Regression with Interactions – Solutions
STAT-UB.0103 – Statistics for Business Control and Regression Models

Regression with Qualitative and Quantitative Predictors

1. Suppose we want to investigate the relationship between starting salary after college and GPA for undergratues in New York City. We have the following variables:

   Response: Salary
   Predictor 1: School (Baruch, Columbia, or NYU)
   Predictor 2: GPA

   We encode School using three dummy variables:

$$\text{Baruch} = \begin{cases} 1 & \text{if School is "Baruch"} \\ 0 & \text{otherwise;} \end{cases}$$

$$\text{Columbia} = \begin{cases} 1 & \text{if School is "Columbia"} \\ 0 & \text{otherwise;} \end{cases}$$

$$\text{NYU} = \begin{cases} 1 & \text{if School is "NYU"} \\ 0 & \text{otherwise.} \end{cases}$$

   We choose to use NYU as the baseline level for School, and we use the model

$$\text{Salary} = \beta_0 + \beta_1 \cdot \text{Baruch} + \beta_2 \cdot \text{Columbia} + \beta_3 \cdot \text{GPA} + \varepsilon$$

   (a) What is the interpretation of β_3?

   Solution: In a regression model with School and GPA, if we hold School constant and increase GPA by 1 unit, then mean salary increases by β_3 units.

   (b) What are the interpretations of β_1 and β_2?

   Solution: In a regression model with School and GPA, holding GPA constant

   - β_1 is the difference in mean salary between Baruch and NYU;
   - β_2 is the difference in mean salary between Columbia and NYU.

   (c) What is the meaning of β_0. Is this interpretable?

   Solution: The coefficient β_0 is the mean salary for NYU graduates with GPA $= 0$. This is not interpretable since there are no NYU graduates with GPA $= 0$.

   (d) What is the null hypothesis for the t test on β_1?

   Solution: After adjusting for GPA, mean salary is the same for Baruch and NYU.

   (e) What is the null hypothesis for the t test on β_2?

Solution: After adjusting for GPA, mean salary is the same for Columbia and NYU.

(f) What is the null hypothesis for the t test on β_3?

Solution: GPA does not help explain Salary beyond what is explained by School.

(g) What is the null hypothesis for the ANOVA F test?

Solution: The model explaining Salary in terms of School and GPA is not useful.

(h) All of the t tests involve comparisons with NYU. What should we do if we want to compare Baruch and Columbia?

Solution: Fit a new regression model, using either Baruch or Columbia as the baseline category.

2. Suppose we want to explain Text (minutes per week) in terms of Cell Type (Blackberry, iPhone, Other Smartphone, Standard Cell) and Audio (minutes per week). Here is the regression output:

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	1446349	361587	0.60	0.668
Cell_Blackberry	1	1353	1353	0.00	0.963
Cell_iPhone	1	415734	415734	0.69	0.413
Cell_Smartphone	1	11342	11342	0.02	0.892
Audio	1	420911	420911	0.69	0.410
Error	41	24878363	606789		
Lack-of-Fit	27	23751716	879693	10.93	0.000
Pure Error	14	1126647	80475		
Total	45	26324711			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
778.967	5.49%	0.00%	0.00%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	90	322	0.28	0.781	
Cell_Blackberry	24	509	0.05	0.963	1.56
Cell_iPhone	299	361	0.83	0.413	2.46
Cell_Smartphone	53	390	0.14	0.892	2.22
Audio	0.801	0.962	0.83	0.410	1.05

Regression Equation

Text = 90 + 24 Cell_Blackberry + 299 Cell_iPhone + 53 Cell_Smartphone + 0.801 Audio

(a) What is the estimated mean of the text usage for all NYU students with a Blackberry phones who communicate via audio chat for 100 minutes per week?

Solution:

$$90 + 24 \cdot 1 + 299 \cdot 0 + 53 \cdot 1 + 0.801 \cdot 100 = 194.1$$

(b) What is the estimated standard deviation of the text usage for all NYU students with Blackberry phones who communicate via audio chat for 100 minutes per week?

Solution:

$$s = 778.967$$

(Note: this does not depend on the values of the predictors.)

(c) What is the interpretation of the p-value for Audio?

Solution: Audio is not useful for explaining Text beyond what is explained by Cell Phone Type.

(d) What is the interpretation of the p-value for iPhone?

Solution: After adjusting for Audio, there is not a significant difference for mean Text between iPhone users and Standard Cell users.

(e) What is the interpretation of the ANOVA F test?

Solution: The model is not useful for explaining Text.

Regression with Interactions

3. Recall the model from Problem 1: If we want the affect of GPA to depend on the school, then we can include interactions between School and GPA in the model:

$$\text{Salary} = \beta_0 + \beta_1 \cdot \text{Baruch} + \beta_2 \cdot \text{Columbia} + \beta_3 \cdot \text{GPA} + \beta_4 \cdot \text{Baruch} \cdot \text{GPA} + \beta_5 \cdot \text{Columbia} \cdot \text{GPA} + \varepsilon$$

   (a) What is the relationship between Salary and GPA for NYU graduates?

   Solution:
   $$\text{Salary} = \beta_0 + \beta_3 \cdot \text{GPA} + \varepsilon.$$

   (b) What is the relationship between Salary and GPA for Baruch graduates?

   Solution:
   $$\text{Salary} = (\beta_0 + \beta_1) + (\beta_3 + \beta_4) \cdot \text{GPA} + \varepsilon.$$

   (c) What is the relationship between Salary and GPA for Columbia graduates?

   Solution:
   $$\text{Salary} = (\beta_0 + \beta_2) + (\beta_3 + \beta_5) \cdot \text{GPA} + \varepsilon.$$

   (d) What is the null hypothesis for the test on β_4?

   Solution: Increasing GPA by one unit has the same affect on mean salary at NYU and Baruch.

   (e) What is the null hypothesis for the test on β_5?

   Solution: Increasing GPA by one unit has the same affect on mean salary at NYU and Columbia.

4. We fit a regression model to the class survey data using Text as the response, with Cell Type and Audio as predictors. We also included an interaction between Cell Type and Audio. Here is the resulting fit:

```
Analysis of Variance

Source                    DF    Adj SS  Adj MS  F-Value  P-Value
Regression                 7   1627178  232454     0.36    0.921
  Cell_Blackberry          1     86634   86634     0.13    0.717
  Cell_iPhone              1    217754  217754     0.34    0.566
  Cell_Smartphone          1     46144   46144     0.07    0.791
  Audio                    1     41481   41481     0.06    0.802
  Cell_Blackberry*Audio    1     64270   64270     0.10    0.755
  Cell_iPhone*Audio        1     12997   12997     0.02    0.888
  Cell_Smartphone*Audio    1     31726   31726     0.05    0.826
Error                     38  24697534  649935
  Lack-of-Fit             24  23570887  982120    12.20    0.000
  Pure Error              14   1126647   80475
Total                     45  26324711

Model Summary

      S   R-sq  R-sq(adj)  R-sq(pred)
806.186  6.18%      0.00%       0.00%

Coefficients

Term                     Coef  SE Coef  T-Value  P-Value    VIF
Constant                   13      573     0.02    0.982
Cell_Blackberry           318      871     0.37    0.717   4.27
Cell_iPhone               353      610     0.58    0.566   6.56
Cell_Smartphone           174      653     0.27    0.791   5.81
Audio                    2.29     9.08     0.25    0.802  87.44
Cell_Blackberry*Audio   -3.10     9.86    -0.31    0.755  16.55
Cell_iPhone*Audio       -1.29     9.15    -0.14    0.888  87.47
Cell_Smartphone*Audio   -2.11     9.55    -0.22    0.826  16.63

Regression Equation

Text = 13 + 318 Cell_Blackberry + 353 Cell_iPhone + 174 Cell_Smartphone + 2.29 Audio
       - 3.10 Cell_Blackberry*Audio - 1.29 Cell_iPhone*Audio - 2.11 Cell_Smartphone*Audio
```

(a) What is the estimated mean of the text usage for all NYU students with Blackberry phones who communicate via audio chat for 100 minutes per week?

Solution:

$$13 + 318 + 2.29 \cdot 100 - 3.10 \cdot 100 = 250.$$

(b) What is the estimated mean of the text usage for all NYU students with standard cell phones who communicate via audio chat for 100 minutes per week?

Solution:

$$13 + 2.29 \cdot 100 = 242$$

(c) What does the result of the t test on the coefficient of `iPhone*Audio` tell us?

Solution: The p-values is 0.888, so at level 5%, we do not reject the null hypothesis that increasing Audio by 1 minute per week has the same effect on mean Text usage for iPhone and standard cell phone users.