MULTIPLE LINEAR REGRESSION

MLR

Multiple Linear Regression: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \epsilon$

is just like

Simple Linear Regression: $y = \beta_0 + \beta_1 x + \epsilon$

But instead of using one predictor variable we will use multiple predictor variables to make a prediction about a single dependent variable.

Most of the SLR concepts still pertain to MLR.

We are still finding a best fit 'line' thru the data.

We are still using the least squares method finding the smallest sum of all the squares of the residuals.

R^2 also exists for MLR

Simple Linear Regression: $y = \beta_0 + \beta_1 x + \epsilon$

Multiple Linear Regression: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \epsilon$

The difference is that now we have multiple predictor variables: $x_1, x_2, x_3, \ldots$
And each predictor variable x_i has a coefficient β_i.

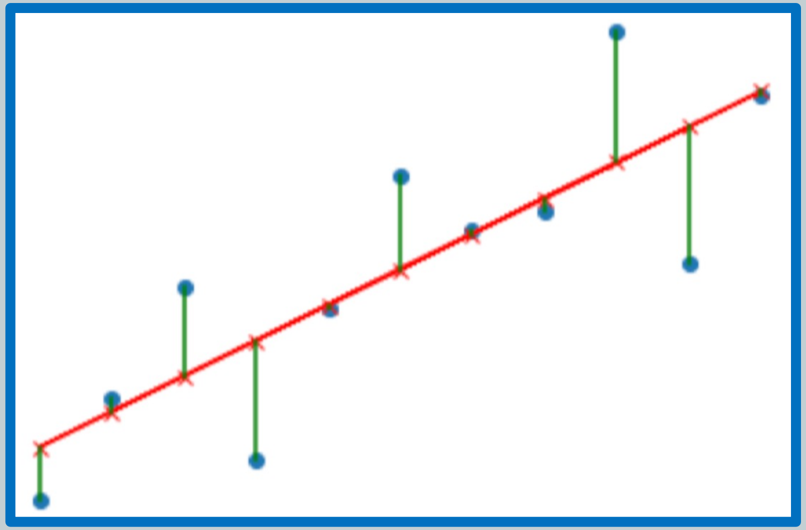
There is still an error term.

In other words, there is still some amount of uncertainty.

The equation is never going to be exactly accurate.

In SLR $y = \beta_0 + \beta_1 x + \epsilon$:

β_0 is the intercept and β_1 is the slope.



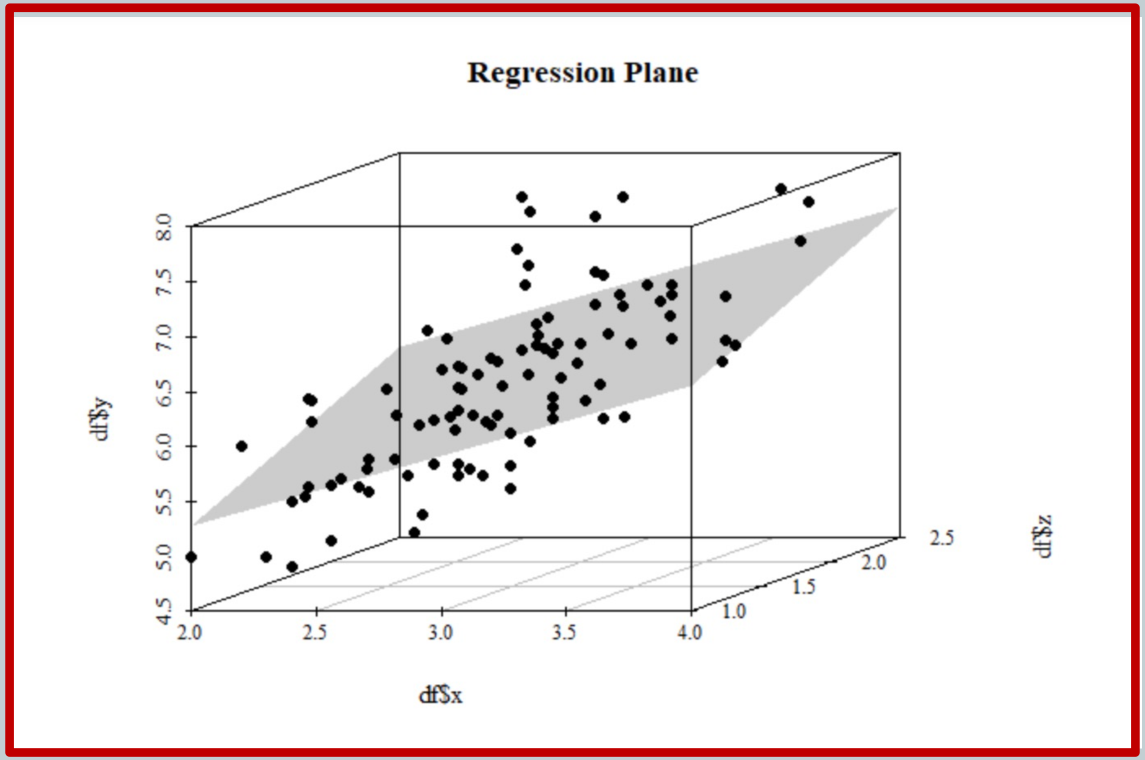
What does each coefficient mean in MLR:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \epsilon$$

β_0 is the y-intercept of the hyperplane.

β_i is the amount of change in the dependent variable with a one-unit change in the predictor variable x_i.

That is to say y will change an amount β_i with a one-unit change in x_i.



The Objectives of MLR:

1] To come up with a predictive model for the dependent variable.
To answer the question, "How is y affected by all of these x_i factors?"

2] To determine the relationship between 1 or 2 predictor variables and the y variable, but we throw in a whole lot of other predictor variables to control for their affects.

In other words, we are really interested in these 1 or 2 predictor variables but we have to throw in all these other predictor variables to get the true affect of the variables we are really interested in.

I want to know the relationship between x_1 and y but I threw in x_2 and x_3 to control for their affects.

"I want to know the relationship between x_1 and y but I threw in x_2 and x_3 to control for their affects."

What does that mean?

Suppose Zillow wants to predict the price of a home in Colorado.

It is decided that home size (square footage) would be a great predictor.

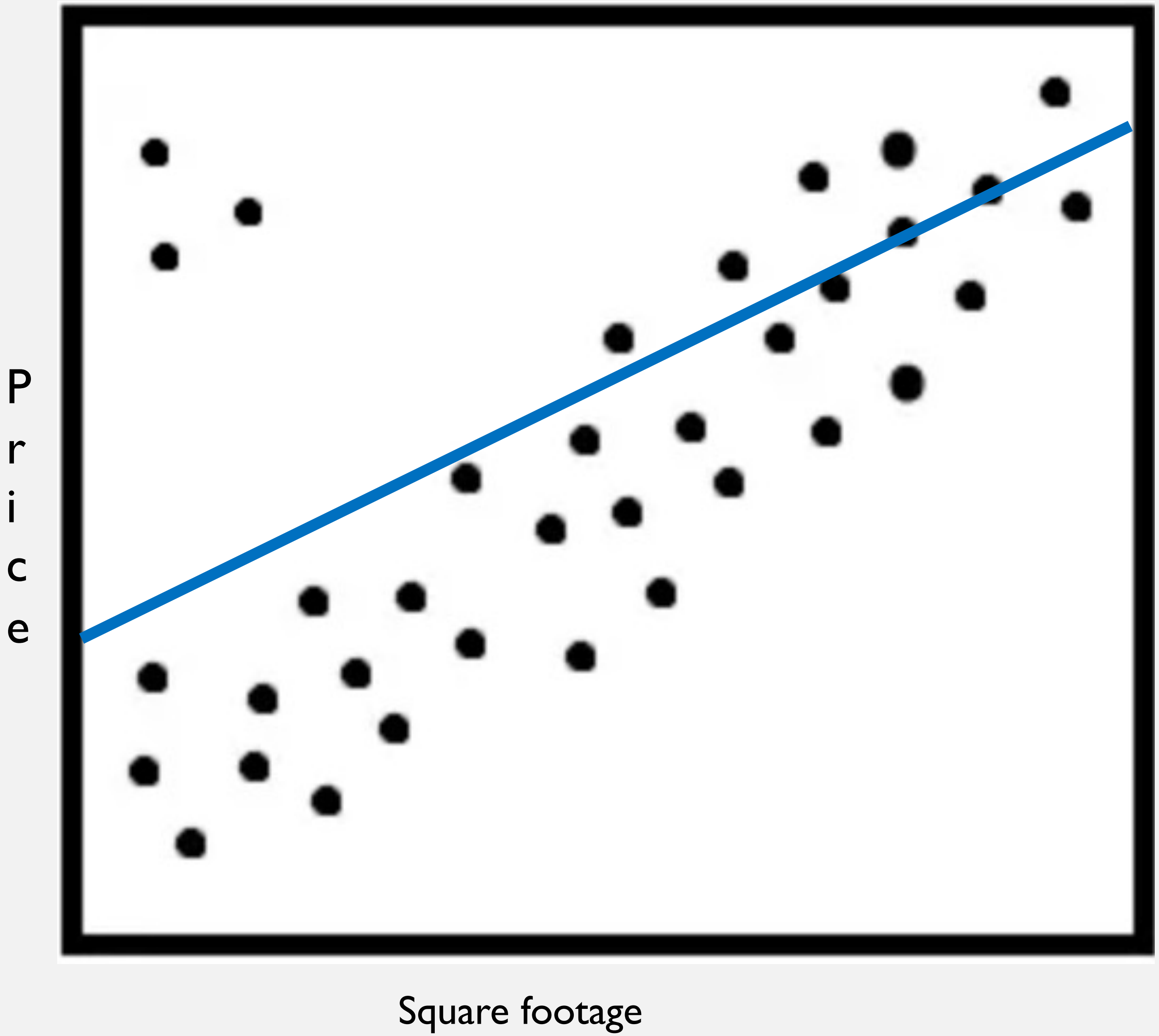
predictor variable:

x_1 = square footage

dependent variable:

y = price (cost of home)

So, Zillow runs a simple linear regression on x and y.



The resulting plot might show some oddities. For instance, there are a number of smaller homes going for larger prices
These homes change the fit of the best line.

The reason these smaller houses are going for big prices might be because they are found in neighborhoods with great schools which puts them in high demand with families.

Therefore, the line of best fit for the data is not giving the best predictions.

Zillow would get a more accurate picture by throwing in additional variables in order to control for their affects.

Small (desirable houses) amongst large expensive homes.

Zillow should control the affect with another variable:

Quality of schools in neighborhood.

A more accurate picture arises by throwing in additional variables in order to control for their affects.

Other possible factors we might add to control for their affect:

Median income of folks who live in the neighborhood
Age of home
Size of lot that the home sits on
as well as Quality of schools in the neighborhood.

These other factors will give us a more accurate account of the affect of square footage on price, which is what we are really interested in.

Each time you add in a new x_i and find a new best fit, the other β_i will change and the equation will become more accurate.

But picking the 'right' predictor variables (as well as the right amount of x_i) to put into the regression equation is an art and a difficult science.

We need to pick predictors that make sense.
Don't pick a predictor on home price like the realtor's sign color.

Perhaps it is a color that catches more eyes and therefore creates a bidding war between multiple customers and drives up the price or there could be a correlation that makes the predictor look significant. Sign color (red and yellow) relates to subconscious hunger which then relates to price.

And if you throw in enough predictors, even at 95%, 1 in 20 will come out as being significant just by chance; type I error.

Picking bad predictor variables can make mess up the regression (overfitting) and the x_i may come out as looking like they are significant when in reality they are not.

There is another, more subtle mistake, which is to pick variables that have redundancy.

For instance, if one of your predictors is house size, then you probably should not also include number of rooms as a second predictor; they are basically the same.

This type of redundancy is called multicollinearity.

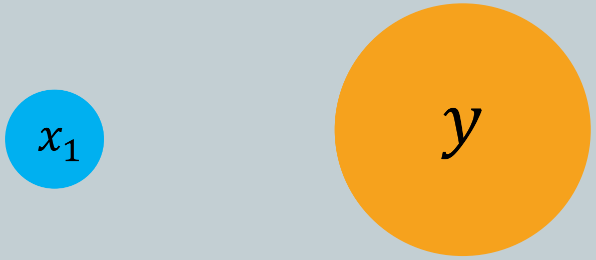
Multicollinearity: Two predictors that do have an affect on y,
but they have the same affect.

This introduces the problem: Which of the factors has more of an affect?
Or do neither of them have any affect?

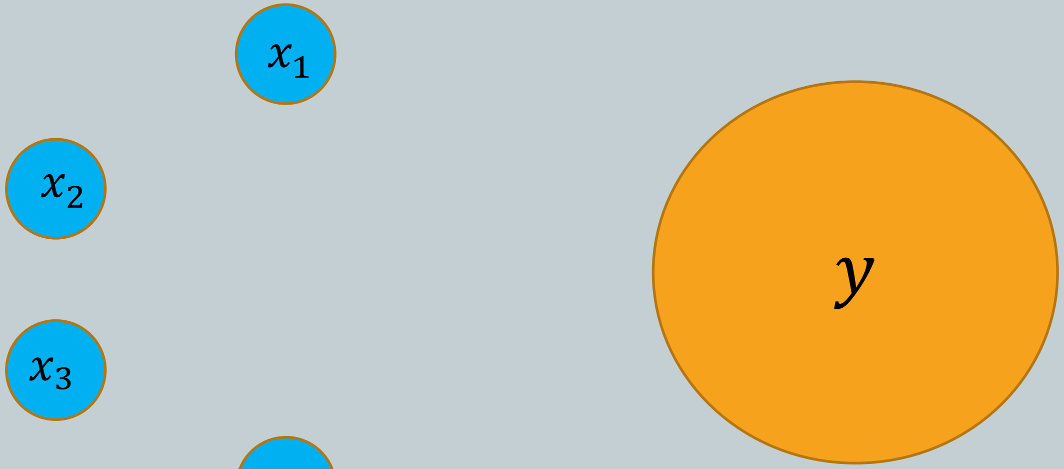
If you are cooking and creating a new recipe and experimenting with several different spices, if you mix pink sea salt and black flake salt into your dinner, then which one caused the saltiness or lack thereof?

Keep in mind that MLR is just an extension of SLR.

And while SLR is a one-to-one relationship,



MLR is a many-to-one relationship.



After picking the predictor variables, one should perform some prep work before doing multiple regression analysis due to overfitting and multicollinearity:

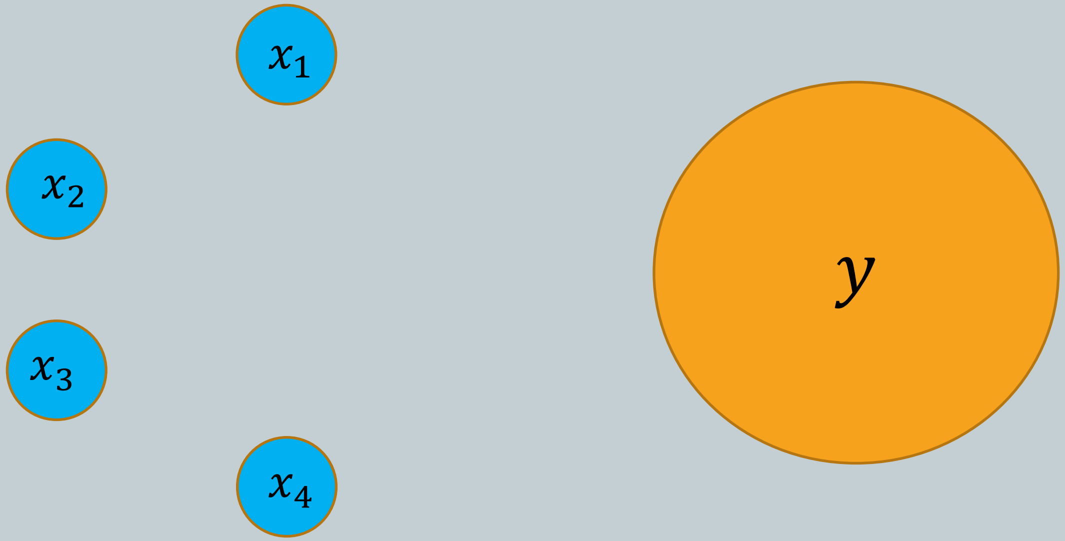
Look at correlations

Look at scatterplots

Perform SLR between each independent variable and the dependent variable.

Look at relations between each of the independent variables as well.

picture the 4 to 1 relations (star picture)



Remember in the many-to-one concept that adding independent variables does not always mean better predictions.

It could make things worse via overfitting.

Independent variables will explain more variation, but this also introduces more problems.

So, try to just pick the ‘best’ variables.
The addition of more variables creates more relationships between them (problem).
They can be related to each other, i.e., Multicollinearity: when the independent variables are related to each other.

Ideally, we want the independent variables, $x_i^{\prime s}$ to be related to the depend variable, y, **but not to each other**.

Salt adds to the recipe, but two different kinds of salt introduce redundancy, aka multicollinearity.

Example:

Let’s say Bob has created a new business: LongLift.

It is like long distance Uber.
Bob oversees several vehicles that make pickups over long distance trips.

Bob wants to know how to predict travel time (in hours) for his business.
He decides on three predictor variables (miles traveled, number of pick-ups, price of gasoline).

He wants to predict the travel time, y, with x_1, x_2, and x_3
He wants to know in what way does y depend on x_1, x_2, and x_3 ?

Bob then collects data: distance in miles, number of pickups during the trip., and the current price of gasoline. All from a random sample of 10 trips.

Dependent Variable: $y =$ travel time in hours.

Miles traveled	Number of Deliveries	Gas Price	Travel Time
x_1	x_2	x_3	y
89	4	3.84	7
66	1	3.19	5.4
78	3	3.78	6.6
111	6	3.89	7.4
44	1	3.57	4.8
77	3	3.57	6.4
80	3	3.03	7
66	2	3.51	5.6
109	5	3.54	7.3
76	3	3.25	6.4

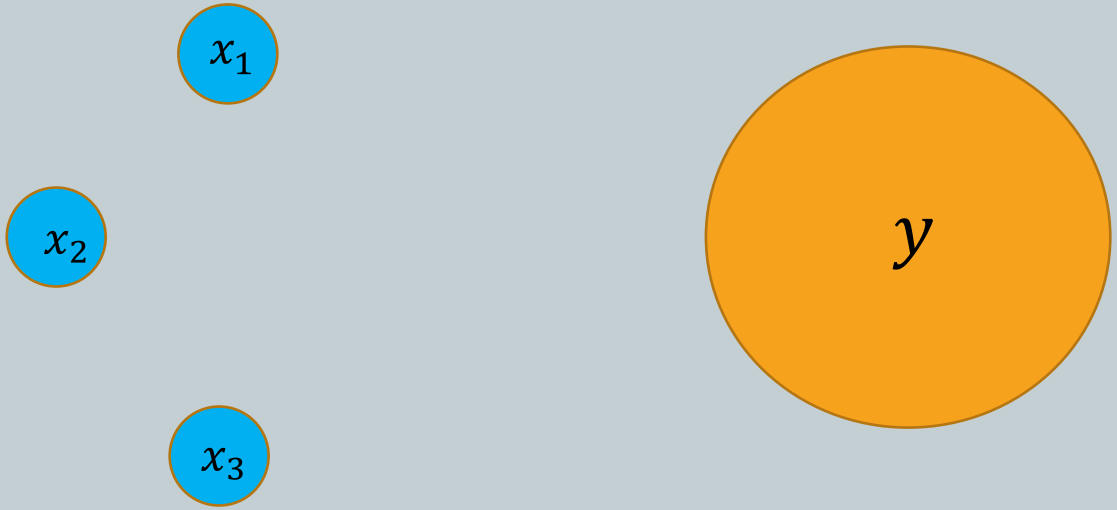
After picking the predictor variables, one should perform some prep work before doing multiple regression analysis due to overfitting and multicollinearity:

Look at correlations

Look at scatterplots

Perform SLR between each independent variable and the dependent variable.

In this LongLift example we need to look at relationships between x_1, x_2, and x_3 to y as well as the x_1 to x_2, x_1 to x_3, x_2 to x_3 relationships.



With numerous predictor variables the art is deciding which independent variables ultimately make the cut.

Some are better at predicting than others.
Some do nothing to help prediction.

We want all of the x_i to be related to y but **not to each other**.

The process of conducting MLR

Generate a list of potential variables; independent and dependent

Collect data on the variables

Check the relationships between each independent variable and the dependent variable using scatterplots and correlations

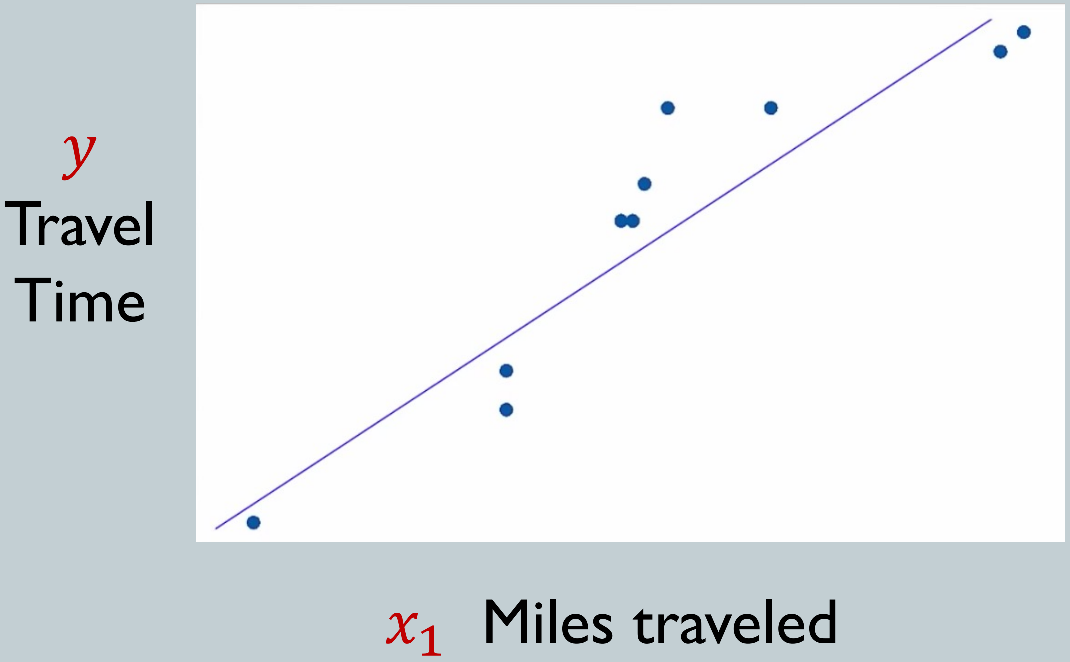
Check the relationships among the independent variables using scatterplots and correlations

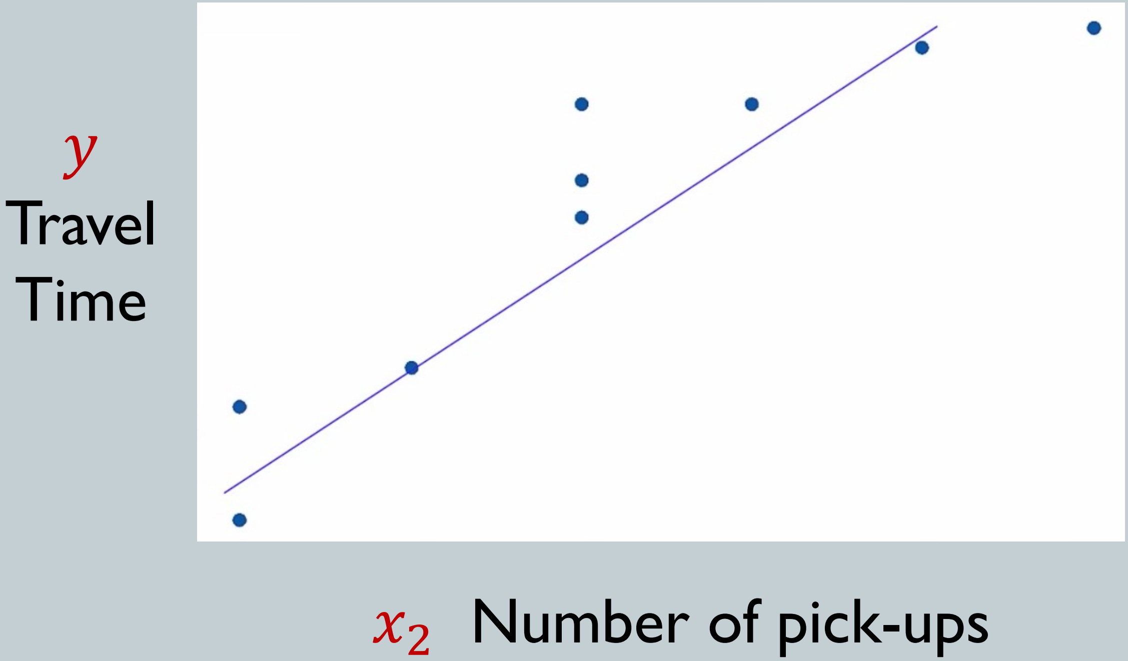
Conduct SLR for each x_i and y pair.

Use the non-redundant independent variables in the analysis to find the best fitting model.

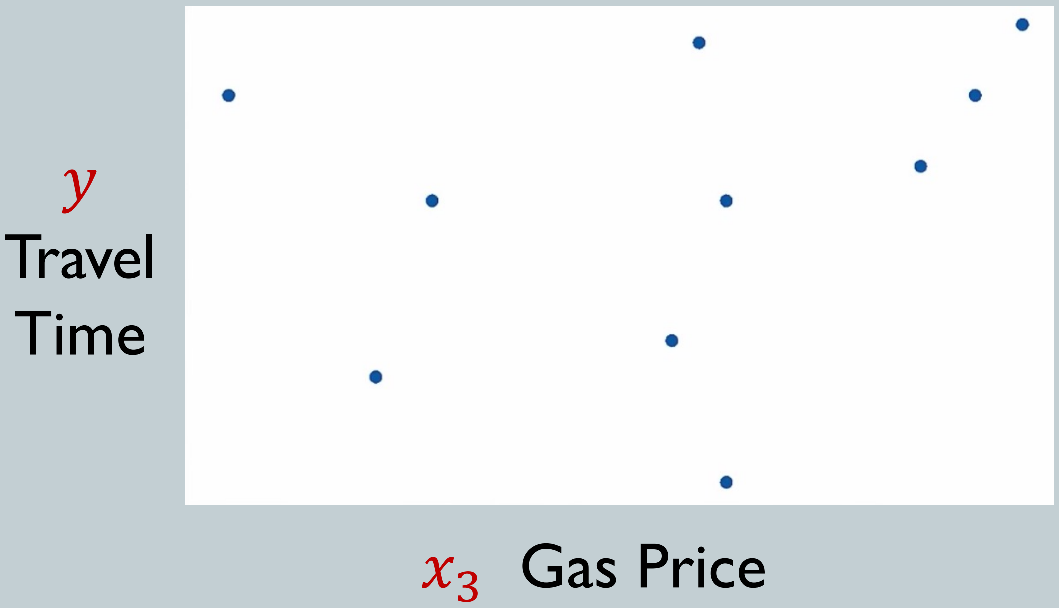
Use the best fitting model to make predictions about the dependent variable.

So, let's first look at the independent variable to dependent variable relationships. We are checking for relevancy of the independent variable.





Relationships between all the independent and dependent variables



There appears to be a strong linear relationship between x_1 and y, and also between x_2 and y.

i.e., between miles traveled and travel time as well as between number of pick-ups and travel time.

There does not seem to be a linear relationship between x_3 and y.

i.e., between gasoline prices and travel time.

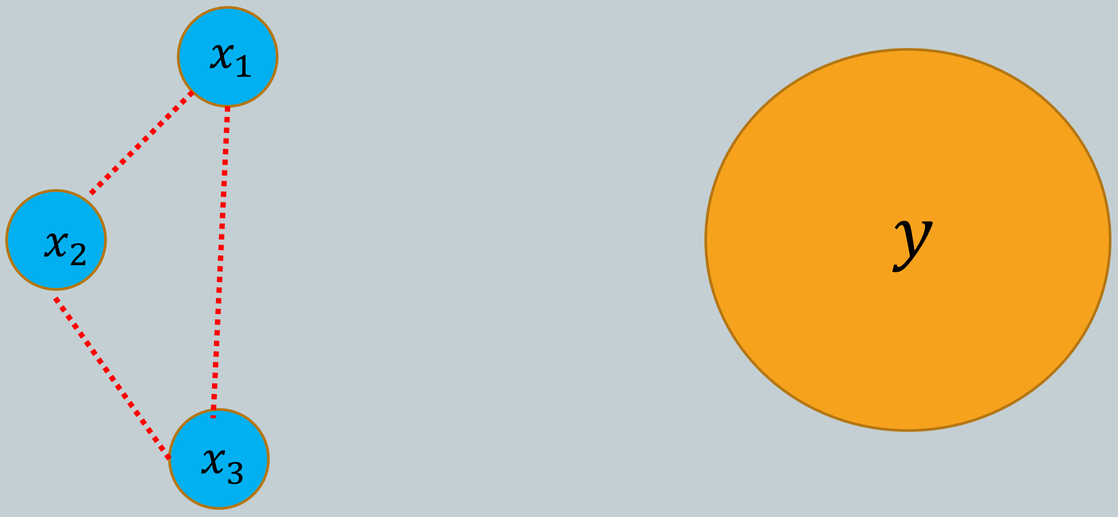
Just from a visual perspective x_3 does not appear to have a strong linear relationship to y.

Travel time, y, appears highly correlated with mile traveled, x_1.
Travel time, y, appears highly correlated with number of pick-ups, x_2.
Travel time, y, does NOT appear to be highly correlated with gas price, x_3.

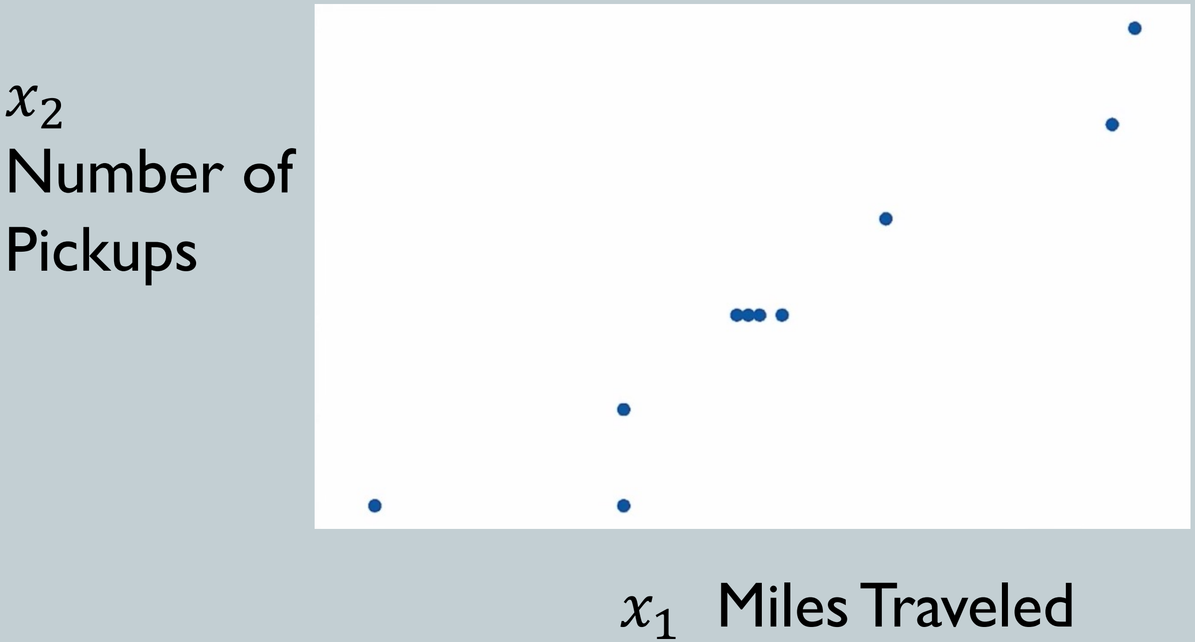
Because x_3 (gas price) does not appear to be correlated with the dependent variable, we should not use that variable in the multiple regression.

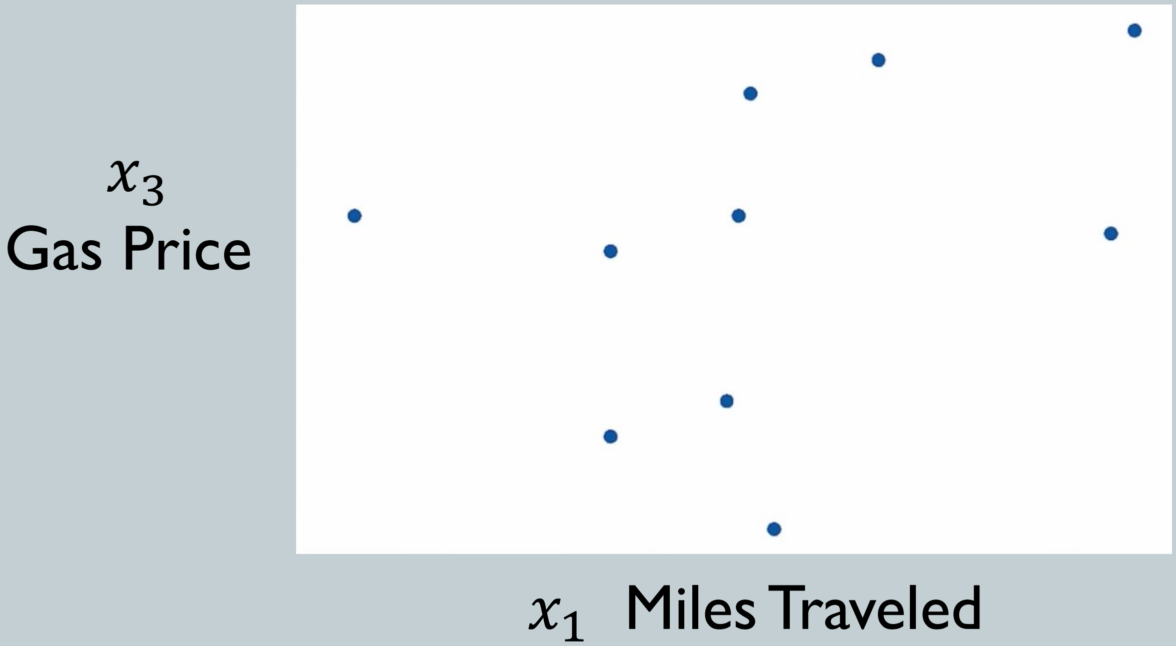
There are still three more relationships to investigate.

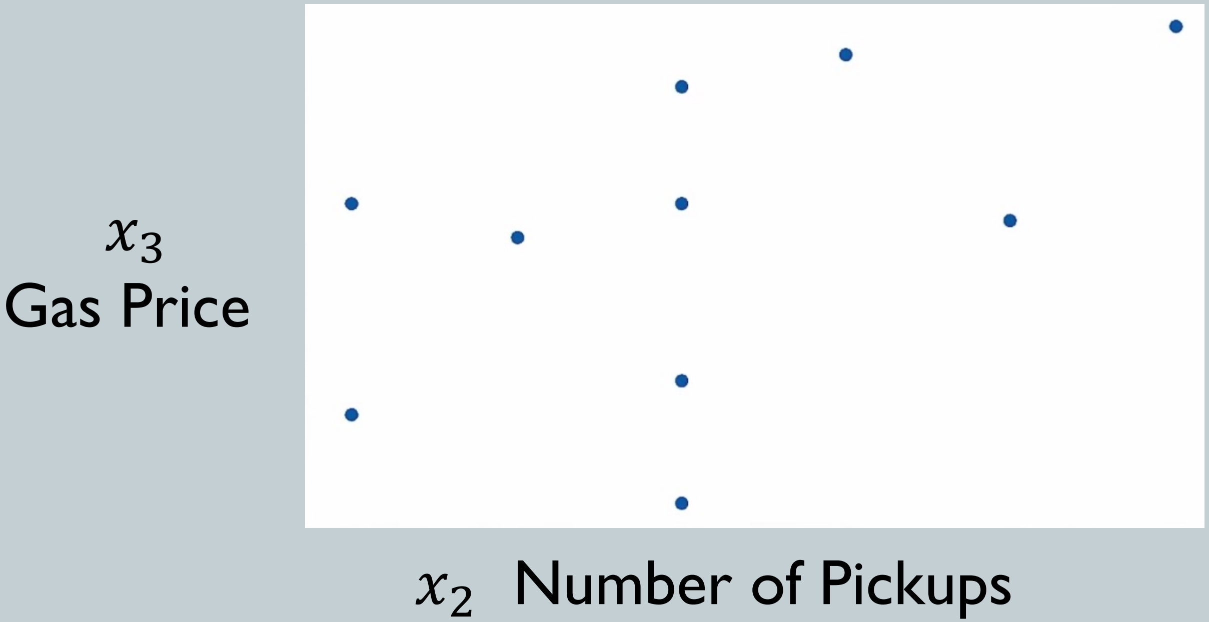
We must check for multicollinearity. Correlation between the independent variables themselves which can cause regression problems.



Collinearity Check







Visually there does not appear to be a linear pattern/relationship between x_1 and x_3 nor is there one between x_2 and x_3, which is good.

Miles traveled and gas price. Number of pick-ups and gas price.

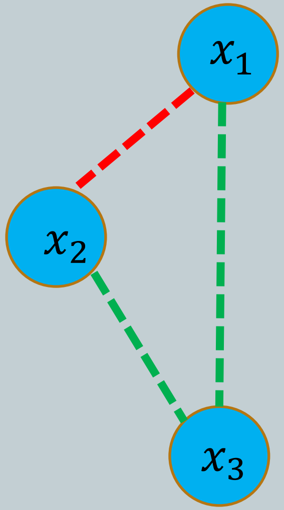
However, there does appear to be a linear relationship between x_1 and x_2, which is bad.

Miles traveled and Pick-ups.

Miles traveled and pick-ups are highly correlated; this is a problem.
they have a highly linear relationship. This is the definition of multicollinearity.

Including both x_1 and x_2 will cause serious problems because it is unknown what coefficients to assign to the two variables since they are so similar.
They are two 'salts' that are too similar in taste to detect a difference.

x_1 and x_3 are not correlated. This is a good thing.
x_2 and x_3 are not correlated. Again, this is good.



But we do have a problem with multicollinearity with x_1 and x_2.

The number of pick-ups, x_2, appears to be highly correlated with miles traveled, x_1; This is multicollinearity.

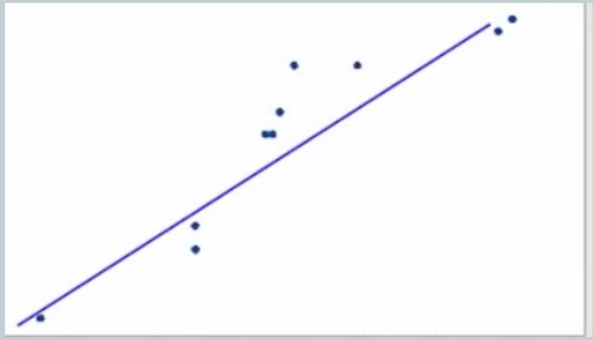
Since these two are highly correlated they should not both be in the multiple regression; they are redundant.

Miles traveled, x_1, does not appear to be highly correlated with gas price, x_3.
Gas price, x_3, does not appear to be correlated with number of pick-ups, x_2.

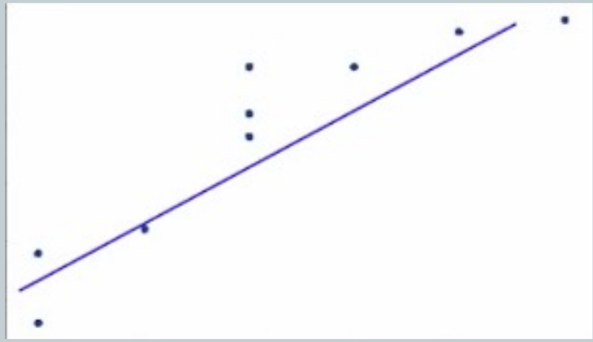
Note that we made this decision just by looking at scatterplots.
We should do a correlation analysis between the variables.
Look at each pairs r and p-value. A stats package will show:

Independent versus Dependent relationships

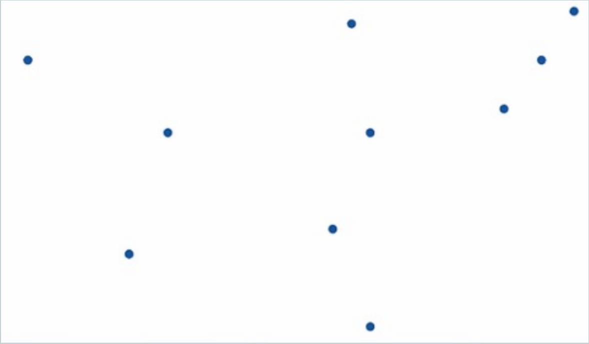
☺ Travel time versus miles traveled:
$r = 0.928$ with a p-value < 0.001



☺ Travel time versus number of pick-ups:
$r = 0.916$ with p-value < 0.001



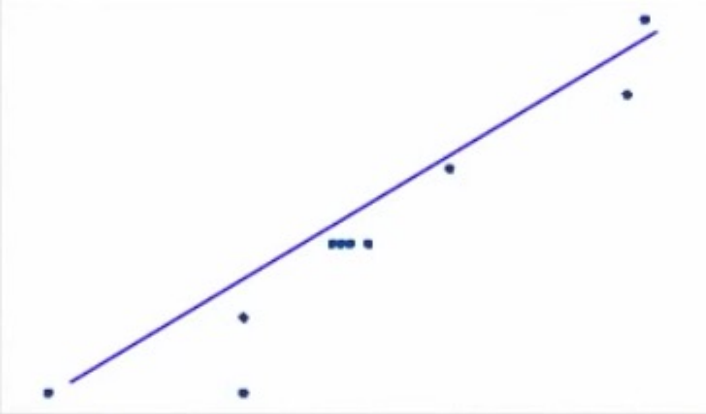
☹ Travel time versus gas prices:
$r = 0.267$ with a p-value 0.455.



Independent variable versus Independent variable

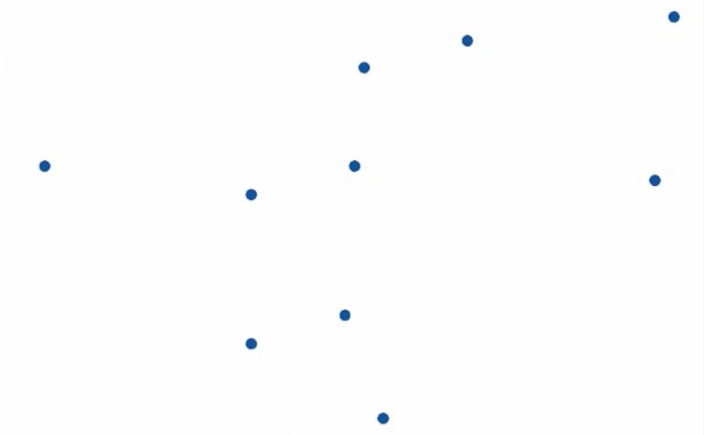
☹ Miles traveled versus Number of Pick-ups:

$r = 0.956$ and p-value < 0.001



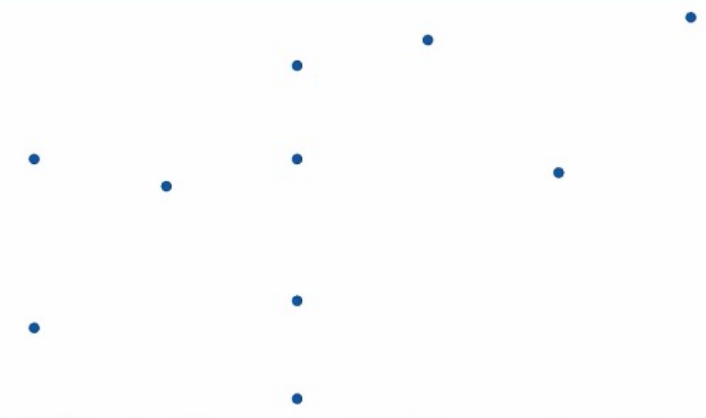
☺ Miles traveled versus Gas Prices:

$r = 0.356$ and p-value $= 0.313$



☺ Number of Pick-ups versus Gas Prices:

$r = 0.498$ and p-value $= 0.143$



Correlation analysis confirms the conclusions reached by visual examination of the scatterplots, with the addition of some quantification.

Miles Traveled and number of pick-ups are both highly correlated with each other and therefore are redundant; only one should be used in the multiple regression analysis.

Gas Price is NOT correlated with the dependent variable and should be excluded.

Always consider whether the assumptions are being met, and what the data looks like. i.e., look at scatterplots and then correlation analysis.

Do not blindly accept values as good or bad without reason.

Why is R^2 good/bad ? Are you sure?

Look at scatterplots, correlation analysis, individual regression, and group regression.

Next Time: MLR Inference.