1 Exploratory Data Analysis and Summary Statistics

1.1 Populations and samples

1.1.1 Definitions

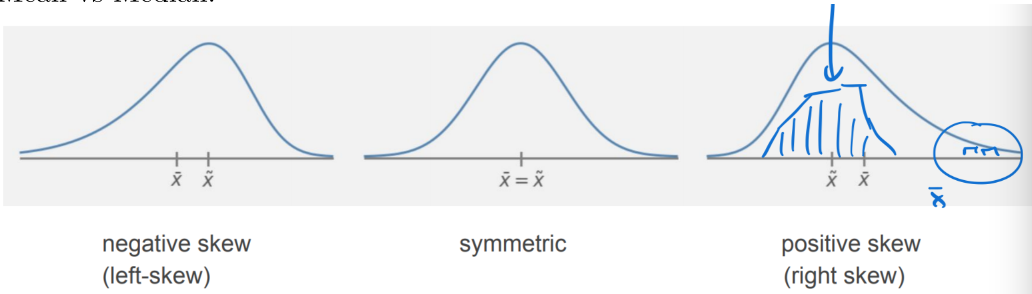
- Population: collection of units
- Sample: subset of population
- Characteristic/Variable of Interest (VOI): something we want to measure for each unit
- Sample frame: source material or device from which sample is drawn
- Sample types:
  - Simple random sample: randomly select people from sample frame
  - Systematic sample: order the sample frame. Choose integer k. Sample every kth unit in the sample frame.
  - Census sample: sample literally everyone/everything in the population
  - Stratified sample: if you have heterogeneous population that can be broken up into homogeneous groups, randomly sample from each group proportionate to their prevalence in the population

1.1.2 Exploratory data analysis (EDA)

[1] Numerical summaries

- Definition: calculation and interpretation of certain summarizing numbers (called: sample statistics)
- Measures of centrality: summarizing the center "central tendency"
- Sample mean (Arithmetic average): for set of numbers x_1, x_2, ..., x_n, sample mean is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Sample median: the middle value when the observations are ordered from smallest to largest
  - if n is odd, then $\tilde{x} = (\frac{n+1}{2})^{th}$ ordered valued
  - if n is even, then $\tilde{x} =$ average of $(\frac{n}{2})^{th}$ and $(\frac{n}{2}+1)^{th}$ ordered values
- Sample mode: the value that occurs the most often in the sample

- Mean vs Median:



- Quartiles: Divide the data into 4 equal parts
  - Lower quartile (Q_1 or P_{25}): splits the lowest 25% of the data from the other 75%
  - Middle quartile (Q_2 or P_{50}): splits the data in half
  - Upper quartile (Q_3 or P_{75}): splits the highest 25% of the data from the lowest 75%
  - Computation:
    1. Use the median to divide the ordered data set into 2 halves: A. if n is odd, include the median in both halves, B. if n is even, split the data exactly in half
    2. The lower quartile is the median of the lower half
    3. The upper quartile is the median of the upper half

- Percentiles: same as quartiles, but can calculate general percentiles, such as (P_{16}) splits off the lower 16% of the data

- Variability (spread)
  - Range: the difference between max and min values
  - Sample variance: $s^2 = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2$
  - Sample standard deviation: $s = \sqrt{s^2}$

- Interquartile range: difference between the upper and lower quartiles, $IQR = Q_3 - Q_1$

[2] Graphical summaries

- Histogram: graphical representation of distribution of numerical data
  - Frequency histogram: count the number of data values that fall into a bin and draw a rectangle over that bin withe height equal to the count.
  - Density histogram: count the number of data values that fall into a bin and adjust the height such that the sum of the area of all bins is equal to 1

- Freedman-Diaconis Rule: bin size $= 2\frac{IQR}{n^{1/3}} = 2\frac{Q_3 - Q_1}{n^{1/3}}$

2 Introduction to Probability

2.1 Basics

- Probability: it is a way of thinking about unpredictable phenomenon as if they were each generated from some random process.
- Sample space Ω: set of all possible outcomes of the experiment.
- For each event in the probability is a measure between 0 and 1 of how likely it is for the event to occur.
- Intersection (and) $\cap$: the subset of outcomes in both events.
- Union (or) $\cup$: the subset of outcomes in one or both events.
- Complement (i.e. A^c): set of outcome in Ω but not in a certain event, let's say A.
- Disjoint or mutually exclusive: when the intersection of two events is empty.
- A is a subset of B, if all outcomes of event A are also outcomes of even B.
- DeMorgan's Laws:
  - Complement of union: $(A \cup B)^c = A^c \cap B^c$
  - Complement of intersection: $(A \cap B)^c = A^c \cup B^c$

2.2 Probability functions

- Two key properties:
  - The probability of the entire sample space is 1
  - The probability of the union of disjoint events is the sum of the probability of each event.
- Probability of disjoint (independent) events: $P(A \cup B) = P(A) + P(B)$
- Probability of non-disjoint (dependent) events: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Probability of the complement: $P(A^c) = 1 - P(A)$
- Independent events probabilities:
  - Multiply the probabilities of two or more independent events.
  - OR means addition
- Bernoulli Trial: a random process with two outcomes with fixed probabilities assigned to each outcome.

3 Probability Theory

3.1 Conditional probability

- The conditional probability of A given C is defined by:
- $P(A|C) = \frac{P(A \cap C)}{P(C)}$ provided that $P(C) > 0$

3.2 Product rule of probability

- $P(A \cap C) = P(A|C)P(C)$

3.3 Independent events - rules

- An event is said to be independent of event B if $P(A|B) = P(A)$
- Combining the definition with product rule and conditional probability:
- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(A \cap B) = P(A)P(B)$

3.4 Subtleties of independence

- Events $A_1, A_2, ..., A_m$ are independent if:
- $P(A_1 \cap A_2 \cap ... \cap A_m) = P(A_1)P(A_2)...P(A_m)$

3.5 Law of total probability (LTP)

- Suppose $C_1, C_2, ..., C_m$ are disjoint event such that $C_1 \cup C_2 \cup ... C_m = \Omega$. Then the probability of an arbitrary event A can be expressed as:
- $P(A) = P(A|C_1)P(C_1) + P(A|C_2)P(C_2) + ... + P(A|C_m)P(C_m)$

3.6 Bayes' Theorem

- Combining conditional probability and product rule.
- $P(A|C) = \frac{P(C|A)P(A)}{P(C)}$
  - $P(A|C)$ = posterior distribution
  - $P(A)$ = prior distribution

  - $P(C|A)$ = likelihood function
  - $P(C)$ = evidence

- Bayes' rule + LTP: $P(A|C) = \frac{P(C|A)P(A)}{P(C)} = \frac{P(C|A)P(A)}{P(C|A)(P(A)+P(C|A^c)P(A^c)}$

4 Random variables

4.1 Discrete random variables

- Definition: a function that maps elements of the sample space Ω to a finite number of values $a_1, a_2, ..., a_n$ or an infinite number of values $a_1, a_2, ...$
- Examples:
  - Sum of dice, difference of the dice, maximum of the dice, ...
  - Number of coin flips until we get heads, number of heads in n flips, ...

4.2 Probability mass function (PMF)

- Definition: the map between the random variable's values and the probability of those values.
- $f(a) = P(X = a)$
- Called a "probability `mass` function" because each of the random variables' values has some probability mass (or weight) associated with it.
- Sum of masses: $\sum_{i=1}^{n} f(a_i) = 1$ because it is a probability function.

4.3 Cumulative distribution function (CDF)

- Definition: a function whose value at a point a is the cumulative sum of probability masses up until a.
- $F(a) = P(X \leq a)$

4.4 Relationship between the PMF and the CDF

- $F(a) = \sum_{x \leq a} f(x)$

5 Counting

5.1 Permutations

- Definition: Counting the number of ways that a set of objects can be ordered(or permuted)
- r-permutations of n objects: $P(n, r) = \frac{n!}{(n-r)!}$

5.2 Combinations

- Definition: Counting the number of ways that a set of objects can be combined into subsets.
- Key difference: When counting combinations, order does not matter.
- r-combinations of n objects: $C(n, r) = C_{n,r} = \binom{n}{r} = \frac{n!}{(n-r)!r!}$

6 Discrete random variables and their distributions

6.1 The Bernoulli distribution

- Definition: A discrete random variable X has a Bernoulli distribution with parameter p, where $0 \le p \le 1$, if its probability mass function is given by: (we denote this distribution by Ber(p))
  $f(1) = p_x(1) = P(X = 1) = p$ and $p_x(0) = P(X = 0) = 1 - p$
- It is used to model experiments with only two possible outcomes.
- If we have $p_x(1) = p$ and $p_x(0) = 1 - p$, then for x in {0, 1}, we have: $p_x(x) = p^x(1 - p)^{1-x}$

6.2 Binomial distribution

- A discrete random variable X has a binomial distribution with parameters n and p, where $n = 1, 2, ...$ and $0 \le p \le 1$, if its probability mass function is given by:
  $p_X(k) = P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$ for $k = 0, 1, 2, ..., n$

6.3 Bernoulli vs. Binomial

- Assumptions in going from Ber(p) to Bin(n, p)
  - Each of n Bernoulli trials are independent.
  - Each of the Bernoulli trials has the same probability of success p.
- Bernoulli distribution – a coin flip // success or failure
- Binomial distribution – how many successes out of n Bernoulli trials

6.4 Discrete uniform distribution

- Definition: A discrete random variable X has a discrete uniform distribution with parameters a, b and $n = b - a + 1$ if:
  $p_X(k) = \frac{1}{n}$ for $k = a, a + 1, a + 2, ..., b$

6.5 Binomial-like distributions

6.5.1 Geometric distribution

- A discrete random variable X has a geometric distribution with parameter p, where $0 \le p \le 1$, if its probability mass function is given by:
  - $f(k) = p_x(k) = P(X = k) = (1 - p)^{k-1}p$, for $k = 1, 2, 3, ...$
  - We say that $X \sim Geo(p)$

- Assumptions:
  - Each trail is independent and "identically distributed"
  - Each trail is a Bernoulli r.v. with probability of success p

6.5.2 Negative Binomial distribution

- A discrete r.v. X has a negative binomial distribution with parameters r and p, where $r > 1$ and $0 \leq p \leq 1$, if its probability mass function is given by:
  - $p_X(k) = P(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$
  - p = probability of success for each trial
  - r = number of successes we want to observe
  - X = number of trials needed before we observe r successes (r.v.)
  - We say that X $\sim$ NB(r, p)
- Assumptions:
  - Each trial a a Bernoulli r.v. with probability of success p
  - Each trial is independent

6.5.3 Poisson distribution

- A discrete r.v. X has a Poisson distribution with parameter λ, where $\lambda > 0$, if its probability mass function is given by:
  - $p_X(k) = P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$, for $k = 0, 1, 2, ...$
  - k = number of successes in the interval Δt
  - λ = average number of success in the interval of Δt
  - We say that X $\sim$ Pois(λ)
- Assumptions:
  - Probability of observing a single event over a small interval is proportional to the size of the interval
  - Each event/interval is independent

7 Continuous Random Variables and Their Distribution

7.1 Continuous Random Variables

- Definition:
  - A random variable X is continuous if for some function $f : R \to R$ and for any numbers a and b with $a \leq b$, $P(a \leq X \leq b) = \int_a^b f(x)dx$
- The function f must satisfy (to be proper pdf, which is pmf but continuous):
  - $f(x) \geq 0$ for all x (nonnegative)
  - $\int_{-\infty}^{\infty} f(x)dx = 1$ (normalized)

7.2 Uniform distribution

- Definition:
  - A continuous random variable has a uniform distribution on the interval $[\alpha, \beta]$ if its probability density function f is given by $f(x) = 0$ if x is not in $[\alpha, \beta]$ and
  - $f(x) = \frac{1}{\alpha-\beta}$ for $\alpha \leq x \leq \beta$
  - We say $X \sim U(\alpha, \beta)$

7.3 Cumulative distribution

- Cumulative distribution function: $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$
- $P(a \leq x \leq b) = \int_a^b f(t)dt = F(b) - F(a)$
- $\frac{d}{dx}F(x) = f(x)$ `This is an important and useful relationship`

7.4 The normal distribution

- Used for: e.g. location, scale
- Definition:
  - A continuous random variable X has a normal (or Gaussian) distribution with parameters μ and σ^2 if its probability density function is given by:
  - $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$

7.5 The exponential distribution

- Used for: e.g. interarrival times

- Sometime it's easier to first find the cdf and then derive the pdf by taking a derivative.
- Definition:
  - A continuous random variable X has an exponential distribution with rate parameter $\lambda > 0$ if its probability density function is given by
  - $f(x) = \lambda e^{-\lambda x}$ for $x > 0$, $f(x) = 0$ for $x < 0$
  - Theorem: (memoryless property)
    * If $T \sim Exp(\lambda)$, then $P(T > t + t_0 | T > t_0) = P(T > t)$

8 Expectation of Discrete and Continuous Random Variables

8.1 Expected value: discrete random variables

- Definition:
  - The expectation or expected value of a discrete random variable X that takes the values $a_1, a_2, ...$ and with pmf p is given by:
  - $E[X] = \sum_i a_i P(X = a_i) = \sum a_i p(a_i)$
  - $\sum$: Sum over all possible values for r.v. X
  - a_i: Possible outcome
  - X: Probability mass (or weight) associated with that outcome
- Intuition: Think of masses of weight $p(a_i)$ placed at the points $a_i \rightarrow E[X]$ is the balancing point.

8.2 Expected value: continuous random variables

- Definition:
  - The expectation, expected value, or mean, of a continuous random variable X with probability density function f is:
  - $E[X] = \int_{-\infty}^{\infty} x f(x) dx$
- Intuition: Think of a single rock balancing on t a fulcrum.

8.3 Change-of-variables formula

- Let X be a random variable and let g: $R \rightarrow R$ be a function
  - If X is a discrete and takes the values $a_1, a_2, ...$ then:
    * $E[g(x)] = \sum_i g(a_i) P(X = a_i)$
  - If X is continuous, with probability density function (pdf) f, then:
    * $E[g(x)] = \int_{-\infty}^{\infty} g(x) f(x) dx$

8.4 Linearity of expectation

- Expectation is a linear function
- $E[aX + b] = aE[x] + b$

9 Variance of Discrete and Continuous Random Variables

9.1 Definition

- The variance Var(X) of a random variables X is the number: $Var(X) = E[(X - E[X])^2] = E[(X - \bar{X})^2] = E[X^2] - E[X]^2$
- The standard deviation of random variable X is the square root of the variance: $SD(X) = \sqrt{Var(X)}$
- How to compute?
  - First, compute E(X)
  - Then, use the definition of Variance and change-of-variables formula (w/ $g(x) = (x - E[X])^2$) to get Var(X):
    * $Var(X) = \sum_i (a_i - E[X])^2 p(a_i)$ or $Var(X) = \int_{-\infty}^{\infty} (x - E[X])^2 f(x)dx$

9.2 Quick notes

- If X and Y are independent, then $Var(X + Y) = Var(X) + Var(Y)$
- If $X \sim Ber(p)$, then:
  - $E[X] = p$
  - $Var(X) = p(1 - p)$
- If $X \sim Bin(n, p)$, then:
  - $E[X] = np$
  - $Var(X) = np(1 - p)$
- If $X \sim U[\alpha, \beta]$, then:
  - $E[X] = \frac{1}{2}(\alpha + \beta)$
  - $Var(X) = \frac{1}{12}(\beta - \alpha)^2$
- Variance is not linear: $Var(aX + b) = a^2 Var(X)$

10 The Normal Distribution

10.1 Definition

- A continuous random variable X has a normal (or Gaussian) distribution with parameters μ(mean) and σ^2(variance) if its probability density function is given by:
- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$
- We say $X \sim N(\mu, \sigma^2)$

10.2 The standard normal distribution

- Definition
  - The normal distribution with parameter values $\mu = 0$ and $\sigma^2 = 1$. $(Z \sim N(0,1))$
- PDF
  - $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$
- CDF
  - $\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} f(x)dx$ (we usually look up values for $\Phi(z)$ in a table)
- The critical value
  - We say z_α is the critical value of Z under the standard normal distribution that gives a certain tail area. In particular, it is the Z values such that exactly $\alpha/2$ of the area under the curve lies to the right of z_α.
  - Relationship between z_α and the CDF
    - $\Phi(z_\alpha) = 1 - \alpha$
  - Relationship between z_α and percentiles
    - z_α is the $100(1-\alpha)^{th}$ percentile

10.3 Non-standard normal distributions

- Proposition
  - If X is a normally distributed random variable with mean μ and standard deviation σ, then Z follows a standard normal distribution if we define: (Box-Muller transformations)
    - $Z = \frac{X-\mu}{\sigma}$
    - $X = \sigma Z + \mu$

11 The Central Limit Theorem

11.1 Random samples

- The random variable $X_1, X_2, ..., X_n$ are said to form a (sample) random sample of size n if:
  - All X_k's are independent
  - All X_k's come from the same distribution
- We say these X_k's are independent and identically distributed $\rightarrow$ iid

11.2 Estimator and their distribution

- We use estimators to summarize our iid sample
- Examples:
  - $\bar{x}$ is the sample mean estimator of the population mean μ
  - $\hat{p}$ is the sample proportion (#in sample satisfying some characteristic of interest / total #)
  - s^2 is the sample estimator for σ^2(unknown true population variance)
- Fun fact: any estimator, including the sample mean $\bar{(}X)$, is a random variable since it is based on a random sample.
- This means that $\bar{X}$ has a distribution of its own, which is referred to as the sampling distribution of the sample mean
- The sampling distribution depend on:
  - Population distribution
  - Sample size n
  - Method of sampling
- Distribution of the sample mean
  - Proposition
    * Let $X_1, X_2, ..., X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, then for any n
    * $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
  - We know everything about the distribution of the sample mean when the population distribution is normal
  - $E[\bar{X}] = E[\frac{1}{n}\sum_{k=1}^{n} X_k] = \frac{1}{n}\sum_{k=1}^{n} E[X_k] = \frac{1}{n}\sum_{k=1}^{n} \mu = \mu$
  - $Var(\bar{X}) = \frac{1}{n^2}Var(\sum_{k=1}^{n} X_k) = \frac{1}{n^2}\sum_{k=1}^{n} Var(X_k) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$

11.3 The Central Limit Theorem

- Important note
  - When the population distribution is non-normal, averaging produces a distribution more normal (bell-shaped) than the one being sampled.
  - A reasonable assumption is that if n is large, a suitable normal curve will approximate well the actual distribution of the sample mean.
- Definition
  - Let $X_1, X_2, ..., X_n$ be iid draws from some distribution. Then, as n becomes large:
  - $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

12 Intro to Statistical Inference and Confidence Intervals

12.1 Goal

- Want to extract properties of an underlying population by analyzing sampled data

12.2 Confidence intervals

- The CLT tell us that as the sample size n increases, the sample mean of X is close to normally distributed with expected value μ and standard deviation $\sigma/\sqrt{n}$
  - $\bar{X} \sim N(\mu, \sigma^2/n)$
- Standardizing the sample mean by first subtracting the expected value and dividing by the standard deviation yeilds a standard normal random variable
  - $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
- How large does the sample need to be if:
  - the population is normally distributed? $n \geq 1$
  - the population is not normally distributed? $n \geq 30$ (roughly)
- The CI varies from sample to sample, so the CI is really a random interval itself.

12.3 Interpreting the confidence level

- In repeated sampling, 95% of all CIs obtained from sampling will actually contain the true population mean. The other 5% of CIs will not.
- The confidence level is not a statement about any one particular interval. Instead, it describes what would happen if a very large number of CIs were computed using the same CI formula.
- A $100 * (1-a)\%$ confidence interval for the mean μ when the value of a is known is given by:
  - $[\bar{x} - z_{a/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{a/2}\frac{\sigma}{\sqrt{n}}]$
- Wide vs. narrow confidence interval:
  - Wider CI: will be successful more often (capture μ)
  - Narrower CI: information could be more actionable

12.4 Confidence intervals for proportions

- The estimator for p is given by $\hat{p} = \frac{X}{n}$
- The estimator is approximately normally distributed with:

- $E[\hat{p}] = p$
- $Var(\hat{p}) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}$
- Standardizing the estimate yields:
  * $Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$
- This gives us a $100 \cdot (1-a)\%$ confidence interval of:
  * $\hat{p} \pm z_{a/2}\sqrt{\frac{p(1-p)}{n}}$

12.5 Difference between population means

- How do two sub-populations compare? Are their means the same?
  - Solution process: collect samples from both sub-pops, and perform inference on both samples to make conclusions about $\mu_1 - \mu_2$
  - Basic assumptions:
    * $X_1, X_2, ..., X_m$ is a random sample from distribution 1 with mean μ_1 and SD σ_1
    * $Y_1, Y_2, ..., Y_n$ is a random sample from distribution 2 with mean μ_2 and SD σ_2
    * The X and Y sample are independent of one another
  - The natural estimator of $\mu_1 - \mu_2$ is the difference in sample means $\bar{x} - \bar{y}$
  - The expected value of $\bar{X} - \bar{Y}$ is given by:
    * $E[\bar{X} - \bar{Y}] = E[\bar{X}] - E[\bar{Y}] = \mu_1 - \mu_2$
  - The SD of $\bar{X} - \bar{Y}$ is given by:
    * $SD = \sqrt{Var(\bar{X} - \bar{Y})} = \sqrt{Var(\bar{X}) + Var(\bar{Y})} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$
- Normal populations with known SDs
  - The difference in sample means is normally distributed, for any sample sizes, with:
  - $\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n})$
- Confidence intervals for the difference
  - Standardized $\bar{X} - \bar{Y}$ gives a standard random variable, so we can compute a $100 \cdot (1-a)\%$ confidence interval for μ_1, μ_2
- Large-sample CIs for the difference
  - If both m and n are large, then the CLT kicks in and our confidence interval for the difference of means is valid, even if the populations are not normally distributed.
  - Furthermore, if m and n are large, and we don't know the SDs, we can replace them with the sample standard deviations
- Difference between population proportions
  - Let $\hat{p}_1 = \frac{X}{m}, \hat{p}_2 = \frac{Y}{n}$, where $X \sim Bin(m, p_1)$ and $Y \sim Bin(n, p_2)$

- Assuming that X and Y are independent, we can show that the expected valued and SD are estimated by:
  * $E[\hat{p_1} - \hat{p_2}] = E[\hat{p_1}] - E[\hat{p_2}] = \frac{1}{m}E[X] - \frac{1}{n}E[Y] = \frac{1}{m}mp_1 - \frac{1}{n}np_2 = p_1 - p_2$
- $Var(\hat{p_1} - \hat{p_2}) = \frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}$
- $SD(\hat{p_1} - \hat{p_2}) = \sqrt{\frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}}$
- The $100 \cdot (1-a)\%$ confidence interval for $p_1 - p_2$ is then given by:
  * $(\hat{p_1} - \hat{p_2}) \pm z_{a/2}\sqrt{\frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}}$

13 Intro to Hypothesis Testing

13.1 Statistical hypothesis

- Definition: A claim about the values of a parameter of a population characteristic.

13.2 Null vs alternative hypotheses

- In any hypothesis-testing problem, there are always two competing hypotheses under consideration:
  - Null hypothesis: H_0 - default, what we believe before collecting data
  - Alternative hypothesis: H_1 or H_A - what we want to find evidence for
- The objective of hypothesis testing is to choose, based on sampled data, between two competing hypotheses about the value of a population parameter.

13.3 Test statistics and evidence

- Definition: A test statistic is a quantity derived from the sample data and calculated assuming that the null hypothesis is true. It is used in the decision about whether or not to reject the null hypothesis
- Intuition:
  - We can think of the test statistics as our evidence about the competing hypotheses
  - We consider the test statistic under the assumption that the null hypothesis is true by asking question like:
    * How likely would we be to obtain this evidence if the null hypothesis were true?

13.4 Rejection regions and significance level

- The rejection region is the range of values of the test statistic that would lead you to reject the null hypothesis
- The significance level α indicates the largest probability of the tests statistic occurring under the null hypothesis that would lead you to reject the null hypothesis.

13.5 Error in hypothesis testing

- A type 1 error occurs when the null hypothesis is incorrectly rejected (it was, in fact, true) $\rightarrow$ false positive
  - Probability of committing type 1 error = the significance level α

- A type 2 error occurs when the null hypothesis is incorrectly not rejected (it was false) $\rightarrow$ false negative

13.6 p-values

- Definition:
  - A p-value is the probability, under the null hypothesis, that we would get a test statistic at least as extreme as the one we calculated
  - For a lower-tailed test with test statistic x, the p-values is equal to $P(X \leq x|H_0)$
- Intuition: the p-value assesses the extremeness of the test statistic. The smaller the p-value, the more evidence we have against the null hypothesis
- Important notes - The p-values is:
  - calculated under the assumption that the null hypothesis is true
  - always a value between 0 and 1
  - `not` the probability that the null hypothesis is true
- The decision rule is:
  - if p-value $\leq \alpha$, then reject the null hypothesis
  - if p-value $> \alpha$, then fail to reject the null hypothesis
  - $H_1 : \theta > \theta_0 \rightarrow 1 - \Phi(z) \leq \alpha$
  - $H_1 : \theta < \theta_0 \rightarrow \Phi(z) \leq \alpha$
  - $H_1 : \theta \neq \theta_0 \rightarrow 2\Phi(-|z|) \leq \alpha$
- Note: the p-value can be thought of as the smallest significance level at which the null hypothesis can be rejected
- Two-sample testing for difference of means:
  - test statistic = $z = \frac{(\mu_1 - \mu_2) - c}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0,1)$
- p-value = 1 - $\Phi(z)$

14 Statistical Inference with Small Samples

14.1 Summary:

- Statistical inference for population mean when data are normal and n is large:
  - σ is known: $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$
  - σ is unknown: $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$
- Statistical inference for population mean when data are NOT normal and n is large:
  - σ is known: $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$ (by CLT)
  - σ is unknown: $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} \sim N(0,1)$ (by CLT)
- Statistical inference when data are normal and n is small:
  - σ is known: $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$
  - σ is unknown: ??? (will use t-distribution instead of z-distribution)

	n ≥ 30	n < 30
Normal data, known σ ✓		
Normal data, unknown σ		
Non-normal data, known σ		
Non-normal data, unknown σ		

z-distribution (past) t-distribution (today!) Bootstrap (later)

6

14.2 Small-sample tests for μ

- When n is small, we can't invoke the Central Limit Theorem
  - If we don't even know if the data are normal, then we can bootstrap
  - But that can be expensive (producing lots of replicates takes time and memory)
- If we have small n and some reason to think our data are (approximately) normal, then:
  - When $\bar{X}$ is the sample mean of a random sample of size n from a normal distribution with mean μ, then random variable:
    - $t = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} \sim t_v$
  - Follows a probability distribution called t-distribution with parameter $v = n-1$ degrees of freedom (df).

14.3 Properties of t-distributions

- Let t_v denote the t-distribution with parameter $v = n - 1$ df
- Each t_v curve is bell-shaped and centered at 0
- Each t_v curve is more spread out than the standard normal distribution
- As v increases, the spread of the corresponding t_v curve decreases
- As $v \to 0$ the sequence of t_v curves approaches the standard normal curve

14.4 The t-critical value

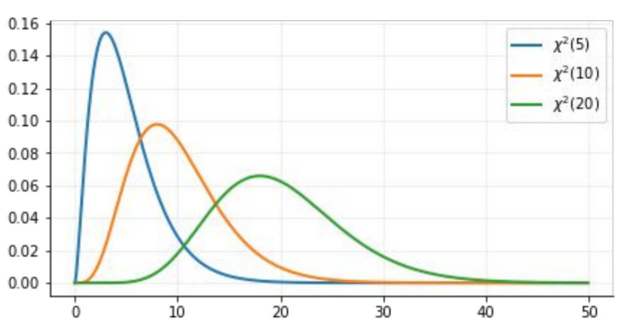
- We can extend all of our inferential mechanics to small-sample case by introducing the so-call t-critical value, which we denote as $t_{a,v}$
- Definition: the t-critical value, $t_{a,v}$, is the point such that are under the t_v-curve to the right of $t_{a,v}$ is equal to α (stats.t.ppf($1 - \alpha$, $v = n - 1$))

14.5 The t-confidence interval for the mean

- Let $\bar{X}$ and s be the sample mean and sample standard deviation computed from a random sample of size n, from a normal population with mean μ
- Then, a $100 \cdot (1 - \alpha)\%$ t-confidence interval for the mean μ is given by:
  - $[\bar{X} - t_{\alpha/2,n-1} \cdot \frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2,n-1} \cdot \frac{S}{\sqrt{n}}]$

14.6 The chi-squared distribution (χ^2)

- The chi-squared distribution is also parameterized by degrees of freedom $v = n - 1$
- The pdfs of the family χ^2_v are pretty nasty. Here is a plot of a few.



14.7 A confidence interval for the variance

- Let $X_1, X_2, ..., X_n$ be a random sample from a normal distribution with mean μ and SD σ.
- The random variable $\frac{(n-1)S^2}{\sigma^2}$ follows the distribution χ^2_{n-1}
- Then it follows that $P(\chi^2_{1-\alpha/2,n-1} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2,n-1}) = 1 - \alpha$
- For a $100 \cdot (1-\alpha)\%$ CI, we choose the two critical values $\chi^2_{1-\alpha/2,n-1}$ and $\chi^2_{\alpha/2,n-1}$, which attributed $\alpha/2$ probability to each the left and right tails. Then with $100 \cdot (1-\alpha)\%$ confidence we can say that:
  - $\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}$

15 The Bootstrap

15.1 Definition

- Bootstrapping means to accomplish what you need with what you've got.
- The statistical bootstrap is to make the most of a smaller dataset without sacrificing statistical rigor or collecting more samples.

15.2 Confidence intervals for the mean

- Consider a sample $X_1, X_2, ..., X_n$, instead of computing the CI analytically from the sample, we instead re-sample the sample many times and examine those.
- Definition: A bootstrapped resample is a set of n draws from the original sample set with replacement. (should contain the same number of observations as the original sample)

15.3 Non-parametric bootstrap

- Parametric statistics assumes that sample data comes from a population that follows a probability distribution based on a fixed set of parameters

15.4 Parametric bootstrap

- The parametric bootstrap estimates a CI for a desired property in two steps:
  1. Repeatedly estimate the parameter(s) of the known distribution via bootstrap
  2. Compute a CI for the desired property by sampling from the know distribution using the parameters that you infered.

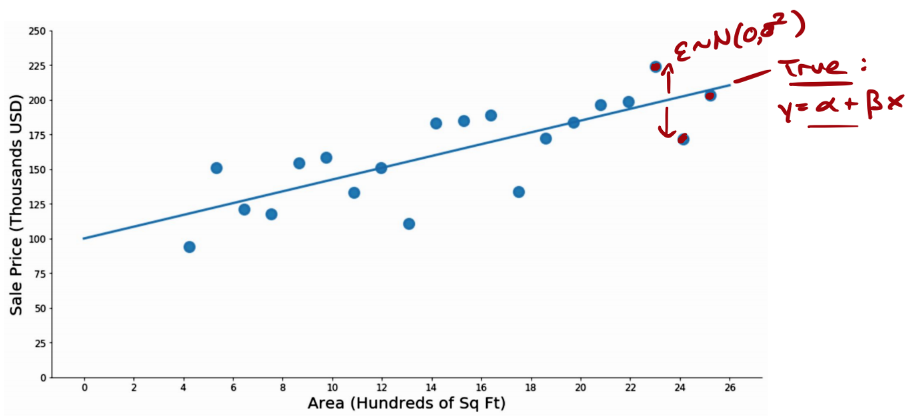
16 Introduction to Regression

16.1 Linear regression for prediction

- Examples:
  - Given a person's age and gender, predict their height
  - Given the area of a house, predict its sale price
  - Given unemployment, inflation, number of wars and economic growth, predict the president's approval rating

16.2 Simple linear regression (SLR) model

- Definitions and assumptions of SLR model:
  - $y_i = \alpha + \beta x_i + \epsilon_i$
  - Each of the ϵ_i are independent
  - $\epsilon_i \sim N(0, \sigma^2)$
- SLR model vocabulary:
  - X: the independent variable, the predictor, the explanatory variable, the feature
  - Y: the dependent variable, the response variable
  - ϵ: the random deviation or random error
- What is ϵ doing?
  - Accounting for X not being a prefect predictor of Y
  - Uncertainty
- The points $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ resulting from n independent observations will be scattered about the true regression line



- How do we know that the SLR model is appropriate?
  - Eyeball metric

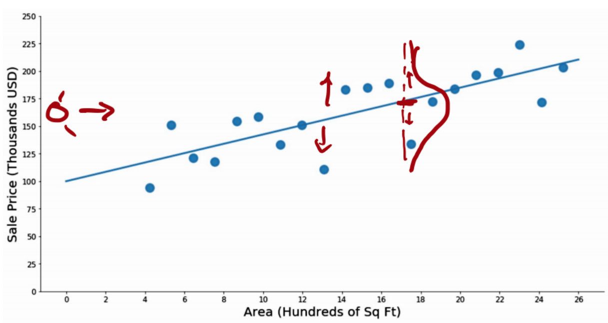
- Experience
- R^2 (later)

16.3 Interpreting SLR parameters

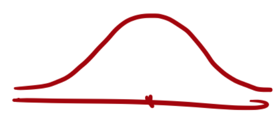
- Y is a random variable $\rightarrow$ what is its expectation?
  - $E[y] = E[\alpha + \beta x + \epsilon]$
  - $E[y] = E[\alpha] + E[\beta x] + E[\epsilon]$
  - $E[y] = \alpha + \beta E[x] + 0$ ($\epsilon \sim N(0, \sigma^2) \rightarrow E[\epsilon] = 0$)
  - $E[y] = \alpha + \beta x$
- α is the intercept of the true regression line (i.e. the baseline average)
- β is the slope of the true regression line

16.4 Interpreting the error term

- The variance parameter σ^2 determines the extent to which each normal curve spreads about the true regression line



As viewed from y-axis:



16.5 Directional considerations

- So far, we've come up with a framework where we can choose the model parameters and then generate random data. Called a generative model.
- But we really want to run this process in reverse. We have data, and we want to find/learn/estimate the parameters that explain the data. (Inference)
- General model + Parameters $\rightarrow$ Data (Sample)
- General model + Parameters $\leftarrow$ Data (Inference)

16.6 How can we estimate parameters from some data?

- Game plan: The variance of our model σ^2 will be smallest if the differences between the estimate of the true regressions line and each point is the smallest.
- This is the goal: minimize σ^2
- We will use our sample data $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ to estimate the parameters of the regression line.
- Assumption about the observations: (x_1, y_1) is collected independently of (x_2, y_2) and others.

16.7 Estimating model parameters

- Definition: The sum of squared-errors for the points $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ to the regressions line is given by:
  - $SSE = \sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2$
- Definition: The point-estimates (single value estimates from the data) of the slope($\hat{\beta}$) and intercept($\hat{\alpha}$) parameters are called least-squares estimates, and are defined to be the values that minimize the SSE:
  - Take derivative, set $= 0$ with respect to α, β
  - $\frac{dSSE}{d\alpha} = 0$ and $\frac{dSSE}{d\beta} = 0$
- Definition: The fitted regressions line or the least-squares line is then the line given by: $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$
- $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$
- $\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$

16.8 Residuals

- Fitted or predicted values $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ are obtained by plugging in the independent data variables into the fitted model
- The residuals are the difference between the observed and the predicted responses:
  - $r_i = y_i - \hat{y}_i = y_i - (\hat{\alpha} + \hat{\beta} x_i)$
- Claim: The residuals r_i are estimates of the (unknown) true error e_i

16.9 Maximum likelihood estimation

- An alternative method for estimating model parameters is to create a likelihood function that quantifies the goodness-of-fit between the model and the data, and choose the values of the parameters that maximize it.

- Likelihood function: the probability that we would observe the data we did if these parameters were true. (did this before, P(y — x))

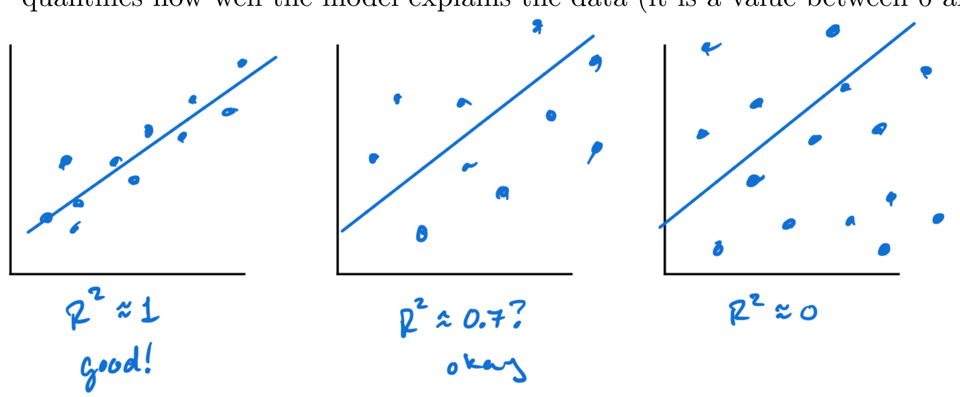
17 Inference in Regression

17.1 Estimating the variance

- The parameter σ^2 determines the spread of the data about the true regression line.
- An estimate of σ^2 will be used in computing confidence intervals and doing hypothesis testing on the estimated regression parameters
- We want answers to questions like:
  - Is the slope $\beta \neq 0$? (is the a linear relationships at all?)
  - Is the intercept $\alpha > 0$?
- Estimate of variance s given by:
  - $\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - (\hat{\alpha} + \hat{\beta}x_i))^2$
- Degrees of freedom (df) is reduced by two in denominator for $\hat{\sigma}^2$ becuase:
  - Estimating each parameter requires one degree of freedom
  - we had to estimate α and β first $\rightarrow$ loss of 2 df

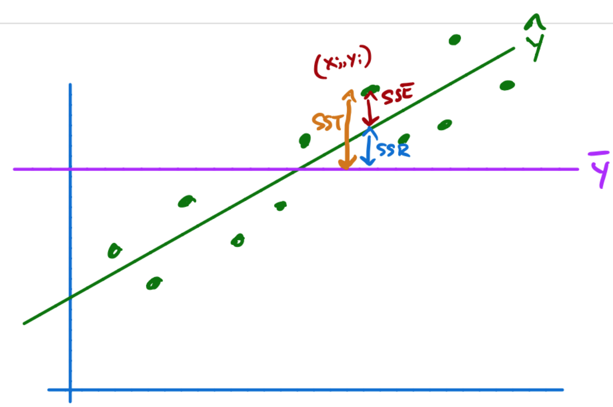
17.2 The coefficient of determination (R^2)

- R^2 quantifies how well the model explains the data (it is a value between 0 and 1)



- The sum of squared errors can be thought of as a measure of how much variation in Y left unexplained by the model. That is, how much cannot be attributed to a linear relationship.
- The regression sum of squares is given by: $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- A quantitative measure of the total amount of variation in observed Y values is given by the total sum of squares:
  - $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- Intuition: SST is what we would get for SSE if we just used the mean of the data as our model.

- The sum of squared deviations about the least-squares line is smaller than the sum of squared deviation about any other line
  - $SSE \leq SST$
  - Concept check: when are they equal?
- The ratio SSE/SST is the proportion of total variation in the data (SST) the cannot be explained by the SLR model (SSE). So we define the coefficient of determination R^2 to be the proportion that can be explained by the model:
  - $R^2 = 1 - \frac{SSE}{SST}$
- Warning: R^2 is the proportion of total variation in the data that is explained by the model. It does not tell you that you necessarily have the correct model.
- Summary:
  - SSE: unexplained variation
  - SST: total variation
  - SSR: explained variation



$$SSR = \sum (\hat{y}_i - \bar{y})^2 \checkmark$$
$$SSE = \sum (y_i - \hat{y}_i)^2$$
$$SST = \sum (y_i - \bar{y})^2$$

$$SST = SSR + SSE$$

17.3 Inference about SLR parameters

- The parameter in the simple linear regressions model have distributions
- From these distributions, we can construct CIs for the parameters, conduct hypothesis tests, and all other things.
- We will focus mainly on the slope parameter β
  - β allow us to ask/answer questions like: Is there really a relationship between the feature and the response?
  - The distribution of the slope is given by:

    - $\beta \sim N(\beta, \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}) \to SE(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{\sum_i (x_i - \bar{x})^2}}$

- The confidence intervals for β is given by:
  - $100(1-\alpha)\%$ for β is: $\hat{\beta} \pm t_{\alpha/2,n-2} \cdot SE(\hat{\beta})$
- A hypothesis testing:
  - $H_0 : \beta = c$
  - $H_1 : \beta \neq c$ (or maybe something line $\beta = c$ against $\beta > c$
  - Test statistic: $t = \frac{\hat{\beta} - c}{SE(\hat{\beta})} \to$ Comapare to $t_{\alpha/2,n-2}$ or compute p-value
  - Concept check: What t critical value would we compate for the test of $\beta = 0$ against $\beta > 0$?
- Workflow: given data(x, y)...
  1. Explore. Plot it, at least.
  2. Fit the model $\to \hat{\alpha}, \hat{\beta} \to$ get the fitted model: $\hat{y_i} = \hat{\alpha} + \hat{\beta} x_i$
  3. Goodness of fit $\to R^2 = 1 - \frac{SSE}{SST}$
  4. Inference/hypothesis testing $\to$ CI for β or hypothesis testing of $(H_0 : B = 0$, $H_1 : b \neq 0)$
     - $\hat{\sigma} = \sqrt{\frac{SSE}{n-2}}$
     - $SE(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{SS_x}}$

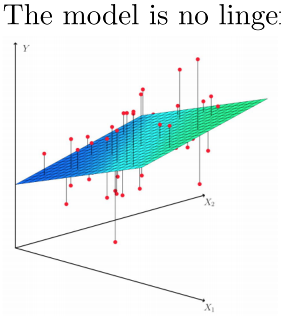
18 Multiple Linear Regression

18.1 Regression with multiple features

- Turns out, in most practical applications, there are multiple features/predictors that potentially have an effect on the response.
- Example: Suppose that Y represent the sale price of a house. What are some reasonable features associated with the sale price?
  - x_1: interior size of the house
  - x_2: size of the lot
  - x_3: number of bedrooms, etc.
- We would like to answer:
  - Is at at least one of the features useful in predicting the response?
  - Do all of the features help to explain the response? Or can we reduce to just a few?
  - How well does the model fit the data? How well does just a subset of features do?
  - Given a set of predictor values, what response should we predict, and how accurate is our prediction?

18.2 Multiple linear regression

- Definition: In MLR, the data assumed to come from a model of the form:
  - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$
- For each of the n data points $(x_{i1}, x_{i2}, ..., x_{ip}, y_i)$, for i = 1, 2, ..., n, we assume:
  - $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon$
- We make similar assumptions as in the case of SLR:
  - Each ϵ_i is independent
  - $\epsilon_i \sim N(0, \sigma^2)$
- The model is no linger a single line, it is a liner surface.



18.3 Estimating the MLR parameters

- Just as in the case of SLR, we have no hope of discovering the true model parameters
- Need to estimate them from the data. Our estimated model will be:
  - $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p$
- As before, we will find the estimated parameters by minimizing the sum of squared errors:
  - $SSE = \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p))^2$
- The SSE is again interpreted as the measure of how much variation is left in the data that cannot be explained by the model.

18.4 Covariance and correlation of features

- On way to discover these relationships among features is to do a correlation analysis.
- We want to know, if the value of one feature changes, how will this affect the other features?
- Definition: Let X and Y be random variables. The covariance between X and Y is given by:
  - $Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$
- Definition: The correlation coefficient $p(X, Y)$ is a measure between -1 and 1, and given by:
  - $p(X, Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$ (use df.corr() in python)
- The sample covariance is given by:
  - $S_{XY}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
- The sample correlation coefficient is given by:
  - $\hat{p}(X, Y) = \frac{S_{XY}^2}{\sqrt{S_X^2 S_Y^2}}$

18.5 Polynomial regression

- For single-feature data, we can fit a polynomial regression model by casting it as a multiple linear regression, where the additional features are powers of the original single feature, x.
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...$
- $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^{0.5} + \beta_4 sin(x)$
- Residual plots, in polynomial regression:
  - Recall that the assumed nature of our true model is:
    - $y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_p x^p + \epsilon$
  - If true model is $y = \beta_0 + \beta_1 x + \epsilon_1$, and our model is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \epsilon$

    - Then, $r = y - \hat{y} \sim N(0, \sigma^2)$
- If true model is $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$, and our model is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \epsilon$
    - Then, $r = y - \hat{y} \sim N(\beta_2 x^2, \sigma^2)$
- In general: if you plot the residuals $r = y - \hat{y}$, the should be normally distribute around the missing feature. So add that to your model.

19 Inference and Model Selection in MLR

19.1 Is at leaset one feature important?

- In the SLR setting, we can do a hypothesis test to determine if $\beta_1 = 0$
- In the MLR setting with p features, we need to check whether all coefficients are 0:
  - $H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$
  - $H_1 : \beta_1 \neq 0$ for at least on values of k in 1, 2, ..., p
- The F-test
  - We test the hypothesis via the F-statistic: $F = \frac{(SST-SSE)/p}{SSE/(n-p-1)}$
  - The F-statistic is a measure of how much better our model is than just using the mean

19.2 Is subset of feature important?

- Full model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ (p=4 feature in full model)
- Reduced model: $y = \beta_0 + \beta_2 x_2 + \beta_4 x_4$ (k=2 feature in reduced model)
- Are the missing features important, or are we okay going with the reduced model? Answer: Partial F-test
  - $H_0 : \beta_1 = \beta_3 = 0$ vs. H_1: at least one of β_1, β_3 is not 0
  - Since the feature in the reduced model are also in the full model, we expect the full model to perform at least as well as the reduced model. SSE(red) $\geq$ SSE(full).
  - Strategy: Fit the full and reduced models. Determine if the difference in performance is real or just due to chance.
  - Intuitively, if SSE(full) is much smaller than SSE(red), the full model fits the data much better than the reduced model.
  - The appropriate test statistic should depend on the difference SSE(red)-SSE(full) in unxplained variation.
  - Test statistic: $F = \frac{(SSE_{red}-SSE_{full})/(p-k)}{SSE_{full}/(n-p-1)} \sim Fp-k, n-p-1$
  - Rejection region: if $F \geq F_{\alpha,p-k,n-p-1}$ then reject H_0

19.3 Why use the F-tests?

- Why compute the p-value for the F-statistic when we could compute p-values for each of the feature slopes?
  - Why, Part A: If we do this, we are testing p different hypotheses instead of a single hypothesis.
  - Why, Part B: At $\alpha = 0.05$, how mane p-values do we expect to be significant if the null hypothesis is in fact true?

    * If we had 100 parameters, about 5 would be significant just by chance.
    * Problem of multiple comparisons.

19.4 Quantifying model goodness-of-fit

- Like in SLR, the MLR sum of squared errors, SSE, is: $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- Like in SLR, the MLR total sum of squares, SST, is: $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- The coefficient of determination, R^2, is: $R^2 = 1 - \frac{SSE}{SST}$
- R^2 interpretation: the fraction of variation that is explained by the model.
- The objective of MLR is not simply to explain the most variation in the data, but to do so with a model with relatively few features that are easily interpreted $\rightarrow$ principle of parsimony.
- It is thus desirable to adjust R^2 to account for the size of the model (i.e. number of features)
  - $R^2 = 1 - SSE/SST$, bu let's adject each of SSE and SST by their degrees of freedom
  - $df_{SSE} = n - p - 1$ and $df_{SST} = n - 1$
- Definition: The adjusted R^2 value is:
  - $R_a^2 = 1 - \frac{SSE/df_{SSE}}{SST/df_{SST}} = 1 - \frac{SSE/(n-p-1)}{SST(n-1)}$

19.5 Model selection: which feature should we keep?

- Forward selection: A greedy algorithm for adding features
  1. Fit model with intercept but no slopes
  2. Fit p individual SLR models – 1 for each possible feature. Add the one that improves the performance the most based on some measure (e.g. decreases SSE the most, or increases F-statistic the most)
  3. Fit p-1 MLR models – 1 for each of the remaining features, adding to the feature you added in Step 2. Add the one that improves model performance the most.
  4. Repeat until some stopping criterion is reached. (e.g. some threshold SSE, or some fixed number of features)
- Backward selection: A greedy algorithm for removing features
  1. Fit model with all available features
  2. Remove the feature with the largest p-value (i.e. the least significant feature)
  3. Repeat until some stopping criterion is reached. (e.g. some threshold SSE, or some fixed number of features)

20 Analysis of Variance (ANOVA)

20.1 Are any of the means different?

- Idea: look at where the variance in the data comes from.
- Suppose we have I groups that we want to compare, each with n_i data points (i = 1, 2, ..., I)
  - The grand mean is the sample mean of all responses.
  - The group means are the sample means within each group.

20.2 The one-way ANOVA model

- Look first at the total sum of squares: $SST = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}})^2$
- A helpful decomposition: $y_{ij} - \bar{\bar{y}} = (y_{ij} - \bar{y_i}) + (\bar{y_i} - \bar{\bar{y}})$
  - $(y_{ij} - \bar{y_i})$: within group
  - $(\bar{y_i} - \bar{\bar{y}})$: between groups
- A minor mathematical miracle:
  - $SST = \sum_{i=1}^{I} \sum_{j=1}^{n_i} [(y_{ij} - \bar{y_i})^2 + (\bar{y_i} - \bar{\bar{y}})^2] = SSW + SSB$
- The between groups degrees of freedom is: $SSB_{df} = I - 1$
  - $SSB = \sum_{i=1}^{I} n_i (\bar{y_i} - \bar{\bar{y}})^2$
- The within groups degrees of freedom is: $SSW_{df} = N - I$
  - $SSW = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y_i})^2$

20.3 Hypothesis testing

- We want to perform a hypothesis test to determine if the group means are equal. We have:
  - $H_0 : \mu_1 = \mu_2 = ... = \mu_I$
  - $H_1 : \mu_i \neq \mu_j$ for some pair i, j
- Our test statistic will be: $F = \frac{SSB/SSB_{df}}{SSW/SSW_{df}} = \frac{SSB/(I-1)}{SSW/(N-I)} \sim F_{I-1,N-1}$
- Rejection region: $F \geq F_{\alpha,I-1,N-I}$
- P-value: 1 - stats.f.cdf(F, I-1, N-I)

20.4 The ANOVA table

- It is common practice to organize all computations into an ANOVA table.

ANOVA	SS	DF	SS/DF	F
between	24	2	12	12
within	6	6	1	p = 0.008
total	30	8		

20.5 ANOVA as a multiple linear regression

- Interestingly, there is a close relationship between One-Way ANOVA and MLR
- Suppose we have I groups that you want to compare. A random sample of size n_i is taken from the i^{th} group. Then:
  - Choose one group as the control.
  - Model: $y_{ij} = \mu_0 + \tau_1 x_{1j} + \tau_2 x2j + ... + \tau_{I-1} x_{I-1,j} + \epsilon_{ij}$
    * y_{ij} is the j^{th} response for the i^{th} group, and
    * $x_{ij} = 1$ if j^{th} response if from i^{th} group, $x_{ij} = 0$ otherwise.

20.6 Tukey's honest significance test

- Suppose we determine that some of the mean are different, how can we tell which ones?
  - Tukey's HST (aka Tukey's Range Test aka Tukey's Honest Significant Difference (HSD))
  - Hypothesis test for pairwise comparison of means (it's just lots of pairwise tests)
    * It's just lots of pairwise tests using what's called the studentized range distribution
  - Adjusts so that prob of making a Type 1 error over all possible pairwise comparisons = α
    * Fixes problem of multiple comparisons.

21 Classification and Logistic Regression

21.1 The sigmoid function

- $sigm(z) = \frac{1}{1-e^{-z}}$
- Behaves like a probability
- Distinguishes between points
- Really smooth (differentiable)

21.2 Logistic regression

- The model: $p(y = 1|x) = sigm(\beta_0 + \beta_1 x)$
  - Learn weights β_0 and β_1 from the data
  - Classify data point x according to: $\hat{y} = 1$ if $sigm(\beta_0 + \beta_1 x) \geq 0.5$, $\hat{y} = 0$ if $sigm(\beta_0 + \beta_1 x) < 0.5$

21.3 An odd(s) view of logistic regression

- Our inevitable path to logistic regressions and the sigmoid function began with out insistence on modeling the relationship between features and the response as a legit probability.
- It turns out that through some basic algebra, we can arrive at an interpretation of logistic regression that is very regression-like.