Code

Has the Mean Annual Temperature in Australia Increased Over the Past 100 years?

A statistical investigation conducted by

Ashley Mallia s3773716 & Daniel Evans s3766654

02/06/2019

Problem Statement

Given the scientific consensus that our climate is changing unusually rapidly.

The Paris Agreement sets a global goal to hold average temperature increase to well below 2°C and pursue efforts to keep warming below 1.5°C above pre-industrial levels.

It is predicted that should warming reach 4°C above preindustrial levels, the earth would be impacted by unprecedented heat waves, severe drought and major floods in many regions.

Given the importance of client we thought it be appropriate to perfom a statistical analysis on Australian climate data from the past 100 years to determine:

1. Does the mean climate temperature from 1910-1950 differ, to a statistically significant degree, from the mean climate temperature from 1978-2018?
2. Does a statistically significant linear relationship exists between Year and Temperature?
3. In what years will we reach global warming of 1.5°C?

- To answer question 1, we split our data into two 40 year ranges: “before” for years 1910-1950,“after” for years 1978-2018. We will use a two sample (independent) t-test between the before and after datasets to determine if the difference between them is statisticly significant or not.
- To answer question 2, we create a least squares fit linear regression model between Year and Temperature and use statistical tests to verify if a statistically significant linear relationship exists.
- To answer question 3, we will use the regression model to predict in what years will we reach global warming of 1.5°C

Data

Our data source is the ACORN-SAT (Australian Climate Observations Reference Network-Surface Air Temperature) homogeneous temperature timeseries accessable through the Bureau of Meteorology (Ref 1).

The BoM collects the ACORN-SAT temperature data from a network of recording stations. They apply advanced statistical techniques to properly normalize and weight this data into daily and monthly averages.

Our dataset is the Australian Monthly Temperature Anomalies (or departures from the 1961–1990 average (21.8°C)) for the time period Jan-1910 to Apr-2019.

Data Variables

1. Year (integer) - from 1910 to 2019
2. Month (factor) - 12 abbriviated month names Jan to Dec
3. Temp (numeric) - temperature departure from long term average for specific year and month

Preprocessing

1. The datasource was supplied in wide format. We used tidyr::gather to convert this into long format
2. We exclude the Year 2019, since it is an incomplete year

Hide

```
(allmonths <- read_csv("data/allmonths.csv", col_types = cols(Year = col_integer())))
```

Year <int>	Jan <dbl>	Feb <dbl>	Mar <dbl>	Apr <dbl>	May <dbl>	Jun <dbl>	Jul <dbl>	Aug <dbl>	Sep <dbl>	Oct <dbl>	Nov <dbl>	Dec <dbl>
1910	-0.28	-0.35	-1.77	-0.28	0.20	-0.02	0.06	0.36	0.64	-1.78	-1.33	-1.57
1911	-0.49	-1.16	-0.88	-1.18	-0.86	-1.68	-0.33	-0.33	-0.59	-0.39	0.40	-0.64
1912	0.18	0.61	0.20	-0.99	-0.37	-0.72	-0.35	0.03	0.00	-0.06	-0.77	-0.21
1913	-0.52	-0.75	-1.67	-0.46	-2.66	-1.97	-0.28	-0.60	-1.21	-0.21	-0.24	0.25
1914	-0.46	0.48	0.31	0.33	-0.18	-0.33	-0.93	0.73	-0.07	0.32	0.87	0.35
1915	-0.86	0.90	0.23	0.15	-1.07	0.78	1.15	-0.58	0.77	-0.46	0.33	-0.43
1916	-0.19	-0.11	-0.03	-1.49	0.35	0.56	-0.61	0.05	0.11	-1.21	-2.55	-1.74
1917	-1.39	-2.83	-1.37	-1.67	-2.40	-1.44	0.69	-0.84	-0.28	-0.73	-1.76	-0.92
1918	-1.01	-1.26	-1.73	-0.80	-0.43	0.23	-1.41	-0.10	-0.22	-0.14	0.16	0.27

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1919	-0.05	-0.48	-0.56	0.56	-0.36	0.33	-0.62	-0.42	-0.57	-0.53	0.33	0.52

1-10 of 110 rows Previous **1** 2 3 4 5 6 ... 11 Next

Hide

```
allmonths <- allmonths %>% gather(Month, Temp, Jan:Dec, factor_key = T) %>% arrange(Year, Month)
(allmonths <- allmonths %>% filter(Year != 2019))
```

Year	Month	Temp
<int>	<fctr>	<dbl>
1910	Jan	-0.28
1910	Feb	-0.35
1910	Mar	-1.77
1910	Apr	-0.28
1910	May	0.20
1910	Jun	-0.02
1910	Jul	0.06
1910	Aug	0.36
1910	Sep	0.64
1910	Oct	-1.78

1-10 of 1,308 rows Previous **1** 2 3 4 5 6 ... 100 Next

Descriptive Statistics and Visualisation

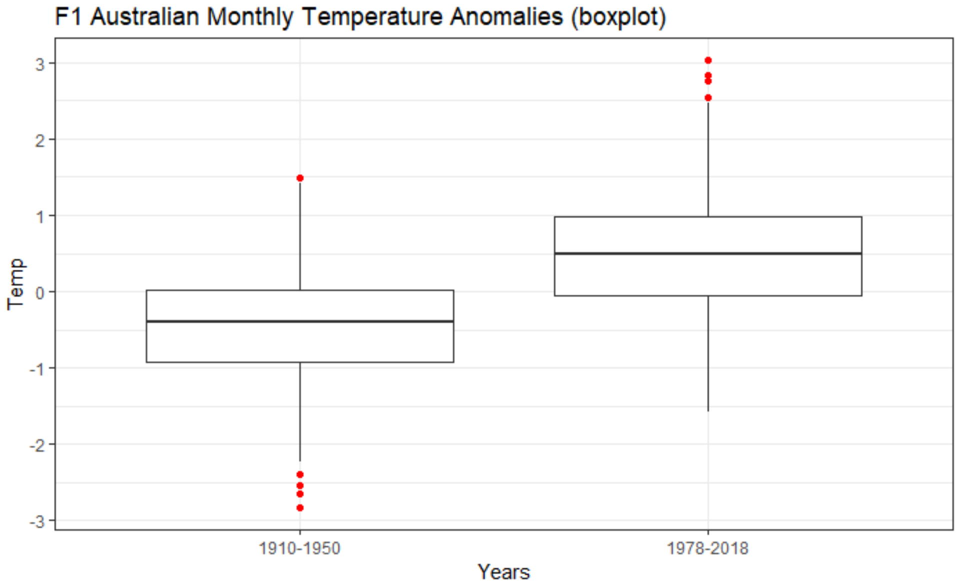
Hide

```
before.range = "1910-1950"
after.range  = "1978-2018"

before <- allmonths %>% filter(Year >= 1910 & Year <= 1950) %>% mutate(Years = before.range)
after  <- allmonths %>% filter(Year >= 1978 & Year <= 2018) %>% mutate(Years = after.range)

group <- bind_rows(before, after)

group %>% group_by(Years) %>% summarise(
  N   = n(),
  MEAN = mean(Temp, na.rm = T),
  SD  = sd(Temp, na.rm = T),
  MIN = min(Temp, na.rm = T),
  Q1  = quantile(Temp, .25, na.rm = T),
  MEDIAN = quantile(Temp, .5, na.rm = T),
  Q3   = quantile(Temp, .75, na.rm = T),
  MAX  = max(Temp, na.rm = T),
  IQR  = Q3-Q1,
  LF = Q1 - 1.5*IQR,
  UF = Q3 + 1.5*IQR,
  LOUT = sum(Temp < LF, na.rm = T),
  UOUT = sum(Temp > UF, na.rm = T)
) %>%
arrange(Years) %>% modify_if(is.numeric, round, 2) %>% kable
```

Years	N	MEAN	SD	MIN	Q1	MEDIAN	Q3	MAX	IQR	LF	UF	LOUT	UOUT
1910-1950	492	-0.43	0.73	-2.83	-0.92	-0.40	0.02	1.49	0.94	-2.33	1.42	4	1
1978-2018	492	0.46	0.80	-1.58	-0.05	0.49	0.98	3.03	1.03	-1.60	2.53	0	4

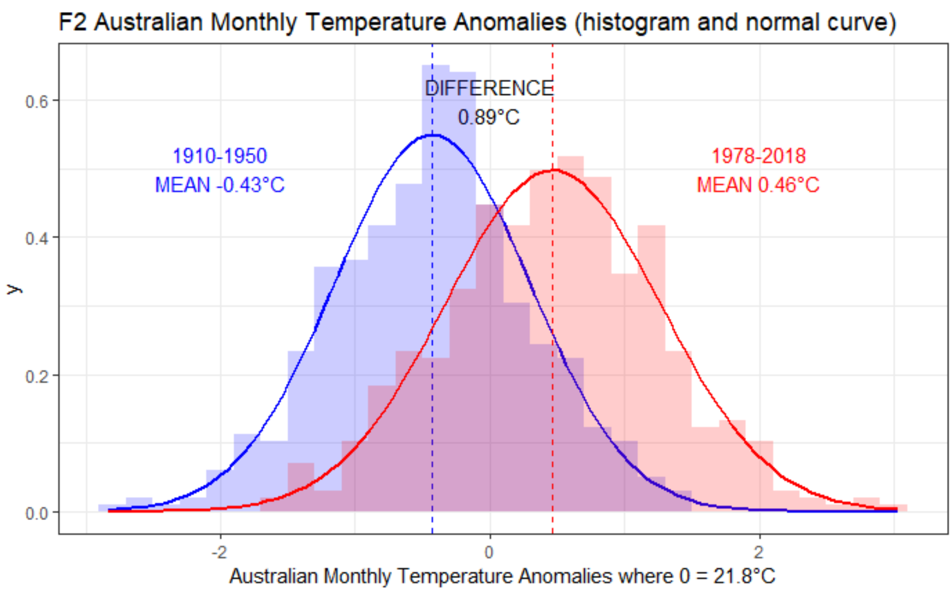
Hide

```
before.label = paste0(before.range, "\nMEAN ", round(mean(before$Temp),2), "°C")
after.label  = paste0(after.range,  "\nMEAN ", round(mean(after$Temp),2), "°C")
abdiff.label = paste0("DIFFERENCE\n", round(mean(after$Temp-before$Temp),2), "°C")

ggplot(group, aes(x = Years, y = Temp)) + theme_bw() +
  ggtitle("F1 Australian Monthly Temperature Anomalies (boxplot)") +
  geom_boxplot(outlier.colour = "red")
```



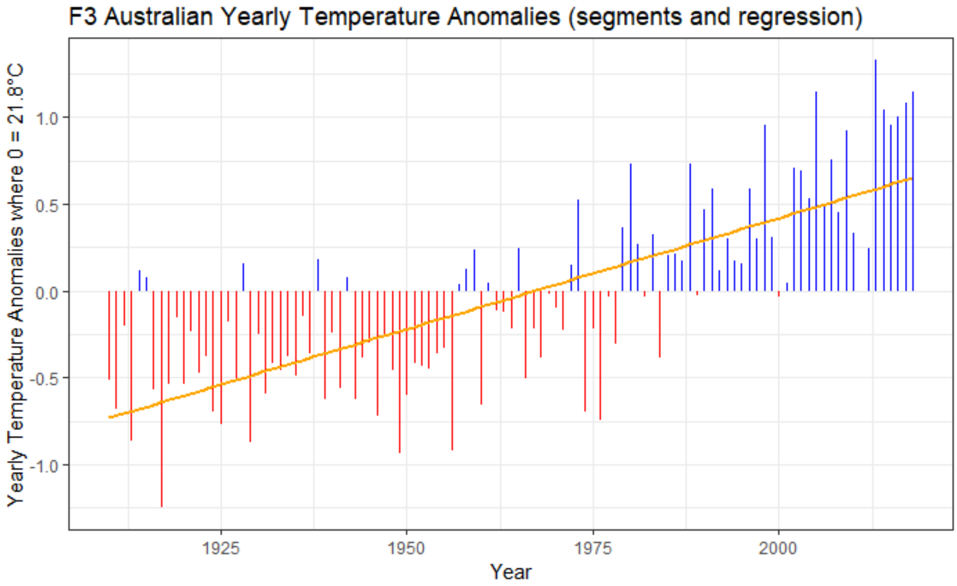
Hide

```
ggplot() + theme_bw() +
  ggtitle("F2 Australian Monthly Temperature Anomalies (histogram and normal curve)") +
  xlab("Australian Monthly Temperature Anomalies where 0 = 21.8°C") +
  annotate(geom="text", x=0, y=.6, label=abdiff.label, color="black") +
  #before
  geom_histogram(aes(x = before$Temp, y=..density..), binwidth = .2, alpha = .2, fill="blue") +
  stat_function(aes(x = before$Temp), fun = function(x) dnorm(x, mean(before$Temp), sd(before$Temp)), color="blue", size=1)
 +
  geom_vline(aes(xintercept=mean(before$Temp)), color="blue", linetype="dashed") +
  annotate(geom="text", x=-2, y=.5, label=before.label, color="blue") +
  #after
  geom_histogram(aes(x = after$Temp, y=..density..), binwidth = .2, alpha = .2, fill="red") +
  stat_function(aes(x = after$Temp), fun = function(x) dnorm(x, mean(after$Temp), sd(after$Temp)), color="red", size=1) +
  geom_vline(aes(xintercept=mean(after$Temp)), color="red", linetype="dashed") +
  annotate(geom="text", x=2, y=.5, label=after.label, color="red")
```



Hide

```
allyears <- allmonths %>% group_by(Year) %>% summarise(Temp = mean(Temp))


ggplot(allyears, aes(x=Year,xend=Year,y=0,yend=Temp,color=ifelse(Temp<0, 'red','blue'))) +
  theme_bw() + ggtitle("F3 Australian Yearly Temperature Anomalies (segments and regression)") +
  xlab("Year") + ylab("Yearly Temperature Anomalies where 0 = 21.8°C") +
  geom_segment() + scale_color_identity() +
  geom_smooth(aes(x=Year, y=Temp),  method = "lm", col = "orange", se=F)
```



Code

The summary statistics show:

- mean temperature from 1910-1950 is -0.43°C while mean temperature from 1978-2018 is 0.46°C
- difference in means of 0.89°C. We will determine if this is a statistically significant difference
- the standard errors of both groups are similar (confirms homogeneity of variance)

Fig 1 and Fig 2 shows:

- the datasets are normally distributed (confirmed visually and since mean similar to median)
- a few outliers are not influential and so have been left in the dataset.
- the only missing data was in the year 2019 and so that year has been excluded

Fig 3 shows:

- 100 years of Australian Yearly Temperature Anomalies
- A regression line calculated later in the document $Temp = -25.156 + 0.0128\, Year$

Hypothesis Testing

Significance Level

- Due to our large sample size we choose to use a significance level $\alpha = 0.01$

Hypotheses Testing - Two Sample t-Test

Hypotheses

- $H_0 : \mu_1 - \mu_2 = 0$
- $H_A : \mu_1 - \mu_2 \neq 0$

Assumptions

- This data is continuous: **TRUE**
- Comparing two independent population means with unknown population variance: **TRUE**
- Population data are normally distributed or large sample used (n>30 for both groups): **TRUE**
- Population homogeneity of variance: **TRUE**

Decision Rules

- **Reject** H_0 if
  - p-value < the significance level α
  - if 95% CI of the parameter does not capture H_0
- Otherwise **Fail to Reject** H_0

Conclusion

- Test will be statistically significant if we reject H_0
- Otherwise, the test is not statistically significant.

Results

Hide

```
t.test(after$Temp, before$Temp, var.equal = T, conf.level = .99)
```

```
	Two Sample t-test

data:  after$Temp and before$Temp
t = 18.338, df = 982, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 0.7679886 1.0195724
sample estimates:
 mean of x  mean of y 
 0.4648374 -0.4289431 
```

- t(df=982) = 18.338, p < 0.001, 99% CI [0.77, 1.02]
- Actual mean difference $\mu_1 - \mu_2 = 0.894$
- Given these findings we **Reject** H_0

The data provides evidence to conclude that the mean temperature from 1910-1950 does differ to a statistically significant degree from the mean temperature from 1978-2018

Hypotheses Testing - Linear Regression: Overall Model

Hypotheses

- H_0 : The data does not fit the linear regression model
- H_A : The data fits the linear regression model
- Test model parameters using F-test

Assumptions

- Independence (check research design)
- Linearity (check scatter plot)
- Normality of residuals (check after model is fitted)
- Homoscedasticity (check after model is fitted)

Decision Rules

- **Reject** H_0 if p-value < the significance level α
- Otherwise **Fail to Reject** H_0

Conclusion * Test will be statistically significant if we reject H_0 * Otherwise, the test is not statistically significant.

Hypotheses Testing - Linear Regression: Model Parameters

Hypotheses

- Intercept: $H_0 : \alpha = 0 \quad H_A : \alpha \neq 0$
- Slope: $H_0 : \beta = 0 \quad H_A : \beta \neq 0$
- Test model parameters using t-test

Assumptions

- Same as in overall model

Decision Rules

- **Reject** H_0 if
    - p-value < the significance level α
    - if 95% CI of the parameter does not capture H_0
- Otherwise **Fail to Reject** H_0

Conclusion

- Test will be statistically significant if we reject H_0
- Otherwise, the test is not statistically significant.

Results

Hide

```
l <- lm(Temp ~ Year, data = allyears)
summary(l)
```

```
Call:
lm(formula = Temp ~ Year, data = allyears)

Residuals:
     Min       1Q   Median       3Q      Max
-0.85958 -0.21854 -0.02217  0.24316  0.79746

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -25.155869   2.123991  -11.84   <2e-16 ***
Year          0.012788   0.001081   11.83   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3552 on 107 degrees of freedom
Multiple R-squared:  0.5666,    Adjusted R-squared:  0.5625
F-statistic: 139.9 on 1 and 107 DF,  p-value: < 2.2e-16
```

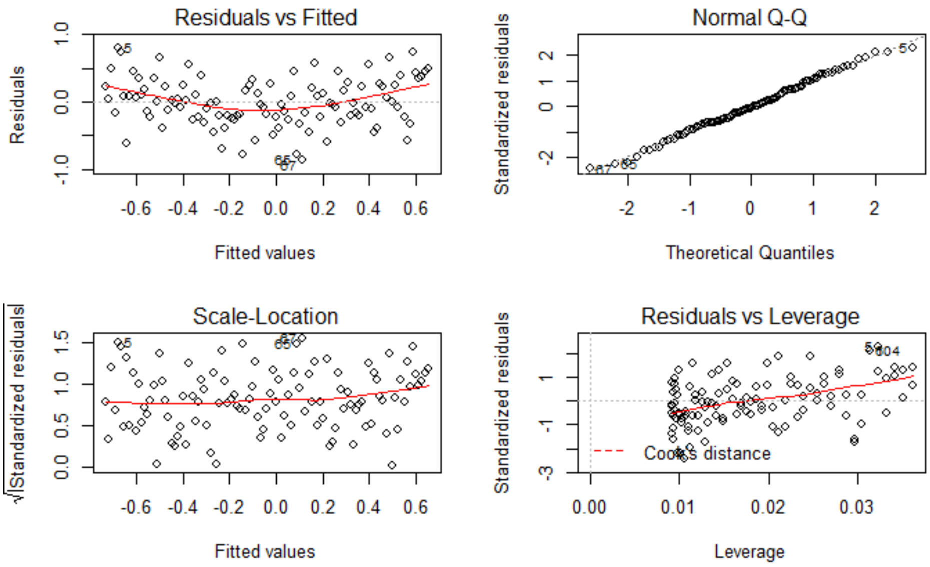
Hide

```
cat("Coefficients 99% Confidence Intervals:\n"); confint(l, level=.99) %>% round(3)
```

```
Coefficients 99% Confidence Intervals:
              0.5 %  99.5 %
(Intercept) -30.726 -19.586
Year          0.010   0.016
```

Hide

```
par(mfrow=c(2,2))
plot(l)
par(mfrow=c(1,1))
```



Diagnostic Plots

- Normality of residuals OK, Linearity OK, Homoscedasticity OK, No influential outliers OK

Overall Model

- f(1,107) = 139.9, p < 0.001. **Reject** H_0

Intercept

- a = -25.156, t(107) = -11.84, p < 0.001, 99% CI [-30.726 -19.586]. **Reject** H_0

Slope

- b = 0.0128, t(107) = 11.83, p < 0.001, 99% CI [0.010 0.016]. **Reject** H_0

Correlation Coe

- $r^2 = 0.5666$ meaning 57% of the variability in temperature can be explained by a linear relationship with the year.

LSR Model

$$Temp = -25.156 + 0.0128\,Year$$

Prediction

Hide

```
yearEst <- function(t) { (t+25.155869)/0.012788 }
tibble(Temp = 1.5, Year = yearEst(Temp) %>% round(0)) %>% kable
```

Temp	Year
1.5	2084

The linear model predicts (based on current data) that we will see global warming of 1.5°C in 2084.

Discussion

This research has answered the questions posed in the problem statement.

1. A two sample t.test was used to show that Question 1 is TRUE. The mean temperature from 1910-1950 differs from the mean temperature from 1978-2018 by 0.894°C. This is statistically significant t(df=982) = 18.338, p < 0.001, 99% CI [0.77, 1.02]
2. F-test and t-tests performed on LSR model showed that question 2 is TRUE. A positive linear relationship between Temperature and Year was found. This is statistically significant. $R^2 = 0.5666$ meaning 57% of the variability in temperature can be explained by a linear relationship with the year. The model formula was discovered to be $Temp = -25.156 + 0.0128\,Year$
3. The linear model predicts (based on current data) that we will see global warming of 1.5°C in 2084.

Discuss any strengths and limitations

A strength of our analysis were our large sample size, clear analysis and beautiful visulisations. A limitation is that we were only able to model linear regression relationships, since even basic climate models are both non-linear and multivariate.

Future investigation

While we have showed that there has been a statisticly significant temperature increase, we have not performed a comparison to correlation to the cause. It would be insightful to perform an analysis between temperature and atmospheric CO_2. Using more advanced non-linear and multivarite statistical methods would be able to find relationships that simple linear regression cannot.

References

1. Australian climate variability & change http://www.bom.gov.au/climate/change/#tabs=Tracker&tracker=timeseries&tQ=graph%3Dtmean%26area%3Daus%26season%3Dallmonths%26ave_yr (http://www.bom.gov.au/climate/change/#tabs=Tracker&tracker=timeseries&tQ=graph%3Dtmean%26area%3Daus%26season%3Dallmonths%26ave_y
2. https://en.wikipedia.org/wiki/Paris_Agreement (https://en.wikipedia.org/wiki/Paris_Agreement)