SOME NOTES ON REGRESSION ANALYSIS - SETUP OF THE MODEL

We have a linear regression model of the following form:

$$\Delta y_i = c + a_1 prod_i + a_2 proc_i + a_3 y_i(-1) + a_4 gen_i + a_5 g24 + e_i \qquad (1)$$

where

$\Delta y_i \equiv y_i - y_i(-1)$ denotes our dependent variable, the cumulative change in platelet counts for donor i, i = 1, 2, . . . , 74. y_i denotes the end count and $y_i(-1)$ denotes the beginning or base count.

c denotes a constant, which you estimate. It turns out to be equal to the mean value across donors of that which you cannot explain with your independent variables (such as number of products and procedures). If you had no independent variables, your estimate of the constant term would equal mean value of Δy, which for your sample equals −5.216.

$prod_i$ denotes the number of products observed for donor i.

$proc_i$ denotes the number of procedures observed for donor i.

gen_i is a variable that takes on one of two values: if the donor is a female, it equals 1, and otherwise it equals 0.

g24 similarly takes on one of two values, 1 if the number of products is equal to or greater than 24 and 0 otherwise.

e_i is the error for donor i, the difference between the true value of the cumulative platelet count (what was recorded at the hospital) and what our estimated model predicts.

The regression technique we use, Ordinary Least Squares, chooses the coefficients, c, a_1, a_2, . . . , a_n, to minimize the sum of squared errors, in our problem, $sigma^2_e \equiv \Sigma^{74}_{i=1} e^2_i$, at the same time, the sum of the errors equals 0: $\Sigma^{74}_{i=1} e_i = 0$.

Some terminology:

R-squared: measures the fraction of the sample variance of Δy, $\hat{\sigma}^2_{\Delta y}$, that can be explained by the estimated model.

$$R - squared = \frac{\Delta y, \hat{\sigma}^2_{\Delta y} - \Delta y, \hat{\sigma}^2_e}{\hat{\sigma}^2_{\Delta y}} \in [0, 1] \qquad (2)$$

If you divide the sum of the squared resid by 74, you will find $\hat{\sigma}^2_e$. And, if you square, S.D. dependent variable you will find $\hat{\sigma}^2_{\Delta y}$. Then you can verify the value of R-squared.

The t-statistic is the value from a certain distribution (a students-t distribution) under the null-hypothesis that the estimated coefficient is equal to zero. The corresponding probability shown tells us how likely it is, in light of the data, that null-hypothesis is true. A low probability (less than 5 percent or 1 percent) means that we can reject the null hypothesis. This means that the variable in question matters, as indicated by the estimated coefficient.

The std. error reported next to each coefficient indicates the precision with which you have estimated that coefficient. If it is large relative to the magnitude of the value of your estimated coefficient, you are not likely to be able to reject the hypothesis that the variable in question has no impact on the dependent variable (Δy). Such an outcome is more likely when the variable (for example dg24) shows very little variation in the sample.

Some interpretation of the coefficients

In the second regression model, the estimated coefficient on the variable proc (procedures) was significantly positive. That the coefficient is positive indicates that, when comparing any two donors who are identical in all respects but the number of procedures, the one with the larger value of proc will likely have a larger value for Δy. This implication says nothing about whether it will be positive or negative.[1] If the two donors had $\Delta y > 0$, then the one with the greater number of procedures will likely have a larger increase in her platelet count. If the two patients had $\Delta y < 0$, then the one with the greater number of procedures will likely have a smaller decrease in her platelet count.

In the first regression, the estimated coefficient on the variable y(−1) is negative, and significantly so. That the coefficient is negative indicates that, when comparing any two donors who are identical in all respects but the beginning platelet count, the patient having the higher count will exhibit a smaller change, Δy. If both patients had negative changes, $\Delta y < 0$, then the one with the larger value for y(−1) will likely show a larger drop in her count. If both have positive values for Δy, then the one with the large value for y(−1) will likely have a smaller positive change. Either way, the estimated coefficient is consistent with a "convergence" story for platelet counts, that there is tendency to move towards some "normal count value."

The fact that when we replace y(−1) with y and the coefficient on y is of the opposite sign would be consistent with this convergence result since, by definition, $\Delta y = y - y(-1)$.

Including the variable gen (gender) allows us to estimate a separate constant term for women.

When gen = 0 including all donors,

$$\Delta y_i = c + a_1 prod_i + a_2 proc_i + a_3 y_i(-1) + a_5 g24_i + e_i \qquad (3)$$

and when gen = 1 (only for the females sampled),

$$\Delta y_i = [c + a_4] + a_1 prod_i + a_2 proc_i + a_3 y_i(-1) + a_5 g24_i + e_i \qquad (4)$$

The estimated constant term for each of the models is positive; that the coefficient on gen is positive indicates that the constant term for women is even larger($c+a_4$). Holding all else the same, a female exhibits a larger Δy.

The other dummy variable, g24 (> 24 products), has no explanatory power, even when included in the later models.

In tables 4-6 we allow the gender variable to interact with the number of procedures, proc. Doing so allows us to estimate a separate coefficient on this variable for females as well as a separate constant term.

When gen = 0

$$\Delta y_i = c + a_1 prod_i + a_2 proc_i + a_3 y_i(-1) + a_5 g24_i + e_i \qquad (5)$$

When gen = 1 (only for the females sampled),

$$\Delta y_i = (c + a_4) + a_1 prod_i + (a_2 + a_6) proc_i + a_3 y_i(-1) + a_5 g24_i + e_i \qquad (6)$$

That the coefficient a_6 is positive indicates that the effect of the number of procedures on Δy is greater for females in our sample. The full effect is $a_2 + a_6$.

[1]Note, however, that the mean value for Δy in your sample is −5.216.