Statistics 512: Solution to HW#9

1. For this problem, the idea is to demonstrate the similarity between regression with dummy variables and ANOVA. To do this run the SAS code `stat512hw9.sas`.

   (a) Compare the ANOVA table and parameter results from the GLM analysis and Parameterization #1. What do the coefficients associated with X_1 and X_2 (i.e. b_1 and b_2) estimate in terms of treatment means? What constraint system does this parameterization correspond to?

   Solution: The ANOVA tables from the two analyses are identical.

   The parameter estimates $\hat{\mu}$, $\hat{\tau}_1$, and $\hat{\tau}_2$ from the `GLM` analysis are the same as the parameter estimates b_0, b_1, and b_2 from the regression model. The fourth parameter estimate $\hat{\tau}_3$ which is set to zero in the `GLM` analysis does not appear in the regression analysis, since it is assumed equal to zero in that model. Parameterization #1 sets

   $$\begin{aligned} \mu_1 &= \beta_0 + \beta_1 = \mu + \tau_1, \\ \mu_2 &= \beta_0 + \beta_2 = \mu + \tau_2, \text{ and} \\ \mu_3 &= \beta_0 = \mu. \end{aligned}$$

   This system of equations can be solved for the β's to obtain

   $$\begin{aligned} \beta_0 &= \mu = \mu_3, \\ \beta_1 &= \tau_1 = \mu_1 - \mu_3, \text{ and} \\ \beta_2 &= \tau_2 = \mu_2 - \mu_3. \end{aligned}$$

   The coefficient associated with X_1 is $b_1 = \hat{\tau}_1$, which thus is an estimate for $\mu_1 - \mu_3$. The coefficient associated with X_2 is $b_1 = \hat{\tau}_2$, which thus is an estimate for $\mu_2 - \mu_3$.

```
                                The GLM Procedure
Dependent Variable: response

                                          Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F
Model                        2     38.88888889     19.44444444      15.91    0.0040
Error                        6      7.33333333      1.22222222
Corrected Total              9     46.22222222

               R-Square     Coeff Var      Root MSE    response Mean
               0.841346      4.522670      1.105542         24.44444

                                            Standard
      Parameter           Estimate             Error    t Value    Pr > |t|
      Intercept        26.66666667 B      0.63828474      41.78      <.0001
      trt       1      -5.00000000 B      0.90267093      -5.54      0.0015
      trt       2      -1.66666667 B      0.90267093      -1.85      0.1144
      trt       3       0.00000000 B       .                .          .
```

The REG Procedure
Model: MODEL1
Dependent Variable: response

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	38.88889	19.44444	15.91	0.0040
Error	6	7.33333	1.22222		
Corrected Total	8	46.22222			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	26.66667	0.63828	41.78	<.0001
x1	1	-5.00000	0.90267	-5.54	0.0015
x2	1	-1.66667	0.90267	-1.85	0.1144

(b) Compare the ANOVA table and parameter results from the GLM analysis and Parameterization #2. What do the coefficients associated with X_1 and X_2 (i.e. b_1 and b_2) estimate in terms of treatment means? What constraint system does this parameterization correspond to?

Solution:

The REG Procedure
Model: MODEL1
Dependent Variable: response

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	38.88889	19.44444	15.91	0.0040
Error	6	7.33333	1.22222		
Corrected Total	8	46.22222			

Root MSE	1.10554	R-Square	0.8413
Dependent Mean	24.44444	Adj R-Sq	0.7885
Coeff Var	4.52267		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	24.44444	0.36851	66.33	<.0001
x1	1	-2.77778	0.52116	-5.33	0.0018
x2	1	0.55556	0.52116	1.07	0.3274

The ANOVA tables from the two analyses are identical. However, the parameter estimates for the two analyses are different. The `GLM` analysis gives parameter estimates $\hat{\mu}$, $\hat{\tau}_1$, and $\hat{\tau}_2$ which estimate the parameters of the factor effects model assuming $\tau_3 = 0$. The regression analysis gives parameter estimates b_0, b_1, and b_2 which estimate the parameters of the factor effects model assuming $\sum_i \tau_i = 0$. Because the two solutions use different constraints, they are estimating different quantities.

Parameterization #2 sets

$$\mu_1 \quad = \quad \beta_0 + \beta_1 = \mu + \tau_1,$$

$$\begin{aligned}\mu_2 &= \beta_0 + \beta_2 = \mu + \tau_2, \text{ and}\\ \mu_3 &= \beta_0 - \beta_1 - \beta_2 = \mu + \tau_3.\end{aligned}$$

This system of equations can be solved for the β's to obtain

$$\begin{aligned}\beta_0 &= \mu = \frac{\mu_1 + \mu_2 + \mu_3}{3},\\ \beta_1 &= \tau_1 = \mu_1 - \mu, \text{ and}\\ \beta_2 &= \tau_2 = \mu_2 - \mu.\end{aligned}$$

In the regression analysis, the constraint $\sum_i \tau_i = 0$ implies $\beta_0 = \mu = \frac{1}{3}\sum_i \mu_i$, which is estimated by b_0; $\beta_1 = \tau_1 = \mu_1 - \mu$ which is estimated by b_1; and $\beta_2 = \tau_2 = \mu_2 - \mu$, which is estimated by b_2.

(c) Show that $b_0 + b_1$ and $b_1 - b_2$ give the same answer regardless of the parameterization. What do these two terms estimate in terms of the treatment means?

Solution: For Parametrization 1:

$$b_0 + b_1 = 26.67 - 5.0 = \mathbf{21.67};\ b_1 - b_2 = -5.0 + 1.67 = \mathbf{-3.33}.$$

For Parametrization 2:

$$b_0 + b_1 = 24.44 - 2.78 = \mathbf{21.66};\ b_1 - b_2 = -2.78 - 0.56 = \mathbf{-3.33}.$$

These are the same (except for roundoff error). The quantity $b_0 + b_1$ estimates $\mu + \tau_1 = \mu_1$, the mean for `trt = 1`. The quantity $b_1 - b_2$ estimates $\tau_1 - \tau_2 = (\mu + \tau_1) - (\mu + \tau_2) = \mu_1 - \mu_2$, the difference in means for the groups `trt = 1` and `trt = 2`.

The next three problems use the dataset from Problem 16.11 described on page 725 of KNNL, and continue the analysis begun on Homework#8.

2. Use the Tukey multiple comparison method to determine which pairs of machines differ significantly. Summarize the results.

Solution: The Tukey comparison method shows that machines 3 and 4 are significantly different from machines 1, 2, 5, and 6 (on a pairwise basis).

```
                              The GLM Procedure
                    Tukey's Studentized Range (HSD) Test for wtdev
NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher
                          Type II error rate than REGWQ.
                     Alpha                                      0.05
                     Error Degrees of Freedom                    114
                     Error Mean Square                       0.03097
                     Critical Value of Studentized Range     4.09949
                     Minimum Significant Difference           0.1613
               Means with the same letter are not significantly different.
```

Tukey Grouping	Mean	N	machine
A	0.46000	20	3
A			
A	0.36550	20	4
B	0.19050	20	2
B			
B	0.15150	20	6
B			
B	0.12500	20	5
B			
B	0.07350	20	1

3. Suppose you want to compare the average of the first two machines with the average of the last four. Use the `estimate` and `contrast` statements in `proc glm` to test the appropriate hypothesis. Report the estimated value of this contrast with its standard error; state the null and alternative hypotheses, the test statistic with degrees of freedom, the p-value and your conclusion.

   Solution: The contrast is $L = \frac{(\mu_1+\mu_2)}{2} - \frac{(\mu_3+\mu_4+\mu_5+\mu_6)}{4}$.

   The estimated value of this contrast is $\hat{L} = -0.1435$ with standard error $s\{\hat{L}\} = 0.03408$. We test the null hypothesis

   $$\mathrm{H}_0 : \frac{(\mu_1+\mu_2)}{2} - \frac{(\mu_3+\mu_4+\mu_5+\mu_6)}{4} = 0 \text{ vs } \mathrm{H}_a : \frac{(\mu_1+\mu_2)}{2} - \frac{(\mu_3+\mu_4+\mu_5+\mu_6)}{4} \neq 0.$$

   [These could also be correctly written as

   $$\mathrm{H}_0 : L = 0 \text{ vs. } \mathrm{H}_a : L \neq 0$$

   or

   $$\mathrm{H}_0 : \frac{(\mu_1+\mu_2)}{2} = \frac{(\mu_3+\mu_4+\mu_5+\mu_6)}{4} \text{ vs } \mathrm{H}_a : \frac{(\mu_1+\mu_2)}{2} \neq \frac{(\mu_3+\mu_4+\mu_5+\mu_6)}{4}.]$$

   The test statistic is either $F = 17.73$ with $(1, 114)$ df or $t = -4.21$ with 114 df, and the p-value is 5.11×10^{-5}. We reject H_0 and conclude that the mean for the first two machines is not the same as the mean for the last four machines.

```
                                The GLM Procedure
Dependent Variable: wtdev
      Contrast                 DF     Contrast SS     Mean Square    F Value    Pr > F
      prob2                     1      0.54912667      0.54912667      17.73    <.0001

                                                      Standard
          Parameter                 Estimate           Error    t Value    Pr > |t|
          prob2                  -0.14350000      0.03407905      -4.21      <.0001
```

4. Check assumptions using the residuals. Turn in the plots/output you used to check the assumptions and state your conclusions.

   Solution: The assumptions of normality and constant variance appear to be satisfied, as shown by the residual plot and qqplot (Figure 1).

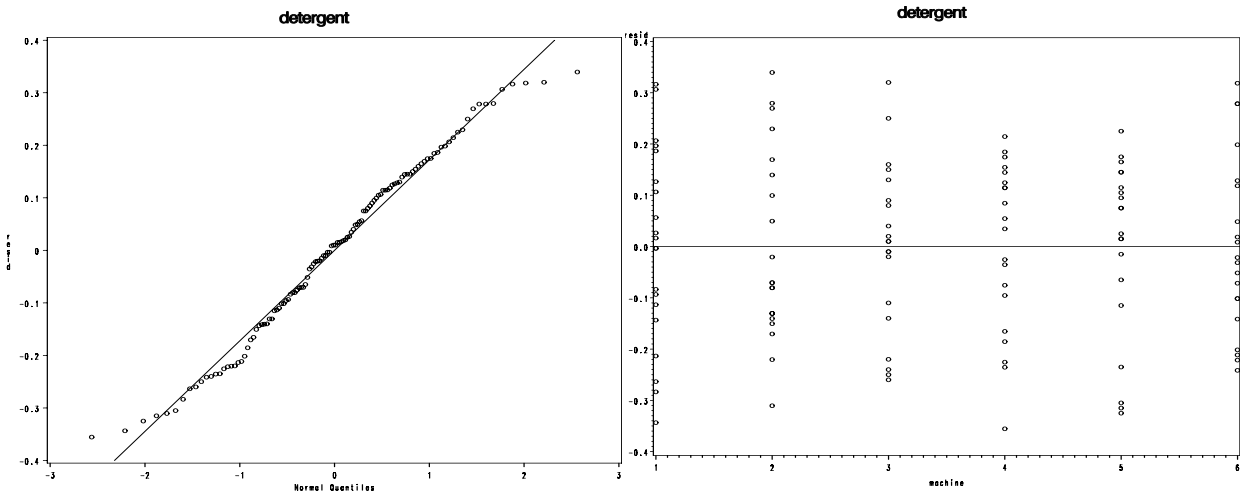

Figure 1: Diagnostic plots for Problem 4

The remaining problems use the dataset from Problem 18.15 on page 804 of KNNL.

5. KNNL 18.15 (Omit part e).) Please do not print out all 80 values for part a); it is sufficient to plot them in part b).

   Helicopter service. An operations analyst in a sheriff's department studied how frequently their emergency helicopter was used during a recent 20-day period, by time of day (shift 1: 2am - 8am; shift 2: 8am - 2pm; shift 3: 2pm - 8pm; shift 4: 8pm - 2am). Since the data are counts, the analyst was concerned about the normality and equal variances assumptions of ANOVA model (16.2).

   (a) Obtain the fitted values and residuals for ANOVA model (16.2).

   Solution: The fitted values are the means for the four levels as shown below. The residuals are shown on the plot in part 5b.

   The GLM Procedure

Level of shift	N	use Mean	use Std Dev
1	20	3.90000000	1.97084006
2	20	1.15000000	1.08942283
3	20	2.00000000	1.45095250
4	20	3.40000000	1.78885438

   (b) Prepare suitable residual plots to study whether or not the error variances are equal for the four shifts. What are your findings?

   Solution: The residual plot (Figure 2) shows that the variances are not constant. The plot of residuals vs. predictors (also Figure 2) shows that the variances tend increase with the predicted value.

   (c) Test by means of the modified Levene test whether or not the treatment error variances are equal; use $\alpha = 0.10$. What is the p-value of the test? Are your results consistent with the diagnosis in part 5b?

   Solution: The p-value of the modified Levene test is 0.1344, so we do not reject the null hypothesis of equal variances at the $\alpha = 0.10$ level; however note

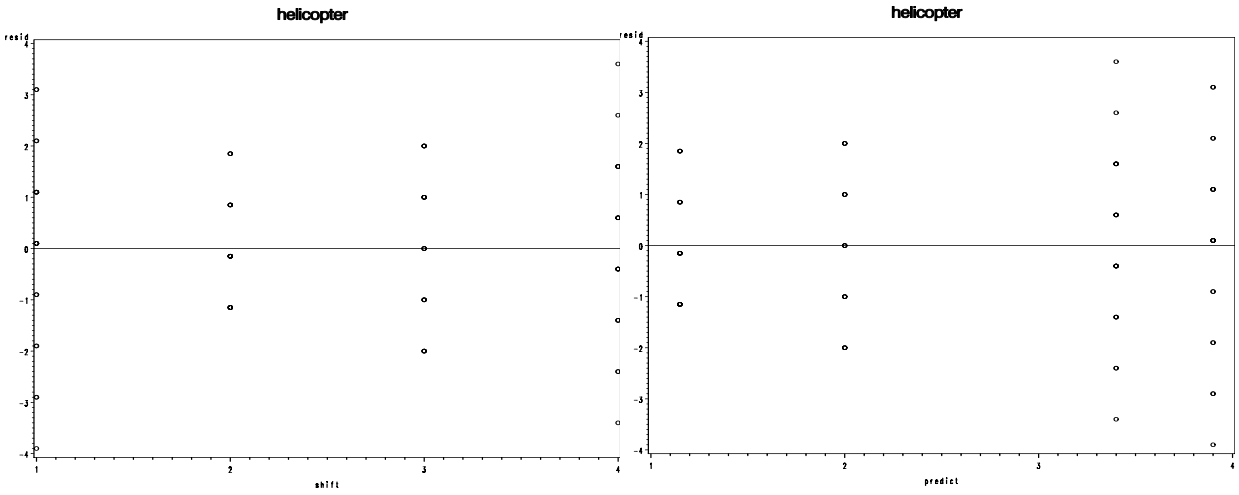

Figure 2: Residual plots for Problem 5b [only one is required]

that the p-value is fairly close to α. This failure to reject does not support the conclusion of unequal variances reached in part 5b.

The GLM Procedure
Levene's Test for Homogeneity of use Variance
ANOVA of Absolute Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
shift	3	4.8394	1.6131	1.91	0.1344
Error	76	64.0395	0.8426		

(d) For each shift, calculate $\bar{Y}_{i.}$ and s_i. Examine the three relations found in the table on page 773 and determine the transformation that is most appropriate here. What do you conclude?

Solution:

shift level	$\bar{Y}_{i.}$	s_i	$\frac{s_i^2}{\bar{Y}_{i.}}$	$\frac{s_i}{\bar{Y}_{i.}}$	$\frac{s_i}{\bar{Y}_{i.}^2}$
1	3.90	1.971	0.996	0.505	0.130
2	1.15	1.089	1.032	0.947	0.824
3	2.00	1.451	1.053	0.725	0.363
4	3.40	1.789	0.941	0.526	0.154

The ratio $\frac{s_i^2}{\bar{Y}_{i.}}$ appears the most consistent across levels. This indicates that a square root transformation may be appropriate.

Obs	shift	_TYPE_	_FREQ_	meanuse	varuse	stduse	meanuseplus1	logmeanuse	logvaruse	logstduse	logmeanuseplus1	reciprocalstduse
1	1	0	20	3.90	3.88421	1.97084	15.2100	1.36098	0.67846	0.99595	0.50534	0.12958
2	2	0	20	1.15	1.18684	1.08942	1.3225	0.13976	0.08565	1.03204	0.94732	0.82376
3	3	0	20	2.00	2.10526	1.45095	4.0000	0.69315	0.37222	1.05263	0.72548	0.36274
4	4	0	20	3.40	3.20000	1.78885	11.5600	1.22378	0.58158	0.94118	0.52613	0.15475

6. A rather simple approximation of the Box-Cox procedure is the following:

   (a) Compute the mean and standard deviation for each treatment factor level.

   (b) Take the log of both the mean and standard deviation.

   (c) Fit the regression model $\log(\sigma_i) = \beta_0 + \beta_1 \log(\mu_i) + \epsilon$ using the observed means and standard deviations as the data for i and i respectively (there are 4 "observations" in this dataset).

   (d) Set $\hat{\lambda} = 1 - b_1$ where b_1 is the estimate for β_1 obtained in (6c).

   Use the Helicopter service data to perform this approximation. What value of λ appears reasonable according to this method?

   Solution The regression model gives the estimate $b_1 = 0.47$, and so $\hat{\lambda} = 1 - 0.47 = 0.53 \approx 0.5$. This also suggests a square root transformation.

The REG Procedure
Model: MODEL1
Dependent Variable: logstduse

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.02766	0.02261	1.22	0.3458
logmeanuse	1	0.47029	0.02305	20.40	0.0024

7. Define a new response variable by adding 1 to the original response. (This will avoid 0's which mess up the log and reciprocal transformations.) Then use SAS's Box-Cox procedure to determine an appropriate transformation. `Proc transreg` can be used to perform ANOVA if we tell it `shift` is a `class` variable, as in the following:

```
proc transreg data=helicopter;
  model boxcox(usesplus1) = class(shift);
```

Solution:

The TRANSREG Procedure

Transformation Information
for BoxCox(usep1)

Lambda	R-Square	Log Like	
-3.00	0.14	-182.106	
-2.75	0.15	-165.922	
-2.50	0.15	-150.178	
-2.25	0.16	-134.942	
-2.00	0.17	-120.291	
-1.75	0.18	-106.320	
-1.50	0.20	-93.144	
-1.25	0.21	-80.897	
-1.00	0.23	-69.735	
-0.75	0.24	-59.830	
-0.50	0.26	-51.364	
-0.25	0.28	-44.517	
0.00	0.30	-39.442	
0.25	0.31	-36.250	*
0.50	+ 0.32	-34.991	<
0.75	0.32	-35.645	*
1.00	0.33	-38.129	
1.25	0.33	-42.313	
1.50	0.32	-48.035	
1.75	0.32	-55.124	
2.00	0.31	-63.413	
2.25	0.30	-72.748	
2.50	0.29	-82.989	
2.75	0.28	-94.017	
3.00	0.26	-105.728	

< - Best Lambda
* - Confidence Interval
+ - Convenient Lambda

The Box-Cox procedure suggests that $\lambda = 0.5$, i.e. a square root transformation, would be appropriate.

8. KNNL 18.16 (Omit the coefficient of correlation in part b).)

Refer to Helicopter service Problem 18.15. The analyst decided to apply the square root transformation and examine its effectiveness.

(a) Obtain the transformed response data, fit ANOVA model (16.2), and obtain the residuals.

Solution: The SAS output for the ANOVA model with the transformed data is given below.

```
                                The GLM Procedure
Dependent Variable: rootuse

                                   Sum of
Source                     DF       Squares     Mean Square    F Value    Pr > F
Model                       3    13.60854916      4.53618305      10.29    <.0001
Error                      76    33.48888826      0.44064327
Corrected Total            79    47.09743742

          R-Square     Coeff Var      Root MSE    rootuse Mean
          0.288945      46.66182      0.663810        1.422597

                                The GLM Procedure
                  Level of                -----------rootuse-----------
                  shift          N              Mean          Std Dev
                  1             20        1.87136414       0.64725840
                  2             20        0.84265033       0.68051097
                  3             20        1.22925287       0.71740569
                  4             20        1.74712038       0.60486655
```

(b) Prepare suitable plots of the residuals to study the equality of the error variances of the transformed response variable for the four shifts. Also obtain a normal probability plot. What are your findings? Does the transformation appear to have been effective?

Solution: There has been an improvement in the variances of the transformed

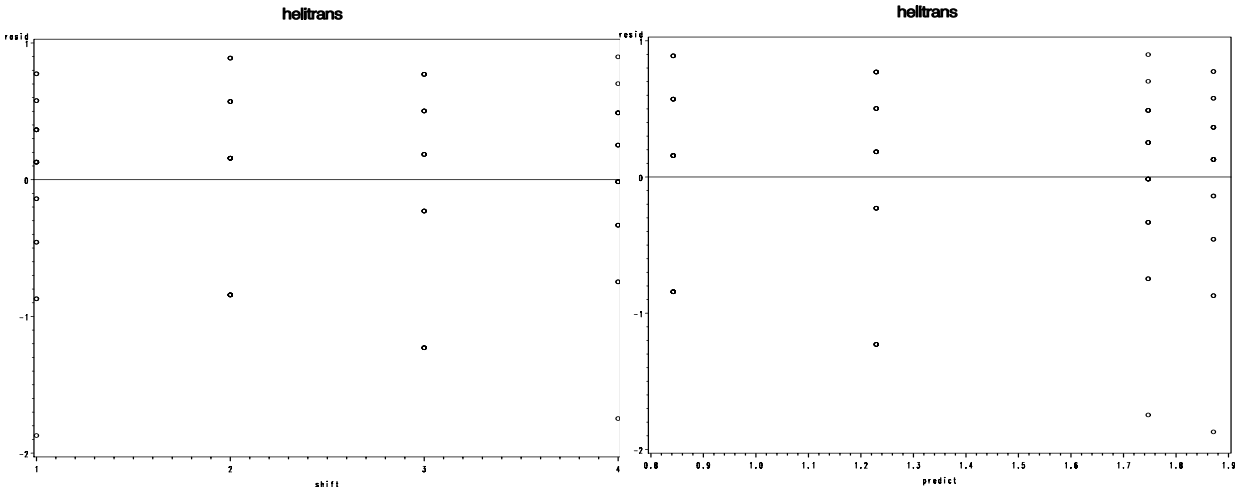

Figure 3: Residual plots for Problem 8b

variable, but there are still signs that the variance increases with the predicted value (see Figure 3). The QQ-plot (Figure 4) shows some skewness away from normality. The transformation stabilized the variance somewhat, however, it also appears to have destroyed the normality of the residuals.

(c) Test by means of the modified Levene test whether or not the treatment error variances for the transformed response variable are equal; use $\alpha = 0.10$. State the alternatives, decision rule, and conclusion. Are your findings in part (b) consistent with your conclusion here?

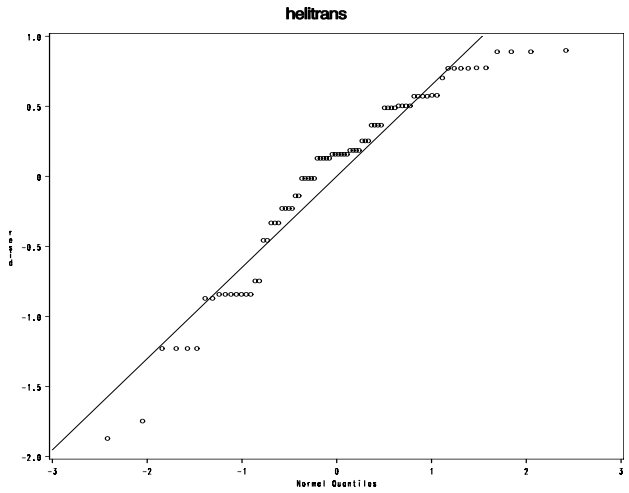

Figure 4: Qqplot for Problem 8b

Solution:

The GLM Procedure
Levene's Test for Homogeneity of rootuse Variance
ANOVA of Absolute Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
shift	3	0.3652	0.1217	0.81	0.4905
Error	76	11.3771	0.1497		

The Levene test is now very far from significant. The p-value is 0.49, which is a big increase over the 0.13 for the untransformed data. This gives some evidence that the errors are more homogeneous after the transformation. (However, the Levene test is sensitive to the normality assumption, and since we know normality is violated, we cannot really trust the results.)

9. Use the Tukey multiple comparison method for differences in means on both the untransformed and transformed Helicopter service data to determine which shifts differ significantly. Summarize and compare the results.

Solution: Using the transformed data we find that the following pairs are significantly different: 1 with 2 and 3; and 2 with 1 and 4.

Tukey's Studentized Range (HSD) Test for rootuse

Alpha	0.05
Error Degrees of Freedom	76
Error Mean Square	0.440643
Critical Value of Studentized Range	3.71485
Minimum Significant Difference	0.5514

Tukey Grouping	Mean	N	shift
A	1.8714	20	1
A			
B A	1.7471	20	4
B			
B C	1.2293	20	3
C			
C	0.8427	20	2

Using the untransformed data we find that shifts 1 and 4 differ significantly from shifts 2 and 3.

```
      Tukey's Studentized Range (HSD) Test for use
      Alpha                                   0.05
      Error Degrees of Freedom                  76
      Error Mean Square                   2.594079
      Critical Value of Studentized Range  3.71485
      Minimum Significant Difference        1.3379

Tukey Grouping          Mean      N    shift
              A        3.9000     20    1
              A
              A        3.4000     20    4

              B        2.0000     20    3
              B
              B        1.1500     20    2
```

It is interesting that in the transformed data 3 and 4 are not significantly different but in the untransformed data they are.