CH 5440 Multivariate Data Analysis

Assignment 1

Due Date: 26/01/20

1. (a) Let $x_1, x_2, \cdots, x_N$ and $y_1, y_2, \cdots, y_N$ be a set of N measurements of two variables x and y which are linearly related. We are interested in determining the linear regression parameter a where $y = ax + b$. Assume that the measurements of x and y contain errors, with standard deviations σ_δ and σ_ε, respectively. (a) If the ratio of the error variances $\alpha = \frac{\sigma_\epsilon^2}{\sigma_\delta^2}$ is known, derive the weighted TLS (WTLS) estimates of a and b in terms of $s_{xx}, s_{yy}, \bar{x}, \bar{y}, \alpha$. (b) How will the solution for α change if it is already known that the constant b is known to be 0?

   Note: The WTLS regression problem when the error variances are known is the solution of the following minimization problem. Multiply the objective function by σ_ε^2 and replace the ratio of the error variances by α. Differentiate the objective function with respect to the decision variables and solve resulting set of nonlinear algebraic equations for obtaining the parameters a and b.

   $$\min_{\alpha,\beta,\hat{x}_i} (y_i - a\hat{x}_i - b)^2/\sigma_\varepsilon^2 + (x_i - \hat{x}_i)^2/\sigma_\delta^2$$

   (b) From the solution of WTLS, obtain the solutions of the regression parameters for IOLS and OLS in the limit as $\alpha \to 0$ and $\alpha \to \infty$. Also obtain the solution for the estimates of x_i and y_i for each case (OLS, IOLS, WTLS) in terms of the regression parameters and measurements

2. The level of phytic acid in urine samples was determined by a catalytic fluorimetric (CF) method and the results were compared with those obtained using an established extraction photometric (EP) technique. The results, in mg/L, are the means of triplicate measurements, as shown in Table 2.
   (a) Is the new method (CF) a good substitute for the established method (EP) for measuring the level of phytic acid in urine? Justify your conclusion using linear regression between the two methods for different modelling assumptions regarding the accuracy of the respective measurement techniques.
   (b) Estimate the level of phytic acid in urine if EP measurement is 2.31 mg/l and CF measurement is 2.20 mg/l, for different modelling assumptions and provide 95% confidence intervals for these estimates, if possible.

3. Carbon-dioxide (CO2) is one of the major greenhouse gases that is implicated in the gradual warming of the earth's temperature. Measured concentrations of CO2 (in ppm) and atmospheric temperature (spatially and temporally averaged over a year) available from USEPA's Climate Change Indicators website (www.epa.gov/climate-indicators) between 1984 and 2014 is given in Table 1. The temperatures are deviation in deg F from the average

temperature in the period 1901-2000. Climate models recommend that the global temperature increase should be kept below 1.5 deg C by cutting down on CO2 emissions. Using linear regression estimate the maximum permissible level of CO2 in the atmosphere that can meet this goal. Your estimate should be conservative (which implies that among all estimates based on different assumptions you should use the least value). Note that this is a simplified analysis because other greenhouse gases such as methane, nitrous oxide, water vapour, etc. have not been considered. In order to improve your model you are encouraged to use other reliable data sources you can find (cite the sources from where you obtain additional data).

Table 2. Comparison of CF versus EP

EP	CF	EP	CF
1.98	1.87	0.13	0.14
2.31	2.20	3.15	3.20
3.29	3.15	2.72	2.70
3.56	3.42	2.31	2.43
1.23	1.10	1.92	1.78
1.57	1.41	1.56	1.53
2.05	1.84	0.94	0.84
0.66	0.68	2.27	2.21
0.31	0.27	3.17	3.10
2.82	2.80	2.36	2.34

Table 1. Measured average atmospheric CO2 concentration and temperature

Year	CO2	Temp (^{0}F)	Year	CO2	Temp (^{0}F)
1984	344.58	0.27	1999	368.33	0.792
1985	346.04	0.234	2000	369.52	0.756
1986	347.39	0.414	2001	371.13	0.972
1987	349.16	0.666	2002	373.22	1.08
1988	351.56	0.666	2003	375.77	1.098
1989	353.07	0.522	2004	377.49	1.026
1990	354.35	0.774	2005	379.8	1.17
1991	355.57	0.72	2006	381.9	1.098
1992	356.38	0.45	2007	383.76	1.098
1993	357.07	0.504	2008	385.59	0.972
1994	358.82	0.612	2009	387.37	1.134
1995	360.8	0.81	2010	389.85	1.26
1996	362.59	0.576	2011	391.63	1.026
1997	363.71	0.918	2012	393.82	1.116
1998	366.65	1.134	2013	396.48	1.188
			2014	398.61	1.332