DISCUSSION 4

- Outliers and Influential points
  1. Regression outliers: unusual Y values (standardized residual $\geq 2$)
  2. High leverage points: unusual X values (leverage $>.5$ or $2p/n$, consider log transformation on X)
  3. Influential points: points that significantly effect the fit (Cook’s distance $> 1$ or $4/(n-p)$)

State Public Expenditures Data

Outlier data Read the dataset from http://lib.stat.cmu.edu/DASL/Datafiles/pubexpendat.html. It contains the information of public expenditures and associated state demographic and economic characteristics during the year of 1960. Exclude the variable STATE in the following analysis.

1. Draw the scatter plot and fit the simple linear model between EX and ECAB in this dataset. Is there any points special to you?
2. Try a transformation to the dataset. Will it be better than the previous one? How to interpret the coefficients?
3. Check the regression diagnostics plots of the two models above. Learn what these plots really say.
4. How to assess the significance of the MET effect using t-tests? Consider 4 different models and 4 separate tests: with and without ECAB, with and without WEST.
5. Use AIC to select one model among all possible models (no interaction effects). First do a stepwise selection from the most simple model, then do a stepwise selection from the full model. Do we arrive at the same model? If not, do the two models selected by these two searches form nested models? If so, perform an F-test to select one of these two models. If not, select one of these two models by other means, e.g. AIC, adjusted $R^2$.
6. Use ANOVA F-tests to select a model among all possible models (no interaction effects). First do a forward selection from the most simple model, then do a backward elimination from the full model. Do we arrive at the model? If not, do the two models selected by these two searches form nested models? If so, perform an F-test to select one of these two models. If not, select one of these two models by other means, e.g. AIC, adjusted $R^2$. 
Solutions

```r
expend = read.table("expend.txt", header=T)
expend$WEST = factor(expend$WEST, labels=c("eastern","western"))
str(expend)

plot(EX ~ ECAB, expend); with(expend, identify(EX ~ ECAB))

fit1=lm(EX ~ ECAB, expend)
abline(fit1, lty=1)
abline(lm(EX ~ ECAB, expend[-47,]), lty=2)
plot(fit1)

plot(EX ~ log(ECAB), expend)
fit2=lm(EX ~ log(ECAB), expend)
abline(fit2, lty=1)
abline(lm(EX ~ log(ECAB), expend[-47,]), lty=2)
plot(fit2)

summary(lm(EX ~ MET, expend))
summary(lm(EX ~ MET + ECAB, expend))
summary(lm(EX ~ MET + WEST, expend))
summary(lm(EX ~ MET + ECAB + WEST, expend))

basic = lm(EX ~ 1, expend)
full = lm(EX ~ ECAB + MET + GROW + YOUNG + OLD + WEST, expend)
fullfo = formula(full)

step(basic, scope = fullfo, test="F")
fitforward=lm(EX ~ ECAB + WEST, data = expend)
step(full, scope = fullfo, test="F")
fitbackward=lm(EX ~ ECAB + MET + GROW + WEST, data = expend)

anova(fitforward, fitbackward)
summary(fitforward)$adj.r.square; summary(fitbackward)$adj.r.square
AIC(fitforward); AIC(fitbackward)

fit=basic; add1(fit, scope = fullfo, test="F")
fit=update(basic, ~ . + ECAB); add1(fit, scope = fullfo, test="F")
fit=update(fit, ~ . + WEST); add1(fit, scope = fullfo, test="F")

fit=full; drop1(fit, test="F")
fit = update(fit, ~ . - OLD); drop1(fit, test="F")
fit = update(fit, ~ . - YOUNG); drop1(fit, test="F")
fit = update(fit, ~ . - GROW); drop1(fit, test="F")
fit = update(fit, ~ . - MET); drop1(fit, test="F")
```