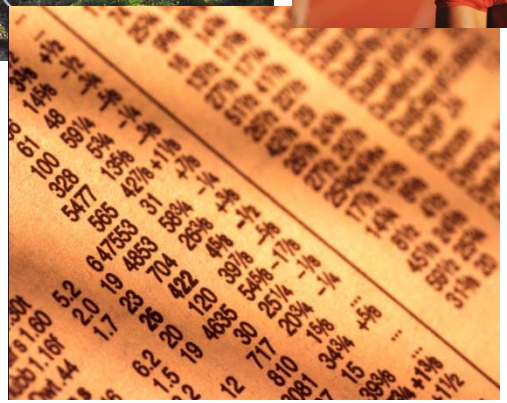
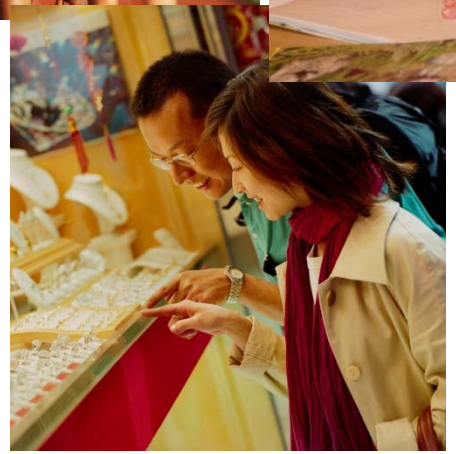
A central question in all human endeavor:

I've done something differently and I got different results… is this significant?

Why is math necessary to answer this question?

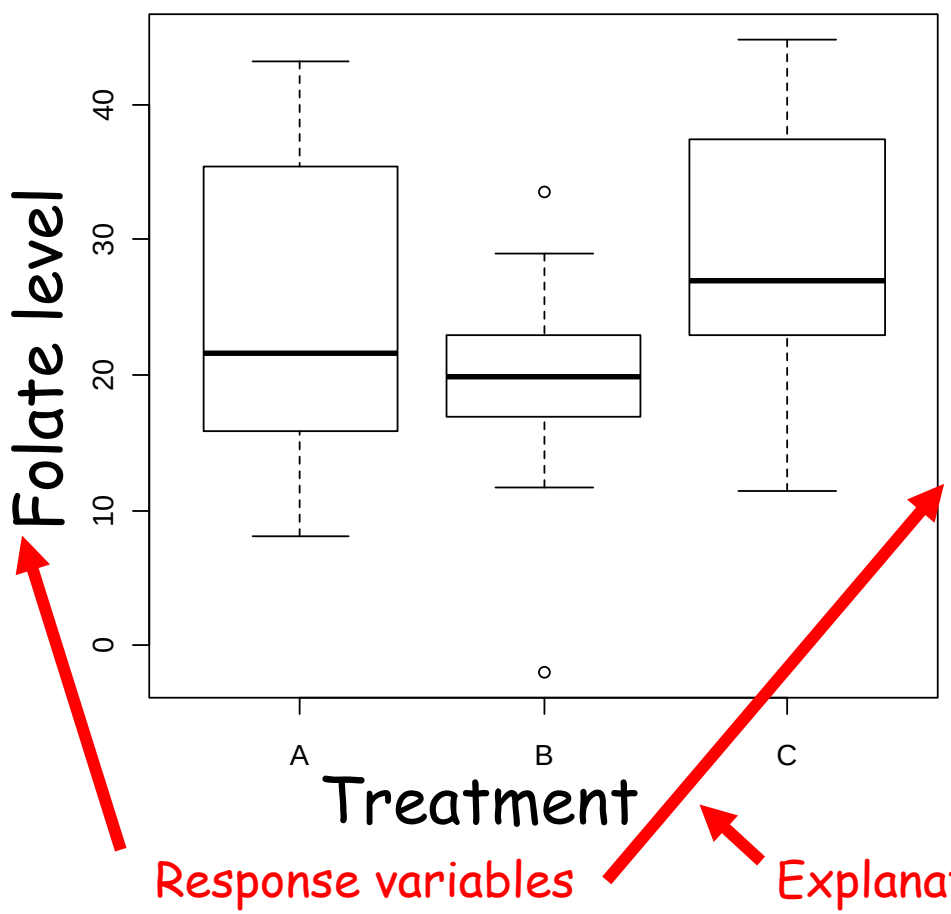


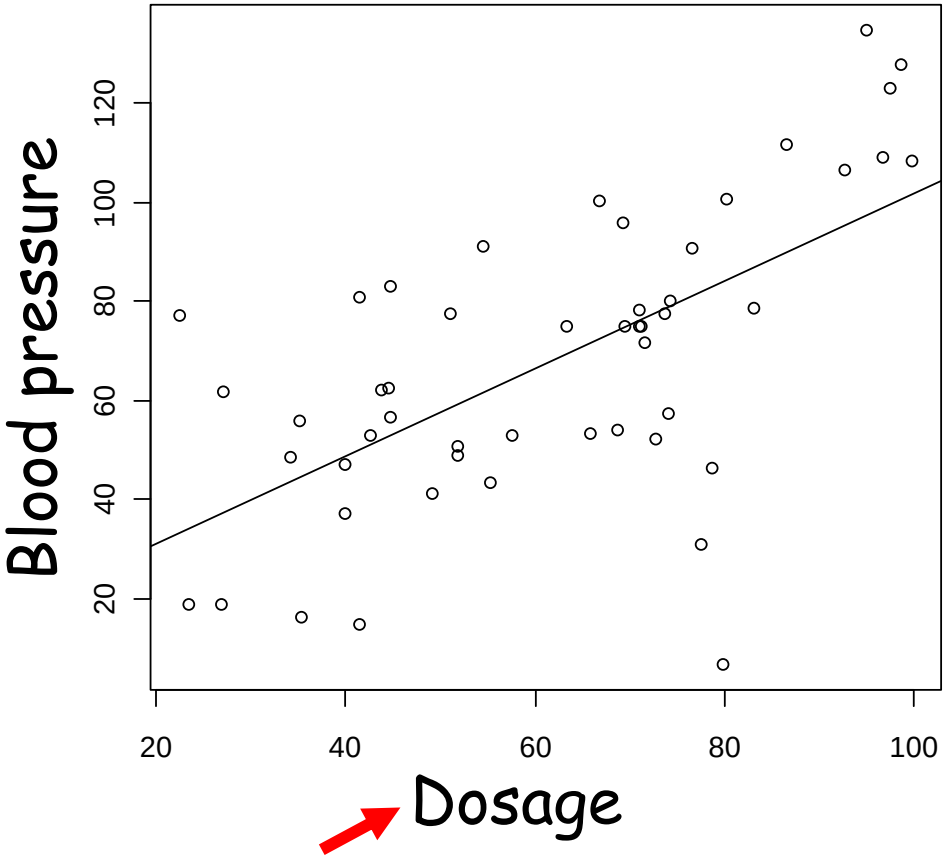


(Images from MS Powerpoint Clipart collection or M. Colvin)

Answering this question involves distinguishing random variability from causal differences





Response variables

Explanatory variables

Key idea – compare how the difference in values between treatments compare to the variation within a treatment group

The kind of statistical model you build depends on the number and type of explanatory variables

Continuous X

- Continuously varying
- Values have meaning as numbers
- Values are ordered
- Interpolation makes sense

Multiple Regression:

$$Y = \alpha + \beta_1 \times X_1 + \beta_2 \times X_2$$

Categorical X

- Discrete values
- Values are just "names" that define subsets
- Values are unordered
- Interpolation is meaningless

Multiway analysis of variance:

$$Y = \mu + \begin{bmatrix} 0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} 0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

A statistical model describes the relationship between a response variable and 1 or more explanatory variables

folate~ventilation

folate: measured levels of folate in red blood cells in post-op patients
ventilation: methods of ventilation while under anesthesia (three types)

met.rate~weight

met.rate: resting metabolic rate (kcal/day)
weight: body weight (kg)

Calories~Sodium+Type

Calories: calorie concentration in meat
Sodium: sodium concentration
Type: type of meat

cost~carat+color+clarity

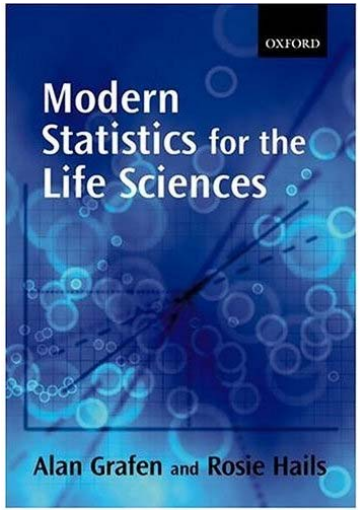
cost: diamond cost ($ Singapore)
carat: diamond carat weight
color: diamond color (D, E, F, …)
clarity: diamond clarity (FL, IF, VVS_1, …)

glucose~pregnant+diastolic+triceps+insulin+bmi+diabetes+age

glucose: 2-hr plasma glucose
pregnant: Number of times pregnant
diastolic: Diastolic blood pressure (mm Hg)
triceps: Triceps skin fold thickness (mm)
insulin: 2-hr serum insulin (mu U/ml)
bmi: Body mass index (weight in kg/(height in m)^2)
diabetes: Diabetes pedigree function (family history)
age: Age (years)

All parametric statistical tests can be included in one framework called the General Linear Model

From the preface of the BIO180/QSB 280 text:



"[This approach] is based on a grand conceptual scheme, called the General Linear Model or GLM. This contains within it all the usual parametric tests, including t-tests, analysis of variance, contrast analysis, linear regression, multiple regression, analysis of covariance, and polynomial regression. *Instead of learning these as separate tests with broadly similar features but maddening differences, this book will teach you a single coherent framework. Instead of a mish-mash of eccentrically named accidents, this book presents statistics as a meaningful whole.*"

In R, the General Linear Model is performed using the `lm()` command

GLM framework covers many statistical tests

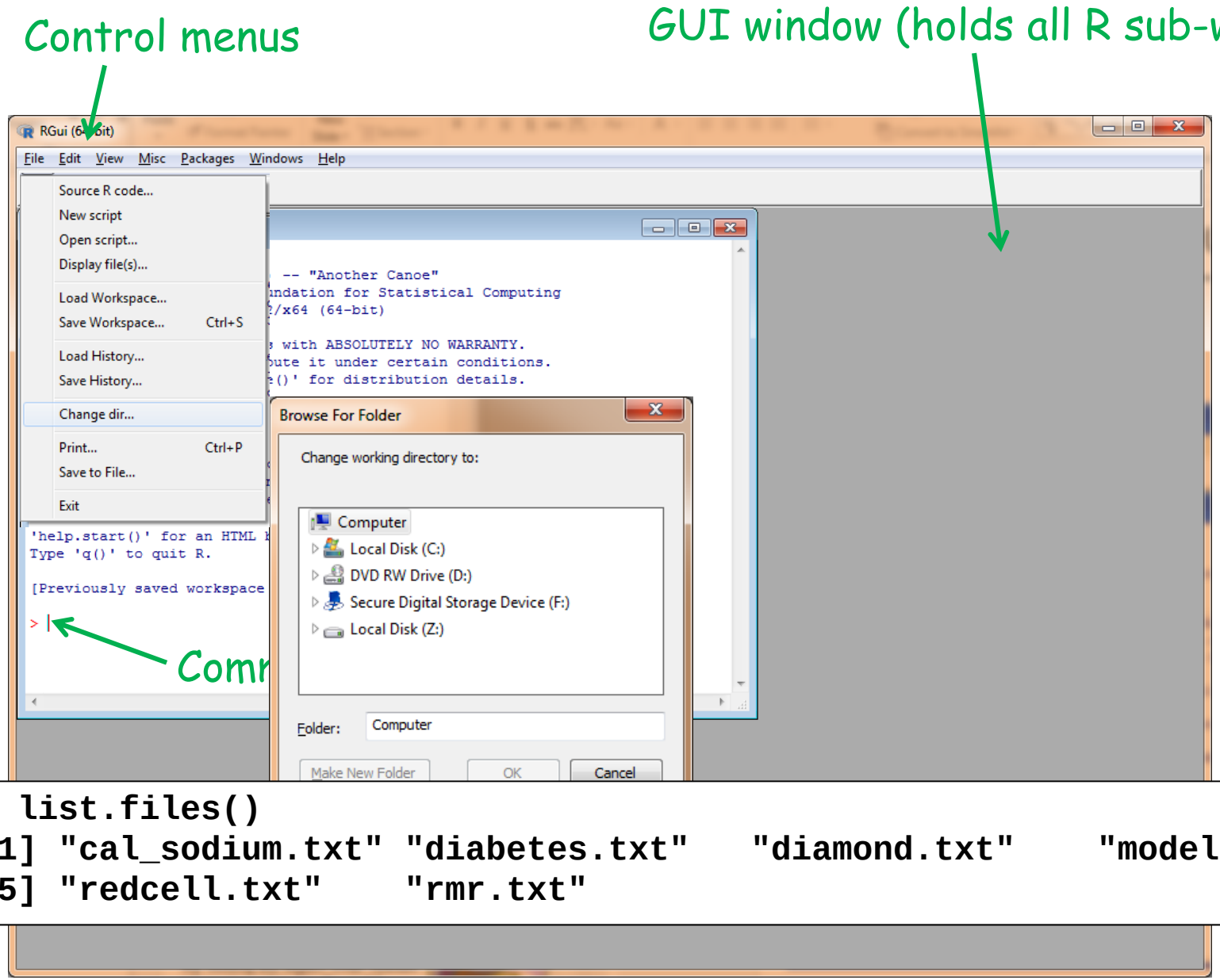
Blue indicates categorical variable ***Red (italic) indicates continuous variable***

Traditional test	Example	Model formula
2-sample *t*-test	Comparing yield of two types of fertilizer	***YIELD*=FERTIL** ("~" instead of "=" in R)
One way ANOVA	Comparing yield of 3 or more types of fertilizer	***YIELD*=FERTIL**
Blocked ANOVA	Comparing yield of fertilizers in blocked experiment	***YIELD*=BLOCK+FERTIL**
Regression	Predicting yield based on amount of water	***YIELD*=*WATER***
Analysis of covariance	Yield predicted from type of fertilizer and amount of water	***YIELD*=FERTIL+ *WATER***
Multiple Regression	Yield predicted from amount of water and sunlight	***YIELD*=*WATER*+*SUN***
Two way ANOVA	Comparing yield of different fertilizers and seed brands	***YIELD*=FERTIL+ SEEDBRAND**

What do we get from our statistical models?

1. A well-defined p-value to address your overall Null Hypothesis that the response variable is not affected by any of your explanatory variables
2. Well-defined p-values for the Null Hypotheses that each individual explanatory variable does not affect your response variable (after statistical elimination)
3. An R^2 value that tells you the fraction of variation in response explained by explanatory variables
4. Optimal fitted parameters for predicting responses from explanatory variables (i.e. your model parameters: slopes, offsets, etc.)

To start using R, double click on the icon



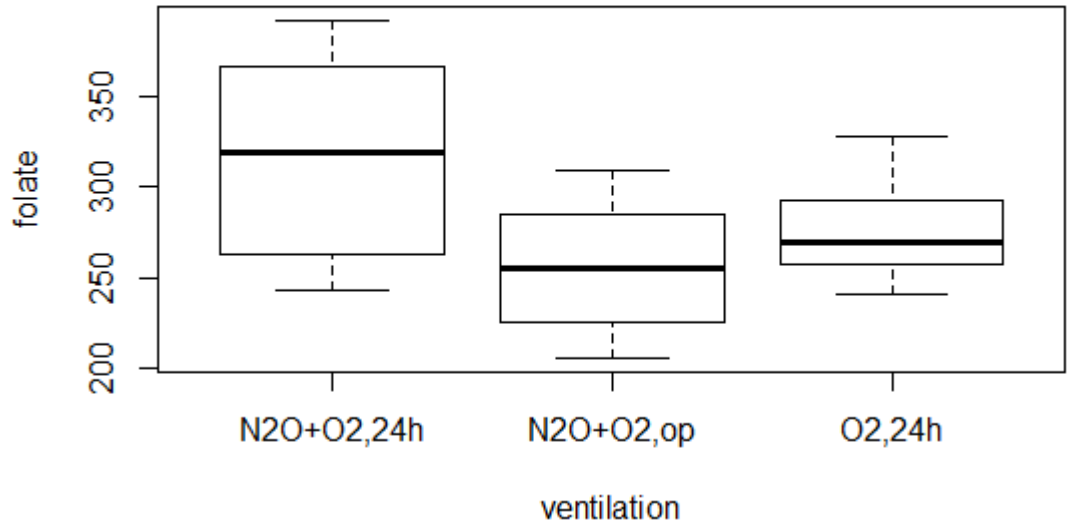

```
> list.files()
[1] "cal_sodium.txt" "diabetes.txt"   "diamond.txt"    "model.R"
[5] "redcell.txt"    "rmr.txt"
```

Build a GLM for the blood folate dataset with 1 categorical explanatory variable

Commands to type at R prompt:

```
redcell<-read.table("redcell.txt", header=T)
redcell
names(redcell)
plot(folate~ventilation, data=redcell)
model<-lm(folate~ventilation, data=redcell)
summary(model)
```

"Box and whiskers" plot:



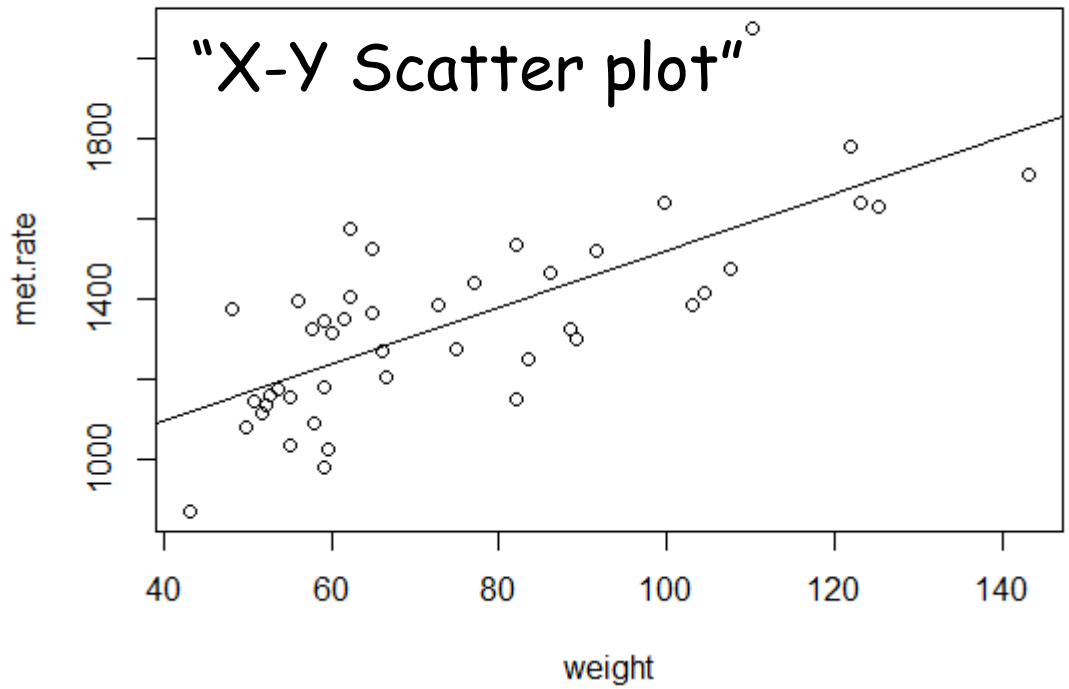
The model "summary" presents the key results from the general linear model

```
Call:
lm(formula = folate ~ ventilation, data = redcell)

Residuals:
    Min      1Q  Median      3Q     Max
-73.625 -35.361  -4.444  35.625  75.375

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            316.62      16.16  19.588 4.65e-14 ***
ventilationN2O+O2,op   -60.18      22.22  -2.709   0.0139 *
ventilationO2,24h      -38.62      26.06  -1.482   0.1548
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 45.72 on 19 degrees of freedom
Multiple R-squared:  0.2809,    Adjusted R-squared:  0.2052
F-statistic: 3.711 on 2 and 19 DF,  p-value: 0.04359
```

Build a GLM for the metabolic rate dataset with 1 continuous explanatory variable

Commands to type at R prompt:

```
rmr<-read.table("rmr.txt", header=T)
names(rmr)
plot(met.rate~weight, data=rmr)
model<-lm(met.rate~weight, data=rmr)
abline(model)
summary(model)
```



The model summary presents the key results from the general linear model

```
lm(formula = met.rate ~ weight, data = rmr)

Residuals:
    Min      1Q  Median      3Q     Max
-245.74 -113.99  -32.05  104.96  484.81

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 811.2267    76.9755  10.539 2.29e-13 ***
weight        7.0595     0.9776   7.221 7.03e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 157.9 on 42 degrees of freedom
Multiple R-squared:  0.5539,    Adjusted R-squared:  0.5433
F-statistic: 52.15 on 1 and 42 DF,  p-value: 7.025e-09
```

The script **model.R** automates these steps and prints the statistical data to **model.out** & the graph to **graph.png**. This can be run from the **file/Source R code…** menu

Next build a model for the meat dataset with 1 continuous & 1 categorical explanatory variable

Commands to type at R prompt:

```
cs<-read.table("cal_sodium.txt", header=T)
plot(Calories~Sodium, data=cs)
plot(Calories~Type, data=cs)
model<-lm(Calories~Sodium, data=cs)
summary(model)
model<-lm(Calories~Type, data=cs)
summary(model)
model<-lm(Calories~Sodium+Type, data=cs)
summary(model)
anova(model)
```

Summary results for all three models

Calories~Sodium

```
Multiple R-squared:  0.2182,    Adjusted R-squared:  0.2032
F-statistic: 14.51 on 1 and 52 DF,  p-value: 0.0003693
```

Calories~Type

```
Multiple R-squared:  0.3866,    Adjusted R-squared:  0.3626
F-statistic: 16.07 on 2 and 51 DF,  p-value: 3.862e-06
```

Calories~Sodium+Type

```
Multiple R-squared:  0.7934,    Adjusted R-squared:  0.781
F-statistic: 64.01 on 3 and 50 DF,  p-value: < 2.2e-16

Analysis of Variance Table
Response: Calories
           Df  Sum Sq Mean Sq F value    Pr(>F)
Sodium      1  9985.5  9985.5  52.815 2.292e-09 ***
Type        2 26320.6 13160.3  69.607 3.551e-15 ***
Residuals 50  9453.2   189.1
```

Build a model for diamond cost with 1 continuous and 2 categorical explanatory variables

```
diamond<-read.table("diamond.txt", header=T)
names(diamond)
model<-lm(cost~carat+color+clarity, data=diamond)
summary(model)
anova(model)
```

```
Multiple R-squared:  0.958,     Adjusted R-squared:  0.9565
F-statistic: 676.7 on 10 and 297 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Response: cost
            Df     Sum Sq    Mean Sq  F value    Pr(>F)
carat        1 1350730005 1350730005 6304.435 < 2.2e-16 ***
color        5   60591911   12118382   56.562 < 2.2e-16 ***
clarity      4   38450149    9612537   44.866 < 2.2e-16 ***
Residuals  297   63632471     214251
```

Are all three explanatory variables essential to get a good model for diamond cost?

```
summary(lm(cost~carat+color, data=diamond))
summary(lm(cost~carat+clarity, data=diamond))
```
Etc.

Variables	R^2
carat+color+clarity	0.958
carat+color	0.933
carat+clarity	0.910
color+clarity	0.114
carat	0.893
color	0.084
clarity	0.030

Finally, a model with 7 explanatory variables

```
diabetes <- read.table("diabetes.txt",header=T)
model<-lm(glucose~.,data=diabetes)
summary(model)
anova(model)
```

```
Multiple R-squared:  0.4012,    Adjusted R-squared:  0.3903
F-statistic: 36.76 on 7 and 384 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table
Response: glucose
           Df Sum Sq Mean Sq  F value    Pr(>F)
pregnant    1  14642   14642  25.2158 7.864e-07 ***
diastolic   1  10975   10975  18.9004 1.766e-05 ***
triceps     1   8067    8067  13.8933 0.0002226 ***
insulin     1 107840  107840 185.7190 < 2.2e-16 ***
bmi         1    127     127   0.2183 0.6406234
diabetes    1   1202    1202   2.0706 0.1509758
age         1   6556    6556  11.2901 0.0008573 ***
```

How can we select the minimal # of expl. vars?

There are a number of automated methods to help in multiple regression

Possible pitfalls:

1. Temptation just to let the computer do the thinking and neglect other relevant information related to the model
2. Slightly different automated procedures can give different models (although usually with similar significance)
3. Don't take overall p-value of final model too literally, due to multiplicity of p-value issues. (To be *really* conservative, you can divide p_{sig} by $2^{(\text{n explanatory vars})}$)

Number of possible subsets of n explanatory variables

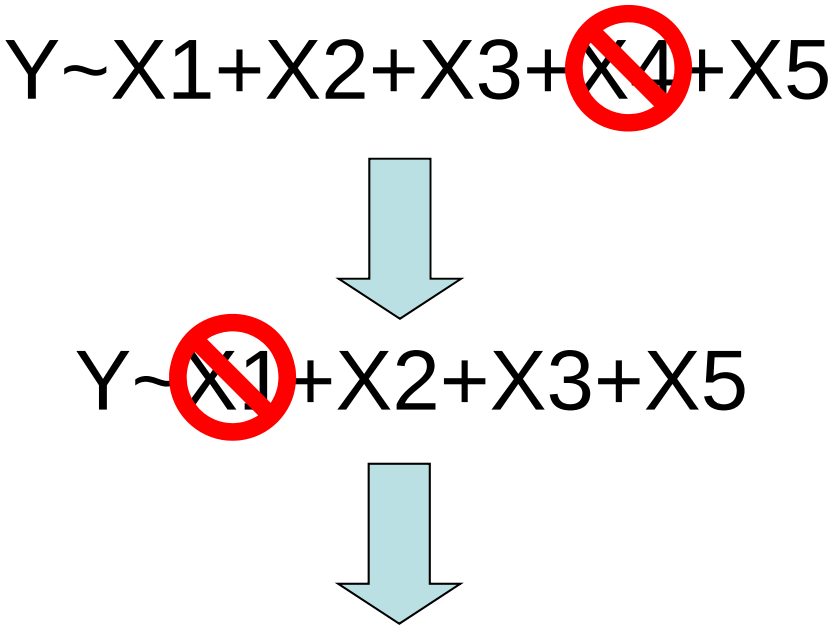
Stepwise regression is an automated procedure for selecting a subset of variables in model

Backwards Stepwise Regression

Example: Model for Y vs. five explanatory variables: X1, X2, X3, X4, X5

Step 1:
Build full model and remove variable that contributes least (by some measure)

Step 2:
Build new model and remove variable that contributes least



Stop when all remaining variables fulfill some criterion

Conversely, there is forward stepwise regression

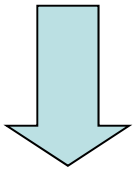
Stepwise regression can also be run in the forward direction

Forwards Stepwise Regression

Example: Model for Y vs. five explanatory variables: X1, X2, X3, X4, X5

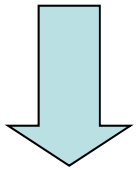
Step 1:
Build 5 models & pick best
(by some metric)

Y~X1 Y~X2 **Y~X3** Y~X4 Y~X5



Step 2:
Build 4 models & pick best
(by some metric)

Y~X3+X1 **Y~X3+X2** Y~X3+X4 Y~X3+X5



Stop when additional variable does not improve Adjusted R^2 (or some other metric for model quality)

Backward stepwise regression on diabetes data

```
model<-lm(glucose~.,data=diabetes)
model2 <- step(model)
summary(model2)
```

```
Step:  AIC=2498.07
glucose ~ diastolic + insulin + diabetes + age

            Df Sum of Sq    RSS    AIC
- diabetes   1      1071 224845 2497.9
<none>                   223773 2498.1
- diastolic  1      3434 227207 2502.0
- age        1     12396 236169 2517.2
- insulin    1     94915 318688 2634.7

Step:  AIC=2497.95
glucose ~ diastolic + insulin + age

            Df Sum of Sq    RSS    AIC
<none>                   224845 2497.9
- diastolic  1      3258 228103 2501.6
- age        1     12964 237809 2517.9
- insulin    1     98886 323731 2638.8
```

7 variable model

0.3903

Multiple R-squared: 0.3962, Adjusted R-squared: 0.3915

Forward stepwise regression on diabetes data

```
model<-lm(glucose~1,data=diabetes)
step(model,direction="forward",
     scope=list(upper=terms(glucose~.,data=diabetes)))
```

```
Step:  AIC=2501.59
glucose ~ insulin + age

             Df Sum of Sq    RSS    AIC
+ diastolic  1    3258.3 224845 2497.9
+ bmi        1    2104.8 225998 2499.9
+ triceps    1    1591.2 226512 2500.8
<none>                   228103 2501.6
+ diabetes   1     896.0 227207 2502.0
+ pregnant   1       1.0 228102 2503.6

Step:  AIC=2497.95
glucose ~ insulin + age + diastolic

            Df Sum of Sq    RSS    AIC
<none>                   224845 2497.9
+ diabetes  1   1071.35 223773 2498.1
+ bmi       1    916.49 223928 2498.3
+ triceps   1    873.04 223972 2498.4
+ pregnant  1      0.00 224845 2499.9
```

Same result we got from backward stepwise regression