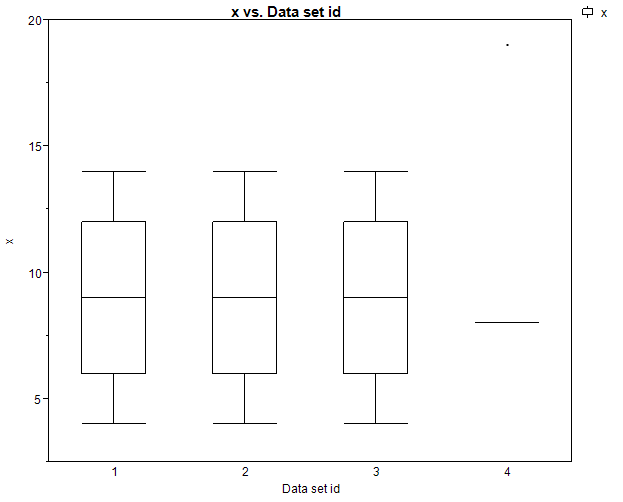
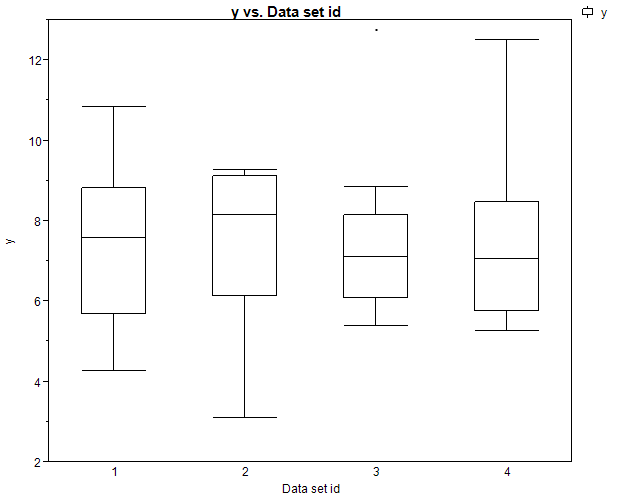
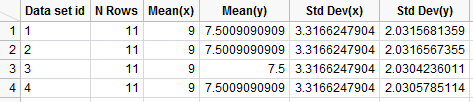
3a) 3pts



3b) 3pts

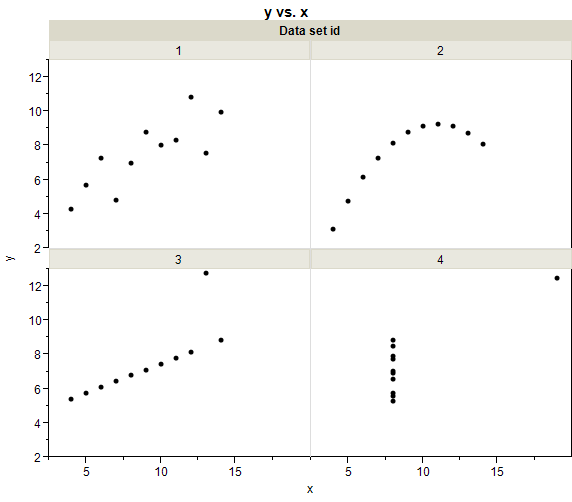


3c) 5pts



The boxplots for the ‘x’ columns looked fairly similar except for the 4th data set. The boxplots for the ‘y’ columns looked somewhat similar. However, the summaries show that all of the data sets have essentially the same values for the means and standard deviations for x and y. So if all we did was look at these summary statistics it would be hard to tell these datasets apart.

3d) 5 pts



3e) 10

**Bivariate Fit of y1 By x1**





**Linear Fit**

y1 = 3.0000909 + 0.5000909\*x1

**Summary of Fit**

|  |  |
| --- | --- |
| RSquare | 0.666542 |
| RSquare Adj | 0.629492 |
| Root Mean Square Error | 1.236603 |
| Mean of Response | 7.500909 |
| Observations (or Sum Wgts) | 11 |

**Bivariate Fit of y2 By x2**





**Linear Fit**

y2 = 3.0009091 + 0.5\*x2

**Summary of Fit**

|  |  |
| --- | --- |
| RSquare | 0.666242 |
| RSquare Adj | 0.629158 |
| Root Mean Square Error | 1.237214 |
| Mean of Response | 7.500909 |
| Observations (or Sum Wgts) | 11 |

**Bivariate Fit of y3 By x3**





**Linear Fit**

y3 = 3.0024545 + 0.4997273\*x3

**Summary of Fit**

|  |  |
| --- | --- |
| RSquare | 0.666324 |
| RSquare Adj | 0.629249 |
| Root Mean Square Error | 1.236311 |
| Mean of Response | 7.5 |
| Observations (or Sum Wgts) | 11 |

**Bivariate Fit of y4 By x4**





**Linear Fit**

y4 = 3.0017273 + 0.4999091\*x4

**Summary of Fit**

|  |  |
| --- | --- |
| RSquare | 0.666707 |
| RSquare Adj | 0.629675 |
| Root Mean Square Error | 1.235695 |
| Mean of Response | 7.500909 |
| Observations (or Sum Wgts) | 11 |

Our regression lines are:

y1 = 3.0000909 + 0.5000909\*x1

y2 = 3.0009091 + 0.5\*x2

y3 = 3.0024545 + 0.4997273\*x3

y4 = 3.0017273 + 0.4999091\*x4

and all of our R^2 values are .666 (to the nearest thousandths place). So the lines and the our “goodness of fit” statistic are essentially identical for all four data sets but looking at the data we can tell that there are big differences between these datasets.

Data set 1: This actually looks good and doing regression with this one is fine

Data set 2: There is a quadratic trend that should definitely be accounted for

Data set 3: This looks perfectly linear except we have one outlier that messes everything up

Data set 4: We only have two values for x. There is only one outcome in the second value for x so the line is almost essentially completely determined by that single point.

3f) 5 pts

Clearly graphical exploration of the data is important because all of our summary statistics for x and y and the relationship between x and y are essentially identical for these four data sets but they are clearly different. We need to graphically explore the data to make sure that the assumptions in our regression are met

4a-c) 6 pts

**Multivariate**

**Correlations**

|  | **Total Bill ($)** | **Tip ($)** | **tip+5** | **tip\*100** |
| --- | --- | --- | --- | --- |
| Total Bill ($) | 1.0000 | 0.6757 | 0.6757 | 0.6757 |
| Tip ($) | 0.6757 | 1.0000 | 1.0000 | 1.0000 |
| tip+5 | 0.6757 | 1.0000 | 1.0000 | 1.0000 |
| tip\*100 | 0.6757 | 1.0000 | 1.0000 | 1.0000 |

The correlations of interest are all .6757. Linear correlation isn’t affected by taking a linear function of the variables of interest which is a nice property because if we want a measure of the strength of the association between different variables we don’t want it to be changed based on the units we measure out outcomes in.

4d) 5 pts

**Bivariate Fit of StdTip By StdBill**





**Linear Fit**

StdTip = -3.53e-15 + 0.6757341\*StdBill

The slope in this regression using the standardized variables equals the correlation between the two variables. This fact holds true in general – If we standardize two variables then the estimated slope for the regression using the two variables (it doesn’t matter which is the predictor and which is the response) will be the correlation between the two variables and the intercept will be 0.

4e) 3 pts

**Linear Fit**

Tip ($) = 0.9202696 + 0.1050245\*Total Bill ($)

4f) 3pts

**Linear Fit**

Total Bill ($) = 6.7502838 + 4.3477142\*Tip ($)

4g) 10 pts

Rearranging Total Bill ($) = 6.7502838 + 4.3477142\*Tip ($) gives:

(Total Bill ($) - 6.7502838)/ 4.3477142 = Tip ($)

Which reduces to

0.2300059\*Total Bill ($) -1.552605 = Tip ($)

Which is not the same as the fit in (4e).

They aren’t the same because the prediction equations were fit using different loss functions so we shouldn’t expect that the lines will be identical. The regression predicting Tip tried to minimize the squared error loss between the predictions and Tip; the regression predicting Bill tried to minimize the squared error loss between the predictions and Bill. So we shouldn’t expect them to give the same fit line.