7.0 Lesson Plan

- Regression
- Residuals
Recall the regression assumptions:

1. Each point \((X_i, Y_i)\) in the scatterplot satisfies:

   \[ Y_i = aX_i + b + \epsilon_i \]

   where the \(\epsilon_i\) have a normal distribution with mean zero and (usually) unknown standard deviation.

2. The errors \(\epsilon_i\) have nothing to do with one another. A large error does not tend to be followed by another large error, for example.

3. The \(X_i\) values are measured without error. (Thus all the error occurs in the vertical direction, and we do not need to minimize perpendicular distance to the line.)
A nutritionist studied how the amount of time that a toddler spends at the table predicts the amount of calories he or she consumes at lunch.

Data were collected on 20 toddlers at a nursery school. “Time” is the number of minutes that each toddler sat when lunch was served. “Calories” is the number of calories that child consumed.

Which is the dependent variable, and which is the independent variable?

Which is the explanatory variable, and which is the response variable?
What would you guess about the shape of the scatterplot and the correlation.
The regression line explains 42.1% (or $(-.65)^2 \times 100\%$) of the variation in calories. Given all the factors that affect a toddler’s eating, this is an impressive percentage. Time is a major factor in the calorie consumption.

From the JMP output, $\hat{b} = 560.65$ and $\hat{a} = -3.0771$. Both are highly significant, since “Prob $>|t|” is less than .01 in each case.

The estimated standard deviation of $\epsilon$ is 23.398. This is the typical vertical distance between a point and the line.
The root mean square (RMSE) is the standard deviation of the vertical distances between each point and the estimated line. It is an estimate of the standard deviation of the vertical distances between the observations and the true line.

Formally,

\[
RMSE = \sqrt{\frac{1}{n} \left[ (Y_1 - (\hat{a}X_1 + \hat{b})^2 + \cdots + (Y_n - (\hat{a}X_n + \hat{b})^2 \right]}
\]

Since \( \hat{a}X_i + \hat{b} \) is the estimate of the average value of the \( Y \)-value at \( X_i \), then RMSE should look like a standard deviation.
The regression model assumes that, at each value of time, the average calorie amount falls on a line. If the model is correct and one plots the average of all children in the universe who sit for 10 minutes, 20 minutes, 30 minutes, etc., then the graph of those averages is a line.
The regression line predicts the average value for the $Y$ values at a given $X$.

In practice, one wants to predict the individual value for a particular value of $X$. For example, if I make Elvis sit at the table for 25 minutes, how much will he eat?

For example, the prediction for the number of calories consumed on average by children who sit for $X = 32$ minutes is:

\[
\hat{Y} = \hat{b} + \hat{a}X
= 560.65 - 3.0771 \times 32
= 462.18
\]
The individual value is less exact than the average value. To predict the average value, the only source of uncertainty is the exact location of the regression line (i.e., \( \hat{a} \) and \( \hat{b} \) are estimates of the true intercept and slope).

In order to predict Elvis’s value, the uncertainty about Elvis’s deviation from the average is added to the uncertainty about the location of the line.

That is, uncertainty in \( \hat{Y} \) at 32 mins plus uncertainty due to Elvis’s variability about \( \hat{Y} \)

Again \( Y = aX + b + \epsilon \). Variability associated with estimating \( aX + b \) plus the uncertainty associated with \( \epsilon \)
Predicting $Y$ values for $X$ values outside the range of $X$ values observed in the data is **extrapolation**.

This is risky, because you have no evidence that the linear relationship you have seen in the scatterplot continues to hold in the new $X$ region. Extrapolated values can be entirely wrong.

What would you think Elvis might eat if he sat for 0 minutes at the table. Or for 10 years?
7.3 Residuals

Estimate the regression line (using JMP software or by calculating \( \hat{a} \) and \( \hat{b} \) by hand).

Then find the difference between each observed \( Y_i \) and the predicted value \( \hat{Y}_i \) using the fitted line. These differences are called the residuals.

Plot each difference against the corresponding \( X_i \) value. This plot is called a residual plot.
If the assumptions for linear regression hold, what should one see in the residual plot?

If the pattern of the residuals around the horizontal line at zero is:

- curved, then the assumption of linearity is violated.
- fan-shaped, then the assumption of constant standard deviation is violated (heteroscedasticity).
- filled with many outliers, then again the assumption of constant standard deviation is violated.
- shows a pattern (e.g., positive, negative, positive, negative, ...) then the assumption of independent errors is violated.
We also typically make a normal plot of the residuals since they are supposed to be independent with the same normal distribution.

When the residuals have a histogram that looks normal and when the residual plot shows no pattern, then we can feel comfortable using the normal distribution to make inferences about individuals.

Estimate what percentage of toddlers who sit for 32 minutes eat less than 425 calories?

Under the regression assumptions, the toddlers who sit for 32 minutes have calorie consumption that is normally distributed with mean 462.18 and standard deviation 23.398.

So the $z$-transform is $(425-462.18)/23.398 = 1.589$. From the table, the area under the curve to the left of 1.589 is about $(100 - 89.04)/2 = 5.48\%$. 

The correlation between father’s heights and son’s heights is, say, $r = .8$

John’s height is 1.5 sd’s above the mean for fathers

John’s son’s height is estimated to be $(.8)(1.5) = 1.2$ sd’s above the mean for sons

change $r$ to $r = .6$, change 1.2 to $(.6)(1.5) = .9$

If John’s son’s height is 1.5 sd’s above the mean for sons, then, with $r = .8$ we estimate John’s height to be 1.2 sd’s above the mean for fathers

if $r < 0$, inverse relationship

Note that we don’t need to specify what the means and sd’s are