Lsn 25 - MA206Y

Clark

Admin

Recall that last class we were exploring 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:

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

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

Let's write out the fitted model:

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:

Remember that the fitted model assumes that we have two parallel lines that best explain our data.

```
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))
```

We sort of hand waved this last class, but let's look at the fitted values vs the residuals

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

What strikes us as odd here?

For houses that are NOT on the lakefront, what happens to the residuals as the fitted values increase?

For houses that ARE on the lakefront, what happens to the residuals as the fitted values increase?

Going back to the picture of our fitted model from above, we can see this. As we move up the red line our points are now over the red line. As we move up the blue line our points are now below the blue line.

Perhaps a better fit would be:

```
house.dat = house.dat %>% group_by(lake)
house.dat %>% ggplot(aes(x=sqft,y=price.1000,color=lake))+geom_point()+
  stat_smooth(method="lm",se=FALSE,fullrange=T)
```

In order to account for this, we might consider adding **an interaction term**. To do this we use the statistical model:

The presence of an interaction helps address the question, does square feet impact price differently depending on whether a house is on the lakefront or not?

Our H_0 is, square feet impacts price the same whether the house is on the lakefront or not, the H_a is square feet impacts price differently depending on whether the house is on the lakefront or not.

The statistic that answers this question is $\hat{\beta}_3$ from our model above. In symbols our H_0 and H_a are:

To fit this model in R we fit:

```
inter.lm<-lm(price.1000~sqft+lake+sqft:lake,data=house.dat)
summary(inter.lm)
```

```
##
## Call:
## lm(formula = price.1000 ~ sqft + lake + sqft:lake, data = house.dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -54.16 -28.60 -14.15  29.64  73.93
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             86.76438   71.60612   1.212  0.25648
## sqft                     0.21990    0.02831   7.769  2.8e-05 ***
## lakenotlakefront       -28.65098   91.32560  -0.314  0.76088
## sqft:lakenotlakefront   -0.13595    0.03895  -3.491  0.00682 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.48 on 9 degrees of freedom
## Multiple R-squared:  0.9684, Adjusted R-squared:  0.9578
## F-statistic: 91.84 on 3 and 9 DF,  p-value: 4.547e-07
```

Our interpretation is:

The most difficult thing about writing out the fitted model is deciding what to do with `sqft:lakenotlakefront`. To see what this value means, let's write out the fitted model for a house on the lakefront.

Now write out the fitted model for a house not on the lakefront.

What does it mean that `lakenotlakefront` has a P value of 0.76 here? Does this mean that there is no difference between the price of houses on the lakefront and houses not on the lakefront?

Let's check our residuals now:

```
inter.lm2=augment(inter.lm)
```

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

Now we see much less of a pattern to our residuals.

```
APFT=read.csv("APFT.csv")
```

Let's build a model for weight and APFT_Score

```
APFT.lm=lm(APFT_Score~weight,data=APFT)
```

Write out the fitted model and interpret your results.

Let's look at the residuals

```
APFT.lm2=augment(APFT.lm)
APFT.lm2 %>%ggplot(aes(x=.fitted,y=.resid,color=APFT$CS))+geom_point()+
  geom_hline(yintercept=0)
```

Here we've colored by Corps Squad status. It sort of looks like the residuals for Corps Squad are higher than not Corps Squad. Perhaps this is an important explanatory variable. Write out a model for weight and Corps Squad

Fit the model and write out the fitted model for Corps Squaders and not Corps Squaders

Perhaps whether someone is Corps Squad impacts how weight effects APFT score. Write out a model that tests this hypothesis

Fit the model and check the assumptions.