

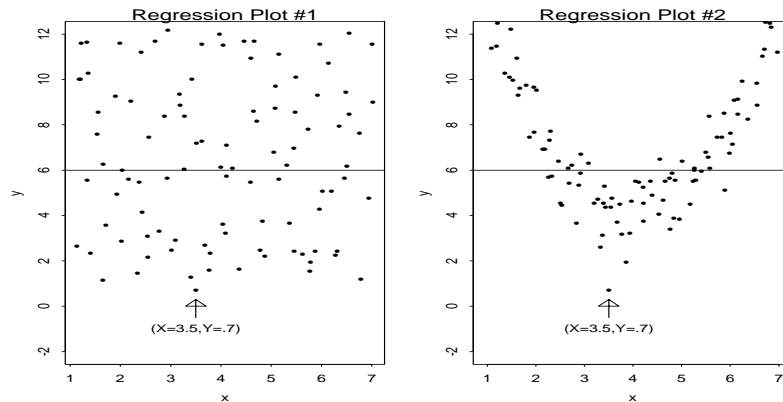
Quantitative Political Methodology II (Fall), Course: 582. Homework #1.

1. Show that the product of idempotent matrices is idempotent.
2. A Hilbert matrix has elements $x_{ij} = 1/(i + j - 1)$ for the entry in row i and column j . Is this always a symmetric matrix? Is it always positive definite?
3. Consider the following analysis that uses Florida county-by-county voting data for Bush and Buchanan in the 2000 Presidential election.

```
> library(UsingR); data(floriga)
> floriga.lm = lm(BUCHANAN ~ BUSH, data=floriga)
> dfbetas(floriga.lm)
> plot(BUCHANAN ~ BUSH, data=floriga,pch="+")
> abline(floriga.lm,lty=1)
> text(floriga$BUSH[13]-10000,floriga$BUCHANAN[13]+150,"Dade")
> text(floriga$BUSH[50],floriga$BUCHANAN[50]-200,"Palm Beach")
> summary(floriga.lm)
```

Is this regression analysis reliable? State specifically why or why not. Using the `dfbetas` output diagnostic and some graphical evidence that you generate, state which cases are the most influential.

4. Below are two sets of data each with least square regression lines calculated ($\hat{y} = 6 + 0x$). Answer the following questions by looking at the plots.



- (a) Does the construction of the least squares line in panel 1 violate any of the Gauss-Markov assumptions?
 - (b) Does the construction of the least squares line in panel 2 violate any of the Gauss-Markov assumptions?
 - (c) Does the identified point (identically located in both panels) have a substantively different interpretation?
5. Cigarette smoking among elderly males has decreased substantially whereas cigarette smoking among elderly females has remained fairly constant. Given the following data, calculate the a regression model for each gender and test for a differences of slopes.

	Proportion of Smokers, 65 and Over											
Year	1965	1974	1979	1983	1985	1990	1992	1993	1994	1995	1997	1998
Male	28.5	9.6	24.9	12.0	20.9	13.2	22.0	13.1	19.6	13.5	14.6	11.5
Female	16.6	12.4	13.5	10.5	13.2	11.1	14.9	11.5	12.8	11.5	10.4	11.2

Source: Centers for Disease Control and Prevention, National Center for Health Statistics. National Health Interview Survey, respective years.

6. Clogg, Petkova, and Haritou (1995) give detailed guidance for deciding between different linear regression models using the same data. In this work they define the matrices \mathbf{X} , which is $n \times (p + 1)$ rank $p + 1$, and \mathbf{Z} , which is $n \times (q + 1)$ rank $q + 1$, with $p < q$. They calculate the matrix $A = [\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}$. Find the dimension and rank of A .
7. Consider the General multiple-regression equation

$$Y = A + B_1X_1 + B_2X_2 + \dots + B_kX_k + E$$

An alternative procedure for calculating the least-squares coefficient B_1 is as follows:

- (a) Regress Y on X_2 through X_k , obtaining residuals $E_{Y|2\dots k}$.
- (b) Regress X_1 on X_2 through X_k , obtaining residuals $E_{1|2\dots k}$.
- (c) Regress the residuals on $E_{Y|2\dots k}$ on the residuals $E_{1|2\dots k}$. The slope of this simple regression is the multiple-regression slope for X_1 , that is, B_1 .

Suppose you had an outcome variable vector Y with n values, and an explanatory variable matrix X that has the k variables in columns plus a leading column of 1's. That is, using R:

```
> length(Y)
[1] 100
> dim(X)
[1] 100 5
```

for 100 data values and 4 explanatory variables. Write the R commands to implement the procedure above for these data. Test your code with artificially created data that you produce. Also, notice that the intercept for the simple regression in step 3 is 0. Why is this the case?