Correlation, Simple Linear Regression, and Prediction

September 22, 2017

Correlation

A correlation coefficient is a symmetric, scale-invariant measure of association between two random variables. It ranges from 1 to +1, where the extremes indicate perfect correlation and 0 means no correlation. The sign is negative when large values of one variable are associated with small values of the other and positive if both variables tend to be large or small simultaneously.

Pearson correlation

The empirical correlation coefficient is $r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i \bar{x})^2 (y_i - \bar{y})^2}}$

It can be shown that $|r|$ will be less than 1 unless there is a perfect linear relation between x_i and y_i, and for that reason the Pearson correlation is sometimes called the *linearcorrelation*. For mean, var, sd, and similar one-vector functions, you can give the argument $na.rm = T$ to indicate that missing values should be removed before the computation. For *cor*, you can write

```
> cor(blood.glucose,short.velocity)
Error in cor(blood.glucose, short.velocity) :
missing observations in cov/cor

> cor(blood.glucose,short.velocity,use="complete.obs")
[1] 0.4167546
```

You can obtain the entire matrix of correlations between all variables in a data frame by saying, for instance,

```
> cor(thuesen,use="complete.obs")
blood.glucose      short.velocity
blood.glucose   1.0000000           0.4167546
short.velocity  0.4167546           1.0000000
```

Simple plot

```
plot(blood.glucose~short.velocity, xlab = "Blood Glucose", ylab = "Short Velocity")
```

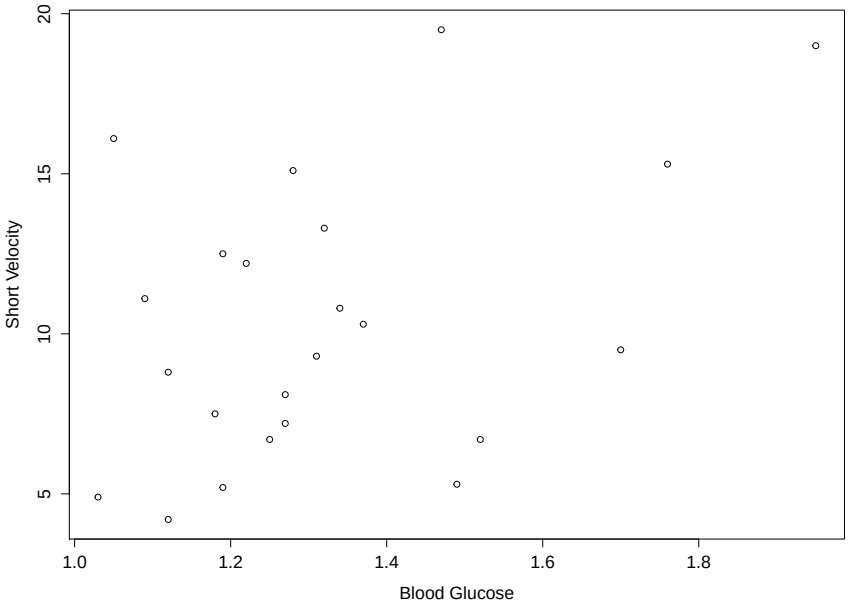

Figure 1: relationship between two variable

However, the calculations above give no indication of whether the correlation is significantly different from zero.

```
> cor.test(blood.glucose,short.velocity)
Pearsons product-moment correlation
data: blood.glucose and short.velocity
t = 2.101, df = 21, p-value = 0.0479
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.005496682 0.707429479
sample estimates:
cor
0.4167546
```

Spearman ρ

As with the one- and two-sample problems, you may be interested in nonparametric variants. These have the advantage of not depending on the normal distribution the test is considered one of several possibilities for testing correlations:

```
> cor.test(blood.glucose,short.velocity,method="spearman")
Spearmans rank correlation rho
data: blood.glucose and short.velocity
S = 1380.364, p-value = 0.1392
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
```

```
0.318002
Warning message:
In cor.test.default(blood.glucose, short.velocity, method="spearman"):
Cannot compute exact p-values with ties
```

Kendall τ

```
> cor.test(blood.glucose,short.velocity,method="kendall")
Kendalls rank correlation tau
data: blood.glucose and short.velocity
z = 1.5604, p-value = 0.1187
alternative hypothesis: true tau is not equal to 0
sample estimates:
tau
0.2350616
Warning message:
In cor.test.default(blood.glucose, short.velocity, method="kendall"):
Cannot compute exact p-value with ties
```

Notice that neither of the two nonparametric correlations is significant at the 5% level, which the Pearson correlation is, albeit only borderline significant.

linear regression

Linear regression can be performed with the *lm* function in the native **stats** package. A robust regression can be performed with the *lmrob* function in the **robustbase** package.

Graphical Analysis

let's try to understand these variables graphically. Typically, for each of the independent variables (predictors), the following plots are drawn to visualize the following behavior:

- Scatter plot: Visualize the linear relationship between the predictor and response
- Box plot: To spot any outlier observations in the variable. Having outliers in your predictor can drastically affect the predictions as they can easily affect the direction/slope of the line of best fit.
- Density plot: To see the distribution of the predictor variable. Ideally, a close to normal distribution (a bell shaped curve), without being skewed to the left or right is preferred.

scatter plot

Scatter plots can help visualize any linear relationships between the dependent (response) variable and independent (predictor) variables. Ideally, if you are having multiple predictor variables, a scatter plot is drawn for each one of them against the response, along with the line of best as seen below.

```
scatter.smooth(x=blood.glucose, y=short.velocity, main="short.velocity ~ blood.glucose")  # sca
```

Figure 2: Scatter Plot

BoxPlot Check for outliers

Generally, any datapoint that lies outside the 1.5 * interquartile-range (1.5*IQR) is considered an outlier, where, IQR is calculated as the distance between the 25th percentile and 75th percentile values for that variable.

```
boxplot(short.velocity , blood.glucose)
```

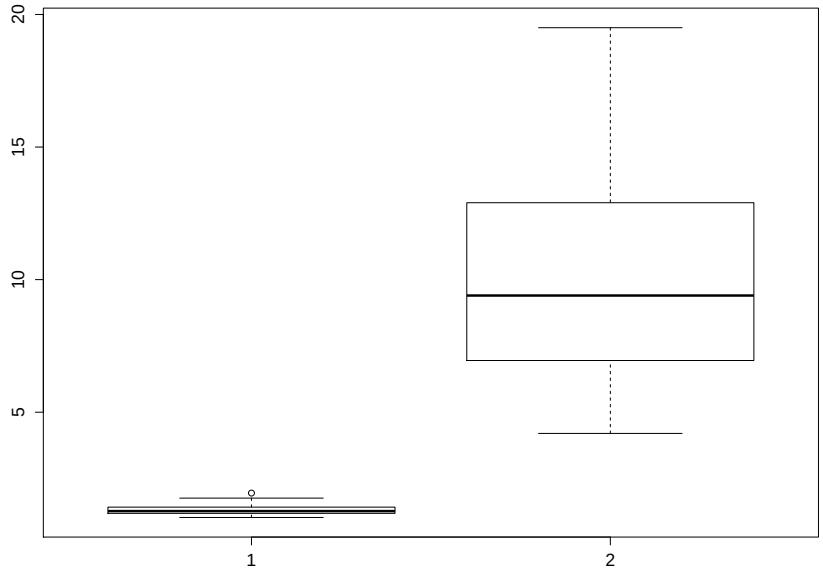

Figure 3: Boxplot

Build Linear Model

```
linearMod

Call:
lm(formula = short.velocity ~ blood.glucose)

Coefficients:
(Intercept)  blood.glucose
1.09781        0.02196
```

Linear Regression Diagnostics

```
summary(lm(short.velocity~blood.glucose))   #### extracting statistical testing hypothesis

Call:
lm(formula = short.velocity ~ blood.glucose)

Residuals:
Min       1Q   Median       3Q      Max
-0.40141 -0.14760 -0.02202  0.03001  0.43490

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.09781    0.11748   9.345 6.26e-09 ***
```

```
blood.glucose  0.02196    0.01045   2.101   0.0479 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.2167 on 21 degrees of freedom
(1 observation deleted due to missingness)
Multiple R-squared:  0.1737,Adjusted R-squared:  0.1343
F-statistic: 4.414 on 1 and 21 DF,  p-value: 0.0479
```

How to know which model is best fit

The most common metrics to look at while selecting the model are:

```
STATISTIC       CRITERION
R-Squared        Higher the better (> 0.70)
Adj R-Squared    Higher the better
F-Statistic       Higher the better
Std. Error       Closer to zero the better
t-statistic      Should be greater 1.96 for p-value to be less than 0.05
AIC              Lower the better
BIC               Lower the better
MSE (Mean squared error) Lower the better
```

AIC and BIC
The Akaikes information criterion - AIC (Akaike, 1974) and the Bayesian information criterion - BIC (Schwarz, 1978) are measures of the goodness of fit of an estimated statistical model and can also be used for model selection.

checking the assumption

Prediction

R Codes

```
?lillie.test  ##############lillie.test{nortest}  for the composite hypothesis of normality
######
################### package {IsWR}
library(ISwR)
attach(thuesen)
?thuesen
###############looking at the data
library(psych)
str(thuesen)
summary(thuesen)
#############################################################correlation
?cor
cor(blood.glucose,short.velocity)   ########### Wrong
```

```
cor(blood.glucose,short.velocity,use="complete.obs") ### taking care of missing values
cor(thuesen,use="complete.obs")  ##################### complete correlation Matrix
############################ploting the data
plot(blood.glucose~short.velocity, xlab = "Blood Glucose", ylab = "Short Velocity")

cor.test(blood.glucose,short.velocity) ############ to have statistical pack up to test H0: co
cor.test(blood.glucose,short.velocity,method="spearman") #### spearman test
cor.test(blood.glucose,short.velocity,method="kendall")  #### Kendall test
############################################################Simple Linear Regression
######################## Some Plots
scatter.smooth(x=blood.glucose, y=short.velocity, main="short.velocity ~ blood.glucose")  # sc
boxplot(short.velocity , blood.glucose)
############################fiting model
linearMod <- lm(short.velocity~blood.glucose)#### simple linear regression line build linear r
linearMod
summary(lm(short.velocity~blood.glucose))   #### extracting statistical testing hypothesis
summary(linearMod)
############################which model
AIC(linearMod)
BIC(linearMod)
```