Linear regression

McMaster University

Caitlin Simopoulos

2017-05-26

Let’s do some actual analysis!

Now that we’ve got the hang of some R commands, let’s do some real analysis

Linear regression

- explains the relationship between dependent (Y) and independent (X) variables
- “line of best fit” to data
- a “simple” linear regression is equivalent to a correlation
- is a *supervised* machine learning technique
- the model learns from known data, and can be used to predict
- sometimes we are looking for trends and don’t care about predictions

Linear regression

linear regression takes the linear equation of:

$$Y = mX + b$$

But we're going to use it more like:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i$$

where:
$\hat{y}_i =$ predicted response for expeimental unit i
$x_i =$ predictor (or independent variable) of experimental unit i
$\hat{\beta}_0 =$ is expected value when $x_i = 0$ (intercept)
$\hat{\beta}_1 =$ slope
$\epsilon =$ some error, because models are never perfect!

β_0 and β_1

β_0 and β_1 are unknown.
We will estimate these coefficients from known data. To do this, we need to estimate a line of best fit than minimizes error

An example: Does head size predict brain size?

```r
library(data.table)
# we've already loaded this package...
head_brain <- fread('http://www.stat.ufl.edu/~winner/data/brainhead.dat
head(head_brain)
```

```
##    V1 V2   V3   V4
## 1:  1  1 4512 1530
## 2:  1  1 3738 1297
## 3:  1  1 4261 1335
## 4:  1  1 3777 1282
## 5:  1  1 4177 1590
## 6:  1  1 3585 1300
```

Brain and head size data

Columns are:

1. gender (1 = male, 2 = female)
2. age (1 = 20-46, 2 = 46+)
3. head size (cm^3)
4. brain weight (g)

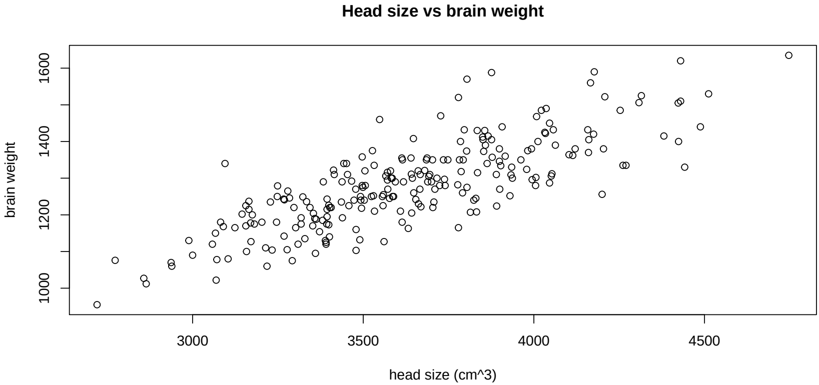
```r
# name columns to reflect what they are
colnames(head_brain) = c("gender", "age", "head", "brain")
```

Change data type

```
head_brain$gender <- as.factor(head_brain$gender)
head_brain$age <- as.factor(head_brain$age)
head_brain$head <- as.numeric(head_brain$head)
head_brain$brain <- as.numeric(head_brain$brain)
```

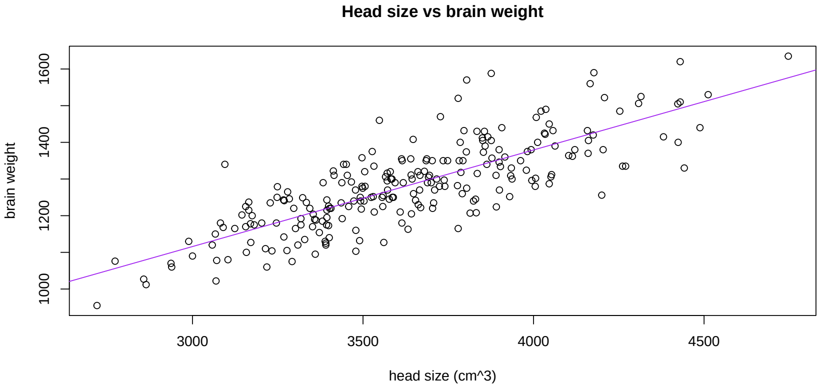
Plot data

```
#attach(head_brain)
plot(head_brain$head, head_brain$brain,
     main="Head size vs brain weight",
     xlab = "head size (cm^3)", ylab = "brain weight")
```



How to do a linear regression

```
set.seed(519)
lm1 = lm(brain ~ head, data=head_brain)
plot(head_brain$head, head_brain$brain,
     main="Head size vs brain weight",
     xlab = "head size (cm^3)", ylab = "brain weight")
abline(lm1, col="purple")
```



Summary of the linear regression

Here's a good explanation of the summary in MUCH more detail.

```
summary(lm1)
```

```
##
## Call:
## lm(formula = brain ~ head, data = head_brain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -175.98  -49.76   -1.76   46.60  242.34
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 325.57342   47.14085   6.906 4.61e-11 ***
## head          0.26343    0.01291  20.409  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 72.43 on 235 degrees of freedom
## Multiple R-squared:  0.6393, Adjusted R-squared:  0.6378
## F-statistic: 416.5 on 1 and 235 DF,  p-value: < 2.2e-16
```

Summary output gives a lot of information

1. Info about distrubution of residuals (errors)
2. The estimates of our coefficients
3. Standard error of coefficient estimates
    - square root of the variance
    - $\sqrt{\sigma}^2$
4. t-values (coefficient estimates/standard error)
5. $R^2 =$ (want closer to 1)
6. p-values (yuck!)
    - testing if your coefficient $= 0$

Quick residual check

What is a residual?

```
Residuals

Min      1Q  Median      3Q      Max  -175.98  -49.76
-1.76   46.60  242.34
```

- Median should be around 0
- 1Q and 3Q should be roughly of same absolute magnitube

Linear equation with coefficients

- $\beta_0 = \text{intercept} = 325.57$
- $\beta_1 = \text{head size coefficient} = 0.26$

$$\hat{y}_i = 325.57 + 0.26x_i + \epsilon_i$$

What does this mean in actual words?

Use confidence intervals instead of p-values

```r
confint(lm1, level=0.95) #95% confidence interval
```

```
##                   2.5 %      97.5 %
## (Intercept) 232.7007553 418.4460868
## head          0.2380003   0.2888584
```

We have more information than just head size!

We also have age range and sex!
We can use this information to potentially improve our regression by using a multiple linear regression model.

Here's a good tutorial.

Multiple linear regression

First, let's change out age and sex categories into 0s and 1s. (This is the convention in programming)

```
head_brain$gender = as.factor(ifelse(head_brain$gender == 1, 1, 0))
# if head_brain$gender[i] == 1
#     head_brain$gender[i] == 1
# else
#     head_brain$gender[i] == 0
head_brain$age = as.factor(ifelse(head_brain$age == 1, 1, 0))
```

Does age have an effect on brain size?

We have some options on possible formulas for the linear regression:

- brain ~ head + age
  - $\hat{brain}_i = \hat{\beta}_0 + \hat{\beta}_1 head_i + \hat{\beta}_2 age_i + \epsilon_i$
- brain ~ head:age
  - $\hat{brain}_i = \hat{\beta}_0 + \hat{\beta}_2 age_i head_i + \epsilon_i$
- brain ~ head*age = brain ~ head + age + head x age
  - $\hat{brain}_i = \hat{\beta}_0 + \hat{\beta}_1 head_i + \hat{\beta}_2 age_i + \hat{\beta}_2 age_i head_i + \epsilon_i$

```
lm2 = lm(brain ~ head + age, data=head_brain)
lm3 = lm(brain ~ head:age, data=head_brain)
lm4 = lm(brain ~ head * age, data=head_brain)
```

How can we interpret this?

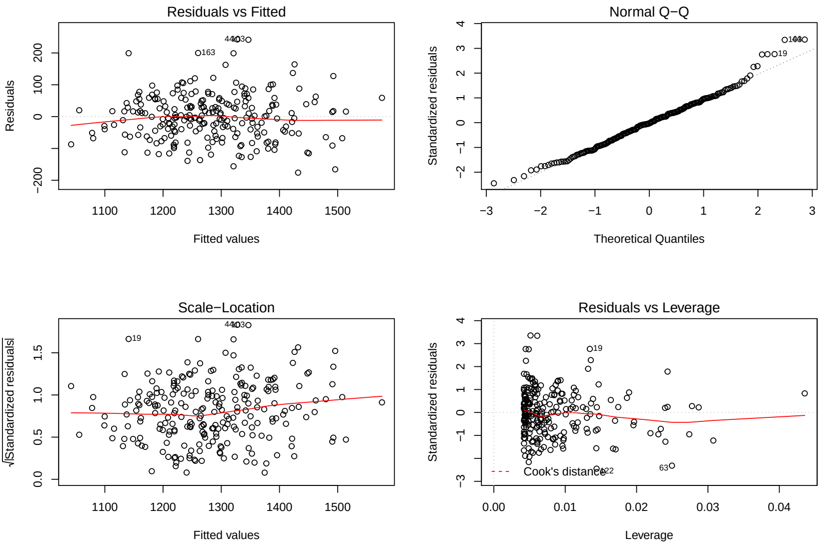
Interactions mean that slopes is different for each age category and each head size.

Evaluate

How can we evaluate these models?

Do these models meet the asumptions of linear models?

```
par(mfrow=c(2,2))
plot(lm1)
```



Diagnostic plot

- Residuals vs. Fitted
  - look for residuals randomly distributed around 0.
- Q-Q plot
  - used to look for normality of residuals
  - want the points to follow the diagonal line.
- Scale-Location plot
  - Similar to residuals vs fitted.
- Residuals vs. Leverage plot
  - Was our model driven by outliers?
    - Cook's distance lines would be visible (they aren't right now)
    - Data points driving the model would be labeled
    - If this happens, try removing the outliers.

PS. http://data.library.virginia.edu/diagnostic-plots/ is helpful for interpretting these plots.

Predictions

What if we are given new data, and want to make predictions from our models?

```r
new_data = data.frame(gender = as.factor(c(1,1,0,0)), age = as.factor(c
print(new_data)
```

```
##   gender age head
## 1      1   0 2265
## 2      1   1  999
## 3      0   0 2775
## 4      0   1 9275
```

Predictions

```
lm1_pred <- predict(lm1,newdata=new_data,
                    interval="confidence", level=.95)
lm2_pred <- predict(lm2,newdata=new_data,
                    interval="confidence", level=.95)
lm3_pred <- predict(lm3,newdata=new_data,
                    interval="confidence", level=.95)
lm4_pred<- predict(lm4,newdata=new_data,
                   interval="confidence", level=.95)
```

Predictions

To see our predictions...

How do they compare?

```
print(lm1_pred)
print(lm2_pred)
print(lm3_pred)
print(lm4_pred)
```

Can we test which model has the best predictions?

How can we do this?

Acknowledgements

This document was inspired by: this tutorial