Statistical Modelling

Kamau Kamau

22/10/2019

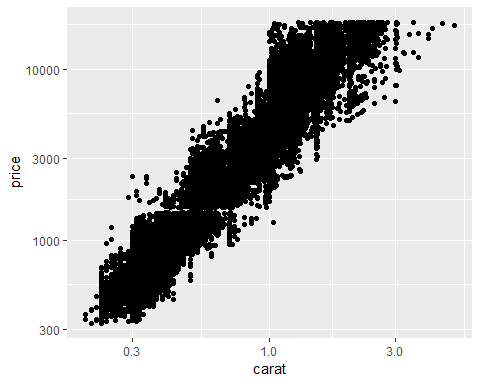
# Assumptions for the Regression Model

What are the assumptions that we have made using this model?

A straight line was the best way to model the overall relationship between log price and log carat. The noise terms εi all have the same variance. The noise terms εi are normally distributed. The error terms are independent. (i.i.d. stands for “independent and identically distributed”.) We have names for each of the assumptions:

* Linearity
* Constant variance (homoscedasticity)
* Normality
* Independence

## # A tibble: 6 x 10  
## carat cut color clarity depth table price x y z  
## <dbl> <ord> <ord> <ord> <dbl> <dbl> <int> <dbl> <dbl> <dbl>  
## 1 0.23 Ideal E SI2 61.5 55 326 3.95 3.98 2.43  
## 2 0.21 Premium E SI1 59.8 61 326 3.89 3.84 2.31  
## 3 0.23 Good E VS1 56.9 65 327 4.05 4.07 2.31  
## 4 0.290 Premium I VS2 62.4 58 334 4.2 4.23 2.63  
## 5 0.31 Good J SI2 63.3 58 335 4.34 4.35 2.75  
## 6 0.24 Very Good J VVS2 62.8 57 336 3.94 3.96 2.48



##   
## Call:  
## lm(formula = log(price) ~ log(carat), data = diamonds)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.50833 -0.16951 -0.00591 0.16637 1.33793   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.448661 0.001365 6190.9 <2e-16 \*\*\*  
## log(carat) 1.675817 0.001934 866.6 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2627 on 53938 degrees of freedom  
## Multiple R-squared: 0.933, Adjusted R-squared: 0.933   
## F-statistic: 7.51e+05 on 1 and 53938 DF, p-value: < 2.2e-16

