Example4_3

Kevin Cummiskey

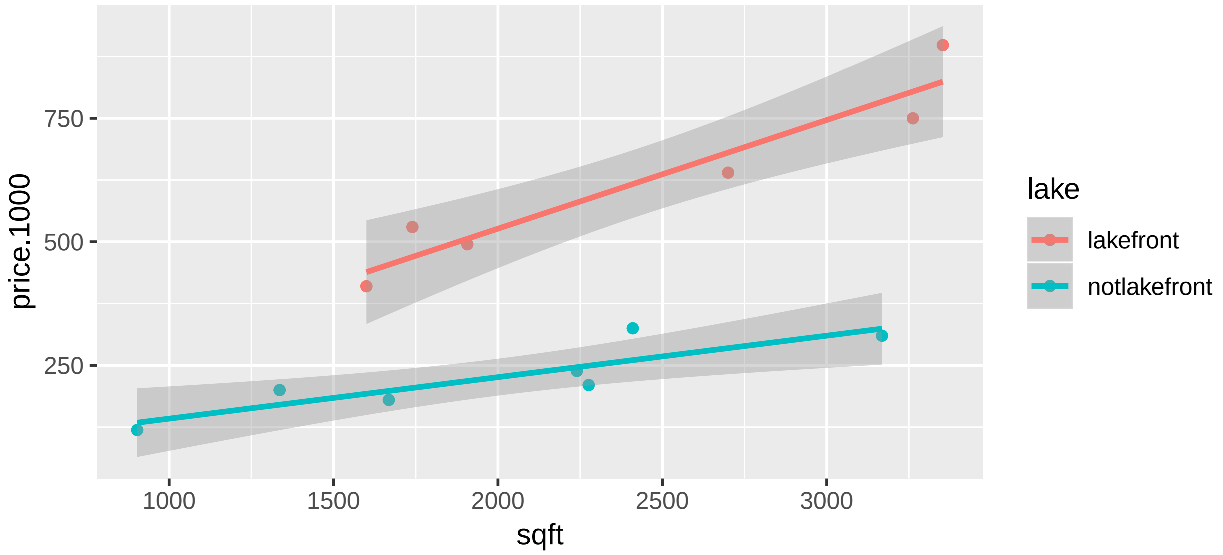
October 18, 2019

Read in the data and perform data analysis.

```r
houses = read.table(file = "http://www.isi-stats.com/isi2/data/housing.txt",
                    header = TRUE)
head(houses)
```

```
##   sqft price.1000      lake
## 1 2700      639.9 lakefront
## 2 3353      898.0 lakefront
## 3 1600      410.0 lakefront
## 4 1740      529.9 lakefront
## 5 1907      495.0 lakefront
## 6 3262      749.9 lakefront
```

```r
houses %>% ggplot(aes(x = sqft, y = price.1000, color = lake)) + geom_point() +
  geom_smooth(method = "lm")
```



What are the observational units? explanatory variable(s)? outcome variable?

What is the study design?

Let's fit a simple linear regression model.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$$

where y_i is the price and x_i is the size (sq ft) of house i.

```
model_simple = lm(price.1000 ~ sqft, data = houses)
summary(model_simple)
```

```
##
## Call:
## lm(formula = price.1000 ~ sqft, data = houses)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -304.70 -128.44  -13.74  128.98  244.04
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -59.36870  161.36807  -0.368   0.7199
## sqft          0.21274    0.06963   3.055   0.0109 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 185.1 on 11 degrees of freedom
## Multiple R-squared:  0.4591, Adjusted R-squared:  0.4099
## F-statistic: 9.335 on 1 and 11 DF,  p-value: 0.01094
```

What is the interpretation of β_1?

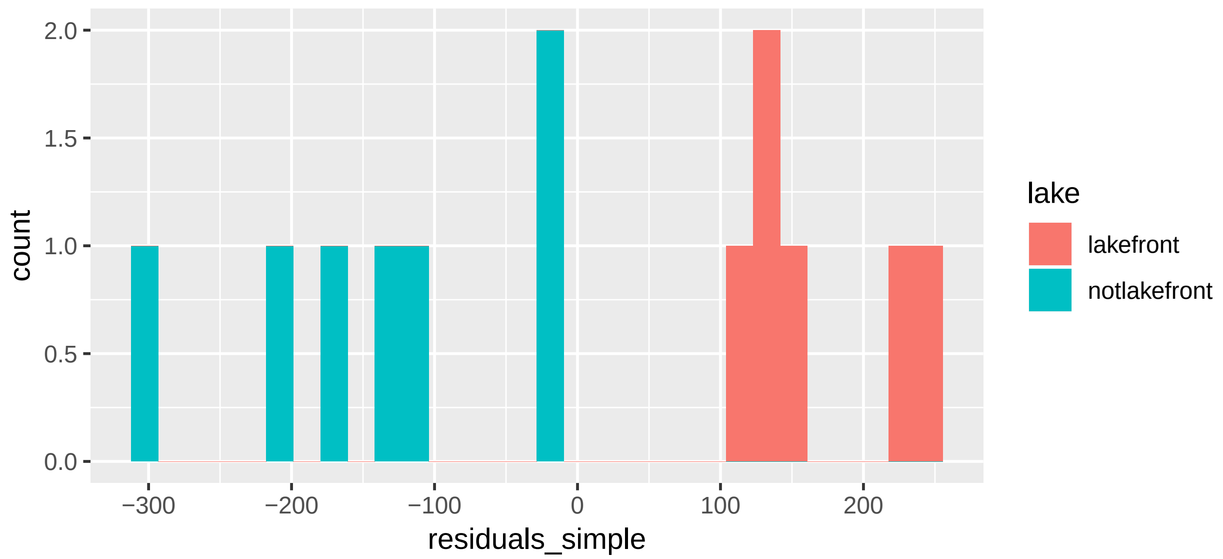
Let's look at the residuals?

```
houses = houses %>% mutate(residuals_simple = residuals(model_simple))
```

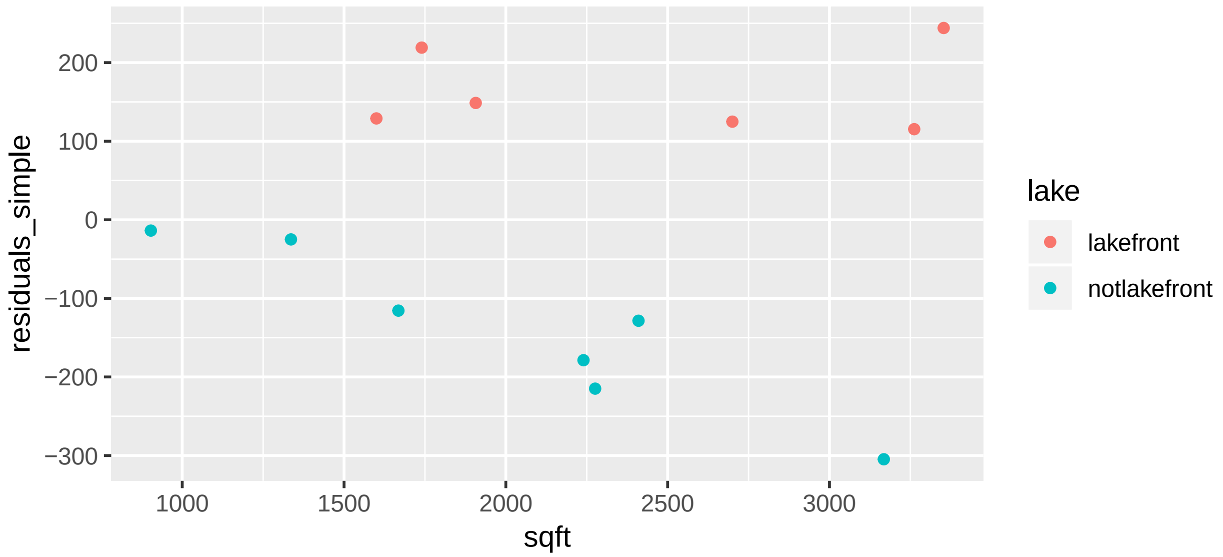
```
## Warning: The `printer` argument is deprecated as of rlang 0.3.0.
## This warning is displayed once per session.
```

```
houses %>% ggplot(aes(x = residuals_simple, fill = lake)) + geom_histogram()
```

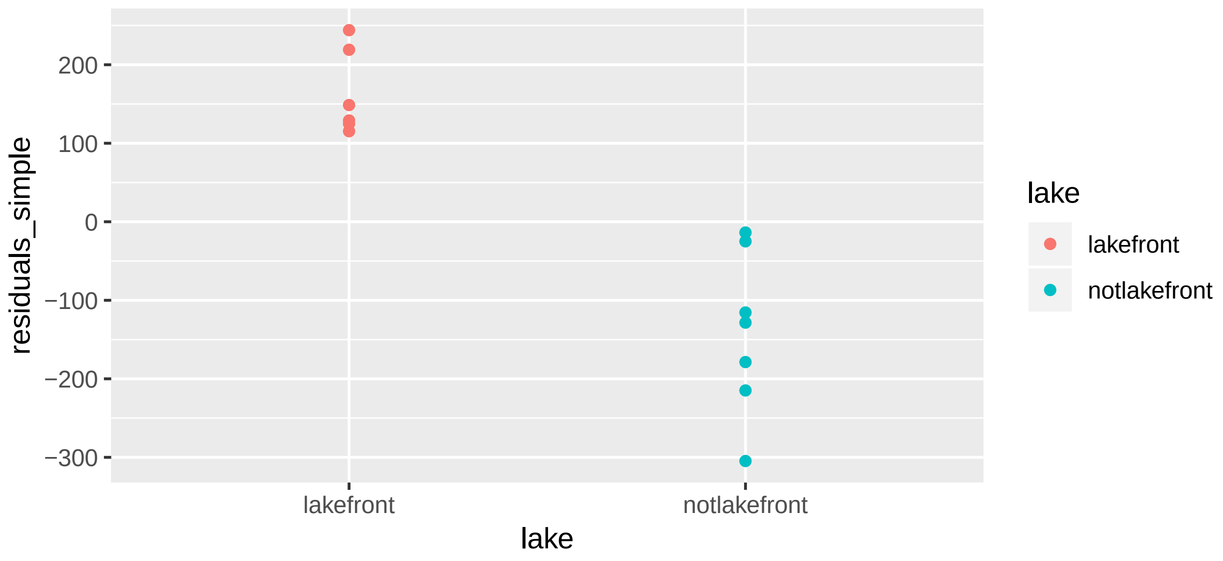
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
houses %>% ggplot(aes(x = sqft, y = residuals_simple, color = lake)) + geom_point()
```



```
houses %>% ggplot(aes(x = lake, y = residuals_simple, color = lake)) + geom_point()
```



What do these residuals tell us?

Next, let's consider a model including location.

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$$

where y_i is the price of house i, $x_{1,i}$ is the size (sq ft) of house i, and $x_{2,i}$ is 1 if house i is lakefront and is 0 otherwise.

$x_{2,i}$ is an indicator variable. How else could we encode this variable?

What assumptions does this model make?

Let's fit the model.

```r
#Reverse coding of lake
houses$lake = factor(houses$lake, levels = c("notlakefront",
                                             "lakefront"))

model_withLake = lm(price.1000 ~ sqft + lake, data = houses)
summary(model_withLake)
```

```
##
## Call:
## lm(formula = price.1000 ~ sqft + lake, data = houses)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -89.059 -48.444   3.072  38.191 140.421
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -70.1821    62.8062  -1.117 0.289933
## sqft            0.1481     0.0283   5.233 0.000383 ***
## lakelakefront 331.2235    41.8470   7.915 1.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 72.02 on 10 degrees of freedom
## Multiple R-squared:  0.9255, Adjusted R-squared:  0.9106
## F-statistic: 62.15 on 2 and 10 DF,  p-value: 2.289e-06
```

```r
anova(model_withLake)
```

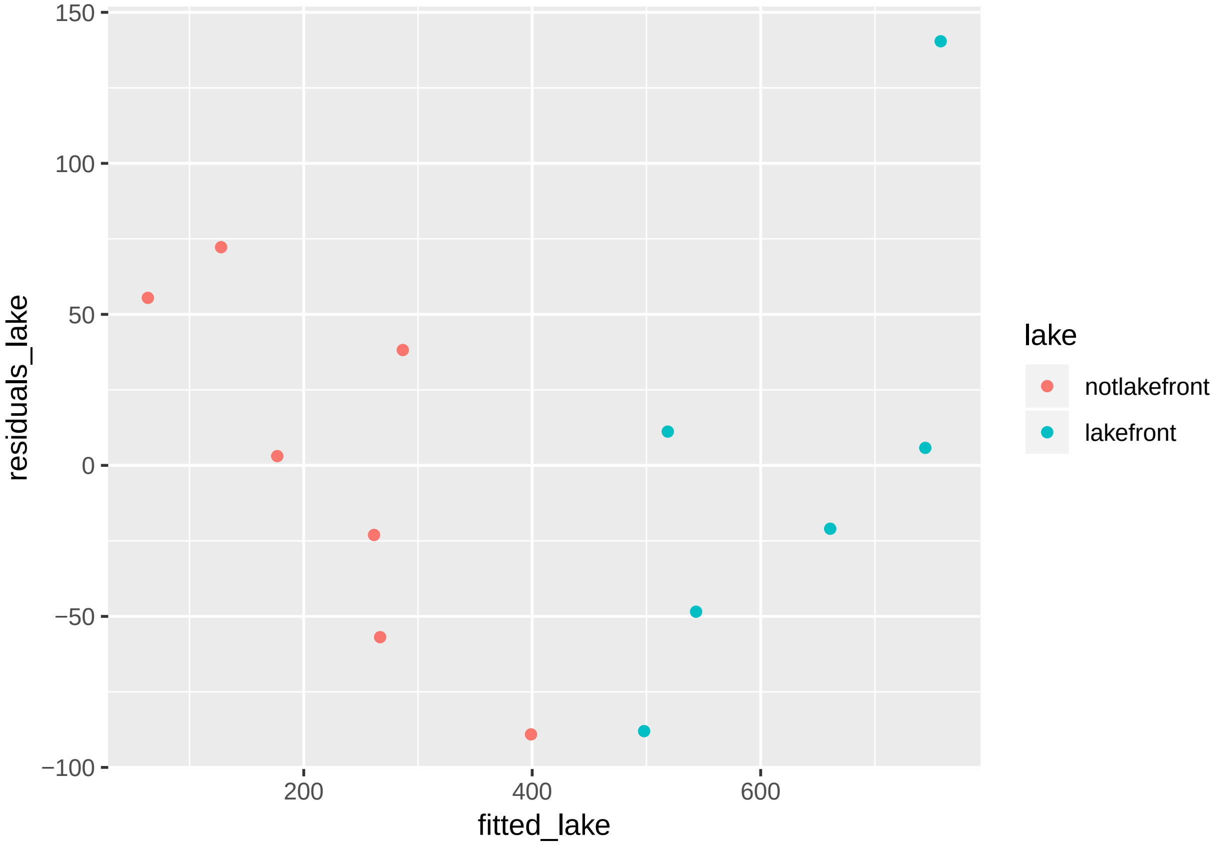
```
## Analysis of Variance Table
##
## Response: price.1000
##           Df Sum Sq Mean Sq F value    Pr(>F)
## sqft       1 319753  319753  61.654 1.386e-05 ***
## lake       1 324911  324911  62.649 1.293e-05 ***
## Residuals 10  51862    5186
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What is the interpretation of β_1?

What conclusions would you make from this model?

Let's look at the residuals.

```
houses = houses %>% mutate(residuals_lake = residuals(model_withLake))
houses = houses %>% mutate(fitted_lake = fitted(model_withLake))
houses %>% ggplot(aes(x = fitted_lake, y = residuals_lake, color = lake)) +
  geom_point()
```



What do the residuals tell us?