Session 2: Cognitive processes and biases in Judgements

Anchoring: Summary

THE BIAS	WHY IS IT A PROBLEM?	EXAMPLES
Human tendency to rely too heavily on the first piece of information offered (the "anchor") when making judgments. A person begins with a first approximation (anchor) and then makes incremental adjustments based on additional information These adjustments are usually insufficient, giving the initial anchor a great deal of influence over the final assessment.	! The anchor might be irrelevant, reflecting only partial information, and could be intentionally manipulated.	**Negotiations** **Sales Forecasting** **Budgeting** **Earning Estimates by Analysts:** positive surprises followed by further positive surprises, and negative surprises followed by further negative surprises. **Predictions:** Asset prices anchored on what a media or a brokerage research report might suggest,.

Availability: Summary

THE BIAS	WHY IS IT A PROBLEM?	EXAMPLES
A mental shortcut that relies on immediate examples that come to a given person's mind when making a specific judgment. A tendency to heavily weigh judgments toward more recent information, making new opinions biased toward that latest information. The easier it is to recall the consequences of something the greater those consequences are often perceived to be.	! Information that comes easily to mind, or is the latest, might only be partial or incomplete information.	**Overreaction in the Financial Markets:** Excessive weight on the most recent information, for example, leading asset prices too high on good news and too low on bad news. **Premium and Discounts in the Closed-End Country Funds** **Home Bias in Investments** **Tendency to Concentrate on Well-Known Stocks:** Baidu, Alibaba, Google, Amazon,..

Representativeness: Summary

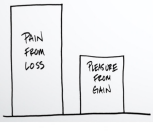
THE BIAS	WHY IS IT A PROBLEM?	EXAMPLES
Individuals tend to categorize a situation based on a pattern of previous experiences or beliefs about the scenario. It can be useful when trying to make a quick decision but it can also be limiting because it leads to close-mindedness such as in stereotypes. There are several types of representative heuristics, including the Gambler's Fallacy, Base Rate Fallacy, Regression To The Mean, and Conjunction Fallacy	! Leads to inaccurate predictions and likelihoods of events, while ignoring relevant information such as base rate	Extrapolation of future results based on a limited set of observations or facts. This is best illustrated by investors seeking a fund in which to invest, and basing their decision on the fund's most recent performance rather than covering a longer duration, especially during bear markets.

Overconfidence: Counting Letters

THE BIAS	WHY IS IT A PROBLEM?	EXAMPLES
Overestimation of one's actual performance Overplacement of one's performance relative to others Overprecision – excessive certainty regarding the accuracy of one's beliefs	! A person's subjective confidence in his or her judgments is much greater than the objective accuracy of those judgments	**Underestimation of Uncertainty** High guesses too low and low guesses too high; underestimating tails. **Underplaying Volatility in the Financial Markets:** **Range Accruals**

Loss Aversion: Summary

THE BIAS	WHY IS IT A PROBLEM?	EXAMPLES
A tendency to strongly prefer avoiding losses to acquiring gains Losses are about twice as powerful, psychologically, as gains	! It can lead to suboptimal or even bad portfolio decisions ! It leads to people chasing losses at great risk ! Leads to risk aversion, which at an aggregate level in an organization is not necessarily optimal	**Dispositon Effect:** Selling performing assets too soon and holding on to losses too long. **Escalation of Commitment:** - Traders chasing losses - Managers holding on to under-performing workers - Inability to terminate losing projects.



Session 3: Binomial and poisson + Session 4: Normal dist.

Binomial Distribution

BINOMIAL DISTRIBUTION IF

1. ***n*** **trials such that each trial has only two possible outcomes: "Success" or "Failure" (*Bernoulli* trials)**
2. **P(Success) = *p* is the same for all trials**

   P(Failure) = 1 - *p* = *q*
3. **All trials are independent.**

Interested in the probability of observing

X = # of Successes in *n* trials

Poisson Distribution

POISSON DISTRIBUTION IF...

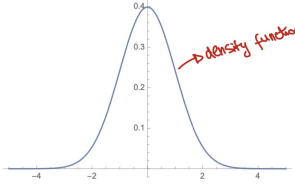
1. **Observe # of "successes" over some continuum (such as time or space)**
2. **Average # of successes is constant in the unit of measure** (λ/unit)
3. **Successes are independent**

Interested in the probability of

X = # of successes in a given period of time or space

Normal Distribution I

- Most important probability distribution you will encounter (due, in part, to the central limit theorem).
- This distribution belongs to the exponential family of distributions, and it has two parameters, its average μ and standard deviation σ.
- Represented by the famous "bell curve": symmetric around its mean



density function

- Given by

$$f(x|\mu,\sigma^2) = (2\pi\sigma^2)^{-1/2} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

variance

- Probabilities correspond to areas
- Probabilities sum to 1: $\Pr(Z < k) = 1 - \Pr(Z > k)$
- Symmetry: $\Pr(Z < -k) = \Pr(Z > k)$
- For intervals, use substraction: $\Pr(a < Z < b) = \Pr(Z > a) - \Pr(Z > b)$

$= P(Z<b) - P(Z<a)$

Binomial Distribution

↳ succession of bernoulli trials

- Suppose we are tossing a coin once, and we want to know the probability that it lands on heads (p), or tails $(1-p)$.
- If we toss the coin n times, then this becomes

$$p^x(1-p)^{n-x}$$

- But we must account for the number of difference ways we can observe x successes in n experiments. So the full distribution becomes,

$$\Pr(X = x) = \binom{n}{x} p^x(1-p)^{n-x}$$

where $\binom{n}{x} = \frac{n!}{x!(n-x)!}$

- Note that $\mu_X = \mathrm{E}[X] = np$, $\sigma_X^2 = np(1-p)$ (μ: mean, σ²: variance), $\sigma_X = \sqrt{np(1-p)}$ (σ: std)
- The parameters that describe this distribution: $\theta = (n, p)'$

Poisson Distribution

! the process is memory-less

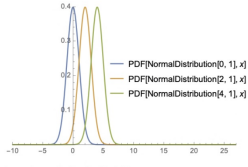
- In a given interval (such as time or space), λ is the average number of successes
- We are interested in X, the number of successes in the interval

$$\Pr(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

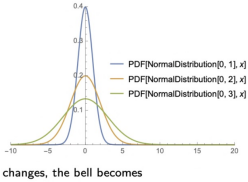
$\mu_X = \mathrm{E}[X] = \lambda$, $\sigma_X^2 = Var(X) = \lambda$, σ_X = std = $\sqrt{\lambda}$

- Why?? We will see later that Poisson distribution is an approximation to the binomial for large n and small p, [n ≥ 20, p ≤ 0.05]
- Recall for the binomial dist. $\mu_X = np$ and $\sigma_X^2 = np(1-p)$
- For small p, $1 - p \approx 1$ so variance is np. ≈ E[expectation]
- The parameters that describe this distribution: λ.

Personality of Normal Parameters



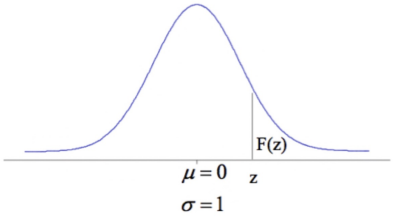
- As the mean changes, the location of the bell shifts
  To the left (for smaller means)
  To the right (for larger means)



- As the standard deviation changes, the bell becomes
  taller and thinner (for smaller standard deviations)
  shorter and thicker (for larger standard deviations)

Transformations III

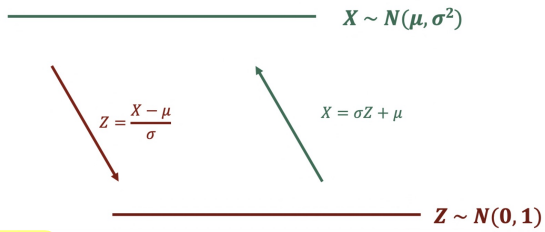
Z-SCORE TRANSFORMATIONS



- Capital F denotes a cumulative distribution:

$$F(z) = Pr(Z \leq z) = \int_{-\infty}^{z} f(z)dz$$

- So $\Pr(a < x < b) = F(b) - F(a)$



Z-Score I

Summary Z-Score

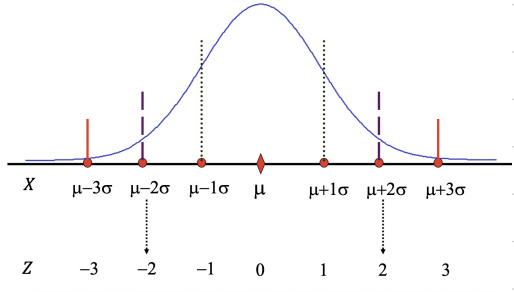
- Ex: You want $\Pr(X > k)$ where $X \sim \mathrm{N}(\mu, \sigma)$.
- Step 1: Transform $X \to Z$

$$\Pr(X > k) = \Pr\left(\frac{X-\mu}{\sigma} > \frac{k-\mu}{\sigma}\right)$$
$$= \Pr\left(Z > \frac{k-\mu}{\sigma}\right)$$

- Step 2: Look up the probability of $F(\frac{k-\mu}{\sigma})$ in R or a standard normal table.
- Step 3: Get the result:

$$\Pr(X > k) = F(\frac{k-\mu}{\sigma})$$

Z-SCORE FOR X: SUBTRACT MEAN, DIVIDE BY STANDARD DEVIATION



Z-score tells you how many standard deviations you are from the mean

Distribution	Parameters	Summary Measures
✓ Binomial	n, p	$\mu = np, \quad \sigma = \sqrt{np(1-p)}$
✓ Poisson	λ	$\mu = \lambda, \quad \sigma = \sqrt{\lambda}$
✓ Normal	μ, σ	μ, σ

Normal Approximation of the Binomial I

- The normal distribution can be used to approximate the binomial when n becomes large.
- To do so, let $\mu = np$, and $\sigma^2 = np(1-p)$
- Then use $Y \sim \mathrm{N}(\mu = np, \sigma^2 = np(1-p))$ to approximate $X \sim \mathrm{Bin}(n, p)$

Normal Approximation of the Poisson

- Similarly, we can approximate a Poisson distribution with a normal as λ gets large.
- In this case, $\mu = \sigma^2 = \lambda$.

LINEAR FUNCTION OF TWO RANDOM VARIABLES

$W = a + bX + cY$

- ❑ $a, b,$ and c are (known) constants,
- ❑ X is a random variable, with $E(X) = \mu_X$ and $Var(X) = {\sigma_X}^2$
- ❑ Y is a random variable, with $E(Y) = \mu_Y$ and $Var(Y) = {\sigma_Y}^2$
- ❑ With X and Y as random variables, W is a random variable

Always true:

1. $E(W) = \mu_W = a + b\mu_X + c\mu_Y$
2. $Var(W) = {\sigma_W}^2 = b^2{\sigma_X}^2 + c^2{\sigma_Y}^2 + 2bc[Cov(X,Y)]$

i. If X and Y are independent, then $Cov(X,Y) = 0$
ii. $Corr(X,Y) = Cov(X,Y)/\sigma_X\sigma_Y$

If X and Y are jointly Normal:

3. If $X \sim N(\mu_X, {\sigma_X}^2)$ and $Y \sim N(\mu_Y, {\sigma_Y}^2)$, then $W \sim N(\mu_W, {\sigma_W}^2)$,

correlation

$\sigma_{XY} = \rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$, $-1 < \rho_{XY} < 1$

$E(X) = \mu_X$, $sd(X) = \sigma_X$, $Var(X) = \sigma_X^2$

$E(Y) = \mu_Y$, $sd(Y) = \sigma_Y$, $Var(Y) = \sigma_Y^2$

$W = a + bX + cY$

$E(W) = \mu_W = a + b\mu_X + c\mu_Y$

$Var(W) = \sigma_W^2 = b^2\sigma_X^2 + c^2\sigma_Y^2 + 2bc[Cov(X,Y)]$

Rules for linear functions of random variables:

(1a)	μ_{aX+b}	$= a(\mu_X) + b$	Expected Value of a Linear Function
(1b)	σ_{aX+b}	$= \lvert a \rvert (\sigma_X)$	Standard Deviation of a Linear Function
(2)	$\rho_{(aX+b,cY+d)}$	$= \rho_{X,Y}$	Correlation of Linear Functions

These equations say the following:

(1a&b) If you multiply a random variable X by any number a, multiply the expected value and the standard deviation by the same amount (if $a < 0$, multiply standard deviation by $-a$).

(1a&b) If you add a constant b, add the same amount to the expected value, but do not change the standard deviation.

(2) Linear transformations do not change the correlation.

Rules for adding random variables: W = aX+bY

(3a)	$\mu_{(aX+bY)}$	$= a(\mu_X) + b(\mu_Y)$	Expected Value of a Sum
(3b)	σ_{aX+bY}	$= \sqrt{a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\,Cov(X,Y)}$	Standard Deviation of a Sum
	σ_{aX+bY}	$= \sqrt{a^2\sigma_X^2 + b^2\sigma_Y^2}$	Special Case of (3b), when X and Y are independent.

* If X and Y are independent, then $Cov(X,Y)$ is zero.

If you use the value 1.0 for a and b, i.e., W = X+Y, then these equations say the following:

(3a) If you add two random variables, you add their expected values.

(3b) If you add two random variables, to get the standard deviation you add their variances, add twice their covariance, then take the square root.

(3b) Special Case: If the variables are independent, the covariance is zero so you can just add the variances and take the square root.

Since these equations involve the covariance, whereas we are mostly familiar with correlation, the relationship between covariance and correlation is given here for convenience, in two versions.

Central limit theorem: the sum of independant variables is normal

$$\text{cov}(X,Y) = \text{E}[XY] - \text{E}[X]\text{E}[Y]$$

- Indicates the direction of a linear relationship:
  X and Y are positively related if $\text{Cov}(X,Y) > 0$.
  X and Y are negatively/inversely related if $\text{Cov}(X,Y) < 0$.
  If X and Y are independent then $\text{Cov}(X,Y) = 0$.

$$\text{corr}(X,Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- It lies between -1 and 1, i.e., $-1 \le \text{corr}(X,Y) \le 1$.
- If $\text{corr}(X,Y) < 0$ then negative relationship.
- If $\text{corr}(X,Y) > 0$ then positive relationship.
- If $\text{corr}(X,Y)$ closer to -1 or 1, then stronger relationship.

corr (X,Y) = 0 → not related (not independant)

Session 5 : CI

Formulas for Population vs. Sample Statistics

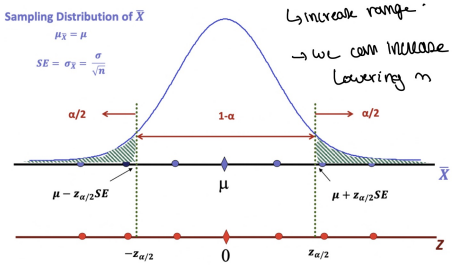
	Population (of size N)	Sample (of size n)
Mean	$\mu = \frac{\sum_i x_i}{N} = \frac{(x_1 + x_2 + \ldots + x_N)}{N}$	$\bar{x} = \frac{\sum_i x_i}{n} = \frac{(x_1 + x_2 + \ldots + x_n)}{n}$
Variance	$\sigma^2 = \frac{\sum_i (x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_i (x_i - \bar{x})^2}{(n-1)}$
Std. deviation	$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_i (x_i - \mu)^2}{N}}$	$s = \sqrt{s^2} = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{(n-1)}}$

"Exact" 95% Confidence Interval: Logic

- ❑ **95% of the time, the sample mean $\bar{x}$ falls within 1.96 standard error of the population mean μ**
- ❑ **95% of the time, $\bar{x}$ is within $\mu \pm 1.96\,SE$**
- ❑ **Hence, 95% of the time, $\bar{x} \pm 1.96\,SE$ contains μ**

In General: 100(1-α)% Confidence Intervals

Increase confidence
↳ increase range.
→ we can increase CI
Lowering n



In General: 100(1-α)% Confidence Intervals

- ❑ **100(1-α)% of the time, the sample mean $\bar{X}$ falls within $z_{\alpha/2}$ standard error (SE) of the population mean μ**
- ❑ **100(1-α)% of the time, $\bar{X}$ is within $\mu \pm z_{\alpha/2}SE$**
- ❑ **Hence, 100(1-α)% of the time, $\bar{X} \pm z_{\alpha/2}SE = \bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ contains μ**

We don't know σ usually $\approx \bar{X} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$, for n ≥ 30

Session 6: Proportions

How much to sample? Sample Size?

Recall that a $100(1-\alpha)\%$ *CI* is

$$x \pm E,$$

Where

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

n↗ → E↓ → CI↓

If we specify *E*, solving for *n* gives:

$$n = \left(\frac{z_{\alpha/2}\,\sigma}{E}\right)^2$$

↳ sample size

Example: Suppose σ = 20,000, and we want a 95% *CI* for μ with *E* = 1000

$$n = \left(\frac{z_{\alpha/2}\,\sigma}{E}\right)^2 = \left(\frac{1.96(20000)}{1000}\right)^2 \approx 1537$$

3 Types of Questions

From a study of 300 randomly selected subscribers, the WSJ reports that the average income is 80,000 with a standard deviation of 15,000.

1. **How close to this number would the *real* average (of *all* WSJ readers) be with 95% probability/confidence?**

$$E = 1.96\frac{15000}{\sqrt{300}} = 1697.41 \text{ plus or minus}$$

2. **What is the probability that the error in the estimation of the mean annual income of *all* WSJ subscribers exceeds 1000?**

$$Z_{\alpha/2}\frac{15000}{\sqrt{300}} = 1000, \text{ so } Z_{\alpha/2} = \frac{1000\sqrt{300}}{15000} = 1.155, \text{ so } \alpha \approx 2(0.1251) = 0.25$$

3. **How many subscribers should we sample to be 95% confident that the average income of *all* subscribers is within 5000 (half the width of a 95% confidence interval) of the sample mean?**

$$1.96\frac{15000}{\sqrt{n}} = 5000, \text{ so } n = \left(1.96\frac{15000}{5000}\right)^2 \approx 35$$

- From a study of 200 randomly selected defendants in the United States, the DA's office found that their average age is 25 with a standard deviation of 3.

(a) **How close to 25 would the population average be with 90% probability?**

$E = 1.65 \times \frac{3}{\sqrt{200}} = 0.35$ (4 months) plus or minus.

(b) **What is the probability that the error in the estimation of the average age exceeds 3 months?**

$z_{\alpha/2}\frac{3}{\sqrt{200}} = 0.25 \rightarrow z_{\alpha/2} = \frac{0.25\sqrt{200}}{3} = 1.18 \rightarrow$ (→ probability) $\boxed{\alpha = 2(0.12) = 0.24.}$

(c) **How many defendants should we sample so that the probability is 90% that the true average age is within 6 months of the sample mean?**

$1.65\frac{3}{\sqrt{n}} = 0.5 \rightarrow n = (1.65\frac{3}{0.5})^2 = 98$

Sampling Distribution of the Proportion

The sample proportion p follows a **Normal distribution*** with

- mean : $\mu_p = E(p) = p$
- standard error : $\sigma_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}$

But, we don't know *p*!

Estimate *p* by p ,

Then, the standard error of p : $\sigma_p \approx \sqrt{\frac{p(1-p)}{n}}$

**This approximation works well if n is large and the proportion not too extreme ($x \geq 5$ and $n\text{-}x \geq 5$).*

Confidence Intervals Derivation

❑ **With probability 0.95, p is within $p \pm 1.96(SE)$**

❑ **Hence, with probability 0.95, p is contained within $p \pm 1.96(SE)$**

❑ **So, a 95% CI for *p* is:**

$p \pm 1.96(SE)$, **where**

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

❑ **In general, a 100(1-α)% CI for *p* is:**

$p \pm z_{\alpha/2}(SE)$, **where**

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

Formulas for 100(1-α)% Confidence Interval

❑ **for the Population Mean (μ), if $n \geq 30$:**

$$X \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \textit{if } \sigma \textit{ is known}$$

$$X \pm z_{\alpha/2}\frac{s}{\sqrt{n}}, \quad \textit{if } \sigma \textit{ is unknown}$$

❑ **for the Population Proportion (*p*), if $x \geq 5$, $n\text{-}x \geq 5$:**

$$p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

↑ $p = \frac{X}{n}$

Proportions: Choosing a Sample Size

→ Choose p = 0.5

Question: how many people *n* to sample to ensure a margin of error E?

$$E = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

E is the desired margin of error, before the sample is selected.

Problem: Desired sample size depends on p , an unknown.

Standard Approach: Assume that p will be 0.5 (worst case/safest)

$$E = z_{\alpha/2}\sqrt{\frac{0.5\,(1-0.5)}{n}} \Leftrightarrow \boxed{n = \frac{0.25 z_{\alpha/2}{}^2}{E^2} = \frac{z_{\alpha/2}{}^2}{4E^2}}$$ → p=0.5

✓*Guarantees desired margin of error.*
✓*Sample size may be larger than needed.*

Procedure for CI for Population Proportion (p)

Step 1*: decide on *n (e.g., how many people to interview).
This depends on desired confidence (α) and margin of error (***E***).

$$\boxed{n = \frac{z_{\alpha/2}{}^2}{4E^2}}$$

***Step 2*: collect the data** (i.e., ask *n* people) **& build 100(1-α)% CI**

$$\boxed{p \pm z_{\alpha/2}(SE), \text{ where } SE = \sqrt{\frac{p(1-p)}{n}}}$$

***Note:* Actual margin of error is going to be *at most* E**

Session 7: small sample

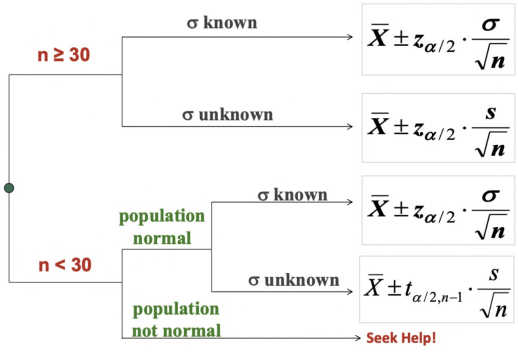
$$\bar{X} + t_{\alpha/2,\,n-1}\frac{S}{\sqrt{n}}$$

$n-1$ = degrees of freedom

$v = n-1$, d.o.f

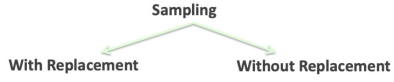
$n < 30$

A 100(1-α)% Confidence Interval for μ



Footnote: Finite samples and replacement

Role of the Population Size



Need to multiply all standard errors by the Finite Population Correction factor:

$$\sqrt{\frac{N-n}{N-1}}$$

Standard Errors with Finite Population Correction (FPC) Factor
sampling w/o replacement

σ known: $SE_{\bar{X}} = \sqrt{\frac{N-n}{N-1}}\,\frac{\sigma}{\sqrt{n}}$

$SE_p = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}}$

σ unknown: $SE_{\bar{X}} = \sqrt{\frac{N-n}{N-1}}\,\frac{s}{\sqrt{n}}$

Why is the FPC factor often ignored?

$$\sqrt{\frac{N-n}{N-1}}$$

- ❑ *This factor is ≈1 if the population size N is much larger than the sample size n (which is usually the case)*
- ❑ *It is always < 1; so if we ignore it our CIs are conservative (too wide)*

Session 8: Hypothesis testing.

- Type I error: reject a true hypothesis
  α prob.
- Type II error: accept a false hypothesis
  β prob

$\Rightarrow$ make α II $\rightarrow \alpha = 0.05$

$\Rightarrow$ minimize type I error

P(Type I error) = P(Reject H_0, given H_0 is true) = α = 0.05

P(Type II error) = P(Accept H_0, given H_A is true) = β

$1-\beta$ = P(Reject H_0, given H_A is true)

• Means: large samples ($n \geq 30$)

Step 1. Formulate the hypotheses. The null hypothesis is of the form

$$H_0 : \mu = \mu_0,$$

where μ_0 stands for the specific value given in H_0 ($\mu_0 = 115$ in the IQ example, for instance). We have considered three types of alternative hypotheses:

$H_A : \mu < \mu_0$ (one-sided test to the left),
$H_A : \mu > \mu_0$ (one-sided test to the right),
$H_A : \mu \neq \mu_0$ (two-sided test).

Step 2. Determine the appropriate test statistic. The test statistic used in this section is

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}.$$

Step 3. Specify a rejection region. For a given value of α, the decision rule is

reject H_0 if $z \leq -z_\alpha$ for a one-sided test to the left;
reject H_0 if $z \geq z_\alpha$ for a one-sided test to the right;
reject H_0 if $z \leq -z_{\alpha/2}$ or if $z \geq z_{\alpha/2}$ for a two-sided test.

Here z_α represents the value of z cutting off the area α in the right tail of the normal curve (and $z_{\alpha/2}$, of course, cuts off $\alpha/2$ in the right tail of the normal curve). When $\alpha = 0.05$, for example, $z_\alpha = 1.64$ and $z_{\alpha/2} = 1.96$; when $\alpha = 0.01$, $z_\alpha = 2.33$ and $z_{\alpha/2} = 2.58$.

If you prefer, the rejection region can be expressed in terms of $\bar{x}$:

reject H_0 if $\bar{x} \leq \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}$ for a one-sided test to the left;
reject H_0 if $\bar{x} \geq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$ for a one-sided test to the right;
reject H_0 if $\bar{x} \leq \mu_0 - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ or if $\bar{x} \geq \mu_0 + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ for a two-sided test.

Step 4. Compute the values of the sample mean $\bar{x}$ and the test statistic z from the data. If the observed z falls in the rejection region, reject H_0. Otherwise, accept H_0 → fail to reject

Means: small samples: $(n \leq 30)$

Step 1. Formulate the hypotheses. The null hypothesis is of the form

$$H_0 : \mu = \mu_0$$

and the alternative hypothesis is either

$H_A : \mu < \mu_0$ (one-sided test to the left),

$H_A : \mu > \mu_0$ (one-sided test to the right);

or $H_A : \mu \neq \mu_0$ (two-sided test).

Step 2. Determine the appropriate test statistic. The test statistic for small-sample tests involving μ is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

with $n - 1$ degrees of freedom.

Step 3. Specify a rejection region. For a given value of α, the decision rule is:

reject H_0 if $t \leq -t_{\alpha, n-1}$ for a one-sided test to the left;

reject H_0 if $t \geq t_{\alpha, n-1}$ for a one-sided test to the right;

reject H_0 if $t \leq -t_{\alpha/2, n-1}$ or if $t \geq t_{\alpha/2, n-1}$ for a for a two-sided test.

Here t represents the value of t cutting off the area α in the right tail of the t curve (and $t_{\alpha/2}$, of course, cuts off $\alpha/2$ in the right tail).

In terms of $\bar{x}$, the rejection region is

reject H_0 if $\bar{x} \leq \mu_0 - t_{\alpha, n-1} \frac{s}{\sqrt{n}}$ for a one-sided test to the left;

reject H_0 if $\bar{x} \geq \mu_0 + t_{\alpha, n-1} \frac{s}{\sqrt{n}}$ for a one-sided test to the right;

reject H_0 if $\bar{x} \leq \mu_0 - t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$ or if $x \geq \mu_0 + t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$ for a two-sided test.

Step 4. Compute the values of the sample mean $\bar{x}$ and the sample standard deviation s from the data. Then compute t. Reject H_0 if the observed t falls in the rejection region, accept H_0 otherwise.

Proportions:

Step 1. Formulate the hypotheses. The null hypothesis is of the form

$H_0 : p = p_0$,

where p_0 stands for the specific value given in H_0 (p_0 is 0.60 in the cancer example, 0.50 in the coin example, and 0.80 in the strike vote example).
We have considered three types of alternative hypotheses:

$H_A : p < p_0$ (one-sided test to the left);
$H_A : p > p_0$ (one-sided test to the right),
and $H_A : p \neq p_0$ (two-sided test).

In summary,

P-value = area to the left of the observed test statistic (z or t, whichever is used) for a one-sided test to the left;
P-value = area to the right of the observed test statistic (z or t, whichever is used) for a one-sided test to the right;
P-value = twice the one-sided P-value for a two-sided test.

Step 2. Determine the appropriate test statistic. The test statistic used in this section is

$$z = \frac{(x/n) - p_0}{\sqrt{p_0(1-p_0)/n}}.$$

Step 3. Specify a rejection region. For given value of α, the decision rule is

reject H_0 if $z \leq -z_\alpha$ for a one-sided test to the left;
reject H_0 if $z \geq z_\alpha$ for a one-sided test to the right;
reject H_0 if $z \leq -z_{\alpha/2}$ or if $z \geq z_{\alpha/2}$ for a two-sided test.

Here z_α represents the value of z cutting off the area α in the right tail of the normal curve (and $z_{\alpha/2}$, of course, cuts off $\alpha/2$ in the right tail of the normal curve).
If your prefer, the rejection region can be expressed in terms of x/n:

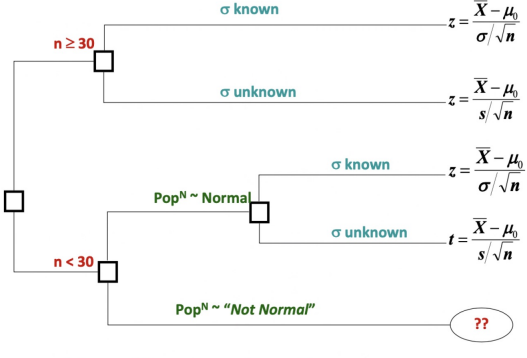
reject H_0 if $x/n \leq p_0 - z_\alpha \sqrt{p_0(1-p_0)/n}$ for a one-sided test to the left;
reject H_0 if $x/n \geq p_0 + z_\alpha \sqrt{p_0(1-p_0)/n}$ for a one-sided test to the right;
reject H_0 if $x/n \leq p_0 - z_{\alpha/2} \sqrt{p_0(1-p_0)/n}$ or if $x/n \geq p_0 + z_{\alpha/2} \sqrt{p_0(1-p_0)/n}$ for a two-sided test.

Step 4. Compute the values of the sample proportion x/n and the test statistic z from the data. If the observed z falls in the rejection region, reject H_0. Otherwise, accept H_0.

P-value < 0.05 → significant = evidence against H_0, P-value ↓ → evidence against H_0 ↑ → reject H_0

P-value > 0.05 → not significant → fail to reject H_0

Determine the Test Statistic



How to Compute P-values

p-value = probability of a result *at least as extreme* (in the direction of H_A) if H_0 is true

For a one-sided test to the right,
the P-value is the probability of a result at least as high as that observed, assuming H_0 is true.
P-value = $P(z \geq z_{observed})$, or *t* if appropriate

For a one-sided test to the left,
the P-value is the probability of a result at least as low as that observed, assuming H_0 is true.
P-value = $P(z \leq z_{observed})$, or *t*

For a two-sided test,
the P-value is the probability of a result at least as extreme (on either side) as that observed, assuming H_0 is true.
P-value = $2\,P(z \geq |z_{observed}|)$, or *t*

Hypothesis Testing Mechanics: P-Value Approach

(1) Set up hypotheses, for example:

H_0: $\mu = \mu_0$ $\quad H_0$: $p = p_0$
H_A: $\mu > \mu_0$ (or $\mu < \mu_0$, or $\mu \neq \mu_0$) $\quad H_A$: $p > p_0$ (or $p < p_0$, or $p \neq p_0$)

(2) Determine the test statistic.

(3) Decide on the significance level α

(4) Collect data, compute the sample mean (or proportion) and, if necessary, the sample standard deviation.

(5) Compute the p-value. Reject H_0 if the p-value < α.

Regression Analysis:

Introduction to Regression Analysis

Regression analysis is used to:

- **Predict the value of a dependent variable based on at least one independent variable**
- **Explain the impact of changes in an independent variable on the dependent variable**

Jargon:

- **Dependent variable: The variable we wish to predict or explain**
- **Independent variable: The variable used to explain the dependent variable**

Types of regression models:

- **Simple Regression: Use one independent variable to predict another (today)**
- **Multiple Regression: Use more than one independent variable to predict another variable**

Postulated Model vs. Estimated Model

Postulated Model (for the *population*)

$$Y_i = A + BX_i + e_i$$

Definitions:

Y Dependent variable
X Independent variable
A, B Unknown Regression Parameters
e_i Random Error Term

Assumptions:

- $e_i \sim Normal\ (0, \sigma^2)$
- e_i's are independent

Goal: Estimate A and B (based on a sample)

Estimated Model (based on a *sample*)

$$\hat{Y}_i = a + bX_i$$

Definitions:

a, b Regression Coefficients (estimates of Regression Parameters)

$\hat{Y}$ is the Predicted Value of Y for a given X

Finding the Estimated Model $\hat{Y}_i = a + bX_i$

a and b are computed such that $\sum(Y_i - \hat{Y}_i)^2$ is minimized

The resulting formulas for the optimal a and b are:

$$a = \bar{Y} - b\bar{X}$$

$$b = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{Cov(X,Y)}{Var(X)}$$

Interpretation:

- The **intercept** a is the estimated average value of Y when the value of X is zero.
- The **slope** b is the estimated average change in the value of Y due to a unit change in X.

Testing B – is there a relationship?

Objective: Is B different from zero? If not, we want it out of the model.

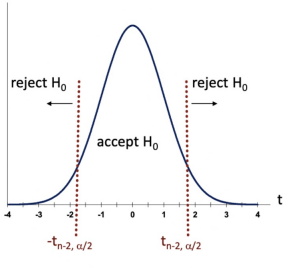
The test: H_0: B = 0 ← no relationship
H_A: B ≠ 0 ← a relationship

The statistic: If H_0 is true, then

$$\text{test statistic (t-stat)} = \frac{b - B_0}{s_b} = \frac{b}{s_b} \sim t_{n-2}$$

The decision: **Reject** at α level

if p-value < α
or if $|t\text{-stat}| > t_{n-2,\,\alpha/2}$



Testing B – is there a relationship?

Alternatively,

Does a 100(1-α)% CI for b contain zero?

Does $b \pm t_{n-2,\alpha/2} s_b$ contain zero?

For large n:

Does $b \pm z_{\alpha/2} s_b$ contain zero?

For large n (α=0.05):

Does $b \pm 1.96 s_b$ contain zero?

Regression: Are the Predictors Significant?

- ***Question:*** ***Does the independent (predictor X) variable have an impact on the dependent variable Y? In other words, is the true slope coefficient really different from zero ?***

  In general, the estimated coefficient will almost surely not be *exactly* zero. Is this due to the just chance (randomness), or is there really a relationship?

- ***Measure*****: p-value of the coefficient tells us how significant the effect is, based on the data**

  The lower the p-value, the more likely it is that the slope is really different from zero, i.e. that there is a significant impact of the explanatory variable.

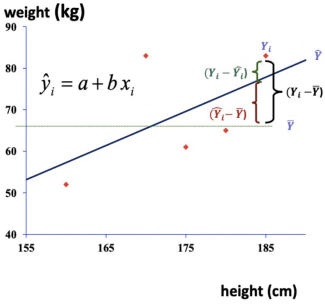
- ***Test*** **(at α = 5% significance level):**

  If p-value < 5%, then there is a significant impact!

- ***Action*****: If the predictor variable *X* is not significant, then remove it from the analysis!**

What explains the variation in Y_is around $\bar{Y}$? And, how much? R^2 (Coefficient of Determination)

So Many Sums of Squares!



$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$ Total variation to explain

$SSR = \sum_{i=1}^{n} (\bar{Y}_i - \bar{Y})^2$ Variation explained by the reg. equatio

$SSE = \sum_{i=1}^{n} (Y_i - \bar{Y}_i)^2$ Variation that is not explained

$SST = SSR + SSE$

$0 \leq R^2 \leq 1$

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

→ Proportion of variation in the dependent variable that is explained by the model

- **SST**: sum of squares total, is the squared differences between the observed dependent variable and its mean. It is the dispersion of the observed variable around the mean similar to variance.
- **SSR**: sum of squares due to regression, is the sum of differences between the predicted value of the dependent variable and its mean. It is a measure that describes how well our line fits the data. If this value of SSR is equal to the sum of squares total, it means our regression model captures all the observed variability and is perfect.
- **SSE**: the sum of squared errors, is the sum of the difference between the observed value and the predicted value. We usually want to minimize the error. The smaller the error, the better the predictive power of the model.
- R^2: An R-squared of zero means our model/line explains none of the variability in the data. An R-squared of 1 would mean our model explains the entire variability of the data. **What is a good R squared?**

Postulated vs. Estimated Model

(Multivariable)

Postulated Model *(based on entire population)*

$$Y_i = A + B_1X_{1i} + B_2X_{2i} + B_3X_{3i} + ... + B_kX_{ki} + e_i$$

- A & the B_j's are unknown regression parameters
- e_i is a random error term

Estimated Model *(based on sample)*

$$\hat{y}_i = a + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} + ... + b_kx_{ki}$$

- a & the b_j's are estimates of A & the B_j's

Testing B_j

Objective: Is B_j different from zero? If not, we want it out of the model

We know that: $b_j \sim N(B_j, {\sigma_{b_j}}^2)$

Therefore: $\frac{b_j - B_j}{\sigma_{b_j}} \sim z$ and $\frac{b_j - B_j}{s_{b_j}} \sim t_{n-k-1}$

where n = the number of observations,
k = the number of independent variables.

Testing B – is there a relationship?

Objective: Is B_j different from zero? If not, we want it out of the model.

The test: H_0: $B_j = 0$ ← no relationship
H_A: $B_j \neq 0$ ← a relationship

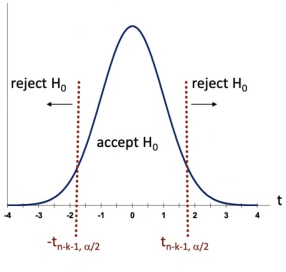
The statistic: If H_0 is true, then

test statistic = (t-stat) $\frac{b_j - B_{j0}}{s_{b_j}} = \frac{b_j}{s_{b_j}} \sim t_{n-k-1}$

The decision: **Reject** at α level

if p-value < α
or if $|t\text{-stat}| > t_{n-k-1, \alpha/2}$

- *k= # of independent variables,*
- *n = # of observations*



Testing B – is there a relationship?

Alternatively,

Does an 100(1-α)% CI for b_j contain zero?

Does $b_j \pm t_{n-k-1,\alpha/2} s_{b_j}$ contain zero?

For large **n**:

Does $b_j \pm z_{\alpha/2} s_{b_j}$ contain zero?

For large n (α=0.05):

Does $b_j \pm 1.96 s_{b_j}$ contain zero?

MULTICOLLINEARITY

Multicolinearity (M/C)

Strong linear relationship between two independent variables

Why is it a problem?

- **One of the independent variables is redundant**
- **Regression coefficients might not make sense**
- **Might be computational problems**

How to catch it?

Rule of thumb: if the correlation btw two independent variables is higher than 0.7 in absolute value,

i.e., greater than 0.7 or less than -0.7.

DEALING WITH MULTICOLLINEARITY

Two independent variables are highly correlated (positively or negatively): $|Corr| > 0.7$

- neither is significant → Remove the one with highest p-value (or smallest absolute t-test; it's the same thing).
- one is significant → Remove the one which is not significant.
- both are significant → Does the model make sense? (Do the signs of the coefficients make sense?) If yes, leave both in. If no, there could be a computational problem. Remove the one with the highest p-value.

F-test in a nutshell

- This is a global test where $H_0 : B_1 = B_2 ... = B_k = 0$.
- If $p < \alpha$ then you can reject H_0 sugesting your model is better than a model with no predictors, i.e., a model where $Y = B_0$.
- If none of the predictors are statistically significant, the overall F-test is unlikely to be statistically significant, but not always, because the **F-test evaluates all predictors jointly**.
- The intercept is equal to the mean of the response, $\bar{Y}$, when $X_1 = X_2 = ... = X_k = 0$.
- The F-test suggests you may be better off using the linear model as opposed to using just the mean of the response.

THE F-TEST

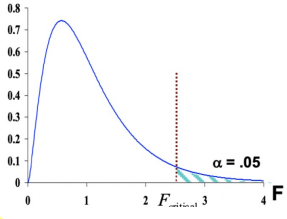
Objective: To provide a global test of the regression equation

The test: H_0: $B_1 = B_2 = \ldots = B_k = 0$
H_A: at least one $B_j \neq 0$

If H_0 were true, this would be the distribution of F.

The statistic: If H_0 is true:

$$F = \frac{R^2}{1-R^2}\left(\frac{n-k-1}{k}\right) \sim F^k_{n-k-1}$$



Interpretation: ratio of explained to unexplained variance

A STEP-BY-STEP APPROACH TO MODEL BUILDING

(1) Choose dependent and independent variables
- choose what you would like to ultimately predict (the dependent variable)

(2) Study scatterplots
- to get a visual feel for the relationships between the dependent variable and each of the independent variable

(3) Review the correlation matrix
- do the correlations make sense?
- check for possible multicollinearity (strong correlations between the independent variable

(4) Run the regression (look at F-Test)

(5) Begin the model building phase

backward elimination: (between independent variables)
- deal with multicollinearity (use multicollinearity tree)
- remove insignificant variables, one by one, starting with the one with the highest p-value (or smallest t-test) (keep step case)

CHECKING ASSUMPTIONS: RESIDUAL ANALYSIS

Assumptions:
$e_i \sim$ normal $(0, \sigma^2)$
e_i's are independent

Residuals

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

Check in Three Steps:
1. e_i's are independent
2. e_i's have the same variance
3. e_i's are normally distributed

Check for problems in the residuals:
1. **Autocorrelation** (not independent)
2. **Heteroskedasticty** (unequal variance)
3. **Non-normality** (not normally distributed)

RESIDUAL ANALYSIS I: AUTOCORRELATION

H_o: No Autocorrelation (random residuals – no systematic pattern)
H_A: Autocorrelation (systematic pattern in residuals) → *problem!*

- mainly an issue (problem) in time series data

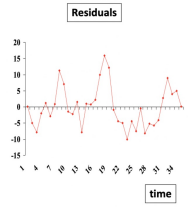
How to check for Autocorrelation?

❑ Visual check:
A plot of Residuals vs. (1) Time, or
(2) Observation Number (after ordering the data in some manner), or
(3) Lagged Residuals

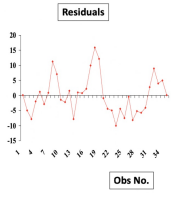
❑ A test: Durbin Watson

CHECKING FOR AUTOCORRELATION

If time series data: Plot residuals vs. time



If not time series data: Order the data in some natural way (or by Y) and then Plot Residuals vs. Observation Number



Reasons for Autocorrelation

(1) An important independent variable is missing

(2) non-linearity in the model (between *Y* and one of the *X*'s)

Residual Analysis 2: Heteroskedasticity

Heteroskedasticity

From Wikipedia, the free encyclopedia

In statistics, a sequence of random variables is **heteroscedastic**, or **heteroskedastic**, if the random variables have different variances. The term means "differing variance" and comes from the Greek "hetero" ('different') and "skedasis" ('dispersion'). In contrast, a sequence of random variables is called homoscedastic if it has constant variance.

The econometrician Robert Engle won the 2003 Nobel Memorial Prize for Economics for his studies on regression analysis in the presence of heteroscedasticity, which led to his formulation of the Autoregressive conditional heteroskedasticity (ARCH) modeling technique.

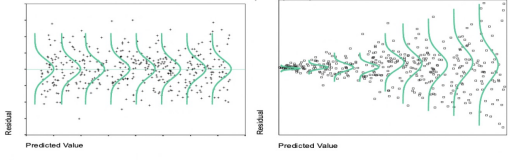
Checking for Heteroskedasticity

- ***Homoskedasticity*** **versus** ***Heteroskedasticity***
- e_i**'s have constant variance vs. they don't**

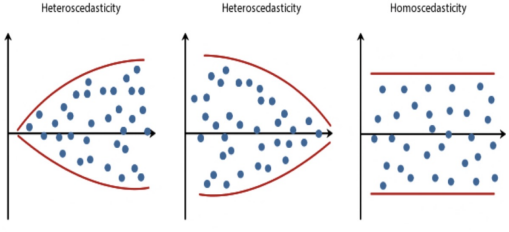
How to check?

- **Plot the residuals (** $\hat{e}_i$**) against the predicted value (** $\hat{y}_i$ **)**
- **If the boundaries are NOT parallel, heteroskedasticity!**

$$\text{Residual} = \hat{e}_i = Y_i - \hat{Y}_i$$



Checking for Heteroskedasticity



Residual Analysis 3: Non-Normality of Residuals

Diagnosis: Draw a histogram of residuals $\hat{e}_i$

Unless you have a strong feeling that it's not normally distributed, ignore it

Residual Analysis I: check for Autocorrelation

Assumption	**The** $e_i s$ **are independent**
Technical Term	No Autocorrelation
Modeling Implications if violated	❑ Mis-specified model ❑ Invalid t-tests ❑ Unreliable standard errors and R-squared
Tests	• Visual inspection: residuals vs. time, or observation number (after data is ordered by Y), or lagged variables • Statistical test: Durbin-Watson test
Remedies for violation	▪ Identify a non-linear relationship ▪ Look for an important missing independent variable

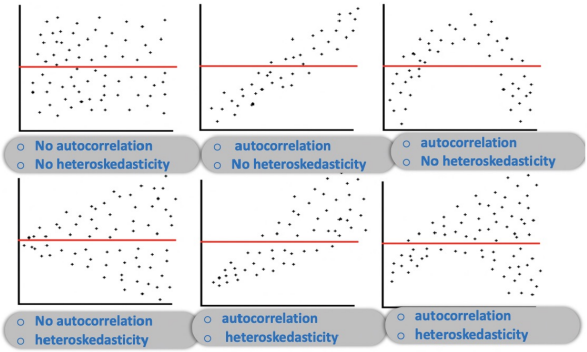
Residual Analysis 2: check for Heteroskedasticity

Assumption	**Variance of** $e_i s$ **is constant**
Technical Term	Homoskedasticity (constant variance)
Modeling Implications if violated	❑ Incorrect confidence intervals
Tests	• Visual inspection: residuals vs. predicted value ($\hat{Y}$)
Remedies for violation	▪ Identify the source (one of the independent variables) ▪ Divide the data and build separate models ▪ Transform the dependent variable

Residual Analysis 3: Check for Non-Normality of Residuals

Assumption	e_is are normally distributed
Technical Term	
Modeling Implications if violated	❑ Incorrect confidence intervals
Tests	• Visual inspection: histogram of residuals • Check what proportion of residuals lie within $\pm 1\ std.dev, \pm 2\ std.dev., \pm 3\ std.dev.$
Remedies for violation	▪ Getting more data often fixes the problem ▪ Fixing autocorrelation, heteroskedasticity will frequently fix non-normality also

Some Residual Patterns (Signals)



Variable Selection: why?

- **Variable selection** is intended to select the **best** subset of predictors. But why bother?
  1. We want to explain the data in the simplest way redundant predictors should be removed. Occam's Razor states that among several plausible explanations for a phenomenon, the simplest is best. Applied to regression analysis, this implies that **the smallest model that fits the data is best.**
  2. Unnecessary predictors will add noise to the estimation of other quantities that we are interested in. Degrees of freedom will be wasted.
  3. Collinearity is caused by having too many variables trying to do the same job.
  4. Cost: if the model is to be used for prediction, we can save time and/or money by not measuring redundant predictors.

Variable Selection: how

Prior to variable selection:

1. Identify outliers and influential points - maybe exclude them at least temporarily.
2. Add in any transformations of the variables that seem appropriate.

Methods

- Backward Elimination
- Forward Selection
- Stepwise Regression
- Criterion-based procedures

Variable Selection: bottom line

- Accept the possibility that several models may be suggested. If this happens, consider:
  1. Do the models have similar qualitative consequences?
  2. Do they make similar predictions?
  3. What is the cost of measuring the predictors?
  4. Which has the best diagnostics?
- If you find models that seem roughly equally as good but lead to quite different conclusions then it is clear that the data cannot answer the question of interest unambiguously.
- Be alert to the danger that a model contradictory to the tentative conclusions might be out there.

- Variable selection is a means to an end and not an end itself.
- **The aim is to construct a model that predicts well or explains the relationships in the data.**
- Automatic variable selections are not guaranteed to be consistent with these goals. Use these methods as a guide only.
- **Stepwise methods** use a restricted search through the space of potential models and use a hypothesis testing based method for choosing between models.
- **Criterion-based methods** typically involve a wider search and compare models in a preferable manner.

Variable Selection: methods

Methods in a nutshell

- Backward Elimination: This is the simplest of all variable selection procedures and can be easily implemented without special software.
  1. Start with all the predictors in the model
  2. Remove the predictor with highest p-value greater than α_{crit}.
  3. Reset the model and go to step 2
  4. Stop when all p-values are less than α_{crit}.

  The α_{crit} is sometimes called the p-to-remove and does not have to be 5%.
- Forward Selection: This just reverses the backward method.
  1. Start with no variables in the model.
  2. For all predictors not in the model, check their p-value if they are added to the model. Choose the one with lowest p-value less than α_{crit}.
  3. Continue until no new predictors can be added.

- Stepwise Regression: This is a combination of backward elimination and forward selection. Stepwise procedures are relatively cheap computationally but they do have some drawbacks.
  1. Because of the one-at-a-time nature of adding/dropping variables, it's possible to miss the optimal model.
  2. The p-values used should not be treated too literally. There is so much multiple testing occurring that the validity is dubious.
  3. Model selection cannot be divorced from the underlying purpose of the investigation.
- Criterion-based methods If there are p potential predictors, then there are 2^p possible models. We fit all and choose the best one according to some criterion (**how to avoid overfitting?**).
  1. The Akaike Information Criterion (AIC), the Bayes Information Criterion (BIC), the Mallow's C_p Statistic.
  2. Adjusted R^2, Predicted R^2 → adjusted R^2 ↑ when new term improves model more than would be expected by chance.

• Session 13: Non-linear models:

A 100(1-α)% CI for Regression Parameters

Regression Statistics	
Multiple R	0.8884
R Square	0.7892
Adjusted R Square	0.7651
Standard Deviation of Regression	248.7039
Observations	40
D.F.Numerator	4
D.F.Denominator	35

Analysis of Variance (ANOVA)

Source	Sum of Squares	d.f.	Mean Square	F	P-value
Regression	8105970.31	4	2026492.577	32.762708	0
Residual	2164877.214	35	61853.63468		
Total	10270847.52	39			

Dependent Variable: SALES

Independent Variable	Coefficient	Standard Error	t-stat	P-value	0.05 Significance?
Constant: a	3333.3884	363.2868	9.17563929	0.0000	Y
PDI	5.5343	0.7056	7.84372269	0.0000	Y
PRICE	-15.3901	6.5801	-2.338874262	0.0252	Y
INVEST	1.9211	0.6879	2.792711301	0.0084	Y
ADVERTIS	7.5453	1.6519	4.567660124	0.0001	Y

Confidence Intervals for Regression Coefficients (B_j)

- b_j *is an estimate of* B_j
- $b_j \sim$ *Normal with mean* B_j *and a SE*
- s_{b_j} *= estimate of the SE of* b_j
- *So,* $\frac{b_j - B_j}{s_{b_j}} \sim t_{n-k-1,\alpha/2}$
- *Hence, a* $100(1-\alpha)\%$ *CI for Regression Parameters*:

$$b_j \pm t_{\frac{\alpha}{2}n-k-1} s_{b_j}$$

- *Note, if* $n-k-1 \geq 29$, *then a* $100(1-\alpha)\%$ *CI for Regression Parameters*:

$$b_j \pm z_{\alpha/2} s_{b_j}$$

A 100(1-α)% CI for Regression Parameters

Regression Statistics	
Multiple R	0.8884
R Square	0.7892
Adjusted R Square	0.7651
Standard Deviation of Regression	248.7039
Observations	40
D.F.Numerator	4
D.F.Denominator	35

Analysis of Variance (ANOVA)

A 95% CI for "impact" of Advertising on Sales:

$$b_j \pm t_{n-k-1,\alpha/2} s_{b_j}$$

If $n\text{-}k\text{-}1 \geq 29$, *then use* $b_j \pm z_{\alpha/2}\, s_{b_j}$

$7.55 \pm 1.96\ (1.65) \approx 7.55 \pm 3.23$

Source	Sum of Squares	d.f.	Mean Square	F	P-value
Regression	8105970.31	4	2026492.577	32.762708	0
Residual	2164877.214	35	61853.63468		
Total	10270847.52	39			

Dependent Variable: SALES

Independent Variable	Coefficient	Standard Error	t-stat	P-value	0.05 Significance?
Constant: a	3333.3884	363.2868	9.17563929	0.0000	Y
PDI	5.5343	0.7056	7.84372269	0.0000	Y
PRICE	-15.3901	6.5801	-2.338874262	0.0252	Y
INVEST	1.9211	0.6879	2.792711301	0.0084	Y
ADVERTIS	7.5453	1.6519	4.567660124	0.0001	Y

Using the regression model for forecasting

- **Goal**: Forecast the value of Y when the values of independent variables are set at

$$X_1 = x_{1f},\ X_2 = x_{2f}, \ldots,\ X_k = x_{kf}$$

- **Point estimate** for the forecast will be:

$$\hat{Y}_f = a + b_1 x_{1f} + b_2 x_{2f} + \cdots + b_k x_{kf}$$

- How confident are we in this estimate? This is the question we can approach with Prediction Intervals.
- A $100(1-\alpha)\%$ **Prediction interval** is an estimate of an interval in which a future observation will fall, with a certain confidence level, given the observations that were already observed. For example, about a 95% prediction interval we can state that if we would repeat our sampling process many times, 95% of the constructed prediction intervals would contain the new observation.

Confidence (Prediction) Intervals for a Forecast

- **A** $100(1-\alpha)\%$ **Prediction Interval for forecast:**

$$\hat{Y}_f \pm t_{n-k-1,\alpha/2} s_f$$

← Standard error of the forecast (hard to compute w/o software)

- If n is large, can use $s_e \approx s_f$, i.e., a $100(1-\alpha)\%$ **Prediction Interval for forecast:**

$$\hat{Y}_f \pm t_{n-k-1,\alpha/2} s_e$$

← Standard error of the regression (standard deviation of the e_is from the regression output - always given in regression output as std. dev. of reg.)

Example: For simple regression ($k = 1$)

$$s_e = \sqrt{\frac{SSE}{n-k-1}} \quad \text{and} \quad s_f = s_e \sqrt{1 + \frac{1}{n} + \frac{(X_f - X)^2}{\sum_{i=1}^{n}(X_i - X)^2}}$$

Using the regression model for forecasting: summary

- **Goal**: Forecast the value of Y when the values of independent variables are set at

$$X_1 = x_{1f},\ X_2 = x_{2f}, \ldots,\ X_k = x_{kf}$$

- **Point estimate** for the forecast will be:

$$\hat{Y}_k = a + b_1 x_{1f} + b_2 x_{2f} + \cdots + b_k x_{kf}$$

- **Approximate**: $100(1-\alpha)\%$ Prediction Interval for this forecast.

$$\hat{Y}_f \pm t_{n-k-1,\alpha/2}\ s_e$$

where s_e is the standard error of the regression.

- For larger samples, say $n-k-1 \geq 29$, the $100(1-\alpha)\%$ Prediction Interval for this forecast is

$$\hat{Y}_f \pm z_{\alpha/2}\ s_e$$

! check multicolinearity between independant variables

F stat $\geq 3.95 \rightarrow$ reject null hypothesis at $\alpha = 0.1$

A 100(1-α)% Prediction Interval for Y

Regression Statistics	
Multiple R	0.8884
R Square	0.7892
Adjusted R Square	0.7651
Standard Deviation of Regression	248.7039
Observations	40
D.F.Numerator	4
D.F.Denominator	35

Analysis of Variance (ANOVA)

Source	Sum of Squares	d.f.	Mean Square	F	P-value
Regression	8105970.31	4	2026492.577	32.762708	0
Residual	2164877.214	35	61853.63468		
Total	10270847.52	39			

Dependent Variable SALES

Independent Variable	Coefficient	Standard Error	t-stat	P-value	0.05 Significance?
Constant: a	3333.3884	363.2868	9.17563929	0.0000	Y
PDI	5.5343	0.7056	7.84372269	0.0000	Y
PRICE	-15.3901	6.5801	-2.338874262	0.0252	Y
INVEST	1.9211	0.6879	2.792711301	0.0084	Y
ADVERTIS	7.5453	1.6519	4.567660124	0.0001	Y

A 95% Prediction Interval for Sales:
(Given PDI = 400, PRICE =70, INVEST = 100, ADVERTIS = 50)

$\widehat{SALES}$ = 3333.39 + 5.53(400) -15.39(70) +1.92(100) +7.55(50) = 5039.19

A 95% PI for SALES: $5039.19 \pm t_{n-k-1,\alpha/2} s_f =$ $5039.19 \pm (1.96)\,254.97$ — Exact, from the computer

$\approx 5039.19 \pm 1.96\, s_e = 5039.19 \pm 1.96\,248.70$ — Approximate, from the Reg Output

How to compare regression models?

- R^2 is ok for models of the same size.
  Problem: it goes up whenever you add another variable, significant or not.
- Adjusted-R^2 is used to compare models with **different no. of variables.** Unlike R^2, it can go up or down when a variable is taken out of the model.
- **Parsimony Principle**: all else equal we prefer models with fewer variables

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad R^2_{adjusted} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

→ eliminate the less significant (higher P value)

Polynomial Regression

- Consider
$$Y = \beta_0 + \beta_1 X_1 + e$$
- Suppose we believe that the change in the response is itself changing as a function of X_1 – i.e., **the slope is different at different values of** X_1. And suppose we think the curve is quadratic. We model this as:
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + e$$
- More generally up to order k:
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + \ldots + \beta_k X_1^k + e$$
- **Important**: Notice that the response is still linear in terms of the unknown parameters, even though the regression line is no longer a straight line.
- **Note**: For higher order, we may be **overfitting** because the slope would be changing very aggressively to fit each point.