Simulate continuous variables dataset

Contents

Purpose

This script shows how to simulate a dataset that can be used in regression problems.

In regression, the variables that we are measuring are continuous, and the variable that we are predicting is continuous as well.

Dependent variables

Simulate the data

Create two Normally distributed datasets that have a relationship.

Play with the number of samples and we move the means around.

```r
# From Harvard data science class (see references at the end of this notebook)
x <- rnorm(10000, mean=10, sd=sqrt(5))

# Initialize y with x...We would have a straight line if plotting y~x
y <- x

# Now inject variability to each, and we will not have a straight line exactly
x <- x + rnorm(10000, sd=2)
y <- y + rnorm(10000, sd=2)
```

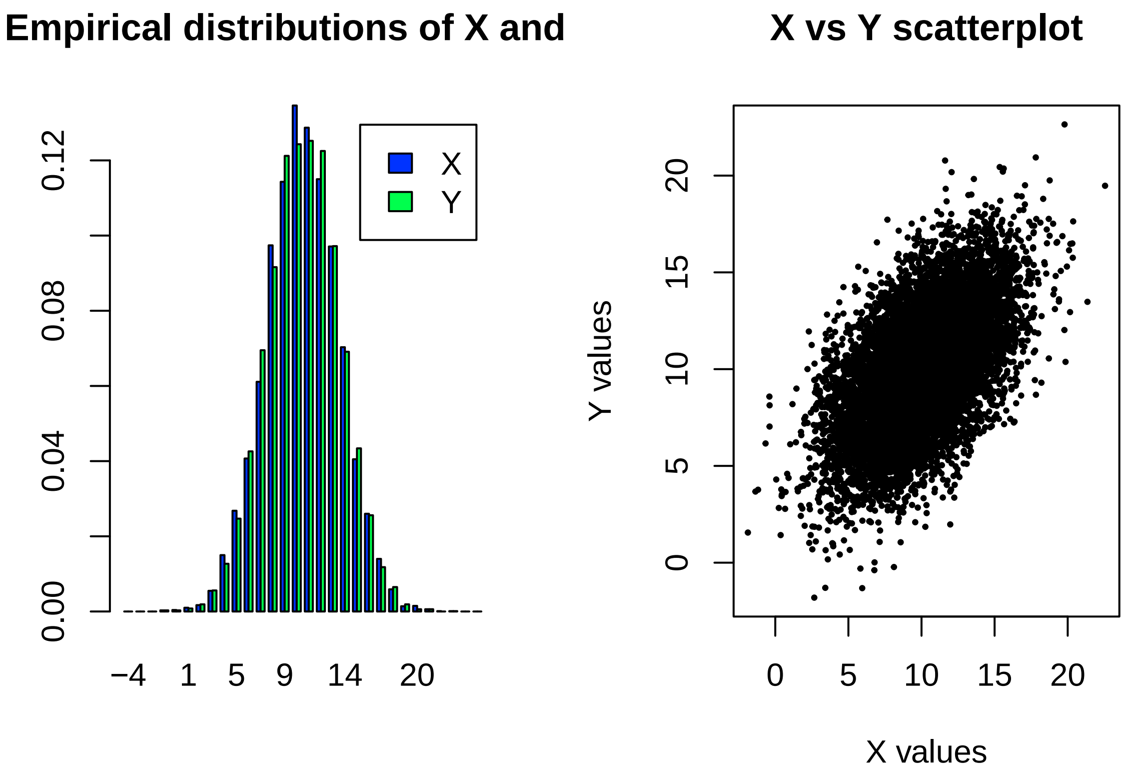
Plot histograms and scatterplot

```r
br<- -5:25 # set manually bins for histograms
# save histograms for X and Y , don't plot yet
hx <- hist(x, breaks=br, plot=F)
hy <- hist(y, breaks=br, plot=F)

# prepare 2 panels in one plot:
old.par <- par(mfrow=c(1,2))

# plot histograms side by side using rbind
barplot(rbind(hx$density,hy$density),
        beside=T,
        col=c(rgb(0,0.2,1), rgb(0,1,0.3)),
        legend=c('X','Y'),
        main='Empirical distributions of X and Y',
        names=br[-1])

# Scatter plot
plot(x,y,
     xlab='X values',
     ylab='Y values',
     main='X vs Y scatterplot',
     pch=19,
     cex=0.3)
```




```
# restore graphical attributes to previous values:
par(old.par)
```

Independent variables

Simulate the data

Create two independent Normally distributed datasets x and y.

Play with the number of samples and we move the means around.

```
# From Harvard data science class (see references at the end of this notebook)
# simulate sampling of 10000 values for X and for Y.
# We can play with the mean and sd. Should have same size to keep it balanced.
x <- rnorm(10000, mean=10, sd=3)
y <- rnorm(10000, mean=10, sd=3)
```

Plot histograms and scatterplot

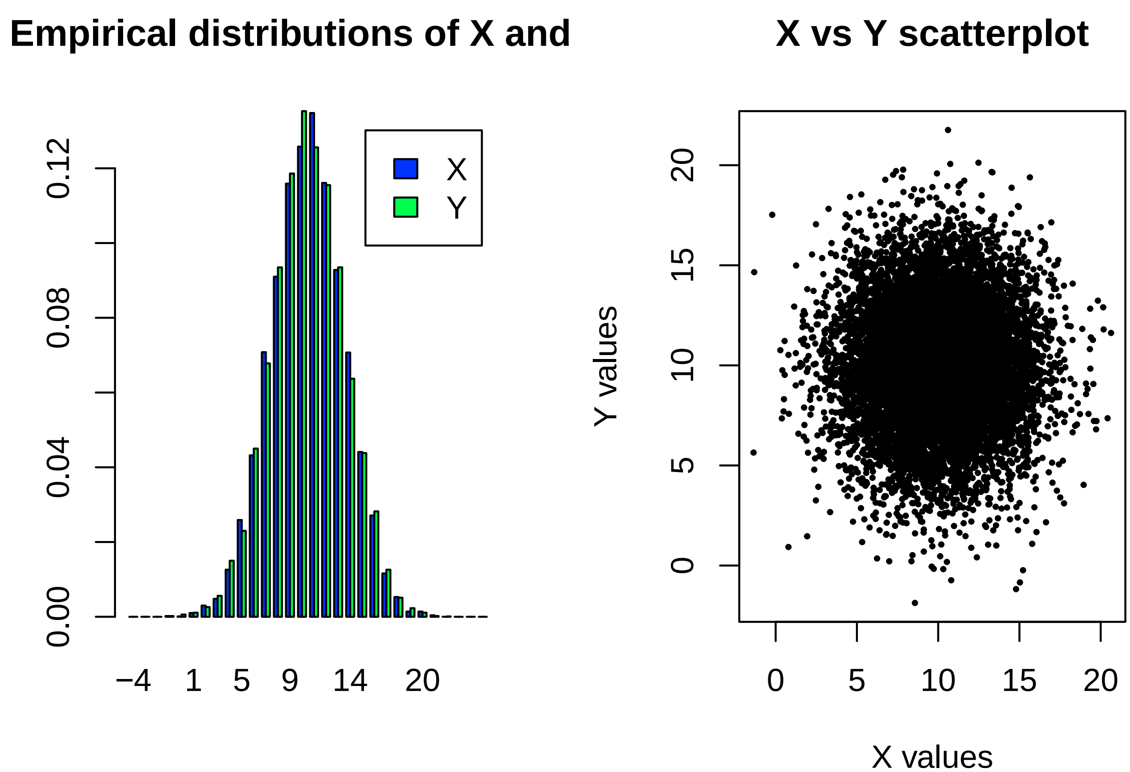
```
# Set manually bins for histograms
br<- -5:25
# Save histograms for X and Y , don't plot yet
```

```
hx <- hist(x, breaks=br, plot=F)
hy <- hist(y, breaks=br, plot=F)

# prepare 2 panels in one plot:
old.par <- par(mfrow=c(1,2))

# plot histograms side by side using rbind
barplot(rbind(hx$density,hy$density),
        beside=T,
        col=c(rgb(0,0.2,1), rgb(0,1,0.3)),
        legend=c('X','Y'),
        main='Empirical distributions of X and Y',
        names=br[-1])

# Scatter plot
plot(x,y,
     xlab='X values',
     ylab='Y values',
     main='X vs Y scatterplot',
     pch=19,
     cex=0.3)
```




```
# restore graphical attributes to previous values
par(old.par)
```

Simulate known linear regression

In this approach, we will simulate data where we know the linear regression parameters.

Simulate the data

Here we simulate X to be Uniformly distributed across a set of values.

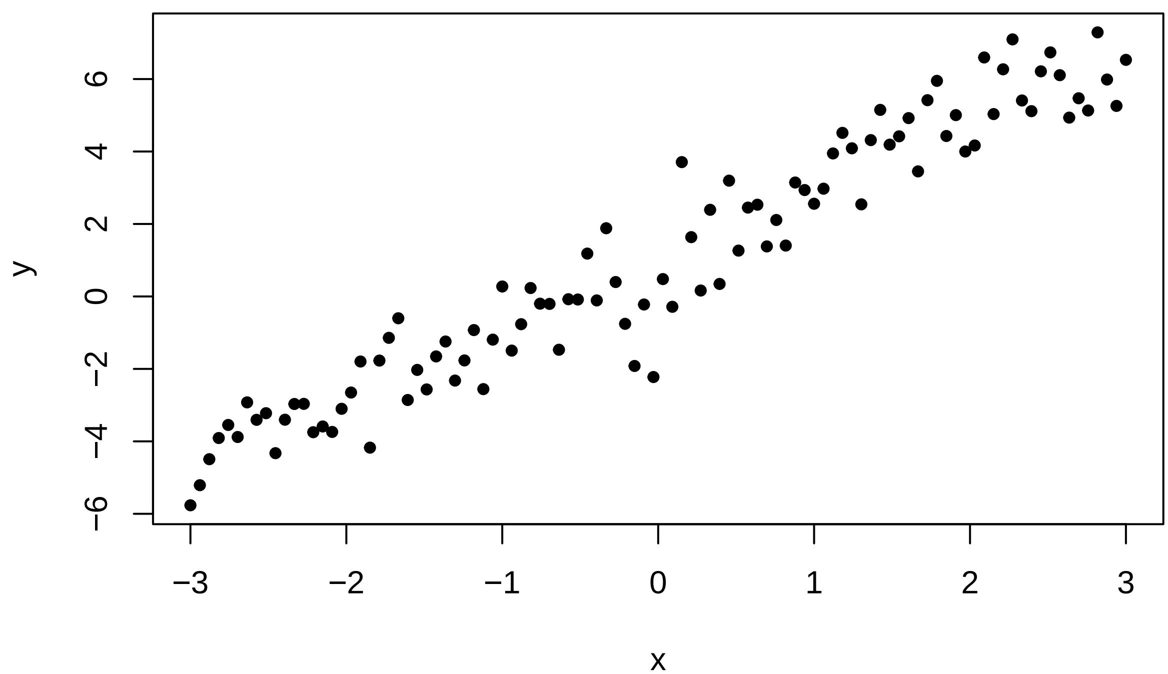
We simulate Y with a known intercept, pus a slope time X with a random variability.

```
# X has a uniform distribution over a sequence over a range
x <- seq(-3, 3, length=100)

# Y is based on X with a slope, an intercept and normal randomness
y <- 1 + 2*x+rnorm(100, sd=1)
```

Scatter plots

```
plot(x, y, pch=19, cex=0.7)
```



Apply linear regression model

```
m <- lm(y~x)
summary(m)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.15705 -0.69116  0.06019  0.69113  2.42504
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.99052    0.09591   10.33   <2e-16 ***
## x            1.93098    0.05482   35.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9591 on 98 degrees of freedom
## Multiple R-squared:  0.9268, Adjusted R-squared:  0.926
## F-statistic:  1241 on 1 and 98 DF,  p-value: < 2.2e-16
```

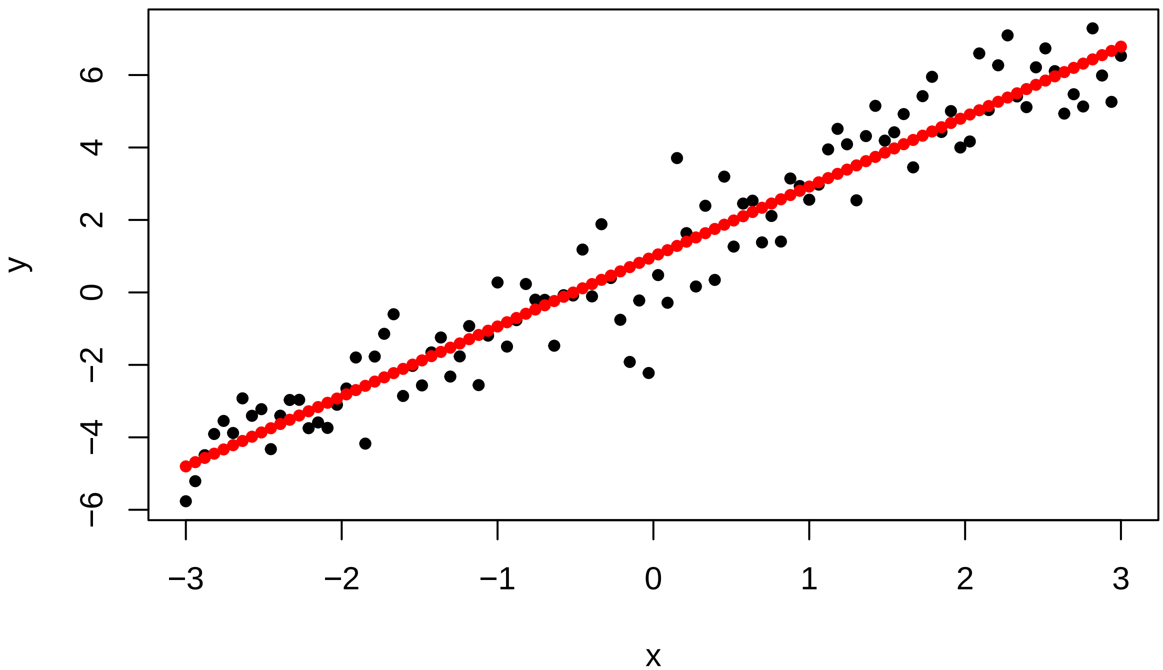
Notice the intercept is 1.01, while our empirical value was 1.

And the slope is 1.96, while our empirical value was 2.

Predict

```
yp <- predict(m)

# Plot the prediction. Replot and add prediction points:
plot(x, y, pch=19, cex=0.7)
points(x,yp,col="red",pch=19,cex=0.7)
```



Prediction metrix

```
rss <- sum((y - yp)^2)                  # Sum of Squares Estimated (aka SSE)
rse <- sqrt(rss/98)
tss <- sum((y - mean(y))^2)             # Sum of Squares Total (aka SST)
ssr <- sum((yp - mean(y))^2)          # Sum of Squares Regression
se <- rse/sqrt(sum((x-mean(x))^2))    # Standard Error
roh_squared <- ssr / tss

cat(' RSS aka SSE = ', rss)
```

```
##  RSS aka SSE =  90.1385
```

```
cat('\n RSE aka ?? = ', rse)
```

```
##
##  RSE aka ?? =  0.9590519
```

```
cat('\n TSS aka SST = ', tss)
```

```
##
##  TSS aka SST =  1231.342
```

```r
cat('\n SSR = ', ssr)
```

```
##
##  SSR =  1141.204
```

```r
cat('\n Coefficient of determination (roh squared) = ', roh_squared)
```

```
##
##  Coefficient of determination (roh squared) =  0.9267965
```

```r
cat('\n roh = ', sqrt(roh_squared))
```

```
##
##  roh =  0.9627027
```

```r
#
# cat('\n SE = ', se)
# (tss-rss)/tss
```

References

- Harvard "Elements of Statistical Learning" (2021) taught by professors Dr. Sivachenko, Dr. Farutin