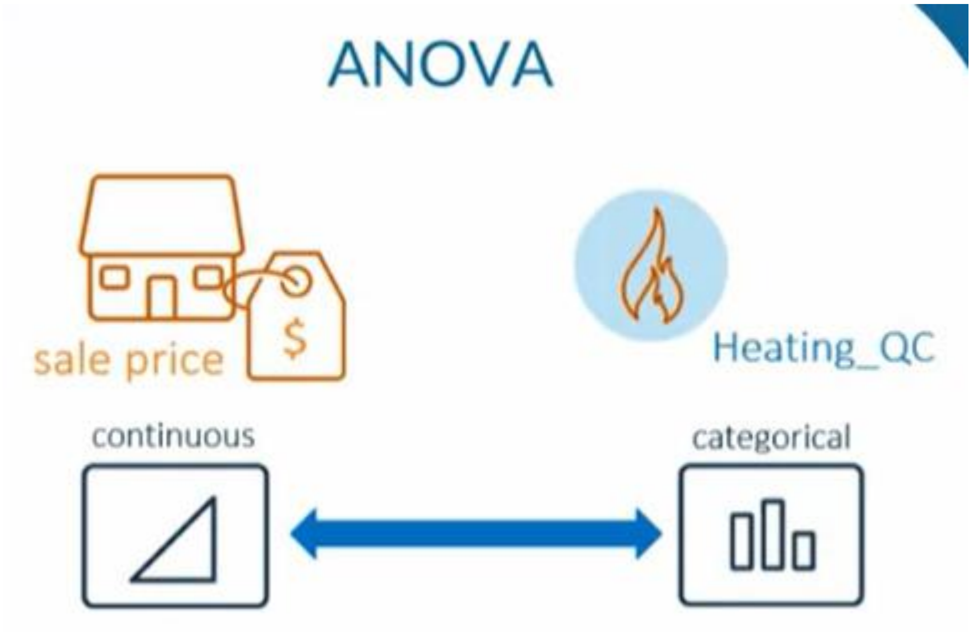
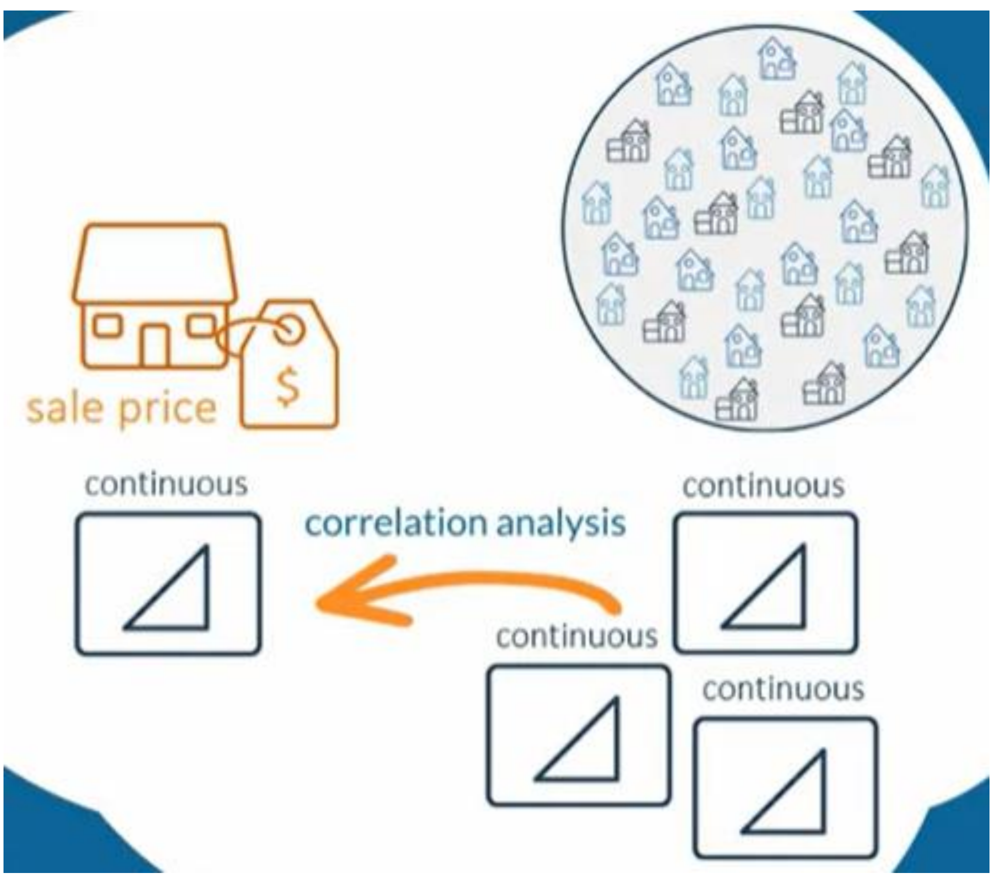
SBA: Statistical Business Analyst with SAS

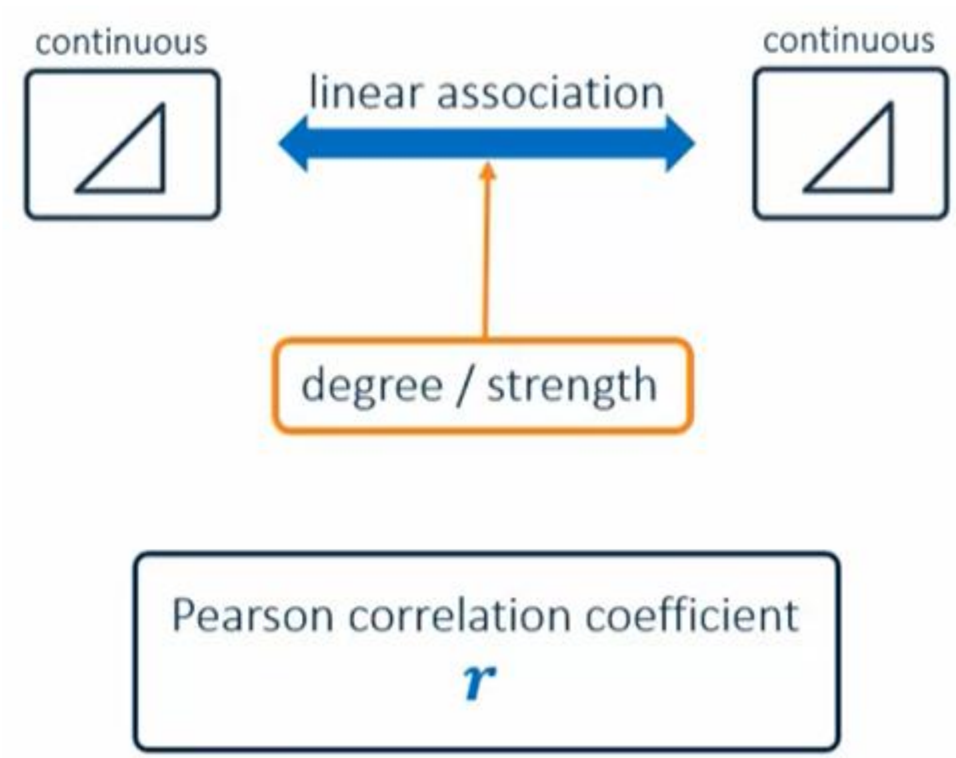
SBA1 Introduction to Statistical Analysis: Hypothesis Testing

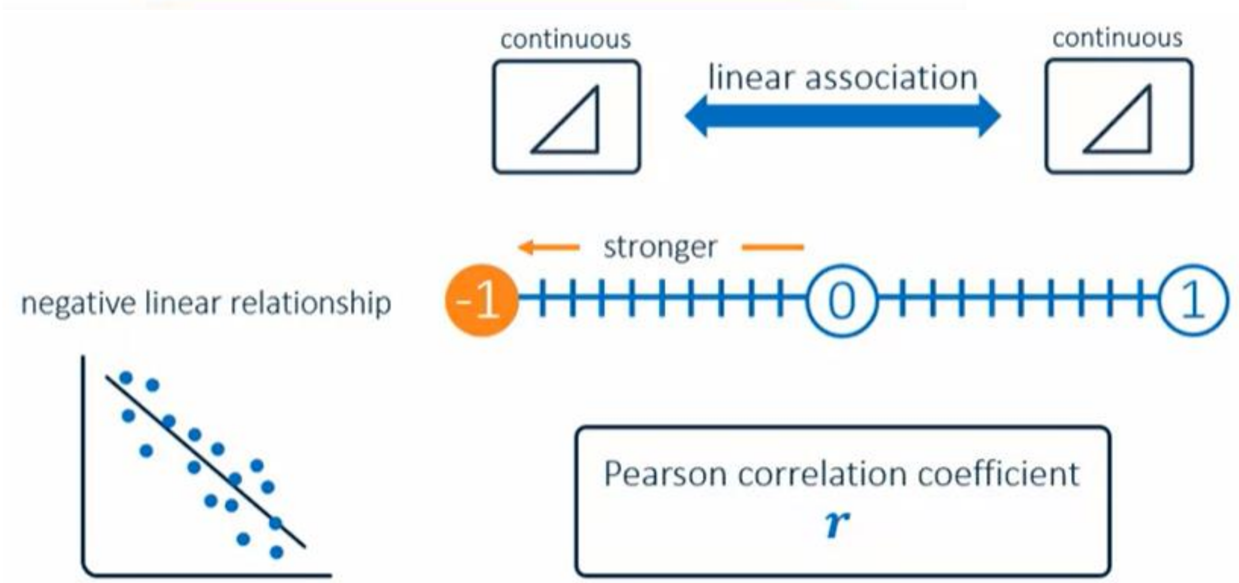
W2 Pearson Correlation and Simple Regression

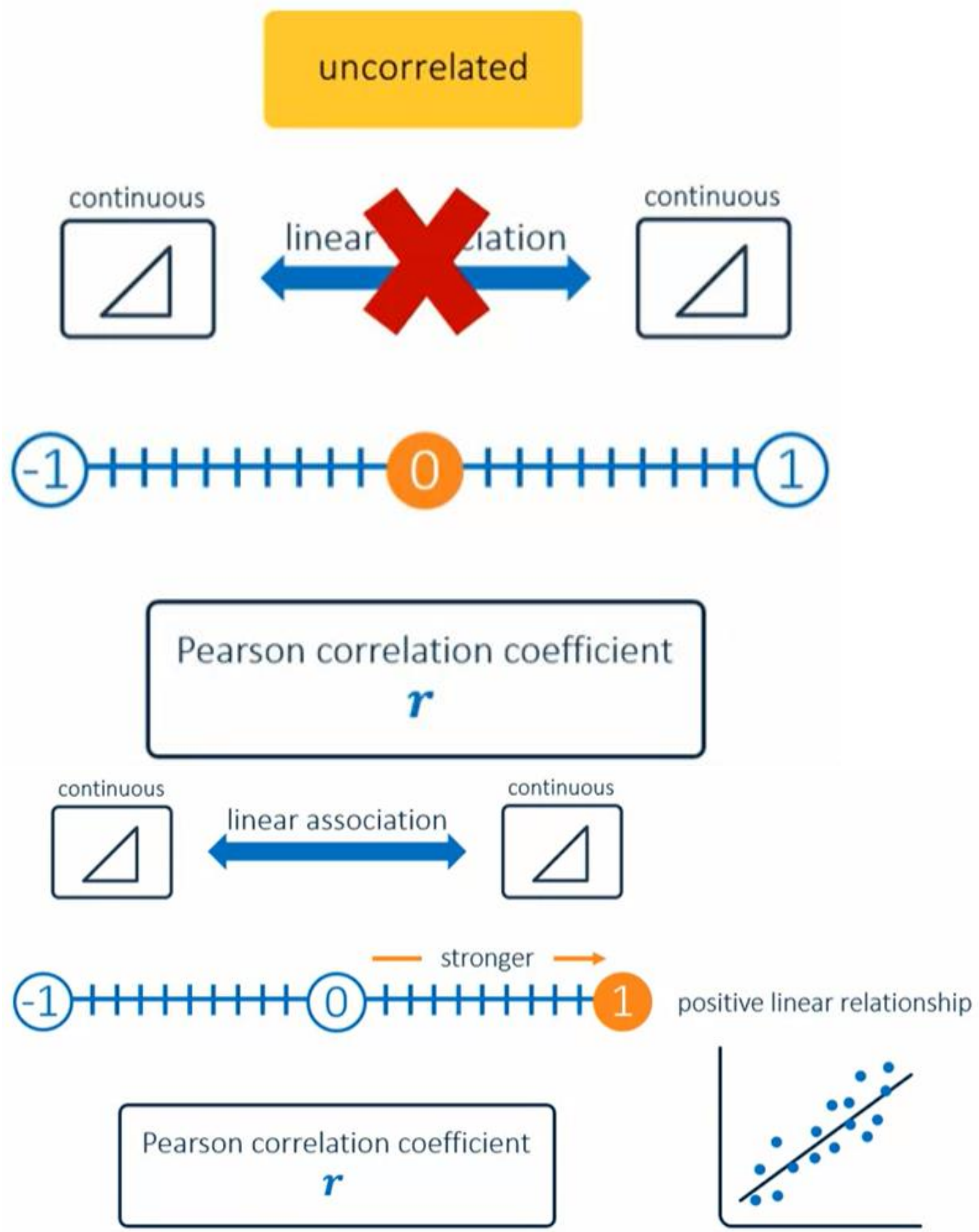




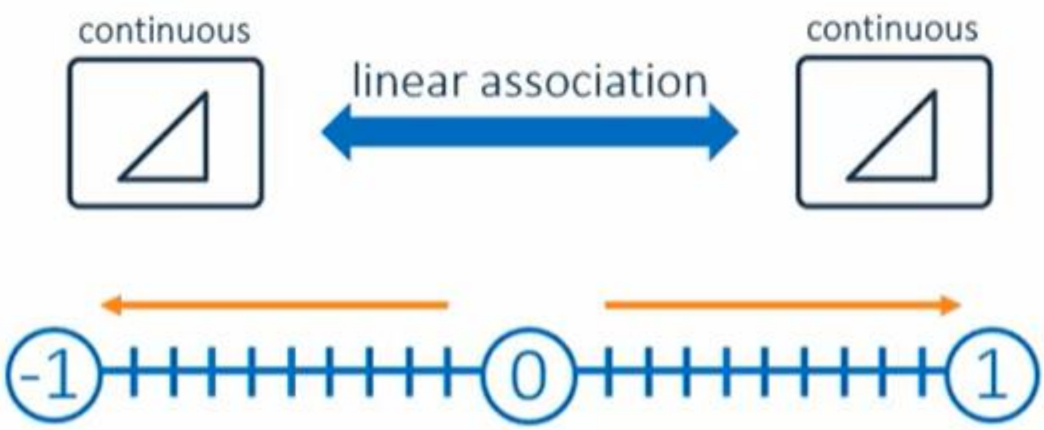
Using Correlation to Measure Relationship Between Continuous Variables






uncorrelated
continuous
linear
continuous
-1
0
1
Pearson correlation coefficient
r
continuous
linear association
continuous
stronger
-1
0
1
positive linear relationship
Pearson correlation coefficient
r

Hypothesis Testing for a Correlation



	population parameter	sample statistic
correlation	ρ	r

$$H_0 : \rho = 0 \qquad H_a : \rho \neq 0$$

	population parameter	sample statistic
correlation	ρ	r

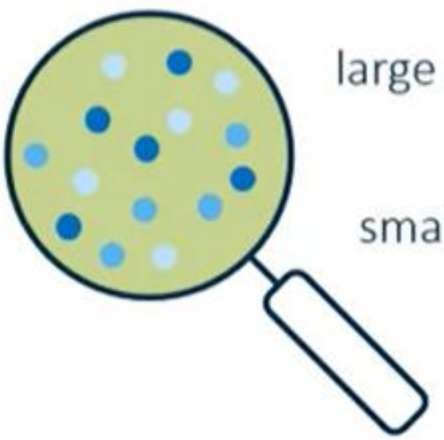
p-value does not measure strength

$H_0 : \rho = 0$ $H_a : \rho \neq 0$

	population parameter	sample statistic
correlation	ρ	r

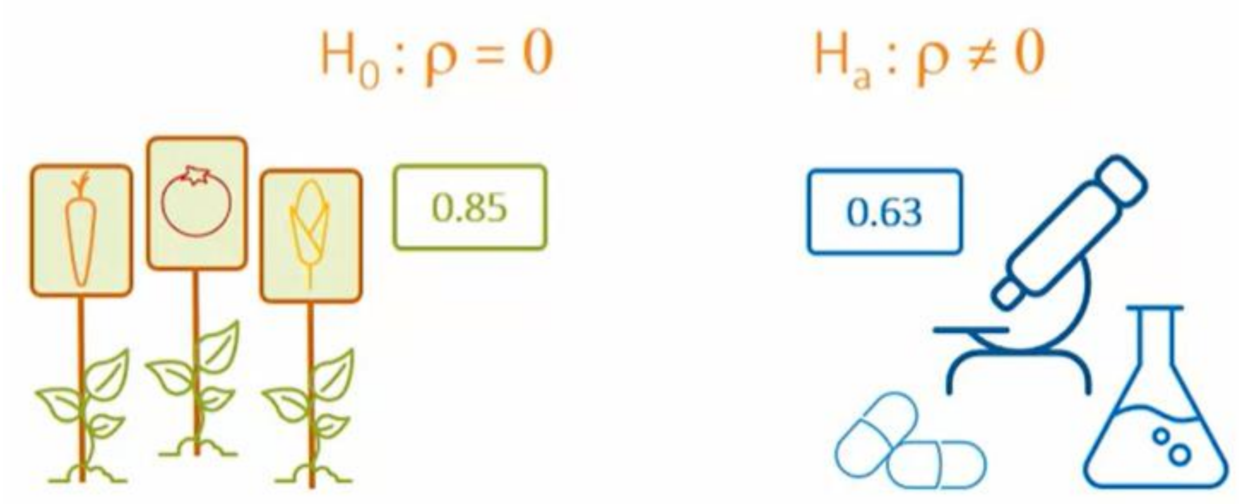
does measure strength

$H_0 : \rho = 0$ $H_a : \rho \neq 0$



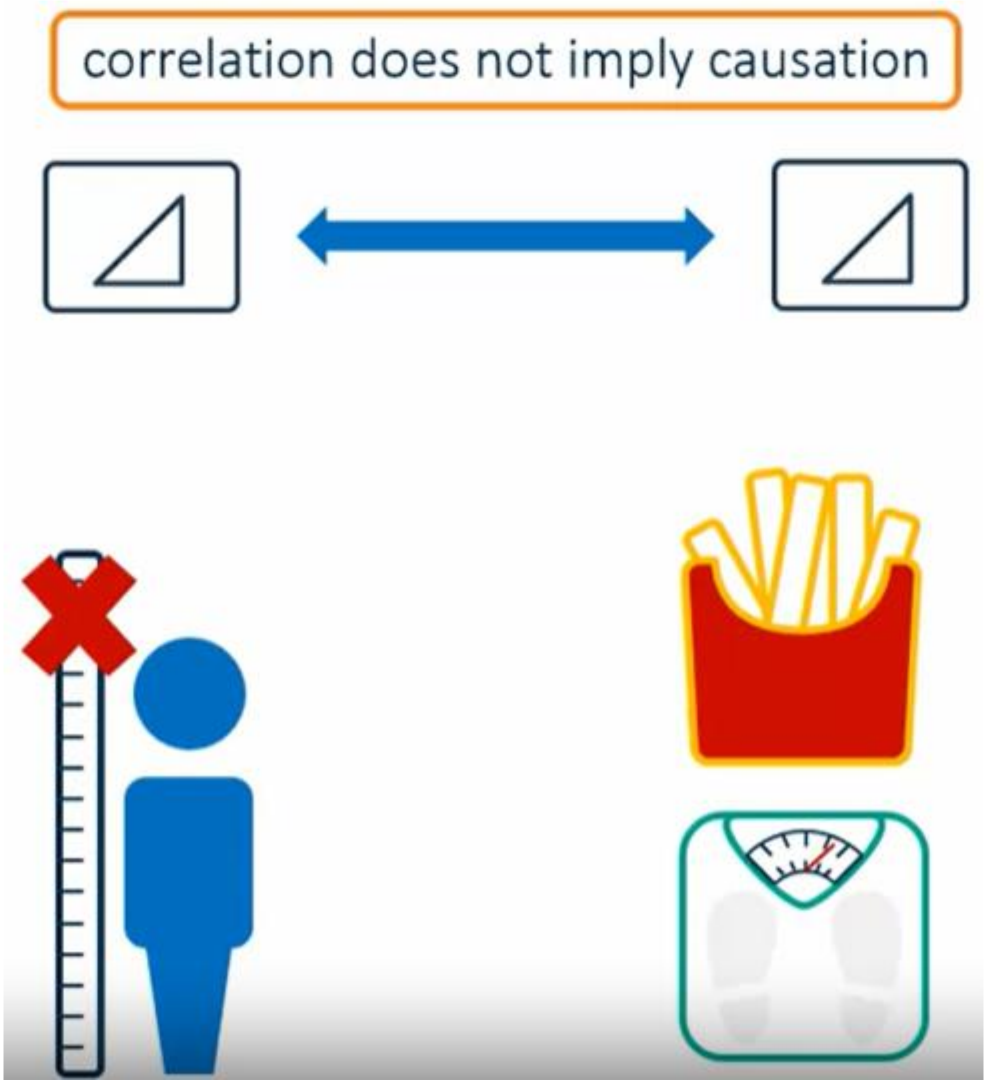
large sample sizes

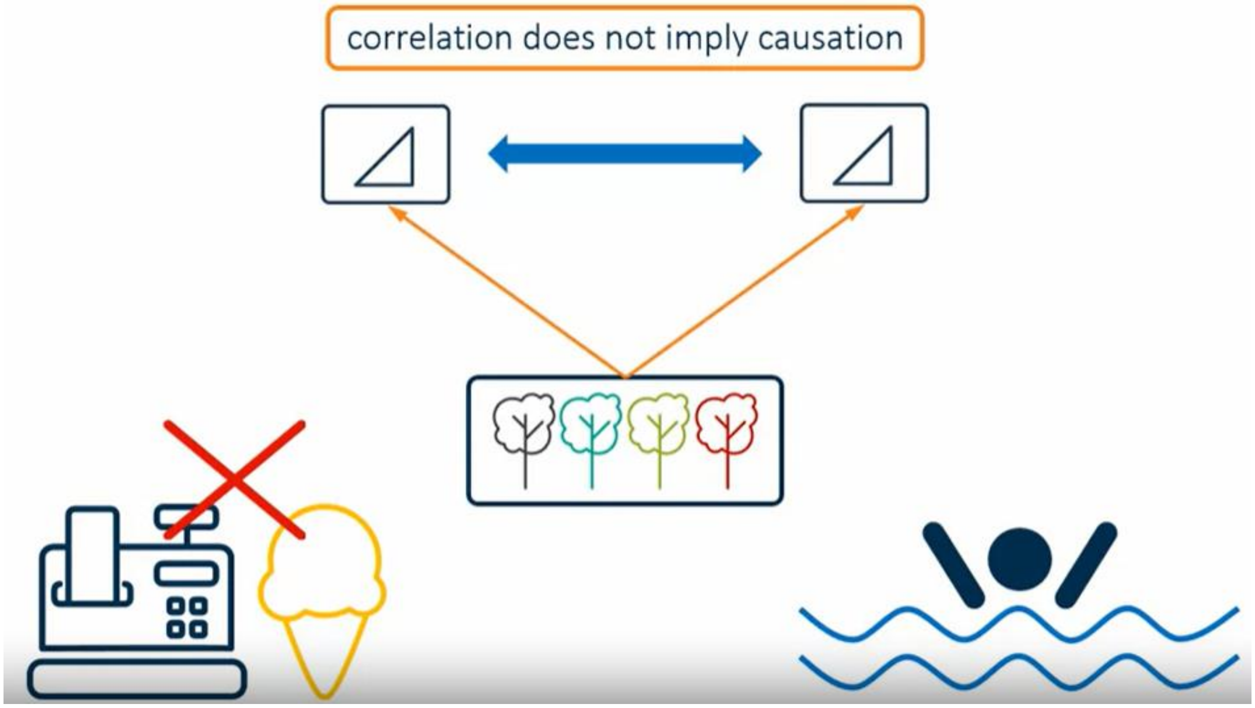
small *p*-values

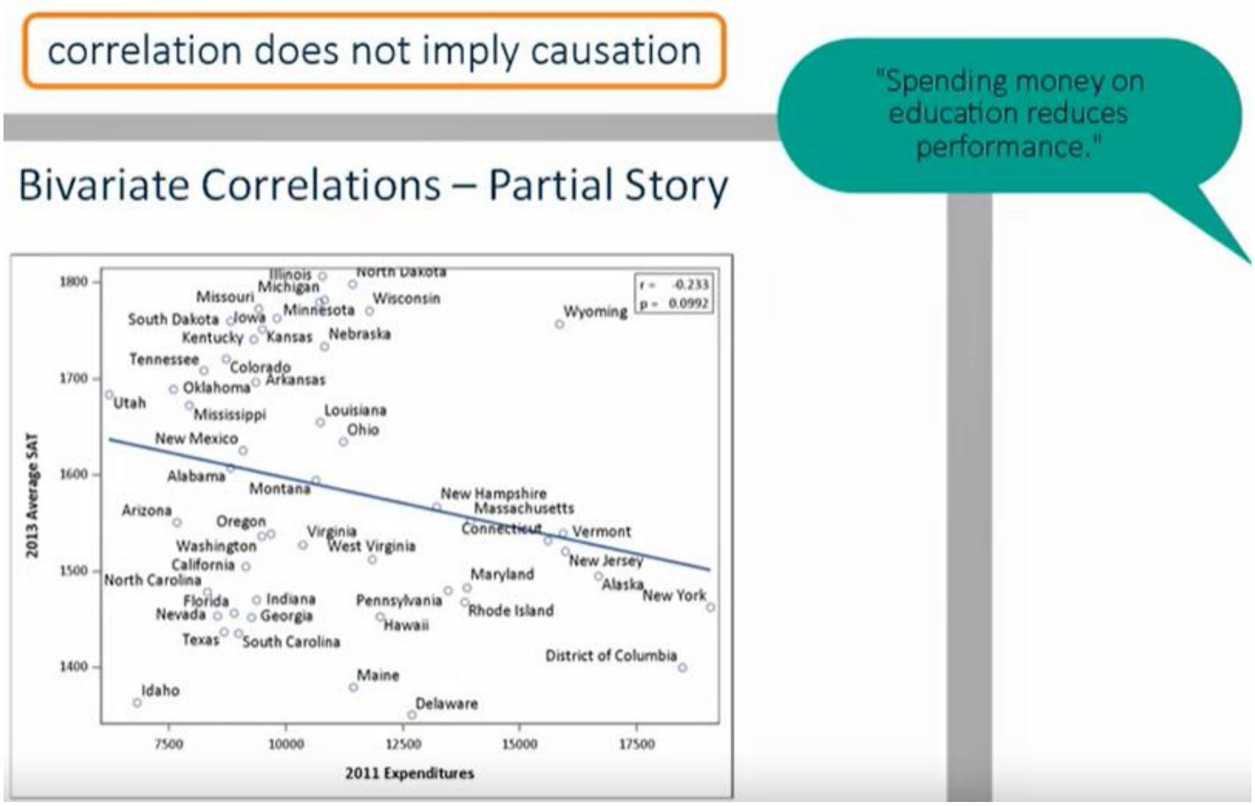


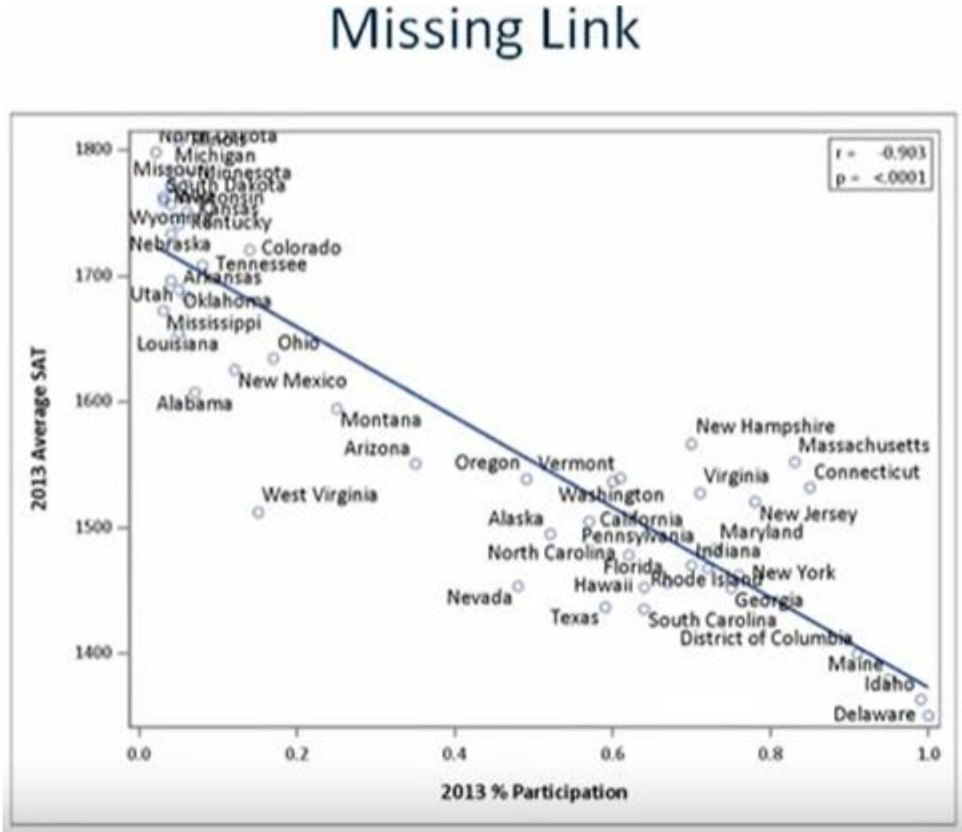
The Pearson correlation statistic is a measure of the linear relationship, or association, between two continuous variables. The closer the value is to -1, the stronger the negative linear relationship is between the two variables. The closer the value is to 0, the weaker the linear relationship. A correlation coefficient of 0 means that no linear relationship or association exists between the two variables.

Avoiding Common Errors When Interpreting Correlations

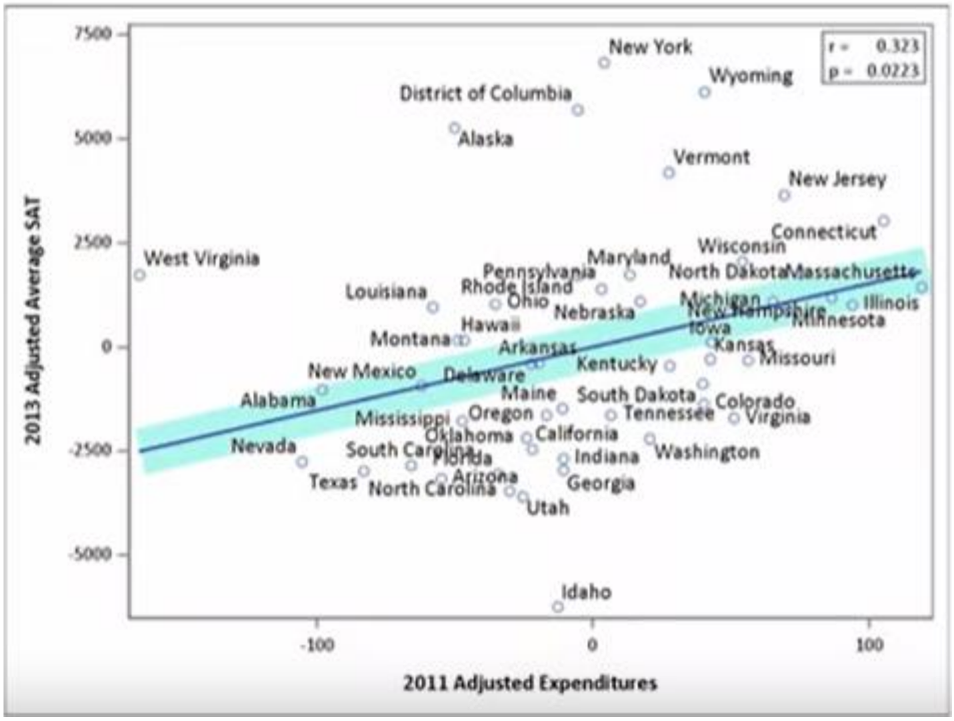

correlation does not imply causation

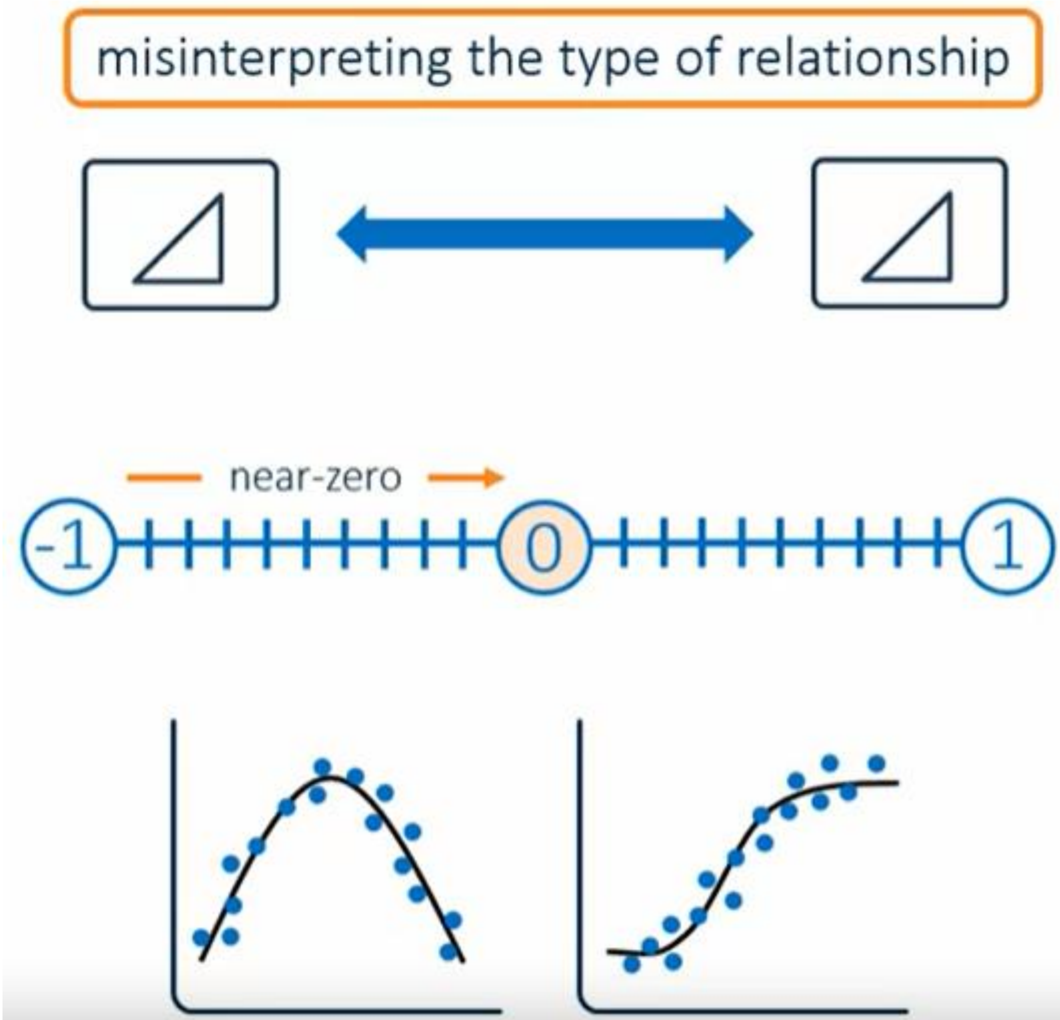

correlation does not imply causation

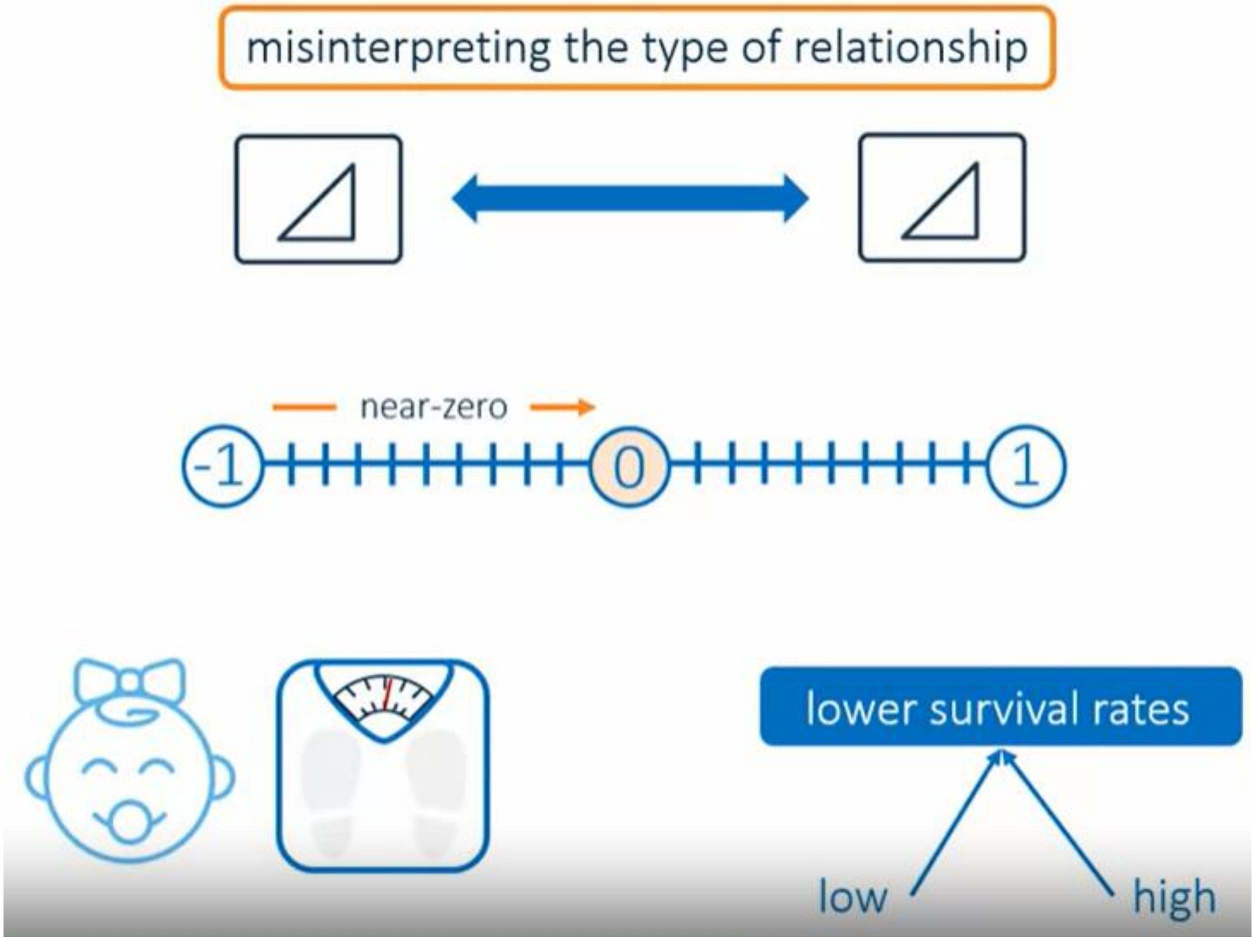

correlation does not imply causation
"Spending money on education reduces performance."
Bivariate Correlations – Partial Story
r = -0.233
p = 0.0992
2013 Average SAT
2011 Expenditures
1800
1700
1600
1500
1400
7500
10000
12500
15000
17500
Illinois
North Dakota
Michigan
Missouri
Wisconsin
Minnesota
South Dakota
Iowa
Wyoming
Kentucky
Kansas
Nebraska
Tennessee
Colorado
Arkansas
Oklahoma
Utah
Mississippi
Louisiana
Ohio
New Mexico
Alabama
Montana
New Hampshire
Massachusetts
Arizona
Oregon
Virginia
Connecticut
Vermont
Washington
West Virginia
California
New Jersey
North Carolina
Maryland
Alaska
New York
Florida
Indiana
Pennsylvania
Nevada
Georgia
Rhode Island
Hawaii
Texas
South Carolina
District of Columbia
Maine
Idaho
Delaware

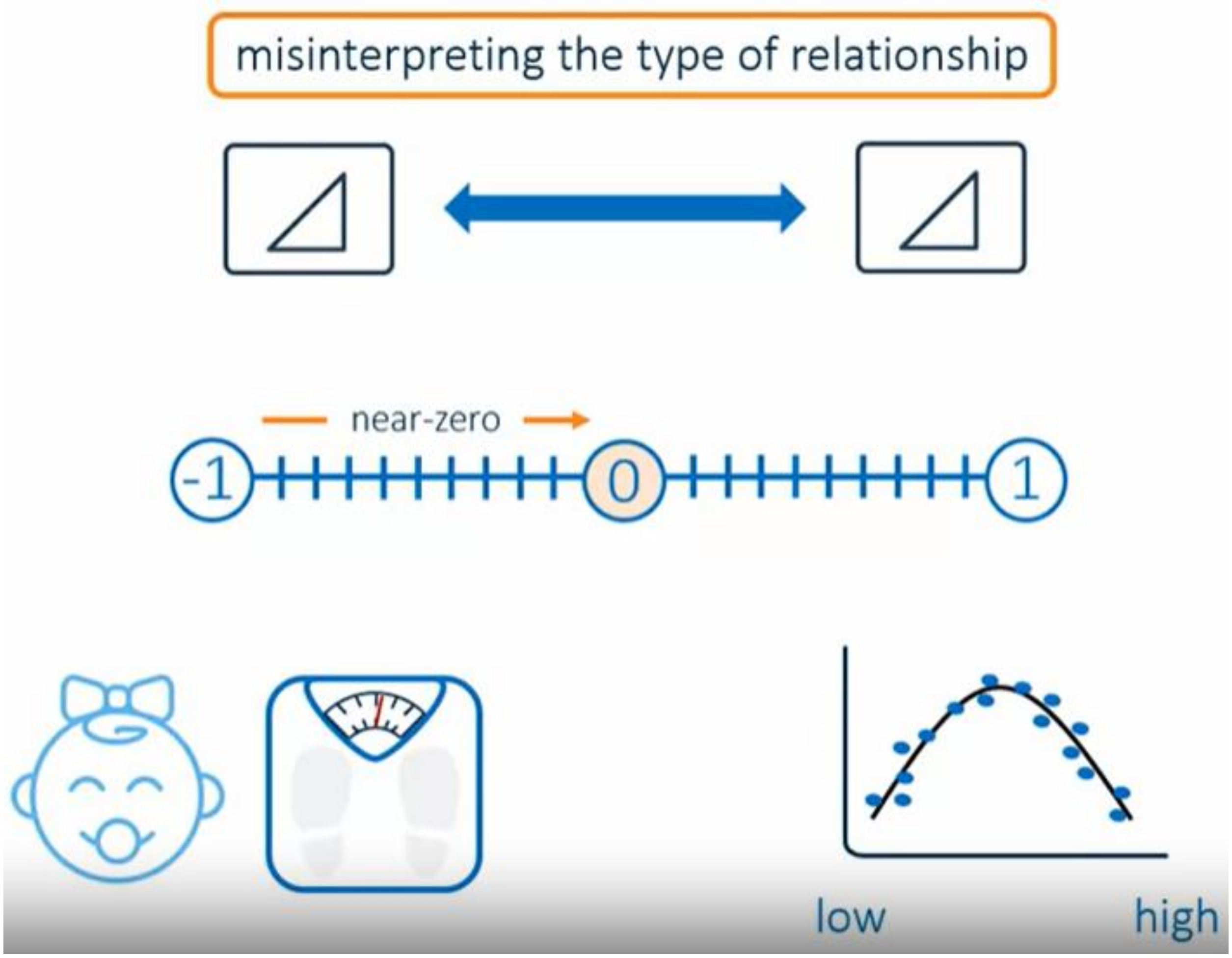

Missing Link
r = -0.903
p = <.0001
2013 Average SAT
2013 % Participation
1800
1700
1600
1500
1400
0.0
0.2
0.4
0.6
0.8
1.0
Michigan
Colorado
Tennessee
Nebraska
Utah
Mississippi
Louisiana
Ohio
New Mexico
Alabama
Montana
Arizona
West Virginia
Oregon
Vermont
Washington
Alaska
North Carolina
Nevada
Hawaii
Texas
South Carolina
District of Columbia
New Hampshire
Massachusetts
Connecticut
Virginia
New Jersey
Maryland
Indiana
New York
Georgia
Maine
Idaho
Delaware

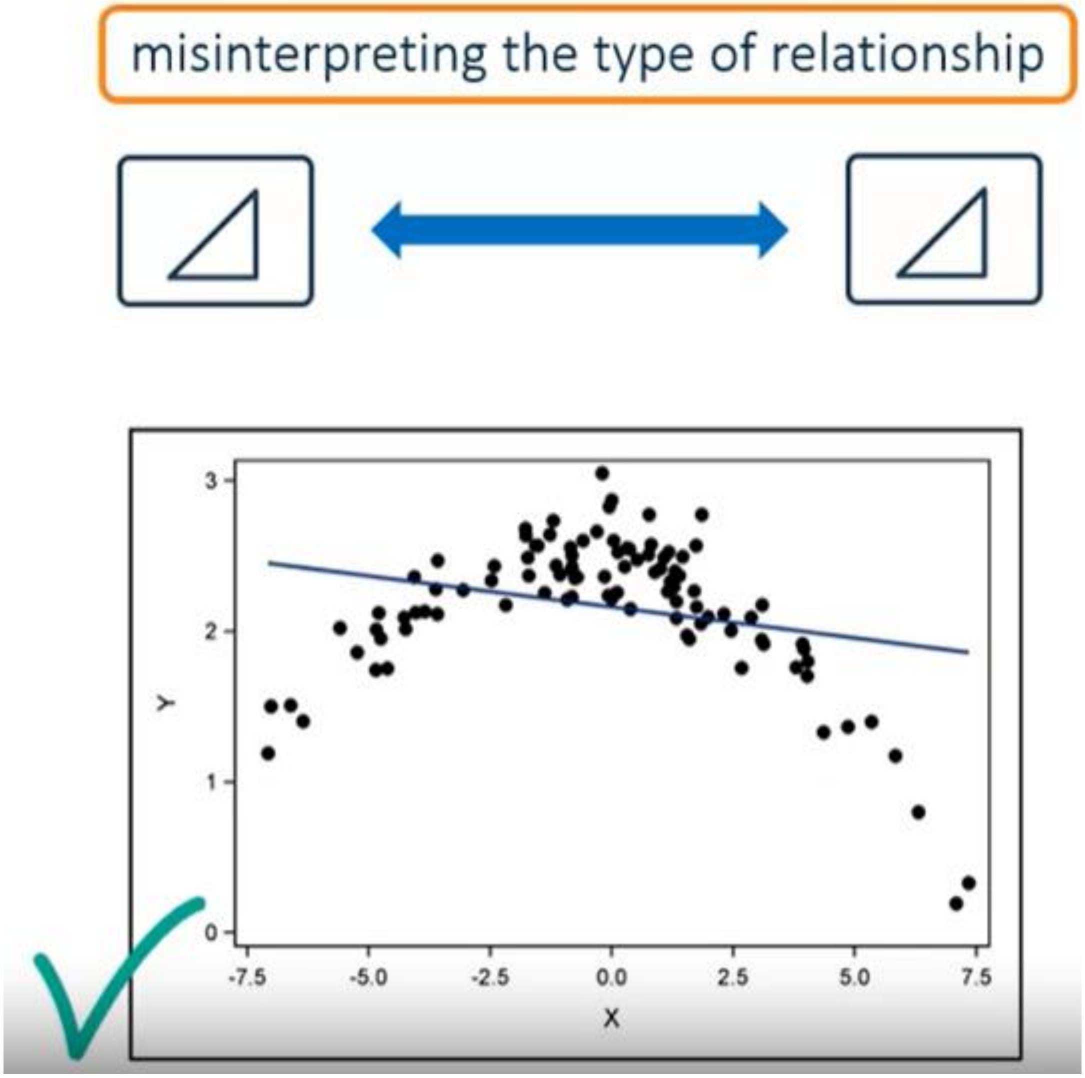
The Truer Story

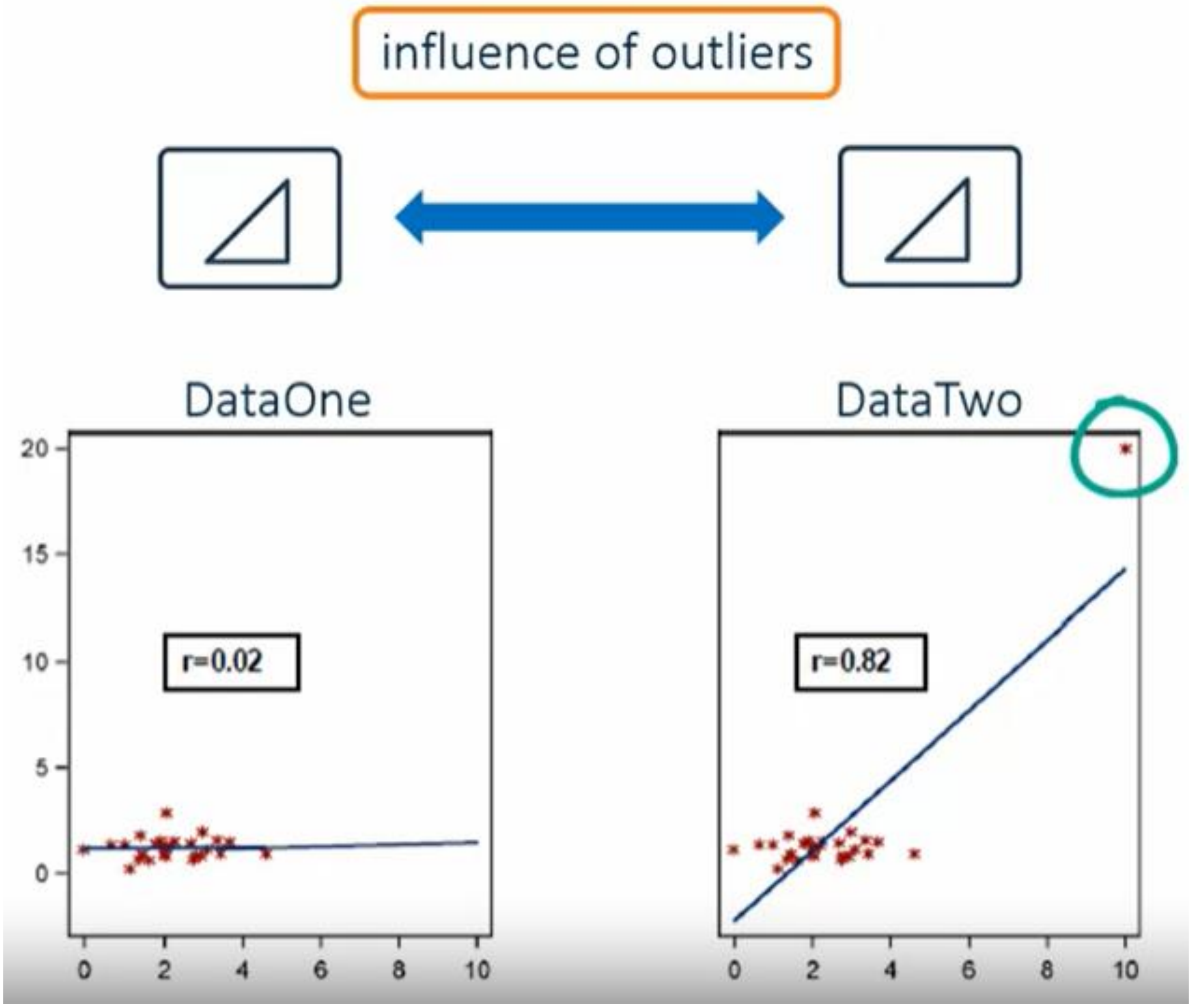


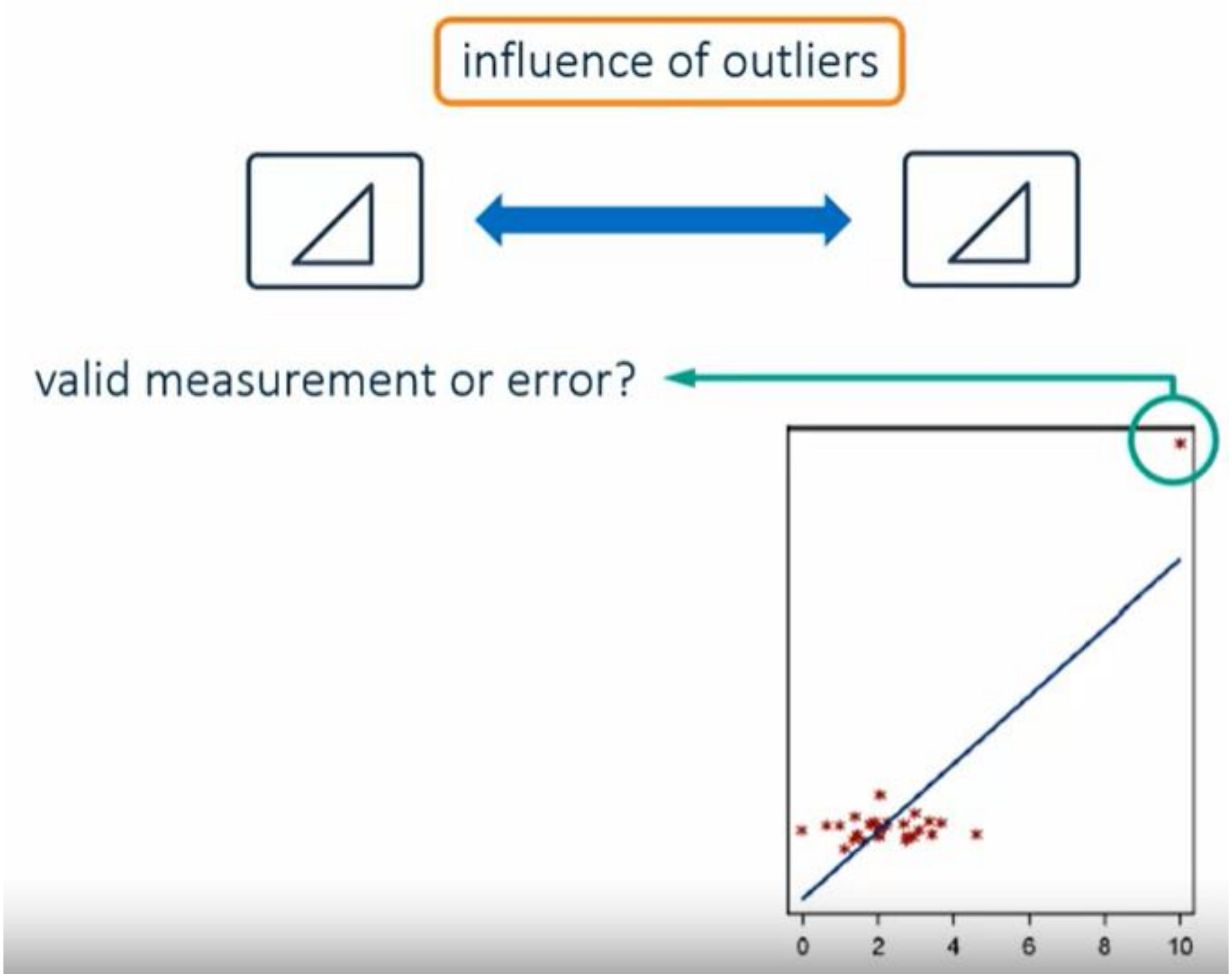

misinterpreting the type of relationship
near-zero
-1
0
1

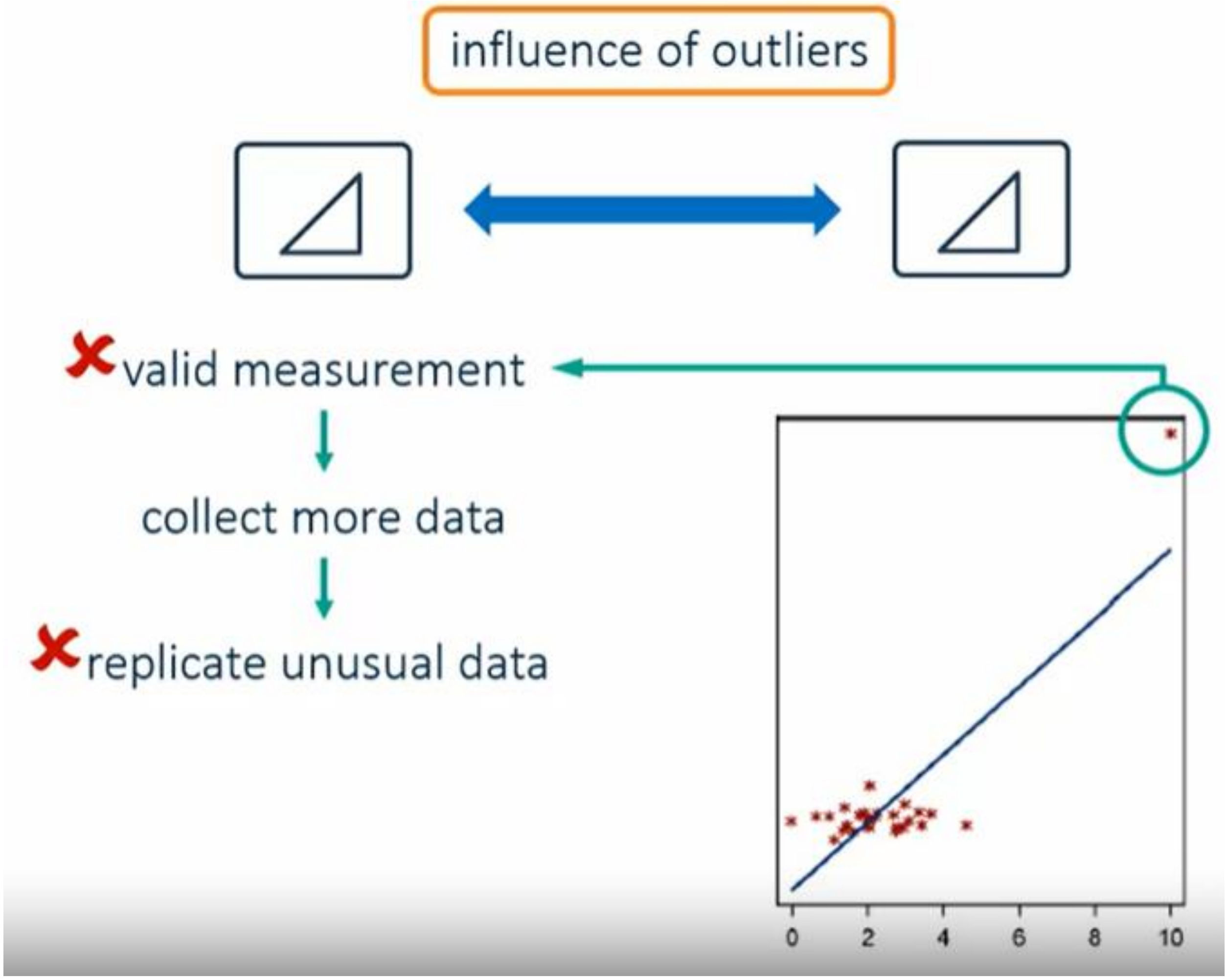

misinterpreting the type of relationship
near-zero
-1
0
1
lower survival rates
low
high

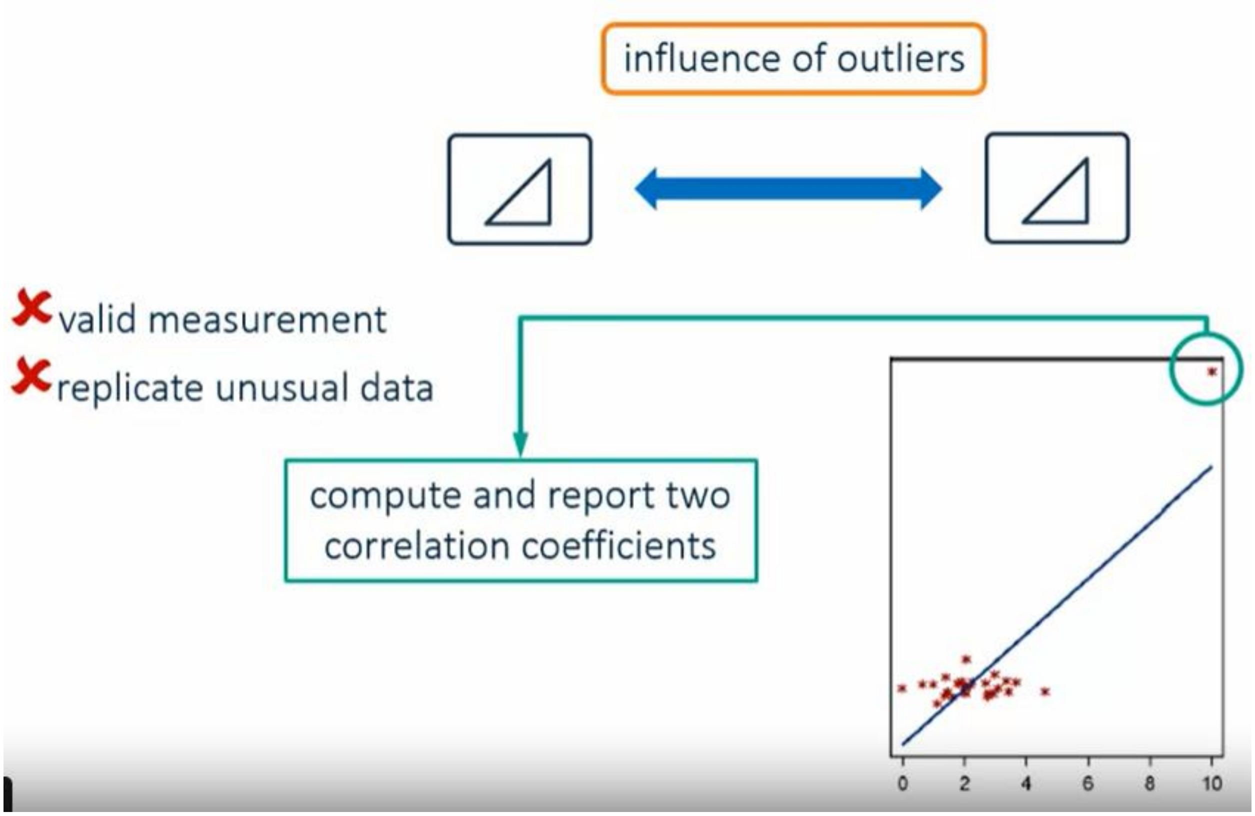

misinterpreting the type of relationship
near-zero
-1
0
1
low
high


misinterpreting the type of relationship
3
2
1
0
Y
-7.5
-5.0
-2.5
0.0
2.5
5.0
7.5
X

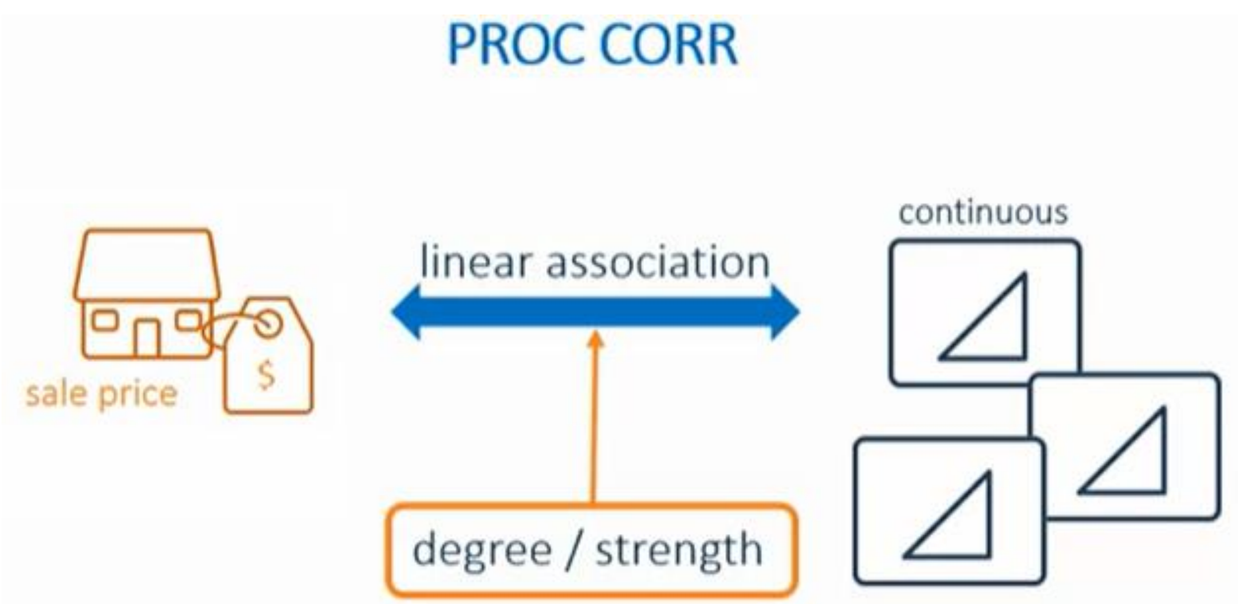

influence of outliers
DataOne
DataTwo
r=0.02
r=0.82
20
15
10
5
0
0
2
4
6
8
10

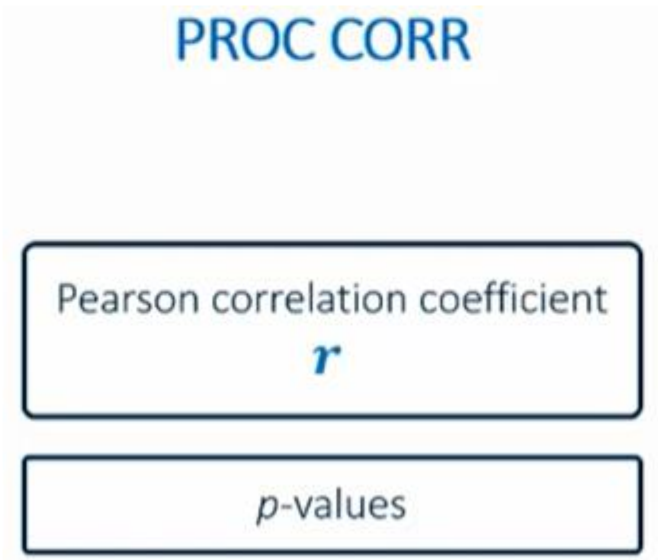

influence of outliers
valid measurement or error?
0
2
4
6
8
10


influence of outliers
valid measurement
collect more data
replicate unusual data
0
2
4
6
8
10



Demo Producing Correlation Statistics and Scatter Plots Using PROC CORR



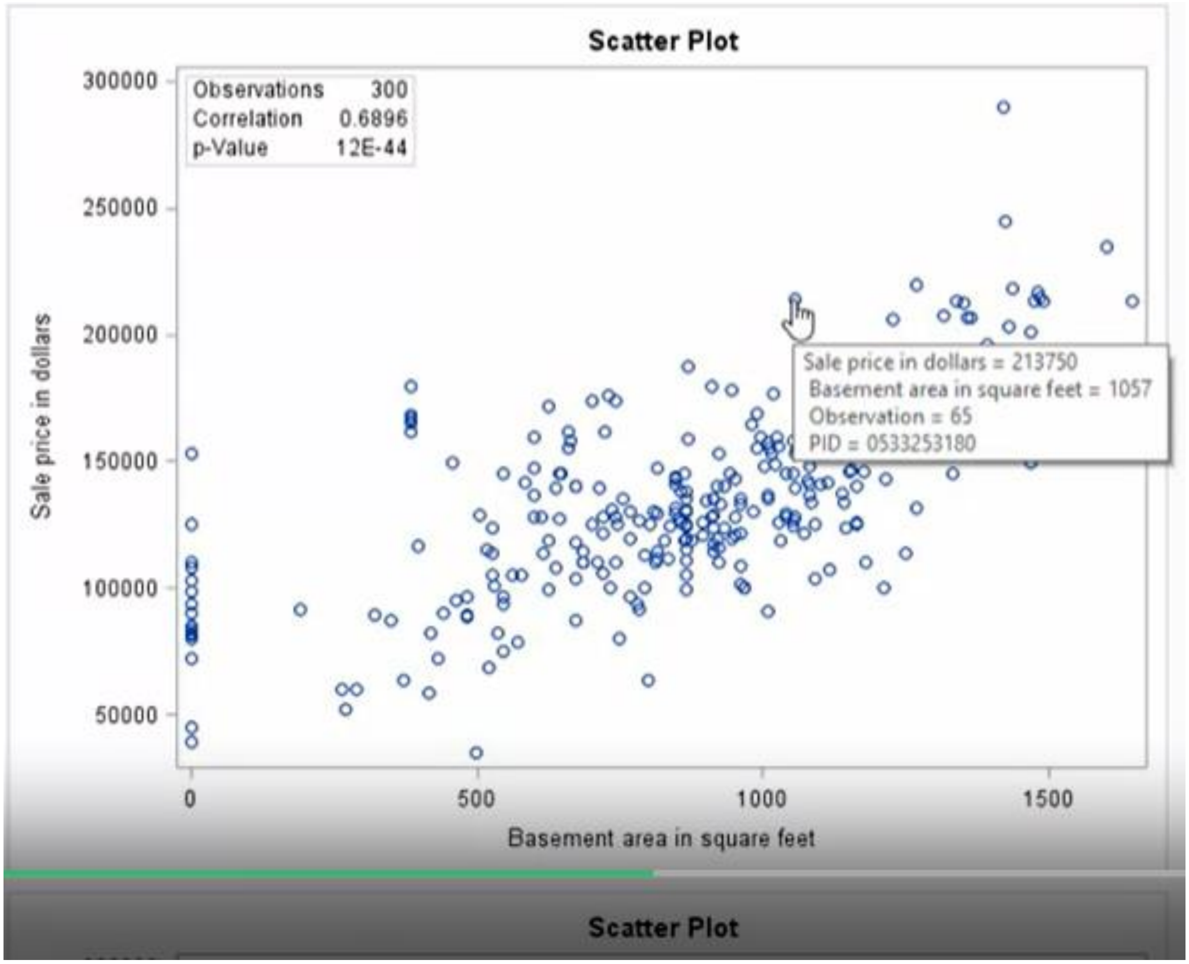


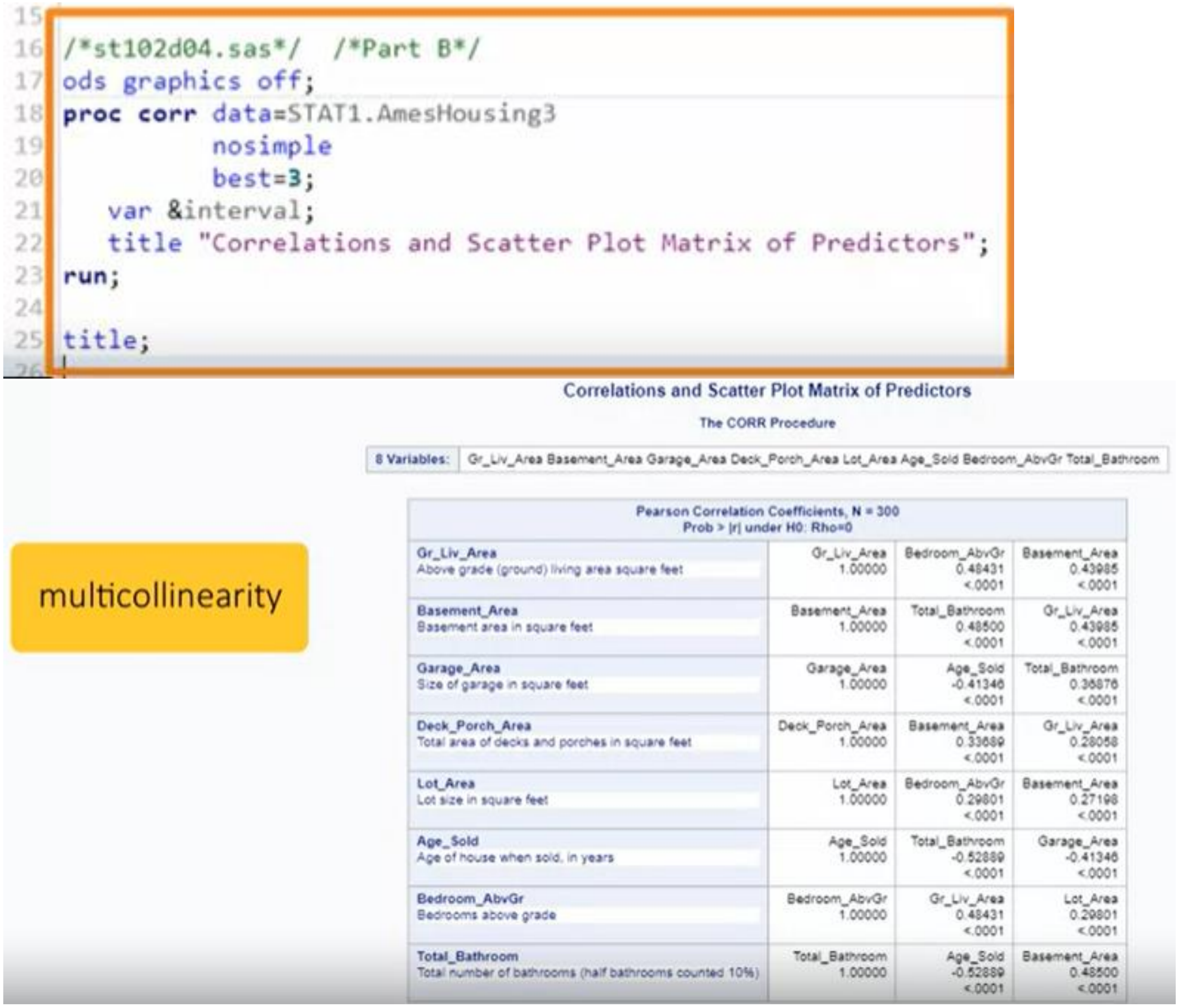

```
/*st102d04.sas*/  /*Part A*/
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
         Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom;

ods graphics / reset=all imagemap;
proc corr data=STAT1.AmesHousing3 rank
          plots(only)=scatter(nvar=all ellipse=none);
   var &interval;
   with SalePrice;
   id PID;
   title "Correlations and Scatter Plots with SalePrice";
run;

title;
```

PROC CORR DATA=*SAS-data-set <options>*;
    VAR *variables*;
    WITH *variables*;
    ID *variables*;
RUN;

Pearson Correlation Coefficients, N = 300
Prob > |r| under H0: Rho=0

	Basement_Area	Gr_Liv_Area	Age_Sold	Total_Bathroom	Garage_Area	Deck_Porch_Area	Lot_Area	Bedroom_AbvGr
SalePrice Sale price in dollars	0.68956 <.0001	0.65046 <.0001	-0.61542 <.0001	0.60043 <.0001	0.57892 <.0001	0.43989 <.0001	0.25335 <.0001	0.16594 0.0040





```
/*st102d04.sas*/  /*Part B*/
ods graphics off;
proc corr data=STAT1.AmesHousing3
          nosimple
          best=3;
   var &interval;
   title "Correlations and Scatter Plot Matrix of Predictors";
run;

title;
```

Correlations and Scatter Plot Matrix of Predictors

The CORR Procedure

8 Variables:	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom

Pearson Correlation Coefficients, N = 300 Prob > |r| under H0: Rho=0			
Gr_Liv_Area Above grade (ground) living area square feet	Gr_Liv_Area 1.00000	Bedroom_AbvGr 0.48431 <.0001	Basement_Area 0.43985 <.0001
Basement_Area Basement area in square feet	Basement_Area 1.00000	Total_Bathroom 0.48500 <.0001	Gr_Liv_Area 0.43985 <.0001
Garage_Area Size of garage in square feet	Garage_Area 1.00000	Age_Sold -0.41346 <.0001	Total_Bathroom 0.36876 <.0001
Deck_Porch_Area Total area of decks and porches in square feet	Deck_Porch_Area 1.00000	Basement_Area 0.33689 <.0001	Gr_Liv_Area 0.28058 <.0001
Lot_Area Lot size in square feet	Lot_Area 1.00000	Bedroom_AbvGr 0.29801 <.0001	Basement_Area 0.27198 <.0001
Age_Sold Age of house when sold, in years	Age_Sold 1.00000	Total_Bathroom -0.52889 <.0001	Garage_Area -0.41346 <.0001
Bedroom_AbvGr Bedrooms above grade	Bedroom_AbvGr 1.00000	Gr_Liv_Area 0.48431 <.0001	Lot_Area 0.29801 <.0001
Total_Bathroom Total number of bathrooms (half bathrooms counted 10%)	Total_Bathroom 1.00000	Age_Sold -0.52889 <.0001	Basement_Area 0.48500 <.0001

/*st102d04.sas*/ /*Part A*/

%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area

        Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom;

ods graphics / reset=all imagemap;

proc corr data=STAT1.AmesHousing3 rank

         plots(only)=scatter(nvar=all ellipse=none);

   var &interval;

   with SalePrice;

   id PID;

   title "Correlations and Scatter Plots with SalePrice";

```
run;

title;

/*st102d04.sas*/  /*Part B*/
ods graphics off;
proc corr data=STAT1.AmesHousing3
        nosimple
        best=3;
   var &interval;
   title "Correlations and Scatter Plot Matrix of Predictors";
run;

title;
```

Correlations and Scatter Plots with SalePrice

The CORR Procedure

1 With Variables:	SalePrice
8 Variables:	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom

Simple Statistics

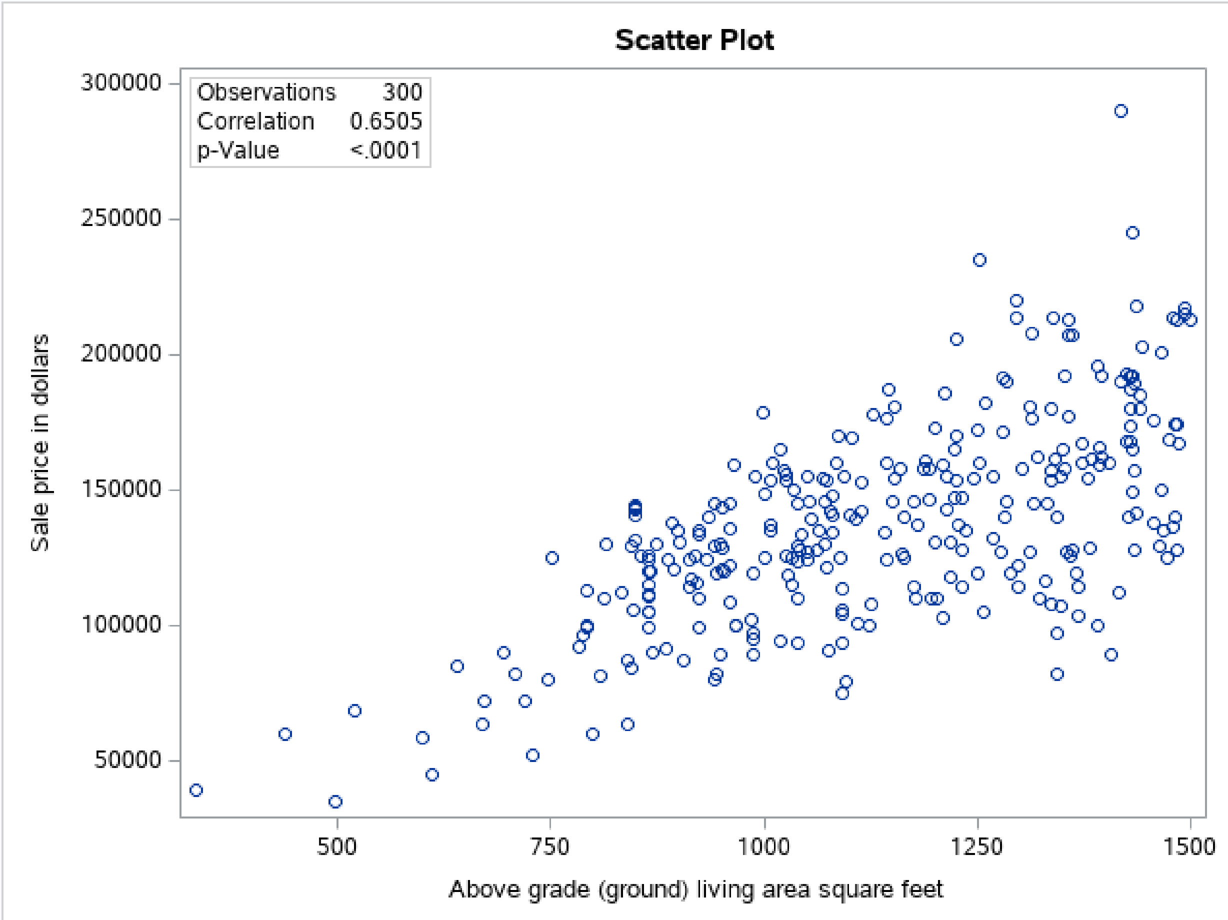
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
SalePrice	300	137525	37623	41257460	35000	290000	Sale price in dollars
Gr_Liv_Area	300	1131	232.64939	339222	334.00000	1500	Above grade (ground) living area square feet
Basement_Area	300	882.31000	359.78397	264693	0	1645	Basement area in square feet
Garage_Area	300	369.45333	176.25309	110836	0	902.00000	Size of garage in square feet
Deck_Porch_Area	300	118.26333	132.61169	35479	0	897.00000	Total area of decks and porches in square feet
Lot_Area	300	8294	3324	2488241	1495	26142	Lot size in square feet
Age_Sold	300	45.88667	27.47697	13766	1.00000	135.00000	Age of house when sold, in years
Bedroom_AbvGr	300	2.51333	0.69144	754.00000	0	4.00000	Bedrooms above grade
Total_Bathroom	300	1.70167	0.65707	510.50000	1.00000	4.10000	Total number of bathrooms (half bathrooms counted 10%)

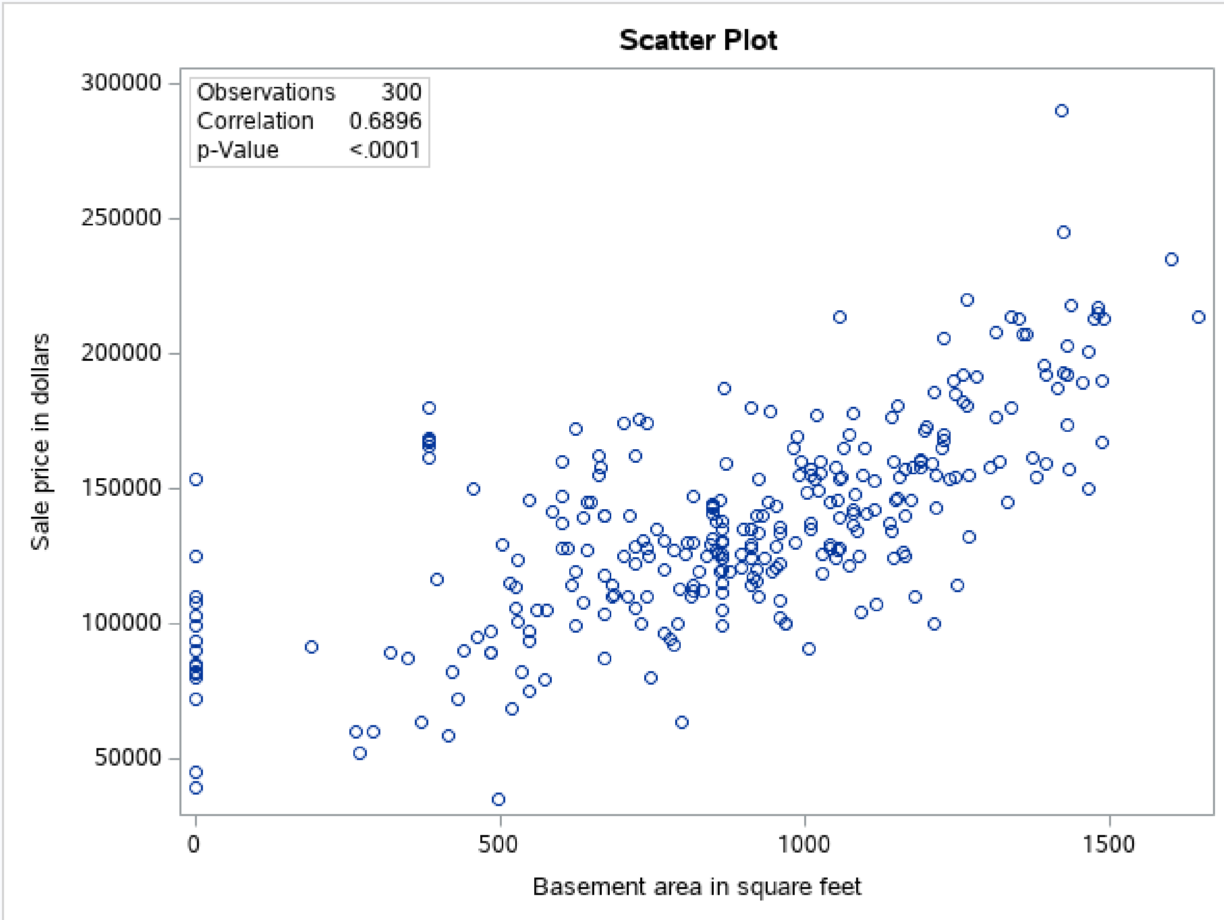
Pearson Correlation Coefficients, N = 300
Prob > \|r\| under H0: Rho=0

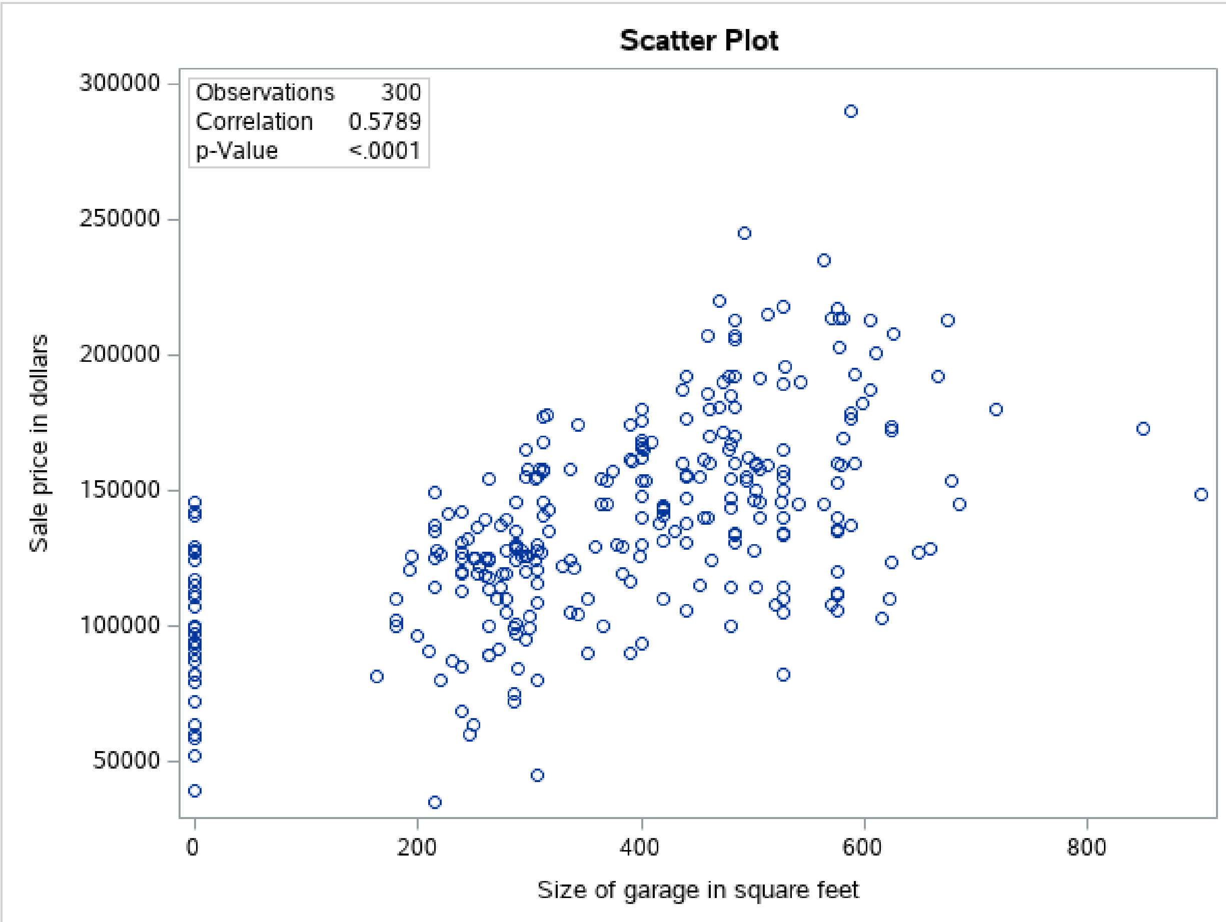
	Basement_Area	Gr_Liv_Area	Age_Sold	Total_Bathroom	Garage_Area	Deck_Porch_Area	Lot_Area	Bedroom_AbvGr
SalePrice Sale price in dollars	0.68956 <.0001	0.65046 <.0001	-0.61542 <.0001	0.60043 <.0001	0.57892 <.0001	0.43989 <.0001	0.25335 <.0001	0.16594 0.0040

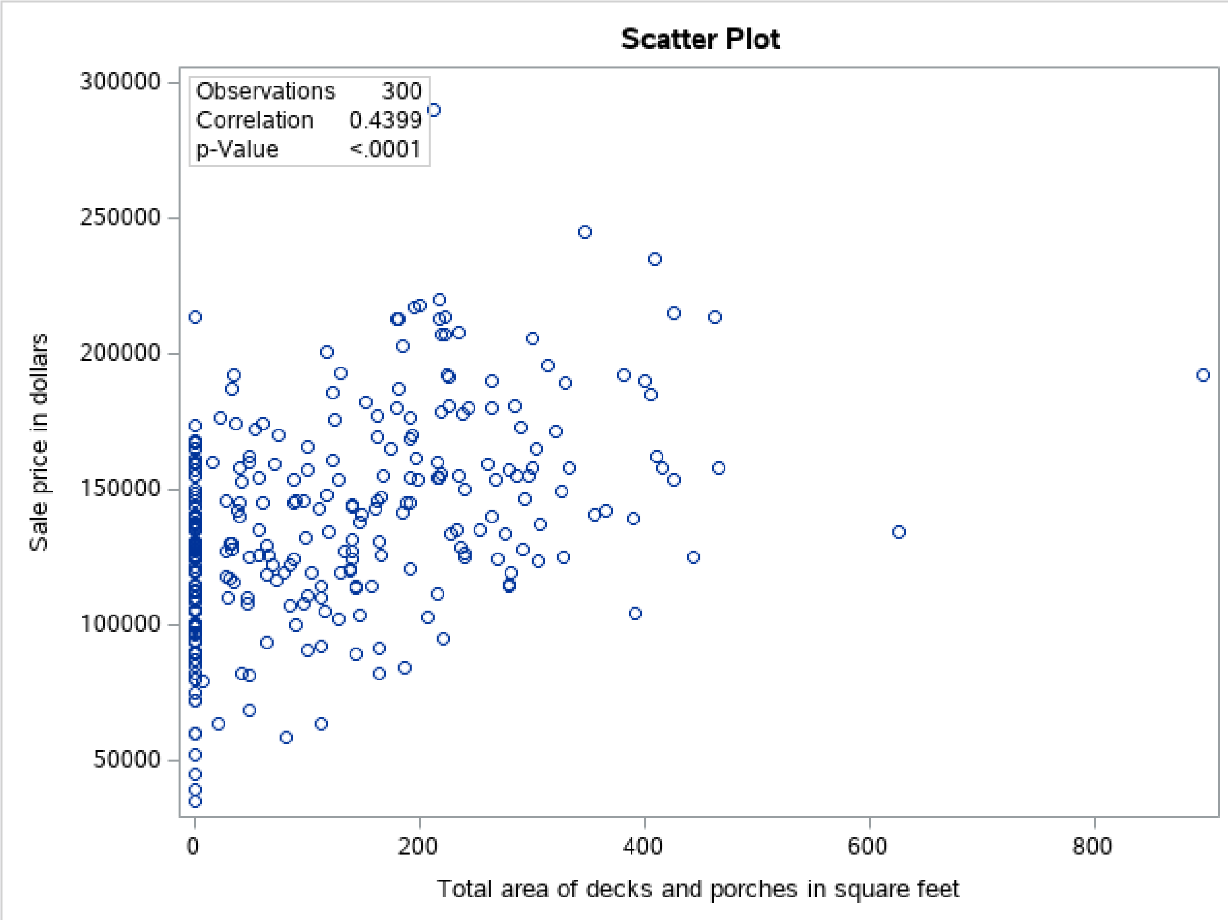
Correlations and Scatter Plots with SalePrice

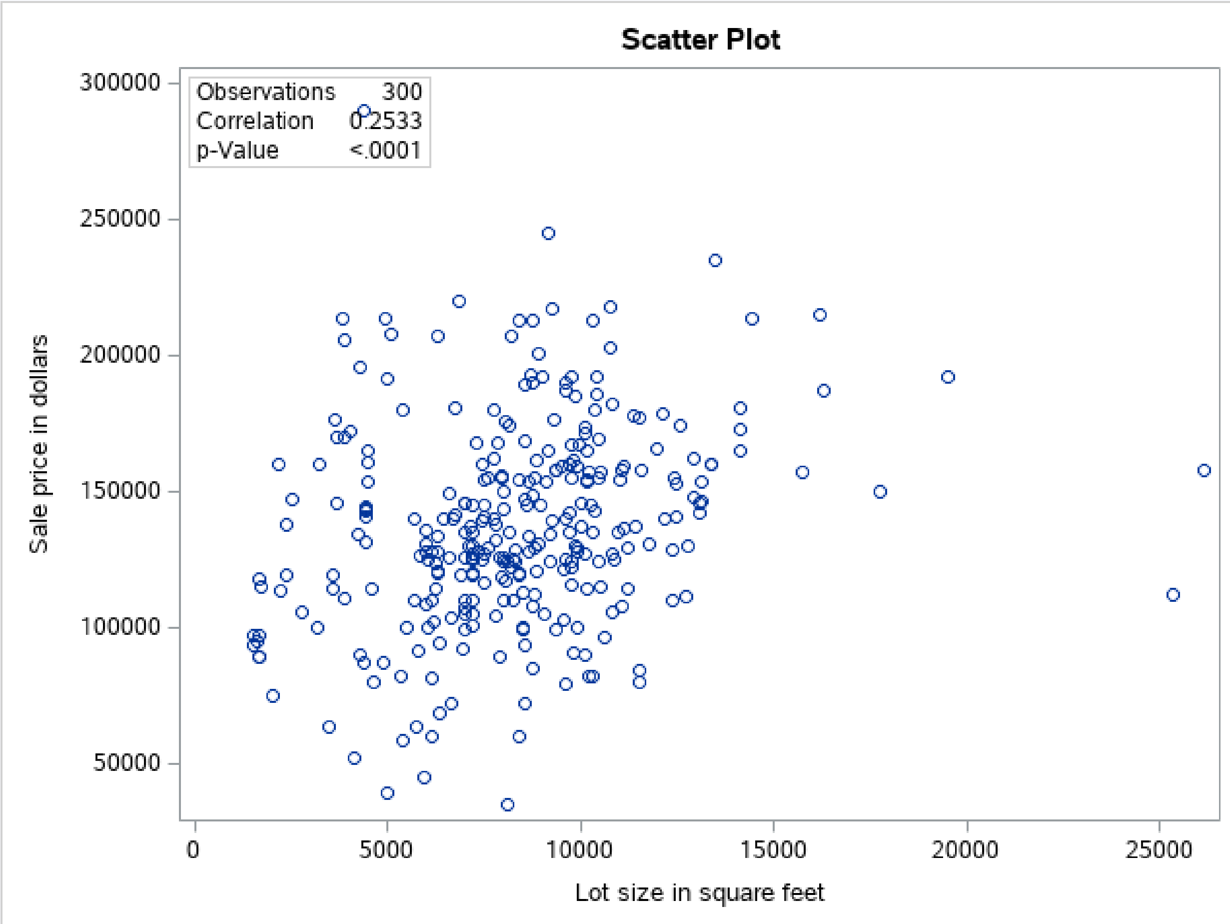
The CORR Procedure

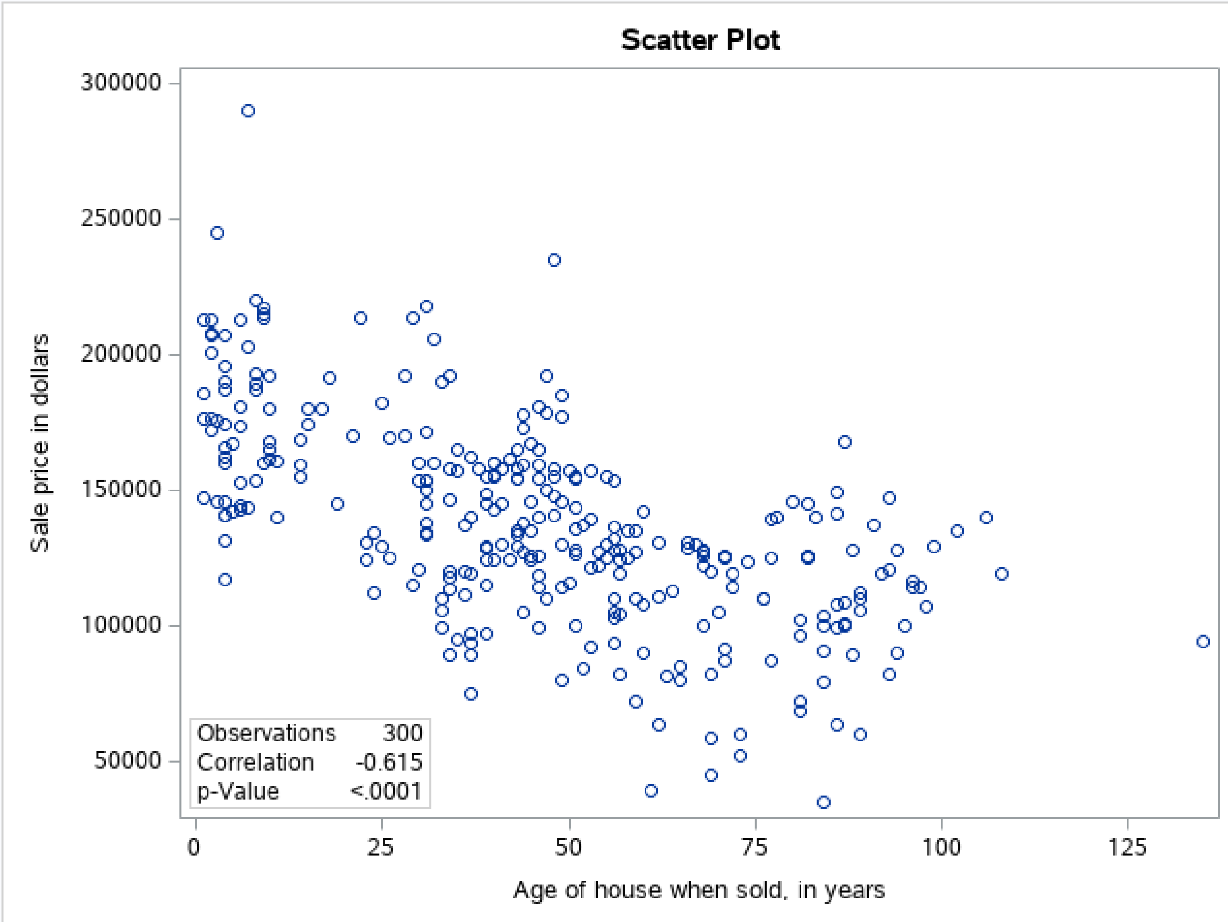


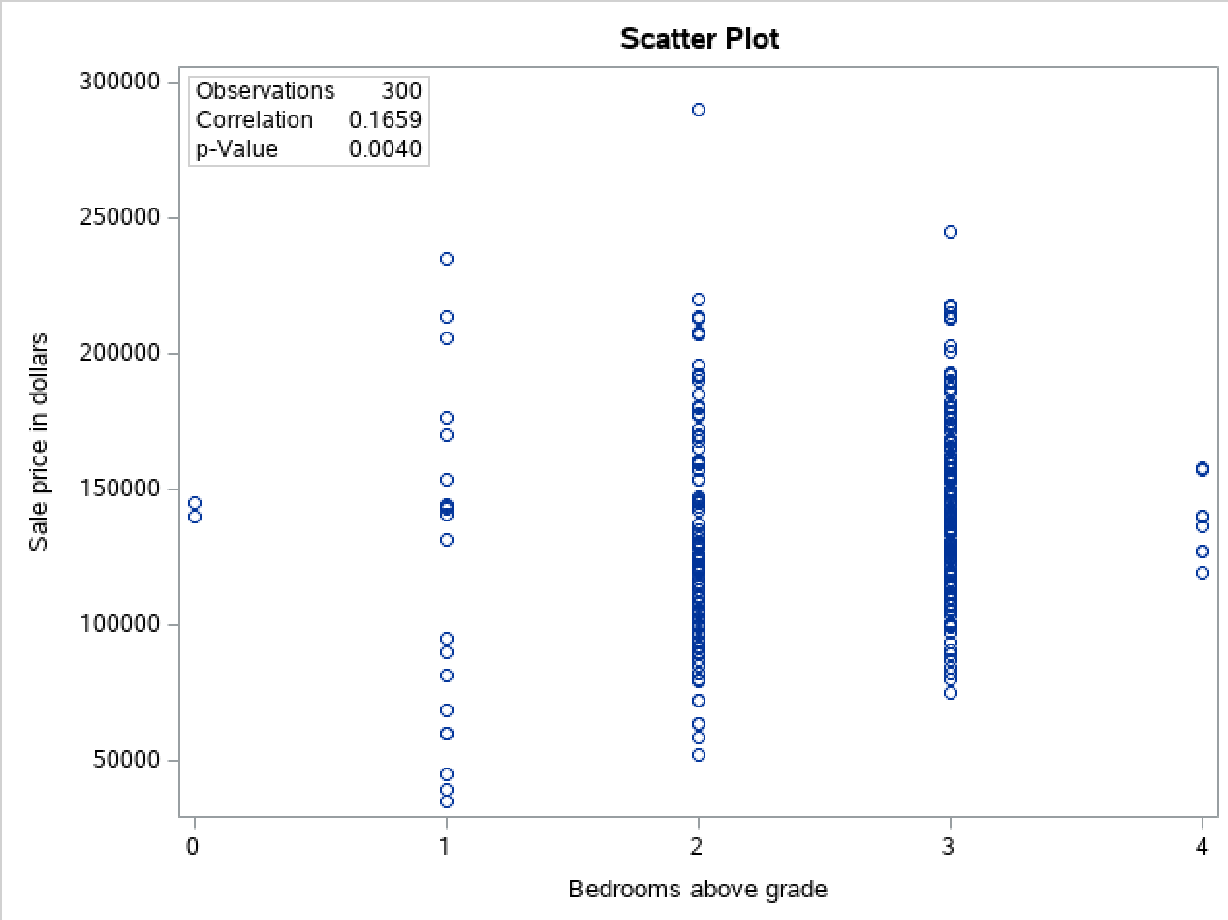

Scatter Plot
Observations 300
Correlation 0.6896
p-Value <.0001
300000
250000
200000
150000
100000
50000
Sale price in dollars
0
500
1000
1500
Basement area in square feet

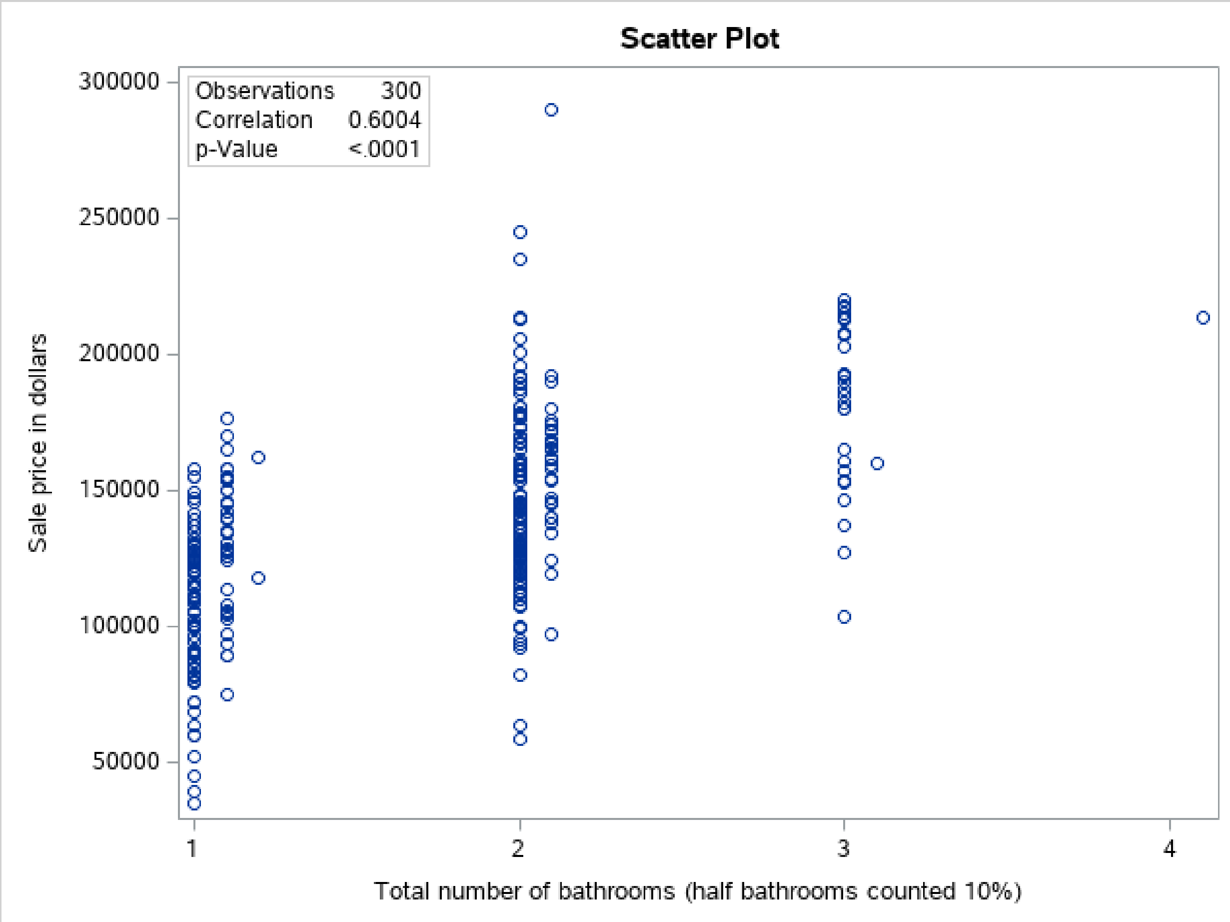

Scatter Plot
Observations 300
Correlation 0.5789
p-Value <.0001
300000
250000
200000
150000
100000
50000
Sale price in dollars
0
200
400
600
800
Size of garage in square feet


Scatter Plot
Observations 300
Correlation 0.4399
p-Value <.0001
300000
250000
200000
150000
100000
50000
Sale price in dollars
0
200
400
600
800
Total area of decks and porches in square feet


Scatter Plot
Observations 300
Correlation 0.2533
p-Value <.0001
Sale price in dollars
300000
250000
200000
150000
100000
50000
0
5000
10000
15000
20000
25000
Lot size in square feet


Scatter Plot
300000
250000
200000
150000
100000
50000
Sale price in dollars
Observations 300
Correlation -0.615
p-Value <.0001
0
25
50
75
100
125
Age of house when sold, in years


Scatter Plot
Observations 300
Correlation 0.1659
p-Value 0.0040
300000
250000
200000
150000
100000
50000
Sale price in dollars
0
1
2
3
4
Bedrooms above grade


Scatter Plot
Observations 300
Correlation 0.6004
p-Value <.0001
Sale price in dollars
300000
250000
200000
150000
100000
50000
1
2
3
4
Total number of bathrooms (half bathrooms counted 10%)

Correlations and Scatter Plot Matrix of Predictors

The CORR Procedure

8 Variables:	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom

Pearson Correlation Coefficients, N = 300 Prob > |r| under H0: Rho=0			
Gr_Liv_Area Above grade (ground) living area square feet	Gr_Liv_Area 1.00000	Bedroom_AbvGr 0.48431 <.0001	Basement_Area 0.43985 <.0001
Basement_Area Basement area in square feet	Basement_Area 1.00000	Total_Bathroom 0.48500 <.0001	Gr_Liv_Area 0.43985 <.0001
Garage_Area Size of garage in square feet	Garage_Area 1.00000	Age_Sold -0.41346 <.0001	Total_Bathroom 0.36876 <.0001
Deck_Porch_Area Total area of decks and porches in square feet	Deck_Porch_Area 1.00000	Basement_Area 0.33689 <.0001	Gr_Liv_Area 0.28058 <.0001
Lot_Area Lot size in square feet	Lot_Area 1.00000	Bedroom_AbvGr 0.29801 <.0001	Basement_Area 0.27198 <.0001
Age_Sold Age of house when sold, in years	Age_Sold 1.00000	Total_Bathroom -0.52889 <.0001	Garage_Area -0.41346 <.0001
Bedroom_AbvGr Bedrooms above grade	Bedroom_AbvGr 1.00000	Gr_Liv_Area 0.48431 <.0001	Lot_Area 0.29801 <.0001
Total_Bathroom Total number of bathrooms (half bathrooms counted 10%)	Total_Bathroom 1.00000	Age_Sold -0.52889 <.0001	Basement_Area 0.48500 <.0001

Correlation Analysis and Model Building

Correlations between the response variable and potential predictors can be useful by suggesting variables that should be included or excluded from model building. Often, modelers have many predictors, and thus, a very large number of possible models to explore. Predictors with a weak or no relationship with the response variable might sometimes be excluded. Typically, the decision to throw out a variable is based on multivariable analyses. However, if the modeler has far more predictors than can be used and variable reduction becomes necessary (often under time pressure), predictors with weak or no correlation with the response variable are good candidates for exclusion. Part of correlation analysis involves visually assessing associations between variables by looking at scatter plots. When these plots reveal patterns in the data, such as curvilinear relationships, a modeler might need to build additional terms into the model, such as polynomials. Another reason to create scatter plots is to assess the linear relationship between pairs of predictor variables. When predictors are highly correlated, they provide redundant information. Multicollinearity (strong correlations among sets of predictors) can destabilize parameter estimates and degrade the ability of model selection routines, such as stepwise selection, to select good variables. Correlation analysis is one of several ways to address collinearity prior to model building.

Practice - Describing the Relationship between Continuous Variables

TOTAL POINTS 4

1.

Question 1

The percentage of body fat, age, weight, height, and 10 body circumference measurements (for example, abdomen) were recorded for 252 men by Dr. Roger W. Johnson of Calvin College in Minnesota. The data are in the **stat1.bodyfat2** data set. Body fat, one measure of health, has been accurately estimated by a water displacement measurement technique.

1. Generate scatter plots and correlations for the VAR variables **Age**, **Weight**, and **Height**, and the circumference measures **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist** versus the WITH variable, **PctBodyFat2**. ****IMPORTANT**: For PROC CORR, ODS Graphics will display a maximum of 10 VAR variable plots at a time. This practice analyzes thirteen variables, so it requires two PROC CORR steps to generate all thriteen plots. This limitation only applies to the ODS graphics. The correlation table displays all variables in the VAR statement by default.
2. Write a PROC CORR step to analyze all thirteen variables (**Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist**). This will generate a correlation table for all of the variables, but it will display plots for only the first ten.
3. Write an ODS statement to limit the graphic output to scatter plots.
4. Write another PROC CORR step, to look at only the last three variables, **Biceps**, **Forearm**, and **Wrist**.
5. Submit the code. The output should include a correlation table for all thirteen variables followed by a plots for the first ten, and then plots for the last three.
6. Examine the plots. Can straight lines adequately describe the relationships?

```
/*st102s03.sas*/  /*Part A*/

%let interval=Age Weight Height Neck Chest Abdomen Hip

          Thigh Knee Ankle Biceps Forearm Wrist;


ods graphics / reset=all imagemap;

proc corr data=STAT1.BodyFat2

        plots(only)=scatter(nvar=all ellipse=none);

  var &interval;
```

```
  with PctBodyFat2;
  id Case;
  title "Correlations and Scatter Plots";
run;

%let interval=Biceps Forearm Wrist;

ods graphics / reset=all imagemap;
ods select scatterplot;
proc corr data=STAT1.BodyFat2
        plots(only)=scatter(nvar=all ellipse=none);
  var &interval;
  with PctBodyFat2;
  id Case;
  title "Correlations and Scatter Plots";
run;
```

Correlations and Scatter Plots

The CORR Procedure

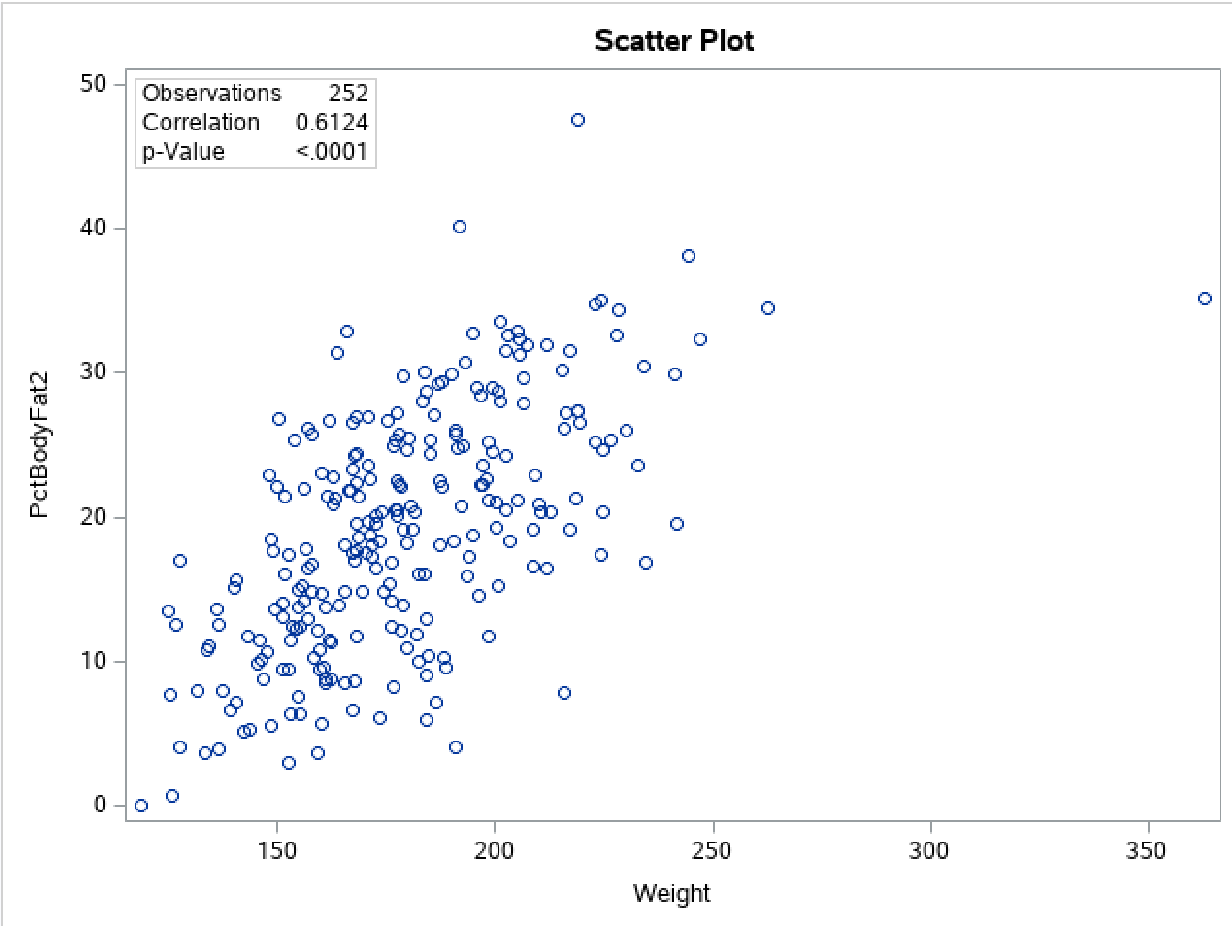
1 With Variables:	PctBodyFat2
13 Variables:	Age Weight Height Neck Chest Abdomen Hip Thigh Knee Ankle Biceps Forearm Wrist

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
PctBodyFat2	252	19.15079	8.36874	4826	0	47.50000
Age	252	44.88492	12.60204	11311	22.00000	81.00000
Weight	252	178.92440	29.38916	45089	118.50000	363.15000
Height	252	70.30754	2.60958	17718	64.00000	77.75000
Neck	252	37.99206	2.43091	9574	31.10000	51.20000
Chest	252	100.82421	8.43048	25408	79.30000	136.20000
Abdomen	252	92.55595	10.78308	23324	69.40000	148.10000
Hip	252	99.90476	7.16406	25176	85.00000	147.70000
Thigh	252	59.40595	5.24995	14970	47.20000	87.30000
Knee	252	38.59048	2.41180	9725	33.00000	49.10000
Ankle	252	23.10238	1.69489	5822	19.10000	33.90000
Biceps	252	32.27341	3.02127	8133	24.80000	45.00000
Forearm	252	28.66389	2.02069	7223	21.00000	34.90000
Wrist	252	18.22976	0.93358	4594	15.80000	21.40000

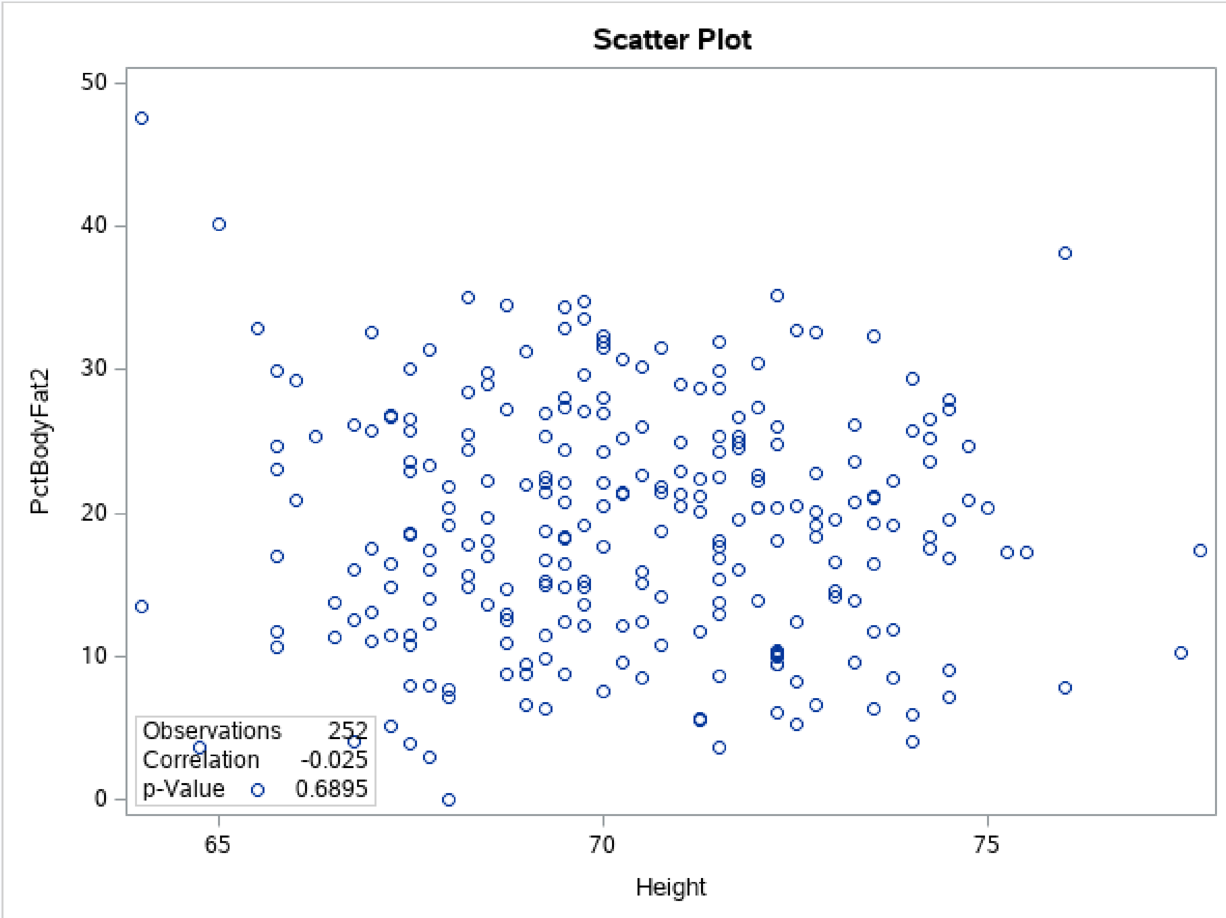
Pearson Correlation Coefficients, N = 252 Prob > |r| under H0: Rho=0													
	Age	Weight	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
PctBodyFat2	0.29146 <.0001	0.61241 <.0001	-0.02529 0.6895	0.49059 <.0001	0.70262 <.0001	0.81343 <.0001	0.62520 <.0001	0.55961 <.0001	0.50867 <.0001	0.26597 <.0001	0.49327 <.0001	0.36139 <.0001	0.34657 <.0001

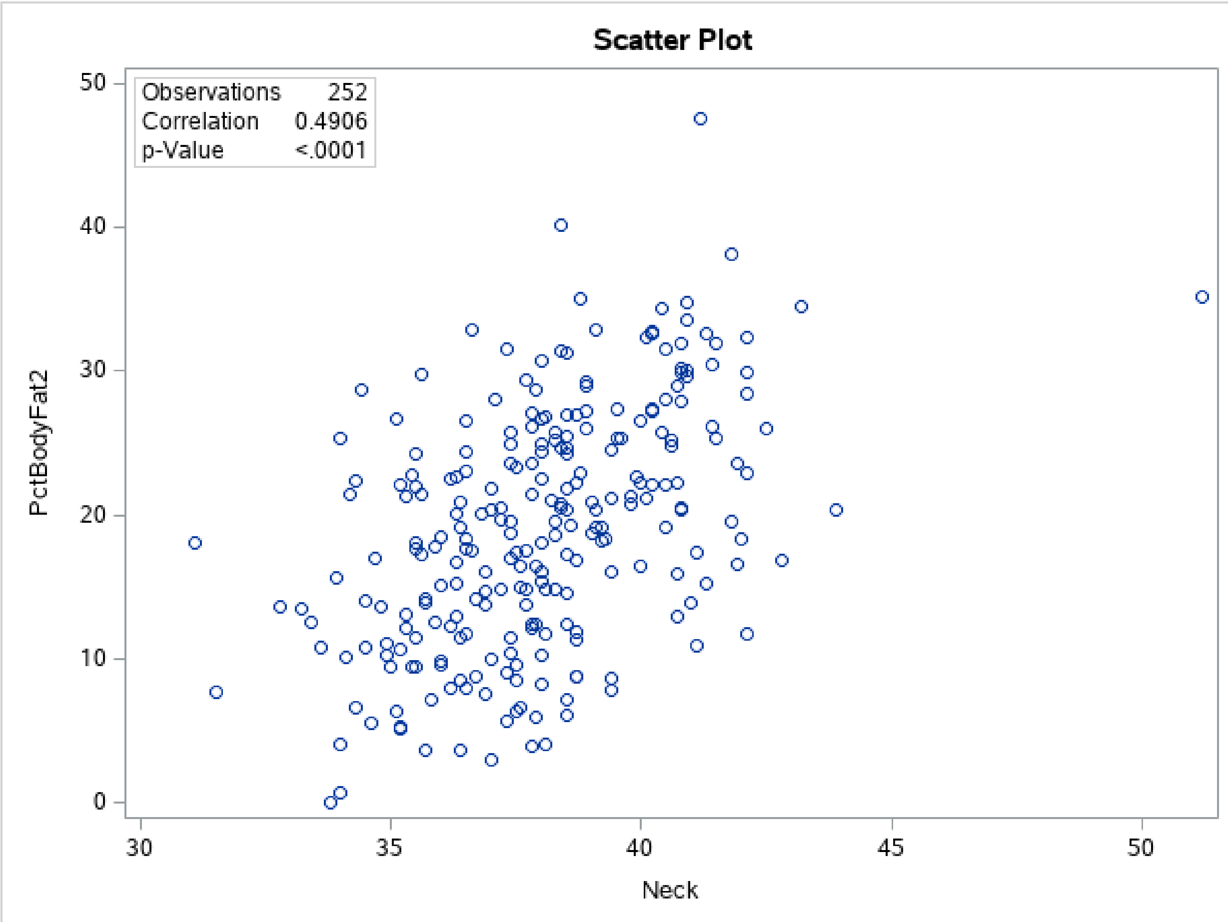
Correlations and Scatter Plots

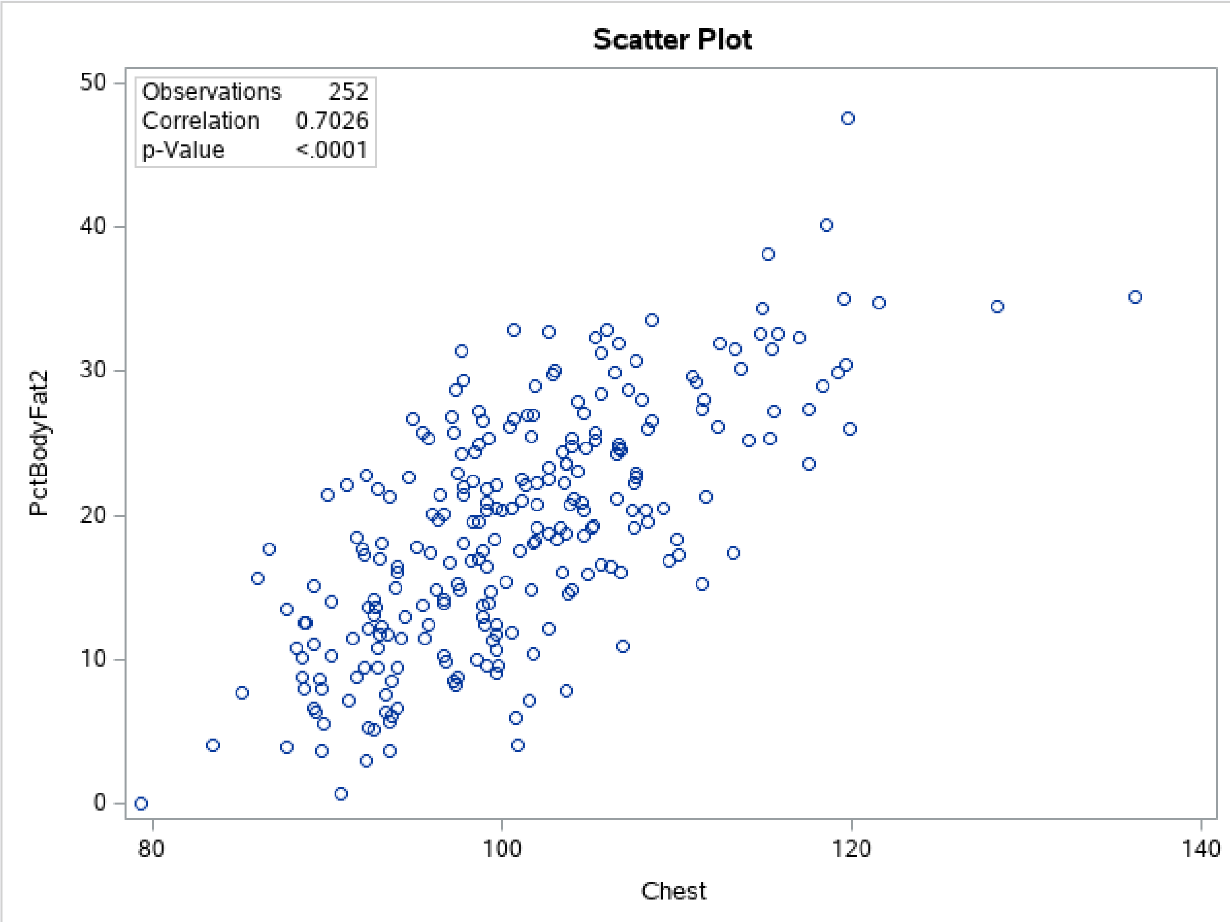
The CORR Procedure

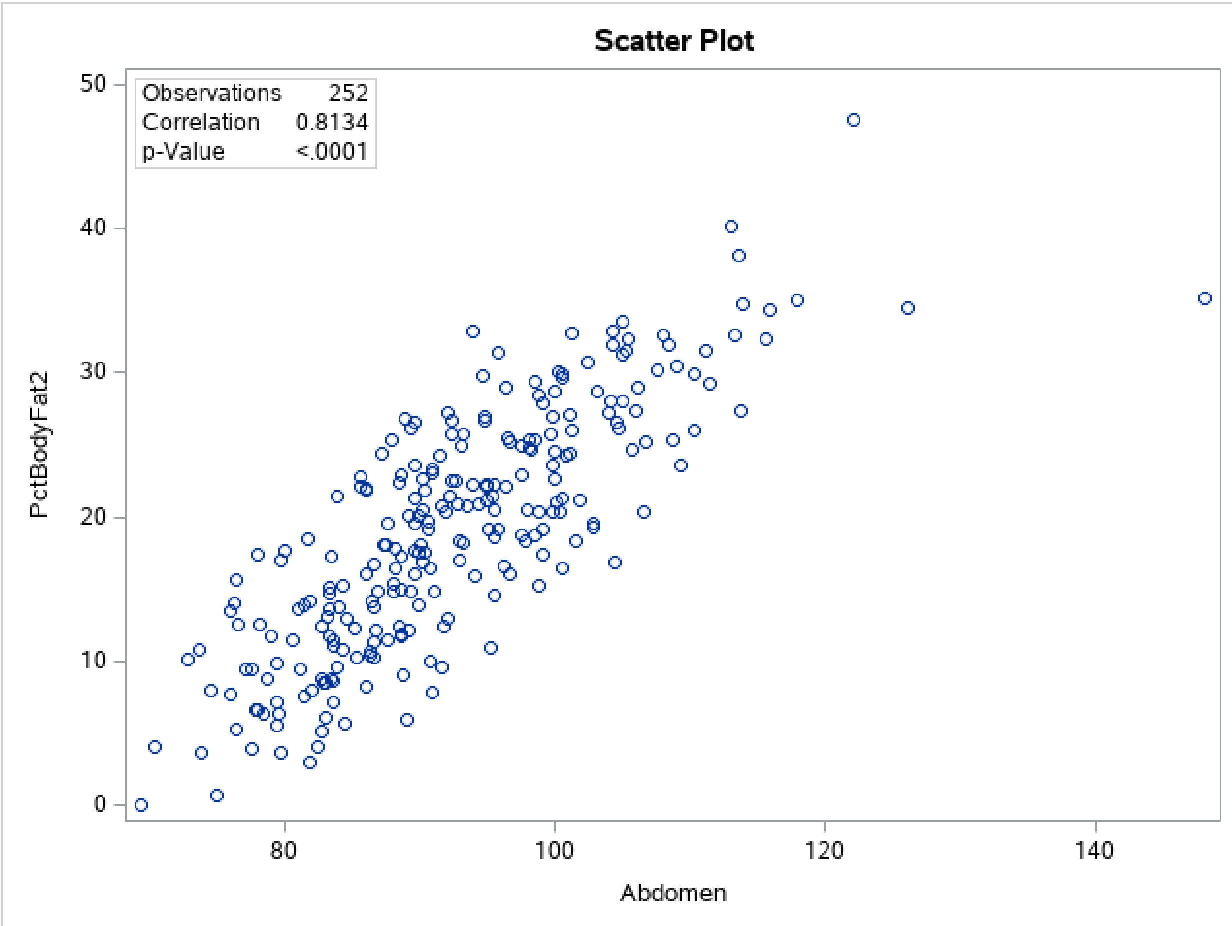


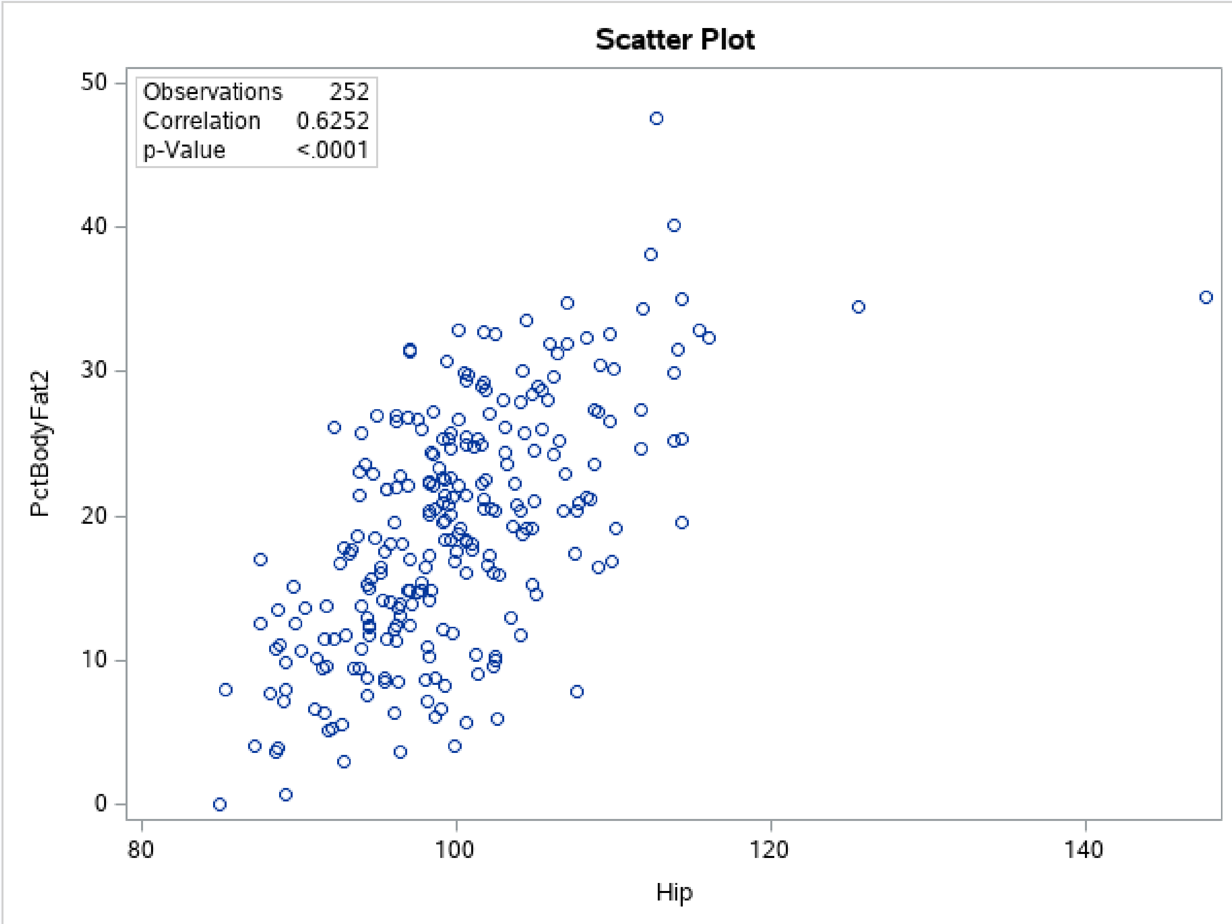
Scatter Plot
Observations 252
Correlation 0.6124
p-Value <.0001
PctBodyFat2
Weight

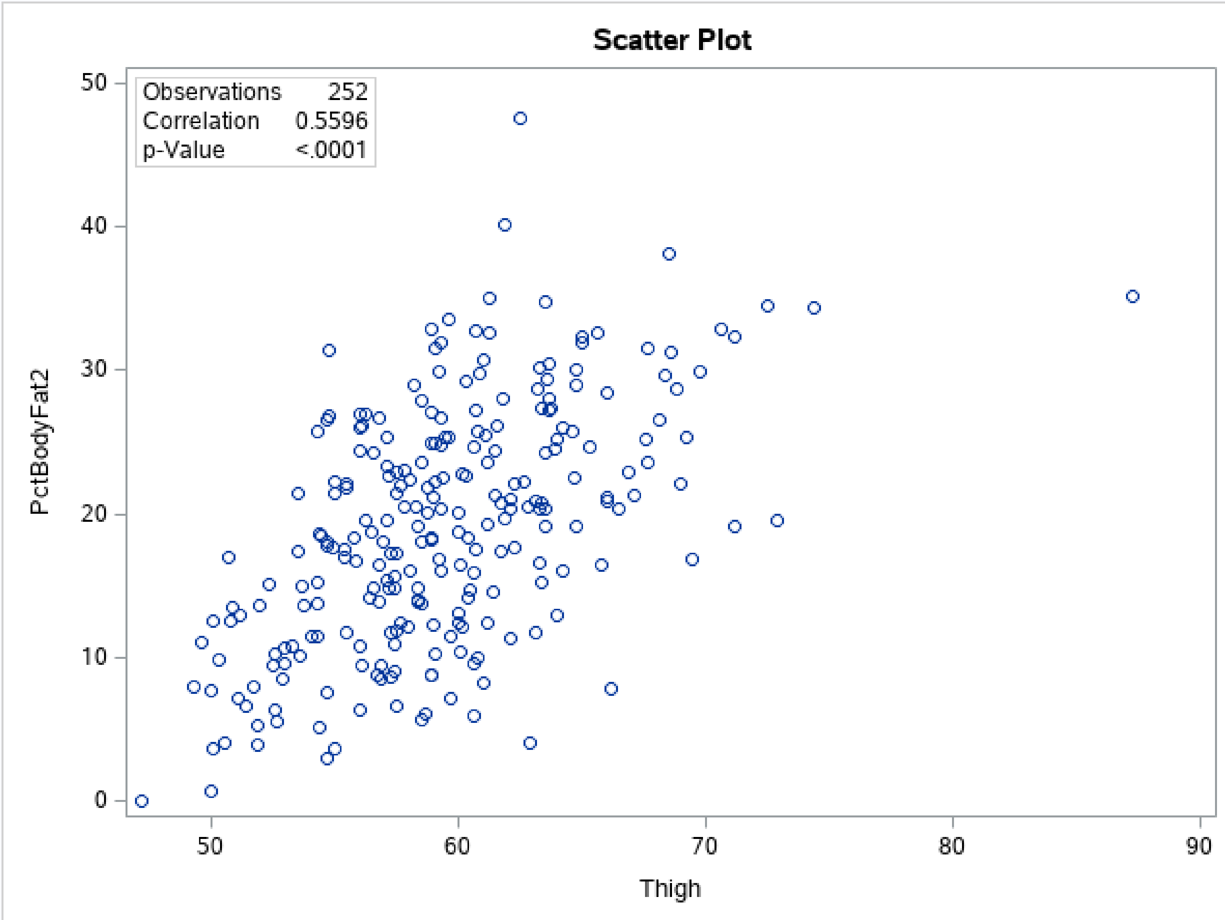

Scatter Plot
PctBodyFat2
Height
50
40
30
20
10
0
65
70
75
Observations 252
Correlation -0.025
p-Value 0.6895

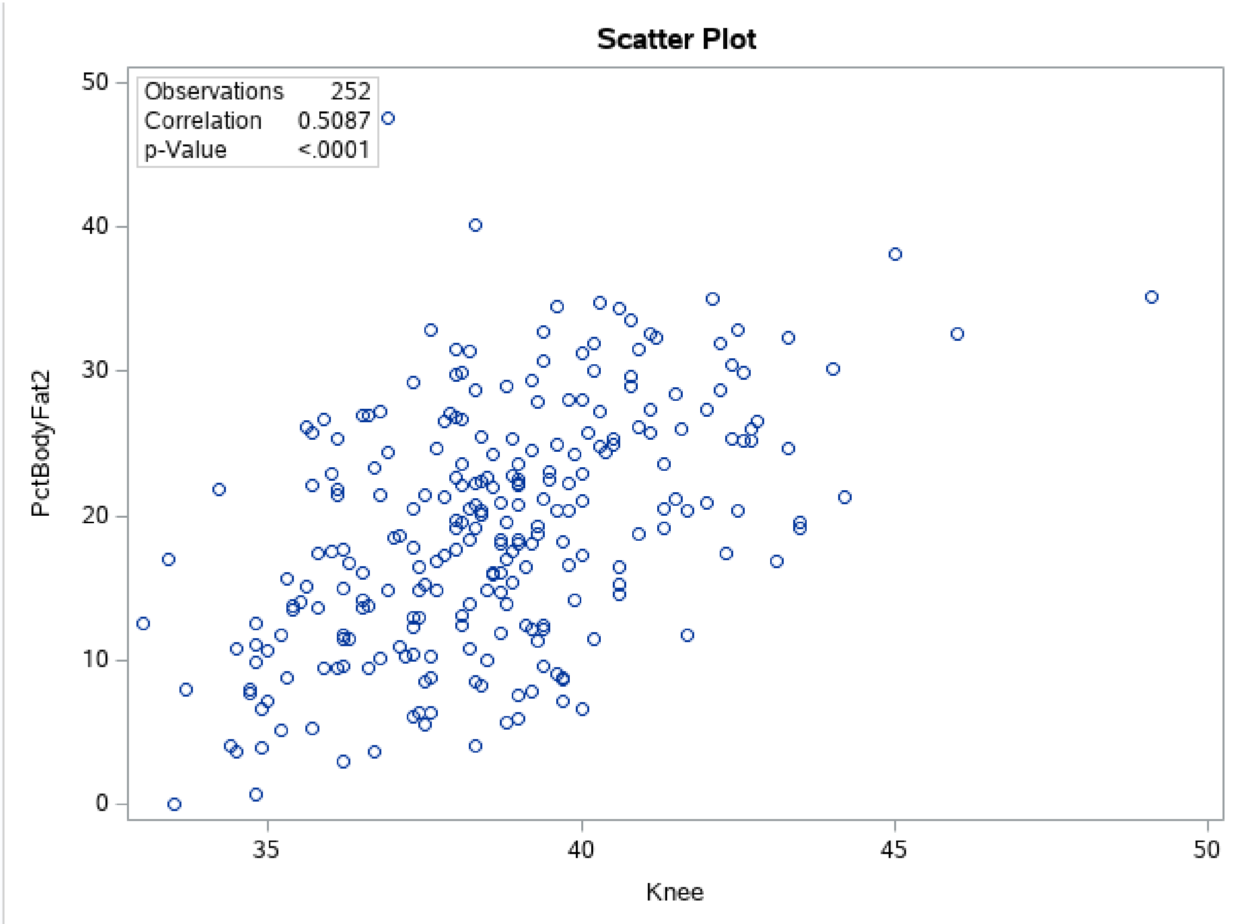

Scatter Plot
Observations 252
Correlation 0.4906
p-Value <.0001
PctBodyFat2
Neck
0
10
20
30
40
50
30
35
40
45
50

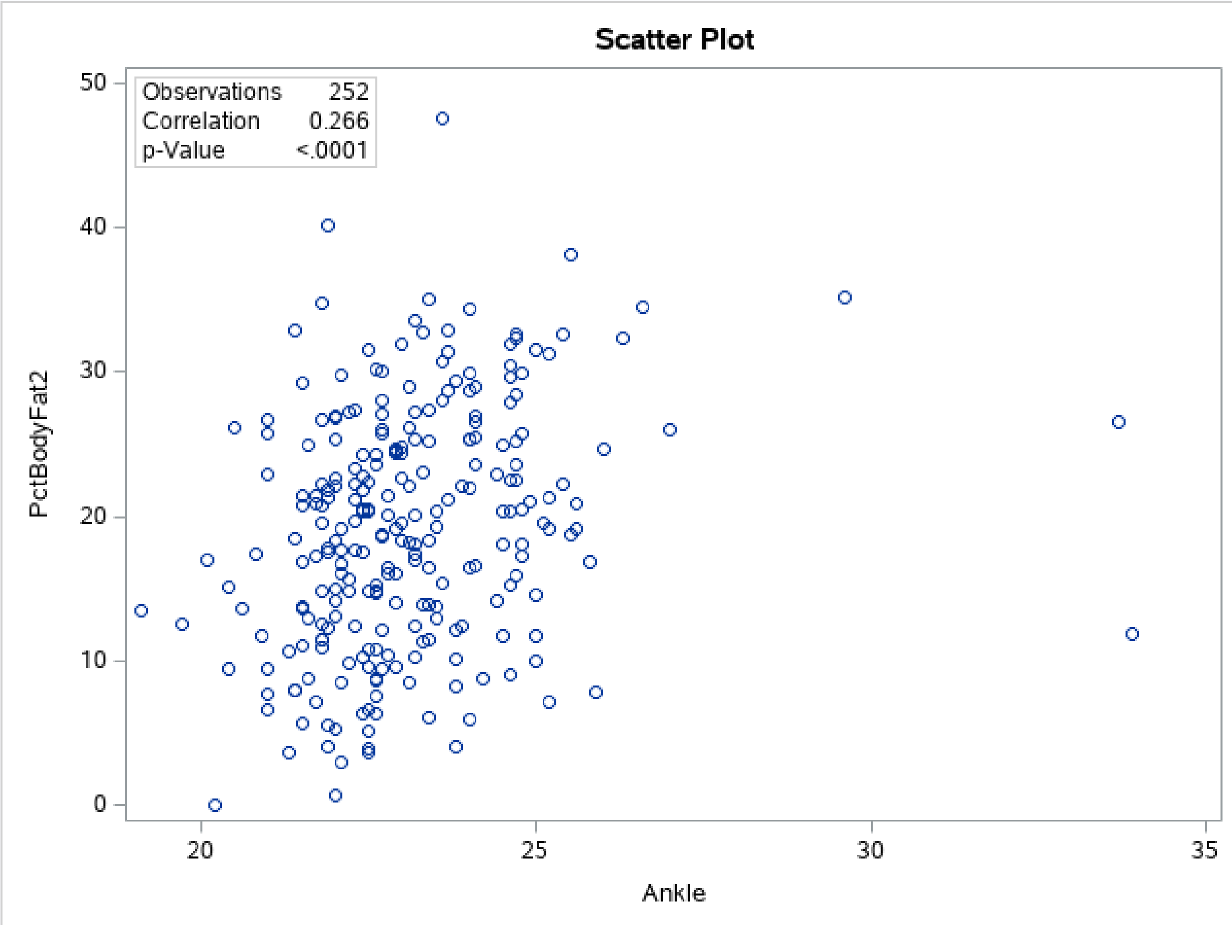

Scatter Plot
Observations 252
Correlation 0.7026
p-Value <.0001
PctBodyFat2
Chest
0
10
20
30
40
50
80
100
120
140


Scatter Plot
Observations 252
Correlation 0.8134
p-Value <.0001
PctBodyFat2
Abdomen
0
10
20
30
40
50
80
100
120
140


Scatter Plot
Observations 252
Correlation 0.6252
p-Value <.0001
PctBodyFat2
Hip
0
10
20
30
40
50
80
100
120
140

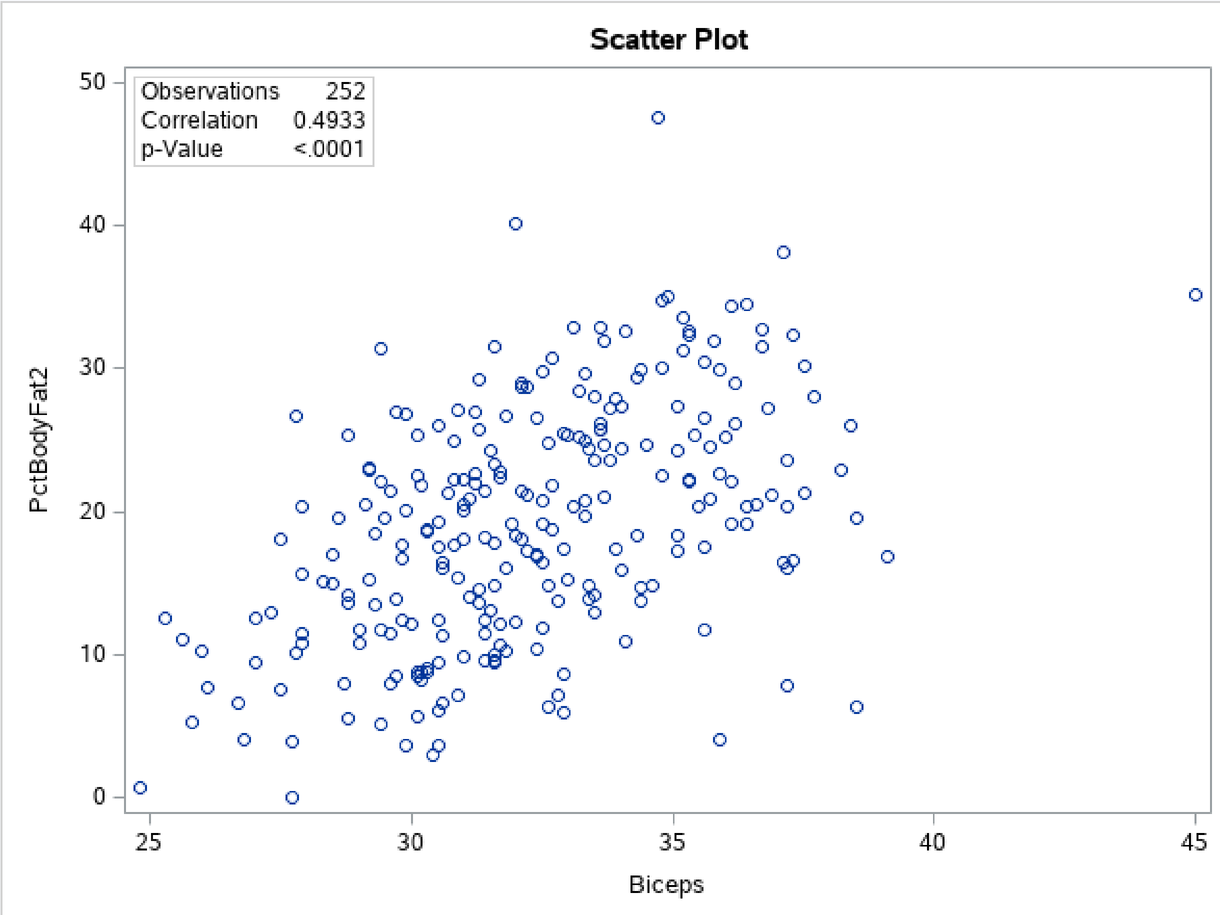

Scatter Plot
Observations 252
Correlation 0.5596
p-Value <.0001
PctBodyFat2
Thigh
0
10
20
30
40
50
50
60
70
80
90

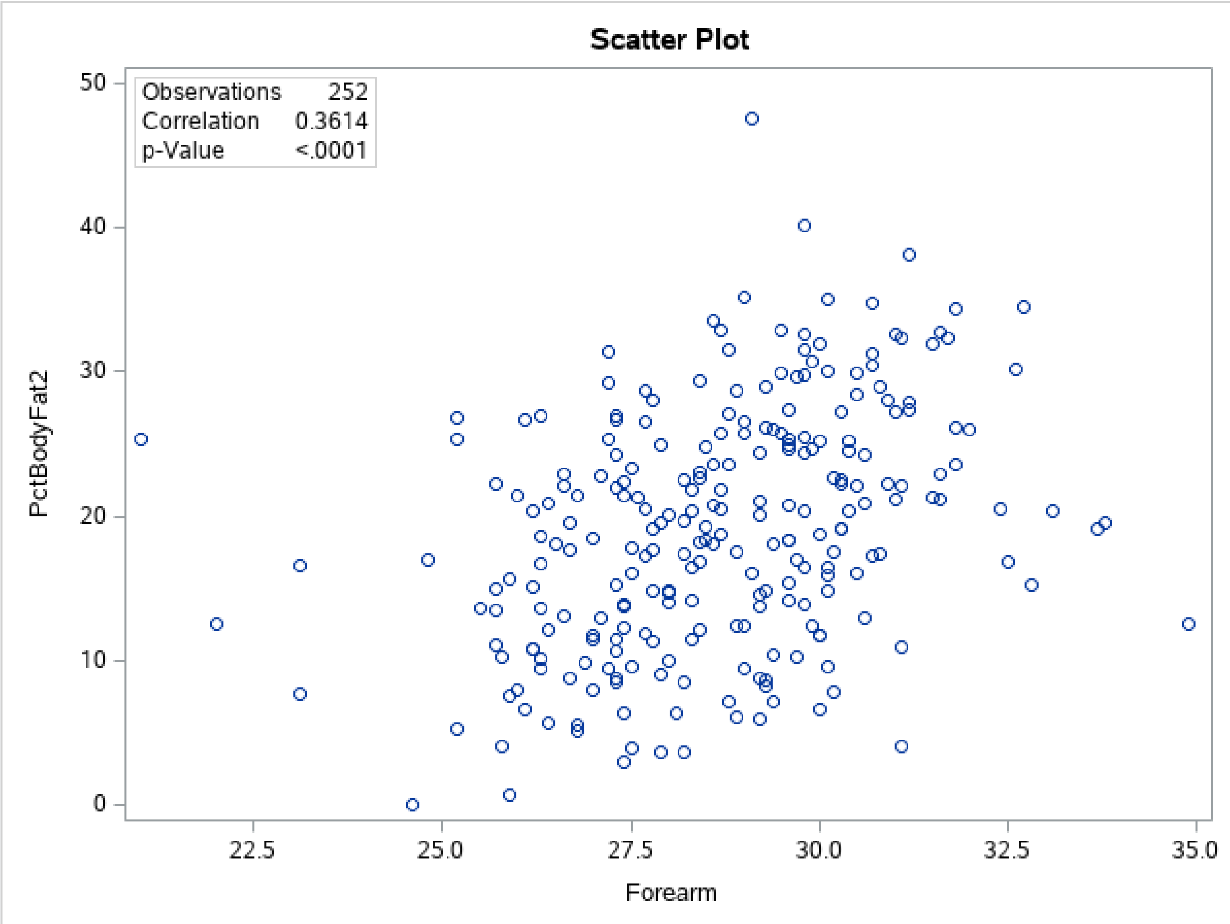

Scatter Plot
Observations 252
Correlation 0.5087
p-Value <.0001
PctBodyFat2
Knee
0
10
20
30
40
50
35
40
45
50

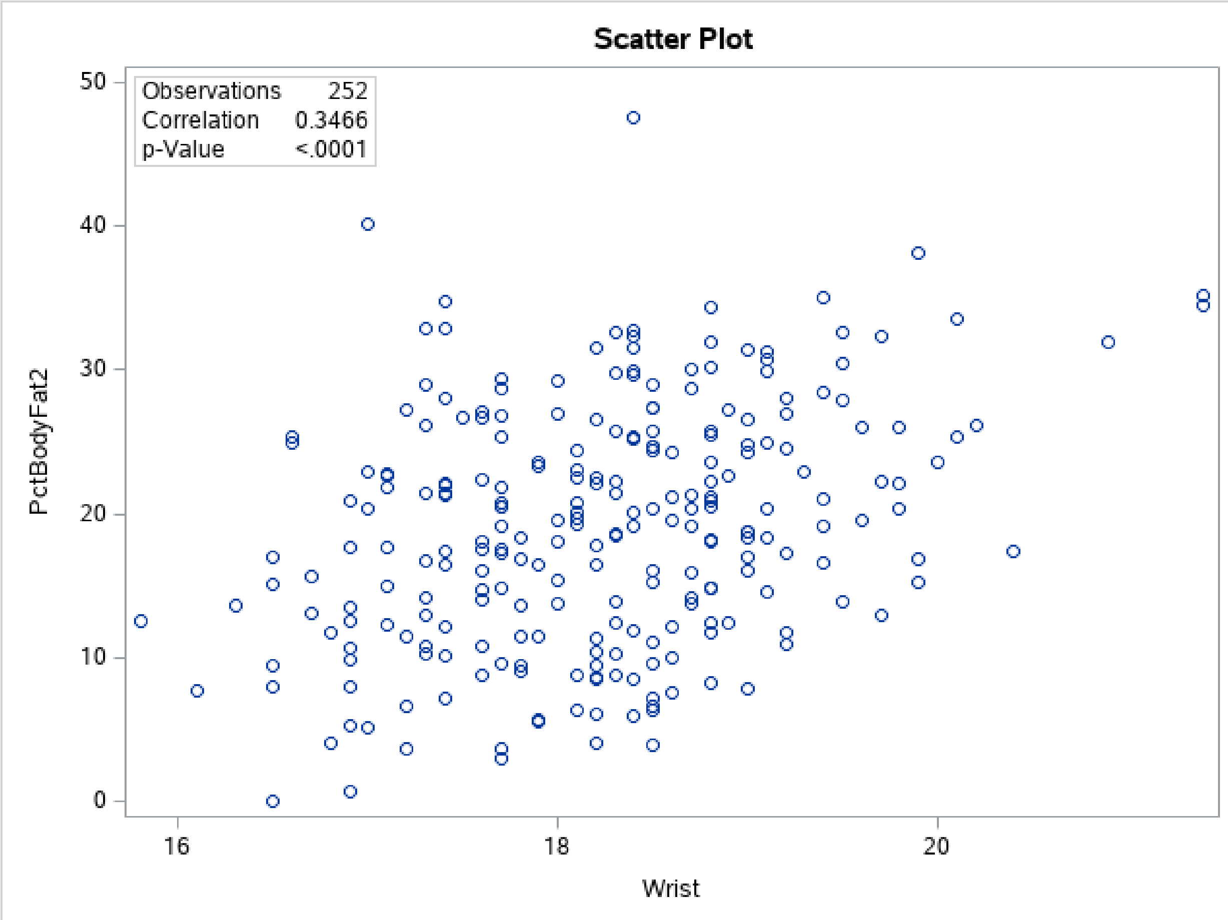

Scatter Plot
Observations 252
Correlation 0.266
p-Value <.0001
PctBodyFat2
Ankle
50
40
30
20
10
0
20
25
30
35

Correlations and Scatter Plots

The CORR Procedure




Scatter Plot
Observations 252
Correlation 0.3614
p-Value <.0001
PctBodyFat2
Forearm
0
10
20
30
40
50
22.5
25.0
27.5
30.0
32.5
35.0



Generate correlations among all the variables previously mentioned (**Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist**) minus **PctBodyFat2**. Use the OUT= option in the PROC CORR statement to output the correlation table into a data set named **pearson**. Use the BEST= option to select only the highest five per variable.

Submit the code and review the results. Are there any notable relationships?

```
/*st102s03.sas*/  /*Part B*/

ods graphics off;

%let interval=Age Weight Height Neck Chest Abdomen Hip Thigh
      Knee Ankle Biceps Forearm Wrist;


proc corr data=STAT1.BodyFat2
     nosimple
     best=5
```

```
        out=pearson;
   var &interval;
   title "Correlations of Predictors";
run;

%let big=0.7;
proc format;
    picture correlations &big -< 1 = '009.99' (prefix="*")
                    -1 <- -&big = '009.99' (prefix="*")
                    -&big <-< &big = '009.99';
run;

proc print data=pearson;
    var _NAME_ &interval;
    where _type_="CORR";
    format &interval correlations.;
run;

%let big=0.7;
data bigcorr;
    set pearson;
    array vars{*} &interval;
    do i=1 to dim(vars);
       if abs(vars{i})<&big then vars{i}=.;
    end;
    if _type_="CORR";
    drop i _type_;
run;
```

```
proc print data=bigcorr;
   format &interval 5.2;
run;

title;
```

Correlations of Predictors

The CORR Procedure

13 Variables:	Age Weight Height Neck Chest Abdomen Hip Thigh Knee Ankle Biceps Forearm Wrist

Pearson Correlation Coefficients, N = 252
Prob > |r| under H0: Rho=0

Age	Age 1.00000	Height -0.24521 <.0001	Abdomen 0.23041 0.0002	Wrist 0.21353 0.0006	Thigh -0.20010 0.0014
Weight	Weight 1.00000	Hip 0.94088 <.0001	Chest 0.89419 <.0001	Abdomen 0.88799 <.0001	Thigh 0.86869 <.0001
Height	Height 1.00000	Knee 0.50050 <.0001	Weight 0.48689 <.0001	Wrist 0.39778 <.0001	Ankle 0.39313 <.0001
Neck	Neck 1.00000	Weight 0.83072 <.0001	Chest 0.78484 <.0001	Abdomen 0.75408 <.0001	Wrist 0.74483 <.0001
Chest	Chest 1.00000	Abdomen 0.91583 <.0001	Weight 0.89419 <.0001	Hip 0.82942 <.0001	Neck 0.78484 <.0001
Abdomen	Abdomen 1.00000	Chest 0.91583 <.0001	Weight 0.88799 <.0001	Hip 0.87407 <.0001	Thigh 0.76662 <.0001
Hip	Hip 1.00000	Weight 0.94088 <.0001	Thigh 0.89641 <.0001	Abdomen 0.87407 <.0001	Chest 0.82942 <.0001
Thigh	Thigh 1.00000	Hip 0.89641 <.0001	Weight 0.86869 <.0001	Knee 0.79917 <.0001	Abdomen 0.76662 <.0001
Knee	Knee 1.00000	Weight 0.85317 <.0001	Hip 0.82347 <.0001	Thigh 0.79917 <.0001	Abdomen 0.73718 <.0001
Ankle	Ankle 1.00000	Weight 0.61369 <.0001	Knee 0.61161 <.0001	Wrist 0.56619 <.0001	Hip 0.55839 <.0001
Biceps	Biceps 1.00000	Weight 0.80042 <.0001	Thigh 0.76148 <.0001	Hip 0.73927 <.0001	Neck 0.73115 <.0001
Forearm	Forearm 1.00000	Biceps 0.67826 <.0001	Weight 0.63030 <.0001	Neck 0.62366 <.0001	Wrist 0.58559 <.0001
Wrist	Wrist 1.00000	Neck 0.74483 <.0001	Weight 0.72977 <.0001	Knee 0.66451 <.0001	Chest 0.66016 <.0001

Correlations of Predictors

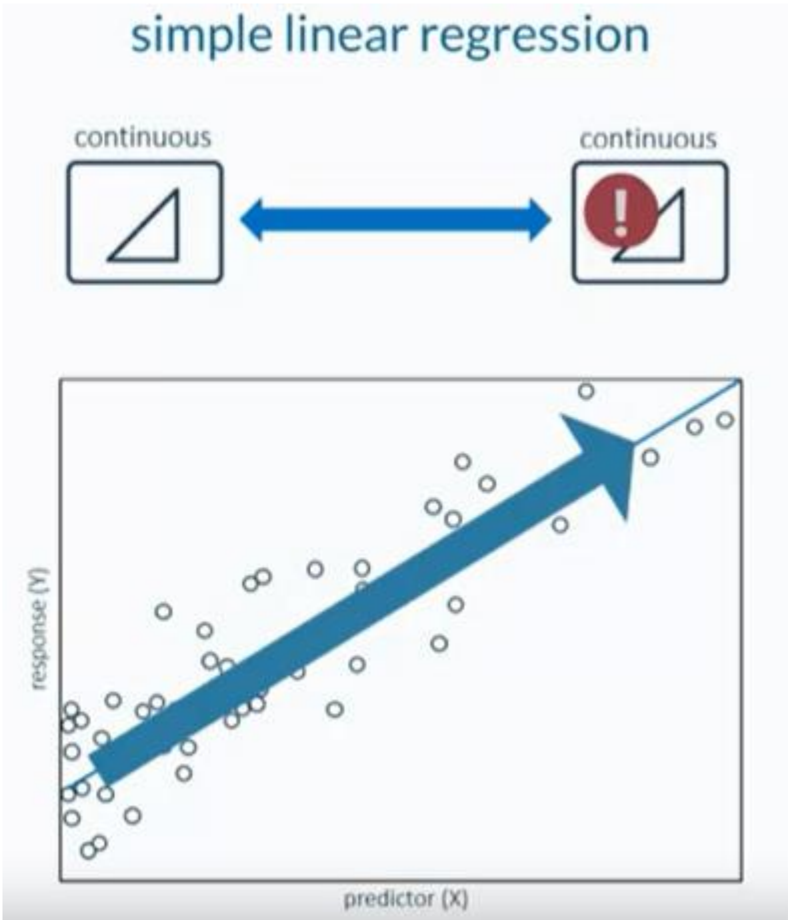
Obs	_NAME_	Age	Weight	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
4	Age	1	0.01	0.24	0.11	0.17	0.23	0.05	0.20	0.01	0.10	0.04	0.08	0.21
5	Weight	0.01	1	0.48	*0.83	*0.89	*0.88	*0.94	*0.86	*0.85	0.61	*0.80	0.63	*0.72
6	Height	0.24	0.48	1	0.32	0.22	0.18	0.37	0.33	0.50	0.39	0.31	0.32	0.39
7	Neck	0.11	*0.83	0.32	1	*0.78	*0.75	*0.73	0.69	0.67	0.47	*0.73	0.62	*0.74
8	Chest	0.17	*0.89	0.22	*0.78	1	*0.91	*0.82	*0.72	*0.71	0.48	*0.72	0.58	0.66
9	Abdomen	0.23	*0.88	0.18	*0.75	*0.91	1	*0.87	*0.76	*0.73	0.45	0.68	0.50	0.61
10	Hip	0.05	*0.94	0.37	*0.73	*0.82	*0.87	1	*0.89	*0.82	0.55	*0.73	0.54	0.63
11	Thigh	0.20	*0.86	0.33	0.69	*0.72	*0.76	*0.89	1	*0.79	0.53	*0.76	0.56	0.55
12	Knee	0.01	*0.85	0.50	0.67	*0.71	*0.73	*0.82	*0.79	1	0.61	0.67	0.55	0.66
13	Ankle	0.10	0.61	0.39	0.47	0.48	0.45	0.55	0.53	0.61	1	0.48	0.41	0.56
14	Biceps	0.04	*0.80	0.31	*0.73	*0.72	0.68	*0.73	*0.76	0.67	0.48	1	0.67	0.63
15	Forearm	0.08	0.63	0.32	0.62	0.58	0.50	0.54	0.56	0.55	0.41	0.67	1	0.58
16	Wrist	0.21	*0.72	0.39	*0.74	0.66	0.61	0.63	0.55	0.66	0.56	0.63	0.58	1

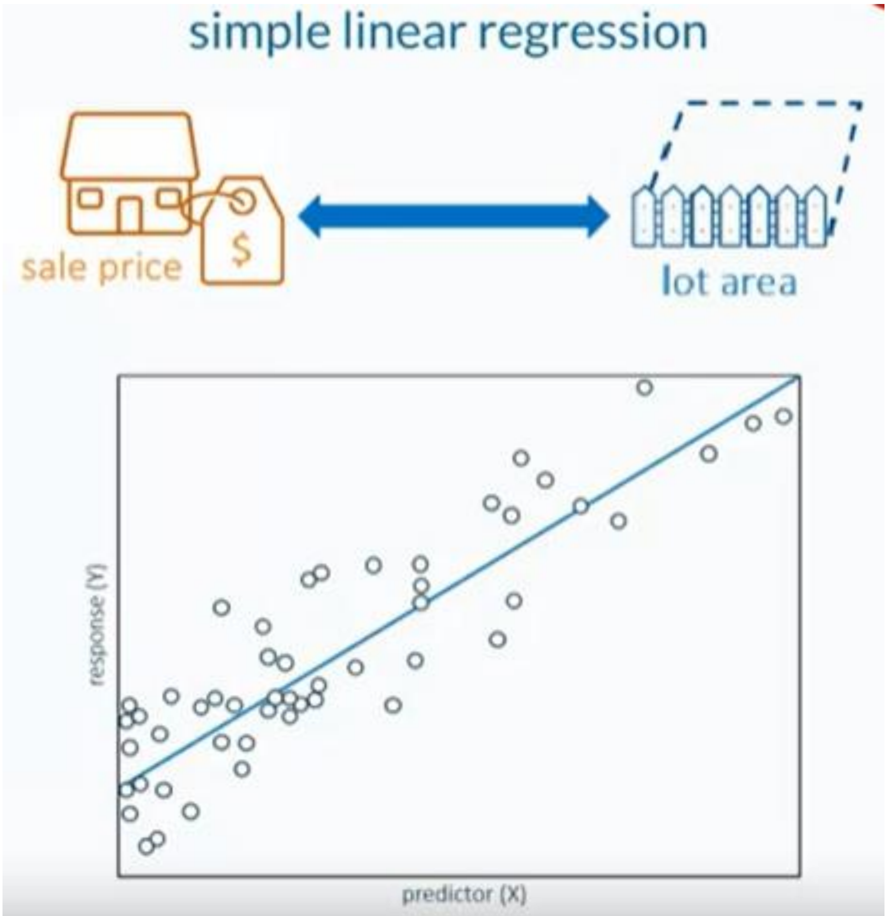
Correlations of Predictors

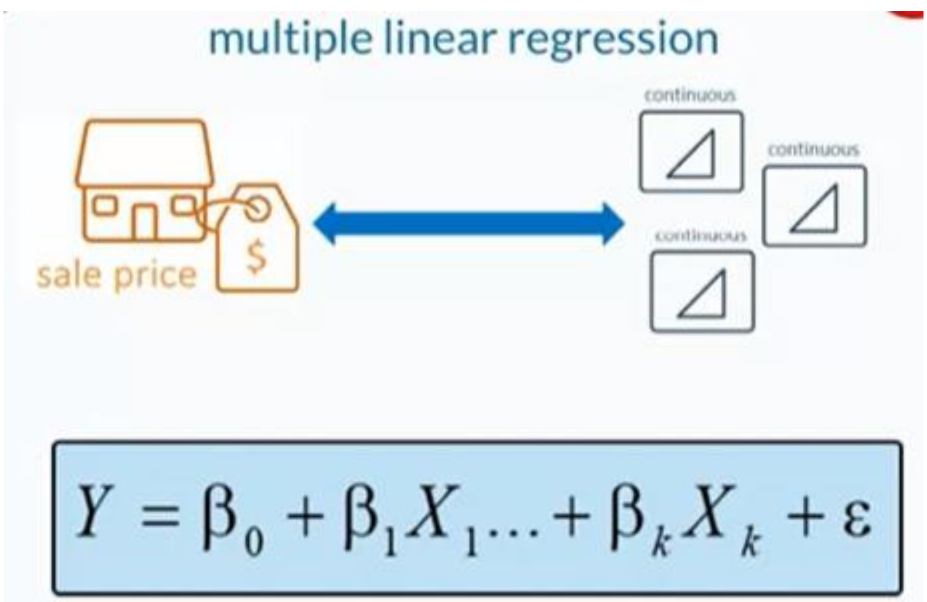
Obs	_NAME_	Age	Weight	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
1	Age	1.00	.	.	.	.	.	.	.	.	.	.	.	.
2	Weight	.	1.00	.	0.83	0.89	0.89	0.94	0.87	0.85	.	0.80	.	0.73
3	Height	.	.	1.00	.	.	.	.	.	.	.	.	.	.
4	Neck	.	0.83	.	1.00	0.78	0.75	0.73	.	.	.	0.73	.	0.74
5	Chest	.	0.89	.	0.78	1.00	0.92	0.83	0.73	0.72	.	0.73	.	.
6	Abdomen	.	0.89	.	0.75	0.92	1.00	0.87	0.77	0.74	.	.	.	.
7	Hip	.	0.94	.	0.73	0.83	0.87	1.00	0.90	0.82	.	0.74	.	.
8	Thigh	.	0.87	.	.	0.73	0.77	0.90	1.00	0.80	.	0.76	.	.
9	Knee	.	0.85	.	.	0.72	0.74	0.82	0.80	1.00	.	.	.	.
10	Ankle	.	.	.	.	.	.	.	.	.	1.00	.	.	.
11	Biceps	.	0.80	.	0.73	0.73	.	0.74	0.76	.	.	1.00	.	.
12	Forearm	.	.	.	.	.	.	.	.	.	.	.	1.00	.
13	Wrist	.	0.73	.	0.74	.	.	.	.	.	.	.	.	1.00

Simple Linear Regression

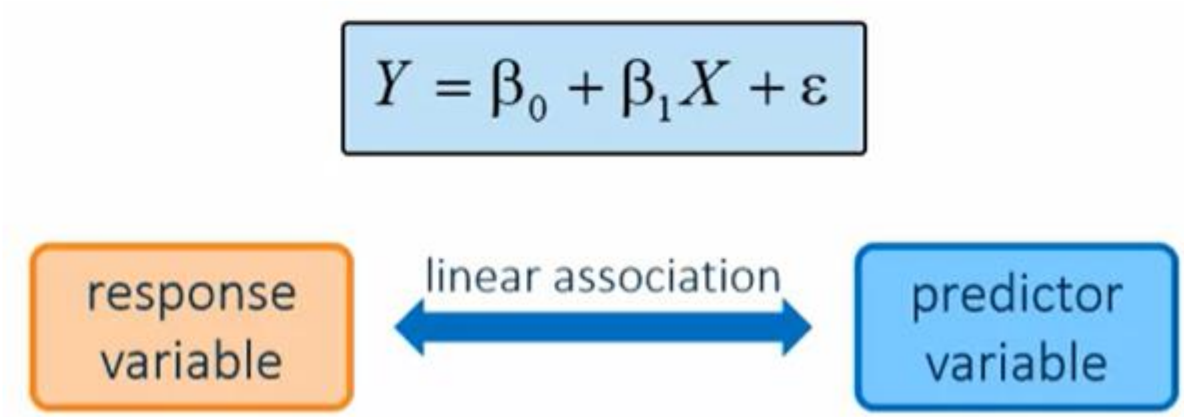
Scenario



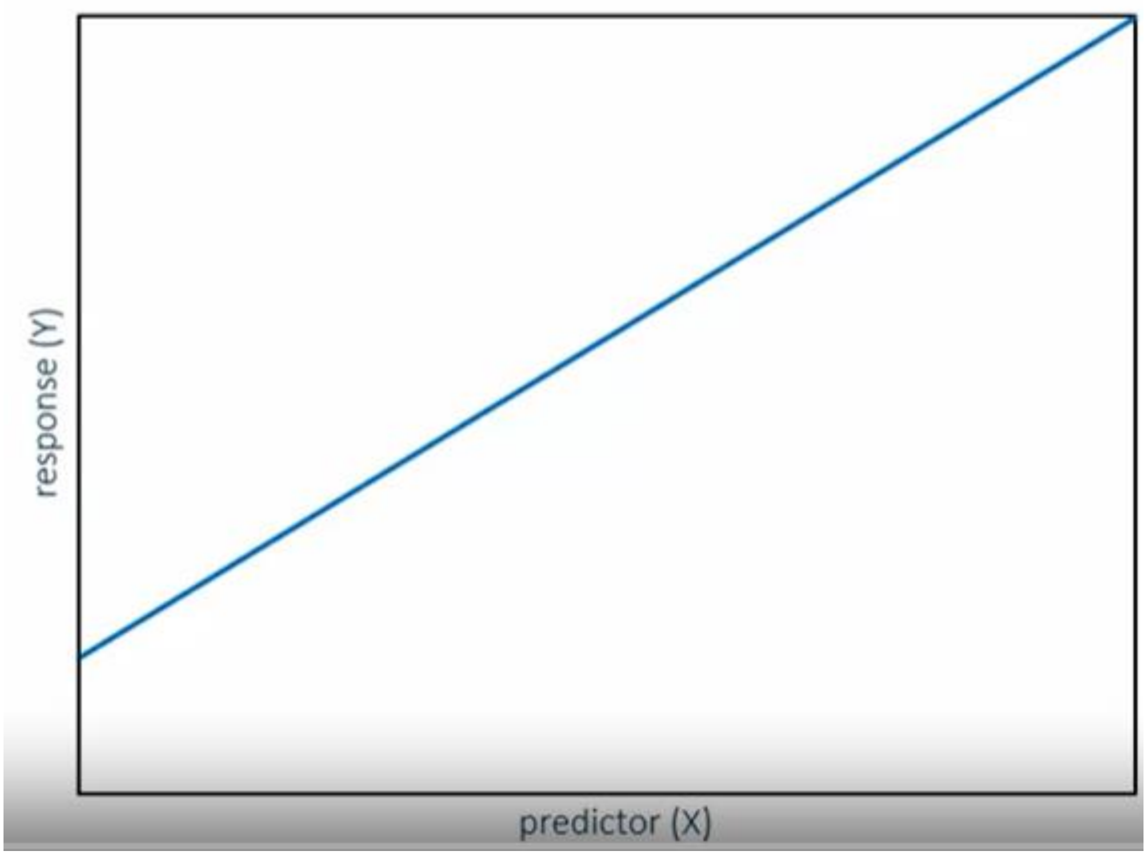

simple linear regression
sale price
$
lot area
response (Y)
predictor (X)

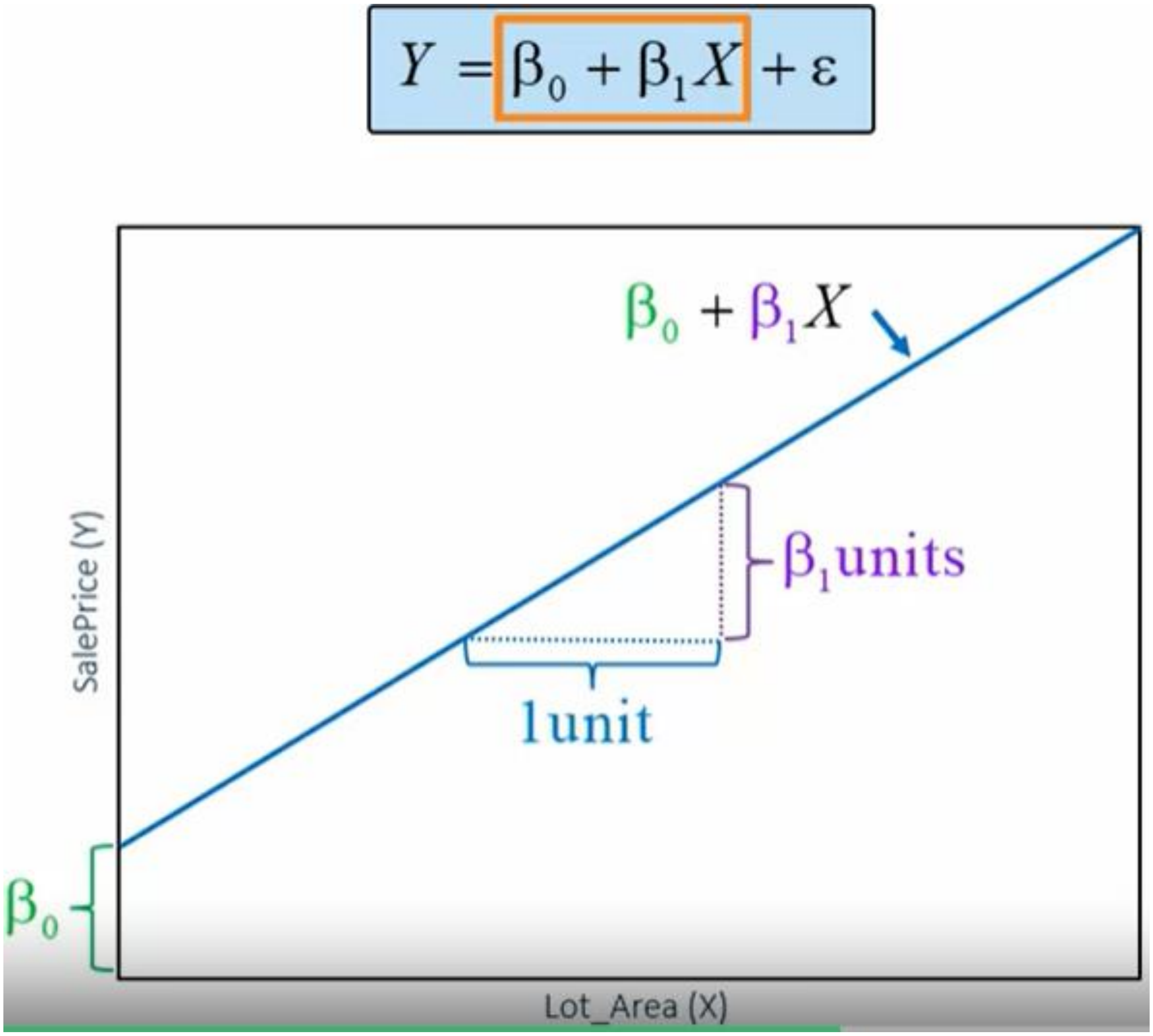

multiple linear regression
sale price
$
continuous
continuous
continuous
$Y = \beta_0 + \beta_1 X_1 \ldots + \beta_k X_k + \varepsilon$

The Simple Linear Regression Model

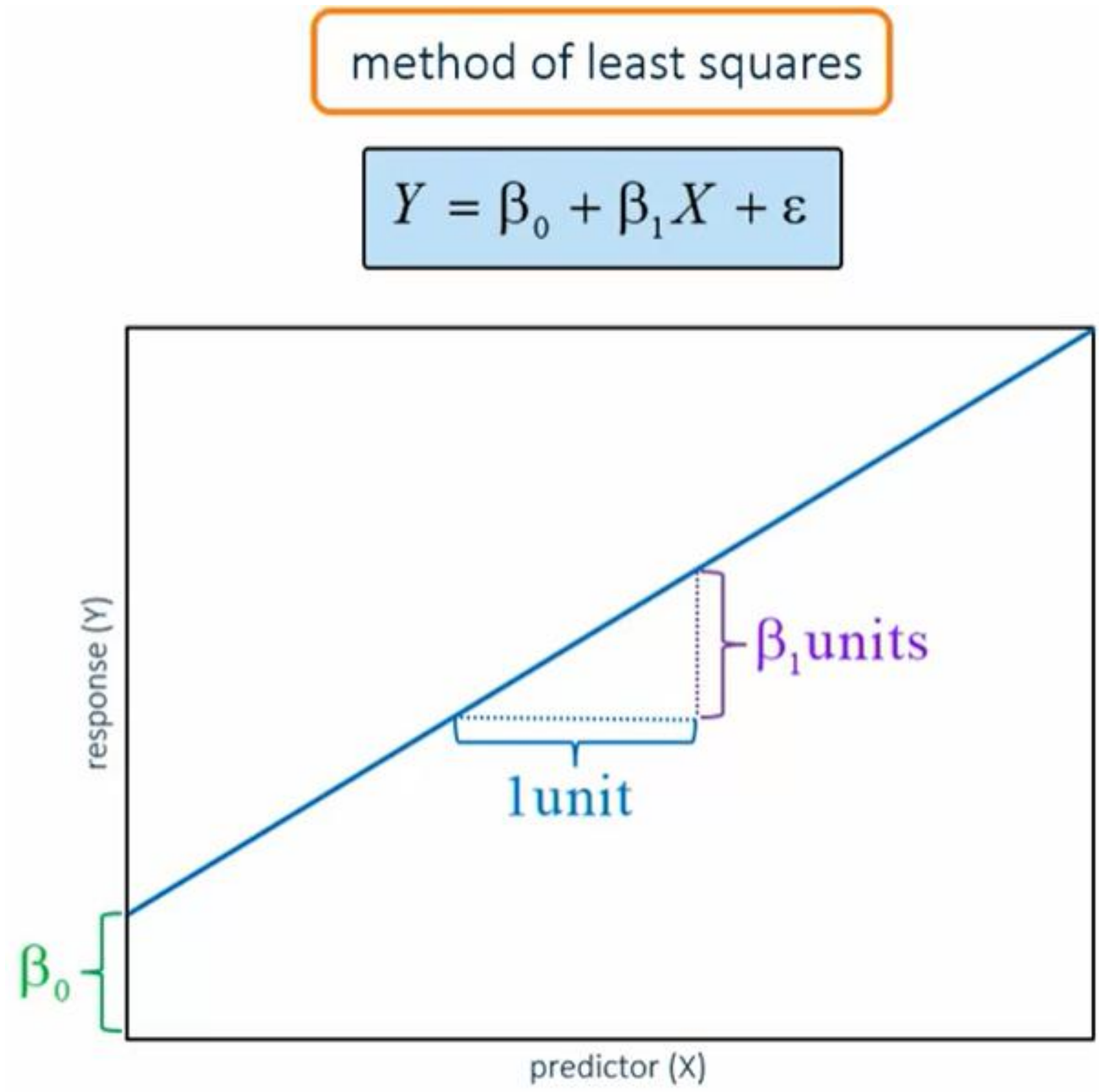


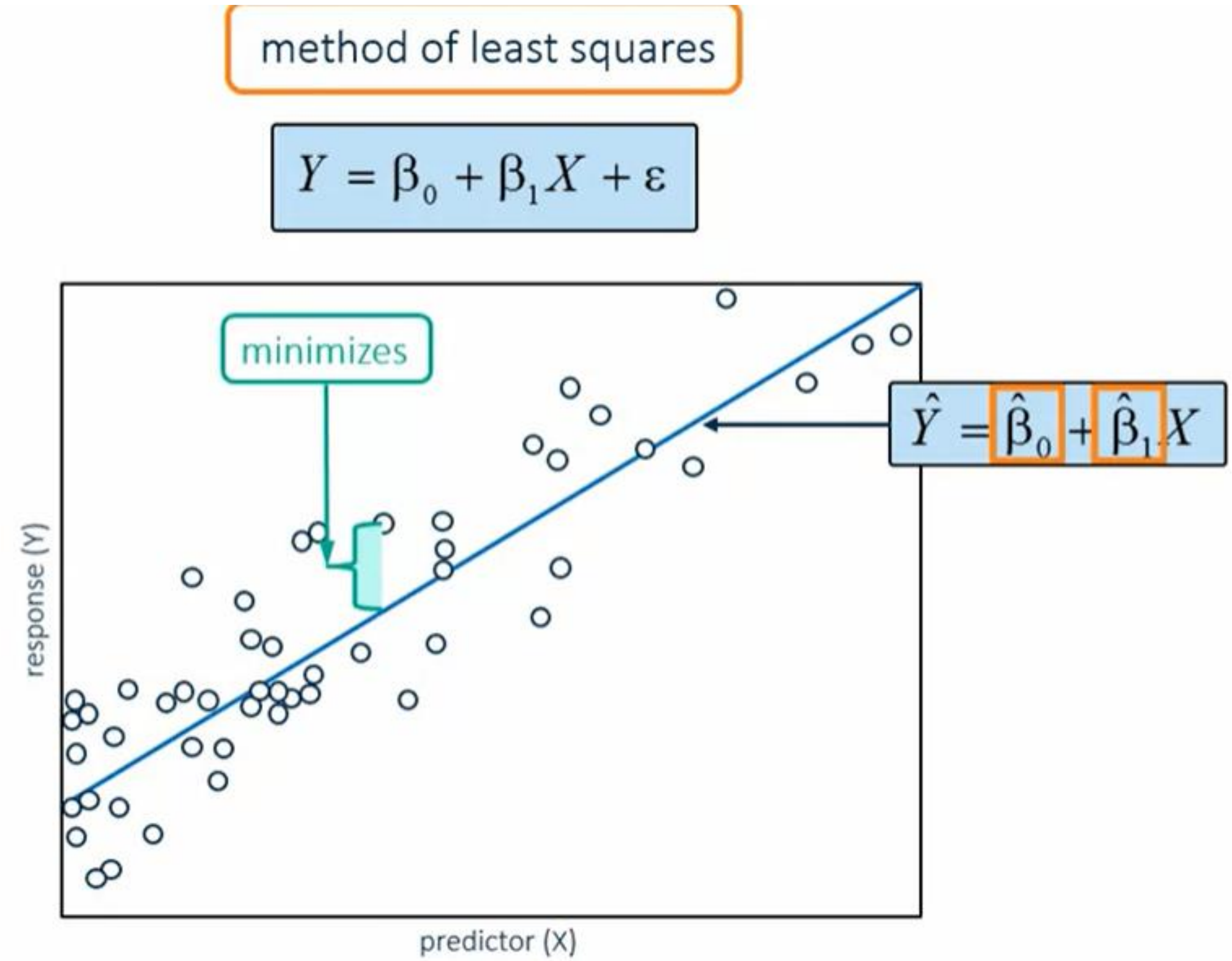
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

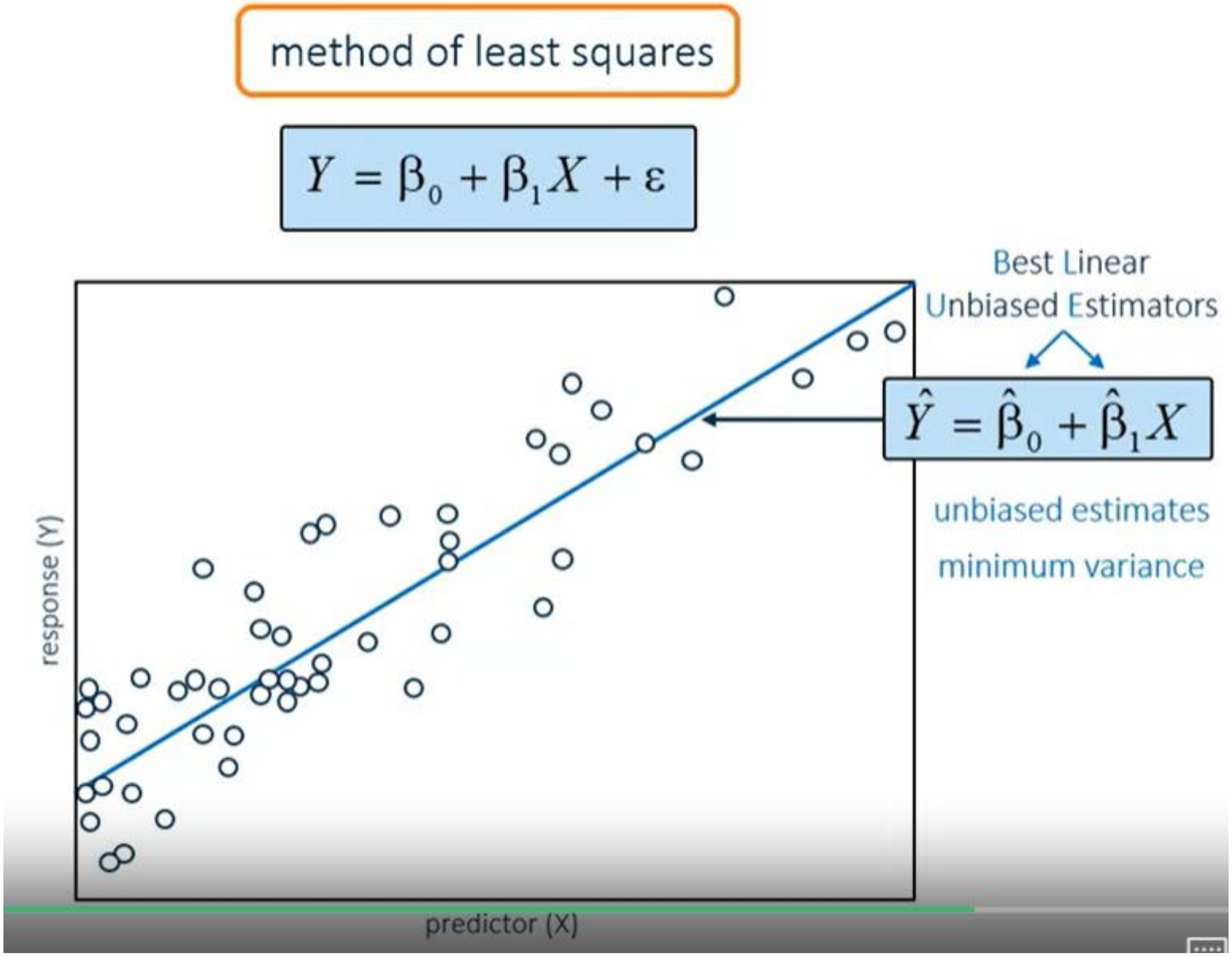


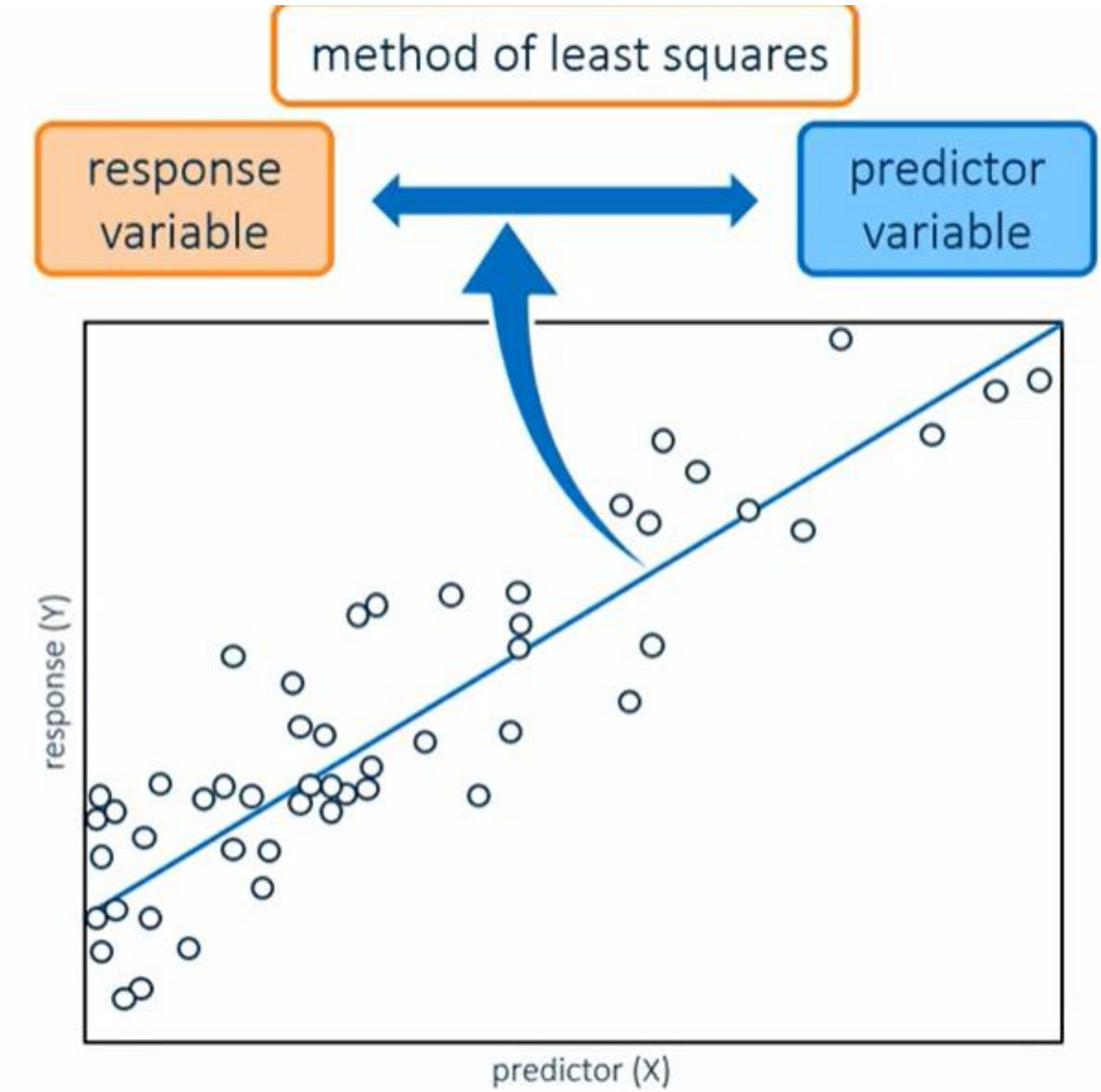


How SAS performs Simple Linear Regression

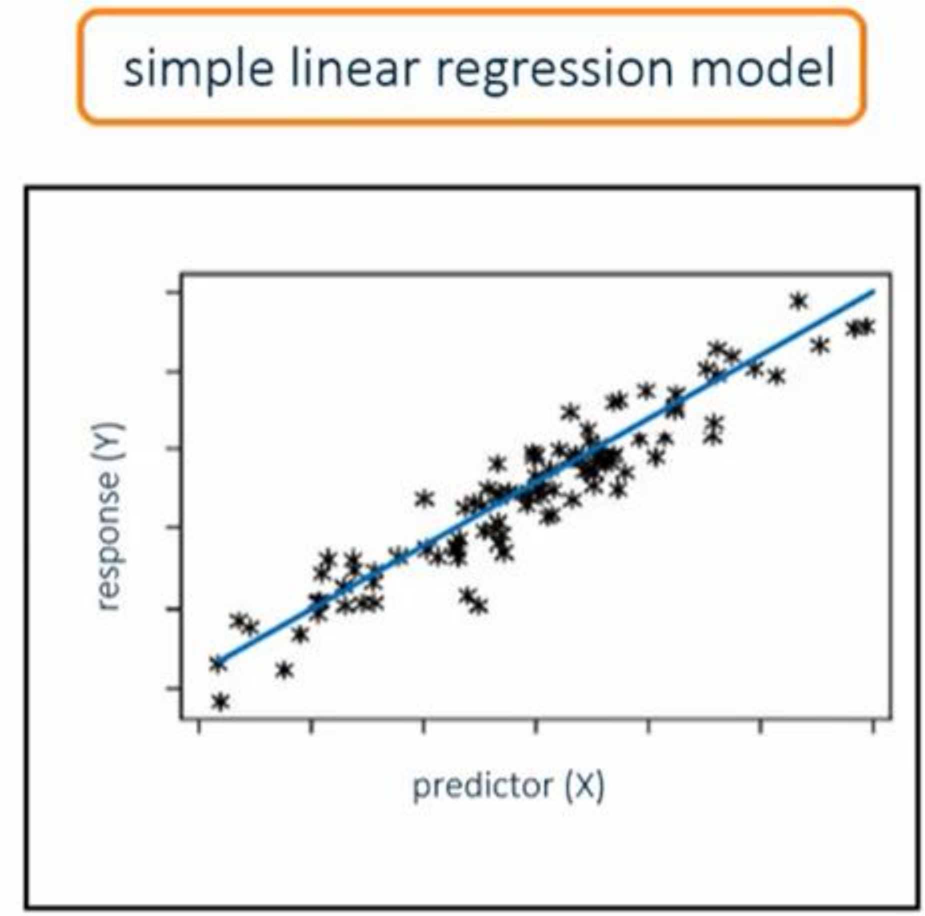

method of least squares
$Y = \beta_0 + \beta_1 X + \varepsilon$
β_1 units
1 unit
β_0
response (Y)
predictor (X)

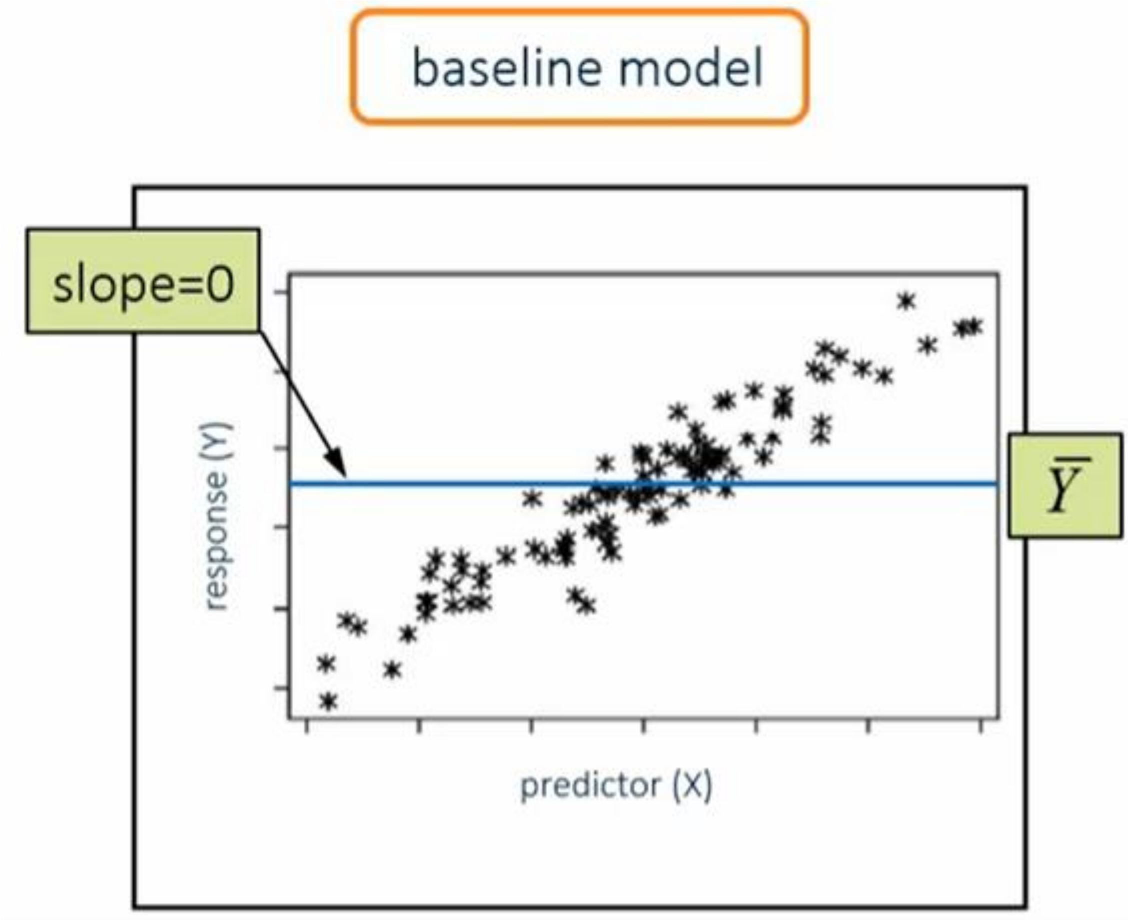

method of least squares
$Y = \beta_0 + \beta_1 X + \varepsilon$
minimizes
$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
response (Y)
predictor (X)

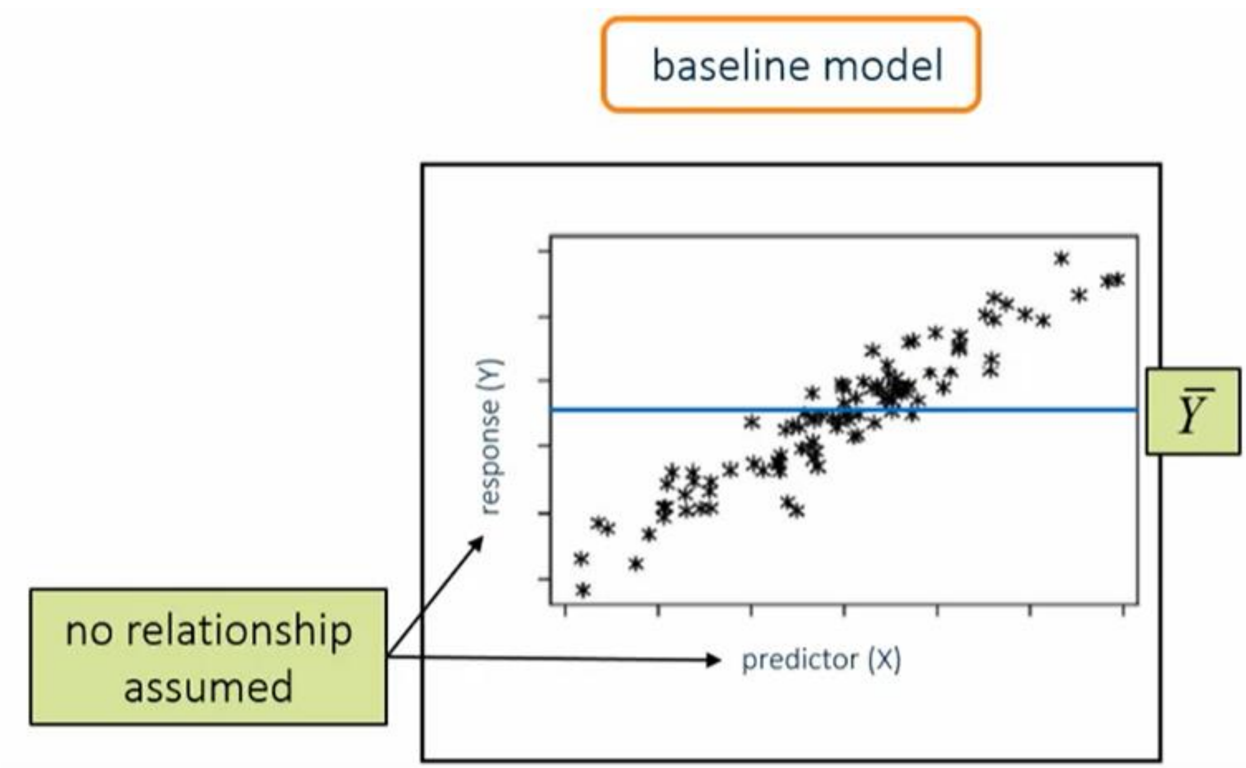

method of least squares
$Y = \beta_0 + \beta_1 X + \varepsilon$
Best Linear
Unbiased Estimators
$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
unbiased estimates
minimum variance
response (Y)
predictor (X)

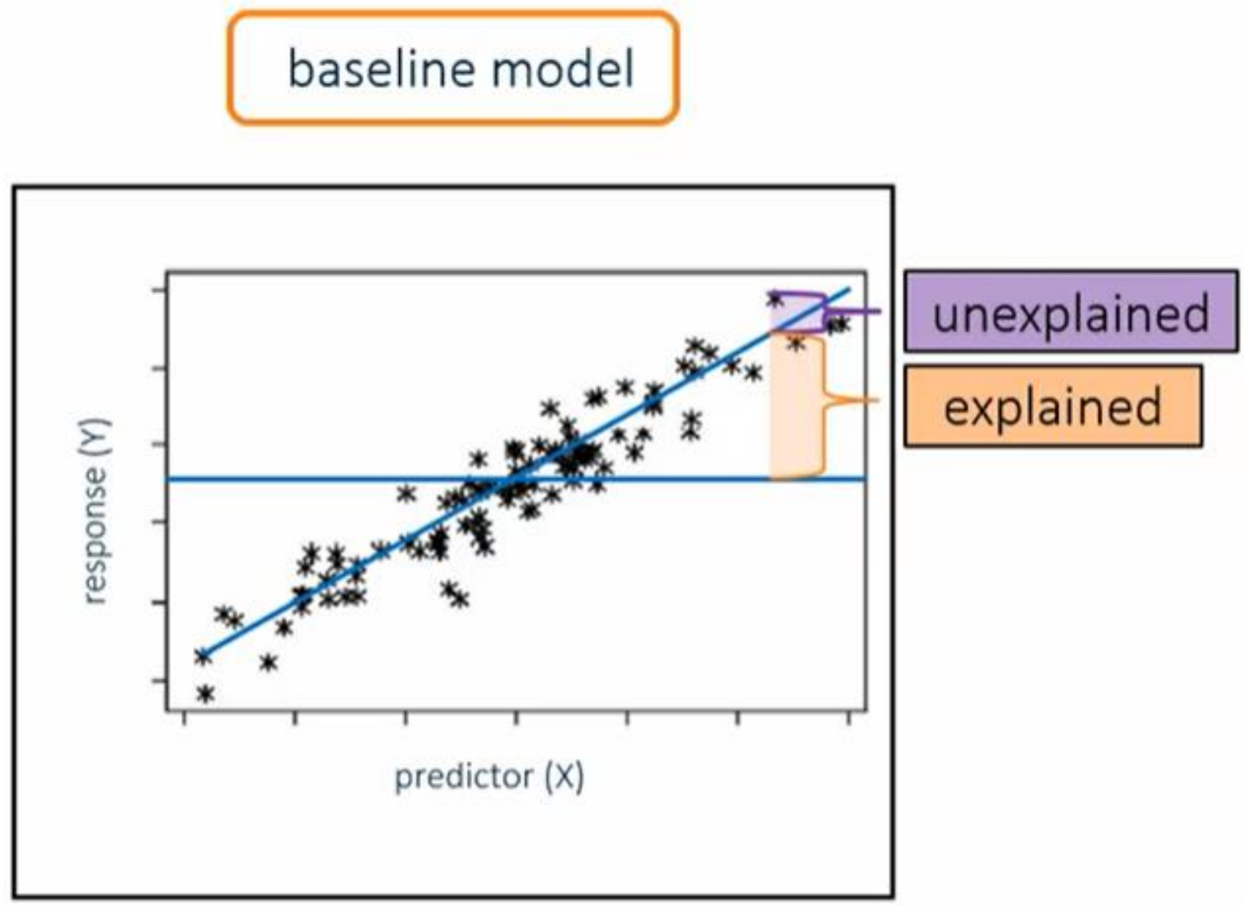

method of least squares
response variable
predictor variable
response (Y)
predictor (X)

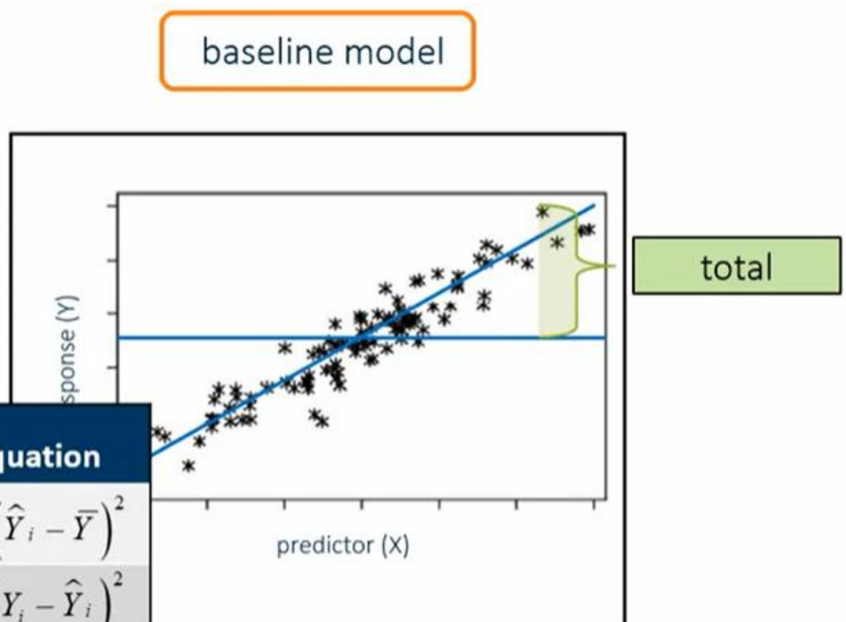
Comparing the Regression Model to a Baseline Model



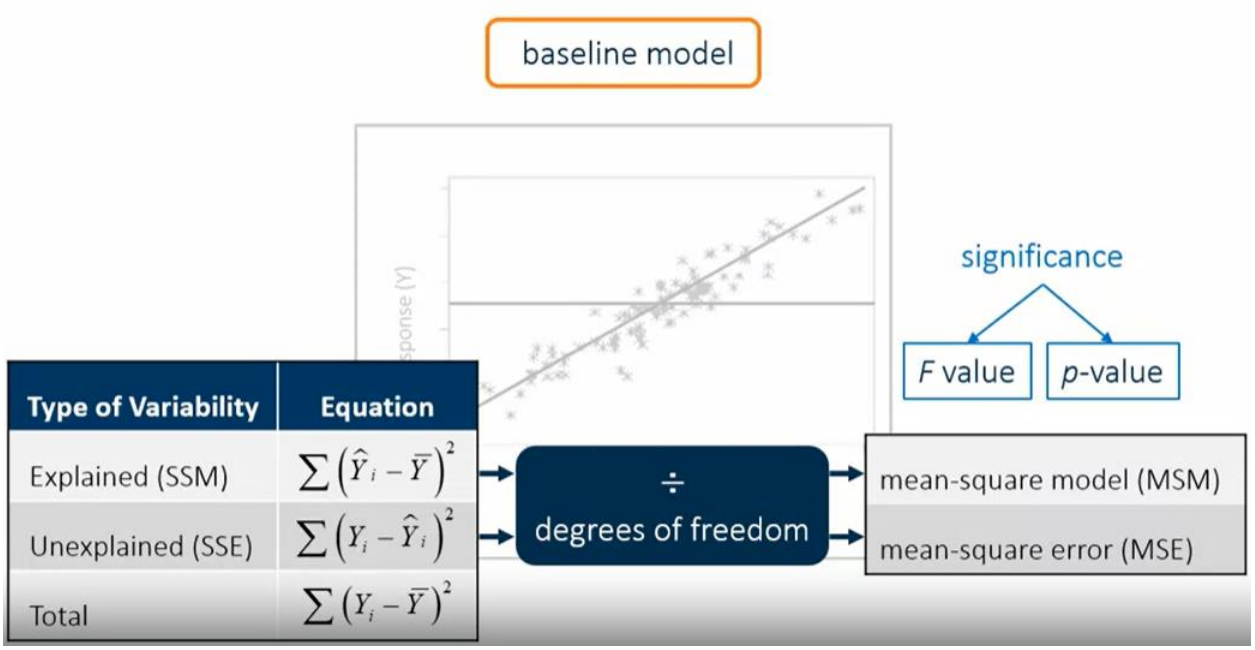

baseline model
slope=0
response (Y)
predictor (X)
$\bar{Y}$


baseline model
response (Y)
predictor (X)
$\bar{Y}$
no relationship assumed

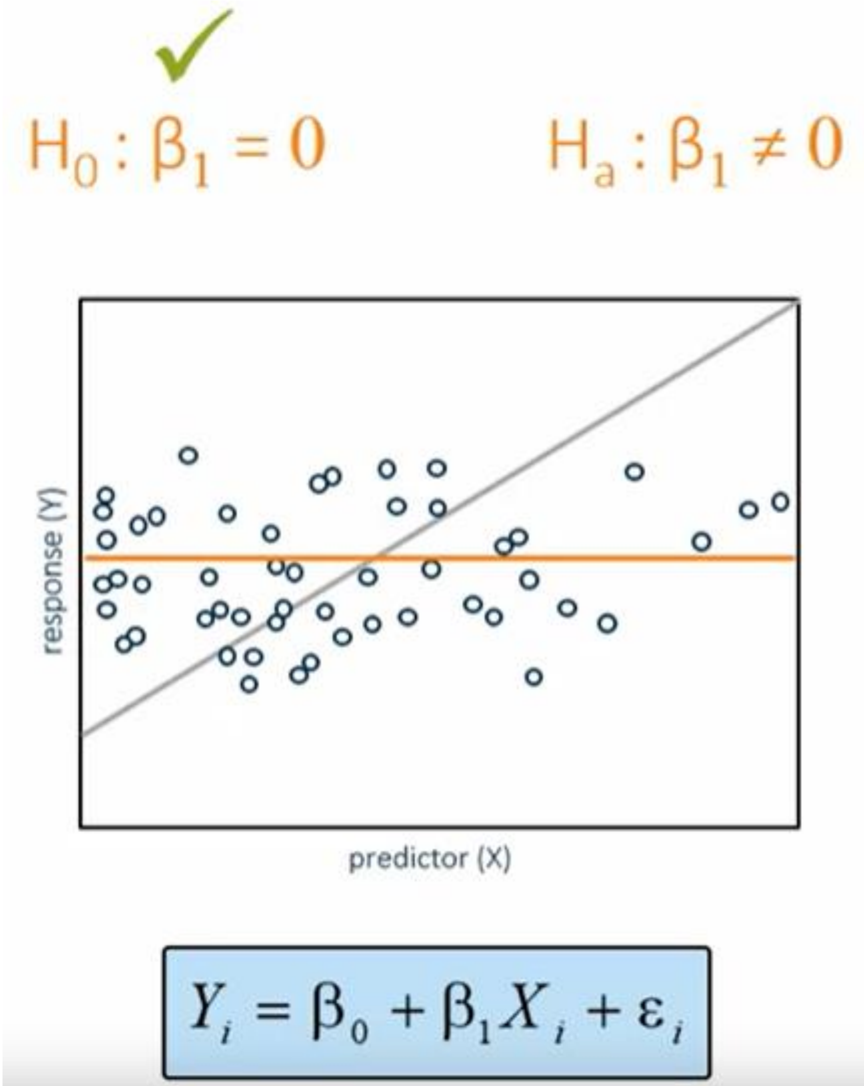


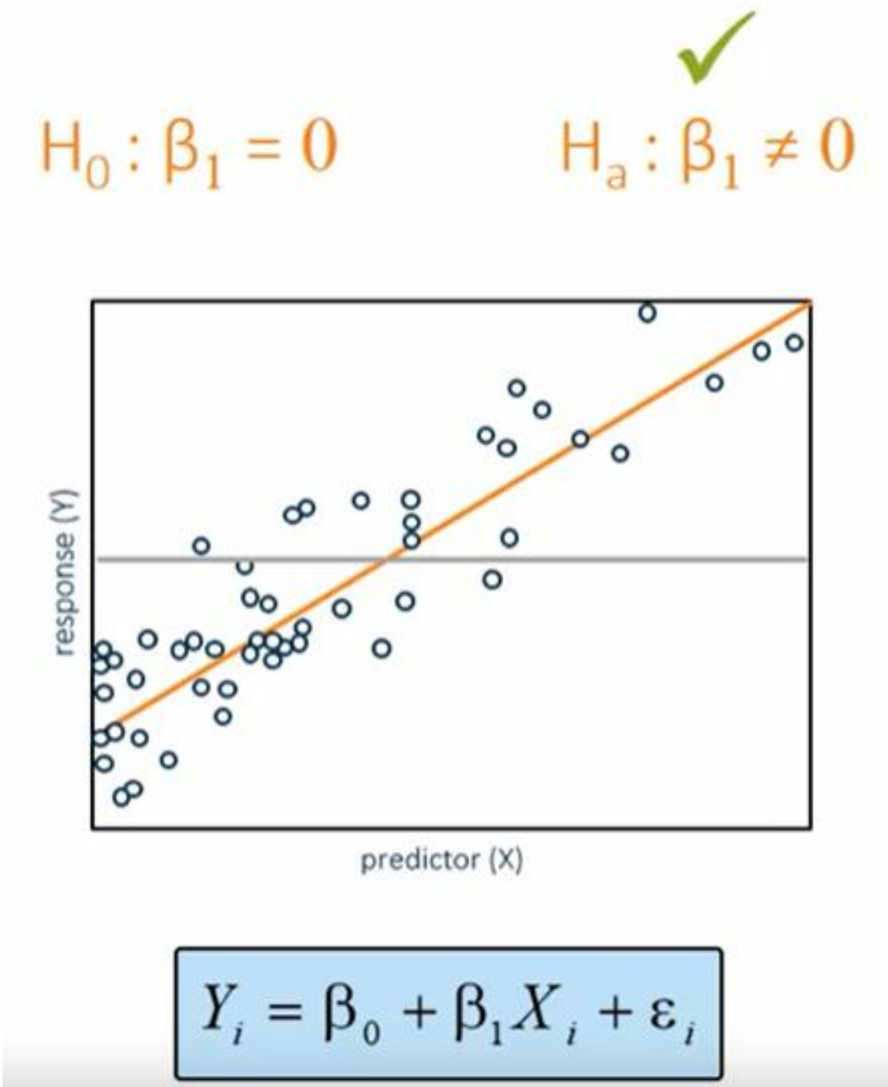


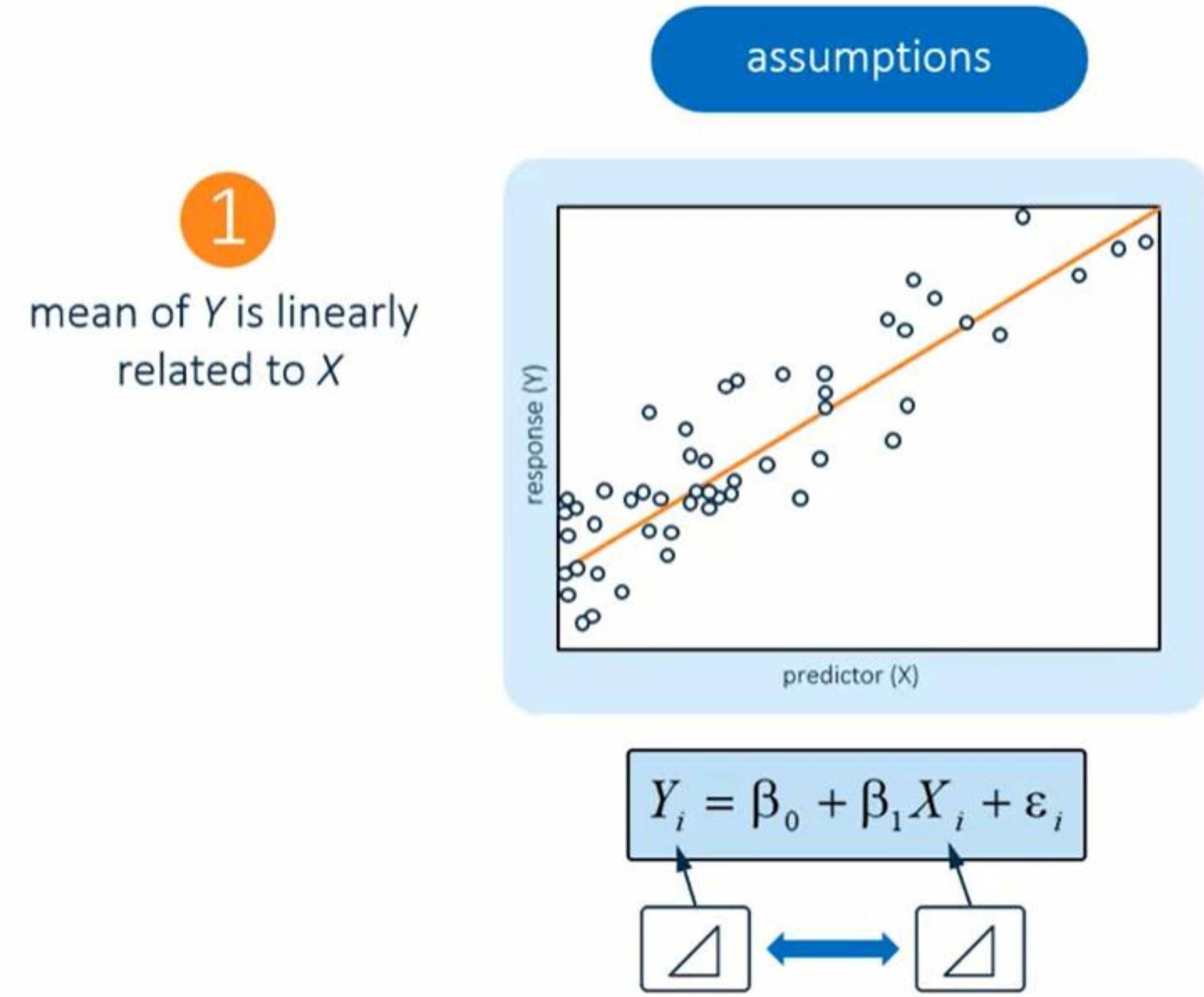
Type of Variability	Equation
Explained (SSM)	$\sum\left(\hat{Y}_i - \bar{Y}\right)^2$
Unexplained (SSE)	$\sum\left(Y_i - \hat{Y}_i\right)^2$
Total	$\sum\left(Y_i - \bar{Y}\right)^2$

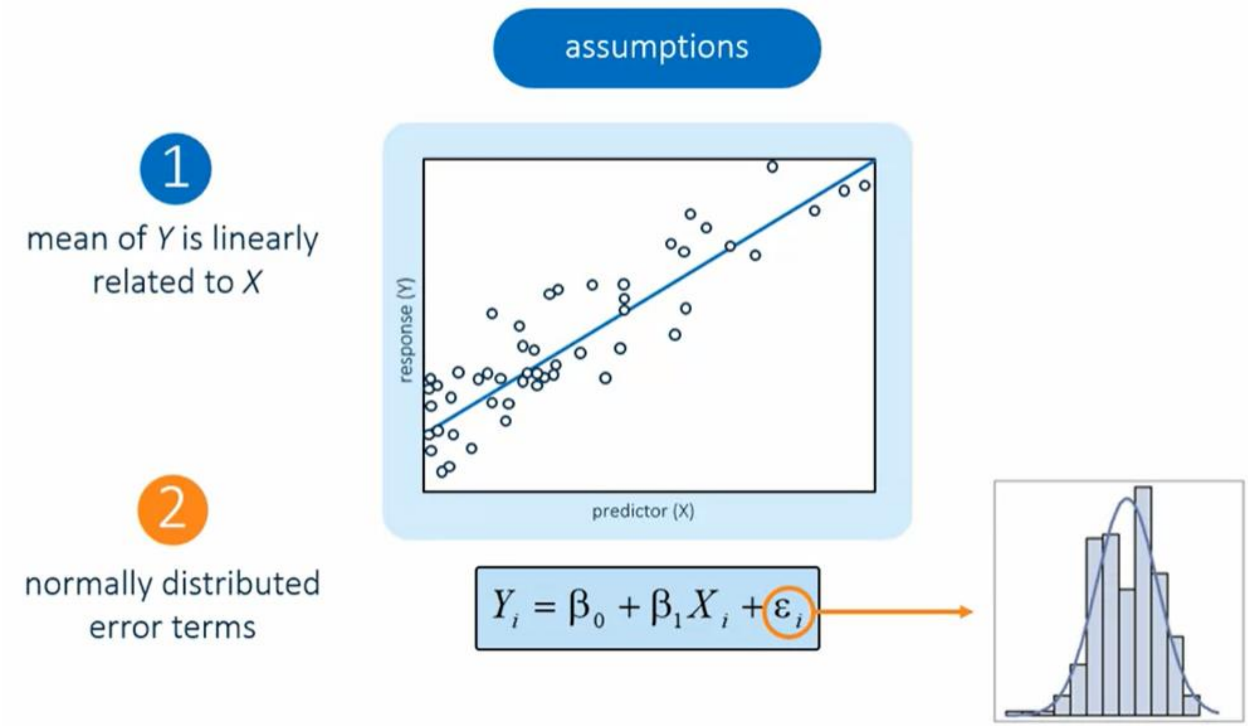

baseline model
sponse (Y)
significance
F value
p-value
Type of Variability
Equation
Explained (SSM)
$\sum\left(\hat{Y}_i-\bar{Y}\right)^2$
Unexplained (SSE)
$\sum\left(Y_i-\hat{Y}_i\right)^2$
Total
$\sum\left(Y_i-\bar{Y}\right)^2$
÷
degrees of freedom
mean-square model (MSM)
mean-square error (MSE)

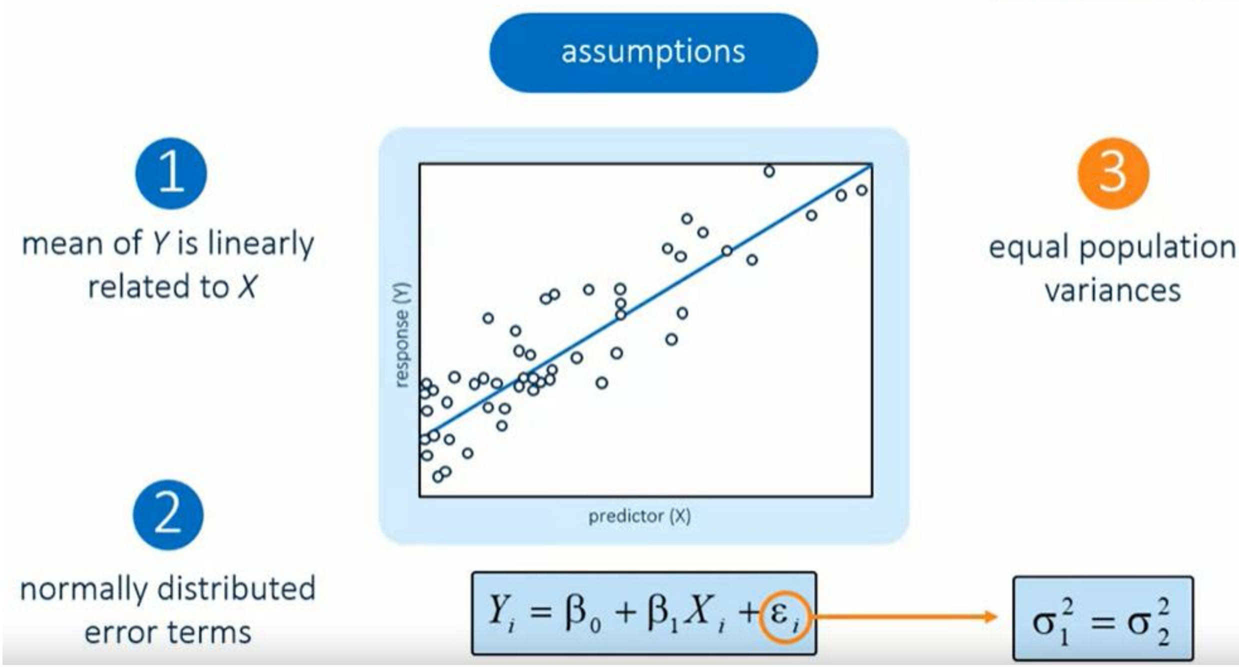
Hypothesis Testing and Assumptions for Linear Regression

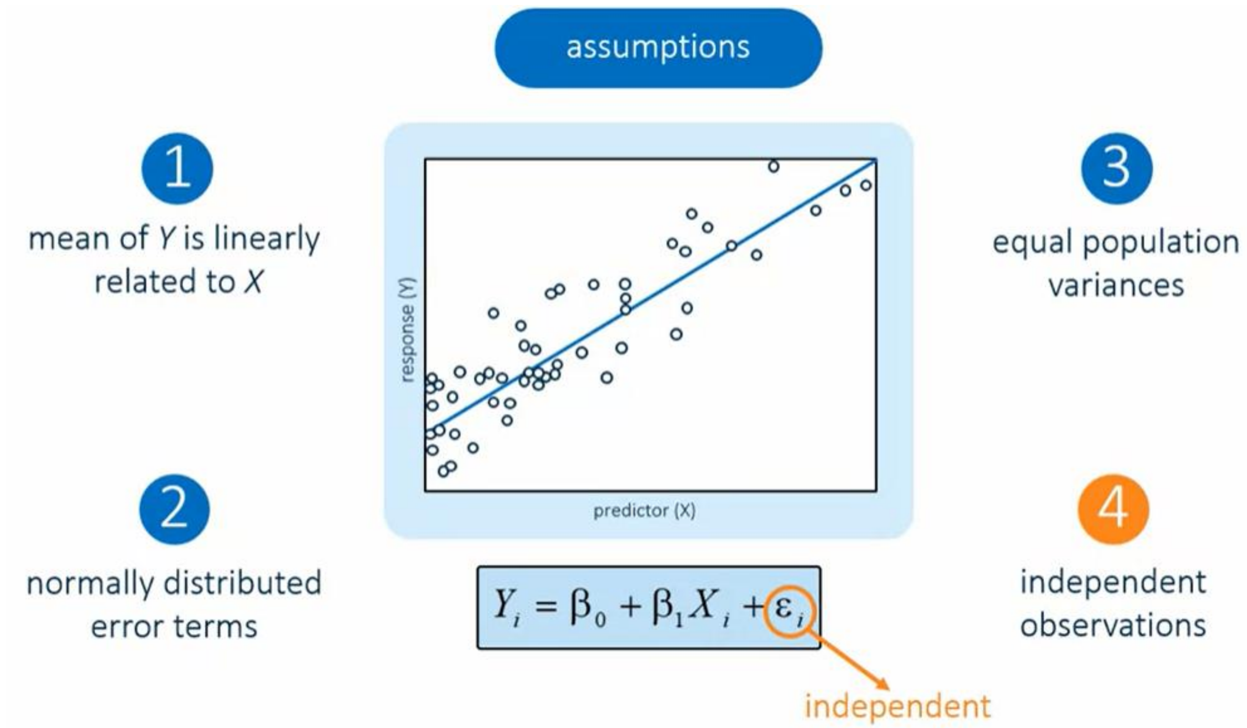



$H_0 : \beta_1 = 0$
$H_a : \beta_1 \neq 0$
response (Y)
predictor (X)
$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

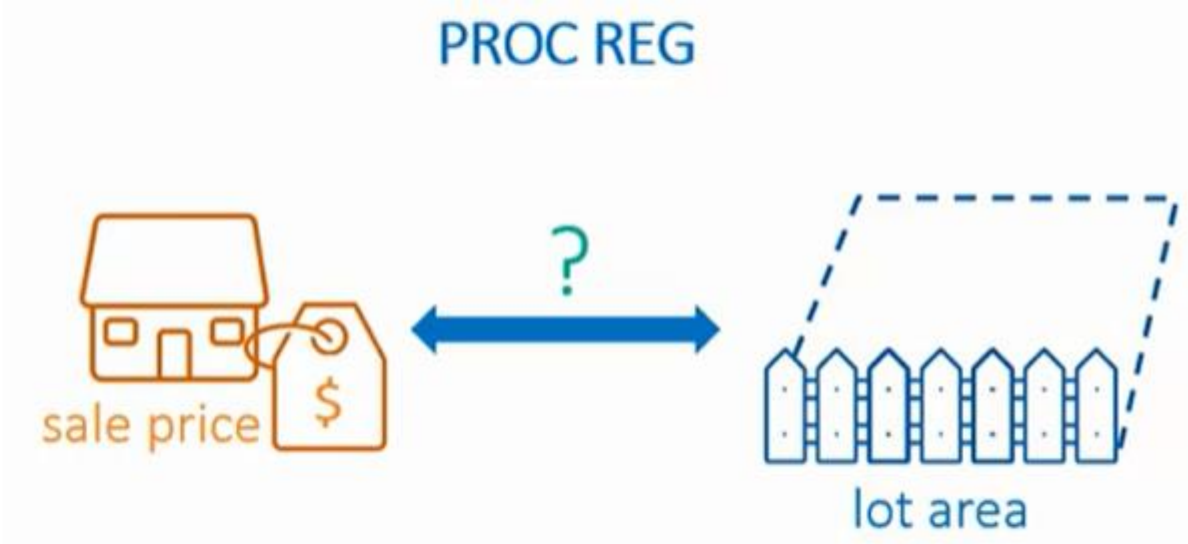

assumptions
1
mean of Y is linearly related to X
response (Y)
predictor (X)
$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$


assumptions
1
mean of Y is linearly related to X
response (Y)
predictor (X)
2
normally distributed error terms
$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$


assumptions
1
mean of Y is linearly related to X
response (Y)
predictor (X)
3
equal population variances
2
normally distributed error terms
$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
$\sigma_1^2 = \sigma_2^2$


assumptions
1
mean of Y is linearly related to X
2
normally distributed error terms
3
equal population variances
4
independent observations
response (Y)
predictor (X)
$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
independent

Demo Performing Simple Linear Regression Using PROC REG




```
/*st102d05.sas*/
ods graphics;

proc reg data=STAT1.ameshousing3;
    model SalePrice=Lot_Area;
    title "Simple Regression with Lot Area as Regressor";
run;
quit;

title;
```

```
PROC REG DATA=SAS-data-set <options>;
    MODEL dependents = <regressors> </ options>;
RUN;
```

Simple Regression with Lot Area as Regressor

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice Sale price in dollars

Number of Observations Read	300
Number of Observations Used	300

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	27164711173	27164711173	20.44	<.0001
Error	298	3.960588E11	1329056404		
Corrected Total	299	4.232235E11			

Root MSE	36456	R-Square	0.0642
Dependent Mean	137525	Adj R-Sq	0.0610
Coeff Var	26.50882		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	113740	5666.48352	20.07	<.0001
Lot_Area	Lot size in square feet	1	2.86770	0.63431	4.52	<.0001

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	113740	5666.48352	20.07	<.0001
Lot_Area	Lot size in square feet	1	2.86770	0.63431	4.52	<.0001

SalePrice = 113740 + 2.86770 * **Lot_Area**

```
/*st102d05.sas*/
ods graphics;

proc reg data=STAT1.ameshousing3;
   model SalePrice=Lot_Area;
   title "Simple Regression with Lot Area as Regressor";
run;
quit;

title;
```

Simple Regression with Lot Area as Regressor

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice Sale price in dollars

Number of Observations Read	300
Number of Observations Used	300

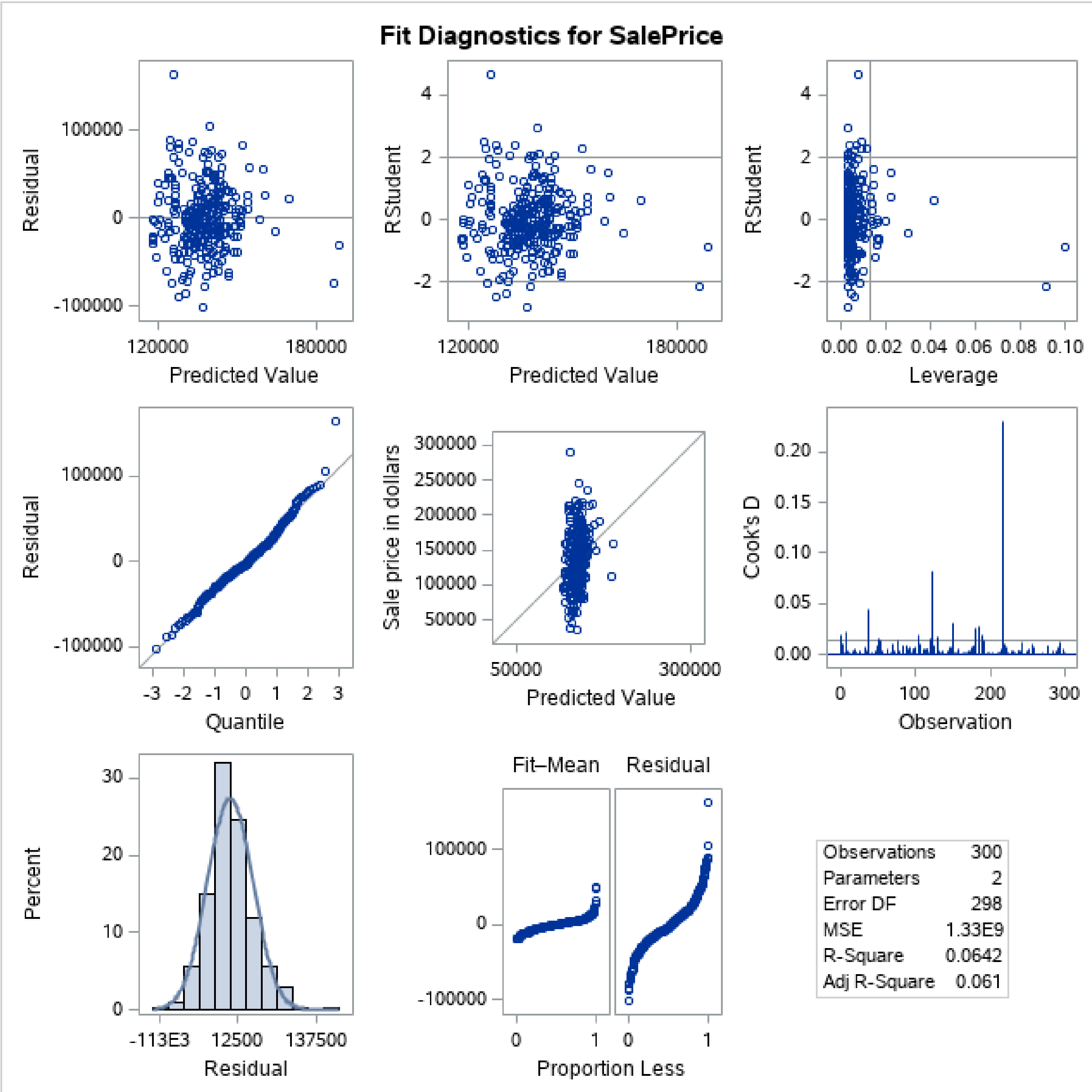
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	27164711173	27164711173	20.44	<.0001
Error	298	3.960588E11	1329056404		
Corrected Total	299	4.232235E11			

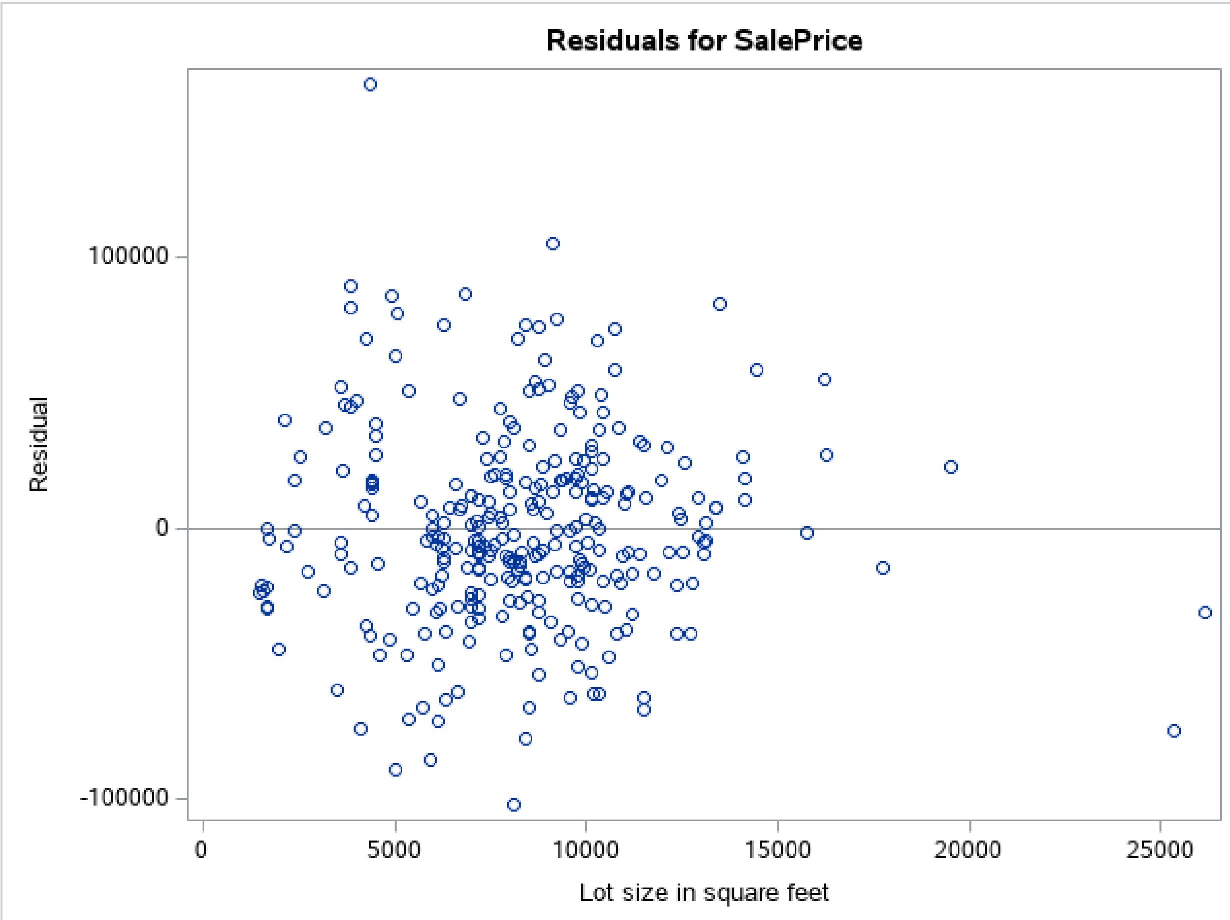
Root MSE	36456	R-Square	0.0642
Dependent Mean	137525	Adj R-Sq	0.0610
Coeff Var	26.50882		

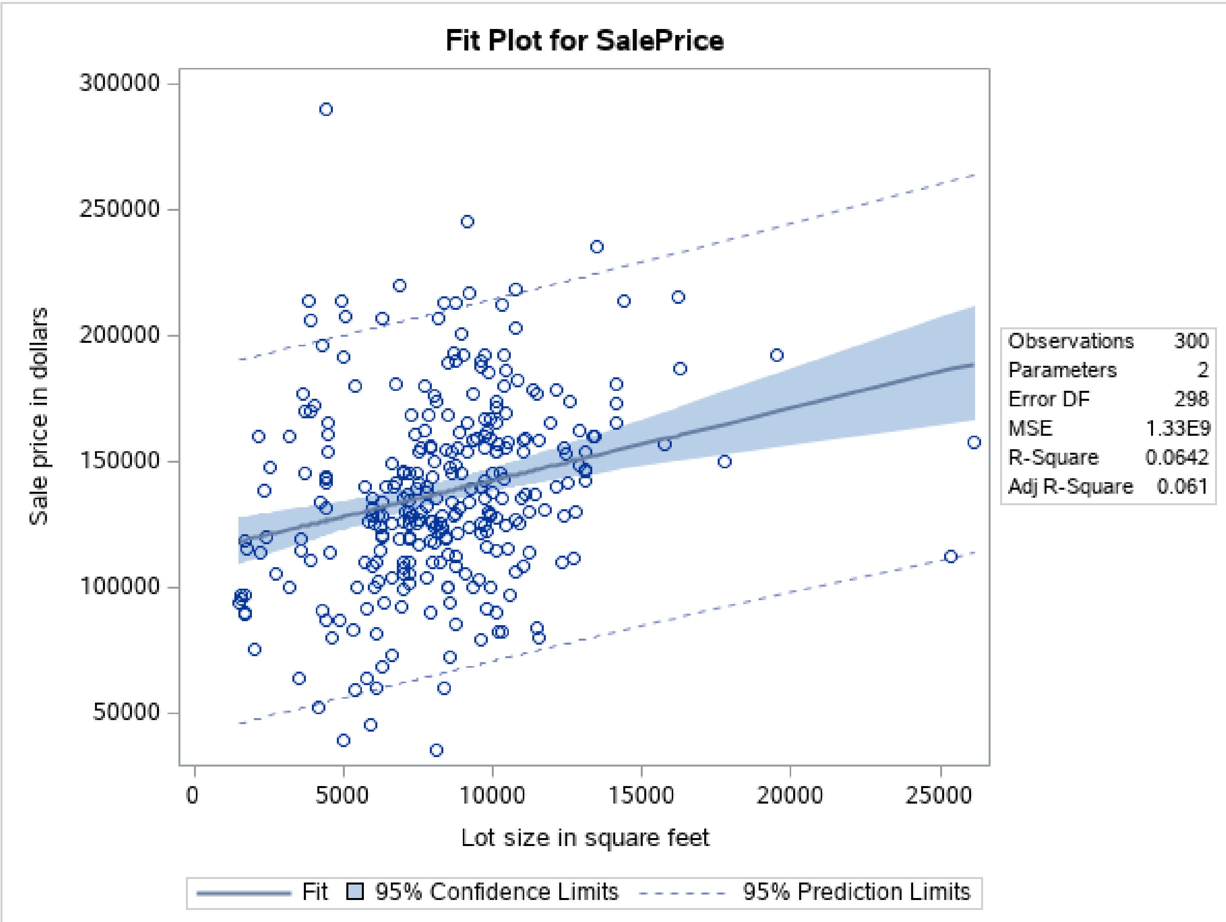
Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	113740	5666.48352	20.07	<.0001
Lot_Area	Lot size in square feet	1	2.86770	0.63431	4.52	<.0001

Simple Regression with Lot Area as Regressor

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice Sale price in dollars




Residuals for SalePrice
100000
0
-100000
Residual
0
5000
10000
15000
20000
25000
Lot size in square feet



Question 1

You just used PROC REG to regress y on X_1 and found the parameter estimates table below. Given this information, what is the best guess (predicted value) of y when X_1 = 13?

Parameter Estimates				
Variable	DF	Parameter Estimate	Standard Error	t Value
Intercept	1	5	1.0	5
X_1	1	10	2.5	4

The best guess of y when X_1 = 13 is the intercept plus the slope times X_1:

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta} X_1$$

$$135 = 5 + 10(13)$$

Practice - Using PROC REG to Fit a Simple Linear Regression Model

TOTAL POINTS 3

1.

Question 1

Using the **bodyfat2** data set, perform a simple linear regression model.

- Perform a simple linear regression model with **PctBodyFat2** as the response variable and **Weight** as the predictor.

What is the value of the *F* statistic and the associated *p*-value? How would you interpret this in connection with the null hypothesis?

F value = 150 and p-value < .001 Reject null hypothesis

The value of the *F* statistic is 150.03 and the *p*-value is <.001. Therefore, you would reject the null hypothesis of no relationship, or a zero slope for **Weight**.

Question 2

Write the predicted regression equation.

y = -12.05158 + 0.17439 X

The prediction regression equation is:

PctBodyFat2 = -12.05158 + 0.17439 * **Weight**.

Question 3

What is the value of R-square? How would you interpret this?

The R-square value of 0.3751 can be interpreted to mean that 37.51% of the variability in **PctBodyFat2** can be explained by **Weight**.

```
/*st102s04.sas*/
ods graphics on;

proc reg data=STAT1.BodyFat2;
```

```
   model PctBodyFat2=Weight;
   title "Regression of % Body Fat on Weight";
run;
quit;

title;
```

Regression of % Body Fat on Weight

The REG Procedure
Model: MODEL1
Dependent Variable: PctBodyFat2

Number of Observations Read	252
Number of Observations Used	252

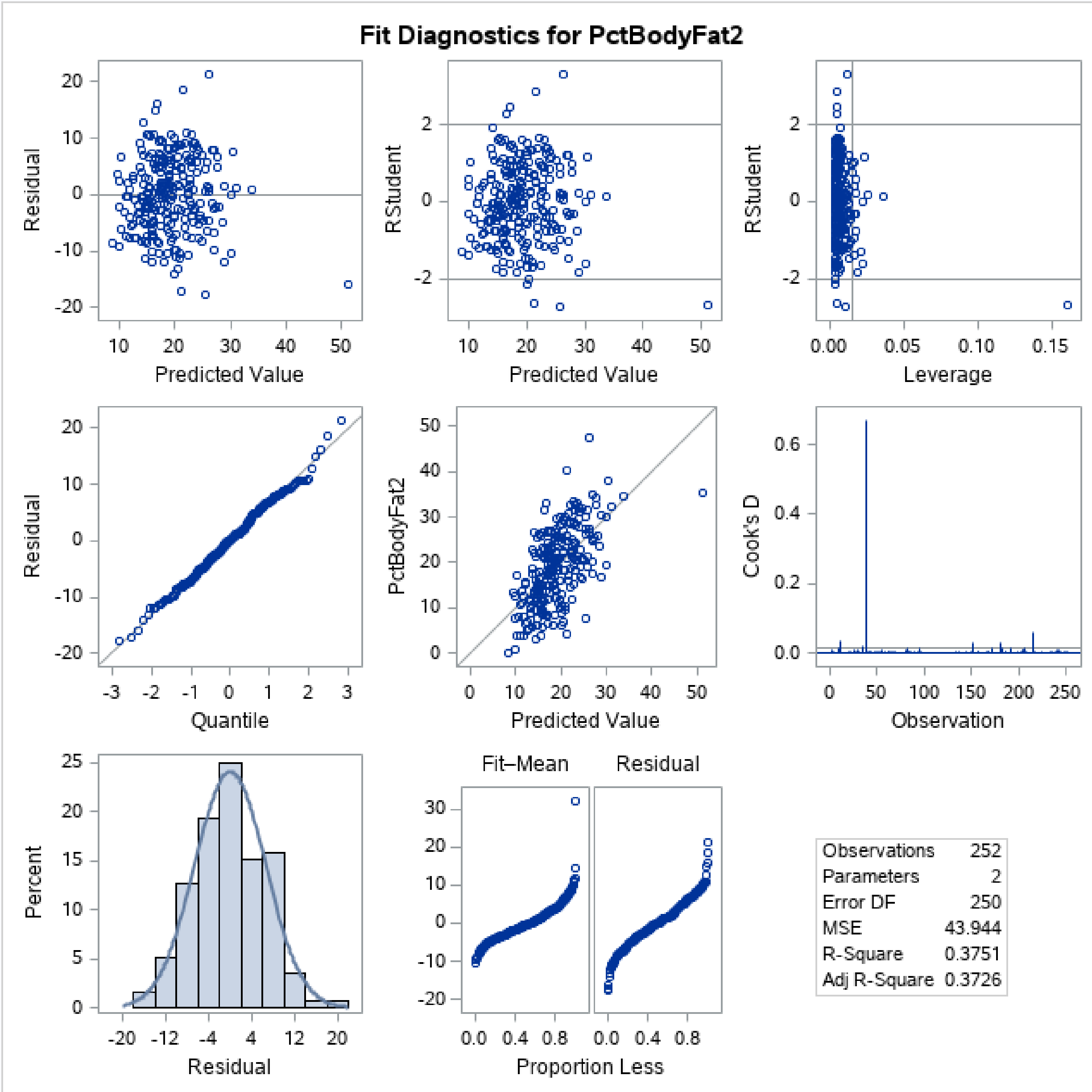
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6593.01614	6593.01614	150.03	<.0001
Error	250	10986	43.94389		
Corrected Total	251	17579			

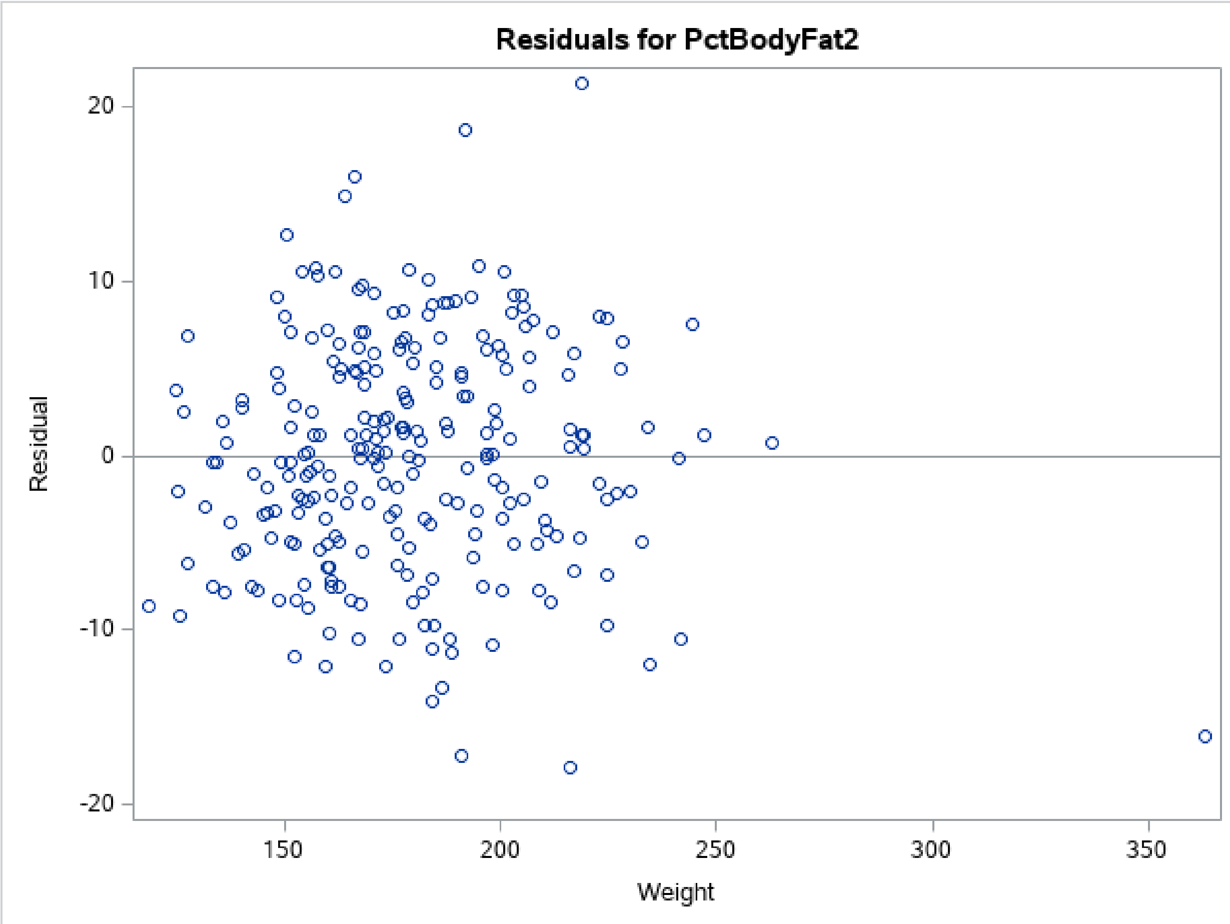
Root MSE	6.62902	R-Square	0.3751
Dependent Mean	19.15079	Adj R-Sq	0.3726
Coeff Var	34.61485		

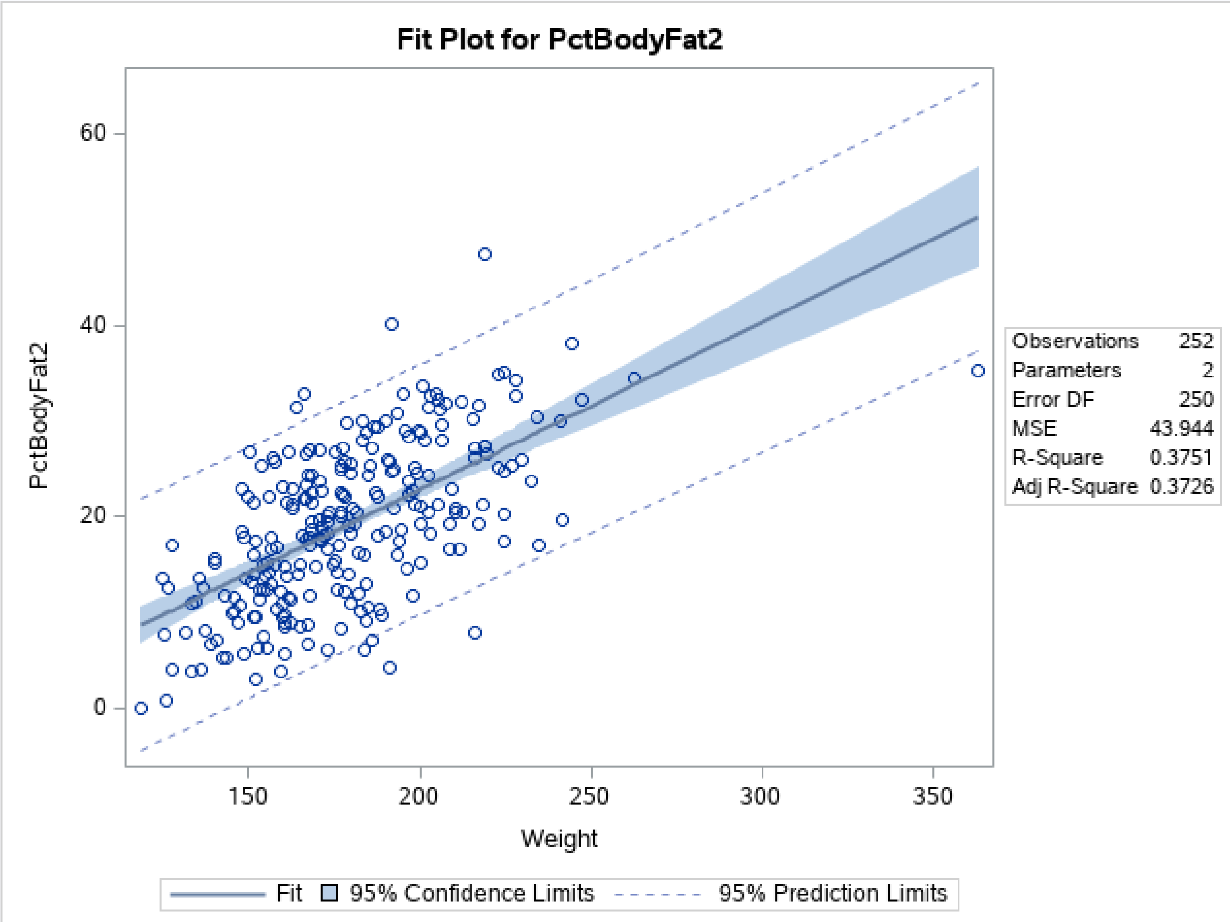
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-12.05158	2.58139	-4.67	<.0001
Weight	1	0.17439	0.01424	12.25	<.0001

Regression of % Body Fat on Weight

The REG Procedure
Model: MODEL1
Dependent Variable: PctBodyFat2




Residuals for PctBodyFat2
Residual
Weight
20
10
0
-10
-20
150
200
250
300
350



Fit Plot for PctBodyFat2
PctBodyFat2
60
40
20
0
150
200
250
300
350
Weight
Observations 252
Parameters 2
Error DF 250
MSE 43.944
R-Square 0.3751
Adj R-Square 0.3726
Fit
95% Confidence Limits
95% Prediction Limits