UNIT 11 HW

1. From Problem 26, Chapter 8:

The Metabolic data set has the average mass, metabolic rate, and average lifespan of 95 different species of mammals. Kleiber's Law states that the metabolic rate of an animal species, on average, is proportional to its mass raised to the power ¾. Judge the adequacy of this theory with these data. Ultimately, for this problem, we want to find the best model. (At this point, you will limit the analysis to the two variables under study, though the data set has more variables.) In the current data set, assume that mass has not yet been raised to the power ¾.

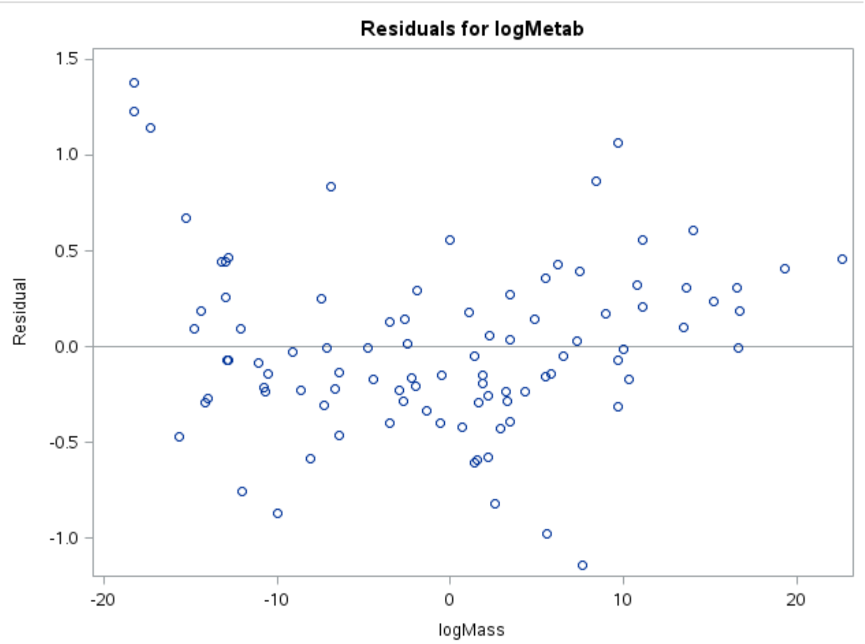
- Use alpha = 0.05.
- Use **SAS** for this problem.
- Include **relevant** code and output. Make sure you directly answer the questions. Do NOT assume the answer is obvious from the output.

Specifically, provide/answer the following:

a. Judging by a scatterplot alone, does it seem reasonable that the metabolic rate of an animal species, on average, is proportional to its mass raised to the power of ¾? (Recall that if some variable y is proportional to the variable x, then $\hat{y} = mx$ (with nonzero m) is a well-fitting model.) In other words, does the data (metabolic rate, mass$^{3/4}$) reasonably fall along a straight (nonhorizontal) line and nearly pass through the origin?

```
proc import datafile = '/Metabolism Data Prob 26.csv'
 out = meta
 dbms = csv
 ;

proc print data = meta;
run;

proc reg data=meta alpha = .05;
model Metab = Mass;
run;


data logmeta;
set meta;
logMetab= log(Metab);
logMass = log(Mass**3/4);
run;

proc print data = logmeta;
```

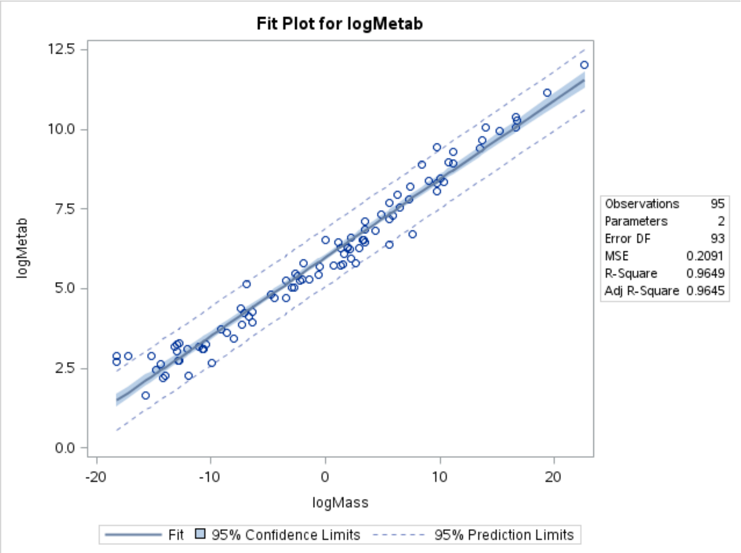
```
run;

proc reg data = logmeta;
model logMetab= logMass / cli;
run;
```



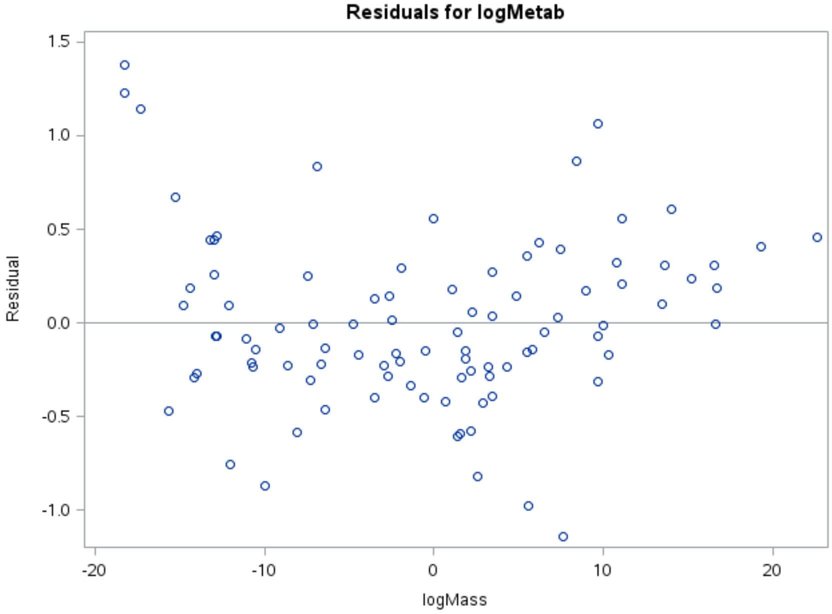
Random, scatter with constant variance, no trend

b. We want to find the "best" model to predict metabolic rate from mass$^{3/4}$ **and** make appropriate statistical inferences. Therefore, address all the assumptions prior to the analysis (using mass$^{3/4}$). If the assumptions are not met, handle the data appropriately. If a transformation is used to satisfy the assumptions, address the assumptions again to ensure that the transformation is logical, and carry out your analysis on your newly transformed data. For example, you should include a scatter plot for the original data AND transformed data, etc. (Hint: if a transformation is necessary, try one of the transformations discussed in class first.) Either way, keep the "mass$^{3/4}$" in the model; do not go back to regular "mass," although mass$^{3/4}$ may be transformed if it makes sense for the assumptions. At minimum, provide and interpret the following elements to address assumptions FOR THE ORIGINAL DATA AND ANY TRANSFORMED DATA (IF you use a transformation). You may include more graphs if you find them useful.

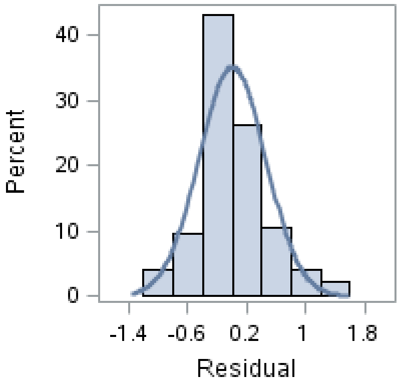
   i. A scatterplot with the following included on the graph: regression line, confidence intervals of the regression line, and prediction intervals of the regression line.



ii. A scatterplot of residuals.

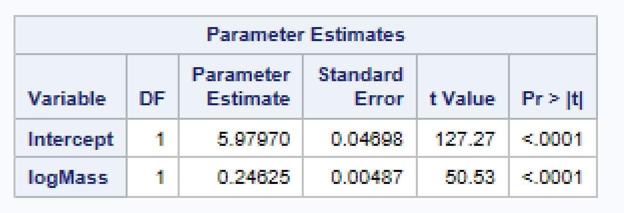


iii. A histogram of residuals with the normal distribution superimposed.



iv. A discussion supporting the use of the model you chose (support that the assumptions are met).

c. Once a reasonable model is found (possibly using a transformation), provide a table showing the t-statistics and p-values for the significance of the regression parameters $\beta_0\ and\ \beta_1$.



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.97970	0.04698	127.27	<.0001
logMass	1	0.24625	0.00487	50.53	<.0001

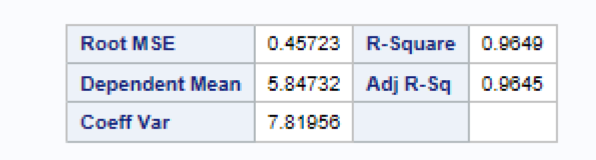
d. The estimated regression equation. Make sure the dependent variable is noted as the predicted value or predicted mean value, not just the dependent variable.

logMetabolism = 5.9797 + 0.24625(logMass)

e. Interpretation of the model, paying special attention if you used a transformation (hint!). That is, interpret the slope as well as the **confidence interval**.

Log transformed both X and Y. A doubling of X is associated with a multiplicative change of $2^{\beta 1}$ in the median of Y. A 95% confidence interval for β1 is (0.24625-1.985802*0.00487, 0.24625+1.985802*0.00487) = (0.2366 , 0.2559), a 95% confidence interval for the multiplicative factor in the median is $2^{.2366}$ to $2^{.2559}$, which is (1.178, 1.194).

f. A measure of the proportion of variation in the response that is accounted for by the explanatory variable. Interpret this measure clearly.



Root MSE	0.45723	R-Square	0.9649
Dependent Mean	5.84732	Adj R-Sq	0.9645
Coeff Var	7.81956		

r2 = 0.9649

It is estimated that mass explains about 96.49% of the variation in metabolic rate for these species.

2. From Problem 29, Chapter 8:

The autism data show the prevalence of autism per 10,000 ten-year-old children in the United States in each of five years. Analyze the data to describe the change in the distribution of autism prevalence per year during this time period.

- Use alpha = 0.05.
- Use **R** for this problem.

- Include **relevant** code and output. Make sure you directly answer the questions. Do NOT assume the answer is obvious in the output.

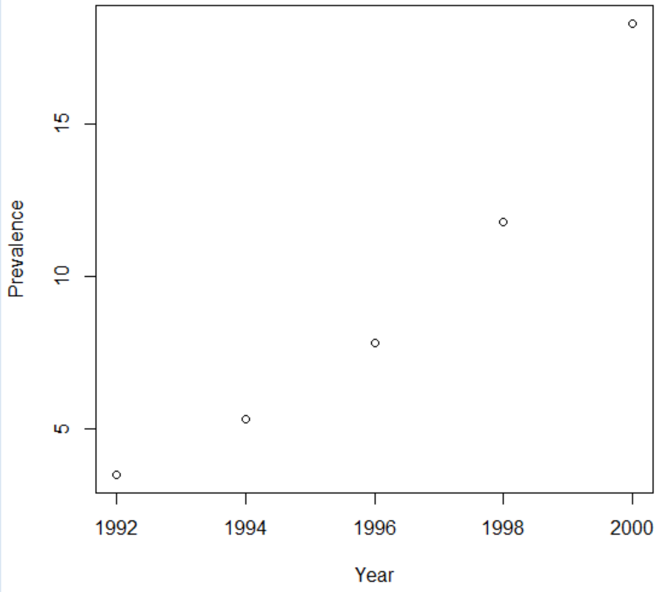
Specifically, provide/answer the following:

a. Address all the assumptions for a linear regression model prior to the analysis. If the assumptions are not met, handle the data appropriately. If a transformation is used, address the assumptions again with the transformed data to ensure that the transformation is logical. The questions below should reflect this. For example, you should include a scatter plot for the original data AND transformed data, etc. (Hint: if a transformation is necessary, try one of the transformations discussed in class first.) At minimum, provide and interpret the following elements to address assumptions FOR THE ORIGINAL DATA AND ANY TRANSFORMED DATA (IF you use a transformation). You may include more graphs if you find them useful.

   i. A scatterplot with the following included on the graph: regression line, confidence intervals of the regression line, and prediction intervals of the regression line.

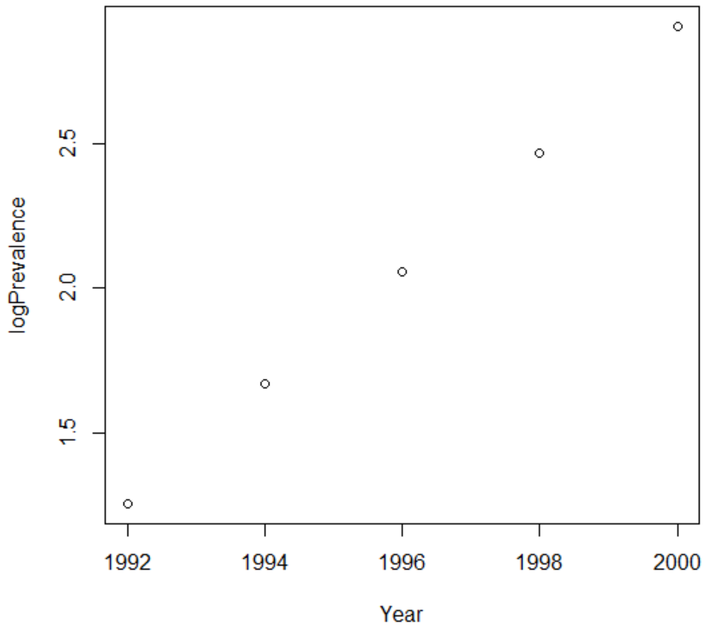
Original data

plot(autism$Year, autism$Prevalence, ylab = "Prevalence", xlab = "Year")



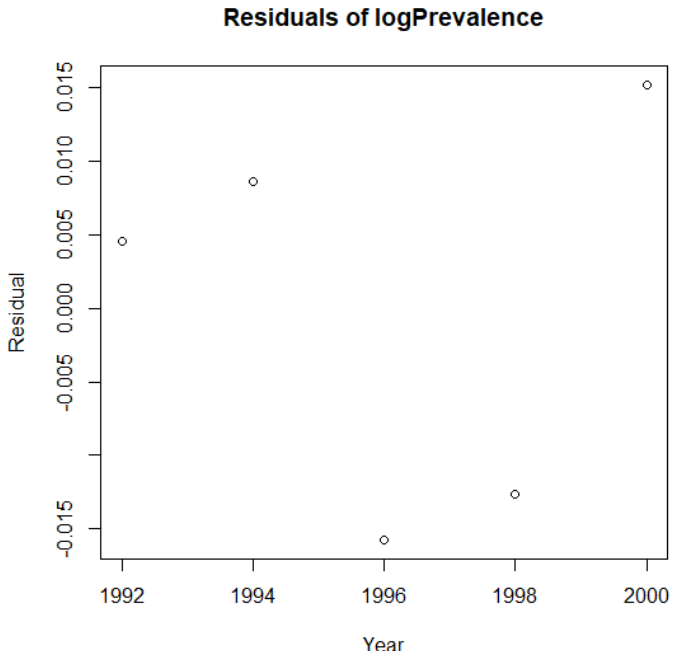
Tranformed data

plot(autism$Year, autism$log.Prevalence, ylab = "logPrevalence", xlab = "Year")



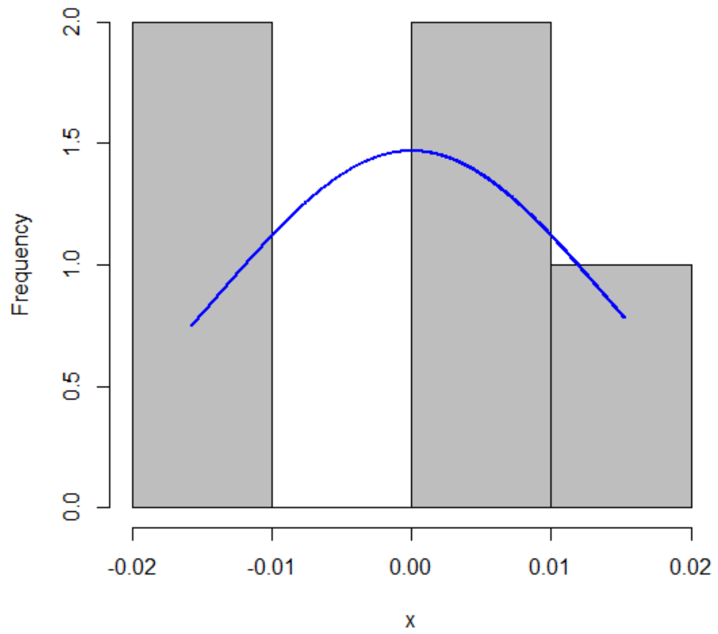
ii. A scatterplot of residuals.

plot(autism$Year, res.autism, main="Residuals of logPrevalence", ylab= "Residual", xlab="Year")



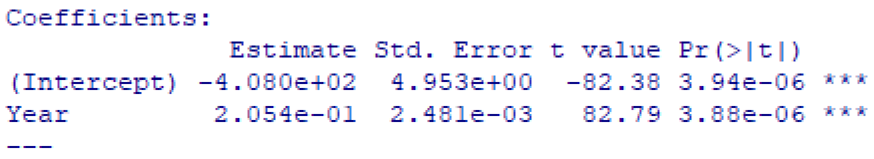
iii. A histogram of residuals with the normal distribution superimposed.

plotNormalHistogram(residuals(lin.reg2))



iv. A discussion supporting the use of the model you chose (support that the assumptions are met).

b. Once a reasonable model is found (possibly using a transformation), provide a table showing the t-statistics and p-values for the significance of the regression parameters $\beta_0 \; and \; \beta_1$.



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.080e+02  4.953e+00  -82.38 3.94e-06 ***
Year         2.054e-01  2.481e-03   82.79 3.88e-06 ***
---
```

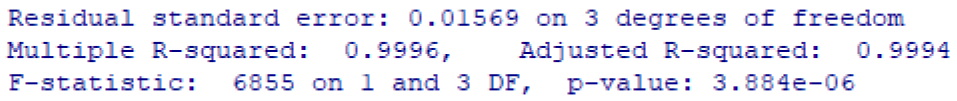
c. The estimate regression equation. Make sure the dependent variable is noted as the predicted value or predicted mean value, not just the dependent variable.

logPrevalence= 0.2054(Year)-408

d. Interpretation of the model, paying special attention if you used a transformation (hint!). That is, interpret the slope as well as the **confidence interval**.

The data suggest that each 1-year increase in Year is associated with an increase in Prevalence multiplicative change of $e^{0.2054}$ = 1.228 in Median(Prevalence|Year). In other words, a one unit increase in Year is associated with a 22.8% increase in Prevalence. A 95% confidence interval for β_1 is (0.2054 – 3.182446*.00248, 0.2054 + 3.182446*.00248) = (0.2384, 0.2541). Therefore, a 95% confidence interval for e^{B1} $2^{.2384}$ to $2^{.2541}$, which is (1.180, 1.193).

e. A measure of the proportion of variation in the response that is accounted for by the explanatory variable. Interpret this measure clearly.



```
Residual standard error: 0.01569 on 3 degrees of freedom
Multiple R-squared:  0.9996,    Adjusted R-squared:  0.9994
F-statistic:  6855 on 1 and 3 DF,  p-value: 3.884e-06
```

R2 =0.9996. It is estimated that Year explains about 99.96% of the variation in Prevalence.

Bonus!

Consider the steer data in Display 7.3 on page 179 (Chapter 7) of the textbook (third edition). Perform a lack of fit test comparing the regression model and a separate means model. Because we have at least two points in at least one group (replication to estimate the variance), we can perform ANOVA. (ANOVA does not make sense if no values of the independent variable are repeated.) During live session, we already addressed the assumptions and determined that a linear-log model is best for regression. Perform this lack of fit test (all parts) on the transformed data. Use the software of your choice. Specifically, include the following:

f. Hypotheses

H_o:linear regression model fits

H_a:variability in the means cannot be explained by the model

g. The ANOVA table you created

Source	DF	SS	MS	F	Pr > F
Model	3	0.00299	.00099693	3.75	0.0944
Error	5	0.00133	.00026581		
Corrected Total	8	0.00432			

h. Decision

there is not enough evidence to suggest the linear regression model has a lack of fit with respect to the separate means model.

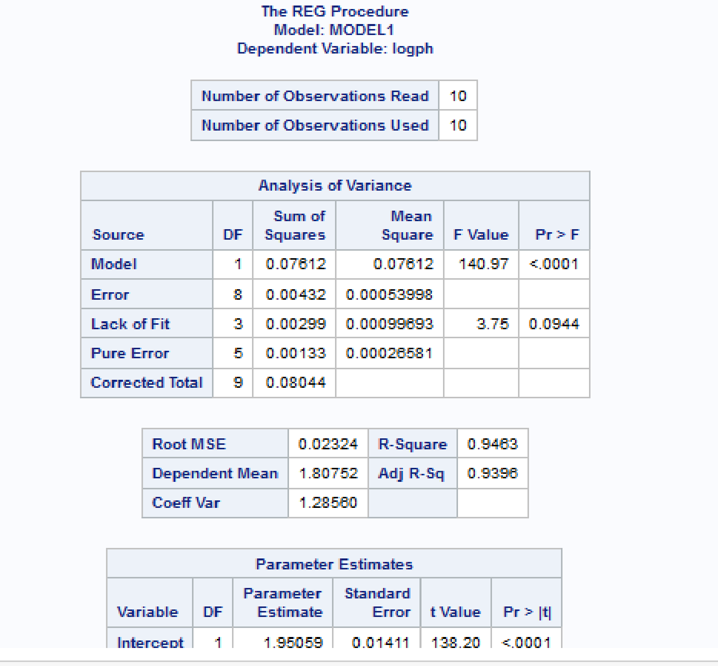
d. Conclusion in non-statistical terms

It appears a straight line approximation is reasonable for hours 1-8,

e. Code and relevant output

```
data meat;
input hour ph;
datalines;
```

```
1 7.02
1 6.93
2 6.42
2 6.51
4 6.07
4 5.99
6 5.59
6 5.80
8 5.51
8 5.36
;

data logPH;
set meat;
logph= log(ph);
run;


proc reg data = logPH;
model logph= hour / lackfit;
run;
```



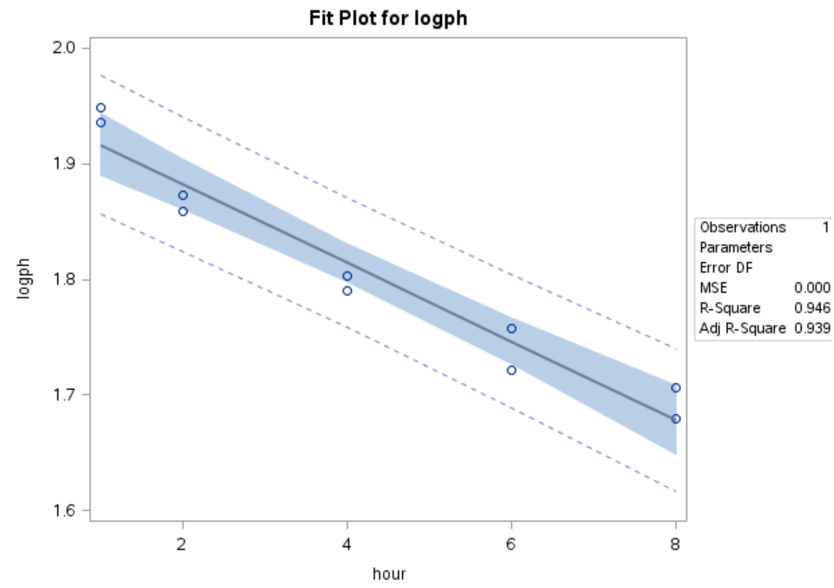
The REG Procedure
Model: MODEL1
Dependent Variable: logph

Number of Observations Read	10
Number of Observations Used	10

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.07612	0.07612	140.97	<.0001
Error	8	0.00432	0.00053998		
Lack of Fit	3	0.00299	0.00099693	3.75	0.0944
Pure Error	5	0.00133	0.00026581		
Corrected Total	9	0.08044			

Root MSE	0.02324	R-Square	0.9463
Dependent Mean	1.80752	Adj R-Sq	0.9396
Coeff Var	1.28560		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.95059	0.01411	138.20	<.0001


Fit Plot for logph
2.0
1.9
1.8
1.7
1.6
logph
2
4
6
8
hour
Observations 1
Parameters
Error DF
MSE 0.000
R-Square 0.946
Adj R-Square 0.939