Using Categorical Variables in Linear Models

Jolene Liu, Sarah Kim, and Cesar Acosta-Mejia
Daniel J. Epstein Department of Industrial & Systems Engineering
University of Southern California
Los Angeles, CA 90089, USA
joleneli@usc.edu, sarahyki@usc.edu, acostame@usc.edu

Abstract

Categorical variables that are numerical are sometimes, included in a regression model, as continuous predictors. In this paper we show that the fit of regression models may be very different when numerical categorical variables are considered as continuous or as factors. With a small example we show that it is possible that the adjusted R-squared can be negative in the former case and close to one in the latter. We use data visualization to explain the difference. We build models with categorical variables, using data from the car's Consumer Reports, to show how to improve the fit and to explain outliers.

Keywords

Categorical variables, Regression Models, Adjusted R-squared, Dummy variables

1. Introduction

In Science and Engineering most regression models use continuous and categorical variables to predict a response. A categorical variable is sometimes referred to as a factor, and the possible values of the variable are referred to as the *levels* of the factor. For instance, the levels of variable origin are *foreign* and *domestic*, and those of variable temperature are *high*, *medium* or *low*. We can associate numbers 0, and 1 to the levels of origin or numbers 0, 1, and 2 to the temperature levels, but the numbers should still be considered as labels.

Many standard textbooks discuss how to build regression models with continuous and categorical variables. We refer the reader to Kutner et al. (2004), Montgomery et al. (2013), Mendenhall and Sincich (2011), and Gujarati and Porter (2009). There is also a large number of programs that can efficiently build statistical models (SAS, SPSS, STATA, and *R*). Actually, most commonly used are spreadsheets since little or no additional training is needed. Unfortunately, some spreadsheets do not have the capability to create and analyze models with categorical variables. As a result, some users tend to treat categorical variables as continuous variables finding models with smaller prediction performance.

In this paper we compare these models. We find that very different models may result. We also show that regression models with categorical variables are useful to identify outliers. Our results are obtained using *R* 3.2.3 (2014). It is a language and an environment useful for statistical computing and graphics. There is a wide variety of sources that help introduce the practitioner to *R*. Many Statistics textbooks introduce the reader to *R*. We refer the reader to Akritas (2015), Gardener (2012), and Crawley (2014).

2. Models with one categorical variable

Suppose a linear regression model is to be defined for estimating a response Y using a continuous predictor X_1. It is of interest to define the model for two populations. These populations are defined by using a categorical variable X_2 with levels a and b.

Let us consider a model with two predictors X_1 and X_2 where X_1 is continuous and X_2 is a categorical variable with two levels defined by 0 (population a) and 1 (population b). Therefore, the populations are classified by means of a binary 0-1 variable. 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. Further we assume that model (1) satisfies standard regression assumptions for each population *a* and *b*. We also assume that the error terms for the two populations have the same variance σ^2.

When X_2 is defined as categorical variable we have two resulting models. For population *a* (when $X_2 = 0$) we have

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

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

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

Both models represent a straight line, with the same slope, but with different intercept. For population *a* the slope is β_0 while for population *b* it is equal to $\beta_0 + \beta_2$.

Model (1) is appropriate when the resulting effect of continuous predictor X_1 on the mean response is the same for both populations. The estimated response of model (1), given by $E[Y]$, is different for the two populations, and β_2 indicates how much higher or lower the estimated response is, for any given value of X_1.

When the assumption that the effect of the predictor on the mean response is not the same for both populations, model (1) can be extended to include an interaction term, but we do not consider those cases in this article.

2.1 More than two levels

If the statistical model includes a categorical variable with more than two levels, then more binary variables are needed. Suppose X_2 is a categorical variable with three levels *a*, *b*, and *c*. Then to appropriately use this categorical variable in the model, two binary variables are required. Let us define them 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.

A statistical model, assuming that the effect of the continuous predictor X_1 on the mean response is the same for all three populations, is

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

The resulting models are

$$\begin{aligned} E[Y] &= \beta_0 + \beta_1 X_1 && \text{for } c && (5) \\ E[Y] &= (\beta_0 + \beta_2) + \beta_1 X_1 && \text{for } b && (6) \\ E[Y] &= (\beta_0 + \beta_3) + \beta_1 X_1. && \text{for } a && (7) \end{aligned}$$

All models represent a straight line, with the same slope, but with different intercept. And 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 for 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, but we do not consider those cases in this article.

3. An Extreme Example

Consider fitting statistical models to a small data set with two predictors, one categorical (X_1) and the other continuous (X_2). Our objective is to compare models (1) and (4) with the observations given in Table 1. To fit model (4) the categorical variable X_1 with three levels is replaced by two binary (dummy) variables X_{11} and X_{12}, as shown in Table 2. Using R these variables are not to be created since R creates them once they are defined as factors. The comparison of models (1) and (4) is as follows.

When both X_1 and X_2 are included in the model as *continuous variables* the coefficients table and the R-squared values are given by

```
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
```

The R-squared is close to 0.05, indicating that the explained variation of the response about the fitted equation is negligible. The Adjusted R-squared is negative and equal to -0.2632. These values show that the fit is poor and the resulting model is not useful for prediction purpose.

Table 1. Data for Model (1)

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

Table 2. Data for Model (4)

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

However, when X_1 is defined using indicator variables X_{11} X_{12}, as shown in Table 2 the results are

```
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 model explains 99.26% of the response variability. The adjusted R-squared is also high, being 0.988. The corresponding fitted equations at each level are given by

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

These two models show very different performance on the same data set, the second being the best of the two. This example shows that treating a categorical numerical variable appropriately may result in a better statistical model for prediction.

4. Predicting City Mileage

The *R* library `MASS` includes the dataframe `Cars93`. It is a selection of 93 car models from the Consumer Reports. It includes 26 variables, such as manufacturer, price, fuel efficiency, engine's size and power, car's size and other properties such as number of airbags, drive train, and origin. The description of each variable is found in `https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/Cars93.html`

We consider predicting the city mileage of a new car based on the number of revolutions per minute at maximum horsepower, and the weight of the car. Let us denote the city mileage `MPG.city` by Y, the `RPM` by X_1, and the `weight` by X_2.

When both predictors are considered in the model as continuous variables, the estimated coefficients and corresponding ANOVA table are

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.688e+01  4.254e+00  11.020   <2e-16 ***
RPM          2.582e-05  5.906e-04   0.044    0.965
Weight      -8.021e-03  5.974e-04 -13.426   <2e-16 ***
Residual standard error: 3.055 on 90 degrees of freedom
Multiple R-squared:  0.7109,    Adjusted R-squared:  0.7045
F-statistic: 110.6 on 2 and 90 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table
          Df  Sum Sq Mean Sq F value    Pr(>F)
RPM        1  382.96  382.96   41.03 6.687e-09 ***
Weight     1 1682.58 1682.58  180.27 < 2.2e-16 ***
Residuals 90  840.03    9.33
```

Based on the p-values the coefficients table shows that `RPM` is not a significant predictor, while the Analysis of variance Table shows the opposite. This contradiction may indicate that including `RPM` as a continuous variable is not appropriate. Note however that the R-squared values are around 0.71. For comparison we record the resulting fitted equation for Y, the city mileage

$$E[Y] = 46.88 + 0.00002582 \text{ RPM} + 0.008021 \text{ Weight} \tag{8}$$

To clarify the contradiction, consider using `RPM` as a categorical variable. Table 3 shows the number of cars grouped by RPM.

Table 3. Number of cars grouped by RPM values

RPM	3800	4000	4100	4200	4400	4500	4600	4800	5000	5100	5200	5300
cars	1	2	1	3	1	1	4	13	10	1	10	1
RPM	5400	5500	5550	5600	5700	5750	5800	5900	6000	6200	6300	6500
cars	4	8	1	6	2	1	4	1	14	1	1	2

There is only one car for some RPM values (3800, 4100, 4400, ... ,6300) and cars with 6000 RPM are the largest group in the data set. Let us consider a new model where RPM is a categorical variable with 3800 RPM as the base level. To build that model 23 binary variables are needed.

After fitting this new model, the resulting table of coefficients is given by

```
Coefficients:
           Estimate  Std. Error  Pr(>|t|)
Intercept 47.0412933  2.8621954  < 2e-16 ***
RPM4000    0.0342904  2.7698998  0.990159
RPM4100   -2.9223233  3.1880935  0.
RPM4200   -0.7249827  2.6034637  0.781498
RPM4400   -1.3397479  3.1883602  0.675664
RPM4500    0.7186849  3.1926716  0.822573
RPM4600   -1.5487233  2.5236976  0.541479
RPM4800   -0.9590356  2.3407744  0.683307
RPM5000   -1.0926181  2.3804357  0.647699
RPM5100   -4.3596932  3.2058604  0.178349
RPM5200   -1.7374400  2.3966732  0.470977
RPM5300    0.2620712  3.1884275  0.934734
RPM5400   -0.3257535  2.5468986  0.898604
RPM5500   -1.3766630  2.4127084  0.570160
RPM5600    1.2049205  2.4703716  0.627297
RPM5700    7.6789698  2.7990959  0.007768 **
RPM5750   -5.0104987  3.2380632  0.126415
RPM5800   -2.6969918  2.5358764  0.291302
RPM5900   13.2127342  3.2469691  0.000125 ***
RPM6000   -0.5621574  2.3544584  0.812008
RPM6200   -1.8352000  3.1930724  0.567361
RPM6300   -1.0297425  3.2185995  0.749999
RPM6500   -5.9714850  2.7955638  0.036278 *
Weight    -0.0077677  0.0004885  < 2e-16 ***
Residual standard error: 2.254 on 68 degrees of freedom
Multiple R-squared:  0.8811,    Adjusted R-squared:  0.8391
F-statistic: 20.99 on 24 and 68 DF,  p-value: < 2.2e-16
```

Clearly the fit is improved, but not all RPM levels are significant. RPM levels not significant (with no * in the right most column) should have the same intercept as that of the base level. We combine all non-significant levels with the base RPM level 3800 and refit. The new estimates are shown below and the three fitted equations are shown in Table 4. This model explains 84.77% of the variability of the city mileage. An improvement of roughly 15% over the model with continuous variables. The adjusted R-square also improves by 14%.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 45.4630971  1.2738638  35.689  < 2e-16 ***
RPM5700      8.7948426  1.6131799   5.452 4.50e-07 ***
RPM5900     14.3836351  2.2755915   6.321 1.04e-08 ***
RPM6500     -4.8634115  1.6113206  -3.018  0.00333 **
Weight      -0.0075944  0.0004038 -18.807  < 2e-16 ***
Residual standard error: 2.243 on 88 degrees of freedom
Multiple R-squared:  0.8477,    Adjusted R-squared:  0.8407
F-statistic: 122.4 on 4 and 88 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table
          Df  Sum Sq Mean Sq F value    Pr(>F)    
RPM        3  683.98  227.99  45.328 < 2.2e-16 ***
Weight     1 1778.97 1778.97 353.686 < 2.2e-16 ***
Residuals 88  442.62    5.03                      
```

Table 4. Fitted equations for City mileage

RPM	fitted equation				
6500	E[Y] = (45.463 – 4.8634)	-	0.0076 Weight	=	40.6000 - 0.0076 Weight
5900	E[Y] = (45.463 + 14.3836)	-	0.0076 Weight	=	59.8460 - 0.0076 Weight
5700	E[Y] = (45.463 + 8.7948)	-	0.0076 Weight	=	54.2578 - 0.0076 Weight
other	E[Y] = 45.463	-	0.0076 Weight		

To explain why the fitted model with categorical variables improves over the model defined by (8) we refer to Table 3. From that table it can be seen that there are two cars with 6500 RPM and 5700 RPM, and one car with 5900 RPM. These cars appear in the data set in rows 32 and 57 (6500 RPM), 5 and 39 (5700 RPM), and row 42 (5900 RPM). Figure 1 shows the fitted equations from Table 4. By using categorical variables new fitted equations were found, which fit better, the cars in categories 5700, 5900, and 6900 RPM. As a result, the residuals for these cars are substantially reduced and the R-squared and adjusted R-squared improved. In this case, the fitted equations proved to be useful to explain the outliers given by cars in rows 39, 42, and 57. These cars belong to a different category.

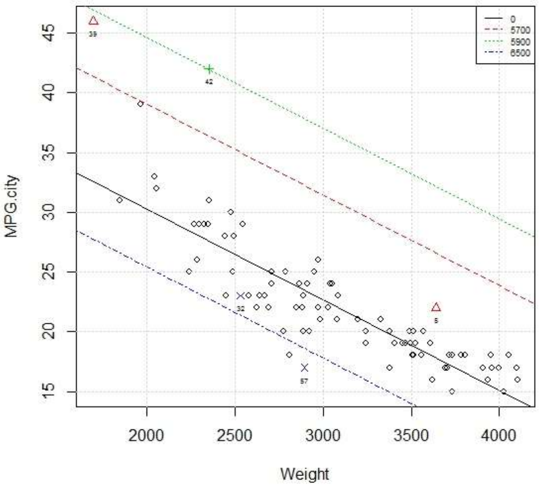

Figure 1. Fitted equations for City mileage

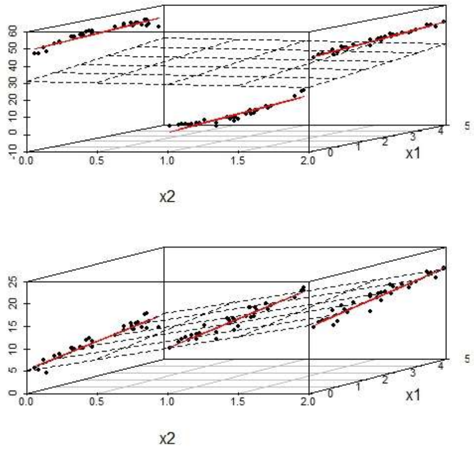

Figure 2. Fitted plane and fitted lines – two cases

5. Why are the models different?

Consider a linear model with two predictors X_1 and X_2, the second one being a categorical variable (a factor) with three levels (see Fig. 2). When the categorical variable is incorporated in the model using binary variables, fitted equations are found for each level of the factor (the fitted lines shown in Fig. 2).

However, if the factor is considered as a continuous variable in the model, a fitted plane is found. If the observations do not lie close to the plane, the resulting fit will be poor and the fitted equations may show a better fit, as in the top box in Figure 2. The bottom box shows a case where the fitted lines and the fitted plane would provide similar predictive performance.

6. Conclusion

Regression models may include categorical predictors that are numerical. As a result, two linear models may be considered. The first would include the categorical variables as continuous predictors. The second model would use binary variables to define the categorical predictors. In this paper we have shown that it is possible that the model adequacy as defined by the R-squared values may be very different, and as a result, considerably different prediction performances. When these two models can be constructed, the data analyst should fit both, compare their goodness of fit and choose the model with binary variables if it clearly outperforms the simpler model (that with continuous predictors). If both models show similar goodness of fit values, the simpler model can be selected for prediction.

References

Akritas M., *Probability & Statistics with R for Engineers and Scientists*, Pearson, London, 2015.

Chapman, C. and Feit E. M., *R for Marketing Research and Analytics*, Springer, New York, 2015.

Crawley M., *Statistics: An Introduction Using R,* 2nd Edition, Wiley, New York, 2014.

Gardener M., *Statistics for Ecologists Using R and Excel: Data Collection, Exploration, Analysis and Presentation*, Pelagic Publishing, London, 2012.

Gujarati D. and Porter D., *Basic Econometrics*, 5th Edition, McGraw-Hill, New York, 2009.

Kuiper S., *Introduction to Multiple Regression: How much is Your Car Worth?,* Journal of Statistics Education, vol. 16, no. 3, 2008.

R Core Team, 2014, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. Available: http://www.r-project.org/.

Kutner M. H., Neter J., Nachsteim C. J., and Li W., *Applied Linear Statistical Models*, McGraw-Hill, New York, 2004.

Montgomery D. C., Peck E. A., and Vining G. G., *Introduction to Linear Regression Analysis*, 5th Edition, Wiley, New York, 2013.

Mendenhall W. and Sincich T., A *Second Course in Statistics: Regression Analysis*, 7th Edition, Pearson, New York, 2011.

Biographies

Sarah Y Kim holds a in BS Industrial & Systems Engineering and a BA in Economics both from the University of Southern California. She is interested in data analytics applications in information technology and healthcare. She has worked for the Enterprise Data and Analytics Team and the Keck School of Medicine, both from USC.

Jolene Liu is an Industrial & Systems Engineering MS student at the University of Southern California. She holds a BS degree in Industrial & Systems Engineering from USC. She is interested in data analytics applications in information technology and supply chain management. She has worked for Protiviti's Internal Audit and Financial Advisory practice, KPMG's KTech Advisory Solutions group, and Lightspeed Aviation's engineering department.

Cesar Acosta-Mejia is a faculty member of the Department of Industrial and Systems Engineering at USC. He holds a Ph.D. in Industrial Engineering from Texas A&M University, and a Ph.D. in Statistics from University of Texas at Dallas. His main interests include predictive and prescriptive analytics using financial data. Dr. Acosta is the instructor of the graduate programs in Financial Engineering and Analytics at USC. He has published in Communications in Statistics Theory and Methods, Communications in Statistics Simulation and Computation, Quality Engineering, and other top Industrial Engineering and Statistics journals.