

1. Consider the 928 bivariate measurements of height (y) and midparent height (x).
 - (a) Use SAS to compute \bar{y} . Obtain a 95% confidence interval for the population mean height, or the expected height of a randomly sampled person from the population, $E(Y)$.
 - (b) Report the following summary statistics:
 - i. $\bar{y} = 1/n \sum y_i = 1/n(y_1 + y_2 + \cdots + y_n)$
 - ii. $\bar{x} = 1/n \sum x_i = 1/n(x_1 + x_2 + \cdots + x_n)$
 - iii. $\sum (x_i - \bar{x})(y_i - \bar{y})$
 - iv. $\sum (x_i - \bar{x})y_i$
 - v. $\sum x_i(y_i - \bar{y})$
 - vi. $s_{xx} = \sum (x_i - \bar{x})^2$
 - vii. $s_x^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$
 - viii. $s_{yy} = \sum (y_i - \bar{y})^2$
 - ix. $s_y^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2$
 - x. r
 - (c) Express the slope from the least squares regression line as a function of r , s_x and s_y above.
2. Consider the chirp frequency data from lecture notes. These are $n = 15$ bivariate measurements on striped ground crickets. (“crickets.dat” online.)
 - (a) Obtain a scatterplot of these measurements (sketched or using software).
 - (b) Specify the simple linear regression model for these data. Identify all parameters in the model, providing the interpretation of each.
 - (c) Explain how the interpretation (and the estimate) of the slope parameter changes if temperature is expressed in Celcius.
 - (d) Estimate the mean chirp frequency among crickets in a temperature of $80^\circ F$. Estimate the standard deviation among chirp frequency measurements made at this fixed temperature.
 - (e) Estimate the mean chirp frequency among crickets in a temperature of $105^\circ F$.
 - (f) Report the sum of squared deviations between the fitted values and the average chirp frequency, \bar{y} ?
 - (g) What proportion of variance in chirp frequencies is explained by the linear regression model?
 - (h) Obtain a plot of the residuals against the fitted values.
 - (i) Obtain a plot of the ordered residuals against the corresponding quantiles from the standard normal distribution.
3. Do the exercises like those on page 15 of the lecture notes:
 - (a) Obtain an approximate 95% confidence interval for the population correlation coefficient ρ when a bivariate random sample of size $n = 20$ results in a sample correlation coefficient of $r_{xy} = -0.45$. Also, conduct a test of $H_0 : \rho = 0$.

- (b) Suppose that two random variables X and Y have correlation $\rho = 0.6$. (That is, the correlation among two quantities in an entire population is $E[(X - \mu_x)(Y - \mu_y)] = 0.6$.) What is the probability that a random sample of $n = 30$ bivariate observations will yield a sample correlation coefficient that exceeds 0.7. Find $\Pr(R > 0.7; \rho = 0.6)$.

4. Rao 11.3bc. In the regression equation:

$$E(Y|a, b, c, d, f) = \beta_0 + \beta_1 a + \beta_2 b + \beta_3 c + \beta_4 d + \beta_5 f$$

(b) Interpret β_1 and β_2

(c) Interpret the quantity $\beta_1 + \beta_2 + \beta_3$

5. 11.11 abdeh (p. 494-495, Rao text needed, data available on website as “ex11.8.dat”, SAS code also available as “rao-ex11-8.sas”)

```

-----
                                The SAS System                                1
                                Model: MODEL1
                                Analysis of Variance
                                Sum of          Mean
Source                          DF          Squares      Square    F Value    Pr > F
Model                            2          4.40039        2.20020    34.56     0.0012
Error                            5          0.31836        0.06367
Corrected Total                   7          4.71875

                                Parameter Estimates

Variable    DF      Parameter      Standard
Intercept   1      -0.86114      0.50012    t Value    Pr > |t|
x1          1       0.80007      0.09658     8.28     0.0004
x2          1      -0.07878      0.03715    -2.12     0.0874

                                Output Statistics

Obs    Dependent Variable    Predicted Value    Residual
1      1.7000                1.9556             -0.2556
2      1.9000                1.8379              0.0621
3      2.1000                2.3004             -0.2004
4      2.8000                2.8423             -0.0423
5      2.2000                2.1811              0.0189
6      2.9000                2.4617              0.4383
7      3.4000                3.3253              0.0747
8      4.1000                4.1958             -0.0958

```

Model: MODEL2
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4.40325	1.46775	18.61	0.0082
Error	4	0.31550	0.07887		
Corrected Total	7	4.71875			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-0.59775	1.49078	-0.40	0.7089
x1	1	0.74355	0.31561	2.36	0.0780
x2	1	-0.12610	0.25186	-0.50	0.6429
x1x2	1	0.00998	0.05238	0.19	0.8582

- (a) Argue that while β_0 is the same in both models, the meanings of β_1 and β_2 are different.
- (b) Determine the least squares prediction equation using model 1.
- (d) Test $H_0 : \beta_1 = \beta_2 = 0$ at level $\alpha = 0.05$ in model 1. Give a brief conclusion.
- (e) Determine the least squares prediction equation using model 2.
- (h) Obtain estimates of the error variance based on the two models. Do the estimates lead you to prefer one model over the other?
6. 11.12 bdfg (p. 495-496 Rao text needed)
7. 11.22: (refer to example 11.17, p.507): Which of models 2-6 is nested in model 1?
1. $Y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + E$
 2. $Y = \beta_0 + \beta_1x_1 + \beta_2x_2 + E$
 3. $Y = \beta_0 + \beta_1x_1 + \beta_3x_3 + E$
 4. $Y = \beta_0 + \beta_1x_1 + E$
 5. $Y = \beta_0 + \beta_1x_1 + \beta_2(x_2 + x_3) + E$
 6. $Y = \beta_0 + \beta_1x_1 + \beta_2x_2^2 + E$
8. 11.25(This question is not required, do not submit it.)
- (a) abcdefg
 - (b) Report the $MS[E]$ and the coefficients of determination for models 1,2 and the full quadratic model.
 - (c) True or false: the coefficient of determination for any of these regression models is the square of the correlation coefficient computed from observed (y) and predicted (\hat{y}) values of the response variable.

9. See Example 11.18 (p. 511) (Data available online as “leafbrn.dat”). This example considers modelling the log of leafburn time in seconds (Y) as a function of percentage nitrogen (X_1), chlorine (X_2) and potassium (X_3) for a sample of $n = 30$ tobacco leaves.

```

-----
                                The SAS System                                1
                                Model: MODEL1
                                Analysis of Variance
                                Sum of          Mean
Source          DF          Squares          Square    F Value    Pr > F
Model           3           5.50473          1.83491    40.27     <.0001
Error          26           1.18479          0.04557
Corrected Total 29           6.68952

                                Parameter Estimates
Variable        DF          Parameter          Standard
Intercept       1           1.81104          0.27952
x1              1           -0.53146         0.06958
x2              1           -0.43964         0.07304
x3              1           0.20898          0.04064
t Value        Pr > |t|
6.48           <.0001
-7.64          <.0001
-6.02          <.0001
5.14           <.0001
-----

```

```

-----
                                Model: MODEL2
                                Analysis of Variance
                                Sum of          Mean
Source          DF          Squares          Square    F Value    Pr > F
Model           1           3.44601          3.44601    29.75     <.0001
Error          28           3.24351          0.11584
Corrected Total 29           6.68952

                                Parameter Estimates
Variable        DF          Parameter          Standard
Intercept       1           2.62570          0.36102
x1              1           -0.59161         0.10847
t Value        Pr > |t|
7.27           <.0001
-5.45          <.0001
-----

```

- Using simple linear regression, report the p -value for a test of no linear association between $\log(\text{leaf burning time})$, y , and potassium percentage, x_3 .
- Using multiple linear regression, report the p -value for a test of no linear association between $\log(\text{leaf burning time})$, y , and potassium percentage, x_3 , after adjusting for dependence on nitrogen (x_1) and chlorine (x_2) percentage.
- Estimate the mean difference in $\log(\text{leaf burning time})$ between two populations; both have fixed x_1 and x_2 , but potassium is increased from $x_3 = 6.0$ to $x_3 = 7.0$. Report a standard error for this difference.
- Consider the difference in the slopes for x_1 and x_2 , for fixed x_3 . Estimate $\beta_1 - \beta_2$ and report a 95% confidence interval.
- Estimate the mean leaf-burning time (on the log scale) among tobacco leaves with $x_1 = 3\%$ nitrogen, $x_2 = 1\%$ chlorine and $x_3 = 7\%$ potassium. Report a standard error and confidence interval for this mean.
- Estimate the standard deviation of this population of leaves.

- (g) Consider a single leaf from this subpopulation with $(x_1, x_2, x_3) = (3, 1, 7)$. Estimate the leaf-burning time on the log-scale. Use a procedure for reporting an interval that will cover the individual leaf's burning time (log-scale) with 95% confidence.
- (h) 11.27, parts a,b,c and d:

Obs	Dependent Variable	Predicted Value	Residual
1	0.3400	0.7375	-0.3975
2	0.1100	-0.009587	0.1196
3	0.3800	0.7973	-0.4173
4	0.6800	0.6300	0.0500
5	0.1800	0.1192	0.0608
6	0	0.1723	-0.1723
7	0.0800	-0.0794	0.1594
8	0.1100	0.3057	-0.1957
9	1.5300	1.2246	0.3054
10	0.7700	1.1695	-0.3995
11	1.1700	1.0758	0.0942
12	1.0100	1.0177	-0.007696
13	0.8900	1.1958	-0.3058
14	1.4000	1.1646	0.2354
15	1.0500	0.9730	0.0770
16	1.1500	1.1535	-0.003523
17	1.4900	1.3723	0.1177
18	0.5100	0.4585	0.0515
19	0.1800	0.2151	-0.0351
20	0.3400	0.4492	-0.1092
21	0.3600	0.1239	0.2361
22	0.8900	0.8503	0.0397
23	0.9100	0.9946	-0.0846
24	0.9200	0.6820	0.2380
25	1.3500	1.0545	0.2955
26	1.3300	1.0967	0.2333
27	0.2300	0.1838	0.0462
28	0.2600	0.3913	-0.1313
29	0.7300	0.8306	-0.1006
30	0.2300	0.2297	0.000321

- a Calculate (or deduce) the variances s_{obs}^2 and s_{pred}^2 of the observed and predicted values, respectively.
- b Calculate the sample coefficient of determination, r^2 as the ratio of these two variances.
- c Verify that the correlation coefficient between the observed and predicted values is the same as the multiple correlation coefficient for the (multiple) linear regression of y on x_1, x_2, x_3 .
- d Will the correlation coefficient between the observed values and the values predicted from the equation below be less than, equal to or more than the correlation coefficient in part [c]?:

$$\hat{y} = 3 - 0.3x_1 - 0.6x_2 + 0.5x_3$$