



More Complex ANOVA & Regression

Institute for Advanced Analytics
MSA Class of 2025

Two-Way ANOVA

n -Way ANOVA

Continuous Target Variable

One-Way ANOVA

- 1 variable
- k categories

Two-Way ANOVA

- 2 variables
- k_1 and k_2 categories

⋮

n -Way ANOVA

- n variables
- k_1, k_2, \dots, k_n categories

Additional Linear Models Terminology

- **Model** – a mathematical relationship between explanatory variables and response variables
- **Effect** – the expected change in the response that occurs with a 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** – next section

Exploring the Data

- Similar to One-Way ANOVA, we need to explore variables to add to our generalized ANOVA model.
- Previously looked at Heating Quality.
- Want to also look at Central Air availability.

Exploring the Data

```
train %>%  
  group_by(Heating_QC, Central_Air) %>%  
  summarise(mean = mean(Sale_Price),  
            sd = sd(Sale_Price),  
            max = max(Sale_Price),  
            min = min(Sale_Price),  
            n = n())
```

Exploring the Data

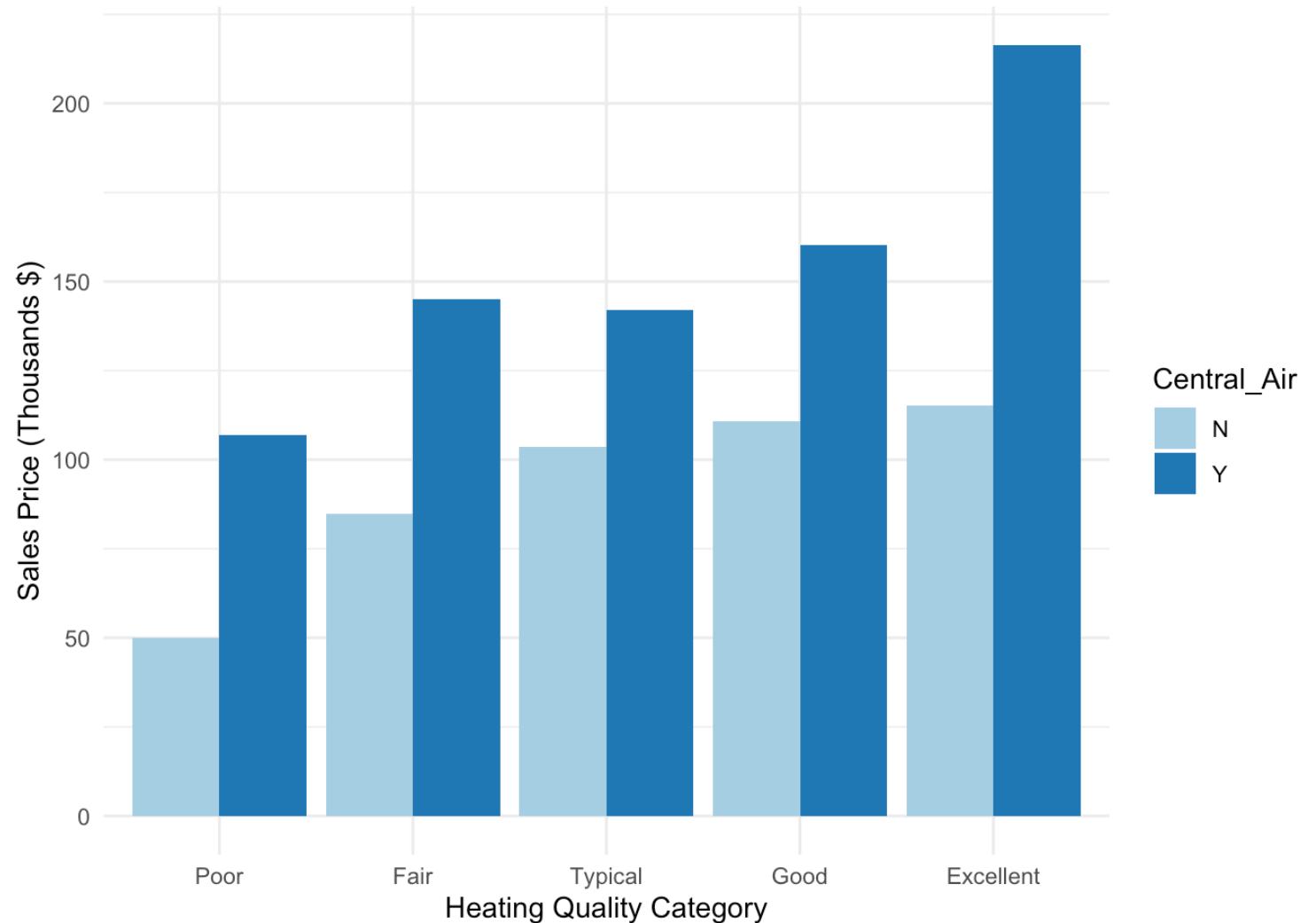
```
## # A tibble: 10 x 6
## # Groups:   Heating_QC [5]
##   Heating_QC Central_Air mean      sd    max    min     n
##   <ord>      <fct>      <dbl>  <dbl>  <int>  <int> <int>
## 1 Poor      N          50050  52255.  87000  13100     2
## 2 Poor      Y          107000    NA  107000  107000     1
## 3 Fair      N          84748.  28267.  158000  37900    29
## 4 Fair      Y          145165.  38624.  230000  50000    36
## 5 Typical   N          103469.  34663.  209500  12789    82
## 6 Typical   Y          142003.  39657.  375000  60000   527
## 7 Good      N          110811.  38455.  214500  59000    23
## 8 Good      Y          160113.  54158.  415000  52000   318
## 9 Excellent N          115062.  33271.  184900  64000    11
## 10 Excellent Y          216401.  88518.  745000  58500  1022
```

Exploring the Data

```
## # A tibble: 10 x 6
## # Groups:   Heating_QC [5]
##   Heating_QC Central_Air mean      sd    max    min     n
##   <ord>      <fct>    <dbl>  <dbl> <int>  <int> <int>
## 1 Poor      N      50050  52255.  87000  13100    2
## 2 Poor      Y      107000  NA     107000 107000    1
## 3 Fair      N      84748. 28267. 158000  37900   29
## 4 Fair      Y     145165. 38624. 230000  50000   36
## 5 Typical   N     103469. 34663. 209500  12789   82
## 6 Typical   Y     142003. 39657. 375000  60000  527
## 7 Good      N     110811. 38455. 214500  59000   23
## 8 Good      Y     160113. 54158. 415000  52000  318
## 9 Excellent N     115062. 33271. 184900  64000   11
## 10 Excellent Y     216401. 88518. 745000  58500 1022
```


Exploring the Data

- Appears to have differences between heating quality levels as well as presence of central air.
- Need statistical proof.



Two-Way ANOVA

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

Sales Price Base Level Heating Quality Central Air Unexplained Error

Two-Way ANOVA

```
ames_aov2 <- aov(Sale_Price ~ Heating_QC + Central_Air, data = train)
summary(ames_aov2)
```

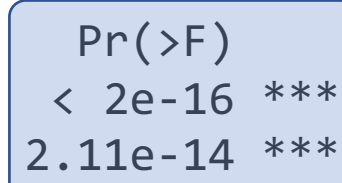
```
##              Df      Sum Sq   Mean Sq F value    Pr(>F)
## Heating_QC      4 2.891e+12 7.228e+11  147.60 < 2e-16 ***
## Central_Air     1 2.903e+11 2.903e+11   59.28 2.11e-14 ***
## Residuals    2045 1.002e+13 4.897e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Two-Way ANOVA

```
ames_aov2 <- aov(Sale_Price ~ Heating_QC + Central_Air, data = train)
summary(ames_aov2)
```

```
##              Df      Sum Sq   Mean Sq F value    Pr(>F)
## Heating_QC      4 2.891e+12 7.228e+11  147.60 < 2e-16 ***
## Central_Air     1 2.903e+11 2.903e+11   59.28 2.11e-14 ***
## Residuals    2045 1.002e+13 4.897e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Each variable has
F test



Variable	F value	Pr(>F)
Heating_QC	147.60	< 2e-16 ***
Central_Air	59.28	2.11e-14 ***

Two-Way ANOVA

```
ames_aov2 <- aov(Sale_Price ~ Heating_QC + Central_Air, data = train)
summary(ames_aov2)
```

5-1 = 4 for Heating Quality
2-1 = 1 for Central Air



	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Heating_QC	4	2.891e+12	7.228e+11	147.60	< 2e-16	***
Central_Air	1	2.903e+11	2.903e+11	59.28	2.11e-14	***
Residuals	2045	1.002e+13	4.897e+09			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

Post-Hoc Testing

- Similar to One-Way ANOVA, if we have statistical differences among the categories, we want to know where these statistical differences exist.
- Use same approaches as before.

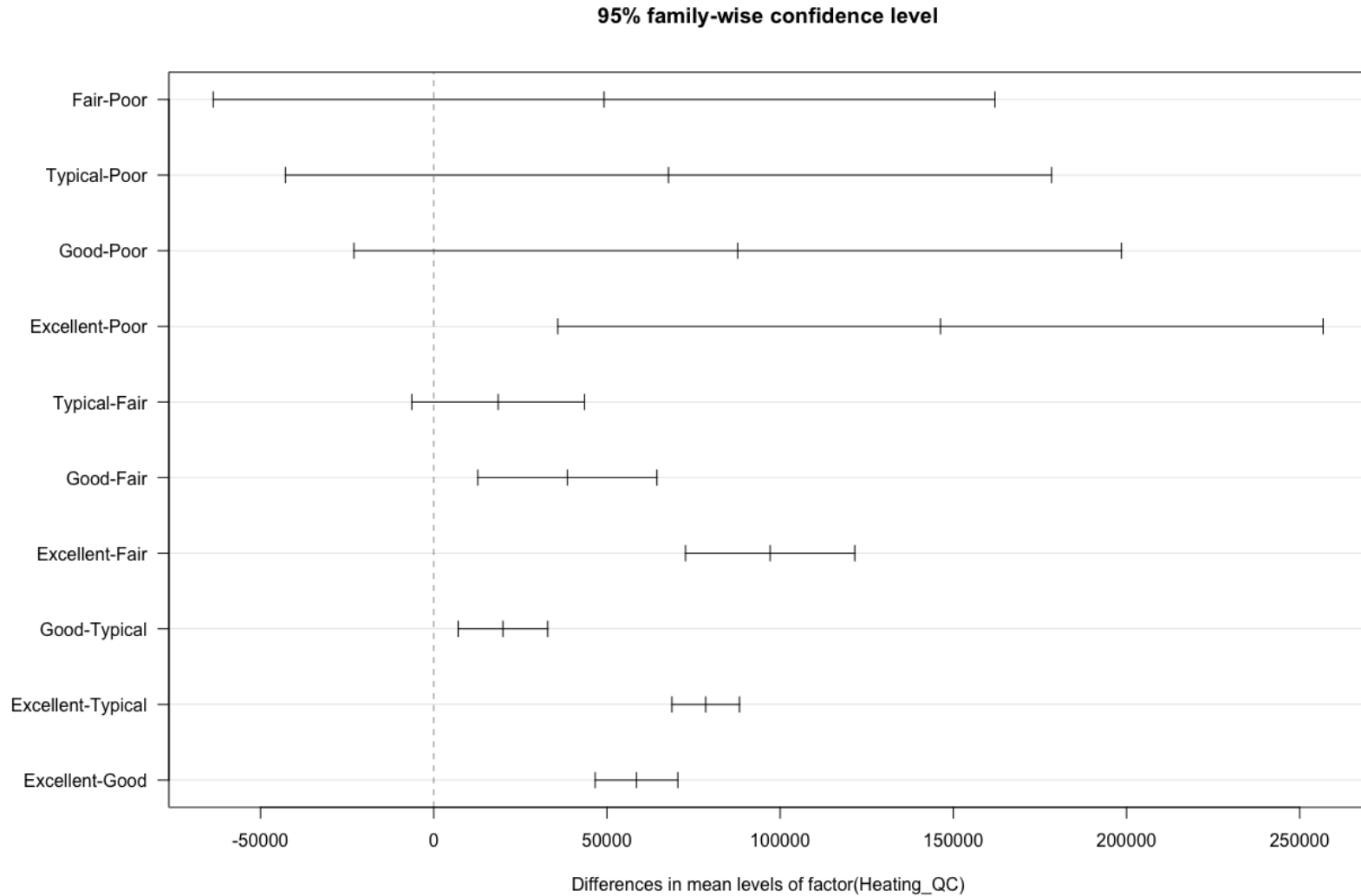
Post-Hoc Testing

```
tukey.ames2 <- TukeyHSD(ames_aov2)
print(tukey.ames2)
plot(tukey.ames2)
```

Post-Hoc Testing

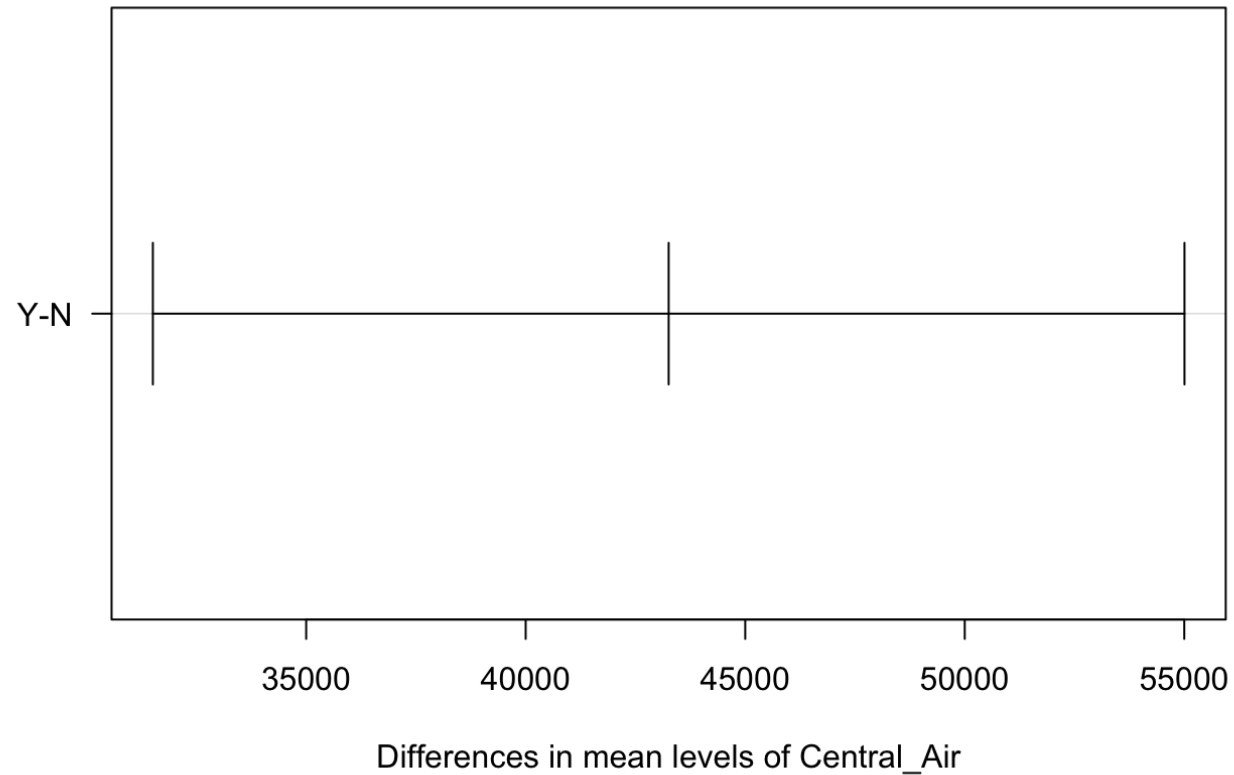
```
## $Heating_QC
##           diff           lwr           upr           p adj
## Fair-Poor      49176.42 -63650.448 162003.29 0.7571980
## Typical-Poor    67781.01 -42800.320 178362.35 0.4506761
## Good-Poor       87753.89 -23040.253 198548.03 0.1945181
## Excellent-Poor 146288.89  35818.859 256758.92 0.0028361
## Typical-Fair    18604.59  -6326.425  43535.61 0.2484556
## Good-Fair       38577.47  12718.894  64436.04 0.0004622
## Excellent-Fair  97112.47  72679.867 121545.07 0.0000000
## Good-Typical    19972.87   7050.230  32895.52 0.0002470
## Excellent-Typical 78507.88  68746.678  88269.07 0.0000000
## Excellent-Good  58535.00  46602.229  70467.78 0.0000000
##
## $Central_Air
##           diff           lwr           upr           p adj
## Y-N 43256.57 31508.27 55004.87 0
```


Post-Hoc Testing



Post-Hoc Testing

95% family-wise confidence level





Two-Way ANOVA with Interactions

Additional Linear Models Terminology

- **Model** – a mathematical relationship between explanatory variables and response variables
- **Effect** – the expected change in the response that occurs with a 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 one variable changes as levels of another variable changes (for example, $x_1 \times x_2, x_1 \times x_2 \times x_3$)

Two-Way ANOVA with Interactions

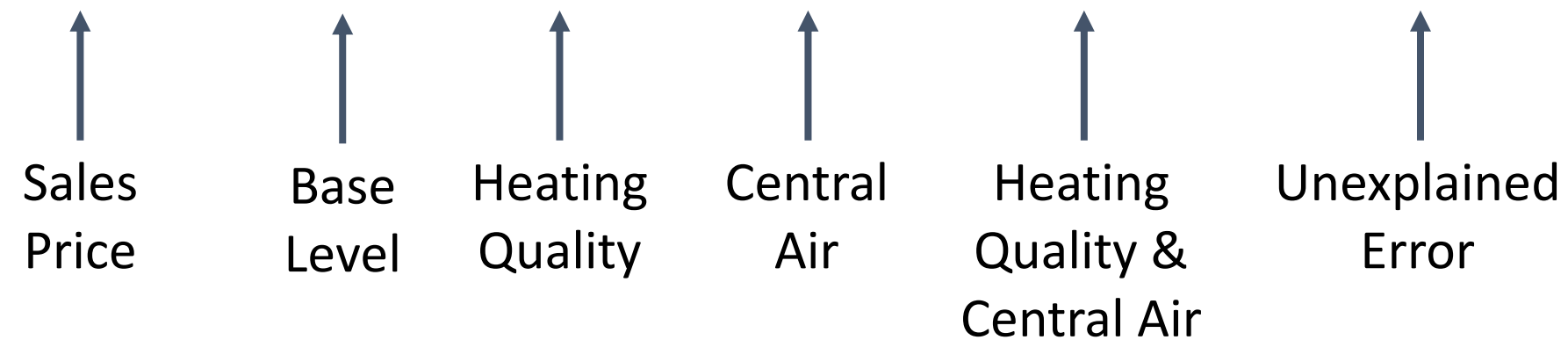
$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$


Diagram illustrating the components of the Two-Way ANOVA model with interactions:

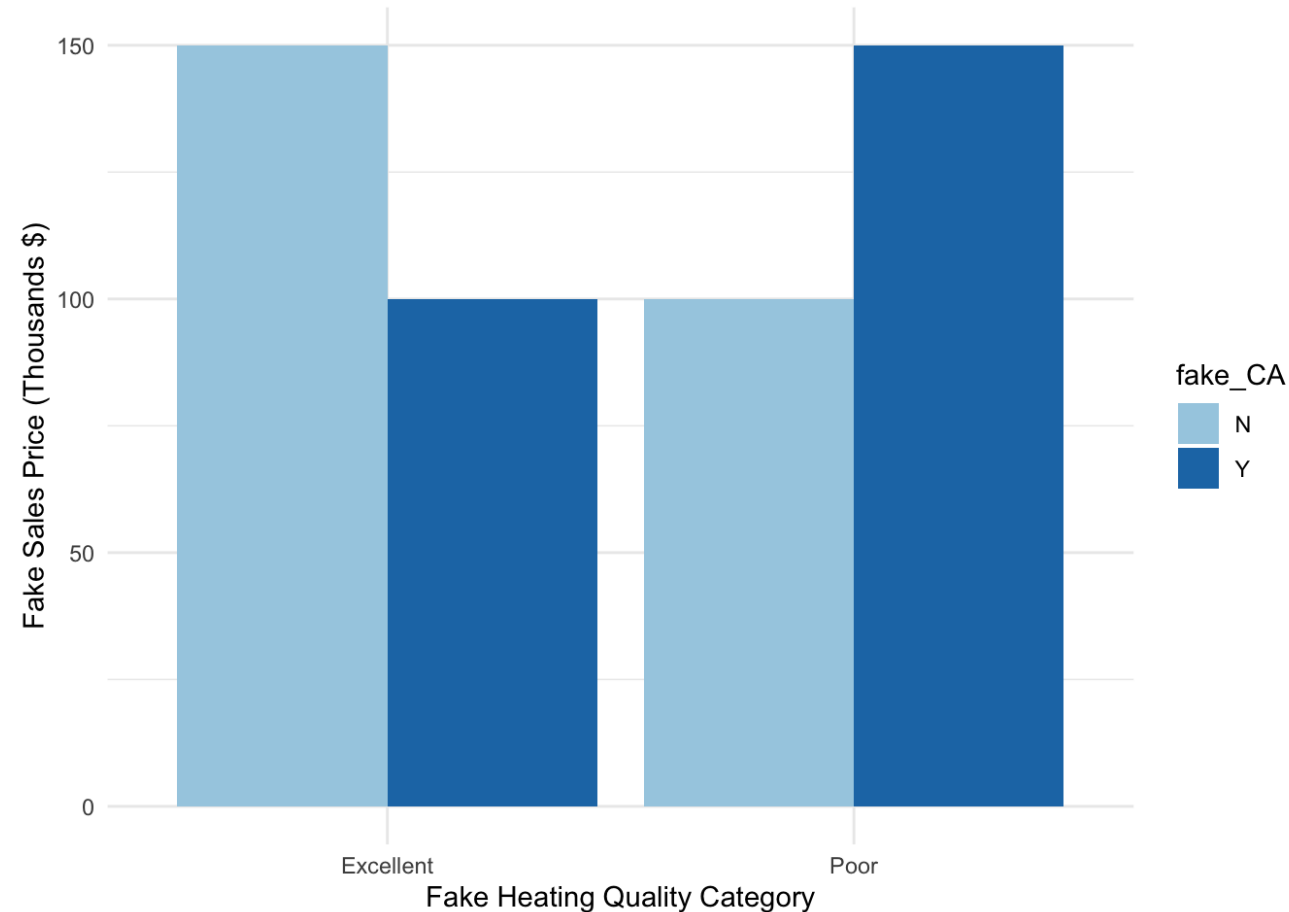
- Y_{ijk} : Sales Price
- μ : Base Level
- α_i : Heating Quality
- β_j : Central Air
- $(\alpha\beta)_{ij}$: Heating Quality & Central Air (Interaction)
- ε_{ijk} : Unexplained Error

Interactions

- For our model, an interaction between Central Air and Heating Quality would imply both of the following:
 - The impact of Heating Quality on Sale Price differs across levels of Central Air (ex. Difference in price between Excellent and Poor HQ changes if Central Air is or is **not** present)
 - The impact of Central Air on Sale Price differs across levels of Heating Quality (ex. Difference in price between having and not having Central Air changes across levels of HQ)

Interactions – Caution

- Interactions can potentially **mask** the effects of the variables.



Interactions

```
ames_aov_int <- aov(Sale_Price ~ Heating_QC*Central_Air, data = train)
summary(ames_aov_int)
```

Interactions

```
ames_aov_int <- aov(Sale_Price ~ Heating_QC*Central_Air, data = train)
```

```
summary(ames_aov_int)
```

Same

```
Heating_QC + Central_Air + Heating_QC:Central_Air
```

Interactions

```
ames_aov_int <- aov(Sale_Price ~ Heating_QC*Central_Air, data = train)
summary(ames_aov_int)
```

```
##              Df      Sum Sq   Mean Sq F value    Pr(>F)
## Heating_QC      4 2.891e+12 7.228e+11 147.897 < 2e-16 ***
## Central_Air      1 2.903e+11 2.903e+11  59.403 1.99e-14 ***
## Heating_QC:Central_Air  4 3.972e+10 9.930e+09   2.032  0.0875 .
## Residuals    2041 9.975e+12 4.887e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interactions

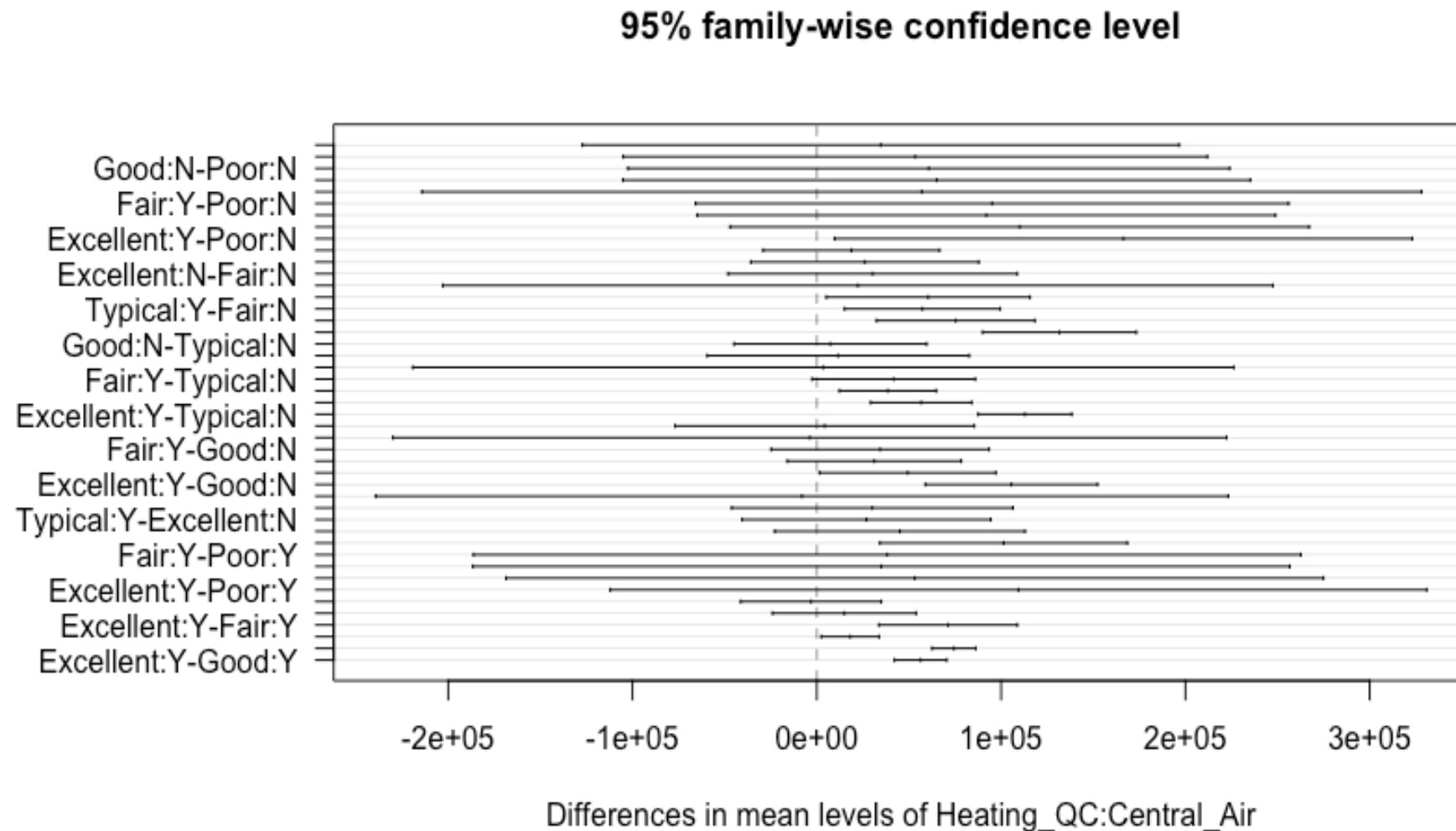
```
ames_aov_int <- aov(Sale_Price ~ Heating_QC*Central_Air, data = train)
summary(ames_aov_int)
```

```
##              Df      Sum Sq   Mean Sq F value    Pr(>F)
## Heating_QC      4 2.891e+12 7.228e+11 147.897 < 2e-16 ***
## Central_Air     1 2.903e+11 2.903e+11  59.403 1.99e-14 ***
## Heating_QC:Central_Air 4 3.972e+10 9.930e+09   2.032  0.0875 .
## Residuals    2041 9.975e+12 4.887e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Post-hoc Testing

```
tukey.ames_int <- TukeyHSD(ames_aov_int)  
plot(tukey.ames_int, las = 1)
```

Post-hoc Testing



Within Effects Testing (Slicing)

- If an interaction exists, we are probably curious to see which of the levels of one variable are different within the level of another variable.
- Testing every possible combination might be overwhelming.
- **Slicing** performs an F-test for means for one variable within the level of another variable.

Sliced ANOVA

```
CA_aov <- train %>%  
  group_by(Central_Air) %>%  
  nest() %>%  
  mutate(aov = map(data, ~summary(aov(Sale_Price ~ Heating_QC, data = .x))))  
print(CA_aov$aov)
```


Sliced ANOVA

```
## [[1]]
##           Df      Sum Sq   Mean Sq F value   Pr(>F)
## Heating_QC    4 2.242e+12 5.606e+11   108.5 <2e-16 ***
## Residuals  1899 9.809e+12 5.165e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## [[2]]
##           Df      Sum Sq   Mean Sq F value   Pr(>F)
## Heating_QC    4 1.774e+10 4.435e+09    3.793 0.00582 **
## Residuals   142 1.660e+11 1.169e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Central Air – YES

Heating Quality levels
different

Central Air – NO

Heating Quality levels
different

Assumptions

- The assumptions of n -Way ANOVA are the same as with One-Way ANOVA.
 - Independence of observations
 - Equality of variance
 - Normality of categories

Assumptions

- The assumptions of n -Way ANOVA are the same as with One-Way ANOVA.
 - Independence of observations
 - Equality of variance – Levene Test only available for interactions
 - Normality of categories

Assumptions

- The assumptions of n -Way ANOVA are the same as with One-Way ANOVA.
 - Independence of observations
 - Equality of variance (of errors from model)
 - Normality of categories (or errors from model)

Since ANOVA is essentially a linear regression, can use diagnostic approaches of linear regression to assess.

Discussed in later section of course!



Randomized Block Design with ANOVA

Observational or Retrospective Studies

- Groups can be naturally occurring.
 - Ex: Gender and ethnicity
- Random assignment might be unethical or untenable.
 - Ex: 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.

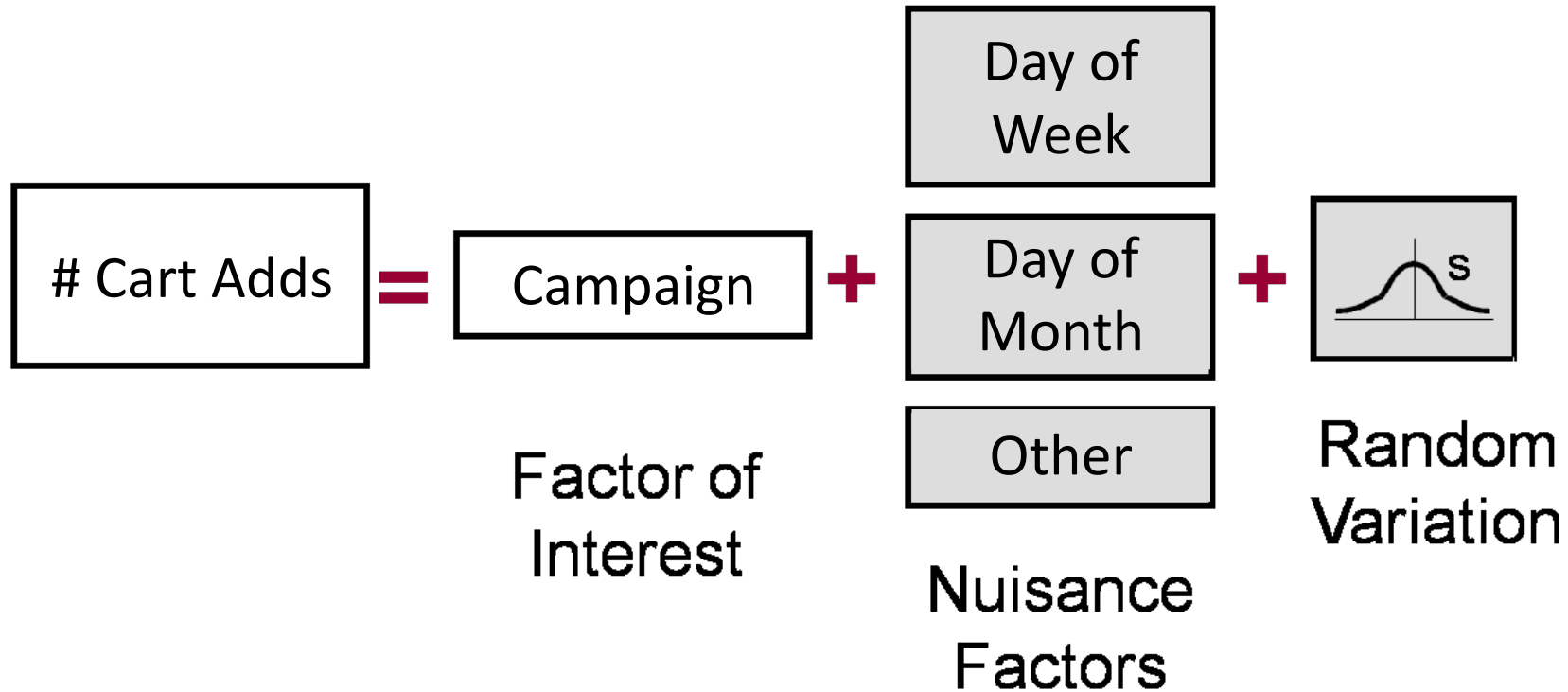
Controlled Experiments

- 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 **nuisance factors** contributing to the outcome measure.

Marketing Data

- This dataset contains many metrics from website usage from different marketing campaigns across a month.
- Compare the effects of different marketing campaigns on average number of times someone added our product to their virtual shopping cart.
- Potential nuisance factors:
 - Day of the week
 - Day of the month
 - External time factors
- **Blocking** to account for these nuisance factors.

Nuisance Factors



Assigning Treatments within Blocks

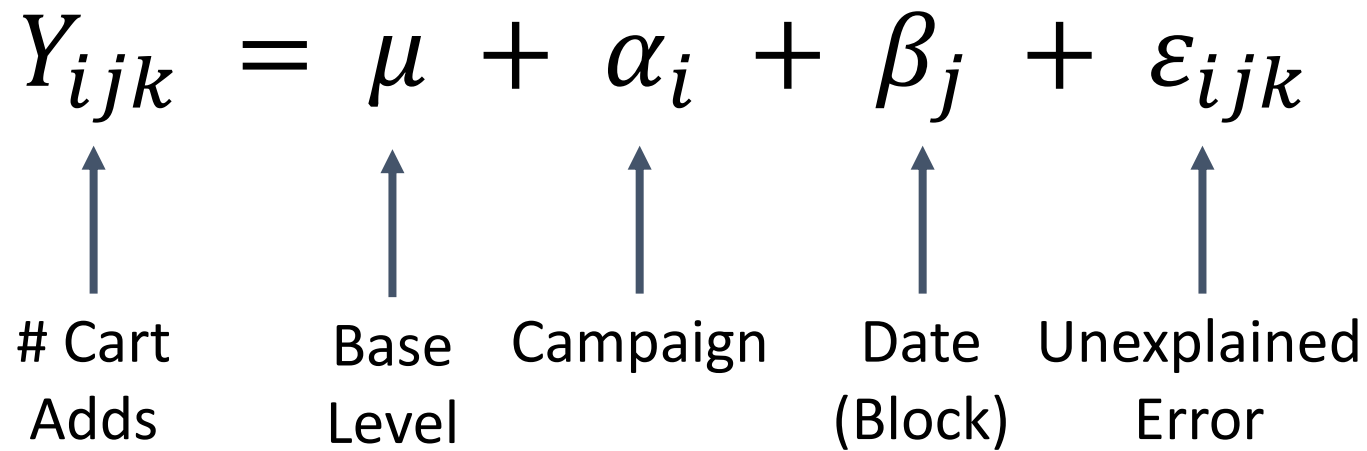
Aug 1 st	Person 1, 5, 7, ...	Person 2, 4, 8, ...	Person 3, 6, 9, ...
Aug 2 nd	Person 1, 6, 9, ...	Person 2, 5, 8, ...	Person 3, 4, 7, ...
Aug 3 rd	Person 1, 3, 9, ...	Person 2, 5, 8, ...	Person 4, 6, 7, ...
Aug 4 th	Person 2, 4, 5, ...	Person 6, 8, 9, ...	Person 1, 3, 7, ...
Aug 5 th	Person 5, 8, 9, ...	Person 1, 2, 6, ...	Person 3, 4, 7, ...

⋮

Exploring Marketing Data

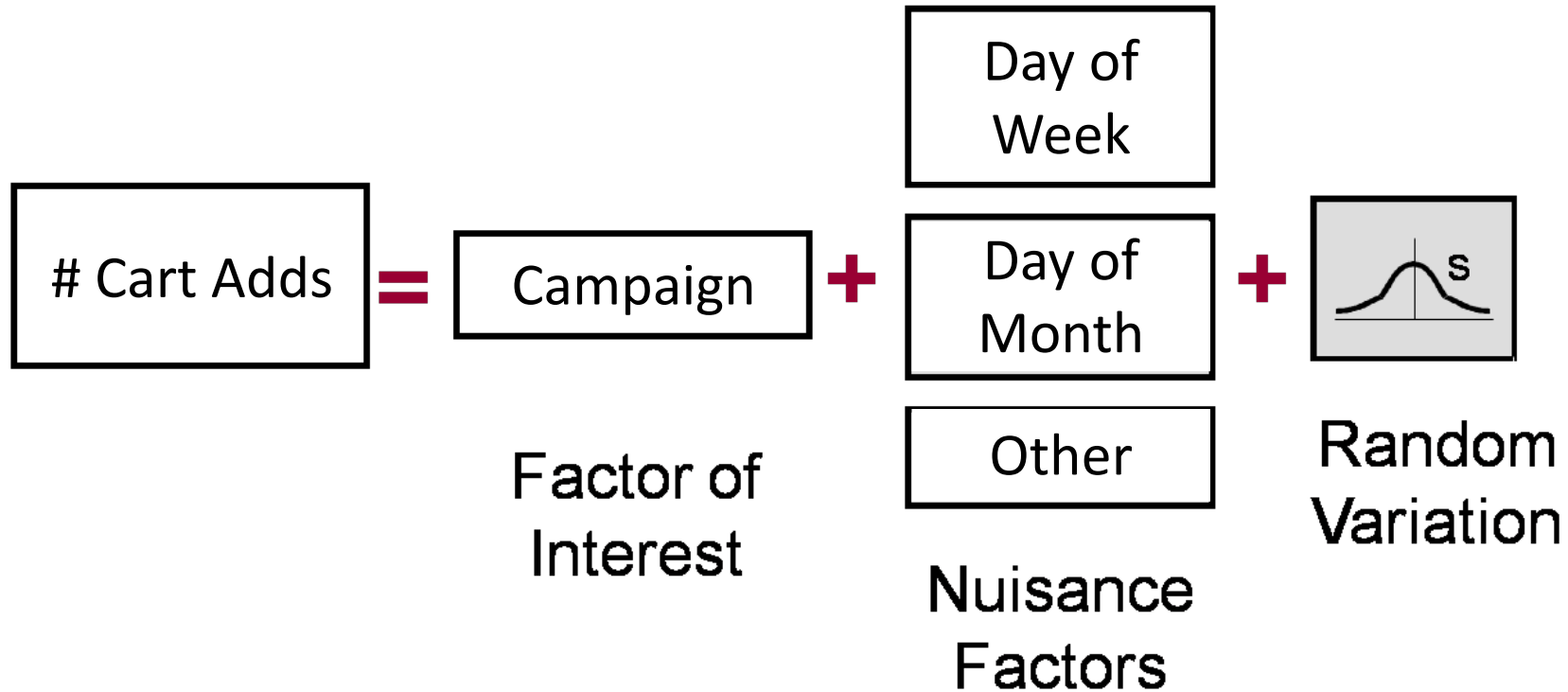
```
## # A tibble: 90 × 8
##   Campaign_Name Date Spend_USD Num_Cart_Adds Num_Purchase Num_Impressions
##   <chr> <chr> <dbl> <dbl> <dbl> <dbl>
## 1 Control Campaign 1.08.2... 3008 894 255 39550
## 2 Control Campaign 2.08.2... 2542 879 677 100719
## 3 Control Campaign 3.08.2... 2365 1268 578 70263
## 4 Control Campaign 4.08.2... 2710 566 340 78451
## 5 Control Campaign 5.08.2... 2297 956 768 114295
## 6 Control Campaign 6.08.2... 2458 882 488 42684
## 7 Control Campaign 7.08.2... 2838 1301 890 53986
## 8 Control Campaign 8.08.2... 2916 1240 431 33669
## 9 Control Campaign 9.08.2... 2652 1200 845 45511
## 10 Control Campaign 10.08.... 2790 424 275 95054
## # 80 more rows
## # 2 more variables: Num_Website_Clicks <dbl>, Num_Views <dbl>
```

Include Blocking Variable in Model

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


Cart Adds Base Level Campaign Date (Block) Unexplained Error

Nuisance Factors



ANOVA with Random Block Design

```
block_aov <- aov(Num_Cart_Adds ~ Campaign_Name + Date, data = block)
```

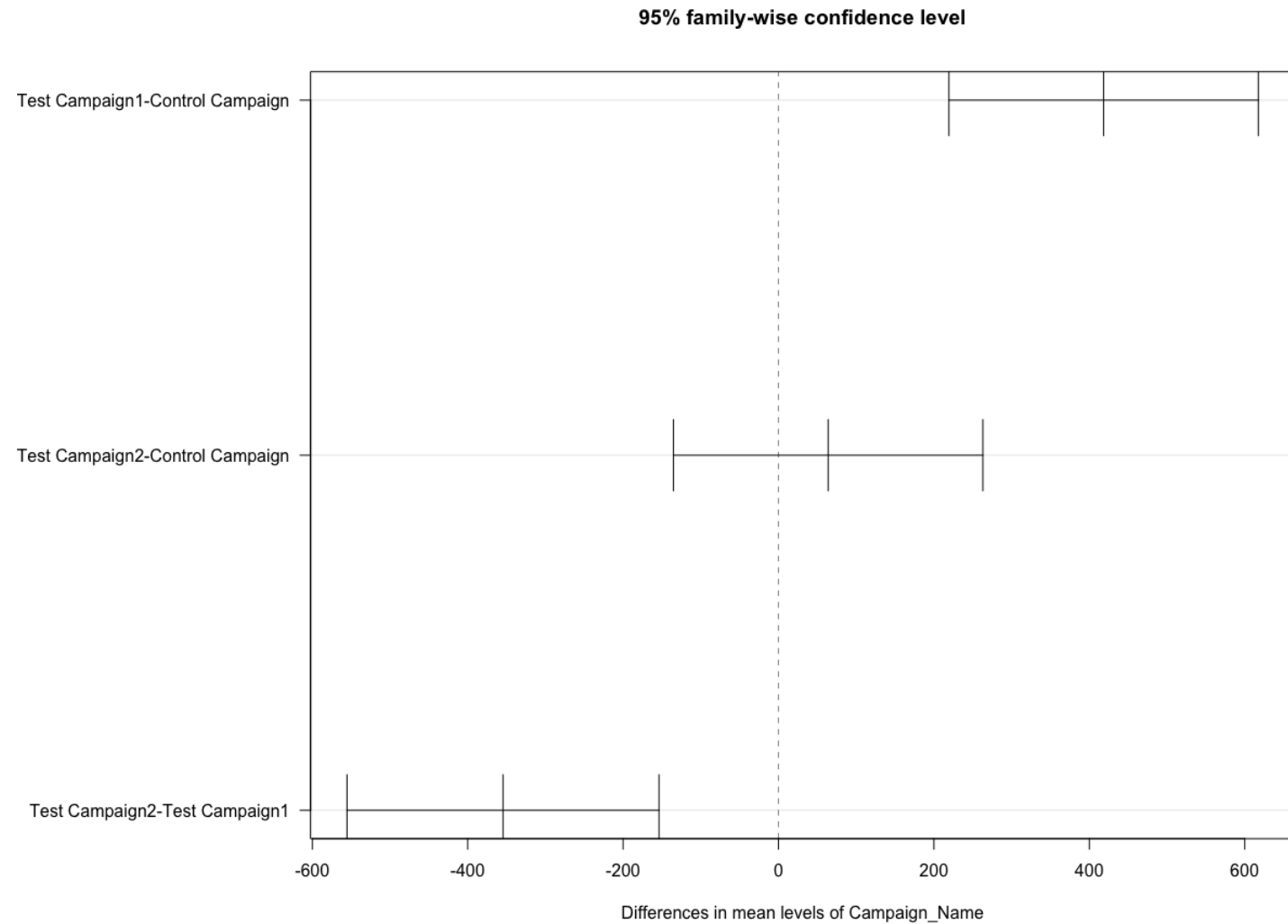
```
summary(block_aov)
```

```
##              Df  Sum Sq Mean Sq F value  Pr(>F)
## Campaign_Name  2 2973370 1486685  14.739 7.2e-06 ***
## Date          29 6315371  217771   2.159 0.00677 **
## Residuals     56 5648612  100868
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2 observations deleted due to missingness
```

Post-hoc Testing

```
tukey.block <- TukeyHSD(block_aov)  
plot(tukey.block, las = 1)
```


Post-hoc Testing



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 marketing example, the design is balanced, which means that there is the same number of customers samples for every **Campaign/Date** combination.

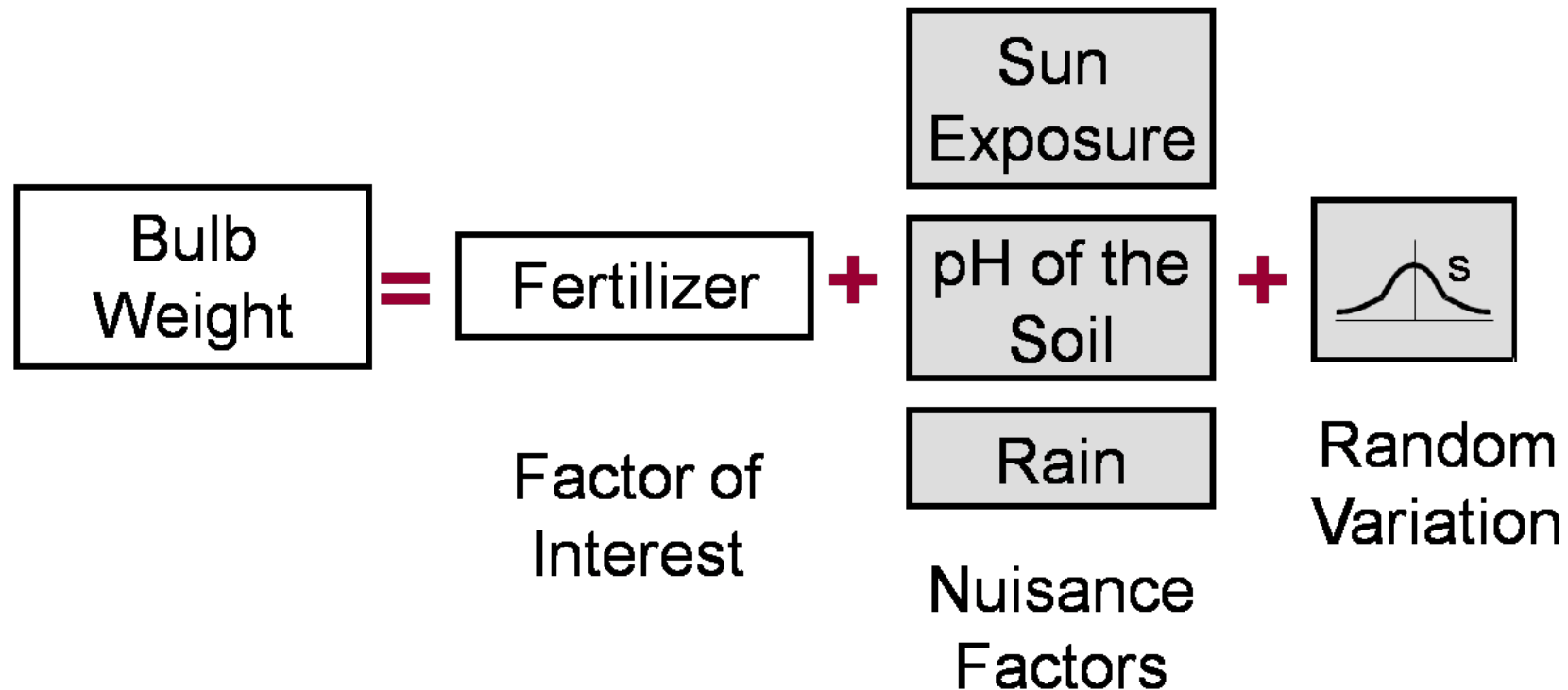
Additional Blocking Data Example

OPTIONAL SELF STUDY

Garlic Bulb Weight Data

- This dataset contains the average garlic bulb weight from different plots of land.
- Compare the effects of fertilizer on average bulb weight.
- Potential nuisance factors:
 - Sun exposure
 - pH for the soil
 - Rain amounts
- **Blocking** to account for these nuisance factors.

Nuisance Factors




Assigning Treatments within Blocks



Exploring Garlic Data

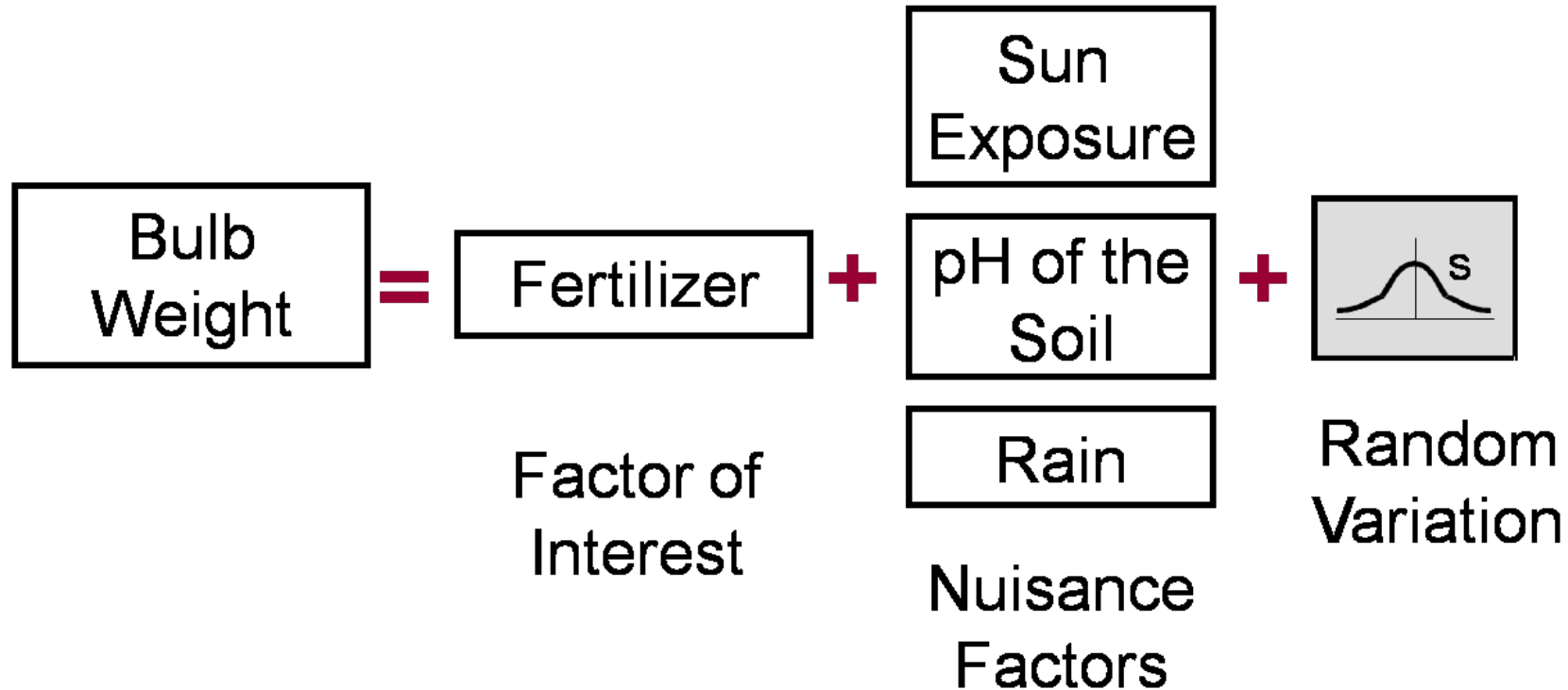
```
## # A tibble: 32 x 6
##   Sector Position Fertilizer BulbWt Cloves BedId
##   <dbl>    <dbl>    <dbl>  <dbl> <dbl> <dbl>
## 1      1      1      3    0.259   11.6 22961
## 2      1      2      4    0.207   12.6 23884
## 3      1      3      1    0.275   12.1 19642
## 4      1      4      2    0.245   12.1 20384
## 5      2      1      3    0.215   11.6 20303
## 6      2      2      4    0.170   12.7 21004
## 7      2      3      1    0.225   12.0 16117
## 8      2      4      2    0.168   11.9 19686
## 9      3      1      4    0.217   12.4 26527
## 10     3      2      3    0.226   11.7 23574
## # ... with 22 more rows
```

Include Blocking Variable in Model

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


Bulb Weight Base Level Fertilizer Sector (Block) Unexplained Error

Nuisance Factors



ANOVA with Random Block Design

```
block_aov <- aov(BulbWt ~ factor(Fertilizer) + factor(Sector), data = block)
summary(block_aov)
```

```
##              Df    Sum Sq   Mean Sq F value    Pr(>F)
## factor(Fertilizer)  3 0.005086 0.0016954    4.307 0.016222 *
## factor(Sector)      7 0.017986 0.0025695    6.527 0.000364 ***
## Residuals          21 0.008267 0.0003937
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA with Random Block Design

```
block_aov <- aov(BulbWt ~ factor(Fertilizer) + factor(Sector), data = block)
summary(block_aov)
```

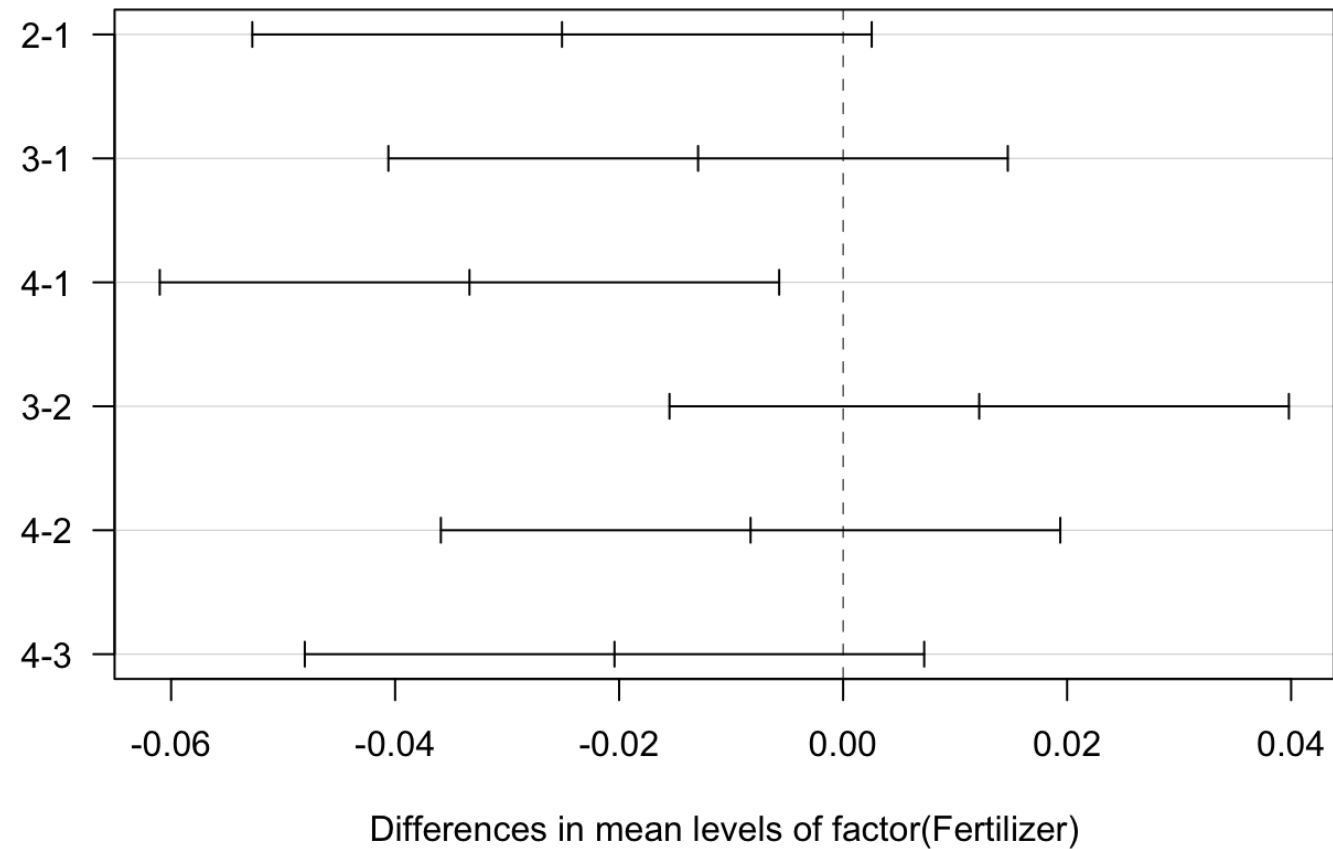
```
##              Df    Sum Sq   Mean Sq F value    Pr(>F)
## factor(Fertilizer)  3 0.005086 0.0016954    4.307 0.016222 *
## factor(Sector)      7 0.017986 0.0025695    6.527 0.000364 ***
## Residuals          21 0.008267 0.0003937
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Post-hoc Testing

```
tukey.block <- TukeyHSD(block_aov)  
plot(tukey.block, las = 1)
```

Post-hoc Testing

95% family-wise confidence level



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.



Multiple Linear Regression

CONCEPTS

Regression Modeling

- Most practical applications of regression modeling involve using more complicated models than the simple linear regression model.
- Typically, it is better to have more than one variable in a regression model.
- Models with more than one predictor variable are called **multiple regression models**.

Multiple Linear Regression

- Models with more than one predictor variable are called **multiple regression models**.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Multiple Linear Regression

- Models with more than one predictor variable are called **multiple regression models**.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

- With multiple variables in the model, the interpretation of $\hat{\beta}_j$ changes slightly.
- The estimate $\hat{\beta}_j$ is the predicted (or expected or average) change in y with a one unit increase in x_j **given all other variables are held constant**.

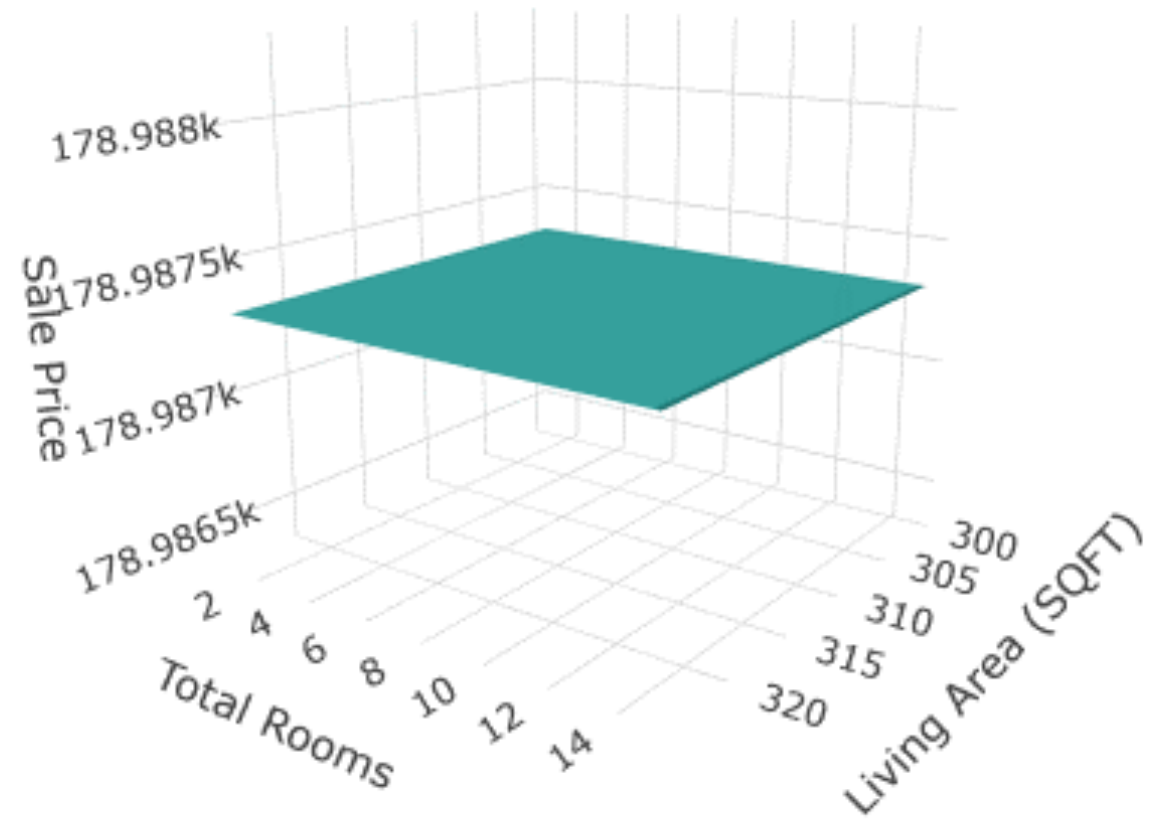
Fitting the Model

- The method for finding the line of best fit for multiple linear regression is the exact same for simple linear regression – the least squares method.
- The only thing that has changed is the predicted value of the response, \hat{y}_i .

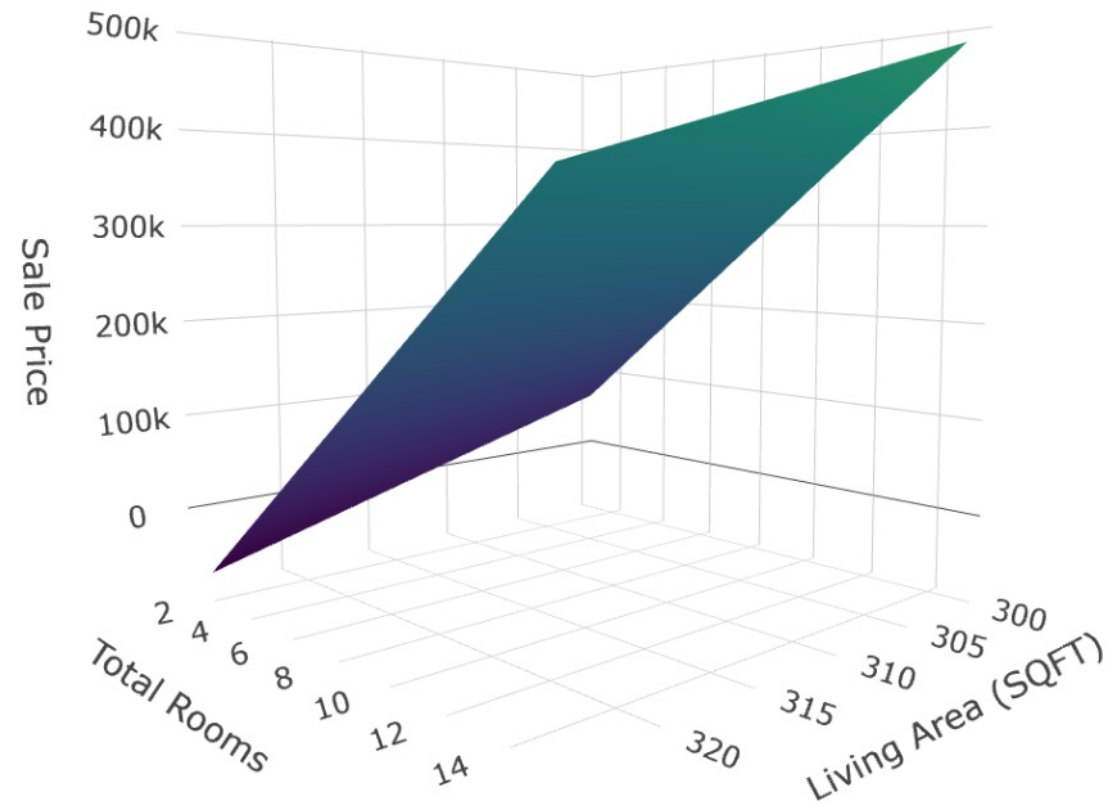
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \cdots + \hat{\beta}_k x_{k,i}$$

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Picturing the Model: No Relationship



Picturing the Model: A Relationship



Multiple *Linear* Regression

- The *linear* in multiple linear regression has nothing to do with the visualization of the fitted plane (or line in 2-dimensions).
- *Linear* refers to the linear combination of variables in the model.
- For example, mathematically, z is a linear combination of x and y (a and b are just constants/numbers):

$$z = ax + by$$

Multiple *Linear* Regression

- The *linear* in multiple linear regression has nothing to do with the visualization of the fitted plane (or line in 2-dimensions).
- *Linear* refers to the linear combination of variables in the model.
- Mathematically, y is a linear combination of the x 's and error term:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Multiple *Linear* Regression

- Linear regression with 4 variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

Multiple *Linear* Regression

- Linear regression with 4 variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

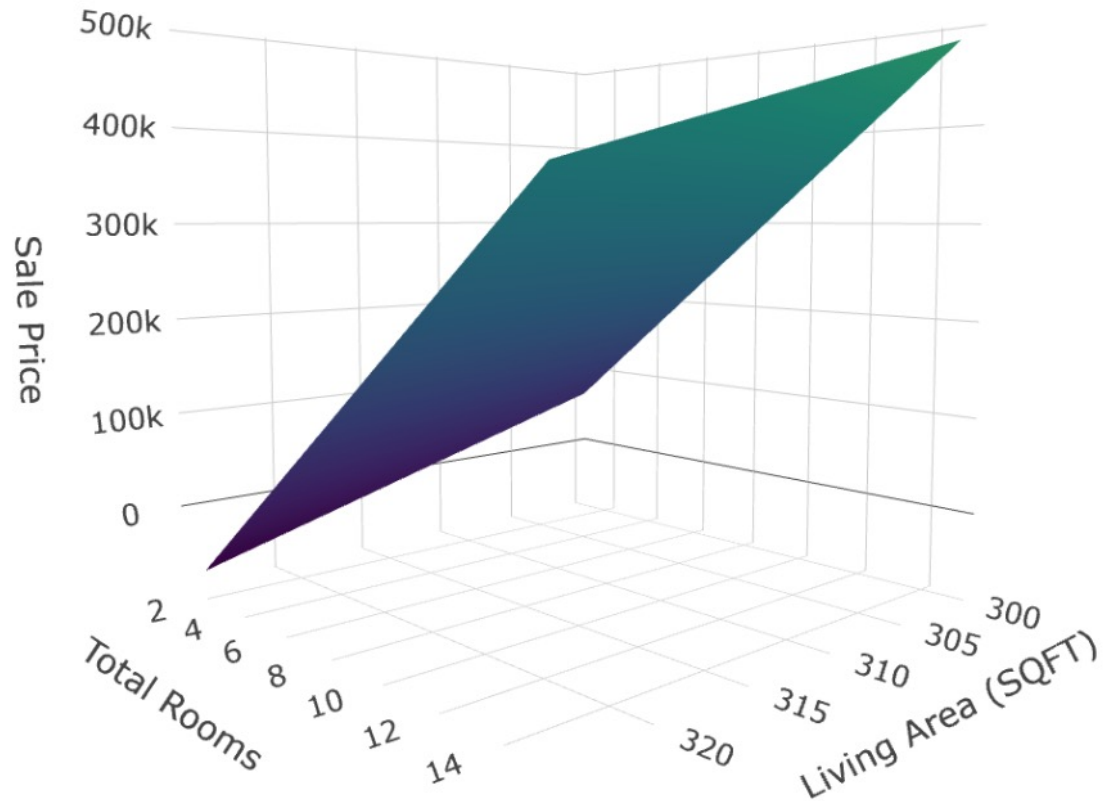
- Let $x_3 = x_1^2$ and $x_4 = x_2^2$:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \varepsilon$$

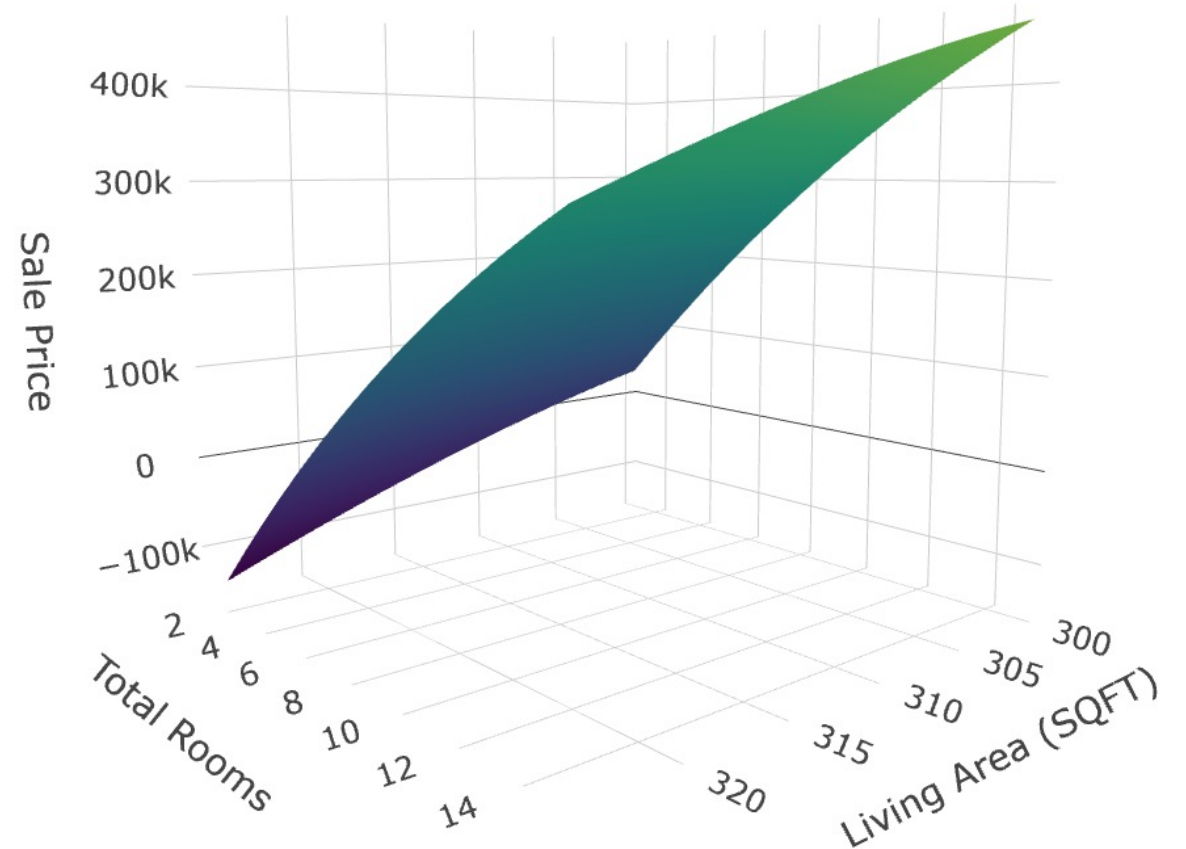
- Still linear regression!

Both Linear Regressions

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$



$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \varepsilon$$





Multiple Linear Regression

GLOBAL & LOCAL INFERENCE

Global vs. Local Test

- In simple linear regression we could just look at the t-test for our slope parameter estimate to determine the utility of our model.
- With multiple parameter estimates comes multiple t-tests.
- Ideally there should be a way of determining whether the model is adequate for predicting y overall, instead of looking at every individual parameter estimate.

Global Hypothesis Test

- **Null Hypothesis:**
 - None of the variables are useful in predicting the target variable.
 - $\beta_1 = \beta_2 = \dots = \beta_k = 0$
- **Alternative Hypothesis:**
 - At least one variable is useful in predicting the target variable.
 - Not all β_i s equal zero.

Global F-test

- The test statistic for this hypothesis test follows an F -distribution and is calculated as follows:

$$F = \frac{\left(\frac{SSR}{k}\right)}{\left(\frac{SSE}{n - (k + 1)}\right)}$$

Global F-test

- The test statistic for this hypothesis test follows an F -distribution and is calculated as follows:

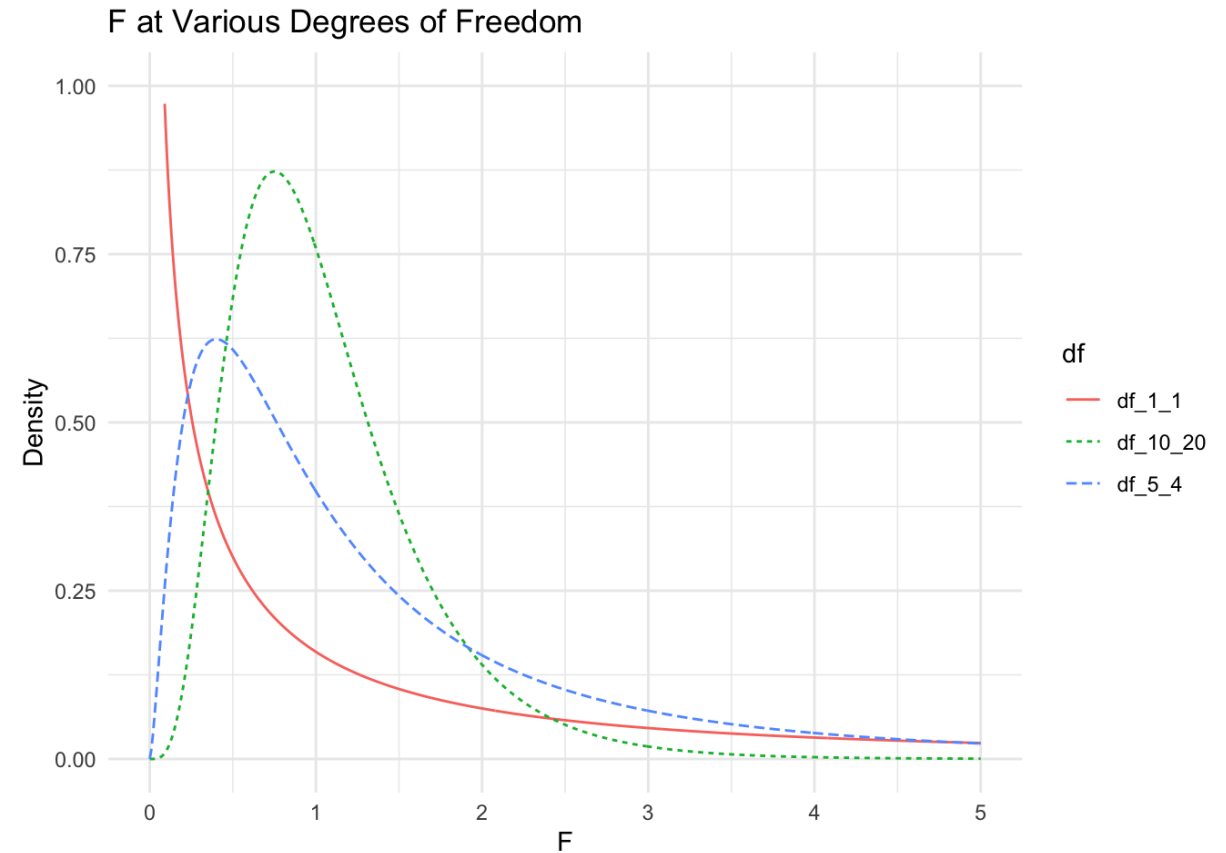
$$F = \frac{\left(\frac{SSR}{k}\right)}{\left(\frac{SSE}{n - (k + 1)}\right)}$$

Average amount of variation each variable explains (i.e. Mean Square Regression)

“Average” amount of variation left per data point (i.e. Mean Square Error)

F-Distribution

- The F-test comes from the **F-distribution**.
- Characteristics of the F-distribution:
 1. Bounded Below By Zero
 2. Right Skewed
 3. Numerator **and** Denominator Degrees of Freedom



Global vs. Local Test

- If the global F-test is significant (at least one variable is useful), then we would dive down into the individual t-tests to find which variables are useful.
- Need to test if the values of each of these coefficients are zero to determine if a relationship exists between the response variable y and that **specific** explanatory variable x .

$$H_0: \beta_j = 0, \quad \text{for } j = 1, \dots, k$$

$$H_a: \beta_j \neq 0, \quad \text{for } j = 1, \dots, k$$

Multiple Linear Regression

```
ames_lm2 <- lm(Sale_Price ~ Gr_Liv_Area + TotRms_AbvGrd, data = train)
summary(ames_lm2)
```

Multiple Linear Regression

```
## Call:
## lm(formula = Sale_Price ~ Gr_Liv_Area + TotRms_AbvGrd, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -528656  -30077   -1230    21427   361465
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   42562.657    5365.721    7.932 3.51e-15 ***
## Gr_Liv_Area     136.982      4.207   32.558 < 2e-16 ***
## TotRms_AbvGrd -10563.324    1370.007   -7.710 1.94e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56630 on 2048 degrees of freedom
## Multiple R-squared:  0.5024, Adjusted R-squared:  0.5019
## F-statistic: 1034 on 2 and 2048 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression

```
## Call:
## lm(formula = Sale_Price ~ Gr_Liv_Area + TotRms_AbvGrd, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -528656  -30077   -1230   21427  361465
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   42562.657    5365.721    7.932 3.51e-15 ***
## Gr_Liv_Area     136.982      4.207   32.558 < 2e-16 ***
## TotRms_AbvGrd -10563.324    1370.007   -7.710 1.94e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56630 on 2048 degrees of freedom
## Multiple R-squared:  0.5024, Adjusted R-squared:  0.5019
## F-statistic: 1034 on 2 and 2048 DF, p-value: < 2.2e-16
```

Multiple Linear Regression

```
## Call:
## lm(formula = Sale_Price ~ Gr_Liv_Area + TotRms_AbvGrd, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -528656  -30077   -1230    21427   361465
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42562.657   5365.721    7.932 3.51e-15 ***
## Gr_Liv_Area    136.982     4.207   32.558 < 2e-16 ***
## TotRms_AbvGrd -10563.324   1370.007   -7.710 1.94e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56630 on 2048 degrees of freedom
## Multiple R-squared:  0.5024, Adjusted R-squared:  0.5019
## F-statistic: 1034 on 2 and 2048 DF,  p-value: < 2.2e-16
```



Multiple Linear Regression

EVALUATING A MODEL

Assumptions for Linear Regression

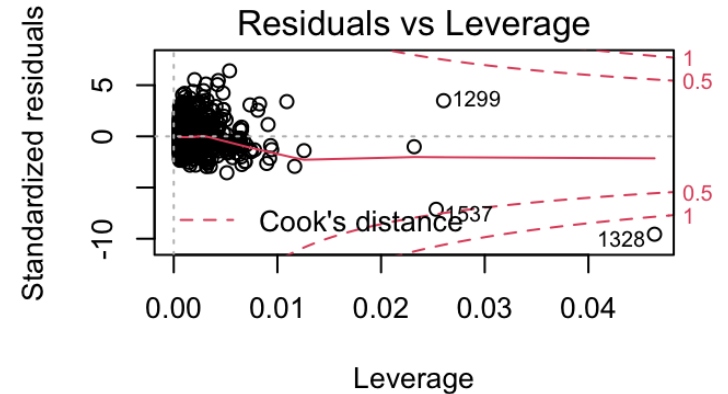
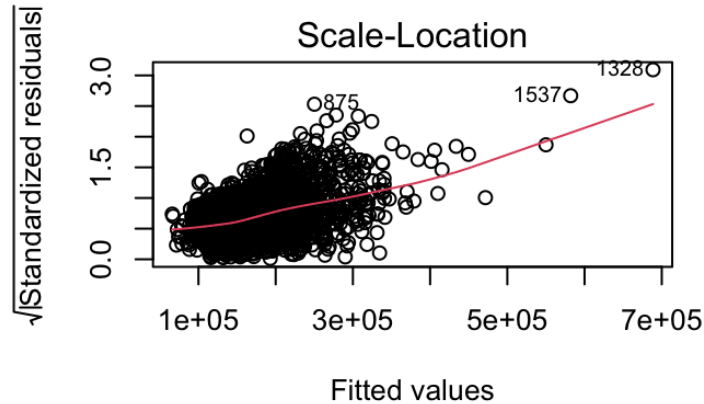
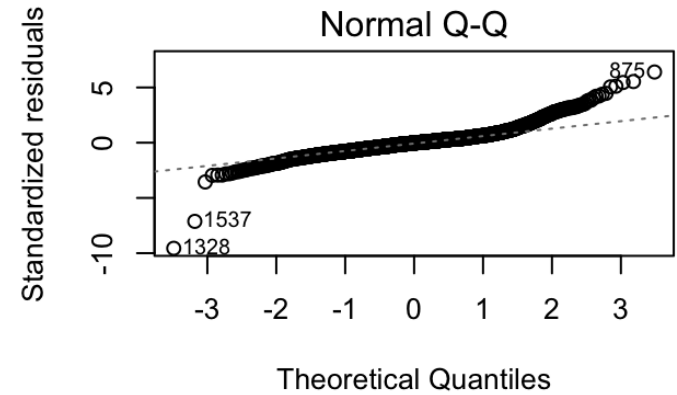
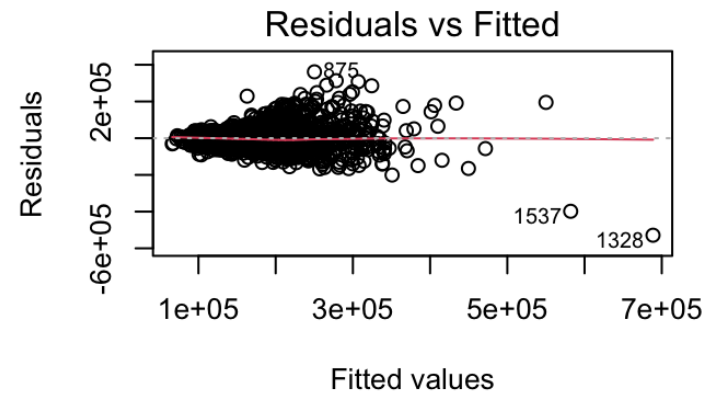
- The mean of the Y s is accurately modeled by a **linear** function of the X s.
- The random error term, ε , is assumed to have a **normal** distribution with a mean of zero.
- The random error term, ε , is assumed to have a **constant variance**, σ^2 .
- The errors are **independent**.
- **No perfect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

Multicollinearity

- **Multicollinearity** – predictor variables are correlated with each other.
- No **perfect** collinearity (multicollinearity) – predictor variables are a perfect linear combination of each other.
 - Only care when collinearity has **drastic** impact.
 - Linear regression only breaks when collinearity is perfect.

Assumptions through Residuals

```
plot(ames_lm2)
```



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 X's. The predicted value of Y is given by this formula:

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

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.

Example with Randomness

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    42562.657   5365.721   7.932 3.51e-15 ***
## Gr_Liv_Area     136.982     4.207  32.558 < 2e-16 ***
## TotRms_AbvGrd  -10563.324  1370.007  -7.710 1.94e-14 ***
## ---
## Multiple R-squared: 0.5024, Adjusted R-squared: 0.502
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    42589.091   5364.877   7.939 3.34e-15 ***
## Gr_Liv_Area     136.927     4.207  32.548 < 2e-16 ***
## TotRms_AbvGrd  -10552.425  1369.808  -7.704 2.05e-14 ***
## rnorm(length(Sale_Price), 0, 1)  1629.854  1259.478   1.294 0.196
## ---
## Multiple R-squared: 0.5028, Adjusted R-squared: 0.502
```

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:

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

OR

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

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.

Example with Randomness

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42562.657   5365.721   7.932 3.51e-15 ***
## Gr_Liv_Area    136.982     4.207  32.558 < 2e-16 ***
## TotRms_AbvGrd -10563.324   1370.007  -7.710 1.94e-14 ***
## ---
## Multiple R-squared: 0.5024, Adjusted R-squared: 0.502
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42589.091   5364.877   7.939 3.34e-15 ***
## Gr_Liv_Area    136.927     4.207  32.548 < 2e-16 ***
## TotRms_AbvGrd -10552.425   1369.808  -7.704 2.05e-14 ***
## rnorm(length(Sale_Price), 0, 1)  1629.854   1259.478   1.294  0.196
## ---
## Multiple R-squared: 0.5028, Adjusted R-squared: 0.502
```



Multiple Linear Regression

CATEGORICAL PREDICTORS

Dummy Variables

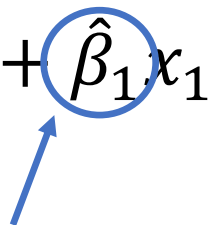
- Categorical variables need to be coded differently because they are not numerical in nature.
- **Dummy variables** are a common way to code categorical variables.
- Central Air Example (Y, N):

$$x_1 = \begin{cases} 1 & \text{if Y} \\ 0 & \text{if N} \end{cases}$$

Dummy Variable Interpretation

- Categorical variables need to be coded differently because they are not numerical in nature.
- **Dummy variables** are a common way to code categorical variables.
- Central Air Example (Y, N):

$$x_1 = \begin{cases} 1 & \text{if Y} \\ 0 & \text{if N} \end{cases}$$

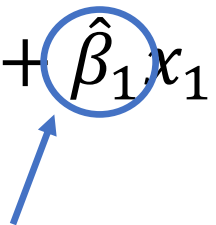
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$


Average difference between
category Y and N.

Dummy Variable Interpretation

- Central Air Example (Y, N):

$$x_1 = \begin{cases} 1 & \text{if Y} \\ 0 & \text{if N} \end{cases}$$

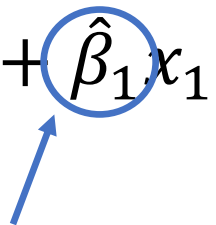
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$


Average difference between
category Y and N. BUT WHY?

Dummy Variable Interpretation

- Central Air Example (Y, N):

$$x_1 = \begin{cases} 1 & \text{if Y} \\ 0 & \text{if N} \end{cases}$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$


Average difference between
category Y and N. **BUT WHY?**

$$\hat{y}_Y = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1 = \hat{\beta}_0 + \hat{\beta}_1$$

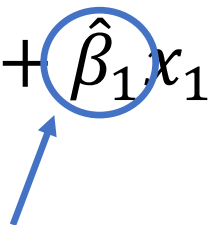
$$\hat{y}_N = \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 = \hat{\beta}_0$$

$$\hat{y}_{Y-N} = (\hat{\beta}_0 + \hat{\beta}_1) - \hat{\beta}_0 = \hat{\beta}_1$$

Effects Coding Interpretation

- Categorical variables need to be coded differently because they are not numerical in nature.
- **Effects coding** is another common way to code categorical variables.
- Central Air Example (Y, N):

$$x_1 = \begin{cases} 1 & \text{if Y} \\ -1 & \text{if N} \end{cases}$$

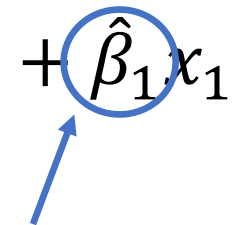
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$


Average difference between
category Y and the overall
average of categories Y & N.

Effects Coding Interpretation

- Central Air Example (Y, N):

$$x_1 = \begin{cases} 1 & \text{if Y} \\ -1 & \text{if N} \end{cases}$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$


Average difference between
category Y and the overall
average of categories Y & N.

BUT WHY?

$$\hat{y}_Y = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1 = \hat{\beta}_0 + \hat{\beta}_1$$

$$\hat{y}_N = \hat{\beta}_0 + \hat{\beta}_1 \cdot (-1) = \hat{\beta}_0 - \hat{\beta}_1$$

$$\hat{y}_{Avg.} = \frac{((\hat{\beta}_0 + \hat{\beta}_1) + (\hat{\beta}_0 - \hat{\beta}_1))}{2} = \hat{\beta}_0$$

$$\hat{y}_{Y-Avg} = (\hat{\beta}_0 + \hat{\beta}_1) - \hat{\beta}_0 = \hat{\beta}_1$$

Multiple Linear Regression

```
ames_lm2 <- lm(Sale_Price ~ Gr_Liv_Area + TotRms_AbvGrd + Central_Air, data = train)
summary(ames_lm2)
```

Multiple Linear Regression

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -510745  -28984   -2317    20273   356742
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -7169.259    6778.879   -1.058    0.29
## Gr_Liv_Area     129.594      4.131   31.374 < 2e-16 ***
## TotRms_AbvGrd  -8980.938    1335.669   -6.724 2.29e-11 ***
## Central_AirY   54513.082    4762.926   11.445 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54910 on 2047 degrees of freedom
## Multiple R-squared:  0.5323, Adjusted R-squared:  0.5316
## F-statistic: 776.6 on 3 and 2047 DF,  p-value: < 2.2e-16
```

