CSU **Hayward**

Statistics *Department*

Minitab Printouts For Simple Linear Regression

1. Data and Data Description

We use Example 10.2 of Ott: *An Introduction to Statistics and Data Analysis (4th ed)* to illustrate regression using Minitab. The data are yields of corn and amounts of fertilizer used for 10 plots. To begin:

- We put *y* = Yield into c1 of a Minitab worksheet and *x* = Fertilizer into c2 (labeling the columns appropriately),
- Print out the data for reference, and
- Make a scatter plot of the data (`gstd` mode). We have used "standard" graphics mode here to avoid graphic images on the web (but see Problem 1.1 below).

The resulting printout and scatter plot of the data are shown below.

```
MTB > print c1 c2

 Row   Yield   Fertlzr

   1     12        2
   2     13        2
   3     13        3
   4     14        3
   5     15        4
   6     15        4
   7     14        5
   8     16        5
   9     17        6
  10     18        6
```

```
MTB > gstd
MTB > plot c1 c2

     18.0+                                                      *
         -
 Yield   -                                                      *
         -
         -
     16.0+                                          *
         -
         -                             2
         -
         -
     14.0+                 *                        *
         -
         -    *            *
         -
         -
     12.0+    *
         -
           --------+---------+---------+---------+---------+--------Fertlzr
               2.40      3.20      4.00      4.80      5.60
```

This character-graphics plot uses the plotting symbol `2` (instead of the usual `*`) to indicate two points that cannot be resolved as separate.

Problems:

1.1. In the data listing, identify the two points that appear at the same plotting position. In this instance are they coincident or are they merely very close neighbors?

In Windows versions of Minitab you can make a graphics plot:

```
MTB > gpro
MTB > plot c1 * c2
```

Notice that a `*` is required on the command line when using "professional graphics" mode. How does this version of the scatter plot handle the "double" point? Make two printouts of this scatter plot. Mark them A and B.

Notice that there seems to be a linear association between the two variables. On Printout A, use a ruler to draw by eye the line you think fits the data best. Is the association positive or negative? (Save Printout B for problems below.)

1.2. Consider the two variables separately, and find their descriptive statistics. The command is `desc c1 c2`.

(a) Find the "center" of the scatter plot: the point corresponding to the two sample means. Mark this point with an `X` on Printout B from Problem 1.

(b) A quantity of use in some regression-related formulas is $S_{xx} = \Sigma x_i^2 - [\Sigma x_i]^2/n$. It is the numerator of the sample variance of the observed values of x. How can you use the sample standard deviation of Fertilizer from the printout to find S_{xx}? Use a calculator to find the two sums in the above formula for S_{xx} and then to verify your answer from the printout.

1.3. The correlation coefficient r is a measure of linear association. Use the command `corr c1 c2` to verify that the correlation between the variables Fertilizer and Yield is $r = 0.91$.

2. Regression Line and Relevant Tests

Next we use Minitab to find the least squares line: $y = 10.1 + 1.15x$. All of the computer printout in this section results from the `regr` command and the two subcommands that follow it. Ignore the subcommands for now, their will be explained later. Notice the number `1` on the command line. It shows that only one predictor variable (Fertilizer) is used here. That is, we are doing *simple* linear regression. (In general, omitting the number of predictor variables in the command results in an error message.)

```
MTB > regr c1 1 c2;
SUBC> resid c3;
SUBC> pred 4.5.

The regression equation is
Yield = 10.1 + 1.15 Fertlzr
```

The regression model is $Y_i = \beta_0 + \beta_1 x_i + e_i$, where the e_i are random observations from a normal distribution with mean 0 and standard deviation σ. The subscript i runs from 0 through $n = 10$. That is, the data fit a line except for normally distributed random noise. In finding the regression line, the y-intercept of the linear model β_0 is estimated by $b_0 = 10.1$, its slope β_1 is estimated by $b_1 = 1.15$, and the standard deviation σ is estimated by $s_{y|x} = 0.8404$ (denoted by `s` in the printout below). Formulas for computing these estimates from the observed numerical values of Fertilizer and Yield are given in your textbook.

Minitab performs two tests of hypothesis to check whether the y-intercept and the slope, respectively, of the regression model are equal to 0 (against the two-sided alternatives).

- The test that the y-intercept is equal to 0 has a t-statistic of 10.1/0.7973 = 12.67 with $n - 2 = 8$ degrees of freedom. With a value of the t-statistic so far from 0, the null hypothesis that $\beta_0 = 0$ is overwhelmingly rejected. A practical interpretation is that *some* corn would grow even with no fertilizer. (If this null hypothesis were accepted, one might want to eliminate β_0 from the regression model and fit a line that is "forced through the origin.")
- The test that the slope is equal to 0 has a t-statistic of 1.15/0.1879 = 6.12, also with 8 degrees of freedom. The null hypothesis null hypothesis that $\beta_1 = 0$ is also overwhelmingly rejected. This means that Yield changes as Fertilizer changes. Hence some of the variability in Yield can be explained in terms of differing levels of Fertilizer applied, and knowing the Fertilizer level is of some use in predicting Yield. (If this null hypothesis were accepted, then we would be obliged to abandon the regression line as a useful way to predict Yield.)

```
Predictor        Coef       StDev          T        P
Constant      10.1000      0.7973      12.67    0.000
Fertlzr        1.1500      0.1879       6.12    0.000

S = 0.8404      R-Sq = 82.4%     R-Sq(adj) = 80.2%
```

The percentage of the variability in Yield that is "explained by the regression" on Fertilizer is expressed by the coefficient of determination r^2 = 0.824 = 82.4%, denoted `R-Sq` in the printout. (For simple linear regression—one predictor variable—you may ignore the adjusted quantity denoted `R-Sq(adj)` in the printout.)

At this stage we will not study Minitab's Analysis of Variance (ANOVA) table in detail, but the problems below will point out a few useful quantities contained in it.

```
Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      26.450      26.450     37.45    0.000
Residual Error     8       5.650       0.706
Total              9      32.100
```

Minitab calls attention to two kinds of "Unusual Observations":

- Those that have disproportionately *large residuals* (marked with `R` at the end of the relevant printout line). These may be viewed as "outliers" from the regression line. In our example one such observation is noted. (Minitab uses standardized residuals to judge which observations to list as unusual; the method of computing standardized residuals is a bit too complex for us to discuss at this point.)
- Those that have *great influence* in determining how the regression line is oriented through the data (marked with `X`). In deciding whether a point is "influential" the issue is whether the regression line would look substantially different if that point were deleted. Our dataset has no influential points.

```
Unusual Observations
Obs    Fertlzr      Yield         Fit    StDev Fit    Residual    St
Resid
  7       5.00     14.000      15.850       0.325      -1.850       -
2.39R
R denotes an observation with a large standardized residual
```

Problems:

2.1. The printout above shows the equation for the regression line (of *y* on *x*).

> (a) Carefully plot this line through the data points in Printout B from Problem 1.1. Where does the x you made in Problem 1.2(a) fall relative to the regression line?
>
> (b) Compare the actual regression line you just drew on Printout B with the line your drew by eye on Printout A. Was your fitting by eye reasonably successful?
>
> (c) On Printout B locate and circle the observation designated by Minitab as unusual. Did your sketch of the line on Printout A give any warning that this point is unusual?

2.2. (Hand calculator and formulas in the text) This dataset is small and suitable for hand computation. Verify the following quantities by hand computation:

> (a) The *y*-intercept b_0 and the slope b_1 of the regression line.
>
> (b) The two t-statistics for testing whether the true *y*-intercept and slope are 0. Also find the critical values of t for these tests from tables of the t-distribution.
>
> (c) The fitted value ("y-hat") corresponding to observation #7 and the residual for this observation.
>
> (d) The estimate $s_{y|x}$ of the standard deviation σ of the errors e_i.

2.3. Compare the formulas for the sample correlation *r* and the slope b_1 of the regression line. How could you use the results from Minitab's `describe` and `correlation` procedures to find the equation of the regression line with minimal additional hand computation?

2.4. Find the regression line for the regression of *x* on *y*. Solve the result for *y* in terms of *x*. Notice that the result is *not* the same as obtained for the regression of *y* on *x*. Explain why not, in terms of the use of each line for prediction, and in terms of which sums of squared distances of points from the line are minimized in each case. When you draw a line through the data by eye, are you able to make such distinctions?

2.5. (ANOVA Table) Verify the following results in the ANOVA table:

(a) The F-statistic in the ANOVA table is the square of the t-statistic used to test whether the slope is 0, i.e., $6.12^2 = 37.45$. The one-sided F-test, shown in the ANOVA table, is mathematically equivalent to the two-sided t-test for the slope, shown previously.

(b) The mean square for error in the ANOVA table is the square of the estimate of σ.

(c) The total sum of squares is $S_{yy} = 32.10$, the numerator of the variance of the Yields.

3. Prediction and Residual Analysis

The Minitab output corresponding to the two subcommands is discussed in this section.

Prediction. Here we exploit the association between Fertilizer and Yield, expressed by a regression line with slope significantly different from 0, to predict the Yield that will result from a particular level of Fertilizer. We assume, of course, that the other growing conditions that prevailed when the data in Section 1 were collected (soil type, irrigation, temperature, etc.) are unchanged.

The `predict` subcommand specifies a particular level of Fertilizer (x_0), and the output corresponding to this command shows the predicted Yield that will result:

$$\text{Predicted Yield} = \text{Fit} = \text{y-hat} = 10.1 + (1.15)(4.5) = 15.275.$$

This is a *point prediction* of the Yield that results from Fertilizer at the level $x_0 = 4.5$. Of course we cannot say that this is the exact yield that will result. Even if the linear regression model is correct, three things can go wrong:

- Random noise about the line prevents exact prediction.
- We may have incorrectly estimated the true *y*-intercept β_0 of the line.
- We may have incorrectly estimated the true slope β_1 of the line.

These concerns lead us to construct a 95% *prediction interval* extending on either side of our point prediction 15.275. Writing $s_{y|x}$ as s for brevity, the *variance* of the point prediction is:

$$s^2[1 + 1/10 + (4.0 - 4.5)^2/S_{xx}],$$

where 4.0 is the mean of the Fertilizer levels in the data, 4.5 is the Fertilizer level contemplated in the prediction, 10 is the number of data points used to fit the line, and S_{xx} is as computed in Problem 1.2(b). Then the margin for prediction error is the square root of this quantity times $t_{0.025}(8)$.

The three terms in brackets in the expression for the variance of the point prediction correspond, in order, to the three possible difficulties in the bulleted list just above. Consequently, we could reduce the error in estimating the *y*-intercept by taking more observations. Furthermore, the error in estimating the slope is smaller for values of x_0 near the center of the data used determine the regression line.

Minitab gives the end points of the 95% prediction interval as shown below. Thus, in the circumstances described above, we expect the Yield to be between 13.2 and 17.3.

```
Predicted Values

     Fit  StDev Fit         95.0% CI             95.0% PI
  15.275      0.282   (  14.625,  15.925)  (  13.231,  17.319)
```

Note: Do not confuse the *prediction* interval just discussed with the *confidence* interval for the height of the regression line above the point $x_0 = 4.5$. The latter interval is shorter because it does not take account of the noise about the regression line.

Residuals. Finally, we look at the residuals that were put into c3 of the worksheet by the second subcommand (`resid`). One should always examine the residuals from a regression line to see if it is credible that they arose as normally distributed random noise. Trends or cycles in location or spread (with increasing *x* or, where applicable, with increasing time order of the observations), outliers, etc. may indicate that the assumptions of the linear model are not valid. Below we show a plot of the residuals against Fertilizer levels, which shows no remarkable difficulties, except for the unusual observation already noted.

```
MTB > plot c3 c2

      1.0+                                                      *
         -
 C3      -    *
         -                 *
         -                             2            *
      0.0+                                                      *
         -
         -    *
         -                 *
         -
     -1.0+
         -
         -
         -
         -                                          *
     -2.0+
         -
           --------+---------+---------+---------+---------+--------
Fertlzr
                2.40      3.20      4.00      4.80      5.60
```

Problems:

3.1. Prediction intervals.

> (a) Verify the computation of the 95% prediction interval shown in the Minitab printout of this section. (Find the variance of the predicted value, etc.) Also find the 99% prediction interval.
>
> (b) Compute the 95% prediction interval for the Yield corresponding to a Fertilizer level of 5.5. [Do this in the same way you verified the result in Part (a).] Why is this interval longer than the 95% interval in Part (a)? Check your computations using Minitab.
>
> (c) Use Minitab to show that the 95% interval for the predicted Yield corresponding to a Fertilizer level of 10 is longer still. Even so, you should not trust it to be "95% accurate" in practice. Why not? (Would you trust the linear to remain valid no matter how much fertilizer is used?)

3.2. Use the menu path STAT > Basic > Normality to do the Anderson-Darling test for normality. [Answer: P = 0.155, so do not reject the null hypothesis that the residuals come from a normal population.]

3.3. Use the menu path STAT > Regression > Regression > Fitted line plot, Option: prediction bands to make a plot of the regression line through the data, with bands on either side of the fitted line to show prediction intervals. Compare the graph with your answers in Problem 3.1 insofar as possible.

3.4. (This problem requires a modification of the Minitab worksheet used for the comments and problems above. Save your worksheet before making modifications.) Suppose that an 11th data point has Fertilizer level 10 and Yield 17.

(a) Does Minitab show that this is an influential point?

(b) What change does this additional point make in your predicted Yield corresponding to Fertilizer level 4.5? Corresponding to Fertilizer level 10? [Answer: The 95% prediction interval for 4.5 is (12.0, 17.7).]