CORRELATION AND LINEAR REGRESSION

- ➢ Scatterplots
- ➢ Covariance and Correlation
- ➢ Linear Regression
- ➢ Residuals
- ➢ R-Squared
- ➢ Types of Observations

University of Pittsburgh

SCATTERPLOTS

- **Scatterplot:** graphical display of the relationship between two quantitative variables
  - **Response variable:** variable plotted along y-axis that we are trying to explain or predict
  - **Predictor variable:** variable plotted along x-axis that we are using to explain changes about the response variable
  - Observations plotted as ordered pairs

DESCRIBING A SCATTERPLOT

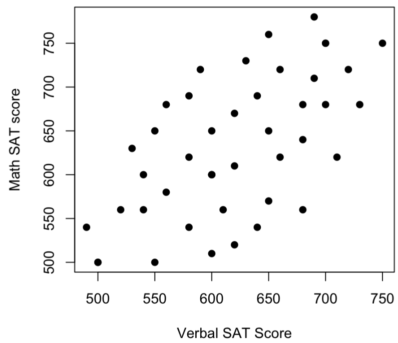
- **Direction:** As predictor variable increases…
  - **Positive:** response tends to increase
  - **Negative:** response tends to decrease
  - **Neither:** no obvious change in response

- **Form:** What is the general trend of the points?
  - **Linear:** response tends to increase at about the same rate across all values of predictor
  - **Curved:** rate at which response changes depends on value of predictor
  - **No pattern**

- **Strength:** How tightly clustered together are the points?
  - Usually described as **strong**, **moderate**, or **weak**

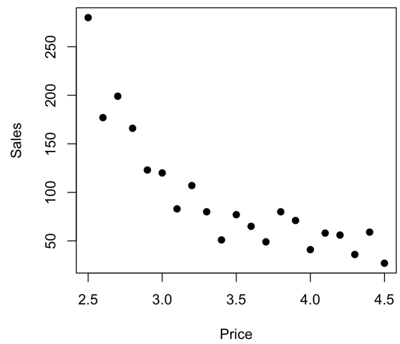
Example: Describing a Scatterplot

- **Scenario:** Want to know if verbal SAT scores yields any information about math SAT scores using a sample of 44 high school seniors
- **Question:** What are the predictor and response variables.
- **Answer:**
  - **Response:** _______ SAT score
  - **Predictor:** _______ SAT score
- **Task:** Describe the relationship.
- **Answer:** ________, _________, and _________
  - Math scores tend to ________ as verbal scores increase
  - Math scores change at _______________ ______ regardless of verbal score



Example: Describing a Scatterplot

- **Scenario:** Supermarket increases the price for a gallon a milk by 10 cents every day for 3 weeks and records number of units sold.
- **Task:** Describe the relationship.
- **Answer:** ________, _________, and ________
  - Sales tend to _______ as price increases
  - Not ____________________
    - Sales decline ______ when ______ prices are increased compared to when ______ prices are increased



Covariance

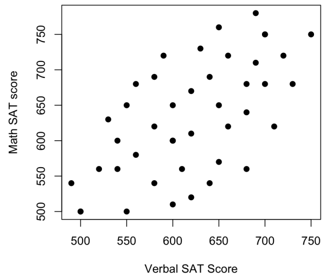
- **Covariance:** measure of the joint variability between two quantitative variables

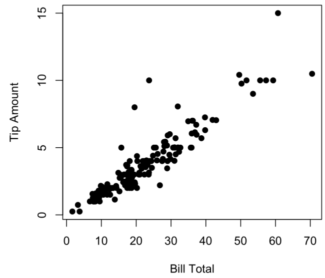
$$s_{XY} = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{n-1}$$

  - Sign of covariance dictates direction of relationship
  - Value is unbounded: Ranges from $-\infty$ to ∞
  - **Problem:** Does not help us interpret the _________ of relationship

Example: Comparing Relationships

- **Scenario:** Two scatterplots shown below:
  - **Left:** Use verbal SAT score to explain math SAT score
  - **Right:** Use restaurant bill to explain amount left as tip
- **Question:** Which scatterplot has the stronger linear relationship?



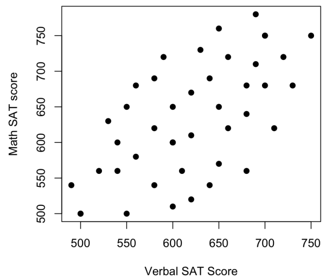


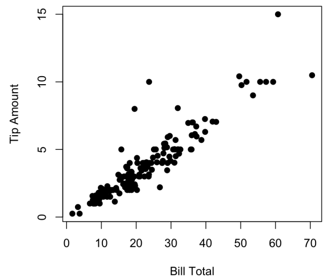
- **Answer:** ____________ vs. ____________________
  - Points are more _______________ around ____________

Example: Understanding Covariance

- **Scenario:** Two scatterplots shown below:
  - **Left:** Use verbal SAT score to explain math SAT score
  - **Right:** Use restaurant bill to explain amount left as tip
- **Question:** What does the covariance reveal?

$s_{XY} = 3331$





$s_{XY} = 26.93$

- **Answer:** _______________ by itself
  - Covariance will be _____ if the ___________ of the observations are large regardless of how _______ the linear relationship is

Correlation

- **Correlation:** measure of the strength and direction of the linear relationship between two quantitative variables
  - **Population correlation:** Denoted by ρ (Greek letter "rho")
    - Parameter → Generally unknown
  - **Sample correlation:** Denoted by r
    - Statistic → Good approximation of ρ

$$r = \frac{s_{XY}}{s_X s_Y}$$

s_{XY} ← Sample covariance

s_X ← Standard deviation of predictor values

s_Y ← Standard deviation of response values

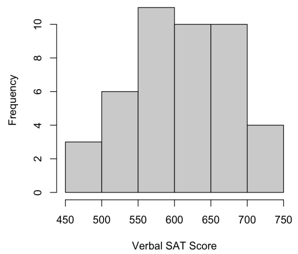
CORRELATION FACTS

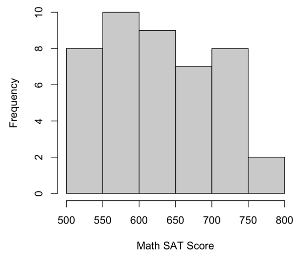
- Bounded between -1 and +1
  - Sign dictates direction of relationship (***positive*** or ***negative***)
  - Magnitude dictates strength of relationship
    - ***Rule of Thumb:*** *If the magnitude is…*
      - *Between 0.70 and 1.00, the strength of the relationship is* ***strong.***
      - *Between 0.40 and 0.70, the strength of the relationship is* ***moderate.***
      - *Between 0.10 and 0.40, the strength of the relationship is* ***weak.***
      - *Between 0.00 and 0.10, there is* ***little to no*** *relationship.*

- Scatterplot must have a **linear** form for the correlation to make sense.
  - As the scatterplot becomes more curved, the correlation becomes less accurate as a means of describing the relationship.

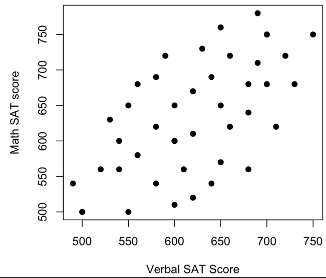
EXAMPLE: CALCULATING CORRELATION

- **Scenario:** Use verbal SAT score to describe math SAT score for a random sample of 44 students





Variable	Mean	Std. Dev.	Covariance
Verbal	619.32	67.35	3331
Math	631.59	79.94	



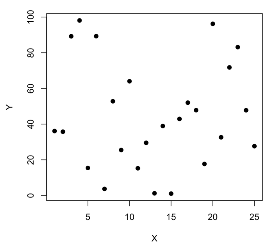
- **Question:** What is the correlation?

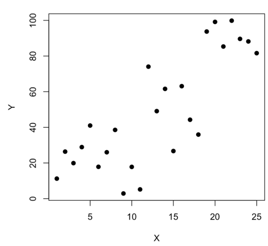
- **Answer:** $r =$ ________________ $=$ ________
  - Indicates ____________________ linear relationship

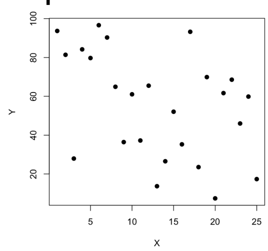
EXAMPLE: APPROXIMATING CORRELATIONS

- **Scenario:** Scatterplots show observations with predictor values from 1 to 25 and responses from 0 to 100.
- **Task:** Sort the scatterplots from weakest correlation to strongest.
- **Answer:** ___, ___, ___

- **Task:** Approximate the correlation in each scatterplot.







- **Answers:** $r =$ _____ $r =$ _____ $r =$ ______

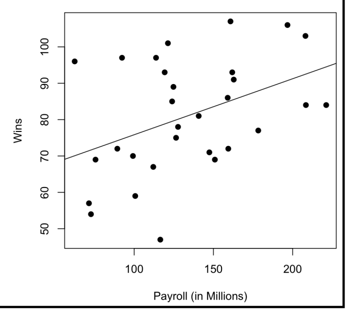
SIMPLE LINEAR REGRESSION LINE

- **Simple linear regression line:** the best fitting line between a single quantitative predictor (X) and a quantitative response (Y)

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Predicted Value of Response: "Y-hat" — Intercept — Slope — Value of predictor

- Describes how Y tends to change as X increases
- Allows us to make predictions about the response given some value of the predictor
- Predictions are not perfect, but provide good estimates



SIMPLE LINEAR REGRESSION LINE

- **Simple linear regression line:** the best fitting line between a single quantitative predictor (X) and a quantitative response (Y)

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Predicted Value of Response: "Y-hat" — Intercept — Slope — Value of predictor

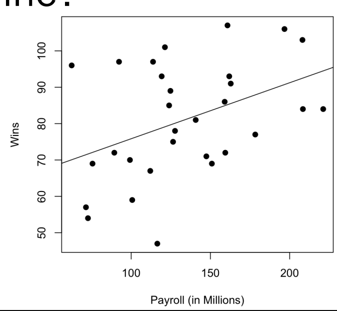
- **Slope:** $\hat{\beta}_1 = r\frac{s_Y}{s_X}$
  - Expected increase in prediction given a one unit increase in the predictor
- **Intercept:** $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$
  - Predicted value of response variable when predictor equals 0

EXAMPLE: CALCULATING REGRESSION LINE EQUATION

- **Scenario:** Use payroll (in millions) to predict wins in Major League Baseball in 2019

Variable	Mean	Std. Dev.	Correlation
Payroll	133.47	42.66	0.4135
Wins	81	15.87	

- **Question:** What is the equation of the regression line?
- **Answer:**
  - **Slope:** $\hat{\beta}_1 =$ ________________ = _______
  - **Intercept:** $\hat{\beta}_0 =$ ____________________ = ______
  - **Regression Line:** ____________________



Example: Interpreting Regression Equation

- **Scenario:** Use payroll (in millions) to predict wins in Major League Baseball in 2019

$$\hat{Y} = 60.47 + 0.1538X$$

- **Question:** What are the interpretations of the slope and intercept?
- **Answer:**
  - **Intercept:** A team with a ____________ would be _______________________ games.
  - **Slope:** For every additional _______________ a team spends on its payroll, their predicted win total _____________________.

Example: Making Predictions

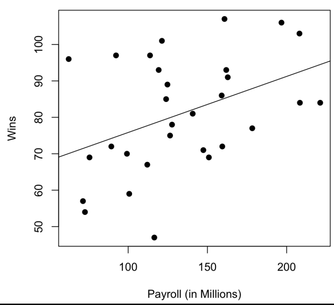
- **Scenario:** Use payroll (in millions) to predict wins in Major League Baseball in 2019

$$\hat{Y} = 60.47 + 0.1538X$$

- **Question:** How many games would we expect a team with a payroll of $100 million to win?
- **Answer:** $\hat{Y} =$ ______________________

$=$ ______________________

$=$ ________



Residuals

- **Residual:** the difference between the predicted value of a response and the observed value from the data

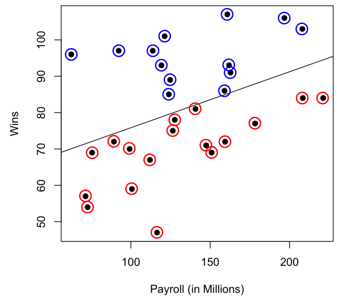
$$e = Y - \hat{Y}$$

Residual (Error) — Observed Value From Data — Predicted Value From Regression Line

- Observations that are:
  - **Close** to the regression line will have ______ residuals
  - **Far away** from the regression line will have ______ residuals
  - **Above** the regression line will have _________ residuals
  - **Below** the regression line will have _________ residuals

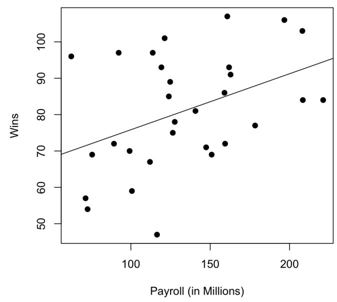
EXAMPLE: RESIDUALS

- **Scenario:** Use payroll (in millions) to predict wins in Major League Baseball in 2019
- **Question:** What can be said about the blue circled observations?
- **Answer:** All have __________ residuals
  - Actual win totals were _______ than their _________ win total
  - ___________________ for the season based on payroll alone
- **Question:** What can be said about the red circled observations?
- **Answer:** All have __________ residuals
  - Actual win totals were _____ than predicted win total



EXAMPLE: RESIDUALS

- **Scenario:** Kansas City Royals had a $100 million payroll in 2019 and won 59 games. (Recall the predicted value was 75.85.)
- **Question:** What was their residual?
- **Answer:** $e =$ _____________ $=$ ______
  - Observation lies ______ the regression line
- **Question:** What does this value mean in context?
- **Answer:** Kansas City _______________ during the season, winning ___________________ than would have been expected given its _______________ payroll.



VARIATION IN THE LINEAR MODEL

- **R-squared:** the percentage of the variability in the response (Y) that is explained by the predictor (X)
  - Calculated by squaring the correlation: r^2
    - Observations in straight line if $r^2 = 1$ → Perfect model
    - No relationship between variables if $r^2 = 0$ → Useless model
  - Higher R-squared indicates a stronger relationship between X and Y
  - Remainder of variation $(1 - r^2)$ is due to natural fluctuations of the response in the sample and comes from the residuals.

Example: R-Squared

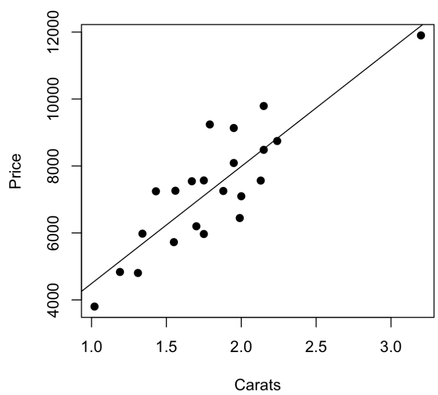
- **Scenario:** Use payroll (in millions) to predict wins in Major League Baseball in 2019. Correlation is 0.4135.
- **Question:** What is the R-squared value?
- **Answer:** $r^2 =$ ________ = ______
- **Question:** What does the R-squared mean in context?
- **Answer:** ________ of the ____________________________ is explained by their ________.
  - Remaining ______ is due to __________________ in the sample, such as:
    - Talented younger, _____________ players
    - Aging, ________ veterans who underperform
    - Randomness within ________________

Types of Points in Regression

- **Outlier:** an observation with an unusually large residual
  - Extreme values of the response, but can take on any predictor value
- **High leverage observation:** an observation where the value of the predictor is unusually far away from the rest of the data
  - Extreme values of the predictor, but can take on any response
  - Have the *potential* to drastically impact the regression line
- **Influential point:** an observation that strongly influences the slope and/or direction of the regression line
  - Extreme values of both the predictor and response

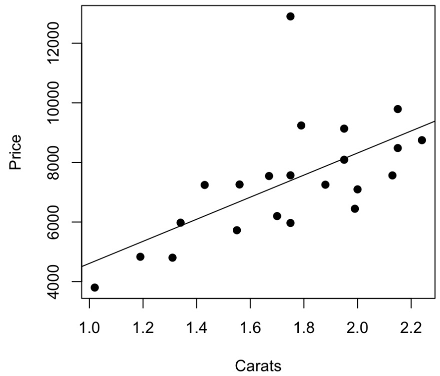
Types of Points in Regression

- **Scenario:** Random sample of 22 clear diamonds, one of which was 3.16 carats and sold for $11,900.
- **Question:** How should we classify the diamond with the weight of 3.16 carats?
- **Answer:** ________________________
  - **Outlier?:** ____
    - Residual is ______
  - **High Leverage?:** ____
    - Predictor value is ________ from the rest of data
  - **Influential Point?:** ____
    - Observation follows _____ of data; removal would __________ regression line by much



Types of Points in Regression

- **Scenario:** Random sample of 21 clear diamonds and one blue diamond that was 1.75 carats but sold for $12,900.
- **Question:** How should we classify the blue diamond?
- **Answer:** ________
  - **Outlier?:** ____
    - ________________ residual
  - **High Leverage?:** ____
    - ________ value of the predictor
  - **Influential Point?:** ____
    - Slope of regression line not _________ impacted – got ___________ slightly



Types of Points in Regression

- **Scenario:** Random sample of 21 clear diamonds and one black diamond that was 2.90 carats but sold for $3,200.
- **Question:** How should we classify the black diamond?
- **Answer:** _____________
  - **Outlier?:** ____
    - _______________ residual
  - **High Leverage?:** ____
    - _______________ predictor value
  - **Influential Point?:** ____
    - Drastically __________ regression line, creating a slope that is closer to ___________