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\CH01PR20a.html' ;

Title1 'Assignment 1 : Copier maintainance example';

DATA ONE; INFILE copier MISSOVER;
      LABEL machines = 'Number of machines serviced';
      LABEL minutes = 'Minutes to service machines';
   INPUT minutes machines;
CARDS; RUN;
;
options ps=45;
PROC PLOT DATA=ONE; PLOT  minutes * machines; run;
options ps=256;

PROC REG DATA=ONE lineprinter; ID machines;
    MODEL  minutes = machines / XPX I P;
    run; options ps=55;
    plot predicted.*machines='X' minutes*machines='o' / overlay;
    output out=next1 p=yhat r=e;
run; options ps=256;

proc sort data=next1; by machines; run;
proc print data=next1; run;

PROC REG DATA=ONE; MODEL minutes = machines; restrict intercept = 0; run;

PROC GLM; MODEL  minutes = machines / XPX I P; run;
```

A1 : Question 1.1 in KNNL) The example given was for sales dollar volume on the number of units sold. If there was no source or errors, this would be a functional relationship, such that Y = 2X
If, however, there were clerical errors in sales, the relationship would no longer be perfectly fitted by this relationship. Although we may feel we know the underlying functional relationship, there would be uncertainty in the fit of the relationship. We would then fit the relationship using

$$Y = \beta_0 + \beta_1 X + \varepsilon_i \qquad \text{for } i = 1, 2, \ , n$$

where, Y = Dollar value of the sale
X = Number of units sold

The random error term ε_i would address the uncertainty, and give the variation due to clerical errors. Additionally, we might expect β_0 to not differ significantly from 0, giving the relationship $Y = \beta_1 X + \varepsilon_i$. We may also hypothesize that β_1 does not differ significantly from 2, if we feel that there is no bias or consistent tendency in the so called “clerical errors”, then the relationship is $Y = 2X + \varepsilon_i$.

A2 : Question 1.2 in KNNL) This function would be would be fixed at Y = 300 + 2X, and would be a functional relationship barring any “clerical errors”.

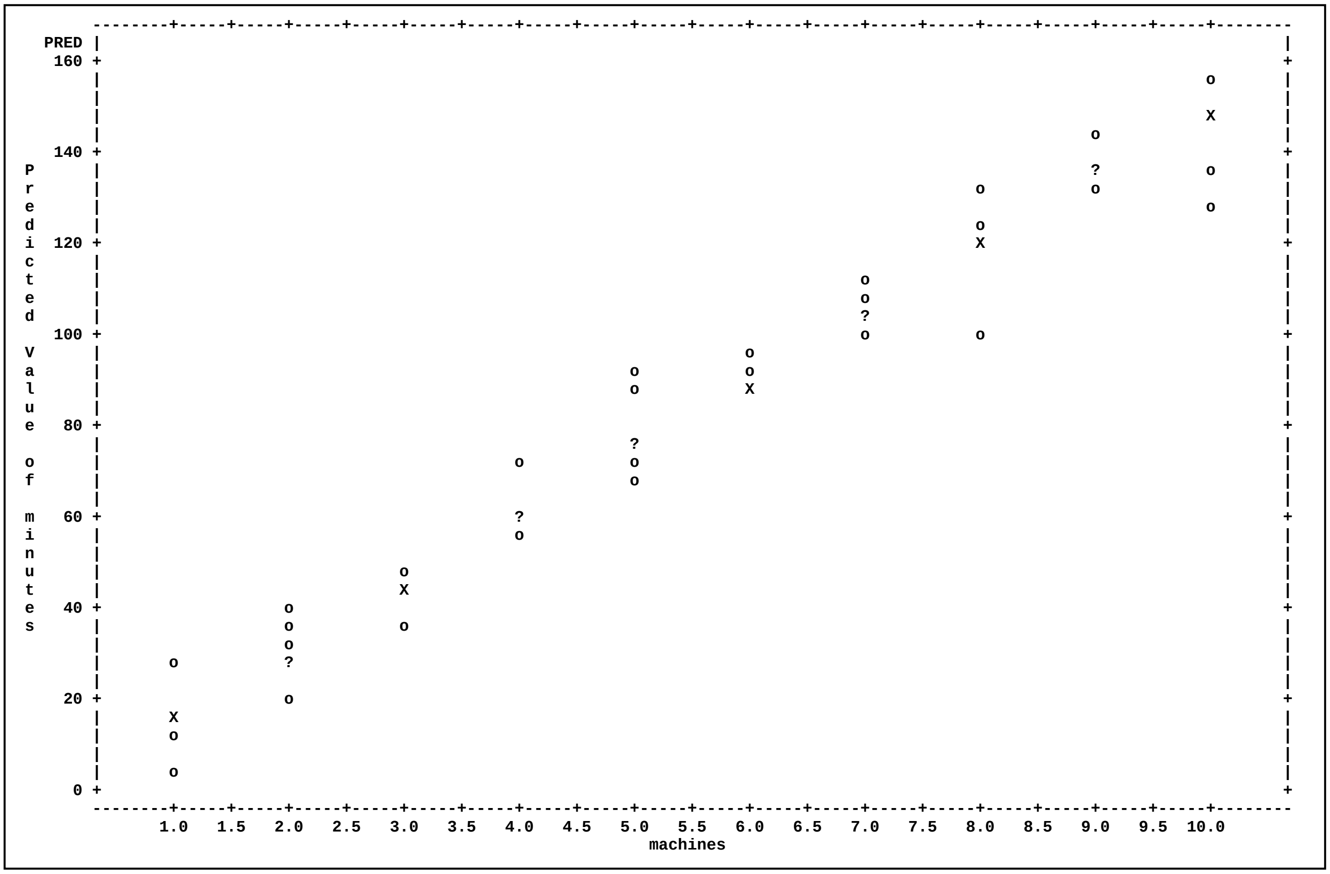
B1 : Question 1.20a in KNNL) Obtain estimated regression function $Y_i = \mathbf{-0.58016} + \mathbf{15.03525} X_i$.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.58016	2.80394	-0.21	0.8371
machines	1	15.03525	0.48309	31.12	<.0001

B1 : Question 1.20b in KNNL) plot reg function and data together

Looks pretty good to me. Notice "X" has been used for the predicted values and "o" for the observed. The questions marks denote where SAS had to place BOTH a "X" and an "o".



B1 : Question 1.20c in KNNL) give an interpretation for b_0

The regression function is $Y_i = -0.58016 + 15.03525X_i$. The intercept is theoretical amount of time required to service a machine when no machine is serviced. We do not know how this company bills for "service time". This value could include travel time, time needed for setting up equipment before actually working on a machine, time to do paperwork after working on a machine. If any of these are included in the time for a call, then we would expect the intercept to be greater than zero, and it would estimated the time needed (in minutes) for these addition tasks. If on the other hand the "service time" includes only time spent working on a machine then we would expect the intercept to be zero (e.g. no machine serviced requires no time).

What we actually observe is a negative number. If the real value is zero then we expect to see a small positive number about half the time and a small negative number about half the time. The question then becomes, does this number differ significantly from zero? SAS tests this (see table for question 1.20a above) and shows that the observed value does not differ significantly from zero (P>F= 0.8371). If the hypotheses was rejected (H_o: $\beta_0 = 0$) and the ***value was negative*** then we may want to question the model adequacy.

The bottom line: YES, I would say that the intercept does tell us something about how this company works and counts the time recorded as "service time". It may also tell us something about the best model (which probably should go through the origin).

Note on testing parameters: Another way to test H_o: $\beta_0 = 0$ is to fit a simple linear regression and "restrict" the model in SAS so that it forces the intercept to be zero (e.g. **`PROC REG; MODEL Y=X; restrict intercept=0;).`** SAS will then report the parameter estimates (see output below) for the model without the intercept ($Y_i = 14.94723X_i$.) and test the restriction (P>|t|=0.8371).

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.20139E-14	0	Infty	<.0001
machines	1	14.94723	0.22642	66.01	<.0001
RESTRICT	-1	-5.86280	28.02544	-0.21	0.8371*

*** Probability computed using beta distribution.**

B1 : Question 1.20d in KNNL) Estimate service time when X = 5

Assignment 1 : Copier maintainance example

Obs	machines	minutes	yhat	e
1	1	12	14.455	-2.4551
2	1	4	14.455	-10.4551
3	1	3	14.455	-11.4551
4	1	27	14.455	12.5449
5	2	20	29.490	-9.4903
6	2	41	29.490	11.5097
7	2	32	29.490	2.5097
8	2	18	29.490	-11.4903
9	2	20	29.490	-9.4903
10	2	28	29.490	-1.4903
11	2	34	29.490	4.5097
12	2	27	29.490	-2.4903
13	3	46	44.526	1.4744
14	3	36	44.526	-8.5256
15	4	60	59.561	0.4392
16	4	72	59.561	12.4392
17	4	57	59.561	-2.5608
18	4	57	59.561	-2.5608
19	4	61	59.561	1.4392
20	**5**	**68**	**74.596**	**-6.5961**
21	**5**	**89**	**74.596**	**14.4039**
22	**5**	**66**	**74.596**	**-8.5961**
23	**5**	**74**	**74.596**	**-0.5961**
24	**5**	**73**	**74.596**	**-1.5961**
25	**5**	**90**	**74.596**	**15.4039**
26	**5**	**86**	**74.596**	**11.4039**
27	**5**	**77**	**74.596**	**2.4039**
28	6	93	89.631	3.3687
29	6	96	89.631	6.3687
30	7	105	104.667	0.3334
31	7	101	104.667	-3.6666
32	7	109	104.667	4.3334
33	7	112	104.667	7.3334
34	7	111	104.667	6.3334
35	7	112	104.667	7.3334
36	8	100	119.702	-19.7018
37	8	131	119.702	11.2982
38	8	123	119.702	3.2982
39	9	144	134.737	9.2629
40	9	134	134.737	-0.7371
41	9	132	134.737	-2.7371
42	9	131	134.737	-3.7371
43	10	137	149.772	-12.7723
44	10	156	149.772	6.2277
45	10	127	149.772	-22.7723

The values to the right are output from PROC REG. PROC REG does not usually include the value of the independent variable in the output. The values of X have been included in the output statistics on the left because the variable X was included in an "ID" statement.

From the output statistics it is clear that there were 8 observations with an X value equal to 5. For these values the point estimate of service time, or predicted value, is equal to 74.596 minutes.

B2 : Question 1.20c+ in KNNL additional request) give an interpretation for b_1

In question 1.20c we had a rather long-winded interpretation of b_0. The value estimated for b_1 has a simpler interpretation. It is the change in "service time" per machine. This would be our best estimate of the time required to service one machine, and was estimated as **15.03525** minutes.

B3 : ANOVA table) This was estimated by the SAS program as

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			

B4 : Normal Equations) The normal equations are provided in SAS in a somewhat indirect form, but they can be obtained. The matrix notation for the normal equations are (X'X)b = X'Y. The model option "XPX" (e.g. `MODEL Y=X / XPX;`) will cause the following listing.

Model Crossproducts X'X X'Y Y'Y

Variable	Intercept	machines	minutes
Intercept	45	230	3432
machines	230	1516	22660
minutes	3432	22660	342124

The 4 values in the upper-left portion of the listing (18, 81, 81 and 439) are the elements of the X'X matrix and the two upper elements in the right column (1152 and 6282) are the two values of the X'Y vector. The lower-right element is the value of Y'Y. The remain two values (the first two elements of the last row, 1152 and 6282) are the (X'Y) '. This causes the 9 values to create a symmetric matrix. Multiplying out the matrix algebra brings the equations more in line with the "algebraic" version of the normal equations.

The normal equations expressed algebraically are:

$$nb_0 + b_1\Sigma X_i = \Sigma Y_i$$

$$b_0\Sigma X_i + b_1\Sigma X^2_i = \Sigma Y_i X_i$$

Filling in the quantitative values for the intermediate calculations we get:

$$45b_0 + 230b_1 = 3432$$

$$230b_0 + 1526b_1 = 22660$$

These are the normal equations, the equations that must be solved to get estimates of b_0 and b_1.

B5 : Regression coefficients and their standard errors) Given directly from the SAS output

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.58016	2.80394	-0.21	0.8371
machines	1	15.03525	0.48309	31.12	<.0001

B6 : Question 1.24a in KNNL) obtain residuals and sum of square of residuals

The residuals can be listed in most SAS procedures (GLM, REG, MIXED, etc.). The residuals were listed in question **1.20d** above. The sum of squared residuals are simply the SSError, and this is an acceptable answer. However, PROC GLM actually calculates the sum of the residuals (squared and unsquared) if the P option is specified on the model. These values were given as:

```
Sum of Residuals                              0.000000
Sum of Squared Residuals                   3416.377023
Sum of Squared Residuals - Error SS           0.000000
```

The value minimized in fitting a least squares regression is the sum of squares deviations (error). The book describes this value as $Q=\sum_{i=1}^{n}\left(Y_i-\beta_0-\beta_1X_i\right)^2$. Since the sum of square of the deviations or residuals are defined as $\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$, and $\hat{Y}_i=\beta_0-\beta_1X_i$, the two values ($Q$ and $\sum_{i=1}^{n}e_i^2$) are equivalent. Both were numerically equal to zero out to 6 decimal places (above).

B6 : Question 1.24b in KNNL) obtain estimates of σ^2 and σ.

These are estimated by the MSE from the SAS Analysis of Variance table (MSE = **79.45063**) and the Root MSE (= 8.91351) also provided by SAS, usually following the ANOVA table. The units on the

```
Root MSE              8.91351    R-Square     0.9575
Dependent Mean       76.26667    Adj R-Sq     0.9565
Coeff Var            11.68729
```

estimate of σ would be the same as the dependent variable, minutes.

Additional questions for chapter 2 were answer with the following statements.

```
OPTIONS LS=111 PS=256 NOCENTER;
PROC REG DATA=ONE lineprinter; ID machines;
    MODEL  minutes = machines / CLM CLI CLB alpha=0.10;
    output out=next1 p=yhat r=e;
    TEST machines = 14;
run;
PROC GLM DATA=ONE; classes anotherx;
    MODEL  minutes = X anotherx;
run;

PROC MEANS DATA=ONE N MEAN SUM VAR USS CSS; VAR machines
minutes; run;

QUIT;
```

Problem 2.5a) From the SAS output – SAS provides a confidence interval for the regression coefficients (b_0 and b_1) and the value for α can be specified, 0.10 in this case. The resulting values were 14.22314 and 15.84735. Joint confidence intervals were not discussed in chapter 2 so I assume

```
Parameter Estimates
                    Parameter    Standard
Variable      DF     Estimate       Error    t Value    Pr > |t|    90% Confidence Limits
Intercept      1     -0.58016     2.80394      -0.21      0.8371    -5.29378         4.13347
machines       1     15.03525     0.48309      31.12      <.0001    14.22314        15.84735
```

they were not the intended answer here. A more complete probability statement would be given as:

$$P(b_1-t_{\alpha/2}S_{b1} \le \beta_1 \le b_1-t_{\alpha/2}S_{b1}) = 1-\alpha$$

Where, n = 45 (d.f. = 43) and α = 0.10 so $t_{\alpha/2}$ = 1.681070704 (from excel).

From previous work b_1 = 15.03525 and S_{b1} = 0.48309.

$$P(15.03525 – 1.746*0.48309 \le \beta_1 \le 15.03525 + 1.746*0.48309) = 0.90 ,$$

and **P(14.22314155 ≤ β_1 ≤ 15.84735845) = 0.90**

Problem 2.5b) The t-test of a linear association is available in the SAS output **(H_0: β_1=0). The alternative is H_1: $\beta_1 \neq 0$**. Rejection would result at the **α = 0.10** level if the calculated value of the t-value was greater than the critical tabular value for 16 d.f., **t = 1.681070704**. SAS reports the actual calculated value to be **31.12** (from the table above) so we clearly reject the null hypothesis.

In the output above SAS also provides a P value that indicates that this t-test is highly significant (**P>|t|)<0.0001**. All this means is that there appears to be a correlation between these two variables, and that values of X_i would provide some utility for estimating Y_i. However, this by no means indicates that this particular "linear association" that we fitted (which is linear) is the best linear model for the relationship or that a nonlinear model might not be even better.

Problem 2.5c) Since the confidence interval in part 2.5a does not include zero, and the results of the t-test in part 2.5b reject zero, **the results are consistent**.

Problem 2.5d) This was done using the TEST statement in SAS (i.e. **`TEST X = 14;`**) However, this is a two tailed test. Since we want a one tailed test (α=0.05) we should look at the upper tail of only. Since one tail equal to α=0.05 corresponds to a two tailed test of α=0.10. Therefore, we would

```
The REG Procedure
Model: MODEL1
      Test 1 Results for Dependent Variable minutes
                                  Mean
Source              DF          Square    F Value    Pr > F
Numerator            1       364.86742       4.59    0.0378
Denominator         43        79.45063
```

reject H_0: β_1=14 with α=0.05 when the two tailed test was less than P(>F)<0.10, AND IF THE RESULT WAS IN THE UPPER TAIL (i.e. H_1: β_1 > 14). To get a one tailed P-value, divide the observed P value by 2 (e.g. P value = 0.0378 / 2 = 0.0189).

In this case the P-value indicates that the results are clearly in the upper tail. We would therefore reject the null hypothesis (H_0: β_1=14 or H_0: $\beta_1 \le 14$), concluding that the alternative hypothesis (H_0: β_1>14) is a more reasonable conclusion than the null hypothesis. This is consistent with the 90% confidence interval for our estimate of β_1, **P(14.22314155 ≤ β_1 ≤ 15.84735845) = 0.90.**

Problem 2.5e) Of course it does (depending on what you call relevant information). We EXPECT a value greater than zero if there is any startup time for the machine repair. However, evidence indicates that there is **no start up time**. In the event there is no startup time, then we might expect the intercept to not be significantly different from zero (which is the case for this data). In the event that we should find a statistically significantly negative intercept it might suggest that the model is wrong or the data is wrong, since for this problem the intercept should not be negative.

Problem 2.14a) The information requested here is a 90% CLM. The SAS GLM output below contains this confidence interval as observation 13 and observation 30. Note that the value of X is listed with each observation. The SAS statements producing this output were: **`MODEL Y = X / CLM CLI CLB alpha=0.10; run;`**

Question 2.14a is answered by the results for observation 2, where 6 machines were serviced. The answer to the question is provided by the CLM (since they ask about the **mean** service time) and the probability statement is: $\mathbf{P(87.2839 \le E(Y_{X=6}) \le 91.9788) = 0.90}$

Output Statistics

Obs	machines	Dependent Variable	Predicted Value	Std Error Mean Predict	90% CL Mean		90% CL Predict		Residual
1	2	20.0000	29.4903	2.0061	26.1180	32.8627	14.1313	44.8494	-9.4903
2	4	60.0000	59.5608	1.4331	57.1517	61.9699	44.3842	74.7375	0.4392
3	3	46.0000	44.5256	1.6750	41.7098	47.3414	29.2791	59.7721	1.4744
4	2	41.0000	29.4903	2.0061	26.1180	32.8627	14.1313	44.8494	11.5097
5	1	12.0000	14.4551	2.3895	10.4381	18.4721	-1.0582	29.9684	-2.4551
6	10	137.0000	149.7723	2.7099	145.2168	154.3278	134.1109	165.4337	-12.7723
7	5	68.0000	74.5961	1.3298	72.3605	76.8316	59.4460	89.7462	-6.5961
8	5	89.0000	74.5961	1.3298	72.3605	76.8316	59.4460	89.7462	14.4039
9	1	4.0000	14.4551	2.3895	10.4381	18.4721	-1.0582	29.9684	-10.4551
10	2	32.0000	29.4903	2.0061	26.1180	32.8627	14.1313	44.8494	2.5097
11	9	144.0000	134.7371	2.3011	130.8688	138.6054	119.2616	150.2126	9.2629
12	10	156.0000	149.7723	2.7099	145.2168	154.3278	134.1109	165.4337	6.2277
13	**6**	**93.0000**	**89.6313**	**1.3964**	**87.2839**	**91.9788**	**74.4643**	**104.7983**	**3.3687**
14	3	36.0000	44.5256	1.6750	41.7098	47.3414	29.2791	59.7721	-8.5256
15	4	72.0000	59.5608	1.4331	57.1517	61.9699	44.3842	74.7375	12.4392
16	8	100.0000	119.7018	1.9270	116.4624	122.9412	104.3714	135.0322	-19.7018
17	7	105.0000	104.6666	1.6119	101.9569	107.3763	89.4393	119.8939	0.3334
18	8	131.0000	119.7018	1.9270	116.4624	122.9412	104.3714	135.0322	11.2982
19	10	127.0000	149.7723	2.7099	145.2168	154.3278	134.1109	165.4337	-22.7723
20	4	57.0000	59.5608	1.4331	57.1517	61.9699	44.3842	74.7375	-2.5608
21	5	66.0000	74.5961	1.3298	72.3605	76.8316	59.4460	89.7462	-8.5961
22	7	101.0000	104.6666	1.6119	101.9569	107.3763	89.4393	119.8939	-3.6666
23	7	109.0000	104.6666	1.6119	101.9569	107.3763	89.4393	119.8939	4.3334
24	5	74.0000	74.5961	1.3298	72.3605	76.8316	59.4460	89.7462	-0.5961
25	9	134.0000	134.7371	2.3011	130.8688	138.6054	119.2616	150.2126	-0.7371
26	7	112.0000	104.6666	1.6119	101.9569	107.3763	89.4393	119.8939	7.3334
27	2	18.0000	29.4903	2.0061	26.1180	32.8627	14.1313	44.8494	-11.4903
28	5	73.0000	74.5961	1.3298	72.3605	76.8316	59.4460	89.7462	-1.5961
29	7	111.0000	104.6666	1.6119	101.9569	107.3763	89.4393	119.8939	6.3334
30	**6**	**96.0000**	**89.6313**	**1.3964**	**87.2839**	**91.9788**	**74.4643**	**104.7983**	**6.3687**
31	8	123.0000	119.7018	1.9270	116.4624	122.9412	104.3714	135.0322	3.2982
32	5	90.0000	74.5961	1.3298	72.3605	76.8316	59.4460	89.7462	15.4039
33	2	20.0000	29.4903	2.0061	26.1180	32.8627	14.1313	44.8494	-9.4903
34	2	28.0000	29.4903	2.0061	26.1180	32.8627	14.1313	44.8494	-1.4903
35	1	3.0000	14.4551	2.3895	10.4381	18.4721	-1.0582	29.9684	-11.4551
36	4	57.0000	59.5608	1.4331	57.1517	61.9699	44.3842	74.7375	-2.5608
37	5	86.0000	74.5961	1.3298	72.3605	76.8316	59.4460	89.7462	11.4039
38	9	132.0000	134.7371	2.3011	130.8688	138.6054	119.2616	150.2126	-2.7371
39	7	112.0000	104.6666	1.6119	101.9569	107.3763	89.4393	119.8939	7.3334
40	1	27.0000	14.4551	2.3895	10.4381	18.4721	-1.0582	29.9684	12.5449
41	9	131.0000	134.7371	2.3011	130.8688	138.6054	119.2616	150.2126	-3.7371
42	2	34.0000	29.4903	2.0061	26.1180	32.8627	14.1313	44.8494	4.5097
43	2	27.0000	29.4903	2.0061	26.1180	32.8627	14.1313	44.8494	-2.4903
44	4	61.0000	59.5608	1.4331	57.1517	61.9699	44.3842	74.7375	1.4392
45	5	77.0000	74.5961	1.3298	72.3605	76.8316	59.4460	89.7462	2.4039

Problem 2.14b) The information requested here is a 90% CLI, since it is a single trip. This is also available in the table above as observation 2. The SAS GLM output below contains this confidence interval as observation 2. The probability statement is : **$P(74.4643 \le E(Y_{X=6}) \le 104.7983) = 0.90$**

Problem 2.14c) The answer to question 2.14a was **$P(87.2839 \le E(Y_{X=6}) \le 91.9788) = 0.90$.** This is the estimated time for the repair of 6 machines. This calculation is done as a confidence interval on the estimated value, $b_1 \pm t_{\alpha/2}S_{b1} = 89.6313 \pm 1.681070704*1.3964$. For this problem the standard error is $MSE\left(\frac{1}{n}+\frac{(X_i-\bar{X})^2}{\Sigma(X_i-\bar{X})^2}\right)$.

From proc means

```
PROC MEANS DATA=ONE N MEAN SUM VAR USS CSS; VAR machines minutes; run;
```

We get the result:

```
The MEANS Procedure
Variable   N        Mean           Sum     Variance          USS    Corrected SS
------------------------------------------------------------------------------
machines  45   5.1111111   230.0000000    7.7373737      1516.00     340.4444444
minutes   45  76.2666667       3432.00      1826.75    342124.00        80376.80
-------------------------------------------------------------------------------
```

$MSE\left(\frac{1}{n}+\frac{(X_i-\bar{X})^2}{\Sigma(X_i-\bar{X})^2}\right)$ is then $79.45063[1/45 + (X_i-5.1111111)^2/340.4444444]$. For $X_i = 6$ this value is 1.396410845. This value is provided by SAS (see STD ERROR MEAN PREDICT in the output statistics).

For a new sample of size m = 6 the calculation is $MSE\left(\frac{1}{m}+\frac{1}{n}+\frac{(X_i-\bar{X})^2}{\Sigma(X_i-\bar{X})^2}\right) =$

$MSE\left(\frac{1}{n}+\frac{(X_i-\bar{X})^2}{\Sigma(X_i-\bar{X})^2}\right)+\frac{MSE}{m}$. In order to estimate the mean time for 6 machines we need modify the previous calculation of the interval of the mean by adding MSE/m = 79.45063/6 = 13.24177167. The variance for 6 machines is then $79.45063[1/45 + (X_i-5.1111111)^2/340.4444444] + 13.24177167 =$ 15.19173491. Then the standard error is $\sqrt{15.19173491} = 3.897657619$.

The interval is then $b_1 \pm t_{\alpha/2}S_{b1} = 89.6313 \pm 1.681070704*3.897657619$, and

$P(83.07906196 \le E(Y_{X=6}) \le 96.18353804) = 0.90$.

Problem 2.14d) This interval for the mean of 6 machines should be wider the interval for the regression line, **$P(87.2839 \le E(Y_{X=6}) \le 91.9788) = 0.90$**, but narrower tan the interval for individual points, **$P(74.4643 \le E(Y_{X=6}) \le 104.7983) = 0.90$**.