STA302/1001 Practice Exam

- Please use $\alpha = 0.05$ for all statistical tests, and keep at least three decimal digits in all your numerical calculations.

1. Use not more than 2 sentences to answer each of the following questions.
   (a) Once an outlier is detected, we should always remove it from the data and re-do the analysis. Comment.
   (b) A data point cannot simultaneously be an outlier and has high leverage value. True or false? (No need to give reasons.)
   (c) If Box-Cox transformation suggests us to transform the response variable Y to $Y^{1.97}$, we may want to consider the transformation Y^2 instead. Why?
   (d) Suppose we want to use multiple linear regression to predict the probability of heart attack for a patient. Do you see any major problem(s) with this?
   (e) Is ordinary least squares a special case of weighted least squares? Why?
   (f) When fitting simple linear regression, why is it necessary to look at diagnostic plots even when R^2 is large?

2. Each of the four diagnostic plots in Figure 1 indicates different potential problems commonly encountered in regression modeling. For each plot state **all** of the potential problems and suggest remedial methods.

3. We want to fit a simple linear regression model $y = \beta_0 + \beta_1 x$ to the following data set. We know $SXX = 4168.82$ and $SXY = 15924.51$.

x	16	19	27	41	42	43	44	45	46	78	80
y	68.4	87.2	110.0	168.0	173.0	170.0	178.0	180.0	192.0	307.0	316.0

   (a) Obtained the OLS estimates for β_0 and β_1.
   (b) Estimate $E(Y|X = 60)$. Attach a 95% confidence interval to your estimate.
   (c) Construct the corresponding ANOVA table and test for the significance of the simple linear regression model.

4. A common statistical tool for speech analysis is sine-cosine regression, from which amplitudes and frequencies of each harmonic can be extracted. Let Y_t, $t = 1, \ldots, n$ be a speech signal, where t denotes time (Y is the response and t is the predictor). A simple regression model is

$$Y_t = a_0 + \sum_{k=1}^{K} a_k \sin(w_k t) + \sum_{k=1}^{K} b_k \cos(w_k t) + e_t, \tag{1}$$

   where $K, a_0, \ldots, a_K, b_0, \ldots, b_K, w_1, \ldots, w_k$ are unknown model parameters and $e_1, \ldots, e_t$ are iid errors. Here K is the number of harmonics, a_k's and b_k's are their amplitudes, while w_k's are the frequencies. Note that, if c is a constant, $\frac{d\sin(cx)}{dx} = c\cos(cx)$ and $\frac{d\cos(cx)}{dx} = -c\sin(cx)$.

   (a) The parameter K in (1) cannot be estimated using OLS. Why? Can you suggest some method for estimating its value?
   (b) Consider the following simpler version of (1):

$$Y_t = 3 + b\cos(4t) + e_t.$$

   Derive the OLS estimator for b.

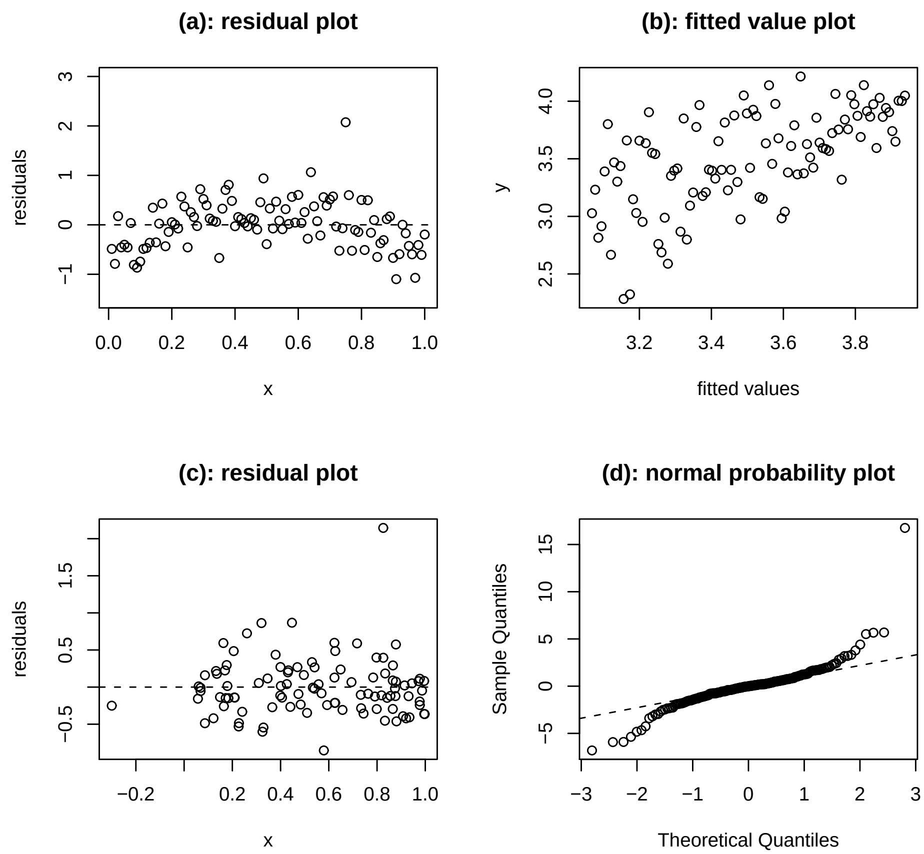

Figure 1: Some diagnostic plots.

(c) Consider yet another simpler version:

$$Y_t = 3 + 2\cos(wt) + e_t.$$

Show that the OLS estimator for w does not have a closed-form. Describe how you would estimate its value in practice.

5. This question is about using the mean-shift model for testing outliers. Suppose we have a data set $(x_1, y_1), \ldots, (x_n, y_n)$ of size n that is well modeled by simple linear regression. Before conducting any statistical analysis, we have reasons to suspect that y_4 and y_7 are outliers. In fact, we have strong reasons to suspect that the mean shifts (i.e., the magnitudes of the outlier) for y_4 and y_7 are δ and 2δ respectively. Describe how you would estimate δ and test if $\delta = 0$. In answering this question you should describe the model that you need to fit (define new variable(s) if necessary), state H_0, H_1, the test statistics, the degrees of freedom involved and so on.

6. Suppose Y is a response variable and X_1 and X_2 are predictor variables. The values of Y, X_1 and X_2 are collected for $n = 100$ subjects and the following regression model was fitted:

$$E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2. \qquad (2)$$

Some R outputs and diagnostic plots are given below.

```
Call:
lm(formula = y ~ x1 + x2)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.06908 -0.28984  0.01082  0.29508  1.07586 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.5393     0.1014  15.187  < 2e-16 ***
x1            0.7445     0.1772   4.202 5.89e-05 ***
x2            0.7962     0.0465  17.121  < 2e-16 ***
---
Residual standard error: 0.4766 on 97 degrees of freedom
Multiple R-squared: 0.8033,     Adjusted R-squared: 0.7992 
F-statistic: 198.1 on 2 and 97 DF,  p-value: < 2.2e-16
```

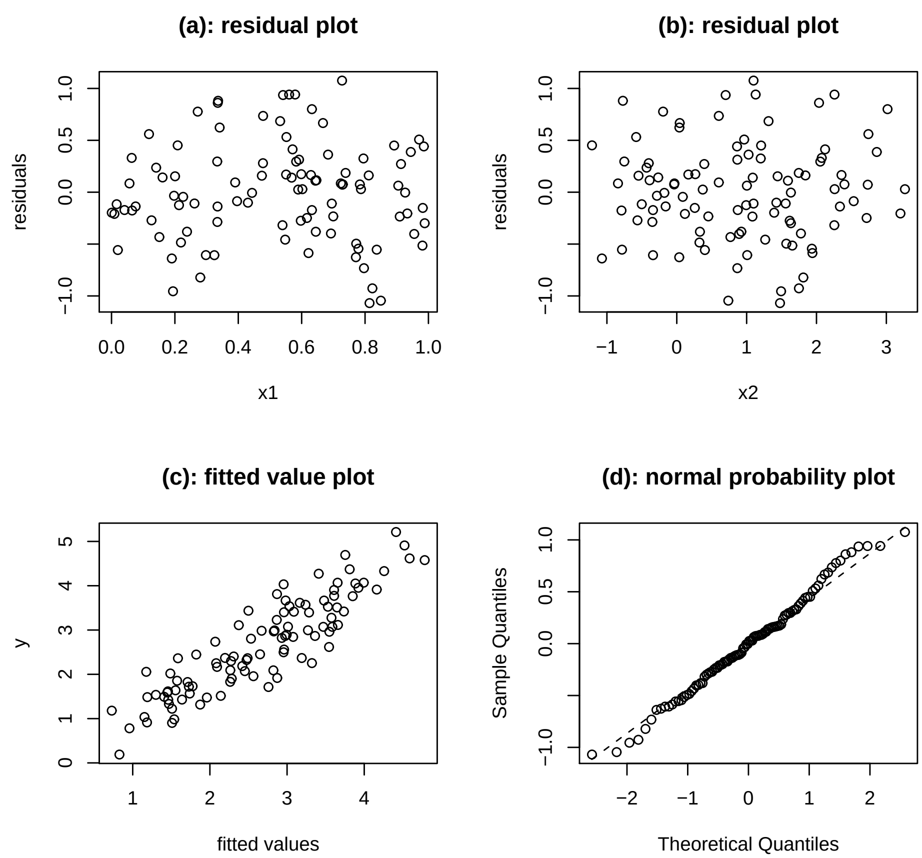

Figure 2: Some diagnostic plots.

Denote the model matrix as $\mathbf{X}$, and $(\mathbf{X}'\mathbf{X})^{-1}$ is

```
                         x1           x2
    0.045229535 -0.06215878 -0.002723652
x1 -0.062158778  0.13820855 -0.011494501
x2 -0.002723652 -0.01149450  0.009521381
```

(a) Let $\theta = 2\beta_1 - \beta_0$ and hence $\hat{\theta} = 2\hat{\beta}_1 - \hat{\beta}_0$. First approximate $\mathrm{Var}(\hat{\theta})$ and then test if $\theta = 0$ against $\theta \neq 0$.

(b) Estimate the variance of $\hat{\beta}_1/\hat{\beta}_2$.

(c) Briefly comment on the diagnostic plots Figures 2(b) to 2(d).

(d) Residuals from Figure 2(a) seem to form a curvature. Can we use the lack-of-fit test to confirm this? Why?

(e) In order to test the existence of the curvature, a second model was fitted and the *R* outputs are given below. Conduct the corresponding test.

```
Call:
lm(formula = y ~ x1 + x2 + I(x1^2))

Residuals:
     Min       1Q   Median       3Q      Max
-1.04242 -0.31034  0.02720  0.25530  1.03528

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.3072     0.1426   9.166 9.19e-15 ***
x1            2.1412     0.6401   3.345  0.00117 **
x2            0.7912     0.0456  17.353  < 2e-16 ***
I(x1^2)      -1.3944     0.6151  -2.267  0.02564 *
---

Residual standard error: 0.4667 on 96 degrees of freedom
Multiple R-squared: 0.8133,     Adjusted R-squared: 0.8075
F-statistic: 139.4 on 3 and 96 DF,  p-value: < 2.2e-16
```