Lsn 22 - MA206Y

Clark

Admin

Recall that the model we are fitting is:

Note that this is the **fitted model**. More correctly, the linear regression model is given by:

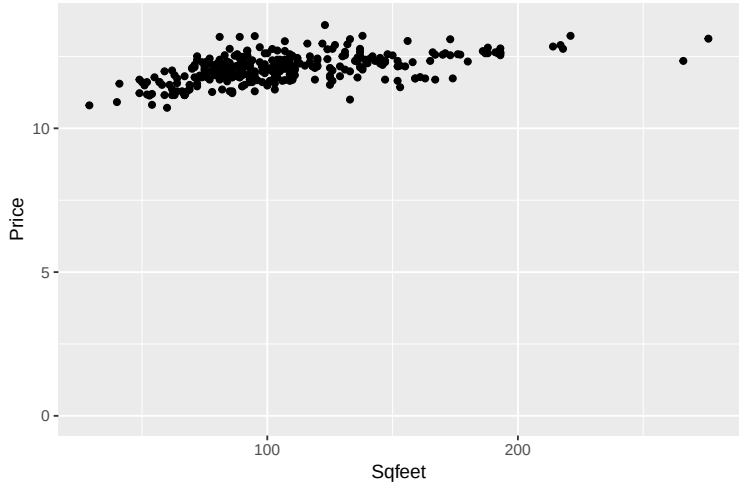
What role does ϵ_i play here?

For instance, awhile ago we were exploring the relationship between home prices as square footage.

```
GreenBay=read.csv("GreenBay.csv")
GreenBay = GreenBay %>% mutate(Sqfeet=X..SQUARE.FEET,Price=log(PRICE))
```

If we look at the data:

```
GreenBay %>% ggplot(aes(x=Sqfeet,y=Price))+geom_point()
```



Does there appear to be a linear relationship between Square Feet and log of price? Though it may seem weird that I'm using log of price here, I'll circle back to this at the end of the lesson.

What is our parameter we are interested in? **Note that our parameter comes from the unfit model**

What is H_0 and H_a?

The next step is to estimate our parameter, β_1. This is done through finding the least squares estimates. Here our parameters are estimated from:

```
house.lm=lm(Price~Sqfeet,data=GreenBay)
coef(house.lm)
```

```
##  (Intercept)       Sqfeet
## 11.352666900  0.006851998
```

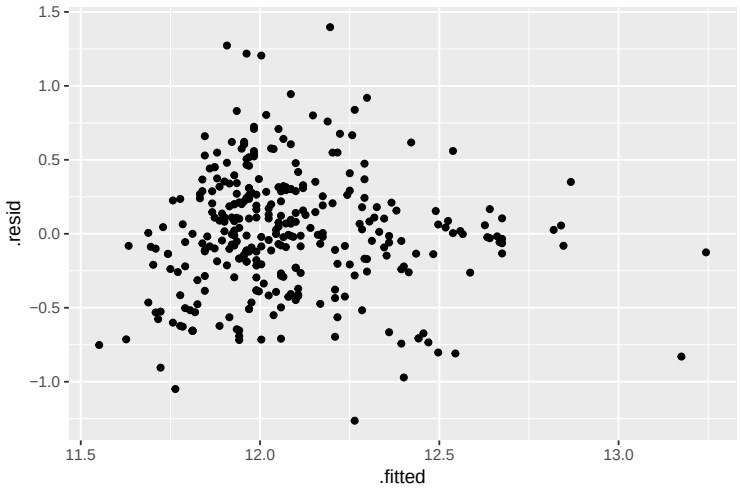
In contest of our problem what does this mean?

Note that we could stop here and find the simulation based p values, or else we could rely on validity conditions. An acronym that helps remember the validity conditions is **LINE**:

To check linearity and equal variance, we plot the **residuals** vs the predicted value. The residuals are:

This can be done by:

```
house.lm %>% ggplot(aes(x=.fitted,y=.resid))+geom_point()
```

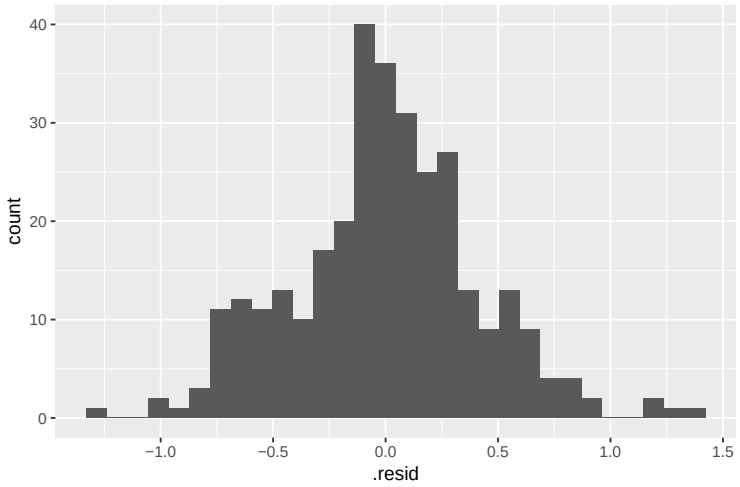


What we generally want to see here is a cloud of points centered at zero.

To check normality we can look at a histogram of the residuals

```
house.lm%>%ggplot(aes(x=.resid))+geom_histogram()
```

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



Which we want to be generally bell shaped.

Overall we're in pretty good shape here! If our validity conditions are met, then we can use the standardized statistic:

Whch has a *t distribution* with $n-2$ degrees of freedom. To find a P value for this, we run:

```
summary(house.lm)
```

```
##
## Call:
## lm(formula = Price ~ Sqfeet, data = GreenBay)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -1.2635 -0.2296 -0.0053  0.2512  1.3968
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.135e+01  6.960e-02   163.1   <2e-16 ***
## Sqfeet      6.852e-03  6.175e-04    11.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4138 on 316 degrees of freedom
##   (32 observations deleted due to missingness)
## Multiple R-squared:  0.2804, Adjusted R-squared:  0.2781
## F-statistic: 123.1 on 1 and 316 DF,  p-value: < 2.2e-16
```

Let's go through this line by line.

Here our conclusion is that there is strong statistical evidence that there is a relationship between square footage and log of price for houses in Green Bay.

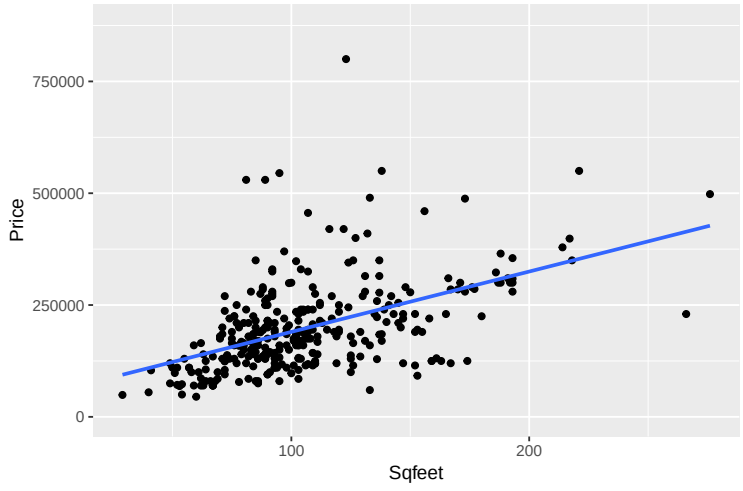
Our **fitted model** is:

Let's go back to our data and see what would have happened if we hadn't taken the log of price.

```
GreenBay=read.csv("GreenBay.csv")
GreenBay = GreenBay %>% mutate(Sqfeet=X..SQUARE.FEET,Price=PRICE)
```

Which we plot:

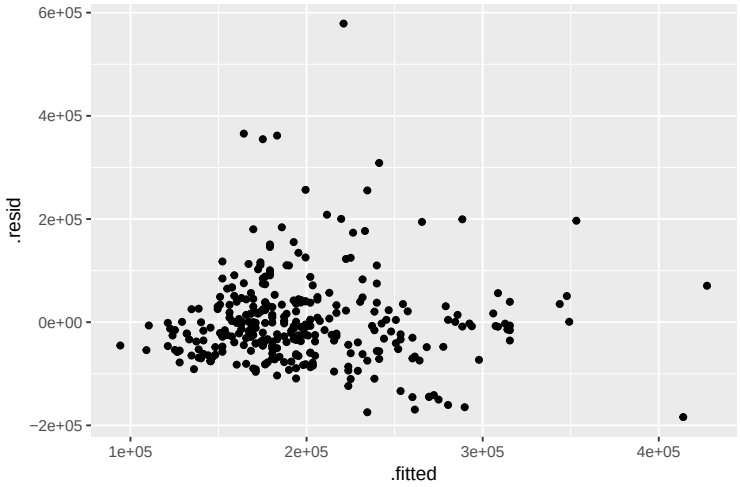
```
GreenBay %>% ggplot(aes(x=Sqfeet,y=Price))+geom_point()+
  stat_smooth(method="lm",se=FALSE)
```



Why might this concern us?

We can see this even more by:

```
house.lm=lm(Price~Sqfeet,data=GreenBay)
house.lm %>% ggplot(aes(x=.fitted,y=.resid))+geom_point()
```



So here we can still get estimates of our parameter, but we cannot use a theory based method as the validity conditions are not met.

To summarize:

Model:

Parameter:

Validity Condtions:

Statistic:

Hypothesis:

Fitted model:

Now you try one, let's consider textbook prices vs. number pages. The data can be found at:

```
textbook.dat=read.table("http://www.isi-stats.com/isi/data/chap10/TextbookPrices.txt",header=T)
```

First write out the population regression model. Make sure you define i, y_i, and x_i. Use Pages as your independent variable and Price as your dependent variable.

Looking at Page 10.5.15, Do parts a-c.

Note that the p value from `summary.lm` gives you the two sided P value. Can you figure out the p value for the hypothesis in part b.?

Write out the fitted model

Interpret the slope coefficient in context of the problem

Confidence intervals can be found using `confint()` of your lm object. For instance we could have done:

```
confint(house.lm)
```

```
##                 2.5 %   97.5 %
## (Intercept) 25817.347 84406.84
## Sqfeet       1088.924  1608.81
```

Use this to find a 95% CI for the slope. Interpret this CI.