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:

<table>
<thead>
<tr>
<th>Question</th>
<th>Possible Points</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>
1. A limnologist is interested in the relationship between clarity of a lake and size of a lake. The following data are recorded on the size and the clarity of 15 lakes.

<table>
<thead>
<tr>
<th>lake ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>size</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>10</td>
<td>3</td>
<td>9</td>
<td>10</td>
<td>7</td>
<td>7</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>7</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>clarity</td>
<td>46</td>
<td>58</td>
<td>55</td>
<td>57</td>
<td>41</td>
<td>57</td>
<td>59</td>
<td>52</td>
<td>55</td>
<td>38</td>
<td>48</td>
<td>39</td>
<td>53</td>
<td>50</td>
<td>57</td>
</tr>
</tbody>
</table>

A simple linear regression (SLR) model $\text{clarity} = b_0 + b_1 \times \text{size} + e$ is fitted to the data set and some of the R code and output are as follows.

```r
> lake;

       size clarity
1        3     46
2        4     58
3        6     55
4       10     57
5        3     41
6        9     57
7       10     59
8        7     52
9        7     55
10       1     38
11       3     48
12       2     39
13       7     53
14       4     50
15       8     57

> lake.lm = lm(clarity~size, data=lake);
> summary(lake.lm);

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.5789   2.2357 17.703  1.76e-10 ***
size        2.0395   0.3559  5.731  6.93e-05 ***
---
Residual standard error: 3.924 on 13 degrees of freedom
Multiple R-Squared: 0.7164, Adjusted R-squared: 0.6946
F-statistic: 32.84 on 1 and 13 DF, p-value: 6.928e-05

> anova(lake.lm);

Response: clarity

Df  Sum Sq Mean Sq  F value Pr(>F)
size  1 505.79  505.79  32.84 0.05 ***
Residuals 13 200.21 15.40
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> lake.aov = lm(clarity~factor(size), data=lake);
> anova(lake.aov);
```
Response: clarity

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>factor(size)</td>
<td>8</td>
<td>641.33</td>
<td>80.17</td>
<td>7.4381</td>
</tr>
<tr>
<td>Residuals</td>
<td>6</td>
<td>64.67</td>
<td>10.78</td>
<td></td>
</tr>
</tbody>
</table>

> anova(lake.lm, lake.aov);

Model 1: clarity ~ size
Model 2: clarity ~ factor(size)

<table>
<thead>
<tr>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>200.211</td>
<td>1</td>
<td>13</td>
<td>135.544</td>
<td>1.7966</td>
</tr>
<tr>
<td>2</td>
<td>64.667</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

> laker = lake[-2,];
> laker.lm = lm(clarity~size, data=laker);
> summary(laker.lm);

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 38.0000 | 1.5636 | 24.304 | 1.42e-11 *** |
| size | 2.1875 | 0.2438 | 8.974 | 1.14e-06 *** |

---

Residual standard error: 2.658 on 12 degrees of freedom
Multiple R-Squared: 0.8703, Adjusted R-squared: 0.8595
F-statistic: 80.53 on 1 and 12 DF, p-value: 1.138e-06

> anova(laker.lm);

Response: clarity

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>size</td>
<td>1</td>
<td>568.75</td>
<td>568.75</td>
<td>80.531</td>
</tr>
<tr>
<td>Residuals</td>
<td>12</td>
<td>84.75</td>
<td>7.06</td>
<td></td>
</tr>
</tbody>
</table>

> predict(laker.lm, data.frame(size=4), se.fit=T);

$fit
[1] 46.75

$se.fit
[1] 0.824067

(a) Perform a lack of fit test (LOF) for the SLR model clarity = b_0 + b_1 \times size + e.

(b) The second observation size = 4, clarity = 58 has the largest studentized residual. Perform a formal test to determine whether this observation is an outlier.

(c) Regardless of your conclusion in (b), is the second observation size = 4, clarity = 58 an influential observation? Give reasons in 1–2 short sentences.
2. The data stored in problem consist of a dependent variable $y$ and 3 independent variables $x_1, x_2, x_3$. The following R output is available.

```r
> problem = read.table("problem.dat", header=T);
> problem.lm1 = lm(y~x1, data=problem);
> summary(problem.lm1);

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 124.4303    7.3788  16.863  1.92e-14 ***
x1          0.5554    0.6378   0.871   0.393

Residual standard error: 15.68 on 23 degrees of freedom
Multiple R-Squared: 0.03191,   Adjusted R-squared: -0.01018
F-statistic: 0.7582 on 1 and 23 DF,  p-value: 0.3929

> problem.lm2 = lm(y~x2, data=problem);
> summary(problem.lm2);

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 250.385     38.921  6.433   1.45e-06 ***
x2         -1.198      0.387  -3.094   0.00512 **

Residual standard error: 13.39 on 23 degrees of freedom
Multiple R-Squared: 0.2939,   Adjusted R-squared: 0.2632
F-statistic: 9.573 on 1 and 23 DF,  p-value: 0.005119

> problem.lm3 = lm(y~x3, data=problem);
> summary(problem.lm3);

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 201.6314    16.5439 12.188   1.62e-11 ***
x3          -1.4093    0.3233  -4.359   0.00023 ***

Residual standard error: 11.79 on 23 degrees of freedom
Multiple R-Squared: 0.4524,   Adjusted R-squared: 0.4286
F-statistic: 19 on 1 and 23 DF,  p-value: 0.0002298

> problem.lm12 = lm(y~x1+x2, data=problem);
> summary(problem.lm12);

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 251.1735    44.8216  5.604   1.24e-05 ***
x1          -0.0226     0.5925  -0.038    0.96984  
x2         -1.2030     0.4210  -2.857    0.00916 **
```

```
---

Residual standard error: 13.69 on 22 degrees of freedom  
Multiple R-Squared: 0.2939, Adjusted R-squared: 0.2298  
F-statistic: 4.58 on 2 and 22 DF, p-value: 0.02174

> problem.lm13 = lm(y~x1+x3, data=problem);  
> summary(problem.lm13);

Coefficients:  
Estimate Std. Error t value Pr(>|t|)  
(Intercept) 203.36968 19.99853 10.169 8.89e-10 ***  
x1 -0.08366 0.51418 -0.163 0.872236  
x3 -1.42629 0.34651 -4.116 0.000454 ***  
---

Residual standard error: 12.05 on 22 degrees of freedom  
Multiple R-Squared: 0.4531, Adjusted R-squared: 0.4034  
F-statistic: 9.113 on 2 and 22 DF, p-value: 0.001309

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

Coefficients:  
Estimate Std. Error t value Pr(>|t|)  
(Intercept) 64.835 58.841 1.102 0.28243  
x2 2.513 1.045 2.405 0.02503 *  
x3 -3.685 0.991 -3.719 0.00120 **  
---

Residual standard error: 10.73 on 22 degrees of freedom  
Multiple R-Squared: 0.5664, Adjusted R-squared: 0.527  
F-statistic: 14.37 on 2 and 22 DF, p-value: 0.0001018

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

Coefficients:  
Estimate Std. Error t value Pr(>|t|)  
(Intercept) 59.2246 63.6825 0.930 0.3629  
x1 0.1273 0.4762 0.267 0.7919  
x2 2.5671 1.0869 2.362 0.0279 *  
x3 -3.7085 1.0164 -3.649 0.0015 **  
---

Residual standard error: 10.96 on 21 degrees of freedom  
Multiple R-Squared: 0.5679, Adjusted R-squared: 0.5062  
F-statistic: 9.2 on 3 and 21 DF, p-value: 0.0004392

Based on the R output, perform a stepwise selection to determine the best regression model for $y$. Give all the steps in detail.
3. A researcher in remote sensing is interested in comparing clarity \((y)\) of 15 lakes in three different ecoregions (A, B, and C). Consider the model \(y = b_0 + b_1w_1 + b_2w_2 + e\). The first 5 rows of the data (\texttt{ecolake}) correspond to ecoregion A, the next 5 rows to ecoregion B, and the last 5 rows to ecoregion C. The following R output may be of use.

\[
\begin{array}{rrr}
  & w1 & w2 & y \\
 1 & 0 & 0 & 37 \\
 2 & 0 & 0 & 41 \\
 3 & 0 & 0 & 37 \\
 4 & 0 & 0 & 46 \\
 5 & 0 & 0 & 41 \\
 6 & 0 & 1 & 17 \\
 7 & 0 & 1 & 22 \\
 8 & 0 & 1 & 23 \\
 9 & 0 & 1 & 22 \\
 10 & 0 & 1 & 19 \\
 11 & 1 & 0 & 56 \\
 12 & 1 & 0 & 52 \\
 13 & 1 & 0 & 48 \\
 14 & 1 & 0 & 41 \\
 15 & 1 & 0 & 54 \\
\end{array}
\]

\[
> \text{ecolake.lm = lm(y~w1+w2, data=ecolake)};
> \text{summary(ecolake.lm)};
\]

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 40.400 | 1.920 | 21.041 | 7.71e-11 *** |
| w1 | 9.800 | 2.715 | 3.609 | 0.00359 ** |
| w2 | -19.800 | 2.715 | -7.292 | 9.58e-06 *** |

Residual standard error: 4.293 on 12 degrees of freedom
Multiple R-Squared: 0.9113, Adjusted R-squared: 0.8966
F-statistic: 61.67 on 2 and 12 DF, p-value: 4.857e-07

\[
> \text{anova(ecolake.lm)};
\]

Analysis of Variance Table

<table>
<thead>
<tr>
<th>Response: y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Df</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>w1</td>
</tr>
<tr>
<td>w2</td>
</tr>
<tr>
<td>Residuals</td>
</tr>
</tbody>
</table>

(a) Briefly interpret \(b_0\) and construct a 90% confidence interval for \(b_0\).

(b) Perform a test to determine whether the expected lake clarity is the same for all three ecoregions.
(c) Assuming that the expected lake clarity is the same in ecoregions A and C, perform a test to determine whether the lake clarity is the same in ecoregions A and B.