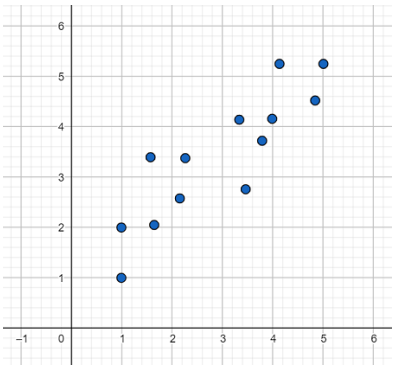
4/25/2023

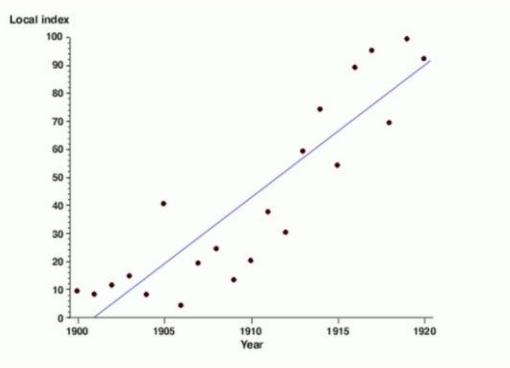
Linear Regression
Review for Final Exam

Recall that scatterplots are plots of two numerical variables that are plotted as pairs of points on a graph.

From our scatterplots there are some things we can assess:

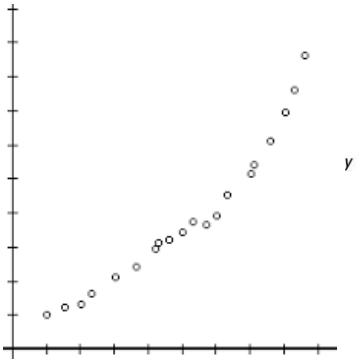
1) Is the relationship between the two variables linear or not linear?
2) How strong is the relationship between the two variables?

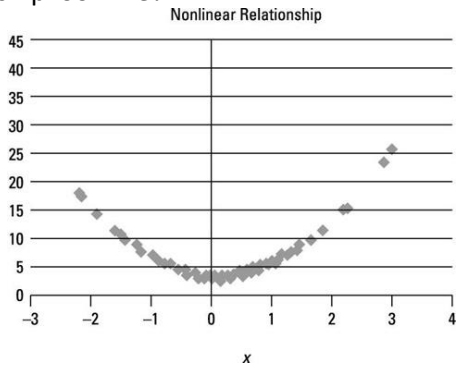


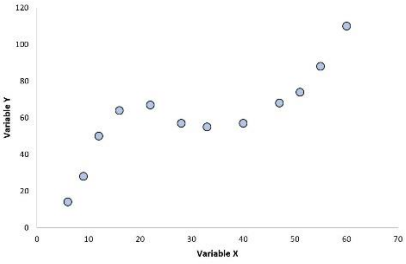


Both of these scatterplots show a linear relationship.

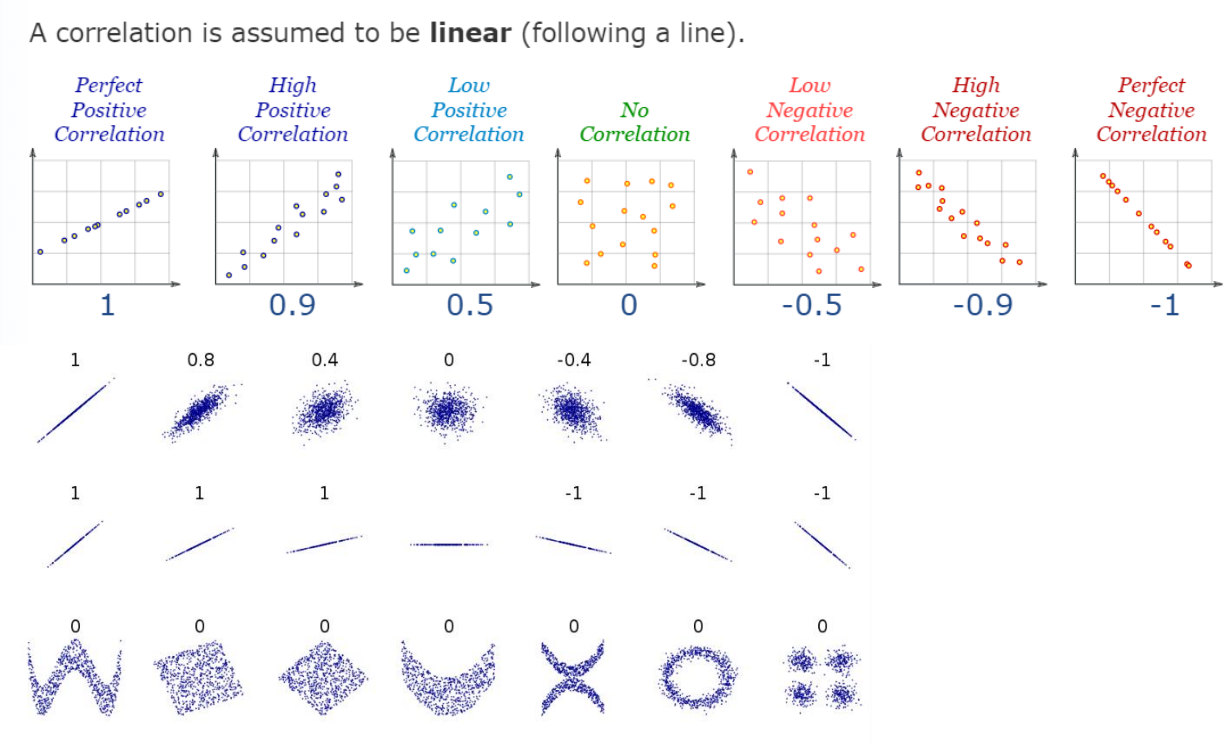
What does a nonlinear relationship look like?







Correlation
Is a value between -1 and 1, and it's always about the linear relationship between two variables. If the correlation is equal to 1, this is a perfect positive correlation. The data falls exactly on a straight line with a positive slope. If the correlation is -1, this a perfect negative correlation. The data falls exactly on a straight line with a negative slope. If the correlation is in between (other than 0) there is some relationship between the variables. Closer to 1 or -1 is stronger, and closer to 0 is weaker. If the correlation is 0, then there is no linear relationship between the variables: can happen if there is no relationship at all, or in some cases, certain non-linear relationships can give a 0 correlation. (If you try to draw a straight line through the middle of the data, the line is horizontal… slope the zero.)



Typically, correlations of $|r| > 0.7$, we call these strong.
Correlations of $0.4 < |r| < 0.7$ are considered moderate.
Correlations of $|r| < 0.4$ are considered weak.

R^2 – coefficient of determination

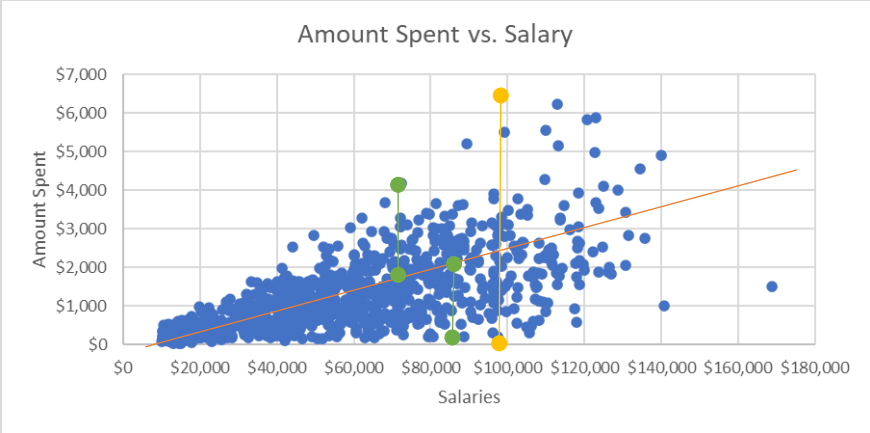
$$R^2 = r^2$$

Can be calculated in other ways, and can be used with nonlinear relationships.
R^2 is a value between 0 and 1. Closer to 1 is a better relationship. Closer to 0 is a weaker relationship.

$R^2 > 0.5$ is considered strong
$0.2 < R^2 < 0.5$ is considered moderate
$R^2 < 0.2$ is considered weak

R^2 is the percent of the variability in the y-value that can be explained by the relationship with the x-value.

The total range of y is from around 0 to 6500. The variability of y relative to the line representing the relationship with x is reduced from the y-variability with no relationship. The percentage is calculated from the variance.



Correlation for this data was approximately 0.7, so the R^2 is around 0.49 (0.489…).
The 49% of the variability in the amount spent is explained by the relationship between amount spent and salary.

Linear Regression is sometimes called ordinary least squares, or lines of best fit, or trendlines.

Takes the form of $y = mx + b$, and it models the straight line that best fit the data (that goes through the middle of the data with roughly half the data on one and half on the other).

The equation we obtained that predicts Amount Spent (y) from Salary (x) is

$$y = 0.022x - 15.332$$

How do we interpret the slope of this equation?

The amount spent increases by 0.022 (on average) for each additional dollar of salary.

The amount spent increase by $22 (on average) for each additional $1000 dollars of salary

How do we interpret the intercept?
We can't in this case. It's negative, and if we make 0 dollars, no one is giving us $15.33.

Final Exam information

Same format as the previous exams.
The material that we covered since the 2nd exam (linear equations, regression, scatterplots, correlation)
The rest of the material will be taken from the previous two exams.

The last quiz is the best preview of regression.
The previous two exams are the best model for review for the rest.