Formula for linear models.
Prediction, extrapolation, significance test against zero slope.
Last time, we looked the linear regression formula.

\[ Y = a + bX + e \]
It’s the line that fits the data best. The Pearson correlation can be considered a measure of how well that line fits.
We can compute the slope $b$ from the correlation coefficient $r$, the standard deviations of the $x$ data and $y$ data.

\[ b = r \frac{S_y}{S_x} \]
The slope is rate that y changes for each unit x changes.

“Y increases by b units when X increases by one unit”

\[
b = r \frac{S_y}{S_x}
\]
The trend is stronger when the correlation is stronger, so the slope increases as the correlation $r$ increases.

\[
b = r \frac{S_y}{S_x}
\]
If $y$ changes a lot relative to $x$, it has to have wider range and more variance than $x$, so $S_y$ will be bigger than $S_x$. 

The slope is given by:

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

- **Slope**
- **Correlation Coefficient**
- **Standard Deviation of $y$**
- **Std. Dev. of $x$**
The simplest way to find the intercept $a$ is to use the slope $b$, and the sample means of $x$ and $y$.

$$a = \bar{y} - b\bar{x}$$
\[ a = \bar{y} - bx \]

- intercept
- slope

Mean of y
Mean of x
Example: We have data with the following properties. Find the slope and intercept.

\[
\bar{x} = 12 \quad \bar{y} = 7 \\
\text{s}_x = 4 \quad \text{s}_y = 22 \\
r = -0.35
\]

Find the regression line from this data.
We start with the slope $b$

$$b = r \frac{S_y}{S_X} = -0.35 \frac{22}{4} = -1.925$$

(Apologies to the alphabet, $b$ comes before $a$ in regression)
Next we get the intercept $a$, using the slope $b$ that we just found.

\[
a = \bar{y} - b\bar{x} \\
= 7 - (-1.925)(12)
\]
a = \bar{y} - b\bar{x}

= 7 - (-1.925)(12)

= 7 + 22.10

= 29.10
Now that we have the slope and intercept, our regression equation is:

\[ \hat{y} = 29.10 - 1.925x \]
If each dot were a note, it would sound like a dragon walking across a piano.
Regression in SPSS.

To find the slope and intercept, also called the __________ we first go to Analyze ➔ ______ ➔ ______

![SPSS Analyze menu](image)
Put the dependent (Y) variable in ____________, and Put the independent (X) variable in ____________.

Later, we’ll put multiple variables in Independent, but not yet. Then click OK. (not shown, at bottom of pop-up)
The results of interest are the unstandardized coefficients. 

- **Constant** is the intercept, \( a = 4.701 \)
- **TV / Day** is the slope, \( b = -0.841 \)
We can also find the p-value of the hypothesis that the slope is zero. (Small sig. means there is significant evidence of a slope)

When there’s only one independent variable, this is the same as the p-value of the correlation.
We might also be interested in drawing the regression line as well as getting its formula. There’s an option to do so with a scatterplot, so build a scatterplot first.

(Graphs ➔ Legacy Dialogs ➔ Scatter/Dot,
Choose Simple Scatter and OK,
Put the Dependent in Y-axis, Independent in X-axis)
Once you have a scatterplot, go to the output window and _______ on it. Choose **Edit Content → In Separate Window**
If a pop-up like the properties window comes up, just ignore it and click "Close".
In the Chart Editor window that comes up, go to _______ ➔
_________________________. Then, click _______ with the window that pops up after.

Clicking on the icon that looks like Fit Line at Total does this too.
Now you have a regression line for your data.

Serves 6-8, best when chilled.
You and SPSS: Making beautiful music together.

Now with watermarks!
With a linear regression, we can use a value of X to predict a value of Y.

Our best prediction would be that any new data is right on the line.

Let’s say, after doing the books vs. TV analysis, we wanted to predict how many books/year were read by a child who watched $x = 2.5$ hours of TV/day.
The regression line was:

\[ \hat{Y} = 4.701 - 0.841 X \]
So, to predict newcomer’s books/year, we replace the TV/year variable $x$ with 2.5.  

$$\hat{Y} = 4.701 - (0.841) \quad = \quad $$

By our estimate, a child watches 2.5 hours of TV/day also reads _______ books/year.
We could choose different $x$ values and get different $y$ predictions.

$\hat{y} = 4.701 - (0.841) (1.5) = 3.440$

$\hat{y} = 4.701 - (0.841) (5) = 0.496$
But we can’t choose any value:

Some predictions don’t make any sense, we won’t know that until after we draw the regression line.

7 hours/day of TV $\rightarrow$ -1.186 books/year

Some predictions won’t make sense no matter what the model is.

-2 hours/day of TV $\rightarrow$ 6.383 books/year

27 hours/day of TV $\rightarrow$ -18.006 books/year
How do we know what’s allowed for prediction?
__________ values, those are predicted from between observed data points, are usually good.
values, those predicted beyond the data, are bad.

A regression line tells us the trend among the data that we see, it tells us nothing of the data beyond.
If you extrapolate, you can get some pretty absurd results.

Source: xkcd.com/605
Sometimes the problem with extrapolating can be more subtle.

_________ is the act of extrapolating in time.

Without data, extrapolation is meaningless.
The intercept is the Y value when X is zero.

When \( X = 0 \) is within the data (or could reasonably be), the intercept means something in the real world.

\[ \hat{Y} = 45.00 - 1.60X \]

Slope:

Intercept:
The intercept is the Y value when X is zero.
When X = 0 is within the data (or could reasonably be), the intercept means something in the real world.

X: Average monthly temperature (Celsius)
Y: Monthly heating bill (CDN $)
\hat{Y} = 45.00 - 1.60X

Slope: For every \__________, the heating bill \__________

Intercept: The average heating bill \__________
When $X = 0$ is beyond the data, the intercept is useful only as a mathematical construct.

$X$: Resting Heart Rate (beats per minute)

$Y$: Body Mass Index (kg / m²)

\[ \hat{Y} = 16.7 + 0.17X \]

Slope:

Intercept:

X: Resting Heart Rate (beats per minute)

Y: Body Mass Index (kg / m²)

\[ \hat{Y} = 16.7 + 0.17X \]

Intercept: The average person with a resting heart rate of zero has a BMI of 16.7.

Slope: For every additional beat-per-minute, BMI increases by 0.17 kg / m².
BMI = 16.7 + 0.17(Heart Rate)

Intercept: The average person with a resting heart rate of zero has a BMI of 16.7.

Is this reasonable? Doesn’t this imply that dead people, on average have a BMI of 16.7?

Was anyone with a heart rate near zero even tested? Go to the source:

http://www.pps.org.pk/PJP/6-1/Talay.pdf
Final note: As long as the relationship is linear (or close), slope always has a real-world interpretation.

The slope interpretation implies that it’s for the interval of observed data and not anything beyond that.

$X = 0$ isn’t always within the observed interval, so the intercept doesn’t always have a real-world interpretation.
Chapter 11 to be continued Mon, July 16.

Next time: Midterm review (send suggestions/requests).