Methods Qualifying Exam 2021

Jack Tubbs

8/10/2021

Contents

Exam

Your exam consists of three parts. The first two involve the analysis of two data sets found in SASHELP, cars and low birth weight data. Each has been modified for this exam. The BOX folder that I created for your exam contains the files that you will need. Read the README file before you begin. The third part of the exam does not involve any analysis. Rather, you will describe your mental process for conducting the analysis that answers the problem at hand[1]. For example, you might believe that the solution involves doing analysis, **A**, for which you will need to assume conditions, **B**. Tell me what you would need to do to verify that conditions **B** hold. What would you hope to determine from your analysis? Can you foresee potential issues that might arise since your study involves human subjects? Your solution will have strengths and weaknesses. Describe each and state why you think the advantages of your proposal out weight the disadvantages.

I have attempted to assist you by creating an Overleaf document with some R and SAS code that illustrates some of the initial steps that one might use in the analysis. You are welcome to use as much or as little of my code for your needs. When answering the questions you can use either SAS or R. I do not need nor want

[1]You have hear people say, “don’t just sit there, do something". Here, I am saying “Don’t do something, just sit there and think before you act. Tell me what you are thinking!

to see the same analysis done using both SAS and R. Do Not exclusively use SAS or R. It is your choice as to how much of each that you use. Your choice will effect your grade! I do not want to see your failed attempts to answer the question at hand, nor do I need to see all the steps that you needed to arrive at your final answer (e.g., model selection procedures).

Answer the questions and present your solutions in a document similar to what I have produced using LaTeX in Overleaf (or your choice of a TeX/LaTeX format). A Word-type document is Not Acceptable! In some cases, the question has been intentionally left vague or open ended. If you encounter these, then tell me what you intend to do, why you did it, and what you found. Hint: Do Not redefine your problem so that it is either trivial or overly complex. Use the methods that were presented in STAT 5380/5381. There shouldn't be any Bayesian solutions. Save those for your PPP!. The level of difficulty for this exam is consistent with what one would expect of an MS in Statistics. You are welcomed to use any of the resources given in the methods sequence (or any other resource that does not consume Oxygen).

Part 1 – Cars Data

I have derived two additional variables; discount from MSRP and overall MPG. They are defined as

$$Discount = \frac{(MSRP - Invoice)}{MSRP} \times 100$$

and

$$MPG\_overall = \theta * MPG\_Highway + (1 - \theta) * MPG\_City$$

where in my example $\theta = .4$.

1. Describe the data using simple descriptive methods
2. Suppose that $Y = Discount$ is the dependent variable of interest where you assume that Y is normally distributed. Is this assumption satisfied? Justify, your answer. If not, what happens if
   - You wish to make Y more normal? What did you do? Did it work (show me)?
   - Since, Hybrid cars are seldom discounted, remove this type and repeat above, what do you have now?
   - How does the class variable **Origin** effect the median or mean discount price for **Type = "sedans"**? Answer this question, if the new Y is a) normal, b) not normal. What did you do with the Hybrid vehicles? Was this necessary when the data are not normally distributed? Justify your answer.
3. Let $Y = MPG\_overall$ with your choice for θ. Determine your "best" least squares linear model for Y when using any of the remaining independent variables (do not use MPG for city or highway). Did your choice of θ make any difference? Explain your answer.
4. How does this model compare with variables found when using CART/RF?
5. Let $high\_discount = I(discount \geq 10)$. Find your "best" logistic model for $high\_discount$. What is your estimate for the probability of a $high\_discount$ for a Ford F-150 Supercab Lariat?

Part 2 – Infant Birth Weight Data

1. Describe the data using descriptive methods.
2. What is the relationship between **Smoking** and **Lowbirthwght**? **Drinking** and **Lowbirthwght**? Does **Race** matter in these relationships?
3. Remove **race = "Native"** from the data and repeat the above question.
4. Let **Death(event='Yes')** be the event of interest, determine your "best" model for this event? How does this model compare with results when using CART/RF?
5. Are either **Drinking** or **Smoking** causal for infant deaths? Explain you answer.

Part 3 – Three-armed Repeated Measures Study

As the lead statistician for the following clinical trial, provide your initial analysis plan for the study in order to get external approval. There is no analysis since there is no data at this time. Your plan should have clearly stated goals and objectives. Hint: concentrate on the whats, whens and whys, rather than the hows.

1. A clinical study consists of randomly assigning subjects to one of three groups, **A, B** and control **C** for which the clinical response of interest is **Y**. Measurements are scheduled to be taken at study onset (time for subject randomization into the three groups) and every 6 weeks for the entire 36 week study. All the clinical subjects enter the trial at day 1 of the study. Additional covariates, **X** are available for which some are time independent and some are measured at each clinical visit. Since the measurement of **Y** is time consuming, the study was conducted at 4 medical centers in Minnesota and Wisconsin. The enrollment is such that the available subjects in each group is about the same at each of the four sites.

Needed R Packages

```r
if(!require(FSA)){install.packages("FSA")}
if(!require(ggplot2)){install.packages("ggplot2")}
if (!require("mosaic")) install.packages("mosaic", dep=FALSE)
if (!require("nortest")) install.packages("nortest", dep=TRUE)
if (!require("epitools")) install.packages("epitools", dep=TRUE)
if (!require("prettyR")) install.packages("prettyR", dep=TRUE)
if (!require("rms")) install.packages("rms", dep=TRUE)
# add other as needed
```

Problem 1 Cars Data

Read data from SAS input file

```r
# this data came from SASHELP.CARS
cars = read.csv('cars.csv', header = TRUE)
cars = data.frame(cars)
summary(cars)
```

```
##      Make              Model               Type              Origin
##  Length:428         Length:428         Length:428         Length:428
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   DriveTrain             MSRP           Invoice          EngineSize
##  Length:428         Min.   : 10280   Min.   :  9875   Min.   :1.300
##  Class :character   1st Qu.: 20334   1st Qu.: 18866   1st Qu.:2.375
##  Mode  :character   Median : 27635   Median : 25294   Median :3.000
##                     Mean   : 32775   Mean   : 30015   Mean   :3.197
##                     3rd Qu.: 39205   3rd Qu.: 35710   3rd Qu.:3.900
##                     Max.   :192465   Max.   :173560   Max.   :8.300
##
##    Cylinders       Horsepower       MPG_City      MPG_Highway
##  Min.   : 3.000   Min.   : 73.0   Min.   :10.00   Min.   :12.00
##  1st Qu.: 4.000   1st Qu.:165.0   1st Qu.:17.00   1st Qu.:24.00
##  Median : 6.000   Median :210.0   Median :19.00   Median :26.00
##  Mean   : 5.808   Mean   :215.9   Mean   :20.06   Mean   :26.84
```

```
##  3rd Qu.: 6.000   3rd Qu.:255.0   3rd Qu.:21.25   3rd Qu.:29.00
##  Max.   :12.000   Max.   :500.0   Max.   :60.00   Max.   :66.00
##  NA's   :2
##      Weight       Wheelbase         Length
##  Min.   :1850   Min.   : 89.0   Min.   :143.0
##  1st Qu.:3104   1st Qu.:103.0   1st Qu.:178.0
##  Median :3474   Median :107.0   Median :187.0
##  Mean   :3578   Mean   :108.2   Mean   :186.4
##  3rd Qu.:3978   3rd Qu.:112.0   3rd Qu.:194.0
##  Max.   :7190   Max.   :144.0   Max.   :238.0
```

I will need these variables to be character variables rather than continuous variables. I have defined 'USA' and 'sedan' as logical variables. I'm not sure I need these in R but will produce something like this when doing SAS

```r
cars = transform(cars, Make.f = as.factor(Make))
cars = transform(cars, Type.f = as.factor(Type))
cars = transform(cars, Origin.f = as.factor(Origin))
cars = transform(cars, DriveTrain.f = as.factor(DriveTrain))
#define binary variables USAS vs non USA
#                         Sedan vs not a sedan
USA = (cars$Origin=='USA')
sedan = (cars$Type=='Sedan')
#define discount
discount = (cars$MSRP-cars$Invoice)/cars$MSRP*100
summary(discount)
```

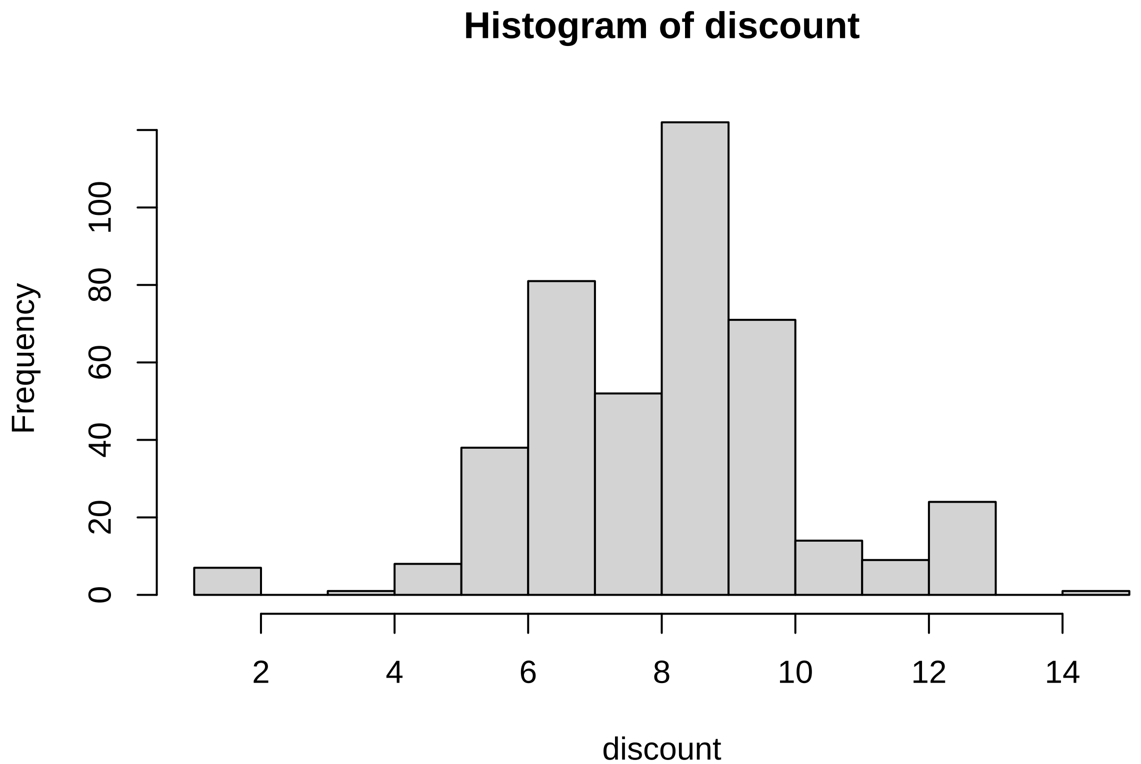
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.211   6.851   8.262   8.064   9.183  14.209
```

```r
#discount = discount[cars$Type != 'Hybrid']
#summary(discount)
#
#Compute MPG if one drive 40% of the time on a highway
#
MPG = .4*cars$MPG_Highw + .6*cars$MPG_City
#remove Hybrid from data
MPG = MPG[cars$Type != 'Hybrid']
summary(MPG)
```

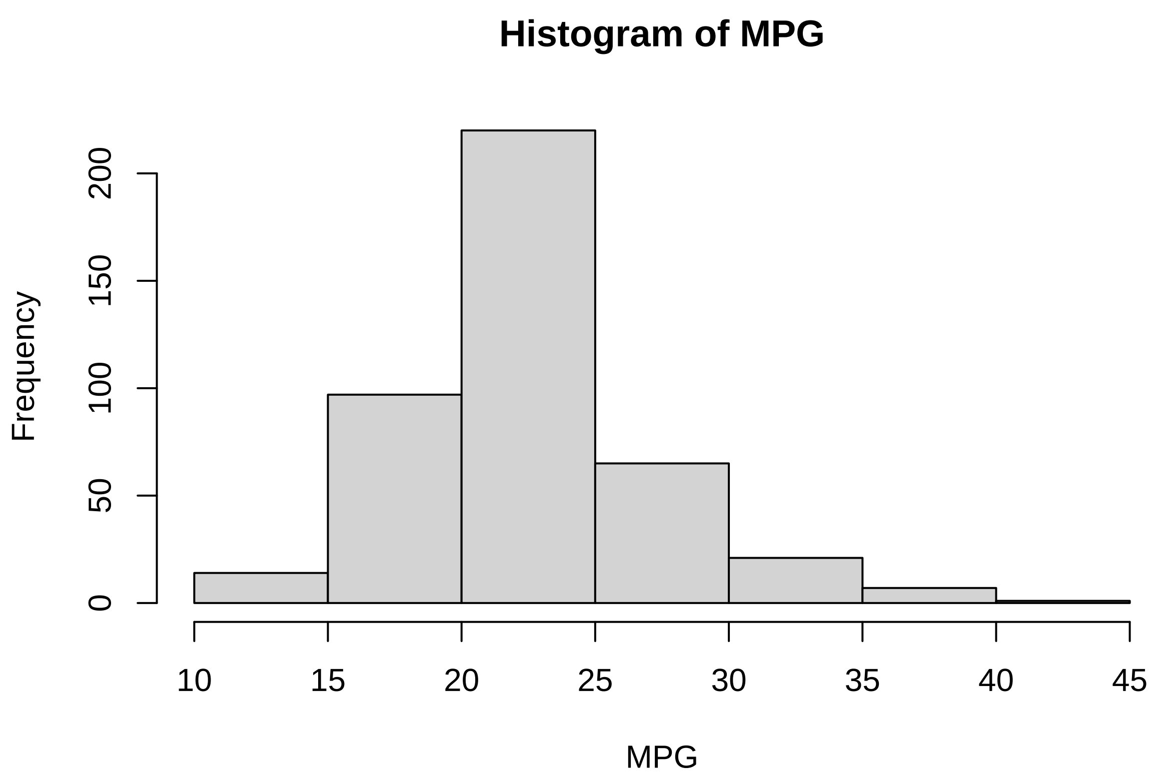
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.80   19.80   22.00   22.54   24.40   41.20
```

Notice the small values of discount in the histogram, these are likely for type = 'Hybrid', one can confirm this by removing these cars. There are large 'MPG' values, which again are probably due to having type = 'Hybrid'

```r
hist(discount)
```




```r
hist(MPG)
```




```r
library("mosaic")
favstats(MPG, data=cars)
```

```
##   min   Q1 median   Q3  max     mean       sd   n missing
```

```
##  10.8 19.8     22 24.4 41.2 22.54353 4.594481 425       0
```

```r
mean(MPG, trim=.05)
```

```
## [1] 22.35039
```

```r
quantile(MPG, seq(from=.025, to= .975, by=.1))
```

```
##  2.5% 12.5% 22.5% 32.5% 42.5% 52.5% 62.5% 72.5% 82.5% 92.5%
## 14.92 17.60 19.48 20.76 21.24 22.20 23.20 24.20 26.56 29.60
```

```r
#test for mu = 20.5
t.test(MPG, mu=22.5)
```

```
##
##  One Sample t-test
##
## data:  MPG
## t = 0.19532, df = 424, p-value = 0.8452
## alternative hypothesis: true mean is not equal to 22.5
## 95 percent confidence interval:
##  22.10547 22.98159
## sample estimates:
## mean of x
##  22.54353
```

```r
library("nortest")

ad.test(MPG)
```

```
##
##  Anderson-Darling normality test
##
## data:  MPG
## A = 4.0611, p-value = 4.061e-10
```

```r
cvm.test(MPG)
```

```
##
##  Cramer-von Mises normality test
##
## data:  MPG
## W = 0.74751, p-value = 2.981e-08
```

```r
lillie.test(MPG)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  MPG
## D = 0.10746, p-value = 8.975e-13
```

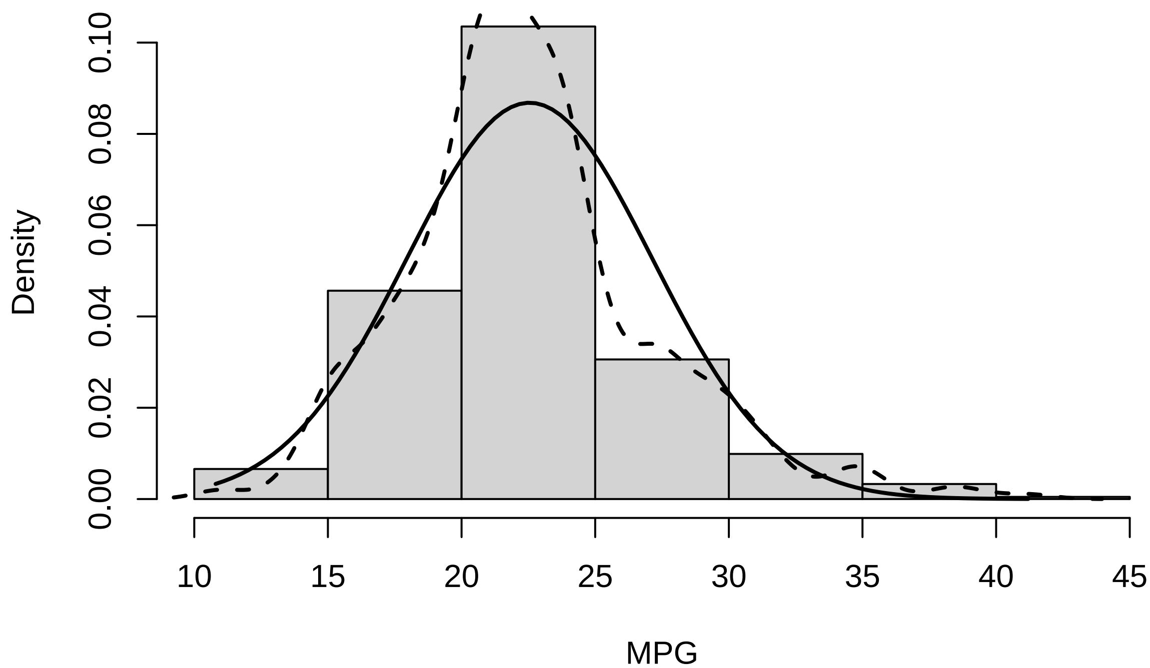
```r
pearson.test(MPG)
```

```
##
##  Pearson chi-square normality test
##
## data:  MPG
## P = 121.32, p-value < 2.2e-16
```

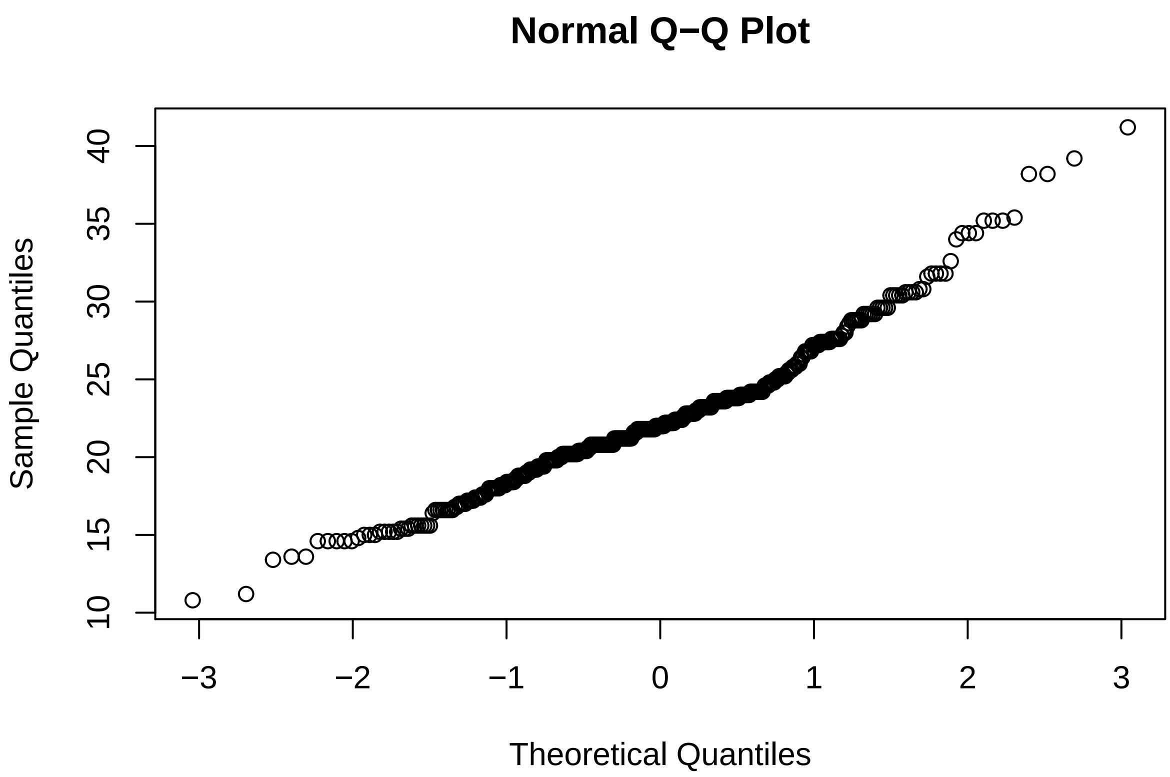
```r
sf.test(MPG)
```

```
##
##  Shapiro-Francia normality test
##
## data:  MPG
## W = 0.96421, p-value = 7.853e-08
```

```r
with(cars, hist(MPG, main="", freq=FALSE))
with(cars, lines(density(MPG), main="MPG", lty=2, lwd=2))
xvals = with(cars, seq(from=min(MPG), to=max(MPG), length=100))
with(cars, lines(xvals, dnorm(xvals, mean(MPG), sd(MPG)), lwd=2))
```




```r
qqnorm(MPG)
```



SAS Code for Cars Data

```
/*
Sashelp 2004 Car Data
The Sashelp.cars data set provides the 2004 car data.
The following steps display information about the data
set. The data set contains 428 observations.
*/
title "Sashelp.cars --- 2004 Car Data";
data cars; set sashelp.cars; run;
proc contents data=cars varnum;
ods select position;
run;
title "The First Five Observations Out of 428";
proc print data=cars(obs=5) noobs;
run;
title "The Type Variable";
proc freq data=cars;
tables Type;
run;

*Define a new variable discount = percent discount on price;
data cars; set cars;
discount = ((msrp - invoice)/msrp)*100;
run;

title 'Histrogram of percent discount';
proc sgplot data=cars;
histogram discount;
density discount/type=kernel;
run;

proc sgpanel data=cars;
panelby origin;
histogram discount;
density discount/type=kernel;
run;
```

```
title 'Descriptive Statistics for Discount';
proc univariate data=cars normal trim=.05 winsor=.05 mu0=7.25;
var discount;
run;

*define some additional variables;
data cars; set cars;
high_dis = (discount ge 9.2);  /*high_dis = 1 iff discount >= 9.2
                                 where 9.2 is Q3 for discount */
USA = (origin = 'USA');
sedan = (type = 'Sedan');
run;

title 'Tables with high discount';
proc freq data=cars; where type ne 'Hybrid';  /* do not include type = hybrid  */
tables high_dis*(origin type)/nopercent norow;
run;

proc freq data=cars; where type ne 'Hybrid';
tables (USA sedan)*high_dis/nopercent norow relrisk;
run;

title 'Modeling High Discount = 1';
proc logistic data=cars desc plots=roc;
class usa sedan DriveTrain;
model high_dis(ref='1') = horsepower drivetrain usa sedan mpg_city/expb;
run;

title 'Modeling discount';
proc glm data=cars plots=diagnostics; where sedan=1;
class usa  DriveTrain;
model discount = horsepower drivetrain usa  mpg_city/e;
run;
```

SAS Output for Cars Data

Sashelp.cars — 2004 Car Data

The CONTENTS Procedure

Variables in Creation Order					
#	*Variable*	*Type*	*Len*	*Format*	*Label*
1	Make	Char	13		
2	Model	Char	40		
3	Type	Char	8		
4	Origin	Char	6		
5	DriveTrain	Char	5		
6	MSRP	Num	8	DOLLAR8.	
7	Invoice	Num	8	DOLLAR8.	
8	EngineSize	Num	8		Engine Size (L)
9	Cylinders	Num	8		
10	Horsepower	Num	8		
11	MPG_City	Num	8		MPG (City)
12	MPG_Highway	Num	8		MPG (Highway)
13	Weight	Num	8		Weight (LBS)
14	Wheelbase	Num	8		Wheelbase (IN)
15	Length	Num	8		Length (IN)

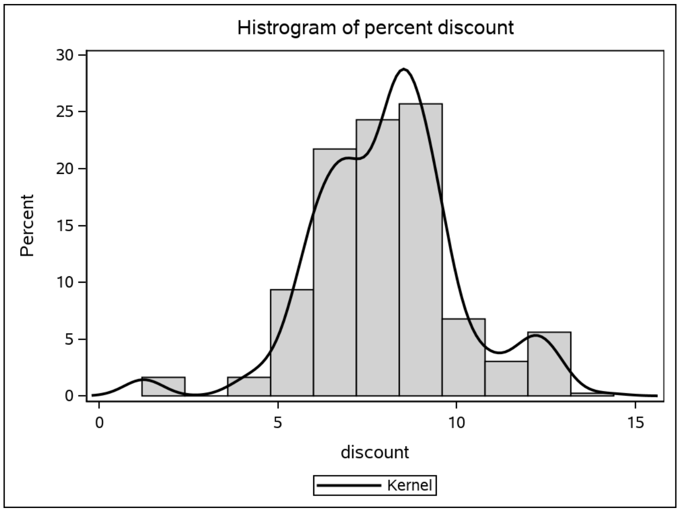
The First Five Observations Out of 428

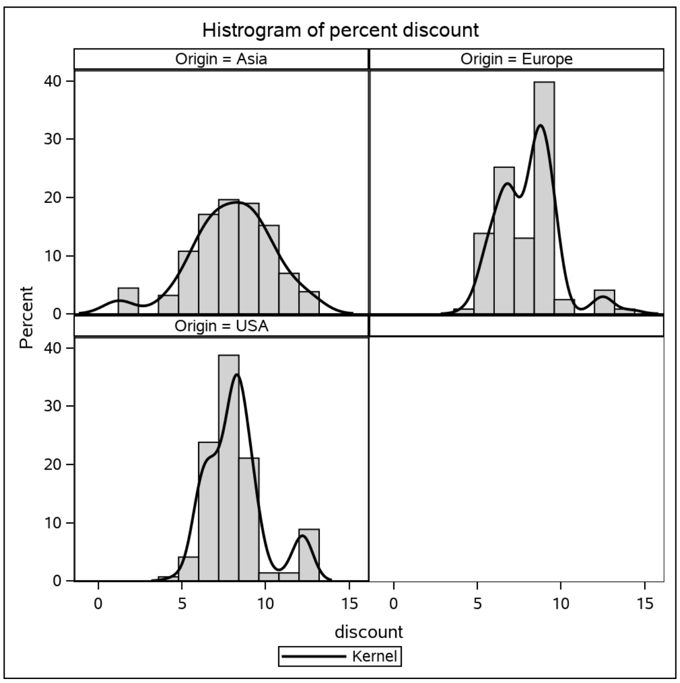
Make	*Type*	*Origin*	*DTrain*	*MSRP*	*Inv*	*ESize*	*Cyl*	*Hpower*	*MPG_C*	*MPG_H*	*Weight*	*Wbase*	*Length*
Acura	SUV	Asia	All	$36,945	$33,337	3.5	6	265	17	23	4451	106	189
Acura	Sedan	Asia	Front	$23,820	$21,761	2.0	4	200	24	31	2778	101	172
Acura	Sedan	Asia	Front	$26,990	$24,647	2.4	4	200	22	29	3230	105	183
Acura	Sedan	Asia	Front	$33,195	$30,299	3.2	6	270	20	28	3575	108	186
Acura	Sedan	Asia	Front	$43,755	$39,014	3.5	6	225	18	24	3880	115	197

The Type Variable

The FREQ Procedure

Type	*Frequency*	*Percent*	*Cumulative Frequency*	*Cumulative Percent*
Hybrid	3	0.70	3	0.70
SUV	60	14.02	63	14.72
Sedan	262	61.21	325	75.93
Sports	49	11.45	374	87.38
Truck	24	5.61	398	92.99
Wagon	30	7.01	428	100.00





When assessing the normality of ‘discount‘, I used goodness of fit tests on page 14 and was able to reject normality using any of the fit tests. I will remove the ‘Hybrid‘ cars and check for fit (not shown). The results did not change so I will try a BOXCOX transformation (not shown)

Descriptive Statistics for Discount

The UNIVARIATE Procedure

Variable: discount

Moments			
N	428	*Sum Weights*	428
Mean	8.0641809	*Sum Observations*	3451.46942
Std Deviation	2.02495566	*Variance*	4.10044544
Skewness	−0.1152351	*Kurtosis*	1.44956429

Moments			
Uncorrected SS	29584.164	Corrected SS	1750.8902
Coeff Variation	25.110494	Std Error Mean	0.09787993

Basic Statistical Measures			
Location		Variability	
Mean	8.064181	Std Deviation	2.02496
Median	8.261789	Variance	4.10045
Mode	6.729993	Range	12.99824
		Interquartile Range	2.33466

Note	Note: The mode displayed is the smallest of 3 modes with a count of 2.

Tests for Location: Mu0=7.25				
Test	Statistic		p Value	
Student's t	t	8.31816	Pr > |t|	<.0001
Sign	M	64	Pr >= |M|	<.0001
Signed Rank	S	21630	Pr >= |S|	<.0001

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.963372	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.07789	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.533154	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	3.863176	Pr > A-Sq	<0.0050

Trimmed Means								
% Trimmed in Tail	# Trimmed in Tail	Trimmed Mean	SE Trimmed Mean	95% CI		DF	t for H0:	Pr > |t|
5.14	22	8.063649	0.094856	7.877146	8.250152	383	8.577768	<.0001

Winsor Means								
% Winsor in Tail	# Winsor in Tail	Winsor Mean	SE Winsor Mean	95% CI		DF	t for H0:	Pr > |t|
5.14	22	8.129760	0.094868	7.943232	8.316288	383	9.273490	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	14.20899

Quantiles (Definition 5)	
Level	Quantile
99%	12.83385
95%	12.15129
90%	10.75741
75% Q3	9.18521
50% Median	8.26179
25% Q1	6.85054
10%	5.83108
5%	5.28645
1%	1.24871
0% Min	1.21075

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1.21075	372	12.8338	226
1.22045	371	12.8556	227
1.24687	367	12.8750	228
1.24705	370	12.8781	220
1.24871	373	14.2090	333

Tables with high discount

The FREQ Procedure

Table of high_dis by Origin				
high_dis	Origin			
	Asia	Europe	USA	Total
0	108 69.68	98 79.67	115 78.23	321
1	47 30.32	25 20.33	32 21.77	104
Total	155	123	147	425

Table of high_dis by Type						
high_dis	Type					
	SUV	Sedan	Sports	Truck	Wagon	Total
0	38 63.33	211 80.53	32 65.31	14 58.33	26 86.67	321
1	22 36.67	51 19.47	17 34.69	10 41.67	4 13.33	104
Total	60	262	49	24	30	425

Tables with high discount

The FREQ Procedure

Table of USA by high_dis			
USA	*high_dis*		
	0	*1*	*Total*
0	206 64.17	72 69.23	278
1	115 35.83	32 30.77	147
Total	321	104	425

Odds Ratio and Relative Risks			
Statistic	*Value*	*95% Confidence Limits*	
Odds Ratio	0.7961	0.4952	1.2800
Relative Risk (Column 1)	0.9472	0.8485	1.0573
Relative Risk (Column 2)	1.1897	0.8257	1.7144

Table of sedan by high_dis			
sedan	*high_dis*		
	0	*1*	*Total*
0	110 34.27	53 50.96	163
1	211 65.73	51 49.04	262
Total	321	104	425

Odds Ratio and Relative Risks			
Statistic	*Value*	*95% Confidence Limits*	
Odds Ratio	0.5017	0.3204	0.7854
Relative Risk (Column 1)	0.8380	0.7417	0.9467
Relative Risk (Column 2)	1.6704	1.1997	2.3258

Modeling High Discount = 1

The LOGISTIC Procedure

Model Information	
Data Set	WORK.CARS
Response Variable	high_dis
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	428
Number of Observations Used	428

Response Profile		
Ordered Value	*high_dis*	*Total Frequency*
1	1	104
2	0	324

Note	Probability modeled is high_dis=0.

Class Level Information			
Class	*Value*	*Design Variables*	
USA	*0*	1	
	1	−1	
sedan	*0*	1	
	1	−1	
DriveTrain	*All*	1	0
	Front	0	1
	Rear	−1	−1

Model Convergence Status
Convergence criterion (GCONV=1E−8) satisfied.

Model Fit Statistics		
Criterion	*Intercept Only*	*Intercept and Covariates*
AIC	476.654	450.140
SC	480.713	478.554
-2 Log L	474.654	436.140

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	38.5139	6	<.0001
Score	35.0948	6	<.0001
Wald	31.8153	6	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Horsepower	1	0.8715	0.3505
DriveTrain	2	4.5041	0.1052
USA	1	2.2883	0.1303
sedan	1	1.0313	0.3099
MPG_City	1	6.6120	0.0101

Analysis of Maximum Likelihood Estimates							
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		1	−0.9337	1.4630	0.4073	0.5234	0.393
Horsepower		1	−0.00234	0.00251	0.8715	0.3505	0.998
DriveTrain	All	1	−0.1770	0.2034	0.7577	0.3840	0.838
DriveTrain	Front	1	−0.2384	0.1872	1.6221	0.2028	0.788
USA	0	1	−0.2039	0.1348	2.2883	0.1303	0.816
sedan	0	1	−0.1361	0.1341	1.0313	0.3099	0.873
MPG_City		1	0.1408	0.0548	6.6120	0.0101	1.151

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Horsepower	0.998	0.993	1.003
DriveTrain All vs Rear	0.553	0.275	1.111
DriveTrain Front vs Rear	0.520	0.273	0.990
USA 0 vs 1	0.665	0.392	1.128
sedan 0 vs 1	0.762	0.450	1.288
MPG_City	1.151	1.034	1.282

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	69.9	Somers' D	0.398
Percent Discordant	30.1	Gamma	0.398
Percent Tied	0.1	Tau-a	0.147
Pairs	33696	c	0.699

Modeling discount

The GLM Procedure

Class Level Information		
Class	*Levels*	*Values*
USA	2	0 1
DriveTrain	3	All Front Rear

Number of Observations Read	262
Number of Observations Used	262

General Form of Estimable Functions	
Effect	*Coefficients*
Intercept	L1
Horsepower	L2
DriveTrain All	L3
DriveTrain Front	L4
DriveTrain Rear	L1−L3−L4
USA 0	L6
USA 1	L1−L6
MPG_City	L8

Modeling discount

The GLM Procedure

Dependent Variable: discount

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	5	116.8853410	23.3770682	8.13	<.0001
Error	256	735.6956393	2.8738111		
Corrected Total	261	852.5809803			

R-Square	*Coeff Var*	*Root MSE*	*discount Mean*
0.137096	21.83059	1.695232	7.765397

Source	*DF*	*Type I SS*	*Mean Square*	*F Value*	*Pr > F*
Horsepower	1	102.0175140	102.0175140	35.50	<.0001
DriveTrain	2	3.6193944	1.8096972	0.63	0.5336
USA	1	0.0025146	0.0025146	0.00	0.9764
MPG_City	1	11.2459180	11.2459180	3.91	0.0490

Source	*DF*	*Type III SS*	*Mean Square*	*F Value*	*Pr > F*
Horsepower	1	10.28669514	10.28669514	3.58	0.0596
DriveTrain	2	0.77901013	0.38950507	0.14	0.8733
USA	1	0.95679935	0.95679935	0.33	0.5644
MPG_City	1	11.24591796	11.24591796	3.91	0.0490

Problem 2 Infant Birth Weight Data

Read data from SAS input file

```
# this data came from SASHELP.BWEIGHT
bw = read.csv('bwgt.csv', header = TRUE)
bw = data.frame(bw)
#summary(bw)
bw = transform(bw, AgeGroup.f = as.factor(AgeGroup))
bw = transform(bw, Race.f = as.factor(Race))
bw = transform(bw, Drinking.f = as.factor(Drinking))
bw = transform(bw, Death.f = as.factor(Death))
bw = transform(bw, Smoking.f = as.factor(Smoking))
bw = transform(bw, SomeCollege.f = as.factor(SomeCollege))
bw = transform(bw, LowBirthWgt.f = as.factor(LowBirthWgt))

tally(~ AgeGroup + Race.f, data=bw)
```

```
##         Race.f
## AgeGroup Asian Black Hispanic Native White
##        1     5   134      164      8   223
##        2   205   514      845     33  2210
##        3    41    62      104      5   447
```

```
tally(~ Race.f | AgeGroup.f, data=bw)
```

```
##           AgeGroup.f
## Race.f        1    2    3
##   Asian       5  205   41
##   Black     134  514   62
##   Hispanic  164  845  104
##   Native      8   33    5
##   White     223 2210  447
```

```
library(mosaic)
mytab = tally(~ Race.f | AgeGroup.f, data=bw)
addmargins(mytab)
```

```
##           AgeGroup.f
## Race.f        1    2    3  Sum
##   Asian       5  205   41  251
##   Black     134  514   62  710
##   Hispanic  164  845  104 1113
##   Native      8   33    5   46
##   White     223 2210  447 2880
##   Sum       534 3807  659 5000
```

```
prop.table(mytab, 1)
```

```
##           AgeGroup.f
## Race.f              1          2          3
##   Asian    0.01992032 0.81673307 0.16334661
##   Black    0.18873239 0.72394366 0.08732394
##   Hispanic 0.14734951 0.75920934 0.09344115
##   Native   0.17391304 0.71739130 0.10869565
##   White    0.07743056 0.76736111 0.15520833
```

```r
library(epitools)
attach(bw)
mytab = tally(~ LowBirthWgt.f | Death.f, data=bw)
addmargins(mytab)
```

```
##              Death.f
## LowBirthWgt.f   No  Yes  Sum
##           No  4544   10 4554
##           Yes  425   21  446
##           Sum 4969   31 5000
```

```r
prop.table(mytab, 1)
```

```
##              Death.f
## LowBirthWgt.f          No         Yes
##           No  0.997804128 0.002195872
##           Yes 0.952914798 0.047085202
```

```r
riskratio(x=Smoking.f, y=Death.f)
```

```
## $data
##          Outcome
## Predictor   No Yes Total
##            278   2   280
##     No    3618  16  3634
##     Yes   1073  13  1086
##     Total 4969  31  5000
##
## $measure
##          risk ratio with 95% C.I.
## Predictor  estimate     lower    upper
##           1.0000000        NA       NA
##       No  0.6164007 0.1424454 2.667336
##       Yes 1.6758748 0.3803905 7.383351
##
## $p.value
##          two-sided
## Predictor midp.exact fisher.exact chi.square
##                   NA           NA         NA
##       No   0.5060677    0.3727212  0.5137826
##       Yes  0.5328608    0.7487939  0.4894594
##
## $correction
## [1] FALSE
##
## attr(,"method")
## [1] "Unconditional MLE & normal approximation (Wald) CI"
```

```r
#GLM Logistic Model
library(rms)
lrm(Death.f ~ LowBirthWgt.f + Smoking.f, data=bw)
```

```
## Logistic Regression Model
##
##  lrm(formula = Death.f ~ LowBirthWgt.f + Smoking.f, data = bw)
##
```

```
##                       Model Likelihood    Discrimination    Rank Discrim.
##                             Ratio Test           Indexes          Indexes
##  Obs          5000    LR chi2      71.64    R2       0.196    C       0.825
##   No          4969    d.f.             3    g        0.810    Dxy     0.651
##   Yes           31    Pr(> chi2) <0.0001    gr       2.248    gamma   0.794
##  max |deriv| 2e-09                          gp       0.008    tau-a   0.008
##                                             Brier    0.006
##
##                     Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept          -6.0476 0.7688 -7.87  <0.0001
##  LowBirthWgt.f=Yes   3.1075 0.3884  8.00  <0.0001
##  Smoking.f=No       -0.4061 0.7634 -0.53  0.5948
##  Smoking.f=Yes       0.5921 0.7751  0.76  0.4449
##
```

SAS Code for Low Birth Weight

```

The Sashelp.BirthWgt data set contains 100,000 random observations
about infant mortality in 2003 from the US National Center for Health Statistics.
Each observation records infant death within one year of birth, birth weight,
maternal smoking and drinking behavior, and other background
characteristics of the mother.

title "Sashelp.bweight --- Infant Birth Weight";
data birthwgt; set sashelp.birthwgt; run;
proc contents data=birthwgt varnum;
ods select position;
run;
title "The First Five Observations Out of 100,000";
proc print data=birthwgt(obs=10);
run;

*Create a new smaller datat set;
title 'New Sample of Size 5,000';
proc surveyselect data=birthwgt out=new_bwgt method=srs n=5000
                  seed=2021;
 *  strata death;
run;

proc freq data=new_bwgt;
tables race*agegroup/norow chisq relrisk riskdiff;
run;

proc freq data=new_bwgt;
tables death*LowBirthWgt*race /nopercent norow cmh;
run;

proc logistic data=new_bwgt plots=roc;where race ne 'Native';
class drinking race smoking lowbirthwgt death/param=glm;
model death(event='Yes') = drinking  smoking  lowbirthwgt/expb;
run;
```

SAS Output for Birth Weight

Sashelp.bweight — Infant Birth Weight

The CONTENTS Procedure

Variables in Creation Order			
#	*Variable*	*Type*	*Len*
1	LowBirthWgt	Char	3
2	Married	Char	3
3	AgeGroup	Num	8
4	Race	Char	9
5	Drinking	Char	3
6	Death	Char	3
7	Smoking	Char	3
8	SomeCollege	Char	3

The First Five Observations Out of 100,000

Obs	*LowBirthWgt*	*Married*	*AgeGroup*	*Race*	*Drinking*	*Death*	*Smoking*	*SomeCollege*
1	No	No	3	Asian	No	No	No	Yes
2	No	No	2	White	No	No	No	No
3	Yes	Yes	2	Native	No	Yes	No	No
4	No	No	2	White	No	No	No	No
5	No	No	2	White	No	No	No	Yes
6	No	No	2	White	No	No	No	
7	No	No	2	Asian	No	No	No	Yes
8	No	No	3	White	No	No	No	Yes
9	No	Yes	1	Black	No	No	No	No
10	No	No	2	Native	No	No	No	Yes

New Sample of Size 5,000

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	BIRTHWGT
Random Number Seed	2021
Sample Size	5000
Selection Probability	0.05
Sampling Weight	20
Output Data Set	NEW_BWGT

New Sample of Size 5,000

The FREQ Procedure

Table of Race by AgeGroup				
Race	*AgeGroup*			
	1	*2*	*3*	*Total*
Asian	5 0.10 0.94	205 4.10 5.38	41 0.82 6.22	251 5.02
Black	134 2.68 25.09	514 10.28 13.50	62 1.24 9.41	710 14.20
Hispanic	164 3.28 30.71	845 16.90 22.20	104 2.08 15.78	1113 22.26
Native	8 0.16 1.50	33 0.66 0.87	5 0.10 0.76	46 0.92
White	223 4.46 41.76	2210 44.20 58.05	447 8.94 67.83	2880 57.60
Total	534 10.68	3807 76.14	659 13.18	5000 100.00

Statistic	*DF*	*Value*	*Prob*
Chi-Square	8	144.4145	<.0001
Likelihood Ratio Chi-Square	8	147.5552	<.0001
Mantel-Haenszel Chi-Square	1	50.5502	<.0001
Phi Coefficient		0.1699	
Contingency Coefficient		0.1675	
Cramer's V		0.1202	

New Sample of Size 5,000

The FREQ Procedure

Table 1 of LowBirthWgt by Race						
Controlling for Death=No						
LowBirthWgt	Race					
	Asian	Black	Hispanic	Native	White	Total
No	219 87.60	597 84.68	1028 92.95	41 91.11	2659 92.87	4544
Yes	31 12.40	108 15.32	78 7.05	4 8.89	204 7.13	425
Total	250	705	1106	45	2863	4969

Table 2 of LowBirthWgt by Race						
Controlling for Death=Yes						
LowBirthWgt	Race					
	Asian	Black	Hispanic	Native	White	Total
No	1 100.00	1 20.00	2 28.57	0 0.00	6 35.29	10
Yes	0 0.00	4 80.00	5 71.43	1 100.00	11 64.71	21
Total	1	5	7	1	17	31

New Sample of Size 5,000

The FREQ Procedure

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)				
Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	30.7786	<.0001
2	Row Mean Scores Differ	1	30.7786	<.0001
3	General Association	4	56.1256	<.0001

New Sample of Size 5,000

The LOGISTIC Procedure

Model Information	
Data Set	WORK.NEW_BWGT
Response Variable	Death
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	4954
Number of Observations Used	4676

Response Profile		
Ordered Value	*Death*	*Total Frequency*
1	No	4647
2	Yes	29

Note	Probability modeled is Death='Yes'.

Note	278 observations were deleted due to missing values for the response or explanatory variables.

Class Level Information					
Class	*Value*	*Design Variables*			
Race	*Asian*	1	0	0	0
	Black	0	1	0	0
	Hispanic	0	0	1	0
	White	0	0	0	1
Smoking	*No*	1	0		
	Yes	0	1		
LowBirthWgt	*No*	1	0		
	Yes	0	1		

Model Convergence Status
Quasi−complete separation of data points detected.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	354.628	304.344
SC	361.078	362.396
-2 Log L	352.628	286.344

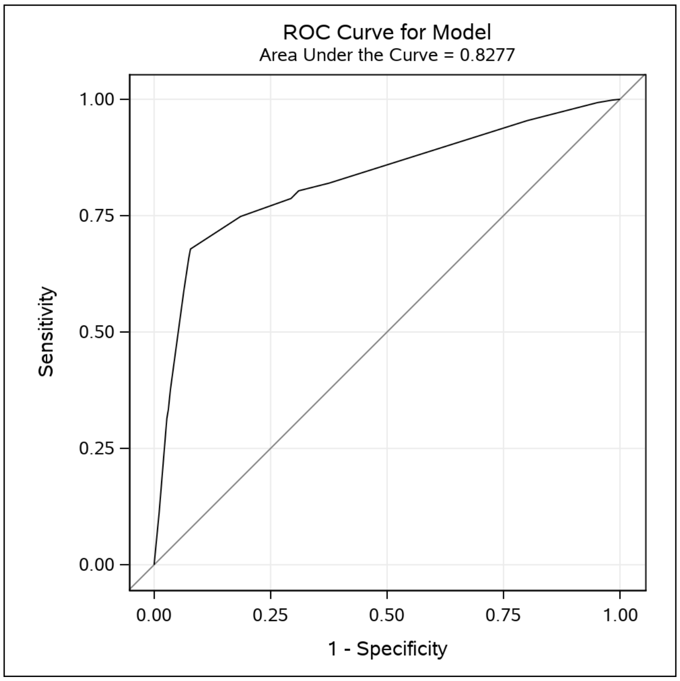
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	66.2838	8	<.0001
Score	126.0425	8	<.0001
Wald	66.0725	8	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Race	3	0.1446	0.9860
Smoking	1	0.0004	0.9851
Race*Smoking	3	0.3734	0.9457
LowBirthWgt	1	57.7302	<.0001

Analysis of Maximum Likelihood Estimates								
Parameter			DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept			1	−2.3701	0.4139	32.7893	<.0001	0.093
Race	Asian		1	−11.9884	404.3	0.0009	0.9763	0.000
Race	Black		1	0.0259	0.8342	0.0010	0.9752	1.026
Race	Hispanic		1	0.3241	0.6498	0.2488	0.6179	1.383
Race	White		0	0	.	.	.	.
Smoking	No		1	−0.8647	0.5160	2.8086	0.0938	0.421
Smoking	Yes		0	0	.	.	.	.
Race*Smoking	Asian	No	1	11.9486	404.3	0.0009	0.9764	154600.6
Race*Smoking	Asian	Yes	0	0	.	.	.	.
Race*Smoking	Black	No	1	−0.3916	1.0705	0.1338	0.7145	0.676
Race*Smoking	Black	Yes	0	0	.	.	.	.
Race*Smoking	Hispanic	No	1	−0.5328	0.9377	0.3228	0.5699	0.587
Race*Smoking	Hispanic	Yes	0	0	.	.	.	.
Race*Smoking	White	No	0	0	.	.	.	.
Race*Smoking	White	Yes	0	0	.	.	.	.
LowBirthWgt	No		1	−3.0566	0.4023	57.7302	<.0001	0.047
LowBirthWgt	Yes		0	0	.	.	.	.

Odds Ratio Estimates			
Effect	*Point Estimate*	*95% Wald Confidence Limits*	
LowBirthWgt No vs Yes	0.047	0.021	0.104

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	79.4	*Somers' D*	0.672
Percent Discordant	12.1	*Gamma*	0.735
Percent Tied	8.5	*Tau-a*	0.008
Pairs	134763	*c*	0.836



Answers

Part 1 – Cars Data

Provide your response to the respective questions here! (I have included some examples – please discard on your work) If you are testing a hypothesis, be sure to state the null and your conclusions.

1. Describe the data using simple descriptive methods I used R to create histogram and KDE for the continuous variables (output pages ? - ??) I observed that variable X was somewhat normal whereas variables Y and Z were highly skewed. Tables were created using SAS. I observed that the hybrids were not trucks or SUVs

2. Suppose that $Y = Discount$ is the dependent variable of interest where you assume that Y is normally distributed. Is this assumption satisfied? Justify, your answer. If not, what happens if

   - You wish to make Y more normal? What did you do? Did it work (show me)?
   - Since, Hybrid cars are seldom discounted, remove this type and repeat above, what do you have now?
   - How does the class variable **Origin** effect the median or mean discount price for **Type = "sedans"**? Answer this question, if the new Y is a) normal, b) not normal. What did you do with the Hybrid vehicles? Was this necessary when the data are not normally distributed? Justify your answer.

3. Let $Y = MPG\_overall$ with your choice for θ. Determine your "best" least squares linear model for Y when using any of the remaining independent variables (do not use MPG for city or highway). Did your choice of θ make any difference? Explain your answer. I found the following model using forward selection in R and this same model using LASSO in SAS. Output for R pages ?-?? and for SAS pages ?-??

4. Let $high\_discount = I(discount \geq 10)$. Find your "best" logistic model for $high\_discount$. What is your estimate for the probability of a $high\_discount$ for a Ford F-150 Supercab Lariat?

5. What are the independent variables of interest when using CART/RF?

Part 2 – Infant Birth Weight Data

Since SAS does not do a very good job with tables , I used R for questions 1 - 3. The answers are found on pages ? - ??

1. Describe the data using descriptive methods.

2. What is the relationship between **Smoking** and **Lowbirthwght**? **Drinking** and **Lowbirthwght**? Does **Race** matter in these relationships?

3. Remove **race = "Native"** from the data and repeat the above question.

4. Let **Death(event='Yes')** be the event of interest, determine your "best" model for this event? How does this model compare with results when using CART/RF? I used SAS and the solution is given on pages ? - ??

5. Are either **Drinking** or **Smoking** causal for infant deaths? Explain you answer. Wouldn't you like to know how I answer this question???

Part 3 – Three-armed Repeated Measures Study

As the lead statistician for the following clinical trial, provide your initial analysis plan for the study in order to get external approval. Your plan should have clearly stated goals and objectives. There is no analysis since there is no data at this time.

1. A clinical study consists of randomly assigning subjects to one of three groups, **A, B** and control **C** for which the clinical response of interest is **Y**. Measurements are scheduled to be taken at study onset (time for subject randomization into the three groups) and every 6 weeks for the entire 36 week study. All the clinical subjects enter the trial at day 1 of the study. Additional covariates, **X** are available for which some are time independent and some are measured at each clinical visit. Since the measurement of **Y** is time consuming, the study was conducted at 4 medical centers in Minnesota and Wisconsin. The enrollment is such that the available subjects in each group is about the same at each of the four sites.

Results of the Exam

I was reminded again as to how difficult our chosen field of statistics is - as compared to the simplistic thinking that one needs in other disciplines, such as, mathematics. This is not to say that mathematics is easy, far from it, but there does seem to be some consensus with regard to how to approach a problem and the answer that will be found. Not so with statistics!

My objective was to have the students perform data analysis with two medium to small data sets; cars and infant birth weight data. The first was primarily continuous variables whereas the second was primarily categorical or binary variables. The cars data was about 500 observations and the infant bw was in excess of 100 thousand from which I selected a subset of about 6000. The students were to use a combination (their choice) of R and SAS for the analysis in which the code and results were to be combined into a single latex pdf document. These activities were well practiced in the methods course and I suspect where not foreign to the David's computational class (sans SAS).

The questions on the exam were in some cases very specific, regression and RF for a continuous response and logistic regression and CART/RF for a binary response variable. In other cases the questions were more open-ended. Examine the data and tell me what you see. Does what you see affect what analysis you will do and what results you will find? The third question was for them to write the statistical design and analysis protocol for a planned experiment or phase 2 clinical trial. No data, just give me your plan.

So how did they do? Each of the 6 students were able to present the analysis from R and SAS in a single latex produced pdf file. So in that sense they did what I wanted. So how did they do with the statistics? They all did some things that were correct but Randy and Jamie performed the best and were somewhat better than their peers. Jamie needs to shake off the public health stuff in order to achieve the PhD in Statistics. She didn't seem to understand the logit transformation when making predictive statements about Ford Ranger trucks! None of the students provided a single model in response to question 3. If there is no data, just talk! Some were able to "see" some issues but again no equations or models. For example, just "use mixed models rather than GLM because they do better with repeated measures" (which is part of the title). Not a single student mentioned the role or need for the control group **C**. Jamie came the closest to having a model when she discussed GEE models for repeated measures. Emmie's answer looked like something you would see on an AP Stat response! Eddie and Sonish were clueless and Brad appears to have come from a non parallel other universe! Other issues that arose is that they are clueless about categorical data analysis (its a good thing we teach it several times) and goodness of fit with "large" data sets! One almost always rejects the null when n is large (or too small).

Conclusions

- None of our students are very good with the language of statistics. Speaking, writing, thinking, communicating or reasoning. The remote classes and lack of community have hurt our students in these areas. Their math skills seem to have diminished. This is difficult for me to say because I have grown to dislike much of what the F2F classes are about.
- I have tried to not be too harsh on their performance because there is a bit of them having to guess what I am thinking. However,
  - Randy and Jamie have a very good chance to be successful in our program.

  - Emmie is a project but appears to be very willing. My comment about AP Stat reflects her lack of maturity in what we are doing, she can grow into it if she is willing.
  - Bradley is a mystery! He appears to be capable yet he is lazy and wants to take short cuts when they are not permitted.
  - Eddie and Sonish performed the worst. I like Eddie but I think the PhD is over his head. I don't know Sonish, is that my issue or his? I suspect the students in this cohort think that Sonish is the best student!
- Unless their math stat performance is problematic, I would not object if each of them would continue in the program.