When variances don't add

Donna Dietz

American University

dietz@american.edu

STAT 202 - Spring 2020

Review Question: LSR does what?

Least Squares Regression (LSR, or a line of best fit) helps us to quantify the connectedness between two variables. So this means they may be somewhat random if taken alone, but taken as pairs, they have something to do with each other.

The whole semester is bringing us here!

The whole semester has been a buildup to this idea, and we've been talking about it all along. These last three discussions should hopefully seal these ideas firmly into your mind.

?!

Are two things connected or aren't they?

Data type gets in the way

Since data can take different forms, this same fundamental question appears to take various forms, so we need to weild a variety of tools to answer still, that same fundamental question.

Examples

- Numerical vs. Numerical (LSR)
- Numerical vs. Categorical (ANOVA and t-tests)
- Categorical vs. Categorical (Chi-Square)
- ... and many more we haven't discussed!

A progression of connectedness

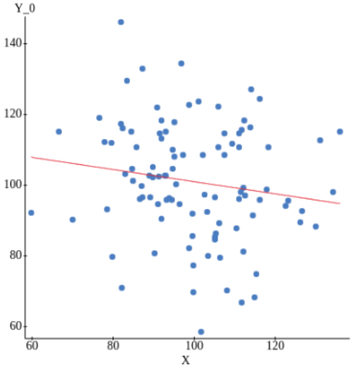
Simple linear regression results:
Dependent Variable: Y_0
Independent Variable: X
Y_0 = 118.17895 - 0.17108076 X
Sample size: 100
R (correlation coefficient) = -0.15690574
R-sq = 0.024619411
Estimate of error standard deviation: 16.141497

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	118.17895	10.967153	≠ 0	98	10.775718	<0.0001
Slope	-0.17108076	0.10877682	≠ 0	98	-1.5727685	0.119

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	644.49158	644.49158	2.4736009	0.119
Error	98	25533.697	260.54793		
Total	99	26178.189			



What I'm doing

With each successive slide, you will see the connectedness increasing.

As we progress, I keep giving the Y variables more of the X value.

Example:

$$Y\_3 = 0.7 \cdot Y\_0 + 0.3 \cdot X$$

Ironically, the first slide in the sequence shows variables which have a negative correlation. So numerically, the least correlation happens in the second slide after I started giving the response variable some of the explanatory variable!

What to watch for

- Horizontal line turns into $y = x$ line.
- Points start to hug the LSR line.
- R^2 increases to 100%
- $p - value$ of the slope increases
- The intercept starts near the average y-value (100).
- The intercept goes to zero progressively.
- The standard error of the intercepts decreases.

A progression of connectedness - 1

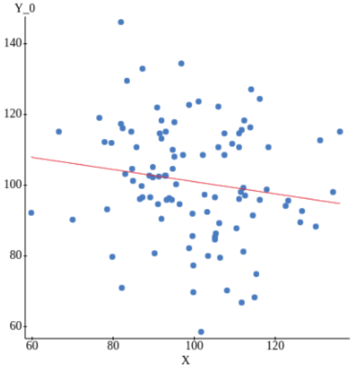
Simple linear regression results:
Dependent Variable: Y_0
Independent Variable: X
Y_0 = 118.17895 - 0.17108076 X
Sample size: 100
R (correlation coefficient) = -0.15690574
R-sq = 0.024619411
Estimate of error standard deviation: 16.141497

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	118.17895	10.967153	≠ 0	98	10.775718	<0.0001
Slope	-0.17108076	0.10877682	≠ 0	98	-1.5727685	0.119

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	644.49158	644.49158	2.4736009	0.119
Error	98	25533.697	260.54793		
Total	99	26178.189			



A progression of connectedness - 2

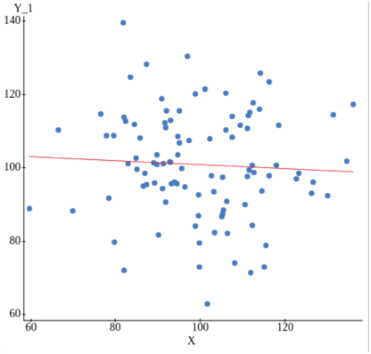
Simple linear regression results:
Dependent Variable: Y_1
Independent Variable: X
Y_1 = 106.36105 - 0.05397268 X
Sample size: 100
R (correlation coefficient) = -0.055604464
R-sq = 0.0030918565
Estimate of error standard deviation: 14.527347

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	106.36105	9.8704377	≠ 0	98	10.775718	<0.0001
Slope	-0.05397268	0.097899136	≠ 0	98	-0.55130906	0.5827

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	64.145013	64.145013	0.30394168	0.5827
Error	98	20682.295	211.04382		
Total	99	20746.44			



A progression of connectedness - 3

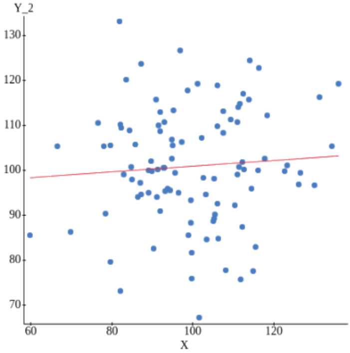
Simple linear regression results:
Dependent Variable: Y_2
Independent Variable: X
Y_2 = 94.543159 + 0.063135395 X
Sample size: 100
R (correlation coefficient) = 0.07309208
R-sq = 0.0053424522
Estimate of error standard deviation: 12.913198

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	94.543159	8.7737224	≠ 0	98	10.775718	<0.0001
Slope	0.063135395	0.087021454	≠ 0	98	0.72551529	0.4699

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	87.772959	87.772959	0.52637244	0.4699
Error	98	16341.566	166.75067		
Total	99	16429.339			



A progression of connectedness - 4

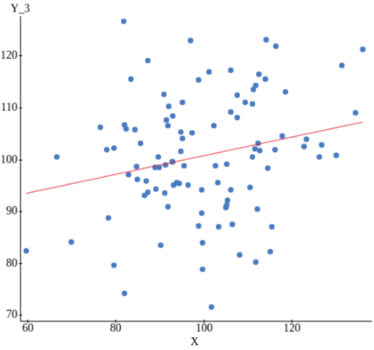
Simple linear regression results:
Dependent Variable: Y_3
Independent Variable: X
Y_3 = 82.725264 + 0.18024347 X
Sample size: 100
R (correlation coefficient) = 0.2325617
R-sq = 0.054084942
Estimate of error standard deviation: 11.299048

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	82.725264	7.6770071	≠ 0	98	10.775718	<0.0001
Slope	0.18024347	0.076143772	≠ 0	98	2.3671466	0.0199

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	715.37542	715.37542	5.603383	0.0199
Error	98	12511.512	127.66848		
Total	99	13226.887			



A progression of connectedness - 5

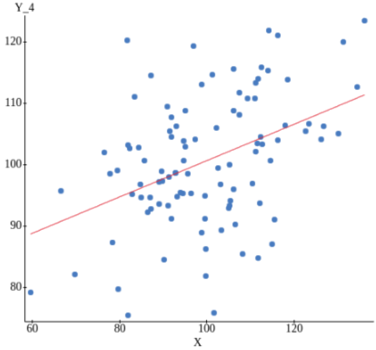
Simple linear regression results:
Dependent Variable: Y_4
Independent Variable: X
Y_4 = 70.907369 + 0.29735155 X
Sample size: 100
R (correlation coefficient) = 0.41807379
R-sq = 0.17478569
Estimate of error standard deviation: 9.6848983

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	70.907369	6.5802918	≠ 0	98	10.775718	<0.0001
Slope	0.29735155	0.06526609	≠ 0	98	4.5559883	<0.0001

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	1946.9524	1946.9524	20.75703	<0.0001
Error	98	9192.1309	93.797254		
Total	99	11139.083			



A progression of connectedness - 6

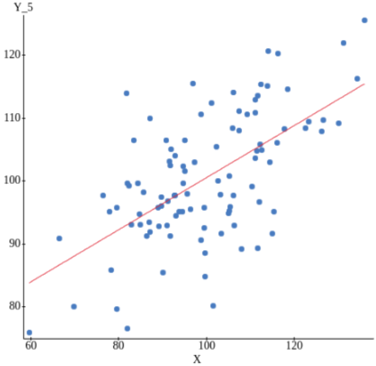
Simple linear regression results:
Dependent Variable: Y_5
Independent Variable: X
Y_5 = 59.089475 + 0.41445962 X
Sample size: 100
R (correlation coefficient) = 0.60998081
R-sq = 0.37207659
Estimate of error standard deviation: 8.0707485

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	59.089475	5.4835765	≠ 0	98	10.775718	<0.0001
Slope	0.41445962	0.054388409	≠ 0	98	7.6203668	<0.0001

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	3782.5039	3782.5039	58.06999	<0.0001
Error	98	6383.4242	65.136982		
Total	99	10165.928			



A progression of connectedness - 7

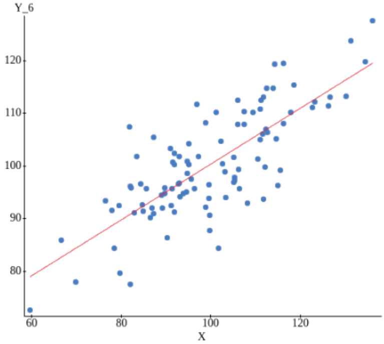
Simple linear regression results:
Dependent Variable: Y_6
Independent Variable: X
Y_6 = 47.27158 + 0.5315677 X
Sample size: 100
R (correlation coefficient) = 0.77694635
R-sq = 0.60364563
Estimate of error standard deviation: 6.4565988

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	47.27158	4.3868612	≠ 0	98	10.775718	<0.0001
Slope	0.5315677	0.043510727	≠ 0	98	12.216934	<0.0001

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	6222.0299	6222.0299	149.25349	<0.0001
Error	98	4085.3915	41.687669		
Total	99	10307.421			



A progression of connectedness - 8

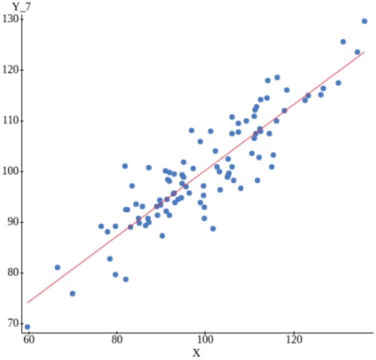
Simple linear regression results:
Dependent Variable: Y_7
Independent Variable: X
Y_7 = 35.453685 + 0.64867577 X
Sample size: 100
R (correlation coefficient) = 0.89513658
R-sq = 0.8012695
Estimate of error standard deviation: 4.8424491

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	35.453685	3.2901459	≠ 0	98	10.775718	<0.0001
Slope	0.64867577	0.032633045	≠ 0	98	19.877881	<0.0001

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	9265.5304	9265.5304	395.13013	<0.0001
Error	98	2298.0327	23.449314		
Total	99	11563.563			



A progression of connectedness - 9

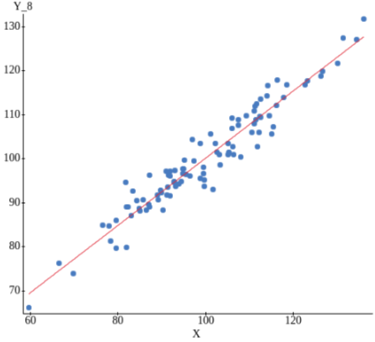
Simple linear regression results:
Dependent Variable: Y_8
Independent Variable: X
Y_8 = 23.63579 + 0.76578385 X
Sample size: 100
R (correlation coefficient) = 0.96265408
R-sq = 0.92670289
Estimate of error standard deviation: 3.2282994

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	23.63579	2.1934306	≠ 0	98	10.775718	<0.0001
Slope	0.76578385	0.021755363	≠ 0	98	35.199773	<0.0001

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	12913.005	12913.005	1239.024	<0.0001
Error	98	1021.3479	10.421917		
Total	99	13934.353			



A progression of connectedness - 10

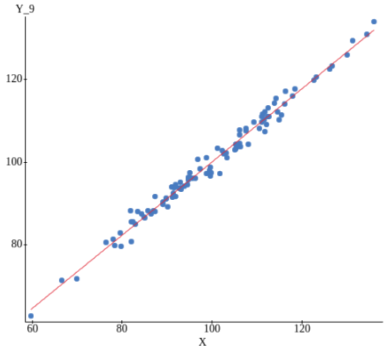
Simple linear regression results:
Dependent Variable: Y_9
Independent Variable: X
Y_9 = 11.817895 + 0.88289192 X
Sample size: 100
R (correlation coefficient) = 0.99264401
R-sq = 0.98534213
Estimate of error standard deviation: 1.6141497

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	11.817895	1.0967153	≠ 0	98	10.775718	<0.0001
Slope	0.88289192	0.010877682	≠ 0	98	81.165449	<0.0001

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	17164.455	17164.455	6587.8301	<0.0001
Error	98	255.33697	2.6054793		
Total	99	17419.792			



A progression of connectedness - 11

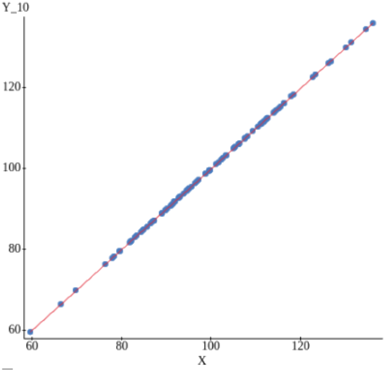
Simple linear regression results:
Dependent Variable: Y_10
Independent Variable: X
Y_10 = 0 + 1 X
Sample size: 100
R (correlation coefficient) = 1
R-sq = 1
Estimate of error standard deviation: 0

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	0	0	≠ 0	98	NaN	NaN
Slope	1	0	≠ 0	98	Infinity	NaN

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	22019.879	22019.879	Infinity	<0.0001
Error	98	0	0		
Total	99	22019.879			



$y = x$ line

The reason we approach the $y = x$ line is that both data sets were originally from the same population, and the points themselves were approaching the $y = x$ line. Any perfectly correlated points would lie on some line, but not necessarily a $y = x$ line.

What does a hypothesis test do?

So, what does this have to do with hypothesis testing?

My goal for you in this course is to see the connections between the concepts, not to just see the course contents as stand-alone tricks you can do to data.

Hypothesis testing in this course is to ask the question of whether two variables are connected or disconnected.

Example

In the progression we just saw, the initial slide showed randomly generated points which had no underlying reason to be correlated, yet they were negatively correlated!

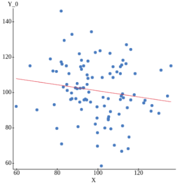
Simple linear regression results:
Dependent Variable: Y_0
Independent Variable: X
Y_0 = 118.17895 - 0.17108076 X
Sample size: 100
R (correlation coefficient) = -0.15690574
R-sq = 0.024619411
Estimate of error standard deviation: 16.141497

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	118.17895	10.967153	≠ 0	98	10.775718	<0.0001
Slope	-0.17108076	0.10877682	≠ 0	98	-1.5727685	0.119

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	644.49158	644.49158	2.4736009	0.119
Error	98	25533.697	260.54793		
Total	99	26178.189			



But the $p - value$ for the slope correctly warned us not to be too confident in making a statement about their connectedness! If we'd had a hypothesis that these variables were connected, the $p - value$ would have kept us from making a false statement.

But...

But, sadly, once I started mixing the variables - and I did indeed do this - the $p - value$ STILL told me to be cautious, because the sample I had did not provide enough evidence that this mixing was taking place!

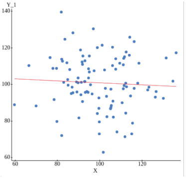
Simple linear regression results:
Dependent Variable: Y_1
Independent Variable: X
Y_1 = 106.36105 - 0.05397268 X
Sample size: 100
R (correlation coefficient) = -0.055604464
R-sq = 0.0030918565
Estimate of error standard deviation: 14.527347

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	106.36105	9.8704377	≠ 0	98	10.775718	<0.0001
Slope	-0.05397268	0.097899136	≠ 0	98	-0.55130906	0.5827

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	64.145013	64.145013	0.30394168	0.5827
Error	98	20682.295	211.04382		
Total	99	20746.44			



I only know it was happening, because I was the one doing it!

Finally

But eventually, the mixing became apparent!

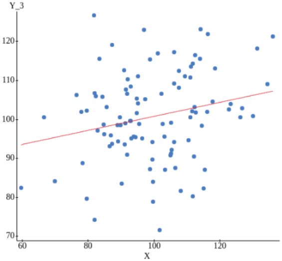
Simple linear regression results:
Dependent Variable: Y_3
Independent Variable: X
Y_3 = 82.725264 + 0.18024347 X
Sample size: 100
R (correlation coefficient) = 0.2325617
R-sq = 0.054084942
Estimate of error standard deviation: 11.299048

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	82.725264	7.6770071	≠ 0	98	10.775718	<0.0001
Slope	0.18024347	0.076143772	≠ 0	98	2.3671466	0.0199

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	715.37542	715.37542	5.603383	0.0199
Error	98	12511.512	127.66848		
Total	99	13226.887			



We could have hypothesized that these variables were connected. We would have said they were, and we'd have been right.

What does the $p - value$ do?

The $p - value$ answers a conditional probability question.

If the two variables (or sets of various variables) are **not** connected, what is the likelihood that we could see a sample with this amount of correlation in it, just because we randomly drew it out that way?

Note: This ignores many important considerations like poorly designed studies and other issues with experimental design and only focuses on the pure theoretical question about sampling errors!

Then...

Then, this gives us a $p - value$ or perhaps a confidence interval that seeks to answer our conditional probablity question.

We wish we could ask...

What it doesn't give us, but what we of course wish we could ask, is "What's the probability that there **is** a connection between these variables?"

Sadly, we can never answer this question, except in quite rare situations.

Lines today, other things later

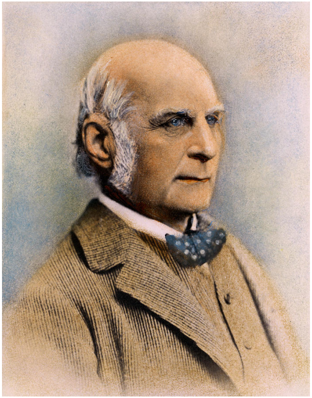
This discussion is about Numerical vs. Numerical data. But the same type of discussion will surround ANOVA, t-tests, Chi-Square and many other similar tests which seek to answer the same fundamental question about connectedness.

Galton Families Data

The data sets in the worksheets are about SAT scores and college GPA, as well as High School GPA. But the data set for this discussion will be a famous data set from Francis Galton (1886).

https://vincentarelbundock.github.io/Rdatasets/doc/HistData/GaltonFamilies.html

Sir Francis Galton (16 February 1822 – 17 January 1911)



Galton produced over 340 papers and books. He also created the statistical concept of correlation and widely promoted regression toward the mean. (Wikipedia)

Data Overview

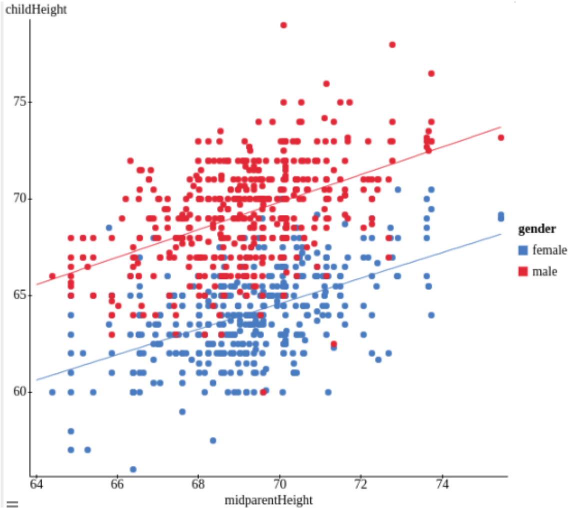
GaltonFamilies.csv

StatCrunch ▾ Applets ▾ Edit ▾ Data ▾ Stat ▾ Graph ▾

Row	family	father	mother	midparentHeigl	children	childNum	gender	childHeight
1	1	78.5	67	75.43	4	1	male	73.2
2	1	78.5	67	75.43	4	2	female	69.2
3	1	78.5	67	75.43	4	3	female	69
4	1	78.5	67	75.43	4	4	female	69
5	2	75.5	66.5	73.66	4	1	male	73.5
6	2	75.5	66.5	73.66	4	2	male	72.5
7	2	75.5	66.5	73.66	4	3	female	65.5
8	2	75.5	66.5	73.66	4	4	female	65.5
9	3	75	64	72.06	2	1	male	71
10	3	75	64	72.06	2	2	female	68
11	4	75	64	72.06	5	1	male	70.5
12	4	75	64	72.06	5	2	male	68.5
13	4	75	64	72.06	5	3	female	67
14	4	75	64	72.06	5	4	female	64.5
15	4	75	64	72.06	5	5	female	63
16	5	75	58.5	69.09	6	1	male	72
17	5	75	58.5	69.09	6	2	male	69
18	5	75	58.5	69.09	6	3	male	68
19	5	75	58.5	69.09	6	4	female	66.5
20	5	75	58.5	69.09	6	5	female	62.5
21	5	75	58.5	69.09	6	6	female	62.5

We are looking at some families and their heights with gender. Birth order is not given, although it appears to be. It's not.

Scatterplot by gender



The average of the two parents' heights is the explanatory variable, and the (adult) child's height is the response.

Some results

male children

$$child = 20 + 0.71 parent$$

female children

$$child = 18 + 0.66 parent$$

Note here how ridiculous the intercepts are, and how utterly meaningless. This says that if a parent is zero inches tall, the male children will end up 2 inches taller than their sisters.

Really, if you look at the range of the data, the difference between the male and female children is between 5.2 and 5.8 inches.

To simplify

To simplify, we don't have to group by gender.

$$child = 22.6 + 0.637 parent$$

Where, again, the parent height means the average of the two parents.

Can we find the LSR line?

There is a process we are not covering in this class, and it feels the same as finding a standard deviation. There are lots of subtractions, squaring, and square rooting.

But if the standard deviations for each variable individually as well as the correlation between the two variables has already been calculated, you can use those values to find the LSR line!

Formula time

$$y = a + bx$$

$$if \quad b = r\frac{s_y}{s_x}$$

$$and \quad a = \bar{y} - b \cdot \bar{x}$$

and r is the correlation between x and y, with s_x and s_y being the sample standard deviations and $\bar{x}$ and $\bar{y}$ being the means.

Let's try it!

Simple linear regression results:
Dependent Variable: childHeight
Independent Variable: midparentHeight
childHeight = 22.636241 + 0.6373609 midparentHeight
Sample size: 934
R (correlation coefficient) = 0.3209499
R-sq = 0.10300884
Estimate of error standard deviation: 3.3917132

Summary statistics:

Column	Mean	Std. dev.	Std. err.
midparentHeight	69.206773	1.8023702	0.058975355
childHeight	66.745931	3.5792512	0.11711668

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	22.636241	4.2651074	≠ 0	932	5.3073084	<0.0001
Slope	0.6373609	0.061607602	≠ 0	932	10.345491	<0.0001

We need $r, \bar{x}, \bar{y}, s_x$, and s_y.
Let's use $0.321, 69.2, 66.75, 1.80$, and 3.58.

$$b = 0.321 \frac{3.58}{1.80} = 0.638$$

$$a = 66.75 - 0.638 \cdot 69.21 = 22.6$$

Which agrees fairly well with $\textit{child} = 22.6 + 0.637\textit{parent}$ from StatCrunch!

Standard Error

Simple linear regression results:
Dependent Variable: childHeight
Independent Variable: midparentHeight
childHeight = 22.636241 + 0.6373609 midparentHeight
Sample size: 934
R (correlation coefficient) = 0.3209499
R-sq = 0.10300884
Estimate of error standard deviation: 3.3917132

Summary statistics:

Column	Mean	Std. dev.	Std. err.
midparentHeight	69.206773	1.8023702	0.058975355
childHeight	66.745931	3.5792512	0.11711668

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	22.636241	4.2651074	≠ 0	932	5.3073084	<0.0001
Slope	0.6373609	0.061607602	≠ 0	932	10.345491	<0.0001

Note: **Standard Error** is another term for the standard deviation of the sampling distribution.

$$s_x = 1.80237 \;\; so \;\; SE_x = 1.80237/\sqrt{934} = 0.058975$$

$$s_y = 3.57925 \;\; so \;\; SE_y = 3.5925/\sqrt{934} = 0.117117$$

Why is s_x so much lower than s_y?

It's the average of the two parents' heights, so it's going to be roughly a factor of $\sqrt{2}$ less than the child's height. (It's actually even less than that though.)

Adding variables that are correlated

Let's try adding the parent height and the child height, which are correlated, just to get some experience adding correlated variables. (On the worksheet, you will add SAT math and SAT verbal scores.)

Data $>$ Compute $>$ Expression and I've called this 'sum'.

Let's see if this formula works!

$$S_{A+B}^2 = S_A^2 + S_B^2 + 2rS_AS_B$$

Remember, of course, if $r = 0$ you get back to the formula you already know how to use!

Testing our new formula

Summary statistics:

Column	Mean	Variance	Std. dev.
midparentHeight	69.206773	3.2485384	1.8023702
childHeight	66.745931	12.811039	3.5792512
sum	135.9527	20.20056	4.4945033

Correlation between childHeight and midparentHeight is:
0.3209499

$$S^2_{A+B} \stackrel{?}{=} S^2_A + S^2_B + 2rS_AS_B$$

$$20.20 \stackrel{?}{=} 3.25 + 12.8 + 2(0.321)(1.80)(3.58)$$

$$20.20 \stackrel{?}{=} 16.05 + 4.14$$

$$20.20 \approx 20.19$$

Close enough! (It's unequal due only to roundoff error.)

Multi-linear models

StatCrunch makes it easy to use multi-linear models. This allows us to use the fathers' and mothers' heights as two different inputs to our model. We can find out whose genes matter more.

Stat > Regression > Multiple Linear

Dad matters more!

Multiple linear regression results:
Dependent Variable: childHeight
Independent Variable(s): father, mother
childHeight = 22.64328 + 0.36828233 father + 0.29050997 mother

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	22.64328	4.2621271	≠ 0	931	5.3126712	<0.0001
father	0.36828233	0.044888233	≠ 0	931	8.2044293	<0.0001
mother	0.29050997	0.048524794	≠ 0	931	5.9868359	<0.0001

Analysis of variance table for multiple regression model:

Source	DF	SS	MS	F-stat	P-value
Model	2	1257.7124	628.8562	54.742013	<0.0001
Error	931	10694.987	11.487634		
Total	933	11952.7			

Summary of fit:
Root MSE: 3.3893412
R-squared: 0.1052
R-squared (adjusted): 0.1033

$$R^2 = 10.5\%$$

So, this is a little bit better than without splitting the parents up. But, we only gain about half a percent on R^2.

Double split!

Let's split the entire problem by gender. We can split that in the model by using the father and mother separately, and we can split on the gender of the child as well.

Multiple linear regression results for gender=male:
Dependent Variable: childHeight
Independent Variable(s): father, mother
childHeight = 19.312813 + 0.41755622 father + 0.32877342 mother

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	19.312813	4.0950285	≠ 0	478	4.7161609	<0.0001
father	0.41755622	0.045612392	≠ 0	478	9.1544468	<0.0001
mother	0.32877342	0.04530087	≠ 0	478	7.257552	<0.0001

Analysis of variance table for multiple regression model:

Source	DF	SS	MS	F-stat	P-value
Model	2	786.32461	393.16231	74.622924	<0.0001
Error	478	2518.4162	5.2686532		
Total	480	3304.7408			

Summary of fit:
Root MSE: 2.2953547
R-squared: 0.2379
R-squared (adjusted): 0.2347

Multiple linear regression results for gender=female:
Dependent Variable: childHeight
Independent Variable(s): father, mother
childHeight = 18.833583 + 0.37254233 father + 0.30348214 mother

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	18.833583	3.6094745	≠ 0	450	5.2178184	<0.0001
father	0.37254233	0.035912028	≠ 0	450	10.373748	<0.0001
mother	0.30348214	0.042080619	≠ 0	450	7.2119219	<0.0001

Analysis of variance table for multiple regression model:

Source	DF	SS	MS	F-stat	P-value
Model	2	672.557	336.2785	82.437552	<0.0001
Error	450	1835.6358	4.0791908		
Total	452	2508.1928			

Summary of fit:
Root MSE: 2.0197007
R-squared: 0.2681
R-squared (adjusted): 0.2649

For males $R^2 = 23.8\%$, and for females $R^2 = 26.8\%$.

Splitting the parents didn't help much

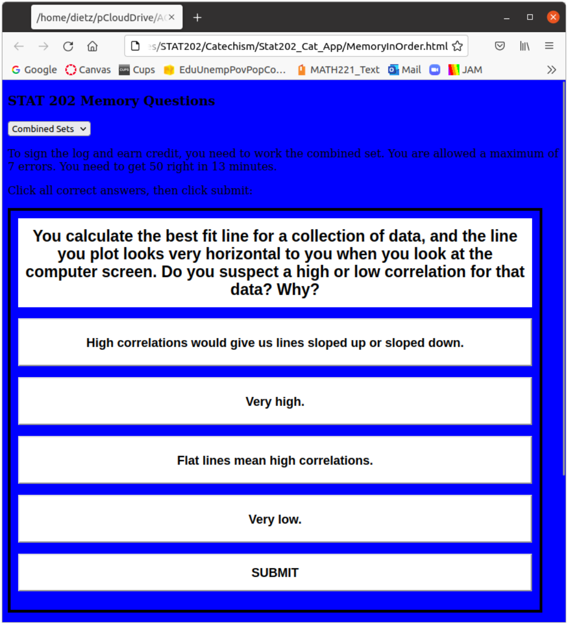
With splitting: for males $R^2 = 23.8\%$, and for females $R^2 = 26.8\%$.
Without splitting: for males $R^2 = 23.3\%$, and for females $R^2 = 26.3\%$.

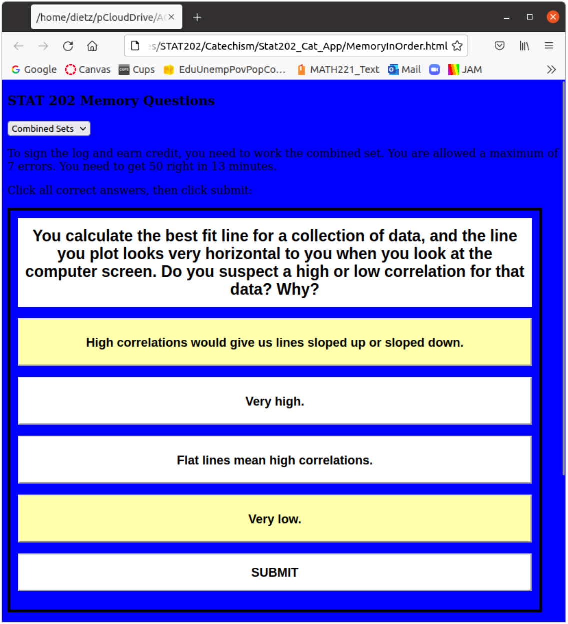
So, we didn't gain much with all that effort. However, it was just clicking an option on the computer, so it really wasn't much effort anyhow.

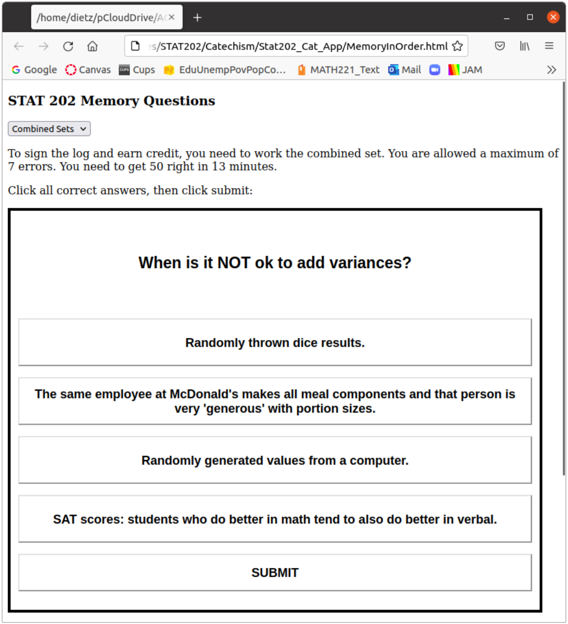
Worksheet time!

Go have fun!

MEMORY QUESTIONs


/home/dietz/pCloudDrive/A
s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
You calculate the best fit line for a collection of data, and the line you plot looks very horizontal to you when you look at the computer screen. Do you suspect a high or low correlation for that data? Why?
High correlations would give us lines sloped up or sloped down.
Very high.
Flat lines mean high correlations.
Very low.
SUBMIT


/home/dietz/pCloudDrive/A
s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
You calculate the best fit line for a collection of data, and the line you plot looks very horizontal to you when you look at the computer screen. Do you suspect a high or low correlation for that data? Why?
High correlations would give us lines sloped up or sloped down.
Very high.
Flat lines mean high correlations.
Very low.
SUBMIT


/home/dietz/pCloudDrive/A
s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
When is it NOT ok to add variances?
Randomly thrown dice results.
The same employee at McDonald's makes all meal components and that person is very 'generous' with portion sizes.
Randomly generated values from a computer.
SAT scores: students who do better in math tend to also do better in verbal.
SUBMIT

STAT 202 Memory Questions

Combined Sets

To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.

Click all correct answers, then click submit:

When is it NOT ok to add variances?

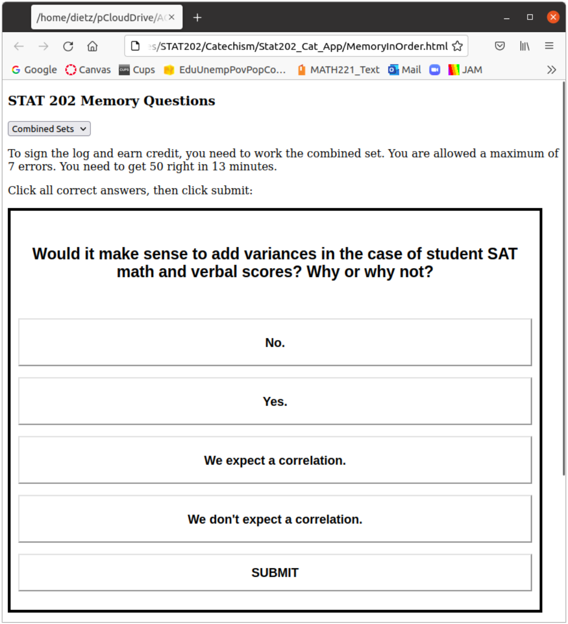
Randomly thrown dice results.

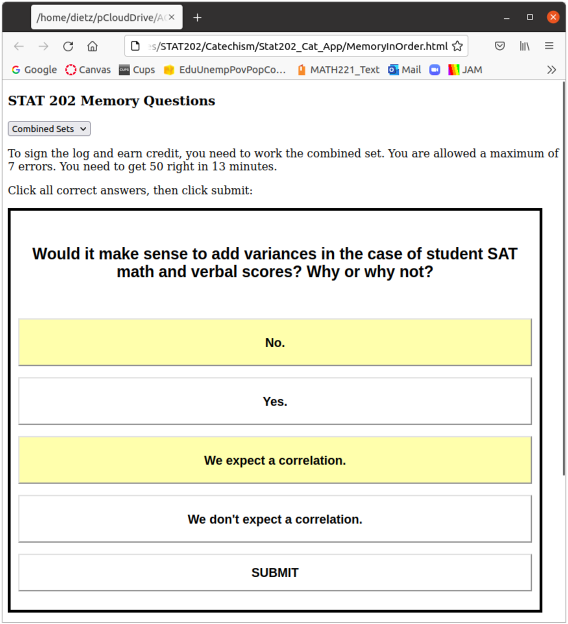
The same employee at McDonald's makes all meal components and that person is very 'generous' with portion sizes.

Randomly generated values from a computer.

SAT scores: students who do better in math tend to also do better in verbal.

SUBMIT


/home/dietz/pCloudDrive/A
s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
Would it make sense to add variances in the case of student SAT math and verbal scores? Why or why not?
No.
Yes.
We expect a correlation.
We don't expect a correlation.
SUBMIT


/home/dietz/pCloudDrive/A
s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
Would it make sense to add variances in the case of student SAT math and verbal scores? Why or why not?
No.
Yes.
We expect a correlation.
We don't expect a correlation.
SUBMIT