Unit 5 – Regression and Correlation
Practice Problems (1 of 3)
Solutions

Download from the course website.
simplelinear.xlsx

# 1.
This exercise gives you practice doing a simple linear regression using **simplelinear.xlsx.** This data set has n=31 observations of boiling points (Y=boiling) and temperature (X=temp). You will be exploring the following two simple linear models:

(i) $Y = b_0 + b_1 X$; where Y=boiling and X=temp

(ii) $newy = b_0 + b_1 X$; where $newy = 100*\log_{10}(y)$ and where y=boiling and X=temp

1a. Create a new variable newy = $100*\log_{10}$(boiling)

1b. For each model, obtain:

i. The fitted line estimates of $\hat{b}_0$ and $\hat{b}_1$
ii. Analysis of variance table
iii. R^2 = % of the variability in the outcome explained by the fitted line
iv. Scatter plot with overlay of fitted line

1c. In 3-5 sentences, write a one-paragraph interpretation of your two model fits.

Art of Stat Users

1a. Create a new variable newy = $100*\log_{10}$(boiling)

1b. For each model, obtain:

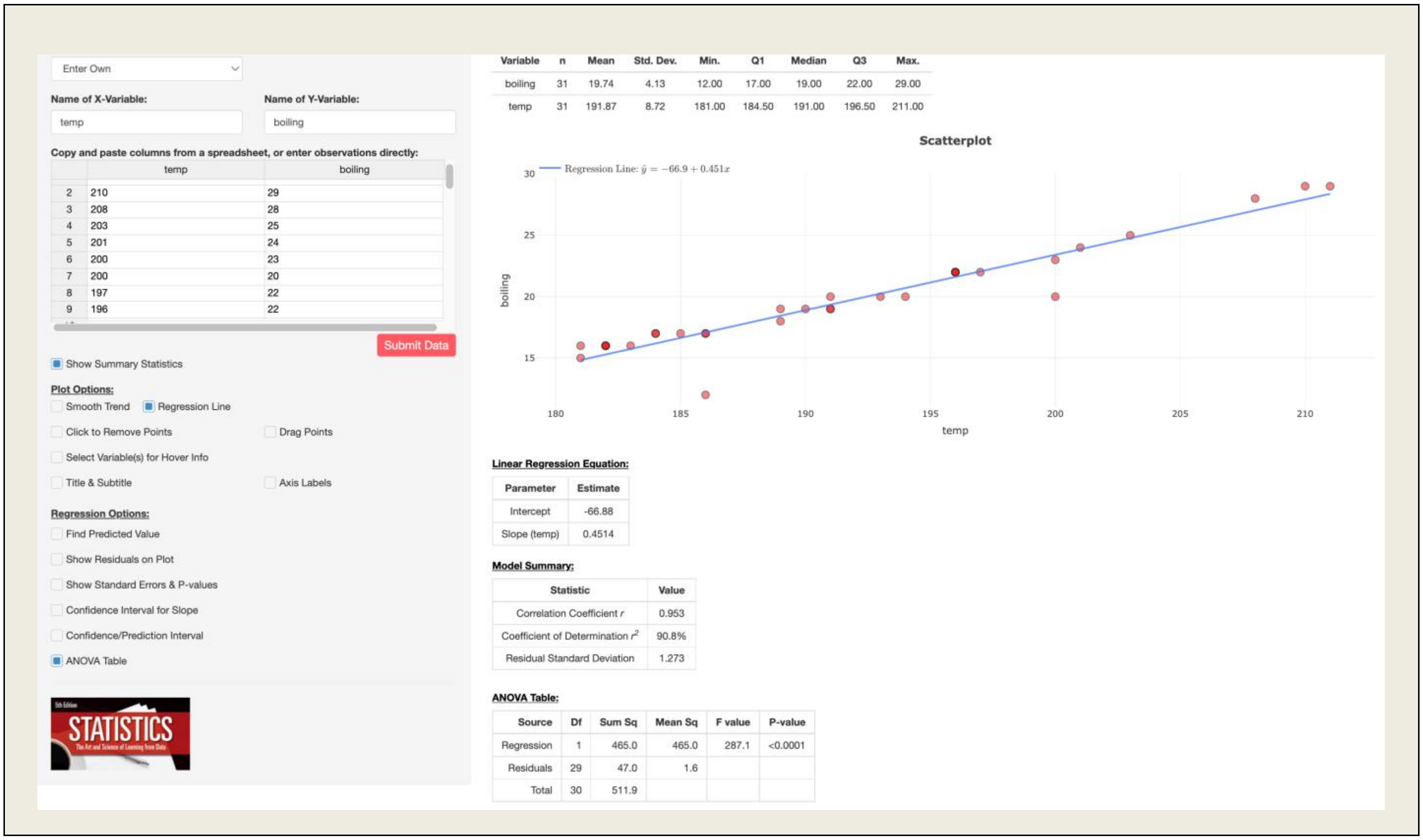
i. The fitted line estimates of $\hat{b}_0$ and $\hat{b}_1$
ii. Analysis of variance table
iii. R^2 = % of the variability in the outcome explained by the fitted line
iv. Scatter plot with overlay of fitted line

```
Fit of Model i:  Y = boiling and X = temp
```

```
__1.  Launch Excel and open simplelinear.xlsx
__2.  Do an EDIT/COPY of cells A2:B32.  Do not include column headings.  Minimize Excel but do not close.
__3.  Launch ArtofStat here www.artofstat.com   >   Online WebApps  >   Linear Regression
__4.  From ENTER DATA drop down, choose:  YOUR OWN
__5.  PASTE your data.
__6.  Click SUBMIT DATA.
```

```
From the options at left, click to display
__7.  the analysis of variance table (ANOVA TABLE)
__8.  the fitted regression line on the plot (REGRESSION LINE)
```

```
You should now see:
```



Name of X-Variable: temp

Name of Y-Variable: boiling

Copy and paste columns from a spreadsheet, or enter observations directly:

	temp	boiling
2	210	29
3	208	28
4	203	25
5	201	24
6	200	23
7	200	20
8	197	22
9	196	22

☑ Show Summary Statistics

Plot Options:

☐ Smooth Trend ☑ Regression Line

☐ Click to Remove Points ☐ Drag Points

☐ Select Variable(s) for Hover Info

☐ Title & Subtitle ☐ Axis Labels

Regression Options:

☐ Find Predicted Value

☐ Show Residuals on Plot

☐ Show Standard Errors & P-values

☐ Confidence Interval for Slope

☐ Confidence/Prediction Interval

☑ ANOVA Table

Variable	n	Mean	Std. Dev.	Min.	Q1	Median	Q3	Max.
boiling	31	19.74	4.13	12.00	17.00	19.00	22.00	29.00
temp	31	191.87	8.72	181.00	184.50	191.00	196.50	211.00

Linear Regression Equation:

Parameter	Estimate
Intercept	-66.88
Slope (temp)	0.4514

Model Summary:

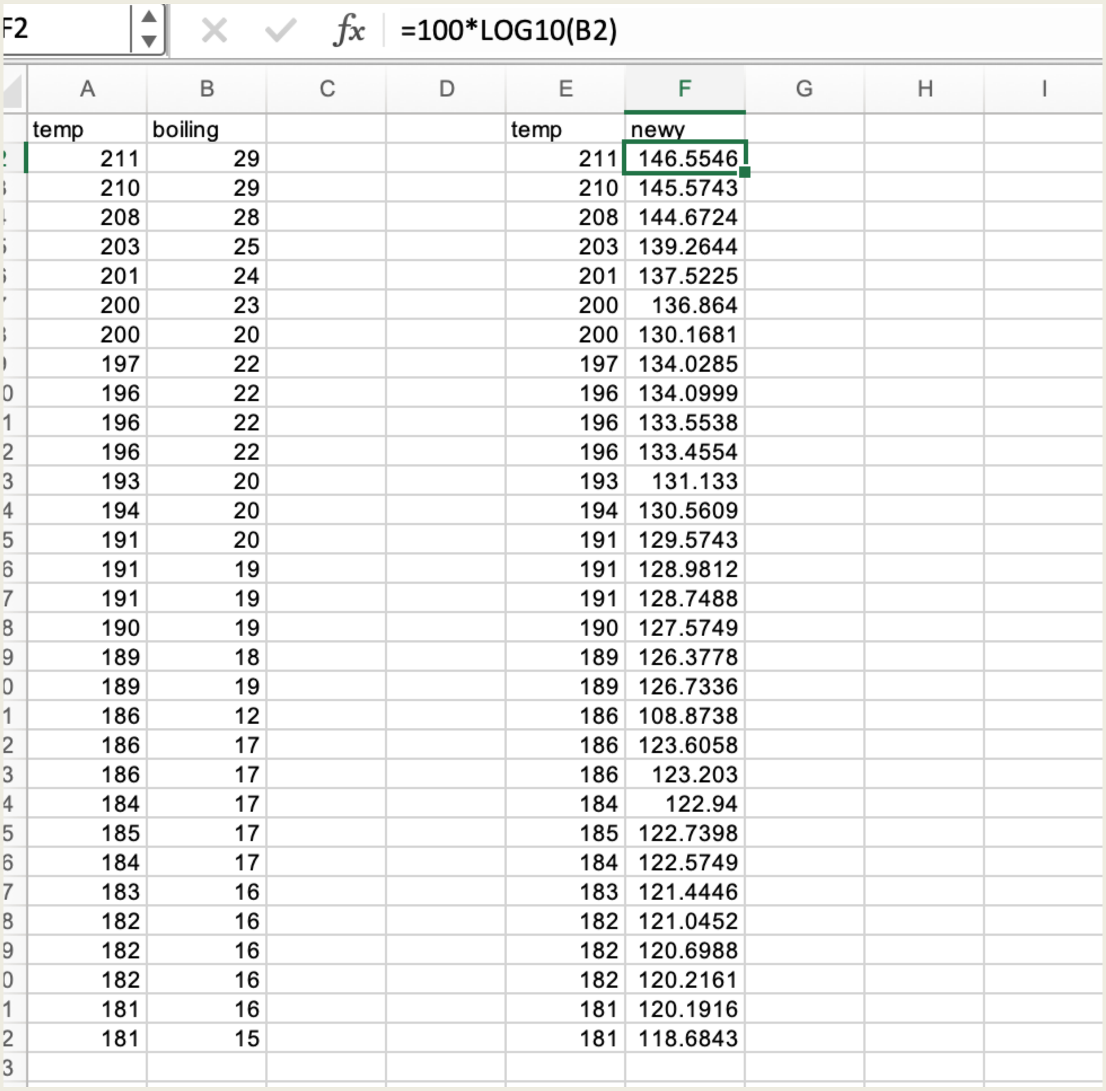
Statistic	Value
Correlation Coefficient r	0.953
Coefficient of Determination r^2	90.8%
Residual Standard Deviation	1.273

ANOVA Table:

Source	Df	Sum Sq	Mean Sq	F value	P-value
Regression	1	465.0	465.0	287.1	<0.0001
Residuals	29	47.0	1.6		
Total	30	511.9			

Fit of Model ii: Y = newy and X = temp

__1. Activate Excel (simplelinear.xlsx should still be open)
__2. Make a copy of the column containing X=temp
__3. Create a new column called newy calculated as newy = 100*log$_{10}$(boiling)

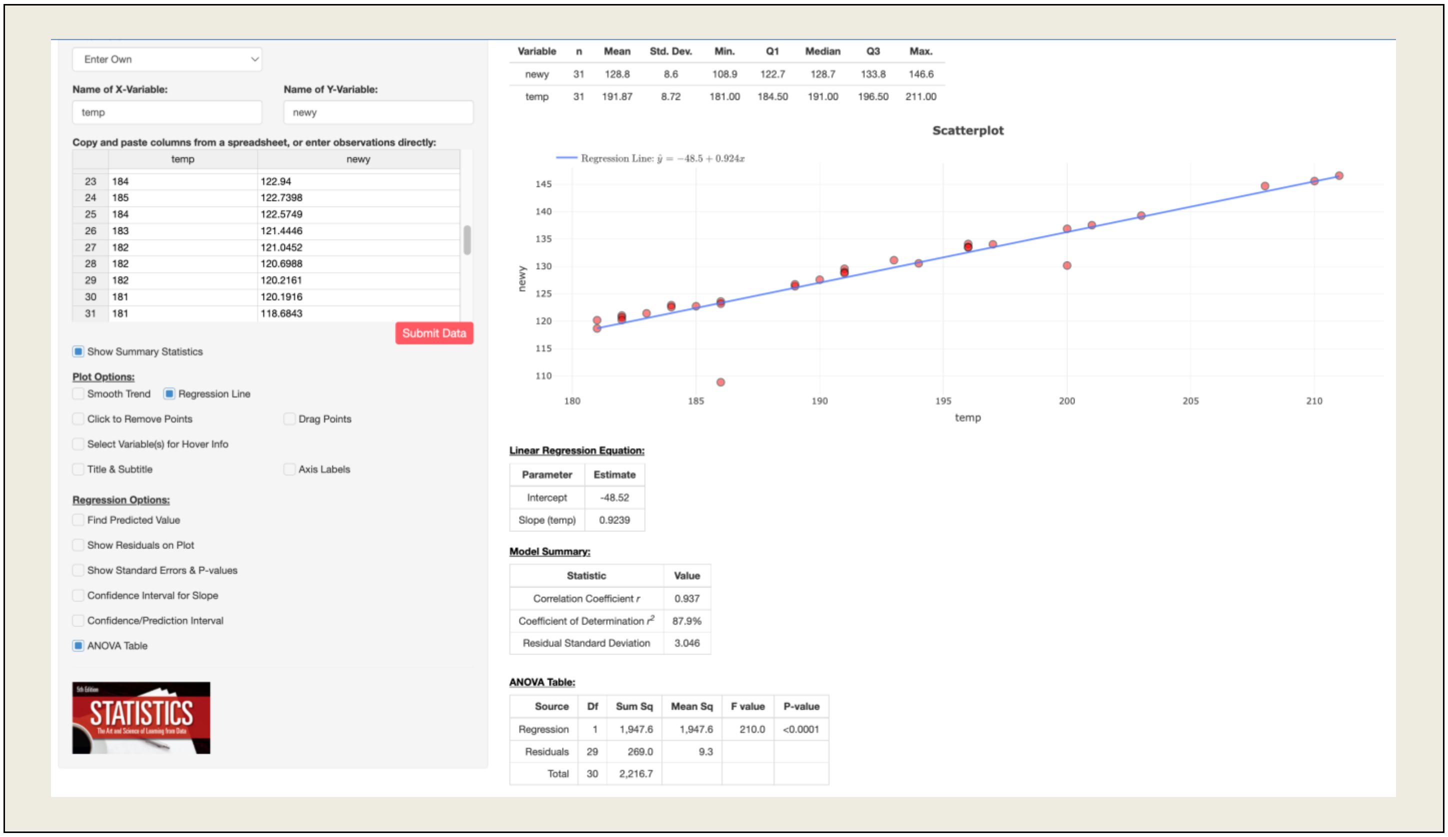


F2 =100*LOG10(B2)

A	B	C	D	E	F	G	H	I
temp	boiling			temp	newy			
211	29			211	146.5546			
210	29			210	145.5743			
208	28			208	144.6724			
203	25			203	139.2644			
201	24			201	137.5225			
200	23			200	136.864			
200	20			200	130.1681			
197	22			197	134.0285			
196	22			196	134.0999			
196	22			196	133.5538			
196	22			196	133.4554			
193	20			193	131.133			
194	20			194	130.5609			
191	20			191	129.5743			
191	19			191	128.9812			
191	19			191	128.7488			
190	19			190	127.5749			
189	18			189	126.3778			
189	19			189	126.7336			
186	12			186	108.8738			
186	17			186	123.6058			
186	17			186	123.203			
184	17			184	122.94			
185	17			185	122.7398			
184	17			184	122.5749			
183	16			183	121.4446			
182	16			182	121.0452			
182	16			182	120.6988			
182	16			182	120.2161			
181	16			181	120.1916			
181	15			181	118.6843			

__4. Do an EDIT/COPY of your two new columns: X = copy of temp and Y = newy (In my excel, this is columns "E" and "F")
__5. Launch ArtofStat here www.artofstat.com > Online WebApps > Linear Regression
__6. As you did for model i, fit model ii.
__7. At left, click to display anova table.
__8. At left, click to display fitted regression line on plot.

You should now see:


Scatterplot
Regression Line: $\hat{y} = -48.5 + 0.924x$
145
140
135
130
125
120
115
110
newy
180
185
190
195
200
205
210
temp
STATISTICS
The Art and Science of Learning from Data

Enter Own

Name of X-Variable: temp

Name of Y-Variable: newy

Copy and paste columns from a spreadsheet, or enter observations directly:

	temp	newy
23	184	122.94
24	185	122.7398
25	184	122.5749
26	183	121.4446
27	182	121.0452
28	182	120.6988
29	182	120.2161
30	181	120.1916
31	181	118.6843

Submit Data

☑ Show Summary Statistics

Plot Options:

☐ Smooth Trend ☑ Regression Line

☐ Click to Remove Points ☐ Drag Points

☐ Select Variable(s) for Hover Info

☐ Title & Subtitle ☐ Axis Labels

Regression Options:

☐ Find Predicted Value

☐ Show Residuals on Plot

☐ Show Standard Errors & P-values

☐ Confidence Interval for Slope

☐ Confidence/Prediction Interval

☑ ANOVA Table

Variable	n	Mean	Std. Dev.	Min.	Q1	Median	Q3	Max.
newy	31	128.8	8.6	108.9	122.7	128.7	133.8	146.6
temp	31	191.87	8.72	181.00	184.50	191.00	196.50	211.00

Linear Regression Equation:

Parameter	Estimate
Intercept	-48.52
Slope (temp)	0.9239

Model Summary:

Statistic	Value
Correlation Coefficient r	0.937
Coefficient of Determination r^2	87.9%
Residual Standard Deviation	3.046

ANOVA Table:

Source	Df	Sum Sq	Mean Sq	F value	P-value
Regression	1	1,947.6	1,947.6	210.0	<0.0001
Residuals	29	269.0	9.3		
Total	30	2,216.7			

R Users

1a. Create a new variable newy = 100*$\log_{10}$(boiling)

1b. For each model, obtain:

i. The fitted line estimates of $\hat{b}_0$ and $\hat{b}_1$
ii. Analysis of variance table
iii. R^2 = % of the variability in the outcome explained by the fitted line
iv. Scatter plot with overlay of fitted line

```
Fit of Models i and ii:
 i:  Y = boiling and X = temp
ii:  Y = newy and X = temp
```

```
setwd("/Users/cbigelow/Desktop/")
options(scipen=1000)                                              # scipen=1000 turns off scientific notation
rm(list=ls())                                                     # rm(list=ls()) clears the workspace/environment


Input data.
library(readxl)
dfboiling <- read_excel("simplelinear.xlsx")                     # During import, I named the dataframe dfboiling for ease

Create newy
dfboiling$newy <- 100*log10(dfboiling$boiling)
```

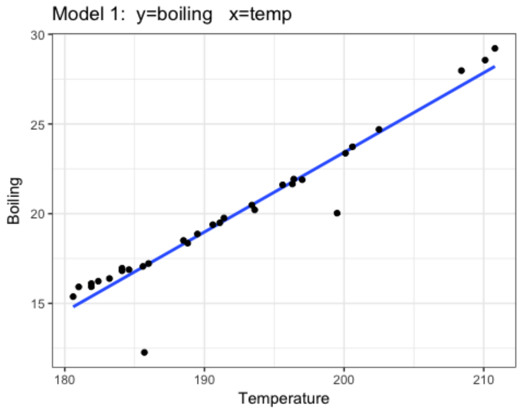
```
Model i: y=boiling x=temp
#a)  Fitted line estimates of intercept and slope
m1 <- lm(boiling ~ temp, data=dfboiling)
coefficients(m1)
## (Intercept)        temp
## -65.3429977   0.4437942

#b) Analysis of Variance (anova) table
temp <- anova(m1)                                    # saving anova(m1) will let me suppress stars in next line of code
print(temp,signif.stars=FALSE)                       # option signif.stars=FALSE suppresses stars
## Analysis of Variance Table
##
## Response: boiling
##           Df Sum Sq Mean Sq F value                Pr(>F)
## temp       1 450.56  450.56   336.6 < 0.00000000000000022
## Residuals 29  38.82    1.34

#c)  R-squared
summary(m1)$r.squared                                # Tip! Issue the command str(m1) to obtain list of stored quantities
## [1] 0.9206773
```

```
Model i: Scatterplot w Overlay of Fitted Line
library(ggplot2)
ggplot(data=dfboiling, aes(x=temp,y=boiling)) +          # ggplot requires data=, aes( ) and geom_NAME()
     geom_smooth(method=lm, se=FALSE) +                  # Plot line first. se=FALSE suppresses the CI band
     geom_point() +                                      # geom_point( ) plots the points on top of line
     xlab("Temperature") +                               # Additional layers are optional (but nice!)
     ylab("Boiling") +
     ggtitle("Model 1:  y=boiling   x=temp") +
     theme_bw()
```

Model 1: y=boiling x=temp



Model ii: newy=100*log10(boiling) x=temp

```
#a)  Fitted line
m2 <- lm(newy ~ temp, data=dfboiling)
coefficients(m2)
## (Intercept)        temp
## -48.8582854   0.9261547
```

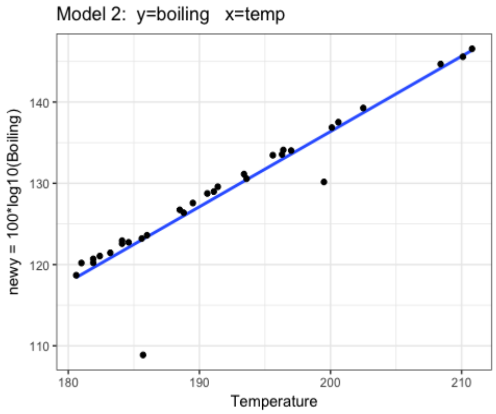
```
#b) Anova table
temp <- anova(m2)
print(temp,signif.stars=FALSE)
## Analysis of Variance Table
##
## Response: newy
##           Df Sum Sq Mean Sq F value                Pr(>F)
## temp       1 1962.2 1962.25  223.69 0.000000000000003623
## Residuals 29  254.4    8.77

#c)  R-squared
summary(m2)$r.squared
## [1] 0.8852331
```

Model ii: Scatterplot w Overlay of Fitted Line

```
library(ggplot2)
ggplot(data=dfboiling, aes(x=temp,y=newy)) +
     geom_smooth(method=lm, se=FALSE) +                 # geom_smooth( ) to plot line first
     geom_point() +                                     # geom_point( ) to layer scatter plot on top
     xlab("Temperature") +
     ylab("newy = 100*log10(Boiling)") +
     ggtitle("Model 2:  y=boiling   x=temp") +
     theme_bw()
```

Model 2: y=boiling x=temp



1c. In 3-5 sentences, write a one-paragraph interpretation of your two model fits.

In this sample of n=31 observations, the scatter plot reveals two outlying values. Their inclusion may or may not be appropriate.

In an analysis that includes all n=31 observations, a simple linear regression of y=boiling point on x=temperature explains more of the variability in the outcome than a simple linear regression of newy = 100*$\log_{10}$(boiling) on x=temperature (R^2 = 92% versus 89%).

Take care! - It would not make sense to compare the residual mean squares of the two models because the scales of measurement involved are different.

#2.

Note – This question does NOT require use of software (R or otherwise!)

This exercise gives you practice working with a fitted model that is provided to you. A psychiatrist wants to know whether the level of pathology (Y) in psychotic patients 6 months after treatment could be predicted with reasonable accuracy from knowledge of pretreatment symptom ratings of thinking disturbance (X_1) and hostile suspiciousness (X_2).

2a. The least squares estimation equation involving both independent variables is given by

$$Y = -0.628 + 23.639(X_1) - 7.147(X_2)$$

Using this equation, determine the predicted level of pathology (Y) for a patient with pretreatment scores of 2.80 on thinking disturbance and 7.0 on hostile suspiciousness. How does the predicted value obtained compare with the actual value of 25 observed for this patient?

$$\widehat{Y} = 15.53, \textit{ obtained as follows:}$$

$$\widehat{Y} = -0.628 + 23.639 \cdot X_1 - 7.147 \cdot X_2$$

$$= -0.628 + (23.639 \cdot 2.80) - (7.147 \cdot 7.0)$$

$$= 15.53$$

The predicted value of 15.53 is lower than the actual value of 25 observed for this patient.

2b. Using the analysis of variance tables below, carry out the *overall* F test for each of three models: i) model with X_1 alone; ii) model with X_2 alone; and iii) model with both X_1 and X_2.

Source	**DF**	**Sum of Squares**
Regression on X_1	1	1546
Residual	51	12246

Source	**DF**	**Sum of Squares**
Regression on X_2	1	160
Residual	51	13632

Source	**DF**	**Sum of Squares**
Regression on X_1 , X_2	2	2784
Residual	50	11008

Model Containing X_1 ALONE

$$F = \left(\frac{\text{SSQ Regression on } X_1 \text{/DF Regression}}{\text{SSQ residual/DF Residual}} \right) = \left(\frac{1546/1}{12,246/51} \right) = 6.4385$$

on DF=1,51
p-value=0.01427

Application of the null hypothesis model has led to an extremely unlikely result (p-value = .014), prompting statistical rejection of the null hypothesis. The fitted linear model in X_1 explains statistically significantly more of the variability in level of pathology (Y) than is explained by $\bar{Y}$ (the intercept model) alone.

R code for p-value

```
pf(6.4385,df1=1,df2=51,lower.tail=FALSE)
[1] 0.01426712
```

Model Containing X_2 ALONE

$$F = \left(\frac{\text{SSQ Regresion on } X_2 \text{/DF Regression}}{\text{SSQ Residual/DF Residual}} \right) = \left(\frac{160/1}{13{,}632/51} \right) = 0.5986$$

on DF=1,51
p-value=0.44268

Here, application of the null hypothesis model has ***not*** led to an extremely unlikely result (p-value = .44). The null hypothesis is therefore **not rejected**. The fitted linear model in X_2 **does not** explain statistically significantly more of the variability in level of pathology (Y) than is explained by $\bar{Y}$ (the intercept model) alone.

R code for p-value

```
pf(0.5986,df1=1,df2=51,lower.tail=FALSE)
[1] 0.4426844
```

Model Containing X_1 and X_2

$$F = \left(\frac{\text{SSQ regression on } X_1 \text{ and } X_2 \text{ /Regression df}}{\text{SSQ residual / Residual df}} \right) = \left(\frac{2{,}784/2}{11{,}008/50} \right) = 6.3227$$

on DF=2,50
p-value=0.00356→

Last but not least, here, application of the null hypothesis model has led to an extremely unlikely result (p-value = .00356), prompting statistical rejection of the null hypothesis. The fitted linear model in X_1 and X_2 explains statistically significantly more of the variability in level of pathology (Y) than is explained by $\bar{Y}$ (the intercept model) alone.

R code for p-value

```
pf(6.3227,df1=2,df2=50,lower.tail=FALSE)
[1] 0.003564679
```

2c. Based on your results in part (b), how would you rate the importance of the two variables in predicting Y?

X_1 explains a significant proportion of the variability in Y when modelled as a linear predictor.
X_2 does not. (However, we don't know if a different functional form might have been important.)

2d. What are the R² values for the three regressions referred to in part (b)?

Total SSQ= (Regression SSQ) + (Residual SSQ) is constant.
Therefore total SSQ can be calculated from just one anova table:

$$\text{Total (SSQ)=} \quad 1{,}546 \ + \ 12{,}246 \ = \ 13{,}792$$

$$\begin{aligned}
R^2(X_1 \text{ only}) &= \text{(Regression SSQ)/(Total SSQ)} \\
&= (1546)/(13{,}792) = \ 0.1121 \\
R^2(X_2 \text{ only}) &= (160)/(13{,}792) = \ 0.0116 \\
R^2(X_1 \text{ and } X_2) &= (2784)/(13{,}792) = \ 0.2019
\end{aligned}$$

2e. Based on the above, in your opinion, which is the best model involving either one or both of the two independent variables?

Eliminate from consideration model with X_2 only.
Compare model with X_1 alone versus X_1 and X_2 using partial F test.

$$\textit{Partial F} = \frac{\{(\text{SSQ Regression on } X_1,X_2) \text{ - } (\text{SSQ Regression on } X_1)\}/\text{VDF}}{\text{SSQ Residual for model w } X_1,X_2/\text{Residual DF}} = \frac{(2784 - 1546)/1}{(11{,}008)/50}$$

$$= 5.6263 \quad \text{on} \quad \text{DF} = 1{,}50$$

$$\text{P-value} = 0.02162$$

Addition of X_2 to model containing X_1 is statistically significant (p-value = .02). →
More appropriate model includes X_1 and X_2

R code for p-value

```
pf(5.6263,df1=1,df2=50,lower.tail=FALSE)
[1] 0.0215831
```

#3.

Note – This question does NOT require use of software (R or otherwise!) with one exception: to obtain p-values for parts a-c. Tip – Use Art of Stat if you like!

This exercise gives you practice working with analysis of variance tables. In an an experiment to describe the toxic action of a certain chemical on silkworm larvae, the relationship of $\log_{10}$(dose) and $\log_{10}$(larva weight) to $\log_{10}$(survival) was sought. The data, obtained by feeding each larva a precisely measured dose of the chemical in an aqueous solution and then recording the survival time (ie time until death) are given in the table. Also given are relevant computer results and the analysis of variance table.

Larva	1	2	3	4	5	6	7	8
Y = $\log_{10}$(survival time)	2.836	2.966	2.687	2.679	2.827	2.442	2.421	2.602
X_1=$\log_{10}$(dose)	0.150	0.214	0.487	0.509	0.570	0.593	0.640	0.781
X_2=$\log_{10}$(weight)	0.425	0.439	0.301	0.325	0.371	0.093	0.140	0.406

Larva	9	10	11	12	13	14	15
Y = $\log_{10}$(survival time)	2.556	2.441	2.420	2.439	2.385	2.452	2.351
X_1=$\log_{10}$(dose)	0.739	0.832	0.865	0.904	0.942	1.090	1.194
X_2=$\log_{10}$(weight)	0.364	0.156	0.247	0.278	0.141	0.289	0.193

$$Y = 2.952 - 0.550\,(X_1)$$
$$Y = 2.187 + 1.370\,(X_2)$$
$$Y = 2.593 - 0.381\,(X_1) + 0.871\,(X_2)$$

Source	DF	Sum of Squares
Regression on X_1	1	0.3633
Residual	13	0.1480

Source	DF	Sum of Squares
Regression on X_2	1	0.3367
Residual	13	0.1746

Source	DF	Sum of Squares
Regression on X_1 , X_2	2	0.4642
Residual	12	0.0471

3a. Test for the significance of the overall regression involving both independent variables X_1 and X_2.

$X_1$and X_2

$$F = \frac{(\text{SSQ regression on } X_1 \text{ and } X_2)/2}{(\text{SSQ Residual})/12} = \frac{(0.4642)/2}{(0.0471)/12} = 59.18 \quad on \quad DF = 2{,}12$$

$$P\text{-}value < 0.0001$$

Application of the null hypothesis model has led to an extremely unlikely result (p-value = .0001), prompting statistical rejection of the null hypothesis. The fitted linear model in X_1 and X_2 explains statistically significantly more of the variability in $\log_{10}$(survival time) (Y) than is explained by $\bar{Y}$ (the intercept model) alone.

R code for p-value

```
pf(59.1818, df1=2, df2=12, lower.tail=FALSE)
[1] 0.0000006083442
```

3b. Test to see whether using X_1 alone significantly helps in predicting survival time.

X_1 alone

$$F = \frac{(0.3633)/1}{(0.1480)/13} = \frac{(\text{SSQ Regression on } X_1)/1}{(\text{SSQ Residual})/13} = 31.9115 \quad on \quad DF = 1{,}13$$

$$P\text{-}value = 0.00008$$

Application of the null hypothesis model has led to an extremely unlikely result (p-value = .00008), prompting statistical rejection of the null hypothesis. The fitted linear model in X_1 explains statistically significantly more of the variability in $\log_{10}$(survival time) (Y) than is explained by $\bar{Y}$ (the intercept model) alone.

R code for p-value

```
pf(31.9115, df1=1, df2=13, lower.tail=FALSE)
[1] 0.00007942699
```

3c. Test to see whether using X_2 alone significantly helps in predicting survival time.

X_2 alone

$$F = \frac{(\text{SSQ Regression on } X_2)/1}{(\text{SSQ Residual})/13} = \frac{(0.3367)/1}{(0.1746)/13} = 25.07 \quad on \quad DF = 1{,}13$$

$$P\text{-}value = 0.00027$$

Application of the null hypothesis model has led to an extremely unlikely result (p-value = .00027), prompting statistical rejection of the null hypothesis. The fitted linear model in X_2 explains statistically significantly more of the variability in $\log_{10}$(survival time) (Y) than is explained by $\bar{Y}$ (the intercept model) alone.

R code for p-value

```
pf(25.07, df1=1, df2=13, lower.tail=FALSE)
[1] 0.0002399706
```

3d. Compute R^2 for each of the three models.

$$\begin{aligned}
&\text{TotalSSQ} &&= && 0.5113 \\
&R^2\left(X_1 \text{ and } X_2\right) = && 0.4642/0.5113 &&= 0.9079 \\
&R^2\ (X_1 \text{ alone}) = && 0.3633/0.5113 &&= 0.7105 \\
&R^2\left(X_2 \text{ alone}\right) = && 0.3367/0.5113 &&= 0.6585
\end{aligned}$$

3e. Which independent predictor do you consider to be the best single predictor of survival time?

Using just the criteria of the overall F test and comparison of R^2, the single predictor model containing X_1 is better.

3f. Which model involving one or both of the independent predictors do you prefer and why?

Partial F for comparing model with X_1 alone versus model with X_1 and X_2

$$= \frac{(\,\Delta\ Regression\ SSQ)\big/(\,\Delta\ Regression\ df\,)}{(\,Full\ model\ Residual\ SSQ)\big/(\,Full\ model\ Residual\ df\,)} = \frac{(\,.4642 - .3633)\big/(2-1)}{.0471\big/12}$$

$= 25.707006$

R code for p-value

```
pf(25.07, df1=1, df2=12, lower.tail=FALSE)
[1] 0.0003057104
```