Chapter 1

Monte-Carlo Simulations

In this chapter I will explore simulations of the bias of estimator (**??**) in comparison to the size of the sample estimated from, with respect to different values of k; by exploring 1-dimensional distributions and then progressing onto 2-dimensional. Firstly, the distributions considered will be analysed to determine if they satisfy the conditions **??**, **??** and **??** stated for Theorems **??** and **??** to hold. Then, I will explore the estimator of entropy for simulations of samples from certain distributions, for different values of k.

The motivation for these simulations is to explore the consistency of this estimator for different values of k; the relationship between the size of the bias of the estimator $\hat{H}_{N,k}$, $Bias(\hat{H}_{N,k})$, and the sample size, N. Throughout this analysis we will be considering the absolute value of this bias, since when taking its logarithm, we need a positive value. Using Theorem **??**, we can write that the bias of the estimator approaches 0 as $N \to \infty$. This is because we can write $Bias(\hat{H}_{N,k}) = \mathbb{E}(\hat{H}_{N,k}) - H$, which in equation (**??**) implies $Bias(\hat{H}_{N,k}) \to 0$ as $N \to \infty$. Thus, there must be a type of inverse relationship between the modulus of the bias of the estimator, $|Bias(\hat{H}_{N,k})|$, and N. We believe this relationship is of the form;

$$|Bias(\hat{H}_{N,k})| = \frac{c}{N^a} \tag{1.1}$$

for $a, c > 0$ [**?**, **?**]. By taking the logarithm of this, we can generate a linear relationship, which is easier to analyse, and is given by;

$$\begin{aligned} log|Bias(\hat{H}_{N,k})| &\approx log(c) - a[log(N)] + \epsilon \\ &\approx \zeta - a[log(N)] \end{aligned} \tag{1.2}$$

where $\epsilon > 0$ is some small error term. I will investigate the consistency of this estimator for a sample from a specified distribution, dependent on the value of k, this mean finding the optimum value of k for which $|Bias(\hat{H}_{N,k})| \to 0$ for $N \to \infty$. For the relationship in equation (1.1), this will happen for larger values of a and relatively small c, as $N \to \infty$. As previously mentioned, there is evidence supporting that the bias becomes either of order $(\frac{1}{N})^a$ (equation (**??**)) or $(\frac{k}{N})^a$ (equation (**??**)). This leads to also examining the dependence of c / ζ on the value of k.

As I wish to consider the difference in accuracy of the estimator when using different values of k, let us denote the approximate values for a and c dependent on k as a_k and c_k.

I will conduct a range of analysis, for each distribution, to consider how this estimator acts in reality, the process of analysis will be as follows;

1. Create a summary table of the mean absolute value of the bias of the estimator for $N = 100, 25000$ and 50000 for all values of k that satisfy Condition **??**. I could also consider the variance of the bias at the values of N stated above, for all applicable values of k. However, we will find that the $Var|Bias(\hat{H}_{50000,k})| \to 0$ for $k \to 10$, by the definition of the estimator using the nearest neighbour method. Taking a larger k in the nearest neighbour method will produce less varied results, this is because more smoothing takes place for a larger k, eventually - if k is made large enough - the output will be constant and the variance negligible regardless of the inputted values. Thus, considering the variance of the bias of the estimator in comparison to k is not necessarily informative.

2. Graphical representations of the linear relationship shown in equation 1.2, of $log(N)$ against $log|Bias(\hat{H}_{N,k})|$ for sample sizes $N = 100, 200, 300, ..., 50000$ (which are taken 500 times and averaged), for each value of k.

3. Tabulate the results from the regression analysis; I will first discuss the coefficient of determination (R^2), this is a measure of how well the regression model describes the observed data [**?**]. Next I will consider the standard error/deviation of the model (σ^2), this is a measure of accuracy of predictions. Lastly, I will go onto consider the values of a_k and c_k from relationship shown in equation 1.1, for each k, which is the regression line that minimizes the sum of squared deviations (σ^2) of prediction.

4. Graphically compare the values of a_k and c_k for each k.

1.1 1-dimensional Gaussian/Normal Distribution

I will begin by exploring entropy of samples from the normal distribution $N(0, \sigma^2)$, where without loss of generality we can use the mean $\mu = 0$ and change the variance σ^2 as needed. The normal distribution has an exact formula to work out the entropy, given the variance σ^2. Using equation (**??**) and the density function for the normal distribution $f(x) = \frac{1}{\sqrt{(2\pi)}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right)$ for $x \in \mathbb{R}$, given $\mu = 0$.

We can write the exact entropy for the normal distribution, using equation (**??**);

$$
\begin{aligned}
H &= -\int_{x\in\mathbb{R}^d} f(x)log(f(x))dx \\
&= -\int_{\mathbb{R}} \frac{1}{\sqrt{(2\pi)}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right) log\left[\frac{1}{\sqrt{(2\pi)}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right)\right] dx \\
&= \int_{\mathbb{R}} \frac{1}{\sqrt{(2\pi)}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right) \left(log(\sqrt{(2\pi)}\sigma) + \frac{x^2}{2\sigma^2}\right) \\
&= \frac{log(\sqrt{(2\pi)}\sigma)}{\sqrt{(2\pi)}\sigma} \int_{\mathbb{R}} \exp\left(\frac{-x^2}{2\sigma^2}\right) dx + \frac{1}{2\sqrt{(2\pi)}\sigma} \int_{\mathbb{R}} \frac{x^2}{2\sigma^2} \exp\left(\frac{-x^2}{\sigma^2}\right) dx \\
&= log(\sqrt{(2\pi)}\sigma) + \frac{1}{2}
\end{aligned}
$$

Thus the exact entropy for the normal distribution is given by

$$H = log(\sqrt{(2\pi e)}\sigma) \tag{1.3}$$

I will first explore samples from 1-dimensional standard normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$, $N(0,1)$, to consider the behavior of the Kozachenko-Leonenko estimator. The exact entropy of this distribution is given by equation (1.3), with $\sigma^2 = 1$;

$$H = log(\sqrt{(2\pi e)}) \approx 1.418939 \tag{1.4}$$

Since, I am first considering the 1-dimensional normal distribution, the estimator takes the form in equation (**??**), which is given by;

$$\hat{H}_{N,k} = \frac{1}{N}\sum_{i=1}^{N} log\left[\frac{2\rho_{(k),i}(N-1)}{e^{\Psi(k)}}\right]$$

1.1.1 Estimator Conditions

The density of the normal distribution satisfies Conditions **??**, **??** and **??**, due to the below analysis. Firstly, to satisfy Condition **??**, for density function $f(x) = \frac{1}{\sqrt{(2\pi)}} \exp\left(\frac{-x^2}{2}\right)$ for $x \in \mathbb{R}$, given $\mu = 0$ and $\sigma^2 = 1$, it must be such that;

- f is bounded - obvious, since for any probability distribution we always have $f(x) \geq 0$, additionally for the normal distribution we have that $f(x) = \frac{1}{\sqrt{(2\pi)}} \exp\left(\frac{-x^2}{2}\right) < 0.4$, $\forall x \in \mathbb{R}$. Hence, f is bounded above and below; so bounded.
- f is m-times differentiable - using Hermite polynomials, defined as;

$$H_m(x) = (-1)^m e^{\frac{x^2}{2}} \frac{d^m}{dx^m}\left(e^{\frac{-x^2}{2}}\right)$$

multiplying this by the coefficient in the distribution of $f(x)$, $\frac{1}{\sqrt{(2\pi}}$, we then get;

$$\frac{d^m}{dx^m} f(x) = \frac{H_m(x)}{(-1)^m} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$
$$= \frac{H_m(x)}{(-1)^m} f(x)$$

where $\frac{H_m(x)}{(-1)^m}$ is a polynomial; thus f is m-times differentiable.

- $\exists r_* > 0$ and a Borel measurable function g_*, with $\|y - x\| \le r_*$ so that $\|f^{(t)}(x)\| \le g_*(x)f(x)$ and $\|f^{(m)}(x) - f^{(m)}(x)\| \le g_*(x)f(x)\|y - x\|^\eta$, for some g_* such that $\sup_{\{x:f(x)<\delta\}} g_*(x) = O(\delta^{-\epsilon})$ as $\delta \searrow 0$ for some $\epsilon > 0$.

  Since we are considering a 1-dimensional distribution, we can write the norms $\|\cdot\|$ as $|\cdot|$. Moreover, considering that for Theorems **??** and **??**, we have the value of $\beta \ge 2$; thus choosing $\beta = 2$, and since $m = \lfloor \beta \rfloor = \lfloor 2 \rfloor = 2 = \beta$ and $\eta = \beta - m$, we have that $\eta = 0$. Thus we need $|f^{(t)}(x)| \le g_*(x)f(x)$, which is obvious by above, in view of writing $|\frac{d^t}{dx^t} f(x)| = g_*(x)f(x)$, where we choose $g_*(x) = |\frac{H_t(x)}{(-1)^t}| = |H_t(x)|$, for $t = 1, 2, ..., m$, and $|f(x)| = f(x)$, since $f(x) > 0$. Also, g_* is a polynomial and is hence Borel measurable over $\mathbb{R}$, and for any polynomial we obviously have $\sup_{\{x:f(x)<\delta\}} g_*(x) = O(\delta^{-\epsilon})$ as $\delta \searrow 0$ for some $\epsilon > 0$. Additionally, we need $|f^{(m)}(x) - f^{(m)}(y)| \le g_*(x)f(x)|y - x|^0 = g_*(x)f(x)$. We currently have;

$$|f^{(m)}(x) - f^{(m)}(y)| = \left| \frac{H_m(x)}{(-1)^m} f(x) - \frac{H_m(y)}{(-1)^m} f(y) \right|$$
$$\le \left| \frac{H_m(x)}{(-1)^m} f(x) \right| + \left| \frac{H_m(y)}{(-1)^m} f(y) \right|$$
$$= g_*(x)f(x) + g_*(y)f(y)$$
$$\le g_*(x)f(x)$$

  since we know that $f(x) > 0$ for all $x \in \mathbb{R}$, and $g_*(x) = |H_m(x)| > 0$, which is similar to the g_* before; thus satisfies the conditions for it.

Next, to satisfy Condition **??**, for the density function f of the normal distribution, must fulfill that;

- The α-moment of f must be finite, so $\int_{\mathbb{R}^d} \|x\|^\alpha f(x)dx < \infty$ - this is always true for the normal distribution, all of its moments are finite, since they are defined with respect to σ^n, for some n, and $\sigma < \infty$.

Lastly, to satisfy Condition **??**, we must find the values of k for which the estimator provides a uniform convergence for Theorems **??** and **??**. To do this we must have, for some $\alpha > d = 1$, let k_0^* and k_1^* denote two deterministic sequences of positive integers with $k_0^* \le k_1^*$. Taking $\alpha := 2$, we must have;

- $k_1^* = O(N^\tau)$, where $\tau < \min\left\{\frac{2\alpha}{5\alpha+3d}, \frac{\alpha-d}{2\alpha}, \frac{4}{4+3d}\right\} = \min\left\{\frac{4}{13}, \frac{1}{4}, \frac{4}{7}\right\} = \frac{1}{4}$, so we can choose $\tau := \frac{2}{9} < \frac{1}{4}$ so that we have $k_1^* = O(N^{\frac{2}{9}})$
- $\frac{k_0^*}{\log^5 N} \to \infty$ - for this to be true we need to choose $k_0^* := N^A$ for some $A > 0$. Considering that $k_0^* \leq k_1^*$ and $k_1^* = O(N^{\frac{2}{9}})$, thus $A \in (0, \frac{2}{9})$. So we can choose $A := \frac{1}{\eta}$ for some large η, which gives that $k_0^* = O(N^{\frac{1}{\eta}}) \approx 1$.

Thus, on account of the values of N being considered in the simulations; $N = 100, 200, ..., 50000$, we have that for the smallest $N = 100$, the values of k for which Theorem **??** and **??** both hold, are $k \in \{k_0^*, ..., k_1^*\} = \{1, ..., 100^{\frac{2}{9}}\} = \{1, ..., 2.782\} \approx \{1, 2\}$. Also, for the middle value $N = 25,000$, we have the values of k to be in $\{k_0^*, ..., k_1^*\}$, where $k_1^* \approx 25000^{\frac{2}{9}} = 9.491 \approx 9$, thus $k \in \{1, ..., 9\}$. Moreover, for the largest $N = 50,000$, we must consider $k \in \{1, ..., k_1^*\} = \{1, ..., 50000^{\frac{2}{9}}\} = \{1, ..., 11.072\} \approx \{1, 2, ..., 11\}$.

Overall, due to Conditions **??**, **??** and **??** being met, we can say that for the normal distribution, Theorems **??** and **??** hold; henceforth, we can say that the Kozachenko-Leonenko estimator, of a sample from the 1-dimensional normal distribution is an asymptotically unbiased and consistent estimator for entropy, for some values of $k \in \{1, 2, ..., 11\}$, depending on the sample size N.

1.1.2 Simulation Results

I will now conduct some simulations to consider this for each value of k separately, each time considering 500 samples of size N from this distribution, finding the estimator in each case and take the average of these estimators to find our entropy estimator. I will then consider the relationship show in equation (1.2) for each sample and work out the average for the values of a and c, for each $k \in \{1, 2, ..., 11\}$.

For $N = 100$, $N = 25,000$ and $N = 50,000$, using the results from **??**, we can create a table to compare the mean values of the bias of the estimator for the different values of k considered.

The results shown in table 1.1 show that for a larger N, the modulus of the bias of the estimator is smaller, this is true for all values of k except when $k = 2, 3, 7, 8$, for which the bias is smaller when $N = 25,000$ in comparison to the larger value of N. There are a number of reasons why this could be; however, it is first important to notice that when finding the values of k that satisfy condition**??**, we found that for $N = 100$, we must have $k \in \{1, 2\}$, for $N = 25,000$ we have $k \in \{1, 2, ..., 9\}$ and for $N = 50,000$ we have $k \in \{1, 2, ..., 11\}$.

For the smallest values of $N = 100$, we expect the best value of k to be either 1 or 2; and the table agrees with this showing that the smallest bias occurs at $k = 1$ for a small sample size.

When $N = 25,000$ we have that for $k \in \{2, ..., 8\}$ that the bias is very small, especially for the values of $k = 3, 4, 7, 8$ with the smallest bias appearing when $k = 3$; which fits with the previous analysis that the best value of k will lie within 1 and 9.

Table 1.1: *1-dimensional normal distribution, comparison of k*

k	$\lvert Bias(\hat{H}_{100,k})\rvert$	$\lvert Bias(\hat{H}_{25000,k})\rvert$	$\lvert Bias(\hat{H}_{50000,k})\rvert$
1	0.0031912	0.0006312	0.0004428
2	0.0195347	0.0000092	0.0003632
3	0.0167902	0.0000056	0.0002278
4	0.0264708	0.0001657	0.0001196
5	0.0238265	0.0002138	0.0000003
6	0.0311576	0.0001546	0.0001471
7	0.0356302	0.0000217	0.0003024
8	0.0396299	0.0000984	0.0001021
9	0.0460706	0.0003620	0.0002070
10	0.0458648	0.0002752	0.0002611
11	0.0387339	0.0003332	0.0002458

This table is comparing the values of $\lvert Bias(\hat{H}_{N,k})\rvert$ *for the values of* k *with* $N = 100$, $N = 25,000$ *and* $N = 50,000$*, when the estimator is taken over* 500 *samples*

Now considering the largest sample size $N = 50,000$, the bias when $k = 5$ sticks out since it is $\approx 10^{-3}$ smaller than the other bias values in the table. However, for all other values of k the bias is still extremely small in comparison to the bias for $N = 100$ and even in comparison to $N = 25,000$ in some places. This extreme difference could be an outlier in my data; thus in table 1.2 I have shown the values for the modulus of the bias, when $k = 5$, for different, also large values of N. This table does indeed show that $\lvert Bias(\hat{H}_{50000,5})\rvert \approx 0.0000003$ is an anomaly in the data, and that $k = 5$ is not necessarily the best value of k for $N = 50,000$. Thus, we cannot yet draw any major conclusions about the best value of k for the estimator of a sample this size.

I now wish to consider the equation 1.2 and plot the simulated data, to fit a regression line for each value of k separately, these are shown in Figures 1.1 and 1.2. All of these graphs agree with the relationship previously stated between the sample size and the bias of the estimator; they all show that the logarithm of this equations gives a negative linear relationship - with relatively small error bars.

Moreover, I would like to consider the coefficient of determination (R^2) for each of the above regression lines, this value provides an estimate of the strength of the relationship between the model and the response variable. Also, I would like to consider the standard error/deviation (σ^2), for each of the different graphs, which shows a measure of the predictions' accuracy. These are all depicted for each value of k in table 1.3.

Both columns of this table essentially point to the same conclusion; the

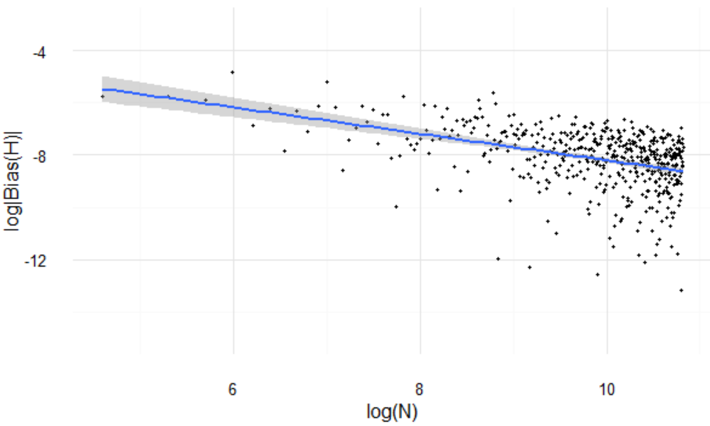
Table 1.2: *1-dimensional normal distribution, $k = 5$ for large N*

N	$\lvert Bias(\hat{H}_{N,5})\rvert$
49100	0.0000639
49200	0.0001463
49300	0.0001700
49400	0.0001037
49500	0.0000711
49600	0.0003221
49700	0.0001047
49800	0.0000644
49900	0.0001240
50000	0.0000003

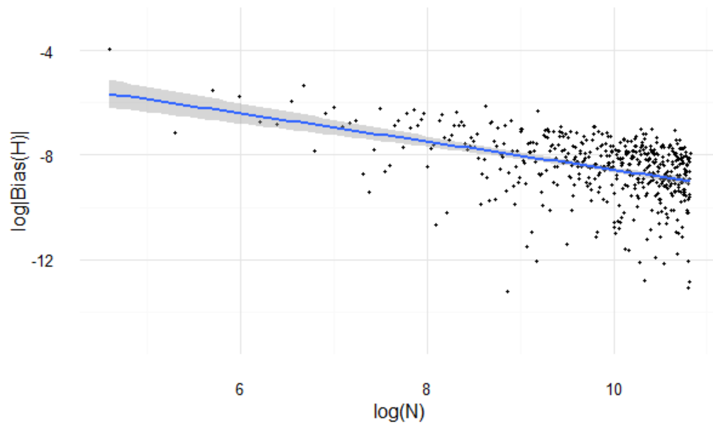
This table is comparing the values of $Var|Bias(\hat{H}_{N,5})|$ for the large values of N.

Table 1.3: *Comparison of the coefficient of determination and the standard deviations of the regression for each value of k for the 1-dimensional normal distribution*

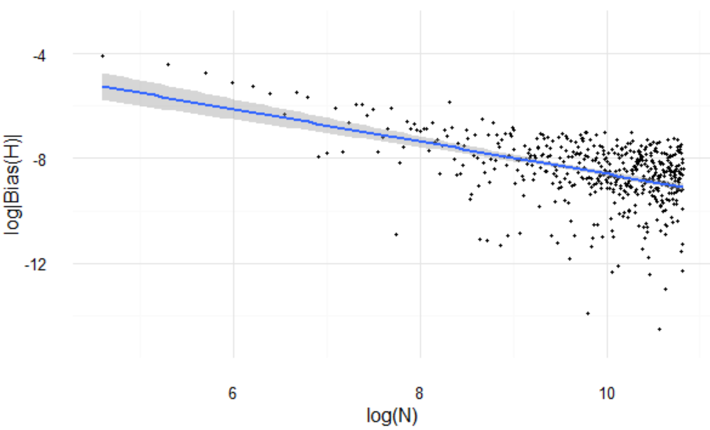
k	R^2	σ^2
1	0.1766	1.0661
2	0.1793	1.1477
3	0.2292	1.1053
4	0.3556	1.0759
5	0.3322	1.1752
6	0.4260	1.0180
7	0.4532	1.0155
8	0.4623	1.0088
9	0.4962	0.9730
10	0.5227	0.9759
11	0.5839	0.8566



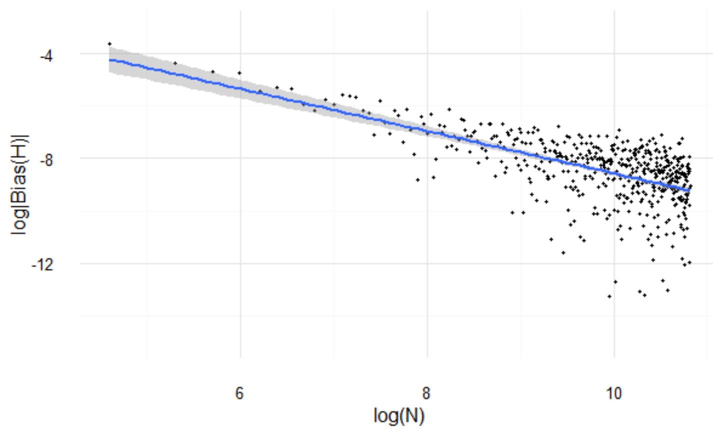
(a) *k=1*



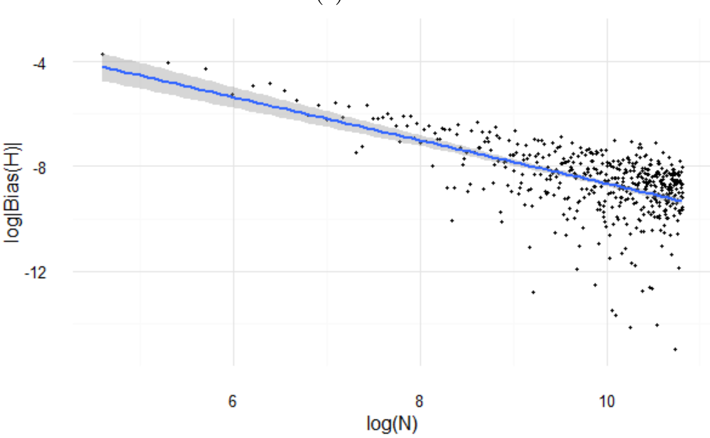
(b) *k=2*



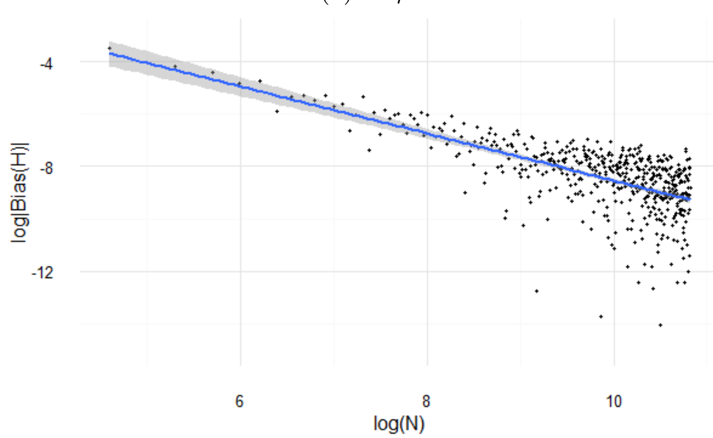
(c) *k=3*



(d) *k=4*

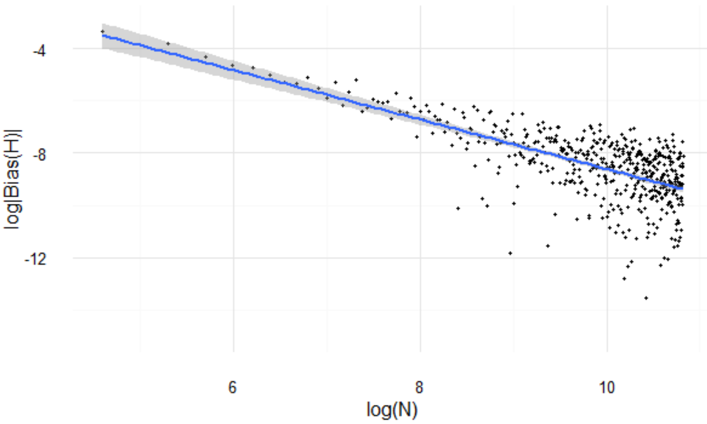


(e) *k=5*

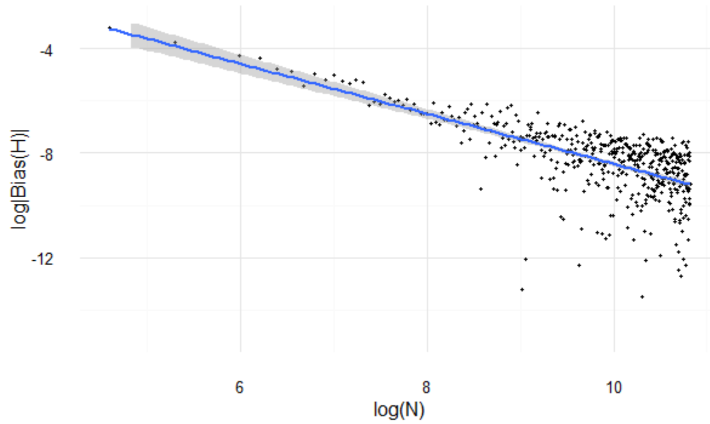


(f) *k=6*

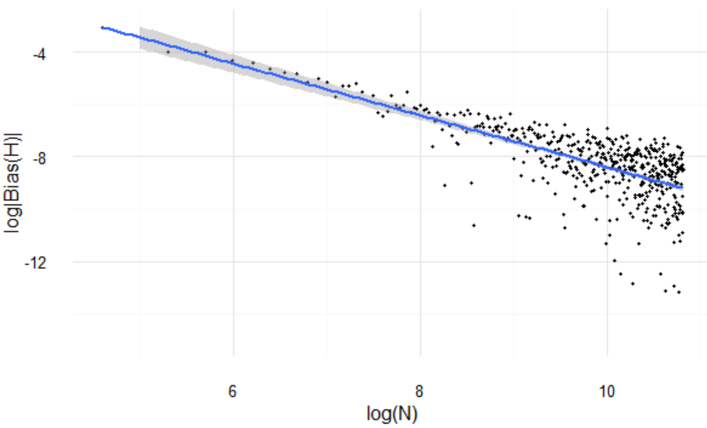
Figure 1.1: *1-dimensional normal distribution with different* $k = 1, ..., 6$



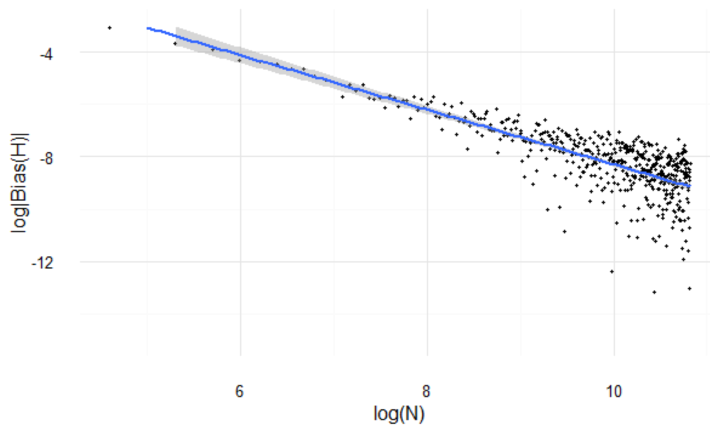
(a) *k=7*



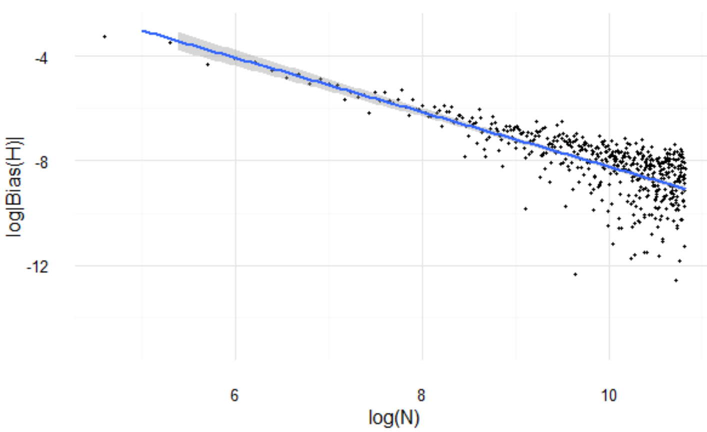
(b) *k=8*



(c) *k=9*



(d) *k=10*



(e) *k=11*

Figure 1.2: *1-dimensional normal distribution with different* $k = 7, ..., 11$

Table 1.4: *Comparison of coefficients of regression a_k and c_k from equation 1.1, for 1-dimensional normal distribution*

k	a_k	c_k
1	0.5054	0.0433
2	0.5490	0.0459
3	0.6169	0.0894
4	0.8181	0.6690
5	0.8486	0.8235
6	0.8976	1.5514
7	0.9464	2.3576
8	0.9574	3.2021
9	0.9883	4.4558
10	1.0454	8.5402
11	1.0386	8.7457

larger the value of k, the more accurate the linear model is to fitting the data. This is shown by the R^2 value increasing towards 1 and the σ^2 values decreasing positively.

The R^2 is very small for $k \leq 3$, which points towards the line being a poor fir to the data; however, due the the standard deviation being $\sigma^2 \approx 1.1$, we cannot say that these lines are poorly fitting; since the majority of the data is within a very small range of the line.

The most important information found from the regression analysis is shown in table 1.4; where the values of a_k and c_k are given for each value of k.

As k runs from $1 \to 11$, we have that a_k and c_k both increase, with smooth values of a_k and a large jump, in the value of c_k, between $k = 3$ and 4, and $k = 9$ and 10. The higher the value of a_k, the stronger the negative relationship is between the two variables in question, so for a larger values of a_k, we have that $|Bias(\hat{H}_{N,k})| \to 0$ for large N faster than smaller values of a_k. This is due to the relationship between $|Bias(\hat{H}_{N,k})|$ and a_k shown in equation (1.1)

Recall, from section **??** we have that the bias acts in one of two ways (equations **??** and **??**); it is either of $O\left(\frac{1}{N^a}\right)$ or $O\left(\left(\frac{k}{N}\right)^a\right)$. Thus we have $|Bias(\hat{H}_{N,k})| \approx \frac{c_k}{N^{a_k}}$ where either c_k is constant or it depends on k and a_k - more specifically is $O(k^{a_k})$. There is evidence here to support the latter claim. If we consider the jump between $k = 3$ and 4 shown in the value of c_k, and consider the results in table 1.5.

This shows that the proportional behaviour between k^{a_k} and c_k also has a large jump when k goes from $3 \to 4$. This agrees with the claim of c_k depending on k in this fashion; however, in table **??** we mentioned another jump between $k = 9$ and $k = 10$, and the evidence here does not show a large jump in the same area. We cannot yet make any conclusions about the dependence of c_k on

Table 1.5: *Considering the dependence of k on c_k*

k	k^{a_k}	c_k	$\frac{k^{a_k}}{c_k}$
1	1	0.0433	23.095
2	1.4631	0.0459	31.875
3	1.9694	0.0894	22.029
4	3.1085	0.6690	4.646
5	3.9187	0.8235	4.759
6	4.9942	1.5514	3.219
7	6.3067	2.3576	2.675
8	7.3218	3.2021	2.287
9	8.7716	4.4558	1.969
10	11.1020	8.5402	1.300
11	12.0668	8.7457	1.380

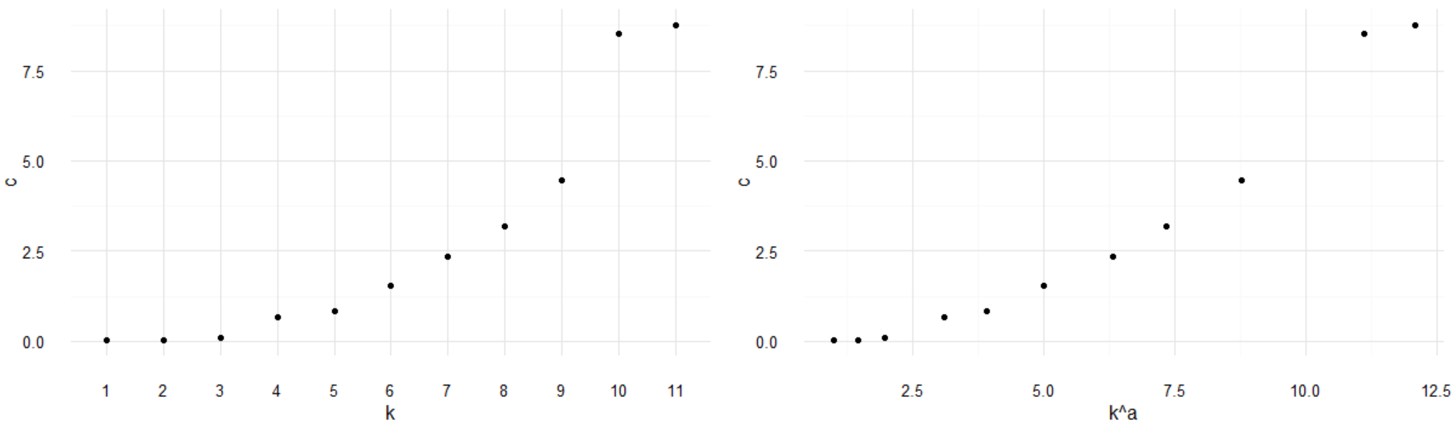
k; this motivates a graphical representation of the value of c_k against k to see if there is any relation, Figure 1.3.

Interestingly, plot 1.3 (a) shows an almost exponential relationship between the values of c_k and the values of k. This leads me to believe that there is some kind of relationship between the two variables, and looking at plot 1.3(b) this shows that there's a strong possibility that the relationship is of the form stated in equation **??**.

To better study the linear relationship between the logarithm of the bias and the logarithm of the sample size, I have generated a comparison plot, shown in Figure 1.4.

From this we can see obviously that for smaller values of N, smaller values of $log(N)$, the smallest bias occurs when $k = 2$, since this line is the lowest for the data up until $log(N) \approx 9$ - i.e. $N \approx 13,000$. For a larger sample size, we cannot accurately see in this graph which line is the best. This motivates us to look at a section of the graph when $9 \leq log(N) \leq 11$ - i.e. $8,000 \leq N \leq 50,000$, which is shown in Figure 1.5.

From this graph we can obviously discount $k = 1$ for large N, since this is the most gradual descent; thus the bias will be largest for this k. Also, both the lines for $k = 2$ and $k = 3$ are more gradual in their descent at larger N, so are probably not the best to choose. Even though, for $k = 9, 10$ and 11, the slope is the steepest - a_k is largest - the intercept is larger so around the biggest sample size considered $N = 50,000$, $log(N) \approx 10.8$, there is not the smallest bias. Actually, for large values of $N \leq 50,000$ we can see from this graph that the best lines appear to be those which are blue/green; $k = 4, 5, 6, 7, 8$. Where the lowest lines around the maximal sample size are those for $k = 5$ and $k = 7$; thus these values of k could possibly be the best nearest neighbour value to choose, when looking at a sample of size $N \approx 50,000$ from the normal distribution.



(a) *The values of k against the values of c_k*

(b) *The values of k^{a_k} against the corresponding values of c_k*

Figure 1.3: *Graphically representing the relationship between c_k and k*

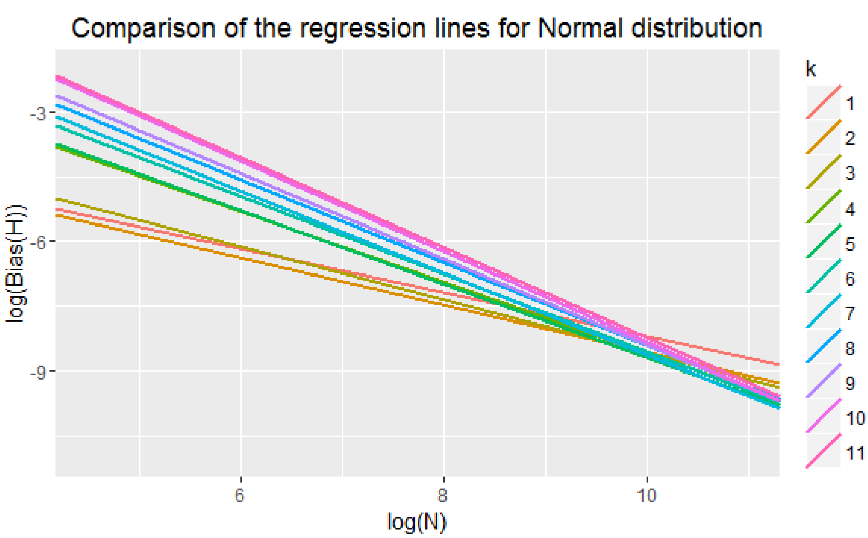


Figure 1.4: *Plot of regression lines for $\log|Bias(\hat{H}_{N,k})|$ against $\log(N)$, for $k = 1, 2, ..., 11$, for samples from the normal distribution*

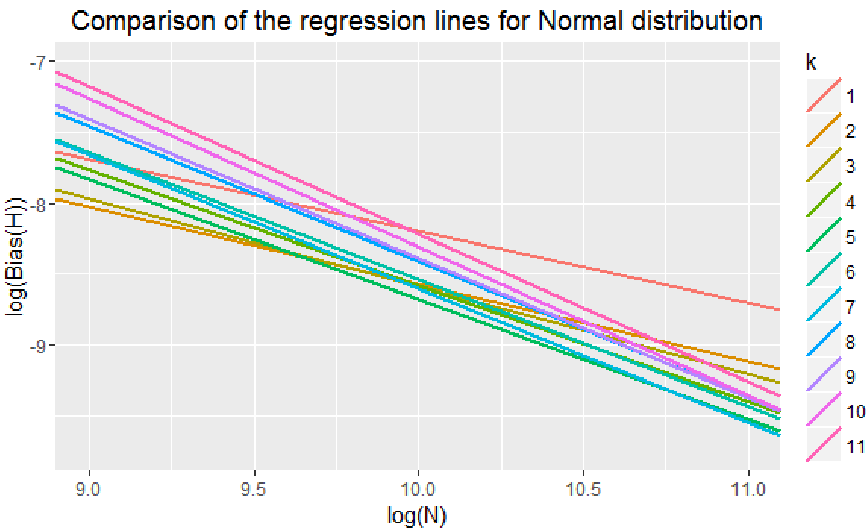


Figure 1.5: *Figure 1.4 zoomed in around large N*

1.2 1-dimensional Uniform Distribution

I will now explore the entropy of samples from the 1-dimensional uniform distribution, $U[a, b]$. This distribution also has an exact formula to work out the entropy for. We can find this formula by considering the density function, f, from the uniform distribution, which is given by;

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & otherwise \end{cases}$$

Using the definition of Shannon entropy given in equation (**??**), we can find the exact entropy for the uniform distribution;

$$\begin{aligned} H &= -\int_{x \in \mathbb{R}^d} f(x) log(f(x)) dx \\ &= -\int_a^b \frac{1}{b-a} log \left[\frac{1}{b-a} \right] dx \\ &= -\frac{1}{b-a} log \left[\frac{1}{b-a} \right] \int_a^b dx \\ &= -log \left[\frac{1}{b-a} \right] \end{aligned}$$

Thus, the actual value of entropy for the uniform distribution is given by;

$$H = log[b-a] \tag{1.5}$$

Similarly to the 1-dimensional normal distribution, we have for the 1-dimensional uniform distribution that $d = 1$ so $V_1 = 2$, thus our estimator takes the form of equation (**??**);

$$\hat{H}_{N,k} = \frac{1}{N}\sum_{i=1}^{N} log\left[\frac{2\rho_{(k),i}(N-1)}{e^{\Psi(k)}}\right]$$

Moreover, the samples considered will not be from the standard uniform, but from the uniform distribution $U[0, 100]$. This is because, using the standard uniform, $U[0, 1]$, would fail since taking $N = 50,000$ samples between 0 and 1 would generate problems as the density function would be $f(x) = 1, \quad 0 \leq x \leq 1$, which would incur working on a very small scale; i.e taking a points with distance between them. Thus, I will be using the density function $f(x) = 0.01, \quad 0 \leq x \leq 100$, which is from the $U[0, 100]$ distribution and gives the exact entropy to be;

$$H = log(100) \approx 4.605170 \tag{1.6}$$

1.2.1 Estimator Conditions

For Theorems **??** and **??** to be satisfied by the estimators generated by samples from the uniform distribution, this density function must meet the Conditions **??**, **??** and **??**. Firstly, to satisfy Condition **??**, for the density function $f(x) = 0.01$ for $0 \leq x \leq 100$, it must be such that;

- f is bounded - obviously, since the density function for the uniform distribution is constant for $x \in [a, b]$ and 0 otherwise; hence is bounded.
- f is m-times differentiable - as f is constant this holds
- $\exists r_* > 0$ and a Borel measurable function g_*, with $\|y - x\| \leq r_*$ so that $\|f^{(t)}(x)\| \leq g_*(x)f(x)$ and $\|f^{(m)}(x) - f^{(m)}(x)\| \leq g_*(x)f(x)\|y - x\|^{\eta}$, for some g_* such that $\sup_{\{x:f(x)<\delta\}} g_*(x) = O(\delta^{-\epsilon})$ as $\delta \searrow 0$ for some $\epsilon > 0$. Since we are considering a 1-dimensional distribution, we can write the norms $\|\cdot\|$ as $|\cdot|$. Moreover, considering that for Theorems **??** and **??**, we have the value of $\beta \geq 2$; thus choosing $\beta = 2$, and since $m = \lfloor \beta \rfloor = \lfloor 2 \rfloor = 2 = \beta$ and $\eta = \beta - m$, we have that $\eta = 0$. Thus we need $|f^{(t)}(x)| \leq g_*(x)f(x)$, which is obvious since f is constant for $f^{(t)}(x) = 0$ for all $t \geq 1$ TODO - sort this out

Next, to satisfy Condition **??**, for the density function f of the uniform distribution, must fulfill that;

- The α-moment of f must be finite, so $\int_{\mathbb{R}^d} \|x\|^{\alpha} f(x)dx < \infty$ - this is true, since for the 1-dimensional uniform distribution, $f(x)$ is constant; thus we would be integrating a polynomial $|x|^{\alpha}$, over a finite interval $a \leq x \leq b$, which is always finite.

Lastly, to satisfy Condition **??**, we must find the values of k for which the estimator provides a uniform convergence for Theorems **??** and **??**. These values

Table 1.6: *1-dimensional uniform distribution, comparison of k*

k	$\lvert Bias(\hat{H}_{100,k})\rvert$	$\lvert Bias(\hat{H}_{25000,k})\rvert$	$\lvert Bias(\hat{H}_{50000,k})\rvert$
1	0.0005189	Inf	Inf
2	0.0047466	0.0001745	0.0001163
3	0.0083912	0.0001776	0.0000899
4	0.0169364	0.0001177	0.0000490
5	0.0152168	0.0000588	0.0000509
6	0.0148205	0.0000817	0.0000538
7	0.0218339	0.0002918	0.0000663
8	0.0250401	0.0001884	0.0000487
9	0.0297655	0.0001406	0.0001184
10	0.0337164	0.0001337	0.0000949
11	0.0381473	0.0001693	0.0000417

This table is comparing the values of $|Bias(\hat{H}_{N,k})|$ for the values of k with $N = 100$, $N = 25,000$ and $N = 50,000$, when the estimator is taken over 500 samples

are independent of the distribution that the sample is from, and only depends on the size of the sample, the dimension of the distribution that sample is taken from and the value chosen for α, where we have chosen $\alpha = 2$. The values of k found in section **??** are $\{1, 2\}$ for $N = 100$, $\{1, 2, ..., 9\}$ for $N = 25,000$ and then $\{1, 2, ..., 11\}$ for $N = 50,000$.

Thus, due to the above conditions for Theorems **??** and **??** being met, we can say that the Kozachenko-Leonenko estimator, of a sample from the uniform distribution is an asymptotically unbiased and consistent estimator for entropy, for specific $k \in \{1, 2, ..., 11\}$, depending on the sample size N.

1.2.2 Simulation Results

I will now conduct simulations, in a similar manner as for the normal distribution, where for each value of k separately, I will consider 500 samples of size N from the uniform distribution, finding the estimator in each case and take the average of these estimators to find our entropy estimator. I will then consider the relationship show in equation (1.2) for each sample and work out the average for the values of a and c, for each $k \in \{1, 2, ..., 11\}$.

For $N = 100$, $N = 25,000$ and $N = 50,000$, using the results from **??**, we can create a table to compare the mean values of the bias of the estimator for the different values of k considered.

Looking at the results shown in table 1.6, for $N = 100$, we can see that the smallest bias, quite obviously, occurs when the estimator is taken with $k = 1$, and that for other values of k there is a significant difference in the size of the bias.

When looking at the larger values of N, it is first important to note that `Inf` $= \infty$ is present in the table as for a large value of N, it is difficult to work out the estimator when $k = 1$. This is because there are only extremely small distances between the closest samples of this distribution, and since any program used to compute these numbers will fail with extremely small numbers and default them to `- Inf`. We only need there to be two samples with an infinitely small distance to make the whole estimator become `- Inf`, which is `Inf` when looking at the modulus. This was mentioned earlier and was the motivation for taking the uniform distribution over $[0, 100]$; however, because of this, from now onwards I will not be considering $k = 1$ in the estimator for the uniform distribution.

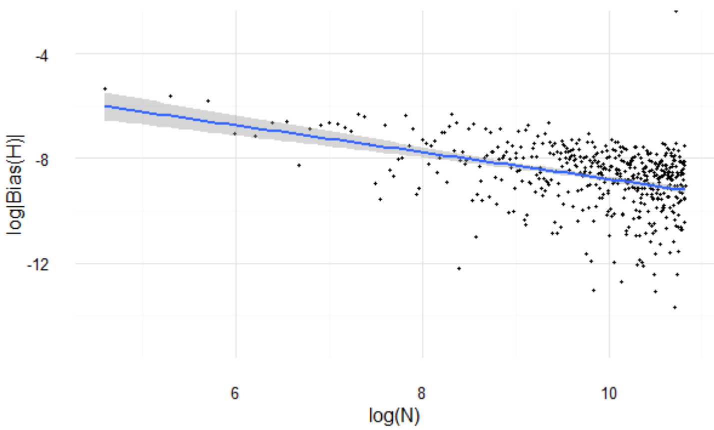
Thus, looking at the results shown in table 1.6, ignoring $k = 1$ and considering $N = 25,000$, it appears to show smallest bias for values of $k \in \{4, 5, 6\}$; more specifically $k = 5$ appears to have the smallest bias. This fits with the previous analysis done on the value of k for different sample sizes, which stated that we must have $k \in \{1, 2, ..., 9\}$. Next, examining the table for $N = 50,000$, we can see that the smallest bias looks to be when $k \in \{4, 5, 6, 7, 8, 11\}$, however, for all other values of k we still have a small bias < 0.000012. Thus, we cannot yet draw any conclusions about the optimal value of k for large N and this distribution.

Next I wish to examine the graphs showing the relationship in equation 1.2, by plotting the simulated data and fitting a regression line for each value of k separately, in figures 1.6 and 1.7. Thus showing a negative linear relationship between the logarithm of the bias of the estimator and the logarithm of the sample size N. Also, looking more closely, the regression lines fitted to the data appear to become more steep for higher values of k, whilst the standard error bars appear to become smaller.

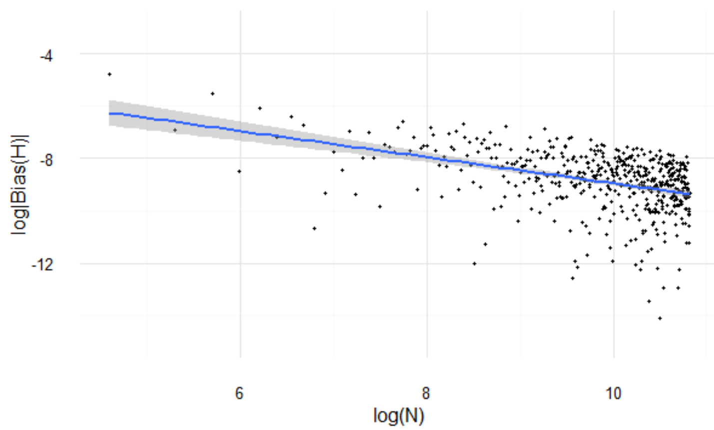
Another important thing to consider, before looking at the equations of the regression lines, is to see how well these lines actually fit the data. To do this I have, as before, examined the coefficient of determination and the standard deviation of the lines, Table 1.7.

This table is very similar to that for the normal distribution in that both columns point towards the same conclusion; the larger that k is, the more accurate the linear model is to fitting the data. This is shown by the R^2 value generally increasing towards 1 and the σ^2 values decreasing positively - the deviation about the line is decreasing for higher k. There is slight fluctuation in the middle values of k, where for $k = \{6, 7, 8\}$ there is not an exact direction that the relationship is going. However, there is nothing to say that this isn't normal behaviour, since we do not yet know the relationship between the bias and k.

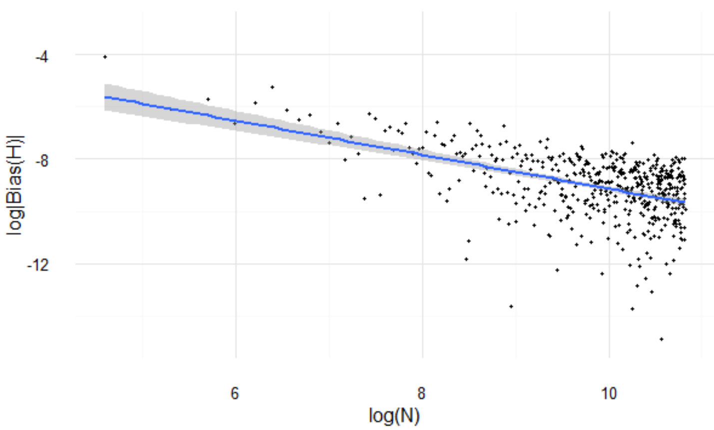
Moreover, R^2 is very small for $k \leq 4$, which points towards the line being a poor fir to the data; however, due the the standard deviation being $\sigma^2 \approx 1.1$, we cannot discard the importance of these lines; since most of the data is in a very small range about the line. Additionally, for $k = 11$ we have the strongest



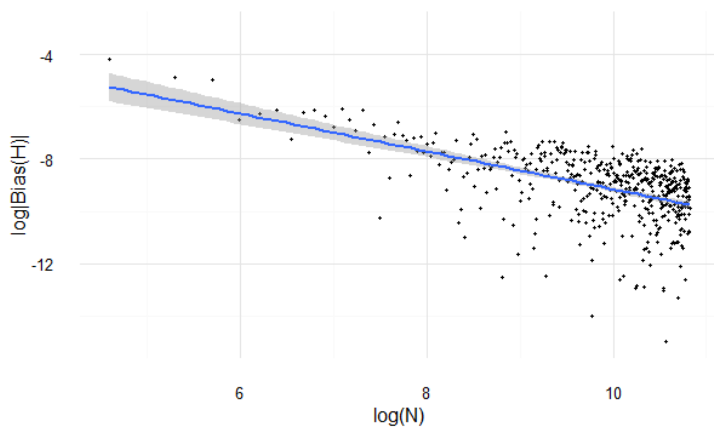
(a) *k=2*



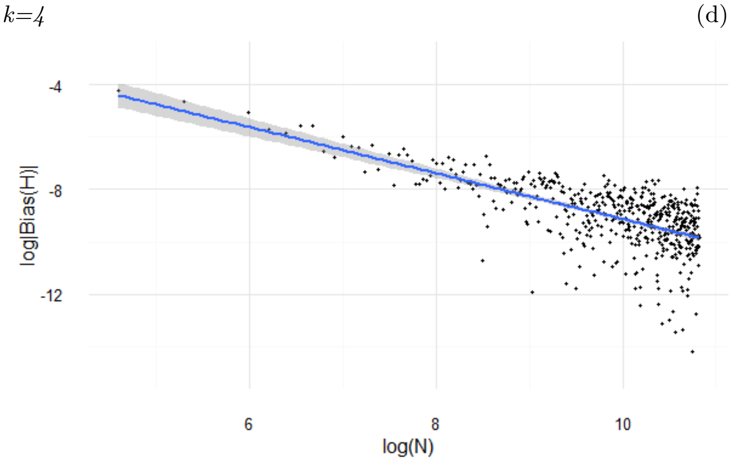
(b) *k=3*



(c) *k=4*

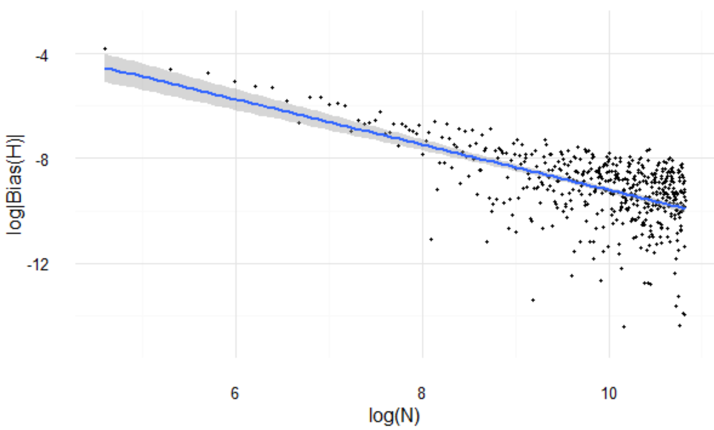


(d) *k=5*

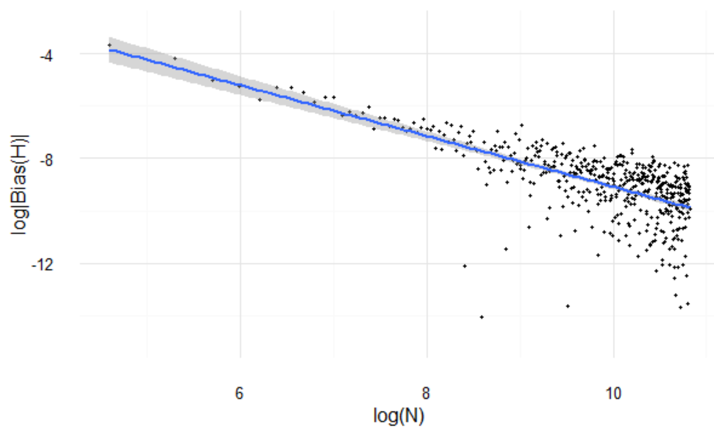


(e) *k=6*

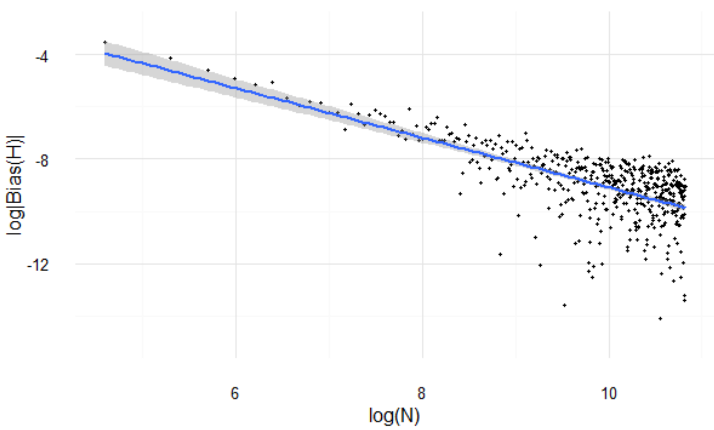
Figure 1.6: *1-dimensional Uniform distribution with different* $k = 2, ..., 6$



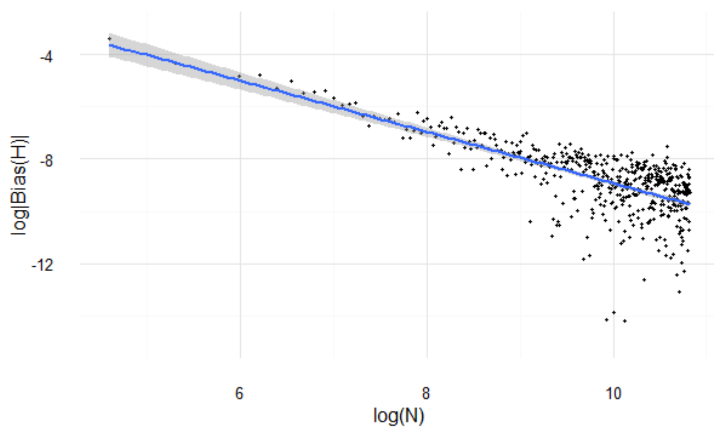
(a) *k=7*



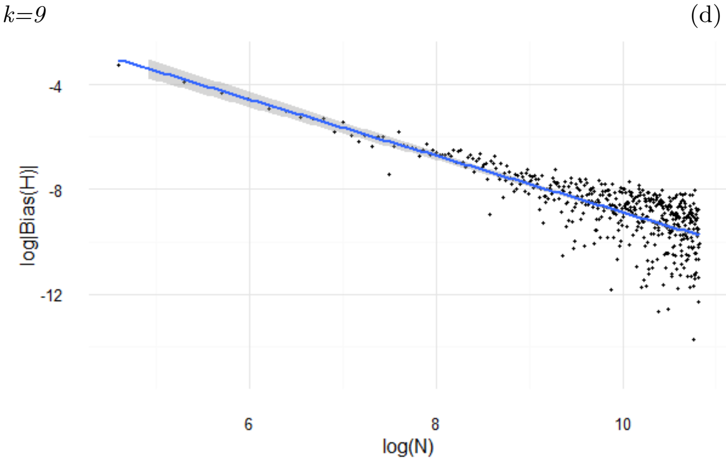
(b) *k=8*



(c) *k=9*



(d) *k=10*



(e) *k=11*

Figure 1.7: *1-dimensional Uniform distribution with different* $k = 7, ..., 11$

Table 1.7: *Comparison of the coefficient of determination and the standard deviations of the regression for each value of k for the 1-dimensional uniform distribution*

k	R^2	σ^2
2	0.1721	1.0982
3	0.1576	1.1402
4	0.2485	1.1200
5	0.2784	1.1459
6	0.4027	1.0286
7	0.3605	1.1252
8	0.4689	1.0073
9	0.4653	0.9940
10	0.5088	0.9408
11	0.6215	0.8208

relationship seen so far - the largest R^2 and smallest σ^2 values appear - this could indicate that for this distribution, when $k = 11$ the linear relationship between the logarithm of the bias and the logarithm of the sample size is most likely to be true.

Possibly the most important information found from the regression analysis, is shown in Table 1.8; where the values of a_k and c_k are given for each value of $k = \{2, 3, ..., 11\}$.

We wish to have a value of $a_k \geq 0.5$ to show the relationship desired, and this is true for all values of k in this table. As k increases, we have that both a_k and c_k increase, expect for $k = 8$, which has larger values that $k = 9$. However, this slight change around $k = 8$ does not necessarily imply anything dramatic since the overall trend seems to fit with that found for the normal distribution.

To better see if the relationship of the bias of the estimator and the sample size is of the form of equations **??** and **??**; either $O\left(\frac{1}{N^a}\right)$ or $O\left(\left(\frac{k}{N}\right)^a\right)$. Thus I will consider the value of c_k for each k, to see if it is dependent on k_k^a or not, shown in table 1.9.

This shows that the proportional behaviour between k^{a_k} and c_k does not imply that $c_k = O(k_k^a)$, since there is no exact trend in the numbers, other than a slight decrease as k increases. However, this increase is not uniform so we can make little assumptions about the behaviour of c_k, perhaps a better way to show this relationship is through a graphical representation of c_k against k to see what form it takes.

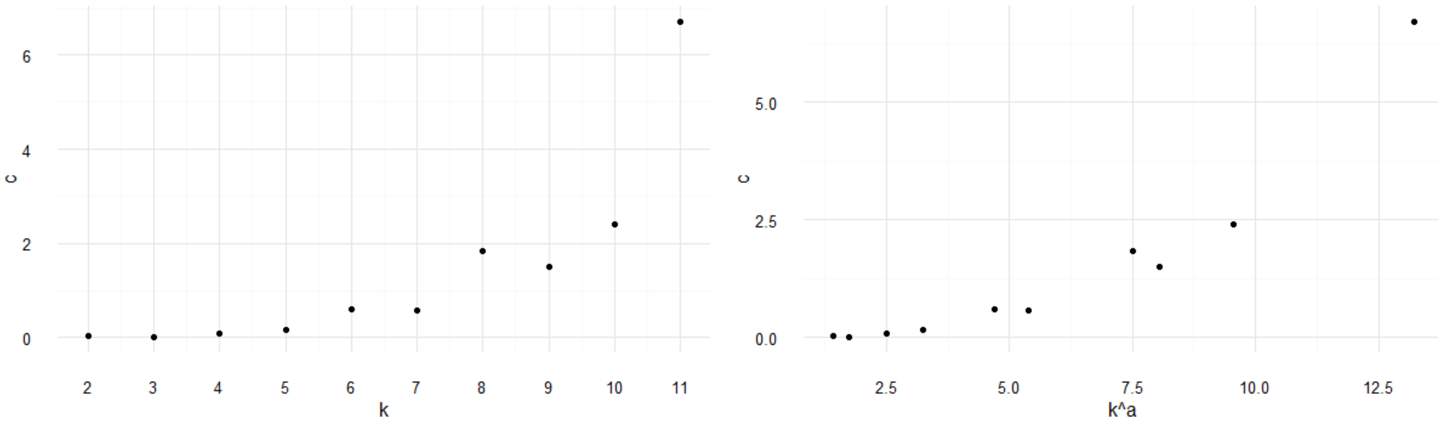
Figure 1.8, for the uniform distribution tells a slightly different story to that of the normal. It does seem to show an almost exponential relationship, as before, however, there are a larger number of outlier values that do not fit within a smooth line in these graphs than in those for the normal distribution,

Table 1.8: *Comparison of coefficients of regression a_k and c_k from equation 1.1, for 1-dimensional uniform distribution*

k	a_k	c_k
2	0.5125	0.0258
3	0.5048	0.0199
4	0.6593	0.0779
5	0.7286	0.1531
6	0.8645	0.6090
7	0.8648	0.5758
8	0.9688	1.8377
9	0.9492	1.5095
10	0.9801	2.4041
11	1.0765	6.6932

Table 1.9: *Considering the dependence of k on c_k*

k	k^{a_k}	c_k	$\frac{k^{a_k}}{c_k}$
2	1.4265	0.0258	55.291
3	1.7412	0.0199	87.498
4	2.4942	0.0779	32.019
5	3.2305	0.1531	21.101
6	4.7066	0.6090	7.729
7	5.3807	0.5758	9.345
8	7.4975	1.8377	4.080
9	8.0495	1.5095	5.332
10	9.5521	2.4041	3.973
11	13.2148	6.6932	1.974



(a) *The values of* k *against the values of* c_k

(b) *The values of* k^{a_k} *against the corresponding values of* c_k

Figure 1.8: *Graphically representing the relationship between* c_k *and* k *for the uniform distribution*

Figure 1.3. However, the relationship that is shown leads me to believe that there is something important between the two variables, which could possibly be of the form stated in equation **??**.

To better study the linear relationship between the logarithm of the bias and the logarithm of the sample size, I have generated a comparison plot, shown in Figure 1.4.

For the smaller values of N, when $N \approx 100$, we can see obviously from this plot that the lowest line occurs from $k = 3$, and this value of k stays the best for the estimator until $log(N) \approx 8.5$, so the sample size $N \approx 5,000$. Above this value, it appears to be that the smallest bias occurs from when $k \in \{7, 8, 9, 11\}$, to see this more accurately consider an enlarged version of this graph for $5,000 < N < 50,000$, Figure 1.10.

This graph shows that for large $N \leq 50,000$, the lowest line on the graph is when the estimator is found using $k = 7$; thus there is a possibility that this value of k could be the best nearest neighbour value to choose, when considering sample size $N \approx 50,000$ from the uniform distribution.

1.3 1-dimensional Exponential Distribution

I will now be looking at the entropy of samples from the exponential distribution $exp(\lambda)$, where $\lambda > 0$ is the rate or inverse scale parameter. In a similar fashion to the previous distributions, the exponential also has an exact formula for the entropy, given the rate parameter λ. Using equation (**??**) and the density function for the exponential distribution $f(x) = \lambda e^{-\lambda x}$ for $x \in [0, \infty)$ and for

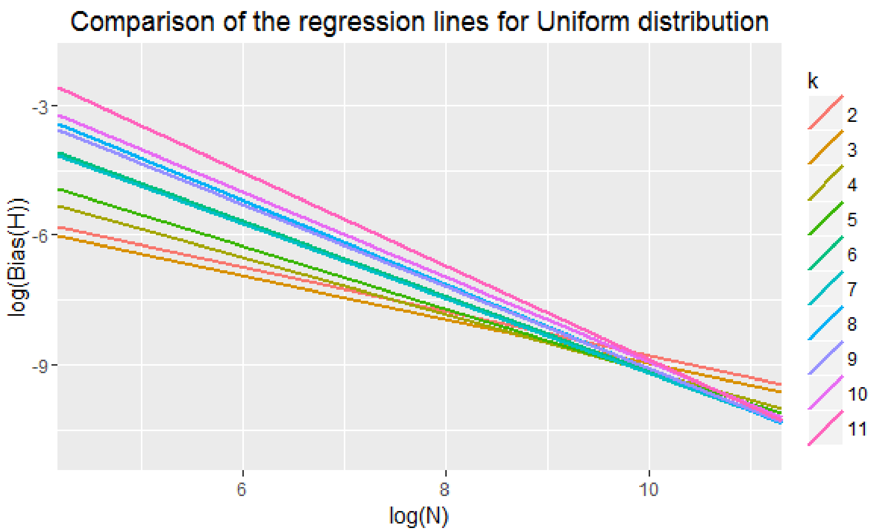


Figure 1.9: *Plot of regression lines for* $\log|Bias(\hat{H}_{N,k})|$ *against* $\log(N)$*, for* $k = 2, 3, ..., 11$*, for samples from the uniform distribution*

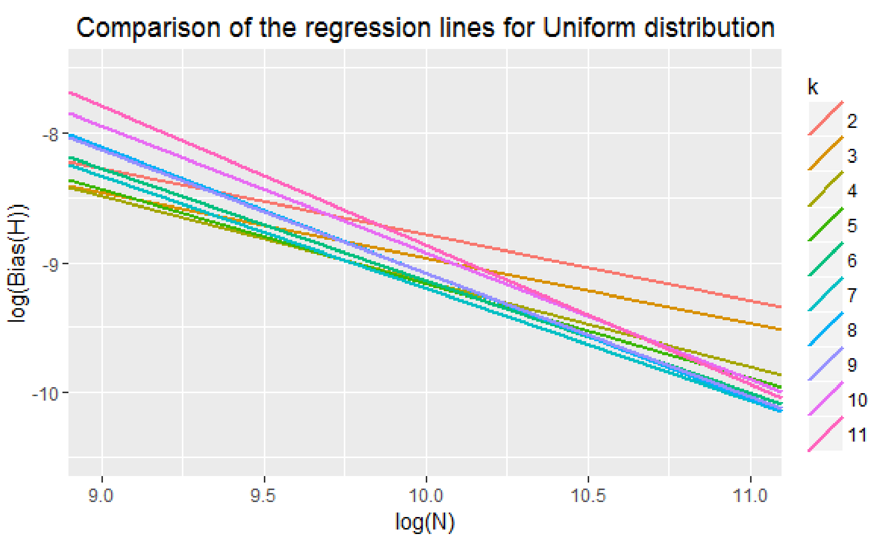


Figure 1.10: *Figure 1.9 zoomed in around large* N*.*

$\lambda > 0$, we can write the exact entropy;

$$\begin{aligned}
H &= -\int_{x\in\mathbb{R}^d} f(x)log(f(x))dx \\
&= -\int_0^{\infty} \lambda e^{-\lambda x} log[\lambda e^{-\lambda x}]dx \\
&= -\lambda \int_0^{\infty} \lambda e^{-\lambda x}[log(\lambda) - \lambda x]dx \\
&= \lambda \int_0^{\infty} \lambda e^{-\lambda x} - log(\lambda)e^{-\lambda x}dx \\
&= -\lambda \left[xe^{-\lambda x}\right]_0^{\infty} + \int_0^{\infty} \lambda e^{-\lambda x}dx + log(\lambda)\left[e^{-\lambda x}\right]_0^{\infty} \\
&= 0 + (log(\lambda) - 1)\left[e^{-\lambda x}\right]_0^{\infty} \\
&= -(log(\lambda) - 1)
\end{aligned}$$

Thus we have the the exact value of entropy for the exponential distribution, given the rate parameter $\lambda > 0$, is;

$$H = 1 - log(\lambda) \tag{1.7}$$

Moreover, I am again considering a 1-dimensional distribution; thus $V_d = V_1 = 2$, and;

$$\hat{H}_{N,k} = \frac{1}{N}\sum_{i=1}^{N} log\left[\frac{2\rho_{(k),i}(N-1)}{e^{\Psi(k)}}\right]$$

is the form of the Kozachenko-Leonenko estimator that I will be considering here, equation (**??**).

I have decided to choose to explore the exponential distribution with rate parameter $\lambda = 0.5$, this is because, we know that for the exponential distribution we must have the rate parameter $\lambda > 0$ and if I choose $\lambda > e \approx 2.7183$ we get a negative values of entropy, $H < 0$. This will introduce problems when considering the modulus of the bias; hence, for this analysis it will be more beneficial to consider a positive value of entropy. Also, for $\lambda \geq 1$, we have a very small value of entropy, $0 \leq H \leq 1$, which would cause problems when calculating large samples and their entropy. Therefore, I have chosen the rate parameter such that $\lambda \in (0, 1)$, so $\lambda = 0.5$, the exact entropy is given by;

$$H = 1 - log(0.5) \approx 1.693147 \tag{1.8}$$

1.3.1 Estimator Conditions

Samples from the exponential distribution must satisfy the conditions of Theorem **??** and **??**, to be an asymptotically unbiased and consistent estimator. For these theorems to hold, this distribution must satisfy the Conditions **??**, **??** and **??**.

Firstly, to satisfy Condition **??**, the density function $f(x) = \frac{e^{\frac{-x}{2}}}{2}$ for $x \in [0, \infty)$, where we have chosen $\lambda = 0.5$, must be such that;

- f is bounded - this is true, since for any probability distribution we have $f(x) \geq 0$, also for the exponential distribution we always have for $x \in [0, \infty)$ that $f(x) \leq 1$, so f is a bounded function.

- f is m-times differentiable - this is obvious, since if we consider the mth derivative of the density function f we get;

$$\begin{aligned}\frac{d^m}{dx^m}(f(x)) &= (-1)^m \lambda^{m+1} e^{\frac{-x}{\lambda}} \\ &= (-1)^m \lambda^m f(x)\end{aligned}$$

  where $(-1)^m \lambda^m$ is a polynomial, which exists for all $x \in [0, \infty)$, thus f is m-times differentiable.

- $\exists r_* > 0$ and a Borel measurable function g_*, with $\|y - x\| \leq r_*$ so that $\|f^{(t)}(x)\| \leq g_*(x)f(x)$ and $\|f^{(m)}(x) - f^{(m)}(x)\| \leq g_*(x)f(x)\|y - x\|^\eta$, for some g_* such that $\sup_{\{x:f(x)<\delta\}} g_*(x) = O(\delta^{-\epsilon})$ as $\delta \searrow 0$ for some $\epsilon > 0$

  Since we are considering a 1-dimensional distribution, the norm $\| \cdot \|$ can be written as $|.|$. Moreover, considering that for Theorems **??** and **??**, we have the value of $\beta \geq 2$; thus choosing $\beta = 2$, and since $m = \lfloor \beta \rfloor = \lfloor 2 \rfloor = 2 = \beta$ and $\eta = \beta - m$, we have that $\eta = 0$, just as previously. Thus we need $|f^{(t)}(x)| \leq g_*(x)f(x)$, which is obvious by above, in view of writing $|\frac{d^t}{dx^t} f(x)| = g_*(x)f(x)$, where we choose $g_*(x) = |(-1)^m \lambda^m| = |\lambda^m| = \lambda^m$, for $t = 1, 2, ..., m$, and $|f(x)| = f(x)$, since $f(x) > 0$. Also, g_* is a polynomial and is hence Borel measurable over $\mathbb{R}$, and for any polynomial we obviously have $\sup_{\{x:f(x)<\delta\}} g_*(x) = O(\delta^{-\epsilon})$ as $\delta \searrow 0$ for some $\epsilon > 0$. Additionally, we need $|f^{(m)}(x) - f^{(m)}(y)| \leq g_*(x)f(x)|y - x|^0 = g_*(x)f(x)$. We currently have;

$$\begin{aligned}|f^{(m)}(x) - f^{(m)}(y)| &= |(-1)^m \lambda^m f(x) - (-1)^m \lambda^m f(y)| \\ &\leq \lambda^m |f(x) + f(y)| \\ &\leq g_*(x)(|f(x)| + |f(y)|) \\ &\leq g_*(x)f(x)\end{aligned}$$

  since we know that $f(x) > 0$ for all $x \in \mathbb{R}$, and $g_*(x) = \lambda^m > 0$, which is the g_* before; thus satisfies the conditions for it.

Next, to satisfy Condition **??**, for the density function f of the exponential distribution, must fulfill that;

- The α-moment of f must be finite, so $\int_{\mathbb{R}^d} \|x\|^\alpha f(x)dx < \infty$ - this is true since the moments of the exponential distribution are given by;

$$\begin{aligned}\int_{\mathbb{R}^d} \|x\|^\alpha f(x)dx &= \int_0^\infty |x|^\alpha \lambda e^{-\lambda x} dx \\ &= \frac{\alpha!}{\lambda^\alpha} < \infty\end{aligned}$$

Table 1.10: *1-dimensional exponential distribution, comparison of k*

k	$\lvert Bias(\hat{H}_{100,k})\rvert$	$\lvert Bias(\hat{H}_{25000,k})\rvert$	$\lvert Bias(\hat{H}_{50000,k})\rvert$
1	0.0008253	Inf	Inf
2	0.0210589	0.0008092	0.0000643
3	0.0173133	0.0001295	0.0004322
4	0.0146824	0.0000769	0.0000023
5	0.0155819	0.0000654	0.0000749
6	0.0194462	0.0000117	0.0001435
7	0.0127644	0.0005499	0.0000490
8	0.0174169	0.0000440	0.0004291
9	0.0216625	0.0003314	0.0000560
10	0.0177220	0.0004974	0.0000544
11	0.0162163	0.0000472	0.0006244

This table is comparing the values of $|Bias(\hat{H}_{N,k})|$ *for the values of k with* $N = 100$*,* $N = 25,000$ *and* $N = 50,000$*, when the estimator is taken over* 500 *samples*

for all $\alpha \in \mathbb{N}$, which is obviously finite.

Lastly, to satisfy Condition **??**, we must find the values of k for which the estimator provides a uniform convergence for Theorems **??** and **??**. As previously, these values are independent of the distribution that the sample is from, and only depends on the size of the sample, the dimension of the distribution that sample is taken from and the value chosen for α, where we have chosen $\alpha = 2$. Thus, the values of k found in section 1.1.1 are $\{1, 2, ..., 11\}$.

Due to the above conditions for Theorems **??** and **??** being met, we can say that the Kozachenko-Leonenko estimator, of a sample from the exponential distribution is an asymptotically unbiased and consistent estimator for entropy.

1.3.2 Simulation Results

I will be exploring the simulation results using the same process as for the previous two distributions, which begins with examining the values of the bias of the estimator at certain values of N for all different $k \in \{1, 2, ..., 11\}$.

In a similar manner to before, using $N = 100$, $N = 25,000$ and $N = 50,000$, from **??**, we can create a table to compare the mean values of the bias of the estimator for the different values of k considered. Recall that the information is taken over 500 sample of size N and the mean of these estimators is considered, then the bias is found by taking the modulus of this value minus the exact entropy, equation 1.8.

For the sample size $N = 100$, Table 1.10 shows that the optimal value of k to choose for the estimator is obviously given by $k = 1$, since the value of the bias is $\approx 10^{-3}$ smaller than that for all the other values of k. Considering $k = 1$ for larger sample sizes, we have ∞ - an error, similar to that for uniform distribution. This happens due to limitations in computer programs; since, we cannot have continuous data, it must be discretised, so for infinitely small distances between two points, we get a $-\infty$ value for the distance to the values closest neighbour, which only has to happen between two points for the whole estimator to become infinite. Because of this, I will not be examining the value of the estimator for $k = 1$, when considering large sample sizes.

Looking at the magnitude of the bias for sample sizes $N = 25,000$, when $k \in \{2, 3, ..., 11\}$, we can see that the smallest bias occurs at $k = 6$; but it is also quite small for the estimator found with $k = 4, 5$ and 11. Because of the closeness of the size of the bias for all the values of k just mentioned, we cannot currently draw up any conclusions about which value of k is definitly better to use - in terms of reducing the bias.

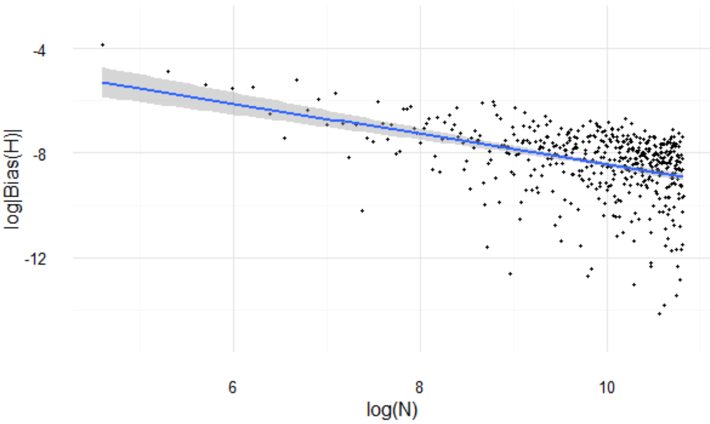
Considering the information found from samples of size $N = 50,000$, it appears to show that the estimator for $k = 4$ has a significantly smaller bias than that for the other values of k. The next smallest bias is $\approx 10^1$ larger than that for $k = 4$, and occurs when $k = 2, 5, 7, 9$ and 10. Since we are only considering the bias at this specific sample size, and not looking at the information about the value of the bias found for samples of a similar size, we cannot yet draw any conclusive decisions from this table.

Considering the relationship shown in equation 1.2, we can plot the simulated data to see if the data does infact show this linear relationship proposed earlier. I have plotted the data seperately for each value of k, and have fitted a linear regression line to each plot to indicate the general trend of the plot. I have not considered $k = 1$ since a large proportion of the simulated values are infinite; thus, the graph does not make much sense. These plots are shown in Figures 1.11 and 1.12.

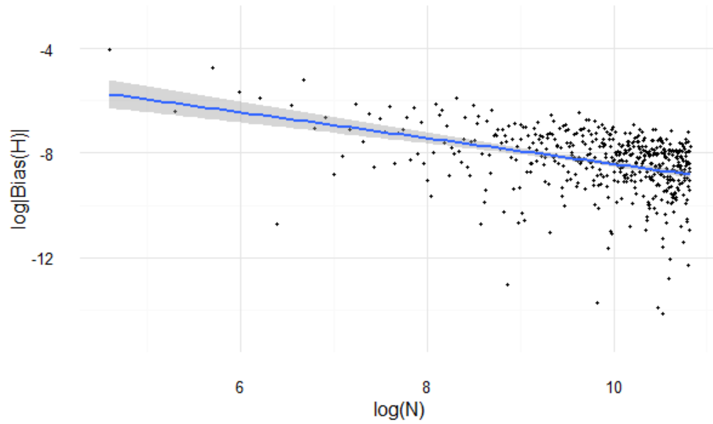
These graphs all show what we expected, and what has been confirmed with the previous two distributions, that the logarithm of the sample size against the logarithm of the bias of the estimator does indeed show a linear relationship. In these graphs there seems to be little difference between the values of k, in both the gradient of the line and in how close the data actually fits the line. This motivates us to consider both the coefficient of determination and the standard deviations of the regression for each value of k, which is given in Table 1.11.

The R^2 value (coefficient of determination) is ≈ 0.207, with very little variation, for all values of k. This is an indication that the regression line could be a poor fit to the data, since a value of 1 means a perfect fit. However, the standard error of the regression line is small, ≈ 1.15 for all k, again with little variation. The standard error for the lines here is similar to that for both the uniform and normal distributions. Thus, this doesn't indicate a poor fit, just that .. TODO wtf does this show???

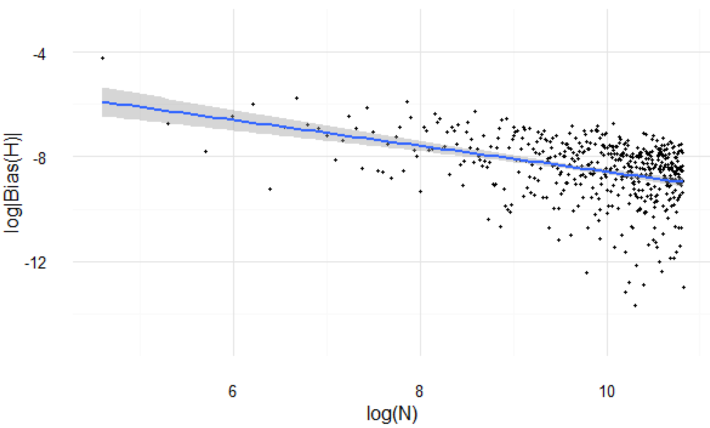
The most important information found from these graphs is the equation of the regression line. I have collated all the data about the coefficients a_k and c_k



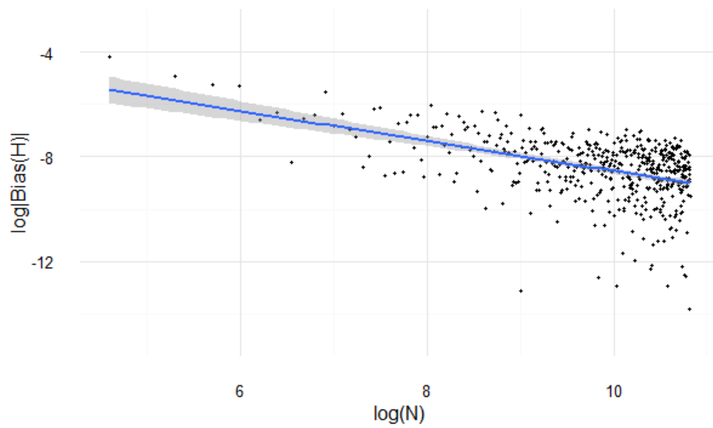
(a) *k=2*



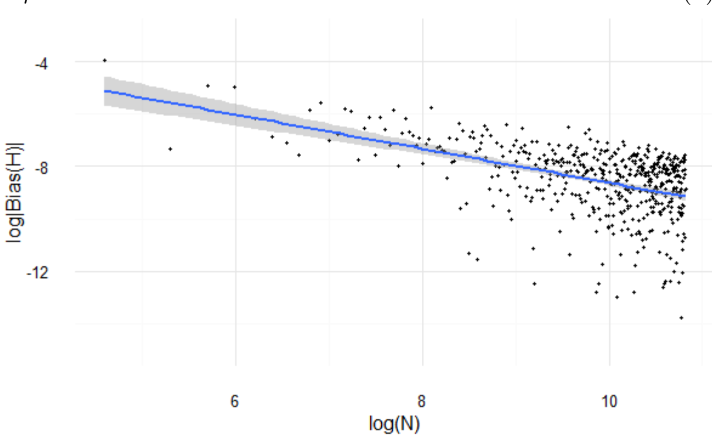
(b) *k=3*



(c) *k=4*

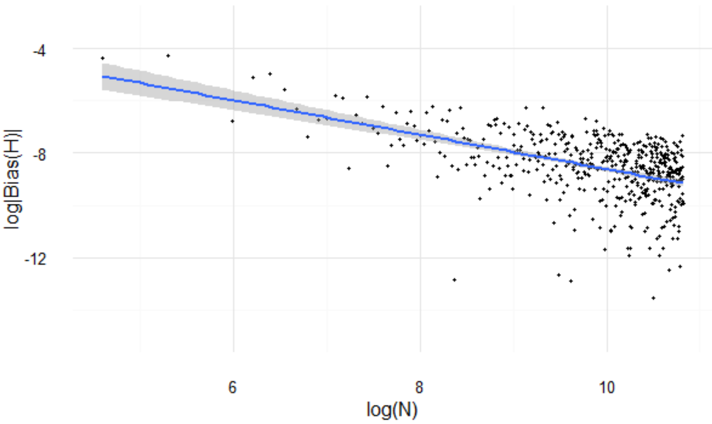


(d) *k=5*

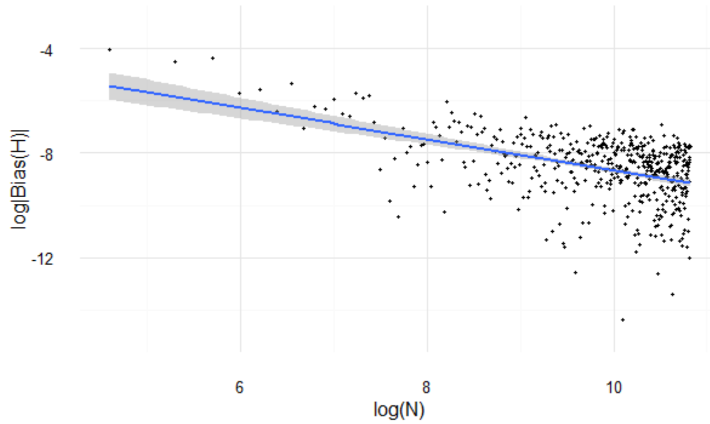


(e) *k=6*

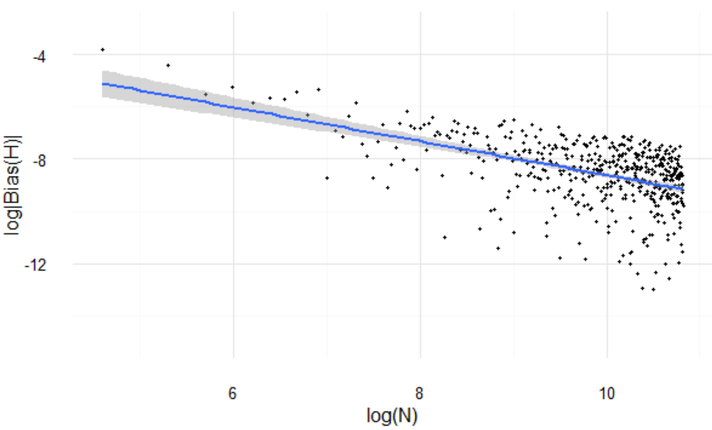
Figure 1.11: *1-dimensional Exponential distribution with different* $k = 2, ..., 6$



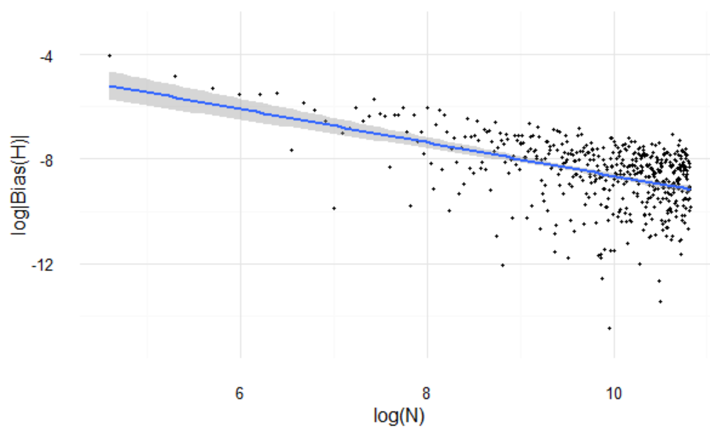
(a) *k=7*



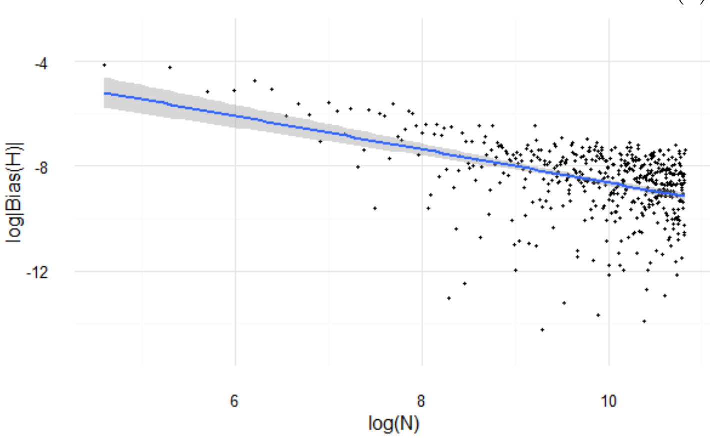
(b) *k=8*



(c) *k=9*



(d) *k=10*



(e) *k=11*

Figure 1.12: *1-dimensional Exponential distribution with different* $k = 7, ..., 11$

Table 1.11: *Comparison of the coefficient of determination and the standard deviations of the regression for each value of k for the 1-dimensional exponential distribution*

k	R^2	σ^2
2	0.1843	1.1972
3	0.1596	1.1073
4	0.1538	1.1321
5	0.2151	1.0687
6	0.2277	1.1694
7	0.2474	1.1253
8	0.2107	1.1473
9	0.2509	1.0939
10	0.2200	1.2150
11	0.2017	1.2369

and displayed them in Table 1.12.

This distribution shows interesting results, that are somewhat different to those previously shown in the other two distributions. In a similar fashion to previously there is a general increase in a_k with k; however, in comparison to before where c_k acted similarly, here c_k does not increase uniformly with k. Both the values of a_k and c_k are unvaried between each k, with a_k ranging from 0.4940 to 0.6606 and c_k ranging from 0.0266 to 0.1324. Interestingly, the smallest values of a_k and c_k, occur at $k = 4$ then at $k = 3$. Additionally, the largest two values occur at $k = 7$ and $k = 10$. The optimal value of k will hopefully become more apparent when plotting all the regression lines against one and other in Figure **??**.

Firstly, I wish to look at the relationship between k and c_k in the exponential distribution. To do this, I will look at the values of c_k for each k and see if it depends on k^{a_k} or not. This should help us to decipher if the bias is of $O\left(\frac{1}{N^a}\right)$ or $O\left(\left(\frac{k}{N}\right)^a\right)$. This information is shown in table 1.13.

For the previous distributions there has been an obvious increase when looking at the values of $\frac{k^{a_k}}{c_k}$ for increasing k. However, for the exponential distribution this is not the case, the values seem to be all over the place, and plotting these in Figure 1.13 confirms this.

These graphs do not show the almost exponential relationship seen before, they don't appear to show any relationship between k and c_k. This could imply that for the exponential distribution we have $Bias|\hat{H}_{N,k}| = O\left(\frac{1}{N^a}\right)$, in comparison the normal and uniform which showed results in favour of the opposing idea; $Bias|\hat{H}_{N,k}| = O\left(\left(\frac{k}{N}\right)^a\right)$.

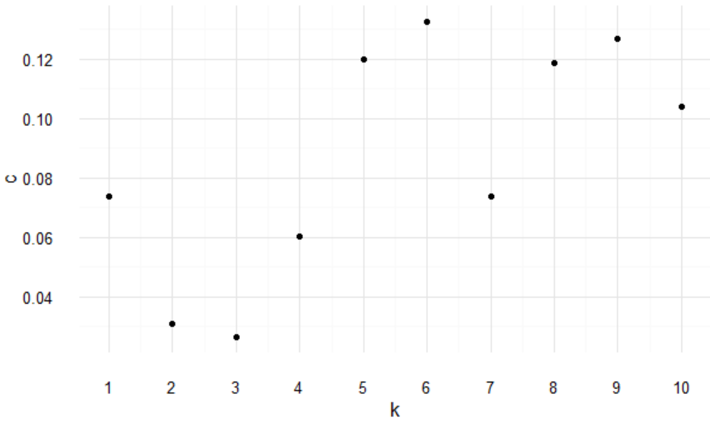
To find the optimal value of k for this distribution, I have generated a comparison plot of all of the regression lines, shown in Figure 1.14. Hopefully this

Table 1.12: *Comparison of coefficients of regression* a_k *and* c_k *from equation 1.1, for 1-dimensional exponential distribution*

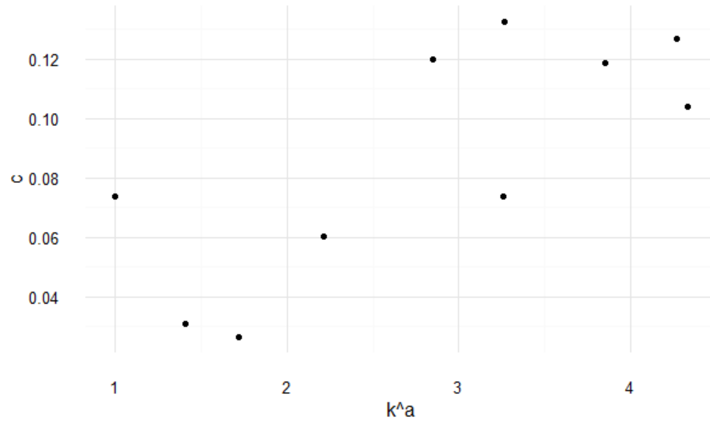
k	a_k	c_k
2	0.5824	0.0739
3	0.4941	0.0310
4	0.4940	0.0266
5	0.5727	0.0602
6	0.6500	0.1199
7	0.6605	0.1324
8	0.6067	0.0739
9	0.6480	0.1189
10	0.6606	0.1267
11	0.6365	0.1042

Table 1.13: *Considering the dependence of* k *on* c_k

k	k^{a_k}	c_k	$\frac{k^{a_k}}{c_k}$
2	1.0000	0.0739	13.532
3	1.4084	0.0310	45.434
4	1.7207	0.0266	64.687
5	2.2121	0.0602	36.745
6	2.8466	0.1199	23.742
7	3.2656	0.1324	24.665
8	3.2563	0.0739	44.063
9	3.8477	0.1189	32.361
10	4.2695	0.1267	33.697
11	4.3301	0.1042	41.556



(a) *The values of k against the values of c_k*



(b) *The values of k^{a_k} against the corresponding values of c_k*

Figure 1.13: *Graphically representing the relationship between c_k and k for the exponential distribution*

can shed some light on which value of k appears to be the best for use for the estimator, depending on the sample size N.

At smaller sample size $N \leq 5,000$, the graph appears to show that the estimator with $k = 4$ has the lowest line; hence, the smallest bias and the best one to use for the estimation. However, for larger N this is not true, to better visualise what is happening for large N, I have created an enlarged version in Figure 1.15.

This graph shows that for large sample sizes, $5,000 \leq N \leq 50,000$ that the estimator with $k = 10$ has the smallest bias; thus appears to be the best one to use for estimating entropy from this distribution.

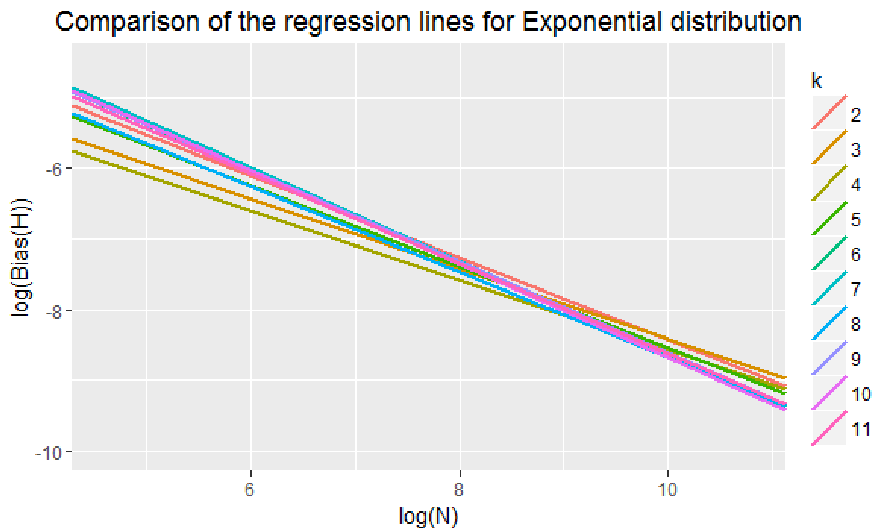


Figure 1.14: *Plot of regression lines for* $\log|Bias(\hat{H}_{N,k})|$ *against* $\log(N)$*, for* $k = 2, 3, ..., 11$*, for samples from the exponential distribution*

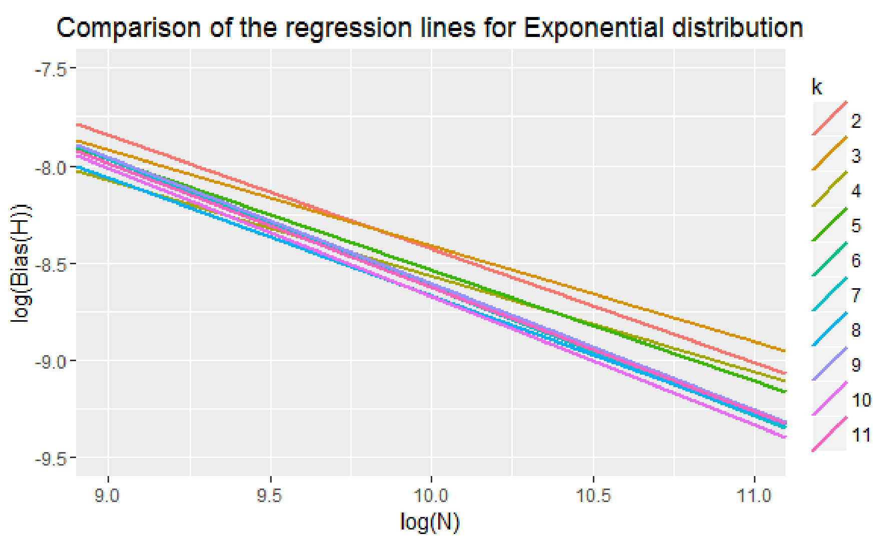


Figure 1.15: *Figure 1.14 zoomed in around large* N*.*