Minilab 8b Worksheet

Statistical Testing: Significance in Correlations and Linear Models

We have seen that linear models can involve any number of quantitative and categorical predictor variables. In this minilab, we consider which of these predictor variables can be considered *statistically significant*, i.e., whether a variable should be retained in the linear model or can be safely dropped from the linear model.

1. Hypothesis Tests for Correlations

The dataset “state.x77” includes several variables (columns) so is a multivariate dataset.

(1) Taking an exploratory data analysis style look at the data, we might look at summary statistics, scatterplots and correlation coefficients.

```
library(tidyverse)
state = as_tibble(state.x77)
summary(state)
library(GGally)
ggpairs(state)
```

(2) Once you have calculated a correlation coefficient (using the sample data), how do you check whether or not there really is enough evidence in the sample data for a correlation in the population?

The typical null hypothesis is no relationship between two particular variables, i.e., the correlation ρ (Greek letter “rho”) between those two particular variables in the *population* is zero, written as $H_0\colon \rho = 0$.

We can then use the `cor.test()` function to carry out a hypothesis test on two particular variables. For example, we can test the correlation between Population and Income.

```
cor.test(state$Population, state$Income)
```

The output of `cor.test()` is as follows.

```
        Pearson's product-moment correlation

data:  state$Population and state$Income
t = 1.475, df = 48, p-value = 0.1467
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.07443435  0.45991855
sample estimates:
      cor
0.2082276
```

This gives $r = 0.2082276$ (or you can see 0.208 in the ggpairs plot). Given $\alpha = 0.05$ and looking at the p-value (highlighted in red), we see that 0.1467>0.05 and so we **fail to reject** the null hypothesis $H_0 : \rho = 0$, i.e., there is **no evidence** from the data for a correlation between Population and Income in the population.

Notice that the confidence interval for ρ includes 0 and hence 0 is a likely (plausible) value for ρ.

(3) Is it possible to test the correlations between all pairs of variables all at once? The R package “psych” has a function that does exactly this. Note this is corr.test() with TWO r’s instead of cor.test() with ONE r.

```
#install.packages("psych")
library(psych)
corr.test(state)
```

To interpret the output, only look at the p-values BELOW-and-LEFT of the diagonal (top-left to bottom-right) entries and pick out the combinations which give p-value ≤ 0.05. We say that only these are **statistically significant**.

Remember that correlation does not imply causation. For example, Population and Frost have a statistically significant correlation, but neither causes the other.

2. Statistical Significance in Linear Models

When fitting linear models to a multivariate dataset, we have just one quantitative *response* variable and several available quantitative and categorical *predictor* variables.

Which of the predictor variables can meaningfully be included in the linear model, i.e., which of the predictor variables are statistically significant in predicting the response variable?

(1) Consider the following age (months) and height (cm) data of children. This data comes from https://www.datacamp.com/community/tutorials/linear-regression-R but has been slightly modified.

```
children = tribble(
  ~age,~height,~siblings,
  18, 76.1, 1,
  19, 77.0, 2,
  20, 78.1, 3,
  21, 78.2, 2,
  22, 78.8, 0,
  23, 79.7, 1,
  24, 79.9, 5,
  25, 81.1, 0,
  26, 81.2, 1,
  27, 81.8, 4,
  28, 82.8, 1,
  29, 83.5, 5 )
```

(2) We can easily plot a scatterplot of height vs age and fit a linear model “height~age”.

```
ggplot(children, aes(x=age,y=height)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE)
model = lm(height~age, data=children)
summary(model)
```

(3) What if we also use the number of siblings as a predictor?

```
model = lm(height~age+siblings, data=children)
summary(model)
```

The output is as follows.

```
> summary(model)

Call:
lm(formula = height ~ age + siblings, data = children)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.26297 -0.22462 -0.02021  0.16102  0.49752 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 64.90554    0.53526 121.260 8.96e-16 ***
age          0.63751    0.02340  27.249 5.85e-10 ***
siblings    -0.01772    0.04735  -0.374    0.717    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2677 on 9 degrees of freedom
Multiple R-squared:  0.9889,	Adjusted R-squared:  0.9865 
F-statistic: 402.2 on 2 and 9 DF,  p-value: 1.576e-09
```

Look at the Coefficients section of the output, and in particular at the column Pr(>|t|), which gives a p-value for the Intercept and each of the predictors (age and siblings). We can conclude that:

- Intercept is significant (p-value is 8.96e-16, i.e., 0.000000000000000896 which is clearly less than 0.05). Notice the three stars (one star would be enough to show the p-value $\leq$ 0.05 according the "signif. codes").
- Age is significant (p-value is 0.000000000585). Notice the three stars.
- Siblings is **not significant** (we see that 0.717>0.05). Notice there are no stars. *We can and should remove Siblings from the model.*

Of course, we might think it very silly to include the number of siblings of a child in attempting to predict their height, but perhaps there is a plausible explanation.

When looking at a multivariate dataset, we need to use statistical tools to determine which predictors are significant.

Exercise. The dataset in the R code below concerns the number of species of tortoise on the various Galapagos Islands. There are 30 cases (Islands) and 7 variables in the dataset. See https://rdrr.io/cran/faraway/man/gala.html for a description of the variables involved.

```
# install.packages("faraway")    # run only once
library(tidyverse)
library(faraway)
gala = as_tibble(gala)
library(GGally)
ggpairs(gala)
model = lm(Species~Area+Elevation+Nearest+Scruz+Adjacent,data=gala)
summary(model)
```

(a) From the scatter matrix, which variables have significant correlation coefficient with Species? Does corr.test() provide the same information?

(b) From fitting the linear model, which predictor variables are NOT significant predictors of Species? What linear model would you suggest instead?

Summary

In this minilab, we have looked at hypothesis tests for population correlation coefficient ρ, and linear model intercept α and slopes β. It is possible to calculate everything we need ourselves, but actually the R functions `cor.test()` and `lm()` do all the calculations required and we can then make conclusions based on the p-value or the confidence interval.

Overall, even though it is easy to fit a linear model using lm(), you must always check that all predictor variables are significant and that the diagnostic plots are ok.