a residual plot, using a one number summary of lack of fit such as the test statistic F_{LF} makes little sense.

Nevertheless, the literature for lack of fit tests for various statistical methods is enormous. See Joglekar et al. (1989), Peña and Slate (2006), and Su and Yang (2006) for references.

For the following homework problems, Cody and Smith (2006) is useful for *SAS*, while Cook and Weisberg (1999a) is useful for *Arc*. Becker et al. (1988) and Crawley (2013) are useful for *R*.

2.13 Problems

Problems with an asterisk * are especially important.

```
Output for Problem 2.1
Full Model Summary Analysis of Variance Table
Source      df    SS        MS           F    p-value
Regression   6  265784.  44297.4     172.14    0.0000
Residual    67  17240.9    257.327

Reduced Model Summary Analysis of Variance Table
Source      df    SS        MS           F    p-value
Regression   1  264621. 264621.     1035.26    0.0000
Residual    72  18403.8    255.608
```

2.1. Assume that the response variable Y is *height*, and the explanatory variables are $X_2 =$ *sternal height*, $X_3 =$ *cephalic index*, $X_4 =$ *finger to ground*, $X_5 =$ *head length*, $X_6 =$ *nasal height*, and $X_7 =$ *bigonal breadth*. Suppose that the full model uses all 6 predictors plus a constant ($= X_1$) while the reduced model uses the constant and *sternal height*. Test whether the reduced model can be used instead of the full model using the output above. The data set had 74 cases.

```
Output for Problem 2.2
Full Model Summary Analysis of Variance Table
Source     df      SS   MS       F          p-value
Regression  9  16771.7  1863.52  1479148.9  0.0000
Residual  235  0.29607  0.00126

Reduced Model Summary Analysis of Variance Table
Source     df      SS    MS          F       p-value
Regression  2  16771.7   8385.85   6734072.0  0.0000
Residual  242  0.301359  0.0012453
```

```
Coefficient Estimates, Response = y, Terms = (x2 x2^2)
Label      Estimate  Std. Error    t-value    p-value
Constant   958.470      5.88584    162.843     0.0000
x2        -1335.39     11.1656    -119.599     0.0000
x2^2       421.881      5.29434     79.685     0.0000
```

2.2. The above output, starting on the previous page, comes from the Johnson (1996) STATLIB data set *bodyfat* after several outliers are deleted. It is believed that $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_2^2 + e$ where Y is the person's bodyfat and X_2 is the person's density. Measurements on 245 people were taken. In addition to X_2 and X_2^2, 7 additional measurements $X_4, \ldots, X_{10}$ were taken. Both the full and reduced models contain a constant $X_1 \equiv 1$.

a) Predict Y if $X_2 = 1.04$. (Use the reduced model $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_2^2 + e$.)

b) Test whether the reduced model can be used instead of the full model.

```
Output for Problem 2.3
Label      Estimate   Std. Error   t-value    p-value
Constant  -5.07459    1.85124       -2.741     0.0076
log[H]     1.12399    0.498937       2.253     0.0270
log[S]     0.573167   0.116455       4.922     0.0000

R Squared: 0.895655 Sigma hat: 0.223658, n = 82
(log[H] log[S]) (4 5)
Prediction = 2.2872, s(pred) = 0.467664,
Estimated population mean value = 2.287, s = 0.410715
```

2.3. The above output was produced from the file *mussels.lsp* in *Arc*. See Cook and Weisberg (1999a). Let Y = log(M) where M is the muscle mass of a mussel. Let $X_1 \equiv 1$, $X_2 = \log(\mathrm{H})$ where H is the height of the shell, and let $X_3 = \log(\mathrm{S})$ where S is the shell mass. Suppose that it is desired to predict Y_f if $\log(H) = 4$ and $\log(S) = 5$, so that $\boldsymbol{x}_f^T = (1, 4, 5)$. Assume that $se(\hat{Y}_f) = 0.410715$ and that se(pred) = 0.467664.

a) If $\boldsymbol{x}_f^T = (1, 4, 5)$ find a 99% confidence interval for $E(Y_f)$.

b) If $\boldsymbol{x}_f^T = (1, 4, 5)$ find a 99% prediction interval for Y_f.

```
Problem 2.4 Output, Coef. Estimates Response = height
Label            Estimate Std. Error  t-value p-value
Constant         227.351  65.1732      3.488  0.0008
sternal height   0.955973  0.0515390  18.549  0.0000
finger to ground 0.197429  0.0889004   2.221  0.0295
```

```
R Squared: 0.879324    Sigma hat:  22.0731

Summary Analysis of Variance Table
Source      df      SS      MS          F    p-value
Regression   2  259167.  129583.    265.96     0.0000
Residual    73  35567.2  487.222
```

2.4. The above output, starting on the previous page, is from the multiple linear regression of the response $Y =$ *height* on the two nontrivial predictors *sternal height* = height at shoulder, and *finger to ground* = distance from the tip of a person's middle finger to the ground.

a) Consider the plot with Y_i on the vertical axis and the least squares fitted values $\hat{Y}_i$ on the horizontal axis. Sketch how this plot should look if the multiple linear regression model is appropriate.

b) Sketch how the residual plot should look if the residuals r_i are on the vertical axis and the fitted values $\hat{Y}_i$ are on the horizontal axis.

c) From the output, are *sternal height* and *finger to ground* useful for predicting *height*? (Perform the ANOVA F test.)

2.5. Suppose that it is desired to predict the weight of the brain (in grams) from the cephalic index measurement. The output below uses data from 267 people.

```
predictor    coef      Std. Error  t-value    p-value
Constant   865.001     274.252       3.154     0.0018
cephalic   5.05961     3.48212       1.453     0.1474
```

Do a 4 step test for $\beta_2 \neq 0$.

2.6. Suppose that the scatterplot of X versus Y is strongly curved rather than ellipsoidal. Should you use simple linear regression to predict Y from X? Explain.

2.7. Suppose that the 95% confidence interval for β_2 is $[-17.457, 15.832]$. In the simple linear regression model, is X a useful linear predictor for Y? If your answer is no, could X be a useful predictor for Y? Explain.

2.8. Suppose it is desired to predict the yearly return from the stock market from the return in January. Assume that the correlation $\hat{\rho} = 0.496$. Using the table below, find the least squares line $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$.

variable	mean $\overline{X}$ or $\overline{Y}$	standard deviation s
January return	1.75	5.36
yearly return	9.07	15.35

2.9. Suppose that $\sum(X_i - \overline{X})(Y_i - \overline{Y}) = 70690.0$, $\sum(X_i - \overline{X})^2 = 19800.0$, $\overline{X} = 70.0$, and $\overline{Y} = 312.28$.

a) Find the least squares slope $\hat{\beta}_2$.

b) Find the least squares intercept $\hat{\beta}_1$.

c) Predict Y if $X = 80$.

x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})(y_i - \overline{y})$	$(x_i - \overline{x})^2$
38	41				
56	63				
59	70				
64	72				
74	84				

2.10. In the above table, x_i is the length of the femur and y_i is the length of the humerus taken from five dinosaur fossils (*Archaeopteryx*) that preserved both bones. See Moore (2000, p. 99).

a) Complete the table and find the least squares estimators $\hat{\beta}_1$ and $\hat{\beta}_2$.

b) Predict the humerus length if the femur length is 60.

2.11. Suppose that the regression model is $Y_i = 7 + \beta X_i + e_i$ for $i = 1, \ldots, n$ where the e_i are iid $N(0, \sigma^2)$ random variables. The least squares criterion is $Q(\eta) = \sum_{i=1}^{n}(Y_i - 7 - \eta X_i)^2$.

a) What is $E(Y_i)$?

b) Find the least squares estimator $\hat{\beta}$ of β by setting the first derivative $\frac{d}{d\eta}Q(\eta)$ equal to zero.

c) Show that your $\hat{\beta}$ is the global minimizer of the least squares criterion Q by showing that the second derivative $\frac{d^2}{d\eta^2}Q(\eta) > 0$ for all values of η.

2.12. The location model is $Y_i = \mu + e_i$ for $i = 1, \ldots, n$ where the e_i are iid with mean $E(e_i) = 0$ and constant variance $\text{VAR}(e_i) = \sigma^2$. The least squares estimator $\hat{\mu}$ of μ minimizes the least squares criterion $Q(\eta) = \sum_{i=1}^{n}(Y_i - \eta)^2$. To find the least squares estimator, perform the following steps.

a) Find the derivative $\frac{d}{d\eta}Q$, set the derivative equal to zero and solve for η. Call the solution $\hat{\mu}$.

b) To show that the solution was indeed the global minimizer of Q, show that $\frac{d^2}{d\eta^2}Q > 0$ for all real η. (Then the solution $\hat{\mu}$ is a local min and Q is convex, so $\hat{\mu}$ is the global min.)

2.13. The normal error model for simple linear regression through the origin is

$$Y_i = \beta X_i + e_i$$

for $i = 1, \ldots, n$ where $e_1, \ldots, e_n$ are iid $N(0, \sigma^2)$ random variables.

a) Show that the least squares estimator for β is

$$\hat{\beta} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}.$$

b) Find $E(\hat{\beta})$.

c) Find VAR$(\hat{\beta})$.

(Hint: Note that $\hat{\beta} = \sum_{i=1}^n k_i Y_i$ where the k_i depend on the X_i which are treated as constants.)

2.14. Suppose that the regression model is $Y_i = 10 + 2X_{i2} + \beta_3 X_{i3} + e_i$ for $i = 1, \ldots, n$ where the e_i are iid $N(0, \sigma^2)$ random variables. The least squares criterion is $Q(\eta_3) = \sum_{i=1}^n (Y_i - 10 - 2X_{i2} - \eta_3 X_{i3})^2$. Find the least squares estimator $\hat{\beta}_3$ of β_3 by setting the first derivative $\frac{d}{d\eta_3}Q(\eta_3)$ equal to zero. Show that your $\hat{\beta}_3$ is the global minimizer of the least squares criterion Q by showing that the second derivative $\frac{d^2}{d\eta_3^2}Q(\eta_3) > 0$ for all values of η_3.

Minitab Problems

"Double click" means press the rightmost "mouse" button twice in rapid succession. "Drag" means hold the mouse button down. This technique is used to select "menu" options.

After your computer is on, get into *Minitab*, often by searching programs and then double clicking on the icon marked "Student Minitab."

i) In a few seconds, the *Minitab* session and worksheet windows fill the screen. At the top of the screen there is a menu. The upper left corner has the menu option "File." Move your cursor to "File" and drag down the option "Open Worksheet." A window will appear. Double click on the icon "Student." This will display a large number of data sets.