Submitted to the Annals of Applied Statistics

STAT 444 FINAL PROJECT PROPOSAL

BY ANGELO CARREON [1] , HOSEOK LEE [1,??]
JOY CHEN [1,??] \AND THIRD AUTHOR [1,??]

[1]*Department of Statistics and Actuarial Science, University of Waterloo, jaccarre@uwaterloo.ca*

This paper contains

1. Exploratory Data Analysis.

```
## Rows: 1460 Columns: 81
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (43): MSZoning, Street, Alley, LotShape, LandContour, Utilities, LotCon
## dbl (38): Id, MSSubClass, LotFrontage, LotArea, OverallQual, OverallCond, Y
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this me
```

1.1. *Summary Statistics.*

- the dataset contains 2930 rows and 82 columns
- 80 explanatory variables, 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variates.

```
nrow(housing)
```

```
## [1] 1460
```

```
ncol(housing)
```

```
## [1] 81
```

```
# remove column Order since it's just row index
housing <- housing[ , !(names(housing) == "Order")]
```

1.2. *Missing/duplicate Values.*

- some columns contain a large number of NAs
- 0 duplicates

```
# Alley: Type of alley access to property
sum(!is.na(housing$Alley))      # only 198 observations
```

```
## [1] 91
```

```
# Pool.QC: Pool quality
sum(!is.na(housing$Pool.QC))    # only 13 observations
```

```
## Warning: Unknown or uninitialised column: `Pool.QC`.
```

```
## [1] 0
```

```
# Misc.Feature: Miscellaneous feature not covered in other categories
sum(!is.na(housing$Misc.Feature)) # only 106 observations
```

Keywords and phrases: We love regression.

```
## Warning: Unknown or uninitialised column: 'Misc.Feature'.
```

```
## [1] 0
```

```
# Fence: Fence quality
sum(!is.na(housing$Fence))  # only 572 observations
```

```
## [1] 281
```

```
# Fireplace.Qu: Fireplace quality
sum(!is.na(housing$Fireplace.Qu)) # only 1508 observations
```

```
## Warning: Unknown or uninitialised column: 'Fireplace.Qu'.
```

```
## [1] 0
```

```
# remove column alley, pool QC, Misc Feature
drops <- c("Alley","Pool.QC", "Fence", "Misc.Feature")
newHousing <- housing[ , !(names(housing) %in% drops)]

# Check for duplicates in the entire dataset (0 duplicates)
duplicates <- housing[duplicated(housing), ]
dim(duplicates)
```

```
## [1]  0 81
```

1.3. *Distribution of dependent variable.*

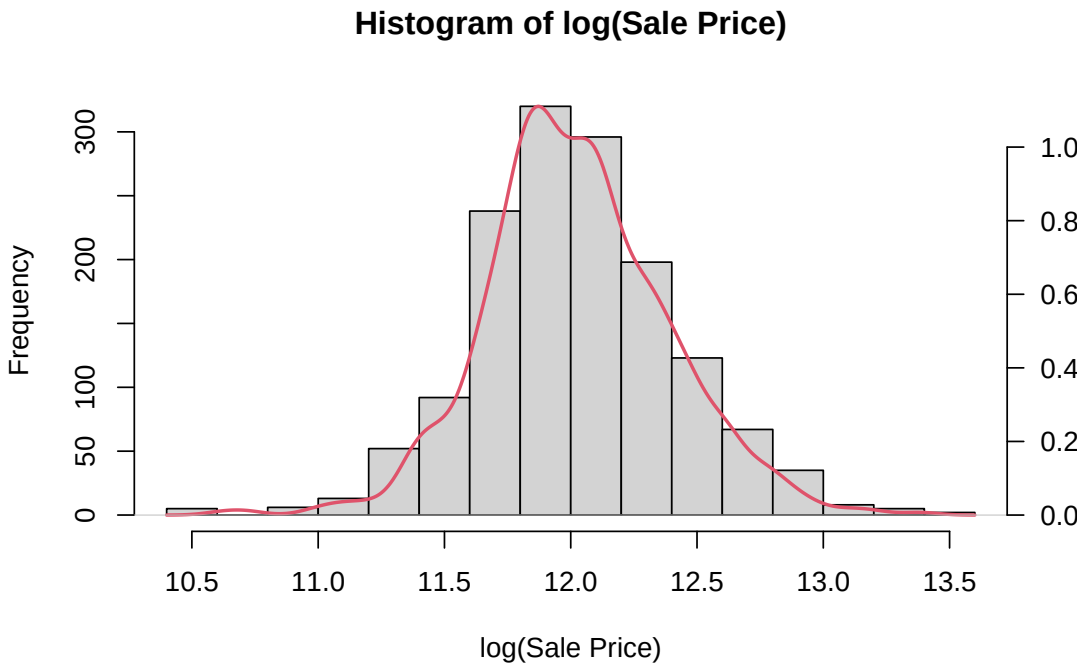
- the distribution of Sale Price looks very right-skewed: most values are clustered around the left tail while the right tail is longer (extreme high values)
- The sale prices range from $12,789 to $755,000 with a mean of $180,796 and a standard deviation of $79,886.69.
- Need to apply natural log transformation to address the non-normal distribution

```
summary(housing$SalePrice)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34900  129975  163000  180921  214000  755000
```

```
options(scipen=10)
hist(housing$SalePrice, xlab = "Property Sale Price", main = "Histogram of Sal
lines(housing$SalePrice, col = 4, lwd = 2)
par(new=TRUE)
plot(density(housing$SalePrice), col=2, lwd = 2, yaxt="n", xaxt="n",
     bty='n', xlab="", ylab="", main='')
axis(4, las=1)
```


```
hist(log(housing$SalePrice), xlab = "log(Sale Price)", main = "Histogram of lo
par(new=TRUE)
plot(density(log(housing$SalePrice)), col=2, lwd = 2, yaxt="n", xaxt="n",
     bty='n', xlab="", ylab="", main='')
axis(4, las=1)
```



1.4. *Correlations with the dependent variable.*

- Top 5 variables with the highest correlations:
  - Overall.Qual: Rates the overall material and finish of the house
  - Gr.Liv.Area: Above grade (ground) living area square feet
  - Garage.Cars: Size of garage in car capacity
  - Garage.Area: Size of garage in square feet
  - Total.Bsmt.SF: Total square feet of basement area

```
# Subset of numeric columns
numericHousingData <- newHousing[, sapply(newHousing, is.numeric)]

# Calculate the correlations between each numeric column and the response vari
correlations <- cor(numericHousingData[, -which(names(numericHousingData) == '
                    numericHousingData$SalePrice, use = "complete.obs")

# rank the correlations into two groups (positive and negative)
posCorrelations <- correlations[correlations[,1] >= 0, ]
negCorrelations <- correlations[correlations[,1] < 0, ]

# Print the correlations
print(sort(posCorrelations,decreasing=TRUE))
```

```
##  OverallQual    GrLivArea   GarageCars   GarageArea  TotalBsmtSF     1stFlr
##   0.79788068   0.70515357   0.64703361   0.61932962   0.61561224   0.607969
##     FullBath TotRmsAbvGrd    YearBuilt YearRemodAdd  GarageYrBlt   MasVnrAr
##   0.56662744   0.54706736   0.52539360   0.52125327   0.50475302   0.488658
##   Fireplaces   BsmtFinSF1  LotFrontage  OpenPorchSF   WoodDeckSF     2ndFlr
##   0.46187269   0.39030052   0.34426977   0.34335381   0.33685512   0.306879
##      LotArea     HalfBath BsmtFullBath    BsmtUnfSF BedroomAbvGr  ScreenPor
##   0.29996221   0.26856030   0.23673741   0.21312868   0.16681389   0.110426
##     PoolArea       MoSold    3SsnPorch
##   0.09248812   0.05156806   0.03077659
```

```
print(sort(negCorrelations,decreasing=FALSE))
```

```
## EnclosedPorch  KitchenAbvGr   OverallCond    MSSubClass            Id
##  -0.154843204  -0.140497445  -0.124391232  -0.088031702  -0.047121850
##  BsmtHalfBath       MiscVal    BsmtFinSF2        YrSold  LowQualFinSF
##  -0.036512665  -0.036041237  -0.028021366  -0.011868823  -0.001481983
```

1.5. *Check if there is a high degree of correlation or linear association among independent variables.*

```
cor_matrix <- cor(numericHousingData, use = "complete.obs")

corr_df = data.frame(cor_matrix)
corr_df$var = row.names(corr_df)

# Positive
corr_df1 = reshape2::melt(corr_df, value_name = "Corr")
```

```
## Using var as id variables
```

```
corr_df1 = corr_df1[(corr_df1$value > 0.7 & corr_df1$value < 1),]
print(corr_df1[order(corr_df1$value,decreasing=TRUE),])
```

```
##               var     variable     value
## 1016   GarageArea   GarageCars 0.8394149
## 1053   GarageCars   GarageArea 0.8394149
## 470      1stFlrSF  TotalBsmtSF 0.8359994
## 507   TotalBsmtSF    X1stFlrSF 0.8359994
## 632  TotRmsAbvGrd    GrLivArea 0.8243121
## 891     GrLivArea TotRmsAbvGrd 0.8243121
## 254   GarageYrBlt    YearBuilt 0.8235195
## 957     YearBuilt  GarageYrBlt 0.8235195
## 190     SalePrice  OverallQual 0.7978807
## 1411  OverallQual    SalePrice 0.7978807
## 646     SalePrice    GrLivArea 0.7051536
## 1423    GrLivArea    SalePrice 0.7051536
```

```
maxpos <- corr_df1[order(corr_df1$value),]

# Negative
corr_df2 = reshape2::melt(corr_df, value_name = "Corr")
```

```
## Using var as id variables
```

```
corr_df2 = corr_df2[(corr_df2$value < -0.7 & corr_df2$value > -1),]
print(corr_df2[order(corr_df2$value,decreasing=TRUE),])
```

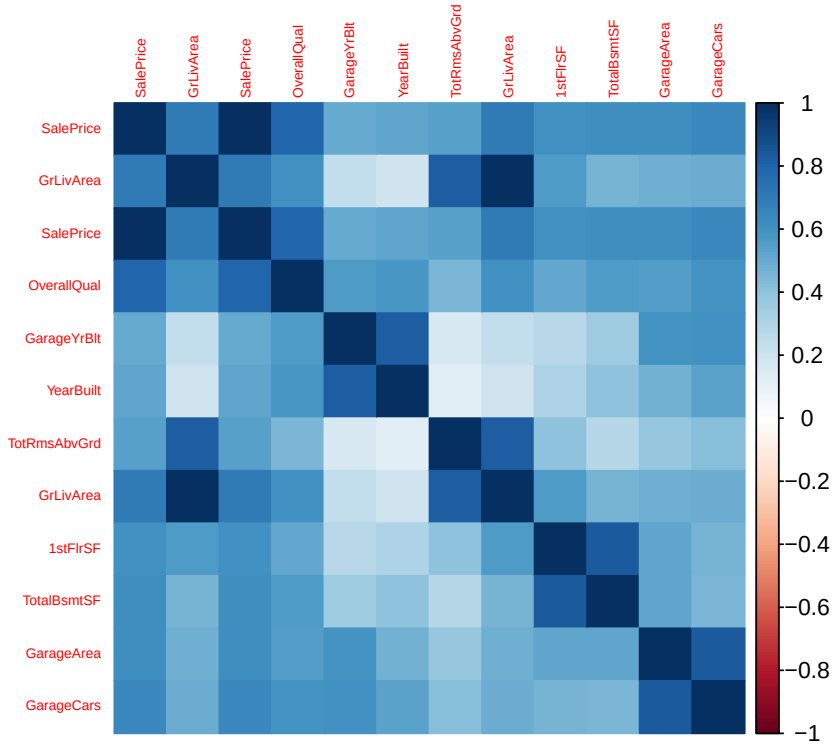
```
## [1] var      variable value
## <0 rows> (or 0-length row.names)
```

```
maxneg <- corr_df2[order(corr_df2$value),]

# Correlation matrix including only the top positive and negative correlations
names <- c(maxpos$var, maxneg$var)
topcormat <- cor_matrix[names, names]

# plot the correlation matrix
# install.packages('corrplot')
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.1
```

```
## corrplot 0.92 loaded
```

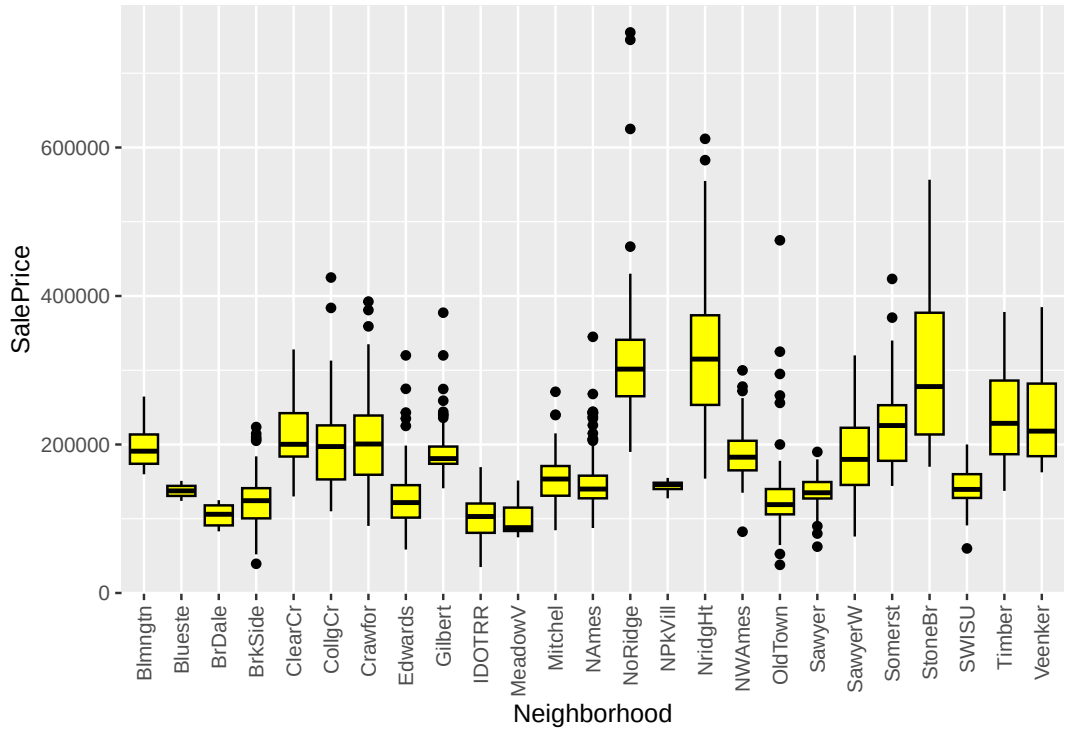
```
corrplot(topcormat, method='color', tl.cex=0.5)
```



1.6. *Other interesting but non-numeric variables.*

- Neighborhood
- Street: Type of road access to property
  - no predictive power; most likely irrelevant to house prices
  - can be dropped
- Utilities: Type of utilities available
  - There is not much variability in the distribution of data. All observations but 3 have "AllPub" as their values for Utilities
  - can be dropped
- Roof.Matl: Roof material
  - The most common data item "CompShg" occurs 2887 times out of 2930 observations
  - can be dropped

```
ggplot(housing, aes(x=Neighborhood,y=SalePrice)) + geom_boxplot(color="black",
```




```
as.data.frame(table(housing$Street))
```

```
##   Var1 Freq
## 1 Grvl    6
## 2 Pave 1454
```

```
as.data.frame(table(housing$Utilities))
```

```
##     Var1 Freq
## 1 AllPub 1459
## 2 NoSeWa    1
```

```
as.data.frame(table(housing$Roof.Matl))
```

```
## Warning: Unknown or uninitialised column: 'Roof.Matl'.
```

```
## [1] Freq
## <0 rows> (or 0-length row.names)
```

2. Introduction. This template helps you to create a properly formatted LaTeX 2_ε manuscript. Prepare your paper in the same style as used in this sample .pdf file. Try to avoid excessive use of italics and bold face. Please do not use any LaTeX 2_ε or TeX commands that affect the layout or formatting of your document (i.e., commands like `\textheight`, `\textwidth`, etc.).

3. Section headings. Here are some sub-sections:

3.1. *A sub-section.* Regular text.

3.1.1. *A sub-sub-section.* Regular text.

4. Text.

4.1. *Lists.* The following is an example of an *itemized* list, two levels deep.

- This is the first item of an itemized list. Each item in the list is marked with a "tick." The document style determines what kind of tick mark is used.
- This is the second item of the list. It contains another list nested inside it.
  - This is the first item of an itemized list that is nested within the itemized list.
  - This is the second item of the inner list. LaTeX allows you to nest lists deeper than you really should.
- This is the third item of the list.

The following is an example of an *enumerated* list of one level.

(i) This is the first item of an enumerated list.
(ii) This is the second item of an enumerated list.

The following is an example of an *enumerated* list, two levels deep.

1. This is the first item of an enumerated list. Each item in the list is marked with a "tick.". The document style determines what kind of tick mark is used.
2. This is the second item of the list. It contains another list nested inside of it.
   (i) This is the first item of an enumerated list that is nested within.
   (ii) This is the second item of the inner list. LaTeX allows you to nest lists deeper than you really should.

This is the rest of the second item of the outer list.
3. This is the third item of the list.

ird item of the list. \end{longlist}

4.2. *Punctuation.* Dashes come in three sizes: a hyphen, an intra-word dash like "U-statistics" or "the time-homogeneous model"; a medium dash (also called an "en-dash") for number ranges or between two equal entities like "1–2" or "Cauchy–Schwarz inequality"; and a punctuation dash (also called an "em-dash") in place of a comma, semicolon, colon or parentheses—like this.

Generating an ellipsis ... with the right spacing around the periods requires a special command.

4.3. *Citation.* Simple author and year cite: Billingsley (1999). Multiple bibliography items cite: Billingsley (1999); Bourbaki (1966) or (Billingsley, 1999; Bourbaki, 1966). Author only cite: Ethier and Kurtz. Year only cite: 1956 or (1956).

5. Fonts. Please use text fonts in text mode, e.g.:

Roman
Italic
Bold
SMALL CAPS
Sans serif
`Typewriter`

Please use mathematical fonts in mathematical mode, e.g.:

$ABCabc123$
$\mathit{ABCabc123}$
$\mathbf{ABCabc123}$
$\boldsymbol{ABCabc123\alpha\beta\gamma}$
$\mathcal{ABC}$
$\mathbb{ABC}$
$\mathsf{ABCabc123}$
$\mathtt{ABCabc123}$
$\mathfrak{ABCabc123}$

Note that `\mathcal, \mathbb` belongs to capital letters-only font typefaces.

6. Notes. Footnotes[1] pose no problem.[2]

7. Quotations. Text is displayed by indenting it from the left margin. There are short quotations

> This is a short quotation. It consists of a single paragraph of text. There is no paragraph indentation.

and longer ones.

> This is a longer quotation. It consists of two paragraphs of text. The beginning of each paragraph is indicated by an extra indentation.
>
> This is the second paragraph of the quotation. It is just as dull as the first paragraph.

8. Environments.

8.1. *Examples for* `plain`-style environments.

AXIOM 1. *This is the body of Axiom 1.*

PROOF. This is the body of the proof of the axiom above. □

CLAIM 2. *This is the body of Claim 2. Claim 2 is numbered after Axiom 1 because we used* `[axiom]` *in* `\newtheorem`.

THEOREM 8.1. *This is the body of Theorem 8.1. Theorem 8.1 numbering is dependent on section because we used* `[section]` *after* `\newtheorem`.

[1] This is an example of a footnote.

[2] Note that footnote number is after punctuation.

THEOREM 8.2 (Title of the theorem). *This is the body of Theorem 8.2. Theorem 8.2 has additional title.*

LEMMA 8.3. *This is the body of Lemma 8.3. Lemma 8.3 is numbered after Theorem 8.2 because we used* `[theorem]` *in* `\newtheorem`.

PROOF OF LEMMA 8.3. This is the body of the proof of Lemma 8.3. □

8.2. *Examples for* `remark`*-style environments.*

DEFINITION 8.4. This is the body of Definition 8.4. Definition 8.4 is numbered after Lemma 8.3 because we used `[theorem]` in `\newtheorem`.

EXAMPLE. This is the body of the example. Example is unnumbered because we used `\newtheorem*` instead of `\newtheorem`.

FACT. This is the body of the fact. Fact is unnumbered because we used `\newtheorem*` instead of `\newtheorem`.

9. Tables and figures. Cross-references to labeled tables: As you can see in Table1 and also in Table2.

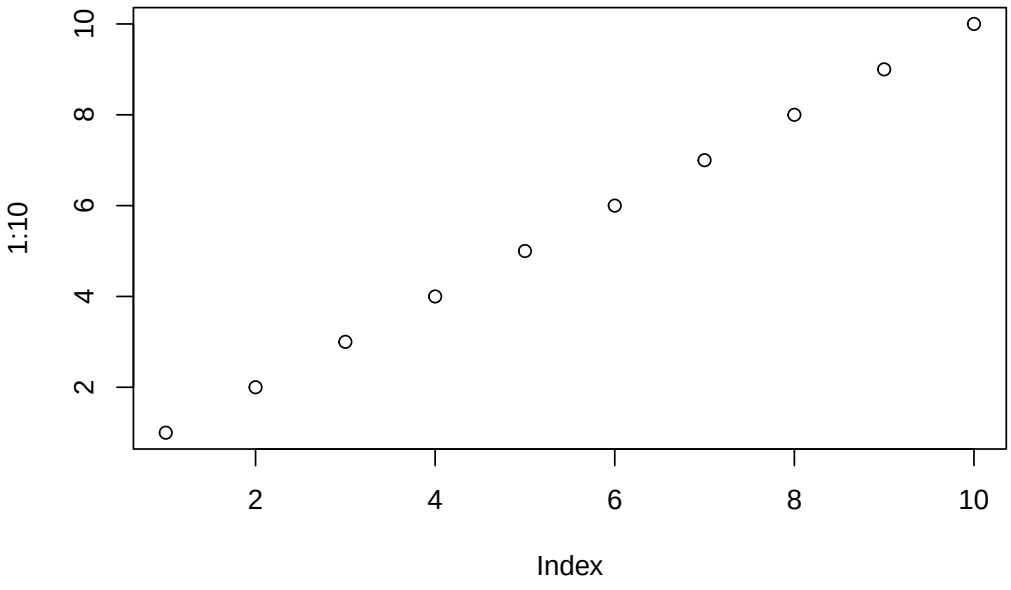


FIGURE 1. *Figure caption*

Sample of cross-reference to figure. Figure1 shows that it is not easy to get something on paper.

TABLE 1
Table caption

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

TABLE 2
Sample posterior estimates for each model

				Quantile		
Model	Parameter	Mean	Std. dev.	2.5%	50%	97.5%
Model 0	β_0	-12.29	2.29	-18.04	-11.99	-8.56
	β_1	0.10	0.07	-0.05	0.10	0.26
	β_2	0.01	0.09	-0.22	0.02	0.16
Model 1	β_0	-4.58	3.04	-11.00	-4.44	1.06
	β_1	0.79	0.21	0.38	0.78	1.20
	β_2	-0.28	0.10	-0.48	-0.28	-0.07
Model 2	β_0	-11.85	2.24	-17.34	-11.60	-7.85
	β_1	0.73	0.21	0.32	0.73	1.16
	β_2	-0.60	0.14	-0.88	-0.60	-0.34
	β_3	0.22	0.17	-0.10	0.22	0.55

10. Equations and the like. Two equations:

$$C_s = K_M \frac{\mu/\mu_x}{1 - \mu/\mu_x} \tag{10.1}$$

and

$$G = \frac{P_{\text{opt}} - P_{\text{ref}}}{P_{\text{ref}}} 100(\%). \tag{10.2}$$

Equation arrays:

$$\frac{dS}{dt} = -\sigma X + s_F F, \tag{10.3}$$

$$\frac{dX}{dt} = \mu X, \tag{10.4}$$

$$\frac{dP}{dt} = \pi X - k_h P, \tag{10.5}$$

$$\frac{dV}{dt} = F. \tag{10.6}$$

One long equation:

$$\begin{aligned} \mu_{\text{normal}} &= \mu_x \frac{C_s}{K_x C_x + C_s} \\ &= \mu_{\text{normal}} - Y_{x/s}\big(1 - H(C_s)\big)(m_s + \pi/Y_{p/s}) \\ &= \mu_{\text{normal}}/Y_{x/s} + H(C_s)(m_s + \pi/Y_{p/s}). \end{aligned} \tag{10.7}$$

APPENDIX: TITLE

Appendices should be provided in `{appendix}` environment, before Acknowledgements.

If there is only one appendix, then please refer to it in text as ... in the Appendix.

APPENDIX A: TITLE OF THE FIRST APPENDIX

If there are more than one appendix, then please refer to it as ... in Appendix A, Appendix B, etc.

APPENDIX B: TITLE OF THE SECOND APPENDIX

B.1. First subsection of Appendix B. Use the standard LaTeX commands for headings in `{appendix}`. Headings and other objects will be numbered automatically.

$$\mathcal{P} = (j_{k,1}, j_{k,2}, \ldots, j_{k,m(k)}). \tag{B.1}$$

Sample of cross-reference to the formula (B.1) in Appendix B.

Acknowledgements. The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

The first author was supported by NSF Grant DMS-??-??????.

The second author was supported in part by NIH Grant ???????????.

SUPPLEMENTARY MATERIAL

Title of Supplement A
Short description of Supplement A.

Title of Supplement B
Short description of Supplement B.

REFERENCES

BILLINGSLEY, P. (1999). *Convergence of Probability Measures*, 2nd ed. ed. John Wiley & Sons.

BOURBAKI, N. (1966). *General Topology* **1**. Addison–Wesley, Reading, MA.

ETHIER, S. N. and KURTZ, T. G. (1985). *Markov Processes: Characterization and Convergence*. Wiley, New York.

PROKHOROV, Y. V. (1956). Convergence of random processes and limit theorems in probability theory. *Theory of Probability & Its Applications* **1** 157–214.