Announcements
① HW 2 Due this Friday 1/22
② Read Ch 1 (end Ch 2)

STAT 342: Design & Analysis of Experiments

Introductory

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

You - name (?
   - where from

Wait list - see - after class

Syllabus

Review of STAT 341

SUVs dataset: old vs. new (2003 vs. 2004 sales), collected by CNN Money & discussed in Damn Post.

① Fit model

② Investigate fit of model, consider model adequacy
   - maybe larger variance as X (more) increases
   - maybe non-linear at low value of X

Model that we fit was

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- fixed
- variance constant
- independent
- (known)
- (known)
Model assumptions:

1. $E[\epsilon_i] = 0$
2. $Var[\epsilon_i] = \sigma^2$ (homo)  
3. $\epsilon_i \perp e_j$
4. (Marginal) $\epsilon_i \sim N(0, \sigma^2)$

$$E[y_i] = E[\beta_0 + \beta_1 x_i + \epsilon_i] = E[\beta_0] + E[\beta_1 x_i] + E[\epsilon_i] = \beta_0 + \beta_1 x_i + 0$$

Fitting method: least squares.

Obs. $y_i, y_i \to y_i = (\hat{\beta}_0), X = \begin{bmatrix} 1 & x_i \end{bmatrix}$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\Rightarrow (y_i - \hat{x}_i) (y_i - \hat{x}_i) = e_i$$

$$\hat{f}(\beta) = \frac{1}{n} y_i - \frac{2}{n} \beta^T X^T y_i + \frac{2}{n} \beta^T X^T X \beta = \frac{1}{n} e_i$$

$$\frac{\partial \hat{f}}{\partial \beta} = -X^T y_i + 2 X^T X \beta = 0$$

$$X^T X \beta = X^T y_i$$

$$\hat{\beta} = (X^TX)^{-1} X^T y_i$$

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n e_i = \frac{1}{n-2} (y_i - \hat{y}_i)$$ (simple variance of residuals)

---

Notice: Such a change in notation

Model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$  
$y_i, \epsilon_i$ random variables

Fitting: $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\epsilon}_i$, after computing $\hat{\beta}_0, \hat{\beta}_1$, we have $\hat{\epsilon}_i$ observed residual
Might just say being overly concerned w/ notation/details but relate to a bigger issue.

\[ Y_i = \beta + \varepsilon_i \]

People did buy SUV's by:

1. Taking income, multiplying by $\beta$, adding $\varepsilon_i$
2. Drawing a random # with mean $\beta$ and $\varepsilon_i$
3. Factoring an SUV with that price.

"All models are wrong. Some are useful." - George Box

Statistics is a bridge between real world & math world.

Our model is probably useful. CNN modeling might use this model to target advertising.

It's wrong. Heteroskedasticity can skew behavior. If we were purely interested in behavior of lower end of income range ($x$), might be misleading.

We could work to improve model but

1. Use new data
2. More explicit models have more utility & finance (low fitting)
Q: What's different between STAT 341 and 342?

A1: How the data are obtained.

- In STAT 341, most of the data obtained were observational. A (hopefully representative) sample was collected, and a model was fit to this data.

- The levels/values of the covariates were not specified by the researcher.
  In one example, researchers did not randomly assign people to different levels of income, but 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 being a higher income cause 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 experimental. Subjects will be randomized and assigned to different treatment groups.
  Ex: Subject: plants, treatment (acorn/leather) fertilizer (type or amount used)
  Response: yield
A2. Methods/mode will be different likely ANOVA rather than regression. But maybe an simpler statistically linear

\[ Y = X\beta + \varepsilon \]

Response didn't have to have linear relationship of covariate

\[ Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \]

Design mat.

\[ X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} \]

\[ \beta = (\beta_0, \beta_1, \beta_2)^T \]

Still: \[ Y_i = X\beta + \varepsilon_i \] (statistically linear)

And still: \[ \hat{\beta} = (X^TX)^{-1}X^TY \]
STAT 342: One way to view things will still be through this linear model framework, but the design matrix will be different, determined by our experimental setup.

Example of what data from an experiment might look like: "levels" of covariate fixed by randomizer.

Q: Why does it not make sense to fit a model $Y_i = \beta + \beta X_i + \epsilon_i$ to this data?

Answer: 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, tee (w/c/e), beverage. Interaction, experimental design.

1.2 Application 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 (new edge or center) noise

How do we design the experiment? Say we can afford to do 16 plots.
One set at a time

8 plots, 3 treatments (2 = each of two levels, one seed [baseline], 3rd = each of two levels, one fertilizer [baseline])

Appealing, allows us to isolate effects of each treatment

DATA: Fertilizer exp: Yijk = \( A, B \) for level \( k = 1, \ldots, 4 \) replicate

Seed Exp: Z.1, 2 = A, B seed
\( k = 1, \ldots, 4 \) replicate

Summary statistics: mean growth of each group

Fertilizer exp: Fertilizer A: \( \bar{y}_{A, k} + \bar{y}_{B, k} + \bar{y}_{w, k} \) for level \( k = 1, \ldots, 4 \) replicate

Seed Exp: Seed A: \( \frac{Z_{A, k} + Z_{B, k} + Z_{w, k} + Z_{w, k}}{4} \)
Seed B: \( \frac{Z_{A, k} + Z_{B, k} + Z_{w, k} + Z_{w, k}}{4} \)

Possible results:

Fertilizer exp: Fertilizer A

Seed exp: Seed A

\[ y_{i, j} = a_i + b_j + \epsilon_{i, j} \]

Decision: Fertilizer B, Seed A

Disadvantages to OFAT

- power
- interaction
Different set up:

\[
\begin{array}{c|cc}
 & \text{Seed} A & \text{Seed} B \\
\hline
\text{Fertilizer} A & 4 & 4 \\
\text{Fertilizer} B & 4 & 4 \\
\end{array}
\]

4 replicates of the experiment. Data: \( y_{ijk} \), \( i = A, B \) (fertilizer)
(1, \ldots, 4) (replicate)

Sowing rate by group

\[
\text{Box} \text{ A}: \frac{y_{AA} + y_{AB} + y_{BB}}{4}
\]
(Similar for other boxes)

One possible outcome

Another possible result

Fertilizer B appears to outperform fertilizer A,

An interaction, fertilizer does not

"across" seed type. Seed A may outperform

seed A (may not be significant). Need to

perform seed test.

Resolves about the issue of OFAT

But experiment is better. How many

plots do we have at each fertilizer level? 4

Decision: fertilizers, seed B.

\[
\begin{array}{c|c}
\text{Fertilizer} & \text{Seed} A & \text{Seed} B \\
\hline
\text{A} & \text{B} & \text{B} \\
\end{array}
\]

\[2^n\text{ factorial experiment}\]

\[
2 = \# \text{of factors to test} \\
2^n = \text{# of combinations} \\
\text{Each combination two level}
\]
Greenhouse fever: What about that? We can control it to some extent, but we can'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 experiment. Don't put all seed A plots at one ad all seed B plots at the edge.

However, can we do this in a way so that we don't increase variability?

If Box A, \( \frac{y_{11} + y_{21} + y_{22} + y_{23}}{4} \) k = 1 center

k = 3.4 edge

Say growth better in center, our survey set 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: \( SS_{total} = SS_{model} + SS_{residual} \).

\( SS_{model} \) explained variability explained by model

Our Box A survey set 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 it into our "noise" term.

Golf Ex. 2nd factorial experiment (drive, pull, twist, bungee).

10 possible combinations. Too many! Says only able to do 5 rounds of golf.

Factorial factorial: how to sensibly strategically leave out groups if still draw conclusions.
1.2 Applications of Exp Design

"Experimentation is part of scientific process."

Both Science & Engineering

Science settings

1. medicine (clinical trials)
2. psychology
3. biology
4. education

Engineering (design, process management)

1. design configurations
2. evaluation of material alternatives
3. robust variability
4. robust (good performance under wide variety of conditions)

Situations where experimentation is possible

1. Smoking
2. Climate science
1.3 Basic Principles

1. Randomization
2. Replication
3. Blocking

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 random-unit 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 this time. But other things
(untold impure) could also drift this time.

Repeated measures: different multiple measurements on same sample unit
(Dog example).

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 unanswered or that unimportant.
Simply means I don't have anything to add to what book says.
Ch 2: Simple Comparison Experiments

- Experiments for Two treatments / four levels

2.1 Introduction

2.2 Basic statistical concepts

2.3 Sampling + Sampling distributions

\[ \text{Important concept} \]

2.4 Inference about difference in mean, randomized design

2.5 Inference about difference in mean, paired design

2.6 Inference about variance (of normal distributions)

2.1 Intro:

Example of type of question we wish to answer.

- Two formulations of cement:
  - 0: unmodified (control)
  - 1: modified (test)

Q. Is Tension bond strength different for modified?

Data: in book

- plot data (R code)
- Check possibilities: box-whisker, histograms (for each group)

Q. Is mean TB strength significantly different?
2.2 Basic Statistical Concepts

For the most part, section 2.2 talks about stuff that

"in

"math world" prob dists, mean, variance (not sample mean/variance.)

Probability dists: two main types

Discrete

Continuous

Exp(λ = 2)

PMF: \[ P(Y = y) = \begin{cases} \frac{1}{5}, & y = 1, 2, 3 \\ \frac{1}{4}, & y = 4 \\ \frac{1}{5}, & y = 5, 6 \end{cases} \]

Density: \[ f(y) = 2e^{-2y} \]

Model for: loaded 6-sided die. Model for: bus waiting time.

Both live in math world.

Model for:

We know everything.

Discrete: \[ P(Y \leq y) = \frac{1}{5} + \frac{1}{5} + \frac{1}{5} + \frac{1}{5} \]

Continuous: \[ P(2 \leq Y \leq 5) = \int_{2}^{5} f(y) \, dy = \int_{2}^{5} 2e^{-2y} \, dy \]

Note: \[ \sum_{y} P(Y = y) = 1 \quad \text{and} \quad \int_{0}^{\infty} f(y) \, dy = 1 \quad \text{(scale)} \]

Notation: I am using \( Y \) for a random variable. Be sure to use care I like to reserve her case for values after they have been observed.
\[ 
E[Y] = \begin{cases} 
\sum_y y f(y) & \text{if discrete} \\
\int_y y \, f(y) \, dy & \text{if continuous} 
\end{cases} 
\]

\[ 
E[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{5} + 4 \cdot \frac{1}{5} + 5 \cdot \frac{1}{4} + 6 \cdot \frac{1}{4} = \frac{2}{6} + \frac{4}{6} + \frac{6}{6} + \frac{9}{5} + \frac{20}{5} = \frac{3}{6} = 0.5 
\]

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

IBP \quad u = y, \quad \frac{dv}{dy} = \exp(-y) \, dy \\
\quad \frac{du}{dy} = 1, \quad v = -\frac{1}{2} \exp(-y) 

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

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

\[ 
= 2 \left[ 0 + \frac{1}{2} \cdot \left( 1 - 0 \right) \right] = \frac{1}{2} 
\]
\[ \text{Variance: (another formula, with side)} \]

\[
\begin{align*}
\text{Var}[Y] &= E[(Y-\mu)^2] \\
&= \sum_{y} (y-\mu)^2 p(y) \\
&= \int (y-\mu)^2 f(y) \, dy
\end{align*}
\]

\[
R\&\text{e. for expectation} + \text{Variance:}
\]

1. \( E[c] = c \quad c = \text{constant} \)
2. \( E[cY] = cE[Y] = c\mu \)
3. \( \text{Var}[c] = 0 \)
4. \( \text{Var}[cY] = c^2 \text{Var}[Y] = c^2 \sigma^2 \)
5. \( E[Y + Y'] = E[Y] + E[Y'] \) \quad (\text{expectation is linear})
6. \( \text{Var}[Y + Y'] = \text{Var}[Y] + 2 \text{Cov}[Y, Y'] + \text{Var}[Y'] \neq \text{Var}[Y] + \text{Var}[Y'] \) \quad \text{in general}

\[
\begin{align*}
\text{However, if } Y \perp Y' &\Rightarrow \text{Cov}[Y, Y'] = 0 \\
\Rightarrow \text{Var}[Y + Y'] &= \text{Var}[Y] + \text{Var}[Y']
\end{align*}
\]

7. \( E[Y, Y'] + E[Y]E[Y'] \) \quad \text{in general}

\[
\begin{align*}
\text{However, if } Y \perp Y' &\Rightarrow E[Y, Y'] = E[Y]E[Y']
\end{align*}
\]

8. \( E\left[\frac{Y}{Y'}\right] = \frac{E[Y]}{E[Y']}, \quad \text{av} \text{ if } Y \perp Y' \)