TAKEHOME EXAM ECON 7800 FALL 2003

ECONOMICS DEPARTMENT, UNIVERSITY OF UTAH

For the computer assignments it is not necessary to submit the printouts, but simply a short verbal answer describing what you see and what the main features of the regressions are. A pdf-version of this exam is in http://www.econ.utah.edu/ehrbar/2003faTH.pdf. After the exam there will be a version with the answers at the same location. You will also find the two earlier exams this Semester at http://www.econ.utah.edu/ehrbar/2003faM1.pdf and http://www.econ.utah.edu/ehrbar/2003faM2.pdf.

For the benefit of those who will take the field exam I also uploaded two old field exams which I had on my hard disk (with possibly updated wording of the questions, and with answers). They are at http://www.econ.utah.edu/ehrbar/2000TFLD.pdf and http://www.econ.utah.edu/ehrbar/2001TFLD.pdf.

Date of exam Due Date is Tuesday, December 2nd, 12:25 pm. No late exams are accepted.

Problem 36. *Two researchers counted cars coming down a road, which obey a Poisson distribution with unknown parameter λ. In other words, in an interval of length t one will have k cars with probability*

$$\frac{(\lambda t)^k}{k!}e^{-\lambda t}. \tag{1}$$

Their assignment was to count how many cars came in the first half hour, and how many cars came in the second half hour. However they forgot to keep track of the time when the first half hour was over, and therefore wound up only with one count, namely, they knew that 213 cars had come down the road during this hour. They were afraid they would get fired if they came back with one number only, so they applied the following remedy: they threw a coin 213 times and counted the number of heads. This number, they pretended, was the number of cars in the first half hour.

● **a.** *6 points Did the probability distribution of the number gained in this way differ from the distribution of actually counting the number of cars in the first half hour?*

Answer. First a few definitions: x is the total number of occurrences in the interval $[0, 1]$. y is the number of occurrences in the interval $[0, t]$ (for a fixed t; in the problem it was $t = \frac{1}{2}$, but we will do it for general t, which will make the notation clearer and more compact. Then we want to compute $\Pr[y{=}m|x{=}n]$. By definition of conditional probability:

$$\Pr[y{=}m|x{=}n] = \frac{\Pr[y{=}m \text{ and } x{=}n]}{\Pr[x{=}n]}. \tag{2}$$

How can we compute the probability of the intersection $\Pr[y{=}m \text{ and } x{=}n]$? Use a trick: express this intersection as the intersection of independent events. For this define z as the number of events in the interval $(t, 1]$. Then $\{y{=}m \text{ and } x{=}n\} = \{y{=}m \text{ and } z{=}n-m\}$; therefore $\Pr[y{=}m \text{ and } x{=}n] = \Pr[y{=}m]\Pr[z{=}n-m]$; use this to get

(3)

$$\Pr[y{=}m|x{=}n] = \frac{\Pr[y{=}m]\Pr[z{=}n-m]}{\Pr[x{=}n]} = \frac{\frac{\lambda^m t^m}{m!}e^{-\lambda t}\frac{\lambda^{n-m}(1-t)^{n-m}}{(n-m)!}e^{-\lambda(1-t)}}{\frac{\lambda^n}{n!}e^{-\lambda}} = \binom{n}{m}t^m(1-t)^{n-m},$$

Here we use the fact that $\Pr[x{=}k] = \frac{t^k}{k!}e^{-t}$, $\Pr[y{=}k] = \frac{(\lambda t)^k}{k!}e^{-\lambda t}$, $\Pr[z{=}k] = \frac{(1-\lambda)^k t^k}{k!}e^{-(1-\lambda)t}$. One sees that a. $\Pr[y{=}m|x{=}n]$ does not depend on λ, and b. it is exactly the probability of having m successes and $n-m$ failures in a Bernoulli trial with success probability t. Therefore the procedure with the coins gave the two researchers a result which had the same probability distribution as if they had counted the number of cars in each half hour separately.

□

• **b.** *2 points Explain what it means that the probability distribution of the number for the first half hour gained by throwing the coins does not differ from the one gained by actually counting the cars. Which condition is absolutely necessary for this to hold?*

Answer. The supervisor would never be able to find out through statistical analysis of the data they delivered, even if they did it repeatedly. All estimation results based on the faked statistic would be as accurate regarding λ as the true statistics. All this is only true under the assumption that the cars really obey a Poisson distribution and that the coin is fair.

The fact that the Poisson as well as the binomial distributions are memoryless has nothing to do with them having a sufficient statistic.

□

Problem 37. *You have two unbiased measurements with errors of the same quantity μ (which may or may not be random). The first measurement y_1 has mean squared error $\mathrm{E}[(y_1 - \mu)^2] = \sigma^2$, the other measurement y_2 has $\mathrm{E}[(y_1 - \mu)^2] = \tau^2$. The measurement errors $y_1 - \mu$ and $y_2 - \mu$ have zero expected values (i.e., the measurements are unbiased) and are independent of each other.*

- **a.** *2 points Show that the linear unbiased estimators of μ based on these two measurements are simply the weighted averages of these measurements, i.e., they can be written in the form $\tilde{\mu} = \alpha y_1 + (1 - \alpha) y_2$, and that the MSE of such an estimator is $\alpha^2\sigma^2 + (1-\alpha)^2\tau^2$. Note: we are using the word "estimator" here even if μ is random. An estimator or predictor $\tilde{\mu}$ is unbiased if $\mathrm{E}[\tilde{\mu} - \mu] = 0$. Since we allow μ to be random, the proof in the class notes has to be modified.*

Answer. The estimator $\tilde{\mu}$ is linear (more precisely: affine) if it can written in the form

$$\tilde{\mu} = \alpha_1 y_1 + \alpha_2 y_2 + \gamma \tag{4}$$

The measurements themselves are unbiased, i.e., $\mathrm{E}[y_i - \mu] = 0$, therefore

$$\mathrm{E}[\tilde{\mu} - \mu] = (\alpha_1 + \alpha_2 - 1)\,\mathrm{E}[\mu] + \gamma = 0 \tag{5}$$

for all possible values of $\mathrm{E}[\mu]$; therefore $\gamma = 0$ and $\alpha_2 = 1 - \alpha_1$. To simplify notation, we will call from now on $\alpha_1 = \alpha$, $\alpha_2 = 1 - \alpha$. Due to unbiasedness, the MSE is the variance of the estimation error

$$\operatorname{var}[\tilde{\mu} - \mu] = \alpha^2\sigma^2 + (1-\alpha)^2\tau^2 \tag{6}$$

□

- **b.** *4 points Define ω^2 by*

$$\frac{1}{\omega^2} = \frac{1}{\sigma^2} + \frac{1}{\tau^2} \qquad \textit{which can be solved to give} \qquad \omega^2 = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}. \tag{7}$$

Show that the Best (i.e., minimum MSE*) linear unbiased estimator (BLUE) of μ based on these two measurements is*

$$\hat{y} = \frac{\omega^2}{\sigma^2}y_1 + \frac{\omega^2}{\tau^2}y_2 \tag{8}$$

i.e., it is the weighted average of y_1 and y_2 where the weights are proportional to the inverses of the variances.

Answer. The variance (6) takes its minimum value where its derivative with respect of α is zero, i.e., where

$$\frac{\partial}{\partial\alpha}\bigl(\alpha^2\sigma^2+(1-\alpha)^2\tau^2\bigr)=2\alpha\sigma^2-2(1-\alpha)\tau^2=0 \tag{9}$$

$$\alpha\sigma^2=\tau^2-\alpha\tau^2 \tag{10}$$

$$\alpha=\frac{\tau^2}{\sigma^2+\tau^2} \tag{11}$$

In terms of ω one can write

$$\alpha=\frac{\tau^2}{\sigma^2+\tau^2}=\frac{\omega^2}{\sigma^2}\qquad\text{and}\qquad 1-\alpha=\frac{\sigma^2}{\sigma^2+\tau^2}=\frac{\omega^2}{\tau^2}. \tag{12}$$

□

- **c.** *2 points Show: the* MSE *of the BLUE* ω^2 *satisfies the following equation:*

$$\frac{1}{\omega^2}=\frac{1}{\sigma^2}+\frac{1}{\tau^2} \tag{13}$$

Answer. We already have introduced the notation ω^2 for the quantity defined by (13); therefore all we have to show is that the MSE or, equivalently, the variance of the estimation error is equal to this ω^2:

$$\operatorname{var}[\tilde\mu-\mu]=\Bigl(\frac{\omega^2}{\sigma^2}\Bigr)^2\sigma^2+\Bigl(\frac{\omega^2}{\tau^2}\Bigr)^2\tau^2=\omega^4\Bigl(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\Bigr)=\omega^4\frac{1}{\omega^2}=\omega^2 \tag{14}$$

□

Problem 38. *2 points Assume each $y_i \sim N(\mu, 1)$, $n = 400$ and $\alpha = 0.05$, and different y_i are independent. Compute the value c which satisfies $\Pr[\bar{y} > c \,|\, \mu = 0] = \alpha$. You shoule either look it up in a table and include a xerox copy of the table with the entry circled and the complete bibliographic reference written on the xerox copy, or do it on a computer, writing exactly which commands you used. In* R*, the function* `qnorm` *does what you need, find out about it by typing* `help(qnorm)`*.*

Answer. In the case $n = 400$, $\bar{y}$ has variance $1/400$ and therefore standard deviation $1/20 = 0.05$. Therefore $20\bar{y}$ is a standard normal: from $\Pr[\bar{y} > c \,|\, \mu = 0] = 0.05$ follows $\Pr[20\bar{y} > 20c \,|\, \mu = 0] = 0.05$. Therefore $20c = 1.645$ can be looked up in a table, perhaps use [JHG+88, p. 986], the row for ∞ d.f.

Let us do this in R. The p-"quantile" of the distribution of the random variable y is defined as that value q for which $\Pr[y \leq q] = p$. If y is normally distributed, this quantile is computed by the R-function `qnorm(p, mean=0, sd=1, lower.tail=TRUE)`. In the present case we need either `qnorm(p=1-0.05, mean=0, sd=0.05)` or `qnorm(p=0.05, mean=0, sd=0.05, lower.tail=FALSE)` which gives the value 0.08224268.

□

Problem 39. *2 points Prove that, if one predicts a random variable y by a constant a, the constant which gives the best* MSE *is $a = \mathrm{E}[y]$, and the best* MSE *one can get is* $\mathrm{var}[y]$.

Answer. $\mathrm{E}[(y - a)^2] = \mathrm{E}[y^2] - 2a\,\mathrm{E}[y] + a^2$. Differentiate with respect to a and set zero to get $a = \mathrm{E}[y]$. One can also differentiate first and then take expected value: $\mathrm{E}[2(y - a)] = 0$. □

Problem 40. *Given two random variables* x *and* y *with finite variances, and* $\operatorname{var}[x] > 0$. *You know the expected values, variances and covariance of* x *and* y*, and you observe* x*, but* y *is unobserved. This question explores the properties of the Best Linear Unbiased Predictor (BLUP) of* y *in this situation.*

● **a.** *4 points Give a mathematical proof of the following: If you want to predict* y *by an affine expression of the form* $a + bx$*, you will get the lowest mean squared error* MSE *with* $b = \operatorname{cov}[x, y]/\operatorname{var}[x]$ *and* $a = \operatorname{E}[y] - b\operatorname{E}[x]$.

Answer. The MSE is variance plus squared bias (see e.g. problem **??**), therefore

$$\operatorname{MSE}[a + bx; y] = \operatorname{var}[a + bx - y] + (\operatorname{E}[a + bx - y])^2 = \operatorname{var}[bx - y] + (a - \operatorname{E}[y] + b\operatorname{E}[x])^2. \tag{15}$$

Therefore we choose a so that the second term is zero, and then you only have to minimize the first term with respect to b. Since

$$\operatorname{var}[bx - y] = b^2 \operatorname{var}[x] - 2b\operatorname{cov}[x, y] + \operatorname{var}[y] \tag{16}$$

the first order condition is

$$2b\operatorname{var}[x] - 2\operatorname{cov}[x, y] = 0 \tag{17}$$

□

● **b.** *2 points For the first-order conditions you needed the partial derivatives* $\frac{\partial}{\partial a}\operatorname{E}[(y - a - bx)^2]$ *and* $\frac{\partial}{\partial b}\operatorname{E}[(y - a - bx)^2]$*. It is also possible, and probably shorter, to interchange taking expected value and partial derivative, i.e., to compute* $\operatorname{E}\left[\frac{\partial}{\partial a}(y - a - bx)^2\right]$

and $\mathrm{E}\Big[\frac{\partial}{\partial b}(y - a - bx)^2\Big]$ *and set those zero. Do the above proof in this alternative fashion.*

Answer. $\mathrm{E}\Big[\frac{\partial}{\partial a}(y - a - bx)^2\Big] = -2\,\mathrm{E}[y - a - bx] = -2(\mathrm{E}[y] - a - b\,\mathrm{E}[x])$. Setting this zero gives the formula for a. Now $\mathrm{E}\Big[\frac{\partial}{\partial b}(y - a - bx)^2\Big] = -2\,\mathrm{E}[x(y - a - bx)] = -2(\mathrm{E}[xy] - a\,\mathrm{E}[x] - b\,\mathrm{E}[x^2])$. Setting this zero gives $\mathrm{E}[xy] - a\,\mathrm{E}[x] - b\,\mathrm{E}[x^2] = 0$. Plug in formula for a and solve for b:

$$b = \frac{\mathrm{E}[xy] - \mathrm{E}[x]\,\mathrm{E}[y]}{\mathrm{E}[x^2] - (\mathrm{E}[x])^2} = \frac{\operatorname{cov}[x, y]}{\operatorname{var}[x]}. \tag{18}$$

□

- **c.** *2 points Compute the* MSE *of this predictor.*

Answer. If one plugs the optimal a into (15), this just annulls the last term of (15) so that the MSE is given by (16). If one plugs the optimal $b = \operatorname{cov}[x, y]/\operatorname{var}[x]$ into (16), one gets

$$\mathrm{MSE} = \Big(\frac{\operatorname{cov}[x, y]}{\operatorname{var}[x]}\Big)^2 \operatorname{var}[x] - 2\frac{(\operatorname{cov}[x, y])}{\operatorname{var}[x]} \operatorname{cov}[x, y] + \operatorname{var}[x] \tag{19}$$

$$= \operatorname{var}[y] - \frac{(\operatorname{cov}[x, y])^2}{\operatorname{var}[x]}. \tag{20}$$

□

- **d.** *2 points Show that the prediction error is uncorrelated with the observed x.*

Answer.

$$\operatorname{cov}[x, y - a - bx] = \operatorname{cov}[x, y] - a\operatorname{cov}[x, x] = 0 \tag{21}$$

□

• **e.** *4 points If* $\operatorname{var}[x] = 0$, *the quotient* $\operatorname{cov}[x, y]/\operatorname{var}[x]$ *can no longer be formed, but if you replace the inverse by the g-inverse, so that the above formula becomes*

$$b = \operatorname{cov}[x, y](\operatorname{var}[x])^{-} \tag{22}$$

then it always gives the minimum MSE *predictor, whether or not* $\operatorname{var}[x] = 0$, *and regardless of which g-inverse you use (in case there are more than one). To prove this, you need to answer the following four questions: (a) what is the BLUP if* $\operatorname{var}[x] = 0$*? (b) what is the g-inverse of a nonzero scalar? (c) what is the g-inverse of the scalar number 0? (d) if* $\operatorname{var}[x] = 0$, *what do we know about* $\operatorname{cov}[x, y]$*?*

Answer. (a) If $\operatorname{var}[x] = 0$ then $x = \mu$ almost surely, therefore the observation of $\boldsymbol{x}$ does not give us any new information. The BLUP of y is ν in this case, i.e., the above formula holds with $b = 0$.

(b) The g-inverse of a nonzero scalar is simply its inverse.

(c) Every scalar is a g-inverse of the scalar 0.

(d) if $\operatorname{var}[x] = 0$, then $\operatorname{cov}[x, y] = 0$.

Therefore pick a g-inverse 0, an arbitrary number will do, call it c. Then formula (22) says $b = 0 \cdot c = 0$. □

Problem 41. *4 points Show that the proportionate reduction in the* MSE *of the best predictor of y, if one goes from predictors of the form $y^* = a$ to predictors of the form $y^* = a + bx$, is equal to the squared correlation coefficient between y and x. You are allowed to use the results of Problems 39 and 40. To set notation, call the minimum* MSE *in the first prediction (Problem 39)* MSE[*constant term*; y], *and the minimum* MSE *in the second prediction (Problem 40)* MSE[*constant term and* x; y]. *Show that*

$$\text{(23)} \qquad \frac{\text{MSE}[\textit{constant term}; y] - \text{MSE}[\textit{constant term and } x; y]}{\text{MSE}[\textit{constant term}; y]} = \frac{(\operatorname{cov}[y, x])^2}{\operatorname{var}[y]\operatorname{var}[x]} = \rho^2_{yx}.$$

Answer. The minimum MSE with only a constant is $\operatorname{var}[y]$ and (20) says that MSE[constant term and $x; y] = \operatorname{var}[y] - (\operatorname{cov}[x, y])^2 / \operatorname{var}[x]$. Therefore the difference in MSE's is $(\operatorname{cov}[x, y])^2 / \operatorname{var}[x]$, and if one divides by $\operatorname{var}[y]$ to get the relative difference, one gets exactly the squared correlation coefficient. □

Problem 42. *4 points The electricity demand date from* [Hou51] *are available on the web at* `www.econ.utah.edu/ehrbar/data/ukelec.txt`. *Import these data into your favorite statistics package. For* R *you need the command* `ukelec <- read.table( "http://www.econ.utah.edu/ehrbar/data/ukelec.txt")`. *Make a scatterplot matrix of these data using e.g.* `pairs(ukelec)` *and describe what you see.*

Answer. inc and cap are negatively correlated. cap is capacity of rented equipment and not equipment owned. Apparently customers with higher income buy their equipment instead of renting it.

gas6 and gas8 are very highly correlated. mc4, mc6, and mc8 are less hightly correlated, the corrlation between mc6 and mc8 is higher than that between mc4 and mc6. It seem electicity prices have been coming down.

kwh, inc, and exp are strongly positively correlated.

the stripes in all the plots which have mc4, mc6, or mc8 in them come from the fact that the marginal cost of electricity is a round number.

electricity prices and kwh are negatively correlated.

There is no obvious positive correlation between kwh and cap or expen and cap.

Prices of electricity and gas are somewhat positively correlated, but not much.

When looking at the correlations of inc with the other variables, there are several outliers which could have a strong "leverage" effect.

in 1934, those with high income had lower electricity prices than those with low income. This effect dissipated by 1938.

No strong negative correlations anywhere.

cust negatively correlated with inc, because rich people live in smaller cities?

□

Problem 43. *2 points How would you answer the question whether marginal prices of gas vary more or less than those of electricity (say in the year 1936)?*

Answer. Marginal gas prices vary a little more than electricity prices, although electricity was the newer technology, and although gas prices are much more stable over time than the electricity prices. Compare `sqrt(var(mc6))/mean(mc6)` with `sqrt(var(gas6))/mean(gas6)`. You get 0.176 versus 0.203. Another way would be to compute `max(mc6)/min(mc6)` and compare with `max(gas6)/min(gas6)`: you get 2.27 versus 2.62. In any case this is a lot of variation. □

Problem 44. *2 points Is electricity a big share of total income? Which command is better: `mean(expen/inc)` or `mean(expen)/mean(inc)`? What other options are there? Actually, there is a command which is clearly better than at least one of the above, can you figure out what it is?*

Answer. The proportion is small, less than 1 percent. The two above commands give 0.89% and 0.84%. The command `sum(cust*expen) / sum(cust*inc)` is better than `mean(expen) / mean(inc)`, because each component in `expen` and `inc` is the mean over many households, the number of households given by `cust`. `mean(expen)` is therefore an average over averages over different population sizes, not a good idea. `sum(cust*expen)` is total expenditure in all households involved, and `sum(cust*inc)` is total income in all households involved. `sum(cust*expen) / sum(cust*inc)` gives the value 0.92%. Another option is `median(expen/inc)` which gives 0.91%. A good way to answer this question is to plot it: `plot(expen,inc)`. You get the line where expenditure is 1 percent of income by `abline(0,0.01)`. For higher incomes expenditure for electricity levels off and becomes a lower share of income. □

Problem 45. *Have your computer compute the sample correlation matrix of the data. The R-command is `cor(ukelec)`*

• **a.** *4 points Are there surprises if one looks at the correlation matrix?*

Answer. Electricity consumption `kwh` is slightly negatively correlated with gas prices and with the capacity. If one takes the correlation matrix of the logarithmic data, one gets the expected positive signs.

marginal prices of gas and electricity are positively correlated in the order of 0.3 to 0.45.

higher correlation between mc6 and mc8 than between mc4 and mc6.

Correlation between `expen` and `cap` is negative and low in both matrices, while one should expect positive correlation. But in the logarithmic matrix, `mc6` has negative correlation with `expen`, i.e., elasticity of electricity demand is less than 1.

In the logarithmic data, `cust` has higher correlations than in the non-logarithmic data, and it is also more nearly normally distributed.

`inc` has negative correlation with `mc4` but positive correlation with `mc6` and `mc8`. (If one looks at the scatterplot matrix this seems just random variations in an essentially zero correlation).

mc6 and expen are positively correlated, and so are mc8 and expen. This is due to the one outlier with high expen and high income and also high electricity prices.

The marginal prices of electricity are not strongly correlated with `expen`, and in 1934, they are negatively correlated with `income`.

From the scatter plot of `kwh` versus `cap` it seems there are two datapoints whose removal might turn the sign around. To find out which they are do `plot(kwh,cap)` and then use the identify function: `identify(kwh,cap,labels=row.names(ukelec))`. The two outlying datapoints are Halifax and Wallase. Wallase has the highest income of all towns, namely, 1422, while Halifax's income of 352 is close to the minimum, which is 279. High income customers do not lease their equipment but buy it. □

• **b.** *3 points The correlation matrix says that* **kwh** *is negatively related with* **cap***, but the correlation of the logarithm gives the expected positive sign. Can you explain this behavior?*

Answer. If one plots the date using `plot(cap,kwh)` one sees that the negative correlation comes from the two outliers. In a logarithmic scale, these two are no longer so strong outliers.

□

Problem 46. *6 points Write up the main results from the regressions which in* R *are run by the commands*

```
houth.olsfit <- lm(formula = kwh ~ inc+I(1/mc6)+gas6+cap)
houth.glsfit <- lm(kwh ~ inc+I(1/mc6)+gas6+cap, weight=cust)
houth.olsloglogfit <- lm(log(kwh) ~
log(inc)+log(mc6)+log(gas6)+log(cap))
```

Instead of `1/mc6` *you had to type* `I(1/mc6)` *because the slash has a special meaning in formulas, creating a nested design, therefore it had to be "protected" by applying the function* `I()` *to it.*

If you then type `houth.olsfit`*, a short summary of the regression results will be displayed on the screen. There is also the command* `summary(houth.olsfit)`*, which gives you a more detailed summary. If you type* `plot(houth.olsfit)` *you will get a series of graphics relevant for this regression.*

Answer. All the expected signs.

Gas prices do not play a great role in determining electricity consumption, despite the "cookers" Berndt talks about on p. 337. Especially the logarithmic regression makes gas prices highly insignificant!

The weighted estimation has a higher R^2. □

Problem 47. *4 points Although there is good theoretical justification for using* ***cust*** *as weights, one might wonder if the data bear this out. How can you check this?*

Answer. Do `plot(cust, rstandard(houth.olsfit))` and `plot(cust, rstandard(houth.glsfit))`. In the first plot, smaller numbers of customers have larger residuals, in the second plot this is mitigated. Also the OLS plot has two terrible outliers, which are brought more into range with GLS. □

Problem 48. *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))`. Pu `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.*

Maximum number of points: 83.

References

[Coo77] R. Dennis Cook, *Detection of influential observations in linear regression*, Technometrics **19** (1977), no. 1, 15–18. 16

[Hou51] H. S. Houthakker, *Some calculations on electricity consumption in Great Britain*, Journal of the Royal Statistical Society (A) (1951), no. 114 part III, 351–371, J. 11

[JHG+88] George G. Judge, R. Carter Hill, William E. Griffiths, Helmut Lütkepohl, and Tsoung-Chao Lee, *Introduction to the theory and practice of econometrics*, second ed., Wiley, New York, 1988. 7

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