Announcements:
① HW 1 Posted. Due ~~this~~ this Friday 1/22
② Read Ch 1 (and Ch 2).

STAT 342: Design + Analysis of Experiments

Introductions

Me: - Assoc Prof, 9th year on faculty
- Family: 2 kids, boys 8+10; ski (telemark), bike (road/mtn)
- Research: extreme values, atmospheric science

You: - name
- where from (?)

Wait list: see me after class

Syllabus

Review of STAT 341

SUV's data set: old now (2003+ 2004 sales), collected by CNW marketing + discussed in Denver Post.

① Fit model

② Investigate fit of model, consider model adequacy

- maybe larger variance as x (income) increases
- maybe non-linear at lower values of x.

Model that we fit was

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

car price
response/
dep var
(random)

income
covariate
indep var
(known)

error
(random)

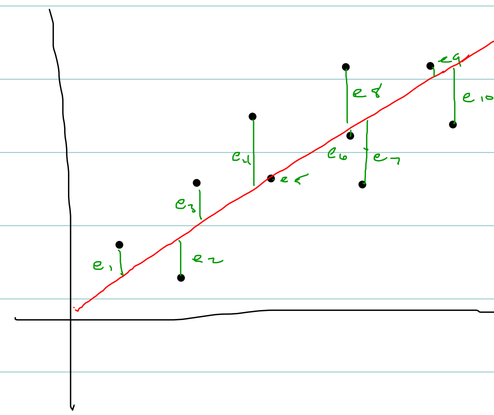
Model assumptions:

1. $E[\varepsilon_i] = 0$
2. $Var[\varepsilon_i] = \sigma^2$ (homoskedastic)
3. $\varepsilon_i \perp \varepsilon_j$
4. (Maybe) $\varepsilon_i \sim N(0, \sigma^2)$

$$E[Y_i] = E[\beta_0 + \beta_1 x_i + \varepsilon_i]$$
$$= E[\beta_0] + E[\beta_1 x_i] + E[\varepsilon_i]$$
$$= \beta_0 + \beta_1 x_i + 0$$

Implies we also assume a linear relationship in mean

Also: $Var[Y_i] = \sigma^2$



Fitting method: least squares:

Obs: $y_1 \dots y_n \Rightarrow \underline{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$; $X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$

$$\underset{n\times 1}{\underline{y}} = \underset{n\times p}{X} \underset{p\times 1}{\beta} + \underset{n\times 1}{\underline{e}}$$

$$\Rightarrow (\underline{y} - X\underline{\beta})^T (\underline{y} - X\underline{\beta}) = \underline{e}^T \underline{e} = \sum e_i^2$$

$$f(\underline{\beta}) = \underline{y}^T \underline{y} - 2\underline{\beta}^T X^T \underline{y} + \underline{\beta}^T X^T X \underline{\beta} = \sum e_i^2$$

$$\frac{df}{d\underline{\beta}} = -2X^T \underline{y} + 2X^T X \underline{\beta} \stackrel{set}{=} 0$$

$$X^T X \underline{\beta} = X^T \underline{y}$$
$$\hat{\underline{\beta}} = (X^T X)^{-1} X^T \underline{y}$$

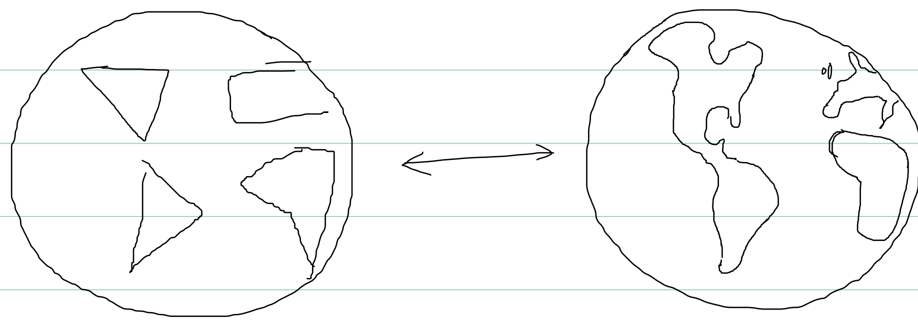
$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{e}_i^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - X_i \hat{\underline{\beta}})$$ (sample variance of residuals)

Notice: subtle change in notation

Model: $\underline{Y_i} = \beta_0 + \beta_1 x_i + \underline{\varepsilon_i}$ — Y_i, ε_i random variables

Fitting: $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \underline{\hat{e}_i}$ — y_i obs; after estimating $\beta_0 + \beta_1$, we have an observed residual

Might just say being overly concerned w/ notation/ details but relates to a bigger issue.



Math world
- random variables
- distributions
- models

$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

happily lives here
- has properties (expected value, variance, etc.)

Real World
- data

People don't buy SUV's by
① Taking income, multiplying by β_1, adding β_0
② Drawing a random # w/ mean 0 & var σ^2
③ Finding an SUV w/ that price.

"All models are wrong. Some are useful." - George Box Statistician, U Wisconsin

Statistics is a bridge between real world & math world.

Our model is probably useful. CNW marketing might use this model to for targeting advertising

It's wrong. Heteroskedacity, non linear behavior. If we were particularly interested in behavior of lower end of income range (x), might be misleading

We could work to improve model, but
① takes more work; ② more complicated models have more ways to be wrong. (over fitting).

Q: What's different between STAT 341 + 342?

A1: How the data are obtained.

- In STAT 341, most of the data obtained were observational. A (hopefully representative) sample was collected + a model was fit to this data.
- The levels/values of the covariate were not specified by the researcher. In our example, researcher did not randomly assign people to different levels of income, + then see how much they spent on an SUV.
- Disadvantage of an observational study: you cannot conclude causality. "Correlation does not imply causation." We cannot prove that having a higher income causes someone to buy a more expensive SUV.

  Ex: Global mean temp + average price of bacon. (data by year)
- In STAT 342 our data will arise from an experiment. Subjects will be randomized + assigned to different treatment groups.

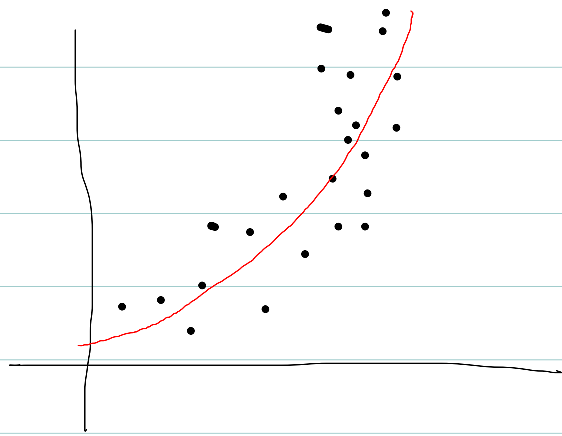
  Ex: Subject: plants, treatment (covariate/indep var) fertilizer (type or amount or both)

  Response: yield

A2: Methods/models will be different. Largely ANOVA rather than regression. But models are similar, statistically linear.

$\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$

Response didn't have to have a linear relationship w/ covariate



$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$

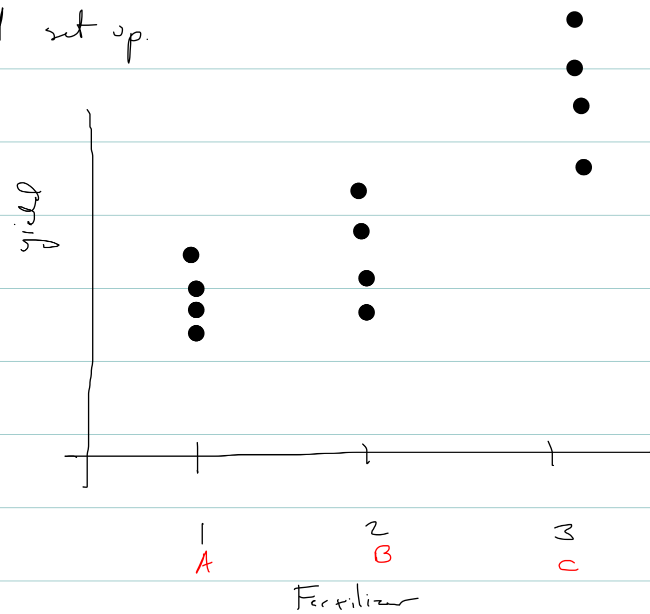
Design mtx: $X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}$

$\underline{\beta} = (\beta_0, \beta_1, \beta_2)^T$

Still: $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$ (statistically linear)

And still: $\hat{\underline{\beta}} = (X^T X)^{-1} X^T y$.

STAT 342: One way to view things will still be thru this linear model framework, but the design matrix will be different, determined by our experimental set up.



Example of what data from an experiment might look like. "Levels" of covariate fixed by researcher.

Q: Why does it not make sense to fit a model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ to this data?

Onto 342 . . .

Ch 1 Introduction

1.1 Strategy of Experimentation

→ golf example. Very good example illustrating many of ideas we will discuss in class. Response: golf score. Multiple covariates: ball, driver, travel (walk/cart), beverage. Interaction, experimental design

1.2 Applications of Experimental Design

1.3 Basic Principles (and definitions)

1.4 Guidelines for Design

1.5 History

1.6 Summary

Consider a simple experiment.

Response: growth of a plant

Variables: ① fertilizer (A+B) controllable
② seed (A+B) controllable
③ place in greenhouse (near edge or center) nuisance

How do we design the experiment? Say we can afford to do 16 plots

W1
D2 One set up: (one factor at a time)

8 plants test fertilizer (3 @ each of two levels, same seed (baseline))

8 plants test seed (3 @ each of two levels, same fertilizer (baseline))

Appealing: allows us to isolate effects of each treatment.

Data: Fertilizer exp: y_{ik} $i = A, B$ fertilizer; $k = 1, \dots, 4$ replicate

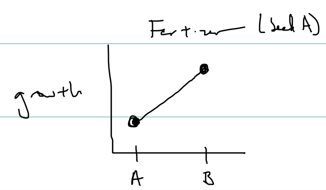
Seed Exp: z_{ik} $i = A, B$ seed; $k = 1, \dots, 4$ replicate

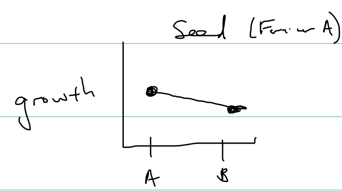
Summary statistic: Mean growth of each group

Fertilizer exp: Fertilizer A $\dfrac{y_{A1}+y_{A2}+y_{A3}+y_{A4}}{4}$; Fertilizer B: $\dfrac{y_{B1}+y_{B2}+y_{B3}+y_{B4}}{4}$

Seed Exp: Seed A $\dfrac{z_{A1}+z_{A2}+z_{A3}+z_{A4}}{4}$; Seed B: $\dfrac{z_{B1}+z_{B2}+z_{B3}+z_{B4}}{4}$

Possible results:





$\boxed{\text{Decision: Fert B, Seed A}}$

Disadvantages to OFAT

- power
- interaction

Different set up:

4 groups:

Fertilizer \ Seed	A	B
A	4	4
B	4	4

4 replicates of the experiment. Data: y_{ijk}, $i = A, B$ (fertilizer), $j = A, B$ (seed), $k = 1, \dots, 4$ (replicate)

Summary stats: by group

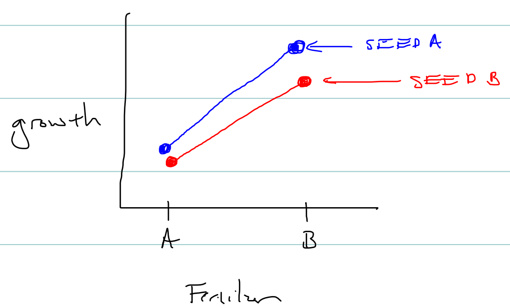
Box AA: $\dfrac{y_{AA1} + y_{AA2} + y_{AA3} + y_{AA4}}{4}$

(similar for other boxes)

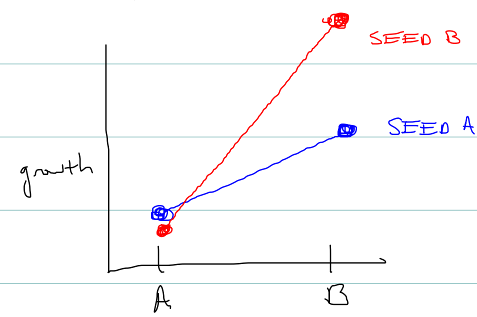
But also Fertilizer effect $\dfrac{y_{AA1} + \dots + y_{AA4} + y_{AB1} + \dots + y_{AB4}}{8}$

(does it make sense to do this? If fertilizer has same effect across seed)

One possible outcome



Another possible result.



Fertilizer B appears to outperform fertilizer A, "across" seed type. Seed A may outperform seed A (may not be significant). Need to perform stat. test.

Results about the same as OFAT. But experiment is better. How many plots do we have at each fertilizer level? Power.

$\boxed{\text{Decision: Fert B Seed A}}$

An interaction. Fertilizer does not have same effect across seed. With OFAT, we never tested seed B w/ fertilizer B.

$\boxed{\text{Decision Fert B, Seed B}}$

$\boxed{\text{"}2^2\text{ factorial experiment"}}$: 2^2 ← # of covariates to test

↑ each covariate @ two levels

Greenhouse factor? What about that? We can control it to some extent, but we aren't going to make a decision about it (we're going to grow plants at all locations in the greenhouse). Still it's something we need to consider when we're designing the experiments: Don't put all seed A plots at center and all seed B plots at edge?

However: Can we do this in a way so that we don't increase variability?

If BoxAA: $\dfrac{y_{AA1}+y_{AA2}+y_{AA3}+y_{AA4}}{4}$ $k=1,2$ center, $k=3,4$ edge

Say growth better in center, our summary stat has known source of variability built in. This is bad. Unexplained/unmodeled variability reduces our power to make a decision.

A lot of our statistical approaches will be designed to separate "signal" (explained variability) from "noise" (unexplained variability.

This is not a new idea to you. In STAT 341 you talked about decomposing the sums of squares: $SS_{TOTAL} = SS_{REG} + SS_{ERROR}$

SS_{REG}: signal due to regression

SS_{ERROR}: variability unexplained by regression model

Our BoxAA summary stat above does not separate out a known source of variability (greenhouse location). Even though we don't want to make a decision about it, we still don't want to lump the variability from location into our "noise" term.

Golf Ex: 2^4 factorial experiment (driver, ball, travel, beverage)

16 possible combinations: Too many! Say only able to do 8 rounds of golf.

Fractional factorial: how to (sensibly/strategically) leave out groups & still draw conclusions.

1.2 Applications of Exp Design

"experimentation is part of scientific process"

book: science + engineering

<u>Science</u> settings

① medicine (clinical trials)

② psychology

③ biology

④ agriculture

<u>engineering</u> (design, process management)

① design configurations

② evaluation of material alternatives

③ reduce variability

④ robust (good performance under wide variety of conditions)

Situations where experimentation is impossible

① Smoking

② climate science

1.3 Basic Principles

① randomization; ② replication; ③ blocking

① Randomization: Seems like it should be easy since we're running an experiment.

Recall: when collecting a sample of observational data, we have to worry about whether we are getting a representative sample for population of interest.

Ex: car data. How was this data obtained? If researcher went door-to-door in some neighborhood (Cherry Hills in Denver) is this representative of population of interest (SUV buyers)?

In an experiment, it's not always easy. Manufacturing process: variable: temp. Might be difficult/impossible to randomize temp thru time. But other things (material composition) could also drift thru time.

② Replication: an indep't repeat run. How many can you run? Cost. Independence

Repeated measurements: different multiple measurements on same sample unit.

(Dog example)

③ Blocking: "design technique used to improve the precision w/ which comparisons among factors of interest are made." Aim: reduce or eliminate variability due to nuisance factors (like greenhouse location).

Will not lecture on 1.4, 1.5, 1.6; but this doesn't mean unimportant, or that won't appear on exam. Simply means I don't have anything to add to what book says.

Ch 2: Simple Comparitive Experiments

- Experiments for <u>TWO</u> treatments/factor levels

2.1 Introduction

2.2 Basic statistical concepts

2.3 Sampling & Sampling distributions.

⚠ Important concept

2.4 Inference about difference in <u>means</u>, randomized designs

2.5 " , paired designs

2.6 Inference about <u>variance</u> (of normal distributions)

2.1 Intro:

Example of type of question we wish to answer

<u>Two</u> formulations of cement: ① unmodified (control); ② modified (treatment)

Q: Is tension bond strength different for modified?

Data: in book.

plot data (R code): Other possibilities: box-whisker, histogram (for each group)

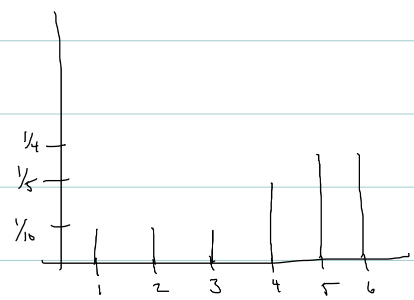
Q: Is mean TB strength significantly different?

2.2 Basic ~~Statistical~~ Probability Concepts

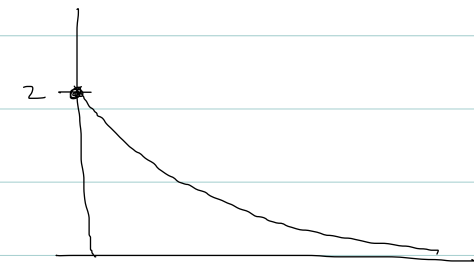
For the most part, section 2.2 talks about stuff that lives in "math world": prob dists, mean, variance (NOT sample mean/variance)

Probability distn: two (main) Types

Discrete



Continuous: (Exponential $\lambda = 2$)



PMF: $p(y) = \begin{cases} \frac{1}{10} & y = 1,2,3 \\ \frac{1}{5} & y = 4 \\ \frac{1}{4} & y = 5,6 \end{cases}$

Density $f(y) = 2\exp(-2y)$

<u>Model</u> for: loaded 6-sided die

<u>Model</u> for: bus waiting time

Both live in math world

⊕ We know <u>everything</u>.

Discrete: $P(Y \le 4) = \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{5}$

Continuous: $P(2 \le Y \le 5) = \int_2^5 f(x)\,dx = \int_2^5 2\exp(-2x)\,dx$ (u-sub)

Note: $\sum_{x=1}^{6} p(y) = 1$ & $\int_0^\infty f(y)\,dy = 1$ (check)

Notation: I am using Y for a random variable. Book uses lower case. I like to reserve lower case for values <u>after</u> they have been observed.

Mean (true mean, population mean) Still on math side — its a parameter: a (numerical) characteristic of a distribution.

$$\mu = E[Y] = \begin{cases} \sum_{\text{all } y} y\, p(y) & \text{if discrete} \\ \int_{\text{all } y:\, f(y) \geq 0} y\, f(y)\, dy & \text{if cnt} \end{cases}$$

Ex: $E[Y] = 1 \cdot \frac{1}{10} + 2 \cdot \frac{1}{10} + 3 \cdot \frac{1}{10} + 4 \cdot \frac{1}{5} + 5 \cdot \frac{1}{4} + 6 \cdot \frac{1}{4} =$

$$\frac{2}{20} + \frac{4}{20} + \frac{6}{20} + \frac{16}{20} + \frac{25}{20} + \frac{30}{20} = \frac{83}{20} = 4.15$$

$$E[Y] = \int_0^\infty y \cdot 2 \exp(-2y)\, dy = 2 \int_0^\infty y \exp(-2y)\, dy$$

IBP $u = y$ $dv = \exp(-2y)\, dy$

$du = dy$ $v = -\frac{1}{2} \exp(-2y) dy$

$$= 2 \left\{ \left[-2y \exp(-2y) \right]_0^\infty + \int_0^\infty \frac{1}{2} \exp(-2y)\, dy \right\}$$

$$= 2 \left\{ 0 + \frac{1}{2} \cdot \left[\frac{-1}{2} \exp(-2y) \right]_0^\infty \right\}$$

$$= 2 \left\{ 0 + \frac{1}{2} \left(0 - \left(-\frac{1}{2} \right) \right) \right\} = \boxed{\frac{1}{2}}$$

Variance: (another parameter, math side!)

$$\sigma^2 = Var[Y] = E[(Y-\mu)^2] = \begin{cases} \sum_{\text{all } y} (y-\mu)^2 p(y) \\ \int_{\text{all } y:\, f(y)\geq 0} (y-\mu)^2 f(y)\,dy \end{cases}$$

Rules for expectation + Variance:

1. $E[c] = c$ c a constant
2. $E[cY] = cE[Y] = c\cdot\mu$
3. $Var[c] = 0$
4. $Var[cY] = c^2\, Var[Y] = c^2\sigma^2$
5. $E[Y_1 + Y_2] = E[Y_1] + E[Y_2]$ (expectation is linear: props 2+5)
6. $Var[Y_1 + Y_2] = Var[Y_1] + 2\,Cov[Y_1, Y_2] + Var[Y_2] \neq Var[Y_1] + Var[Y_2]$ in general.

   However, if $Y_1 \perp Y_2 \Rightarrow Cov[Y_1, Y_2] = 0 \Rightarrow Var[Y_1 + Y_2] = Var[Y_1] + Var[Y_2]$
7. $E[Y_1 Y_2] \neq E[Y_1]E[Y_2]$ in general

   However: if $Y_1 \perp Y_2 \Rightarrow E[Y_1 Y_2] = E[Y_1]E[Y_2]$
8. $E\left[\frac{Y_1}{Y_2}\right] \neq \frac{E[Y_1]}{E[Y_2]}$ even if $Y_1 \perp Y_2$

2.2 Sampling & Sampling Distributions.