ME3-HSTAT Statistics COURSEWORK
Henry Hart - 01190775
Monday 9th March 2020

1 Exploratory data analysis

This task was completed using the provided data set named 'hh2616.csv' which provides the costs of maintenance, length and age of bridges. The aim of this task is to create a model for the costs as predicted by length and age.

1.1 Summary statistics

Table 1 summarises the summary statistics of the data set. Figure 1 shows a histogram of bridge maintenance costs. The histogram is the most useful measurement of location in this case because it contains the most

Table 1: Data set summary statistics;

Statistic	Value ('000 per m)
Mean cost	4.7718
Median cost	4.4000
10% trimmed mean	4.6500
Cost interquartile range	1.6500
Cost standard deviation	1.3859

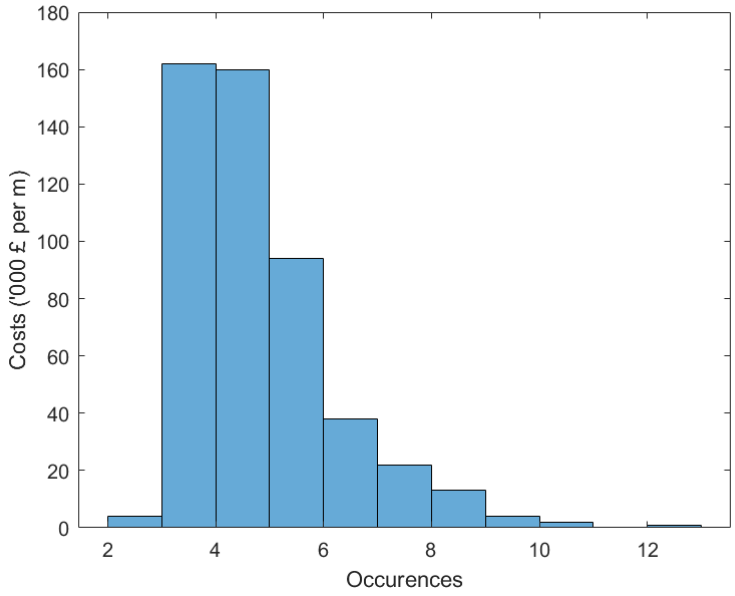

Figure 1: *Histogram of bridge maintenance costs;*

information about the distribution of costs. That being said, it is not the most compact of representations. Mean cost and cost standard deviation are the only numeric measures that are sensitive to all members of the data set and so are useful summary statistics.

Figure 2 shows scatter plots of the data in two dimensions. Effectively, these three plots are snapshots of the three dimensional plot taken from the front, top and side. It seems that cost is related to length and age, whereas there is little apparent relation between length and age themselves.

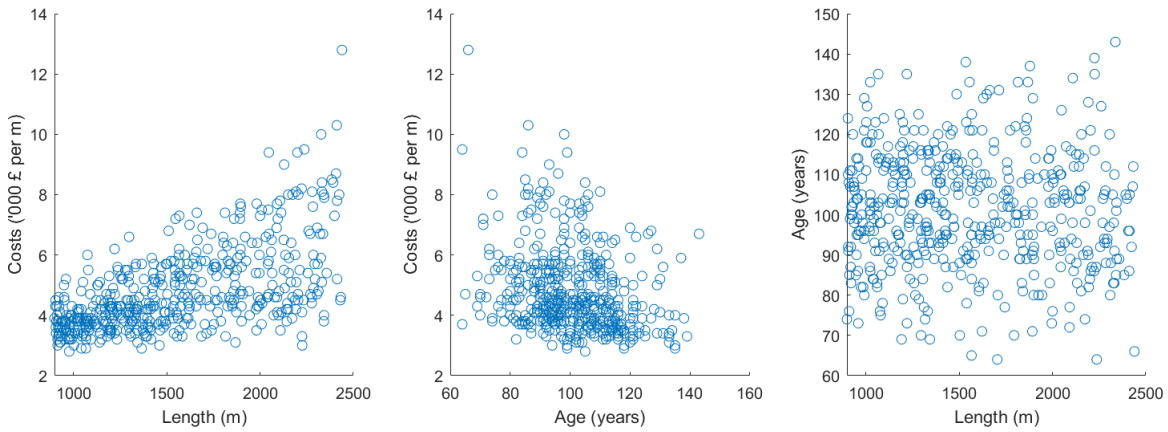

Figure 2: *Scatter plots of data;*

2 Modelling

Figure 3 shows a linear fit for the data overlaid on the cost to length scatter plot. The model seems reasonable for shorter lengths, however scatter and non-linearity appear to disrupt this model at higher lengths.

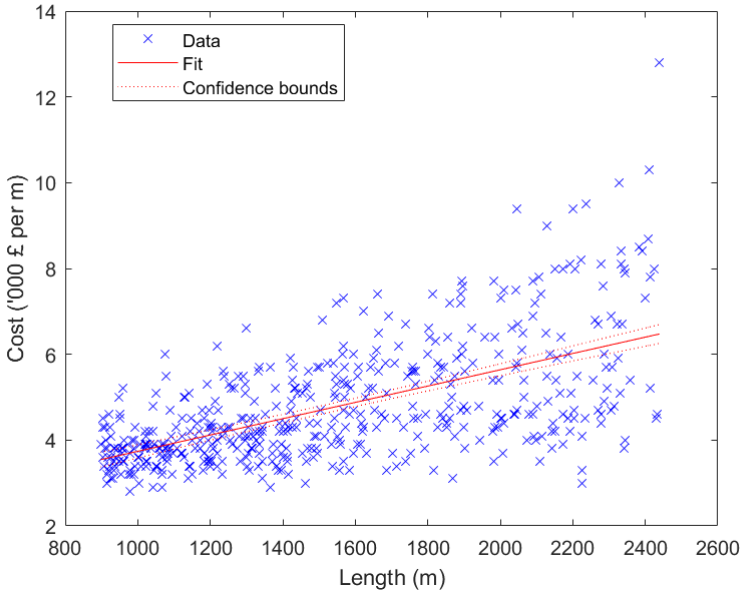

Figure 3: *Linear model for costs as a function of length;*

A number of models to predict bridge length were considered and compared to find the best fit. Table 2 summarises the models considered and the values found for comparative measures of fit.

Table 2: Models and comparative measures of fit;

Response variable	Predictors	R^2 [a]	R_a^2 [b]	MSE [c]	AIC [d]
Cost	Length	0.3646	0.3620	1.218	$1.5216 * 10^3$
log(Cost)	log(Length)	0.3703	0.3677	0.0425	-156.0804
Cost	Length^2	0.3687	0.3661	1.2102	$1.5183 * 10^3$
log(Cost)	log(Length)$^{0.5}$	0.3694	0.3669	0.0426	-155.4029
Cost	Length & Age	0.434	0.431	1.1025	$1.4641 * 10^3$
Cost	Length & Age^{-1}	0.432	0.429	1.1025	$1.4658 * 10^3$
Cost	Length & Age x Length $^{0.5}$	0.439	0.437	1.0816	$1.4589 * 10^3$

[a] Multiple coefficient of determination
[b] Adjusted multiple coefficient of determination
[c] Mean squared error
[d] Akaike Information Criterion

The coefficients of determination are a measure of fit that are close to unity when the fit is very close to the data. As such, a higher value is desired. The mean squared error is the mean of the squares of the differences between the data and the fitted model. Therefore, smaller values are desire. Finally, the Akaike information criterion measures how much information is lost when the data is represented only by the model, rather than the data points themselves. A number closer to $-\infty$ shows that more information has been retained.

Upon consideration of the values found for the comparative measures of fit across all considered models, the chosen model uses length and age x length $^{0.5}$ as the predictors for cost. This model shall be considered in greater detail henceforth.

3 Model checking and interpretation

For the chosen model, which uses length and age x length $^{0.5}$ as the predictors for cost, the residuals were calculated as the difference between the data points and the fitted values. Figure 4 shows a plot of the fitted values against the residuals.

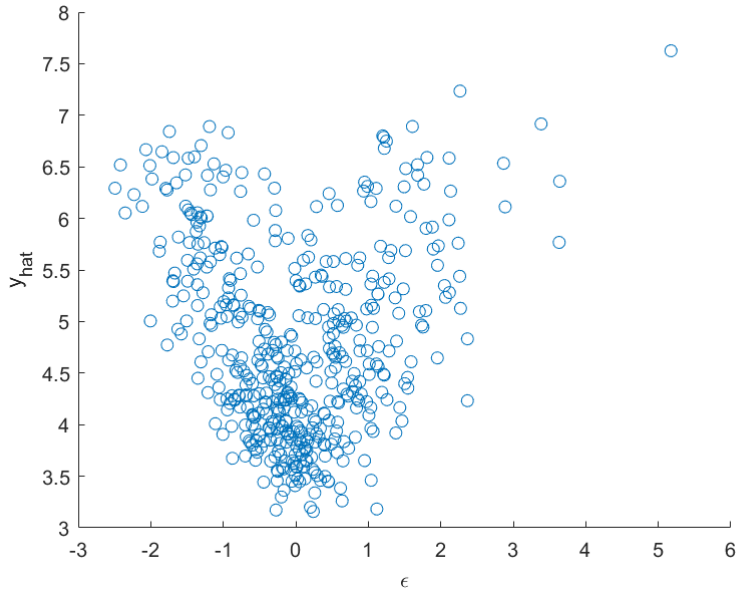

Figure 4: *Scatter plot of fitted values against residuals;*

Figure 5 shows a Q-Q plot of the quantiles of residuals taken from the data against theoretical quantiles taken from a normal distribution. Figure 4 shows that there is little relation between fitted values and residuals which is desirable. Figure 5 shows a linear trend in the quantile-quantile relation of the residuals to a normal distribution, indicating that the residuals are indeed normally distributed, at least away from the tails of the

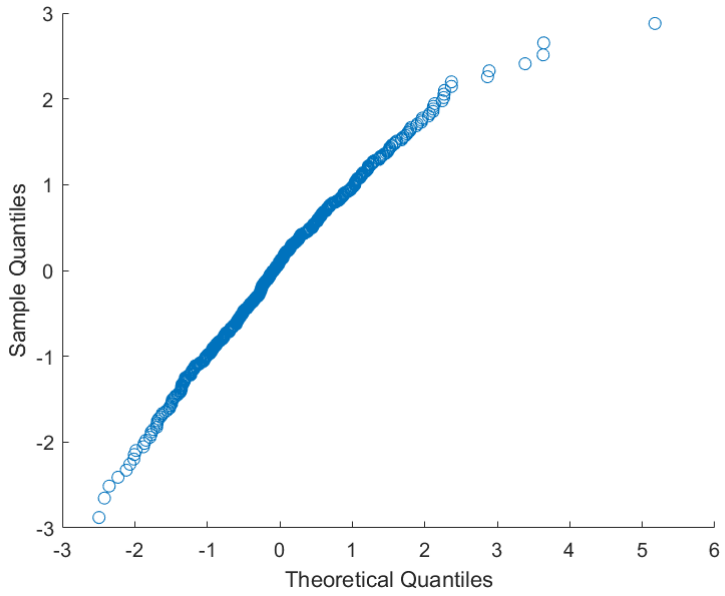

Figure 5: *Q-Q plot of residuals;*

distribution.

4 Bootstrapping

A 95% confidence band of bridge maintenance costs as a function of age was calculated using the bootstrapping method described in the coursework instructions. The code used to achieve this is shown in appendix A. The predicted cost, 95% confidence band and raw data for the model selected in section 2 is shown in figure 6.

As suggested, the inbuilt *MATLAB* function *predict*() was also used to automatically calculate confidence limits, dispensing with the use of the bootstrap method to generate a sample. The results of this method are shown in figure 7. The results of this method are very similar to those generated by the bootstrap method, but do vary slightly. This difference is likely due to the limited number of samples that were generated in the bootstrap loop.

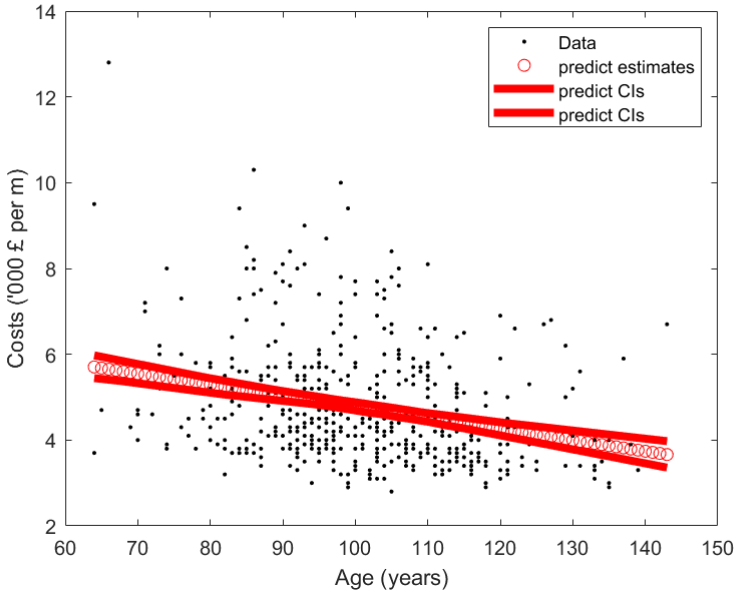

Figure 6: *Predicted cost, 95% confidence band and raw data for the model selected using the bootstrap method;*

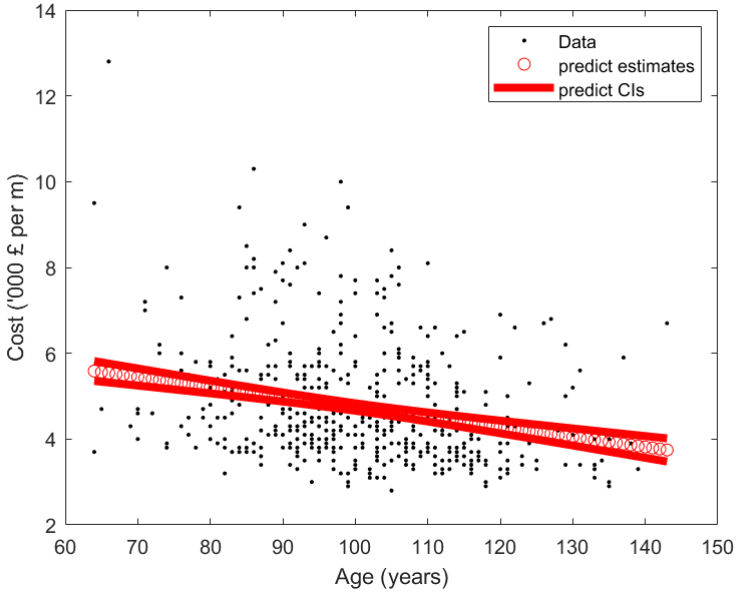

Figure 7: *Predicted cost, 95% confidence band and raw data for the model selected using the predict() function;*

A Source Code

The source code to produce results is given below:

```
% ME3-HSTAT Statistics COURSEWORK
% Issued: 24/02/2020 @ 1400 in Lecture
% Due: 09/03/2020 @ 1200 on Blackboard
% Name: Henry Hart
% CID: 01190775

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% 0.1) Setup
clear all
CID = 01190775;
seed = rng(CID); %This sets the seed for the rng

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% 0.2) Import data
data_table_header = readtable('hh2616.csv', 'HeaderLines',0);
data_table = readtable('hh2616.csv', 'HeaderLines',1);
data = table2array(data_table); %cost, length, age
cost = data(:,1);
Length = data(:,2);
age = data(:,3);


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% 1) Exploratory data analysis
% 1a)
mean_cost = mean(cost)
median_cost = median(cost)
trimmed_mean_10_cost = trimmean(cost,10)
IQR_cost = iqr(cost)
std_cost = std(cost)
edges = min(floor(cost)):1:max(ceil(cost));
histogram(cost,edges)
ylabel('Costs (000  per m)')
xlabel('Occurences')
% 1b)
figure
subplot(1,3,1)
scatter(Length,cost);
xlabel('Length (m)');
ylabel('Costs (000  per m)');
subplot(1,3,2)
scatter(age,cost);
xlabel('Age (years)');
ylabel('Costs (000  per m)');
subplot(1,3,3)
scatter(Length,age);
ylabel('Age (years)');
xlabel('Length (m)');

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% 2) Modelling
% 2a)
figure
```

```
fitlm(Length,cost);
plot(ans)
fit = ans;
xlabel('Length (m)')
ylabel('Cost (000  per m)')
title('')
% 2b)
alpha = table2array(fit.Coefficients(1,1));
beta = table2array(fit.Coefficients(2,1));
% multiple coefficient of determination
R_squared = 1 - sum((cost - (alpha + beta.*Length)).^2)...
/sum((cost - mean(cost)).^2)
% adjusted multiple coefficient of determination
n = length(cost);
q = 2;
R_squared_a = 1 - (n-1)*(1-R_squared)/(n-(q+1))
% mean squared error
MSE = sum((cost - (alpha + beta.*Length)).^2)/n
% Akaike information criterion
AIC = 2*q - 2*fit.LogLikelihood

%
% Test different models
% y = log(cost);
y = cost;
% x = log(Length);
% x = Length.^2;
% x = log(Length).^2;
% x = [age,Length];
% x = [Length,age.^-1]:
x = [Length,age.*Length.^0.5];
fitlm(x,y)
fit = ans;
AIC = 2*q - 2*fit.LogLikelihood

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% 3) Model checking and interpretation
% 3a) Residuals
y_hat = table2array(fit.Coefficients(1,1))...
+ table2array(fit.Coefficients(2,1))*x(:,1)...
+ table2array(fit.Coefficients(3,1))*x(:,2);
epsilon = y - y_hat;
figure
scatter(epsilon,y_hat);
ylabel('y_{hat}')
xlabel('\epsilon')
title('')
% Q-Q plot
epsilon_ordered = sort(epsilon);
normal_quantiles = transpose(norminv([1:500]/(n+1)));
figure
scatter(epsilon_ordered,normal_quantiles);
xlabel('Theoretical Quantiles')
ylabel('Sample Quantiles')
title('')
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
% 4) Bootstrapping
age_vector = transpose(min(age):max(age));
for k = 1:1000
    for i = 1:length(epsilon)
        epsilon_star(i,1) = epsilon(randi([1,length(epsilon)]));
    end
    y_star = epsilon_star + y_hat;
    fitty = fitlm(x,y_star);
    y_hat_j(:,k) = table2array(fitty.Coefficients(1,1))...
    + table2array(fitty.Coefficients(2,1))*mean(Length)...
    + table2array(fitty.Coefficients(3,1))...
    *(mean(Length).^0.5)*age_vector;
end
for i = 1:length(age_vector)
    ci1(i) = prctile(y_hat_j(i,:),2.5);
    ci2(i) = prctile(y_hat_j(i,:),97.5);
    pred(i) = mean(y_hat_j(i,:));
end
figure
h1 = plot(age,cost,'k.');
hold on
h2 = plot(age_vector,pred,'ro');
h3 = plot(age_vector,ci1,'r-','LineWidth',4);
h4 = plot(age_vector,ci2,'r-','LineWidth',4);
legend([h1; h2; h3; h4], ...
    {'Data','predict estimates','predict CIs', 'predict CIs'});
xlabel('Age (years)');
ylabel('Costs (000  per m)');
[ypred,yci] = predict(fitty,[mean(Length)*ones(length(age_vector),1),...
    (mean(Length)^0.5)*age_vector])
figure
h1 = plot(age,cost,'k.');
hold on
h2 = plot(age_vector,ypred,'ro');
h3 = plot(age_vector,yci,'r-','LineWidth',4);
legend([h1; h2; h3], ...
    {'Data','predict estimates','predict CIs'});
xlabel('Age (years)');
ylabel('Cost (000  per m)');
```