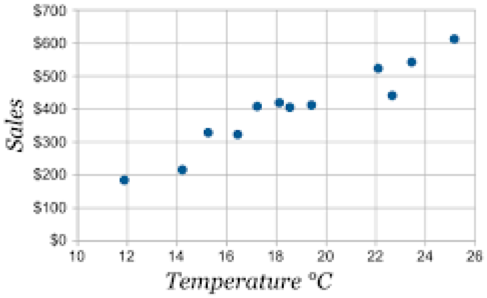
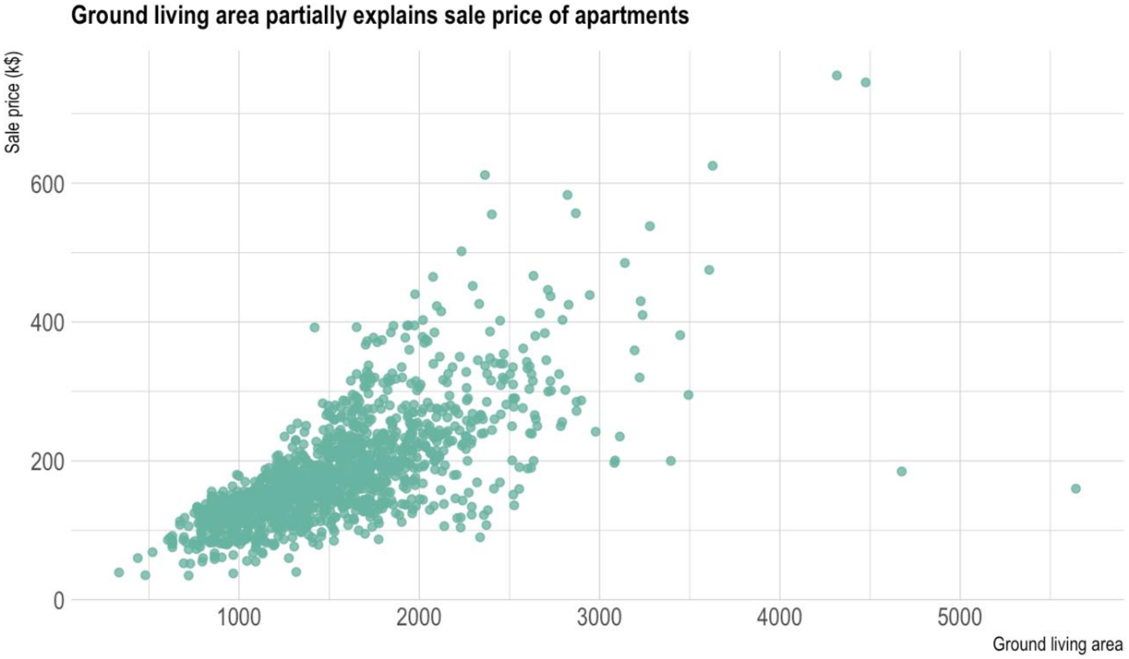
5/3/2022

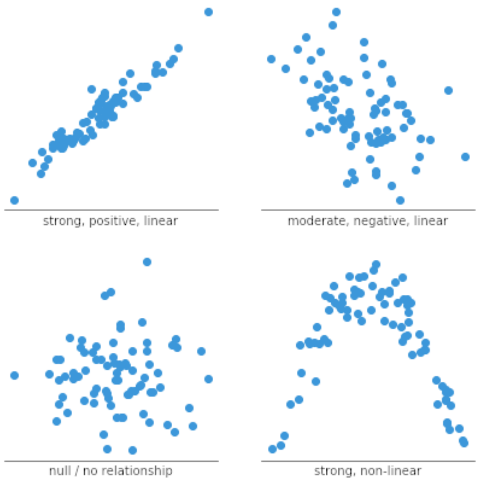
Correlation and Regression (Ch 12)

Data that comes in pairs, an x-coordinate and a y-coordinate. One is the input (x, independent variable, explanatory variable), and one is the output (y, dependent variable, response variable).

Create scatterplots, with x-variable on the horizontal axis and the y-variable on the vertical axis. Each pair is plotted as a point.

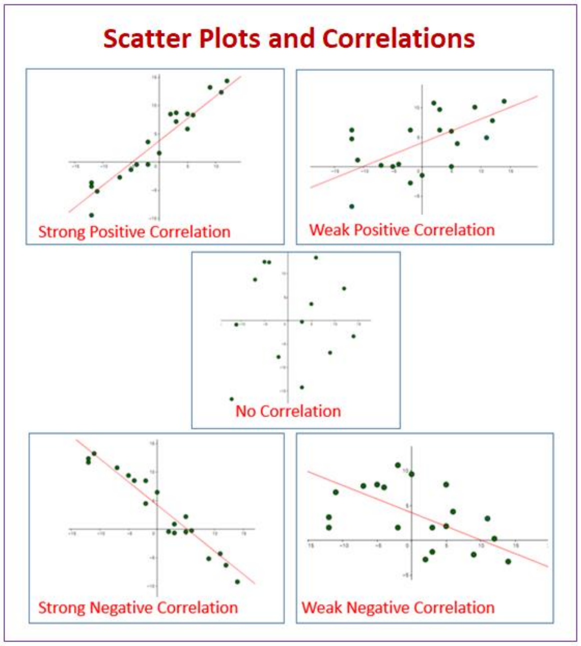






One characteristic of scatterplots is whether or not the data has a linear relationship.
Be able to look at a scatterplot to determine whether the data is roughly in a linear relationship or whether it is nonlinear.

Care about the strength of the relationship. If the data is close to the line, then the relationship is strong (errors are small). If the data is spread out from the line then the relationship is weaker (moderate, or weak). Correlation. Specifically linear. (correlation game online)



Correlation (linear), the values fall between -1 and 1. A 1 correlation is a perfect fit to a straight with a positive slope. A -1 correlation is a perfect fit to a straight line with a negative slope. 0 is no correlation.

Excel examples.

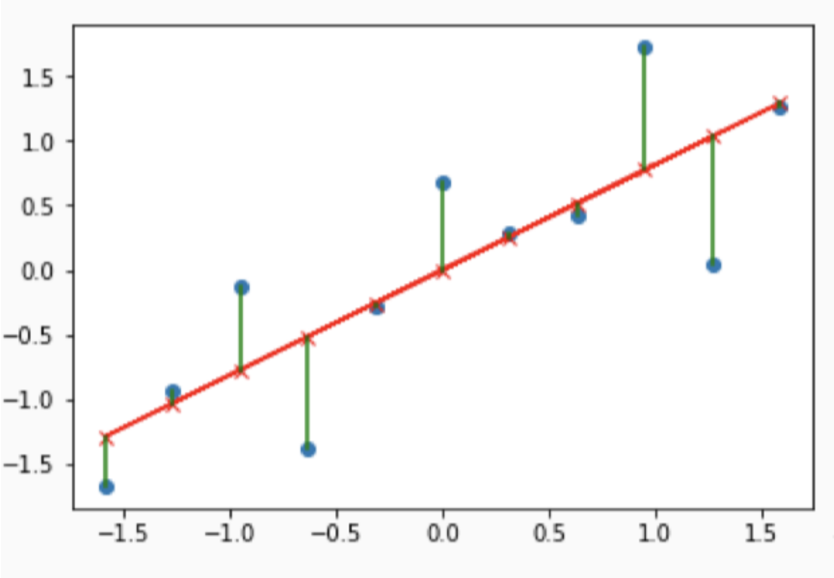
For Correlation to be considered strong, it generally 0.7 or 0.8 or higher
For correlation to be consider weak, it generally lower than 0.4

And in between is moderate.

(in absolute values).

Linear Regression Equation
Ordinary Least Squares
Line of Best Fit
Trendline

Finding a line (linear equation) that is as close to the data points as possible.
Trying to minimize the error between the observed values for each point (at very x), compared to the values that predicted from the line of best-fit.



The blue points are the original data. The red line is the line of best fit (regression equation), the green segments are the distances between the prediction line and the observations. We want to make those green distances as short as possible.
Green segments are called residuals. The error in a regression line is based on these residuals.

Regression equation from sheet 1

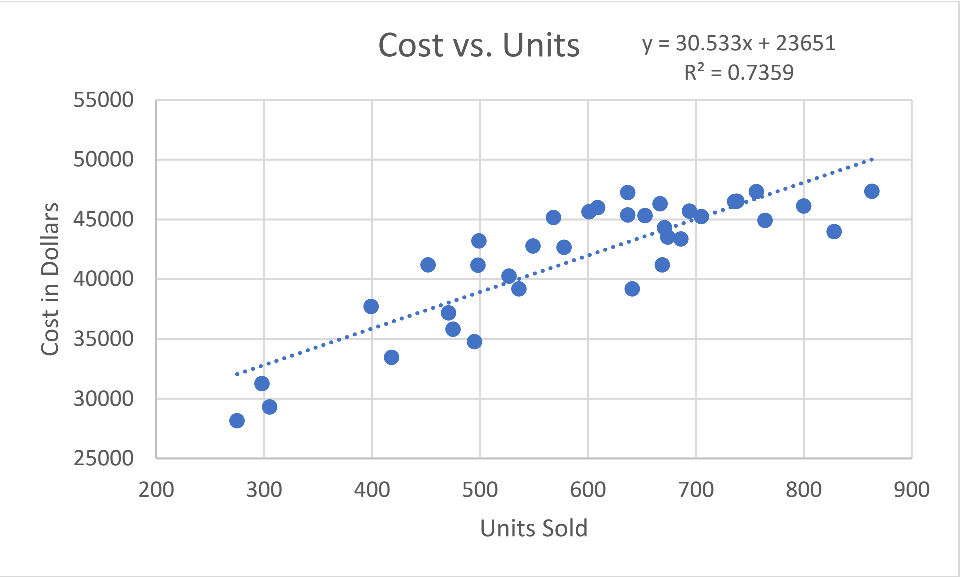
$$y = 30.533x + 23651$$

X was the number of units sold, and y was the cost to produce those units.
If I want to predict how much 500 units would cost to produce, I estimate it using the equation.

$$cost = 30.533(500) + 23651 = \$38{,}917.5$$

Estimate of the cost, it some margin of error (prediction error).
We can think of this estimate as an average: if I produced 500 units several times, what would be the mean cost to produce those units?

One of the assumptions that make to employ this regression equation for predictions is that the errors are normally distributed (a common standard deviation, a common variance). And that the regression equation is roughly valid within the range of the data we've collected. We shouldn't make predictions far outside the original range of the data.



Within the range of the data (or nearby), it is safe to assume that the trendline (roughly) continues to hold. But the further outside that range you go, the more suspect your prediction in because we don't know that the trend will continue.

R^2 value what it means and how it is related to correlation.
The variable name for correlation is r for sample data. The population correlation is ρ. Think of r as an estimate of ρ. (Just like s is an estimate of σ). Sometimes this is called the correlation coefficient. If we square r to get r^2, in the linear case this is equal to R^2.
In the linear, if we take the square root of R^2, we can get back to the correlation: with one fix: Correlation can be either positive or negative. The correlation r takes the same sign as the slope of the regression equation (not the same value, only the sign).

The R^2 value has a meaning all by itself. Can only be a value between 0 and 1. **Represents the percent of the variability in the original y-values that can be explained by their relationship to the model x-values**.
It is a measure of the improvement in our predictions based on establishing the regression equation.

One side imagine calculating the variance (or standard deviation) in the original y-value all by themselves and seeing how big that number is.
Then calculate the variance of the residuals (residuals are all smaller than the original errors). The variance of our predictions is reduced.
The R^2 value compares these two values. The bigger R^2 is, the more the residuals are reduced (they are smaller). If R^2 is small, then the regression equation eliminates very little error (the error is still large).
R^2 close to 1 is good: very helpful in making better predictions. R^2 close to 0 helps very little, the predictions are almost as bad as they were without any additional information.

Next time: hypothesis testing, prediction intervals and review for final.