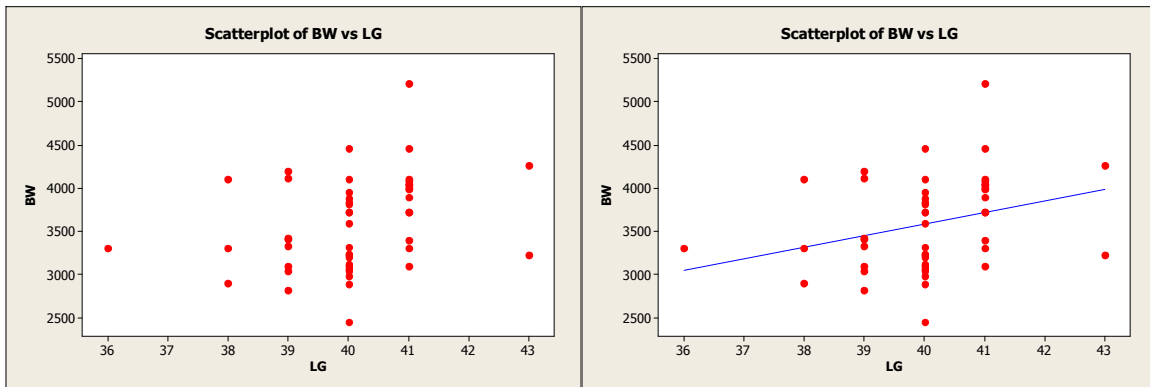


Homework 9, due Thursday November 17 in lab **Solutions**

(assignment for these solutions can be found on the last page)

1. Plot the birth weight (BW) against the length of gestation (LG). Describe the relationship. Looking at the plot, should the sample correlation between BW and LG be positive, negative, or nearly zero?

There appears to be a very slight positive correlation between BW and LG, since there is a slight increasing trend in the graph. The second plot is the same, but with the regression line asked for in problem (3).



2. Compute the Pearson and Spearman correlations between BW and LG. Comment. Test the hypothesis that the population correlation between BW and LG is zero. Comment on the tests.

Below, the Pearson correlation is 0.3, and the Spearman correlation is 0.33.

Testing the hypothesis of no correlation:

$H_0: \rho=0$ (no correlation between BW and LG), vs.

$H_A: \rho \neq 0$ (a non-zero correlation between BW and LG).

Because the p -values below (0.039 or 0.022) are less than $\alpha=0.05$, we reject H_0 in favor of H_A concluding that there is a non-zero correlation between length of gestation and birth weight.

Correlations: LG, BW

Pearson correlation of LG and BW = 0.300
 P-Value = 0.039

Correlations: R_LG, R_BW

Spearman using ranks

Pearson correlation of R_LG and R_BW = 0.330
 P-Value = 0.022

3. Provide an equation for the least squares line for predicting BW from LG. Test the hypothesis that the slope of the population regression line is zero. [We can think of this as a test that LG is important for explaining the observed variation in

BW]. Superimpose the LS line on the data plot and comment on whether the simple linear regression model appears to adequately summarize the relationship between BW and LG.

The least squares line is $BW = -1787 + 134 LG$. The test for whether the slope parameter of the population regression line is given by the p-value on the LG line. Note that this p-value is the same as the p-value for the correlation coefficient above. At 0.039, it is small enough to reject H_0 in favor of H_A that the slope is different from 0. The plot in question (1) has the least squares line superimposed. It appears to explain the mild increasing relationship between BW and LG.

Regression Analysis: BW versus LG

The regression equation is
BW = - 1787 + 134 LG

| Predictor | Coef | SE Coef | T | P |
|-----------|--------|---------|-------|--------------|
| Constant | -1787 | 2527 | -0.71 | 0.483 |
| LG | 134.28 | 63.07 | 2.13 | 0.039 |

S = 519.765 **R-Sq = 9.0%** R-Sq(adj) = 7.0%

4. *What percentage (or proportion) of the variability in BW is explained by the linear relationship between BW and LG?*

$R^2=0.09$, which is a very weak linear relationship between BW and LG.

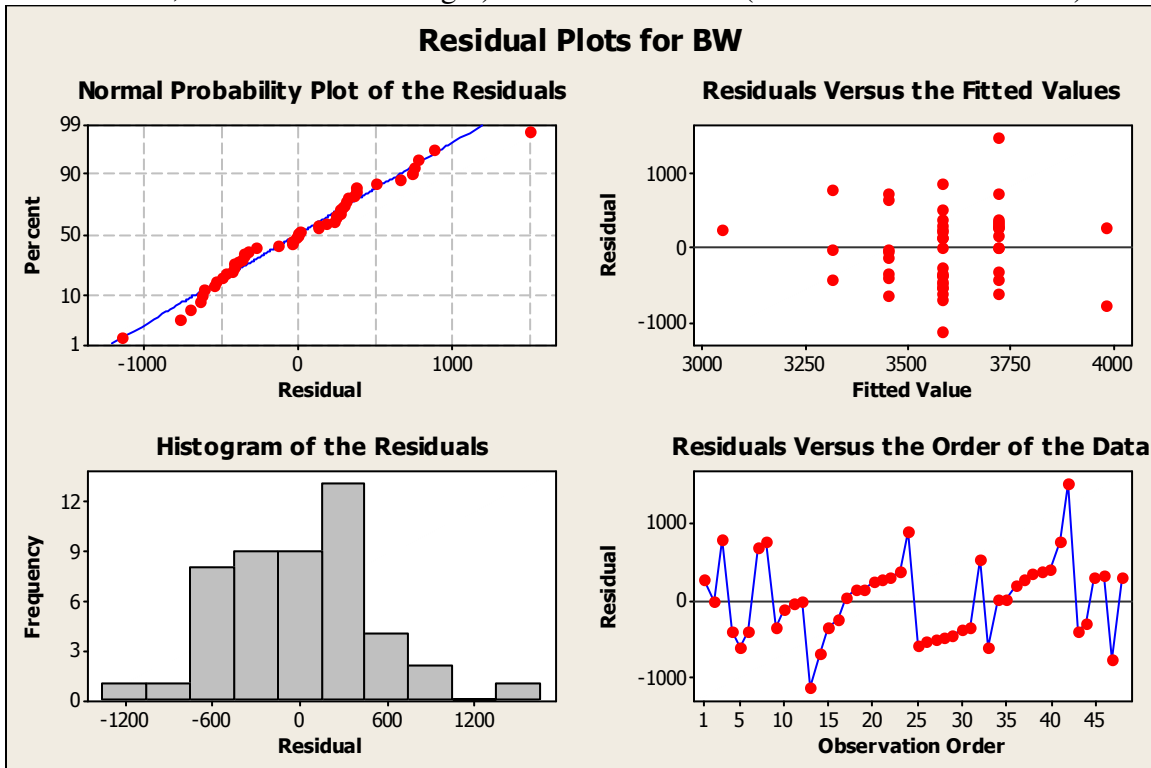
5. *Make appropriate residual plots to check for inadequacies with the model. Comment on the plots.*

The two plots on the left are checking for whether the residuals are normally distributed. The Normal probability plot shows the points following the line rather closely, suggesting that they are normal. The Histogram looks unimodal, symmetric, with no outliers, so it too suggests the residuals are normal.

The residuals versus the fitted values scatter plot is checking for constant variance across values of the x-variable and whether there is any pattern. We certainly have more values in the middle than in the tails of this plot, however, there doesn't appear to be strong evidence that the variance is different for different values of LG. Also, the values have no apparent structure above and below the 0-line, which is what we want. There may be a few outliers, however. The lowest value in the middle group and the highest value in the middle-right group appear far from the center. We could check whether these values are correct by double-checking data-entry, checking with the mother, etc.

The last plot by order of the data is checking for trends in our data collection. Because our data has been sorted, this plot is not useful to us. Otherwise, we might look for systematic patterns in the data.

Below the plot I included the unusual observation output from the regression procedure. It finds the two outliers (R) (outliers in the y direction) and points out the values (the one to the far left, and two to the far right) that are influential (outliers in the x direction).



Unusual Observations

| Obs | LG | BW | Fit | SE Fit | Residual | St Resid |
|-----|------|--------|--------|--------|----------|----------|
| 1 | 36.0 | 3300.0 | 3047.3 | 265.7 | 252.7 | 0.57 X |
| 13 | 40.0 | 2450.0 | 3584.4 | 75.1 | -1134.4 | -2.21R |
| 42 | 41.0 | 5220.0 | 3718.7 | 96.3 | 1501.3 | 2.94R |
| 47 | 43.0 | 3220.0 | 3987.3 | 201.1 | -767.3 | -1.60 X |
| 48 | 43.0 | 4270.0 | 3987.3 | 201.1 | 282.7 | 0.59 X |

R denotes an observation with a large standardized residual.
 X denotes an observation whose X value gives it large influence.

Stat 538, 2005: Homework 9, due November 17th in lab

SIMPLE LINEAR REGRESSION:

The following data were collected from 48 women who were at least 40 years old when they gave birth to their first child. The data concern the gestation period of that pregnancy, and related variables on the child and mother.

The columns are, from left to right:

- 1) ID
- 2) The child's gestation period, in weeks
- 3) Sex of the child (0=Male, 1=Female)
- 4) Birth Weight of child, in grams
- 5) Number of cigarettes smoked per day (on average) by the mother
- 6) Height of mother in cm
- 7) Weight of mother in kilograms at first prenatal visit
- 8) Weight of mother in kilograms at final prenatal visit

| | | | | | | | |
|----|----|---|------|----|-------|------|-------|
| 1 | 36 | 0 | 3300 | 0 | 160.0 | 67.3 | 82.7 |
| 2 | 38 | 0 | 3300 | 60 | 167.6 | 52.7 | 76.0 |
| 3 | 38 | 0 | 4100 | 20 | 167.6 | 64.2 | 79.6 |
| 4 | 38 | 1 | 2900 | 10 | 163.9 | 72.7 | 95.8 |
| 5 | 39 | 0 | 2820 | 0 | 161.3 | 50.0 | 63.3 |
| 6 | 39 | 0 | 3040 | 0 | 158.8 | 49.1 | 61.5 |
| 7 | 39 | 0 | 4120 | 0 | 160.0 | 57.7 | 73.5 |
| 8 | 39 | 0 | 4200 | 0 | 174.0 | 68.0 | 86.8 |
| 9 | 39 | 1 | 3100 | 0 | 171.5 | 67.3 | 85.6 |
| 10 | 39 | 1 | 3330 | 0 | 160.0 | 74.0 | 90.5 |
| 11 | 39 | 1 | 3410 | 0 | 165.1 | 55.9 | 70.7 |
| 12 | 39 | 1 | 3420 | 0 | 162.6 | 52.3 | 66.0 |
| 13 | 40 | 0 | 2450 | 20 | 167.6 | 61.4 | 72.5 |
| 14 | 40 | 0 | 2885 | 0 | 167.7 | 60.0 | 78.6 |
| 15 | 40 | 0 | 3235 | 0 | 170.2 | 50.0 | 65.5 |
| 16 | 40 | 0 | 3320 | 0 | 165.1 | 63.6 | 80.2 |
| 17 | 40 | 0 | 3600 | 0 | 165.1 | 53.2 | 68.7 |
| 18 | 40 | 0 | 3720 | 0 | 165.0 | 57.7 | 74.4 |
| 19 | 40 | 0 | 3720 | 0 | 172.7 | 61.4 | 80.0 |
| 20 | 40 | 0 | 3820 | 0 | 175.3 | 60.8 | 78.1 |
| 21 | 40 | 0 | 3840 | 0 | 167.0 | 60.5 | 83.9 |
| 22 | 40 | 0 | 3880 | 0 | 156.2 | 57.3 | 73.7 |
| 23 | 40 | 0 | 3960 | 0 | 157.5 | 52.7 | 68.2 |
| 24 | 40 | 0 | 4465 | 0 | 157.5 | 51.4 | 66.4 |
| 25 | 40 | 1 | 2980 | 0 | 160.0 | 47.7 | 55.2 |
| 26 | 40 | 1 | 3040 | 0 | 162.0 | 49.0 | 60.3 |
| 27 | 40 | 1 | 3060 | 20 | 157.5 | 61.0 | 75.0 |
| 28 | 40 | 1 | 3100 | 0 | 170.2 | 55.5 | 64.6 |
| 29 | 40 | 1 | 3120 | 0 | 160.3 | 56.8 | 75.4 |
| 30 | 40 | 1 | 3205 | 0 | 172.7 | 58.2 | 75.5 |
| 31 | 40 | 1 | 3220 | 0 | 170.0 | 64.6 | 86.0 |
| 32 | 40 | 1 | 4100 | 40 | 167.0 | 67.0 | 85.0 |
| 33 | 41 | 0 | 3100 | 0 | 168.9 | 61.4 | 69.2 |
| 34 | 41 | 0 | 3720 | 0 | 170.2 | 57.7 | 67.7 |
| 35 | 41 | 0 | 3720 | 20 | 170.2 | 57.7 | 80.5 |
| 36 | 41 | 0 | 3900 | 0 | 167.0 | 68.0 | 85.4 |
| 37 | 41 | 0 | 3990 | 0 | 165.1 | 52.3 | 71.2 |
| 38 | 41 | 0 | 4050 | 0 | 167.6 | 61.0 | 78.5 |
| 39 | 41 | 0 | 4080 | 0 | 162.6 | 59.1 | 83.1 |
| 40 | 41 | 0 | 4100 | 0 | 165.1 | 60.5 | 86.5 |
| 41 | 41 | 0 | 4460 | 20 | 165.1 | 56.8 | 88.0 |
| 42 | 41 | 0 | 5220 | 0 | 157.5 | 56.8 | 68.2 |
| 43 | 41 | 1 | 3300 | 40 | 162.6 | 74.1 | 89.7 |
| 44 | 41 | 1 | 3400 | 0 | 172.7 | 71.4 | 87.8 |
| 45 | 41 | 1 | 4000 | 0 | 165.1 | 90.0 | 100.8 |
| 46 | 41 | 1 | 4030 | 0 | 166.0 | 63.0 | 95.3 |
| 47 | 43 | 1 | 3220 | 0 | 166.4 | 60.9 | 72.0 |
| 48 | 43 | 1 | 4270 | 0 | 162.6 | 54.5 | 70.3 |

- 1) Plot the birth weight (BW) against the length of gestation (LG).

Describe the relationship. Looking at the plot, should the sample correlation between BW and LG be positive, negative, or nearly zero?

- 2) Compute the Pearson and Spearman correlations between BW and LG. Comment. Test the hypothesis that the population correlation between BW and LG is zero. Comment on the tests.
- 3) Provide an equation for the least squares line for predicting BW from LG. Test the hypothesis that the slope of the population regression line is zero. [We can think of this as a test that LG is important for explaining the observed variation in BW]. Superimpose the LS line on the data plot and comment on whether the simple linear regression model appears to adequately summarize the relationship between BW and LG.
- 4) What percentage (or proportion) of the variability in BW is explained by the linear relationship between BW and LG?
- 5) Make appropriate residual plots to check for inadequacies with the model. Comment on the plots.