Advanced Data Analysis 1, Stat 427/527 Homework 8 Name: Answer Key
Fall 2014 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.

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

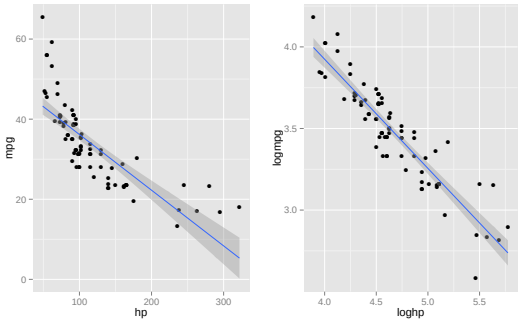
```
# read the table in as a data.frame
cars <- read.table("http://statacumen.com/teach/ADA1/ADA1_HW_08_F14-1.txt", header=TRUE)
```

(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?

Solution: It is more likely that horsepower (hp) influcences miles per gallon (mpg). The plots indicate substantial curvature in the original units. The plot of the variables on the log scale are much more linear, therefore we will perform the analysis using the log-transformed variables.

```
cars$loghp  <- log(cars$hp)
cars$logmpg <- log(cars$mpg)

library(ggplot2)
p1 <- ggplot(cars, aes(x = hp, y = mpg))
p1 <- p1 + geom_point()
p1 <- p1 + geom_smooth(method = lm, se = TRUE)
#print(p1)

p2 <- ggplot(cars, aes(x = loghp, y = logmpg))
p2 <- p2 + geom_point()
p2 <- p2 + geom_smooth(method = lm, se = TRUE)
#print(p2)

library(gridExtra)
grid.arrange(p1, p2, nrow=1)
```



(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?

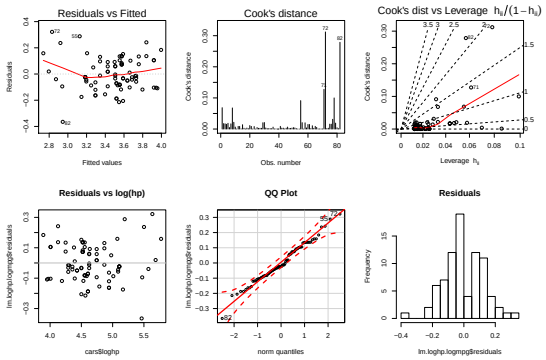
Solution: The normal probability plot (bottom-center) indicates the residuals look very close to normal, and indicates one possible outlier in the left tail, with this shape supported by the histogram (bottom-left).
The residuals versus fitted value are basically random, though there may be a very slight upward curvature and more variability for smaller fitted values than larger fitted values, neither of these are large enough to cause concern.
Because the cars are sorted by wt, the observation order is meaningless.

```
# fit model
lm.loghp.logmpg <- lm(logmpg ~ loghp, data = cars)

# plot diagnistics
par(mfrow=c(2,3))
plot(lm.loghp.logmpg, which = c(1,4,6))

# residuals vs loghp
plot(cars$loghp, lm.loghp.logmpg$residuals, main="Residuals vs log(hp)")
  # horizontal line at zero
  abline(h = 0, col = "gray75")
```

```
# Normality of Residuals
library(car)
qqPlot(lm.loghp.logmpg$residuals, las = 1, id.n = 3, main="QQ Plot")
```

```
## 82 72 55
##  1 82 81
```

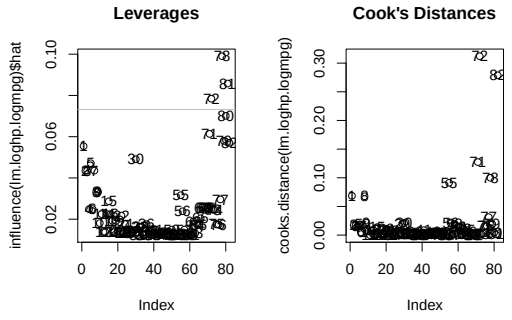
```
# # residuals vs order of data
# plot(lm.loghp.logmpg$residuals, main="Residuals vs Order of data")
#   # horizontal line at zero
#   abline(h = 0, col = "gray75")
hist(lm.loghp.logmpg$residuals, breaks=15, main="Residuals")
```



(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.

Solution: The leverages cutoff is $3p/n = 3(2)/82 = 0.073$, observations exceeding this cutoff are 78, 81, and 72. Observation 72 has the largest Cook's D, but it is much less than 1 (it is 0.31). Therefore, all model assumptions appear to be satisfied.

```
# plot diagnistics
par(mfrow=c(1,2))

plot(influence(lm.loghp.logmpg)$hat, main="Leverages")
text(1:nrow(cars), influence(lm.loghp.logmpg)$hat, label=paste(1:nrow(cars)))
  # horizontal line at zero
  abline(h = 3*2/82, col = "gray75")

plot(cooks.distance(lm.loghp.logmpg), main="Cook's Distances")
text(1:nrow(cars), cooks.distance(lm.loghp.logmpg), label=paste(1:nrow(cars)))
  # horizontal line at zero
  abline(h = 1, col = "gray75")
```



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

Solution: The extremely large F-statistic (409.68) and tiny p-value (< 0.0005) indicates the model is highly significant. That is, the regression slope is significantly different from zero.
The $R^2 = 0.837$, indicating that the regression model explains 83.7% of the observed variability in the mpg response y. This is pretty high, giving the model high predictive ability of mpg for a car with a given hp.

```
# ANOVA table of the simple linear regression fit
anova(lm.loghp.logmpg)
```

```
## Analysis of Variance Table
##
## Response: logmpg
##           Df Sum Sq Mean Sq F value Pr(>F)
## loghp      1   6.28    6.28     410 <2e-16 ***
## Residuals 80   1.23    0.02
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(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.

Solution: Let β_1 represent the slope of the regression line. We're testing whether the slope of the regression line, β_1 is different from 0. That is $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$. The p-value< 0.0001, therefore we reject the null in favor of the alternative, concluding that there is sufficient evidence that the slope is different from zero. Furthermore, the slope is clearly negative.
The slope -0.669 indicates that for each increase of 1 log(hp), the expected decrease in log(mpg) is 0.669.

```
# the last row of output has the F-stat and p-value of the ANOVA table
summary(lm.loghp.logmpg)
```

```
##
## Call:
## lm(formula = logmpg ~ loghp, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3650 -0.0818 -0.0233  0.0920  0.3218
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.599      0.155    42.6   <2e-16 ***
## loghp         -0.669      0.033   -20.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.124 on 80 degrees of freedom
## Multiple R-squared:  0.837,Adjusted R-squared:  0.835
## F-statistic:  410 on 1 and 80 DF,  p-value: <2e-16
```

(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)?

Solution: $r = \text{sign(slope)}\sqrt{R^2} = -\sqrt{0.837} = -0.915$. This is a very high correlation.

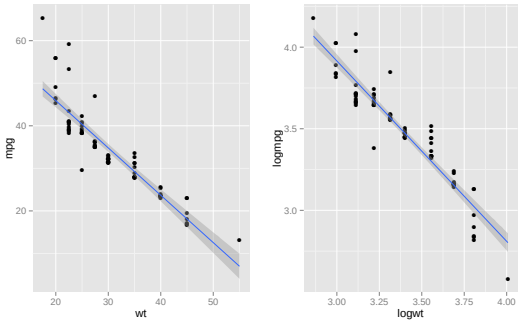
(30 pts) **2. Gas milage and automobile weight**

(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?

Solution: It is more likely that weight (wt) influcences miles per gallon (mpg). The plots indicate substantial curvature in the original units. The plot of the variables on the log scale are much more

linear, therefore we will perform the analysis using the log-transformed variables.

```
cars$logwt  <- log(cars$wt)

library(ggplot2)
p1 <- ggplot(cars, aes(x = wt, y = mpg))
p1 <- p1 + geom_point()
p1 <- p1 + geom_smooth(method = lm, se = TRUE)
#print(p1)

p2 <- ggplot(cars, aes(x = logwt, y = logmpg))
p2 <- p2 + geom_point()
p2 <- p2 + geom_smooth(method = lm, se = TRUE)
#print(p2)

library(gridExtra)
grid.arrange(p1, p2, nrow=1)
```



(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?

Solution: The normal probability plot (bottom-center) indicates the residuals look different from normal, peaky with long tails, with this shape supported by the histogram (bottom-right). The normality test of the standardized residuals indicates sufficient evidence that the residuals are not normal (Anderson-Darling p-value= 0.0002). This is a violation of the normality assumption.
The residuals versus fitted value are basically random, though there is a slight downward curvature. This curvature indicates that a straight line does not fit these data well.
Because the cars are sorted by wt, the observation order is meaningless.

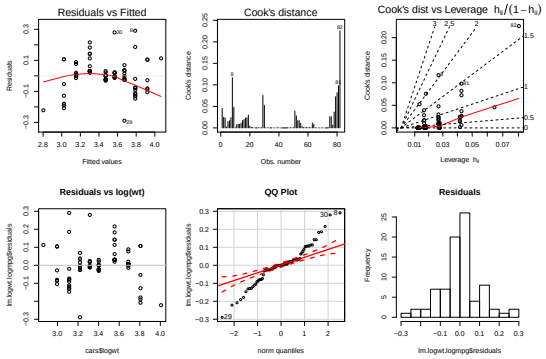
```
# fit model
lm.logwt.logmpg <- lm(logmpg ~ logwt, data = cars)

# plot diagnistics
par(mfrow=c(2,3))
plot(lm.logwt.logmpg, which = c(1,4,6))

# residuals vs logwt
plot(cars$logwt, lm.logwt.logmpg$residuals, main="Residuals vs log(wt)")
  # horizontal line at zero
  abline(h = 0, col = "gray75")

# Normality of Residuals
library(car)
qqPlot(lm.logwt.logmpg$residuals, las = 1, id.n = 3, main="QQ Plot")
```

```
##  8 29 30
## 82  1 81
```

```
# # residuals vs order of data
```

```
# plot(lm.logwt.logmpg$residuals, main="Residuals vs Order of data")
#   # horizontal line at zero
#   abline(h = 0, col = "gray75")
hist(lm.logwt.logmpg$residuals, breaks=15, main="Residuals")

# normality test
library(nortest)
ad.test(lm.logwt.logmpg$residuals)
```

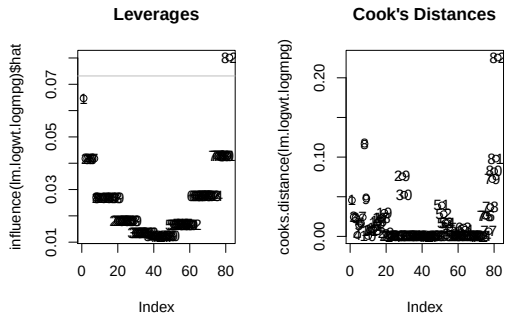
```
##
##  Anderson-Darling normality test
##
## data:  lm.logwt.logmpg$residuals
## A = 1.744, p-value = 0.0001669
```



(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.

Solution: The leverages cutoff is $3p/n = 3(2)/82 = 0.073$, the one observation exceeding this cutoff is 82 (the heaviest automobile). Observation 82 also has the largest Cook's D, but it is much less than 1 (it is 0.225). Therefore, the poor model fit (from the residuals) is not a result of influential observations.

```
# plot diagnostics
par(mfrow=c(1,2))

plot(influence(lm.logwt.logmpg)$hat, main="Leverages")
text(1:nrow(cars), influence(lm.logwt.logmpg)$hat, label=paste(1:nrow(cars)))
  # horizontal line at zero
  abline(h = 3*2/82, col = "gray75")

plot(cooks.distance(lm.logwt.logmpg), main="Cook's Distances")
text(1:nrow(cars), cooks.distance(lm.logwt.logmpg), label=paste(1:nrow(cars)))
  # horizontal line at zero
  abline(h = 1, col = "gray75")
```



(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).

Solution: Though beyond the scope of this class, if we fit a quadratic model by including a squared term in the model, $\beta_2(\log(wt))^2$, then we could capture the curvature that we see in the residuals. It's possible this also improves the normality of the residuals, since they depend on model fit.