STAT 224 Lecture 9
Models with Ordinal Categorical Predictors

Yibi Huang
Department of Statistics
University of Chicago

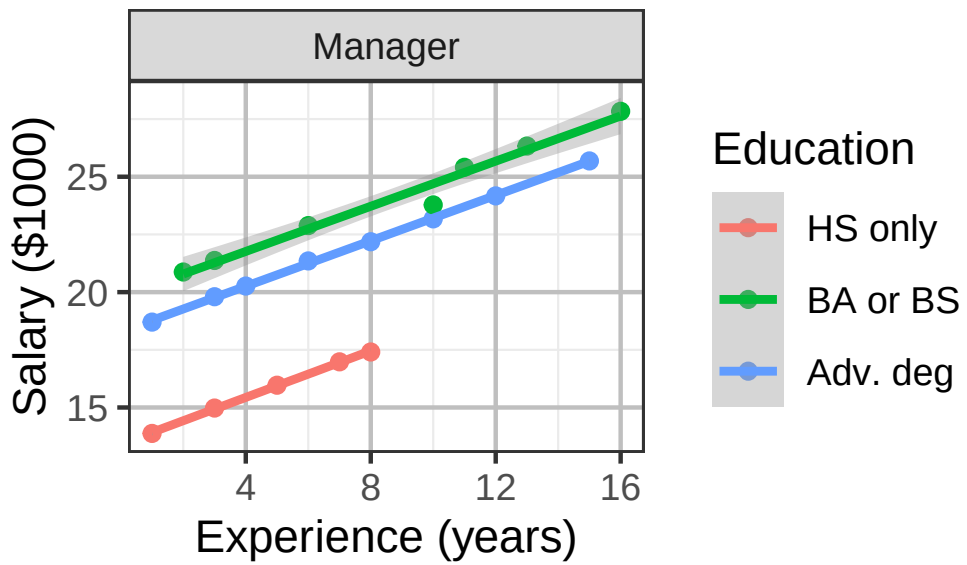
Ordinal Categorical Predictors

An **ordinal** variable is a categorical variable w/ **ordered** categories.

- e.g., E = education (HS only, BA or BS, advance degree) in the salary survey data is an ordinal variable

When we create indicators for E and include them in a model, we *ignore* the fact that the 3 education levels are *ordered*. The estimated salary might not be ordered by education levels.

e.g., in the salary survey data, managers w/ a Bachelor's degree earn more than managers w/ an advanced degree



Ordinal Categorical Predictors

We can incorporate the ordinal info of a ordinal predictor by assigning a *score* to each its category like

$$E = \begin{cases} 1 & \text{if HS only,} \\ 2 & \text{if Bachelor's degree,} \\ 3 & \text{if Advanced degree} \end{cases}$$

and include `E` as a **numerical** predictor in the model

$$S = \beta_0 + \alpha M_1 + \beta_1 E + \beta_2 X + \varepsilon$$

instead of including the indicator variables for `E` in the model.

This way, the fitted salary will always be ordered by education levels

- a BA or BS increases salary by β_1
- an advanced degree increases salary by another β_1

Models w/ Ordinal Predictors

If one believe the salary gap btw a Bachelor's deg. and HS diploma is greater than that btw a Bachelor's deg. and an adv. deg., one may try a different scoring $(1, 2.5, 3)$, i.e.,

$$E = \begin{cases} 1 & \text{if HS only,} \\ 2.5 & \text{if Bachelor's degree,} \\ 3 & \text{if Advanced degree} \end{cases}$$

Then based on the model $S = \beta_0 + \alpha M_1 + \beta_1 E + \beta_2 X + \varepsilon$, the salary gaps would be

- $1.5\beta_1$ between HS only and BA or BS,
- $0.5\beta_1$ between BA or BS and adv. deg.

Models w/ Ordinal Predictors

If one use the scoring $(1, 3, 4)$, i.e.,

$$E = \begin{cases} 1 & \text{if HS only,} \\ 3 & \text{if Bachelor's degree,} \\ 4 & \text{if Advanced degree} \end{cases}$$

the gaps would be

- $2\beta_1$ between HS only and BA or BS, and
- β_1 between BA or BS and adv. deg.

Fitting Models w/ E As An Ordinal Predictor

Say we want to fit the model $S = \beta_0 + \alpha M_1 + \beta_1 E + \beta_2 X + \varepsilon$ while E is scored as $(1, 2.5, 3)$.

```
p130 = read.table("P130.txt",header=TRUE)
p130$E.score1 = ifelse(p130$E == 2, 2.5, p130$E)
summary(lm(S ~ M + E.score1 + X, data=p130))$coef
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)   6401.1     561.46   11.40 1.943e-14
M             6741.1     330.94   20.37 2.184e-23
E.score1      1703.3     204.22    8.34 1.883e-10
X              562.2      32.01   17.57 5.779e-21
```

- a BA or BS increases mean salary by $1.5\widehat{\beta_1} = 1.5 \times 1703 = 2554.5$
- an advanced degree increases mean salary by another $0.5\widehat{\beta_1} = 0.5 \times 1703 = 851.5$

Here is another model with a different scoring $(1, 3, 4)$ for `E`:

```
p130$E.score2 = ifelse(p130$E >= 2, p130$E + 1, p130$E)
summary(lm(S ~ M + E.score2 + X, data=p130))$coef
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)      7072     535.55  13.205 1.517e-16
M                6712     349.59  19.199 2.074e-22
E.score2         1135     148.44   7.647 1.751e-09
X                 566      33.82  16.736 3.442e-20
```

- a BA or BS increases mean salary by $2\widehat{\beta_1} = 2 \times 1135 = 2270$
- an advanced degree increases mean salary by another $\widehat{\beta_1} = 1135$

Here is another model with a different scoring $(1, 3, 4)$ for `E`:

```
p130$E.score2 = ifelse(p130$E >= 2, p130$E + 1, p130$E)
summary(lm(S ~ M + E.score2 + X, data=p130))$coef
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)      7072     535.55  13.205 1.517e-16
M                6712     349.59  19.199 2.074e-22
E.score2         1135     148.44   7.647 1.751e-09
X                 566      33.82  16.736 3.442e-20
```

- a BA or BS increases mean salary by $2\widehat{\beta_1} = 2 \times 1135 = 2270$
- an advanced degree increases mean salary by another $\widehat{\beta_1} = 1135$

Which model fits better? Compare their multiple R^2:

```
summary(lm(S ~ M + E.score1 + X, data=p130))$r.squared
[1] 0.9493
summary(lm(S ~ M + E.score2 + X, data=p130))$r.squared
[1] 0.9437
```

Comparison of Models Ordinal and Nominal Predictors

Whatever scoring one uses for `E`, the model

$$S = \beta_0 + \alpha M_1 + \beta_1 E + \beta_2 X + \varepsilon$$

is always nested to the model that treats `E` as nominal

$$S = \beta'_0 + \alpha M_1 + \delta_2 E_2 + \delta_3 E_3 + \beta_2 X + \varepsilon.$$

Why?

Comparison of Models Ordinal and Nominal Predictors

Whatever scoring one uses for `E`, the model

$$S = \beta_0 + \alpha M_1 + \beta_1 E + \beta_2 X + \varepsilon$$

is always nested to the model that treats `E` as nominal

$$S = \beta_0' + \alpha M_1 + \delta_2 E_2 + \delta_3 E_3 + \beta_2 X + \varepsilon.$$

Why?

e.g., for the scoring (1, 3, 4), the second model become the first model when

$$\beta_0' = \beta_0 + \beta$$
$$\delta_2 = 2\beta$$
$$\delta_3 = 3\beta$$

```
anova(lm(S ~ M + E.score1 + X, data=p130),
      lm(S ~ M + as.factor(E) + X, data=p130))
Analysis of Variance Table

Model 1: S ~ M + E.score1 + X
Model 2: S ~ M + as.factor(E) + X
  Res.Df      RSS Df Sum of Sq    F Pr(>F)
1     42 50750487
2     41 43280719  1   7469768 7.08  0.011
```

Pros and Cons of Using Ordinal Predictors

Pros: Simplicity, fewer parameters

- A categorical predictor with c categories requires $c - 1$ parameters if regarded as nominal since each indicator variable needs 1 parameter
- An ordinal predictor that uses the scores as the numerical values for its categories of need only 1 parameter

Cons:

- The choice of scores seems arbitrary
- If the scores are not chosen properly, the model might not be reasonable.

Interactions between Ordinal and Numerical Predictors

If we use `E` as an ordinal predictor with the scoring (1, 2.5, 3), we may as well consider models with `E*X` interactions like

$$S = \beta_0 + \alpha M_1 + \beta_1 E + \beta_2 X + \beta_3 (E \cdot X) + \varepsilon$$

Based on this model, every extra year of experience X increases mean salary by

- $\beta_2 + \beta_3$ if HS only
- $\beta_2 + 2.5\beta_3$ if BA or BS only
- $\beta_2 + 3\beta_3$ if advanced degree

Example (Diamonds Data)

Records of 3000 diamonds, randomly sampled from `http://www.diamondse.info/` in 2008.

You can download the data at

`http://www.stat.uchicago.edu/~yibi/s224/data/diamonds3000.txt`

change working directory, and load the data by

```
diamonds  = read.table("diamonds3000.txt", header = TRUE, sep="\t")
```

Diamonds Data

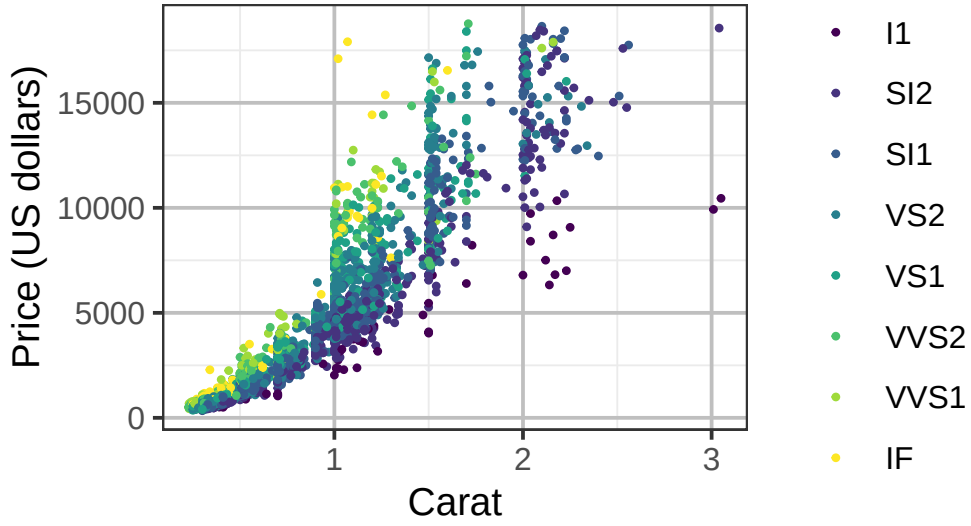
The variables include

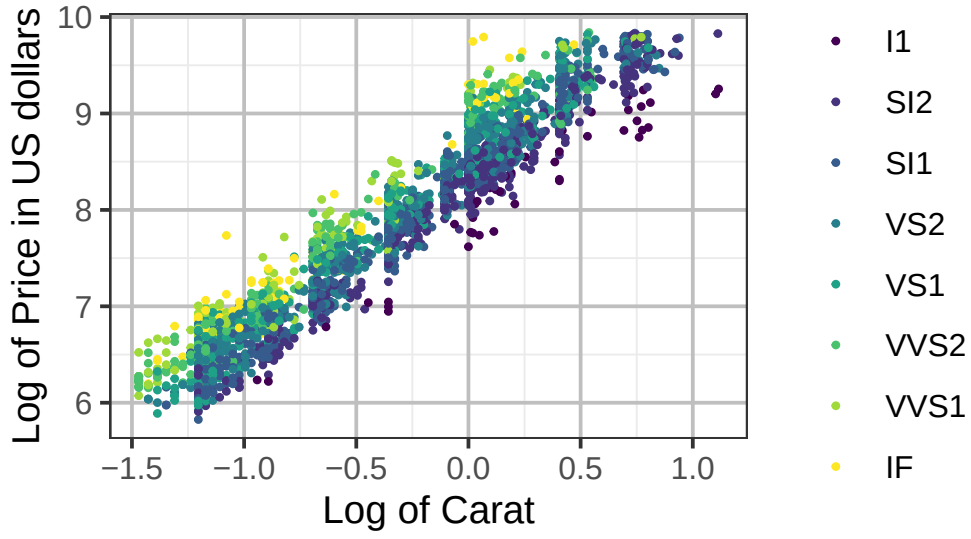
- `price`: price in US dollars
- `carat`: weight of the diamond
- `clarity`: a measurement of how clear the diamond is ordinal with 8 categories from

  I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)

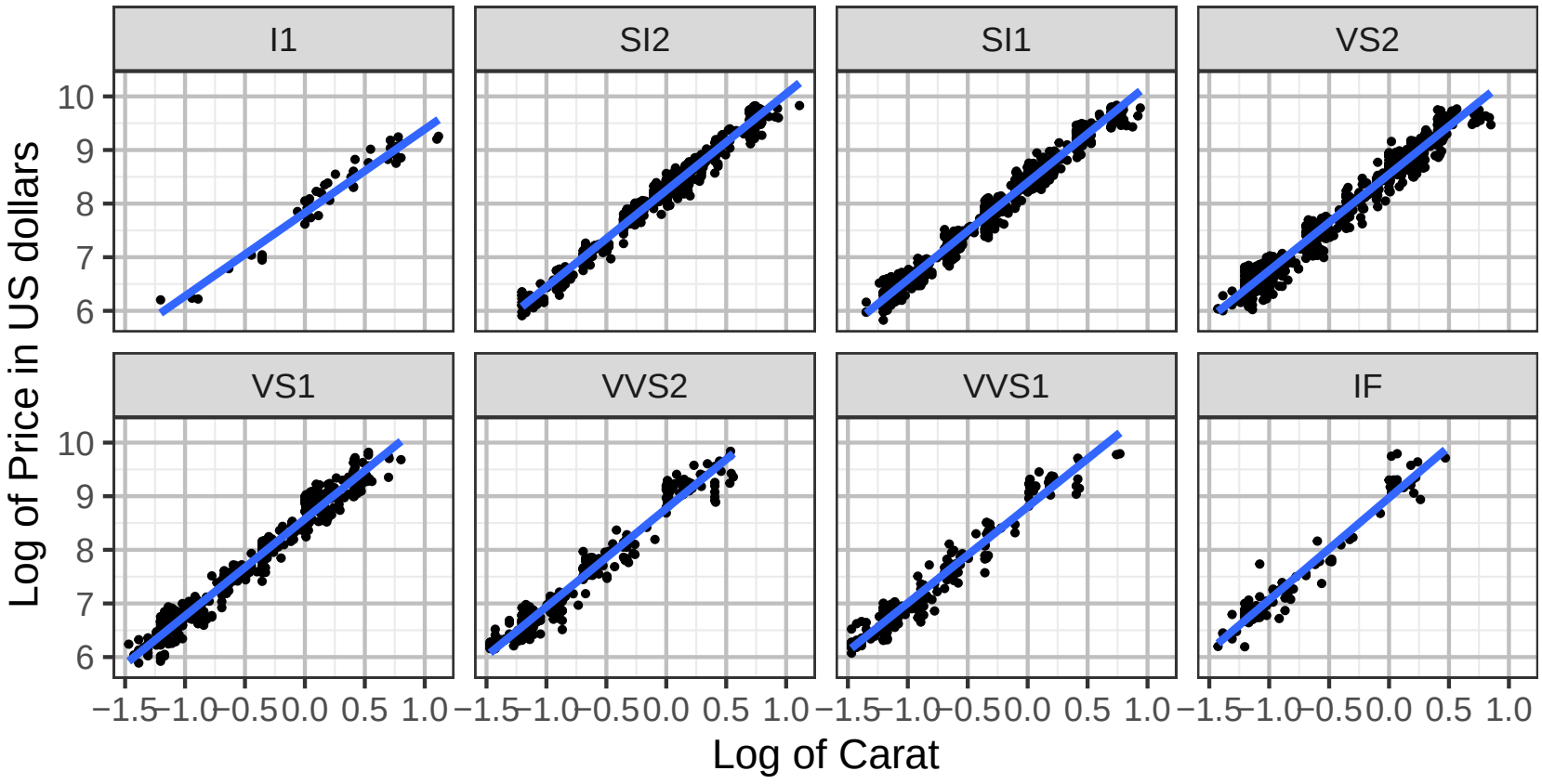
```
diamonds$clarity =
  ordered(diamonds$clarity,
          levels = c("I1", "SI2", "SI1", "VS2", "VS1",
                     "VVS2", "VVS1", "IF"))
```

Linear assumption seems more appropriate when `price` and `carat` are both log-transformed.



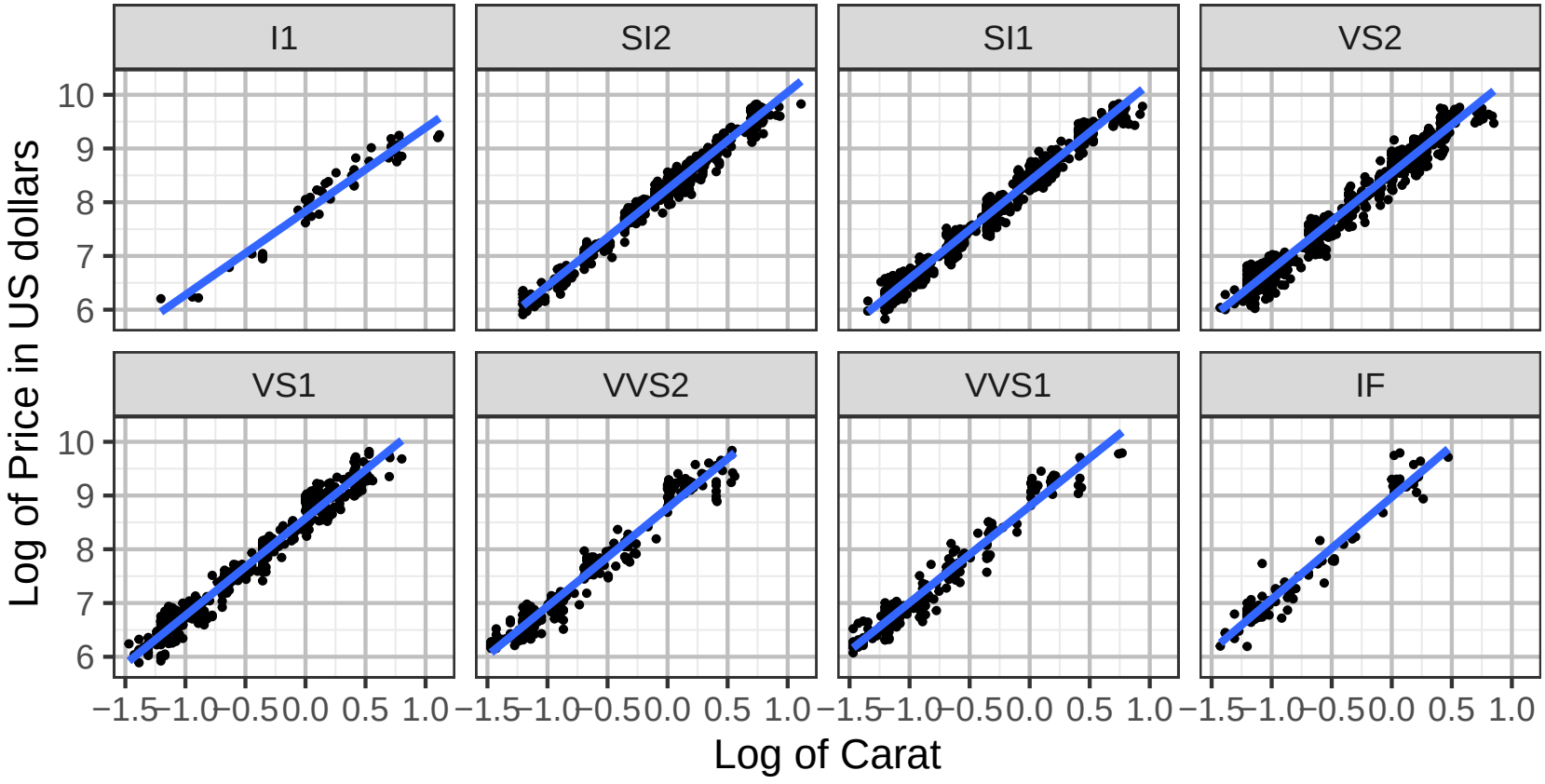


```
ggplot(diamonds, aes(x = log(carat), y = log(price))) +
    geom_point(cex=0.5) + facet_wrap(~clarity, ncol=4) +
  geom_smooth(method='lm', formula='y~x', se=FALSE) +
  labs(x="Log of Carat", y="Log of Price in US dollars")
```



- `log(price)` is linear in `log(carat)` for each level of `clarity`

```
ggplot(diamonds, aes(x = log(carat), y = log(price))) +
    geom_point(cex=0.5) + facet_wrap(~clarity, ncol=4) +
  geom_smooth(method='lm', formula='y~x', se=FALSE) +
  labs(x="Log of Carat", y="Log of Price in US dollars")
```



- `log(price)` is linear in `log(carat)` for each level of `clarity`
- Are the regression lines *parallel* w/ the *same slope*?

```
lm1 = lm(log(price) ~ log(carat) + clarity, data=diamonds)
lm2 = lm(log(price) ~ log(carat) * clarity, data=diamonds)
anova(lm1,lm2)
Analysis of Variance Table

Model 1: log(price) ~ log(carat) + clarity
Model 2: log(price) ~ log(carat) * clarity
  Res.Df RSS Df Sum of Sq    F   Pr(>F)
1   2991 103
2   2984 102  7      1.12 4.67 0.000032
```

Why do the two models differ by 7 degrees of freedom?

The tiny P-value means the slopes of the 8 regression lines are not all the same.

```
summary(lm2)$coef
                          Estimate Std. Error   t value   Pr(>|t|)
(Intercept)               8.51757   0.007078 1203.3895  0.000e+00
log(carat)                1.78802   0.009646  185.3573  0.000e+00
clarity.L                 0.91905   0.026399   34.8135 3.685e-223
clarity.Q                -0.18910   0.025316   -7.4697  1.050e-13
clarity.C                 0.15422   0.022522    6.8475  9.088e-12
clarity^4                -0.05699   0.018976   -3.0035  2.691e-03
clarity^5                 0.04321   0.016258    2.6577  7.909e-03
clarity^6                 0.02735   0.014192    1.9274  5.402e-02
clarity^7                 0.07001   0.011603    6.0334  1.804e-09
log(carat):clarity.L      0.18983   0.036943    5.1384  2.950e-07
log(carat):clarity.Q     -0.08784   0.036047   -2.4368  1.488e-02
log(carat):clarity.C      0.14147   0.031073    4.5529  5.502e-06
log(carat):clarity^4     -0.04252   0.025032   -1.6985  8.952e-02
log(carat):clarity^5      0.06662   0.020431    3.2605  1.124e-03
log(carat):clarity^6      0.01802   0.017530    1.0282  3.040e-01
log(carat):clarity^7      0.00581   0.015163    0.3832  7.016e-01
```

Oops! What are those `clarity.L`, `clarity.Q`, `clarity.C`, `clarity^4`, etc?

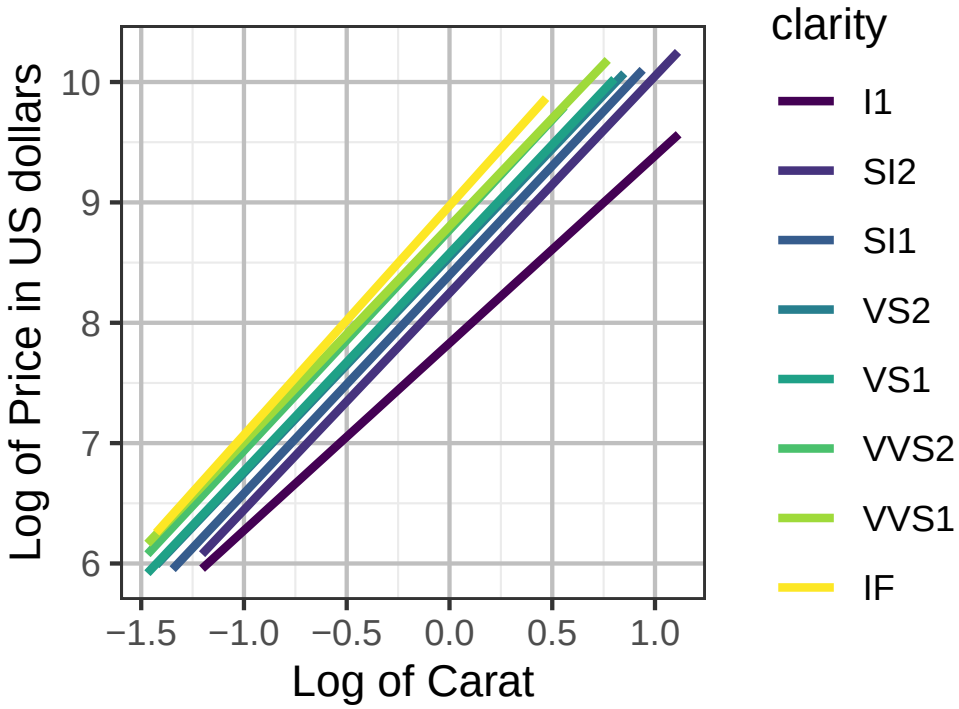
```
diamonds$clarity2 = factor(diamonds$clarity, ordered=FALSE)
summary(lm(log(price) ~ log(carat) * clarity2, data=diamonds))$coef
                       Estimate Std. Error t value   Pr(>|t|)
(Intercept)              7.8306    0.03000 261.037  0.000e+00
log(carat)               1.5562    0.05327  29.210 3.405e-165
clarity2SI2              0.4164    0.03110  13.387  9.814e-40
clarity2SI1              0.5657    0.03097  18.268  9.324e-71
clarity2VS2              0.7164    0.03130  22.889 5.820e-107
clarity2VS1              0.7453    0.03218  23.160 2.857e-109
clarity2VVS2             0.9399    0.03491  26.925 3.973e-143
clarity2VVS1             0.9713    0.03795  25.594 8.684e-131
clarity2IF               1.1409    0.04478  25.475 1.056e-129
log(carat):clarity2SI2   0.2456    0.05571   4.409  1.077e-05
log(carat):clarity2SI1   0.2580    0.05474   4.713  2.548e-06
log(carat):clarity2VS2   0.2383    0.05473   4.355  1.376e-05
log(carat):clarity2VS1   0.2512    0.05555   4.522  6.350e-06
log(carat):clarity2VVS2  0.2766    0.05691   4.861  1.231e-06
log(carat):clarity2VVS1  0.2380    0.05862   4.061  5.019e-05
log(carat):clarity2IF    0.3471    0.06425   5.402  7.099e-08
```

- Which indicator variable of `clarity` is dropped?
- What are the slopes for
  - `clarity` = I1?
  - `clarity` = SI2?
- Do the slopes of the lines increase with the clarity level?
- All terms are significant. Can we simplify the model further?

- Which indicator variable of `clarity` is dropped? I1
- What are the slopes for
  - `clarity` = I1?
  - `clarity` = SI2?
- Do the slopes of the lines increase with the clarity level?
- All terms are significant. Can we simplify the model further?

- Which indicator variable of `clarity` is dropped? I1
- What are the slopes for
  - `clarity` = I1? 1.5562
  - `clarity` = SI2? 1.5562 + 0.2456 = 1.8018
- Do the slopes of the lines increase with the clarity level?
- All terms are significant. Can we simplify the model further?

```
ggplot(diamonds, aes(x = log(carat), y = log(price), color=clarity)) +
  geom_smooth(method='lm', formula='y~x', se=FALSE) +
  labs(x="Log of Carat", y="Log of Price in US dollars")
```



The slopes for diamonds of the worst (I1) and the best (IF) clarity might be different from the rest.

```
diamonds$clarityI1 = ifelse(diamonds$clarity=="I1", 1, 0)
diamonds$clarityIF = ifelse(diamonds$clarity=="IF", 1, 0)
lm2a = lm(log(price) ~ log(carat) + clarity2 +
            log(carat):clarityI1 +log(carat):clarityIF, data=diamonds)
anova(lm2a,lm2)
Analysis of Variance Table

Model 1: log(price) ~ log(carat) + clarity2 + log(carat):clarityI1 + lo
Model 2: log(price) ~ log(carat) * clarity
  Res.Df RSS Df Sum of Sq    F Pr(>F)
1   2989 102
2   2984 102  5     0.115 0.67   0.65
```

```
summary(lm2a)$coef
                     Estimate Std. Error t value   Pr(>|t|)
(Intercept)           7.83060   0.029990 261.109  0.000e+00
log(carat)            1.80662   0.006369 283.674  0.000e+00
clarity2SI2           0.41661   0.031084  13.402  8.066e-40
clarity2SI1           0.56363   0.030817  18.290  6.415e-71
clarity2VS2           0.72163   0.030932  23.330 9.834e-111
clarity2VS1           0.74493   0.031393  23.729 3.474e-114
clarity2VVS2          0.92184   0.032356  28.491 3.666e-158
clarity2VVS1          0.98116   0.032953  29.775 7.899e-171
clarity2IF            1.14091   0.044772  25.482 8.709e-130
log(carat):clarityI1 -0.25046   0.053639  -4.669  3.154e-06
log(carat):clarityIF  0.09663   0.036465   2.650  8.093e-03
```

```
lm2a$coef
         (Intercept)          log(carat)         clarity2SI2
             7.83060             1.80662             0.41661
         clarity2SI1         clarity2VS2         clarity2VS1
             0.56363             0.72163             0.74493
        clarity2VVS2        clarity2VVS1         clarity2IF
             0.92184             0.98116             1.14091
log(carat):clarityI1 log(carat):clarityIF
            -0.25046             0.09663
```

$$\widehat{\log(\text{price})} = \begin{cases} 7.8306 + (1.8066 - 0.2505)\log(\text{carat}) & \text{for I1} \\ 7.8306 + 0.4166 + 1.8066\log(\text{carat}) & \text{for SI2} \\ 7.8306 + 0.5636 + 1.8066\log(\text{carat}) & \text{for SI1} \\ 7.8306 + 0.7216 + 1.8066\log(\text{carat}) & \text{for VS2} \\ 7.8306 + 0.7449 + 1.8066\log(\text{carat}) & \text{for VS1} \\ 7.8306 + 0.9218 + 1.8066\log(\text{carat}) & \text{for VVS2} \\ 7.8306 + 0.9812 + 1.8066\log(\text{carat}) & \text{for VVS1} \\ 7.8306 + 1.1409 + (1.8066 + 0.0966)\log(\text{carat}) & \text{for IF} \end{cases}$$

“clarity” as a Ordinal Predictor — First Attempt

Regarding `clarity` as an ordinal predictor with scores from 1 to 8 for the 8 levels of clarity.

```
diamonds$clarity3 = as.numeric(diamonds$clarity)
table(diamonds$clarity3,diamonds$clarity)
```

```
     I1 SI2 SI1 VS2 VS1 VVS2 VVS1  IF
  1  44   0   0   0   0    0    0   0
  2   0 513   0   0   0    0    0   0
  3   0   0 724   0   0    0    0   0
  4   0   0   0 689   0    0    0   0
  5   0   0   0   0 448    0    0   0
  6   0   0   0   0   0  267    0   0
  7   0   0   0   0   0    0  213   0
  8   0   0   0   0   0    0    0 102
```

Then we fit the model

```
lm2b = lm(log(price) ~ log(carat) + clarity3 +
          log(carat)*clarityI1 +log(carat)*clarityIF, data=diamonds)
anova(lm2b, lm2)
Analysis of Variance Table

Model 1: log(price) ~ log(carat) + clarity3 + log(carat) * clarityI1 +
    log(carat) * clarityIF
Model 2: log(price) ~ log(carat) * clarity
  Res.Df RSS Df Sum of Sq    F Pr(>F)
1   2993 106
2   2984 102  9      3.31 10.7 <2e-16
```

- Model `lm2b` says that the `log(price)` gap between adjacent levels of `clarity` are not all the same
- The gap in `log(price)` between adjacent levels of `clarity` are not all the same.

“clarity” as a Ordinal Predictor — Second Attempt

If we change the scoring of `clarity` to the following

clarity	I1	SI2	SI1	VS2	VS1	VVS2	VVS1	IF
score	0	4	5.5	7	7.5	9	9.5	11

```
diamonds$clarity4 = rep(0, length(diamonds$clarity))
diamonds$clarity4[diamonds$clarity == "SI2"] = 4
diamonds$clarity4[diamonds$clarity == "SI1"] = 5.5
diamonds$clarity4[diamonds$clarity == "VS2"] = 7
diamonds$clarity4[diamonds$clarity == "VS1"] = 7.5
diamonds$clarity4[diamonds$clarity == "VVS2"] = 9
diamonds$clarity4[diamonds$clarity == "VVS1"] = 9.5
diamonds$clarity4[diamonds$clarity == "IF"] = 11
```

```
lm2c = lm(log(price) ~ log(carat) + clarity4 +
          log(carat):clarityI1 +log(carat):clarityIF, data=diamonds)
anova(lm2c, lm2a)
Analysis of Variance Table

Model 1: log(price) ~ log(carat) + clarity4 + log(carat):clarityI1 + lo
Model 2: log(price) ~ log(carat) + clarity2 + log(carat):clarityI1 + lo
  Res.Df RSS Df Sum of Sq    F Pr(>F)
1   2995 103
2   2989 102  6      0.29 1.41   0.21
```

Conclusion

- Only the slopes of `log(carat)` for diamonds of the best and worst clarity are significantly different from the other clarity levels The slopes for diamonds at other `clarity` levels do not differ significantly
- Prices of diamonds increases as the clarity increases. The price gaps between some levels of `clarity` are closer like (`VS2` and `VS1`) and (`VVS2` and `VVS1`).