CSAMA 2024 lecture

regression analysis with application to genomic data science

Vince Carey

Objectives

- form small groups, discuss and submit a handout related to deficiencies of reliance on statistical summaries (15 min, +5 min discussion)
- understand the informal equation Data = Fit + Residual
- formalism: variables and indices for data elements, model stipulation, parameters and their estimates
- distinguishing multivariate and conditional models
- principles of estimation: OLS, LS, WLS, GLS
- figures of merit encountered with regression models
    - Yes it is a lot of material, but the concepts will re-emerge frequently through the course

Forming groups

- Student that is in frontmost row, furthest to speaker's left is "1"
- student to left of "1" is "2", to left of "2" is "3", …
- continue in a snaking manner up to "13"
- next student is "1"
- when all students have numbers, please form groups of like-numbered students and receive the handout
- identify a group leader
- discuss the handout, write responses on sheet in ~10 minutes of discussion

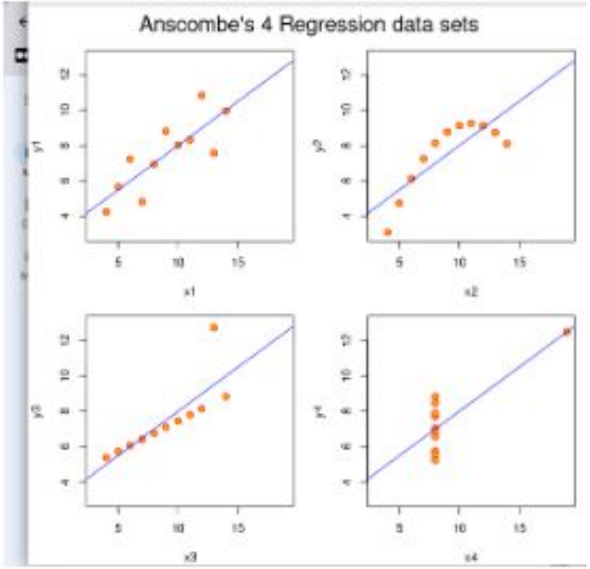
Discuss

CSAMA 2024 Regression session – Group commentary

1. Our definition of "statistical analysis" is:

2. Two key roles statistical analysis has played in work done by the group.

3. Briefly describe in lay terms your view of relationships implied by the graphs below



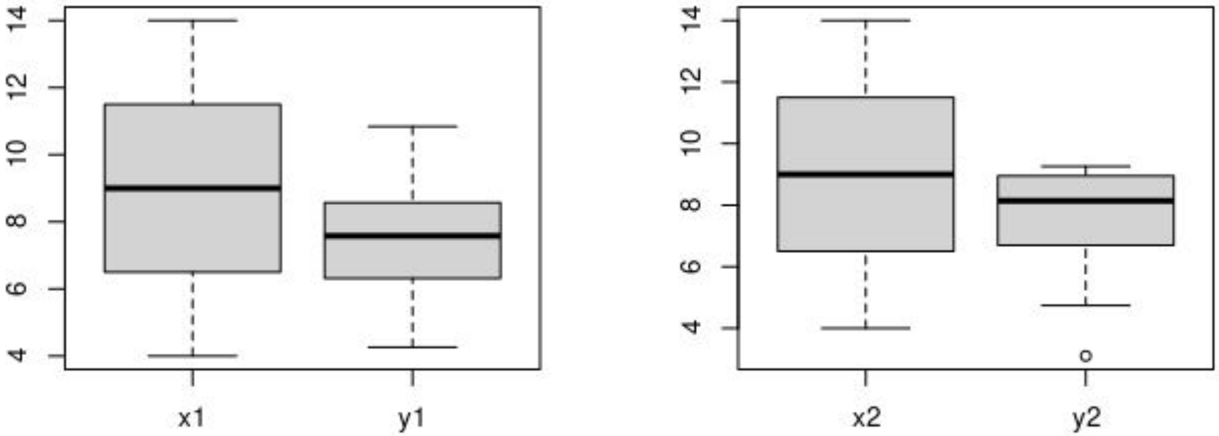
y1:x1

y2:x2

y3:x3

y4:x4

purpose of anscombe's dataset



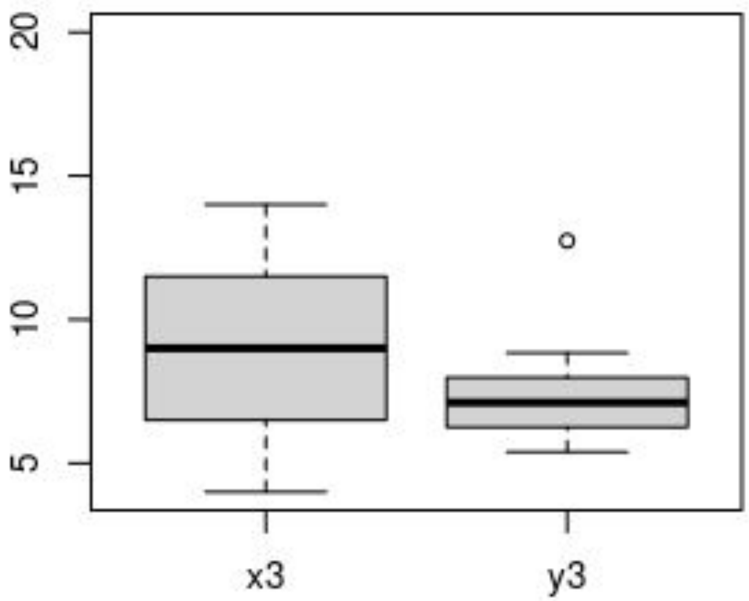
regarding the variables as separate samples to be compared in terms of average values of reponses (labeled x,y), we find identical results for manifestly different data configurations. **How would you do the comparisons? What are the manifest differences in data configurations?**

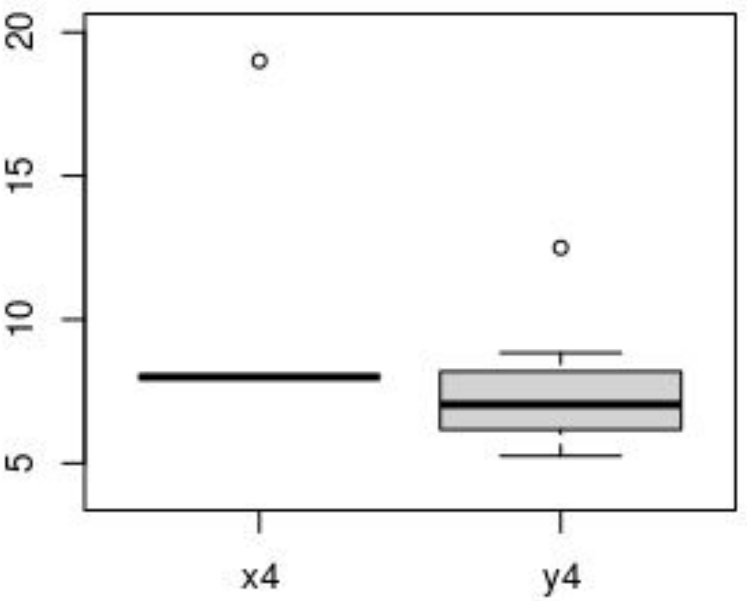
One way of comparing x1,y1, and then x2,y2

```
     Welch Two Sample t-test

data:  x1 and y1
t = 1.2783, df = 16.578, p-value = 0.2187
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.9798783  3.9780601
sample estimates:
mean of x mean of y 
 9.000000  7.500909 

> with(anscombe, t.test(x2,y2))

     Welch Two Sample t-test

data:  x2 and y2
t = 1.2783, df = 16.578, p-value = 0.2187
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.9799027  3.9780845
sample estimates:
mean of x mean of y 
 9.000000  7.500909 
```

these configurations also have identical t test outputs

to 3 decimal places





Upshots

- Data visualization is an important complement to statistical summary reports
- Verify on your own:
  - standard correlation tests have identical results for all four pairs of variables
  - summaries with lm(y1~x1, data=anscombe), etc. have identical results

Let's develop some formalism so we can be precise about our interpretations of the anscombe data configurations

- Basic rubric: Data = Fit + Residual (Daryl Pregibon)
  - "Data" is a collection of numbers or codes with metadata
  - "Fit" is a structured or simplified presentation of the process giving rise to the data
  - "Residual" is the discrepancy between reality (data) and the idealized "fit"
  - Our models and methods must address all of these components

Data = Fit + Residual

Simplest model

$$y_i = \mu + e_i, \;\; i = 1, \ldots, N$$

$$e_i \sim_{iid} N(0, \sigma^2)$$

The disturbances e_i are statistically independent and identically distributed

Estimation; principle of least squares

$$y_i = \mu + e_i,\ \ i = 1, \ldots, N$$

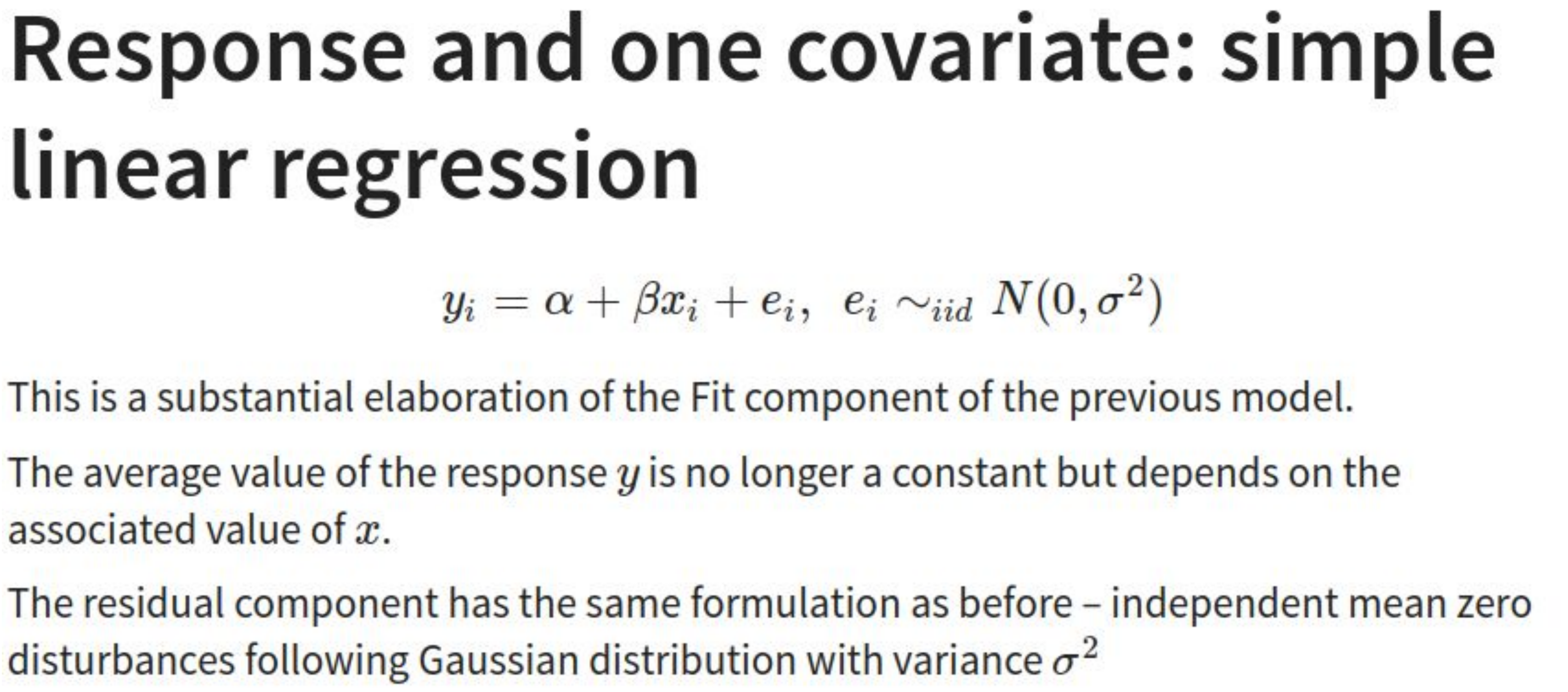
Write the “loss function”

$$L(\mu) = \sum_i (y_i - \mu)^2$$

Use calculus to show that

$$\hat{\mu} = \sum y_i / N$$

is the minimizer.



Response and one covariate: simple linear regression

$$y_i = \alpha + \beta x_i + e_i, \;\; e_i \sim_{iid} N(0, \sigma^2)$$

This is a substantial elaboration of the Fit component of the previous model.

The average value of the response y is no longer a constant but depends on the associated value of x.

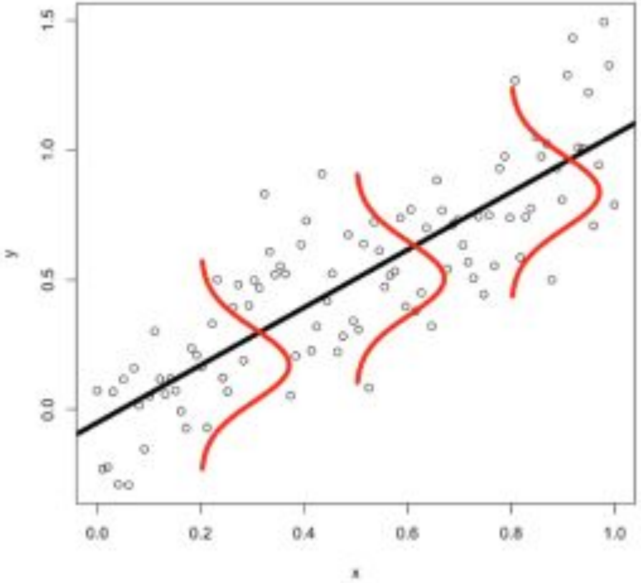
The residual component has the same formulation as before – independent mean zero disturbances following Gaussian distribution with variance σ^2

Critical distinction to be made: If we regard (y,x) as a bivariate response, correlation analysis is relevant, treating the variables symmetrically. If we regard y as a univariate response and x as a fixed covariate, then we use regression to model the conditional mean of y given x.

Assumptions

(Robert G's slide)

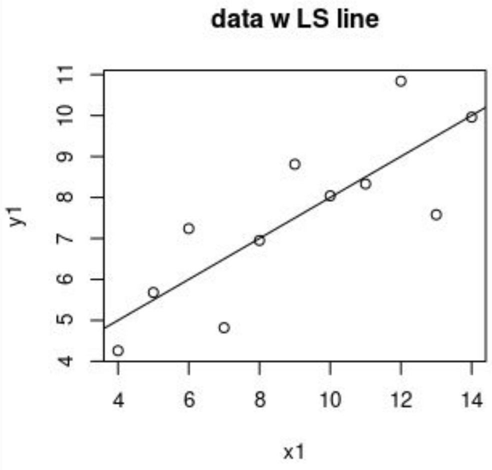
- For a specified value of x, the distribution of the y values is normal with mean $y = \alpha + \beta x$ and standard deviation σ

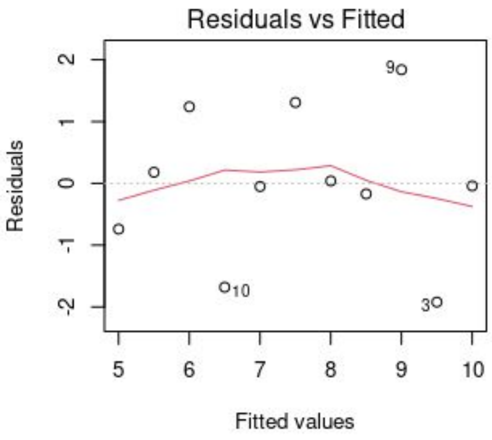


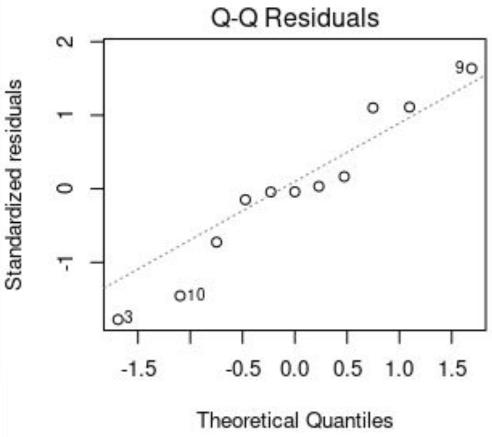
- For any specified value of x, σ is constant
- This assumption of constant variability across all values of x is known as **homoscedasticity**

We use m1 = lm(y1~x1, data= anscombe) to estimate intercept and slope for the fitted line. abline(m1) will plot the estimated line over the data.

plot(m1) has arguments to produce diagnostic plots for any linear regression.


data w LS line

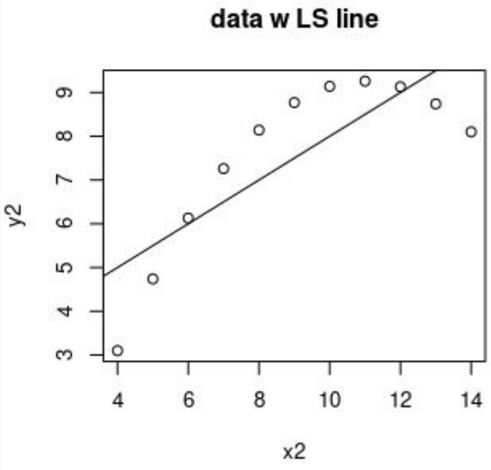

Residuals vs Fitted

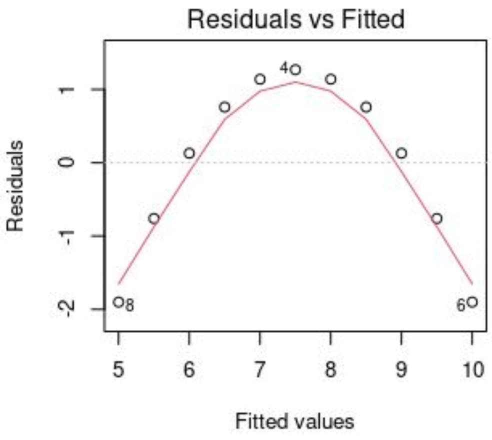

Q-Q Residuals

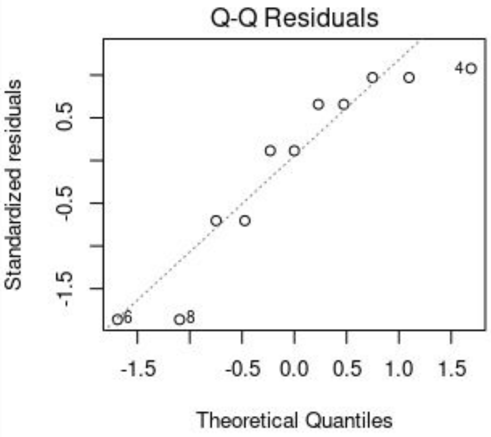
Q-Q plots are used to assess suitability of parametric models such as Gaussian model for residuals

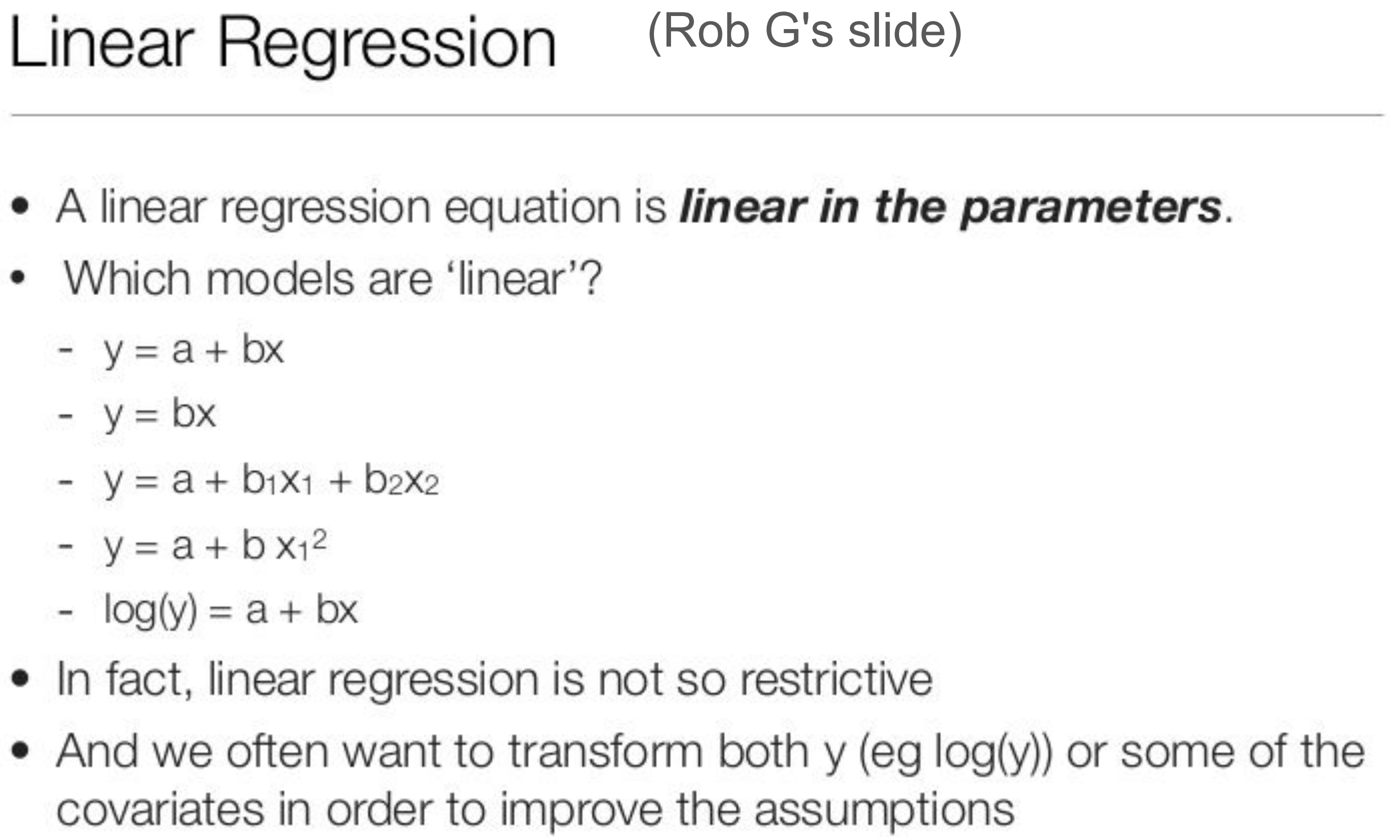
For x2, y2
the simple linear regression fits very poorly. Residuals show substantial "structure"

plot(y2~x2,data=anscombe, main="data w LS line"); abline(lm(y2~x2,data=anscombe));plot(m2,which=1); plot(m2, which=2)









Linear Regression

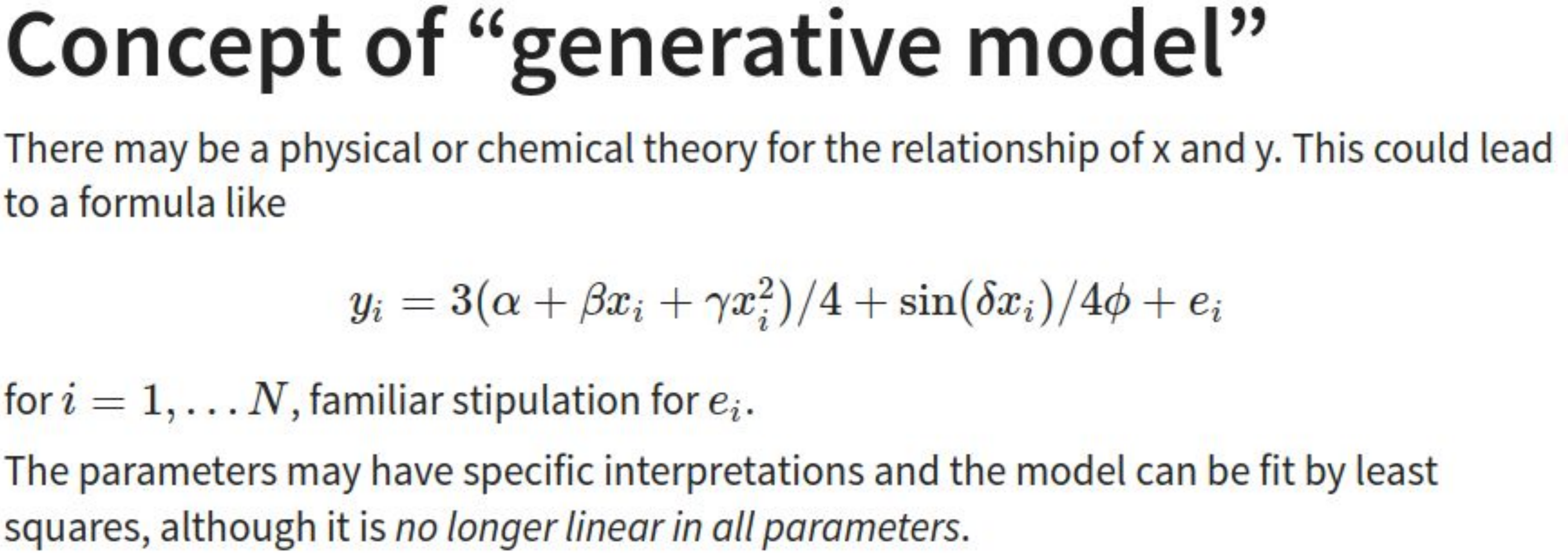
(Rob G's slide)

- A linear regression equation is ***linear in the parameters***.
- Which models are 'linear'?
  - $y = a + bx$
  - $y = bx$
  - $y = a + b_1x_1 + b_2x_2$
  - $y = a + b\,x_1^2$
  - $\log(y) = a + bx$
- In fact, linear regression is not so restrictive
- And we often want to transform both y (eg log(y)) or some of the covariates in order to improve the assumptions

^

[our capacity to be faithful to]

This slide is a "critique" of the proposition that y2 depends "quadratically" on x2 … and also of the corresponding proposition about linear dependence of y1 on x1



Concept of “generative model”

There may be a physical or chemical theory for the relationship of x and y. This could lead to a formula like

$$y_i = 3(\alpha + \beta x_i + \gamma x_i^2)/4 + \sin(\delta x_i)/4\phi + e_i$$

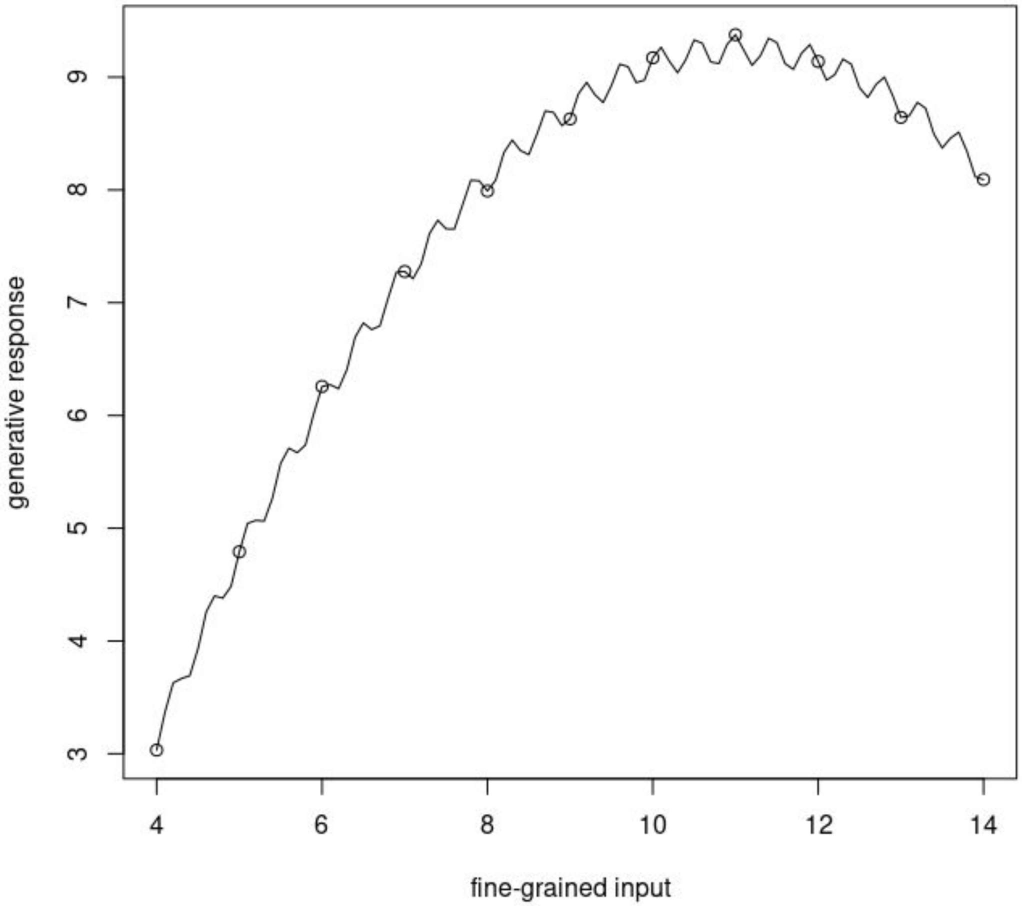
for $i = 1, \ldots N$, familiar stipulation for e_i.

The parameters may have specific interpretations and the model can be fit by least squares, although it is *no longer linear in all parameters*.

A quadratic regression
lm(y2~poly(x2,2),data=
anscombe)
fits the available data well,
but a rather different reality
might produce the data, and
would only be detectable
with finer-grained sampling
on x

A similar situation emerges
for y1, x1, because we
interpolate where no x is
observed

Care is warranted in all
applications of models

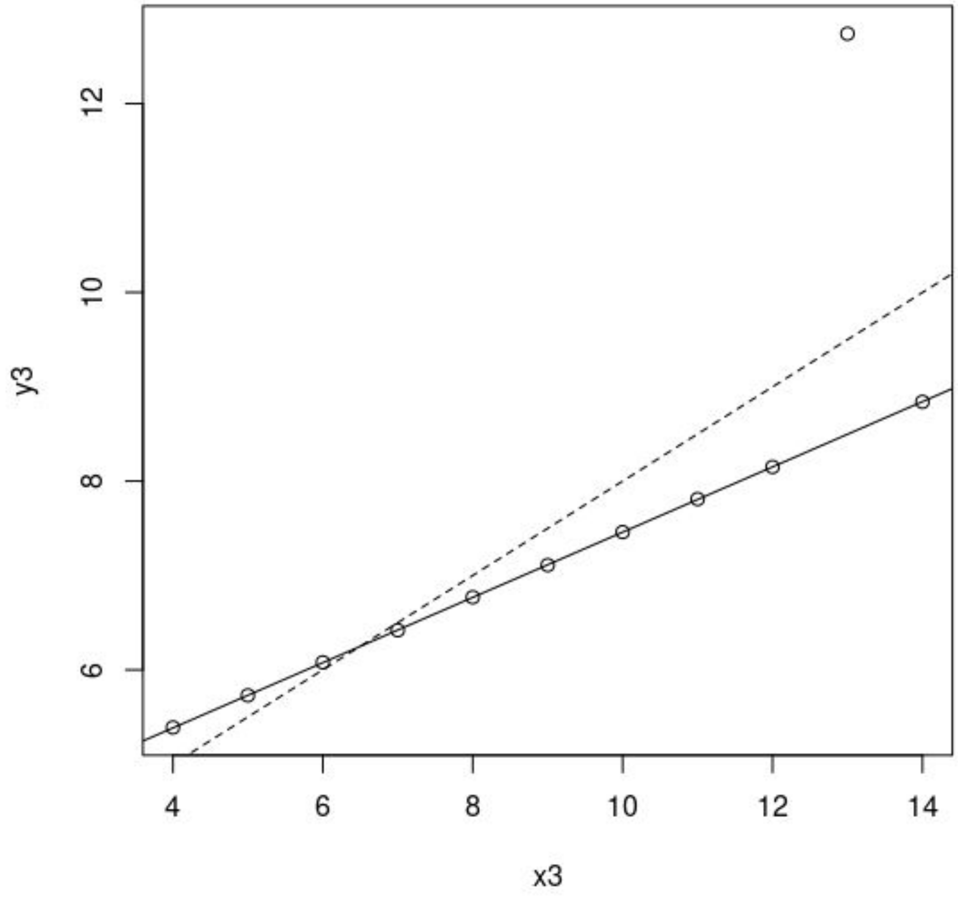


What about y3, x3?

If the value of y at x3==13 is anomalous, then a robust procedure can be used to produce a seemingly accurate estimate of the slope of the relationship

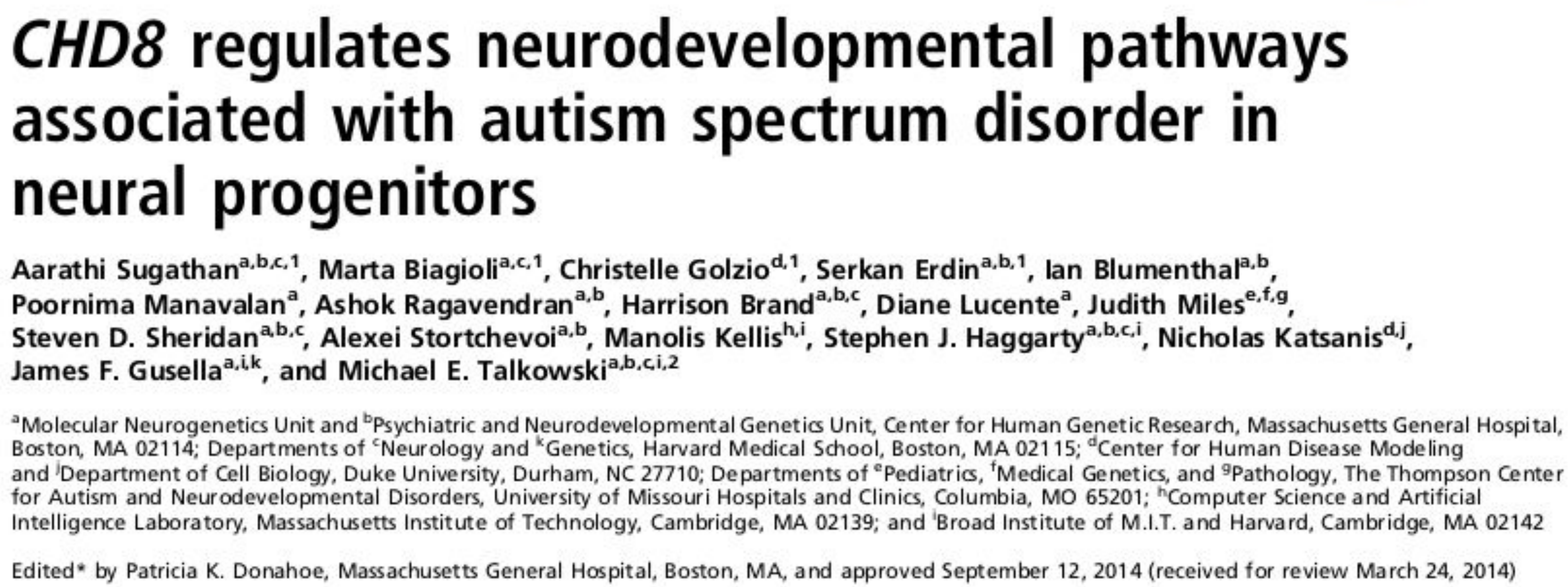
OR, the observation at x3==13 may be extremely important to explain – might be a sweet spot for producing more y3

always look, never accept the opaque report without checking



Final remarks: multiple regression and confounding

- The scale of data complexity in genomics has led to a tendency to use opaque "genome-wide" tools to report on possible signals
- There is no reason to avoid detailed exploration of possible relationships of interest in different ways to chase different theories
- We'll conclude with one example of multiple regression to explore effects of batch and shRNA species in a 2014 study of autism



CHD8 regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors

Aarathi Sugathan[a,b,c,1], Marta Biagioli[a,c,1], Christelle Golzio[d,1], Serkan Erdin[a,b,1], Ian Blumenthal[a,b], Poornima Manavalan[a], Ashok Ragavendran[a,b], Harrison Brand[a,b,c], Diane Lucente[a], Judith Miles[e,f,g], Steven D. Sheridan[a,b,c], Alexei Stortchevoi[a,b], Manolis Kellis[h,i], Stephen J. Haggarty[a,b,c,i], Nicholas Katsanis[d,j], James F. Gusella[a,i,k], and Michael E. Talkowski[a,b,c,i,2]

[a]Molecular Neurogenetics Unit and [b]Psychiatric and Neurodevelopmental Genetics Unit, Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114; Departments of [c]Neurology and [k]Genetics, Harvard Medical School, Boston, MA 02115; [d]Center for Human Disease Modeling and [j]Department of Cell Biology, Duke University, Durham, NC 27710; Departments of [e]Pediatrics, [f]Medical Genetics, and [g]Pathology, The Thompson Center for Autism and Neurodevelopmental Disorders, University of Missouri Hospitals and Clinics, Columbia, MO 65201; [h]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139; and [i]Broad Institute of M.I.T. and Harvard, Cambridge, MA 02142

Edited* by Patricia K. Donahoe, Massachusetts General Hospital, Boston, MA, and approved September 12, 2014 (received for review March 24, 2014)

"CHD8 is an ATP-dependent chromatin remodeler of the SNF2 family (8). CHD8 was identified as one of the genes in the minimal region of overlap of de novo 14q11.2 microdeletions in two children with developmental delay and cognitive impairment (9). We previously detected direct disruption of CHD8 by a de novo balanced translocation, with concomitantly reduced mRNA expression, in a patient diagnosed with ASD, intellectual disability, …"

Bulk RNA-seq was used with shRNA knockdown of CHD3. To get an instance of the data with Bioconductor recount3

```
autrse = recount3::create_rse_manual(
    project = "SRP047233",
    project_home = "data_sources/sra",
    organism = "human",
    annotation = "gencode_v26",
    type = "gene"
)
```

We'd like to explore the data patterns underlying these findings:

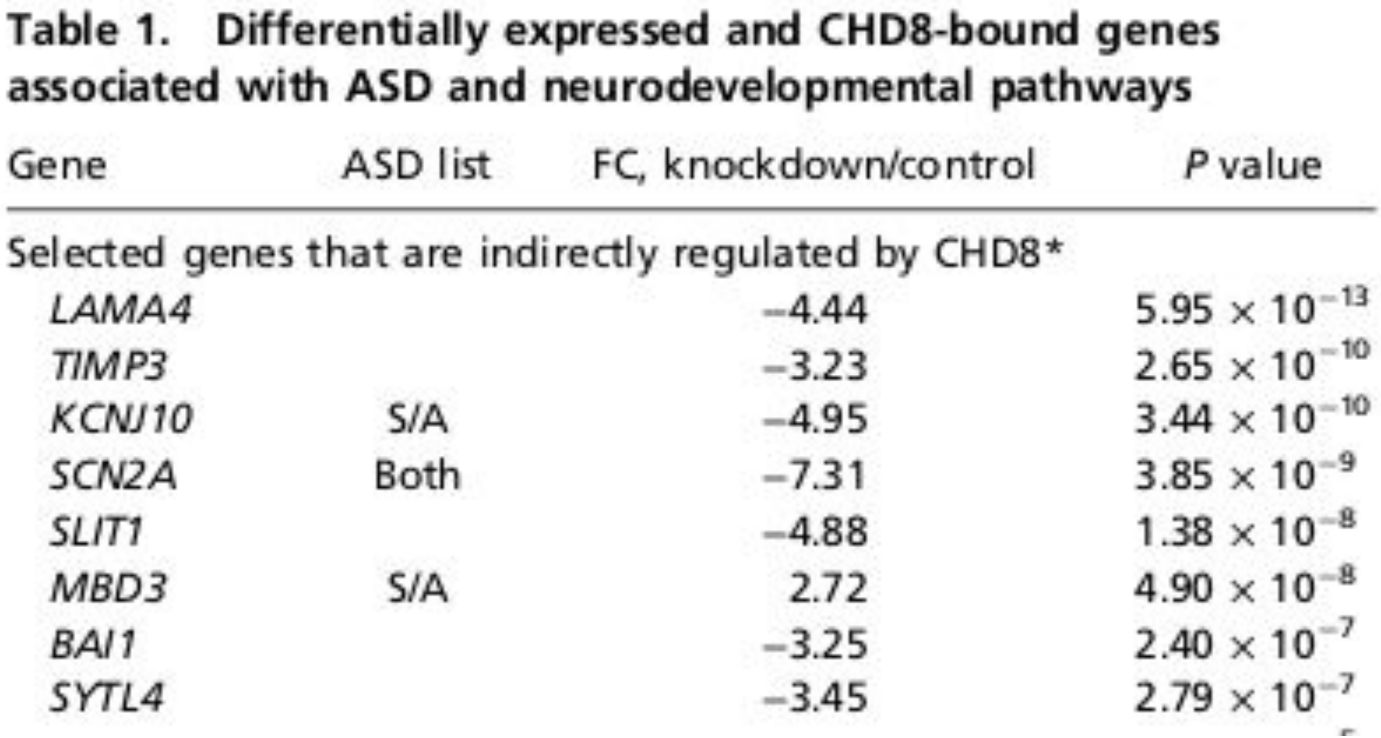


Table 1. Differentially expressed and CHD8-bound genes associated with ASD and neurodevelopmental pathways

Gene	ASD list	FC, knockdown/control	*P* value
Selected genes that are indirectly regulated by CHD8*			
LAMA4		–4.44	5.95×10^{-13}
TIMP3		–3.23	2.65×10^{-10}
KCNJ10	S/A	–4.95	3.44×10^{-10}
SCN2A	Both	–7.31	3.85×10^{-9}
SLIT1		–4.88	1.38×10^{-8}
MBD3	S/A	2.72	4.90×10^{-8}
BAI1		–3.25	2.40×10^{-7}
SYTL4		–3.45	2.79×10^{-7}

```
> summary(lm(log(x[52569,]+1)~x[20762,]))

Call:
lm(formula = log(x[52569, ] + 1) ~ x[20762, ])

Residuals:
     Min       1Q   Median       3Q      Max 
-1.16858 -0.44534  0.08686  0.47548  1.06017 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.010e+01  1.461e-01  69.115  < 2e-16 ***
x[20762, ]  5.042e-06  6.486e-07   7.773 2.92e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6188 on 52 degrees of freedom
Multiple R-squared:  0.5374,  Adjusted R-squared:  0.5285 
F-statistic: 60.41 on 1 and 52 DF,  p-value: 2.917e-10
```

Estimate effect of level of CHD8 on mean of LAMA4

Sign seems right: knockdown of CDH8 leads to reduction in LAMA4 expression

```
> summary(lm(log(x[52569,]+1)~x[20762,]+tt+bat))

Call:
lm(formula = log(x[52569, ] + 1) ~ x[20762, ] + tt + bat)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.43029 -0.20244 -0.03879  0.12816  0.67864 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              1.096e+01  1.336e-01  82.021  < 2e-16 ***
x[20762, ]               3.239e-06  5.651e-07   5.731 8.37e-07 ***
ttCHD8 (TRCN0000016511) -1.470e-01  2.334e-01  -0.630 0.532005    
ttCHD8 (TRCN0000016512) -1.489e+00  2.339e-01  -6.367 9.73e-08 ***
ttCHD8 (TRCN0000360108) -7.771e-01  1.678e-01  -4.631 3.23e-05 ***
ttCHD8 (TRCN0000360109) -9.418e-01  1.837e-01  -5.126 6.36e-06 ***
ttCHD8 (TRCN0000367896) -9.409e-01  2.326e-01  -4.045 0.000208 ***
ttGFP (TRCN0000072181)   1.763e-01  2.090e-01   0.844 0.403328    
ttLacZ (TRCN0000072236)  4.704e-02  1.931e-01   0.244 0.808694    
batbatch;;2             -2.991e-01  1.290e-01  -2.319 0.025121 *  
```

Direct adjustment for batch and shRNA species attenuates effect size and degree of significance

Exercise: are modeling assumptions reasonable for this redo?

Additionally: when batch and shRNA species are included as predictors we have

```
Residual standard error: 0.3113 on 44 degrees of freedom
Multiple R-squared:  0.901,	Adjusted R-squared:  0.8807
F-statistic: 44.47 on 9 and 44 DF,  p-value: < 2.2e-16
```

For the simple model:

```
Residual standard error: 0.6188 on 52 degrees of freedom
Multiple R-squared:  0.5374,	Adjusted R-squared:  0.5285
F-statistic: 60.41 on 1 and 52 DF,  p-value: 2.917e-10
```

```
Analysis of Variance Table

Model 1: log(x[52569, ] + 1) ~ x[20762, ]
Model 2: log(x[52569, ] + 1) ~ x[20762, ] + tt + bat
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     52 19.911
2     44  4.263  8    15.648 20.189 2.172e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1
```

We made a more complex model by introducing batch and shRNA species and so we must pay a penalty. The F-test from anova() suggests the increased complexity is worthwhile.

Conclusions

- The anscombe dataset is not just about the value of visualization. Various problems with opacity of statistical reports and models are exposed
  - simple statistics can hide meaningful patterns
  - model "exceptions" may be important
  - whereof we have no data, thereof we might need to be silent
- In many situations Data = Fit + Residual is a useful rubric
  - Residual is a "structureless" quantity defying explanation apart from scale of variability
  - When structure is found, elaborate the model to remove it, but pay a price for added complexity
- Principle of Least Squares is useful for a variety of linear modeling problems
- "Control" of additional factors like batch and experimental factor details can be explored with multiple regression