Correlation & Regression

Correlation

It is critical that when "interpreting" the association between 2 variables *via* a scatterplot, to employ "weasel words" such as ***in general*** and ***on average*** and ***tends to***.

Why? Because if you dont, youre claiming that there are absolutely NO COUNTEREXAMPLES and that the relationship described in strictly / always true...which is very rarely the case! For instance, this is from a UCLA statistics curriculum: *there is a moderate linear relationship between writing scores and reading scores with a positive association suggesting that those with higher reading scores* **tended to** have *higher writing scores...specifically, as reading scores rose from ~30-75, the writing scores rose from ~30-68.*

Why is the **bold phrase** important? Because if you didnt include it, youd be stating that ALWAYS and WITHOUT EXCEPTION fellows with higher reading scores had higher writing scores! But that's clearly <u>not</u> the case from examining the scatterplot: there are numerous instances of individuals with higher reading scores with THE SAME or LOWER writing scores!

Basic expectations for the CSET:

1. describe bivariate relationships
2. interpret the correlation coefficient
3. know and apply the properties of the correlation coefficient
4. calculate the Line of Best Fit and make predictions

How to describe bivariate relationships?
In terms of ***Strength, Form, Association*** and ***Clusters / Outliers***, if any.

- Strength: indicates whether the relationship is Strong, Moderate or Weak
- Form: indicates whether the relationship is Linear or non-Linear
- Association: indicates whether the direction of the relationship is Positive or Negative or Constant over specific domains of the X-variable

MADLIBS: The relationship between [Y] and [X] is [**STRENGTH** <u>moderately? strongly? weakly?</u>] [**FORM** <u>linear? nonlinear?</u>] with a [**ASSOCIATION** positive? negative?] association, which means that as [X] increases, [Y] [increases? decreases?], ***on average***. ***If applicable***, the value of ([X], [Y]) is an outlier since it falls outside the overall pattern of the distribution. There are clusters [between X values from #-# and another from #-#].

Example. The relationship between Number of Calories Consumed and Time at the Table by toddlers is moderately linear with a negative association: as the amount of time spent by toddlers rose from ~20 to ~50min, ***in general***, the calories consumed fell from ~525 to ~400cal.

Example. The relationship between Income and Age is strongly linear with the association being positive from 15-55years: as individuals grow older from 15 to 55 years, their incomes ***tended to*** rise from $12,000 to $62,000. From 55 to 65years, the relationship was moderately linear with negative association: as individuals grow older from 55 to 65 years, their incomes ***tended to*** decline from $62,000 to $58,000. Beyond 65years, incomes were relatively flat at about $58,000.

Bottomline: youre describing the relationship between Y and X in detail -- you write what you see!

What is the correlation coefficient? How to interpret the correlation coefficient, r?
The correlation coefficient is a measure of how close the points of the scatterplot are to the Line of Best Fit i.e. it measures the strength (and direction) of *linear* relationship between the 2 variables.

There are 2 components to *r*:

- Number:
    - close to ±1 → strong *linear* relationship;
    - close to 0 → weak *linear* relationship
- Sign [positive or negative association].

In general, *both* in the positive and negative direction:

- $0.8 < |r| < 1$ ~ VERY STRONG;
- $0.6 < |r| < 0.8$ ~ MODERATELY STRONG
- $0.4 < |r| < 0.6$ ~ MODERATELY WEAK
- $0 < |r| < 0.4$ ~ VERY WEAK

How to make a scatterplot on the calculator?
Enter X-Variable: L1 and Y-variable: L2
Use **2nd + STATPLOT [Y =** on the top left**] → Select the 1st plot → Xlist: L1 → Ylist: L2**
Hit **ZOOM** [on the top row]→ **ZOOMSTAT** to view the scatterplot.

How to calculate the corr. coeff., r, on the calculator?
After entering the X-values in L1, and Y-values in L2, use
STAT → CALC → Option8: **LinReg L1, L2.**
Note: If your calculator does not display *r*, do this:
2nd 0 → Scroll Down to Diagnostic On → Hit ENTER twice. Then, follow the above instructions.

What are the properties of the Correlation Coefficient, r?
Definition: *r* is a measure of the strength [#] and association [sign] of the *linear* relationship between 2 variables. Alternately, *r* is a measure of how closely the points are clustered to the Line of Best Fit. For it to be reasonable to use *r*, the scatterplot must *appear* linear.

Caution! It makes no sense to calculate *r* for a relationship that is non-linear [a u-shaped pattern]...or for one where there is no clear association between Y and X [a haphazard plot].

1. The formula to calculate r is $r = (\sum Z_x Z_y) / (n - 1)$ i.e. it is the average of the product of the z-scores of the X and Y variables.
IMPORTANT! *r* has no units, being derived from Z-scores; Z-scores are pure unit-less *numbers* [~ number of s.d. a number is from the mean].

2. *r* has the same sign of the slope of the Line of Best Fit.
Positive association: when X increases, so does Y;
Negative association: when X increases, Y decreases

3. $-1 \leq r \leq 1$

4. An Important Property:

r is NOT affected by changes in SCALE [Translation: multiplying or dividing *every* X- or Y-value by a constant does not affect *r*!] and

r is NOT affected by changes in ORIGIN [adding a constant to *each* X- or Y- value does not affect *r*!].

Why? Because

A) *r* only tells you how closely clustered the data is to the Line of Best fit i.e. *r* measures the strength and association of the *linear* relationship.

Changing the origin [adding / subtracting each X- or Y-value by a constant] does NOT affect the relative cluster of the points to the Line of Best Fit since all it does is shift the *all* the points leftward / rightward [if a constant is added / subtracted to all the X-values], or upwards / downwards [if a constant is added / subtracted to all the Y-values]...but it doesnt affect how closely the points are ***relative to the Line of Best Fit***!

Changing the scale [multiplying / dividing X- or Y-value by a constant] also does NOT affect the relative cluster of the points to the Line of Best Fit since all it does is move the points closer or further apart (multiplying / dividing). There's simply more or less "empty space" in the plot: it doesnt change the orientation of the Line of Best Fit!

B) *r* is the average of the product of the Z-scores of the X- and Y-variables...and how are Z-scores calculated? By subtracting a constant [Mean!] from each value and dividing each value by another constant [the s.d.!].

Hence, *other* linear transformations do <u>not</u> affect *r*.

5. Correlation does *not* imply causation.

Translation: simply because the correlation coefficient, *r,* is close to ±1, say, does *not* indicate that X *caused* Y. A [strong] relationship between Y and X does <u>not</u> automatically indicate a <u>causal</u> relationship.

Example. There is a strong correlation between Hair length and Shoe-size [after all, as babies grow, they have longer hair and bigger shoe-sizes]…but it'd be absurd to argue that bigger shoe size ***causes*** hair to grow longer!

Example. There is a strong correlation between SAT scores and Family Income [wealthier families tend to be well-educated *or* well-educated families tend to be wealthier...and those families may invest in education more, leading to higher SAT scores amongst their young]...but it'd be ridiculous to that the higher Income *caused* the higher scores or that to do better on the SATs one would *need* wealthier parents!

6. *r* being close to zero indicates a weak *linear* relationship [for instance, a relatively haphazard scatterplot!] ***or*** may suggest a strong *non*-linear relationship [for instance, a U-shaped scatterplot]; on the other hand, *r* being close to ±1 indicates a strong *linear* relationship.

Case I: haphazard scatter, showing NO relationship at all
Case II: U-shaped / upside-down-U shaped curves with strong *non*-linear relationships; even polynomial curves...however, some curves MAY given a HIGH *r.*

Observe this! A Haphazard scatter indicates no relationship between X and Y and Zero Correlation.

Observe this! But Zero Correlation Does <u>NOT</u> Mean No Relationship...
it MIGHT mean which might be curvilinear.

7. *r* requires the 2 variables to be numeric.

Caution! It does not make sense to calculate *r* for categorical variables: "Income and ***Gender*** are strongly correlated" is an absurd statement from a statistical standpoint!

How is the correlation coefficient, r, *interpreted?*
Example. The correlation coefficient of r = -0.14 indicates that the relationship between Sodium content in Hot Dogs and Calories is weakly ***linear*** [$r \rightarrow 0$], with negative association suggesting that as the calories in the hot dogs rose from 280 to 720cal, the Sodium content fell from 420g to 370g, ***on average***. Yes, interpreting ***r*** is just like describing the linear relationship!

Example. The correlation coefficient of r = +0.68 indicates that the relationship between Income and Education Level is moderately strong and ***linear*** [$r \sim 0.7$] with positive association [r, being +]: as the Education Levels of individuals rose from 2years to 8years, so did their incomes ***in general***, from $23,450/year to $73,861.

What is the Regression Line or the Line of Best Fit?

1. The regression line is: Y^ = *a* + *b*X, with the slope, *b* [the # next to X] and Y-intercept, *a* [the # all by itself] is used to predict Y-values for given X-values.
2. For the Line of Best Fit, i.e. regression, it *matters* which is the Independent variable (X), and which, the Dependent variable (Y). So, it is vital to identify the independent (X) and dependent (Y) variables accurately. Y *depends* on X. Alternately, we use X to *predict* Y. Read the Q *carefully*, and examine the given plots make a judgment.

Observe this!

- It is convention to use a ^ after the Y to suggest the idea that the Line of Best Fit is used to ***predict*** Y-values given an X-value, so Y^ ~ *predicted Y e.g.* **Y^ = 3.1452 – 4.1562X**
- It is convention to write the Line of Best Fit "in context" i.e. <u>without</u> using the Y and X letters but with mnemonic variables e.g. **Height^ = 3.1452 – 4.1562Age** or **HT^ = 3.1452 – 4.1562AGE**

How to calculate the Regression Line ~ Line of Best Fit on the calculator?
After entering the X-values in L1, and Y-values in L2, use
STAT → CALC → Option8: LinReg L1, L2

Problem. Fill in the blanks.
1. *r* is a measure of the _____ between 2 variables in terms of their strength [#] and association [sign]. Alternately, *r* is a measure of how close the scatterplot is clustered to the ___.
2. *r* requires X and Y to be ___ variables.
3. The formula for *r* is ___.
4. The correlation coefficient, *r*, has the same sign as the ___ of the LSRL.
5. Suppose we are estimating the Height (cm) of children based on their Age (months).
a) The Explanatory variable is __, and the response variable is __.
b) The unit of the slope of the LSRL is __.
c) The unit of the slope of the y-intercept is __.
6. *r* is *not* affected by changes in __ or __.
7. Correlation does not imply __.
8. LSRL stands for ___.
9. The Least Squares Regression Line or LSRL describing the linear relationship between Y and X minimizes ___.
10. For the LSRL, Y^ = *a* + *b*X, the formula for the slope is __ and the formula for the Y-intercept is __.
11. The Residual is simply the __ [definition in simple language, not the formula].
12. The LSRL of Y on X minimizes the Sum of the Squares of the Residuals in the [vertical / horizontal] __ direction.
13. The letter [or variable] used to describe the slope is __.
14. The letter [or variable] used to describe the slope is __.
15. The formula for Residual, R = __.

Solution.
1. *r* is a measure of the strength of the linear relationship between 2 variables in terms of their strength [#] and association [sign]. Alternately, *r* is a measure of how close the data are clustered to the **LSRL**.
2. *r* requires X and Y to be **numeric / quantitative** variables.
3. The formula for *r* is **$\sum Z_x Z_y/(n - 1)$**.
4. *r* has the same sign of the **slope** of the LSRL.
5. Suppose we are predicting the Height (cm) of children based on their Age (months).
a) The Explanatory variable is **Age** and the response variable is **Height**.
b) The unit of the slope of the LSRL is **cm/months**.
c) The unit of the y-intercept is **cm**.
6. *r* is *not* affected by changes in **origin** or **scale**.
7. Correlation does not imply **causation**.
8. LSRL stands for **Least Squares Regression Line**.
9. The Least Squares Regression Line or LSRL describing the linear relationship between Y and X minimizes **the sum of the squares of residuals** OR **the sum of the squares of vertical distances** OR **the sum of the squares of differences between Actual and Predicted Y-values.**
10. For the LSRL, Y^ = *a* + *b*X, the formula for the slope, ***b* = ΔY/ ΔX = r·Sy/Sx** *and* Y-intercept, ***a* = YB – b·XB.**
11. The Residual is simply the **prediction error**.
12. The LSRL of Y on X minimizes the Sum of the Squares of the Residuals in the **vertical** direction.

13. The letter [or variable] used to describe the slope is ***b***.
14. The letter [or variable] used to describe the Y-intercept is ***a***.
15. The formula for Residual, R = **Actual Y-value – Predicted Y-value**

Problem. Researchers suspect that there's a relationship between Per Capita Alcohol Consumption and Heart Disease, specifically, that they could employ Per Capita Alcohol Consumption to predict incidence of Heart Disease. It is found that for a data-set of 19 well-developed countries, the average Per Capita Alcohol Consumption from wine was 3.0263litres/year with a s.d. of 2.5097litres/year; and that the average heart disease death rate was 191.0526 (per 100,000) with a s.d. of 68.3963 (per 100,000). Further, the correlation between per capita wine consumption and heart disease was -0.8428.

Country	**Per Capita Alcohol Consumption (litres / year)**	**Heart Disease Death Rate (per 100,000)**
Australia	2.5	211
Austria	3.9	167
Belgium	2.9	131
Canada	2.4	191
Denmark	2.9	220
Finland	0.8	297
France	9.1	71
Iceland	0.8	211
Ireland	0.7	300
Italy	7.9	107
Netherlands	1.8	167
New Zealand	1.9	266
Norway	0.8	227
Spain	6.5	86
Sweden	1.6	207
Switzerland	5.8	115
United Kingdom	1.3	285
United States	1.2	199
West Germany	2.7	172

a) Determine the Explanatory and Response variables. What are the *units* in which each is measured? [Be detailed.]
b) Attach symbols the statistics above. Then use the statistics to calculate the slope and y-intercept of the LSRL between Heart Disease Death Rate and Per Capita Alcohol Consumption.
c) Write the LSRL in context. **Show formulas / work.**

d) Interpret the correlation coefficient between Per Capita Alcohol Consumption and Heart Disease Death Rate in context.
e) What are the units of the slope of the LSRL? That for the y-intercept of the LSRL? **Tip!** The units of slope, *b* are units of [rise / run ~ Y / X] whereas the unit of the y-intercept, *a*, is simply that of the Y-variable.
f) Predict the death rate for Sweden. Is this an overestimate / underestimate?
g) Calculate and interpret the Residual for Sweden.

Solution.

a) The Explanatory or X-variable is the Per Capita Wine Consumption ~ PCWC (litres / year) and Response or Y-variable is the Heart Disease Death Rate ~ HDDR (per 100,000).
b) XB = 3.0263litres/year, Sx = 2.5097litres/year;
YB = 191.0526 (per 100,000); Sy = 68.3963 (per 100,000);
r = -0.8428.
b = *r*·Sy/Sx = -0.8428·68.3963/2.5097 = -22.9686 per 100,00 / (litres/year)
a = YB – *b*·XB = 191.0526 – (-22.9686)·(3.0263) = 260.5633 per 100,000
c) HDDR**^** (per 100,000) = 260.5633 – 22.9687·PCAC (litres per year).
d) The relationship between Heart Disease Death Rate and Per Capita Wine Consumption is *strongly* linear with a negative association: as the Per Capita Wine Consumption of counties increased [from 0.7 to 9.1 litres per year], **in general**, the Heart Disease Death Rate fell [from 300 to 71 deaths per 100,000]. **WATCH THE UNITS!**
e) Units of Slope **(Y / X)**: per 100,00 / (litres/year)
Unit of y-intercept (**Y)**: per 100,000
f) For Sweden, given: PCAC = 1.6 litres / year
Therefore, HDDR^ = 260.5633 – 22.9687·1.6 = 224.2 deaths / 100,000 **WATCH THE UNITS!**
Since the actual HDDR is higher than the predicted HDDR, it is an over-estimate.
g) Residual Death Rate = Actual Death Rate – Predicted Death Rate [Observe context]
= 207 **from the table!** – 224.2 = -17.2 deaths / 100,000 **WATCH THE UNITS!**
Interpretation: A residual of -17.2 deaths / 100,000 indicates the our LSRL model overestimates the Actual Death Rate of 204 deaths / 100,000 for Sweden by 17.2 deaths / 100,000.
Alternately: A residual of -17.2 deaths / 100,000 is a measure of the prediction error – an overestimate – when predicting the Death Rate for Sweden using the LSRL.

Problem. Obesity is a growing problem around the world. A study sought to shed some light on gaining weight. Some people don't gain weight even when they overeat. Perhaps, fidgeting and other non-exercise activity (NEA) explains why – some people may spontaneously increase non-exercise activity when fed more. Researchers deliberately overfed 16 healthy young adults for 8 weeks. They wished to determine if the change in energy use (in calories) from Non-Exercise Activity (NEA) i.e. activity other than deliberate exercise – fidgeting, daily living, etc. could predict the fat gain (in kgs).
a) Identify the explanatory and response variables, and the units they are measured in.
b) The following summary statistics were obtained:
Mean NEA change: 324.8cal, s.d. of NEA change = 257.66cal;
Mean fat gain = 2.388Kg; s.d. of fat gain = 1.1389kg;
the correlation between fat gain and NEA change was -0.7786.
Calculate the slope of the LSRL and the y-intercept of the LSRL. State their units.
c) Write the equation of the LSRL, in context. Mention units in ( ) beneath the Y^ and X variables.

d) The actual data-set is:

NEA change (cal)	Fat Gain (Kg)
-94	4.2
-57	3
-29	3.7
135	2.7
143	3.2
151	3.6
245	2.4
355	1.3
392	3.8
473	1.7
486	1.6
535	2.2
571	1
580	0.4
620	2.3
690	1.1

Interpret the correlation coefficient r = -0.7786, in context, to describe the relationship between fat gain and NEA change, describing the association between them.
e) Calculate and interpret the residual for (355, 1.3). **Show formulas / work.**

Solution.

a) Explanatory Variable: NEA change (cal); Response Variable: fat gain (kgs)
b) The following summary statistics were obtained:
Mean NEA change: 324.8cal = **XB,** s.d. of NEA change = 257.66cal = **Sx**
Mean fat gain = 2.388Kg = **YB,** s.d. of fat gain = 1.1389kg = **Sy**
correlation between fat gain and NEA change was -0.7786 = ***r***
b = r·Sy/Sx = -0.7786·1.1389/257.66 = -0.00344 Kg / Cal
a = YB – b·XB = 2.388 - (-0.00344)·(324.8) = 3.505Kg
c) FG^ (Kg) = 3.505 – 0.00344·NEAC (Cal)
d) The correlation coefficient of -0.7786 indicates that the relationship between Fat Gain and NEA change is strongly linear with a negative association, indicating that as the NEA change increased for the subjects [from -94 to 690cal], the Fat Gain they experienced fell, in general [from 4.2 to 0.4Kgs].
e) For NEAC = 355, FG^ (Kg) = 3.505 – 0.00344·355 = 2.2838 Kg
Residual FG = Actual FG – Predicted FG
= 1.3 **Given!** – 2.2838 = -0.9838 Kg **WATCH THE UNITS!**
A residual of -0.9838 indicates that the LSRL **overestimates** the actual FG for an NEAC of 355Cal by 0.9838 Kgs **OR** the residual of -0.9838 is a measure of the prediction error when using the LSRL for predicting the FG for an NEAC of 355Cal.

NOTES on Regression

How are the slope and y-intercept interpreted?
The LSRL is Y^ = a + bX with the slope being b and the y-intercept, a.

Concept of slope, b: The slope tells us how fast is Y changing when X changes by a certain amount.
Mathematically, b = rise / run = $\Delta Y / \Delta X$ where Δ stands for change.

UNDERSTAND THIS WELL What $b = \Delta Y / \Delta X$ implies is that ***if*** $\Delta X = 1$, ***then*** clearly, $b = \Delta Y$ or what is the same thing, $\Delta Y = b$. That is the essence of the **interpretation of the slope**: a slope of b (Y / X units) indicates that when X changes by 1 unit, then Y is estimated to change by b units. **MASTER THIS.**

Concept of Y-intercept, a: The y-intercept tells us what the Y-value is when X = 0.
UNDERSTAND THIS WELL → in the LSRL, if you substitute X = 0, then Y = a → the y-intercept is the point (X = 0, Y = a). That is the essence of the **interpretation of the Y-intercept**: A Y-intercept of a (Y units) ~ (X = 0 units, Y = a units) indicates that when X = 0 units, Y is predicted to be a units. ← **MASTER THIS.**

Example. If
SalesVolume**^** ($mn) = -410.2365 + 26.8934Time (years since 1995), then

- a slope of **setting up the interpretation** b = 263.8936 **$mn / year** = **ΔSalesVolume^** / **Δtime = 1** indicates that every successive **year since 1995**, the Sales Volume has been estimated to rise by about $26.8934**mn**.
- A y-intercept of -$410.2365**mn** ~ **setting up the interpretation** (Time = 0**years since 1995**, SalesVolume^ = **-$**410.2365**mn**) indicates that in 1995 (can you see why?!), the Sales Volume was predicted to be -**$**410.2365**mn**...which, however does not make sense in the context of the problem.

Example. If
WaterUsage**^** ('000 gallons) = 410.2365 – 26.8934RelativeHumidity(%), then

- a slope of **setting up the interpretation** -26.8936 **'000 gallons / %** = **ΔWaterUsage^** / **ΔRelativeHumidity = 1** indicates that when the Relative Humidity increases by 1**%**, the Water Usage is estimated to rise by 410.2365 **'000 gallons**.
- A y-intercept of 410.2365 **'000 gallons** ~ **setting up the interpretation** (RelativeHumidity = **0%**, WaterUsage ^ = 410.2365 **'000 gallons**) indicates that when the Relative Humidity is 0**%**, then the Water Usage is predicted to be 410.2365 **'000 gallons**. **Note:** this does makes sense in the context of the problem but you dont have to state *that* when it does make sense, haha.

Caution!

- The slope deals with the idea of change: you must employ terms such as *increase, decrease, rise, fall* etc. when interpreting the slope.
- Only slope deals with the idea of change, not the Y-intercept. So dont use the terms *increase, decrease, rise, fall* for the Y-intercept.
- Both, slope and y-intercept, relate to the LSRL -> use the terms *estimated* or *predicted*.
- You **need** to mention the UNITS of the slope and y-intercept *throughout*.

Psst!

- Read the Q and identify the X-variable and the Y-variable 1st. Remember: Y depends on X; also, we predict Y when X is given!

- If the problem refers to "predict" or "estimate", *that* refers to predicted Y-value -> use the LSRL!

Example. The LSRL relating Value of a car, V(t) with time, *t*, measured in terms of Years since 2002 is V(t)^ = $34,500 – 2,300t.
Set up and interpret the slope of the LSRL in context, using UNITS.
Set up and interpret the y-intercept of the LSRL in context, using UNITS.
[Power Tip! It enormously helps to "set up" the interpretation as we did in class 1st!]

Solution.
Interpretation Guide for Slope! *b* = $2300 / year = ΔValue of Car ($)/ ΔTime (Years since 2002) **Mentioning UNITS is vital!**

A slope of $2300 / year indicates that *every year* since 2002, the Value of the car is estimated to *fall* by $2300, *on average*. **Mentioning UNITS is vital!**
OR
A slope of $2300 / year indicates that as the number of years increases by 1 beyond 2002 [how awkward!], the Value of the car is estimated to *fall* by *about* $2300. **Mentioning UNITS is vital!**

Interpretation Guide for Y-intercept! (Time = 0 years since 2002, Value of Car = $34,500) **Mentioning UNITS is vital!**

A y-intercept of $34,500 indicates that in 2002, the Value of the car was estimated to be *about* $34,500. **Mentioning UNITS is vital!**

Example: The LSRL relating the Number of women in the labour force in millions, W, and Years since 1998, t is W^ = 0.9286t + 63.7
Set up and interpret the slope of the LSRL in context. **Mentioning UNITS is vital!**
Set up and interpret the y-intercept of the LSRL in context. **Mentioning UNITS is vital!**

Solution.
Interpretation Guide for slope! *b* = 0.9286mn/year = ΔNumber of women in the labour force (mn) / ΔTime (years since 1998) **= 1**
A slope of 0.9286mn/year indicates that for every (successive) year since '98, the **estimated** number of women in the labour force in the U.S. rose by ***about*** 0.9286mn. **Mentioning UNITS is vital!**

Interpretation Guide for Y-intercept! (0 years since 1998, 63.7mn women in the US labour force)
A y-intercept of 63.7mn indicates that in 1998 there were an **estimated** 63.7mn women in the labour force in the U.S. ***on average...*****[which makes sense, by the way**!]. **Mentioning UNITS is vital!**

Example. The distribution of Heart Disease Death Rates and Per Capita Alcohol Consumption, the LSRL was HDDR**^** (per 100,000) = 260.5633 – 22.9687·PCAC (litres per year). Set up and interpret the slope of the LSRL in context. **Mentioning UNITS is vital!**

Set up and interpret the y-intercept of the LSRL in context. **Mentioning UNITS is vital!**

Solution.

Mentioning UNITS is vital! Interpretation Guide for slope! m = –22.9687 per 100,000 / lires per year = ΔHDDR (per 100,000) /Δ PCWC (in litres / year) = 1

A slope of –22.9687 **per 100,000 / litres per year** indicates that when the Per Capita Alcohol Consumption of countries rose by 1 **litre / year**, the *estimated* Heart Disease Death Rate fell by *about* 22.9687 deaths / 100,000 [**OR**...the Heart Disease Death Rate was *estimated* to fall by *about* 22.9687 deaths / 100,000] **OR**

A slope of –22.9687 **per 100,000 / litres per year** indicates that countries that had a Per Capita Alcohol Consumption of 1 litre / year more than another, had *about* 22.9687 deaths / 100,000 lower due to Heart Disease.

Interpretation Guide for Y-intercept! (0 litres/year, 260.5633 deaths / 100,000)
A y-intercept of 260.5633 **deaths per 100,000** indicates that when the Per Capita Wine Consumption of countries was close to 0 **litres / year**, the *estimated* Heart Disease Death Rate was 260.5633 **deaths per 100,000** [**or** the Heart Disease Death Rate was *estimated* to be 260.5633 deaths per 100,000]...[**which makes sense, by the way**!].

Example. A certain teacher wishes to predict GPA of her students based on their IQ and finds that their Mean IQ was 108.9, the s.d. of IQ was 13.17, the Mean GPA was 7.447, with the s.d. of GPAs being 2.10. If the correlation coefficient between GPA and IQ is r = 0.6337, determine the equation of the LSRL and write it in context. **Show relevant formulas and calculations.**

Solution. Since we are predicting GPA based on IQ, Y ~ GPA and X ~ IQ so that:
Given: XB = 108.9, Sx = 13.17, YB = 7.447, Sy = = 2.10 with r = 0.6337
Slope, $b = r\cdot Sy/Sx = 0.6337\cdot 2.1/13.17 = 0.1010$
and Y-intercept: $a = YB - b\cdot XB = 7.447 - (0.1010)\cdot(108.9) = -3.557$
and LSRL: GPA^ = -3.557 + 0.101·IQ

Example. An electricity utility would like to examine the relationship between daily temperature and electricity consumption and records the following data:

Average Daily Temperature (F)	KiloWatts (KW) of Electricity Consumed (in Million)
77	10
84	12.1
85	13.1
90	14.2
92	15.6

91	14.1
81	9.7
88	10.7
79	8.1
86	11.5
78	8.4
93	9.9
105	16.3
95	12.7

a) Identify the Explanatory and Response variables [be detailed], and state their units.
b) Use the calculator and calculate the Correlation Coefficient, *r*, between Electricity Consumption and Average Daily Temperature and interpret it in context.
c) Write the LSRL *in context*. Mention units for the Y and X variables in ( ) next to them.
d) Calculate the Residual for a temperature of 81F. Show all work. Interpret the Residual in #14. Give adequate context.

Solution.

a) E: Average Daily Temperature (F);
R: Electricity Consumption: KiloWatts (KW) (in Million)
b) *r* = 0.7712 suggests that the relationship between Electricity Consumption and Average Daily Temperature is strongly linear with a positive association indicating that Average Daily Temperature rose from 77 to 105 F, the Electricity Consumption rose from 8.14 to 16.3 KiloWatts (KW) (in Million). **Mentioning UNITS is vital!**
c) EC^ (kW mn) = -10.6924 + 0.2582·ADT (F)
d) The LSRL for predicting EC for a given ADT is: EC (kW mn) = -10.6924 + 0.2582*ADT (F)
Given: ADT = 81F → *Predicted* EC = EC (kW mn) = -10.6924 + 0.2582*81 = 10.22**mn kW**
Residual EC = Actual EC – Predicted EC
= 9.7 – 10.22 = -0.52 **mn kW**
A residual EC of -0.52mn kW indicates that the LSRL *overestimates* the EC for an ADT of 81F by 0.52mn kW **OR** A residual EC of -0.52mn kW is a measure of the prediction error when using the LSRL for an ADT of 81F.

Problem. We want to employ Amount of Vegetables and Fruits Consumed (in grams) to estimate the Time to Lose 5lbs (in Months). Suppose R^2 = 58.96% with the LSRL: **TL5LBS^ = 45.5964 – 2.2353AVFC**
1. Identify the Explanatory and Response variables and their units.
2. Calculate *r*. Interpret *r* in context.
3. Interpret the slope in context.
4. Interpret the y-intercept in context.

Solution.

1. Explanatory variable (X): Amount of Vegetables and Fruits Consumed (g)
Response Variable (Y): Time Taken to Lose 5lbs (months)
2. $r = \pm\sqrt{0.5896} = \mathbf{-}0.7679$ [since (think about it) there's a *negative* relationship between Time to Lose 5lbs and the Amount of Vegetables and Fruits Consumed: after all, as the Amount of Vegetables and Fruits Consumed, X, increases, the Time to Lose 5lbs, Y, decreases. Also, the slope is negative, so...]
Interpretation: r = -0.7679 suggests that there is a reasonably strong linear relationship between the Time to Lose 5lbs and the Amount of Vegetables and Fruits Consumed [**Mentioning detailed CONTEXT is vital!**], indicating that as the Amount of Vegetables and Fruits Consumed, X, increases, the Time to Lose 5lbs, Y, decreases. ← **some of you are forgetting to explain the association.**
3. **Interpretation Guide for slope** b = -2.2353 months / grams = ΔTL5LBS^/ΔAVFC = 1g
Mentioning CONTEXT and UNITS is vital! A slope of -2.2353 **months / grams** indicates that when the Amount of Vegetables and Fruits that an individual Consumes rises by **1gram**, the Time to Lose 5lbs is estimated to fall by 2.2353**months OR AP 5 students!** A slope of -2.2353 **months / grams** indicates that every additional gram of vegetable and fruit consumed is associated with 2.2353 *fewer* **months,** *on average,* for a person to lose 5lbs.
4. **Interpretation Guide for Y-intercept!** (AVFC = 0grams, TL5LBS^ = 45.5964months)
Mentioning CONTEXT and UNITS is vital! A y-intercept of 45.5964months indicates that when the Amount of Vegetables and Fruits Consumed is 0g [or when an individual consumes no vegetables or fruits], then the Time to Lose 5lbs is estimated to take about 45.5964months...[**which makes sense, by the way**!].

Example. The relationship between the amount of Nicotine and Tar in cigarettes is given by the regression line: Nicotine^ (mg) = 0.154030 + 0.065052Tar (mg).
a) Predict the Nicotine content for a cigarette with 4mg of Tar.
b) Interpret the slope of the regression line.
c) Interpret the y-intercept of the regression line.

Solution.

a) For Tar = 4mg:
Nicotine^ (mg) = 0.154030 + 0.065052·4 (mg) **Showing Substitution is vital!**
= 0.414mg
b) **Set-up:** b = 0.065052 mg / mg = ΔNicotine^ mg/ ΔTar = 1mg
Mentioning CONTEXT and UNITS is vital! A slope of b = 0.065052 mg / mg [**Note:** you may omit the units because they cancel or leave them alone...] indicates that as the Tar content in cigarettes increased by 1mg, the Nicotene content was estimated to rise by about 0.065052mg **OR AP 5 students!** in general, cigarettes that had 1mg of tar more, were estimated to have about 0.065052mg more of nicotine.
c) **Set-up:** (Tar = 0mg, Nicotine = 0.154030mg)
Mentioning CONTEXT and UNITS is vital! A y-intercept of 0.154030mg indicates that for cigarettes with 0mg of Tar [or **no tar** ~ 0mg] content, the *predicted* Nicotene content was about 0.154030mg.

Example. The relationship between Mortality Rate and Calclium Content in the water supply for a group of cities is given by the regression line, MortalityRate^ (deaths / 100,000) = 1676 – 3.23CalciumContent (ppm) with the correlation coefficient, *r* = 0.6557.
a) Interpret the correlation coefficient in context.
b) Interpret the slope of the regression line. Then , interpret the y-intercept.

Solution.

a) The relationship between Mortality Rate (per 100,000) and Calcium (ppm) is moderately linear with a negative association: as the calcium content of the water in the towns rose, the mortality rate declined, in general.
b) **Set-up:** *b* = 3.23 deaths per 100000 / ppm = ΔMortality Rate^/ ΔCalcium Content = 1ppm
Mentioning CONTEXT and UNITS is vital! A slope of 3.23 ***deaths per 100000 / ppm*** indicates that as the Calcium Content of the water **in the towns** rose by **1*ppm***, the Mortality Rate was ***estimated*** to fall by **about** 3.23 ***deaths / 100,000***.
Set-up: [(Calcium Content = 0ppm, Mortality Rate = 1676 deaths / 100000)]
Mentioning CONTEXT and UNITS is vital! A Y-Intercept of 1676 deaths per 100,000 indicates that for a town with NO CC in the water, the *predicted* MR is about 1676 deaths/100,000.