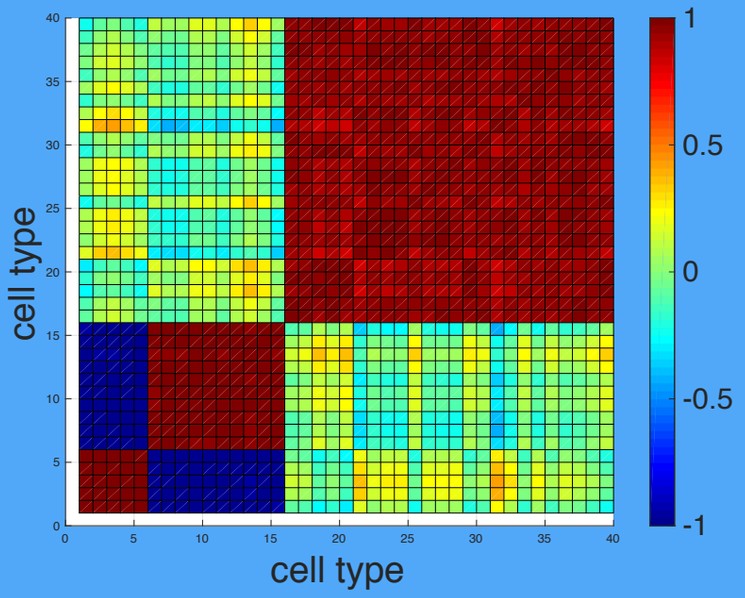
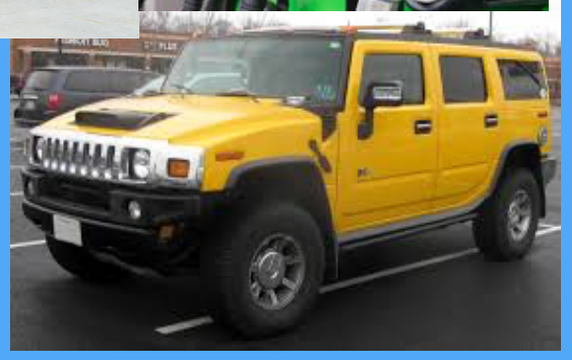
Introduction to Scientific Computation

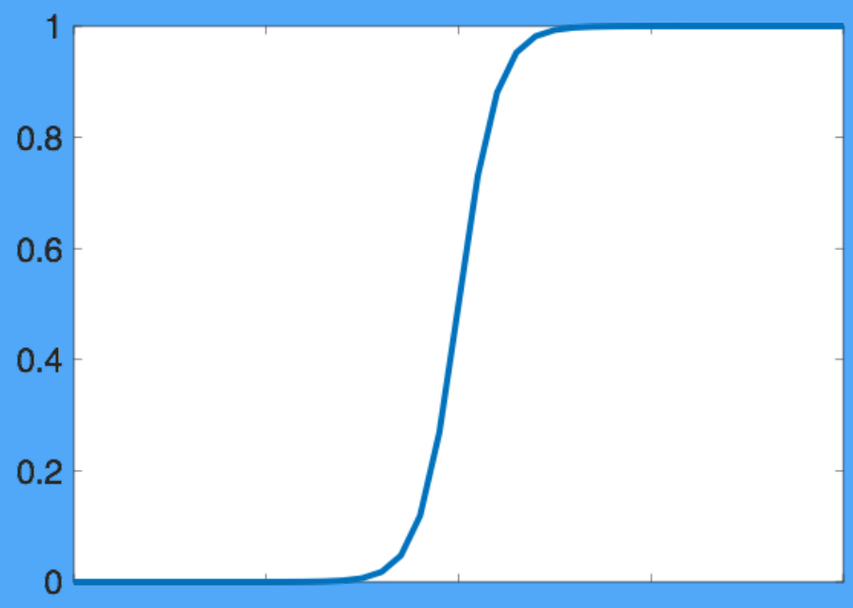
Halil Bayraktar

Lecture 9 – Multivariable regression and Logistic Regression





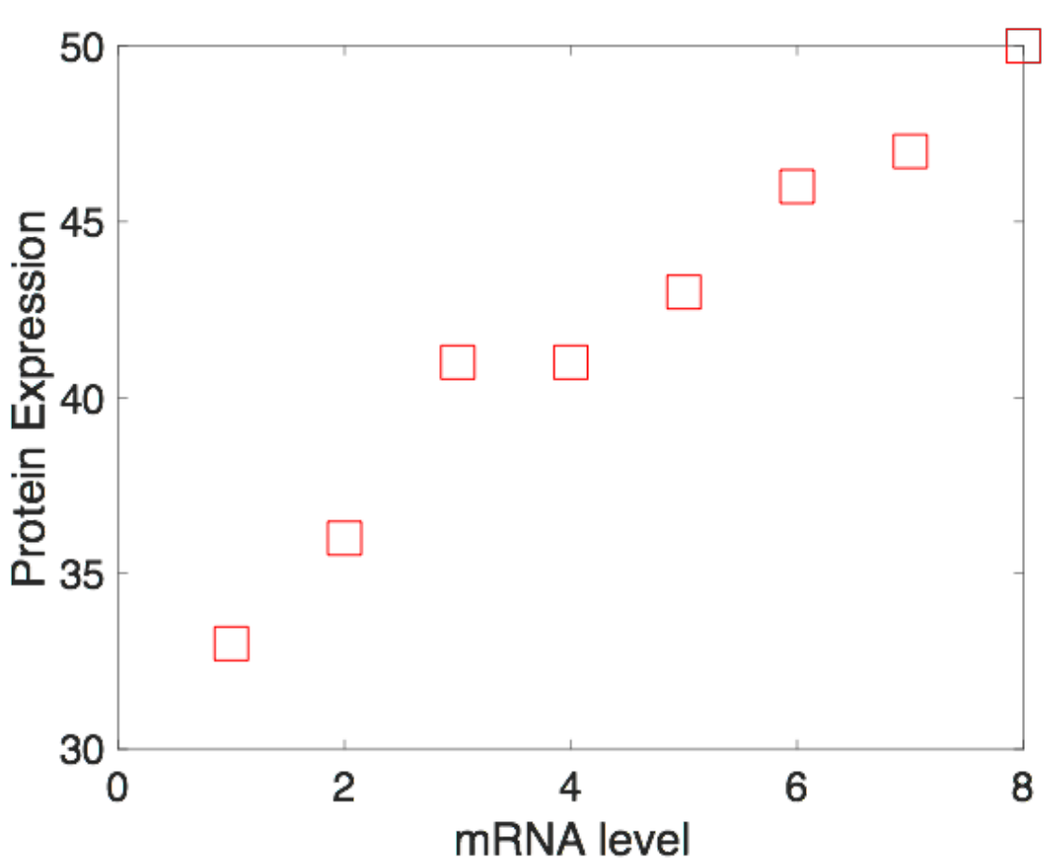
```
Linear regression model:
    y ~ 1 + x1 + x2 + x3

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
>>
```

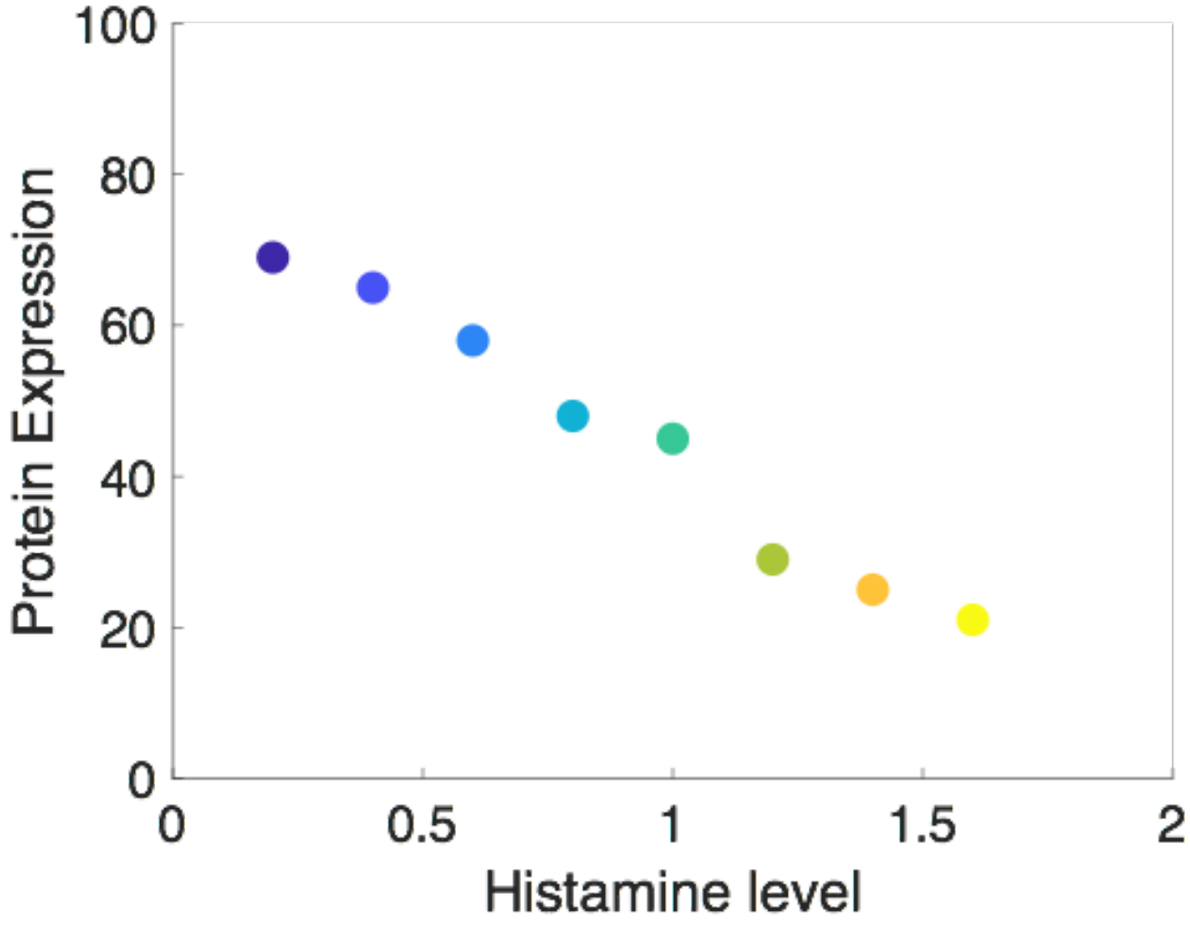


Two variable vs multivariable regression

Scatter plot

Shows the relation between two variables



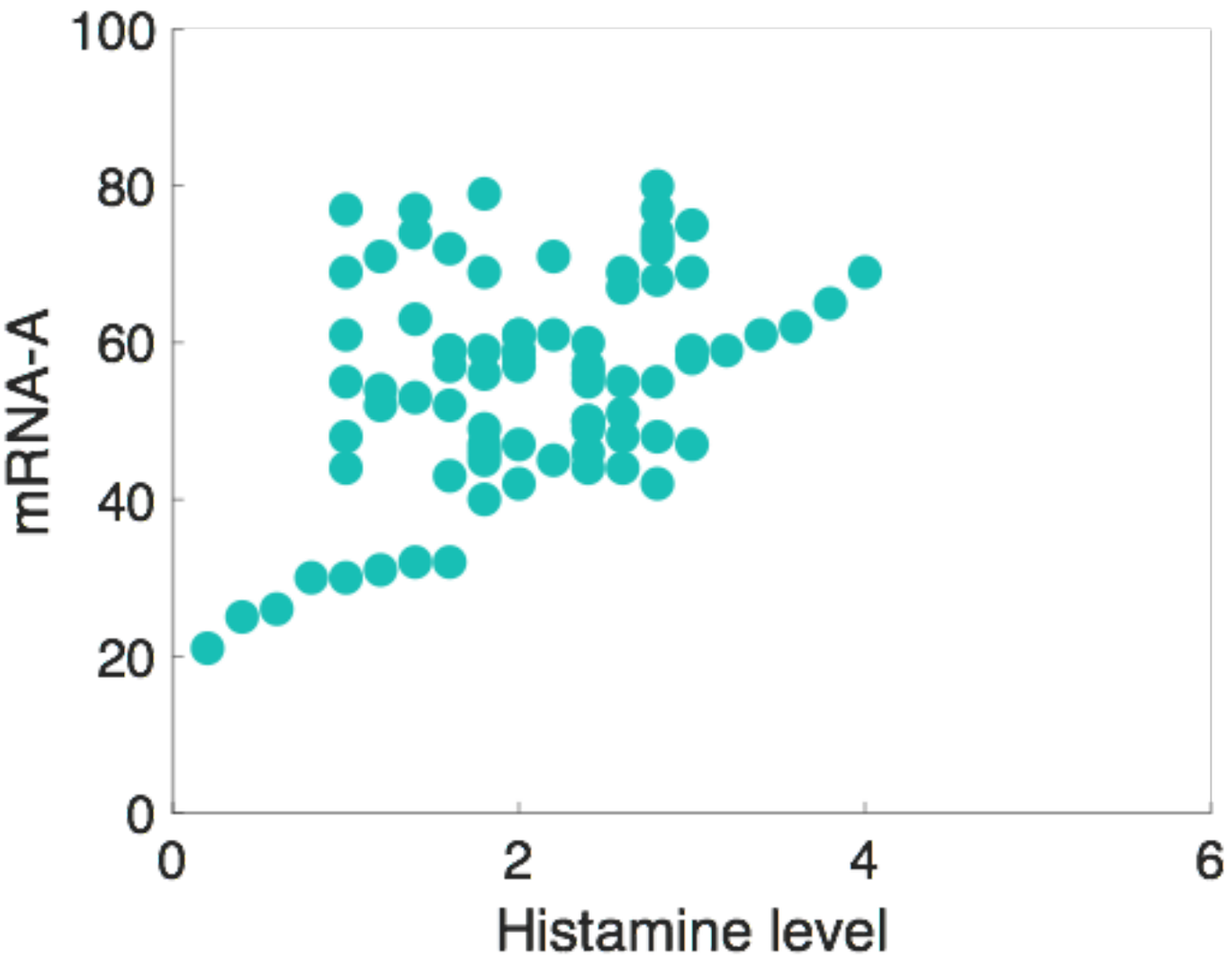


Can we quantitatively measure the strength of relationship between variables?

Covariance and Correlation

They are indicators of how strong relationship is present between two variables

Is it a positive or negative association?

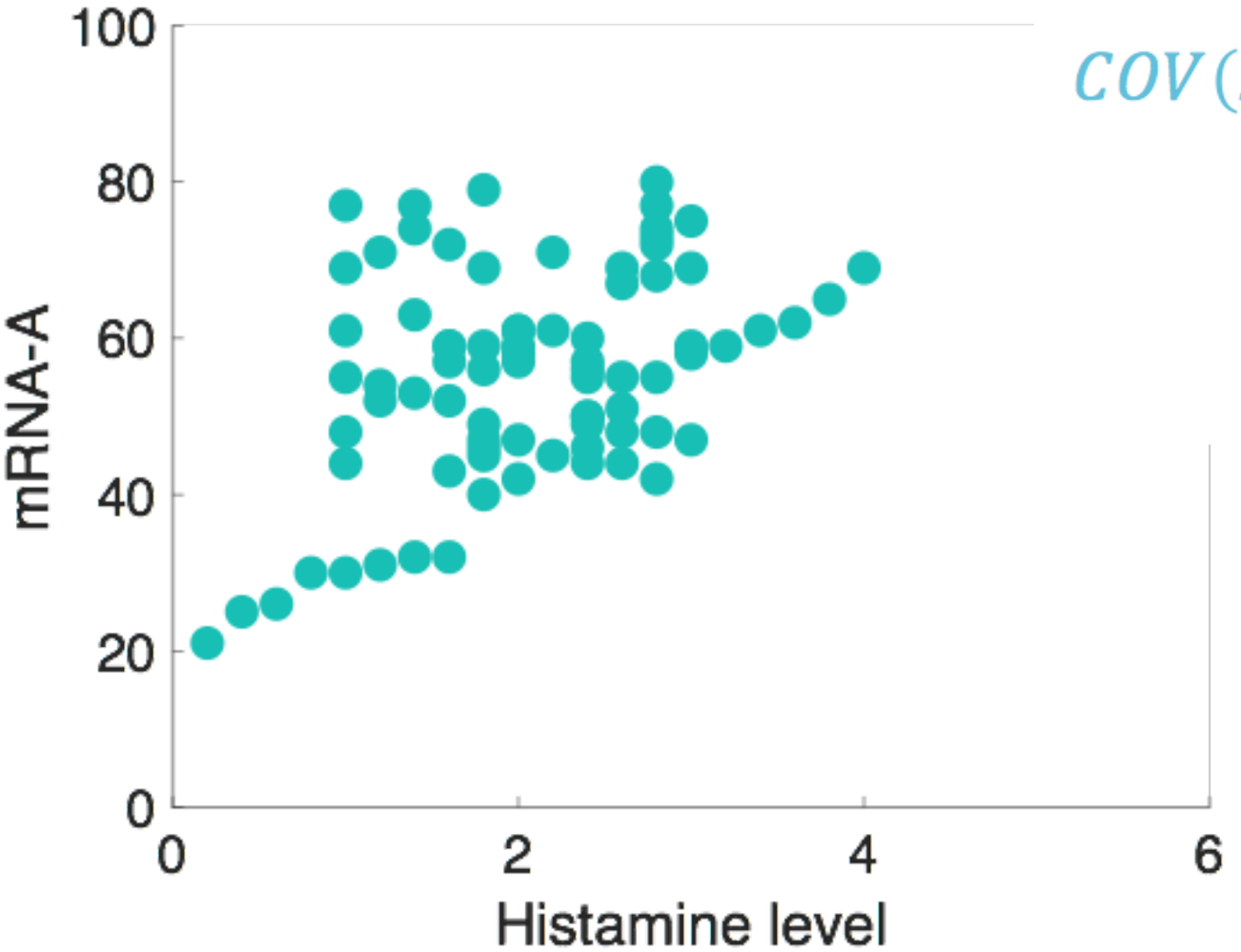


n=80

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

Covariance

Does Y get larges (smalleR) as Y increase?



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

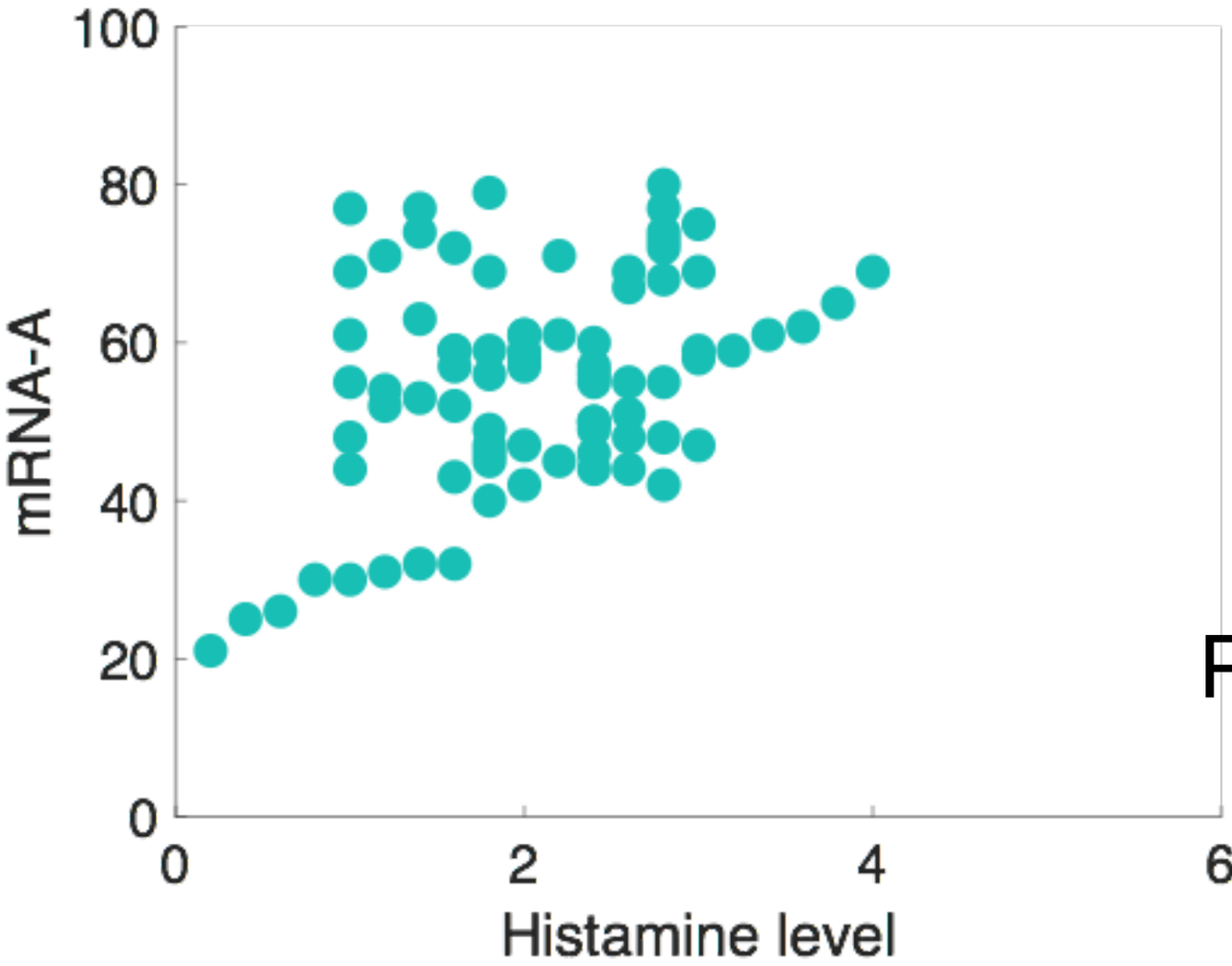
mean of x= 2.0525
mean of y = 55.4125
Sx = 0.7916
Sy= 13.6537

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.

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

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

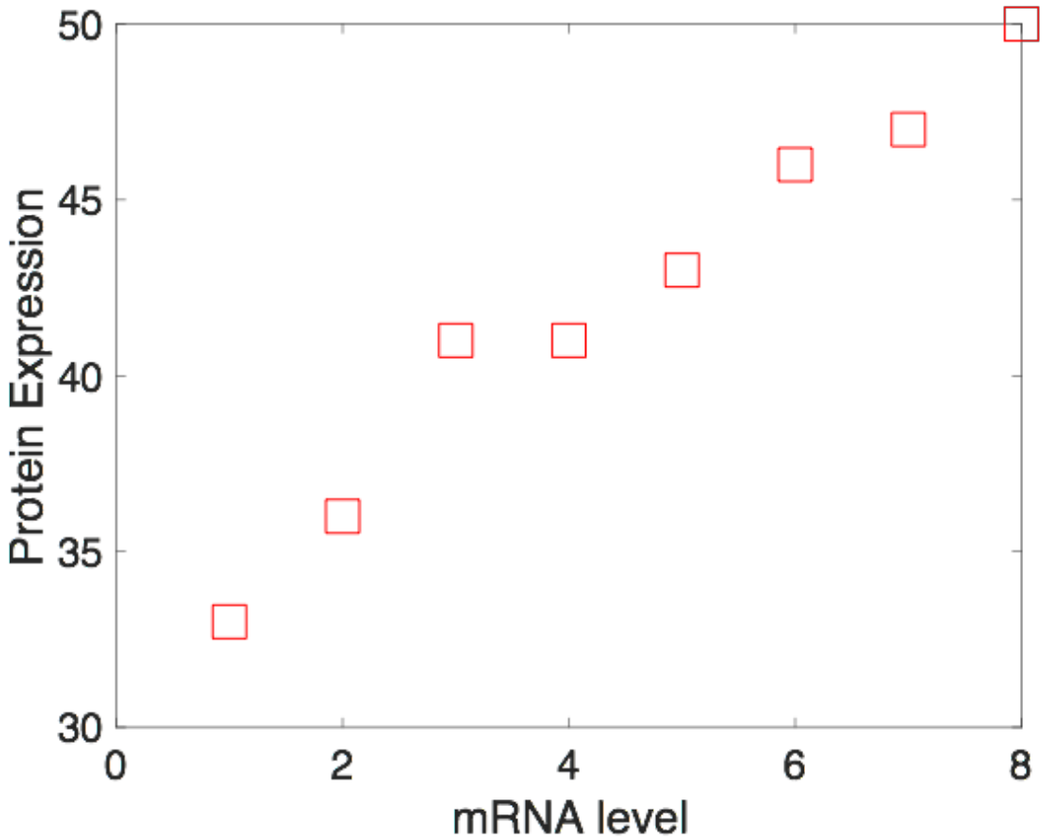
close to n-1 if x and y have

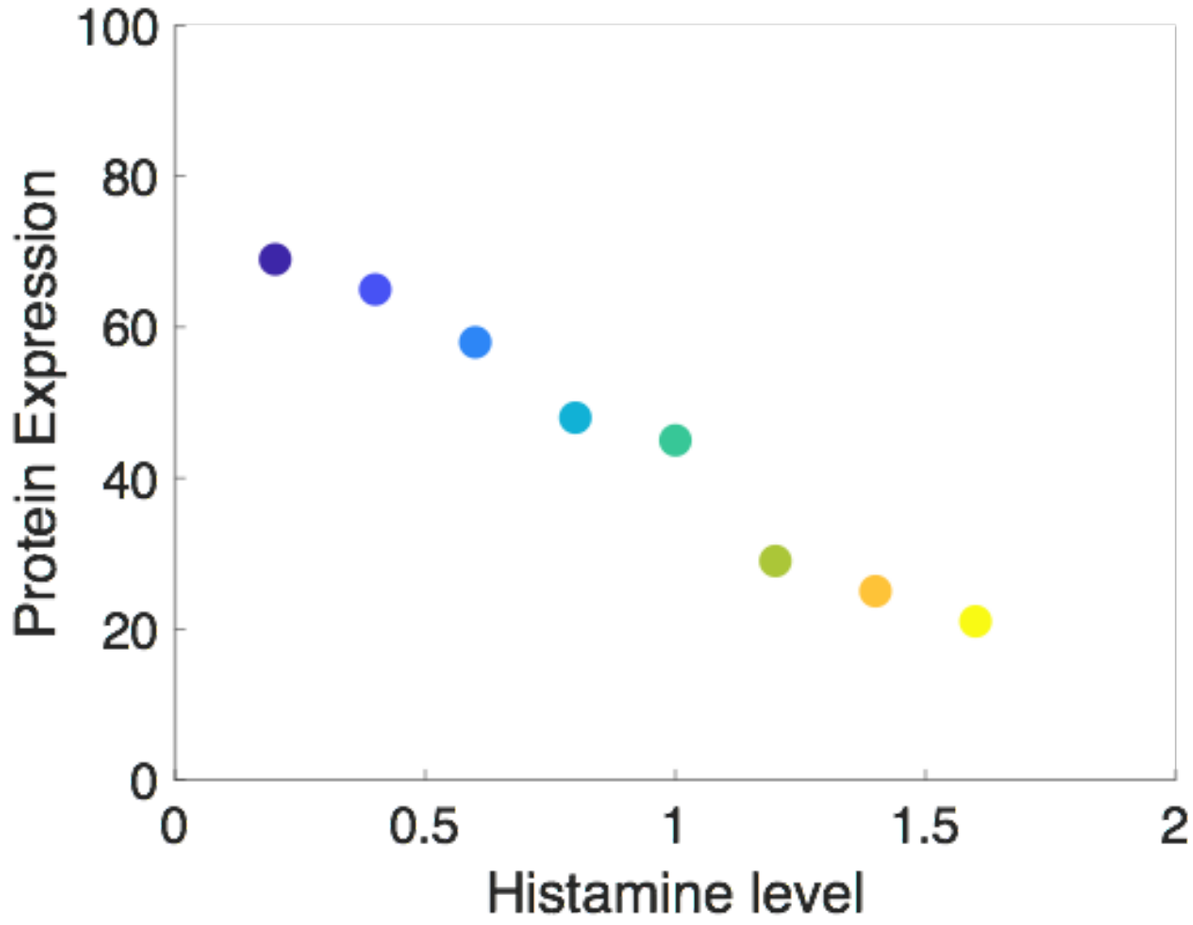
$\bar{x}$ = the sample mean of $x_1,...,x_n$,

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

s_x = the standard deviation of $x_1, \ldots, x_n$,

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





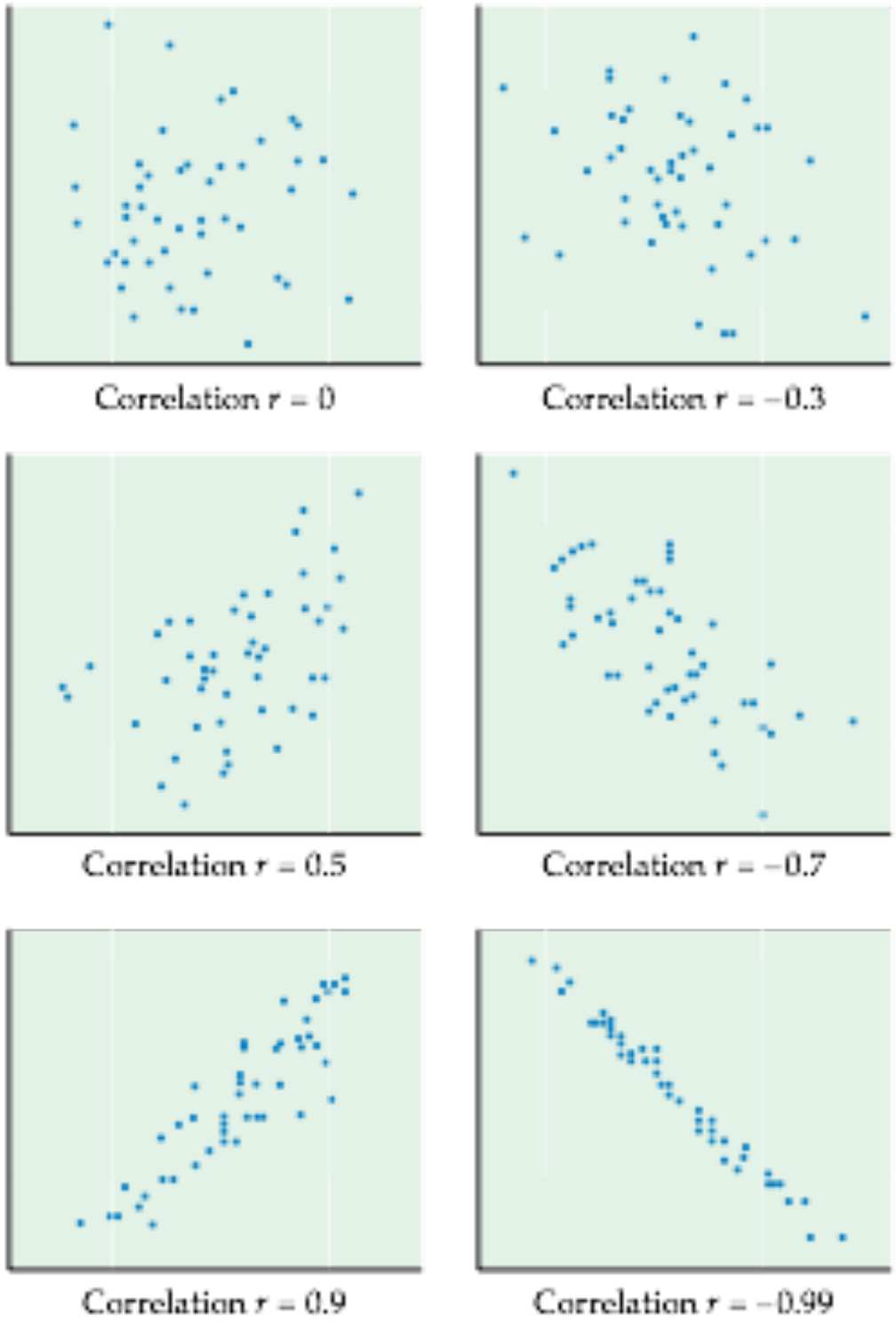
r=+0.9538

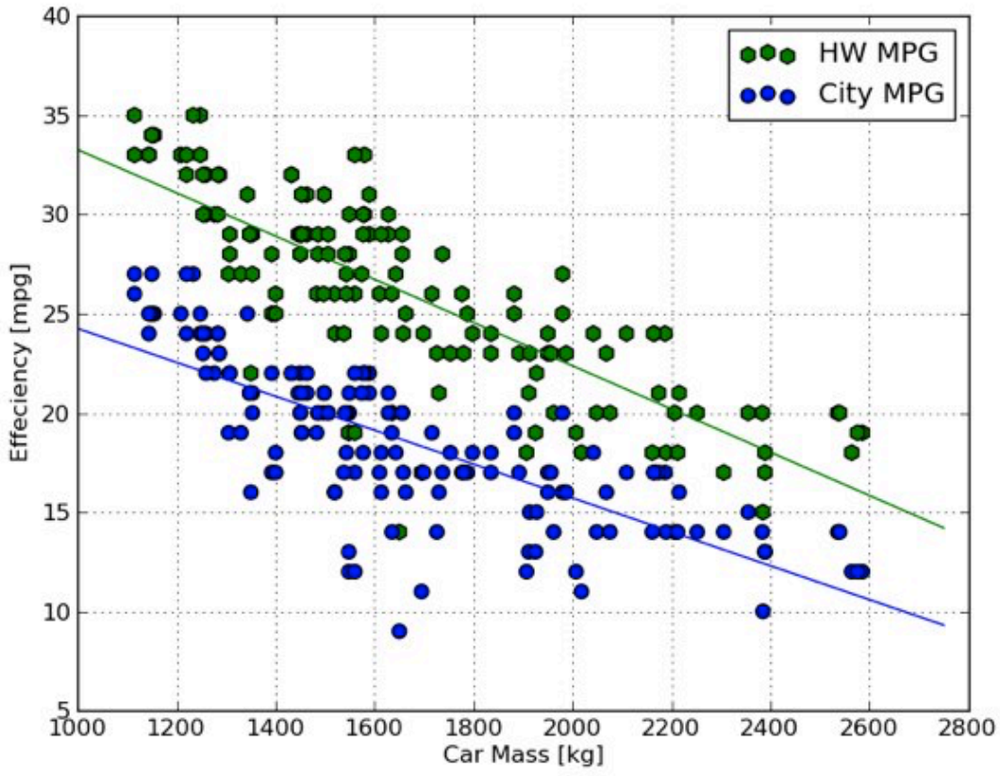
r=-0.987

Correlation coefficient always lies between -1 to +1

Types of correlation

Positive, negative and no correlation





Strong negative relationship between car weight and fuel consumption.

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).

Summary of Correlation and Covariance

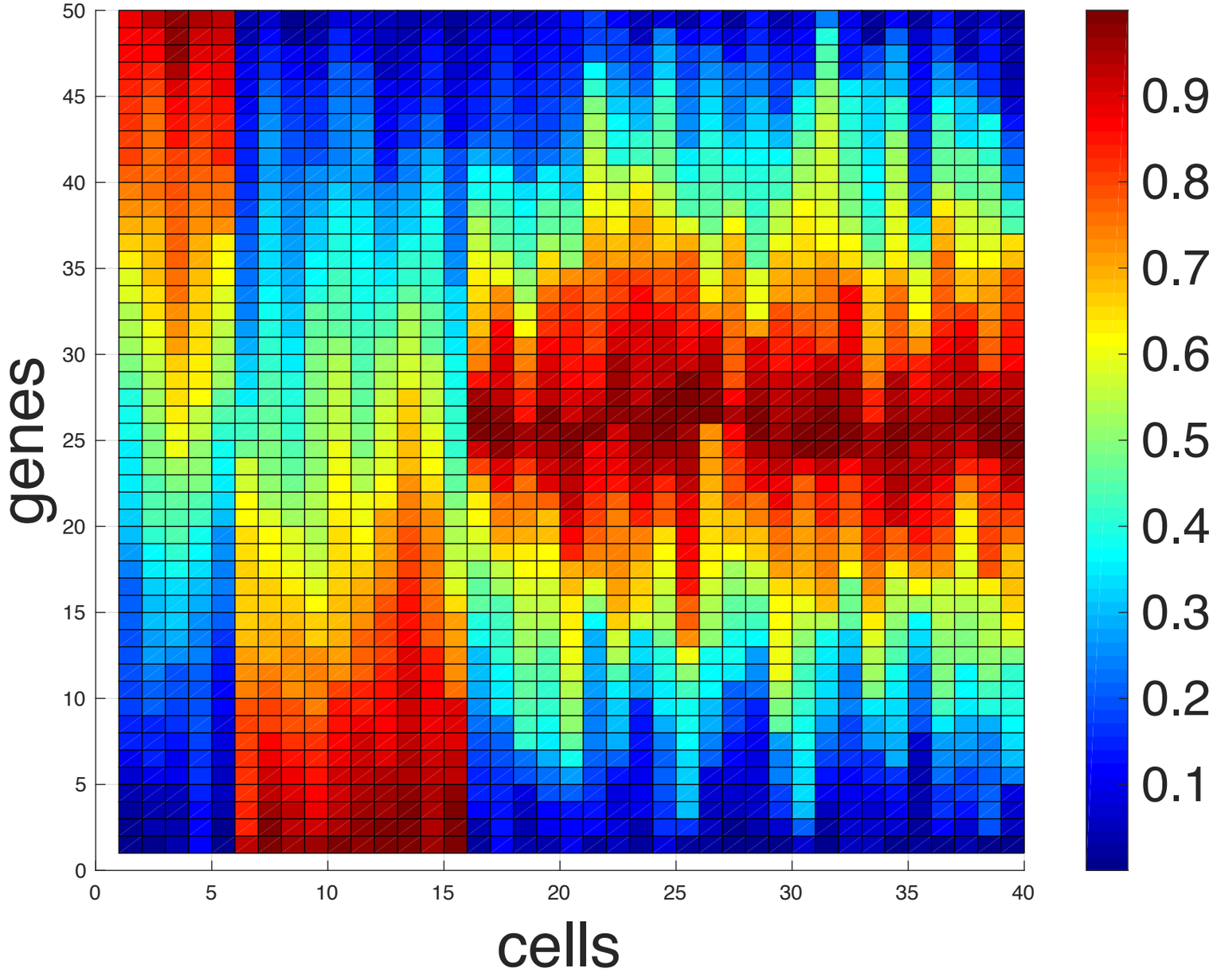
Correlation ranges from -1 to 1
Covariance can be any number

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

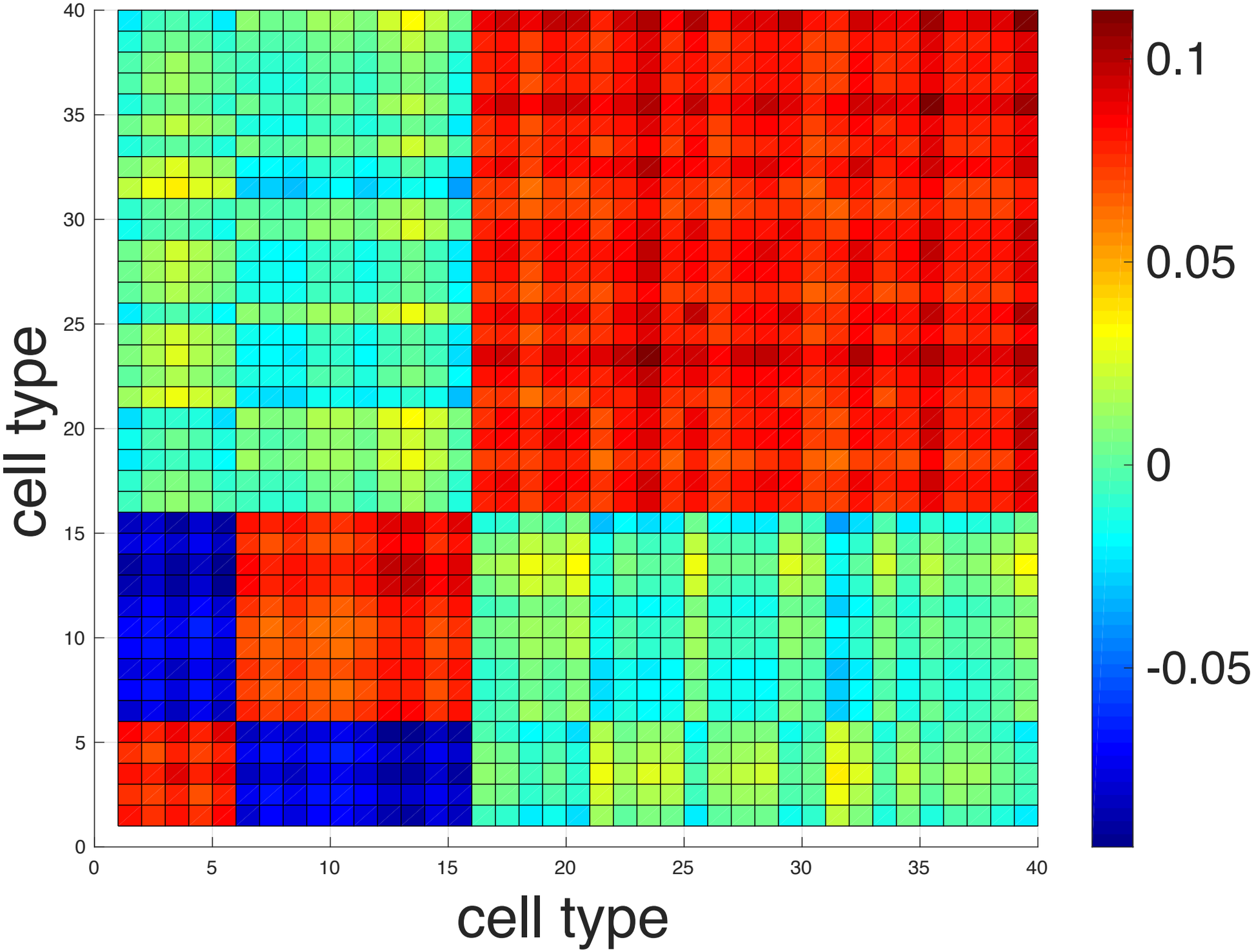
if there are outliers or a nonlinear relation, the results can be wrong!

Correlation does not imply causation!

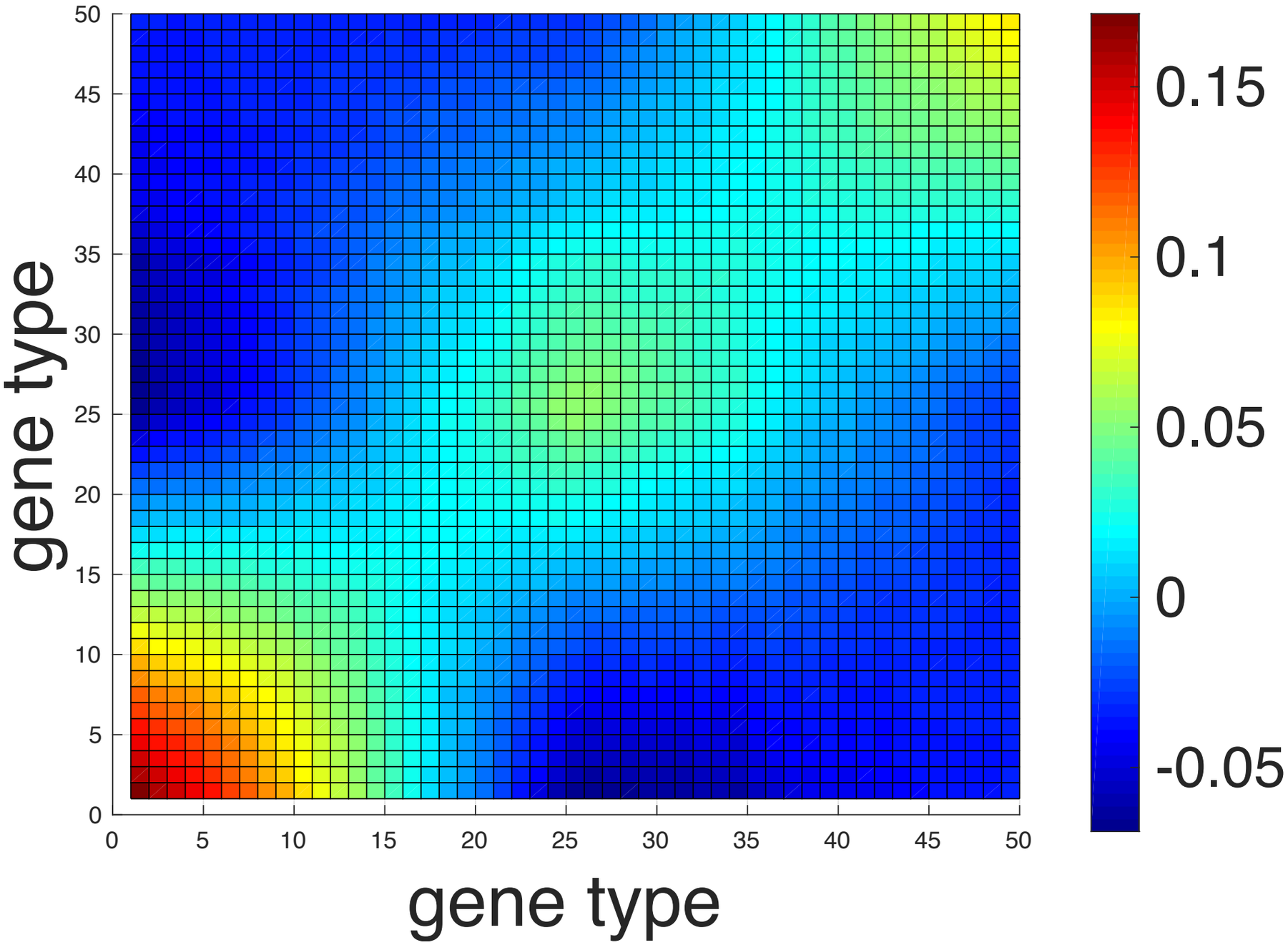
Gene expression in different cells



What is the covariance between different cell types?

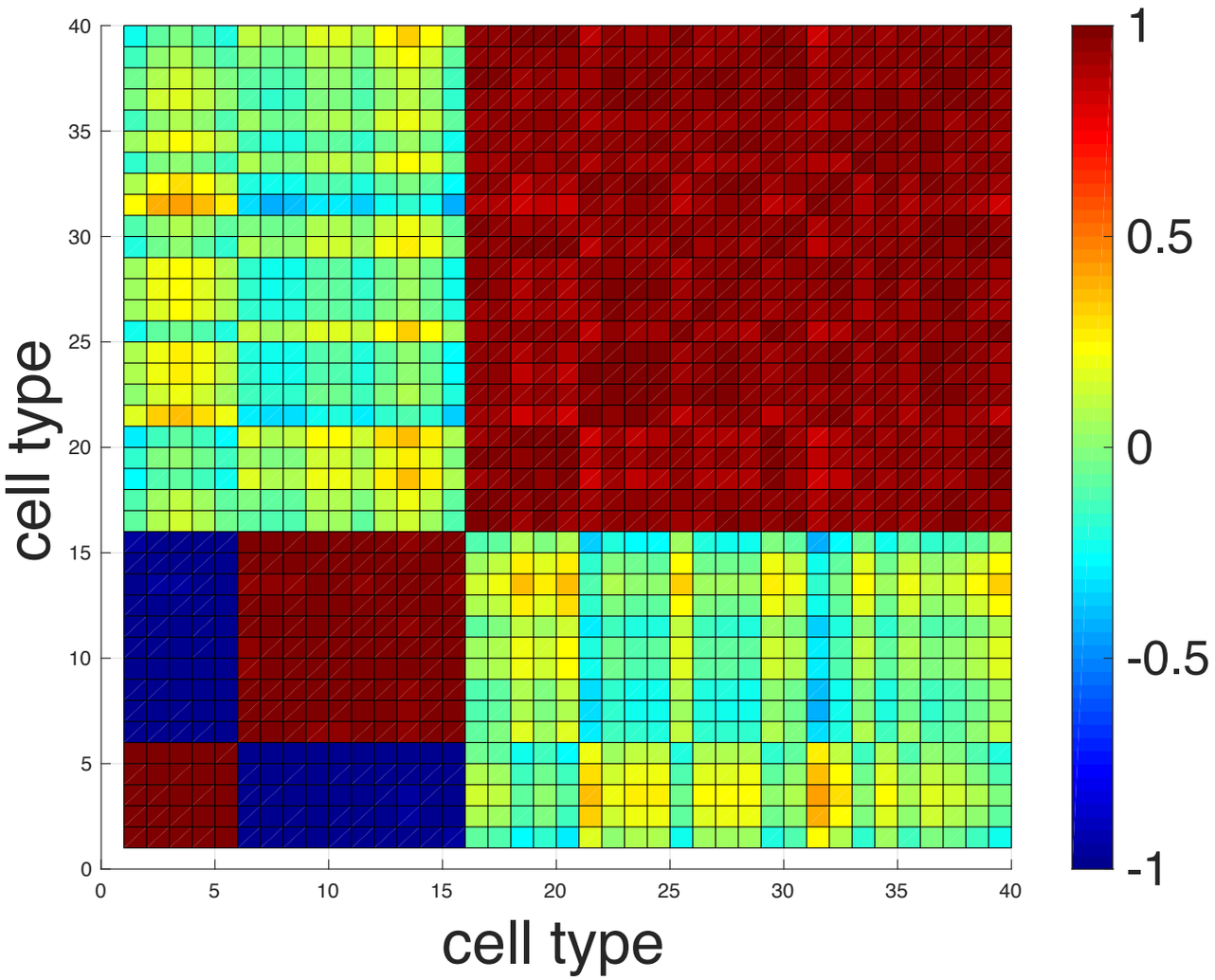


What is the covariance between different genes?



Correlation of difference cell types

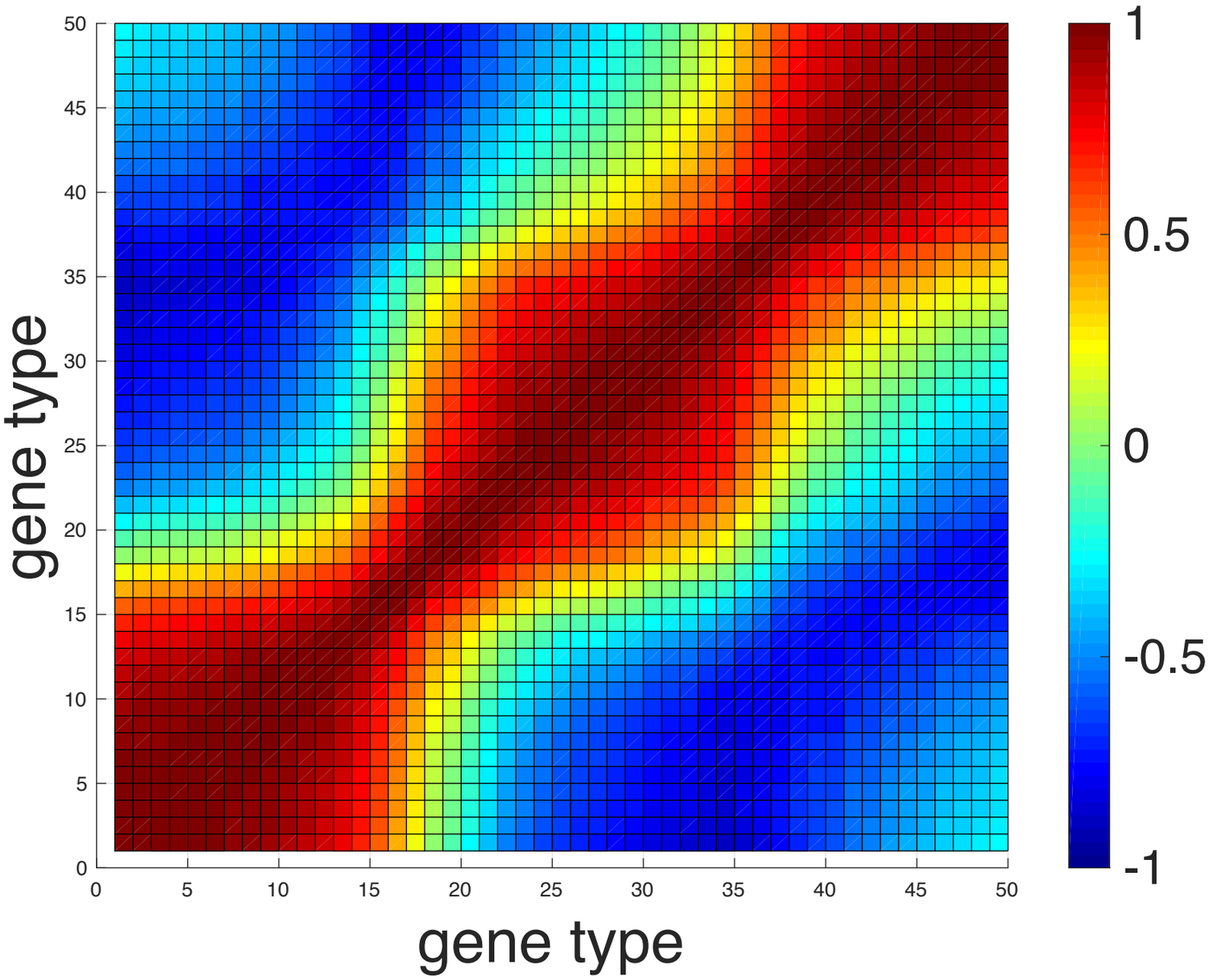
Liver cells, kidney cells and neurons



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

Covariance returns the direction of relation while the correlation

Correlation of difference genes



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

- $x_1, x_2, x_3 \ldots, x_n$ are called independent variables

or explanatory variables.

- X values can either quantitative or categorical.

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

The **statistical model for multiple linear regression** is

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

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

Parameter coefficients of the model are β_0, β_1, $\beta_2,\ldots,\beta_p$, and σ.

For the ith 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 = observed response − 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)

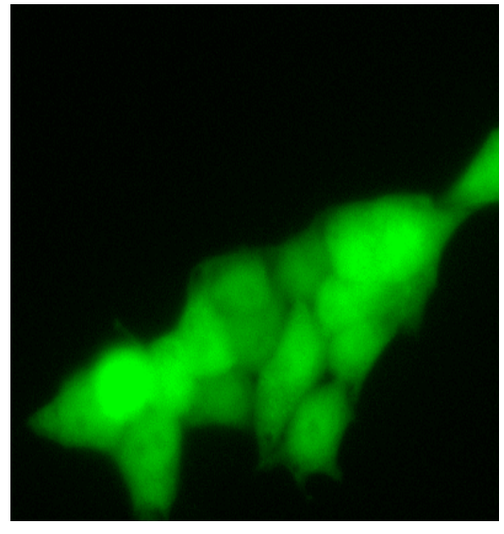
4. Dependence of hormone levels to genes (health)

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

In Matlab

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

Dependence of hormone levels 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
>>

Examples hormones are thyroid hormone, steroids and retinoic acid.

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

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

Test for regression coefficients and predict cell growth

We conclude that geneX and geneZ contain useful information for predicting cell growth.

Lets predict cell growth

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

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

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

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

Relationship between pairs of variables

Dependence of cell growth to mRNA expression features,
Dependence of average exam score to study time, exercise
Time etc.

Step 1. Determine the relationship between all pairs

Compute correlation and use scatter plots

Step 2. Compute correlation values, P values

For example we see that the correlation between cell growth and geneB is 0.65, with a *P*-value of 0.0001, whereas the correlation between cell growth and GeneA is 0.13, with a *P*-value of 0.076. Former one is significant

Thus, we see that the correlation between cell growth and geneB is 0.64 with P value 0.0005 and gene C 0.58 with P value of 0.0003 are cell growth and GeneD is 0.13, with a *P*-value of 0.076

What does that mean?

The first one (geneB and C) is statistically significant, and the third is barely significant.

we can say that gene B and C all have higher correlations with cell growth than do the geneD.

Step 3. Check if genes have any correlation

gene B and gene C have high correlation 0.56, gene D have low correlation with both B and C

Regression analysis of cell growth

To explore the relationship between the explanatory variables and our response variable such as cell growth etc. , multiple regressions can be performed.

Suppose that We select geneB and Gene C

ANOVA F statistic is 18.86, with a P-value of 0.0001.

Under the null hypothesis

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

at least one of the two regression coefficients

for cell growth is different from 0 in the population regression equation

Example 2.

Dependence of fuel consumption to car features (weight, horse power, model year etc.)

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

Estimated Coefficients:
                   Estimate        SE        tStat        pValue
                   ________    ________    _________    __________

    (Intercept)      47.977      3.8785       12.37     4.8957e-21
    x1              -6.5416      1.1274      -5.8023     9.8742e-08
    x2              -0.042943    0.024313    -1.7663       0.08078
    x3              -0.011583    0.19333     -0.059913     0.95236


Number of observations: 93, Error degrees of freedom: 89
Root Mean Squared Error: 4.09
R-squared: 0.752,  Adjusted R-Squared 0.744
F-statistic vs. constant model: 90, p-value = 7.38e-27
>>
```

A one-unit difference in the rating of weight corresponds to a 6.5 point difference in fuel consumption.

What does car model year affect the fuel consumption?

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    ________    ________    _________

    (Intercept)      -7.5209      7.2521    -1.0371      0.30255
    x1               -7.0765     0.84904    -8.3347    9.5728e-13
    x2             0.0057071    0.019164     0.29781       0.76655
    x3             -0.058512     0.14528    -0.40274       0.68812
    x4               0.68988     0.08256      8.3561    8.6529e-13


Number of observations: 93, Error degrees of freedom: 88
Root Mean Squared Error: 3.07
R-squared: 0.862,  Adjusted R-Squared 0.855
F-statistic vs. constant model: 137, p-value = 6.01e-37
```