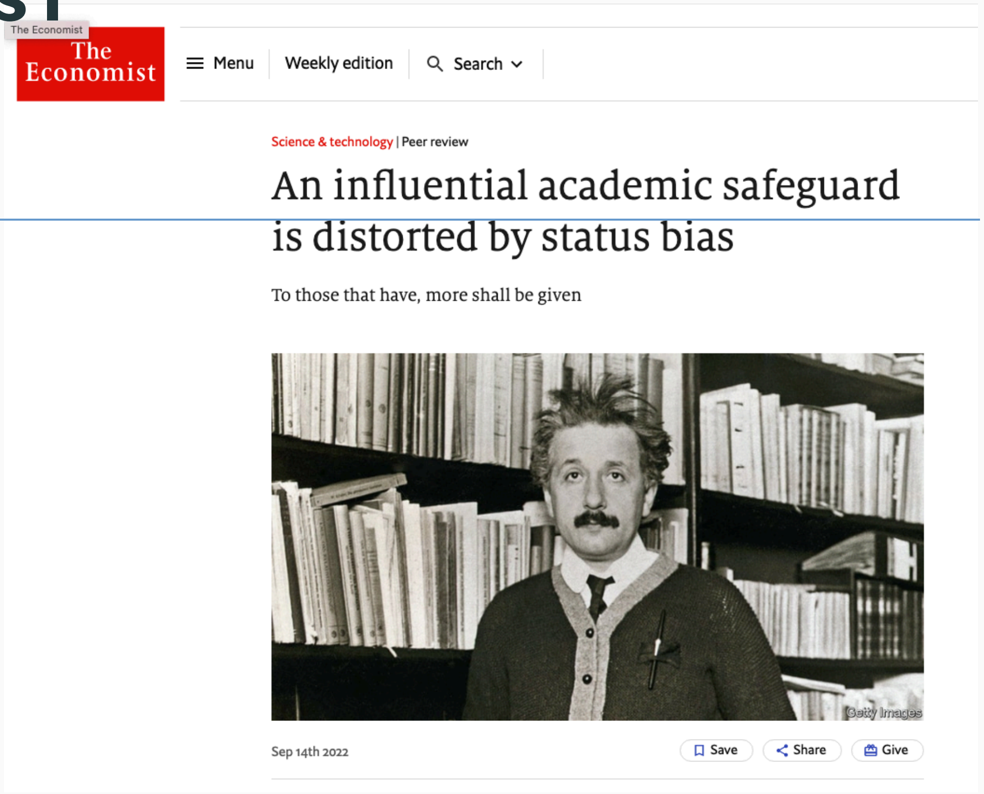
Methods of Applied Statistics I

STA2101H F LEC9101

Week 3

September 28 2022



Science & technology | Peer review

An influential academic safeguard is distorted by status bias

To those that have, more shall be given

Sep 14th 2022

Today

Start Recording

1. Upcoming events
2. Comments re HW
3. Linear Regression Part 3: recap, checking model assumptions, collinearity, model-building, $p > n$
4. In the News

Today

Start Recording

1. Upcoming events
2. Comments re HW
3. Linear Regression Part 3: recap, checking model assumptions, collinearity, model-building, $p > n$
4. In the News

Upcoming

- September 29: CANSSI Ontario Research Day   Schedule and Registration

- Distinguished Lecture Series in Statistical Sciences
- Xihong Lin, Harvard U   Details and Registration

- September 29 3.30 89 Chestnut Street, 3rd Floor
  Lessons learned from the COVID-19 Pandemic: a statistician's reflection

- September 30 3.30 UY9014
  Ensemble methods for testing a global null hypothesis

- September 30 1.00 Zoom data_4_lyf
  Toronto Data workshop
  "How the NFL blocks black coaches"

2022 DLSS: Xihong Lin

Professor, Department of BiostatisticsCoordinating Director, Program in Quantitative Genomics; Harvard T.H. Chan School of Public Health; Professor of Statistics, Department of Statistics, Harvard University

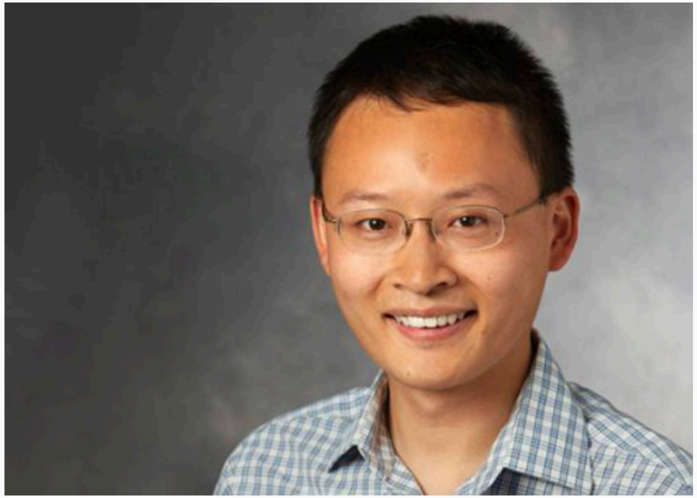
… upcoming

- October 3 3:30 Data Science ARES online

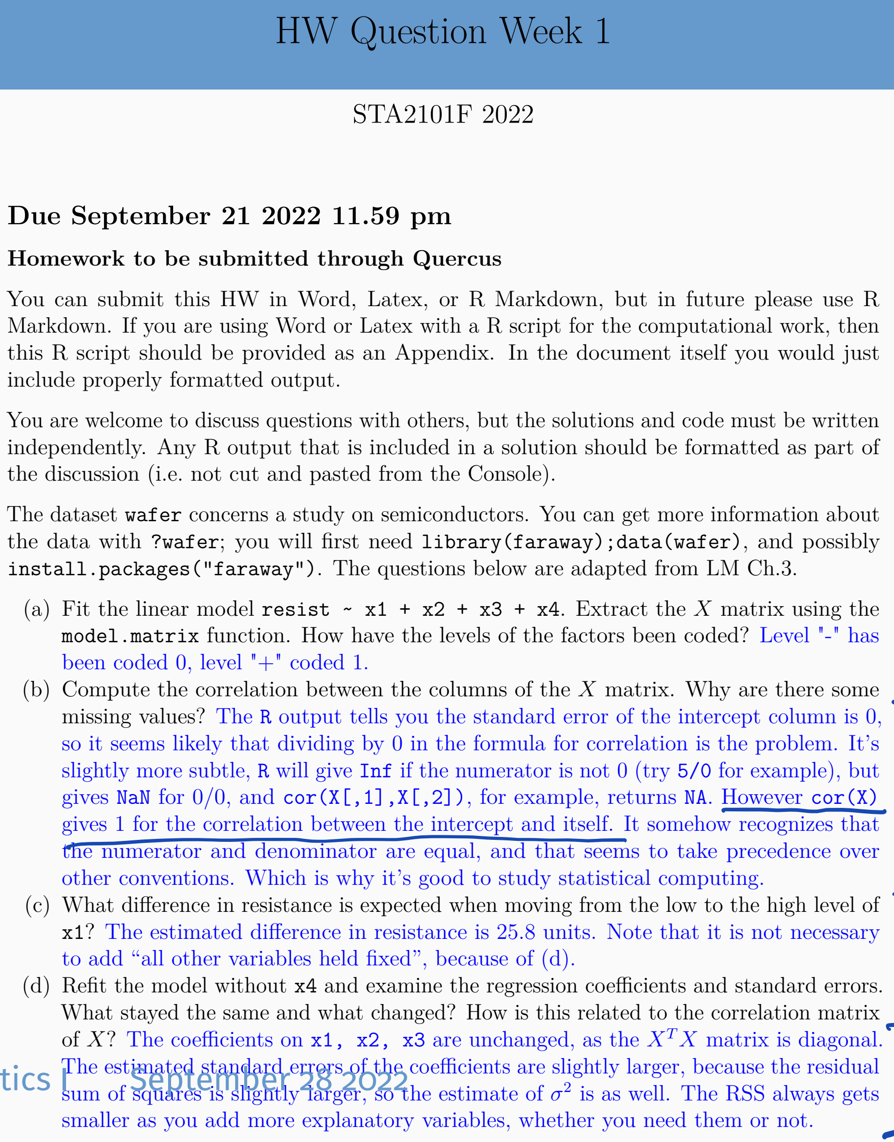
  James Zou, Stanford

  “AI for clinical trials and clinical trials for AI”

Register here



HW 1



HW Question Week 1

STA2101F 2022

Due September 21 2022 11.59 pm

Homework to be submitted through Quercus

You can submit this HW in Word, Latex, or R Markdown, but in future please use R Markdown. If you are using Word or Latex with a R script for the computational work, then this R script should be provided as an Appendix. In the document itself you would just include properly formatted output.

You are welcome to discuss questions with others, but the solutions and code must be written independently. Any R output that is included in a solution should be formatted as part of the discussion (i.e. not cut and pasted from the Console).

The dataset `wafer` concerns a study on semiconductors. You can get more information about the data with `?wafer`; you will first need `library(faraway);data(wafer)`, and possibly `install.packages("faraway")`. The questions below are adapted from LM Ch.3.

(a) Fit the linear model `resist ~ x1 + x2 + x3 + x4`. Extract the X matrix using the `model.matrix` function. How have the levels of the factors been coded? Level "-" has been coded 0, level "+" coded 1.

(b) Compute the correlation between the columns of the X matrix. Why are there some missing values? The `R` output tells you the standard error of the intercept column is 0, so it seems likely that dividing by 0 in the formula for correlation is the problem. It's slightly more subtle, `R` will give `Inf` if the numerator is not 0 (try `5/0` for example), but gives `NaN` for 0/0, and `cor(X[,1],X[,2])`, for example, returns `NA`. However `cor(X)` gives 1 for the correlation between the intercept and itself. It somehow recognizes that the numerator and denominator are equal, and that seems to take precedence over other conventions. Which is why it's good to study statistical computing.

(c) What difference in resistance is expected when moving from the low to the high level of `x1`? The estimated difference in resistance is 25.8 units. Note that it is not necessary to add "all other variables held fixed", because of (d).

(d) Refit the model without `x4` and examine the regression coefficients and standard errors. What stayed the same and what changed? How is this related to the correlation matrix of X? The coefficients on `x1, x2, x3` are unchanged, as the X^TX matrix is diagonal. The estimated standard errors of the coefficients are slightly larger, because the residual sum of squares is slightly larger, so the estimate of σ^2 is as well. The RSS always gets smaller as you add more explanatory variables, whether you need them or not.

ntbc

bonus, rmd, soln's ✓

Linear regression recap

- Analysis of variance: $y^Ty = \overset{\text{left over}}{(y - X\hat{\beta})^T(y - X\hat{\beta})} + \overset{\text{model}}{\hat{\beta}^TX^TX\hat{\beta}}$

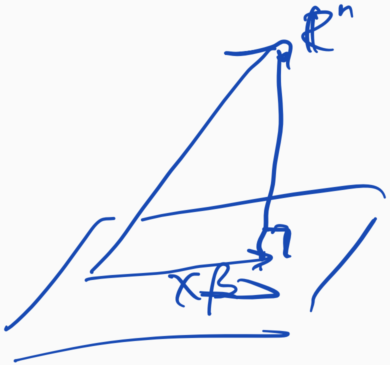
Source	DF	SS	MS
Regression	$p-1$	SS_{REG}	$RegMS = SS_{REG}/(p-1)$
Residual	$n-p$	RSS	$ResMS = RSS/(n-p)$
Total (corrected)	$n-1$	TSS	

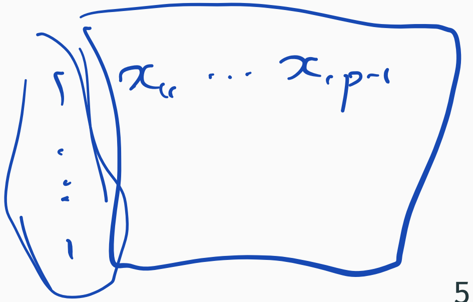
$$\sum(y_i - \bar{y})^2$$

$$\underbrace{(y - \bar{y}\underset{\sim}{1})^T}(y - \bar{y}\underset{\sim}{1})$$

anova

$\cancel{\beta_0}\ \beta_1 \cdots \beta_{p-1}$





Linear regression recap

- Analysis of variance: $y^T y = (y - X\hat{\beta})^T (y - X\hat{\beta}) + \hat{\beta}^T X^T X \hat{\beta}$

Source	DF	SS	MS	
Regression	$p - 1$	SS_{REG}	$RegMS = SS_{REG}/(p-1)$	model
Residual	$n - p$	RSS	$ResMS = RSS/(n-p)$	error
Total (corrected)	$n - 1$	TSS		

$$\rightarrow F = \frac{RegMS}{ResMS} \sim F_{p-1,n-p} \quad \text{under } H_0 : \beta_1 = \dots = \beta_{p-1} = 0$$

- regression SS can be further partitioned

depends on the order

$$\frac{SS_{reg}/(p-1)}{RSS/(n-p)} =$$

$$\frac{\partial}{\partial\beta}\left\{(y - X\beta)^{T}(y - X\beta)\right\}\Bigg|_{\beta=\hat{\beta}} = 0$$

$$\begin{pmatrix} \frac{\partial}{\partial\beta_1} \\ \vdots \\ \frac{\partial}{\partial\beta_p} \end{pmatrix} \quad - - -$$

… Linear regression recap

- same principle can be used to test for sets of variables
- or for testing any linear constraint on β $H_0: A\beta = c$

$A\beta = c$

- numerator degrees of freedom for F-statistic depend on the rank of A

Fit full model $\rightarrow$ RSS_{full}

Fit Hyp. model RSS_{red} $\sim$ $F_{p-q,n-p}$ under H_0

reduced

$(p-q)$ restrictions

$\sim$

$$\frac{(RSS_{red} - RSS_{Full}) / \overset{(p-q)}{(n-q-(n-p))}}{RSS_{full} / (n-p)}$$

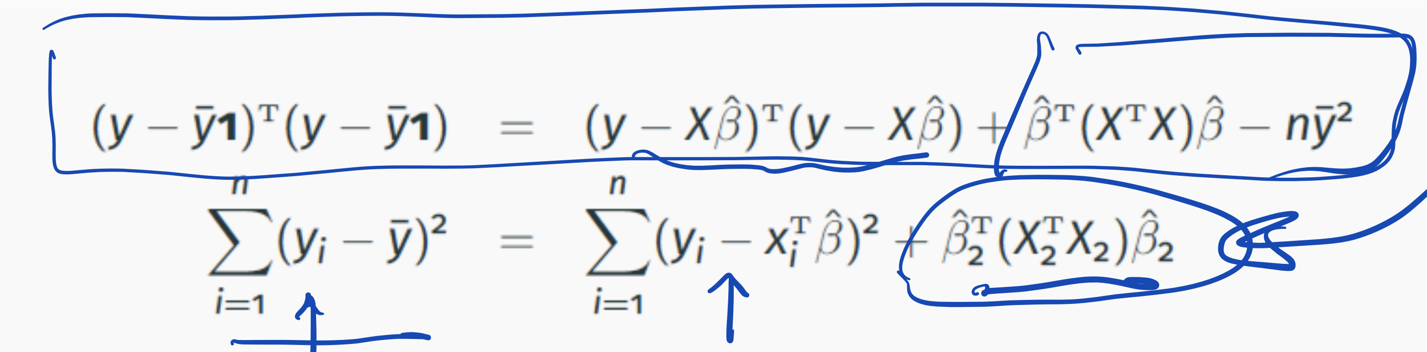
... Linear regression recap

- same principle can be used to test for sets of variables
- or for testing any linear constraint on β $A\beta = c$
- numerator degrees of freedom for F-statistic depend on the rank of A $(p-q)$
- $$F_{1,\nu} \equiv t^2_{\nu}$$ Same as t-test when $H_0: \beta_5 = 0$
- sometimes only an F-test can be used to assess the effect of an explanatory variable

when?

Q on Piazza

$$y^T y - n\bar{y}^2 = (y-\hat{y})^T(y-\hat{y}) + \boxed{\hat\beta^T X^T X \hat\beta - n\bar{y}^2}$$



$$(y-\bar{y}\mathbf{1})^T(y-\bar{y}\mathbf{1}) = (y-X\hat\beta)^T(y-X\hat\beta)+\hat\beta^T(X^TX)\hat\beta - n\bar{y}^2$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-x_i^T\hat\beta)^2+\hat\beta_2^T(X_2^TX_2)\hat\beta_2$$

?

$$= \sum y_i^2 - n\bar{y}^2$$

✓

$$\beta = (\beta_0, \dots, \beta_{p-1})$$

$$\beta_2 = (\beta_1, \dots, \beta_{p-1})$$

~~$p\times p$~~

$$\hat\beta^T_{1\times p} \quad X^TX_{p\times p} \quad \hat\beta_{p\times 1}$$

$$\hat\beta_0^2 \sum \dots \qquad \hat\beta_{p-1}^2 \sum x_{i,p-1}^2 \qquad - n\bar{y}^2$$

Factor variables

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \dots - \hat{\beta}_{p-1} \bar{x}_{p-1}$$

- *F*-tests are used when the columns to be removed form a group

- if a covariate is a factor, i.e. categorical, then `lm` will construct a set of dummy variables as part of the model matrix

- these variables should either all be in, or all be out

in most cases

Factor variables

- *F*-tests are used when the columns to be removed form a group

- if a covariate is a factor, i.e. categorical, then `lm` will construct a set of dummy variables as part of the model matrix

- these variables should either all be in, or all be out

in most cases

6, 7, 8, 9

-
```
  prostate$gleason_factor <- factor(prostate$gleason)
  levels(prostate$gleason_factor)
[1] "6" "7" "8" "9"
  model_fac <- lm(lpsa ~ .-gleason, data=prostate)
```

... factor variables

```
  model_fac <- lm(lpsa ~ .-gleason, data=prostate)
  sumary(model_fac)
  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.91328    0.84084    1.09   0.2804
lcavol            0.56999    0.09010    6.33  1.1e-08
lweight           0.46879    0.16961    2.76   0.0070
age              -0.02175    0.01136   -1.91   0.0589
lbph              0.09968    0.05898    1.69   0.0946
svi               0.74588    0.24740    3.01   0.0034
lcp              -0.12511    0.09559   -1.31   0.1941
pgg45             0.00499    0.00467    1.07   0.2885
gleason_factor7   0.26761    0.21942    1.22   0.2259
gleason_factor8   0.49682    0.76927    0.65   0.5201
gleason_factor9  -0.05621    0.50020   -0.11   0.9108

n = 97, p = 11, Residual SE = 0.70, R-Squared = 0.67
```

→ diff. from level 6

→ " " "

→ " " "

check `model.matrix`

... factor variables

```
model_nog <- lm(lpsa ~ . - ~~gleason~~ - gleason_factor, data = prostate)

anova(model_fac, model_nog) # compare two models

Analysis of Variance Table

Model 1: lpsa ~ (lcavol + lweight + age + lbph + svi + lcp + gleason +
    pgg45 + gleason_factor) - gleason - gleason_factor
Model 2: lpsa ~ (lcavol + lweight + age + lbph + svi + lcp + gleason +
    pgg45 + gleason_factor) - gleason
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     89 44.2
2     86 42.7  3      1.48 0.99    0.4
```

$$\frac{1.48/3}{42.7/86} =$$

data don't contradict

→ $H_0: \beta_8 = \beta_9 = \beta_{10} = 0$

Model checking

LM-2 Ch. 6, LM-1 Ch. 4; SM 8.6

```
plot(model1)
```

, 6

$\varepsilon_i = y_i - E(y_i) = y_i - x_i^T\beta$

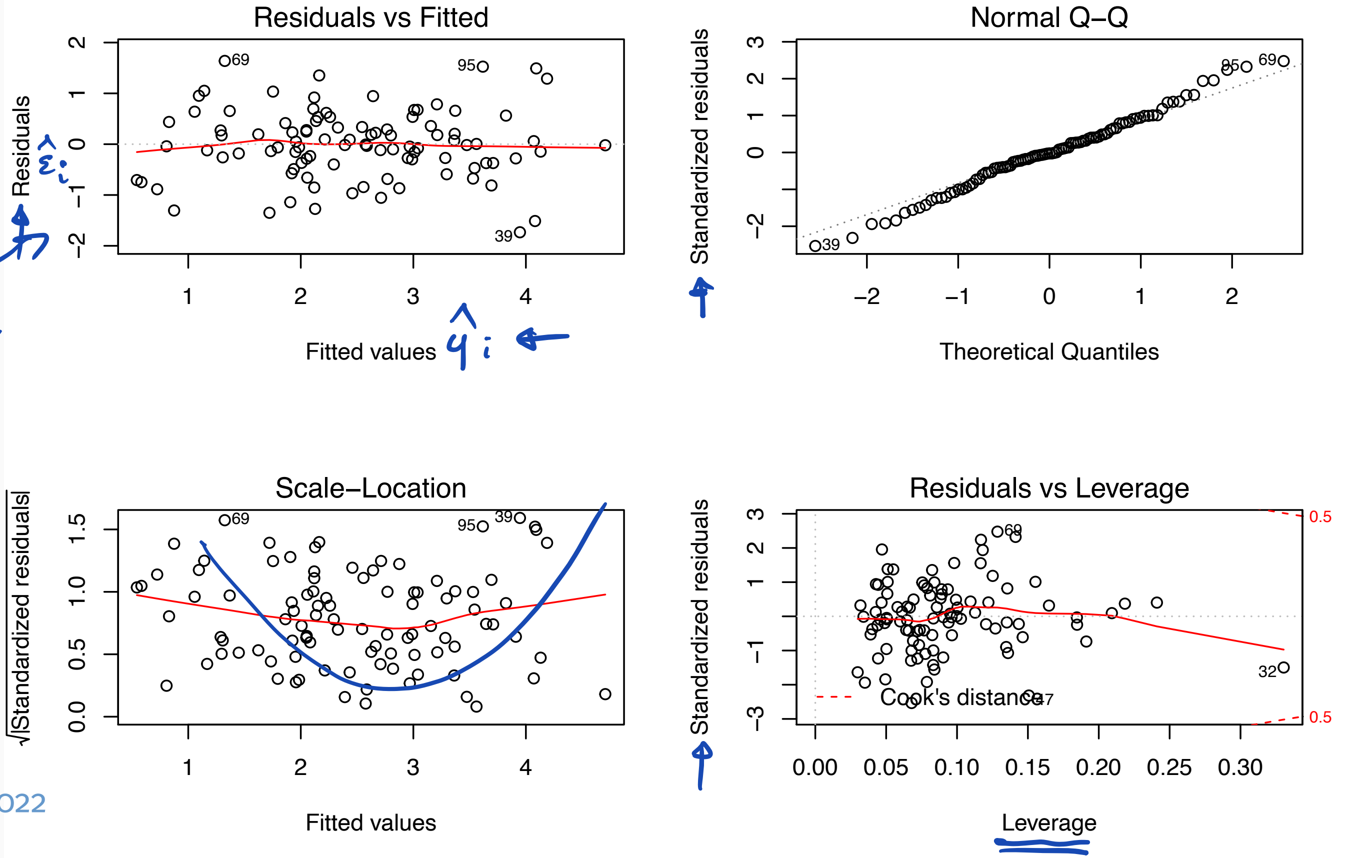
$\hat{\varepsilon}_i = y_i - \hat{y}_i$

$= y_i - x_i^T\hat{\beta}$

constant variance

$\varepsilon_i \sim (0, \sigma^2)$

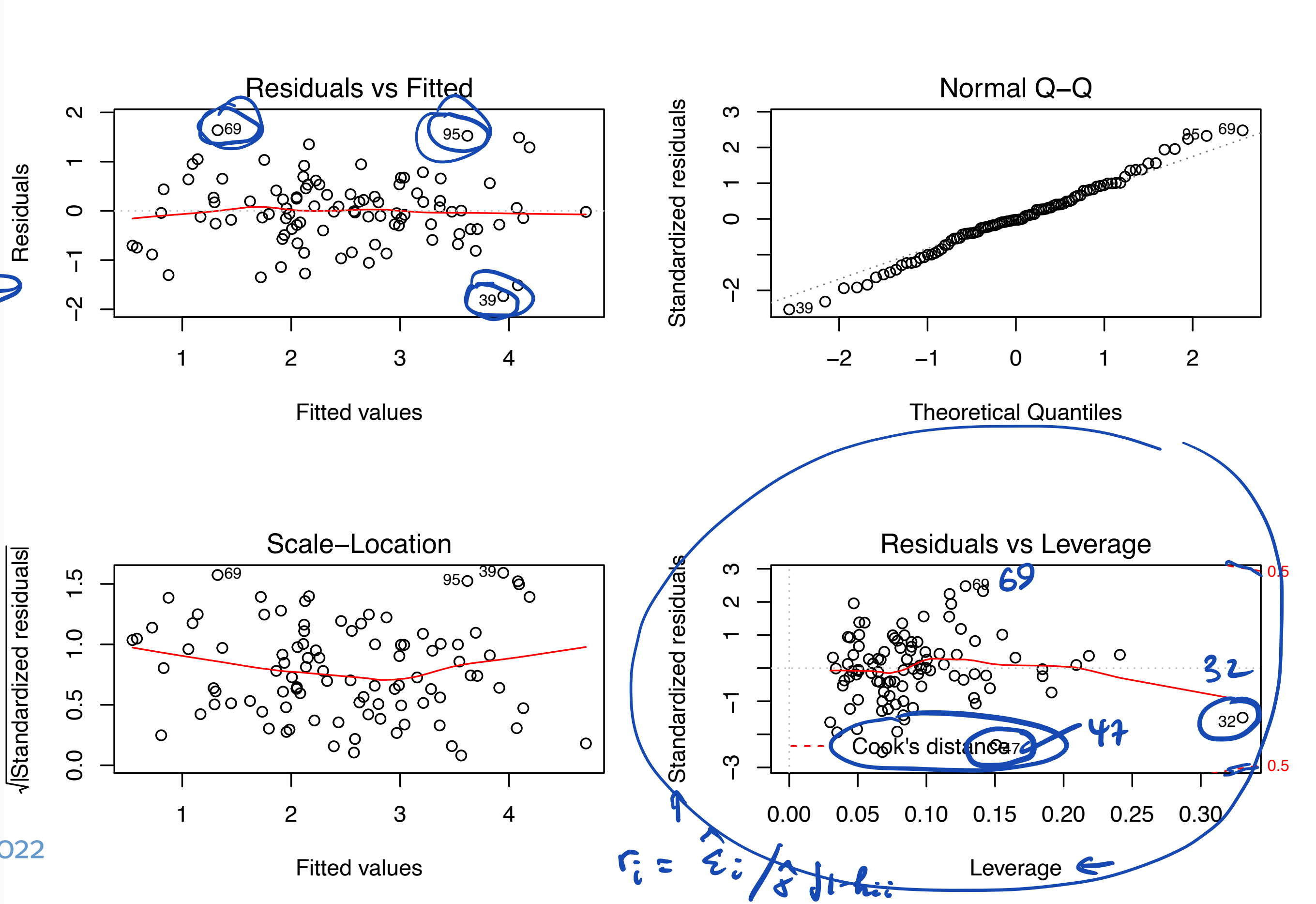
$\varepsilon_i \sim N(0, \sigma^2)$



Model assumptions

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

might see outliers here



$r_i = \hat{\varepsilon}_i / \hat{\sigma}\sqrt{1-h_{ii}}$

… Model checking

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

  matrix

- $\text{Var}(\hat{\epsilon}) =$
- i.e. don't all have the same variance

$$\hat{\varepsilon} = \begin{pmatrix} \hat{\varepsilon}_1 \\ \vdots \\ \hat{\varepsilon}_n \end{pmatrix}$$

$$\text{cov}(\hat{\varepsilon}) = \text{cov}(y - \hat{y}) = \text{cov}(y - Hy)$$

$$= \sigma^2 (I - H)$$

$$\hat{y} = \boxed{Hy}$$

$$\boxed{H = X(X^T X)^{-1} X^T}_{n \times n}$$

hat matrix

$$\text{cov}(y(I-H)) = \underline{(I-H)^T}\,\underline{\text{cov}(y)}\,\underline{(I-H)} = (I-H)^T \sigma^2 (I-H)$$

$$\text{var}\,\underline{\underline{\hat{\varepsilon}_i}} = \sigma^2 (1 - \boxed{\underline{\underline{h_{ii}}}})$$

n

$$\text{trace}\, H = p$$

↑

average value

… Model checking

$\sum_{i=1} \downarrow = \sigma^2 (n-p)$

average value of $H_{ii} \approx p/n$

- residuals: $\hat{\epsilon}_i =$ $y_i - \hat{y}_i$
- $\mathrm{Var}(\hat{\epsilon}) =$ $\sigma^2 (I - H)$
- i.e. don't all have the same variance
- hat matrix $H =$ $X(X^TX)^{-1}X^T$
- standardized residuals: $r_i =$ $\hat{\varepsilon}_i / \sqrt{1-h_{ii}} \cdot \hat{\sigma}$

$= \tilde{\sigma}$

$\boxed{\hat{\sigma}} = \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{n-p}$

approx same variance $= 1$

- Cook's distance $C_i =$

influential ← affects the fit

outlier ← affects the ~~est~~ estimate of variance

… Model checking

- 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

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

$\hat{\sigma} = \tilde{\sigma} = \sqrt{MSE}$

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

$X\hat{\beta}$ $\qquad X_{(-i)}\hat{\beta}$

$= \dots = \left(\dfrac{r_i^2}{p}\right)\left(\dfrac{h_{ii}}{1-h_{ii}}\right)$

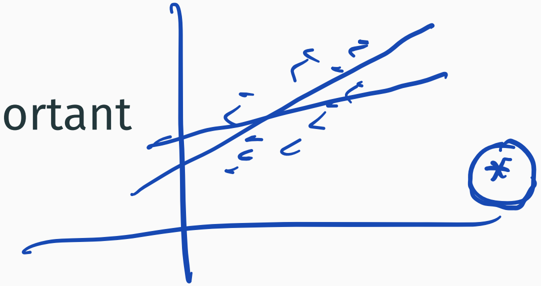
h_{ii} is a "leverage" of

... Model checking

ith obs =

- standard diagnostics check for non-constant variance, influential observations
- and for normality of residuals

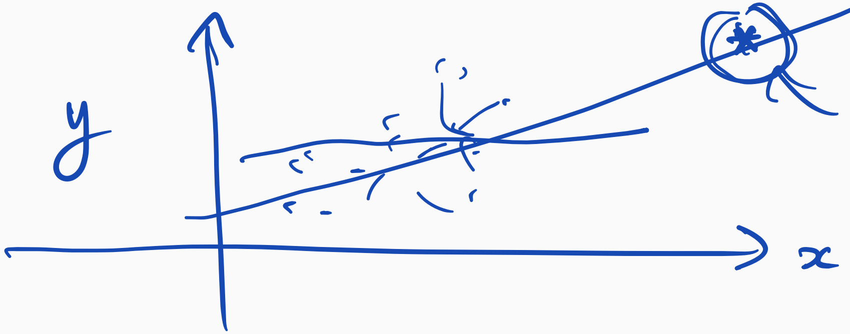
using qqnorm



- assumption of independence across *i* may be more important
- but more difficult to assess
- exception: observations collected over time LM-2, §6.1.3, LM-1 §4.1.3

$\sigma^2 / \underline{\sum (x_i - \bar{x})^2}$

var $\hat{\beta}_1$



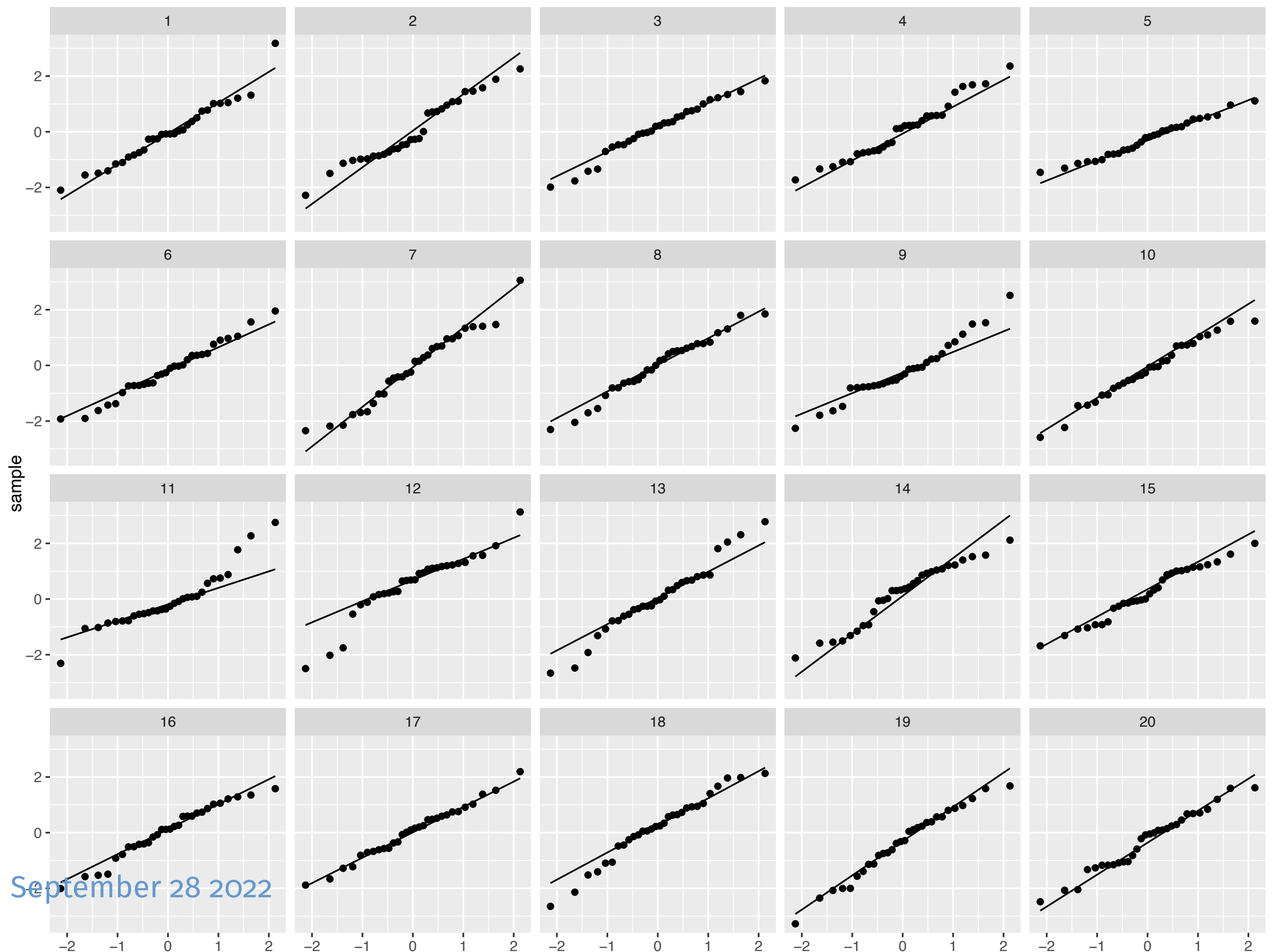
high leverage

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

>> avg

Aside on normal plots

$(x_i - x)$



… Aside

```
library(ggplot2); library(nullabor); library(tidyverse)
df5_frame <- data.frame(x = rt(30, df = 5))
lineup_df5_data <- lineup(
  method = null_dist("x", dist = "norm", params = list(mean = 0, sd = 1)),
  true = df5_frame, n=12)

lineup_df5_data %>%
  ggplot(aes(sample = x)) +
  geom_qq_line() +
  geom_qq() +
  facet_wrap(˜ .sample)
```

Model structure

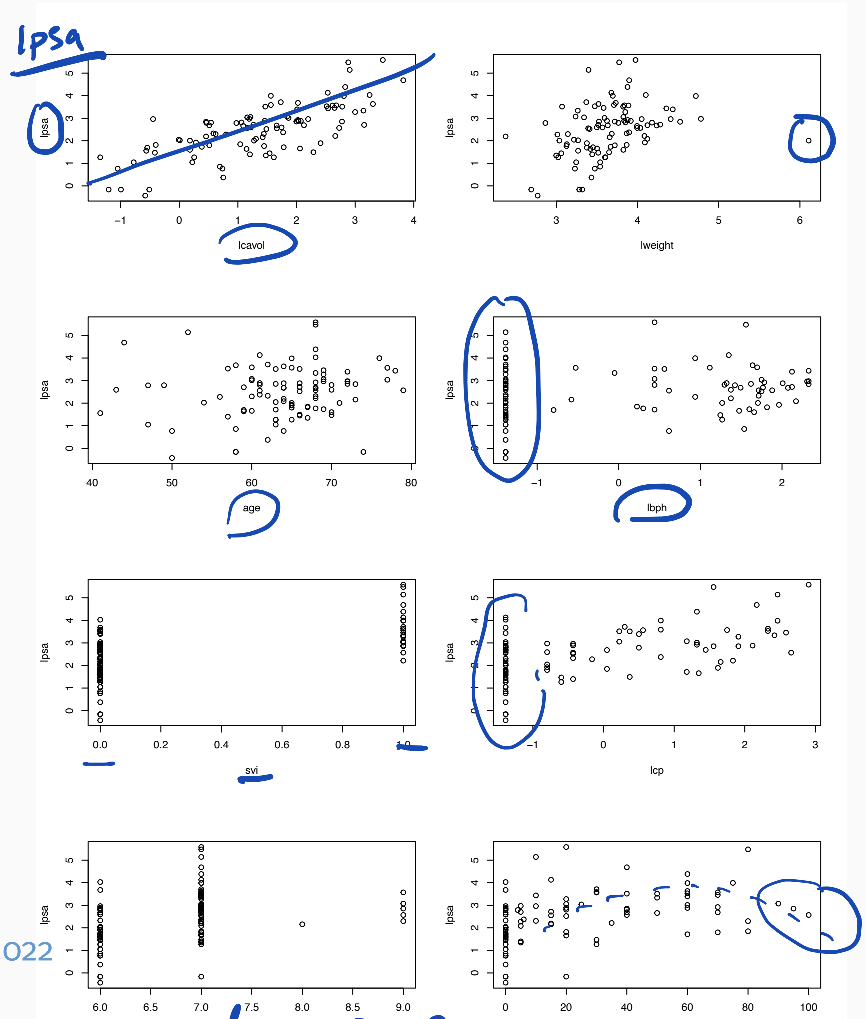
LM-2 §6.3, LM-1 §4.3

- Model $y = X\beta + \epsilon$, alternatively,
- $E(y \mid X) = X\beta, \quad \mathrm{Var}(Y \mid X) = \sigma^2 I$
- plots of y against each column of x can be helpful
- `for(i in 1:8){plot(prostate[,i],prostate[,9]... }`
- added variable plots can be more helpful

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

  slope of this line is $\hat{\beta}_j$

Prostate data



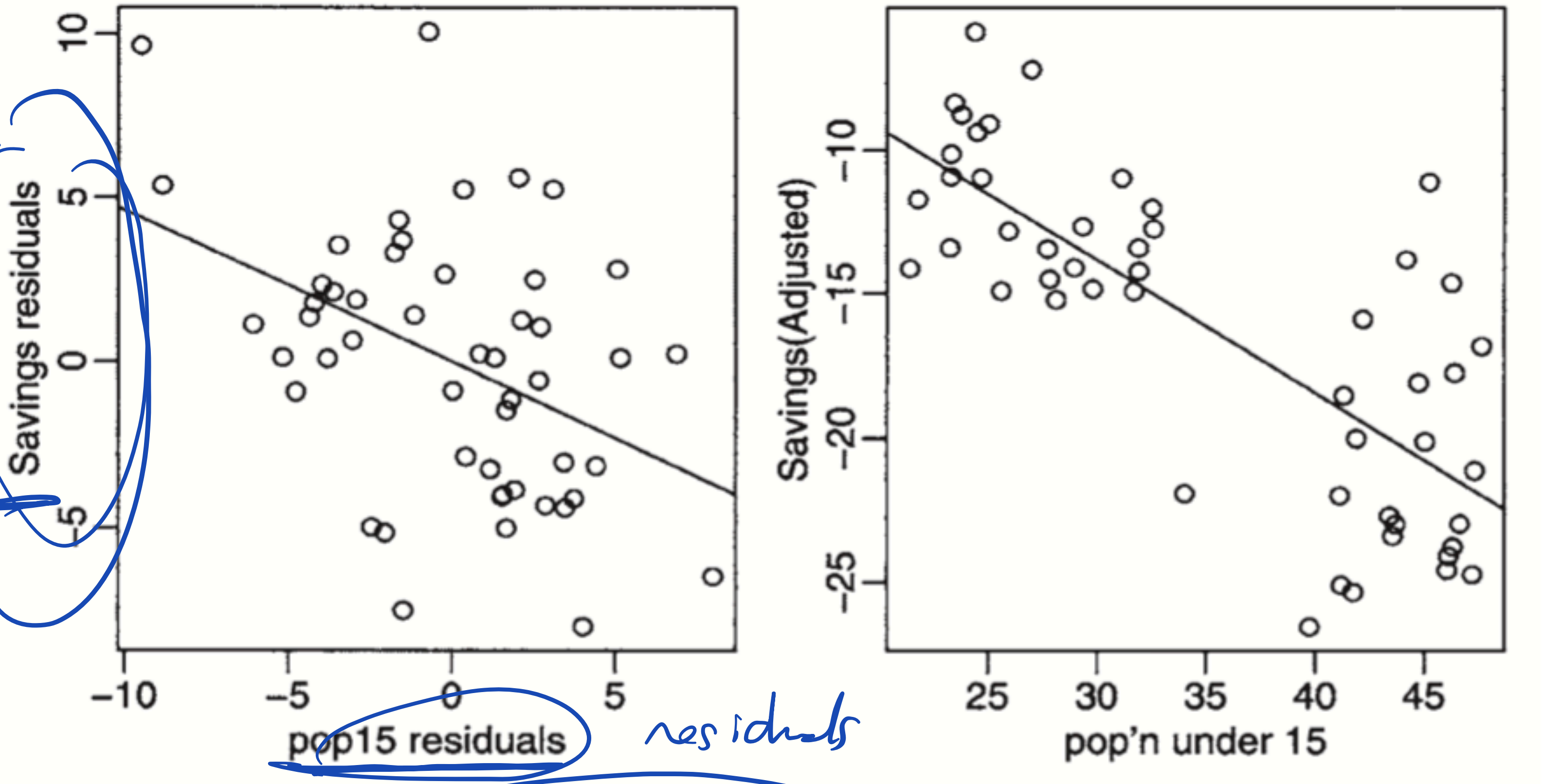
Pgg45

gleason score

Partial residual plots

LM-2 Fig 6.13, LM-1 Fig 4.13

resid

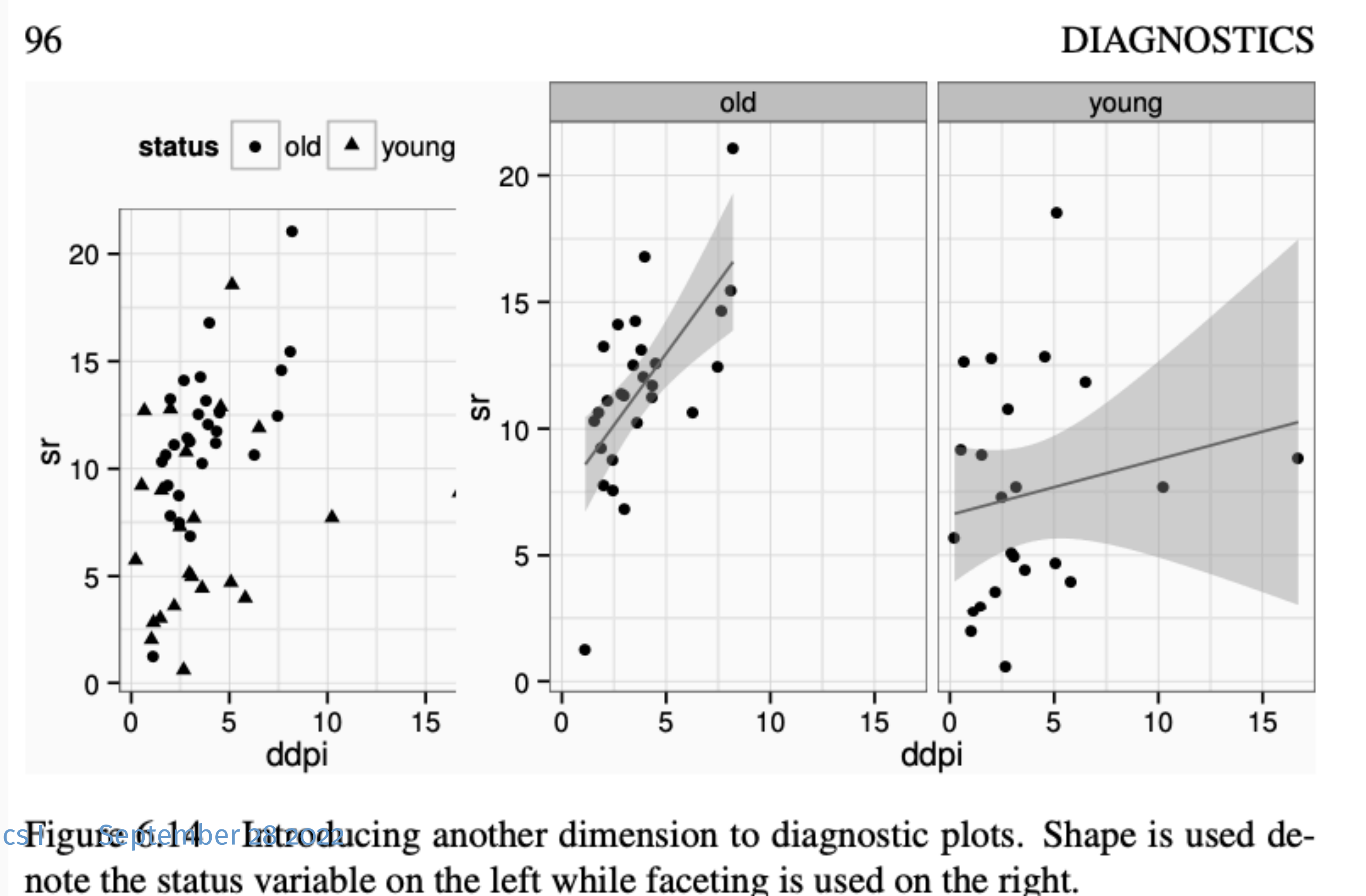


residuals

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

Partial residual plots

LM-2 Fig 6.13, LM-1 Fig 4.13



96 DIAGNOSTICS

Figure 6.14 Introducing another dimension to diagnostic plots. Shape is used denote the status variable on the left while faceting is used on the right.

partial ssion plots

pop15

reg. $\overset{\text{savings}}{y}$ on everything but 5th x

reg ~~other x's " "~~

5th x " " all other x's

pop15

Fit y on $X - x_5$ $\hat{\delta}$

Fit x_5 on $X - x_5$ $\hat{\gamma}$

regress $\hat{\delta}$ on $\hat{\gamma}$ slope $\equiv \hat{\beta}_5$ from full fit

$\hat{\beta}_5$ has an interpretation as what's "left over" bet y & x_5 after other X's fitted,

Collinearity

LM-2 §7.3, LM-1 §5.3

- 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

Collinearity

LM-2 §7.3, LM-1 §5.3

- 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 $\underline{X^TX \text{ not invertible}}$
- usually we're somewhere in between, at least in observational studies
- may be very difficult to dis-entangle effects of correlated covariates

condition number of model.matrix

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

$$(X^TX) = \begin{bmatrix} n & \cdot & 0 \\ & \ddots & \\ & & \ddots \end{bmatrix}$$

$\hat\beta_1$ is same whether or not x_2 is included

Collinearity

LM-2 §7.3, LM-1 §5.3

- 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
- 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

$$y_t \sim \beta_0 + \boxed{\beta_1} PM_{10,t} + \boxed{\beta_2 \, temp_t} + \varepsilon_t$$

$\hat\beta_1 \quad \hat\beta_2$ corr'd

$\beta_2 S\,(temp) + \beta_3 S_2\,(temp)$

Collinearity

LM-2 §7.3, LM-1 §5.3

- 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
- 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

Three tasks related to linear regression

- Estimation of β, and estimation of its standard error – for inference about $\mathbb{E}(y \mid x)$

  alternatively comparing sub-models using F-tests

- Prediction of y_+, say, given a new vector of explanatory variables x_+

  LM-2 Ch.4, LM-1 §3.5, SM §8.3.2

- Model Selection: which explanatory variables do we need for prediction or inference?

Three tasks related to linear regression

- Estimation of β, and estimation of its standard error – for inference about $\mathbb{E}(y \mid x)$

  alternatively comparing sub-models using *F*-tests

- Prediction of y_+, say, given a new vector of explanatory variables x_+

  LM-2 Ch.4, LM-1 §3.5, SM §8.3.2

- Model Selection: which explanatory variables do we need for prediction or inference?

These same questions arise in other models such as logistic regression, analysis of survival data, and so on, but the generic linear model is often a good starting point

Three tasks related to linear regression

- Estimation of β, and estimation of its standard error – for inference about $\mathbb{E}(y \mid x)$

alternatively comparing sub-models using F-tests

- Prediction of y_+, say, given a new vector of explanatory variables x_+

LM-2 Ch.4, LM-1 §3.5, SM §8.3.2

- Model Selection: which explanatory variables do we need for prediction or inference?

These same questions arise in other models such as logistic regression, analysis of survival data, and so on, but the generic linear model is often a good starting point

- Prediction: $y_+ = x_+^{\mathrm{T}}\beta + \epsilon$; $\quad \hat{y}_+ = x_+^{\mathrm{T}}\hat{\beta}$; $\quad \mathrm{var}(\hat{y}_+) = \sigma^2 x_+ (X^{\mathrm{T}}X)^{-1} x_+$

assuming ...

Three tasks related to linear regression

- Estimation of β, and estimation of its standard error – for inference about $\mathbb{E}(y \mid x)$

alternatively comparing sub-models using F-tests

- Prediction of y_+, say, given a new vector of explanatory variables x_+

LM-2 Ch.4, LM-1 §3.5, SM §8.3.2

- Model Selection: which explanatory variables do we need for prediction or inference?

These same questions arise in other models such as logistic regression, analysis of survival data, and so on, but the generic linear model is often a good starting point

- Prediction: $y_+ = x_+^{\mathrm{T}}\beta + \epsilon$; $\quad \hat{y}_+ = x_+^{\mathrm{T}}\hat{\beta}$; $\quad \text{var}(\hat{y}_+) = \sigma^2 x_+ (X^{\mathrm{T}}X)^{-1} x_+$

assuming ...

- error in expected response different from

prediction error $\mathbb{E}(y_+ - \hat{y}_+)^2 = \sigma^2 + \text{var}(\hat{y}_+)$

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, ...
- 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

LM-2 Ch.10; LM-1 Ch.8; SM, Ch.8.7

- “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 $x = 0/1$
- $y = \beta_0 + \beta_1 \sin(2\pi x) + \beta_2 \cos(2\pi x) + \beta_3 \sin(4\pi x) + \beta_4 \cos(4\pi x) + \epsilon$
- $y_t = \beta_0 + \alpha y_{t-1} + \epsilon$ $y_t = \beta_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} \epsilon$ *not* $y_t = \beta_0 + \alpha_2 y_{t-2} + \epsilon$

… Model Selection

LM-2 Ch.10; LM-1 Ch.8; SM, Ch.8.7

- testing procedures: forward selection, backward selection, stepwise selection
- it is quite common to fit all explanatory variables, and then drop if $p > 0.05$

... Model Selection

LM-2 Ch.10; LM-1 Ch.8; SM, Ch.8.7

```
step(model1)
...
Step:  AIC=-61.37
lpsa ˜ lcavol + lweight + age + lbph + svi

          Df Sum of Sq    RSS     AIC
<none>                 45.526 -61.374
- age      1    0.9592 46.485 -61.352
- lbph     1    1.8568 47.382 -59.497
- lweight  1    3.2251 48.751 -56.735
- svi      1    5.9517 51.477 -51.456
- lcavol   1   28.7665 74.292 -15.871

Call:
lm(formula = lpsa ˜ lcavol + lweight + age + lbph + svi, data = prostate)

Coefficients:
(Intercept)       lcavol      lweight          age         lbph          svi
    0.95100      0.56561      0.42369     -0.01489      0.11184      0.72095
```

… Model Selection

LM-2 Ch.10; LM-1 Ch.8; SM, Ch.8.7

- Criterion-based procedures
- AIC, BIC, Mallows C_p, R_a^2

most widely used

RSS: residual sum of squares

… Model Selection

LM-2 Ch.10; LM-1 Ch.8; SM, Ch.8.7

- Criterion-based procedures

most widely used

- AIC, BIC, Mallows C_p, R_a^2

RSS: residual sum of squares

-

$$AIC = n\log(RSS/n) + 2p$$

… Model Selection

LM-2 Ch.10; LM-1 Ch.8; SM, Ch.8.7

- Criterion-based procedures

most widely used

- AIC, BIC, Mallows C_p, R_a^2

RSS: residual sum of squares

- $$AIC = n\log(RSS/n) + 2p$$

- $$BIC = n\log(RSS/n) + \log(n)p$$

… Model Selection

LM-2 Ch.10; LM-1 Ch.8; SM, Ch.8.7

- Criterion-based procedures

most widely used

- AIC, BIC, Mallows C_p, R_a^2

RSS: residual sum of squares

- $$AIC = n\log(RSS/n) + 2p$$

- $$BIC = n\log(RSS/n) + \log(n)p$$

- $$C_p = RSS_p/\tilde{\sigma}^2 + 2p - n$$

… Model Selection

LM-2 Ch.10; LM-1 Ch.8; SM, Ch.8.7

- Criterion-based procedures

most widely used

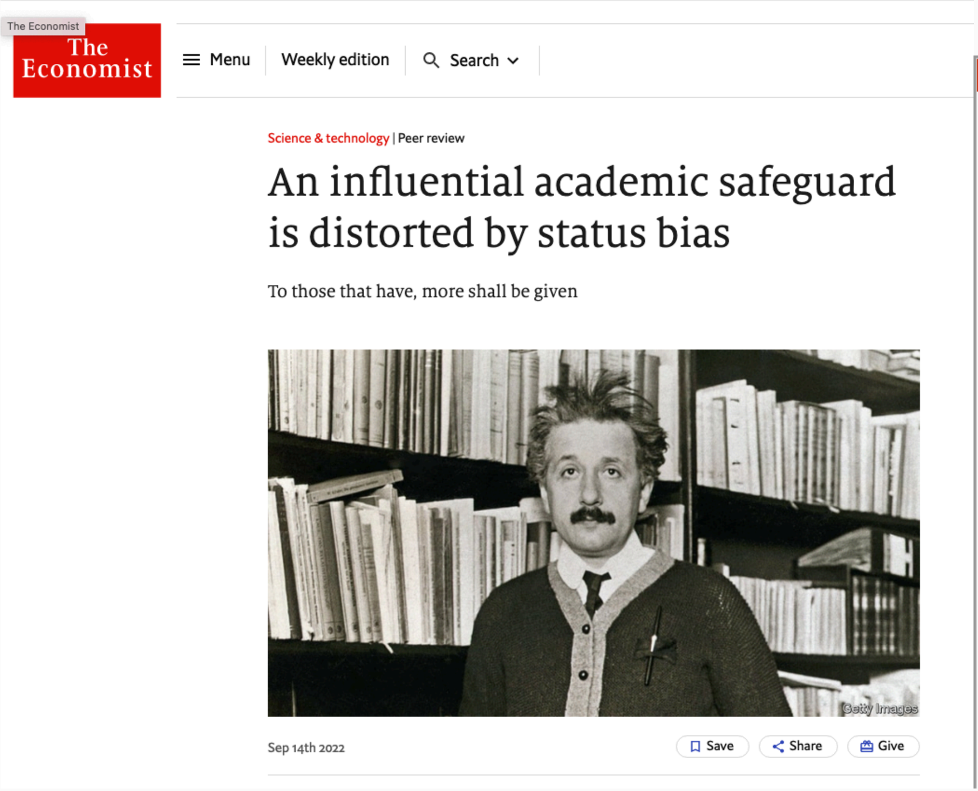
- AIC, BIC, Mallows C_p, R_a^2

RSS: residual sum of squares

- $$AIC = n\log(RSS/n) + 2p$$
- $$BIC = n\log(RSS/n) + \log(n)p$$
- $$C_p = RSS_p/\tilde{\sigma}^2 + 2p - n$$
- $$R_a^2 = 1 - \frac{\tilde{\sigma}^2_{model}}{TSS/(n-1)}$$
- SM has yet another version AIC_c which may be better than AIC for linear models

In the News

Economist, Sep 14

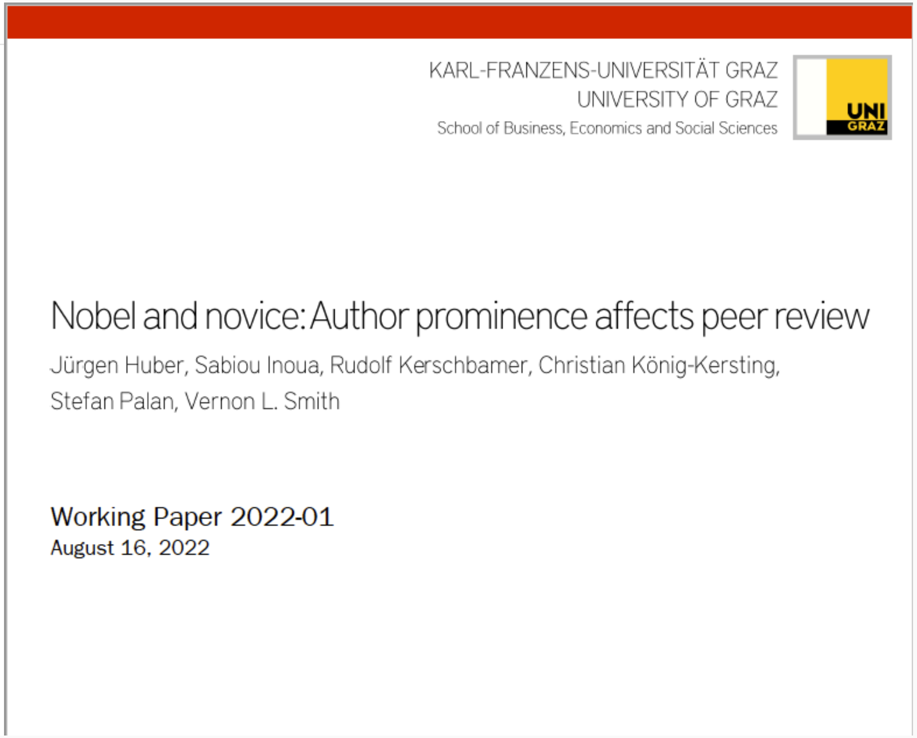


Science & technology | Peer review

An influential academic safeguard is distorted by status bias

To those that have, more shall be given

Sep 14th 2022



KARL-FRANZENS-UNIVERSITÄT GRAZ
UNIVERSITY OF GRAZ
School of Business, Economics and Social Sciences

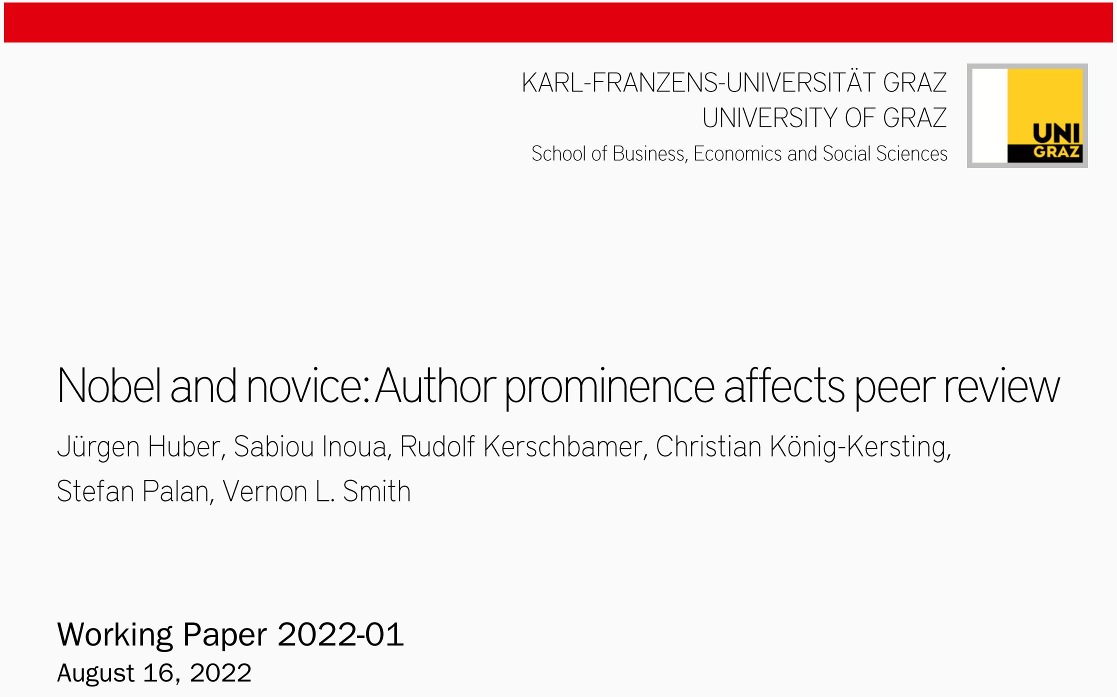
Nobel and novice: Author prominence affects peer review

Jürgen Huber, Sabiou Inoua, Rudolf Kerschbamer, Christian König-Kersting, Stefan Palan, Vernon L. Smith

Working Paper 2022-01
August 16, 2022

link

A randomized experiment



KARL-FRANZENS-UNIVERSITÄT GRAZ
UNIVERSITY OF GRAZ
School of Business, Economics and Social Sciences

Nobel and novice: Author prominence affects peer review

Jürgen Huber, Sabiou Inoua, Rudolf Kerschbamer, Christian König-Kersting,
Stefan Palan, Vernon L. Smith

Working Paper 2022-01
August 16, 2022

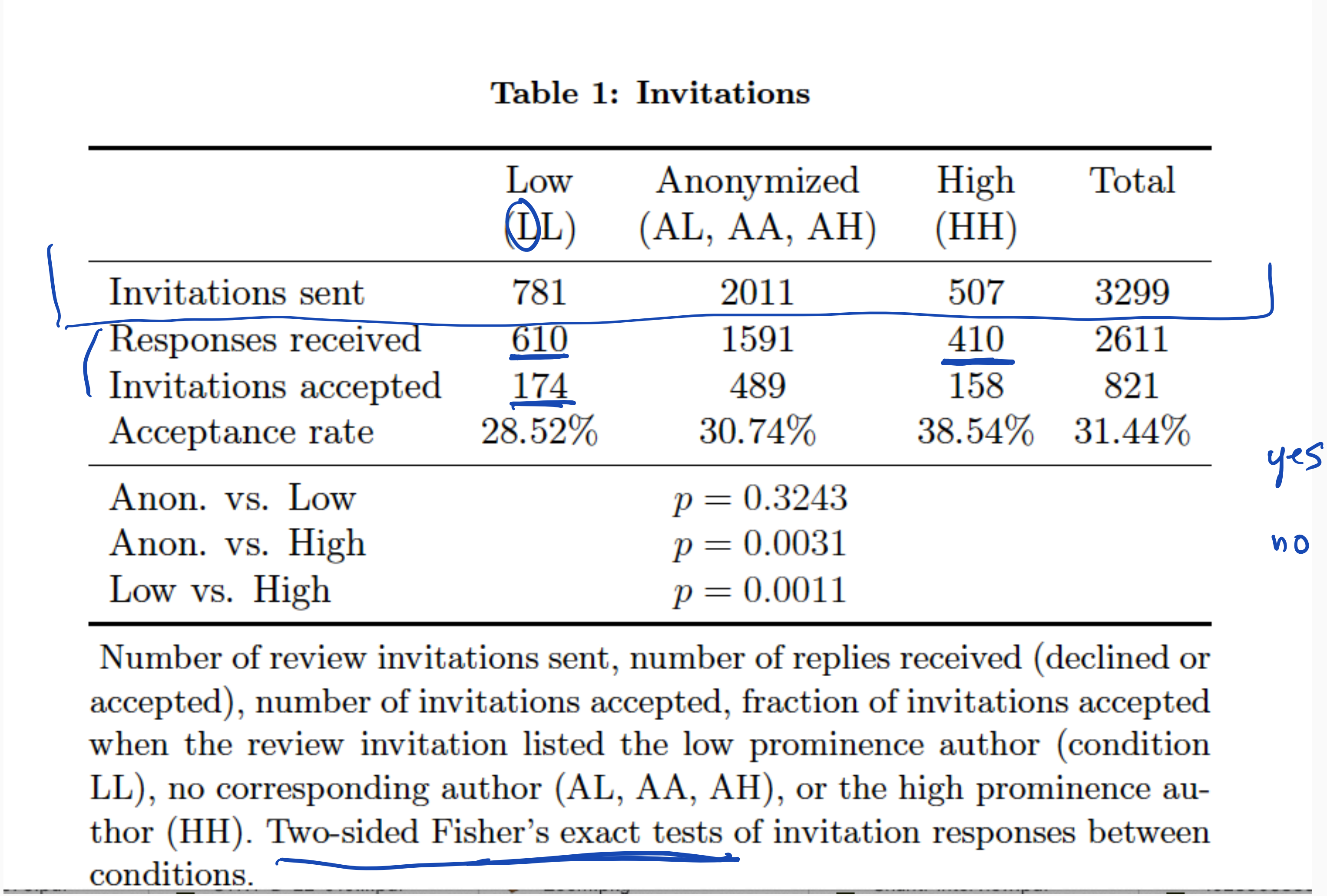


Table 1: Invitations

	Low (LL)	Anonymized (AL, AA, AH)	High (HH)	Total
Invitations sent	781	2011	507	3299
Responses received	610	1591	410	2611
Invitations accepted	174	489	158	821
Acceptance rate	28.52%	30.74%	38.54%	31.44%
Anon. vs. Low		$p = 0.3243$		
Anon. vs. High		$p = 0.0031$		
Low vs. High		$p = 0.0011$		

Number of review invitations sent, number of replies received (declined or accepted), number of invitations accepted, fraction of invitations accepted when the review invitation listed the low prominence author (condition LL), no corresponding author (AL, AA, AH), or the high prominence author (HH). Two-sided Fisher's exact tests of invitation responses between conditions.

	L	A	H
yes	174		
no			

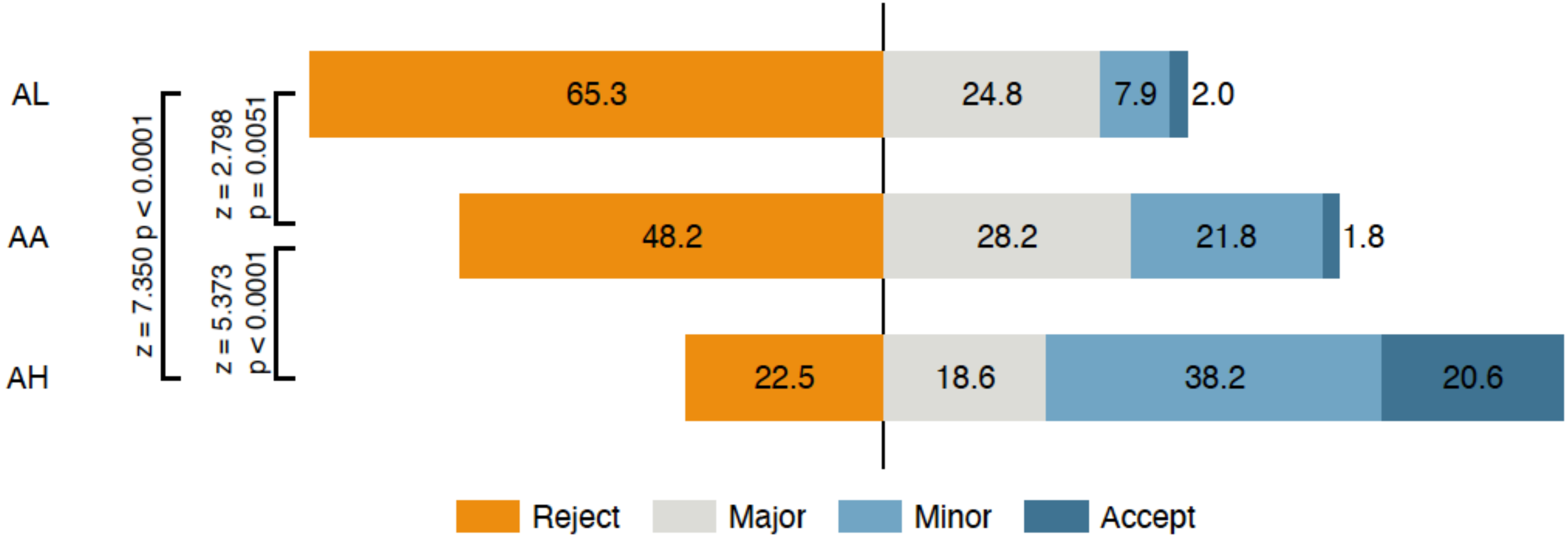


Figure 1: Recommendation percentages by condition. L stands for the relatively unknown author, A stands for anonymized and H stands for the highly prominent author. In conditions AL and AH, the invitation email is anonymized, but the respective corresponding author's name appears on the manuscript, while in AA both the invitation and the paper are anonymized. The tests are pairwise, two-sided Mann-Whitney U tests.

Design of Studies

- 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

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

Design of Studies

CD, Ch.2

- 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

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
- to ensure that the scale of effort is appropriate

… design of studies

- we concentrate largely on the careful analysis of individual studies

… design of studies

- we concentrate largely on the careful analysis of individual studies
- in most situations synthesis of information from different investigations is needed

… 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

… 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

… design of studies

- formulation of a plan of analysis
- establish and document that proposed data are capable of addressing the research questions of concern

… 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

… 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

… 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

… 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

… 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

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)

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/...
- split plot experiments have two classes of units of study and analysis

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/...
- split plot experiments have two classes of units of study and analysis
- 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

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/...
- split plot experiments have two classes of units of study and analysis
- 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/...
- split plot experiments have two classes of units of study and analysis
- 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

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)

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

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

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

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
- experiment: investigator has complete control over treatment assignment

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
- experiment: investigator has complete control over treatment assignment
- census

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
- experiment: investigator has complete control over treatment assignment
- census
- meta-analysis: statistical assessment of a collection of studies on the same topic

Avoidance of systematic error

CD §2.4

- "distortion in the conclusions arising from irrelevant sources that do not cancel out in the long run"

Avoidance of systematic error

CD §2.4

- “distortion in the conclusions arising from irrelevant sources that do not cancel out in the long run”
- can arise through systematic aspects of, for example, a measuring process, or the spatial or temporal arrangement of units

Avoidance of systematic error

CD §2.4

- “distortion in the conclusions arising from irrelevant sources that do not cancel out in the long run”
- can arise through systematic aspects of, for example, a measuring process, or the spatial or temporal arrangement of units
- this can often be avoided by design, or adjustment in analysis

Avoidance of systematic error

CD §2.4

- “distortion in the conclusions arising from irrelevant sources that do not cancel out in the long run”
- can arise through systematic aspects of, for example, a measuring process, or the spatial or temporal arrangement of units
- this can often be avoided by design, or adjustment in analysis
- can arise by the entry of personal judgement into some aspect of the data collection process

Avoidance of systematic error

CD §2.4

- “distortion in the conclusions arising from irrelevant sources that do not cancel out in the long run”
- can arise through systematic aspects of, for example, a measuring process, or the spatial or temporal arrangement of units
- this can often be avoided by design, or adjustment in analysis
- can arise by the entry of personal judgement into some aspect of the data collection process
- this can often be avoided by randomization and blinding