CATEGORICAL VARIABLES
-ENCODING-

Cesar Acosta, PhD

Department of Industrial and Systems Engineering
University of Southern California

Numerical Categorical Predictors – EXAMPLE 2

Consider the following dataset

X_1	X_2	Y
S	-0.10	19.19
S	2.53	22.74
S	4.86	23.91
M	0.26	7.07
M	2.55	7.93
M	4.87	8.93
L	0.08	20.63
L	2.62	23.46
L	5.09	25.75

Numerical Categorical Predictors – EXAMPLE 2

Consider the following dataset

X_1	X_2	Y
S	-0.10	19.19
S	2.53	22.74
S	4.86	23.91
M	0.26	7.07
M	2.55	7.93
M	4.87	8.93
L	0.08	20.63
L	2.62	23.46
L	5.09	25.75

LABEL ENCODING

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

X_1 and X_2 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

X_1 and X_2 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

The fitted plane is

$$E[Y] = 15.1678 + 0.6019\ X1\ - 0.7769\ \mathrm{X2}$$

Numerical Categorical Predictors – EXAMPLE 2

Replace X_1 with binary variables X_{11} and X_{12}

X_1	X_2	Y
S	-0.10	19.19
S	2.53	22.74
S	4.86	23.91
M	0.26	7.07
M	2.55	7.93
M	4.87	8.93
L	0.08	20.63
L	2.62	23.46
L	5.09	25.75

ONE-HOT ENCODING

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

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
```

Numerical Categorical Predictors – EXAMPLE 2

The fitted equations for each level are

$$E[Y] = \begin{cases} 19.9650 + 0.8155X_2 & \text{when } X_1 = S \\ (19.9650 - 14.076) + 0.8155X_2 & \text{when } X_1 = M \\ (19.9650 + 1.1974) + 0.8155X_2 & \text{when } X_1 = L \end{cases}$$

What encoding is better?

Numerical Categorical Predictors – EXAMPLE 2

LABEL ENCODING

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

ONE-HOT ENCODING

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

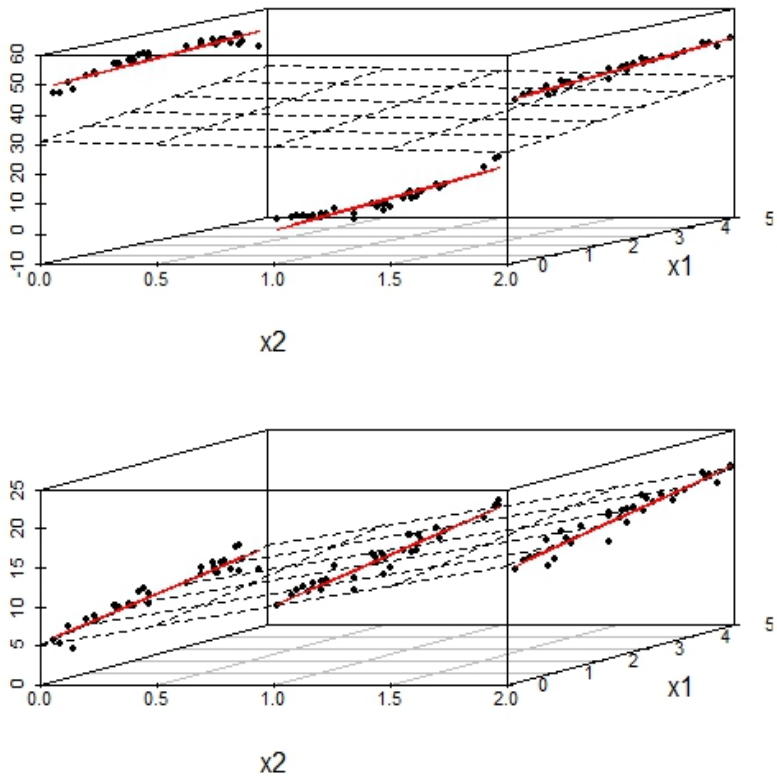
Numerical Categorical Predictors – EXAMPLE 2

	LABEL ENCODING	ONE-HOT ENCODING
R-squared	0.05259	0.9926
Adjusted R-squared:	-0.2632	0.9882

Why are the models different?

Why are the models different?

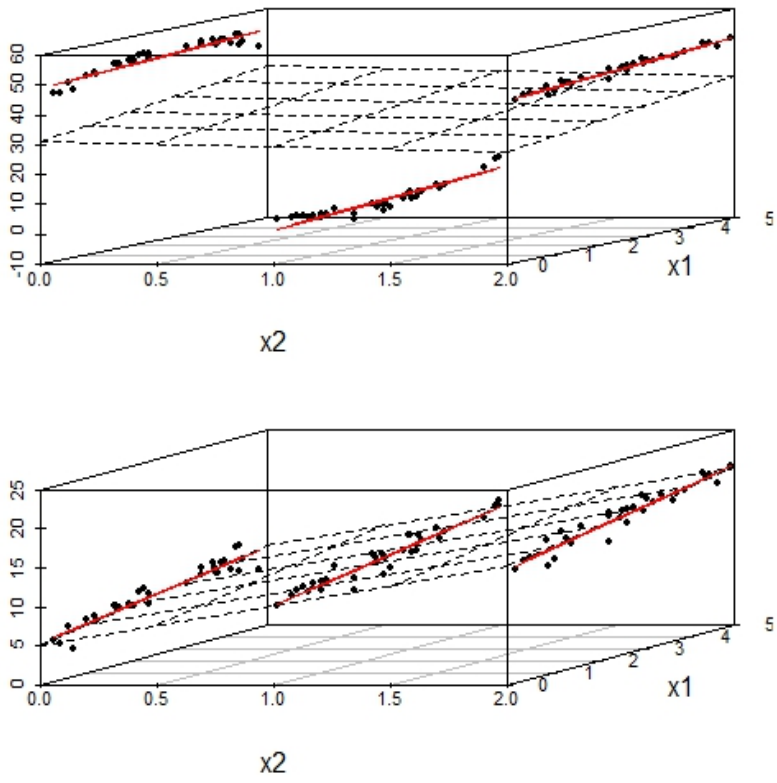
A fitted plane is found if both variables are included in the model as continuous.



Why are the models different?

A fitted plane is found if both variables are included in the model as continuous.

If X_2 is included in the model using indicator variables,
for each level j = 0, 1, 2, a fitted equation is found.



Why are the models different?

The fitted lines may be away from the fitted plane

or

may be close to the fitted plane