STAT588/BIOL588: Genomic Data Science
Lecture 6: Review Statistics (Part II)

Dr. Yen-Yi Ho (hoyen@stat.sc.edu)

Objectives of Lecture 6

- Association between Variables
  - Goodness of Fit Test
  - Pearson χ^2 Test of Association
  - Relative Risk
  - Odds Ratio
- Statistical Models
- Linear Regression
- Multiple Linear Regression
- Interaction
- Likelihood Ratio Test for Model Seletion
- Logistic Regression

Association between Variables

		Independent Categorical	Variable Continuous
Outcome Variable	Continuous	T-Test, ANOVA (A)	Regression (C)
	Categorical	χ^2, Fisher (B)	GLM (D)

- Difference in gene expression in patients with/without a mutation (yes/no): A
- Determine the association between disease Status (yes/no) and genotype (AA, Aa, aa): B
- Predict father's height from daughter's height: C
- Determine the relationship between smoking status (yes/no) and lung cancer (yes/no): B

Goodness of Fit Test

	Count
AA	30
Aa	55
aa	15
Total	100

- What is the allele frequency of A allele?

Goodness of Fit Test

	Count
AA	30
Aa	55
aa	15
Total	100

- What is the allele frequency of A allele?

$$p(A) = \frac{30 \times 2 + 55}{2 \times (30 + 55 + 15)} = 0.575$$

- What is the expected counts if this locus is in Hardy-Weinberg equilibrium?

Goodness of Fit Test

	Count	Expected	$\frac{(O_i-E_i)^2}{E_i}$
AA	30	$100 \times 0.575^2 = 33$	0.28
Aa	55	$100 \times 2 \times 0.575 \times 0.425 = 49$	0.77
aa	15	18	0.52
Total	100	100	1.57

- What is the expected counts if this locus is in Hardy-Weinberg equilibrium?

$$\chi^2 = \Sigma_i \frac{(O_i - E_i)^2}{E_i} = 1.57 < \chi^2_{1,0.95} = 3.84$$

Since $\chi^2 = 1.57 < 3.841$, we conclude that the genotype frequencies in this population are not significantly different than what would be expected if the population is in Hardy-Weinberg equilibrium.

Assumptions for Hardy-Weinberg Equilibrium

- Random Mating
- No Nature Selection: neither allele confers a selective advantage or disadvantage
- No Migration: no one enters or leaves the population
- No Mutation: an A allele will never mutate into an a allele, and vice versa
- Infinite Population size: no genetic drift

Pearson χ^2 Test of Association

FAMuSS Data Example

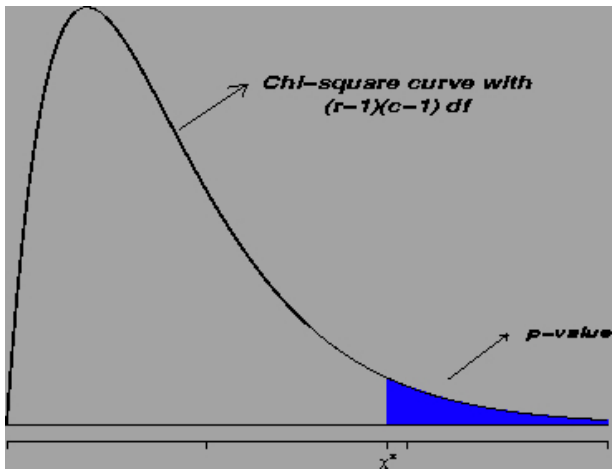
	Genotype			
BMI $>$ 25	AA	GA	GG	Total
0	30	246	380	656
1	30	130	184	344
Total	60	376	564	1000

Test of Association

Hypothesis: no association between genotype and disease

$$\chi^2 = \Sigma_{allcells} \frac{(oberseved - expected)^2}{expected}$$

$$\text{p value} = \Pr(\chi^2_{df} > \chi^2_{obs})$$



$\rightarrow$ If p value is small, reject H_0 Hypothesis.

Expected Cell Count

Observed				
	Genotype			
	AA	GA	GG	Total
0	30	246	380	656
1	30	130	184	344
Total	60	376	564	1000

Expected				
	Genotype			
	AA	GA	GG	Total
0	$1000 \times 0.656 \times 0.06$			656
1				344
Total	60	376	564	1000

Degree of freedom

Pearson's χ^2 test for association

	Observed		
	Genotype		
	AA	GA	GG
0	30	246	380
1	30	130	184

	Expected		
	Genotype		
	AA	GA	GG
0	39.36	246.66	369.98
1	20.64	129.34	194.02

$$\chi^2_{obs} = \frac{(30-39.36)^2}{39.36} + \frac{(246-246.66)^2}{246.66} + \frac{(380-369.998)^2}{369.98}$$
$$+ \frac{(30-20.64)^2}{20.64} + \frac{(130-129.34)^2}{129.34} + \frac{(184-194.02)^2}{194.02} \approx 7.26$$

```
>tab< −matrix(c(30, 30, 246, 130, 380, 184), nrow=2)
>chisq.test(tab)
```

Pearson's Chi-squared test

data: tab, X-squared = 7.2638, df = 2, p-value = 0.02647

Relative Risk

	Smoker	Nonsmoker
Cancer	89	37
Normal	6063	5711

- $p_1 = \Pr(\text{Cancer}|\text{Smoker})$
- $\hat{p}_1 - \hat{p}_2 = 0.0145 - 0.00644 = 0.008$.
- Relative Risk$=\frac{\hat{p}_1}{\hat{p}_2} = \frac{0.0145}{0.00644} = 2.25$. The probability of cancer is 2.25 times greater in smokers.
- To estimate p_1, p_2, we need to follow up many smokers and nonsmokers in a prospective study.

Relative Risk

	Smoker	Nonsmoker
Cancer	89	37
Normal	6063	5711

- $p_1 = \Pr(\text{Cancer}|\text{Smoker})$
- $\hat{p}_1 - \hat{p}_2 = 0.0145 - 0.00644 = 0.008$.
- Relative Risk$=\frac{\hat{p}_1}{\hat{p}_2} = \frac{0.0145}{0.00644} = 2.25$. The probability of cancer is 2.25 times greater in smokers.
- To estimate p_1, p_2, we need to follow up many smokers and nonsmokers in a prospective study.
- In retrospective study, we can use odds ratio.

Odds

The odds in favor of an event are the ratio of the probability that the event will happen to the probability that it will not happen.

$$Odds = \frac{p}{1-p}$$

What does "3 to 1 odds the Gamecocks will win" mean?

Apollo 13

NASA Director : He specifically wanted a quote from a flight director.
Gene Kranz : Who wanted a quote?
Deke Slayton : The president.
Gene Kranz : The president?
Glynn Lunney : Nixon. He wants odds.
Gene Kranz : We are not losing the crew.
NASA Director : Gene, I gotta give him odds. Five to one against? Three to one?
Glynn Lunney : I don't think they're that good.
Gene Kranz : [firmly] We are not losing those men!

Odds ratio: Measuring Association

	Genotype	
BMI > 25	AA	(GA or GG)
1	a	c
0	b	d
	a+b	c+d

$$\text{Odds of disease among AA} = \frac{\Pr(D^+|E^+)}{[1-\Pr(D^+|E^+)]} = \frac{\frac{a}{(a+b)}}{\frac{b}{(a+b)}} = \frac{a}{b},$$

$$\text{Odds of disease among GA and GG} = \frac{\Pr(D^+|E^+)}{[1-\Pr(D^+|E^+)]} = \frac{\frac{c}{(c+d)}}{\frac{d}{(c+d)}} = \frac{c}{d}.$$

Odds ratio (OR)

	Genotype	
BMI > 25	AA	(GA and GG)
1	30	314
0	30	626
	60	940

$$\text{OR}_{\frac{\text{AA}}{\text{GA and GG}}} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc}$$

$$= \frac{30 \times 626}{30 \times 314} \approx 1.99$$

		Independent Variable	
		Categorical	Continuous
Outcome Variable	Continuous	T-Test, ANOVA (A)	Regression (C)
	Categorical	χ^2, Fisher (B)	GLM (D)

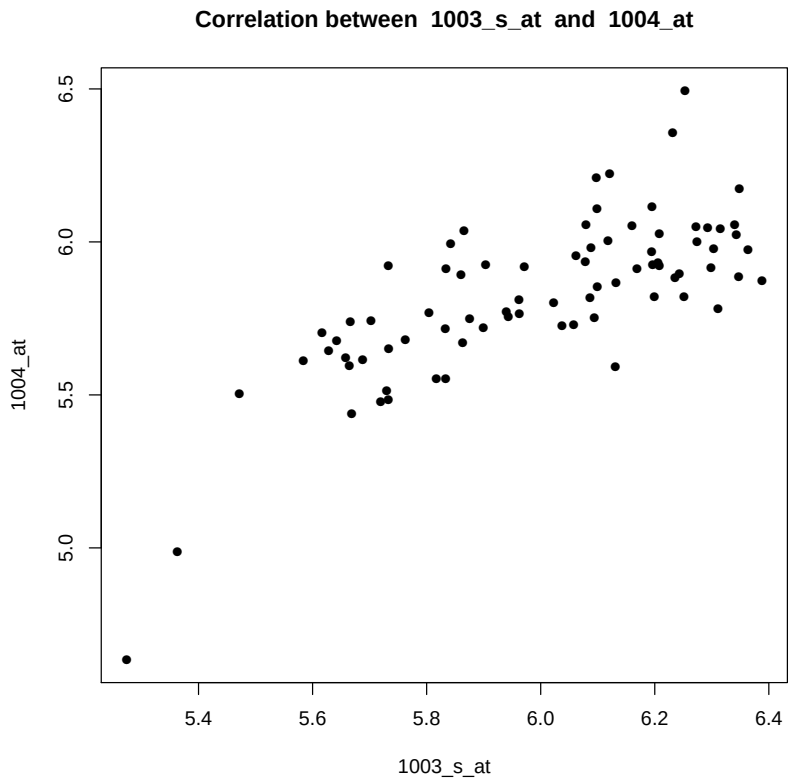
Statistical Models

Statistical models can be powerful tools for understanding complex relationship among variables. First, we will start by looking at 2 continuous variables. Typically, we explore data by a scatter plot.

Gene Expression Example

```
>library(``Biobase")
>library(``annotate")
>library(``hgu95av2.db")
>library(ALL)
>data<-exprs(ALL_bcrneg)
>probename<-rownames(data)
>genename<-mget(probename, hgu95av2SYMBOL)
>genename[1:5]
>plot(data[4,], data[5,], pch=16)
```

Correlation



Probe ("1003_s_at" and "1004_at") are mapped to the same gene (*CXCR5*), are their expression measures correlated?

Pearson Correlation

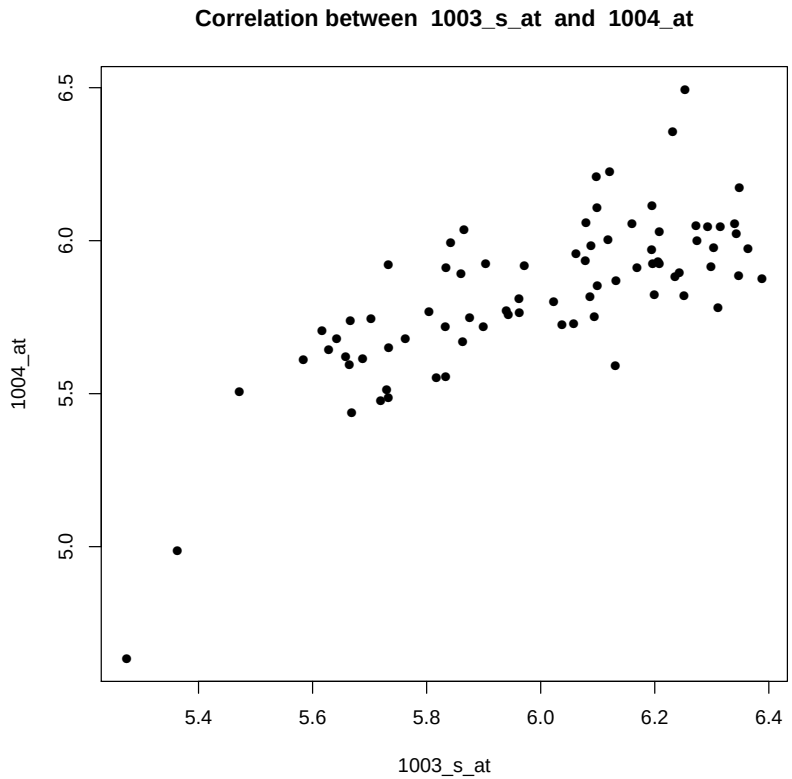
Consider n pairs of data: (x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_n, y_n)

$$r = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)s_x s_y}$$

s_x, s_y: SD of x and y.
This is sometimes also called the correlation coefficient;
$-1 \leq r \leq 1$.

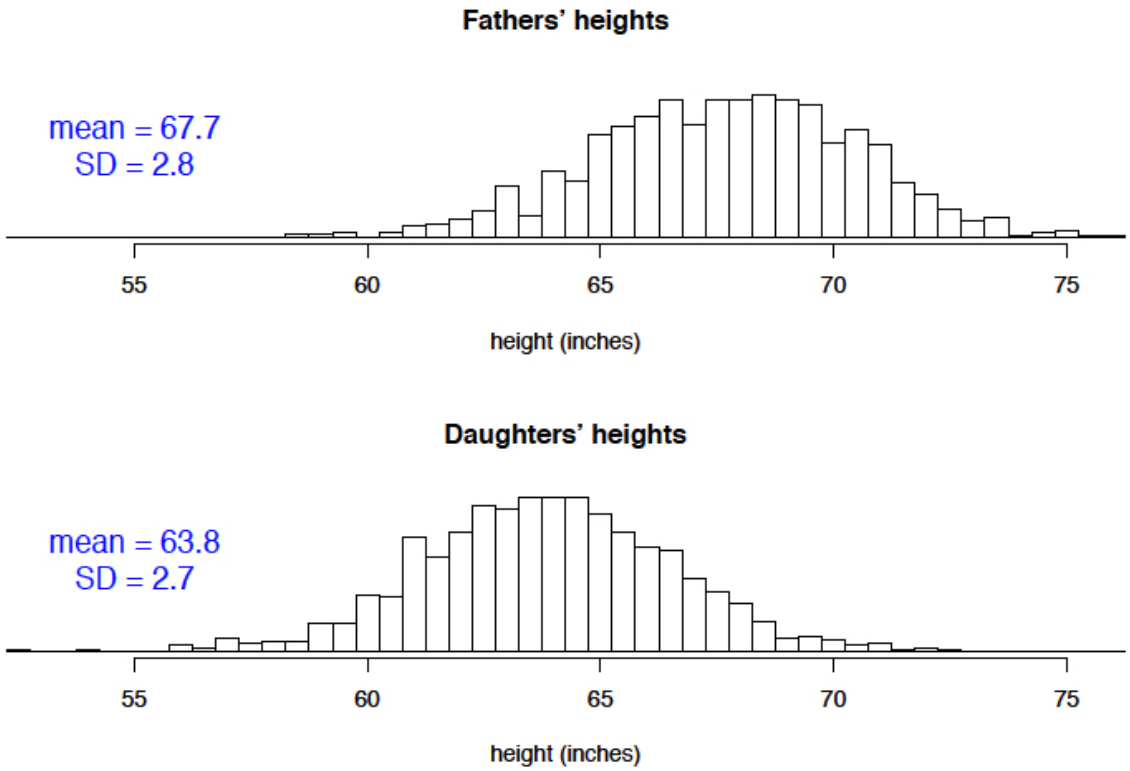
- r=0 : no correlation
- r > 0: positive correlation; Y increases with increasing X.
- r<0: negative correlation.
- $|r| > 0.7$, strong correlation
- $0.3 < |r| < 0.7$, moderate correlation
- $|r| < 0.3$, weak correlation

Gene Expression Example



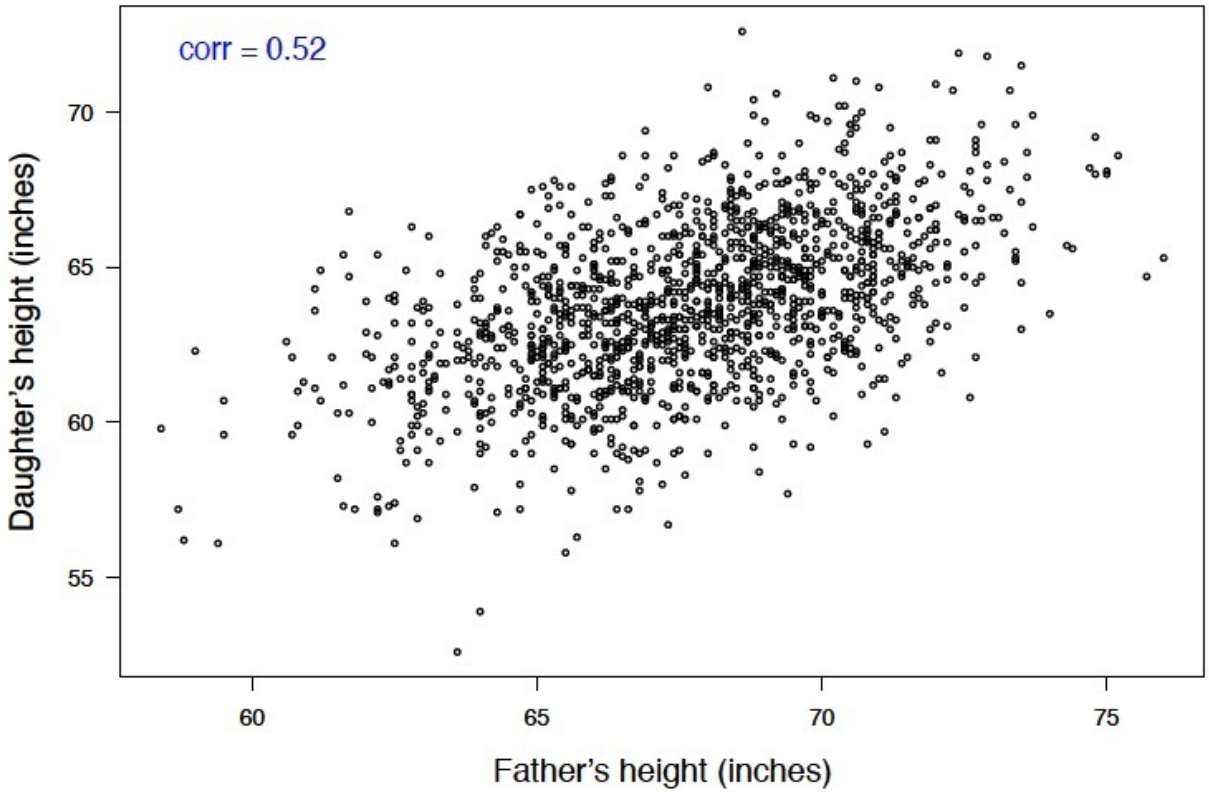

```
> cor(data[4,], data[5,])
[1] 0.7499144
```

Example 2: Fathers' and daughters' heights



Reference: Pearson and Lee (1906) Biometrika 2:357-462
1376 pairs

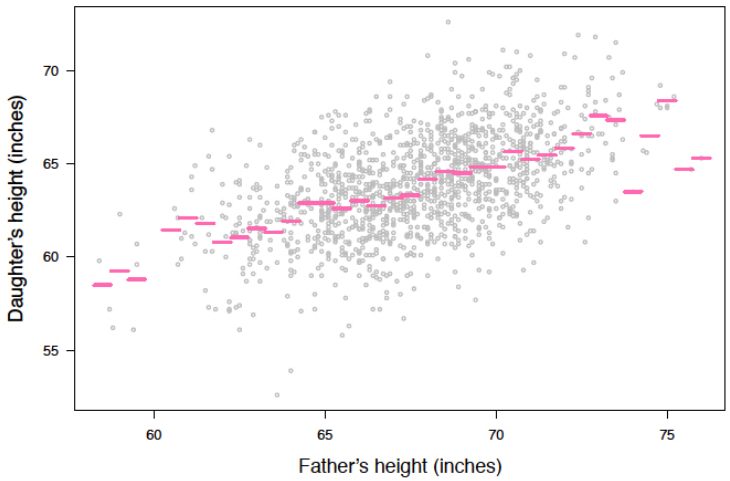
Fathers' and daughters' heights

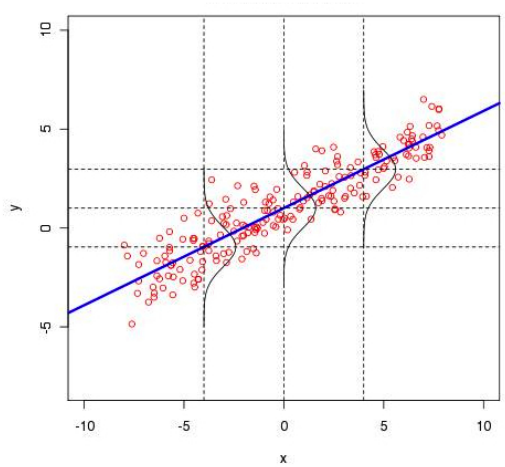


Reference: Pearson and Lee (1906) Biometrika 2:357-462
1376 pairs

Linear Regression

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$





The regression model

Let X be the predictor and Y be the response. Assume we have n observations $(x_1, y_1), \ldots, (x_n, y_n)$ from X and Y. THe simple linear regression model is

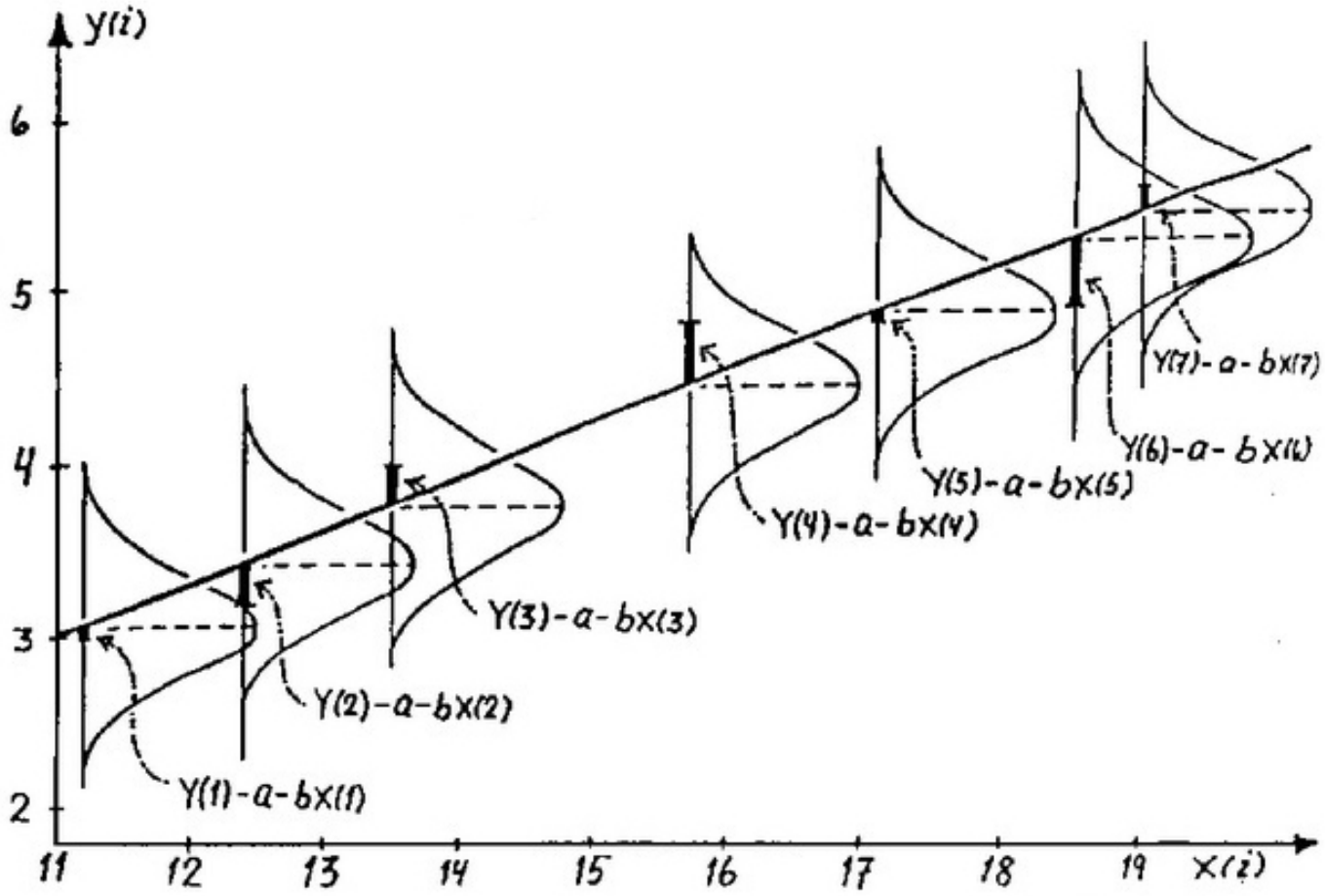
$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2), \\ &\text{or} \\ \hat{Y} &= \beta_0 + \beta_1 X. \end{aligned}$$

$\hat{Y}$ is the fitted value of Y.

$\rightarrow$ How do we decide the values β_0, β_1, and σ^2?

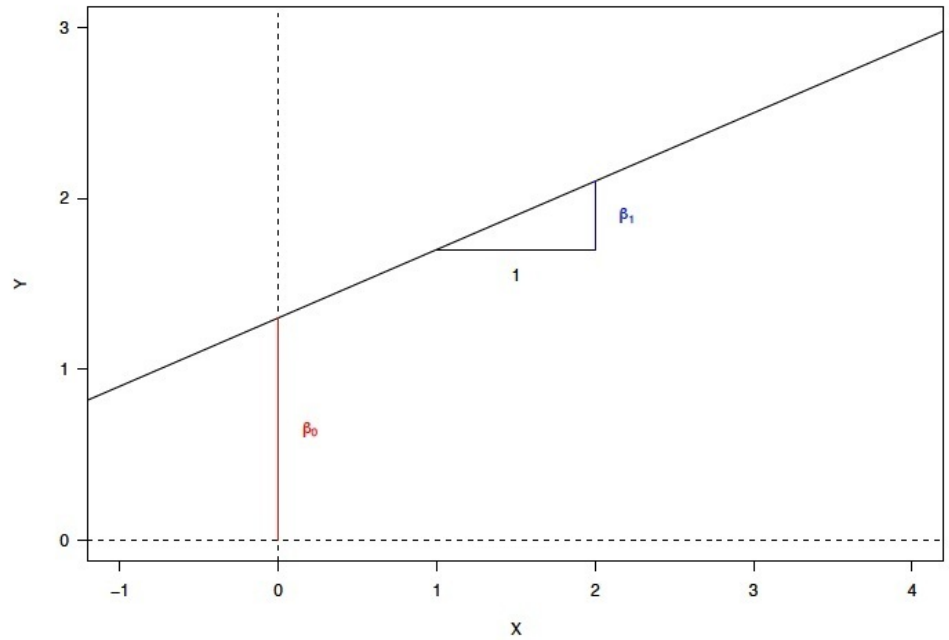
Residuals

$$\epsilon_i = y_i - (\beta_0 + \beta_1 x_i)$$



Regression Coefficients

$$\hat{Y} = \beta_0 + \beta_1 X$$



- β_1: the amount of change in y that occurs with on unit change in x.
- β_0: the fitted value of y when x=0.

Fitting Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X + \epsilon_i$$

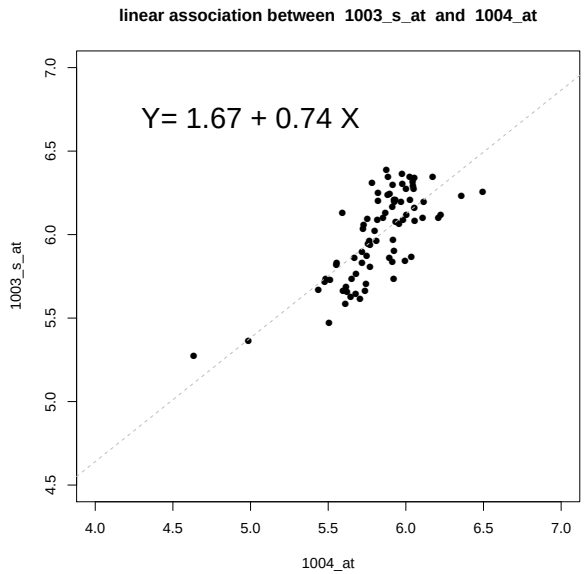
Data:

Obs	y	x
1	0.72	0.43
2	0.65	1.51
3	0.81	-0.63
4	-0.06	-0.73
5	1.39	0.27
6	-0.04	0.13
7	-0.09	0.65
8	-0.31	-0.83
9	0.85	-0.54
10	0.35	0.04
...		

```
fit<-lm(y ~ x)
```

Gene Expression Example

$$\hat{Y} = \beta_0 + \beta_1 X_1$$



$$H_0 : \beta_i = 0 \quad \text{vs} \quad H_a : \beta_i \neq 0$$

$$t = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

```
>fit2<-lm(data[4,] ~ data[5,])
>aa<-summary(fit2)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.6740	0.4348	3.85	0.0002
‘‘1004_at"	0.7416	0.0746	9.95	0.0000

Matrix Multiplication

$$x = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \times \begin{pmatrix} 9 \\ 8 \\ 7 \end{pmatrix}$$

$$1 \times 9 + 2 \times 8 + 3 \times 7 = 46$$

$$4 \times 9 + 5 \times 8 + 6 \times 7 = 118$$

$$x = \begin{pmatrix} 46 \\ 118 \end{pmatrix}$$

Dimension: $(2 \times 3) \times (3 \times 1) = (2 \times 1)$

Fitting Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Design Matrix

$$Y = X\beta + \epsilon$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Design Matrix

More than one predictor

Data

	y	x_1	z
1	0.72	0.37	0
2	0.65	0.19	0
3	0.81	0.11	0
4	-0.06	-0.44	0
5	1.39	-0.31	0
6	-0.04	-0.39	1
7	-0.09	-0.20	1
8	-0.31	-0.23	1
9	0.85	-0.01	1
10	0.35	-0.45	1
...			

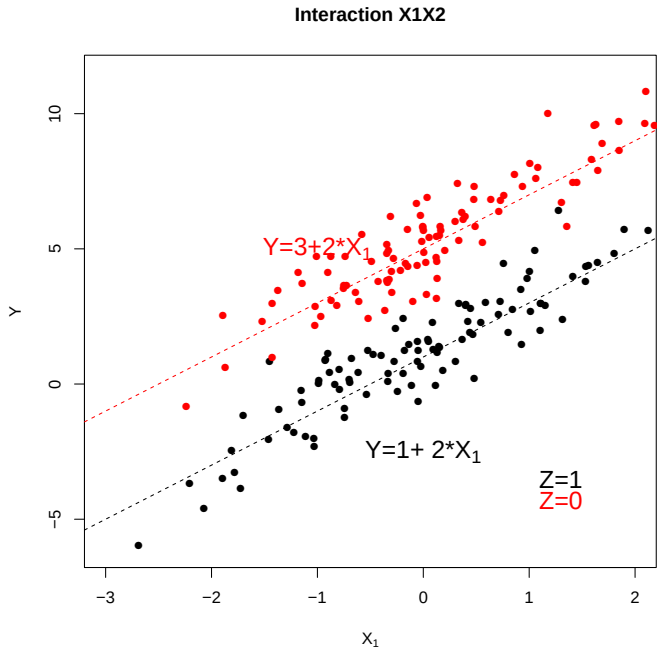
$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 Z + \epsilon_i$$

In other words (or, equations):

$$Y_i = \begin{cases} \beta_0 + \beta_1 X_1 + \epsilon_i, & \text{if } Z = 0 \\ (\beta_0 + \beta_2) + \beta_1 X_1 + \epsilon_i, & \text{if } Z = 1 \end{cases}$$

Multiple Linear Regression

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 Z + \epsilon_i$$



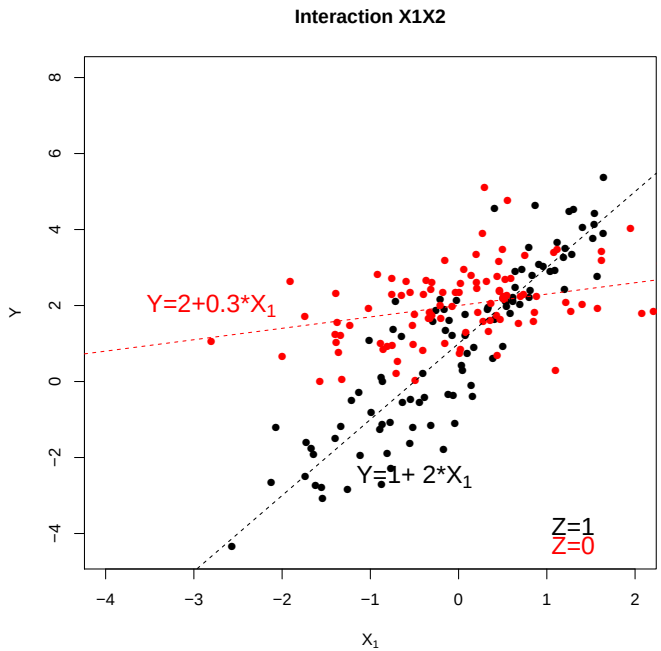
$$Y_i = \begin{cases} \beta_0 + \beta_1 X_1 + \epsilon_i, & \text{if } Z = 0 \\ (\beta_0 + \beta_2) + \beta_1 X_1 + \epsilon_i, & \text{if } Z = 1 \end{cases}$$

$\rightarrow$ Assuming the same slope for both $Z = 0$ and $Z = 1$.

Multiple Linear Regression: Interaction

When slopes are different in $Z = 0$ vs. $Z = 1$,

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 Z + \beta_3 X_1 \times Z + \epsilon_i$$



$$Y_i = \begin{cases} \beta_0 + \beta_1 X_1 + \epsilon_i, & \text{if } Z = 0 \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 + \epsilon_i, & \text{if } Z = 1 \end{cases}$$

Gene Expression Example

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 Z + \beta_3 X_1 \times Z + \epsilon_i$$

Y: measure of "1003_s_at" probe
X: measure of " 1004_at" probe
Z: molecular type (BCR/ABL=0 or NEG=1)

Intercept	X_1	Z	$X_1 \times Z$
1	5.93	0	0.00
1	5.91	1	5.91
1	5.89	0	0.00
1	5.62	1	5.62
1	5.92	1	5.92
...			

Table: Design Matrix

Gene Expression Example

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 Z + \beta_3 X_1 \times Z + \epsilon_i$$

Y: measure of "1003_s_at" probe
X: measure of " 1004_at" probe
Z: molecular type (BCR/ABL=1 or NEG=0)

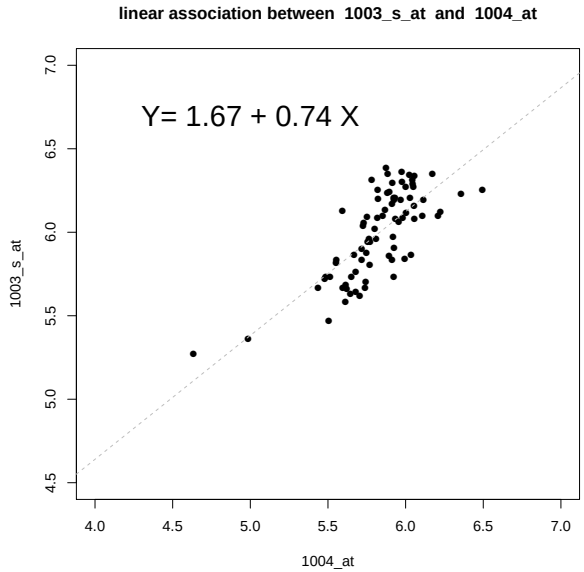
```
> int <- as.numeric(ALL_bcrneg$mol.biol) * data[5,]
> fit1<- lm(data[4,] ~ data[5,] +
ALL_bcrneg$mol.biol + int)
> fitout <- summary(fit1)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.5971	0.6249	2.56	0.0126
"1004_at"	0.7815	0.2398	3.26	0.0017
mol.biolNEG	0.1388	0.8821	0.16	0.8754
int	-0.0257	0.1513	-0.17	0.8656

Table: Linear regression model with interaction term

Gene Expression Example: Simplified model

$$Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i$$




```
>fit2<-lm(data[4,] ~ data[5,])
>aa<-summary(fit2)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.6740	0.4348	3.85	0.0002
"1004_at"	0.7416	0.0746	9.95	0.0000

Model Selection: Likelihood Ratio Test

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_1 + \beta_2 Z + \beta_3 X_1 \times Z + \epsilon_i \\
&\text{or} \\
Y_i &= \beta_0 + \beta_1 X_1 + \epsilon_i
\end{aligned}
$$

```
> anova(fit1, fit2)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	75	2.31				
2	77	2.31	-2	-0.00	0.05	0.9491

p value > 0.05 suggests that both models fit data equally well. We choose the simple over the complicated model.

		Independent Variable	
		Categorical	Continuous
Outcome Variable	Continuous	T-Test, ANOVA (A)	Regression (C)
	Categorical	χ^2, Fisher (B)	GLM (D)

For Binary Response

Y = 0 or 1, a binary response

$$\begin{aligned} \hat{Y} &= \beta_0 + \beta_1 X \quad ? \text{ Y=1.2 } ? \\ \Pr(Y = 1) &= \beta_0 + \beta_1 X \quad ? \text{ Pr(Y=1) =1.1 } ? \end{aligned}$$

The problem:
$\rightarrow$ the right hand side, $\beta_0 + \beta_1 X \in (-\infty, \infty)$

Logistic Regression

$$\log\left[\frac{\Pr(Y=1)}{1-\Pr(Y=1)}\right] = \beta_0 + \beta_1 X$$

or

$$\text{logit}[\Pr(Y=1)] = \beta_0 + \beta_1 X$$

$\text{logit}(z) = \log \frac{z}{1-z}$

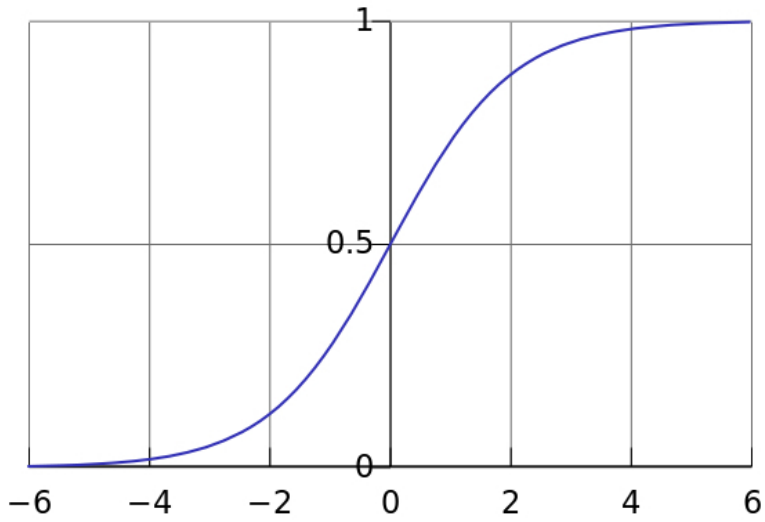


Figure: The logistic function

Interpretation of β's

$$\log\left[\frac{\Pr(Y=1)}{1-\Pr(Y=1)}\right] = \beta_0 + \beta_1 X$$

β_0: log odds when X=0

β_1: change in log odds with 1 unit increase in X.

For example:

X=4, odds = $e^{\beta_0+\beta_1\times 4}$

X=3, odds = $e^{\beta_0+\beta_1\times 3}$

$$OR_{\frac{X=4}{X=3}} = \frac{e^{\beta_0+\beta_1\times 4}}{e^{\beta_0+\beta_1\times 3}} = e^{\beta_1}$$

With 1 unit increase in X, odds of Y=1 increases e^{β_1} times.

FAMuSS Example

	Genotype	
BMI $>$ 25	AA	(GA and GG)
1	30	314
0	30	626
	60	940

$$\text{OR}_{\frac{AA}{other}} = \frac{ad}{bc} = 1.99 = e^{0.69}$$

```
>geno<-ifelse(Geno=="AA", 1, 0)
>fit4<-glm(trait ~ geno, data=fms,
family=binomial(link=logit))
```

	Estimate	Std. Error	z value	Pr($>$|z|)
(Intercept)	-0.69	0.0692	-9.98	0.0000
geno	0.69	0.2673	2.58	0.0098