MULTIPLE LINEAR REGRESSION WITH NUMERICAL CATEGORICAL VARIABLES

Cesar Acosta

Department of Industrial and Systems Engineering

University of Southern California

October 6, 2017

Models with one categorical variable

Consider a model with two predictors

- X_1, continuous
- X_2, categorical with two levels, 0 (population a) and 1 (population b)

Consider the linear statistical model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \tag{1}$$

where ε_i is the error term.

- Assume model (1) satisfies standard regression assumptions for each population a and b.
- Also assume that the error terms for the two populations have the same variance σ^2.

Models with one categorical variable

When X_2 is defined as categorical variable we have two resulting models.

For population a (when $X_2 = 0$) we have

$$E[Y] = \beta_0 + \beta_1 X_1 \tag{2}$$

for population b (when $X_2 = 1$) we have

$$E[Y] = (\beta_0 + \beta_2) + \beta_1 X_1 \tag{3}$$

- Two straight lines, with the same slope, but with different intercept.
- For population a slope is β_0 while for population b it is equal to $\beta_0 + \beta_2$

More than two levels

- Suppose X_2 is a categorical variable with three levels a, b and c.
- Consider using two binary variables defined as,

$$W_2 = \begin{cases} 1 & \text{for population } a \\ 0 & \text{otherwise,} \end{cases}$$

$$W_3 = \begin{cases} 1 & \text{for population } b \\ 0 & \text{otherwise.} \end{cases}$$

No additional binary variable is needed since population c is identified when W_2 and W_3 are both equal to zero.

More than two levels

Consider the model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 W_{i2} + \beta_3 W_{i3} + \varepsilon_i \tag{4}$$

The resulting fitted equations (one for each population) are

- For population c
$$E[Y] = \beta_0 + \beta_1 X_1 \tag{5}$$
- for population a
$$E[Y] = (\beta_0 + \beta_2) + \beta_1 X_1 \tag{6}$$
- for population b
$$E[Y] = (\beta_0 + \beta_3) + \beta_1 X_1. \tag{7}$$

More than two levels

- β_2 and β_3 indicate how much different the mean response of populations a and b are from that of population c, for any given value of X_1.
- In this case, the model from population c is referred to as the *base model*, and the parameters β_2 and β_3 are the incremental effects of populations a and b.
- Model (4) can be expanded to include interaction terms to allow for different slopes

More than two levels

- When a categorical variable is numeric is used as continuous variable, no binary variables are needed.
- For example,
  - X_2 is the number of doors of a car, and the mileage of cars is the response, $X_2 = 2$ or $X_2 = 4$
  - X_2 is the number of bedrooms of an apartment and the price is the response, $X_2 = 1$, $X_2 = 2$ or $X_2 = 3$
- As a result the fitted equation is that of a plane fitted on the $X_1 - X_2$ plane as opposed to m fitted lines on X_1, where m is the number of levels of X_2.

We now discuss on the benefits or adverse effects on this practice.

Numerical Categorical Predictors - EXAMPLE 1

Consider fitting a linear model with a categorical variable X_1 with three levels (1,7,13) and a continuous variable X_2.

X_1	X_2	Y
1	1.0	4.31
1	3.5	7.70
1	6.0	9.08
7	1.0	4.25
7	3.5	5.36
7	6.0	6.60
13	1.0	0.54
13	3.5	3.31
13	6.0	5.63

Numerical Categorical Predictors - EXAMPLE 1

Consider fitting a linear model with a categorical variable X_1 with three levels (1,7,13) and a continuous variable X_2.

X_1	X_2	Y
1	1.0	4.31
1	3.5	7.70
1	6.0	9.08
7	1.0	4.25
7	3.5	5.36
7	6.0	6.60
13	1.0	0.54
13	3.5	3.31
13	6.0	5.63

First consider X_1 as continuous

Numerical Categorical Predictors - EXAMPLE 1

When both X_1 and X_2 are included in the model *as continuous variables*

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.60628    0.59018   7.805 0.000233 ***
x1          -0.32250    0.04926  -6.547 0.000607 ***
x2           0.81400    0.11822   6.886 0.000463 ***

Residual standard error: 0.7239 on 6 degrees of freedom
Multiple R-squared:  0.9377,    Adjusted R-squared:  0.9169
F-statistic: 45.14 on 2 and 6 DF,  p-value: 0.000242
```

Numerical Categorical Predictors - EXAMPLE 1

Now consider X_1 as categorical variable

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.1810     0.6245   6.695  0.00112 **
x17          -1.6267     0.6276  -2.592  0.04873 *
x113         -3.8700     0.6276  -6.166  0.00163 **
x2            0.8140     0.1255   6.485  0.00130 **

Residual standard error: 0.7687 on 5 degrees of freedom
Multiple R-squared:  0.9414,    Adjusted R-squared:  0.9063
F-statistic:  26.8 on 3 and 5 DF,  p-value: 0.001654
```

In this example, both models fit the data well, showing similar adequacy of fit values.

Numerical Categorical Predictors - EXAMPLE 2

Now let us consider the following observations

Table 1: Example 2 data set

X_1	X_2	Y
0	-0.10	19.19
0	2.53	22.74
0	4.86	23.91
1	0.26	7.07
1	2.55	7.93
1	4.87	8.93
2	0.08	20.63
2	2.62	23.46
2	5.09	25.75

Numerical Categorical Predictors - EXAMPLE 2

If both X_1 and X_2 are included in the model *as continuous variables*

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.1678     5.6816   2.670    0.037 *
x1            0.6019     3.4742   0.173    0.868
x2            0.7769     1.4275   0.544    0.606

Residual standard error: 8.505 on 6 degrees of freedom
Multiple R-squared:  0.05259,   Adjusted R-squared:  -0.2632
F-statistic: 0.1665 on 2 and 6 DF,  p-value: 0.8504
```

Numerical Categorical Predictors - EXAMPLE 2

If both X_1 and X_2 are included in the model *as continuous variables*

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.1678     5.6816   2.670    0.037 *
x1            0.6019     3.4742   0.173    0.868
x2            0.7769     1.4275   0.544    0.606

Residual standard error: 8.505 on 6 degrees of freedom
Multiple R-squared:  0.05259,   Adjusted R-squared:  -0.2632
F-statistic: 0.1665 on 2 and 6 DF,  p-value: 0.8504
```

- R^2 is close to 0.05, the explained variation of the response about the fitted equation is negligible
- The Adjusted R-squared is negative and equal to -0.2632.
- Both predictors X_1, X_2 seem not to be useful for predicting Y.

Numerical Categorical Predictors - EXAMPLE 2

When factor X_1 is properly defined using indicator variables X_{11} and X_{12}, as shown

X_{11}	X_{12}	X_2	Y
0	0	-0.10	19.19
0	0	2.53	22.74
0	0	4.86	23.91
1	0	0.26	7.07
1	0	2.55	7.93
1	0	4.87	8.93
0	1	0.08	20.63
0	1	2.62	23.46
0	1	5.09	25.75

the result is

Numerical Categorical Predictors - EXAMPLE 2

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.9650     0.5802  34.413 3.90e-07 ***
x11         -14.0760     0.6703 -20.998 4.54e-06 ***
x12           1.1974     0.6705   1.786  0.13418
x2            0.8155     0.1378   5.920  0.00196 **
Residual standard error: 0.8207 on 5 degrees of freedom
Multiple R-squared:  0.9926,    Adjusted R-squared:  0.9882
F-statistic:   225 on 3 and 5 DF,  p-value: 9.416e-06
```

Numerical Categorical Predictors - EXAMPLE 2

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.9650     0.5802  34.413 3.90e-07 ***
x11         -14.0760     0.6703 -20.998 4.54e-06 ***
x12           1.1974     0.6705   1.786  0.13418
x2            0.8155     0.1378   5.920  0.00196 **
Residual standard error: 0.8207 on 5 degrees of freedom
Multiple R-squared:  0.9926,    Adjusted R-squared:  0.9882
F-statistic:   225 on 3 and 5 DF,  p-value: 9.416e-06
```

- These values show that the fitted model is highly significant.
- The R-squared is very close to 1.
- The set of two predictors explain 99.26% of the response variability. The adjusted R-squared is also high, being 0.988.
- After validation this model can and should be used.

Numerical Categorical Predictors - EXAMPLE 2

The corresponding fitted equations for prediction at each level are given by

$$E[Y] = \begin{cases} 19.9650 \qquad\qquad\qquad + 0.8155X_2 & \text{when } X_1 = 0 \\ (19.9650 - 14.076) + 0.8155\,X_2 & \text{when } X_1 = 1 \\ (19.9650 + 1.1974) + 0.8155\,X_2 & \text{when } X_1 = 2 \end{cases}$$

Numerical Categorical Predictors - EXAMPLE 2

- This is an extreme example
- But it shows that care show be taken, when dealing with *numerical* categorical variables
- Do not use a *numerical* categorical variable as continuous, ... blindly
- Fit both models (factor as continuous, factor as a factor), ... then compare

Examples from the literature - Example 1

- To predict the price of a GM car, Kuiper (2008) used data collected from the Kelly Blue Book to build multivariate regression models.
- Those models were developed to show how to check for standard regression assumptions, multicollinearity, and to explain variable selection methods.
- The data set, available from the ASA journals website, includes 13 variables from 804 GM cars.
- The variables are shown in the following output.
- There are four factors and seven numerical predictors.
- Some of these predictors are actually categorical variables (Cylinder, Doors, Cruise, Sound, and Leather)
- leaving just two continuous predictors (Mileage and Liter) available for modeling.

Examples from the literature - Example 1

```
'data.frame':   804 obs. of  12 variables:
 $ Price   : num  17314 17542 16219 16337 16339 ...
 $ Mileage : int  8221 9135 13196 16342 19832 22236 22576 22964 ...
 $ Make    : Factor w/ 6 levels "Buick","Cadillac",..: 1 1 1 1 1 ...
 $ Model   : Factor w/ 32 levels "9-2X AWD","9_3",..: 9 9 9  9 9 ...
 $ Trim    : Factor w/ 47 levels "Aero Conv 2D",..: 40 40 40 40 ...
 $ Type    : Factor w/ 5 levels "Convertible",..: 4 4 4 4 4 4 4 4 ...
 $ Cylinder: int  6 6 6 6 6 6 6 6 6 6 ...
 $ Liter   : num  3.1 3.1 3.1 3.1 3.1 3.1 3.1 3.1 3.1 3.1 ...
 $ Doors   : int  4 4 4 4 4 4 4 4 4 4 ...
 $ Cruise  : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Sound   : int  1 1 1 0 0 1 1 1 0 1 ...
 $ Leather : int  1 0 0 0 1 0 0 0 1 1 ...
```

Examples from the literature - Example 1

To show variable selection techniques, Minitab was used to identify the best subset of predictors. As a result the author builds a model that includes variables Cylinder, Doors, Cruise, Sound, and, Leather.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.323e+03  1.771e+03   4.135 3.92e-05 ***
Mileage     -1.705e-01  3.186e-02  -5.352 1.14e-07 ***
Cylinder     3.200e+03  2.030e+02  15.765  < 2e-16 ***
Doors       -1.463e+03  3.083e+02  -4.747 2.45e-06 ***
Cruise       6.206e+03  6.515e+02   9.525  < 2e-16 ***
Sound       -2.024e+03  5.707e+02  -3.547 0.000412 ***
Leather      3.327e+03  5.971e+02   5.572 3.45e-08 ***
Residual standard error: 7387 on 797 degrees of freedom
Multiple R-squared:  0.4457,    Adjusted R-squared:  0.4415
F-statistic: 106.8 on 6 and 797 DF,  p-value: < 2.2e-16
```

Examples from the literature - Example 1

These predictors explain roughly 44.5% of the Price variability. The fittedd equation to predict the price Y of a GM car is given by

$$\begin{aligned} E[Y] = 7323 - 0.171 \text{ Mileage} + 3200 \text{ Cylinder} \\ + 1463 \text{ Doors} + 6206 \text{ Cruise} - 2024 \text{ Sound} + 3327 \text{ Leather} \end{aligned} \tag{8}$$

Note this model assumes that all predictors are continuous variables

Examples from the literature - Example 1

In the data set the categorical variable **Cylinder** has three levels (4, 6, or 8 cylinders). This variable be can included in the model as a factor,

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 18858.8032  1371.2165  13.753  < 2e-16 ***
Mileage        -0.1743     0.0283  -6.159 1.16e-09 ***
Cylinder6     613.7968   535.3060   1.147   0.2519
Cylinder8   17725.4990   795.9888  22.269  < 2e-16 ***
Doors        -538.1149   281.0627  -1.915   0.0559 .
Cruise       6821.0008   580.2470  11.755  < 2e-16 ***
Sound        -789.7920   513.9682  -1.537   0.1248
Leather      1107.8339   551.7116   2.008   0.0450 *
Residual standard error: 6562 on 796 degrees of freedom
Multiple R-squared:  0.5631,    Adjusted R-squared:  0.5593
F-statistic: 146.6 on 7 and 796 DF,  p-value: < 2.2e-16
```

This new model explains roughly 56% of the Price variability.

Examples from the literature - Example 1

This new model explains roughly 56% of the Price variability. The fitted equations for cars with different number of cylinders are

$$
\begin{aligned}
E[Y] &= 18859 - 0.1743 \text{ Mileage} - 538 \text{ Doors} + 6821 \text{ Cruise} - 790 \text{ Sound} + 1108 \text{ Leather} && \text{for 4 cyl.}\\
E[Y] &= 19473 - 0.1743 \text{ Mileage} - 538 \text{ Doors} + 6821 \text{ Cruise} - 790 \text{ Sound} + 1108 \text{ Leather} && \text{for 6 cyl.}\\
E[Y] &= 36584 - 0.1743 \text{ Mileage} - 538 \text{ Doors} + 6821 \text{ Cruise} - 790 \text{ Sound} + 1108 \text{ Leather} && \text{for 8 cyl.}
\end{aligned}
$$

- Similarly binary 0-1 variables can be defined for the other categorical variables and refit
- We included `Doors, Sound, Leather` as categorical variables, however no improvement was found
- In fact a simpler model can be found by excluding these predictors

Examples from the literature - Example 1

We found that the model with predictors `Mileage, Cruise and Cylinder` has about the same prediction performance as model (8). For this simplified model

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.722e+04  7.350e+02  23.433  < 2e-16 ***
Mileage     -1.724e-01  2.839e-02  -6.074 1.93e-09 ***
Cylinder6    2.817e+02  5.239e+02   0.538    0.591
Cylinder8    1.826e+04  7.732e+02  23.620  < 2e-16 ***
Cruise1      6.858e+03  5.775e+02  11.875  < 2e-16 ***

Residual standard error: 6585 on 799 degrees of freedom
Multiple R-squared:  0.5584,    Adjusted R-squared:  0.5562
F-statistic: 252.6 on 4 and 799 DF,  p-value: < 2.2e-16
```

Examples from the literature - Example 1

The resulting fitted equations for cars with different number of cylinders and with or with no cruise control are as shown

Cylinder	Cruise	fitted equation
4	0	E[Y] = 17222 -0.1724 Mileage
4	1	E[Y] = 24080 -0.1724 Mileage
6	0	E[Y] = 17504 -0.1724 Mileage
6	1	E[Y] = 24362 -0.1724 Mileage
8	0	E[Y] = 35485 -0.1724 Mileage
8	1	E[Y] = 42344 -0.1724 Mileage

This set of equations explain an additional 11% of the variability of the car's Price.

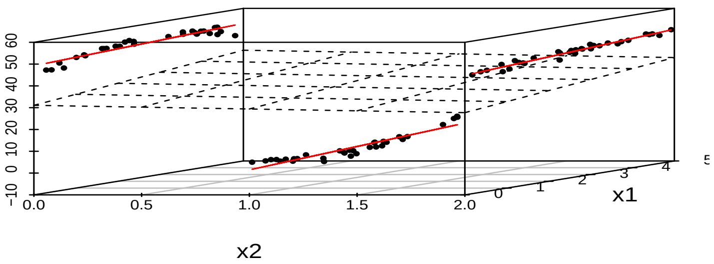
Why are the models different?

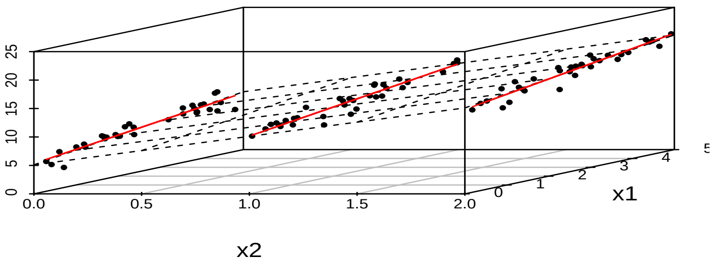
Consider a model with two predictors,

- X_1 is continuous
- X_2 a categorical with three levels 0, 1, and 2

Why are the models different?

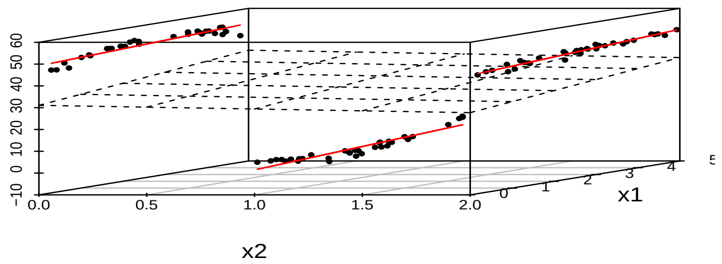
If both variables are included in the model as continuous then a fitted plane is found. Residuals are computed by the squared distance of each observation from that plane.

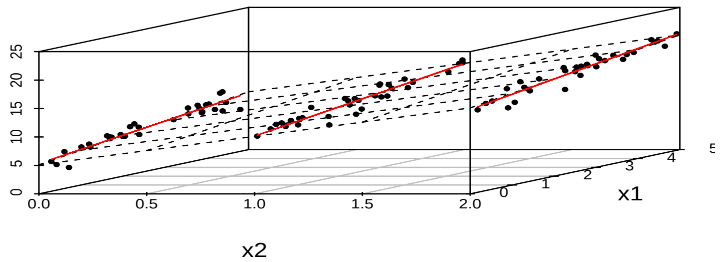




Why are the models different?

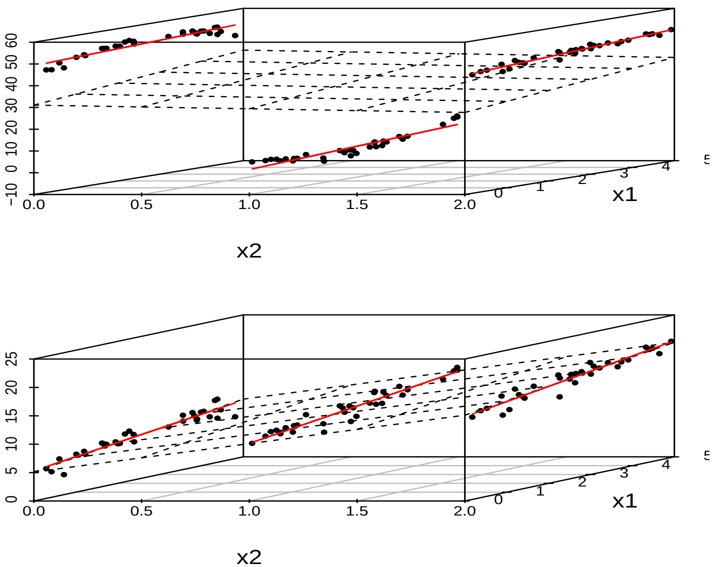
If X_2 is included in the model using binary variables, then for each level $j = 0, 1, 2$ a fitted equation is found. Residuals are computed by the squared distance of each observation from the fitted equation associated with that level j.





Why are the models different?

In the lower plot both models provide about the same fit. In this case defining the categorical variable as continuous or as categorical does not change the model performance



REFERENCES

- Kuiper S. (2008), Introduction to Multiple Regression: How much is Your Car Worth? *Journal of Statistics Education*, Vol. 16(3).
- `http://www.amstat.org/publications/jse/datasets/`