BUS 232 chapter 2 - data and relationships

Summary

Now there are **two variables** instead of one in chapter 1. **Normal quantile plots** show whether the data is **skewed** or **normal**. A **scatterplot** shows the relationship between **two** quantitative variables. A scatterplot can have a **strong** or **weak relationship**. A stronger relationship will have an **r-value close to 1 or -1**. We can get a line called the **least squares regression** line that will act as a line through the data points. We can tell the **"goodness of fit"** based on how close **r^2 is to 1**.

Terminology

Pattern

- form
- direction (if any)
- strength of pattern
  Linear, Nonlinear, No relationship
- Different relationships
  Strength of the association
- weak relationship
- strong relationship
  outlier
- a data value that has a low probability of occurence (it is unusual or unexpected)

MISC

- z scores do not have units

Relationships

Usually looking for *positive relationships*. So if one **increases**, the other should also **increase**.
eg) what relationship does cigarettes smoked and lung cancer rate have.

Normal Quantile

Shows whether the curve is skewed right, left, or normal.
plot the y-axis as the 5-number summary, plot the x-axis as the z-value normal distribution plot.

- will always be a positive curve
- will concave up if skewed right
- will concave down if skewed left
- if normal curve then linear

Scatterplot

- shows the relationship between two quantitative variables measured on the same individuals
- independent variable is `x axis`, response or dependent variable is y
- can add categorical variables by classifying the data into multiple groups *(x's instead of dots for*

saturday sales)

- look at the overall pattern and striking deviations from the pattern
- Linear, Nonlinear, No relationship
- Strength of the association (weak relationship, strong relationship)
- cannot use a Scatterplot for a quantitative variable and a categorical variable
  Nonlinear relationship
- can use a log transformation to make it linear

Measuring linear relationships

```
r = 1/(n-1) sigma( (xi-Xmean)/Sx * (yi-Ymean)/Sy )
```

The correlation measures the direction and strength of the linear relationship between two quantitative variables

- uses standardized values
- measures strength and direction of linear relationship
  Graph mean of x and y to divide the scatterplot into 4 regions.
  If most of the points are in quadrent 1 and 3, then it is a positive relationship.
  In quadrent 1, the increase from the mean x and the mean y, it is a positive increase.
- if this sum is positive it is a positive relationship, if negative then negative.
- Always between -1 and 1.
- The closer it is to -1 or 1 it becomes a stronger linear relationship.
- If 0, then there is no relationship

Least Squares Regression line

```
y = a + bx; b = r * Sy/Sx; a = ybar - b * xbar
```

- get a straight line to a scatterplot

- there is one line that minimizes the distances from all the points to the line

- minimization problem

- the slope b and r are directly proportional and b is positive iff r is positive

- regression line passes through point xbar, ybar

- regressions of x on y and y on x could be different

"Goodness of fit" - the coefficient of determination - r^2

```
r^2 = 0.x = x% certainty that the predicted value is correct
```

- when the linear relationship is strong it is good
- the linear relationship is strong when r -> 1, r -> -1
- because we want one value to represent the "goodness of fit" we square this value
- gives fraction of variation of y that is explained by the regression line