Data Analysis and Machine-Learning

Chapter 11.3.

*Dummy Variables – Application*



1. Interpretation

The effect of ‘credit card’ can be calculated as: 2790 + 1044 = 3834

The effect of ‘electronic check’ can be calculated as: 2709 + (-1196) = 1513

The regression coefficient of ‘credit card’ implies the difference of independent variable mean to the baseline category(‘bank transfer’). For example, the regression coefficient of credit card is the value of mean total charges of credit card minus the value of mean total charges of bank transfer. The statistical significance for such difference is denoted as p-value as follows.

Ex) assume that the independent variable is the number of harmful insects survived, and independent variables are insect sprays A, B, C, and D (with A as baseline dummy). If p<.05 for spray C, it can be stated that there is a statistical difference between the effects of sprays C and A.

2. Application

2.1. Importing dummy variables (inputting nominal values in program)

**Results**

**Linear Regression**

| Model Fit Measures | | | | | |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| **Model** | | **R** | | **R²** | |
| 1 |  | 0.515 |  | 0.266 |  |
|  | | | | | |

| Model Coefficients - TotalCharges | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |
| **Predictor** | | **Estimate** | | **SE** | | **t** | | **p** | |
| Intercept ᵃ |  | 2790 |  | 655 |  | 4.26 |  | < .001 |  |
| PaymentMethod: |  |  |  |  |  |  |  |  |  |
| Credit card (automatic) – Bank transfer (automatic) |  | 1044 |  | 858 |  | 1.22 |  | 0.230 |  |
| Electronic check – Bank transfer (automatic) |  | -1196 |  | 858 |  | -1.39 |  | 0.170 |  |
| Mailed check – Bank transfer (automatic) |  | -2073 |  | 905 |  | -2.29 |  | 0.027 |  |
| ᵃ Represents reference level | | | | | | | | | |

2.2. Results via directly generating dummy variables

Since there are four categories, three variables should be generated.

For example, when ‘bank transfer’ is set as baseline category:

Bank transfer would be: 0 0 0,

Electronic check: 1 0 0,

Mailed check: 0 1 0,

Credit card: 0 0 1

Ex)



As follows, the regression result, of course, would be the same:

**Linear Regression**

| Model Fit Measures | | | | | |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| **Model** | | **R** | | **R²** | |
| 1 |  | 0.515 |  | 0.266 |  |
|  | | | | | |

| Model Coefficients - TotalCharges | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |
| **Predictor** | | **Estimate** | | **SE** | | **t** | | **p** | |
| Intercept |  | 2790 |  | 655 |  | 4.26 |  | < .001 |  |
| Credit Card (Automatic) |  | 1044 |  | 858 |  | 1.22 |  | 0.230 |  |
| Electronic Check |  | -1196 |  | 858 |  | -1.39 |  | 0.170 |  |
| Mailed Check |  | -2073 |  | 905 |  | -2.29 |  | 0.027 |  |
|  | | | | | | | | | |