TEXAS A&M UNIVERSITY

DEPARTMENT OF STATISTICS

STAT 608 - Regression Analysis
Homework VII

Salih Kilicli

July 20, 2019

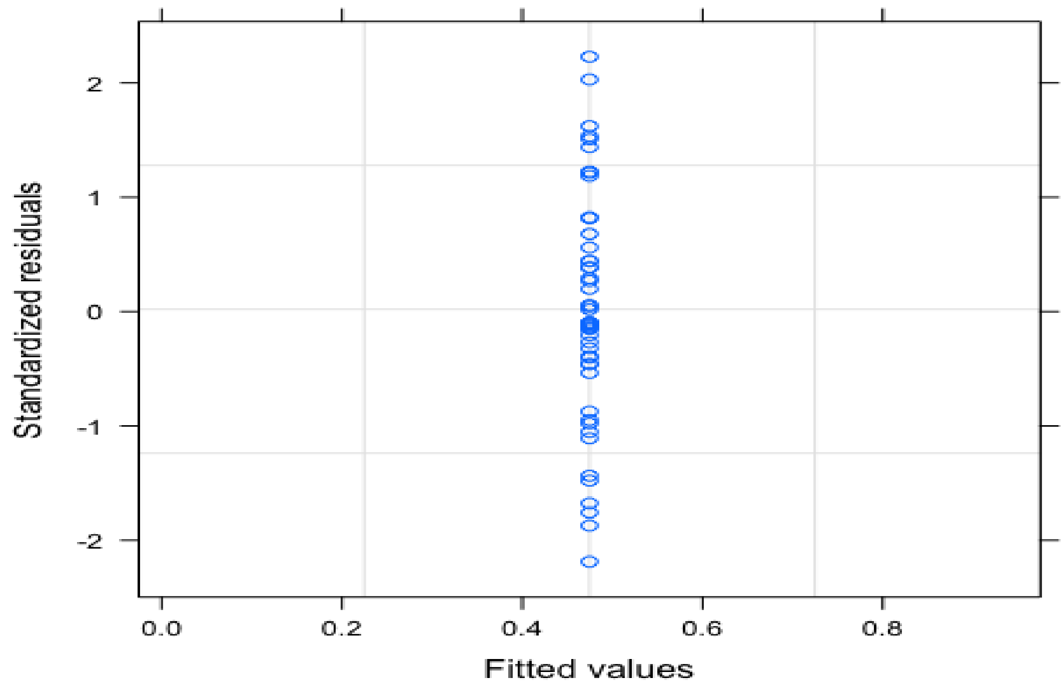
Question 1:

Solution: (1.1) Testing Null Hypothesis ignoring the possibility of serial correlation, we fail to reject it since the p-value of the model is given by $p = 0.128 > 0.05$. The summary of the model is given below:

```
Call:
lm(formula = Q1[, 1] ~ 1)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6827 -0.9847 -0.2116  1.1366  4.7700

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.4762     0.3074   1.549    0.128

Residual standard error: 2.174 on 49 degrees of freedom
```

(1.2) Testing Null Hypothesis after checking for correlation we reject it since the p-value of the model is given by $p = 0.007 < 0.05$. The summary and plot of the model is given below:

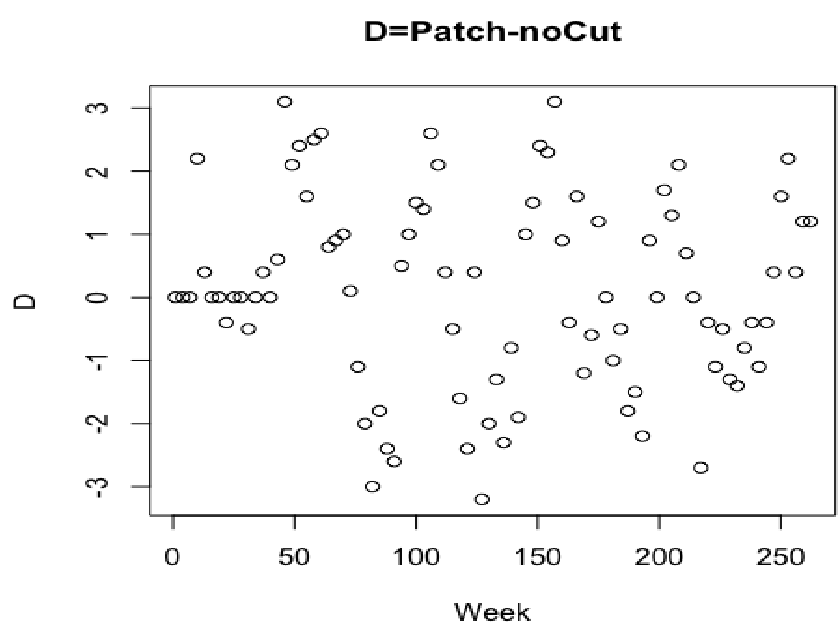
```
Generalized least squares fit by maximum likelihood
  Model: x ~ 1
  Data: Q1
       AIC      BIC    logLik
  207.9588 213.6949 -100.9794

Correlation Structure: AR(1)
 Formula: ~day
 Parameter estimate(s):
       Phi
-0.5279798

Coefficients:
                Value Std.Error  t-value p-value
(Intercept) 0.4746875 0.1710985 2.774352  0.0078

Standardized residuals:
        Min          Q1         Med          Q3         Max
-2.18754096 -0.45944270 -0.09817887  0.53185051  2.22976090

Residual standard error: 2.139934
Degrees of freedom: 50 total; 49 residual
```



Question 2:

Solution :

(2.1) The plot of the series of D-values is given below.



Since 1 year consists of approximately 52 weeks. 2 years is 104 weeks, 3 years is 156 weeks, 4 years is 208 weeks and 5 years is 260 weeks etc. The plot therefore shows a periodicity of approximately 1 year.

(2.2) First, ignoring the auto-correlation I fitted the model, and the summary is given below. Coefficient estimates without AR(1) adjustment are:

$$\hat{\beta}_0 = 0.10455,\ \hat{\beta}_1 = 0.09822,\ \hat{\beta}_2 = 1.39739$$

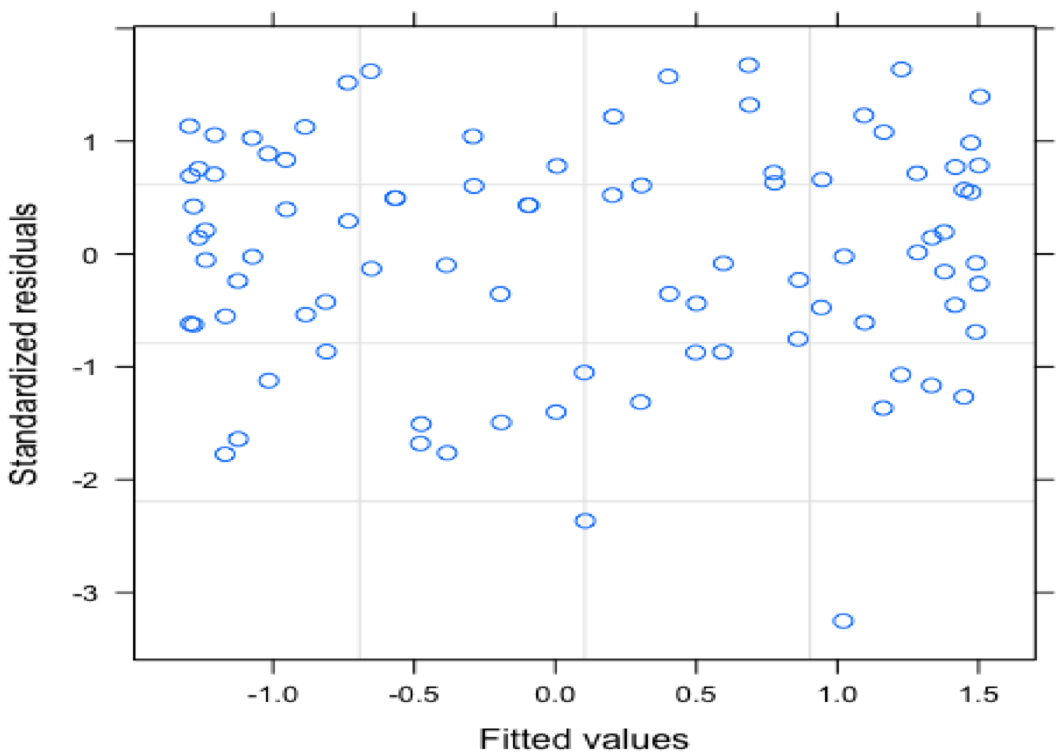
.

```
Generalized least squares fit by maximum likelihood
  Model: D ~ x1 + x2
  Data: Q1
       AIC      BIC    logLik
  283.4914 295.8781 -136.7457

Correlation Structure: ARMA(1,0)
 Formula: ~Week
 Parameter estimate(s):
Phi1
   0

Coefficients:
                Value Std.Error  t-value p-value
(Intercept) 0.1045455 0.1241412 0.842150  0.4021
x1          0.0982249 0.1755622 0.559488  0.5773
x2          1.3973928 0.1755622 7.959532  0.0000

 Correlation:
   (Intr) x1
x1 0
x2 0      0

Standardized residuals:
        Min          Q1         Med          Q3         Max
-3.25078410 -0.62027157  0.07894603  0.72972181  1.67325607

Residual standard error: 1.144525
Degrees of freedom: 88 total; 85 residual
```

After adjustment of AR(1), the new fitted model gives parameters estimate $\phi = 0.3996349$. The summary of AR(1) model given below:

(2.3) Looking at the summary and the plot given below, the model looks valid, since Std residuals vs Fitted Values shows a random pattern and the p-values aren't so big.

```
Generalized least squares fit by maximum likelihood
  Model: D ~ x1 + x2
  Data: Q2
       AIC      BIC    logLik
  268.3022 280.6889 -129.1511

Correlation Structure: AR(1)
 Formula: ~1:88
 Parameter estimate(s):
      Phi
0.3996349

Coefficients:
                Value Std.Error  t-value p-value
(Intercept) 0.0918267 0.1881097 0.488155  0.6267
x1          0.1005348 0.2511485 0.400300  0.6899
x2          1.3746588 0.2479961 5.543067  0.0000

 Correlation:
   (Intr) x1
x1 -0.003
x2 -0.019 -0.005

Standardized residuals:
        Min          Q1          Med           Q3          Max
-3.23031329 -0.62898246  0.08931067  0.74240666  1.68990576

Residual standard error: 1.144196
Degrees of freedom: 88 total; 85 residual
```



Question 3: **Exercise 8.3.1, page 294, in the textbook. (Note: ”Traditional” linear regression methodology won’t get you anywhere with this question, but a logistic regression approach will.)**

Solution: (3.1) The author used the traditional least squares model regression on a binomial response variable. The validity of this model is therefore threatened by the fact that the mean of y, $E[y] = p(x)$, therefore the variance of the response variable y, $Var(y) = p(x)(1 - p(x))$ is not constant (depends on x, predictor variable) and is unknown.

(3.2) Fitting a logistic regression model to data we see that there is a strong evidence of a relationship between Y and x, since the estimates have highly significant coefficients. The summary of the model is given below:

```
Call:
glm(formula = cbind(playoff, noplayoff) ~ x, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4876  -2.0968  -0.4703   1.0666   5.3057

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.45843    0.21102  -6.911  4.8e-12 ***
x            0.07807    0.02751   2.838  0.00455 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 124.10  on 29  degrees of freedom
Residual deviance: 116.22  on 28  degrees of freedom
AIC: 170.33

Number of Fisher Scoring iterations: 4
```

Question 4:

Solution: 4.1 $log\Big(\dfrac{p(3)}{1-p(3)}\Big) = -2.643 + 0.674 \times 3 = -0.621$ implies;

$$\Big(\frac{p(3)}{1-p(3)}\Big) = exp(-0.621) = 0.5374068$$

Solving the equation given above for p(3) we get:

$$p(3) = \frac{0.5374068}{1+0.5374068} = 0.3495541 \approx 0.35 = \frac{35}{100}$$

Therefore for 200 insect exposed to the dosage level x=3, $200 \times p(3) = 70$ insects are expected to die. Consequently, 130 of them will survive.

4.2 $\dfrac{p(x)}{1-p(x)} = e^{-2.643+0.674(x+1)} = e^{-2.643+0674x}e^{0.674} = e^{-2.643+0674x}\phi$

Therefore, $\phi = e^{\hat{\beta}_1} = e^{0.674} = 1.96207$ is the estimated factor.

4.3 A 90% confidence interval for β_1 is given by:

$$\hat{\beta}_1 \pm z_{0.95} se(\hat{\beta}_1) = 0.674 \pm 1.644854(0.039) = (0.6098507, 0.7381493)$$

Therefore, a 90% confidence interval for ϕ can be found as:

$$(e^{0.6098507}, e^{0.7381493}) = (1.840157, 2.09206)$$

Question 5:

Solution: (5.1) $V_i = arcsin(\sqrt{z_i}), \quad z_i = \dfrac{y_i}{m_i}$ where $E[y_i] = m_i\theta_i, \quad Var(y_i) = m_i\theta_i(1-\theta_i)$. Therefore we have:

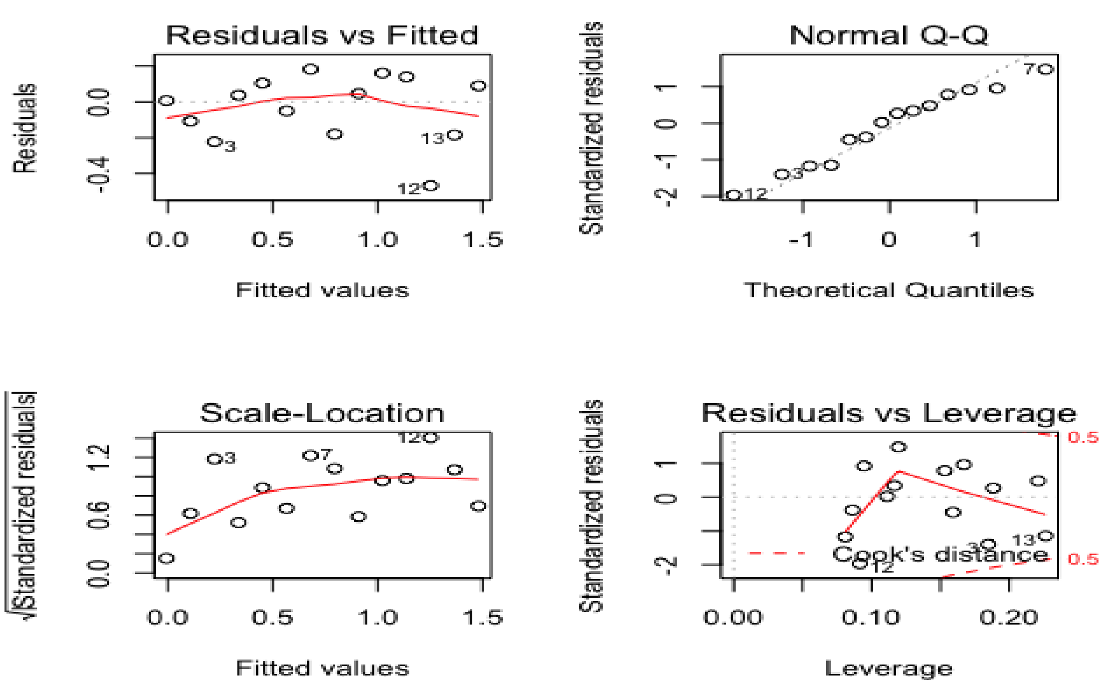
$$E[z_i] = \theta_i, \quad Var(z_i) = \frac{\theta_i(1-\theta_i)}{m_i}$$

Now, from section 3.3.1 we have: (Note that, $dV_i/dz_i = (2\sqrt{z_i(1-z_i)})^{-1}$)

$$\begin{aligned} Var(V_i) &\approx \Big[dV_i/dz_i(E[z_i])\Big]^2 Var(z_i) \\ &\approx \Big[\frac{1}{2\sqrt{E[z_i](1-E[z_i])}}\Big]^2 Var(z_i) \\ &\approx \frac{1}{4\theta_i(1-\theta_i)}\frac{\theta_i(1-\theta_i)}{m_i} \\ &\approx \frac{1}{4m_i} \end{aligned}$$

(5.2) $\hat{\gamma}_0 = -1.72348, \ \hat{\gamma}_1 = 0.11447, \ S^2 = (0.2967)^2 = 0.8804$

(5.3) All of the estimates are highly significant, $R^2 = 0.8563$, the F-statistic of the model is 71.5 with a highly significant $p = 2.122e{-}06$ value. o, the model is valid; however, the Residual vs Fitted and scale-location plots show slightly increasing variance and non-linear relationship, so model definitely can be improved by other means.



(5.4) 90% prediction interval for V_i* is:

$$V_i* = sin^{-1}(\sqrt{z_i*}) \in (0.145305, 1.215334) \quad \text{fitted on } 0.683195.$$

Therefore, 90% prediction interval for z_i* can be found as:

$$\begin{aligned} z_i* = sin^2(V_i*) &\in \Big(sin^2(0.145305),\ sin^2(1.215334)\Big) \\ &\in (0.0203241,\ 0.8788794) \end{aligned}$$

Question 6:

Solution: (a) Using logistic regression we get the parameter estimates as below:

$$\hat{\beta}_0 = 15.07,\ \hat{\beta}_1 = -113.63,\ \hat{\beta}_2 = 255.29,\ \hat{\beta}_3 = -183.14$$

(b) There are two ways to answer this question. First, simply looking at Wald p-value of $\hat{\beta}_3$ given in the summary $0.00316 < 0.1$ we see that the coefficient is significant therefore we reject the Null Hypothesis. Alternatively, using the difference of deviances and calculating p-value from χ^2 distribution with 1 dof we get:

$$P\{G^2_{H_0} - G^2_{H_A} > 11.68\} = 0.0006333219 < 0.1$$

Therefore, we reject the Null Hypothesis.