

ST 625: Exam 2

Steven Truong

Bentley Honor Code

As a Bentley student, I promise to act honorably in my courses and my professional endeavors, adhering to both the letter and spirit of Bentley's academic integrity system. I will neither take advantage of my classmates nor betray the trust of my professors. My work will be honest and transparent, and I will hold myself and my peers accountable to the highest ethical standards.

Instructions

The **Bentley Honor Code** always applies. In particular, you are expected to keep the content of this exam confidential. You cannot discuss the particulars of this exam or the types of questions that it contains with another student until you receive your graded exam back or solutions to the exam are posted by your instructor.

1. This is a take-home exam and it is **due on Wednesday, April 28th at 7:00 PM Eastern Day Saving Time**. No late exams will be accepted.
2. You have 48 hours to complete the exam.
3. Write your name at the top of this file under the **author**.
4. You are only allowed to consult your book, your notes, and class materials.
5. You are **not allowed** to discuss your work with another student. If you have any questions regarding problems, please post them on Discussion Board. Make sure you ask with plenty of time before the deadline.
6. You are **forbidden** to use a service (human or machine, paid or unpaid) that will help you solve a problem. **The solutions you provide must be your own.**
7. Write enough details so I can follow your thought process and you can double-check your work. **Unsubstantiated answers will receive zero points.** If I cannot follow your solution strategy, you will lose points. The final solutions/conclusion should be presented in a complete sentence.
8. **Please rename the final file "SN2-Exam2-FirstName-LastName.pdf" with your name and submit PDF files to Blackboard (go to Blackboard -> Assignment -> Quizzes and Exams).**
9. Before working on the file, please restart your Rstudio to make sure there are no other variables that you used before in the Environment and then compile the file to PDF to check any potential issues. My suggestion is to knit the file as often as you can while working on the R markdown.
10. **Avoid directly copying any symbols from other PDF files because Unicode characters are not allowed in R markdown.**
11. Before submitting your answers, make sure the final PDF file can be successfully open in the Adobe Acrobat Reader and read through the PDF file to ensure that your solutions can be found easily.

There are 3 problems for a total of 30 points.

```
# install.packages("alr4") ## install alr4 package if it is the first time use it.
library(alr4) ## load package
```

```
## Warning: package 'alr4' was built under R version 4.0.4
```

```
## Warning: package 'car' was built under R version 4.0.3
```

```
## Warning: package 'carData' was built under R version 4.0.3
```

```
## Warning: package 'effects' was built under R version 4.0.4
```

Problem 1: [10 points]

The data file *salary* (in the *alr4* library) concerns salary and other characteristics of all faculty in a small Midwestern college collected in the early 1980s for presentation in legal proceedings for which discrimination against women in salary was at issue. All persons in the data hold tenured or tenure track positions; temporary faculty are not included. The variables include *degree*, a factor with levels PhD and MS; *rank*, a factor with levels Asst, Assoc, and Prof; *sex*, a factor with levels Male and Female (Male is the reference level); *year*, years in current rank; *ysdeg*, years since highest degree, and *salary*, academic year salary in dollars.

```
data(salary) ## import the data: "salary" which stored in alr4
```

- (2 points) Fit a simple linear regression: regress *salary* on *sex* and write down the estimated equation. Interpret the estimated coefficient of *sex*.
- (2 points) Fit a multiple regression: regress *salary* on *sex* and *year*. Interpret again the estimated coefficient of *sex*. Do you have the same interpretation as in (a) and can you explain the difference if there is any?
- (2 points) Now regress *salary* on *sex*, *year*, *degree*, *rank*, and *ysdeg*. Based on this model, report the test statistic, p-value, decision, and the final conclusion for testing the hypothesis that the mean salary for men and women is the same.
- (2 points) Using the regression in (c), obtain and interpret a 99% confidence interval for the difference in salary between males and females.
- (2 points) Regress *salary* on *sex*, *year* and their interactions. Provide an interpretation for the estimated coefficient of the interaction term.

Problem 2: [10 points]

The data file *Highway* (in the *alr4* library) contains the automobile accident rate in 39 sections of large highways in the state of Minnesota in 1973. The variables are described below.

1. *rate*: Accident rate per million vehicle miles
2. *len*: Length of the segment in miles
3. *adt*: Average daily traffic count in thousands
4. *trks*: Truck volume as a percentage of the total volume
5. *slim*: Speed limit
6. *shld*: Shoulder width in feet of outer shoulder on the roadway
7. *sigs*: Number of signalized interchanges per mile in the segment

```
data("Highway")
```

The variable *sigs* is 0 for freeway-type road segments but can be well over 2 for other segments. We can't use logarithms because of the 0 values. The variable *sigs* is a rate per mile, so we add the constant to the number of signals in the segment, and then recompute a rate.

```
Highway$sigs1 <- with(Highway, (sigs * len + 1)/len)
```

We consider the logarithms of *len*, *adt*, *trks*, and *sigs1*. Now, we have eight predictors after transformation: $\log(\text{len})$, *shld*, $\log(\text{adt})$, $\log(\text{trks})$, *lane*, *slim*, *lwid*, *itg*, $\log(\text{sigs1})$, *acpt*, and *htype*. Because there are too many potential predictors, we would like to conduct variable selection first before using the model.

- (a) (2 points) With the Highway data, start with the full model

$$\log(\text{rate}) = \log(\text{len}) + \text{shld} + \log(\text{adt}) + \log(\text{trks}) + \text{lane} + \text{slim} + \text{lwid} + \text{itg} + \log(\text{sigs1}) + \text{acpt} + \text{htype} + \epsilon$$

and do a backward elimination using BIC as the criterion. Write down the final estimated model.

- (b) (2 points) With the Highway data, start with a model without any predictor and do a forward addition using BIC as the criterion. Write down the final estimated model.
- (c) (2 points) Based on models from part (a) and (b), compare their adjusted R² and MSE, which model is better?
- (d) (2 points) Using the model you think is better based on part (c), make the residual vs fitted values plot. Do you see any problems?
- (e) (2 points) Using the same model as part (d), plot the quantile-quantile plot to check normality. Do you see any problems?

Problem 3 [10 points]

Breakdowns of machines that produce steel cans are very costly. The more breakdowns, the fewer cans produced, and the smaller the company's profits. To help anticipate profit loss, the owners of a can company would like to find a model that will predict the number of breakdowns on the assembly line. The model proposed by the company's statisticians is the following:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

where *y* is the number of breakdowns per 8-hour shift, $x_1 = \begin{cases} 1, \text{afternoon shift} \\ 0, \text{otherwise} \end{cases}$, $x_2 = \begin{cases} 1, \text{midnight shift} \\ 0, \text{otherwise} \end{cases}$,

*x*₃ is the temperature of the plant (°F), and *x*₄ is the number of inexperienced personnel working on the assembly line. After the model is fit using the least squares procedure, the residuals are plotted against \hat{y} , as shown in the accompanying figure.

- (a) (2 points) Do you detect a pattern in the residual plot? What does this suggest about the least squares assumptions?
- (b) (2 points) Given the nature of the response variable *y* and the pattern detected in part a, what model adjustments would you recommend.

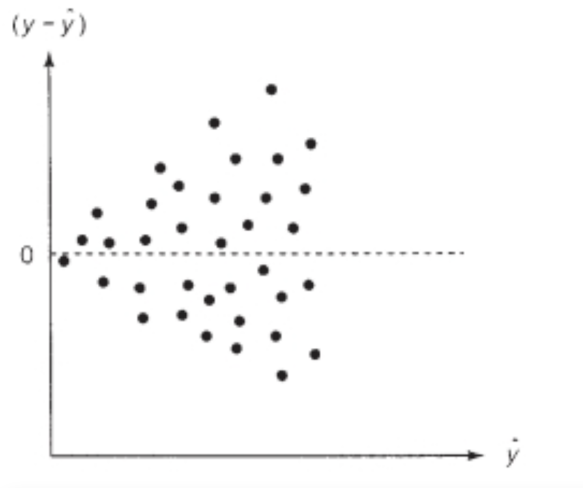


Figure 1: Residual Plot.

The regression analysis for the transformed model (where y^* is the transformed variable)

$$y^* = \sqrt{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

produces the prediction equation

$$\hat{y}^* = 1.3 + 0.008x_1 - 0.13x_2 + 0.0025x_3 + 0.26x_4$$

- (c) (2 points) Use the equation above to interpret the estimated coefficient of x_1 and x_2 .
- (d) (2 points) Use the equation to predict the number of breakdowns during the midnight shift if the temperature of the plant at that time is 87°F and if there is only one inexperienced worker on the assembly line.
- (e) (2 points) A 95% prediction interval for y^* when $x_1 = 0$, $x_2 = 0$, $x_3 = 90^\circ\text{F}$, and $x_4 = 2$ is $(1.965, 2.125)$. For those same values of the independent variables, find a 95% prediction interval for y , the number of breakdowns per 8-hour shift.