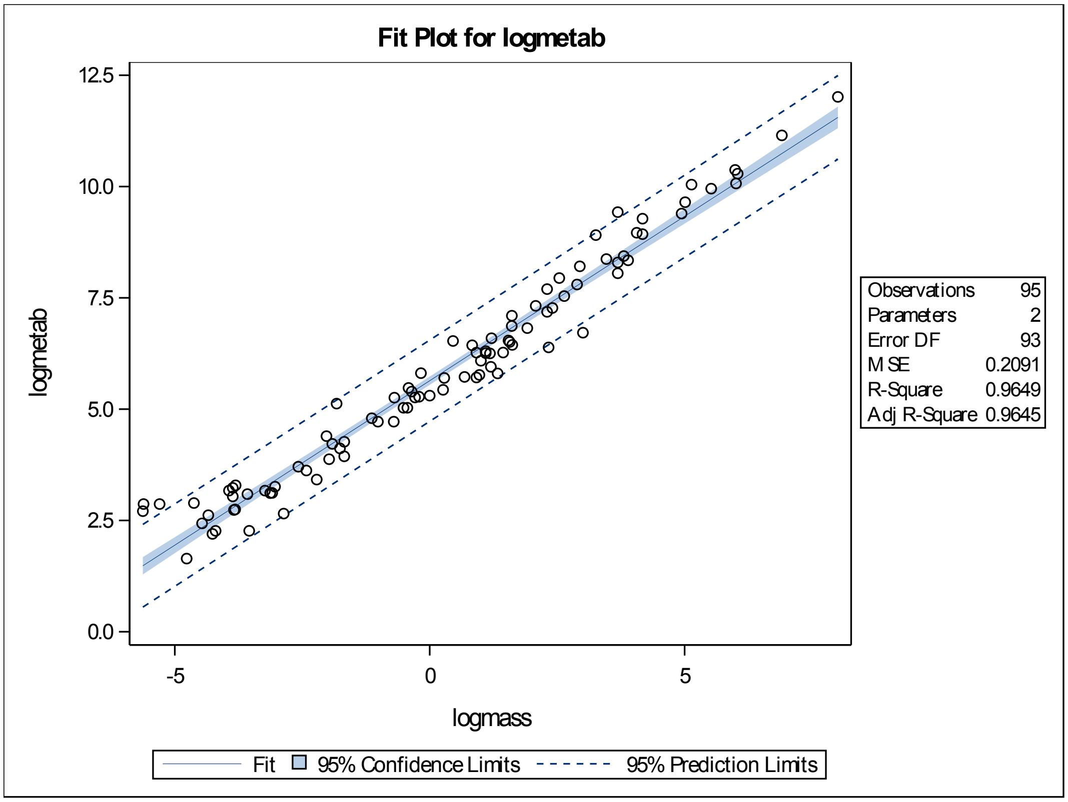
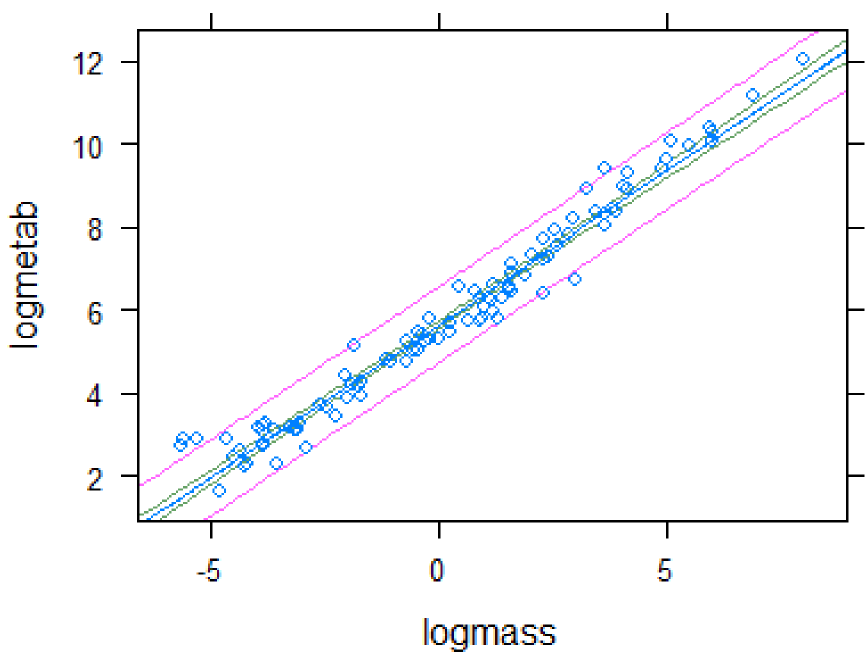
Question 1)

Part i) Scatterplot with Regression Confidence & Prediction Intervals (SAS/R)





Part ii) Tabular Regression T-stats and P-Values (SAS/R)

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	5.638330664	0.04709325	119.73	<.0001	5.544812798	5.731848530
logmass	0.738743639	0.01461957	50.53	<.0001	0.709712080	0.767775198

```
Call:
lm(formula = logmetab ~ logmass, data = ex0826)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1422 -0.2647 -0.0489  0.2531  1.3762

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.6383     0.0471   119.7   <2e-16 ***
logmass       0.7387     0.0146    50.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.457 on 93 degrees of freedom
Multiple R-squared:  0.965, Adjusted R-squared:  0.964
F-statistic: 2.55e+03 on 1 and 93 DF,  p-value: <2e-16
```

Part iii) Regression Equation

Log Transforming both X and Y:

ln y-hat = 5.6383 + .7387 ln x
y-hat = e^(5.6383) * (e^(ln x))^.7387
y-hat = 280.984638 * x^(.7387)

--> **when x (Mass) is 1, y-hat (Metabolism) = 281;**
-->**multiplying x by 2, multiplies y-hat by 2^(.7387) or 1.6686715**
-->**multiplying x by n, multiplies y-hat by n^(.7387)**

Part iv) Model Interpretation

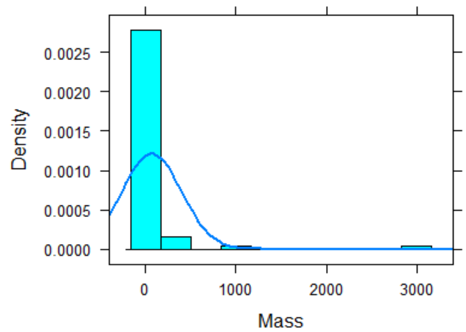
The regression coefficients and CIs confirm the adequacy of the theory espoused by Kleiber's Law that: the metabolic rate of an animal species is, on average, proportional to its mass raised to the power of ¾.

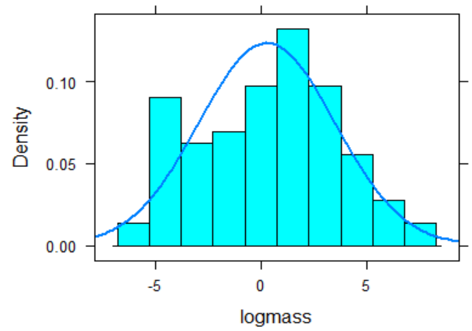
```
> coef(logmetab_logmass)
(Intercept)     logmass
     5.6383      0.7387
```

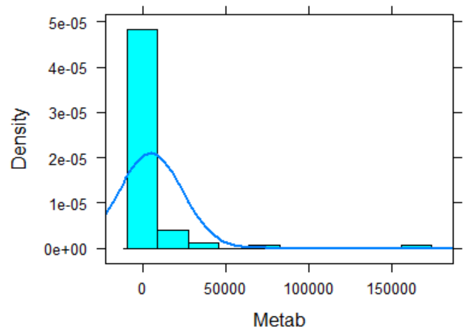
```
> confint(logmetab_logmass)
               2.5 %    97.5 %
(Intercept) 5.5448    5.7318
logmass       0.7097    0.7678
```

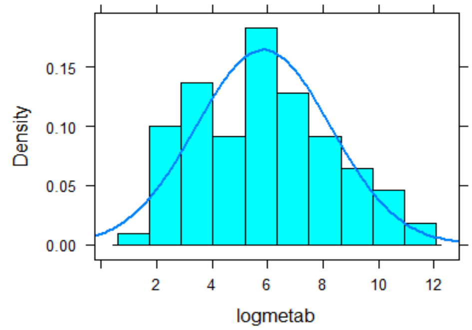
Part v) Residual Scatterplots (R/SAS)

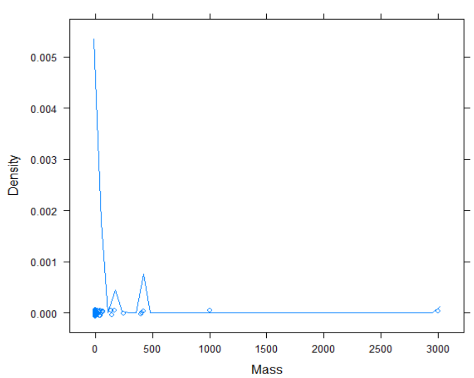
The log transformations of X and Y produce the following figures and residual scatterplots (R/SAS):

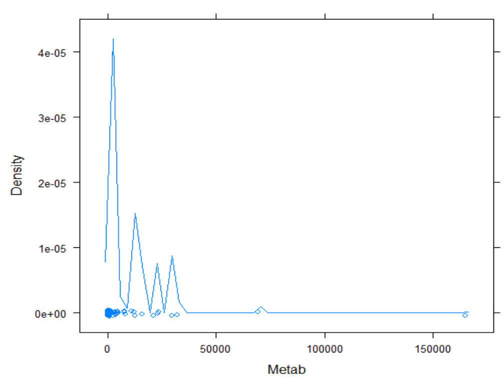

Density
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
1000
2000
3000
Mass

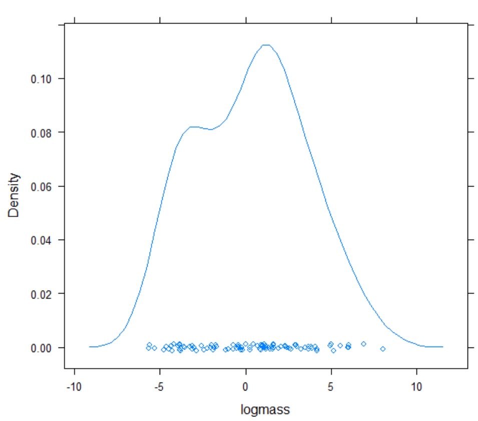

Density
0.10
0.05
0.00
-5
0
5
logmass

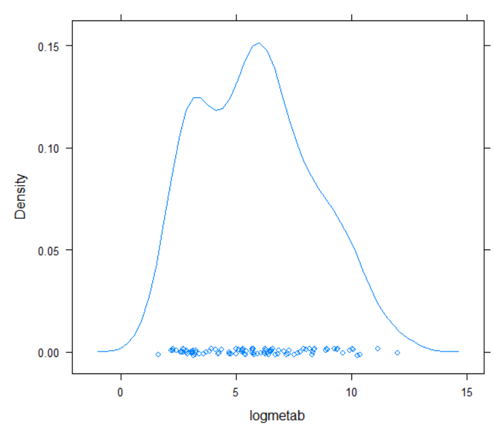

Density
5e-05
4e-05
3e-05
2e-05
1e-05
0e+00
0
50000
100000
150000
Metab

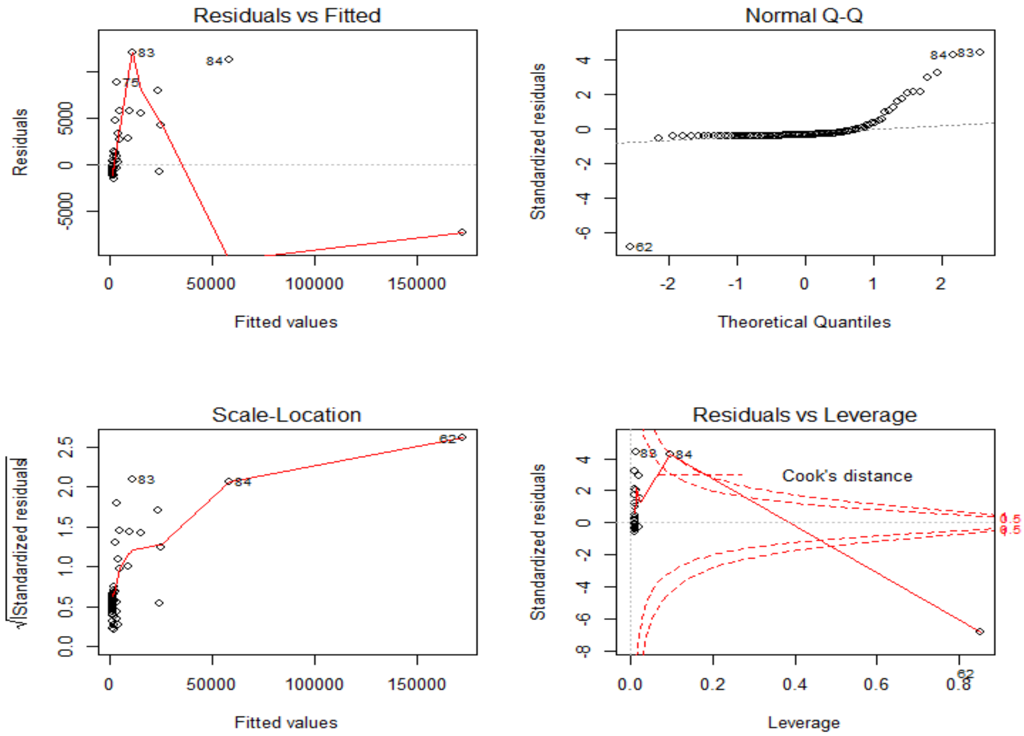

Density
0.15
0.10
0.05
0.00
2
4
6
8
10
12
logmetab

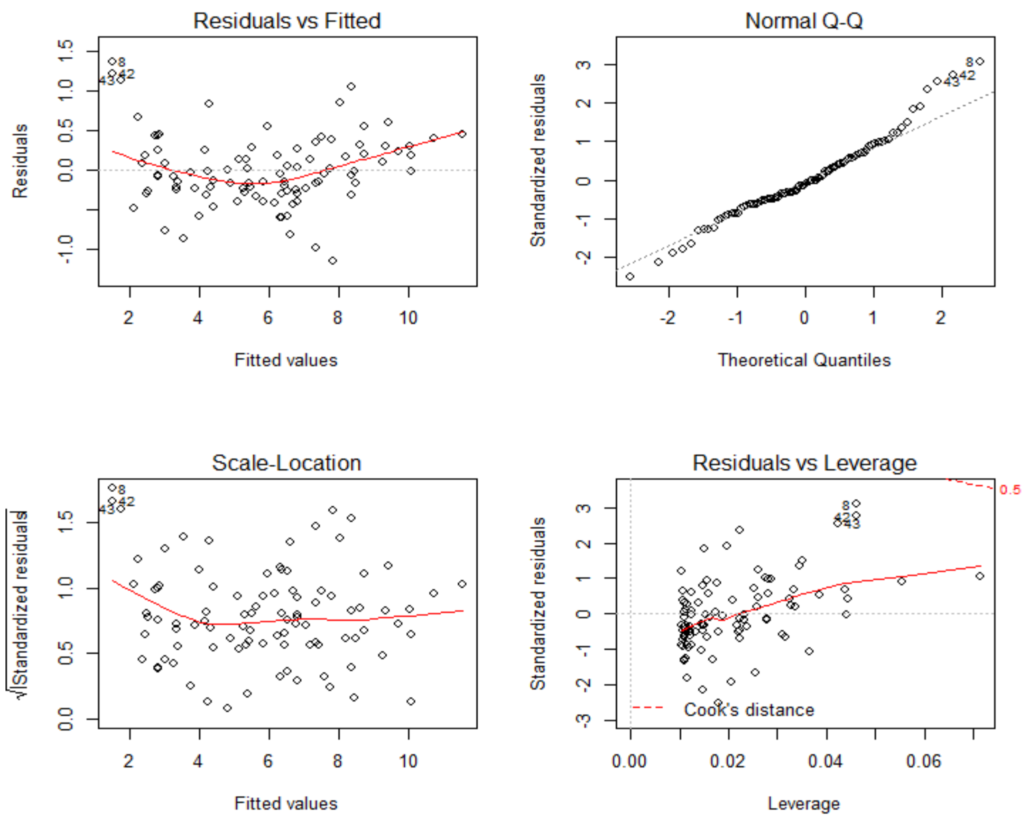

Density
0.005
0.004
0.003
0.002
0.001
0.000
0
500
1000
1500
2000
2500
3000
Mass

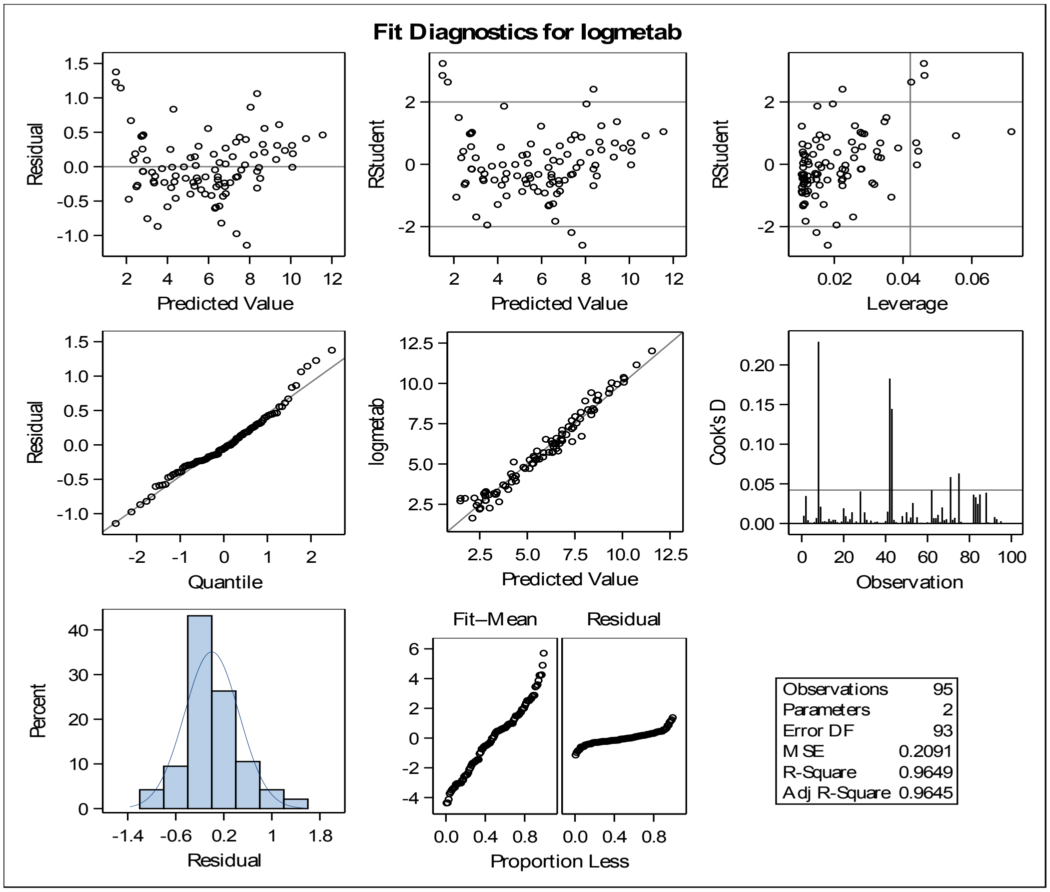

Density
4e-05
3e-05
2e-05
1e-05
0e+00
0
50000
100000
150000
Metab


Density
0.10
0.08
0.06
0.04
0.02
0.00
-10
-5
0
5
10
logmass

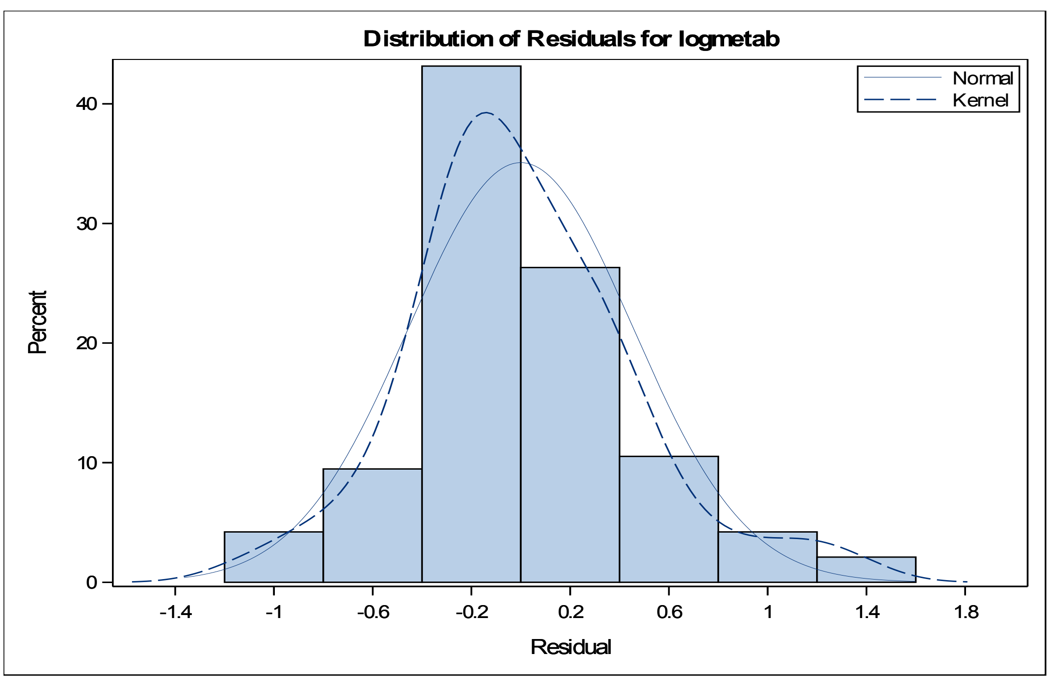

Density
0.15
0.10
0.05
0.00
0
5
10
15
logmetab

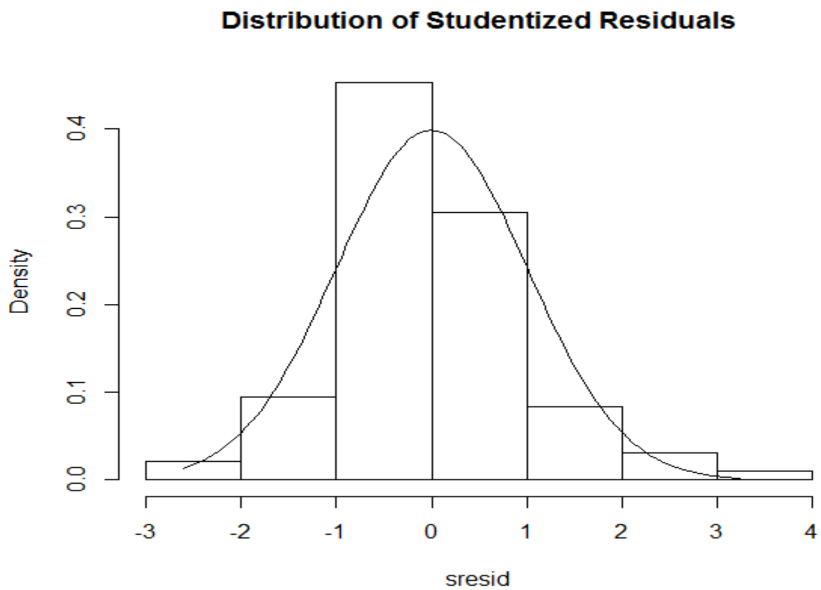

Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage



<u>Part vi) Residual Histograms with Superimposed Normal Distributions (SAS/R)</u>





Part vii) Variation in Y accounted for by X → R-squared (SAS/R)

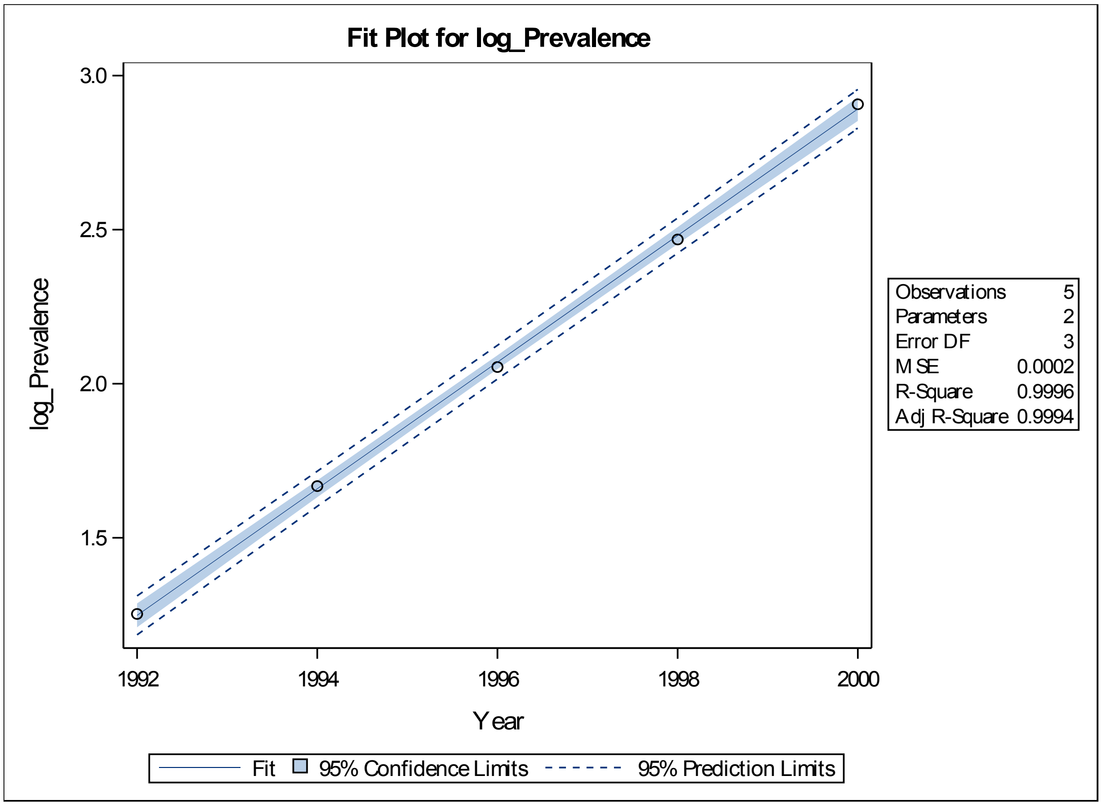
Root MSE	0.45723	**R-Square**	0.9649
Dependent Mean	5.84732	**Adj R-Sq**	0.9645
Coeff Var	7.81956		

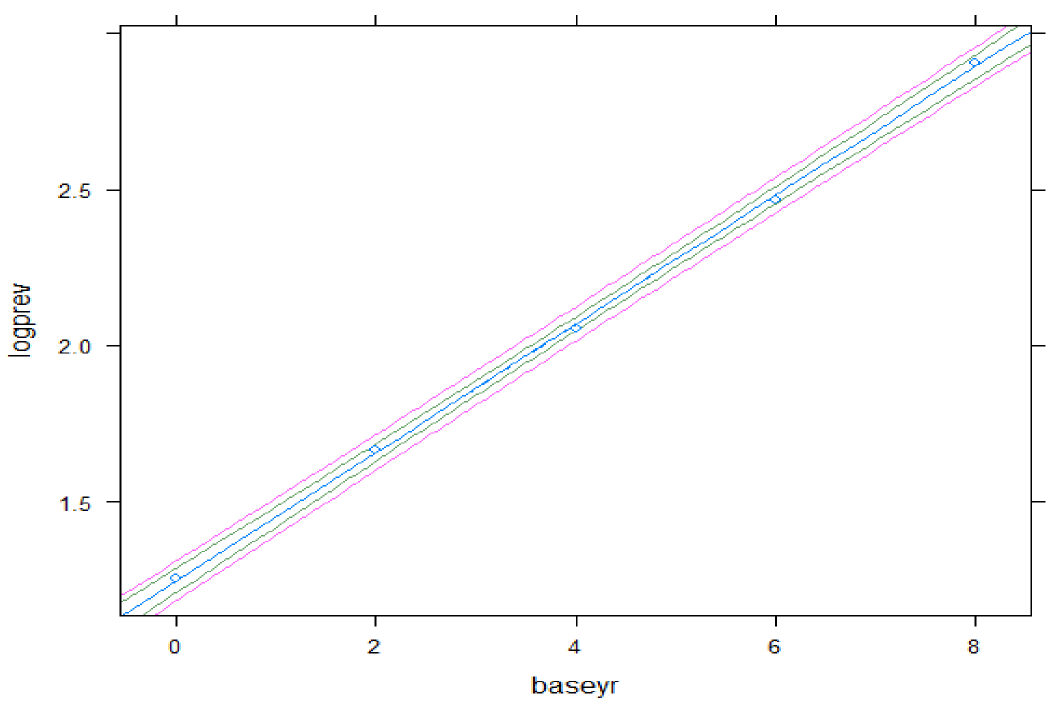
```
Call:
lm(formula = logmetab ~ logmass, data = ex0826)
...

Residual standard error: 0.457 on 93 degrees of freedom
Multiple R-squared:  0.965, Adjusted R-squared:  0.964
F-statistic: 2.55e+03 on 1 and 93 DF,  p-value: <2e-16

> SSR = sum(resid(logmetab_logmass)^2)
> SSR
[1] 19.44

> SSE = sum((fitted(logmetab_logmass) - mean(~logmetab, data = ex0826))^2)
> SSE
[1] 533.8

> R_sq = 1 - (SSR/(SSE + SSR))
> R_sq
[1] 0.9649
```

Approximately 96.49% of the variation in the response (log Metabolism; avg. basal metabolic rate in kJ/day) was explained by the linear regression on the explanatory variable (log Mass; avg. mass in kg.).

Question 2)

Part i) Scatterplot with Regression Confidence & Prediction Intervals (SAS/R)





Part ii) Tabular Regression T-stats and P-Values (SAS/R)

Parameter Estimates										
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS	Standardized Estimate	95% Confidence Limits	
Intercept	1	1.24819	0.01216	102.68	<.0001	21.42282	2.59661	0	1.20950	1.28687
Year2	1	0.20543	0.00248	82.79	<.0001	1.68812	1.68812	0.99978	0.19754	0.21333

```
Call:
lm(formula = logprev ~ baseyr, data = ex0829)

Residuals:
       1        2        3        4        5
 0.00458  0.00865 -0.01580 -0.01269  0.01525

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.24819    0.01216   102.7  2.0e-06 ***
baseyear     0.20543    0.00248    82.8  3.9e-06 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0157 on 3 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:  0.999
F-statistic: 6.85e+03 on 1 and 3 DF,  p-value: 3.88e-06
```

Part iii) Regression Equation

Log Transforming Y:

ln y-hat = 1.24819 + .2054x
y-hat = e^(1.24819) * (e^.2054)^x
y-hat = 3.484031 * (1.228016)^x

--> **when x is 0, y-hat = 3.484031**
--> **increasing x by 1, multiplies y-hat by 1.228016 or (e^.2054)**
--> **increasing x by n, multiplies y-hat by 1.228016^n**

Part iv) Model Interpretation

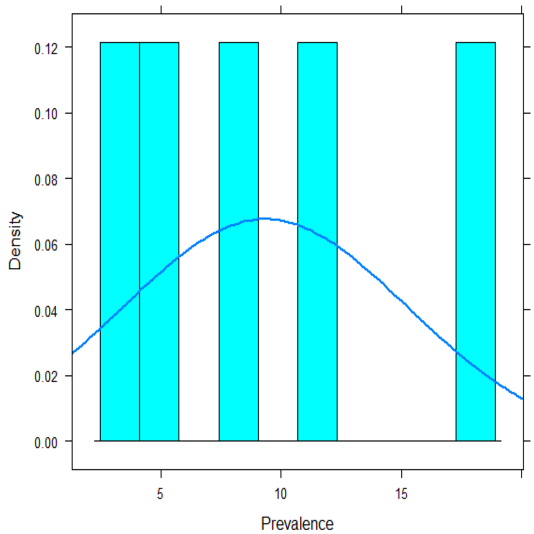
From 1992-2000, for every unit increase in x (Year), the prevalence of autism per 10,000 ten-year-olds increases by a factor of (e^.2054), or (1.228016).

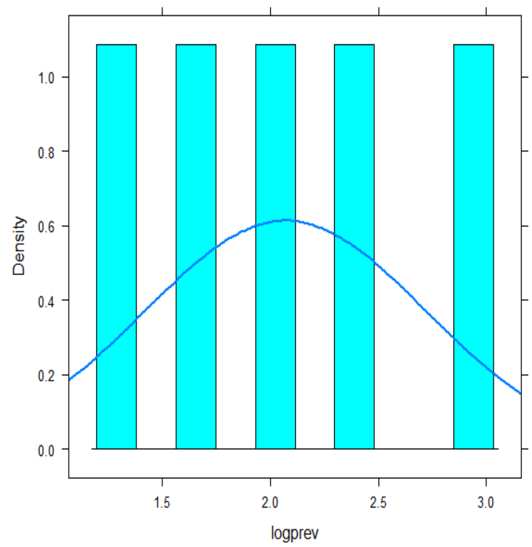
```
> exp(coef(logprev_baseyr))
(Intercept)      baseyr
      3.484       1.228
```

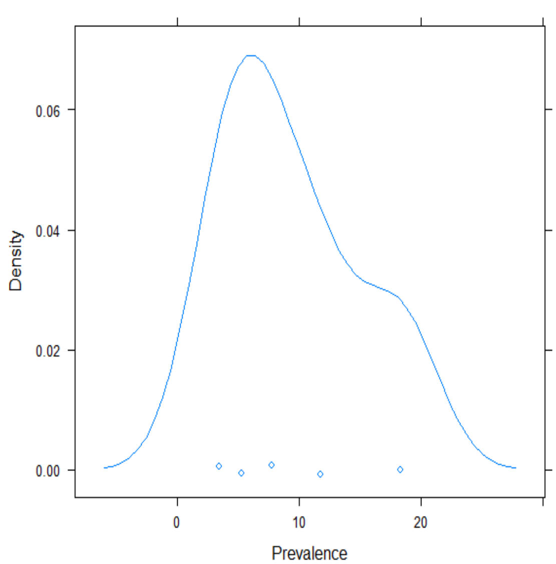
```
> exp(confint(logprev_baseyr))
               2.5 %      97.5 %
(Intercept) 3.352         3.621
baseyr      1.218         1.238
```

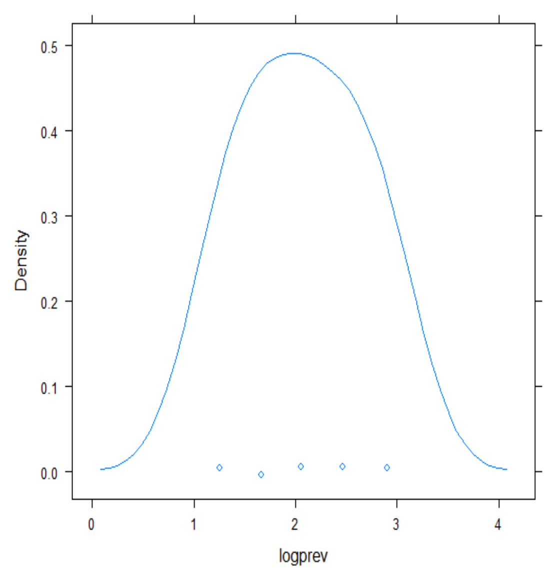
Part v) Residual Scatterplots (R/SAS)

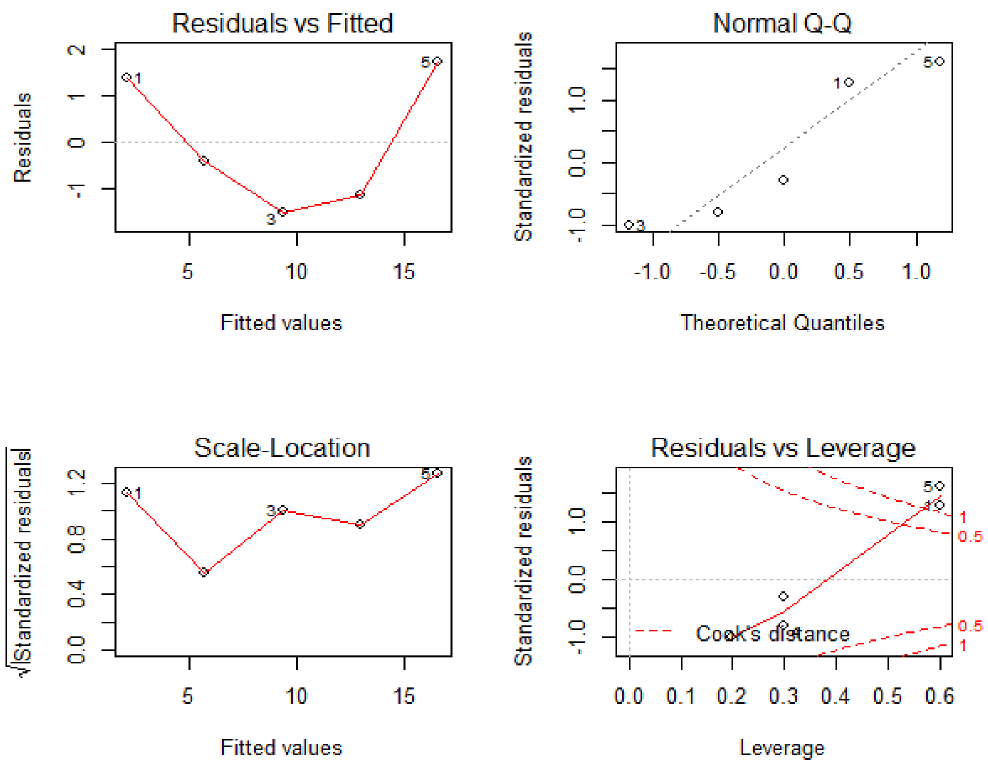
The log transformations of X and Y produce the following figures and residual scatterplots:

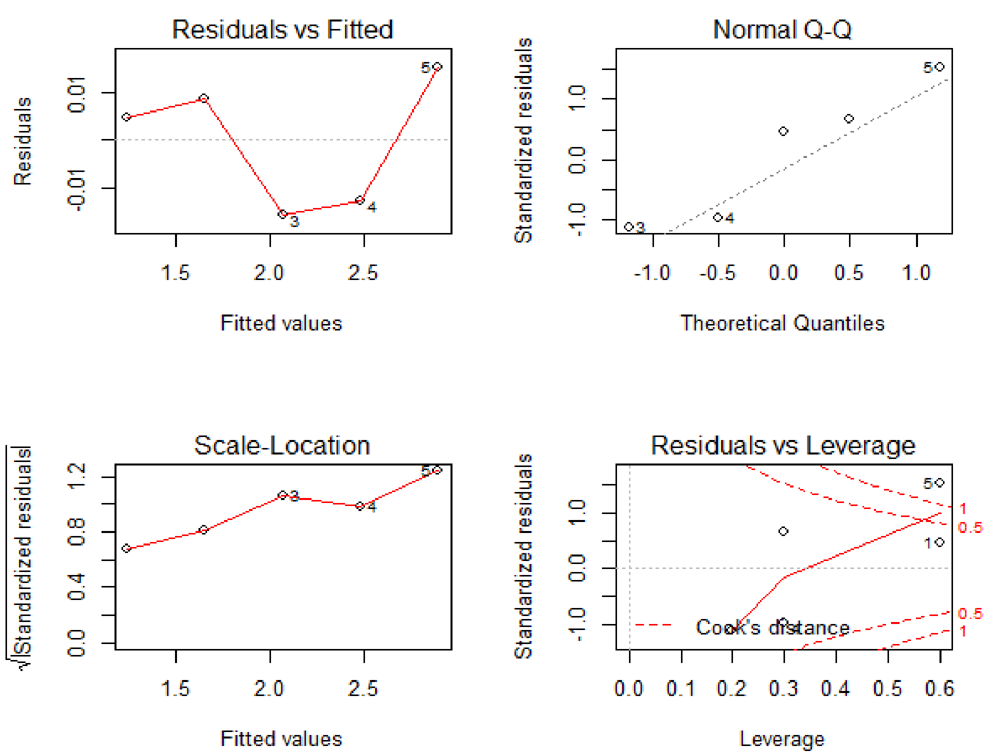


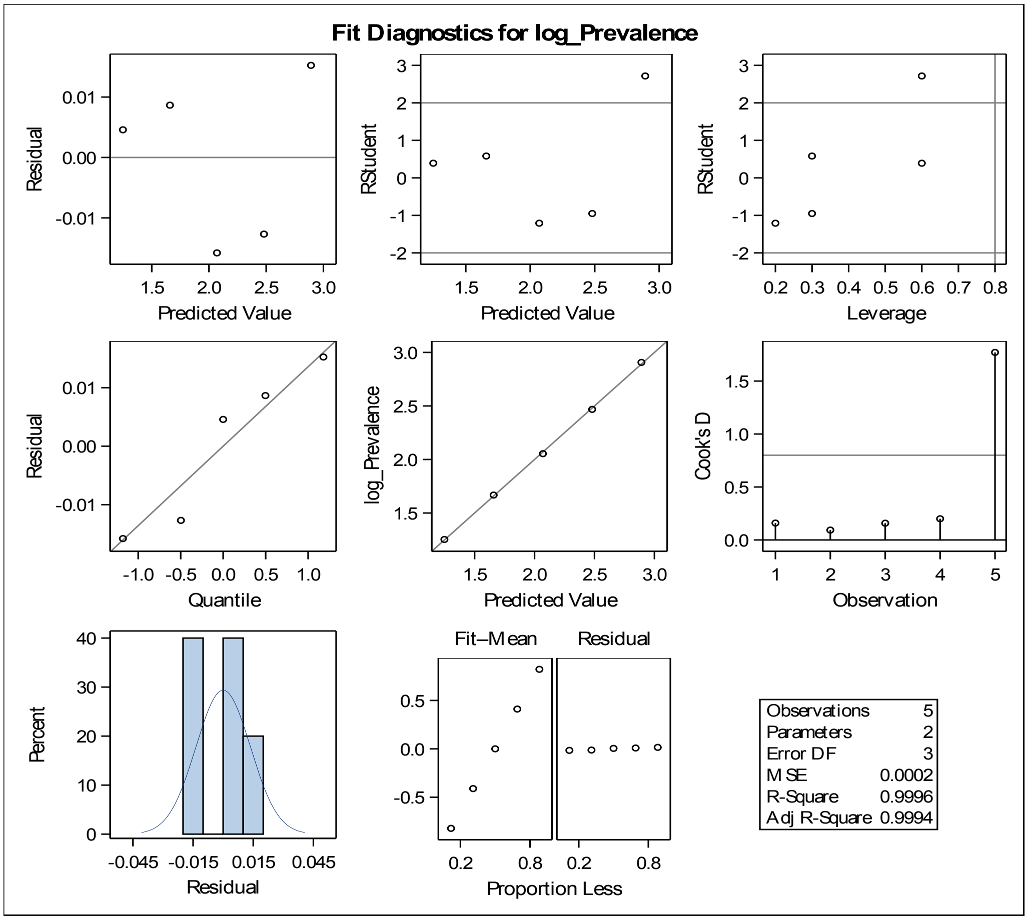




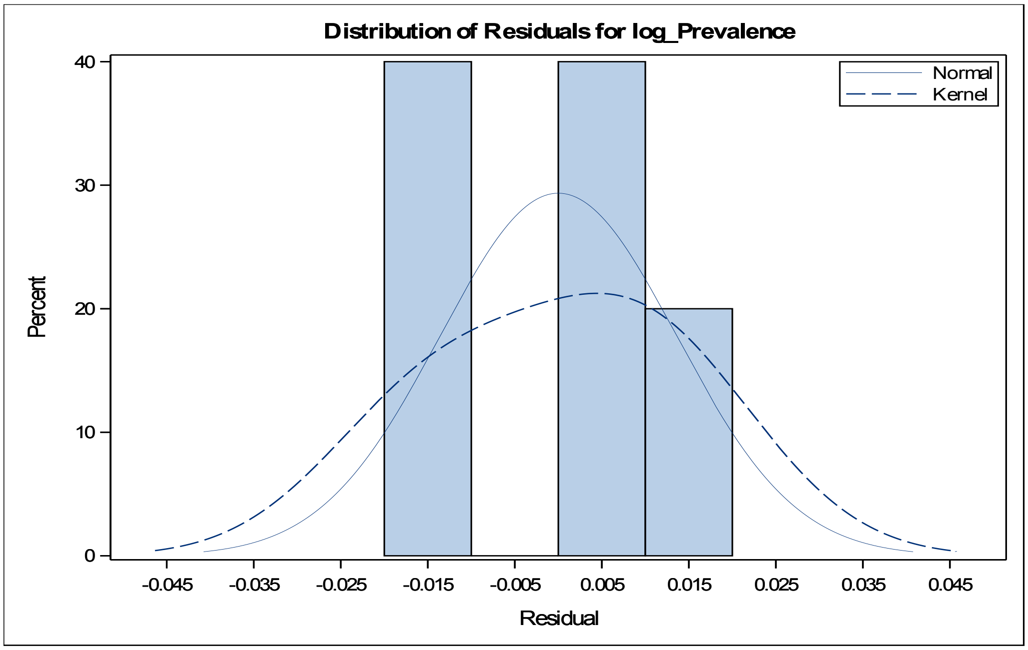


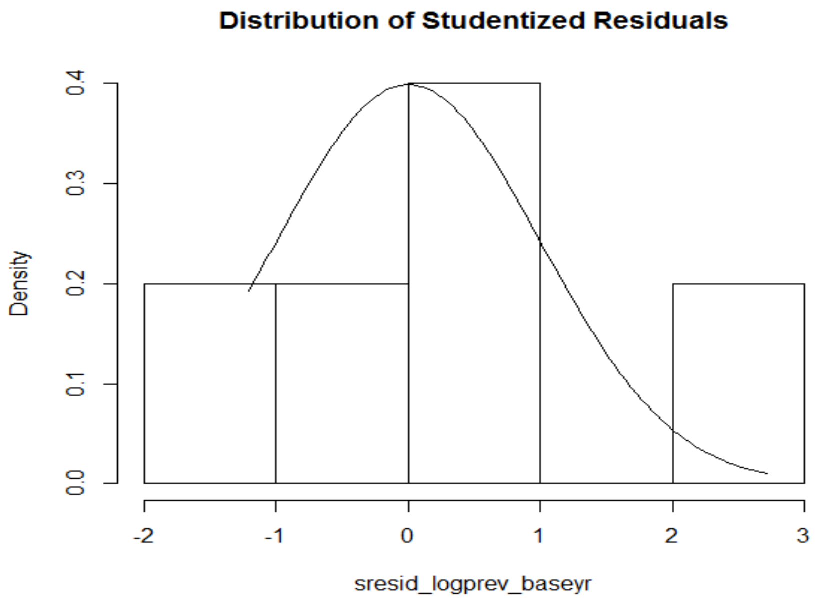

Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage



Part vi) Residual Histograms with Superimposed Normal Distributions (SAS/R)





Part vii) Variation in Y accounted for by X → R-squared (SAS/R)

Root MSE	0.01569	**R-Square**	0.9996
Dependent Mean	2.06992	**Adj R-Sq**	0.9994
Coeff Var	0.75814		

```
Call:
lm(formula = logprev ~ baseyr, data = ex0829)
…

Residual standard error: 0.0157 on 3 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:  0.999
F-statistic: 6.85e+03 on 1 and 3 DF,  p-value: 3.88e-06

> SSR_prevbase = sum(resid(logprev_baseyr)^2)
> SSR_prevbase
[1] 0.0007388

> SSE_prevbase = sum((fitted(logprev_baseyr) - mean(~logprev, data = ex0829))^2)
> SSE_prevbase
[1] 1.688

> R_sq_prevbase = 1 - (SSR_prevbase/(SSE_prevbase + SSR_prevbase))
> R_sq_prevbase
[1] 0.9996
```

Approximately 99.96% of the variation in the response (log Prevalence per 10,000 ten-year-old children) was explained by the linear regression on the explanatory variable (Year; 1992-2000; base year = 1992)

SAS Code:

Question 1)

```
FILENAME REFFILE '/home/jrasmusvorrath0/ex0826.csv';

PROC IMPORT DATAFILE=REFFILE
DBMS=CSV
OUT=WORK.IMPORT;
GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.IMPORT; RUN;

data metabolism; set work.import; run;

proc print data = metabolism; run;

proc reg data = metabolism;
model Metab = Mass / ss1 ss2 clb stb r cli clm;
run; quit;

data log_metab; set metabolism;                    *Log Transform (X), (Y)
logmetab = log(Metab);
logmass = log(Mass);
run;

proc print data = log_metab; run;

proc reg data = log_metab;
model logmetab = logmass / ss1 ss2 clb stb r cli clm;
run; quit;

*proc glm data = log_metab;                        *Validation by Alternative Procedure
*model logmetab = logmass / clparm;
*output out = log_metab_resid residual = Residuals;
*run; quit;
```

Question 2)

```
FILENAME REFFILE '/home/jrasmusvorrath0/ex0829.csv';

PROC IMPORT DATAFILE=REFFILE
DBMS=CSV
OUT=WORK.IMPORT1;
GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.IMPORT1; RUN;

data autism; set work.import1; run;

proc print data = autism; run;

proc reg data = autism;
model Prevalence = Year / ss1 ss2 clb stb r cli clm;
run; quit;

data log_autism; set autism;                    *Log Transform (Y)
log_Prevalence = log(Prevalence);
run;

proc print data = log_autism; run;

proc reg data = log_autism;
model log_Prevalence = Year / ss1 ss2 clb stb r cli clm;
run; quit;

data log_autism2; set log_autism;          *Adjusting X-axis Values for an Interpretable Intercept
input Year2 Prevalence;
datalines;
0 3.5
2 5.3
4 7.8
6 11.8
8 18.3
;
run;

proc print data = log_autism2; run;

proc reg data = log_autism2;
model log_Prevalence = Year2 / ss1 ss2 clb stb r cli clm;
run; quit;

*proc transreg data = autism;                   *Validation by Alternative Procedure
*model log(Prevalence) = linear(Year) / CI;
*output out = autism_transreg CLI CLM Coefficients Predicted Residuals;
*run;

*proc print data = autism_transreg; run;
```

R Code:

Question 1)

```
> require(Sleuth 3)
> require(mosaic)
> options(digits = 4)

> summary(ex0826)                                                    *Exploratory Data Analysis
> xyplot(Metab ~ Mass, type = c("p", "r"), data = ex0826)
> metab_mass <- lm(Metab ~ Mass, data = ex0826)
> summary(metab_mass)
> anova(metab_mass)
> par(mfrow = c(2, 2))
> plot(metab_mass)

> par(mfrow = c(1, 1))                                               *Log Transform (X)
> densityplot(~Mass, data = ex0826)
> histogram(~Mass, type = "density", density = TRUE, nint = 10, data = ex0826)
> ex0826$logmass = with(ex0826, log(Mass))
> densityplot(~logmass, data = ex0826)
> histogram(~logmass, type = "density", density = TRUE, nint = 10, data = ex0826)

> metab_logmass <- lm(Metab ~ logmass, data = ex0826)
> summary(metab_logmass)

> densityplot(~Metab, data = ex0826)                                 *Log Transform (Y)
> histogram(~Metab, type = "density", density = TRUE, nint = 10, data = ex0826)
> ex0826$logmetab = with(ex0826, log(Metab))
> densityplot(~logmetab, data = ex0826)
> histogram(~logmetab, type = 'density', density = TRUE, nint = 10, data = ex0826)

> logmetab_logmass <- lm(logmetab ~ logmass, data = ex0826)          *Regression & Residual Plot
> summary(logmetab_logmass)
> anova(logmetab_logmass)
> xyplot(logmetab ~ logmass, panel = panel.lmbands, data = ex0826)
> par (mfrow = c(2, 2))
> plot(logmetab_logmass)

> coef(logmetab_logmass)                                             *Back-transform Coefficient Values & CIs
> exp(coef(logmetab_logmass))
> confint(logmetab_logmass)
> exp(confint(logmetab_logmass))

> par(mfrow = c(1,1))                                                *Studentized Residual Histogram
> library(MASS)
> sresid <- studres(logmetab_logmass)
> hist(sresid, freq = FALSE, main = "Distribution of Studentized Residuals")
> xfit <- seq(min(sresid), max(sresid), length = 100)
> yfit <- dnorm(xfit)
> lines(xfit, yfit)


> SSR = sum(resid(logmetab_logmass)^2)                               *R-Squared Calculation
> SSE = sum((fitted(logmetab_logmass) - mean(~logmetab, data = ex0826))^2)
> R_sq = 1 - (SSR/(SSE + SSR))
```

Question 2)

```
> require(Sleuth3)
> require(mosaic)
> options(digits = 4)
                                                                    *Exploratory Data Analysis
> summary(ex0829)
> xyplot(Prevalence ~ Year, type = c("p", "r"), data = ex0829)
> prev_yr <- lm(Prevalence ~ Year, data = ex0829)
> summary(prev_yr)
> anova(prev_yr)
> par(mfrow = c(2, 2))
> plot(prev_yr)

> par(mfrow = c(1, 1))                                              *Log Transform (Y)
> densityplot(~Prevalence, data = ex0829)
> histogram(~Prevalence, type = "density", density = TRUE, nint = 10, data = ex0829)
> ex0829$logprev = with(ex0829, log(Prevalence))
> densityplot(~logprev, data = ex0829)
> histogram(~logprev, type = 'density', density = TRUE, nint = 10, data = ex0829)

> ex0829$baseyr = with(ex0829, c(0, 2, 4, 6, 8))                    *Adjust X for 1992 Base Year

> logprev_baseyr <- lm(logprev ~ baseyr, data = ex0829)             *Regression & Residual Plot
> summary(logprev_baseyr)
> anova(logprev_baseyr)
> xyplot(logprev ~ baseyr, panel = panel.lmbands, data = ex0829)
> par(mfrow = c(2, 2))
> plot(logprev_baseyr)

> coef(logprev_baseyr)                                              *Back-transform Coefficient Values & CIs
> exp(coef(logprev_baseyr))
> confint(logprev_baseyr)
> exp(confint(logprev_baseyr))

> par(mfrow = c(1, 1))                                              *Studentized Residual Histogram
> library(MASS)
> sresid_logprev_baseyr  <- studres(logprev_baseyr)
> hist(sresid_logprev_baseyr, freq = FALSE, main = "Distribution of Studentized Residuals")
> xfit_logprev_baseyr <- seq(min(sresid_logprev_baseyr), max(sresid_logprev_baseyr), length = 100)
> yfit_logprev_baseyr <- dnorm(xfit_logprev_baseyr)
> lines(xfit_logprev_baseyr, yfit_logprev_baseyr)

> SSR_prevbase = sum(resid(logprev_baseyr)^2)                       *R-Squared Calculation
> SSE_prevbase = sum((fitted(logprev_baseyr) - mean(~logprev, data = ex0829))^2)
> R_sq_prevbase = 1 - (SSR_prevbase/(SSE_prevbase + SSR_prevbase))
```