The SAS program I used to obtain the analyses for my answers is given below.

```
dm'log;clear;output;clear';
*************************************************************;
*** EXST7034 Homework Example 1                           ***;
*** Problem from Neter, Wasserman & Kuttner 1989, #2.18   ***;
*************************************************************;
OPTIONS LS=132 PS=256 NOCENTER NODATE NONUMBER nolabel;
filename copier 'C:\Geaghan\Current\EXST7034\Fall2005\SAS\CH01PR20.txt';
ODS HTML style=minimal rs=none
body='C:\Geaghan\Current\EXST7034\Fall2005\SAS\CH01PR20b.html' ;

Title1 'Assignment 2 : Copier maintenance example';

DATA ONE; INFILE copier MISSOVER;
      LABEL machines = 'Number of machines serviced';
      LABEL minutes = 'Minutes to service machines';
   INPUT minutes machines;
   X_number_2 = machines;
CARDS; RUN;
;

OPTIONS LS=99 PS=256;
proc univariate data=ONE; var minutes; run;
PROC REG DATA=ONE lineprinter; ID machines;
    MODEL  minutes = machines / XPX I P;
    output out=next1 p=yhat r=e;
run;
proc univariate data=next1 plot normal; var e; run;
proc univariate data=ONE plot normal; var machines; run;

OPTIONS LS=99 PS=56;
proc plot data=next1;
   plot e*machines / vref=0;
run;
OPTIONS LS=99 PS=256;

PROC GLM DATA=ONE; classes X_number_2;
    MODEL  minutes = machines X_number_2;
run;

proc transreg data=one;
   title2 'Box-Cox transformation with PROC TRANSREG';
   MODEL  BOXCOX(minutes) = identity(machines);
run;
```

Problem 2.24a) The ANOVA table from the SAS output is given below.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	76960	76960	968.66	<.0001
Error	43	3416.37702	79.45063		
Corrected Total	44	80377			

This output is in the style of Table 2.2. The columns for the d.f. and Sum of Squares are additive. The other style of table (2.3) includes rows for the correction factor and the uncorrected total SS. There are many ways to get these values. Since SAS provides an estimate of the "Dependent mean" (mean of the values of the dependent variable), I used this value (76.2666667) to calculate the correction factor. $CF = n*76.2666667^2 = 45*76.2666667^2 = 261747.2002$. The uncorrected total can be obtained from SAS PROC UNIVARIATE, the X′X matrix in PROC REG or by adding corrected total and the correction factor (UTotal = CTotal + CF = 80376.7998+261747.2002 = 342124). This information

was added to the table. The first two columns of numbers are additive in that the df and SS for Model and Error should add to the C Total, and the C Total and CF should add to the U Total.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	76960	76960	968.66	<.0001
Error	43	3416.37702	79.45063		
Corrected Total	44	80377			
Correction Factor	1	73728			
Uncorrected Total	45	342124			

Problem 2.24b) The requested F value is given in the tables above **(F = 968.66)**. The requested test is of the "linear association" (H_0: $\beta_1=0$ versus H_1: $\beta_1\neq0$). This is the same as previously done in Problem 2.5b, except that here it is to be done with an F test. The P value indicates that the relationship is significant well above the requested $\alpha = 0.10$ level (i.e. **P(>F) < 0.0001**). See additional notes about the "linear association" in the answer to the previously assigned problem 2.5b.

Problem 2.24c) The "Total" variation to be partitioned is the USSTotal = 342124 (see problem 2.24a). First the total is adjusted for the mean (i.e. the correction factor). This step is almost always done, and not usually considered in subsequent calculations. After the correction for the mean, the remaining variation is **CSSTotal = 80377** (note that this is the corrected total). When the X-variable is entered into the model, this is reduced by an amount equal to the **SSRegression = 76960**. This is then the relative reduction. The relative reduction is expressed as a proportion or percent, and is called R^2, where **R^2 = 76960 / 80377 = 0.9575 or 95.75%**. This is a relatively large value by almost any standard, and is in fact much larger than would be expected for many types of studies.

Question KNNL 3.8 modified parts a & b) The Stem and Leaf plots and Box plots are given below. There are not assumptions about the distribution of X_i, so the stem leaf plot for the independent variable is not too important.

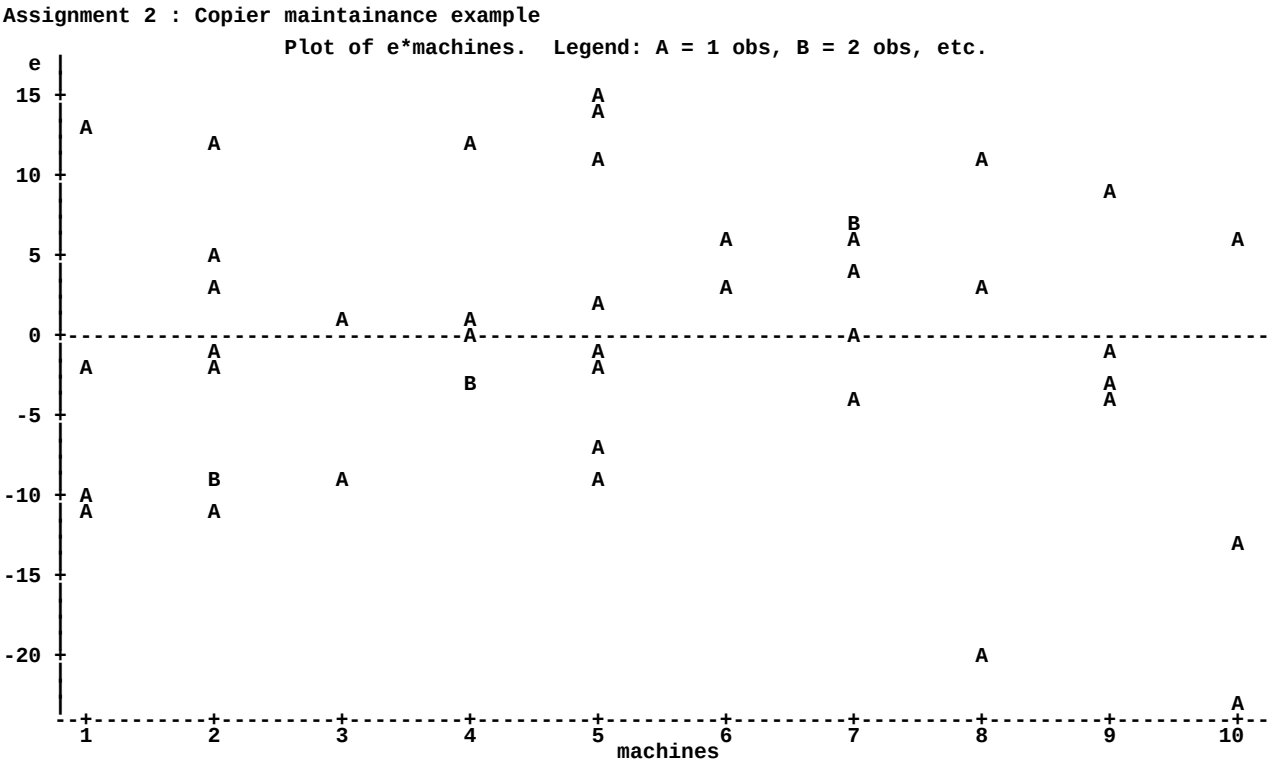
The UNIVARIATE Procedure
Variable: machines

```
Stem Leaf                     #  Boxplot
  10 000                      3     |
   9                          0     |
   9 0000                     4     |
   8                          0     |
   8 000                      3     |
   7                          0     |
   7 000000                   6  +-----+
   6                          0  |     |
   6 00                       2  |     |
   5                          0  |     |
   5 00000000                 8  *--+--*
   4                          0  |     |
   4 00000                    5  |     |
   3                          0  |     |
   3 00                       2  |     |
   2                          0  |     |
   2 00000000                 8  +-----+
   1                          0     |
   1 0000                     4     |
     ----+----+----+----+
```

The stem leaf plot for the residuals shows that the data is spread pretty well with a few values that are possible outliers. It does not follow a somewhat normal distribution, and the Shapiro-Wilk tests confirms this (P=0.4614). The boxplot indicates a possible negative skew, or a couple of potential outliers. For this rather small sample we see not clear indication of problems.

```
The UNIVARIATE Procedure
Variable:  e
   Stem Leaf                     #  Boxplot
     14 44                       2     |
     12 45                       2     |
     10 345                      3     |
      8 3                        1     |
      6 23433                    5  +-----+
      4 35                       2  |     |
      2 4534                     4  |     |
      0 3445                     4  *--+--*
     -0 6576                     4  |     |
     -2 7776655                  7  +-----+
     -4                             |
     -6 6                        1   |
     -8 5565                     4   |
    -10 555                      3   |
    -12 8                        1   |
    -14
    -16
    -18 7                        1   0
    -20
    -22 8                        1   0
        ----+----+----+----+
```

Question KNNL 3.8c) The residual plot shows that the two possible outliers indicate by the boxplot occurred at higher values of X_i. , There is no clear sign of nonhomogeniety, but there is a suggestion of possible curvature. There is not clear departure from the random pattern desired, so there is no evidence of problems from this plot.

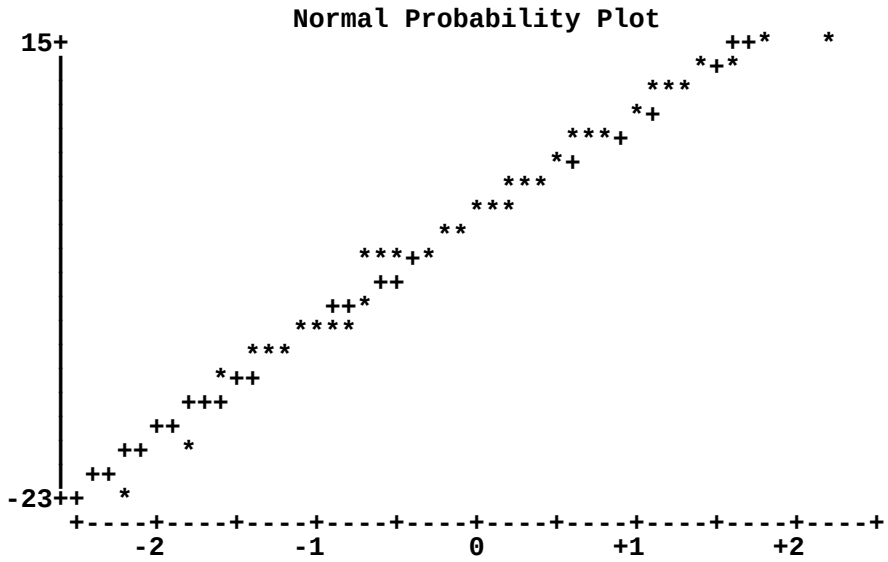


Question KNNL 3.8d) As mentioned in the answer to 3.8b, the residual do not show a significant departure from normality when judged by the Shapiro-Wilk test.

Tests for Normality

Test		--Statistic---	-----p Value------	
Shapiro-Wilk	W	0.975828	Pr < W	0.4614
Kolmogorov-Smirnov	D	0.091299	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.037232	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.278721	Pr > A-Sq	>0.2500

Question KNNL 3.8d) The normal probability plot shows a number of observations (asterisks) which do not fall on the line representing expected values (plus signs). The points depart in the middle of the range instead of the extremes, indicating that the departure is not due to outliers or skew. This is an additional indication that the residuals do not follow a normal curve.

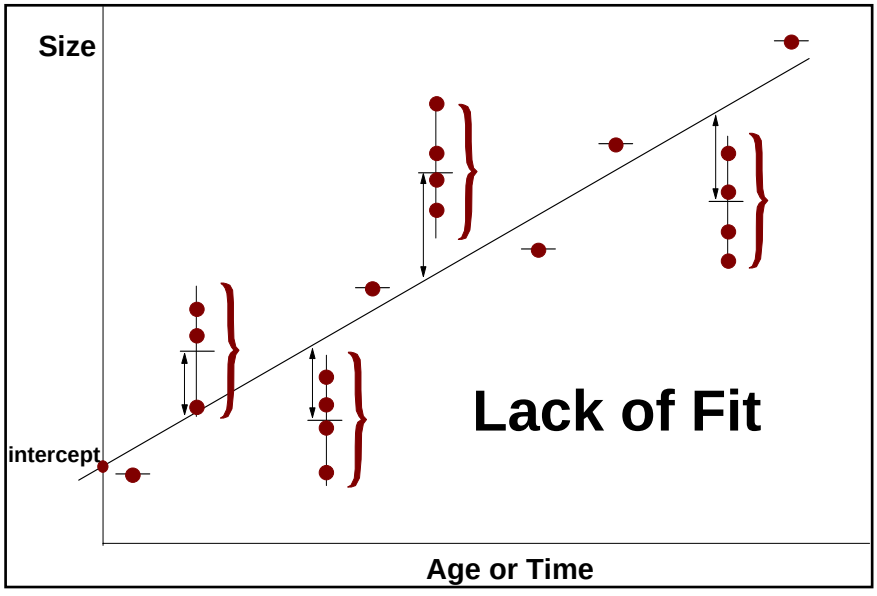


My question number 3 is answered below. I actually provide answers to question KNNL 3.13, Parts a, b and c) The usual regression error (total deviations from regression) is;

$$\sum_{i=1}^{m}\sum_{j=1}^{n_m}(Y_{ij}-\hat{Y}_i)^2$$

Which can be broken down into two parts, (1) Pure error (deviations of individual Y values from individual mean Y– values at each value of X) and (2) LOF, deviations of the means, Y–, from the regression line, Y^.). The formulas are;

$$\sum_{i=1}^{m}\sum_{j=1}^{n_m}(Y_{ij}-\bar{Y}_i)^2 - \sum_{i=1}^{m}(\bar{Y}_i-\hat{Y}_i)^2$$



Question KNNL 3.13 Part a-c)

a) The hypotheses are; $H_0: E(Y) = \beta_0 + \beta_1 X_i$ versus $H_1: E(Y) \neq \beta_0 + \beta_1 X_i$,

although the hypotheses could also be expresses as $H_0: \sigma^2_{LOF} = 0$ versus $H_0: \sigma^2_{LOF} > 0$,

or as $H_0: \mu_i = \mu_{y.x}$. Rejection of this hypothesis implies that there is significant departure of some of the $\bar{Y}_i$ from the regression line, with the implication that the regression is not adequate to describe the variation in the means.

b) Test if Lack of Fit

Reduced model (using PROC REG)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	76960	76960	968.66	<.0001
Error	43	3416.37702	79.45063		
Corrected Total	44	80377			

Full Model (using PROC GLM)

Assignment 2 : Copier maintenance example

The GLM Procedure

Dependent Variable: minutes

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	77579.14167	8619.90463	107.84	<.0001
Error	35	2797.65833	79.93310		
Corrected Total	44	80376.80000			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
machines	1	76960.42298	76960.42298	962.81	<.0001
X_number_2	8	618.71869	77.33984	0.97	0.4766

Source	DF	Type III SS	Mean Square	F Value	Pr > F
machines	0	0.0000000	.	.	.
X_number_2	8	618.7186903	77.3398363	0.97	0.4766

The test can be done with the full and reduced models above using a General Linear test approach (below). However, the model was run such that the GLM procedure provides Lack of Fit (MSLOF = 5.838) and is tested with the full model error. The P value is P(>F)=0.9677. For <0.05, there is no evidence that the line does not fit the data adequately.

Source	df	SSE	MS	F	P>F
Reduced model	**16**	**321.39597**			
Full model	**10**	**286.36667**			
Difference	**6**	**35.02931**	**5.83822**	**0.20**	**0.9677**
Full model	**10**	**286.36667**	**28.63667**		

c) The test of LOF does not directly test normality or nonhomogeniety of variance. The calculation of Pure Error is essentially a pooling of "within X variances", and as such requires the usual assumptions.

Under some circumstances, for example if Pure Error was homogeneous and LOF was not, then the larger, heterogeneous variance in LOF may show up as significant. However, this would be unusual, and proper interpretation would not be possible without further examination of the data.

My question number 4 is answered below. There is no comparable questions for copier maintenance.

The Box-Cox analysis was done in PROC TRANSREG. It could be done in PROC REG with a series of dependent variables created in the datastep (e.g. Y1=Y**3; Y2=Y**2; Y3=Y; Y4=Y**0.5; etc). The best model appears to be the one we ran, with Y as the dependent variable.

```
Assignment 2 : Copier maintenance example
Box-Cox transformation with PROC TRANSREG

The TRANSREG Procedure

     Transformation Information
         for BoxCox(minutes)

  Lambda      R-Square    Log Like

   -3.00          0.09    -450.963
   -2.75          0.10    -421.823
   -2.50          0.10    -393.082
   -2.25          0.11    -364.803
   -2.00          0.12    -337.058
   -1.75          0.13    -309.936
   -1.50          0.16    -283.546
   -1.25          0.19    -258.018
   -1.00          0.25    -233.509
   -0.75          0.34    -210.205
   -0.50          0.46    -188.305
   -0.25          0.59    -167.991
    0.00          0.73    -149.381
    0.25          0.83    -132.532
    0.50          0.90    -117.622
    0.75          0.94    -105.516
    1.00 +        0.96     -98.441 <
    1.25          0.96     -98.645 *
    1.50          0.95    -105.060
    1.75          0.94    -114.484
    2.00          0.92    -124.767
    2.25          0.90    -135.069
    2.50          0.88    -145.160
    2.75          0.86    -155.017
    3.00          0.83    -164.679

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