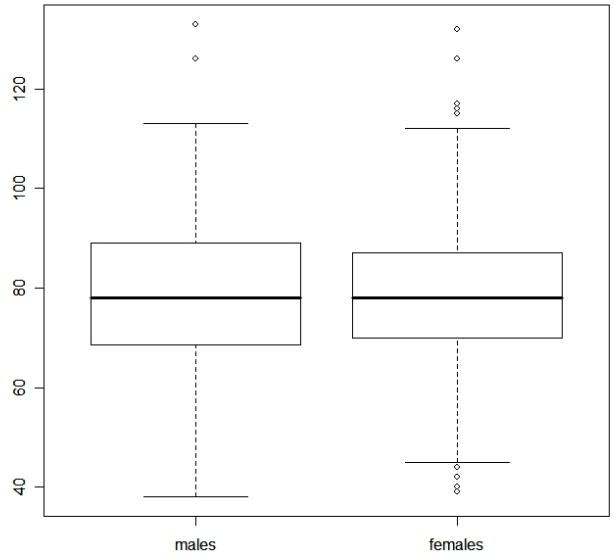
Suppose that instead of a continuous X variable such as age, we have a dichotomous variable, such as gender, or a multiple categorical variable, such as country of birth, or type of employment (e.g., student, office worker, or miner, etc.). How does the nature of the regression change when such variables are used?

We actually saw an example of this at the end of last lecture, when we used gender as a variable in our multiple regression predicting *mphr*, the percentage of maximum predicted heart rate achieved. Let's revisit that example, using a simple linear regression of these two variables.

First, let's see some descriptive statistics about these two variables. Note that we coded the gender variable as males=0 and females=1. This is called a "dummy variable', where the 0's and 1's represent males and females, respectively. It is quite common to code variables in this way, because internally, statistical programs use numbers in all of their calculations, not words. So, even if you did not code your dichotomous or multi-categorical variables using numbers, they will be converted by the program into numbers.

```
#  Input data set as a data frame in R

> heart <- read.table(file="g:\\HeartData.txt", header=T)

> attach(heart)  #  For convenience, to allow direct access to variables
                 #   outside of the data frame.

> summary(gender)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  1.0000  0.6057  1.0000  1.0000
```

```
> summary(mphr)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  38.00   69.00   78.00   78.57   88.00  133.00

> summary(mphr[gender==0])   #  Look at males only
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  38.00   68.75   78.00   78.70   89.00  133.00

> summary(mphr[gender==1])   #  Look at females only
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  39.00   70.00   78.00   78.49   87.00  132.00

> boxplot(list(males=mphr[gender==0],females=mphr[gender==1]))
```



There does not seem to be much difference between males and females, but, in any case, we will next run a simple linear regression of these two variables:

```
> regression.out <- lm(mphr ~ gender)

> multiple.regression.with.ci(regression.out)
$regression.table

Call: lm(formula = mphr ~ gender)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-40.6955  -9.4882  -0.4882   9.5118  54.3045

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  78.6955     1.0204  77.125   <2e-16 ***
gender       -0.2073     1.3110  -0.158    0.874
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.13 on 556 degrees of freedom Multiple
R-Squared: 4.496e-05,  Adjusted R-squared: -0.001754 F-statistic:
0.025 on 1 and 556 DF,  p-value: 0.8744

$intercept.ci [1] 76.69123 80.69968

$slopes.ci [1] -2.782458  2.367880
```

As expected, we see no evidence of a difference in *mphr* between genders. Looking at the confidence interval, we see that the difference is at most 2 or 3 points in either direction, a clinically negligible difference. Therefore, we can strongly conclude no gender difference exists in *mphr*, at least from these data.

Because the data were coded as 0/1, the coefficient is directly interpreted as the female (1) – male (0) difference. In other words, females on average were -0.2073 points lower than males, a negligible difference.

Coding as Factor Variables

One can have R automatically do dummy variable coding, using a "factor" declaration, as follows:

```
#  Create a new gender variable that will be coded as
#  male/female, rather than 0/1

> gender2 <- gender   #  Make a copy of the variable gender
> gender2[gender2==1] <- "female"   #  Change 1's to female
> gender2[gender2==0] <- "male"     #  Change 0's to male

#  Check a few values

> gender2[1:10]
```

```
 [1] "male"   "male"   "male"   "female" "male"   "male"   "female" "female"
 [9] "female" "male"

> gender[1:10]
 [1] 0 0 0 1 0 0 1 1 1 0

#  Everything seems to work, so convert gender2 to be a factor variable:

> gender2 <- as.factor(gender2)

#  Do some comparisons of the different types of variables:

> summary(gender2)
female   male
   338    220

> summary(gender)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  1.0000  0.6057  1.0000  1.0000

#  Finally, rerun regression with this new factor variable:

> regression.out <- lm(mphr ~ gender2)
> multiple.regression.with.ci(regression.out)
$regression.table

Call:
lm(formula = mphr ~ gender2)

Residuals:
     Min       1Q   Median       3Q      Max
-40.6955  -9.4882  -0.4882   9.5118  54.3045

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  78.4882     0.8232  95.345   <2e-16 ***
gender2male   0.2073     1.3110   0.158    0.874
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.13 on 556 degrees of freedom
Multiple R-Squared: 4.496e-05,  Adjusted R-squared: -0.001754
F-statistic: 0.025 on 1 and 556 DF,  p-value: 0.8744
```

```
$intercept.ci
[1] 76.87120 80.10513

$slopes.ci
[1] -2.367880  2.782458
```

Results are the same as before, but note that R has recognized that *gender*2 is a factor variable, and has in fact chosen males as the variable to report, indicated by the name "gender2male" it used, meaning that the coefficient and CI are reversed but numerically the same as the previous regression run with the original "gender" variable.

Another Example With Multi-Level Factor Variables

We will now go through another multiple regression example, this time with several factor variables, some with more than two levels.

Background: Government statisticians in England conducted a study of the relationship between smoking and lung cancer. The data concern 25 occupational groups (here categorized into three main categories, for convenience, not necessarily scientific reasons!) and are condensed from data on thousands of individual men. The explanatory variables are work category and the number of cigarettes smoked per day by men in each occupation relative to the number smoked by all men of the same age. This smoking ratio is 100 if men in an occupation are exactly average in their smoking, it is below 100 if they smoke less than average, and above 100 if they smoke more than average. The response variable is the standardized mortality ratio for deaths from lung cancer. It is also measured relative to the entire population of men of the same ages as those studied, and is greater or less than 100 when there are more or fewer deaths from lung cancer than would be expected based on the experience of all English men.

Variable Names:

Occupation:	occupation category of the subjects (Outdoor/Factory/Office)
Smoke.Index:	relative ranking of amount of smoking, with 100 being average
Cancer.Index:	relative ranking of deaths due to lung cancer, with 100 being average

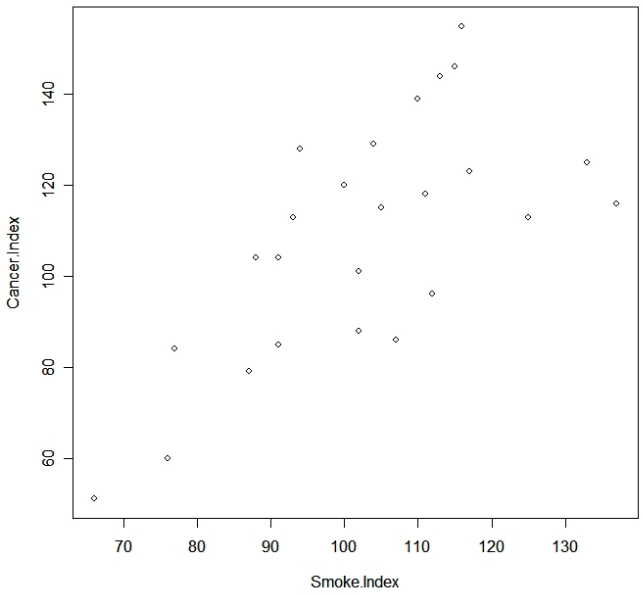
The actual data set is:

Smoke.Index	Cancer.Index	Occupation
77	84	Outdoor
137	116	Outdoor
117	123	Factory
94	128	Factory
116	155	Factory
102	101	Factory
111	118	Office
93	113	Outdoor
88	104	Factory
102	88	Factory
91	104	Factory
104	129	Factory
107	86	Factory
112	96	Factory
113	144	Factory
110	139	Office
125	113	Outdoor
133	125	Outdoor
115	146	Outdoor
105	115	Office
87	79	Office
91	85	Office
100	120	Outdoor
76	60	Office
66	51	Office

We will perform a regression analysis to determine the effects of Occupation category and Smoke.Index on the outcome, Cancer.Index.

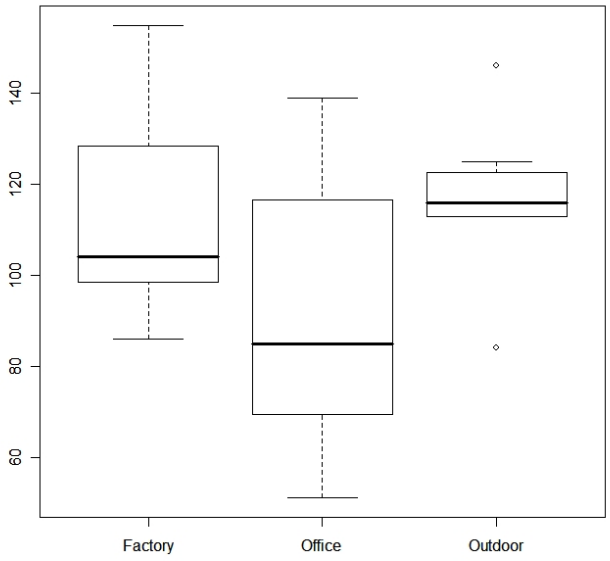
```
#  Start by entering the data set:

> Smoke.Index <- c( 77 , 137 , 117 , 94 , 116 , 102 , 111 , 93 , 88 , 102,
 91 , 104 , 107 , 112 , 113 , 110 , 125 , 133 , 115 , 105 , 87 , 91 ,
100 , 76 , 66 )

> Cancer.Index <- c( 84 , 116 , 123 , 128 , 155 , 101 , 118 , 113 , 104 ,
88 , 104 , 129 , 86 , 96 , 144 , 139 , 113 , 125 , 146 , 115 , 79 , 85 ,
120 , 60 , 51 )

> Occupation <- c("Outdoor" , "Outdoor" , "Factory" , "Factory" ,
"Factory" , "Factory" , "Office" , "Outdoor" , "Factory" , "Factory" ,
"Factory" , "Factory" , "Factory" , "Factory" , "Factory" , "Office" ,
"Outdoor" , "Outdoor" , "Outdoor" , "Office" , "Office" , "Office" ,
"Outdoor" , "Office" , "Office" )
```

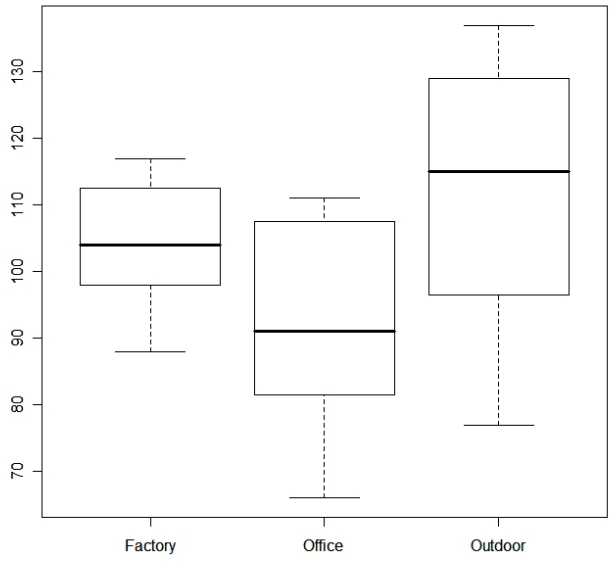
```
#  Convert Occupation to a factor variable

> Occupation <- as.factor(Occupation)

# Create a data frame with all variables

> cancer <- data.frame(Smoke.Index, Cancer.Index, Occupation)

#  Quick summary of the data

> summary(cancer)
  Smoke.Index     Cancer.Index     Occupation
 Min.   : 66.0   Min.   : 51.0   Factory:11
 1st Qu.: 91.0   1st Qu.: 88.0   Office : 7
 Median :104.0   Median :113.0   Outdoor: 7
 Mean   :102.9   Mean   :108.9
 3rd Qu.:113.0   3rd Qu.:125.0
 Max.   :137.0   Max.   :155.0

#  Do some preliminary plots to investigate relationships

> plot(Smoke.Index, Cancer.Index)
```




```
#  Create some boxplots for the factor variable
```

```
> boxplot(list(Factory=Cancer.Index[Occupation=="Factory"],
   Office=Cancer.Index[Occupation=="Office"],
   Outdoor=Cancer.Index[Occupation=="Outdoor"]))

> boxplot(list(Factory=Smoke.Index[Occupation=="Factory"],
   Office=Smoke.Index[Occupation=="Office"],
   Outdoor=Smoke.Index[Occupation=="Outdoor"]))
```





Notice that there seems to be a very strong correlation between Smoke.Index and Cancer.Index, while the boxplots indicate possible effects for Occupation as well, on *both* variables. The possible relationship between Smoke.Index and Occupation has implications for possible confounding (we will discuss confounding in detail later).

Now to perform the regression calculations:

```
#  First, two simple linear regressions (for later comparison)

> regression.out <- lm(Cancer.Index ~ Smoke.Index)
> multiple.regression.with.ci(regression.out)
$regression.table

Call:
lm(formula = Cancer.Index ~ Smoke.Index)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-27.9950 -18.6450   0.7632  14.2881  32.6174

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0000    24.1231   0.124 0.902110
Smoke.Index   1.0292     0.2314   4.448 0.000184 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.5 on 23 degrees of freedom
Multiple R-Squared: 0.4624,     Adjusted R-squared: 0.439
F-statistic: 19.78 on 1 and 23 DF,  p-value: 0.0001845

$intercept.ci
[1] -46.90241  52.90234

$slopes.ci
[1] 0.5504845 1.5078365

> regression.out <- lm(Cancer.Index ~ Occupation)
> multiple.regression.with.ci(regression.out)
$regression.table

Call:
lm(formula = Cancer.Index ~ Occupation)

Residuals:
    Min      1Q  Median      3Q     Max
-41.429 -13.429  -3.714  14.636  46.571

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        114.364      7.498  15.252 3.51e-13 ***
OccupationOffice   -21.935     12.024  -1.824   0.0817 .
OccupationOutdoor    2.351     12.024   0.196   0.8468
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.87 on 22 degrees of freedom
Multiple R-Squared: 0.1633,     Adjusted R-squared: 0.08721
F-statistic: 2.146 on 2 and 22 DF,  p-value: 0.1407
```

```
$intercept.ci
[1]  98.81348 129.91379

$slopes.ci
          [,1]      [,2]
[1,] -46.87079  3.000659
[2,] -22.58507 27.286373

> regression.out <- lm(Cancer.Index ~ Smoke.Index + Occupation)

> multiple.regression.with.ci(regression.out)
$regression.table

Call:
lm(formula = Cancer.Index ~ Smoke.Index + Occupation)

Residuals:
     Min       1Q   Median       3Q      Max
-31.0627 -16.2543   0.2586  14.2311  29.6061

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        14.5868    27.8911   0.523  0.60645
Smoke.Index         0.9577     0.2615   3.663  0.00145 **
OccupationOffice  -10.5420    10.1037  -1.043  0.30864
OccupationOutdoor  -4.5897     9.7980  -0.468  0.64430
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.88 on 21 degrees of freedom
Multiple R-Squared: 0.4895,     Adjusted R-squared: 0.4165
F-statistic: 6.711 on 3 and 21 DF,  p-value: 0.002375


$intercept.ci
[1] -43.41585  72.58948

$slopes.ci
            [,1]      [,2]
[1,]   0.4139858  1.501451
[2,] -31.5537844 10.469884
[3,] -24.9657680 15.786372
```

What can be concluded from these analyses (mostly from looking at the confidence

intervals)?

1. First, examine the "full model", the one with both independent variables included. The confidence interval for Smoke.Index goes from about 0.4 to 1.5. This means (roughly speaking, remember the Bayesian interpretation of frequentist confidence intervals) that there is a 95% chance that the true effect of Smoke.Index on Cancer.Index is between these numbers *for each unit increase* in Smoke.Index. In other words, for every 1% rise, cancer also increases by about 1% (point estimate was 0.9577), and even the lower limit of 0.4 is quite high. Therefore the effect is very strong.

2. On the other hand, there is much less evidence for an effect from Occupation, at least as crudely categorized here. Both confidence intervals are very wide, both potentially clinically interesting and the null (0) effect. All we can conclude here is that these results are inconclusive. More research will need to be done to further investigate this relationship.

3. Note that the estimated values for OccupationOffice and OccupationOutdoor are both negative (in the multiple regression model), and that they are both interpreted in relation to the Factory workers. Therefore, the Factory workers had the highest cancer levels, once we have adjusted for Smoke.Index.

4. In the univariate analysis, and also seen in the boxplots, it appears that Factory had lower CancerIndex compared to Outdoor, which were highest. However, in the second boxplot, we saw that Outdoor workers also smoked the most. Thus, once adjusting for the effects of smoking, the effect of Outdoor occupation is lower. This is an example of confounding, which we will cover in more detail later.

5. While for continuous variables like Smoke.Index, coefficients (and CI limits) are interpreted as effects per unit change, for factor variables, they are interpreted as the difference between having that value of the factor versus the "reference category", that is, the one left out in the regression table.

6. Finally, let's compare the univariate results to the multivariate. For Smoke.Index, the CI from the univariate results were (0.55, 1.51), similar to but slightly higher than the multivariate result. This is owing to the confounding with the second independent variable, Occupation. Similarly, as already discussed, Occupation estimates change because of confounding with Smoke.Index.

In the next few lectures, we will investigate various issues in multiple regression in detail, including confounding, interactions, goodness of fit of a model, and model selection.