Midterm I

NAME: _____

Instructions:

1. For hypothesis testing, follow the standard procedures. That is, state the H_0 and the H_A, select and compute the test statistic, state the null distribution, report the p-value, draw a conclusion at the $\alpha = 0.05$ level, and interpret the testing result.
2. This exam is open book. You may use textbooks, notebooks, and a calculator.
3. Do all your work in the spaces provided. If you need additional space, use the back of the preceding page, indicating *clearly* that you have done so.
4. To get full credit, you must show your work. Partial credit will be awarded.
5. Do not dwell too long on any one question. Answer as many questions as you can.
6. Note that some questions have multiple parts. For some questions, these parts are independent, and so, for example, you can work on parts (b) or (c) separately from part (a).

For grader's use:

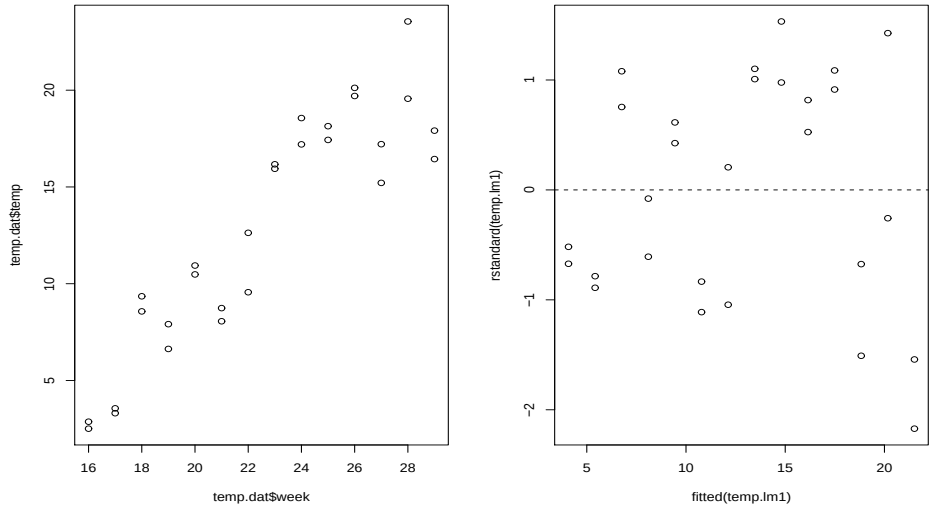
Question	Possible Points	Score
1	35	
2	35	
3	30	
Total	100	

1. In a study of air temperature pattern over time, a researcher obtained the weekly average temperature in two different study sites from the 16^{th} week to the 29^{th} week of a given year. The data set (`temp.dat`) consists of three variables: weekly average temperature in ^{o}C (`temp`), week of the year (`week`), and the square of the week (`weeksq`). The following (edited) R code and output may be used to answer Questions 1(a)–(d).

```
> temp.dat;
    temp week weeksq
1   2.87   16    256
2   3.31   17    289
3   8.57   18    324
4   7.91   19    361
5  10.48   20    400
6   8.74   21    441
7  12.63   22    484
8  15.94   23    529
9  18.56   24    576
10 18.14   25    625
11 20.12   26    676
12 15.21   27    729
13 19.56   28    784
14 16.44   29    841
15  2.51   16    256
16  3.56   17    289
17  9.35   18    324
18  6.63   19    361
19 10.94   20    400
20  8.06   21    441
21  9.56   22    484
22 16.17   23    529
23 17.20   24    576
24 17.43   25    625
25 19.70   26    676
26 17.21   27    729
27 23.55   28    784
28 17.91   29    841

> temp.lm1 = lm(temp~week, data=temp.dat);
> summary(temp.lm1);
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.3763     2.6803  -6.483 7.17e-07 ***
week          1.3409     0.1173  11.436 1.21e-11 ***
---

Residual standard error: 2.501 on 26 degrees of freedom
Multiple R-Squared: 0.8342,Adjusted R-squared: 0.8278
F-statistic: 130.8 on 1 and 26 DF,  p-value: 1.209e-11
```

```
> anova(temp.lm1);
Analysis of Variance Table

Response: temp
          Df Sum Sq Mean Sq F value    Pr(>F)
week       1 818.15  818.15  130.78 1.209e-11 ***
Residuals 26 162.65    6.26
---

> predict(temp.lm1, data.frame(week=18), se.fit=T);
$fit
[1] 6.760747

$se.fit
[1] 0.7084004

$df
[1] 26

> temp.aov = lm(temp~factor(week), data=temp.dat);
> anova(temp.lm1, temp.aov);
Analysis of Variance Table

Model 1: temp ~ week
Model 2: temp ~ factor(week)
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     26 162.648
2     14  18.601 12   144.047 9.0348 0.0001198 ***
---

> temp.lm2 = lm(temp~week+weeksq, data=temp.dat);
> summary(temp.lm2);
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -55.90319   14.65120  -3.816 0.000794 ***
week          4.87913    1.33135   3.665 0.001165 **
weeksq       -0.07863    0.02949  -2.666 0.013260 *
---

Residual standard error: 2.251 on 25 degrees of freedom
Multiple R-Squared: 0.8709,Adjusted R-squared: 0.8605
F-statistic: 84.31 on 2 and 25 DF,  p-value: 7.719e-12

> anova(temp.lm2);
Analysis of Variance Table

Response: temp
          Df Sum Sq Mean Sq  F value    Pr(>F)
week       1 818.15  818.15 161.5064 2.080e-12 ***
weeksq     1  36.00   36.00   7.1075   0.01326 *
Residuals 25 126.64    5.07
```

(a) Consider the model `temp` $= b_0 + b_1 \times$ `week` $+e$. According to the scatter plot of `temp` versus `week` and the residual plot below, perform model diagnostics.



(b) For the model `temp` $= b_0 + b_1 \times$ `week` $+e$, perform a lack of fit (LOF) test.

(c) For the model `temp` $= b_0 + b_1 \times$ `week` $+e$, construct a 95% confidence interval for the expected weekly average temperature in the 18^{th} week.

(d) Now consider the model `temp` $= b_0 + b_1 \times$ `week` $+ b_2 \times$ `weeksq` $+e$. Perform a lack of fit (LOF) test.

2. An experiment was carried out on two turkey farms to understand the relation between weight gain of turkey in kg (`y`) and the level of a diet supplement in g (`w1`). The first 5 rows of the data `turkey` corresponds to farm A and the next 5 rows correspond to farm B. The following (edited) R code and output may be used to answer Questions 2(a)–(d).

```
> turkey;
    w1 w2  w3    y
1  2.0  0 0.0 3.74
2  2.5  0 0.0 4.31
3  3.0  0 0.0 4.50
4  3.5  0 0.0 5.39
5  4.0  0 0.0 5.54
6  2.0  1 2.0 5.71
7  2.5  1 2.5 6.37
8  3.0  1 3.0 6.82
9  3.5  1 3.5 7.19
10 4.0  1 4.0 7.41

> turkey.lm = lm(y~w1+w2+w3, data=turkey);
> summary(turkey.lm);

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.8880     0.3299   5.723 0.001234 **
```

```
w1            0.9360     0.1070   8.745 0.000124 ***
w2            2.2800     0.4665   4.887 0.002747 **
w3           -0.0920     0.1514  -0.608 0.565604
---

Residual standard error: 0.1692 on 6 degrees of freedom
Multiple R-Squared: 0.9879,     Adjusted R-squared: 0.9818
F-statistic: 163.1 on 3 and 6 DF,  p-value: 3.873e-06

> anova(turkey.lm);
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq  F value    Pr(>F)
w1         1  3.9605  3.9605 138.2856 2.283e-05 ***
w2         1 10.0400 10.0400 350.5601 1.499e-06 ***
w3         1  0.0106  0.0106   0.3694    0.5656
Residuals  6  0.1718  0.0286
---
```

(a) Consider the model $y = b_0 + b_1 w_1 + b_2 w_2 + b_3 w_3 + e$. Briefly interpret b_3 and perform a test for $H_0 : [b_3 = 0 | b_0, b_1, b_2]$.

(b) Consider the model $y = b_0 + b_1 w_1 + b_2 w_2 + b_3 w_3 + e$. Briefly interpret b_0 and construct a 90% confidence interval for b_0.

(c) Perform a test to determine whether the relation between weight gain and diet supplement level is the same on the two farms.

(d) Perform a test for $H_0 : [b_2 = 0 | b_0, b_1]$ and briefly interpret the result under the context of this experiment.

3. The data `study` consists of a response variable y and three explanatory variables x_1, x_2, x_3. In addition, an explanatory variable x_4 is created such that it has the value of 1 for the third observation and 0 for all other observations. Use the following (edited) R code and output to answer Questions 3(a)–(c).

```
> cor(study);
           x1         x2         x3         x4          y
x1  1.0000000  0.7971536 -0.9561373 -0.4294967 -0.4332871
x2  0.7971536  1.0000000 -0.7317592 -0.1929603  0.1068198
x3 -0.9561373 -0.7317592  1.0000000  0.3117618  0.5926742
x4 -0.4294967 -0.1929603  0.3117618  1.0000000  0.2514859
y  -0.4332871  0.1068198  0.5926742  0.2514859  1.0000000

> study.lm23 = lm(y~x2+x3, data=study);
> summary(study.lm23);

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.5130     1.3892   8.288 1.67e-05 ***
x2            3.1730     0.1877  16.905 3.98e-08 ***
x3            3.3085     0.1577  20.981 5.95e-09 ***
---
```

```
Residual standard error: 0.7802 on 9 degrees of freedom
Multiple R-Squared: 0.9802,     Adjusted R-squared: 0.9758
F-statistic: 222.7 on 2 and 9 DF,  p-value: 2.166e-08

> anova(study.lm23);
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq  F value   Pr(>F)
x2         1   3.156   3.156   5.1847   0.0488 *
x3         1 267.924 267.924 440.1964 5.95e-09 ***
Residuals  9   5.478   0.609
---

> study.lm123 = lm(y~x1+x2+x3, data=study);
> summary(study.lm123);

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.4114     1.6571   5.679 0.000465 ***
x1            1.8841     0.9994   1.885 0.096138 .
x2            2.9982     0.1898  15.794 2.58e-07 ***
x3            3.8700     0.3287  11.772 2.48e-06 ***
---

Residual standard error: 0.6886 on 8 degrees of freedom
Multiple R-Squared: 0.9863,     Adjusted R-squared: 0.9811
F-statistic: 191.8 on 3 and 8 DF,  p-value: 8.659e-08

> anova(study.lm123);
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x1         1  51.920  51.920  109.51 6.041e-06 ***
x2         1 155.141 155.141  327.22 8.955e-08 ***
x3         1  65.704  65.704  138.58 2.481e-06 ***
Residuals  8   3.793   0.474

> study.lm1234 = lm(y~x1+x2+x3+x4, data=study);
> summary(study.lm1234);

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.8393     1.4579   4.691  0.00223 **
x1            3.3599     0.8645   3.886  0.00600 **
x2            2.8381     0.1450  19.577 2.26e-07 ***
x3            4.2151     0.2603  16.192 8.34e-07 ***
x4            1.9560     0.6557   2.983  0.02043 *
---
```

```
Residual standard error: 0.4884 on 7 degrees of freedom
Multiple R-Squared: 0.994,      Adjusted R-squared: 0.9905
F-statistic: 288.1 on 4 and 7 DF,  p-value: 7.664e-08

> anova(study.lm1234);
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq  F value    Pr(>F)
x1         1  51.920  51.920 217.6265 1.574e-06 ***
x2         1 155.141 155.141 650.2793 3.643e-08 ***
x3         1  65.704  65.704 275.4013 7.046e-07 ***
x4         1   2.123   2.123   8.8981   0.02043 *
Residuals  7   1.670   0.239
---
```

(a) Consider the three simple linear regression models: $y = b_0 + b_1x_1 + e$, $y = b_0 + b_1x_2 + e$, and $y = b_0 + b_1x_3 + e$. Which model has the smallest MSE? Briefly explain why.

(b) For the model $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + e$, the third observation has the largest studentized residual. Perform an outlier test on this observation.

(c) Perform a backward elimination for model selection, starting with the full model $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + e$.