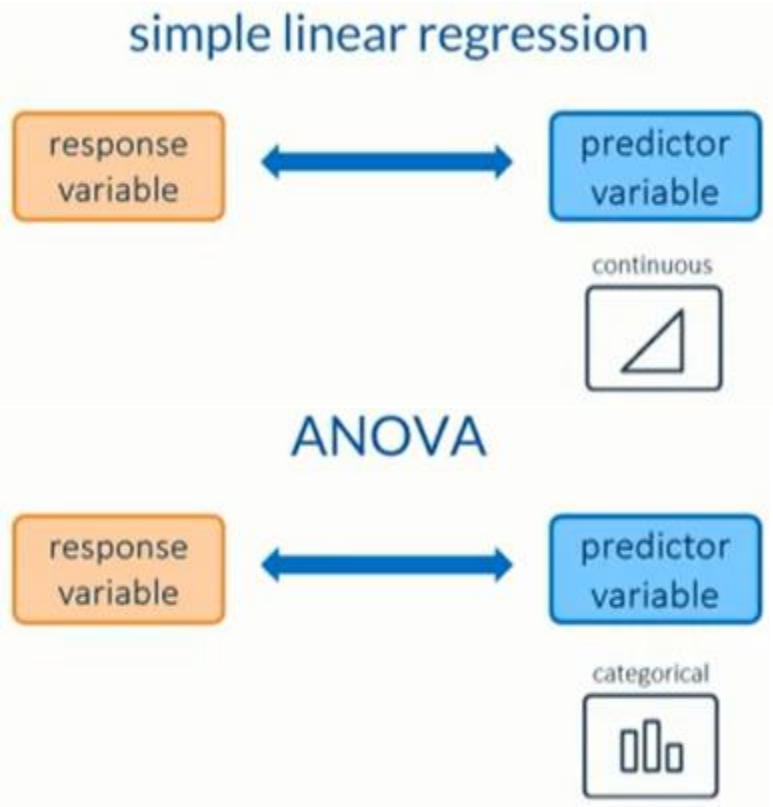
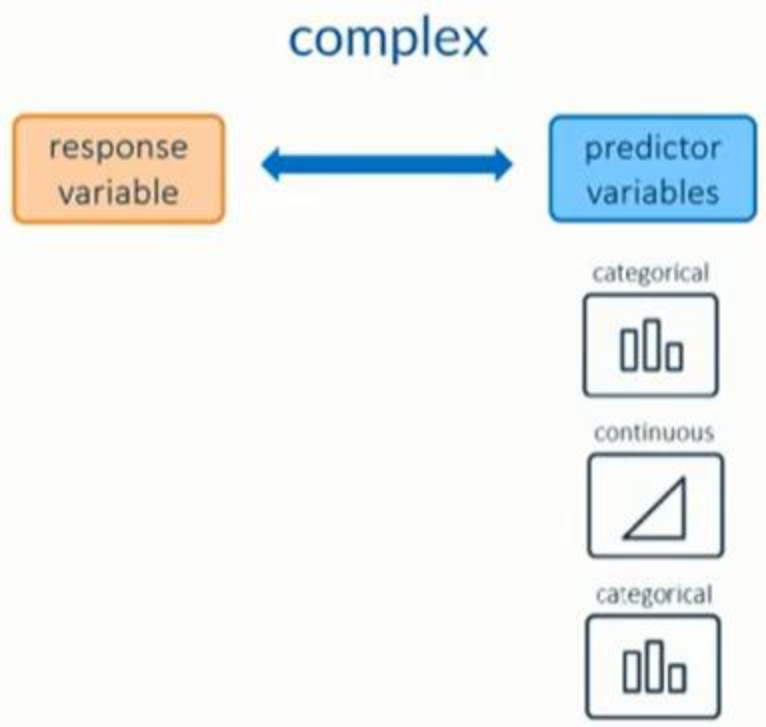
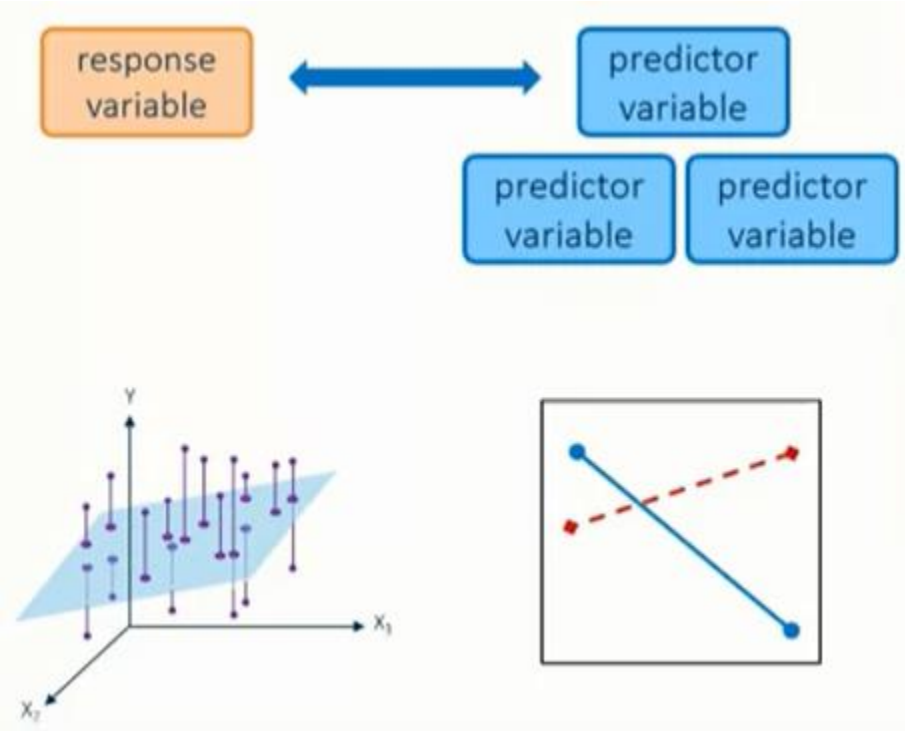
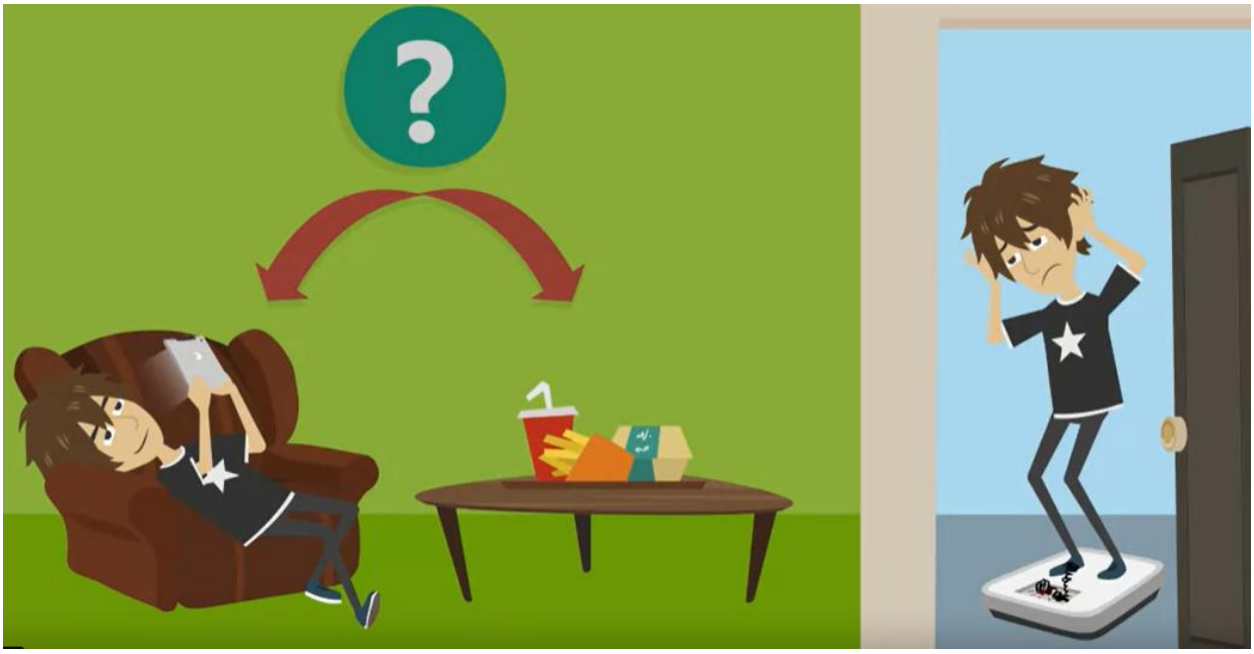
SBA1 Introduction to Statistical Analysis: Hypothesis Testing

Overview

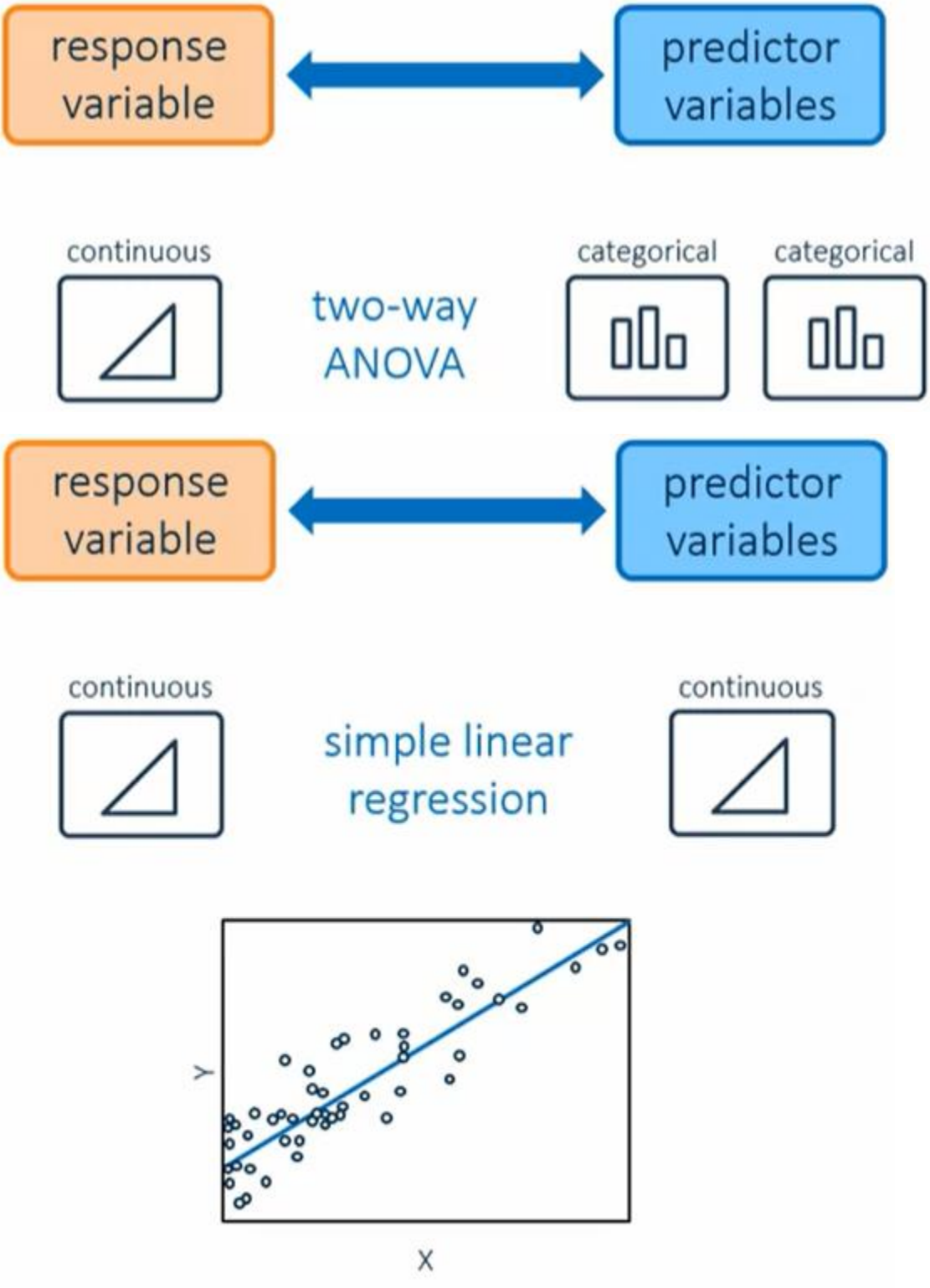


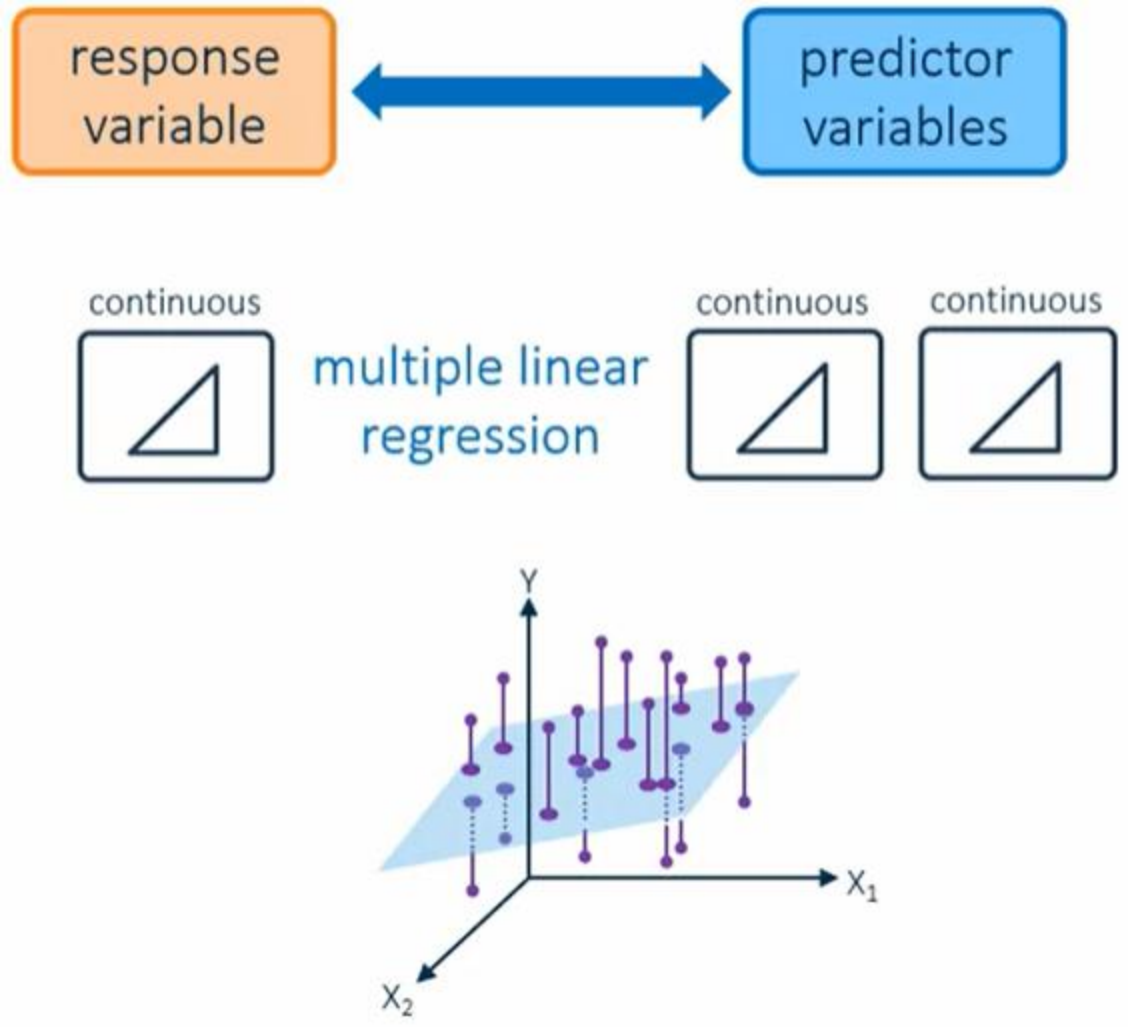

complex
response variable
predictor variables
categorical
continuous
categorical

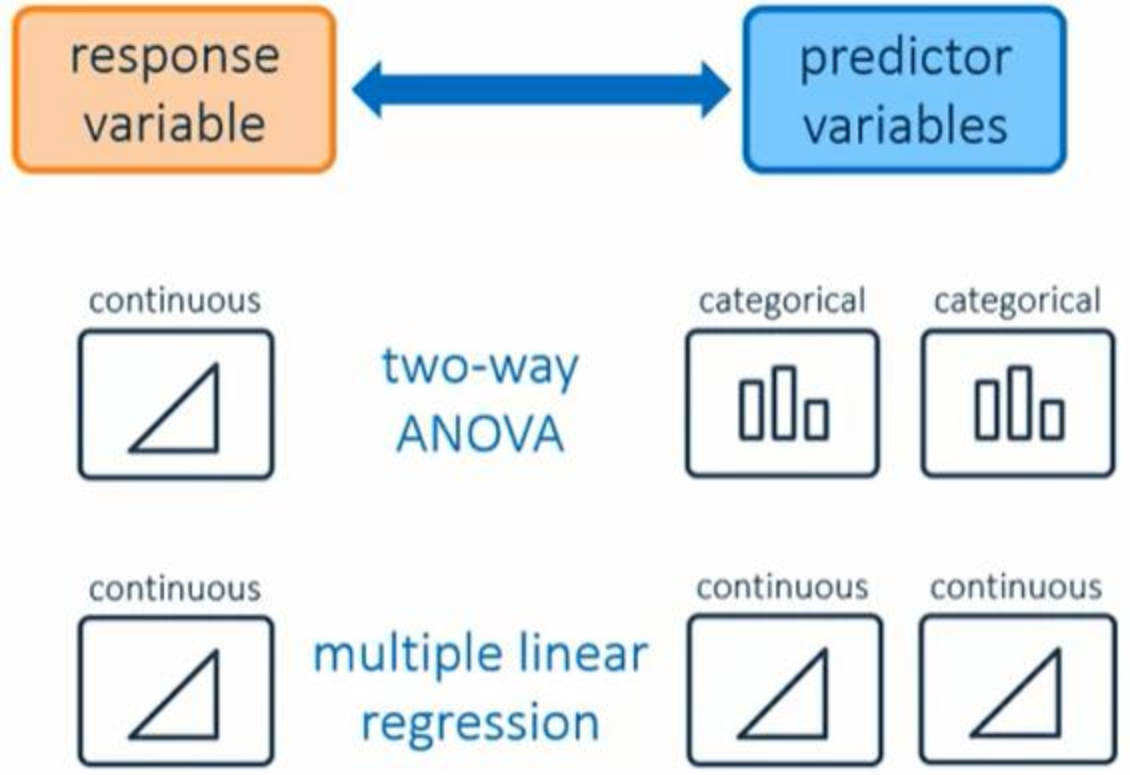

response variable
predictor variable
predictor variable
predictor variable
Y
X1
X2


?

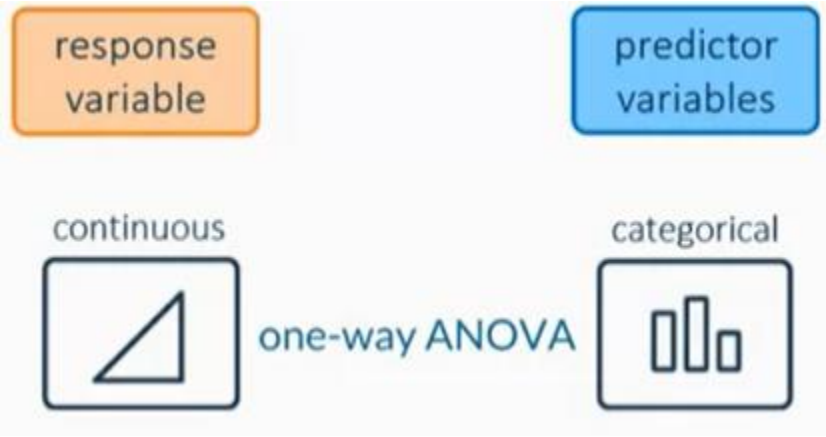

sale price
$
lot area
basement area
heating quality
number of bathrooms

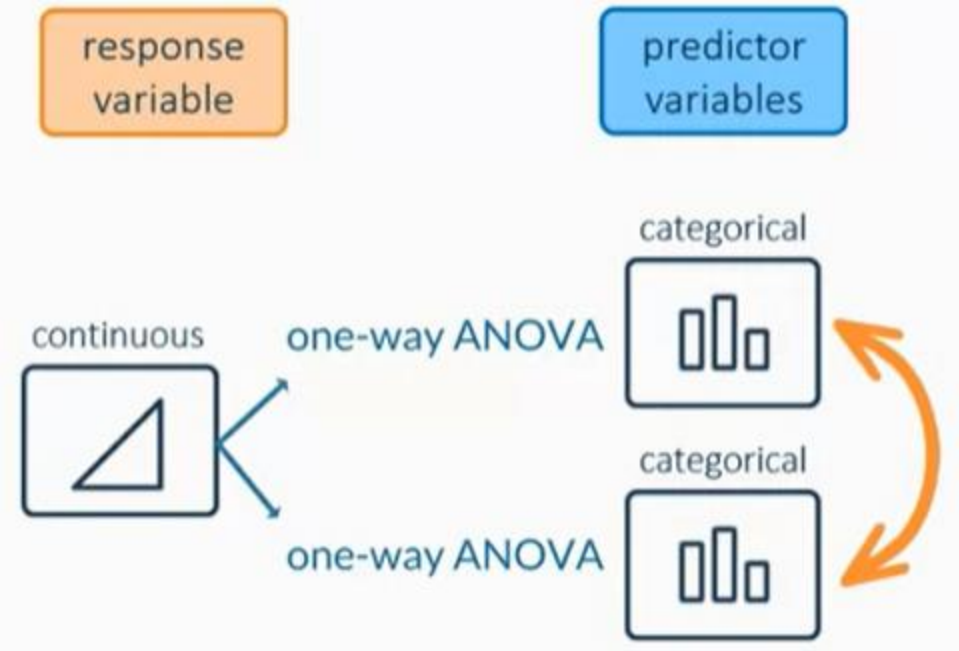

response variable
predictor variables
continuous
two-way ANOVA
categorical
categorical
response variable
predictor variables
continuous
simple linear regression
continuous
Y
X

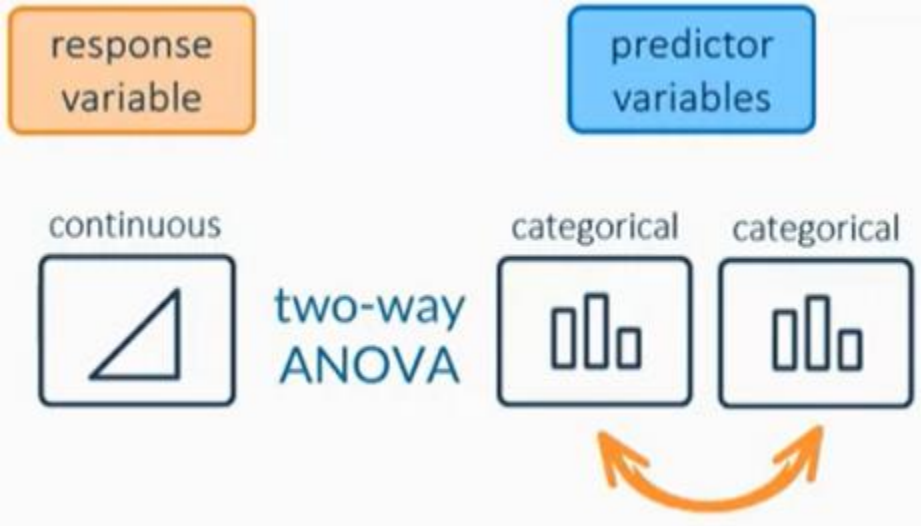

response variable
predictor variables
continuous
multiple linear regression
continuous
continuous
Y
X1
X2



Two-Way ANOVA and Interactions Scenario




response variable
predictor variables
categorical
continuous
one-way ANOVA
categorical
one-way ANOVA


response variable
predictor variables
continuous
two-way ANOVA
categorical
categorical

response variable

predictor variables

continuous

gender

age

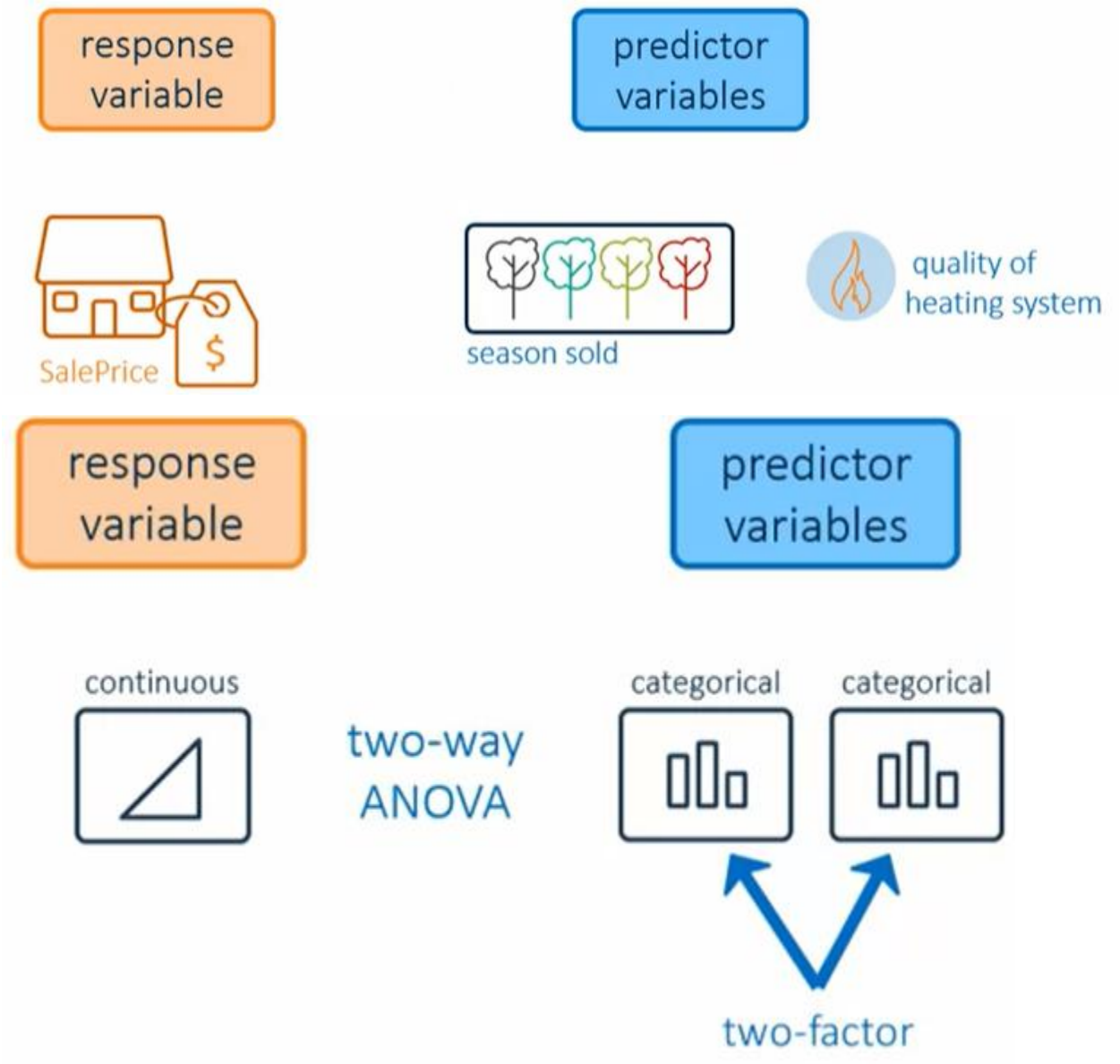
response variable

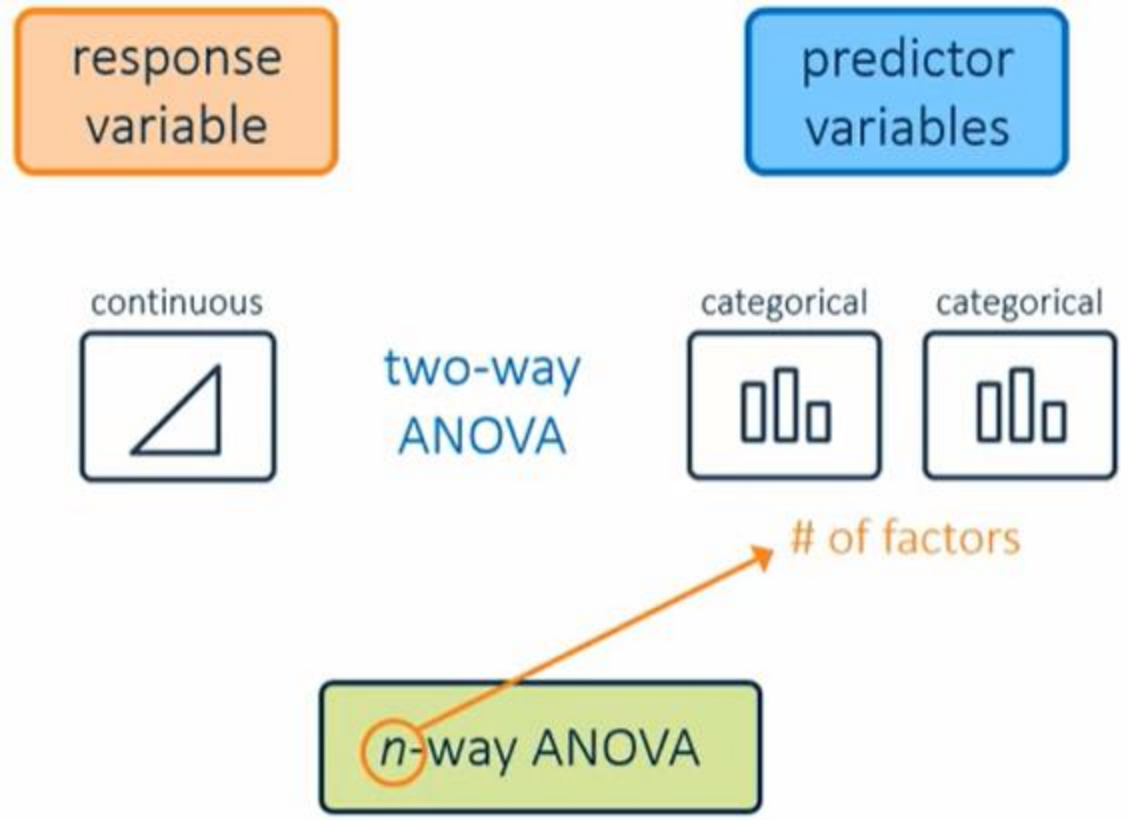
predictor variables

blood pressure

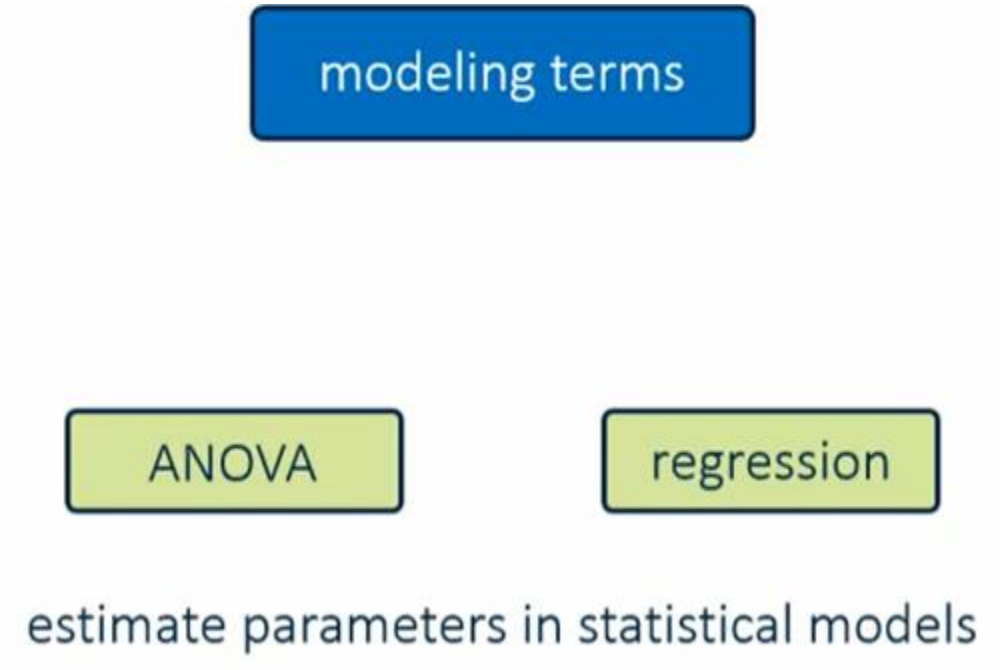
drug dosages

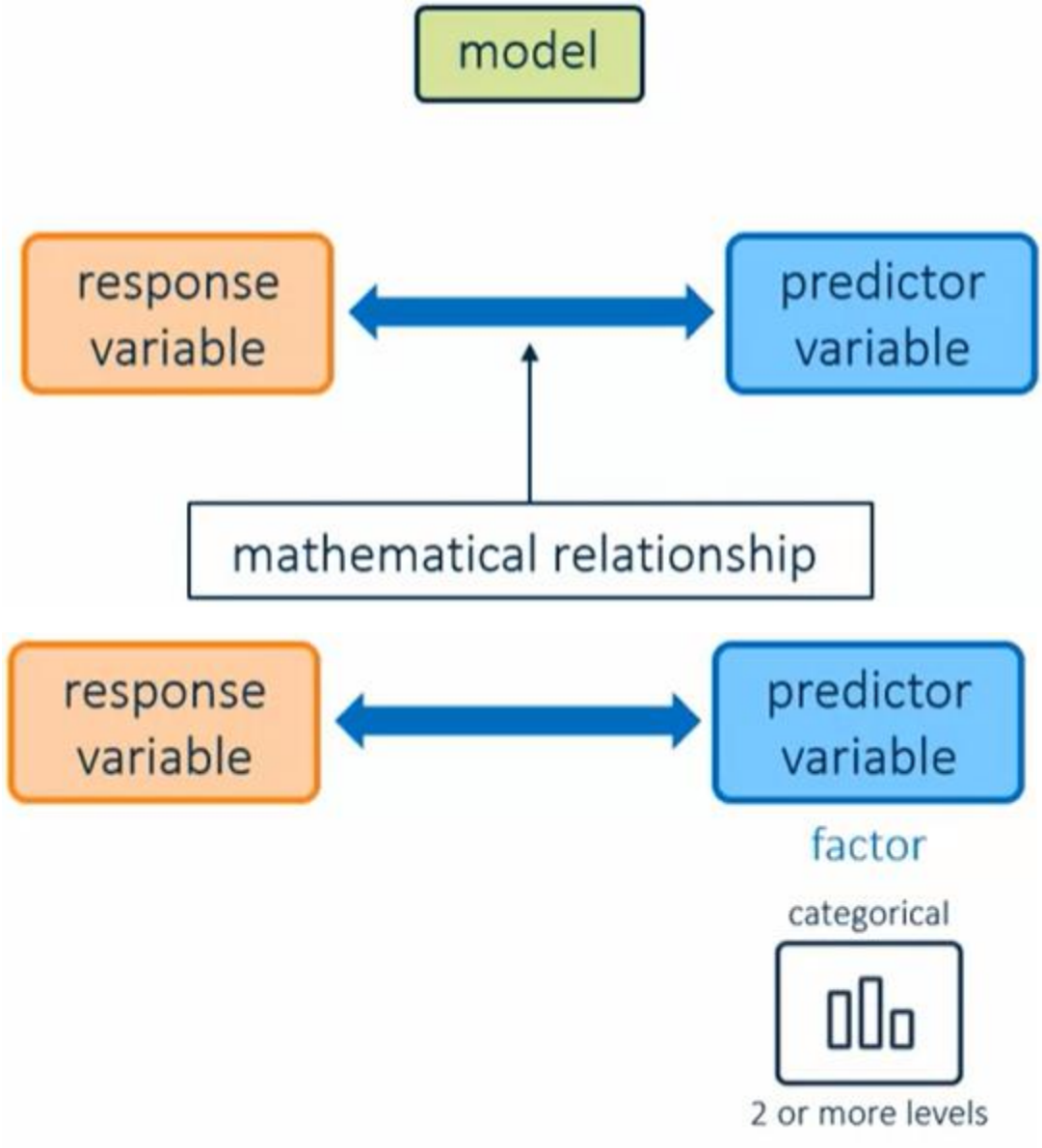
heart disease

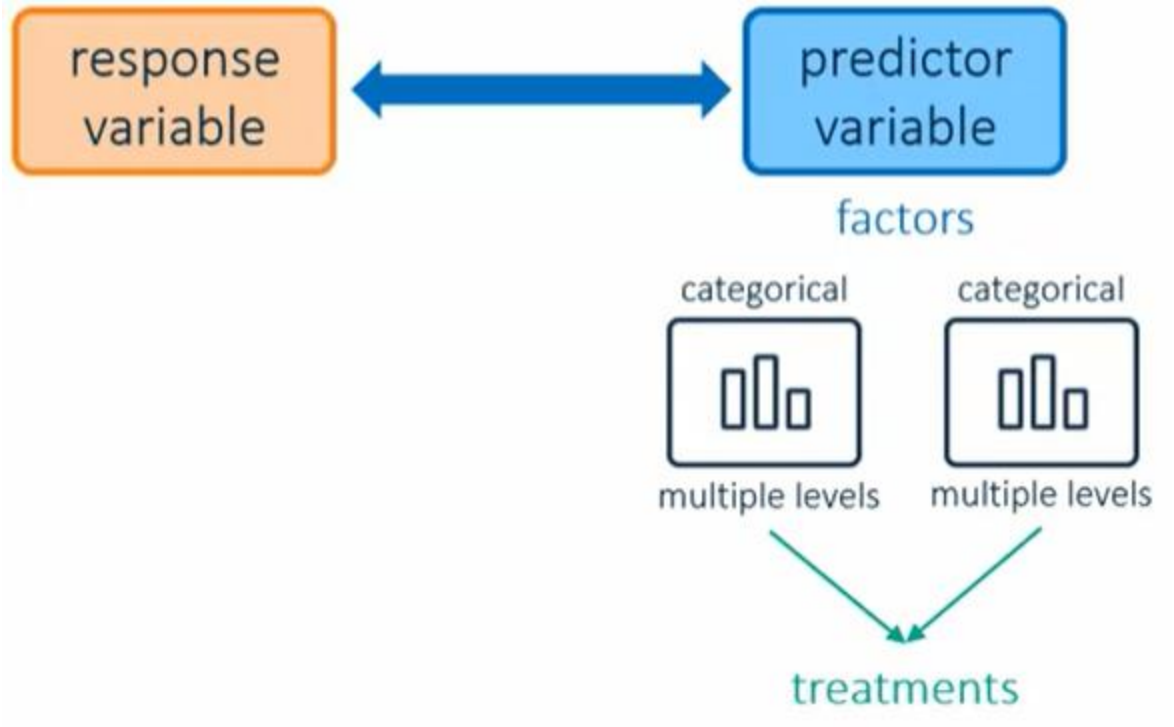

response variable
predictor variables
SalePrice
season sold
quality of heating system
response variable
predictor variables
continuous
two-way ANOVA
categorical
categorical
two-factor

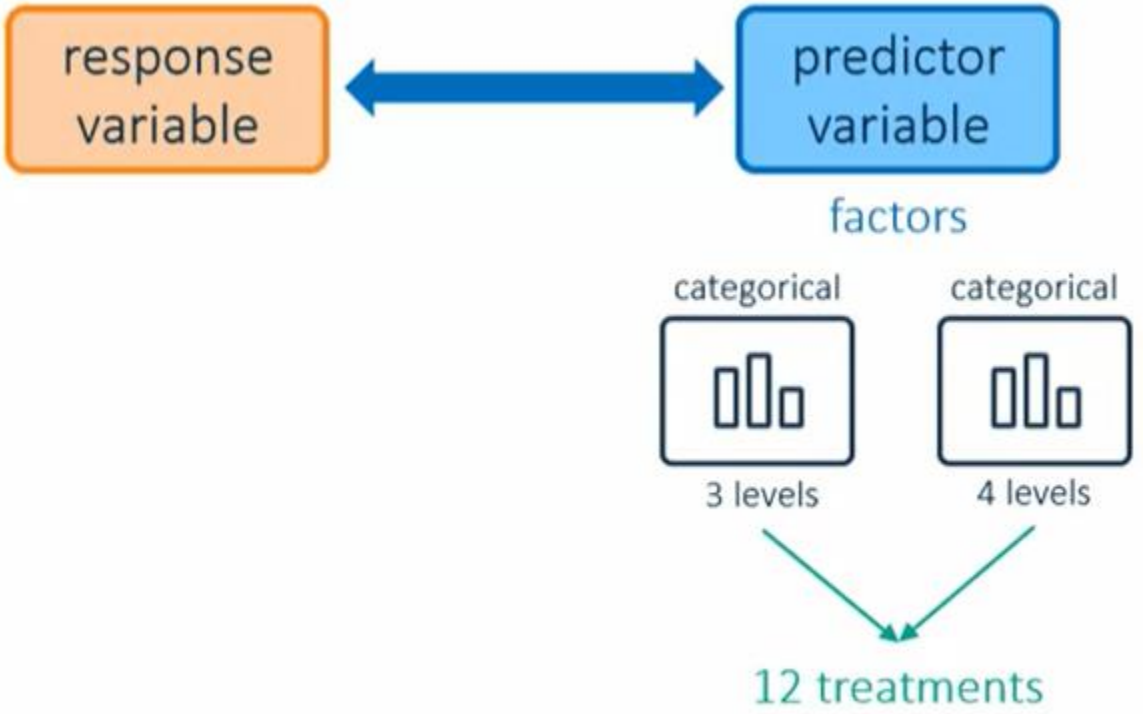


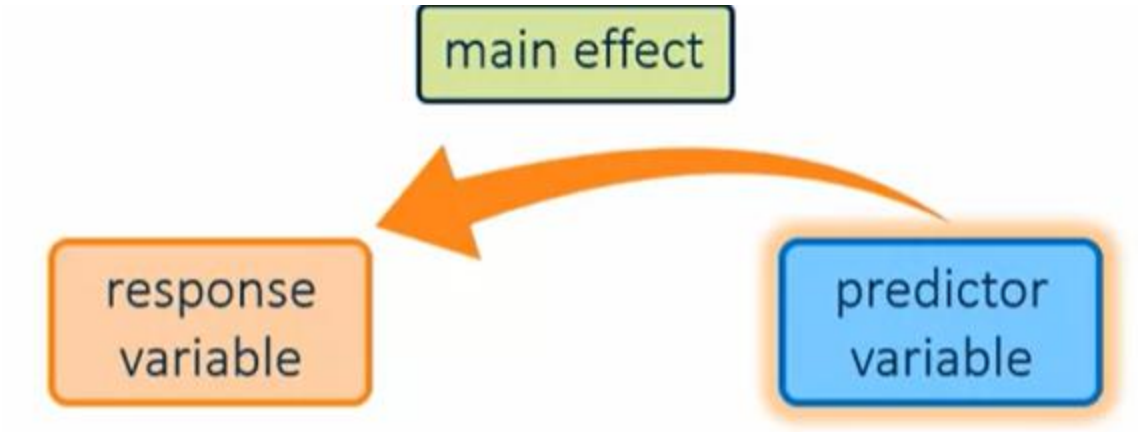
Applying the Two-Way ANOVA Model



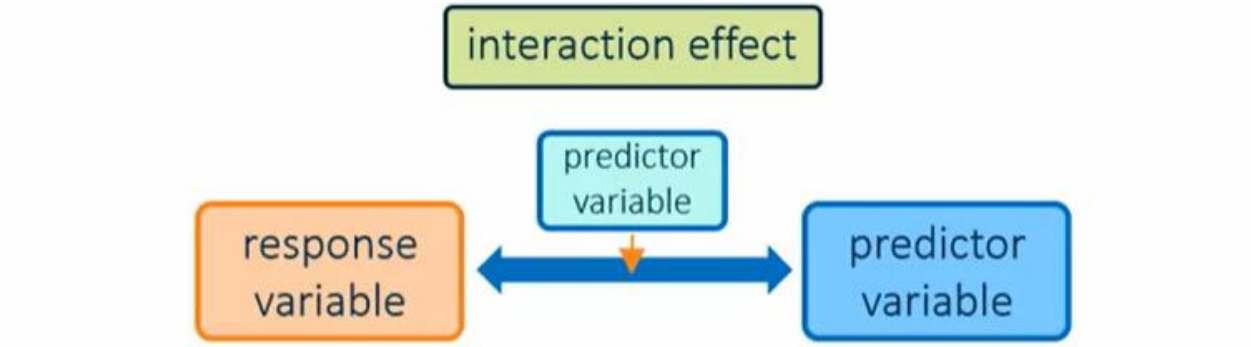

model
response variable
predictor variable
mathematical relationship
response variable
predictor variable
factor
categorical
2 or more levels


response variable
predictor variable
factors
categorical
categorical
multiple levels
multiple levels
treatments

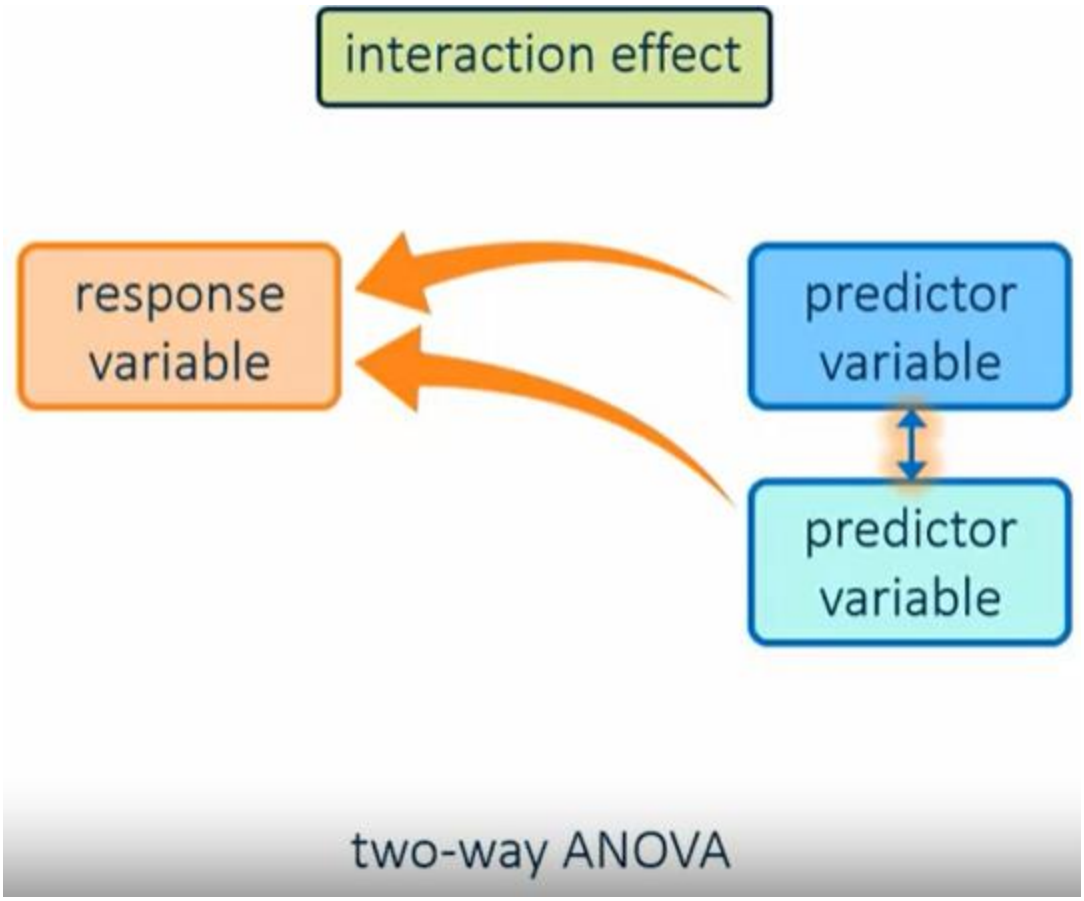

response variable
predictor variable
factors
categorical
categorical
3 levels
4 levels
12 treatments



$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} \cdots$$



$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{1i} X_{2i} \cdots$$

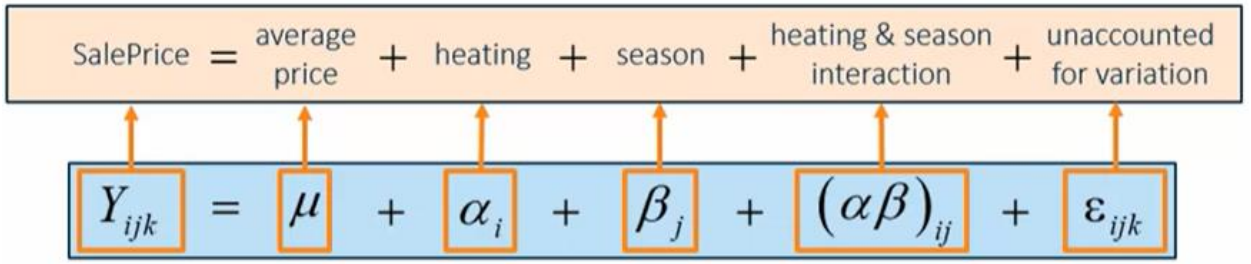


product terms

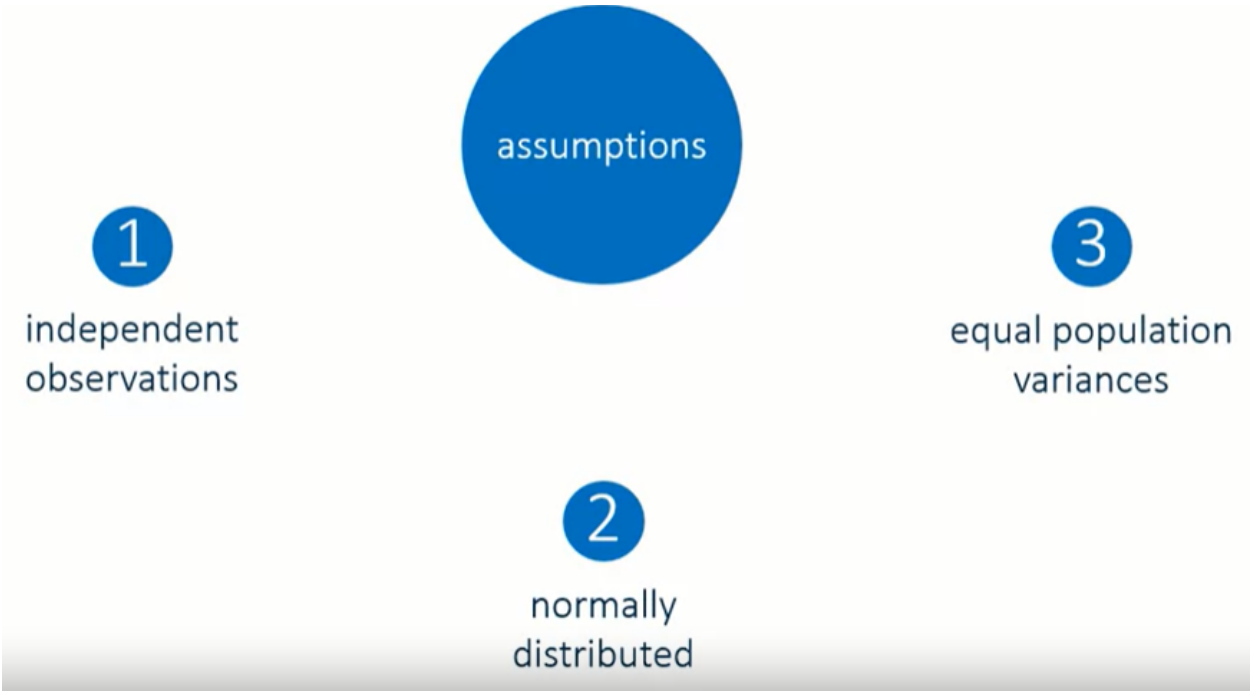
crossed effects

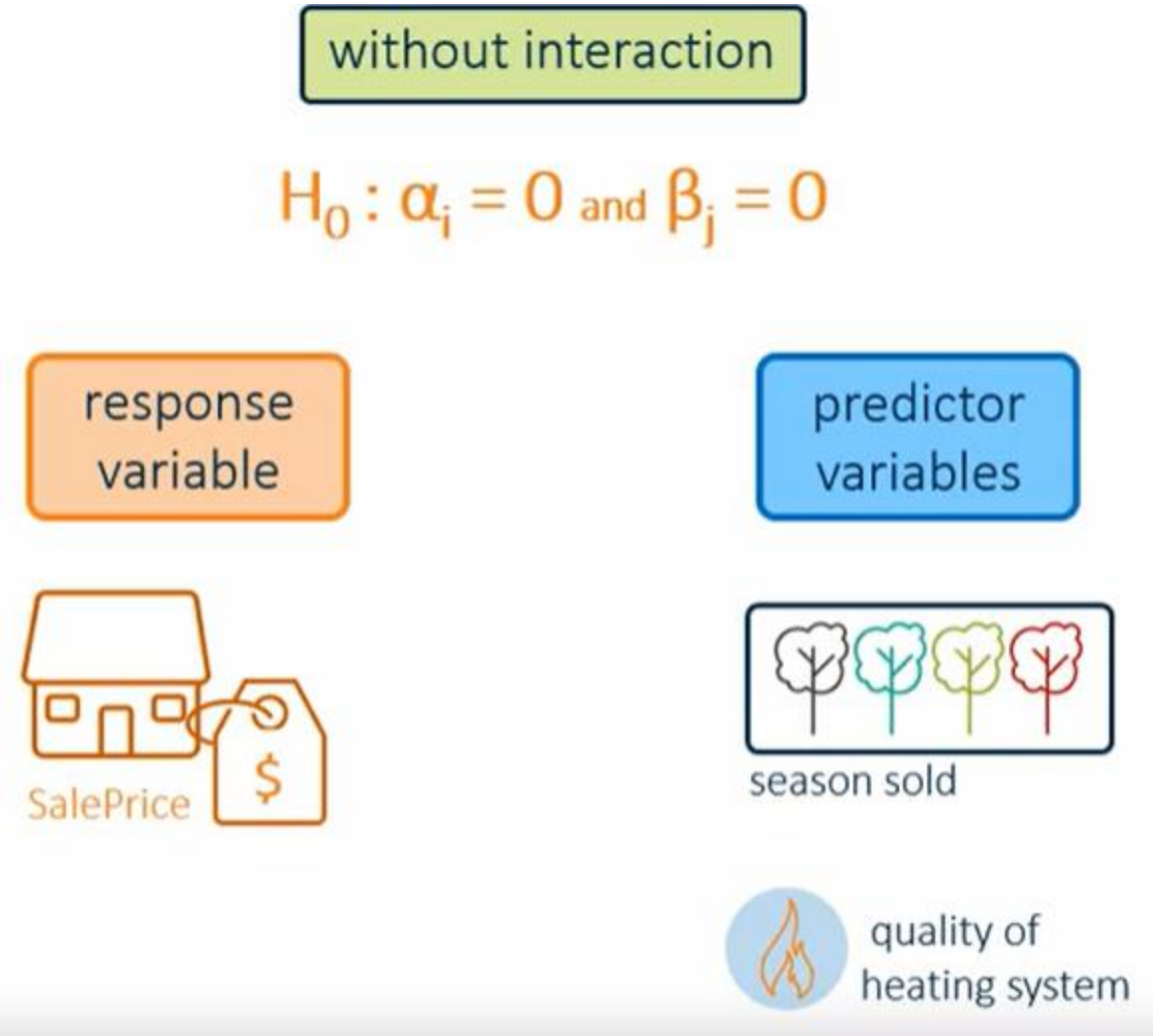
$$\text{SalePrice} = \underset{\text{price}}{\text{average}} + \text{heating} + \text{season} + \underset{\text{interaction}}{\text{heating \& season}}$$

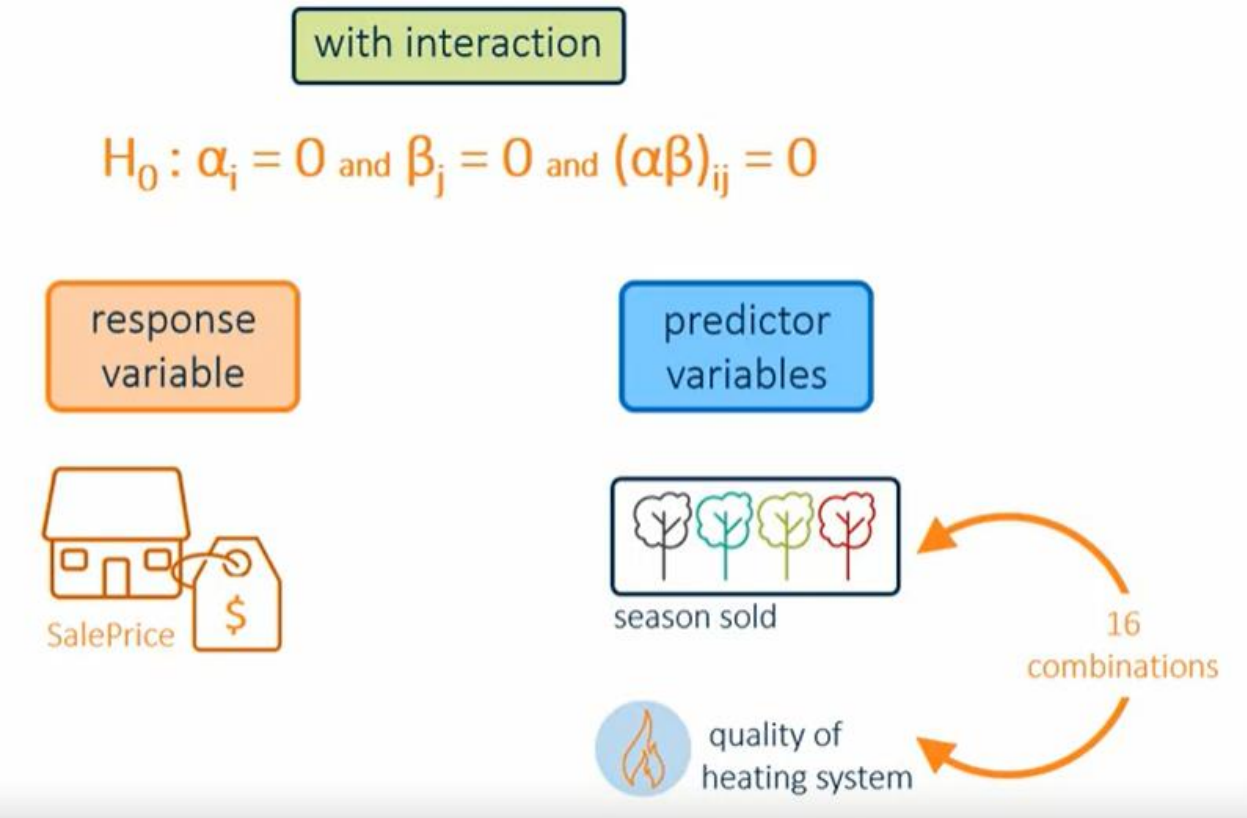
$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$



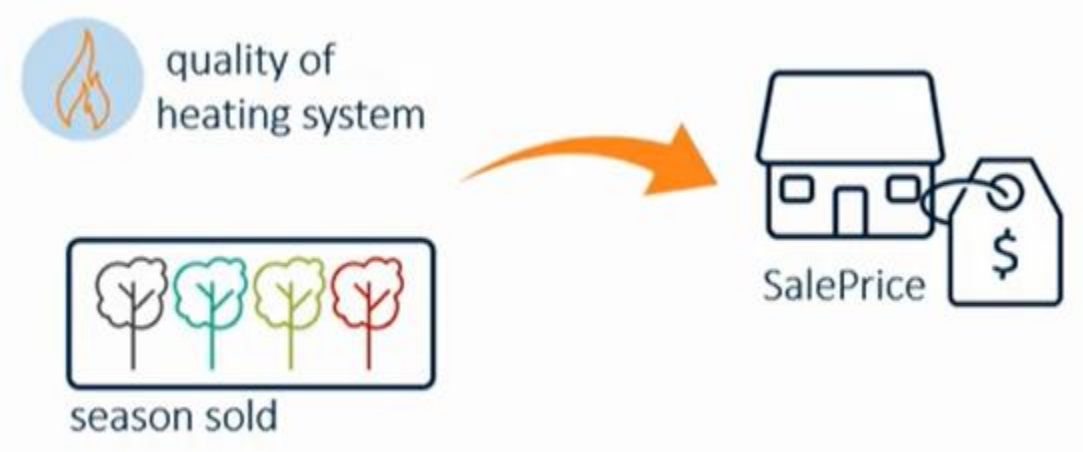
two-way ANOVA




without interaction
$H_0 : \alpha_i = 0$ and $\beta_j = 0$
response variable
predictor variables
SalePrice
$
season sold
quality of heating system


with interaction
$H_0 : \alpha_i = 0$ and $\beta_j = 0$ and $(\alpha\beta)_{ij} = 0$
response variable
predictor variables
SalePrice
$
season sold
16 combinations
quality of heating system

Demo Performing a Two-way ANOVA Using PROC GLM




```
/*st103d01.sas*/  /*Part A*/
ods graphics off;
proc means data=STAT1.ameshousing3
           mean var std nway;
    class Season_Sold Heating_QC;
    var SalePrice;
    format Season_Sold Season.;
    title 'Selected Descriptive Statistics';
run;
```

PROC MEANS DATA=*SAS-data-set <statistic-keyword(s)>***;**
CLASS *variable(s) < / option(s)>***;**
VAR *variable(s)***;**
RUN;

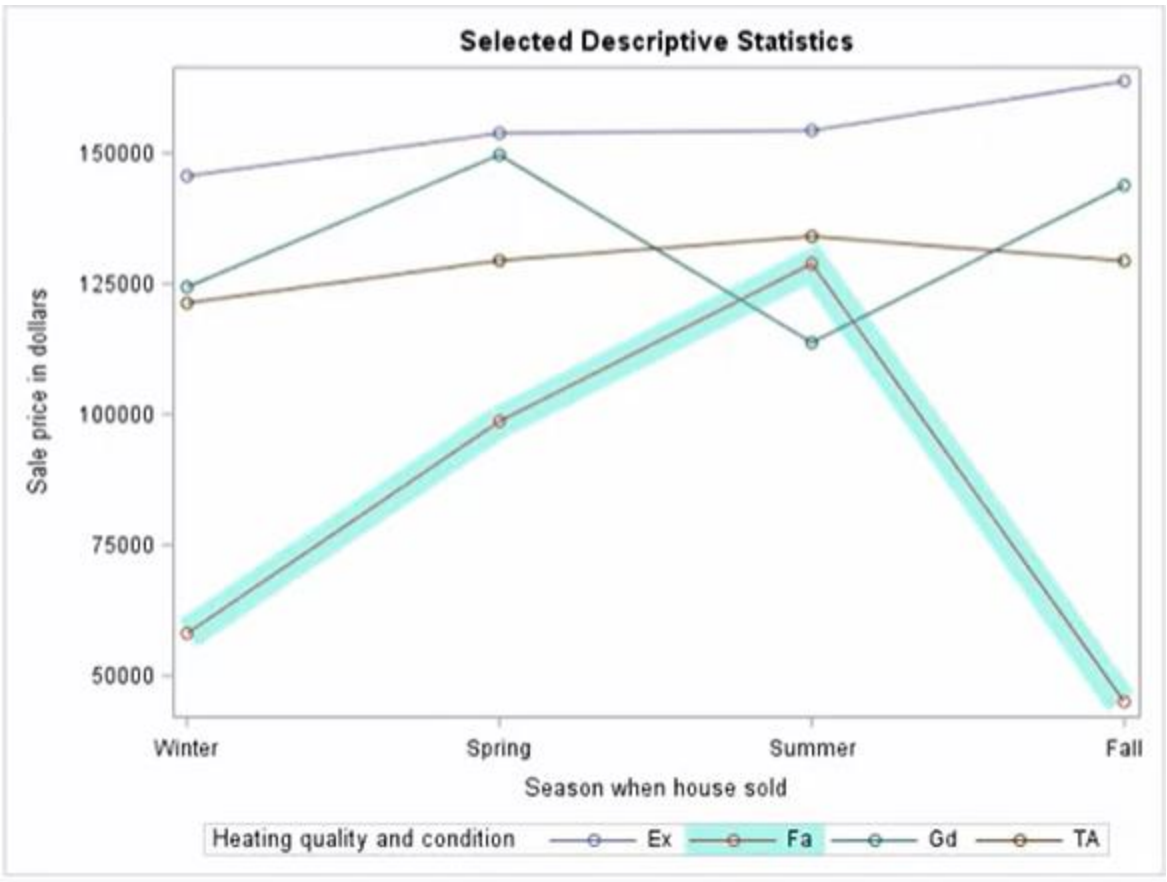
```
/*st103d01.sas*/  /*Part B*/
proc sgplot data=STAT1.ameshousing3;
    vline Season_Sold / group=Heating_QC
                        stat=mean
                        response=SalePrice
                        markers;
    format Season_Sold season.;
run;
```

PROC SGPLOT DATA=*SAS-data-set < option(s)>***;**
VLINE *category-variable < / option(s)>***;**
RUN;

Selected Descriptive Statistics

The MEANS Procedure

Analysis Variable : SalePrice Sale price in dollars					
Season when house sold	Heating quality and condition	N Obs	Mean	Variance	Std Dev
Winter	Ex	6	145583.33	1579141667	39738.42
	Fa	3	58100.00	321330000	17925.68
	Gd	10	124330.00	935189000	30580.86
	TA	16	121312.50	1679295833	40979.21
Spring	Ex	41	153765.24	1129742652	33611.64
	Fa	7	98857.14	452506190	21272.19
	Gd	18	149619.83	1082782633	32905.66
	TA	34	129404.41	767370965	27701.46
Summer	Ex	45	154279.42	1244833504	35282.20
	Fa	5	128800.00	1332625000	36507.88
	Gd	22	113727.27	1155184935	33988.01
	TA	58	134046.55	1138642444	33743.78
Fall	Ex	15	163726.93	2436449681	49360.41
	Fa	1	45000.00	.	.
	Gd	8	143812.50	547495536	23398.62
	TA	11	129345.45	462560727	21507.23




```
/*st103d01.sas*/  /*Part C*/
ods graphics on;

proc glm data=STAT1.ameshousing3 order=internal;
    class Season_Sold Heating_QC;
    model SalePrice = Heating_QC Season_Sold;
    lsmeans Season_Sold / diff adjust=tukey;
    format Season_Sold season.;
    title "Model with Heating Quality and Season as Predictors";
run;
quit;
```

PROC GLM DATA=*SAS-data-set <options>*;
CLASS *variable(s)*;
MODEL *dependent-variable = independent-effects < / options>*;
LSMEANS *effects < / options>*;
RUN;

Model with Heating Quality and Season as Predictors

The GLM Procedure
Dependent Variable: SalePrice Sale price in dollars

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	72774816066	12129136011	10.14	<.0001
Error	293	350448703445	1196070660.2		
Corrected Total	299	423223519511			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.171954	25.14764	34584.25	137524.9

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Heating_QC	3	66835556221	22278518740	18.63	<.0001
Season_Sold	3	5939259845	1979753282	1.66	0.1768

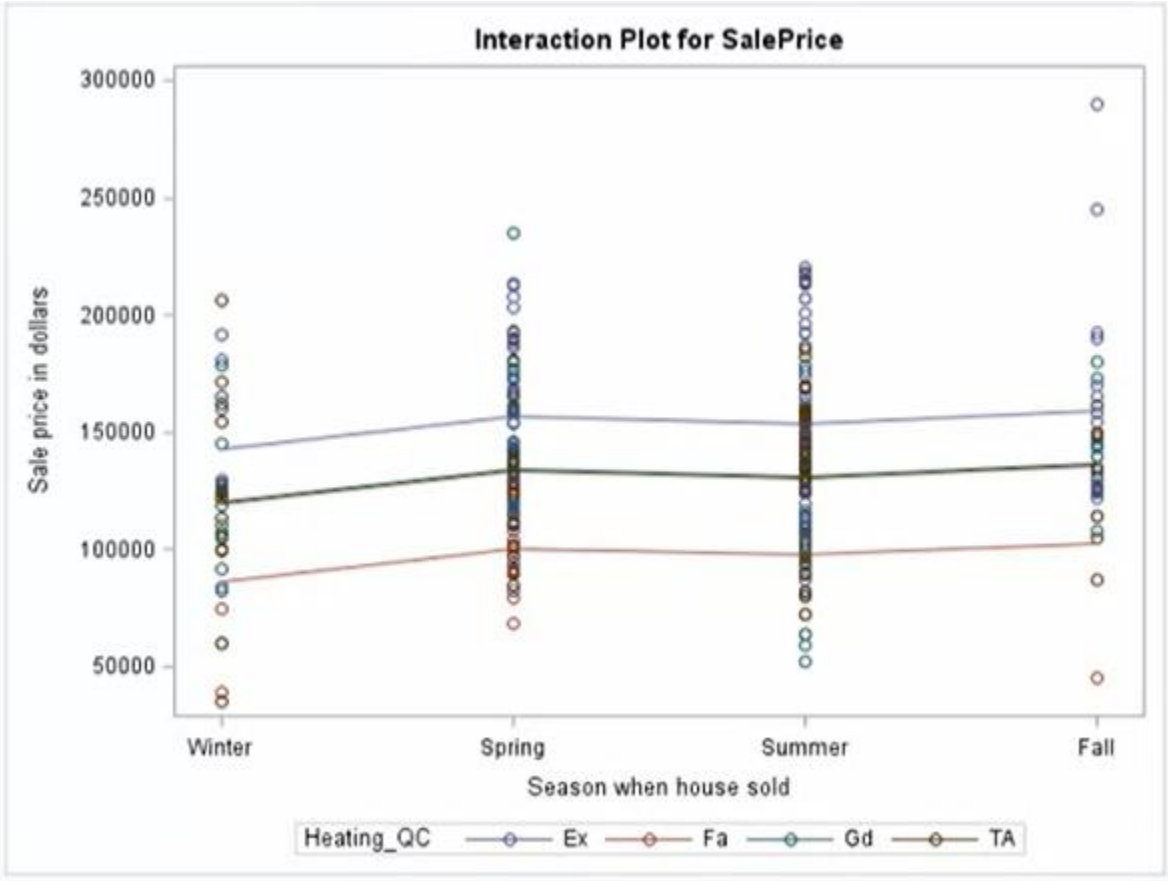
Source	DF	Type III SS	Mean Square	F Value	Pr > F
Heating_QC	3	60050783038	20016927679	16.74	<.0001
Season_Sold	3	5939259845	1979753282	1.66	0.1768

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Heating_QC	3	66835556221	22278518740	18.63	<.0001
Season_Sold	3	5939259845	1979753282	1.66	0.1768

order matters

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Heating_QC	3	60050783038	20016927679	16.74	<.0001
Season_Sold	3	5939259845	1979753282	1.66	0.1768

order is not important



Model with Heating Quality and Season as Predictors

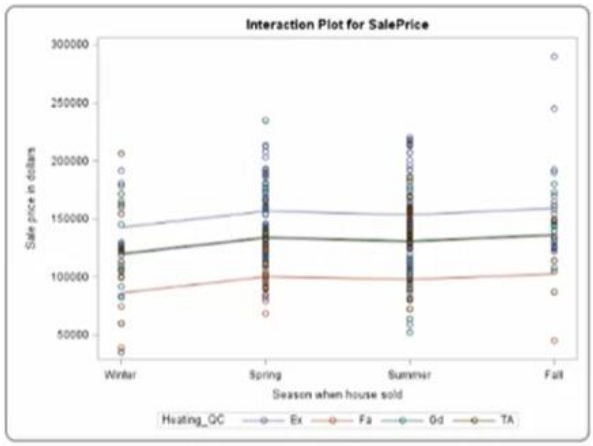
The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

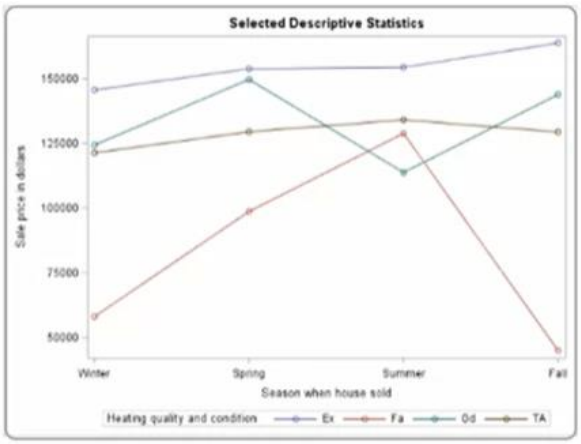
Season_Sold	SalePrice LSMEAN	LSMEAN Number
Winter	117255.605	1
Spring	131263.281	2
Summer	128216.231	3
Fall	133543.394	4

Least Squares Means for effect Season_Sold
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: SalePrice

i/j	1	2	3	4
1		0.1760	0.3529	0.2089
2	0.1760		0.9124	0.9870
3	0.3529	0.9124		0.8517
4	0.2089	0.9870	0.8517	





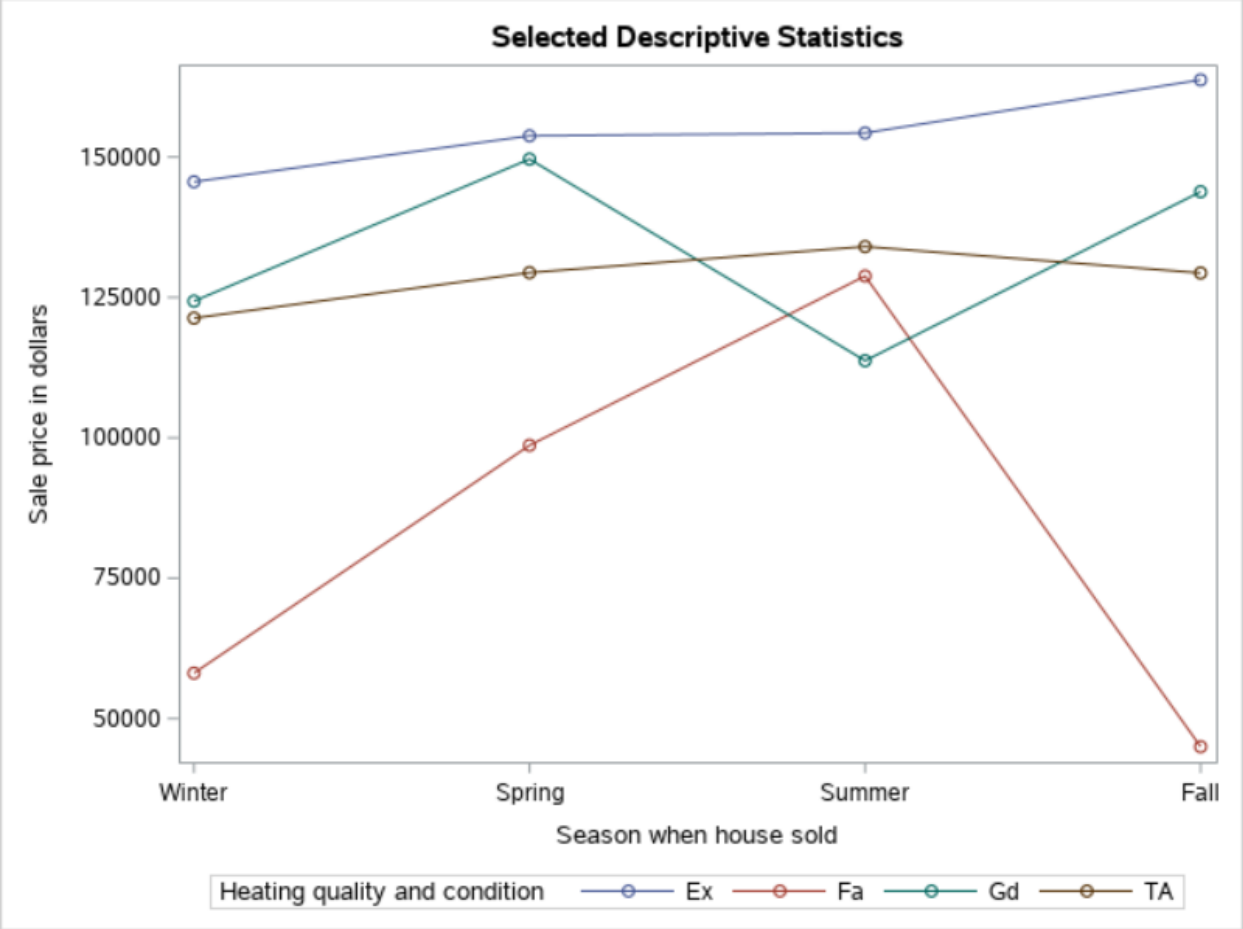

```
/*st103d01.sas*/  /*Part A*/
ods graphics off;
proc means data=STAT1.ameshousing3
          mean var std nway;
   class Season_Sold Heating_QC;
   var SalePrice;
   format Season_Sold Season.;
   title 'Selected Descriptive Statistics';
run;
```

```
/*st103d01.sas*/  /*Part B*/
proc sgplot data=STAT1.ameshousing3;
   vline Season_Sold / group=Heating_QC
                       stat=mean
                       response=SalePrice
                       markers;
   format Season_Sold season.;
run;

/*st103d01.sas*/  /*Part C*/
ods graphics on;

proc glm data=STAT1.ameshousing3 order=internal;
   class Season_Sold Heating_QC;
   model SalePrice = Heating_QC Season_Sold;
   lsmeans Season_Sold / diff adjust=tukey;
   format Season_Sold season.;
   title "Model with Heating Quality and Season as Predictors";
run;
quit;

title;
```

Selected Descriptive Statistics

The MEANS Procedure

Analysis Variable : SalePrice Sale price in dollars					
Season when house sold	Heating quality and condition	N Obs	Mean	Variance	Std Dev
Winter	Ex	6	145583.33	1579141667	39738.42
	Fa	3	58100.00	321330000	17925.68
	Gd	10	124330.00	935189000	30580.86
	TA	16	121312.50	1679295833	40979.21
Spring	Ex	41	153765.24	1129742652	33611.64
	Fa	7	98657.14	452506190	21272.19
	Gd	18	149619.83	1082782633	32905.66
	TA	34	129404.41	767370965	27701.46
Summer	Ex	45	154279.42	1244833504	35282.20
	Fa	5	128800.00	1332825000	36507.88
	Gd	22	113727.27	1155184935	33988.01
	TA	58	134046.55	1138642444	33743.78
Fall	Ex	15	163726.93	2436449681	49360.41
	Fa	1	45000.00	.	.
	Gd	8	143812.50	547495536	23398.62
	TA	11	129345.45	462560727	21507.23


Selected Descriptive Statistics
Sale price in dollars
150000
125000
100000
75000
50000
Winter
Spring
Summer
Fall
Season when house sold
Heating quality and condition
Ex
Fa
Gd
TA

Model with Heating Quality and Season as Predictors

The GLM Procedure

Class Level Information		
Class	Levels	Values
Season_Sold	4	Winter Spring Summer Fall
Heating_QC	4	Ex Fa Gd TA

Number of Observations Read	300
Number of Observations Used	300

Model with Heating Quality and Season as Predictors

The GLM Procedure

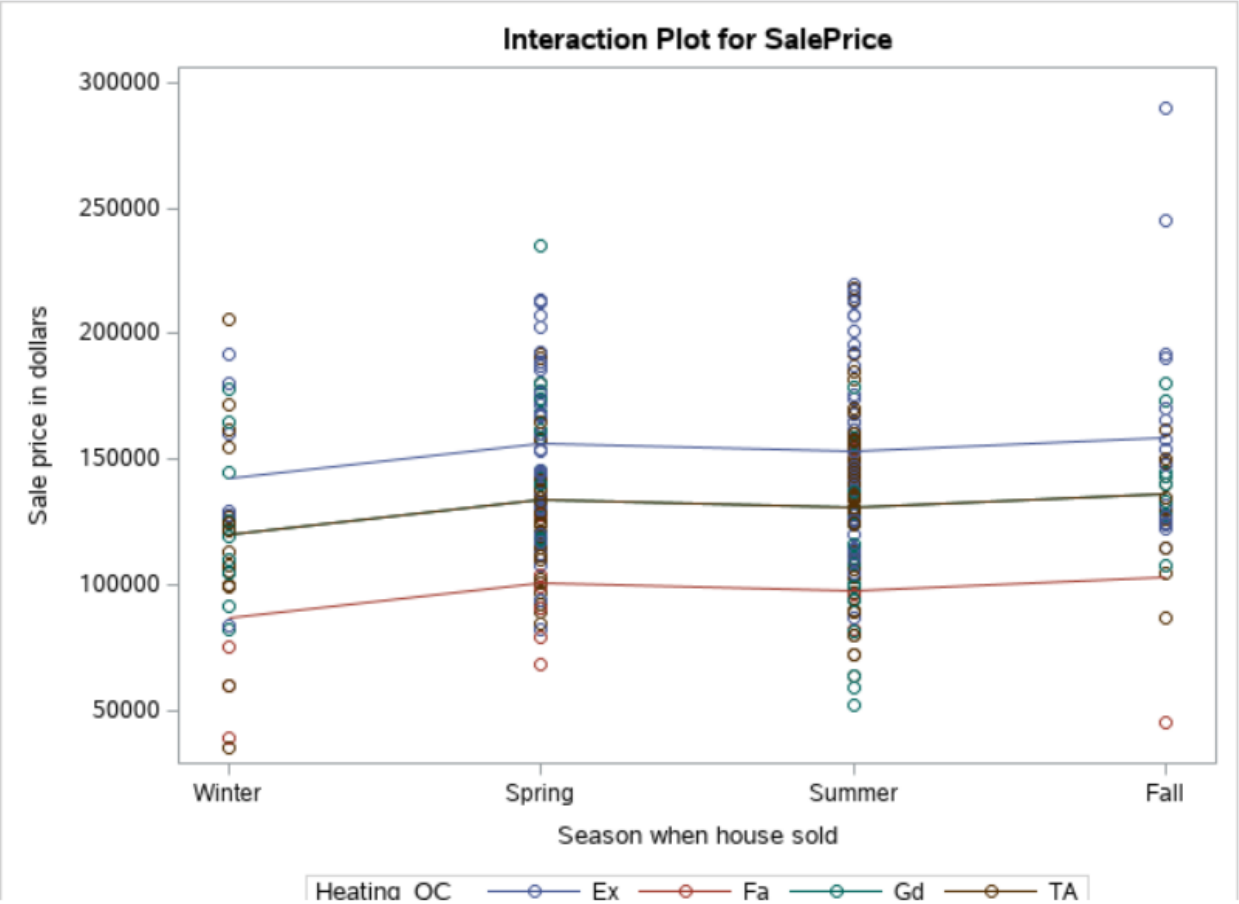
Dependent Variable: SalePrice Sale price in dollars

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	72774816066	12129136011	10.14	<.0001
Error	293	350448703445	1196070660.2		
Corrected Total	299	423223519511			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.171954	25.14764	34584.25	137524.9

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Heating_QC	3	66835556221	22278518740	18.63	<.0001
Season_Sold	3	5939259845	1979753282	1.66	0.1768

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Heating_QC	3	60050783038	20016927679	16.74	<.0001
Season_Sold	3	5939259845	1979753282	1.66	0.1768



Model with Heating Quality and Season as Predictors

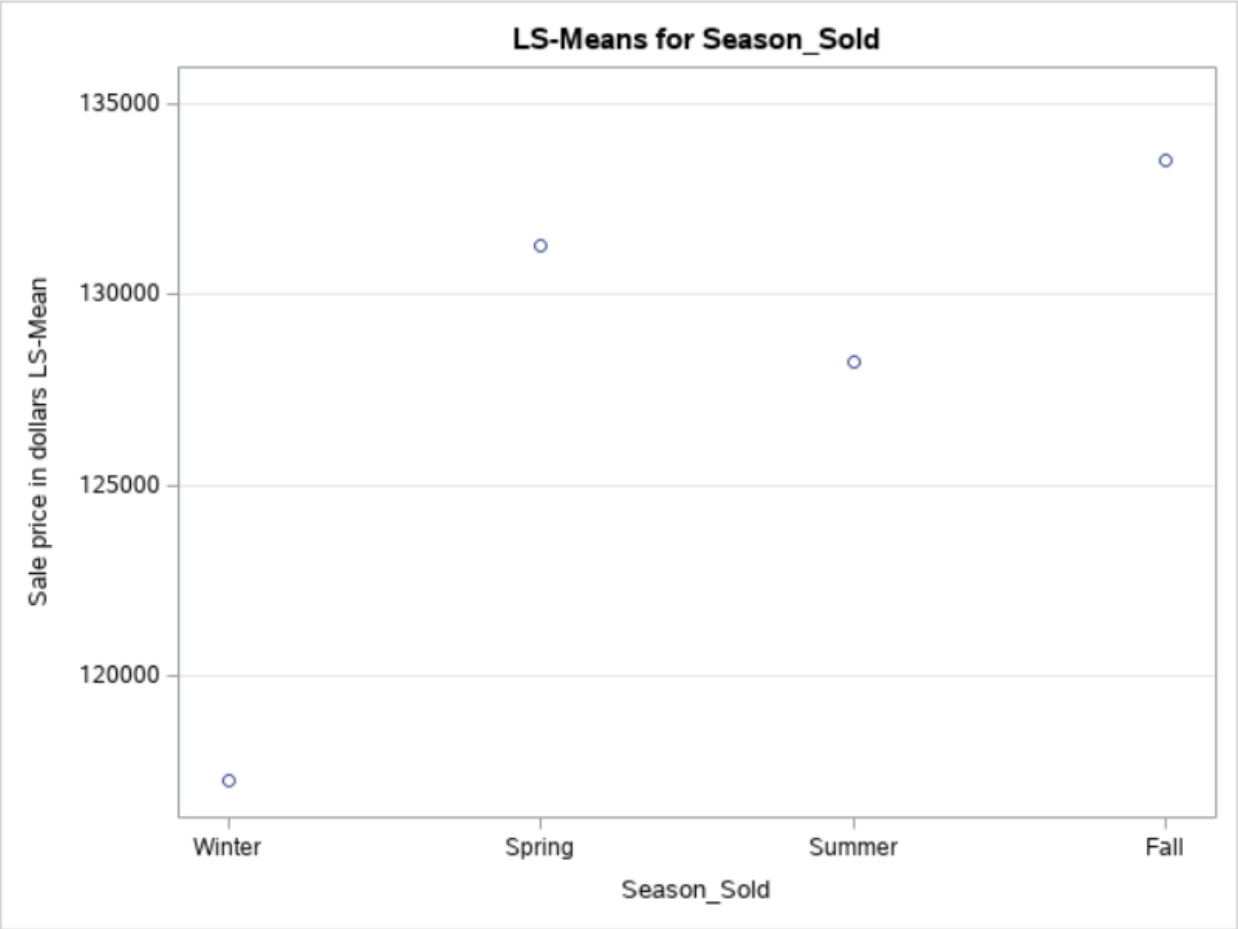
The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

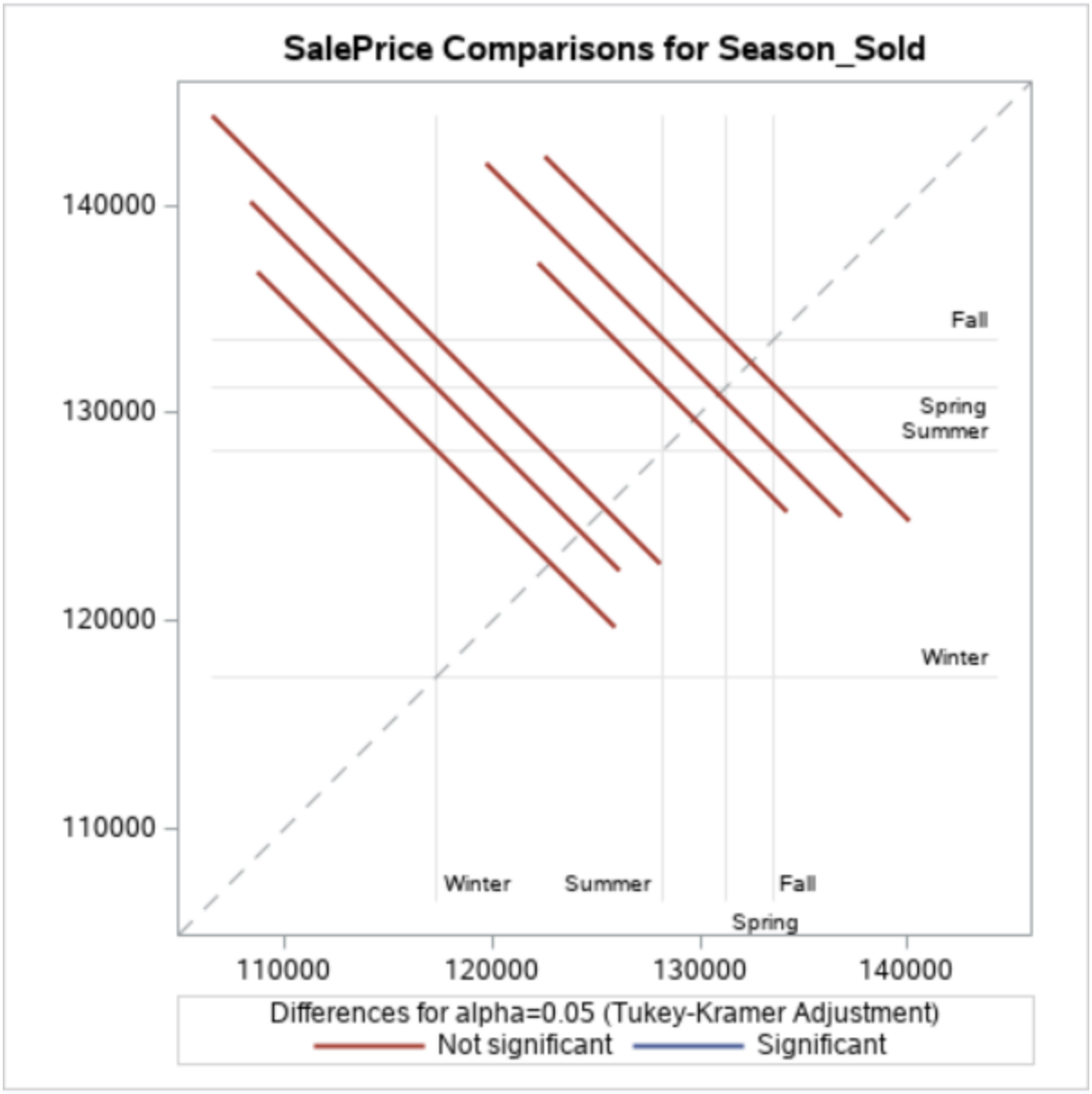
Season_Sold	SalePrice LSMEAN	LSMEAN Number
Winter	117255.605	1
Spring	131263.281	2
Summer	128216.231	3
Fall	133543.394	4

Least Squares Means for effect Season_Sold
Pr > |t| for H0: LSMean(i)=LSMean(j)

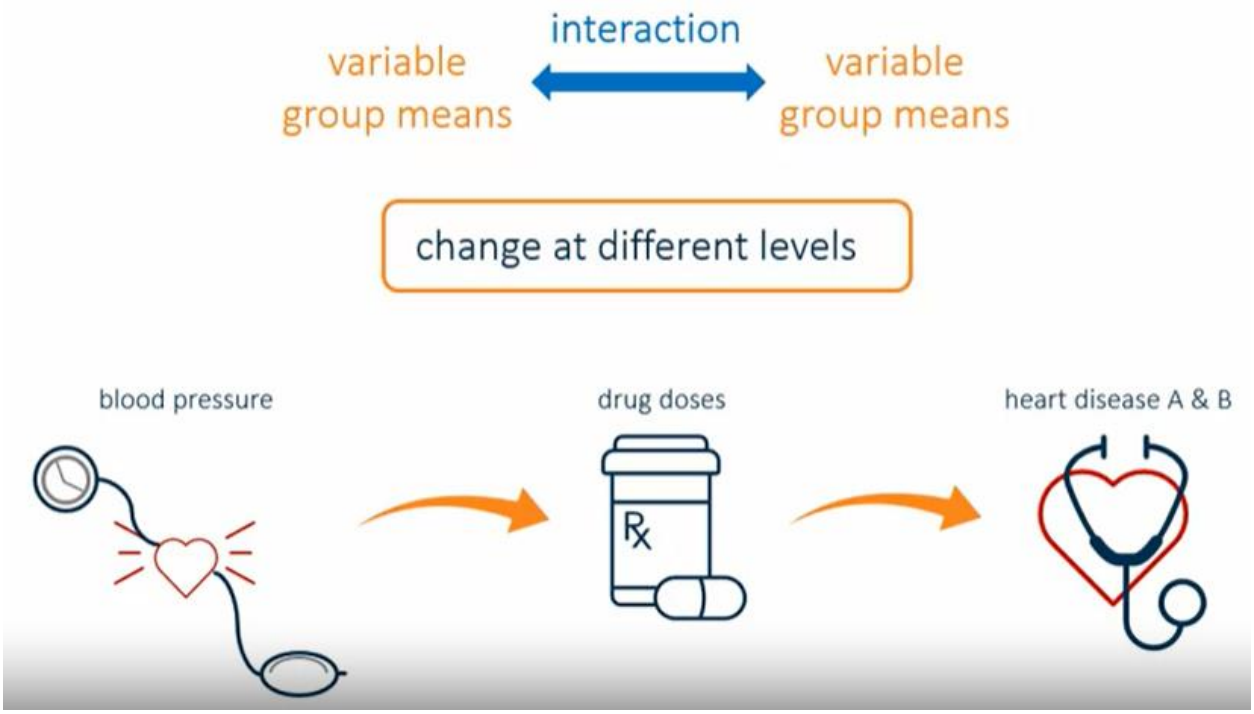
Dependent Variable: SalePrice

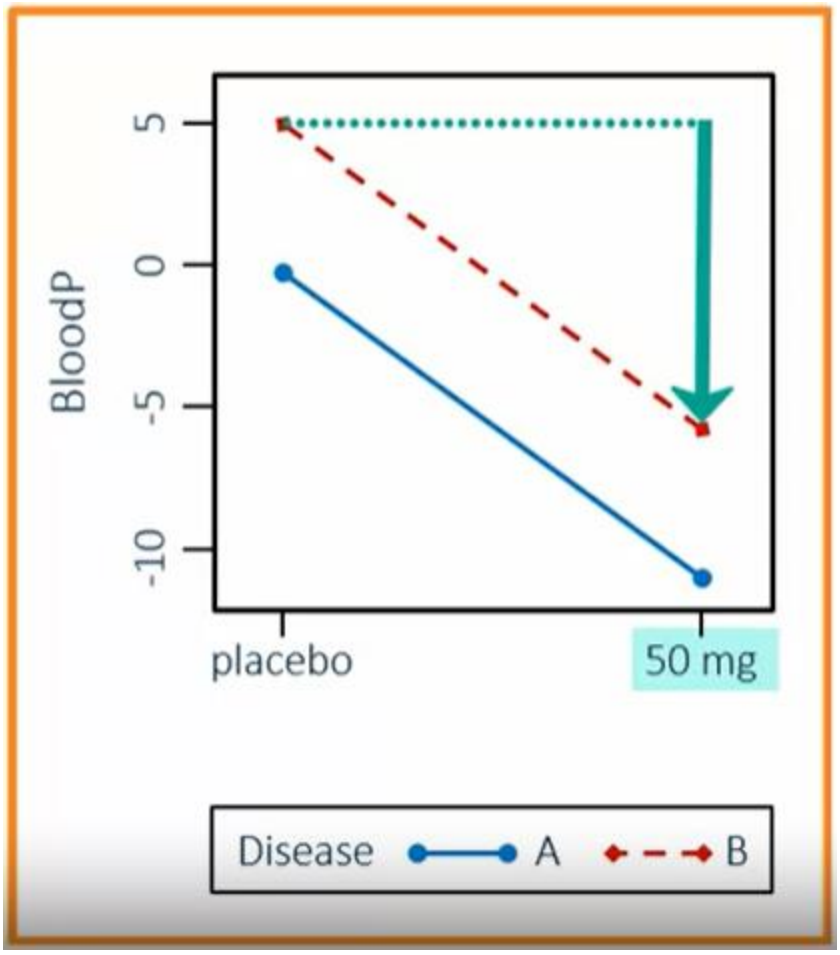
i/j	1	2	3	4
1		0.1760	0.3529	0.2089
2	0.1760		0.9124	0.9870
3	0.3529	0.9124		0.8517
4	0.2089	0.9870	0.8517	

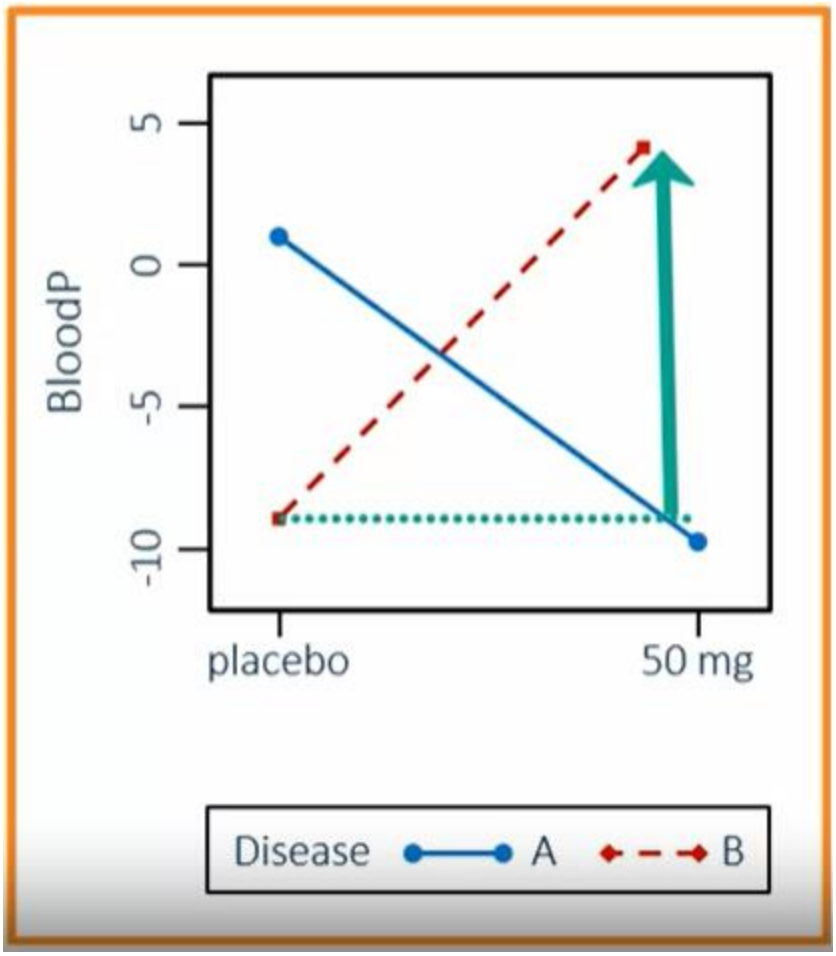

LS-Means for Season_Sold
135000
130000
125000
120000
Sale price in dollars LS-Mean
Winter
Spring
Summer
Fall
Season_Sold

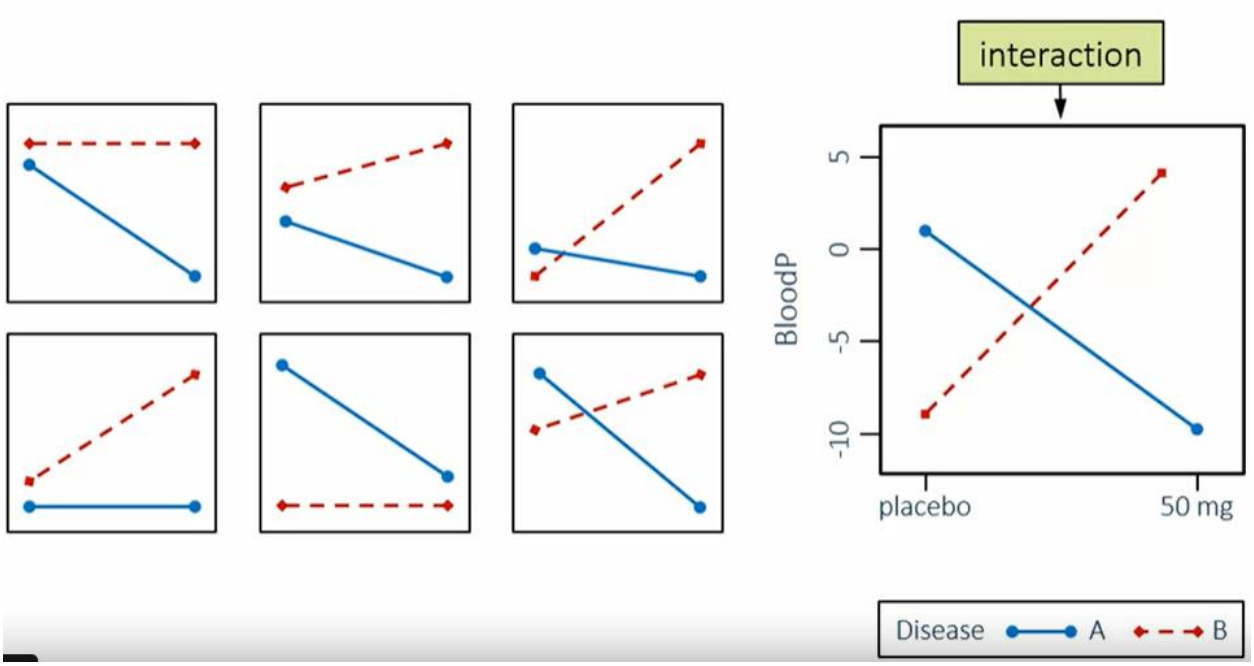

SalePrice Comparisons for Season_Sold
140000
130000
120000
110000
Fall
Spring
Summer
Winter
Winter
Summer
Fall
Spring
110000
120000
130000
140000
Differences for alpha=0.05 (Tukey-Kramer Adjustment)
Not significant
Significant

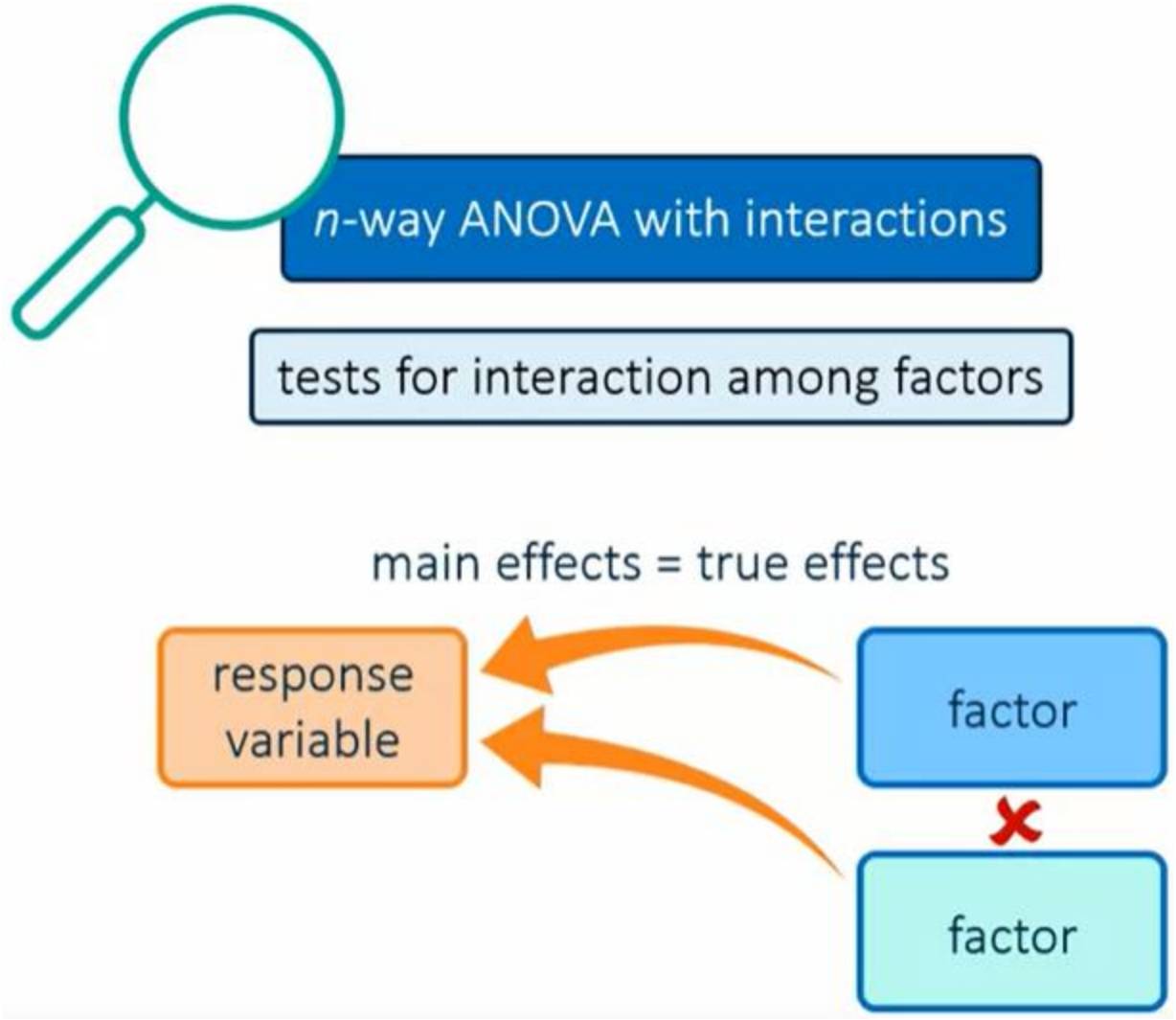
Interactions

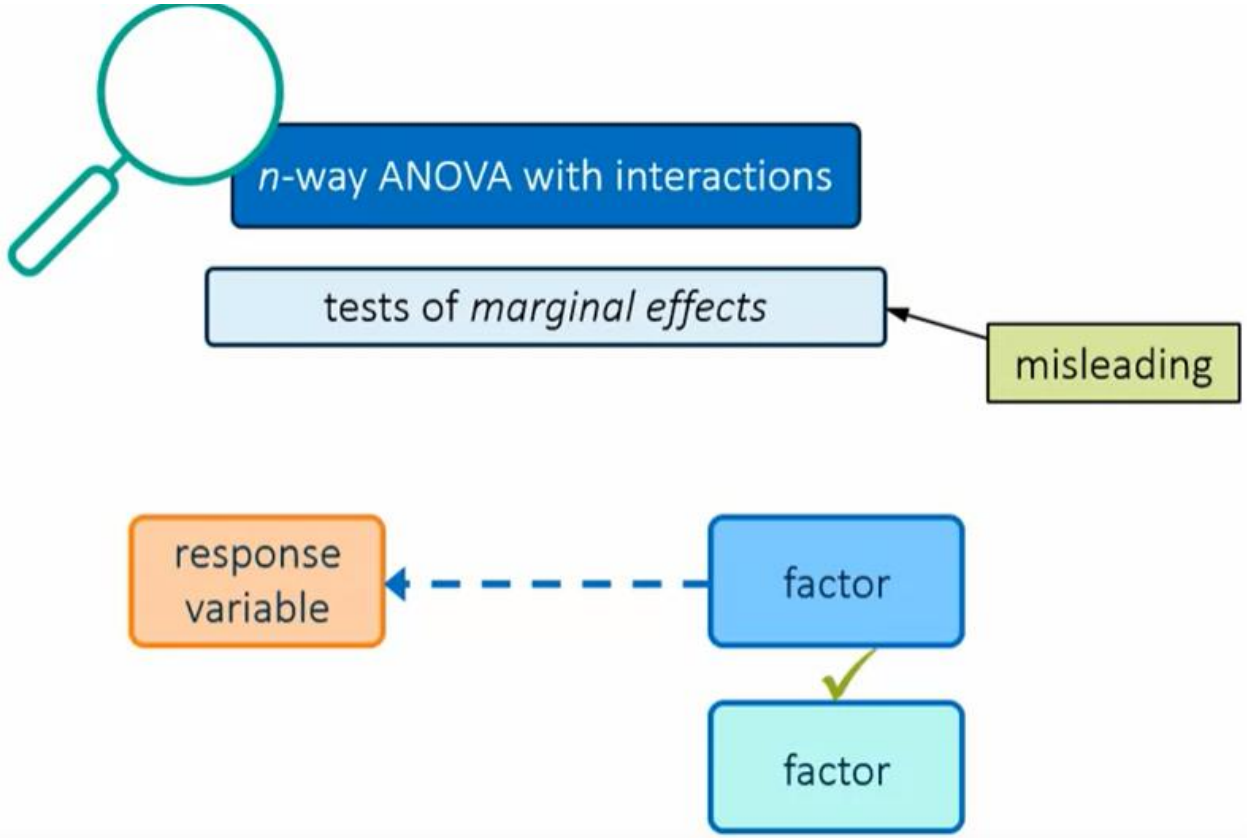


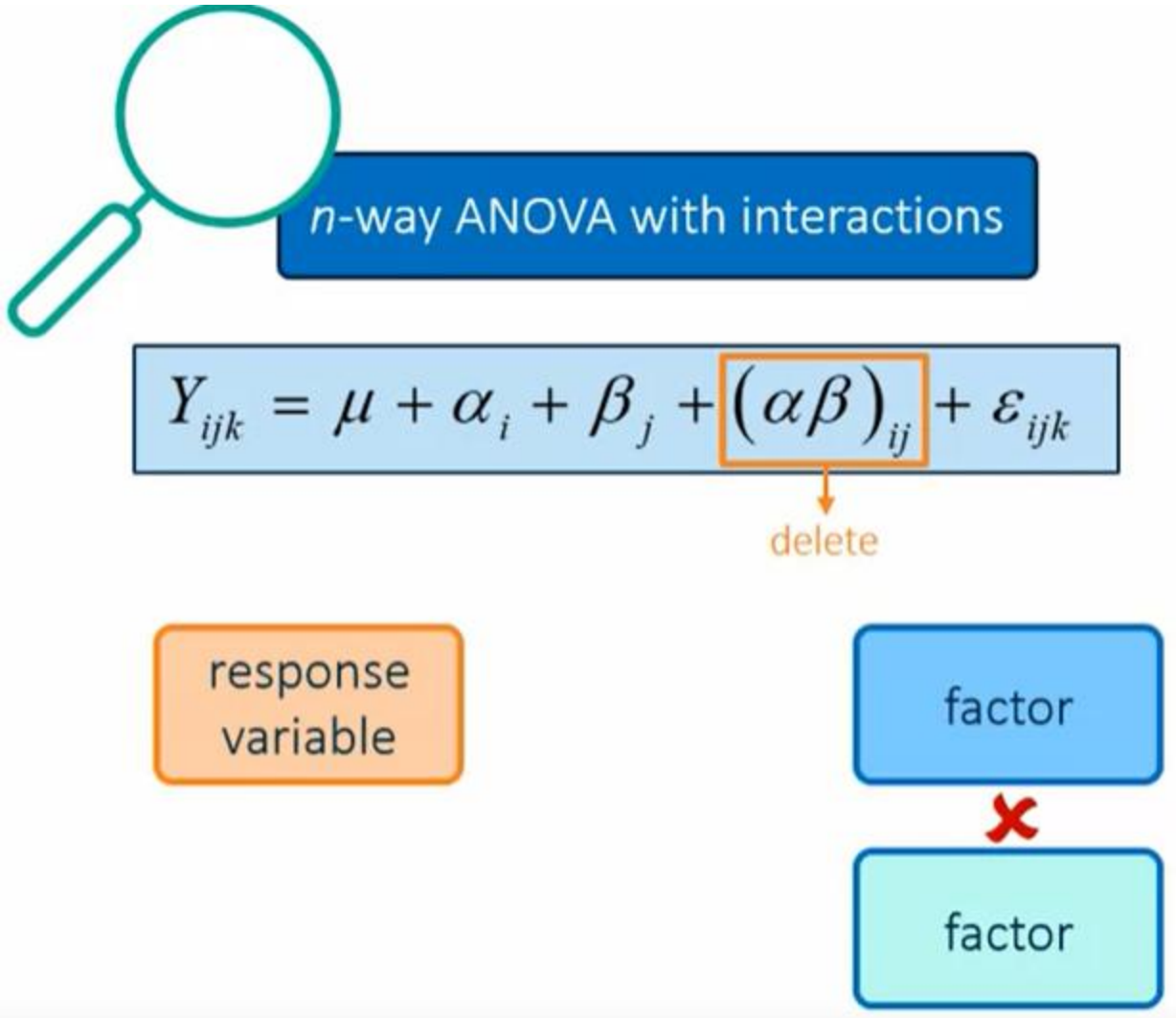

BloodP
5
0
-5
-10
placebo
50 mg
Disease
A
B


BloodP
5
0
-5
-10
placebo
50 mg
Disease
A
B


interaction
BloodP
5
0
-5
-10
placebo
50 mg
Disease
A
B

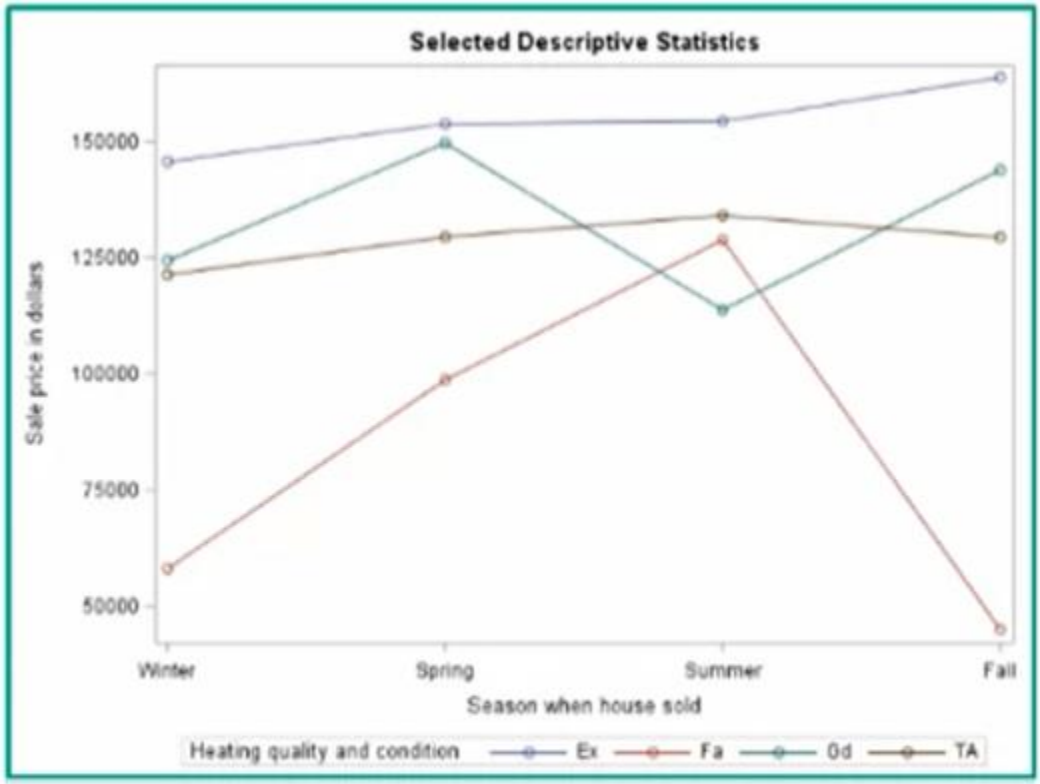

n-way ANOVA with interactions
tests for interaction among factors
main effects = true effects
response variable
factor
factor


n-way ANOVA with interactions
tests of marginal effects
misleading
response variable
factor
factor



You can recognize an interaction by comparing the effect of one variable on the response at different levels of another using group means, or plotting the means to investigate different effect patterns on the response for a variable at different levels of another.

Demo Performing a Two-Way ANOVA with an Interaction Using PROC GLM



Heating_QC*Season_Sold

```
1  /*st103d02.sas*/  /*Part A*/
2  ods graphics on;
3
4  proc glm data=STAT1.ameshousing3
5           order=internal
6           plots(only)=intplot;
7      class Season_Sold Heating_QC;
8      model SalePrice = Heating_QC Season_Sold Heating_QC*Season_Sold;
9      lsmeans Heating_QC*Season_Sold / diff slice=Heating_QC;
10     format Season_Sold Season.;
11     store out=interact;
12     title "Model with Heating Quality and Season as Interacting Predictors";
13 run;
14 quit;
15
```

Heating_QC | Season_Sold

PROC GLM DATA=*SAS-data-set <options>***;**
CLASS *variables*;
MODEL *dependent-variables = independent-effects*;
LSMEANS *effects < / options>*;
STORE <OUT=> *item-store-name < /* **LABEL=***'label'>*;
RUN;

Model with Heating Quality and Season as Interacting Predictors

The GLM Procedure
Dependent Variable: SalePrice Sale price in dollars

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	97609874155	6507324943.7	5.68	<.0001
Error	284	325613645356	1146526920.3		
Corrected Total	299	423223519511			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.230634	24.62130	33860.40	137524.9

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Heating_QC	3	66835556221	22278518740	19.43	<.0001
Season_Sold	3	5939259845	1979753282	1.73	0.1617
Season_So*Heating_QC	9	24835058089	2759450899	2.41	0.0121

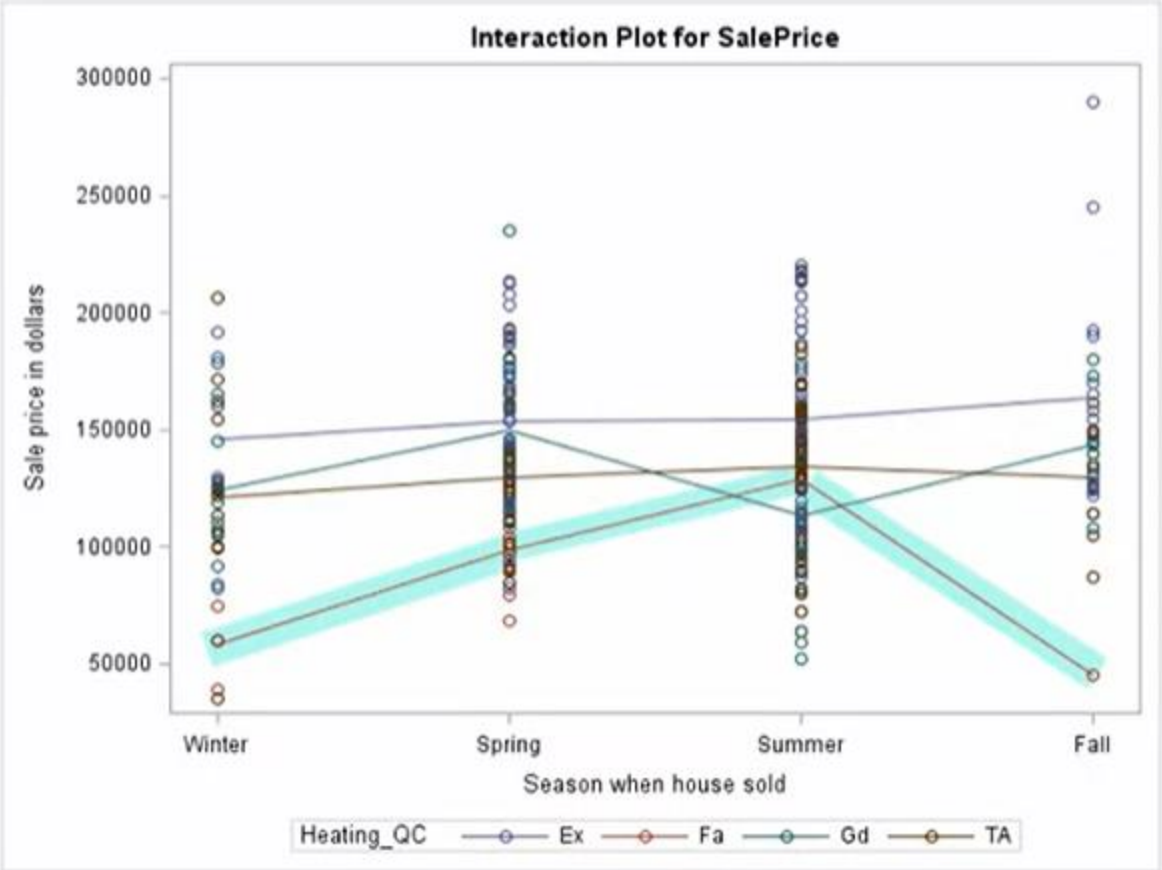
Source	DF	Type III SS	Mean Square	F Value	Pr > F
Heating_QC	3	51116493768	17038831256	14.86	<.0001
Season_Sold	3	9318181844	3106060615	2.71	0.0455
Season_So*Heating_QC	9	24835058089	2759450899	2.41	0.0121

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Heating_QC	3	51116493768	17038831256	14.86	<.0001
Season_Sold	3	9318181844	3106060615	2.71	0.0455
Season_So*Heating_QC	9	24835058089	2759450899	2.41	0.0121

plot of SeasonSold*Heating_QC

pairwise comparisons

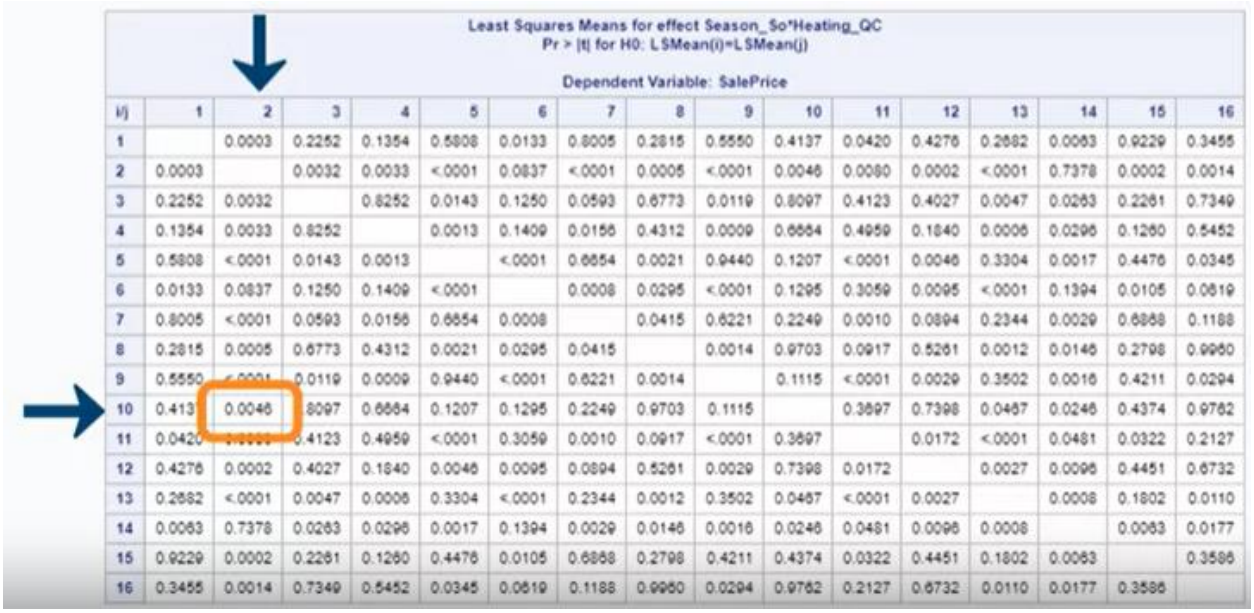
tests of simple effects



Model with Heating Quality and Season as Interacting Predictors

The GLM Procedure
Least Squares Means

Season_Sold	Heating_QC	SalePrice LSMEAN	LSMEAN Number
Winter	Ex	145583.333	1
Winter	Fa	58100.000	2
Winter	Gd	124330.000	3
Winter	TA	121312.500	4
Spring	Ex	153765.244	5
Spring	Fa	98657.143	6
Spring	Gd	149619.833	7
Spring	TA	129404.412	8
Summer	Ex	154279.422	9
Summer	Fa	128800.000	10
Summer	Gd	113727.273	11
Summer	TA	134046.552	12
Fall	Ex	163726.933	13
Fall	Fa	45000.000	14
Fall	Gd	143812.500	15
Fall	TA	129345.455	16



Least Squares Means for effect Season_So*Heating_QC
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: SalePrice

i/j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1		0.0003	0.2252	0.1354	0.5808	0.0133	0.8005	0.2815	0.5550	0.4137	0.0420	0.4276	0.2682	0.0063	0.9229	0.3455
2	0.0003		0.0032	0.0033	<.0001	0.0837	<.0001	0.0005	<.0001	0.0046	0.0080	0.0002	<.0001	0.7378	0.0002	0.0014
3	0.2252	0.0032		0.8252	0.0143	0.1250	0.0593	0.6773	0.0119	0.8097	0.4123	0.4027	0.0047	0.0263	0.2261	0.7349
4	0.1354	0.0033	0.8252		0.0013	0.1409	0.0156	0.4312	0.0009	0.6684	0.4959	0.1840	0.0006	0.0296	0.1260	0.5452
5	0.5808	<.0001	0.0143	0.0013		<.0001	0.6654	0.0021	0.9440	0.1207	<.0001	0.0046	0.3304	0.0017	0.4476	0.0345
6	0.0133	0.0837	0.1250	0.1409	<.0001		0.0008	0.0295	<.0001	0.1295	0.3059	0.0095	<.0001	0.1394	0.0105	0.0619
7	0.8005	<.0001	0.0593	0.0156	0.6654	0.0008		0.0415	0.6221	0.2249	0.0010	0.0894	0.2344	0.0029	0.6868	0.1188
8	0.2815	0.0005	0.6773	0.4312	0.0021	0.0295	0.0415		0.0014	0.9703	0.0917	0.5261	0.0012	0.0146	0.2798	0.9960
9	0.5550	<.0001	0.0119	0.0009	0.9440	<.0001	0.6221	0.0014		0.1115	<.0001	0.0029	0.3502	0.0016	0.4211	0.0294
10	0.413	0.0046	.8097	0.6684	0.1207	0.1295	0.2249	0.9703	0.1115		0.3697	0.7398	0.0467	0.0246	0.4374	0.9762
11	0.0420	[illegible]	0.4123	0.4959	<.0001	0.3059	0.0010	0.0917	<.0001	0.3697		0.0172	<.0001	0.0481	0.0322	0.2127
12	0.4276	0.0002	0.4027	0.1840	0.0046	0.0095	0.0894	0.5261	0.0029	0.7398	0.0172		0.0027	0.0096	0.4451	0.6732
13	0.2682	<.0001	0.0047	0.0006	0.3304	<.0001	0.2344	0.0012	0.3502	0.0467	<.0001	0.0027		0.0008	0.1802	0.0110
14	0.0063	0.7378	0.0263	0.0296	0.0017	0.1394	0.0029	0.0146	0.0016	0.0246	0.0481	0.0096	0.0008		0.0063	0.0177
15	0.9229	0.0002	0.2261	0.1260	0.4476	0.0105	0.6868	0.2798	0.4211	0.4374	0.0322	0.4451	0.1802	0.0063		0.3596
16	0.3455	0.0014	0.7349	0.5452	0.0345	0.0619	0.1188	0.9960	0.0294	0.9762	0.2127	0.6732	0.0110	0.0177	0.3596	

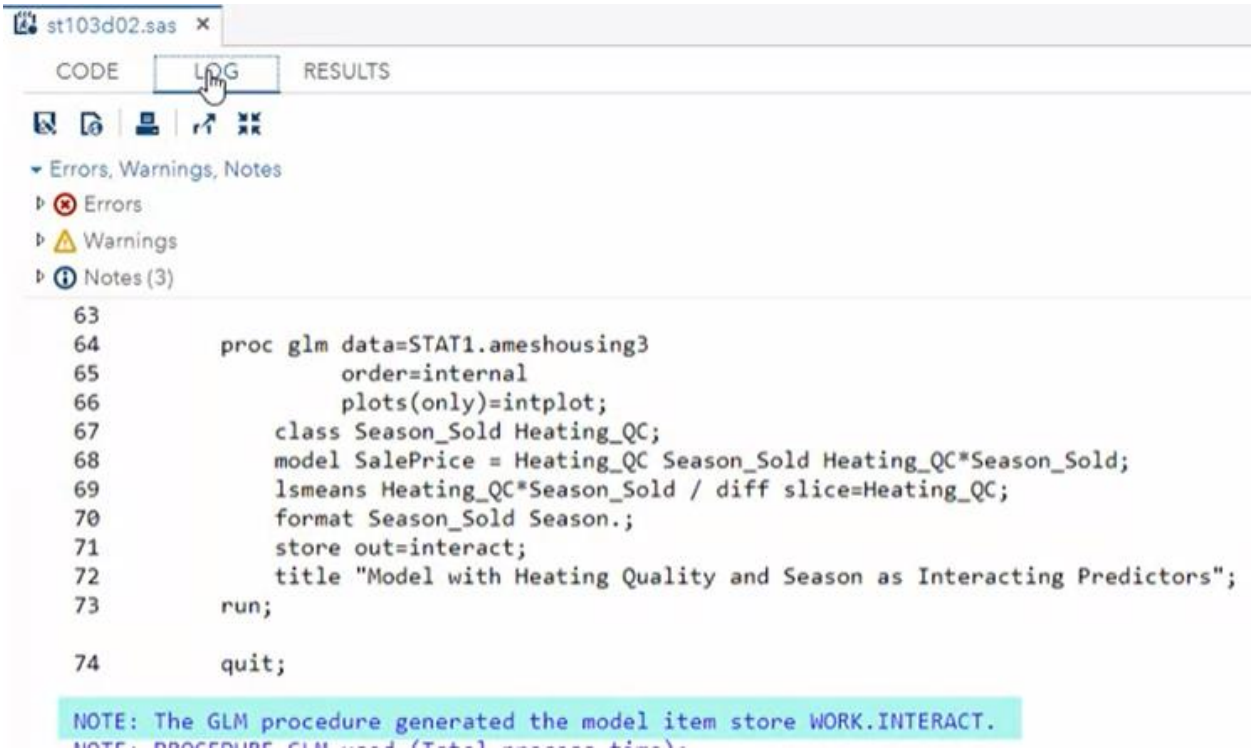
Model with Heating Quality and Season as Interacting Predictors

The GLM Procedure
Least Squares Means

Season_So*Heating_QC Effect Sliced by Heating_QC for SalePrice

Heating_QC	DF	Sum of Squares	Mean Square	F Value	Pr > F
Ex	3	1759608339	586536113	0.51	0.6746
Fa	3	12318827232	4106275744	3.58	0.0143
Gd	3	14560964166	4853654722	4.23	0.0060
TA	3	2134918196	711639399	0.62	0.6021

Note: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used.




```
/*st103d02.sas*/  /*Part A*/
ods graphics on;

proc glm data=STAT1.ameshousing3
        order=internal
        plots(only)=intplot;
    class Season_Sold Heating_QC;
    model SalePrice = Heating_QC Season_Sold Heating_QC*Season_Sold;
    lsmeans Heating_QC*Season_Sold / diff slice=Heating_QC;
    format Season_Sold Season.;
    store out=interact;
    title "Model with Heating Quality and Season as Interacting Predictors";
run;
quit;
```

Model with Heating Quality and Season as Interacting Predictors

The GLM Procedure

Class Level Information		
Class	Levels	Values
Season_Sold	4	Winter Spring Summer Fall
Heating_QC	4	Ex Fa Gd TA

Number of Observations Read	300
Number of Observations Used	300

Model with Heating Quality and Season as Interacting Predictors

The GLM Procedure

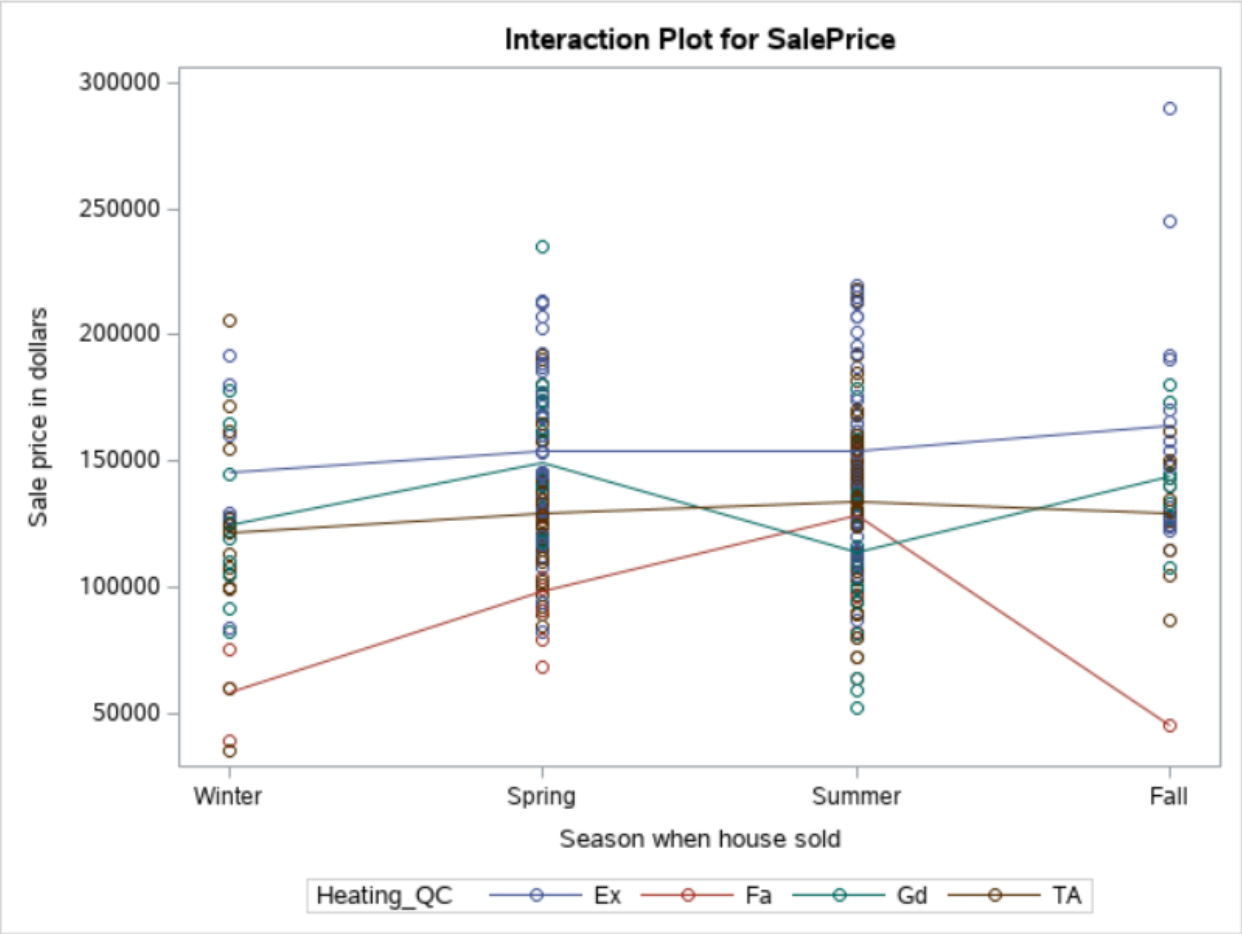
Dependent Variable: SalePrice Sale price in dollars

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	97609874155	6507324943.7	5.68	<.0001
Error	284	325613645356	1146526920.3		
Corrected Total	299	423223519511			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.230634	24.62130	33860.40	137524.9

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Heating_QC	3	66835556221	22278518740	19.43	<.0001
Season_Sold	3	5939259845	1979753282	1.73	0.1617
Season_So*Heating_QC	9	24835058089	2759450899	2.41	0.0121

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Heating_QC	3	51116493768	17038831256	14.86	<.0001
Season_Sold	3	9318181844	3106060615	2.71	0.0455
Season_So*Heating_QC	9	24835058089	2759450899	2.41	0.0121


Interaction Plot for SalePrice
300000
250000
200000
150000
100000
50000
Sale price in dollars
Winter
Spring
Summer
Fall
Season when house sold
Heating_QC
Ex
Fa
Gd
TA

Model with Heating Quality and Season as Interacting Predictors

The GLM Procedure
Least Squares Means

Season_Sold	Heating_QC	SalePrice LSMEAN	LSMEAN Number
Winter	Ex	145583.333	1
Winter	Fa	58100.000	2
Winter	Gd	124330.000	3
Winter	TA	121312.500	4
Spring	Ex	153765.244	5
Spring	Fa	98657.143	6
Spring	Gd	149619.833	7
Spring	TA	129404.412	8
Summer	Ex	154279.422	9
Summer	Fa	128800.000	10
Summer	Gd	113727.273	11
Summer	TA	134046.552	12
Fall	Ex	163726.933	13
Fall	Fa	45000.000	14
Fall	Gd	143812.500	15
Fall	TA	129345.455	16

Least Squares Means for effect Season_So*Heating_QC
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: SalePrice

i/j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1		0.0003	0.2252	0.1354	0.5808	0.0133	0.8005	0.2815	0.5550	0.4137	0.0420	0.4276	0.2682	0.0063	0.9229	0.3455
2	0.0003		0.0032	0.0033	<.0001	0.0837	<.0001	0.0005	<.0001	0.0046	0.0080	0.0002	<.0001	0.7378	0.0002	0.0014
3	0.2252	0.0032		0.8252	0.0143	0.1250	0.0593	0.6773	0.0119	0.8097	0.4123	0.4027	0.0047	0.0263	0.2261	0.7349
4	0.1354	0.0033	0.8252		0.0013	0.1409	0.0156	0.4312	0.0009	0.6664	0.4959	0.1840	0.0006	0.0296	0.1260	0.5452
5	0.5808	<.0001	0.0143	0.0013		<.0001	0.6654	0.0021	0.9440	0.1207	<.0001	0.0046	0.3304	0.0017	0.4476	0.0345
6	0.0133	0.0837	0.1250	0.1409	<.0001		0.0008	0.0295	<.0001	0.1295	0.3059	0.0095	<.0001	0.1394	0.0105	0.0619
7	0.8005	<.0001	0.0593	0.0156	0.6654	0.0008		0.0415	0.6221	0.2249	0.0010	0.0894	0.2344	0.0029	0.6868	0.1188
8	0.2815	0.0005	0.6773	0.4312	0.0021	0.0295	0.0415		0.0014	0.9703	0.0917	0.5261	0.0012	0.0146	0.2798	0.9960
9	0.5550	<.0001	0.0119	0.0009	0.9440	<.0001	0.6221	0.0014		0.1115	<.0001	0.0029	0.3502	0.0016	0.4211	0.0294
10	0.4137	0.0046	0.8097	0.6664	0.1207	0.1295	0.2249	0.9703	0.1115		0.3697	0.7398	0.0467	0.0246	0.4374	0.9762
11	0.0420	0.0080	0.4123	0.4959	<.0001	0.3059	0.0010	0.0917	<.0001	0.3697		0.0172	<.0001	0.0481	0.0322	0.2127
12	0.4276	0.0002	0.4027	0.1840	0.0046	0.0095	0.0894	0.5261	0.0029	0.7398	0.0172		0.0027	0.0096	0.4451	0.6732
13	0.2682	<.0001	0.0047	0.0006	0.3304	<.0001	0.2344	0.0012	0.3502	0.0467	<.0001	0.0027		0.0008	0.1802	0.0110
14	0.0063	0.7378	0.0263	0.0296	0.0017	0.1394	0.0029	0.0146	0.0016	0.0246	0.0481	0.0096	0.0008		0.0063	0.0177
15	0.9229	0.0002	0.2261	0.1260	0.4476	0.0105	0.6868	0.2798	0.4211	0.4374	0.0322	0.4451	0.1802	0.0063		0.3586
16	0.3455	0.0014	0.7349	0.5452	0.0345	0.0619	0.1188	0.9960	0.0294	0.9762	0.2127	0.6732	0.0110	0.0177	0.3586	

Model with Heating Quality and Season as Interacting Predictors

The GLM Procedure
Least Squares Means

Season_So*Heating_QC Effect Sliced by Heating_QC for SalePrice					
Heating_QC	DF	Sum of Squares	Mean Square	F Value	Pr > F
Ex	3	1759608339	586536113	0.51	0.6746
Fa	3	12318827232	4106275744	3.58	0.0143
Gd	3	14560964166	4853654722	4.23	0.0060
TA	3	2134918196	711639399	0.62	0.6021

Note: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used.

The STORE Statement

The STORE statement saves model fit information in an item store. It can be used in many SAS/STAT procedures including GENMOD, GLIMMIX, GLM, GLMSELECT, LOGISTIC, MIXED, ORTHOREG, PHREG, PROBIT, SURVEYLOGISTIC, SURVEYPHREG, and SURVEYREG. Later, you can use PROC PLM to access the item store and perform new tests and analyses on the fit model.

For example, suppose you need to perform an analysis that will take several hours. Perhaps you have access to the data only for a limited time, and you know that you'll need to do further analysis at a later date. You can use a STORE statement to save the results in an item store, and later use PROC PLM to perform additional analysis on the saved results without needing to access the original data, or fit the model again. This can be a great time saver!

Here's the syntax:

STORE <OUT=>*item-store-name* </ LABEL='*label*'>;

- *Item-store-name* is a one- or two-level SAS name that is similar to the names that are used for SAS data sets. If you specify a one-level name, then the item store resides in the Work library and is deleted at the end of the SAS session. Because item stores are usually used to perform postprocessing tasks, typical usage specifies a two-level name in the form *libname.membername* where *libname* is a permanent SAS library.
- *Label* identifies the estimate on the output. A label is optional but must be enclosed in quotation marks. When the PLM procedure processes an item store, the label appears in the PROC PLM output along with other identifying information.

Note: If an item store by the same name as specified in the STORE statement already exists, the existing store is replaced. For more information about postprocessing tasks based on item stores, see the documentation for the PLM procedure.

Demo Performing Post-Processing Analysis Using PROC PLM

```
/*st103d02.sas*/  /*Part A*/
ods graphics on;

proc glm data=STAT1.ameshousing3
         order=internal
         plots(only)=intplot;
    class Season_Sold Heating_QC;
    model SalePrice = Heating_QC Season_Sold Heating_QC*Season_Sold;
    lsmeans Heating_QC*Season_Sold / diff slice=Heating_QC;
    format Season_Sold Season.;
    store out=interact;
    title "Model with Heating Quality and Season as Interacting Predictors";
run;
quit;

/*st103d02.sas*/  /*Part B*/
proc plm restore=interact plots=all;
    slice Heating_QC*Season_Sold / sliceby=Heating_QC adjust=tukey;
    effectplot interaction(sliceby=Heating_QC) / clm;
run;

title;
```

PROC PLM RESTORE=*item-store-specification <options>*;
SLICE *model-effect < / options>*;
EFFECTPLOT *<plot-type <(plot-definition-options)>> < / options>*;
RUN;

F Test for Season_So*Heating_QC Least Squares Means Slice

Slice	Num DF	Den DF	F Value	Pr > F
Heating_QC Ex	3	284	0.51	0.6746

Simple Differences of Season_So*Heating_QC Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

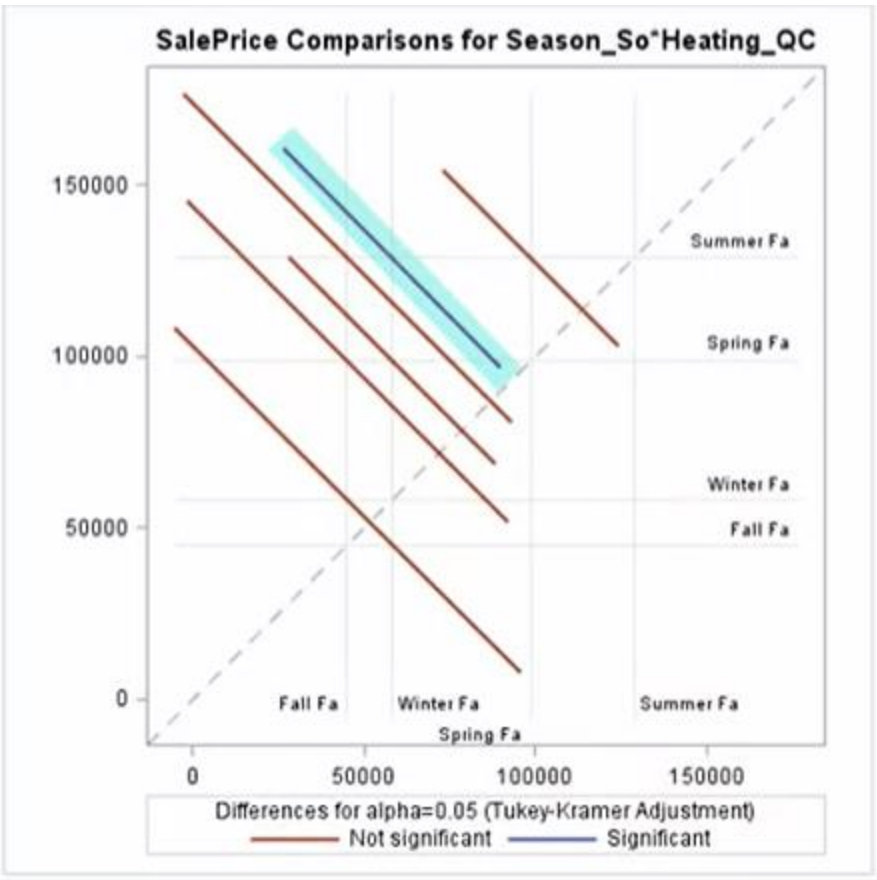
Slice	Season when house sold	Season when house sold	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P
Heating_QC Ex	Winter	Spring	-8181.91	14800	284	-0.55	0.5808	0.9457
Heating_QC Ex	Winter	Summer	-8696.09	14716	284	-0.59	0.5550	0.9348
Heating_QC Ex	Winter	Fall	-16144	16356	284	-1.11	0.2682	0.6841
Heating_QC Ex	Spring	Summer	-514.18	7310.43	284	-0.07	0.9440	0.9999
Heating_QC Ex	Spring	Fall	-7962.09	10218	284	-0.97	0.3304	0.7638
Heating_QC Ex	Summer	Fall	-7447.91	10095	284	-0.94	0.3502	0.7856

F Test for Season_So*Heating_QC Least Squares Means Slice

Slice	Num DF	Den DF	F Value	Pr > F
Heating_QC Fa	3	284	3.58	0.0143

Simple Differences of Season_So*Heating_QC Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Slice	Season when house sold	Season when house sold	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P
Heating_QC Fa	Winter	Spring	-40557	23366	284	-1.74	0.0837	0.3071
Heating_QC Fa	Winter	Summer	-70700	24728	284	-2.86	0.0046	0.0235
Heating_QC Fa	Winter	Fall	13100	39099	284	0.34	0.7378	0.9870
Heating_QC Fa	Spring	Summer	-30143	19827	284	-1.52	0.1295	0.4267
Heating_QC Fa	Spring	Fall	53657	36198	284	1.48	0.1394	0.4495
Heating_QC Fa	Summer	Fall	83800	37092	284	2.26	0.0246	0.1102

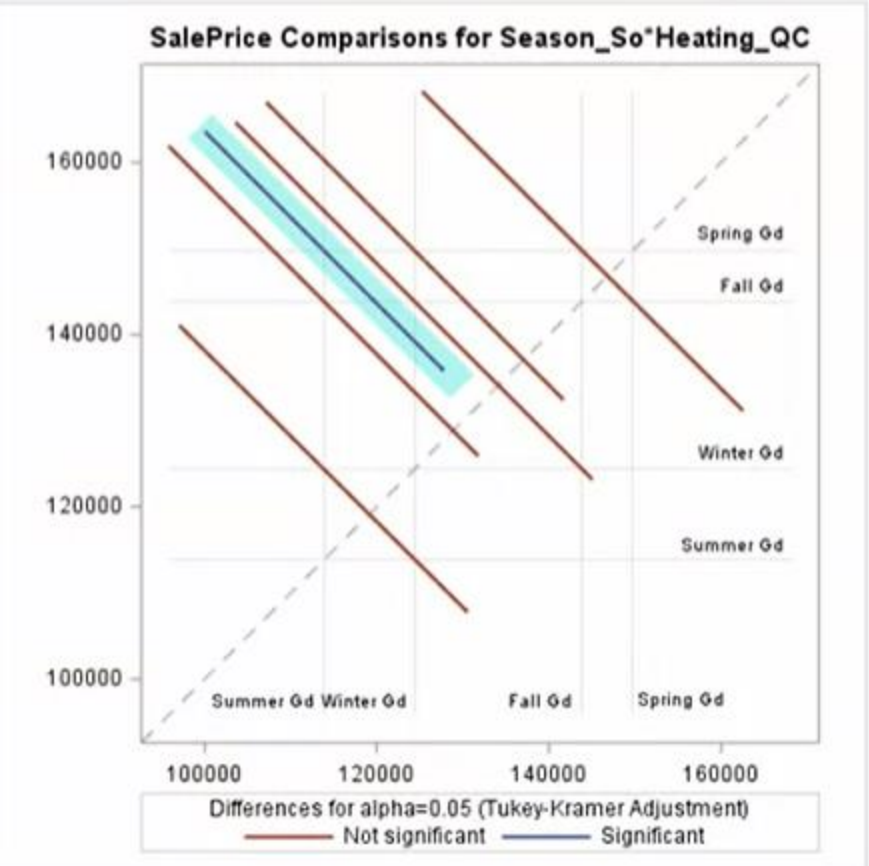


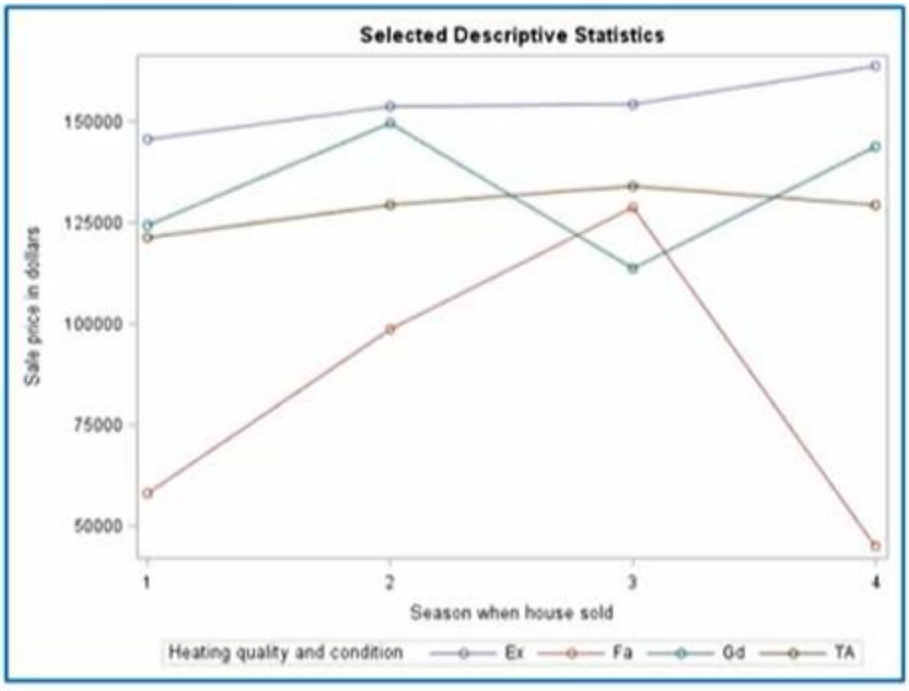
F Test for Season_So*Heating_QC Least Squares Means Slice

Slice	Num DF	Den DF	F Value	Pr > F
Heating_QC Gd	3	284	4.23	0.0060

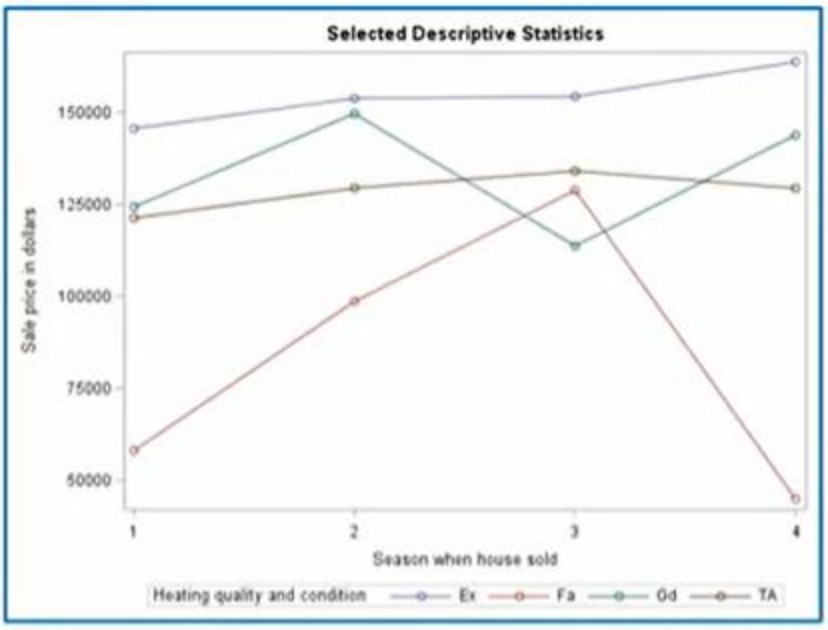
Simple Differences of Season_So*Heating_QC Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Slice	Season when house sold	Season when house sold	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P
Heating_QC Gd	Winter	Spring	-25290	13355	284	-1.89	0.0593	0.2330
Heating_QC Gd	Winter	Summer	10603	12914	284	0.82	0.4123	0.8445
Heating_QC Gd	Winter	Fall	-19483	16061	284	-1.21	0.2261	0.6191
Heating_QC Gd	Spring	Summer	35893	10762	284	3.34	0.0010	0.0053
Heating_QC Gd	Spring	Fall	5807.33	14388	284	0.40	0.6868	0.9777
Heating_QC Gd	Summer	Fall	-30085	13980	284	-2.15	0.0322	0.1394



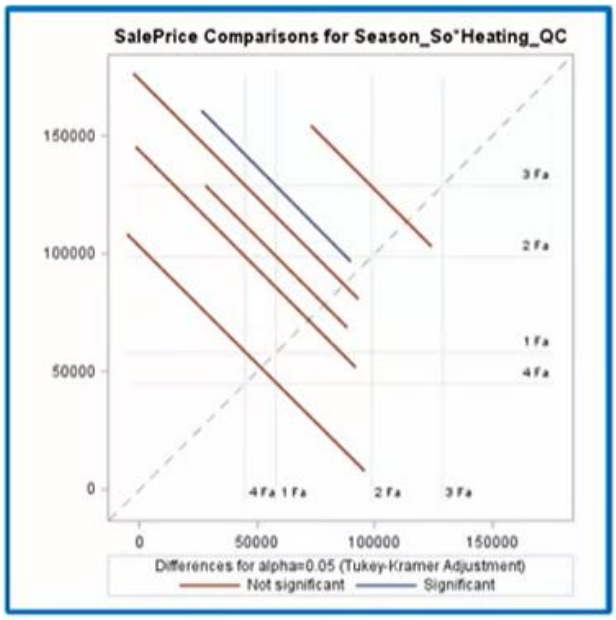


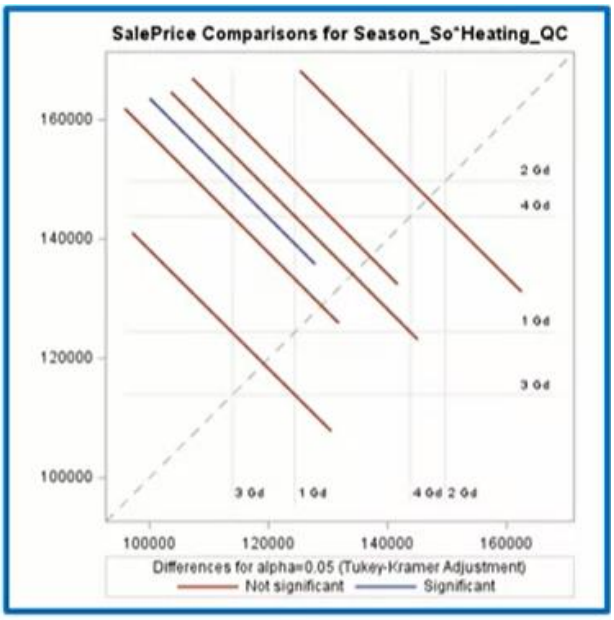
Begin by visually investigating the data.



Explore significant interactions

- additional graphics
- differences of least squares means
- tests of simple effects





Fair

Good

```
/*st103d02.sas*/  /*Part B*/
proc plm restore=interact plots=all;
    slice Heating_QC*Season_Sold / sliceby=Heating_QC adjust=tukey;
    effectplot interaction(sliceby=Heating_QC) / clm;
run;

title;
```

Model with Heating Quality and Season as Interacting Predictors

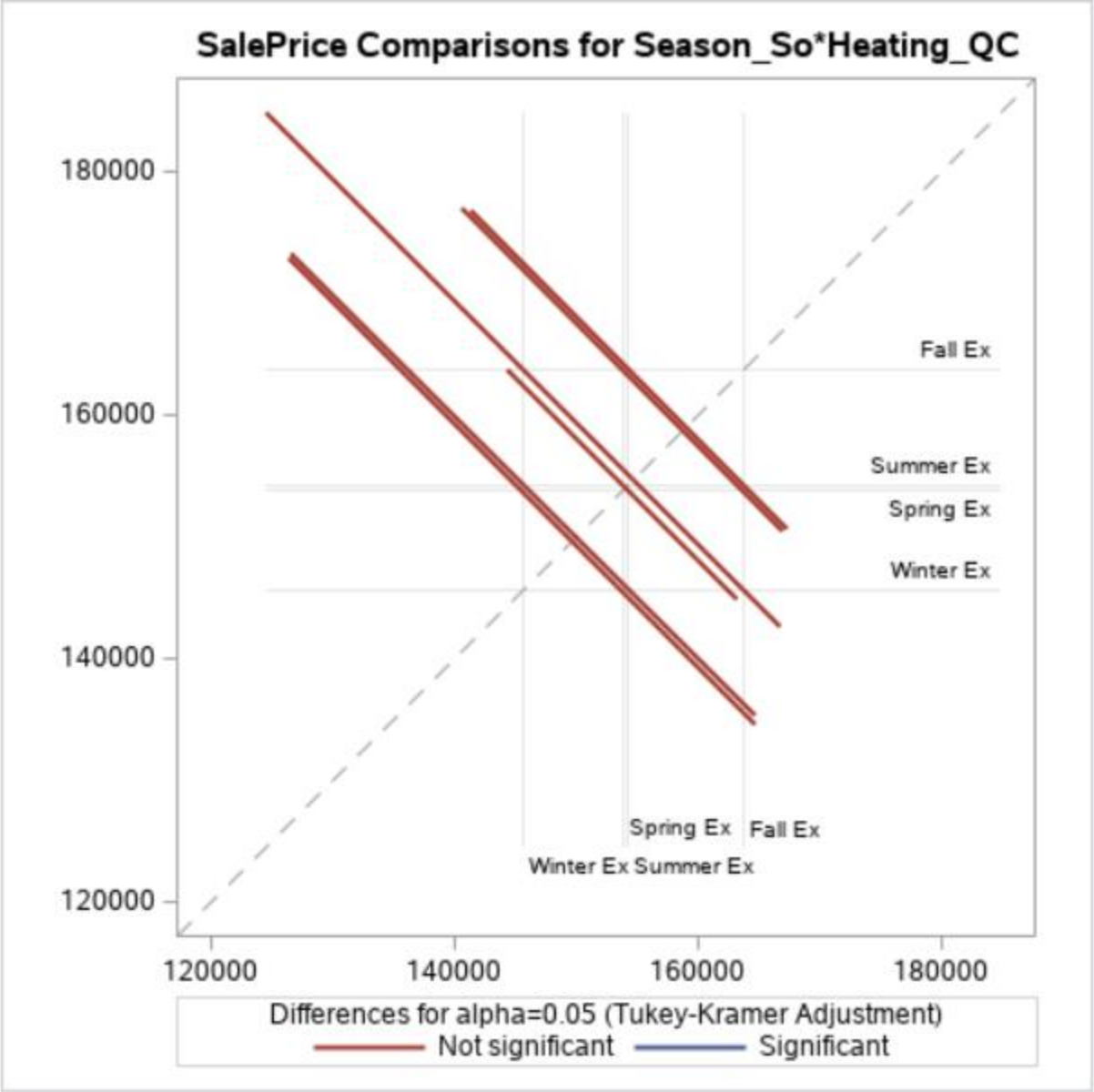
The PLM Procedure

Store Information	
Item Store	WORK.INTERACT
Data Set Created From	STAT1.AMESHOUSING3
Created By	PROC GLM
Date Created	22AUG21:04:57:28
Response Variable	SalePrice
Class Variables	Season_Sold Heating_QC
Model Effects	Intercept Heating_QC Season_Sold Season_So*Heating_QC

Class Level Information		
Class	Levels	Values
Season_Sold	4	Winter Spring Summer Fall
Heating_QC	4	Ex Fa Gd TA

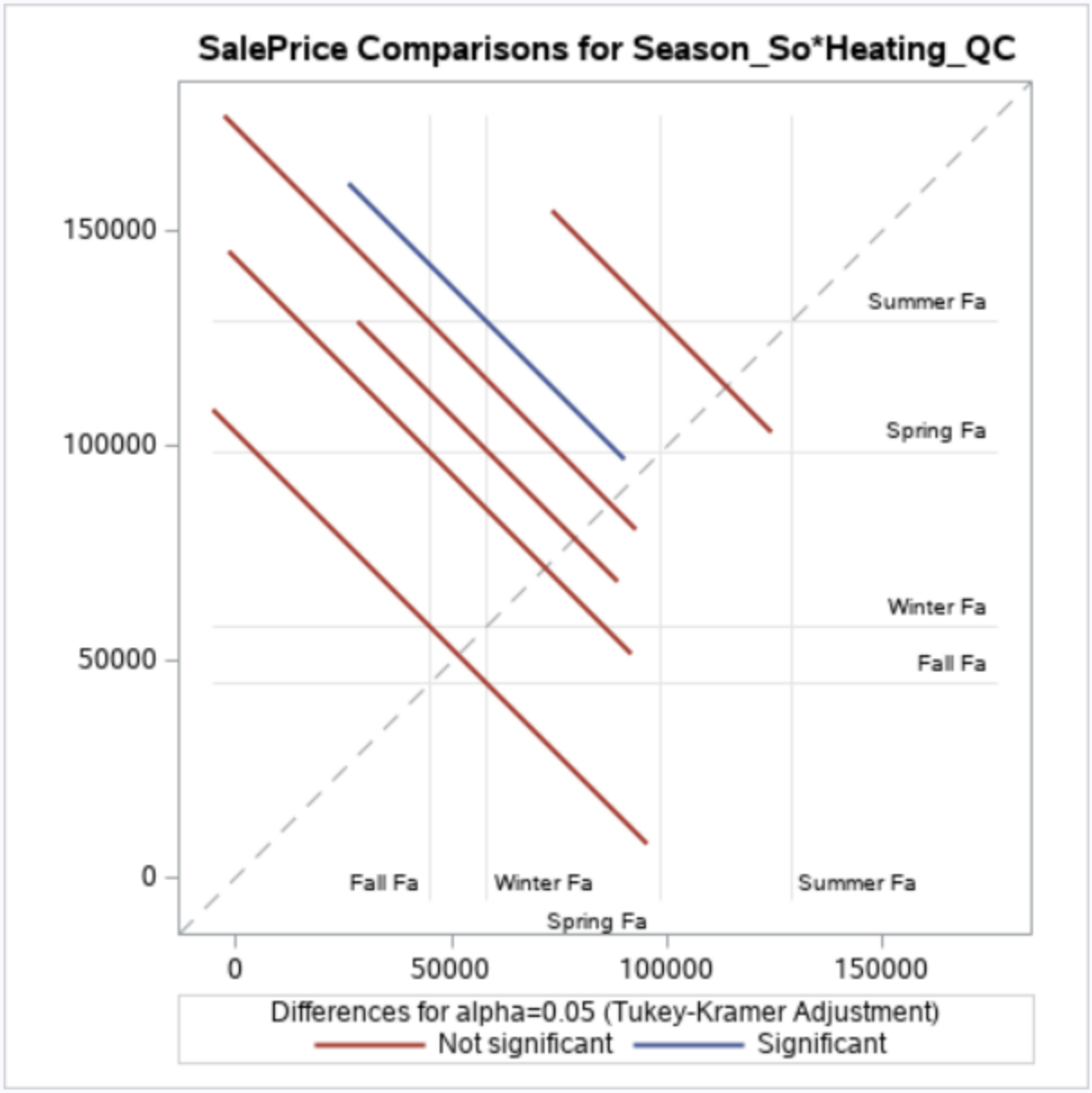
F Test for Season_So*Heating_QC Least Squares Means Slice				
Slice	Num DF	Den DF	F Value	Pr > F
Heating_QC Ex	3	284	0.51	0.6746

Simple Differences of Season_So*Heating_QC Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer								
Slice	Season when house sold	Season when house sold	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P
Heating_QC Ex	Winter	Spring	-8181.91	14800	284	-0.55	0.5808	0.9457
Heating_QC Ex	Winter	Summer	-8696.09	14716	284	-0.59	0.5550	0.9348
Heating_QC Ex	Winter	Fall	-18144	16356	284	-1.11	0.2682	0.6841
Heating_QC Ex	Spring	Summer	-514.18	7310.43	284	-0.07	0.9440	0.9999
Heating_QC Ex	Spring	Fall	-9961.69	10218	284	-0.97	0.3304	0.7638
Heating_QC Ex	Summer	Fall	-9447.51	10095	284	-0.94	0.3502	0.7856



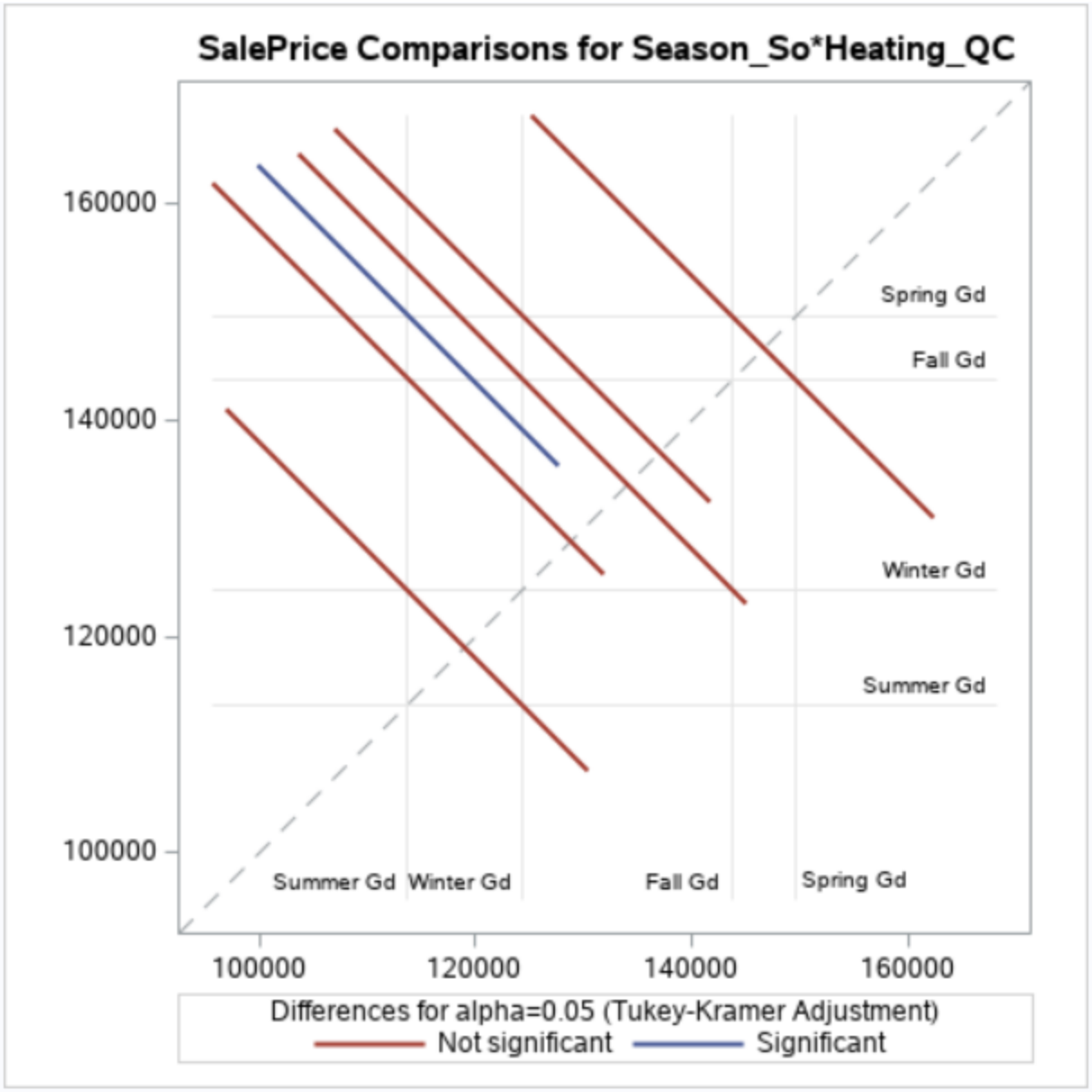
F Test for Season_So*Heating_QC Least Squares Means Slice				
Slice	Num DF	Den DF	F Value	Pr > F
Heating_QC Fa	3	284	3.58	0.0143

Simple Differences of Season_So*Heating_QC Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer								
Slice	Season when house sold	Season when house sold	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P
Heating_QC Fa	Winter	Spring	-40557	23366	284	-1.74	0.0837	0.3071
Heating_QC Fa	Winter	Summer	-70700	24728	284	-2.86	0.0046	0.0235
Heating_QC Fa	Winter	Fall	13100	39099	284	0.34	0.7378	0.9870
Heating_QC Fa	Spring	Summer	-30143	19827	284	-1.52	0.1295	0.4267
Heating_QC Fa	Spring	Fall	53657	36198	284	1.48	0.1394	0.4495
Heating_QC Fa	Summer	Fall	83800	37092	284	2.26	0.0246	0.1102



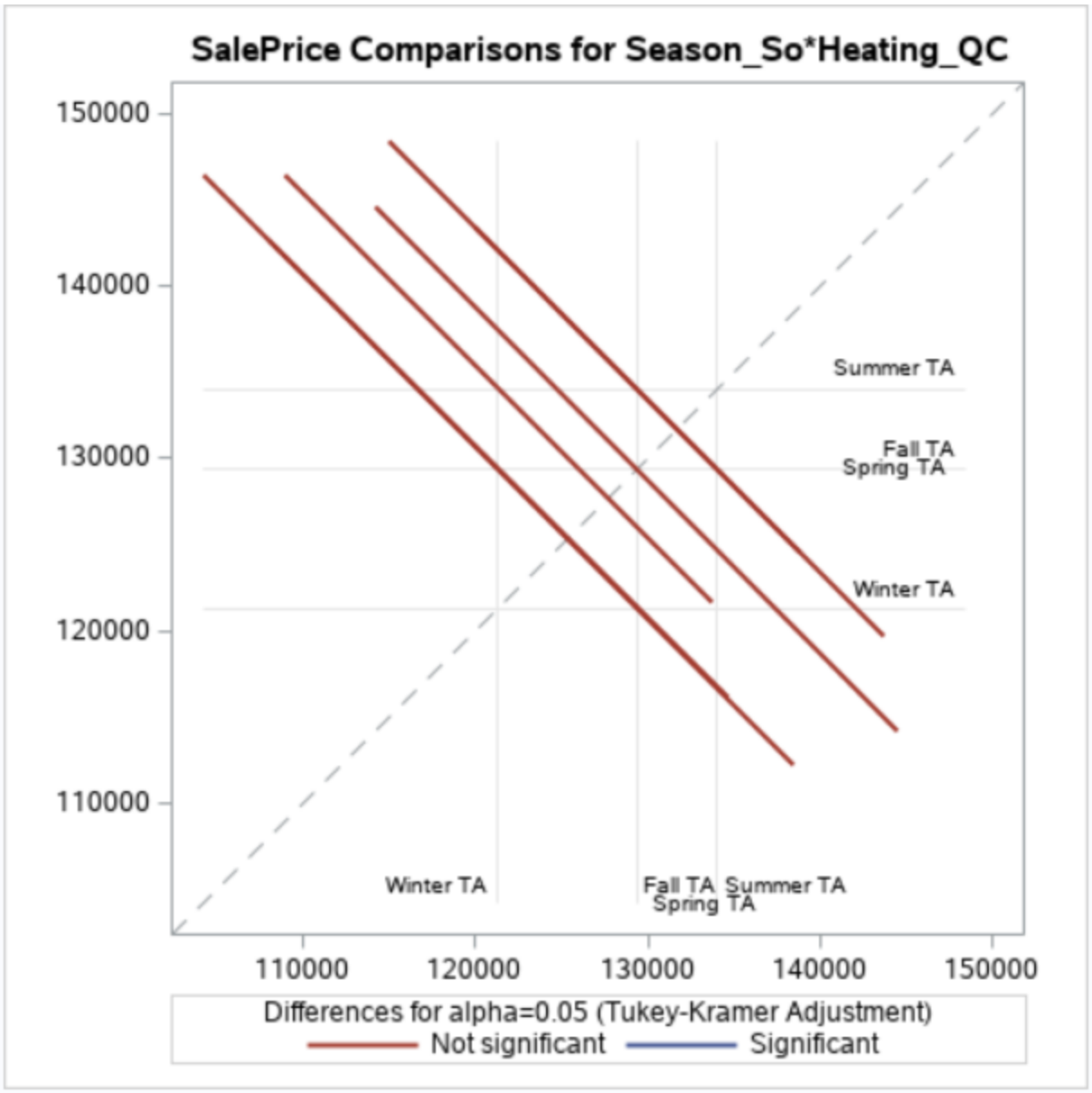
F Test for Season_So*Heating_QC Least Squares Means Slice				
Slice	Num DF	Den DF	F Value	Pr > F
Heating_QC Gd	3	284	4.23	0.0060

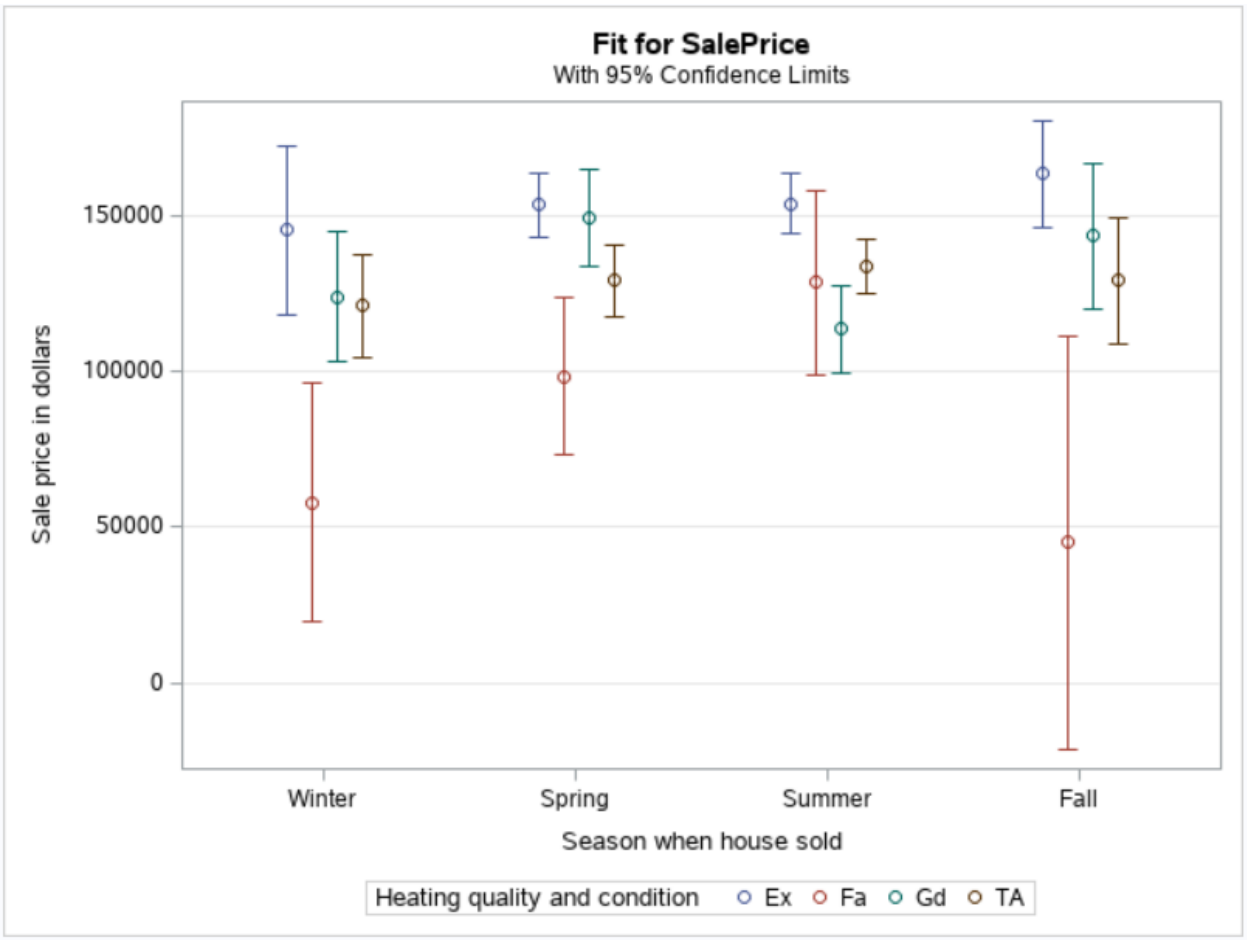
Simple Differences of Season_So*Heating_QC Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer								
Slice	Season when house sold	Season when house sold	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P
Heating_QC Gd	Winter	Spring	-25290	13355	284	-1.89	0.0593	0.2330
Heating_QC Gd	Winter	Summer	10603	12914	284	0.82	0.4123	0.8445
Heating_QC Gd	Winter	Fall	-19483	16061	284	-1.21	0.2261	0.6191
Heating_QC Gd	Spring	Summer	35893	10762	284	3.34	0.0010	0.0053
Heating_QC Gd	Spring	Fall	5807.33	14388	284	0.40	0.6868	0.9777
Heating_QC Gd	Summer	Fall	-30085	13980	284	-2.15	0.0322	0.1394



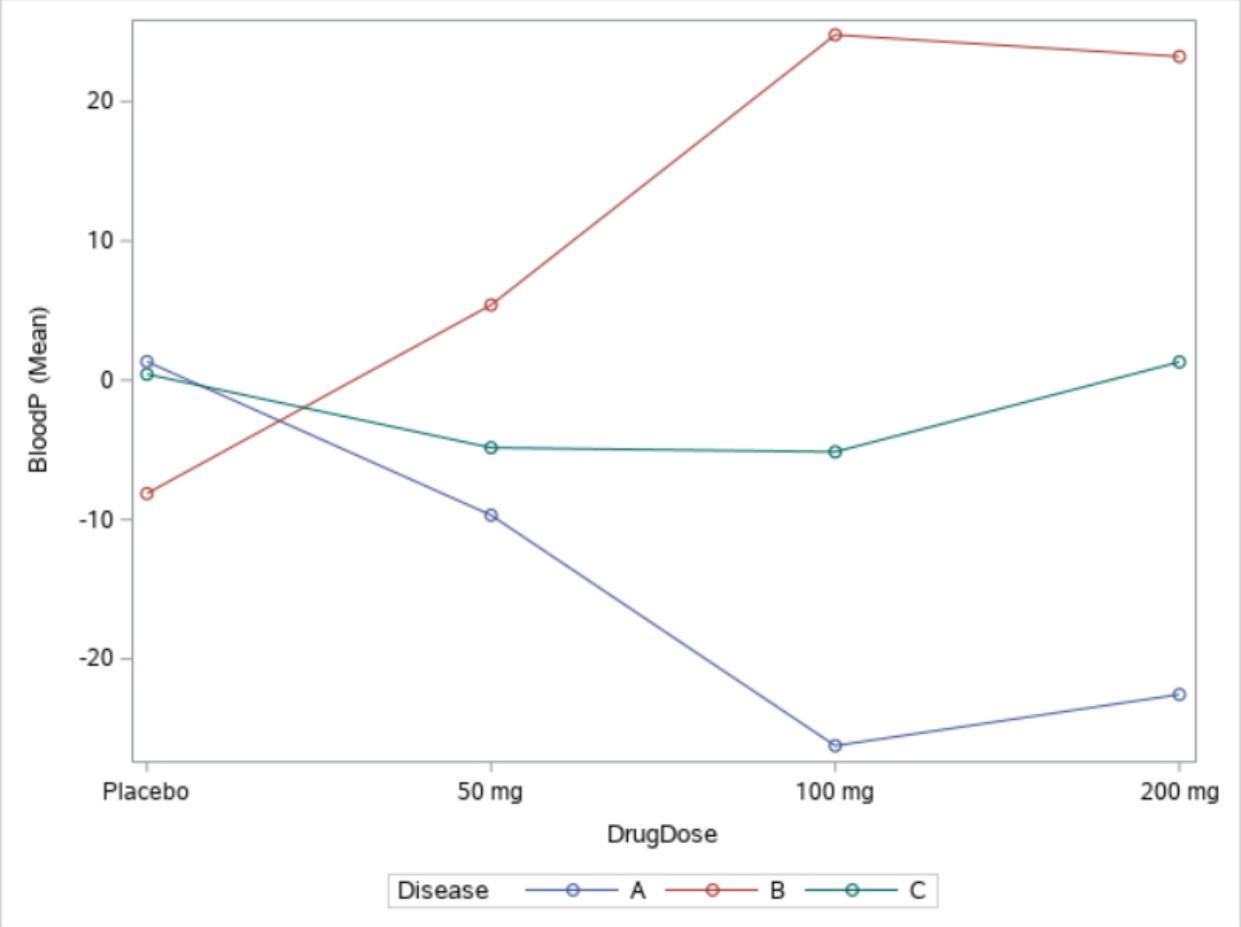
F Test for Season_So*Heating_QC Least Squares Means Slice				
Slice	Num DF	Den DF	F Value	Pr > F
Heating_QC TA	3	284	0.62	0.6021

Simple Differences of Season_So*Heating_QC Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer								
Slice	Season when house sold	Season when house sold	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P
Heating_QC TA	Winter	Spring	-8091.91	10265	284	-0.79	0.4312	0.8598
Heating_QC TA	Winter	Summer	-12734	9561.68	284	-1.33	0.1840	0.5434
Heating_QC TA	Winter	Fall	-8032.95	13262	284	-0.61	0.5452	0.9302
Heating_QC TA	Spring	Summer	-4642.14	7313.62	284	-0.63	0.5261	0.9207
Heating_QC TA	Spring	Fall	58.9572	11745	284	0.01	0.9960	1.0000
Heating_QC TA	Summer	Fall	4701.10	11135	284	0.42	0.6732	0.9747


SalePrice Comparisons for Season_So*Heating_QC
150000
140000
130000
120000
110000
Summer TA
Fall TA
Spring TA
Winter TA
Winter TA
Fall TA
Spring TA
Summer TA
110000
120000
130000
140000
150000
Differences for alpha=0.05 (Tukey-Kramer Adjustment)
Not significant
Significant


Fit for SalePrice
With 95% Confidence Limits
Sale price in dollars
150000
100000
50000
0
Winter
Spring
Summer
Fall
Season when house sold
Heating quality and condition
Ex
Fa
Gd
TA

```
/*st103s01.sas*/  /*Part A*/
proc sgplot data=STAT1.drug;
    vline DrugDose / group=Disease
                     stat=mean
                     response=BloodP
                     markers;
    format DrugDose dosefmt.;
run;
```



```
/*st103s01.sas*/  /*Part B*/
ods graphics on;

proc glm data=STAT1.drug plots(only)=intplot;
    class DrugDose Disease;
    model BloodP = DrugDose|Disease;
    lsmeans DrugDose*Disease / slice=Disease;
run;
quit;

title;
```

The GLM Procedure

Class Level Information		
Class	Levels	Values
DrugDose	4	1 2 3 4
Disease	3	A B C

Number of Observations Read	170
Number of Observations Used	170

The GLM Procedure

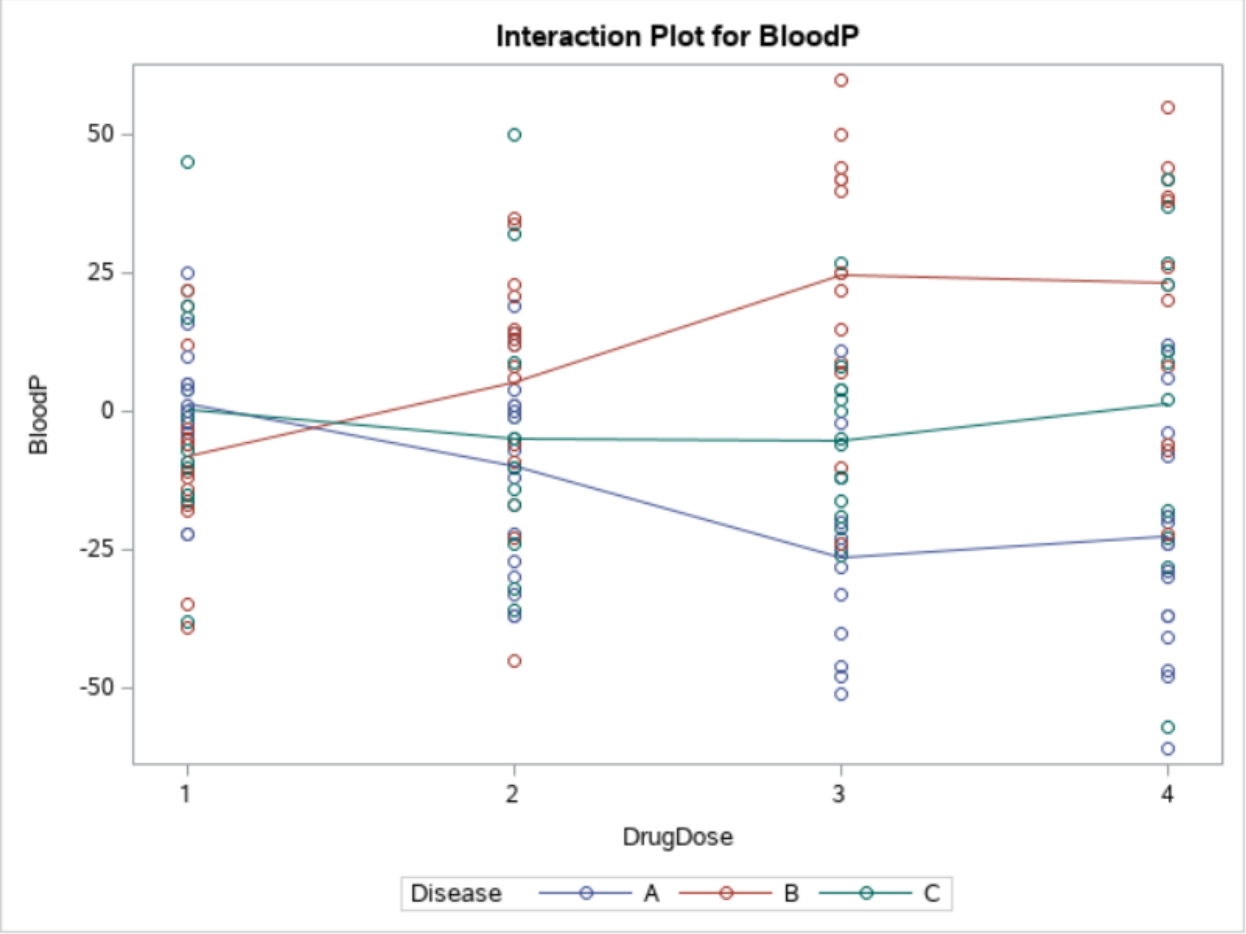
Dependent Variable: BloodP

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	36476.8353	3316.0759	7.66	<.0001
Error	158	68366.4589	432.6991		
Corrected Total	169	104843.2941			

R-Square	Coeff Var	Root MSE	BloodP Mean
0.347918	-906.7286	20.80142	-2.294118

Source	DF	Type I SS	Mean Square	F Value	Pr > F
DrugDose	3	54.03137	18.01046	0.04	0.9886
Disease	2	19276.48690	9638.24345	22.27	<.0001
DrugDose*Disease	6	17146.31698	2857.71950	6.60	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
DrugDose	3	335.73526	111.91175	0.26	0.8551
Disease	2	18742.62386	9371.31193	21.66	<.0001
DrugDose*Disease	6	17146.31698	2857.71950	6.60	<.0001

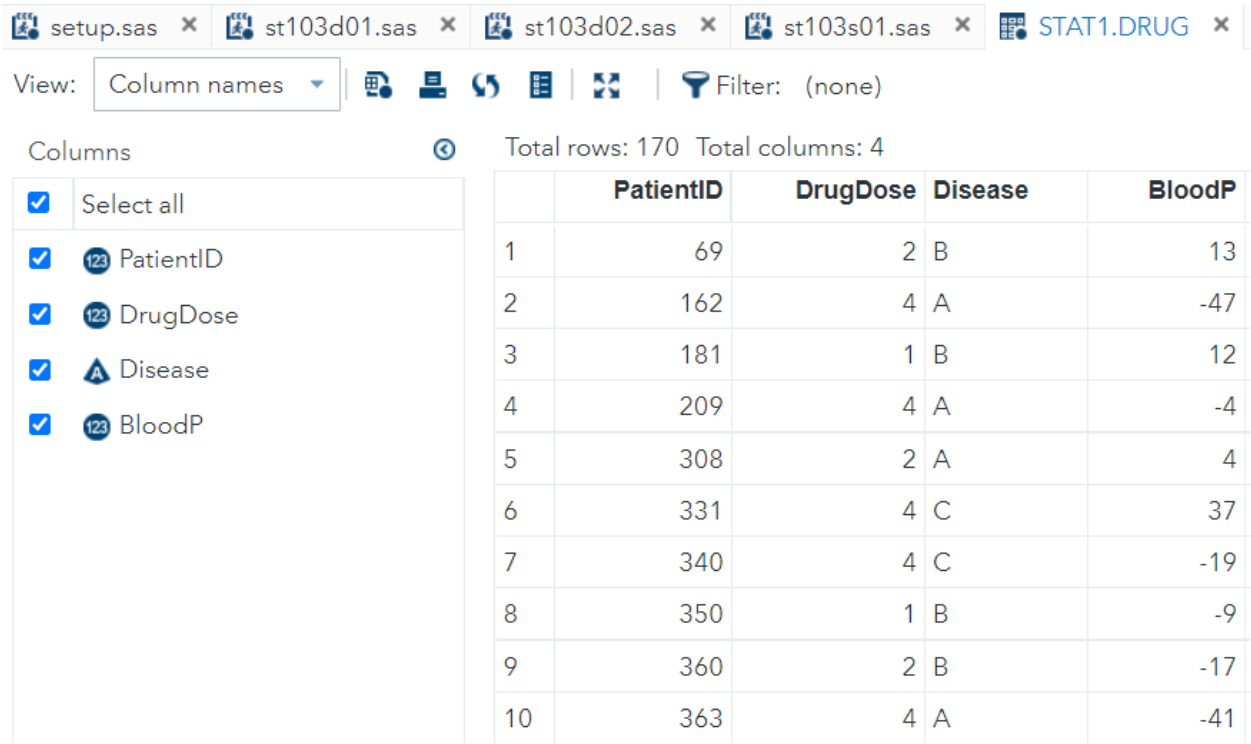

Interaction Plot for BloodP
BloodP
50
25
0
-25
-50
1
2
3
4
DrugDose
Disease
A
B
C

The GLM Procedure
Least Squares Means

DrugDose	Disease	BloodP LSMEAN
1	A	1.3333333
1	B	-8.1333333
1	C	0.4285714
2	A	-9.6875000
2	B	5.4000000
2	C	-4.8461538
3	A	-26.2307692
3	B	24.7857143
3	C	-5.1428571
4	A	-22.5555556
4	B	23.2307692
4	C	1.3076923

The GLM Procedure
Least Squares Means

DrugDose*Disease Effect Sliced by Disease for BloodP					
Disease	DF	Sum of Squares	Mean Square	F Value	Pr > F
A	3	6320.126747	2106.708916	4.87	0.0029
B	3	10561	3520.222833	8.14	<.0001
C	3	468.099308	156.033103	0.36	0.7815



	PatientID	DrugDose	Disease	BloodP
1	69	2	B	13
2	162	4	A	-47
3	181	1	B	12
4	209	4	A	-4
5	308	2	A	4
6	331	4	C	37
7	340	4	C	-19
8	350	1	B	-9
9	360	2	B	-17
10	363	4	A	-41

Practice - Performing a Two-Way ANOVA Using PROC GLM

Question 1

Data were collected to determine whether different dosage levels of a drug have an effect on blood pressure for people with one of three types of heart disease. The data are in the **stat1.drug** data set.

1. Examine the data with a vertical line plot. Put **BloodP** on the Y axis, and **DrugDose** on the X axis, and then stratify by **Disease**.
2. What information can you obtain by looking at the data?

Given DrugDose Placebo, Disease B shows BloodP mean of -10 while Disease A and C show close to zero. Given DrugDose 50mg ,Disease B BloodP mean jumped to +7 while Disease A dropped to -10 and Disease B dropped to -5 Given DrugDose 100mg and 200mg, Disease B BloodP mean jumped to more than +20 while Disease A dropped to below -20 and Disease C shows -5 and 0, respectively.

Correct

It seems that the drug dose affects a change in blood pressure. However, that effect is not consistent across diseases. Higher doses result in increased blood pressure for patients with disease B, decreased blood pressure for patients with disease A, and little change in blood pressure for patients with disease C.

Solution code:

```
/*st103s01.sas*/ /*Part A*/

proc sgplot data=STAT1.drug;
```

```
   vline DrugDose / group=Disease stat=mean response=BloodP markers;

   format DrugDose dosefmt.;

run;
```

Question 2

1. Test the hypothesis that the means are equal. Be sure to include an interaction term if the graphical analysis that you performed indicates that would be advisable.
2. What conclusions can you reach at this point?

R-square = 0.35

DrugDose alone, the hypothesis that the means are equal is not significant. With interaction term of DrugDose*Disease, the hypothesis that the means are equal is significant.

Correct

The global *F* test indicates a significant difference among the different groups. Because the interaction is in the model, this is a test of all combinations of **DrugDose*Disease** against all other combinations. The R-square value implies that approximately 35% of the variation in **BloodP** can be explained by variations in the explanatory variables. The interaction term is statistically significant, as predicted by the plot of the means.

Solution code:

```
/*st103s01.sas*/ /*Part B*/

ods graphics on;

proc glm data=STAT1.drug plots(only)=intplot;

   class DrugDose Disease;

   model BloodP=DrugDose|Disease;

   lsmeans DrugDose*Disease;

run;

quit;
```

Question 3

1. To investigate the interaction effect between the two factors, include the SLICE option in the code you just wrote.

2. Is the effect of **DrugDose** significant?

The effect of DrugDose is significant for Disease A and B, but not significant for Disease C.

Correct

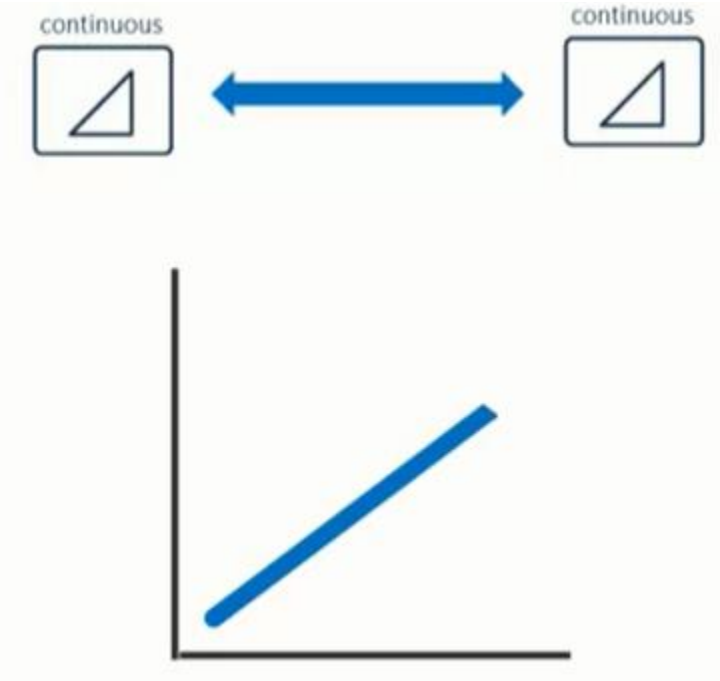
Modify the LSMEANS statement to include the *slice=Disease* option preceded by a slash, as shown in the code below.

```
/*st103s01.sas*/ /*Part B*/

ods graphics on;

proc glm data=STAT1.drug plots(only)=intplot;

   class DrugDose Disease;

   model BloodP=DrugDose|Disease;

   lsmeans DrugDose*Disease / slice=Disease;

run;

quit;
```

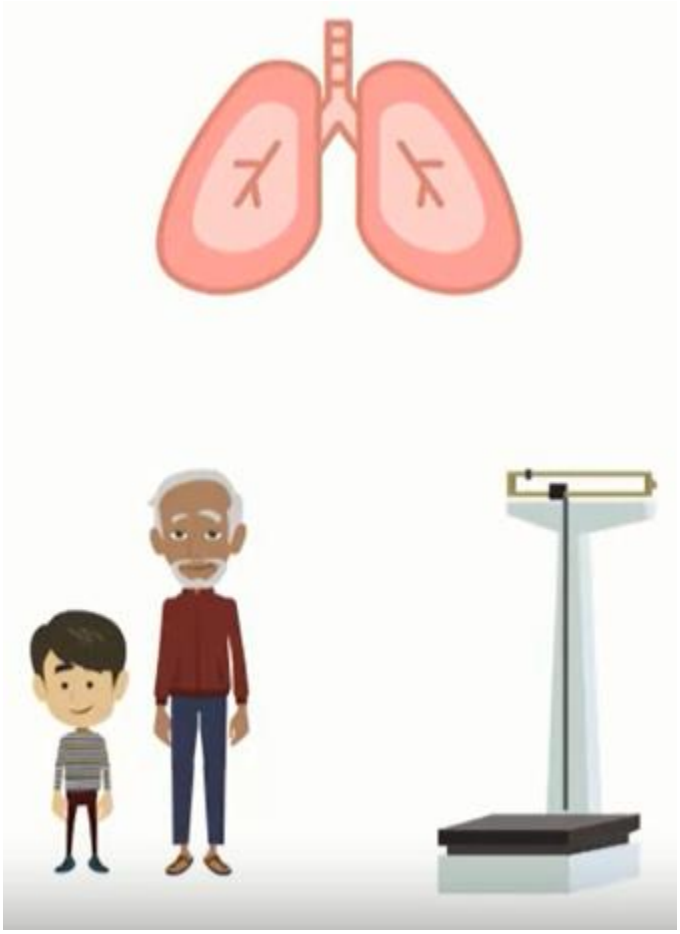
The slice table shows the effect of **DrugDose** at each level of the disease. The effect is significant for all, except Disease C.

Multiple Regression

Scenario

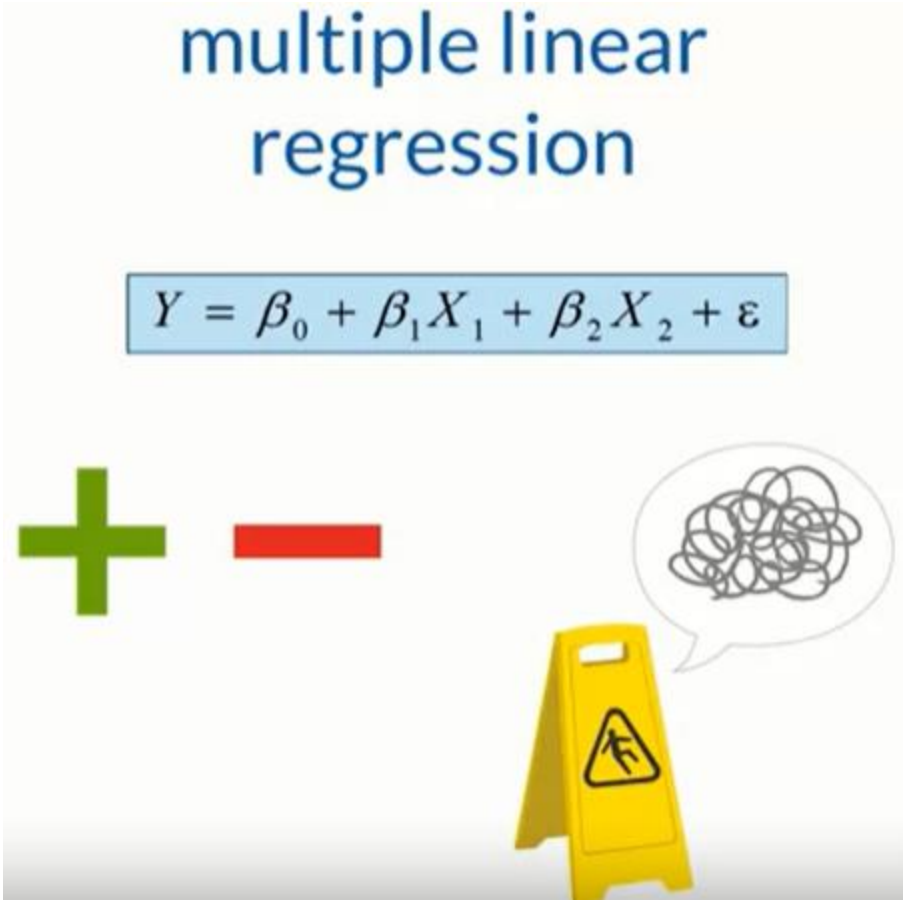



continuous
continuous
continuous

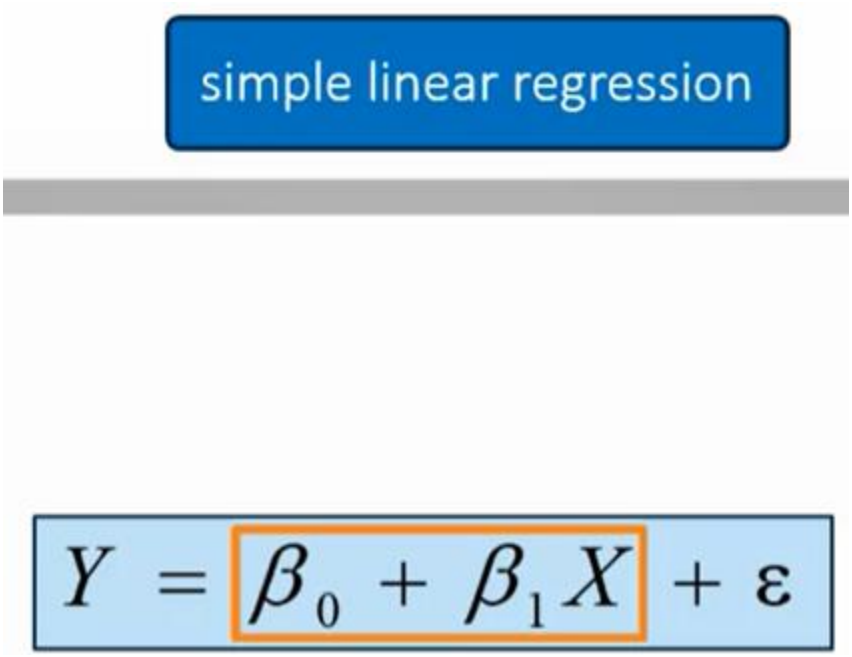



$$
SOLD
square feet
lot size


$$
SOLD
age
bathrooms


multiple linear regression
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

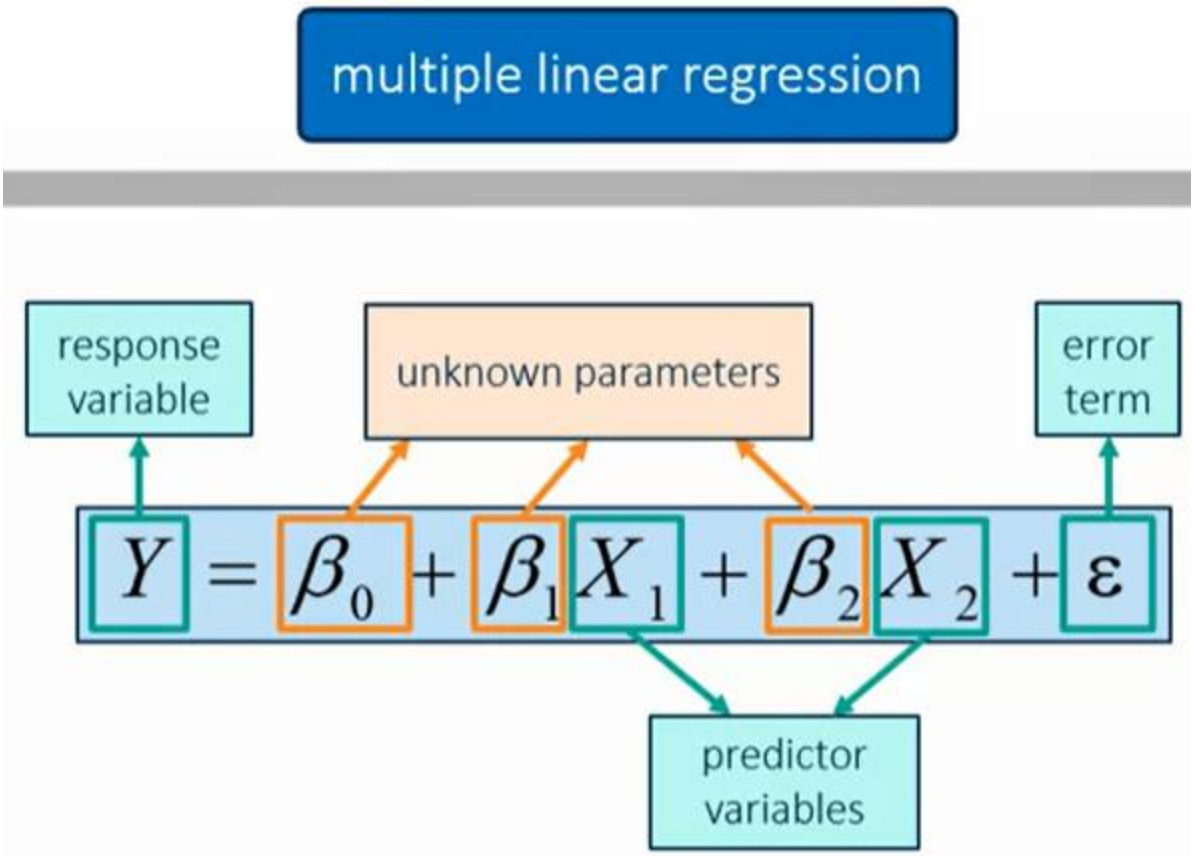
The Multiple Linear Regression Model

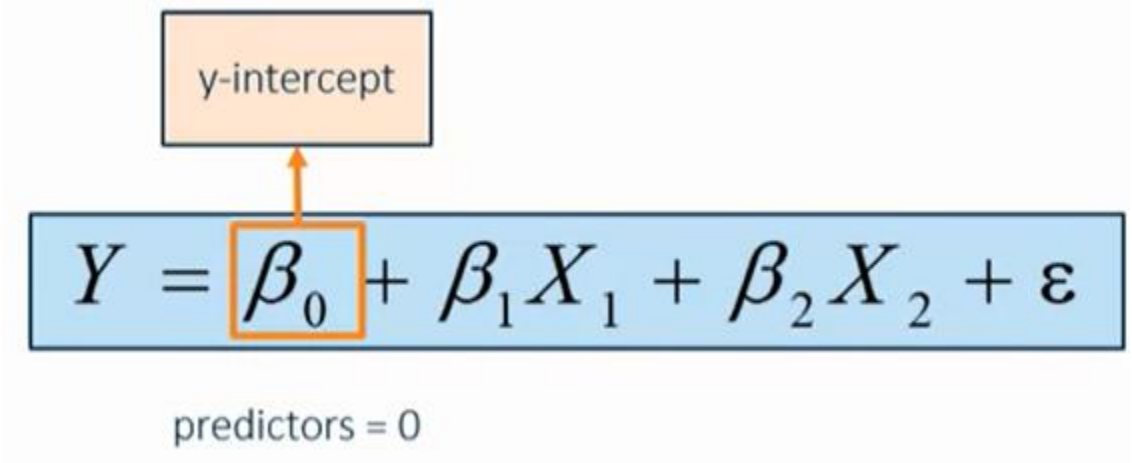


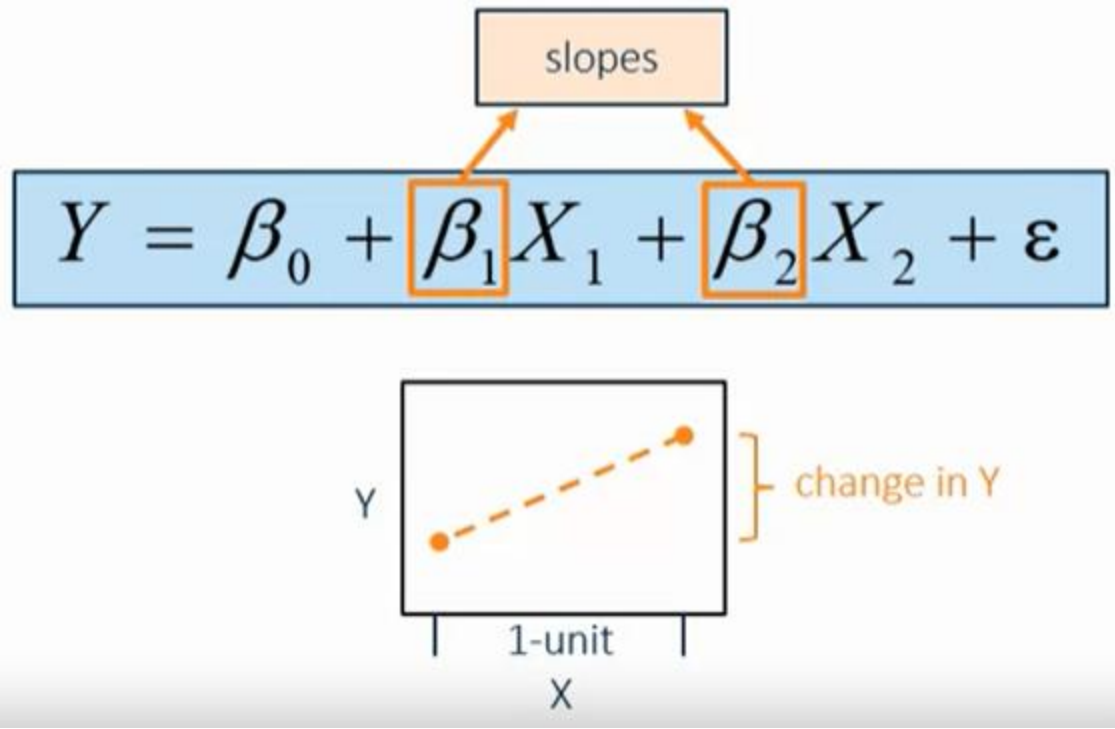
multiple linear regression

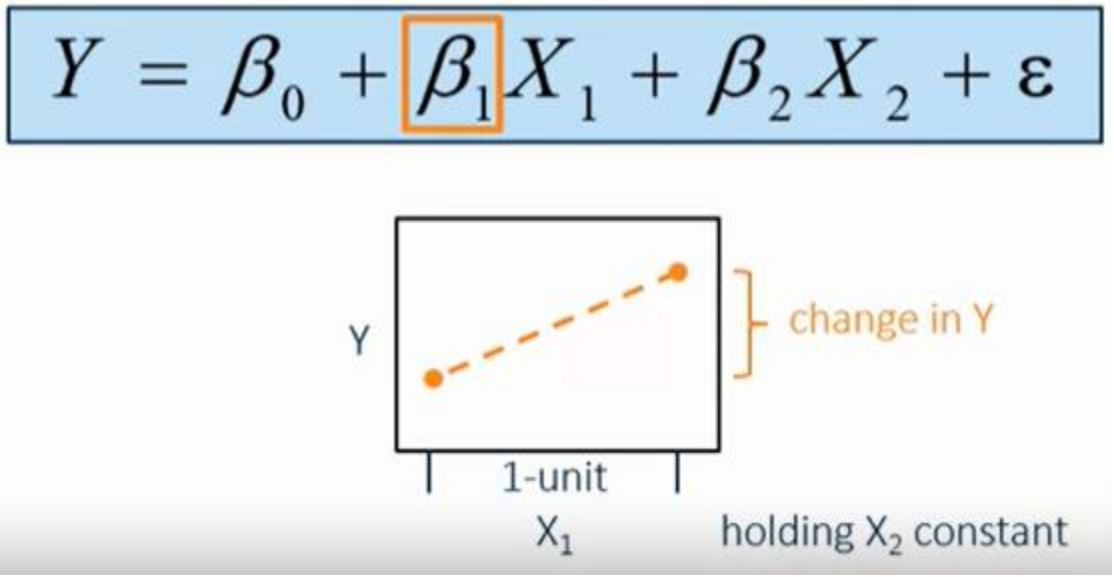
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

plane

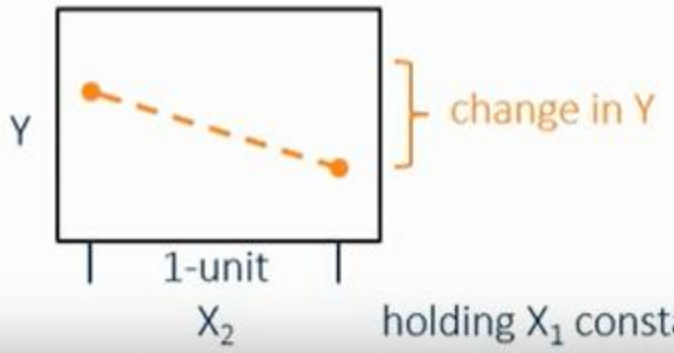

multiple linear regression
response variable
unknown parameters
error term
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
predictor variables


y-intercept
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
predictors = 0


slopes
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
Y
change in Y
1-unit
X


$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
Y
change in Y
1-unit
X_1
holding X_2 constant

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

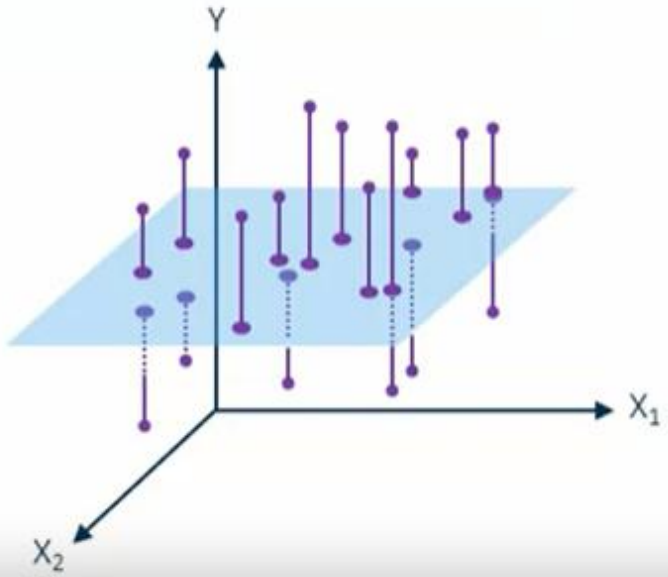


multiple linear regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Y ✗ X_1

✗ X_2



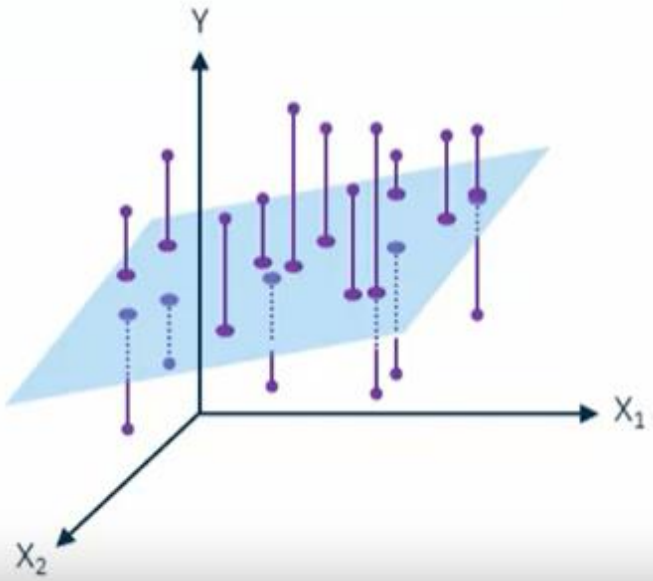
multiple linear regression

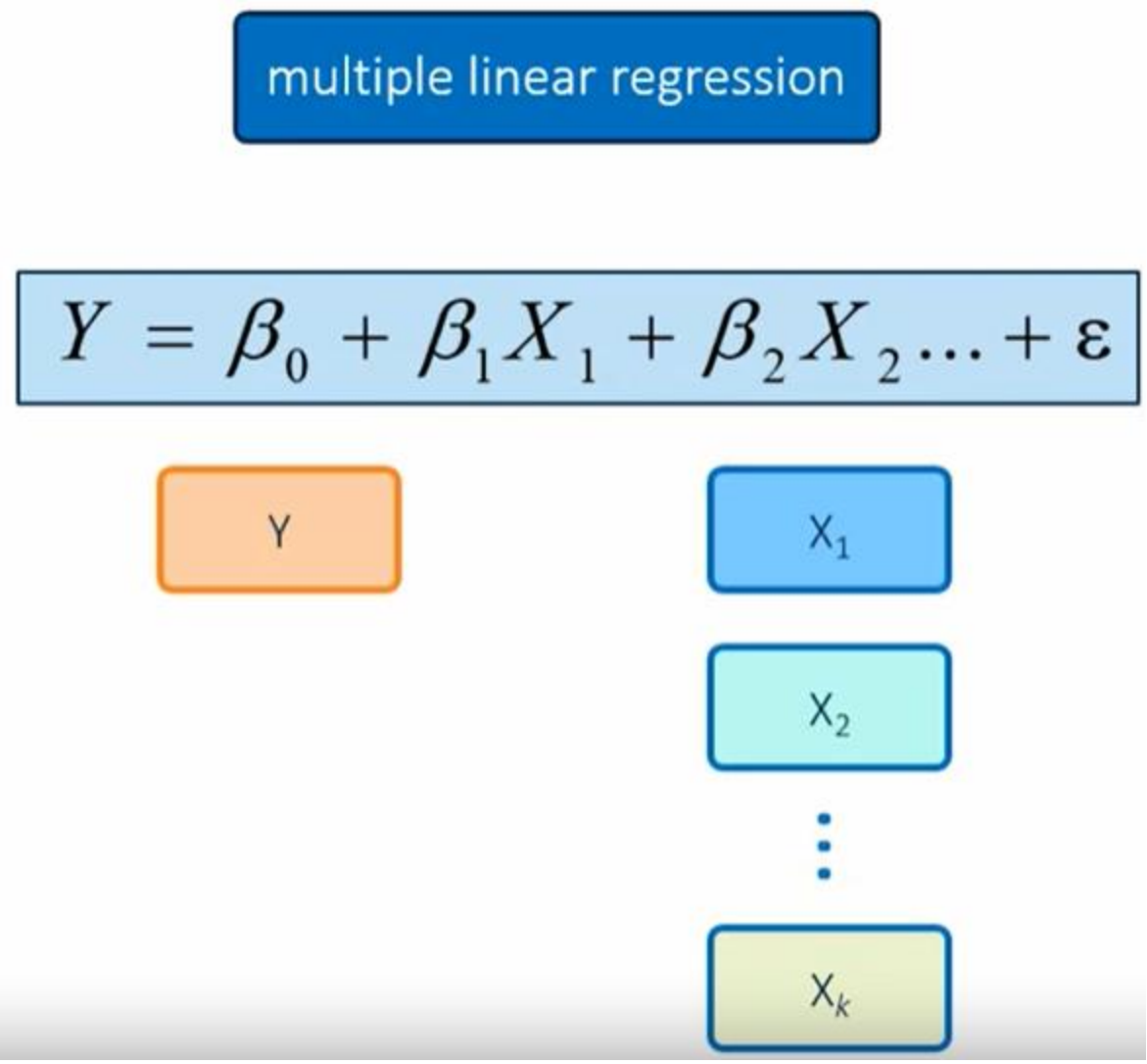
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Y

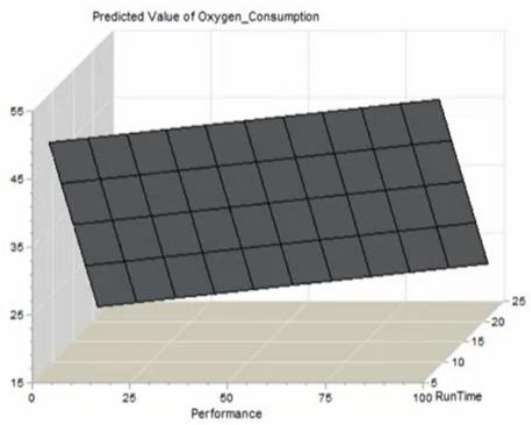
✓ X_1

✓ X_2

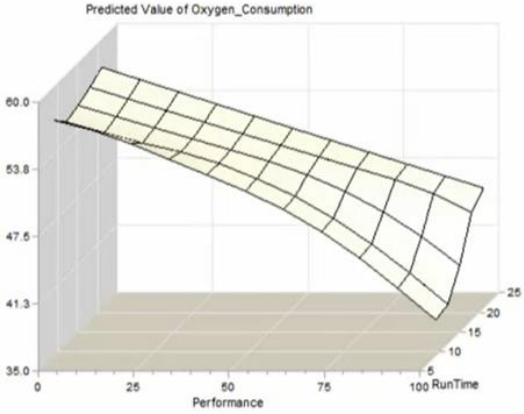



multiple linear regression
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \varepsilon$
Y
X_1
X_2
X_k

multiple linear regression

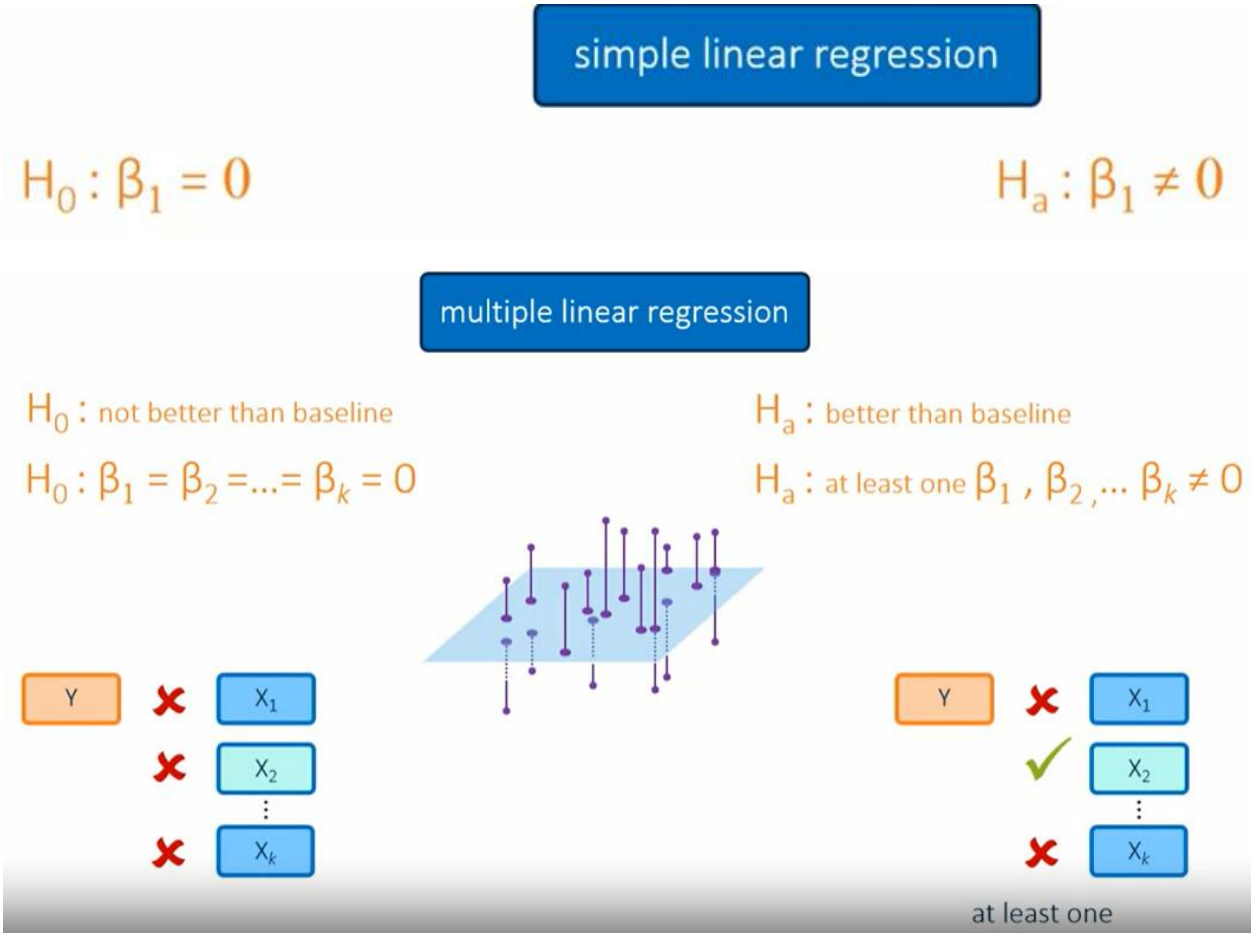


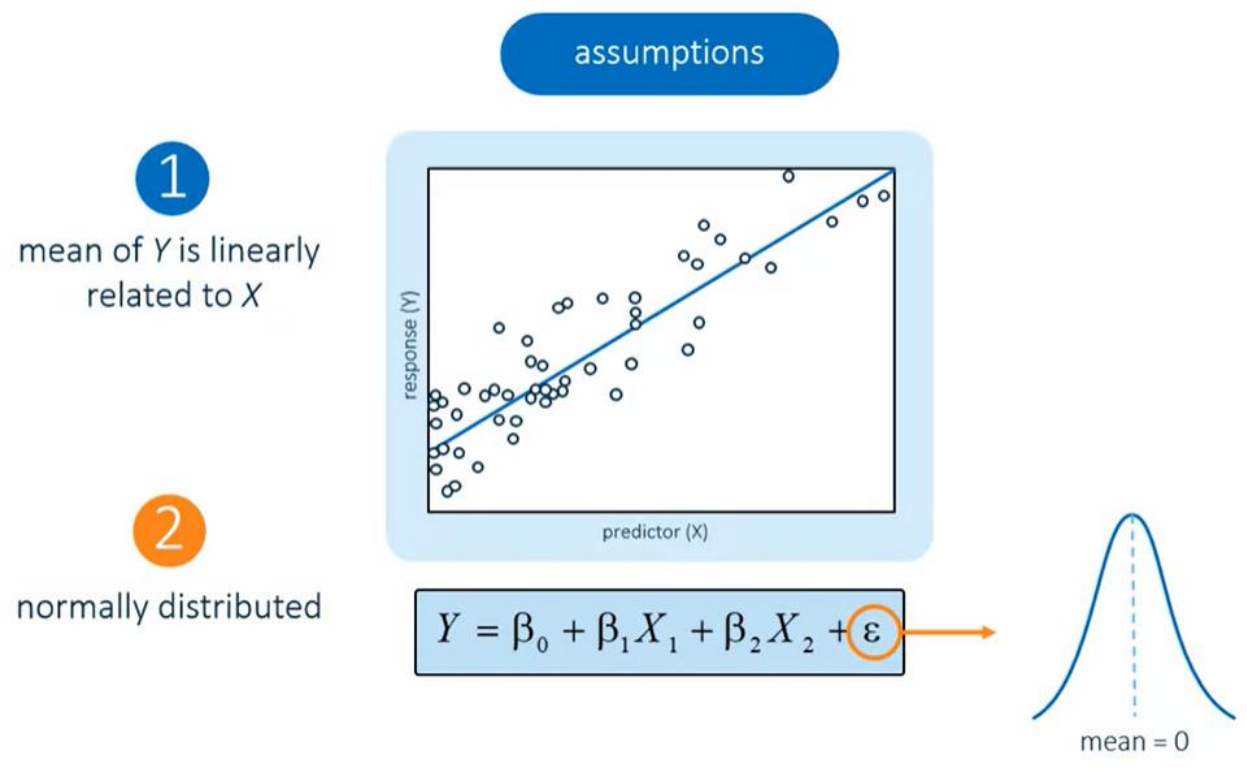
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

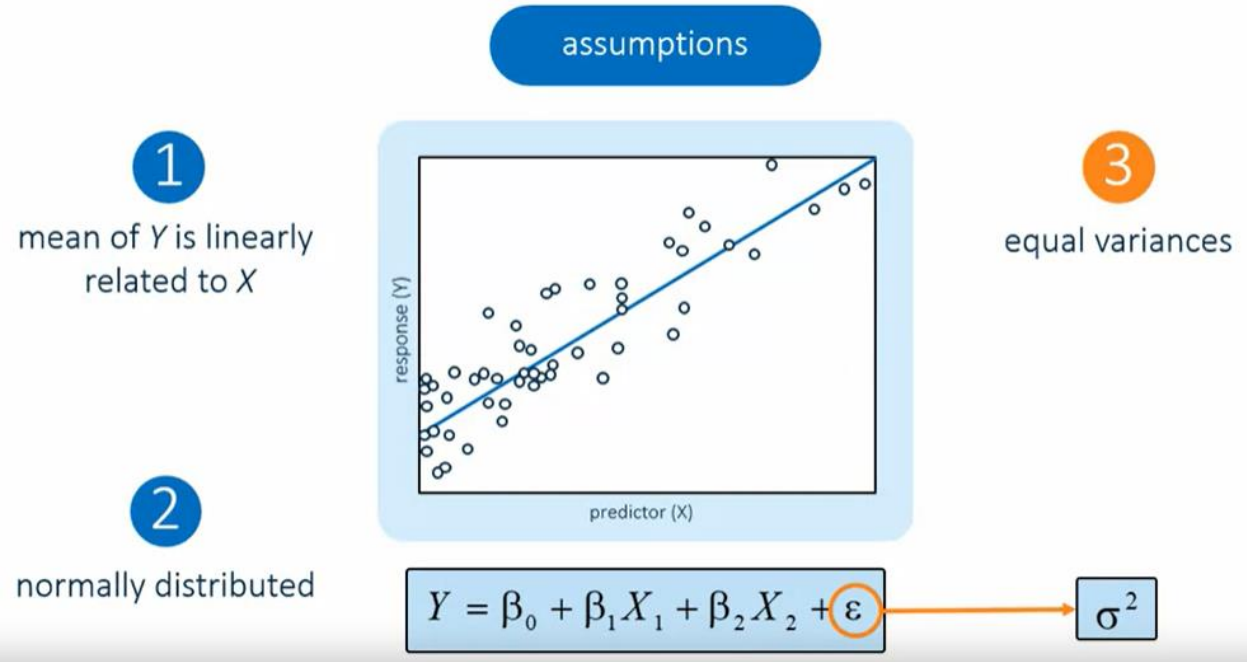


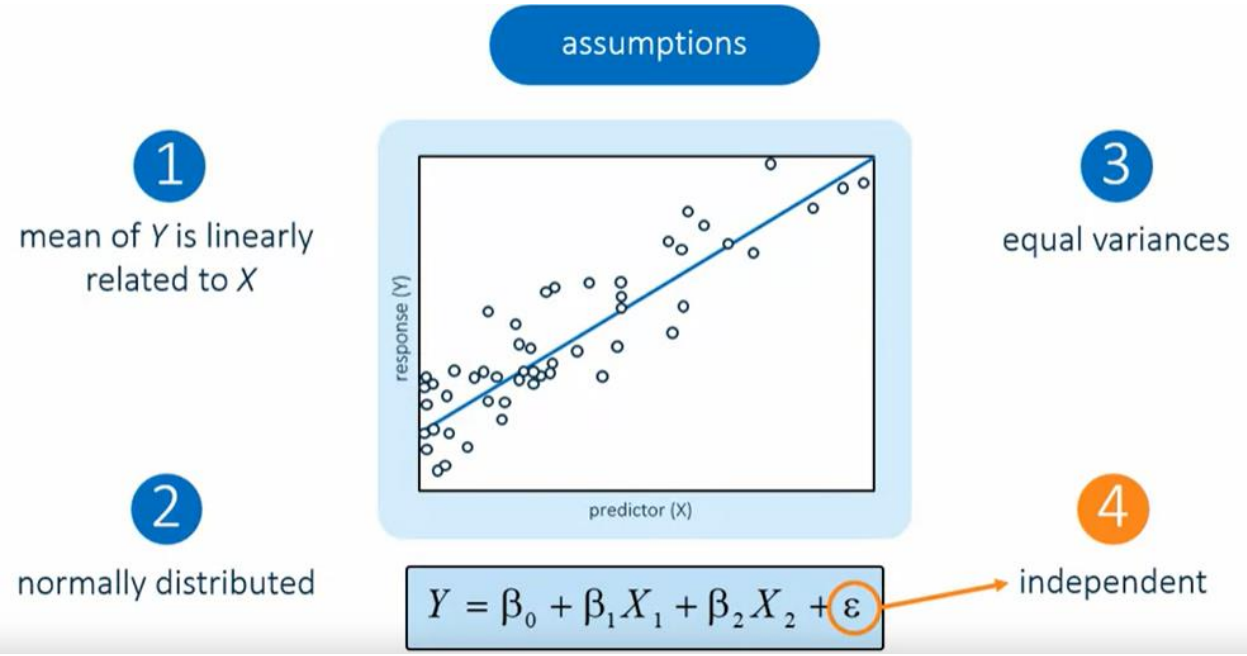
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^{\,2} + \beta_3 X_2 + \beta_4 X_2^{\,2} + \varepsilon$$

Hypothesis Testing for Multiple Regression

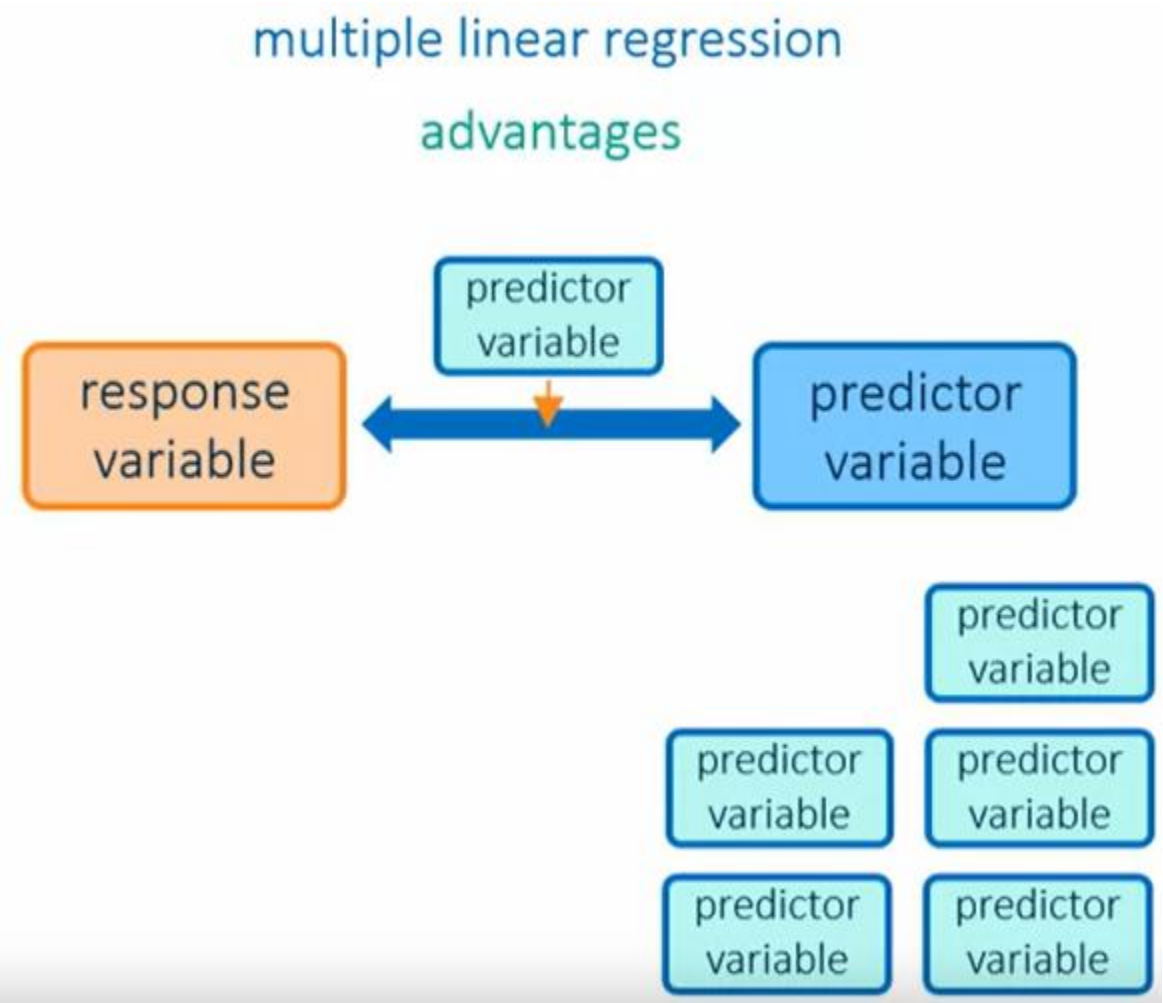


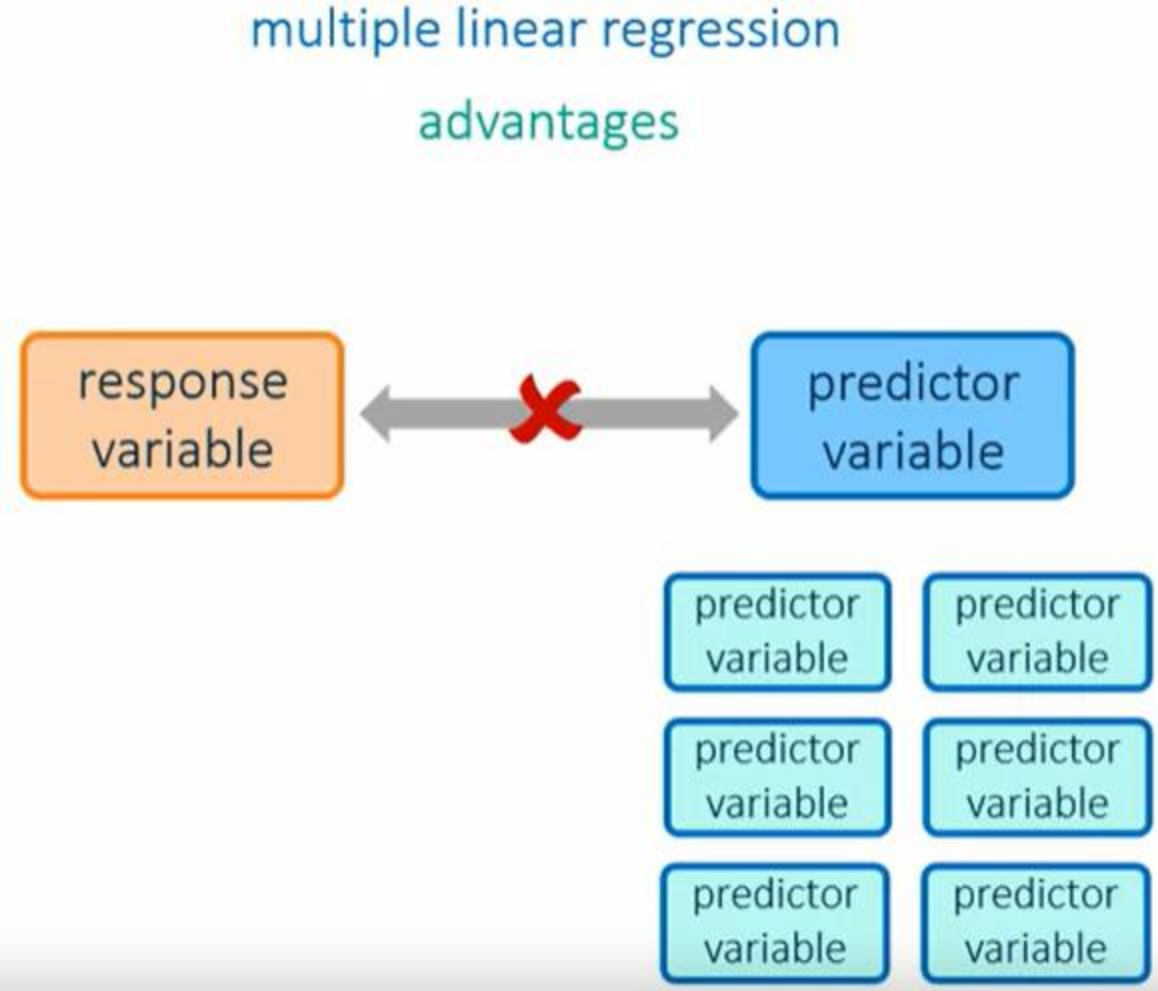

assumptions
1
mean of Y is linearly related to X
response (Y)
predictor (X)
2
normally distributed
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
mean = 0

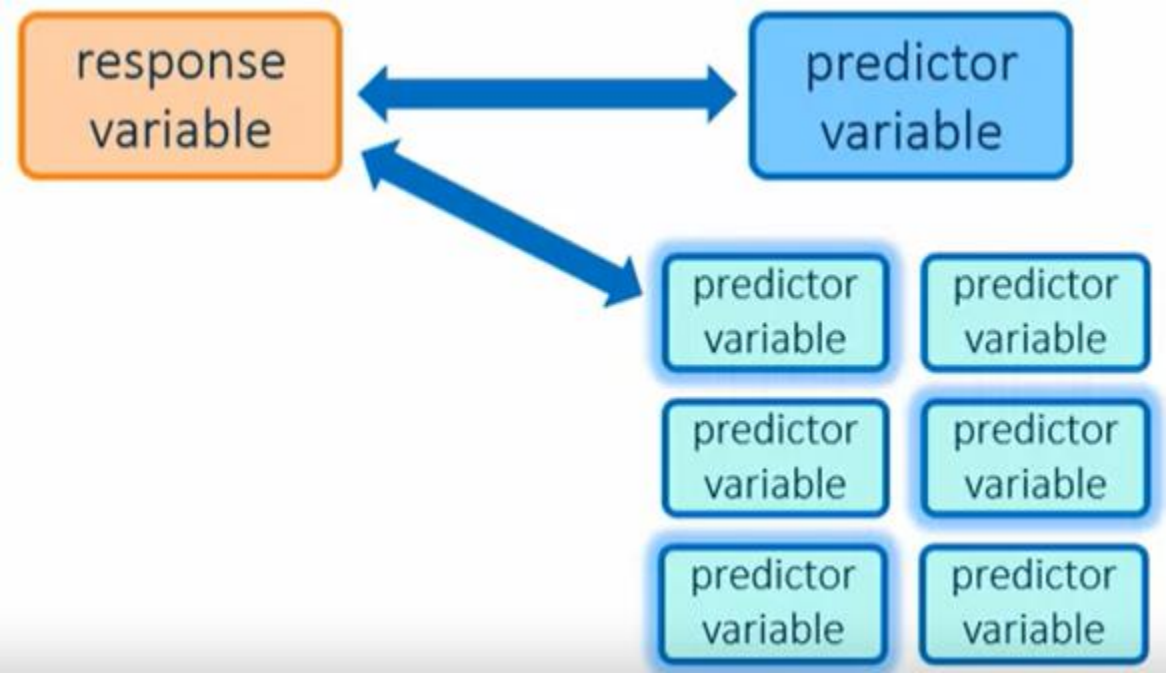

assumptions
1
mean of Y is linearly related to X
response (Y)
predictor (X)
3
equal variances
2
normally distributed
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
σ^2

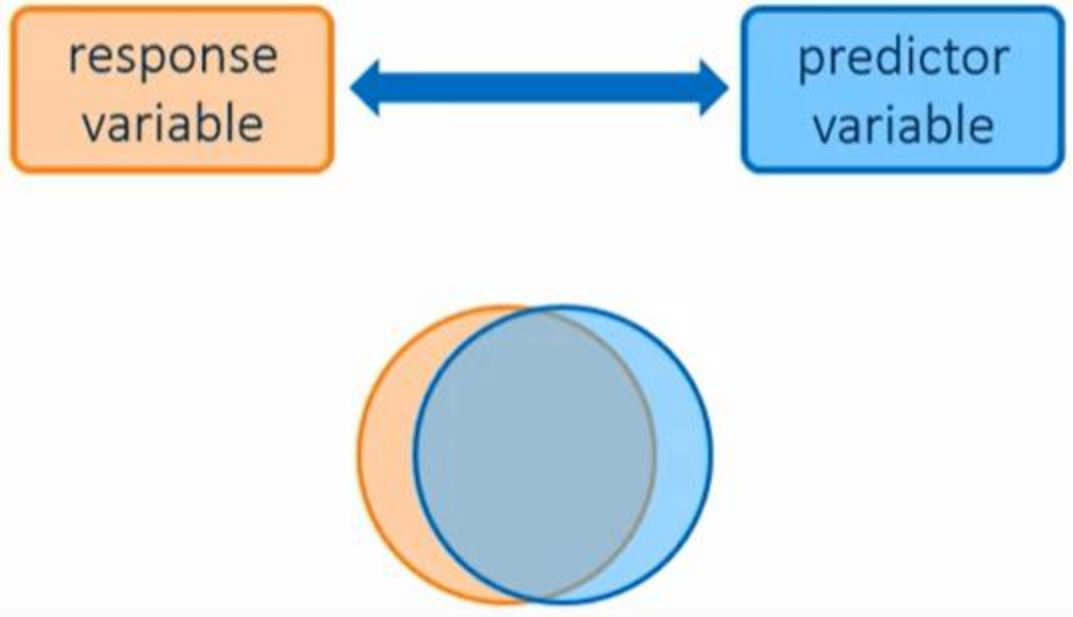

assumptions
1
mean of Y is linearly related to X
3
equal variances
response (Y)
predictor (X)
2
normally distributed
4
independent
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

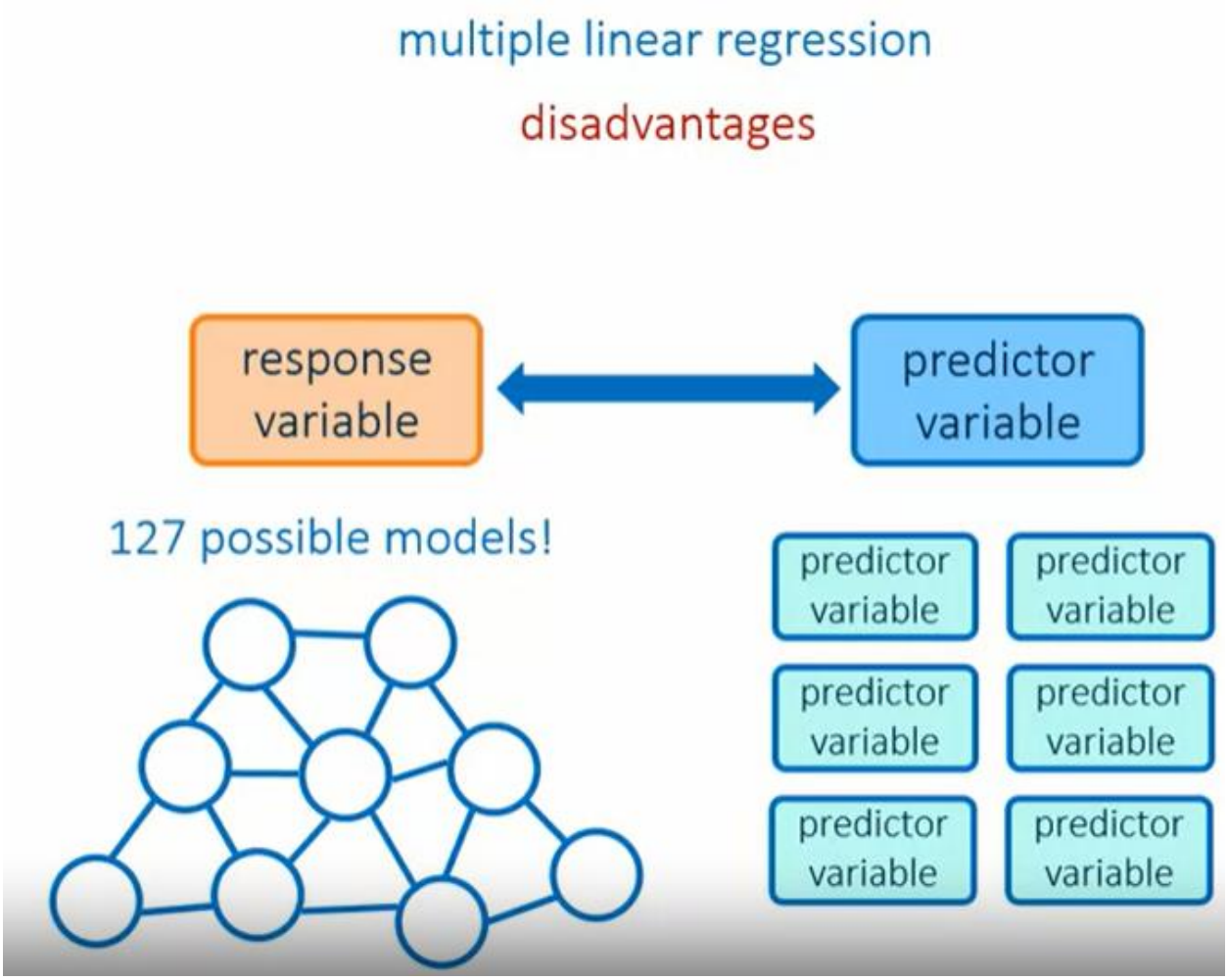
Multiple Linear Regression versus Simple Linear Regression

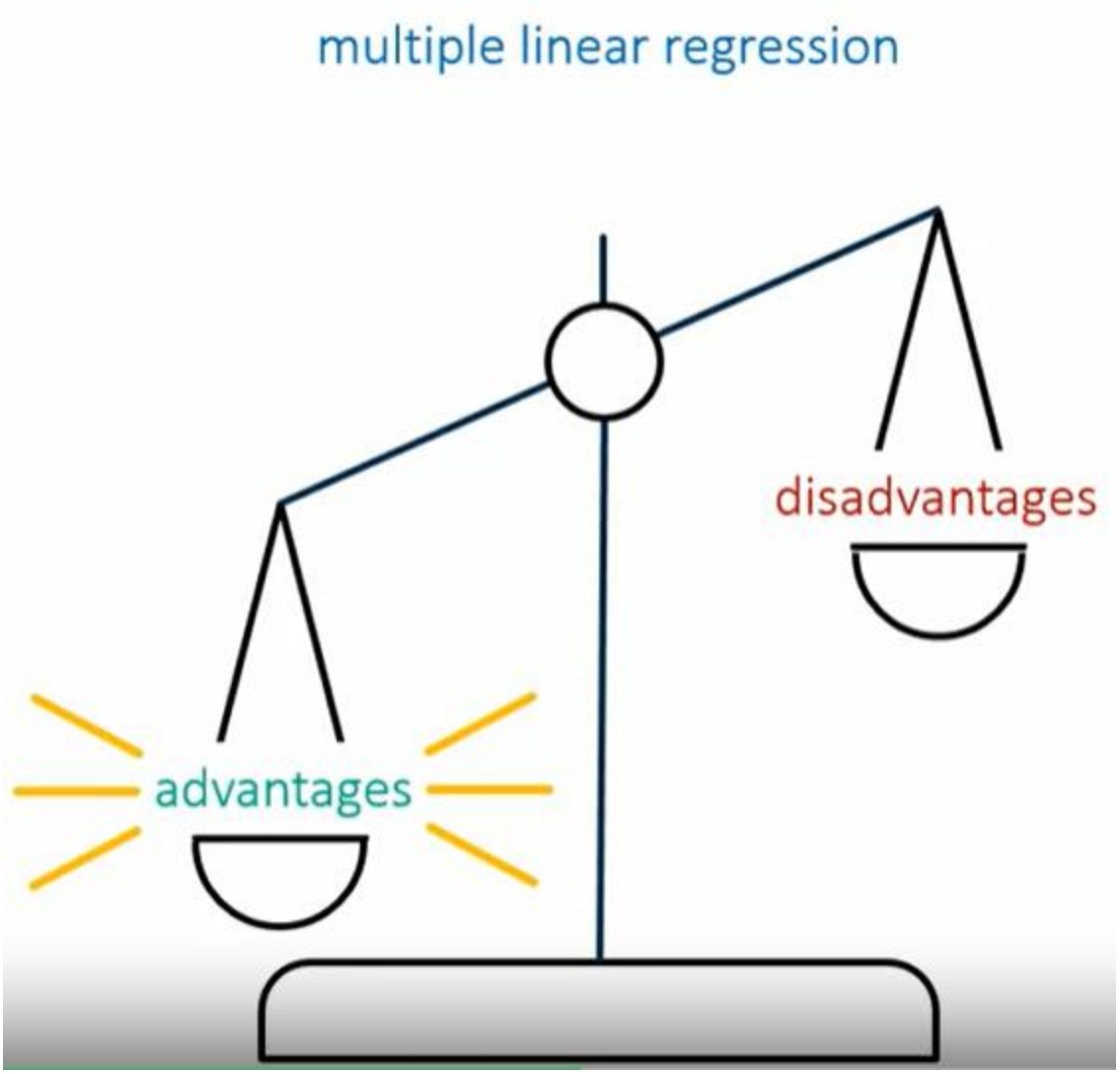



multiple linear regression
advantages
response variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable


response variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable


response variable
predictor variable


multiple linear regression
disadvantages
response variable
predictor variable
127 possible models!
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable


multiple linear regression
disadvantages
advantages

multiple linear regression

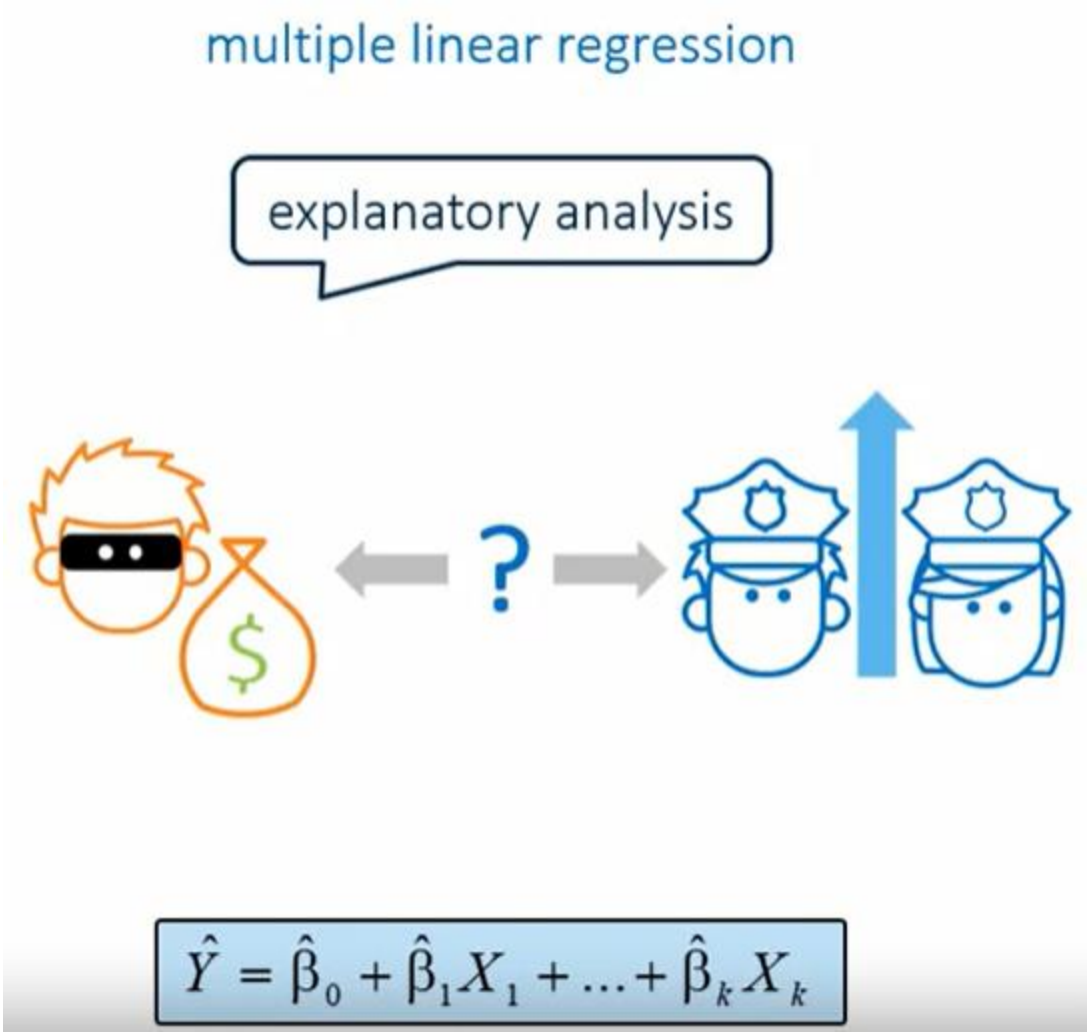
explanatory analysis

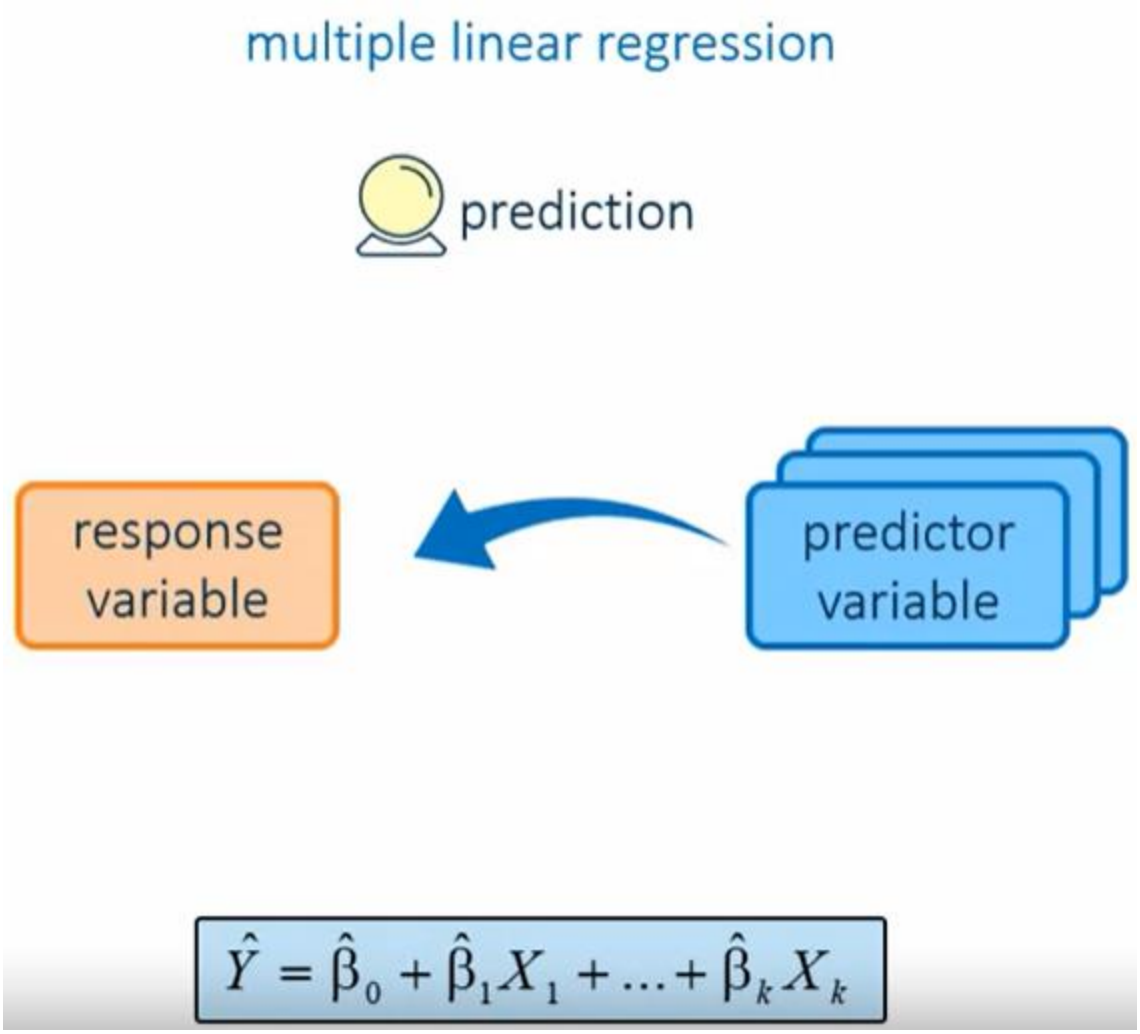
prediction

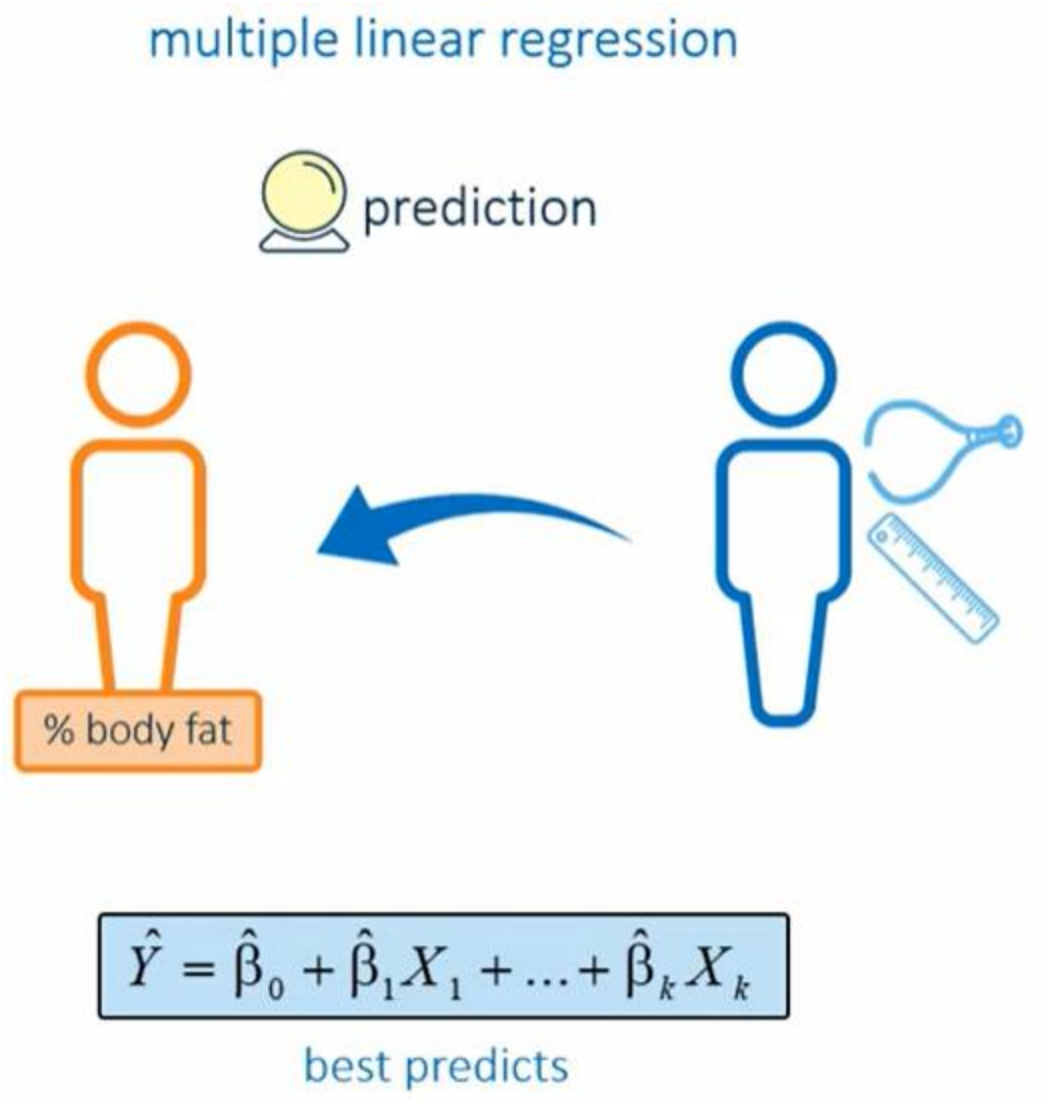
multiple linear regression

explanatory analysis

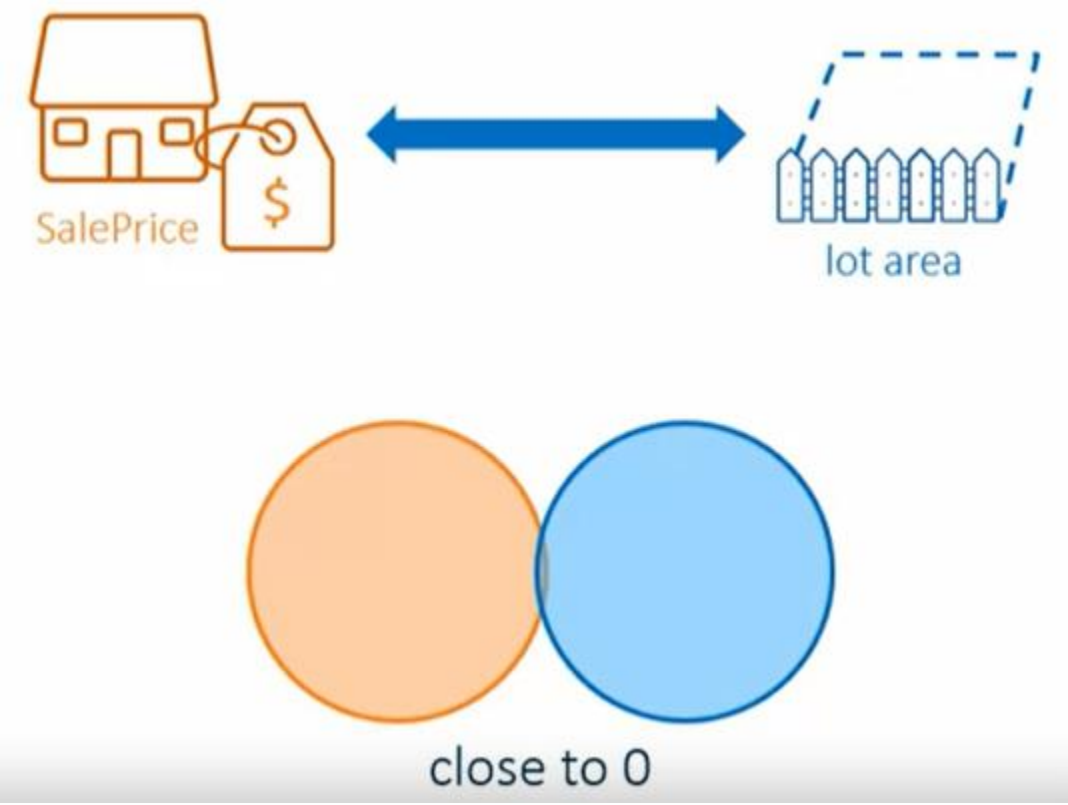
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k$$

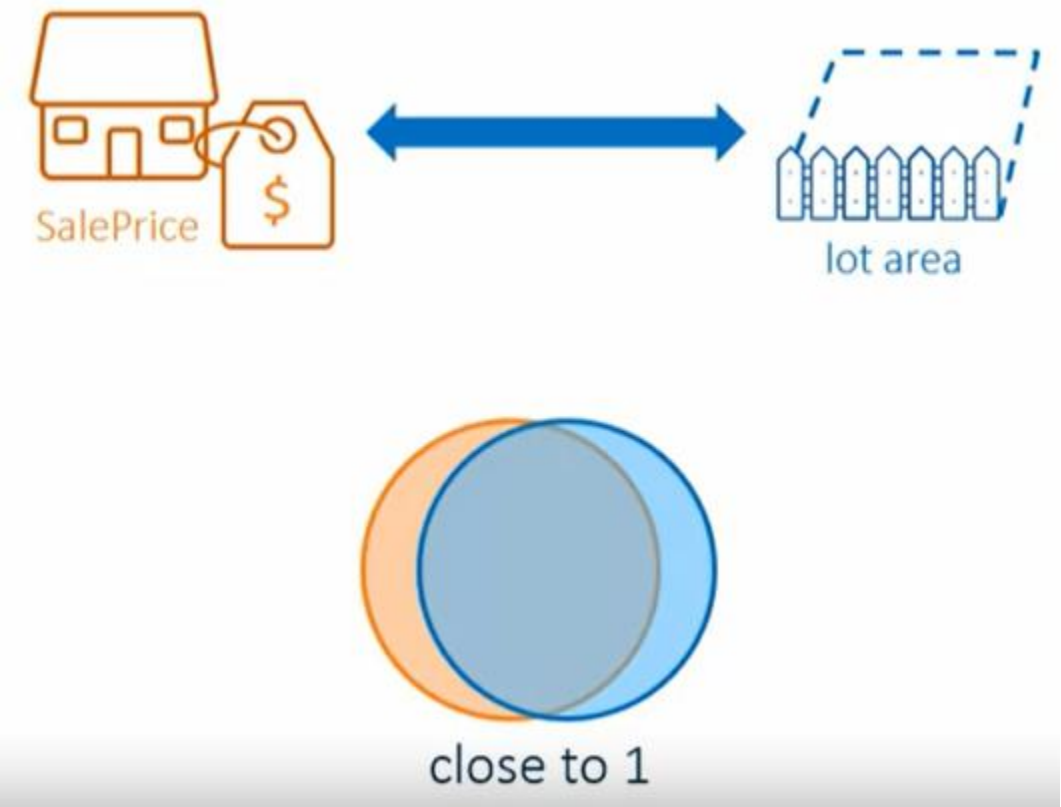

multiple linear regression
explanatory analysis
?
$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k$


multiple linear regression
prediction
response variable
predictor variable
$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k$

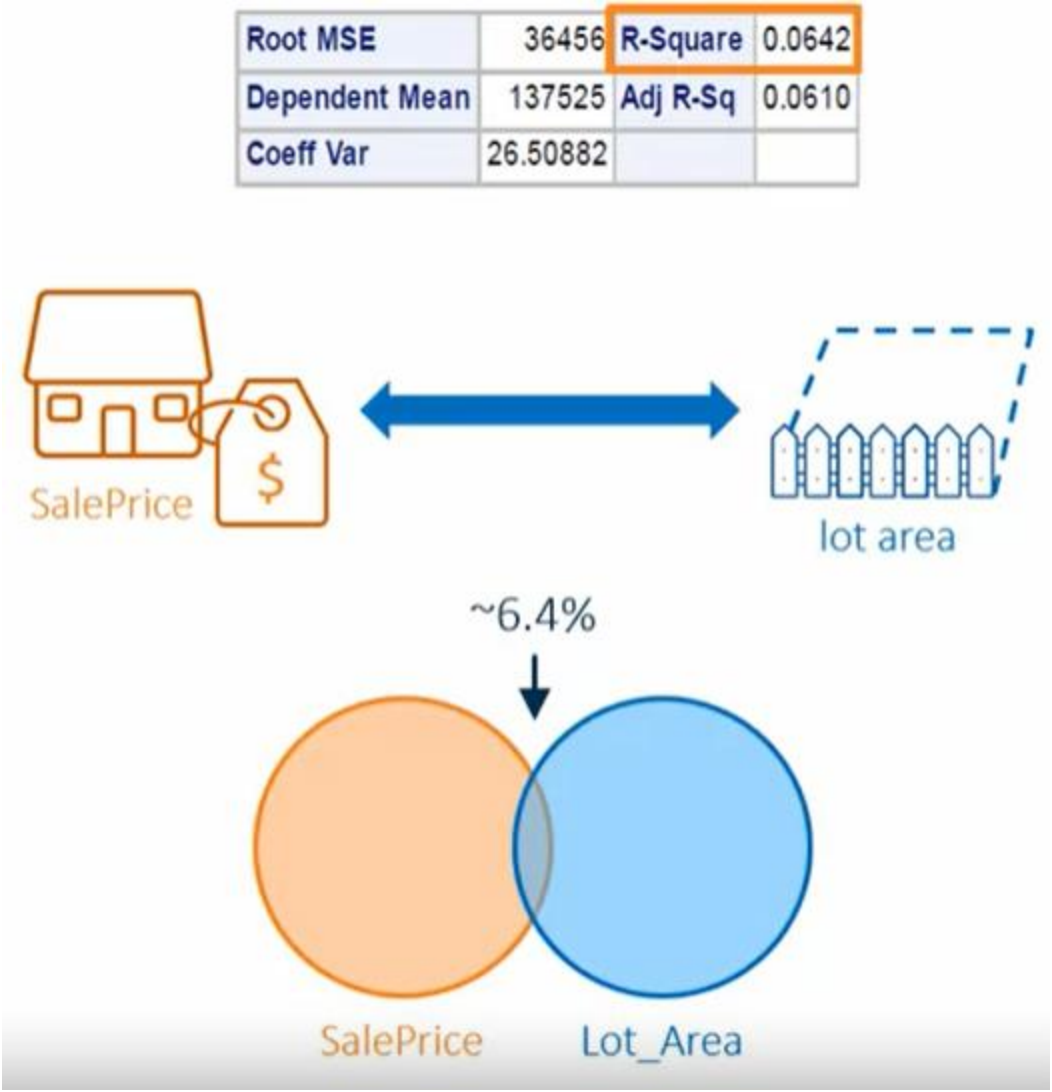

multiple linear regression
prediction
% body fat
$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k$
best predicts

Adjusted R-square

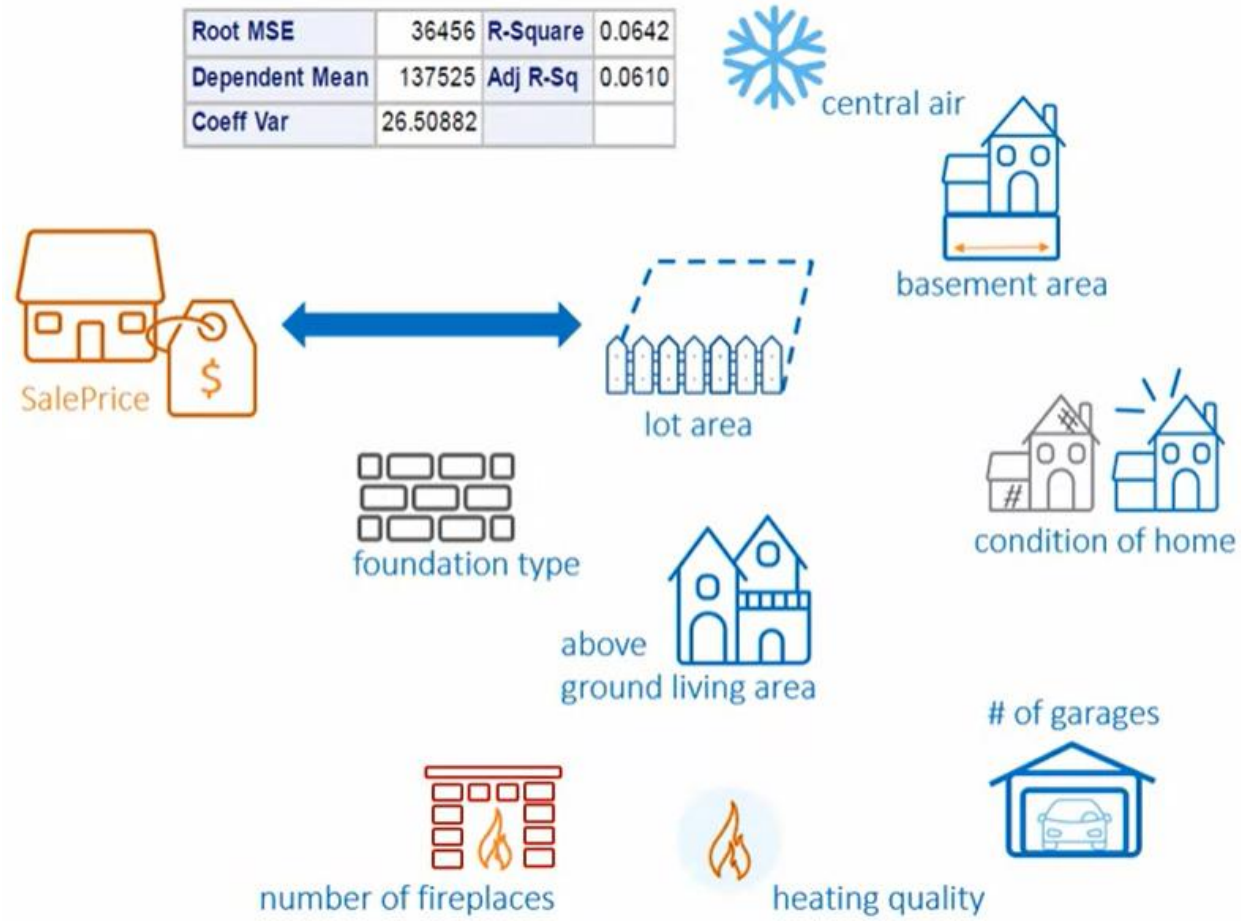




Root MSE	36456	R-Square	0.0642
Dependent Mean	137525	Adj R-Sq	0.0610
Coeff Var	26.50882		

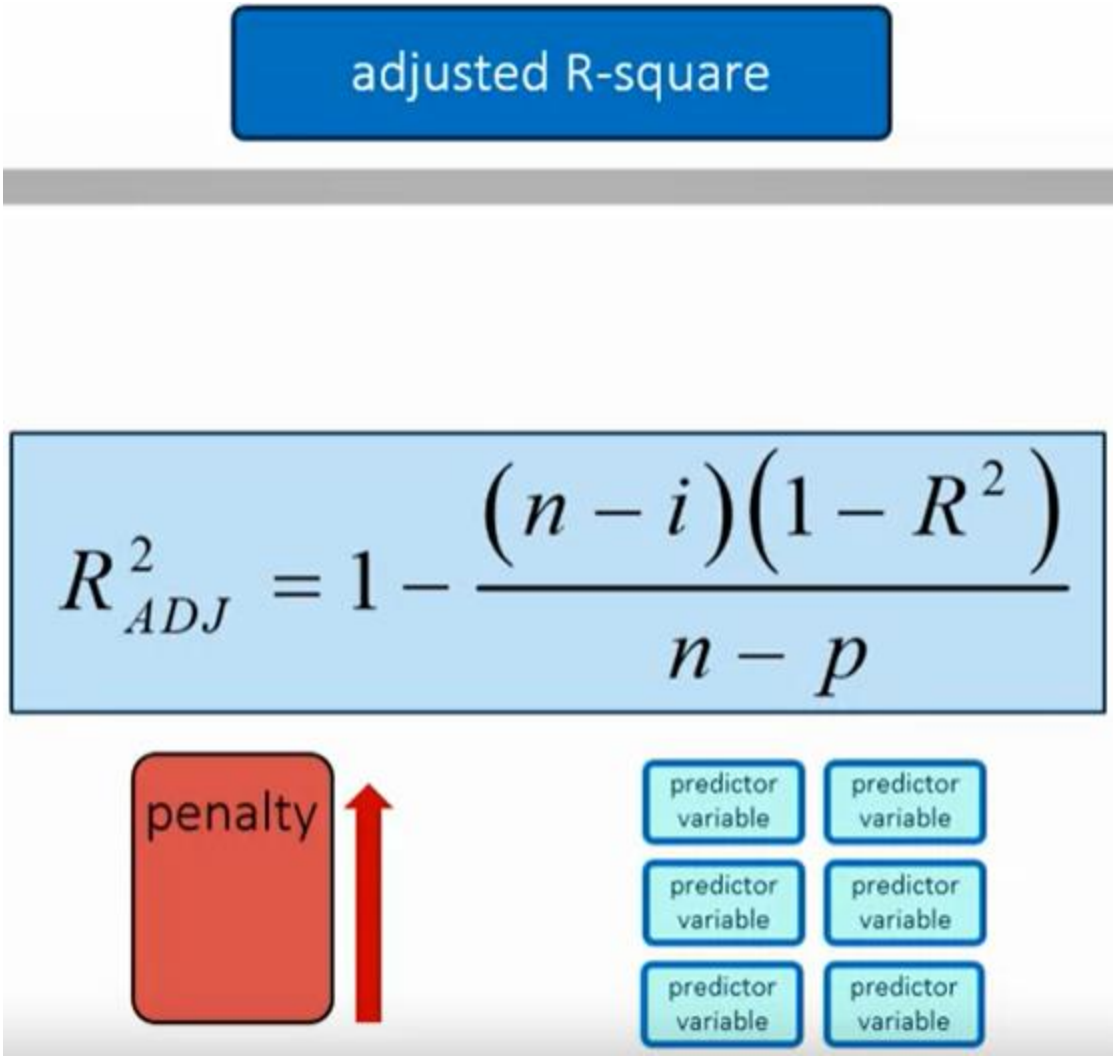


Root MSE	36456	R-Square	0.0642
Dependent Mean	137525	Adj R-Sq	0.0610
Coeff Var	26.50882		



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

considers number of terms


adjusted R-square
$R^2_{ADJ} = 1 - \frac{(n-i)(1-R^2)}{n-p}$
penalty
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable

Demo Fitting a Multiple Linear Regression Model Using PROC REG

```
/*st103d03.sas*/  /*Part A*/
ods graphics on;

proc reg data=STAT1.ameshousing3 ;
    model SalePrice=Basement_Area Lot_Area;
    title "Model with Basement Area and Lot Area";
run;
quit;
```

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

Model with Basement Area and Lot Area

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	2	2.032206E11	1.016103E11	137.17	<.0001
Error	297	2.200029E11	740750509		
Corrected Total	299	4.232235E11			

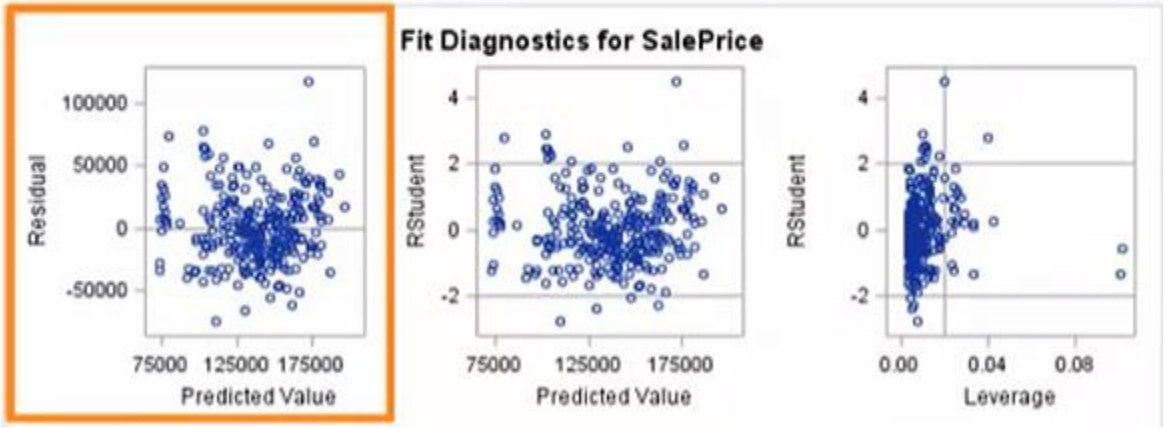
Root MSE	27217	R-Square	0.4802
Dependent Mean	137525	Adj R-Sq	0.4767
Coeff Var	19.79041		

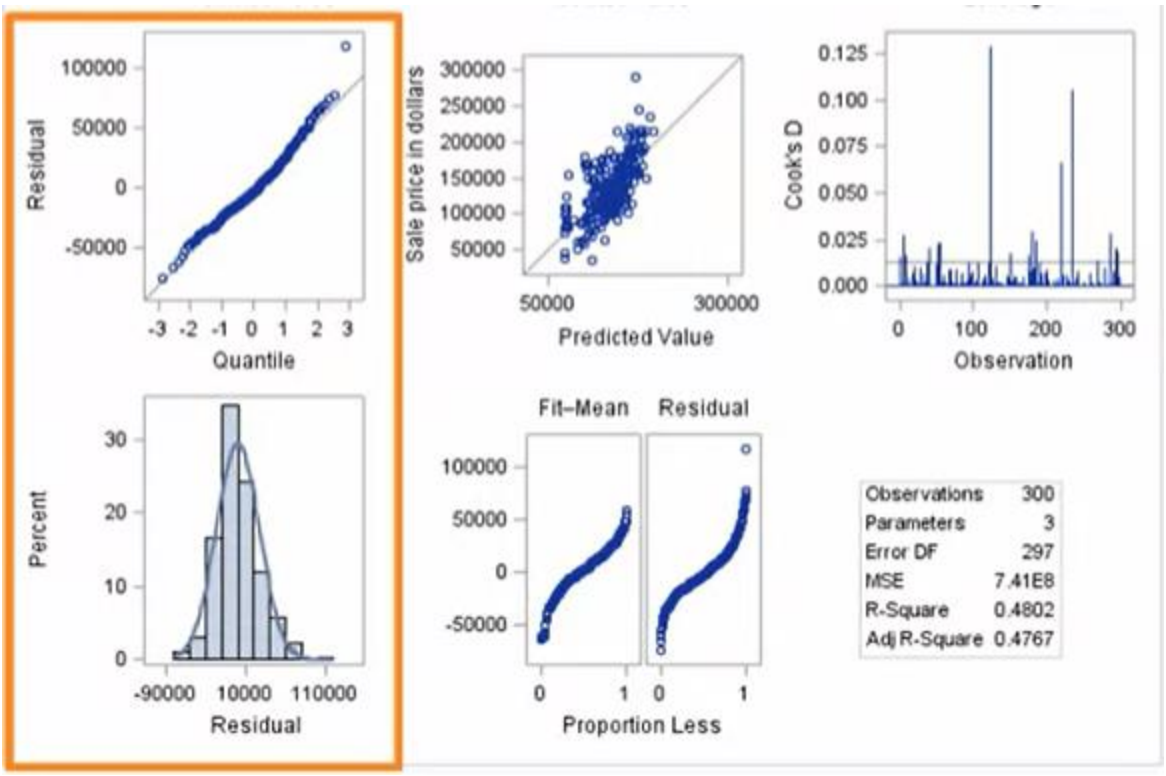
Parameter Estimates

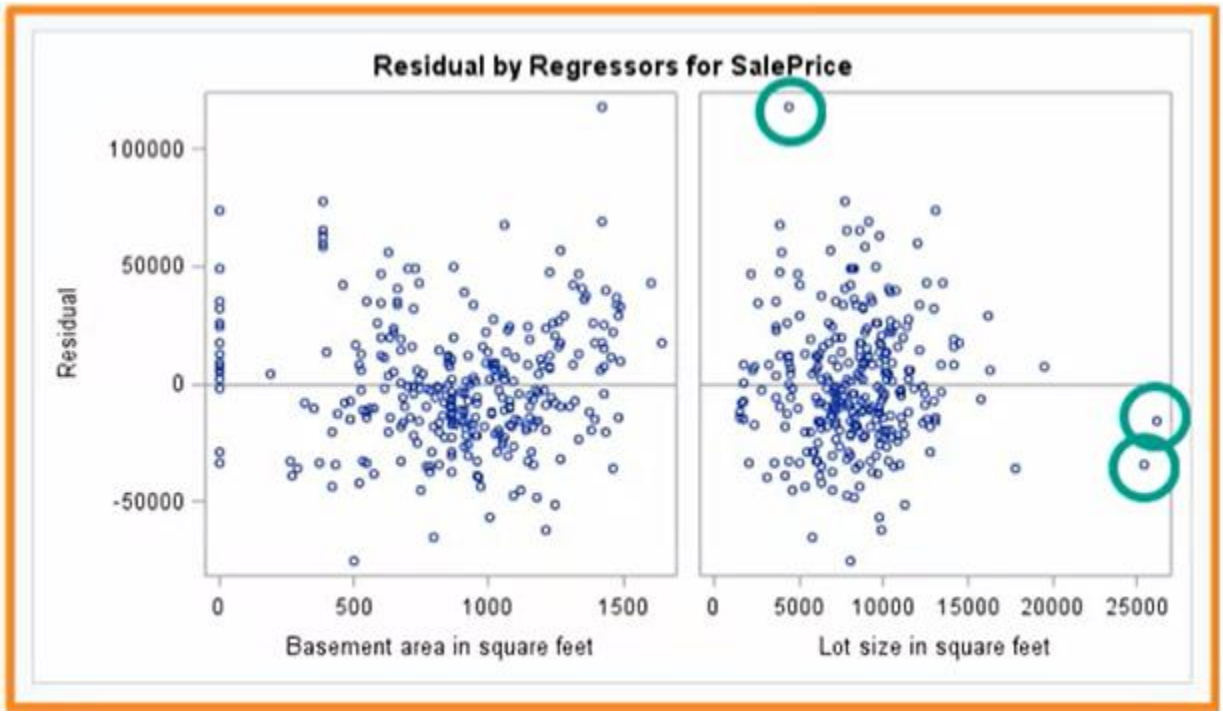
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	69016	5129.52179	13.45	<.0001
Basement_Area	Basement area in square feet	1	70.08680	4.54618	15.42	<.0001
Lot_Area	Lot size in square feet	1	0.80430	0.49210	1.63	0.1032

Model with Basement Area and Lot Area

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

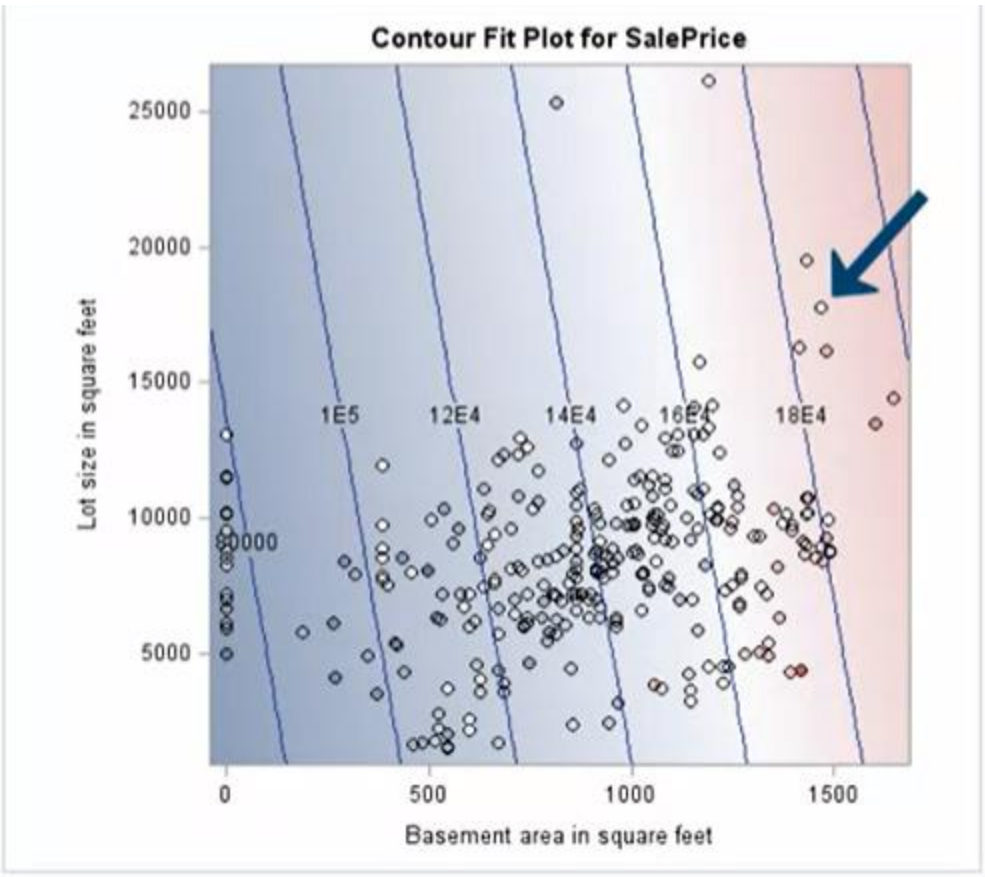






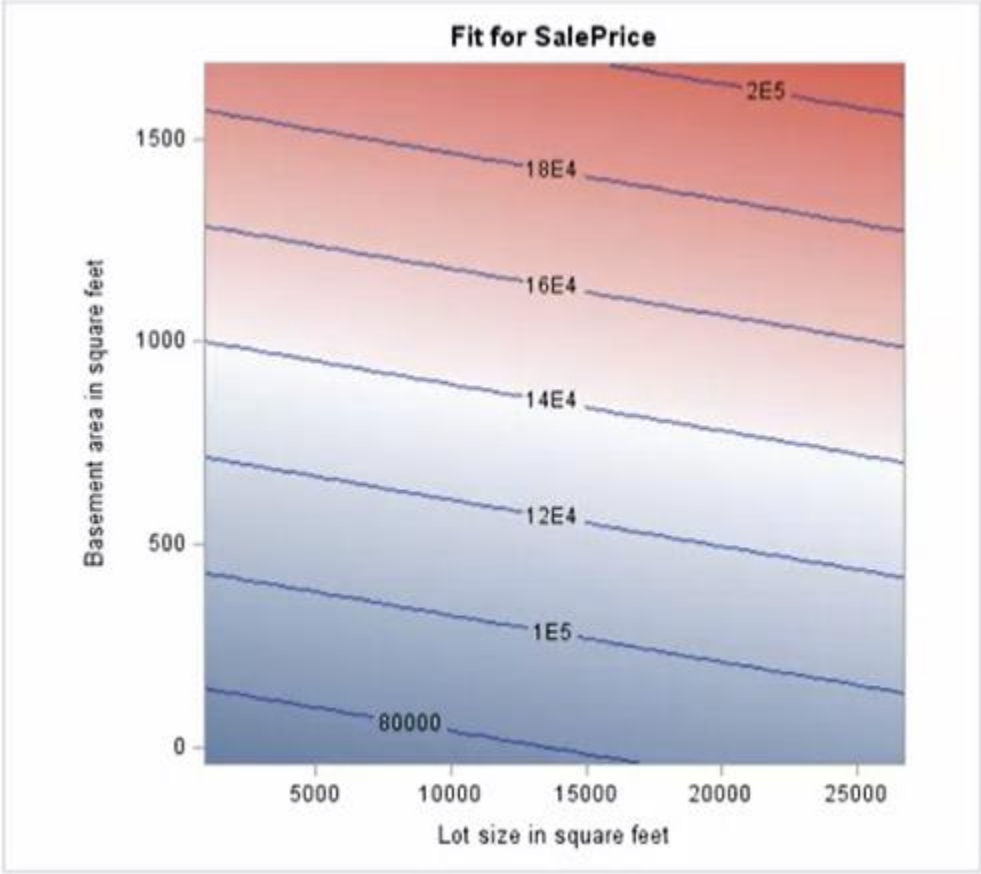

```
/*st103d03.sas*/  /*Part B*/
proc glm data=STAT1.ameshousing3
         plots(only)=(contourfit);
    model SalePrice=Basement_Area Lot_Area;
    store out=multiple;
    title "Model with Basement Area and Gross Living Area";
run;
quit;
```

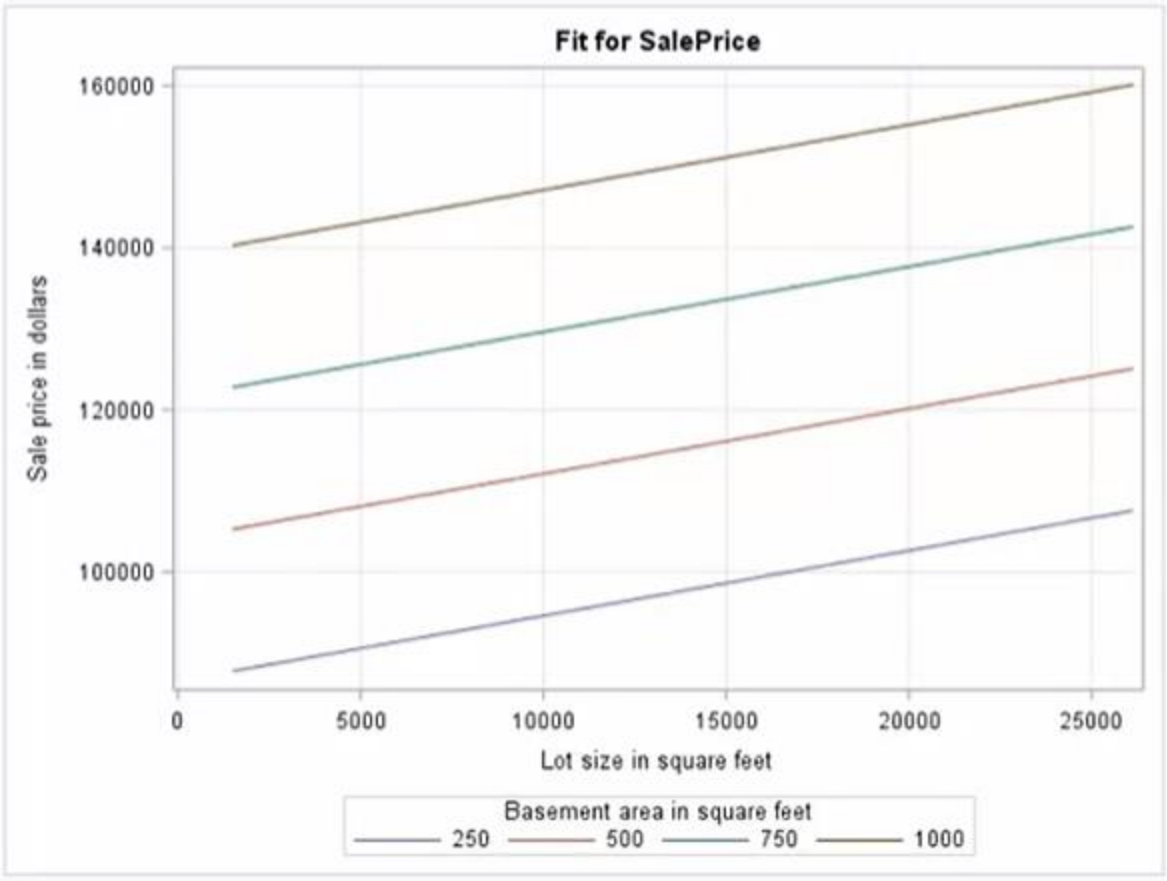
PROC GLM DATA=*SAS-data-set <options>***;**
MODEL *dependent-variables = independent-effects***;**
STORE <OUT=> *item-store-name* **< / LABEL=***'label'***>;**
RUN;


Contour Fit Plot for SalePrice
25000
20000
15000
10000
5000
Lot size in square feet
1E5
12E4
14E4
16E4
18E4
0
500
1000
1500
Basement area in square feet

```
/*st103d03.sas*/  /*Part B*/
proc glm data=STAT1.ameshousing3
         plots(only)=(contourfit);
    model SalePrice=Basement_Area Lot_Area;
    store out=multiple;
    title "Model with Basement Area and Gross Living Area";
run;
quit;

/*st103d03.sas*/  /*Part C*/
proc plm restore=multiple plots=all;
    effectplot contour (y=Basement_Area x=Lot_Area);
    effectplot slicefit(x=Lot_Area sliceby=Basement_Area=250 to 1000 by 250);
run;

title;
```

PROC PLM RESTORE=*item-store-specification <options>*;
EFFECTPLOT <plot-type <(plot-definition-options)>> < / options>;
RUN;






```
/*st103d03.sas*/  /*Part A*/
ods graphics on;

proc reg data=STAT1.ameshousing3 ;
   model SalePrice=Basement_Area Lot_Area;
   title "Model with Basement Area and Lot Area";
run;
quit;

/*st103d03.sas*/  /*Part B*/
proc glm data=STAT1.ameshousing3
        plots(only)=(contourfit);
   model SalePrice=Basement_Area Lot_Area;
   store out=multiple;
```

```
   title "Model with Basement Area and Gross Living Area";
run;
quit;

/*st103d03.sas*/  /*Part C*/
proc plm restore=multiple plots=all;
   effectplot contour (y=Basement_Area x=Lot_Area);
   effectplot slicefit(x=Lot_Area sliceby=Basement_Area=250 to 1000 by 250);
run;

title;
```

Model with Basement Area and Lot Area

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	2	2.032206E11	1.016103E11	137.17	<.0001
Error	297	2.200029E11	740750509		
Corrected Total	299	4.232235E11			

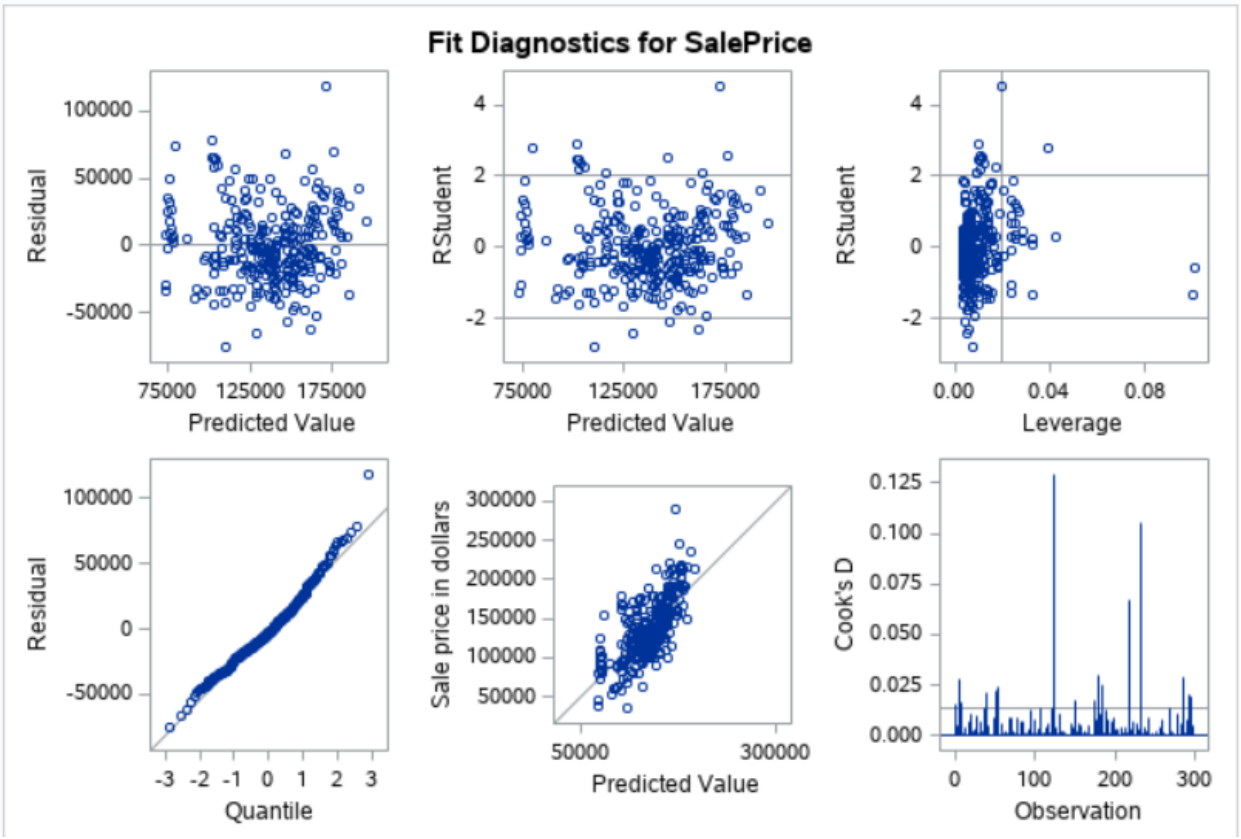
Root MSE	27217	R-Square	0.4802
Dependent Mean	137525	Adj R-Sq	0.4767
Coeff Var	19.79041		

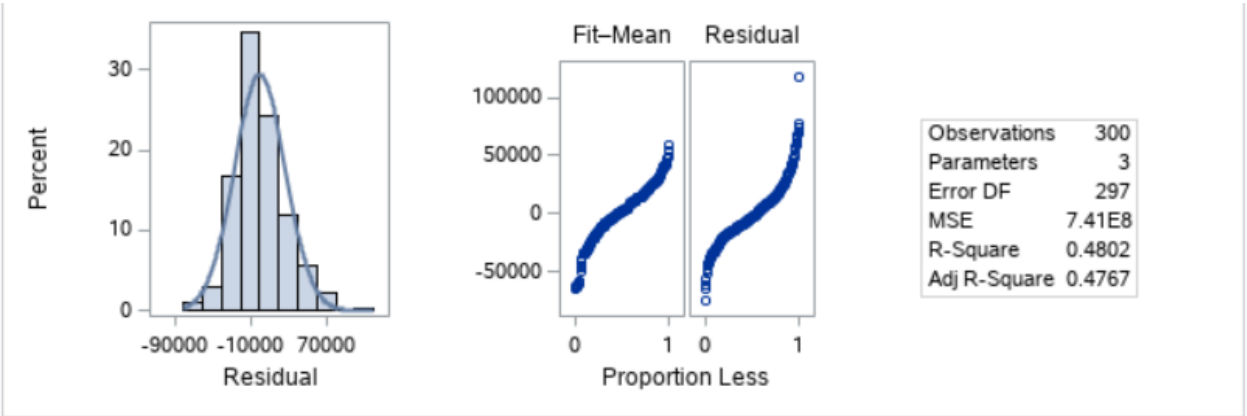
Parameter Estimates

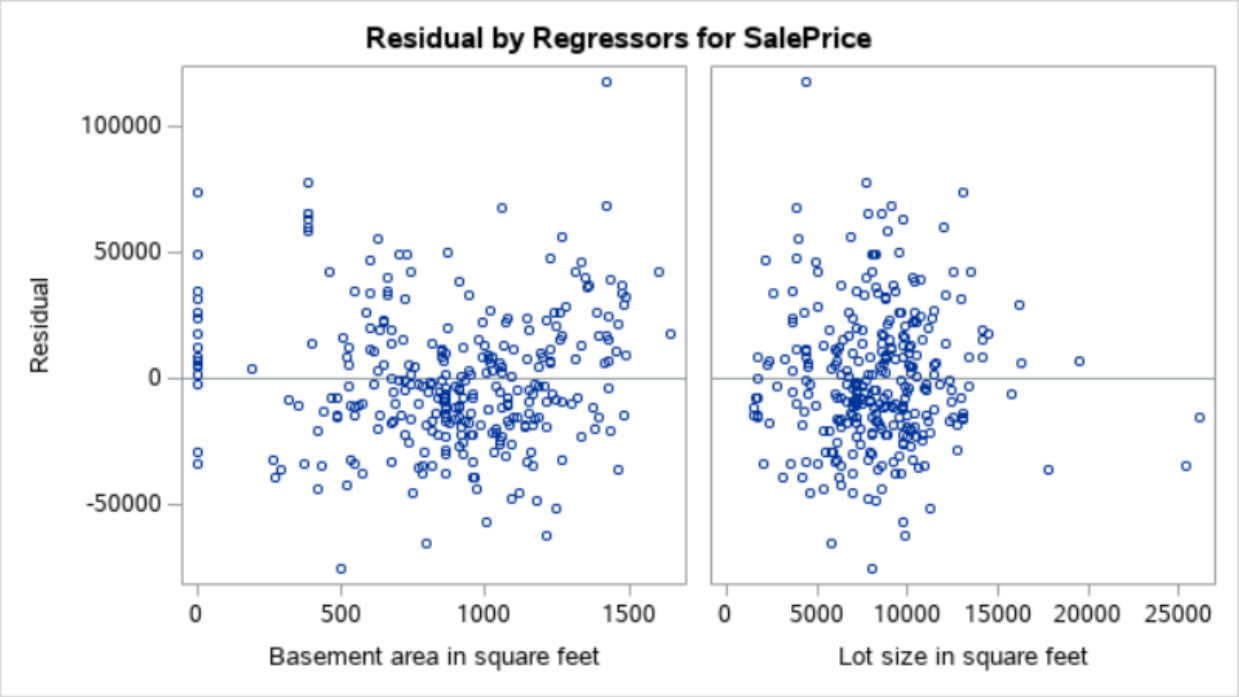
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	69016	5129.52179	13.45	<.0001
Basement_Area	Basement area in square feet	1	70.08680	4.54618	15.42	<.0001
Lot_Area	Lot size in square feet	1	0.80430	0.49210	1.63	0.1032

Model with Basement Area and Lot Area

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







Model with Basement Area and Gross Living Area

The GLM Procedure

Number of Observations Read	300
Number of Observations Used	300

Model with Basement Area and Gross Living Area

The GLM Procedure

Dependent Variable: SalePrice Sale price in dollars

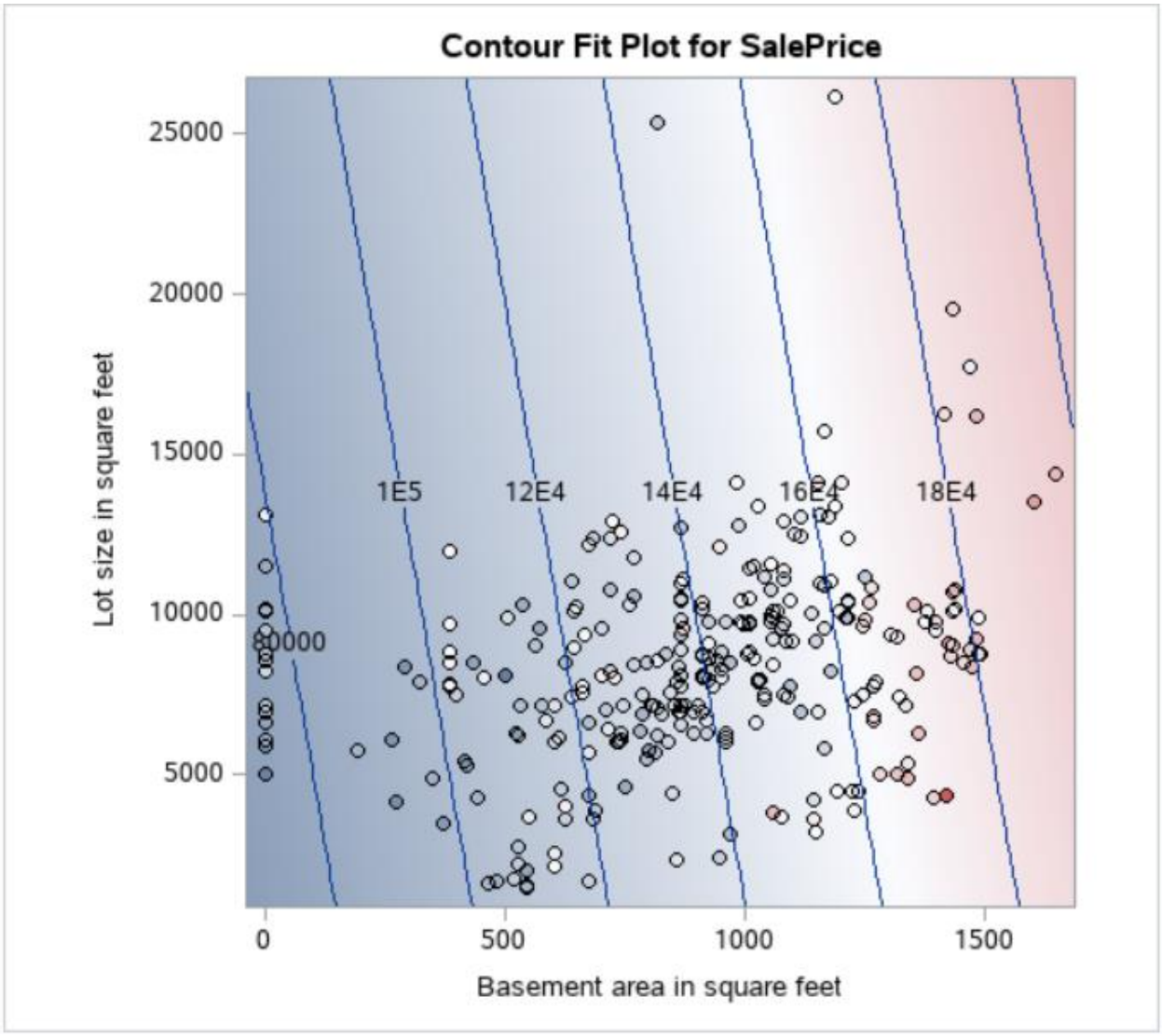
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	203220618262	101610309131	137.17	<.0001
Error	297	220002901249	740750509.26		
Corrected Total	299	423223519511			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.480173	19.79041	27216.73	137524.9

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Basement_Area	1	201241844480	201241844480	271.67	<.0001
Lot_Area	1	1978773781.7	1978773781.7	2.67	0.1032

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Basement_Area	1	176055907089	176055907089	237.67	<.0001
Lot_Area	1	1978773781.7	1978773781.7	2.67	0.1032

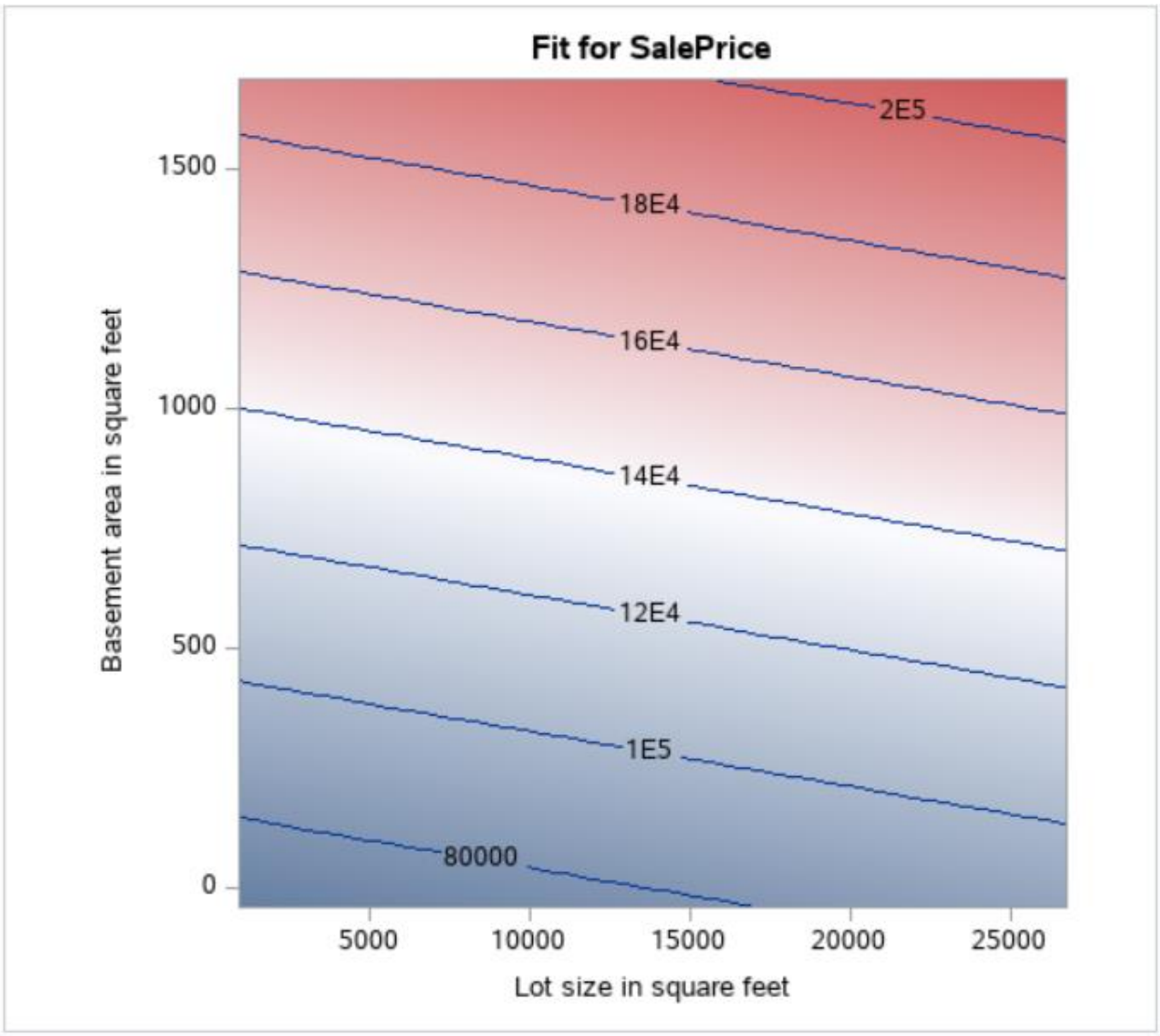
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	69015.61360	5129.521790	13.45	<.0001
Basement_Area	70.08680	4.546183	15.42	<.0001
Lot_Area	0.80430	0.492102	1.63	0.1032

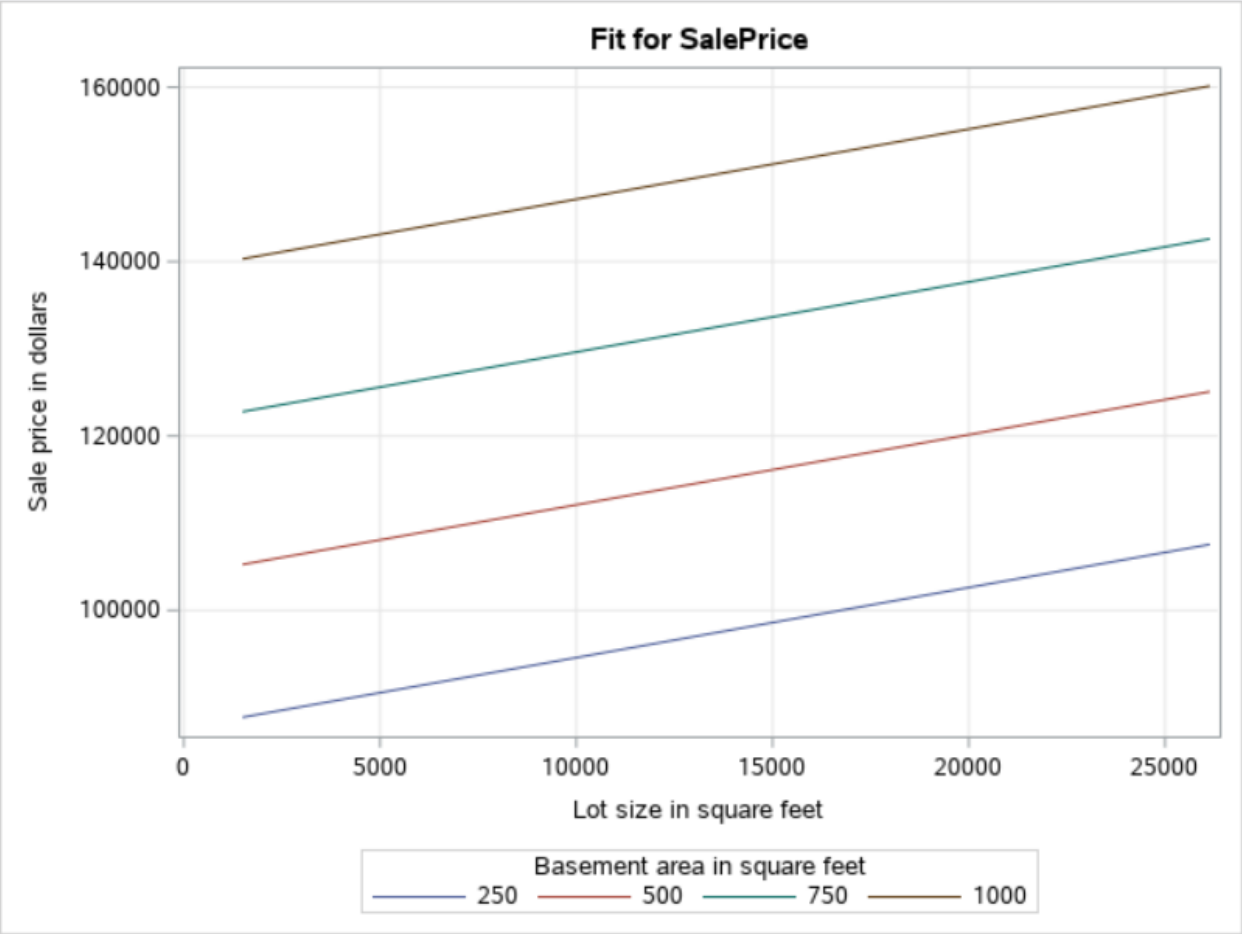


Model with Basement Area and Gross Living Area

The PLM Procedure

Store Information	
Item Store	WORK.MULTIPLE
Data Set Created From	STAT1.AMESHOUSING3
Created By	PROC GLM
Date Created	23AUG21:05:24:56
Response Variable	SalePrice
Model Effects	Intercept Basement_Area Lot_Area


Fit for SalePrice
2E5
18E4
16E4
14E4
12E4
1E5
80000
1500
1000
500
0
Basement area in square feet
5000
10000
15000
20000
25000
Lot size in square feet



Which statistic is used to test the null hypothesis that all regression slopes are zero, against the alternative hypothesis that they are not all zero?

The *F* test in the ANOVA table tests the global hypothesis for the model. *F* tests in the Type I and Type III tables, as well as the *t* tests in the parameter estimates table only test individual effects. The R-square and Adjusted R-square are measures of model fit.

```
/*st103s02.sas*/  /*Part A*/
ods graphics off;
proc reg data=STAT1.BodyFat2;
   model PctBodyFat2=Age Weight Height
         Neck Chest Abdomen Hip Thigh
         Knee Ankle Biceps Forearm Wrist;
   title 'Regression of PctBodyFat2 on All '
         'Predictors';
```

run;

quit;

Regression of PctBodyFat2 on All Predictors

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	13	13159	1012.22506	54.50	<.0001
Error	238	4420.06401	18.57170		
Corrected Total	251	17579			

Root MSE	4.30949	R-Square	0.7486
Dependent Mean	19.15079	Adj R-Sq	0.7348
Coeff Var	22.50293		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-21.35323	22.18616	-0.96	0.3368
Age	1	0.06457	0.03219	2.01	0.0460
Weight	1	-0.09638	0.06185	-1.56	0.1205
Height	1	-0.04394	0.17870	-0.25	0.8060
Neck	1	-0.47547	0.23557	-2.02	0.0447
Chest	1	-0.01718	0.10322	-0.17	0.8679
Abdomen	1	0.95500	0.09016	10.59	<.0001
Hip	1	-0.18859	0.14479	-1.30	0.1940
Thigh	1	0.24835	0.14617	1.70	0.0906
Knee	1	0.01395	0.24775	0.06	0.9552
Ankle	1	0.17788	0.22262	0.80	0.4251
Biceps	1	0.18230	0.17250	1.06	0.2917
Forearm	1	0.45574	0.19930	2.29	0.0231
Wrist	1	-1.65450	0.53316	-3.10	0.0021

```
/*st103s02.sas*/  /*Part B*/
proc reg data=STAT1.BodyFat2;
   model PctBodyFat2=Age Weight Height
         Neck Chest Abdomen Hip Thigh
            Ankle Biceps Forearm Wrist;
   title 'Regression of PctBodyFat2 on All '
         'Predictors, Minus Knee';
run;
quit;
```

Regression of PctBodyFat2 on All Predictors, Minus Knee

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	12	13159	1096.57225	59.29	<.0001
Error	239	4420.12286	18.49424		
Corrected Total	251	17579			

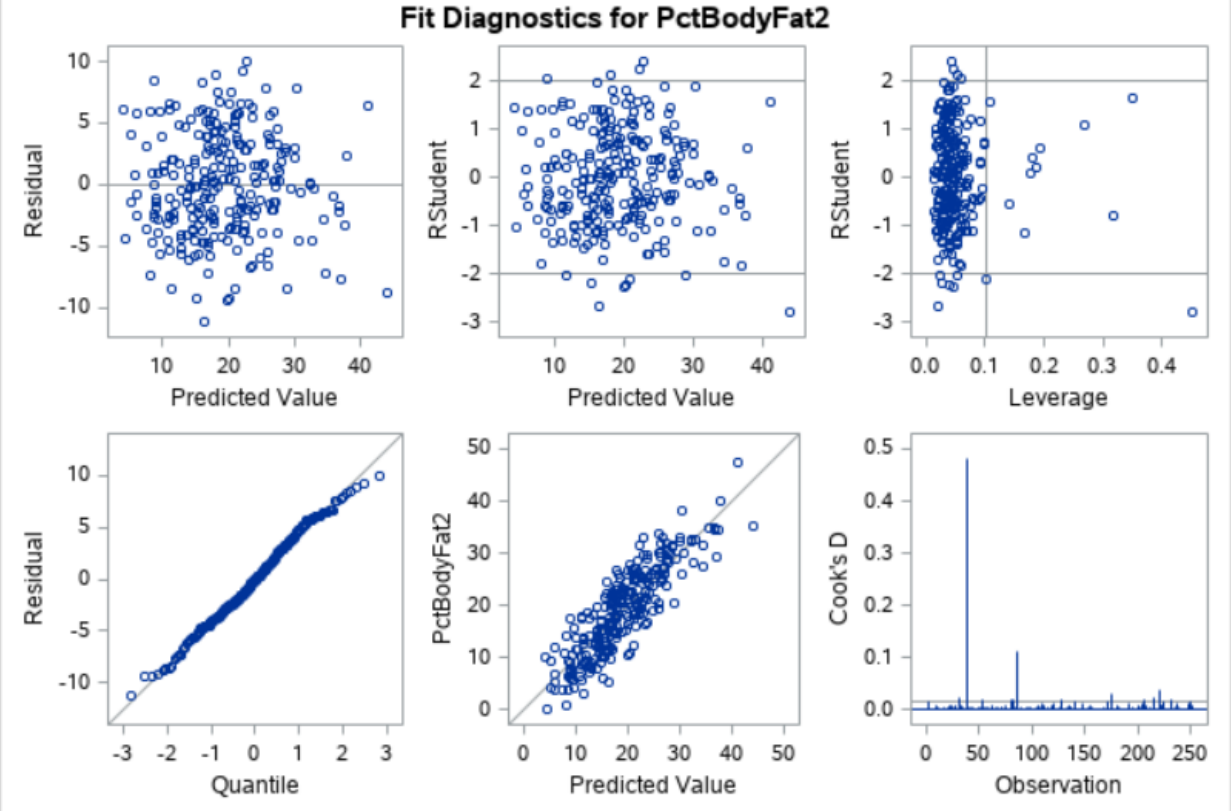
Root MSE	4.30049	R-Square	0.7486
Dependent Mean	19.15079	Adj R-Sq	0.7359
Coeff Var	22.45595		

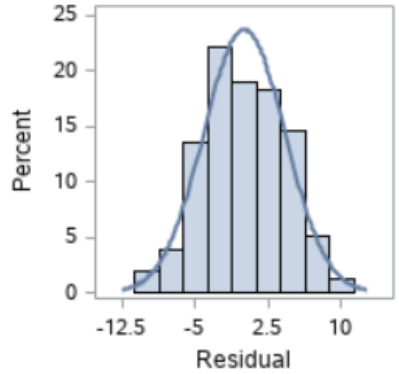
Parameter Estimates

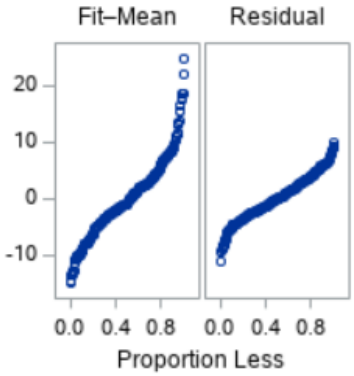
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-21.30204	22.12123	-0.96	0.3365
Age	1	0.06503	0.03108	2.09	0.0374
Weight	1	-0.09602	0.06138	-1.56	0.1191
Height	1	-0.04166	0.17369	-0.24	0.8107
Neck	1	-0.47695	0.23361	-2.04	0.0423
Chest	1	-0.01732	0.10298	-0.17	0.8666
Abdomen	1	0.95497	0.08998	10.61	<.0001
Hip	1	-0.18801	0.14413	-1.30	0.1933
Thigh	1	0.25089	0.13876	1.81	0.0719
Ankle	1	0.18018	0.21841	0.82	0.4102
Biceps	1	0.18182	0.17193	1.06	0.2913
Forearm	1	0.45667	0.19820	2.30	0.0221
Wrist	1	-1.65227	0.53057	-3.11	0.0021

Regression of PctBodyFat2 on All Predictors, Minus Knee

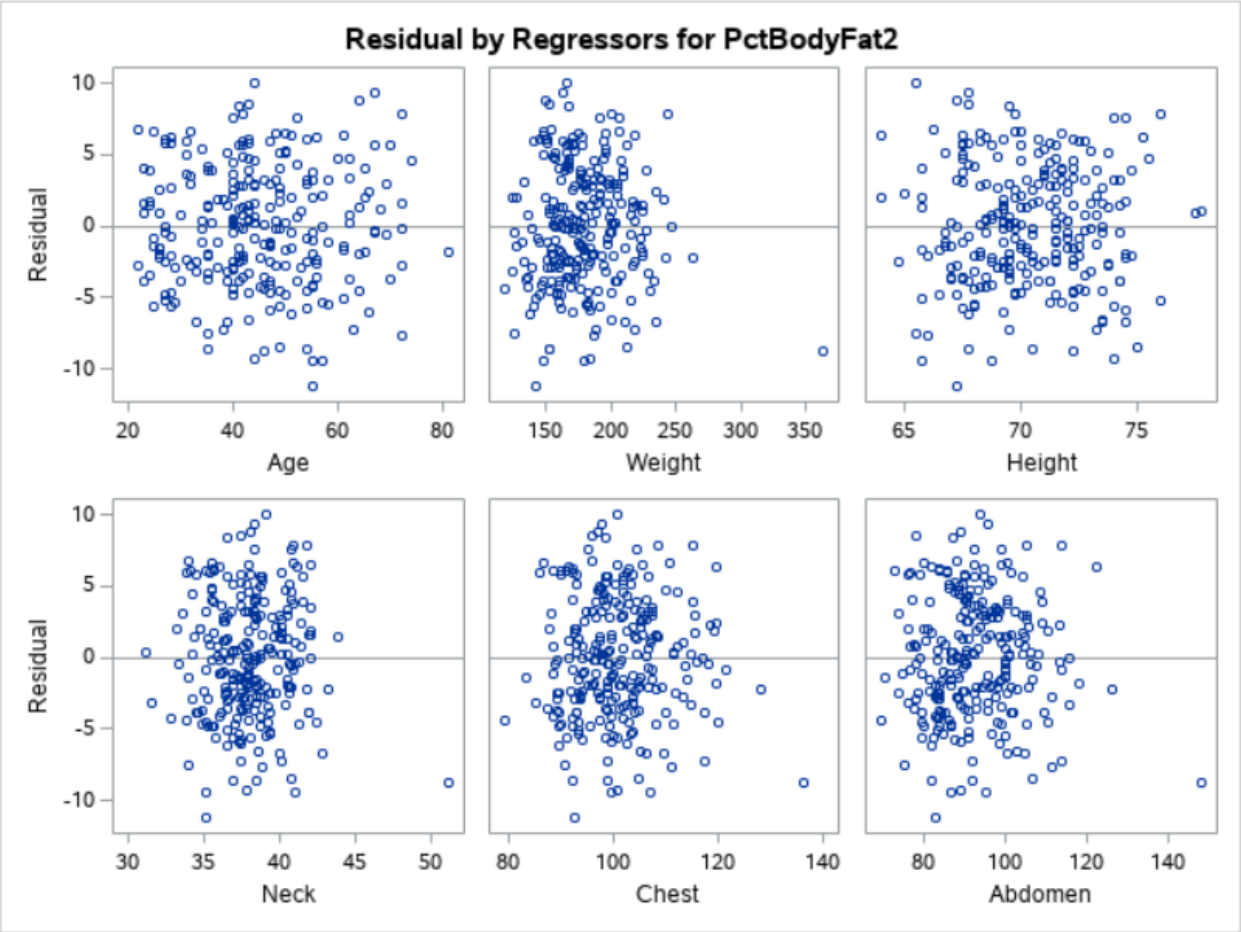
The REG Procedure
Model: MODEL1
Dependent Variable: PctBodyFat2

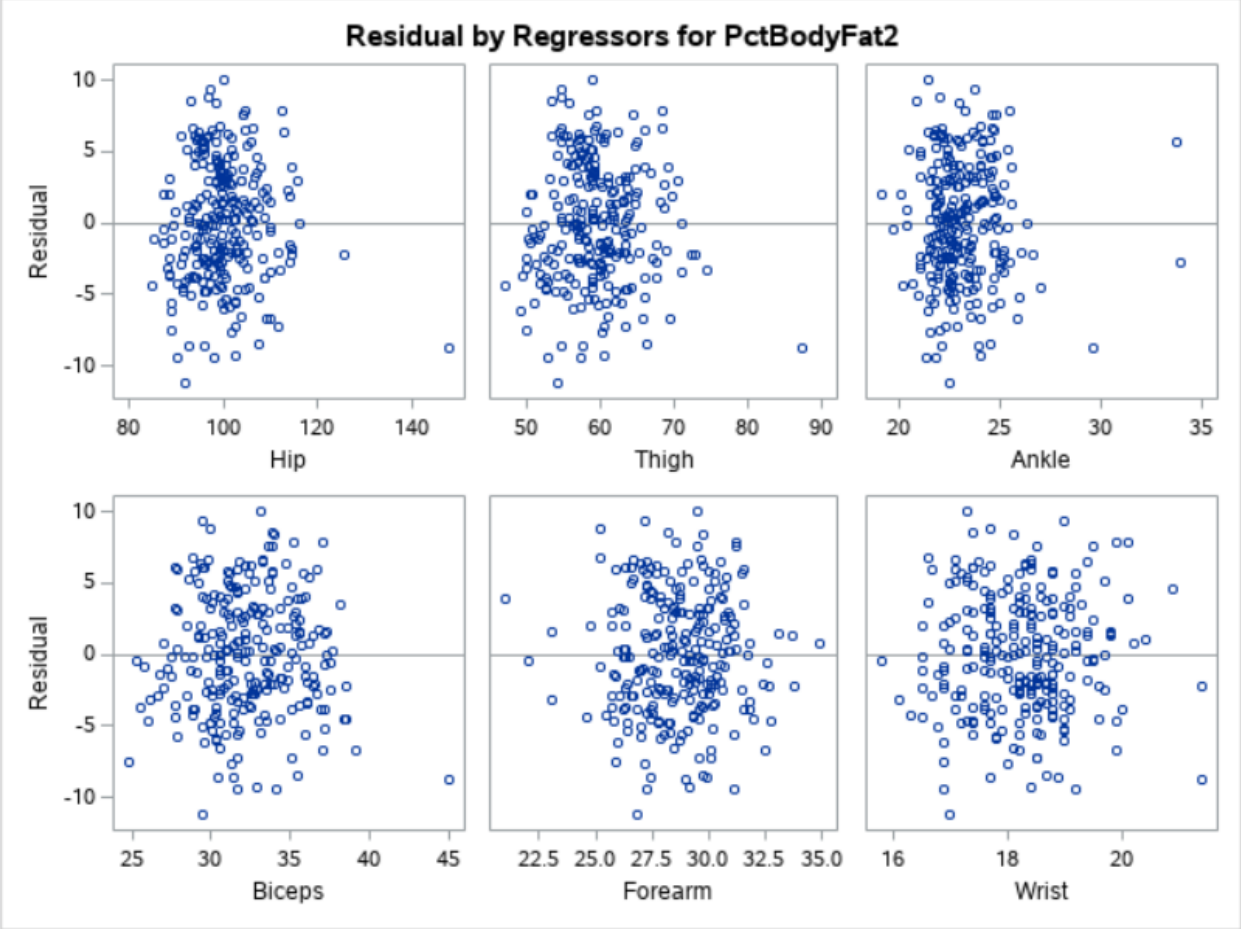






Observations	252
Parameters	13
Error DF	239
MSE	18.494
R-Square	0.7486
Adj R-Square	0.7359


Residual by Regressors for PctBodyFat2
Residual
10
5
0
-5
-10
20
40
60
80
Age
150
200
250
300
350
Weight
65
70
75
Height
30
35
40
45
50
Neck
80
100
120
140
Chest
Abdomen




```
/*st103s02.sas*/  /*Part C*/
proc reg data=STAT1.BodyFat2;
   model PctBodyFat2=Age Weight Height
         Neck      Abdomen Hip Thigh
            Ankle Biceps Forearm Wrist;
   title 'Regression of PctBodyFat2 on All '
         'Predictors, Minus Knee, Chest';
run;
quit;
```

Regression of PctBodyFat2 on All Predictors, Minus Knee, Chest

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	11	13158	1196.21310	64.94	<.0001
Error	240	4420.64572	18.41936		
Corrected Total	251	17579			

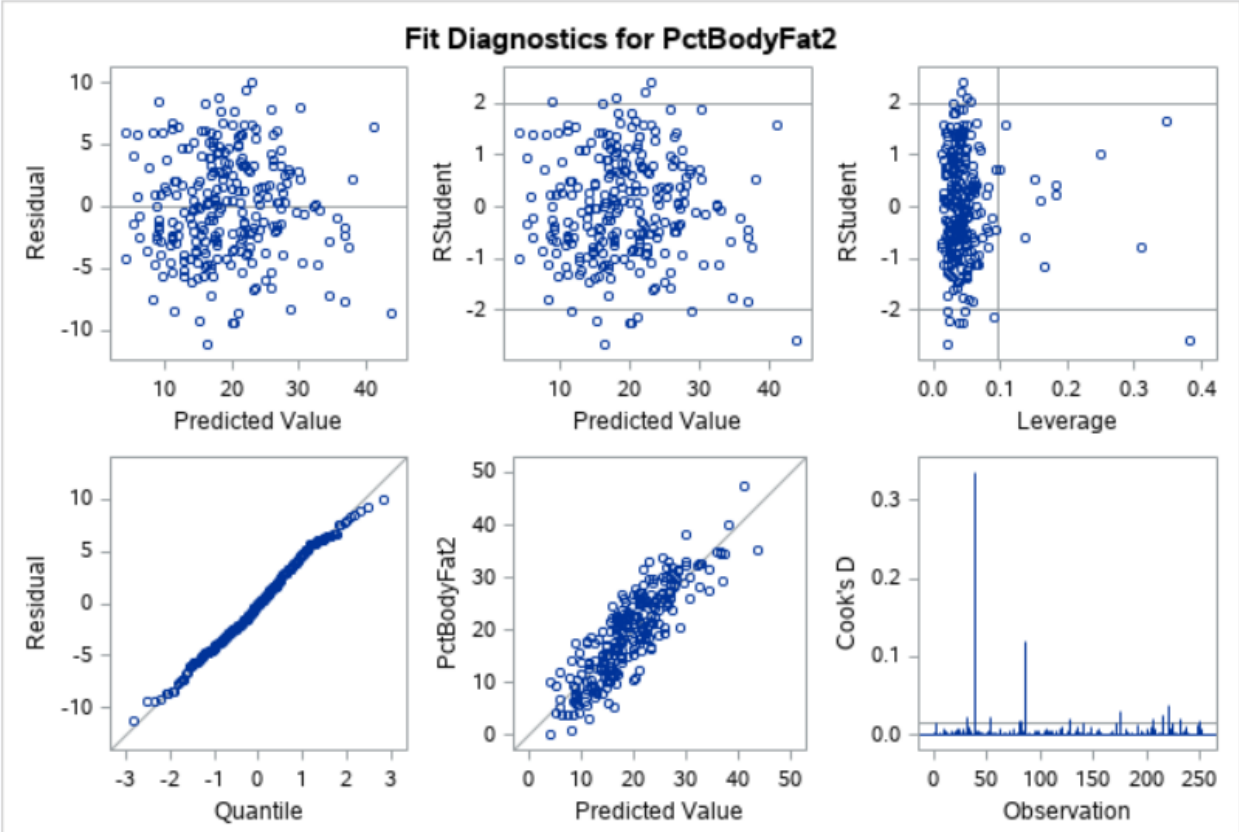
Root MSE	4.29178	R-Square	0.7485
Dependent Mean	19.15079	Adj R-Sq	0.7370
Coeff Var	22.41044		

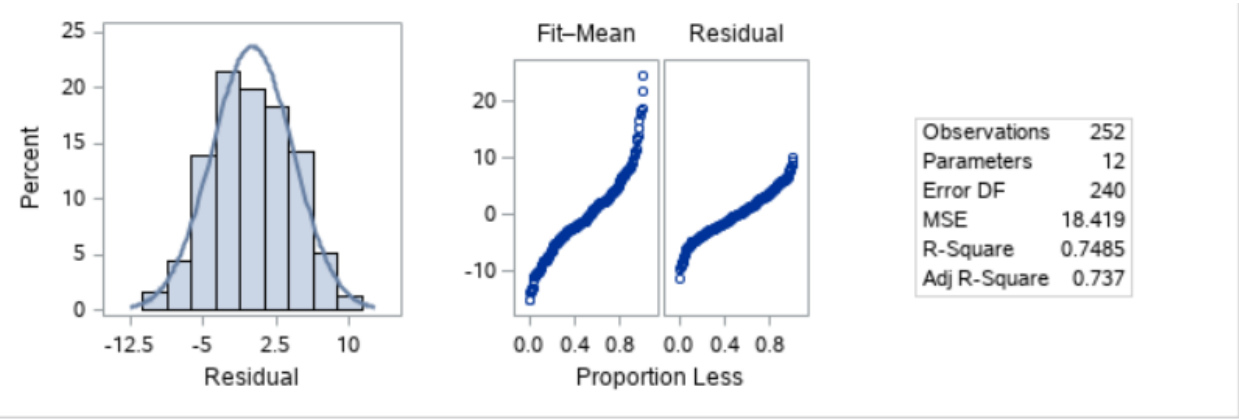
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-23.13736	19.20171	-1.20	0.2294
Age	1	0.06488	0.03100	2.09	0.0374
Weight	1	-0.10095	0.05380	-1.88	0.0618
Height	1	-0.03120	0.16185	-0.19	0.8473
Neck	1	-0.47631	0.23311	-2.04	0.0421
Abdomen	1	0.94965	0.08406	11.30	<.0001
Hip	1	-0.18316	0.14092	-1.30	0.1950
Thigh	1	0.25583	0.13534	1.89	0.0599
Ankle	1	0.18215	0.21765	0.84	0.4035
Biceps	1	0.18055	0.17141	1.05	0.2933
Forearm	1	0.45262	0.19634	2.31	0.0220
Wrist	1	-1.64984	0.52930	-3.12	0.0020

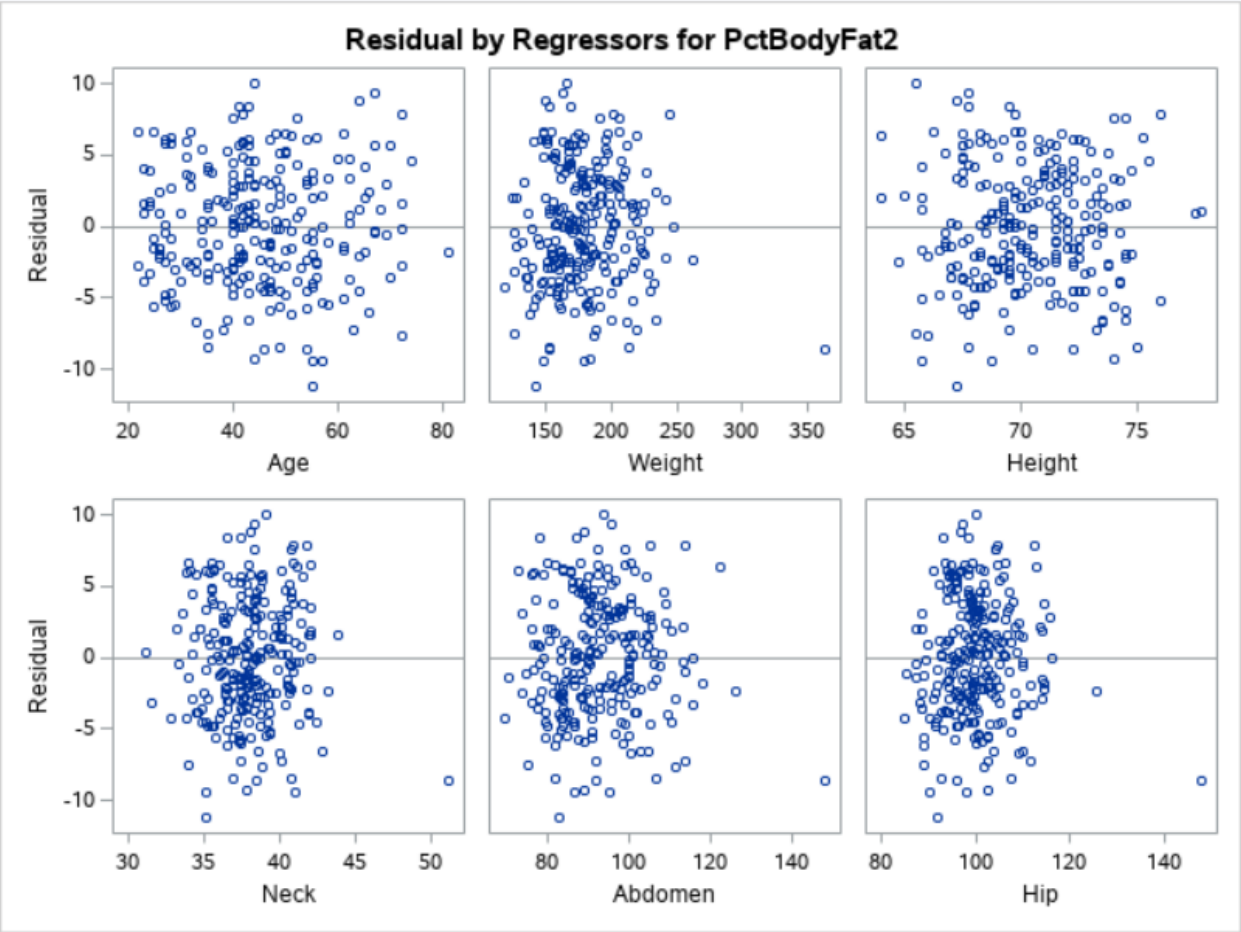
Regression of PctBodyFat2 on All Predictors, Minus Knee, Chest

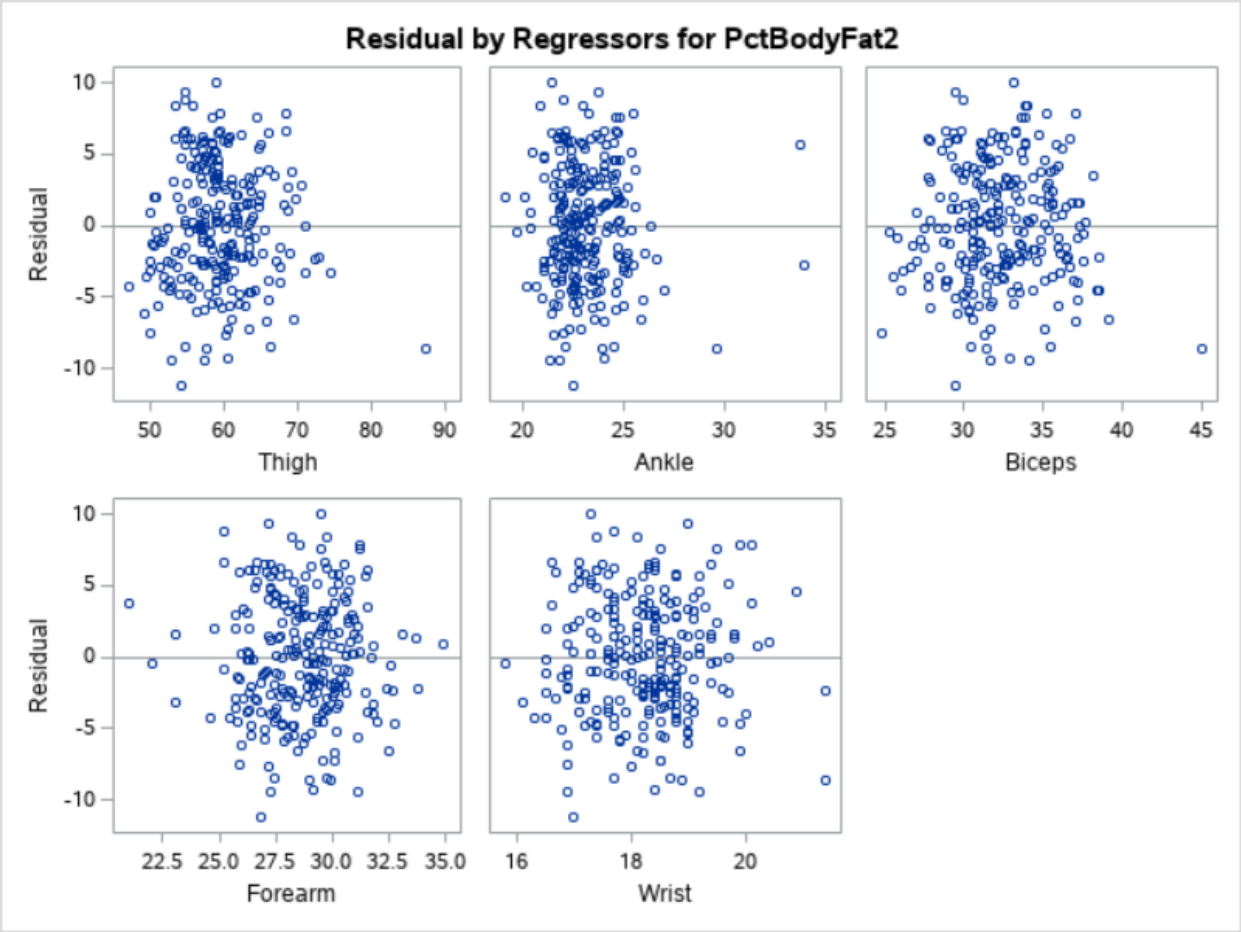
The REG Procedure
Model: MODEL1
Dependent Variable: PctBodyFat2





Observations	252
Parameters	12
Error DF	240
MSE	18.419
R-Square	0.7485
Adj R-Square	0.737


Residual by Regressors for PctBodyFat2
Residual
10
5
0
-5
-10
20
40
60
80
Age
150
200
250
300
350
Weight
65
70
75
Height
30
35
40
45
50
Neck
80
100
120
140
Abdomen
Hip



Practice - Performing Multiple Regression Using PROC REG

TOTAL POINTS 8

Question 1

Using the **stat1.bodyfat2** table, fit a multiple regression model with multiple predictors, and then modify the model by removing the least significant predictors.

Note: Turn off ODS Graphics.

1. Run a regression of **PctBodyFat2** on the variables **Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist**.
2. Compare the ANOVA table with this one from the model with only **Weight**. What is different?

F Value from ANOVA table is 54.50 versus 150.03 from the model with only Weight. Sum of Squares and Mean Square from ANOVA table for Model is almost double from the model with only Weight above. Meanwhile, the Sum of Squares and Mean Square from ANOVA table for Error is more than double from the error with only Weight.

Correct

Solution code:

```
/*st103s02.sas*/  /*Part A*/

ods graphics off;
proc reg data=STAT1.BodyFat2;
    model PctBodyFat2=Age Weight Height
          Neck Chest Abdomen Hip Thigh
          Knee Ankle Biceps Forearm Wrist;
    title 'Regression of PctBodyFat2 on All '
          'Predictors';
run;
quit;
```

There are key differences between the ANOVA table for this model and the one for the simple linear regression model. The degrees of freedom for the model are much higher, 13 versus 1. Also, the Mean Square model and the *F* ratio are much smaller.

Question 2

How do the R-Square and the adjusted R-Square compare with these statistics for the **Weight** regression?

The R-Square and the adjusted R-Square are more than double than the statistics for the Weight regression. Both R-Square are higher than their adjusted R-square, respectively.

Correct

Both the R-Square and the adjusted R-Square for the full models are larger than the simple linear regression. The multiple regression model explains almost 75% of the variation in the **PctBodtFat2** variable versus approximately 37.5% that is explained by the simple linear regression model.

Question 3

Did the estimate for the intercept change? Did the estimate for the coefficient of **Weight** change?
Yes, both estimate for the intercept and coefficient of Weight changed.

Correct

Yes, including the other variables in the model changed both the estimate of the intercept and the slope for **Weight**. Also, the *p*-values for both changed dramatically. The slope of **Weight** is now not significantly different from zero.

Question 4

To simplify the model, rerun the model from Question 1, but eliminate the variable with the highest *p*-value. Compare the output with the model from Question 1.

Did the p-value for the model change?
No, the p-value for the model did not change.

Correct

Solution code:

Knee was removed because it has the largest *p*-value (0.9552).

```
/*st103s02.sas*/  /*Part B*/

ods graphics off;
proc reg data=STAT1.BodyFat2;
    model PctBodyFat2=Age Weight Height
          Neck Chest Abdomen Hip Thigh
          Ankle Biceps Forearm Wrist;
    title 'Regression of PctBodyFat2 on All '
          'Predictors, Minus Knee';
run;
quit;
```

The *p*-value for the model did not change to four decimal places.

Question 5
Did the R-Square and the adjusted R-Square values change?
The R-Square value did not change, but the adjusted R-Square value did change a little bit higher to 0.7359 from 0.7348

Correct

The R-Square showed essentially no change. The adjusted R-Square increased from .7348 to .7359. When an adjusted R-Square increases by removing a variable from the model, it strongly implies that the removed variable was not necessary.

Question 6
Did the parameter estimates and their *p*-values change?
Yes, the parameter estimates and their p-values changed slightly.

Correct

Some of the parameter estimates and their *p*-values changed slightly, but none to any large degree.

Question 7
To simplify the model further, rerun the model from Question 4, but eliminate the variable with the highest *p*-value. How did the output change from the previous model?
The p-value remains the same. The R-Square and adjusted R-Square changed a little bit. The new R-Square is slightly lower, but the new adjusted R-Square is slightly higher.

Correct

Solution code:

Chest was removed because it is the variable with the highest *p*-value in the previous model.

```
/*st103s02.sas*/  /*Part C*/

ods graphics off;
proc reg data=STAT1.BodyFat2;
    model PctBodyFat2=Age Weight Height
          Neck Abdomen Hip Thigh
          Ankle Biceps Forearm Wrist;
```

```
    title 'Regression of PctBodyFat2 on All '
          'Predictors, Minus Knee, Chest';
run;
quit;
```

The ANOVA table did not change significantly. The R-Square remained essentially unchanged. The adjusted R-Square increased again. This confirms that the variable **Chest** did not contribute to explaining the variation in **PctBodyFat2** when the other variables were in the model.

Question 8

Did the number of parameters with *p*-values less than 0.05 change?

Yes, some number of parameters with p-values less than 0.05 changed a little bit.

Correct

The *p*-value for **Weight** changed more than any other and is now slightly more than 0.05. The *p*-values and parameter estimates for other variables changed much less. There are no more variables in this model with *p*-values below 0.05, compared with the previous one.