Example4_5

Kevin Cummiskey

October 28, 2019

Review

Let y_i and $x_{1,i}$ be quantative variables and $x_{2,i}$ be a categorical variable with two levels.

Model 1: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i$

Model 1: $y_i = \alpha_0 + \alpha_1 x_{1,i} + \alpha_2 x_{2,i} + \alpha_3 x_{1,i} x_{2,i} + \epsilon_i$

For each of the following tests, which model and parameter would you use?

1. There is a linear association between x_2 and y after adjusting for x_1.
2. There is a linear association between x_1 and y for subjects in the reference group of x_2.
3. There is a linear assocation between x_1 and y for subjects not in the reference group of x_2.
4. The effect of x_1 on y differs by level of x_2.

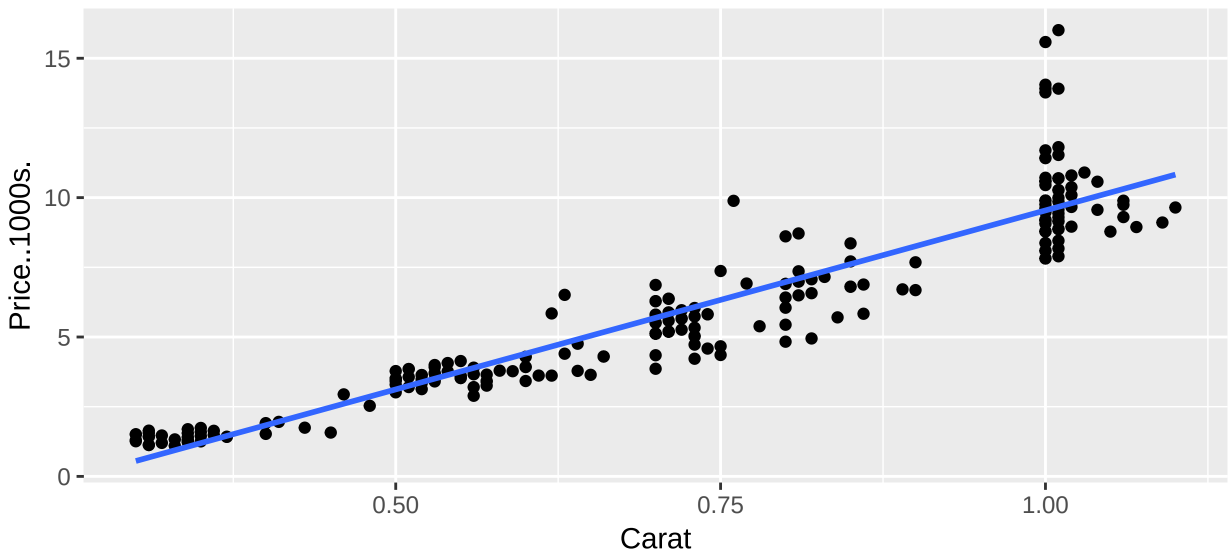
Why are we typically not interested in α_2?

One variable analyses

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

Price vs Weight

```
diamonds %>% ggplot(aes(x = Carat, y = Price..1000s.)) +
  geom_point() + geom_smooth(method = "lm", se = F)
```



```r
model_weight = lm(Price..1000s. ~ Carat, data = diamonds)
summary(model_weight)
```

```
##
## Call:
## lm(formula = Price..1000s. ~ 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
```

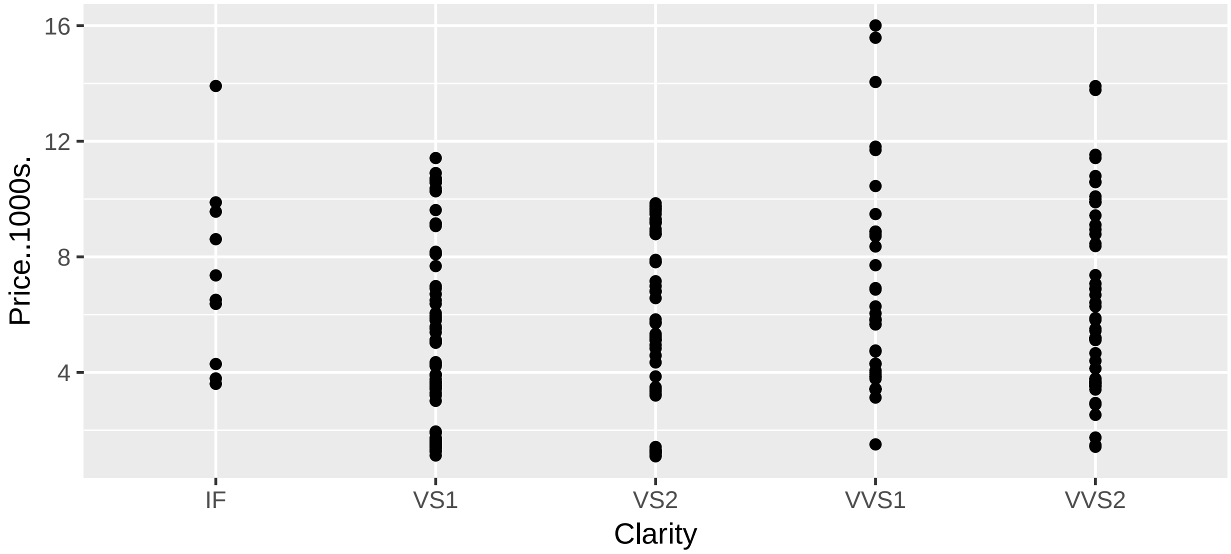
```r
anova(model_weight)
```

```
## Analysis of Variance Table
##
## Response: Price..1000s.
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## Carat       1 2092.26 2092.26  1465.1 < 2.2e-16 ***
## Residuals 228  325.59    1.43
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What would you conclude from this model?

Price vs clarity

```r
diamonds %>% ggplot(aes(x = Clarity, y = Price..1000s.)) +
  geom_point()
```



How many indicator variables do we need?

Here is the model:

$$y_i = \beta_0 + \beta_1 VS1_i + \beta_2 VS2_i + \beta_3 VVS1 + \beta_4 VVS2 + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)$$

What is the reference category?

```r
model_Clarity = lm(Price..1000s. ~ Clarity, data = diamonds)
summary(model_Clarity)
```

```
## 
## Call:
## lm(formula = Price..1000s. ~ 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
```

```
anova(model_Clarity)
```

```
## Analysis of Variance Table
##
## Response: Price..1000s.
##            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
```

What would you conclude from this model? Why are all the clarity coefficients negative?

Two variable analysis

Are there differences in price across clarity categories after adjusting for weight?

Here is the model:

$$y_i = \beta_0 + \beta_1 Carat_i + \beta_2 VS1_i + \beta_3 VS2_i + \beta_4 VVS1 + \beta_5 VVS2 + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)$$

```
model_ClarityWeight = lm(Price..1000s. ~ Carat + Clarity, data = diamonds)
summary(model_ClarityWeight)
```

```
##
## Call:
## lm(formula = Price..1000s. ~ 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
```

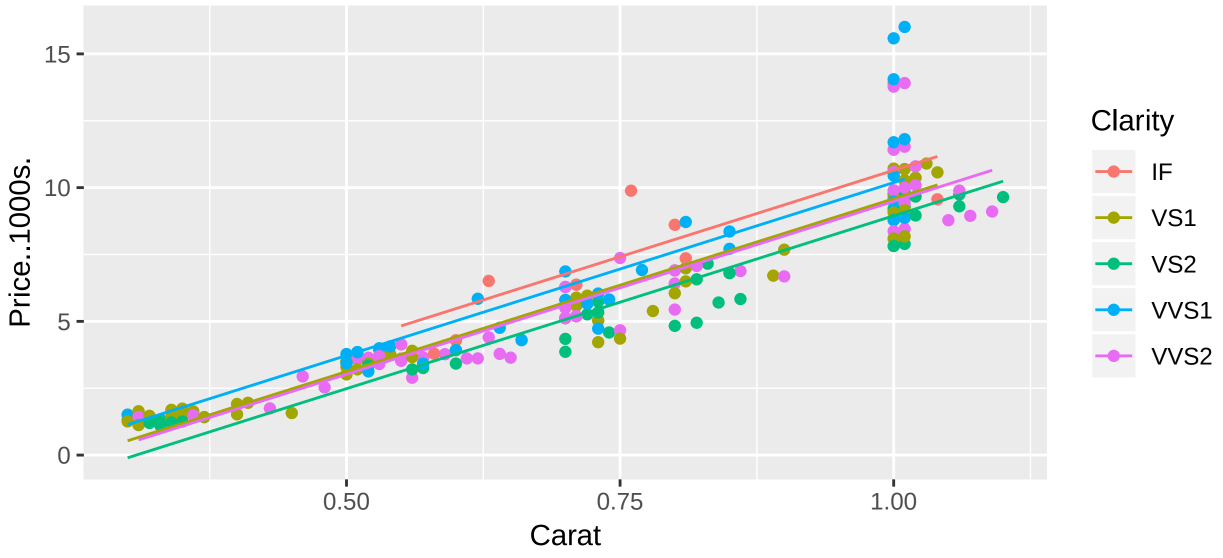
```
anova(model_ClarityWeight)
```

```
## Analysis of Variance Table
##
## Response: Price..1000s.
##            Df  Sum Sq Mean Sq   F value    Pr(>F)
## Carat       1 2092.26 2092.26 1677.4324 < 2.2e-16 ***
## Clarity     4   46.20   11.55    9.2601 6.084e-07 ***
## Residuals 224  279.39    1.25
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How have the coefficients changed from the one-variable models?

Write out the regression equations for the five categories of clarity. What is the relationship among these regression lines?

```
diamonds = diamonds %>%
  mutate(predicted2 = predict(model_ClarityWeight, diamonds))
diamonds %>% ggplot(aes(x = Carat, y = Price..1000s., color = Clarity)) +
  geom_point() + geom_line(aes(y = predicted2))
```



Draw inference.

H_0 : There is no linear association between clarity and price, after adjusting for diamond weight.

H_a : There is a linear association between clarity and price, after adjusting for diamond weight.

Why not use the p-values for each indicator variable?

Perform the partial F-test (pg 344).

What do the p-values of the indicator variable coefficients tell us?

Does the price increase associated diamond weight differ across clarity categories?

H_0 : There is no interaction between clarity and price, after adjusting for diamond weight and diamond clarity.

H_a : There is an interaction between clarity and price, after adjusting for diamond weight and diamond clarity.

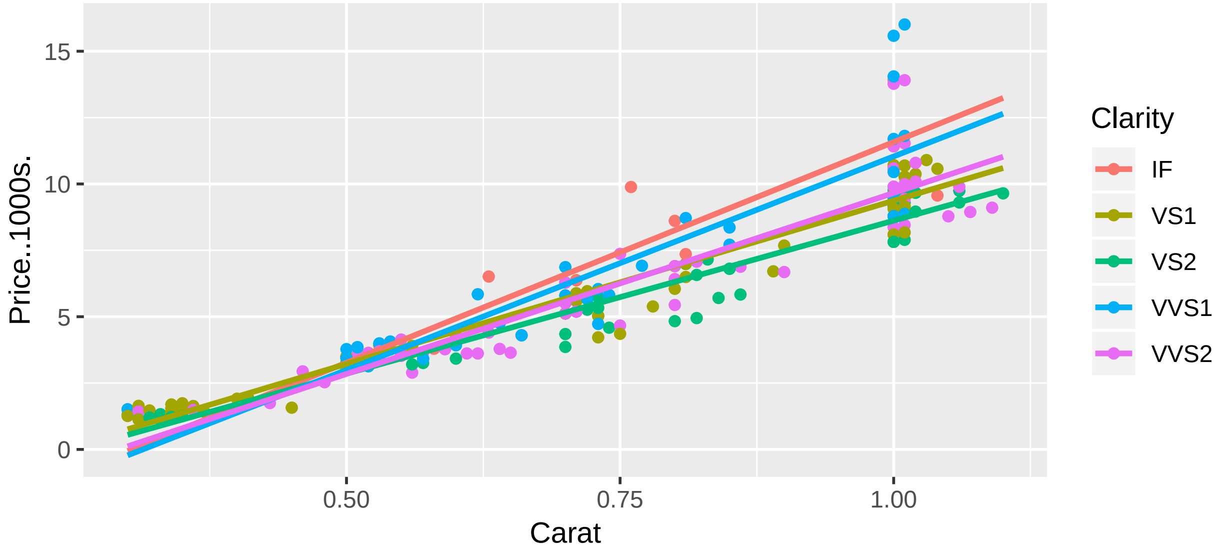
```r
model_interaction = lm(Price..1000s. ~ Carat*Clarity, data = diamonds)
summary(model_interaction)
```

```
##
## Call:
## lm(formula = Price..1000s. ~ Carat * Clarity, data = diamonds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6811 -0.5578 -0.0072  0.5191  4.8095
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -5.043673   1.605603  -3.141  0.00191 **
## Carat            16.622824   2.098298   7.922 1.15e-13 ***
## ClarityVS1        2.110958   1.642652   1.285  0.20011
## ClarityVS2        2.131208   1.676210   1.271  0.20491
## ClarityVVS1      -0.005889   1.740908  -0.003  0.99730
## ClarityVVS2       1.067706   1.687518   0.633  0.52758
## Carat:ClarityVS1  -4.319286   2.155500  -2.004  0.04631 *
## Carat:ClarityVS2  -5.091072   2.181988  -2.333  0.02054 *
## Carat:ClarityVVS1 -0.535659   2.278844  -0.235  0.81438
## Carat:ClarityVVS2 -2.985106   2.200818  -1.356  0.17637
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.07 on 220 degrees of freedom
## Multiple R-squared:  0.8958, Adjusted R-squared:  0.8915
## F-statistic: 210.1 on 9 and 220 DF,  p-value: < 2.2e-16
```

```
anova(model_interaction)
```

```
## Analysis of Variance Table
##
## Response: Price..1000s.
##                Df Sum Sq Mean Sq   F value    Pr(>F)
## Carat           1 2092.3 2092.26 1826.5840 < 2.2e-16 ***
## Clarity         4   46.2   11.55   10.0835 1.638e-07 ***
## Carat:Clarity   4   27.4    6.85    5.9793 0.0001384 ***
## Residuals     220  252.0    1.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
diamonds %>% ggplot(aes(x = Carat,y = Price..1000s., color=Clarity)) +
  geom_point() + geom_smooth(method = "lm", se = F, fullrange = T)
```



Perform the partial F-test.

Confidence Intervals

```
confint(model_interaction)
```

```
##                         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
```