Advanced Statistical Regression Analysis: Comprehensive Final Examination

Name: Answer Key

Instructions: Read each question carefully before determining the best answer. Show all work; supporting computer code and output must be attached to the question for which it is used. Use computer output as a guide only; all questions must be answered in full on the exam paper. Report all final numerical answers to a precision of 4 units past the decimal point (e.g., 18.1234 or 1.1234 \times 10^{-5}). There are 100 total points on this exam.

Do not discuss this exam or its components with anyone besides the course instructor or the course TA. This exam is due by 1:00 PM, December 18, 2018.

1. Return to the dispersal rate data from the previous exam, where the dispersal rate (Y) and the temperature (X_1) were observed in n = 9 samples of tea. Also recorded in that study was a second predictor variable, the time (X_2) the tea was brewed. Download the data from http://math.arizona.edu/~piegorsch/571A/finalprob1.csv.

a. (10 points) Construct a scatterplot matrix to visualize the data.

Sample R commands:

```
dispersal.df = read.csv( file.choose() )
attach( dispersal.df ); Y = dispersal.rate
X1 = temperature; X2 = time
pairs(cbind(Y, X1, X2), pch=19)
```

which produces the following plot. The only strong (linear) relationship appears to be between Y and X_1.

[Scatterplot matrix image]

cont’d
1. Tea data (cont’d)

b. (10 points) Assume a multiple linear regression (MLR) model for these data, with \( Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}X_{i2} + \epsilon_i \) and \( \epsilon_i \sim \text{i.i.d.} \text{N}(0, \sigma^2) \), \( i=1,...,n \). Assess whether the inclusion of \( X_2 \) and the \( X_1X_2 \) interaction significantly improves the model fit. State the null hypothesis and employ a single test statistic for performing the test. Operate at a false positive rate of 0.5%. (Note: for purposes of this problem, you do not need to center \( X_1 \) or \( X_2 \).)

Sample R commands:

```
dispersalFM.lm = lm(Y ~ X1*X2)
dispersalRM.lm = lm(Y ~ X1)
anova( dispersalRM.lm, dispersalFM.lm )
```

These produce the following output (edited):

```
Analysis of Variance Table
Model 1: Y ~ X1
Model 2: Y ~ X1 * X2

<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>0.00165707</td>
<td>2</td>
<td>0.00073684</td>
<td>2.0018</td>
<td>0.2298</td>
</tr>
</tbody>
</table>
```

To test the 2 d.f. hypothesis \( H_0: \beta_2 = \beta_3 = 0 \) we see the p-value is \( P = 0.2289 \). As \( P = 0.2298 > \alpha = 0.005 \), we fail to reject \( H_0 \) and conclude that, jointly, \( X_2 \) and the \( X_1X_2 \) interaction do not significantly improve the model fit.

c. (10 points) Are there any concerns with multicollinearity with the full MLR model fit in part (b)?

Find the VIFs. Sample R code and (edited) output:

```
require( car )
vif( dispersalFM.lm )
```

```
X1       X2    X1:X2
14.50000 30.85135 44.35135
```

```
mean( vif(dispersalFM.lm) )
```

```
[1] 29.9009
```

Yes! There appear to be problems with multicollinearity, since all VIFs are above 10 and \( \sqrt{VIF} = 29.9009 \geq 6 \).

(However, as both \( X_2 \) and \( X_1X_2 \) appear to be jointly insignificant, we should simply remove them from the model rather than attempt to correct the multicollinearity.)
1. Tea data (cont’d)

d. (10 points) Find the Studentized deleted residuals for the full MLR model fit in part (b) and plot them against the fitted values. Also include an assessment of possible outlying residuals (set your \( \alpha \)-level to 1%, but remember to adjust for multiplicity). Are any untoward patterns evident? Are there any potential outliers?

Sample R code and consequent plot (dashed lines form exceedance limits):

```r
plot( rstudent(dispersalFM.lm)~fitted(dispersalFM.lm), ylim=c(-8.5,8.5) )
abline( h=0 )

n = length(Y); p = length(coef(dispersalFM.lm))
tcrit = qt( 1-(.01/(2*n)), df=n-p-1 )
abline( h=tcrit, lty=2 ); abline( h=-tcrit, lty=2 )
```

The plot is relatively stable with no gross patterns. No points lie above or below the exceedance limits at \( t_{\text{crit}} = \pm 8.376331 \), so there are no potential outliers.
2. The effectiveness of a new experimental pharmaceutical in reducing anxiety was studied in 12 patients. In the data that follow, $X_i$ denotes the dose (in mg.) of the $i$th patient and $Y_i$ denotes % relief reported.

<table>
<thead>
<tr>
<th>$i$</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>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_i$</td>
<td>22</td>
<td>20</td>
<td>28</td>
<td>31</td>
<td>37</td>
<td>38</td>
<td>46</td>
<td>44</td>
<td>54</td>
<td>51</td>
<td>59</td>
<td>62</td>
</tr>
<tr>
<td>$X_{i1}$</td>
<td>35</td>
<td>35</td>
<td>40</td>
<td>40</td>
<td>45</td>
<td>45</td>
<td>50</td>
<td>50</td>
<td>55</td>
<td>55</td>
<td>60</td>
<td>60</td>
</tr>
</tbody>
</table>

These data are found in the file [http://math.arizona.edu/~piegorsch/571A/finalprob2.csv](http://math.arizona.edu/~piegorsch/571A/finalprob2.csv).

It is felt that a simple linear regression model of the form $Y_i \sim \text{indep. } \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$ ($i = 1, \ldots, n$) is a reasonable approximation for these data.

a. (10 points) Fit the linear model suggested above. Overlay a plot of the original data and the fitted curve. Comment on the visual adequacy of the fit.

*Always Plot the Data! Sample R code:*

```r
anxiety.df = read.csv( file.choose() )
attach( anxiety.df )
n = length(Y)
anxiety.lm = lm (Y ~ X1)
plot( Y~X1, pch=19 ); abline( anxiety.lm )
```

The overlay implies an acceptable fit of the SLR model.

cont’d
2. Anxiety data (cont’d)

b. (10 points) Corroborate the indication given in part (a) and display a (raw) residual plot for your SLR fit. Comment on its visual adequacy.

Clearly, a residual analysis should also be performed. Sample R code is

```r
plot( resid(anxiety.lm) ~ fitted(anxiety.lm), pch=19 )
abline( h=0 )
```

The (raw) residual plot shows a generally good fit with no strong anomalies.

c. (10 points) It is important to estimate the dose at which a 50% effect is attained. (This is termed the median effective dose.) Call this \( \theta \). An expression for \( \theta \) in terms of the \( \beta \)-parameters under this simple linear model is known to be \( \theta = (50 - \beta_0)/\beta_1 \). For the data above, find an estimate of \( \theta \)

Sample R code:

```r
thetahat = (50 - coef(anxiety.lm))[1]/coef(anxiety.lm)[2]
```

which gives \( \hat{\theta} = 53.2482 \text{ mg} \).
2. Anxiety data (cont’d)

d. (10 points) Use bootstrap resampling to find a 95% confidence interval for $\theta$ in part (c) via the percentile method. Set your seed to 2019 and employ 5000 bootstrap resamples.

Sample R code for the bootstrap confidence limits:

```R
ei = resid( anxiety.lm )
Yhat = fitted( anxiety.lm )
B = 5000
theta = numeric(B)
set.seed( 2019 ) #sets seed for random sampler
for( b in 1:B ) {
    estar = sample( ei, n, replace=T )
    Ystar = Yhat + estar
    temp.lm = lm( Ystar ~ X1 )
    theta[b] = (50 - coef(temp.lm)[1])/coef(temp.lm)[2]
}  #end for loop
theta = sort(theta)  #order theta from small-to-large
c( theta[126], theta[4875] )
```

The consequent output is

```
[1] 52.72446  53.82118
```

suggesting that the 95% bootstrap confidence limits are $52.7245 < \theta < 53.8212$ mg.
2. Anxiety data (cont’d)

   e. (10 points) An additional predictor variable, $X_2 =$ “patient age”, was also available (see the full data set in the link given above). Thus the data become

   $\begin{array}{c|cccccccccccc}
   i & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\
   \hline
   Y_i & 22 & 20 & 28 & 31 & 37 & 38 & 46 & 44 & 54 & 51 & 59 & 62 \\
   X_{i1} & 35 & 35 & 40 & 40 & 45 & 45 & 50 & 50 & 55 & 55 & 60 & 60 \\
   X_{2i} & 22 & 37 & 17 & 31 & 37 & 29 & 29 & 38 & 37 & 30 & 26 & 25 \\
   \end{array}$

   Construct an added variable plot to determine graphically if the additional variable $X_2$ should be considered

   **Sample R code:**
   ```r
   library( car )
   avPlot( model=lm( Y~X1+X2 ), variable=X2 )
   ```

   This produces the following graphic:

   ![Added variable plot](image)

   The plot is essentially flat, with little pattern. This suggests that the addition of $X_2$ would not provide sufficient improvement of the fit over the existing $X_1$ variable.
3. (10 points) A sample of n = 2000 data pairs is available on US airline on-time performance, with X = Flight distance and Y = log\{Arrival delay\}. (Download the data from [http://math.arizona.edu/~piegorsch/571A/finalprob3.csv](http://math.arizona.edu/~piegorsch/571A/finalprob3.csv). The Y values have already been log-transformed. Values with Y=0 indicate no arrival delay.) Construct a loess smooth of Y against X. Use second-order smoothing and employ robust fitting with a smoothing parameter of $q = 1/3$. Overlay the loess smooth on a scatterplot of the data. Comment on any patterns in the plot.

Sample R code is

```r
airline.df = read.csv(file.choose()); attach( airline.df )
X = dist; Y = logArrivalDelay
fit.loess = loess( Y ~ X, span=1/3, degree=2, family='symmetric' )
Ysmooth = predict( fit.loess, data.frame(X = seq(min(X),max(X)) ) )
plot( Y ~ X, xlab='X', ylab='Y' , ylim=c(min(Y),max(Y)), xlim = c(min(X),max(X)), cex.lab=1.35, col='gray'); par( new=T )
plot( Ysmooth ~ seq(min(X),max(X)), type='l', lwd=2, xlab='', ylab='', ylim=c(min(Y),max(Y)), xlim = c(min(X),max(X)) )
```

See left panel below. One could also use `scatter.smooth()`:

```r
scatter.smooth(fit.loess, xlab='X', ylab='Y', span=1/3, degree=2, family='symmetric', col='gray')
```

See right panel below.

In either plot, the loess smooth shows a varying, slightly curvilinear pattern, but relative to the amount of variation in the observations the pattern is essentially flat.