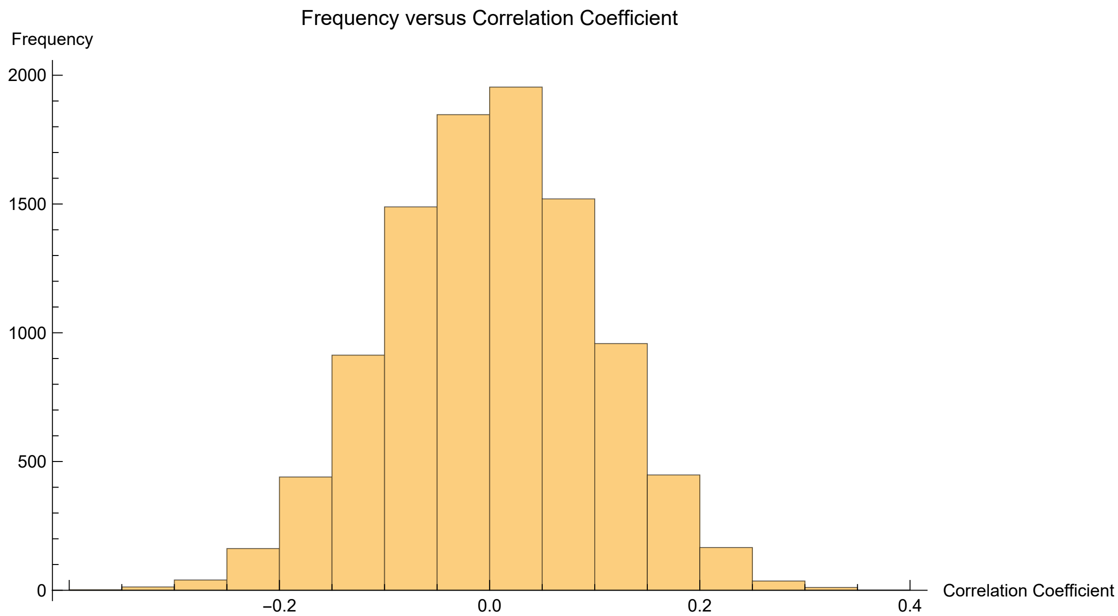
Fooled By Correlation

Correlation is a stochastic variable. The mean may be zero. Very often correlation is not even correlation.

[Image]

```
Histogram[Table[X = RandomVariate[NormalDistribution[], 100];
   Y = RandomVariate[NormalDistribution[], 100];
   Correlation[X, Y], {10^4}], AxesLabel → {"Correlation Coefficient", "Frequency"},
 ImageSize → Large, PlotLabel → "Frequency versus Correlation Coefficient"]
```

Out[•]=

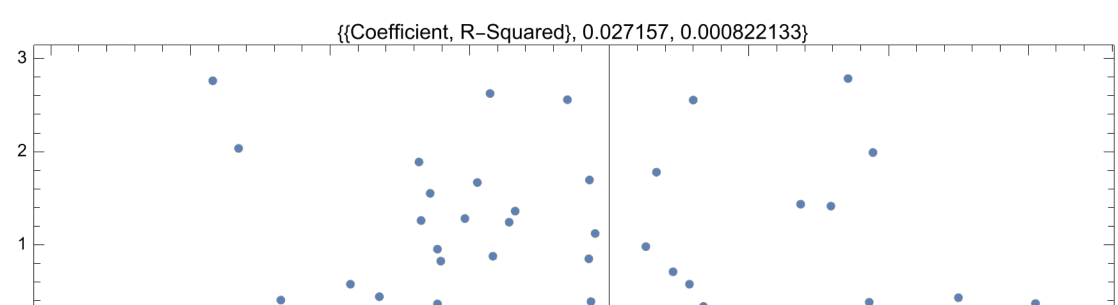


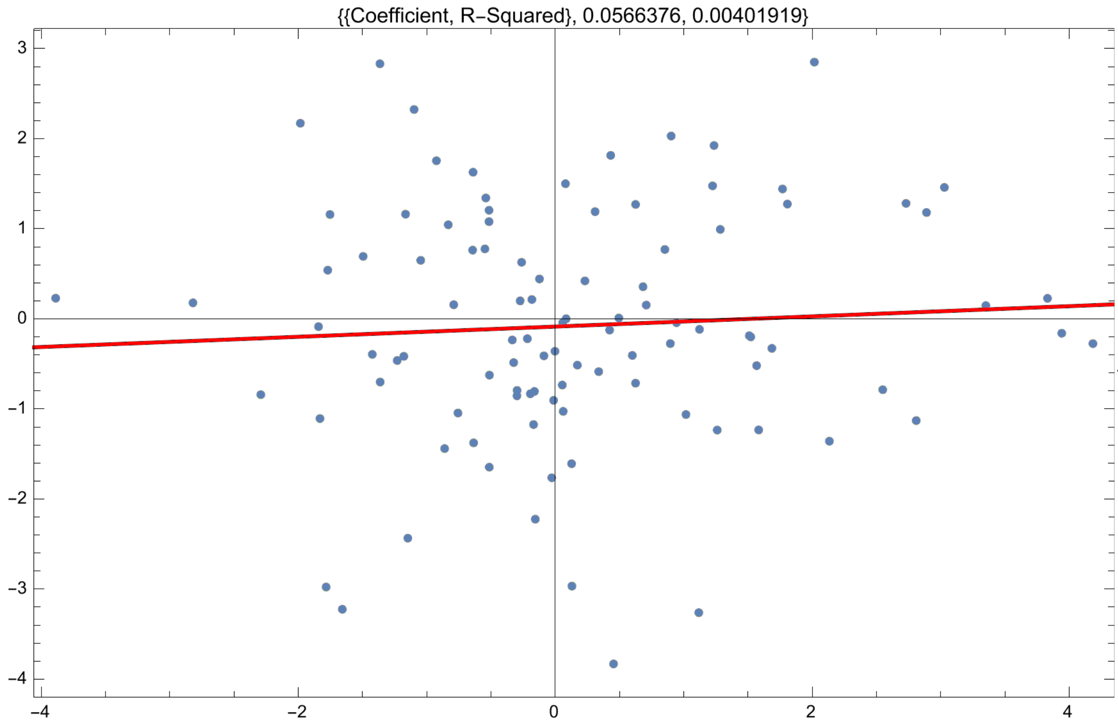
Comments

Correlations of from 0 - 0.3 is closer to zero. Entropy methods are superior.

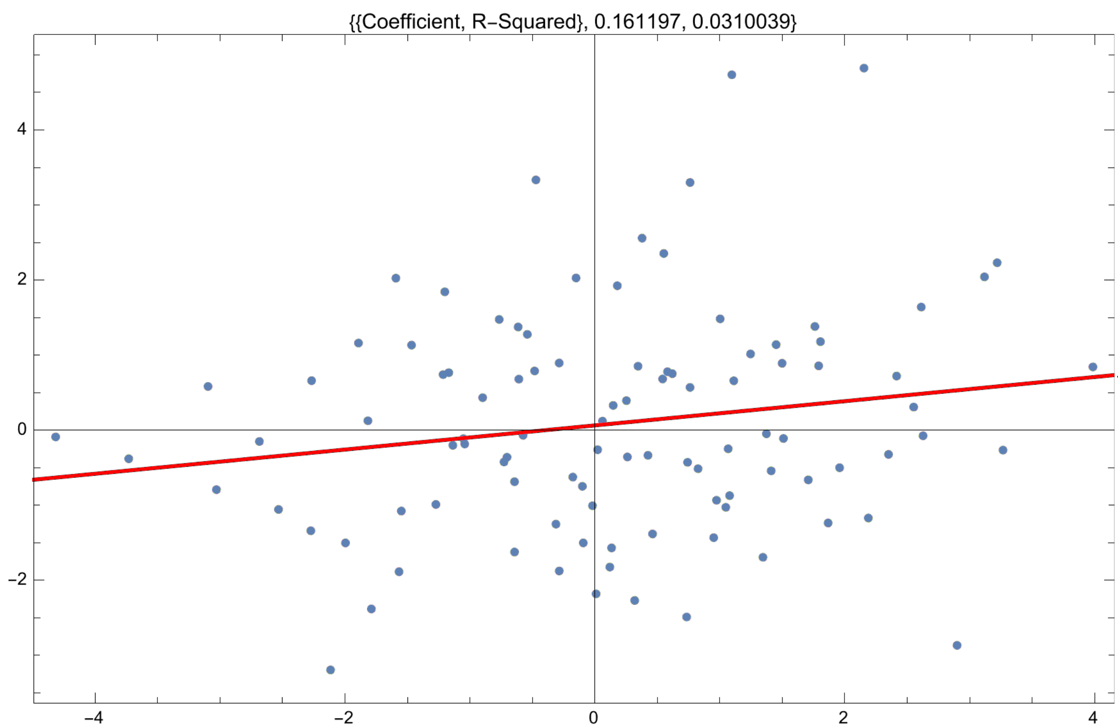
```
Table[
 data = Table[{RandomVariate[NormalDistribution[RandomVariate[NormalDistribution[]], 1]],
     RandomVariate[NormalDistribution[RandomVariate[NormalDistribution[]], 1]]}, {100}];
 lm = LinearModelFit[data, x, x];
 Show[ListPlot[data, PlotStyle → PointSize[Medium], ColorFunction → Blue,
    ImageSize → Large], Plot[lm[x], {x, -5, 5}, PlotStyle → Red], Frame → True,
  PlotLabel → {"{Coefficient, R-Squared}", (lm // Normal)〚2〛〚1〛, lm["RSquared"]}], {12}]
```

Out[•]=



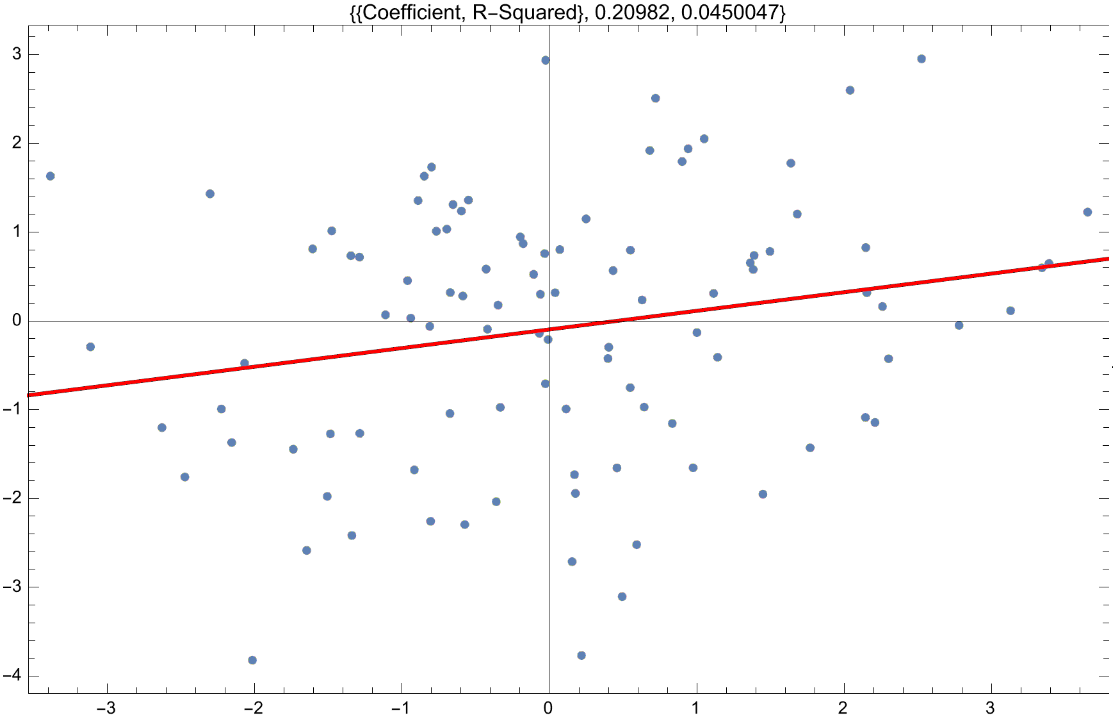

{{Coefficient, R-Squared}, 0.0566376, 0.00401919}
3
2
1
0
−1
−2
−3
−4
−4
−2
0
2
4

,


{{Coefficient, R-Squared}, 0.161197, 0.0310039}
4
2
0
−2
−4
−2
0
2
4

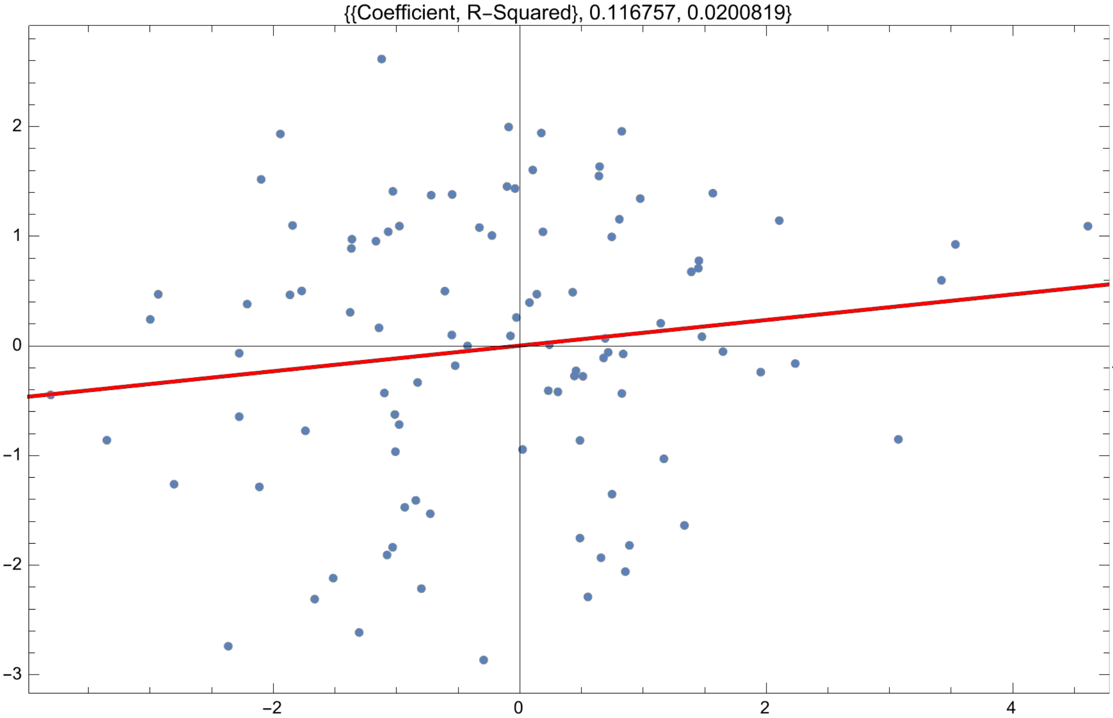
,

{{Coefficient, R-Squared}, 0.20982, 0.0450047}

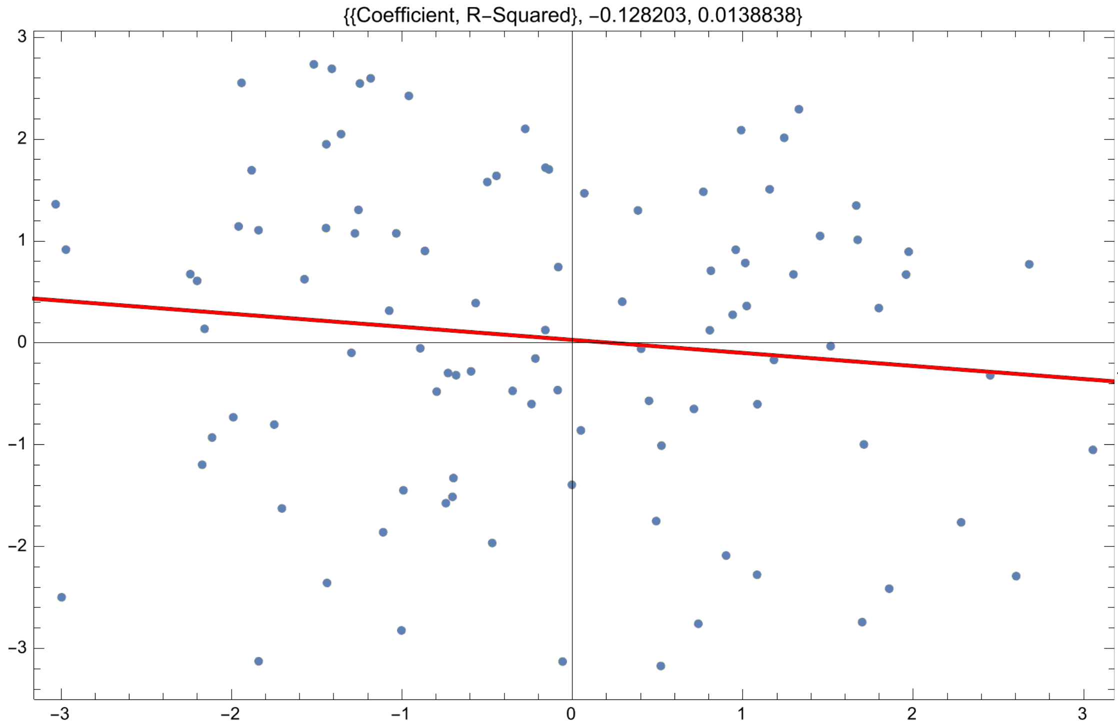


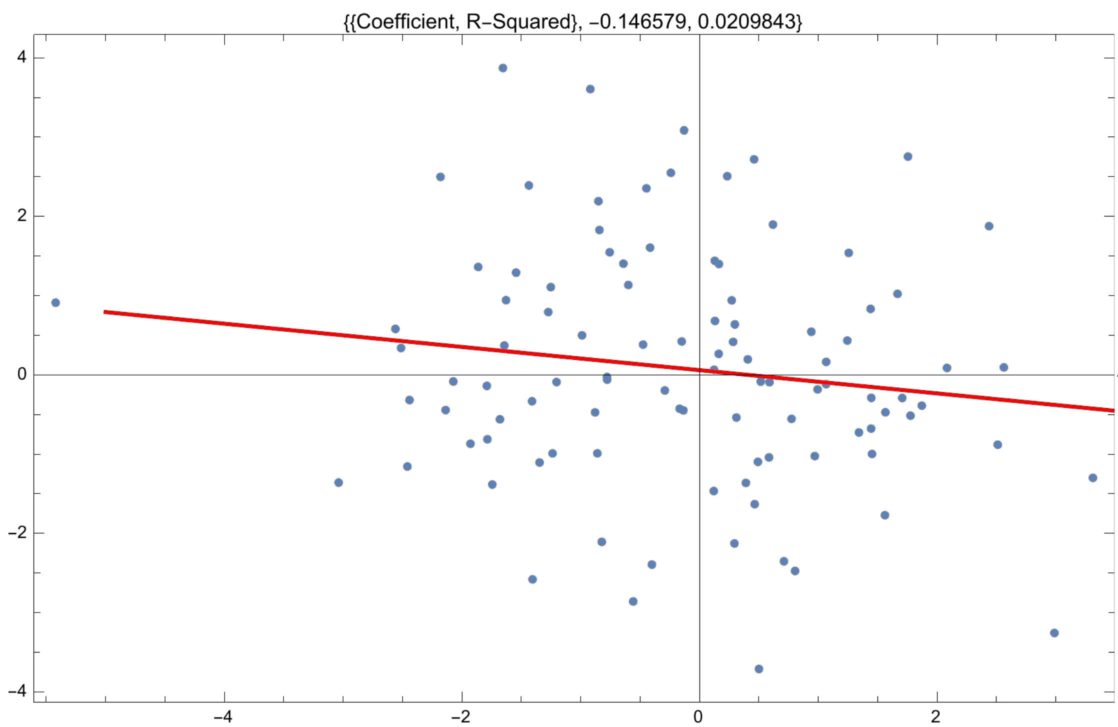
,

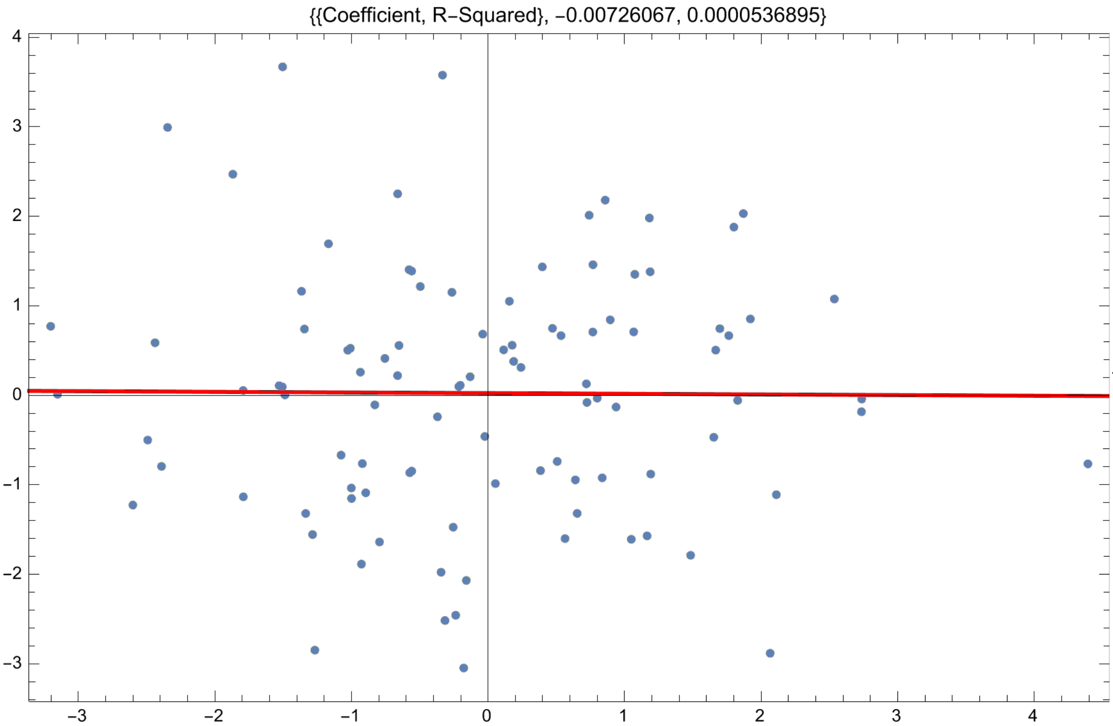
{{Coefficient, R-Squared}, 0.116757, 0.0200819}



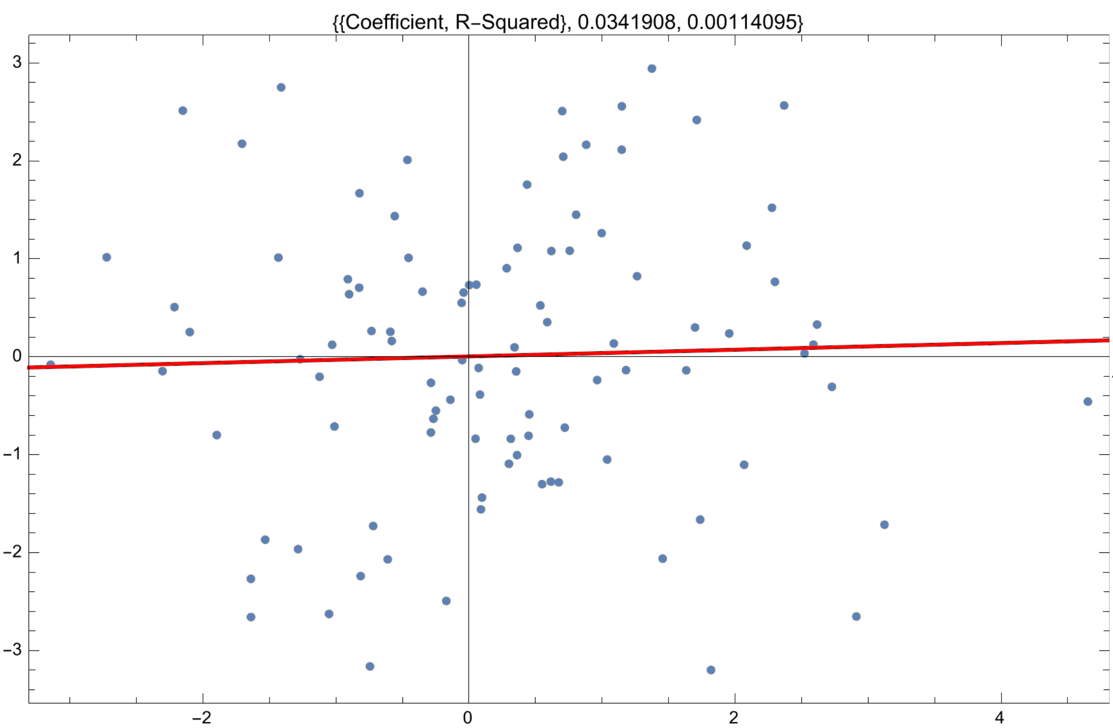
,


{{Coefficient, R-Squared}, -0.128203, 0.0138838}
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

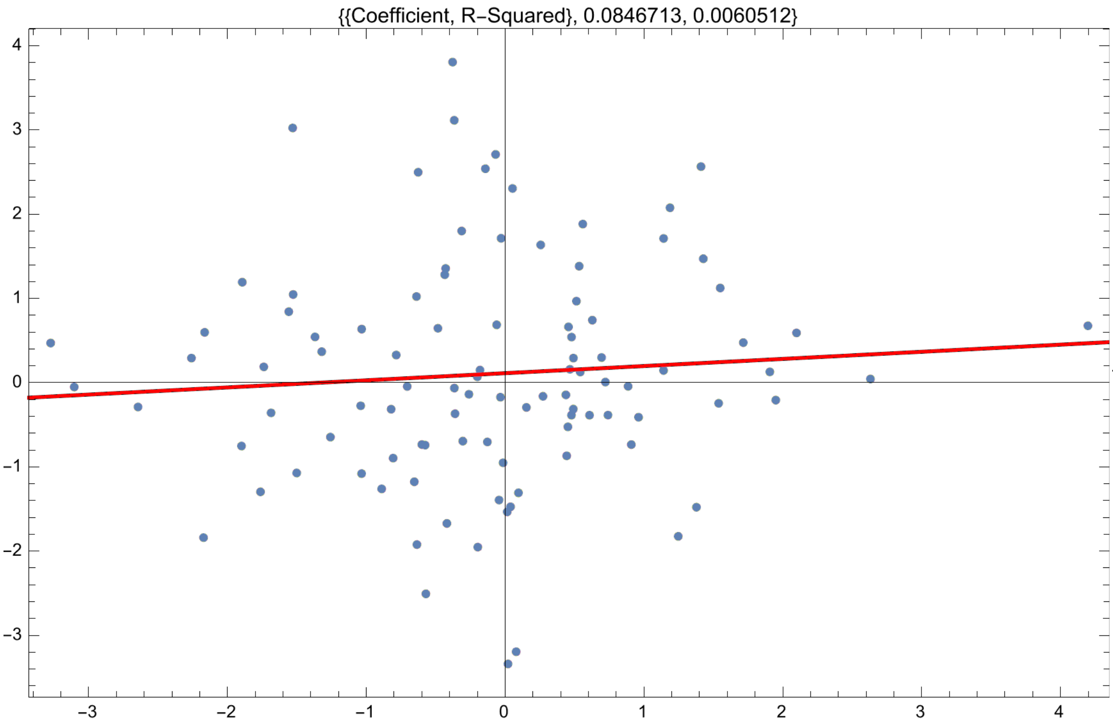

{{Coefficient, R-Squared}, -0.146579, 0.0209843}
4
2
0
-2
-4
-4
-2
0
2


{{Coefficient, R-Squared}, -0.00726067, 0.0000536895}
4
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
4

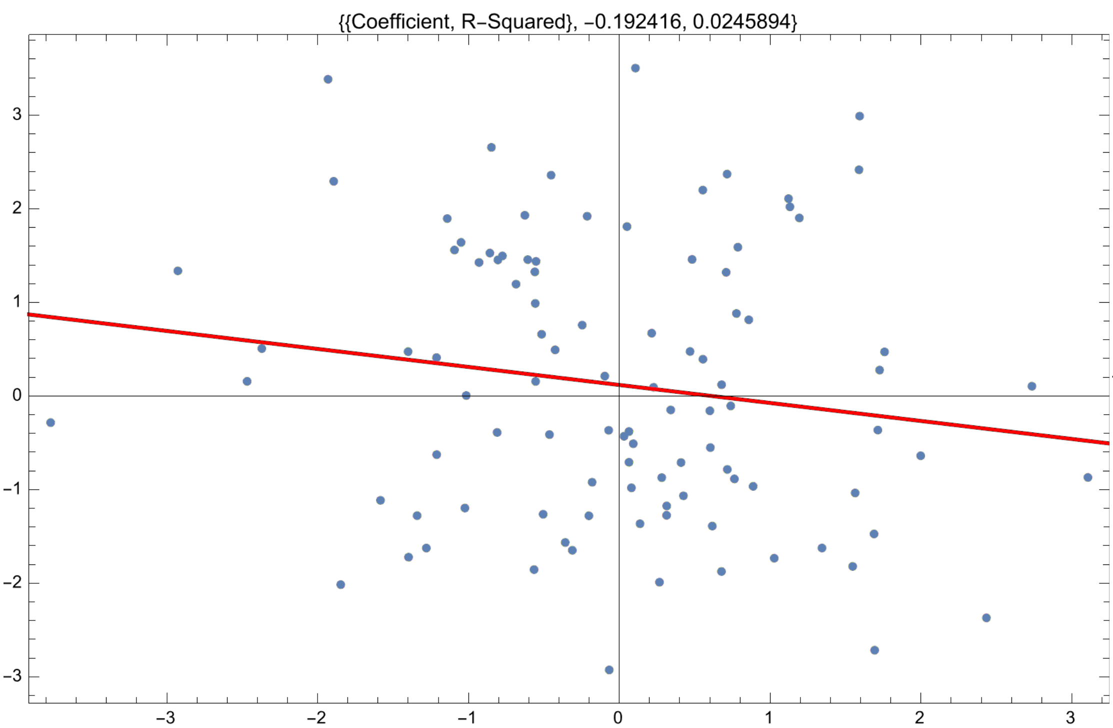
,


{{Coefficient, R-Squared}, 0.0341908, 0.00114095}
3
2
1
0
-1
-2
-3
-2
0
2
4

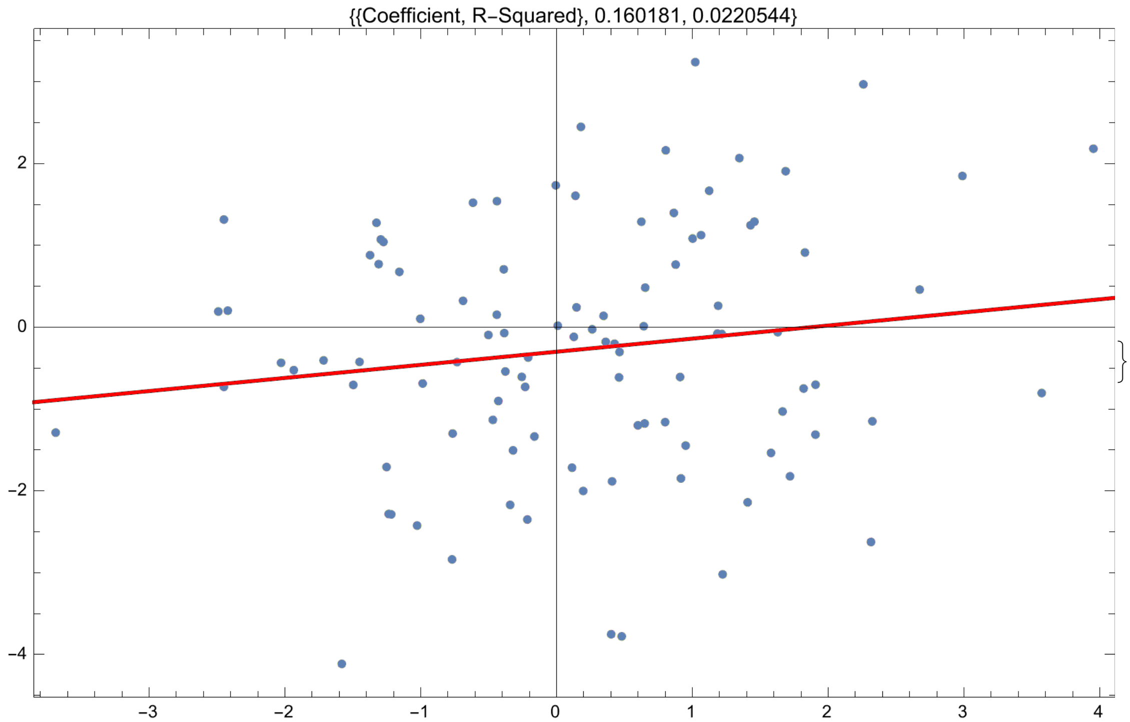
,


{{Coefficient, R-Squared}, 0.0846713, 0.0060512}
4
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
4

,


{{Coefficient, R-Squared}, −0.192416, 0.0245894}
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

,



[Image]

The researcher can choose from many correlations and take the upper bound. We can find the distributions of that upper bound.

There are too many correlations in finance.

Suppose I have do correlation matrix for $0 < i \leq p$, X_i, X_j.

How many correlations we should have? It is $p(p-1)/2$. This is p^2, remove diagonal, then take half. If you add one observation, then it will become $p(p+1)/2$.

Spuriousness decreases at $n^{1/2}$ But spuriousness increases at p^2. This trade off is quite bad.

```
Clear[X]
```

```
f[r_] := ((1 - r^2)^((n - 4) / 2)) / Beta[1 / 2, (1 / 2) (n - 2)]
```

```
dist0 = TransformedDistribution[X^2, X ≈
   ProbabilityDistribution[((1 - r^2)^((n - 4) / 2)) / Beta[1 / 2, (1 / 2) (n - 2)], {r, -1, 1}]]
```

Out[∘]=

TransformedDistribution$\left[X^2, X \approx \text{ProbabilityDistribution}\left[\frac{(1-x^2)^{\frac{1}{2}(-4+n)}}{\text{Beta}\left[\frac{1}{2}, \frac{1}{2}(-2+n)\right]}, \{x, -1, 1\}\right]\right]$

```
PDF[TransformedDistribution[Sqrt[X], X ≈ dist0], r]
```

Out[∘]=

$$\begin{cases} \frac{2\,(1-r^2)^{\frac{1}{2}(-4+n)}}{\text{Beta}\left[\frac{1}{2},\frac{1}{2}(-2+n)\right]} & 0<r<1 \\ 0 & r>1\,||\,r<0 \\ \text{Indeterminate} & \text{True} \end{cases}$$

```
CDF[TransformedDistribution[Sqrt[X], X ≈ dist0], r]
```

Out[∘]=

$$\begin{cases} 1 & r\geq 1 \\ \frac{2\,r\,\text{Hypergeometric2F1}\left[\frac{1}{2},2-\frac{n}{2},\frac{3}{2},r^2\right]}{\text{Beta}\left[\frac{1}{2},\frac{1}{2}(-2+n)\right]} & 0<r<1 \\ 0 & \text{True} \end{cases}$$

```
TransformedDistribution[Sqrt[X], X ≈ dist0] // Mean
```

Out[∘]=

$$\frac{2}{(-2+n)\,\text{Beta}\left[\frac{1}{2},\frac{1}{2}(-2+n)\right]}$$

```
MeanAbs = 2 / ((n - 2) Beta[1 / 2, (1 / 2) (n - 2)]);
MeanAbs /. n → 18 // N
```

Out[∘]=

0.196381

```
(*TransformedDistribution[Sqrt[X],X≈dist0]//Variance*)
```

```
g = (2 * r * HyperGeometric2F1[1 / 2, (2 - n / 2), 3 / 2, r^2]) / (Beta[1 / 2, (1 / 2) * (n - 2)])
```

Out[∘]=

$$\frac{2\,r\,\text{HyperGeometric2F1}\left[\frac{1}{2},2-\frac{n}{2},\frac{3}{2},r^2\right]}{\text{Beta}\left[\frac{1}{2},\frac{1}{2}(-2+n)\right]}$$

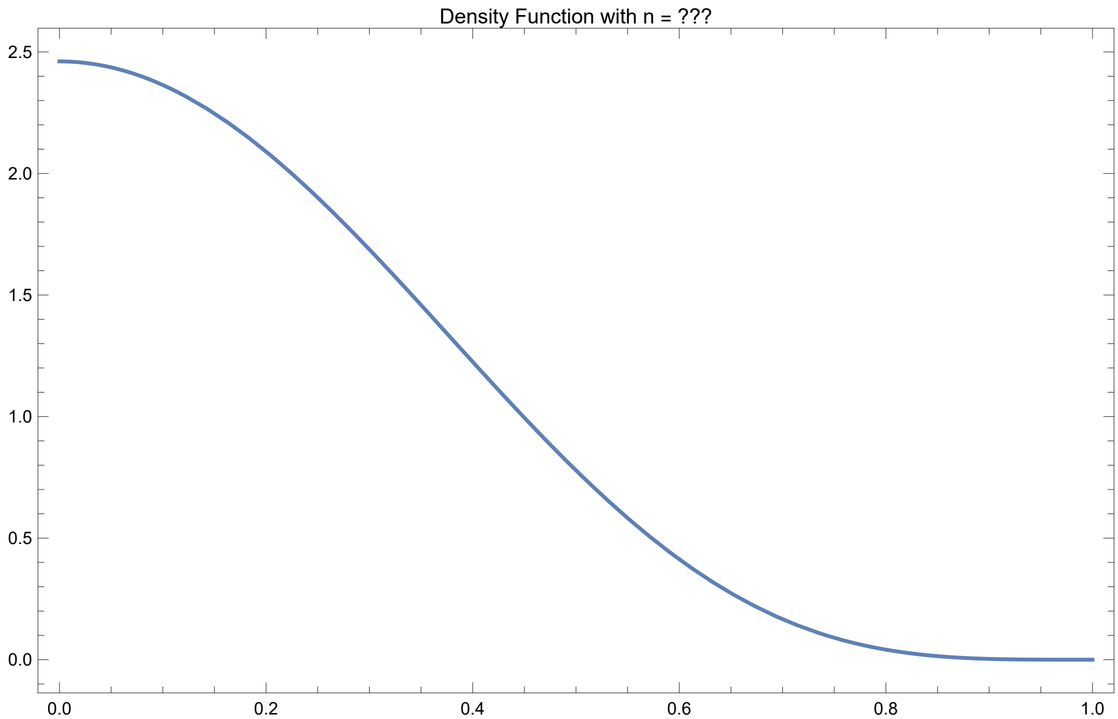
```
fabs[r_] := (2 ((1 - r^2) ^ ((n - 4) / 2))) / (Beta[1 / 2, (1 / 2) (n - 2)])
```

For the special case of $\rho = 0$.

We can determine the upper bound.

```
Plot[fabs[r] /. n → 12, {r, 0, 1},
 AxesLabel → {"Coefficient Coefficient (r)", "Density Function"},
 PlotLabel → "Density Function with n = ???", Frame → True, ImageSize → Large]
```

Out[∘]=



Example of Bad Linear Regressions

[Image]

All you need is a heavy aircraft with a low max-speed for special purposes to debunk this trend line!

[Image]

I had Gell-Mann Amnesia from narrative fallacies watching this guy.

[Image]