R for Business 3

1201 Data Science: Tim Raiswell

2018/12/08

Setting Up in R

Getting Started

1. Open R Studio. In the console type: `install.packages("pacman")`.
2. Now go to `File > New File > R Markdown > Document > PDF`. Name the file and hit 'Okay'.
3. In your new R Markdown file delete eveything after the `##R Markdown` heading.
4. Now change the heading to `##Loading Libraries`.
5. Underneath your heading create a new R code chunk by hitting CTRL + ALT + "I". Type the following and hit CTRL + Enter to run it in the chunk.

```
pacman::p_load(ggthemes, psych, gridExtra, ggcorrplot)
```

This code checks for the presence of the named packages on your machine. If thyey are present, it loads them. If they are not installed, `pacman` installs and loads them.

Getting to Know Markdown

Learn some basic markdown here: https://www.markdowntutorial.com. It takes about 5 minutes to learn and master the basics and about 30 mins to master the entire format.

Markdown is a stylized text formatting method to help scientists, academics and students combine technical and non-technical expression with visualization tools. It's basically like fast HTML.

Tip: Markdown text is written outside of code chunks. Text written inside code chunks needs to be code, unless it is preceded by a hashtag.

Outside a code chunk, type:

```
# Markdown
## Markdown
### Markdown
#### Markdown
```

And hit the `knit` button.

A Little More Markdown

Type and knit afterwards:

```
Strikethrough: ~~The cat sat on the mat.~~
Italics: *The cat sat on the mat.*
Bold: **The cat sat on the mat.**
```

Now for a list:

```
* Item one
* Item two
[tab] + Item two (a)
```

Indented quote:

```
> blah blah blah
```

Cheatsheet here: https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf

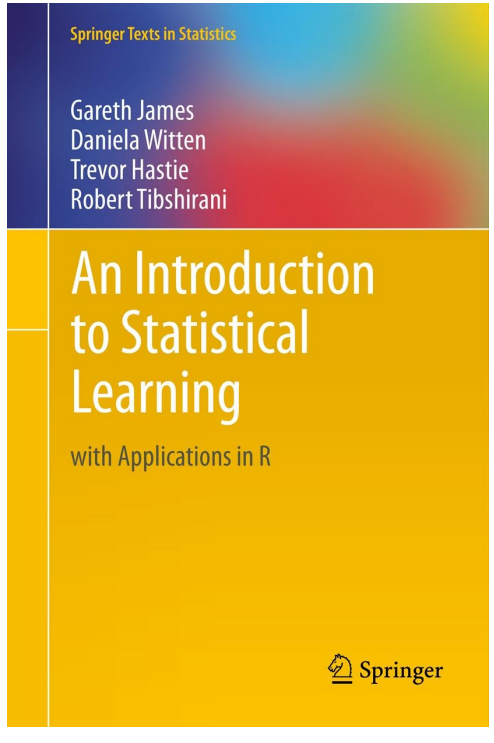
C is For Correlation

Some Learning Resources

1. https://newonlinecourses.science.psu.edu/statprogram/ The undergraduate and graduate stats courses and materials are excellent. I frequently find myself relying on it because it's written so clearly.

2. I love this book and you will too: https://www-bcf.usc.edu/~gareth/ISL/ (FREE!)

But buy the hardcopy so you can look bad-ass in Starbucks.



When Two Things Are Correlated (in Machine Learning)

1. They share a quantifiable mathematical relationship.
2. They do not necessarily belong to the same causal set.

An Unpopular View on Correlation and Causation

1. Correlation $\neq$ Causation is the single most cited statistical principle (by anyone who suspects you may be up to dark magic with your analysis).

2. In some branches of statistics, the principle matters a lot. See medical experimentation.

3. In other branches, it's just a distraction from getting the job done. See digital marketing.

Tip: Establish whether real or intabgible harm can be caused by an incorrect causal interpretation of model results. If there is no harm or cost, there is no need to spend time convincing yourself or others that causation matters.

This is not an invitation to abandon ethical data science. Conduct ethical analysis above all else. Rather, it is a recognition that the quesiton of causality can be immaterial to the outcome of many data science projects.

For some interesting and very current Bayesian theory of causal modeling outside experimentation, see: https://bit.ly/2REDkVS

A Note on Confounding

A Note on Confounding

Why do people who use the most sunscreen experience more instances of skin cancer?

A Note on Confounding

Why do people who use the most sunscreen experience more instances of skin cancer?

> Confounding occurs when the experimental controls do not allow the experimenter to reasonably eliminate plausible alternative explanations for an observed relationship between independent and dependent variables. Source: **stattrek.com**

Loading the Data

Type the name of the dataset. It has 10 x observations of 8 variables.

```
library(tidyverse) # load the swiss-army knife of data management and analysis packages
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ tibble  1.4.2     ✔ purrr   0.2.5
## ✔ tidyr   0.8.2     ✔ dplyr   0.7.8
## ✔ readr   1.1.1     ✔ stringr 1.3.1
## ✔ tibble  1.4.2     ✔ forcats 0.3.0
```

Describe the Data

What do you notice?

```
library(psych) # load the library with the describe function
describe(anscombe) %>% head(6) # return a compact output of the describe function.
```

```
##    vars  n mean   sd median trimmed  mad  min   max range  skew kurtosis
## x1    1 11  9.0 3.32   9.00    9.00 4.45 4.00 14.00 10.00  0.00    -1.53
## x2    2 11  9.0 3.32   9.00    9.00 4.45 4.00 14.00 10.00  0.00    -1.53
## x3    3 11  9.0 3.32   9.00    9.00 4.45 4.00 14.00 10.00  0.00    -1.53
## x4    4 11  9.0 3.32   8.00    8.00 0.00 8.00 19.00 11.00  2.47     4.52
## y1    5 11  7.5 2.03   7.58    7.49 1.82 4.26 10.84  6.58 -0.05    -1.20
## y2    6 11  7.5 2.03   8.14    7.79 1.47 3.10  9.26  6.16 -0.98    -0.51
##      se
## x1 1.00
## x2 1.00
## x3 1.00
## x4 1.00
## y1 0.61
## y2 0.61
```

Let's Build Some Linear Regression Models

```
lm(y1 ~ x1, data = anscombe) %>% summary() # creates a linear model and summary statistics
```

```
##
## Call:
## lm(formula = y1 ~ x1, data = anscombe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9212 -0.4558 -0.0414  0.7094  1.8388
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0001     1.1247   2.667  0.02573 *
## x1            0.5001     0.1179   4.241  0.00217 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.237 on 9 degrees of freedom
## Multiple R-squared:  0.6665,    Adjusted R-squared:  0.6295
## F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

What is R and Why do We Square it?

R is the correlation coefficient.

The correlation coefficient measures the strength and direction of a **LINEAR** relationship between two variables. The value is always between +1 and –1.

R^2 is the coefficient of determination.

When we square R, we get a value that explains the **proportion of the variance in the dependent variable that is predictable from the independent variable(s).**

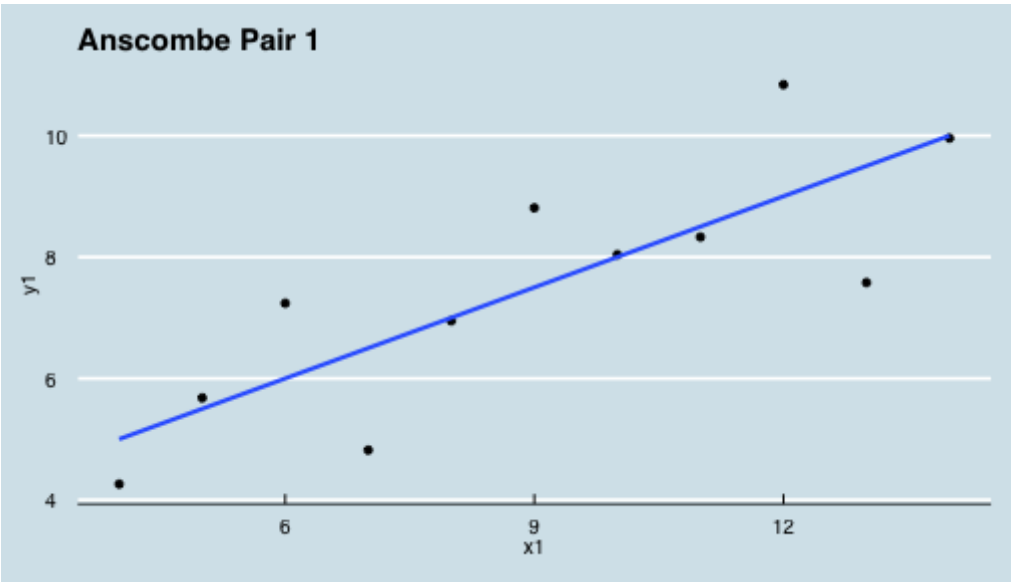
That is why the R^2 statistic is sometimes referred to as "goodness of fit".

Let's Look at the Data Again

```
library(ggthemes) # load some cool themes
(pair_one <- ggplot(anscombe, aes(x1, y1)) + # plot anscombe data with x and y variables specifie
  geom_point() + # scatter plot
  geom_smooth(method = "lm", se = FALSE) + # line of best fit using a linear model method
  theme_economist() + #economist theme
  labs(title = "Anscombe Pair 1")) # add the title
```

Let's Look at the Data Again

```
library(ggthemes) # load some cool themes
(pair_one <- ggplot(anscombe, aes(x1, y1)) + # plot anscombe data with x and y variables specifie
  geom_point() + # scatter plot
  geom_smooth(method = "lm", se = FALSE) + # line of best fit using a linear model method
  theme_economist() + #economist theme
  labs(title = "Anscombe Pair 1")) # add the title
```

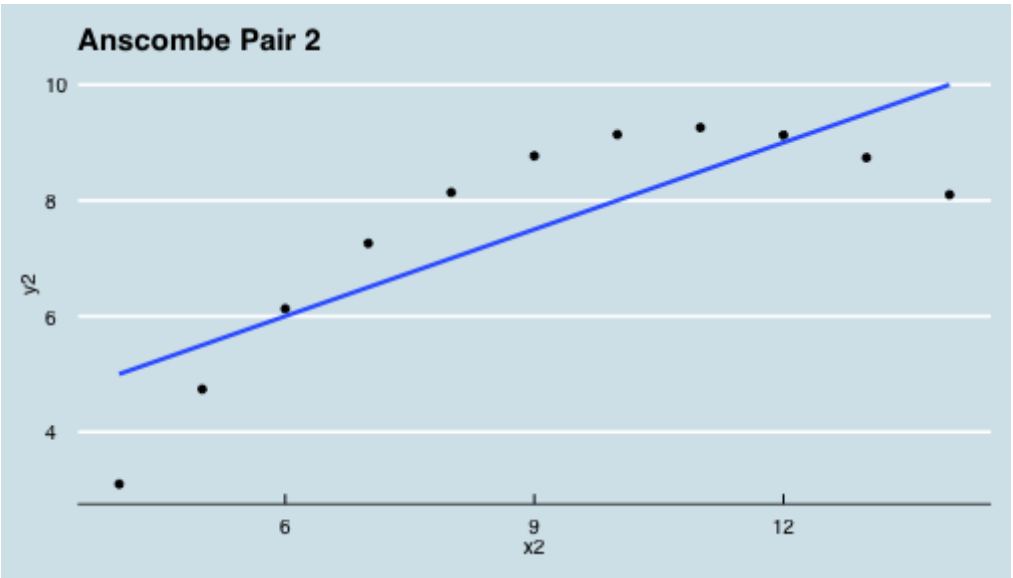


Pair 2

```
(pair_two <- ggplot(anscombe, aes(x2, y2)) + # plot anscombe data with x and y variables specifie
  geom_point() + # scatter plot
  geom_smooth(method = "lm", se = FALSE) + # line of best fit using a linear model method
  theme_economist() + #economist theme
  labs(title = "Anscombe Pair 2")) # add the title
```

Pair 2

```
(pair_two <- ggplot(anscombe, aes(x2, y2)) + # plot anscombe data with x and y variables specifie
  geom_point() + # scatter plot
  geom_smooth(method = "lm", se = FALSE) + # line of best fit using a linear model method
  theme_economist() + #economist theme
  labs(title = "Anscombe Pair 2")) # add the title
```

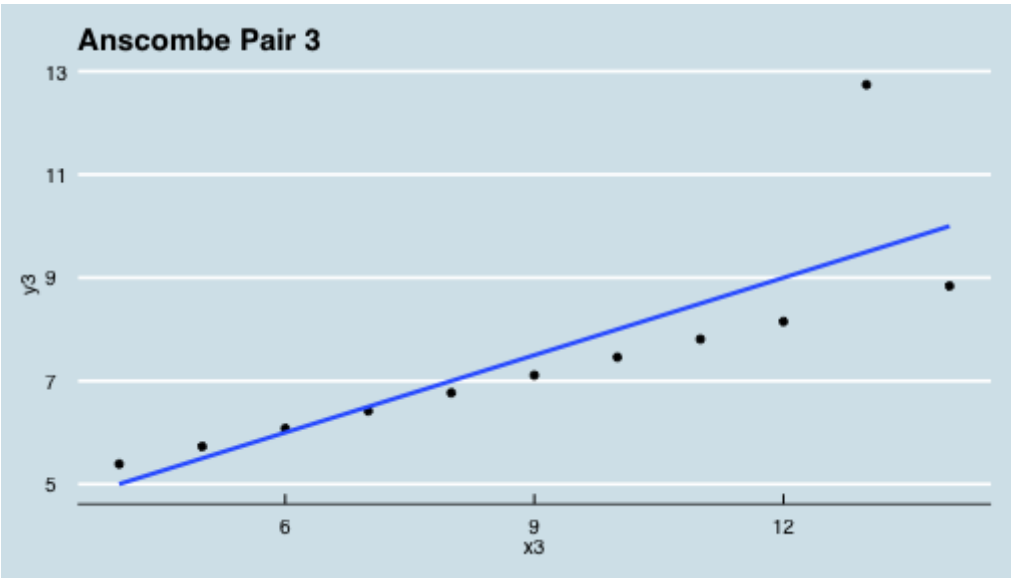


Pair 3

```
(pair_three <- ggplot(anscombe, aes(x3, y3)) + # plot anscombe data with x and y variables speci
  geom_point() + # scatter plot
  geom_smooth(method = "lm", se = FALSE) + # line of best fit using a linear model method
  theme_economist() + #WSJ theme
  labs(title = "Anscombe Pair 3")) # add the title
```

Pair 3

```
(pair_three <- ggplot(anscombe, aes(x3, y3)) + # plot anscombe data with x and y variables speci
  geom_point() + # scatter plot
  geom_smooth(method = "lm", se = FALSE) + # line of best fit using a linear model method
  theme_economist() + #WSJ theme
  labs(title = "Anscombe Pair 3")) # add the title
```

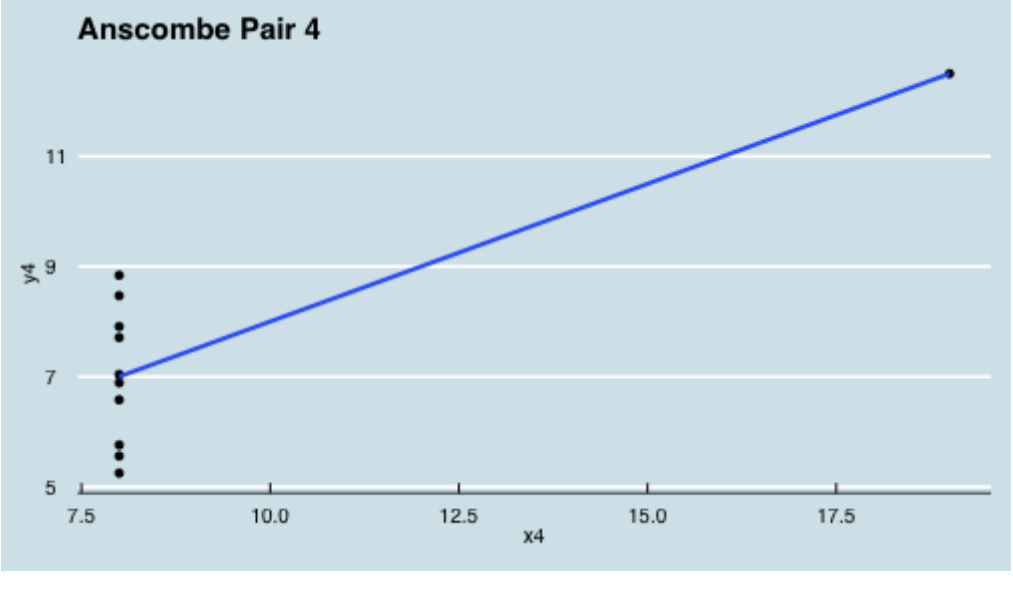


Pair 4

```
(pair_four <- ggplot(anscombe, aes(x4, y4)) + # plot anscombe data with x and y variables specif
  geom_point() + # scatter plot
  geom_smooth(method = "lm", se = FALSE) + # line of best fit using a linear model method
  theme_economist() +
  labs(title = "Anscombe Pair 4")) # add the title
```

Pair 4

```
(pair_four <- ggplot(anscombe, aes(x4, y4)) + # plot anscombe data with x and y variables specif
  geom_point() + # scatter plot
  geom_smooth(method = "lm", se = FALSE) + # line of best fit using a linear model method
  theme_economist() +
  labs(title = "Anscombe Pair 4")) # add the title
```

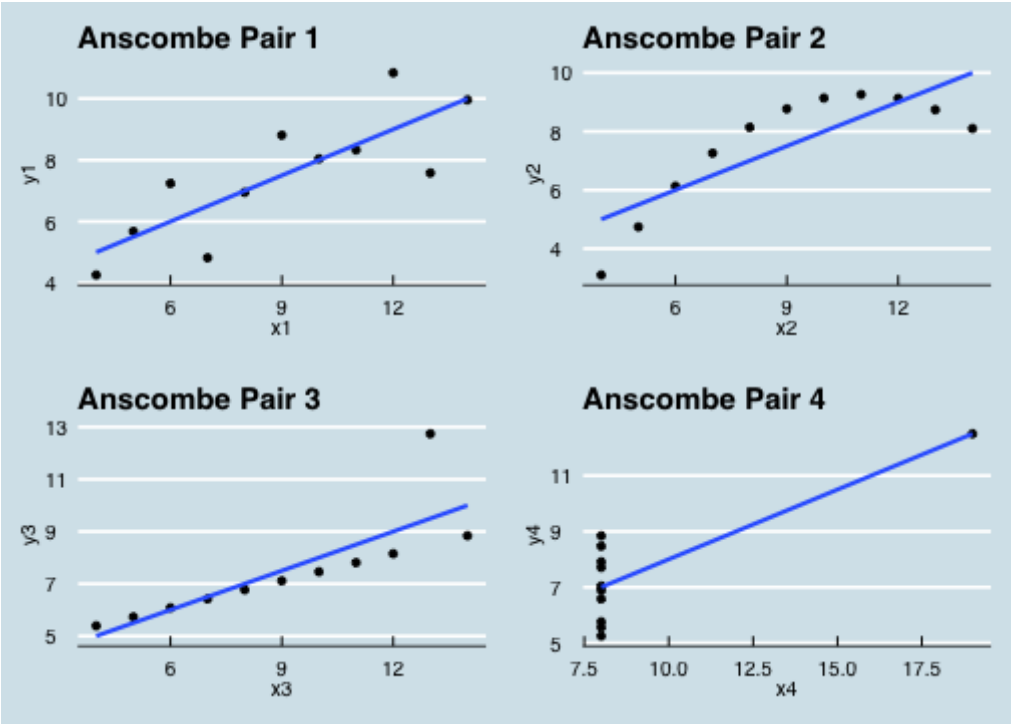


What Do We Notice?

```
library(gridExtra)
grid.arrange(pair_one, pair_two, pair_three, pair_four, nrow=2)
```

What Do We Notice?

```
library(gridExtra)
grid.arrange(pair_one, pair_two, pair_three, pair_four, nrow=2)
```

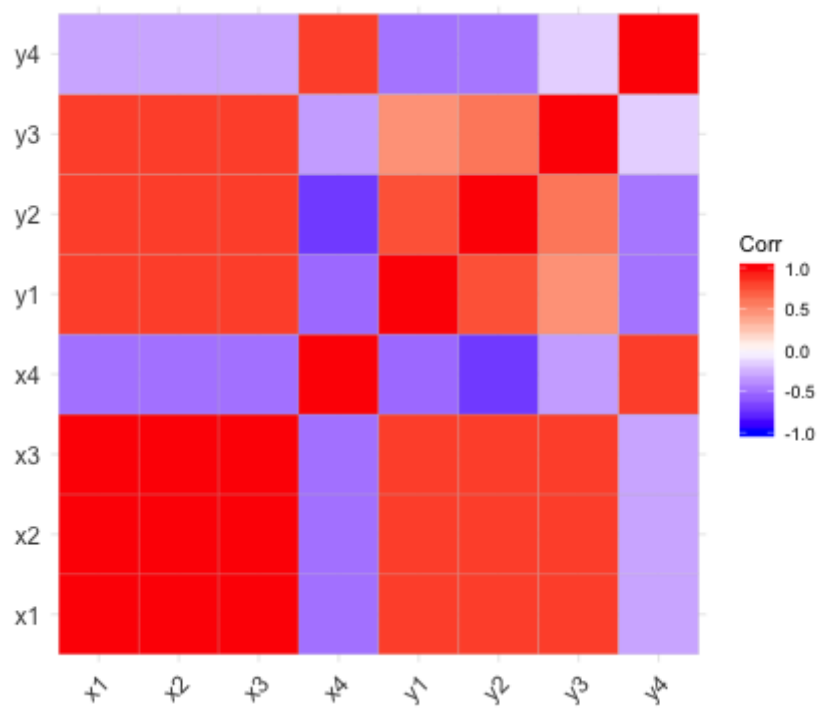


Correlation Plot 1

```
anscombe %>%
  cor() %>%
  ggcorrplot()
```

Correlation Plot 1

```
anscombe %>%
  cor() %>%
  ggcorrplot()
```



Correlation Plot 2

```
anscombe %>% pairs.panels(lm = TRUE)
```

Correlation Plot 2

```
anscombe %>% pairs.panels(lm = TRUE)
```