Linear Models

Two sample tests

- tests such as the t-test or Wilcoxon are used to compare two samples
- there is no obvious way to adjust for, or control for other variables
- eg we might want to adjust for age and sex when comparing gene expression values across human samples
- to do that we consider more general regression models

A simple experiment

- we are interested in comparing gene expression between two groups of people (n=10 in each group)
- blood is drawn and baseline for RNA-seq analysis
- participants are randomly split into two groups,
- Group 1 and Group 2
  - Group 1 goes for 1 week to a resort at an altitude of 5K ft.
  - Group 2 goes for 1week to a resort at sea level
  - both groups go through the same amount of exercise and are given the same diet
- RNA is extracted
- we sequence, get counts and want to compare the changes in gene expression
  - so we have 20K genes, and for each one 10 measurements for each group
- careful examination of the data suggests that we:
  - add one to the counts and then use the log of RNA count
- that we model difference in the log of the (counts + 1) pre and post test
- we consider these are our responses (one test for each gene)

The t-test as linear regression

- for each gene, the t-test is then the difference in means between the two groups divided by an estimate of the standard error

$$\frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}_p}$$

- An equivalent form of the t-test for two samples (compare Group 1 to Group 2)

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon$$

  - where $x_i = 0$ if the i^{th} person is in Group 1 and $x_i=1$ if the i^{th} person is in Group 2
  - and $\varepsilon \sim N(0,\sigma^2)$
  - $E[Y|X=0] = \beta_0 = \mu_1$
  - $E[Y|X=1] = \beta_0+\beta_1=\mu_2$
- So a test of $\beta_1=0$, is the same as the t-test that the means in the two groups are the same
- We can show that the two tests are identical

Linear Models

- the main reason to consider the linear model approach is that it allows us to easily include other variables

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon$$

- where β_2 could be sex and β_3 could be age, for example
- sex could be encoded as 1 for Female, 0 Male, then β_2 will be the mean change in response for Females.
- β_3 tells us the mean change in *y* for a one unit change in *x* (could be years, if age is measured in years)
- we would then think of β_1 as the effect of our treatment, adjusted for age and sex

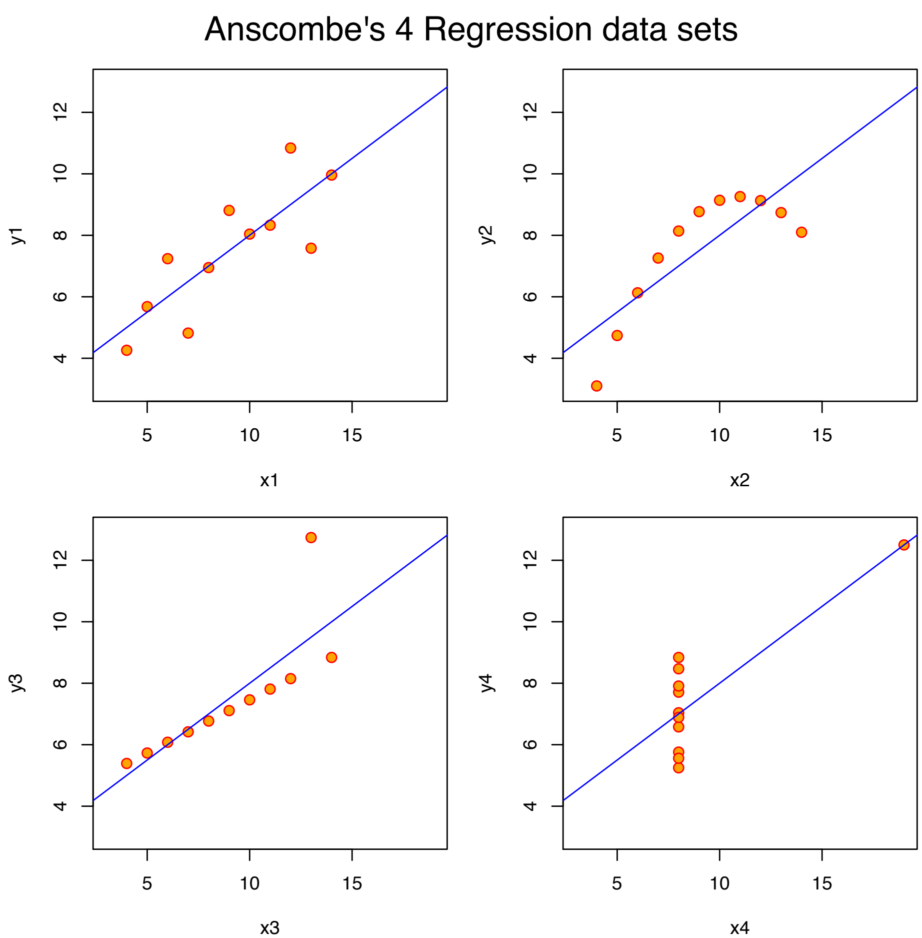
Some assumptions

- that the model holds, at least approximately

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots \beta_k x_{k,i} + \varepsilon_i$$

- that the response *y* is linearly associated with the *x*'s, there are *k* covariates
- that the errors are approximately Normal with approximately constant variance (over all *x*'s)
- Anscombe devised a simple example with four different sets of data, but where the estimates are identical [HW: data(anscombe)….]

Which one is appropriate for linear regression



The outputs:

```
anscmb> lapply(mods, function(fm) coef(summary(fm)))
$lm1
             Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 3.0000909  1.1247468 2.667348 0.025734051
x1          0.5000909  0.1179055 4.241455 0.002169629

$lm2
            Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 3.000909  1.1253024 2.666758 0.025758941
x2          0.500000  0.1179637 4.238590 0.002178816

$lm3
             Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 3.0024545  1.1244812 2.670080 0.025619109
x3          0.4997273  0.1178777 4.239372 0.002176305

$lm4
             Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 3.0017273  1.1239211 2.670763 0.025590425
x4          0.4999091  0.1178189 4.243028 0.002164602
```

Caution

- lm does not check the assumptions of the linear model – nor does it check whether the model actually fit the data
- that is YOUR JOB!
- if your model does not fit the data, or if any of the assumptions are not valid then the parameters really have no interpretation
- your p-values are not interpretable

Some special cases

- Analysis of Variance: ANOVA models
  - usually refer to the case where X specifies a number of different groups
  - typically including interactions
- eg: we want to study the yield from two types of wheat, in two fields
- $Y=\beta_0+\beta_1 X_1 + \beta_2 X_2 + \varepsilon$
- where X_1 is coded 0 for Field 1 and 1 for Field 2
- and X_2 is coded as 0 for Type 1 and 1 for Type 2
- so β_0 is the mean yield for Field 1, Type 1
- $\beta_0+\beta_1$ is the mean yield for Field 2, Type 1
- $\beta_0+\beta_2$ is the mean yield for Field 1, Type 2
- $\beta_0+\beta_1+\beta_2$ is the mean yield for Field 2, Type 2

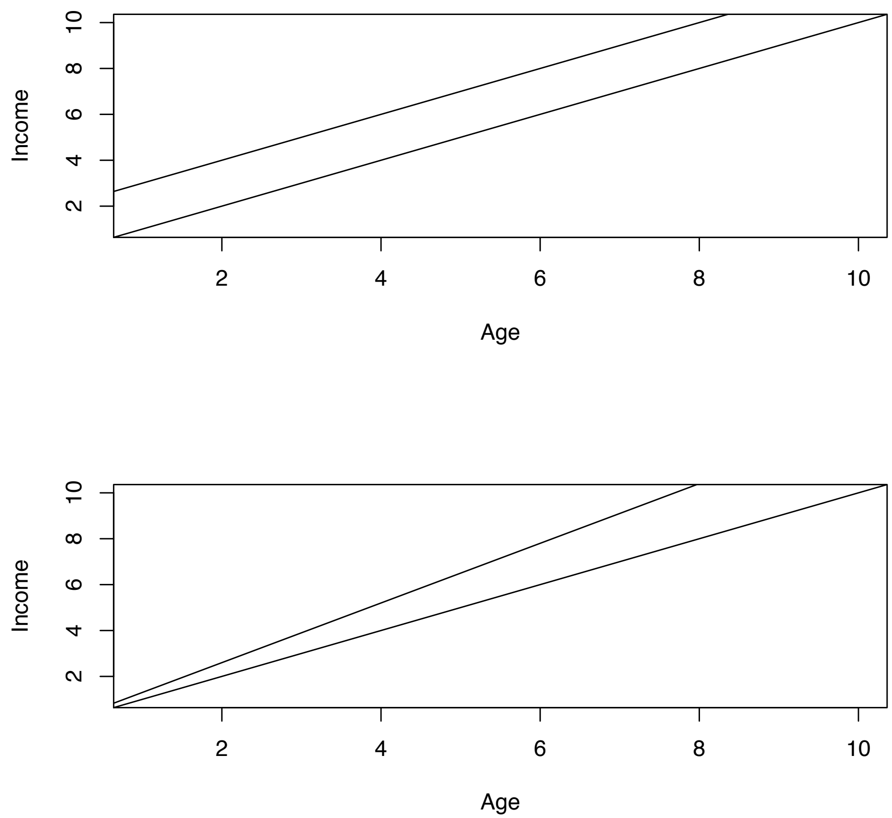
ANOVA

- two types of wheat, two fields we got the model
- $Y=\beta_0+\beta_1 X_1 + \beta_2 X_2 + \varepsilon$
  - where X_1 is coded 0 for Field 1 and 1 for Field 2
  - and X_2 is coded as 0 for Type 1 and 1 for Type 2
- what else are we assuming in this model?
- that there is no interaction! that the effect of the field and that of the type of wheat are the same
- suppose that field 2 is much wetter than field 1
- and suppose that Type 1 likes dry weather, type 2 likes more moisture
- we can model this by adding in one more term to our model
- $Y=\beta_0+\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2+\varepsilon$
  - here β_3 requires both X_1 and X_2 to be 1
  - so it captures those data points for Field 2 and Type 2 simultaneously

Mix continuous and discrete

- income as a function of age (continuous) and sex (M/F)
- $y = \beta_0 + \beta_1 X_A + \beta_2 X_M + \beta_3 X_A{*}X_M + \varepsilon$
  - now β_1 is the effect on income of Age, if β_1 is positive then income increases with age
  - β_2 is the effect for sex (suppose X_M =1 if Male), then that represents the difference between males and females
  - β_3 is the interaction, it allows the slope of the age relationship to be different for men and women

Interactions: mean income by age



- In the top panel we see two parallel lines
- the effect of age is the same for both sexes
- In the bottom panel the lines diverge
- the effect of sex is different for each age

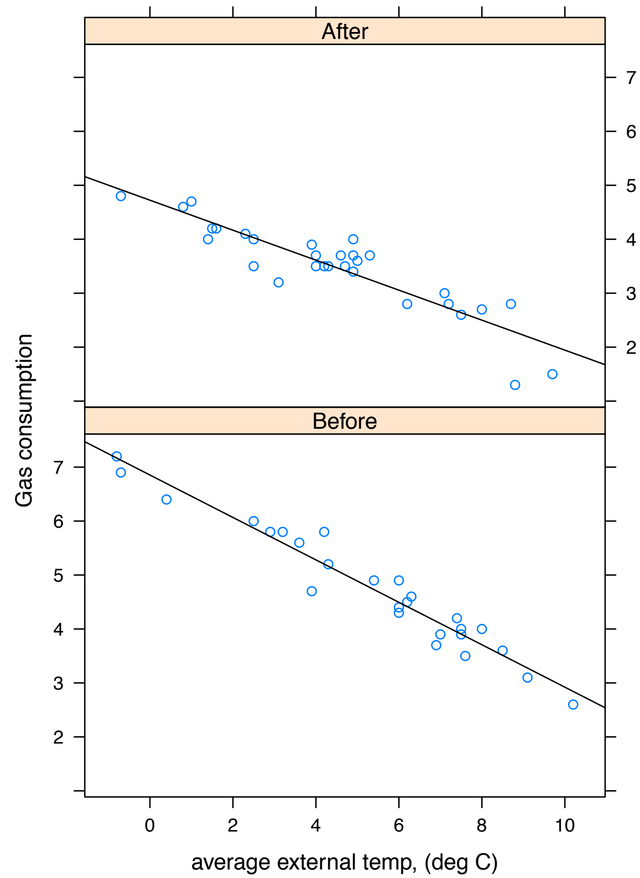
More assumptions

- we assume that the X's are measured without error (there are other models, *errors-in-variables*, that can be used)
- we assume that the *y* measurements are independent
  - this fails when we measure the same person over and over (repeated measures)
  - it fails for almost all mouse experiments (litter effects, shared cages and so on)
  - addressing these concerns usually requires the use of so-called random effects models, or mixed-effects models

Modeling in R

- lm is the main function
- a simple example from Modern Applied Statistics, Chapter 6 (Venables andRipley)
- library(MASS); data(whiteside)
- the data consist of measurements before and after Mr. Whiteside added insulation to his home
  - mean temperature in degrees C for the week
  - gas consumption for the week
  - before and after insulation

Plot the data



Now fit some models

- gasA = lm(Gas~Temp, data=whiteside, subset=Insul==“Before”)
- gasB = lm(Gas~Temp, data=whiteside, subset=Insul==“After”)
- summary(gasA)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.72385    0.12974   36.41  < 2e-16 ***
Temp        -0.27793    0.02518  -11.04 1.05e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3548 on 28 degrees of freedom
Multiple R-squared:  0.8131,    Adjusted R-squared:  0.8064 
F-statistic: 121.8 on 1 and 28 DF,  p-value: 1.046e-11
```

Model before

- summary(gasB)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.85383    0.11842   57.88   <2e-16 ***
Temp        -0.39324    0.01959  -20.08   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2813 on 24 degrees of freedom
Multiple R-squared:  0.9438,    Adjusted R-squared:  0.9415
F-statistic: 403.1 on 1 and 24 DF,  p-value: < 2.2e-16
```

Fit them together

- gasBA = lm(Gas ~ Insul/Temp - 1, data = whiteside)
- summary(gasBA)

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
InsulBefore        6.85383    0.13596   50.41   <2e-16 ***
InsulAfter         4.72385    0.11810   40.00   <2e-16 ***
InsulBefore:Temp -0.39324    0.02249  -17.49   <2e-16 ***
InsulAfter:Temp  -0.27793    0.02292  -12.12   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.323 on 52 degrees of freedom
Multiple R-squared:  0.9946,	Adjusted R-squared:  0.9942
F-statistic:  2391 on 4 and 52 DF,  p-value: < 2.2e-16
```

- the parameter estimates are the same
- their standard errors are different because we are now estimating them jointly

Explain the model forumla

- lm(formula = Gas ~ Insul/Temp - 1, data = whiteside)
- the Insul/Temp: says fit a model of the form 1+Temp, separately for each level of Insul
- Insul has two levels (Before and After)
- the last term, -1, means do not fit an overall intercept
- we don’t need one in this case because there is a separate intercept for each level of Insul

Why would we do this?

- Why would we want to combine the two sets of observations?
- Mostly because, if they error terms are roughly similar then having more data improves our estimate of the standard error of the β's
- this improves our power and uses all of our data

Even more complicated

- gasBA2 = lm(Gas ~ Insul/(Temp + I(Temp^2)) - 1, data = whiteside)
- what do you think this means?
- summary(gasBA2)$coef

```
>   summary(gasBA2)$coef
                         Estimate  Std. Error   t value     Pr(>|t|)
InsulBefore           6.759215179 0.150786777 44.826312 4.854615e-42
InsulAfter            4.496373920 0.160667904 27.985514 3.302572e-32
InsulBefore:Temp     -0.317658735 0.062965170 -5.044991 6.362323e-06
InsulAfter:Temp      -0.137901603 0.073058019 -1.887563 6.489554e-02
InsulBefore:I(Temp^2) -0.008472572 0.006624737 -1.278930 2.068259e-01
InsulAfter:I(Temp^2)  -0.014979455 0.007447107 -2.011446 4.968398e-02
```

Things to notice

- when we added the terms Temp^2 to the model we could test for linearity
- which we did not see – and indeed we lost the effects for Temp altogether
- Why?
- Collinearity and its effects

Linear Models and Collinearity

- the easiest models to interpret are those where the columns of X are orthogonal to each other
- in that case the estimate of β_i does not change depending on which other variables are in the model
- but this is seldom ever true
- when the columns of X are related to each other, we say they are collinear

Collinearity Example

- BPdat= read.delim("BPex.txt")
  - measure blood pressure (BP), Age, Weight, body surface area (BSA), …
- cor(BPdat)

	BP	Age	Weight	BSA	Dur	Pulse	Stress
BP	1.0000000	0.6590930	0.95006765	0.86587887	0.2928336	0.7214132	0.16390139
Age	0.6590930	1.0000000	0.40734926	0.37845460	0.3437921	0.6187643	0.36822369
Weight	0.9500677	0.4073493	1.00000000	0.87530481	0.2006496	0.6593399	0.03435475
BSA	0.8658789	0.3784546	0.87530481	1.00000000	0.1305400	0.4648188	0.01844634
Dur	0.2928336	0.3437921	0.20064959	0.13054001	1.0000000	0.4015144	0.31163982
Pulse	0.7214132	0.6187643	0.65933987	0.46481881	0.4015144	1.0000000	0.50631008
Stress	0.1639014	0.3682237	0.03435475	0.01844634	0.3116398	0.5063101	1.00000000

- BPW = lm(BP ~ Weight, data=BPdat)
- BPBSA = lm(BP ~ BSA, data=BPdat)
- BPboth = lm(BP ~Weight + BSA, data=BPdat)

What happens

```
> summary(BPW)$coef
              Estimate Std. Error    t value     Pr(>|t|)
(Intercept) 2.205305 8.66333119  0.2545563 8.019513e-01
Weight      1.200931 0.09297008 12.9173953 1.527885e-10
> summary(BPBSA)$coef
              Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 45.18326   9.391857 4.810897 1.400279e-04
BSA         34.44281   4.690245 7.343499 8.114254e-07
> summary(BPboth)$coef
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 5.653398  9.3924833 0.6019067 5.551796e-01
Weight      1.038734  0.1926583 5.3915869 4.870718e-05
BSA         5.831250  6.0626938 0.9618250 3.496199e-01
>
```

- the estimates depend on what variables are in the model
- BSA is hard to interpret

A medical example

- suppose we are interested in different measures of cholesterol in humans
- LDL, HDL and Triglycerides are all measured and important
- but they are correlated in most healthy individuals
- therefore it seldom makes sense to talk about a one unit change in LDL holding HDL constant.

Good sources

- https://onlinecourses.science.psu.edu/stat501/node/2/
- has very good lessons and examples