NTHU STAT 5410 Final Examination Jan. 12, 2023

Instructions: Attempt all questions. Short and specific answers are preferred. Given explanation when required, but keep it as short and simple as possible. Give only one answer to each question – if you give alternative answers, the worst answer will be graded.

Question A.

The question concerns data collected on 38 countries in 1993. The variables are:

`life` – Life expectancy in years,
`tv` – Number of people per television set,
`doctor` – Number of people per doctor.

Here is the numerical summary information of these variables:

```
      life                  tv                       doctor
Min.     :51.50    Min.      :  1.30      Min.      :  226.0
1st Qu.  :64.12    1st Qu.   :  3.35      1st Qu.   :  456.8
Median   :70.00    Median    :  6.30      Median    :  824.5
Mean     :67.76    Mean      : 51.98      Mean      : 2933.8
3rd Qu.  :74.12    3rd Qu.   : 23.00      3rd Qu.   : 2861.8
Max.     :79.00    Max.      :592.00      Max.      :36660.0
```

(1) (1 pt) Which variables, if any, have skewed distributions?

A regression model with `life` as the response and `tv` and `doctor` as predictors, called **`Model_1`**, was fit and the following output obtained:

```
Coefficients:
             Estimate     Std. Error   t value   Pr(>|t|)
(Intercept)  70.2519573   1.0877047    64.587    <2e-16
tv           -0.0234954   0.0096469    -2.436    0.0201
doctor       -0.0004320   0.0002023    -2.136    0.0398
---
Residual standard error: 6.003 on 35 degrees of freedom
Multiple R-squared:  0.44,    Adjusted R-squared:  0.408
F-statistic: 13.75 on 2 and 35 DF,  p-value: 3.916e-05
```

(2) (1 pt) The country Ethiopia had the largest (in absolute value) jackknife residual. Its observed value was `4.31` making it an outlier. Does this imply that Ethiopia is also an influential point? Explain.

A plot of the residuals and fitted values is shown in Figure 1. A Q-Q normal plot of the residuals is shown in Figure 2.

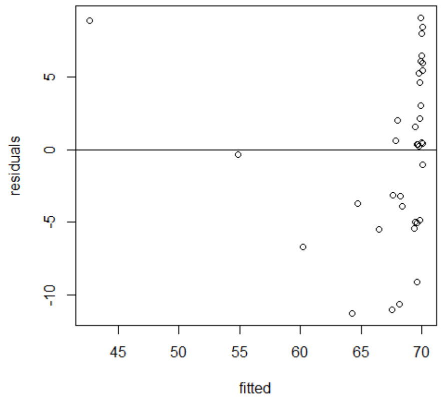

Figure 1

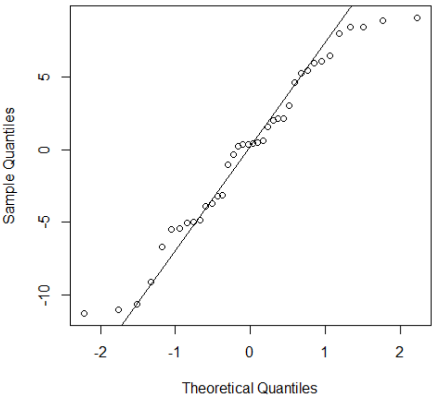

Figure 2

(3) (1 pt) What is the most important conclusion that can be drawn from Figure 1?

(4) (1 pt) The slope of the line in Figure 2 could be used as a rough estimate of which *parameter* in the regression model?

Both predictors were logged and the model refit, called **`Model_2`**, resulted in the following output:

```
Coefficients:
              Estimate  Std. Error   t value   Pr(>|t|)
(Intercept)  90.6222   4.3557       20.806    < 2e-16
log(tv)      -2.9156   0.5907       -4.936    1.95e-05
log(doctor)  -2.2589   0.7474       -3.022    0.00467
---
Residual standard error: 3.704 on 35 degrees of freedom
Multiple R-squared:  0.7868,    Adjusted R-squared:  0.7747
F-statistic:  64.6 on 2 and 35 DF,  p-value: 1.788e-12
```

(5) (1 pt) What previous finding(s) supports using logged predictors? Explain why.

(6) (1 pt) Is `Model_2` better than `Model_1`? Explain.

A plot of the residuals and fitted values under `Model_2` is shown below in Figure 3.

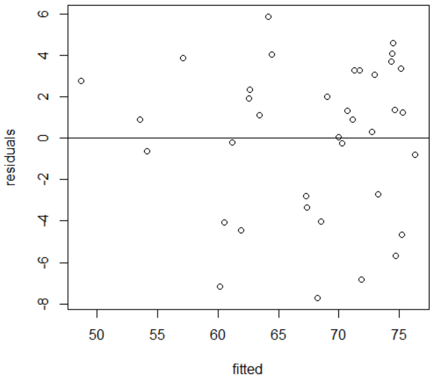

Figure 3

(7) (1 pts) Does the unwanted pattern you observed in Figure 1 (i.e., problem (3)) disappear in Figure 3? If yes, explain why it can happen.

(8) (1 pt) Suppose that we used number of TVs per person rather than number of persons per TV in constructing the predictor `log(tv)` in the `Model_2`. What can we say about the regression coefficient for this transformed predictor?

(9) (1 pt) Two countries A and B are otherwise identical except that A has twice as many TVs as B. What does the `Model_2` predict about the life expectancy in A compared to B?

Question B.

A study investigated whether babies take longer to learn to crawl in cold months when they are often bundled in clothes that restrict their movement, than in warmer months. The study sought an association between babies' first crawling age and the average temperature during the month they first try to crawl (about 6 months after birth). Parents brought their babies into the University of Denver Infant Study Center between 1988-1991 for the study. The parents reported the birth month and age at which each of their children was first able to creep or crawl a distance of four feet in one minute. Data were collected on 215 boys and 216 girls (40 pairs of which were twins). Some data were missing and the cases with missing values were deleted. The rest data are shown below:

```
   BirthMonth   CrawlingAge   SD    n     temperature
1  January      29.84         7.08  32    66
2  February     30.52         6.96  36    73
3  March        29.70         8.33  23    72
4  April        31.84         6.21  26    63
5  May          28.58         8.07  27    52
6  June         31.44         8.10  29    39
7  July         33.64         6.91  21    33
8  August       32.82         7.61  45    30
9  September    33.83         6.93  38    33
10 October      33.35         7.29  44    37
11 November     33.38         7.42  49    48
12 December     32.32         5.71  44    57
```

The average crawling age (`CrawlingAge`) is in weeks. The `temperature` (°F) is the average monthly temperature six months after birth month. Standard deviation (`SD`) for the crawling ages and sample sizes (`n`) for each month are given.

(10) (1 pt) What assumptions on the missing mechanism can guarantee that the deletion of the cases with missing values would not bias our conclusions?

An unweighted regression was fit with the average crawling age in months as the response and the sixth month temperature as the predictor. Here is the summary:

```
Coefficients:
              Estimate      Std. Error    t value    Pr(>|t|)
(Intercept)   35.6781       1.3175        27.1       1.1e-10
temperature    -0.0777      0.0251        -3.1       0.011
```

(11) (1 pt) The sixth month temperature is used to represent the ambient temperature at the time the baby first tries to crawl. But since the time and temperature will not be exactly correct for any given baby, there is some inaccuracy. Suppose we had been able to record the correct information for each individual baby. Explain how you would expect the fitted slope of the regression to change --- more negative, less negative, or about the same? Explain.

(12) (2 pts) We have information about non-constant variance in the response so a weighted regression should be used. Which choice of weights would be appropriate? Explain.

(13) (2 pts) In each birth month, the distribution of first crawling age is skewed because a few babies take much longer to learn to crawl. Hence this is not a normal distribution. Explain why this would likely not result in a serious violation of the usual assumptions on the error distribution.

(14) (2 pts) No correlation is the third assumption on the error distribution. What feature of this data would lead us to question this assumption? Explain.

(15) (2 pts) Suppose girls crawl earlier on average than boys, but there is no association between gender and birth month. What advantage concerning the temperature effect, if any, would be gained by including gender in the model? Explain.

Question C.

Fortune magazine publishes a list of the world’s billionaires each year. The 1992 list includes 225 individuals. Their wealth, age, and geographic location (Asia, Europe, Middle East, United States, and Other) are reported. A numerical summary is given below where the location are coded with the first letter of the name of the region:

```
         wealth              age           region
Min.     : 1.000   Min.     :  7.00   A   :37
1st Qu.  : 1.300   1st Qu.  : 56.00   E   :76
Median   : 1.800   Median   : 65.00   M   :22
Mean     : 2.726   Mean     : 64.03   O   :28
3rd Qu.  : 3.000   3rd Qu.  : 72.00   U   :62
Max.     :37.000   Max.     :102.00
```

The histogram of `wealth` is shown below in Figure 4. The model, called **`Model_3`**,

`wealth ~ age+region+age:region`

was fit. A plot of the residuals and fitted values under `Model_3` is shown below in Figure 5.

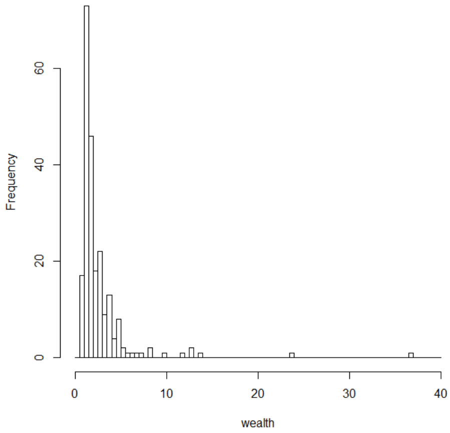

Figure 4

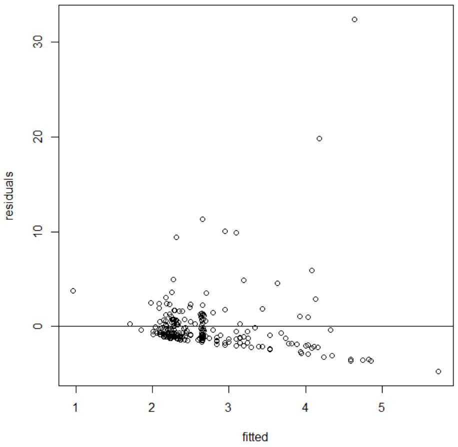

Figure 5

(16) (2 pts) Suppose that a robust estimator is used to estimate the coefficients under `Model_3`. Would you expect their estimates to be very different from the ordinary least square estimates? Explain.

(17) (2 pts) A transformation of the response seemed necessary. An analysis of Box-Cox regression showed that the inverse transformation of the response (i.e., `wealth`$^{-1}$) would be optimal for the data. What outputs given here also support the use of inverse transformation? Explain.

The model, called **`Model_4`**,

wealth^{-1} ~ age+region+age:region

was refit. The summary of this model is shown below.

```
Coefficients:
              Estimate     Std. Error   t value    Pr(>|t|)
(Intercept)   0.4104503    0.2767447    1.483      0.1395
age           0.0015257    0.0042971    0.355      0.7229
regionE       0.1875133    0.3068232    0.611      0.5417
regionM       0.6348091    0.3369673    1.884      0.0609 .
regionO       0.0647260    0.3710286    0.174      0.8617
regionU       0.0551576    0.3304029    0.167      0.8676
age:regionE   -0.0017634   0.0047472    -0.371     0.7107
age:regionM   -0.0092215   0.0051670    -1.785     0.0757
age:regionO   -0.0002263   0.0057217    -0.040     0.9685
age:regionU   -0.0002221   0.0051112    -0.043     0.9654
---
Residual standard error: 0.2568 on 215 degrees of freedom
Multiple R-squared: 0.04388,    Adjusted R-squared: 0.003853
F-statistic: 1.096 on 9 and 215 DF,  p-value: 0.3667
```

(18) (1 pt) The adjusted R-squared is very small. Could it be negative for some other datasets? Explain.

(19) (1 pt) What is the fitted regression line for the Middle East region under Model_4?

(20) (1 pt) A half-normal plot for the Cook's distances are shown in Figure 6. Since the points follow a curve, the Cook's distance do not follow a half-normal distribution. What is the practical consequence of this fact?

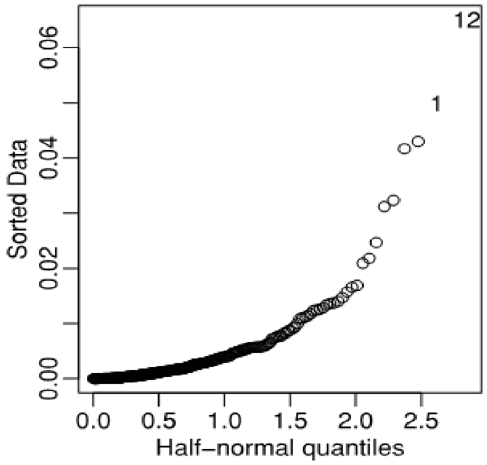

Figure 6

The model, called **Model_5**,

$\texttt{wealth}^{-1}$ ~ age+regionM+age:regionM

was fit, where the regionM is identical to the regionM effect used in Model_4. A summary of the *sequential* ANOVA under Model_5 is given below:

```
Analysis of Variance Table
              Df     Sum Sq     Mean Sq    F value    Pr(>F)
age           1      0.0434     0.04340    0.6690     0.41430
regionM       1      0.0004     0.00036    0.0056     0.94065
age:regionM   1      0.4424     0.44237    6.8180     0.00964
Residuals     221    14.3391    0.06488
```

and a regression summary for Model_5 is given below:

```
Coefficients:
              Estimate      Std. Error    t value    Pr(>|t|)
(Intercept)   0.5177302     0.0915148     5.657      4.73e-08
age           0.0005895     0.0014022     0.420      0.67456
regionM       0.5275292     0.2115365     2.494      0.01337
age:regionM   -0.0082853    0.0031731     -2.611     0.00964
```

(21) (2 pts) Obtain the *F*-statistic for testing H_0: Model_5 versus H_1: Model_4.

(22) (1 pt) Explain why the two tests for regionM (the one in sequential ANOVA and the one in regression summary) under Model_5 have different *p*-values.

(23) (2 pts) The Model_5 can be further simplified. Based on the sequential ANOVA and the regression outputs, which simplified model of Model_5 would you suggest, i.e.,

wealth^{-1} ~ age:regionM (based on sequential ANOVA)

or

wealth^{-1} ~ regionM+age:regionM (based on regression outputs)?

Explain why.

(24) (1 pt) For the simplified model chosen in the previous problem, what does its intercept parameter represent (i.e., how to interpret its intercept parameter)?

Question D.

Moore (1975) reported the results of an experiment to construct a model for total oxygen demand in dairy wastes as a function of five laboratory measurements. Data were collected on samples kept in suspension in water in a laboratory for 220 days. Although all observations reported were taken on the same sample over time, assume that they are independent. The measured variables are:

y – log (oxygen demand, mg oxygen per minute),
x1 – biological oxygen demand, mg/liter,
x2 – total Kjeldahl nitrogen, mg/liter,
x3 – total solids, mg/liter,
x4 – total volatile solids, a component of x3, mg/liter,
x5 – chemical oxygen demand, mg/liter.

A model with y as the response and x1, …, x5 as the predictors, called **Model_6**, was fit to the data, and the following output was obtained:

```
Coefficients:
             Estimate     Std. Error   t value   Pr(>|t|)
(Intercept)  -2.16e+00    9.13e-01     -2.36     0.033
x1           -9.01e-06    5.18e-04     -0.02     0.986
x2           1.32e-03     1.26e-03     1.04      0.315
x3           1.28e-04     7.69e-05     1.66      0.119
x4           7.90e-03     1.40e-02     0.56      0.582
x5           1.42e-04     7.38e-05     1.92      0.075
```

The variance inflation factors (VIFs) of the five predictors are shown below:

```
x1        x2        x3        x4        x5
7.1348    1.2984    4.4552    2.4377    4.3662
```

(25) (1 pt) The predictor x1 has the most serious collinearity in this data. Suppose that we plan to conduct the same experiment again in the future to collect a new data in which x1 is almost orthogonal to the other predictors and has a similar sample variance as the x1 in the old data. What value would you expect to be the standard error of the coefficient estimate of x1 in the new data?

A principal component analysis was performed on the *covariance matrix* of the five predictors and the following output was obtained:

```
$values
[1] 5.04e+06   6.51e+05   1.79e+04   2.17e+03   1.84e+01

$vectors
   PC1       PC2       PC3       PC4        PC5
x1 0.1250    0.0112    0.9684    0.2156    -0.0034
x2 -0.0014   -0.0045   -0.2171   0.9761    -0.0038
x3 0.6868    -0.7226   -0.0761   -0.0192   -0.0021
x4 0.0022    -0.0012   0.0023    0.0044    1.0000
x5 0.7161    0.6911    -0.0966   -0.0173   -0.0005
```

A *sequential* ANOVA for the model y ~ PC5 + PC4 + PC3 + PC2 + PC1, called **Model_7**, was obtained:

```
Analysis of Variance Table
          Df     Sum Sq    Mean Sq   F value   Pr(>F)
PC5       1      0.02      0.02      0.32      0.58
PC4       1      0.07      0.07      1.04      0.33
PC3       1      0.03      0.03      0.45      0.52
PC2       1      0.00      0.00      0.02      0.90
PC1       1      3.98      3.98      58.13     2.4e-06
Residuals 14     0.96      0.07
```

(26) (1 pt) Justify the use of the covariance matrix of the five predictors in constructing the principal components, i.e., explain why using covariance matrix is a reasonable approach for this data?

(27) (2 pts) Find the sum of the sample variances of x1, …, x5. Which of x1, …, x5 would you expect to have the smallest sample variance? Explain.

(28) (1 pt) The R^2 of Model_6: y ~ x1 + x2 + x3 + x4 + x5 is 0.811. Explain how you can use the analysis results given above to obtain (an approximation of) this value.

(29) (2 pts) A model selection based on the adjusted-R^2 criterion was applied on Model_7 and the best sub-model is a model with two principal components. Identify this best sub-model and explain why it is the best one under adjusted-R^2 criterion based on the information available.

(30) (1 pt) Let z=x3+x5 (i.e., a new predictor which is the sum of the two predictors x3 and x5). Would you expect y ~ z to be a good fitted model? Explain.