Regression with Qualitative and Quantitative Variables – Solutions
STAT-UB.0103 – Statistics for Business Control and Regression Models

Multiple Regression with Qualitative Predictors (Review)

1. We asked 46 NYU students how much time they spend on social media, and what their primary computer is (Mac or PC). We are going to use regression to find out if one type of computer associated is with more social media usage. We have the response variable

   $$\text{Social} = \text{amount of time (in minutes per week) using social media}$$

   We would like to use "OS" as a predictor variable, which is a categorical (qualitative) variable taking values in the set $\{\text{Mac}, \text{PC}\}$.

   (a) How can we encode the OS qualitative variable in terms of one or more quantitative variables?

   Solution: Define two dummy variables for OS:

   $$\text{PC} = \begin{cases} 1 & \text{if OS} = \text{PC} \\ 0 & \text{otherwise;} \end{cases}$$

   $$\text{Mac} = \begin{cases} 1 & \text{if OS} = \text{Mac} \\ 0 & \text{otherwise.} \end{cases}$$

   (b) Give a model that relates OS to Social media usage, using an intercept term and a dummy variable for "PC".

   Solution:

   $$\text{Social} = \beta_0 + \beta_1 \text{PC} + \varepsilon$$

   (c) What is the interpretation of β_0 and β_1?

   Solution: For the model $\text{Social} = \beta_0 + \beta_1 \text{PC} + \varepsilon$ The coefficient β_0 is equal to the mean social usage for Mac users. The coefficient β_1 is the difference in mean social usage between PC and Mac users.

2. We use the same data, but now we are interested in whether or not texting behavior differs by cell phone type (Blackberry, iPhone, other smart phone, or standard cell phone).

   (a) Introduce dummy variables to encode cell phone type.

Solution: We can encode cell phone type using four dummy variables

$$\text{Blackberry} = \begin{cases} 1 & \text{if Cell = Blackberry} \\ 0 & \text{otherwise;} \end{cases}$$

$$\text{iPhone} = \begin{cases} 1 & \text{if Cell = iPhone} \\ 0 & \text{otherwise;} \end{cases}$$

$$\text{Other} = \begin{cases} 1 & \text{if Cell = Other smart phone} \\ 0 & \text{otherwise;} \end{cases}$$

$$\text{Standard} = \begin{cases} 1 & \text{if Cell = Standard cell phone} \\ 0 & \text{otherwise.} \end{cases}$$

   (b) Using the variables you defined in part (a), devise a regression model which explains text usage in terms of cell phone type.

Solution: We can choose to use any of the categories as the baseline. For example, if we choose "Standard" as the baseline, then the model is

$$\text{Text} = \beta_0 + \beta_1 \text{Blackberry} + \beta_2 \text{iPhone} + \beta_3 \text{Other} + \varepsilon.$$

Different choices of the baseline category give different models (all are valid).

   (c) What is the interpretation of β_0, the intercept?

Solution: The coefficient β_0 is the mean value of Text for the baseline category (Standard cell phone, in our case).

   (d) What are the interpretations of the other coefficients in your model?

Solution: We first note that the mean value of Text for Blackberry owners is $\beta_0 + \beta_1$. Thus, β_1 is the difference in the mean value of Text between Blackberry owners and Standard cell phone owners. The meanings of β_2 and β_3 can be similarly derived.

3. We fit a model that explains Text in terms of cell phone type using dummy variables for cell phone type.

```
Analysis of Variance

Source             DF    Adj SS  Adj MS  F-Value  P-Value
Regression          3   1025437  341812     0.57    0.640
  Cell_Blackberry   1     19802   19802     0.03    0.857
  Cell_iPhone       1    584505  584505     0.97    0.330
  Cell_Smartphone   1     18678   18678     0.03    0.861
Error              42  25299274  602364
Total              45  26324711

Model Summary

      S   R-sq  R-sq(adj)  R-sq(pred)
776.121  3.90%      0.00%       0.00%

Coefficients

Term             Coef  SE Coef  T-Value  P-Value   VIF
Constant          132      317     0.42    0.680
Cell_Blackberry    91      501     0.18    0.857  1.52
Cell_iPhone       349      354     0.99    0.330  2.39
Cell_Smartphone    68      388     0.18    0.861  2.22

Regression Equation

Text = 132 + 91 Cell_Blackberry + 349 Cell_iPhone + 68Cell_Smartphone
```

(a) What is the estimated mean Text usage for people without smart phones?

Solution: $\hat{\beta}_0 = 132$.

(b) What is the estimated mean Text usage for people with iPhones?

Solution: $\hat{\beta}_0 + \hat{\beta}_2 = 132 + 349 = 481$.

(c) Is there statistically significant evidence that people with iPhones exhibit different texting behavior (volume) than people without smart phones?

Solution: We note that β_2 is equal to the difference in the mean value of Text between people with iPhones and people without smart phones. This question asks us to test the hypotheses

$$H_0 : \beta_2 = 0 \quad \text{(no difference in means)}$$
$$H_a : \beta_2 \neq 0$$

We use a t test on the coefficient of iPhone; the p-values is 0.330. Since this is above .05, there is not significant evidence of a difference (we do not reject H_0).

(d) Is cell phone type useful for predicting Text?

Solution: For this question, we are asked to test the hypotheses

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \quad \text{(cell phone type is useless for predicting Text)}$$
$$H_a : \beta_j \neq 0 \text{ for some } j = 1, 2, 3$$

We use an F test for this. The p-value is 0.640, which is above .05, so we do not reject the null. There is not significant evidence that cell phone type is useful for predicting Text.

Multiple Regression with Qualitative and Quantitative Predictors

4. Suppose we want to explain Social (minutes per week) in terms of OS (PC or Mac) and Email (minutes per week). Here is the regression output:

```
Analysis of Variance

Source         DF   Adj SS  Adj MS  F-Value  P-Value
Regression      2   390597  195299     2.47    0.096
  OS_PC         1   293693  293693     3.72    0.060
  Email         1   190702  190702     2.42    0.127
Error          43  3394150   78934
  Lack-of-Fit  29  2762459   95257     2.11    0.071
  Pure Error   14   631692   45121
Total          45  3784748

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
280.951  10.32%      6.15%       0.64%

Coefficients

Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant   249.0     63.6     3.92    0.000
OS_PC     -165.7     85.9    -1.93    0.060  1.07
Email      0.729    0.469     1.55    0.127  1.07

Regression Equation

Social = 249.0 - 165.7 OS_PC + 0.729 Email
```

(a) Interpret the estimated regression coefficients in the context of the model.

> **Solution:** In the context of the estimated regression model, $\hat{\beta}_2 = 0.729$ gives the mean increase in Social associated with holding OS constant and increasing Email by one minuted per week.
> $\hat{\beta}_1 = -165.7$ gives the difference in the mean value of Social between a PC user and a Mac user with identical values of Email.
> $\hat{\beta}_0$ does not have an interpretation. The estimated coefficient $\hat{\beta}_0$ would be the mean value of Social for Mac users who don't communicate via Email. Since everybody in the population communicates via email, these values have no direct interpretation.

(b) Interpret the p-value for the coefficient of PC.

> **Solution:** After adjusting for Email, there not a statistically significant difference between mean Text usage for PC and Mac users at level 5%. However, there is a

significant difference at level 10%.

(c) Interpret the p-value for the coefficient of Email.

Solution: Email is not useful for explaining Social usage beyond what is explained by OS type.

(d) Interpret the p-value for the ANOVA F test.

Solution: The model is not useful for explaining Social usage.

(e) What assumptions to the various regression hypothesis tests rely on?

Solution: That the mean of ε is 0; that the standard deviation of ε is constant; that ε is normally distributed; and that the values of ε are independent of each other.