USING CATEGORICAL VARIABLES IN LINEAR MODELS

Jolene Liu, Sarah Kim, and Cesar Acosta-Mejia

Department of Industrial and Systems Engineering
University of Southern California

September 18, 2018

INTRODUCTION

- Categorical variables can be numerical/non-numerical
  - N. doors of a car
- Regression models with categorical variables
  - Numerical
  - Factors (using binary variables)
- Does it matter?

INTRODUCTION

- Models may be similar or very different
- 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 show that the fit of regression models may be very different
    - When numerical categorical variables are considered as continuous or as factors

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

First consider X_1 as continuous

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

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 *
x17          -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

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 -0.23632
- Both predictors X_1 and 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_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

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, 0.988

Numerical Categorical Predictors – EXAMPLE 2

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

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

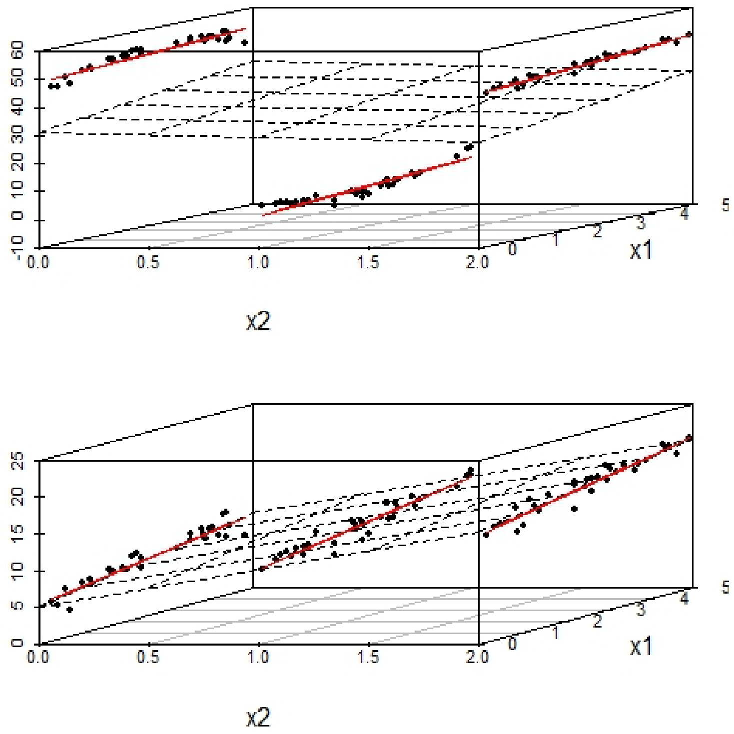
Why are the models different?

Consider a model with two predictors,

- Continuous
- Categorical with three levels 0, 1, and 2

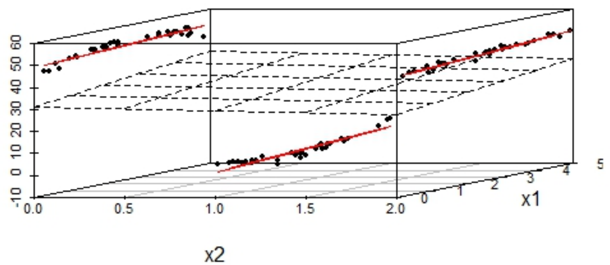
Why are the models different?

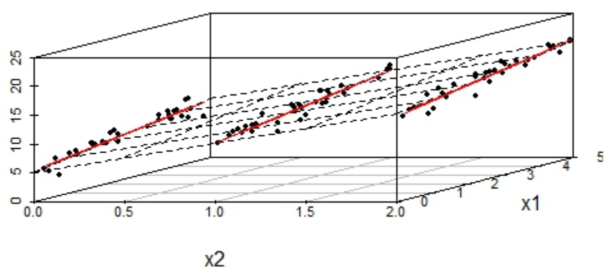
If both variables are included in the model as continuous then a fitted plane is found. Residuals are computed by the squared distance of each observation from that plane



Why are the models different?

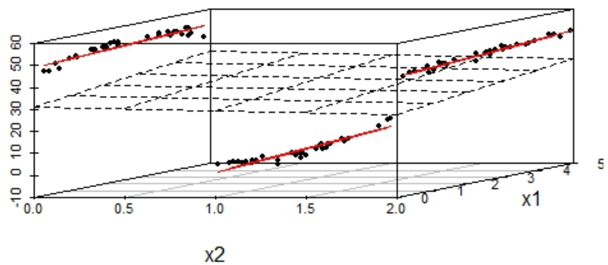
If X_2 is included in the model using binary variables, then for each level j = 0, 1, 2 a fitted equation is found. Residuals are computed by the squared distance of each observation from the fitted equation associated with that level j

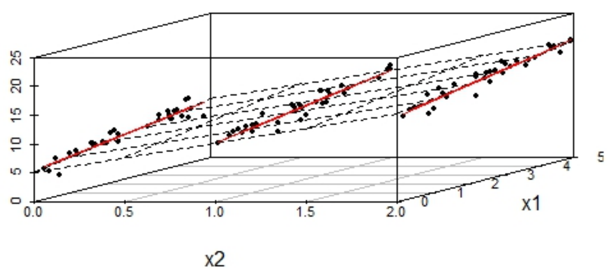




Why are the models different?

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

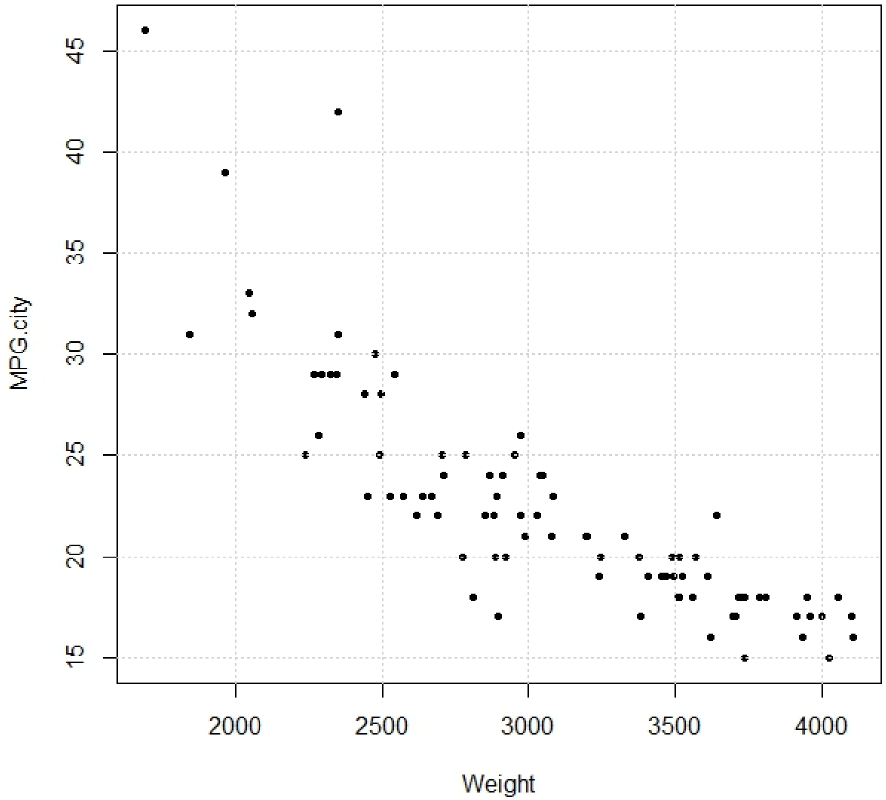




Predicting City Mileage – EXAMPLE 3

- 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
- `stat.ethz.ch/R-manual/R-devel/library/MASS/html/Cars93.html`

Predicting City Mileage – EXAMPLE 3



Predicting City Mileage – EXAMPLE 3

- 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
- Denote the city mileage MPG.city by Y, the RPM by X_1, and the weight by X_2

Predicting City Mileage – EXAMPLE 3

If both predictors are considered in the model as *continuous variables*

```
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
```

Predicting City Mileage – EXAMPLE 3

```
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
```

The coefficients table shows that RPM is not significant, while the Analysis of Variance Table shows the opposite.

Predicting City Mileage – EXAMPLE 3

```
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
```

This contradiction may indicate that the model with two continuous variables is not appropriate.

Predicting City Mileage – EXAMPLE 3

Consider RPM as a categorical variable

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

with 3800 RPM as the base level

To build that model, 23 binary variables are needed

Predicting City Mileage – EXAMPLE 3

```
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
```

The fit has improved, R-squared 0.88, but not all RPM levels are significant.

Predicting City Mileage – EXAMPLE 3

Combining non significant levels with the base RPM level 3800

```
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
```

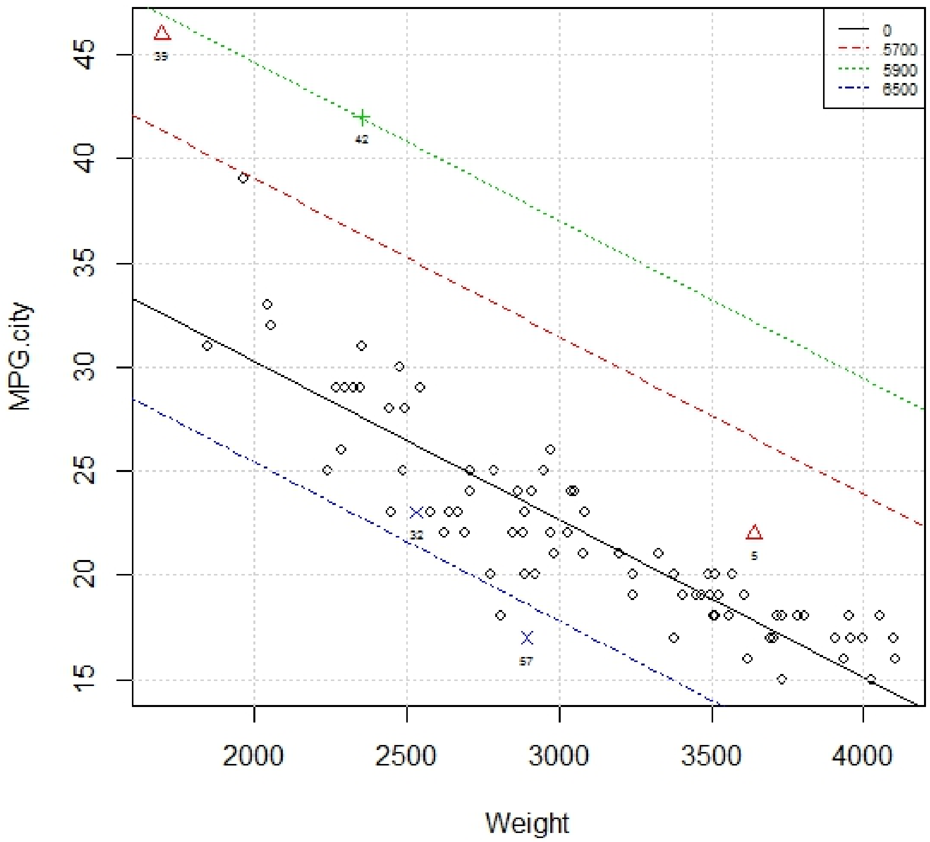
Predicting City Mileage – EXAMPLE 3

```
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
```

Model explains 84.77% of the variability of city mileage (15% over first model)

Predicting City Mileage – EXAMPLE 3



CONCLUSIONS

- Care should 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
- Spreadsheets
- Useful to identify outliers

REFERENCES

- Akritas M. (2015), Probability & Statistics with *R* for Engineers and Scientists, Pearson.
- Chapman, C., Feit E. M. (2015), *R* for Marketing Research and Analytics, Springer, New York.
- Crawley M. (2014), Statistics: An Introduction Using *R*, 2nd Edition, New York: Wiley.
- Gardener M. (2012), Statistics for Ecologists Using *R* and Excel: Data Collection, Exploration, Analysis and Presentation, Pelagic Publishing.
- Gujarati D. and Porter D. (2009), Basic Econometrics, 5th Edition, New York: McGraw Hill.
- 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/
- R Core Team (2014), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. URL www.r-project.org/.
- Kutner M.H., Neter J., Nachsteim C.J., Li W. (2004), Applied Linear Statistical Models, New York: McGraw-Hill.
- Montgomery D.C., Peck E.A., Vining G.G. (2013), Introduction to Linear Regression Analysis, 5th Edition, New York: Wiley.
- Mendenhall W. and Sincich T., (2011), A Second Course in Statistics: Regression Analysis, 7th Edition, Pearson.

USING CATEGORICAL VARIABLES IN LINEAR MODELS

Questions?

USING CATEGORICAL VARIABLES IN LINEAR MODELS

Thank you!