# ECEN321 - Lab 5
# Regression

Joshua Benfell - 300433229

June 22, 2020

## 1 Introduction

Regression is a common method used to analyse the correlation between two variables. In this report, we will be using regression methods to perform analysis on the temperature data collected around Wellington, New Zealand over a period of 161 years [1]. From here we will be able to find out how well the data correlates and in which direction and potentially determine how strong this correlation is.

## 2 Method

Before working with the temperature data, a trial run will be conducted using a sample set of data where the outcome is known. For this the line $y = x$ will have gaussian distributed noise added to it and act as the trial data, as we know what the original line looks like.

To perform the regression, we will first compute the sample correlation coefficient. This will provide an indication of correlation and direction. This will be confirmed graphically by computing the least-squares line. This is done by using the equations eqs. (1) to (3).

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x} \tag{1}$$

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2}$$

$$\hat{y_i} = \hat{\beta_0} + \hat{\beta_1}x_i \tag{3}$$

$$s = \sqrt{(1-r^2)\sum(\frac{(y-\bar{y})^2}{n-2}} \tag{4}$$

Once plotted the standard deviation can be estimated with eq. (4). This estimation can be used to confirm the methods work by comparing it to the standard deviation of the imposed noise on the line $y = x$.

Finally, we will determine the confidence that the relation found for the temperature is this. This will be done through the method of hypothesis testing. We have $H_0 : \rho = 0$, indicating no correlation between the year and temperature and $H_a : \rho \neq 0$, indicating that there is a correlation between the year and the temperature.

## 3 Results

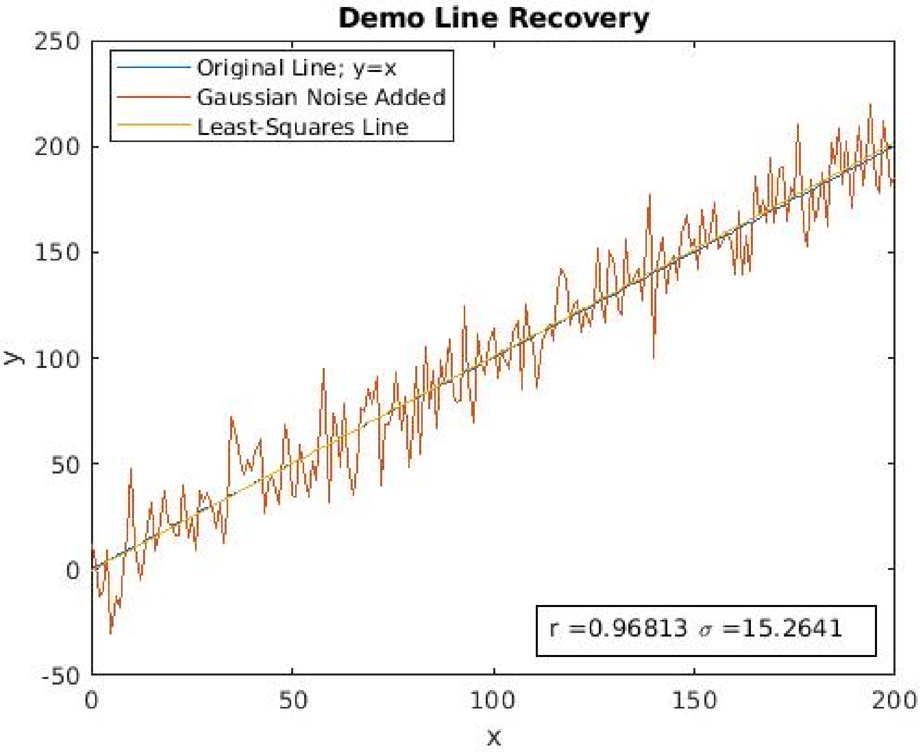

Figure 1: Regression Methods Applied to a Noisy $y = x$ with $\sigma = 16$

fig. 1 Shows the results of the trial run of the used regression methods. We can see that the resulting line matches closely to the original and the correlation coefficient, being close to 1, indicates that this line is similar to the original. Furthermore, the retrieved standard deviation from this plot is fairly close to the standard deviation that was provided to the gaussian noise. Therefore we are able to conclude that the chosen methods work. fig. 2 shows the results of performing the same methods as performed above. From this we can see that there is in fact a positive trend and that the year and temperature are reasonably positively correlated with a $r = 0.68023$. This indicates that the average temperature in wellington is gradually increasing each year.

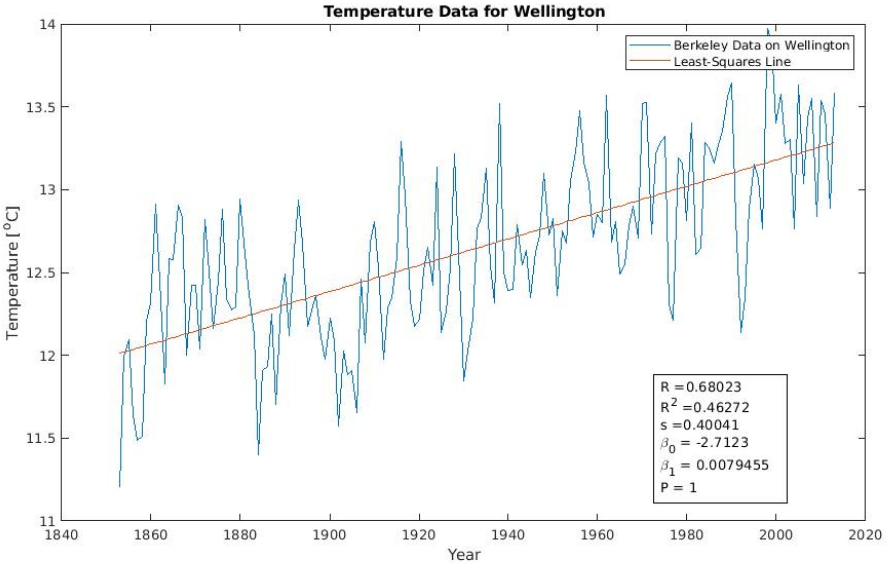

Figure 2: Regression Methods Applied to Wellington Temperature Data

$$U = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{5}$$

Next the hypothesis test is performed. For this, we are going to use eq. (5), which is t distributed and can be used as a way of performing the test of the correlation coefficient. Using the t distribution, as seen in fig. 2, we get a P of 1. This indicates that we can fully reject the null hypothesis as this sits outside of the t curve entirely and any confidence intervals we would normally compare this P to. So it is possible to confirm, very confidently, that the average temperature is correlated to the year, whether that be a positive or negative relationship.

Another thing fig. 2 shows is that the standard deviation of the temperature is 0.4004 indicating that it doesn't vary by much, making the trend appear stronger.

## 4 Conclusion

In conclusion, the world is on fire and will slowly cook us alive if we don't take early action to reduce the temperature of the planet. The regression methods were certainly useful, however, I am not entirely certain that the method used for obtaining the standard devitation. Although it appears correct, based off multiple reruns of the program resulting in values between 15 and 17 for a provided standard deviation of 16, there are other equations used in the lecture slides to come to many different expressions for the standard deviation. The equation used in this report is even used in them, indicating that it may not be the actual standard deviation of the plot, but more, a stepping stone to get to the estimate. Additionally, I could not get the 95% confidence intervals working as they appeared at temperatures of -188, which is clearly not right as that is an unlikely temperature to reach unless we were in an iceage, which I'm fairly confident we are currently not. So this method would need some more tweaking to be able to show those confidence intervals.

## References

[1] 2020. [Online]. Available: http://berkeleyearth.org/

# Appendices

## A Regression Testing Code

```
clear
close all
x = 0:200;
ss = 16; %std dev
y = x + randn(1, length(x)) * ss; % add noise to line y
    =x
plot(x,x) % Original Line y=x
hold on
plot(x,y) % noisy line
rr = corrcoef(x,y);
rr = rr(1,2);

xbar = mean(x);
ybar = mean(y);
B_1hat = sum((x-xbar).*(y-ybar)) ./ sum(power((x-xbar),2));
B_0hat = ybar - B_1hat * xbar;
yhat = B_0hat + B_1hat * x;
plot(x, yhat);

n = length(y);
ss_yhat = sqrt(((1-rr^2) .* sum(power((y - ybar),2))) ./(n-2)) %estimate of the std dev

legend("Original Line; y=x", "Gaussian Noise Added", "Least-Squares Line");
title("Demo Line Recovery")
xlabel("x")
ylabel("y")

dim = [.65 0 .9 0.25];
str = strcat('r = ', num2str(rr),newline,' \sigma = ', num2str(ss_yhat));
annotation('textbox', dim, 'String', str, 'FitBoxToText', 'on');
```

## B MATLAB Code

```
clear
close all
data = readtable('data.csv');
data.Unc__4 = str2double(data.Unc__4); % last column was a string for some reason
averageTemp = 12.93;
h = height(data);
temps = zeros(1, 12);
years = (data.Year(1):data.Year(end))';
currentYear = data.Year(1);
yearlyAverageTemps = zeros(length(years), 1);

YATIndex = 1;
```

```
index = 1; % temps index; index - 1 = num of non-NaN values for a given year values.
for r = 1:h
    if(data.Year(r) ~= currentYear || r == h)
        yearlyAverageTemps(YATIndex) = mean(temps);
        yearlyAverageTemps(YATIndex) = yearlyAverageTemps(YATIndex) * 12/(index-1); % fixes the fact that temps is always 12 long
        temps = zeros(1, 12); % reset
        currentYear = data.Year(r); %increment
        index = 1; % reset
        YATIndex = YATIndex + 1; %Increment
    end
    anom = data.Anomaly(r);
    if not(isnan(anom))
        temps(index) = anom; % Get current months anomaly
        index = index + 1; % increment
    end

end

yearlyAverageTemps = yearlyAverageTemps + averageTemp;
plot(years, yearlyAverageTemps);

RR = corrcoef(years,yearlyAverageTemps);
RR = RR(1,2);
RR_sq = RR^2;

% Least Squares Line
xbar = mean(years);
ybar = mean(yearlyAverageTemps);

B_1hat = sum((years - xbar) .* (yearlyAverageTemps - ybar)) ./ sum(power((years-xbar),2));
B_0hat = ybar - B_1hat * xbar;
yhat = B_0hat + B_1hat * years;

n = length(yearlyAverageTemps);
ss_yhat = sqrt(((1-RR^2) .* sum(power((
    yearlyAverageTemps - ybar),2)))./(n-2)); %estimate of the std dev

hold on
plot(years, yhat);
legend("Berkeley Data on Wellington", "Least-Squares Line")
xlabel("Year")
ylabel("Temperature [^oC]")
title("Temperature Data for Wellington")

% Null Hypothesis: Rho = 0
% Alternate Hypothesis: Rho ~= 0
U = (RR * sqrt(n-2)) / sqrt(1-RR^2);
dof = n -1;
P = tcdf(U, dof);

dim = [.7 0.1 .9 0.25];
str = {strcat('R = ', num2str(RR)),strcat('R^2 = ', num2str(RR_sq)), strcat('s = ', num2str(ss_yhat)), strcat("\beta_0 = ", num2str(B_0hat)), strcat("\beta_1 = ", num2str(B_1hat)), strcat("P = ", num2str(P))};
annotation('textbox', dim, 'String', str, 'FitBoxToText', 'on');
set(gcf, 'Position', [400, 400, 1000, 600]);
```