Unit 11 HW Solutions

## Question 1 (52 points total)

From Problem 26, Chapter 8:

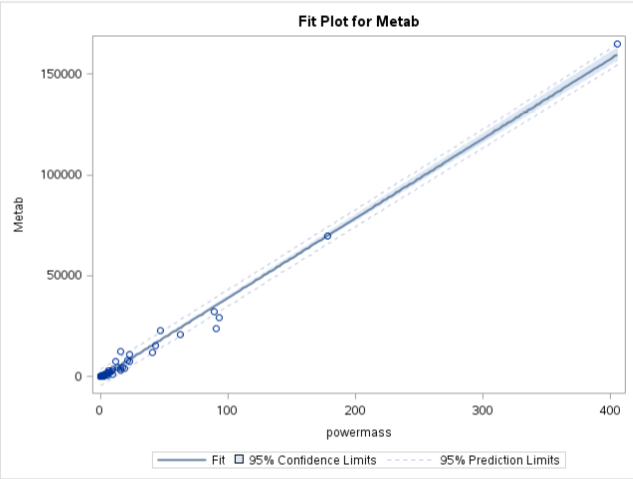
The Metabolic data set has the average mass, metabolic rate, and average lifespan of 95 different species of mammals. Kleiber’s Law states that the metabolic rate of an animal species, on average, is proportional to its mass raised to the power ¾. Judge the adequacy of this theory with these data. Ultimately, 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 3/4.

* 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:

### Part A (4 points)

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 3/4? (Recall that if some variable y is proportional to the variable x, then (with nonzero m) is a well-fitting model.) In other words, does the data (metabolic rate, ) reasonably fall along a straight (nonhorizontal) line and nearly pass through the origin?



**The scatterplot looks linear and nearly crosses through the origin.**

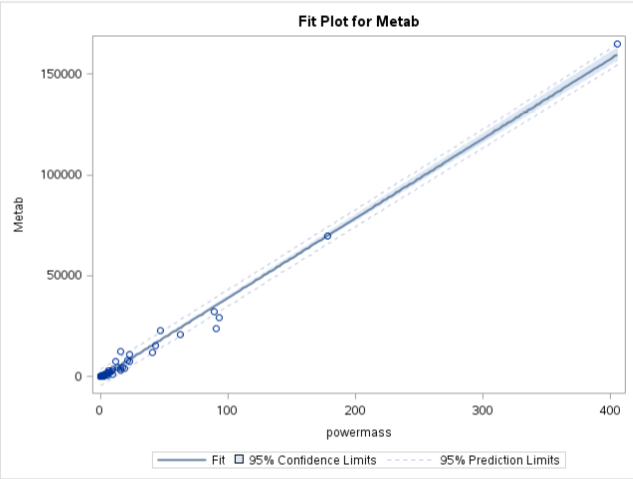
### Part B (12 points total)

We want to find the “best” model to predict metabolic rate from and make appropriate statistical inferences. Therefore, address all the assumptions prior to the analysis (using ). 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 “” in the model; do not go back to regular “mass,” although 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. (2 points)

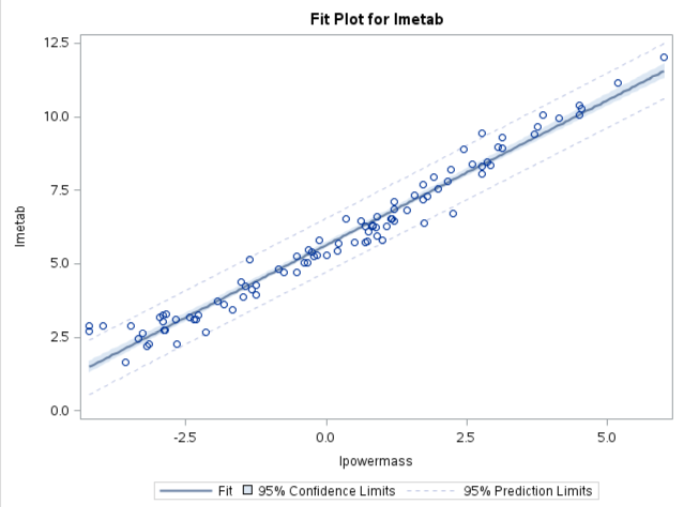
A scatterplot with the following included on the graph: regression line, confidence intervals of the regression line, and prediction intervals of the regression line.

**Original data:**



**Looks linear.**

**Log-log data:**

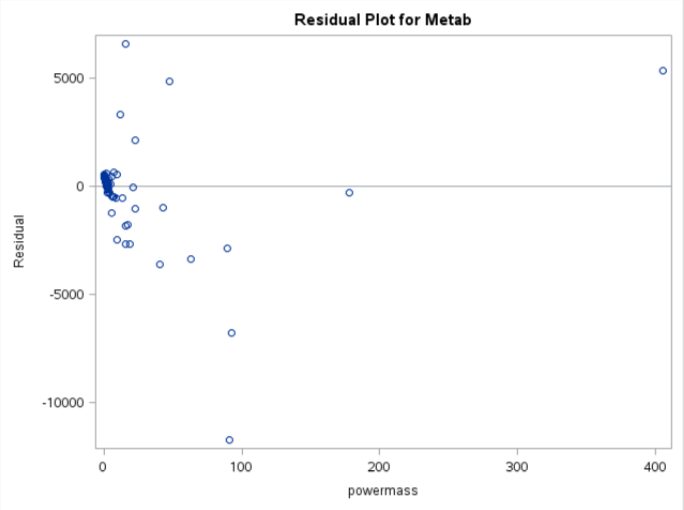


**Data still looks linear, but more spread out along the x-axis.**

#### ii. (2 points)

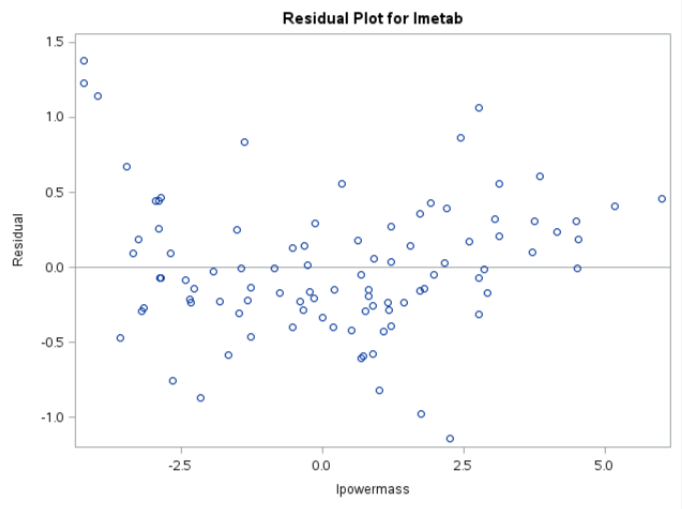
A scatterplot of residuals.

**Original data:**



**Data is very clustered around low values of mass. The residuals definitely don’t resemble a random cloud.**

**Log-log data:**

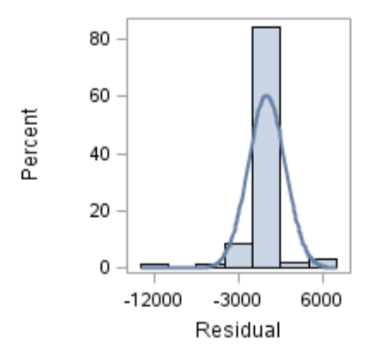


**The data is much more spread out. The residuals look equally spread out in a random cloud.**

#### iii. (2 points)

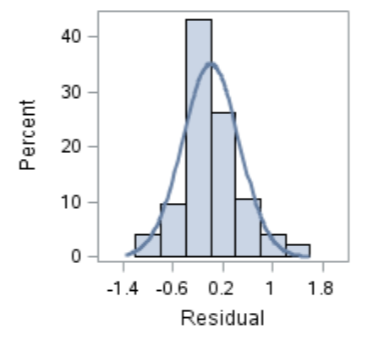
A histogram of residuals with the normal distribution superimposed.

**Original data:**



**The residuals don’t look very non-normal, but there may be a slight skew.**

**Log-log data:**



**The residuals look much more normally distributed.**

#### iv. (6 points total)

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

**(1 point) Linearity: Met with original and log-log model. (view scatterplots)**

**(1 point) Normality: Log-log model looks slightly better. (view histograms)**

**(1 point) Equal standard deviations: Log-log model looks much better. (view residual scatter plots)**

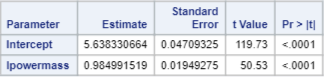
**(1 point) Independence: We will assume independence, although not much is known about how these species were chosen.**

**(1 point) Outliers: There are no major outliers to exclude.**

**(1 point) We will proceed to make inferences on a log-log model.**

### Part C (9 points)

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 and .



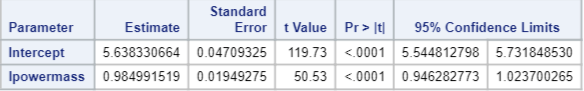
### Part D (9 points)

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

### Part E (9 points)

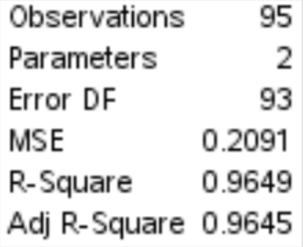
Interpretation of the model, paying special attention if you used a transformation (hint!). That is, interpret the slope as well as the **confidence interval**.

**When is doubled, the median increases by a factor of . In other words, when is doubled, the median increases by 98%. A 95% confidence interval for this is or (93%, 103%) for the percent increase.**



### Part F (9 points)

**96.49% of the variability in metabolic rate can be accounted for by the mass.**



SAS Code:  
\*To import data;  
FILENAME REFFILE ‘/home/sadiet0/Metabolism Data Prob 26.csv’;  
PROC IMPORT DATAFILE=REFFILE  
DBMS=CSV  
OUT=metabolism;  
GETNAMES=YES;  
RUN;

proc print data = metabolism;  
run;

\*To raise mass to the 3/4 power;  
data metabolism;  
set metabolism;  
powermass = mass\*\*0.75;  
run;

proc print data = metabolism;  
run;

\*To check assumptions;  
proc glm data = metabolism plots = all;  
model metab = powermass/solution;  
run;

\*To log the powermass and metab;  
data metabolism;  
set metabolism;  
lpowermass=log(powermass);  
lmetab=log(metab);  
run;

\*To check assumptions and perform regression with log-log data;  
proc glm data = metabolism plots = all;  
model lmetab = lpowermass/solution clparm;  
run;

## Question 2 (48 points total)

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:

### Part A (12 points total)

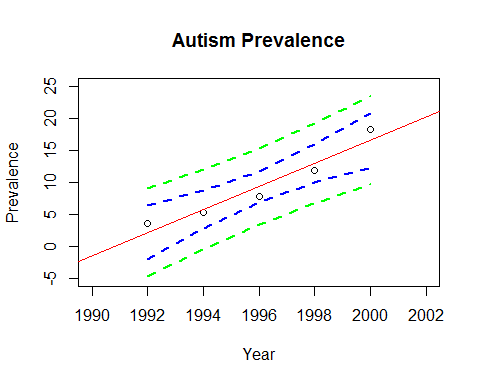
Address all the assumptions for a linear regression model prior to the analysis. If the assumptions are not met, handle the data appropriately. If a transformation is used, address the assumptions again with the transformed data to ensure that the transformation is logical. The questions below should 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. (2 points)

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:**

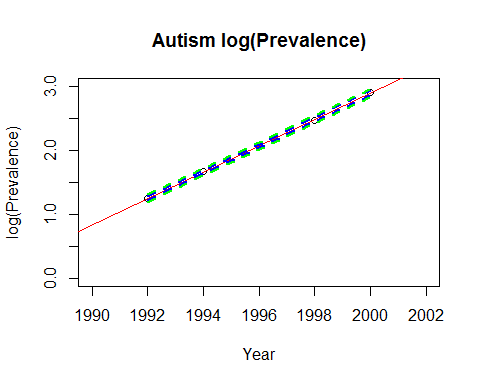
#CI's & PI's with original data & regression line  
autism <- read.csv("C:/Users/Charles/Documents/SMU/Online Teaching/MSDS 6371 - Statistical Foundations for Data Science/UNIT 11/Autism Data Prob 29.csv")  
  
autism.lm <- lm(Prevalence ~ Year, data = autism)  
newx <- autism$Year  
newx <- sort(newx)  
prd\_c <- predict(autism.lm, newdata= data.frame(Year = newx), interval=c("confidence"),   
 type = c("response"), level=0.95)  
prd\_p <- predict(autism.lm, newdata= data.frame(Year = newx), interval=c("prediction"),   
 type = c("response"), level=0.95)  
  
#Plot with confidence and prediction intervals  
plot(autism[,1], autism[,2],xlim = c(1990,2002), ylim = c(-5,25), xlab = "Year",   
 ylab = "Prevalence", main = "Autism Prevalence")  
abline(autism.lm, col = "red")  
lines(newx,prd\_c[,2],col = "blue",lty = 2, lwd = 2)  
lines(newx,prd\_c[,3],col = "blue", lty = 2, lwd = 2)  
lines(newx,prd\_p[,2],col = "green", lty = 2, lwd = 2)  
lines(newx,prd\_p[,3],col = "green", lty = 2, lwd = 2)



**Data looks curved, non-linear.**

**Log-linear data:**

##Transform prevalence  
autism$log.prevalence <- log(autism$Prevalence)  
  
#New linear model, PI/CI  
  
log.autism.lm <- lm(log.prevalence ~ Year, data = autism)  
prd\_c2 <- predict(log.autism.lm, newdata= data.frame(Year = newx), interval=c("confidence"),   
 type = c("response"), level=0.95)  
prd\_p2 <- predict(log.autism.lm, newdata= data.frame(Year = newx), interval=c("prediction"),   
 type = c("response"), level=0.95)  
  
#Plot with confidence and prediction intervals  
plot(autism[,1], autism[,3],xlim = c(1990,2002), ylim = c(0,3), xlab = "Year",   
 ylab = "log(Prevalence)", main = "Autism log(Prevalence)")  
abline(log.autism.lm, col = "red")  
lines(newx,prd\_c2[,2],col = "blue",lty = 2, lwd = 2)  
lines(newx,prd\_c2[,3],col = "blue", lty = 2, lwd = 2)  
lines(newx,prd\_p2[,2],col = "green", lty = 2, lwd = 2)  
lines(newx,prd\_p2[,3],col = "green", lty = 2, lwd = 2)



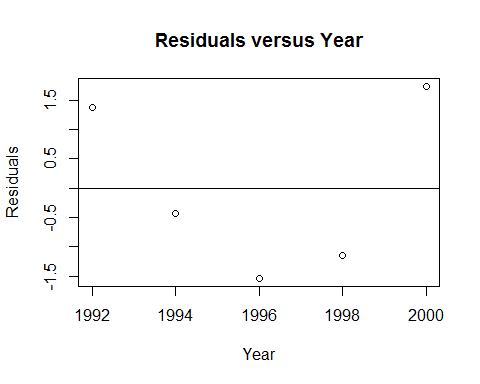
**The data is much more linear.**

#### ii. (2 points)

A scatterplot of residuals.

**Original data:**

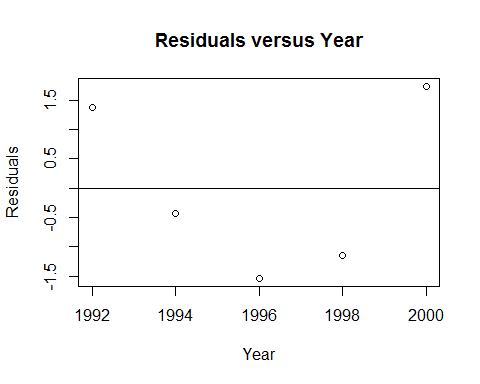
#Store residuals from linear model  
autismres <- resid(autism.lm)  
  
#Get scatterplot of residuals  
plot(autism$Year, autismres, ylab="Residuals", xlab="Year", main="Residuals versus Year")  
abline (0,0)



**Data looks very non-linear.**

**Log-linear data:**

#Store residuals from linear model  
autismres2 <- resid(log.autism.lm)  
  
#Get scatterplot of residuals  
plot(autism$Year, autismres, ylab="Residuals", xlab="Year", main="Residuals versus Year")  
abline (0,0)



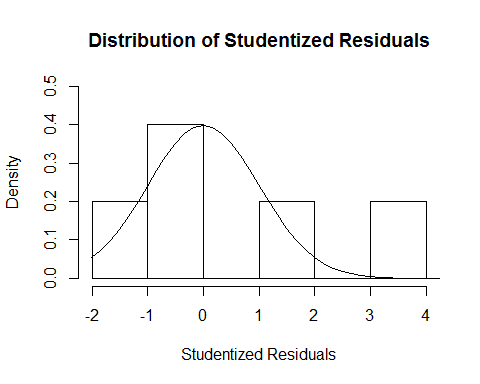
**There is much less of a pattern to these residuals.**

#### iii. (2 points)

A histogram of residuals with the normal distribution superimposed.

**Original data:**

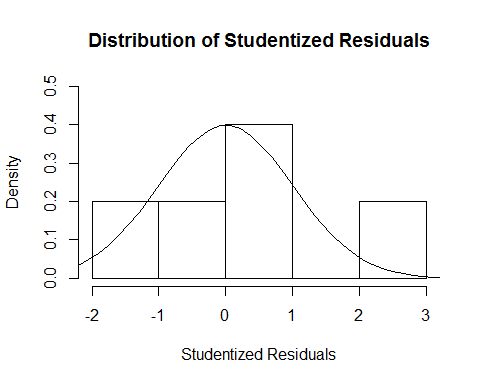
##Store studentized residuals  
studresautism <- rstudent(autism.lm)  
  
##Histogram  
hist(studresautism, freq=FALSE, main="Distribution of Studentized Residuals",  
xlab="Studentized Residuals", ylab="Density", ylim=c(0,0.5))  
  
##Create range of x-values for normal curve  
xfit <- seq(min(studresautism)-1, max(studresautism)+1, length=40)  
  
##Generate values from the normal distribution at the specified values  
yfit <- (dnorm(xfit))  
  
##Add the normal curve  
lines(xfit, yfit, ylim=c(0,0.5))



**Very difficult to check normality with five data points.**

**Log-linear data:**

##Store studentized residuals  
studresautism2 <- rstudent(log.autism.lm)  
  
##Histogram  
hist(studresautism2, freq=FALSE, main="Distribution of Studentized Residuals",  
xlab="Studentized Residuals", ylab="Density", ylim=c(0,0.5))  
  
##Create range of x-values for normal curve  
xfit2 <- seq(min(studresautism2)-1, max(studresautism2)+1, length=40)  
  
##Generate values from the normal distribution at the specified values  
yfit2 <- (dnorm(xfit2))  
  
##Add the normal curve  
lines(xfit2, yfit2, ylim=c(0,0.5))

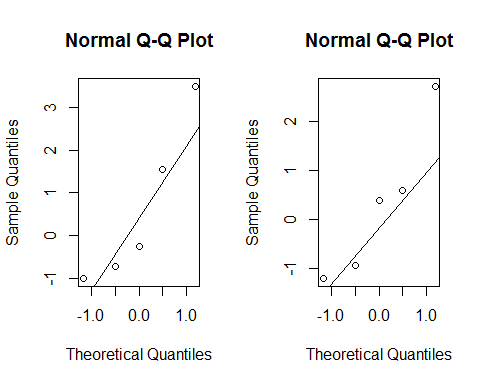


**The data looks slightly more normal, but it is very hard to tell with five data points.**

#### iv. (6 points total)

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

##Check q-q plots as well  
par(mfrow=c(1,2))  
qqnorm(studresautism)  
qqline(studresautism)  
qqnorm(studresautism2)  
qqline(studresautism2)



**For the original data (left), there is curvature suggesting non-normality (even with the small data set). In the log-linear plot (right), there is less curvature.**

**(1 point) Linearity: Met with log-linear model. (view scatterplots)**

**(1 point) Normality: Log-linear model looks better. (view histograms, q-q plots)**

**(1 point) Equal standard deviations: Neither original nor log-linear model looks like the equal standard deviations assumption is not met. It’s very hard to tell with 5 data points (view residual scatter plots)**

**(1 point) Independence: We will assume independence, although this is unlikely given the time series nature of the data.**

**(1 point) Outliers: There are no major outliers to exclude.**

**(1 point) We will proceed to make inferences on a log-linear model.**

**Note that the log-log model looks appropriate here as well. However, all things being equal, the log-linear model may be preferred given the “for every one year increase” interpretation of the slope rather than the “for every doubling of time” interpretation of the slope.**

### Part B (9 points)

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 and .

summary(log.autism.lm)

##   
## Call:  
## lm(formula = log.prevalence ~ Year, data = autism)  
##   
## Residuals:  
## 1 2 3 4 5   
## 0.004578 0.008655 -0.015795 -0.012686 0.015248   
##   
## 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 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## 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

### Part C (9 points)

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

### Part D (9 points)

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

**After each subsequent year, there is an approximate multiplicative change of in the median. That is, it is estimated that each year the median prevalence increases by 23%. A 95% confidence interval for the multiplicative increase in median is , or an increase of (22%, 24%).**

### Part E (9 points)

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

summary(log.autism.lm)$r.squared

## [1] 0.9995625

**About 99.96% of the variation in lPrevalence is accounted for by the year.**

## Bonus (+8 points total)

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:

### Part A (+1 point)

Hypotheses

**: Linear regression performs as well as the separate means model. (Or linear regression shows no lack of fit with respect to the separate means model.)**

**: Linear regression performs worse than the separate means model. (Or linear regression shows lack of fit with respect to the separate means model.)**

### Part B (+4 points)

The ANOVA table you created.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | DF | Sum of Squares | Mean Squares | F-statistic | p-Value |
| Model | 3 | 0.00953 | 0.003177 | 0.356 | 0.788 |
| Error (SM) | 0.0446 | 0.00892 |  |  |  |
| Total (LR) | 0.05413 |  |  |  |  |

### Part C (+1 point)

Decision.

**Fail to reject .**

### Part D (+1 point)

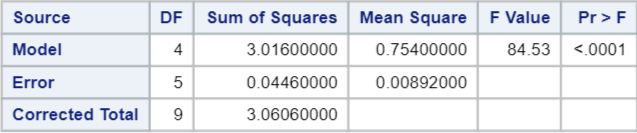
Conclusion in non-statistical terms.

**There is insufficient evidence to support the claim that the linear regression model shows lack of fit with respect to the separate means model (p-value = 0.788 from a lack of fit test).**

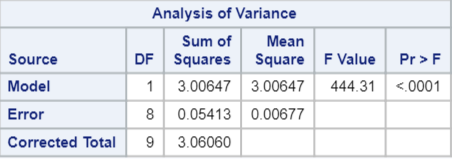
### Part E (+1 point)

Code and relevant output.

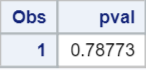
**ANOVA:**



**Regression:**



**p-Value:**



SAS Code:  
\*To bring in the data;  
data steer;  
input time 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  
;  
run;

proc print data = steer;  
run;

\*To get log of time;  
data steer;  
set steer;  
ltime=log(time);  
run;

\*To perform regression on transformed data;  
proc reg data = steer;  
model pH=ltime;  
run;

\*To perform ANOVA on transformed data;  
proc glm data = steer;  
class ltime;  
model pH=ltime;  
run;

\*To get p-value for ANOVA (lack of fit) table;  
data mypval;  
pval=1-probf(0.356, 3, 5);  
run;

proc print data = mypval;  
run;

time <- c(1,1,2,2,4,4,6,6,8,8)  
ph <- c(7.02, 6.93, 6.42, 6.51, 6.07, 5.99, 5.59, 5.80, 5.51, 5.36)  
ltime <- log(time)  
steer <- data.frame(ltime, ph, time)  
  
#To get separate means model ANOVA table  
#Note that we have to treat log time as a factor or  
#character variable otherwise R will treat it as continuous  
steer.aov <- aov(ph ~ as.factor(ltime), data = steer)  
summary(steer.aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## as.factor(ltime) 4 3.0160 0.7540 84.53 8.88e-05 \*\*\*  
## Residuals 5 0.0446 0.0089   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#To get regression table  
steer.lm <- lm(ph ~ ltime, data = steer)  
summary(aov(steer.lm))

## Df Sum Sq Mean Sq F value Pr(>F)   
## ltime 1 3.0065 3.0065 444.3 2.7e-08 \*\*\*  
## Residuals 8 0.0541 0.0068   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

pf(0.356, 3, 5, lower.tail=F)

## [1] 0.7877274