An Analysis of Credits to Graduation at the University of Utah

by
Jeremy Morris

Supervisory Committee:
Lajos Horváth (Committee Chair)
Gary Levy
Jingyi Zhu

A project submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of
Master of Statistics
Department of Mathematics

Abstract

The purpose of this project is to discuss current trends in graduation data at the University of Utah. Administrators and staff have a need to understand what graduation trends are occuring, this research will serve to specify, using statistical methods, those trends. It will be left up to the staff and administration to interpret these results in the larger context of their knowledge about the institution.

Contents

List of Figures

List of Tables

1 Data Collection and Cleaning

The data used for this analysis was collected from the course and graduation tables maintained by the Office of Budget and Analysis. These tables are extracted from the campus computing system every year in August and include all degrees posted to the system by the graduation office beginning July 1 of the previous year and ending June 30 of the current year. Each graduation file includes one line for each degree that was given by the university during a given year. Students earning two degrees in one year will appear in the file twice.

The variable of interest is the number of credits students have accumulated at the time of graduation. We are only interested in the number of credits undergraduates have taken. It became evident early on that some additional cleaning would be necessary. University graduation requirements state that a student graduating with a bachelor's degree must have at least 122 semester credits (183 quarter credits). The graduation files have many students who graduate with fewer than the minimum number of credits for a variety of reasons. In order to remove the students not meeting the graduation requirements, students who were working on second bachelor's degrees were excluded. This criteria resulted in much fewer students not meeting the graduation requirements. Later it was decided to also relax the credit requirement to 120 and 180 credits, which also reduced the number of students not meeting minimum requirements in the data set. The remaining students who did not meet the minimum requirements were eliminated.

The university went from a quarter system to a semester system starting in the 1998/1999 academic year. All credits were converted with a factor of 2/3 before this date. Before this switch the university only recorded credit earned at the University of Utah under the total credit category, for each of the years before the switch all other credit was added into the total credit amount before making the calculations for minimum credit requirements. Consequently transfer credit, test credit and other credit were included in the final data set. The minimum credit requirement was subtracted from the number of credits earned since it is assumed that all students have earned above the minimum.

2 Credits at Graduation

Graduation requirements state that a student must have at least 120 credits to graduate. We could view the number of credits taken as a waiting time, once a student earns 120 credits, they continue to take courses until they meet their particular graduation requirements. For these reasons, it makes sense to use the exponential distribution to model number of credits to graduation. The plots in Figure 1 and Figure 2 show kernel density estimators for credits to graduation for each academic year. Each density plot shows something close to an exponential shape, the values in Table 1 show that the mean and standard deviation of the total credit hours are roughly the same, another feature of exponential data. A generalized likelihood ratio test will be derived to determine if the yearly averages are all the same after the data is tested for the exponential assumption.

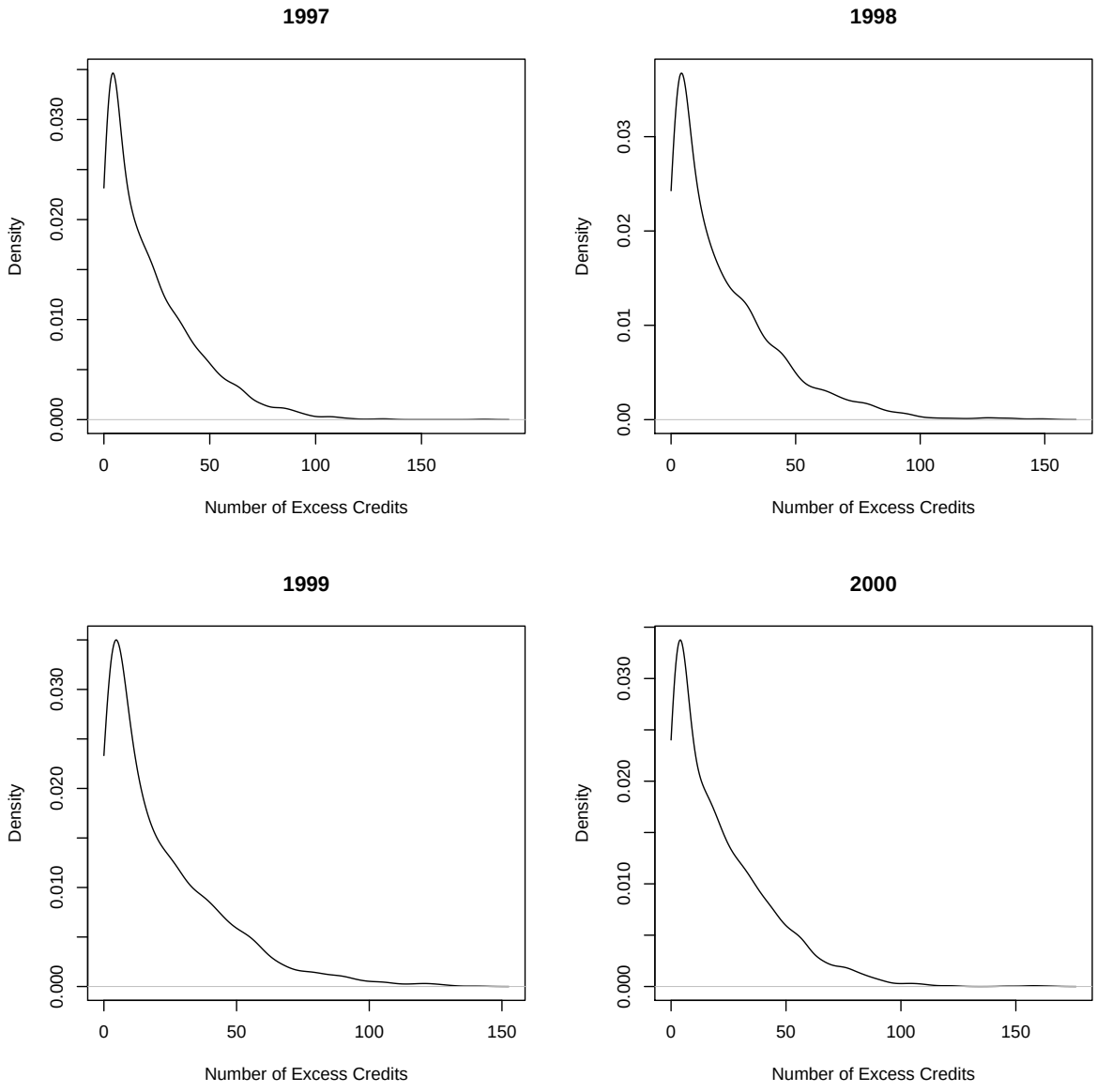

Figure 1: Density Plots for Academic Years 1997 - 2000

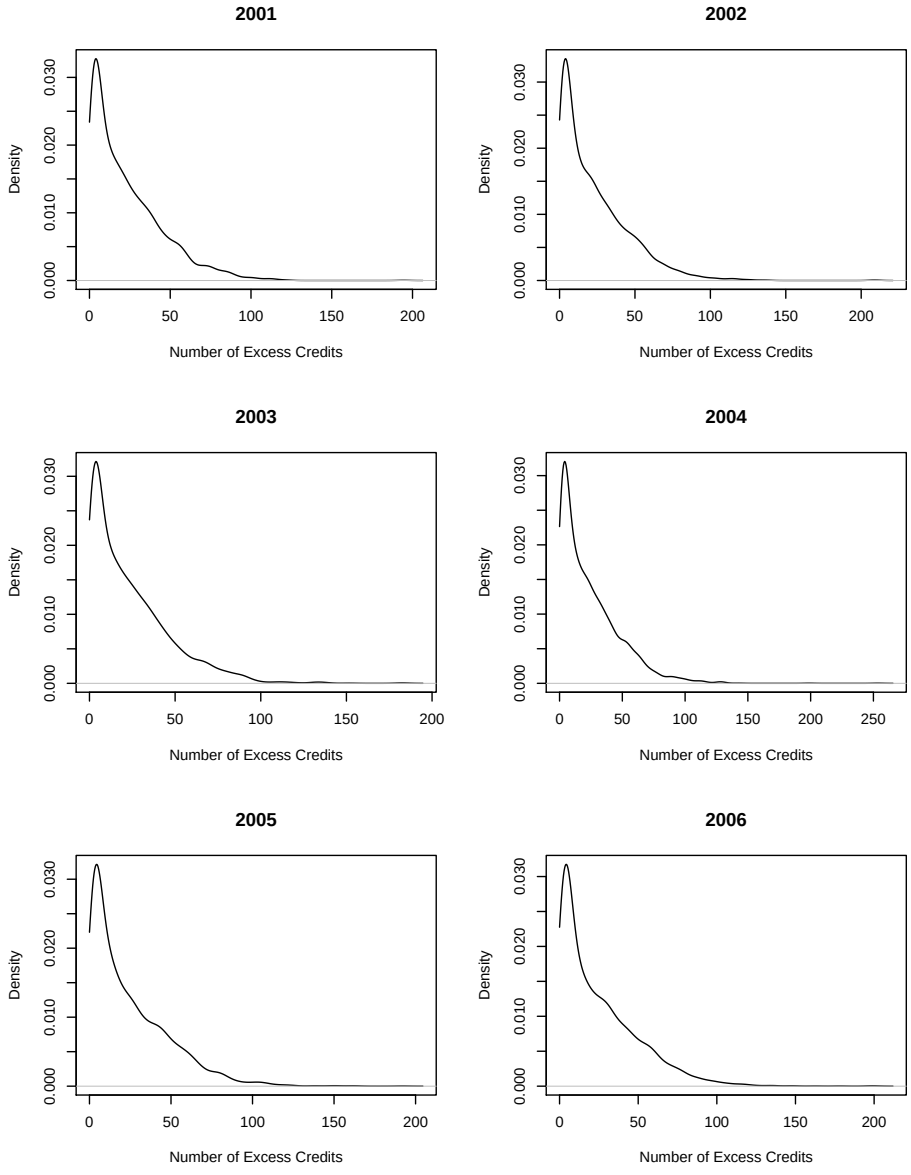

Figure 2: Density Plots for Academic Years 2001 - 2006

Year	Mean	St Deviation
1997	22.119	21.436
1998	21.828	22.066
1999	22.465	22.595
2000	22.359	21.648
2001	23.009	21.983
2002	22.985	22.671
2003	23.533	23.007
2004	24.009	23.392
2005	24.544	23.886
2006	25.148	24.670

Table 1: Mean and Standard Deviation of Excess Credits

3 Tests for Exponentially Distributed Data

We would like to test the graduation data for the assumption that it is exponentially distributed. Three tests will be derived and performed to verify this assumption.

3.1 χ^2 Test

Assume that $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables with distribution funcion F and let F_0 be a distribution function. We would like to test

$$H_0 : \text{there is a } \theta \text{ such that } (\theta \in \Theta, \dim \Theta = d), F(t) = F_0(t;\theta) \text{ for all } t$$

against the alternative $H_A : H_0$ not true. We would like to show that our data comes from the exponential distribution so that

$$F_0(x;\theta) = 1 - e^{-x/\theta}$$

and $d = \dim \Theta = 1$. Then we divide the data into K cells so that $0 = t_0 < t_1 < \cdots < t_K = \infty$ and define

$$Y_i = \sum_{1 \le j \le n} I\{t_{i-1} < X_j \le t_i\}.$$

Then $(Y_1, Y_2, \ldots, Y_K)$ is multinomial with parameters $(n, p_1, p_2, \ldots, p_K)$ where $p_i = F(t_i) - F(t_{i-1})$. If H_0 holds, then there is $\theta \in \Theta$ such that

$$p_i = F_0(t_i;\theta) - F_0(t_{i-1};\theta),$$

which, under H_0, becomes

$$p_i = (1 - e^{-t_i/\theta}) - (1 - e^{-t_{i-1}/\theta}) = e^{-t_{i-1}/\theta} - e^{-t_i/\theta}$$

The multinomial random sample depends on the parameter θ. We will estimate θ from $y_1, y_2, \ldots, y_K$ using the maximum likelihood method, the estimator is denoted $\hat{\theta}$. The multinomial distribution $(Y_1, Y_2, \ldots, Y_K)$ has the density function

$$P\{Y_1 = y_1, \ldots, Y_K = y_k\} = \frac{n!}{y_1! \cdots y_K!} p_1^{y_1} \cdots p_K^{y_K}$$

which we need to maximize with respect to θ. Under H_0, we have the likelihood function

$$L = \frac{n!}{y_1! \cdots y_K!} (F_0(t_1; \theta) - F_0(t_0; \theta))^{y_1} \cdots (F_0(t_{K-1}; \theta) - F_0(t_K; \theta))^{y_K}$$

and the log-likelihood is

$$\ell = \log n! - \sum_{i=1}^{K} \log y_i! + \sum_{i=1}^{K} y_i \log \left(F_0(t_i; \theta) - F_0(t_{i-1}; \theta)\right).$$

After substituting the distribution function for the exponential distribution, the log-likelihood becomes

$$\ell = \log n! - \sum_{i=1}^{K} \log y_i! + \sum_{i=1}^{K} y_i \log \left(e^{-t_{i-1}/\theta} - e^{-t_i/\theta}\right).$$

In order to find the maximum likelihood estimate, we take the derivative of ℓ with respect to θ and set to zero, which gives the expression

$$\frac{\partial \ell}{\partial \theta} = \sum_{i=1}^{K} \frac{1}{\theta^2} \left(\frac{t_{i-1} e^{-t_{i-1}/\theta} - t_i e^{-t_i/\theta}}{e^{-t_{i-1}/\theta} - e^{-t_i/\theta}} \right) = 0.$$

We can eliminate the θ^2 factor so that the expression becomes

$$\frac{\partial \ell}{\partial \theta} = \sum_{i=1}^{K} \left(\frac{t_{i-1} e^{-t_{i-1}/\theta} - t_i e^{-t_i/\theta}}{e^{-t_{i-1}/\theta} - e^{-t_i/\theta}} \right) = 0.$$

There is no simple solution for this equation. On general advice from Bain and Englehart [1], we use $\hat{\theta} = \bar{X}$. The test statistic is then defined as

$$Q = \sum_{1 \le i \le K} \frac{(Y_i - n[F_0(t_i, \hat{\theta}) - F_0(t_{i-1}, \hat{\theta})])^2}{n[F_0(t_i, \hat{\theta}) - F_0(t_{i-1}, \hat{\theta})]}.$$

If H_0 holds, then Q can be approximated with the χ^2 distribution with $K - d - 1$ degrees of freedom. Here we take $d = 1$ for the one estimated paramter $\hat{\theta}$.

Table 2 shows the results of the χ^2 test when performed on the data as a whole and when performed on each year's credits separately. We see that the χ^2 test does not reject for any of these cases meaning that we can assume that the data is exponentially distributed.

Year	Q	df	p value
1997	1.721	16	1.000
1998	2.838	14	0.999
1999	3.770	13	0.993
2000	2.015	15	1.000
2001	0.961	8	0.998
2002	0.925	9	1.000
2003	2.762	17	1.000
2004	0.885	11	1.000
2005	1.097	8	0.998
2006	0.995	8	0.998
All	5.462	11	0.907

Table 2: Results for χ^2 test for exponentially distributed data

3.2 Transformation into uniform order statistics

Let $X_1, \ldots, X_n$ be independent, identically distributed exponential random variables with mean θ and define

$$S(i) = X_1 + X_2 + \cdots + X_i.$$

Then

$$\left(\frac{S(1)}{S(n)}, \frac{S(2)}{S(n)}, \ldots, \frac{S(n-1)}{S(n)}\right)$$

has the same distribution as the order statistics of $n-1$ independent uniform $[0,1]$ random variables. So we can apply the general results on the uniform empirical distribution function. For example,

$$\sqrt{n} \max_{1 \le k \le n-1} \left| \frac{S(k)}{S(n)} - \frac{k}{n} \right|$$

gives the Kolmogorov-Smirnov statistic and

$$\sum_{1 \le k \le n-1} \left(\frac{S(k)}{S(n)} - \frac{k}{n} \right)^2$$

gives the Cramér-von Mises statistic. Critical values for the following modification of the Cramér-von Mises statistic can be found in the Appendix of Bain and Englehardt [1]:

$$CM = \frac{1}{12n} + \sum_{i=1}^{n} \left(F_0(x_i; \hat{\theta}) - \frac{i - 0.5}{n} \right)^2.$$

However, the critical values for the Kolmogorov-Smirnov statistic are only calculated for a fully specified distribution and not for the case where the parameters are estimated. The

critical values for the Kolmogorov-Smirnov statistic are taken from Lilliefors [5] and we use the following modification as our test statistic :

$$D = \max_{1 \le i \le n} |F_0(X_i) - S_n(X_i)|$$

where F_0 is the cumulative distribution function for the exponential distribution with parameter $1/\theta = \bar{X}$ and S_n is the sample cumulative distribution function.

Year	D	Critical Value	CM	Critical Value
1997	0.0787	0.0564	1.56	0.224
1998	0.0779	0.0556	1.29	0.224
1999	0.0646	0.0314	2.13	0.224
2000	0.0734	0.0294	2.44	0.224
2001	0.0755	0.0287	1.90	0.224
2002	0.0855	0.0275	2.11	0.224
2003	0.0885	0.0276	2.37	0.224
2004	0.0798	0.0294	1.88	0.224
2005	0.0796	0.0309	2.39	0.224
2006	0.0938	0.0355	2.00	0.224
All	0.0787	0.0150	17.38	0.224

Table 3: Results for KS and Cramér-von Mises Tests

Table 3 shows the results for the Kolmogorov-Smirnov tests and Cramér-von Mises tests. As in the previous subsection, these tests were performed for each individual year and all years combined. In this case both tests reject the exponential assumption under all circumstances.

3.3 Total time on test

For this test, we note that $f_0(t) = \exp(-t)I(t \ge 0)$ and $F_0^{-1}(t) = -\log(1-t)$ and therefore

$$f_0(F_0^{-1}(t)) = 1 - t, \quad 0 \le t \le 1.$$

Then $\{T_k : 1 \le k \le n-1\}$ is the Total Time on Test transform, where

$$T_k = \frac{\sum_{i=1}^{k}(n-i+1)(X_{i+1,n} - X_{i,n})}{\sum_{i=1}^{n-1}(n-i+1)(X_{i+1,n} - X_{i,n})}, \quad 1 \le k \le n-1.$$

It can be shown that if the observations are exponential then

$$t_1 = \sqrt{n} \max_{1 \le k \le n-1} \left| T_k - \frac{k}{n} \right| \quad \xrightarrow{d} \quad \sup_{0 \le t \le 1} |B(t)|$$

and

$$t_2 = \sum_{k=1}^{n-1} \left(T_k - \frac{k}{n} \right)^2 \quad \xrightarrow{d} \quad \int_0^1 B^2(t) \quad dt,$$

where $\{B(t) : 0 \leq t \ \leq 1\}$ is a Brownian bridge.

Table 4 shows values for t_1 and t_2 for each year and for all years combined. The critical values for t_1 and t_2 are found in Bain and Englhardt [1]. The $\alpha = 0.05$ critical value for t_1 is 0.461 and for $\alpha = 0.01$ the critical value is 0.743. In all cases we reject the assumption that the data comes from the exponential distribution. The $\alpha = 0.05$ critical value for t_2 is 1.094, the $\alpha = 0.01$ critical value is 1.298. Again, we see that in all cases the exponential assumption is rejected.

Year	t_1	t_2
1997	2.906	3.946
1998	2.805	4.038
1999	3.709	5.855
2000	4.016	7.547
2001	3.534	6.060
2002	3.559	4.905
2003	3.669	6.217
2004	3.617	5.382
2005	4.008	6.714
2006	3.930	6.154
All	9.835	50.233

Table 4: Values for Total Time on Test Transformation

The density plots in Figure 1 and 2, show the reason why the Cramér-von Mises, Kolmogorov-Smirnov and total time on test transformation tests reject the exponential assumption. If the data were exponentially distributed, these density plots would start at a maximum value of zero and decrease from that point on, instead the maximum is not reached until around four credits (in most cases). All of these tests determine the deviation of the empirical distribution function from the assumed exponential distribution function. Figures 3 and 4 show QQ-plots for all academic years, each plot shows a red line showing where the data points should be located if they come from an exponential distribution. In all cases, the points representing the empirical distribution function dip below the expected line at the beginning. This dip corresponds to the point where the maximum is reached in the density plots of Figures 1 and 2. The the χ^2 test does not reject the hypothesis that the data come from an exponential distribution due to the fact that the large amount of data that appear at about four credits all fit in one cell. The more sophisticated tests, however, are able to determine the deviation from the exponential distribution.

Since the Gamma distribution is an abstraction of the exponential distribution, it is the next obvious choice. We note here that $\text{Exp}(\theta) = \text{Gamma}(1, \theta)$ and that for Gamma distributed data $X_i > 0$. Because of the way the data was extracted we recognize that there

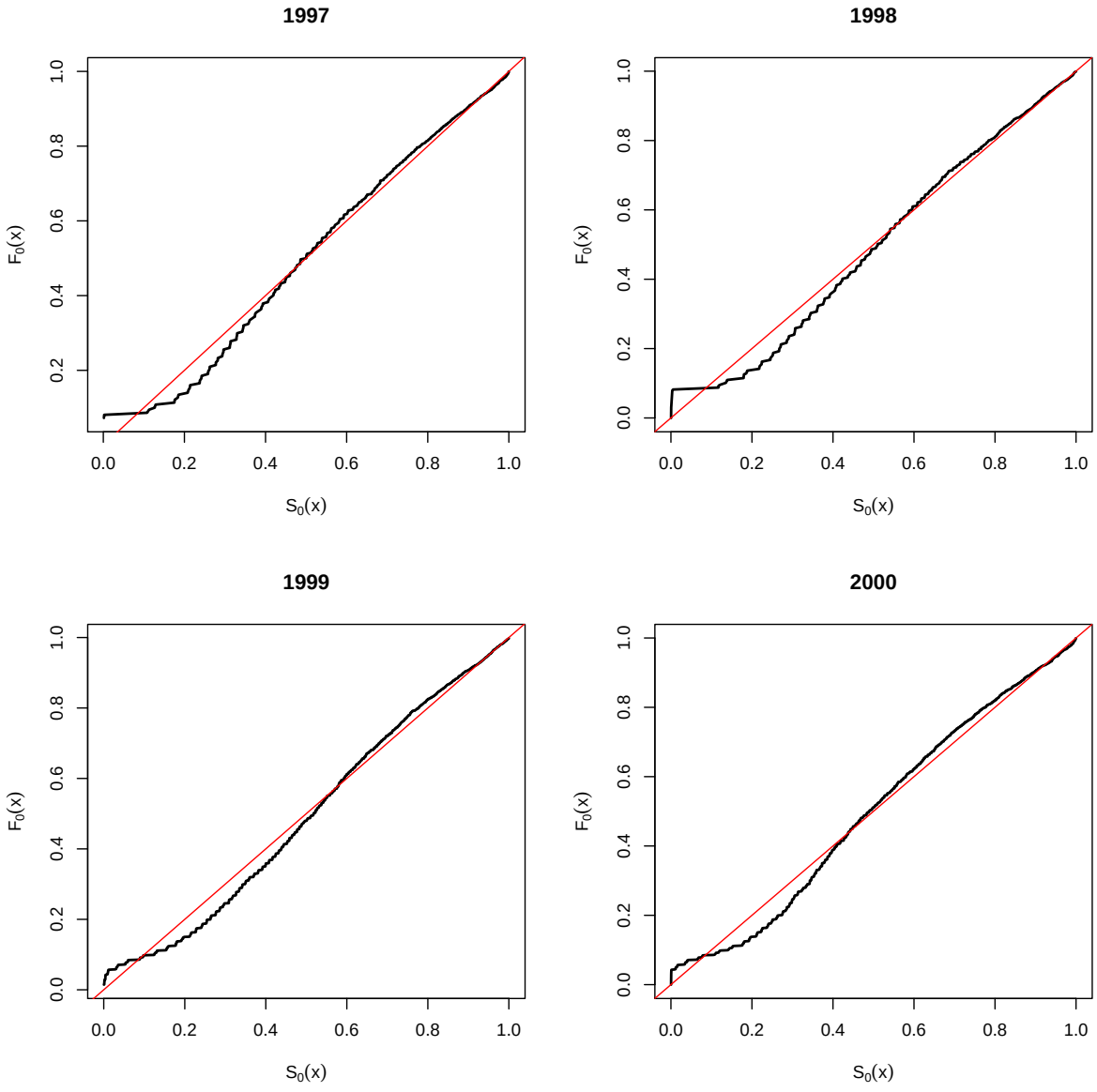

Figure 3: QQ-Plots for Academic Years 1997 - 2000

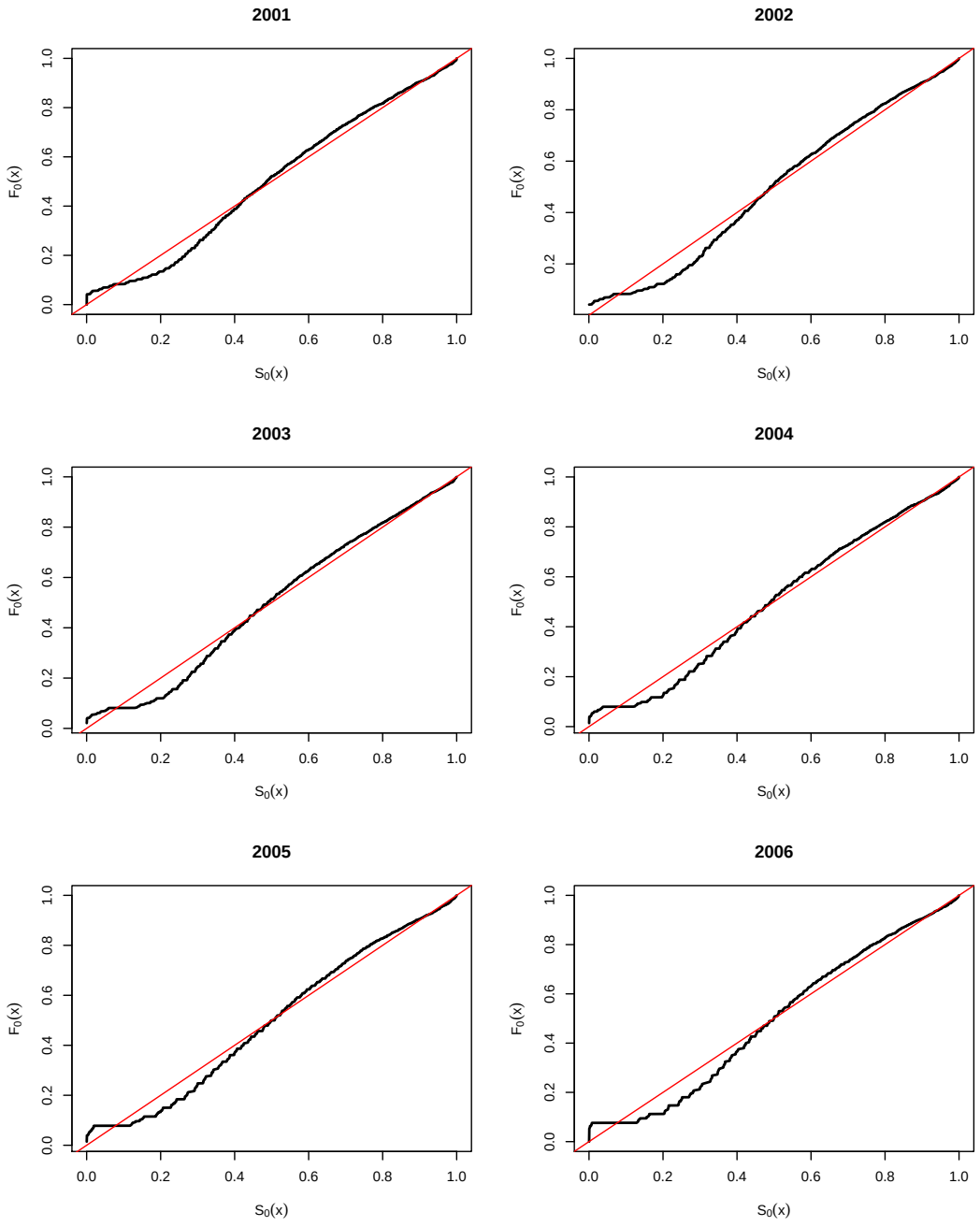

Figure 4: QQ-Plots for Academic Years 2001 - 2006

are some data points that are zero. We will attempt to show that the simple transformation $Y = X + 1$ follows the Gamma distribution. We will attempt to show this assumption using a simple graphical method. We previously determined the reason the data does not follow the exponential distribution is due to the increase in the empirical density function around four credits. Figure 5 and 6 show density plots of the underlying data in red with a density plot of the Gamma density function in blue for the same data points. The density function of the Gamma distribution is given as

$$f(x) = \frac{1}{\theta^{\kappa}\Gamma(\kappa)} x^{\kappa-1} e^{-x/\theta}, \quad x > 0.$$

Here we use the maximum likelihood estimates of the parameters κ and θ. The maximum likelihood extimate of θ is given by Bain and Englhardt [1] as

$$\hat{\theta} = \frac{\bar{x}}{\hat{\kappa}}.$$

The equation for the maximum likelihood estimate of κ is also given by Bain and Englehardt [1] and is the solution to

$$\log \hat{\kappa} - \Psi(\hat{\kappa}) - \log \bar{x}/\tilde{x} = 0$$

where $\tilde{x}$ is the geometric mean of the data and the psi function is defined as

$$\Psi(x) = \frac{\Gamma'(\kappa)}{\Gamma(\kappa)}.$$

Here we see that the maximum likelihood estimate of κ cannot be be solved in closed form. Bain and Englehardt [1] note that the following approximation can be used for $\hat{\kappa}$

$$\hat{\kappa} = \frac{0.5000876 + 0.1648852M - 0.0544274M^2}{M} \qquad 0 < M \leq 0.5772, \tag{3.1}$$

where $M = \log \bar{x}/\tilde{x}$. Table 5 shows the parameters for the Gamma distributions plotted in Figures 5 and 6. Notice that the values for M fall in the interval specified by (3.1), so that we will not note the equations for values of M that fall outside this interval.

Figures 5 and 6 show a strong correspondance between the empirical density function and the Gamma density. The values for κ in Table 5 show that the data deviate from the exponential distribution by only a small amount. On the surface, the exponential assumption is a good one however, the deviation from the exponential distribution shows that by and large students do not graduate with the minimum number of credits, but are more likely to take slightly more than the minimum requirement.

4 Generalized Likelihood Ratio Test

Here we derive the generalized likelihood ratio test for exponentially distributed data to test homogeneity. Even though the previous subsection discussed the fact that the data

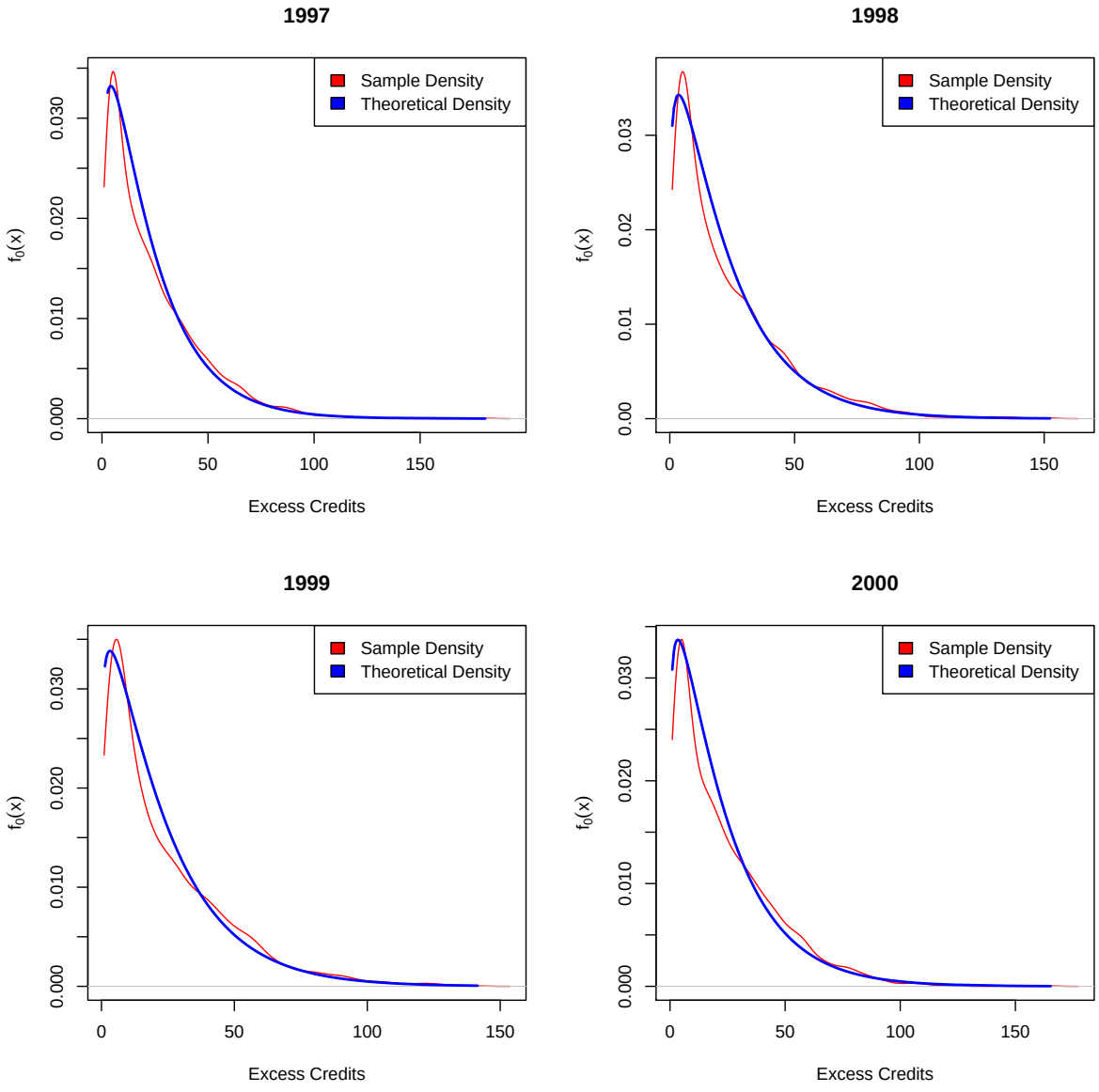

Figure 5: Comparison to Gamma Distribution for Academic Years 1996 - 2000

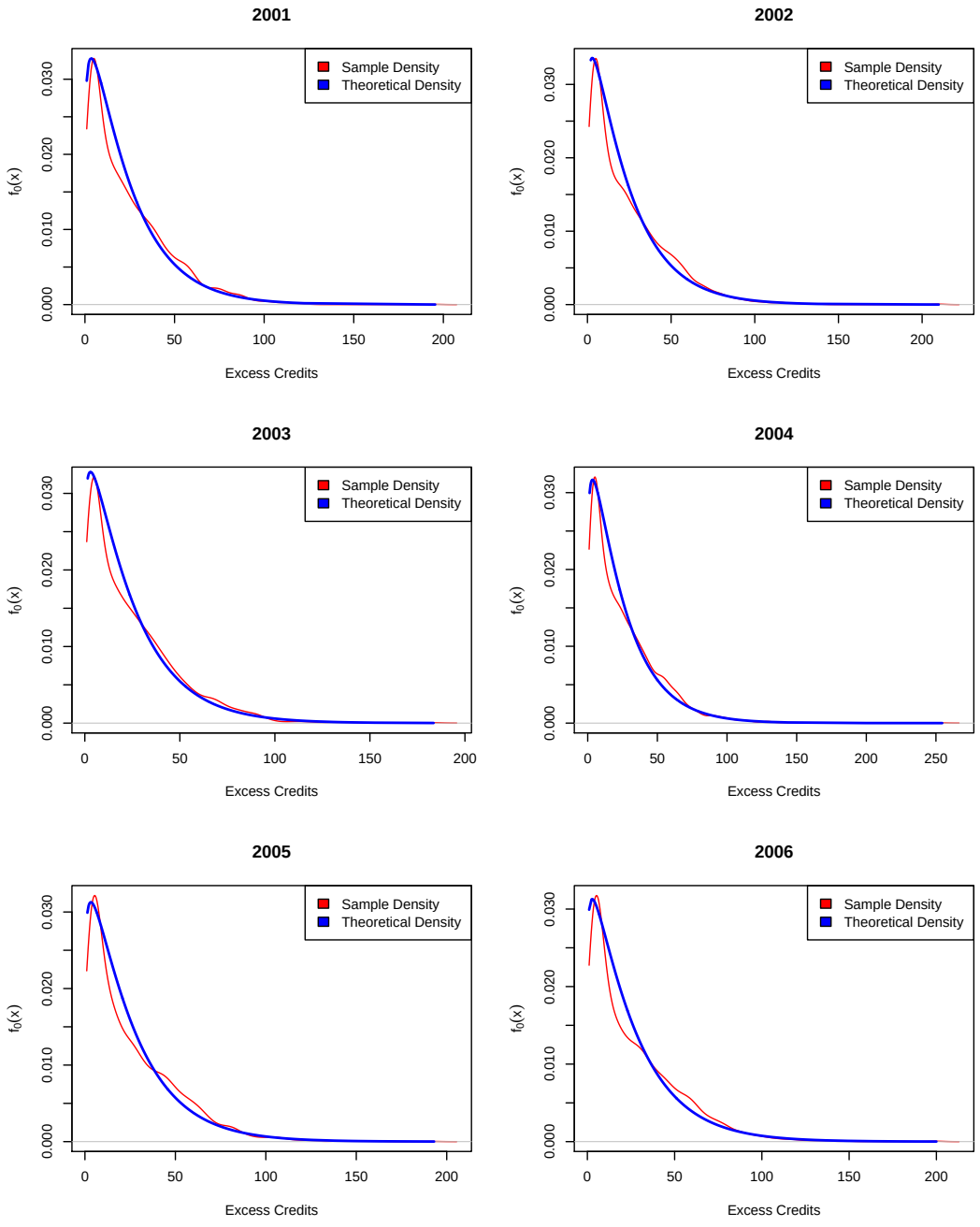

Figure 6: Comparison to Gamma Distribution for Academic Years 2001 - 2006

Year	M	$\hat{\kappa}$	$\hat{\theta}$
1997	0.46180	1.2227	18.909
1998	0.47887	1.1831	19.294
1999	0.49017	1.1584	20.256
2000	0.48407	1.1716	19.938
2001	0.48335	1.1732	20.464
2002	0.50057	1.1367	21.101
2003	0.49979	1.1383	21.553
2004	0.48835	1.1623	21.516
2005	0.49477	1.1487	22.237
2006	0.50994	1.1178	23.392
All	0.49107	1.1565	21.013

Table 5: Parameter values for the Gamma distribution

deviate from the exponential distribution, it will be used since we are dealing with averages and the deviation is not large. We will make the following assumption about our data

$$X_{ij} \sim Exp(\theta_i) \qquad 1 \leq i \leq k, \quad 1 \leq j \leq n_i$$

where the data represent independent random variables. This means that $\{X_{ij}\}$ are independent. We let, $k = 10$, for each of the years under consideration, n_i is the number of students graduating in the i^{th} year and X_{ij} is the number of credit hours (above the minimum requirement) for the j^{th} student in the i^{th} year. We will also denote the total number of observations as follows.

$$N = \sum_{i=1}^{k} n_i$$

A test will be derived for the following hypothesis.

$$H_0 : \theta_1 = \theta_2 = \ldots = \theta_k \qquad H_a : \theta_1 \neq \theta_2 \neq \ldots \neq \theta_k$$

The generalized likelihood ratio as given by Bain and Englehardt [1] is defined as

$$\Lambda(\boldsymbol{X}) = \frac{\max\limits_{\theta \in \Omega_0} f(\boldsymbol{X}; \theta)}{\max\limits_{\theta \in \Omega} f(\boldsymbol{X}; \theta_1, \ldots, \theta_k)} = \frac{f(\boldsymbol{X}; \hat{\theta})}{f(\boldsymbol{X}; \hat{\theta}_1, \ldots, \hat{\theta}_k)},$$

where $\boldsymbol{X} = \{X_{ij}; 1 \leq j \leq n_i, 1 \leq i \leq k\}$, $\Omega_0 = \{\theta_1 = \theta_2 = \cdots = \theta_k > 0\}$ and $\Omega = \{\theta_1 > 0, \theta_2 > 0, \ldots, \theta_k > 0\}$. We also note that $\hat{\theta}$ and $\hat{\theta}_1, \ldots, \hat{\theta}_k$ denote the corresponding maximum likelihood estimators. The log likelihood function is given as

$$\log \Lambda(\boldsymbol{X}) = \log f(\boldsymbol{X}; \hat{\theta}) - \log f(\boldsymbol{X}; \hat{\theta}_1, \ldots, \hat{\theta}_k).$$

If H_0 holds, then $-2 \log \Lambda(\boldsymbol{X}) \sim \chi^2_{k-1}$.

Next we find the individual log likelihood functions under the null and under the alternative. We will derive equations for the maximum likelihood estimates and log likelihood functions. Under H_0 we have

$$L(X;\theta) = \prod_{i=1}^{k}\prod_{j=1}^{n_i}\frac{1}{\theta}e^{-X_{ij}/\theta} = \prod_{i=1}^{k}\frac{1}{\theta^{n_i}}\exp\left(-\frac{1}{\theta}\sum_{j=1}^{n_i}X_{ij}\right),$$
$$= \frac{1}{\theta^{n_1+\ldots+n_k}}\exp\left(-\frac{1}{\theta}\sum_{i=1}^{k}\sum_{j}^{n_i}X_{ij}\right)$$

and therefore

$$\ell(X;\theta) = -\sum_{i=1}^{k} n_i \log\theta - \frac{1}{\theta}\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}. \tag{4.1}$$

Then we get an expression for the maximum likelihood estimate by differentiating (4.1) and solving for zero. Namely,

$$\frac{\partial\ell}{\partial\theta} = -\frac{1}{\theta}\sum_{i=1}^{k} n_i + \frac{1}{\theta^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij} = 0$$

and therefore

$$\frac{1}{\theta}\sum_{i=1}^{k} n_i = \frac{1}{\theta^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}$$

so that we get

$$\hat{\theta} = \left(\sum_{k=1}^{k} n_i\right)^{-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}.$$

Then we substitute $\hat{\theta}$ into the log likelihood for the final calculations

$$\ell(X;\hat{\theta}) = -\sum_{i=1}^{k} n_i \log\hat{\theta} - \frac{1}{\hat{\theta}}\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}$$

so that

$$\ell(X;\hat{\theta}) = -\sum_{i=1}^{k} n_i \log\hat{\theta} - \sum_{i=1}^{k} n_i.$$

Next, we derive the likelihood function under the alternative. By independence we obtain that

$$L(\boldsymbol{X};\theta_1,\ldots,\theta_k) = \prod_{i=1}^{k}\prod_{j=1}^{n_i}\frac{1}{\theta_i}e^{-X_{ij}/\theta_i} = \prod_{i=1}^{k}\frac{1}{\theta_i^{n_i}}\exp\left(-\frac{1}{\theta_i}\sum_{j=1}^{n_i}X_{ij}\right),$$

which is the same as

$$L(\boldsymbol{X};\theta_1,\ldots,\theta_k)=\left(\prod_{i=1}^{k}\frac{1}{\theta_i^{n_i}}\right)\exp\left(-\sum_{i=1}^{k}\frac{1}{\theta_i}\sum_{j=1}^{n_i}X_{ij}\right).$$

Then for the log-likelihood function we get that

$$\ell(\boldsymbol{X};\theta_1,\ldots,\theta_k)=-\sum_{i=1}^{k}n_i\log\theta_i-\sum_{i=1}^{k}\frac{1}{\theta_i}\sum_{j=1}^{n_i}X_{ij}. \tag{4.2}$$

Solving the partial derivatives of (4.2) with respect to $\theta_1,\theta_2,\ldots,\theta_k$ for zero we conclude

$$\frac{\partial\ell}{\partial\theta_i}=-\frac{n_i}{\theta_i}+\frac{1}{\theta_i^2}\sum_{j=1}^{n_i}X_{ij}=0.$$

Solving for $\hat{\theta}_i$, we get

$$\hat{\theta}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}X_{ij}.$$

Finally, the maximum of the log likelihood has the following form:

$$\begin{aligned}\ell(\boldsymbol{X};\hat{\theta}_1,\ldots,\hat{\theta}_k)&=-\sum_{i=1}^{k}n_i\log\hat{\theta}_i-\sum_{i=1}^{k}\frac{1}{\hat{\theta}_i}\sum_{j=1}^{n_i}X_{ij}\\&=-\sum_{i=1}^{k}n_i\log\hat{\theta}_i-\sum_{i=1}^{k}n_i.\end{aligned}$$

Then we have the following generalized log likelihood function

$$\log\Lambda(\boldsymbol{X})=\sum_{i=1}^{k}n_i\log\hat{\theta}_i-\left(\sum_{i=1}^{k}n_i\right)\log\hat{\theta}. \tag{4.3}$$

As mentioned above, we have that $-2\log\Lambda(\boldsymbol{X})\approx\chi^2(k-1)$. This will now be shown. First, we will use the following approximation:

$$\log(x)=\log(1+x-1)\approx(x-1)-\frac{1}{2}(x-1)^2$$

as $x\to 1$. Then we have the following Taylor expansion for $-2\log\Lambda(x)$:

$$\begin{aligned}-2\log\Lambda(\boldsymbol{X})&=-2\sum_{i=1}^{k}n_i\log\left(\frac{\hat{\theta}_i}{\hat{\theta}}\right),\\&\approx-2\sum_{i=1}^{k}n_i\left(\frac{\hat{\theta}_i}{\hat{\theta}}-1\right)+2\frac{1}{2}\sum_{i=1}^{k}n_i\left(\frac{\hat{\theta}_i}{\hat{\theta}}-1\right)^2.\end{aligned} \tag{4.4}$$

The first term in (4.4) is actually zero, since

$$-2\sum_{i=1}^{k} n_i \left(\frac{\hat{\theta}_i}{\hat{\theta}} - 1\right) = -2\sum_{i=1}^{k} n_i \left(\frac{\hat{\theta}_i - \hat{\theta}}{\hat{\theta}}\right)$$
$$= -\frac{2}{\hat{\theta}} \sum_{i=1}^{k} n_i \left(\hat{\theta}_i - \hat{\theta}\right). \tag{4.5}$$

Equation (4.5) is zero regardless of the leading term, since by the definitions of $\hat{\theta}_i$ and $\hat{\theta}$ we have

$$\sum_{i=1}^{k} \left(n_i\hat{\theta}_i - n_i\hat{\theta}\right) = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij} - \sum_{i=1}^{k} n_i\hat{\theta},$$

which simplifies to

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} - \sum_{i=1}^{k} n_i \left(\sum_{k=1}^{k} n_i\right)^{-1} \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} = 0.$$

This means that $-2\log\Lambda(\boldsymbol{X})$ can be approximated with

$$-2\log\Lambda(\boldsymbol{X}) \approx \sum_{i=1}^{k} n_i \left(\frac{\hat{\theta}_i - \hat{\theta}}{\hat{\theta}}\right)^2.$$

We use the test statistic

$$T = \sum_{i=1}^{k} n_i \left(\frac{\hat{\theta}_i - \hat{\theta}}{\hat{\theta}}\right)^2 \tag{4.6}$$

to obtain results. Some experimentation with (4.3) revealed that, due to the uneven sample sizes, negative numbers were occasionally returned. Equation (4.6) will not exhibit this behavior and, therefore, is a better choice for our statistic. We note that T is approximately $\chi^2(k-1)$ according to the general theory of the generalized likelihood method since

$$\sum_{i=1}^{k} n_i \left(\frac{\hat{\theta}_i - \hat{\theta}}{\hat{\theta}}\right) = 0.$$

4.1 Results

The generalized likelihood ratio test in Section 4 was derived to test for homogeneity of the average number of credits students took to graduate. The first test performed was to determine whether there is one mean for all years or not. A p-value of 1.084847×10^{-10} resulted, the hypothesis that the mean remains the same for all years is rejected at any reasonable level of significance.

Next, the following procedure was used, we start with the data from 1997 and keep adding the data from consecutive years until a p-value less than some α_i is achieved. The

index i is used to specify which of the tests was performed, in this case $i = 1$. Then we say that the years 1997 through the last year before α_1 is achieved are homogeneous. Then we take a new set of data starting with the year that made the previous tests fail and add the data from consecutive years until the test rejects H_0 again at the confidence level α_2 and so on. We would also like to have joint coverage for all of the tests performed. This is done using the Bonferroni method of multiple comparisons from Johnson and Wichern [2]. Let C_i denote the i^{th} test perfomed where

$$P\{C_i \text{ true}\} = 1 - \alpha_i, \quad i = 1, 2, \ldots, m.$$

We use the Bonferroni inequality to have joint coverage for all of the tests performed. We have the following relation

$$P\{\text{all } C_i \text{ true}\} = 1 - P\{\text{at least one } C_i \text{ false}\},$$

and by the Bonferonni inequality

$$1 - P\{\text{at least one } C_i \text{ false}\} \geq 1 - \sum_{i=1}^{m} P\{C_i \text{ false}\}$$

furthermore,

$$1 - \sum_{i=1}^{m} P\{C_i \text{ false}\} = 1 - \sum_{i=1}^{m} (1 - P\{C_i \text{ true}\})$$

and finally we have

$$P\{\text{all } C_i \text{ true}\} \geq 1 - \sum_{i=1}^{m} \alpha_i$$

Then we say that there is joint coverage for all tests where

$$\alpha = \sum_{i=1}^{m} \alpha_i.$$

We will take $\alpha_i = \alpha/m$ as our choice for the α_i and we assume that $m = 2$, or that there will only be two tests necessary. We would like $\alpha = 0.05$, so that $\alpha_i = 0.025$.

The p-values for this procedure are found in Table 6. These tests indicate that the average of excess credits is the same from the years 1997 through 2003 and again from 2004 to 2006. We conclude that the number of excess credits is growing over the last ten years. However, it appears that the average is not growing steadily, as we expect, but jumps up every few years. Figure 7 shows a plot of the average excess credits by year along with lines representing the averages from the hypothesis tests that were performed. It shows that the underlying data actually does display an upward trend, however, the statistical tests show two change points within the data. The next section will explore ways to test validity of the the change points.

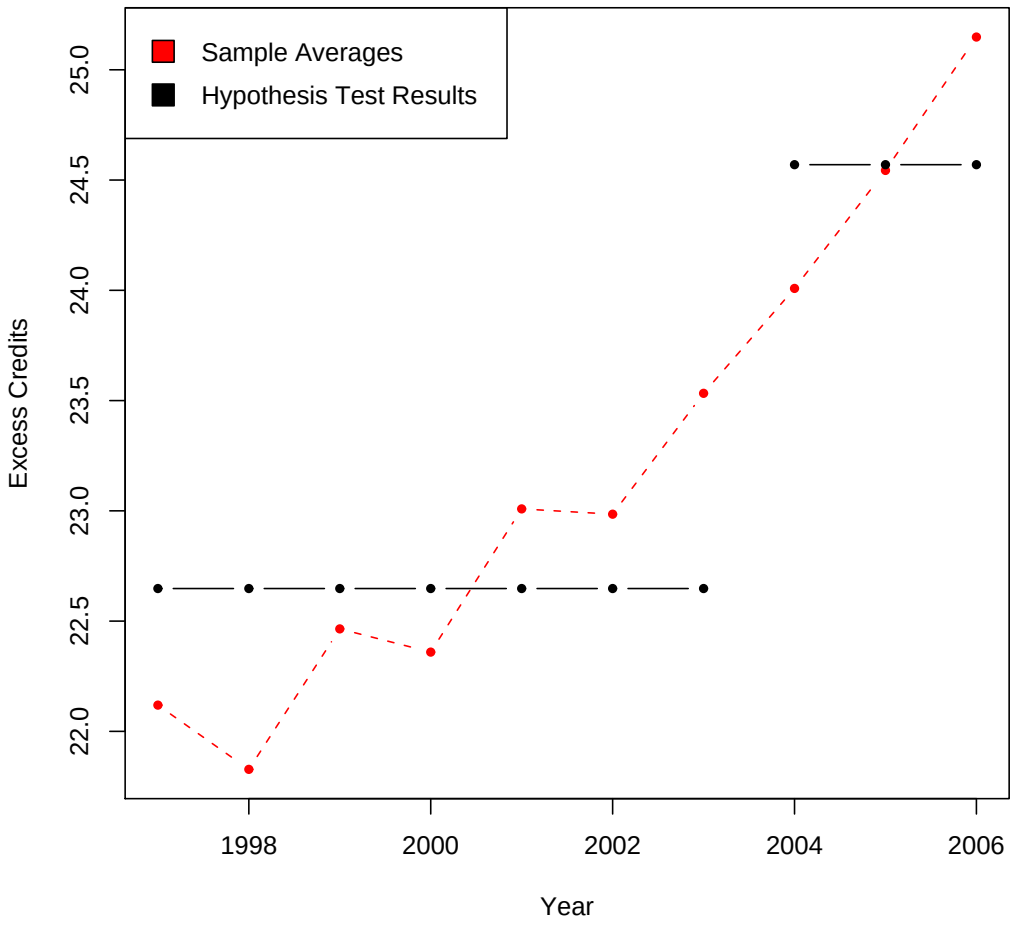

Figure 7: Average excess credits by year

Start Year	End Year	p-value
1997	2003	0.047
1997	2004	0.001
2004	2006	0.148

Table 6: Tests for Homogeneity

5 Regression Obeying Two Different Regimes

Figure 7 suggests that there is a change in the average number of excess credits from the period 1997 through 2003 to the period 2003 through 2005. Quandt [3] suggested some methods for testing the significance of such a jump using regression. Under regression, we make the assumption that our data is normally distributed. A likelihood ratio test will be derived using the ratio

$$\Lambda_k = \max_{2 \le k \le N-2} \frac{L_k(\boldsymbol{y})}{L(\boldsymbol{y})}. \tag{5.1}$$

We reject H_0, that no regime shift occurred, if Λ_k is large.

In order to develop $L_k(\boldsymbol{y})$, we assume the existence of two relationships within the data, the relationships are of the form

$$\begin{aligned} y_i &= \alpha_1 x_i + \beta_1 + \varepsilon_i \qquad 1 \le i \le k^* \\ y_i &= \alpha_2 x_i + \beta_2 + \varepsilon_i \qquad k^* < i \le N, \end{aligned}$$

where ε_i are normally and independently distributed error terms with zero mean and variances σ_1^2 and σ_2^2. Likelihood ratios will be developed for a three different scenarios. In our case,

$$y_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij},$$

the average credit hours to graduate, x_i denotes the year associated with each average. Since y_i is an average, normality of the y's can be assumed. Since randomness comes from the ε_i's, the normality of the errors is also assumed.

5.1 Variances unknown and unequal

First we look at the scenario that σ_1^2 and σ_2^2 are unknown and $\sigma_1^2 \neq \sigma_2^2$. Using the normality assumption, we derive a maximum likelihood function to estimate the unknown parameters $\alpha_1, \beta_1, \sigma_1^2, \alpha_2, \beta_2$ and σ_2^2 assuming that k^* is some known value k. The likelihood function under the alternative is given as

$$L_k(\boldsymbol{y}) = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(y_i - \alpha_1 x_i - \beta_1)^2} \prod_{i=k+1}^{N} \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2\sigma_2^2}(y_i - \alpha_2 x_i - \beta_2)^2}.$$

We say that a change (or regime shift) occurred at time k where the parameters changed from $(\alpha_1, \beta_1, \sigma_1^2)$ to $(\alpha_2, \beta_2, \sigma_2^2)$. In order to find the maximum likelihood estimates for each of the parameters, we find the log-likelihood assuming independence

$$\ell_k(\boldsymbol{y}) = -\frac{N}{2} - \frac{k}{2}\log\sigma_1^2 - \frac{N-k}{2}\log\sigma_2^2 - \frac{1}{2\sigma_1^2}\sum_{i=1}^{k}(y_i - \alpha_1 x_i - \beta_1)^2 - \frac{1}{2\sigma_2^2}\sum_{i=k+1}^{N}(y_i - \alpha_2 x_i - \beta_2)^2.$$

To find the maximum likelihood estimate for $(\alpha_1, \beta_1, \sigma_1^2, \alpha_2, \beta_2, \sigma_2^2)$, we solve the following partial derivatives for zero:

$$\begin{aligned}
\frac{\partial \ell_k(\boldsymbol{y})}{\partial \alpha_1} &= -\frac{1}{2\sigma_1^2}\sum_{i=1}^{k} 2(y_i - \alpha_1 x_i - \beta_1)(-x_i) = 0 \\
\frac{\partial \ell_k(\boldsymbol{y})}{\partial \beta_1} &= -\frac{1}{2\sigma_1^2}\sum_{i=1}^{k} 2(y_i - \alpha_1 x_i - \beta_1)(-1) = 0 \\
\frac{\partial \ell_k(\boldsymbol{y})}{\partial \sigma_1^2} &= -\frac{k}{2\sigma_1^2} + \frac{1}{2\sigma_1^4}\sum_{i=1}^{k}(y_i - \alpha_1 x_i - \beta_1)^2 = 0 \\
\frac{\partial \ell_k(\boldsymbol{y})}{\partial \alpha_2} &= -\frac{1}{2\sigma_2^2}\sum_{i=k+1}^{N} 2(y_i - \alpha_2 x_i - \beta_2)(-x_i) = 0 \\
\frac{\partial \ell_k(\boldsymbol{y})}{\partial \beta_2} &= -\frac{1}{2\sigma_2^2}\sum_{i=k+1}^{N} 2(y_i - \alpha_2 x_i - \beta_2)(-2) = 0 \\
\frac{\partial \ell_k(\boldsymbol{y})}{\partial \sigma_2^2} &= -\frac{N-k}{2\sigma_2^2} + \frac{2}{2\sigma_2^4}\sum_{i=k+1}^{N}(y_i - \alpha_2 x_i - \beta_2)^2 = 0.
\end{aligned}$$

To calculate the maximum likelihood estimate for α_1, we solve

$$\frac{\partial \ell_k(\boldsymbol{y})}{\partial \alpha_1} = -\frac{1}{2\sigma_1^2}\sum_{i=1}^{k} 2(y_i - \alpha_1 x_i - \beta_1)(-x_i),$$

collecting terms and setting equal to zero, we get

$$\frac{1}{\sigma_1^2}\sum_{i=1}^{k}(y_i x_i - \alpha_1 x_i^2 - \beta_1 x_i) = 0,$$

moving the α_1 terms to the other side, we get

$$\frac{\alpha_1 \sum_{i=1}^{k} x_i^2}{\sigma_1^2} = \frac{1}{\sigma_1^2}\sum_{i=1}^{k}(y_i x_i - \beta_1 x_i).$$

Then we solve for the maximum likelihood estimate $\hat{\alpha}_{1,k}$ and get

$$\hat{\alpha}_{1,k} = \frac{\sum_{i=1}^{k}(y_i x_i - \hat{\beta}_{1,k} x_i)}{\sum_{i=1}^{k} x_i^2} = \frac{\sum_{i=1}^{k}(y_i - \hat{\beta}_{1,k}) x_i}{\sum_{i=1}^{k} x_i^2}$$

where $\hat{\beta}_{1,k}$ is the maximum likelihood estimate for β_1. Next, we find the maximum likelihood estimate for β_1 by taking the partial derivative with respect to β_1

$$\frac{\partial \ell_k(\boldsymbol{y})}{\partial \beta_1} = -\frac{1}{2\sigma_1^2} \sum_{i=1}^{k} 2(y_i - \alpha_1 x_i - \beta_1)(-1) = 0$$

collecting terms and setting equal to zero yields

$$\frac{1}{\sigma_1^2} \sum_{i=1}^{k} (y_i - \alpha_1 x_i - \beta_1) = 0,$$

collecting the β_1 terms and moving this term to the other side we achieve

$$\frac{k\beta_1}{\sigma_1^2} = \frac{1}{\sigma_1^2} \sum_{i=1}^{k} (Y_i - \alpha_1 x_i),$$

which yields the maximum likelihood estimate

$$\hat{\beta}_{1,k} = \frac{1}{k} \sum_{i=1}^{k} (y_i - \hat{\alpha}_{1,k} x_i).$$

Using the linear equations for $\hat{\alpha}_{1,k}$ and $\hat{\beta}_{1,k}$, we can rewrite the expression for $\hat{\alpha}_{1,k}$ as follows:

$$\hat{\alpha}_{1,k} = \frac{k \sum_{i=1}^{k} y_i x_i - \sum_{i=1}^{k} y_i \sum_{i=1}^{k} x_i}{k \sum_{i=1}^{k} x_i^2 - \left(\sum_{i=1}^{k} x_i \right)^2},$$

and

$$\hat{\beta}_{1,k} = \frac{1}{k} \sum_{i=1}^{k} (y_i - \hat{\alpha}_{1,k} x_i).$$

We see that $\hat{\alpha}_{1,k}$ and $\hat{\beta}_{1,k}$ are the usual least-squares estimates. We also note that the maximum likelihood estimates $\hat{\alpha}_{2,k}$ and $\hat{\beta}_{2,k}$ are also their respective least-squares estimates

and have similar equations. Next, we derive the maximum likelihood estimate for σ_1^2 from the equation

$$\frac{\partial \ell_k(\boldsymbol{y})}{\partial \sigma_1^2} = -\frac{k}{2\sigma_1^2} + \frac{1}{2\sigma_1^4}\sum_{i=1}^{k}(y_i - \alpha_1 x_i - \beta_1)^2 = 0.$$

Solving for zero, we get the following as the maximum likelihood estimate for σ_1^2 :

$$\hat{\sigma}_{1,k}^2 = \frac{1}{k}\sum_{i=1}^{k}(y_i - \hat{\alpha}_1 x_i - \hat{\beta}_1)^2.$$

Again, there is a similar expression for $\hat{\sigma}_{2,k}^2$. After substituting the maximum likelihood estimates, we get

$$\ell_k(\boldsymbol{y}) = -\frac{N}{2}\log 2\pi - \frac{k}{2}\log \hat{\sigma}_{1,k}^2 - \frac{N-k}{2}\log \hat{\sigma}_{2,k}^2 - \frac{k}{2} - \frac{N-k}{2}.$$

This completes the calculations of the log-likelihood function under the alternative hypothesis that exactly one switch occurred within the data set. It remains to calculate the log-likelihood function under the null hypothesis that no change occurs. Under H_0, the likelihood function is given as

$$L(\boldsymbol{y}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \alpha x_i - \beta)^2},$$

which can be rewritten as follows

$$L(\boldsymbol{y}) = \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \alpha x_i - \beta)^2.$$

The log-likelihood function is given as

$$\ell(\boldsymbol{y}) = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \alpha x_i - \beta)^2.$$

To find the maximum likelihood estimate of $(\alpha, \beta, \sigma^2)$, we take the partial derivatives

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{y})}{\partial \alpha} &= -\frac{1}{2\sigma^2}\sum_{i=1}^{N} 2(y_i - \alpha x_i - \beta)(-x_i) \\
\frac{\partial \ell(\boldsymbol{y})}{\partial \beta} &= -\frac{1}{2\sigma^2}\sum_{i=1}^{N} 2(y_i - \alpha x_i - \beta)(-1) \\
\frac{\partial \ell(\boldsymbol{y})}{\partial \sigma^2} &= -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{N}(y_i - \alpha x_i - \beta)^2
\end{aligned}$$

and solve for zero. We begin by finding the maximum likelihood estimate of α, by multiplying through by $-2x_i$ to get

$$\frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i x_i - \alpha x_i^2 - \beta x_i) = 0.$$

Separating out the α terms and adding to zero, we get

$$\frac{\alpha\sum_{i=1}^{N} x_i^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i x_i - \beta x_i).$$

So that the maximum likelihood estimate $\hat{\alpha}$ is defined as

$$\hat{\alpha} = \frac{\sum_{i=1}^{N} x_i(y_i - \hat{\beta})}{\sum_{i=1}^{N} x_i^2},$$

where $\hat{\beta}$ is the maximum likelihood estimate of β. To obtain the maximum likelihood estimate of β, we solve

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{N} 2(y_i - \alpha x_i - \beta)(-1) = 0.$$

Multiplying through by -2 and moving the β term out of the summation, we get

$$\frac{N\beta}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i - \alpha x_i).$$

So that the maximum likelihood estimate $\hat{\beta}$ is given as

$$\hat{\beta} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\alpha} x_i).$$

To derive the maximum likelihood estimate for σ^2, we solve the equation

$$-\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{N}(y_i - \alpha x_i - \beta)^2 = 0.$$

The solution to this gives us the definition

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\alpha} x_i - \hat{\beta})^2.$$

Plugging the maximum likelihood estimates into $\ell(\boldsymbol{y})$ yields the maximum of the log-likelihood function

$$\ell(\boldsymbol{y}) = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\hat{\sigma}^2 - \frac{N}{2} \tag{5.2}$$

Next, we determine an expression for the log-likelihood ratio, which is given as

$$\lambda_k = \log\Lambda_k = \ell_k(\boldsymbol{y}) - \ell(\boldsymbol{y}),$$

so that we get

$$\lambda_k = -\frac{N}{2}\log 2\pi - \frac{k}{2}\log\hat{\sigma}^2_{1,k} - \frac{N-k}{2}\log\hat{\sigma}^2_{2,k} - \frac{N}{2} - (-\frac{N}{2}\log 2\pi - \frac{N}{2}\log\hat{\sigma}^2 - \frac{N}{2}).$$

After some simplification, we get the form:

$$\lambda_k = \frac{N}{2}\log\hat{\sigma}^2 - \frac{k}{2}\log\hat{\sigma}^2_{1,k} - \frac{N-k}{2}\log\hat{\sigma}^2_{2,k}, \tag{5.3}$$

with the likelihood ratio

$$\Lambda_k = \left(\frac{\hat{\sigma}^N}{\hat{\sigma}^k_{1,k}\hat{\sigma}^{N-k}_{2,k}}\right). \tag{5.4}$$

5.2 Variance equal and unknown

In the second scenario, the variances remain unknown but $\sigma_1^2 = \sigma_2^2 = \sigma_k^2$. In this case we have the following regression parameters under each hypothesis :

$$H_0 : (\alpha, \beta, \sigma^2) \qquad\qquad H_a : (\alpha_1, \beta_1, \sigma^2), (\alpha_2, \beta_2, \sigma^2).$$

We proceed by using the likelihood ratio in (5.1) to define the test. First, we make the comment that the likelihood function for this scenario under H_0 will be the same as (5.2), no calculations are needed to derive the likelihood function under H_0.

Under the alternative, again assuming $k^* = k$, we have the likelihood function

$$L_k(\boldsymbol{y}) = \prod_{i=1}^{k}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(y_i-\alpha_1 x_i-\beta_1)^2}\prod_{i=k+1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(y_i-\alpha_2 x_i-\beta_2)^2}$$

which becomes

$$L_k(\boldsymbol{y}) = \left(\frac{1}{2\pi\sigma^2}\right)^{N/2}\exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{k}(y_i-\alpha_1 x_i-\beta_1)^2 + \sum_{i=k+1}^{N}(y_i-\alpha_2 x_i-\beta_2)^2\right)\right]$$

after some simplification. The log-likelihood function is given as

$$\ell_k(\boldsymbol{y}) = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{k}(y_i-\alpha_1 x_i-\beta_1)^2 + \sum_{i=k+1}^{N}(y_i-\alpha_2 x_i-\beta_2)^2\right).$$

To find the maximum likelihood estimates, we begin by taking the partial derivatives with respect to $(\alpha_1, \beta_1, \alpha_2, \beta_2, \sigma^2)$ and solving for zero. The partial derivatives are defined as

$$\begin{aligned}
\frac{\partial \ell_k}{\partial \alpha_1} &= -\frac{1}{2\sigma^2} \sum_{i=1}^{k} 2(y_i - \alpha_1 x_i - \beta_1)(-x_i) \\
\frac{\partial \ell_k}{\partial \beta_1} &= -\frac{1}{2\sigma^2} \sum_{i=1}^{k} 2(y_i - \alpha_1 x_i - \beta_1)(-1) \\
\frac{\partial \ell_k}{\partial \alpha_2} &= -\frac{1}{2\sigma^2} \sum_{i=k+1}^{N} 2(y_i - \alpha_2 x_i - \beta_2)(-x_i) \\
\frac{\partial \ell_k}{\partial \beta_2} &= -\frac{1}{2\sigma^2} \sum_{i=k+1}^{N} 2(y_i - \alpha_2 x_i - \beta_2)(-1) \\
\frac{\partial \ell_k}{\partial \sigma^2} &= -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \left(\sum_{i=1}^{k} (y_i - \alpha_1 x_i - \beta_1)^2 + \sum_{i=k+1}^{N} (y_i - \alpha_2 x_i - \beta_2)^2 \right).
\end{aligned}$$

We note that the expressions for $\frac{\partial \ell_k}{\partial \alpha_1}$, $\frac{\partial \ell_k}{\partial \beta_1}$, $\frac{\partial \ell_k}{\partial \alpha_2}$ and $\frac{\partial \ell_k}{\partial \beta_2}$ are all similar to the ones developed in the previous section and independent of σ_k^2. We note here that the maximum likelihood estimates are

$$\hat{\alpha}_{1,k} = \frac{k \sum_{i=1}^{k} (y_i x_i) - \sum_{i=1}^{k} y_i \sum_{i=1}^{k} x_i}{k \sum_{i=1}^{k} x_i^2 - \left(\sum_{i=1}^{k} x_i \right)^2},$$

$$\hat{\beta}_{1,k} = \frac{1}{k} \sum_{i=1}^{k} (y_i - \hat{\alpha}_{1,k} x_i),$$

$$\hat{\alpha}_{2,k} = \frac{(N-k) \sum_{i=k+1}^{N} (y_i x_i) - \sum_{i=k+1}^{N} y_i \sum_{i=k+1}^{N} x_i}{(N-k) \sum_{i=k+1}^{N} x_i^2 - \left(\sum_{i=k+1}^{N} x_i \right)^2}$$

and

$$\hat{\beta}_{2,k} = \frac{1}{N-k} \sum_{i=k+1}^{N} (y_i - \hat{\alpha}_{2,k} x_i).$$

To calculate the maximum likelihood estimate for σ^2, we solve

$$-\frac{N}{2\sigma^2}+\frac{1}{2\sigma^4}\left(\sum_{i=1}^{k}(y_i-\alpha_1 x_i-\beta_1)^2+\sum_{i=k+1}^{N}(y_i-\alpha_2 x_i-\beta_2)^2\right)=0.$$

Then the maximum likelihood estimate $\hat{\sigma}_k^2$ is given as

$$\hat{\sigma}_k^2=\frac{1}{N}\left(\sum_{i=1}^{k}(y_i-\hat{\alpha}_{1,k}x_i-\hat{\beta}_{1,k})^2+\sum_{i=k+1}^{N}(y_i-\hat{\alpha}_{2,k}x_i-\hat{\beta}_{2,k})^2\right).$$

When the maximum likelihood estimates are substituted into ℓ_k, we get the following expression

$$\ell_k(\boldsymbol{y})=-\frac{N}{2}\log 2\pi-\frac{N}{2}\log\hat{\sigma}_k^2-\frac{N}{2}.$$

We then calculate the value of the log-likelihood ratio

$$\lambda_k(\boldsymbol{y})=\ell_k(\boldsymbol{y})-\ell(\boldsymbol{y}),$$

which is given as

$$\lambda_k(\boldsymbol{y})=-\frac{N}{2}\log 2\pi-\frac{N}{2}\log\hat{\sigma}_k^2-\frac{N}{2}-(-\frac{N}{2}\log 2\pi-\frac{N}{2}\log\hat{\sigma}^2-\frac{N}{2})$$

and can be simplified to give the expression

$$\lambda_k(\boldsymbol{y})=\frac{N}{2}\log\hat{\sigma}^2-\frac{N}{2}\log\hat{\sigma}_k^2. \tag{5.5}$$

And the likelihood ratio can be written

$$\Lambda_k=\left(\frac{\hat{\sigma}^2}{\hat{\sigma}_k^2}\right)^{N/2}. \tag{5.6}$$

5.3 Variances equal and known

If σ^2 is known, we have the following parameters under the null and alternative

$$H_0:(\alpha,\beta)\qquad\qquad H_a:(\alpha_1,\beta_1),(\alpha_2,\beta_2).$$

Under H_0, the likelihood function is given as

$$L(\boldsymbol{y};\alpha,\beta)=\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(y_i-\alpha x_i-\beta)^2}$$

and can be rewritten in the form

$$L(\boldsymbol{y};\alpha,\beta)=\left(\frac{1}{2\pi\sigma^2}\right)^{N/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\alpha x_i-\beta)^2\right).$$

To find the maximum, we will use the log-likelihood which is given as

$$\ell(\boldsymbol{y}; \alpha, \beta) = -\frac{N}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \alpha x_i - \beta)^2.$$

The maximum is achieved when the partial derivatives are equal to zero. We solve the following equations to find the maximum :

$$\frac{\partial \ell}{\partial \alpha} = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} 2(y_i - \alpha x_i - \beta)(-x_i) = 0,$$
$$\frac{\partial \ell}{\partial \beta} = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} 2(y_i - \alpha x_i - \beta)(-1) = 0.$$

Then the maximum likelihood estimate for α is given by solving the equation

$$\frac{1}{\sigma^2} \sum_{i=1}^{N} (y_i x_i - \alpha x_i^2 - \beta x_i) = 0$$

for α. Taking the term involving α to the other side, we get

$$\alpha \frac{\sum_{i=1}^{N} x_i^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{N} (y_i x_i - \beta x_i).$$

Then the maximum likelihood estimate $\hat{\alpha}$ is

$$\hat{\alpha} = \frac{\sum_{i=1}^{N} x_i (y_i - \hat{\beta})}{\sum_{i=1}^{N} x_i^2}. \tag{5.7}$$

Next, we find the maximum over β by solving the equation

$$\frac{1}{\sigma^2} \sum_{i=1}^{N} (y_i - \alpha x_i - \beta) = 0.$$

Isolating the β terms, we get

$$\frac{N\beta}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{N} (y_i - \alpha x_i).$$

And, finally, the maximum is achieved at

$$\hat{\beta} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{\alpha} x_i). \tag{5.8}$$

Then the maximum of the log-likelihood function under H_0 is given by

$$\ell(\boldsymbol{y}; \hat{\alpha}, \hat{\beta}) = -\frac{N}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \hat{\alpha} x_i - \hat{\beta})^2,$$

where $\hat{\alpha}$ and $\hat{\beta}$ are given by (5.7) and (5.8).

Under the alternative, there is one shift in regime. If the location of the shift is at some known time k, then the likelihood function is given as

$$L_k(\boldsymbol{y}) = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \alpha_1 x_i - \beta_1)^2} \prod_{i=k+1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \alpha_2 x_i - \beta_2)^2}.$$

With the log-likelihood function

$$\ell_k(\boldsymbol{y}) = -\frac{N}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{k} (y_i - \alpha_1 x_i - \beta_1)^2 - \frac{1}{2\sigma^2} \sum_{i=k+1}^{N} (y_i - \alpha_2 x_i - \beta_2)^2.$$

To find the maximum, we solve the partial derivatives for zero

$$\begin{aligned}
\frac{\partial \ell_k}{\partial \alpha_1} &= -\frac{1}{2\sigma^2} \sum_{i=1}^{k} 2(y_i - \alpha_1 x_i - \beta_1)(-x_i) = 0, \\
\frac{\partial \ell_k}{\partial \beta_1} &= -\frac{1}{2\sigma^2} \sum_{i=1}^{k} 2(y_i - \alpha_1 x_i - \beta_1)(-1) = 0, \\
\frac{\partial \ell_k}{\partial \alpha_2} &= -\frac{2}{2\sigma^2} \sum_{i=k+1}^{N} 2(y_i - \alpha_2 x_i - \beta_2)(-x_i) = 0, \\
\frac{\partial \ell_k}{\partial \beta_2} &= -\frac{2}{2\sigma^2} \sum_{i=k+1}^{N} 2(y_i - \alpha_2 x_i - \beta_2)(-2) = 0.
\end{aligned}$$

Next, we find the maximum likelihood estimates for α_1 and β_1. The maximum likelihood estimate for α_1 is calculated by solving the equation

$$\frac{1}{\sigma^2} \sum_{i=1}^{k} (y_i - \alpha_1 x_i - \beta_1)(x_i) = 0.$$

The solution to this equation has been derived in previous sections and is

$$\hat{\alpha}_{1,k} = \frac{\sum_{i=1}^{k} x_i (y_i - \hat{\beta}_1)}{\sum_{i=1}^{k} x_i^2}.$$

Similarly, the expression for $\hat{\alpha}_{2,k}$ can be written as

$$\hat{\alpha}_{2,k} = \frac{\sum_{i=k+1}^{N} x_i(y_i - \hat{\beta}_2)}{\sum_{i=k+1}^{N} x_i^2}.$$

To find the maximum likelihood estimate for β_1, we solve

$$\frac{1}{\sigma^2}\sum_{i=1}^{k}(y_i - \alpha_1 x_i - \beta_1) = 0.$$

Again, from the previous sections we have

$$\hat{\beta}_{1,k} = \frac{1}{k}\sum_{i=1}^{k}(y_i - \hat{\alpha}_{1,k}x_i)$$

and

$$\hat{\beta}_{2,k} = \frac{1}{N-k}\sum_{i=k+1}^{N}(y_i - \hat{\alpha}_{1,k}x_i).$$

Then the maximum of the log-likelihood ratio under the alternative can be written

$$\ell_k(\boldsymbol{y}) = -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{k}(y_i - \hat{\alpha}_{1,k}x_i - \hat{\beta}_{1,k})^2 - \frac{1}{2\sigma^2}\sum_{i=k+1}^{N}(y_i - \hat{\alpha}_{2,k}x_i - \hat{\beta}_{2,k})^2.$$

The generalized log-likelihood ratio for our test is

$$\lambda_k = \log\Lambda_k = \ell_k(\boldsymbol{y}) - \ell(\boldsymbol{y}).$$

Substituting the expressions for ℓ_k and ℓ, we get

$$\begin{aligned}\lambda_k = &-\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{k}(y_i - \hat{\alpha}_{1,k}x_i - \hat{\beta}_{1,k})^2 - \frac{1}{2\sigma^2}\sum_{i=k+1}^{N}(y_i - \hat{\alpha}_{2,k}x_i - \hat{\beta}_{2,k})^2 \\ &- \left(-\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \hat{\alpha}x_i - \hat{\beta})^2\right).\end{aligned}$$

So that the generalized log-likelihood is

$$\lambda_k = \frac{1}{2\sigma^2}\left(\sum_{i=1}^{N}(y_i - \hat{\alpha}x_i - \hat{\beta})^2 - \sum_{i=1}^{k}(y_i - \hat{\alpha}_{1,k}x_i - \hat{\beta}_{1,k})^2 - \sum_{i=k+1}^{N}(y_i - \hat{\alpha}_{2,k}x_i - \hat{\beta}_{2,k})^2\right).$$

5.4 Results

We are interested in whether or not a regime shift occurs for the average credits at graduation. Table 1 shows the values under consideration. To be more specific, the y_i will come from the means column of Table 1 and the x_i are the years associated with each average. Figure 8 shows the log-likelihood values for each k between 1997 and 2005 and under each of the three scenarios developed previously[1]. All three show that, if the value for ℓ_{2001} is large, there was a regime shift in the year 2001.

5.4.1 Critical Values

When we assume that the split point k is unknown, the critical values cannot be assumed to come from the χ^2 distribution. Critical values for these tests will be derived using a resampling method. In this method, we will use the model

$$y_i = \beta + \alpha x_i + \varepsilon_i, \quad 1 \leq i \leq 10,$$

where the ε_i will be independent normally distributed random numbers with mean 0 and variance $\sigma^2 = 0.04$. Then, we calculate $\{\lambda_k : 2 \leq k \leq 8\}$. Under the three scenarios discussed in sections 5.1 through 5.3. This process is repeated n times, which we denote

$$\varphi_j = \max_{2 \leq k \leq 8} \lambda_k, \quad 1 \leq j \leq n.$$

The φ_j are then ordered, according to each of the three scenarios, and the critical values for the test are given as $\varphi_{(1-\alpha)n}$. In our case we take $n = 1000$.

	$\alpha = 0.10$	$\alpha = 0.05$
Variance Unequal and Unknown	11.596	14.052
Variance Equal and Unknown	8.011	9.174
Variance Equal and Known	12.886	14.468

Table 7: Critical Values for Regime Change Test

The critical values can be found in Table 7. Figure 9 shows log-likelihood plots in each of the three scenarios with critical values. Critical values for $\alpha = 0.10$ are in red, and values for $\alpha = 0.05$ are in blue. We see that only in the case where the variances are unknown and unequal do we reject for $\alpha = 0.10$. This indicates that, under most circumstances, only one model is necessary to summarize the rise in credits to graduation. We note here that the scenario where variances are unequal and unknown is the more realistic of the three. It has been shown previously that we can assume either an exponential or gamma distribution and that the scale parameter is not staying constant over the time period considered. This means that the variance is also assumed to be changing over that same time period.

Notice that the observed values in Figure 7 for 2002 through 2006 fall nearly along the same line, whereas the values from 1997 through 2001 are not rising as steadily. This seems

[1] We note that in the third scenario where the variances are equal and known, $\sigma^2 = 0.04$. This value was taken from the variance of the original samples divided by n_i, or the number of students in each year.

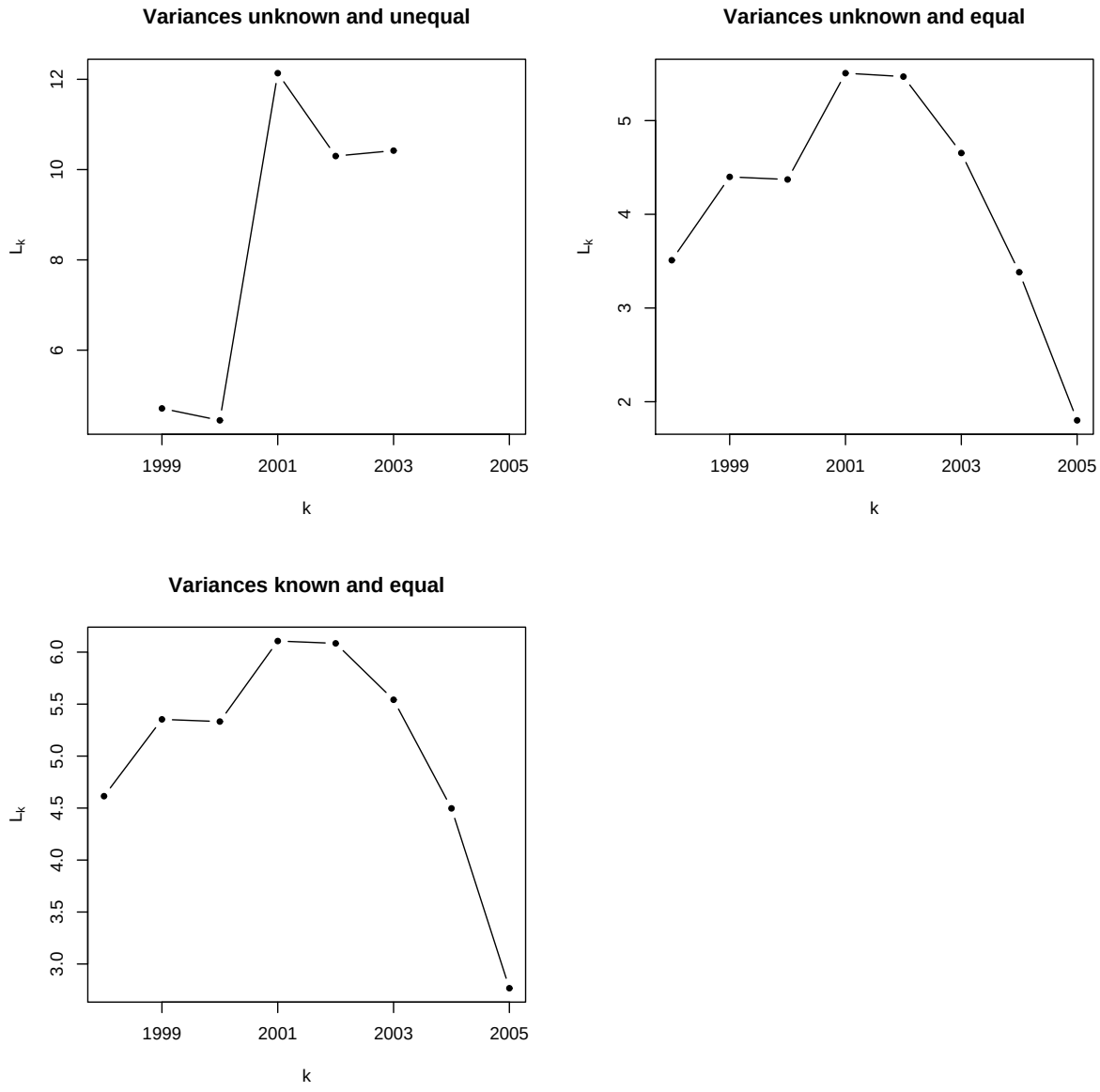

Figure 8: Log-likelihood Values

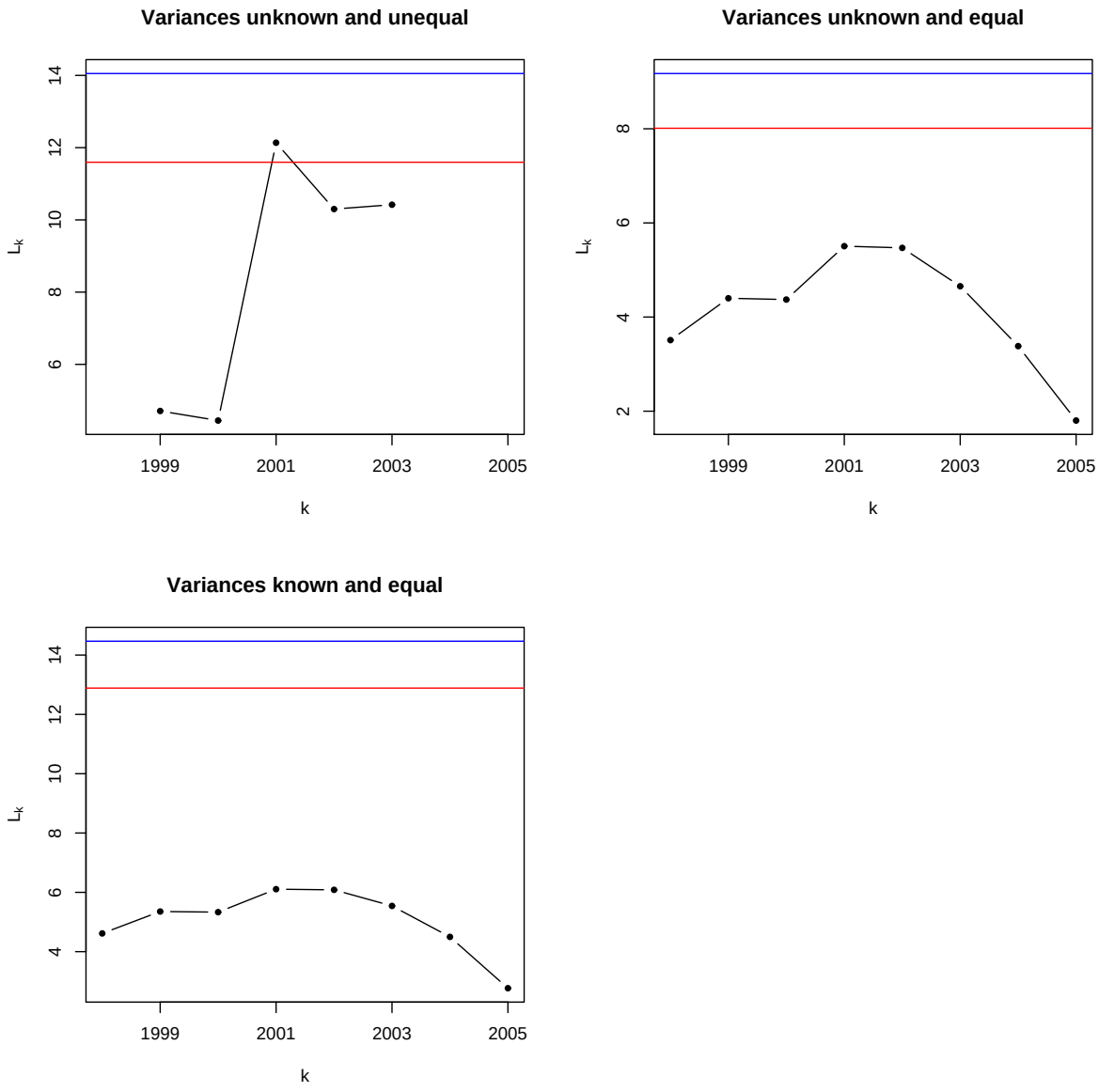

Figure 9: Log-likelihood Values with Critical Values

to indicate that the regime change test should have concluded that two regression lines are necessary. To investigate further, we will assume two different scenarios and compare the R^2 value for each. In each scenario, we use the normal regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

In the first scenario, we assume that one model appropriately describes the trend, so that we have the following definitions for $\boldsymbol{Y}$ and $\boldsymbol{X}$

$$\boldsymbol{Y} = \begin{pmatrix} 22.119 \\ 21.828 \\ 22.465 \\ 22.359 \\ 23.009 \\ 22.985 \\ 23.533 \\ 24.009 \\ 24.544 \\ 25.148 \end{pmatrix}$$

and

$$\boldsymbol{X} = \begin{pmatrix} 1997 & 1 \\ 1998 & 1 \\ 1999 & 1 \\ 2000 & 1 \\ 2001 & 1 \\ 2002 & 1 \\ 2003 & 1 \\ 2004 & 1 \\ 2005 & 1 \\ 2006 & 1 \end{pmatrix}.$$

In the second scenario, we assume that there are actually two models for the data. We keep the same definition for $\boldsymbol{Y}$ and take

$$\boldsymbol{X} = \begin{pmatrix} 1997 & 1 & 0 & 0 \\ 1998 & 1 & 0 & 0 \\ 1999 & 1 & 0 & 0 \\ 2000 & 1 & 0 & 0 \\ 2001 & 1 & 0 & 0 \\ 0 & 0 & 2002 & 1 \\ 0 & 0 & 2003 & 1 \\ 0 & 0 & 2004 & 1 \\ 0 & 0 & 2005 & 1 \\ 0 & 0 & 2006 & 1 \end{pmatrix}.$$

Using least squares regression, we get

$$\boldsymbol{\beta}' = (0.348, -674.207) \tag{5.9}$$

with $R^2 = 0.9319$ for the first scenario and

$$\boldsymbol{\beta}' = (0.231, -439.683, 0.533, -1045.667) \tag{5.10}$$

with $R^2 = 1$ for the second. We would say that the second scenario of two models results in a better fit judging from the increase in R^2. The regime change test may not agree because it does not recognize the small difference in R^2 as being significant. The results from performing regression under both assumptions show that the average number of credits students are graduating with is increasing between 0.348 (5.9) and 0.533 (5.10) starting in 2001. While this shows a rise in credits at graduation, it is not an incredibly large increase from year to year. At worst, it shows that after 2001 students will take an extra three credit course every six years. It is also unclear whether or not this trend will continue forever.

Figure 10 shows the observed average excess credits from 1997 through 2006 along with predicted values using one model and two models. These prediction lines show that by 2010 students should be taking between 25 and 27 extra credits on average. This equates to roughly two extra semesters (or one academic year) for the average student. It should be noted that we make no assumptions about how many terms it will take students to graduate. We assume that, due to lifestyle differences, terms to graduation and credits to graduation can be considered independent.

5.5 Conclusions and further research

The main conclusion of this project is that students are indeed taking an increasing number of credits to graduate; but that the increase is not very large. Further research could be performed to determine the cause of the increase. Information may or may not be available to determine the cause. For example, economic factors may be the cause for the increase. It is widely known that when the economy is not performing well students continue to attend college trying to wait out the downturn, this information is not easily obtained. There is also an indication that students are working more while studying. This fact might force students to take longer to graduate and possibly to change majors more, forcing students to take more credits. However, the institution does not have data on how much or how often students work making this a hard variable to study. There are also indications that colleges have a strong effect on average credits to graduation. A small change in a department or college's curriculum could affect the average credits to graduation for the institution. Also, shifts in student population from one college or department to another could affect the change in average credits at graduation. The number of double/multiple majors and the number of students working on combined BS/MS degrees should also be investigated as a reason for the increase. It is our conclusion that although credits at graduation are increasing, the explanation for the increase is complicated and merits further research.

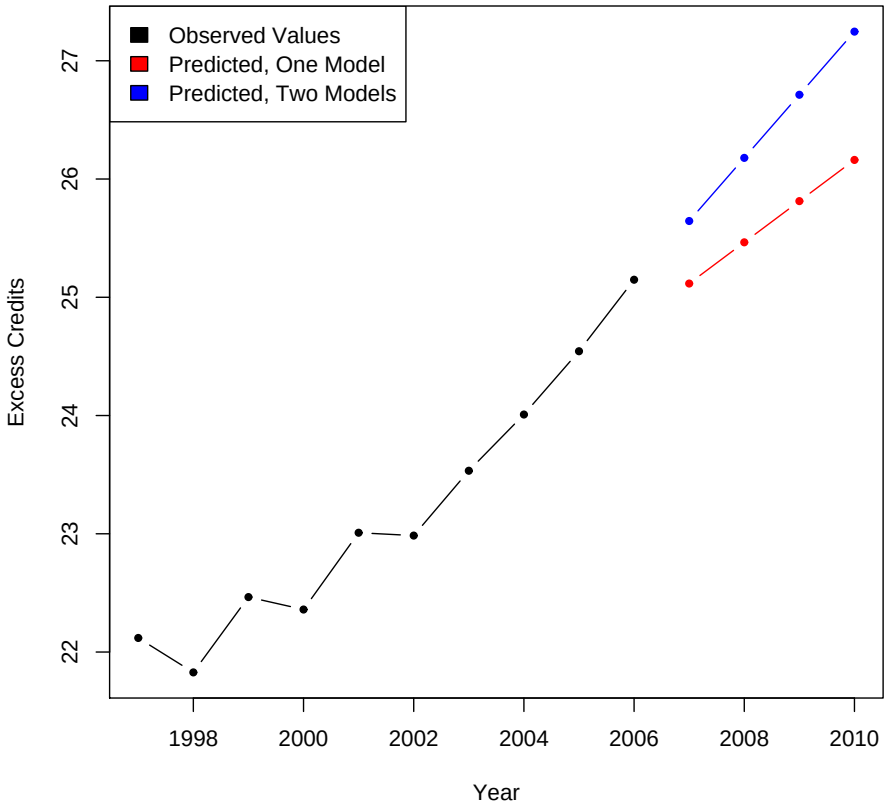

Figure 10: Predicted Values for Excess Credits

A Code Written for Data Extraction and Analysis

This section includes most of the code written for this project. Generally, Perl was used for data extraction from the MySQL databases maintained by the Office of Budget and Analysis. Some of the data cleaning was done in the Perl scripts as the data came out of the databases. All of the data analysis and some of the data cleaning was done in R. Some of the more mundane code was left out, for example, code creating most of the plots and the code for performing the tests for each year.

A.1 Data extraction and cleaning

This package was written to make the job of extraction easier. Moving these functions into a package made the extraction cleaner.

```
package MastersTools;
sub getSemTables {
    my ($startyear,$endyear) = @_;
    my $dbh = DbTools::connectDb(@DbTools::sherman,"obia_prod");
    my $tables_sql = DbTools::getDbHandle($dbh,"show tables");
    $tables_sql->execute();

    my %tables = ();

    while( my $t = $tables_sql->fetchrow_array() ){
        if( index($t,'dem') > -1 ){
            $tables{$t} = 1;
        }
    }

    #figure out tables to use and create the rest of the select
    # and join statements.
    my $num_sem = 4*(1998 - $startyear);
    $num_sem += 3*($endyear - 1998);
    if( $startyear > 1998 ){
        $num_sem = 3*($endyear - $startyear);
    }
    if( $endyear <= 1998 ){
        $num_sem = 4*(1998 - $startyear);
    }


    my @terms = ('W','S','U','F');
    my @semesters = ('S','U','F');
    my $thisyr = $startyear;

    my @tbls = ();

    for(my $i = $startyear;$i <= $endyear;$i++){
        if($i == $startyear){
```

```
            push @tbls,'F' . substr($i,2,2);
        }elsif($i == $endyear){
            push @tbls,'S' . substr($i,2,2);
        }elsif($i <= 1998){
            for(my $j=0;$j < 4;$j++){
                push @tbls,$terms[$j] .  substr($i,2,2);
            }
        }else{
            for(my $j=0;$j < 3;$j++){
                push @tbls,$semesters[$j] .  substr($i,2,2);
            }
        }

    }
    push @tbls,'U' . substr($endyear,2,2);

    my @results = ();

    foreach my $i (@tbls){
        my $extract = 'E';
        if( !  defined( $tables{'dem'.$i.$extract}) ){
            $extract = 'C';
        }
        push @results, 'dem' .  $i .  $extract;
    }

    return @results;
}
sub getDegTables {
    my ($startyear,$endyear) = @_;

    my @results = ();
    for(my $i=$startyear;$i<$endyear;$i++){
        my $nxtyr = $i+1;
        push @results,'deg' .  substr($i,2,2) .  substr($nxtyr,2,2);
    }

    return @results;
}
return 1;
```

The following code does the bulk of the data extraction and cleaning. Usernames and passwords are located in another package that has not been included here for security.

```
#!/usr/bin/perl -w
use strict;
use DBI;
use lib qw(/home/jmorris/Documents/scripts/);
use lib qw(/home/jmorris/MastersProject/scripts/);
use tools;
```

```
use mtools;

my $dbh = DbTools::connectDb(@DbTools::sherman,"obia_prod");

my @degreeTbls = MastersTools::getDegTables(1993,2006);

open OUT,">../data/credits.csv";

print OUT "year,degree,terms,total_credits,transfer_credits\n";

foreach my $i (@degreeTbls){
    my $yr = substr($i,5,2) + 1900;
    if( $yr < 1930 ){
        $yr += 100;
    }
    my $mincredits = 122;
    my $crfactor = 1;
    my $tr = 0;
    if( $yr < 1999 ){
        $mincredits = $mincredits * 3 / 2;
        $crfactor = 2/3;
        $tr = 1;
    }
    my $query = "select `DEGREE TYPE`,NBRTERMSATTENDED,TOTCUM,
        TOTTRANSFER,TOTTEST,TOTOTHER from $i ";
    $query .= "where TOTCUM+$tr >= $mincredits";
    my $data_sql = DbTools::getDbHandle($dbh,$query);
    $data_sql->execute();

    my $nstudents = 0;
    while(( my ($degree,$terms,$credits,$transfer,$test,$other) = $data_sql->fetchrow_array() )){
        if( $tr == 1 ){
            if( defined($transfer) ){
                $credits += $transfer;
            }
            if( defined($test) ){
                $credits += $test;
            }
            if( defined($other) ){
                $credits += $other;
            }
        }
        if( $credits >= $mincredits ){
            $credits -= $mincredits;
        }
        $credits *= $crfactor;
        $terms *= $crfactor;
        if(($credits >= 0)&&(( substr($degree,0,1) eq 'B' )||( substr($degree,0,1) eq 'H' ))){
            print OUT "$yr,$degree,$terms,$credits";

            if( defined($transfer) ){
```

```
                print OUT ",$transfer\n";
            }else{
                print OUT ",\n";
            }
            $nstudents++;
        }
    }
}

close OUT;
```

This code was written in R and takes care of the last part of the data cleaning.

```
#here we define who gets to stay in the data set and who doesn't

tmp <- data.frame(all_grads,keep=0)
acad.level <- c('01','02','03','04','05')
for( i in 9:11 ){
    for( j in seq(along=acad.level) ){
        tmp$keep[ tmp[,i] == acad.level[j] ] <- 1
    }
}

yrs <- c(1997:2006)

for( i in seq(along=yrs) ){
    fac <- 1
    if( yrs[i] < 1999 ){
        fac <- 3/2
    }
    tmp$keep[ tmp$year == yrs[i] & tmp$total_credit < 120*fac ] <- 0

    tmp$total_credit[ tmp$year == yrs[i] ] <- (1/fac) * (tmp$total_credit[ tmp$year == yrs[i] ]
        - (120*fac))
    tmp$terms[ tmp$year == yrs[i] ] <- (1/fac) * tmp$terms[ tmp$year == yrs[i] ]
}

tmp <- tmp[ tmp$keep == 1, ]
final.all_grads <- data.frame(tmp[, c(1:8)])
```

A.2 Testing the Exponential Assumption

Code written in R testing data for exponentiality.

```
exp.chisq <- function(x,...){
    n <- length(x)
    h <- hist(x,plot=FALSE,...)
    b <- h$breaks[-1]
```

```
    K <-length(b)
    prob <- numeric(K)
    theta <- mean(x)

    prob[1] <- pexp(b[1],1/theta)
    for( i in 2:K ){
        prob[i] <- pexp(b[i],1/theta) - pexp(b[i-1],1/theta)
    }
    prob[K] <- pexp(b[K-1],1/theta,FALSE)

    prob[ is.na(prob) ] <- 0

    Q <- sum( ((h$counts - n*prob)^2 / n*prob) )

    list(counts=h$counts,p=prob,breaks=b,Q=Q,df=(K-2),
        p.value=pchisq(Q,df=(K-2),lower.tail=FALSE))
}

exp.unif <- function(x,mn=mean(x),shift=0){
    x.s <- as.numeric(levels(factor(x))) #removes all duplicates
    n <- length(x.s)
    Sn <- sum(x.s)
    ks <- numeric(n-1)
    cvm <- 0

    for( k in 1:(n-1) ){
        Sk <- sum(x.s[1:k])
        tmp <- (Sk / Sn) - (k/n)
        ks[k] <- abs(tmp)
        cvm <- cvm + tmp^2
    }

    F <- ecdf(x)
    D <- abs( pexp(x - shift,rate=1/(mn - shift)) - F(x) )

    CM <- 1/(12*n) + sum( (pexp(x.s - shift,rate=1/mean(x.s - shift)) - (1:n - 0.5)/n )^2 )

    list(ks=(sqrt(n)*max(ks)),cvm=cvm,D=max(D),CM=((1 + 0.16/n)*CM),n=n)
}

ttot <- function(x) {
    X <- as.numeric(levels(factor(x))) #removes all duplicates
    n <- length(X)
    T <- numeric(n - 1)
    denom <- 0

    for( i in 1:(n-1) ){
        denom <- denom + (n - i - 1)*(X[i+1] - X[i])
    }

    s <- 0
```

```
    for( i in 1:(n-1) ){
        s <- s + (n - i - 1)*(X[i+1] - X[i])
        T[i] <- s / denom
    }

    br1 <- sqrt(n) * max( abs(T - (1:(n-1))/n) )
    br2 <- sum( (T - (1:(n-1))/n)^2 )

    list(t1=br1,t2=br2)
}
```

Code written in R to plot the empirical density function against the density function for the Gamma distribution.

```
y <- data.frame( matrix(0,ncol=3,nrow=11) )
names(y) <- c('M','kappa','theta')
row.names(y) <- c(1997:2006,'All')

x11(w=9,h=9)
par(mfrow=c(2,2))
for( i in 1:4 ){
    x <- sort(total_credit[ year == i + 1996 ]) + 1
    x.den <- density(x,from=1)

    s <- log(mean(x)) - (1/length(x))*sum( log(x) )
    cat(s,'\n')
    k <- (0.5000876 + 0.1648852*s - 0.0544274*s^2) / s
    th <- mean(x) / k

    y[i,] <- c(s,k,th)

    plot(x.den,type='l',col='red',lwd=1,main=(i + 1996),xlab='Excess Credits',
        ylab=expression(f[0](x)))
    points( x,dgamma(x,shape=k,scale=th),type='l',lwd=2,lty=1,col='blue')
}
dev.copy2eps(device='x11',file='../writeup/gamma1.eps',horizontal=FALSE)
dev.off()

x11(w=8,h=10)
par(mfrow=c(3,2))
for( i in 5:10 ){
    x <- sort(total_credit[ year == i + 1996 ]) + 1
    x.den <- density(x,from=1)

    s <- log(mean(x)) - (1/length(x))*sum( log(x) )
    cat(s,'\n')
    k <- (0.5000876 + 0.1648852*s - 0.0544274*s^2) / s
    th <- mean(x) / k
```

```
    y[i,] <- c(s,k,th)

    plot(x.den,type='l',col='red',lwd=1,main=(i + 1996),xlab='Excess Credits',
        ylab=expression(f[0](x)))
    points( x,dgamma(x,shape=k,scale=th),type='l',lwd=2,lty=1,col='blue')
}
dev.copy2eps(device='x11',file='../writeup/gamma2.eps',horizontal=FALSE)
dev.off()

x11(w=9,h=9)
x <- sort(total_credit) + 1
x.den <- density(x,from=1)

s <- log(mean(x)) - (1/length(x))*sum( log(x) )
cat(s,'\n')
k <- (0.5000876 + 0.1648852*s - 0.0544274*s^2) / s
th <- mean(x) / k

y[11,] <- c(s,k,th)

plot(x.den,type='l',col='red',lwd=1,main='1997 - 2006',xlab='Excess Credits',
    ylab=expression(f[0](x)))
points( x,dgamma(x,shape=k,scale=th) , type='l',lwd=2,lty=1,col='blue')
dev.copy2eps(device='x11',file='../writeup/gamma3.eps',horizontal=FALSE)
dev.off()
```

A.3 Testing for Homogeneity of the Mean

Code written in R.

```
theta.hat <- function(x,index,sub=c(1997:2006),...){
    mns <- numeric(length(sub))
    for( i in seq(along=sub) ){
        mns[i] <- mean( x[ index == sub[i] ],...  )
    }
    ret <- mean(mns)
    return( ret )
}

loglke.2 <- function(x,index,sub=c(1997:2006)){
    tot.mean <- theta.hat(x,index,sub)
    n <- numeric(length(sub))
    ss <- numeric(length(sub))
    for( i in seq(along=sub) ){
        ss[i] <- ((theta.hat(x,index,sub[i]) - tot.mean)/tot.mean )^2
        n[i] <- length( x[ index == sub[i] ] )
    }
```

```
    sum( n*ss )
}

loglke <- function(x,index){
    grps <- factor(index)
    k <- length(levels(grps))
    mns <- sapply(split(x,grps),mean)
    lns <- sapply(split(x,grps),length)
    mn <- mean(x)

    null <- sum(dexp(x,rate=mn,log=T))

    alt <- numeric(k)
    for( i in 1:k ){
        cat(x[index == levels(grps)[i]],'\n')
        alt[i] <- sum(dexp(x[index == levels(grps)[i]],rate=mns[i],log=T))
    }

    cat(null,'\n',alt,'\n')
    a <- null - sum(alt)

    alt <- 0
    for( i in seq(along=levels(grps)) ){
        alt <- alt + lns[i]*log(mns[i])
    }
    null <- length(x)*log(mn)

    b <- null - alt

    list(a=a,b=b)

}

get.pvalue <- function(x,index,sub=c(1997:2006),...){
    ll <- loglke(x,index,sub,...)
    p.value <- pchisq(-2*ll,length(sub)-1,lower.tail=F)
    list(ll=ll,p.value=p.value)
}
get.pvalue.approx <- function(x,index,sub=c(1997:2006)){
    ll <- loglke.2(x,index,sub)
    pchisq(ll,length(sub)-1,lower.tail=F)
}
```

A.4 Regime Change Functions

Code written in R.

```
regime.likelihood <- function(x,y,f=NULL,v=0.4) {
```

```
    N <- length(x)
    idx <- x
    if( !  is.null(f) ){
        N <- length(f)
        idx <- f
    }

    RSS1 <- numeric(N)
    RSS2 <- numeric(N)
    L <- numeric(N)

    RSSp <- numeric(N)
    Lp <- numeric(N)

    L.kn <- numeric(N)

    RSS <- sum(lm(y   x)$residuals^2)
    for( k in 2:(N-1) ){
        RSS1[k] <- sum(lm(y   x,subset=(x <= idx[k]))$residuals^2)
        RSS2[k] <- sum(lm(y   x,subset=(x > idx[k]))$residuals^2)
        RSSp[k] <- (RSS1[k] + RSS2[k])

        L[k] <- (N/2)*log(RSS/N) - (k/2)*log(RSS1[k]/k) - ((N-k)/2)*log(RSS2[k]/(N-k))
        Lp[k] <- (N/2)*(log(RSS/N) - log(RSSp[k]/N))
        L.kn[k] <- (1/(2*v))*( RSS - RSS1[k] - RSS2[k])
    }

    list(RSS=RSS,RSS1=RSS1[c(-1,-N)],RSS2=RSS2[c(-1,-N)],
        RSSp=RSSp[c(-1,-N)],L=L[c(-1,-N)],Lp=Lp[c(-1,-N)],
        L.kn=L.kn[c(-1,-N)])

}

regime.critical <- function(x,y,N=500,er.var=0.4,...){
    cr <- data.frame(L=numeric(N),Lp=numeric(N),L.kn=numeric(N))
    n <- length(x)

    for(i in 1:N ){
        errors <- rnorm(n,mean=0,sd=sqrt(er.var))
        y.  <- y + errors
        tmp <- regime.likelihood(x,y.,...)

        cr$L[i] <- max(tmp$L[tmp$L < Inf])
        cr$Lp[i] <- max(tmp$Lp)
        cr$L.kn[i] <- max(tmp$L.kn)
    }

    cr <- apply(cr,2,sort)

    list(a90=cr[N*0.9,],a95=cr[N*0.95,])
}
```

References

[1] Bain, Lee, and Max Engelhardt. **Introduction to Probability and Mathematical Statistics**. 2nd ed. Belmont: Duxbury Press, 1992.

[2] Johnson, Richard A., and Dean W. Wichern. **Applied Multivariate Statistical Analysis**. 5th ed. Upper Saddle River, NJ: Pearson Education, 2002.

[3] Quandt, Richard E.. "The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes." **Journal of the American Statistical Association** 53(1958): 873-880.

[4] Quandt, Richard E.. "Test of the Hypothesis that a Linear Regression System Obeys Two Separate Regimes." **Journal of the American Statistical Association** 55(1960): 324-330.

[5] Lilliefors, Hubert W.. "On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown." **Journal of the American Statistical Association** 64(1969): 387-389.