

Stat 538 - Biostatistics I - Fall 2005

Lab 10

Correlation and Regression

Minitab can provide scatterplots, correlation coefficients, and perform regression. Scatterplots are found under Graph/Scatterplot, Correlation is found under Stat/Basic Statistics/Correlation, and Regression is found under Stat/Regression.

Today's lab we do twice, once detailed with one dataset, then a second time with another dataset. The two datasets are listed here, and are also on the Labs website in the Ch 12 data.

Ch 12/peakflow Ex. 12.10 p.540

peakflow	ht
733	174
572	183
500	176
738	169
616	183
787	186
866	178
670	175
550	172
660	179
575	171
577	184
783	200
625	195
470	176
642	176
856	190

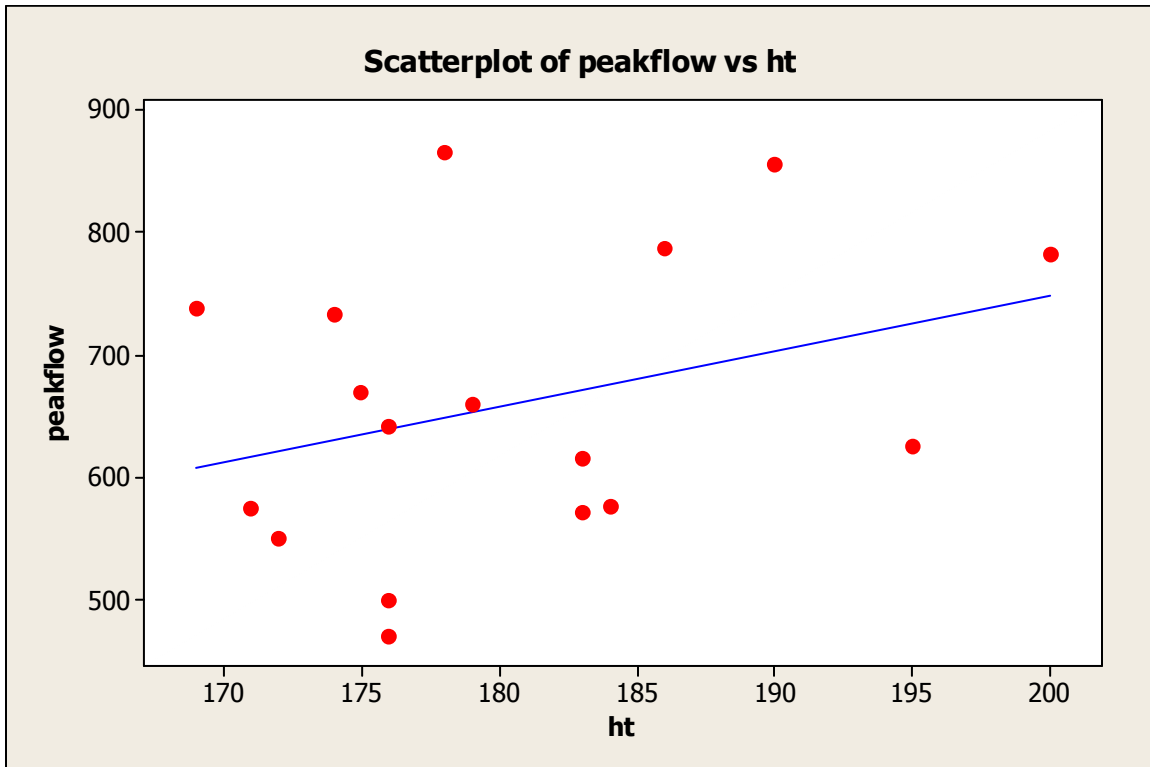
Ch 12/rowan Ex. 12.8 p.539

resp-rate	altitude
.11	90
.20	230
.13	240
.15	260
.18	330
.16	400
.23	410
.18	550
.23	590
.26	610
.32	700
.37	790

Scatterplot/Correlation

The first thing one should do when looking at data is plot it, and this goes doubly so for relationships between variables within a dataset.

Below we're interested in whether there is a *linear* relationship between the peakflow and ht variables. Below peakflow is plotted against ht, and I have chosen to display the least squares regression line on the plot (by choosing with regression under scatterplot). Without the line there it might be difficult to discern whether there was a positive relationship or no relationship, it seems nearly flat.

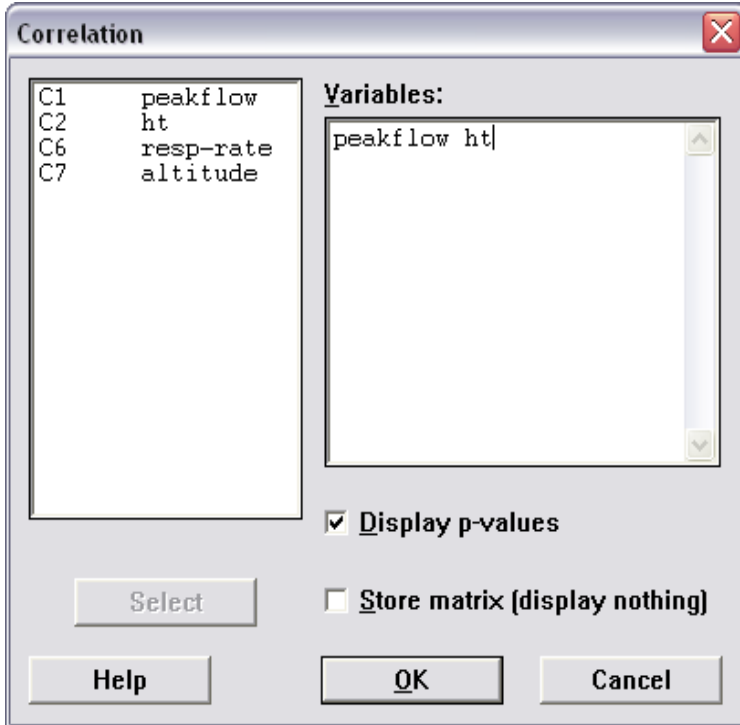


We have been introduced to two correlation coefficients:

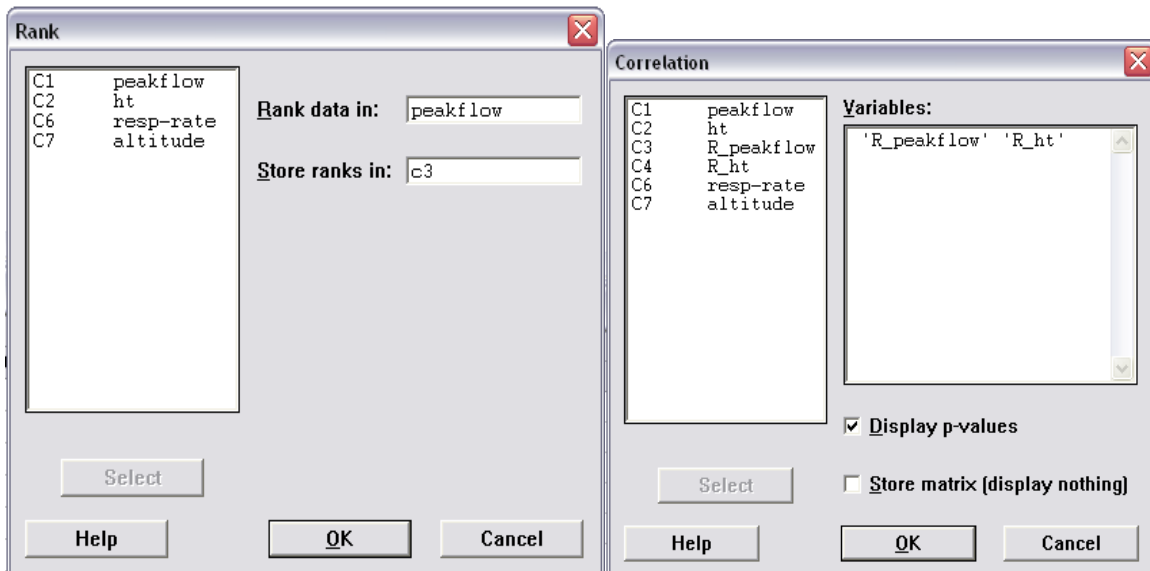
The **Pearson correlation coefficient** measures the linear association between the x and y variables. It is available under Stat/Basic Statistics/Correlation. Simply select the two variables you are interested in.

The **Spearman Rank correlation coefficient** measures the relationship between the ranks of the x and y variables, as a nonparametric test of association. There is no direct way to get these, however it is not difficult. First use Data/Rank, select each variable individually and output the ranks to an empty column. Then calculate the correlation between the two ranked variables.

Here, we find the **Pearson** correlation coefficient between peakflow and ht.



Here we find the **Spearman** rank correlation coefficient by first obtaining the ranks for the two variables (Data/Rank one at a time), then performing the correlation between the newly created rank variables. Remember to label your new variables in a way that reminds you they are the ranks of the original variables. I place a “R_” before each variable name. Choose a convention that works for you.



Below are the results of the two correlations above. The first is the Pearson on the original variables, and the second is the Spearman on the ranks of the variables.

The Pearson correlation coefficient for linear association between peakflow and ht is 0.327. The p -value is testing whether the population correlation coefficient ρ “rho” is different from 0.

$H_0: \rho=0$ (there is no linear relationship between the x and y variables)

$H_1: \rho \neq 0$ (there is a linear relationship between the x and y variables)

In this case, because our p -value is large (p -value=0.2>0.05= α), we fail to reject H_0 concluding there is insufficient evidence that there is a linear relationship between peakflow and ht.

The Spearman correlation is 0.284, and with the large p -value, the same conclusion is reached.

Correlations: peakflow, ht

Pearson correlation of peakflow and ht = 0.327
P-Value = 0.200

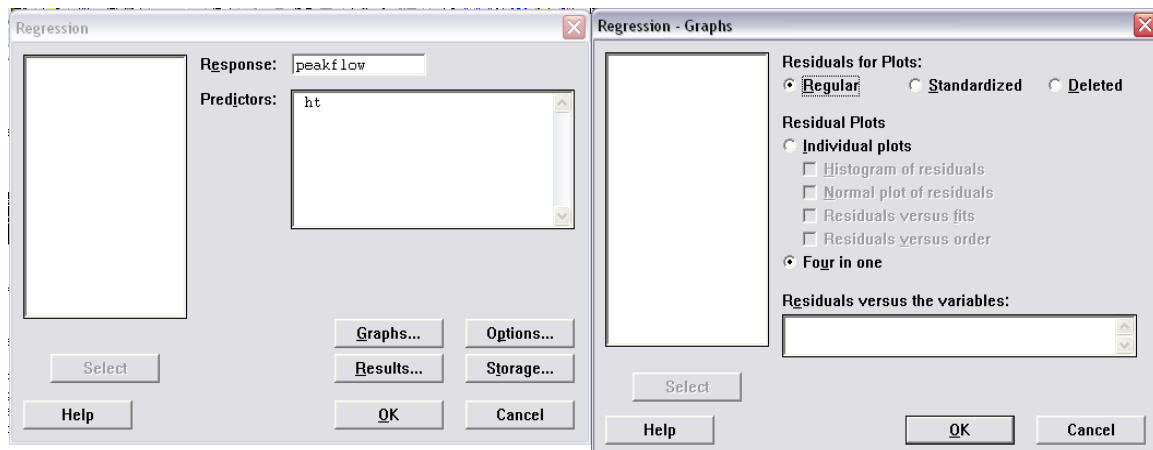
Correlations: R_peakflow, R_ht

Pearson correlation of R_peakflow and R_ht = 0.284
P-Value = 0.269

Regression

The line you see superimposed on the scatterplot at the beginning of the lab is the line that best describes the relationship between the two variables. It is called the least squares regression line because the line is determined by minimizing the sum of the squares of the vertical distances between the points and the line.

Let's regress peakflow on ht (we regress the y response variable on the x predictor variable). In Stat/Regression/Regression, we put peakflow in the response variable and ht in the predictor variable. Under graphs I choose four in one. If desired, you can output the residuals into a column of the worksheet to do a normality test on them. Under the Storage button, select Residuals.



The output below begins with the regression equation. This is the line that appears in the scatterplot. Below that is a parameter estimate table which consists of a constant (the y intercept) and ht (the slope). Notice the Coef values are those from the regression equation. t tests are used to test whether either of these are equal to 0. So there are two hypotheses in this table, testing does the intercept equal 0 or not equal 0, and testing does the slope equal 0 or not equal 0. Typically we are not interested in the intercept, but only the slope. Because the p-values are large for both the slope (0.803) and intercept (0.200), we fail to reject H_0 in both cases, so there is insufficient evidence to conclude the intercept (or slope) is different from 0. Notice that the p-value in this test is the same as the p-value for the Pearson correlation coefficient (0.200). Therefore, the slope we see in the scatterplot is not statistically different from 0.

The $R\text{-Sq}=R^2=0.107$ is the coefficient of determination. It is the proportion of variation in the y variable (peakflow in this case) explained by the regression on x (ht). That is, by knowing the value of x , how much better is your prediction of y – about 10.7% better in this case. It is usually useful for telling you how good your model is for prediction of new observations.

The ANOVA table in the case when we have only 1 x variable (it is possible to have many more) gives us the same result as the t -test above it, with the same p-value of 0.200. It is testing whether the model describes a significant amount of the variation in the y variable. In this case we say it the model does not.

At the end, Minitab points out unusual observations. Observations with an X usually are very different in the x direction, and so will have a strong influence on the regression line. Observations with an R (not shown in this example) are outliers in the y direction. In this case we have one influential observation in the x direction. The ht=200 tells us that it is the observation to the far right in the scatterplot.

Regression Analysis: peakflow versus ht

The regression equation is
peakflow = - 154 + 4.51 ht

Predictor	Coef	SE Coef	T	P
Constant	-153.9	607.5	-0.25	0.803
ht	4.511	3.364	1.34	0.200

S = 115.155 **R-Sq = 10.7%** R-Sq(adj) = 4.8%

Analysis of Variance

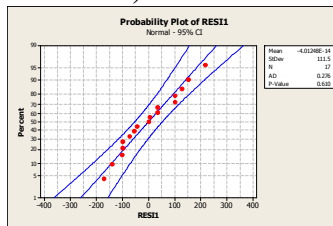
Source	DF	SS	MS	F	P
Regression	1	23857	23857	1.80	0.200
Residual Error	15	198909	13261		
Total	16	222766			

Unusual Observations

Obs	ht	peakflow	Fit	SE Fit	Residual	St Resid
13	200	783.0	748.4	71.6	34.6	0.38 X

X denotes an observation whose X value gives it large influence.

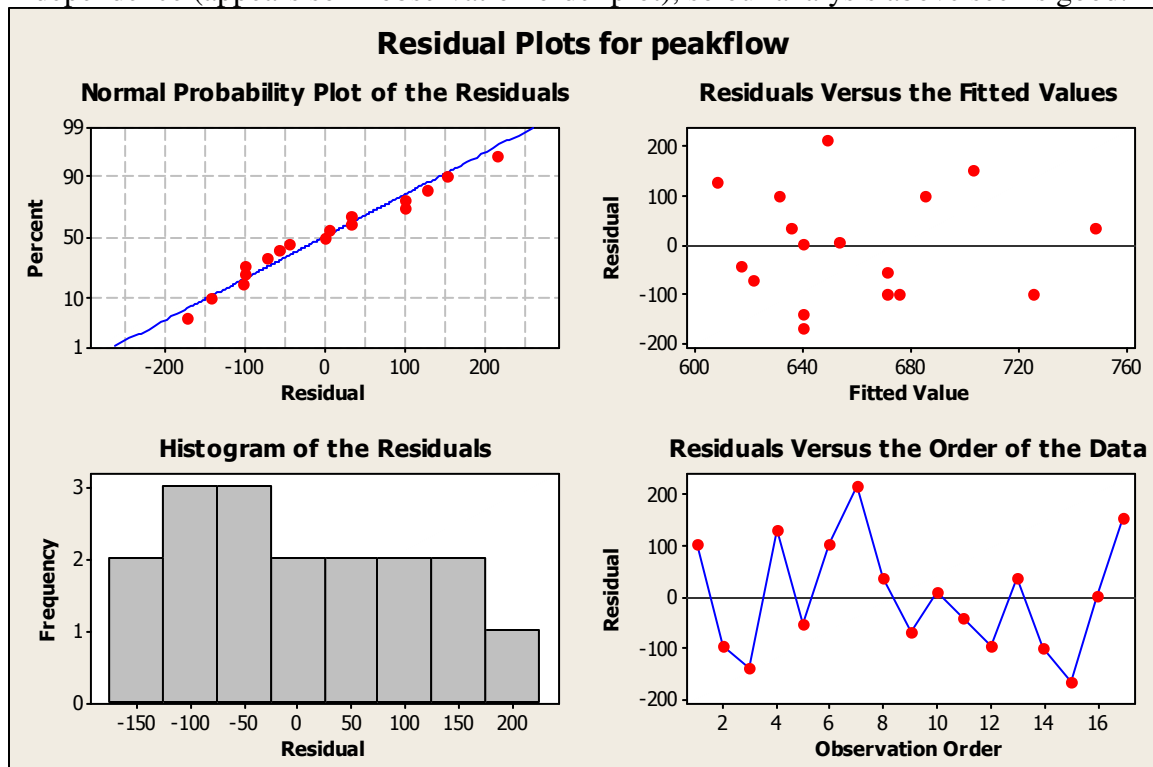
Below is the four in one residual plot. The two plots on the left are checking for whether the residuals are normally distributed. The Normal probability plot shows the points have a downward curvature, so do not follow the line closely, suggesting that they are not normal. The Histogram below it looks unimodal, rather uniform and mildly skewed right, with no outliers, so it too suggests the residuals are not quite normal. If we wanted to do a formal Anderson-Darling normality test, we could output the residuals and to the normal probability plot on the residuals. (If you do this, you find the residuals are sufficiently normal with a $p\text{-value}=0.610$.)



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. There doesn't appear to be strong evidence that the variance is different for different values of x . Also, the values have no apparent structure above and below the 0-line, which is what we want. There does not appear to be any outliers.

The last plot by order of the data is checking for trends in our data collection. There does not appear to be any systematic patterns in the data.

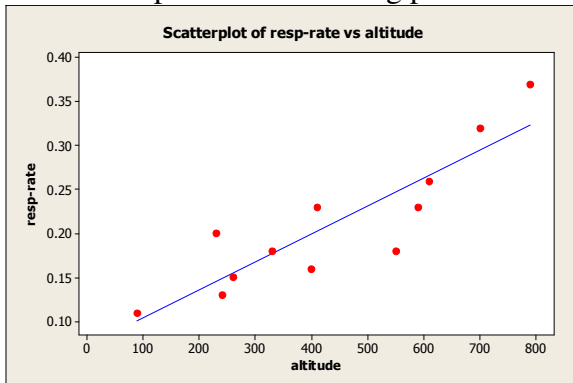
In conclusion, the residuals satisfy our assumptions of normality, constant variance, and independence (appears so in observation order plot), so our analysis above seems good.



End of the first example.

Let's repeat the above analysis with the second dataset.

This scatterplot shows a strong positive linear association between resp-rate and altitude.



Both of the Pearson and Spearman correlation coefficients are larger than 0.85, and are highly significant ($p\text{-value}=0.000 < 0.05 = \alpha$).

Correlations: resp-rate, altitude

Pearson correlation of resp-rate and altitude = 0.887
P-Value = 0.000

Correlations: R_resp-rate, R_altitude

Pearson correlation of R_resp-rate and R_altitude = 0.860
P-Value = 0.000

The regression below gives the linear relationship between our two variables,
 $\text{resp-rate} = 0.0720 + 0.000319 \text{ altitude}$
 and a test for whether either the intercept or slope parameters are 0 rejects in both cases, concluding that both are different from 0. The $R^2=0.786$, so much of the variation in the response variable resp-rate is described by its linear relationship with predictor variable altitude. The ANOVA $p\text{-value}=0.000$ also confirms that this model describes a significant amount of variation in our y variable.

Also, there are no unusual observations, which is why there's no output beyond the ANOVA.

Regression Analysis: resp-rate versus altitude

The regression equation is
 $\text{resp-rate} = 0.0720 + 0.000319 \text{ altitude}$

Predictor	Coef	SE Coef	T	P
Constant	0.07196	0.02520	2.86	0.017
altitude	0.00031855	0.00005254	6.06	0.000

S = 0.0373973 R-Sq = 78.6% R-Sq(adj) = 76.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.051414	0.051414	36.76	0.000
Residual Error	10	0.013986	0.001399		
Total	11	0.065400			

The residual plots show normal residuals, there's not strong evidence for any non-constant variance or patterns. All good signs. A double-check of normality indicates it's ok (p-value=.984).

All indications suggest that the linear relationship in the regression equation describe well the relationship between resp-rate and altitude.

