Lsn 20

Clark

Admin

Diamonds are expensive… But there's a lot of potential reasons why. A common way to examine the reasons for diamonds cost is the 4Cs (cut, clarity, color, and carat). In general, a bigger, more clear, colorless diamond is prefered. But we can explore this a bit more.

```
diamonds=read.table("http://www.isi-stats.com/isi2/data/diamonds.txt",header=T)
diamonds = diamonds %>% mutate(Price=Price..1000s.)%>% select(-Price..1000s.)
```

Conducting a unifariate analysis we might have:

```
single.lm<-lm(Price~Carat,data=diamonds)
summary(single.lm)
```

```
##
## Call:
## lm(formula = Price ~ Carat, data = diamonds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2819 -0.6242 -0.0978  0.3977  6.3380
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.3010     0.2543  -12.98   <2e-16 ***
## Carat        12.8426     0.3355   38.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.195 on 228 degrees of freedom
## Multiple R-squared:  0.8653, Adjusted R-squared:  0.8647
## F-statistic:  1465 on 1 and 228 DF,  p-value: < 2.2e-16
```

Here we see that 38.28^2=1465.4, which makes sense because the F test is comparing:

Which is the same thing as testing:

We can do another univariate analysis:

```r
clarity.lm<-lm(Price~Clarity,data=diamonds)
summary(clarity.lm)
```

```
##
## Call:
## lm(formula = Price ~ Clarity, data = diamonds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2052 -2.6522 -0.7788  2.5254  9.2928
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.390      1.016   7.278 5.62e-12 ***
## ClarityVS1    -2.246      1.082  -2.076   0.0391 *
## ClarityVS2    -1.489      1.111  -1.341   0.1814
## ClarityVVS1   -0.675      1.141  -0.591   0.5548
## ClarityVVS2   -1.102      1.101  -1.001   0.3179
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.211 on 225 degrees of freedom
## Multiple R-squared:  0.04039,    Adjusted R-squared:  0.02333
## F-statistic: 2.368 on 4 and 225 DF,  p-value: 0.05363
```

This model is:

So the F-statistic is conducting the ANOVA test

```r
anova(clarity.lm)
```

```
## Analysis of Variance Table
##
## Response: Price
##            Df  Sum Sq Mean Sq F value  Pr(>F)
## Clarity     4   97.66  24.415  2.3676 0.05363 .
## Residuals 225 2320.19  10.312
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Which doesn't have the same relationshp with the t statistic before. Note here we are using indicator coding instead of effects coding which means our intercept is:

In R we can see the levels using `levels(diamonds$Clarity)`. If we want to change which category is our reference category we can explicity relevel within R

```r
diamonds$modClarity<-factor(diamonds$Clarity,levels=c("VS1","VS2","VVS1","VVS2","IF"))
clarity2.lm<-lm(Price~modClarity,data=diamonds)
```

```r
summary(clarity2.lm)
```

```
##
## Call:
## lm(formula = Price ~ modClarity, data = diamonds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2052 -2.6522 -0.7788  2.5254  9.2928
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.1444     0.3733  13.781   <2e-16 ***
## modClarityVS2    0.7570     0.5844   1.295   0.1965
## modClarityVVS1   1.5708     0.6409   2.451   0.0150 *
## modClarityVVS2   1.1438     0.5659   2.021   0.0444 *
## modClarityIF     2.2458     1.0819   2.076   0.0391 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.211 on 225 degrees of freedom
## Multiple R-squared:  0.04039,    Adjusted R-squared:  0.02333
## F-statistic: 2.368 on 4 and 225 DF,  p-value: 0.05363
```

Which we can see doesn't change our group estimates, but does change our P values. Why?

But what model is better? Maybe we should use a model with both?

Which we can fit:

```r
both.lm<-lm(Price~Carat+Clarity,data=diamonds)
summary(both.lm)
```

```
##
## Call:
## lm(formula = Price ~ Carat + Clarity, data = diamonds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9982 -0.6078 -0.0376  0.4914  5.6904
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.2787     0.4265  -5.343 2.24e-07 ***
## Carat        12.9264     0.3196  40.450  < 2e-16 ***
## ClarityVS1   -1.0629     0.3774  -2.816  0.00529 **
## ClarityVS2   -1.6997     0.3863  -4.400 1.67e-05 ***
```

```
## ClarityVVS1  -0.4593     0.3970  -1.157  0.24850
## ClarityVVS2  -1.1619     0.3829  -3.034  0.00270 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.117 on 224 degrees of freedom
## Multiple R-squared:  0.8844, Adjusted R-squared:  0.8819
## F-statistic: 342.9 on 5 and 224 DF,  p-value: < 2.2e-16
```

Certainly our R^2 increased, but if we know Carat do we gain anything by knowing Clarity? To answer this we can look at the ANOVA model:

```
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
```

```
Anova(both.lm,type=3)
```

```
## Anova Table (Type III tests)
##
## Response: Price
##              Sum Sq  Df   F value    Pr(>F)
## (Intercept)   35.61   1   28.5518 2.240e-07 ***
## Carat       2040.80   1 1636.1764 < 2.2e-16 ***
## Clarity       46.20   4    9.2601 6.084e-07 ***
## Residuals    279.39 224
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that our SST is:

```
SST=sum((diamonds$Price-mean(diamonds$Price))^2)
SST
```

```
## [1] 2417.85
```

So, after adjusting for Carat, Clarity explains 46.2/2417.85 or 2% of the remaining variability whereas after adjusting for clarity, carat explains alost 85% of the remainng variability.

The question we want to ask is whether this model is better than the model with only Carat in this. Why can't we answer this question with the output we obtained in `both.lm`?

Since we have **nested models** we can statistically compare the two models. By nested models I mean:

Assuming our validity conditions are met, we can form the F statistic:

In R this is done through:

```
anova(single.lm,both.lm)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ Carat
## Model 2: Price ~ Carat + Clarity
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    228 325.59
## 2    224 279.39  4    46.201 9.2601 6.084e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note for the parital F test we aren't concerned with types of Sums of Squares as we are, by default, conducting a conditional test.

Because the F statistic is statistically significant, our conclusion is that the model with Clarity is prefered.

But perhaps we want a model that considers the interactions.

What, in words, does this model say about the relationship between Carat and price?

We can test if this model is prefered to a model with out interactions by:

```
inter.lm<-lm(Price~Carat*Clarity,data=diamonds)
anova(both.lm,inter.lm)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ Carat + Clarity
## Model 2: Price ~ Carat * Clarity
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    224 279.39
## 2    220 252.00  4    27.396 5.9793 0.0001384 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can get 95% CI for each of our β terms in this model by:

```
confint(inter.lm)
```

```
##                         2.5 %      97.5 %
```

```
## (Intercept)       -8.208004 -1.87934173
## Carat             12.487486 20.75816258
## ClarityVS1        -1.126390  5.34830568
## ClarityVS2        -1.172277  5.43469216
## ClarityVVS1       -3.436879  3.42510158
## ClarityVVS2       -2.258064  4.39347497
## Carat:ClarityVS1  -8.567357 -0.07121415
## Carat:ClarityVS2  -9.391346 -0.79079804
## Carat:ClarityVVS1 -5.026816  3.95549878
## Carat:ClarityVVS2 -7.322490  1.35227777
```

But, why might we want to adjust these CIs? One way to adjust is to use what is called Bonferonni Corrections. This technique uses α/k in lieu of α where k is the number of comparisons or tests being performed. Here we have 10 Confidence intervals, so Bonferonni corrections would say to use $.05/10 = .005$, or in order to guarantee an overall $\alpha = 0.05$, we should use 99.5 CI instead of 95 CI. This can be modified by:

```r
confint(inter.lm,level=0.995)
```

```
##                       0.25 %    99.75 %
## (Intercept)        -9.596560 -0.4907854
## Carat              10.672837 22.5728116
## ClarityVS1         -2.546988  6.7689030
## ClarityVS2         -2.621896  6.8843111
## ClarityVVS1        -4.942449  4.9306721
## ClarityVVS2        -3.717462  5.8528729
## Carat:ClarityVS1  -10.431476  1.7929040
## Carat:ClarityVS2  -11.278371  1.0962273
## Carat:ClarityVVS1  -6.997604  5.9262867
## Carat:ClarityVVS2  -9.225799  3.2555874
```

Interestingly if we use this technique what can we say? This in general is a very very conservative approach.

Looking at page 348 do we have concerns over our assumptions about ϵ_i?