MLR With Qualitative Predictors

An input variable with c classes will be respresented via $c-1$ binary input variables (this is effectively one-hot encoding). When all of said inputs are 0, this represents the *baseline* case, which we can choose arbitrarily. Then, each class is represented by making one of the variables 1 and all the others 0 (one is "hot").

With this scheme, we effectively end up with c regression models. The pseudo-variables going to 0 causes certain terms to drop out and the pseudo-variable going to 1 causes a term to be added to the intercept.

If we have 2 inputs, $X1, X2$ where $X2$ is a pseudo-variable plot indicating one of two classes, we can plot Y against $X1$ and leave the categorical variables off of the x-axis and instead just distinguishing between classes via a color, we can observe how a change in class affects the linear relationship.

```
df <- read.csv("../datasets/insurance.csv")
model_sum <- lm(month ~ size + factor(type), df)
summary(model_sum)
```

```
##
## Call:
## lm(formula = month ~ size + factor(type), data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6915 -1.7036 -0.4385  1.9210  6.3406
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   33.874069   1.813858  18.675 9.15e-13 ***
## size          -0.101742   0.008891 -11.443 2.07e-09 ***
## factor(type)1  8.055469   1.459106   5.521 3.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.221 on 17 degrees of freedom
## Multiple R-squared:  0.8951, Adjusted R-squared:  0.8827
## F-statistic:  72.5 on 2 and 17 DF,  p-value: 4.765e-09
```

This model has an absence of any interaction effect between X_1 and X_2 because we have not modeled it in. The change in mean adoption time between the two classes is called the main effect (β_2). We can add in an X_1X_2 to the model and observe its change:

```
df <- read.csv("../datasets/insurance.csv")
model_prod <- lm(month ~ size + factor(type) + size * factor(type), df)
summary(model_prod)
```

```
##
## Call:
## lm(formula = month ~ size + factor(type) + size * factor(type),
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7144 -1.7064 -0.4557  1.9311  6.3259
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        33.8383695  2.4406498  13.864 2.47e-10 ***
```

```
## size               -0.1015306  0.0130525  -7.779 7.97e-07 ***
## factor(type)1       8.1312501  3.6540517   2.225   0.0408 *
## size:factor(type)1 -0.0004171  0.0183312  -0.023   0.9821
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.32 on 16 degrees of freedom
## Multiple R-squared:  0.8951, Adjusted R-squared:  0.8754
## F-statistic: 45.49 on 3 and 16 DF,  p-value: 4.675e-08
```

The model is similar and the estimate of the interaction effect is very close to zero. We can test the hypothesis that it is zero using the results of the table, dividing the estimate by its stderr to get t^* and comparing to $t(1-\alpha/2, n-p)$. Equivalently, we can also perform a GLT where the reduced model is the original and the full is the new one:

```r
anova(model_sum, model_prod)
```

```
## Analysis of Variance Table
##
## Model 1: month ~ size + factor(type)
## Model 2: month ~ size + factor(type) + size * factor(type)
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     17 176.39
## 2     16 176.38  1 0.0057084 5e-04 0.9821
```

So the term is insignificant.

Here, $\beta_0 + \beta_1 X_1$ describes the linear model on the baseline category. β_1 describes the linear impact of X_1 on Y, β_2 describes the main effect of the difference in categories (associated with X_2, not X_1), and β_3 describes the interaction effect between X_1 and X_2, which is associated with X_1.

So the model for the other category is $(\beta_0 + \beta_2) + (\beta_1 + \beta_3)X_1$.

We might also call β_3 β_{12} and will do this when there are multiple interaction terms to keep track of.

Main effects describe a difference in *intercept* between categories, while interaction effects describe a difference in *linear impact* between categories.