The Chapman Valley Study

Student Project by: Winnie Kwan
Course: VEE – Regression Analysis
Semester: Winter 2008
Submitted on: June 6, 2010

Introduction

Wheat yield in Western Australia have been found to depend on the period of observation. The goal of this project is to select a model based on time that can be used to represent past or predict future wheat yield.

The Chapman Valley Shire is located in the Mid West of Western Australia and will be the subject of our analysis.

We will estimate three regression equations, then evaluate and select the most preferable by considering a few evaluation methods.

Data and Method

Our data consists of 48 annual observations for the years 1950 to 1997 for the Chapman Valley Shire. This data provides for the wheat yield over time, where time is measured from 1 through 48.

Our data analysis will be conducted using Excel.

Model Selection Procedures

We will determine a reqression equation for the wheat yield in Chapman Valley Shire based on time. The logical starting point is to consider a simple linear regression model of the data. We expect that there is a relation between time and wheat yield, but we may also suspect that this relationship is not a straight line. We will consider two other transformations of the simple linear regression model: a linear log model and a squared model. Determining fitted equations of these three models will allow us to further examine their predictability power.

Our model selection can depend on a number of criteria. For our project, we will consider three evaluation methods: plots of the fitted equations, plots of residuals and values of R^2, which is a measure of goodness-of-fit.

Estimating the Regression Equations

We will first consider a simple linear regression for the Chapman Valley shire. Our goal is to estimate the equation:

$$y_t = \beta_0 + \beta_1 t + e_t \qquad \text{(Eq 1)}$$

Using Excel, a regression analysis of time on wheat yield gives us the following summary statistics:

Regression Statistics	
Multiple R	0.677853518
R Square	0.459485393
Adjusted R Square	0.447735075
Standard Error	0.247323851
Observations	48

ANOVA

	Df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2.391961064	2.391961064	39.10408297	1.20657E-07
Residual	46	2.813778015	0.061169087		
Total	47	5.20573908			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.67759539	0.0725266	9.342715522	3.37934E-12
Time	0.016113879	0.002576849	6.253325752	1.20657E-07

	Lower 95%	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.531606919	0.823583861	0.531606919	0.823583861
Time	0.01092695	0.021300807	0.01092695	0.021300807

Table 1 – Summary statistics of the simple linear equation (Eq 1)

The parameter estimates are:

$$b_0 = 0.6776$$
$$b_1 = 0.0161$$

The estimated simple linear regression equation, R^2 and standard errors (given in parenthesis) for the Chapman Valley shire is:

$$\hat{y}_t = 0.6776 + 0.0161\, t \qquad R^2 = 0.4595$$
$$\text{(se)} \quad (0.0725) \quad (0.0026)$$

* * * * *

Turning to the linear-log model, we will estimate the equation:

$$y_t = \alpha_0 + \alpha_1 \ln(t) + e_t \qquad \text{(Eq 2)}$$

The regression analysis of log time on wheat yield gives us the following summary statistics:

Regression Statistics	
Multiple R	0.49410133
R Square	0.244136124
Adjusted R Square	0.227704301
Standard Error	0.292471854
Observations	48

ANOVA

	Df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.270908964	1.270908964	14.85751878	0.00035807
Residual	46	3.934830116	0.085539785		
Total	47	5.20573908			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.528697478	0.147232805	3.590894571	0.0007977
Time_ln	0.185514276	0.048128707	3.854545211	0.00035807

	Lower 95%	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.232333219	0.825061738	0.232333219	0.825061738
Time	0.088636215	0.282392336	0.088636215	0.282392336

Table 2 – Summary statistics of the linear-log equation (Eq 2)

The parameter estimates are:

$$a_0 = 0.5287$$
$$a_1 = 0.1855$$

The estimated linear-log regression equation, R^2 and standard errors (given in parenthesis) for the Chapman Valley shire is:

$$\hat{y}_t = 0.5287 + 0.1855 \ln(t) \qquad R^2 = 0.2441$$
(se) (0.1472) (0.0481)

* * * * *

Lastly, we will estimate the equation with time raised to the power of 2:

$$y_t = \gamma_0 + \gamma_1 t^2 + e_t \qquad \text{(Eq 3)}$$

The regression analysis of time squared on wheat yield gives us the following summary statistics:

Regression Statistics	
Multiple R	0.754018909
R Square	0.568544516
Adjusted R Square	0.559165049
Standard Error	0.220968455
Observations	48

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2.959694404	2.959694404	60.61586579	6.13369E-10
Residual	46	2.246044675	0.048827058		
Total	47	5.20573908			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.791438513	0.048159982	16.43353001	7.07179E-21
Time_sq	0.000354656	4.55527E-05	7.785619166	6.13369E-10

	Lower 95%	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.694497499	0.888379528	0.694497499	0.888379528
Time	0.000262963	0.000446349	0.000262963	0.000446349

Table 3 – Summary statistics of the power equation (Eq 3)

The parameter estimates are:

$$g_0 = 0.7914$$
$$g_1 = 0.0004$$

The estimated regression equation, R^2 and standard errors (given in parenthesis) of the power model for the Chapman Valley shire is:

$$\hat{y}_t = 0.7914 + 0.0004\, t^2 \qquad R^2 = 0.5685$$
(se) (0.0482) (0.000)

The estimated equations are for the linear, linear-log and power model are again:

$$\hat{y}_t = 0.6776 + 0.0161\, t \qquad \text{(Eq 1)}$$

$$\hat{y}_t = 0.5287 + 0.1855 \ln(t) \qquad \text{(Eq 2)}$$

$$\hat{y}_t = 0.7914 + 0.0004\, t^2 \qquad \text{(Eq 3)}$$

Evaluating a preferable model

To determine the most preferable regression model for wheat yield over time based on data from the Chapman Valley Shire, we will consider (i) plots of the fitted equations, (ii) plots of the residuals, and (iv) values for R^2.

The summary statistics for each of the three equations considered are provided in Tables 1 to 2 from Part (a). In addition, the scatter plots and fitted equations for each of the three equations are provided below in Figures 1 to 3. The blue diamonds represent the 48 actual observed values of time, logarithm of time and time squared of equations 1, 2 and 3 respectively. The pink squares represent the predicted values of the estimated regression equations as given in Part (a).

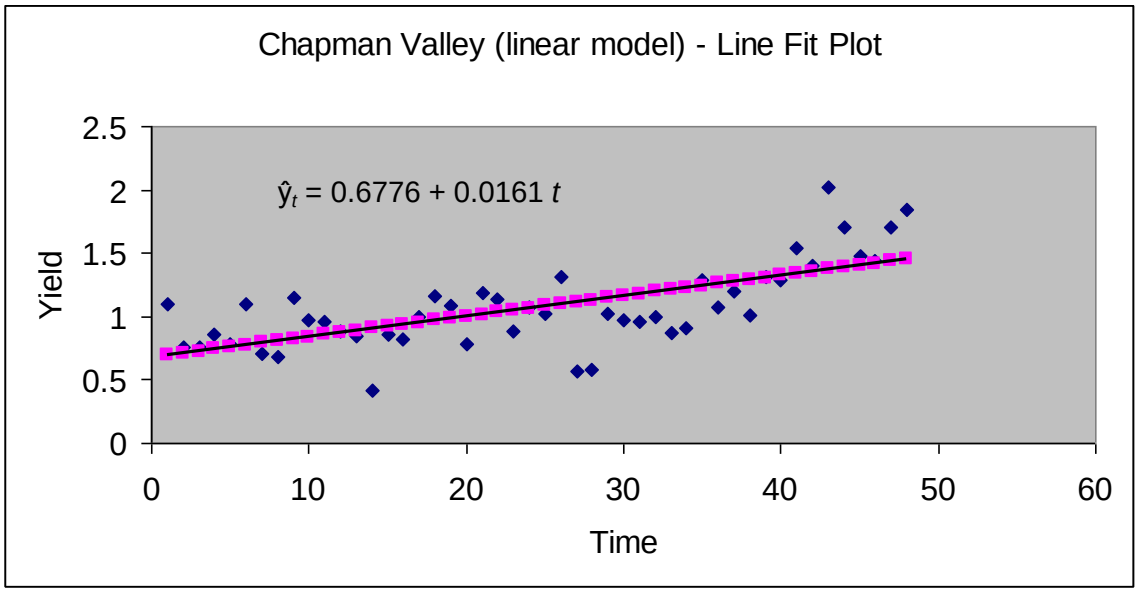

Figure 1 – Scatter plot, predicted values and fitted line of wheat yield over time (Eq 1)

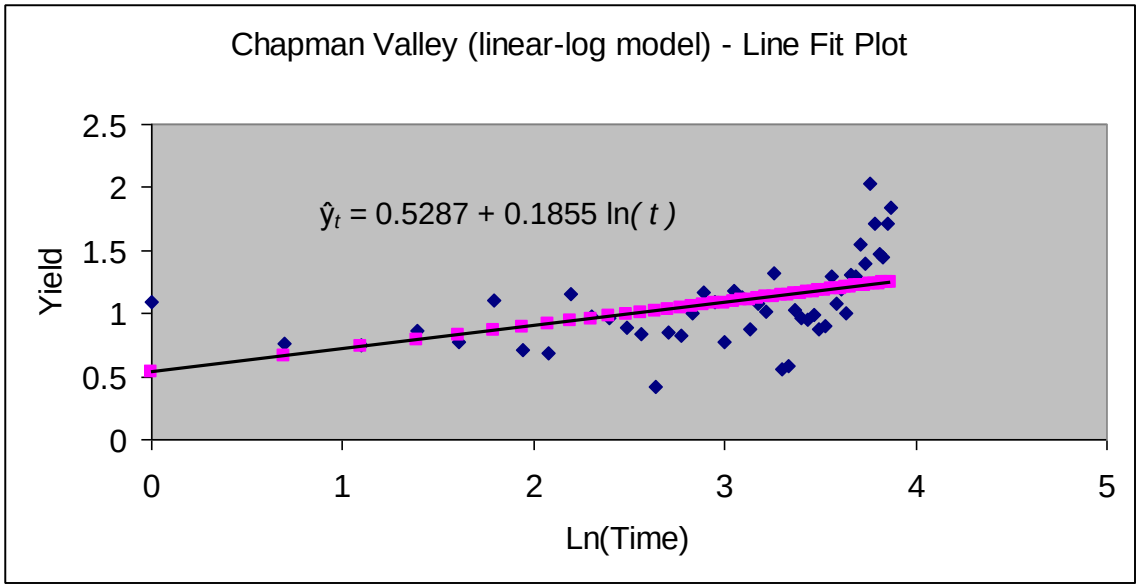

Figure 2 – Scatter plot, predicted values and fitted line of wheat yield over logarithm of time (Eq 2)

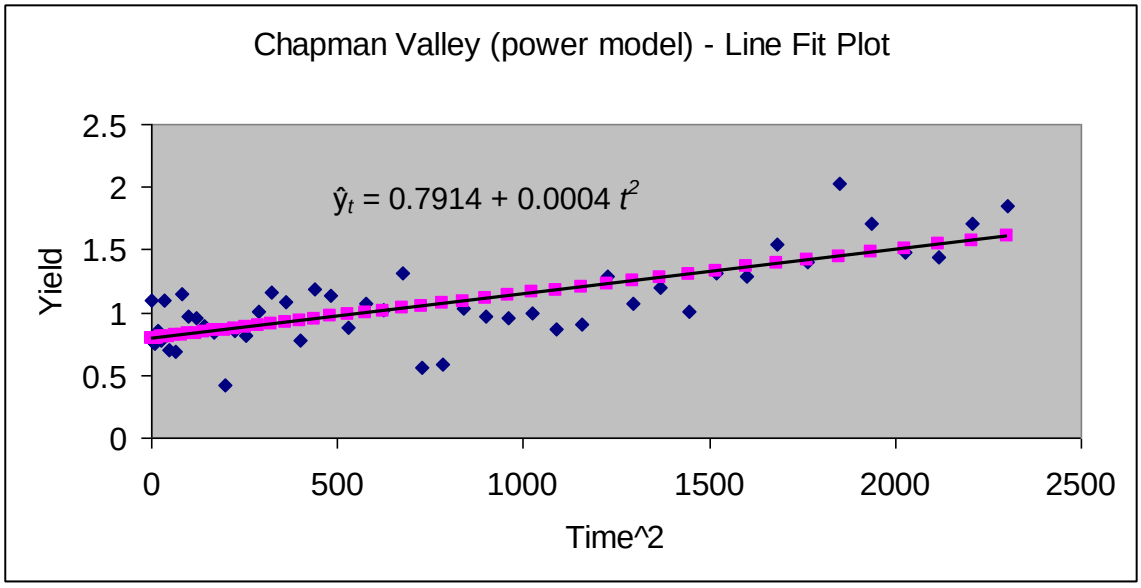

Figure 3 – Scatter plot, predicted values and fitted line of wheat yield over time squared (Eq 3)

Note that the fitted lines in Figures 2 and 3 should not be straight lines, but the curves are not visible on the graphs.

We can also examine the residuals plots of the three equations. The scatter plots and bar charts for each equation are provided in Figures 4 to 9 below.

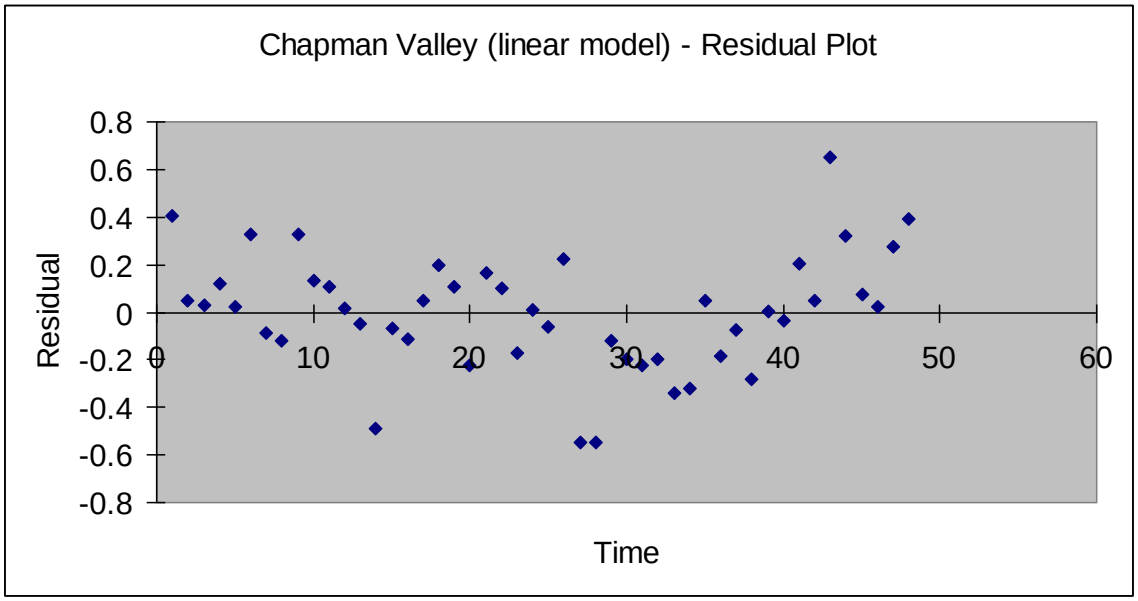

Figure 4 – Scatter plot of residuals from linear equation (Eq 1)

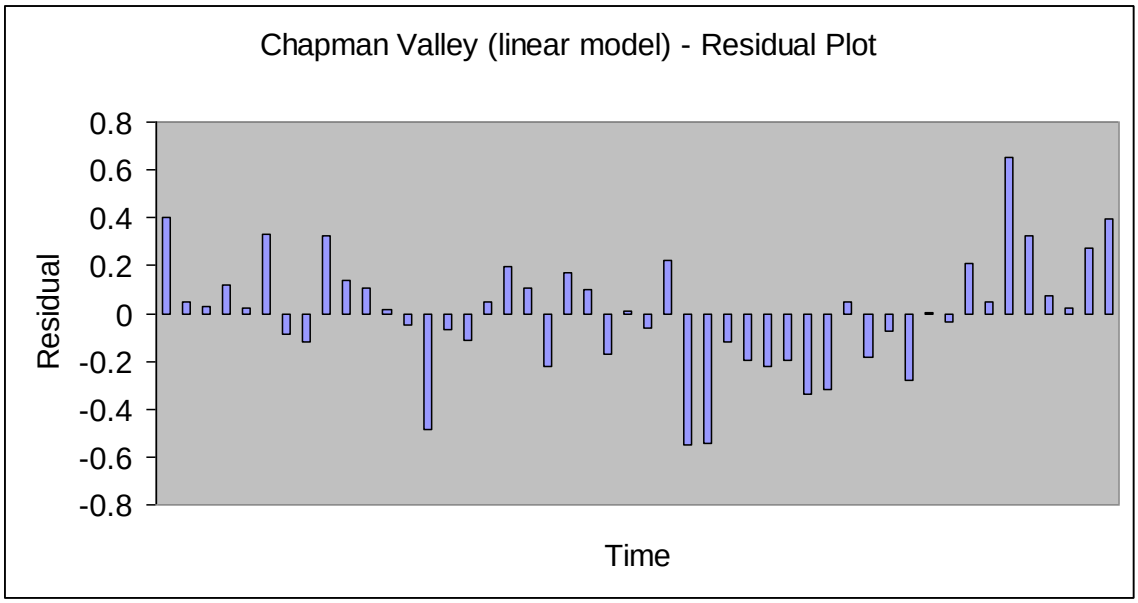

Figure 5 – Bar chart of residuals from linear equation (Eq 1)

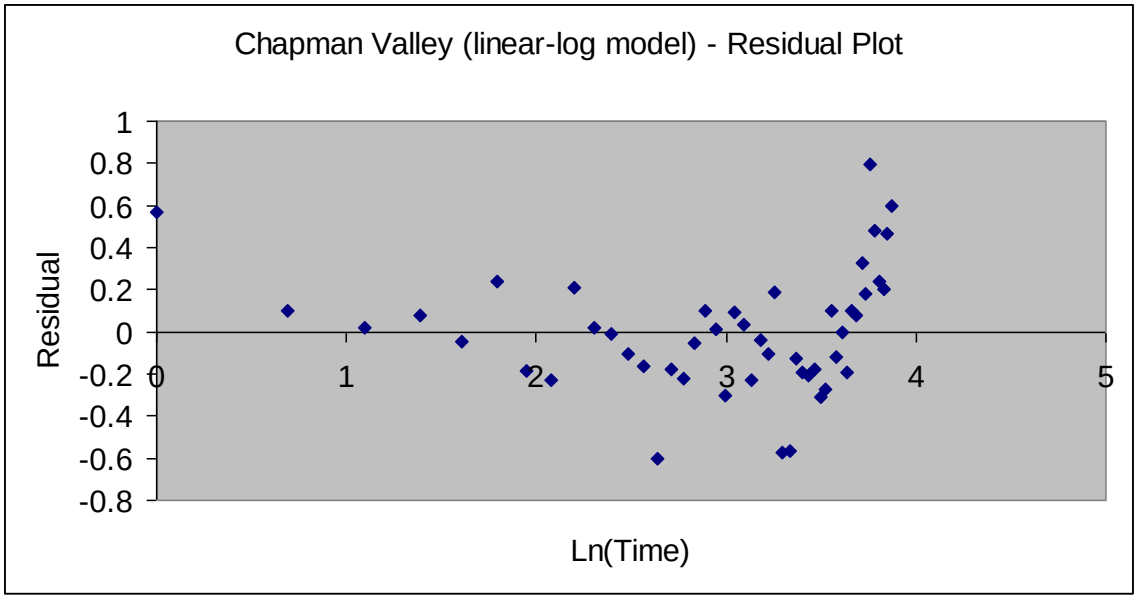

Figure 6 – Scatter plot of residuals from linear-log equation (Eq 2)

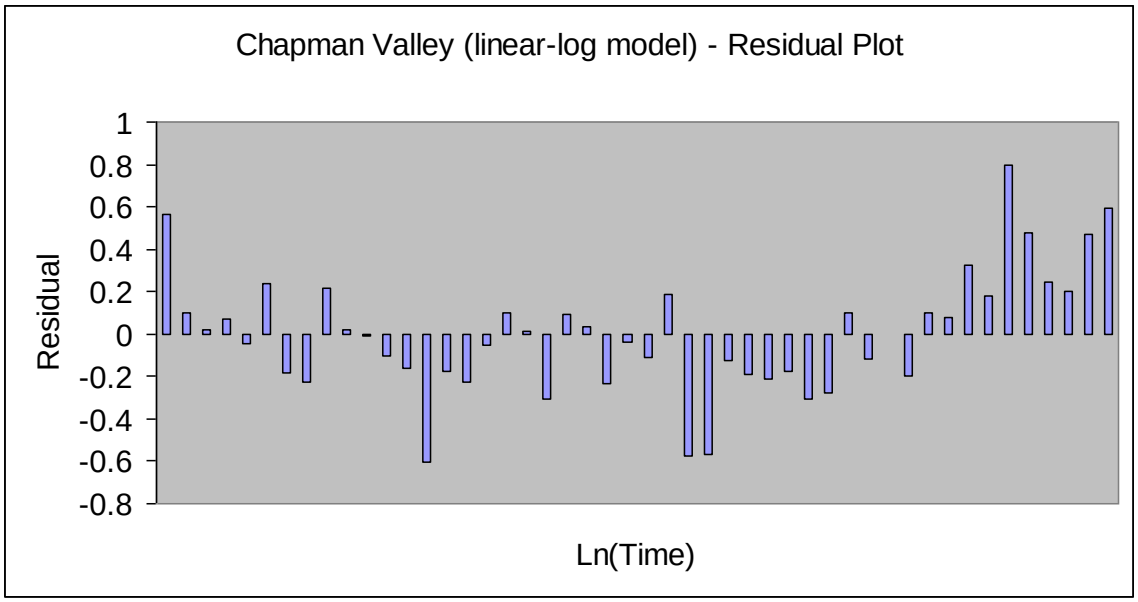

Figure 7 – Bar chart of residuals from linear-log equation (Eq 2)

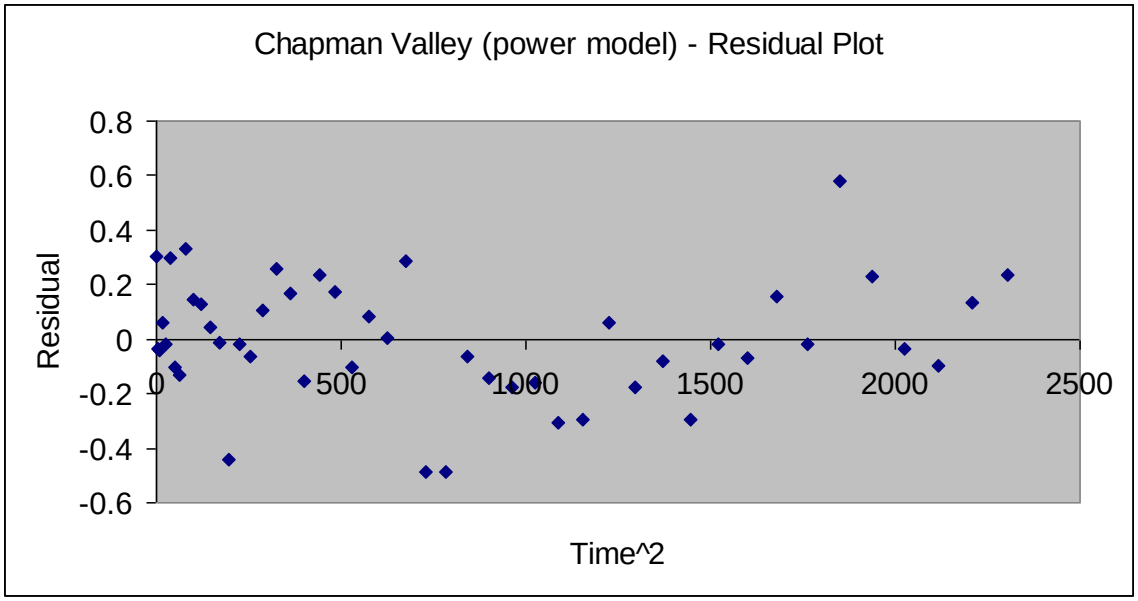

Figure 8 – Scatter plot of residuals from power equation (Eq 3)

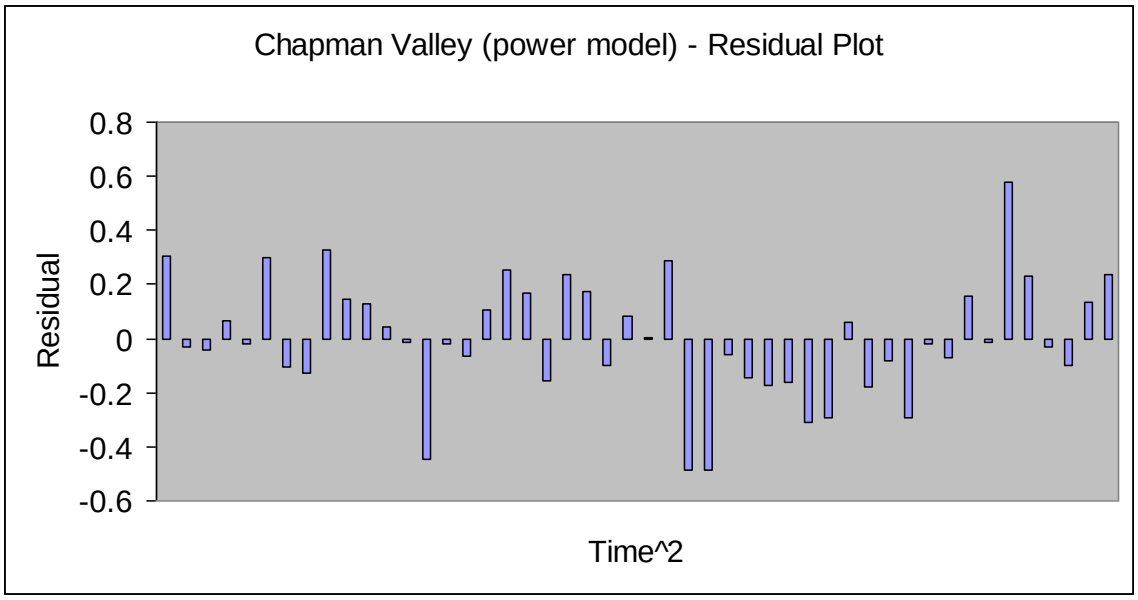

Figure 9 – Bar chart of residuals from power equation (Eq 3)

Analysis

(i) Plots of the fitted equations

The actual values of wheat yield over time, time log and time squared are dispersed around the fitted line of each equation. For the linear regression model, there appears to be a concentration of actual values below the fitted line near the center. This is also true, although less apparent, in Figure 3 for the equation raised to the power of 2. In Figure 3, under careful inspection it can be seen that the predicted values are increasing at an increasing rate and, as such, the slope of the fitted line is positive and increasing. The power model (to the exponent of 2) is consistent with the observed values of wheat yield from time 1 to time 48.

For the linear-log model, it is apparent in Figure 2 that the actual values of wheat yield increases more rapidly for greater values of time log. There is also more dispersion as time log increases. However, the predicted values are displaying the opposite effect, with greater increases lower values of time log and smaller increases for higher values of time log.

This analysis power regression model (Eq 3) is preferred to over the linear and linear-log models. It provides a better functional form. We will continue investigation by analyzing the residuals.

The actual and predicted values of wheat yield over time over the sample time horizon are given in Tables 4 to 6.

(i) Plots of the residuals

For the linear regression equation, it can be seen from Figures 4 and 5 that there is a concentration of positive residuals near the ends of the sample time horizon and negative residuals in the middle. This is caused by the fact that wheat yield is actually increasing at an increasing rate which can not be reflected by a straight line. A transformation of the linear regression should provide a better fit.

Figures 6 and 7 representing the linear-log regression equation display similar results as the linear regression equation. The linear-log model does not provide a good fit for increasing wheat yield at an increasing rate.

Figures 8 and 9 of the power regression equation possess more desirable observations with residuals being evenly dispersed around zero and no apparent pattern of positive or negative residuals clustered around an area.

The residual values of wheat yield over time over the sample time horizon are given in Tables 4 to 6.

(iv) Values for R^2

Referring to Tables 1 to 3 from Part (a), the values of R^2 are 0.4595, 0.2441 and 0.5685 for equations 1, 2 and 3 respectively. This indicates that the time squared (Eq 3) fits the data better than time (Eq 1) which fits the data better than time log (Eq

2). All three equations have the same dependent variable and the same number of explanatory variably (one only) and, as such, R^2 can be used to compare their goodness of fit.

Conclusion
Based on our analysis, the plot of the time squared fitted equation shows a better functional form as it models increasing yield at an increasing rate, consistent with the actual observed values. The residual plots for time squared are also the only plots of the three equations that do not display patterns of positive or negative residuals. Lastly, R^2 is the highest for the time squared fitted equation so that more variation in wheat yield can be explained by the time squared model over the time or time log models. As such, the time squared fitted equation is the most preferable of the three examined.

Interval estimates for coefficients

Our selected equation is the time squared fitted equation. We can find the 95% interval estimates for time squared.

The number of degrees of freedom for the t-statistic is:

$$N - K = 48 - 2 = 46$$

We need to find the value from the $t_{(48)}$-distribution at t_C such that:

$$P(-t_C < t_{(48)} < t_C) = 0.95,$$

where $t_C = t_{(0.975,\, N-K)}$ is the 97.5-percentile.

Using the TINV function in Excel, at 46 degrees of freedom, we find $t_C = 2.0129$.

Solving the interval endpoints for time squared:

$$[g_1 - 2.0129 \text{ x } se(g_1),\ g_1 + 2.0129 \text{ x } se(g_1)]$$
$$[0.0004 - 2.0129 \text{ x } 0.00005,\ 0.0004 + 2.0129 \text{ x } 0.00005]$$

The 95% interval estimate for time squared is given by (the 95% interval estimates for each coefficient are provided in the summary statistics in Table 3 and the equations above are off due to rounding):

$$(0.0003,\ 0.0005)$$

This interval is relatively narrow, so the point estimate $g_2 = 0.0004$ is precise, as its standard error is relatively small.

Similarly, from Table 3 the 95% interval estimate for the intercept can also be found:

$$(0.6945,\ 0.8884)$$

This interval is also narrow, but relatively wider than the 95% interval estimate for time squared.

Economic and statistical interpretation

We will use the statistics from the Excel output in Table 3 to discuss the implication of the regression results of each estimate in the time squared model.

The estimated intercept is 0.7914. We estimate that wheat yield independent of the time that has elapsed is 0.7914. This is the yield that does not vary with time. The corresponding standard error is 0.0482.

The coefficient for time squared is 0.0004. We estimate that as time squared increases, wheat yield will also increase. Since the $0 < 0.0004 < 1$, wheat yield is an increasing function of time squared, but at a decreasing rate. The corresponding standard error is 0.00005.

Closing Remarks

Our analysis is constrained by assumptions and limitations.

Take data for example, could we have used more exhaustive data if available? We expect that there are other determinants that affect of wheat yield in Chapman Valley. Studies show that wheat yield depends on the amount of rainfall during germination, growing and flowering, and therefore would likely be a valuable addition to the model and increase the goodness-to-fit. Multiple regression analysis could have been desirable.

Our project is also a highly simplified study by the choices we decide to conduct the study. We examined three criteria in our model selection, but there could have been other ones chosen. We could also examine various scenarios for hypothesis testing, or consider other transformations such as a log-log model.

In reality, we are constrained by time and resources. There is no single correct fitted model or procedure in which to pursue it.

Appendix

Time	Actual yield	Predicted yield	Residual	Time	Actual yield	Predicted yield	Residual
1.0000	1.0955	0.6937	0.4018	25.0000	1.0182	1.0804	-0.0622
2.0000	0.7595	0.7098	0.0497	26.0000	1.3192	1.0966	0.2226
3.0000	0.7527	0.7259	0.0268	27.0000	0.5640	1.1127	-0.5487
4.0000	0.8603	0.7421	0.1182	28.0000	0.5827	1.1288	-0.5461
5.0000	0.7796	0.7582	0.0214	29.0000	1.0282	1.1449	-0.1167
6.0000	1.1023	0.7743	0.3280	30.0000	0.9662	1.1610	-0.1948
7.0000	0.7057	0.7904	-0.0847	31.0000	0.9568	1.1771	-0.2203
8.0000	0.6855	0.8065	-0.1210	32.0000	0.9945	1.1932	-0.1987
9.0000	1.1493	0.8226	0.3267	33.0000	0.8702	1.2094	-0.3392
10.0000	0.9746	0.8387	0.1359	34.0000	0.9063	1.2255	-0.3192
11.0000	0.9611	0.8548	0.1063	35.0000	1.2883	1.2416	0.0467
12.0000	0.8872	0.8710	0.0162	36.0000	1.0739	1.2577	-0.1838
13.0000	0.8401	0.8871	-0.0470	37.0000	1.1976	1.2738	-0.0762
14.0000	0.4167	0.9032	-0.4865	38.0000	1.0084	1.2899	-0.2815
15.0000	0.8536	0.9193	-0.0657	39.0000	1.3095	1.3060	0.0035
16.0000	0.8200	0.9354	-0.1154	40.0000	1.2885	1.3222	-0.0337
17.0000	1.0014	0.9515	0.0499	41.0000	1.5444	1.3383	0.2061
18.0000	1.1627	0.9676	0.1951	42.0000	1.4005	1.3544	0.0461
19.0000	1.0888	0.9838	0.1050	43.0000	2.0244	1.3705	0.6539
20.0000	0.7796	0.9999	-0.2203	44.0000	1.7095	1.3866	0.3229
21.0000	1.1841	1.0160	0.1681	45.0000	1.4769	1.4027	0.0742
22.0000	1.1344	1.0321	0.1023	46.0000	1.4430	1.4188	0.0242
23.0000	0.8776	1.0482	-0.1706	47.0000	1.7107	1.4349	0.2758
24.0000	1.0768	1.0643	0.0125	48.0000	1.8435	1.4511	0.3924

Table 4 – Actual, predicted and residual values of the simple linear equation (Eq 1)

Time log	Actual yield	Predicted yield	Residual	Time log	Actual yield	Predicted yield	Residual
0.0000	1.0955	0.5287	0.5668	3.2189	1.0182	1.1258	-0.1076
0.6931	0.7595	0.6573	0.1022	3.2581	1.3192	1.1331	0.1861
1.0986	0.7527	0.7325	0.0202	3.2958	0.5640	1.1401	-0.5761
1.3863	0.8603	0.7859	0.0744	3.3322	0.5827	1.1469	-0.5642
1.6094	0.7796	0.8273	-0.0477	3.3673	1.0282	1.1534	-0.1252
1.7918	1.1023	0.8611	0.2412	3.4012	0.9662	1.1597	-0.1935
1.9459	0.7057	0.8897	-0.1840	3.4340	0.9568	1.1658	-0.2090
2.0794	0.6855	0.9145	-0.2290	3.4657	0.9945	1.1716	-0.1771
2.1972	1.1493	0.9363	0.2130	3.4965	0.8702	1.1773	-0.3071
2.3026	0.9746	0.9559	0.0187	3.5264	0.9063	1.1829	-0.2766
2.3979	0.9611	0.9735	-0.0124	3.5553	1.2883	1.1883	0.1000
2.4849	0.8872	0.9897	-0.1025	3.5835	1.0739	1.1935	-0.1196
2.5649	0.8401	1.0045	-0.1644	3.6109	1.1976	1.1986	-0.0010
2.6391	0.4167	1.0183	-0.6016	3.6376	1.0084	1.2035	-0.1951
2.7081	0.8536	1.0311	-0.1775	3.6636	1.3095	1.2083	0.1012
2.7726	0.8200	1.0431	-0.2231	3.6889	1.2885	1.2130	0.0755
2.8332	1.0014	1.0543	-0.0529	3.7136	1.5444	1.2176	0.3268
2.8904	1.1627	1.0649	0.0978	3.7377	1.4005	1.2221	0.1784
2.9444	1.0888	1.0749	0.0139	3.7612	2.0244	1.2265	0.7979
2.9957	0.7796	1.0844	-0.3048	3.7842	1.7095	1.2307	0.4788
3.0445	1.1841	1.0935	0.0906	3.8067	1.4769	1.2349	0.2420
3.0910	1.1344	1.1021	0.0323	3.8286	1.4430	1.2390	0.2040
3.1355	0.8776	1.1104	-0.2328	3.8501	1.7107	1.2430	0.4677
3.1781	1.0768	1.1183	-0.0415	3.8712	1.8435	1.2469	0.5966

Table 5 – Actual, predicted and residual values of the linear-log equation (Eq 2)

Time squared	Actual yield	Predicted yield	Residual	Time squared	Actual yield	Predicted yield	Residual
1.0000	1.0955	0.7918	0.3037	625.0000	1.0182	1.0131	0.0051
4.0000	0.7595	0.7929	-0.0334	676.0000	1.3192	1.0312	0.2880
9.0000	0.7527	0.7946	-0.0419	729.0000	0.5640	1.0500	-0.4860
16.0000	0.8603	0.7971	0.0632	784.0000	0.5827	1.0695	-0.4868
25.0000	0.7796	0.8003	-0.0207	841.0000	1.0282	1.0897	-0.0615
36.0000	1.1023	0.8042	0.2981	900.0000	0.9662	1.1106	-0.1444
49.0000	0.7057	0.8088	-0.1031	961.0000	0.9568	1.1323	-0.1755
64.0000	0.6855	0.8141	-0.1286	1024.0000	0.9945	1.1546	-0.1601
81.0000	1.1493	0.8202	0.3291	1089.0000	0.8702	1.1777	-0.3075
100.0000	0.9746	0.8269	0.1477	1156.0000	0.9063	1.2014	-0.2951
121.0000	0.9611	0.8344	0.1267	1225.0000	1.2883	1.2259	0.0624
144.0000	0.8872	0.8425	0.0447	1296.0000	1.0739	1.2511	-0.1772
169.0000	0.8401	0.8514	-0.0113	1369.0000	1.1976	1.2770	-0.0794
196.0000	0.4167	0.8610	-0.4443	1444.0000	1.0084	1.3036	-0.2952
225.0000	0.8536	0.8712	-0.0176	1521.0000	1.3095	1.3309	-0.0214
256.0000	0.8200	0.8822	-0.0622	1600.0000	1.2885	1.3589	-0.0704
289.0000	1.0014	0.8939	0.1075	1681.0000	1.5444	1.3876	0.1568
324.0000	1.1627	0.9063	0.2564	1764.0000	1.4005	1.4171	-0.0166
361.0000	1.0888	0.9195	0.1693	1849.0000	2.0244	1.4472	0.5772
400.0000	0.7796	0.9333	-0.1537	1936.0000	1.7095	1.4781	0.2314
441.0000	1.1841	0.9478	0.2363	2025.0000	1.4769	1.5096	-0.0327
484.0000	1.1344	0.9631	0.1713	2116.0000	1.4430	1.5419	-0.0989
529.0000	0.8776	0.9791	-0.1015	2209.0000	1.7107	1.5749	0.1358
576.0000	1.0768	0.9957	0.0811	2304.0000	1.8435	1.6086	0.2349

Table 6 – Actual, predicted and residual values of the power of 2 equation (Eq 3)