TEXAS A&M UNIVERSITY

Department of Statistics

STAT 608 - Regression Analysis

Homework VI

Salih Kilicli

July 18, 2019

Question 1: **Download the attached data Oral.xls. The data come from a study of the oral condition of cancer patients. Twenty five patients were divided into two groups (TRT) at random: One group received a placebo $(TRT = 0)$ while the other group received aloe juice treatment $(TRT = 1)$. The initial cancer stage $(STAGE)$ of each patient, was recorded at the beginning of the study. The oral condition at the end of the sixth week $(TOTALCW6)$, which is the dependent variable, was also recorded. Assume that in both treatment groups the regression of $TOTALCW6$ on $STAGE$ is linear and that all the usual distributional assumptions are satisfied. Test at the 5% level of significance the null hypothesis that the slopes of the two regression lines are identical. Describe concisely, in the form of a few bulleted remarks, the method that you apply and report any statistics that you consider relevant in making the hypothesis test.**

Solution: There are multiple ways to solve this problem.

Method 1: Consider the full model:

$$TOTALCW6 = \beta_0 + \beta_1\, TRT + \beta_2\, STAGE + \beta_3\, TRT \times STAGE + error$$

For $TRT = 1$ the slope is $(\beta_2 + \beta_3)$ whereas it is β_3 when $TRT = 0$. Therefore the Null Hypothesis in this case is given by:

$$H_0 : \beta_2 + \beta_3 = \beta_3 \Rightarrow H_0 : \beta_3 = 0$$

Fitting the model given above we get $\hat{\beta}_3 = -2.0149 \quad \text{and} \quad se(\hat{\beta}_3) = 1.0123$. Therefore, $T-$score of the Null Hypothesis is:

$$T = \frac{|\beta_3 - 0|}{se(\beta_3)} = \frac{2.0149}{1.0123} \approx 1.9904 < 2.0930 = t_{0.05,19}$$

which implies we fail to reject the Null Hypothesis since $T < t_{critical}$.

Method 2: We can also create a reduced model and compare it with the full model where the reduced model is:

$$TOTALCW6 = \beta_0 + \beta_1\, TRT + \beta_2\, STAGE + error$$

$ANOVA(Model_{reduced}, Model_{full})$ yields $p = 0.06112 > 0.05 = \alpha$ which implies the reduced model is preferred and therefore we fail to reject $H_0 : \beta_3 = 0$

Question 2: **Work Exercise 1 on page 252 of the textbook. (Chapter 7)**

Solution :

a) Looking at table 7.4 Model 2 and Model 3 are pretty close but AIC and BIC values are smaller for the model with two predictor X1 and X2. Both of them look optimal, but if one needs to pick one of them Model with 2 variables looks slightly better.

b) Looking at AIC and BIC values of forward selection it is clear that model with only X3 variable have the smallest values. So it is the optimal model based on AIC and BIC in terms of forward selection.

c) The reason models are different is that Model with 2 predictors has a multicollinearity issue where $VIF = \frac{1}{1-(0.999987)^2} \approx 44248.04$ for both variables which is much bigger than even 5.

d) I would recommend the simplest model with only X3 in it. X1 and X2 carries very similar information since they are almost perfectly correlated. In such case, coefficients are estimated poorly and p-values might be misleading. Also, R^2 and R^2_{adj} being equal to 1 in models that includes X1, X2 together is not realistic.

Question 3: **Work Exercise 3 on page 261 in the textbook. (Chapter 7)**

Solution: a) In the table below the values of R^2_{adj} , AIC, AICc and BIC given for number of predictors listed. Optimal values of each column is given by bold color.

Model	R^2_{adj}	AIC	AICc	BIC
1	0.25	-62.52	-62.39	-55.96
2	0.49	-135.22	-135.01	-125.39
3	0.54	-155.31	-154.99	**-142.20**
4	0.54	-156.29	-155.85	-139.90
5	**0.55**	**-156.64**	**-156.05**	-136.97
6	0.54	-154.73	-153.96	-131.78
7	**0.55**	-152.74	-151.77	-126.51

Therefore, the optimal model based on BIC is:

$$log(PrizeMoney) = \beta_0+\beta_1(GIR)+\beta_2(BirdieConversion)+\beta_3(Scrambling)+error$$

Rest of the other criterions agree that the best model is:

$$\begin{aligned} log(PrizeMoney) = \beta_0 &+ \beta_1(GIR) + \beta_2(BirdieConversion) + \beta_3(Scrambling) \\ &+ \beta_4(PuttsPerRound) + \beta_5(SandSaves) + error \end{aligned}$$

b) The final model by backwards selection in the last model described above:

$$\begin{aligned} log(PrizeMoney) = \beta_0 &+ \beta_1(GIR) + \beta_2(PuttsPerRound) + \beta_3(BirdieConversion) \\ &+ \beta_4(Scrambling) + \beta_5(SandSaves) + error \end{aligned}$$

Removing any fifth predictor in the model **only increases AIC**. The final model by backwards selection in the last model described above:

$$log(PrizeMoney) = \beta_0+\beta_1(GIR)+\beta_2(BirdieConversion)+\beta_3(Scrambling)+error$$

Removing any third predictor in the model **only increases BIC**.

c) The final model selected in the last model described above:

$$\begin{aligned} log(PrizeMoney) = \beta_0 &+ \beta_1(GIR) + \beta_2(PuttsPerRound) + \beta_3(BirdieConversion) \\ &+ \beta_4(Scrambling) + \beta_5(SandSaves) + error \end{aligned}$$

Adding any predictor as a sixth predictor in the model **only increases AIC**. The final model selected in the last model described above:

$$\begin{aligned} log(PrizeMoney) = \beta_0 &+ \beta_1(GIR) + \beta_2(PuttsPerRound) + \beta_3(BirdieConversion) \\ &+ \beta_4(Scrambling) + error \end{aligned}$$

Adding any predictor as a fifth predictor in the model **only increases BIC**.

d) Both forward and backward selection are approximate methods to find the "best" model. Neither method is guaranteed to find the optimal subset. On the other hand, an exhaustive search (best subsets) will find the optimal predictors given the data and a class of models. Moreover, the forward method starts with less information (only one predictor, then two predictors, etc) while backwards has all

the information to consider before the first selection is made. This would seem to be a favorable situation for the backward selection process if computationally feasible. Then it is not surprising that backwards and best subsets algorithms agree while forward selection does not.

e) Recommending a final model is important in terms of your goal. In Ch.6 (p. 224), the research question was stated as **"what is the relative importance of each different aspect of the game on average prize money in professional golf"**? Therefore, I would prefer a fuller model (model with 5 predictors) to a more simpler model so that parameter estimates corresponding to the different aspects can be compared. If the goal is finding better prediction, then I would naively go with the BIC model with 3 predictors. In general, it's best to use a model from best subsets as discussed above.

f) Both GIR and BirdieConversion coefficients are highly significant (slopes are different from zero) and larger values correspond to more prize money. Since the model is a result of data-driven variable selection, I would say that the other three predictors are **not different** from zero at the 5% significance level. Therefore, there is a trend towards better Scrambling/sandSaves and fewer Putts correspond to more prize money.

Question 4: **Work Exercise 2 on page 122 of the textbook (Chapter 4).**

Solution: Let $W^{1/2} = \begin{bmatrix} \frac{1}{x_1} & & \\ & \ddots & \\ & & \frac{1}{x_n} \end{bmatrix}$, $W = \begin{bmatrix} \frac{1}{x_1^2} & & \\ & \ddots & \\ & & \frac{1}{x_n^2} \end{bmatrix}$, $X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$, $Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, $e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$.

and the matrix form of the model is given by $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$. Then multiplying both sides of the matrix equation by $W^{1/2}$ we get:

$$W^{1/2}Y = W^{1/2}X\beta + W^{1/2}e$$

One can show then, the matrix form of WLS estimator β is given by:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= ((W^{1/2}X)'W^{1/2}X)^{-1}(W^{1/2}X)'W^{1/2}Y = (X'W^{1/2}W^{1/2}X)^{-1}X'W^{1/2}W^{1/2}Y \\
&= (X'WX)^{-1}X'WY \\
&= \begin{bmatrix} x_1, & x_2, \ldots, & x_n \end{bmatrix} \begin{bmatrix} \frac{1}{x_1^2} & & \\ & \ddots & \\ & & \frac{1}{x_n^2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \begin{bmatrix} x_1, & x_2, \ldots, & x_n \end{bmatrix} \begin{bmatrix} \frac{1}{x_1^2} & & \\ & \ddots & \\ & & \frac{1}{x_n^2} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \\
&= \Big(\frac{1}{x_1^2}\, x_1^2 + \cdots + \frac{1}{x_n^2}\, x_n^2\Big)\Big(\frac{1}{x_1^2}\, x_1 y_1 + \cdots + \frac{1}{x_n^2}\, x_n y_n\Big) \\
&= n\sum_{i=1}^{n} \frac{y_i}{x_i}
\end{aligned}$$

Therefore, the WLS estimator is $\hat{\boldsymbol{\beta}} = n\sum\limits_{i=1}^{n} \frac{y_i}{x_i}$.

Question 5: **A group of $m+n$ arthritis sufferers was randomly divided into two groups, A and B. The patients in group A were treated with a placebo while those in group B were treated with a prescription drug. After one month of treatment, the times in seconds taken to walk 10 yards was measured. The measured times in the two groups were $A_1, \ldots, A_m$ and $B_1, \ldots, B_n$ .It is required to test the hypothesis $H0 : \mu = \eta$ where μ and η denote respectively the mean walking times in the (hypothetical) populations of sufferers treated with the placebo and prescription drug respectively. Express the data in linear model form:**

$$E[Y_i|x_i] = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, (m+n)$$

in such a manner that the hypothesis $H_0 : \mu = \eta$ is equivalent to the hypothesis $H_0 : \beta_1 = 0$. Express clearly $\hat{\beta}_0$ and $\hat{\beta}_1$ in terms of μ and η and express $Y_1, \ldots, Y_{m+n}$ in terms of the A_i and B_i. Also give the numerical values of $x_1, \ldots, x_{m+n}$.

Solution: 5.1

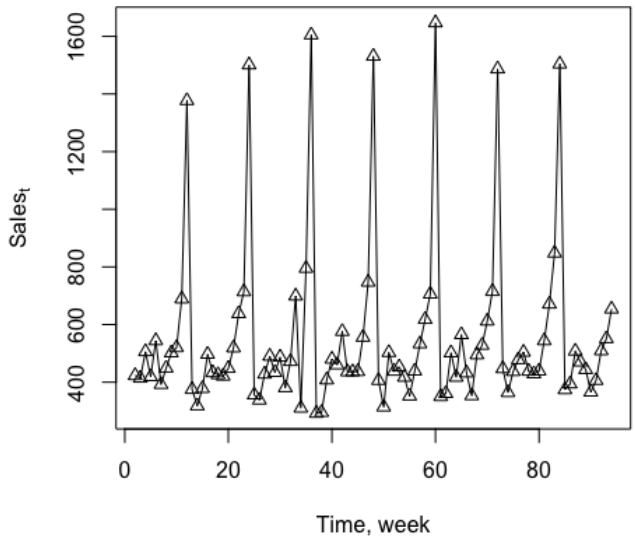
$$\begin{aligned} E[\hat{\alpha}|X] &= E[(X'X)^{-1}X'Y|X] = (X'X)^{-1}X'E[Y|X] = (X'X)^{-1}X'(X\beta + Z\gamma) \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'Z\gamma = \beta\mathbf{I} + X^{-1}\mathbf{I}Z\gamma \\ &= \beta + \gamma X^{-1}Z \end{aligned}$$

5.2 If Z is orthogonal to every column of X, then $X'Z = \mathbf{0} = 0_{m\times 1}$. Therefore,

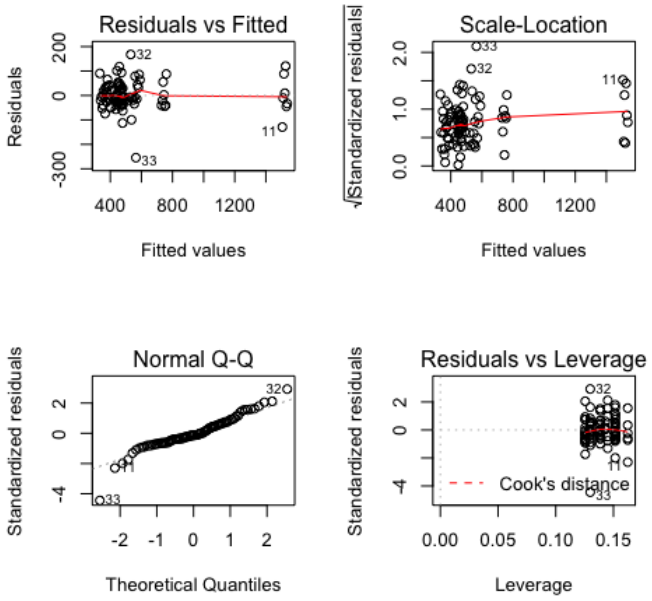
$$\begin{aligned} E[\hat{\alpha}|X] &= E[(X'X)^{-1}X'Y|X] = (X'X)^{-1}X'E[Y|X] = (X'X)^{-1}X'(X\beta + Z\gamma) \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'Z\gamma = \beta\mathbf{I} + (X'X)^{-1}\mathbf{0}\gamma \\ &= \beta \end{aligned}$$

Question 6: **Work Exercise 9.4.2, page 328-329, in the textbook. (Chapter 9)**

Solution: a) A model is fitted for Sales ignoring the effects due to Advert and Lag1Advert. First it is important to see seasonal trend in the Sales date with annual repetition. Below is the plot showing the seasonality in the data:



Here are the diagnostic plots and summary of the modeled regression.

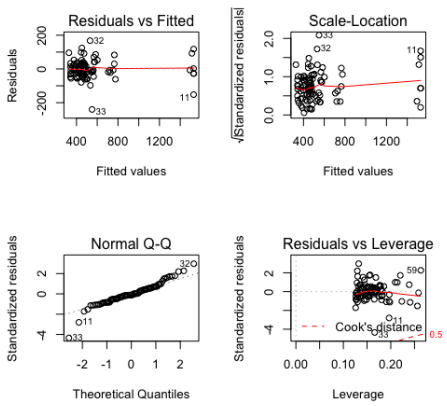


Looking at diagnostic plot we can see a slight increase in the variance of errors from top right plot but it can be ignored for now. On the other hand, there are

```
Call:
lm(formula = Sales ~ Time + Month_2 + Month_3 + Month_4 + Month_5 +
    Month_6 + Month_7 + Month_8 + Month_9 + Month_10 + Month_11 +
    Month_12)

Residuals:
     Min       1Q   Median       3Q      Max
-254.638  -33.188   -7.513   31.888  167.612

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  350.3667    25.9841  13.484  < 2e-16 ***
Time           0.4298     0.2381   1.805  0.07485 .
Month_2      -18.4044    31.8128  -0.579  0.56454
Month_3       77.5408    31.8048   2.438  0.01698 *
Month_4      101.8609    31.7985   3.203  0.00195 **
Month_5       93.4311    31.7941   2.939  0.00431 **
Month_6       89.7513    31.7914   2.823  0.00600 **
Month_7       24.8214    31.7905   0.781  0.43724
Month_8       89.0166    31.7914   2.800  0.00640 **
Month_9      166.8368    31.7941   5.247 1.24e-06 ***
Month_10     199.6569    31.7985   6.279 1.66e-08 ***
Month_11     374.2882    32.8366  11.399  < 2e-16 ***
Month_12    1150.7155    32.8340  35.047  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 61.43 on 80 degrees of freedom
Multiple R-squared:  0.9634,    Adjusted R-squared:  0.9579
F-statistic: 175.3 on 12 and 80 DF,  p-value: < 2.2e-16
```

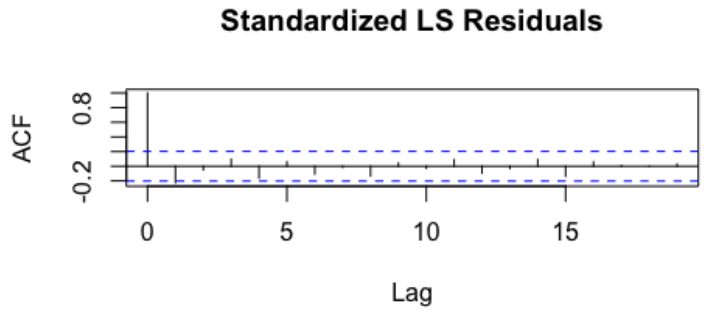
2 points with high leverage values (32, and 33) one of which is a ”bad” leverage point due to its Cook’s distance. Meanwhile, almost all of the coefficients (except $Month_2$ and $Month_7$) in the summary look highly significant and overall R^2 and F-statistics are satisfying. Also, the ACF doesn’t look significant for LAG1 so errors are not serially correlated.

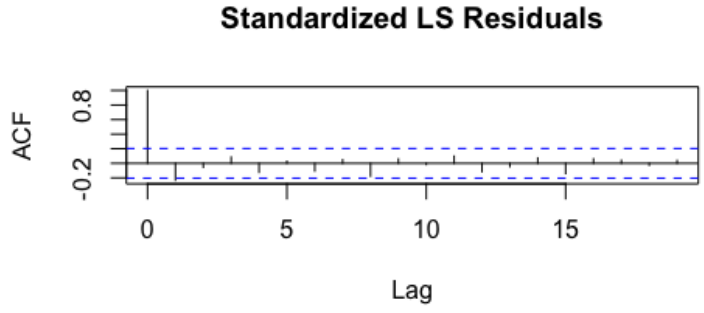
b) A new model is fitted for Sales adding the effects due to Advert and Lag1Advert. Here are the diagnostic plots and summary of the modeled regression.



```
Call:
lm(formula = Sales ~ Time + Advert + Lag1Advert + Month_2 + Month_3 +
    Month_4 + Month_5 + Month_6 + Month_7 + Month_8 + Month_9 +
    Month_10 + Month_11 + Month_12)

Residuals:
     Min       1Q   Median       3Q      Max
-240.283  -28.203   -6.308   28.352  167.624

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  392.3619    40.8122   9.614 6.99e-15 ***
Time           0.4322     0.2377   1.818 0.072853 .
Advert         2.2999     2.5743   0.893 0.374385
Lag1Advert    -4.5174     2.6625  -1.697 0.093744 .
Month_2      -45.8840    37.8276  -1.213 0.228799
Month_3       26.6907    40.5869   0.658 0.512720
Month_4       81.0497    33.5470   2.416 0.018030 *
Month_5       59.3321    36.7394   1.615 0.110361
Month_6       59.3564    36.4203   1.630 0.107184
Month_7      -12.4035    37.4233  -0.331 0.741203
Month_8       60.0235    35.9118   1.671 0.098648 .
Month_9      129.0218    37.2445   3.464 0.000867 ***
Month_10     165.7306    35.4326   4.677 1.20e-05 ***
Month_11     324.7993    43.5922   7.451 1.08e-10 ***
Month_12    1149.1212    39.4806  29.106  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.55 on 78 degrees of freedom
Multiple R-squared:  0.9653,    Adjusted R-squared:  0.9591
F-statistic:   155 on 14 and 78 DF,  p-value: < 2.2e-16
```

Diagnostic plots are pretty much same with the first model. However, there are many estimates with insignificant coefficients in the summary and overall R^2 and F-statistics are satisfying. In order to compare models we present ANOVA (run a partial F-test) and ACF plots of each case below.





Clearly, there is no significant LAG1 issue, so it is not helpful to include the lagged variable in the model. Moreover, partial F-test has a insignificant p-value which sug-

```
Analysis of Variance Table

Model 1: Sales ~ Time + Month_2 + Month_3 + Month_4 + Month_5 + Month_6 +
    Month_7 + Month_8 + Month_9 + Month_10 + Month_11 + Month_12
Model 2: Sales ~ Time + Advert + Lag1Advert + Month_2 + Month_3 + Month_4 +
    Month_5 + Month_6 + Month_7 + Month_8 + Month_9 + Month_10 +
    Month_11 + Month_12
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     80 301843
2     78 285962  2     15881 2.1659 0.1215
```

gests reduced model is better. So, I prefer the Model 1 since there is no serially correlated errors and no need to include a lagged term.

Question 7: **(Section 9.1 and 9.2, pages 305-311) A linear regression model**

$$Y_t = \beta_0 + \beta_1 X_t + e_t \qquad (2)$$

was fit to some time-series data by ordinary least squares. The residuals $\hat{e_1}, \ldots, \hat{e_n}$ from the fit were then used to create two new variables, namely Y with values $\hat{e_2}, \ldots, \hat{e_n}$ and X with values $\hat{e_1}, \ldots, \hat{e_{n-1}}$. A linear "regression through the origin" was then run with Y as dependent variable and X as predictor. The slope estimate was 0.412. Assume that the e_t in (2) follow a standard AR(1) model

$$e_t = \rho e_{t-1} + \epsilon_t,$$

where the ϵ_t are independent with a common standard normal distribution. Using ideas from regression analysis, or otherwise, estimate the first order autocorrelation ρ in the AR model. Explain clearly your calculations reasoning.

Solution: $\hat{e_t} = \hat{e}_{t-1} + \epsilon_t = 0.412\hat{e}_{t-1} + \epsilon_t$ due to linear regression estimate $\mathbf{Y} = 0.412\,\mathbf{X} + error$:

$$Y_t - \hat{Y}_t = \hat{e}_t = 0.412\hat{e}_{t-1} + error \quad \text{for} \quad t = 2, \ldots, n.$$

Therefore, using the same idea in page 311, we calculate: $\sigma_{e_t}^2 = \dfrac{\sigma_\epsilon^2}{1-\rho^2} = \dfrac{1}{1-0.412^2}$ and similary the first order ACF is:

$$Corr(e_t, e_{t-1}) = \frac{E(e_t e_{t-1})}{\sqrt{\sigma_{e_t}^2 \sigma_{e_{t-1}}^2}} = \rho = 0.412$$

.