Lsn 23 - MA206Y

Clark

Admin

Previously we have been fitting linear regression models:

Recall here we are interested in determining if there is a linear relationship between x_i and y_i.

For instance, we might be interested in explaining the relationship between square footage and price of a house.

The model here is:

Our hypothesis that we are testing is:

We fit this doing:

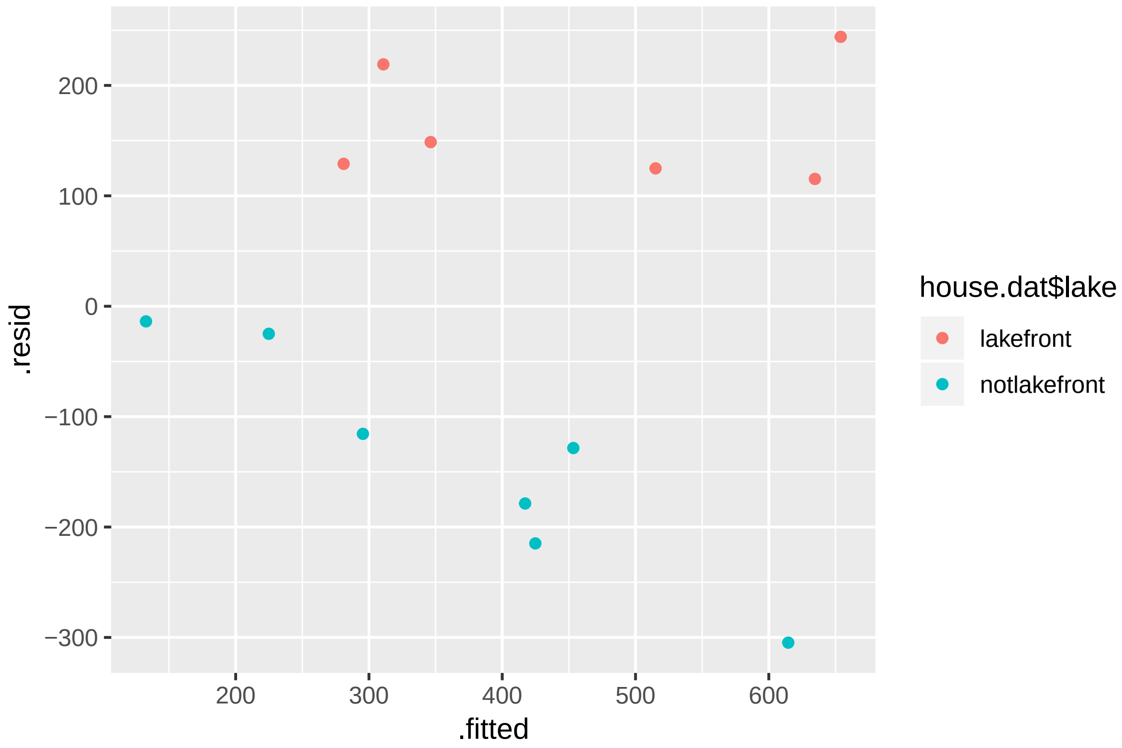
```r
house.dat=read.table("http://www.isi-stats.com/isi2/data/housing.txt",header=T)

sq.lm=lm(price.1000~sqft,data=house.dat)
summary(sq.lm)
```

```
##
## Call:
## lm(formula = price.1000 ~ sqft, data = house.dat)
##
## 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
```

So what does this mean?

Recall that we had assumptions to check.

```
sq.lm %>% ggplot(aes(x=.fitted,y=.resid,color=house.dat$lake))+geom_point()
```



What do we see here?

Here it might be more appropriate to build a multiple regression model. THe model is:

Now, the most important part of understanding this model is that the effects are **conditional**. Meaning, what we are testing is, **If we have a model for square footage and price does lake front have a relationship with price**. Also, **If we have a model for lake front and price, does square footage have a relationship with price**. Our residual plot suggests that this is true.

Our H_0 and H_a are now:

We can fit this with:

```
full.lm=lm(price.1000~sqft+lake,data=house.dat)
summary(full.lm)
```

```
##
```

```
## Call:
## lm(formula = price.1000 ~ sqft + lake, data = house.dat)
##
## 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)         261.0413    74.7091   3.494 0.005783 **
## sqft                  0.1481     0.0283   5.233 0.000383 ***
## lakenotlakefront   -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
```

Our fitted model is then:

Our conclusion is:

This whole business of conditional might not seem important, but sometimes it is. One thing we note is that the impact of square footage on price depends on whether lakefront is in our model or not. If we are just looking at square footage we would say that an increase in 1 sqft adds .21274 price. However, when we add lakefront to the model, we see that an increase in 1 sq ft adds .1481 price to the model. Why is this? Well, it's a bit tricky, but if we look at squarefootage and lakefront we see:

```
house.dat %>% group_by(lake)%>%
  summarize(mu.sqft=mean(sqft),mu.price=mean(price.1000))
```

```
## # A tibble: 2 x 3
##   lake         mu.sqft mu.price
##   <fct>          <dbl>    <dbl>
## 1 lakefront      2427      620.
## 2 notlakefront   2000.     226.
```

So houses on the lakefront are *also bigger houses*. So, here we have potential confounding. Recall that confounding is:

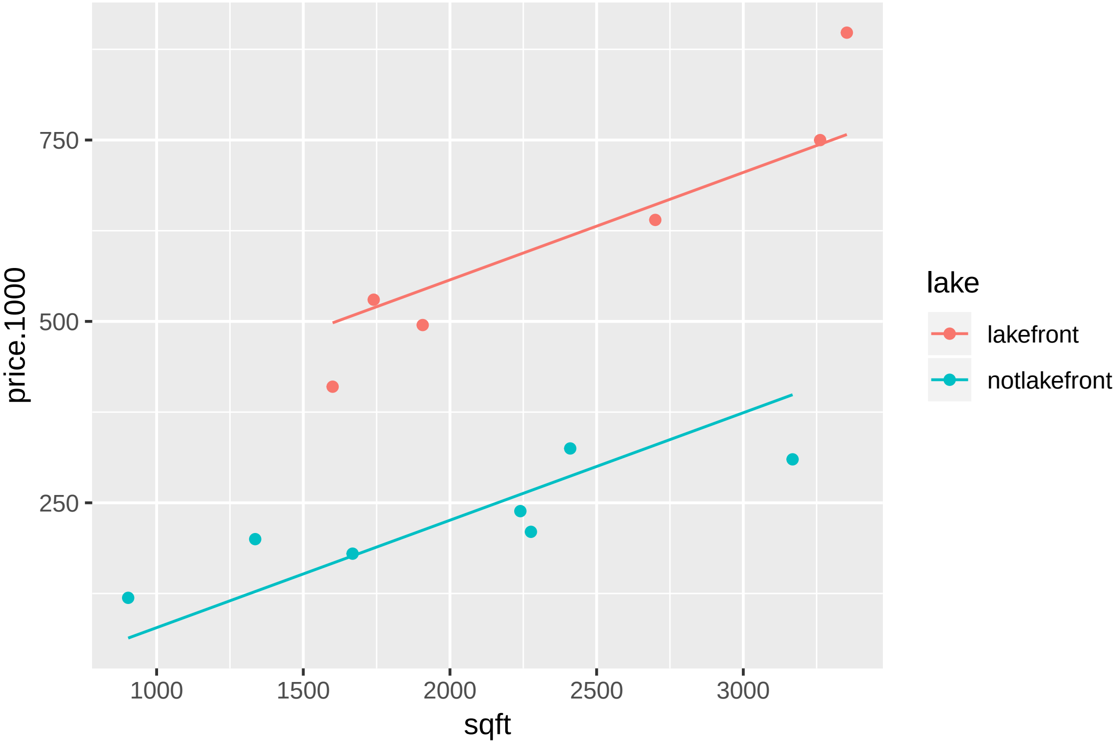
Our diagram is:

So when we don't have lakefront in our model, we attribute an effect due to square footage that is *actually due to lakefront!*

What we want then is the conditional effect.

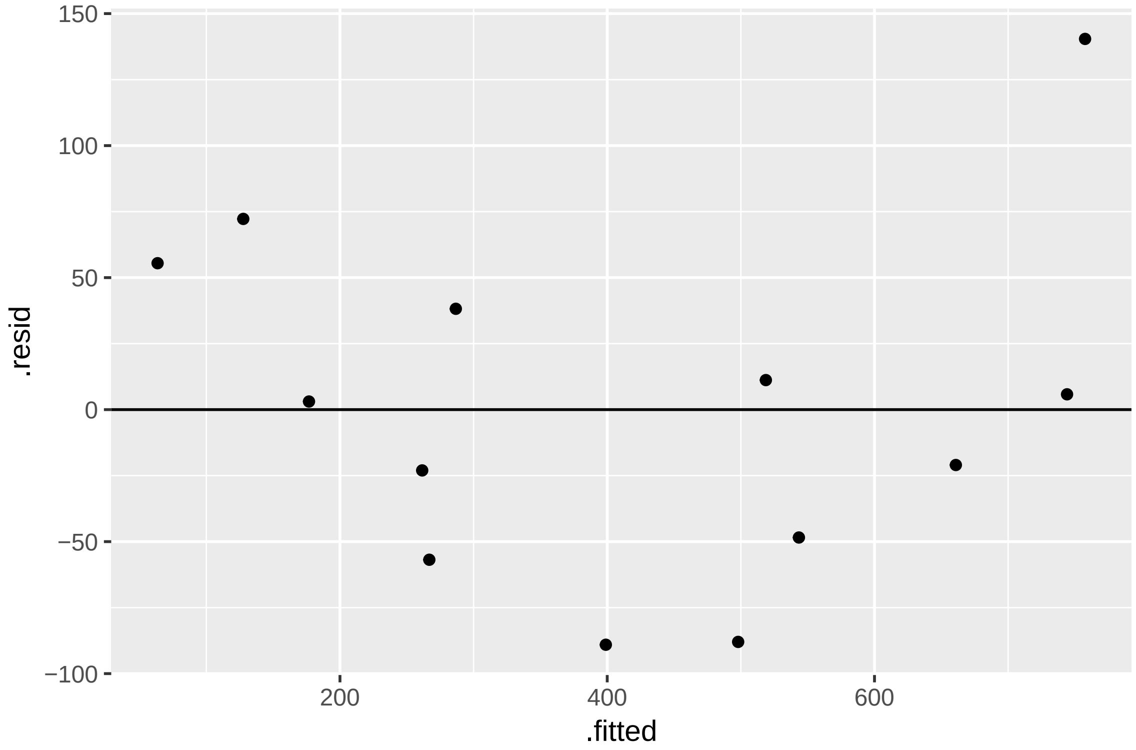
To visualize the fitted model we can do:

```r
library(broom) #Need to install this library
full.lm2=augment(full.lm)
full.lm2 %>% ggplot(aes(x=sqft,y=price.1000,color=lake))+
  geom_point()+
  geom_line(aes(y=.fitted,color=lake))
```



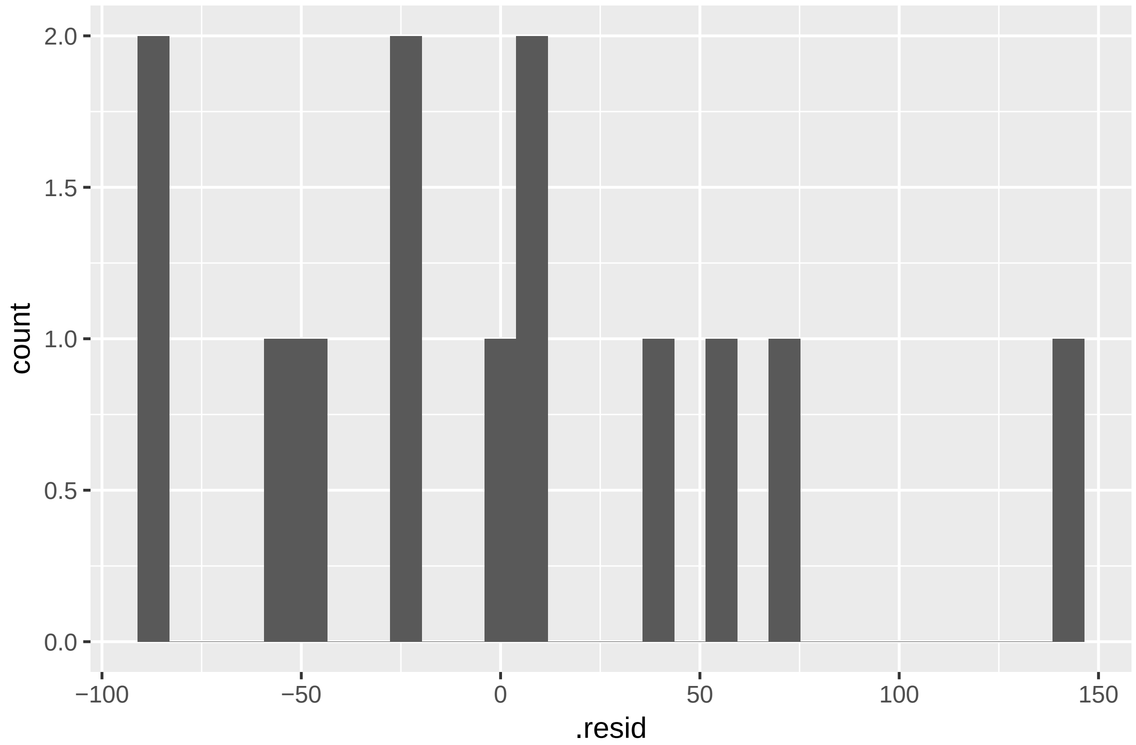
Note we still have to check our assumptions about ϵ_i. Recall that these are LINE

```r
full.lm2%>%ggplot(aes(x=.fitted,y=.resid))+geom_point()+
  geom_hline(yintercept=0)
```



```
full.lm2 %>% ggplot(aes(x=.resid))+geom_histogram()
```

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



Now, using what we've learned. Write out, using the language of statistics, *your* multiple regression model for your project:

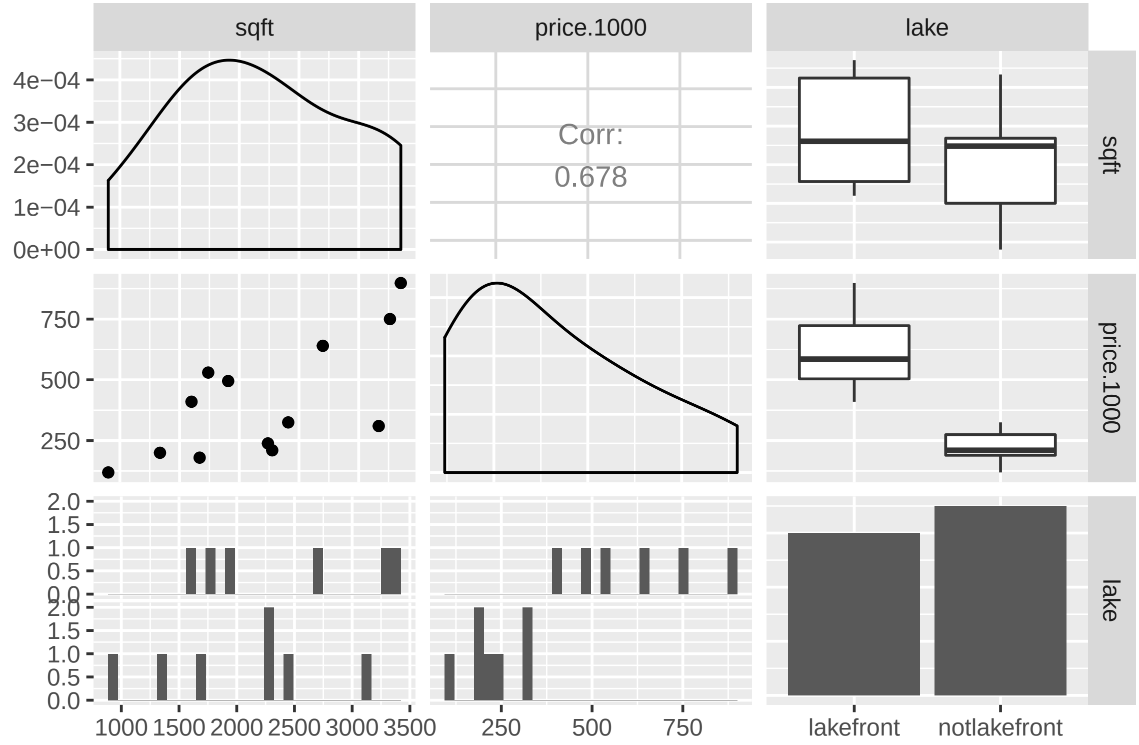
If you have the data start exploring your data. What plots might you want to use? What data might you want to show? Think about the relationships that might exist. Here's one visualization. Some of these are helpful, some are not.

```
library(GGally)
```

```
##
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
ggpairs(house.dat)
```

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

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


sqft
price.1000
lake
4e−04
3e−04
2e−04
1e−04
0e+00
Corr:
lakefront
notlakefront
sqft
price.1000
lake