Consider predicting variable `y` using predictors `x1` and `x2`. The first predictor `x1` is a three-level categorical variable. In this example we compare two models. One in which `x1` is categorical but included as continuous (numerical) variable. The other in which `x1` is properly included as categorical variable.

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

1. Substitute the levels of x_1 with $0, 1, 2$.

2. Fit a linear regression model. What is the R^2 of this model?

3. Substitute the three levels of the original variable x_1 with binary (dummy) variables $0, 1$. Use the `as.factor` function.

4. Fit a linear regression model. What is the R^2 of this model?

5. Substitute the three levels of the original variable x_1 with binary (dummy) variables $0, 1$. Do not use the `as.factor` function. Verify that the last two models agree.

```
setwd("C:/Users/USC Guest/Downloads2")
d0 = read.table("example2b.txt",header=T)
#  x1    x2     y
#1  S -0.10 19.19
#2  S  2.53 22.74
#3  S  4.86 23.91
#4  M  0.26  7.07
#5  M  2.55  7.93
#6  M  4.87  8.93
#7  L  0.08 20.63
#8  L  2.62 23.46
#9  L  5.09 25.75
str(d0)
#'data.frame':   9 obs. of  3 variables:
# $ x1: Factor w/ 3 levels "L","M","S": 3 3 3 2 2 2 1 1 1
# $ x2: num  -0.1 2.53 4.86 0.26 2.55 4.87 0.08 2.62 5.09
# $ y : num  19.19 22.74 23.91 7.07 7.93 ...

d1 = d0
d1$x1 = rep(c(0,1,2),each=3)
#  x1    x2     y
#1  0 -0.10 19.19
#2  0  2.53 22.74
#3  0  4.86 23.91
#4  1  0.26  7.07
#5  1  2.55  7.93
#6  1  4.87  8.93
#7  2  0.08 20.63
#8  2  2.62 23.46
#9  2  5.09 25.75
str(d1)
# 'data.frame':   9 obs. of  3 variables:
#  $ x1: int  0 0 0 1 1 1 2 2 2
#  $ x2: num  -0.1 2.53 4.86 0.26 2.55 4.87 0.08 2.62 5.09
#  $ y : num  19.19 22.74 23.91 7.07 7.93 ...

m1= lm(y~.,d1)
summary(m1)

#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
```

```
#============================================================
d2 = d0
d2$x1 = as.factor(d2$x1)
str(d2)
# 'data.frame':   9 obs. of  3 variables:
# $ x1: Factor w/ 3 levels "0","1","2": 1 1 1 2 2 2 3 3 3
# $ x2: num  -0.1 2.53 4.86 0.26 2.55 4.87 0.08 2.62 5.09
# $ y : num  19.19 22.74 23.91 7.07 7.93 ...

m2= lm(y~.,d0)
summary(m2)

#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

big change in R-squared


#============================================================
# by hand

d3 = d0
d3$x11 = rep(c(0,1,0),each=3)
#  x1    x2     y x11
#1  0 -0.10 19.19   0
#2  0  2.53 22.74   0
#3  0  4.86 23.91   0
#4  1  0.26  7.07   1
#5  1  2.55  7.93   1
#6  1  4.87  8.93   1
#7  2  0.08 20.63   0
#8  2  2.62 23.46   0
#9  2  5.09 25.75   0
```

```
d3$x12 = rep(c(0,0,1),each=3)
d3$x1=NULL
d3
#     x2     y x11 x12
#1 -0.10 19.19   0   0
#2  2.53 22.74   0   0
#3  4.86 23.91   0   0
#4  0.26  7.07   1   0
#5  2.55  7.93   1   0
#6  4.87  8.93   1   0
#7  0.08 20.63   0   1
#8  2.62 23.46   0   1
#9  5.09 25.75   0   1

# reorder
d3 = d3[,c(3,4,1,2)]
#  x11 x12    x2     y
#1   0   0 -0.10 19.19
#2   0   0  2.53 22.74
#3   0   0  4.86 23.91
#4   1   0  0.26  7.07
#5   1   0  2.55  7.93
#6   1   0  4.87  8.93
#7   0   1  0.08 20.63
#8   0   1  2.62 23.46
#9   0   1  5.09 25.75

m3= lm(y~.,d3)
summary(m3)
#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

# same as model m2
```