Unit 2: Exploring Two-Variable Data

Merrick Fanning

September 27, 2025

Unit 2 Overview

- Representing Relationships in Bivariate Categorical Variables
- Statistics for Two Categorical Variables
- Representing Relationships Between Bivariate Quantitative Variables
- Covariance
- Correlation
- Linear Regression Models
- Residuals
- Least Squares Regression
- Analyzing Departures from Linearity

Are redshirts doomed?

Question: In the original series of Star Trek, red-uniformed crew members were said to have a higher fatality rate during missions. Is there statistical evidence that redshirts are more likely to die?

Where Does the Data Come From?

- The dataset was compiled by **Matthew Barsalou** and featured in *Significance Magazine*.
- It analyzes **Star Trek** Enterprise NCC-1701 casualties from episodes aired between September 8, 1966 and June 3, 1969.
- Casualty data were based on fan-curated records from **Memory Alpha**, a Star Trek wiki.
- Full article: *Keep Your Redshirt On: A Bayesian Exploration*

Crew Member	**Area**	**Shirt Color**	**Status**
Talia	Operations, Engineering	Red	DEAD
Matthew	Command and Helm	Gold	DEAD
Nolan	Science and Medical	Blue	Alive
...	...	...	...

Table: Dataset includes information on all 430 crew members over the time interval.

We are interested in exploring the relationship between two categorical variables:

- X : Shirt Colour (Red, Gold, or Blue)
- Y : Status (Dead or Alive)

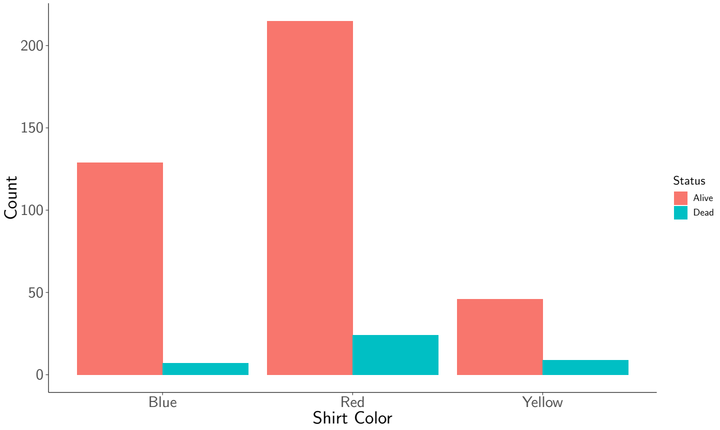
We can tabulate the data in a table:

	Alive	Dead	Total
Blue	129	7	136
Yellow	46	9	55
Red	215	24	239
Total	390	40	430

Table: Contingency table of shirt color vs. crew status aboard the Enterprise

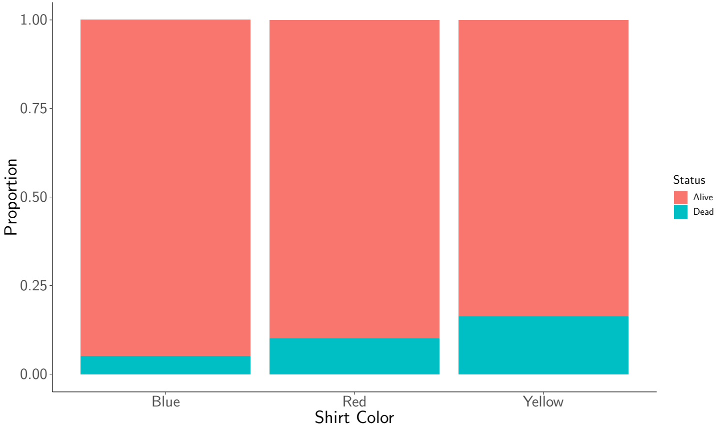
- The table shows the **marginal** and **joint** distributions.
- Using the table we can estimate conditional probabilities.

Bar Charts



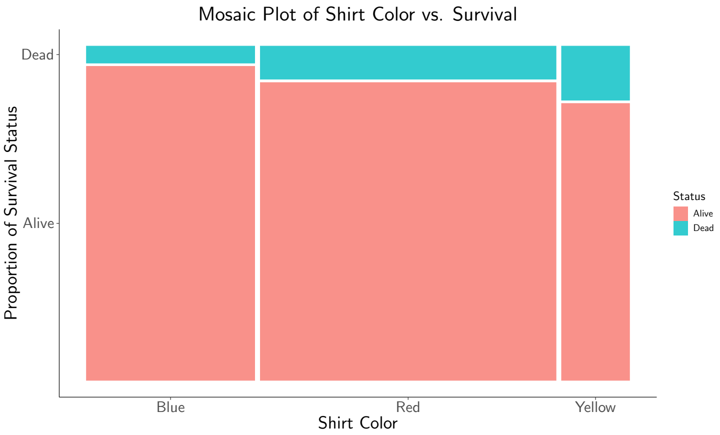
- With bar charts it can be difficult to see relationships between variables.

Relative Bar Charts



- How would you detect an association between variables by observing a relative bar chart?
- What would the bar chart look like for variables that are independent?

Mosaic Plots

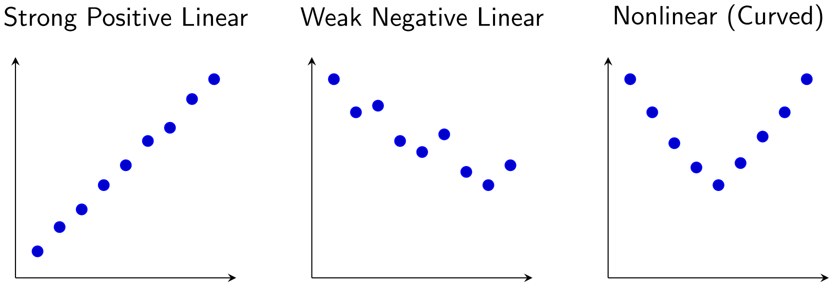


- Why might a mosaic plot be preferred over a relative bar chart?

Describing Relationships Between Two Numeric Variables

We represent relationships between two numeric variables (sample data) using scatter plots. To describe a relationship between two quantitative variables, consider:

- **Form:** Linear, curved, or no pattern
- **Direction:** Positive or negative trend
- **Strength:** Strong if points closely follow a pattern

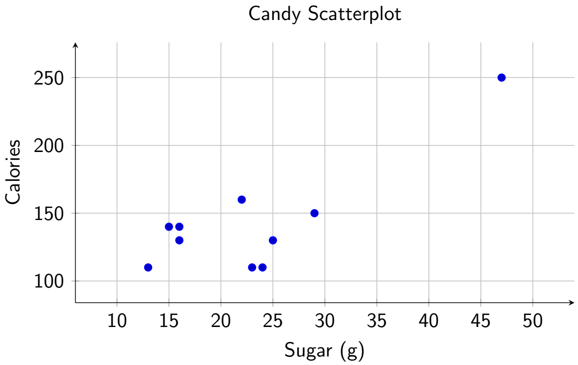


Candy, Sugar, and Calories

Graph sugar content vs. calories for the following dataset with your Ti-84 Calculator. Describe the relationship.

Candy	**Sugar (g)**	**Calories**
Skittles	47	250
Peanut M&M's	15	140
Twizzlers	13	110
Sour Patch Kids	24	110
Milk Duds	16	130
Reese's Pieces	16	140
Junior Mints	25	130
Swedish Fish	23	110
Starburst	22	160
Mike and Ike	29	150

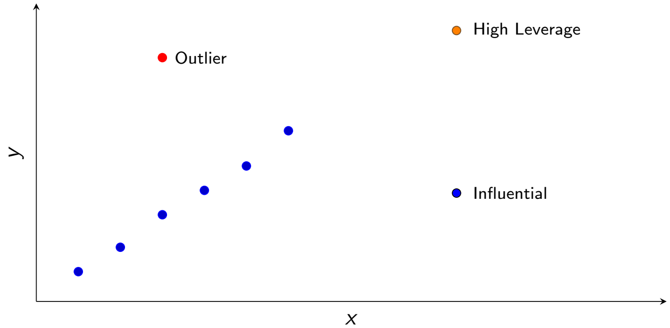
Scatterplot: Sugar vs. Calories in Movie Candy



Description: There is a moderately strong, positive, and roughly linear relationship between sugar content and calorie count in movie theatre candies.

Outliers and Influential Points in Regression

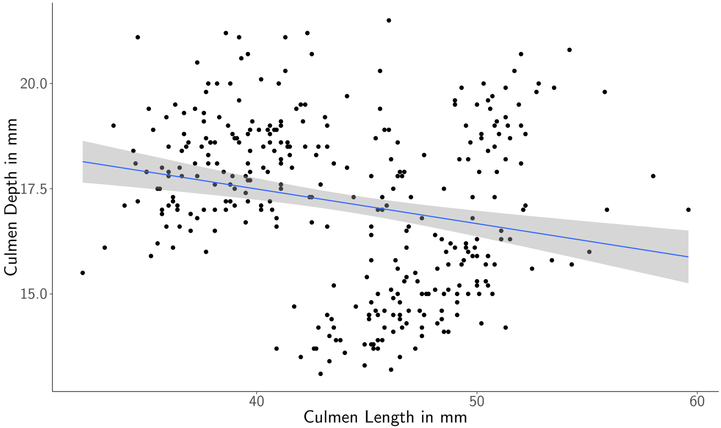
- **Outlier:** A point that deviates from the overall y-pattern.
- **High-leverage:** A point with an extreme x-value.
- **Influential:** A point that substantially changes the regression line if removed (often high-leverage and far from the line).



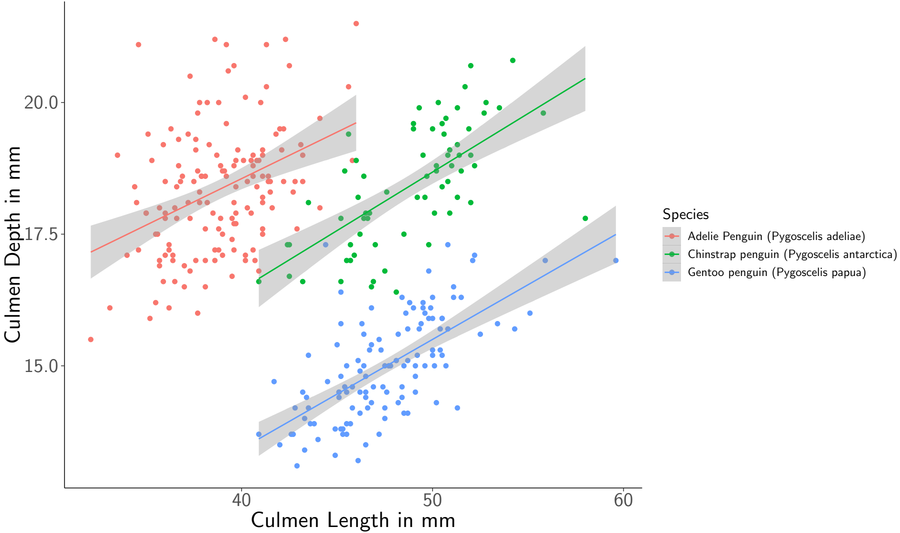
Simpson's Paradox with Two Numeric Variables

Simpson's Paradox: A trend that appears in several groups of data reverses when the groups are combined.

Example: Culmen length vs depths for penguins.



Simpson's Paradox with Two Numeric Variables



Simpson's Paradox with Two Categorical Variables

Simpson's Paradox: A trend that appears in separate groups reverses when the data are combined.

Example: Admission by Gender and Department

Department A (Easier)

	Admitted	Total
Men	80/100	80%
Women	18/20	90%

Department B (Harder)

	Admitted	Total
Men	20/100	20%
Women	54/180	30%

Combined Totals

	Admitted	Total
Men	100/200	**50%**
Women	72/200	**36%**

Paradox: Women have a higher acceptance rate in both departments, but a lower overall acceptance rate because more women applied to the more competitive department.

Understanding Covariance

Covariance measures the direction of a linear relationship between two quantitative variables.

- If **positive**, large values of x tend to go with large y, and small with small.
- If **negative**, large values of x go with small y, and vice versa.
- If close to **zero**, there is no linear association.

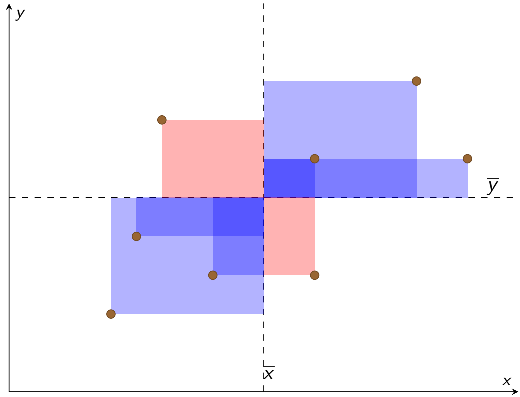
Formula:

$$\mathrm{Cov}(X, Y) = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$

1. What does this formula mean geometrically?
2. Why is the value for covariance hard to interpret?

Geometric Interpretation of Covariance

Covariance as Signed Area: Each point contributes a value $(x_i - \overline{x})(y_i - \overline{y})$, interpreted as the signed area of a rectangle.



Key Idea:

- Blue rectangles (Quadrants I & III): contribute positively.
- Red rectangles (Quadrants II & IV): contribute negatively.
- Covariance is the average of these signed areas.

From Covariance to Correlation

Correlation standardizes covariance:

Covariance

$$\mathrm{Cov}(X, Y) = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$

Correlation

$$r = \frac{\mathrm{Cov}(X, Y)}{s_x s_y} = \frac{1}{n-1} \sum \frac{(x_i - \bar{x})}{s_x} \frac{(y_i - \bar{y})}{s_y}$$

Interpretation of r:

- r measures the **direction** and **strength** of a linear relationship.
- $r > 0$: positive association; $r < 0$: negative association.
- $|r|$ close to 1: strong linear pattern; close to 0: weak linear pattern.
- r has no units and is always between -1 and $+1$.

From Covariance to Correlation

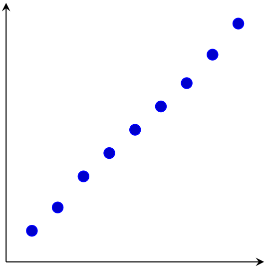
Guidelines for describing strength:

Absolute Value of r	**Interpretation**
$0.0 \leq \lvert r \rvert < 0.3$	Weak linear relationship
$0.3 \leq \lvert r \rvert < 0.7$	Moderate linear relationship
$0.7 \leq \lvert r \rvert \leq 1.0$	Strong linear relationship

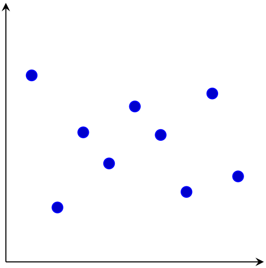
Visualizing Correlation (r) Values

How does the strength and direction of a linear relationship look for different values of r?

Strong Positive Correlation



No Linear Correlation

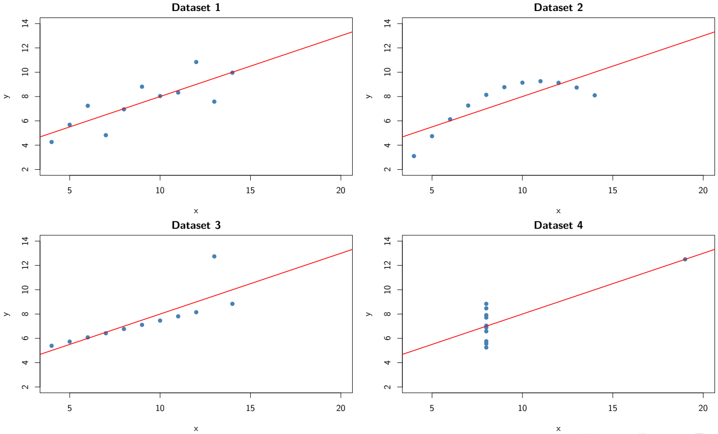


Try it yourself: Use the online applet to practice estimating correlation:
`Guess the Correlation Applet`

Anscombe's Quartet: Why Graphing Matters

Anscombe's Quartet consists of four datasets that have:

- The same mean and standard deviation for x and y
- The same correlation $r \approx 0.816$
- The same regression line



Francis John Anscombe (1918-2001)

- **Born**: May 13, 1918, in Hove, East Sussex, England
- **Education**: Trinity College, Cambridge (B.A. 1939, M.A. 1943)
- **Career Highlights**:
  - Lecturer in mathematics at Cambridge University (1948-1956)
  - Moved to the United States in 1956; became a professor at Princeton University
  - Founding chair of the Department of Statistics at Yale University (1963-1988)
- **Notable Contributions**:
  - Developed **Anscombe's Quartet** to illustrate the importance of data visualization
  - Co-authored foundational work on subjective probability with Robert Aumann

Correlation Does Not Imply Causation

Just because two variables are correlated doesn't mean one causes the other! Is Global Warming causing the number of Pirates to decline? Are Ice-cream sales responsible for higher frequency of drowning rates? The rise of the NBA beard.
Real (but ridiculous) examples from `SpuriousCorrelations.com`:

- Per capita **cheese consumption** correlates with deaths by **bedsheet entanglement**.
- The number of **people who drowned in a pool** tracks with the number of **Nicolas Cage films released**.

Key AP Statistics Message

- Correlation does not imply causation. How do we "prove" something is a causal relationship?
- A strong correlation may be due to:
  - Coincidence
  - A lurking variable
  - Or utter nonsense!

The Simple Linear Regression Model

We model the relationship between two quantitative variables using the equation:

$$\hat{y} = a + bx$$

This is an *estimate* of the true relationship $y = \alpha + \beta x + \epsilon$ where $\epsilon \sim \mathrm{Normal}(0, \sigma)$.

- $\hat{y}$: predicted value of the response variable
- a: y-intercept (predicted average value of y when $x = 0$)
- b: slope (amount on average y changes for each one unit increase in x). Remember **rise over one**
- x: explanatory variable
- y: response variable

Formulas:

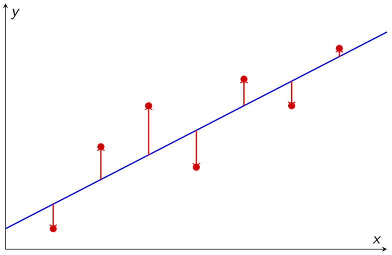
$$b = r \cdot \frac{s_y}{s_x} \qquad \text{and} \qquad a = \bar{y} - b\bar{x}$$

Least squares regression line will **always** pass through $(\overline{x}, \overline{y})$

Visualizing Residuals

Each residual is the vertical distance between the observed value and the predicted value on the regression line.

$$i^{th} \text{ Residual} = \hat{e}_i = y - \hat{y}$$

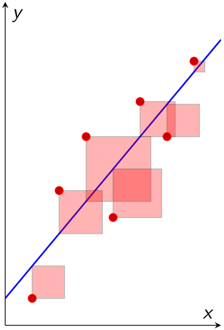


1. What do positive / negative residuals mean?
2. How do we arrive at our equations using the residuals?
3. Why can't we simply minimize the sum of residuals?

Least Squares Regression: Minimizing Squared Residuals

The least squares regression line minimizes the sum of squared residuals:

$$\text{Minimize} \quad \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$



Minimization:

$$\frac{\partial}{\partial a} \sum (y_i - a - bx_i)^2 = 0$$

$$\frac{\partial}{\partial b} \sum (y_i - a - bx_i)^2 = 0$$

Solving these gives the least squares estimates. Check out Interactive Least Squares on Desmos

Exploring Dolbear's Law

Can crickets be used as thermometers? In 1897, physicist Amos Dolbear proposed a formula relating cricket chirps to air temperature:

$$\text{Temperature } (^\circ\text{F}) \approx 40 + \frac{\text{Chirps per minute} - 40}{4}$$

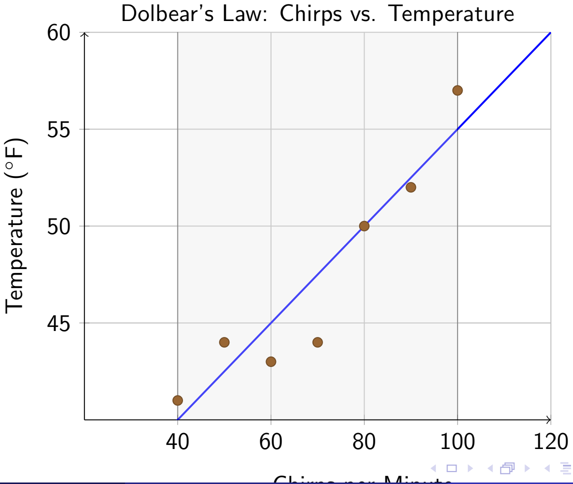
- This works reasonably well for the snowy tree cricket.
- **Why does this relationship exist?** - Crickets are cold-blooded; their metabolism speeds up with temperature.
- The formula is only accurate in a specific range (about 55-100°F).

Dolbear's original paper: "The Cricket as a Thermometer," The American Naturalist (1897).

Extrapolation vs. Interpolation

Why am I telling you about cricket chirps?

- **Interpolation:** Predicting within the observed range of data.
- **Extrapolation:** Predicting outside the observed range - risky or misleading.



Departures from Linearity

How do we tell when data doesn't fit the linear regression model?

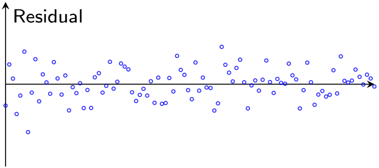
1. Residual plots allow us to see hidden patterns in relationship that isn't clear in scatter plot.
2. r^2 tells us how much of the variation in the response variable is captured or explained by the model.

Diagnosing Residual Plots

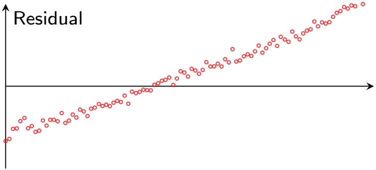
Residual plots help assess the appropriateness of a linear model.

- A good model has **no pattern** (unbiased)
- And **constant vertical spread** (homoscedastic)

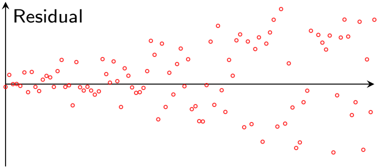
Unbiased & Homoscedastic



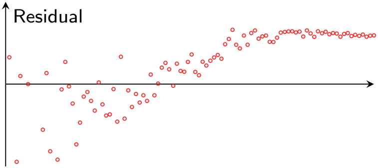
Biased & Homoscedastic



Unbiased & Heteroscedastic



Biased & Heteroscedastic



The Coefficient of Determination (r^2)

What is r^2?

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}} = \frac{\text{SSR}}{\text{SST}}$$

Where:

- SSE $= \sum (y_i - \hat{y}_i)^2$ - Sum of Squared Errors (Residuals)
- SST $= \sum (y_i - \bar{y})^2$ - Total Sum of Squares
- SSR $= \sum (\hat{y}_i - \bar{y})^2$ - Regression Sum of Squares

Interpretation:

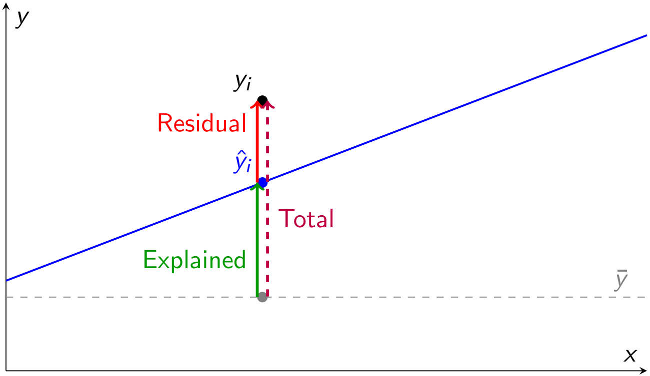
- r^2 measures the proportion of variability in the response variable explained by the least squares regression model.
- If $r^2 = 0.82$, then **82% of the variation in** y is explained by its linear relationship with x.
- A higher r^2 means a better fit, but it **does not prove causation**.

Visual Breakdown of Variability for r^2

For a single point, total variability can be broken into:

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

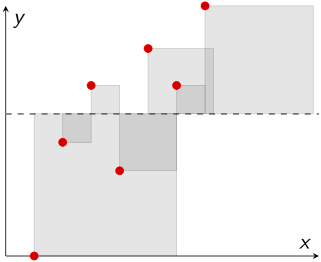
One Observation's Contribution to SST, SSR, SSE



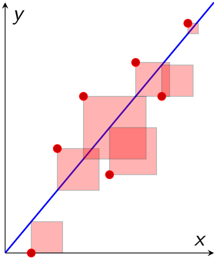
Visualizing r^2: SST vs. SSE

Comparing Total vs. Unexplained Variability

Total Variability SST $= \sum(y_i - \overline{y})^2$



Unexplained Variability SSE $= \sum(y_i - \hat{y}_i)^2$



SSR is the variability that is explained by the model SSR $=$ SST$-$SSE.

$$\frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} = r^2$$

Standard Error of the Regression Line

What is the Standard Error of the Regression Line s?

$$s = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

- s is the **standard deviation of the residuals**.
- It tells us, on average, how far the actual values y_i are from the predicted values $\hat{y}_i$.
- In other words: **how far off our model tends to be** when making predictions.
- It is measured in the same units as the response variable.
- It is an estimate of σ for the regression model $\epsilon \sim \mathrm{Normal}(0, \sigma)$.

Interpretation: If $s = 2.3$, then our model typically over or under predicts y by about 2.3 units on average.

Conditions for Linear Regression: LINER

Before using a least-squares regression model, we must check the following conditions:

L - Linearity

The relationship between the explanatory and response variables should be linear. *Check:* Scatterplot and residual plot (look for no curves).

I - Independence

The observations should be independent of each other.
Check: Study design (e.g., random sampling or random assignment).

N - Normality of Residuals

The residuals should be roughly normally distributed.
Check: Histogram or normal probability plot of residuals.

E - Equal Variance (Homoscedasticity)

The spread of residuals should be roughly constant across all values of x.
Check: Residual plot should show consistent vertical spread.

R - Randomness

The data should come from a random process.
Check: Look for mention of random sampling or random assignment.

Example: Modelling Penguin Body Mass with Linear Regression

Experimental Data (Adelie Penguins):

Culmen Depth (mm)	Mass (g)	Culmen Depth (mm)	Mass (g)
17.0	3750	19.1	3875
18.1	3800	19.4	4050
18.3	3700	19.5	4000
18.6	3850	19.7	4025
18.7	3850	19.9	4250
18.8	3700	20.1	4400
18.9	3700	20.3	4500
19.0	3950	20.4	4450
19.0	4000	20.6	4550
19.1	3950	20.8	4600

1. Determine the regression equation using culmen depth to predict body mass.
2. Determine and interpret a and b for the regression model.
3. Determine and interpret r^2 and s for the regression model.

Example: Modeling Heart Rate Response to Caffeine

Context: A medical researcher investigates how caffeine affects resting heart rate. A group of adults is given increasing doses of caffeine, and their heart rate (in bpm) is measured after 30 minutes.

Experimental Data:

Caffeine (mg)	Heart Rate (bpm)
0	68
50	72
100	75
150	79
200	83
250	84

1. Determine the regression equation from the experimental data.
2. Determine and interpret a and b for the regression model.
3. Determine and interpret r^2 and s for the regression model.

Example: Modeling Hooke's Law with Linear Regression

Context: A physics student investigates Hooke's Law, which says that the force needed to stretch a spring is proportional to how far it's stretched:

$$F = k\Delta x$$

where F is the applied force (N), Δx is the displacement (m), and k is the spring constant.

Experimental Data:

Displacement (m)	**Force (N)**
0.01	0.18
0.02	0.41
0.03	0.60
0.04	0.83
0.05	1.03
0.06	1.21

1. Determine the regression equation from the experimental data.
2. Determine and interpret a and b for the regression model.
3. Determine and interpret r^2 and s for the regression model.

Transformations in Regression Models

Why transform data in regression?

- The standard linear model assumes:
  - A linear relationship between variables
  - Constant variability (equal spread)
  - Normally distributed residuals
- When these assumptions are violated, a transformation can help:
  - Make the relationship more linear
  - Stabilize the spread of residuals
  - Improve interpretability or predictive accuracy

When might we need a transformation?

- The residual plot shows a curved pattern $\rightarrow$ consider taking $\log(x)$, $\sqrt{x}$, or $1/x$
- The spread of residuals increases with x $\rightarrow$ consider transforming y with $\log(y)$ or $\sqrt{y}$
- The relationship is multiplicative or exponential $\rightarrow$ log-log or semi-log transformations can help

Common Transformations in Regression

Transformations help linearize relationships and stabilize variance.

Method	**Transform**	**Regression Equation**	**Predicted Value**
Linear	None	$y = a + bx$	$\hat{y} = a + bx$
Exponential	Take $\log(y)$	$\log(y) = a + bx$	$\hat{y} = 10^{a+bx}$
Square Root	Take $\sqrt{y}$	$\sqrt{y} = a + bx$	$\hat{y} = (a + bx)^2$

Note: Use ln and e^x if working with natural logarithms instead of common logs.

Transformations in Regression: What You Need to Know

You do NOT need to memorize specific transformation formulas for the AP Exam.

What you <u>do</u> need to know:

- Be able to recognize when a transformation might help, based on:
  - A curved pattern in the residual plot ($\rightarrow$ nonlinearity)
  - A fanning or shrinking spread in residuals ($\rightarrow$ changing variability)
- Know that transformations are used to:
  - Make a relationship more linear
  - Stabilize the variability of the residuals
- If given a transformed model like $\log(y) = a + bx$, you should:
  - Interpret a, b, and r^2 in context
  - Understand what $\hat{y}$ means after back-transforming

Introducing the `mtcars` Dataset

- **Dataset:** `mtcars` (Motor Trend Car Road Tests)
- **Source:** 1974 issue of *Motor Trend* magazine
- **Description:**
    - Contains data on **32 cars** from the 1973-74 model year
    - Variables include engine specs, fuel consumption, and performance

Key Variables:

- `mpg`: Miles per gallon (fuel efficiency)
- `hp`: Gross horsepower
- `wt`: Weight (1000 lbs)
- `qsec`: 1/4 mile time
- `cyl`: Number of cylinders
- `am`: Transmission (0 = automatic, 1 = manual)

Reference: Henderson, H. V. and Velleman, P. F. (1981). *Building multiple regression models interactively*. *Biometrics*, 37, 391-411.

R Documentation

Transforming Nonlinear Data: MPG vs. Horsepower

Raw Data (32 Cars):

Car	hp	mpg	Car	hp	mpg
Mazda RX4	110	21.0	Dodge Challenger	150	15.5
Mazda RX4 Wag	110	21.0	AMC Javelin	150	15.2
Datsun 710	93	22.8	Camaro Z28	245	13.3
Hornet 4 Drive	110	21.4	Pontiac Firebird	175	19.2
Hornet Sportabout	175	18.7	Fiat X1-9	66	27.3
Valiant	105	18.1	Porsche 914-2	91	26.0
Duster 360	245	14.3	Lotus Europa	113	30.4
Merc 240D	62	24.4	Ford Pantera L	264	15.8
Merc 230	95	22.8	Ferrari Dino	175	19.7
Merc 280	123	19.2	Maserati Bora	335	15.0
Merc 280C	123	17.8	Volvo 142E	109	21.4
Merc 450SE	180	16.4	Chrysler Imperial	230	14.7
Merc 450SL	180	17.3	Lincoln Continental	215	10.4
Merc 450SLC	180	15.2	Cadillac Fleetwood	205	10.4
Fiat 128	66	32.4	Toyota Corolla	65	33.9
Honda Civic	52	30.4	Toyota Corona	97	21.5

Observation: The plot of hp vs. mpg is nonlinear and decreasing.
Transformation Idea: Try `mpg` vs. `log(hp)` or `log(mpg)` vs. `log(hp)`
Which model has the best fit?

Linear Regression Shows Up in AP Science

Linear regression helps identify patterns, determine constants, and verify models across the sciences.

AP Biology

- Lineweaver-Burk plot: $\frac{1}{v}$ vs. $\frac{1}{[S]}$
- Population growth: $\ln(N)$ vs. t
- Photosynthesis rate: O_2 vs. time

AP Chemistry

- First-order kinetics: $\ln[A]$ vs. t
- Beer's Law: A vs. $[C]$
- Boyle's Law: P vs. $\frac{1}{V}$

AP Physics

- Hooke's Law: F vs. x
- Ohm's Law: V vs. I
- Kinematics: v vs. t
- Pendulum period: T^2 vs. L

Many of these applications require you to linearize non-linear data.

Integrated Rate Laws as Linear Models

Many chemical reactions can be modeled with linearized equations. Here's how different reaction orders relate to linear regression.

Zeroth Order

$$\frac{d[A]}{dt} = -k$$

$$\int d[A] = -k \int dt$$

$$[A] = -kt + [A]_0$$

First Order

$$\frac{d[A]}{dt} = -k[A]$$

$$\int \frac{1}{[A]} d[A] = -k \int dt$$

$$\ln[A] = -kt + \ln[A]_0$$

Second Order

$$\frac{d[A]}{dt} = -k[A]^2$$

$$\int \frac{1}{[A]^2} d[A] = -k \int dt$$

$$\frac{1}{[A]} = kt + \frac{1}{[A]_0}$$

These transformations allow rate laws to be analyzed using linear regression techniques.

Reaction 1: Decomposition of Nitrous Oxide (N_2O)

Objective: Determine the reaction order and calculate the rate constant using regression.

$$2N_2O\ (g) \xrightarrow{\text{Pt catalyst}} 2N_2\ (g) + O_2\ (g)$$

Time (s)	**$[N_2O]$ (mol/L)**
0	0.80
10	0.715
20	0.636
30	0.550
40	0.491
50	0.405
60	0.333
70	0.235
80	0.150
90	0.086

Instructions:

- Plot $[A]$, $\ln[A]$, and $1/[A]$ vs. time.
- Identify the most linear plot to determine the reaction order.
- Use linear regression to find the rate constant k from the slope.

Reaction 2: Decomposition of Hydrogen Peroxide (H_2O_2)

$$2H_2O_2\ (aq) \xrightarrow{\text{KI catalyst}} 2H_2O\ (l) + O_2\ (g)$$

Time (s)	$[H_2O_2]$ (mol/L)
0	1.00
20	0.815
40	0.665
60	0.531
80	0.441
100	0.352
120	0.285
140	0.230
160	0.185
180	0.145

Instructions:

- Plot $[A]$, $\ln[A]$, and $1/[A]$ vs. time.
- Identify which plot is linear to determine the reaction order.
- Use the slope to calculate k.

Reaction 3: Iodide and Peroxydisulfate Reaction

$$2I^- \text{ (aq)} + S_2O_8^{2-} \text{ (aq)} \rightarrow I_2 \text{ (aq)} + 2SO_4^{2-} \text{ (aq)}$$

Time (s)	**[I⁻] (mol/L)**
0	0.50
10	0.397
20	0.319
30	0.263
40	0.205
50	0.179
60	0.151
70	0.130
80	0.112
90	0.093

Instructions:

- Graph all three transformed plots: $[A]$, $\ln[A]$, $1/[A]$ vs. time.
- Determine the best fit and use regression to find k.