

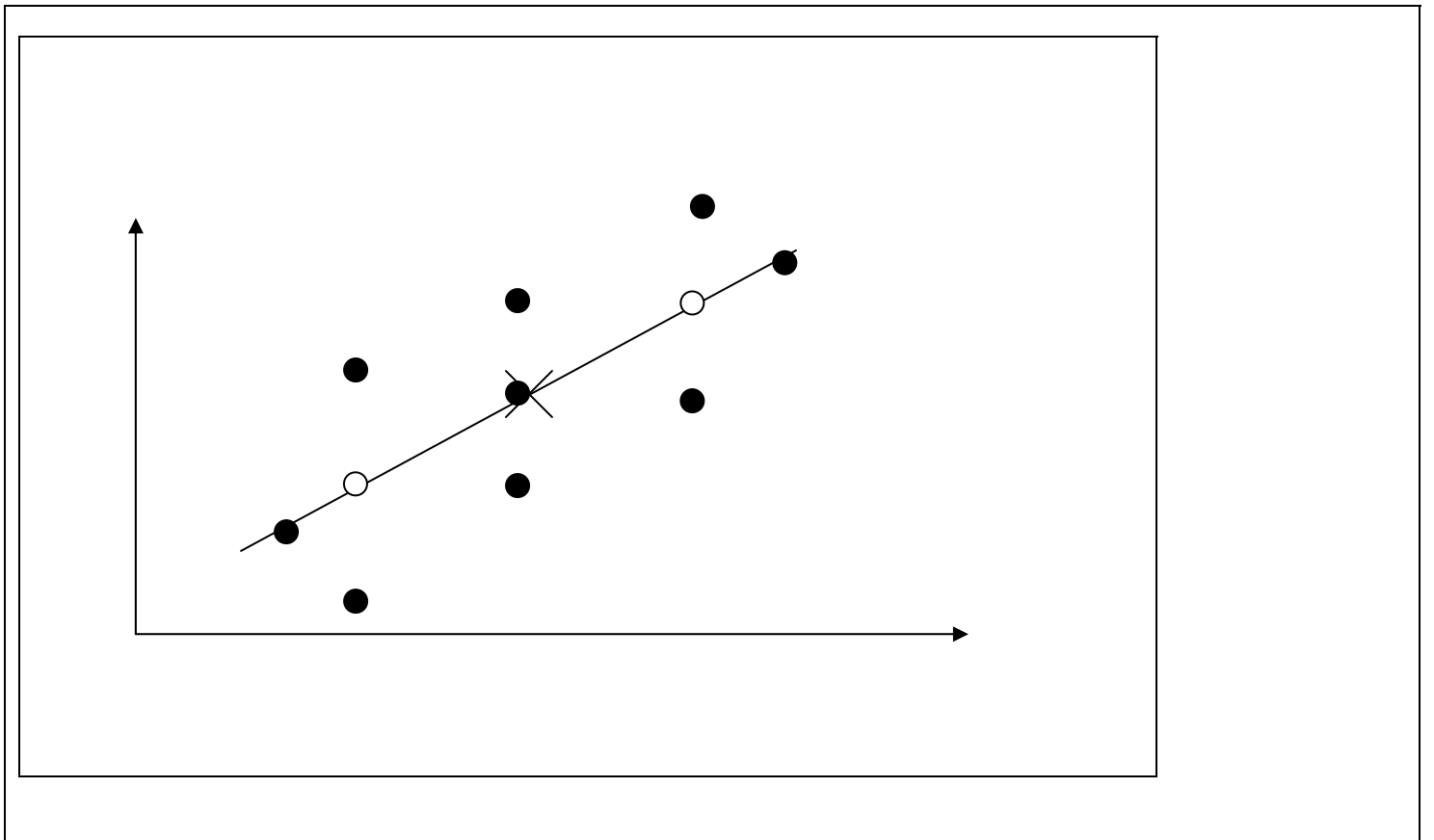
Unit 2 – Regression and Correlation
Practice Problems

SOLUTIONS
Version STATA

1. A regression analysis of measurements of a dependent variable Y on an independent variable X produces a statistically significant association between X and Y. Drawing upon your education in introductory biostatistics, the theory of epidemiology, the scientific method, etc – see how many explanations you can offer for this finding. *Hint – I get seven (7)*

- (i) **X causes Y**
Example: X=smoking Y=cancer
- (ii) **Y causes X**
Example: X=cancer Y=smoking
- (iii) **Positive confounding**
Example: X=alcohol Y=cancer Confounder = Z = smoking
- (iv) **Necessary but not sufficient**
Example: X=sun exposure Y=melanoma
Sun exposure alone does not cause melanoma. Melanoma is the result of a gene-environment interaction.
- (v) **Intermediary**
Example: X = asthma Y=lesions on the lung
X=asthma is an intermediary in the pathway coal exposure → asthma → lesions on lung
- (vi) **Confounding**
Example: X=yellow finger Y=lung cancer
Z=smoking causes both yellow finger and lung cancer.
- (vii) **Anomoly**
There is no association. An event of low probability has occurred.

2. Below is a figure summarizing some data for which a simple linear regression analysis has been performed. The point denoted X that appears on the line is (x,y). The two points indicated by open circles were NOT included in the original analysis.



Multiple Choice (Choose ONE):

Suppose the two points represented by the open circles are added to the data set and the regression analysis is repeated. What is the effect of adding these points on:

1. The estimated slope.
 (a) increase
 (b) decrease
 (c) no change
2. The residual sum of squares.
 (a) increase
 (b) decrease
 (c) no change
3. The degrees of freedom.
 (a) increase
 (b) decrease
 (c) no change
4. Standard error of the estimated slope.
 (a) increase
 (b) decrease
 (c) no change
5. The predicted value of y at $x=24$.
 (a) increase
 (b) decrease
 (c) no change

3. The course website page [REGRESSION AND CORRELATION](#) will have some examples of code to produce regression analyses in STATA and SAS.

The data in the table below are values of boiling points (Y) and temperature (X). Carry out an exploratory analysis to determine whether the relationship between temperature and boiling point is better represented using

- (i) $Y = \beta_0 + \beta_1 X$ or
 (ii) $100 \log_{10}(Y) = \beta'_0 + \beta'_1 X$

In developing your answer, use whatever statistical software you like (SAS, STATA, or Minitab). Try your hand at producing

- (a) Estimates of the regression line parameters
 (b) Analysis of variance tables
 (c) R^2
 (d) Scatter plot with overlay of fitted line.

Complete your answer with a one paragraph text that is an interpretation of your work. Take your time with this and have fun.

X=Temp	Y=Boiling Pt	X=Temp	Y=Boiling Pt	X=Temp	Y=Boiling Pt
210.8	29.211	193.6	20.212	184.1	16.817
210.1	28.559	191.4	19.758	183.2	16.385
208.4	27.972	191.1	19.490	182.4	16.235
202.5	24.697	190.6	19.386	181.9	16.106
200.6	23.726	189.5	18.869	181.9	15.928
200.1	23.369	188.8	18.356	181.0	15.919
199.5	20.030	188.5	18.507	180.6	15.376
197.0	21.892	185.7	12.267		
196.4	21.928	186.0	17.221		
196.3	21.654	185.6	17.062		
195.6	21.605	184.1	16.959		
193.4	20.480	184.6	16.881		

Solution (one paragraph of text that is interpretation of analysis):

Did you notice that the scatter plot of these data reveal two outlying values? Their inclusion may or may not be appropriate.

If all n=31 data points are included in the analysis, then the model that explains more of the variability in boiling point is Y=boiling point modeled linearly in X=temperature. It has a greater R² (92% v 89%).

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

(a). Estimates of Regression Line Parameters

i. $Y = -65.34 + 0.44 * X$

ii. $100 \log_{10}(Y) = -48.85830 + 0.93 * X$

Table 1. Parameters estimations for dependent = Boiling Point

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	-65.34300	4.64382	-14.07	<.0001
Temp	Temperature	1	0.44379	0.02419	18.35	<.0001

Table 2. Parameters estimations for dependent = 100 log₁₀ (Boiling Point)

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	-48.85830	11.88807	-4.11	0.0003
Temp	Temperature	1	0.92615	0.06192	14.96	<.0001

(b). Analysis of Variance Tables

Y=Boiling Point

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	450.55876	450.55876	336.60	<.0001
Error	29	38.81874	1.33858		
Corrected Total	30	489.37750			

$Y = 100 \log_{10}(\text{Boiling Point})$

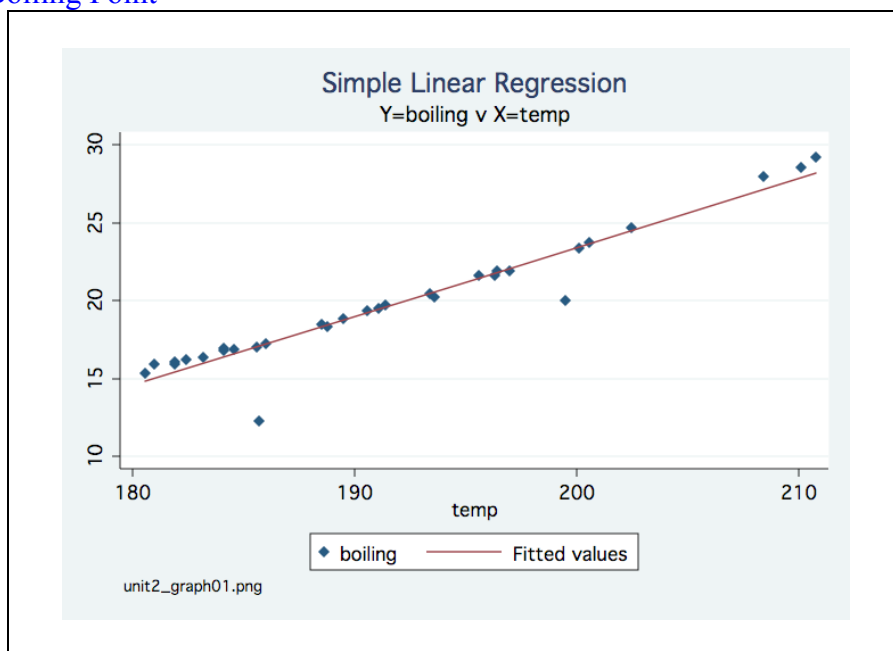
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1962.25366	1962.25366	223.69	<.0001
Error	29	254.39829	8.77235		
Corrected Total	30	2216.65195			

(c). R^2

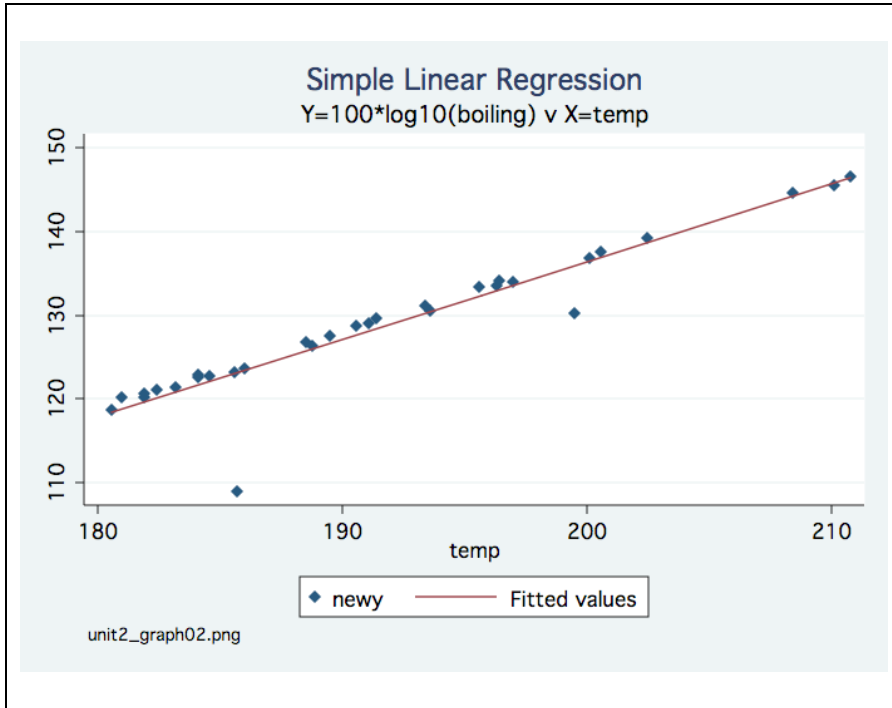
$Y = \text{Boiling Point}$:	$R^2 = 0.9207$
$Y = 100 \log_{10}(\text{Boiling Point})$:	$R^2 = 0.8852$

(d) **Scatterplot with Overlay of Fitted Line**

$Y = \text{Boiling Point}$



$$Y = 100 \log_{10}(\text{Boiling Point})$$



For STATA users (version 10.1)

*comments begin with an asterisk and are shown in green

Read in data and create NEWY = 100 log₁₀(BOILING)

```
. * toggle off the screen by screen pausing of results
. set more off
. * read into memory the data for exercise #3. It is named week02.dta
. use "http://people.umass.edu/biep640w/datasets/week02.dta"
. * Use the command GENERATE to create a new variable called newy
. generate newy=100*log10(boiling)
```

Model 1: Y=Boiling Point**Least squares estimation and analysis of variance table**

```
. regress boiling temp
```

You should see ..

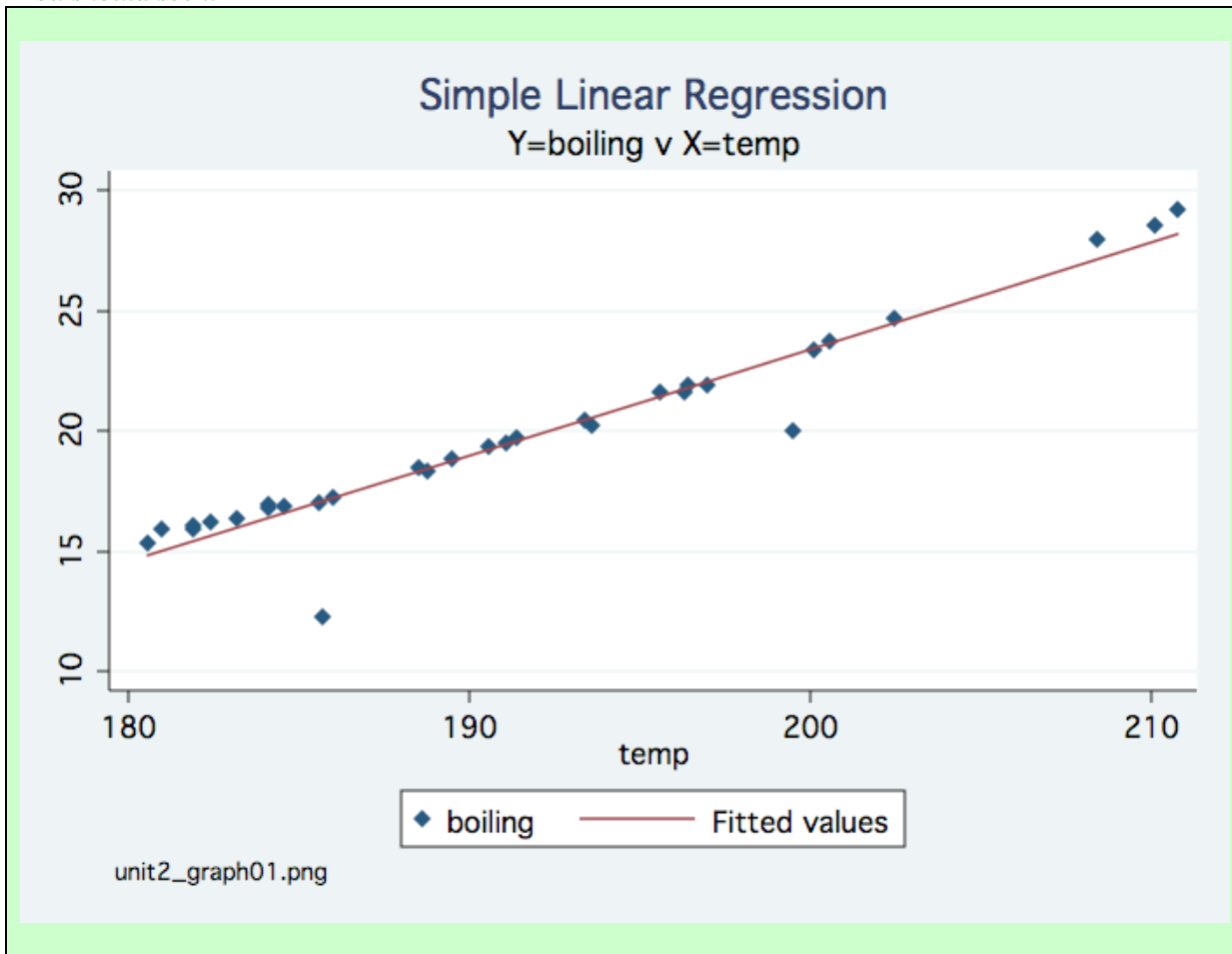
Source	SS	df	MS			
Model	450.558755	1	450.558755	Number of obs =	31	
Residual	38.8187237	29	1.33857668	F(1, 29) =	336.60	
Total	489.377479	30	16.3125826	Prob > F =	0.0000	
				R-squared =	0.9207	
				Adj R-squared =	0.9179	
				Root MSE =	1.157	

boiling	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
temp	.4437942	.0241895	18.35	0.000	.3943211	.4932674
_cons	-65.343	4.643815	-14.07	0.000	-74.84067	-55.84533

Scatter plot with overlay of fitted line.

```
. graph twoway (scatter boiling temp, msymbol(d)) (lfit boiling temp), title("Simple Linear Regression") subtitle("Y=boiling v X=temp") note("unit2_graph01.png")
```

You should see ..



Model 2: NEWY=100*log10(Boiling Point)
Least squares estimation and analysis of variance table

```
. regress newy temp
```

You should see ..

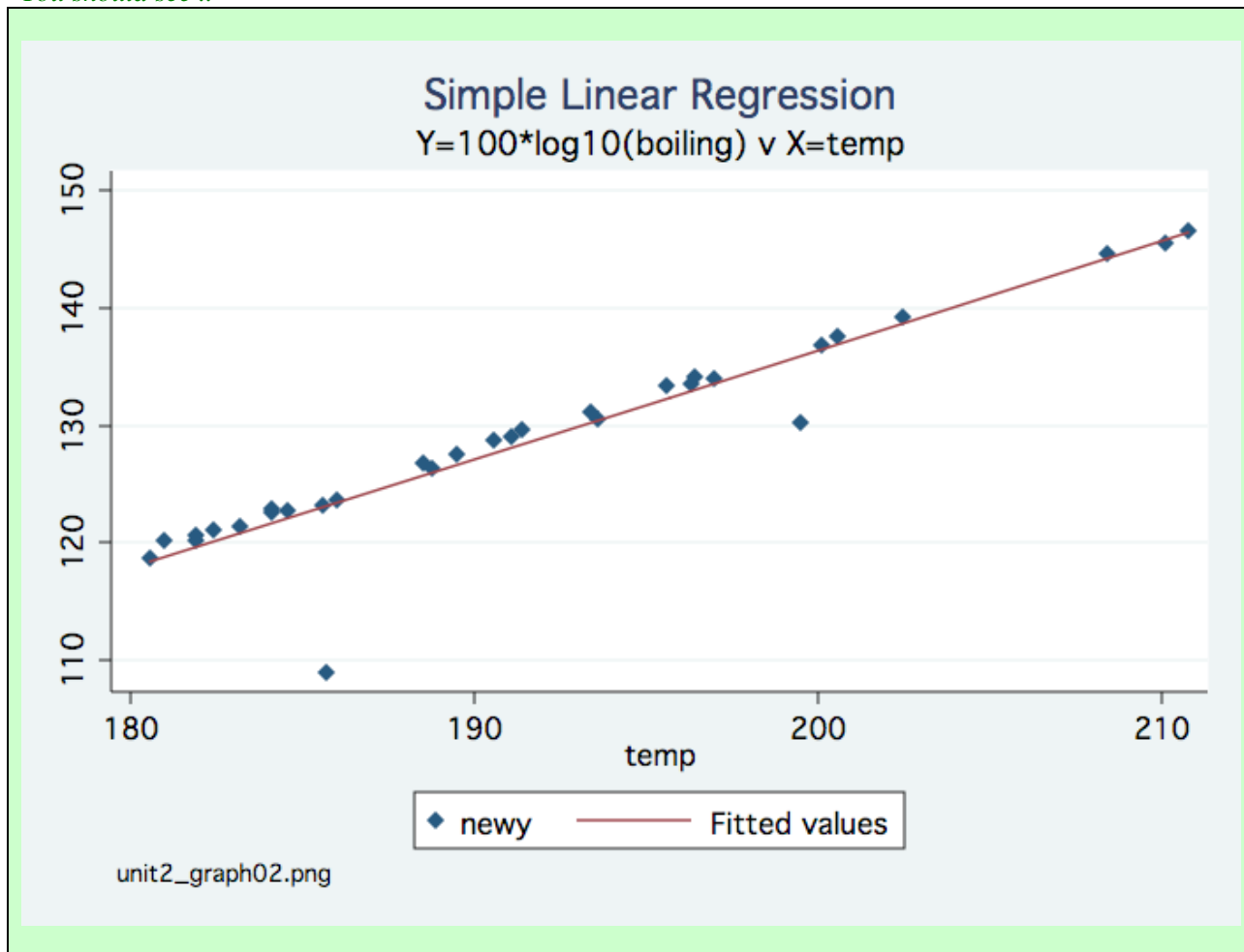
Source	SS	df	MS			
Model	1962.25312	1	1962.25312	Number of obs =	31	
Residual	254.398153	29	8.7723501	F(1, 29) =	223.69	
Total	2216.65127	30	73.8883757	Prob > F =	0.0000	
				R-squared =	0.8852	
				Adj R-squared =	0.8813	
				Root MSE =	2.9618	

newy	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
temp	.9261545	.0619247	14.96	0.000	.7995043	1.052805
_cons	-48.85826	11.88807	-4.11	0.000	-73.17209	-24.54444

Scatter plot with overlay of fitted line.**Note - This requires saving the predicted values. Here I've saved them as `hat_temp`**

```
. graph twoway (scatter newy temp, msymbol(d)) (lfit newy temp), title("Simple Linear  
Regression") subtitle("Y=100*log10(boiling) v X=temp") note("unit2_graph02.png")
```

You should see ..



4.

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).

(a) 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?

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

$$\text{with } X_1=2.80 \text{ and } X_2=7.0 \Rightarrow$$

$$\hat{Y} = 15.5322$$

This value is lower than the observed value of 25

(b) Using the analysis of variance tables below, carry out the overall regression F tests for models containing both X_1 and X_2 , X_1 alone, and X_2 alone.

Source	DF	SS
Regression on X_1	1	1546
Residual	51	12246

Source	DF	SS
Regression on X_2	1	160
Residual	51	13632

Source	DF	SS
Regression on X_1, X_2	2	2784
Residual	50	11008

Model Containing X_1 and X_2

$$F = \left(\frac{2,784/2}{11,008/50} \right) = 6.3227 \quad \text{on DF}=2,50$$

p-value=0.00356 → Conclude 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.

Model Containing X_1 ALONE

$$F = \left(\frac{1546/1}{12,246/51} \right) = 6.4385 \quad \text{on DF}=1,51$$

p-value=0.01427 → Conclude 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.

Model Containing X_2 ALONE

$$F = \left(\frac{160/1}{13,632/51} \right) = 0.5986 \quad \text{on DF}=1,51$$

p-value=0.44268 → Conclude 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.

(c) 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.)

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

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

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

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

$$R^2(X_2 \text{ only}) = (160) / (13,792) = 0.0116$$

$$R^2(X_1 \text{ and } X_2) = (2784) / (13,792) = 0.2019$$

(e) What 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.

$$\begin{aligned} \text{Partial } F &= \frac{\square \text{Regression SSQ} / \square DF}{\text{Residual SSQ}(\text{full}) / DF} = \frac{(2784 - 1546) / 1}{(11,008) / 50} \\ &= 5.6263 \quad \text{on} \quad DF = 1, 50 \end{aligned}$$

P-value = 0.02162

Most appropriate model includes X_1 and X_2

5.

In an experiment to describe the toxic action of a certain chemical on silkworm larvae, the relationship of $\log_{10}(\text{dose})$ and $\log_{10}(\text{larva weight})$ to $\log_{10}(\text{survival})$ 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
$Y = \log_{10}(\text{survival time})$	2.836	2.966	2.687	2.679	2.827	2.442	2.421	2.602
$X_1 = \log_{10}(\text{dose})$	0.150	0.214	0.487	0.509	0.570	0.593	0.640	0.781
$X_2 = \log_{10}(\text{weight})$	0.425	0.439	0.301	0.325	0.371	0.093	0.140	0.406

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

$$Y = 2.952 - 0.550 (X_1)$$

$$Y = 2.187 + 1.370 (X_2)$$

$$Y = 2.593 - 0.381 (X_1) + .871 (X_2)$$

Source	DF	SS
Regression on X_1	1	0.3633
Residual	13	0.1480

Source	DF	SS
Regression on X_2	1	0.3367
Residual	13	0.1746

Source	DF	SS
Regression on X_1, X_2	2	0.4642
Residual	12	0.0471

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

$$\begin{aligned}
 & X_1 \text{ and } X_2 \\
 F &= \frac{(0.4642)/2}{(0.0471)/12} = 59.18 \quad \text{on } DF = 2, 12 \\
 P\text{-value} &< 0.0001
 \end{aligned}$$

- (b) Test to see whether using X_1 alone significantly helps in predicting survival time.

$$\begin{aligned}
 & X_1 \text{ alone} \\
 F &= \frac{(0.3633)/1}{(0.1480)/13} = 31.9115 \quad \text{on } DF = 1, 13 \\
 P\text{-value} &= 0.00008
 \end{aligned}$$

- (c) Test to see whether using X_2 alone significantly helps in predicting survival time.

$$\begin{aligned}
 & X_2 \text{ alone} \\
 F &= \frac{(0.3367)/1}{(0.1746)/13} = 25.07 \quad \text{on } = 1, 13 \\
 P\text{-value} &= 0.00027
 \end{aligned}$$

- (d) Compute R^2 for each of the three models.

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

- (e) Which independent predictor do you consider to be the best single predictor of survival time?

Based on overall F test and comparison of R^2 , the single predictor model containing X_1 is better.

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

$$\begin{aligned} & \text{Partial F for comparing model with } X_1 \text{ alone versus model with } X_1 \text{ and } X_2 \\ & = \frac{(\Delta \text{Regression SSQ}) / \Delta \text{Re g DF}}{(\text{Residual SSQ}) / \text{Residual DF}} = \frac{(0.4637 - 0.3633) / 1}{0.04706 / 12} \\ & = 25.6014 \quad \text{on} \quad \text{DF} = 1, 12 \\ & \text{P-value} = 0.0003 \quad \text{Choose model with both } X_1 \text{ and } X_2 \end{aligned}$$

6. Using whatever software package you like, try your hand at reproducing the analysis of variance tables you worked with in problem #5.

For Stata users

Green – comments (note: comments begin with an asterisk)

Black – stata syntax

Blue – output (note: I have shown only the analysis of variance table portions of the output)

```
. * Use FILE > OPEN to read in data from week03.dta
. use "http://people.umass.edu/biep640w/datasets/week03.dta"
. * Anova table for X1
. regress y x1
```

Source	SS	df	MS	Number of obs =	15
Model	.363274049	1	.363274049	F(1, 13) =	31.91
Residual	.147992891	13	.011384069	Prob > F =	0.0001
				R-squared =	0.7105
				Adj R-squared =	0.6883
Total	.51126694	14	.036519067	Root MSE =	.1067

```
. * Anova table for X2
. regress y x2
```

Source	SS	df	MS	Number of obs =	15
Model	.336674125	1	.336674125	F(1, 13) =	25.07
Residual	.174592816	13	.013430217	Prob > F =	0.0002
				R-squared =	0.6585
				Adj R-squared =	0.6322
Total	.51126694	14	.036519067	Root MSE =	.11589

```
. * Anova table for X1 and X2
. regress y x1 x2
```

Source	SS	df	MS	Number of obs =	15
Model	.46420206	2	.23210103	F(2, 12) =	59.18
Residual	.04706488	12	.003922073	Prob > F =	0.0000
Total	.51126694	14	.036519067	R-squared =	0.9079
				Adj R-squared =	0.8926
				Root MSE =	.06263

7. An educator examined the relationship between number of hours devoted to reading each week (Y) and the independent variables social class (X₁), number of years school completed (X₂), and reading speed measured by pages read per hour (X₃). The analysis of variance table obtained from a stepwise regression analysis on data for a sample of 19 women over the age of 60 is shown.

Source		DF	SSQ
Regression	(X ₃)	1	1058.628
	(X ₂ X ₃)	1	183.743
	(X ₁ X ₂ ,X ₃)	1	37.982
Residual		15	363.300

(a) Test the significance of each variable as it enters the model.

Total SSQ = 1643.653

Step 1 X₃ Enters

Regression SSQ = 1058.628 on DF=1

Residual SSQ = (363.300) + (37.982) + (183.743)=585.0250 on DF=17

$$F = \frac{1058.628/1}{585.025/17} = 30.7622 \text{ on } DF=1,17$$

P-value = 0.00004

Step 2 X₃ already in X₂ Enters

Additional Regression SSQ = 183.743 on DF=1

Residual SSQ = (363.300) + (37.982)=401.282 on DF=16

$$F = \frac{183.743/1}{401.282/16} = 7.3262 \text{ on } DF=1,16$$

P-value = 0.01556

Step 3 X₂ and X₃ already exist in X₁ Enters

Additional Regression SSQ = 37.982 on DF=1

Residual SSQ = 363.300 on DF=15

$$F = \frac{37.982/1}{363.3/15} = 1.5682 \quad \text{on} \quad DF=1,15$$

P-value=0.22964

- (b) Test $H_0: \beta_1 = \beta_2 = 0$ in the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E$.

Models to Compare: X_3 alone versus $(X_1, X_2 \text{ and } X_3)$

$$\text{Partial } F = \frac{\text{Regression SSQ} / \text{Reg DF}}{\text{Residual SSQ} / \text{Res DF}}$$

$$= \frac{(183.743 + 37.982) / 2}{363.300 / 15} = 4.5773 \quad \text{on} \quad DF$$

P-value= 0.02837

- (c) Why can't we test $H_0: \beta_1 = \beta_3 = 0$ using the ANOVA table given? What formula would you use for this test?

The regression SSQ for the model containing X_2 alone is not available.

$$\text{Partial } F = \frac{\text{Reg SSQ}(\text{model w } X_1 \text{ and } X_2 \text{ and } X_3) - \text{Reg SSQ}(X_2 \text{ alone}) / 2}{\text{Resid SSQ}(X_1, X_2 \text{ and } X_3) / 15}$$

- (d) What is your overall evaluation concerning the appropriate model to use given the results in parts (a) and (b)?

The most appropriate model is the one with two predictors, X_3 and X_2 ($R^2 = 0.7559$). The additional predictive information in X_1 (change in $R^2 = 0.0231$) is not statistically significant ($p=0.23$)

8. Consider the following analysis of variance table.

Source		DF	SS
Regression	(X ₁)	1	18,953.04
	(X ₃ X ₁)	1	7,010.03
	(X ₂ X ₁ ,X ₃)	1	10.93
Residual		16	2,248.23
			28,222.23

Using a type I error of 0.05,

(a) Provide a test to compare the following two models:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E. \text{ VERSUS}$$

$$Y = \beta_0 + \beta_1 X_1 + E.$$

$$\text{Partial F} = \frac{\text{Reg SSQ}(X_1, X_2, X_3) - \text{Reg SSQ}(X_1)/2}{\text{Res SSQ}(X_1, X_2, X_3)/16}$$

$$= \frac{(7010.03 + 10.93)/2}{(2248.23)/16} = 24.983 \text{ on DF} = 2, 16$$

P-value < 0.0001

(b) Provide a test to compare the following two models:

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + E. \text{ VERSUS}$$

$$Y = \beta_0 + E.$$

$$\text{Overall F} = \frac{\text{Reg SSQ}(X_1, X_2)/2}{\text{Res SSQ}(X_1, X_3)/17} = \frac{(18,953.04 + 7010.03)/2}{(2,248.23 + 10.93)/17}$$

$$= 97.685 \quad \text{on} \quad \text{DF} = 2, 17$$

P-value < 0.00001

(c) State which two models are being compared in computing:

$$F = \frac{(18,953.04 + 7,010.03 + 10.93)/3}{(2248.23)/16}$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E$$

versus

$$Y = \beta_0 + E$$