# Homework 02 Solutions

9/14/2020

## 1 (Sheather 2.8.6)

### 1.1 (a)

*Proof.*

$$\begin{aligned}(y_i - \hat{y}_i) &= (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) && \text{since } \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \\ &= (y_i - (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i)) && \text{since } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ &= (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})\end{aligned}$$

□

### 1.2 (b)

*Proof.*

$$\begin{aligned}(\hat{y}_i - \bar{y}) &= ((\hat{\beta}_0 + \hat{\beta}_1 x_i) - \bar{y}) \\ &= ((\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i) - \bar{y}) \\ &= \hat{\beta}_1 (x_i - \bar{x})\end{aligned}$$

□

### 1.3 (c)

*Proof.* Substituting the equalities from parts (a) and (b) above, we have:

$$\begin{aligned}\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n} [(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})] \hat{\beta}_1 (x_i - \bar{x}) \\ &= \hat{\beta}_1 \left[ \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) - \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})^2 \right] \\ &= \hat{\beta}_1 \left[ SXY - \frac{SXY}{SXX} SXX \right] \\ &= \hat{\beta}_1 [SXY - SXY] \\ &= 0\end{aligned}$$

□

# 2 (Gelman & Hill 2.6)

## 2.1 (a)

```
var1 <- rnorm(1000, 0, 1)
var2 <- rnorm(1000, 0, 1)
model <- lm(var2 ~ var1)
summary(model)
```

```
##
## Call:
## lm(formula = var2 ~ var1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.13418 -0.71644 -0.03454  0.72114  2.91643
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.01492    0.03250   0.459    0.646
## var1         0.05713    0.03228   1.770    0.077 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.027 on 998 degrees of freedom
## Multiple R-squared:  0.003129,   Adjusted R-squared:  0.00213
## F-statistic: 3.133 on 1 and 998 DF,  p-value: 0.07705
```

The slope is not significantly different from 0, as we can see from the p-value in the `var1` row, `Pr(>|t|)` column in the model summary.

## 2.2 (b)

```
library(arm)  # Needed for the se.coef() function.
z.scores <- rep(NA, 100)
for (k in 1:100) {
  var1 <- rnorm(1000, 0, 1)
  var2 <- rnorm(1000, 0, 1)
  fit <- lm(var2 ~ var1)
  z.scores[k] <- coef(fit)[2]/se.coef(fit)[2]
}

type1_errors <- sum(abs(z.scores) > 2)
print(type1_errors)
```

```
## [1] 5
```

There were 5 z-scores that were significant.

### 2.3 (c)

The rule of thumb to consider a z-score significant if the absolute value exceeds 2 is roughly equivalent to setting $\alpha = 0.05$, since 95% of the standard normal distribution falls within 1.96 standard deviations of the mean of 0. Therefore, we should expect the coefficients to be significant 5 out of 100 times on average, corresponding to a Type 1 error rate of 0.05.

## 3 (Sheather 2.8.7)

The regression line is defined as $\mu(x) = E(Y|X = x)$. In other words, it's the mean value of $Y$ at a given value of the covariates $X$. The confidence interval is for the regression line itself; that is, it's a confidence interval around an estimated mean. It's entirely possible for the data points themselves to fall outside the 95% confidence interval for a mean.

To gain some intuition, imagine that you have data from a normal distribution and you want to estimate the mean of that distribution. The more data you gather, the narrower the confidence interval around your estimate will be. The underlying distribution hasn't changed, however, so the more data you gather, the more of the data will fall outside the confidence interval for the mean.

## 4 (Sheather 3.4.8)

### 4.1 Part 1

#### 4.1.1 (a)

Given the problem constraints, the only flexibility we have is in specifying the intercept. We could (1) fix the intercept to some value, or (2) estimate the intercept via least squares.

Option (1) typically means setting the intercept to 0, which can be nice for interpretability in cases where the value (0, 0) is meaningful and where a different intercept would not make sense. For example, suppose we were asked to regress the number of words typed in a paper against the number of hours of work put in. Zero hours of work would necessarily produce zero words, so in that case there is a good reason to set the intercept to 0.

Here, however, the size of a diamond is presumably bounded away from 0. The notion of a zero-carat diamond doesn't make sense, and no one will manufacture a ring with a diamond that's so small as to be invisible to the naked eye. Indeed, the sizes of diamonds in the data range from 0.12 to 0.35 carats. If this is the range we're interested in modeling, then we may as well give the model the additional flexibility of a non-zero intercept, since the y-value at the intercept will not be meaningful regardless.

Hence, we take option (2) and estimate a model of the form

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Although the problem dicatates that we use a linear model, you should always visualize the data before specifying a model to help you decide whether that model is appropriate. Since we only have two variables, let's simply look at a scatter plot of Price against Size (Figure 1).

```
diamonds <- read.table("../diamonds.txt", header = TRUE)
plot(Price ~ Size, data = diamonds)
```

It looks like there's a more or less linear relationship, so let's fit the model.

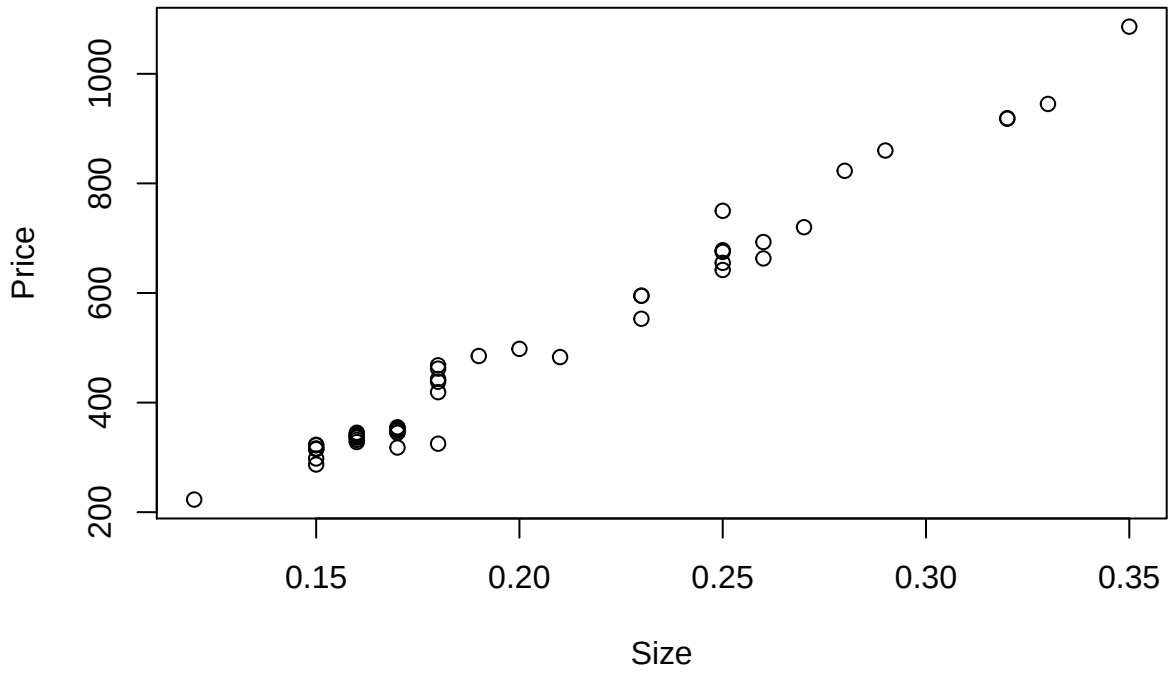


Figure 1: Scatter plot of diamond Price against Size

```
model5a <- lm(Price ~ Size, data = diamonds)
```

The coefficients of the model are

$\hat{\beta}_0$: -258.05

$\hat{\beta}_1$: 3715.02

Let's add the best fit line to the scatter plot (Figure 2).

```
plot(Price ~ Size, data = diamonds)
abline(a = coef(model5a)[1], b = coef(model5a)[2], col = "red", lty = "dashed")
```

The model looks reasonable. We'll examine diagnostic plots in Part 2 below.

#### 4.1.2 (b)

One weakness of the model is that by definition, it cannot capture nonlinearities in the relationship. Additionally, it has only one predictor; additional predictors could provide additional information. It also cannot deal with nonconstant variance.

### 4.2 Part 2

#### 4.2.1 (a)

In order to decide whether and how to modify the model above, let's examine some diagnostic plots. First, the residuals against the predictor (Figure 3).

```
plot(model5a$residuals ~ diamonds$Size, xlab = "Size", ylab = "Residuals")
```

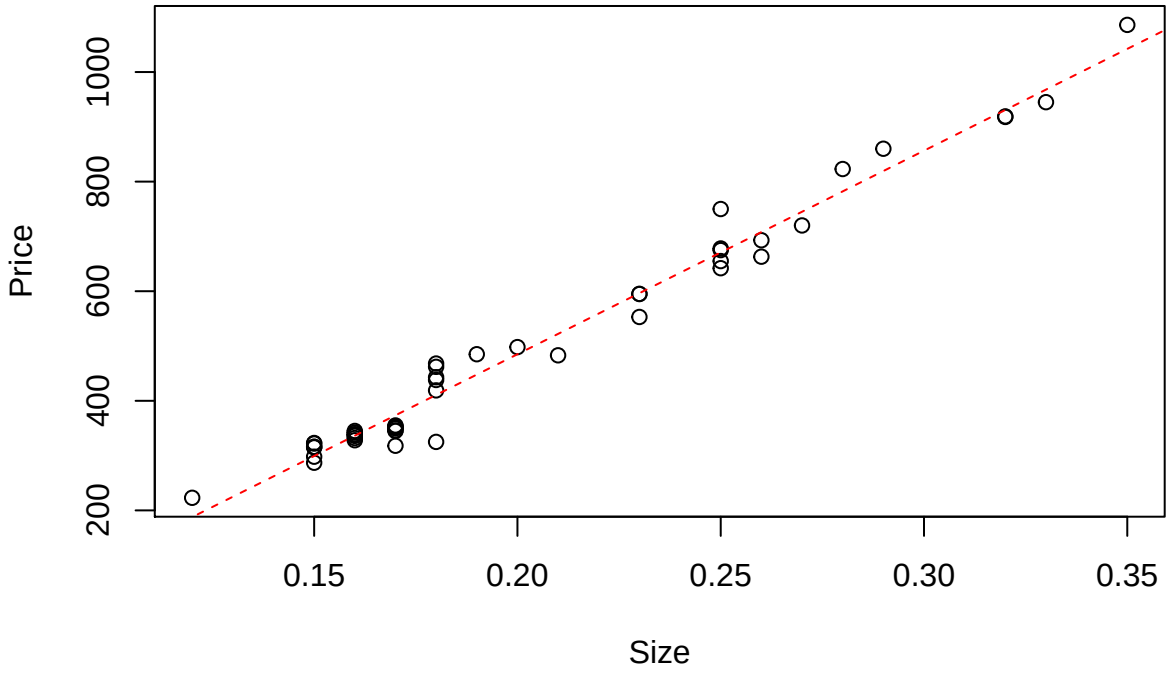


Figure 2: Scatter plot of diamond Price against Size with least squares line

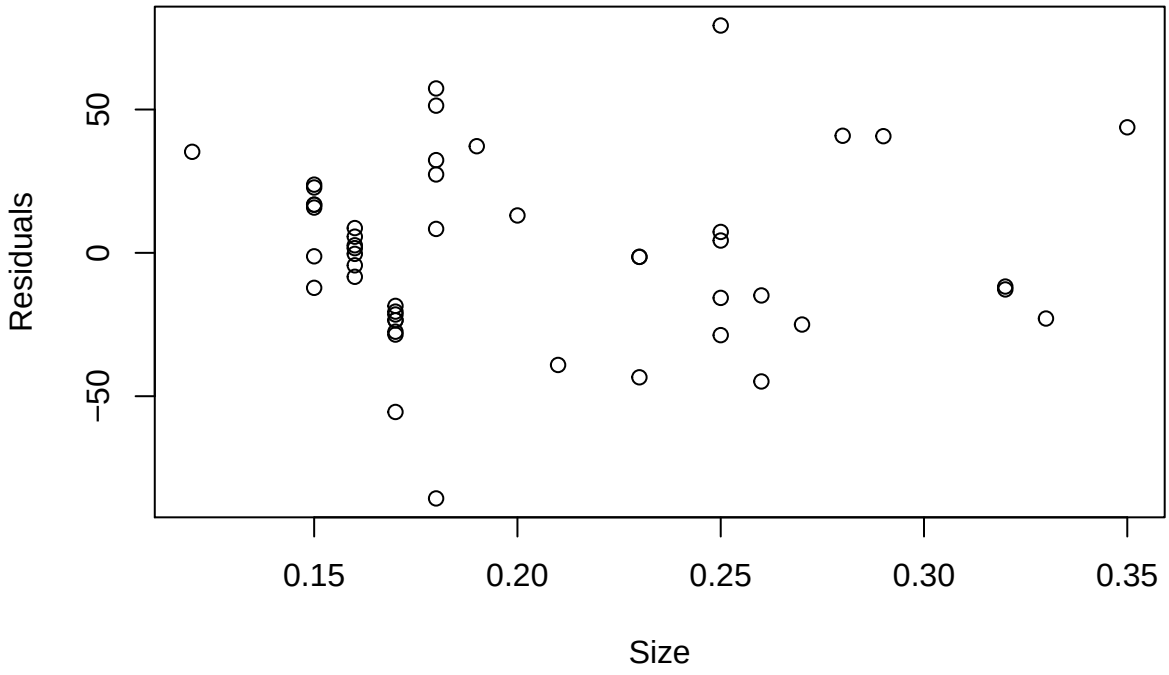


Figure 3: Model 5a residuals against predictor

There's no clear relationship (e.g. the residuals don't seem to be increasing or decreasing with Size), so that's good. Now let's examine the diagnostic plots that R produces for us automatically (Figure 4).

```
par(mfrow = c(2, 2), mar = c(4, 4.5, 2.5, 4)) # reducing the margins slightly
plot(model5a)                                  # generating diagnostic plots
```

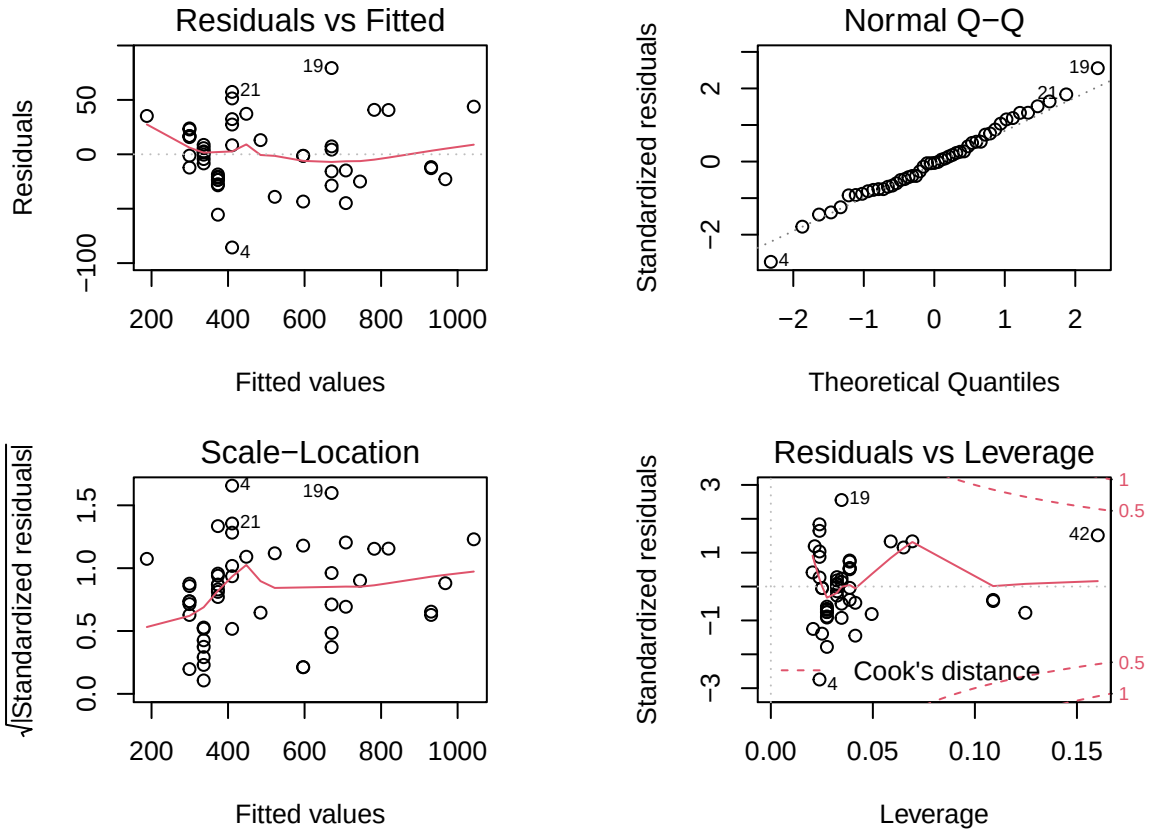


Figure 4: Diagnostic plots for model 5a

Let's discuss each of these in turn:

- The top left plot shows no relationship between the residuals and the fitted values (the $\hat{y}_i$s), so that's good.
- The top right plot shows that the residuals are approximately normally distributed, which is also good.
- The nonparametric red line in the bottom left plot shows a slight increasing trend between the fitted values and the standardized residuals, but there aren't many data points on the right side of the plot, so I wouldn't take this too seriously. (Recall that the variance of the residuals $\hat{e}_i$ is nonconstant, *even if the variance of the underlying errors $e_i$ is constant.* The standardized residuals have approximately constant variance, so it is useful to examine these instead of the raw residuals, particularly when points of high leverage exist.)
- The bottom right plot shows that there are a couple outliers in the sense defined in Sheather: points whose standardized residual falls outside the range [-2, 2]. R has helpfully labeled these points with their row numbers in the data, so we can examine them. Additionally, recall the rule of thumb that a point is "high leverage," or equivalently a "leverage point," if its leverage value $h_{ii}$ is greater than $4/n$, where $n$ is the number of data points. In this case, $4/n = 4/49 \approx 0.082$, so we have a couple of leverage points in the data. One of these, point 42, has a relatively high Cook's distance; it lies close to the contour line that represents a Cook's distance of 0.5.

Since R has helpfully labeled points 4, 19, and 42 for us in the bottom right diagnostic plot, let's examine these on the original graph (Figure 5).

```r
plot(Price ~ Size, data = diamonds)
abline(a = coef(model5a)[1], b = coef(model5a)[2], col = "red", lty = "dashed")
points(diamonds[c(4, 19, 42), ], col = "red")
```

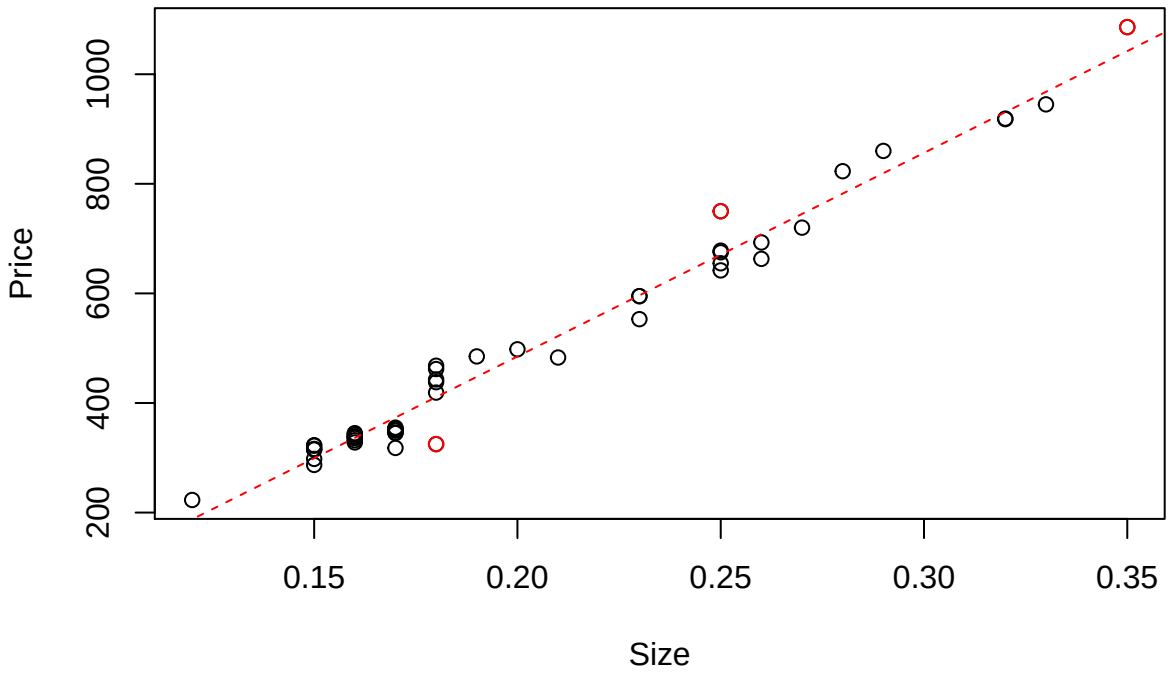


Figure 5: Scatter plot of diamond Price against Size with least squares line and points 4, 19, 42 highlighted

None of these points is obviously problematic. Let's check one more diagnostic, the inverse response plot:

```r
# install.packages("alr3")  # if you don't already have it installed
library(alr3)
invisible(inverseResponsePlot(model5a))
```

Based on this plot, it looks like the original model, represented by the light blue line ($\lambda = 1$) fits the data well. It doesn't look like we would do better transforming the response variable, at least not with a scaled power transformation.

Given these results, it seems reasonable to keep the original model in 5(a)

#### 4.2.2 (b)

Since the model hasn't changed, see 5(a).

### 4.3 Part 3

The model did not change.

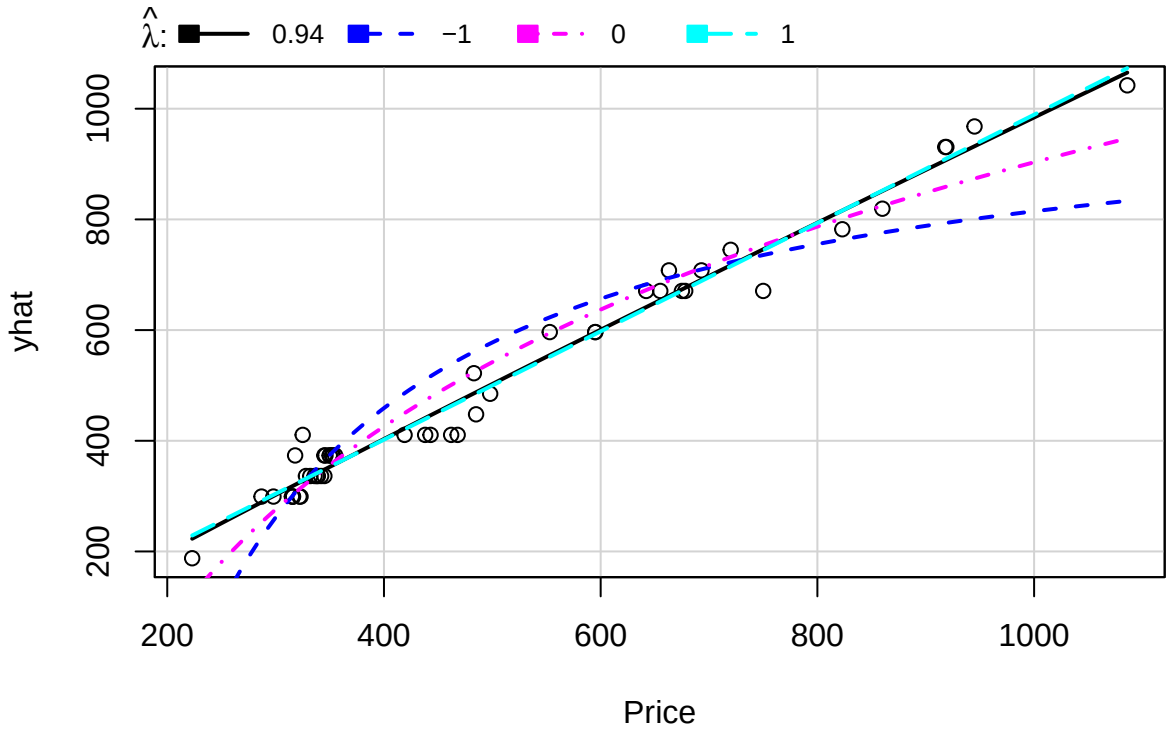


Figure 6: Inverse response plot for the diamonds model5a

# 5 5. IMRAD and IDMRAD analysis

## 5.1 (a)

### 5.1.1 Abstract

**Yes,** *this section has the right content.* The section nicely summarizes the later sections (introduction, method, results and discussion) and provides a clear overview.

### 5.1.2 Introduction

**Yes,** *this section has the right content.* This section does well at motivating why we might be interested in examining the relationship between a critic's rating and the wine's price. It also clearly highlights the 3 questions that the paper will attempt to answer.

This introduction could be improved by more clearly highlighting the connection between the second and third questions to the first question (which is directly related to the motivation in the first paragraph)[1].

### 5.1.3 Methods

**No,** *this section's content does not meet expectations.* This section doesn't highlight (in writing) a lot of the methods that were used to to answer the questions posed. Specifically, all 3 analytics tools are is only mentioned in passing in the final sentence of the section. No motivation is provided to suggest why such

[1]This comment could suggest that our answer should be actually "No, the introduction section does not contain the expected content".

analysis was done, which should have at least occured for Figure 2's log-transformation of the price and scores.

In terms of the organization advice given on slide 9 of lecture 02 (week01), the method section failed to use a "context, content, and conclusion" structure in presenting what methods were used. This is especially true for this section as the figures are presented first without any discussion of what was done.

As most of this section presents data information, it might be useful for the author to switch to an IDMRAD approach and put that information into a Data section. In the commentary on the **results** section of the paper in section 5.1.4, I will discuss parts of results that really should be in the methods section in more details.

#### 5.1.4 Results

**No,** *this section's content does not meet expectations.* As alluded to above, the results section contains several pieces that really should be in the **methods** section. Specifically, the descriptions of the methods / figures used should be in the method section. Additionally, the discussion of the motivation of using a log transformation should have introduced Figure 2 in the methods section.

Additionally, this section presents the idea that the slope of regression line (relating the critic's scores to the price point) is a valuable measure to compare the critic's impact even though the scales are different. This is hard to justify and muddies the water in terms of making sure the reader agrees with the conclusions (#7 of slide 9). Another, slightly nit-picky critique of this section, would be that the introduction failed to well motivate the later 2 questions and the author spent a lot of time discussion the results related to the questions without well motivating the results. This might lose a reader and relates to #9 of slide 9 - allocation the time of the reader wisely.

This section's commentary on the spread of the prices relative to the critic's score does attempt to well address the question relative to the methods used, and if I was an editor / advisor I would encourage that to be expanded on.

### 5.2 Discussion

**Yes,** *this section has the right content.* Conditional on the comments above, the **discussion** section does highlight overarching conclusions, and this section spends a lot of time discussion limitations and future improvements. The author did convert potential criticisms of the work into limitations and future work ideas.

If anything, this section spends the vast majority of presenting limitations and could have been better structured to capture that idea.

## 5.3 (b)

*A rewrite of the **Data** and **Methods** section. Stylistically, I tried to stay as close to the current write-up as possible.*

**Data**

The data for this study come from Parker [2] and Coates [1]. The prices (in pounds sterling) include duty but exclude shipping and VAT in London in September 2003 [3, pp. 8]. Parker's rating system is 100-point based and the scale can be seen in Figure 7. Whereas Coates' rating system is 20-point based and can be seen in Figure 8. The dataset contains prices for 72 wines from the 2000 vintage in Bordeaux, and all the variables in Figure 9, and was obtained from supplementary material provided by Sheather [3].

| | |
|---|---|
| 96-100 points | Extraordinary |
| 90-95 points | Outstanding |
| 80-89 points | Above average to very good |
| 70-79 points | Average |
| 50-69 points | Below average to poor |

Figure 7: Parker's rating system for wine

| | | | |
|---|---|---|---|
| 20 | Excellent. 'Grand vin' | 16 | Very good |
| 19.5 | Very fine indeed | 15.5 | Good plus |
| 19 | Very fine plus | 15 | Good |
| 18.5 | Very fine | 14.5 | Quite good plus |
| 18 | Fine plus | 14 | Quite good |
| 17.5 | Fine | 13.5 | Not bad plus |
| 17 | Very good indeed | 13 | Not bad |
| 16.5 | Very good plus | 12.5 | Poor |

Figure 8: Coates' rating system for wine

$$
\begin{aligned}
Y &= \text{Price} = \text{the price (in pounds sterling) of 12 bottles of wine} \\
x_1 &= \text{ParkerPoints} = \text{Robert Parker's rating of the wine (out of 100)} \\
x_2 &= \text{CoatesPoints} = \text{Clive Coates' rating of the wine (out of 20)} \\
x_3 &= \text{P95andAbove} = 1\ (0) \text{ if the Parker score is 95 or above (otherwise)} \\
x_4 &= \text{FirstGrowth} = 1\ (0) \text{ if the wine is a First Growth (otherwise)} \\
x_5 &= \text{CultWine} = 1\ (0) \text{ if the wine is a cult wine (otherwise)} \\
x_6 &= \text{Pomerol} = 1\ (0) \text{ if the wine is from Pomerol (otherwise)} \\
x_7 &= \text{VintageSuperstar} = 1\ (0) \text{ if the wine is a superstar (otherwise)}
\end{aligned}
$$

Figure 9: Description of dataset's variables

**Methods**

In order to understand the relationship between the prices of wine and critics scores we creates a set of scatter plots relating these two variables together. Figure 10 shows these relationships and the relationship between Parker and Coates' critic scores of the same wines. From 10's plots relating both critic's scores

to the price of wine, we saw that the data points demonstrate an exponential trend, as such, we created a log-log transformation of plots in Figure 10 as can be seen in Figure 11.

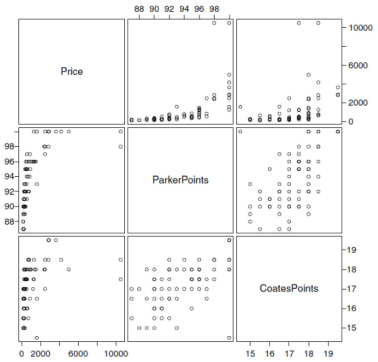


Figure 1: Scatterplot matrix of Price, ParkerPoints, & CoatesPoints. From Sheather (2009, p. 10).

Figure 10: Pairwise scatterplots of critic scores and prices of wines.

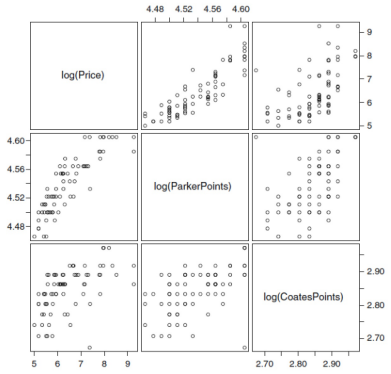


Figure 2: Scatterplot matrix of log(Price), log(ParkerPoints), & log(CoatesPoints). From Sheather (2009, p. 12).

Figure 11: Pairwise scatterplots of critic scores and prices of wines (log-log transformed).

Beyond critic's scores, we explored the relationship between the price of wine and some binary variables that might also effect the wine's price in Figure 12 by comparing the distribution of prices conditional on these binary variables values.

# References


[1] Clive Coates. *The Wines of Bordeaux*. 01 Edition. London: Weidenfeld & Nicolson, 2004.

[2] Robert Parker. "Robert Parker Wine Advocate". In: *The Wine Advocate, Inc* (2003).

[3] S.J. Sheather. *A Modern Approach to Regression with R*. New York: Springer Science + Business Media LLC, 2009.

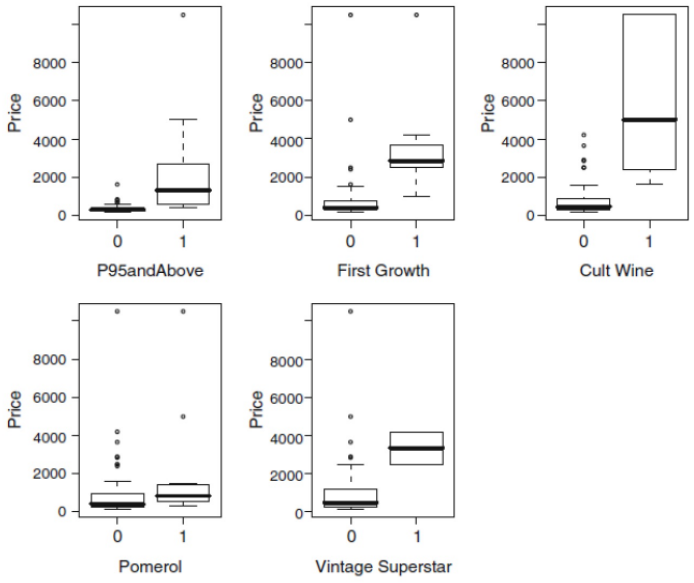


Figure 3: Box plots of Price against each of the dummy variables. From Sheather (2009, p. 11).

Figure 12: Boxplots comparing the price of wine conditional on additional features about the wine (see Figure 9 for more details).