STATS 500, HOMEWORK #4, due Wednesday, Oct, 13, 3PM

```
set.seed(48109711)
n<-200; x<-rnorm(2*n); x1<-x[1:n]; x2<-x[(n+1):(2*n)]
x3<-.7*x1+.9*x2+rnorm(n)*.1
beta<-c(1, 1, 1, 2)
y1<-beta[1]+beta[2]*x1+beta[3]*x2+rnorm(n)*.3
y2<-beta[1]+beta[2]*x1+beta[3]*x2+beta[4]*x3+rnorm(n)*.3
dat<-data.frame(y1=y1,y2=y2,x1=x1,x2=x2,x3=x3)
outx12 <- lm(y1~x1+x2, data=dat)
summary(outx12)
outx123 <- lm(y2~x1+x2+x3, data=dat)
library(ellipse)
## Plot the confidence region
par(mfrow=c(1,2))
plot(ellipse(outx12, c("x1", "x2")), type="l")
points(coef(outx12)[2], coef(outx12)[3], pch=18)
plot(ellipse(outx123, c("x1", "x2")), type="l")
points(coef(outx123)[2], coef(outx123)[3], pch=18)
D1<- model.matrix(outx12);D2<- model.matrix(outx123)
cov(x1,x2);
V1<-solve(t(D1)%*%D1); V2<-solve(t(D2)%*%D2);
round(V1[2:3,2:3],4); round(V2[2:3,2:3],4);
```

1. In the codes given above, we generate $n = 200$ $x1$, $x2$ and $x3$, and $x1$, $x2$ are independent of each other. In model A, we generate $y1$ with the predictors $x1$ and $x2$. In model B, we generate $y2$ with all three predictors. In the two plots we generate, we obtain the 95% confidence regions for the coefficients, β_{x1}, β_{x2} from each of the two models. Note that even you generate $x1$ and $x2$ independently, the sample correlation still may not be 0, but it tends to be small. Consequently, the major axis and minor axis of the ellipse are roughly parallel with the x-axis and y-axis corresponding to $\hat{\beta}_{x1}$, $\hat{\beta}_{x2}$. In model B, you are using the same $x1$ and $x2$ and you still consider the 95% confidence ellipse for β_{x1}, β_{x2}, do you expect the major axis and minor axis of the ellipse are roughly parallel with the x-axis and y-axis corresponding to $\hat{\beta}_{x1}$, $\hat{\beta}_{x2}$? Please briefly explain your answer using the output you obtain. Basically, in a multiple regression, unless the predictors are completely orthogonal to each other, the estimated coefficients and the corresponding inference are adjusted to the existence of other predictors.

2. Using the `teengamb` dataset, fit a model to predict gambling expenditure from all other available variables. Use "help("teengamb") to see more descriptions about this data set. You can start with a linear regression with the predictors `sex, status, income` and `verbal`. Consider the following two Y:

   - Y_{org} = teengamb$gamble.
   - $Y_{\log} = log(teengamb\$gamble + 1)$.

Perform regression diagnostics on these models and **compares the answers to the following questions based on the two models.** Display **only** those plots that are relevant to the questions below. Present your diagnostics in a logical order.

(a) Check the existence of patterns in the residual against fitting values. Are the residuals centering at 0 in all areas of fitted values?

(b) Check the constant variance assumption for the errors.

(c) Check the normality assumption. Do you need to check normality of both models (one only check normality when the residuals center at zero and roughly share a constant variance.)

3. Now focus on the regression model with the response $Y_{log} = log(teengamb\$gamble+1)$. What would be the 95% prediction interval for the gambling expenditure per year for a boy with status = 40, income =2, verbal score = 6? (Hint: you should transform the prediction interval back to the original scale).