Econ 103 – Statistics for Economists

Chapter 2 feat. Z Scores and OLS

Mallick Hossain

University of Pennsylvania

Survey Results

Problems with Surveys

Raw Econ 103 Fall 2016 Survey.csv - LibreOffice Calc

	A	B	C	D	E	F
1	Gender	Jobs	Salary	Retirement	Drink	Team
2	Female	No idea, does this mean I won't do well in OCR interviews?	150k	1 million	Coffee	What are you talking about?
3	Male	100	500000	5000000	Coffee	What are you talking about?
4	Male		100000		Coffee	What are you talking about?
5	Female	40 billion (so 4,000,000 k)	85k	200k	Coffee	What are you talking about?
6	Male	120000	100000	10000000	Coffee	Mystic
7	Male	200000	300000	3000000	Neither	Valor
8	Female	231	115000	20% of salary	Coffee	What are you talking about?
9	Female	180 < x < 300	>=100K a year	Enough to live comfortably and happily - I don't know how much tha	Tea	Valor
10	Female	245K	120000	401K	Tea	What are you talking about?
11	Male	60	$180,000	$2 million	Tea	Mystic
12	Male	30	1,000,000+	10000000	Coffee	What are you talking about?
13	Female	77000	200,000 yearly	1.5 million	Coffee	What are you talking about?
14	Male	12	50000	0	Coffee	Instinct
15	Male	100	10000	1000000	Coffee	What are you talking about?
16	Male	255000	800000	250000	Coffee	What are you talking about?
17	Female	9	$140,000	$800,000	Coffee	Valor
18	Male	$40,000	$400,000	$2,000,000	Coffee	Valor
19	Male	1	150000	100000	Tea	What are you talking about?
20	Female	200000	200000	500000	Coffee	What are you talking about?
21	Male		100000	I don't know	Coffee	What are you talking about?
22	Male				Tea	Valor
23	Male	4	100000	$4	Tea	What are you talking about?
24	Male		150000		Tea	What are you talking about?
25	Male	35	250000	2000000	Coffee	What are you talking about?
26	Male	0	120000	960000	Coffee	What are you talking about?
27	Female	2000	80000	1000000	Neither	What are you talking about?
28	Male		700000	3000000	Coffee	What are you talking about?

Types of Variables

Discussion!

- What are the differences between the following variables?
  - “Age” and “gender”
  - “Gender” and “class standing”
  - “SAT score” and “job creation”

A Few Definitions: A Taxonomy of Variables

- Variables
  - Categorical (Qualitative)
    - Discrete
      - Nominal
      - Ordinal
  - Numerical (Quantitative)
    - Continuous
      - Ordinal
      - Interval
      - Ratio
    - Discrete
      - Ordinal
      - Interval
      - Ratio

Definitions

- **Discrete:** takes a countable number of values
- **Continuous:** takes on any value
- **Nominal:** no order to the categories
- **Ordinal:** categories with natural order
- **Interval:** only differences meaningful, no natural zero
- **Ratio:** differences and ratios meaningful, natural zero

Summary Statistics

Definitions

1. Measures of Central Tendency
    - **Mean:** the average ("balance point")

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

    - **Median:** the middle observation (if data has even number of observations, take the mean of the middle two observations)

Definitions

2. Percentiles

$$P^{th}\text{Percentile = Value in}\,(P/100)\cdot(n+1)^{th}\text{Ordered Position}$$

Note: this is not the only definition of percentiles, but it is the one we will use for this course!

Percentile Example: n = 12

60 63 65 67 70 72 75 75 80 82 84 85

$$
\begin{aligned}
Q_1 &= \text{value in the } 0.25(n+1)^{th} \text{ ordered position} \\
&= \text{value in the } 3.25^{th} \text{ ordered position} \\
&= 0.75 * 65 + 0.25 * 67 \\
&= 65.5
\end{aligned}
$$

Definitions

3. Measures of Spread
   - **Variance:** the spread from the mean

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

   - **Standard Deviation:** another way to measure the spread

$$s = \sqrt{s^2}$$

   - **Range:** the distance between the highest and lowest value

$$Range = |x_{max} - x_{min}|$$

   - **Interquartile Range (IQR):** the distance between the upper and lower quartiles

$$IQR = |x_{75\%} - x_{25\%}|$$

Why Is Variance Squared?

What happens if we do not square the deviations?

$$
\begin{aligned}
\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x}) = \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}\bar{x}\right] &= \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i-n\bar{x}\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i-n\cdot\frac{1}{n}\sum_{i=1}^{n}x_i\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}x_i\right] = 0
\end{aligned}
$$

4. Measure of Symmetry
   - **Skewness:** a measure of symmetry, positive values means the right tail is longer and vice versa

$$Skewness = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

Skewness – A Measure of Symmetry

$$\boxed{\text{Skewness} = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}}$$

What do the values indicate?

Zero $\Rightarrow$ symmetry, positive right-skewed, negative left-skewed.

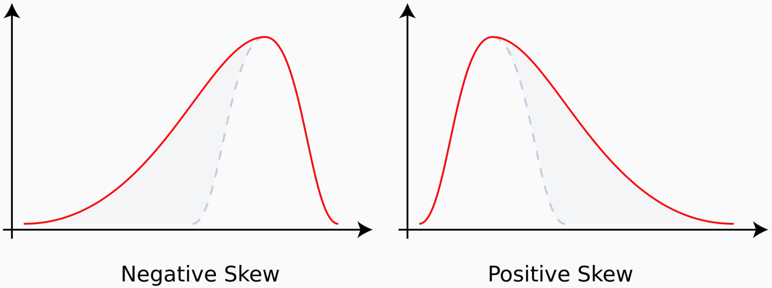
Why cubed?

To get the desired sign.

Why divide by s^3?

So that skewness is unitless

Skewness – A Measure of Symmetry



Definitions

5. Relationship between variables
    - **Covariance:** how two variables vary together
    - **Correlation:** normalized version of covariance (ranges between -1 and +1)
    - **Regression:** an estimation of how two variables are related

Charts

Some Data Visualization Quotes

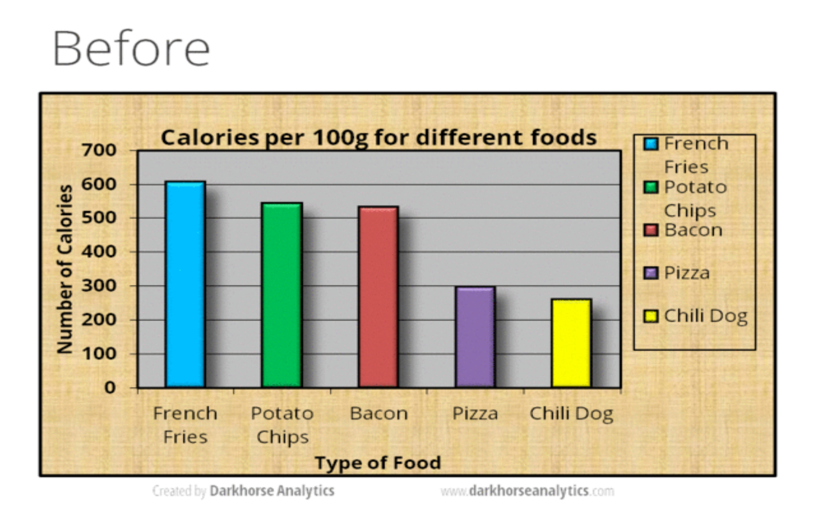
1. “Overload, clutter, and confusion are not attributes of information, they are failures of design” –Edward Tufte
2. “...few people will appreciate the music if I just show them the notes. Most of us need to listen to the music to understand how beautiful it is. But often, that’s how we present statistics; we just show the notes we don’t play the music.” –Hans Rosling

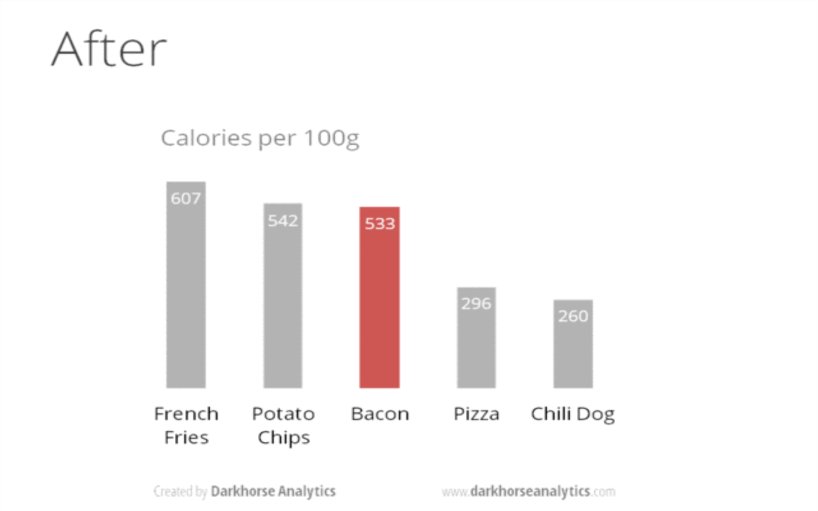
Simplicity

Simplicity is about subtracting the obvious and adding the meaningful

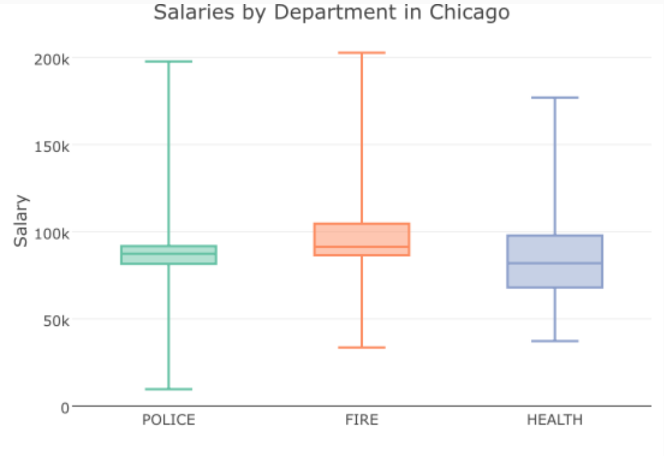
An Illustration

https://i.imgur.com/W4BKCVU.gif


Before
Calories per 100g for different foods
Number of Calories
700
600
500
400
300
200
100
0
French Fries
Potato Chips
Bacon
Pizza
Chili Dog
Type of Food
French Fries
Potato Chips
Bacon
Pizza
Chili Dog
Created by Darkhorse Analytics
www.darkhorseanalytics.com

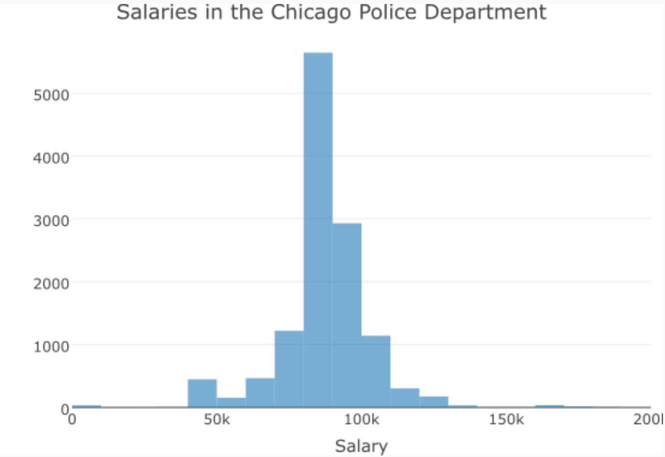

After
Calories per 100g
607
542
533
296
260
French Fries
Potato Chips
Bacon
Pizza
Chili Dog
Created by Darkhorse Analytics
www.darkhorseanalytics.com

Illustrating with Charts (Box and Whisker Chart)



What summary statistics can you infer from this chart?

Illustrating with Charts (Histogram)

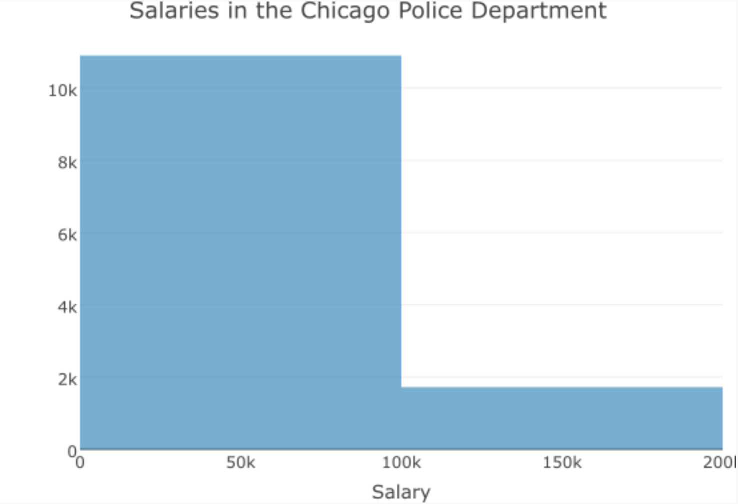


What summary statistics can you infer from this chart?

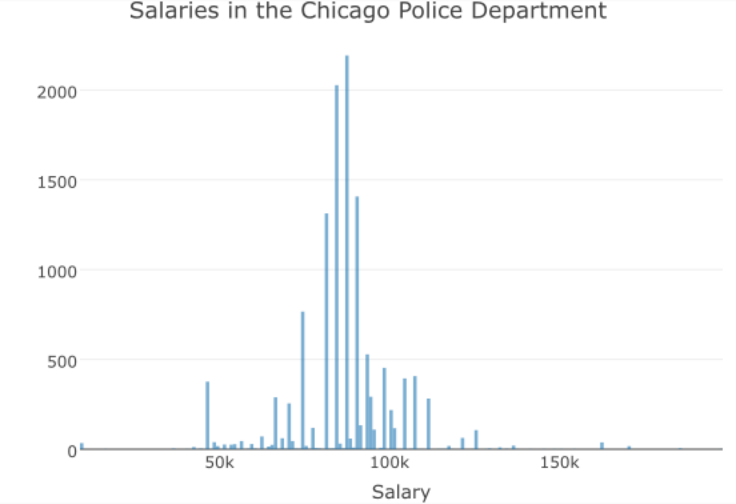
Histograms are *Really* Important

1. Histograms show the frequency of different observations
2. Important Choice: How many bins?

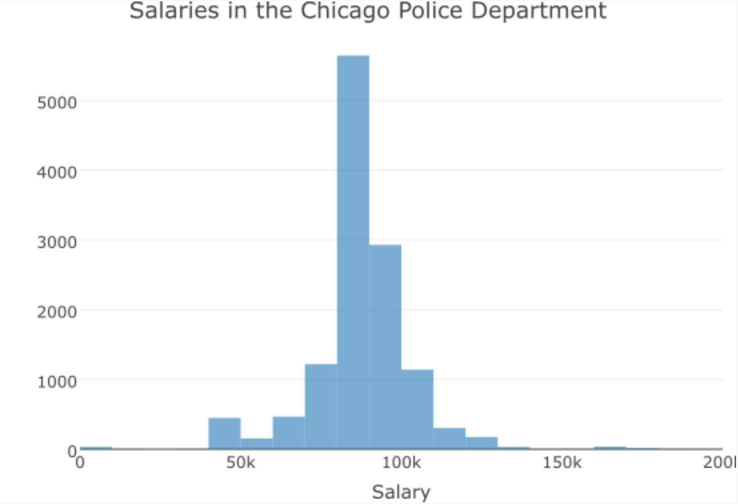
Too Few Bins (Oversmoothing)



Too Many Bins (Undersmoothing)



Just Right! (Usually around 20 bins or so)



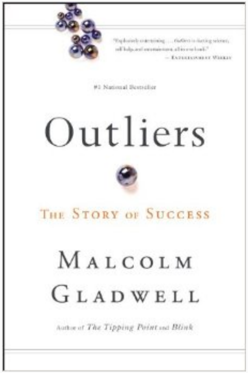
Questions to Ask Yourself about Each Summary Statistic

1. What does it measure?
2. What are its units compared to those of the data?
3. How do its units change if those of the data change?
4. What are the benefits and drawbacks of this statistic?

Some of the information regarding items 2 and 3 is on the homework rather than in the slides because working it out for yourself is a good way to check your understanding.

Outliers

What is an Outlier?



Outlier: A very unusual observation relative to the other observations in the dataset (i.e. very small or very big).

Which Summary Stats are Sensitive to Outliers?

- Assume our data is 1, 2, 3, 4, 5, 6, 7, 8, 9. What are our summary stats (mean, median, variance, range, IQR)?
- What will be affected if the data includes an outlier and becomes 1, 2, 3, 4, 5, 6, 7, 8, 63?

Which Summary Stats are Sensitive to Outliers?

- Mean changes from 5 to 11
- Median remains at 5
- Variance changes from 7.5 to 385.5
- Range changes from 8 to 62
- IQR remains at 5
- When Does the Median Change? IQR?
    - Ranks would have to change.

Variance

Essentially the average squared distance from the mean. Sensitive to both skewness and outliers.

Standard Deviation

$\sqrt{\text{Variance}}$, but more convenient since same units as data

Range

Difference between larges and smallest observations. *Very* sensitive to outliers.

Interquartile Range

Range of middle 50% of the data. Insensitive to outliers, skewness.

Sample vs. Population

Essential Distinction: Sample vs. Population

For now, you can think of the population as a list of N objects:

Population: $x_1, x_2, \ldots, x_N$

from which we draw a sample of size $n < N$ objects:

Sample: $x_1, x_2, \ldots, x_n$

Important Point:

Later in the course we'll be more formal by considering probability models that represent the *act of sampling* from a population rather than thinking of a population as a list of objects. Once we do this we will no longer use the notation N as the population will be *conceptually infinite*.

Essential Distinction: Parameter vs. Statistic

N individuals in the Population, n individuals in the Sample:

	Parameter (Population)	**Statistic** (Sample)
Mean	$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Var.	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
S.D.	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$

Notation Matters!

We use a sample $x_1, \ldots, x_n$ to calculate statistics (e.g. $\bar{x}$, s^2, s) that serve as estimates of the corresponding population parameters (e.g. μ, σ^2, σ).

Why Do Sample Variance and Std. Dev. Divide by $n - 1$?

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

There is an important reason for this, but explaining it requires some concepts we haven't learned yet.

Some Intuition

- **Intuition 1:** If we only had one data point, what would be the sample variance? Would it even be defined?
- **Intuition 2:** We know that the deviations from the sample mean sum to zero (see discussion of why variance is squared). Hence, we only need to know $n-1$ of the deviations since the last one will be whatever it takes to make the sum of them equal to 0. Hence, it would be proper to divide by $n-1$ instead of n

Z Scores

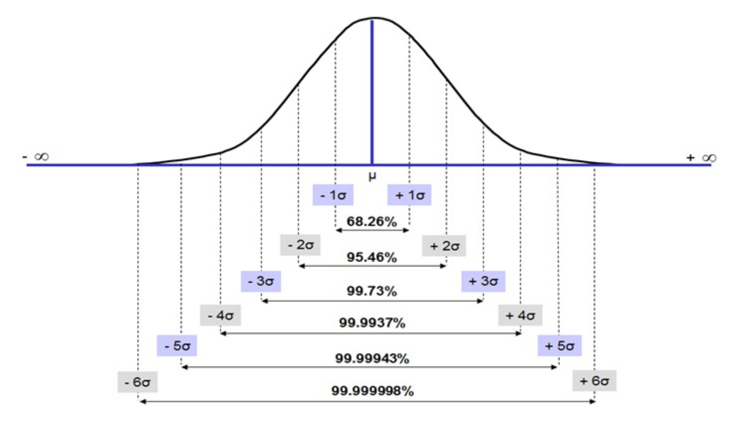
Why Mean and Variance (and Std. Dev.)?

Empirical Rule

For large populations that are approximately bell-shaped, standard deviation tells where most observations will be relative to the mean:

- $\approx$ 68% of observations are in the interval $\mu \pm \sigma$
- $\approx$ 95% of observations are in the interval $\mu \pm 2\sigma$
- Almost all of observations are in the interval $\mu \pm 3\sigma$

Standard Deviations



Physics uses a five-sigma rule (i.e. this could only happen normally 0.00057% of the time!)

Z-scores: How many standard deviations from the mean?

$$z_i = \frac{x_i - \bar{x}}{s}$$

Unitless

Allows comparison of variables with different units.

Detecting Outliers

Measures how “extreme” one observation is relative to the others.

Linear Transformation of x

What is the sample mean of the z-scores?

$$\begin{aligned}
\bar{z} &= \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i - \bar{x}}{s} = \frac{1}{n \cdot s}\left[\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\bar{x}\right] \\
&= \frac{1}{n \cdot s}\left[\sum_{i=1}^{n} x_i - n\bar{x}\right] = \frac{1}{n \cdot s}\left[\sum_{i=1}^{n} x_i - n \cdot \frac{1}{n}\sum_{i=1}^{n} x_i\right] \\
&= \frac{1}{n \cdot s}\left[\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i\right] = 0
\end{aligned}$$

What is the variance of the z-scores?

$$s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2 = \frac{1}{n-1}\sum_{i=1}^{n} z_i^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)^2$$

$$= \frac{1}{s_x^2}\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = \frac{s_x^2}{s_x^2} = 1$$

So what is the *standard deviation* of the z-scores?

Population Z-scores and the Empirical Rule: $\mu \pm 2\sigma$

If we knew the population mean μ and standard deviation σ we could create a *population version* of a z-score. This leads to an important way of rewriting the Empirical Rule:

Bell-shaped population $\Rightarrow$ approx. 95% of observations x_i satisfy

$$\mu - 2\sigma \leq x_i \leq \mu + 2\sigma$$

$$-2\sigma \leq x_i - \mu \leq 2\sigma$$

$$-2 \leq \frac{x_i - \mu}{\sigma} \leq 2$$

Covariance and Correlation

Two Samples of Numeric Data

$x_1, \ldots, x_n$ and $y_1, \ldots, y_n$

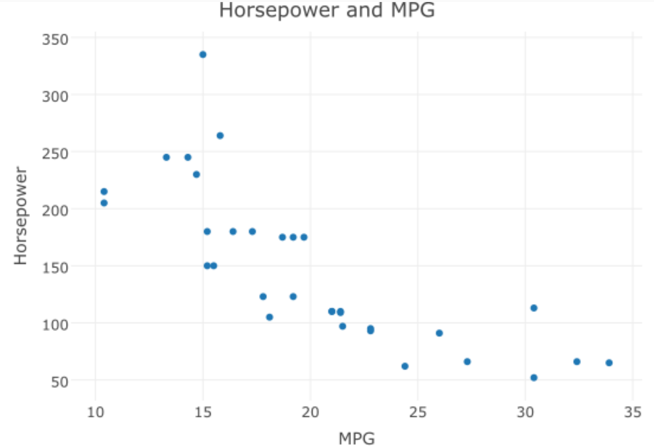
Dependence

Do x and y both tend to be large (or small) at the same time?

Key Point

Use the idea of centering and standardizing to decide what "big" or "small" means in this context.

Are Engine Cylinders and Horsepower Related?



How do we quantify this relationship?

Covariance

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- Centers each observation around its mean and multiplies.
- Zero $\Rightarrow$ no linear dependence
- Positive $\Rightarrow$ positive linear dependence
- Negative $\Rightarrow$ negative linear dependence
- Population parameter: σ_{xy}
- Units?
- Protip: If you know this, you know variance!

Correlation

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{s_{xy}}{s_x s_y}$$

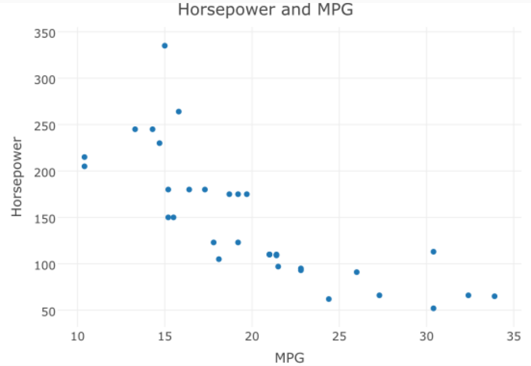
- Centers *and* standardizes each observation
- Bounded between -1 and 1
- Zero $\Rightarrow$ no linear dependence
- Positive $\Rightarrow$ positive linear dependence
- Negative $\Rightarrow$ negative linear dependence
- Population parameter: ρ_{xy}
- Unitless

Game!

`guessthecorrelation.com`

Introduction to Regression

Least Squares Regression – Predict Using a Line



- In order to fit a line through this, we need to estimate $\hat{y}_i = a + bx_i$
- How do we find a and b?

Finding a and b

- Given our data (x_i, y_i), linear regression chooses the slope (b) and intercept (a) that minimize the sum of squared vertical deviations

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

- Why do we square the deviations?
- Note: Once we have our model, we can make predictions for any x and y!

Solving for a

Differentiate with respect to a

$$
\begin{aligned}
-2\sum_{i=1}^{n}(y_i - a - bx_i) &= 0 \\
\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a - b\sum_{i=1}^{n} x_i &= 0 \\
\frac{1}{n}\sum_{i=1}^{n} y_i - \frac{na}{n} - \frac{b}{n}\sum_{i=1}^{n} x_i &= 0 \\
\bar{y} - a - b\bar{x} &= 0 \\
\bar{y} - b\bar{x} &= a
\end{aligned}
$$

Regression Line Goes Through the Means!

$$\boxed{\bar{y} = a + b\bar{x}}$$

Solving for b

Substitute for a

$$\sum_{i=1}^{n}(y_i - a - bx_i)^2 = \sum_{i=1}^{n}(y_i - \bar{y} + b\bar{x} - bx_i)^2$$
$$= \sum_{i=1}^{n}\left[(y_i - \bar{y}) - b\left(x_i - \bar{x}\right)\right]^2$$

FOC wrt b

$$-2\sum_{i=1}^{n}\left[(y_i - \bar{y}) - b\left(x_i - \bar{x}\right)\right]\left(x_i - \bar{x}\right) = 0$$

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right) - b\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = 0$$

$$b = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

Simple Linear Regression

Problem

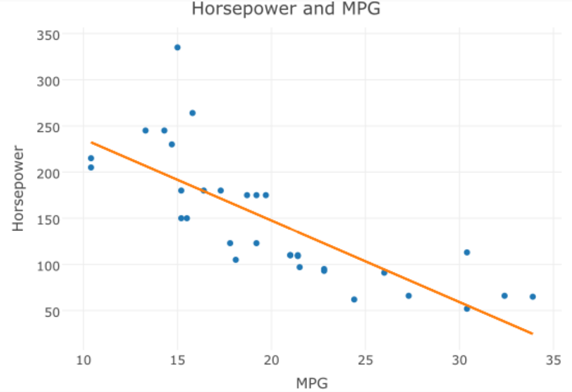
$$\min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Solution

$$b = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

Fitting a Line



$$\widehat{hp} = 324.08 - 8.83mpg$$

Relating Regression to Covariance and Correlation

$$b = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\frac{1}{n-1}\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}$$

$$r = \frac{s_{xy}}{s_x s_y} = b\frac{s_x}{s_y}$$

Units

Correlation is unitless, covariance and regression coefficients (a, b) are not. (What are the units of these?)

Symmetry

Correlation and covariance are symmetric, regression isn't. (Switching x and y axes changes the slope and intercept.)

Checking Our Results

Model: $hp = a + b * mpg$

Facts:

$s_{mpg,hp} = -321 \quad s_{mpg} = 6 \quad s_{hp} = 69 \quad \bar{mpg} = 20, \quad \bar{hp} = 147$

What is the sample correlation between MPG (x) and horsepower (y)?

$$r = \frac{s_{xy}}{s_x s_y} = \frac{-321}{6 \times 69} \approx -0.78$$

Checking our Results

Model: $hp = a + b * mpg$

Facts:

$s_{mpg,hp} = -321 \quad s_{mpg} = 6 \quad s_{hp} = 69 \quad \bar{mpg} = 20, \quad \bar{hp} = 147$

What is the value of b for the regression?

$$b = \frac{s_{xy}}{s_x^2} = \frac{-321}{6^2} \approx -8.9$$

Checking our Results

Model: $hp = a + b * mpg$

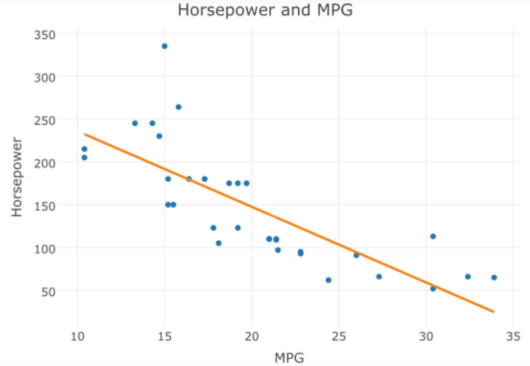
Facts:

$s_{mpg,hp} = -321 \quad s_{mpg} = 6 \quad s_{hp} = 69 \quad \bar{mpg} = 20, \quad \bar{hp} = 147$

What is the value of a for the regression ($b = -8.9$)?

$$a = \bar{y} - b\bar{x} = 147 - (-8.9) \times 20 \approx 325$$

Checking our Results

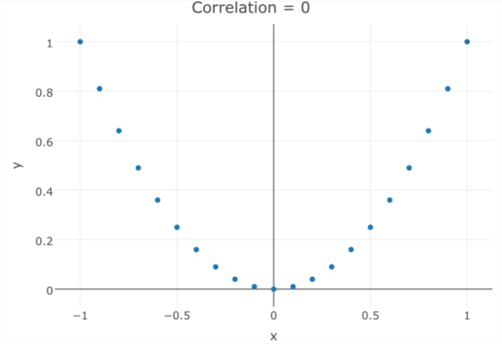


$$a \approx 325 \quad b \approx -8.9$$

$$\widehat{hp} = 324.08 - 8.83mpg$$

EXTREMELY IMPORTANT

- Regression, Covariance, and Correlation: linear association.
- Linear association $\neq$ causation.
- Linear is not the only kind of association!



Review

Essential Distinction: Parameter vs. Statistic

N individuals in the Population, n individuals in the Sample:

	Parameter (Population)	**Statistic** (Sample)
Mean	$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Var.	$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$	$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
S.D.	$\sigma_x = \sqrt{\sigma_x^2}$	$s_x = \sqrt{s^2}$
Cov.	$\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$	$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$
Corr.	$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$	$r = \frac{s_{xy}}{s_x s_y}$

Related Reading

- Wonnacott: Chapter 2, Section 4-5 A+B, Section 5-3, 11-1, 11-2, Appendices to 2-2, 2-5, 11-1, and 11-2
- How to Lie with Statistics: Chapters 2, 5, and 6
- If you're rusty on permutations, combinations, etc., read the "Permutations and Combinations" document on the course page (`mallickhossain.com/econ-103`)

Homework

- Chapter 2 Problems
- Additional Problems
- R Tutorial 2