# Lecture 18

## Regression Diagnostics

Regression diagnostics include using Cook's Distances to find points that exert undue influence over the regression line, looking at your residuals for evidence of lack of normality and heterogeneity of variance, and testing whether the model is really linear or something else. It incorporates topics from both the previous lecture on outliers and leverage points and the next section on transformations.

### Is a line the correct model?

Minitab can perform two different tests to determine if a line is a good model for your data. One looks to see if a separate means model fits the data better. To conduct this test, you need repeated values for your predictor variable - something you can guarantee if you choose the predictor values but not if you don't. We don't get to choose the biparietal diameter of fetuses, but, with 1000 data points, we do have many biparietal diameters that are repeated in the data set. The other looks to see if their is curvature in the data. It divides the predictor values in half (or so) and looks to see if fitting a line to the first half and a line to the second half which meets up with the first half appropriately leads to lines with the same slope or different slopes. If the slopes are different, then the data has curvature which is reported in the output.

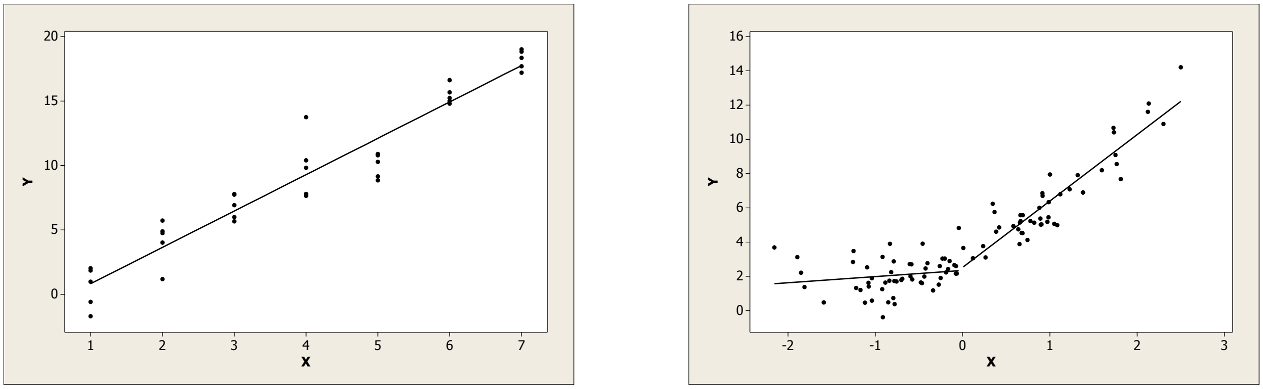

Figure 18.1: Examples when a separate means model may be better than a linear one and when there is curvature in the data that can be determined by the data subsetting procedure.

### Pure Error Lack of Fit Test

For the first method of assessing whether a line is the best model, Minitab considers two separate models. For the Full Model, Minitab fits a separate mean to each distinct predictor ($X$) value. That is, Minitab does a one-way ANOVA with populations determined by the different $X$ values. This is a large model with a lot of parameters generally since we expect a large number of different $X$ values. It also requires the $X$ values to be repeated many times in the data set. For the second model, the Reduced Model, Minitab fits a regression line to the data. Larger models fit better and we expect the sums of squares of their errors to be smaller than that for small models with fewer parameters. This situation is analogous to our discussion of ANOVA from a modeling viewpoint:

| | | | | |
|---|---|---|---|---|
| Population determined by $X$: | $X_1$ | $X_2$ | $\cdots$ | $X_I$ |
| Full Model (separate means): | $\mu_1$ | $\mu_2$ | $\cdots$ | $\mu_I$ |
| Reduced Model (regression): | $a + bX_1$ | $a + bX_2$ | $\cdots$ | $a + bX_I$ |

To compare nested models, one constructs a nested-models F-test. The (mean) difference in the sums of squares explained by the two models goes in the numerator, and the (mean) sums of squares that is the best/smallest (from the full model, of course) goes in the denominator:

$$
\begin{aligned}
F &= \frac{\text{SSE}_{\text{Extra}}/\text{d.f.}_{\text{Extra}}}{\text{SSE}_{\text{Full}}/\text{d.f.}_{\text{Full}}} \\
&= \frac{\left(\text{SSE}_{\text{Reduced}} - \text{SSE}_{\text{Full}}\right) / \left(\text{difference in the number of parameters between the models}\right)}{\text{SSE}_{\text{Full}}/\text{d.f.}_{\text{Full}}}
\end{aligned}
$$

In class, we will examine the pure error lack of fit for predicting femur length from head width. If the data entry error is included, the pure error lack of fit is statistically significant. If we delete the data entry error, then the pure error lack of fit becomes insignificant (p-value $> 0.20$). We will see that the full model sum of squares is exactly what a one-way ANOVA gives.

```
Regression Analysis: FEMUR(normal) versus BPD(normal)

The regression equation is
FEMUR(normal) = - 9.91 + 0.883 BPD(normal)

Predictor          Coef   SE Coef       T      P
Constant        -9.9118    0.3142  -31.55  0.000
BPD(normal)    0.883074  0.008719  101.28  0.000

S = 1.99035   R-Sq = 92.0%   R-Sq(adj) = 91.9%
```

```
Analysis of Variance

Source           DF     SS     MS          F      P
Regression        1  40638  40638   10258.16  0.000
Residual Error  897   3553      4
  Lack of Fit    44    203      5       1.18  0.205
  Pure Error    853   3350      4
Total           898  44191

 7 rows with no replicates
```

What this table says is that the regression on biparietal diameter explains a lot of the variance in the response, femur length. Of the variance that is not explained by the regression, only a statistically insignificant portion would be explained if we allowed separate means for each predictor value. Thus the regression model fits statistically as well as a separate means model.

### Data Subsetting Lack of Fit Test

The data subsetting lack of fit test in Minitab is a combination of several different tests. For simple linear regression, the data subsetting lack of fit test works like this: to check for curvature, break up $X$'s at their center $\bar{X}$. For the full model, fit separate line to the left and right of $\bar{X}$. For the reduced model, fit one line to the data. Compare the two methods using a Nested ANOVA as in the last example. To check for lack of fit at the outer values, fit a line to the central region only. The central region is determined by the leverage of the points, but contains roughly half the data. The full model here uses this central line to estimate the central points and allows every point outside this line to be estimated by itself. The reduced model fits one line to all of the data. Then a Nested ANOVA as in the last example is used. Results from these two tests are combined into one overall p-value using a Bonferoni correction. Since there are 2 tests, the smaller p-value gets doubled in the case of simple linear regression.

```
Lack of fit test
Possible curvature in variable BPD(norm  (P-Value = 0.014 )

Overall lack of fit test is significant at P = 0.014
```

In class, we will see how one determines this value by considering separate regressions to the left and to the right of $\bar{X} = 35.22$ (BPD).

## Exercises for Lecture 18

1. –

2. –