Introduction to ANOVA and Regression

MSA 2024

Overview of Models in this Course

Type of Predictors / Type of Response	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

Population Model

- Linear Model: $Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon = E(Y|\mathbf{x})$
- Linear Model: $Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$

Linear Models Terminology

- Linear Model: $Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon$
  - **Error (Random or Stochastic part of the equation)**

Linear Models Terminology

- Linear Model: $Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k = E(Y|\mathbf{x})$
  - **Explanatory** (Input, Independent, Predictor) **Variables**

Linear Models Terminology

- Linear Model: $Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k = E(Y \mid \mathbf{x})$
  - **Explanatory** (Input, Independent, Predictor) **Variables**
  - **Response** (Target, Dependent) **Variable**

Linear Models Terminology

- Linear Model: $Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k = E(Y \mid \mathbf{x})$
  - **Explanatory** (Input, Independent, Predictor) **Variables**
  - **Response** (Target, Dependent) **Variable**
  - **Parameter Estimates**

Linear Models Terminology

- Linear Model: $Y = \beta_0 + \beta_1 x_1 + \ \ldots \ + \beta_k x_k = E(Y \mid \mathbf{x})$
  - **Explanatory** (Input, Independent, Predictor) **Variables**
  - **Response** (Target, Dependent) **Variable**
  - **Parameter Estimates**
  - **Residuals/Random Error**
- Types of Modeling
  - Explanatory Modeling (Statistical Inference/Hypothesis Testing)
  - Predictive Modeling

Explanatory vs. Predictive Modeling

Explanatory Modeling

- *How* is x_i related to y_i?
- Descriptions
- Fewer Variables/Simpler Model
- Tends to focus on p-values and confidence intervals (interpretations)

Predictive Modeling

- Given x_i can I predict y_i?
- Predictions
- Many Variables and complex models

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon$$

Honest Model Assessment

Train-Validation-Test

Data Preprocessing

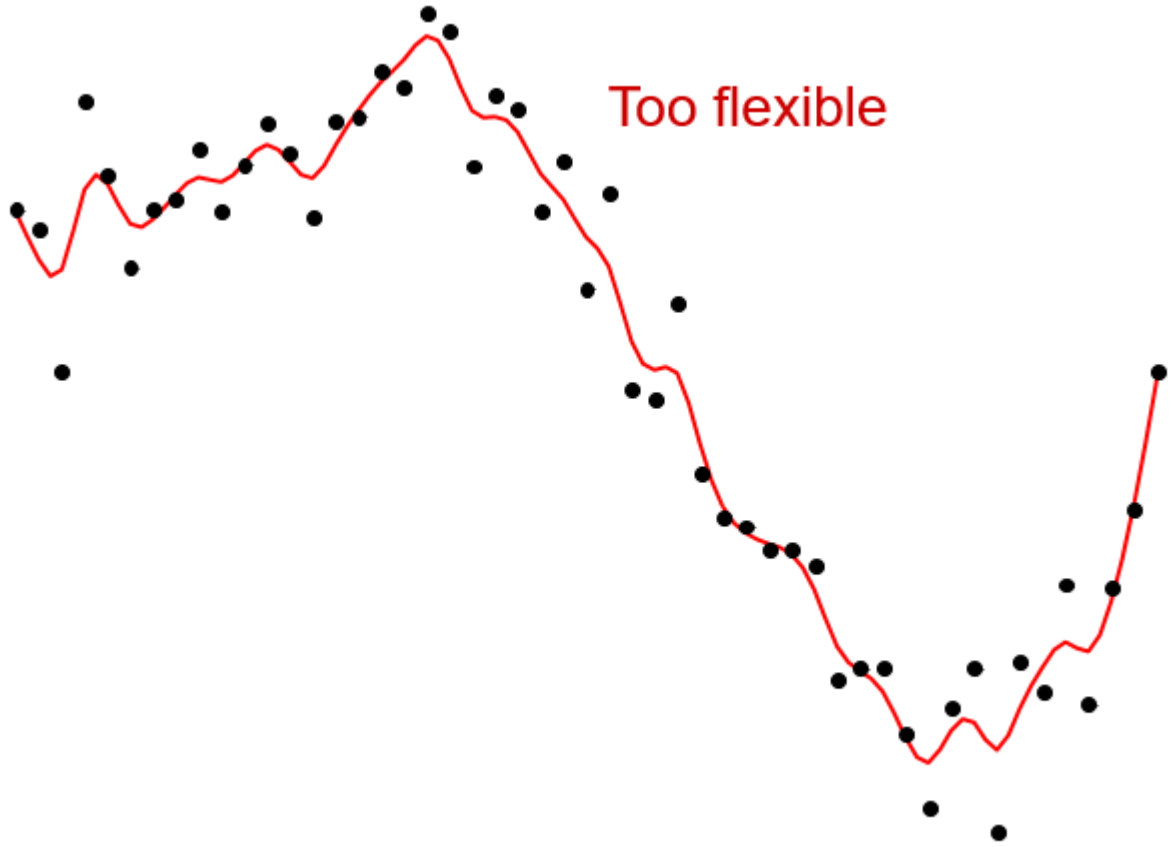
- When you first receive your data, you'll explore for distributions/ outliers, and missing values.
- ***Before*** you look at any relationships between input variables and **target variables**, you should split into training, validation and test samples.

  (Or decide on Cross-Validation / Testing)

The Problem of Overfitting

- Left unchecked, models will capture nuances of the data on which they're built (the training data).
- When these "patterns" do not hold up in validation or test data, the model performance suffers.
  - We say the model **does not generalize well**.
  - The model is **overfit**.

Overfitting



The Problem of Overfitting

- Want to make sure your models are generalizable
  - Not just good models of training sample.
  - Can predict equally well on out-of-sample data
- Split into Training + Validation + Test sets is necessary. Somewhere around 2/3 training, 1/3 validation/test is typical.
    - Lots of data? 50-40-10 split
    - Not so much data? 70-20-10 split
    - Not enough data? Use Cross-Validation

Train-Test Split in R

- For illustrations this summer, we will only divide data into a train and test data
- R Code is shown below:

```
library(tidyverse)
set.seed(123)
ames <- ames %>% mutate(id = row_number())
train <- ames %>% sample_frac(0.7)
test <- anti_join(ames, train, by = 'id')

dim(train)

## [1] 2051   82

dim(test)

## [1] 879  82
```

Bivariate Exploratory Data Analysis (EDA)

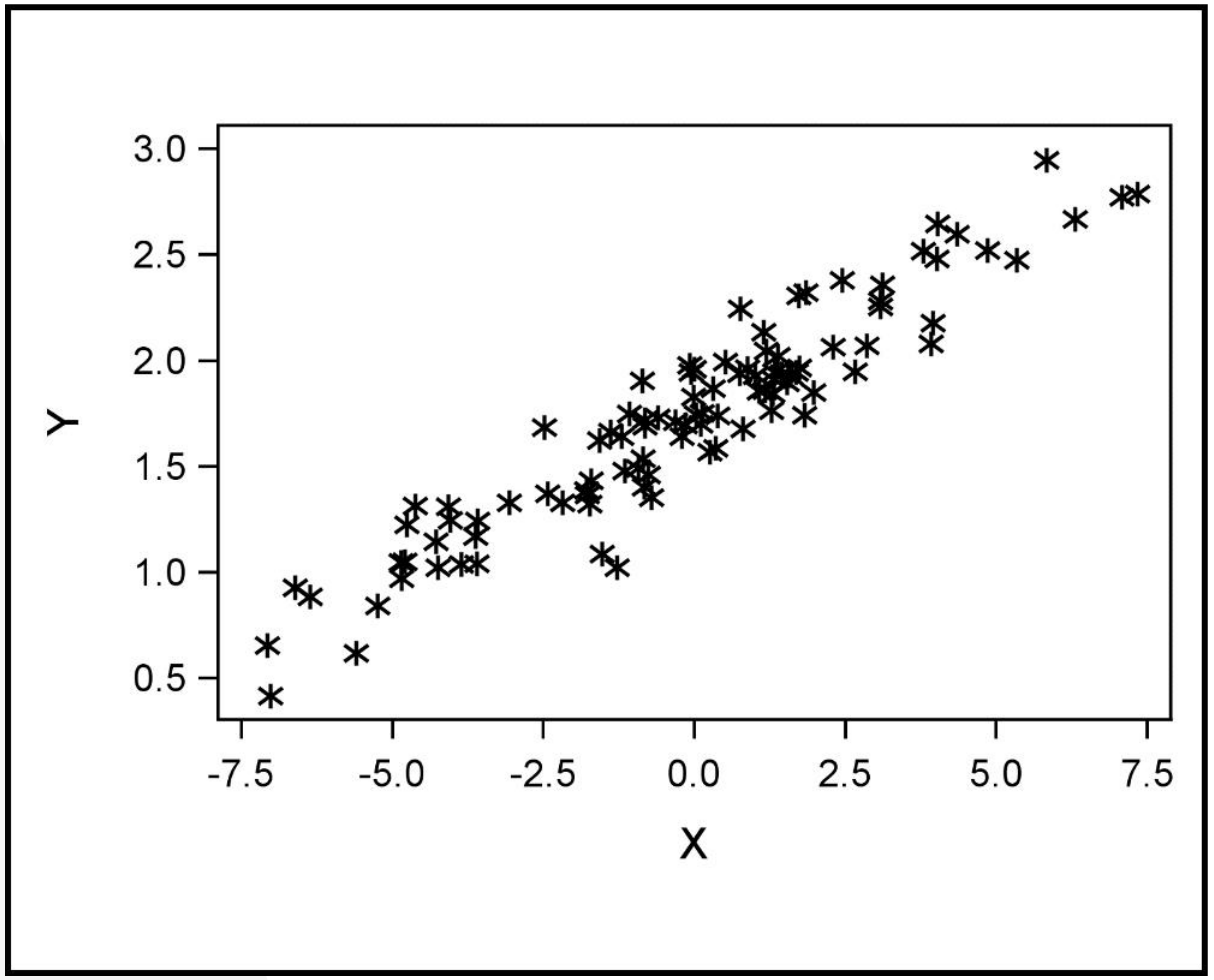
Exploring Associations

Associations between two variables

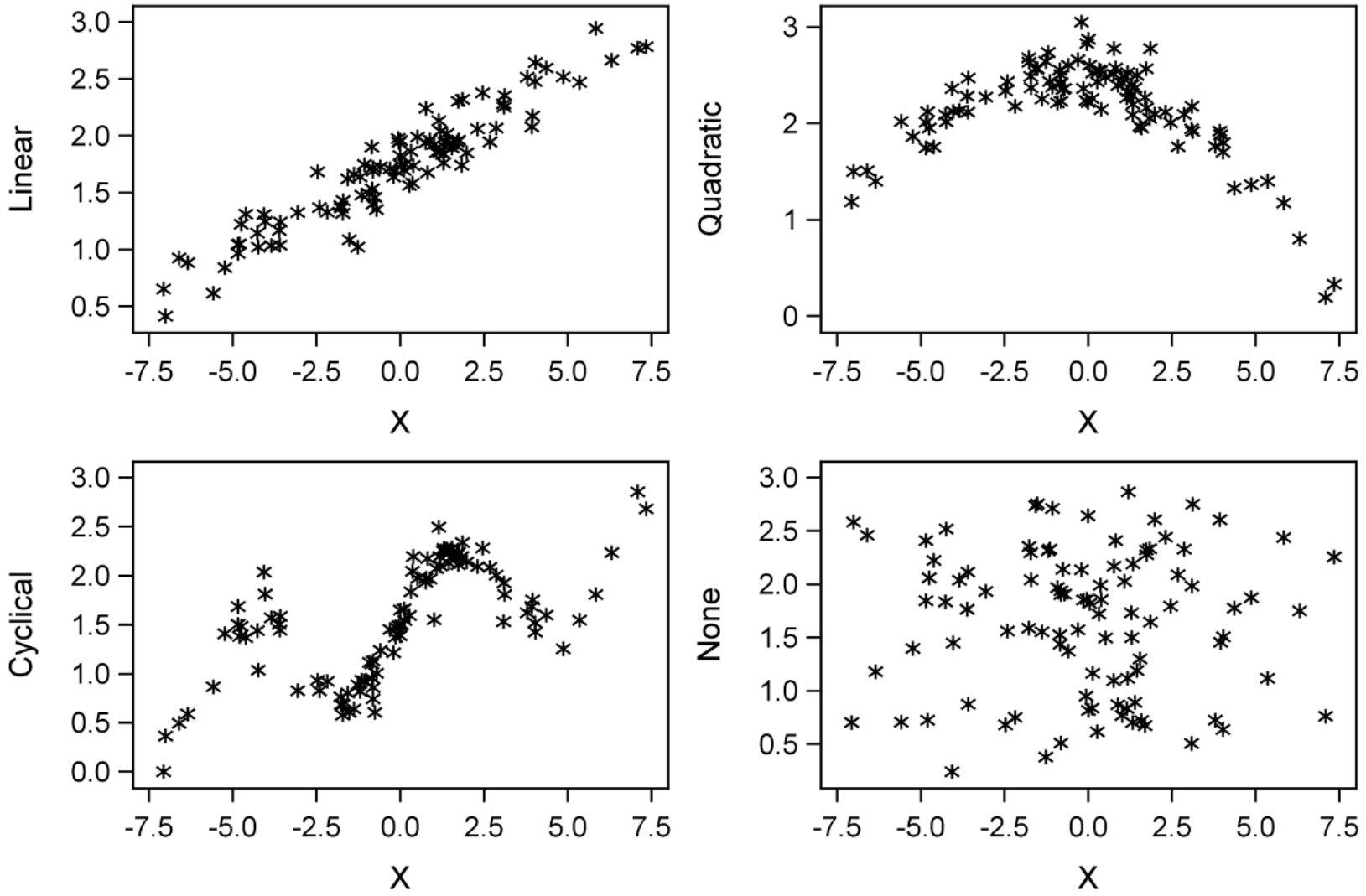
- **Association**: The expected value of one variable changes at different levels of the other variable
- A ***linear* association** between two continuous variables can be inferred when the general shape of a scatter plot of the two variables portrays a straight line.

Scatter Plot - Linear Association

As x increases, y tends to increase at a constant rate.

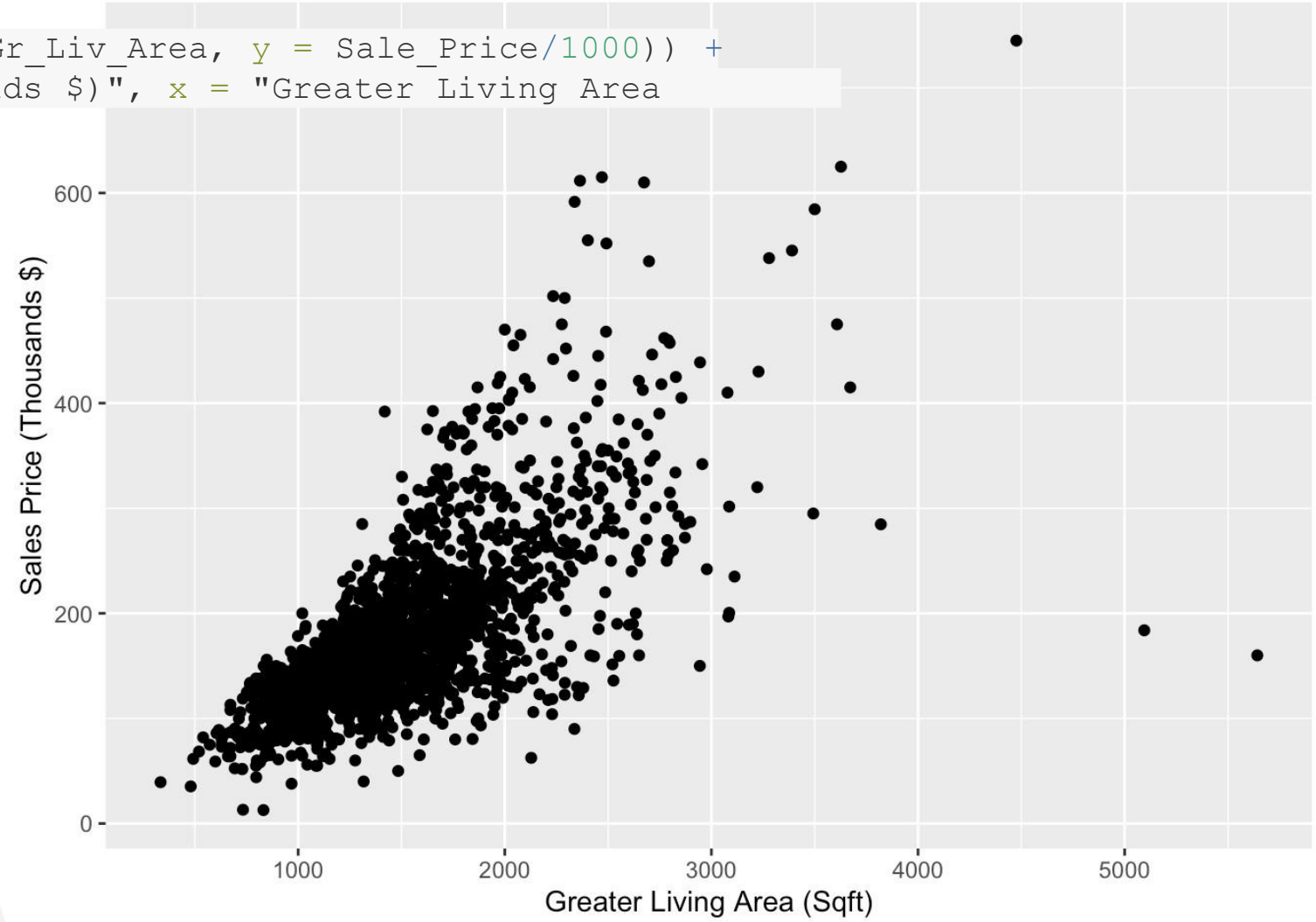


Scatter Plots - Other Association

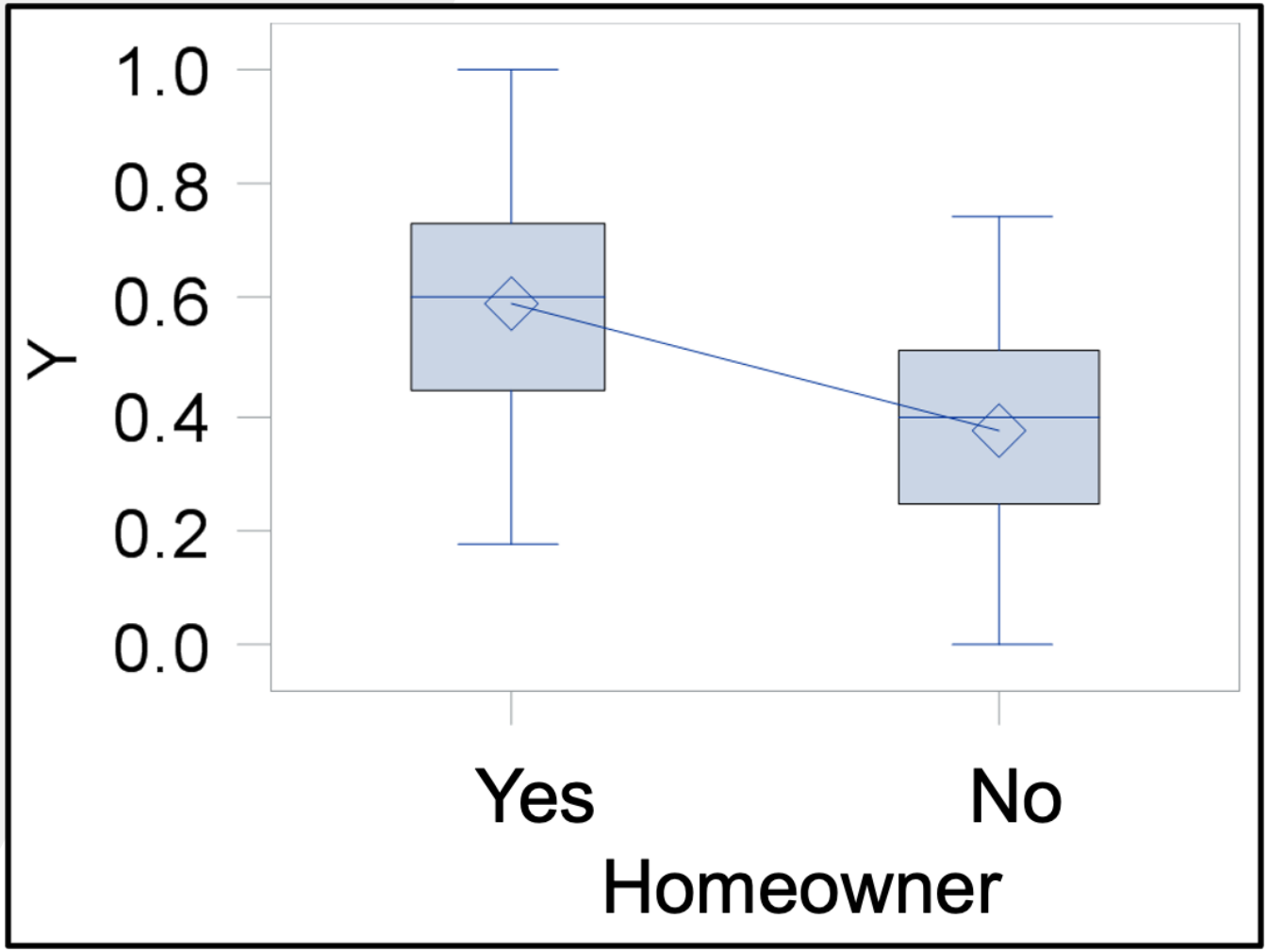


Scatter Plots in R

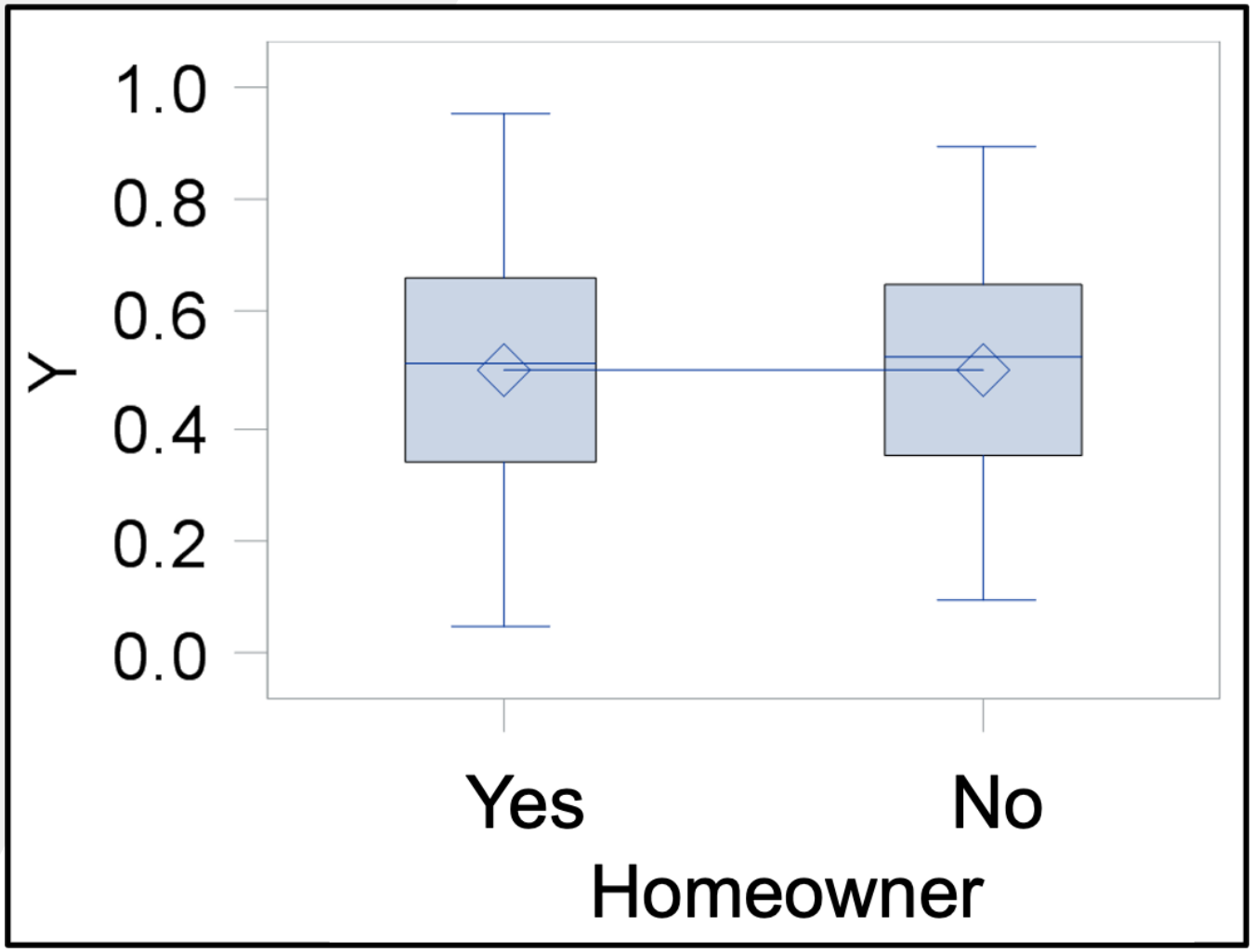
```
ggplot(data = train) +
    geom_point(mapping = aes(x = Gr_Liv_Area, y = Sale_Price/1000)) +
    labs(y = "Sales Price (Thousands $)", x = "Greater Living Area
    (Sqft)")
```



Association - Categorical and Continuous Variables

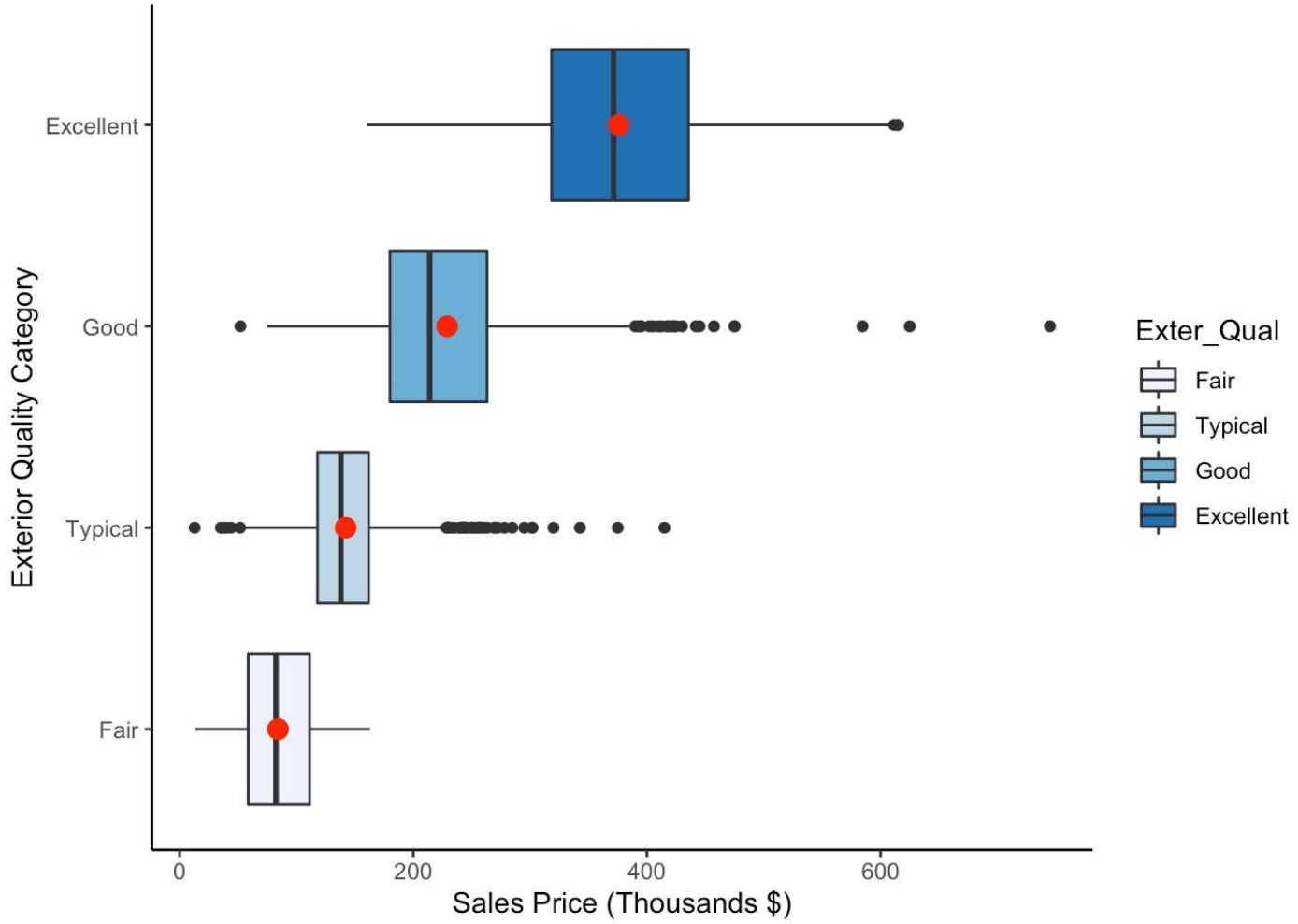


(No) Association - Categorical and Continuous Variables



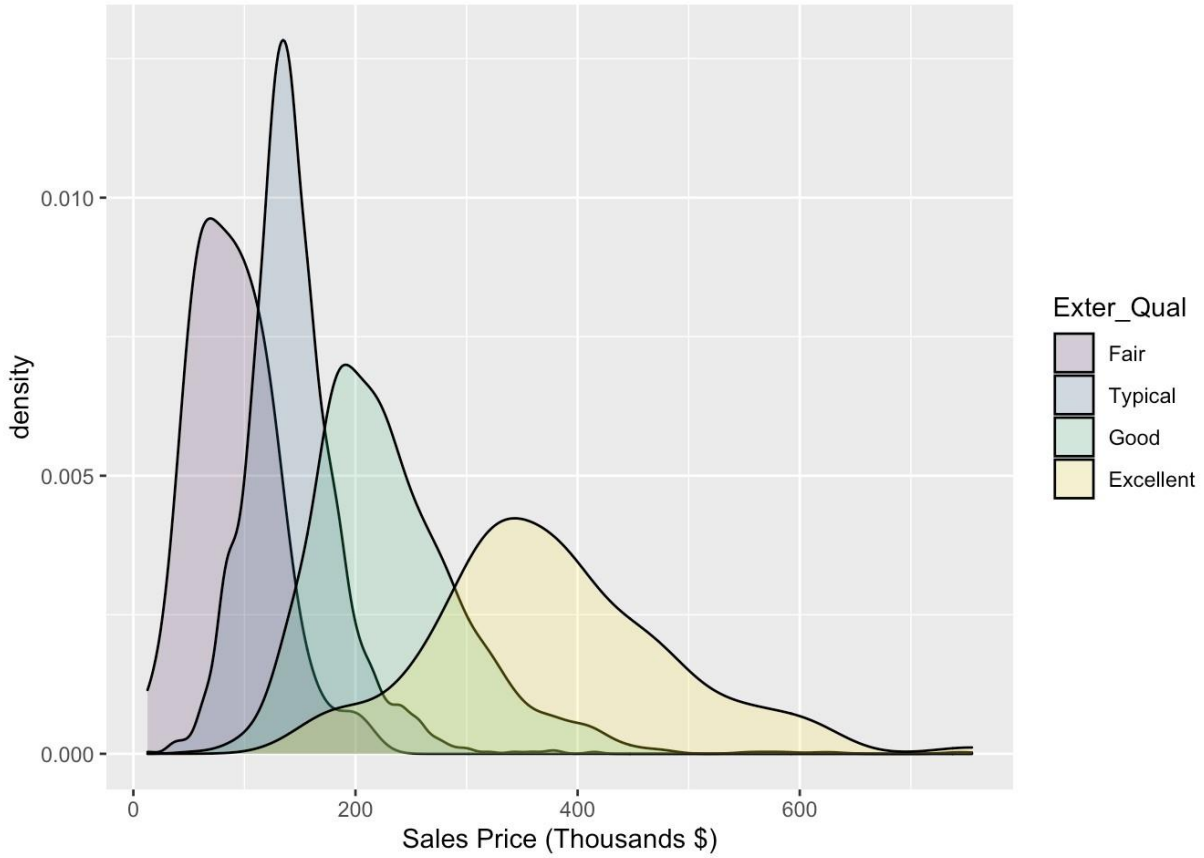
Grouped Box-Plots in R

```
ggplot(data = train, aes(y = Sale_Price/1000,
                         x = Exter_Qual,
                       fill = Exter_Qual)) +
  geom_boxplot() +
  labs(y = "Sales Price (Thousands $)",
       x = "Exterior Quality Category") +
  stat_summary(fun = mean,
               geom = "point",
              shape = 20,
               size = 5,
              color = "red",
               fill = "red") +
  scale_fill_brewer(palette="Blues") +
  theme_classic() + coord_flip()
```



Overlaid histograms in R

```
ggplot(ames,aes(x = Sale_Price/1000,
                fill = Exter_Qual)) +
  geom_density(alpha = 0.2,
               position = "identity") +
  labs(x = "Sales Price (Thousands $)")
```



One-Way ANOVA

A *test* of relationship between categorical input and quantitative response

Overview of Models in this Course

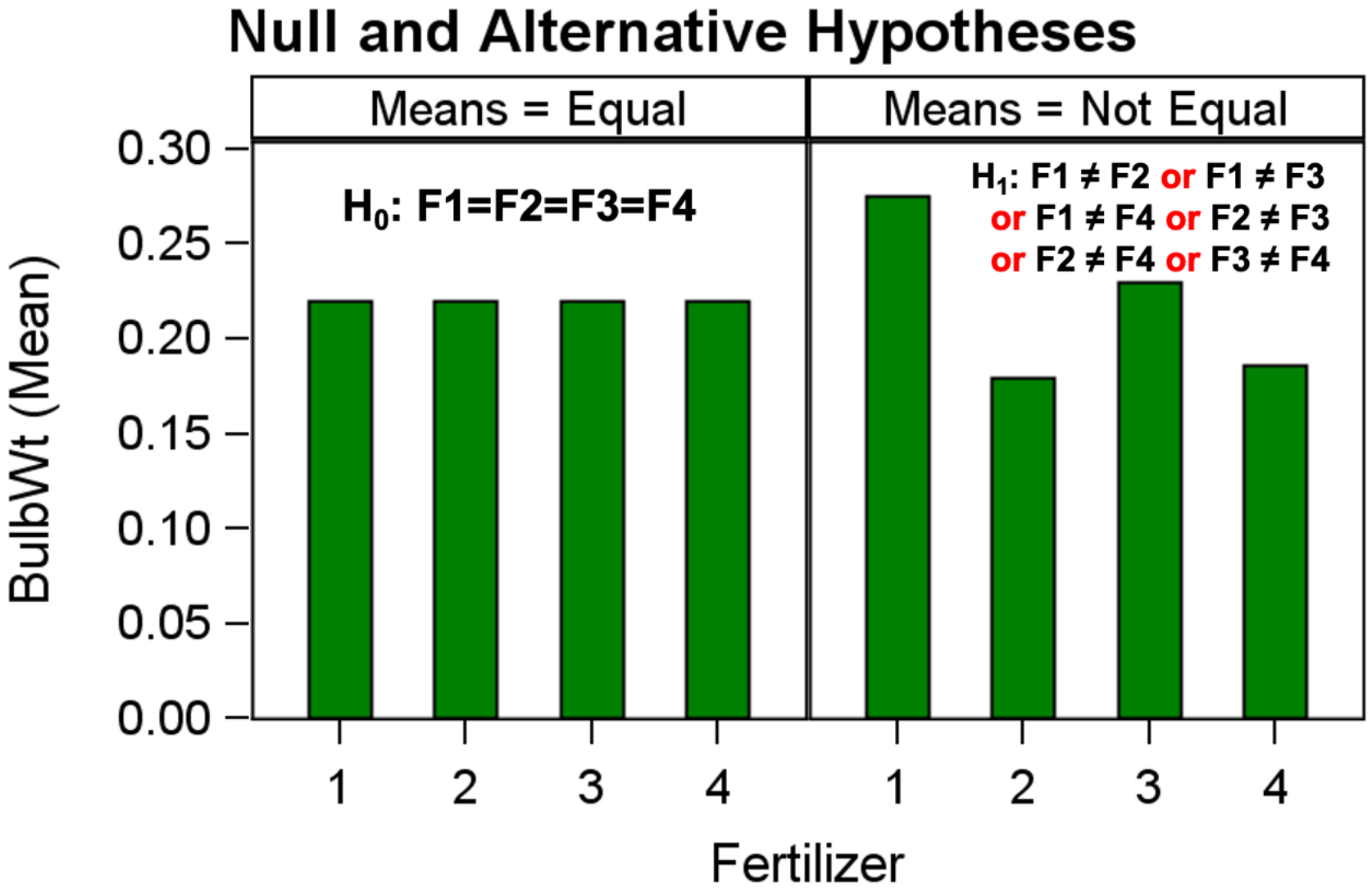
Type of Predictors / Type of Response	Categorical	Continuous	Continuous and Categorical
Continuous	Analysis of Variance (ANOVA)	Ordinary Least Squares (OLS) Regression	Ordinary Least Squares (OLS) Regression
Categorical	Logistic Regression	Logistic Regression	Logistic Regression

ANOVA Overview

Are there any (statistically significant) differences among the groups in terms of their mean response? Similar to the two-sample t-test, but with three groups.

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

The ANOVA Hypothesis



The ANOVA Model

Let's say we are interested in looking at 3 different fertilizers to see their affect on the yield of crops (A, B, C).

Want to know if there is a significant difference in the mean yield of the crop.
H_0: $\mu_A = \mu_B = \mu_C$ versus H_A: At least one mean different

The linear model is:

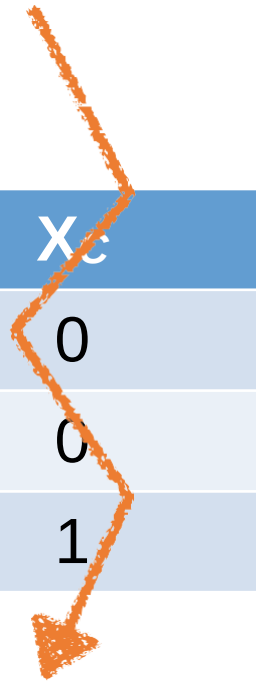
$$Y = \beta_0 + \beta_A x_A + \beta_B x_B + \varepsilon$$

$$x_A = \begin{cases} 1 & If\ A \\ 0 & otherwise \end{cases}$$

$$x_B = \begin{cases} 1 & If\ B \\ 0 & otherwise \end{cases}$$

The ANOVA Model (Reference Coding)

Fertilizer	X_A	X_B	X_C
A	1	0	0
B	0	1	0
C	0	0	1

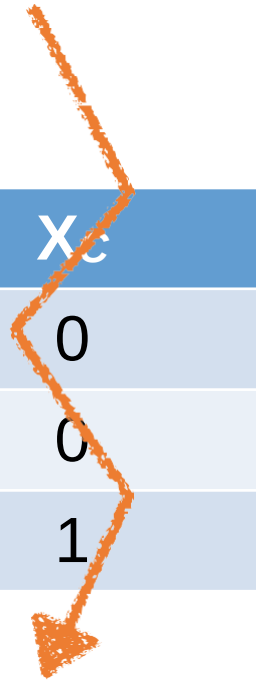


$$y = \beta_0 + \beta_A x_A + \beta_B x_B + \varepsilon$$

For Fertilizer C, $y = \beta_0$

The ANOVA Model

Fertilizer	X_A	X_B	X_C
A	1	0	0
B	0	1	0
C	0	0	1

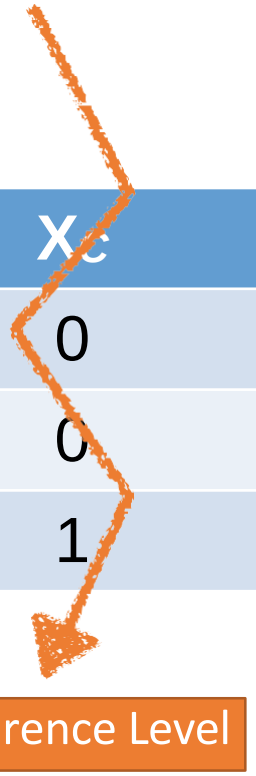


Reference Level

β_A is the difference between mean response in fertilizer A vs. fertilizer C

The ANOVA Model

Fertilizer	X_A	X_B	X_C
A	1	0	0
B	0	1	0
C	0	0	1



β_B is difference between mean response in fertilizer B vs. fertilizer C

The ANOVA Model

For this example with 3 types of fertilizer, what will the *predictions* from this model look like?

How many *unique* values are possible for the predicted value?

Assumptions for ANOVA

- Observations are independent.
- Each group is normally distributed.
  - Often replaced by the statement that the *residuals* of the ANOVA model are normally distributed
- All groups have equal variances (homoskedasticity).
  - If this holds, we use “pooled” variance (classical ANOVA)
  - If this does *not* hold, we use Welch’s ANOVA

Assessing ANOVA Assumptions

- **Good data collection** designs help ensure the **independence** assumption.
- **QQ-Plots, histograms, and/or formal tests** can be used to verify the assumption that the error is approximately **normally distributed**.
- A **formal test of equal variances or viewing the residual plot** is typically done to assess **homoskedasticity**.

ANOVA Hypothesis Test in R

H_0 The means of each level of `Exter_Qual` are equal

H_A At least one mean is different

```
## Analysis of Variance Table
##
## Response: Sale_Price
##               Df     Sum Sq    Mean Sq F value    Pr(>F)
## Exter_Qual     3 6.6913e+12 2.2304e+12  701.83 < 2.2e-16 ***
## Residuals   2047 6.5054e+12 3.1780e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1
```

```
ames_lm <- lm(Sale_Price ~ Exter_Qual, data = train)

anova(ames_lm)
```

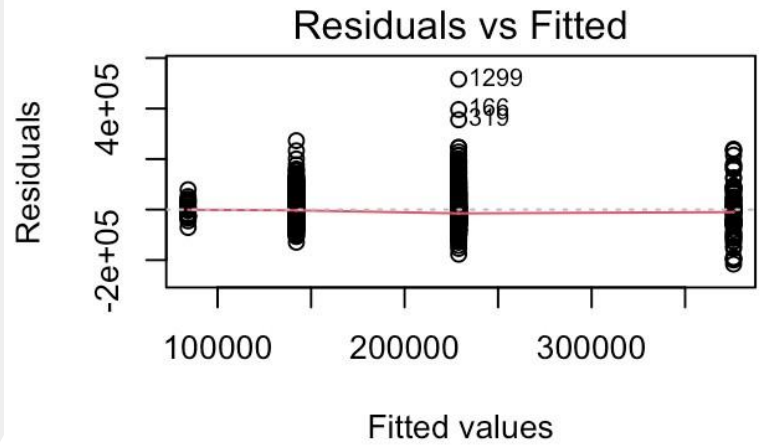
ANOVA Predictions in R

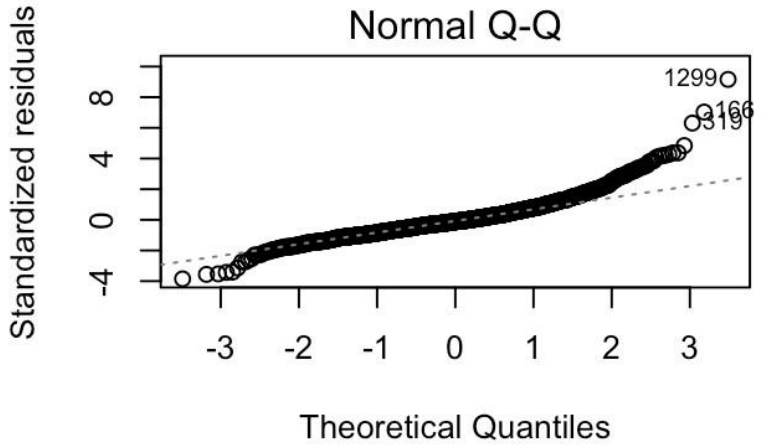
```
train$pred_anova <- predict(ames_lm, data = train)

train$resid_anova <- resid(ames_lm, data = train)

(model_output = train %>%
       select(Sale_Price, pred_anova, resid_anova))
```

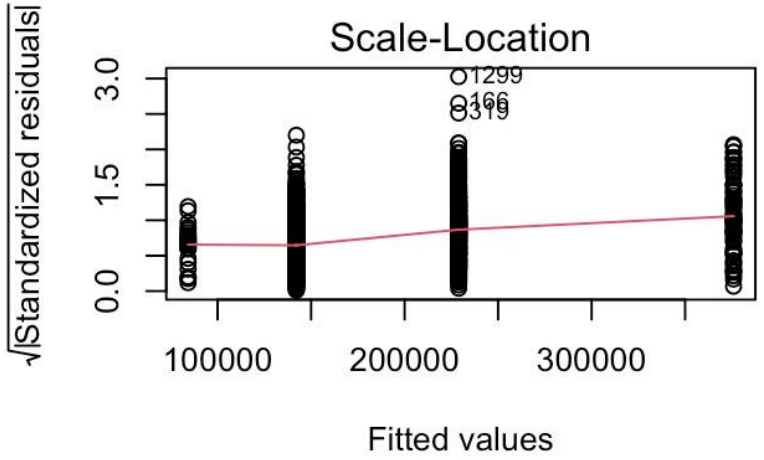
		Sale_Price	pred_anova	resid_anova
1		232600	228910	3690
2		166000	228910	-62910
3		170000	142107	27893
4		252000	228910	23090
5		134000	142107	-8107
6		164700	228910	-64210
7		193500	142107	51393
8		118500	142107	-23607
9		94000	142107	-48107
10		111250	142107	-30857
…	…	…	…	…

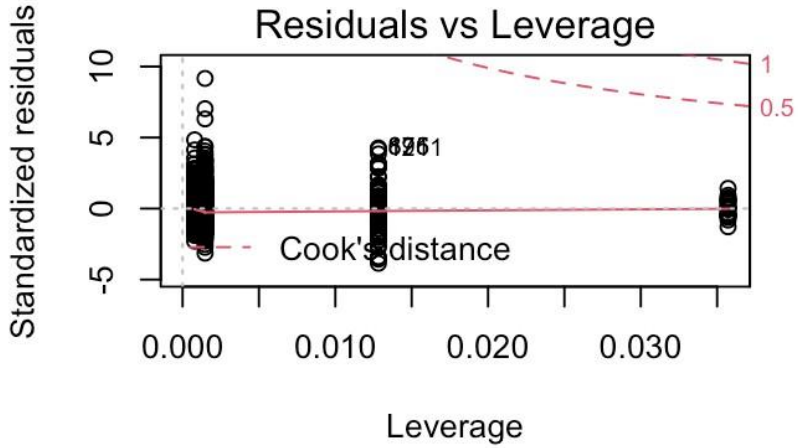
Testing Assumptions in R

```
par(mfrow=c(2,2))
plot(ames_lm)
par(mfrow=c(1,1))
```









Testing Assumptions in R

```
#install.packages('car')
#install.packages('stats')
library(car)
library(stats)

leveneTest(Sale_Price ~ Exter_Qual, data = train)
```

H_0 The group variances are equal

H_A The group variances are NOT equal

Levene's Test *requires* normality of underlying data

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group    3  76.879 < 2.2e-16 ***
##       2047
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fligner.test(Sale_Price ~ Exter_Qual, data = train)
```

Fligner's Test does *not require* normality of underlying data

```
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Sale_Price by Exter_Qual
## Fligner-Killeen:med chi-squared = 206.26, df = 3, p-value < 2.2e-16
```

When the Equal Variances Assumption Fails… Welch's ANOVA

Welch's ANOVA is similar to the two-sample t-test with unequal variances

```
oneway.test(Sale_Price ~ Exter_Qual, data = train, var.equal = FALSE)
```

One-way analysis of means (not assuming equal variances)

data: Sale_Price and Exter_Qual
F = 431.82, num df = 3.00, denom df = 102.11, p-value < 2.2e-16

When the Normality Assumption Fails… Kruskal-Wallis

Kruskal-Wallis is a **nonparametric** test for more than two groups.

Conditions	**Interpretation of Significant Kruskal-Wallis Test**
Group distributions are identical in shape,variance, and symmetric	Difference in means
Group distributions are identical in shape, variance, but not symmetric	Difference in medians
Else	Difference in location. (**distributional dominance**)

A random variable A has **distributional dominance** over random variable B if $P(A \geq x) \geq P(B \geq x)$ for all x, and for *some* x, $P(A > x) > P(B > x)$ (i.e. they're not equal for all x.)

Kruskal-Wallis Nonparametric ANOVA in R

```
kruskal.test(Sale_Price ~ Exter_Qual, data = train)
```

```
##  Kruskal-Wallis rank sum test
##
## data:  Sale_Price by Exter_Qual
## Kruskal-Wallis chi-squared = 975.98, df = 3, p-value < 2.2e-16
```

Conclusion: The distribution of Sale Price is different for the different levels of Exterior Quality

ANOVA Analysis Plan Summary

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

1. Produce Descriptive Statistics (EDA)
2. Verify Assumptions
    A. Independence
    B. Errors Normally Distributed
    C. Equal Variance for all groups
3. Examine the p-value for overall F-test in the ANOVA table. If p-value is less than α, reject the null hypothesis.

ANOVA Post-Hoc Tests

ANOVA Post-Hoc Testing

- The ANOVA F-Test will tell us if at least one group mean is different from the others.
- Next natural question? Which groups are different!
- Want to compare the groups pairwise (two-sample t-tests) to determine if they are different.
- Issue? Testing multiple hypotheses compounds error.

Experimentwise Error Rates

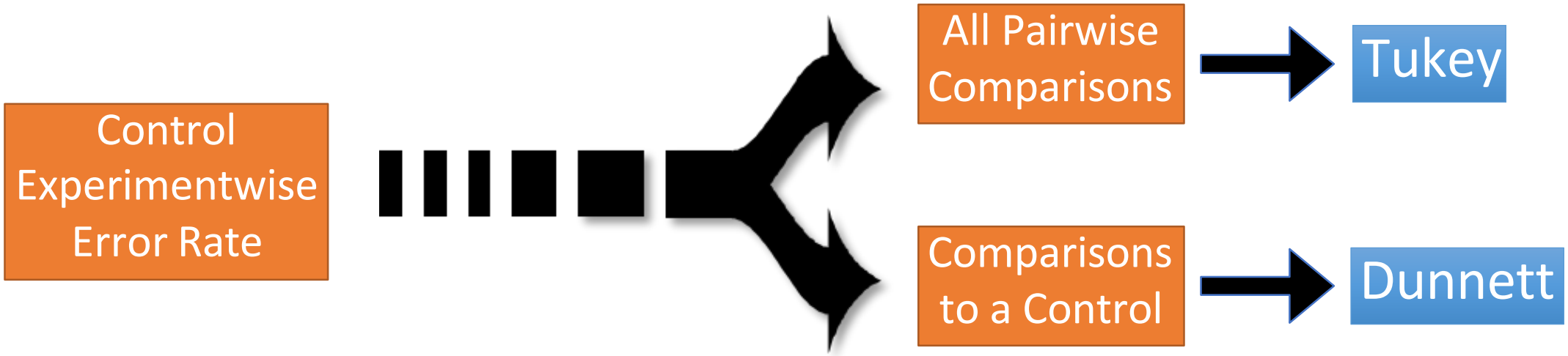
If we make comparisons, each having an error rate of , what is the probability that I make *at least one* error?

Number of Groups Compared	Number of Comparisons	Experimentwise Error Rate (α=0.05)
2	1	0.05
3	3	0.14
4	6	0.26
5	10	0.40 😱

Comparisonwise error rate α

Experimentwise error rate $= 1 - (1 - \alpha)^n$

Two Common Methods for Multiple Comparisons*



* Not an exhaustive list. There are *many* more options in the literature.

Tukey's Honest Significant Difference (HSD)

- Also known as the Tukey-Kramer Test
- Appropriate when you want to make ***all*** pairwise comparisons between groups
- Experimentwise error rate is…
  - Equal to α when ***all*** pairwise comparisons made
  - Less than α if ***fewer than all*** comparisons are made (overly conservative)

Tukey's Test in R

```r
ames_aov <- aov(Sale_Price ~ Exter_Qual, data = train)
tukey.ames <- TukeyHSD(ames_aov)
print(tukey.ames)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Sale_Price ~ Exter_Qual, data = train)
##
## $Exter_Qual
##                          diff       lwr       upr p adj
## Typical-Fair         57887.91  30194.31  85581.52 5e-07
## Good-Fair           144690.25 116739.87 172640.63 0e+00
## Excellent-Fair      291684.79 259752.41 323617.16 0e+00
## Good-Typical         86802.34  79910.03  93694.64 0e+00
## Excellent-Typical   233796.87 216886.62 250707.12 0e+00
## Excellent-Good      146994.54 129666.98 164322.10 0e+00
```

Conclusion: All pairs are significantly different

Dunnett's Test for Control Comparison

- If you're *not* making all pairwise comparisons, Tukey's test is overly conservative
- Dunnett's test is designed to correct for L-1 tests, where L is the number of levels including the control
- Developed for situations where you're only interested in comparisons to a control/placebo group.

Dunnett's Test in R

```
library(DescTools)
DunnettTest(x = train$Sale_Price, g = train$Exter_Qual, control = 'Typical')
```

```
##
##   Dunnett's test for comparing several treatments with a control :
##     95% family-wise confidence level
##
## $Typical
##                          diff    lwr.ci    upr.ci    pval
## Fair-Typical      -57887.91 -83628.55 -32147.28 2.6e-07 ***
## Good-Typical       86802.34  80396.08  93208.59 < 2e-16 ***
## Excellent-Typical 233796.87 218079.15 249514.60 < 2e-16 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion: All levels are significantly different from "Typical"

Break out and Lab 3

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

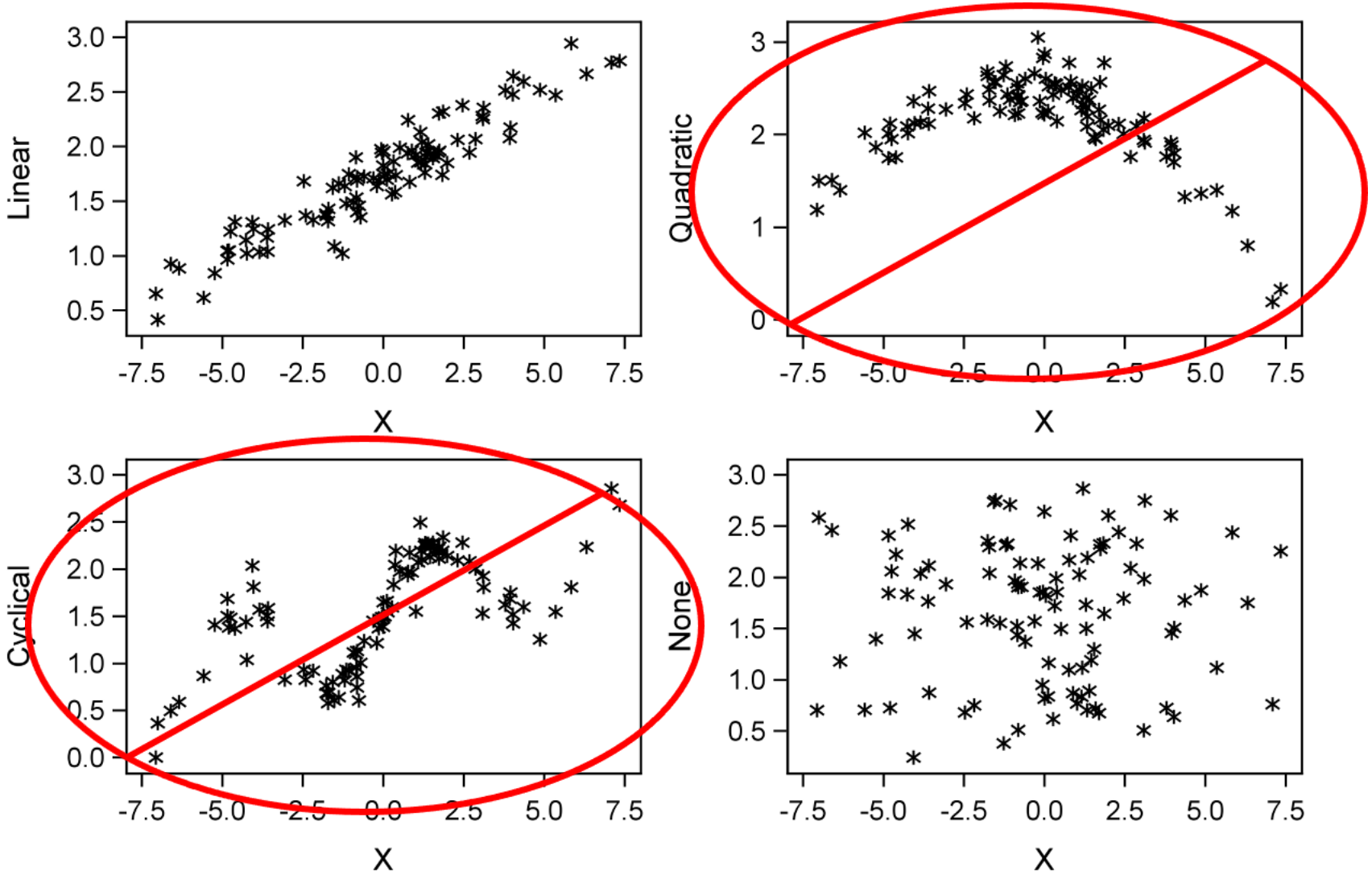
Overview of Models in this Course

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

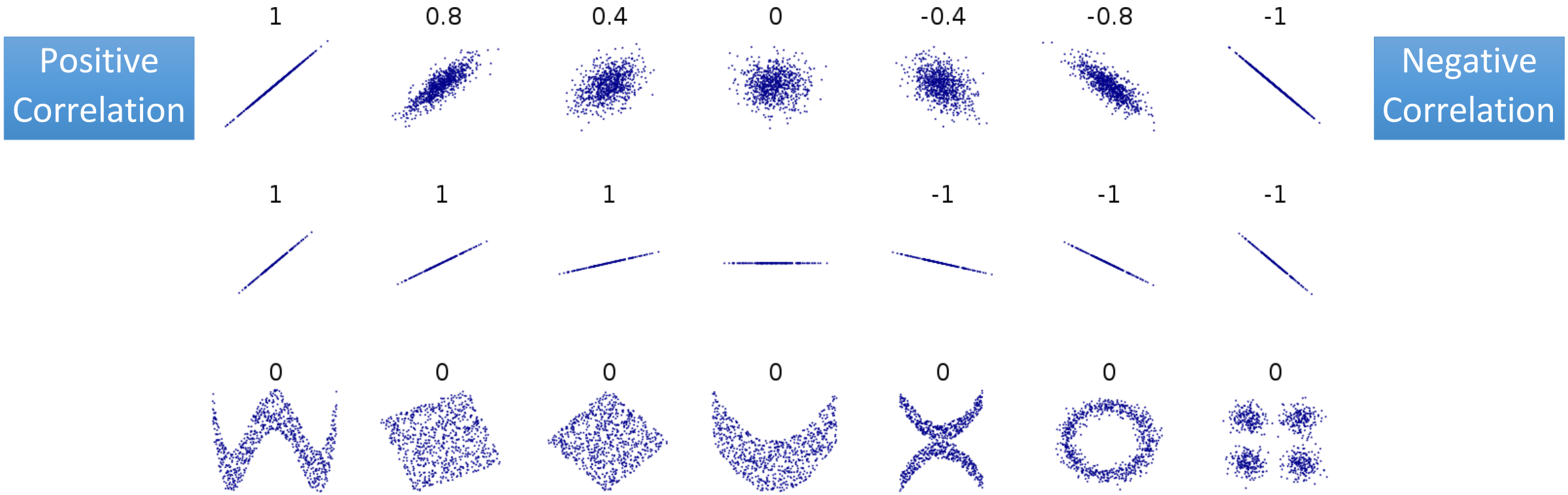
Pearson’s Correlation,

Measuring Strength of Linear Relationships

Pearson's Correlation measures Linear Relationships

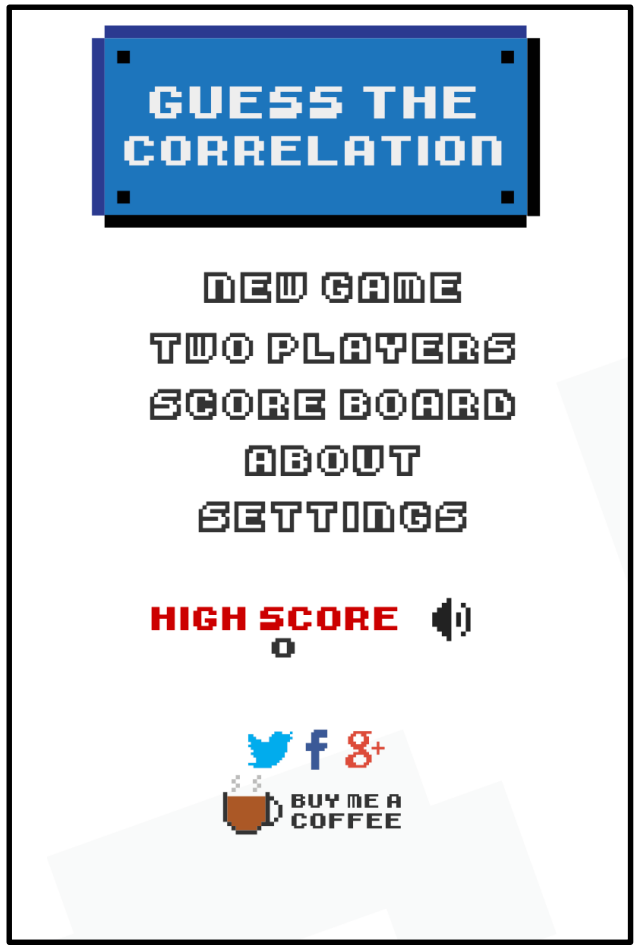


$-1 \leq \rho \leq 1$

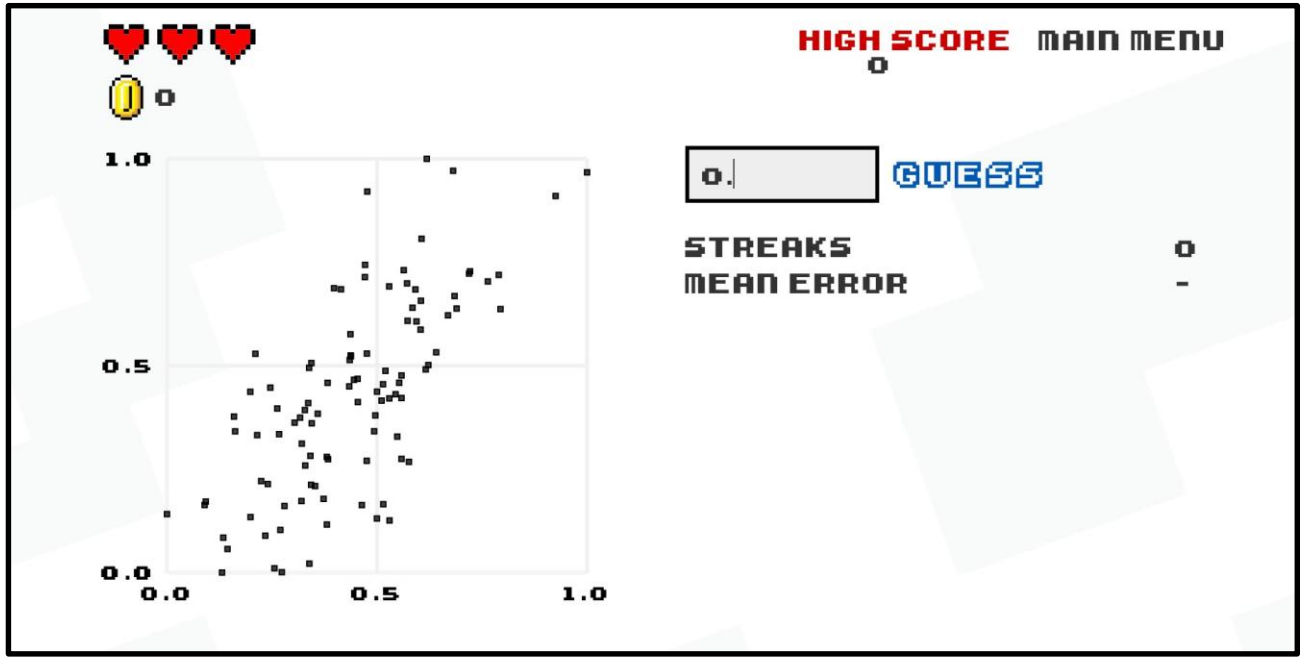


Guess the Correlation!

www.guessthecorrelation.com



- Guess within 0.05: +1 life (❤) and +5 coins (🪙)
- Guess within 0.10: +1 coin
- Guess error > 0.10: -1 life

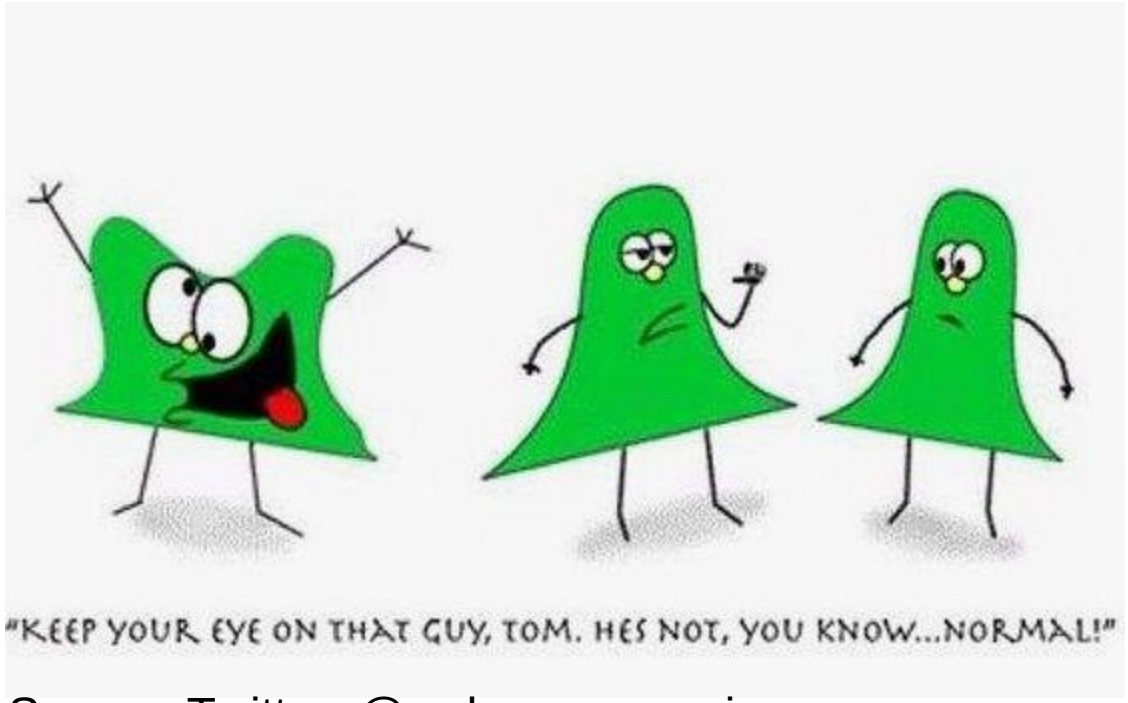


Hypothesis Test for a Correlation

- The parameter representing population correlation is ρ.
- ρ is estimated by the sample statistic *r.*
- $H_0 : \rho = 0$
- Rejecting H_0 only indicates confidence that ρ is not exactly zero - it doesn't mean that the relationship is *practically* significant.

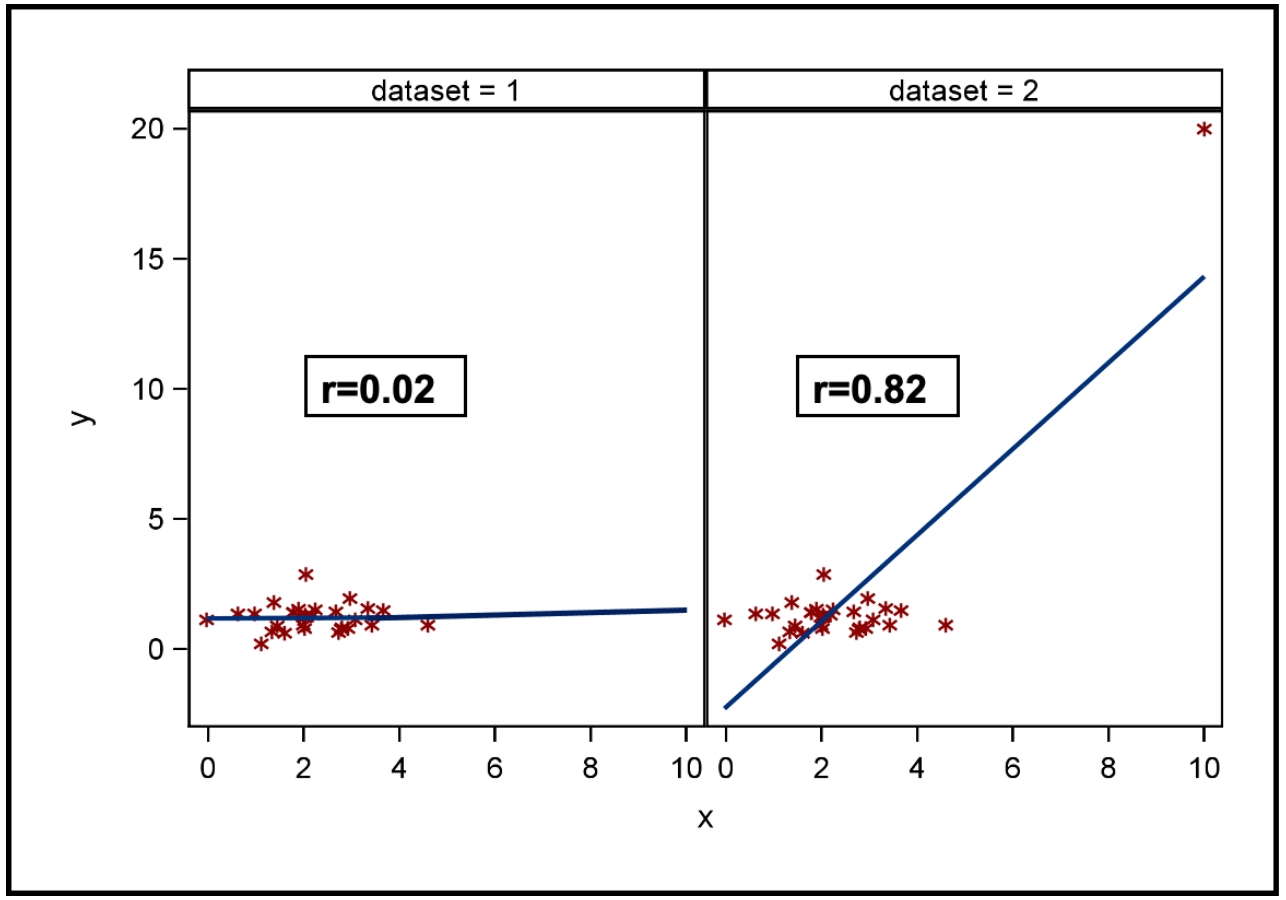
Outliers affect correlations

- Correlations can be skewed by outliers



Source: Twitter: @weloveeconomics

Correlation does NOT imply causation



Test of Correlation in R

```
cor.test(train$Gr_Liv_Area,train$Sale_Price)
```

```
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 44.185, df = 2049, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6756538 0.7200229
## sample estimates:
##       cor
## 0.698509
```

Conclusion: There is a linear relationship between a home's area and its sale price.

The Correlation Matrix

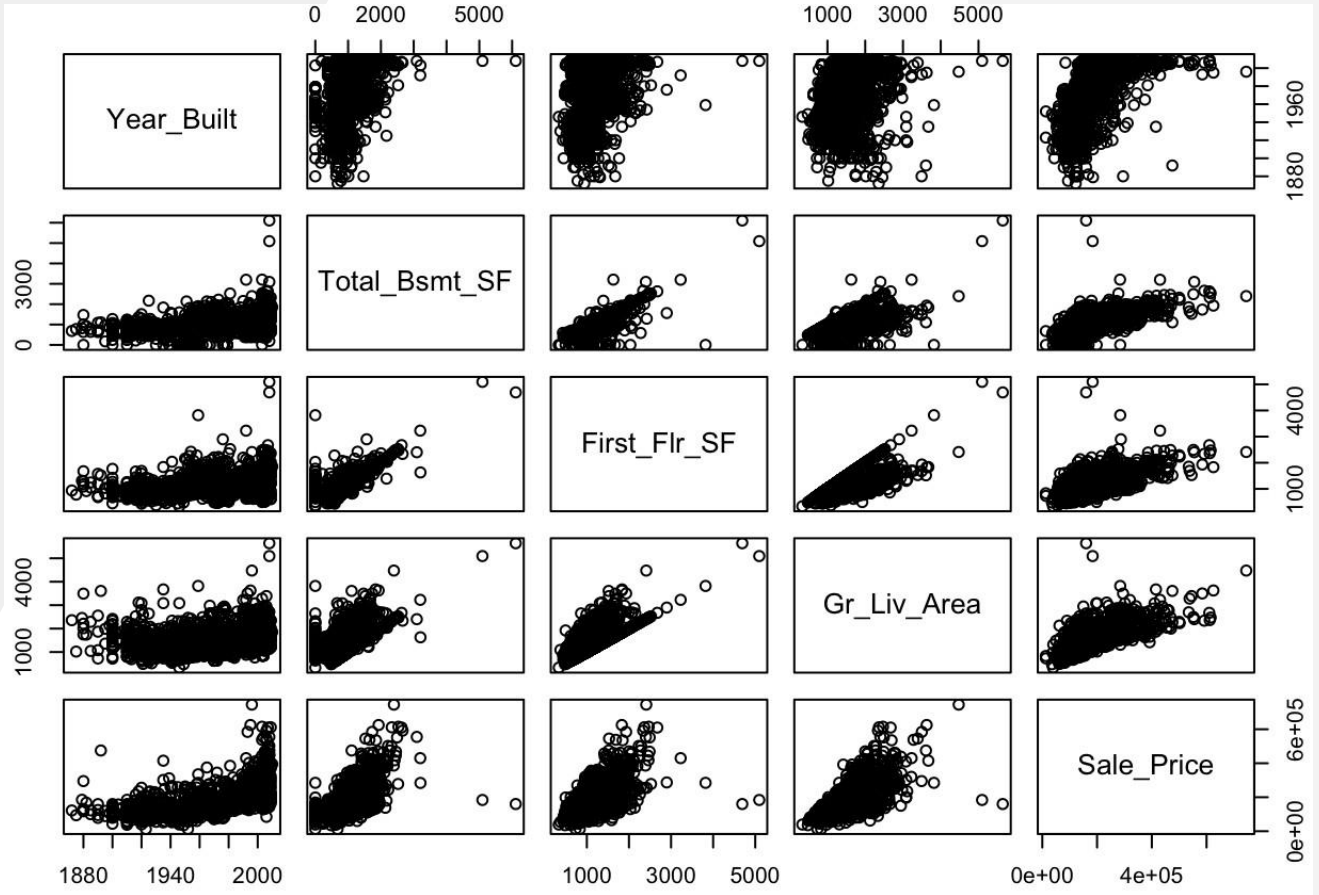
$$C_{ij} = correlation(x_i, x_j)$$

```
cor(train[, c('Year_Built','Total_Bsmt_SF','First_Flr_SF','Gr_Liv_Area','Sale_Price')])
```

	Year_Built	Total_Bsmt_SF	First_Flr_SF	Gr_Liv_Area	Sale_Price
Year_Built	1.0000000	0.4037104	0.3095407	0.2454325	0.5668889
Total_Bsmt_SF	0.4037104	1.0000000	0.8120419	0.4643838	0.6276502
First_Flr_SF	0.3095407	0.8120419	1.0000000	0.5707205	0.6085229
Gr_Liv_Area	0.2454325	0.4643838	0.5707205	1.0000000	0.6985090
Sale_Price	0.5668889	0.6276502	0.6085229	0.6985090	1.0000000

The Corresponding Plot Matrix

```
pairs(train[, c('Year_Built','Total_Bsmt_SF','First_Flr_SF','Gr_Liv_Area','Sale_Price')])
```

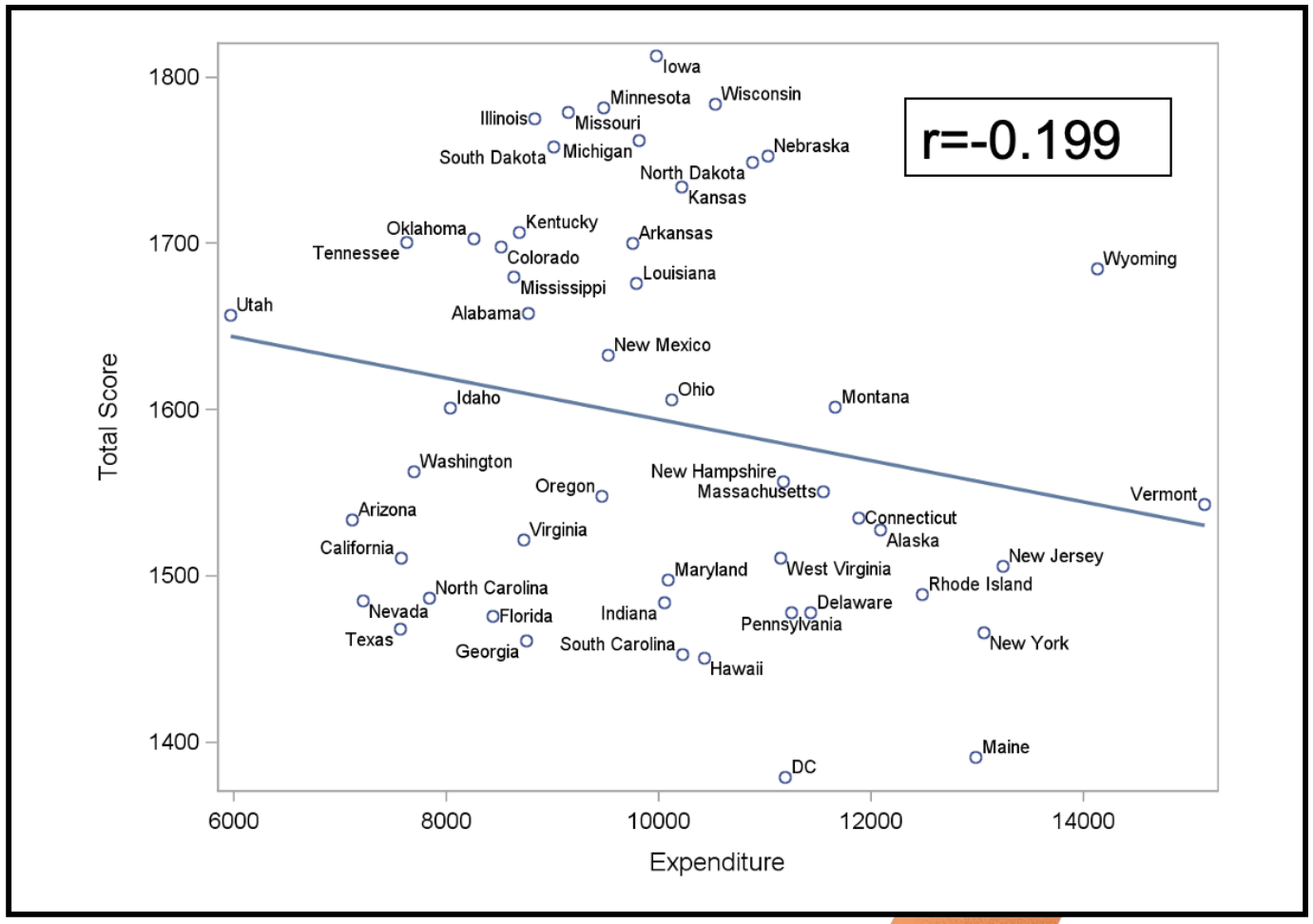


Correlation does NOT imply causation

- A strong correlation ***does not mean*** that a change in one variable *causes* a change in the other
- Correlations can be misleading when both variables are affected by other variables

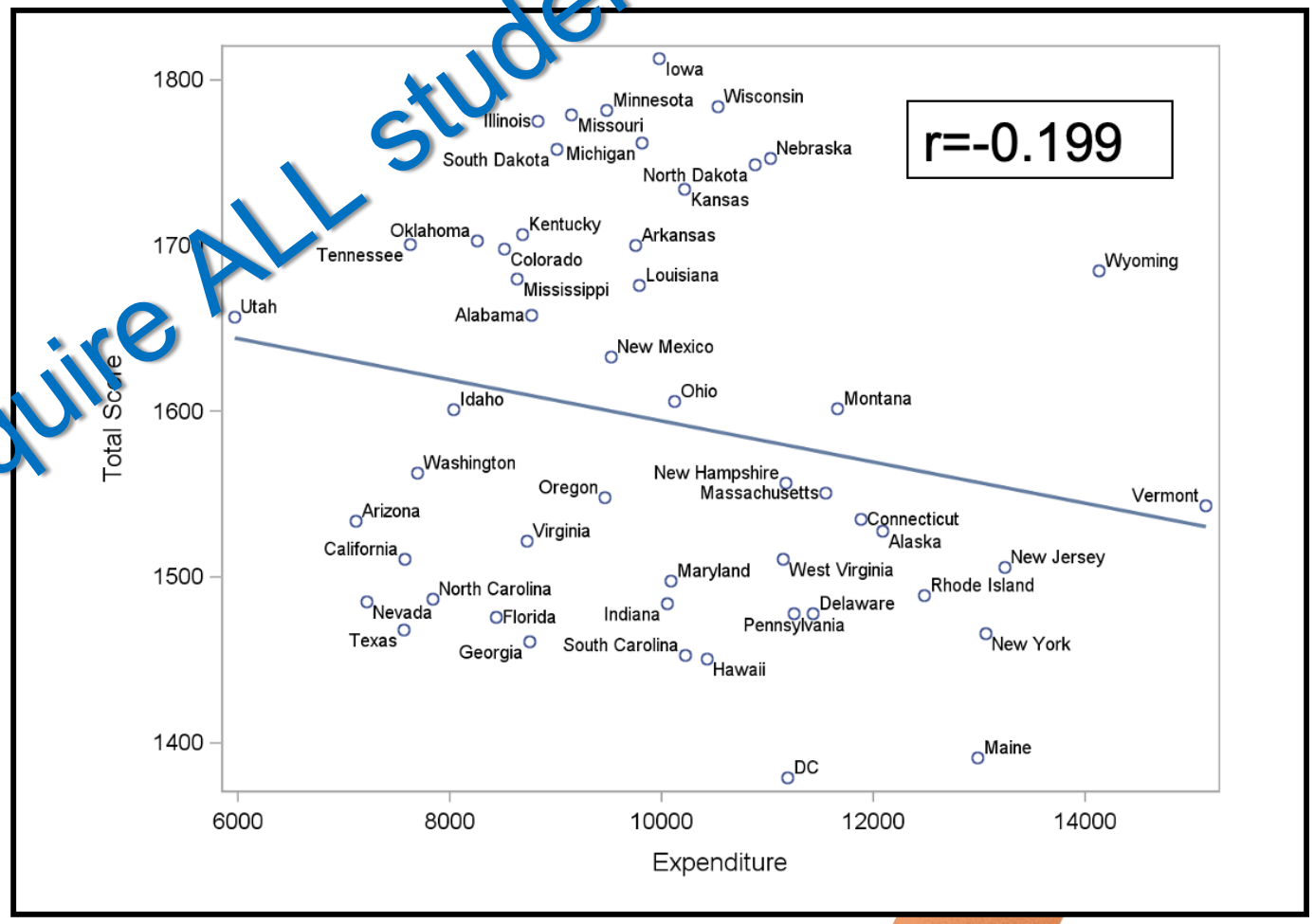
Correlation does NOT imply causation

US Department of Education:
“Expenditure has negative impact on SAT scores” ?



Correlation does NOT imply causation

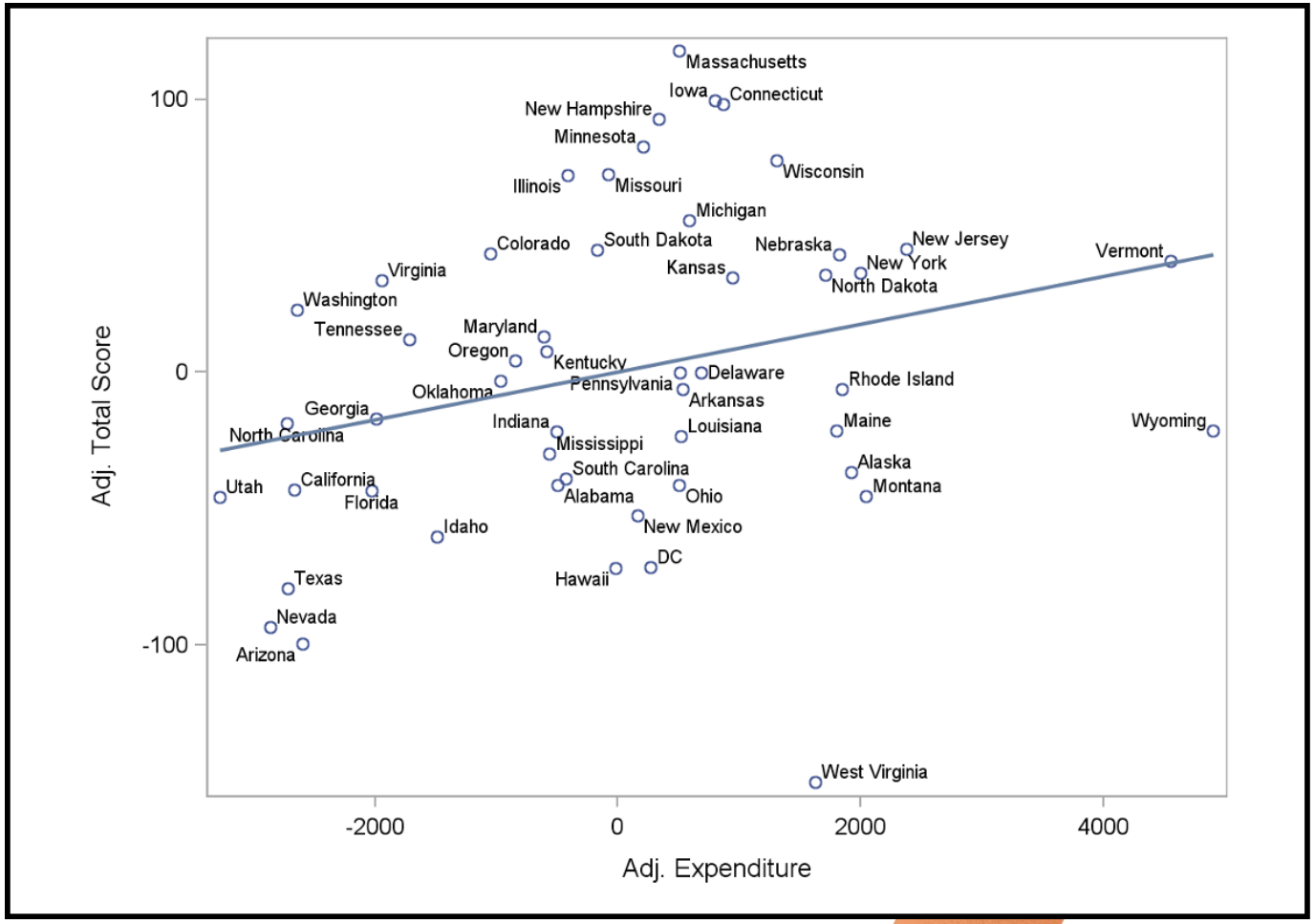
US Department of Education:
"Expenditure has negative impact on SAT scores" ?



Some states require ALL students to take SAT exams

Correlation does NOT imply causation

More Accurate Story:
Once you adjust for the participation rate, you find this positive relationship between expenditure and total score

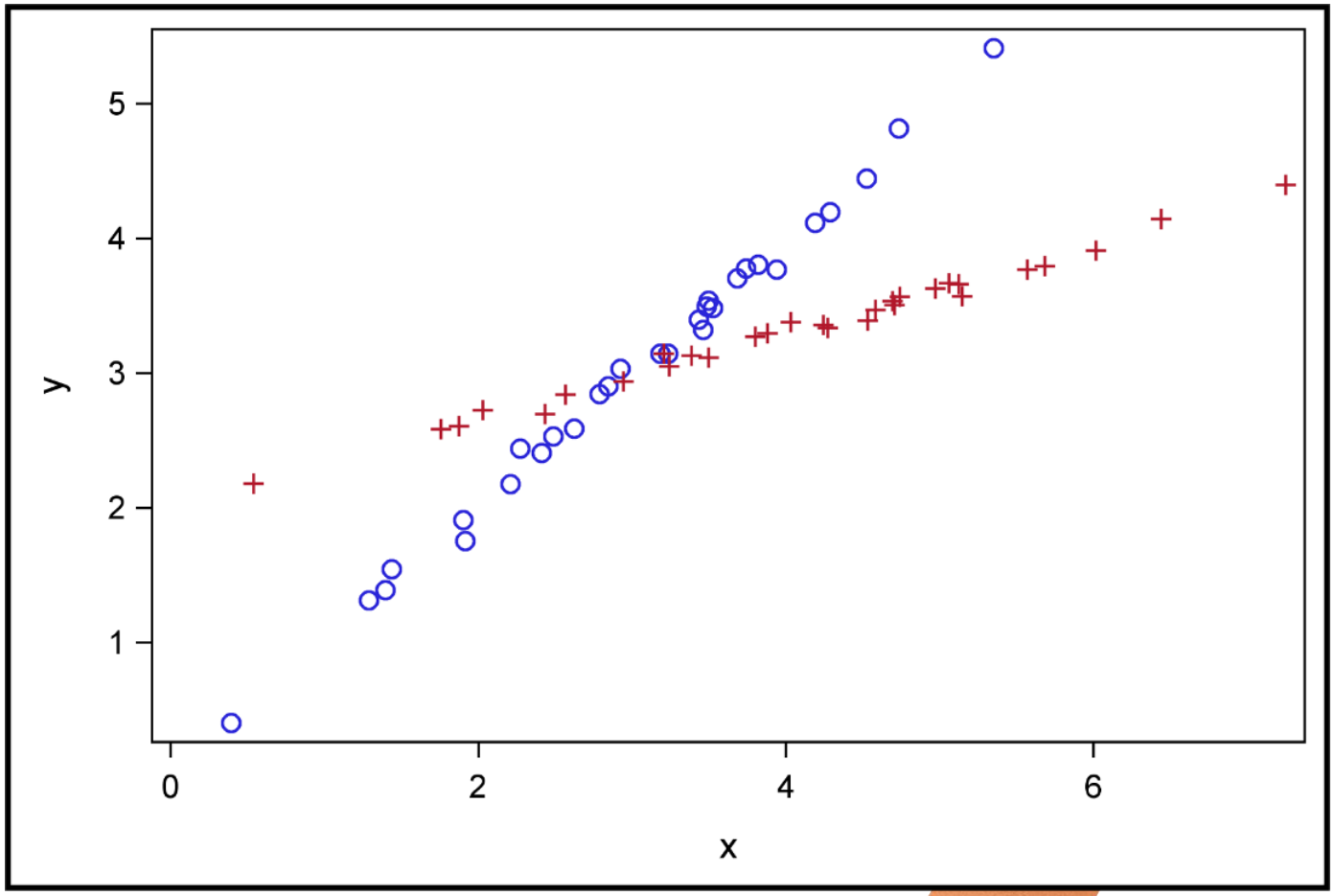


Simple Linear Regression

SLR - Correlation ≠ Slope!

$$y = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Two pairs of variables may have the same correlation coefficient, but different linear relationships.

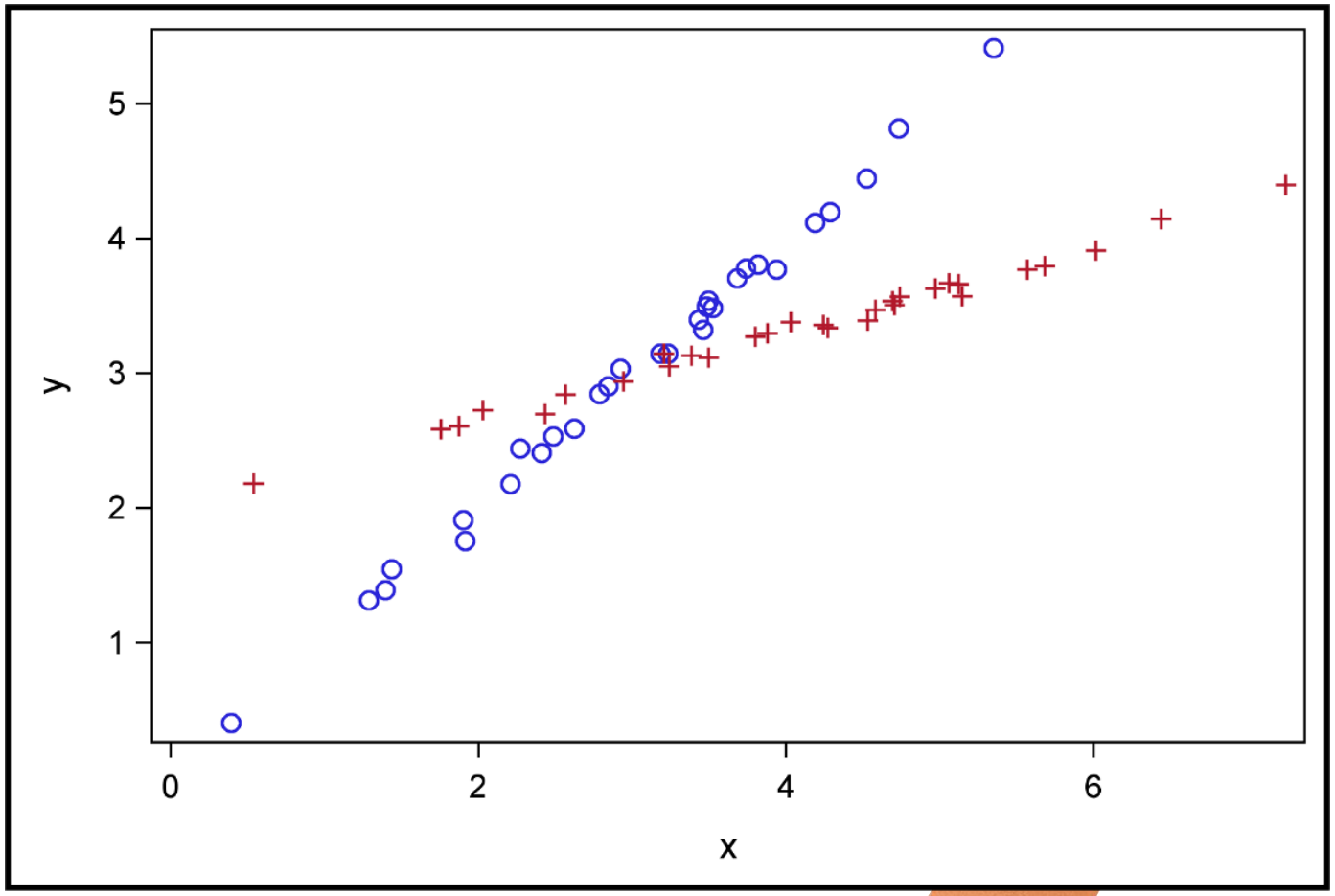


SLR - Correlation ≠ Slope!

$$y = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Deterministic component

Two pairs of variables may have the same correlation coefficient, but different linear relationships.

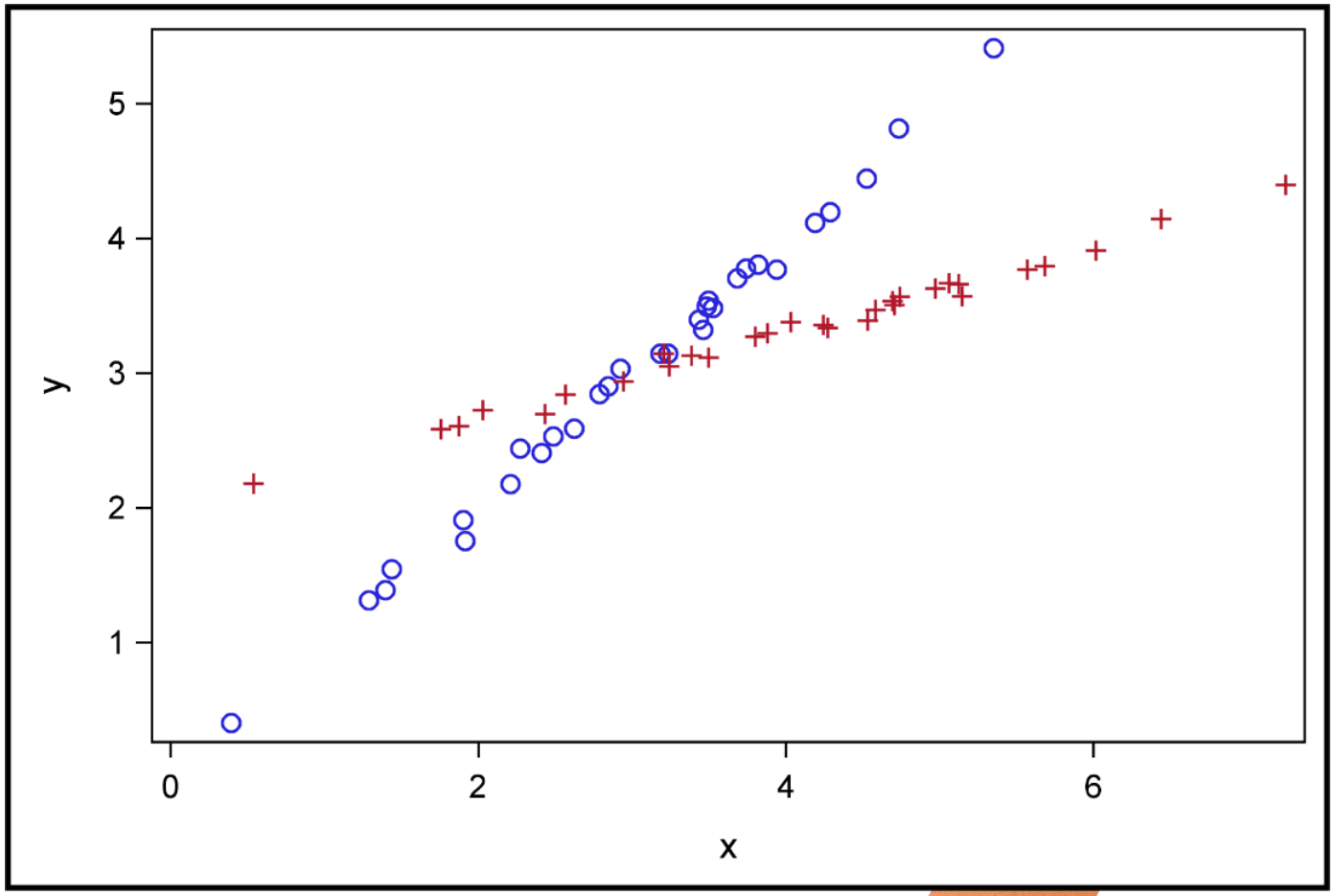


SLR - Correlation ≠ Slope!

$$y = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Random component

Two pairs of variables may have the same correlation coefficient, but different linear relationships.

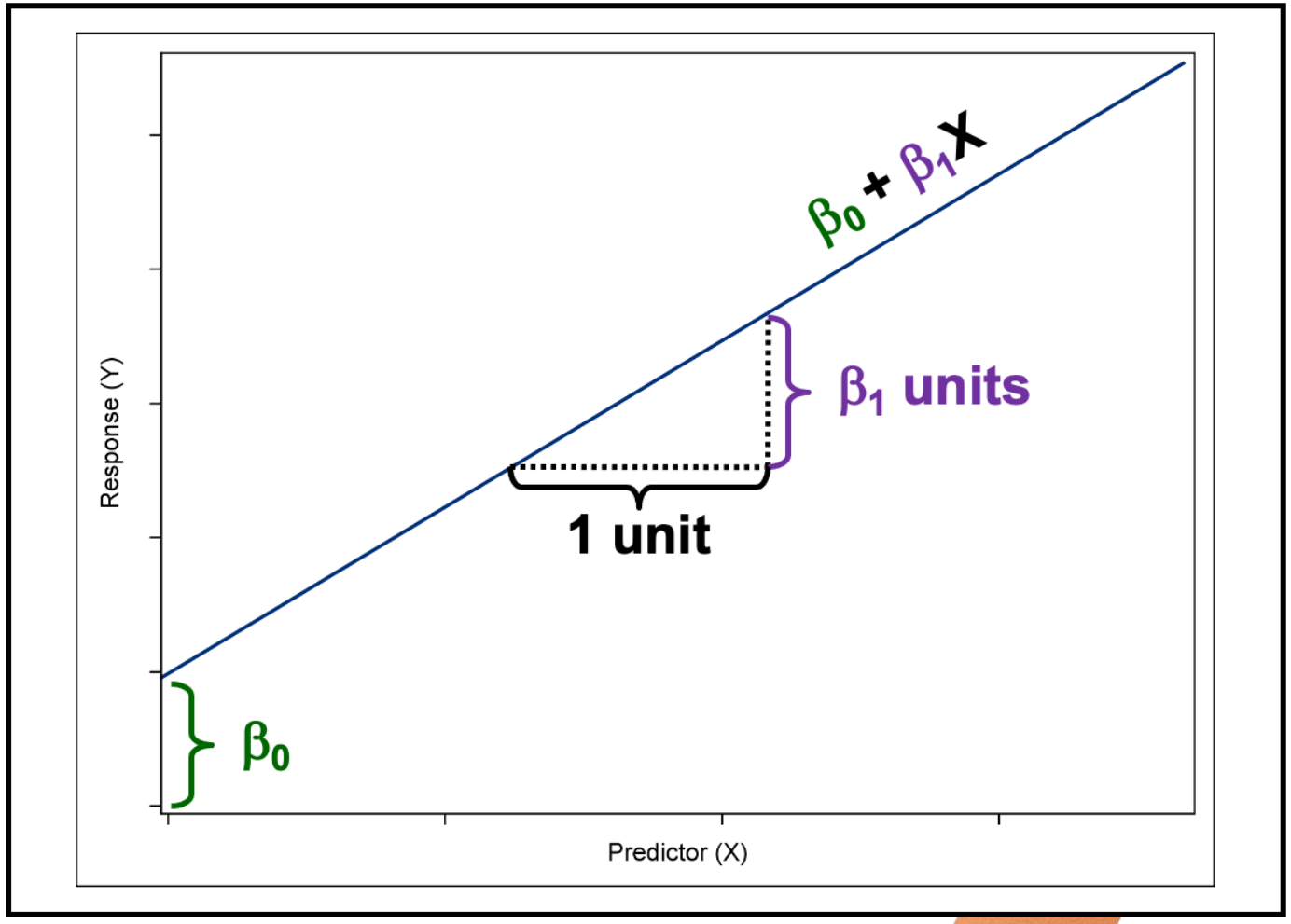


Simple Linear Regression Analysis

Population Parameters
(Unknown "True" Relationship)

β_0 is the ***intercept***

β_1 is the ***slope***

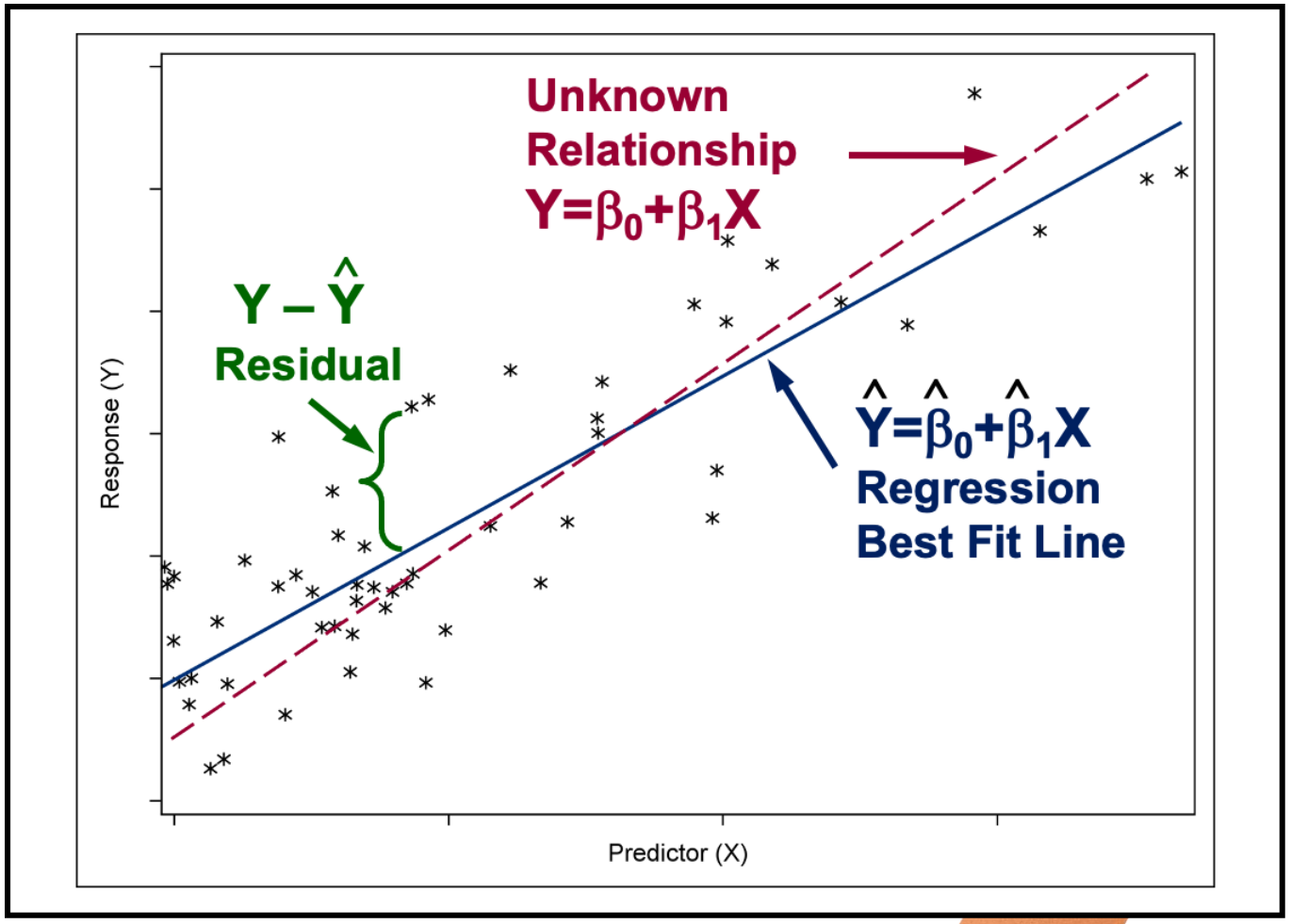


Simple Linear Regression Analysis

Sample Statistics

$\hat{\beta}_0$ is the ***intercept estimate***

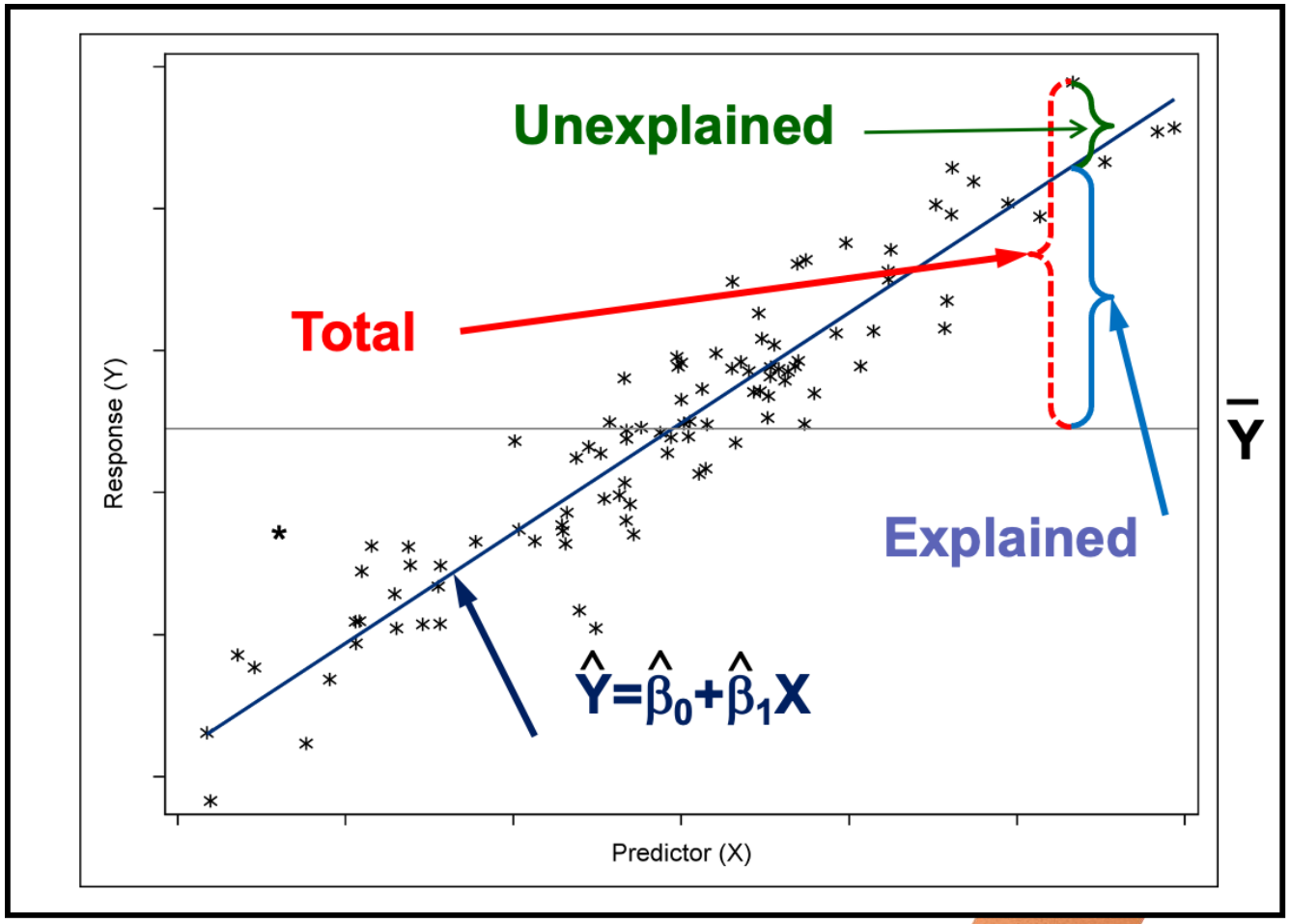
$\hat{\beta}_1$ is the ***slope estimate***



Explained vs. Unexplained Variability

Our objective is to explain the variation in the response variable.

We'll never be able to explain *all* of it, because some is due to random, uncontrollable error (or unobserved factors).

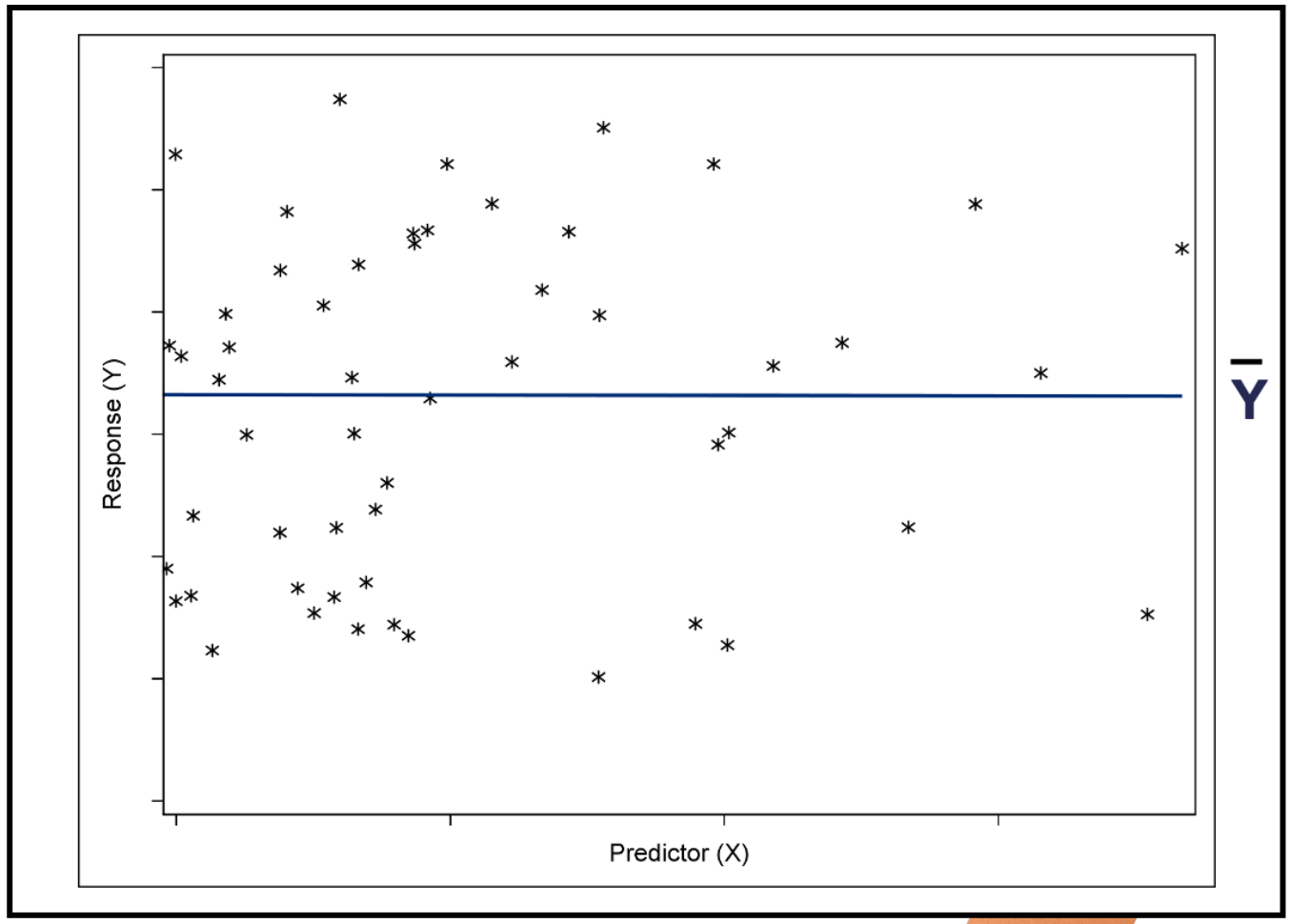


The Baseline Model (Null Hypothesis)

$H_0 \quad \beta_1 = 0$

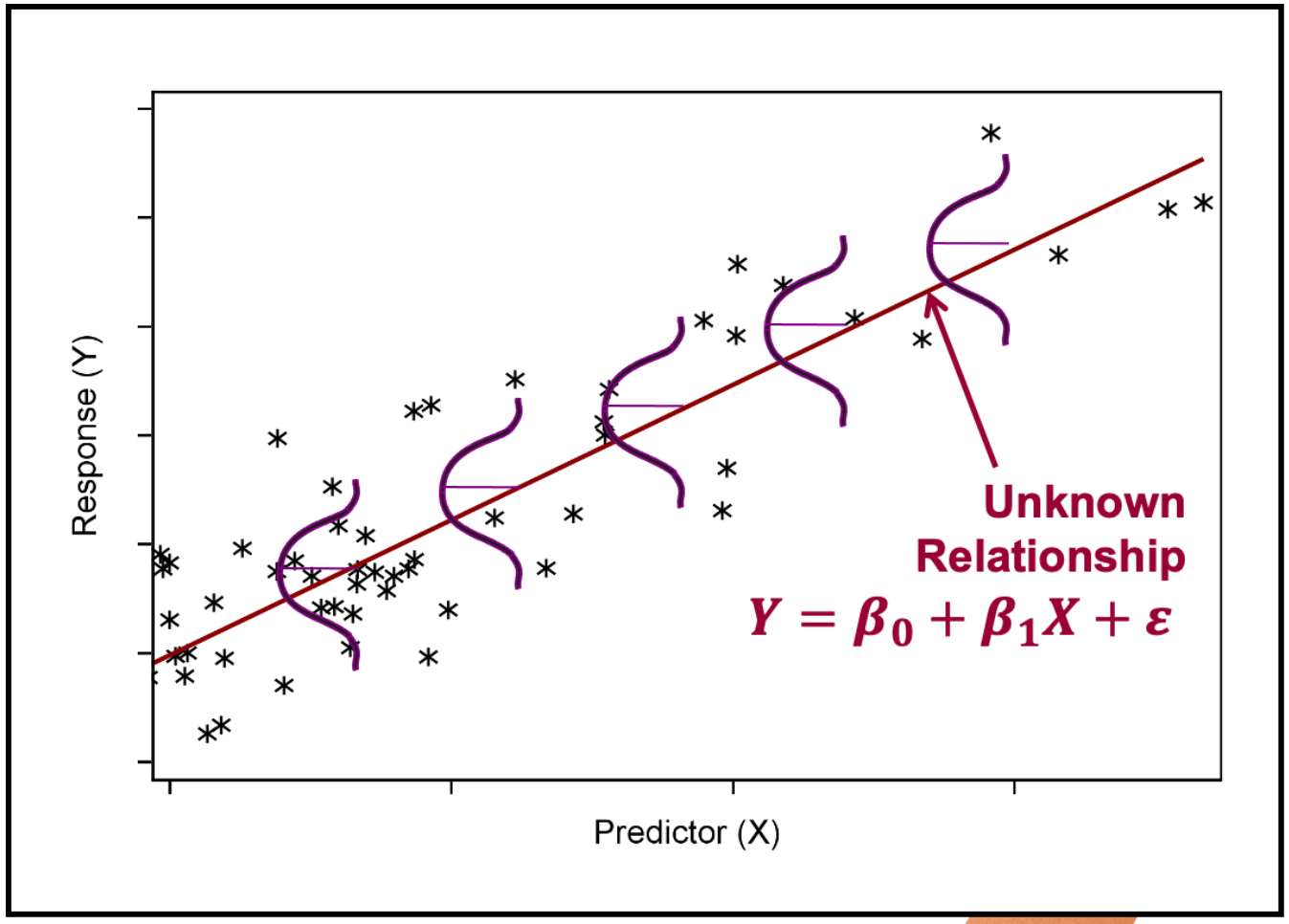
$H_a \quad \beta_1 \neq 0$

For Simple Linear Regression,
The global F-Test, the parameter t-test,
and the test of Pearson's correlation
are all equivalent!



4 Assumptions of Simple Linear Regression

1. Linearity of the mean
2. Errors are normally distributed
3. Errors have equal variance
4. Errors are independent



Testing of assumptions

Normality:

Histogram, QQ-plot or normality test (on residuals)

Equal variances:

Residual plot (residuals versus predicted values – want a nice “band” across all predicted values)

Independence:

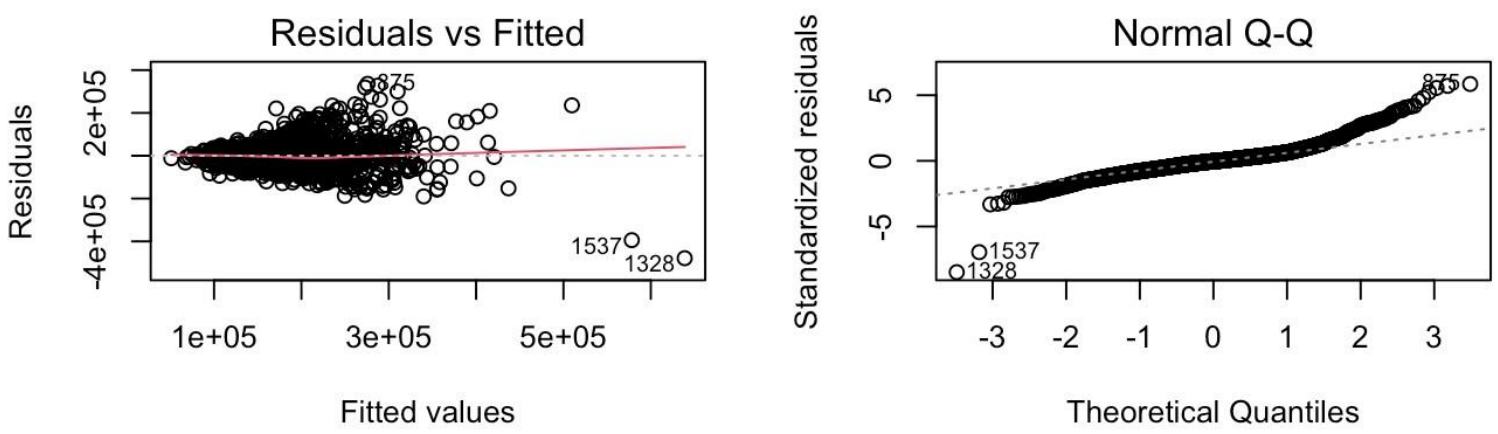
Data collection (can look at residual plots for a autocorrelation….will be discussed later)

Linearity in the mean:

Residual plot (no pattern in residual plot!! More to come on this later)

Simple Linear Regression in R

```
slr <- lm(Sale_Price ~ Gr_Liv_Area, data=train)
par(mfrow=c(2,2))
plot(slr)
```




```
summary(slr)
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14045.872    3942.503    3.563 0.000375 ***
## Gr_Liv_Area    110.726       2.506   44.185  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 57430 on 2049 degrees of freedom
## Multiple R-squared:  0.4879, Adjusted R-squared:  0.4877
## F-statistic:  1952 on 1 and 2049 DF,  p-value: < 2.2e-16
```

Break out and Lab 4

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