TAKEHOME EXAM STAT 6869 SPRING 2000

ECONOMICS DEPARTMENT, UNIVERSITY OF UTAH

You are allowed and encouraged to cooperate while working on this exam. You may submit solutions with more than one name on them, which will count equally for all authors. But you must understand the solution you are handing in. I will perhaps ask you to demonstrate and explain the solutions in class. The exam is due back at the beginning of class on Wednesday, March 29, 2000, at the beginning of class at 3:10 pm.

Date of exam March 22, 2000.

Problem 16. *Here is a more detailed description of the Wichmann-Hill generator: Its seed is a 3-vector* $\begin{bmatrix} x_1 & y_1 & z_1 \end{bmatrix}^\top$ *satisfying*

$$0 < x_1 \leq 30269 \tag{1}$$

$$0 < y_1 \leq 30307 \tag{2}$$

$$0 < z_1 \leq 30323 \tag{3}$$

A call to the random generator updates the seed as follows:

$$x_2 = 171x_1 \bmod 30269 \tag{4}$$

$$y_2 = 172y_1 \bmod 30307 \tag{5}$$

$$z_2 = 170z_1 \bmod 30323 \tag{6}$$

and then it returns

$$\left(\frac{x_2}{30269} + \frac{y_2}{30307} + \frac{z_2}{30323}\right) \bmod 1 \tag{7}$$

as its latest drawing from a uniform distribution. If you have R *on your computer, do parts b and c, otherwise do a and b.*

• **a.** *4 points Program the Wichmann-Hill random generator in the programming language of your choice.*

Answer. A random generator does two things:

- It takes the current seed (or generates one if there is none), computes the next seed from it, and stores this next seed on disk as a side effect.
- Then it converts this next seed into a number between 0 and 1.

The `ecmet` package has two demonstration functions which perform these two tasks separately for the Wichmann-Hill generator, without side effects. The function `next.WHseed()` computes the next seed from its argument (which defaults to the seed stored in the official variable `.Random.seed`), and the function `WH.from.current.seed()` gets a number between 0 and 1 from its argument (which has the same default). Both functions are one-liners:

```
next.WHseed <- function(integer.seed = .Random.seed[-1])
  (c(  171,   172,   170) * integer.seed) %% c(30269, 30307, 30323)

WH.from.current.seed <- function(integer.seed = .Random.seed[-1])
  sum(integer.seed / c(30269, 30307, 30323)) %% 1
```

□

• **b.** *2 points Check that the 3 first numbers returned by the Wichmann-Hill random number generator after setting the seed to 1 10 2000 are 0.2759128 0.8713303 0.6150737. (one digit in those 3 numbers is wrong; which is it, and what is the right digit?)*

Answer. The R-code doing this is `ecmet.script(wichhill)`:

```
##This script generates 3 consecutive seeds, with the
##initial seed set as (1, 10, 2000), puts them into a matrix,
```

```
##and then generates the random numbers from the rows of
##this matrix:

  my.seeds <- matrix(nrow=3, ncol=3)

  my.seeds[1,] <- next.WHseed(c(1, 10, 2000))
  my.seeds[2,] <- next.WHseed(my.seeds[1,])
  my.seeds[3,] <- next.WHseed(my.seeds[2,])

  my.unif <- c(WH.from.current.seed(my.seeds[1,]),
               WH.from.current.seed(my.seeds[2,]),
               WH.from.current.seed(my.seeds[3,]))
```

□

• **c.** *4 points Check that the Wichmann-Hill random generator built into* R *is identical to the one described here.*

Answer. First make sure that R will actually use the Wichmann-Hill generator (since it is not the default): `RNGkind("Wichmann-Hill")`. Then call `runif(1)`. (This sets a seed if there was none, or uses the existing seed if there was one.) `.Random.seed[-1]` shows present value of the random seed associated with this last call, dropping 1st number which indicates which random generator this is for, which is not needed for our purposes. Therefore `WH.from.current.seed()`, which takes `.Random.seed[-1]` as default argument, should give the same result as the last call of the official random generator. And `WH.from.current.seed(next.WHseed())` takes the current seed, computes the next seed from it, and converts this next seed into a number between 0 and 1. It does not write

the updated random seed back. Therefore if we issue now the official call `runif(1)` again, we should get the same result. □

Problem 17. [Sta99, p. 200] *As an example showing what is involved in the RSA algorithm, first generate the private and public keys as follows:*

- **a.** *2 points Select two primes, $p = 3$ and $q = 11$. The modulus of our encryption algorithm is their product $r = pq = 33$. Enumerate all numbers < 33 which are coprime to 33. You should come up with $\phi(r) = (3-1)(11-1) = 20$ numbers.*

Answer. 1, 2, 4, 5, 7, 8, 10, 13, 14, 16, 17, 19, 20, 23, 25, 26, 28, 29, 31, 32. □

- **b.** *2 points Now we have to select s such that s is relatively prime to $\phi(r) = 20$ and less than $\phi(r)$; a possible choice which we will use here is $s = 7$. To get a t such that $st \bmod 20 = 1$ we have to compute $t = s^{\phi(\phi(r))-1} \bmod \phi(r) = s^{\phi(20)-1} \bmod \phi(r)$. First compute $\phi(20)$ and then t.*

Answer. The numbers coprime with 20 are 1, 3, 7, 9, 11, 13, 17, 19. Therefore $\phi(20) = 8$. Therefore $t = 7^7 \bmod 20 = 823543 \bmod 20 = 3$. One easily verifies that $t = 3$ is correct because $st = 7 \cdot 3 = 20 + 1$. □

- **c.** *2 points Therefore the public key is $\{7, 33\}$ and the private key $\{t, 33\}$ with the t just computed. Now take a plaintext consisting of the number 5, use the public key to encrypt it. What is the encrypted text? Use the private key to decrypt it again.*

Answer. If the plaintext $= 5$, then encryption is the computation of $5^7 \bmod 33 = 78125 \bmod 33 = 14$. Decryption is the computation of $14^3 \bmod 33 = 2744 \bmod 33 = 5$. □

• **d.** *1 point This procedure is only valid if the plaintext is coprime with t. What should be done about this?*

Answer. Nothing. t is huge, and if it is selected in such a way that it does not have many different prime multipliers, the chance that a text happens to be not coprime with it is minuscule. □

• **e.** *2 points Now take the same plaintext and use the private key to encrypt it. What is the encrypted text? Then use the public key to decrypt it.*

Answer. If the plaintext $= 5$, then encryption is the computation of $5^3 \bmod 33 = 125 \bmod 33 = 26$. Decryption is the computation of $26^7 \bmod 33 = 8031810176 \bmod 33 = 5$. □

Problem 18. *This is an example adapted from* [GG95]*, which is also discussed in* [Spr98, pp. 2/3 and 375–379]*. Table 1 contains artificial data about two firms hiring in the same labor market. For the sake of the argument it is assumed that both firms receive the exact same number of applications (100), and both firms hire 11 new employees. Table 1 shows how many of the applicants and how many of the new hires were Minorities.*

• **a.** *3 points Let p_1 be the proportion of Minorities hired, and p_2 the proportion Majorities hired. Compute the difference $p_1 - p_2$ and the odds ratio $\big(p_1/(1-p_1)\big)/(p_2/(1-$*

Hirings by Two Different Firms with 100 Applications Each				
	Firm A		*Firm B*	
	Minority	*Majority*	*Minority*	*Majority*
Hired	*1*	*10*	*2*	*9*
Not Hired	*31*	*58*	*46*	*43*

TABLE 1. Which Firm's Hiring Policies are More Equitable?

$p_2)\big)$ *for each firm. Which of the two firms seems to discriminate more? Is the difference of probabilities or the odds ratio the more relevant statistic here?*

Answer. In firm A, 3.125% of the minority applicants and 14.7% of the Majority applicants were hired. The difference of the probabilities is 11.581% and the odds ratio is $\frac{29}{155} = 0.1871$. In firm B, 4.167% of the minority applicants and 17.308% of the majority applicants were hired. The difference is 13.141% and the odds ratio $\frac{43}{207} = 0.2077$. On both accounts, firm A seems to discriminate more.

In order to decide which statistic is more relevant we need to know the purpose of the comparison. The *difference* is more relevant if one wants to assess the macroeconomic implications of discrimination. The *odds ratio* is more relevant if one wants to know the impact of discrimination on one individual. □

• **b.** *1 point Government agencies enforcing discrimination laws traditionally have been using the* selection ratio p_1/p_2. *Compute the selection ratio for both firms.*

Answer. In firm A, the selection ratio is $\frac{1}{32}\frac{68}{10} = \frac{17}{80} = 0.2125$. In firm B, it is $\frac{13}{54} = 0.2407$. □

• **c.** *3 points Statisticians argue that the selection ratio is a flawed measure of discrimination, see* [Gas88, pp. 207–11 of vol. 1]. *Demonstrate this by comparing firm A with firm C which hires 5 out of 32 black and 40 out of 68 white applicants.*

Hirings by Two Different Firms with 100 Applications Each				
	Firm A		Firm C	
	Minority	Majority	Minority	Majority
Hired	1	10	5	40
Not Hired	31	58	27	28

TABLE 2. Selection Ratio gives Conflicting Verdicts

Answer. In Firm C the selection ratio is $\frac{5}{32}\frac{68}{40} = \frac{17}{64} = 0.265625$. In firm A, the chances for blacks to be hired is 24% that of whites, and in firm C it is 26%. Firm C seems better. But if we compare the chances *not* to get hired we get a conflicting verdict: In firm A the ratio is $\frac{31}{32}\frac{68}{58} = 1.1357$. In firm C it is $\frac{27}{32}\frac{68}{28} = 2.0491$. In firm C, the chances not to get hired is twice as high for Minorities as it is for Whites, in firm A the chances not to get hired are more equal. Here A seems better.

This illustrates an important drawback of the selection ratio: if we compare the chances of *not* being hired instead of those of being hired, we get $(1 - p_1)/(1 - p_2)$ instead of p_1/p_2. There is no simple relationship between these two numbers, indeed $(1 - p_1)/(1 - p_2)$ is not a function of p_1/p_2, although both ratios should express the same concept. This is why one can get conflicting information if one looks at the selection ratio for a certain event or the selection ratio for its complement.

The odds ratio and the differences in probabilities do not give rise to such discrepancies: the odds ratio for not being hired is just the inverse of the odds ratio for being hired, and the difference in the probabilities of not being hired is the negative of the difference in the probabilities of being hired.

As long as p_1 and p_2 are both close to zero, the odds ratio is approximately equal to the selection ratio, therefore in this case the selection ratio is acceptable despite the above criticism.

□

• **d.** *3 points Argue whether Fisher's exact test, which is a conditional test, is appropriate in this example.*

Answer. The firms do not have control over the number of job applications, and they also do not have control over how many job openings they have. Here is a situation in which Fisher's exact test, which is conditional on the row sums and column sums of the table, is entirely appropriate. Note that this criterion has nothing to do with sample size. □

• **e.** *4 points Compute the significance levels for rejecting the null hypothesis of equal treatment with the one-sided alternative of discrimination for each firm using Fisher's exact test. You will get a counterintuitive result. How can you explain this result?*

Answer. The R-commands can be run as `ecmet.script(hiring)`. Although firm A hired a lower percentage of applicants than firm B, the significance level for discrimination on Fisher's exact test is 0.07652 for firm A and 0.03509 for firm B. I.e., in a court of law, firm B might be convicted of discrimination, but firm A, which hired a lower percentage of its minority applicants, could not.

[Spr98, p. 377] explains this as follows: "the smaller number of minority hirings reduces the power of Fisher's exact test applied to firm A relative to the power where there is a surplus of minority hirings (firm B). This extra power is enough to produce a significant result despite the higher percentage of promotions among minority hirings (or the higher odds ratio if one makes the comparison on that basis)."

□

Promotions by Two Different Firms with 100 Employees Each				
	Firm A		*Firm B*	
	Minority	*Majority*	*Minority*	*Majority*
Promoted	*1*	*10*	*2*	*9*
Not Promoted	*31*	*58*	*46*	*43*

TABLE 3. Which Firm's Promotion Policies are More Equitable?

• **f.** *5 points Now let's change the example. Table 3 has the same numbers as Table 1, but now these numbers do not count hirings but promotions from the pool of existing employees, and instead of the number of applicants, the column totals are the total numbers of employees of each firm. Let us first look at the overall race composition of the employees in each firm. Let us assume that 40% of the population are minorities, and 32 of the 100 employees of firm A are minorities, and 48 of the 100 employees*

of firm B are minorities. Is there significant evidence that the firms discriminated in hiring*?*

Answer. Assuming that the population is infinite, the question is: if one makes 100 independent random drawings from a population that contains 40% minorities, what is the probability to end up with 32 or less minorities? The R-command is `pbinom(q=32,size=100,prob=0.4)` which is 0.06150391. The other firm has more than 40 black employees; here one might wonder if there is evidence of discrimination against whites. `pbinom(q=48,size=100,prob=0.4)` gives $0.9576986 = 1 - 0.0423$, i.e., it is significant at the 5% level. But here we should apply a two-sided test. A one-sided test about discrimination against Blacks can be justified by the assumption "if there is discrimination at all, it is against blacks." This assumption cannot be made in the case of discrimination against Whites. We have to allow for the possibility of discrimination against Minorities *and* against Whites; therefore the critical value is at probability 0.975, and the observed result is not significant.

□

• **g.** *2 points You want to use Table 3 to investigate whether the firms discriminated in* promotion*, and you are considering Fisher's exact test. Do the arguments made above with respect to Fisher's exact still apply?*

Answer. No. A conditional test is no longer appropriate here because the proportion of candidates for promotion is under control of the firms. Firm A not only promoted a smaller percentage of their minority employees, but it also hired fewer minority workers in the first place. These two acts should be considered together to gauge the discrimination policies. The above Sprent-quote [Spr98, p. 377] continues: "There is a timely warning here about the need for care when using conditional tests when the marginal totals used for conditioning may themselves be conditional upon a further factor, in this case hiring policy."

□

Problem 19. *Suppose that 60% of whites are hired, while only 40% of a minority group are hired. Suppose that a certain type of training or education was related to the job in question, and it is believed that at least 10% of the minority group had this training.*

- **a.** *3 points Assuming that persons with this training had twice the chance of getting the job, which percentage of whites would have had this qualification in order to explain the disparity in the hiring rates?*

Answer. Since 60% of whites are hired and 40% of the minority group, $r_m = 60/40 = 1.5$. Training is the factor x. Sind persons with training had twice the chance of getting the job, $r_x = 2$. Since 10% of the minority group had this training, $f_1 = 0.1$. Therefore (**??**) implies that at least $1.5 \cdot 0.1 + \frac{0.5}{1} = 65\%$ of whites had to have this qualification in order to explain the observed disparity in hiring rates. □

- **b.** *1 point What would this percentage have to be if training tripled (instead of doubling) one's chances of getting the job?*

Answer. If training tripled one's chances of being hired, then the training would explain the disparity if $1.5 \cdot 0.1 + \frac{0.5}{2} = 40\%$ or more of whites had this training. □

Problem 20. *The following is the example given in* [Coo77]. *In* R, *the command* `data(longley)` *makes the data frame* `longley` *available, which has the famous*

Longley-data, a standard example for a highly multicollinear dataset. These data are also available on the web at `www.econ.utah.edu/ehrbar/data/longley.txt`. `attach(longley)` *makes the individual variables available as* R*-objects.*

• **a.** *3 points Look at the data in a scatterplot matrix and explain what you see. Later we will see that one of the observations is in the regression much more influential than the rest. Can you see from the scatterplot matrix which observation that might be?*

Answer. In linux, you first have to give the command `x11()` in order to make the graphics window available. In windows, this is not necessary. It is important to display the data in a reasonable order, therefore instead of `pairs(longley)` you should do something like `attach(longley)` and then `pairs(cbind(Year, Population, Employed, Unemployed, Armed.Forces, GNP, GNP.deflator))`. P `Year` first, so that all variables are plotted against `Year` on the horizontal axis.

Population vs. year is a very smooth line.

Population vs GNP also quite smooth.

You see the huge increase in the armed forced in 1951 due to the Korean War, which led to a (temporary) drop in unemployment and a (not so temporary) jump in the GNP deflator.

Otherwise the unemployed show the stop-and-go scenario of the fifties.

unemployed is not correlated with anything.

One should expect a strong negative correlation between employed and unemployed, but this is not the case. □

• **b.** *4 points Run a regression of the model* `Employed ~ GNP.deflator + GNP + Unemployed + Armed.Forces + Population + Year` *and discuss the result.*

Answer. To fit a regression run `longley.fit <- lm(Employed ~ GNP + Unemployed + Armed.Forces + Population + Year)`. You can see the regression results by typing `summary(longley.fit)`.

Armed forces and unemployed are significant and have negative sign, as expected.

GNP and Population are insignificant and have negative sign too, this is not expected. GNP, Population and Year are highly collinear.

□

• **c.** *3 points Make plots of the ordinary residuals and the standardized residuals against time. How do they differ? In* R*, the commands are* `plot(Year, residuals(lo type="h", ylab="Ordinary Residuals in Longley Regression")`*. In order to get the next plot in a different graphics window, so that you can compare them, do now either* `x11()` *in linux or* `windows()` *in windows, and then* `plot(Year, rstandard(longley.fit), type="h", ylab="Standardized Residuals in Longl Regression")`*.*

Answer. You see that the standardized residuals at the edge of the dataset are bigger than the ordinary residuals. The datapoints at the edge are better able to attract the regression plane than those in the middle, therefore the ordinary residuals are "too small." Standardization corrects for this. □

• **d.** *4 points Make plots of the predictive residuals. Apparently there is no special command in* R *to do this, therefore you should use formula* (**??**)*. Also plot the standardized predictive residuals, and compare them.*

Answer. The predictive residuals are `plot(Year, residuals(longley.fit)/(1-hatvalues(longley.f type="h", ylab="Predictive Residuals in Longley Regression")`. The standardized predictive residuals are often called studentized residuals, `plot(Year, rstudent(longley.fit), type="h", ylab="Standardized predictive Residuals in Longley Regression")`.

A comparison shows an opposite effect as with the ordinary residuals: the predictive residuals at the edge of the dataset are too *large*, and standardization corrects this.

Specific results: standardized predictive residual in 1950 smaller than that in 1962, but predictive residual in 1950 is very close to 1962.

standardized predictive residual in 1951 smaller than that in 1956, but predictive residual in 1951 is larger than in 1956.

Largest predictive residual is 1951, but largest standardized predictive residual is 1956.

□

- **e.** *3 points Make a plot of the leverage, i.e., the h_{ii}-values, using* `plot(Year, hatvalues(longley.fit), type="h", ylab="Leverage in Longley Regression"` *and explain what leverage means.*

- **f.** *3 points One observation is much more influential than the others; which is it? First look at the plots for the residuals, then look also at the plot for leverage, and try to guess which is the most influential observation. Then do it the right way. Can you give reasons based on your prior knowledge about the time period involved why an observation in that year might be influential?*

Answer. The "right" way is to use Cook's distance: `plot(Year, cooks.distance(longley.fit), type="h", ylab="Cook's Distance in Longley Regression")`

One sees that 1951 towers above all others. It does not have highest leverage, but it has second-highest, and a bigger residual than the point with the highest leverage.

1951 has the largest distance of .61. The second largest is the last observation in the dataset, 1962, with a distance of .47, and the others have .24 or less. Cook says: removal of 1951 point will move the least squares estimate to the edge of a 35% confidence region around $\hat{\beta}$. This point is probably so influential because 1951 was the first full year of the Korean war. One would not be able to detect this point from the ordinary residuals, standardized or not! The predictive residuals are a little better; their maximum is at 1951, but several other residuals are almost as large. 1951 is so influential because it has an extremely high hat-value, and one of the highest values for the ordinary residuals! □

At the end don't forget to `detach(longley)` *if you have attached it before.*

Problem 21. *7 points One general inequality measure which generalizes both Theil's and (up to an affine transformation) Herfindahl's indices is constructed as follows: choose* $\beta \geq 0$ *and set* $h(s) = \ln s$ *if* $\beta = 0$ *and* $h(s) = -\frac{1}{\beta}s^\beta$ *otherwise; then the inequality measure is*

$$I = \frac{1}{1+\beta}\Big(\sum_{i=1}^{n} \tfrac{1}{n} h(\tfrac{1}{n}) - \sum_{i=1}^{n} s_i h(s_i)\Big) \tag{8}$$

where s_i *is the ith person's share in total income. Show that if one takes a small amount of income share* ds *from person 2 and adds it to person 1, then the inequality*

measure defined in (8) *changes by* $\big(h(s_2) - h(s_1)\big)ds$. *Hint: if* $\beta \neq 0$,

$$\frac{\partial I}{\partial s_i} = \frac{\partial}{\partial s_i} s_i h(s_i) = \frac{1}{\beta} s_i^{\beta} = -h(s_i). \tag{9}$$

If one therefore takes ds *away from 2 and gives it to 1,* I *changes by*

$$dI = \Big(-\frac{\partial I}{\partial s_2} + \frac{\partial I}{\partial s_1}\Big)ds = \Big(h(s_2) - h(s_1)\Big)ds \tag{10}$$

If $\beta = 0$, *only small modifications apply.*

Answer. If $\beta = 0$, then

$$I = \Big(\ln(\tfrac{1}{n}) - \sum_{i=1}^{n} s_i \ln(s_i)\Big) \tag{11}$$

therefore one has in this case

$$\frac{\partial I}{\partial s_i} = \frac{\partial}{\partial s_i} s_i h(s_i) = -\ln(s_i) - 1 = -h(s_i) - 1 \tag{12}$$

But the extra -1 cancels in the difference. □

Interpretation: if $h(s_2) - h(s_1) = h(s_4) - h(s_3)$ *then for the purposes of this inequality measure, the distance between 2 and 1 is the same as the distance between*

4 and 3. These inequality measures are therefore based on very specific notions of what constitutes inequality.

-

Maximum number of points: 71.

References

[Coo77] R. Dennis Cook, *Detection of influential observations in linear regression*, Technometrics **19** (1977), no. 1, 15–18. 12

[Gas88] Joseph L. Gastwirth, *Statistical reasoning in law and public policy*, Statistical modeling and decision science, Academic Press, Boston, 1988. 8

[GG95] Joseph L. Gastwirth and S. W. Greenhouse, *Biostatistical concepts and methods in the legal setting*, Statistics in Medicine **14** (1995), 1641–53. 6

[Spr98] Peter Sprent, *Data driven statistical methods*, 1st ed. ed., Texts in statistical science, Chapman & Hall, London; New York, 1998. 6, 10, 11

[Sta99] William Stallings, *Cryptography and network security: Principles and practice*, 2nd ed., Prentice Hall, Upper Saddle River, N.J., 1999. 5

Economics Department, University of Utah, 1645 Campus Center Drive, Salt Lake City, UT 84112-9300, U.S.A

E-mail address: ehrbar@econ.utah.edu

URL: http://www.econ.utah.edu/ehrbar