Read M&M §2.1 through M&M §2.4, and M&M §2.7. Skip M&M §2.5 (maybe discussion of logarithmic transformation in Sections?). Postpone M&M §2.6

Association between two variables. Scatterplots. Linear association. Correlation. Least squares fit of a straight line. Relationship between slope of least squares line and correlation. Interpretation of r^2. Traps and difficulties: outliers, influential points, lurking variables.

1. Reminder about standardized variables

For values $x_1, x_2, \ldots, x_N$,

$$\text{mean of the } x\text{ 's} = \bar{x} = \frac{1}{N}\sum_i x_i$$

$$\text{variance of the } x\text{ 's} = s_x^2 = \frac{1}{N-1}\sum_i (x_i - \bar{x})^2$$

Notice that I have added a subscript x to the s^2 to indicate that it is calculated from the x's, because for today's lecture there will be several variables floating about. The standard deviation s_x equals the square root of the variance. To standardize the x's we subtract off the mean then divide by the standard deviation. I will write a twiddle (a tilde, that is) over the variable to indicate that it has been standardized:

$$\widetilde{x}_i = \frac{x_i - \bar{x}}{s_x}$$

Recall that the standardized values have zero mean and variance one:

$$\frac{1}{N}\sum_i \widetilde{x}_i = 0 \qquad \text{and} \qquad \frac{1}{N-1}\sum_i (\widetilde{x}_i - 0)^2 = 1$$

For most of today's lecture, for the purpose of explicit plots or calculations, $x_1, x_2, \ldots, x_N$, with $N = 39$, will denote the classwork scores for 39 students in the Yale grades data set. Of course the general principles will also apply to other data sets, but I find it helps to keep a concrete case in mind. Similarly I will write $y_1, y_2, \ldots, y_{39}$ for the final exam scores, with $\widetilde{y}_i = (y_i - \bar{y})/s_y$ for the standardized values. That is,

$$x_i = \text{ classwork score for } i\text{th student}$$
$$y_i = \text{ score on final exam for } i\text{th student.}$$

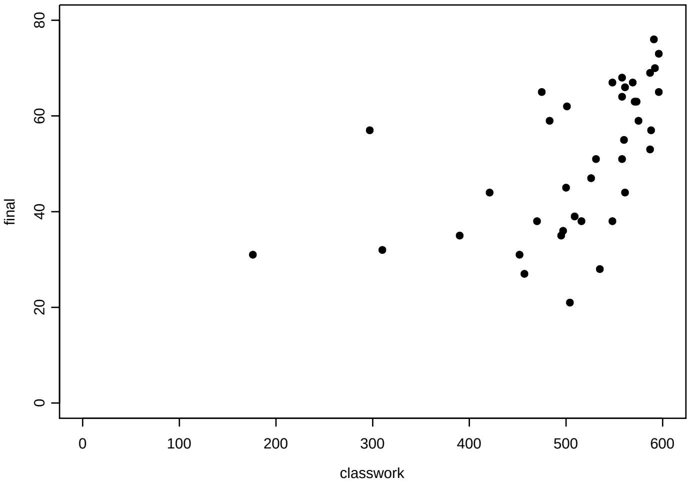
Yale grades	$i=1$	$i=2$	...	$i=38$	$i=39$
x_i	176	297	...	596	596
$\widetilde{x}_i$	-3.71	-2.37	...	0.94	0.94
y_i	31	57	...	73	65
$\widetilde{y}_i$	-1.32	0.40	...	1.45	0.92

For these variables, $\bar{x} = 510.8$, $\bar{y} = 51$, $s_x = 90.3$, and $s_y = 15.1$.

2. Association between variables

I am fond of telling students that their performances on the final exam can be well predicted by their performances on the homework. For the Yale grades data, how well is my assertion borne out?

Look at the ***scatterplot***: represent the ith student by a point with coordinates x_i and y_i. can you find (x_1, y_1) in the plot?



It is clear that there is a positive association between the classwork score and the final exam score: students who do better on the homework tend to do better on the final. In the language of M&M page 105, the classwork score "explains" some of the variation amongst the final exam scores. The relationship is not perfect. It is not invariably true that a larger classwork score implies a larger final score.

How strongly are the classwork and final exam scores associated?

It is traditional to start by looking for simple relationships, perhaps with the idea that inadequacies of simple explanations might suggest more refined descriptions. In general, if the points of a scatterplot were to lie exactly along some straight line we would say that the two variables represented by the coordinates were linearly related. If the points were to lie *close* to some straight line we would be able to describe the relationship as *roughly linear*.

3. Correlation

The correlation, r, between two variables gives a measure of how close they are to being linearly related. It is defined via the standardized variables:

$$r = \frac{1}{N-1}\sum_{i=1}^{N} \widetilde{x}_i \widetilde{y}_i$$

The $N-1$ in the denominator comes from the same place as the $N-1$ in the definition of variance.

To see the intuition behind the formula, suppose the x's and y's had a strong positive association: larger x_i (which give positive values for $\widetilde{x}_i = x_i - \bar{x}$) tend to pair with larger y_i (which give positive values for $\widetilde{y}_i = y_i - \bar{y}$), and smaller x_i tend to pair with smaller y_i. Thus the products $\widetilde{x}_i\widetilde{y}_i$ will all tend to be positive, and r will be tend to be positive. If, however, the x's and y's are not associated we will get positive and negative $\widetilde{x}_i$ each paired up with positive and negative $\widetilde{y}_i$. There will tend to be a lot of cancellation between positive and negative contributions to $\sum_{i=1}^{N} \widetilde{x}_i\widetilde{y}_i$, which will tend to make r close to zero.

If $r = 0$, the variables are said to be uncorrelated.

As you will see below, in general r can never lie outside the range from -1 to $+1$. You will also see why a correlation of exactly $+1$ or -1 implies that the variables are linearly related.

For the Yale grades data, Minitab gives a correlation of 0.53 between classwork and final exam scores. What does that mean?

4. Least squares

Another way to quantify the degree of linear association between two variables is to "fit" a straight line to the scatterplot. In general there will be no straight line that passes exactly through each of the points. Instead we try to find the line $y = a + bx$ that makes the residuals

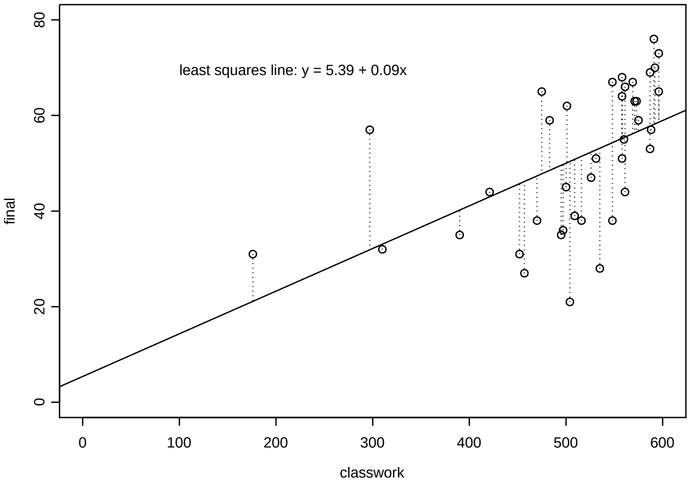
$$\text{RESID}_i = y_i - a - bx_i$$

small. Notice that the residual depends on the choice of the constants a and b, which define the intercept and slope of the line.

The method of least of least squares prescribes that we choose a and b to minimize the sum of squares of the residuals, RESID_i. That is, the method prescribes that we fit the line $y = a + bx$ to the scatterplot by choosing the constants a and b to minimize

$$\sum_{i=1}^{N}(y_i - a - bx_i)^2$$

The residuals, for the a and b that define the elast squares line, are represented by the dotted segments in the following picture.



Why the sum of squares? Why not some other function of the residuals? Primarily it is for mathematical convenience. There are known expressions for the minimizing a and b by least squares, and it is easy to get packages like Minitab to find these values and draw the least squares line. Also the least squares method has some good theoretical properties under assumptions about the mechanisms by which data are generated. Lecture 9 will talk about this theory.

I will explain a little about how the "square" in least squares is related to the "square" in the definition of variance, and how the least squares line is related to the correlation. It is not important that you be able to reproduce all the algebra. It is not important that you memorize formulae for the least squares line, because Mintab will do all the calculaations for you. It is important that you realize that least squares is mostly just a matter of working with the right units, and that the calculations are easy if you look at them the right way. It is important that for you to learn that least squares is a mathematical method that has reasonable properties under a variety of assumptions, but that it can be downright deceptive when misapplied.

The only piece of mathematics you need to know is

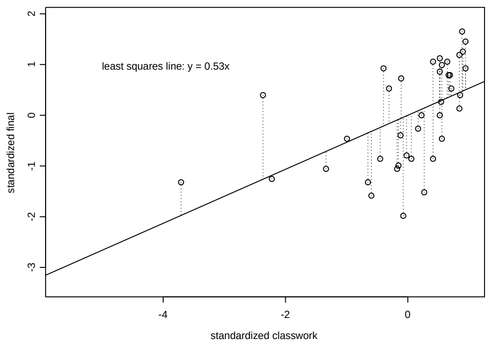
$$(A+B+C)^2 = A^2 + B^2 + C^2 + 2AB + 2BC + 2AC$$

for all choices of A and B and C.

	A	**B**	**C**
A	**A^2**	**AB**	**AC**
B	**AB**	**B^2**	**BC**
C	**AC**	**BC**	**C^2**

5. The least squares line for standardized data

Rescaling of variables should not destroy a linear association between two variables. Think of how it would sound to have a linear association between temperature and libido if temperatures are expressed in degrees fahrenheit but not if temperatures are expressed in degrees celsius. Section 6 argues more precisely, by means of pictures or algebra, but for the moment please just accept that the ground rules are not too much changed if we standardize variables before fitting straight lines.



Least squares is much easier to explain when the variables are standardized, because then the sum of squared residuals can be written out as a very simple expression involving the slope and intercept of the least squares line. As you will see soon, it takes only three lines of algebra to show that the least squares line for the standardized variables has slope equal to the correlation r and that it passes through the origin for the scatterplot. It is also much easier to interpret r^2 for standardized data. It is not vital that you follow the algebra, but it would be a shame if you were to be intimidated by a formula that is so easy to derive.

Simplifications occur because the standardized variables have the following properties:

$$\boxed{\begin{aligned}
&\sum_{i=1}^N \tilde{x}_i = 0 = \sum_{i=1}^N \tilde{y}_i \\
&\frac{1}{N-1}\sum_{i=1}^N \tilde{x}_i^2 = 1 = \frac{1}{N-1}\sum_{i=1}^N \tilde{y}_i^2 \\
&\frac{1}{N-1}\sum_{i=1}^N \tilde{x}_i\tilde{y}_i = r
\end{aligned}}$$

The least squares problem for the standardized data consists of finding constants $\tilde{a}$ and $\tilde{b}$ to minimize

$$\sum_{i=1}^N (\tilde{y}_i - \tilde{a} - \tilde{b}x_i)^2$$

I put the twiddle over the constants so that you will not confuse them with the corresponding constants that define the least squares line for the original data.

Use the only piece of mathematics you need to know to expand the squared residual into a sum of terms:

$$(\widetilde{y}_i - \widetilde{a} - \widetilde{b}\widetilde{x}_i)^2 = \widetilde{y}_i^2 + \widetilde{a}^2 + \widetilde{b}^2\widetilde{x}_i^2 - 2\widetilde{a}\widetilde{y}_i - 2\widetilde{b}\widetilde{x}_i\widetilde{y}_i + 2\widetilde{a}\widetilde{b}\widetilde{x}_i$$

Sum over i, using the relationships in the box to simplify. Also divide by a factor of $N - 1$ to make things look tidier.

$$\begin{aligned}\frac{1}{N-1}\sum_{i=1}^{N}(\widetilde{y}_i - \widetilde{a} - \widetilde{b}\widetilde{x}_i)^2 &= 1 + \widetilde{a}^2\frac{N}{N-1} + \widetilde{b}^2 - (2\widetilde{a})(0) - 2\widetilde{b}r + (2\widetilde{a}\widetilde{b})(0) \\ &= 1 + \widetilde{a}^2\frac{N}{N-1} + (\widetilde{b} - r)^2 - r^2\end{aligned}$$

The last expression is easy to minimize. The contribution from $\widetilde{a}^2$ is always nonnegative, as is the contribution from $(\widetilde{b} - r)^2$. To make both terms as small as possible choose $\widetilde{a} = 0$ and $\widetilde{b} = r$.

Notice that the residuals from the least squares fit sum to zero: by the relationships in the box,

$$\sum_i(\widetilde{y}_i - r\widetilde{x}_i) = \sum_i \widetilde{y}_i - r\sum_i \widetilde{x}_i = 0$$

The mean of the residuals is zero.

Interpretation of r^2

With the standardized variables, the variance of the $\widetilde{y}$'s equals

$$\frac{1}{N-1}\sum_{i=1}^{N}(\widetilde{y}_i - 0)^2 = 1$$

After we remove the linear fit we are left with residuals $\widetilde{y}_i - r\widetilde{x}_i$, with variance

$$\begin{aligned}\frac{1}{N-1}\sum_{i=1}^{N}(\widetilde{y}_i - r\widetilde{x}_i - 0)^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(\widetilde{y}_i^2 - 2r\widetilde{x}_i\widetilde{y}_i + r^2\widetilde{x}_i^2\right) \\ &= \frac{1}{N-1}\sum_{i=1}^{N}\widetilde{y}_i^2 - 2r\frac{1}{N-1}\sum_{i=1}^{N}\widetilde{x}_i\widetilde{y}_i + r^2\frac{1}{N-1}\sum_{i=1}^{N}\widetilde{x}_i^2 \\ &= 1 - 2r^2 + r^2 \qquad \text{by the relationships in the box.}\end{aligned}$$

The ratio of the variances before and after removal of the linear fit equals $1 - r^2$. That is,

> a fraction r^2 of the variance of the (standardized) y's is removed by the fitting of the least squares line.

From this interpretation of r^2 it is clear why r cannot lie outside the range from -1 to $+1$: there is no way to remove more than 100% of the variance.

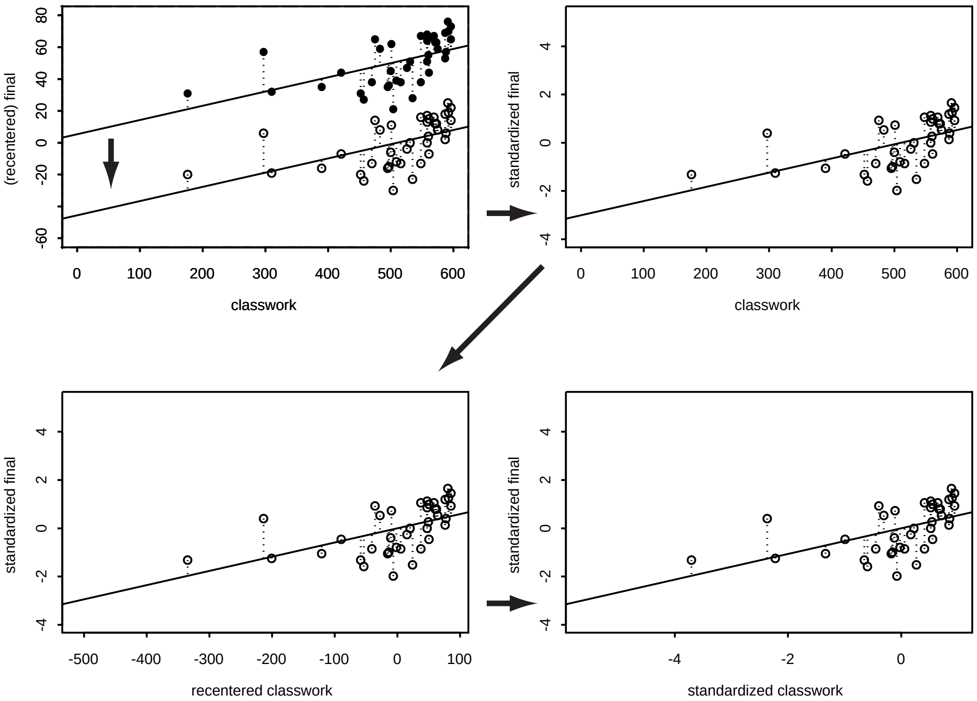
Notice that, in general, if $r = \pm 1$, the variance for the residuals would be zero, and thus every residual would be zero. All the points in the scatterplot would lie along a straight line.

6. Reduction to the standardized case

Let me convince you that we only need to think about the case of standardized variables when we want to understand least squares.

The pictures show the effect on the least squares line of the steps involved in standardizing the variables. The first picture shows that subtraction of $\bar{y}$ from each y_i

merely drops the least squares line by the same amount. The second picture (follow the arrows) shows the effect of dividing each $y_i - \bar{y}$ by s_y: the scale on the vertical axis changes, but all vertical distances change in the same proportion. If I had used the same vertical scale on the top two pictures the slope of the least squares line would have been much smaller—for, indeed, dividing vertical distances by s_y does change the slope of the least squares line by a factor of $1/s_y$. By expanding the vertical scale I have traded a change in the slope of the line, as printed on the page, for a change in the the labels along the vertical axis. If you are confused, try calculating the slope of the line directly from the scales marked along he axes. The bottom two pictures show the analogous effect of centering and rescaling the classroom variable.



If you prefer to think about it algebraically, notice that there is a one-to-one correspondence between the least squares lines for the standardized data and the original (unstandardized) data.

$$\sum_i (\widetilde{y}_i - r\widetilde{x}_i)^2 = \sum_i \left(\frac{y_i - \bar{y}}{s_y} - r\frac{x_i - \bar{x}}{s_x} \right)^2 = \frac{1}{s_y^2} \sum_i \left(y_i - \left(\bar{y} - \frac{r s_y \bar{x}}{s_x} \right) - \frac{r s_y}{s_x} x_i \right)^2$$

The constants

$$a = \bar{y} - \frac{r s_y \bar{x}}{s_x} \qquad \text{and} \qquad b = \frac{r s_y}{s_x}$$

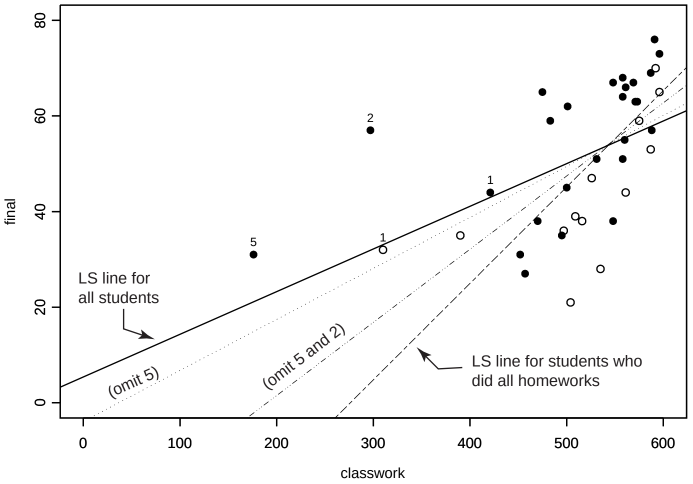
minimize the sum of squares of residuals for the (x_i, y_i)'s. These are the values given by M&M page 141.

7. Departures from the linear fit

What might be disturbing a possible linear association between classwork and final exam scores?

Consider the effect of missed homework sheets. The numbers 5, 2, 1,1 sitting above four of the points in the scatterplot indicate the numbers of missed sheets. If those students had handed in all of the homework, their classwork scores would undoubtedly have been larger. Their points on the scatterplot would have shifted further to the right. (Maybe their final exam scores would have been larger as well, because they would have learned more of the course material.)

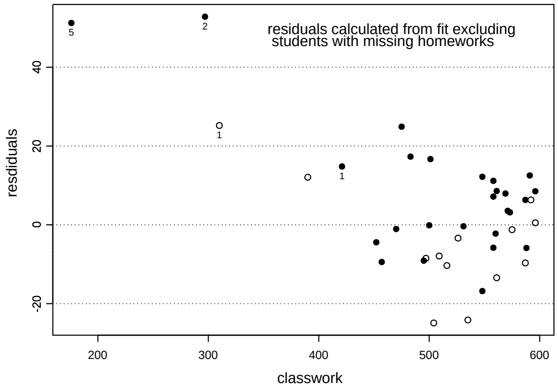
If we are trying to make predictions about exam performance based on classwork scores *for students who did all the homework*, the four students who missed homework will have a distorting effect on the prediction. To illustrate the effect, I have fitted four different least squares lines: one for all students; one based on all except the student who missed 5 sheets; one based on all except the students who missed 5 or 2 sheets; and one for just those students who handed in all homeworks. Notice how the least squares lines respond to the exclusions.



The plot also gives some information that is not contained in the Yale grades data set. The students represented by the solid dots (first exam) took a different final exam from the students represented by the unfilled dots (second exam).

The points labelled 5 and 2 (and the two labelled 1?) are ***outliers*** from the main distribution of classwork and final exam scores. They have a large effect (they have high ***influence***) on the fitted least squares line.

The presence of the outliers was masking an important fact, which can be seen most clearly from an examination of residuals. For this plot I have calculated the least squares fit using only the scores for the 35 students who handed in all the homework, but I have calculated residuals for all 39 students. It is no surprise that the students with missed homeworks should have large residuals. What disturbs me more is the residuals for the students who took the second exam. It appears that almost all of them were doing worse than would be expected fom their classwork scores. The pattern suggests the possibility that the second exam was actually much harder than the first exam.

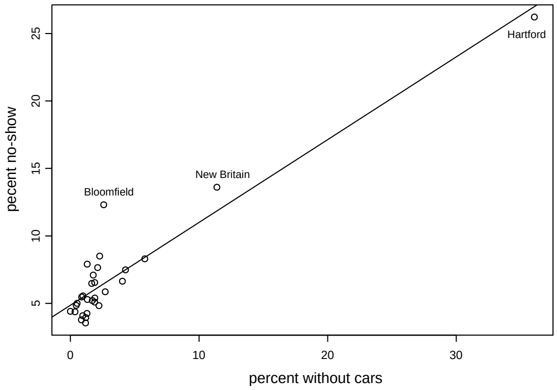


Question: What should I have done after looking at these plots? Is it fair to treat the two exams as equivalent?

8. Causation versus association

Clearly there is quite a strong association between performance on classwork and performance on the final exam. Does it then follow that classwork scores measure something that “causes” final exam scores? Or are both scores merely reflecting some other attribute, such as ability to think statistically, or ability to express oneself clearly in mathematical language? Read M&M §2.7.

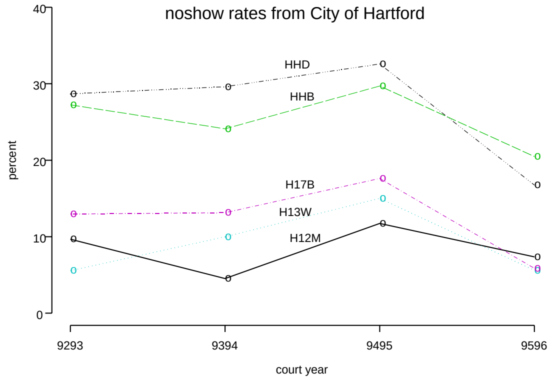
The question of causation is not easily settled. The plot at left was derived from Census data and from juror summonses for the 1994-95 court year, mailed to persons in each of the 29 towns comprising the Hartford-New Britain judicial district. The least square line is shown. The squared correlation (r^2) equals 0.86.



The percentage “no-shows” takes as denominator the number of persons in a town who were assumed to have received a summons and who were not excused for one of a number of acceptable reasons, and as numerator it takes the number of those person who did not present themselves at the court house for jury service. The percentage without cars was calculated from 1990 Census data: for each town it gives the percentage of households, with householder aged between 15 and 64 years, that did not have access to a motor vehicle.

Would one be justified in interpreting such a fit to mean that lack of access to cars was causing the noshow problem?

What other (***lurking***) variables might be affecting the association between no-show rates and lack of access to cars? Is Hartford having too great an influence on the fit? Should we be impressed by the large r^2? Are percentages the right way to measure ‘no show’ and ‘no car’?



What more would one need to look at to gain confidence in a possible causal link between lack of access to cars and a propensity not to turn up at court when summoned? One relevant piece of evidence would be persistance of the effect over time. Another would be the noshow rates by courthouse. If it were lack of access to a car that were causing the no show rate, I would expect to see higher noshow rates at more distant courthouses. For summonses sent to persons in Hartford, just the opposite appears true. (HHD = the main court house, located in the City of Hartford; HHB = the courthouse in the nearby City of New Britain; the other three courthouses are in towns lying further from the City of Hartford.) The data for 1995-96 were incomplete. The apparent downturn in the noshow rates was probably not real.