Advanced Data Analysis 1, Stat 427/527
Fall 2014

Homework 8

Name: ______________________________
Prof. Erik B. Erhardt

Part I. (80 points) Do all calculations in R. All R code for the assignment should be included with the part of the problem it addresses (for code and output use a fixed-width font, such as Courier). Code is used to calculate result. Text is used to report and interpret results. Do not report or interpret results in the code. Also:

1. Clearly define population parameters in each problem. That is, give a verbal description of what the population mean is in the context of the problem.
2. Clearly specify hypotheses when appropriate (not every problem involves a test of hypothesis).
3. Write a coherent conclusion based on each CI or test.

(50pts) **1. Gas milage and automobile horsepower**

Variation in gasoline mileage among makes and models of automobiles is influenced substantially by the weight and horsepower of the vehicles. Eighty-two automobile makes and models through 1991 were compared for fuel economy. The data below are provided by the US Environmental Protection Agency.

```
Reference: R.M. Heavenrich, J.D. Murrell, and K.H. Hellman.
  Light Duty Automotive Technology and Fuel Economy Trends Through 1991,
  U.S. Environmental Protection Agency, 1991 (EPA/AA/CTAB/91-02).
Variable Names:
  vol: Cubic feet of cab space
   hp: Engine horspower
  mpg: Average miles per gallon
   sp: Top speed (mph)
   wt: Vehicle weight (100 lb)
```

make/model	vol	hp	mpg	sp	wt
GM/GeoMetroXF1	89	49	65.4	96	17.5
GM/GeoMetro	92	55	56.0	97	20.0
GM/GeoMetroLSI	92	55	55.9	97	20.0
SuzukiSwift	92	70	49.0	105	20.0
DaihatsuCharade	92	53	46.5	96	20.0
GM/GeoSprintTurbo	89	70	46.2	105	20.0
GM/GeoSprint	92	55	45.4	97	20.0
HondaCivicCRXHF	50	62	59.2	98	22.5
HondaCivicCRXHF	50	62	53.3	98	22.5
DaihatsuCharade	94	80	43.4	107	22.5
SubaruJusty	89	73	41.1	103	22.5
HondaCivicCRX	50	92	40.9	113	22.5
HondaCivic	99	92	40.9	113	22.5
SubaruJusty	89	73	40.4	103	22.5
SubaruJusty	89	66	39.6	100	22.5
SubaruJusty4wd	89	73	39.3	103	22.5
ToyotaTercel	91	78	38.9	106	22.5
HondaCivicCRX	50	92	38.8	113	22.5
ToyotaTercel	91	78	38.2	106	22.5
FordEscort	103	90	42.2	109	25.0
HondaCivic	99	92	40.9	110	25.0
PontiacLeMans	107	74	40.7	101	25.0
IsuzuStylus	101	95	40.0	111	25.0
DodgeColt	96	81	39.3	105	25.0
GM/GeoStorm	89	95	38.8	111	25.0
HondaCivicCRX	50	92	38.4	110	25.0
HondaCivicWagon	117	92	38.4	110	25.0
HondaCivic	99	92	38.4	110	25.0
SubaruLoyale	102	90	29.5	109	25.0
VolksJettaDiesel	104	52	46.9	90	27.5
Mazda323Protege	107	103	36.3	112	27.5
FordEscortWagon	114	84	36.1	103	27.5
FordEscort	101	84	36.1	103	27.5
GM/GeoPrism	97	102	35.4	111	27.5
ToyotaCorolla	113	102	35.3	111	27.5
EagleSummit	101	81	35.1	102	27.5
NissanCentraCoupe	98	90	35.1	106	27.5
NissanCentraWagon	88	90	35.0	106	27.5
ToyotaCelica	86	102	33.2	109	30.0
ToyotaCelica	86	102	32.9	109	30.0
ToyotaCorolla	92	130	32.3	120	30.0
ChevroletCorsica	113	95	32.2	106	30.0
ChevroletBeretta	106	95	32.2	106	30.0
ToyotaCorolla	92	102	32.2	109	30.0
PontiacSunbirdConv	88	95	32.2	106	30.0
DodgeShadow	102	93	31.5	105	30.0
DodgeDaytona	99	100	31.5	108	30.0
EagleSpirit	111	100	31.4	108	30.0
FordTempo	103	98	31.4	107	30.0
ToyotaCelica	86	130	31.2	120	30.0
ToyotaCamry	101	115	33.7	109	35.0
ToyotaCamry	101	115	32.6	109	35.0
ToyotaCamry	101	115	31.3	109	35.0
ToyotaCamryWagon	124	115	31.3	109	35.0
OldsCutlassSup	113	180	30.4	133	35.0
OldsCutlassSup	113	160	28.9	125	35.0
Saab9000	124	130	28.0	115	35.0
FordMustang	92	96	28.0	102	35.0
ToyotaCamry	101	115	28.0	109	35.0
ChryslerLebaronConv	94	100	28.0	104	35.0
DodgeDynasty	115	100	28.0	105	35.0
Volvo740	111	145	27.7	120	35.0
FordThunderbird	116	120	25.6	107	40.0
ChevroletCaprice	131	140	25.3	114	40.0
LincolnContinental	123	140	23.9	114	40.0
ChryslerNewYorker	121	150	23.6	117	40.0
BuickReatta	50	165	23.6	122	40.0
OldsTrof/Toronado	114	165	23.6	122	40.0
Oldsmobile98	127	165	23.6	122	40.0
PontiacBonneville	123	165	23.6	122	40.0
LexusLS400	112	245	23.5	148	40.0
Nissan300ZX	50	280	23.4	160	40.0
Volvo760Wagon	135	162	23.4	121	40.0
Audi200QuatroWag	132	162	23.1	121	40.0
BuickElectraWagon	160	140	22.9	110	45.0
CadillacBrougham	129	140	22.9	110	45.0
CadillacBrougham	129	175	19.5	121	45.0
Mercedes500SL	50	322	18.1	165	45.0
Mercedes560SEL	115	238	17.2	140	45.0
JaguarXJSConvert	50	263	17.0	147	45.0
BMW750IL	119	295	16.7	157	45.0
Rolls-RoyceVarious	107	236	13.2	130	55.0

```
# read the table in as a data.frame
cars <- read.table("http://statacumen.com/teach/ADA1/ADA1_HW_08_F14-1.txt", header=TRUE)
```

(a) (10 pts) Plot a scatterplot of miles per gallon against horsepower, choosing the variable to plot on the x-axis so that it makes sense that it should affect the variable plotted on the y-axis (that is, x should seem to "influence" y more than y "influences" x). Compute the natural logarithm of both, log(mpg) and log(hp), and plot a second scatterplot of these log-transformed variables against each other. Which variables would be more appropriate for a straight-line regression?

(b) (10 pts) Using the more appropriate of the two pairs of variables from (a), that is original or log-transformed variables, fit a simple linear regression model.
Present and interpret the residual plots with respect to model assumptions. If the normality assumption seems to be violated, perform a normality test on the standardized residuals.
Do the residuals versus the fitted values appear random? Or is there a pattern?

(c) (5 pts) Investigate the leverages and Cook's D. Use the $3p/n$ cutoff for large leverages, and the cutoff of 1 for large Cook's D values. Interpret the leverages and Cook's D values with respect to whether any observations are having undue influence on model fit.

(d) (10 pts) Assuming the model fits well, present and interpret the ANOVA table and R^2 value.

(e) (10 pts) Present the parameter estimate table and estimated regression equation. State what the hypothesis test is related to the log(hp) line in the parameter estimate table. State the conclusion of the hypothesis test. Interpret the slope coefficient in the context of the model.

(f) (5 pts) Using the R^2 statistic and the slope of the regression line, what is the correlation between log(hp) and log(mpg)?

(30^{pts}) **2. Gas milage and automobile weight**

(a) (10 pts) Plot a scatterplot of miles per gallon against weight, compute the natural logarithm of both, log(mpg) and log(wt), and plot a second scatterplot of these log-transformed variables against each other. Which variables would be more appropriate for a straight-line regression?

(b) (10 pts) Using the more appropriate of the two pairs of variables from (a), that is original or log-transformed variables, fit a simple linear regression model.
Present and interpret the residual plots with respect to model assumptions. If the normality assumption seems to be violated, perform a normality test on the standardized residuals.
Do the residuals versus the fitted values appear random? Or is there a pattern?

(c) (5 pts) Use the $3p/n$ cutoff for large leverages, and the cutoff of 1 for large Cook's D values. Interpret the leverages and Cook's D values with respect to whether any observations are having undue influence on model fit.

(d) (5 pts) The model doesn't fit well (as we learned from the residual plots above). Therefore, it doesn't make sense to present and interpret the parameter estimates.
Without doing anything more, suggest one or two things we *could* do to find a better relationship between log(mpg) and log(wt).