STAT 626, Summer 2019: Homeworks

NOTES:
(i) The bonus problems tend to get deeper into some statistical concepts, they will encourage you to review/study the relevant sections of this or your earlier textbook/notes more systematically.
(ii) To get the bonus points for a problem, a complete solution must be given. No partial credits for the Bonus problems.

1. W#6 Due Wed July 17, 10;00 AM CST,

   Do Problems 4.1, 4.3, 4.4 , 4.10 (bonus), 5.2, 5.3, 5.6, 5.7 and 5.14 (bonus) from the textbook.

HW#5 : Due Wed. July 3, 10:00 AM CST

I. Let $\{w_{t1}\}, \{w_{t2}\}, \{w_t\}$ be three independent WN$(0, 1)$ series and define

$$y_{t1} = x_{t1} + \sum_{j=1}^{t} w_j, \quad y_{t2} = x_{t2} + 5\sum_{j=1}^{t} w_j,$$

where

$$x_{t1} = 0.5x_{t-1,1} + w_{t1}, \quad x_{t2} = 0.9x_{t-1,2} + w_{t2},$$

are two AR(1) time series.
(a) Simulate and plot $n = 100$ values of the three time series $\{y_{t1}\}, \{y_{t2}\}$ and $z_t = 5y_{t1} - y_{t2}$. Do they appear to be stationary?
(b) Compute the autocovariance function of $\{y_{t1}\}$. Is $\{y_{t1}\}$ stationary?
(c) Compute the autocovariance function of the time series $z_t = 5y_{t1} - y_{t2}$. Is it stationary?
(d) Compare and explain your findings in parts (a)-(c) regarding (non)stationarity of the three time series involved.
(e) Compute the cross-covariance function and cross-correlation function (CCF) between $\{y_{t1}\}$ and $\{y_{t2}\}$, see Definitions 2.22-2.23, p.25. Are the two time series $\{y_{t1}\}$ and $\{y_{t2}\}$ jointly stationary?

II. Do Problems 3.5-3.8, and 4.2 from the textbook.

III. Compute the mean and autocovariance functions of

$$x_t = \sum_{j=0}^{\infty} \phi^j w_{t-j} - \frac{\phi}{1-\phi^2} w_{t+1}, \quad |\phi| < 1,$$

and decide if it is stationary. Is the process causal?

IV. (Bonus) Let x_t be the standard AR(p) model with coefficients $\phi_1, \ldots, \phi_p$. Show that

$$\nabla x_t = \gamma x_{t-1} + \sum_{j=1}^{p-1} \psi_j \nabla x_{t-j} + w_j,$$

where $\gamma = \sum_{j=1}^{p} \phi_j - 1$ and $\psi_j = -\sum_{i=j}^{p} \phi_i$ for $j = 2, \ldots, p$.
Hint: Prove the statement for $p = 1, 2$ first, and then the general case.

HW# 4: Due Wed. June 19, 10:00 am CST.

I. Do Problems 2.8, 2.11-2.15, 3.2 and 3.3 from the textbook.

II. (a) Simulate two random walks $y_t, x_t, t = 1, \ldots, 100$, with initial values $x_0 = y_0 = 0$, using two independent $N(0, 1)$ white noises.

(i) Plot y_t vs x_t in the (x, y)-plane. Describe the pattern of dependence (if any) you notice in the scatterplot.

(ii) Consider the linear regression model $y_t = \beta_0 + \beta_1 x_t + w_t$, and the null hypothesis $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ at $\alpha = 0.05$. Do you expect H_0 would be rejected? Why?

(iii) Perform the test and state your conclusion.

(b) Repeat the above experiment 1000 times and count the number of times H_0 is rejected in (iii). Does it support your expectation in (ii)? If not, find an explanation for this phenomenon based on possible violations of assumptions of inference in regression models. (Hint: Recall the three assumptions of inference for linear regression models: (a) Independence, (b) Homogeniety of variances, (c) Normality. Check whether they hold for the data here.)

III. Multiple choice questions: Circle the correct answer(s).

1. Suppose a reasonable model for a time series x_t is $\beta_0 + \beta_1 t$ plus a white noise. Then, the time series is :
(a) uncorrelated,
(b) independent,
(c) normally distributed,
(d) nonstationary,
(e) random walk.

Regression Problems: For the next few problems, we consider a dataset coming from a study entitled ”Getting What You Pay For: The Debate Over Equity in Public School Expenditure.” The names of the variables measured along with their summaries are given next. The investigators fit a multiple regression model with the total SAT scores as the response, and expend, salary, ratio, and takers as the predictors and obtain the output (SATREG) reported in the next page.

```
  expend          ratio           salary          takers     
 Min.   :3.656   Min.   :13.80   Min.   :25.99   Min.   : 4.00  
 1st Qu.:4.882   1st Qu.:15.22   1st Qu.:30.98   1st Qu.: 9.00  
 Median :5.768   Median :16.60   Median :33.29   Median :28.00  
 Mean   :5.905   Mean   :16.86   Mean   :34.83   Mean   :35.24  
 3rd Qu.:6.434   3rd Qu.:17.57   3rd Qu.:38.55   3rd Qu.:63.00  
 Max.   :9.774   Max.   :24.30   Max.   :50.05   Max.   :81.00  
     verbal           math           total       
 Min.   :401.0   Min.   :443.0   Min.   : 844.0  
 1st Qu.:427.2   1st Qu.:474.8   1st Qu.: 897.2  
 Median :448.0   Median :497.5   Median : 945.5  
 Mean   :457.1   Mean   :508.8   Mean   : 965.9  
 3rd Qu.:490.2   3rd Qu.:539.5   3rd Qu.:1032.0  
 Max.   :516.0   Max.   :592.0   Max.   :1107.0  
```

```
lm(formula~ total ~ expend + salary + ratio + takers, data)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1045.9715    52.8698  19.784  < 2e-16 ***
expend         4.4626    10.5465   0.423    0.674    
salary         1.6379     2.3872   0.686    0.496    
ratio         -3.6242     3.2154  -1.127    0.266    
takers        -2.9045     0.2313 -12.559 2.61e-16 ***
---

Residual standard error: 32.7 on 45 degrees of freedom
Multiple R-squared:  0.8246,    Adjusted R-squared:  0.809 
F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16
```

```
  ## Prediction using the fitted regression model
> x0=data.frame(expend =115, salary =38.55, ratio =16.60, takers =28.00)
>
> predict(SATREG,x0, interval="confidence")
       fit       lwr      upr
  1480.824 -818.8357 3780.483
>
 predict(SATREG,x0, interval="prediction")
       fit       lwr      upr
  1480.824 -819.7787 3781.426
```

2. To three decimal place accuracy the percentage of the variation in the response explained by these four predictors is

   a. 0.880,

   b. 0.825,

   c. 0.809,

   d. 0.327,

   e. none of the above.

3. To two decimal place accuracy the estimated slope for salary and its standard error are:

   a. -1.13 and 2.37,

   b. 0.69 and 0.49,

   c. 1.83 and 8.40,

   d. 1.64 and 2.39,

   e. 0.83 and 0.04.

4. The estimated slope in the previous question should be interpreted as:

   a. an unbiased estimate of the SAT total when the salary is fixed at 1 unit,

   b. an extrapolation of the SAT total when $x = 1$.

   c. the expected change in SAT total when the salary is increased by 1 unit.

   d. the expected change in SAT total when the salary is increased by 1 unit and other variables are fixed.

   e. the change in SAT total when the salary is increased by 1 unit and other variables are fixed

5. The sample size n for the data set analyzed is:

a. 25,

b. 50,

c. 45,

d. $n - 1$,

e. 15.

6. The estimated value of σ (under the the standard assumptions for linear regression models) from the above output is

a. 512.7,

b. 5.24,

c. 32.7,

d. 13,

e. 0.524.

7. The Department of Education is interested in predicting the SAT total when expend =115, salary =38.55, ratio =16.60, takers =28.00. To three decimal point accuracy (using the output above), a 95 percent prediction interval for the SAT total is: :

a. (-819.779, 3781.426),

b. not possible to compute because it represents an extrapolation of the data beyond the bulk of the data.

c. 1480.824,

d. (-818.836, 3780.983).

e. average of the intervals in a. and d.

8. If you computed a 95 percent confidence interval for the average SAT total for expend =115, salary =38.55, ratio =16.60, takers =28.00, then:

a. it will be narrower than the 95 percent prediction interval,

b. it will be wider than the 95 percent prediction interval,

c. the computed confidence interval has probability 0.95 of capturing the SAT total,

d. it will be the same width as the 95 percent prediction interval,

e. none of the above.

9. Data were collected on y, income and x, education of several college graduates. Both quadratic (Model 1) and linear (Model 2) models were fit to the data. The R^2 for Model 2 is

a. the same as that for Model 1,

b. more than that for Model 1,

c. equal to that for Model 1 plus 1,

d. not comparable to that for Model 1,

e. less than that for Model 1.

10. For the least squares estimator of the slope in the model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ (under the standard assumptions) to be more **accurate** it is advisable to have

a. half of the x_i's to be equal, and others be as dispersed as possible,

b. $\sum x_i$ to be large,

c. $\sum x_i^2$ to be large,

d. x_i's be scattered as much as possible,

e. none of the above.

HW#3: Due Wed. June 12, 10:00 am CST,

A. Do Problems 2.1-2.7, 2.9-2.10 from the textbook.

B. (Bonus) Let $w_t \sim i.i.d.\ N(0, \sigma_w^2)$, compute the mean and autocovariance functions of

$$x_t = w_{t-1} w_{t-2} (w_{t-1} + w_t + t),$$

and decide if $\{x_t\}$ is stationary.

HW#2: Due Wed., May 29, 10:00 am CST,

I. Do Problems 1.1, 1.3, and 1.4 from the textbook.

II. In the simple regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, one usually deals with pairs of observations $(x_i, y_i), i = 1, \cdots, n$. Show that

1. $\sum_{i=1}^n (x_i - \bar{x}) = 0$, and $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x}) y_i$.

2. $S_x^{-2} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n c_i y_i$,

where $c_i = \frac{x_i - \bar{x}}{S_x^2}$, $S_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2$.

3. (Bonus) For the simple linear regression model, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$,

(a) Derive the equations for minimizing the residual sum of squares (RSS):

$$Q(\beta_0, \beta_1) = \sum (y_i - \beta_0 - \beta_1 x_i)^2,$$

and give the formulae for the least squares estimates (LSE) $\hat{\beta}_0, \hat{\beta}_1$ of the regression coefficients.

(b) Show that $\hat{\beta}_1 = \sum_{i=1}^n c_i y_i$, with c_i's as above. If $\varepsilon_i \sim N(0, \sigma^2)$, what is the distribution of $\hat{\beta}_1$?

(c) Let $e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i,\ i = 1, 2, \ldots, n$, be the (estimated regression) residuals. Show that

$$\sum_{i=1}^n e_i = 0, \text{ and } \sum_{i=1}^n e_i x_i = 0.$$

What is the (geometrical) interpretation of the above identities?