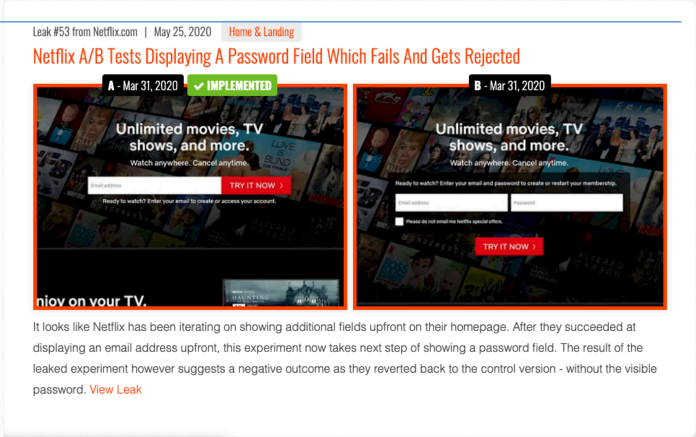
Methods of Applied Statistics I

STA2101H F LEC9101

Week 4

October 6 2021



Today

Start Recording

1. Upcoming events, HW 4
   Office Hour Monday Oct 11 7pm-8.30pm
2. Project and HW 4
3. Linear Regression Part 4: recap, collinearity, model-building, $p > n$
4. Types of studies
5. Third hour – ~~HW 2 Comments~~, HW 3 help

Upcoming

- Bayesian inference for star clusters Thursday Oct 7 3.30 Zoom Link

Gwendolyn Eadie, University of Toronto

Short Bio

My research is in the interdisciplinary field of astrostatistics, and I am jointly-appointed between the Department of Astronomy & Astrophysics and the Department of Statistical Sciences. I am interested in using and developing modern statistical methods for astronomy applications to answer fundamental questions about the universe. For example, I use hierarchical Bayesian analysis to study the dark matter halo of the Milky Way and other galaxies, and am developing new time series analysis methods to learn about the internal structure of stars.

Upcoming

- Bayesian inference for star clusters Thursday Oct 7 3.30 Zoom Link

Gwendolyn Eadie, University of Toronto

Short Bio

My research is in the interdisciplinary field of astrostatistics, and I am jointly-appointed between the Department of Astronomy & Astrophysics and the Department of Statistical Sciences. I am interested in using and developing modern statistical methods for astronomy applications to answer fundamental questions about the universe. For example, I use hierarchical Bayesian analysis to study the dark matter halo of the Milky Way and other galaxies, and am developing new time series analysis methods to learn about the internal structure of stars.

- Friday Oct 8 Toronto Data Workshop Zoom link

Toronto Data Workshop this Friday, 8 October, at noon (Toronto time) hosts Fedor Dokshin, on the intersection of data science and sociology.

Fedor Dokshin - http://www.fedordokshin.org - is an Assistant Professor of Sociology at the University of Toronto. He is a computational social scientist with research interests in social networks, organizations, and energy and the environment. Across these domains, Fedor leverages data science methods and novel data sources to im
existing measurement strategies.

Link: https://utoronto.zoom.us/j/84277066292
Meeting ID: 842 7706 6292
Passcode: data_4_lyf

Project

(UC Irvine)

1. The data source
2. The size of the data – number of observations and number of covariates
3. the response variable(s) y
4. a description of the potential covariates – explanatory vars
   LM predictors $\rightarrow X$
   indep. vars
5. the scientific questions of interest

Project

1. The data source
2. The size of the data – number of observations and number of covariates
3. the response variable(s)
4. a description of the potential covariates
5. the scientific questions of interest

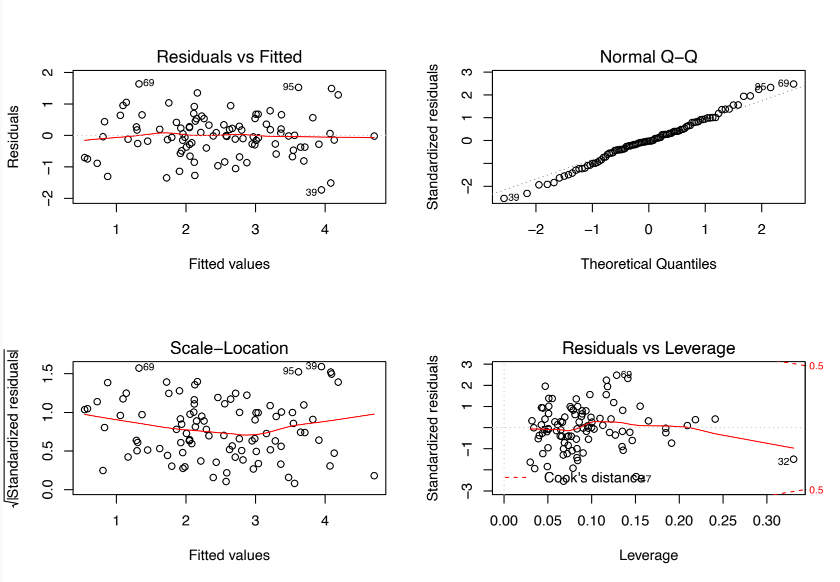
When you submit your final project, it will consist of (at least) the following parts:

1. a description of the scientific problem of interest
2. how (and why) the data being analyzed was collected
3. preliminary description of the data (plots and tables)
4. models and analysis
5. summary for a statistician of the analysis and conclusions
6. non-technical summary for a non-statistician of the analysis and conclusions

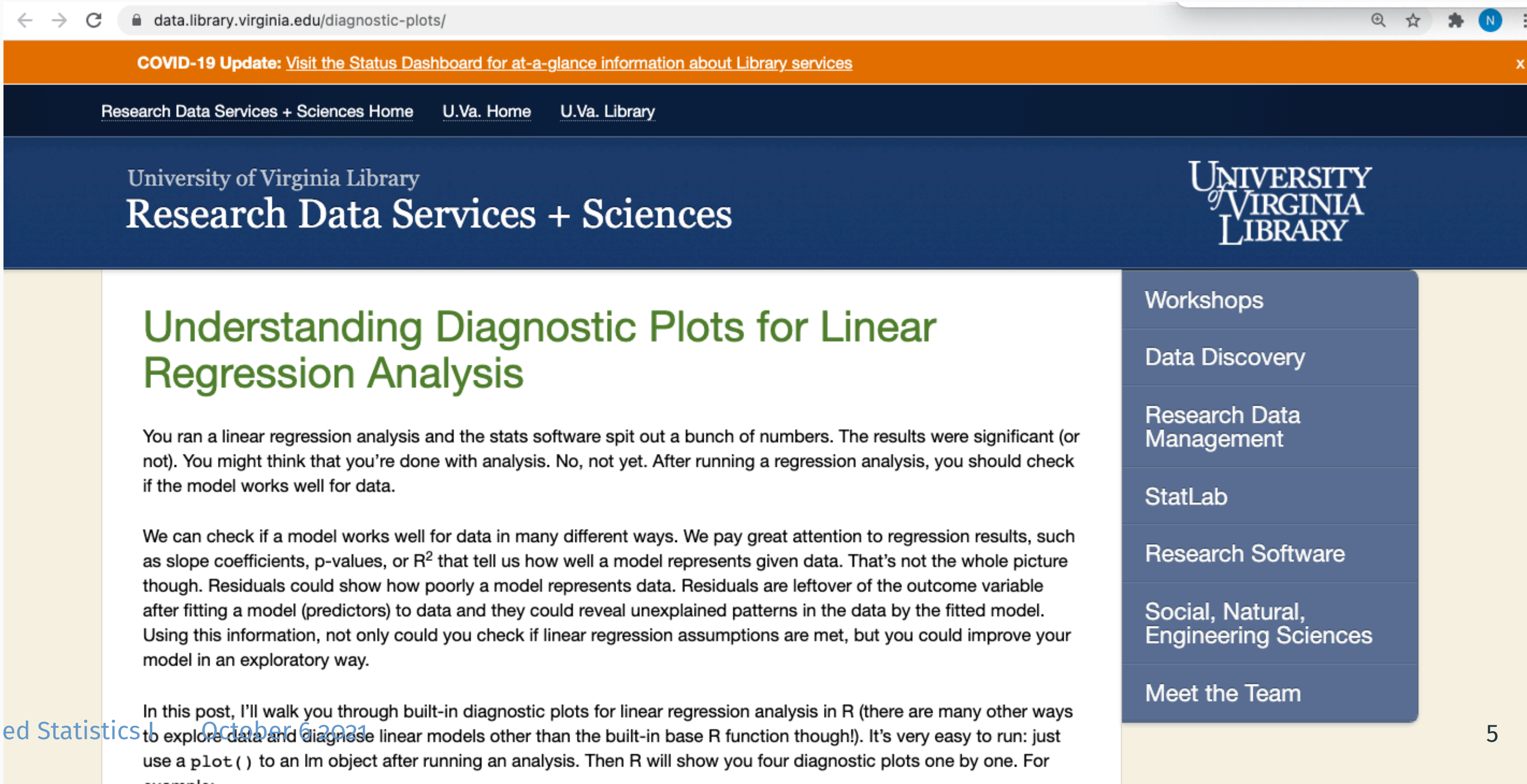
Linear regression recap

- `plot(model1)`

https://data.library.virginia.edu/diagnostic-plots/



… recap


data.library.virginia.edu/diagnostic-plots/

COVID-19 Update: Visit the Status Dashboard for at-a-glance information about Library services

Research Data Services + Sciences Home U.Va. Home U.Va. Library

University of Virginia Library
Research Data Services + Sciences

University of Virginia Library

- Workshops
- Data Discovery
- Research Data Management
- StatLab
- Research Software
- Social, Natural, Engineering Sciences
- Meet the Team

Understanding Diagnostic Plots for Linear Regression Analysis

You ran a linear regression analysis and the stats software spit out a bunch of numbers. The results were significant (or not). You might think that you're done with analysis. No, not yet. After running a regression analysis, you should check if the model works well for data.

We can check if a model works well for data in many different ways. We pay great attention to regression results, such as slope coefficients, p-values, or R^2 that tell us how well a model represents given data. That's not the whole picture though. Residuals could show how poorly a model represents data. Residuals are leftover of the outcome variable after fitting a model (predictors) to data and they could reveal unexplained patterns in the data by the fitted model. Using this information, not only could you check if linear regression assumptions are met, but you could improve your model in an exploratory way.

In this post, I'll walk you through built-in diagnostic plots for linear regression analysis in R (there are many other ways to explore data and diagnose linear models other than the built-in base R function though!). It's very easy to run: just use a `plot()` to an lm object after running an analysis. Then R will show you four diagnostic plots one by one. For

... Recap

- residuals: $\hat{\epsilon}_i = y_i - \hat{y}_i$

- $\text{Var}(\hat{\epsilon}) = \sigma^2(I - H)$, $\quad \text{Var}(y_i - \hat{y}_i) = \sigma^2(1 - h_{ii})$ $\qquad 0 < h_{ii} < 1, \Sigma h_{ii} = p$
- i.e. don't all have the same variance

large h_{ii} suggests case i is influential

- hat matrix $H = \underset{n\times p}{X}(X^{\mathrm{T}}X)^{-1}\underset{p\times n}{X^{\mathrm{T}}}$ $\qquad Hy = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = X\hat{\beta} = \hat{y}$

$n\times n$

- standardized residuals: $r_i = \dfrac{\hat{\epsilon}_i}{\tilde{\sigma}(1 - h_{ii})^{1/2}}$ $\qquad$ approx var 1

- Cook's distance $C_i = \dfrac{(\hat{y} - \hat{y}_{-i})^{\mathrm{T}}(\hat{y} - \hat{y}_{-i})}{p\tilde{\sigma}^2} = \dfrac{r_i^2 h_{ii}}{p(1 - h_{ii})}$ $\qquad$ measure of influence

high leverage or high residual

leave out ith case

... Recap

- Model structure: $\underbrace{E(y \mid X) = X\beta}_{\text{systematic part}}, \quad \overbrace{\text{Var}(Y \mid X) = \sigma^2 I}^{\text{stochastic part}}$

- added variable plots:

  plot residuals from y on X_{-j} against residuals from x_j on X_{-j}

  (slope of this line is $\hat\beta_j$ – nice exercise)

  res res $\hat\beta_j$

  partial regression plots

  note: all components obtained from original fit

- ~~partial residual plots:~~

  $\boxed{\text{plot } \hat\beta_j x_j + \hat\epsilon \text{ against } x_j}$

predict $= \hat y$ vs x_j

x_j

$$y - \sum_{j \neq j'} x_{ij} \hat\beta_j$$

$$= \hat y + \hat\epsilon - \sum_{j \neq j'} x_{ij} \hat\beta_j$$

$$= x_{ij'}\hat\beta_{j'} + \hat\epsilon$$

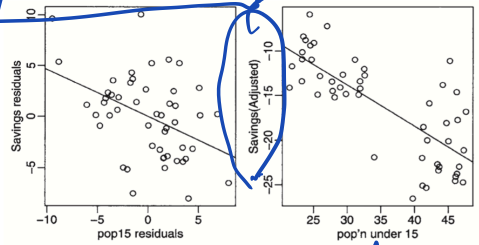


Figure 4.13 *Partial regression (left) and partial residual (right) plots for the savings data.*

x_j pop15

better for outliers better for nonlinearity

Collinearity

- simple model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad i = 1, \ldots n$
- if $x_1 \perp x_2$, then interpretation of β_1 and β_2 clear and $\hat\beta_1, \hat\beta_2$
- if $x_1 = x_2$ then β_1 and β_2 not separately identifiable

$\sum_{i=1}^{n} x_{1i} x_{2i} = 0$

$$\begin{bmatrix} 15 & 15 \\ 17 & 17 \\ 13 & 13 \\ \vdots & \vdots \end{bmatrix}$$

$\hat\beta_1$ or $\hat\beta_1$

not $(\hat\beta_1, \hat\beta_2)$

$$\hat y = \hat\beta_0 + \hat\beta_1 x_2 + \hat\beta_2 x_2$$

Collinearity

- simple model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad i = 1, \ldots n$
- if $x_1 \perp x_2$, then interpretation of β_1 and β_2 clear
- if $x_1 = x_2$ then β_1 and β_2 not separately identifiable
- usually we're somewhere in between, at least in observational studies
- may be very difficult to dis-entangle effects of correlated covariates

Collinearity

- simple model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad i = 1, \dots n$
- if $x_1 \perp x_2$, then interpretation of β_1 and β_2 clear
- if $x_1 = x_2$ then β_1 and β_2 not separately identifiable
- usually we're somewhere in between, at least in observational studies
- may be very difficult to dis-entangle effects of correlated covariates
- example: health effects of air pollution
- measurable increase in mortality on high-pollution days
- measurable increase in mortality on high-temperature days
- high temperatures and high levels of pollutants tend to co-occur

Collinearity

- simple model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad i = 1, \dots n$
- if $x_1 \perp x_2$, then interpretation of β_1 and β_2 clear
- if $x_1 = x_2$ then β_1 and β_2 not separately identifiable
- usually we're somewhere in between, at least in observational studies
- may be very difficult to dis-entangle effects of correlated covariates
- example: health effects of air pollution
- measurable increase in mortality on high-pollution days
- measurable increase in mortality on high-temperature days
- high temperatures and high levels of pollutants tend to co-occur +++
- mathematically, $X^{\mathrm{T}}X$ is nearly singular, or at least ill-conditioned, so calculation of its inverse is subject to numerical errors
- if $p > n$ then $X^{\mathrm{T}}X$ not invertible, no LS solution ridge, Lasso more next week

```
> model1

Call:
lm(formula = lpsa ~ ., data = prostate)

> X <- model.matrix(model1)
> X[1,]
(Intercept)      lcavol     lweight         age        lbph         svi         lcp 
  1.0000000  -0.5798185   2.7695000  50.0000000  -1.3862940   0.0000000  -1.3862900 
    gleason       pgg45 
  6.0000000   0.0000000 
> e <- eigen(t(X[,-1])%*%X[,-1])

[1]   1.00000   2.78186  47.66094  52.22787  85.98499 103.73114 153.85414 243.30248
> vif(X)
(Intercept)      lcavol     lweight         age        lbph         svi         lcp 
   2.004951    2.054115    1.363704    1.323599    1.375534    1.956881    3.097954 
    gleason       pgg45 
   2.473411    2.974361 
```

largest

$\frac{\lambda_p}{\lambda_1}$ = cond = number

eigenvalues of (X^TX)

$\vec{\beta}_j$ as nr vr nr.

- $$\text{var}(\hat{\beta}_j) = \sigma^2 \left(\frac{1}{1-R_j^2}\right) \underbrace{\frac{1}{\Sigma_i(x_{ij}-\bar{x}_j)^2}}$$

  R_j^2 from x_j on X_{-j}

  LM-2 §7.3; LM-1 §5.3

- variance inflation factor

  symmetric

  $$\frac{1}{1-R_j^2}$$

  `vif(X[,-1])`

- $$X^TX = U\Lambda U^T, \quad \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p), \quad \underline{\lambda_1 \geq \ldots \geq \lambda_p \geq 0}$$

  $U^TU = I$

- $\underline{X^TX \text{ invertible}} \iff \underline{\lambda_p > 0}$, but if several λ's are small, it is nearly singular

- condition number (of X): λ_1/λ_p "$>$ 30 considered large"; LM
- $$(\hat{\beta}-\beta)^T(\hat{\beta}-\beta) = ||\hat{\beta}-\beta||_2^2 \stackrel{d}{=} \sigma^2 \sum_{j=1}^{p} Z_j^2/\lambda_j, \quad Z_1,\dots,Z_p \stackrel{iid}{\sim} N(0,1)$$
  L_2 norms
- $$\mathrm{E}(\hat{\beta}-\beta)^T(\hat{\beta}-\beta) = \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1}, \quad \mathrm{var}(\hat{\beta}-\beta)^T(\hat{\beta}-\beta) = 2\sigma^4 \sum_{j=1}^{p} \lambda_j^{-2}$$
  SM, but $d_1 = \lambda_p$
- "statistical interpretation of condition number is not clear-cut" SM
- "a more systematic approach to dealing with weak design matrices is ridge regression" SM, choose regularization parameter by cross-validation

X^TX nearly singular

$$\hat{\beta}_{ridge} = (X^TX + \alpha I)^{-1} X^T y$$

Aside: standardizing dummy variables

- ridge regression: $\arg\min_{\beta}(y - X\beta)^T(y - X\beta) + \lambda\sum_{j=1}^{p}\beta_j^2$
- lasso regression $\arg\min_{\beta}(y - X\beta)^T(y - X\beta) + \lambda\sum_{j=1}^{p}|\beta_j|$
- need to center and scale columns of X so that β's are all on the same scale
- what about dummy variables?
- Hesterburg, 2021: don't scale dummy variables; instead scale other variables to match the SD of dummy variables with the same standardized skewness

  handles highly unbalanced dummy covariates

- LM-2 §7.2: "A binary predictor taking the values of 0/1 with equal probability has a standard deviation of 1/2. This suggests scaling the other continuous predictors by two SDs rather than one."

  $x = \pm 1$?

Model Selection

- "analyses should be as simple as possible, but no simpler"

Model Selection

- "analyses should be as simple as possible, but no simpler"
- What variables should we keep in the model ?

Model Selection

- "analyses should be as simple as possible, but no simpler"
- What variables should we keep in the model ?
- Hierarchical models: some models have a natural hierarchy: polynomials, factorial structure, auto-regressive, sinusoidal, ...

Model Selection

- “analyses should be as simple as possible, but no simpler”
- What variables should we keep in the model ?
- Hierarchical models: some models have a natural hierarchy: polynomials, factorial structure, auto-regressive, sinusoidal, ...
- in these models the ‘highest’ level of the hierarchy is removed first

Model Selection

- “analyses should be as simple as possible, but no simpler”
- What variables should we keep in the model ?
- Hierarchical models: some models have a natural hierarchy: polynomials, factorial structure, auto-regressive, sinusoidal, ...
- in these models the ‘highest’ level of the hierarchy is removed first
- e.g. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ should *not* be simplified to $y = \beta_0 + \beta_2 x^2 + \epsilon$

Model Selection

- “analyses should be as simple as possible, but no simpler”
- What variables should we keep in the model ?
- Hierarchical models: some models have a natural hierarchy: polynomials, factorial structure, auto-regressive, sinusoidal, ...
- in these models the ‘highest’ level of the hierarchy is removed first
- e.g. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ should *not* be simplified to $y = \beta_0 + \beta_2 x^2 + \epsilon$
- e.g. if interaction terms are included, then main effects and other 2nd-order terms also need to be included: $y = \beta_0 + \underline{\beta_1 x_1 + \beta_2 x_2} + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon$

- “analyses should be as simple as possible, but no simpler”
- What variables should we keep in the model ?
- Hierarchical models: some models have a natural hierarchy: polynomials, factorial structure, auto-regressive, sinusoidal, ...
- in these models the ‘highest’ level of the hierarchy is removed first
- e.g. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ should *not* be simplified to $y = \beta_0 + \beta_2 x^2 + \epsilon$
- e.g. if interaction terms are included, then main effects and other 2nd-order terms also need to be included: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon$
- *not* $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$ unless $x = 0/1$

Model Selection

- "analyses should be as simple as possible, but no simpler"
- What variables should we keep in the model ?
- Hierarchical models: some models have a natural hierarchy: polynomials, factorial structure, auto-regressive, sinusoidal, ...
- in these models the 'highest' level of the hierarchy is removed first
- e.g. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ should *not* be simplified to $y = \beta_0 + \beta_2 x^2 + \epsilon$
- e.g. if interaction terms are included, then main effects and other 2nd-order terms also need to be included: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon$
- *not* $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$ unless $x = 0/1$
- $y = \beta_0 + \beta_1 \sin(2\pi x) + \beta_2 \cos(2\pi x) + \beta_3 \sin(4\pi x) + \beta_4 \cos(4\pi x) + \epsilon$

lower order

Model Selection

- "analyses should be as simple as possible, but no simpler"
- What variables should we keep in the model ?
- Hierarchical models: some models have a natural hierarchy: polynomials, factorial structure, auto-regressive, sinusoidal, ...
- in these models the 'highest' level of the hierarchy is removed first
- e.g. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ should *not* be simplified to $y = \beta_0 + \beta_2 x^2 + \epsilon$
- e.g. if interaction terms are included, then main effects and other 2nd-order terms also need to be included: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon$
- *not* $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$ unless $x = 0/1$
- $y = \beta_0 + \beta_1 \sin(2\pi x) + \beta_2 \cos(2\pi x) + \beta_3 \sin(4\pi x) + \beta_4 \cos(4\pi x) + \epsilon$
- $y_t = \beta_0 + \alpha y_{t-1} + \epsilon$ $y_t = \beta_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \epsilon_t$ *not* $y_t = \beta_0 + \alpha_2 y_{t-2} + \epsilon$

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ $\beta_0 = 0$

... Model Selection

- testing procedures: forward selection, backward selection, stepwise selection

- it is quite common to fit all explanatory variables, and then drop if $p > 0.05$
- if estimates and estimated standard errors don't change very much, may be okay
- if estimates and estimated standard errors change a lot, cause for concern
- if estimates change sign, points to possibly extreme confounding

- testing procedures: forward selection, backward selection, stepwise selection

- it is quite common to fit all explanatory variables, and then drop if $p > 0.05$
- if estimates and estimated standard errors don't change very much, may be okay
- if estimates and estimated standard errors change a lot, cause for concern
- if estimates change sign, points to possibly extreme confounding

CD Ch. — ?

- importance of retained explanatory variables probably overstated *p*-values
- procedures not directly linked to final objectives of prediction or explanation
- tends to pick models that are smaller than desirable for prediction LM-2 10.2, LM-1, 8.2
- "should be discouraged" ← using automatically LM-2 10.2

- Criterion-based procedures: AIC, BIC, Mallows C_p, R_a^2 most widely used

"+const"

- $AIC = n\log(RSS/n) + 2p$ balance between fit and simplicity

RSS: residual sum of squares

- $BIC = n\log(RSS/n) + \log(n)p$ choose models with smallest AIC or BIC

Akaike

Bayes

1 for each model

compare AIC_1

AIC_2

want AIC small

- Criterion-based procedures: AIC, BIC, Mallows C_p, R_a^2 most widely used
- $AIC = n\log(RSS/n) + 2p$ balance between fit and simplicity

RSS: residual sum of squares

- $BIC = n\log(RSS/n) + \log(n)p$ choose models with smallest AIC or BIC
- $C_p = RSS_p/\tilde{\sigma}^2 + 2p - n$: estimates average MSE of prediction $C_p \simeq p$

RSS biggest model

- $R_a^2 = 1 - \dfrac{\tilde{\sigma}^2_{model}}{TSS/(n-1)}$

minimizing $\hat{se}(\hat{y})$ means minimizing $\tilde{\sigma}^2_{model}$

choose a model 'automatically'

Summary (lm)

- Criterion-based procedures: AIC, BIC, Mallows C_p, R_a^2 — most widely used
- $AIC = \underline{n \log(RSS/n) + 2p}$ — balance between fit and simplicity

  RSS: residual sum of squares
- $BIC = n \log(RSS/n) + \log(n)p$ — choose models with smallest AIC or BIC
- $C_p = RSS_p/\tilde{\sigma}^2 + 2p - n$: — estimates average MSE of prediction
- $$R_a^2 = 1 - \frac{\tilde{\sigma}^2_{model}}{TSS/(n-1)}$$

  minimizing $\hat{se}(\hat{y})$ means minimizing $\tilde{\sigma}^2_{model}$
- SM has yet another version AIC_c which may be better than AIC for linear models
- C_p and R_a^2 are only useful for linear models; AIC and BIC more general

$-2\ell(\hat{\theta}) + 2p$ (log-likelihood)

Model Building

- Hierarchical principle, testing procedures, criterion-based procedures all provide guidance on how to choose x's
- in a linear regression model — and extensions
- rote application of any of these methods gives little insight into the structure of the model

- Hierarchical principle, testing procedures, criterion-based procedures all provide guidance on how to choose x's
- in a linear regression model

and extensions

- rote application of any of these methods gives little insight into the structure of the model

- Empirical models: "In many fields of study the models used as a basis for interpretation do not have a speical subject-matter base, but, rather, represnt broad patterns of haphazard variation quite widely see in at least approximate form."
- This is typically combined with a specification of the systematic part of the variation, which is often, although not always, the primary focus of interest."
- $\mathrm{E}(y \mid X) = X\beta$

how to choose the x's

"Suppose that, at some point in the analysis, interest is focused on the role of a particular explanatory variable or variables, x_j say, on the response y. Then the following points are relevant.

- the value, standard error, and interpretation of $\hat{\beta}_j$ depends on the other variables in the model

"Suppose that, at some point in the analysis, interest is focused on the role of a particular explanatory variable or variables, x_j say, on the response y. Then the following points are relevant.

- the value, standard error, and interpretation of $\hat{\beta}_j$ depends on the other variables in the model
- relatively mechanical methods of choosing which explanatory variables to use may be helpful in preliminary exploration, especially if p is quite large, but are insecure as a basis for a final interpretation

"Suppose that, at some point in the analysis, interest is focused on the role of a particular explanatory variable or variables, x_j say, on the response y. Then the following points are relevant.

- the value, standard error, and interpretation of $\hat{\beta}_j$ depends on the other variables in the model
- relatively mechanical methods of choosing which explanatory variables to use may be helpful in preliminary exploration, especially if p is quite large, but are insecure as a basis for a final interpretation
- explanatory variables not of direct interest but known to have a substantial effect should be included

"Suppose that, at some point in the analysis, interest is focused on the role of a particular explanatory variable or variables, x_j say, on the response y. Then the following points are relevant.

- the value, standard error, and interpretation of $\hat{\beta}_j$ depends on the other variables in the model
- relatively mechanical methods of choosing which explanatory variables to use may be helpful in preliminary exploration, especially if p is quite large, but are insecure as a basis for a final interpretation
- explanatory variables not of direct interest but known to have a substantial effect should be included
- it may be essential to recognize that several different models are potentially equally effective

"Suppose that, at some point in the analysis, interest is focused on the role of a particular explanatory variable or variables, x_j say, on the response y. Then the following points are relevant.

- the value, standard error, and interpretation of $\hat{\beta}_j$ depends on the other variables in the model
- relatively mechanical methods of choosing which explanatory variables to use may be helpful in preliminary exploration, especially if p is quite large, but are insecure as a basis for a final interpretation
- explanatory variable not of direct interest but known to have a substantial effect should be included
- it may be essential to recognize that several different models are potentially equally effective
- ...

"The choice of a regression model is sometimes presented as a search for a model with as few explanatory variables as reasonably necessary to give an adequate empirical fit. … This approach, which we do not .. in general recommend, may sometimes by appropriate for developing simple empirical prediction equations, although even then the important aspect of the stability ~~to~~ of the prediction equation is not directly addressed"

Example

SM Eg 8.29

- nuclear plant data
- `> library(SMPracticals); data(nuclear); head(nuclear)`

Cox & Snell 1981

y = log(cost)

Table 8.13 Data on light water reactors (LWR) constructed in the USA (Cox and Snell, 1981, p. 81). The covariates are `date` (date construction permit issued), `T1` (time between application for and issue of permit), `T2` (time between issue of operating license and construction permit), `capacity` (power plant capacity in MWe), `PR` (=1 if LWR already present on site), `NE` (=1 if constructed in north-east region of USA), `CT` (=1 if cooling tower used), `BW` (=1 if nuclear steam supply system manufactured by Babcock–Wilcox), `N` (cumulative number of power plants constructed by each architect-engineer), `PT` (=1 if partial turnkey plant).

	cost	date	T_1	T_2	capacity	PR	NE	CT	BW	N	PT
1	460.05	68.58	14	46	687	0	1	0	0	14	0
2	452.99	67.33	10	73	1065	0	0	1	0	1	0
3	443.22	67.33	10	85	1065	1	0	1	0	1	0
4	652.32	68.00	11	67	1065	0	1	1	0	12	0
5	642.23	68.00	11	78	1065	1	1	1	0	12	0
6	345.39	67.92	13	51	514	0	1	1	0	3	0
7	272.37	68.17	12	50	822	0	0	0	0	5	0
8	317.21	68.42	14	59	457	0	0	0	0	1	0
9	457.12	68.42	15	55	822	1	0	0	0	5	0
10	690.19	68.33	12	71	792	0	1	1	1	2	0
11	350.63	68.58	12	64	560	0	0	0	0	3	0
12	402.59	68.75	13	47	790	0	1	0	0	6	0
13	412.18	68.42	15	62	530	0	0	1	0	2	0
14	495.58	68.92	17	52	1050	0	0	0	0	7	0
15	394.36	68.92	13	65	850	0	0	0	1	16	0
16	423.32	68.42	11	67	778	0	0	0	0	3	0
17	712.27	69.50	18	60	845	0	1	0	0	17	0
18	289.66	68.42	15	76	530	1	0	1	0	2	0
19	881.24	69.17	15	67	1090	0	0	0	0	1	0
20	490.88	68.92	16	59	1050	1	0	0	0	8	0
21	567.79	68.75	11	70	913	0	0	1	1	15	0
22	665.99	70.92	22	57	828	1	1	0	0	20	0
23	621.45	69.67	16	59	786	0	0	1	0	18	0
24	608.80	70.08	19	58	821	1	0	0	0	3	0
25	473.64	70.42	19	44	538	0	0	1	0	19	0
26	697.14	71.08	20	57	1130	0	0	1	0	21	0
27	207.51	67.25	13	63	745	0	0	0	0	8	1
28	288.48	67.17	9	48	821	0	0	1	0	7	1
29	284.88	67.83	12	63	886	0	0	0	1	11	1
30	280.36	67.83	12	71	886	1	0	0	1	11	1
31	217.38	67.25	13	72	745	1	0	0	0	8	1
32	270.71	67.83	7	80	886	1	0	0	1	11	1

26

6

... Example

	Full model		Backward		Forward	
	Est (SE)	t	Est (SE)	t	Est (SE)	t
Constant	−14.24 (4.229)	−3.37	−13.26 (3.140)	−4.22	−7.627 (2.875)	−2.66
date	0.209 (0.065)	3.21	0.212 (0.043)	4.91	0.136 (0.040)	3.38
log(T1)	0.092 (0.244)	0.38				
log(T2)	0.290 (0.273)	1.05				
log(cap)	0.694 (0.136)	5.10	0.723 (0.119)	6.09	0.671 (0.141)	4.75
PR	−0.092 (0.077)	−1.20				
NE	0.258 (0.077)	3.35	0.249 (0.074)	3.36		
CT	0.120 (0.066)	1.82	0.140 (0.060)	2.32		
BW	0.033 (0.101)	0.33				
log(N)	−0.080 (0.046)	−1.74	−0.088 (0.042)	−2.11		
PT	−0.224 (0.123)	−1.83	−0.226 (0.114)	−1.99	−0.490 (0.103)	−4.77
Residual SE (df)	0.164 (21)		0.159 (25)		0.195 (28)	

Table 8.14 Parameter estimates and standard errors for linear models fitted to nuclear plants data; forward and backward indicate models fitted by forward selection and backward elimination.

C & Snell

– could also use `stepAIC` or `leaps::regsubsets`

LM-2 10.3, LM-1 8.3

- transformation of variables: `cost, T1, T2, cap, cum.n` all converted to log
- "partly to lead to unit-free parameters whose values can be interpreted in terms of power-law relations between the original variables" Cox & Snell
- "Costs are typically relative. Moreover large costs are likely to vary more than small ones. For consistency we also take logs of the other quantitative covariates" Davison

$E(y|x) = \beta x^{\alpha}$

$\log E(y|x) \simeq \log\beta + \alpha \underbrace{\log x}$

IJALM

... Example

- transformation of variables: `cost`, `T1`, `T2`, `cap`, `cum.n` all converted to log
- "partly to lead to unit-free parameters whose values can be interpreted in terms of power-law relations between the original variables" Cox & Snell
- "Costs are typically relative. Moreover large costs are likely to vary more than small ones. For consistency we also take logs of the other quantitative covariates" Davison

- backward elimination leaves six variables with residual mean square $0.0253 = 0.159^2$; none of the eliminated variables is significant if re-introduced

stepwise

- transformation of variables: `cost`, `T1`, `T2`, `cap`, `cum.n` all converted to log
- "partly to lead to unit-free parameters whose values can be interpreted in terms of power-law relations between the original variables" Cox & Snell
- "Costs are typically relative. Moreover large costs are likely to vary more than small ones. For consistency we also take logs of the other quantitative covariates" Davison

- backward elimination leaves six variables with residual mean square $0.0253 = 0.159^2$; none of the eliminated variables is significant if re-introduced

- variable `PT` is unbalanced 6 1's 26 0's
- check on the model includes interaction with `PT` one variable at a time

T1 / T2 ↗ out

after stepwise

e.g.

```
> nuclear.lm3 <- lm(log(cost) ~ date + log(cap) + NE + CT + log(cum.n) + PT,
data = nuclear); nuclear.lm3$coef

(Intercept)        date    log(cap)          ne          ct  log(cum.n)
  -13.26031     0.21241     0.72341     0.24902     0.14039    -0.08758
         pt
   -0.22610
> update(nuclear.lm3, . ~ . + pt*log(cap))$coef
(Intercept)        date    log(cap)          ne          ct  log(cum.n)
  -13.08645     0.21044     0.71761     0.24841     0.13998    -0.08683
         pt log(cap):pt
   -2.18759     0.29159
```

allow $\hat\beta$ for l(capacity) to change with PT

↓ factor var.

$$\beta_0 + \beta_1 d_i + \beta_2 x_i + \beta_3 (d_i x_i)$$

$p > n$

- if $p > n$ then $X^T X$ is not invertible
- β is not estimable
- residual sum of squares will be 0 with n explanatory variables
- no reduction in complexity; nothing learned about the relationship between y and x

$p > n$

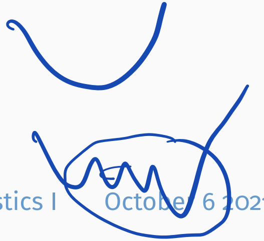
- if $p > n$ then $X^T X$ is not invertible
- β is not estimable
- residual sum of squares will be 0 with n explanatory variables
- no reduction in complexity; nothing learned about the relationship between y and x

- we expect that few variables are "active", i.e. are useful for explaining the variation in y
- number of active variables usually called s, assumed $s < n$ — also $s << p$
- how do we find them?

- if $p > n$ then $X^T X$ is not invertible
- β is not estimable
- residual sum of squares will be 0 with n explanatory variables
- no reduction in complexity; nothing learned about the relationship between y and x

- we expect that few variables are "active", i.e. are useful for explaining the variation in y
- number of active variables usually called s, assumed $s < n$ — also $s << p$
- how do we find them?

RSS



-
$$\arg\min_{\beta}\{\underbrace{(y - X\beta)^T(y - X\beta)}_{\text{RSS}} + \lambda||\beta||_0\}$$

-
$$||\beta_0|| = \#\{j : \beta_j \neq 0\}$$

Constrain to have few active

$p > n$

components

- $$\arg\min_{\beta}\{(y - X\beta)^T(y - X\beta) + \lambda||\beta||_0\}$$
- $$||\beta_0|| = \#\{j : \beta_j \neq 0\}$$
- non-convex optimization; a convex relaxation of this problem is
  $$\arg\min_{\beta}\{(y - X\beta)^T(y - X\beta) + \lambda||\beta||_1\}$$
- $$||\beta||_1 = \sum_j |\beta_j|$$

$$||\beta||_2 = \left(\sum \beta_j^2\right)^{1/2}$$

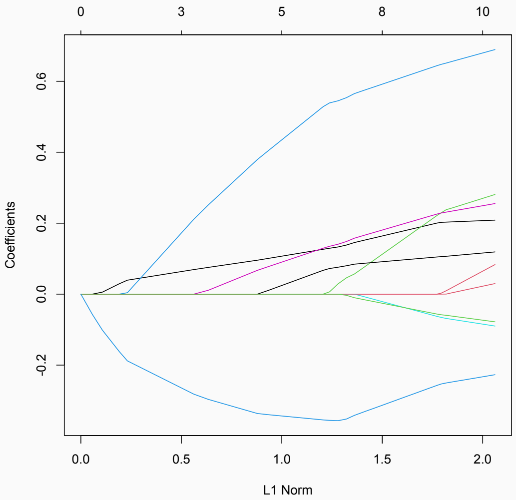
$p > n$

- $$\arg\min_{\beta}\{(y - X\beta)^T(y - X\beta) + \lambda||\beta||_0\}$$
- $$||\beta_0|| = \#\{j : \beta_j \neq 0\}$$
- non-convex optimization; a convex relaxation of this problem is
$$\arg\min_{\beta}\{(y - X\beta)^T(y - X\beta) + \lambda||\beta||_1\}$$
- $$||\beta||_1 = \sum_j |\beta_j|$$
- the resulting estimate $\hat{\beta}_\lambda$ is called the Lasso estimate
- has many components $\hat{\beta}_{\lambda,k} = 0$
- there are many other approaches to regression with $p > n$

$p > n$

```
> require(glmnet)
> x <- model.matrix(nuclear.lm)
> y <- log(nuclear$cost)
> nuclear.lasso <- glmnet(x,y)
> cv.glmnet(x,y)
...
      Lambda Index Measure      SE Nonzero
min  0.0295    24  0.0367 0.0105       6
1se  0.0566    17  0.0462 0.0115       5
> nuclear.lasso2 <- glmnet(x,y,lambda=0.0566)
> coef(nuclear.lasso2)
0.1055 . . 0.4276 . 0.08728 0.02109 . . -0.3426
```

date T1 T2 log(cap) PT



$\hat{var}(\hat{\beta}_{\lambda^*, Lasso})$?

- common objectives

Design of Studies

- common objectives
- to avoid systematic error, that is distortion in the conclusions arising from sources that do not cancel out in the long run

- common objectives
- to avoid systematic error, that is distortion in the conclusions arising from sources that do not cancel out in the long run
- to reduce the non-systematic (random) error to a reasonable level by replication and other techniques

Design of Studies

- common objectives
- to avoid systematic error, that is distortion in the conclusions arising from sources that do not cancel out in the long run
- to reduce the non-systematic (random) error to a reasonable level by replication and other techniques
- to estimate realistically the likely uncertainty in the final conclusions

- common objectives
- to avoid systematic error, that is distortion in the conclusions arising from sources that do not cancel out in the long run
- to reduce the non-systematic (random) error to a reasonable level by replication and other techniques
- to estimate realistically the likely uncertainty in the final conclusions
- to ensure that the scale of effort is appropriate

… design of studies

- we concentrate largely on the careful analysis of individual studies

- in most situations synthesis of information from different investigations is needed

- but even there the quality of individual studies remains important
- examples include overviews (such as the Cochrane reviews)

… design of studies

- we concentrate largely on the careful analysis of individual studies

- in most situations synthesis of information from different investigations is needed

- but even there the quality of individual studies remains important
- examples include overviews (such as the Cochrane reviews)

- in some areas new investigations can be set up and completed relatively quickly; design of individual studies may then be less important

… design of studies

- formulation of a plan of analysis
- establish and document that proposed data are capable of addressing the research questions of concern
- main configurations of answers likely to be obtained should be set out
- level of detail depends on the context

… design of studies

- formulation of a plan of analysis
- establish and document that proposed data are capable of addressing the research questions of concern
- main configurations of answers likely to be obtained should be set out
- level of detail depends on the context

- even if pre-specified methods must be used, it is crucial not to limit analysis
- planned analysis may be technically inappropriate
- more controversially, data may suggest new research questions or replacement of objectives
- latter will require confirmatory studies

Unit of study and analysis

- smallest subdivision of experimental material that may be assigned to a treatment

  context: Expt
- Example: RCT – unit may be a patient, or a patient-month (in crossover trial)
- Example: public health intervention – unit is often a community/school/...

Unit of study and analysis

- smallest subdivision of experimental material that may be assigned to a treatment

  context: Expt
- Example: RCT – unit may be a patient, or a patient-month (in crossover trial)
- Example: public health intervention – unit is often a community/school/...

- in investigations that are not randomized, it may be helpful to consider what the primary unit of analysis would have been, had a randomized experiment been feasible
- the unit of analysis may not be the unit of interpretation – ecological bias

  systematic difference between impact of x at different levels of aggregation

Unit of study and analysis

- smallest subdivision of experimental material that may be assigned to a treatment
  context: Expt
- Example: RCT – unit may be a patient, or a patient-month (in crossover trial)
- Example: public health intervention – unit is often a community/school/...

- in investigations that are not randomized, it may be helpful to consider what the primary unit of analysis would have been, had a randomized experiment been feasible
- the unit of analysis may not be the unit of interpretation – ecological bias
  systematic difference between impact of x at different levels of aggregation

- on the whole, limited detail is needed in examining the variation within the unit of study

Types of observational studies

- secondary analysis of data collected for another purpose
- estimation of a some feature of a defined population (could in principle be found exactly)
- tracking across time of such features

Types of observational studies

- secondary analysis of data collected for another purpose
- estimation of a some feature of a defined population (could in principle be found exactly)
- tracking across time of such features

- study of a relationship between features, where individuals may be examined
    - at a single time point
    - at several time points for different individuals
    - at different time points for the same individual

Types of observational studies

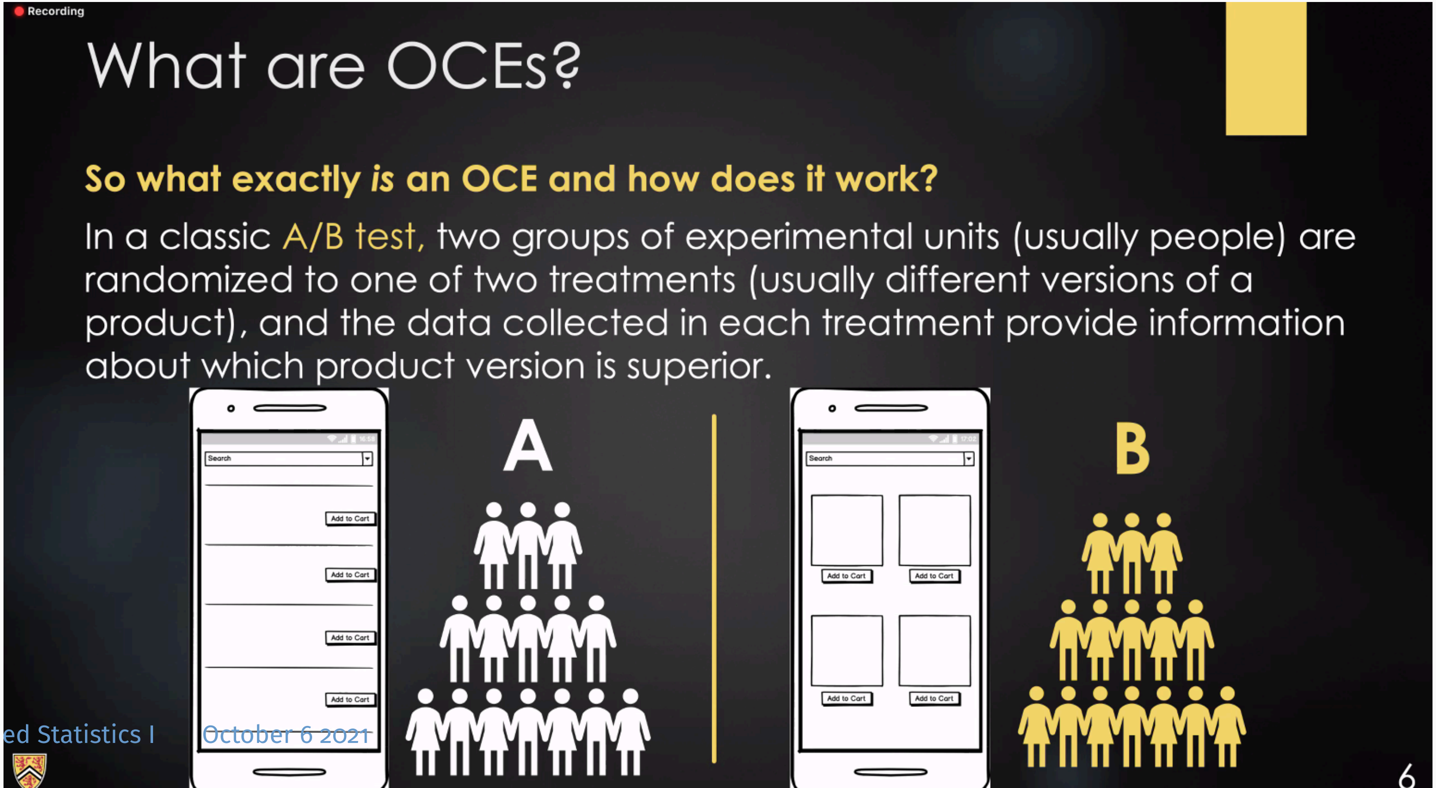
- secondary analysis of data collected for another purpose
- estimation of a some feature of a defined population (could in principle be found exactly)
- tracking across time of such features

- study of a relationship between features, where individuals may be examined
  - at a single time point
  - at several time points for different individuals
  - at different time points for the same individual

- census
- meta-analysis: statistical assessment of a collection of studies on the same topic

Computational advertising

David Banks, Duke University: The statistical challenges of computational advertising


Recording
What are OCEs?
So what exactly is an OCE and how does it work?
In a classic A/B test, two groups of experimental units (usually people) are randomized to one of two treatments (usually different versions of a product), and the data collected in each treatment provide information about which product version is superior.
Search
Add to Cart
Add to Cart
Add to Cart
Add to Cart
A
Search
Add to Cart
Add to Cart
Add to Cart
Add to Cart
B
6

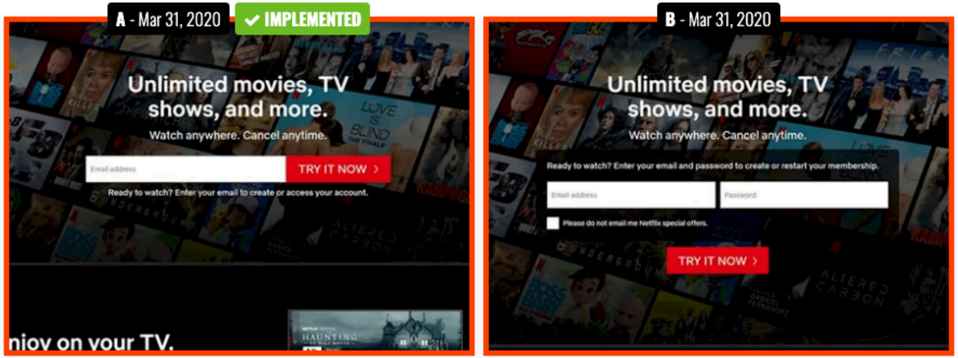

Recording
What are OCEs?
What kinds of things are companies experimenting with?
User acquisition funnels
User engagement mechanics
User retention mechanics
Email promotions and headlines
Website layout
Esthetic features
Checkout experience
Freemium conversion
Branding
Ad Campaigns
Call to action language
ML algorithms
For some real-life examples, checkout the "Leaks" on GoodUI:
https://goodui.org/leaks/


Recording
What are OCEs?
Concrete Examples:
Obama's 2008 campaign increased donations by $60M using a factorial experiment[6]
OBAMA'08
CHANGE
WE CAN BELIEVE IN
JOIN THE MOVEMENT
Email Address
Zip Code
LEARN MORE

https://goodui.org/leaks/

Leak #53 from Netflix.com | May 25, 2020 Home & Landing

Netflix A/B Tests Displaying A Password Field Which Fails And Gets Rejected



It looks like Netflix has been iterating on showing additional fields upfront on their homepage. After they succeeded at displaying an email address upfront, this experiment now takes next step of showing a password field. The result of the leaked experiment however suggests a negative outcome as they reverted back to the control version - without the visible password. View Leak paper

Stanford talk

$\hat{y}_+ =$

$$\mathrm{var}\{ y_+ - \hat{\gamma}_0 - \hat{\gamma}_1 (x_+ - \bar{x}) \} = \cdots$$

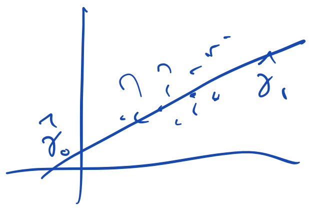
$$\widehat{\mathrm{var}}\{ \} = \tilde{\sigma}^2 \cdots$$

$$\frac{y_+ - \hat{\gamma}_0 - \hat{\gamma}_1 (x_+ - \bar{x})}{\sqrt{\widehat{\mathrm{var}}}} \sim t_{n-2}$$

$$\mathcal{X} = \{x_* : t_1 \le T(x^*) \le t_2\}$$

$$\text{length}(\mathcal{X}) = (t_2 - t_1) \qquad \Pr(t_2 - t_1) = \infty$$

$$= \Pr(t_2 = \infty \text{ or } t_1 = -\infty)$$



$$x_+ = \frac{y_+ - \text{something}}{\text{something}}$$

$$\Pr \; T(x_*) \le$$

$$\lim_{n\to\infty} \Pr(t_2 = \infty \text{ or } t_1 = -\infty)$$

$$Y_+ = \gamma_0 + \gamma_1(x_+ - \bar{x}) + \varepsilon_+ \qquad \varepsilon_+ \sim N(0, \sigma^2)$$

indep of $\varepsilon_1 \dots \varepsilon_n$

$$\hat{y}_+ = \hat{\gamma}_0 + \hat{\gamma}_1(x_+ - \bar{x})$$

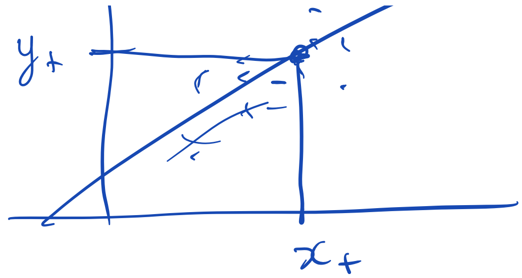
$$\hat{\text{var}}(Y_+ - \hat{y}_+) = \underset{\uparrow \; \frac{RSS}{n-2}}{\tilde{\sigma}^2}\left(1 + \frac{1}{n} + \frac{(x_+ - \bar{x})^2}{S_x^2}\right)$$

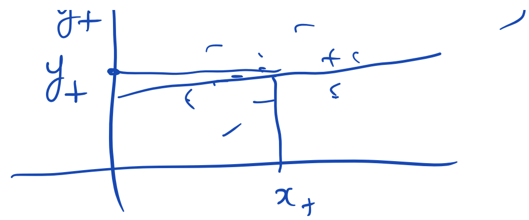
$$T(x^+) = \frac{Y_+ - (\hat{\gamma}_0 + \hat{\gamma}_1(x_+ - \bar{x}))}{\sqrt{\hat{\text{var}}}}$$

$$\Pr\{T(x^+) \le k\} = \Pr\{t_{n-2} \le k\}$$

change $\{t_1 \le T(x_+) \le t_2\} \Rightarrow x_+ \in (\;,\;)$

$$\Pr\{t_1\sqrt{\hat{\text{var}}} + \hat{\gamma}_0 + \hat{\gamma}_1(x_+ - \bar{x}) \le Y_+ \le t_2\sqrt{\hat{\text{var}}} + \dots\}$$

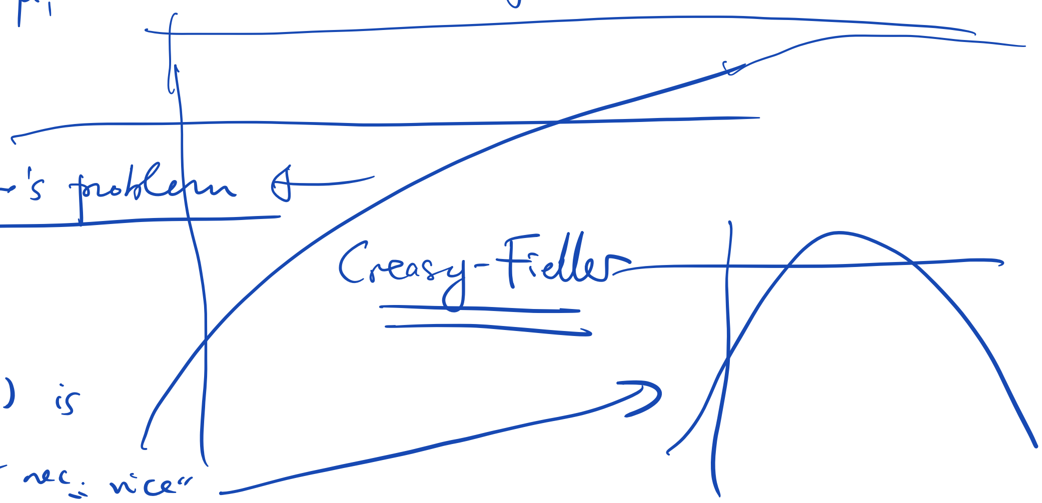




$X_1 \sim N(\mu_1, 1)$ $\qquad X_2 \sim N(\mu_2, 1)$

$\theta = \dfrac{\mu_2}{\mu_1}$ $\qquad \hat{\theta} = \dfrac{X_2}{X_1}$ $\qquad \cancel{X_1 = 0}$

μ_1 ~ nr 0 then CI for θ can be $(-\infty, \infty)$



Fieller's problem

→ (1)

$g(u)$ is "not nec. nice"

Creasy-Fieller

$$Y_+ = \gamma_0 + \gamma_1 (x_+ - \bar{x}) + \varepsilon_+$$

once y_+ obs'd → interval $\{x_* : t_1 \le T(x_*) \le t_2\}$

w. some prob. y_+ will be such that ↑ is ∞

$\underset{Y_+ \mid x_+}{\text{prob.}}$

$\rightarrow \bar{X} \sim N(\mu, 1) \qquad \hat{\mu} = \bar{X}$

$$\left(\bar{X} - z_{\alpha/2} \cdot \cancel{1} \le \mu \le \bar{X} + z_{\alpha/2} \cdot \cancel{1} \right)$$

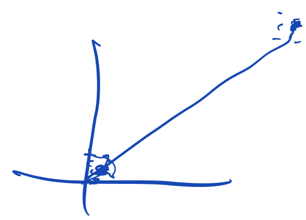
$$Pr_{\bar{X}}\left(\qquad \right) = 1-2\alpha$$

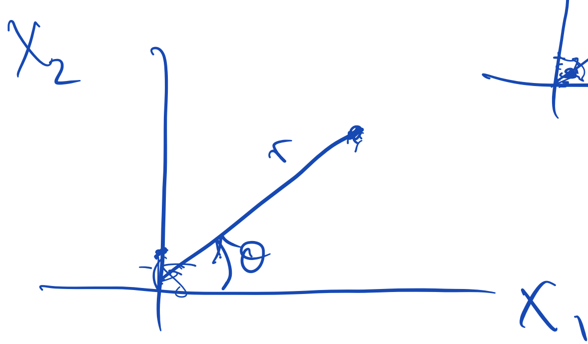
$$Pr_{Y_+|x_+}\left\{ t_1 \le T_{Y_+}(x_+) \le t_2 \right\} = 1-2\alpha$$

$$\left[t_1 \le T_{y_+}(x_+) \le t_2 \right] \quad 1-2\alpha$$

CI for x_+

$$\left\{-b \pm \sqrt{b^2 - 4ac}\right\} / 2a$$

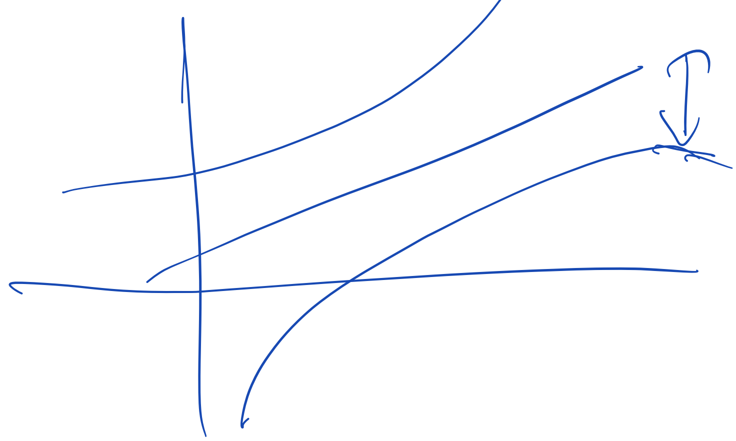




$$\frac{\mu_1}{\mu_2} = \theta$$

$\mu_1 = r\cos\theta$

$\mu_2 = r\sin\theta \qquad \frac{\mu_1}{\mu_2} = \tan\theta$



$(x-\bar{x})^2$