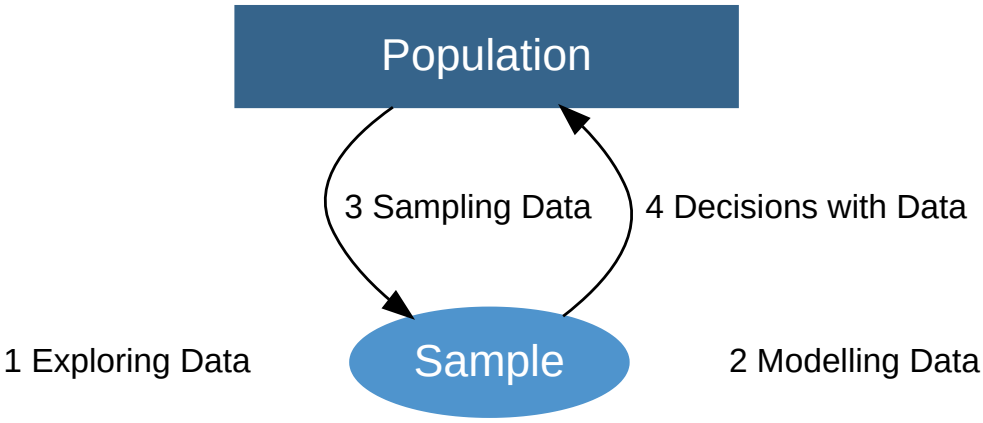
Regression Test

Decisions with Data | Tests for relationships

Unit Overview



Module4 Decisions with Data

Test for a Proportion

How can we make evidence based decisions? Is an observed result due to chance or something else? How can we test whether a population has a certain proportion?

Tests for a Mean

How can we test whether a population has a certain mean? Or whether 2 populations have the same mean?

Tests for a Relationship [DATA1001/MATH1115]

How can we test whether 2 variables are linearly related? How can we test whether a categorical variable is in certain proportions?

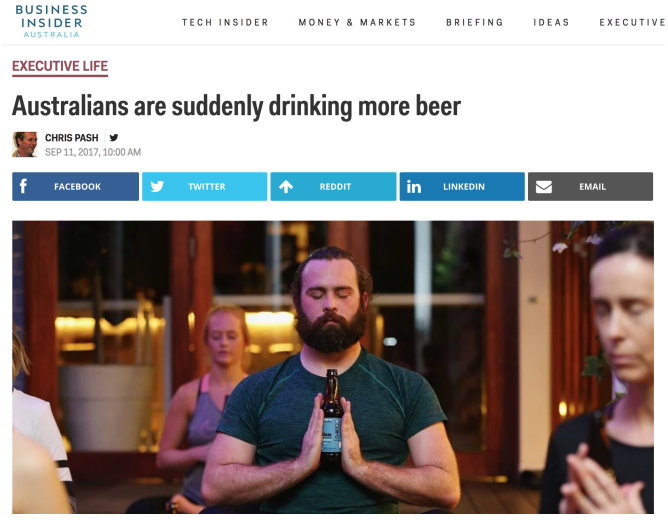
Topic 36 Regression test

Data Stories

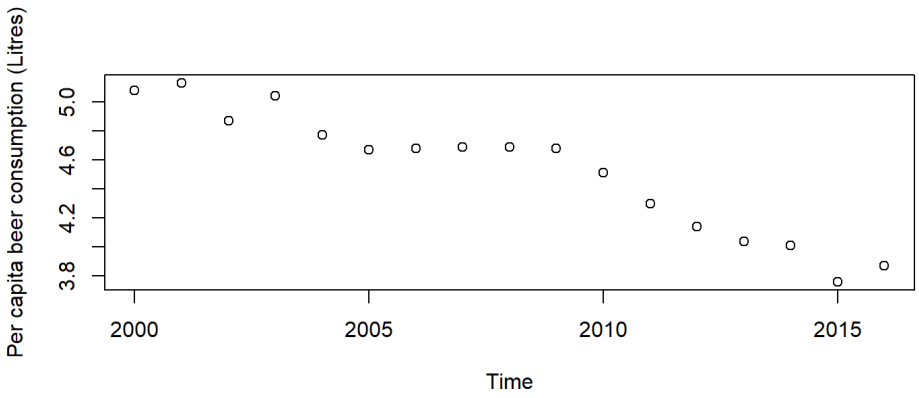
What do beer and rainfall have in common?

Trends in beer and rainfall

- Recent articles suggest that beer consumption is on the increase. Is that true?
- Is the decreasing trend in rainfall in Perth significant and linked to climate change?



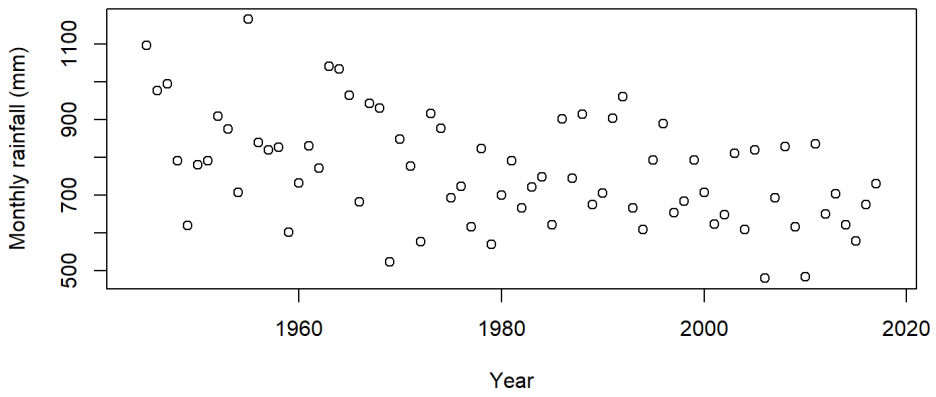
Beer



Data: http://stat.data.abs.gov.au/Index.aspx?DataSetCode=ALC

```
BeerWineSpirit = read.csv("data/WineBeerSpirits.csv")
PCBeer = subset(BeerWineSpirit, Measure == "Per capita apparent consumption (litres)" &
    Beverage.Type == "Beer")
```

Rainfall



Data: http://www.bom.gov.au/climate/data/

```
Rain = read.csv("data/PerthRain.csv")
```

Statistical Thinking

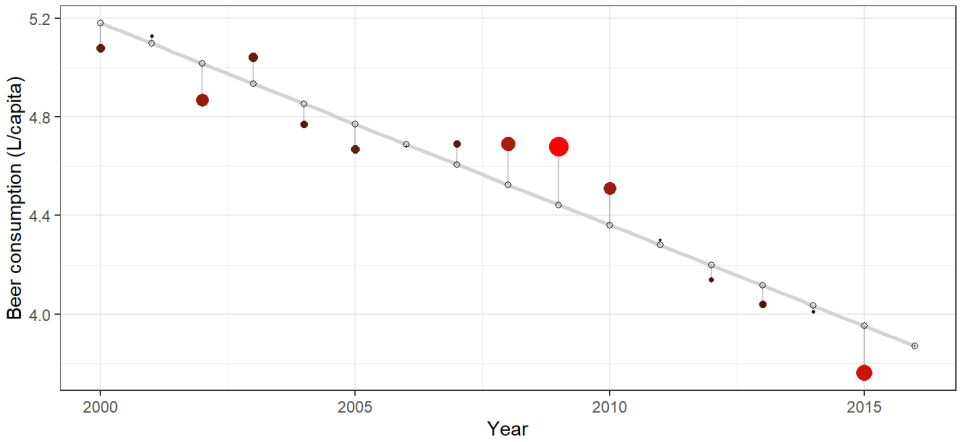
With the person next to you:

- What do the beer and rainfall data have in common?
- What is a hypothesis you could test for these datasets?

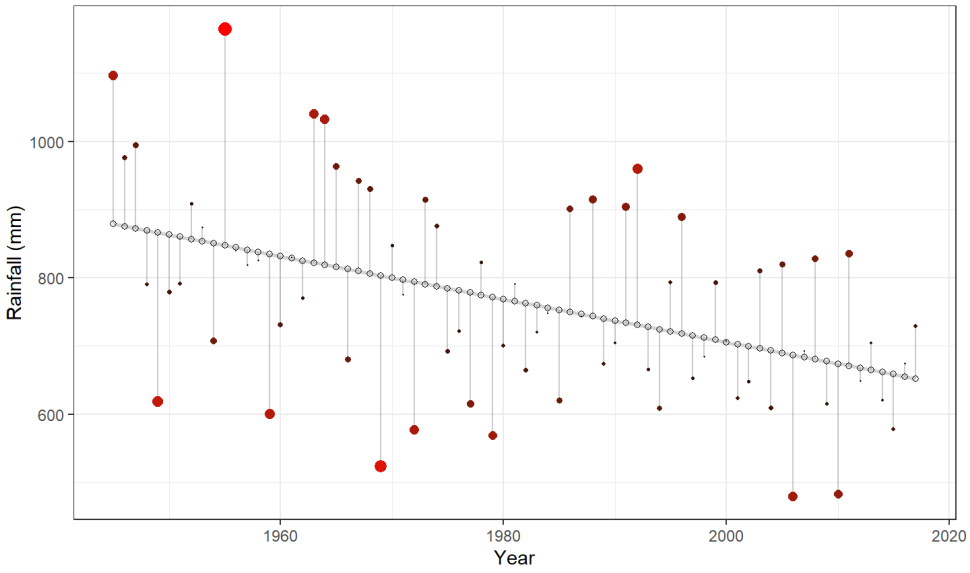
Is it a good line?

Fit a line to beer data

- We have already learnt how to fit a line to bivariate data.
- To find the “best” line, we summed the square of the residuals (“gaps”) between our line (model) and our observations (data), and looked for the line which gave the minimum sum.



Fit a line to rainfall data



Is it a good line?

- Just because we can find the “best” line, doesn’t mean it’s a “good” line? It could be the “best” of a whole lot of poor lines because the data doesn’t have a clear linear trend.
- The equation of the population line is: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

where:

- Y_i is the dependent variable;
- X_i is the independent variable;
- β_0 is the y-intercept;
- β_1 is the slope parameter;
- ϵ_i is the residual or “gap”.

β_0 and β_1 are known as the regression coefficients.

- From our sample, we can work out the sample line: $y = a + bx$
- For Beer data, we get

$$y = 168.95 - 0.08x$$

```
lm(Value ~ Time, PCBeer)
```

```
##
## Call:
## lm(formula = Value ~ Time, data = PCBeer)
##
## Coefficients:
## (Intercept)         Time
##   168.95490     -0.08189
```

1-Sample T Test to Test the Slope

Hypothesis Test for the slope

We want to test if the slope is significant. Then we could make statements like:

- “Perth rainfall has decreased significantly over the last 65 years”.
- “Beer consumption has decreased significantly over the last 15 years”.

Note: here we focus on the slope β_1 as the parameter of interest. Sometimes we also focus on the intercept β_0.

3 main steps for any hypothesis test

1. Set up research question

H: Hypothesis H_0 vs H_1

2. Weigh up evidence

A: Assumptions
T: Test Statistic
P: P-value

3. Explain conclusion

C: Conclusion

H: Hypotheses

Our hypotheses are:

- $H_0 : \beta_1 = 0$ [Or: There is not a linear trend.]
- $H_1 : \beta_1 \neq 0$ [Or: There is a linear trend.]

We can use the 1-Sample T Test to test this hypothesis.

A: Assumptions

There are 2 assumptions that we need to check.

- The residuals should be independent, normal, with constant variance (homoscedasicity). Check: Residual Plot, QQ Plot, Shapiro-Wilk Test
- The relationship between the dependent and independent variable should look linear. Check: Scatter Plot

T: Test statistic

Formula

The test statistic is

$$T = \frac{OV - EV}{SE} = \frac{\hat{\beta}_1 - 0}{SE_{\beta_1}} = \frac{b}{SE_{\beta_1}}$$

with $n - 2$ degrees of freedom.

Note: In general, we have $n - p - 1$ degrees of freedom where n is the sample size and p is the number of independent variables, but we are just dealing with $p = 1$ (simple linear regression).

In R

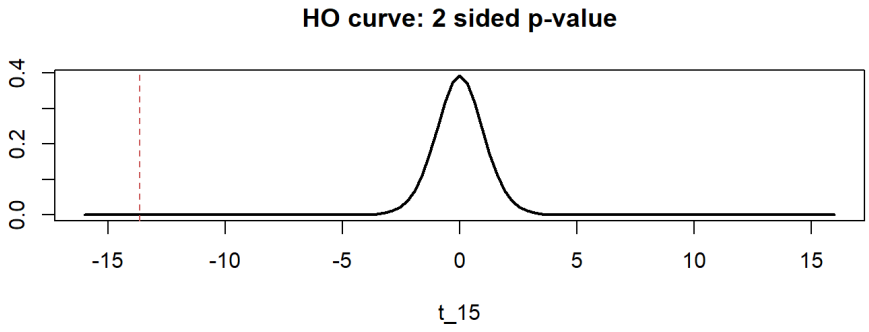
Use `summary()` output, gives $b = \hat{\beta}_1 = -0.082$ (3 dp) and $SE_{\beta_1} = -0.006$ (3 dp), so $t = -0.082/0.006 \approx -13.66$ as given.

```
summary(lm(Value ~ Time, PCBeer))
```

```
##
## Call:
## lm(formula = Value ~ Time, data = PCBeer)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.192083 -0.082843 -0.009069  0.082819  0.236593
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 168.954902  12.032998   14.04 4.92e-10 ***
## Time         -0.081887   0.005993  -13.66 7.18e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.121 on 15 degrees of freedom
## Multiple R-squared:  0.9256, Adjusted R-squared:  0.9207
## F-statistic: 186.7 on 1 and 15 DF,  p-value: 7.183e-10
```

P&C: P-value and Conclusion

- From `summary()`, p-value=7.18×10^{-10}.



- Therefore, we reject the null hypothesis and conclude that there is strong evidence to suggest that the slope is significant. (“As the p-value is low, the null hypothesis must go!”)
- Hence, there is strong evidence to suggest that beer consumption has been changing in Australia since the year 2000. Consumption is *decreasing* from the scatter plot.

Summary: 1 Sample T-Test for Regression (Slope)

Statistical Thinking

H: $H_0 : \beta_1 = 0$ [Or: There is not a linear trend] vs $H_1 : \beta_1 \neq 0$ [Or: There is a linear trend.]

A: The residuals should be independent, normal, with constant variance (homoscedasicity), and the relationship between the dependent and independent variable should look linear.

T: $T = \frac{b}{SE_{\beta_1}}$

P: Use t_{n-2} curve to find tail areas.

C: Retain or Reject H_0.

Hypothesis Test for Rainfall

```
##
## Call:
## lm(formula = Annual ~ Year, data = Rain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -279.55  -79.65  -12.36  100.58  317.30
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6999.4260  1414.3117   4.949 4.84e-06 ***
## Year          -3.1468     0.7139  -4.408 3.63e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 128.5 on 71 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.2149, Adjusted R-squared:  0.2038
## F-statistic: 19.43 on 1 and 71 DF,  p-value: 3.628e-05
```

From the output, $\hat{y} = 6999.4 - 3.1x$. For $H_0 : \beta_1 = 0$, the p-value=3.6×10^{-5}. Hence, there is strong evidence to suggest that annual rainfall has been changing (*decreasing*) in Perth since the 1940's.

Summary

Using the 1 Sample T Test we can test whether the slope is significant in a simple linear regression.

Further Thinking

We can extend the concepts we have learned to multiple linear regression.