

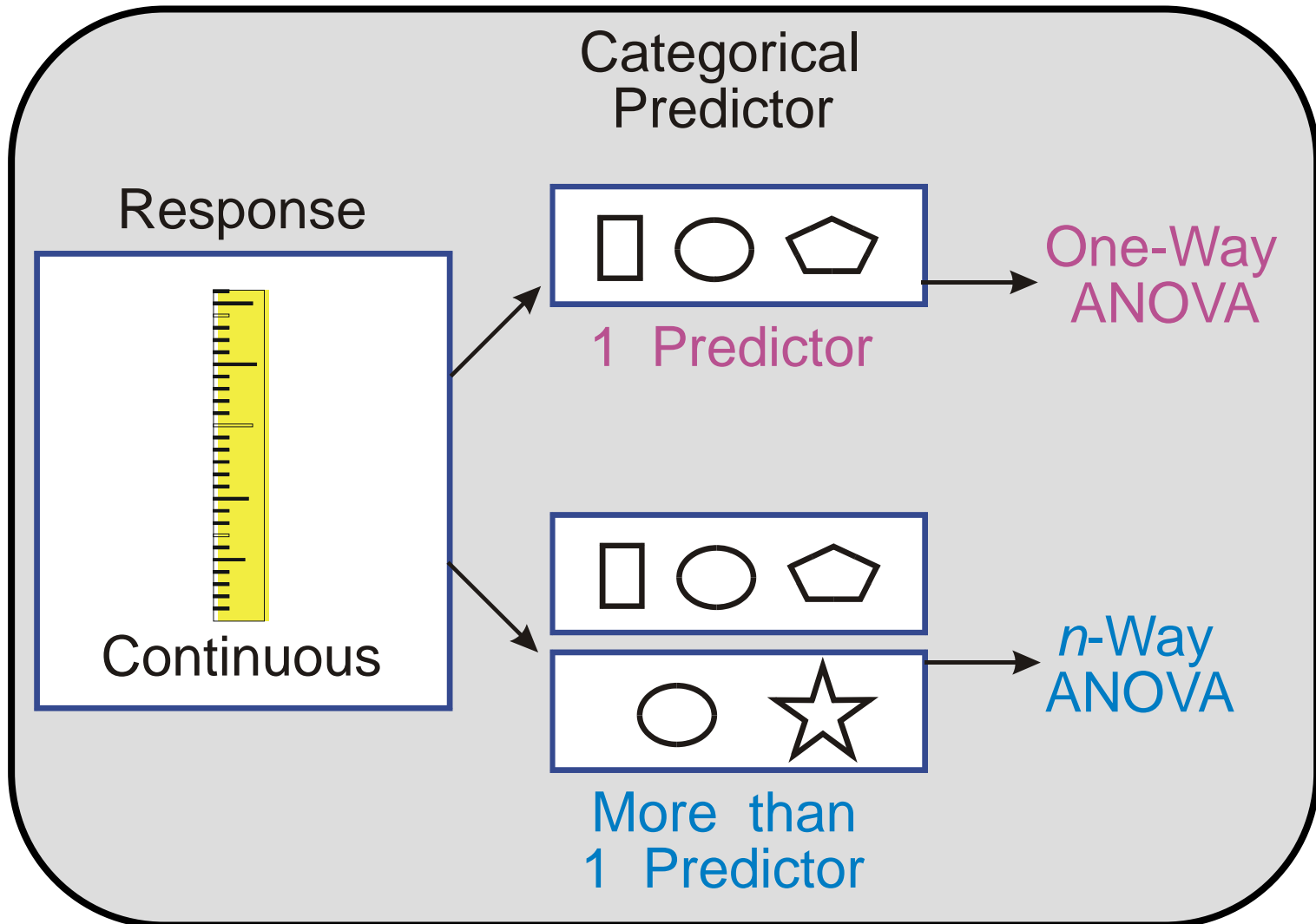
MORE COMPLEX ANOVA & REGRESSION

Institute for Advanced Analytics

MSA Class of 2020

TWO-WAY ANOVA WITH INTERACTIONS

n-Way ANOVA



Additional Linear Models Terminology

- Model – a mathematical relationship between explanatory variables and response variables
- Effect – the expected change in the response that occurs due to the change in the value of an explanatory variable
 - Main Effect – the effect of a single explanatory variable (for example, x_1 , x_2 , x_3)
 - Interaction Effect – the effect of a simultaneous change of two or more explanatory variables (for example, $x_1 * x_2$, $x_1 * x_2 * x_3$)

The Model

BloodP = Base Level + Disease + Drug Dose + DrugDose and Disease + Unaccounted for Variation

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

The Model - Assumptions

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

- Assumptions:
 1. Independent observations
 2. Normality for each treatment
 3. Equal variances

Exploring Data for Two-Way ANOVA

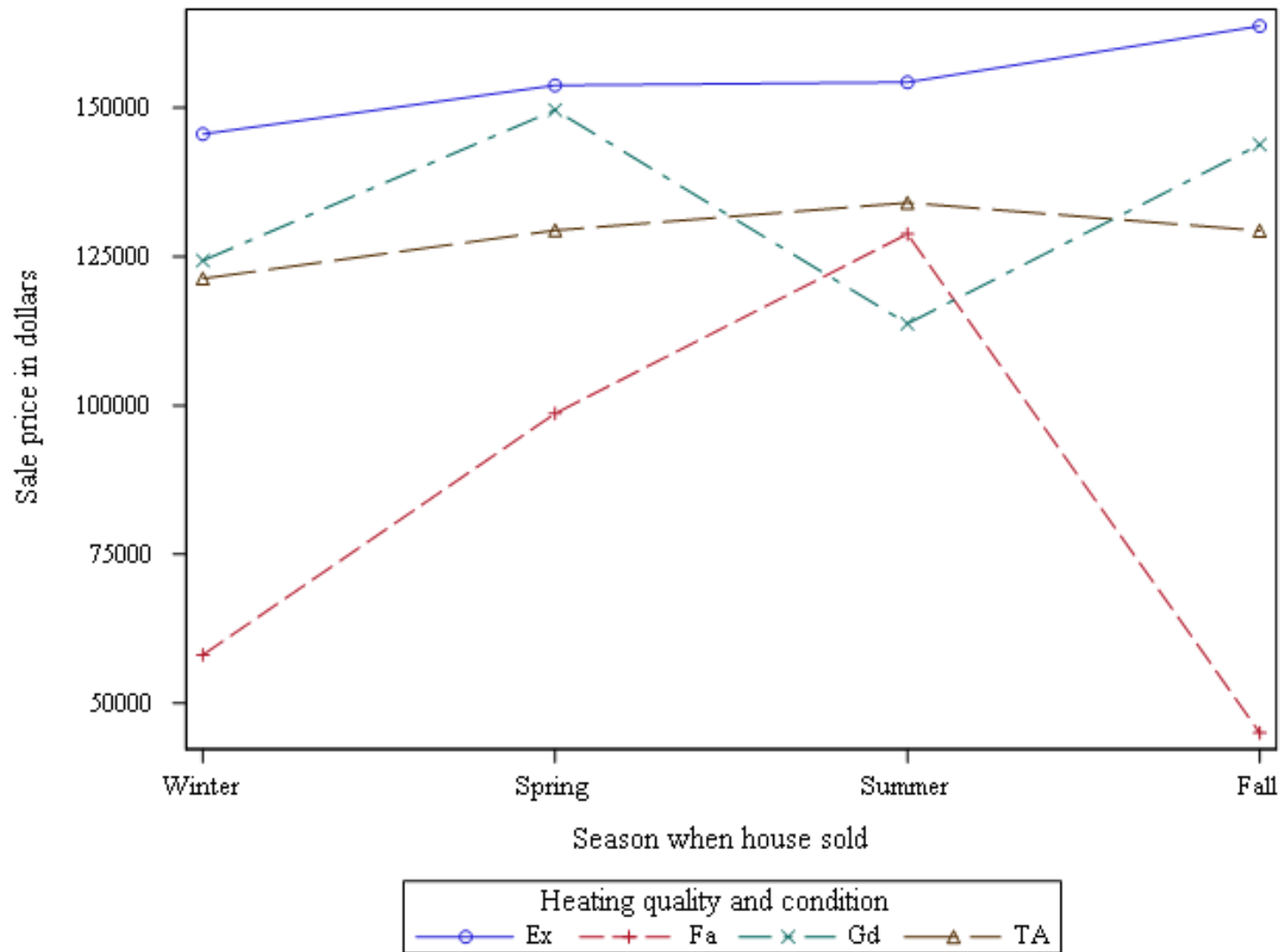
```
proc means data=bootcamp.ameshousing3  
            mean var std nway;  
    class Season_Sold Heating_QC;  
    var SalePrice;  
    format Season_Sold Season.;  
    title 'Selected Descriptive Statistics';  
run;
```

```
proc sgplot data=bootcamp.ameshousing3;  
    vline Season_Sold / group=Heating_QC  
                        stat=mean  
                        response=SalePrice  
                        markers;  
run;
```

Exploring Data for Two-Way ANOVA

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

Exploring Data for Two-Way ANOVA



Two-Way ANOVA

```
proc glm data=bootcamp.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;
```

Two-Way ANOVA

**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

Two-Way ANOVA

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			

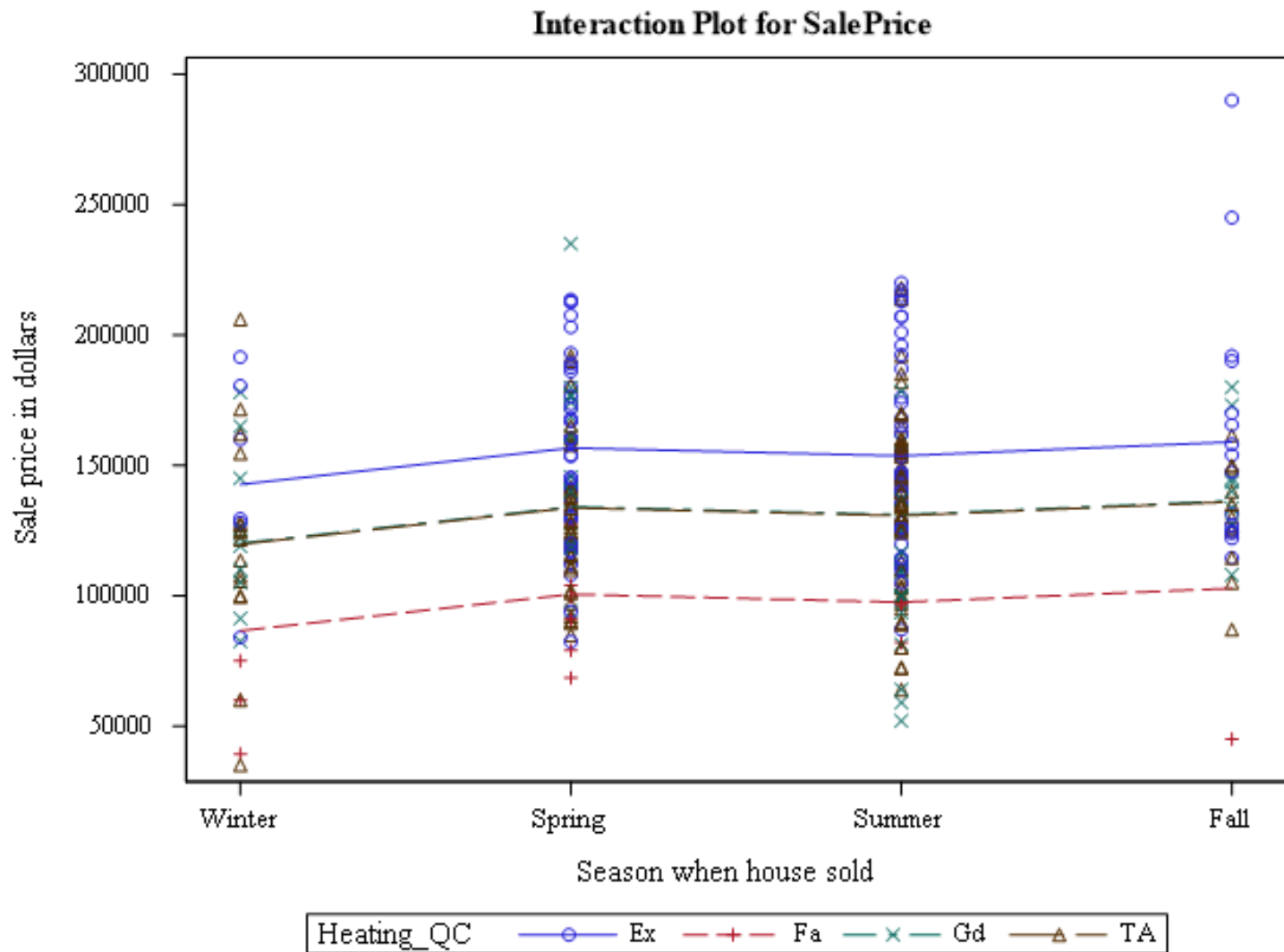
R-Square	Coeff Var	Root MSE	SalePrice Mean
0.171954	25.14764	34584.25	137524.9

Two-Way ANOVA

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

Two-Way ANOVA



Two-Way ANOVA

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

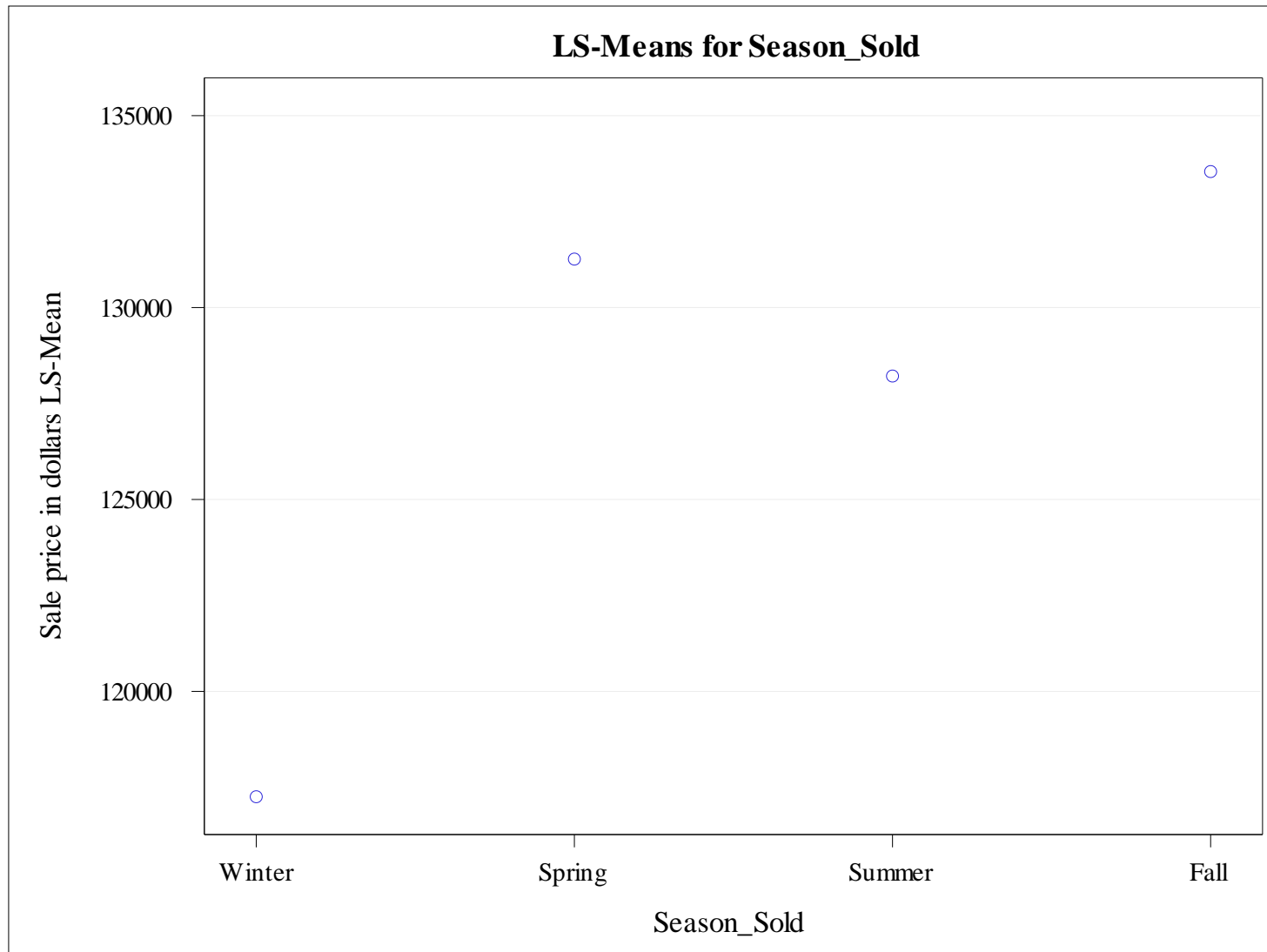
Least Squares Means for effect Season_Sold

Pr > |t| for H0: LSMean(i)=LSMean(j)

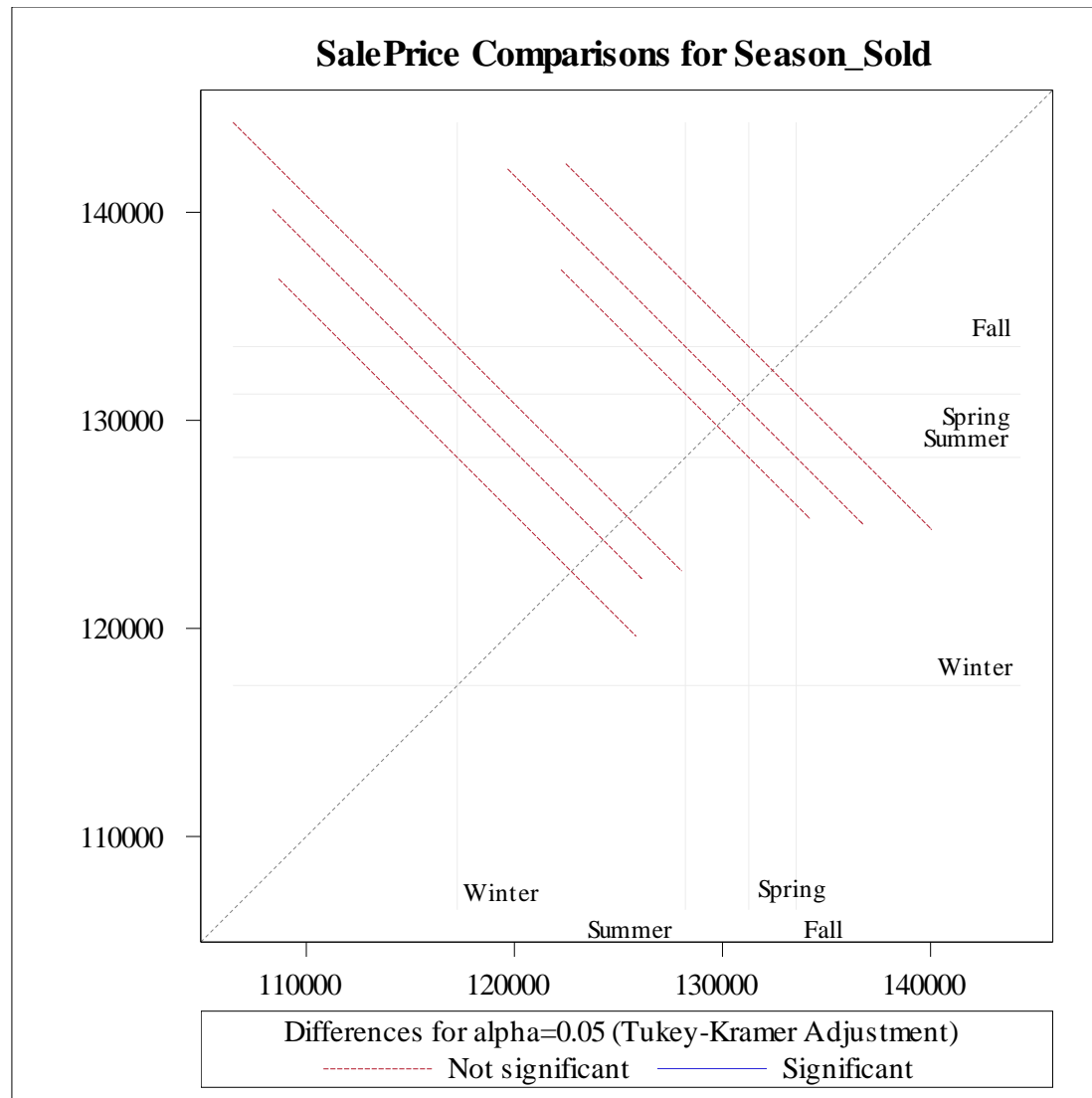
Dependent Variable: SalePrice

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	

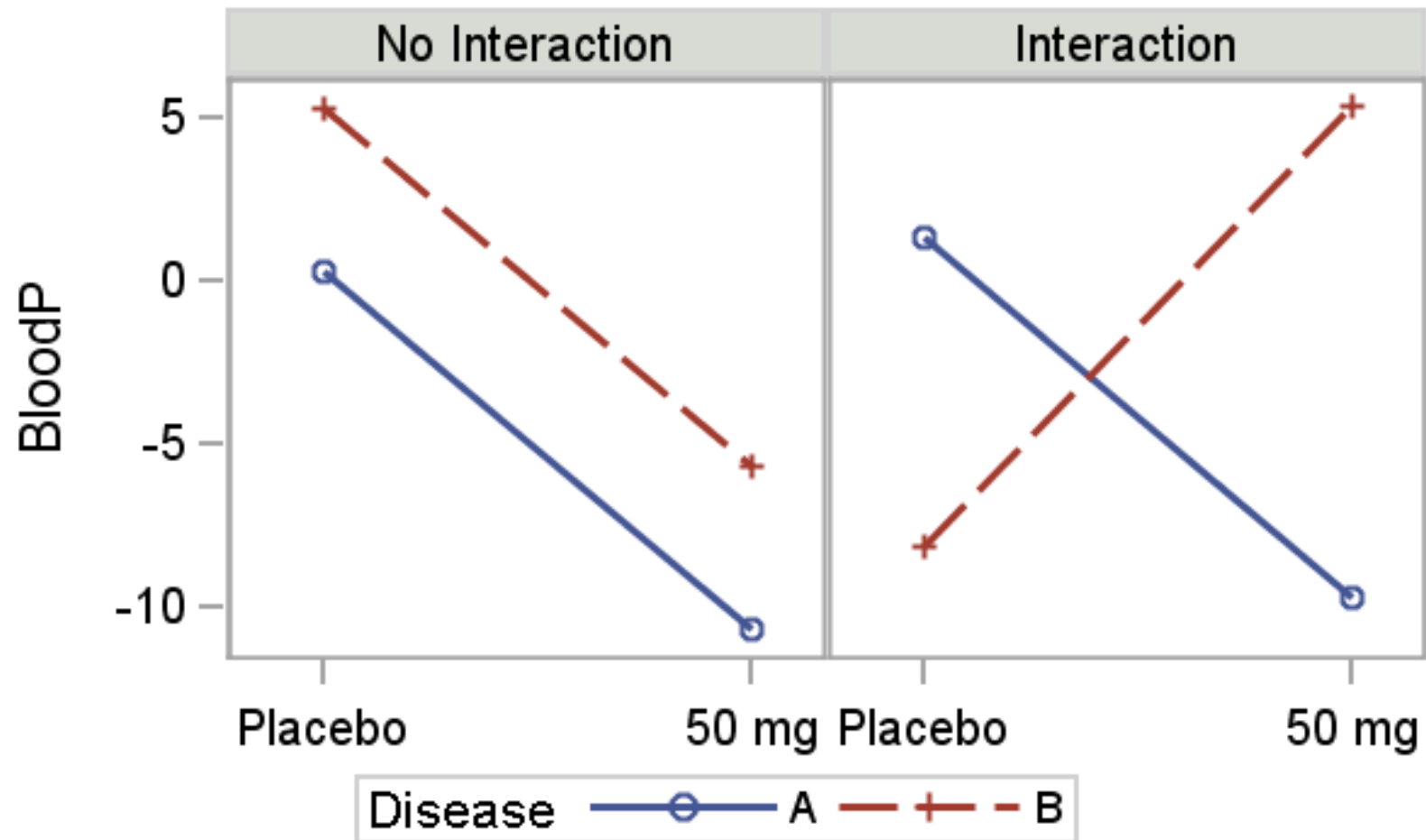
Two-Way ANOVA



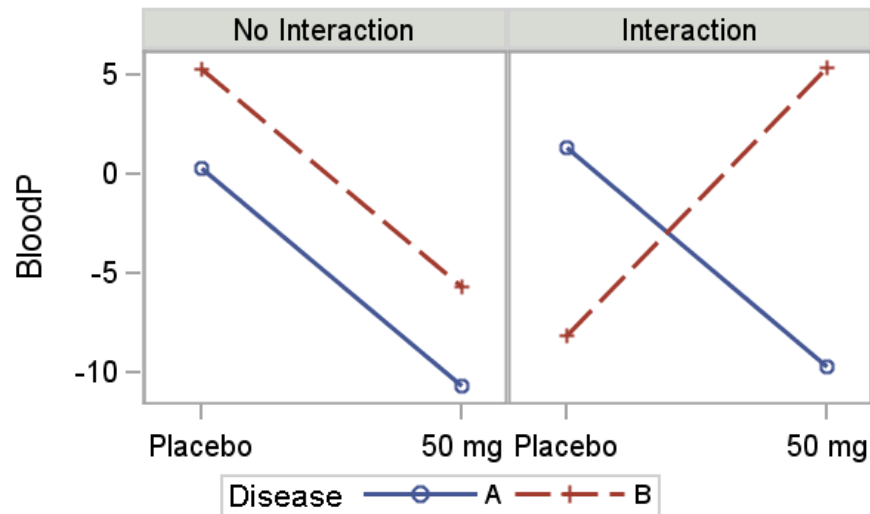
Two-Way ANOVA



Interactions



Interactions



- Interactions occur when changing the level of a factor changes the differences (the relationship) between levels in the other factor.
- Might **MASK** effects of factors if interaction significant!

Nonsignificant Interaction

Analyze the main effects with the interaction in the model.

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

...or...

Delete the interaction from the model, and then analyze the main effects.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

Two-Way ANOVA with Interactions

```
proc glm data=bootcamp.ameshousing3
    order=internal
    plots(only)=intplot;
class Season_Sold Heating_QC;
model SalePrice = Heating_QC Season_Sold
                Heating_QC*Season_Sold / ss1 ss3;
format Season_Sold Season.;
title "Model with Heating Quality and Season as
      Interacting Predictors";

run;
quit;
```

Two-Way ANOVA with Interactions

**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			

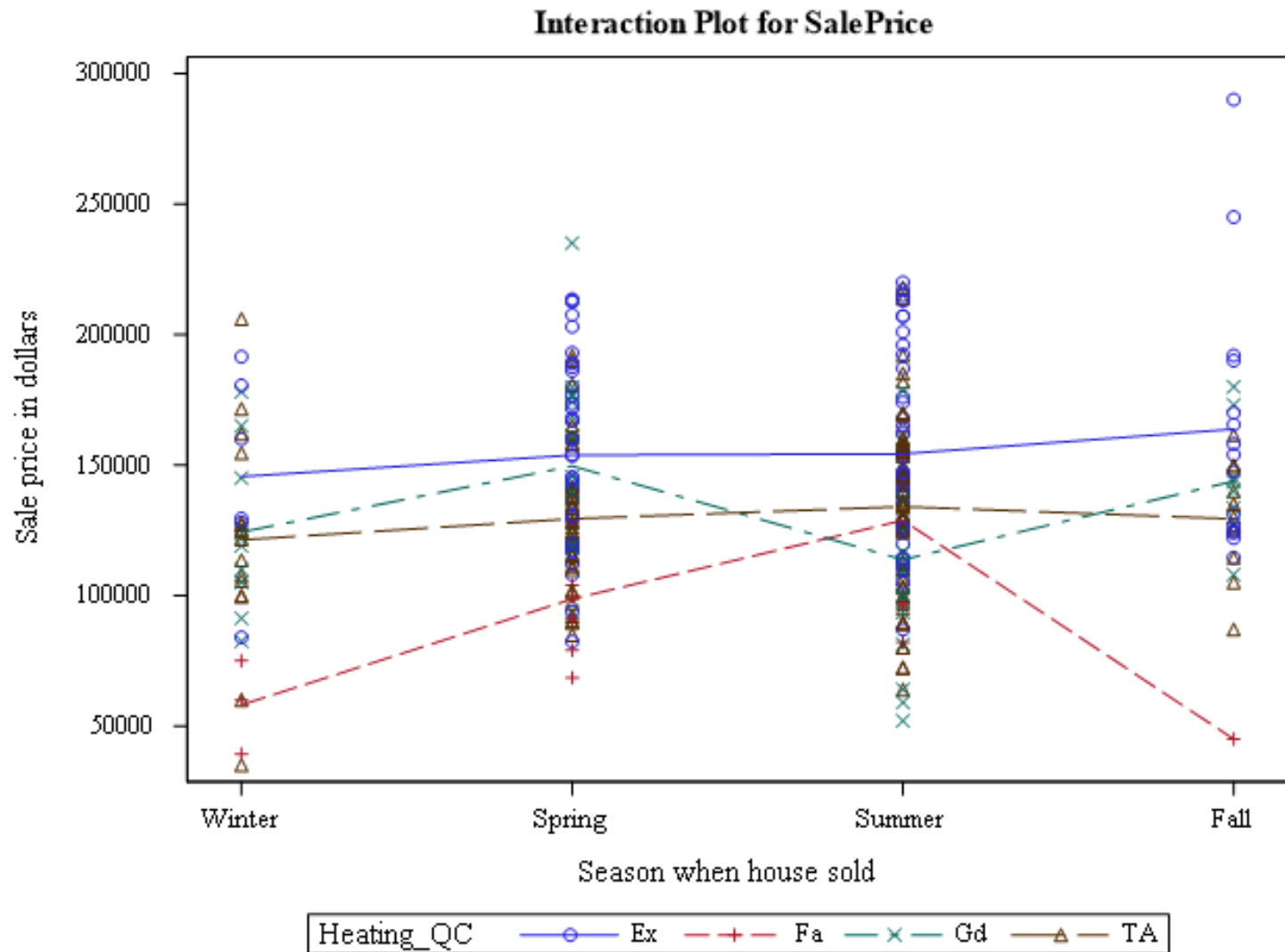
R-Square	Coeff Var	Root MSE	SalePrice Mean
0.230634	24.62130	33860.40	137524.9

Two-Way ANOVA with Interactions

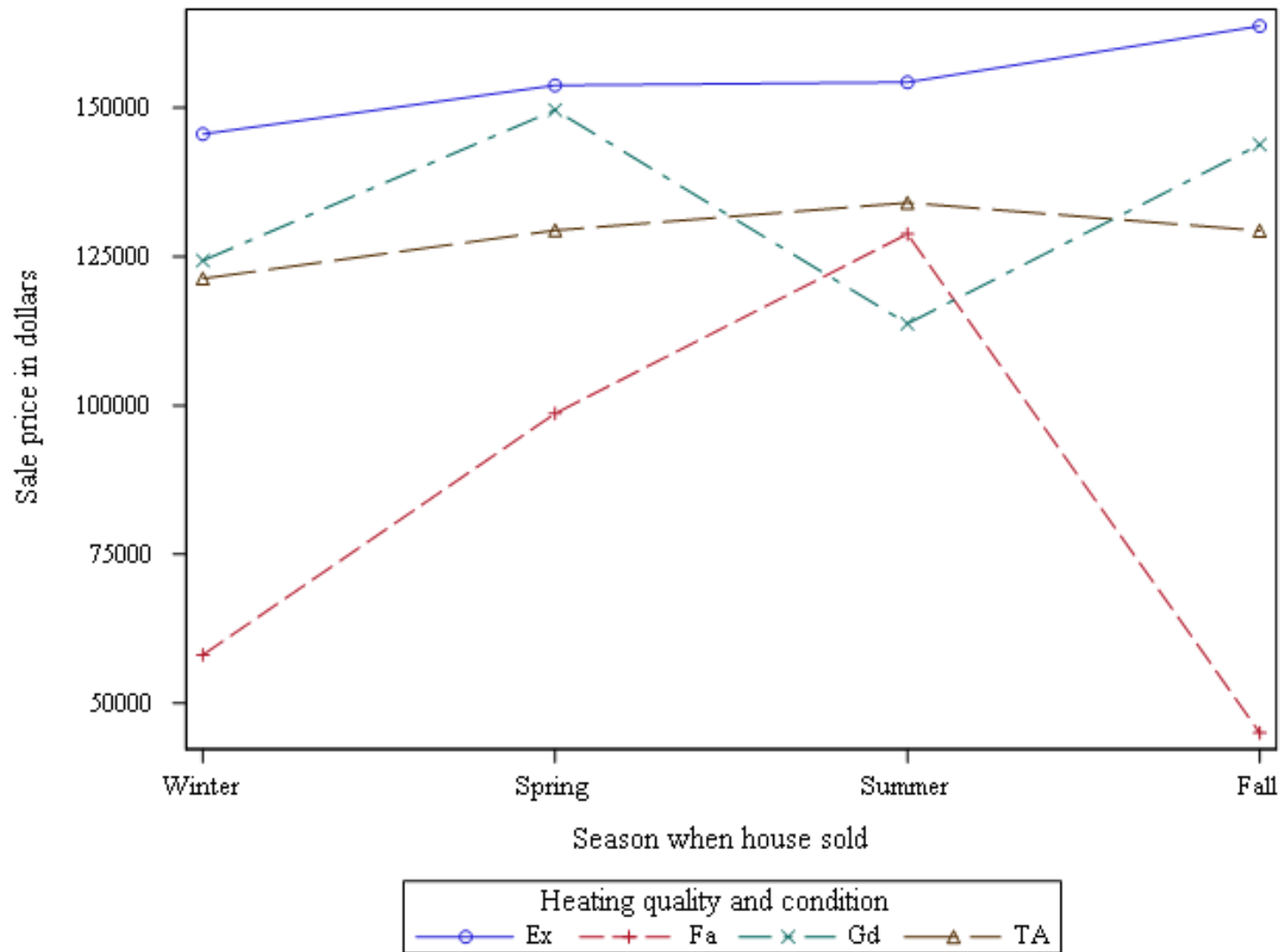
Source	DF	Type I SS	Mean Square	F Value	Pr > F
Heating_QC	3	66835556221	22278518740	19.43	<.0001
Season_Sold	3	5939259845	1979753282	1.73	0.1617
Season_So*Heating_QC	9	24835058089	2759450899	2.41	0.0121

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

Two-Way ANOVA with Interactions



Exploring Data for Two-Way ANOVA



Sliced ANOVA Results

```
proc glm data=bootcamp.ameshousing3
    order=internal
    plots(only)=intplot;
    class Season_Sold Heating_QC;
    model SalePrice = Heating_QC|Season_Sold;
    lsmeans Heating_QC*Season_Sold / diff slice = Heating_QC;
    store out = interact;
    title "Analyze the Effects of Season";
    title2 "at Each Level of Heating Quality";
    format Season_Sold Season.;
run;
quit;
```

Sliced ANOVA Results

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

Sliced ANOVA Results

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	

Sliced ANOVA Results

Analyze the Effects of Season
at Each Level of Heating Quality

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.

Further Slicing Results

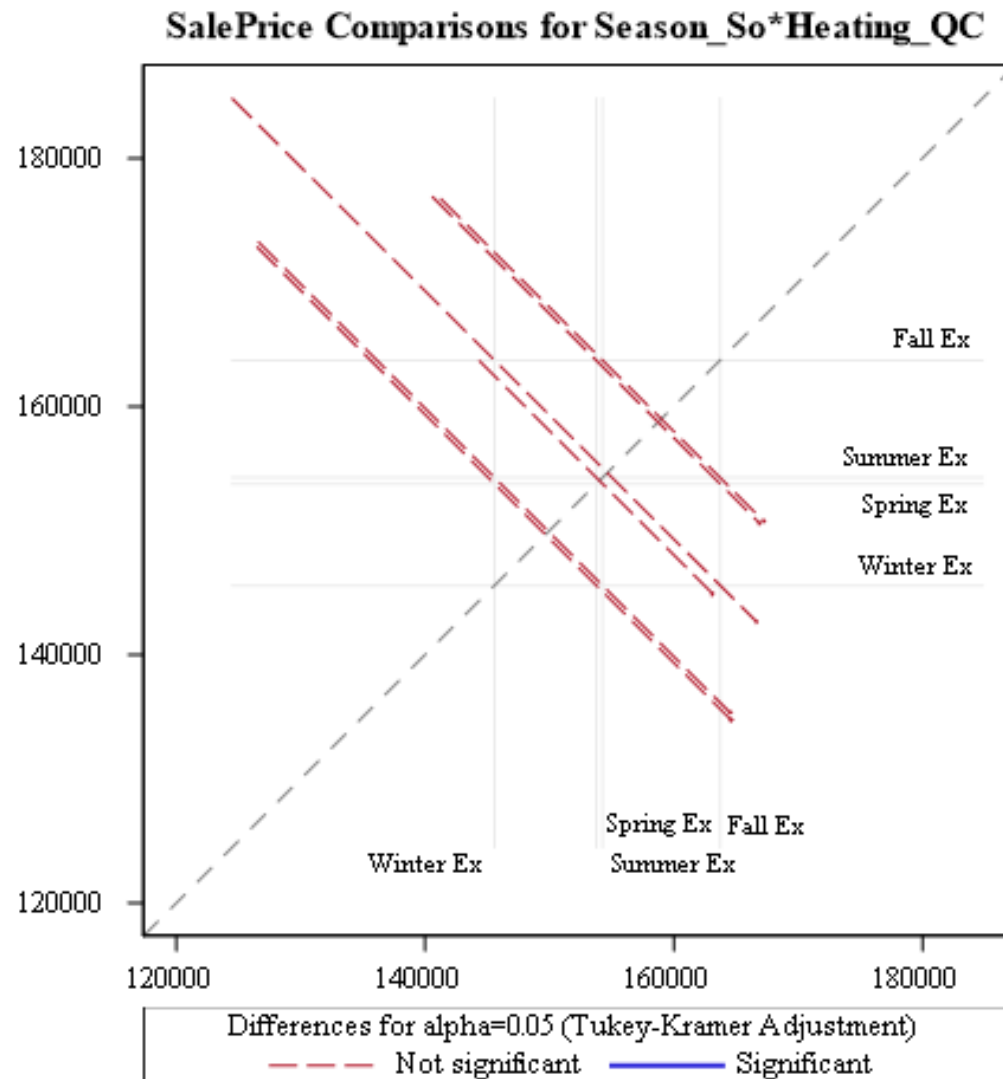
```
proc plm restore = interact plots = all;  
    slice Heating_QC*Season_Sold / sliceby = Heating_QC  
                                   adjust=tukey;  
    effectplot interaction(sliceby = Heating_QC) / clm;  
run;
```

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

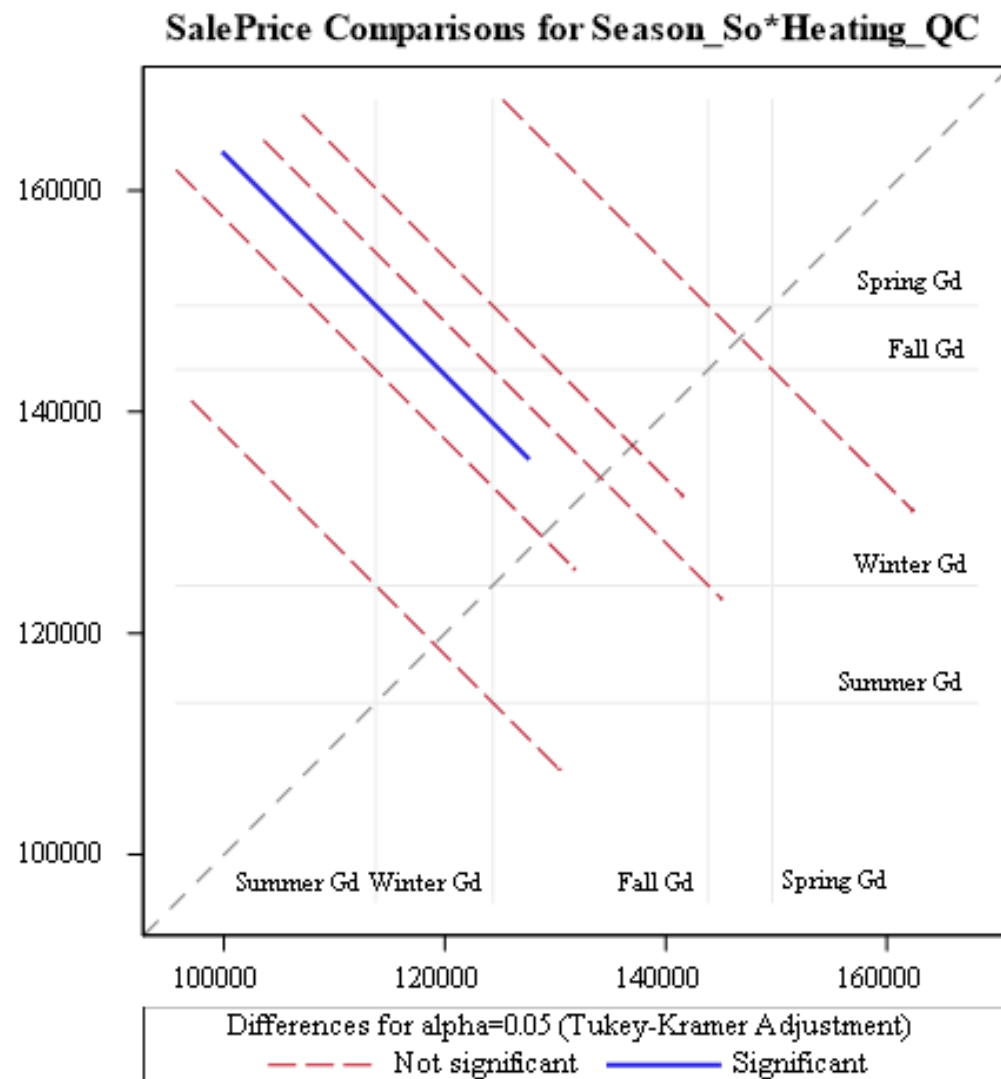
**Simple Differences of Season_So*Heating_QC Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer**

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	-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

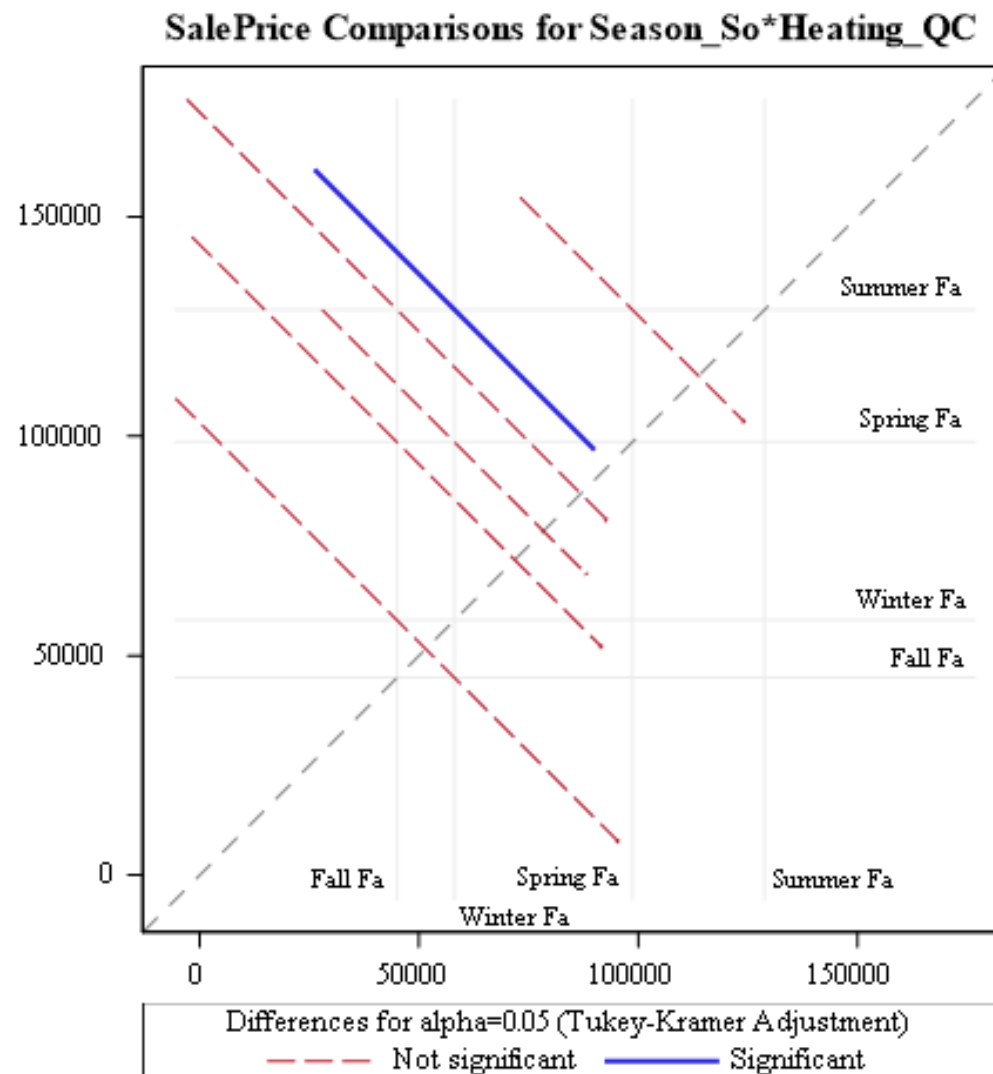
Further Slicing Results



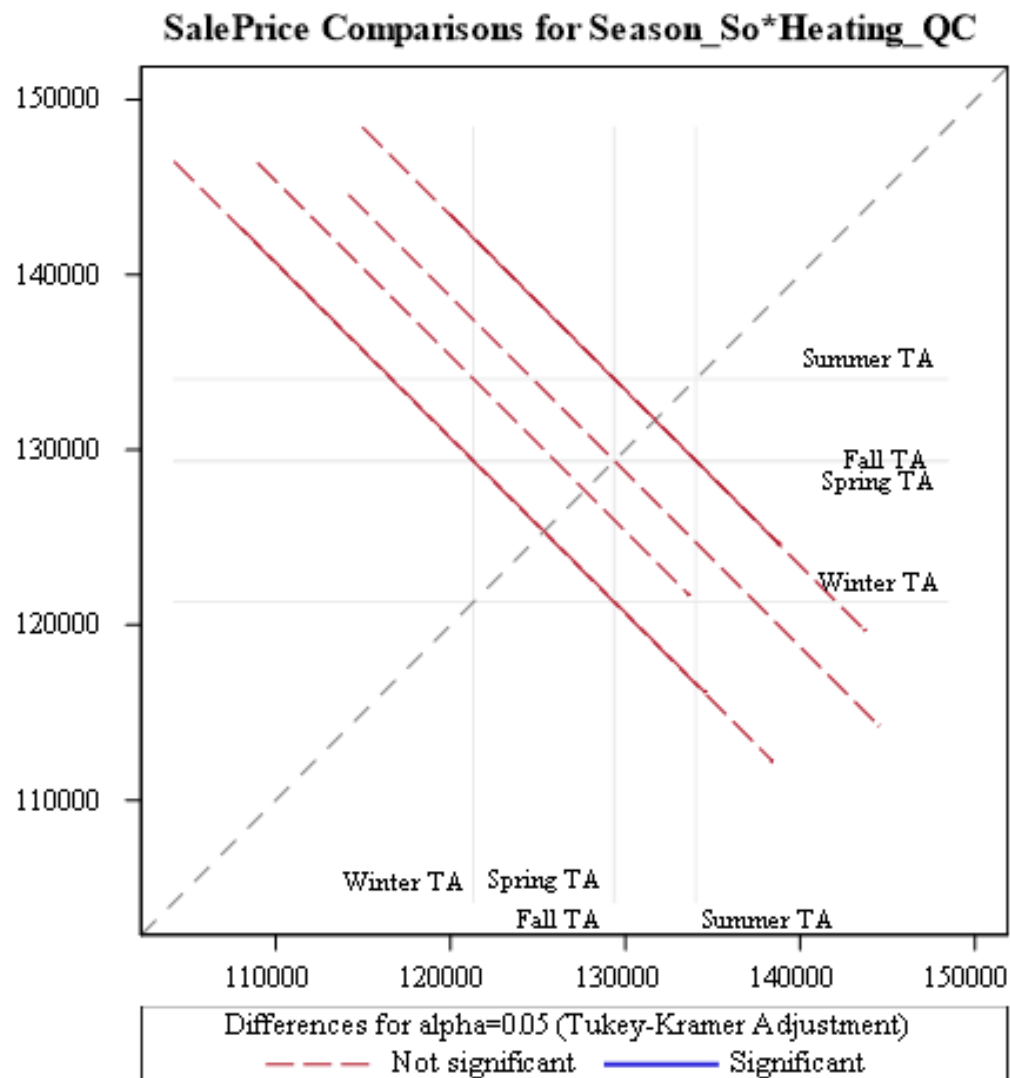
Further Slicing Results



Further Slicing Results



Further Slicing Results





RANDOMIZED BLOCK DESIGN FOR ANOVA

OPTIONAL SELF STUDY

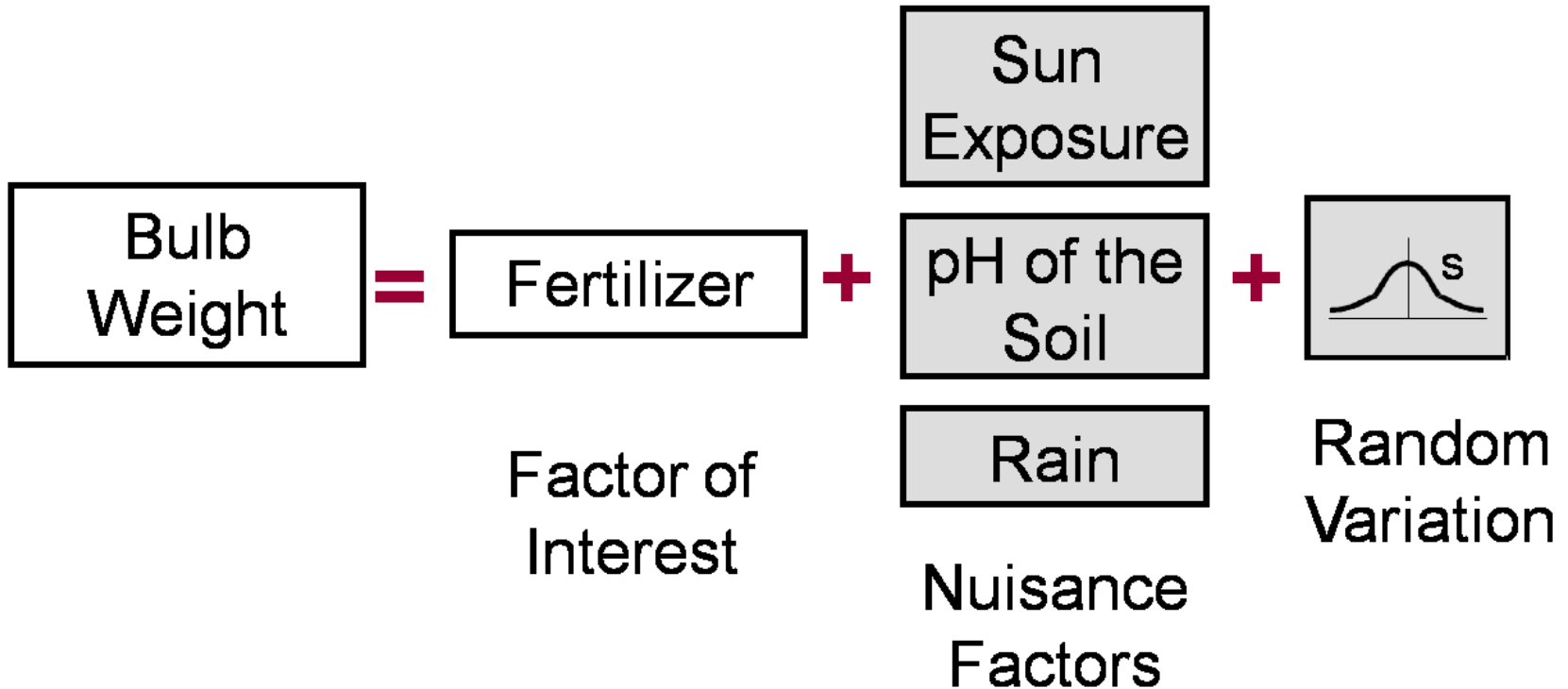
Observational or Retrospective Studies

- Groups can be naturally occurring.
 - Gender and ethnicity
- Random assignment might be unethical or untenable.
 - Smoking or credit risk groups
- Often you look at what already happened (retrospective) instead of following through to the future (prospective).
- You have little control over other factors contributing to the outcome measure.

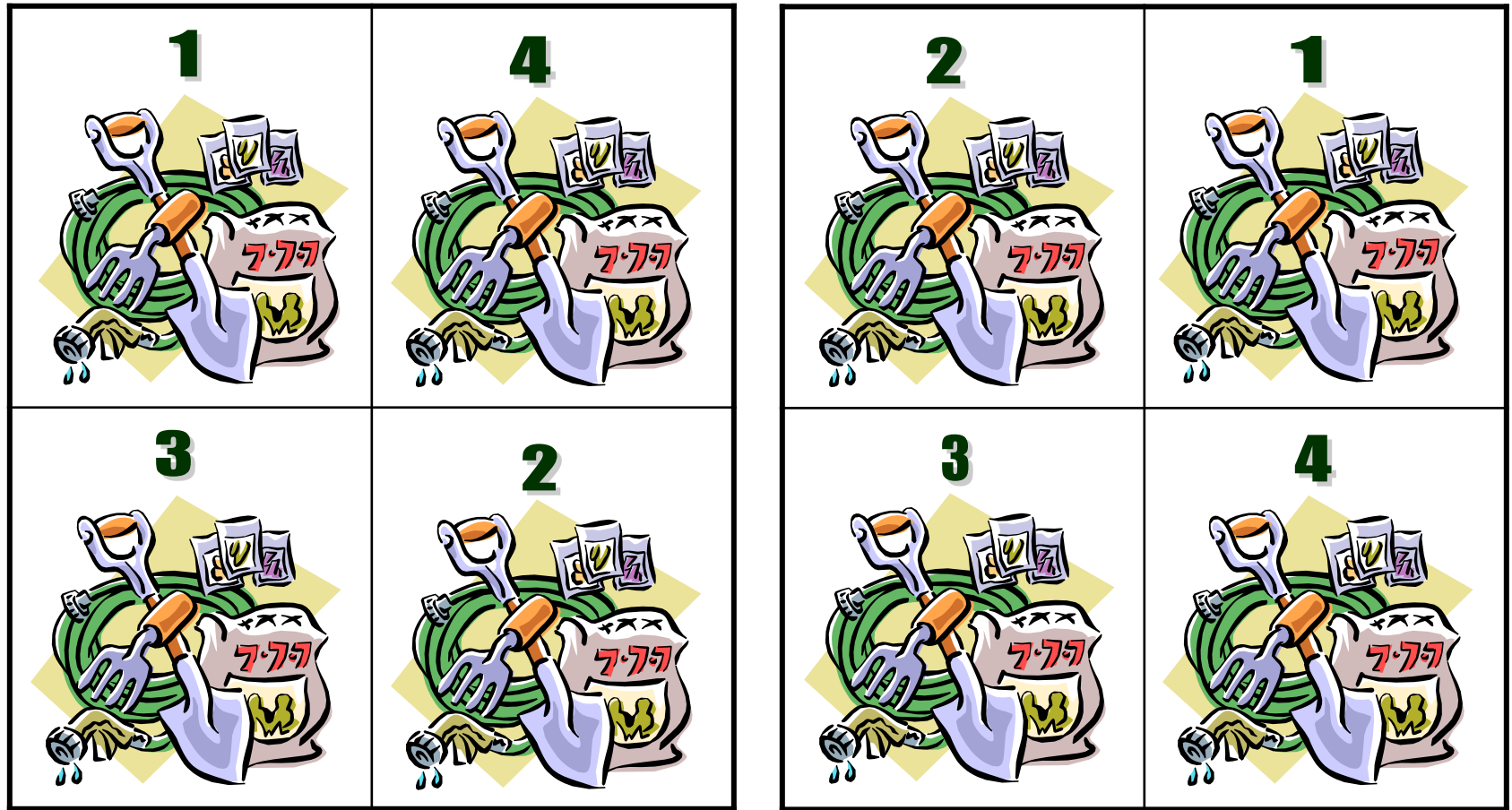
Controlled Experiments

- Random assignment might be desirable to eliminate selection bias.
- You often want to look at the outcome measure prospectively.
- You can manipulate the factors of interest and can more reasonably claim causation.
- You can design your experiment to control for other factors contributing to the outcome measure.

Nuisance Factors



Assigning Treatments within Blocks

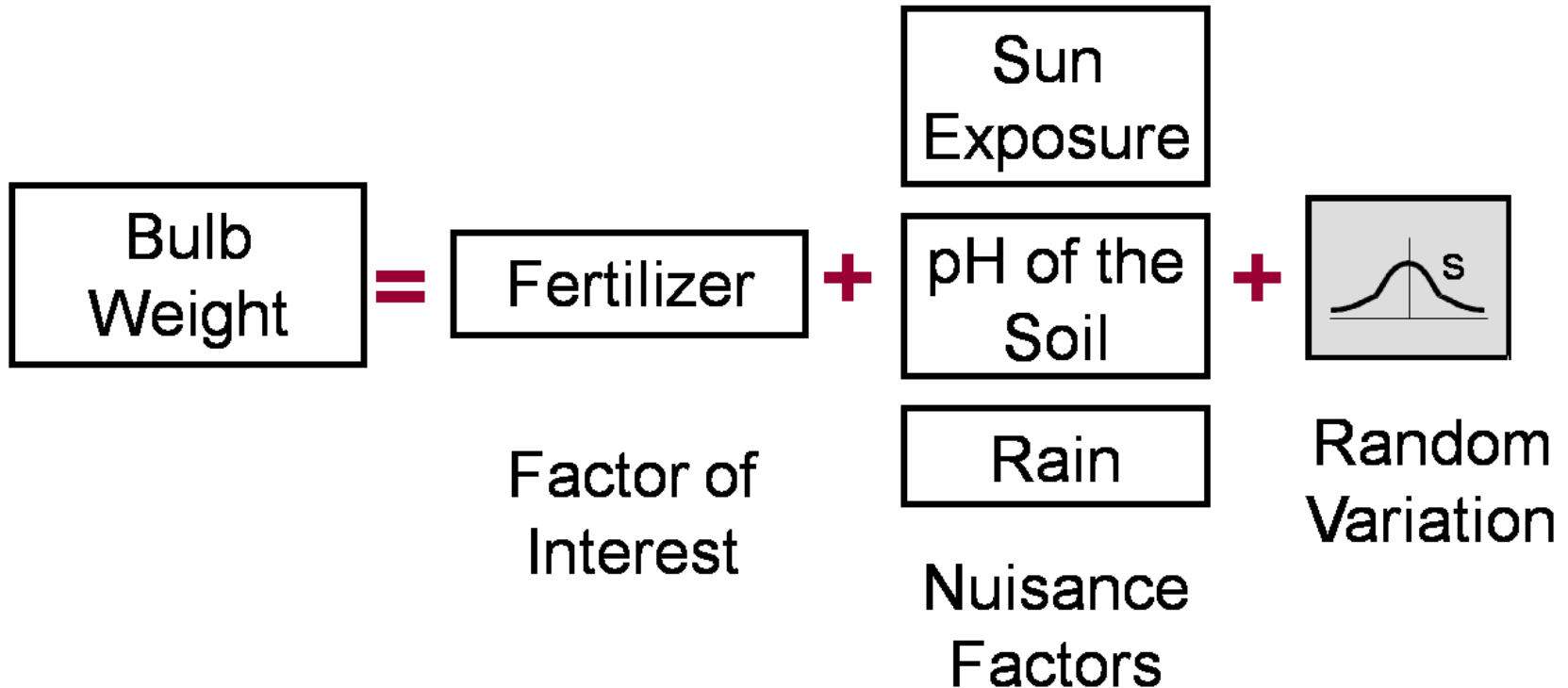


Including a Blocking Variable in the Model

Bulb Weight = Base Level + Sector + Fertilizer Type + Unaccounted for Variation

$$Y_{ijk} = \mu + \alpha_i + \tau_j + \varepsilon_{ijk}$$

Nuisance Factors



Including a Blocking Variable in the Model

- Additional assumptions are as follows:
 - Treatments are randomly assigned within each block.
 - The effects of the treatment factor are constant across the levels of the blocking variable.
- In the garlic example, the design is balanced, which means that there is the same number of garlic samples for every **Fertilizer/Sector** combination.

ANOVA with Blocking

```
proc glm data=sasuser.MGGarlic_Block plots (only)=diagnostics;  
  class Fertilizer Sector;  
  model BulbWt=Fertilizer Sector;  
  title 'ANOVA for Randomized Block Design';  
run;  
quit;
```

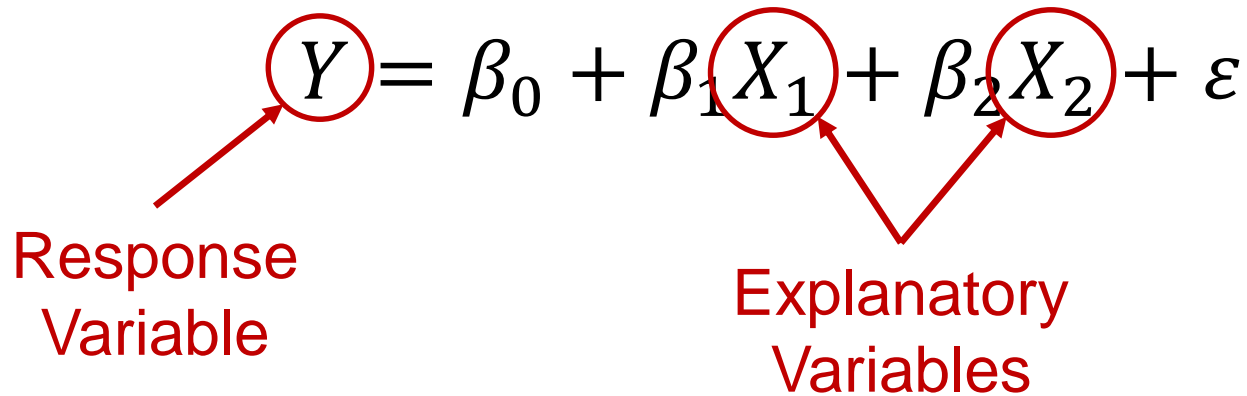
CONCEPTS OF MULTIPLE LINEAR REGRESSION

Multiple Linear Regression with Two Variables

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

Response Variable

Explanatory Variables



The diagram shows the equation $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. The variable Y is circled in red, and a red arrow points from the text 'Response Variable' to it. The variables X_1 and X_2 are also circled in red, and a red arrow points from the text 'Explanatory Variables' to both of them.

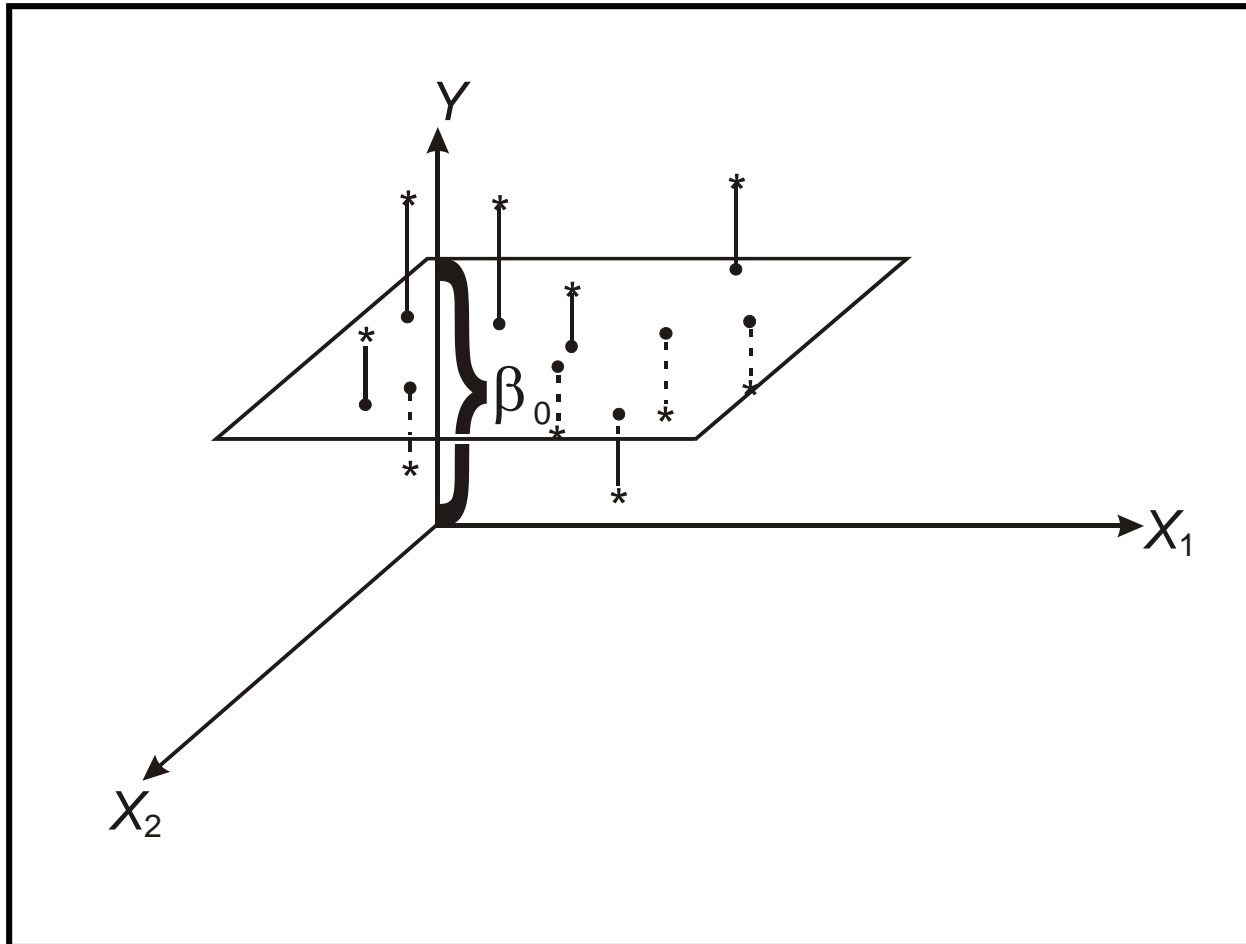
Multiple Linear Regression with Two Variables

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

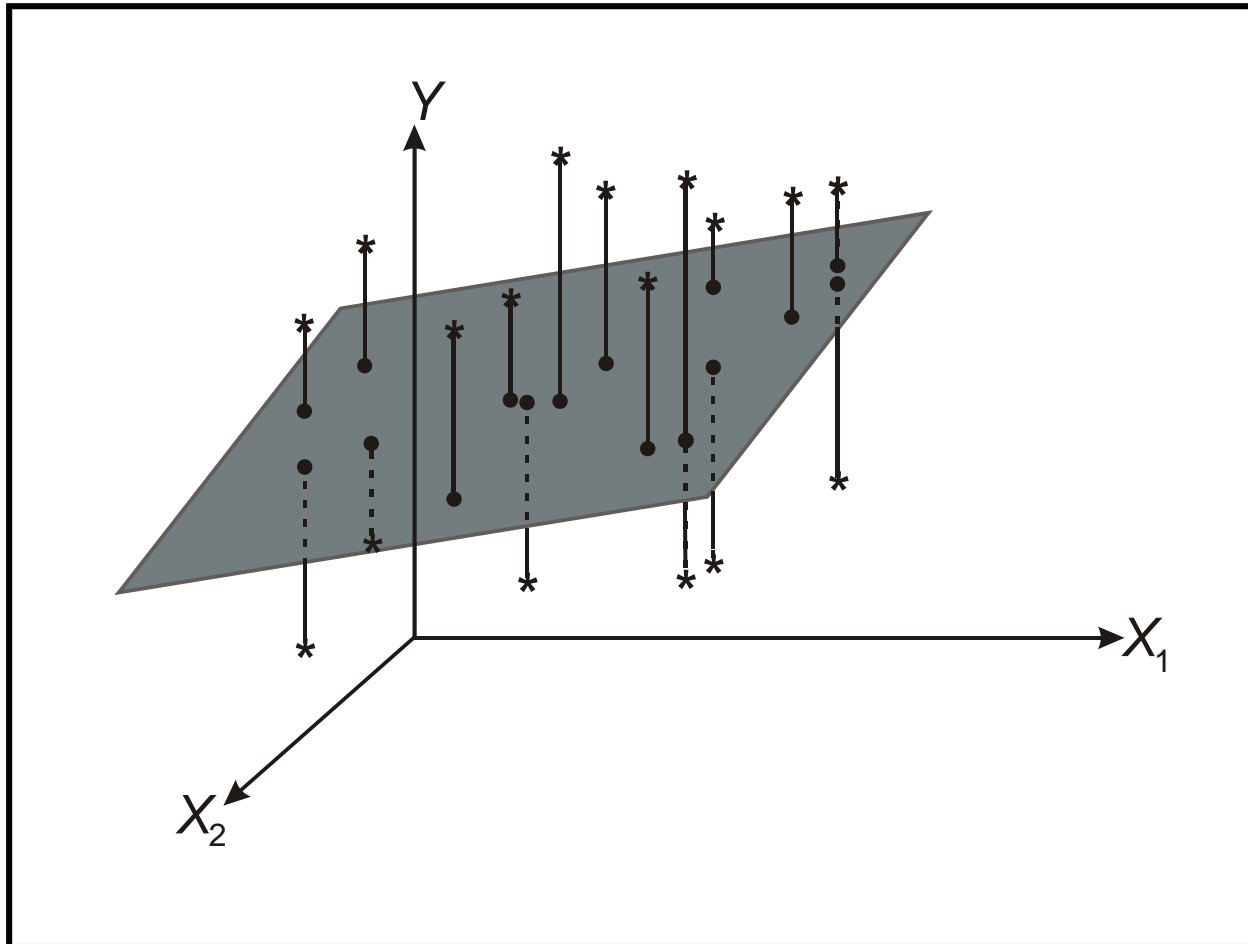
Unknown Coefficients

Error Term

Picturing the Model: No Relationship



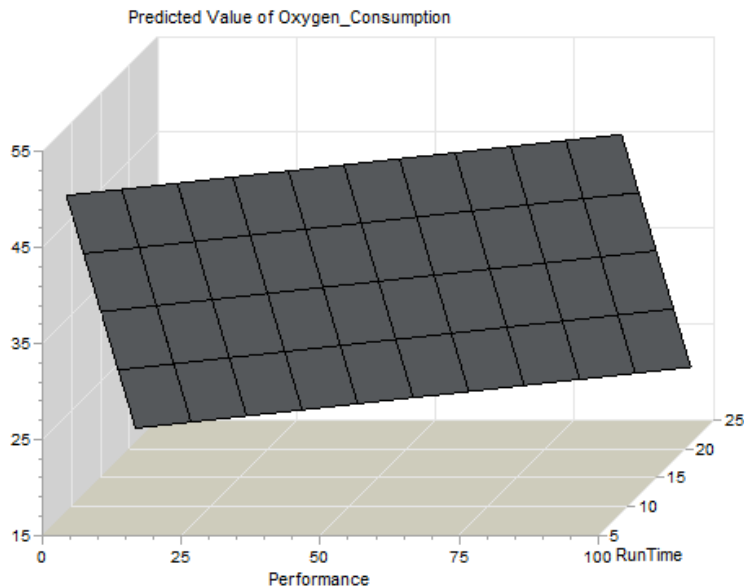
Picturing the Model: A Relationship



The Multiple Linear Regression Model

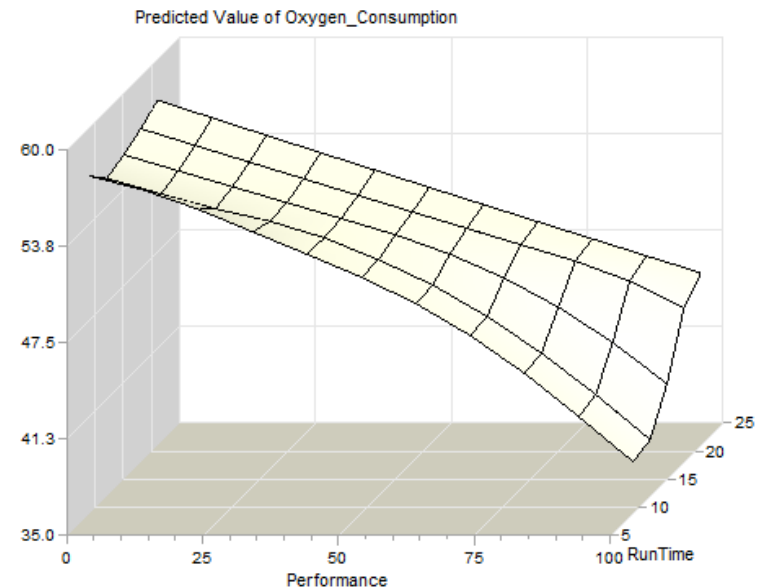
- In general, you model the dependent variable, Y , as a linear function of k independent variables, X_1 through X_k :

- $$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon$$



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

Linear Model with
only Linear Effects



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

Linear Model with
Nonlinear Effects

Model Hypothesis Test

- **Null Hypothesis:**

- The regression model does ***not*** fit the data better than the baseline model.
- $\beta_1 = \beta_2 = \dots = \beta_k = 0$

- **Alternative Hypothesis:**

- The regression model does fit the data better than the baseline model.
- Not all β_i s equal zero.

Assumptions for Linear Regression

- The mean of the Ys is accurately modeled by a **linear** function of the Xs.
- The random error term, ε , is assumed to have a **normal** distribution with a mean of zero.
- The random error term, ε , is assumed to have a **constant variance**, σ^2 .
- The errors are **independent**.
- **No perfect collinearity**

Multiple Linear vs. Simple Linear Regression

- **Main Advantage**

- Multiple linear regression enables you to investigate the relationship among Y and several independent variables simultaneously.

- **Main Disadvantages**

- Increased complexity makes it more difficult to do the following:
 - ascertain which model is “best”
 - interpret the models

Common Applications

- Multiple linear regression is a powerful tool for the following tasks:
 - **Predict** – to develop a model to predict future values of a response variable (Y) based on its relationships with other predictor variables (X s)
 - **Explain** – to develop an understanding of the relationships between the response variable and predictor variables

Predict

- The terms in the model, the values of their coefficients, and their statistical significance are of secondary importance.
- The focus is on producing a model that is the best at predicting future values of Y as a function of the Xs. The predicted value of Y is given by this formula:

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

Explain

- The focus is on understanding the relationship between the dependent variable and the independent variables.
- Consequently, the statistical significance of the coefficients is important as well as the magnitudes and signs of the coefficients.

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

Problem with R^2

- The problem with the calculation of R^2 in a multiple linear regression is that the addition of any variable (good or bad) will make the R^2 value increase if even slightly.

$$R^2 = 1 - \frac{SSE}{TSS}$$

Will never increase with the addition of a variable.

Adjusted Coefficient of Determination

- To account for this problem, most people use the **adjusted coefficient of determination**, R_a^2 .
- The calculations are as follows:

$$R_a^2 = 1 - \left(\frac{n - 1}{n - (k + 1)} \right) \left(\frac{SSE}{TSS} \right)$$

OR

$$R_a^2 = 1 - (1 - R^2) \left(\frac{n - 1}{n - (k + 1)} \right)$$

Adjusted Coefficient of Determination

- The R_a^2 penalizes a model for adding a variable that does not provide any useful information.

$$R_a^2 \leq R^2$$

- The adjusted coefficient of determination loses its interpretation (because it could be negative!), but is better at determining utility of a model.

Fitting a Multiple Linear Regression

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

Fitting a Multiple Linear Regression

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

Fitting a Multiple Linear Regression

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			

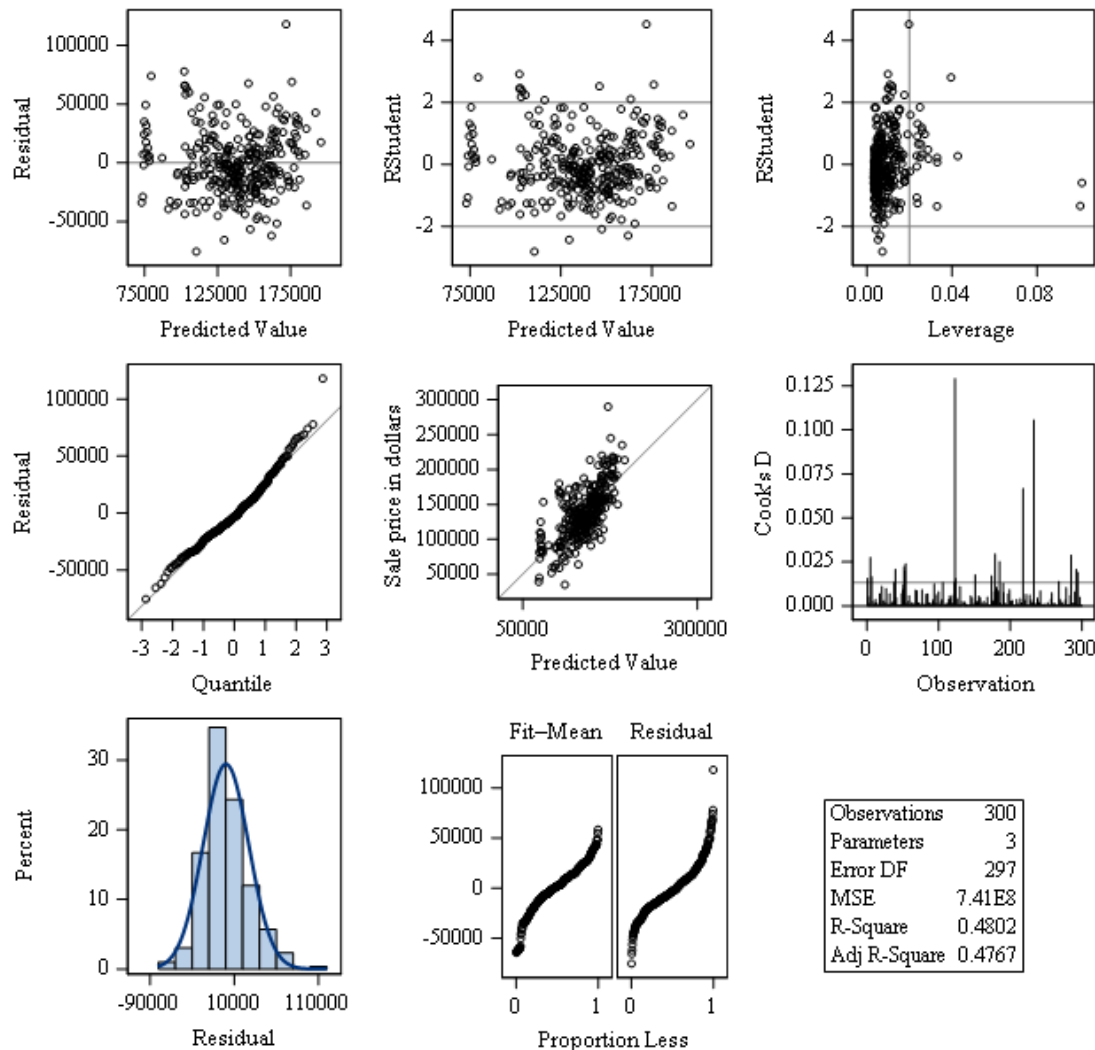
Fitting a Multiple Linear Regression

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

Fitting a Multiple Linear Regression

Fit Diagnostics for SalePrice



Fitting a Multiple Linear Regression

