Lesson 18

Clark

Admin

Recall that a simple linear regression model is written as:

And it explains how the variability in a quantitative outcome is explained through a quantitative explanatory variable. Often times though, we have multiple explanatory variables. For instance, we previously had models where we had two categorical explanatory variables:

But let's consider housing prices. It makes sense that the variability in prices could be explained through the square footage of the house. Why couldn't we use an ANOVA model for this?

Using square footage we could write:

Fitting the model is then done through:

```r
house.dat<-read.table("http://www.isi-stats.com/isi2/data/housing.txt",header=T)
house.dat<- house.dat %>% mutate(price=price.1000)%>% select(-price.1000)
sqft.lm<-lm(price~sqft,data=house.dat)
summary(sqft.lm)
```

```
##
## Call:
## lm(formula = price ~ 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
```

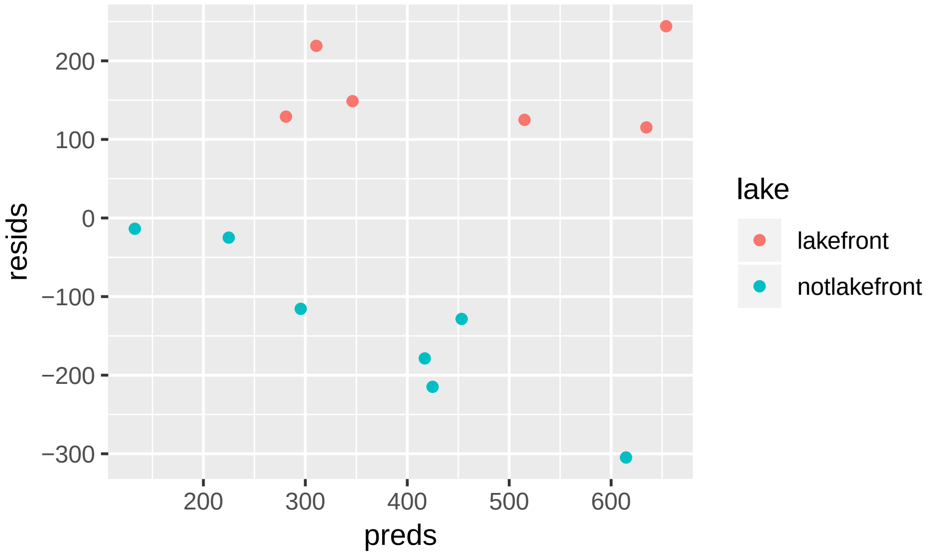
```
anova(sqft.lm)
```

```
## Analysis of Variance Table
##
## Response: price
##           Df Sum Sq Mean Sq F value  Pr(>F)
## sqft       1 319753  319753  9.3353 0.01094 *
## Residuals 11 376773   34252
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What are we testing with the two tests above?

Remember here the choices of using an ANOVA test or using the output of the linear regression model is a choice in using the F statistic or the $\hat{\beta}_1$ statistic. Either way, we have validity conditions.

```
fit.house <- house.dat %>% mutate(resids=sqft.lm$residuals,preds=sqft.lm$fitted.values)
fit.house %>% ggplot(aes(x=preds,y=resids,color=lake))+geom_point()
```



What do we notice here? Are we concerned?

If we want to see the effect of being a lakefront house or the effect of being a not lake front we could calculate

```
fit.house %>% group_by(lake)%>%summarize(mean=mean(resids))
```

```
## # A tibble: 2 x 2
##   lake          mean
##   <fct>        <dbl>
## 1 lakefront     163.
## 2 notlakefront -140.
```

However, our analysis is still not entirely straight forward, because if we know whether a house is lake front or not lakefront do we know anything about the square footage of the house?

```
fit.house %>% group_by(lake)%>%summarize(mean=mean(sqft))
```

```
## # A tibble: 2 x 2
##   lake          mean
##   <fct>        <dbl>
## 1 lakefront    2427
## 2 notlakefront 2000.
```

Note here I'm going to deviate slightly from the text. One way to adjust for one of the vairables that's in our model is to consider the statistical model for lakefront and price:

If we want to adjust for lakefront we would do:

```
contrasts(house.dat$lake)=contr.sum
lake.lm<-lm(price~lake,data=house.dat)
coef(lake.lm)
```

```
## (Intercept)       lake1
##    423.2321    197.2179
```

Thus we can adjust by adding 197.2 to every nonlakefront house and subtracting 197.2 to every lakefront house

```
house.dat.mod<-house.dat%>%mutate(price=ifelse(lake=="lakefront",price-197.2,price+197.2))

mod.lm<-lm(price~sqft,data=house.dat.mod)
coef(mod.lm)
```

```
## (Intercept)        sqft
## 124.9553631   0.1357554
```

Note that this is slightly different then what the book gives, the reason here is the meaning of μ vs β_0 when we have unbalanced design. Recall that when we were unbalanced μ wasn't the overall mean, but rather the mean of the means. This means when we built our model above and subtracted off the effects of lakefront or not lakefront we are left with $\mu + \epsilon_{i,j}$, which is fine, but that μ isn't the μ that we want for square footage…

So let's do this another way. Instead of subtracting off just the effects, let's subtract off everything except the unexplained varation.

```
house.dat.mod<-house.dat%>%mutate(lake.adj.price=lake.lm$residuals)
mod.lm<-lm(lake.adj.price~sqft,data=house.dat.mod)
coef(mod.lm)
```

```
##  (Intercept)         sqft
## -298.2600886    0.1357484
```

So not quite what our book has, but if we now add back μ, which is 408 to the intercept we get 110, which is what our book has.

The bottom line is this, to adjust for effects, fit a model and regress the new, additional variable on the residuals. A plot of this is called and added variable plot and is implimented in `library(car)` using `avPlot`

```
library(car)
```

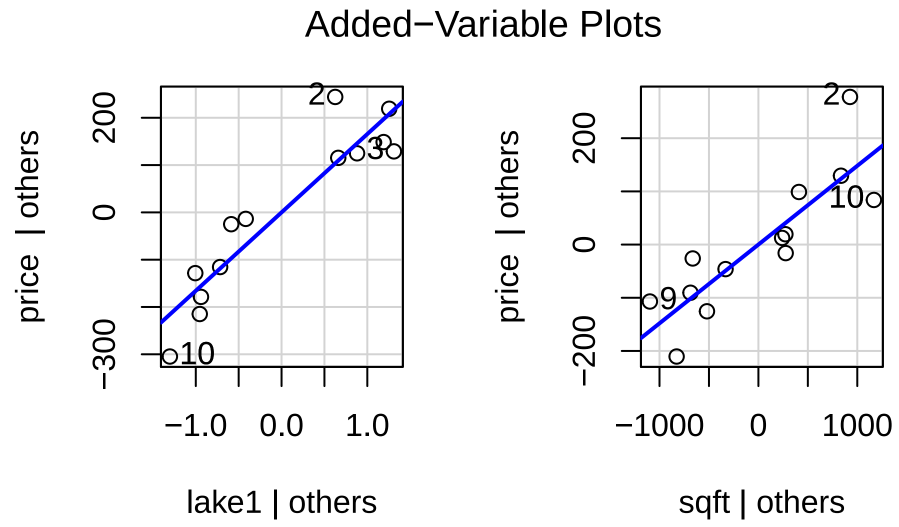
```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some
```

```
full.lm<-lm(price~lake+sqft,house.dat)
avPlots(full.lm)
```



Note that these values are centered, which we'll talk about later, but the bottom line is we are adjusting both square feet and price by lake effect and determining whether after accounting for lake effect is there still a relationship between square feet and price. As our book points out, these are useful if you want to visually explore whether a new explanatory variable explains additional variation.

Here we might decide that square feet does explain variation, so it makes sense to write a new model as:

Up to now we have been using effect coding as is natural for ANOVA, this was done by setting `contrasts(house.dat$lake)=contr.sum` when we do this we are saying $x_{2,i} = -1$ if observation i is lake front, -1 otherwise. Naturally R uses indicator variables instead, $x_{2,i} = 1$ if lakefront, 0 otherwise. As explained previously, it doesn't really matter. Here we'll stick with effect coding.

```r
summary(full.lm)
```

```
## 
## Call:
## lm(formula = price ~ lake + sqft, 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)  95.4296    65.7665   1.451 0.177405    
## lake1       165.6117    20.9235   7.915 1.29e-05 ***
## sqft          0.1481     0.0283   5.233 0.000383 ***
## ---
## 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
```

What are we testing here?

What is the fitted model for lakefront?

What is the fitted model for not lakefront?

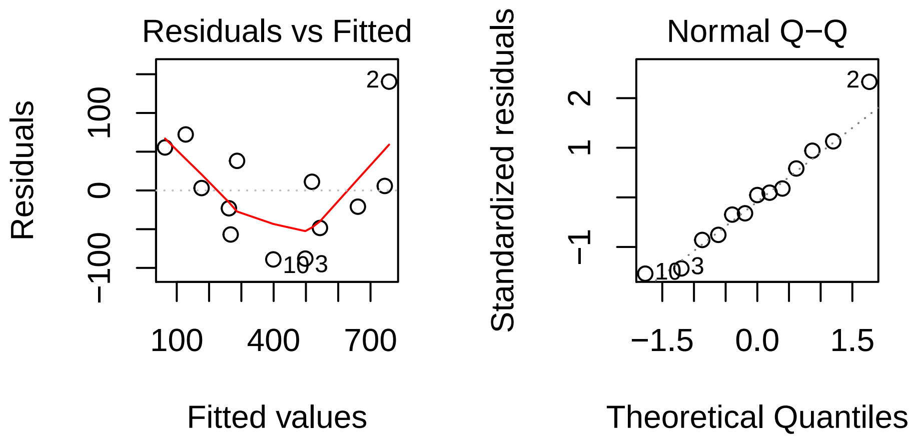
As we see in the fitted models what we have essentially done is fit two lines with the same slope but different intercepts

```r
Anova(full.lm,type=2)
```

```
## Anova Table (Type II tests)
## 
## Response: price
##           Sum Sq Df F value    Pr(>F)    
## lake      324911  1  62.649 1.293e-05 ***
## sqft      142022  1  27.384 0.0003826 ***
## Residuals  51862 10                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What is being tested here? Why do we need `type=2`?

```
par(mfrow=c(1,2))
plot(full.lm,which=c(1:2))
```


Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles