Question 1

Critical Reading - Young (2019)

Summary. Conventional t-tests have distorted size when some observations possess high leverage. On the contrary, randomization inference shows correct nominal size - without reducing power - even in such contexts. This observation is particularly true for small sample sizes ($N < 2.000$), when inference is conducted on sub-samples or when the treatment measure is interacted with pre-treatment covariates because size distorsions are magnified in multiple and joint hypothesis testing problems. The results are presented using a sample of real-world hypothesis tests obtained from 53 peer-reviewed articles and using simulated data. Accordingly, researchers should always be careful when analyzing datasets that depart from the uniform leverage standard and possibly switch to

Main strength. The main strength of the paper is that it brings back to discussion the problem of sensitivity of point estimation and inference to units with high leverage. This issue is clearly underestimated in current applied research and the paper points out that it changes the main results of a substantive amount of papers present in the literature.

Weakness. Divergence between conventional t-tests and randomization inference seems to be pretty severe for samples with size smaller than 2.000. For larger samples, both approaches present the extremely similar nominal sizes and power curves. First, the main results seem to be valid only in the small sample case. It would be informative to check for which of the 53 experimental papers randomization inference and conventional t-tests disagree and subset this information based on the sample size of the regression. Furthermore, I am not convinced that the two procedures are really comparable, in particular their power curves. Indeed, a Fisherian sharp null always implies a Neymanian null hypothesis, but the converse is not true. Intuitively, one might expect that rejecting the former is easier, hence it could be that it is easier to achieve the same power with a sharp null hypothesis. However, Ding (2017) shows that the p-values of the two testing procedures depend on different variance estimators and their difference is only asymptotically negligible and, as a general statement, cannot be ranked in finite samples. As such a rejection of a Neymanian hypothesis does not imply the rejection of a Fisherian hypothesis. The author shows that this logical paradox often takes place in common settings (e.g. constante treatment effects, stratified experiments, ...). Finally, the author uses only HC1 estimators and does not show how HC2, HC3 (even though jackknife is used, which is almost the same), and HC4 estimators behave in those cases. Since they have been introduced in the literature as a way to make sandwich estimators more robust to the presence of outliers, including them in the comparison could have been of interest. Probably, these estimators are also badly behaved in small sample sizes as those for which randomization inference clearly dominates conventional t-tests.

Contribution. The paper contributes to the literatures of conventional t-tests and randomization t-tests in the spirit of Fisher. In particular, the author proposes randomization t-tests as a solution to well known problem of size distortion in conventional t-tests caused by high leverage observations. Furthermore, randomization inference is shown to provide a means to build statistical tests with credible finite-sample rejection probabilities in the case of high-dimensional joint and multiple testing.

Critical Reading - Young (2022)

Summary. The author relies on a sample of 1359 Instrumental Variable regressions (of which 1087 are just identified) to show how high leverage observations affect the size and the power of conventional t-tests for the significance of the IV estimator. First, results obtained using MonteCarlo simulations of DGP that resemble real-world data show that high leverage observations sweep away almost all of the (truncated) bias advantage that 2SLS has with respect to OLS, thus making the latter preferable in an MSE sense. Second, the combined presence of outliers and non iid errors distorts the size of classical tests for weak IV Stock & Yogo (2005), hence leading to deem as relevant instruments that are should not be considered as such. Third, the jackknife and various types of bootstrap are shown to restore nominal size and have larger power than conventional t-tests for both the first and second stages hypothesis tests.

Main strength. The paper brings on the main stage the fact that high leverage observations have the potential to invalidate typical "screening" tests for the strength of the instrument and/or for the IV coefficient. The presence of such observations might induce the researcher to deem as strong an instrument that produces an unidentifed first-stage or interpret as significant a coefficient that is not. Not surprisingly since Young (2019), size distortions are present also in IV inference when outliers are present in the data. Therefore, the paper gives great practical advice to pay the deserved attention to outliers and also proposes a feasible and easy to implement solution, which is use the bootstrap or the jackknife to estimate standard errors used for the first stage F-statistic or the second stage t-test on the IV coefficient.

Weakness. The main weakness of the paper, which is stressed multiple times by the author, is that weak IV scenarios are de-facto ruled away. This is of course a sufficient condition to talk about quantities of interest such as expected bias or mean squared error that would have not been defined otherwise. Also, the DGPs used in the MonteCarlo are obtained by fitting 2SLS regressions in the data and then using the estimated parameters as the true ones to obtain a feasible DGP to draw data from. Since these estimates are obtained using results of published papers and the distribution of first stage coefficients used across the 1327 DGPs is not reported, it is impossible to gauge the extent to which the results might be extended to weak IV settings. Probably just a few of these DGPs can be interpreted as having a weak first stage population regression since the bootstrap and the jackknife show correct nominal size. However, having said that, this paper probably set the ground for future research on the topic, but in general, it is not hard to imagine that if outliers are problematic for strong IVs they must be such also for weak IVs.

Contribution. Extending the previous work of Young (2019) to the IV case is the main contribution of the paper. In hindsight the result is not surprising, but this is just because of the first paper and the fact that conventional IV inference relies typically on sandwich estimators for the standard errors. The paper has also a constructive part, in which advice is given to practitioners on how to tackle the issue of high leverage observations in the data in an IV setting.

Question 2

a) Suppose that we are interested in the population regression $\mathbb{E}\left[Y_i \mid X_i = x\right]$ of log-earnings Y_i on education W_i and a constant, that is

$$\theta_{ds} = \arg\min_{\theta} \mathbb{E}\left[(Y_i - X_i\theta)^2\right],$$

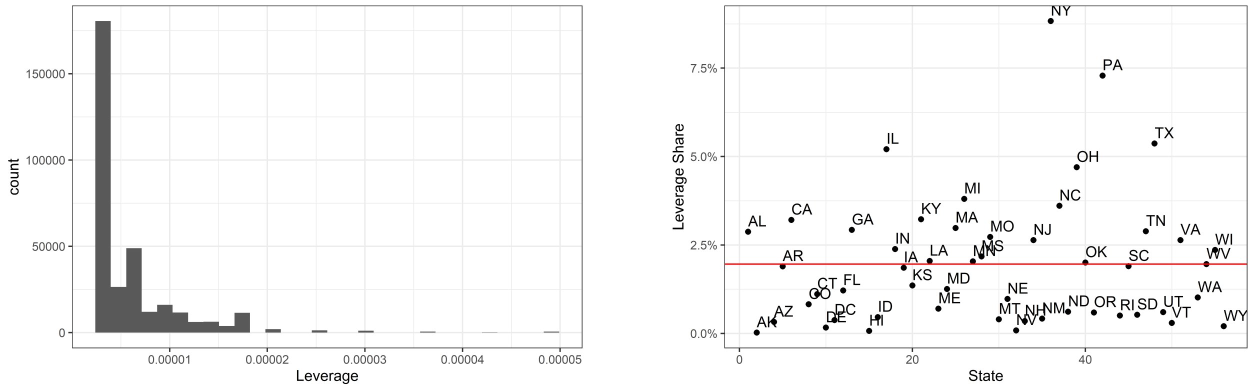
where $X_i = (1, W_i)'$. The data we are using to construct an estimator for θ_{ds} is an extract of the census, implying that we observe just a fraction of the entire population and that sampling uncertainty must be taken into account. The sampling scheme in our case consists in randomly drawing N individuals from a superpopulation of units. If we worry about the correlation between the idiosyncratic individual shocks of units living in the same states, the fact that we observe units from **all** clusters (states) implies that we do not need to cluster our standard errors (Abadie, Athey, Imbens & Wooldridge 2017). Intuitively, we are interested in a descriptive estimand, thus estimation uncertainty arises only from the sampling procedure we use. However, we observe all the clusters in our sample, that is we draw units from all the states, thus there is no sampling error (at this level) to be taken into account. Finally, since we just observe a fraction of the population, we can use the EHW estimator to take into account potential heteroskedasticity.

Let's first estimate θ_{ds} and then compare the standard errors clustered at the state of birth level with the EHW with the correction for degrees of freedom (HC1). Table 1 reports the results.

Table 1: *Comparison of different standard error estimators.*

	Dependent variable:	
	log(wage)	
	HC0	Liang & Zeger (1986)
Education	0.071 (0.0004)	0.071 (0.002)
Observations	329,509	329,509
Clusters	-	51

The two estimates for the standard errors are an order of magnitude apart from each other, with the clustered ones being larger. Both are extremely small and results are unaffected by the choice of the standard error estimator.



(i) *Unit leverage H.*

(ii) *Cluster-leverage H_g.*

Notes: We follow MacKinnon et al. (2022) and compute the cluster leverage as the nuclear norm of H_s and then divide it by k to get the share of total leverage. The horizontal solid red line indicates the ideal benchmark of balanced leverage. In this setting, the benchmark is k/S, rather than k/N. The ideal share is therefore $1/S$.

Finally, the maximum leverage for individual observations is smaller than 1e-5 (see Figure 1i). This allows us to exclude the presence of outliers that would violate the assumptions for the validity of the EHW estimator. If we were to used the LZ estimator, we should have paid attention to the presence of high-leverage clusters (see Figure 1ii). The figure reports the leverage of each cluster, that is

$$L_s = \text{Tr}\left(H_s\right) = \text{Tr}\left(X_s'X_s\left(X'X\right)^{-1}\right), \qquad H_s = X_s\left(X'X\right)^{-1}X_s', \quad s = 1, \dots, S,$$

as a fraction of the total leverage. We can see that many clusters have a leverage high enough to distort EHW standard errors (Young 2019). If we were to cluster standard errors, an outlier-robust estimator such as Bell & McCaffrey (2002) would have been more appropriate.

b) Now we are interested in the impact of growing in NJ on earnings

$$\theta_{ds} = \arg\min_{\theta} \mathbb{E}\left[\left(Y_i - X_i\theta\right)^2\right], \quad X_i = \left(1, \mathbb{1}\left(S_i = \text{NJ}\right)\right),$$

where S_i is the state of residence.
Again, we observe all the clusters, hence there is no need to cluster the standard errors at the state level. Differently from before, if we try to cluster the standard errors at the state level, we will encounter another problem, due to the fact that clustering happens at the same level as one of the covariates is varying, namely $Z_i = \mathbb{1}\left(S_i = \text{NJ}\right)$. This implies that the least squares residuals $\widehat{\epsilon}_{i,s}$ are by construction orthogonal to $Z_{i,s}$ for each cluster. To see this, first note that $Z_{i,s}\widehat{\epsilon}_{i,s} = 0$ whenever $s \neq \text{NJ}$ because $Z_{i,s} = 0$. Second, least squares algebra gives us that $\boldsymbol{Z}\widehat{\boldsymbol{\epsilon}} = 0$. Putting the two things together we get

$$0 = \boldsymbol{Z}\widehat{\boldsymbol{\epsilon}} = \sum_{s=1}^{S}\sum_{i:s(i)=s} Z_{i,s}\widehat{\epsilon}_{i,s} = \sum_{i:s(i)=\text{NJ}} Z_{i,\text{NJ}}\widehat{\epsilon}_{i,\text{NJ}}.$$

In turn, this implies that the "meat" matrix of the cluster-robust standard error estimator will be singular

$$\sum_s \sum_{i,j:s(i)=s(j)=s} \hat{\epsilon}_i \hat{\epsilon}_j X_i X_j' = \begin{bmatrix} \sum_s \sum_{i,j:s(i)=s(j)=s} \hat{\epsilon}_i \hat{\epsilon}_j & \sum_s \sum_{i,j:s(i)=s(j)=s} \hat{\epsilon}_i \hat{\epsilon}_j Z_{i,s} \\ \sum_s \sum_{i,j:s(i)=s(j)=s} \hat{\epsilon}_i \hat{\epsilon}_j Z_{i,s} & \sum_s \sum_{i,j:s(i)=s(j)=s} \hat{\epsilon}_i \hat{\epsilon}_j Z_{i,s} \end{bmatrix}$$

$$\approx \begin{bmatrix} \sum_s \sum_{i,j:s(i)=s(j)=s} \hat{\epsilon}_i \hat{\epsilon}_j & 0 \\ 0 & 0 \end{bmatrix}.$$

Indeed, our estimate for the meat matrix is

$$\begin{bmatrix} 147.8 & 1e-26 \\ 1e-26 & 1e-26 \end{bmatrix}$$

To compute appropriate clustered standard errors that are robust to outliers we need to use the Bell & McCaffrey (2002) estimator

$$\hat{V}_{BM} = n \left(X'X\right)^{-1} \sum_{s=1}^{S} X_s' A_s \hat{\epsilon}_s \hat{\epsilon}_s A_s' X_s' \left(X'X\right)^{-1}, \quad A_s = (I - H_s)^{-1/2}.$$

Furthermore, we take the square root of the pseudo-inverse of $I - H_s$ since such matrix is singular as we showed above. Table 2 shows the estimated standard errors obtained using the EHW sandwich estimator (our preferred choice), the Liang & Zeger (1986) estimator, and the Bell & McCaffrey (2002) estimator.

Table 2: *Comparison of different standard error estimators.*

	Dependent variable:		
	lwage		
	Heteroskedasticity	Liang & Zeger (1986)	Bell & McCaffrey (2002)
NJindic	0.107 (0.007)	0.107 (0.0217)	0.107 (0.0221)
Observations	329,509	329,509	329,509
Clusters	-	51	51

Switching from HC0 to HC2 with clustered standard error does not change much the results. Clustered standard errors are in general one order of magnitude higher than the EHW.

References

Abadie, A., Athey, S., Imbens, G. W. & Wooldridge, J. (2017), When should you adjust standard errors for clustering?, Technical report, National Bureau of Economic Research.

Bell, R. M. & McCaffrey, D. F. (2002), 'Bias reduction in standard errors for linear regression with multi-stage samples', *Survey Methodology* **28**(2), 169–182.

Ding, P. (2017), 'A paradox from randomization-based causal inference', *Statistical science* pp. 331–345.

Liang, K.-Y. & Zeger, S. L. (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**(1), 13–22.

MacKinnon, J. G., Nielsen, M. Ã., Webb, M. D. et al. (2022), Leverage, influence, and the jackknife in clustered regression models: Reliable inference using summclust, Technical report.

Stock, J. & Yogo, M. (2005), *Asymptotic distributions of instrumental variables statistics with many instruments*, Vol. 6, Chapter.

Young, A. (2019), 'Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results', *The Quarterly Journal of Economics* **134**(2), 557–598.

Young, A. (2022), 'Leverage, heteroskedasticity and instrumental variables in practical application', *Working Paper, LSE February* .

Main Code

```
library(haven)
library(dplyr)
library(ggplot2)
library(latex2exp)
library(labelled)
library(sandwich)
library(MASS)
library(dfadjust)
library(stargazer)

setwd("D:/Dropbox/Universit/PhD/II Year/Spring/ECO519 - Non-linear Econometrics/psets/ps2")
data <- haven::read_dta("ak91.dta")
data.use <- subset(data, census == 1980 & cohort == 2)

##########################################################################
## Exercise 1
# Estimate regression of earnings on a constant and education level
earn <- lm(lwage ~ 1 + educ, data = data.use)

# Clustered standard errors (Liang and Scott, 1986)
clSE <- sandwich::vcovCL(earn, cluster = data.use$SOB, type = 'HC0')

# EHW standard errors
EHW <- sandwich::vcovHC(earn, type = 'HC0')

stargazer(earn, earn, earn, se = list(NULL, sqrt(diag(EHW)), sqrt(diag(clSE))))

# Compute leverage
hats <- as.data.frame(hatvalues(earn))
theme_set(theme_bw())
ggplot(hats) + geom_histogram(aes(x=hatvalues(earn))) + xlab("Leverage")
ggsave('lvg_all.png', height = 4, width = 6, dpi = 1000)

# Compute within group leverage
sobs <- unique(data.use$SOB)
G <- length(sobs)
labs <- val_labels(sobs)
storelev <- matrix(NA, G, 3)
X <- cbind(1,data.use$educ)
XX <- solve(t(X)%*%X)

i <- 1
for (sob in sobs) {
  datacl <- subset(data.use, data.use$SOB == sob)
  Xg <- cbind(1,datacl$educ)
  XXg <- t(Xg)%*%Xg

  storelev[i, 2] <- sum(diag(XXg %*% XX)) # compute nuclear norm
  storelev[i, 1] <- sob
```

```r
  storelev[i, 3] <- names(labs)[labs == sob]
  i <- i + 1
}

toplot <- as.data.frame(storelev)
toplot$V1 <- as.numeric(toplot$V1)
toplot$V2 <- as.numeric(toplot$V2)
toplot$V4 <- toplot$V2/sum(toplot$V2)

ggplot(toplot, aes(x=V1, y=V4, label=V3)) + geom_point() +
  xlab("State") + ylab("Leverage Share") +
  geom_hline(yintercept = 1/G, color = "red") +
  geom_text(hjust=0, vjust=-1/2) +
  scale_y_continuous(labels = scales::percent)
ggsave('lvg_cls.png', height = 4, width = 6, dpi = 1000)


############################################################################
## Exercise 2
data.use$NJindic <- data.use$SOB == 34
NJimpact <- lm(lwage ~ NJindic, data = data.use)

ehw <- sandwich::vcovHC(NJimpact, type = 'HC0')
lz.meat <- sandwich::meatCL(NJimpact, cluster = data.use$SOB, type = 'HC0')
lz <- sandwich::vcovCL(NJimpact, cluster = data.use$SOB, type = 'HC0')
bm <- dfadjust::dfadjustSE(NJimpact, clustervar=as.factor(data.use$SOB),
                           IK = FALSE)

stargazer(NJimpact, NJimpact, NJimpact, NJimpact,
          se = list(NULL, sqrt(diag(ehw)), sqrt(diag(lz)),
                    sqrt(diag(bm$vcov))))
```