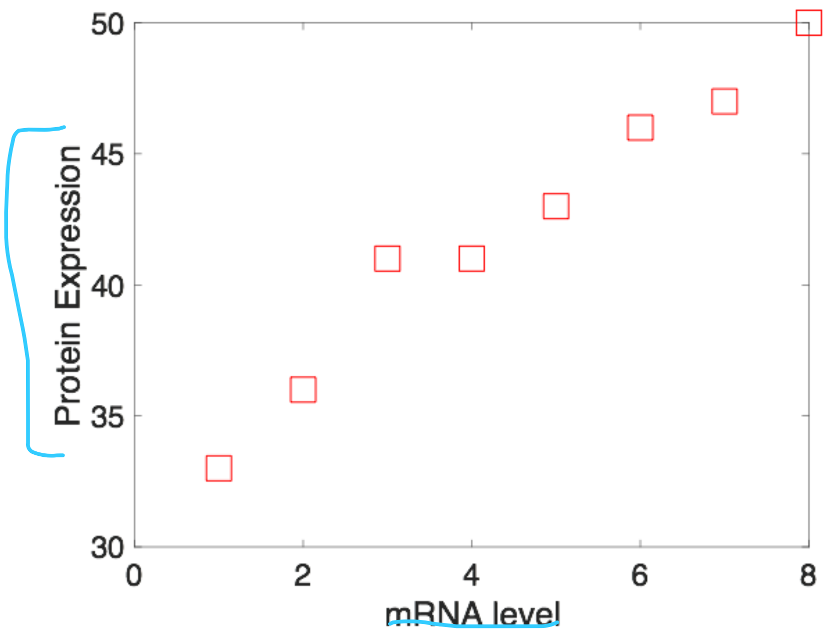
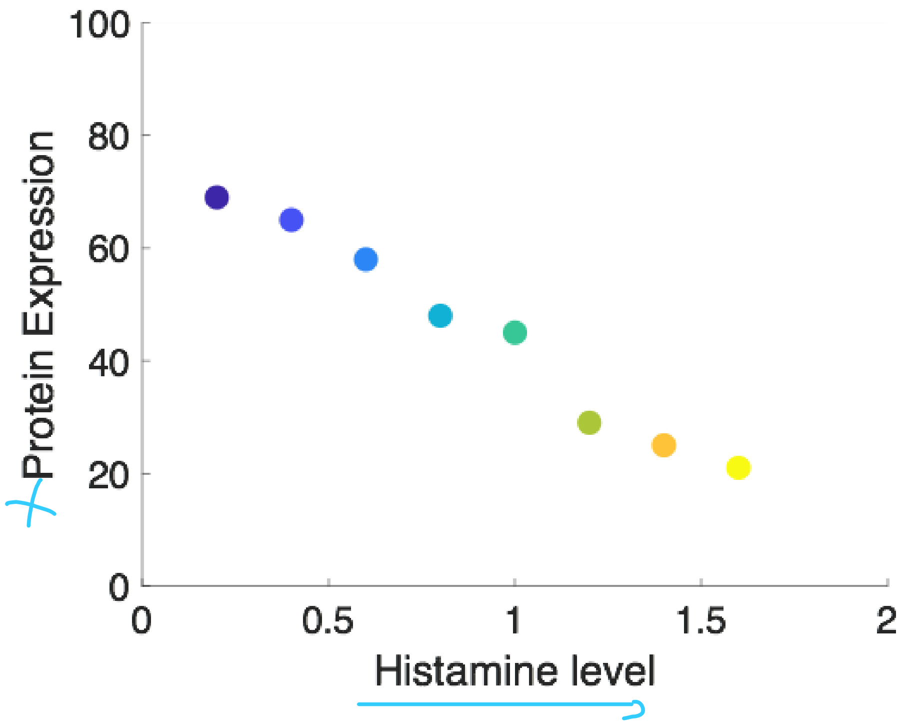
Week 13

Regression analysis in biology

Scatter plot

Shows the relation between two variables



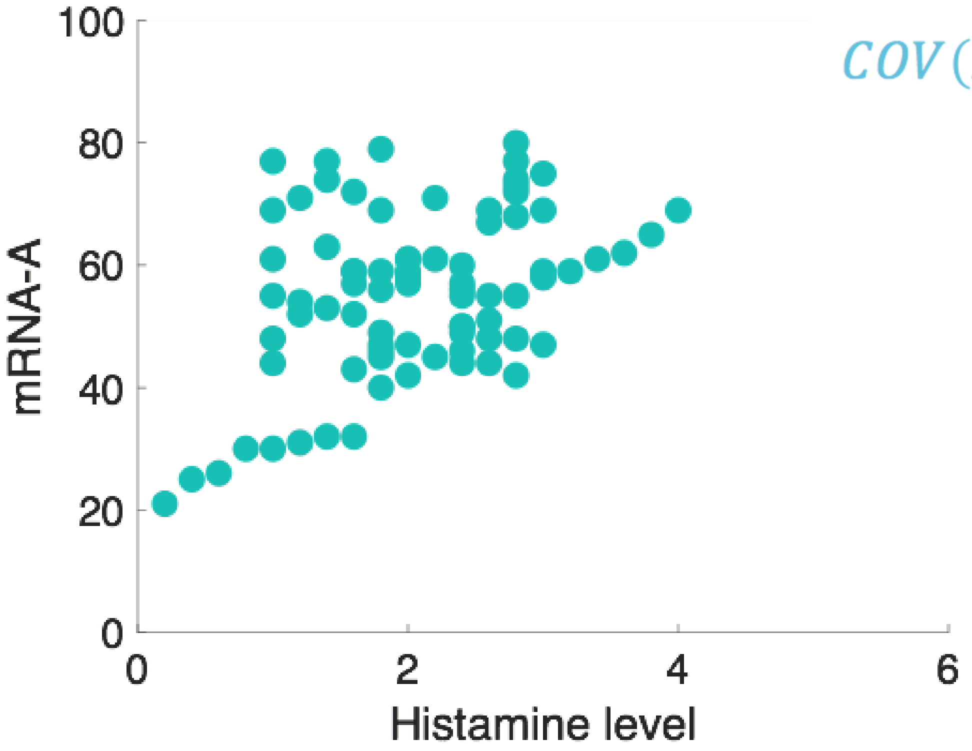


Can we quantitatively measure the strength of relationship between variables?

Linear regression is a form of regression in which one exploratory variable is used to predict the outcome of a response variable.

Covariance

Does Y get larges (smalleR) as Y increase?



$$COV(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

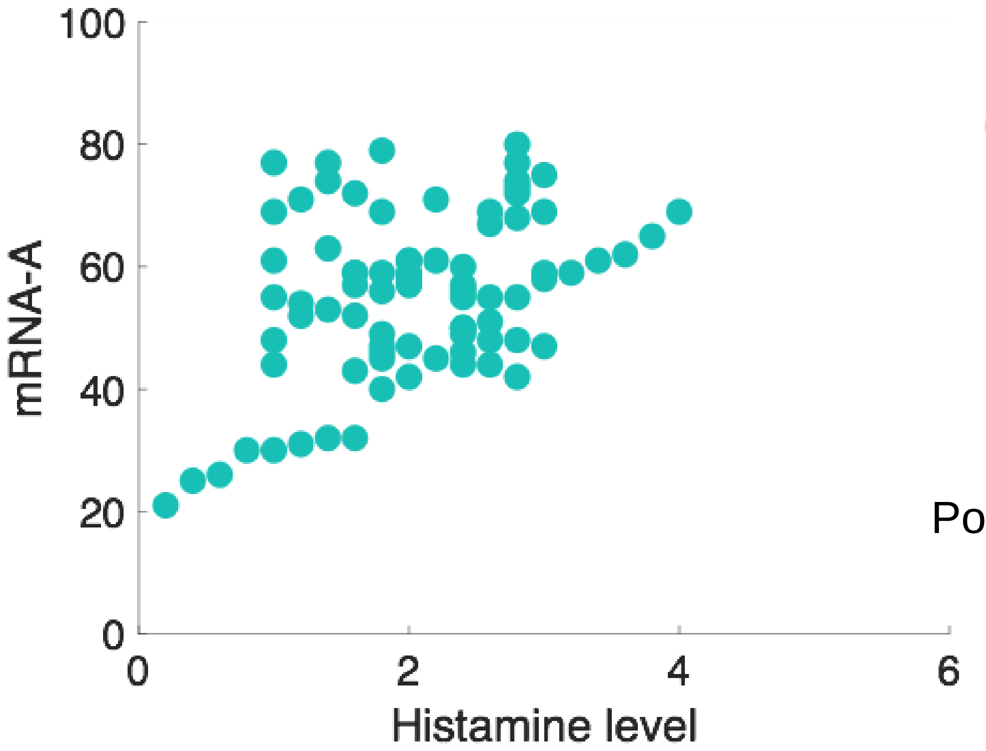
mean of x= 2.0525
mean of y = 55.4125
Sx = 0.7916
Sy= 13.6537

n=80

Covariance > 0 if X and Y variables gets larger

Covariance < 0 if X and Y variables moves opposite direction

Covariance of Histamine vs mRNA levels



$$COV(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Cov=1.65

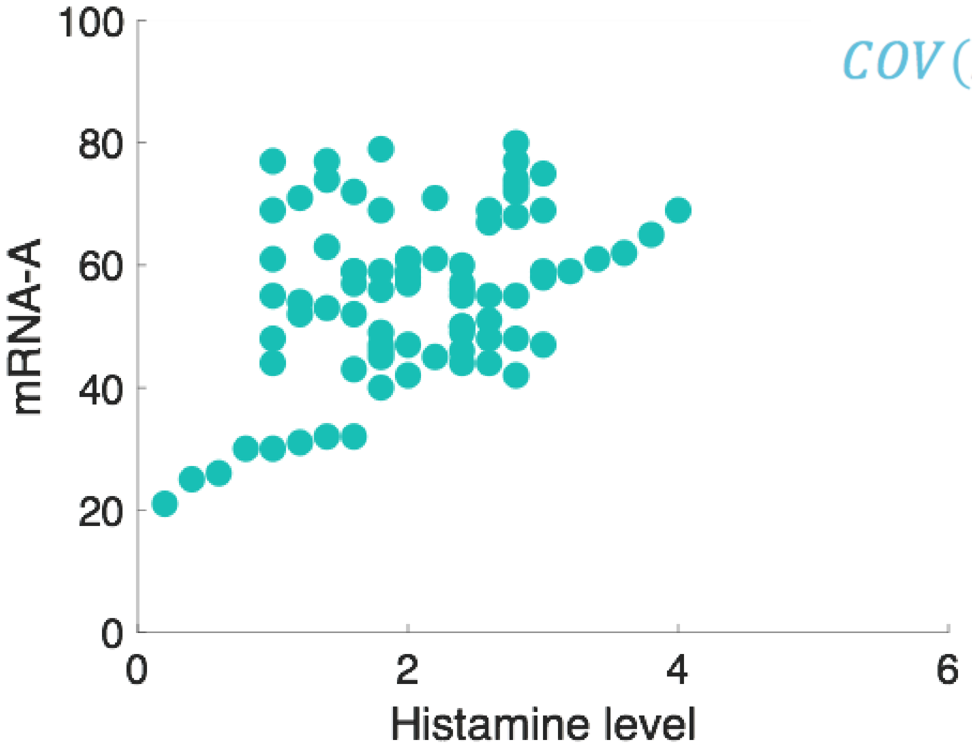
Positive correlation

Sign is a good indicator of relationship but
what is the meaning of 1.65? is it a strong or weak relationship?

To determine the strength of relation, Correlation coefficient
is needed?

Covariance

Does Y get larges (smalleR) as Y increase?



$$COV(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

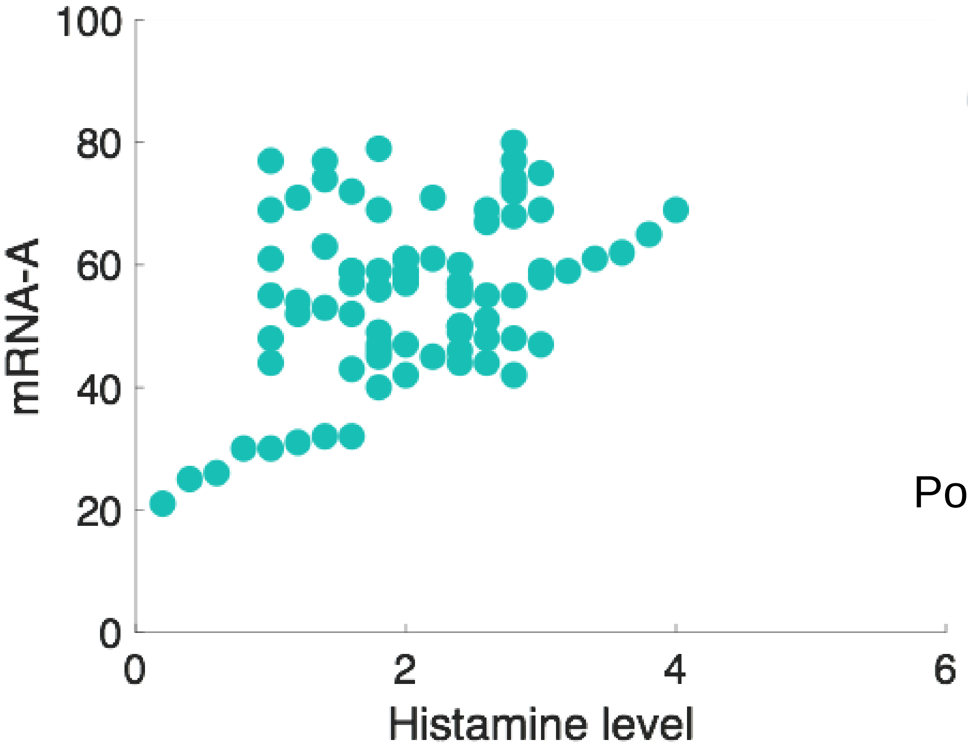
mean of x= 2.0525
mean of y = 55.4125
Sx = 0.7916
Sy= 13.6537

n=80

Covariance > 0 if X and Y variables gets larger

Covariance < 0 if X and Y variables moves opposite direction

Covariance of Histamine vs mRNA levels



$$COV(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Cov=1.65

Positive correlation

Sign is a good indicator of relationship but
what is the meaning of 1.65? is it a strong or weak relationship?

To determine the strength of relation, Correlation coefficient
is needed?

Correlation (r)

measures the direction and strength of relationship between two quantitative variable.
The correlation *r* measures the direction and strength of the linear (straight line) association between two quantitative variables *x* and *y*.
Although you can calculate a correlation for any scatterplot, *r* measures only linear relationships.

$$r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

$$\begin{aligned} r &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i-\bar{x})^2)(\sum_{i=1}^{n}(y_i-\bar{y})^2)}} \\ &= \frac{1}{n-1}\cdot\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{s_x s_y} \end{aligned}$$

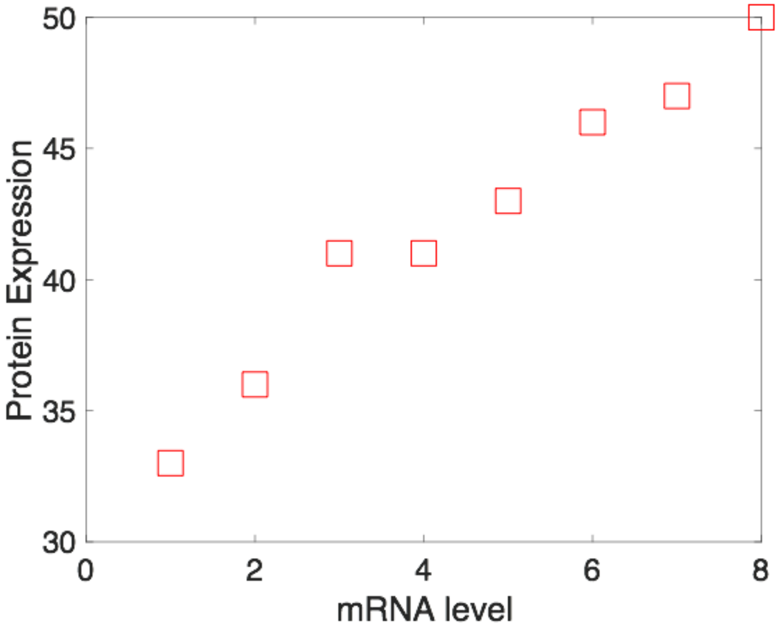
close to n-1 if x and y have

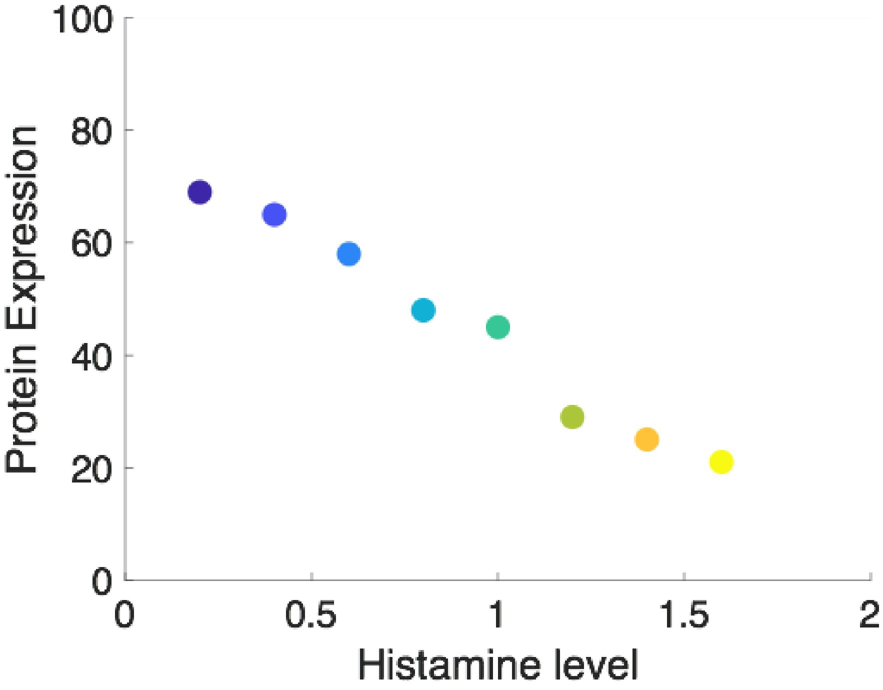
$\bar{x}$ = the sample mean of $x_1,...,x_n$,

$\bar{y}$ = the sample mean of $y_1,...,y_n$,

s_x = the standard deviation of $x_1, . . . , x_n$,

s_y =thestandarddeviationof$y_1,...,y_n$.





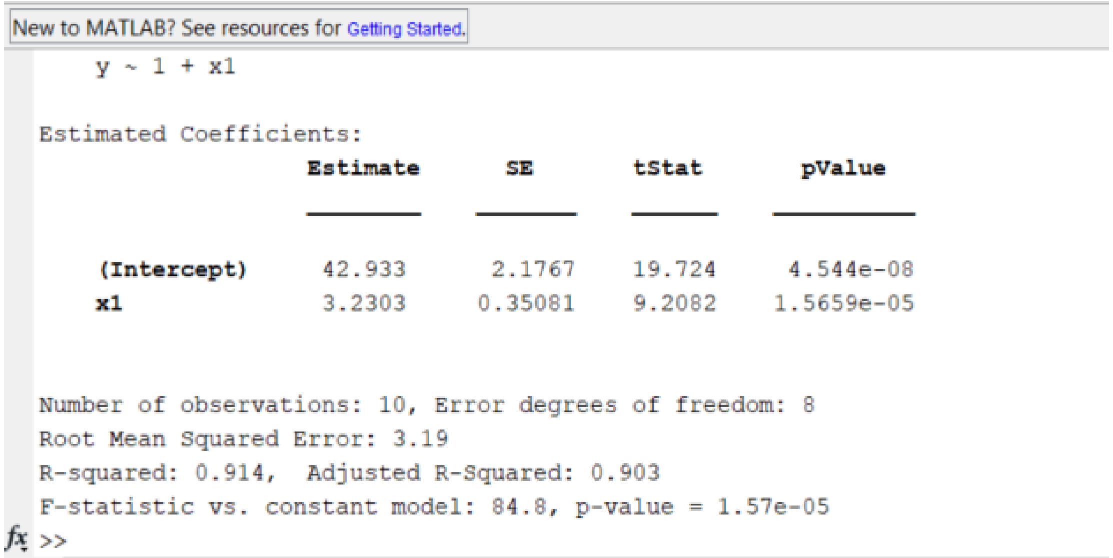
r=+0.9538

r=-0.987

Correlation coefficient always lies between -1 to +1

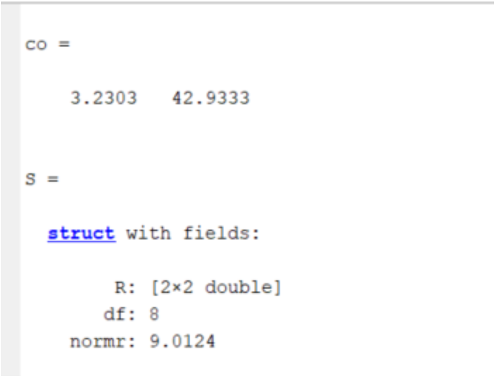
Fitlm and polyfit functions

b = fitlm(hist',genetrial')



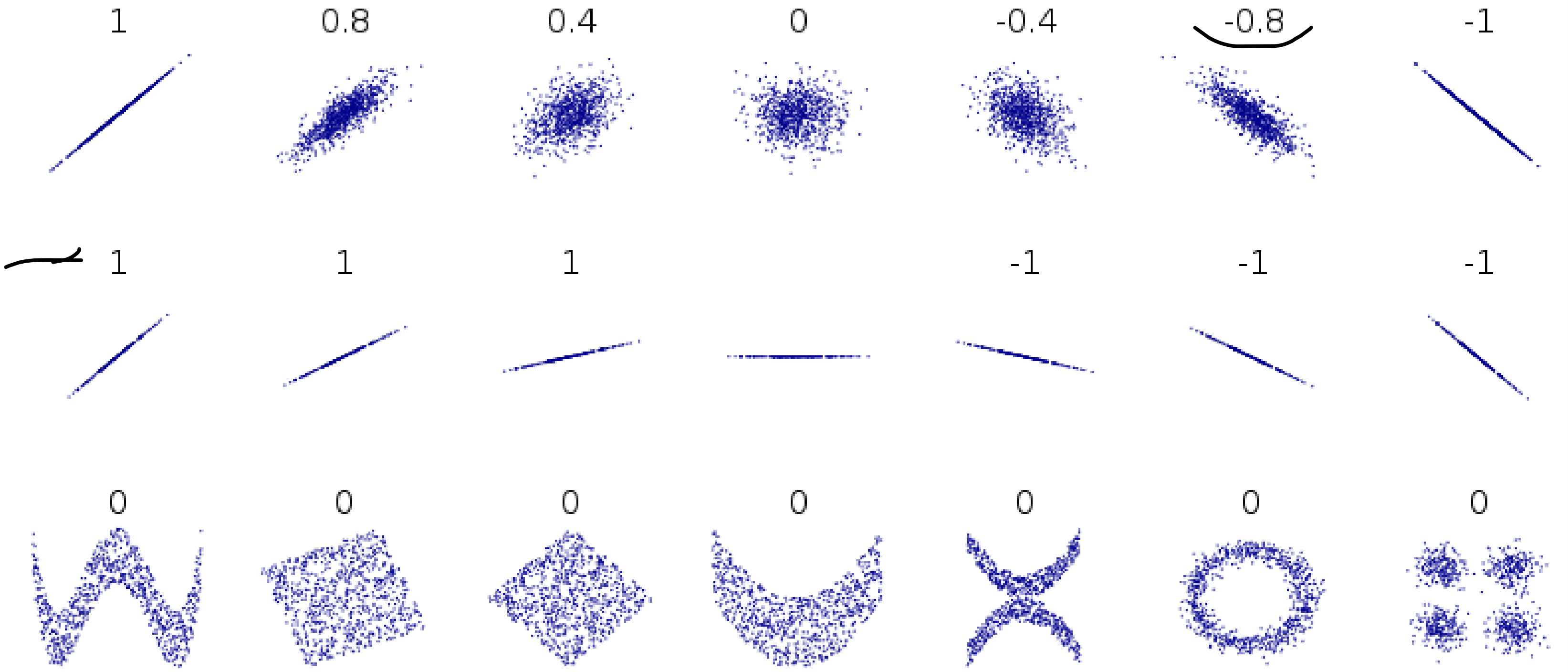
```
New to MATLAB? See resources for Getting Started.
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate      SE        tStat       pValue  
                   ________    _______    ______    __________

    (Intercept)     42.933      2.1767    19.724     4.544e-08
    x1              3.2303     0.35081    9.2082    1.5659e-05


Number of observations: 10, Error degrees of freedom: 8
Root Mean Squared Error: 3.19
R-squared: 0.914,  Adjusted R-Squared: 0.903
F-statistic vs. constant model: 84.8, p-value = 1.57e-05
fx >>
```

[co,S]=polyfit(hist,genetrial,1)



```
co =

    3.2303   42.9333


S =

  struct with fields:

        R: [2×2 double]
       df: 8
    normr: 9.0124
```

Correlation sets



Remember that correlation coefficient is an indicator of the strength of a *linear* relationship between two variables, but its value generally does not completely characterize their relationship

Summary of Correlation between two variables

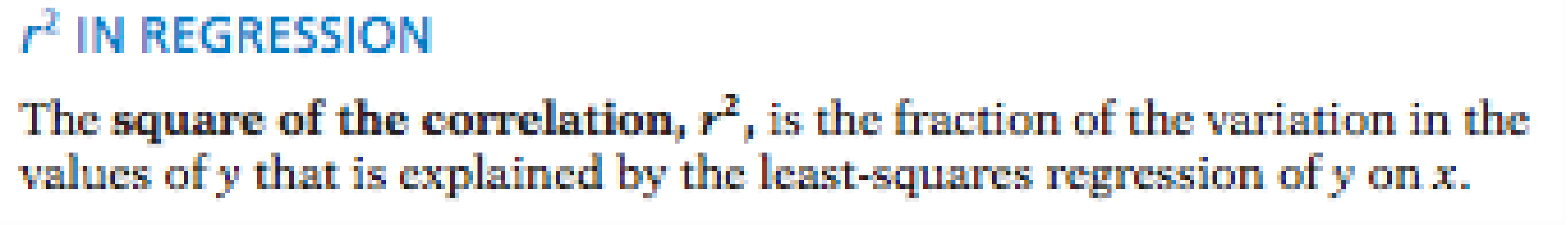
- $-1 \leq r \leq 1$ always
- $r = 1$ when all the points (x_i, y_i) lie on a line with positive slope
- $r = -1$ when all the points (x_i, y_i) lie on a line with negative slope
- When $r = 0$, then there is no positive or negative linear association between the two variables (though the two variables may have a non-linear relationship).

How to find a best fit line? how do you know if these coefficients are right? What does software magically return the coefficients?

Data parameters

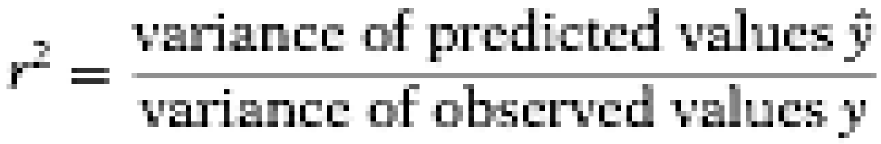
Equation
$y=ax+b+e$

x = independent variable
y = dependent variable (maybe not dependent who knows)
b = intercept
a = slope
e = error



r^2 IN REGRESSION

The square of the correlation, r^2, is the fraction of the variation in the values of y that is explained by the least-squares regression of y on x.



$$r^2 = \frac{\text{variance of predicted values } \hat{y}}{\text{variance of observed values } y}$$

Properties of r2

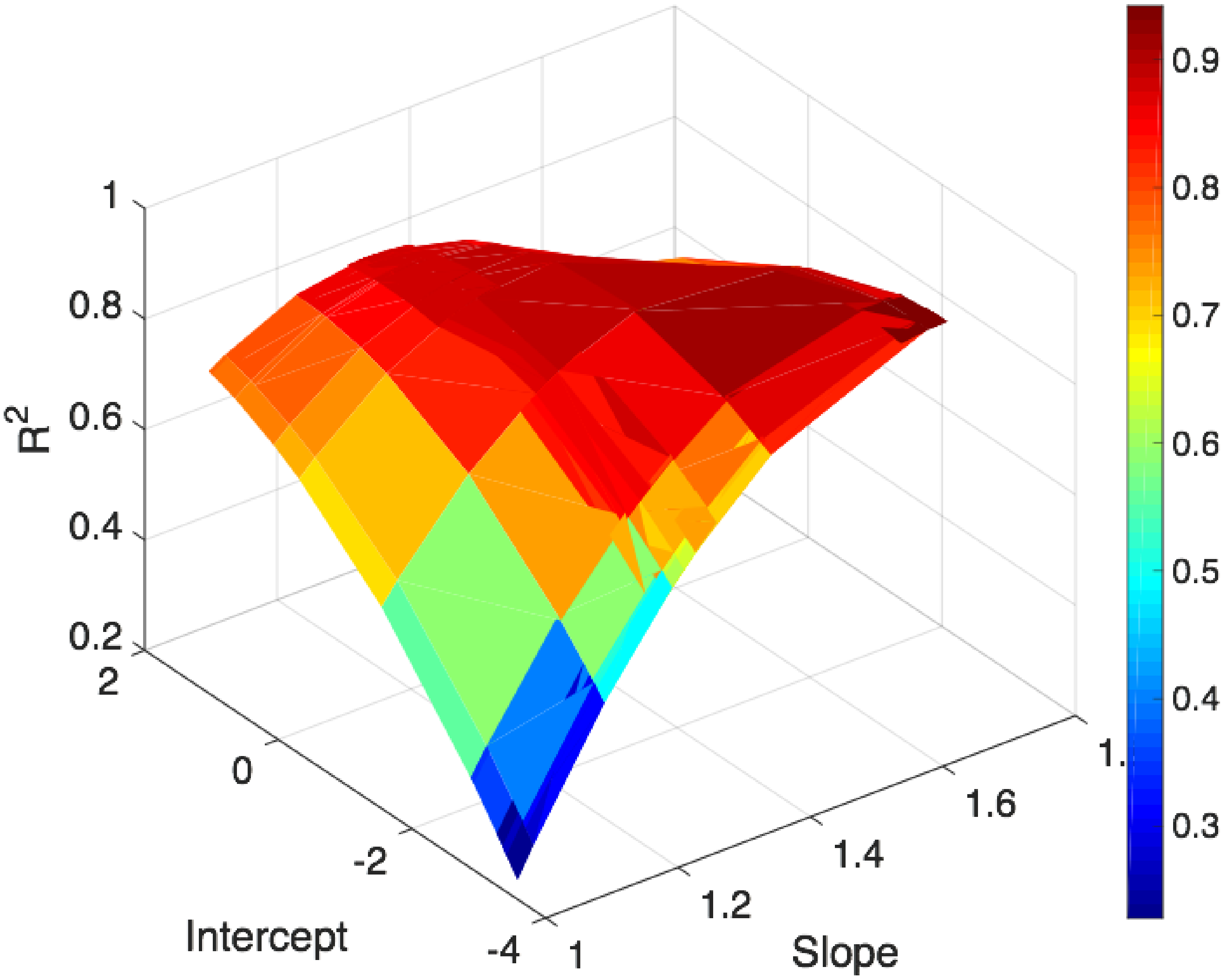
0 =< r^2 =< 1

if r^2 = 1, it represents a straight line
if r^2 = 0, it indicates no correlation between y and x

Larger the r^2 means higher correlation, but not always

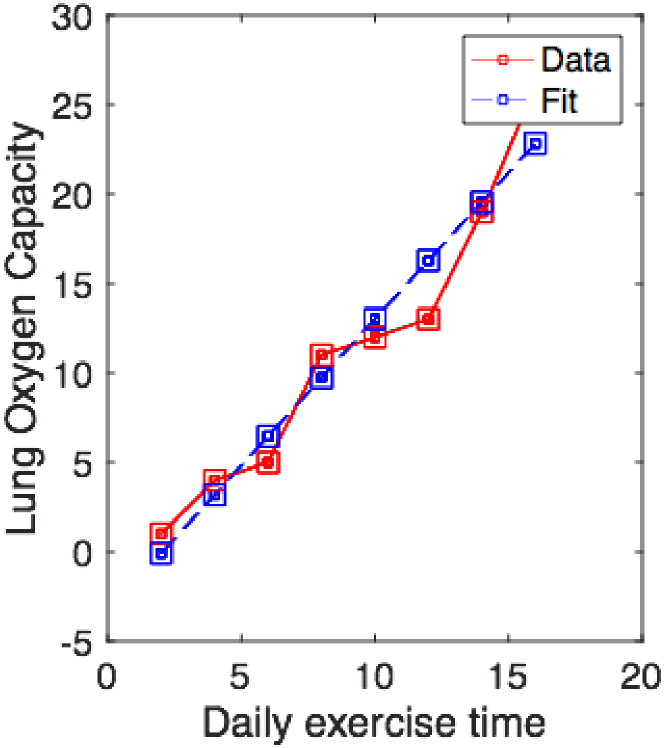
R2 gets smaller by the size of

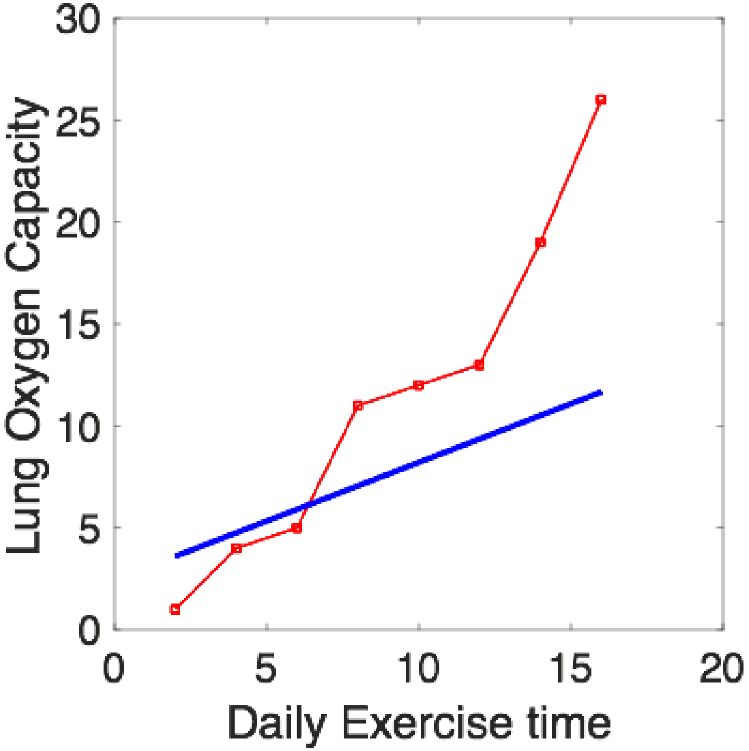
Slope = 1.63
Intercept = -3.35



Compare data fittings

Perfect Fit





R^2=0.94

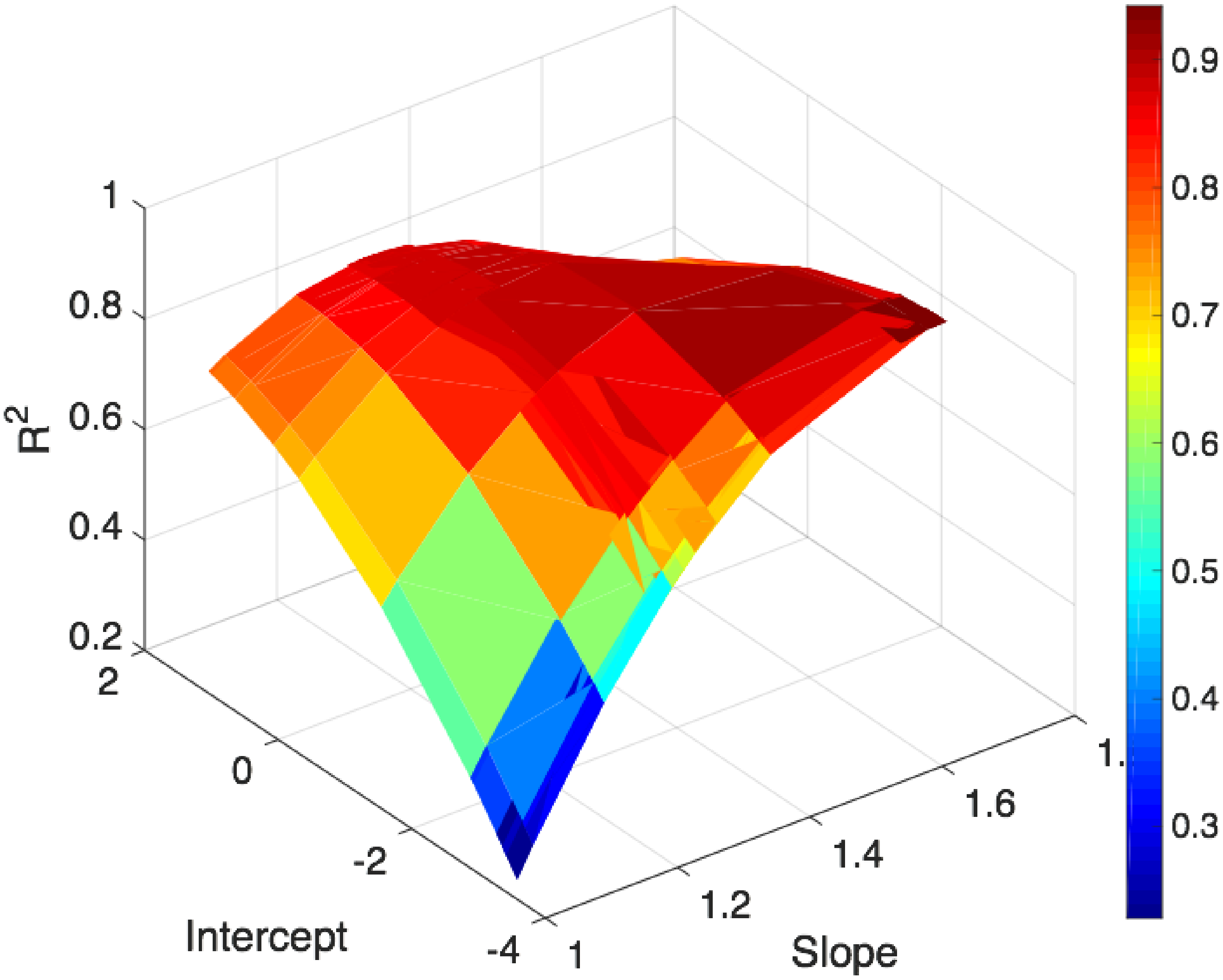
Slope = 1.63
Intercept = -3.35

R^2=0.77

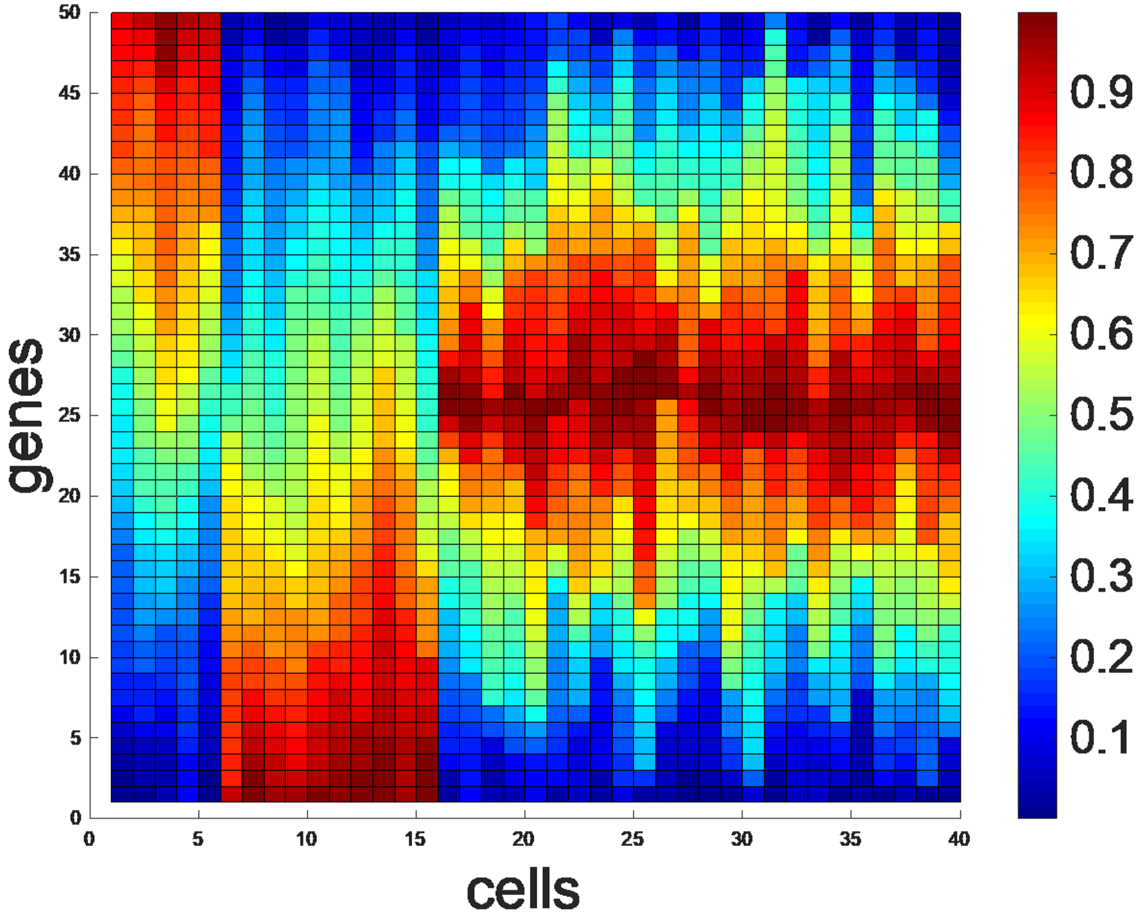
Slope = 0.57
Intercept = 2.45

R2 gets smaller by the size of

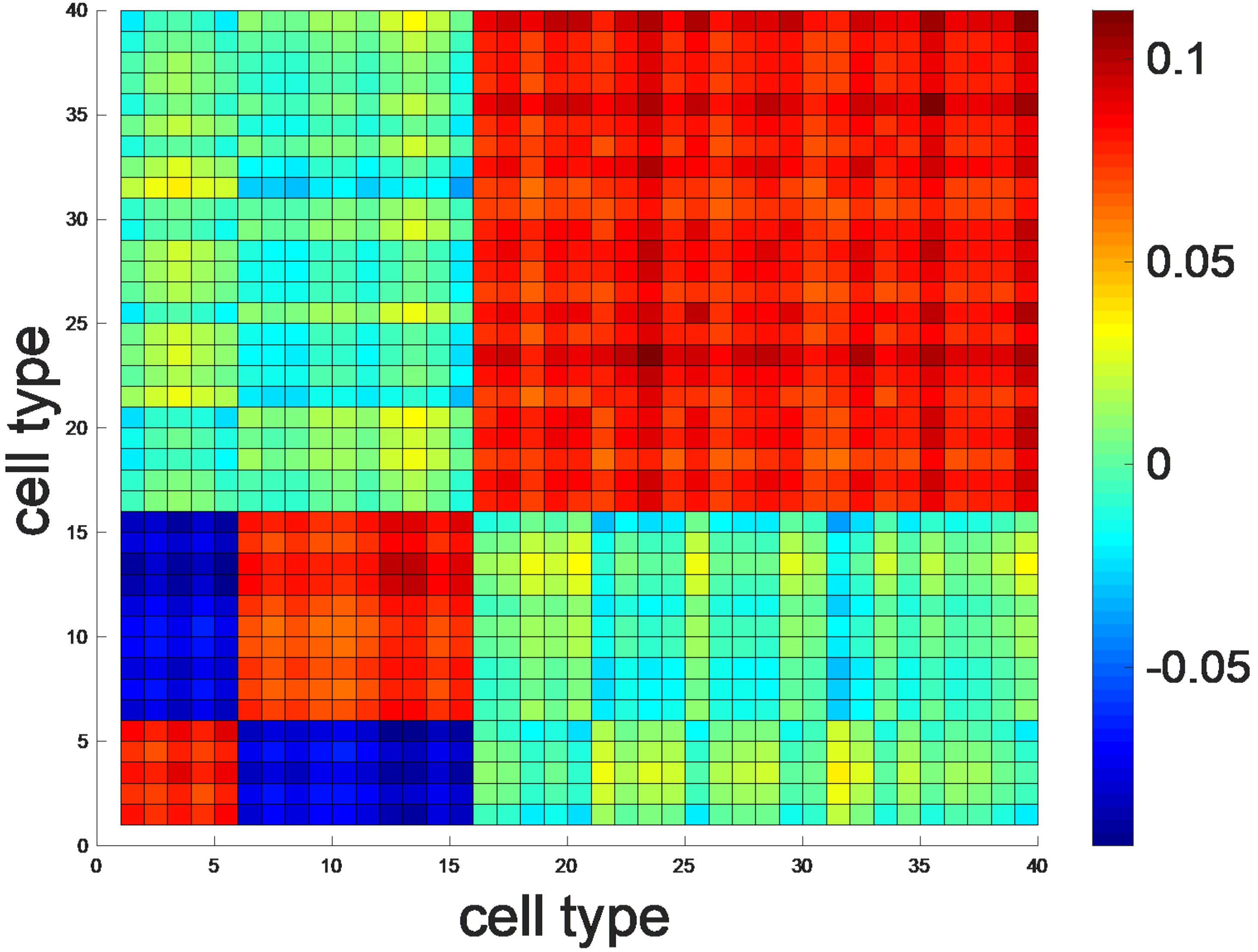
Slope = 1.63
Intercept = -3.35



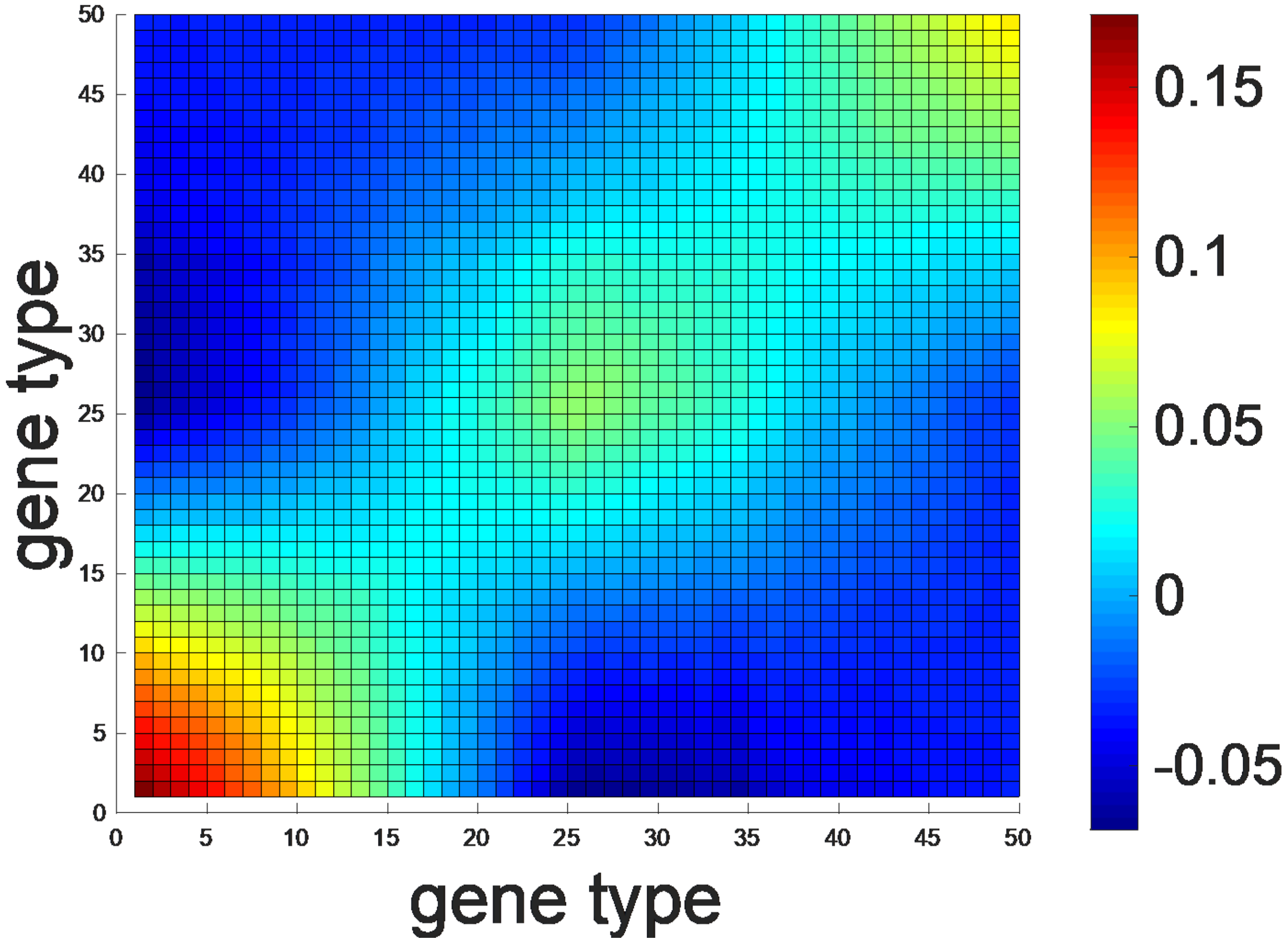
Gene expression in different cells



What is the covariance between different cell types?

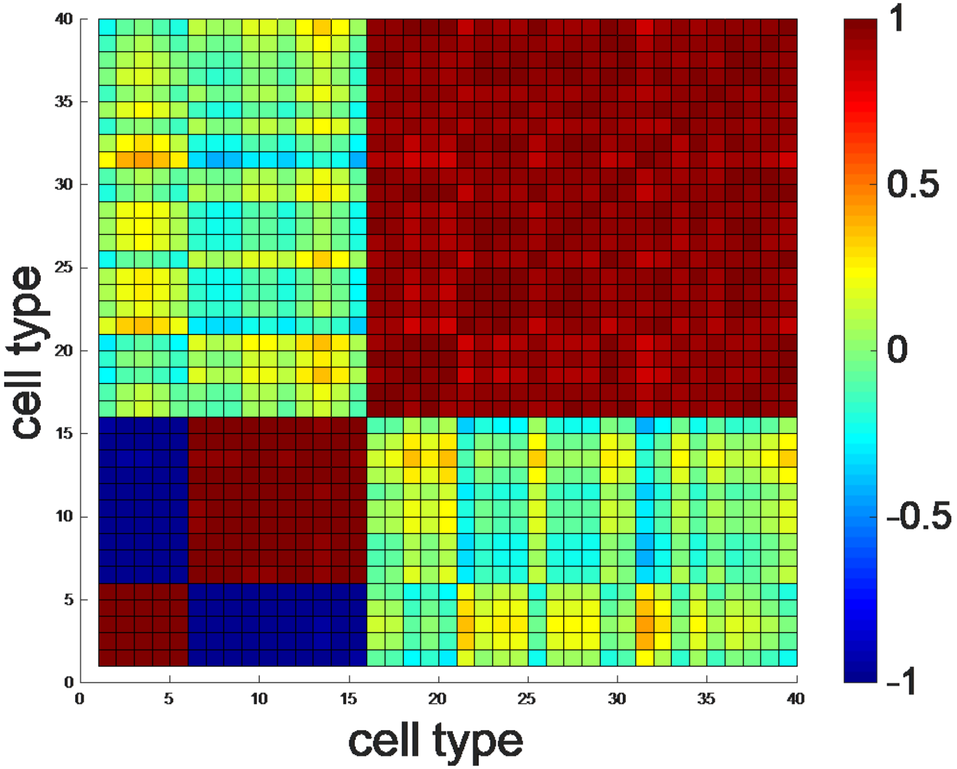


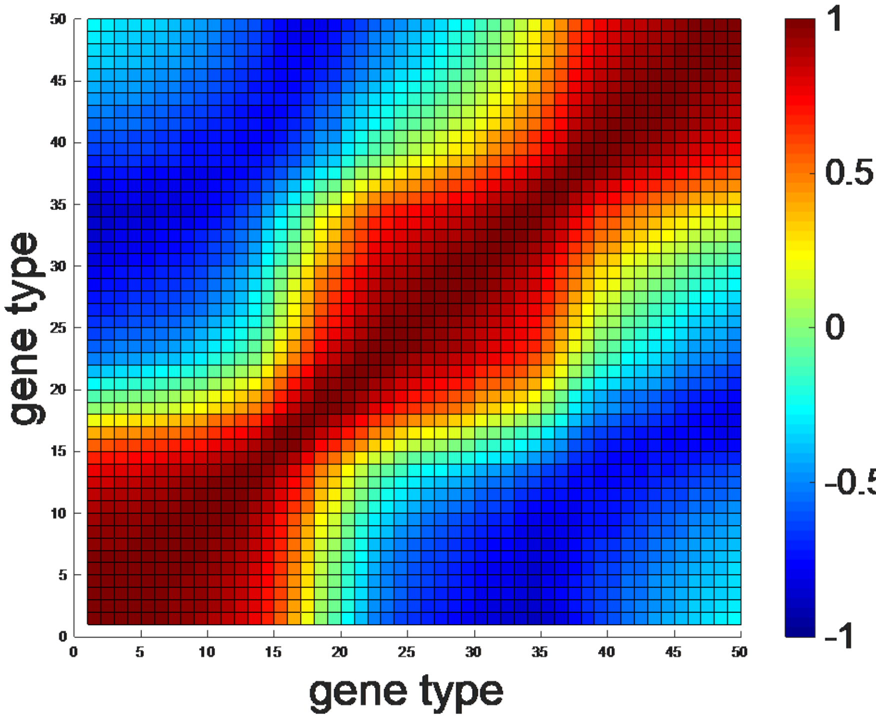
What is the covariance between different genes?



Correlation of difference cell types

Liver cells, kidney cells and neurons





Correlation ranges from -1 to 1
Covariance can be any number

Covariance returns the direction of relation while the correlation returns the strength of relationship

MULTIVARIATE REGRESSION

In linear regression, a single independent variable was present. A total of two variables. In multiple regression, y dependent variable (response variable) depends on a many explanatory independent variables.

Now we can define linear function as

$$Y = \text{constant (a)} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \;....\; + \beta_k x_n$$

It is also called as population regression equation.

y varies normally with a mean given by the population
regression equation

MULTIVARIATE REGRESSION

- y - dependent variable or also called response variable

- x_1, x_2 , x_3... , x_n are called independent variables

or explanatory variables.

- X values can either quantitative or categorical.

$$Y = \text{constant (a)} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \text{ ....} + \beta_k x_n$$

The statistical model for multiple linear regression is

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i}$$

for $i = 1,2,...,n$.

Parameter coefficients of the model are $\underline{\beta_0}$, $\underline{\beta_1}$, $\beta_2,...,\beta_p$, and σ.

For the *i*th observation, the predicted response is

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip}$$

$$e_i = \text{observed response} - \text{predicted response} = y_i - \hat{y}_i$$

$$= y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_p x_{ip}$$

Examples of multivariate regression

1. Dependence of fuel consumption in cars to horsepower, accelaration and weight (engineering)

2. Dependence of cancer risk to several genes (biology)

3. Dependence of home price to location, size, type etc. (home market)

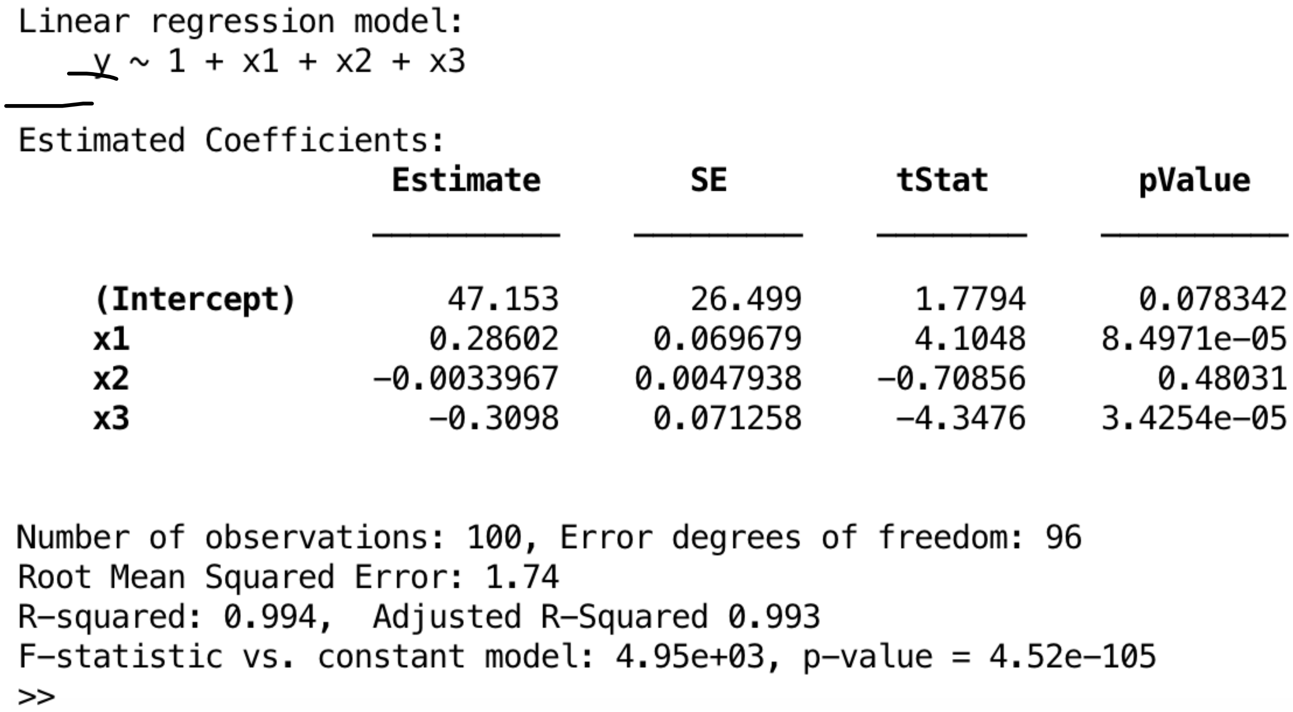
4. Dependence of hormone levels to genes (health)

5. Dependence of reading score to mothers educaiton, age, gender, family income etc. (social science)

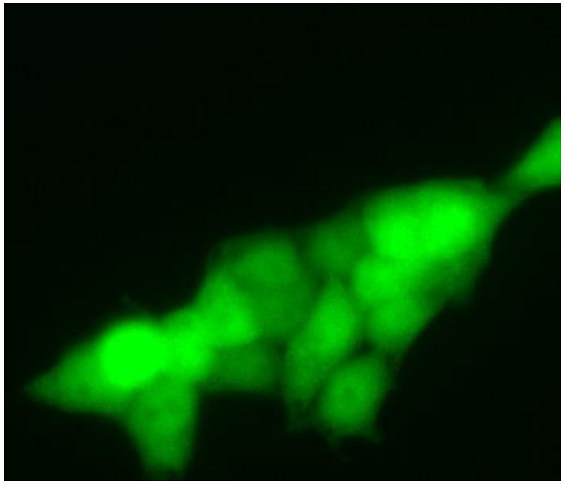
In Matlab

```
mdl = fitlm(X,Y)
```

Dependence of cell growth to expression of geneX, geneY and geneZ

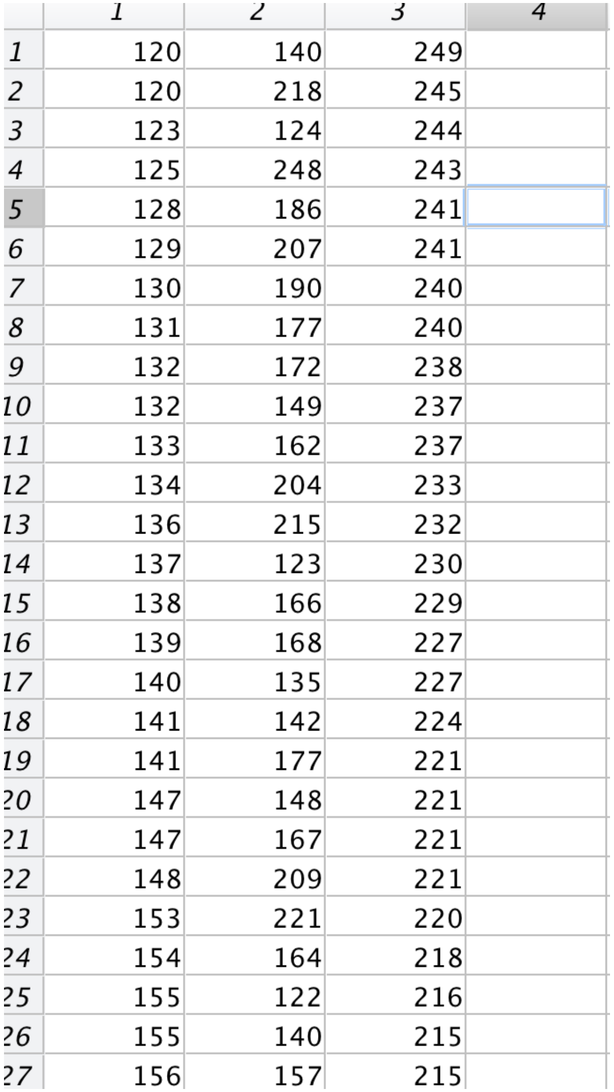


```
Linear regression model:
    y ~ 1 + x1 + x2 + x3

Estimated Coefficients:
                   Estimate        SE         tStat       pValue
                  __________    _________    ________    __________

    (Intercept)       47.153       26.499      1.7794      0.078342
    x1               0.28602     0.069679      4.1048    8.4971e-05
    x2            -0.0033967    0.0047938    -0.70856       0.48031
    x3               -0.3098     0.071258     -4.3476    3.4254e-05


Number of observations: 100, Error degrees of freedom: 96
Root Mean Squared Error: 1.74
R-squared: 0.994,  Adjusted R-Squared 0.993
F-statistic vs. constant model: 4.95e+03, p-value = 4.52e-105
>>
```

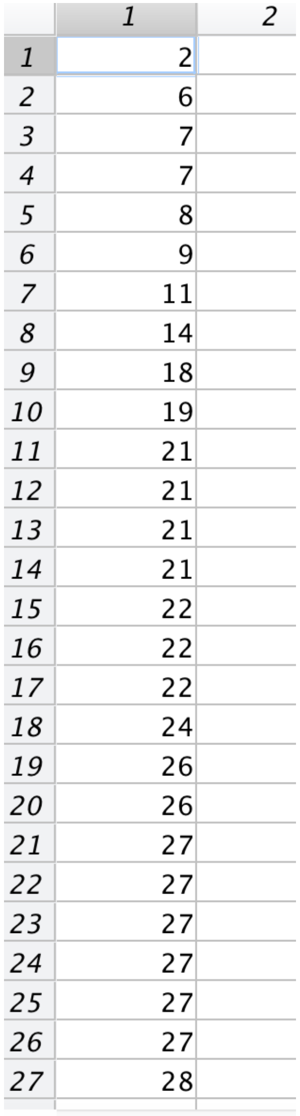


Cell growth = 47 + 0.28geneX -0.003geneY-0.30geneZ

Dependence of hormone levels to expression of geneX, geneY and geneZ



	1	2	3	4
1	120	140	249	
2	120	218	245	
3	123	124	244	
4	125	248	243	
5	128	186	241	
6	129	207	241	
7	130	190	240	
8	131	177	240	
9	132	172	238	
10	132	149	237	
11	133	162	237	
12	134	204	233	
13	136	215	232	
14	137	123	230	
15	138	166	229	
16	139	168	227	
17	140	135	227	
18	141	142	224	
19	141	177	221	
20	147	148	221	
21	147	167	221	
22	148	209	221	
23	153	221	220	
24	154	164	218	
25	155	122	216	
26	155	140	215	
27	156	157	215	



	1	2
1	2	
2	6	
3	7	
4	7	
5	8	
6	9	
7	11	
8	14	
9	18	
10	19	
11	21	
12	21	
13	21	
14	21	
15	22	
16	22	
17	22	
18	24	
19	26	
20	26	
21	27	
22	27	
23	27	
24	27	
25	27	
26	27	
27	28	

Lets predict cell growth

We conclude that geneX and gene Z contain useful information for predicting cell growth

Let’s find the predicted cell growth for a sample with an 0.3 average in geneX and 0.6 in geneZ.

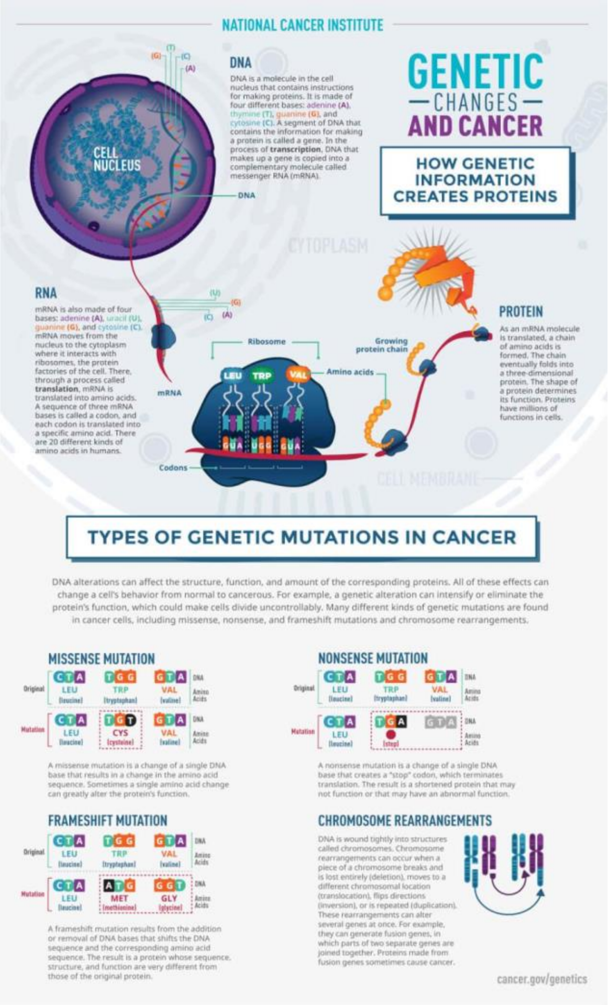
The explanatory variables are geneX and geneY. The predicted cell growth is

$$\text{Cell growth} = 47 + 0.28\text{geneX} - 0.003\text{geneY} - 0.30\text{geneZ}$$

$$\text{Cell growth} = 47 + 0.28\text{gene} - 0.3\text{geneZ}$$

$$\text{Cell growth} = 47 + 0.28(0.3) - 0.30(0.6)$$

Logistic Regression



NATIONAL CANCER INSTITUTE

GENETIC CHANGES AND CANCER

HOW GENETIC INFORMATION CREATES PROTEINS

DNA

DNA is a molecule in the cell nucleus that contains instructions for making proteins. It is made of four different bases: adenine (A), thymine (T), guanine (G), and cytosine (C). A segment of DNA that contains the information for making a protein is called a gene. In the process of **transcription**, DNA that makes up a gene is copied into a complementary molecule called messenger RNA (mRNA).

RNA

mRNA is also made of four bases: adenine (A), uracil (U), guanine (G), and cytosine (C). mRNA moves from the nucleus to the cytoplasm where it interacts with ribosomes, the protein factories of the cell. There, through a process called **translation**, mRNA is translated into amino acids. A sequence of three mRNA bases is called a codon, and each codon is translated into a specific amino acid. There are 20 different kinds of amino acids in humans.

PROTEIN

As an mRNA molecule is translated, a chain of amino acids is formed. The chain eventually folds into a three-dimensional protein. The shape of a protein determines its function. Proteins have millions of functions in cells.

TYPES OF GENETIC MUTATIONS IN CANCER

DNA alterations can affect the structure, function, and amount of the corresponding proteins. All of these effects can change a cell's behavior from normal to cancerous. For example, a genetic alteration can intensify or eliminate the protein's function, which could make cells divide uncontrollably. Many different kinds of genetic mutations are found in cancer cells, including missense, nonsense, and frameshift mutations and chromosome rearrangements.

MISSENSE MUTATION

A missense mutation is a change of a single DNA base that results in a change in the amino acid sequence. Sometimes a single amino acid change can greatly alter the protein's function.

NONSENSE MUTATION

A nonsense mutation is a change of a single DNA base that creates a "stop" codon, which terminates translation. The result is a shortened protein that may not function or that may have an abnormal function.

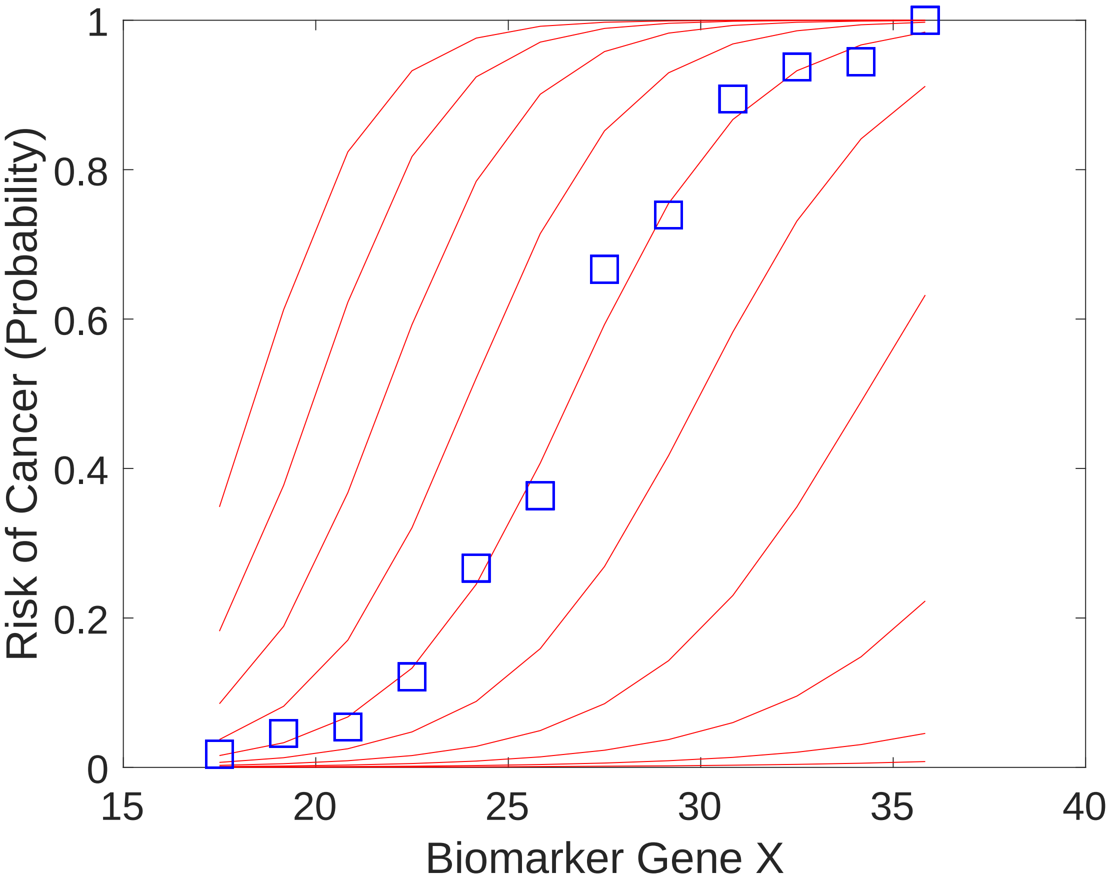
FRAMESHIFT MUTATION

A frameshift mutation results from the addition or removal of DNA bases that shifts the DNA sequence and the corresponding amino acid sequence. The result is a protein whose sequence, structure, and function are very different from those of the original protein.

CHROMOSOME REARRANGEMENTS

DNA is wound tightly into structures called chromosomes. Chromosome rearrangements can occur when a piece of a chromosome breaks and is lost entirely (deletion), moves to a different chromosomal location (translocation), flips directions (inversion), or is repeated (duplication). These rearrangements can alter several genes at once. For example, they can generate fusion genes, in which parts of two separate genes are joined together. Proteins made from fusion genes sometimes cause cancer.

cancer.gov/genetics



What is logistic regression?

It is used to determine model parameters when dependent variables are binary rather than continuous

For example,
cell division, 0 or 1
Cancer diagnostic, cancer/not
Voting yes/no
Mortality alive/death
Product-marketing, sold/not sold
Arrived/delayed

The results of these data is not continuous as you have seen in multivariable linear regression

Logistic model can be used to make prediction for binary results

Logistic Regression

If a response variable such as yes/no or success/failure response variables., we cannot use linear regression models where it assumes a normal distribution.

Think about a cancer patient diagnosis whether a patient either have a cancer or not a cancer

One type of model that can be used is called **logistic regression.** We think in terms of a binomial model for the two possible values of the response variable and use one or more explanatory variables to explain the probability of success.

$P(Y=1|beta)= exp(b(1)+b(2)x) / 1+exp(b(1)+b(2)x)$

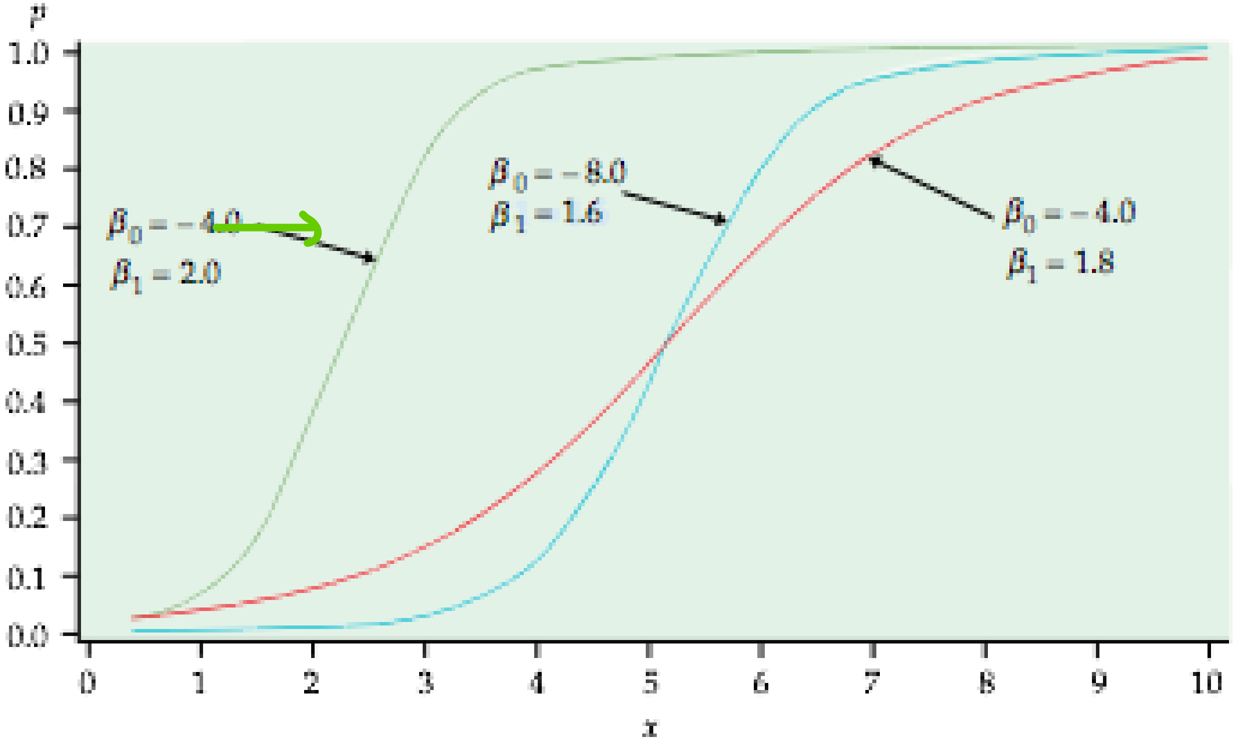
x= binary or cont

y= binary

$b(1)$ and $b(2)$ are coefficients

if y response variable is discrete

Y= P(Y=0) + P(Y=1)



Logistic function

it can be defined as

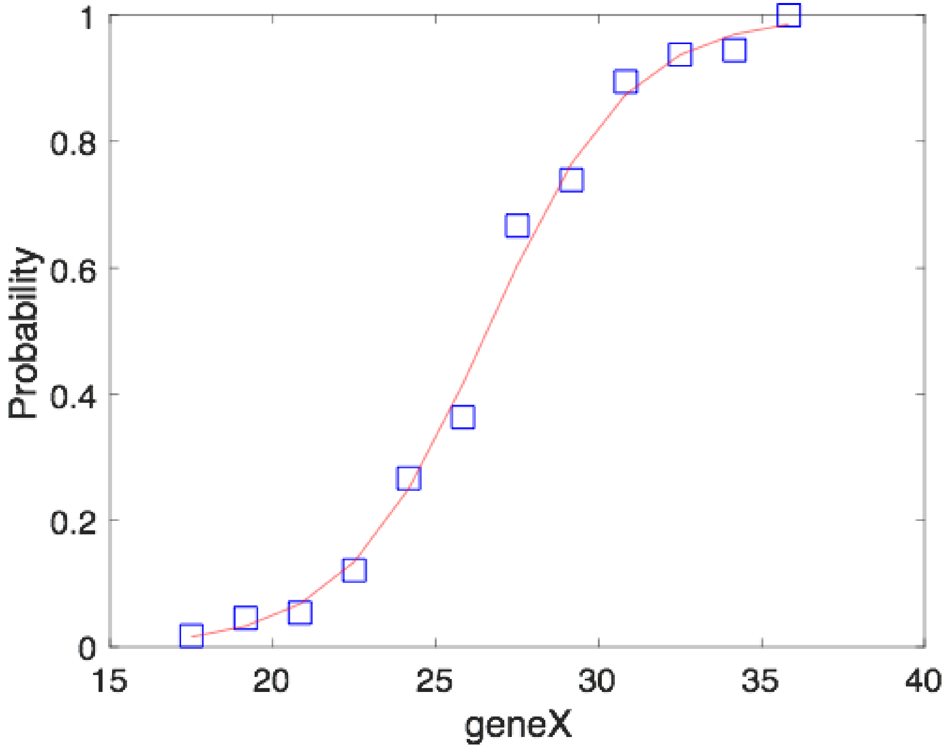
f(x)= exp(x) / 1+exp(x)

f(x) or y values always falls in range between 0 and 1

Solutions: Logistic regression

Logistic regression is the best model if response variable is binomial. Because it uses a fitting method that is appropriate for the binomial distribution.

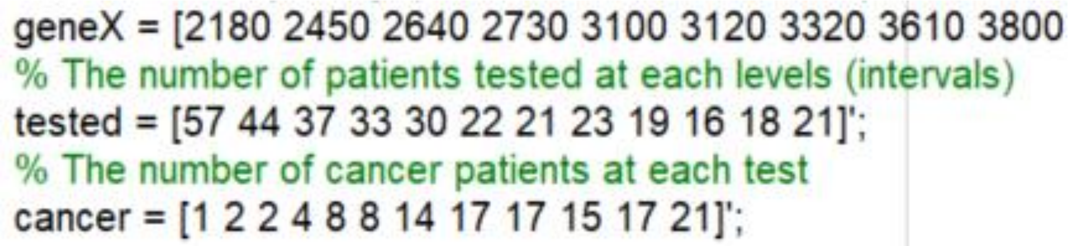
Predicted proportions/probability values are present in the range from 0 to 1.



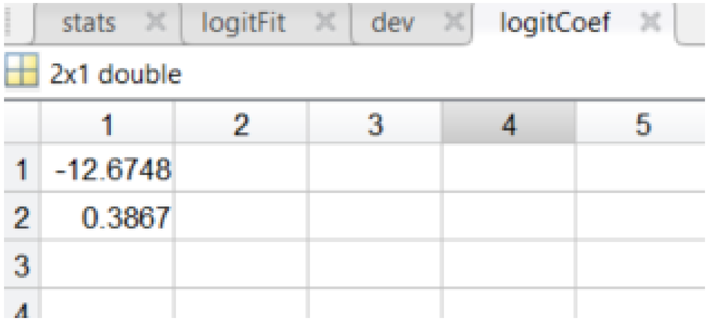
In matlab we use glmfit function to fit our data to a logistic model. This function returns coefficient estimates for a linear regression of the responses Y (f(x)) on the independent variable X

In Matlab,

```
%logistic regression

[logitCoef,dev,stats] = glmfit(geneX,[cancer
tested],'binomial','logit');
```



```
geneX = [2180 2450 2640 2730 3100 3120 3320 3610 3800
% The number of patients tested at each levels (intervals)
tested = [57 44 37 33 30 22 21 23 19 16 18 21]';
% The number of cancer patients at each test
cancer = [1 2 2 4 8 8 14 17 17 15 17 21]';
```

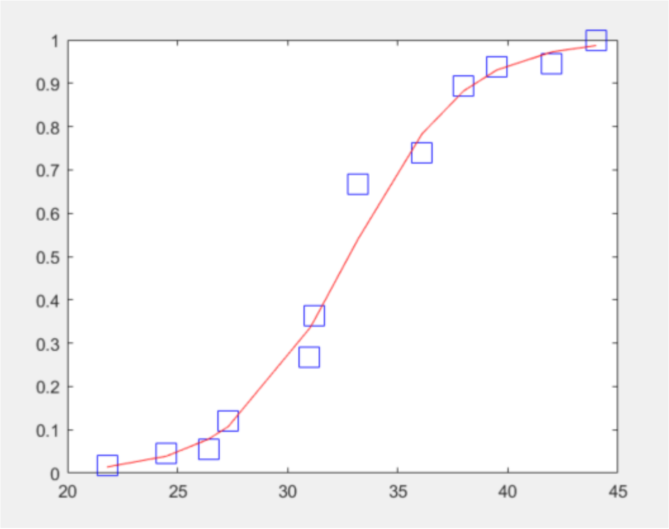


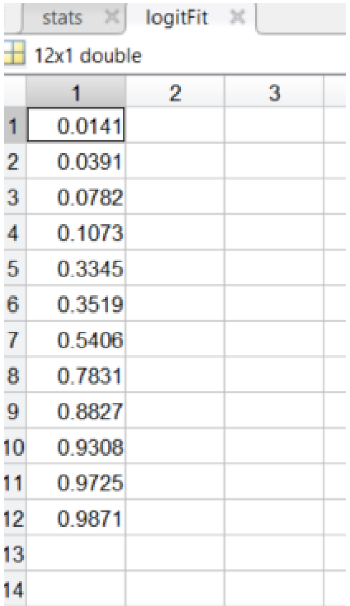
2x1 double

	1	2	3	4	5
1	-12.6748				
2	0.3867				
3					
4					

```
%logistic regression

[logitCoef,dev,stats] = glmfit(geneX,[cancer tested],'binomial','logit');
logitFit = glmval(logitCoef,geneX,'logit');

figure(3)
plot(geneX,proportion,'bs', geneX,logitFit,'r-','markersize',16);
```

Glmval is uses to compute the predicted values for the model

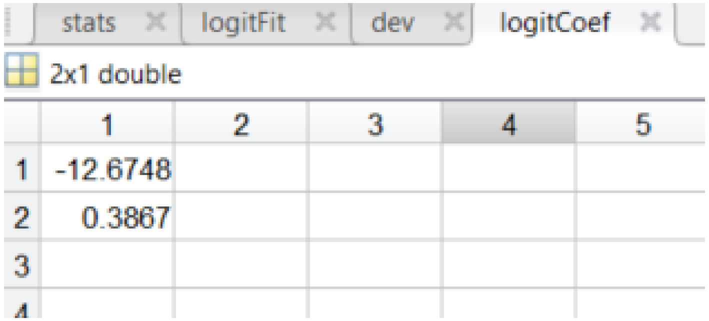




12x1 double

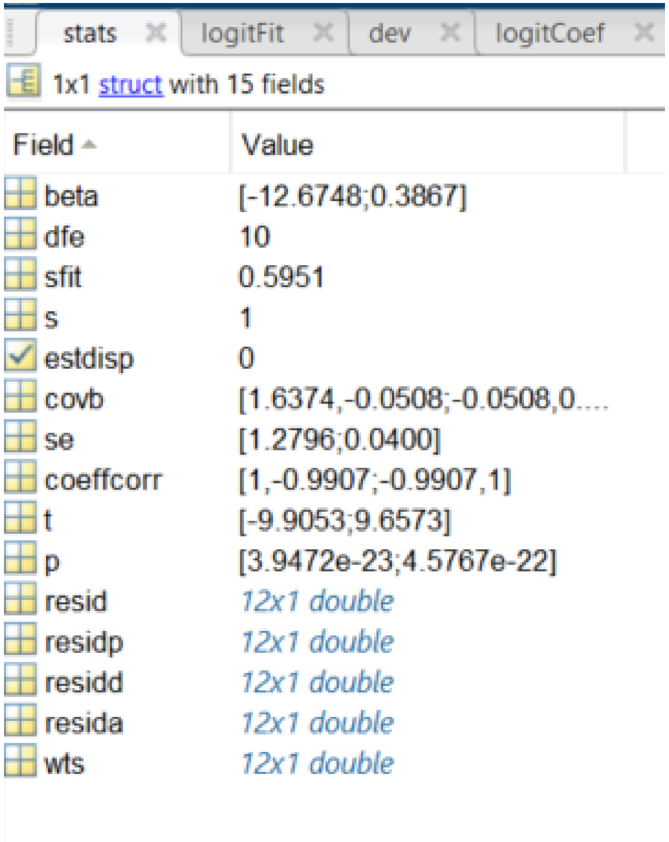
	1	2	3
1	0.0141		
2	0.0391		
3	0.0782		
4	0.1073		
5	0.3345		
6	0.3519		
7	0.5406		
8	0.7831		
9	0.8827		
10	0.9308		
11	0.9725		
12	0.9871		
13			
14			

glmfint: Logistic model coefficients



2x1 double

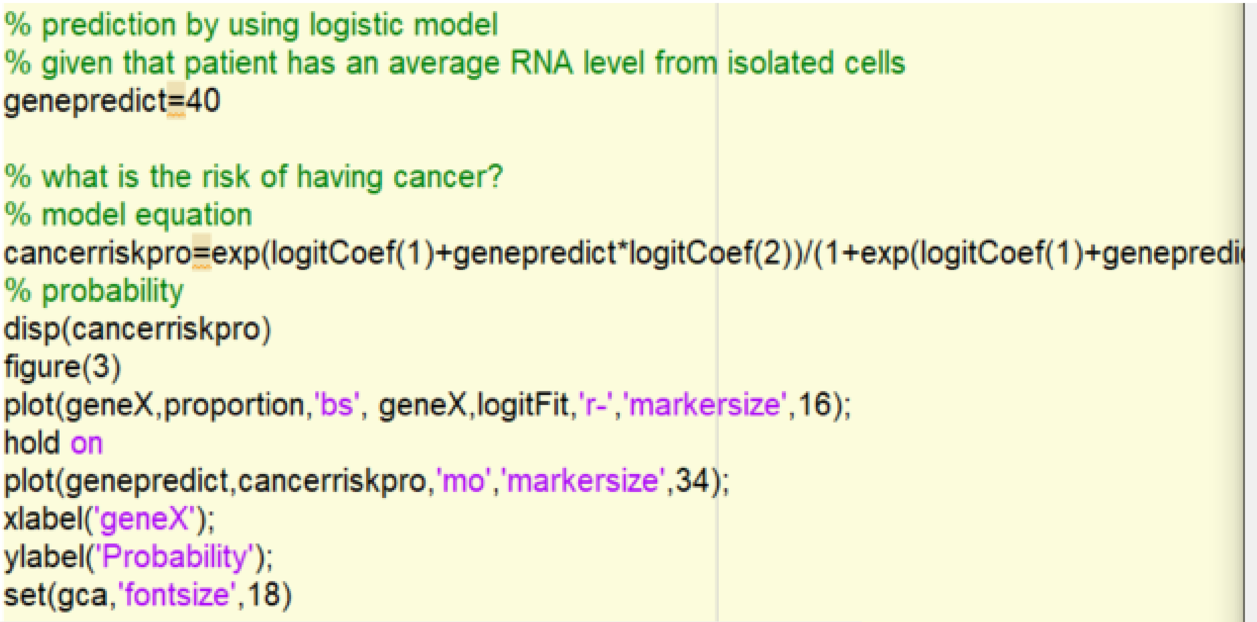
	1	2	3	4	5
1	-12.6748				
2	0.3867				
3					
4					



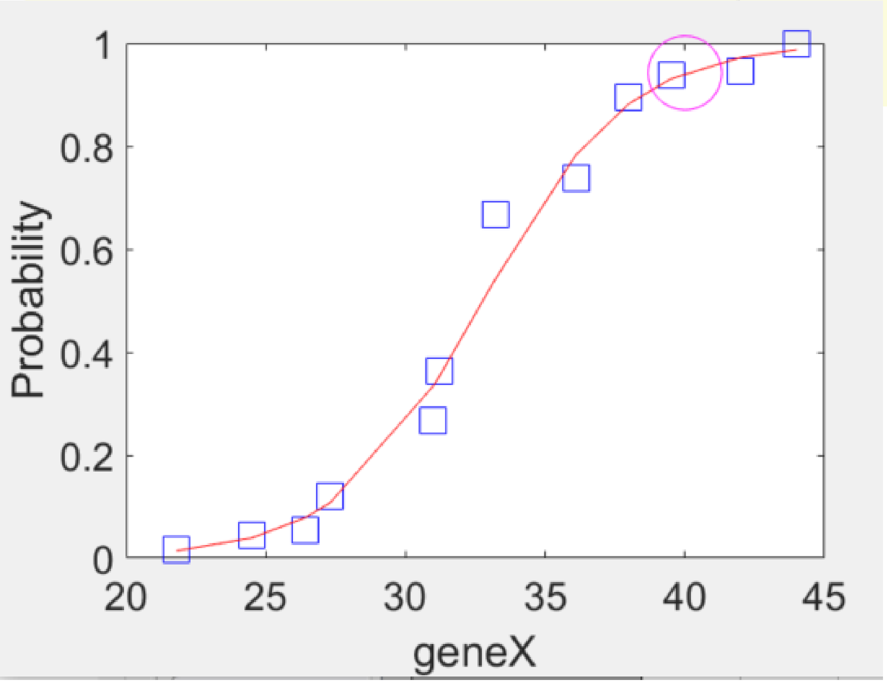
1x1 struct with 15 fields

Field	Value
beta	[-12.6748;0.3867]
dfe	10
sfit	0.5951
s	1
estdisp	0
covb	[1.6374,-0.0508;-0.0508,0....
se	[1.2796;0.0400]
coeffcorr	[1,-0.9907;-0.9907,1]
t	[-9.9053;9.6573]
p	[3.9472e-23;4.5767e-22]
resid	*12x1 double*
residp	*12x1 double*
residd	*12x1 double*
resida	*12x1 double*
wts	*12x1 double*

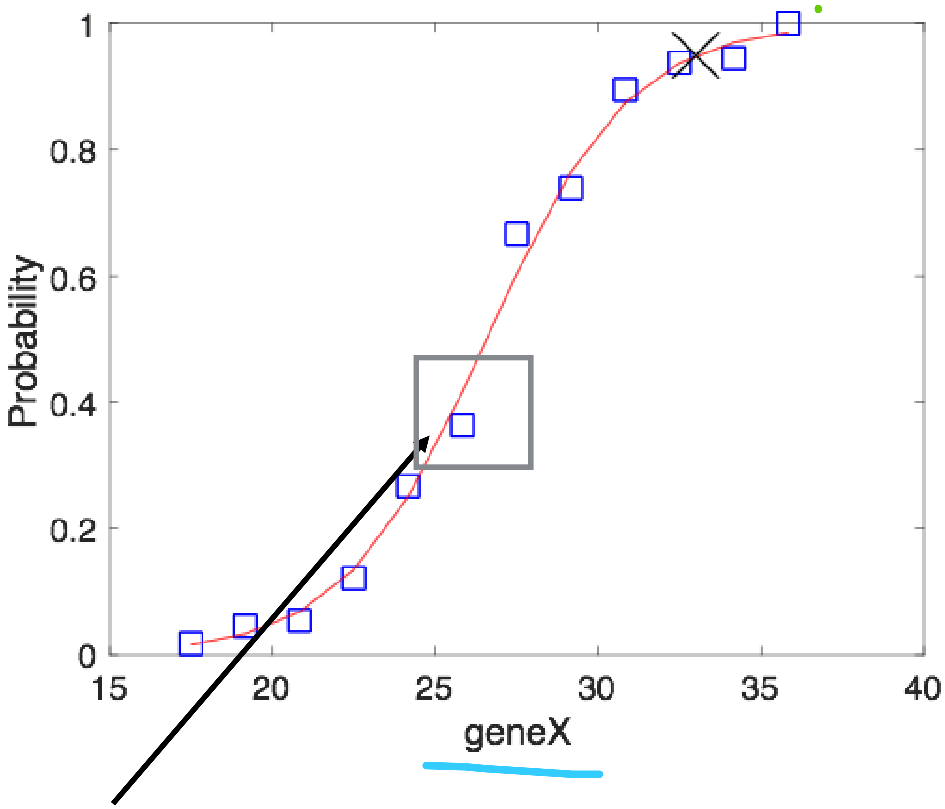
P(Y=1|beta)= exp(b(1)+b(2)x) / 1+exp(b(1)+b(2)x)

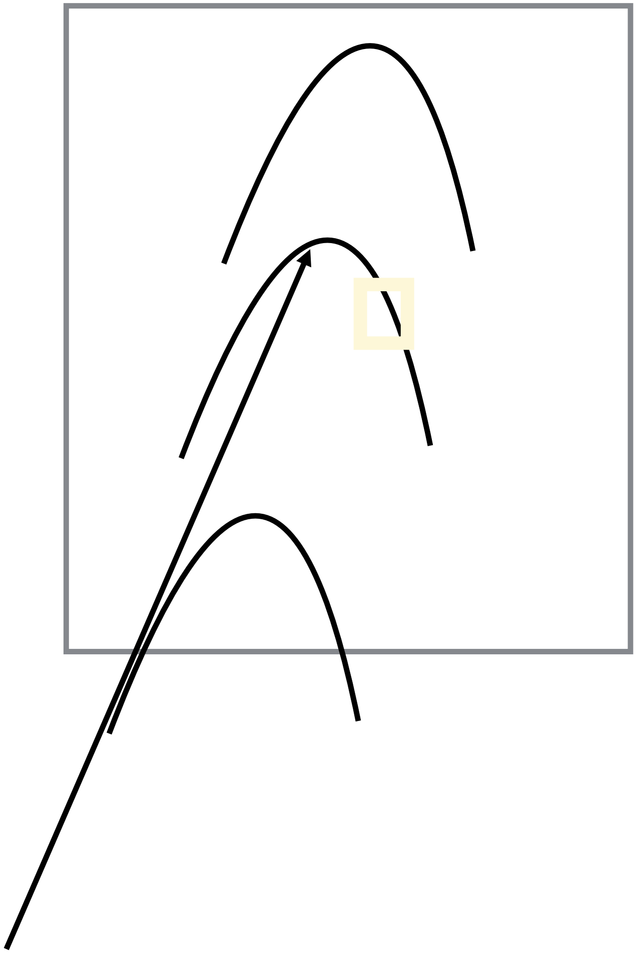


```
% prediction by using logistic model
% given that patient has an average RNA level from isolated cells
genepredict=40

% what is the risk of having cancer?
% model equation
cancerriskpro=exp(logitCoef(1)+genepredict*logitCoef(2))/(1+exp(logitCoef(1)+genepredi
% probability
disp(cancerriskpro)
figure(3)
plot(geneX,proportion,'bs', geneX,logitFit,'r-','markersize',16);
hold on
plot(genepredict,cancerriskpro,'mo','markersize',34);
xlabel('geneX');
ylabel('Probability');
set(gca,'fontsize',18)
```



Coefficients are estimated by using a maximum likelihood estimation method where coefficients maximizes the prediction of observed values in the data

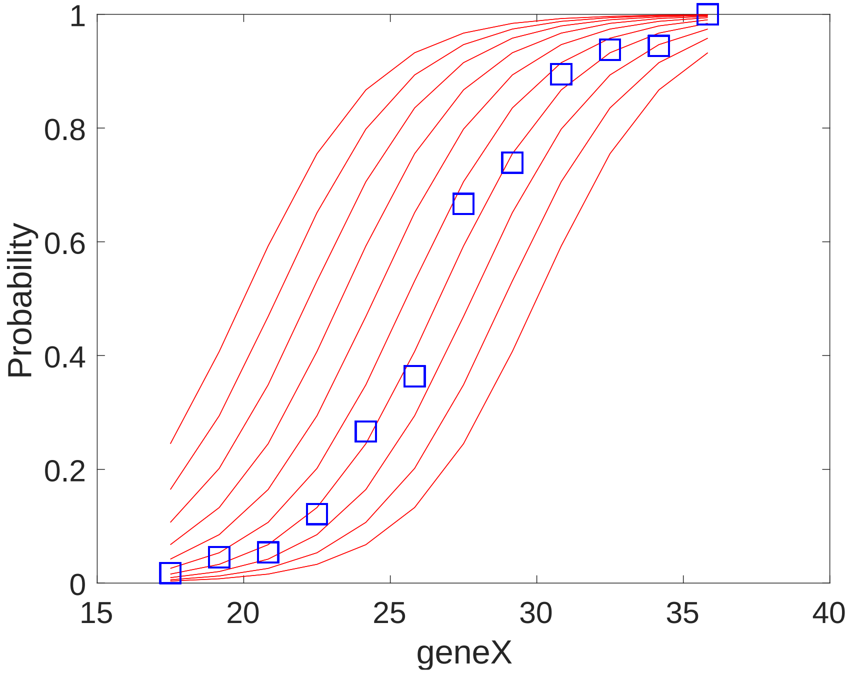


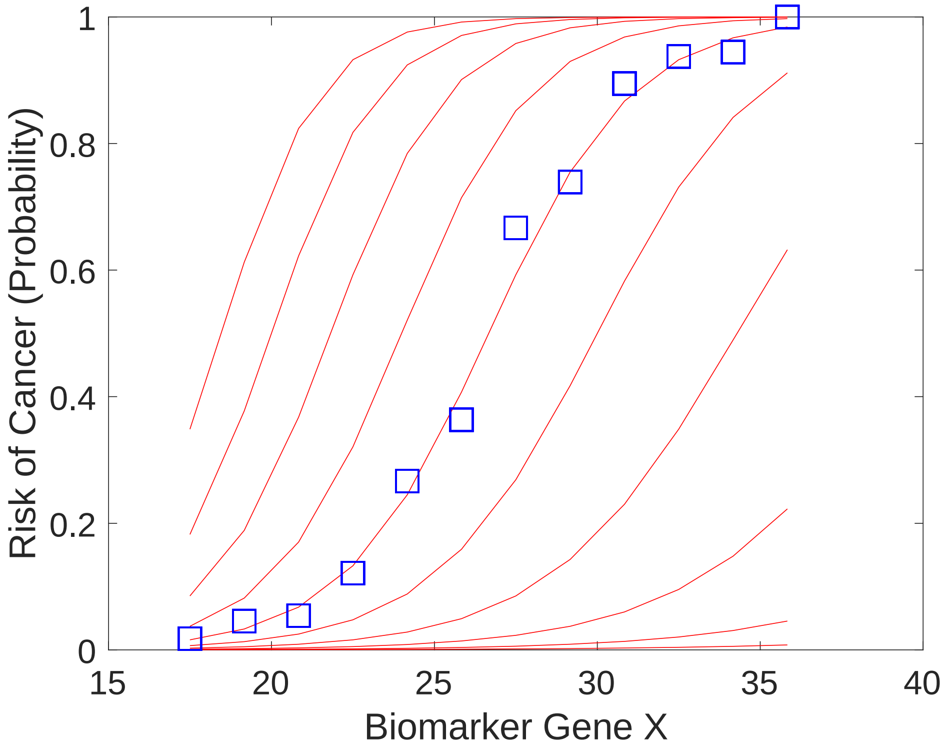


points on a line represents the highest points in the probability distribution

$$\log(\text{odds}) = b_0 + b_1 x = -12.12 + 0.45x$$

The effect of coefficients on the shape of logistic model


Probability
geneX


Risk of Cancer (Probability)
Biomarker Gene X