DEFENDING DEFENDERS: PROVING LESS IS MORE WITH REGRESSION ANALYSIS

Maths governs everything. From the speed of the sunrise to the eyes that observe it, all can be modelled with simple numbers layered on top of each other to make complex observations. This was, however, probably not the thought going through Liverpool FC's Virgil Van Dijk's mind as he was last bastion of resistance to 6 foot 5 "Viking goal machine" Erling Haaland. Instead, he was furiously sorting through the decades of his experience spent playing the game to calculate how not to let Haaland bury the ball in the back of the net. Curiously enough, the conclusion he came to was to do… nothing! Instead of making a tackle, he used his body as a shield until Erling had to try and shoot around him before he dribbled into the goalkeeper. The shot was tame, and Van Dijk came out on top.[1]

Classical football analysis will tell you that tackling is always a good thing; if you're not getting stuck in enough, you're a liability. This idea is present through all of football, but the most prominent propagation of this story occurs in the very statistics sites that claim to be objective: of the 8 defensive actions (as defined by Opta match actions)[2] measured on the official Premier League stats comparison site, 3 of them are measuring tackles[3]. In other words, the biggest football league in the world says 37.5 percent of a player's defensive capability is purely based on their tackling ability. As I said earlier, maths rules all. So how can we use the numbers available to us to test the assumption that tackling is king?

PART ONE: DO THE ENDS JUSTIFY THE MEANS?

To get our answer we must start at the question. What exactly are we comparing here? Well, we're seeing if more tackling leads to better defending. Obviously, we want to know something about tackles over a set period of time. As football is helpfully played at one game per time, we can use that as our starting point.

First things first we need data. The information in this table was provided by whoscored.com[4] and footystats.org[5]

Team	Avg Tackles pg	Avg conceded pg	Avg Possesion pg
1. Manchester City	12.4	0.87	65.2
2. Arsenal	14.9	1.13	59.7
3. Newcastle	16	0.87	52.2
4. Manchester United	17.3	1.13	53.8
5. Liverpool	15.5	1.24	60.6
6. Brighton	16.2	1.39	60.5
7. Tottenham	16.2	1.66	49.8
8. Brentford	15.4	1.21	43.3
9. Aston Villa	16.7	1.21	49.2
10. Chelsea	19.5	1.24	58.8
11. Everton	18.6	1.5	42.5
12. Crystal Palace	18.2	1.29	45.8
13. Fulham	16.5	1.39	48.6
14. West Ham	16	1.45	41.4
15. Leicester	18.4	1.79	47.7
16. Bournemouth	16.3	1.87	40
17. Wolves	17.4	1.53	50.1
18. Southampton	18.8	1.92	44.1
19. Leeds	22.1	2.05	46.3
20. Nottingham Forest	17.3	1.79	37.2

Second, we want to see how many tackles a team made in all their games in a season (we're only measuring over one season because between seasons a team's quality and style can change so much, it's impossible to directly compare any statistic to quality of defence). Let's say that the number of tackles made in a particular game by a particular team in a particular season is t_i, with *i* representing whichever number game in the season you're selecting or our "index number".

Then, to see the number of tackles made in that season, we add all the t_i values up. In notation, we represent the operation of summing all of a type of thing together as $\sum_{i=z}^{n}$, with *z* representing the index number where you start adding, and *n* being where you stop. So in

our case, we want to start from the first game so *i*=1, and finish at the last game, so we're going to *n*. As the number of games in a season can vary from country to country or even level to level, we just keep *n* as a variable. This means all the tackles a team made over a season can be represented as $\sum_{i=1}^{n} t_i$.

Next, we want to divide this number by the number of data points you're summing, or in our case all the games played in a season (*n*). This will give us the number of tackles normally made in a game, or their mean tackle rate. We will write this as $\bar{t}$ with the bar representing that this new number is the mean value for any given set of numbers. This gives us the equation

$$\frac{\sum_{i=1}^{n} t_i}{n} = \bar{t}$$

For instance, Nottingham Forest tackled 660 times in 38 games[6] in the 2022/23 season, so *n*=38 and

$\sum_{i=1}^{38} t_i$ = 660. Therefore, in Forest's case $\bar{t}=\frac{660}{38} = 17.4$

So now we have our first value; how much is a team normally tackling. Next we want to find how good a team's defence is, which we can measure by goals conceded; if you're defending well, you'll concede less. So according to normal theory, as tackles go up, this number should go down. Replacing tackles with goals conceded per season ($\sum_{i=1}^{n} g_i$) into our previous equation, we get

$$\frac{\sum_{i=1}^{n} g_i}{n}=\bar{g}$$

We're not done yet though. What we've done is found how good a particular team's defence is and how often that particular team tackles. This information is useless on its own, because we need to see how effective one team is compared to everyone else. So, we'll find an average of averages, which we'll represent by prefixing the mean values we already have with *μ*.

$\frac{\sum_{i=1}^{n} \bar{t}_i}{n}=\mu(\bar{t})$ and $\frac{\sum_{i=1}^{n} \bar{g}_i}{n}=\mu(\bar{g})$ which are 16.985 and 1.4265 respectively.

Please note three things: normally, we do not average averages because often different sets of data are larger than others, so it unfairly weights toward some values and skews our final number meaning we almost never see $\bar{}$ and *μ* next to one another. However in this case every team has played the same number of games so they are all equally weighted, so it is

faster to just average the averages (if you want to test it, go to a Premier League team stats site and first average any stat per game for the league average and then average all the stats per team per game for the league average. You get the same number) and that is the best notation I could think of.

Secondly, *n* now represents the number of teams who's averages you're summing, not the number of games in the season. This is because we're summing team averages together now, not game statistics.

Finally, the index number is no longer which game you are looking at, it's now which team you are looking at. This means the index number of a team's tackles and goals conceded will be the same number, because that number now represents the team no matter what order you put the teams in.

Now we know how much a team tackles, how good they are at defending, the league standard for defending and the league standard for tackling, how do we see if tackling correlates with good defence?

PART TWO: MISSING PEARSONS

There are vast numbers of ways to calculate a correlatory link between data points, but we are going to attempt to derive a relatively simple one. First, we want to see if one variable changes, how much can we expect the other to. Or in other words, if a team is tackling more, will they concede more or less goals, and by how much?

This can be found using the covariance formula. When we're calculating covariance, we simply pick a team, see how different their number of tackles and goals conceded are from the mean by subtracting the mean from both.

$(\bar{t}_i - \mu\bar{t})\ and\ (g_i - \mu\bar{g})$

Then, we multiply both differences together to see how much one has influenced the other.

$(\bar{t}_i - \mu\bar{t})(g_i - \mu\bar{g})$

This can be intuitively explained by the following: if a deviation of tackles from the mean only leads to a small deviation of goals conceded from the mean, then when you multiply both deviations together you will get a smaller number than you would if a large number of tackles lead to a larger deviation of goals conceded as a small number multiplied by a big number gives a smaller value than a big number multiplied by a big number.

We add all of these "influence numbers" together, and then divide them by *n-1* to get an "average influence number", which we call our covariance, giving the formula.

$$cov(x, y)(with\ x\ and\ y\ being\ any\ given\ variables) = \frac{\sum_{i=1}^{n}((x_i - \bar{x})(y_i - \bar{y}))}{n-1}$$

Or in our case

$$cov(\bar{g}, \bar{t}) = \frac{\sum_{i=1}^{n}((\bar{t}_i - \mu\bar{t})(g_i - \mu\bar{g}))}{n-1}$$

Note we use *n-1* not *n* as we are using sample means to find our covariance, so we lose a bit of information. Subtracting 1 from *n* accounts for this loss in information. We didn't do this earlier as all the means were equally weighted, so we had lost no information on the data.

If we use the 2022/23 Premier League season statistics, we get a covariance of 0.425 tackle-goals per team per game. This isn't very helpful though, as it only gives us a number specifically in goals to tackles, not just general correlation. It would be much easier to interpret our results if it were a number between 1 and -1, with the closeness to 1 or -1 representing how strong the correlation was. So let's do that.

This can be accomplished by dividing covariance by the standard deviation of each number multiplied. Standard deviation means the usual amount of deviation from the mean that a number has, so the average result of $(x_i - \bar{x})\ or\ (y_i - \bar{y})$. Unfortunately, we cannot find the average result by just adding all the deviations and dividing them by *n*-1 this time, as by definition all the positive deviations and negative deviations from the mean just sum to 0. However, we can square and then square root all deviations to make them all positive and then divide them by $\sqrt{n-1}$ (we divide by $\sqrt{n-1}$ not *n* - 1 because we are still using samples, and we want to account for the fact that the data will probably be much more spread than our little sample, so we want our final answer to be bigger, so we make our denominator smaller and square root it) to find a positive deviation average,

So $\sigma_x(we\ represent\ the\ standard\ deviation\ as\ \sigma) = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$

And

$$\sigma_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}$$

All we need to do now to find our correlation number is divide our covariance by the product of both standard deviations. Intuitively this is because we are taking the real product of our deviations and dividing it by the product of what they should be, which both cancels our units as we're dividing x and y by x and y giving us a value between 1 and -1 and tells us how much one influences the other no matter how random our spread of data should be. We can quantify this as

$$r_{xy}(representing\ the\ correlation\ number) = \frac{cov(x,y)}{\sigma_x\sigma_y}$$

or

$$r_{xy} = \frac{\frac{\sum_{i=1}^{n}((x_i-\bar{x})(y_i-\bar{y}))}{n-1}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}\sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}}$$

We can multiply the $\sqrt{n-1}$ values to get the equation

$$r_{xy} = \frac{\frac{\sum_{i=1}^{n}((x_i-\bar{x})(y_i-\bar{y}))}{n-1}}{\frac{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}{n-1}}$$

We cancel the *n* - 1 values and expand our square root to get

$$r_{xy} = \frac{\sum_{i=1}^{n}((x_i-\bar{x})(y_i-\bar{y}))}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

This is called Pearson's correlation formula, and the number we are calculating is called Pearson's correlation coefficient. This is a branch of statistics called regression analysis[7], because we are measuring how numbers regress from the mean.

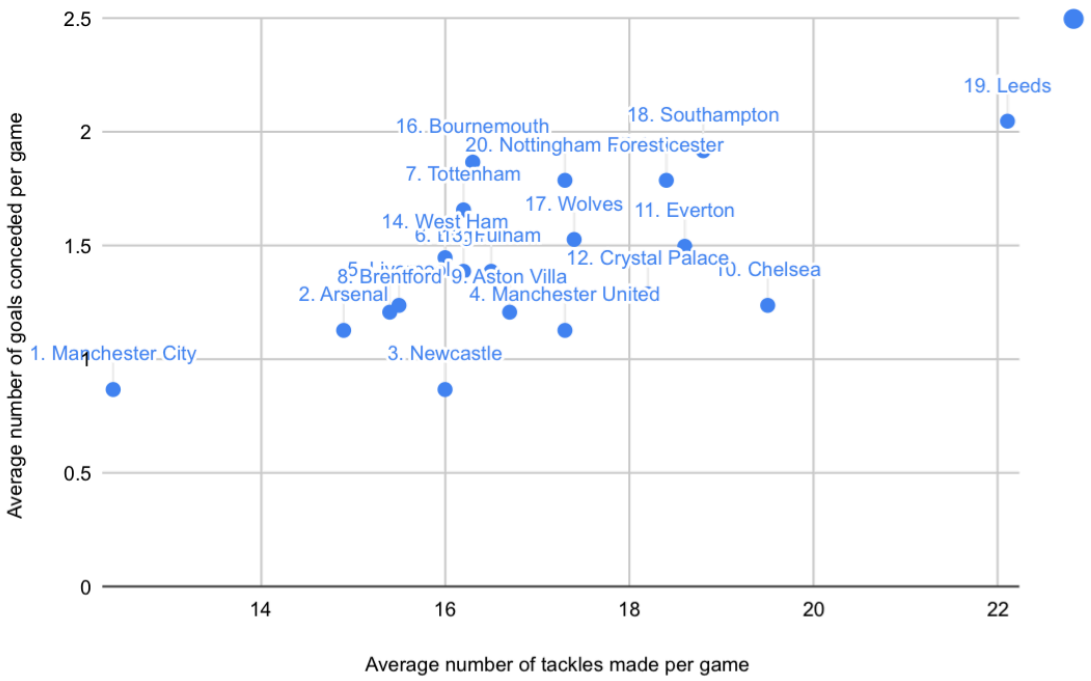
Plugging our Premier League data into the equation, we get $r_{t,g}$=0.632.

With this number, however, there is a very small chance that this could be a coincidence. We must note how small that chance is to show that our answers remain valid, and also to calculate the strength of our correlation. The boundaries of different probabilities, or "critical values" can helpfully be looked up in a table, where you subtract two from the *n* value of numbers in your set and go along from the side. This table is provided by the University of Sussex[8]

df:	0.1	0.05	0.02	0.01
1	.988	.997	.9995	.9999
2	.9	.95	.98	.99
3	.805	.878	.934	.959
4	.729	.811	.882	.917
5	.669	.754	.833	.874
6	.622	.707	.789	.834
7	.582	.666	.75	.798
8	.549	.632	.716	.765
9	.521	.602	.685	.735
10	.497	.576	.658	.708
11	.476	.553	.634	.684
12	.458	.532	.612	.661
13	.441	.514	.592	.641
14	.426	.497	.574	.623
15	.412	.482	.558	.606
16	.4	.468	.542	.59
17	.389	.456	.528	.575
18	.378	.444	.516	.561
19	.369	.433	.503	.549
20	.36	.423	.492	.537
21	.352	.413	.482	.526
22	.344	.404	.472	.515
23	.337	.396	.462	.505
24	.33	.388	.453	.496

If we subtract two from our *n* value, we get 18. Going along, we see that the probability of a Pearson number to be coincidence beyond .561 is <0.01, and as our value is 0.632 we and 0.632>0.561 there is less than 1% chance of our number being random and we have a strong correlation.

This means when you tackle more, you concede more. Plotting our points onto a graph we can visualise the data like this:



Hm. Now that's surprising. Or is it?

PART 3: RUSSIAN DOCTORS

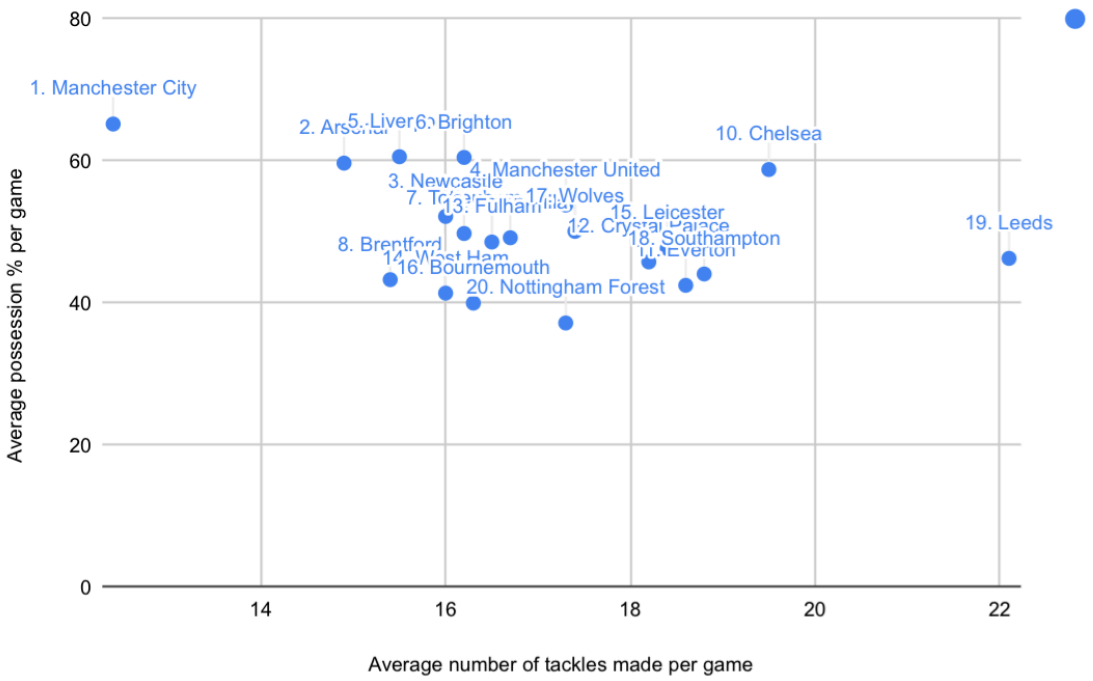
In the 19th century, Russia underwent a cholera epidemic. This was devastating to the nation, so doctors were sent from the capital to find and eliminate outbreaks as soon as they were reported. The issue is, they were too good at rapid response. When locals began noticing that as soon as doctors turned up to a village, people began dying of cholera they put two and two together to make five and began chasing the doctors from their villages. Needless to say, this did not bode well for the doctors or the villagers.[9] So, how can we avoid their mistakes?

The point of that story is to illustrate that correlation does not necessarily imply causation, and even though one thing might happen increasingly with something else, it does not cause the other to occur. While it is true that you require correlation for causation to be drawn, we do not always draw causation from correlation. So, while we have already proved that a team that tackles more probably has a worse defence, with the numbers we have we cannot say tackling is making that defence better or worse. That, however, is what we want to find out. To do this, we must separate the strength of defence from the strength of the team.

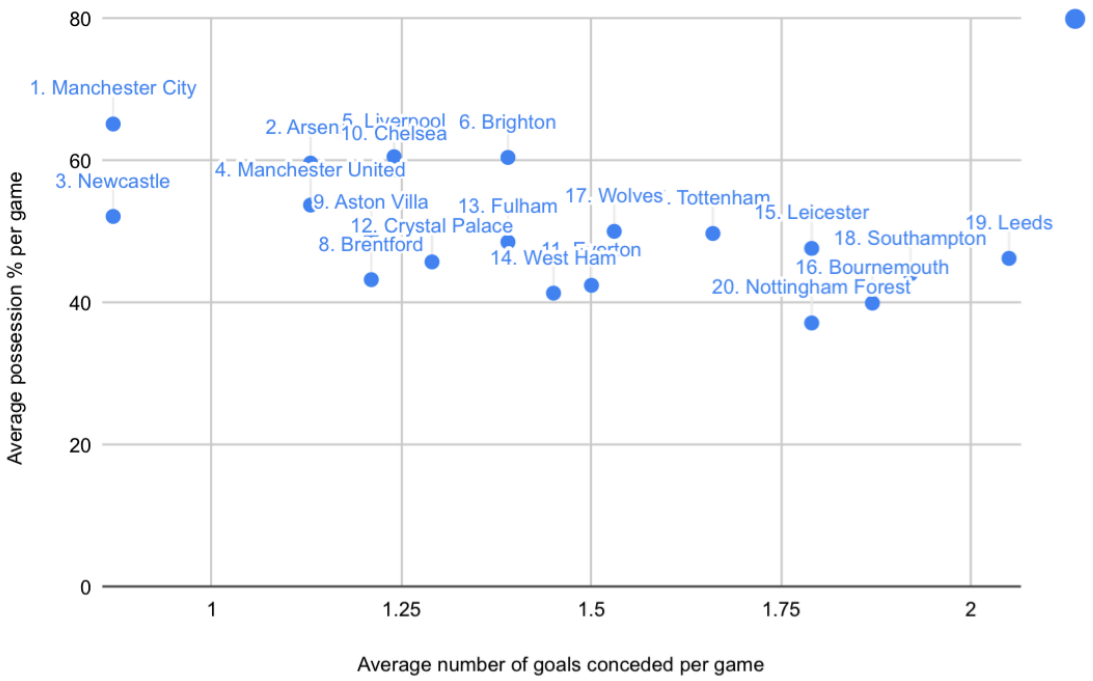
To measure the strength of a team I will use ball possession per game, or for how much of a game is a team on the ball for. This is for two reasons: firstly, possession has a very strong influence on both goals conceded and tackling. If a team is on the ball a lot, they both do not tackle the ball. On top of this, when they have possession, they can control where the ball goes, namely the not the back of their own net. Secondly, possession is a very good

indicator for team strength as that is almost always the objective of top teams in modern football, especially English football, so the better teams will have more possession. As possession influences both of our statistics, we call it a confounding variable. Let us define percentage possession for a particular team for a particular game as p_i.

Using the maths we did in the last two parts we can calculate $r_{\overline{p,t}} = -0.4033$ which is graphically represented as:



As predicted, there is an inverse correlation between tackling and possession. Next, we calculate $r_{\overline{p,g}} = -0.6298$, which is graphically represented as (turn to the next page):



What we want to see is how many goals a team's defence concedes against tackles, bearing in mind possession.

To calculate correlation between x, y bearing in mind the influence of z *(possession)*, we take $r_{x,y}$ and subtract the product of $r_{x,z}$ and $r_{y,z}$ from it. This is because we are doing a very similar thing as what we were doing in Pearson's earlier: the product of $r_{x,z}$ and $r_{y,z}$ represents how much they change with one another, so subtracting them from the general correlation should account for their influence. Then, as we did with Pearson's, we divide by $\sqrt{(1 - r_{x,z}^{\ 2})(1 - r_{y,z}^{\ 2})}$ to control for how far away from 1 possession based *r* values should be and to normalise the equation to between 1 and -1 again. This gives us the equation

$$\frac{r_{x,y} - (r_{x,z} r_{y,z})}{\sqrt{(1 - r_{x,z}^{\ 2})(1 - r_{y,z}^{\ 2})}} = r_{x,y.z}$$

where $r_{x,y.z}$ = the correlation bearing in mind the influence of *z,* also called the partial correlation coefficient. Finally, the effect of tackles to goals conceded bearing in mind team strength is $\frac{0.632 - (-0.6298 \cdot -0.4033)}{\sqrt{(1 - (-0.4033)^2)(1 - (-0.6298)^2}} = 0.5318$

CONCLUSION:

Calculating the critical values for the partial correlation coefficient is much more complex than for our *r* value and there are no readily available tables we can use online, but merely looking at our value and the reasonable size of our dataset shows us that this number is statistically significant.

This number suggests to us that between the ranges of average tackles teams made in the 2022/23 season in the Premier League, even accounting for different team strengths influencing rate of tackling and goals conceded naturally, telling your team to tackle has a correlative relationship with conceding goals. As we have removed our confounding variable, the data suggests a causal relationship between tackling and conceding. This may be because when Premier League teams encourage tackling, they neglect other defensive techniques or because a team that is tackling is a team that is not employing a proper "defence in depth" strategy and is pushing too high up to win the ball. Whatever the reason, tackling more is not a good thing, in fact, it is a bad thing and telling everyone to get their boots muddy will often spell doom for the manager.

REFERENCES

[1]https://www.skysports.com/football/video/33653/13092078/two-juggernauts-going-for-it-haaland-v-van-dijk

[2]https://www.statsperform.com/opta-event-definitions/

[3]https://www.premierleague.com/stats/player-comparison

[4]https://www.whoscored.com/Regions/252/Tournaments/2/Seasons/9075/Stages/20934/TeamStatistics/England-Premier-League-2022-2023

[5]https://footystats.org/england/premier-league/goals-conceded-table (2022/23 season)

[6]https://www.premierleague.com/clubs/15/Nottingham-Forest/stats?se=489 (2022/23 season)

[7]https://www.biostat.jhsph.edu/courses/bio653/misc/JMPer%20Cable%20Summer%2098%20Why%20is%20it%20called%20Regression.htm#:~:text=For%20example%2C%20if%20parents%20were,meaning%20to%20come%20back%20to.

[8]https://users.sussex.ac.uk/~grahamh/RM1web/Pearsonstable.pdf

[9]https://homework.study.com/explanation/in-the-early-19th-century-the-russian-government-sent-doctors-to-southern-russian-villages-to-provide-assistance-during-a-cholera-epidemic-the-villagers-noticed-that-wherever-doctors-appeared-people-died-therefore-many-doctors-were-chased-away-from-v.html

BIBLIOGRAPHY

https://www.scribbr.com/statistics/pearson-correlation-coefficient/#:~:text=The%20Pearson%20correlation%20coefficient%20(r,the%20relationship%20between%20two%20variables.&text=When%20one%20variable%20changes%2C%20the,changes%20in%20the%20same%20direction.

https://www.biostat.jhsph.edu/courses/bio653/misc/JMPer%20Cable%20Summer%2098%20Why%20is%20it%20called%20Regression.htm#:~:text=For%20example%2C%20if%20parents%20were,meaning%20to%20come%20back%20to.

https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#

https://www.youtube.com/watch?v=N7bZnC_a01M

https://www.investopedia.com/terms/c/covariance.asp

https://www.socscistatistics.com/tests/pearson/default2.aspx

https://www.youtube.com/watch?app=desktop&v=OpAf4N582bA