DSSH 6301 - HW 10 Solutions

Read in data and define some useful functions.

```
data <- read.csv("anes_2008tr.csv")
head(data)
```

```
##   age race_white gender_male education income ideology_con partyid_rep
## 1  35          0           1         5      4            4           5
## 2  58          1           0         3      3            6           1
## 3  39          1           0         6      4            6           6
## 4  50          1           1         3      4            4           4
## 5  72          1           1         6      4            6           7
## 6  71          1           0         5      3            6           6
##   vote_rep voted
## 1        1     1
## 2       NA     0
## 3        1     1
## 4       NA     0
## 5        1     1
## 6       NA     0
```

```
invlogit <- function(x) exp(x) / (1+exp(x))
library(stargazer)
```

Problem 1

```
lr <- glm(vote_rep ~ age + race_white + income + ideology_con, data=data,
          family="binomial")
stargazer(lr, align=TRUE, no.space=TRUE, omit.stat=c("LL","ser","f"), header=FALSE)
```

```
## SEE TABLE 1
```

Part a

```
# Get the mean values for the inputs to the model.
means <- sapply(lr$model, mean)
means
```

```
##     vote_rep          age   race_white       income ideology_con
##    0.3339831   48.0194932    0.5204678    2.8206628    4.0708252
```

```
n <- length(means)

lr$coefficients
```

Table 1:

	Dependent variable:
	vote_rep
age	0.005
	(0.004)
race_white	2.437***
	(0.168)
income	0.405***
	(0.075)
ideology_con	1.042***
	(0.066)
Constant	−8.077***
	(0.463)
Observations	1,539
Akaike Inf. Crit.	1,180.284
Note:	*p<0.1; **p<0.05; ***p<0.01

```
##  (Intercept)          age   race_white       income ideology_con
## -8.076556122  0.004737975  2.436605297  0.404586967  1.041976990
```

```
c(1, means[2], 1, means[4:n])
```

```
##                      age                   income ideology_con
##     1.000000    48.019493     1.000000     2.820663     4.070825
```

```
# Plug in our x values and use the inverse logit function.
# Note the use of vector multiplication to speed things along.
# For binary variables, we use the modal value rather than the mean value,
#   since it doesn't make as much sense to talk about someone of intermediate
#   race (for these purposes).
x <- sum(lr$coefficients * c(1, means[2], 1, means[4:n]))
invlogit(x)
```

```
## [1] 0.492619
```

The probability that a white person votes republican, holding other variables at their mean value, is about 0.49.

Part b

```
lr$coefficients
```

```
##  (Intercept)          age   race_white       income ideology_con
## -8.076556122  0.004737975  2.436605297  0.404586967  1.041976990
```

```r
# Black
x1 <- sum(lr$coefficients * c(1, means[2], 0, means[4:n]))

# White
x2 <- sum(lr$coefficients * c(1, means[2], 1, means[4:n]))

invlogit(x2) - invlogit(x1)
```

```
## [1] 0.4143521
```

When a person switches from black to white, there is about a 41.4% increased chance of voting Republican.

Part c

```r
exp(lr$coefficients["race_white"])
```

```
## race_white
##   11.43416
```

Shifting from black to white increases the odds by a factor of 11.43.

Part d

```r
# Age term set to mean.
x1 <- lr$coefficients * c(1, means[2:length(means)])

x2 <- x1

# Increase age term by 50 from mean.
x2["age"] <- x2["age"] + 50*lr$coefficients["age"]

x3 <- x1

# Increase income bracket by 1 from mean.
x3["income"] <- x3["income"] + 1*lr$coefficients["income"]

# Difference by increasing age by 50
invlogit(sum(x2)) - invlogit(sum(x1))
```

```
## [1] 0.04482752
```

```r
# Difference by increasing income by 1 bracket
invlogit(sum(x3)) - invlogit(sum(x1))
```

```
## [1] 0.07960667
```

There is a larger positive difference in the probability of voting republican by increasing the income bracket by 1.

Part e

```
lr_all <- glm(vote_rep ~ age + race_white + income + ideology_con + gender_male +
              education + ideology_con + partyid_rep, data=data,
              family="binomial")

stargazer(lr_all, align=TRUE, no.space=TRUE, omit.stat=c("LL","ser","f"), header=FALSE)
```

Table 2:

	Dependent variable:
	vote_rep
age	0.015***
	(0.005)
race_white	1.627***
	(0.203)
income	0.250***
	(0.096)
ideology_con	0.511***
	(0.086)
gender_male	−0.140
	(0.188)
education	0.019
	(0.064)
partyid_rep	0.894***
	(0.057)
Constant	−8.646***
	(0.604)
Observations	1,539
Akaike Inf. Crit.	830.593
Note:	*p<0.1; **p<0.05; ***p<0.01

```
## SEE TABLE 2 and compare with TABLE 1
```

```
lr$coefficients
```

```
##  (Intercept)          age   race_white       income ideology_con
## -8.076556122  0.004737975  2.436605297  0.404586967  1.041976990
```

```
lr_all$coefficients[names(lr$coefficients)]
```

```
##  (Intercept)          age   race_white       income ideology_con
##  -8.64609889   0.01498968   1.62656986   0.24967983   0.51075810
```

```
lr$coefficients / lr_all$coefficients[names(lr$coefficients)]
```

```
##  (Intercept)          age   race_white       income ideology_con
##    0.9341272    0.3160825    1.4980022    1.6204231    2.0400597
```

Age becomes significant when all variables are added. You can also see there are differences in the β estimates themselves. Note how party ID is also quite significant. One reason why age may now be significant where it was not before is that age both causes people to be more conservative (more likely to vote Republican), and is also correlated with being a Democrat (older people are more Democratic because being a Democrat was more popular decades ago), which of course is correlated with voting Democrat. But once you control for party ID, the second effect is controlled for, so the effect of age is no longer pulled in two directions, revealing its true effect: to make one more likely to vote Republican. Note also that ideology has a strong effect independent of party ID – there are still conservative Democrats out there who vote Republican.

Problem 2

```
require(forecast)
set.seed(10)
y <- 1
for (i in 2:1000) y[i] <- 0.8*y[i-1] + 0.2*rnorm(1)

auto.arima(y)
```

```
## Series: y
## ARIMA(1,0,0) with zero mean
##
## Coefficients:
##          ar1
##       0.8353
## s.e.  0.0175
##
## sigma^2 estimated as 0.03951:  log likelihood=196.1
## AIC=-388.19   AICc=-388.18   BIC=-378.38
```

This is an Arima(1,0,0). That is there is an AR(1) component with no differences and no MA component. This can be interpreted as there is no long term history effect, while the immediate past values influence the current value, and there is no noticeable trend in y. Note the coefficient estimates. The coefficient on ar1 is about 0.8. This is the relative amounts of the current value that depends on the previous value (80%). Note that you may have gotten a somewhat different answer, since auto.arima sometimes gets it wrong due to the randomness of the simulated data.

Problem 3

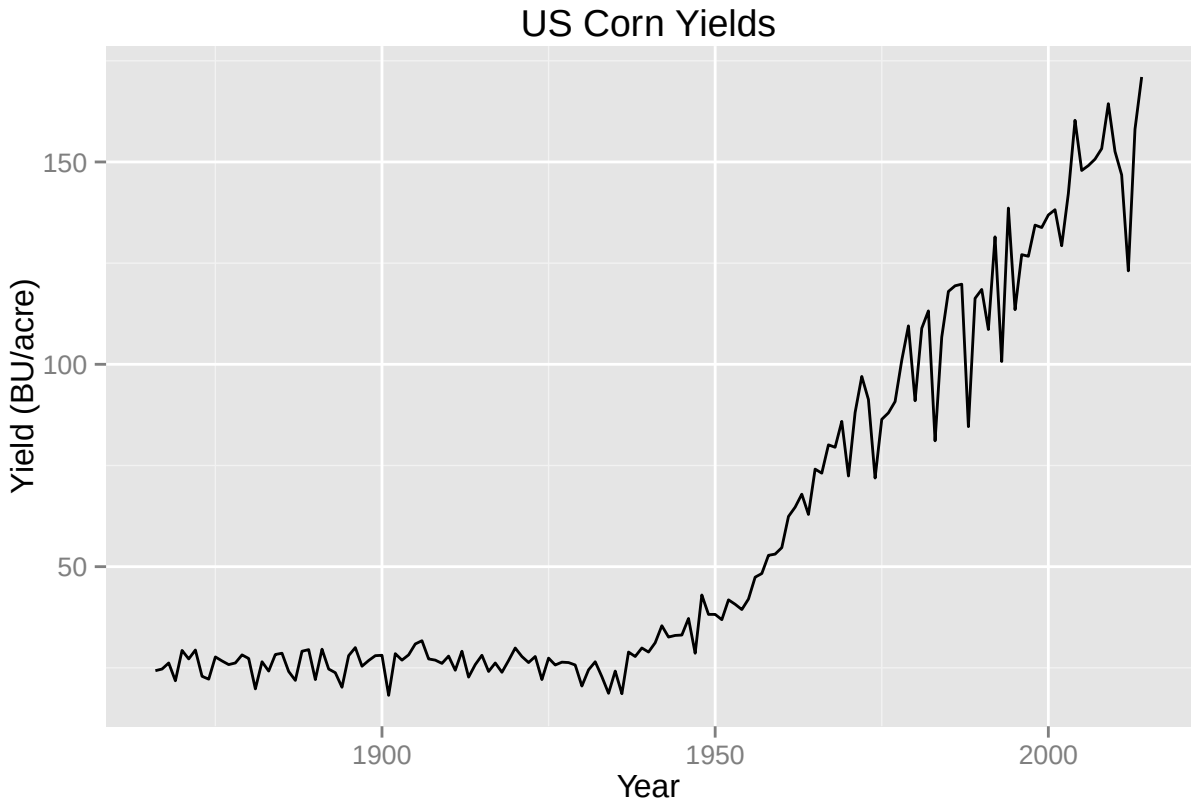
The USDA aggregates data for crops in the US going back to the 1800s.

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
data <- read.csv("HW10_US_corn_yields.csv")
names(data)[names(data)=="CORN..GRAIN...YIELD..MEASURED.IN.BU...ACRE.....b.VALUE..b."] = "yield"

ggplot(data=data, aes(x=data$Year, y=data$yield)) + geom_line() +
  xlab("Year") + ylab("Yield (BU/acre)") + ggtitle("US Corn Yields")
```




```
# Pre-1950, the agricultural revolution in the US occured. Results before this
# time are not comparable to those after.
data <- data[data$Year > 1950,]

auto.arima(data$yield[order(data$Year)])
```

```
## Series: data$yield[order(data$Year)]
## ARIMA(4,1,0) with drift
##
## Coefficients:
##           ar1      ar2      ar3      ar4   drift
##       -0.8950  -0.7790  -0.6536  -0.4417  1.8721
## s.e.   0.1174   0.1495   0.1489   0.1196  0.3512
##
## sigma^2 estimated as 104.4:  log likelihood=-232.42
## AIC=476.84   AICc=478.34   BIC=489.7
```

We have a ARIMA(4,1,0) here. There is a AR(4) component, with yields being dependent on the 4 immediate past values, and a first-differencing. There is no long term dependence here, with MA(0).