

Lesson 14: Explain the quote (by George Box, a famous statistician), “All models are wrong, but some are useful.”

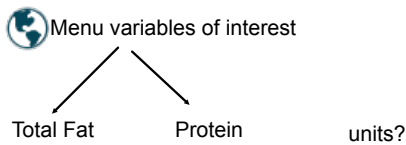
Q: Enter the BK menu data into your calculator.

 BURGERS	 HOT DOGS	 CHICKEN & MORE
 SALADS & VEGGIES	 BREAKFAST	 BEVERAGES
 COFFEE & FRAPPES	 SIDES	 SWEETS
 VALUE MENU	 KING JR. MEALS	

Examining the relationship between two quantitative variables.

1. Are both variables quantitative?

Quantitative variables generally come with units or units are implied.



Labels for Variables:

Response Variable: assigned to y-axis. This is the variable that you hope to predict or explain.

Explanatory Variable: assigned to x-axis. This variable accounts for, explains, predicts, or is otherwise responsible for the y-variable.

Why not use the names **independent** and **dependent** variable?

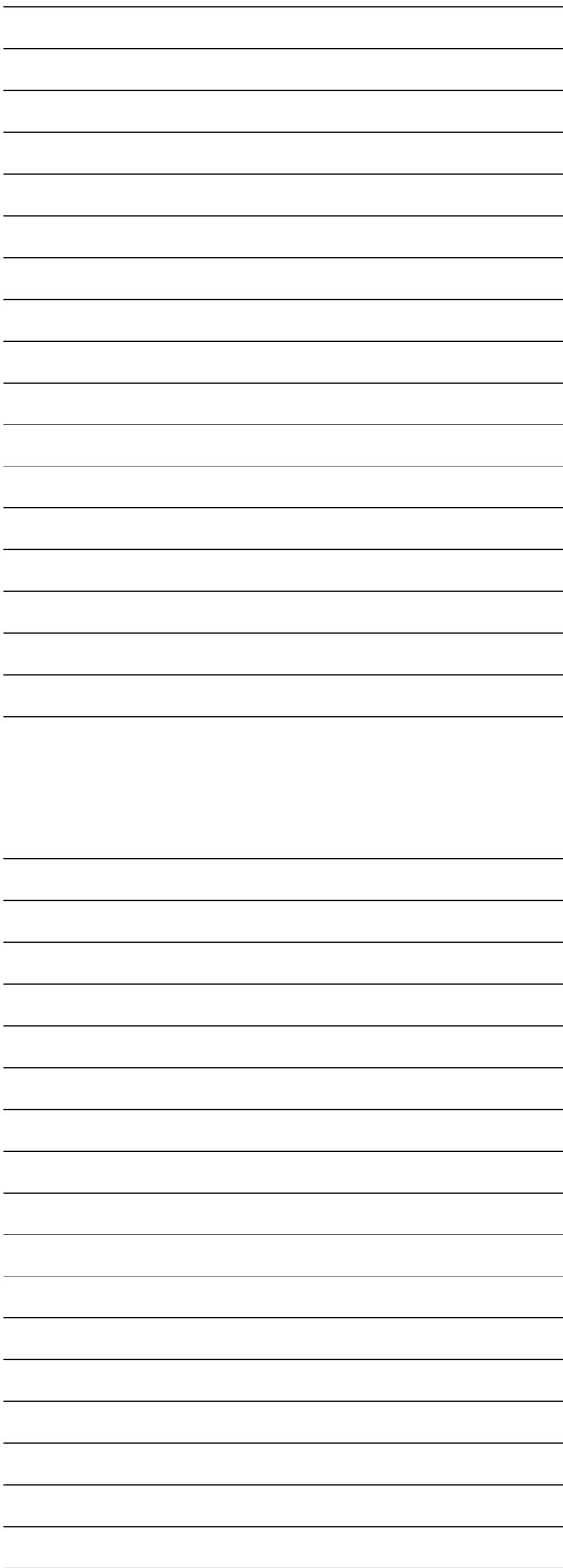
2. Make the scatterplot. Analyze it!

Scatterplot: A graph in which the values of two variables are plotted along two axes, the pattern of the resulting points revealing any correlation present. Outliers become easier to spot as well.

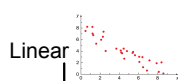
[illegible]

For a scatterplot, we think about it's direction, form, strength, and unusual features.

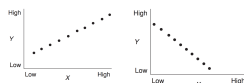
For a scatterplot, we think about it's direction, form, strength, and unusual features.



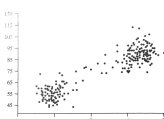
Form:
patterns you
see in the scatterplots



Perfect linear



Clusters



The form is roughly linear with two clusters.

Curvilinear

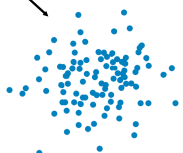


Strength

How much scatter?

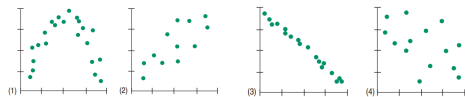


Tightly clustered
(strong \pm , weak \pm ,
moderately strong/weak)



We can barely discern any
trend or pattern

Check of understanding



Scatterplots. Which of the scatterplots at the top of the next column show

- a) little or no association?
- b) a negative association?
- c) a linear association?
- d) a moderately strong association?
- e) a very strong association?


Unusual features:

outliers - points that do not fit the overall pattern

Or other departures from a pattern

Once you have determined that the scatterplot shows a linear trend → Perform Linear Regression Analysis

Correlation (r) is a number (-1 to 1) that measures the direction and strength of a linear association between two quantitative variables.



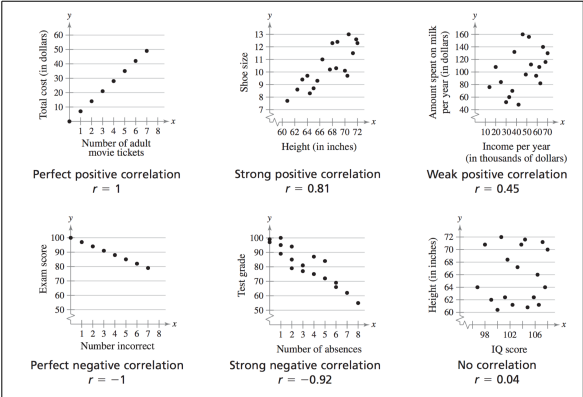
Did you know that there's a strong correlation between playing an instrument and drinking coffee?

Association vs Correlation

“Association” is a deliberately vague term describing the relationship between two variables.

As a rule, we will interpret r values as follows:

- .0 to .2 No relationship to very weak association
- .2 to .4 Weak association
- .4 to .6 Moderate association
- .6 to .8 Strong association
- .8 to 1.0 Very strong to perfect association



Correlation guessing game.



smann/chance applet collection

Correlation Coefficient (r) is unit less.

$$r = \frac{\sum (Z_X \cdot Z_Y)}{n - 1}$$

X and Y data values
converted to z-score
pairs

** r is **not affected** by a change in units of the
explanatory or response variables.

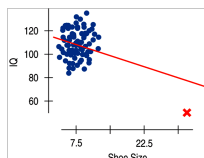
** r is **affected** by outliers



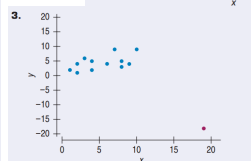
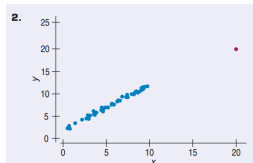
Notice that the best fit line
always goes through
the means of the x and y

Point has x-value that is far from the mean of x-values. This point has potential to change the regression line.

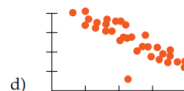
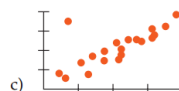
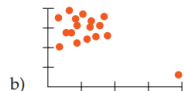
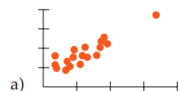
A scatter plot showing IQ (Y-axis) versus Shoe Size (X-axis). The Y-axis has labels at 100, 125, 150, and 175. The X-axis has labels at 7.5 and 22.5. A cluster of blue data points is centered around a shoe size of 7.5 and an IQ of 120. A red regression line is drawn through the data, showing a positive correlation. A single red 'X' marks an outlier at approximately (30, 180).



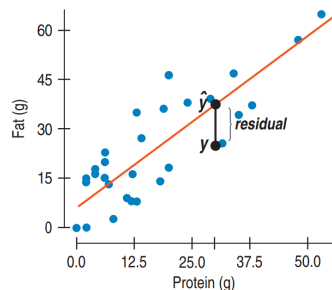
Each of these scatterplots shows an unusual point. For each, tell whether the point is a high-leverage point, would have a large residual, or is influential.



- 1) In what way is the point unusual? Does it have high leverage, a large residual, or both?
- 2) Do you think that point is an influential point?
- 3) If that point were removed, would the correlation become stronger or weaker? Explain.
- 4) If that point were removed, would the slope of the regression line increase or decrease? Explain.

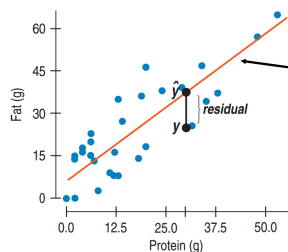
[illegible]

3. Fit a regression model and find the residuals, e , and predicted values \hat{y} .



Does the line match reality?

$\text{residual} = \text{observed value} - \text{predicted value}$
 A *negative* residual means the predicted value is too big—an overestimate. And a *positive* residual shows that the model makes an underestimate. These may seem backwards until you think about them.



best fit line may not hit any point

Why is a best fit then??

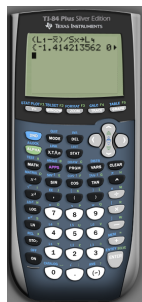
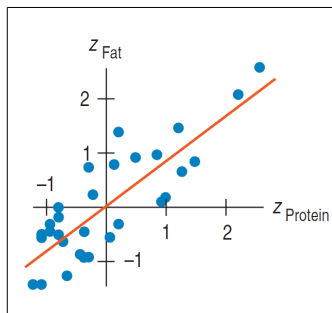
How can we assess how well the line fits the data?

Does adding up residuals make sense?

The line of best fit is the line for which the sum of the squared residuals is smallest, the least squares line.

squared because otherwise the sum would just be zero...positive and negative residuals would cancel

Let's create the BK scatterplot in z-scores



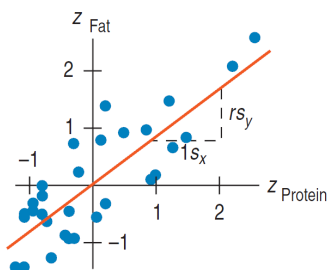
The line must go through (\bar{x}, \bar{y}) so in z-scores(0,0)
The equation of the line passing through the origin is $y=mx$

In terms of z-scores, since our coordinates are (z_x, z_y)

$$\hat{z}_y = mz_x$$

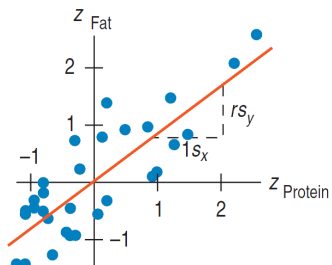
what's the slope??

Each one standard deviation change in protein results in a predicted change of r standard deviations in fat.



How big can the predicted values get?

Notice that if x is 2 SDs above its mean, we won't ever guess more than 2 SDs away for y , since r can't be bigger than 1.0.



Regression line in real units

$$\hat{y} = b_0 + b_1x$$

Protein	Fat
$\bar{x} = 17.2 \text{ g}$	$\bar{y} = 23.5 \text{ g}$
$s_x = 14.0 \text{ g}$	$s_y = 16.4 \text{ g}$
$r = 0.83$	

Slope

$$b_1 = \frac{rs_y}{s_x}$$

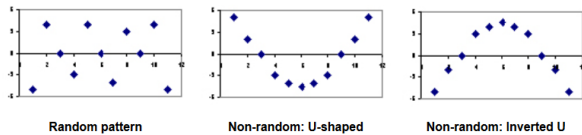
Intercept

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

regression line
goes through
the mean

4. Make a scatterplot of the residuals against x or the predicted values. This plot should have no pattern. Check for any bend in the residuals. This would suggest that the data were not straight enough after all and our visual check was wrong. Also check for any thickening or thinning of data as well as outliers. (If there are outliers, and you can correct them or justify removing them, do so and go back to step 1, or consider performing two regressions - one with and one without the outliers.

We want a boring residual plot



Residuals Revisited

The linear model assumes that the relationship between the two variables is a perfect straight line. The residuals are the part of the data that *hasn't* been modeled.

$$\text{Data} = \text{Model} + \text{Residual}$$

or (equivalently)

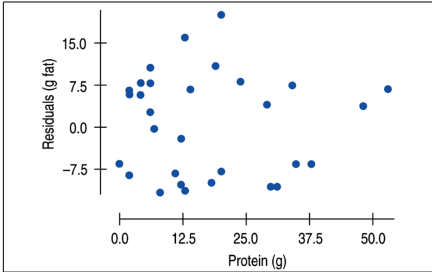
$$\text{Residual} = \text{Data} - \text{Model}$$

Or, in symbols,

$$e = y - \hat{y}$$

- Residuals help us to see whether the model makes sense.
- When a regression model is appropriate, nothing interesting should be left behind.
- After we fit a regression model, we usually plot the residuals in the hope of finding...nothing.

The residuals for the BK menu regression



The Residual Standard Deviation

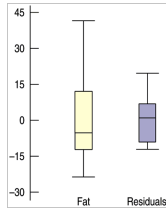
- The standard deviation of the residuals, s_e , measures how much the points spread around the regression line.
- Check to make sure the residual plot has about the same amount of scatter throughout. Check the **Equal Variance Assumption** with the **Does the Plot Thicken? Condition**.
- We estimate the SD of the residuals using:

$$s_e = \sqrt{\frac{\sum e^2}{n-2}}$$

R^2 —The Variation Accounted For

- The variation in the residuals is the key to assessing how well the model fits.

In the BK menu items example, total *fat* has a standard deviation of 16.4 grams. The standard deviation of the residuals is 9.2 grams.

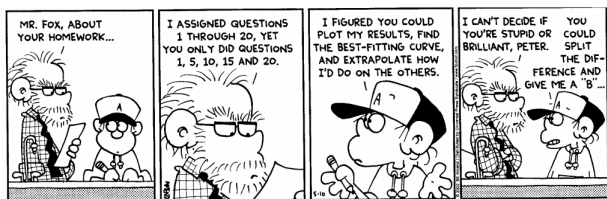


- If the correlation were 1.0 and the model predicted the *fat* values perfectly, the residuals would all be zero and have no variation.
- As it is, the correlation is 0.83—not perfection.
- However, we did see that the model residuals had less variation than total *fat* alone.
- We can determine how much of the variation is accounted for by the model and how much is left in the residuals.

R^2 —The Variation Accounted For

- The squared correlation, r^2 , gives the fraction of the data's variance accounted for by the model.
 - Thus, $1 - r^2$ is the fraction of the original variance left in the residuals.
 - For the BK model, $r^2 = 0.83^2 = 0.69$, so 31% of the variability in total *fat* has been left in the residuals.
- When interpreting a regression model you need to *Tell* what R^2 means.
 - In the BK example, 69% of the variation in total *fat* is accounted for by variation in the protein content.

Warnings!

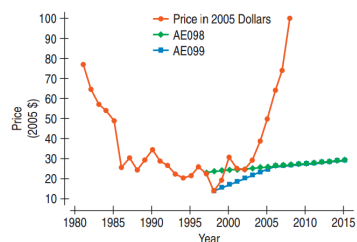


FOXTROT © 2002 Bill Amend. Reprinted with permission of UNIVERSAL PRESS SYNDICATE. All rights reserved.

Predicting values
beyond the data

Warning about extrapolation:

It requires an assumption that nothing about the relationship between the variables changes as we go to extreme values of x .

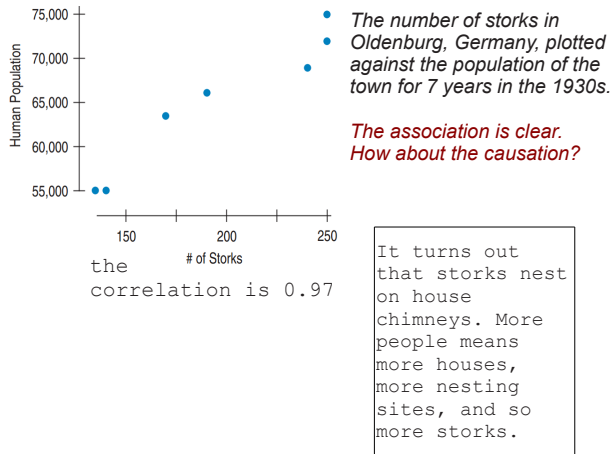


If you must extrapolate into the future, at least don't believe that the prediction will come true.

Here are the EIA forecasts with the actual prices from 1981 to 2008. Neither forecast predicted the sharp run-up in the past few years.



Storks bring babies?



Correlation \neq Causation

- Whenever we have a strong correlation, it is tempting to explain it by imagining that the predictor variable has **caused** the response to help.
- Scatterplots and correlation coefficients **never** prove causation.
- A hidden variable that stands behind a relationship and determines it by simultaneously affecting the other two variables is called a **lurking variable**.

Assign homework with problems where you sift residuals for groups!

[illegible]