Qualitative Predictors & Interaction Model

Links: MATH 4322

(Data Science & Machine Learning lecture 7; textbook section 3.3)

Categorical (Qualitative) Predictors

We can may have a quantitative response variable and use a regression model but there is a possibility to have categorical (or qualitative) predictors.

As an example: Using `mtcars` dataset, say we wanted to investigate the difference in the `mpg` based on the transmission `am`. The transmission, `am`, variable is categorical having two categories: *automatic* (represented by 0) and *manual* (represented by 1).

Since our response, `mpg`, is quantitative this is a regression problem, we also know that our predictor `am` is qualitative with the following info:

$$x = \text{am} = \begin{cases} 0 & \text{automatic} \\ 1 & \text{manual} \end{cases}$$

Based on this, we know that our regression model is:
$y = \beta_0 + \beta_1 \times \text{am} + \epsilon$.

Dummy Variables

Because we have these two categories (0's and 1's) this gives us a *dummy* or *indicator* variable.

$$x_1 = \begin{cases} 1 & \text{if } i\text{th car has a manual transmission} \\ 0 & \text{if } i\text{th car has an automatic transmission} \end{cases}$$

This results in the following model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th car has manual} \\ \beta_0 + \epsilon_i & \text{if } i\text{th car has automatic} \end{cases}$$

It can be interpreted that β_0 means the average mpg among automobiles with automatic transmission, that $\beta_0 + \beta_1$ is the average mpg among automobiles with manual transmission. β_1 can be interpreted as the average difference in mpg between automobiles with automatic and manual transmission.

When doing `summary(lm(mpg~am, data = mtcars))` in R it will work as normal, in this case we must look at both our estimates in the *intercept* section and the `am` section.

```
> #Use mtcars
> summary(lm(mpg~am,data = mtcars))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   17.147      1.125  15.247 1.13e-15 ***
am             7.245      1.764   4.106 0.000285 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.902 on 30 degrees of freedom
Multiple R-squared:  0.3598,Adjusted R-squared:  0.3385
F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

The average `mpg` for an automobile with automatic transmission is 17.147, if it has a manual transmission you *add 7.245* to the mpg on average. Thus the average *mpg* for an automobile with manual

transmission is $17.147 + 7.245 = 24.392$.

$$\hat{mpg} = \begin{cases} 17.147 & \text{if automatic} \\ 243.92 & \text{if manual} \end{cases}$$

(Note the categorical variables could have been coded differently, for instance they could have been 1 or -1)

More than Two Levels

Suppose we want to use the number of cylinders as predictors. These variables are also categorical. The model for that would be as follows

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & i\text{th car has 6 cylinders} \\ \beta_0 + \beta_2 + \epsilon_i & i\text{th car has 8 cylinders} \\ \beta_0 + \epsilon_i & i\text{th car has 4 cylinders} \end{cases}$$

We encode it by using two *x* variables, one for if the car has 6 cylinders, one for if the car has 8 cylinders and if both are 0 it would mean there are 4 cylinders.

$$x_{i1} = \begin{cases} 1 & i\text{th car has 6 cylinders} \\ 0 & i\text{th car does not have 6 cylinders} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & i\text{th car has 8 cylinders} \\ 0 & i\text{th car dpes not have 8 cylinders} \end{cases}$$

You have to tell R to use the cylinder variable as factors, by using `as.factor` on the variable.

```
> cyl.fact = as.factor(mtcars$cyl)
> summary(lm(mtcars$mpg~cyl.fact))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.6636     0.9718  27.437  < 2e-16 ***
cyl.fact6    -6.9208     1.5583  -4.441 0.000119 ***
cyl.fact8   -11.5636     1.2986  -8.905 8.57e-10 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.223 on 29 degrees of freedom
Multiple R-squared:  0.7325,Adjusted R-squared:  0.714
F-statistic:  39.7 on 2 and 29 DF,  p-value: 4.979e-09
```

$$\hat{mpg} = \begin{cases} 26.6636 & \text{if } cyl = 4 \\ 26.6636 - 6.9208 & \text{if } cyl = 6 \\ 26.6636 - 11.5636 & \text{if } cyl = 8 \end{cases}$$

$$\hat{mpg} = \begin{cases} 26.6636 & \text{if } cyl = 4 \\ 19.7428 & \text{if } cyl = 6 \\ 15.1 & \text{if } cyl = 8 \end{cases}$$

① F-statistic $H_0: \beta_1 = \beta_2 = 0$ H_A: At least one $\beta_j \neq 0$
p-value ≈ 0 Reject H_0 and determine that at least one predictor is significant.

② Adjusted R^2: About 71.4% of the variance in mpg can be explained by this equation.

③ Indival T-tests

$H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$ $t = -4.44$ p-value = 0.0001

$H_0: \beta_2 = 0$ $H_A: \beta_2 \neq 0$ $t = -8.91$ p-value ≈ 0

Since the p-values are small, (p-value ≤ 0.05)
the number of cylinders is significant in predicting mpg.

(recall: Multiple Linear Regression > Important Questions for Multivariate Regression)

Interaction

Two Important Assumptions

The **additive** assumption means that the effect of changes in a predictor X_j on the response Y is independent of the values of the other predictors. (covered in this note)

The **linear** assumption means that the change in the response Y due to a one-unit change in X_j is constant, regardless of the value of X_j. (Linearity will be covered in the next note)

Removing the Additive Assumption (Stock Price Example)

Recall the stock price example where we found that the best subset of predictors of the model is `Interest_Rate` and `Unemployment_Rate` and our model was:

$$\hat{\text{stock\_price}} = 1798.4 + 345.5 \times \text{Interest\_Rate} - 250.1 \times \text{Unemployment\_Rate}$$

We concluded that both *unemployment rate* and *interest rate* seem to be associated with *stock index price*.
The linear model that we formed assumes that the effect on the stock index price of increasing one percent of the interest rate is independent of the unemployment rate.

However if we want to be sure that *interest rate* is independent of *unemployment rate* we have to use an **interaction term** between them. The simplest method to construct an interaction term is to multiply the two predictors together.

So for our model we can do:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

$$\begin{aligned}\text{stock\_price} = \beta_0 + \beta_1 \times \text{intr\_rate} + \beta_2 \times \text{unemp\_rate} \\ + \beta_3 \times \text{unemp\_rate} \times \text{intr\_rate} + \epsilon\end{aligned}$$

If we factor our unemployment rate we get this:

$$= \beta_0 + \beta_1 \times \text{intr\_rate} + (\beta_2 + \beta_3 \times \text{intr\_rate}) \times \text{unemp\_rate} + \epsilon$$

Once we factor that out we can see that we have $(\beta_2 + \beta_3 \times \text{intr\_rate})$ as a coefficient for *unemployment rate*. So we can interpret β_3 as the increase in the effectiveness of *unemployment rate* for a one unit increase in *interest rate* (or vice versa).

In R:

```
> #Stock Price data
> #Model with Interaction term
> stock.int = lm(Stock_Index_Price~Interest_Rate*Unemployment_Rate)
> summary(stock.int)
Call:
lm(formula = Stock_Index_Price ~ Interest_Rate * Unemployment_Rate)
Residuals:
     Min       1Q   Median       3Q      Max
-156.009  -40.238   -8.873   52.131  122.073

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   2522.85    2634.04   0.958    0.350
Interest_Rate                  -32.49    1293.06  -0.025    0.980
Unemployment_Rate             -380.76     461.09  -0.826    0.419
Interest_Rate:Unemployment_Rate   68.54     233.53   0.293    0.772

Residual standard error: 72.15 on 20 degrees of freedom
Multiple R-squared:  0.8981,Adjusted R-squared:  0.8828
F-statistic: 58.74 on 3 and 20 DF,  p-value: 4.266e-10
```

(Instead of a `*` in the `lm` command a `:` can also work to tell R that we are looking for interaction. The second you put a star or colon R will recognize the interaction and also do the rest of the coefficients so there is no need to list them all out).

Based on that estimate we get the following output:

$$\begin{aligned}\text{stock\_price} \approx 2522.85 - 32.49 \times \text{intr\_rate} - 308.76 \\ \times \text{unemp\_rate} + 68.54 \\ \times (\text{intr\_rate} \times \text{unemp\_rate})\end{aligned}$$

$$\begin{aligned}= 2522.85 - 32.49 \times \text{intr\_rate} + (-380.76 + 68.54 \times \text{intr\_rate}) \\ \times \text{unemp\_rate}\end{aligned}$$

This means that increasing the unemployment rate by 1% will result in the stock index price increasing by $63.54 \times \text{intr\_rate} - 380.76$.

However, we can see from the *p-value* of the f-statistic that at least one of the predictors is needed, however from the *p-value of the t-test* for each predictor *none* of them are significant. This means that at least one of these terns are *not needed* in the model.

(Recall the T-test for multiple regression models).
You can use the step function on this model, for this example you will see the interaction term is removed and the AIC gets smaller this means that the interaction is not as significant.