

STEVE SCHERRER

DEMYSTIFYING ANOVA

Linear Regression Review

Univariate Regression

$$y = mx + B$$

Univariate Regression

$$y = mx + B$$

slope

A diagram illustrating the components of the univariate regression equation $y = mx + B$. The equation is displayed in a large, bold, light gray font. Below the equation, the word "slope" is written in a bold, orange-red font. A thin, light orange-red arrow points from the word "slope" to the coefficient m in the equation. Similarly, the word "intercept" is written in a bold, orange-red font to the right of "slope". A thin, light orange-red arrow points from the word "intercept" to the constant term B in the equation.

intercept

Univariate Regression

Predicted Values

x values

$$f(x) = \beta_1 X + \beta_0 + \varepsilon$$

slope

intercept

error

Univariate Regression

Predicted values

Values - Feature

$$f(x) - \varepsilon = \beta_1 X + \beta_0$$

error

slope

intercept

Univariate Regression

Values - Y (target) Values - Feature

The diagram illustrates the univariate regression equation $f(x) - \epsilon = \beta_1 X + \beta_0$. A red bracket above the left side of the equation is labeled "Values - Y (target)". A red arrow points from the label "Values - Feature" to the X term. Another red arrow points from the label "slope" to the coefficient β_1 . A final red arrow points from the label "intercept" to the coefficient β_0 .

$$f(x) - \epsilon = \beta_1 X + \beta_0$$

slope intercept

Multivariate Regression

The diagram illustrates the multivariate regression equation $f(x) - \varepsilon = \beta_1 X_1 + \beta_2 X_2 + \beta_0$. A red curly brace is positioned above the left side of the equation, labeled "values - y (target)". Red arrows point from descriptive labels to the coefficients: "values - Feature 1" points to β_1 , "values - Feature 2" points to β_2 , "slope - Feature 1" points to β_1 , "slope - Feature 2" points to β_2 , and "intercept" points to β_0 .

values - y (target)

values - Feature 1

values - Feature 2

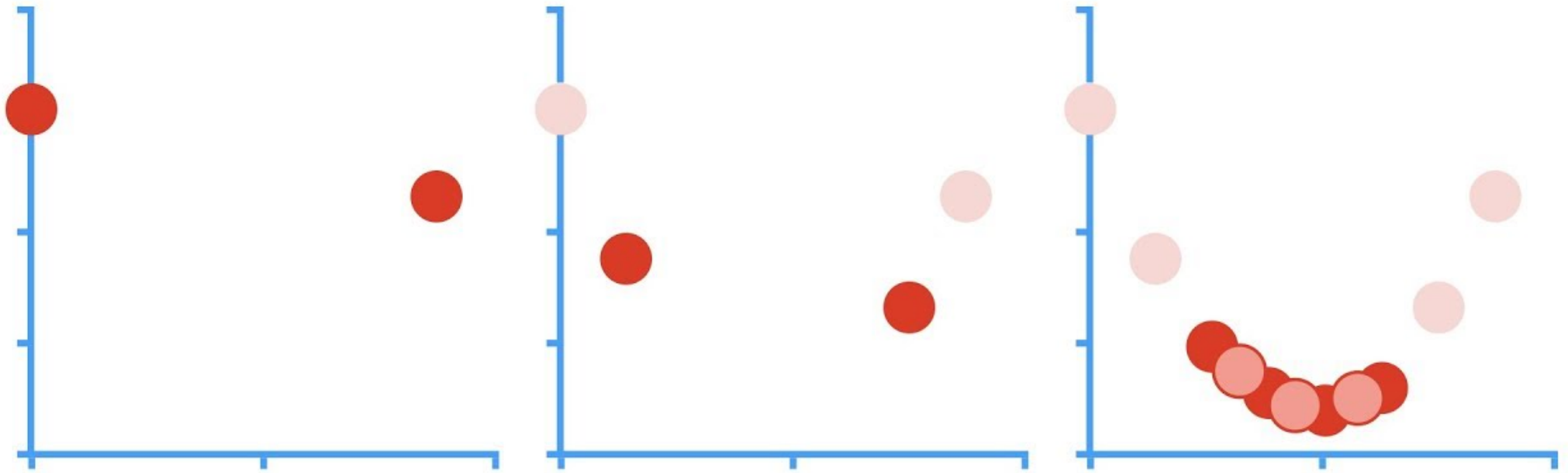
$$f(x) - \varepsilon = \beta_1 X_1 + \beta_2 X_2 + \beta_0$$

slope - Feature 1

slope - Feature 2

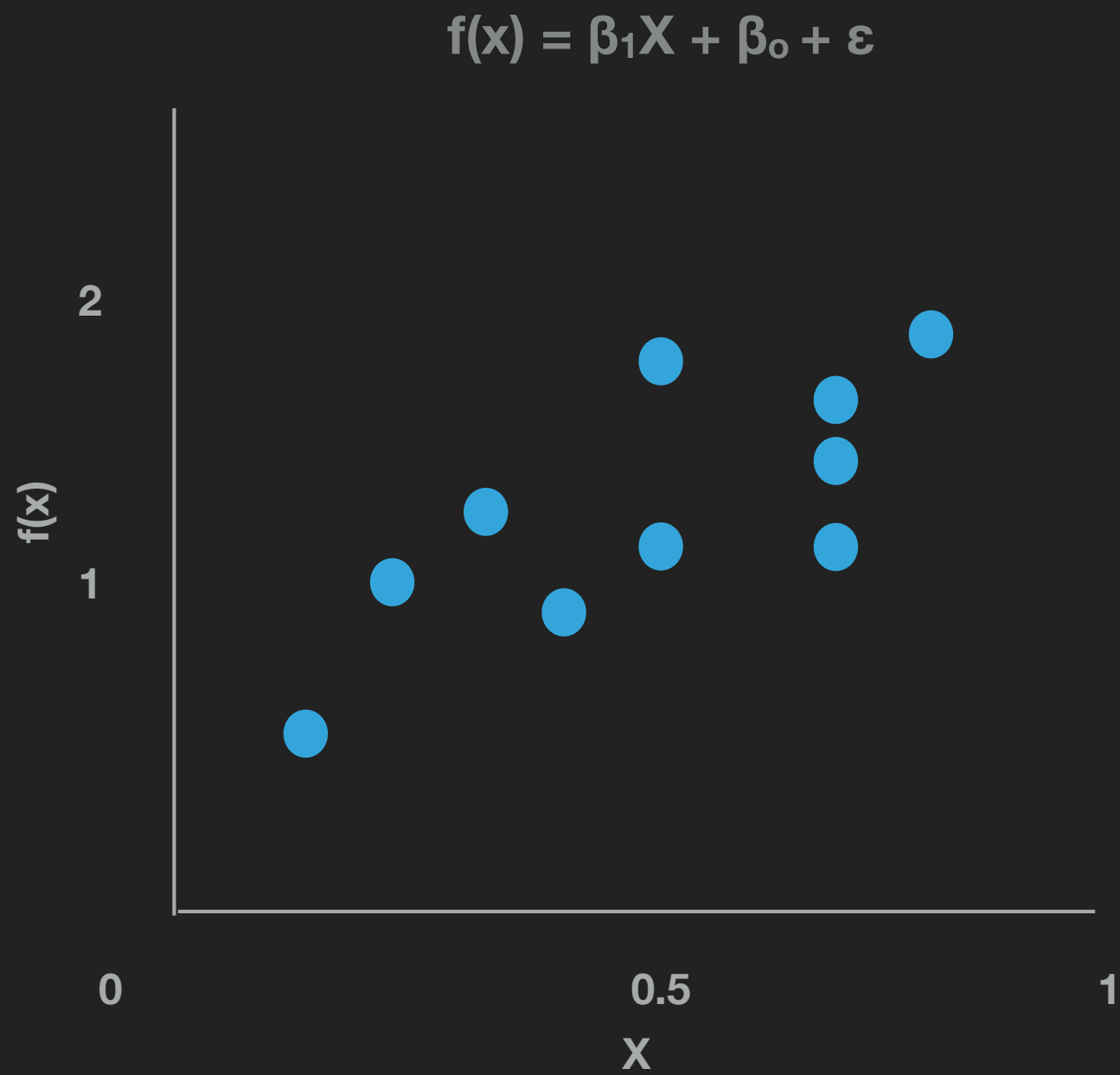
intercept

Gradient Descent....



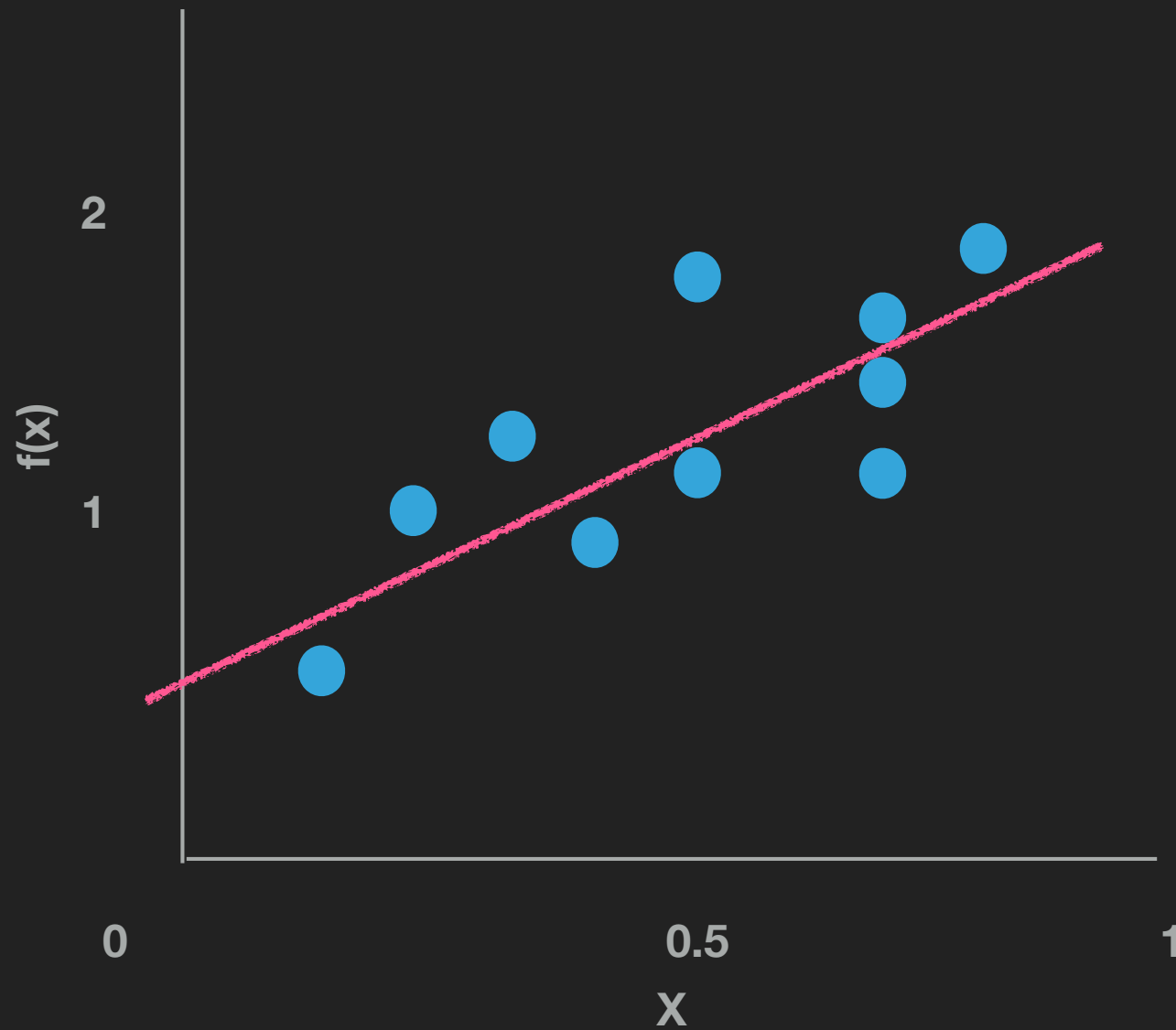
...Step-by-Step!!!

y	x
0.5	0.1
0.9	0.4
1.0	0.2
1.1	0.7
1.1	0.5
1.2	0.3
1.5	0.7
1.6	0.7
1.7	0.5
1.8	0.8



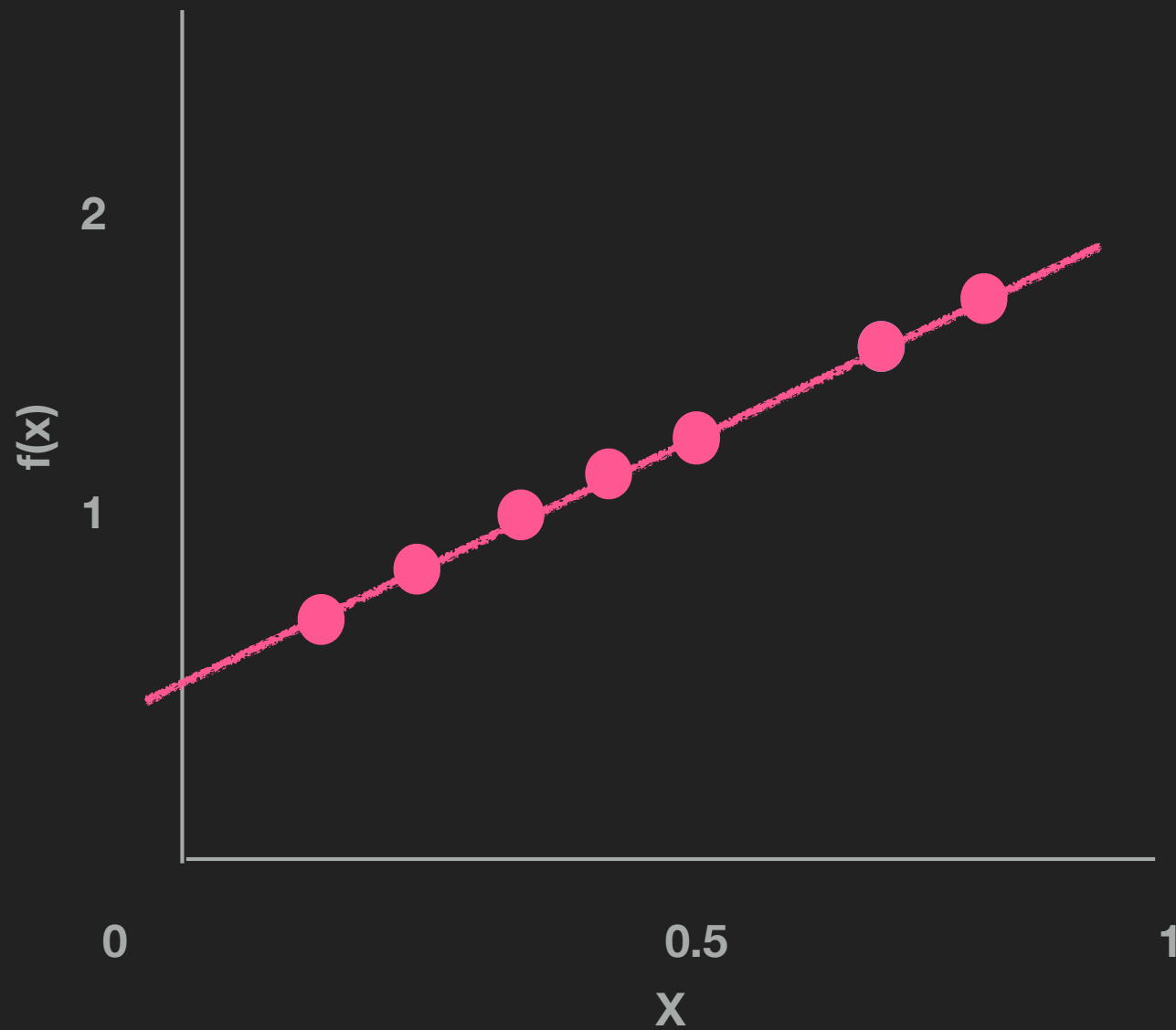
y	x
0.5	0.1
0.9	0.4
1.0	0.2
1.1	0.7
1.1	0.5
1.2	0.3
1.5	0.7
1.6	0.7
1.7	0.5
1.8	0.8

$$f(x) = \beta_1 X + \beta_0 + \varepsilon$$



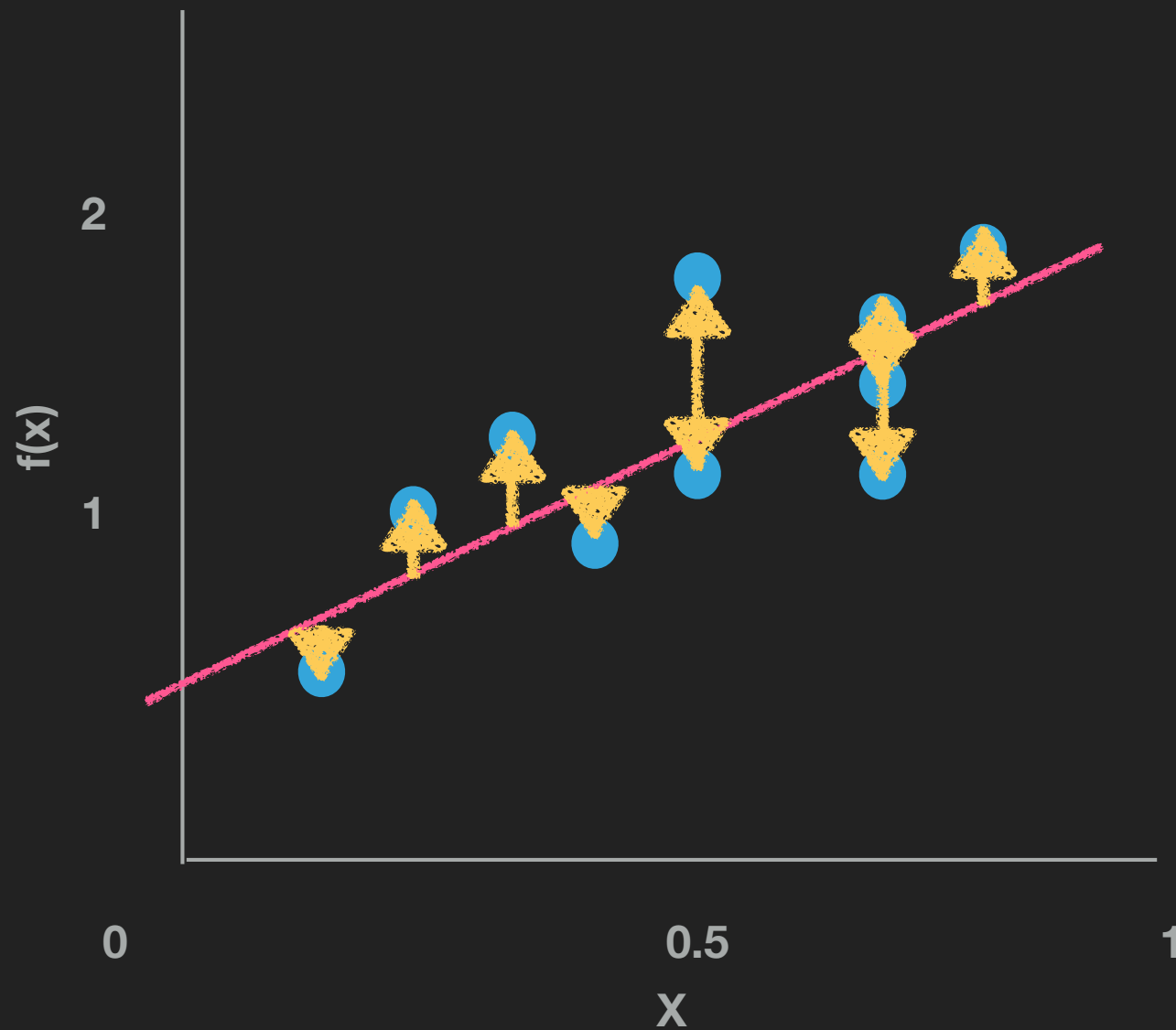
y	x
0.5	0.1
0.9	0.4
1.0	0.2
1.1	0.7
1.1	0.5
1.2	0.3
1.5	0.7
1.6	0.7
1.7	0.5
1.8	0.8

$$f(x) = \beta_1 X + \beta_0 + \varepsilon$$



y	x	f(x)
0.5	0.1	0.73
0.9	0.4	1.1
1.0	0.2	0.86
1.1	0.7	1.50
1.1	0.5	1.24
1.2	0.3	0.99
1.5	0.7	1.50
1.6	0.7	1.50
1.7	0.5	1.24
1.8	0.8	1.62

$$f(x) = \beta_1 X + \beta_0 + \varepsilon$$



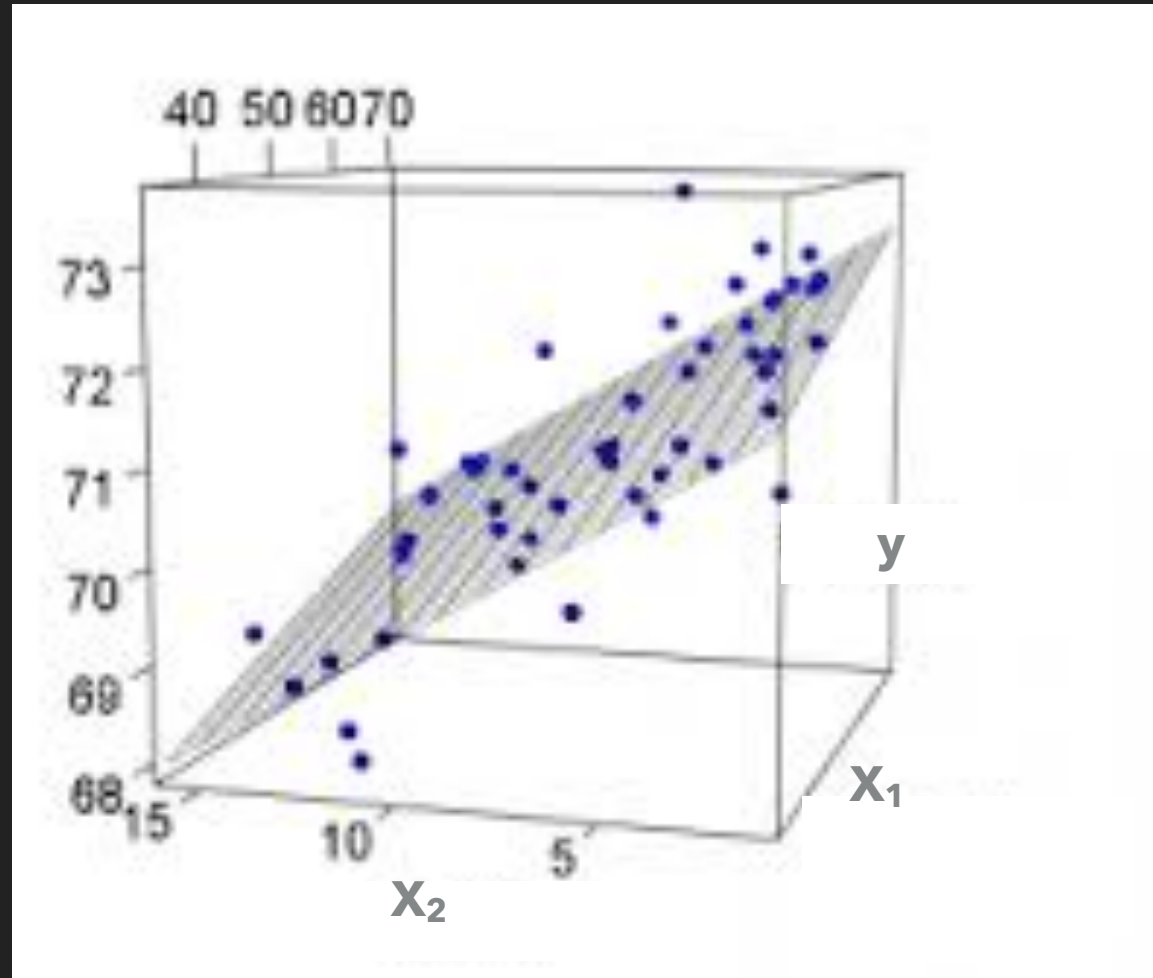
y	x	f(x)	ε
0.5	0.1	0.73	-0.23
0.9	0.4	1.1	-0.2
1.0	0.2	0.86	0.14
1.1	0.7	1.50	-0.4
1.1	0.5	1.24	-0.14
1.2	0.3	0.99	0.21
1.5	0.7	1.50	0
1.6	0.7	1.50	0.1
1.7	0.5	1.24	0.46
1.8	0.8	1.62	0.18

Assumptions of Linear Regression

1. Observations are independent
2. Residuals are normally distributed and centered around zero
 - ▶ Shapiro-Wilk's test
3. Residuals are homoskedastic (no underlying pattern)
 - ▶ - Breusch- Pagan test
4. If using multiple features (multivariate regression), features are not correlated

Multivariate Regression

$$f(\mathbf{x}) = \beta_1 X_1 + \beta_2 X_2 + \beta_0 + \varepsilon$$



ANOVA is a Special Case of Multivariate Regression

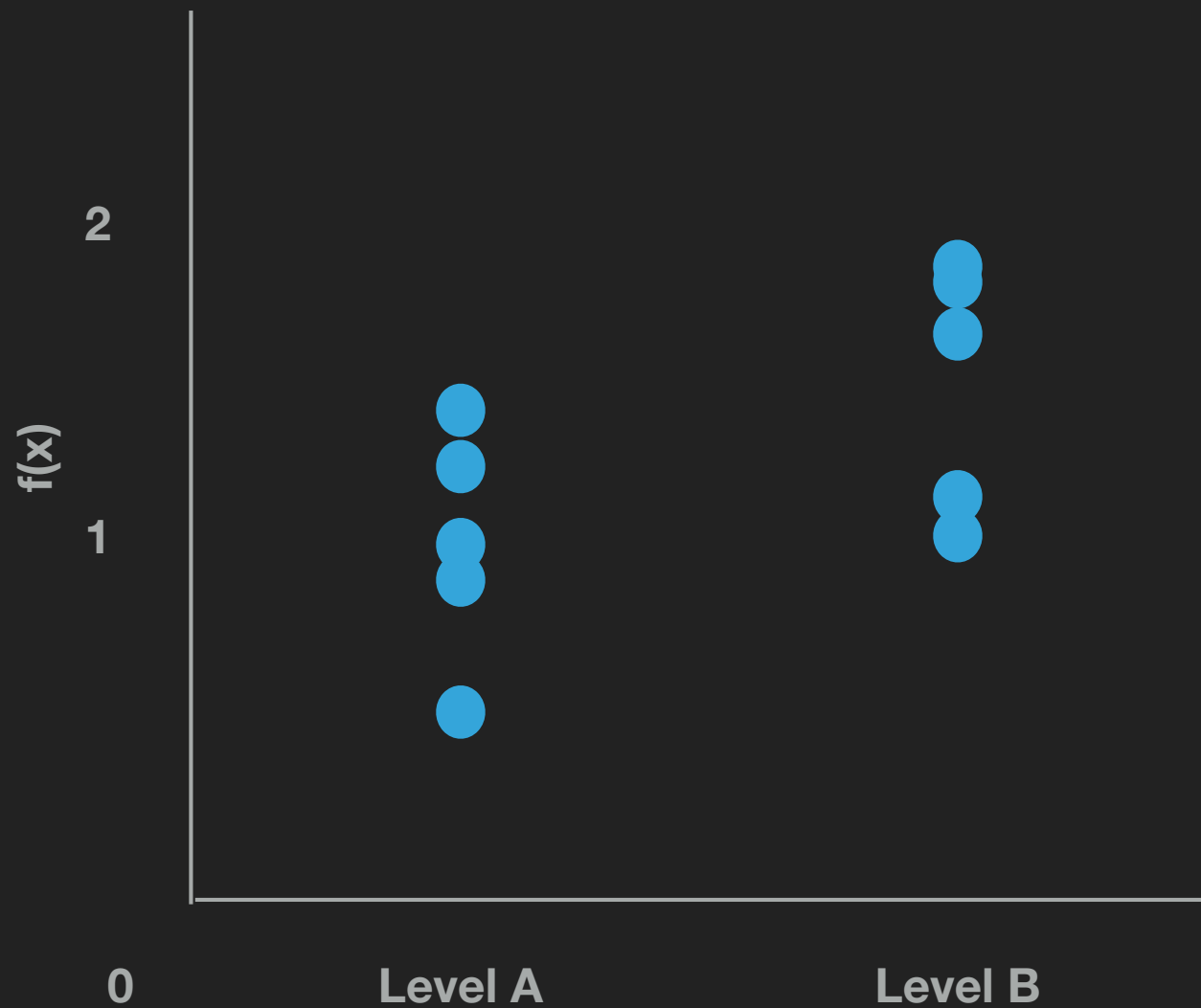
y	x
0.5	A
0.9	A
1.0	A
1.1	B
1.1	A
1.2	B
1.5	A
1.6	B
1.7	B
1.8	B

One Hot Encoded



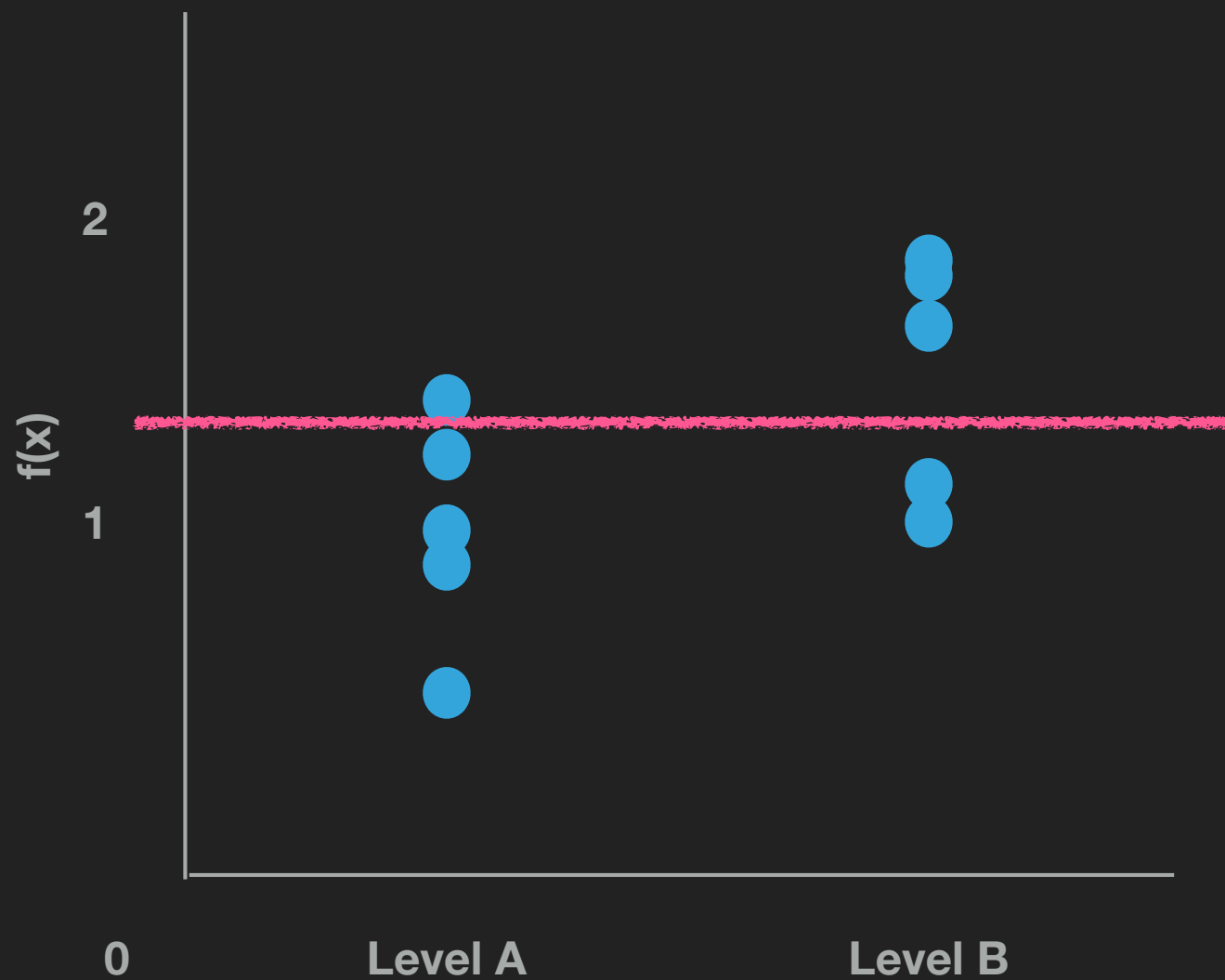
y	X _a	X _b
0.5	1	0
0.9	1	0
1.0	1	0
1.1	0	1
1.1	1	0
1.2	0	1
1.5	1	0
1.6	0	1
1.7	0	1
1.8	0	1

$$Y = \beta_a X_a + \beta_b X_b + \beta_o$$



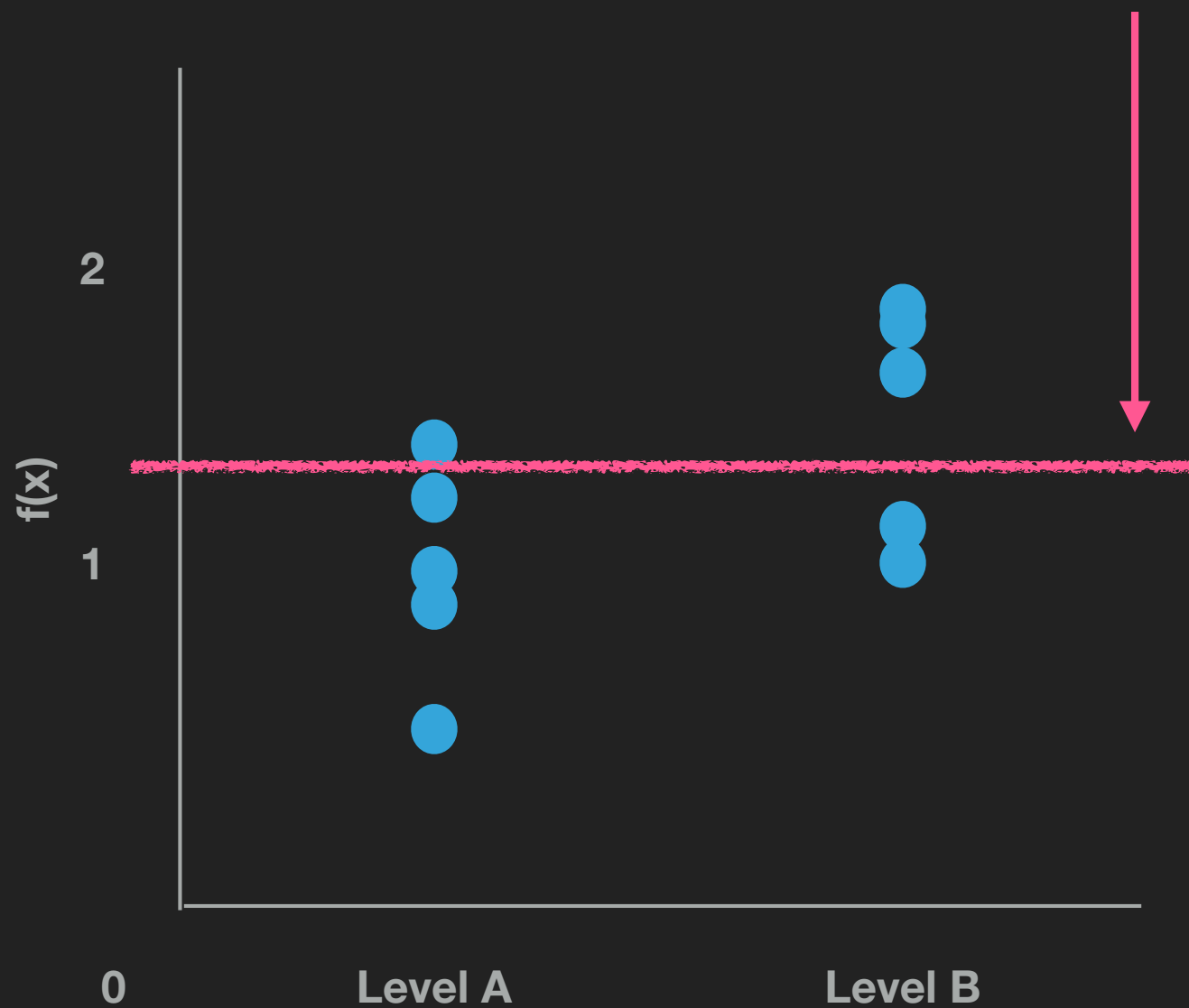
y	X _a	X _b
0.5	1	0
0.9	1	0
1.0	1	0
1.1	0	1
1.1	1	0
1.2	0	1
1.5	1	0
1.6	0	1
1.7	0	1
1.8	0	1

$$Y = \beta_a X_a + \beta_b X_b + \beta_o$$



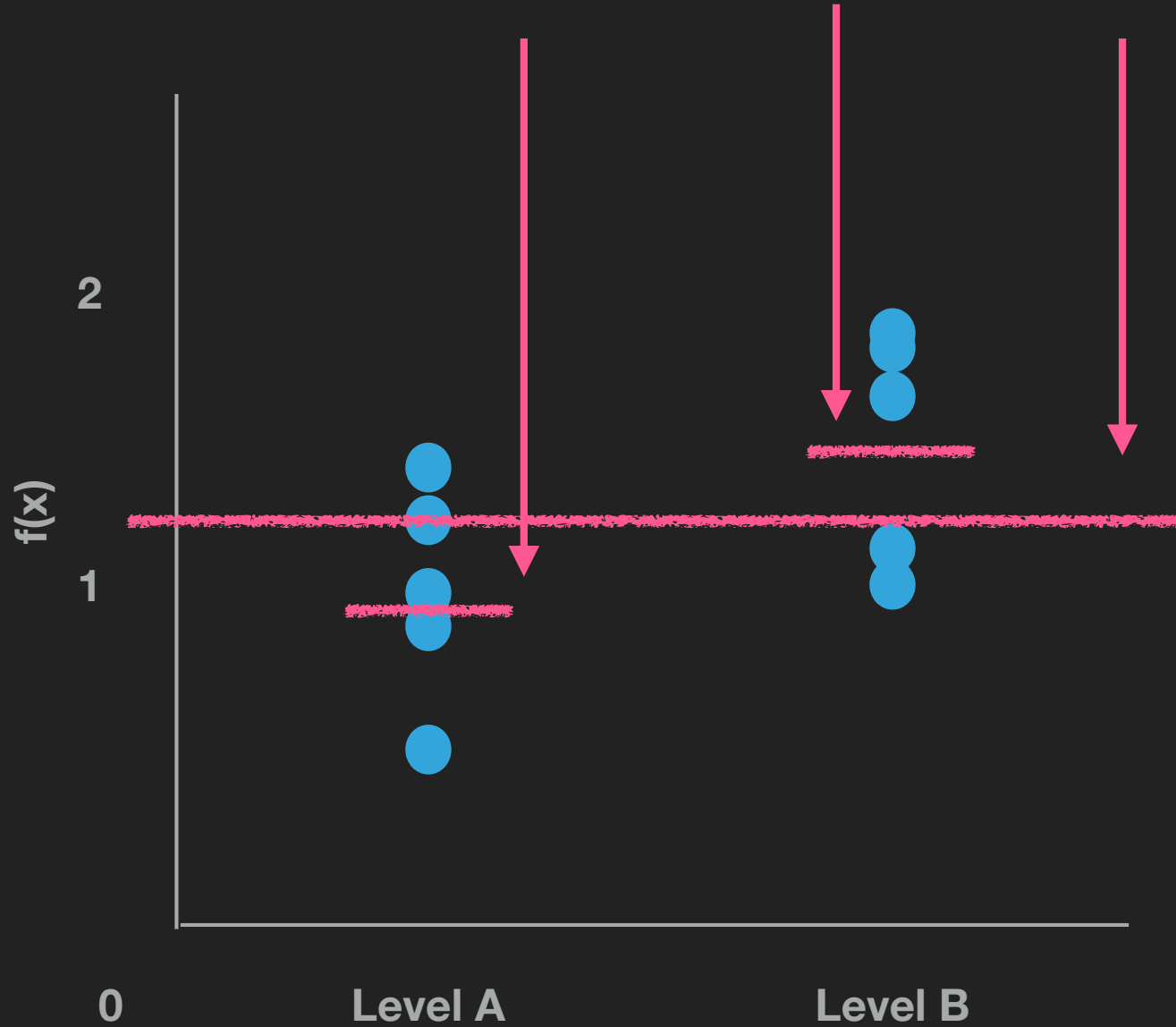
y	X_a	X_b
0.5	1	0
0.9	1	0
1.0	1	0
1.1	0	1
1.1	1	0
1.2	0	1
1.5	1	0
1.6	0	1
1.7	0	1
1.8	0	1

$$Y = \beta_a X_a + \beta_b X_b + \beta_o$$



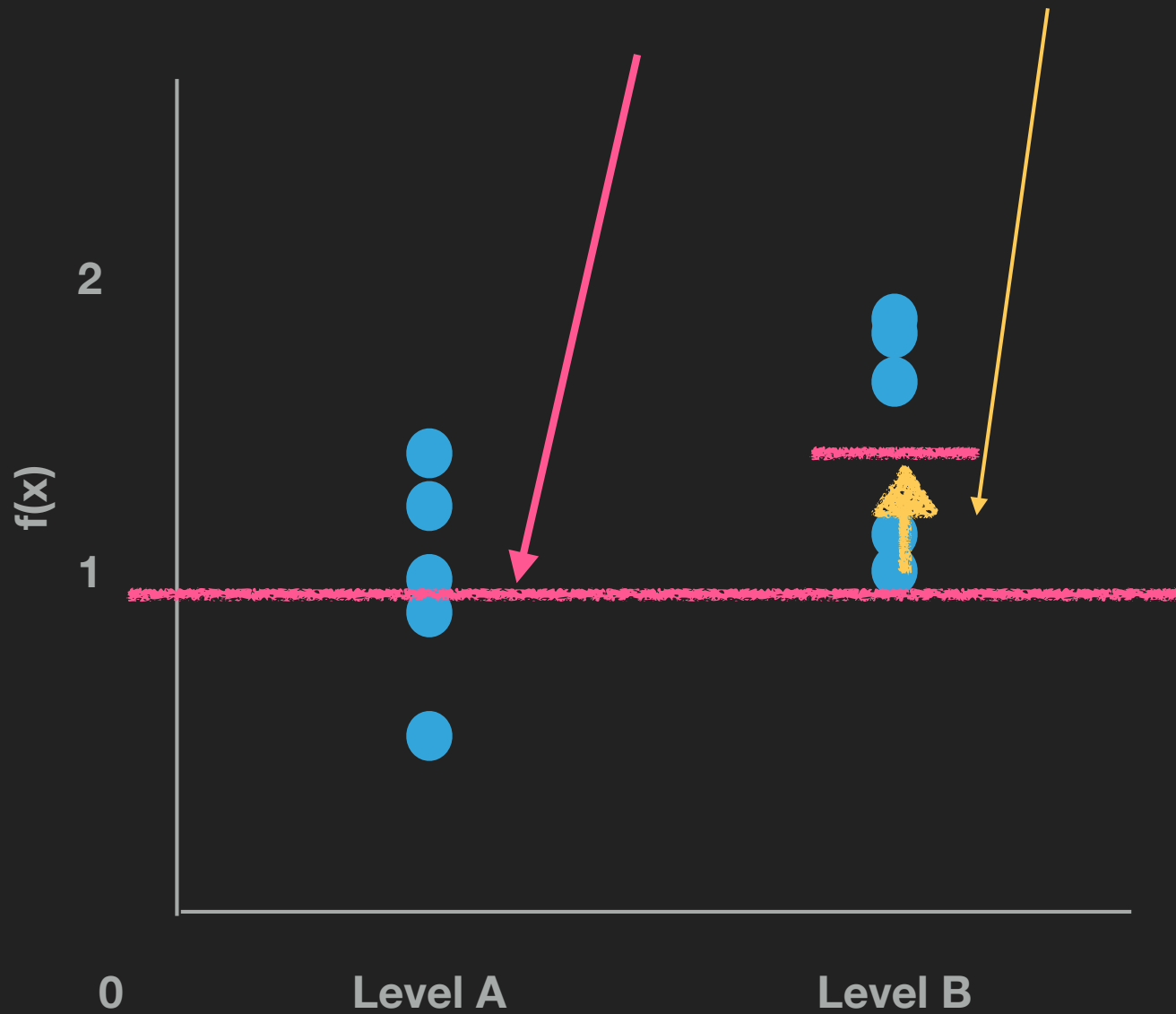
y	X _a	X _b
0.5	1	0
0.9	1	0
1.0	1	0
1.1	0	1
1.1	1	0
1.2	0	1
1.5	1	0
1.6	0	1
1.7	0	1
1.8	0	1

$$Y = -0.24X_a + 0.24X_b + 0.74$$



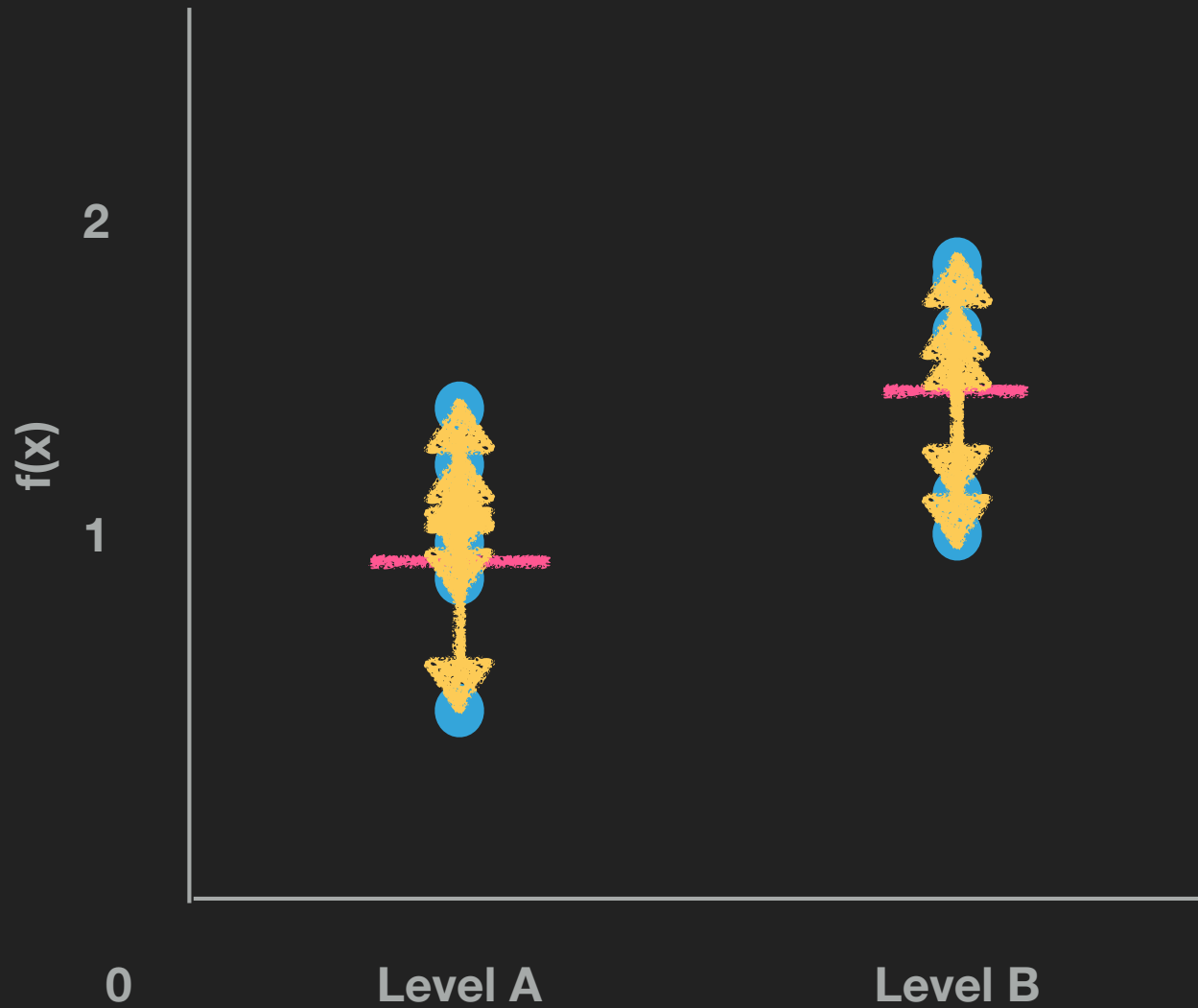
y	X _a	X _b
0.5	1	0
0.9	1	0
1.0	1	0
1.1	0	1
1.1	1	0
1.2	0	1
1.5	1	0
1.6	0	1
1.7	0	1
1.8	0	1

$$Y = 0.98 + 0.5 X_b$$



y	X_a	$f(x)$
0.5	1	0.98
0.9	1	1.1
1.0	1	0.98
1.1	0	1.48
1.1	1	0.98
1.2	0	1.48
1.5	1	0.97
1.6	0	1.48
1.7	0	1.48
1.8	0	1.48

$$Y = 0.98 + 0.5 X_b$$



y	X_a	$f(x)$	ϵ
0.5	1	0.98	-0.48
0.9	1	1.1	-0.2
1.0	1	0.98	0.02
1.1	0	1.48	-0.38
1.1	1	0.98	0.12
1.2	0	1.48	-0.28
1.5	1	0.97	0.53
1.6	0	1.48	0.12
1.7	0	1.48	0.22
1.8	0	1.48	0.32

$$Y = 0.98 + 0.5 X_b$$

OLS Regression Results

Dep. Variable:	y	R-squared:	0.434			
Model:	OLS	Adj. R-squared:	0.363			
Method:	Least Squares	F-statistic:	6.127			
Date:	Wed, 07 Oct 2020	Prob (F-statistic):	0.0384			
Time:	14:04:18	Log-Likelihood:	-1.6598			
No. Observations:	10	AIC:	7.320			
Df Residuals:	8	BIC:	7.925			
Df Model:	1					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]
Intercept	0.9800	0.143	6.861	0.000	0.651	1.309
z[T.b]	0.5000	0.202	2.475	0.038	0.034	0.966
=====						
Omnibus:	0.887	Durbin-Watson:	1.642			
Prob(Omnibus):	0.642	Jarque-Bera (JB):	0.670			
Skew:	-0.287	Prob(JB):	0.715			
Kurtosis:	1.870	Cond. No.	2.62			
=====						

Assumptions of Linear Regression

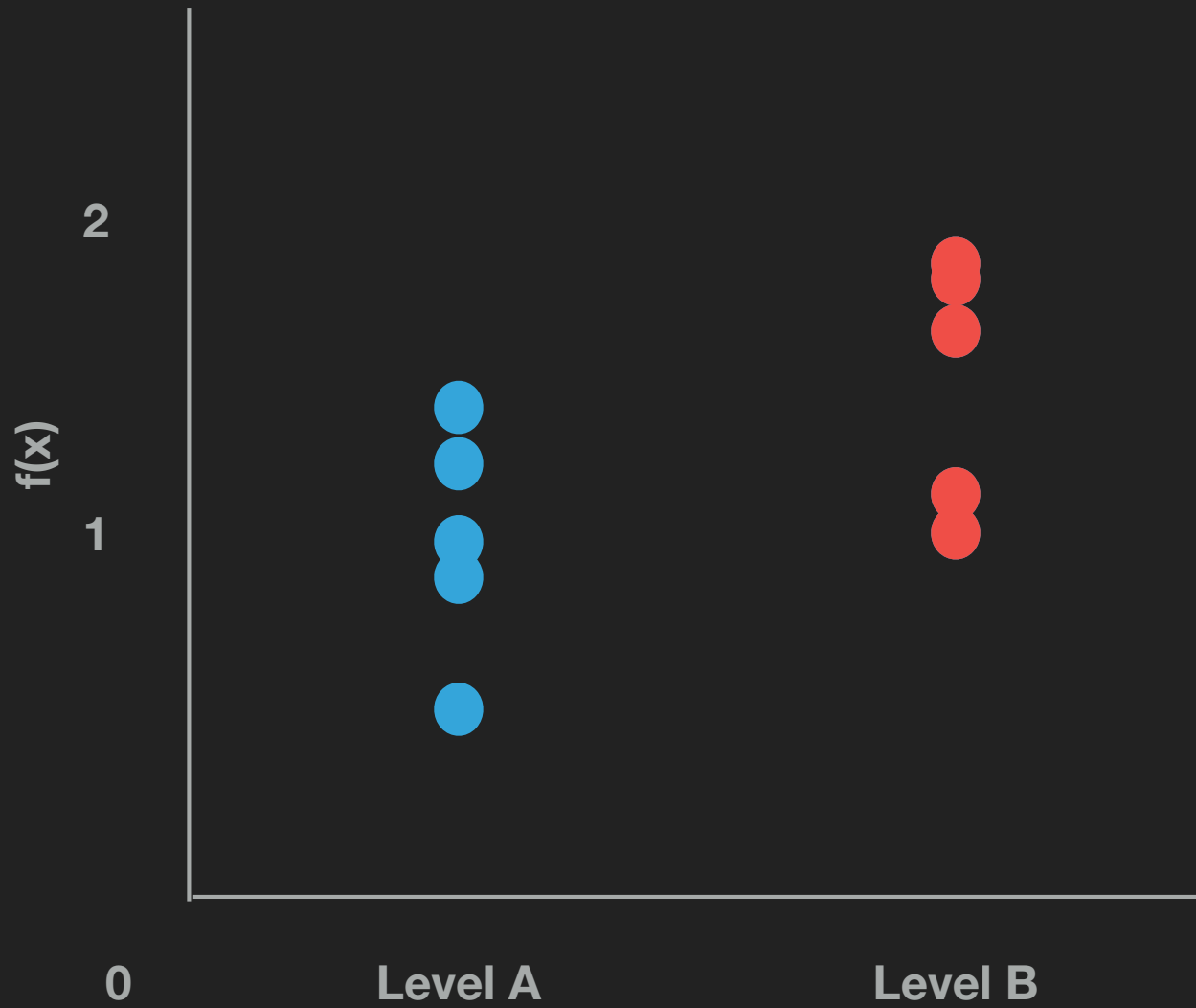
1. Observations are independent and normally distributed
2. Residuals are normally distributed and centered around zero
 - ▶ Shapiro-Wilk's test
3. Residuals are homoskedastic (no underlying pattern)
 - ▶ - Breusch- Pagan test
4. If using multiple features (multivariate regression), features are not correlated

Assumptions of ANOVA

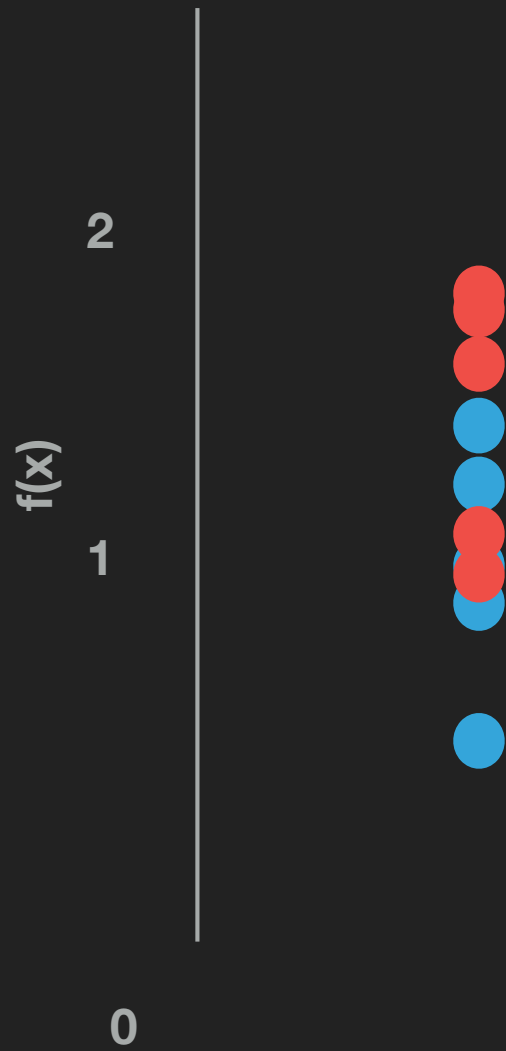
1. Observations are independent and normally distributed within groups
2. Residuals are normally distributed and centered around zero
 - ▶ Shapiro-Wilk's test
3. Variance between groups is similar ($\sim < 2x$)
 - ▶ $|\max([\sigma_a^2 - \sigma_b^2]) / \min([\sigma_a^2 - \sigma_b^2])| < 2$
4. If using multiple features (multivariate regression), features are not correlated

1. Observations are independent and normally distributed within groups

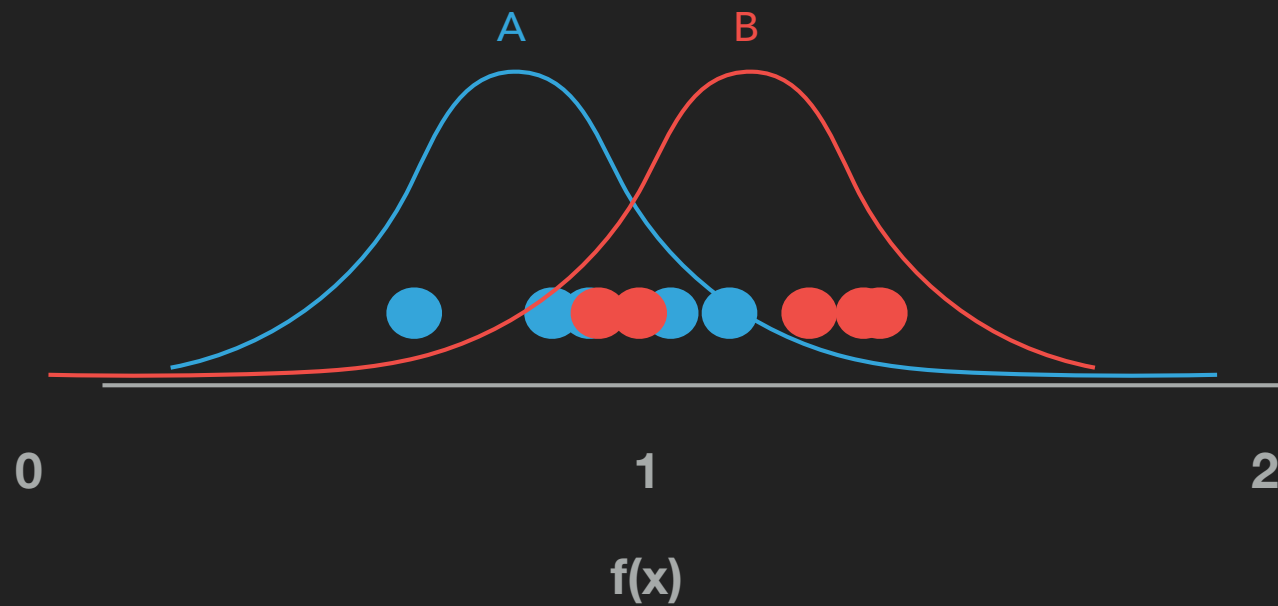
1. Observations are independent and normally distributed within groups



1. Observations are independent and normally distributed within groups



1. Observations are independent and normally distributed within groups



1. Observations are independent and normally distributed within groups



1. Observations are independent and normally distributed within groups

OLS Regression Results

```
=====
Dep. Variable:          y      R-squared:          0.434
Model:                  OLS    Adj. R-squared:       0.363
Method:                 Least Squares    F-statistic:      6.127
Date:                  Sun, 04 Oct 2020    Prob (F-statistic): 0.0384
Time:                  14:37:07    Log-Likelihood:    -1.6598
No. Observations:      10      AIC:              7.320
Df Residuals:          8      BIC:              7.925
Df Model:              1
Covariance Type:       nonrobust
=====
```

	coef	std err	t	P> t	[0.025	0.975]
Intercept	0.9800	0.143	6.861	0.000	0.651	1.309
z[T.b]	0.5000	0.202	2.475	0.038	0.095	0.905

click to scroll output; double click to hide

```
=====
Omnibus:                0.887    Durbin-Watson:      1.642
Prob(Omnibus):          0.642    Jarque-Bera (JB):    0.670
Skew:                   -0.287    Prob(JB):            0.715
Kurtosis:               1.870    Cond. No.            2.62
=====
```

```
1 scipy.stats.ttest_ind(a = aov_data.y.loc[aov_data.z == 'a'], b = aov_data.y.loc[aov_data.z == 'b'])
```

executed in 6ms, finished 13:58:17 2020-10-04

Ttest_indResult(statistic=-2.475368857441685, pvalue=0.03838779259171958)

1. Observations are independent and normally distributed within groups

OLS Regression Results

Dep. Variable:

y

R-squared:

0.434

Model:

OLS

Adj. R-squared:

0.363

Method:

Least Squares

F-statistic:

6.127

Date:

Sun, 04 Oct 2020

Prob (F-statistic):

0.0384

Time:

14:37:07

Log-Likelihood:

-1.6598

No. Observations:

10

AIC:

7.320

Df Residuals:

8

BIC:

7.925

Df Model:

1

Covariance Type:

nonrobust

coef

std err

t

P>|t|

[0.025

0.975]

Intercept

0.9800

0.143

6.861

0.000

0.651

1.309

z[T.b]

0.5000

0.202

2.475

0.038

click to scroll output; double click to hide

Omnibus:

0.887

Durbin-Watson:

1.642

Prob(Omnibus):

0.642

Jarque-Bera (JB):

0.670

Skew:

-0.287

Prob(JB):

0.715

Kurtosis:

1.870

Cond. No.

2.62

For more information, look into contrast coding

Setting your intercept as your "Control" will allow you to see how your different candidates (A, B, ...) compete