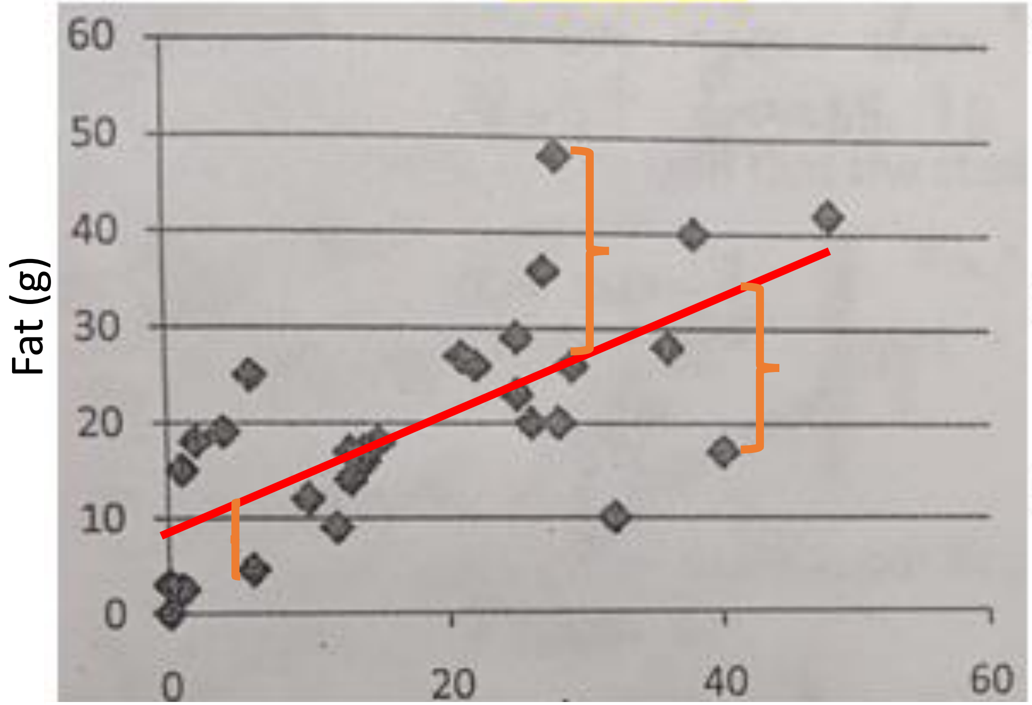
You are given the nutrition facts for 29 items on McDonald's menu. McDonalds has added a new sandwich to its menu, the McMouthful. How much fat should we predict if it has 43 grams of protein?

Steps: 1) make a __scatterplot__ to assess potential association between ___protein__ and ___fat__




2) Describe the association

Form = fairly linear, Direction = Positive,
Strength = moderately strong, Unusual Features = no

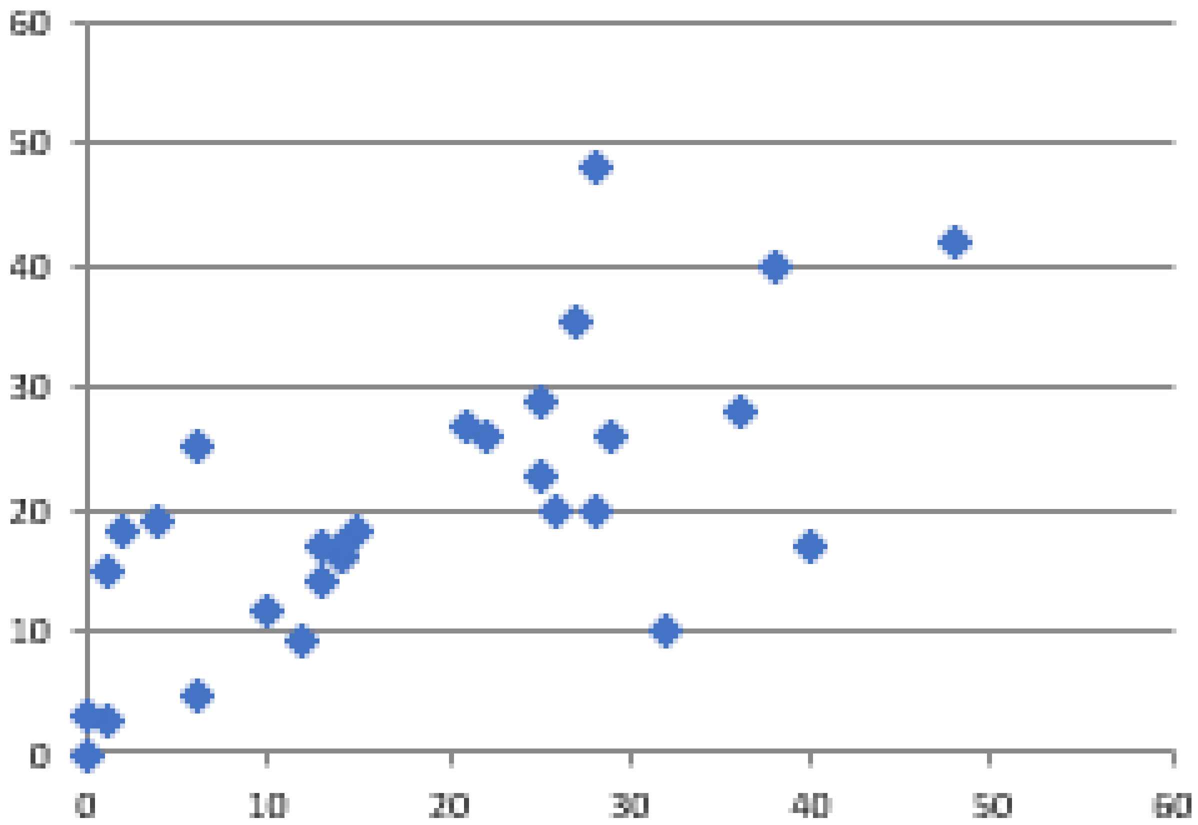
3) Obtain summary statistics.

$\bar{x}_{protein} = 18.\bar{3}$ g  $S_x = 13.2$ g
$\bar{x}_{fat} = 19.9$ g  $S_y = 11.6$ g

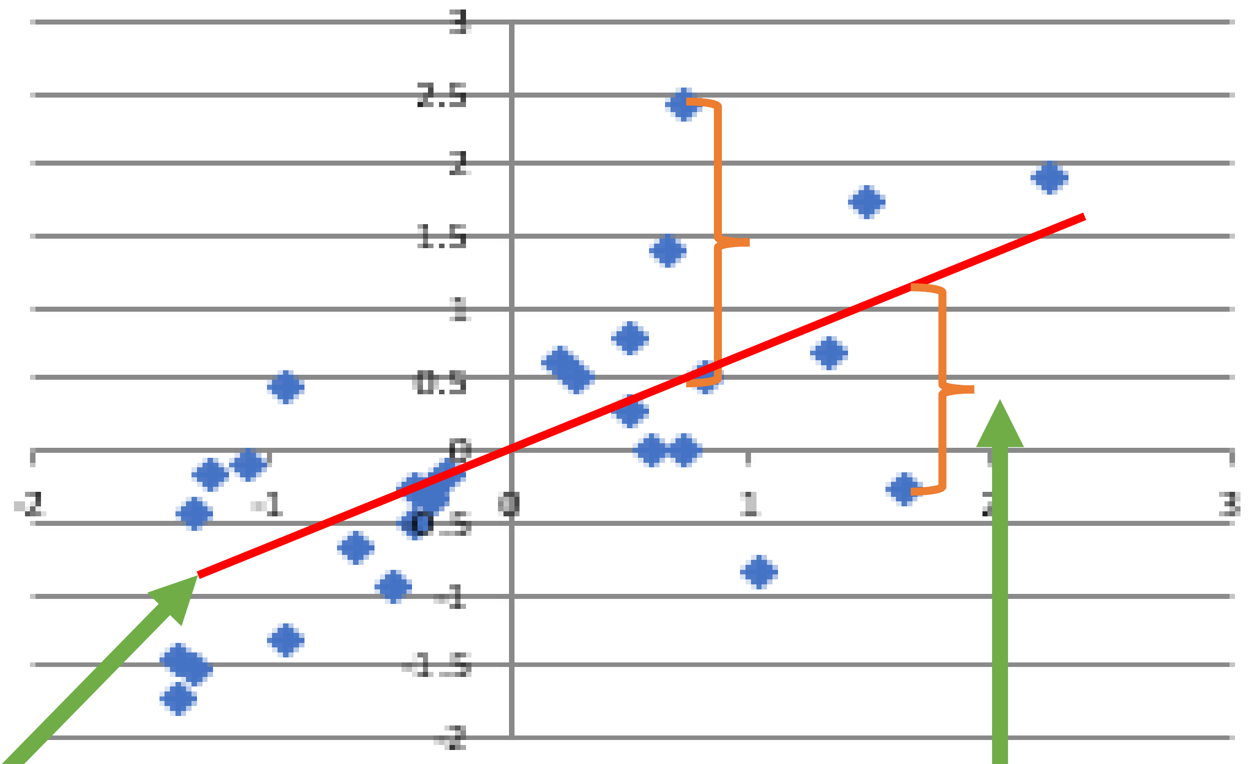
4) Check __conditions__ for correlation:

Quantitative Data
Straight Enough – scatterplot shows linear form
No outliers

| | |
|---|---|
| Residual | Observed value – predicted value<br>$y - \hat{y}$ |
| If positive<br>If negative | Then the model makes an underestimate.<br>Then the model makes an overestimate. |
| Regression line<br>Line of best fit | The unique line that minimizes the sum of the squared residuals (the variance of the residuals). |




# Original




# Standardized

We seek the line $\hat{z}_y = a + mz_x$ that minimizes $\sum\left[z_y - \hat{z}_y\right]^2$

Just like y = mx + b from algebra

The z has a hat because it is a prediction.

We seek the line $\hat{z}_y = a + mz_x$ that minimizes $\sum\left[z_y - \hat{z}_y\right]^2$

Substitute the equation: $\sum\left[z_y - (a + mz_x)\right]^2$

Rearrange terms: $\sum\left[(z_y - mz_x) - a\right]^2$

Square the binomial: $\sum\left[(z_y - mz_x)^2 - 2a(z_y - mz_x) + a^2\right]$

Consider the "middle term": $\sum\left[2a(z_y - mz_x)\right]$

Rewrite the sum: $2a\sum z_y - 2am\sum z_x$

But the mean (and thus the sum) of a set of *z*-scores must be zero. Hence this whole middle term is zero and we can turn our attention to the task of trying to minimize what's left: $\sum\left[(z_y - mz_x)^2 + a^2\right]$

By choosing $a = 0$ we can be sure that the sum will be minimal. (Adding the square of any other value would make it bigger.) Hence the best line must have a *y*-intercept of 0 in the standardized plane, proving that the line of regression goes through the mean-mean point. !

Now we just have to find the slope that minimizes $\sum\left[(z_y - mz_x)^2\right]$, ~~and~~

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.

Givens

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 $\hat{z}_y = mz_x$. We want to find the value for $m$ that will minimize the sum of the 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 $\hat{z}_y = \overset{\text{slope}}{m} z_x$: $$MSR = \frac{\sum(z_y - mz_x)^2}{n-1}$$

↳ from D20 D21

$(y - mx)(y - mx)$

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

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

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

Σ of differences = differences of Σ factoring & Commutative Prop

Wow! That simplified nicely! And as a bonus, the last expression is 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 $m = \frac{-(-2r)}{2(1)} = r$.

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

**Amazing! The slope of the best fit line for z-scores is the correlation, 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 $b_1 = \frac{rs_y}{s_x}$. !

A note of caution in statistics we write the equation a bit differently: $\hat{y} = b_0 + b_1x$

| | |
|---|---|
| Residual<br><br>If positive<br>If negative | Observed value – predicted value<br>$y - \hat{y}$<br>Then the model makes an underestimate.<br>Then the model makes an overestimate. |
| Regression line<br>Line of best fit<br>For standardized values<br>For actual $x$ and $y$ values | The unique line that minimizes the sum of the squared residuals (the variance of the residuals).<br>$\hat{z}_y = rz_x$<br>$\hat{y} = b_0 + b_1x$ |
| To calculate the regression line in real units (actual $x$ and $y$ values) | 1. Find slope, $b_1 = \frac{rs_y}{s_x}$<br>2. Find y-intercept, plug $b_1$ and point (x, y) [usually $(\bar{x}, \bar{y})$] into $\hat{y} = b_0 + b_1x$ and solve for $b_0$<br>3. Plug in slope, $b_1$, and y-intercept, $b_0$, into $\hat{y} = b_0 + b_1x$ |
| 3 conditions needed for Linear Regression Models:<br>/* same as correlation */ | 1. Quantitative Variables<br>2. Straight Enough – check original scatterplot & residual scatterplot<br>3. Outlier (clusters) –points with large residuals and/or high leverage |

Let's model the association between protein and fat of McDonald's menu items by writing an equation for the line of best fit:

$\hat{y} = b_0 + b_1 x$

1. $b_1 = ?$

$= \frac{r S_y}{S_x}$

$= \frac{(0.69)(11.6g)}{(13.2g)}$

$= 0.61$

2. $b_0 = ?$

$\bar{y} = b_0 + b_1 \bar{x}$

$19.9 = b_0 + (0.61)(18.3)$

$8.7 = b_0$

3. $\hat{y} = 8.7 + 0.61x$

$\widehat{fat} = 8.7 + 0.61\ protein$

Just as the mean summarizes a variable and the standard deviation tells how well. The regression line (line of best fit) summarizes the response variable in term of the explanatory variable and $R^2$ tells how well.

- We know choosing $m = r$ minimizes the sum of the squared residuals, but how small does that sum get? Equation (4) told us that the mean of the squared residuals is $1 - 2mr + m^2$. When $m = r$, $1 - 2mr + m^2 = 1 - 2r^2 + r^2$ $= 1 - r^2$. This is the percentage of variability *not* explained by the regression line. Since $1 - r^2$ of the variability is *not* explained, the percentage of variability in $y$ that is explained by $x$ is $r^2$. This important fact will help us assess the strength of our models.

And there's still another bonus. Because $r^2$ is the percent of variability explained by our model, $r^2$ is at most 100%. If $r^2 \leq 1$, then $-1 \leq r \leq 1$, proving that correlations are always between $-1$ and $+1$.

| | |
|---|---|
| $R^2$ | The square of the correlation, *r*, between *x* and *y*<br>The success of the regression model in terms of the fraction of the variation of *y* accounted for by the model.<br>* Differences in *x* explain XX% of the variability in *y*<br>or The model explains XX% of the variability in *y* |