Packet 13: Correlation & Regression

After completing this material, you should be able to:

- identify the response and predictor variables
- interpret a scatterplot in context
- obtain the correlation from StatCrunch and interpret the strength of the linear relationship
- use StatCrunch output to obtain the line of best fit and interpret the coefficients (where appropriate)
- use a hypothesis test to determine if there is a significant linear relationship between two variables
- use the regression equation for prediction
- interpret the appropriate interval (confidence interval or prediction interval) for a given scenario.

Goal:

Because we now have two quantitative variables collected, we need a way to distinguish between them:

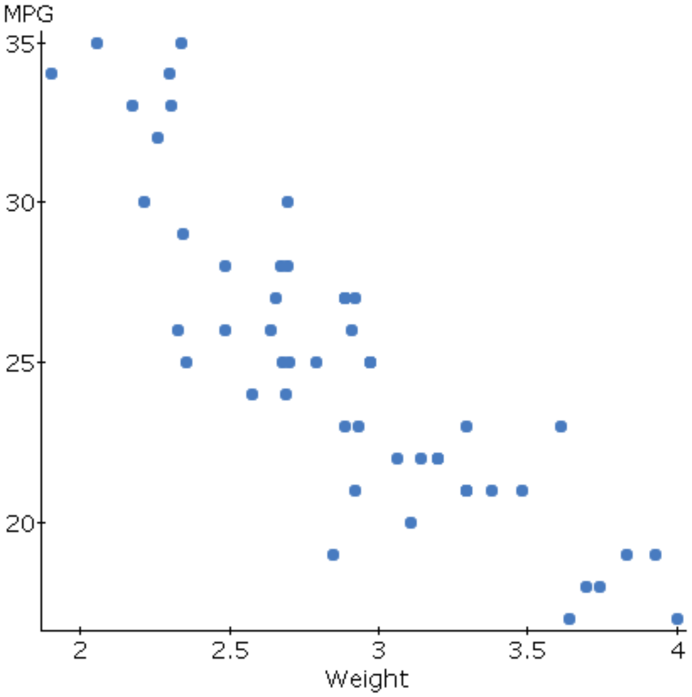
response variable:

predictor variable:

Example: A consumer organization would like to develop a model to predict a car's gas mileage (in miles per gallon) based upon its weight (in thousands of pounds). In order to do this, a random sample of 50 cars was selected with weights ranging between 1900 pounds and 4000 pounds.

What two variables were recorded for this sample of 50 cars?

To graphically inspect the relationship between these two variables, we can create a **scatterplot**. When given a scatterplot, the following three characteristics should be discussed:



-
-
-

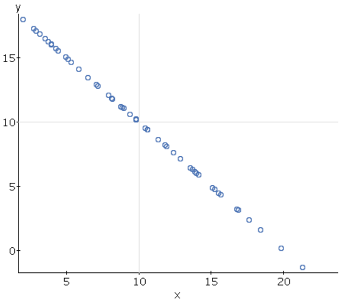
Using the scatterplot, discuss the relationship (in context) between the variables.

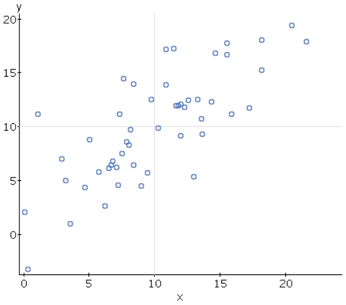
After we inspect a scatterplot and realize that a linear relationship exists, we need a way to quantify the strength of the relationship – this is done using the **correlation**.

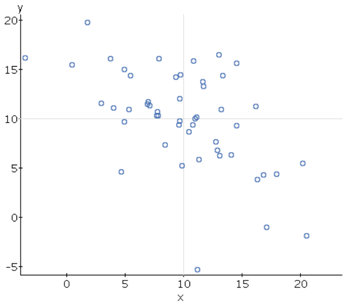
notation:

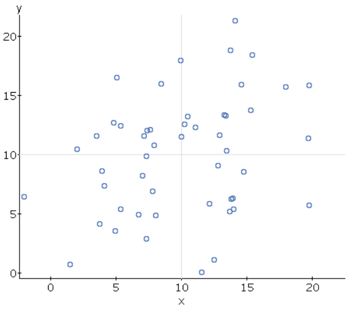
Properties:

Four scatterplots are shown below. Match the following correlation coefficients to the appropriate scatterplot: *r* = 0.76, *r* = 0.25, *r* = -0.55, *r* = -1.

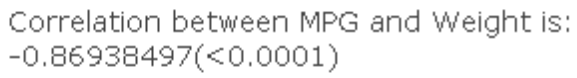








Because the correlation is difficult to calculate, we will rely upon StatCrunch for the calculation. The output for the correlation between weight & gas mileage is shown below:

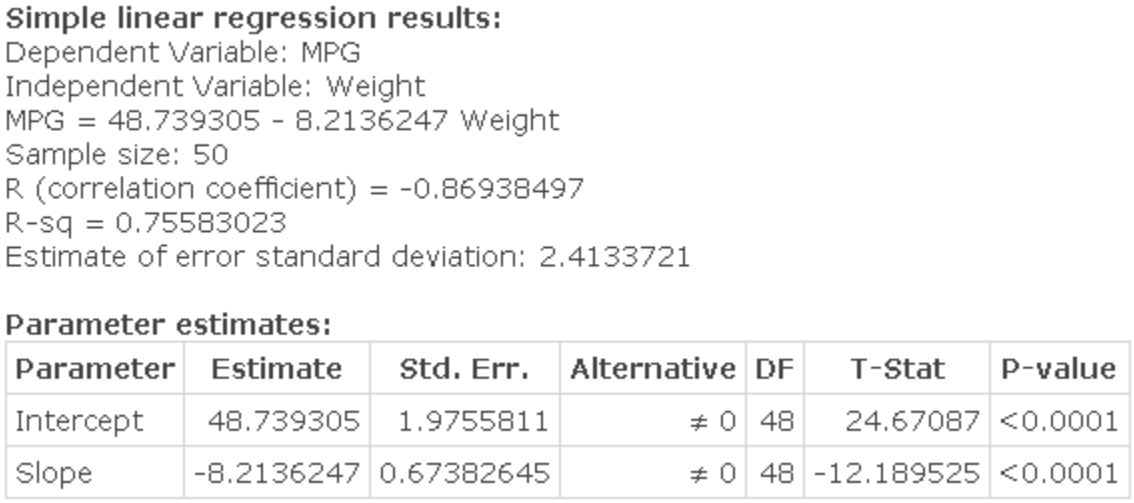


```
Correlation between MPG and Weight is:
-0.86938497(<0.0001)
```

The correlation reported is just the *sample* correlation – we need to determine if the correlation is significant. This will require a hypothesis test – determine if the correlation is significant using a significance level of 0.05:

If the scatterplot shows a linear relationship and the correlation is significant, then it seems reasonable to try to fit a straight line to the data – but, this doesn't mean connecting all the points! We want to find the equation of a line that fits the data the "best." This is done through a method called *least-squares regression* – this technique minimizes the squared distances between each point and the line. Before we fit the linear regression model, let's quickly review linear equations:

To find the **regression equation** or **line of best fit**, we will rely on StatCrunch. The regression output is given below:

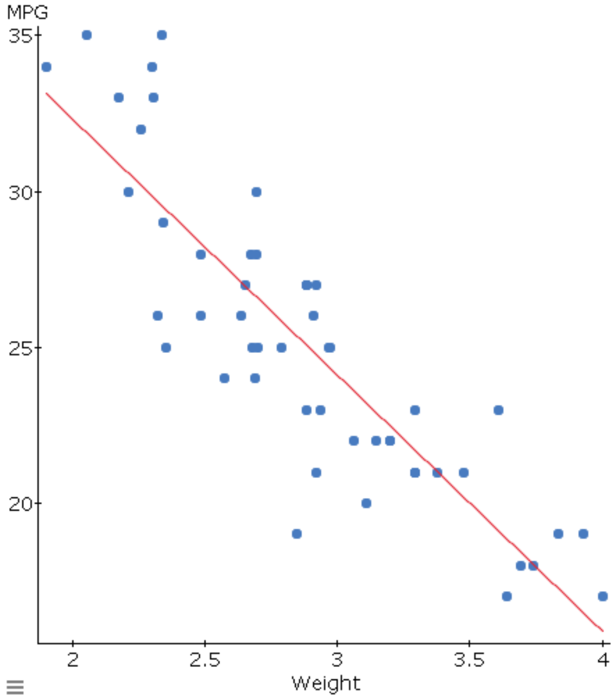


Simple linear regression results:
Dependent Variable: MPG
Independent Variable: Weight
MPG = 48.739305 - 8.2136247 Weight
Sample size: 50
R (correlation coefficient) = -0.86938497
R-sq = 0.75583023
Estimate of error standard deviation: 2.4133721

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	48.739305	1.9755811	≠ 0	48	24.67087	<0.0001
Slope	-8.2136247	0.67382645	≠ 0	48	-12.189525	<0.0001

What is the estimated regression equation?



Interpret the slope in context.

Can the intercept be interpreted?

Let’s briefly recap what we’ve done so far:

- two quantitative variables were collected from a sample of 50 cars
- a scatterplot visually showed a linear relationship between the two variables
- the correlation between the two variables was strong and the test showed it was also *significant*
- regression was used to find the *line of best fit* for the data

How can the regression equation be used?

Suppose we are interested in estimating the gas mileage for a car with a weight of 3236lbs.

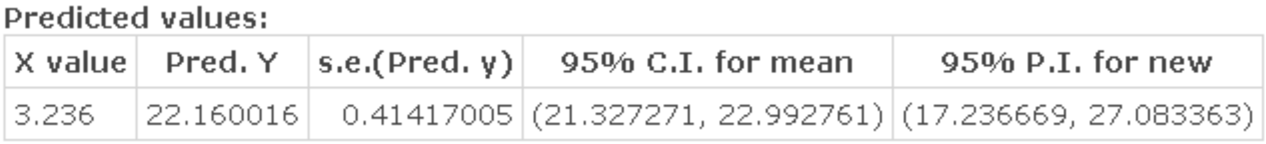
The estimate provided by the regression equation actually estimates the following two quantities:

-
-

Unfortunately, neither of these estimates has been made with any *confidence* – to introduce confidence into our estimate, we need to use *intervals*. There will be two different intervals reported by StatCrunch, which estimate the two quantities above:

-
-

These intervals will always be provided by StatCrunch. Our focus will be on interpreting the intervals using the output provided.

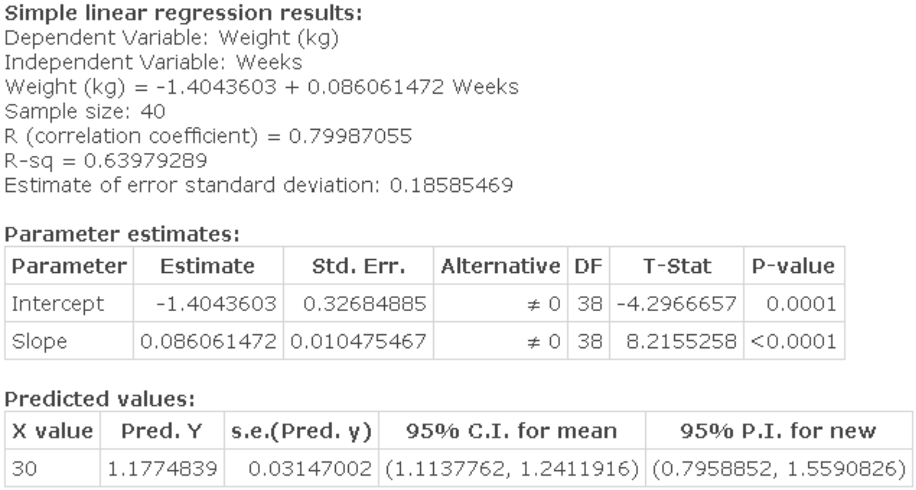


Predicted values:

X value	Pred. Y	s.e.(Pred. y)	95% C.I. for mean	95% P.I. for new
3.236	22.160016	0.41417005	(21.327271, 22.992761)	(17.236669, 27.083363)

Example: Data was collected on preterm infants at three different hospitals. The birthweight (in kg) of each baby was recorded, as well as the gestational age. After developing and testing a model for this data, we are interested in the birthweight for a gestational age of 30 weeks.

- Identify the two variables which were recorded, determining which is the predictor and which is the response.
- Interpret the scatterplot provided *in context*.
- Determine if the linear relationship between the gestational age & birthweight is significant. Use α = 0.05.
- What is the regression equation? Interpret the estimate for the slope in context.
- A researcher is interested in estimating the average birthweight for babies born at 30 weeks. Provide a point estimate for this age, as well as interpreting the appropriate interval.



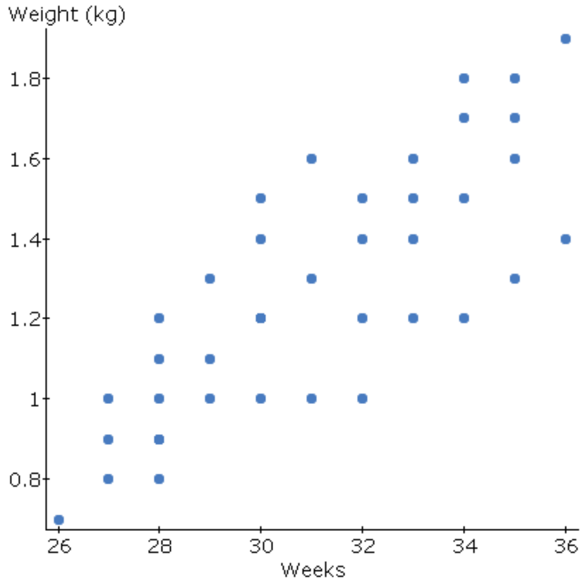
Simple linear regression results:
Dependent Variable: Weight (kg)
Independent Variable: Weeks
Weight (kg) = -1.4043603 + 0.086061472 Weeks
Sample size: 40
R (correlation coefficient) = 0.79987055
R-sq = 0.63979289
Estimate of error standard deviation: 0.18585469

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	-1.4043603	0.32684885	≠ 0	38	-4.2966657	0.0001
Slope	0.086061472	0.010475467	≠ 0	38	8.2155258	<0.0001

Predicted values:

X value	Pred. Y	s.e.(Pred. y)	95% C.I. for mean	95% P.I. for new
30	1.1774839	0.03147002	(1.1137762, 1.2411916)	(0.7958852, 1.5590826)


Weight (kg)
Weeks