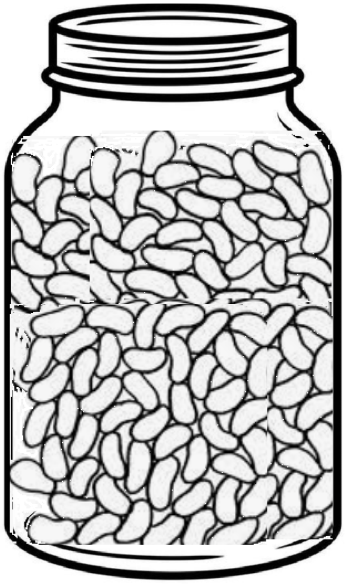
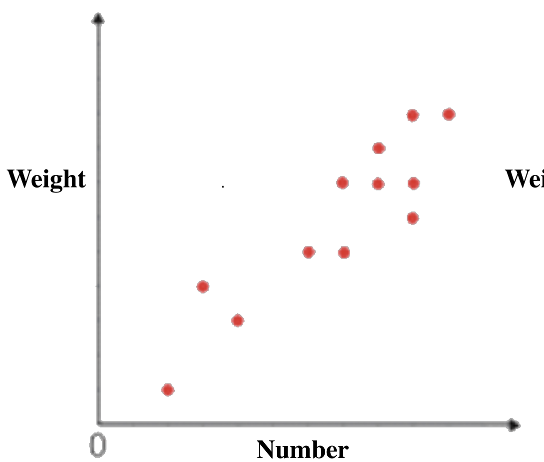
Correlation and Regression

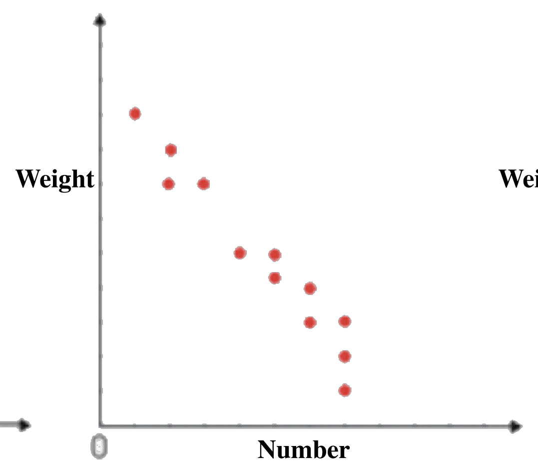


$$\sigma^2_{X\pm Y} = \sigma^2_X \pm \sigma^2_Y \pm \varepsilon\,\sigma_X\sigma_Y$$

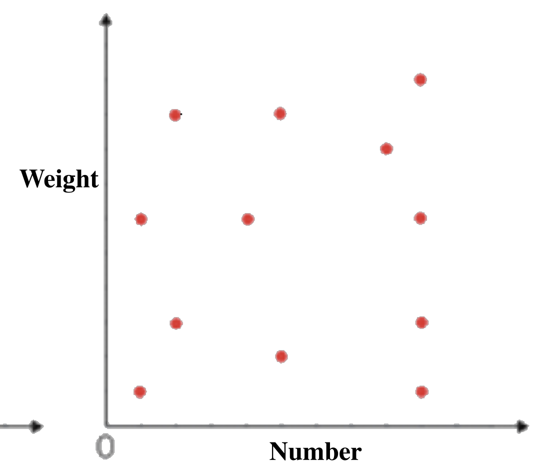
$$\varepsilon = 2\rho$$



Positive Relationship

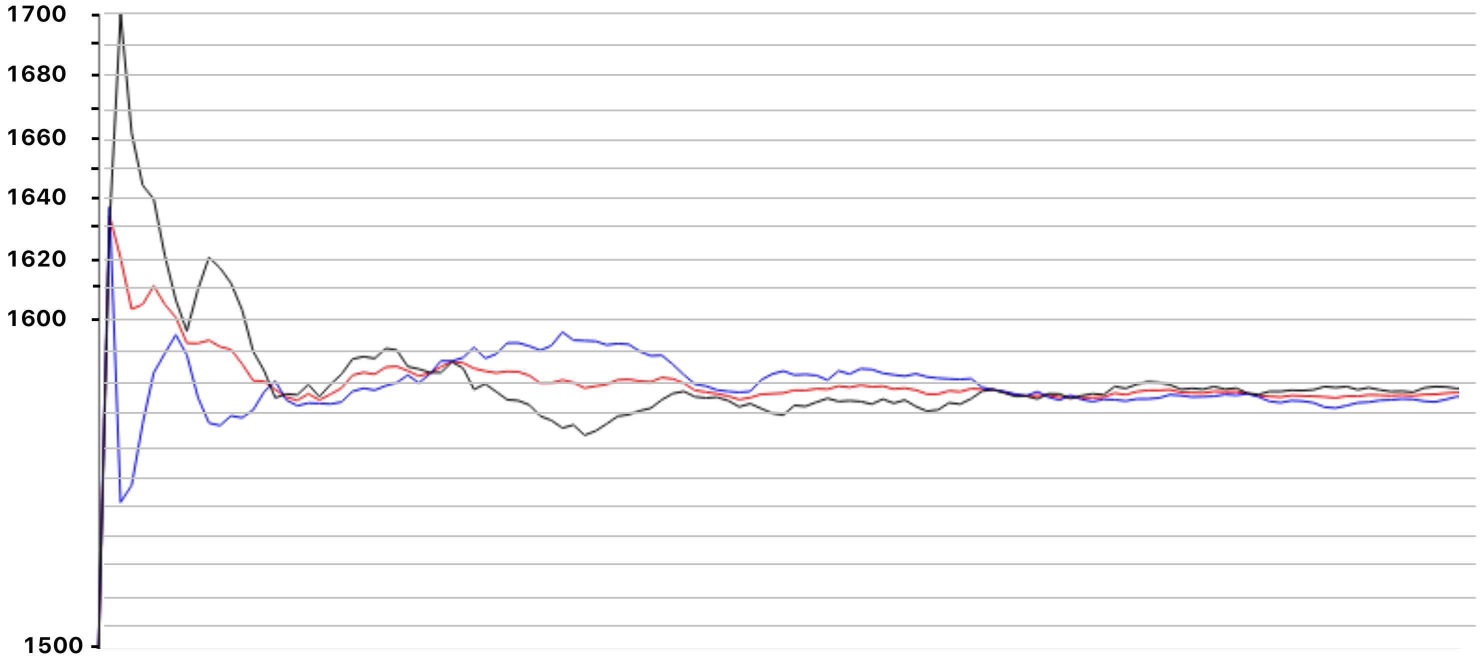


Negative Relationship



No Relationship

VARIANCE SUM LAW I



$$\sigma^2_{X\pm Y} = \sigma^2_X + \sigma^2_Y$$

125000

G O

VAR(X)

832.215397

VAR(Y)

831.661516

VAR(X + Y)

1664.097485

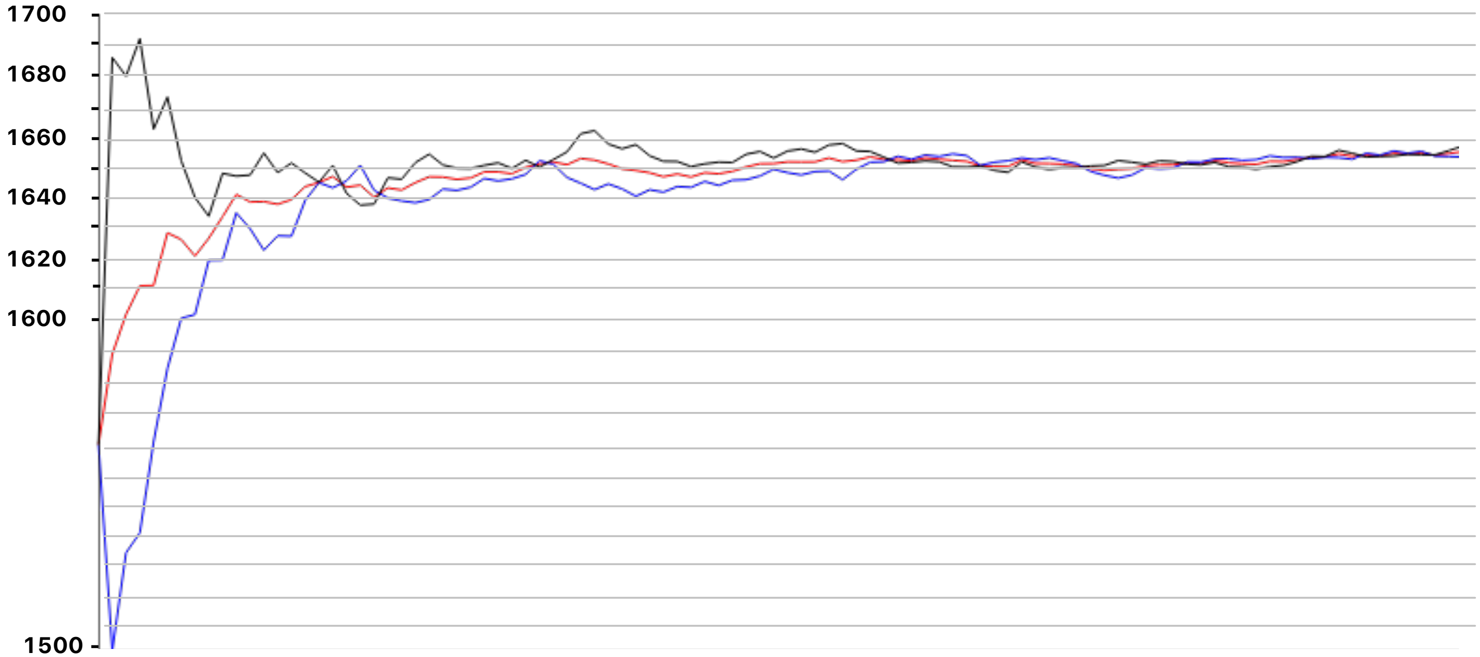
VAR(X - Y)

1663.656341

VAR(X) + VAR(Y)

1663.876913

VARIANCE SUM LAW II



$$\sigma^2_{X\pm Y} = \sigma^2_X + \sigma^2_Y \pm 2\rho\sigma_X\sigma_Y$$

100000

G O

VAR(X)	VAR(Y)	VAR(X + Y)	VAR(X - Y)
835.490535	833.068994	1669.373704	1667.745356

VAR(X) + VAR(Y)

1668.55953

r(X, Y)

-0.000488

$2\rho\sigma_X\sigma_Y$

-0.814174

VARIANCE SUM LAW II

GO

$$\sigma^2_{X\pm Y} = \sigma^2_X + \sigma^2_Y \pm 2\rho\sigma_X\sigma_Y$$

202
121
113
55
142
23
13
170
18
130
134
140
101

100000

GO

VAR(X)	VAR(Y)	VAR(X + Y)	VAR(X - Y)
830.716984	834.060045	1671.22008	1658.333978

VAR(X) + VAR(Y)	$\rho(X,Y)$	$2\rho\sigma_X\sigma_Y$
1664.777029	-0.00387	-6.443051

err added	err subtracted
1658.333978	1671.22008

$$\sigma^2_{X\pm Y} = \sigma^2_X + \sigma^2_Y$$

$$\sigma^2_{X\pm Y} = \sigma^2_X + \sigma^2_Y \pm 2\rho\sigma_X\sigma_Y$$

98	51	149
65	29	94
59	112	171
25	23	48
72	123	195
9	34	43
2	38	40
81	85	166
8	64	72
66	141	207
63	133	196
73	80	153
55	124	179
37	91	128
79	55	
	116	
876.771055	1429.544549	2354.260661

GO

0.021413

2306.315604

Correlation and Regression

Variance Sum Laws

$$\sigma^2_{X\pm Y} = \sigma^2_X + \sigma^2_Y$$

$$\sigma^2_{X\pm Y} = \sigma^2_X + \sigma^2_Y \pm 2\rho\sigma_X\sigma_Y$$

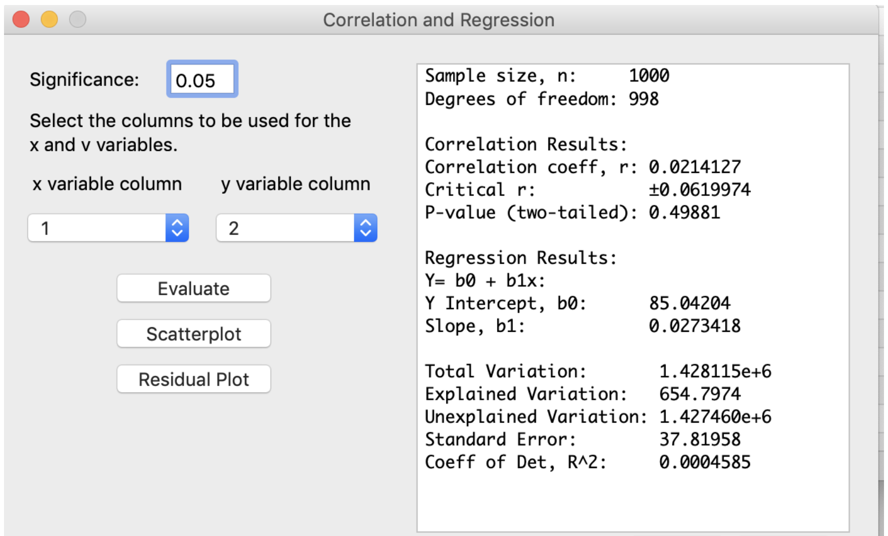
X	Y	X+Y
98	51	149
65	29	94
59	112	171
25	23	48
72	123	195
9	34	43
2	38	40
81	85	166
8	64	72
66	141	207
63	133	196
73	80	153
55	124	179
37	91	128
79	55	
	116	
876.771055	1429.544549	2354.260661

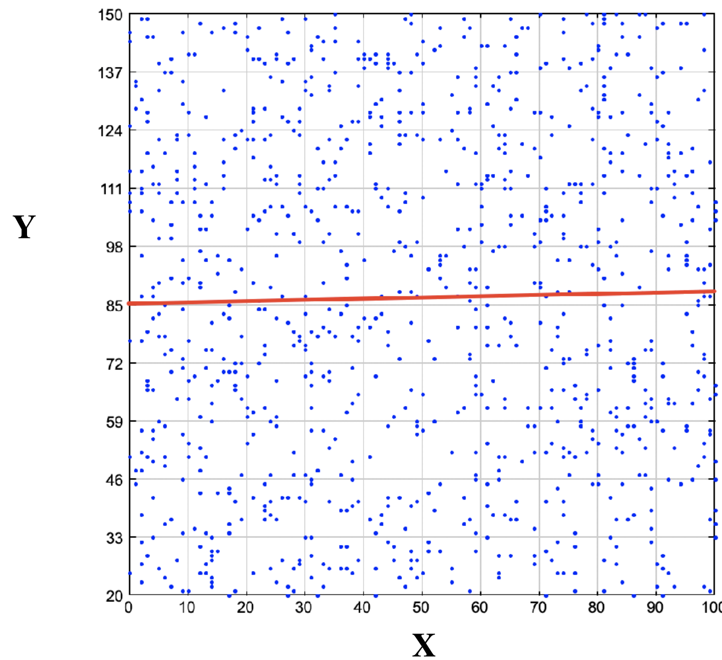
GO

ρ

0.021413

2306.315604





Correlation and Regression

Variance Sum Laws

$$\sigma^2_{X\pm Y} = \sigma^2_X + \sigma^2_Y$$

$$\sigma^2_{X\pm Y} = \sigma^2_X + \sigma^2_Y \pm 2\rho\sigma_X\sigma_Y$$

X	Y	X+Y
98	202	300
65	121	186
59	113	172
25	55	80
72	142	214
9	23	32
2	13	15
81	170	251
8	18	26
66	130	196
63	134	197
73	140	213
55	101	156
37	73	110
79		
876.771055	3556.273513	7945.84462

4433.044568

GO

ρ

0.99467

Correlation and Regression - 1

Significance: 0.05

Select the columns to be used for the x and y variables.

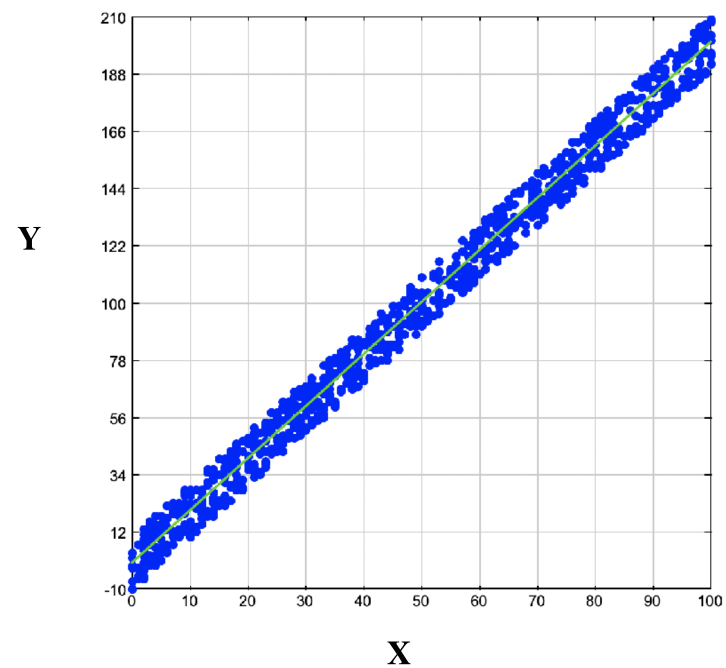
x variable column: 4 y variable column: 6

Evaluate

Scatterplot

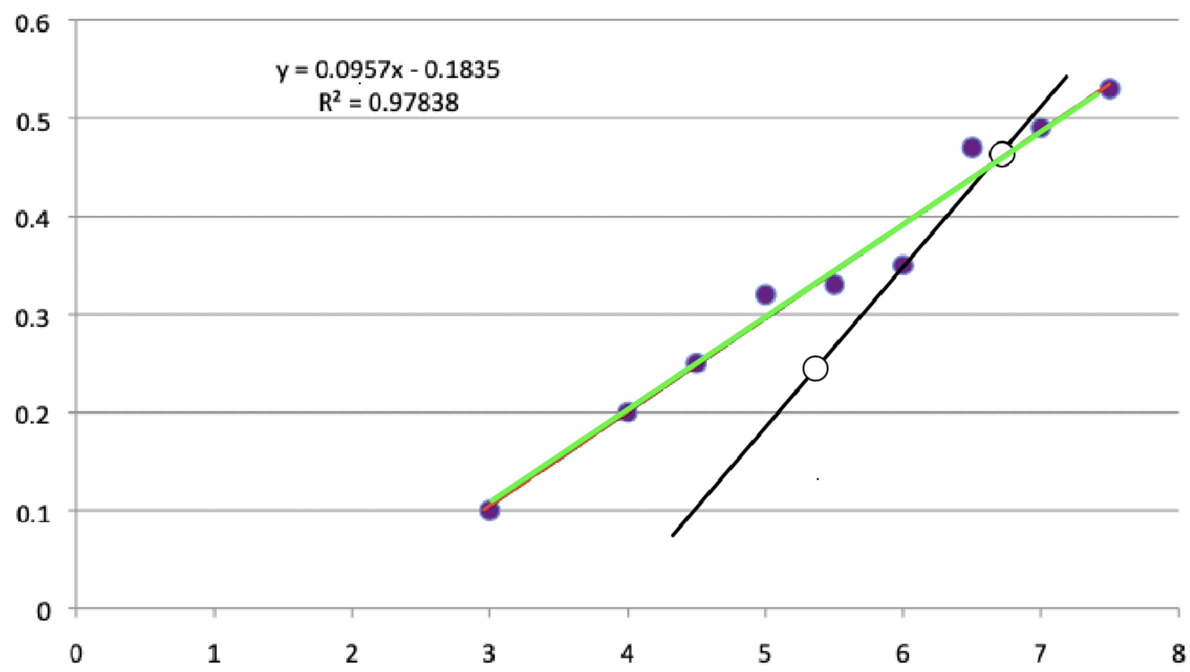
Residual Plot

```
Sample size, n:          1000
Degrees of freedom:      998

Correlation Results:
Correlation coeff, r:    0.9946786
Critical r:              ±0.0619974
P-value (two-tailed):    0.000

Regression Results:
Y= b0 + b1x:
Y Intercept, b0:         -0.037222
Slope, b1:               2.00326

Total Variation:         3.552717e+6
Explained Variation:     3.515007e+6
Unexplained Variation:   37710.56
Standard Error:          6.147043
Coeff of Det, R^2:       0.9893854
```



Correlation and Regression

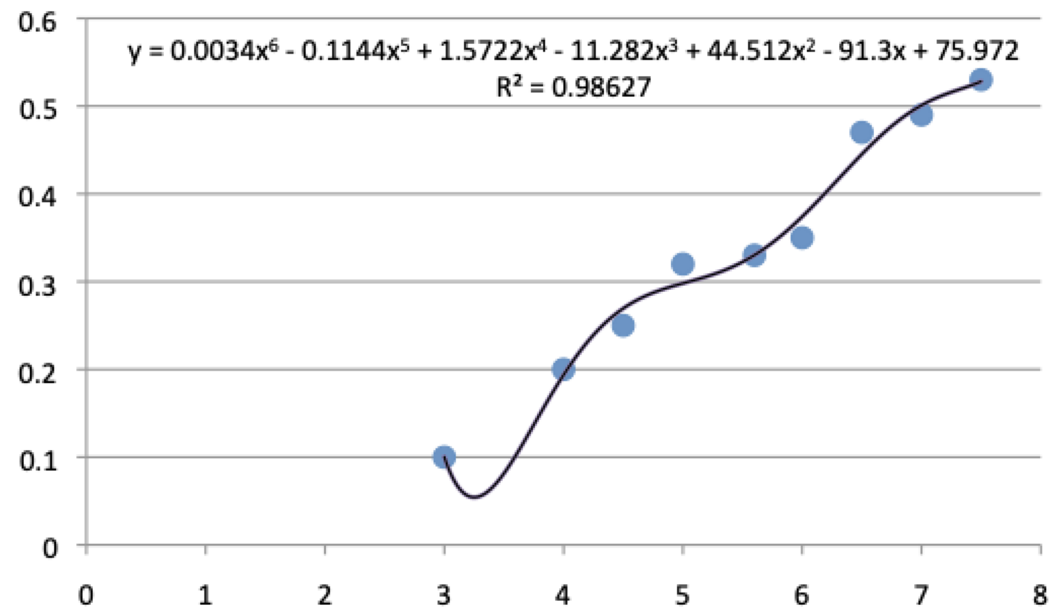
PH	Optical Density
3.0	0.10
4.0	0.20
4.5	0.25
5.0	0.32
5.5	0.33
6.0	0.35
6.5	0.47
7.0	0.49
7.5	0.53




```
Sample size, n:     9
Degrees of freedom: 7

Correlation Results:
Correlation coeff, r: 0.9891333
Critical r:           ±0.666384
P-value (two-tailed): 0.000

Regression Results:
Y= b0 + b1x:
Y Intercept, b0:      -0.1834839
Slope, b1:            0.0957419

Total Variation:        0.1613556
Explained Variation:    0.1578678
Unexplained Variation:  0.0034877
Standard Error:         0.0223215
Coeff of Det, R^2:      0.9783847
```

Correlation and Regression

PH	Optical Density
3.0	0.10
4.0	0.20
4.5	0.25
5.0	0.32
5.5	0.33
6.0	0.35
6.5	0.47
7.0	0.49
7.5	0.53



Correlation and Regression

Assume $f(x) = mx + b$ exists. Now calculate *residuals*

$$\begin{cases} r_1 = y_1 - f(x_1) \\ r_2 = y_2 - f(x_2) \\ . \quad . \quad . \quad . \quad . \\ r_n = y_n - f(x_n) \end{cases} \Rightarrow \begin{cases} r_1 = y_1 - (mx_1 + b) \\ r_2 = y_2 - (mx_2 + b) \\ . \quad . \quad . \quad . \quad . \\ r_n = y_n - (mx_n + b) \end{cases} \Rightarrow \begin{cases} r_1^2 = m^2 x_1^2 + 2bmx_1 + b^2 - 2mx_1y_1 - 2by_1 + y_1^2 \\ r_2^2 = m^2 x_2^2 + 2bmx_2 + b^2 - 2mx_2y_2 - 2by_2 + y_2^2 \\ . \quad . \quad . \quad . \quad . \\ r_n^2 = m^2 x_n^2 + 2bmx_n + b^2 - 2mx_ny_n - 2by_n + y_n^2 \end{cases}$$

Adding we get $\sum r^2 = m^2 \sum x^2 + 2bm \sum x + b^2 - 2m \sum xy - 2b \sum y + \sum y^2$

taking derivatives we have

$$\begin{aligned} m\sum x^2 + b\sum x &= \sum xy \\ m\sum x + bn &= \sum y \end{aligned}$$

Correlation and Regression

Correlation Coefficient: a measure of the **strength** of the (linear) relationship between x and y.

$$r = \frac{1}{n-1}\sum_k \left(\frac{x_k - \bar{x}}{s_x}\right)\left(\frac{y_k - \bar{y}}{s_y}\right) \quad \Rightarrow \quad r = \frac{\dfrac{\sum_k (x_k - \bar{x})(y_k - \bar{y})}{n-1}}{\sqrt{\dfrac{\sum (x_k - \bar{x})^2}{n-1}}\sqrt{\dfrac{\sum (y_k - \bar{y})^2}{n-1}}}$$

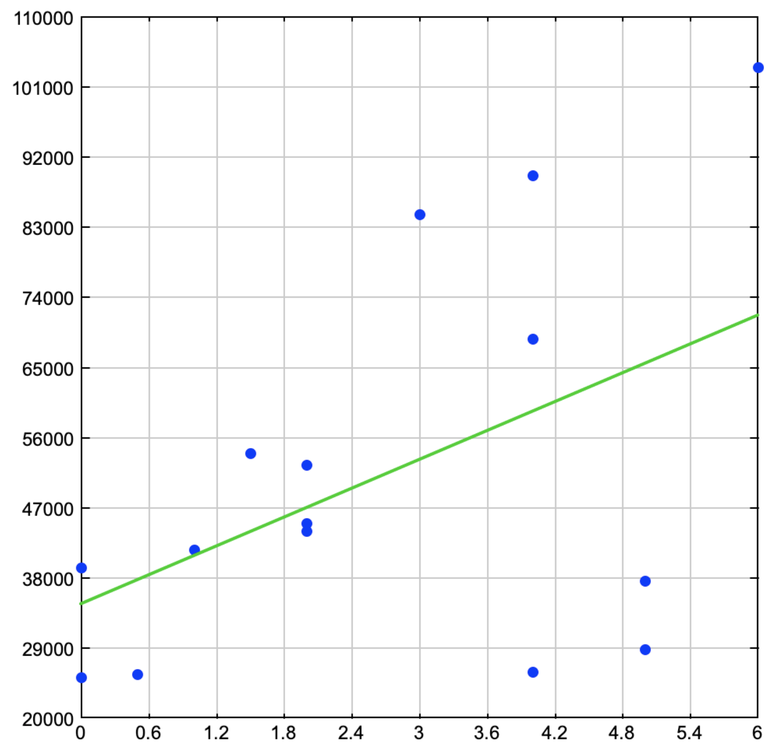
$$\Rightarrow \quad r = \frac{\dfrac{\sum xy}{n-1}}{\sqrt{\dfrac{\sum x^2}{n-1}}\sqrt{\dfrac{\sum y^2}{n-1}}} = \hat{m}\left(\frac{s_x}{s_y}\right)$$

where $\hat{m}\left(\frac{s_x}{s_y}\right)$ is the slope of the sample multiplied by the ratio of the standard deviations of the x and y values.

Coefficient of Determination: r^2

Correlation and Regression

School (Yrs)	Income
0.0	39,300
0.0	25,200
0.5	25,550
1.0	41,600
1.5	54,000
2.0	52,500
2.0	45,000
2.0	44,000
3.0	84,600
4.0	68,600
4.0	25,900
4.0	89,600
5.0	37,600
5.0	28,800
6.0	103,500



Correlation and Regressi

Significance: 0.05

```
Sample size, n:      15
Degrees of freedom: 13

Correlation Results:
Correlation coeff, r: 0.478036
Critical r:           ±0.5139779
P-value (two-tailed): 0.07149

Regression Results:
Y= b0 + b1x:
Y Intercept, b0:      34608.84
Slope, b1:            6165.434

Total Variation:        8.622145e+9
Explained Variation:    1.970319e+9
Unexplained Variation: 6.651826e+9
Standard Error:         22620.32
Coeff of Det, R^2:      0.2285184
```

DO NOT Reje

$$TS = \frac{|r|}{\sqrt{\frac{1-r^2}{v}}} = 1.96212696$$

$$CV = \frac{t_{\alpha,v}}{\sqrt{\left(t_{\alpha,v}\right)^2 + v}} = 0.513977484$$

```
t Value:   1.962000
Prob Dens: 0.0636899

Cumulative Probs
Left:     0.964233
Right:    0.035767
2 Tailed: 0.071533
Central:  0.928467

13 Degrees of Freedom
```

Ho : There is no correlatio

Other Correlation Measures

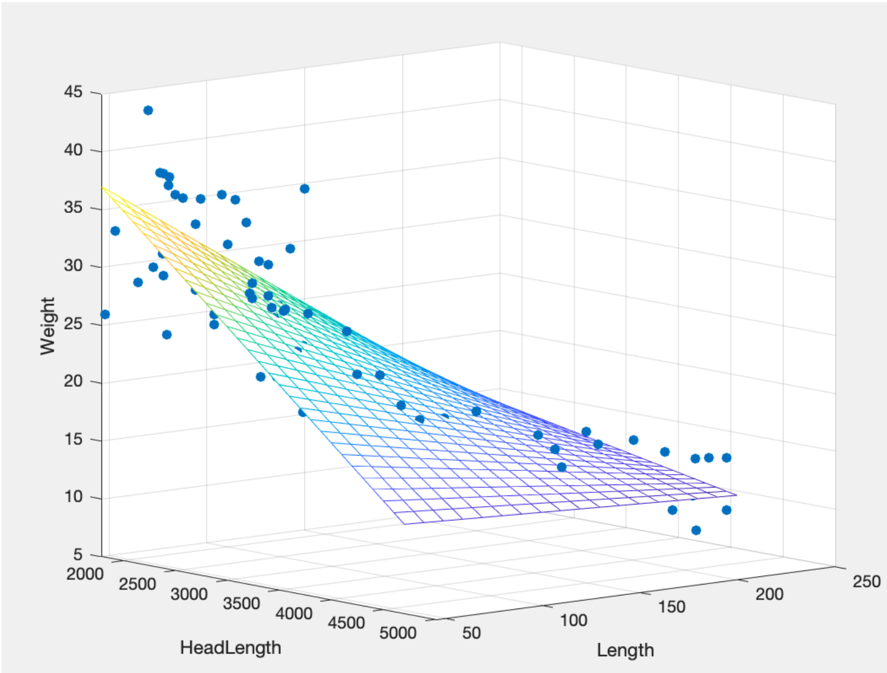
Situation / Data Type	Recommended Coefficient	Why
Both variables are **continuous** and **approximately normally distributed**, relationship looks **linear**	**Pearson r**	Most powerful for linear relationships with normal data
At least one variable is **ordinal** (ranks, Likert scales), or the relationship is **monotonic** but not necessarily linear, or data has outliers	**Spearman ρ (rho)**	Non-parametric, uses ranks, robust to outliers and non-normality

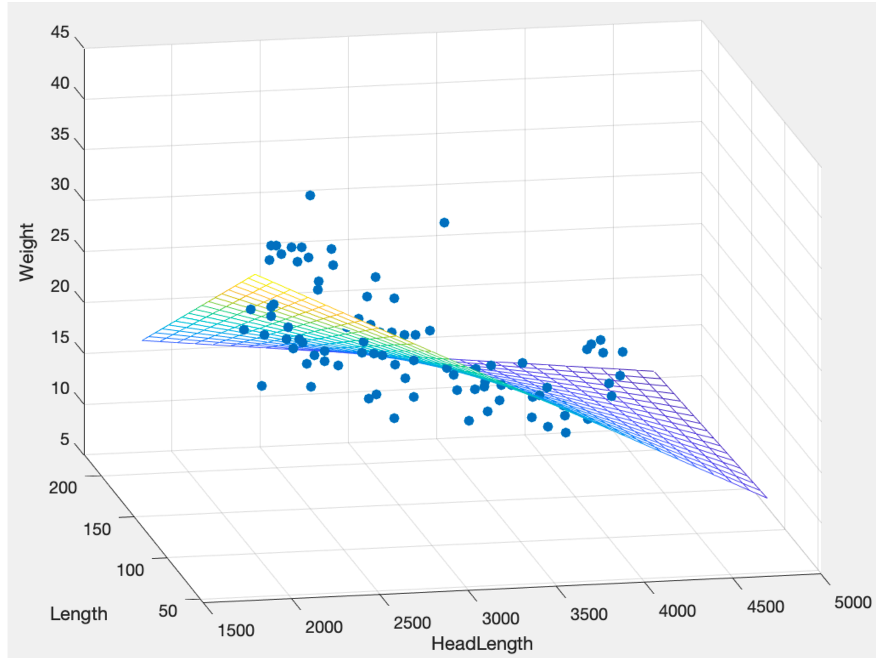
Most Common Real World Choices

- Psychology / social sciences with Likert scales → **Spearman** or **Kendall**
- Medical/biological data with continuous outcomes → **Pearson if** normality holds, otherwise **Spearman**
- When you have a true dichotomy (sex, treatment vs control) and a continuous outcome → **Point-biserial** (or just run a t-test; they're mathematically equivalent)

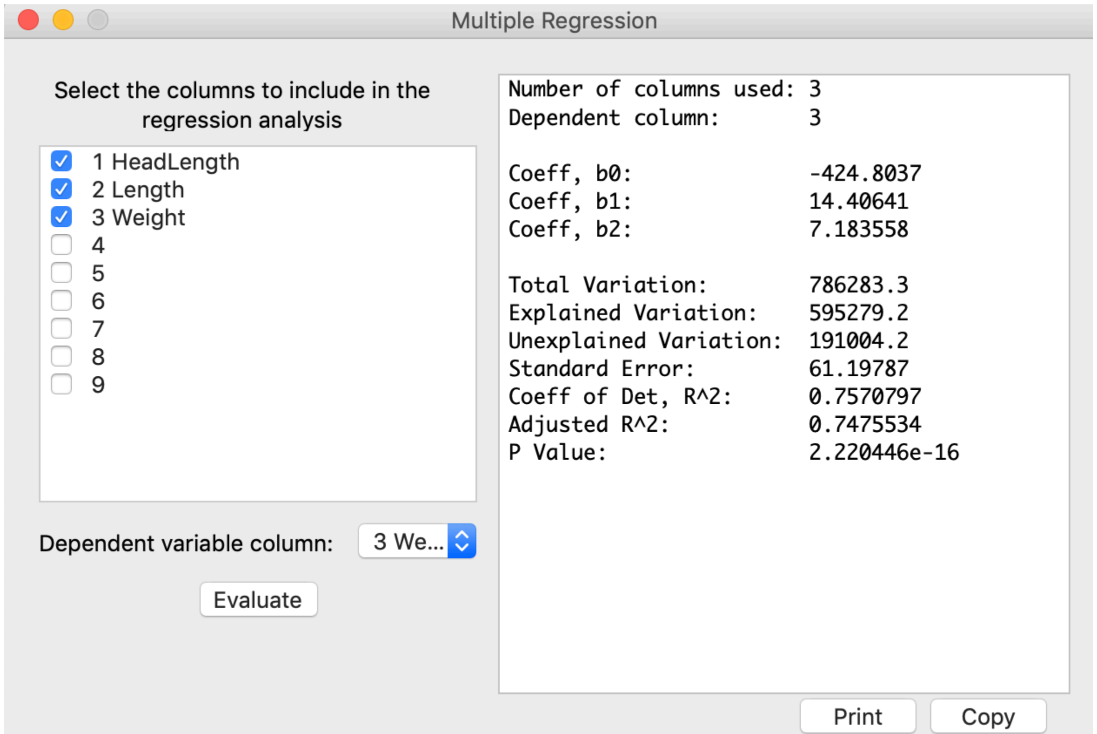
Situation / Data Type	Recommended Coefficient	Why
You have **ordinal data** and want a more conservative estimate, or sample size is small ($n < 30$), or many ties in ranks	**Kendall τ (tau)**	Also rank-based, less sensitive to ties, more conservative than Spearman
One variable is **truly dichotomous** (only 2 real categories, e.g., yes/no, dead/alive) and the other is continuous and roughly normal	**Point-biserial r**	Special case of Pearson; treats the dichotomy as 0/1
One variable is **artificially dichotomized** continuous variable (e.g., pass/fail based on a test score cut-off) and the other is continuous	**Biserial r**	Estimates what Pearson would be if the variable hadn't been dichotomized (rarely used today)

Multiple Correlation and Regression





Multiple Correlation and Regression



y = b0 + b1 * x1 + b2* x2