4.28 The file *wa-wheat.dat* contains observations on wheat yield in Western Australian shires. There are 48 annual observations for the years 1950–1997. For the Northampton shire, consider the following four equations:

$$YIELD_t = \beta_0 + \beta_1 TIME + e_t$$
$$YIELD_t = \alpha_0 + \alpha_1 \ln(TIME) + e_t$$
$$YIELD_t = \gamma_0 + \gamma_1 TIME^2 + e_t$$
$$\ln(YIELD_t) = \phi_0 + \phi_1 TIME + e_t$$

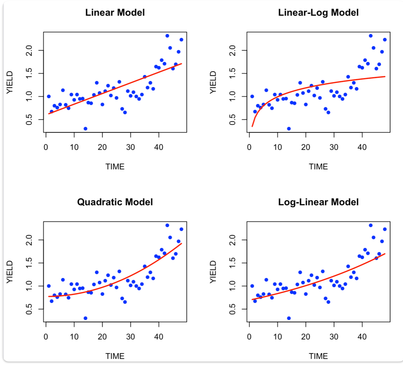
a. Estimate each of the four equations. Taking into consideration (i) plots of the fitted equations, (ii) plots of the residuals, (iii) error normality tests, and (iii) values for R^2, which equation do you think is preferable? Explain.

b. Interpret the coefficient of the time-related variable in your chosen specification.

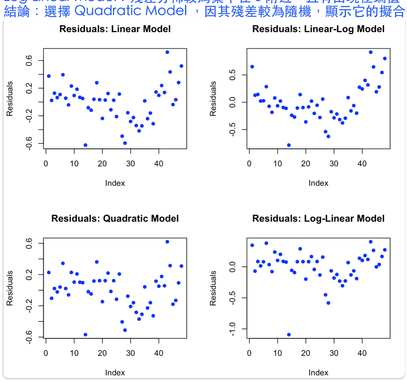
c. Using your chosen specification, identify any unusual observations, based on the studentized residuals, *LEVERAGE*, *DFBETAS*, and *DFFITS*.

d. Using your chosen specification, use the observations up to 1996 to estimate the model. Construct a 95% prediction interval for *YIELD* in 1997. Does your interval contain the true value?

a. (i) 根據擬合圖，
Linear Model：無法捕捉資料在後期的加速增長趨勢。
Linear-Log Model：無法捕捉資料在後期的加速增長趨勢。
Quadratic Model：能夠捕捉小麥產量的加速增長，適合當 YIELD 呈現曲線上升的情況。
Log-Linear Model：雖然模型能夠描述非線性增長，但與二次模型相比，它的增長速度相對較為平滑，無法完全捕捉 YIELD 的加速變化。
結論：選擇 Quadratic Model ，因其最能貼合數據趨勢，能捕捉 YIELD 增長的加速度變化。



(ii) 根據殘差圖
Linear Model：殘差呈現輕微的曲線模式
Linear-Log Model：殘差呈現明顯的 U 型模式
Quadratic Model：殘差分佈較為隨機，沒有明顯的模式
Log-Linear Model：殘差分佈較為集中在 0 附近，且有出現極端值
結論：選擇 Quadratic Model ，因其殘差較為隨機，顯示它的擬合效果較好。



(iii) 根據 Jarque-Bera Test ，若 p-value > 0.05，表示它們的殘差可能符合常態性。
結論：選擇 Linear Model ，因其 p 值最高，表示殘差高度符合常態分布，模型滿足常態假設的程度。

```
> cat("Jarque-Bera Normality Test (p-values):\n")
Jarque-Bera Normality Test (p-values):
> print(jb_values)
          Model JB_p_value
1        Linear  0.9258650
2 Linear-Log    0.2512080
3  Quadratic    0.8504138
4 Log-Linear    0.0000000
```

(iv) 根據 R^2 值，選擇 Quadratic Model ，因其 R^2 值最高，表示該模型對 YIELD 變化的解釋能力最強。

```
> cat("R² Values:\n")
R² Values:
> print(r2_values)
       Model        R2
1     Linear 0.5778369
2 Linear-Log 0.3385733
3  Quadratic 0.6890101
4 Log-Linear 0.5073566
```

b. $TIME^2$ 的係數為 0.0004986，且為正數，表示 YIELD 隨著時間的推移呈加速成長，而非線性增加。

$TIME^2$ 係數的 t 值 = 10.10，表示該變數在回歸中極為顯著。

P-value = 3.01e-13，遠小於 0.05，表示在 99.9% 信心水準下，可以拒絕「$TIME^2$ 係數為 0」的虛無假設。

$TIME^2$ 變數在該模型中極為顯著，顯示其對 YIELD 的影響極為強烈。

```
> summary(model3)

Call:
lm(formula = YIELD ~ TIME2, data = northampton_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.56899 -0.14970  0.03119  0.12176  0.62049

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.737e-01  5.222e-02   14.82  < 2e-16 ***
TIME2       4.986e-04  4.939e-05   10.10 3.01e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2396 on 46 degrees of freedom
Multiple R-squared:  0.689,     Adjusted R-squared:  0.6822
F-statistic: 101.9 on 1 and 46 DF,  p-value: 3.008e-13
```

c.
- Studentized Residuals (學生化殘差)：
若 | Studentized Residual | >2，則該觀測值可能是 異常值 (Outlier)。
觀測值 14、28、43 超過 2，可能是異常值。
- Leverage (槓桿值，影響回歸的程度)：
判斷標準：h_bar > k/n。（其中 k 是變數數量，n 是樣本數）
觀測值 45、46、47、48 的槓桿值較高，可能對回歸影響較大。
- DFBETAS (單個變數影響回歸係數的變化程度)：
判斷標準：| DFBETAS | >2/√n 。
觀測值 14、43、44、48 對 TIME2 變數影響顯著。
- DFFITS (刪除該觀測值後，模型擬合值的變化程度)：
判斷標準：| DFFITS | >2√(k/n) 。（其中 k 是變數數量，n 是樣本數）
觀測值 14、43、48 影響整體回歸擬合結果。

```
> # 顯示異常學生化殘差的觀測點
> print(abnormal_resid_points)
   Observation Studentized_Residual Outlier_Residual
14          14            -2.560682              YES
28          28            -2.246847              YES
43          43             2.889447              YES
```

```
> cat("h_bar:", h_bar, "\n")
h_bar: 0.04166667
> # 印出異常槓桿值的觀測點
> print(abnormal_leverage_points)
   Observation   Leverage High_Leverage_2x
45          45 0.08542511              YES
46          46 0.09531255              YES
47          47 0.10614453              YES
48          48 0.11796846              YES
```

```
> # 顯示異常 DFBETAS 觀測點
> print(abnormal_dfbetas_points)
   Observation DFBETAS_TIME2 Influential_DFBETAS
14          14     0.3205200                 YES
43          43     0.6521798                 YES
44          44     0.3383169                 YES
48          48     0.4607666                 YES
```

```
> # 顯示異常 DFFITS 觀測點
> print(abnormal_dffits_points)
   Observation      DFFITS High_DFFITS_2x
14          14 -0.4944002              YES
43          43  0.7823199              YES
48          48  0.5077802              YES
```

d. 95% 預測區間 = [1.372403, 2.389819]
1997 年的真實值 =
1.372403 ⩽ 2.2318 ⩽ 2.389819
1997 年的真實值落在 95% 預測區間內，表示模型對該年度的預測是合理的。

```
> # 計算 95% 預測區間
> prediction_1997 <- predict(model_restricted, newdata = new_data_1997, interval =
"prediction", level = 0.95)
> print(prediction_1997)  # 顯示預測結果
       fit      lwr      upr
1 1.881111 1.372403 2.389819
>
> # 檢查 1997 年的真實值是否落在預測區間內
> true_value_1997 <- northampton_data$YIELD[northampton_data$TIME == 48]
> cat("1997 年的真實值:", true_value_1997, "\n")
1997 年的真實值: 2.2318
```