# 36-617: Applied Linear Models
# Fall 2021
# HW02 – Solutions

## Exercises

1. *Sheather, Ch 3, pp 109ff, #5*

   *An analyst for the auto industry has asked for your help in modeling data on the prices of new cars. Interest centers on modeling suggested retail price as a function of the cost to the dealer for 234 new cars. The data set, which is available on the book website in the file* `cars04.csv`*, is a subset of the data from* `http://www.amstat.org/publications/jse/datasets/04cars.txt` *(Accessed March 12, 2007)*

   *The first model fit to the data was*

   $$\textit{Suggested Retail Price} = \beta_0 + \beta_1 \ \textit{Dealer Cost} + \varepsilon \tag{3.10}$$

   *On the following pages is some output from fitting model (3.10) as well as some plots (Figures 1 and 2 below reproduce the figures in the book, for easy reference).*

   (a) *Based on the output for model (3.10) the analyst concluded the following: Since the model explains just more than 99.8% of the variability in Suggested Retail Price and the coefficient of Dealer Cost has a t-value greater than 412, model (1) [the model in (3.10)] is a highly effective model for producing prediction intervals for Suggested Retail Price.*

   *Provide a detailed critique of this conclusion.*

   Although the regression output in Figure 1 looks good, there are several volations of the modeling assumptions revealed in the plots in Figure 2:

   - SuggestedRetailPrice, DealerCost and the residuals, are all skewed right
   - The residuals have nonconstant variance
   - The aggregation of the data around different lines in the plot suggests that some predictor variable(s) may be missing from the model

   Violations of the modeling assumptions undermine the validity of inferences we can make from Figure 1.

   (b) *Carefully describe all the shortcomings evident in model (3.10). For each shortcoming, describe the steps needed to overcome the shortcoming.*

   Here are four possible shortcomings *(you may have found others! Name any two legitimate criticisms, and ways to fix them, for full credit)*:

   - The "SuggestedRetailPrice vs DealerCost", "Standardized Residuals vs DealerCost", and "Sqrt(|Standardized Residuals|) vs DealerCost" plots all show that Dealer Cost is substantially right-skewed. Right skewing tends to create high-leverage points in the data. A transformation of DealerCost to reduce the skewing would help: usually a fractional power, or a logarithm are good fixes for this.
   - The Normal QQ Plot shows that the residuals are also right-skewed. The same sort of transformation (log or fractional power) of SuggestedRetailPrice will help to reduce this skewing.

```
Call:
lm(formula = SuggestedRetailPrice ~ DealerCost, data = cars04)

Residuals:
     Min       1Q   Median       3Q      Max
-1743.52  -262.59    74.92   265.98  2912.72

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -61.904248  81.801381  -0.757     0.45
DealerCost    1.088841   0.002638 412.768   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 587 on 232 degrees of freedom
Multiple R-squared:  0.9986,        Adjusted R-squared:  0.9986
F-statistic: 1.704e+05 on 1 and 232 DF,  p-value: < 2.2e-16
```

Figure 1: Regression output for model (3.10).

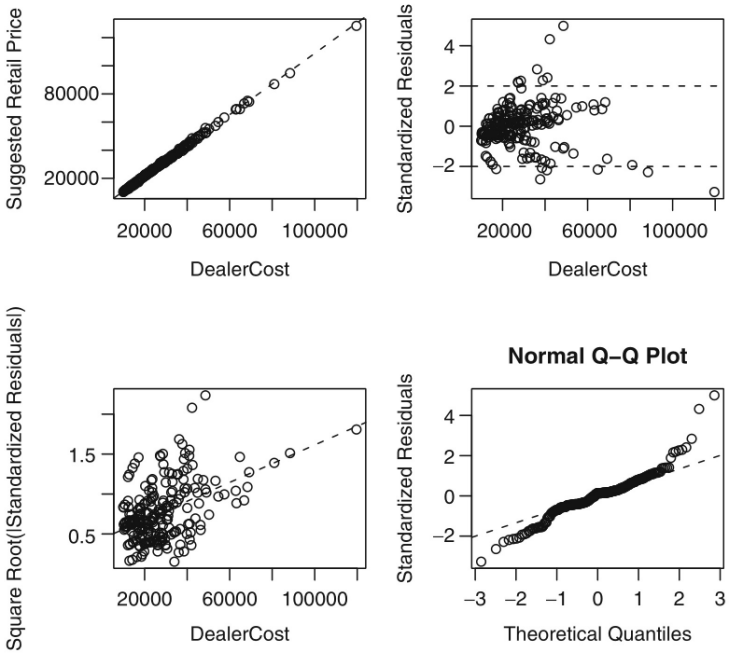


Figure 2: Some plots for model (3.10).

- Both the Standardized Residuals plot and the "Sqrt(|Standardized Residuals|) vs DealerCost" plot show that the residuals have non-constant variance. You could suggest a variance-stabilizing transformation, or a log or power transformation, to help fix this problem.
- The Standardized Residual plot shows the data clustering along several different lines, suggesting that perhaps different vehicle types or brands have different relationships between SuggestedRetailPrice and DealerCost. One could explore this idea by considering an ANCOVA model, which we will talk about in later lectures.

*The second model fitted to the data was*

$$\log(\textit{Suggested Retail Price}) = \beta_0 + \beta_1 \ \log(\textit{Dealer Cost}) + \varepsilon \tag{3.11}$$

*Output from model (3.11) and plots are shown below (Figures 3 and 4 below reproduce the figures in the book, for easy reference.)*

(c) *Is model (3.11) an improvement over model (3.10) in terms of predicting Suggested Retail Price? If so, please describe all the ways in which it is an improvement.*

Model (3.11) is definitely an improvement over model (3.10). The regression output for both models is about the same, so the differences are in the plot in Figures 2 and 4. Here are some ways in which the plots are better for (3.11) than for (3.10) *(You may have found other reasons. Name any two legitimate reasons for full credit).*

- Both DealerCost and SuggestedRetailPrice exhibit less right-skewing in model (3.11) than in model (3.10).
- While there are still outliers in the residual vs fitted plot for model (3.11), they are less extreme than for model (3.10).
- In addition to less right-skew, the normal QQ plot suggests more nearly-normal residuals for model (3.11) than for model (3.10). We have pushed some values out into the left tail, but there are fewer of these in the QQ plot for model (3.11) than there are for model (3.10).

(d) *Interpret the estimated coefficient of log(Dealer Cost) in model (3.11).*

As we know from the text, or from the handout "log xform and percent interpretation.pdf" in the week03 folder in the files area, $\beta_1$ is the expected percent change in $y$ for a 1% change in $x$. Since $\hat{\beta}_1 = 1.015$ in Figure 3, there is about a 1% change in SuggestedRetailPrice for every 1% change in DealerCost, according to the fitted model.

(e) *List any weaknesses apparent in model (3.11).*

Here are some weaknesses we can see in Figure 4 (you may have discovered others; list any two legitimate weaknesses to get full credit):

- From the QQ plot, both tails of the residual distribution are a bit long. Although the deviation is more impressive in the lower tail, there are more data points in the upper tail.
- The scale-location plot suggests that the variance of residuals may increase as DealerCost increases.
- Although it is not as evident in Figure 4 as it is in Figure 2, it still looks like the data aggregates along definable curves, which suggests that perhaps different car types have different relationships between DealerCost and SuggestedRetailPrice.

(Aside: If we were to successfully model different DealerCost vs. SuggestedRetailPrice relationships for different car types, that might remove the other two problems above as well.)

```
Call:
lm(formula = log(SuggestedRetailPrice) ~ log(DealerCost), data = cars04)

Residuals:
      Min        1Q    Median        3Q       Max
-0.062920 -0.008694  0.000624  0.010621  0.048798

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.069459   0.026459  -2.625  0.00924 **
log(DealerCost)  1.014836   0.002616 387.942  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01865 on 232 degrees of freedom
Multiple R-squared:  0.9985,        Adjusted R-squared:  0.9985
F-statistic: 1.505e+05 on 1 and 232 DF,  p-value: < 2.2e-16
```

Figure 3: Regression output for model (3.11).

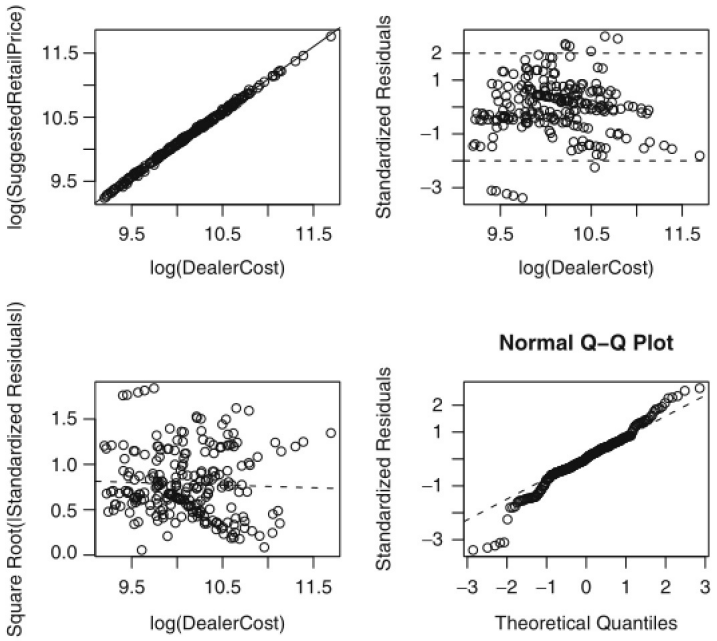


Figure 4: Some plots for model (3.11).

2. *In the folder for this hw assignment you will find a pdf called "COVID breakthrough rates in England". This is a recent article from the medical journal* The Lancet.

   (a) *Write a one-paragraph abstract with exactly four sentences, one for each section of the paper: Introduction, Methods, Results and Discussion. Each sentence should highlight the main point of each section, and together the four sentences should tell the story of the paper. The last sentence should include the main result of the paper.*

   (For most of this abstract I simply copied or merged sentences from the structured abstract on p. 1 of the article [you could of course write your own sentences summarizing each part of the article]. I thought it would be helpful if I added a footnote for the place I got each sentence [you do not have to do this for your answer].)

   This study aimed to identify risk factors for post-vaccination SARS-CoV-2 infection and describe the characteristics of post-vaccination illness[1]. We used univariate logistic regression models (adjusted for age, BMI, and sex) to analyse the associations between risk factors and post-vaccination infection, and the associations of individual symptoms, overall disease duration, and disease severity with vaccination status, in self-report data from UK-based, adult (≥18 years) users of the COVID Symptom Study mobile phone app[2]. Vaccination (compared with no vaccination) was associated with reduced odds of hospitalisation or having more than five symptoms in the first week of illness following the first or second dose, and long-duration (≥28 days) symptoms following the second dose; following first dose only, older adults, individuals living in deprived areas, and obese individuals all experienced higher risk of breakthrough infection[3]. Our findings might support caution around relaxing physical distancing and other personal protective measures in the post-vaccination era, particularly around frail older adults and individuals living in more deprived areas, even if these individuals are vaccinated, and might have implications for strategies such as booster vaccinations[4].

   (b) *Now, imagine that this is an IDMRAD paper, with the data section containing the material before the "Statistical analysis" subhead on p. 4 (as described above). Write a one-paragraph abstract with exactly five sentences, one for each section of the paper: Introduction, Data, Methods, Results and Discussion. Each sentence should highlight the main point of each section, and together the five sentences should tell the story of the paper. The last sentence should include the main result of the paper.*

   (For most of this abstract I simply copied or merged sentences from the structured abstract on p. 1 of the article [you could of course write your own sentences summarizing each part of the article]. I thought it would be helpful if I added a footnote for the place I got each sentence [you do not have to do this for your answer].)

   This study aimed to identify risk factors for post-vaccination SARS-CoV-2 infection and describe the characteristics of post-vaccination illness[5]. This prospective, community-based, nested, case-control study used self-reported data (eg, on demographics, geographical location, health risk factors, and COVID-19 test results, symptoms, and vaccinations) from UK-based, adult (≥18 years) users of the COVID Symptom Study mobile phone app[6]. We used univariate logistic regression models (adjusted for age, BMI, and sex) to analyse the associations between risk factors

[1] Last sentence of summary paragraph, p. 1
[2] Combining first and last sentences of Methods paragraph, p. 1
[3] Second-to-last sentence of Findings paragraph, combined with a summary of the risk factor analysis.
[4] Last sentence of Interpretation paragraph, p. 1
[5] Last sentence of summary paragraph, p. 1
[6] First sentence of Methods paragraph, p. 1

and post-vaccination infection, and the associations of individual symptoms, overall disease duration, and disease severity with vaccination status[7]. Vaccination (compared with no vaccination) was associated with reduced odds of hospitalisation or having more than five symptoms in the first week of illness following the first or second dose, and long-duration (≥28 days) symptoms following the second dose; following first dose only, older adults, individuals living in deprived areas, and obese individuals all experienced higher risk of breakthrough infection[8]. Our findings might support caution around relaxing physical distancing and other personal protective measures in the post-vaccination era, particularly around frail older adults and individuals living in more deprived areas, even if these individuals are vaccinated, and might have implications for strategies such as booster vaccinations[9].

3. *In the folder for this hw assignment you will find a pdf called "An IMRAD paper on wine", based on Example 1.2.4 in Sheather. This paper is based only on EDA, not on any more sophisticated methods.*

   (a) *Does the paper appropriately address each of the parts of an IMRAD paper as described on slide 3, lecture 04 from week02 of class? If you need more detail on the sections of an IMRAD paper, see* `http://www.jpgmonline.com/documents/author/24/2_Aggarwal_10.pdf`.

   *For each section below, either say "yes this section has the right content", or say "no" and describe what is missing and/or what needs to be moved to another section of the paper or deleted.*

   - *Abstract*
     **Yes, this section has the right content.** The first sentence summarizes I; the second and third sentences summarize M; Sentences 4-7 summarize the results; and sentence 8 summarizes the discussion. (It is a little awkward because the answer to the main question of the paper, that Parker's ratings have a bigger impact than Coates', comes in the middle of the abstract instead of toward the end where the summary of the discussion is, but otherwise it seems fine.)
   - *Introduction*
     **Yes**, the introduction addresses "why read the paper" as well as what the main questions to be addressed are.
   - *Methods*
     **Yes**, the text of the methods section is fine: It says where the data came from, and briefly describes the variables, and the data analysis methods (all EDA). It would probably be better if the graphs themselves appeared in the Results section.
   - *Results*
     **Yes this section has the right content.** There is one paragraph for each research question announced in the Introduction (in a paper with more involved analyses, there might be one subsection for each research question, rather than just one paragraph). The paragraphs summarize the analyses and give appropriate conclusions (in a paper with more involved analyses, e.g. regression analysis etc., there might also be tables listing coefficient estimates & standard errors, graphs showing predicted values, etc.).
   - *Discussion*
     **Not quite the right content.** The first paragraph should summarize the main results of the paper. The first paragraph here is a mixture of new results (should be in the results section!) and study limitations (should come later in the discussion!). The first sentence of the last

[7] Last sentence of Methods paragraph, p. 1
[8] Second-to-last sentence of Findings paragraph, combined with a summary of the risk factor analysis.
[9] Last sentence of Interpretation paragraph, p. 1

paragraph would be better as the first sentence of the Discussion. The remaining text describes further limitations and other considerations, which is fine.

(b) *Please write the appropriate* **Data** *and* **Methods** *sections for this paper (just those two sections), as an IDMRAD paper instead of an IMRAD paper (include appropriate text, figures and tables in the two sections). For your convenience the three figures and three tables of this paper are saved as jpg files in the folder for this assignment.*

A suitable Data and Methods section appear on the next page. I have just taken sentences as written in the IMRAD paper and moved them to the appropriate Data and Methods sections here.

## Data

The data for this study come from Parker (2003) and Coates (2004). The prices (in pounds sterling) include duty but exclude shipping and VAT in London in September 2003 (Sheather, 2009, pp. 8). Parker's rating system is 100-point based and the scale is as follows:

| | |
|---|---|
| 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 |

Whereas Coates' rating system is 20-point based and is scaled as follows:

| | | | |
|---|---|---|---|
| 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 |

The dataset contains prices for 72 wines from the 2000 vintage in Bordeaux, and all the variables are listed as follows:

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

The data are available in the file Bordeaux.csv, in the online supplement accompanying Sheather (2009).

(Given that the EDA plots are the only "analysis" here, they probably should go in the results section. In a more elaborate analysis, some EDA could go in the data section, and the results of other EDA and/or more sophisticated statistical methods would go in the Results section. [It is fine, however, if you put the EDA in the Data section for this hw question.])

## Methods

To analyze effects of ratings and other predictor variables on wine prices, we used scatterplot matrix and box plots provided by Sheather (2009, pp. 10-13), which were made possible by the R language (R Core Team, 2017).

(The methods section is very short, because there are basically no methods in this paper!)