Assert that best fit line must go through the mean-mean point

Relate back to A20 problems: given $\bar{x}$ and needed to predict $\hat{y}$ [A20pg.3/5a]

In case you're not convinced by the standardized scatterplot or our A20pg.3/5a ...

Here is a formal proof: (Shown in class) The next one we both do!

Where does the equation of the line of best fit come from? To write the equation of any line, we need to know a point on the line and the slope. The point is easy. Consider the protein-fat example. Since it is logical to predict that a sandwich with average protein will contain average fat, the line passes through the point $(\bar{x}, \bar{y})$

To think about the slope, we look once again at the z-scores. We need to remember a few things.

1. The mean of any set of z-scores is 0. This tells us that the line that best fits the z-scores passes through the origin (0, 0)
2. The standard deviation of a set of z-scores is 1, so the variance is also 1. This means that $\frac{\sum(z_y - \bar{z}_y)^2}{n-1} = \frac{\sum(z_y - 0)^2}{n-1} = \frac{\sum z_y^2}{n-1} = 1$, a fact that will be important soon.
3. The correlation is $r = \frac{\sum z_x z_y}{n-1}$, also important soon.

Ready? Remember that our objective is to find the slope of the best-fit line. Because it passes through the origin, its equation will be of the form $z_y = m z_x$. We want to find the value for m that will minimize the sum of squared residuals. Actually we'll divide that sum by $n-1$ and minimize this "mean squared residual," or MSR. Here goes:

Minimize: $MSR = \frac{\sum(z_y - \hat{z}_y)^2}{n-1}$

Since $\hat{z}_y = mz_x$ (slope): $MSR = \frac{\sum(z_y - mz_x)^2}{n-1}$

Square the binomial: $= \frac{\sum(z_y^2 - 2mz_xz_y + m^2z_x^2)}{n-1}$ } Σ of diff diff of Σ

Rewrite the summation: $= \frac{\sum z_y^2}{n-1} - 2m\frac{\sum z_x z_y}{n-1} + m^2\frac{\sum z_x^2}{n-1}$

4. Substitute from (2) and (3): $= 1 - 2mr + m^2$

Wow! That simplified nicely! And as a bonus, the last expression is a quadratic. Remember parabolas from algebra class? A parabola in the form $y = ax^2 + bx + c$ reaches its minimum at its turning point, which occurs when $x = \frac{-b}{2a}$. We can minimize the mean of squared residuals by choosing $m = \frac{-(-2r)}{2(1)} = r$

$x = \frac{-b \pm \sqrt{b^2-4a}}{2a}$

Amazing! The slope of the best fit line for z-scores is the correlation coefficient, r

That's great, but we still need to figure out a way to work with the original units so we can avoid converting back and forth to z-scores.

A slope of r for z-scores means that for every increase of 1 standard deviation in z_x there is an increase of r standard deviations in $\hat{z}_y$. "Over one, up r," as you probably said in algebra class. Translate that back to the original x and y values: "Over one standard deviation in x, up r standard deviations in $\hat{y}$."

That's it! The slope of the regression line is $b_1 = \frac{rs_y}{s_x}$!!!