**Q1 What are the assumptions of linear regression regarding residuals?**

Ans: Assumptions about the residuals:

1. **Normality assumption**: The error terms are normally distributed.
2. **Zero mean assumption**: The residuals have a mean value of zero, which means the error terms are normally distributed around zero.
3. **Constant variance assumption**: Residual terms have the same (but unknown) variance, σ2.
   This assumption is also known as the assumption of homogeneity or homoscedasticity.
4. **Independent error assumption**: Residual terms are independent of each other.

**Q 2 What is the coefficient of correlation and the coefficient of determination?**

Ans

**Coefficient of Correlation:** The quantity r, called the linear correlation coefficient,
measures the strength and the direction of a linear relationship between two variables.

1. The mathematical formula for computing r is:
   **For population**

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

where $n$ is the number of pairs of data.

**For Sample**

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

- rxy – the correlation coefficient of the linear relation between the variables x and y
- xi – the values of the x-variable in a sample
- $\bar{x}$ – the mean of the values of the variable x
- yi – the values of the y-variable in a sample
- $\bar{y}$ – the mean of the values of the variable y

2. The value of r is such that -1 < r < +1. The + and – signs are used for positive linear correlations and negative linear correlations, respectively.

3. **Positive correlation:** If x and y have a strong positive linear correlation, r is close to +1. An r value of exactly +1 indicates a perfect positive fit. Positive values indicate a relationship between x and y variables such that as values for x increases, values for y also increase.

4. **Negative correlation:** If x and y have a strong negative linear correlation, r is close to -1. An r value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between x and y such that as

values for x increase, values
for y decrease.

5. **No correlation:** If there is no linear correlation or a weak linear correlation, r is
close to 0. A value near zero means that there is a random, nonlinear relationship
between the two variables.

6. Note that r is a dimensionless quantity; that is, it does not depend on the units
employed.

7. A perfect correlation of ± 1 occurs only when the data points all lie exactly on a
straight line. If r = +1, the slope of this line is positive. If r = -1, the slope of this
line is negative.

8. A correlation greater than 0.8 is generally described as strong, whereas a correlation
less than 0.5 is generally described as weak.

**Coefficient of determination** (r square) is the square of coefficient of correlation which means

Multiply R times R to get the R square value. In other words, Coefficient of Determination is the
square of Coefficient of Correlation.

Let's say, A data set has n values marked y1,...,yn (collectively known as yi , each associated with a fitted
(or modeled, or predicted) value f1,...,fn

Define the residuals as ei = yi – fi, If $\bar{Y}$ is the mean of the observed data:

then the variability of the data set can be measured using **sums of squares** formulas:

- The **total sum of squares** (proportional to the variance of the data):

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2,$$

- The **sum of squares of residuals**, also called the residual sum of squares:

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

- The most general definition of the coefficient of determination is

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

1. The coefficient of determination, r 2, is useful because it gives the proportion of
   the variance (fluctuation) of one variable that is predictable from the other variable.
   It is a measure that allows us to determine how certain one can be in making
   predictions from a certain model/graph.
2. The coefficient of determination is the ratio of the explained variation to the total
   variation.
3. The coefficient of determination is such that 0 < r 2 < 1, and denotes the strength
   of the linear association between x and y.

4. The coefficient of determination represents the percent of the data that is the closest
   to the line of best fit. For example, if r = 0.922, then r 2 = 0.850, which means that
   85% of the total variation in y can be explained by the linear relationship between x
   and y (as described by the regression equation). The other 15% of the total variation
   in y remains unexplained.
5. The coefficient of determination is a measure of how well the regression line
   represents the data. If the regression line passes exactly through every point on the
   scatter plot, it would be able to explain all of the variation. The further the line is
   away from the points, the less it is able to explain.

**Q 3 Explain the Anscombe's quartet in detail.**

Ans :

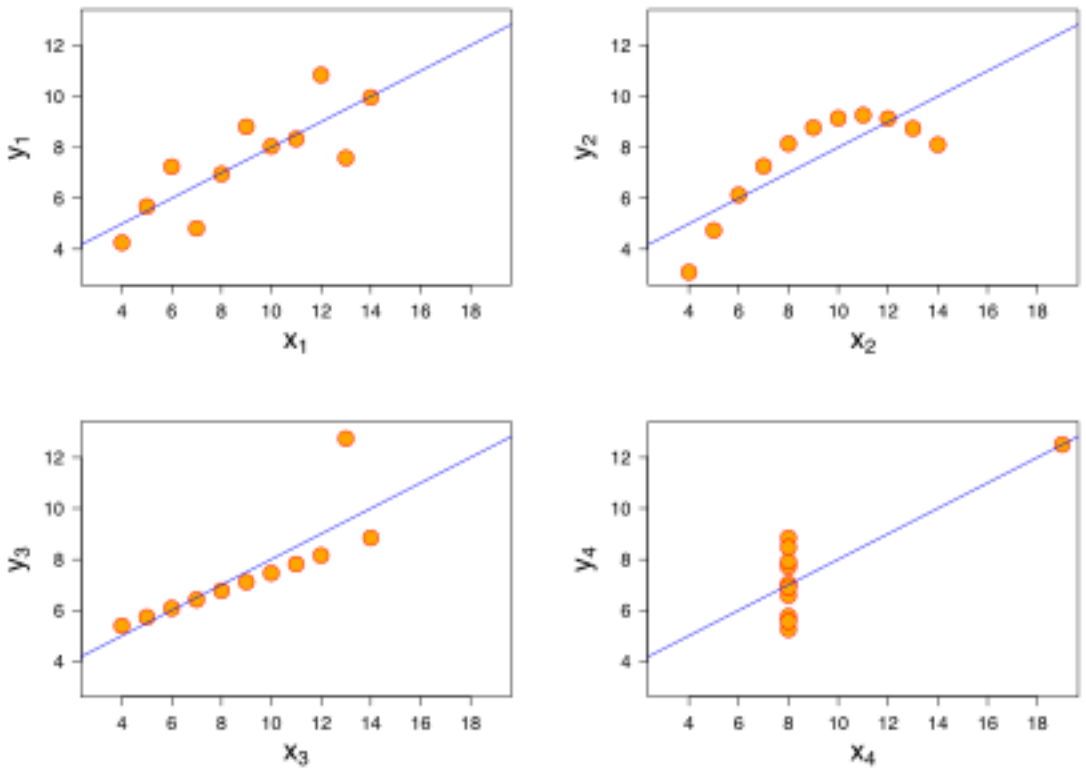
Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics,

yet have very different distributions and appear very different when plotted the graph.

Each dataset consists of eleven (x,y) points. These were formulated in 1973 by the statistician Francis Anscombe to

demonstrate both the importance of graphing data before analyzing it and

the effect of outliers and other influential observations on statistical properties.




As we can see, all the four linear regression are exactly the same. But there are some difference in the data sets that have fooled the regression line.

While the first one seems to be doing a decent job, the second one clearly shows that linear regression can only model linear relationships and

is incapable of handling any other kind of data.

The third and fourth images shows the linear regression model's sensitivity to outliers. If the outlier were not been present,

There could be a great line fitted through the data points. So, we should never run a regression without having a good look at our data.

Below are the Shortcomings of the Linear regression model which are highlighted by these quartets.

- It is sensitive to outliers.

- It models linear relationships only.
- A few assumptions are required to make the inference.

**Q4 What is Pearson's R?**

Ans

The Pearson product-moment correlation coefficient is a measure of the **strength of the linear relationship** between two variables.

It is referred to as Pearson's correlation or simply as the correlation coefficient. If the relation between the variables is not linear,

then the correlation coefficient does not represent the strength of the relation between the variables.

The symbol for Pearson's correlation is **"ρ" for population and "r" for a sample.**

An r of -1 shows a perfect negative linear relationship between variables, an r of 0 indicates

no linear relationship between variables, and an r of 1 shows a perfect positive linear relationship between variables.

Figure 1 shows a scatter plot for which r = 1.

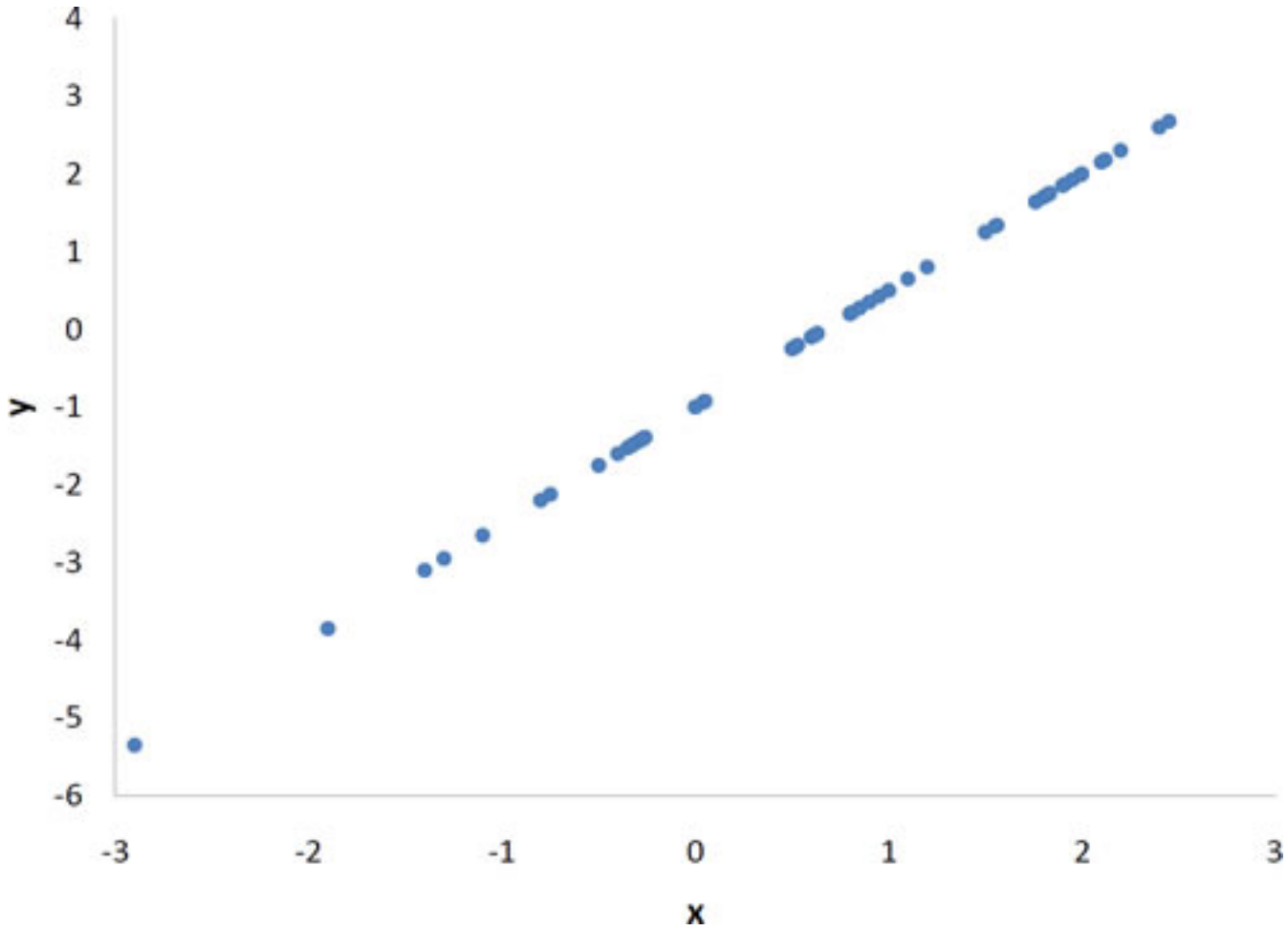


Figure 1. A perfect positive linear relationship, r = 1.

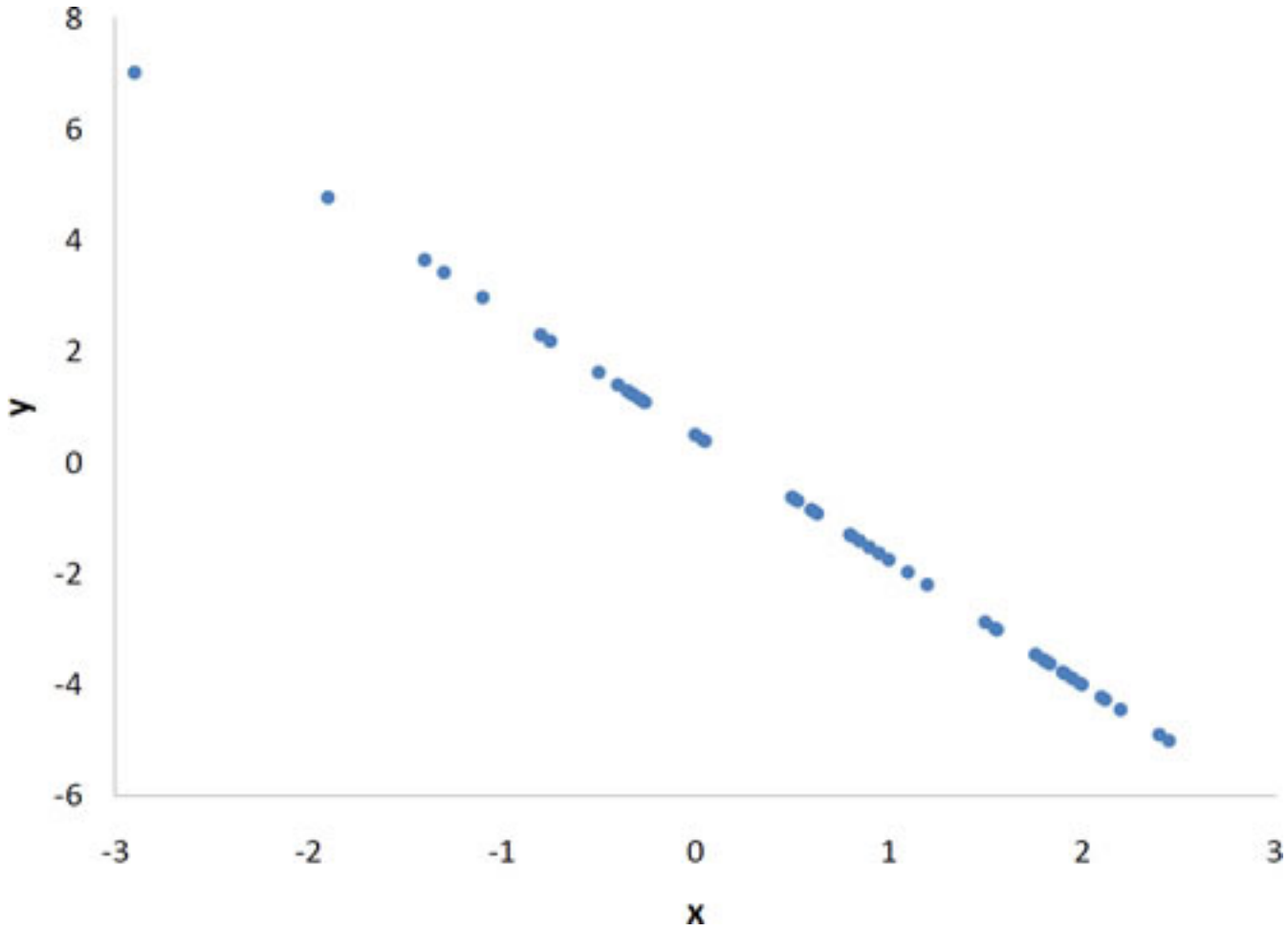



Figure 2. A perfect negative linear relationship, r = -1.

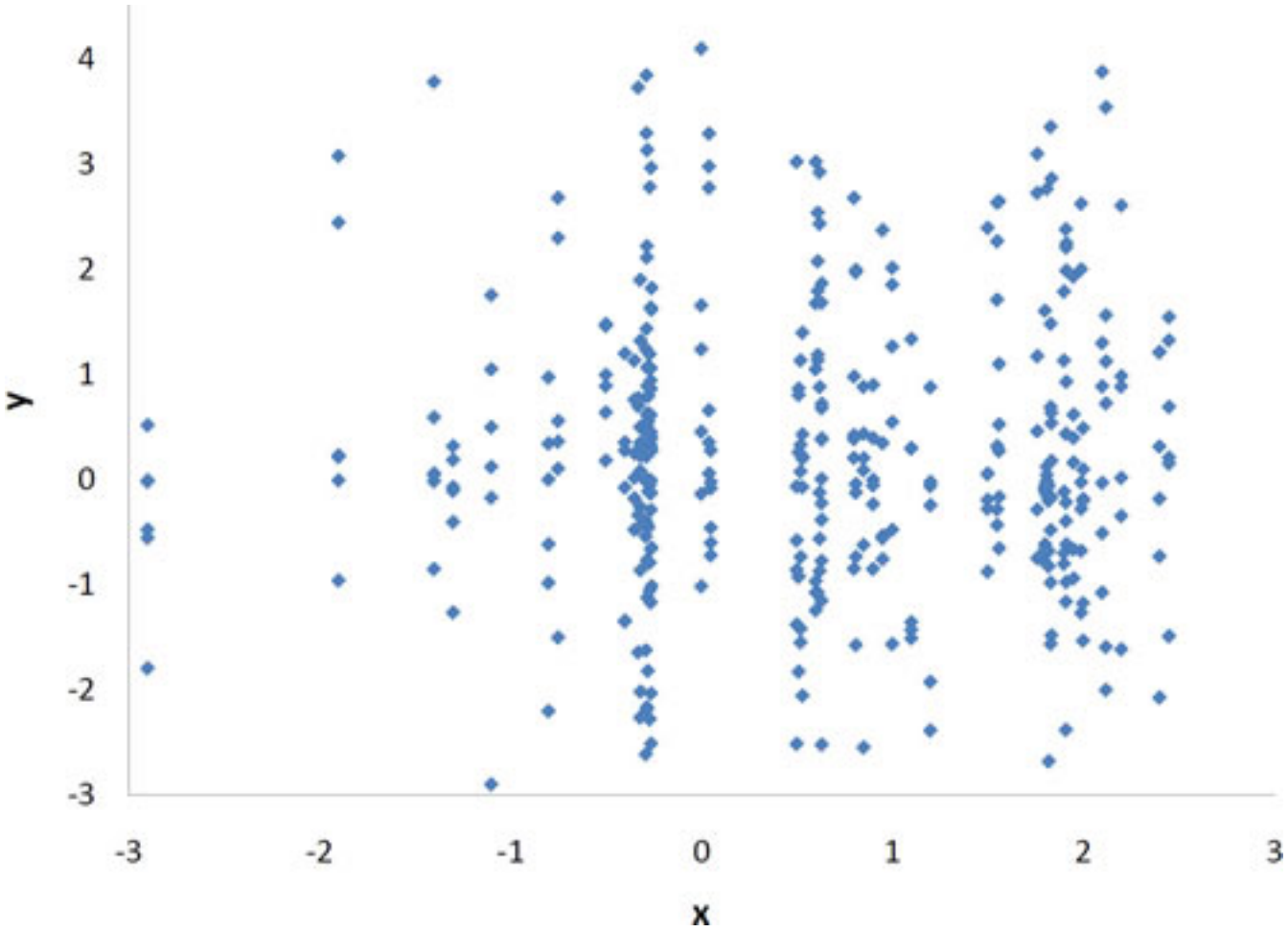



Figure 3. A scatter plot for which r = 0. Notice that there is no relationship between X and Y.

With real data, you would not expect to get values of r of exactly -1, 0, or 1.
The data for spousal ages shown in Figure 4 has an r of 0.97.

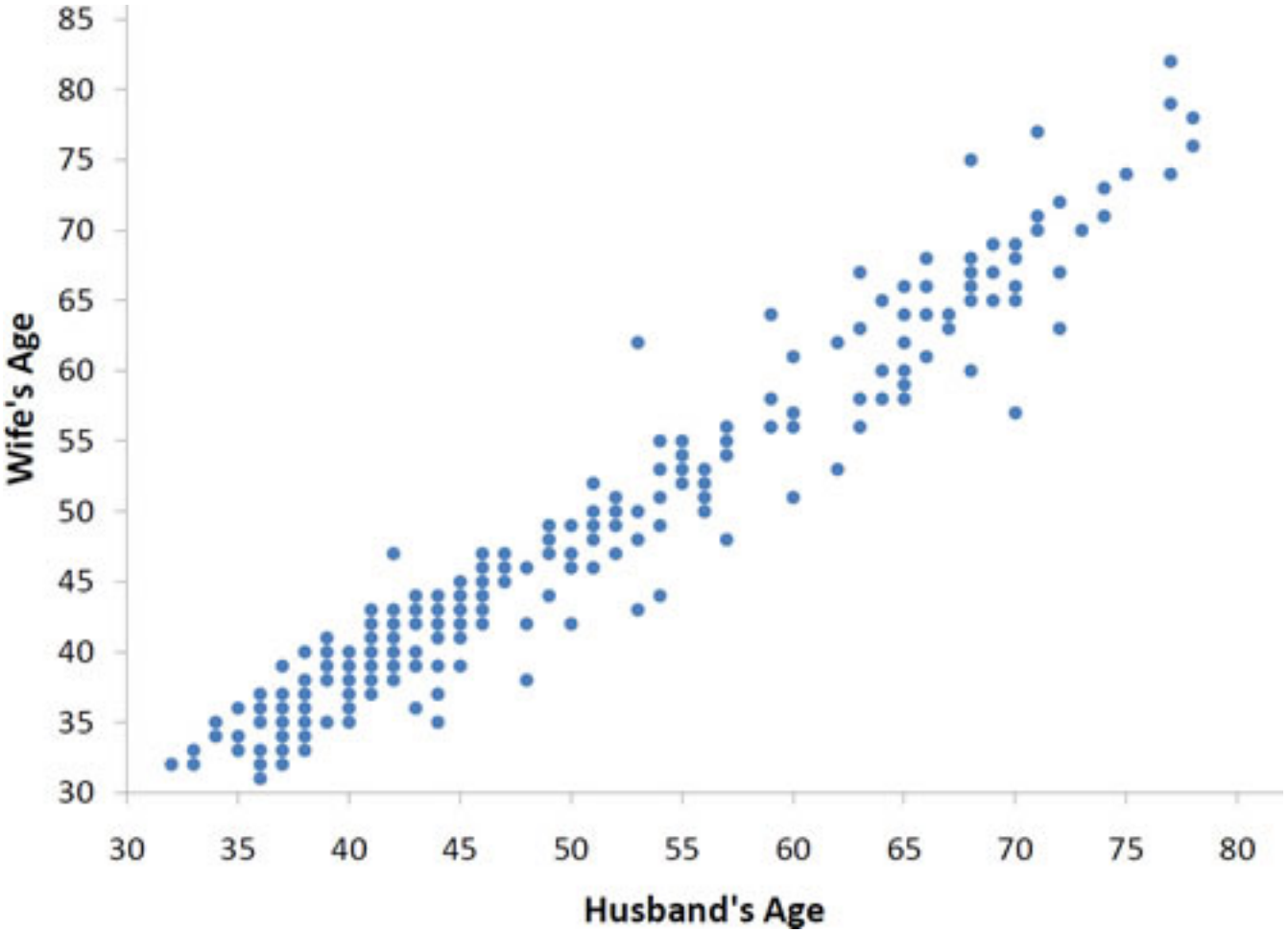



Figure 4. Scatter plot of spousal ages, r = 0.97.

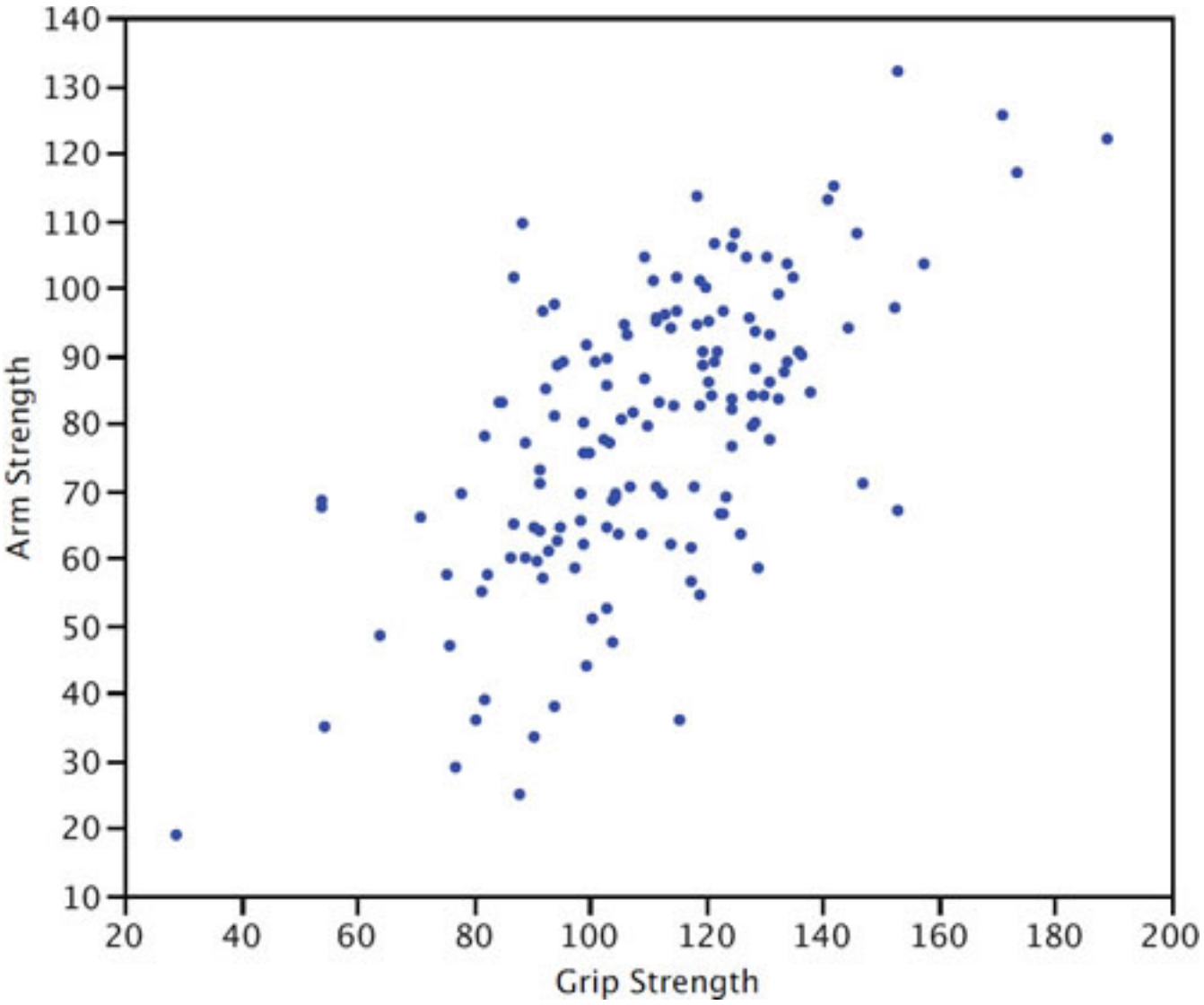



Figure 5. Scatter plot of Grip Strength and Arm Strength, r = 0.63.

The relation between grip strength and arm strength shown in Figure 5 is 0.63.

**Q 5 What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Ans

**1.Scaling**: - Scaling is a method used to standardize the range of independent variables.

The goal of scaling is to change the values of numeric columns in

the dataset to a common scale, without distorting differences in the ranges of values.

2. **Normalization**:

- Assume that the given population (list of values) is in the range of [x_min, x_max]
- $$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$
- Here we subtract min value of the population from all values so that the new range of population will be [0, (x_max - x_min)]
- Now we will divide each element by (x_max - x_min) to convert the range of population into [0, 1]
- Summary: Normalization actually used where we want to scale down the population to [0, 1]

**3. Standardization:**

- Assume the same population again with range [x_min, x_max]
- First calculate standard deviation (x_bar means average of population)

$$\sigma = \sqrt{\frac{\sum (x - \overline{x})^2}{n}}$$

$\sigma$ = **lower case sigma**
$\sum$ = **capital sigma**
$\overline{x}$ = **x bar**

- Now standardize the population with following equation (μ = average of population that we computed above):

- $$x_{new} = \frac{x - \mu}{\sigma}$$
- **Summary**: With standardization we can transform the data into a range such that the new population has mean (average) = 0 and standard deviation = 1.

**Q 6 You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
**Ans**

A variance inflation factor(VIF)
detects multicollinearity in regression analysis.
Multicollinearity is when there's correlation between predictors (i.e. independent variables) in a model;
it's presence can adversely affect your regression results.

VIFs are calculated by taking a predictor and regressing it against every other predictor in the model.
This gives you the R-squared values, which can then be put into the VIF formula. "i" is the predictor you're looking at (e.g. x1 or x2):

$$\text{VIF} = \frac{1}{1 - R_i^2}$$

A rule of thumb for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

Variance inflation factors show the degree to which a regression coefficient will be affected because of the variable's redundancy with other independent variables. As the r square of any predictor
variable with the other predictors approaches unity, the corresponding VIF becomes infinite.

**VIF = 1/(1-1)= 1/0 = infinity that is the estimate is as imprecise as it can be.**

An infinite VIF will be returned for two variables that are exactly collinear. Variables that are exactly the same or linear transformations of each other.

**Some methods that can be used to deal with multicollinearity are as follows:**

- **Dropping variables**

  Drop the variable that is highly correlated with others. Pick the business interpretable variable.
- **Creating a new variable** using the interactions of the older variables

  Add interaction features, i.e., features derived using some of the original features.
- **Variable transformations**