**Statistics 512: Solution to Homework#11**

**Problems 1 - 3 refer to the soybean sausage dataset of Problem 20.8 (`ch21pr08.dat`).**

1. Perform the two-way ANOVA without interaction for this model. Use the results of hypothesis tests to determine whether main effects are present (significant).

   **Solution:** According to the hypothesis tests, main effects for temperature level are present, but not for humidity level ($p = 6.78 \times 10^{-5}$ for temperature and $p = 0.4326$ for humidity.)

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 5 | 204.3216667 | 40.8643333 | 37.23 | 0.0002 |
| Error | 6 | 6.5850000 | 1.0975000 | | |
| Corrected Total | 11 | 210.9066667 | | | |

| R-Square | Coeff Var | Root MSE | propcc Mean |
|---|---|---|---|
| 0.968778 | 5.533185 | 1.047616 | 18.93333 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| humidity | 2 | 2.1216667 | 1.0608333 | 0.97 | 0.4326 |
| temperature | 3 | 202.2000000 | 67.4000000 | 61.41 | <.0001 |

2. Plot the data vs. the temperature factor using three different lines for the three humidity levels. Based on your graph, do you think that interaction is important for this problem?

   **Solution:** The graph (Figure 1) shows that the lines for the three levels of humidity do seem to interact across different levels of temperature. (The interaction is less pronounced when temperature is on the $x$-axis.) It looks like interaction may be important for this problem. (The Tukey Test for additivity, however, is not significant ($p = 0.2118$).)

3. Use the Tukey comparison to determine all significant differences in means in the main effect for temperature.

   **Solution:** Temperatures 1 and 2 are not significantly different from each other, according to the Tukey comparison. Everything other pairwise comparison is significant.

| Tukey Grouping | Mean | N | temperature |
|---|---|---|---|
| A | 24.8333 | 3 | 4 |
| B | 20.7000 | 3 | 3 |
| C | 15.3000 | 3 | 2 |
| C | | | |
| C | 14.9000 | 3 | 1 |

**For problem 4, use the case hardening assembly data described in Problem 24.6 on page 1022 of the text (`CH23PR06.DAT`).**

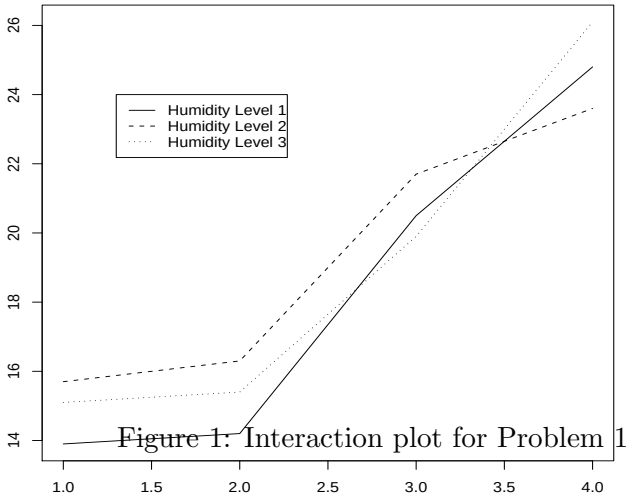


Figure 1: Interaction plot for Problem 1

4. Run the full three-way analysis of variance for these data, and check the assumptions. Summarize the results of the hypothesis tests for main and interaction effects, and your conclusions regarding the assumptions.

**Solution:** The main effects for chemical agent ($A$), temperature ($B$), and duration ($C$) are all significant, but none of the interactions are significant.

```
Dependent Variable: hardness
                                      Sum of
Source                        DF         Squares     Mean Square    F Value    Pr > F
Model                          7     4772.258333      681.751190     202.98    <.0001
Error                         16       53.740000        3.358750
Corrected Total               23     4825.998333

              R-Square     Coeff Var      Root MSE    hardness Mean
              0.988864      3.056605      1.832689         59.95833

Source                        DF       Type I SS     Mean Square    F Value    Pr > F
A                              1      788.906667      788.906667     234.88    <.0001
B                              1     1539.201667     1539.201667     458.27    <.0001
A*B                            1        0.240000        0.240000       0.07    0.7926
C                              1     2440.166667     2440.166667     726.51    <.0001
A*C                            1        0.201667        0.201667       0.06    0.8095
B*C                            1        2.940000        2.940000       0.88    0.3634
A*B*C                          1        0.601667        0.601667       0.18    0.6778
```

The variance appears to be fairly constant, though in the residual plot show that the variance in the higher level appears somewhat smaller than the lower level (Figure 2). This is due to the presence of two potential outliers.

A Box-Cox analysis suggests the squaring transformation. The result of such a transformation is that all but the three-way interaction ($A \times B \times C$) are significant.

(All diagnostic plots in this case are very much improved.)

Dependent Variable: hardnesssq

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 7 | 68976850.93 | 9853835.85 | 278.17 | <.0001 |
| Error | 16 | 566772.49 | 35423.28 | | |
| Corrected Total | 23 | 69543623.42 | | | |

| R-Square | Coeff Var | Root MSE | hardnesssq Mean |
|---|---|---|---|
| 0.991850 | 4.958022 | 188.2107 | 3796.085 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| A | 1 | 11433218.22 | 11433218.22 | 322.76 | <.0001 |
| B | 1 | 21814783.54 | 21814783.54 | 615.83 | <.0001 |
| A*B | 1 | 247063.22 | 247063.22 | 6.97 | 0.0178 |
| C | 1 | 34754721.51 | 34754721.51 | 981.13 | <.0001 |
| A*C | 1 | 377102.94 | 377102.94 | 10.65 | 0.0049 |
| B*C | 1 | 339773.57 | 339773.57 | 9.59 | 0.0069 |
| A*B*C | 1 | 10187.94 | 10187.94 | 0.29 | 0.5991 |

The normality assumption appears fine (both in the original and squared data), with a fairly linear qqplot and bell-shaped histogram. See Figure 3.

**Problems 5 - 7 refer to the hay fever relief problem 19.14 on page 868 (CH19PR14.DAT). The two active ingredients occur in the following quantities in the study:**

| | Quantity (in milligrams) | |
|---|---|---|
| Factor Level | $X_1$ (ingredient 1) | $X_2$ (ingredient 2) |
| Low | 5.0 | 7.5 |
| Medium | 10.0 | 10.0 |
| High | 15.0 | 12.5 |

5. Treating the quantities of each ingredient as quantitative variables, analyze these data using linear regression. Include linear and centered quadratic terms for each predictor and the product of the centered linear terms:

$$Y_{i,j,k} = \beta_0 + \beta_1 x_{i,j,k,1} + \beta_2 x_{i,j,k,2} + \beta_3 x_{i,j,k,1}^2 + \beta_4 x_{i,j,k,2}^2 + \beta_5 x_{i,j,k,1} x_{i,j,k,2} + \epsilon_{i,j,k}.$$

Summarize the results of this analysis.

**Solution:** The levels for factor $A$ are 5, 10, and 15, and the levels for factor $B$ are 7.5, 10, and 12.5. With regression we find that the linear, quadratic, and interaction terms are all significant in the model. For this model, $R^2 = 0.9886$, indicating an excellent fit. The regression coefficients are positive for the linear terms, and negative for the quadratic terms. This indicates that there is an increase in relief with increasing amounts of the ingredients, but that the relief increase levels off with the higher amounts. The coefficient of the interaction term is positive, suggesting that the relief increases with increased levels of both $A$ and $B$.

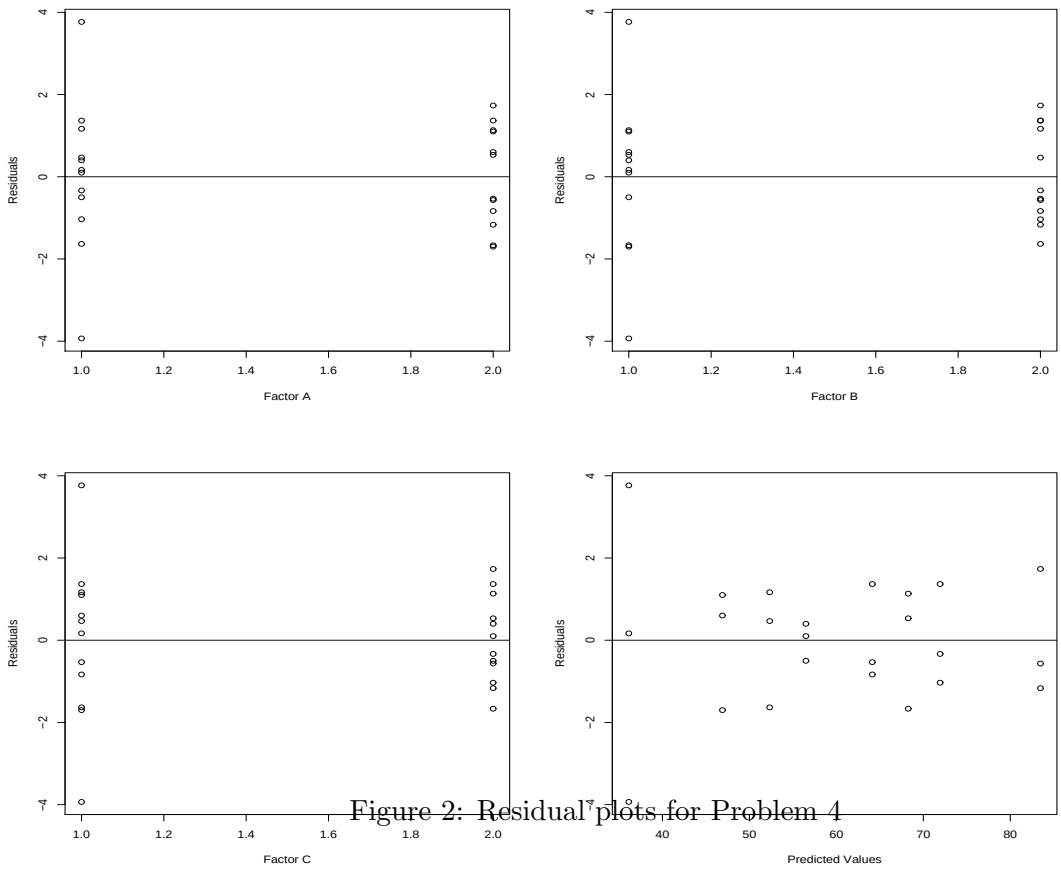


Figure 2: Residual plots for Problem 4

The REG Procedure
Dependent Variable: relief

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 5 | 370.46063 | 74.09213 | 520.63 | <.0001 |
| Error | 30 | 4.26937 | 0.14231 | | |
| Corrected Total | 35 | 374.73000 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.37724 | R-Square | 0.9886 |
| Dependent Mean | 7.18333 | Adj R-Sq | 0.9867 |
| Coeff Var | 5.25165 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | -6.06667 | 0.37197 | -16.31 | <.0001 |
| amt1 | 1 | 0.59500 | 0.01540 | 38.63 | <.0001 |
| amt2 | 1 | 0.87000 | 0.03080 | 28.25 | <.0001 |
| amt1sq | 1 | -0.03900 | 0.00534 | -7.31 | <.0001 |

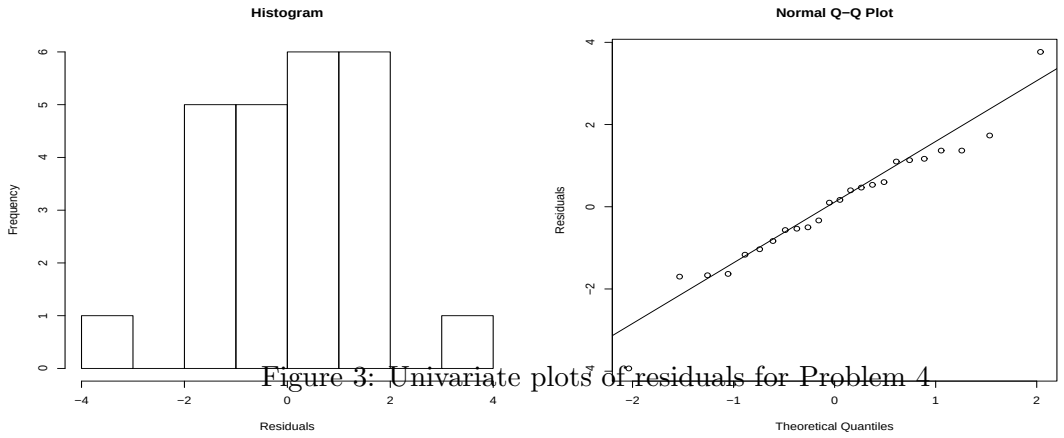


Figure 3: Univariate plots of residuals for Problem 4.

| | | | | | |
|---|---|---|---|---|---|
| amt2sq | 1 | -0.18000 | 0.02134 | -8.43 | <.0001 |
| amt12 | 1 | 0.10350 | 0.00754 | 13.72 | <.0001 |

6. Check the assumptions of the regression model used in Problem 5.

**Solution:** Scatterplots: Since we have significant quadratic terms we expect to see some deviations from linearity vs. the first order terms (Figure 4). Since the quadratic terms have only two $X$ values each, it is not really helpful to check for linearity with those terms (Figures 5 and 6). There appear to be some problems with constant variance.

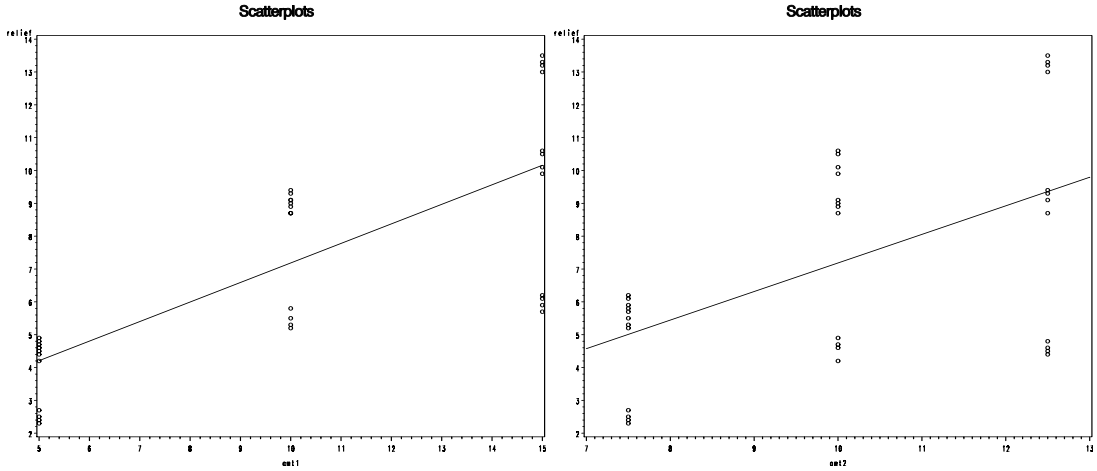


Figure 4: Scatterplots for Problem 6

Residual Plots: The residual plots (Figures 7, 8, and 9) show that there may be some problems with the constant variance assumption. The variance appears smaller for low levels of the ingredients, and there are some strange patterns on the residual vs. predictor and residual vs. interaction plots, which indicate a poor fit (linearity problems).

Normality: The qqplot (Figure 10) shows a fairly good fit to normality, with slight

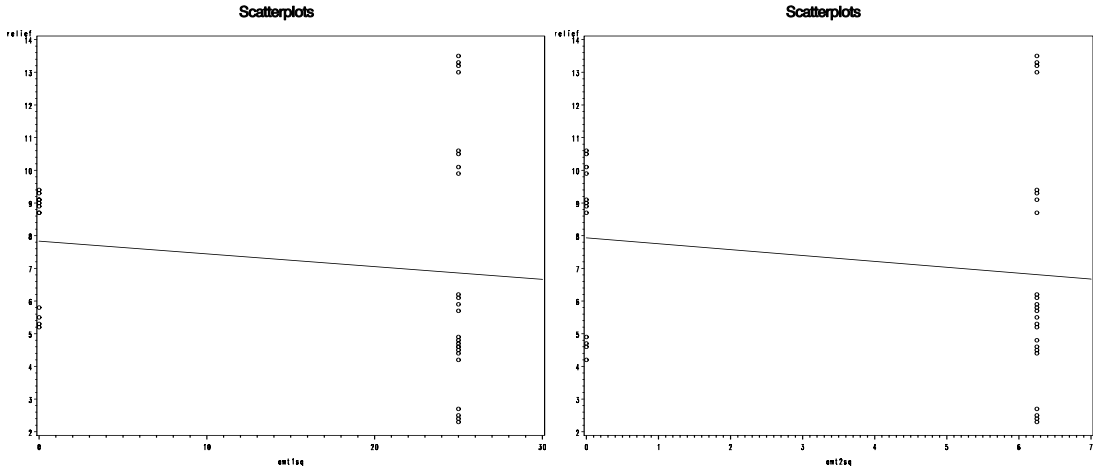


Figure 5: Scatterplots for Problem 6

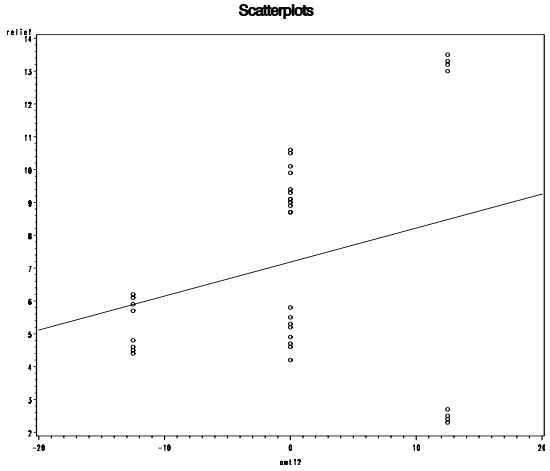


Figure 6: Scatterplot for Problem 6

deviations at the tails.

7. Give a discussion comparing the two-way ANOVA model and the regression (Problem 5) approaches to this analysis. Include a comparison of the values of $R^2$ for the two analyses and the conclusions drawn concerning interactions.

   **Solution:** Both models showed that the two ingredients and their interactions were highly significant. In addition, both models provided a good fit to the data with $R^2 = 0.995664$ for the ANOVA model and $R^2 = 0.9886$ for the regression model. The fit is slightly better for the ANOVA model, which indicates that there is some behavior that is not captured by the linear and quadratic terms in the regression model. However, this is a very tiny difference, indicating that the regression model also fits very well. The ANOVA model is simpler and easier to understand than the regression model which involves quadratic terms. Also, the assumptions of constant variance and normality seem to be better satisfied by the ANOVA model. The ANOVA model has four degrees of freedom (four parameters) for modeling the interaction, while the regression model only has one. As we saw in Problem

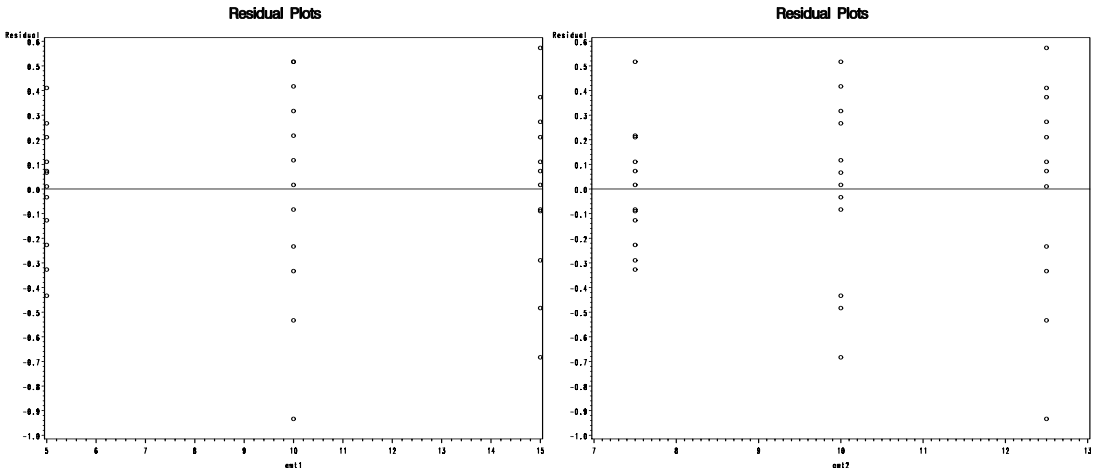


Figure 7: Residual Plots for Problem 6

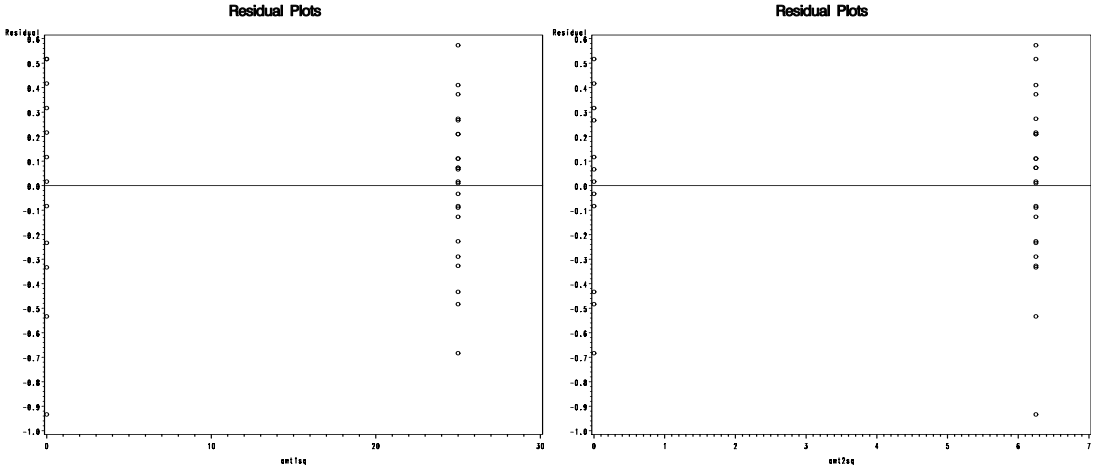


Figure 8: Residual Plots for Problem 6

Set 9, the interaction is of a specific form where it appears primarily at the high levels of both ingredients. Since the regression model only allows one parameter for interaction, it must necessarily be an "average" interaction, as opposed to the more specific four-parameter interaction of ANOVA, which treats the "high-high" combination separately. Thus the ANOVA model seems to be preferable.

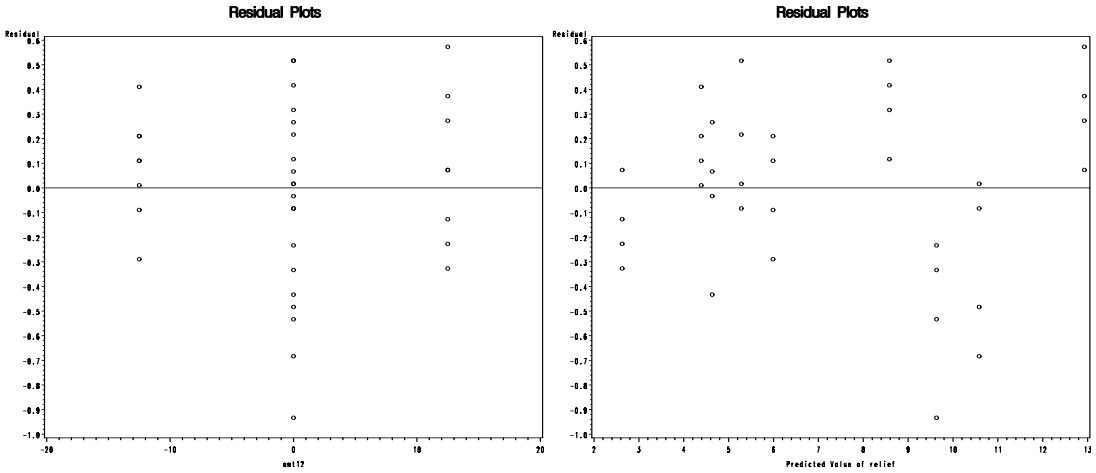


Figure 9: Residual Plots for Problem 6

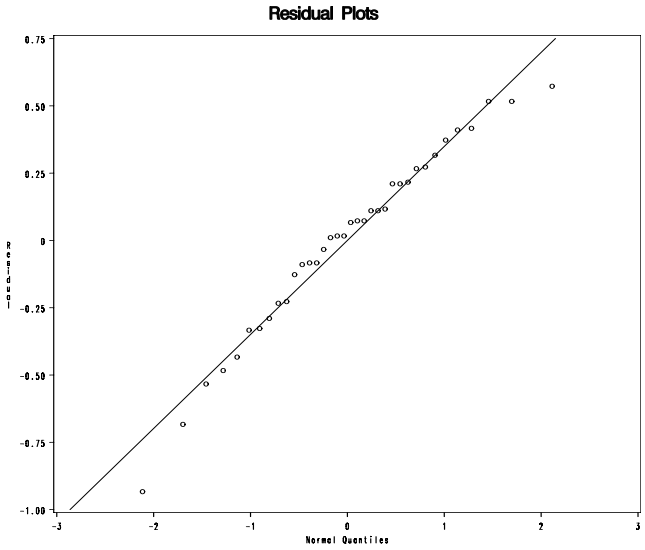


Figure 10: Qqplot for Problem 6