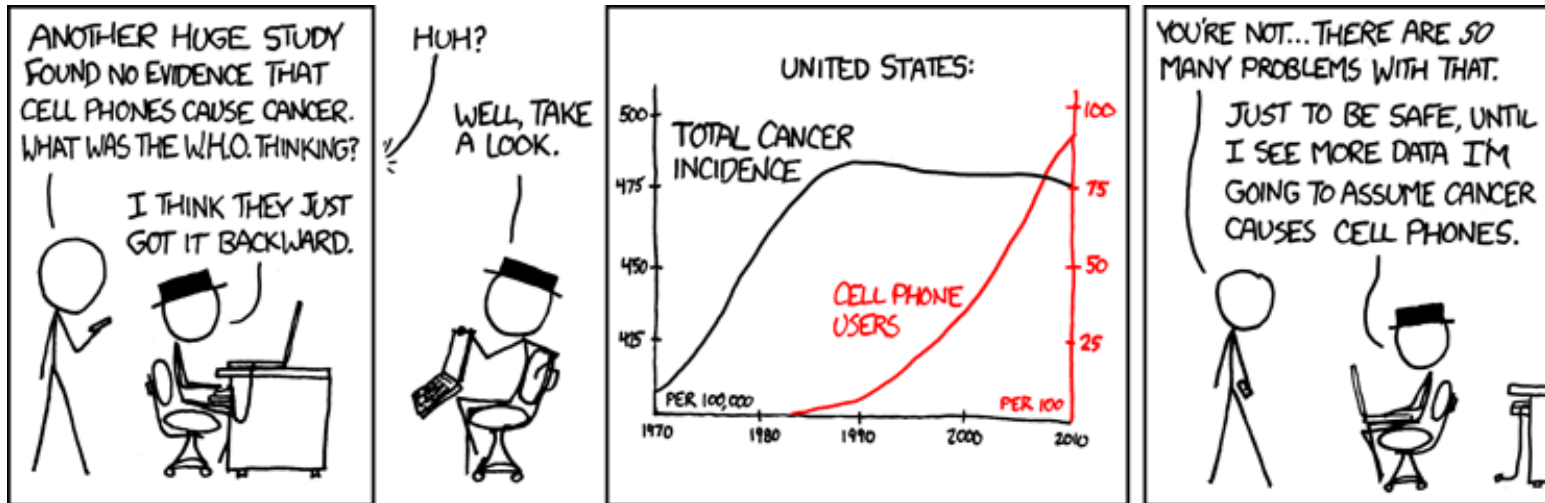


Multiple Linear Regression & AIC

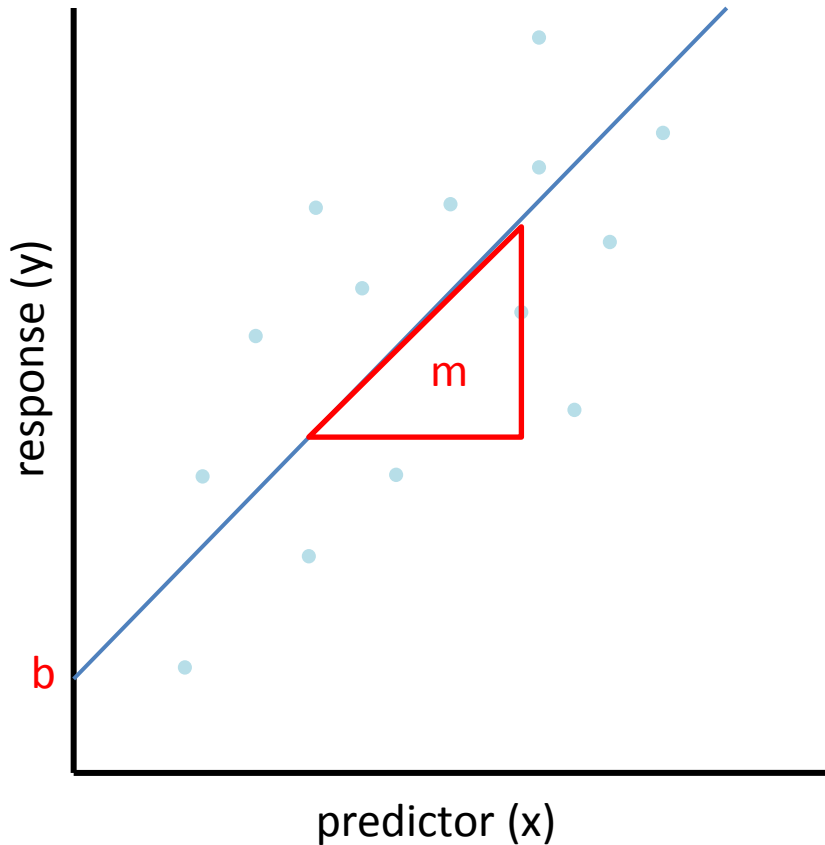


“I've come loaded with statistics, for I've noticed that a man can't prove anything without statistics. No man can.”

Mark Twain (Humorist & Writer)

Linear Regression

Linear relationships



Regression Analysis

PART 1: find a relationship between response variable (Y) and a predictor variable (X)
(e.g. $Y \sim X$)

PART 2: use relationship to predict Y from X

Equation of a line: $y = mx + b$

m = slope of the line $\left(\frac{RISE}{RUN}\right)$

b = y-intercept

Simple Linear Regression in R:

```
lm(response~predictor)  
summary(lm(response~predictor))
```

Multiple Linear Regression

Linear relationship developed from more than 1 predictor variable

Simple linear regression: $y = b + m * x$
 $y = \beta_0 + \beta_1 * x_1$

Multiple linear regression: $y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 \dots + \beta_n * x_n$

β_i is a parameter estimate used to generate the linear curve

Simple linear model: β_1 is the slope of the line

Multiple linear model: β_1, β_2 , etc. work together to generate a linear curve

β_0 is the y-intercept (both cases)

Multiple Linear Regression

Relating model back to data table

ID	DBH	VOL	AGE	DENSITY
1	11.5	1.09	23	0.55
2	5.5	0.52	24	0.74
3	11.0	1.05	27	0.56
4	7.6	0.71	23	0.71
5	10.0	0.95	22	0.63
6	8.4	0.78	29	0.63

Response variable (Y)

Predictor variable 1 (x_1)

Predictor variable 2 (x_2)

Multiple linear regression:

$$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2$$

$$\text{DENSITY} = \text{Intercept} + \beta_1 * \text{AGE} + \beta_2 * \text{VOL}$$

β_1, β_2 : What I need to multiply AGE and VOL by (respectively) to get the value in DENSITY (predicted)

Remember the difference between the observed and predicted DENSITY are our regression residuals

Smaller residuals = Better Model

Multiple Linear Regression

Linear relationship developed from more than 1 predictor variable

Simple linear regression: $y = b + m * x$
 $y = \beta_0 + \beta_1 * x_1$

Multiple linear regression: $y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 \dots + \beta_n * x_n$

β_i is a parameter estimate used to generate the linear curve

Simple linear model: β_1 is the slope of the line

Multiple linear model: β_1, β_2 , etc. work together to generate a linear curve

β_0 is the y-intercept (both cases)

Multiple Linear Regression in R:

```
lm(response~predictor1+predictor2+...+predictorN)  
summary(lm(response~predictor1+predictor2+...+predictorN))
```

Multiple Linear Regression

Output from R

```
R Console
> output2=lm(DENSITY~VOL+AGE,data=data)
> summary(output2)

Call:
lm(formula = DENSITY ~ VOL + AGE, data = data)

Residuals:
    1     2     3     4     5     6     7 
-8.404e-03  5.282e-05  6.304e-03  2.761e-02  2.148e-02 -2.913e-03 -3.170e-02
    8 
-1.243e-02

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.016176   0.080493   12.624 5.54e-05 ***
VOL          -0.326286   0.045100   -7.235 0.000787 ***
AGE          -0.004440   0.002997   -1.481 0.198610
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02237 on 5 degrees of freedom
Multiple R-squared:  0.9187,    Adjusted R-squared:  0.8861 
F-statistic: 28.23 on 2 and 5 DF,  p-value: 0.001887

> |
```

Estimate of model parameters
(β_i values)

Standard error of estimates

Coefficient of determination
a.k.a “Goodness of fit”

Measure of how close the data are to the
fitted regression line
 R^2 and Adjusted R^2

The significance of the overall
relationship described by the
model

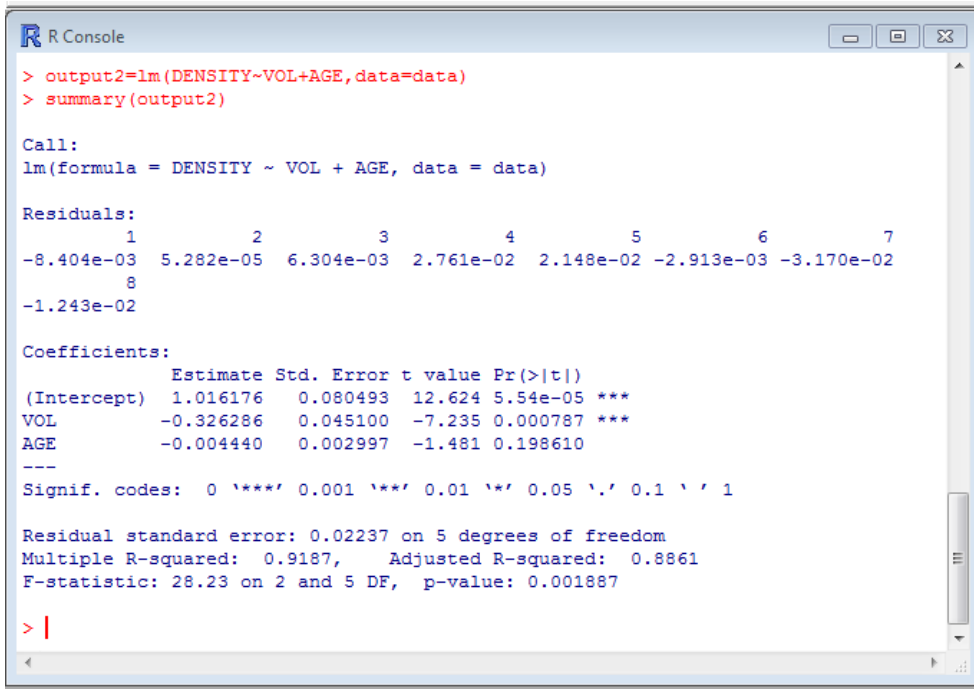
Tests the null hypothesis that the coefficient is equal to zero (no effect)

A predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable

A large p-value suggests that changes in the predictor are not associated with changes in the response

Multiple Linear Regression

Adjusted R-squared



```
R Console
> output2=lm(DENSITY~VOL+AGE,data=data)
> summary(output2)

Call:
lm(formula = DENSITY ~ VOL + AGE, data = data)

Residuals:
    1      2      3      4      5      6      7      8 
-8.404e-03  5.282e-05  6.304e-03  2.761e-02  2.148e-02 -2.913e-03 -3.170e-02 -1.243e-02 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.016176   0.080493   12.624 5.54e-05 ***
VOL          -0.326286   0.045100   -7.235 0.000787 ***
AGE          -0.004440   0.002997   -1.481 0.198610
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02237 on 5 degrees of freedom
Multiple R-squared:  0.9187,    Adjusted R-squared:  0.8861 
F-statistic: 28.23 on 2 and 5 DF,  p-value: 0.001887

> |
```

Test statistic:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

$$R^2 = \frac{SS_{regression}}{SS_{total}}$$

$$R^2_{adj} = R^2(1 - R^2) \left(\frac{p}{n - p - 1} \right)$$

p = number of predictor variables
(regressors, not including intercept)
 n = sample size

- Adjusted R^2 is always positive
- Ranges from 0 to 1 with values closer to 1 indicating a stronger relationship
- Adjusted R^2 is the value of R^2 which has been penalized for the number of variables added to the model
- Therefore Adjusted R^2 is always smaller than R^2

Multiple Linear Regression

Adjusted R-squared

Why do we have to Adjust R^2 ?

For multiple linear regression there are 2 problems:

- **Problem 1:** Every time you add a predictor to a model, the R-squared increases, even if due to chance alone. It never decreases. Consequently, a model with more terms may appear to have a better fit simply because it has more terms.
- **Problem 2:** If a model has too many predictors and higher order polynomials, it begins to model the random noise in the data. This condition is known as **over-fitting the model** and it produces misleadingly high R-squared values and a lessened ability to make predictions.

Therefore for Multiple Linear Regression you need to report the Adjusted R^2 which accounts for the number of predictors you had to added

Akaike's Information Criterion (AIC)

How do we decide what variable to include?



Hirotugu Akaike, 1927-2009

In the 1970s he used information theory to build a numerical equivalent of Occam's razor

Occam's razor: All else being equal, the simplest explanation is the best one

- In statistics, this means a model with fewer parameters is to be preferred to one with more
 - Of course, this needs to be weighed against the ability of the model to actually predict anything
-
- AIC considers both the fit of the model and the number of parameters used
 - More parameters result in a *penalty*

Akaike's Information Criterion (AIC)

How do we decide what variable to include?

- The *model fit* (AIC value) is measured as likelihood of the parameters being correct for the population based on the observed sample
- The number of parameters is derived from the degrees of freedom that are left
- AIC value roughly equals the number of parameters minus the likelihood of the overall model
 - Therefore the smaller the AIC value the better the model
- Allows us to balance over- and under-fitting in our modelled relationships
 - We want a model that is as simple as possible, but no simpler
 - A reasonable amount of explanatory power is traded off against model size
 - AIC measures the balance of this for us

Akaike's Information Criterion (AIC)

AIC in R

- Stepwise model comparison is an iterative model evaluation that will either:
 1. Starts with a single variable, then adds variables one at a time (“forward”)
 2. Starts with all variables, iteratively removing those of low importance (“backward”)
 3. Run in both directions (“both”)
- The order of the variables matters – therefore it is best to run the stepwise model comparison in all directions and compare AIC values

Akaike's Information Criterion in R to determine predictors:

```
step(lm(response~predictor1+predictor2+predictor3), direction="backward")  
step(lm(response~predictor1+predictor2+predictor3), direction="forward")  
step(lm(response~predictor1+predictor2+predictor3), direction="both")
```

Akaike's Information Criterion (AIC)

AIC Output from stepwise procedure

```
R Console
> stepAIC(lm(DENSITY~VOL+AGE+DBH,data=data), direction="both")
Start:  AIC=-57.32
DENSITY ~ VOL + AGE + DBH

   Df Sum of Sq    RSS   AIC
- VOL  1 0.00002009 0.0022951 -59.251
- DBH  1 0.00022625 0.0025013 -58.563
<none>                 0.0022750 -57.322
- AGE  1 0.00131371 0.0035887 -55.675

Step:  AIC=-59.25
DENSITY ~ AGE + DBH

   Df Sum of Sq    RSS   AIC
<none>                 0.0022951 -59.251
- AGE  1 0.0013601 0.0036552 -57.528
+ VOL  1 0.0000201 0.0022750 -57.322
- DBH  1 0.0263906 0.0286857 -41.046

Call:
lm(formula = DENSITY ~ AGE + DBH, data = data)

Coefficients:
(Intercept)      AGE      DBH 
  1.037405   -0.004935   -0.031821

> |
```

AIC value for full model
(starting point)

Backward Selection

If I remove VOL

What happens to my AIC? ($\sim \beta_0 + \text{AGE} + \text{DBH}$)

Now remove DBH

What happens to my AIC? ($\sim \beta_0 + \text{AGE}$)

Now remove AGE

What happens to my AIC? ($\sim \beta_0$)

Best model (lowest AIC) is when
VOL is removed

Test this model again

Forward Selection

Start with $\text{DENSITY} \sim \beta_0 + \text{AGE} + \text{DBH}$

If I remove AGE

What happens to my AIC? ($\sim \beta_0 + \text{DBH}$)

If I add VOL

What happens to my AIC? ($\sim \beta_0 + \text{DBH} + \text{VOL}$)

If I remove DBH

What happens to my AIC? ($\sim \beta_0 + \text{VOL}$)

What is the best model to use? (What has the lowest AIC?)

What are the parameter estimates of this model?

Multiple Linear Regression Assumptions

1. For any given value of X , the distribution of Y must be normal
 - BUT Y does not have to be normally distributed as a whole
2. For any given value of X , Y must have equal variances

You can again check this by using the Shapiro Test, Bartlett Test, and residual plots on the residuals of your model

What we have all ready been doing!

No assumptions for X – but be conscious of your data

Collinearity a.k.a Multicollinearity

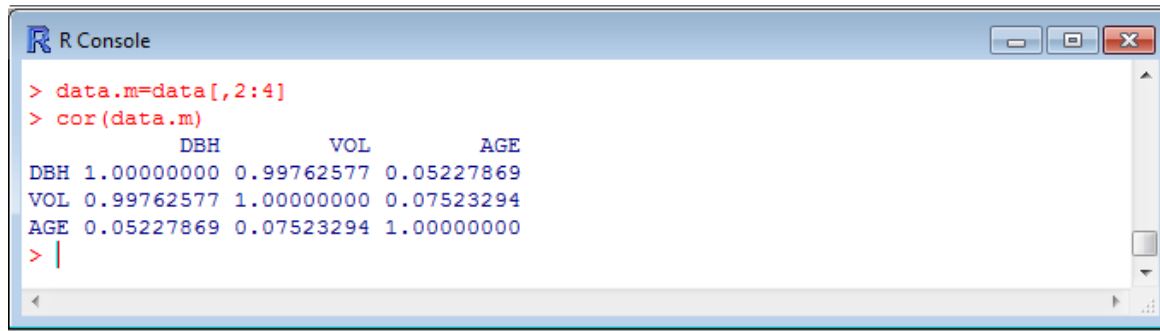
Problem with predictors

- Occurs when predictor variables are related (linked) to one another
 - Meaning that one predictor can be linearly predicted from the others with a non-trivial degree of accuracy
 - E.g. Climate (mean summer precipitation and heat moisture index)
- Does not reduce the predictive power or reliability of the model as a whole
 - *BUT* the coefficient estimates may change erratically in response to small changes in the model or the data
- A model with correlated predictors CAN indicate how well the combination of predictors predicts the outcome variable
 - *BUT* it may not give valid results about any individual predictor, or about which predictors are redundant with respect to others

Collinearity a.k.a Multicollinearity

What to do

- If you suspect your predictor variables are correlated you can calculate a matrix of correlation values between your variables to confirm



```
> data.m=data[,2:4]
> cor(data.m)
```

	DBH	VOL	AGE
DBH	1.00000000	0.99762577	0.05227869
VOL	0.99762577	1.00000000	0.07523294
AGE	0.05227869	0.07523294	1.00000000

Here VOL and DBH are highly correlated

- But it is up to you to judge what might be a problematic relationship

Collinearity a.k.a Multicollinearity

What to do

- Whether or not you choose to use Multiple Regression Models depends on the question you want to answer
 - Are you interested in establishing a relationship?
 - Are you interested in which predictors are driving that relationship?
- There are alternative techniques that can deal with highly correlated variables – these are mostly multivariate
 - Regression Trees = can handle correlated data well

Important to Remember

A multiple linear relationship **DOES NOT** imply causation!

Adjusted R^2 implies a relationship rather than one or multiple factors causing another factor value

Be careful of your interpretations!