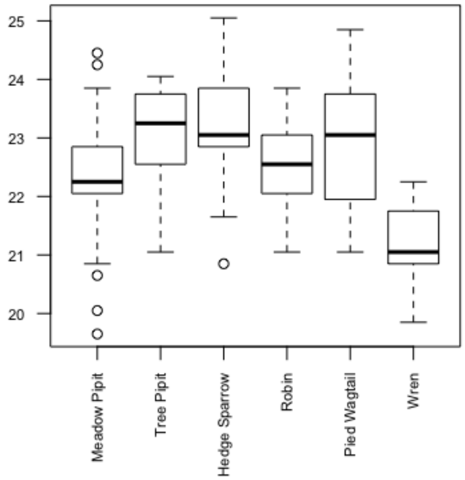
Shane McIntyre
Final Report Stats 351

1. Cuckoos are known to lay their eggs in the nests of other (host) birds. The eggs are then adopted and hatched by the host birds. Is there any relationship between the length of eggs and the host birds? What do we conclude? All data are lengths in millimeters. The data file "Cuckoo_Egg_Lengths_data.pdf" is uploaded at Canvas.

For number 1 my first idea was to create a boxplot to look at the median and spread for each bird.



From this boxplot, it appears the means are distributed randomly. However, this is just a hypothesis from looking at the graph, so I ran a runs up and downs test to verify my hypothesis that the bird and egg size were random. My reasoning for using a runs up and downs test over a K-sample test is the raw data was very large, and every group had a different number of observations. A runs up and downs test allows me to look at each bird as a group. My first step was to calculate the mean for each group

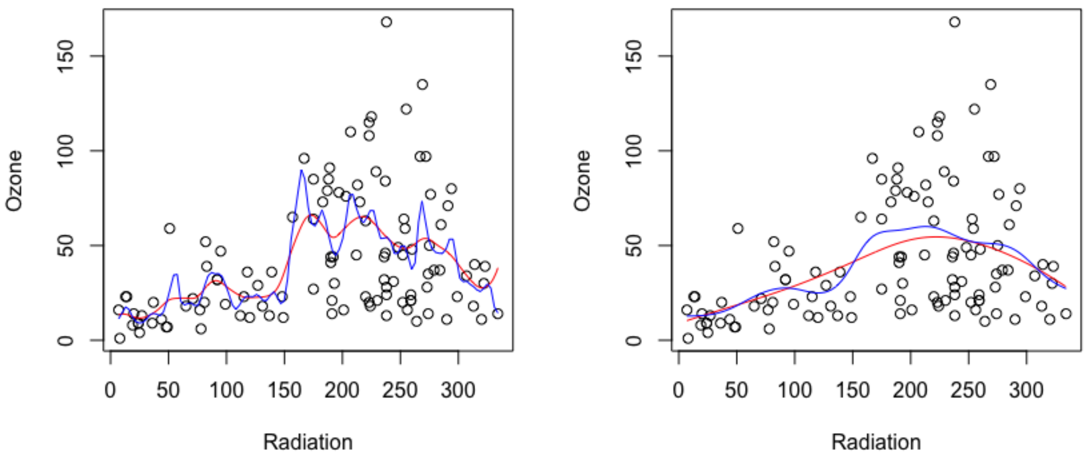
Bird	Mean	+,-
Meadowed Pipit	22.30	+
Tree Pipit	23.09	+
Hedge Sparrow	23.12	+
Robin	22.58	-
Pied Wagtail	22.90	+
Wren	21.13	-

This makes N = 6, and V = 4, from Table E for our class this gives p = .5861. The original hypothesis were: H0: The means are sequenced randomly vs HA: The means are not sequenced randomly. From the p value, H0 is not rejected, so the means are sequenced randomly. This supports the idea that the relationship between the bird and the egg size are random.

2. A data set on ozone can be downloaded from the web sitehttp://web.stanford.edu/~hastie/ElemStatLearn/. Let Y denote the ozone concentration level (given in the 1st column),X1denote “radiation” (given in the 2ndcolumn),X2denote “temperature” (given in the 3rd column), andX3denote “‘windspeed” (given in the 4th column).
    a. (a) Obtain the kernel regression fit to measurements of (X1,Y)
    b. .(b) Obtain the kernel regression fit to measurements of (X2,Y).
    c. (c) Obtain the kernel regression fit to measurements of (X3,Y).
    d. (d) Interpret the results obtained in parts (a), (b), and (c).

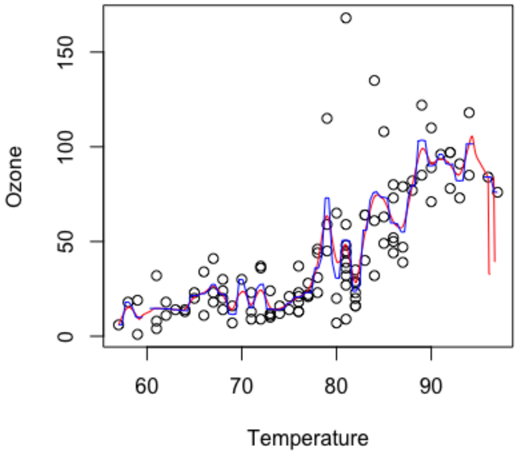
In R two functions locpoly() and ksmooth() create kernel regression data from the given data. The function locpoly() uses Gaussian kernel while ksmooth() uses Nadarya-Watson estimator. The bandwidth (h) limits the spread of the estimators. In the following graphs, the red line is the locpoly() function, Gaussian kernel, and the ksmooth() function, Nadarya-Watson estimator, is the blue line.

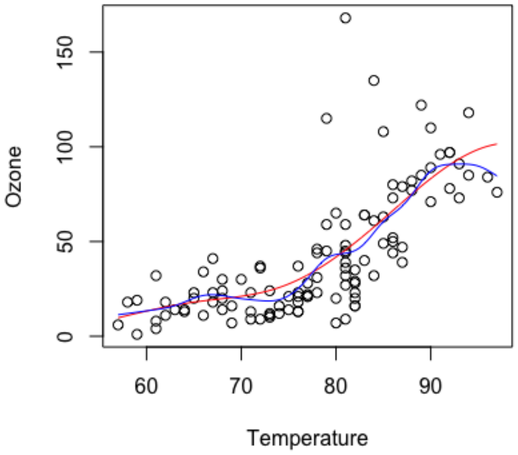
For part a.)



The first graph uses bandwidths of 10, which are already large to begin with. This is due to a lack of correlation in the data. The second graph uses bandwidths of 50. In this graph, the blue line is still bumpy while the red has smoothed out. This suggests a Gaussian kernel estimator is better for this data.

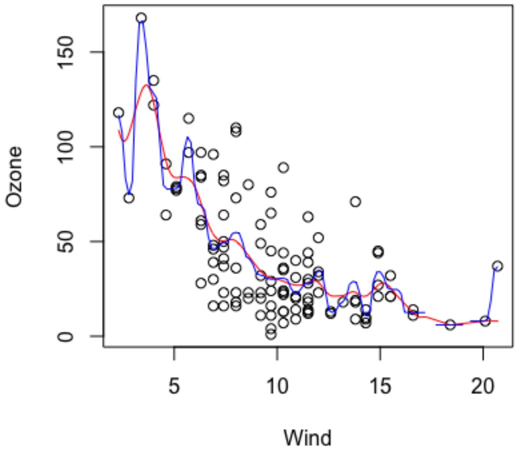
For part b.)

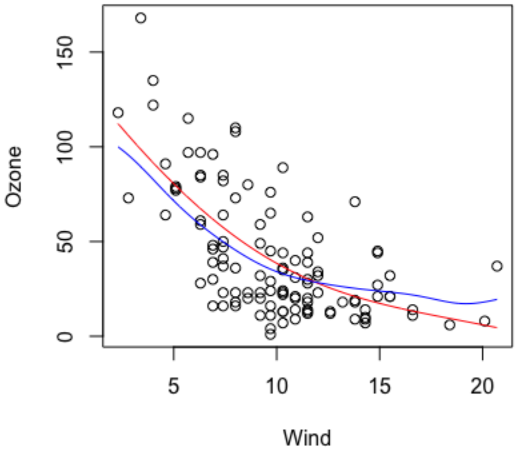




The first graph uses bandwidths of .5. I started small because the data seemed to have a more positive correlation than the data in part a. The second graph uses bandwidths of 5, and again the red line is more smooth than the blue, suggesting that a Gaussian kernel estimator is better for this data.

For part c.)





The first graph again uses bandwidth of .5. The second uses bandwidth of 5. Again, the Gaussian Kernel estimator produces a smoother line than the Nadarya-Watson.

Overall, if I'm wanting to find a kernel to make the most smooth line for this data, I will use a Gaussian Kernel estimator. If I'm wanting a line that fits the data more accurately, but isn't as smooth, I would use the Nadarya-Watson estimator.

3. . We wish to learn the relationship between ranks and variate values via an empirical study. LetX1,...,Xn i.i.d.∼X, where X has a continuous C.D.F.and E(X2)<∞. LetR1,...,Rn be the ranks of X1,...,Xn, i.e.,Ri= rank(Xi),i= 1,...,n. Please simulate dataX1,...,Xn, with n= 100, and three types of distributions of X,X∼Unif(0,1) ,X∼Exp(1), X∼N(0,1), respectively; for each distribution of X, compute the Pearson “product-moment” sample correlation coefficientŝ ρ1̂, ρ2anđ ρ3 as follows:
    a. (a)̂ρ1using{(Xi,Ri) :i= 1,...,n};
    b. (b)̂ρ2using{(Xi,Ri+1) :i= 1,...,n−1};
    c. (c)̂ρ3using{(Xi,Ri−1) :i= 2,...,n}.
    d. Please summarize your results and conclusion.

$$\boxed{\widehat{\rho} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \in [-1, 1]}.$$

The above equation is the Pearson “Product-Moment” sample correlation coefficient. For the problem, I simulated each distribution according to the parameters, and my Y variables were coded to represent the ranks of the simulated data. So instead of having R_1,…,R_n represent the ranks, I coded them as Y_1,…,Y_n. I did so to match the given equation.

For part a.)

Distribution	p coefficient
X ~ Unif(0,1)	1
X ~ Exp(1)	-1
X ~ N(0,1)	1

For part b.)

Distribution	p coefficient
X ~ Unif(0,1)	1
X ~ Exp(1)	1
X ~ N(0,1)	1

For part c.)

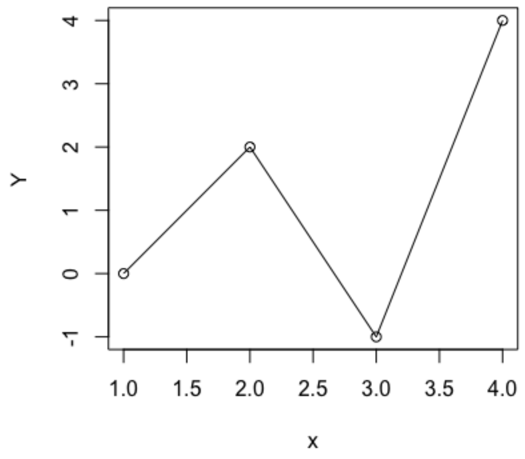
Distribution	p coefficient
X ~ Unif(0,1)	1
X ~ Exp(1)	-1
X ~ N(0,1)	1

From this, these distributions have perfect correlations. By perfect I mean they are either -1, or 1, which is the max values they can be. The graphs of Uniform and Normal distributions both go up to a peak as X increases, and then falls back down as X continues to get larger. This can be seen in Normal distributions bell curve or Uniform distributions box shaped graph. What these graphs do is for the term $(X_i – X)(Y_i – Y)$, the X and Y are the means, the product of that will always be + when the Y's are represented by the ranks because of the nature of the graphs going up and falling down. However, for Exponential distribution, because the graph grows exponentially, the graph is always rising, which allows the products of $(X_i – X)(Y_i – Y)$ to be either + or – depending on which ranks are accounted for in the calculation of p. Because ranks have perfect correlation with the distributions, and the nature of the graphs, Uniform and Normal distributions will always have a p coefficient of 1, while Exponential distributions can have a p coefficient of +1 or -1.

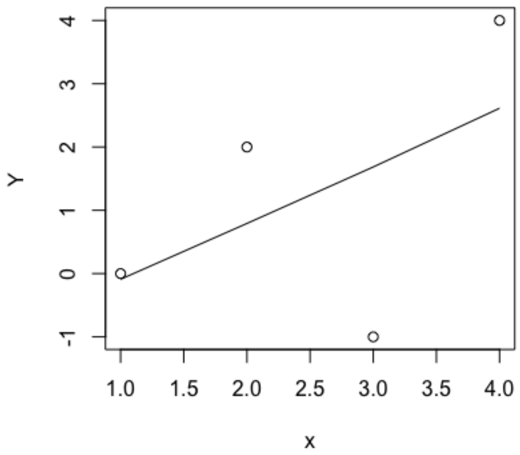
4. Let Y1= 0, Y2= 2, Y3=−1, and Y4= 4. For t ∈ [1,4], plot a curve μ(t), which is assumed to have two continuous derivatives and minimizes the penalized sum of squares, ¼ 4∑i=1 {Yi−μ(i)}^2 + λ∫41 {μ''(t)}^2 dt for λ= 0, λ= 1 and λ= 1000 separately. (Warning: the MATLAB cubic spline minimizes a slightly different function.)

The lambda in the above equation tells the spline how much to penalize the sum of squares in the first part of the equation. When lambda = 0, the minimized sum of squares will have no affect on the curve, and the higher the lambda the bigger the penalized sum of squares will be, creating a smoother line.

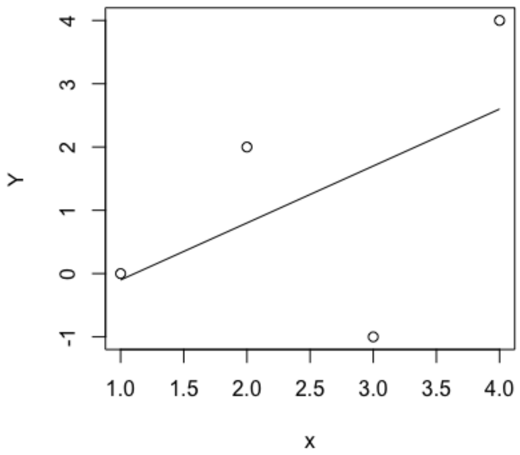
When lambda = 0



When lambda = 1



When lambda = 1000


Y
4
3
2
1
0
-1
1.0
1.5
2.0
2.5
3.0
3.5
4.0
x