INTRODUCTION TO ANOVA & REGRESSION

Institute for Advanced Analytics

MSA Class of 2020

Overview of Models in This Course

Type of Response \ Type of Predictors	Categorical	Continuous	Continuous and Categorical
Continuous	Analysis of Variance (ANOVA)	Ordinary Least Squares (OLS) Regression	Ordinary Least Squares (OLS) Regression
Categorical	Logistic Regression	Logistic Regression	Logistic Regression

Linear Models Terminology

- **Linear Model** $Y = \beta_0 + \beta_1 X_1 \ldots + \beta_k X_k + \varepsilon$

- **Explanatory** (*Input, Predictor, Independent*) **Variable**
- **Response** (*Target, Dependent*) **Variable**

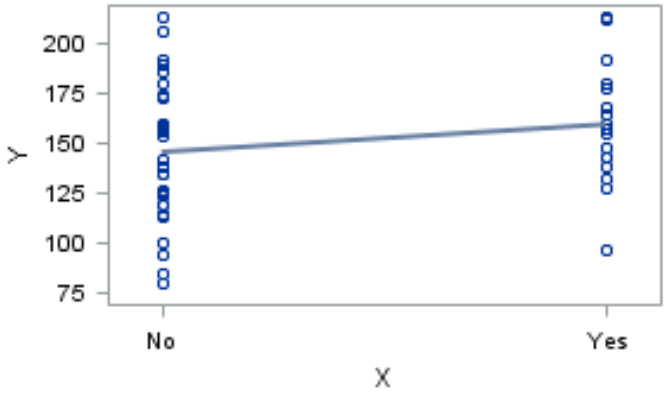
- **Explanatory Modeling**
  - Inferential Statistics
  - (→ Hypothesis Testing)
- **Predictive Modeling**

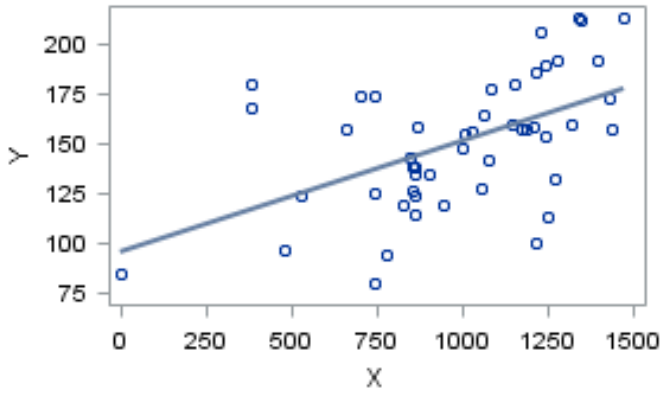
Overview of Models

- General Linear Models

$$Y = \beta_0 + \beta_1 X_1 \ldots + \beta_k X_k + \varepsilon$$

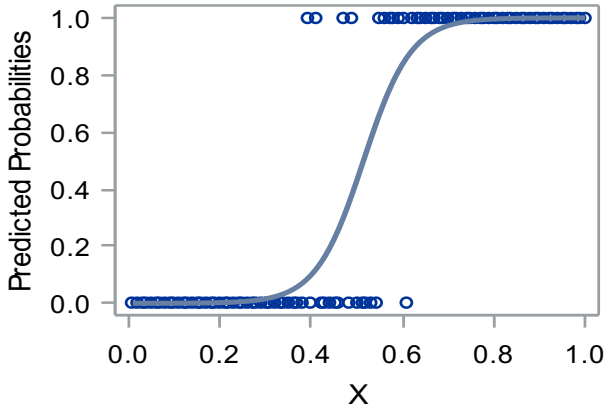
  - Analysis of Variance (ANOVA)
  - Regression





  - Logistic Regression

$$\text{logit}(Y) = \beta_0 + \beta_1 X_1 \ldots + \beta_k X_k$$



Explanatory versus Predictive Modeling

Explanatory Modeling

- How is x_i related to y_i?
- Descriptions
- Small Samples (sometimes)
- Fewer Variables
- Assessed using p-values and confidence intervals

Predictive Modeling

- Given x_i, can I predict y_i?
- Predictions
- Large Samples (ideally)
- Many Variables
- Assessed using error metrics on a holdout sample

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Explanatory versus Predictive Modeling

Explanatory Modeling

- How is x_i related to y_i?
- Descriptions
- Small Samples (sometimes)
- Fewer Variables
- Assessed using p-values and confidence intervals

Predictive Modeling

- Given x_i, can I predict y_i?
- Predictions
- Large Samples (ideally)
- Many Variables
- Assessed using error metrics on a holdout sample

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

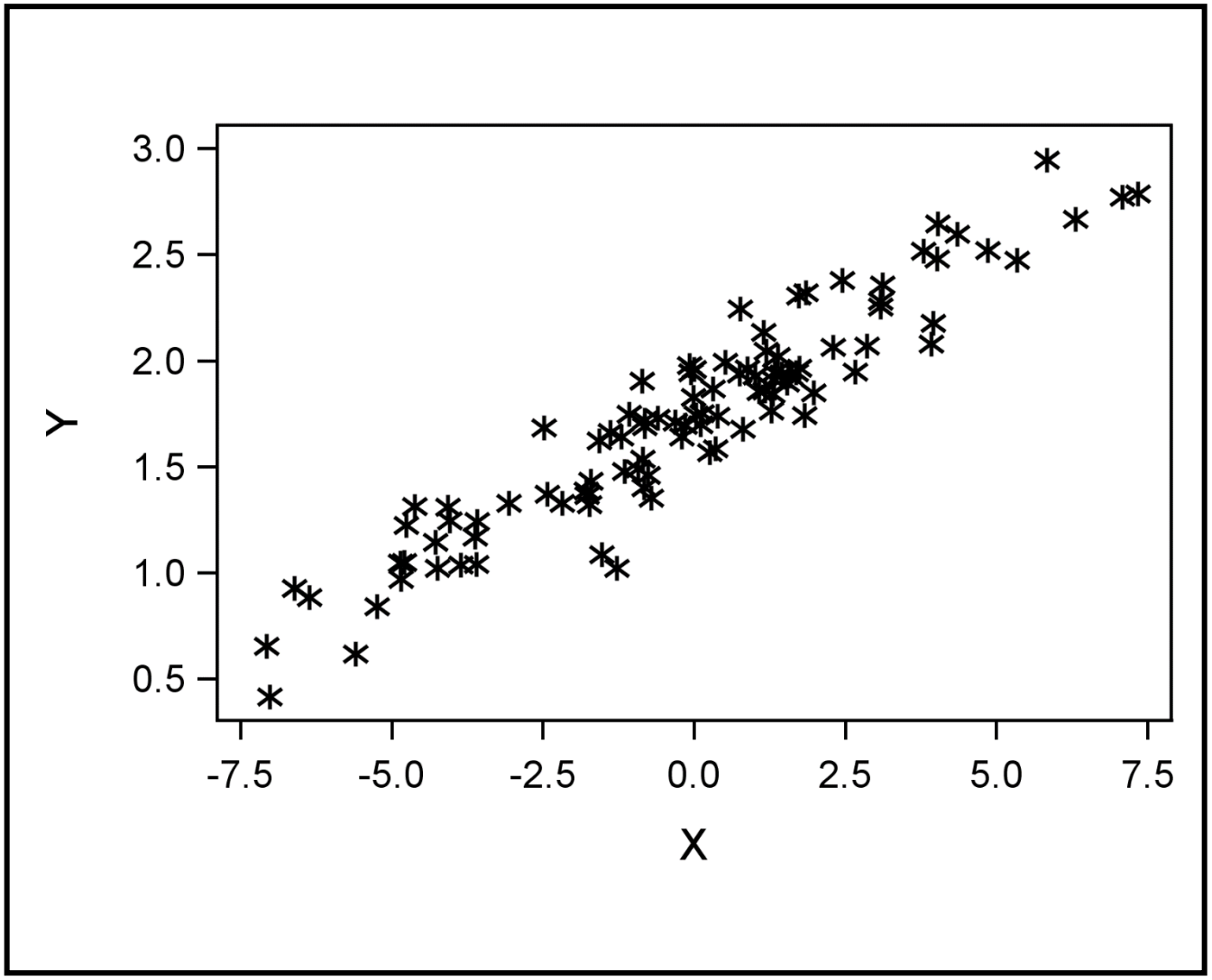
EXPLORATORY DATA ANALYSIS

Associations between 2 variables

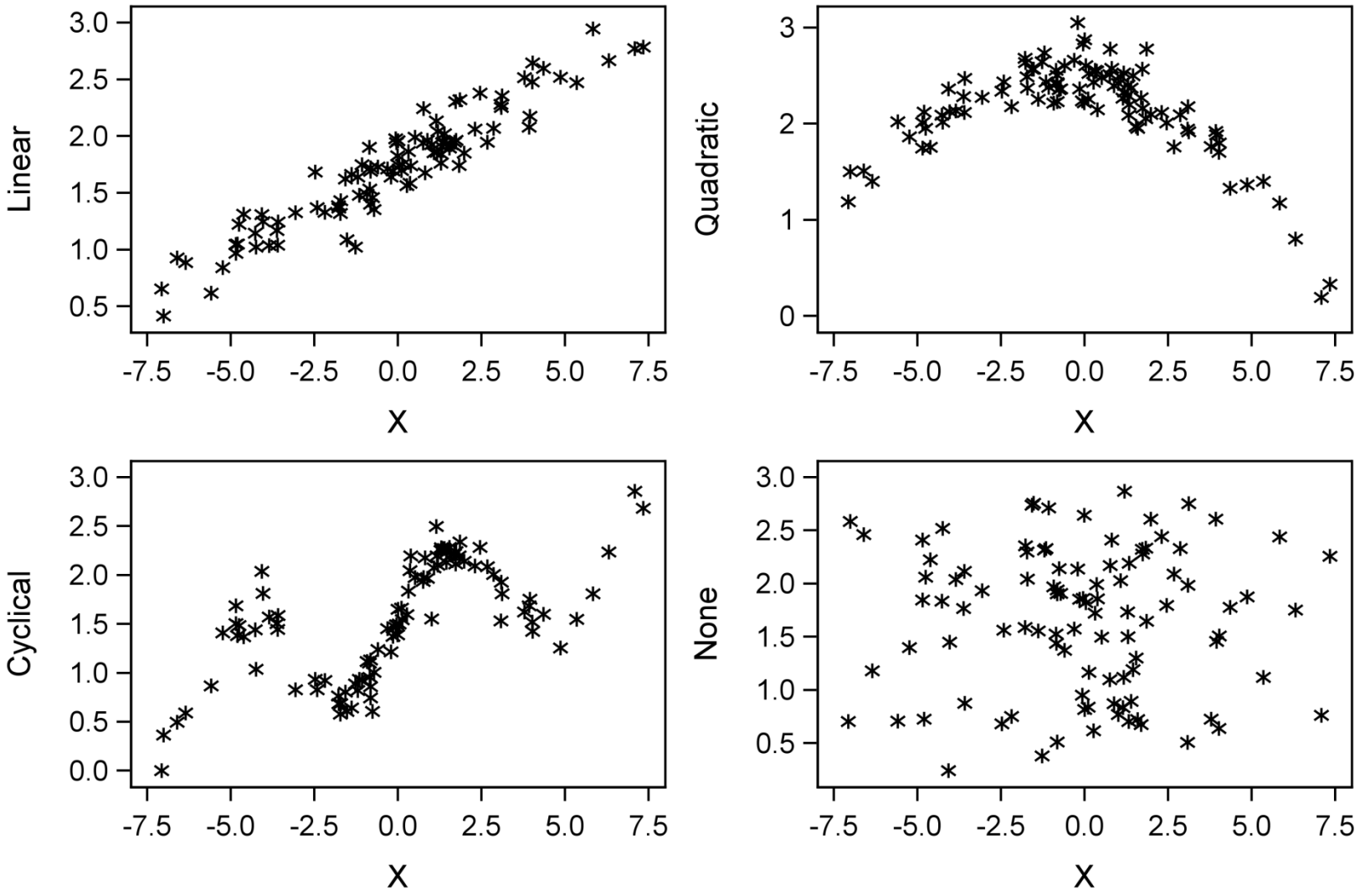
- **Association:** The expected value of one variable differs at different levels of the other variable.

- A ***linear association*** between two continuous variables can be inferred when the general shape of a scatter plot of the two variables is a straight line.

Scatter Plot – Linear Association



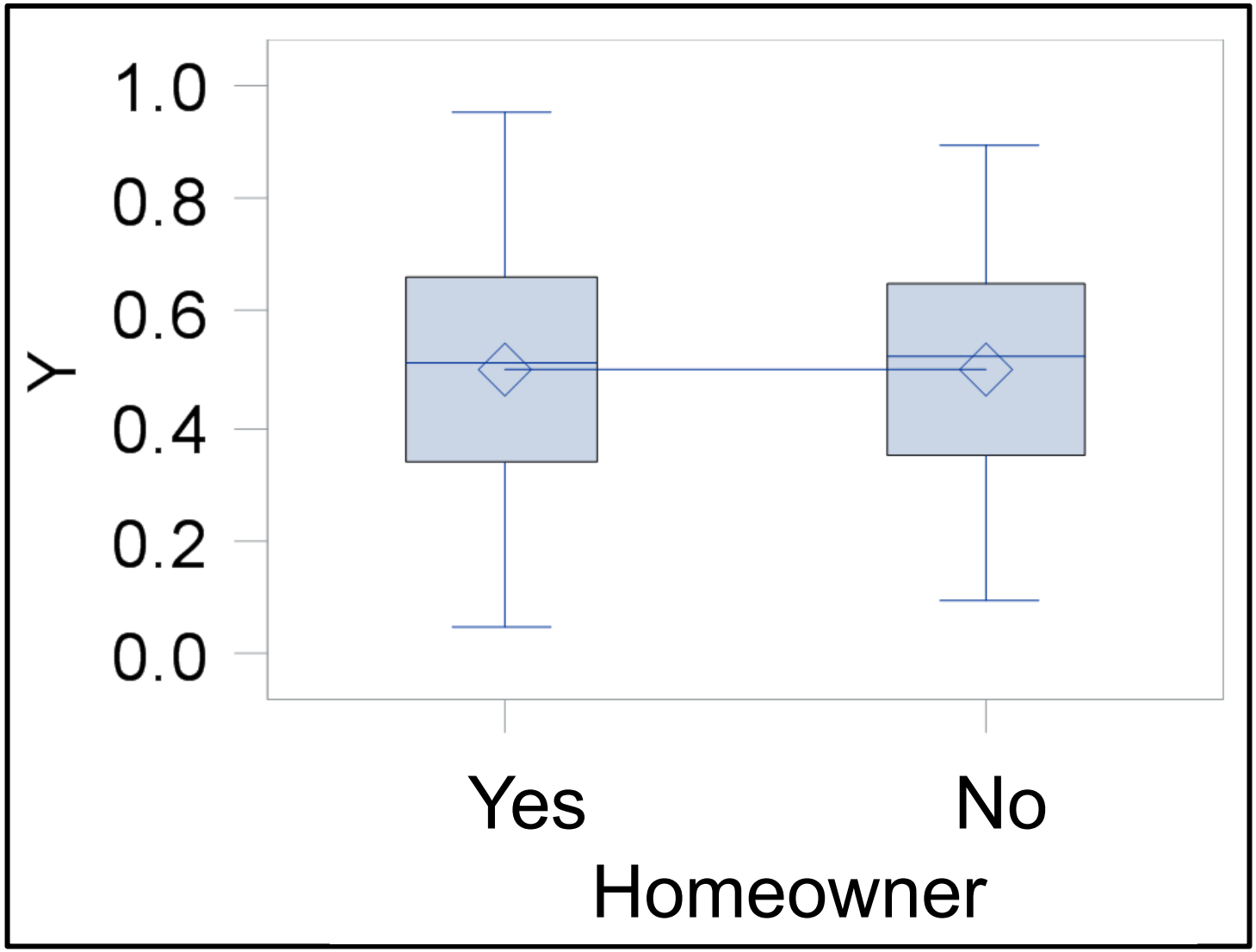
Relationships between Variables



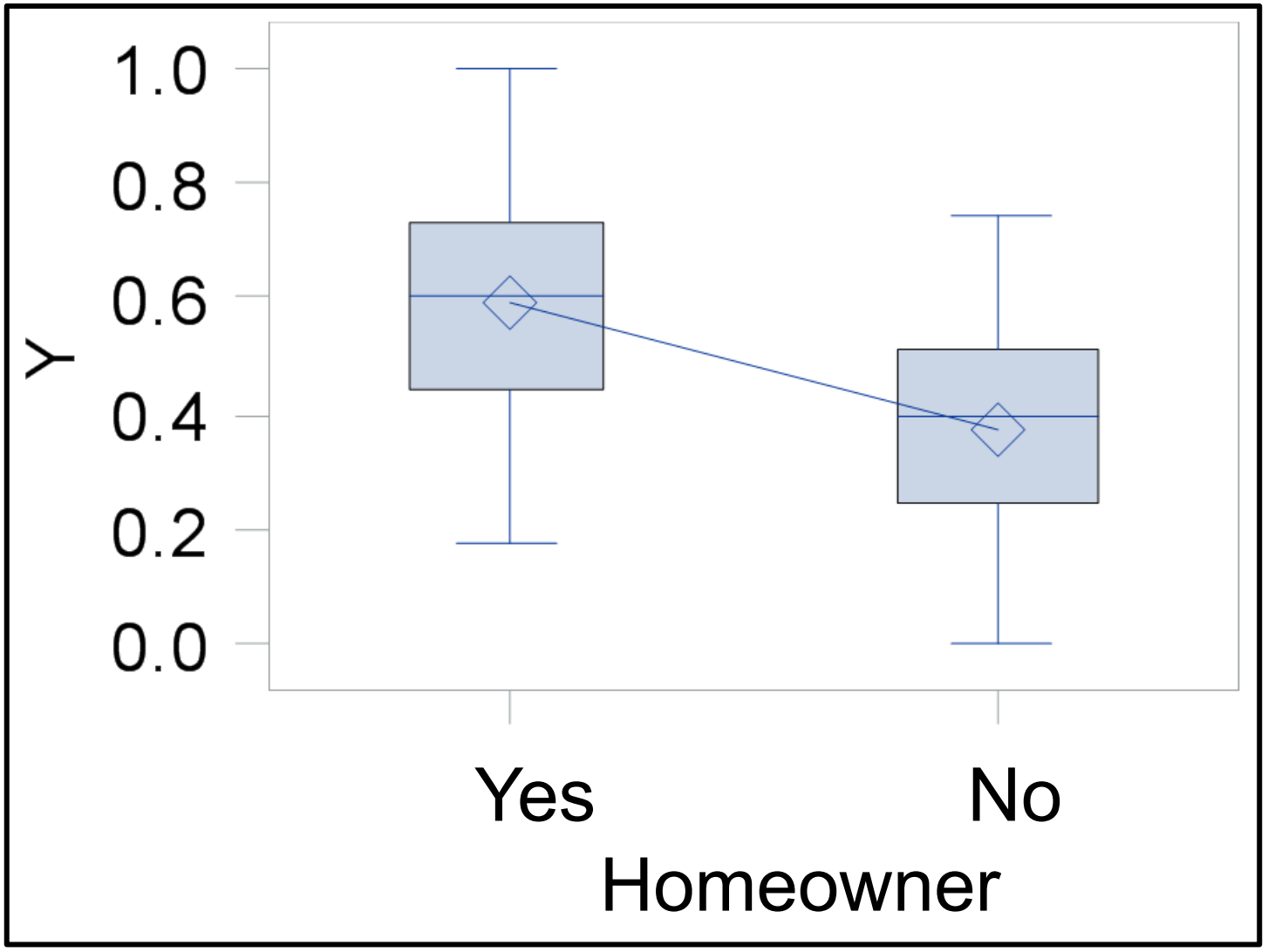
Scatter Plots of Continuous Variables

```
proc sgscatter data=bootcamp.ameshousing3;
     plot SalePrice*Gr_Liv_Area / reg;
     title "Associations of Above Grade Living"
           " Area with Sale Price";
run;
```

No Association – Categorical Predictor



Association – Categorical Predictor



Box Plots of Categorical Variables

```
proc sgplot data=bootcamp.ameshousing3;
    vbox SalePrice / category=Central_Air
                     connect=mean;
    title "Sale Price Differences across Central Air";
run;
```

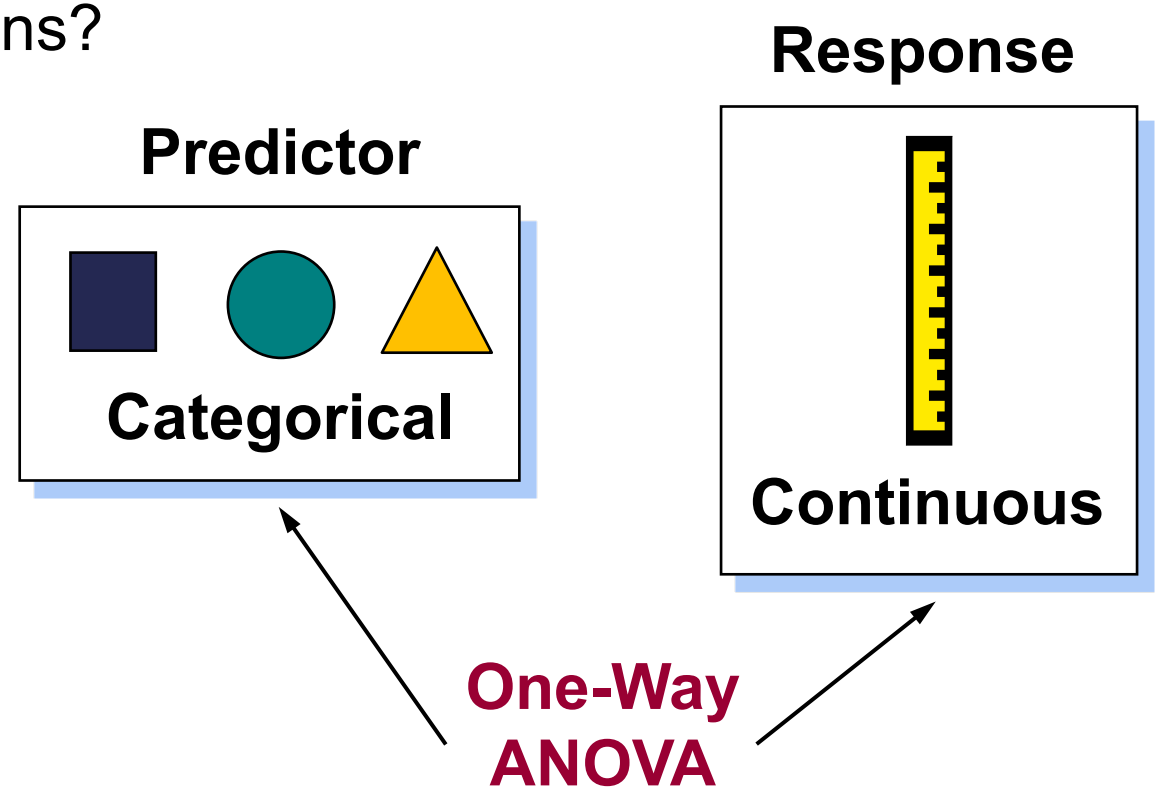
ONE-WAY ANOVA

Overview of Statistical Models

Type of Response \ Type of Predictors	Categorical	Continuous	Continuous and Categorical
Continuous	**Analysis of Variance (ANOVA)**	Ordinary Least Squares (OLS) Regression	Analysis of Covariance (ANCOVA)
Categorical	Contingency Table Analysis or Logistic Regression	Logistic Regression	Logistic Regression

Overview

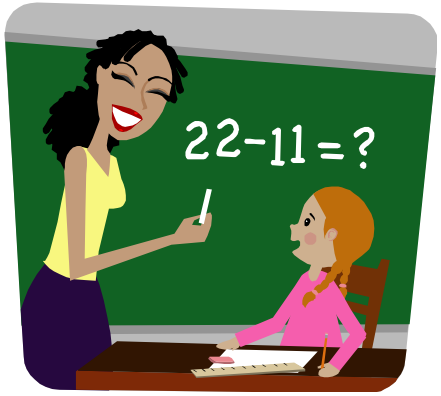
- Are there any differences among the population means?



Another way of asking: Does information about group membership help predict the level of a numeric response?

Research Questions for One-Way ANOVA

- Do accountants, on average, earn more than teachers? *



- *** Is this a case for a *t* test?**

Research Questions for One-Way ANOVA

- Do people treated with one of two new drugs have higher average T-cell counts than people in the control group?

Placebo

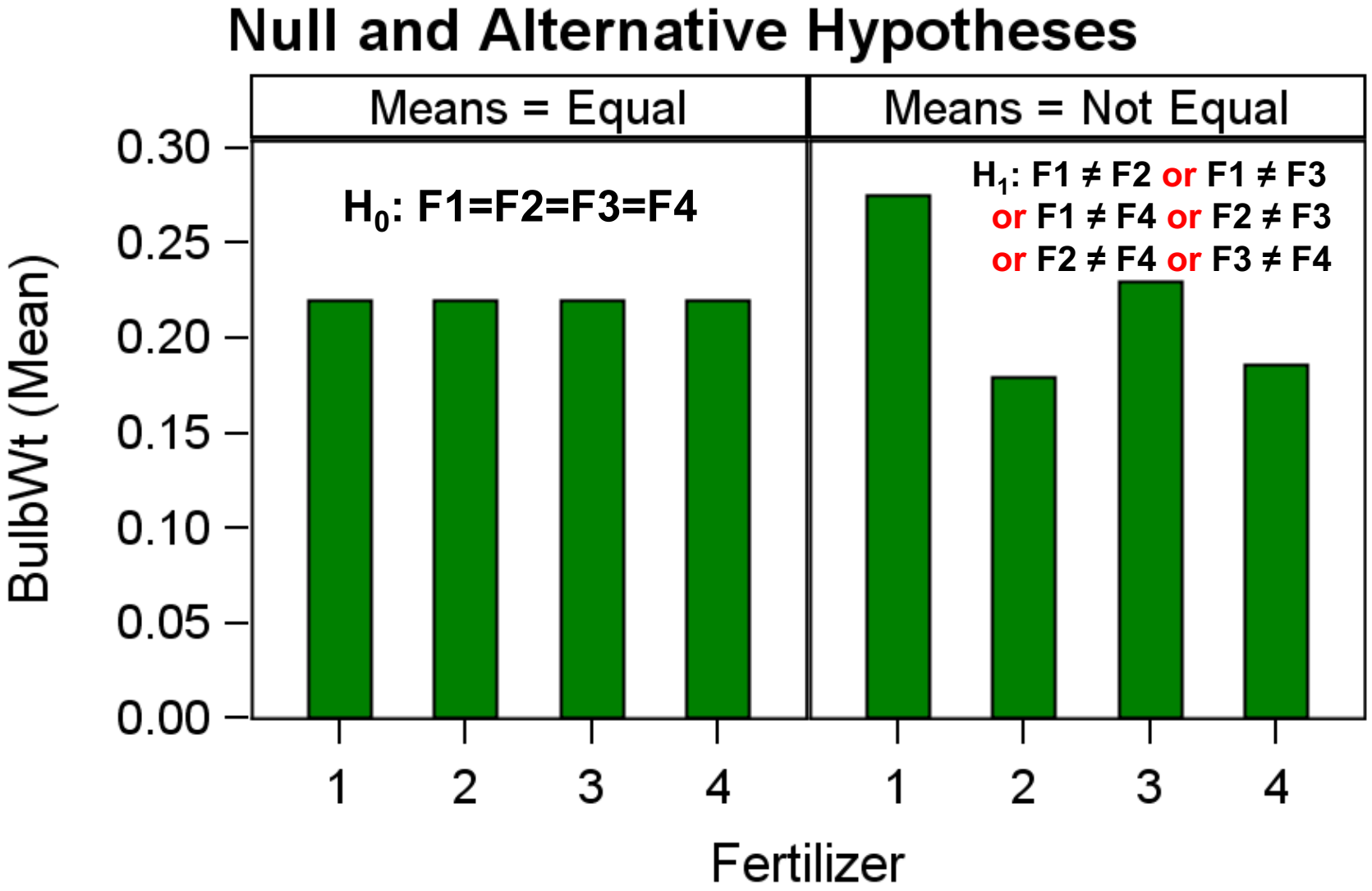
Treatment 1

Treatment 2

Research Questions for One-Way ANOVA

- Do people spend different amounts depending on which type of credit card they have?

The ANOVA Hypothesis

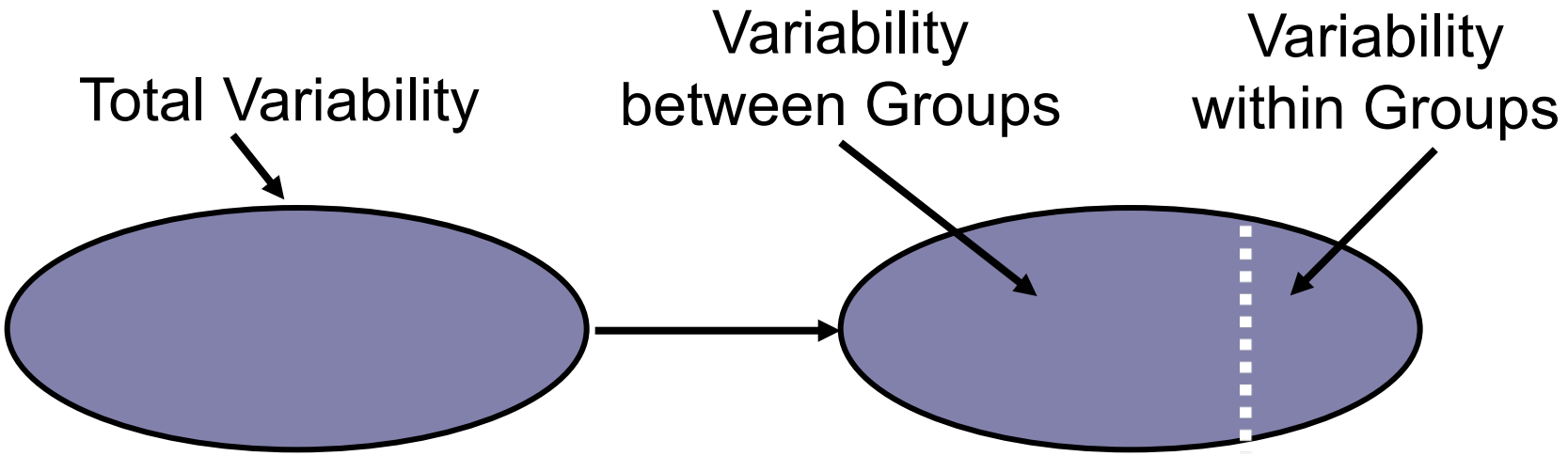


Descriptive Statistics Across Groups

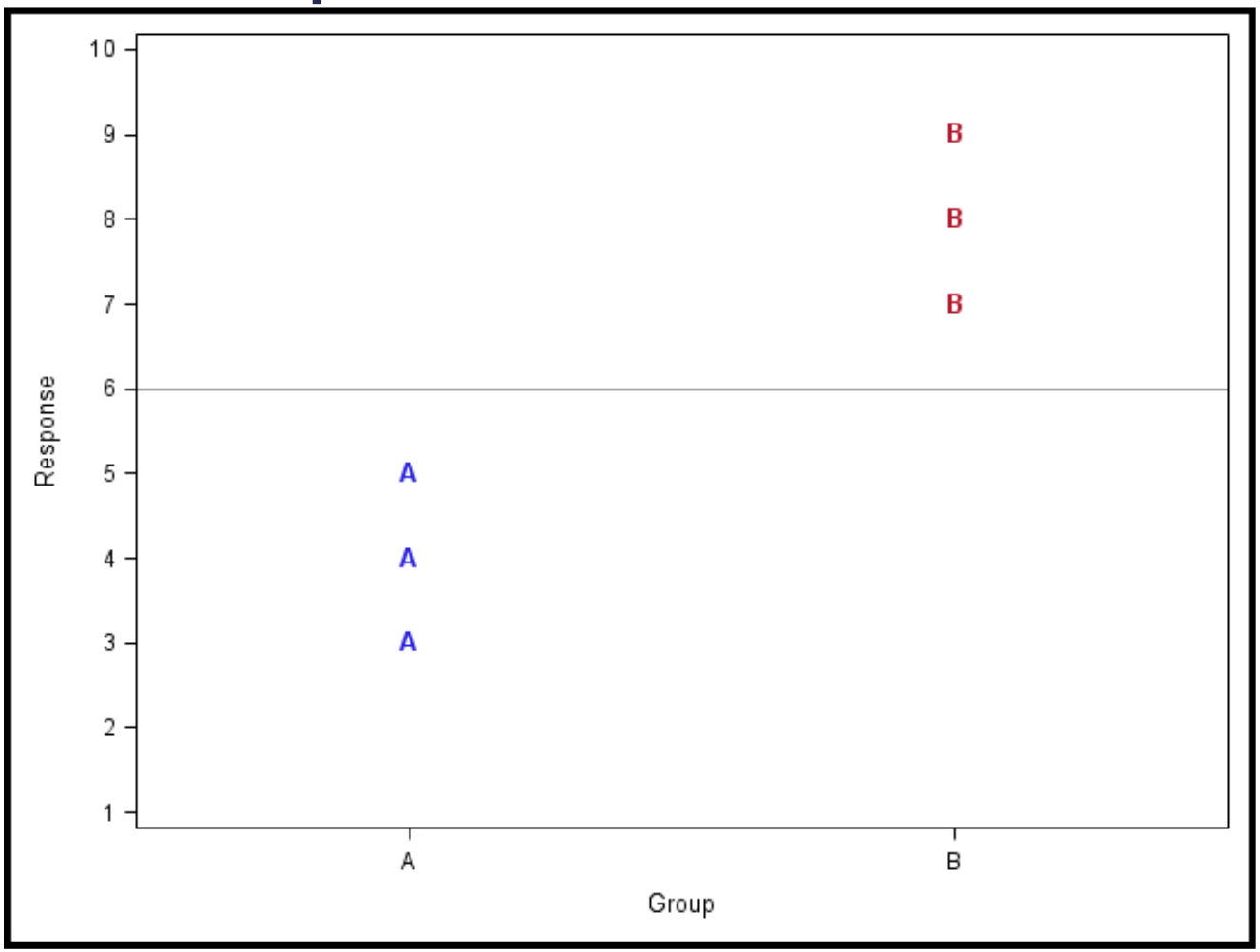
```
proc means data=bootcamp.ameshousing3 printalltypes
maxdec=3;
    var SalePrice;
    class Heating_QC;
    title 'Descriptive Statistics of Sales Price';
run;

proc sgplot data=bootcamp.ameshousing3;
    vbox SalePrice / category=Heating_QC
                     connect=mean;
    title "Sale Price Differences across Heating";
run;
```

Partitioning Variability in ANOVA

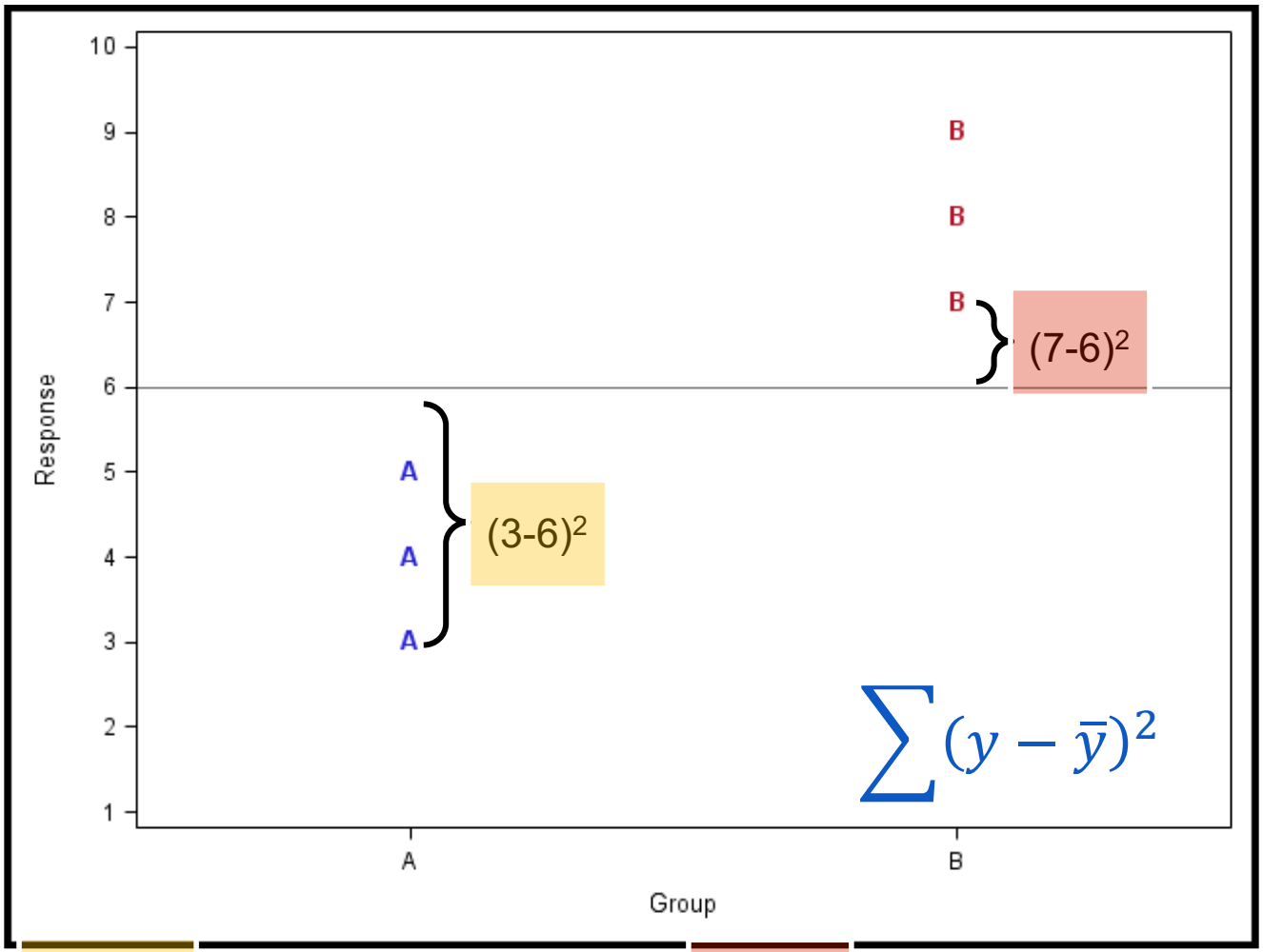


Sums of Squares



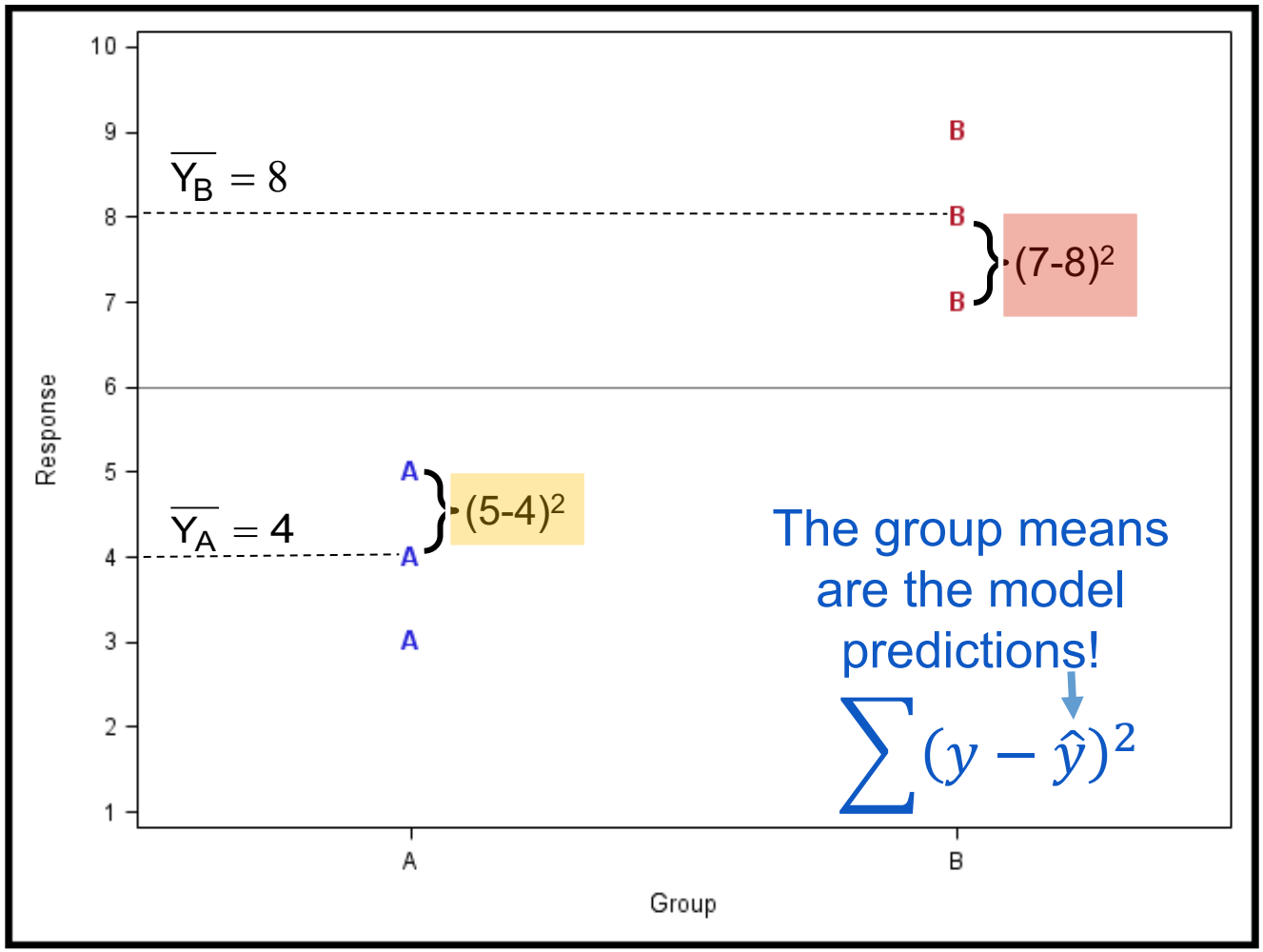
$$\text{Overall mean} = \bar{\bar{y}} = \frac{3+4+5+7+8+9}{6} = \mathbf{6}$$

Total Sum of Squares



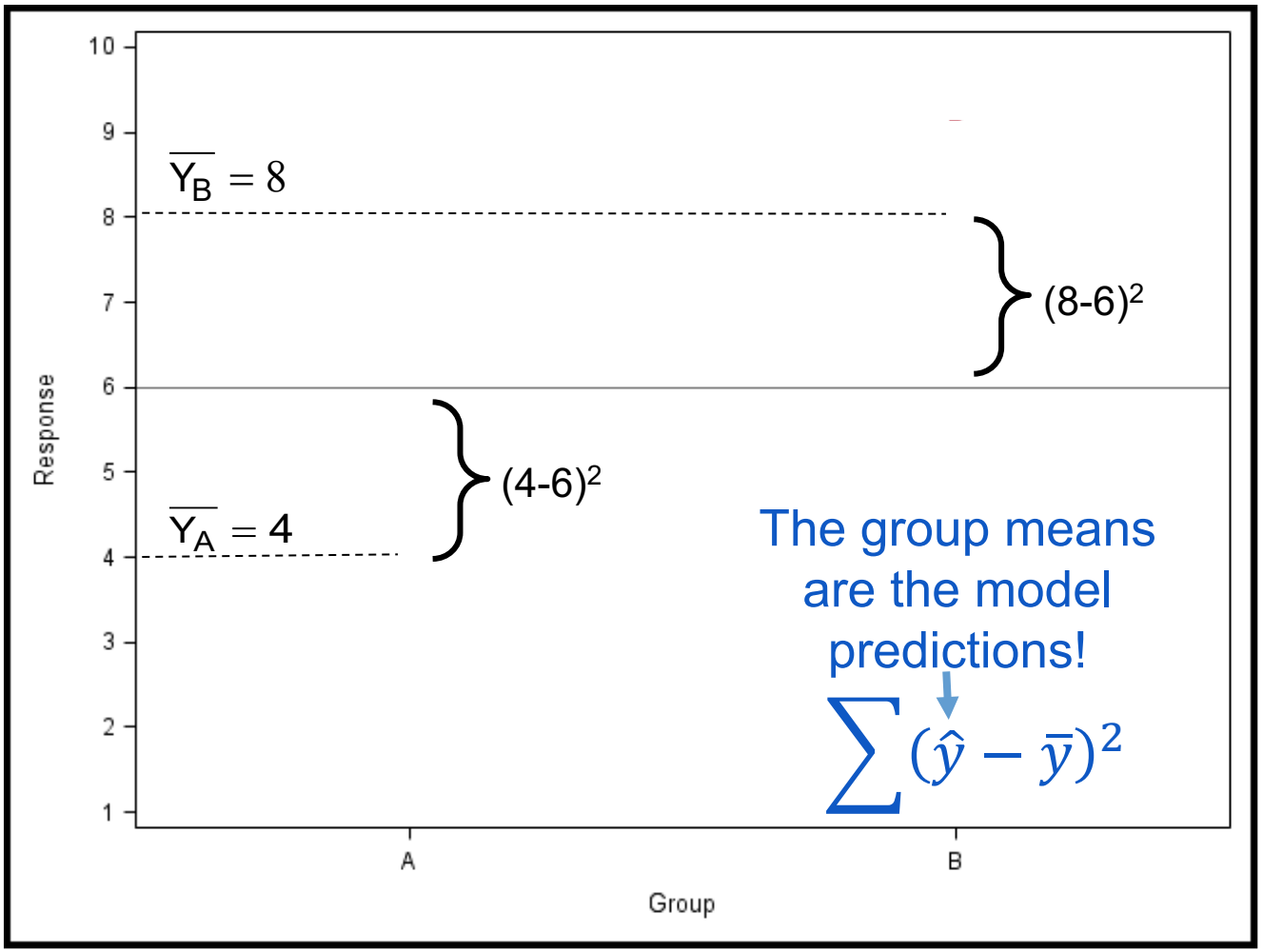
SS_T = (3-6)² + (4-6)² + (5-6)² + (7-6)² + (8-6)² + (9-6)² = **28**

Error Sum of Squares



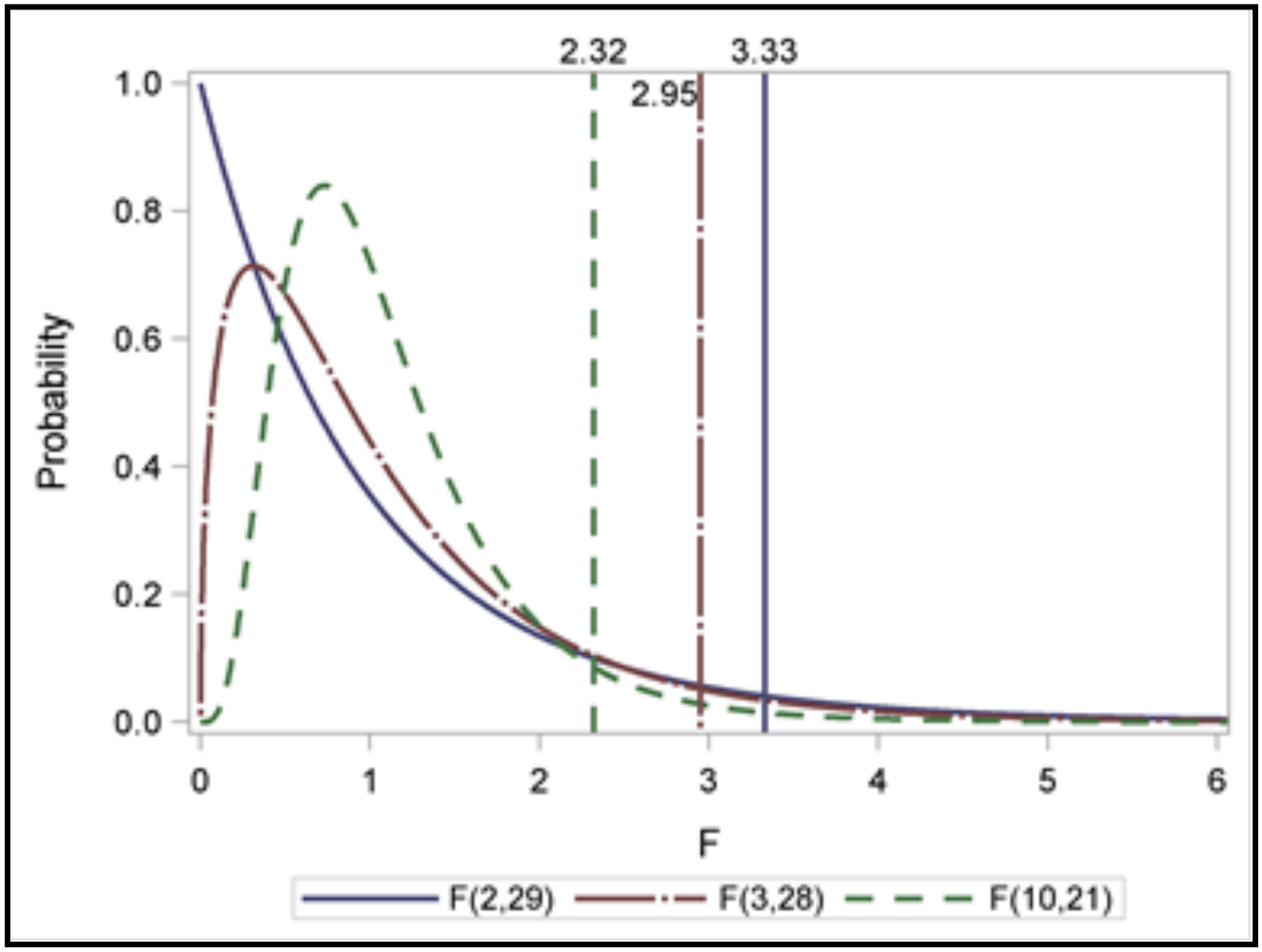
$\mathbf{SS_E} = (3-4)^2 + (4-4)^2 + (5-4)^2 + (7-8)^2 + (8-8)^2 + (9-8)^2 = \mathbf{4}$

Model Sum of Squares



$\mathbf{SS_M}$ = 3*(4-6)² + 3*(8-6)² = **24**

F Statistic and Critical Values at α=0.05



F(Model df, Error df)=MS_M / MS_E

Coefficient of Determination

$$R^2 = \frac{SS_M}{SS_T}$$

“Proportion of variance accounted for by the model”

The ANOVA Model

BulbWt = Base Level + Fertilizer + Unaccounted for Variation

$$Y_{ik} = \mu + \tau_i + \varepsilon_{ik}$$

One-Way ANOVA

```
proc glm data=bootcamp.ameshousing3;
     class Heating_QC;
     model SalePrice=Heating_QC;
     format Heating_QC $Heating_QC.;
     title "One-Way ANOVA with Heating Quality"
           " as Predictor";
run;
quit;
```

What Does a CLASS Statement Actually Do?

- The **CLASS statement creates dummy variables** for the levels of categorical variables.
- PROC GLM performs linear regression on the dummy variables, but reports the output in a manner interpretable as group mean differences.
- There is only one “parameterization” available in PROC GLM.

GLM Default Coding of CLASS Variables

			Design Variables		
CLASS	Value	Label	1	2	3
IncLevel	1	Low Income	1	0	0
	2	Medium Income	0	1	0
	3	High Income	0	0	1

Assumptions for ANOVA

- Observations are independent.
- Errors are normally distributed.
- All groups have equal error variances.

Assessing ANOVA Assumptions

- **Good data collection** designs help ensure the **independence** assumption.

- **Diagnostic plots** from PROC GLM can be used to verify the assumption that the error is approximately **normally distributed.**

- PROC GLM produces **a test of equal variances** with the HOVTEST option in the MEANS statement.
  H0 for this hypothesis test is that the variances are equal for all populations.

One-Way ANOVA – Assumptions

```
proc glm data=bootcamp.ameshousing3;
     class Heating_QC;
     model SalePrice=Heating_QC;
     means Heating_QC / hovtest=levene;
     format Heating_QC $Heating_QC.;
     title "One-Way ANOVA Equal Variance Test";
run;
quit;
```

Predicted and Residual Values

- **The predicted value in ANOVA is the *group mean*.**

- The ***residual*** is the difference between the observed value of the response and the predicted value of the response variable.

Observation		Heating_QC	Observed	Predicted	Residual
1		Ex	213500.0000	154919.1869	58580.8131
2		Ex	191500.0000	154919.1869	36580.8131
3		TA	115000.0000	130573.5294	-15573.5294
4		Ex	160000.0000	154919.1869	5080.8131
5		Ex	180000.0000	154919.1869	25080.8131
6		TA	125000.0000	130573.5294	-5573.5294
7		TA	206000.0000	130573.5294	75426.4706
8		Gd	159000.0000	130844.0862	28155.9138
9		TA	180500.0000	130573.5294	49926.4706
10		Gd	142125.0000	130844.0862	11280.9138

Analysis Plan for ANOVA – Summary

- Null Hypothesis: All means are equal.
- Alternative Hypothesis: At least one mean is different.

1. Produce descriptive statistics.
2. Verify assumptions.
   - Independence
   - Errors are normally distributed.
   - Error variances are equal for all groups.
3. Examine the *p*-value for overall F-Test in the ANOVA table. If the *p*-value is less than alpha, reject the null hypothesis.

Poll

Quiz

ANOVA POST-HOC TESTS

Poll

Quiz

Multiple Comparison Methods

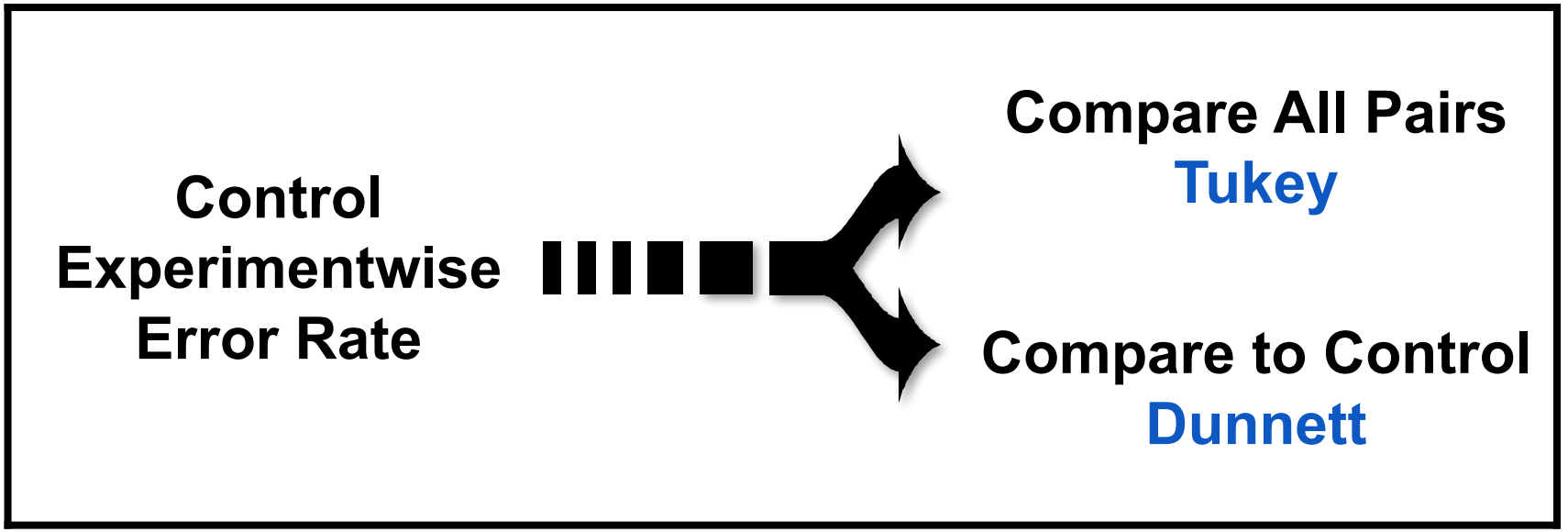
Number of Groups Compared	Number of Comparisons	Experimentwise Error Rate (α=0.05)
2	1	.05
3	3	.14
4	6	.26
5	10	.40

Comparisonwise Error Rate = α = 0.05

Experimentwise Error Rate $\leq 1 - (1 - \alpha)^{nc}$

where *nc*=number of comparisons

Multiple Comparison Methods



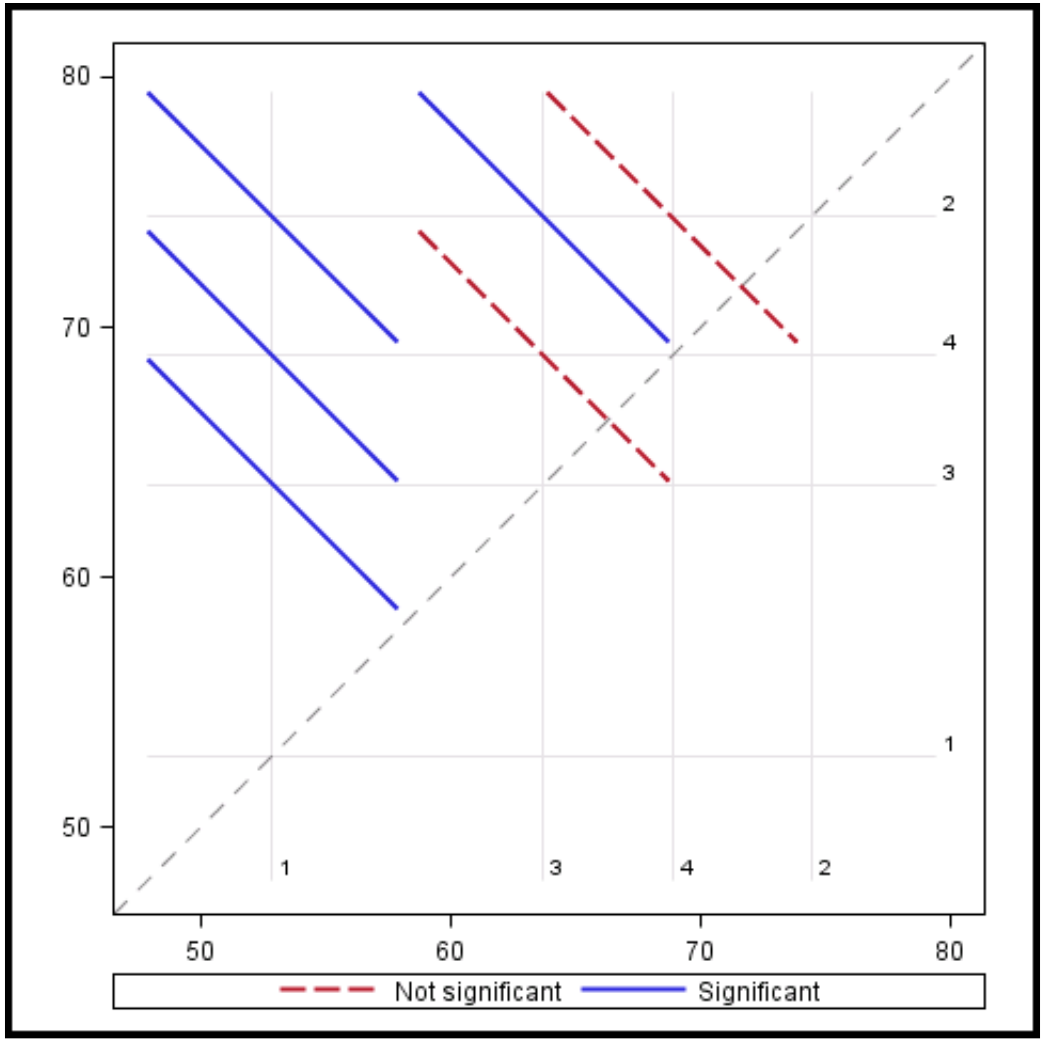
Tukey's Honest Significant Difference Test

- Appropriate when you consider many pairwise comparisons.

- The experimentwise error rate is:
  - equal to α when ***all*** pairwise comparisons are considered
  - less than α when ***fewer*** than all pairwise comparisons are considered.

Also known as the Tukey-Kramer Test

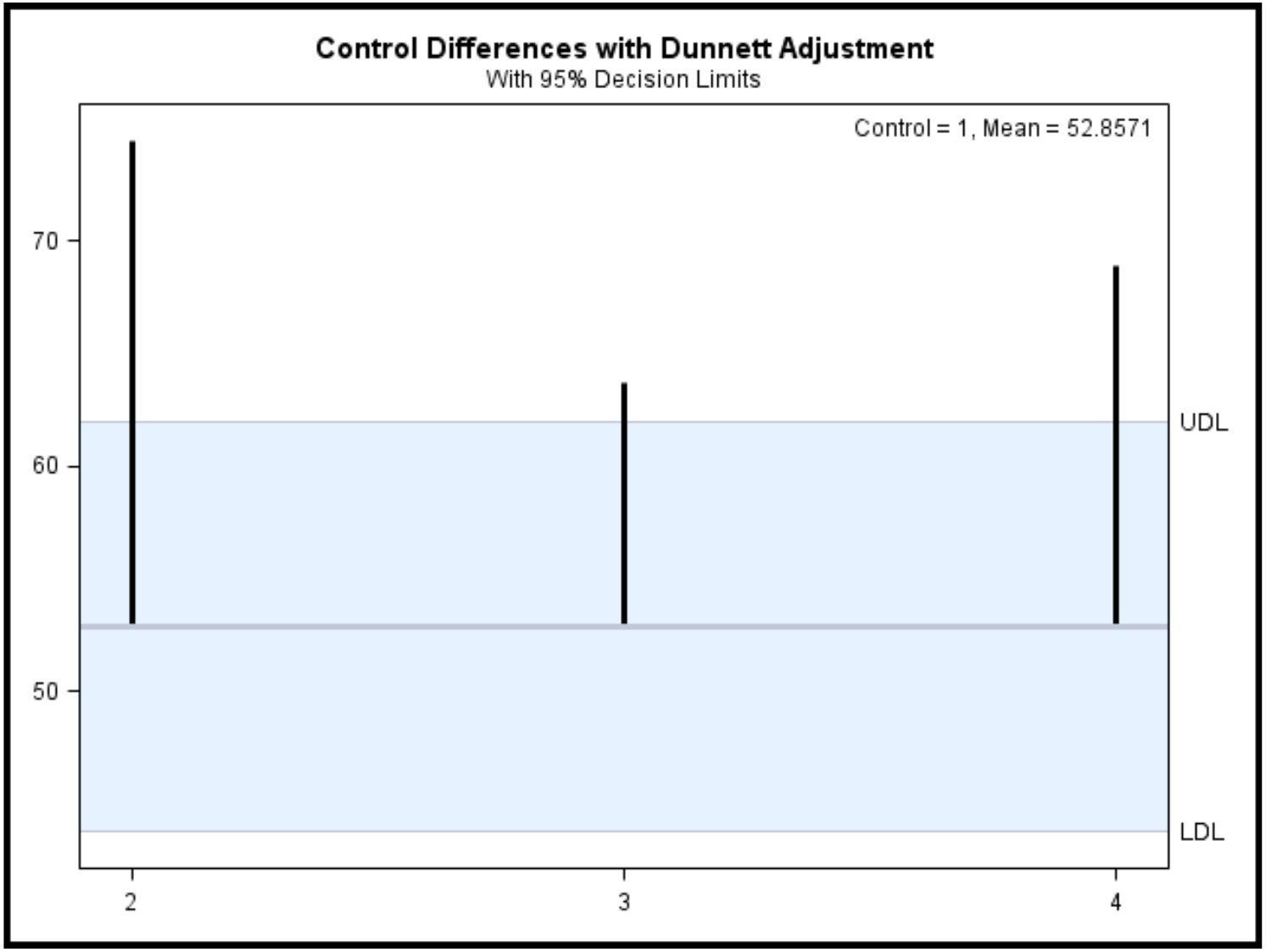
Diffograms



Special Case: Comparing to a Control

- Comparing to a control is appropriate when there is a natural reference group, such as a placebo group in a drug trial.
  - **Control comparison computes and tests *k*-1 differences**, where *k* is the number of levels of the CLASS variable (rather than all pairwise).
  - One-sided hypothesis tests against a control group can be performed.
  - Comparing to a control takes into account the correlations among tests.
  - An example is the **Dunnett method**.

Control Plots



Post Hoc Pairwise Comparisons

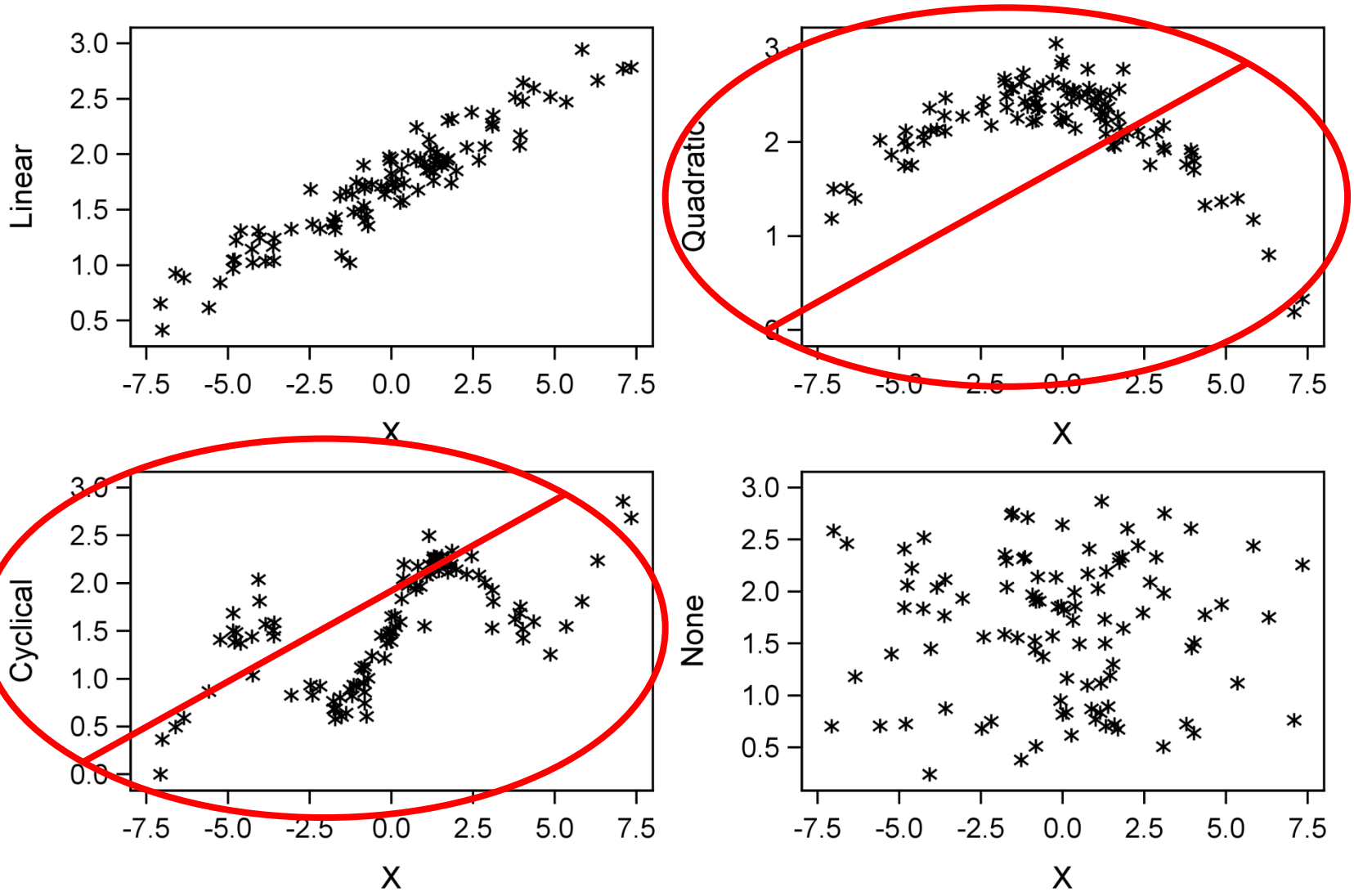
```
proc glm data=bootcamp.ameshousing3
         plots(only)=(diffplot(center) controlplot);
    class Heating_QC;
    model SalePrice=Heating_QC;
    lsmeans Heating_QC / pdiff=all
                         adjust=tukey;
    lsmeans Heating_QC / pdiff=control('Average/Typical')
                         adjust=dunnett;
    format Heating_QC $Heating_QC.;
    title "Post-Hoc Analysis of ANOVA - HQ as Predictor";
run;
quit;
```

LAB 3

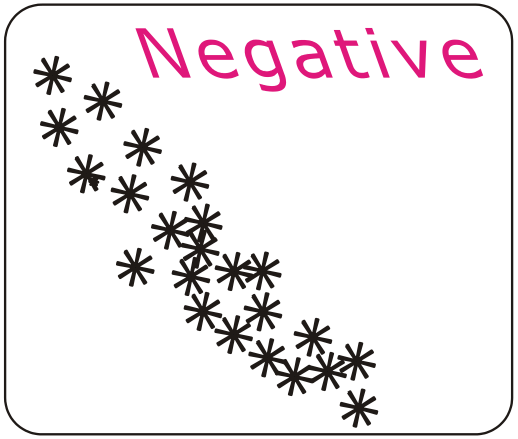
Don’t forget to take the lab check on Moodle!

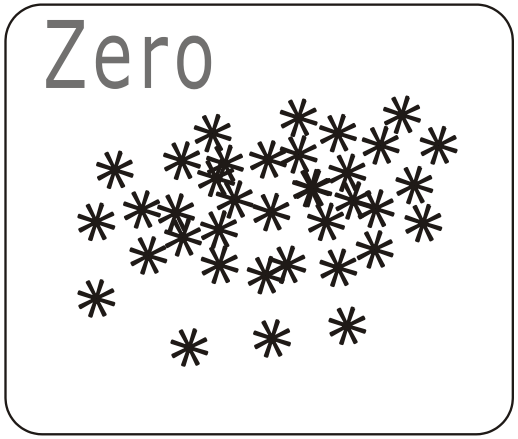
PEARSON CORRELATION

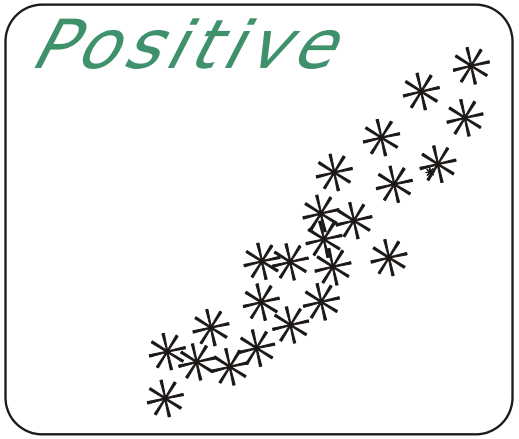
Pearson Correlation – Linear Relationships

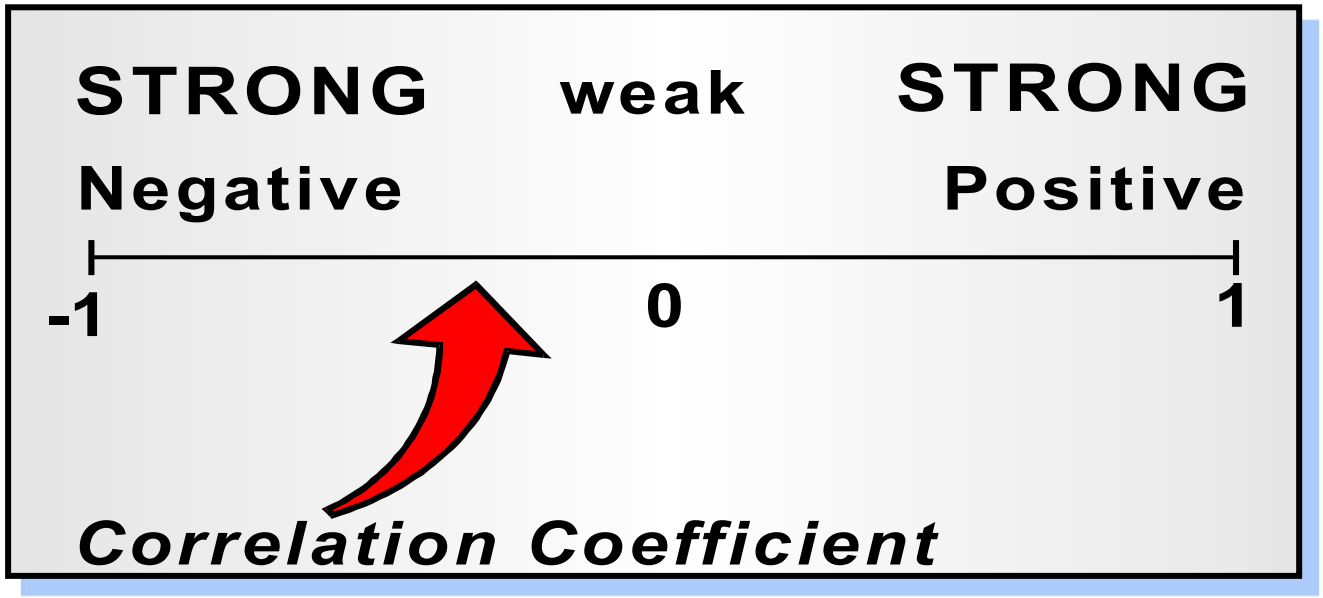


Correlation





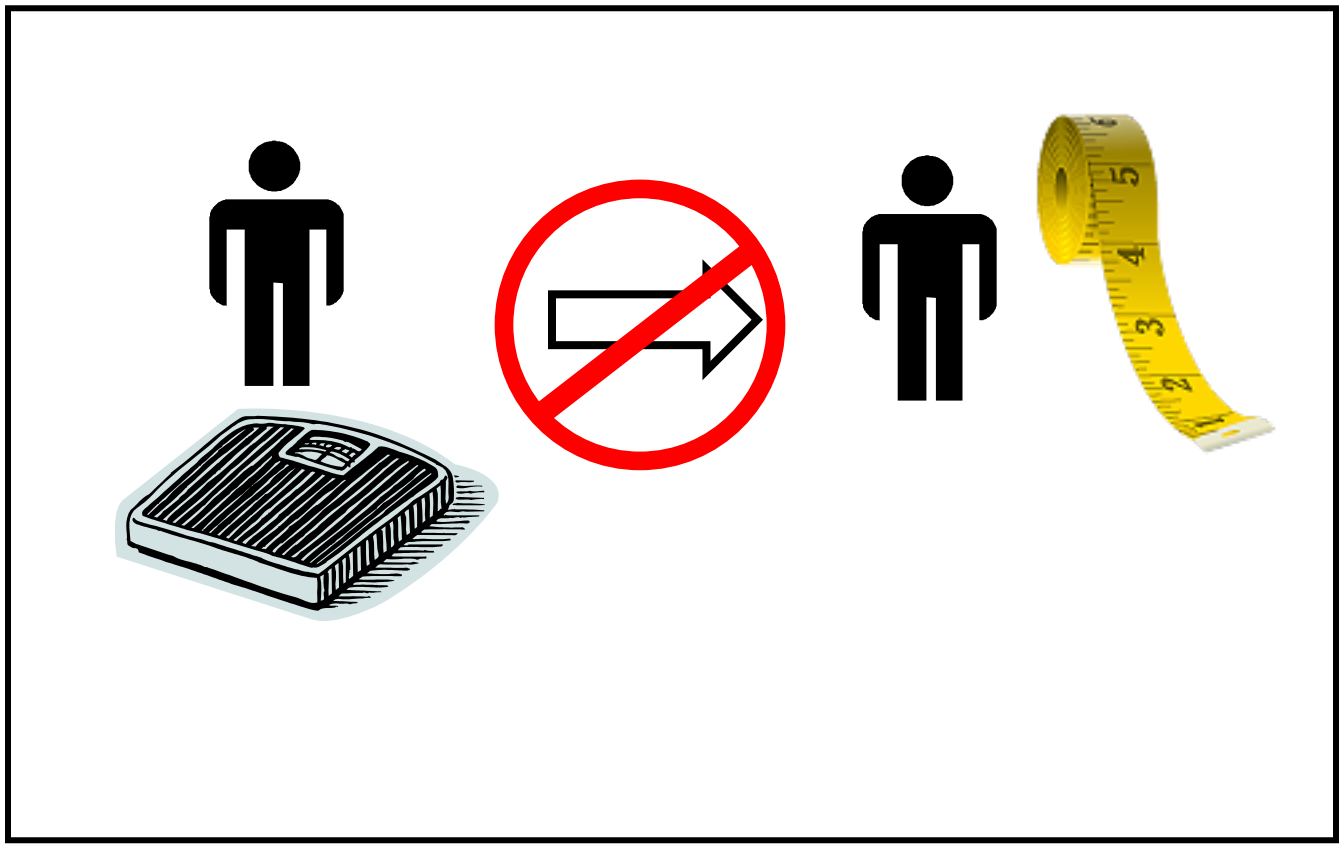




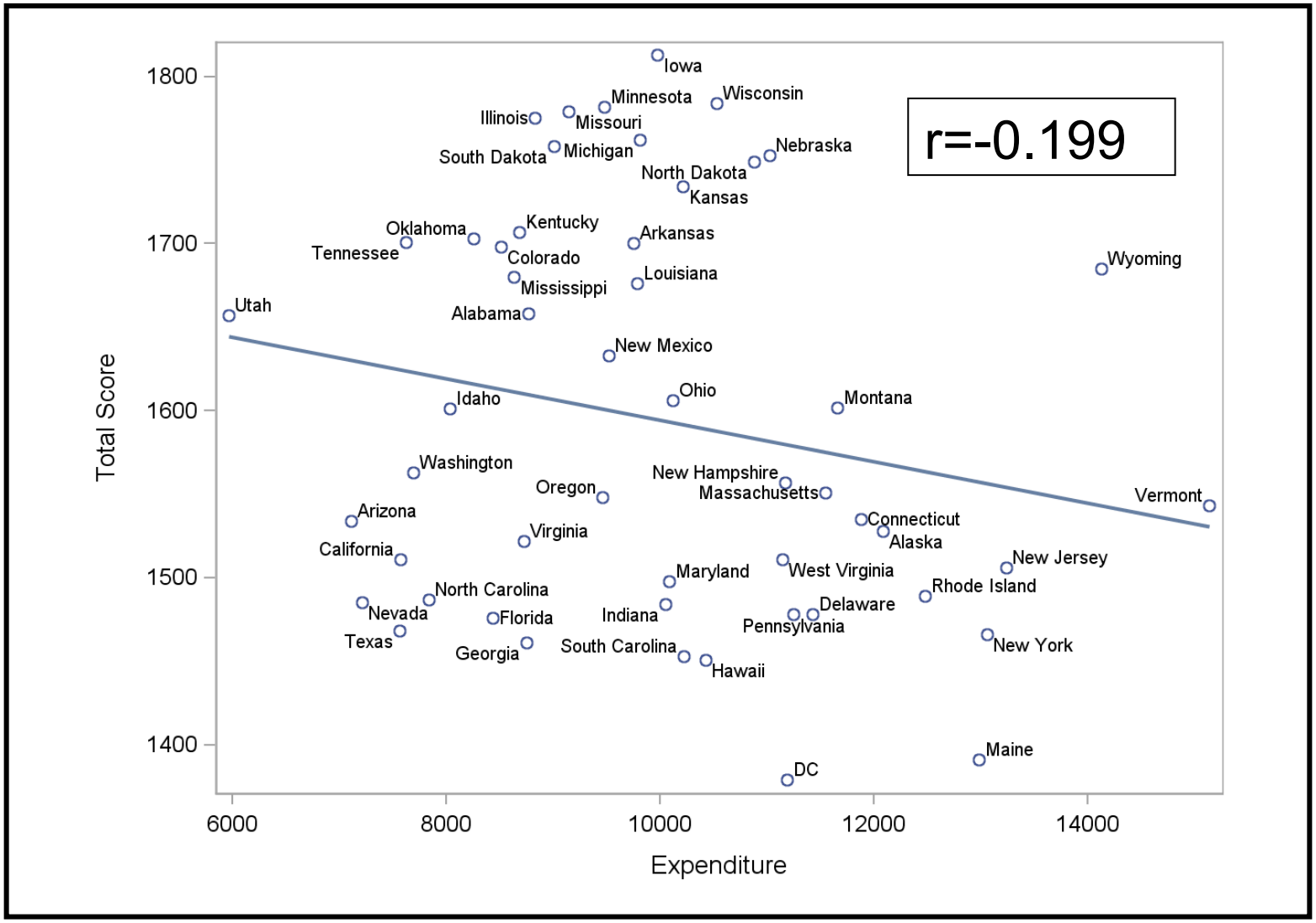
Hypothesis Test for a Correlation

- The parameter representing population correlation is ρ.
- ρ is estimated by the sample statistic *r*.
- H_0: $\rho = 0$
- Rejecting H_0 indicates only great confidence that ρ is not exactly zero.
- A *p*-value **does not** measure the *magnitude* of the association.
- Sample size affects the *p*-value.

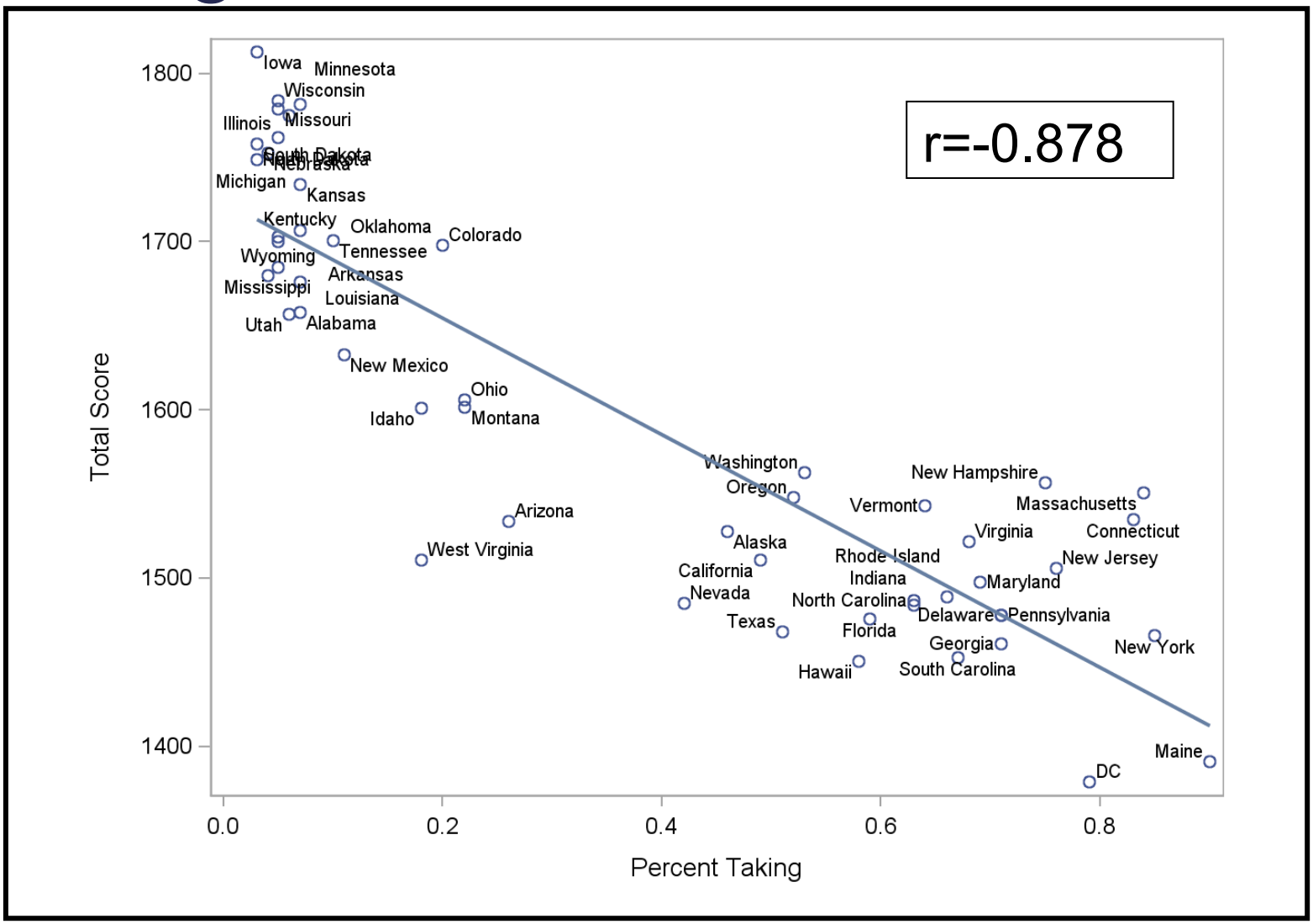
Correlation versus Causation



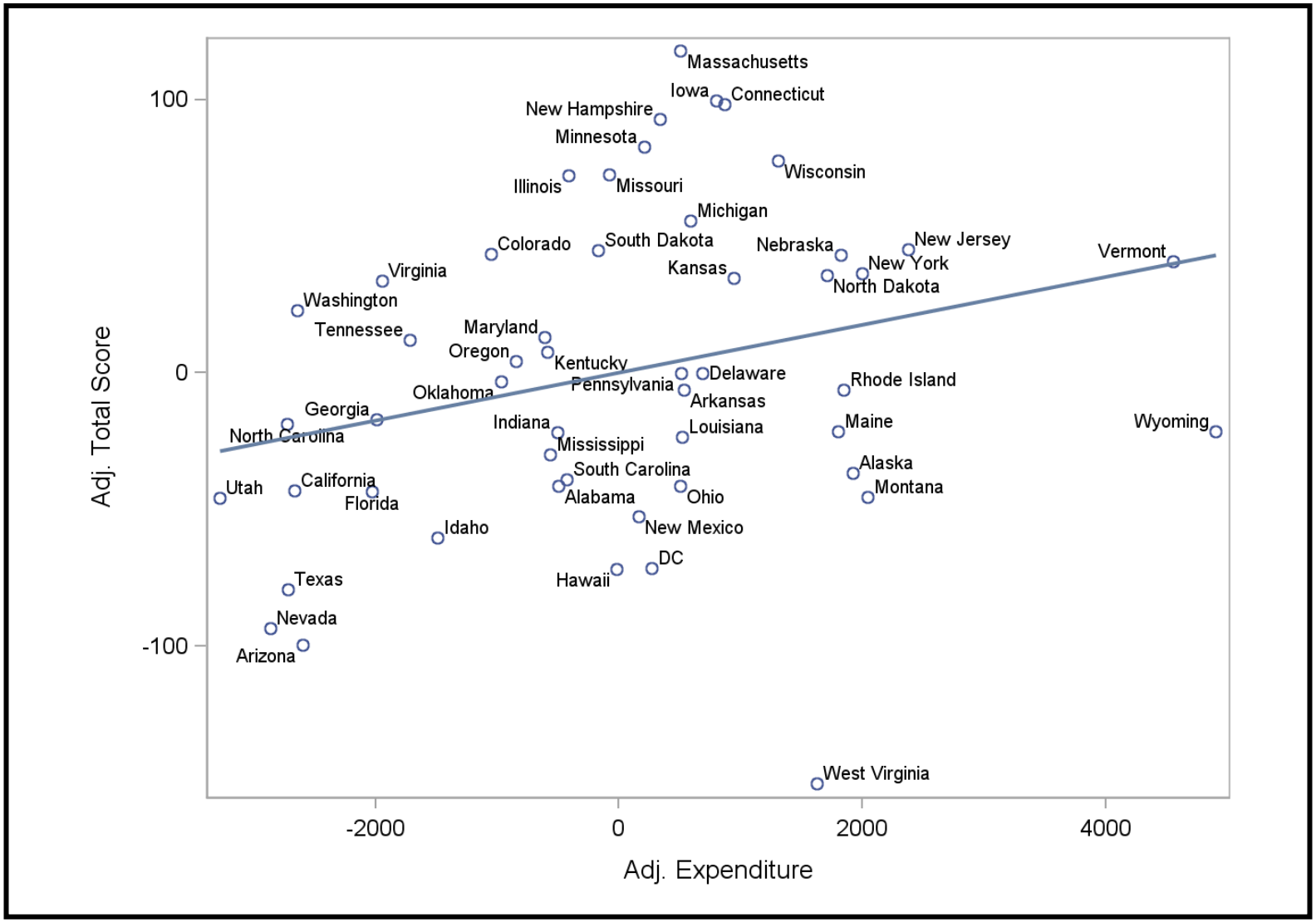
Apparent Relationship



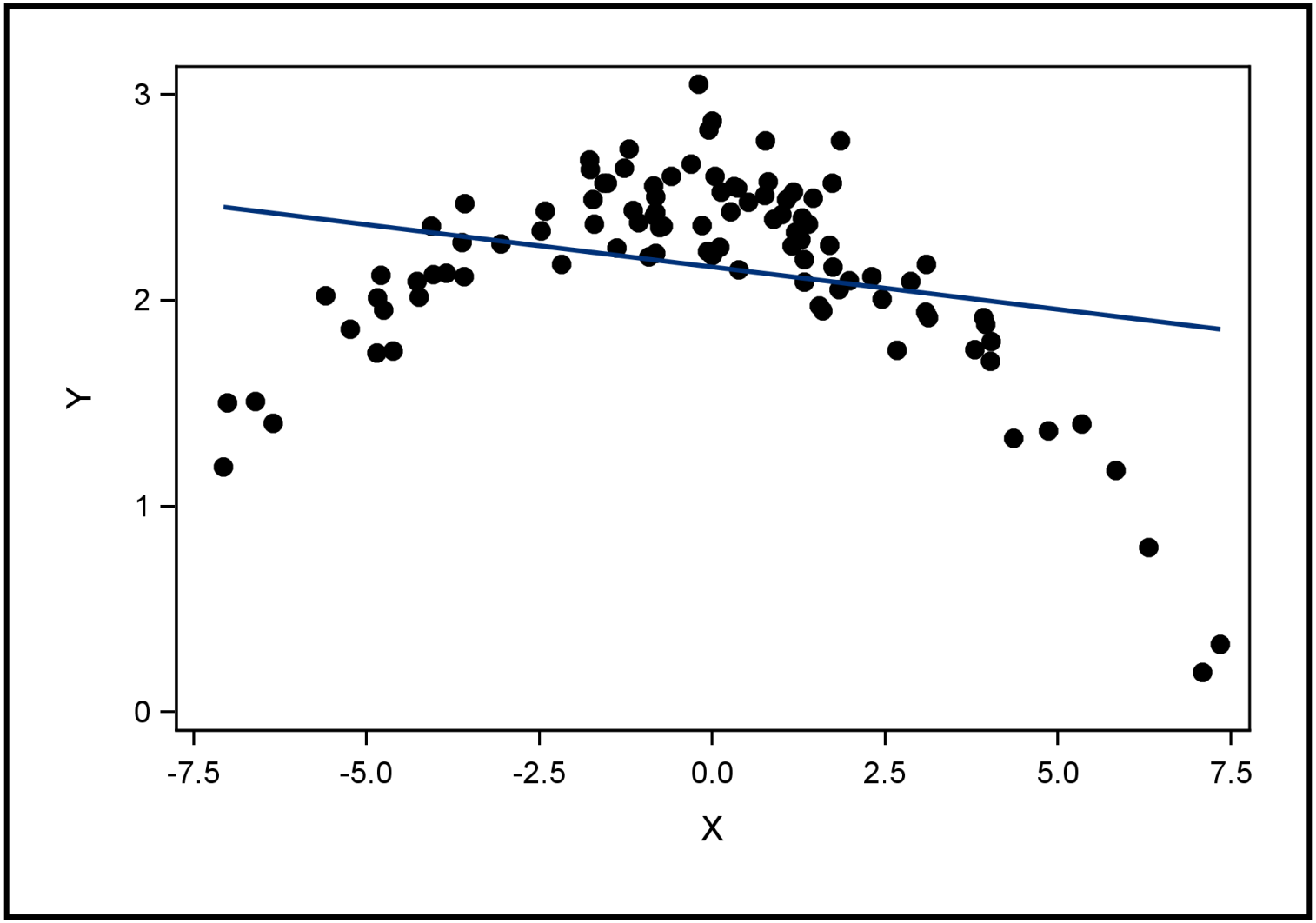
Missing Link



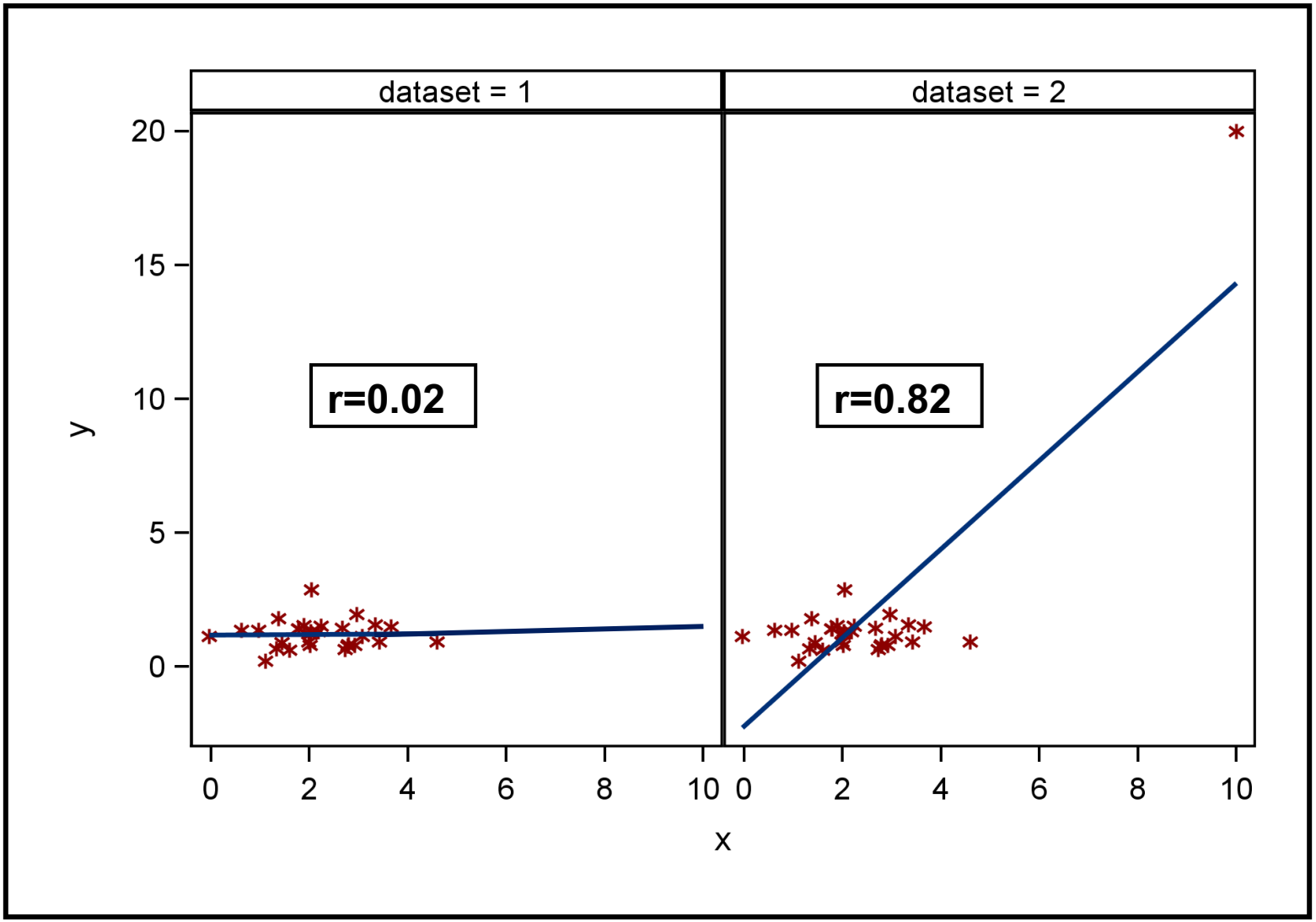
The Truer Story



Missing Another Type of Relationship



Extreme Data Values



Guess the Correlation!

http://guessthecorrelation.com

- guess within 0.05: +1 life and +5 coins
- guess within 0.10: +1 coin
- guess within >0.10: -1 life

Correlations and Plots

```
proc corr data=bootcamp.ameshousing3 rank
          plots(only)=scatter(nvar=all ellipse=none);
   var &interval;
   with SalePrice;
   id PID;
   title "Correlations and Scatter Plots with SalePrice";
run;
```

Correlation Matrix

```
proc corr data=bootcamp.ameshousing3
          nosimple
           plots=matrix(nvar=all histogram);
   var SalePrice Gr_Liv_Area Basement_Area;
   title "Scatter Plot Matrix of Predictors";
run;
```

Poll

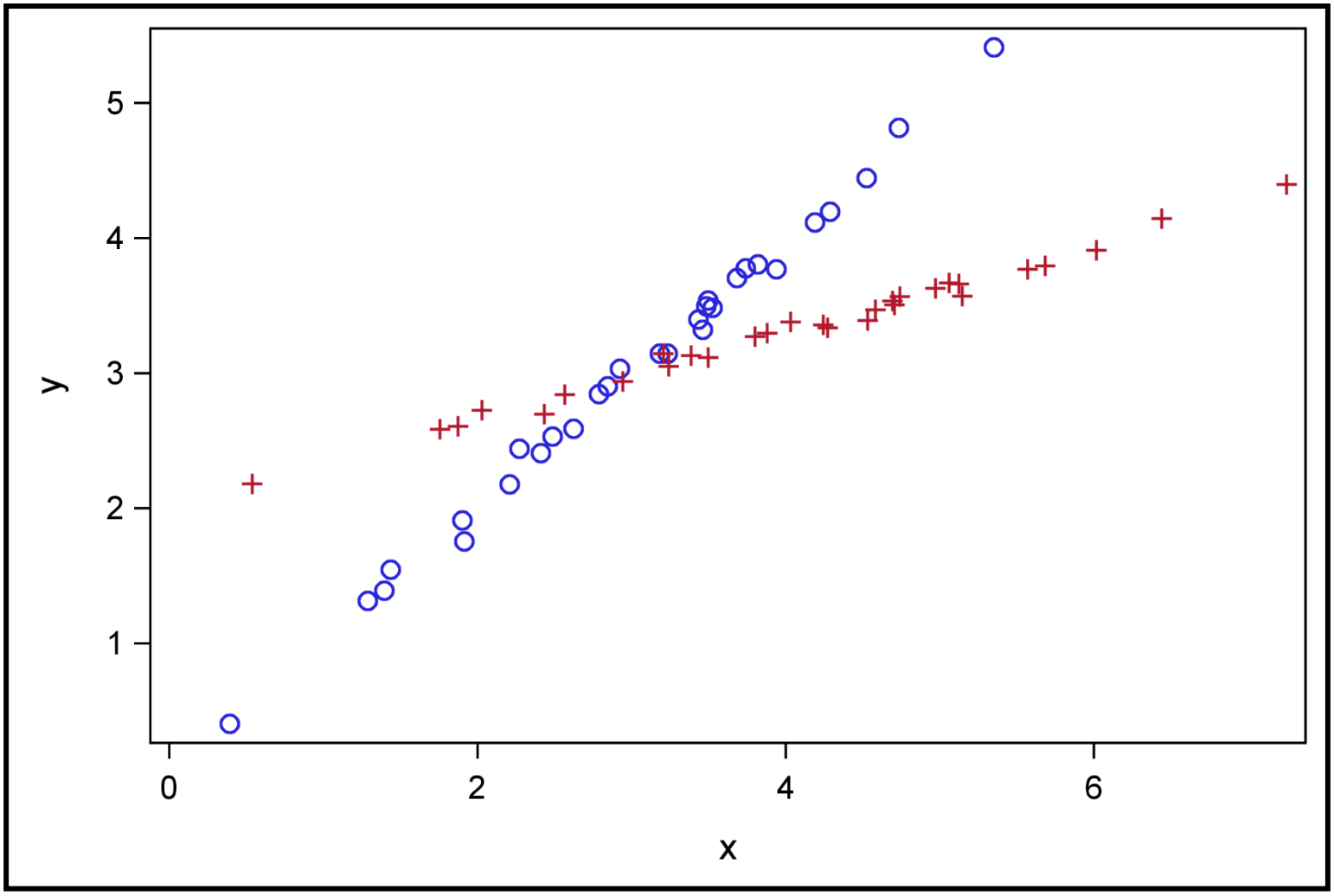
Quiz

SIMPLE LINEAR REGRESSION

Overview of Statistical Models

Type of Predictors / Type of Response	Categorical	Continuous	Continuous and Categorical
Continuous	Analysis of Variance (ANOVA)	Ordinary Least Squares (OLS) Regression	Analysis of Covariance (ANCOVA)
Categorical	Contingency Table Analysis or Logistic Regression	Logistic Regression	Logistic Regression

Overview



Simple Linear Regression Analysis

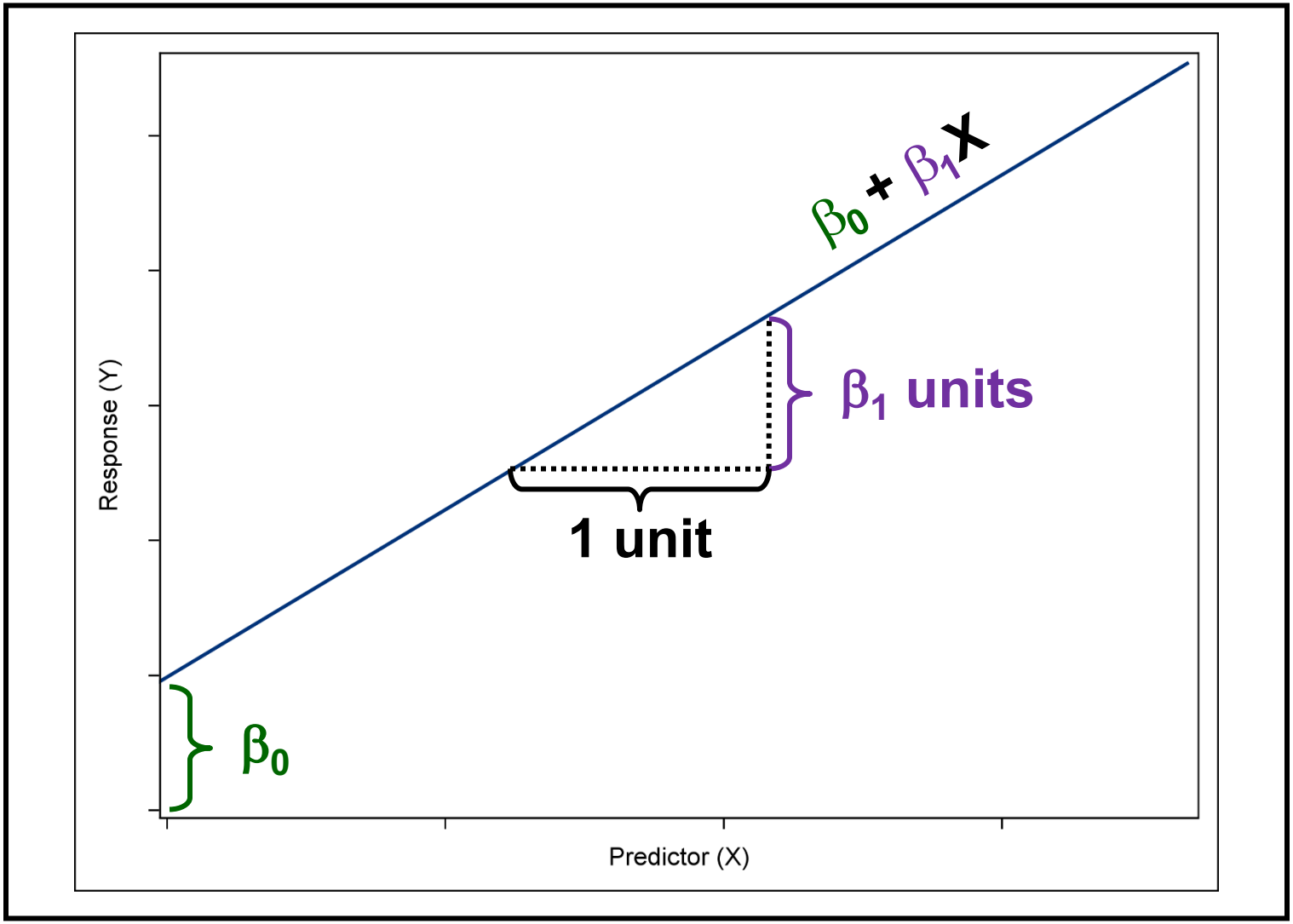
- The objectives of simple linear regression are as follows:

  - **Explain**: assess the significance of the predictor variable in explaining the variability or behavior of the response variable

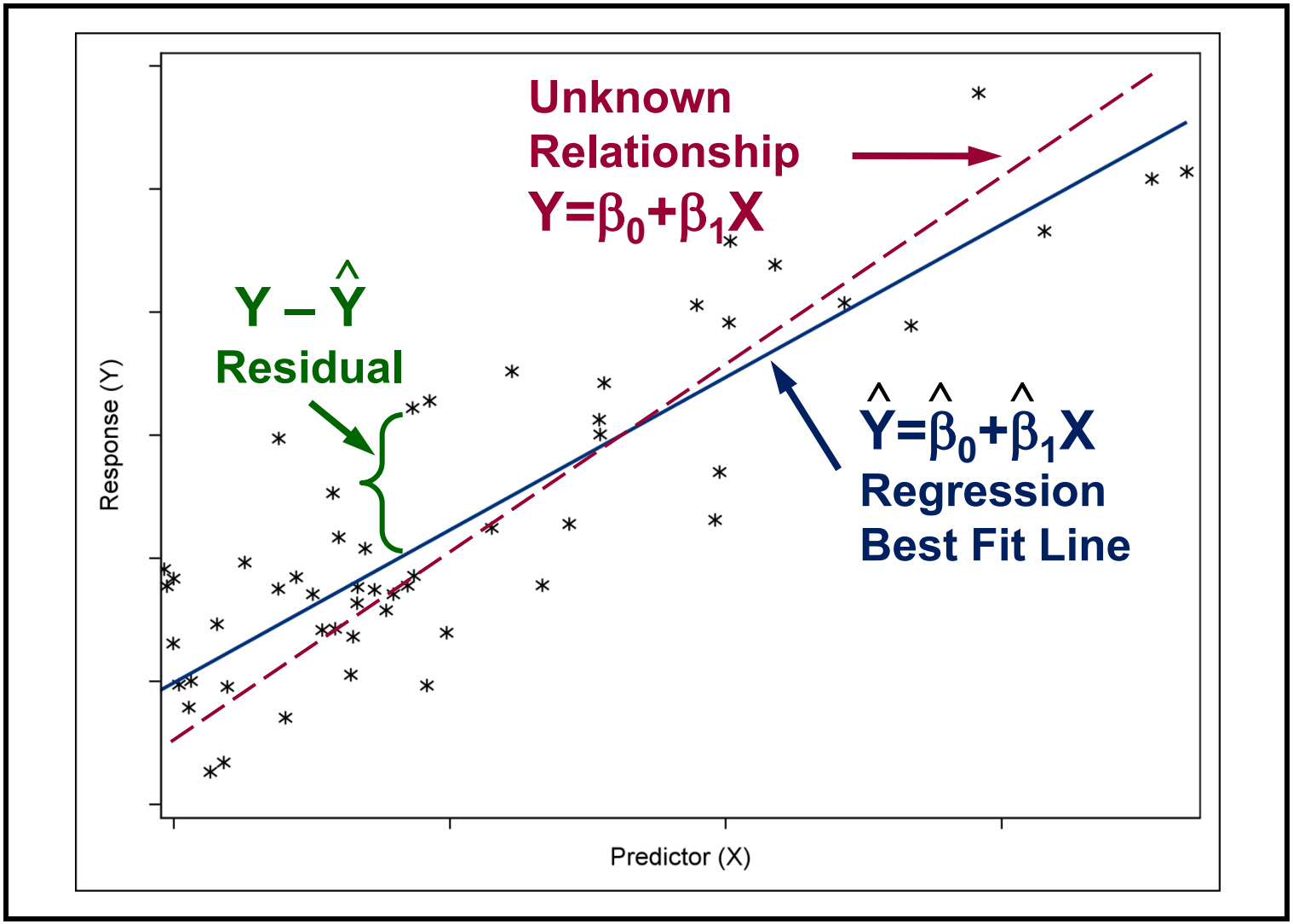
  …and/or…

  - **Predict**: predict the values of the response variable given the values of the predictor variable

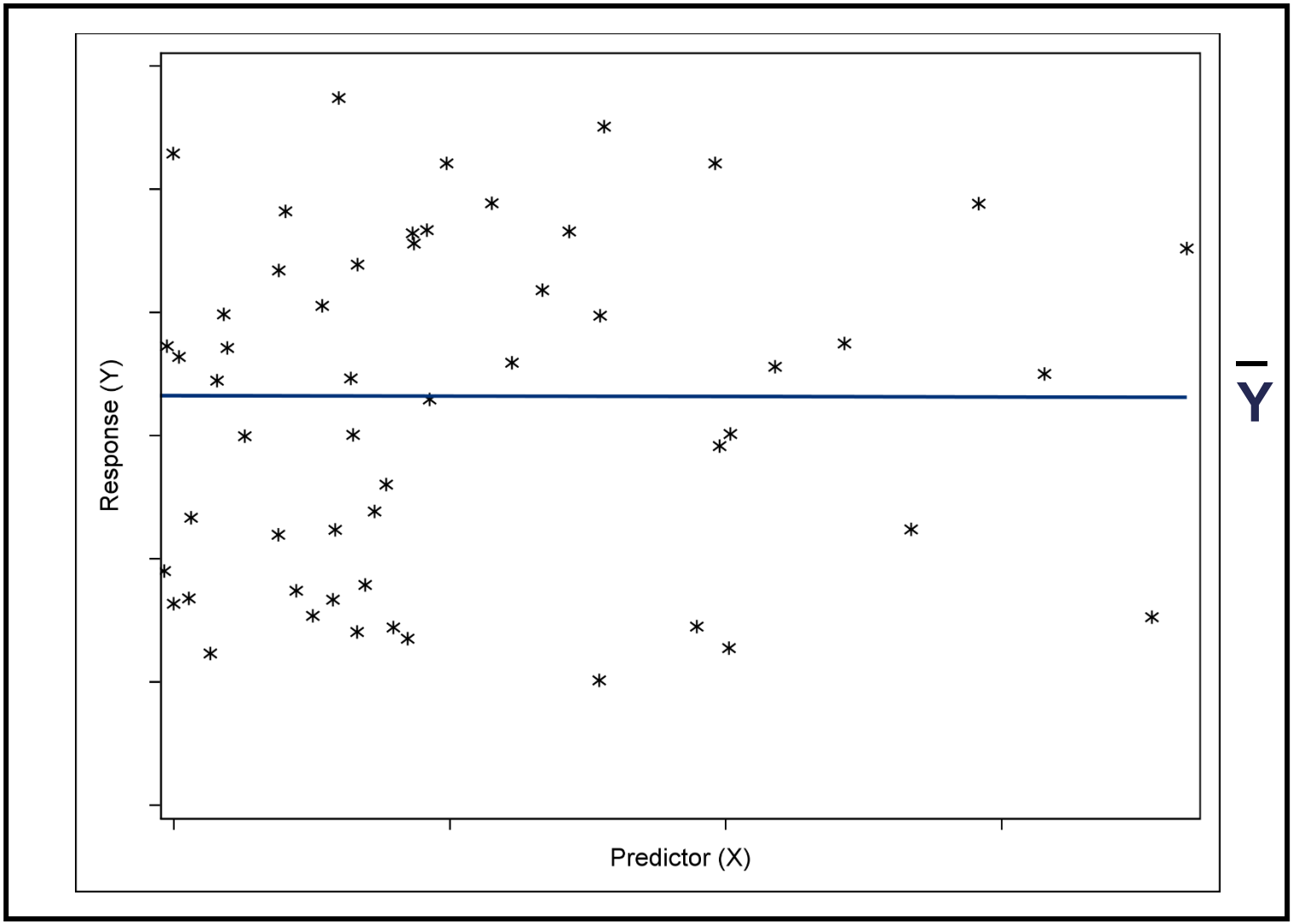
Simple Linear Regression Model



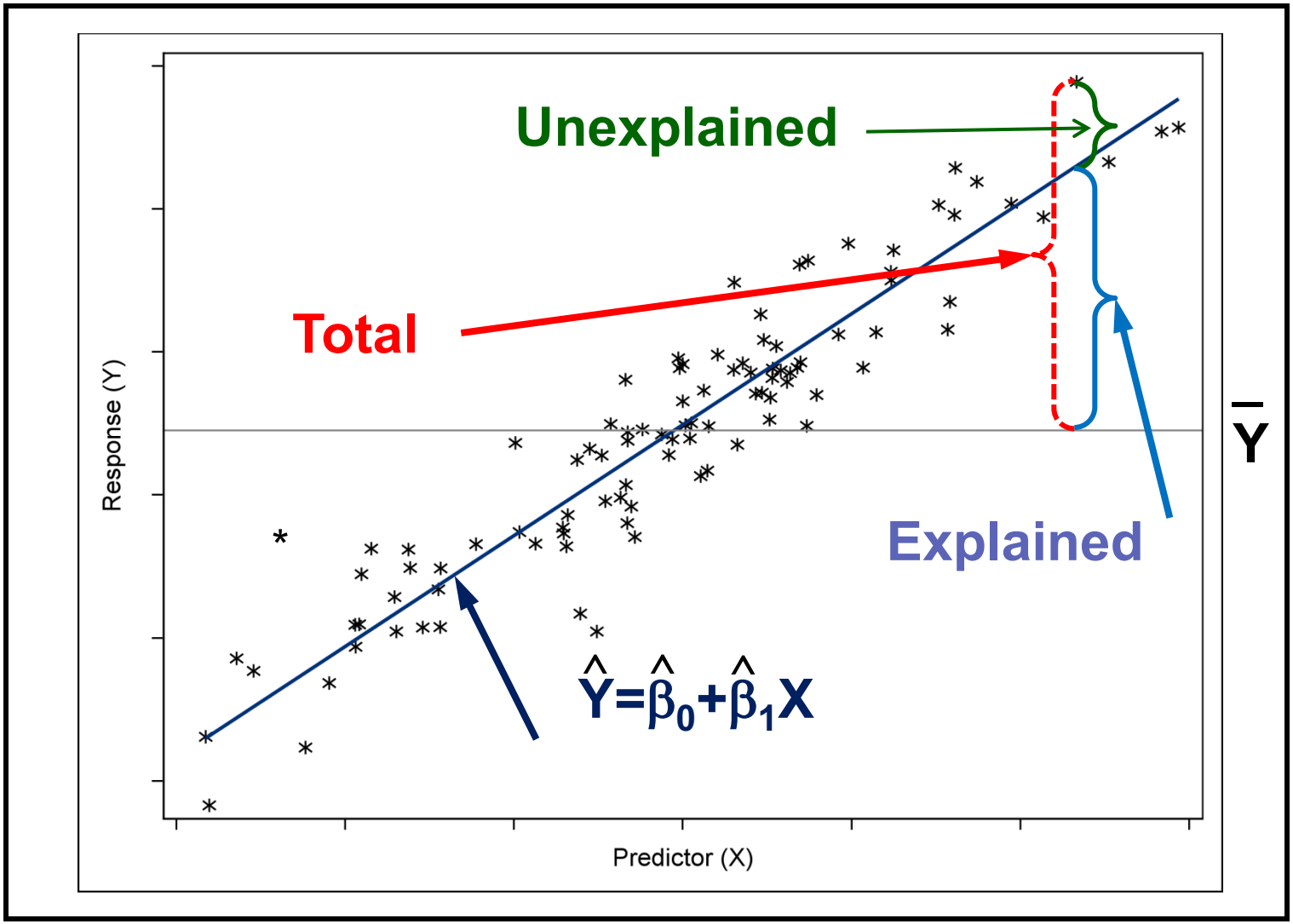
Simple Linear Regression Model



The Baseline Model (Null Hypothesis)



Explained vs. Unexplained Variability

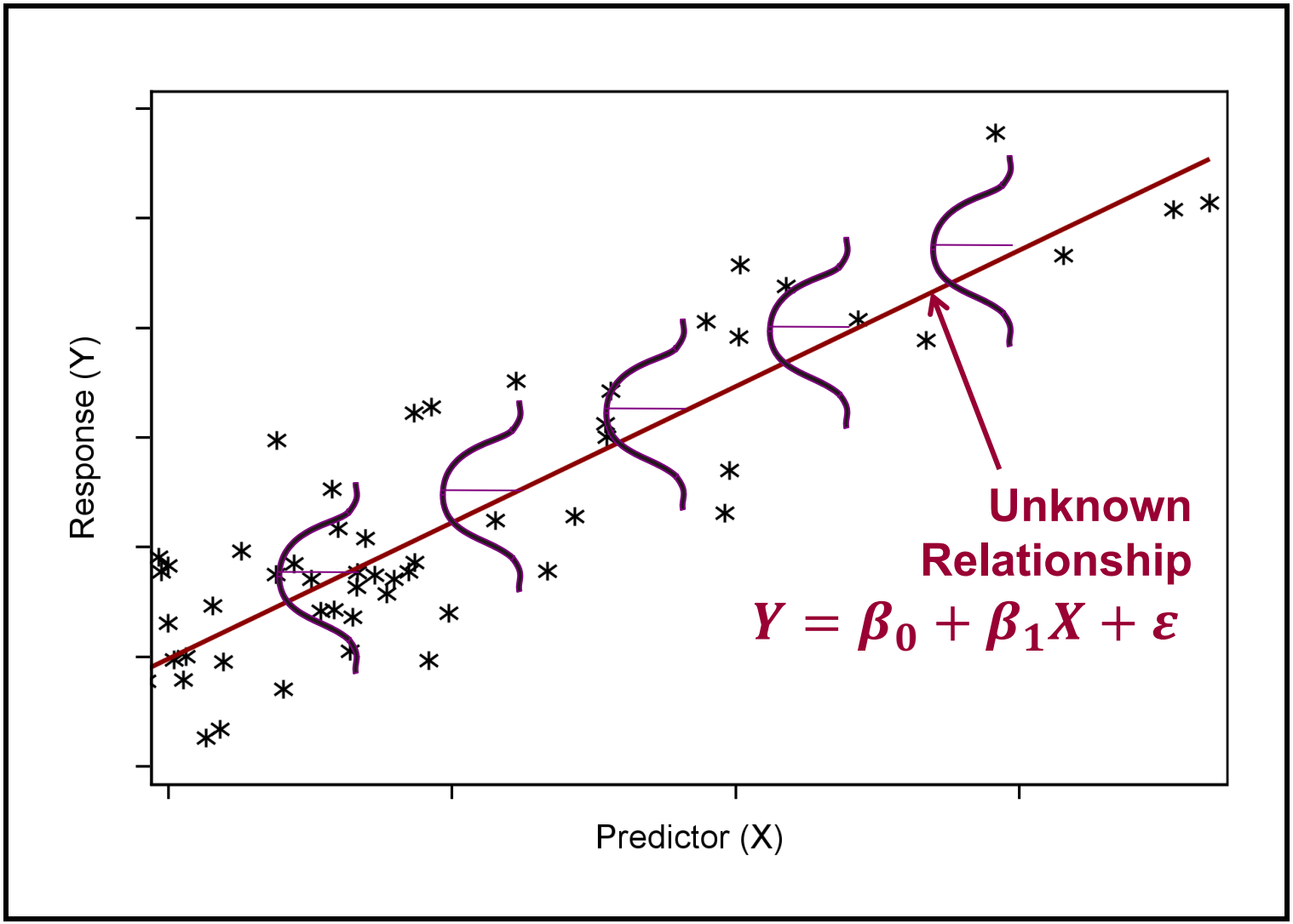


Model Hypothesis Test

- **Null Hypothesis:**
  - The simple linear regression model does ***not*** fit the data better than the baseline model.
  - $\beta_1 = 0$

- **Alternative Hypothesis:**
  - The simple linear regression model does fit the data better than the baseline model.
  - $\beta_1 \neq 0$

Assumptions of Simple Linear Regression



Simple Linear Regression

```
proc reg data=bootcamp.ameshousing3;
     model SalePrice=Lot_Area;
     title "Regression with Lot Area as Predictor";
run;
quit;
```

Poll

Quiz

LAB 4

Don’t forget to take the lab check on Moodle!