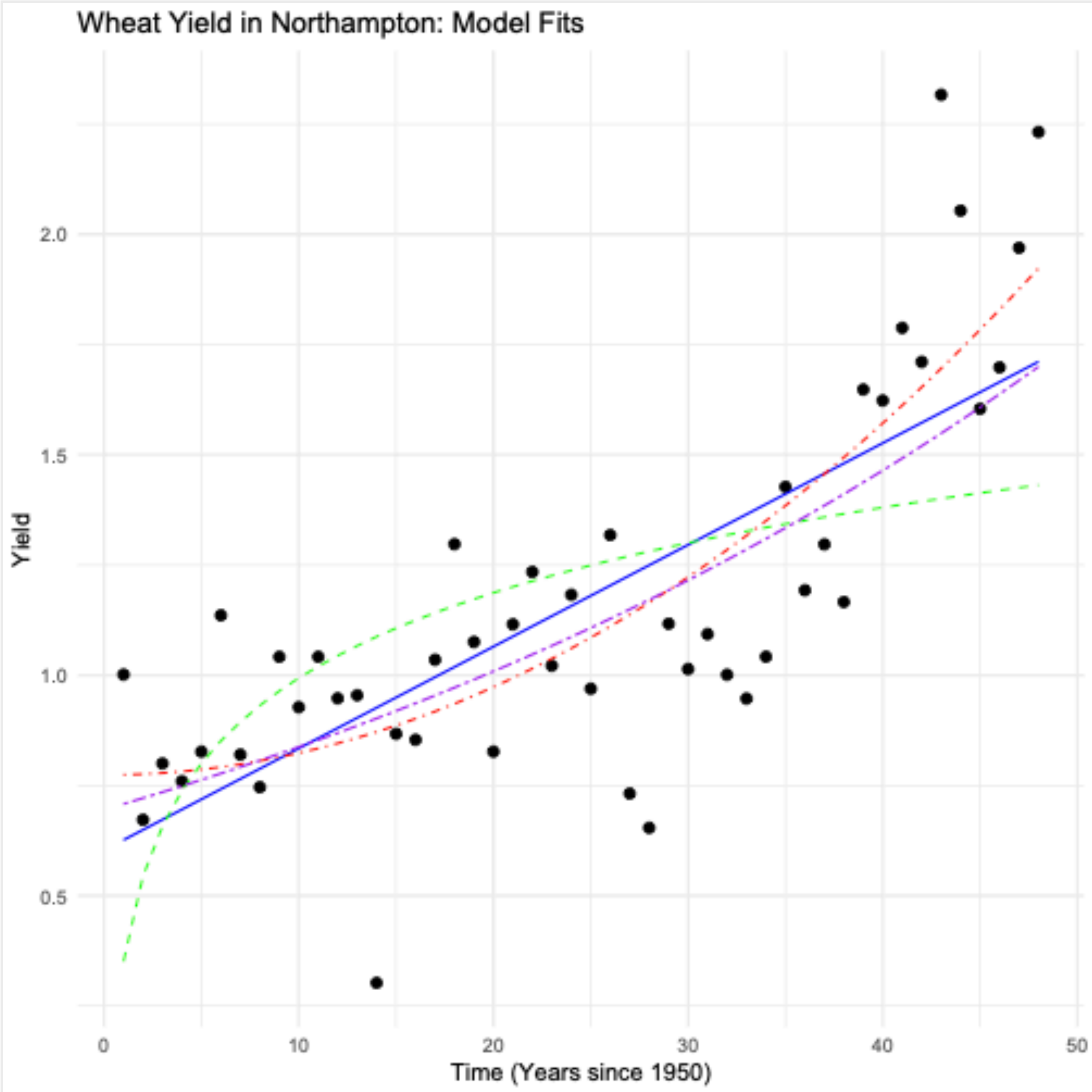
HW: week 4
Question 28 a

These were the results of the code which can be found in the respective R data file.

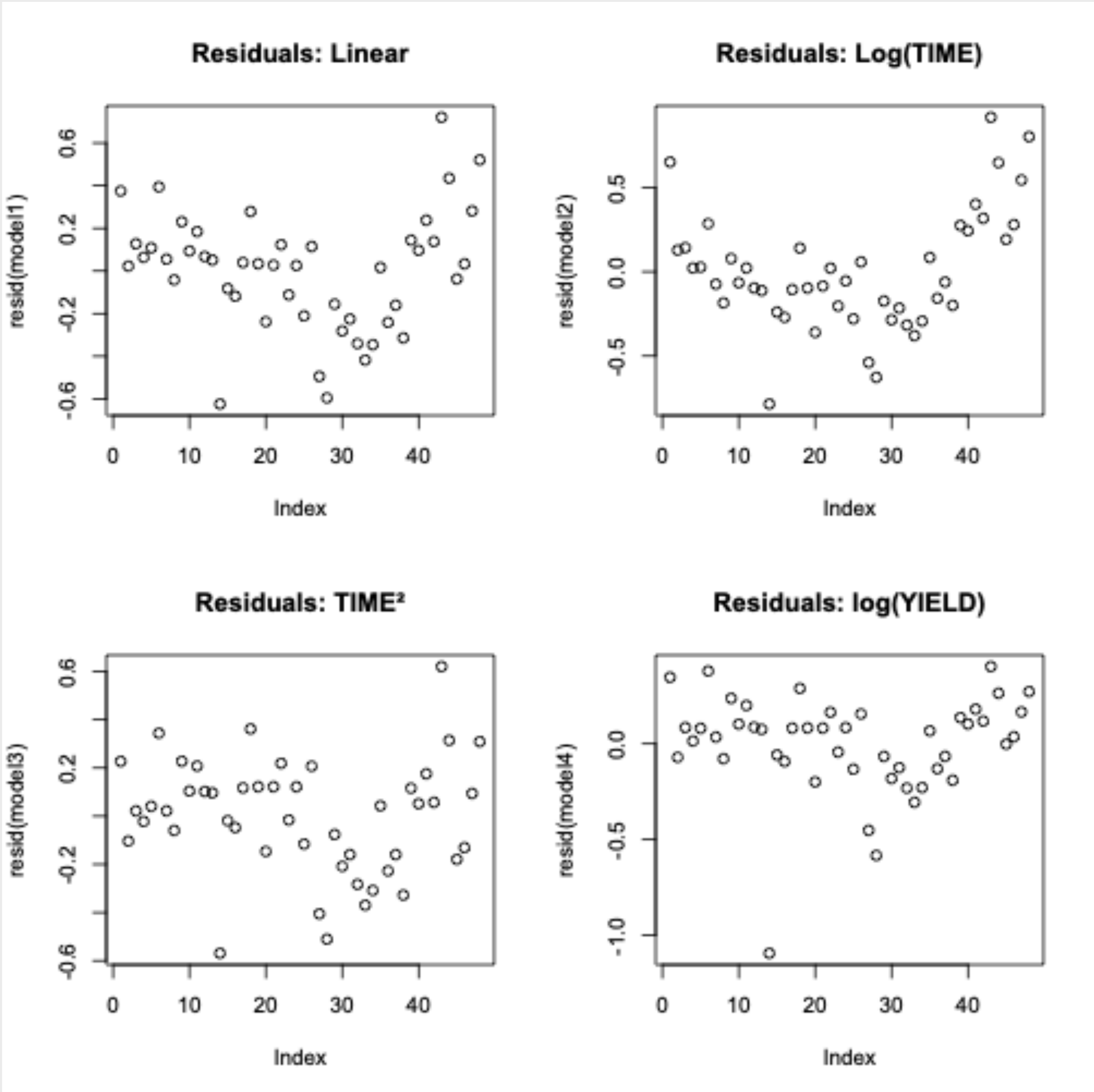
I think that the Quadratic model has the best, because first of all the **quadratic model** most accurately follows the curvature in the yield data. The **log(TIME)** and **log(YIELD)** models underfit or distort the actual yield pattern.

Model	R^2	Shapiro-Wilko p-value	Residuals	Fit Curve
1. Linear	0.578	0.679	Moderate pattern	Decent
2. Log(TIME)	0.339	0.186	Slight curvature remains	Poor fit
3. $TIME^2$ (Quadratic)	**0.689**	0.827	Most random residuals	Best fit
4. log(YIELD)	0.507	0.000072	Heteroskedastic + skewed	Unreliable



Looking at the residuals:

- The **quadratic model's residuals** are the most randomly distributed with minimal structure.
- The **log(YIELD)** model has residuals that clearly show **non-normality and heteroskedasticity**.
- inear and log(TIME) models show some mild patterns, suggesting misspecification.



In term of the normality test we see that:

Model	W Statistic	p-value
1. Linear	0.982	0.679
2. Log(TIME)	0.967	0.186
3. TIME² (Quadratic)	**0.986**	**0.827**
4. log(YIELD)	0.869	0.000072

Out of the for models only the log model has a p-value less than 0.05, in other words the residuals are not normally distributed.

In terms of R-squares we see that:

```
summary(model1)$r.squared
[1] 0.5778369
```

summary(model2)$r.squared
[1] 0.3385733
summary(model3)$r.squared —> this being the Quadratic model and one which seems to have the highest R^2
[1] 0.6890101
summary(model4)$r.squared
[1] 0.5073566

Question 28 b

I used this code to fit a quadratic regression model: model3 <- lm(YIELD ~ I(TIME^2), *data* = df)

After I summarised the model I was left with this output:

```
lm(formula = YIELD ~ I(TIME^2), data = df)

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 ***
I(TIME^2)   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
```

Which can be simplified into: **YIELD**t = 0.7737 + 0.0004986 * TIME^2 + e_t

From which we get: **Intercept** = 0.7737 and **TIME coefficient**: = 0.0004986

This is means that the **wheat yield** at TIME = 0, or in this case 1950, is equivalent to 0.7737. The the TIME coefficient show the exponential growth rate of the wheat yield of 0.0004986. In other words for every increase in T + 1 the wheat yield experiences a corresponding increase of 0.0004986.