# Lecture 3: Intro 506

**These notes are modified from Mark Greenwood's 2019 STAT 506 course**

## General schedule for class:

Weekly homework (some short answer questions and other more involved problems).

Lectures will generally involve PDF documents with a skeleton where we

we fill in the blanks together.

Lectures will also have an active learning component with think-pair-share exercises as well as some computer coding exercises.

In class group work will be required - real statistical analyses are never done alone - both as informal discussion and occasionally as formal lab exercises

In general, I'd suggest bringing your computer daily. However, I'll try to mention when we will have coding-heavy lectures.

Midterm exam around middle of semester prior to spring break (take home and in class).

maybe week before spring break TBD

Final exam during finals week (take home and in class).

There will be a series of three projects to be completed over the course of the semester. Some projects will be more scripted where you are required to answer specific questions and others will allow you to select the dataset and/or research question(s).

### Regression version of linear models in R

$$Y_i \sim N(\mu, \sigma^2)$$

$$\mu(Y | X_1, X_2, \ldots) = B_0 + B_1 X_1 + B_2 X_2$$

$$Var(Y | X_1, X_2) = \sigma^2$$

Or

$$\underline{Y} = X\underline{B} + \underline{\varepsilon} \qquad \underline{Y}\ n\times 1 ;\ X\ n\times p ,\ \underline{B}\ p\times 1 ,\ \underline{\varepsilon}\ n\times 1$$

$$\hat{\underline{B}} = (X^T X)^{-1} X^T \underline{Y}$$

$$\begin{pmatrix} 1 & x_{11} & x_{21} \\ \vdots & \vdots & \vdots \\ 1 & x_{1i} & x_{2i} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{pmatrix} \qquad \underline{B} = (B_0\ B_1\ \ldots\ B_{p-1})$$

$$V(\hat{\underline{B}}) = \hat{\sigma}^2 (X^T X)^{-1}$$

$$V(\underline{\varepsilon}) = \sigma^2 I$$

**Assumptions:**

$Y_i, Y_j$ are conditionally independent for different $i, j$ (after controlling for X1, X2,. . . ),

this implies $\varepsilon_i$ and $\varepsilon_j$ are independent.

1-22-2020

**Indicator / dummy variables:**

You must define indicator variables and we will use $\boxed{I_{VarName=Cat}}$ otherwise.

if var name = cat , this is a 1

Otherwise a 0

Suppose we have a model with a factor variable predictor:

$$\boxed{Y_{ij} = \mu + \alpha_i + \epsilon_{ij},}$$

where $\alpha_1 = 0$ or $\alpha_k = 0$. This is a one-way ANOVA and can be expressed as a linear model.

$$Y_{ij} = B_0 + B_1 I_{var=cat(2),i} + B_2 I_{var=cat(3),i} + \cdots + \varepsilon_{ij}$$

This can be expressed with `lm(y ~ factor_var, data =)` in R.

An alternative is the cell means specification, where

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

or equivalently

$$Y_{ij} = B_1 I_{var=cat(1),i} + B_2 I_{var=cat(2),i} + \cdots + \varepsilon_{ij}$$

which can be written as `lm(y ~ var_factor - 1, data = )` in R.

## Interactions

With two X variables, consider

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + B_3 X_{1i} X_{2i} + \varepsilon_i$$

If $X_{1i}, X_{2i}$ are indicator variables, then $\beta_3$ is

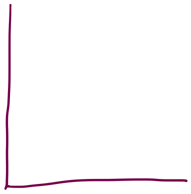
$Q$: write the mean response for - $(X_{1i} = 0, X_{2i} = 0) \rightarrow$ $E[Y | X_{1i} = 0, X_{2i} = 0] = B_0$

- $(X_{1i} = 1, X_{2i} = 0) \rightarrow$ $E[Y | X_{1i} = 1, X_{2i} = 0] = B_0 + B_1$

- $(X_{1i} = 0, X_{2i} = 1) \rightarrow$ $E[Y | X_{1i} = 0, X_{2i} = 1] = B_0 + B_2$

- $(X_{1i} = 1, X_{2i} = 1) \rightarrow$ $E[Y | X_{1i} = 1, X_{2i} = 1] = \underbrace{B_0 + B_1 + B_2}_{\text{model without an interaction}} + \overset{\text{Interaction}}{\boxed{B_3}}$

Interpretation:



First test of interest is the one on the interaction part(s) of the model

Use `effects` plots to generate interpretations of combined results of the two main effects and interaction coefficient(s) (more later)

## Scope of inference:

assume $X_1$ and $X_2$ are variabls related to $y$

Is a variable(s) randomized to subjects or not?

- Randomly assign subjects to levels: $X_1 \rightarrow$ is a casual effect $y$
- Do not randomly assign subjects to levels: $X_2 \rightarrow$ is a non-casual impact on $y$

Are the subjects a random sample or not?

- Random sample → inferences apply to the Population
- Not a random sample → inferences only apply to the Sample

What if you have a quantitative explanatory variable and decide to remove observations that are extreme on the response?

- Don't do this (without justification of a data entry error).
- Consider an analysis with and without the point.

Time period of measurements matters:

- Years of the study: note that inferences to those year ONLY in scope of inference (unless analyzing a random sample of years and then it would be to the population of years sampled from)
- Example: Data collected yearly from 1950 to 1980, only make inferences to 1950 to 1980 (and must note this in scope of inference discussion)
- If data are just from a year or two, note those in study description if known: no implication that that year or two is representative of what would have happened in many others

Demographics of subjects similarly useful to describe.

01-29-20

## Reporting of results:

Report sample size as obtained and analyzed in reports, noting reasons for any missing/deleted observations

"Evidence" sentence contains something like: We found XXX evidence against the null hypothesis of XXX (test statistic, dist under $H_0$, and p-value) controlled for …, so we conclude that [something about alternative hypothesis].

In this, report the value of the test statistic, its distribution under the null hypothesis (t, F, Chi-squared, permutation, standard normal for Z) AND that distribution's degrees of freedom (if exist).

"Evidence" sentence (if not one-sided test) does not test for specific direction - direction is part of "size" sentence or general summary of results.

DO **NOT** just report degrees of freedom without a named distribution to associate it with.

Do report them efficiently as F(2,20) as opposed to F with 2 numerator and 20 denominator degrees of freedom.

You should always report results that correspond to the test being done (one-sided or two-sided) and **"controlling for" anything else in the model (be specific)**.

"Size" sentences interpret slope coefficient with CI where possible for any discussions of specific coefficients.

In complex models (say with interactions or with many levels of categorical variables), "size" is related to pattern of results and so may not involve very specific numerical results and CIs.

Follow up analyses with contrasts or pairwise comparisons if of interest in the application.

**NEVER use the word "significant"** in this class. Discuss strength of evidence that tests provide against the null and then conclude what that means. The "s" word is loaded and often misunderstood. Find a way to make your point without that word!

**"Data"** is a plural word unless you only have one subject and then you should be saying *datum*. This is hard. You are bombarded with the colloquial use of the term (and in Excel) that means "information" which is OK to use in a singular sense. Whenever writing "data" as a *now* sophisticated, statistically-trained scientist, in your mind replace it with "things" and see how your sentence sounds.

Run spell check on everything before submitting it, whether writing reports or short answers.

R stuff (more to come too!): - You will use R either exclusively or nearly exclusively for the course

- You will practice good data management, wrangling, visualization, analysis, and model interrogation and exploration using reproducible methods in R via R-studio.
- Do not `attach()` data sets. This is extremely dangerous when working with complicated data sets.
- Use `<-` instead of "=" to assign things in R ("=" is for options within function calls)

## R Markdown and Reproducible Research

R-markdown provides an easy to use venue where you can write code as *chunks* and run the chunk entirely and provide annotations/writing around the code and output to document and explain what you are doing

Cheat sheet for R-markdown: https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf

Make sure to have a miktex (https://miktex.org/) installation on your computer, miktex allows you to compile to pdf documents and more latex functionality in the documents

**Results Reproducibility/ replicability**: - "ability to recreate all of the figures and numbers in a scientific publication from the code and data provided by the authors" (Leek and Jager, 2017).

- Code and data together and can be easily re-run and provide same answers
- Output usually contains code run and output but some may be suppressed in output (echo=F codechunk option) to clean up output - but only do this if instructed to.
- Allows other researchers to verify results and for you to repeat analyses and get same results
- Anything that is "random" should have `set.seed(XXX)` specified in code so re-running the code will always produce same results

**Inferential replicability/ reproducibility**:

- "ability to reperform the experiments and computational analyses in a scientific publication and arrive at consistent results" (Leek and Jager, 2017).
- Repeat same study and get similar (same?) results
- Remember that p-values are random if the data are random (see Amrhein et al., p 5)
- Null Hypothesis Significance Testing with fixed $\alpha$ levels struggles in replication studies with significant or not interpretation (p-value less than 0.05 once, repeat - what are the chances of a p-value less than 0.05?)
- Consider: what is the chance in two independent studies of the same NHST to both have a p-value less than 0.05 if null hypothesis is true?

Pr(Both $H_0$ rejected| $H_0$ is true) = $.05^2 = .0025$

Pr(At lease one false rejection with two tests| $H_0$ is true) = $1 - Pr(\text{none}) = 1 - .95^2 = .098$

With 10 tests, this rises to about $0.4$

## Data Visualization

Consider a dataset that contains housing prices in King County, WA.

```
Seattle <- read_csv('http://math.montana.edu/ahoegh/teaching/stat408/datasets/SeattleHousing.csv')
```

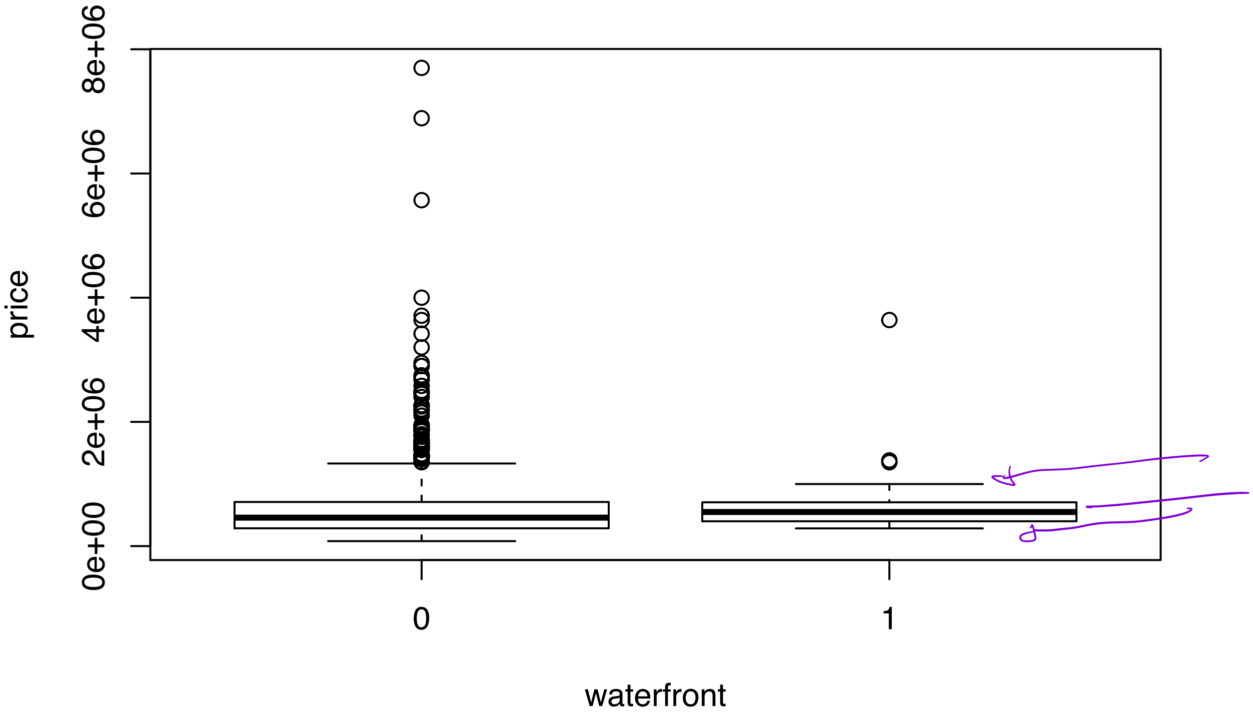
```
## Parsed with column specification:
## cols(
##   price = col_double(),
##   bedrooms = col_double(),
##   bathrooms = col_double(),
##   sqft_living = col_double(),
##   sqft_lot = col_double(),
##   floors = col_double(),
##   waterfront = col_double(),
##   sqft_above = col_double(),
##   sqft_basement = col_double(),
##   zipcode = col_double(),
##   lat = col_double(),
##   long = col_double(),
##   yr_sold = col_double(),
##   mn_sold = col_double()
## )
```

```
Seattle <- Seattle %>% mutate(waterfront = as.factor(waterfront),
            bedrooms = as.factor(bedrooms),
            multiple_floors = as.factor(floors > 1))
```

## Alternatives to boxplots…

Boxplots display the 5 number summary (min, Q1, median, Q3, and max) and potentially some outlier information.

```
boxplot(price ~ waterfront, data = Seattle)
```



Single observations may be flagged by outlier "rules" in boxplots that are not really outside the overall pattern (so not really outliers) and can't see points that are just inside cut-offs

Boxplots will not detect multi-modal distributions, can have issues when multiple quartiles are tied, look the same regardless of sample size, and can be made with less than 5 observations in R (even though 5 lines are displayed)

Reasons for making single and side-by-side boxplots? To compare:

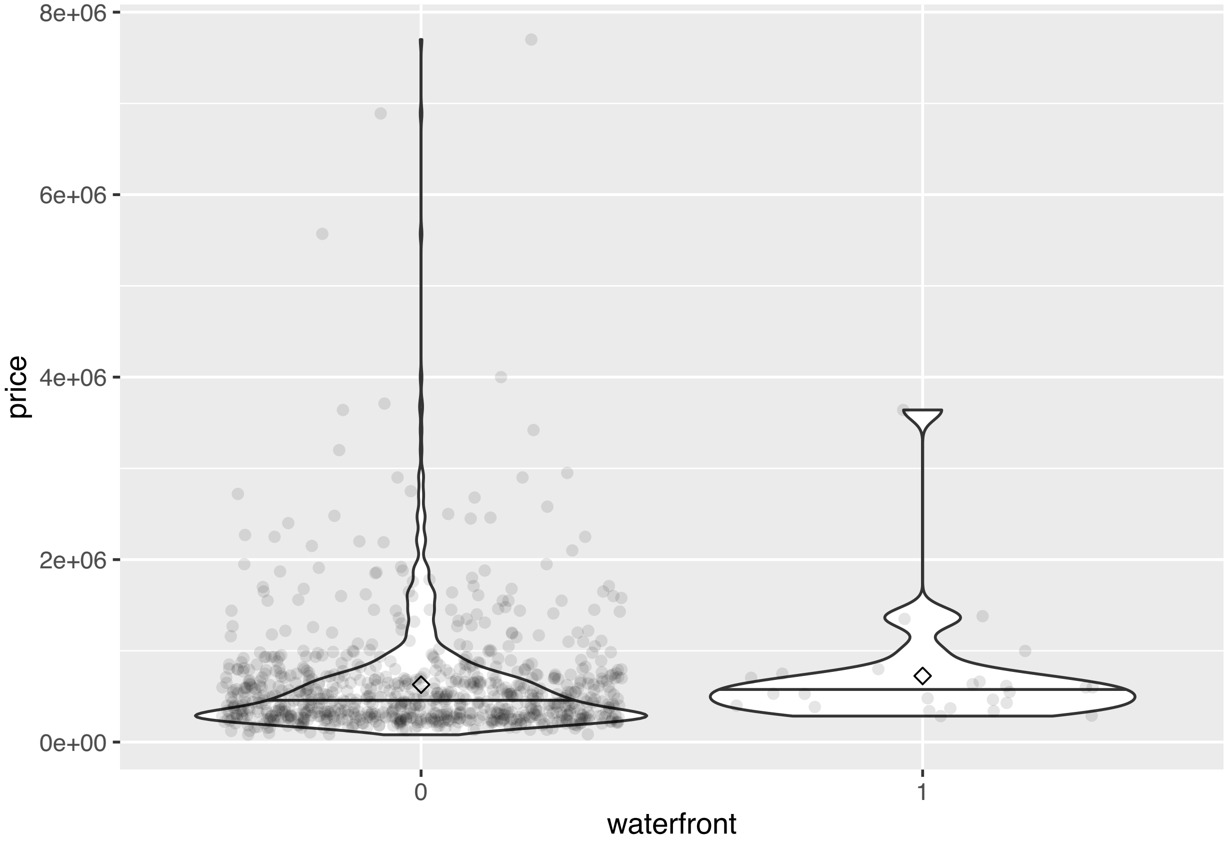
- center
- shape
- spread

In this course, use `geom_violin()` or at least overlay points on top of boxplots.

```r
library(dplyr)
Seattle %>% ggplot(aes(y = price, x = waterfront)) +
  geom_violin(draw_quantiles = c( 0.5)) +
  geom_jitter(alpha = .1) +
  stat_summary(fun.y=mean, geom="point", shape=23, size=2)
```

```
## Warning in regularize.values(x, y, ties, missing(ties)): collapsing to unique
## 'x' values

## Warning in regularize.values(x, y, ties, missing(ties)): collapsing to unique
## 'x' values
```



Nonparametric density plots help to visualize shape (both default to using normal kernels so may look at bit more normal than really are)

### Two-sample t-tests in R:

The goal of the t-test is to Compare the mean response of two populations.

This can be written as a linear model

$$Y_i = B_0 + B_1 I_{var=cat(2)_i} + \varepsilon_i$$

```
t.test(price~waterfront,data=Seattle)
```

```
##
##  Welch Two Sample t-test
##
## data:  price by waterfront
## t = -0.70093, df = 25.286, p-value = 0.4897
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -375116.5  184537.4
## sample estimates:
## mean in group 0 mean in group 1
##        629850.1        725139.7
```

01-31-2020

## One-Way ANOVA in R:

Situation: k - factor predictor with a quantitive response

$$Y_i = \beta_0 + \beta_1 I_{cat(2)i} + \cdots \beta_{k-1} I_{cat(k)i} + \varepsilon_i$$

Model:

model.matrix(      )

Cell-means vs reference and deviation coding:

The cell means notation `model.matrix(y ~ x -1)` results in estimates for the group means $\mu_i$.

```r
lm_beds <- lm(price~ bedrooms - 1,data=Seattle )
summary(lm_beds)
```

```
##
## Call:
## lm(formula = price ~ bedrooms - 1, data = Seattle)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2219025  -265459  -105509   144541  5215975
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## bedrooms0   502333     329487   1.525   0.1277
## bedrooms1   367427     158281   2.321   0.0205
## bedrooms2   438209      48580   9.020  < 2e-16
## bedrooms3   505459      27176  18.600  < 2e-16
## bedrooms4   825598      38921  21.212  < 2e-16
## bedrooms5  1182603      83244  14.207  < 2e-16
## bedrooms6  2484025     201769  12.311  < 2e-16
## bedrooms7  2133333     329487   6.475  1.6e-10
## bedrooms9   700000     570689   1.227   0.2203
##
## Residual standard error: 570700 on 860 degrees of freedom
## Multiple R-squared:  0.5985, Adjusted R-squared:  0.5943
## F-statistic: 142.4 on 9 and 860 DF,  p-value: < 2.2e-16
```

```r
anova(lm_beds)
```

```
## Analysis of Variance Table
##
## Response: price
##            Df     Sum Sq    Mean Sq F value    Pr(>F)
## bedrooms    9 4.1747e+14 4.6385e+13  142.42 < 2.2e-16
## Residuals 860 2.8009e+14 3.2569e+11
```

```r
#
```

```r
lm_beds2 <- lm(price ~ bedrooms, data = Seattle)
summary(lm_beds2)
```

```
##
## Call:
## lm(formula = price ~ bedrooms, data = Seattle)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2219025  -265459  -105509   144541  5215975
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   502333     329487   1.525 0.127729
## bedrooms1    -134906     365533  -0.369 0.712168
## bedrooms2     -64124     333050  -0.193 0.847368
## bedrooms3       3126     330606   0.009 0.992459
## bedrooms4     323265     331778   0.974 0.330161
## bedrooms5     680270     339840   2.002 0.045627
## bedrooms6    1981692     386358   5.129  3.6e-07
## bedrooms7    1631000     465966   3.500 0.000489
## bedrooms9     197667     658975   0.300 0.764279
##
## Residual standard error: 570700 on 860 degrees of freedom
## Multiple R-squared:  0.1993, Adjusted R-squared:  0.1919
## F-statistic: 26.76 on 8 and 860 DF,  p-value: < 2.2e-16
```

```r
anova(lm_beds2)
```

```
## Analysis of Variance Table
##
## Response: price
##            Df     Sum Sq    Mean Sq F value    Pr(>F)
## bedrooms    8 6.9718e+13 8.7147e+12  26.758 < 2.2e-16
## Residuals 860 2.8009e+14 3.2569e+11
```

## Contrasts (Tukey-Kramer) in R:

In situations with a $k$ level ($k$>2) categorical variable and we find evidence against the null that the levels of that variable have the same mean, researchers often want to know which levels are different

F-test just tells us that there is some difference somewhere in the means across the levels

We get tests of each level versus baseline from the t-tests in the model summary, but no direct comparisons of the non-baseline levels with each other.

And the default t-tests fail to account for all the tests we might be conducting, especially when there are many levels being compared:

The number of pairs to compare can be calculated based on taking $k$ choose 2 (in R: `choose(k,2)`), and this grows quickly as $k$ increases - `choose(3,2)`=3, `choose(5,2)`=10, `choose(10,2)`=45

Tukey-Kramer is one approach to performing all pairwise comparisons and controlling the overall (family-wise) error rate across all these tests

Family-wise error control relates to controlling the probability of at least one false detection

Output also contains estimated differences in means between the groups - that is also useful information in many situations

Interpretations can use "detected to be different" of "not detected to be different" wording for pairs and discussing groups of levels that were not detectably different but were detected to be different from other groups

Can be run in more complex models (mixed, glms) and in situations with other variables in the model (differences in levels controlled for . . . )

**Two-Way ANOVA:**

$\boxed{\text{Additive Model}}$

$$Y_{ijk} = B_0 + \alpha_i + B_j + \varepsilon_{ijk}$$

$\boxed{\alpha B_{ij} \quad \text{interaction term}}$

$i = 1, \ldots, I$ & $j = 1, \ldots, J$

groups across the categories

$\circledast$ , +

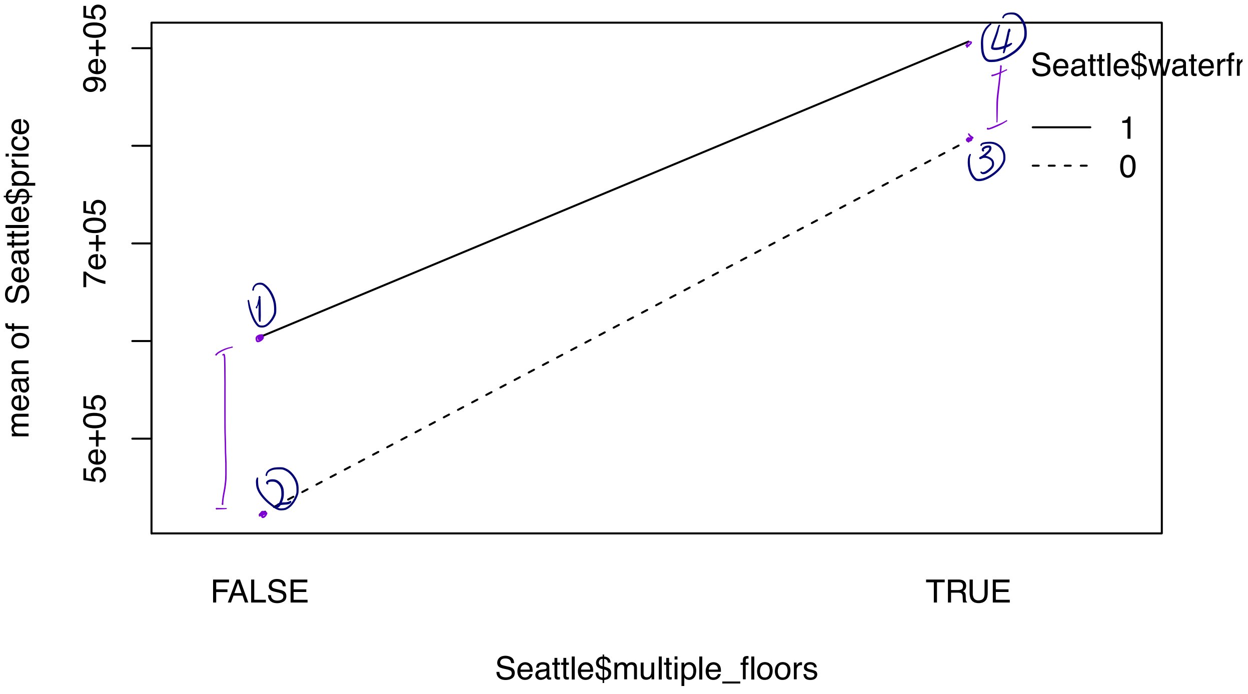
```r
lm_anova2<- lm(price ~ multiple_floors + waterfront, data = Seattle)
summary(lm_anova2)
```

```
##
## Call:
## lm(formula = price ~ multiple_floors + waterfront, data = Seattle)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -681249 -261552 -138552   55451 6895451
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           423552      30585  13.848   <2e-16
## multiple_floorsTRUE   380996      41294   9.226   <2e-16
## waterfront1           149189     123173   1.211    0.226
##
## Residual standard error: 606300 on 866 degrees of freedom
## Multiple R-squared:  0.09008,    Adjusted R-squared:  0.08797
## F-statistic: 42.86 on 2 and 866 DF,  p-value: < 2.2e-16
```

single floor / non-waterfront house

Interaction plots provide visual information to assess whether the effects (components) “look” like they interact in their impacts on the response.

```
interaction.plot(Seattle$multiple_floors,Seattle$waterfront, Seattle$price)
```



Plot of the means of responses by all combinations of different variables (one variable on x-axis and one as lines/symbols)

Look for non-parallel lines in these plots - that suggests that differences at different levels of the x-axis variable are not all the same on the response.

Lines do not need to cross to create an interaction that matters

Importance of the interaction relates to precision in the responses at each combination so add some measure of precision to plot to get some sense of whether the interaction looks “big”/ clearly non-parallel