Section 1.5 Outliers and Influential Points

Load needed packages.

```
library(mosaic)
library(Stat2Data)
```

EXAMPLE 1.10 Olympic long jump

Create dataframe for **LongJumpOlympics2016** and look at the structure of the data.

```
data(LongJumpOlympics2016)
str(LongJumpOlympics2016)
```

```
## 'data.frame':    28 obs. of  2 variables:
##  $ Year: int  1900 1904 1906 1908 1912 1920 1924 1928 1932 1936 ...
##  $ Gold: num  7.18 7.34 7.2 7.48 7.6 ...
```

Fit a model to predict Gold (winning distance) using Year.

```
m1=lm(Gold~Year,data=LongJumpOlympics2016)
summary(m1)
```

```
##
## Call:
## lm(formula = Gold ~ Year, data = LongJumpOlympics2016)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39610 -0.15495 -0.00137  0.11606  0.75349
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -16.470194   2.666282  -6.177 1.56e-06 ***
## Year          0.012508   0.001361   9.191 1.19e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2595 on 26 degrees of freedom
## Multiple R-squared:  0.7646, Adjusted R-squared:  0.7556
## F-statistic: 84.47 on 1 and 26 DF,  p-value: 1.192e-09
```

FIGURE 1.22 Gold-medal-winning distances (m) for the men's Olympic long jump, 1900-2016

```
plot(Gold~Year,data=LongJumpOlympics2016)
abline(m1,lwd=2,col="darkblue")
```

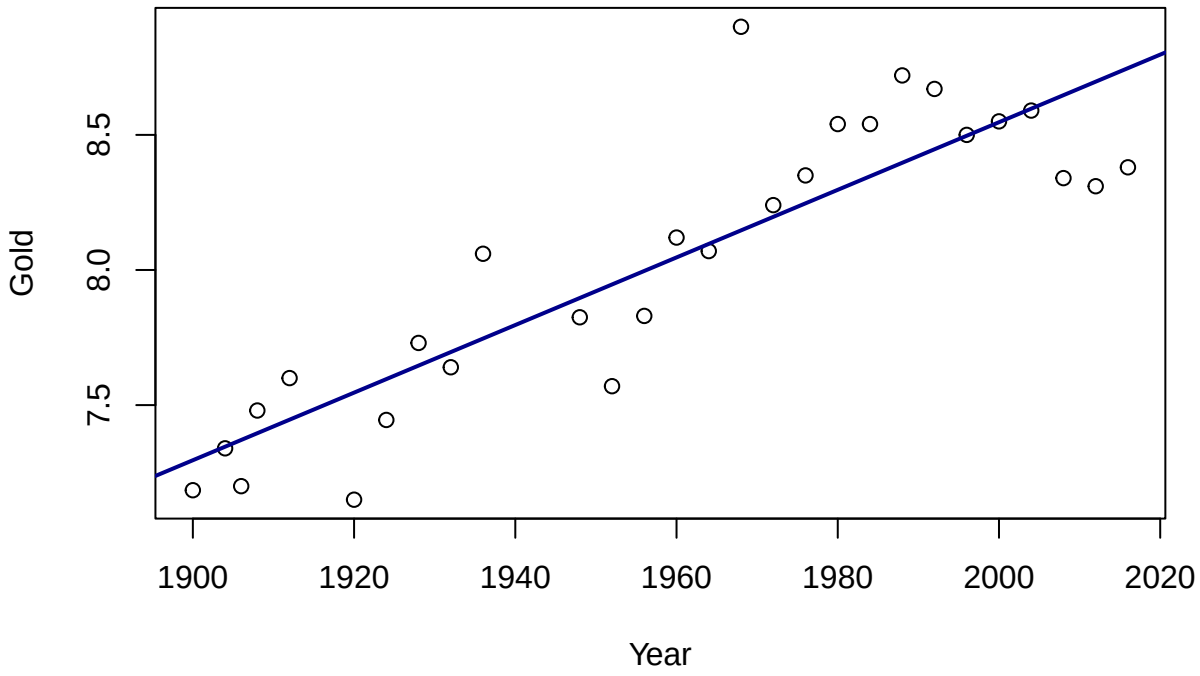


FIGURE 1.23 Residual plot for long jump model

```
plot(m1$residuals~m1$fitted.values,xlab="Predicted",ylab="Residuals")
abline(h=0)
```

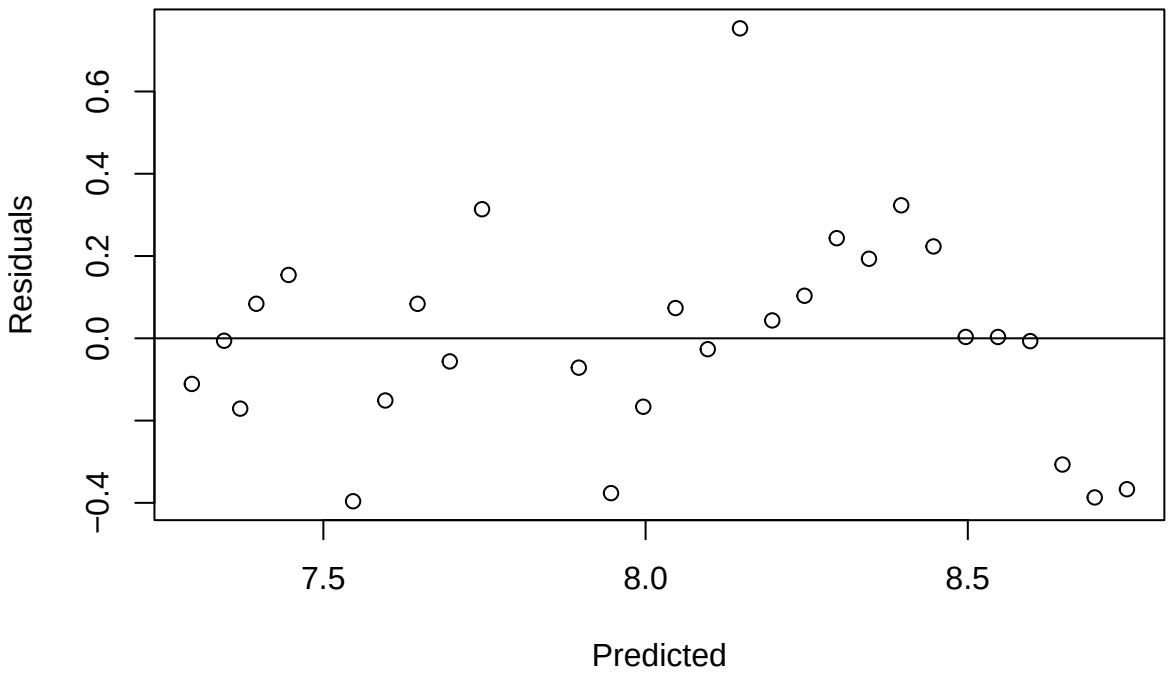
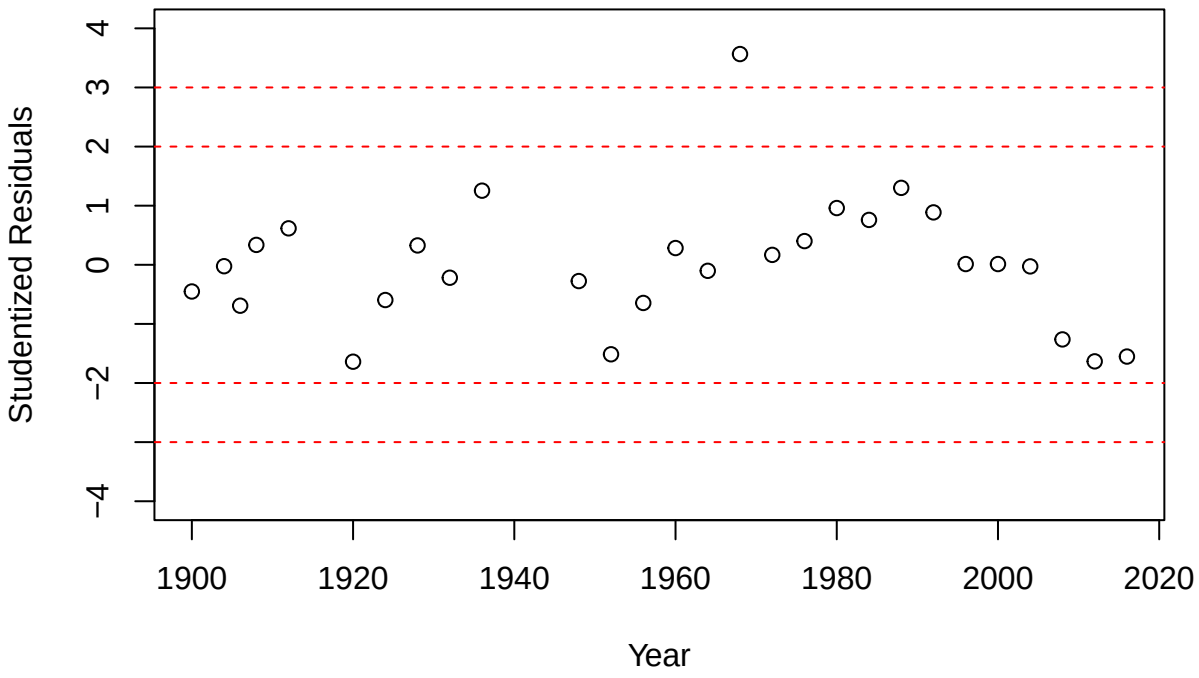


FIGURE 1.24 Studentized residuals for the long jump model

```
studs=rstudent(m1)
plot(studs~LongJumpOlympics2016$Year, xlab="Year",ylab="Studentized Residuals",ylim=c(-4,4),yaxp=c(-4,4
abline(h=c(-3,-2,2,3),lty=2,col="red")
```



Note: For standarized residuals, use rstandard(m1) in place of rstudent(m1).

Bob Beaman is the 16th observation in the dataframe. Here is an easy way to get his studentized residual.

```
studs[16]
```

```
##       16
## 3.565083
```

EXAMPLE 1.11 Butterfly ballot

Create a data frame for PalmBeach.

```
data("PalmBeach")
str(PalmBeach)
```

```
## 'data.frame':    67 obs. of  3 variables:
##  $ County  : Factor w/ 67 levels "ALACHUA","BAKER",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Buchanan: int  262 73 248 65 570 789 90 182 270 186 ...
##  $ Bush    : int  34062 5610 38637 5413 115185 177279 2873 35419 29744 41745 ...
```

Fit a model to predict Buchanan votes using Bush votes.

```
regall=lm(Buchanan~Bush,data=PalmBeach)
summary(regall)
```

```
##
## Call:
## lm(formula = Buchanan ~ Bush, data = PalmBeach)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -907.50  -46.10  -29.19   12.26 2610.19
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.529e+01  5.448e+01   0.831    0.409
## Bush        4.917e-03  7.644e-04   6.432 1.73e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 353.9 on 65 degrees of freedom
## Multiple R-squared:  0.3889, Adjusted R-squared:  0.3795
## F-statistic: 41.37 on 1 and 65 DF,  p-value: 1.727e-08
```

FIGURE 1.25 2000 presidential election totals in Florida counties

```
plot(Buchanan~Bush,data=PalmBeach)
abline(regall,lwd=2,col="darkblue")
```

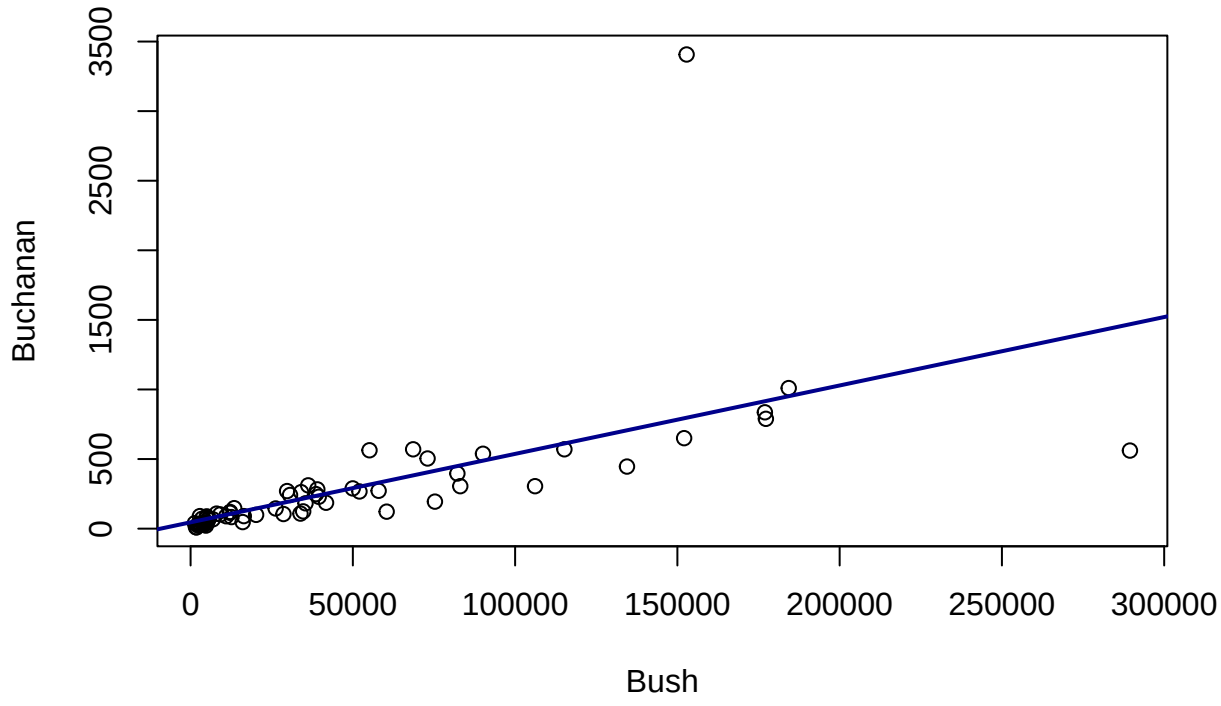
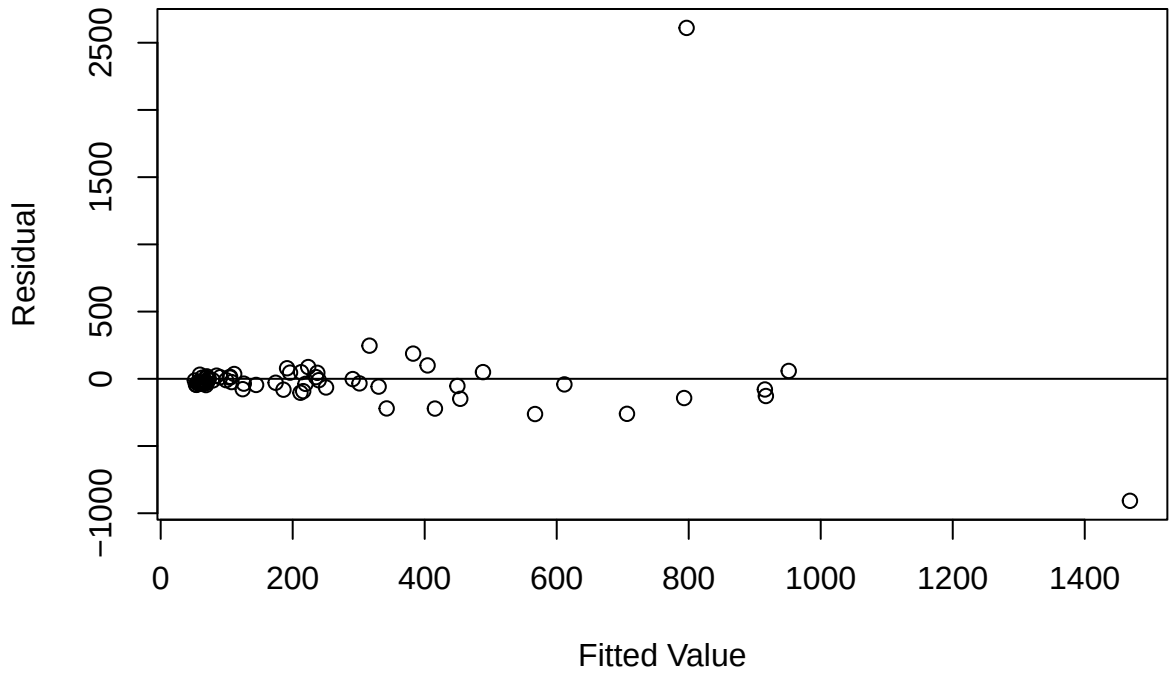


FIGURE 1.26 Residual plot for butterfly ballot data

```
plot(regall$residuals~regall$fitted.values,xlab="Fitted Value",ylab="Residual")
abline(h=0)
```



Palm Beach County is observation 50, so we can find the standardized and studentized residuals as follows.

```
sresid=rstandard(regall)
studresid=rstudent(regall)
sresid[50]
```

```
##       50
## 7.651072
```

```
studresid[50]
```

```
##       50
## 24.08014
```

Dade County is observation 13, so we can find the residual and standardized residual as well.

```
regall$residuals[13]
```

```
##        13
## -907.4953
```

```
sresid[13]
```

```
##       13
## -3.05918
```

FIGURE 1.27 Regression lines with and without Palm Beach

Remove Palm Beach county using the subset command.

```
NoPalmBeach <- PalmBeach[-50,]
```

Fit the regression to the model without Palm Beach.

```
regnoPB=lm(Buchanan~Bush,data=NoPalmBeach)
summary(regnoPB)
```

```
##
## Call:
## lm(formula = Buchanan ~ Bush, data = NoPalmBeach)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -512.43  -47.97  -17.09   41.78  305.45
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.557e+01  1.733e+01   3.784 0.000343 ***
## Bush        3.482e-03  2.501e-04  13.923  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 112.5 on 64 degrees of freedom
## Multiple R-squared:  0.7518, Adjusted R-squared:  0.7479
## F-statistic: 193.8 on 1 and 64 DF,  p-value: < 2.2e-16
```

FIGURE 1.27 Regression lines with and without Palm Beach

```
plot(Buchanan~Bush,data=PalmBeach)
abline(regall,lwd=2,col="darkblue")
abline(regnoPB,lty=2, lwd=2, col="darkblue")
```

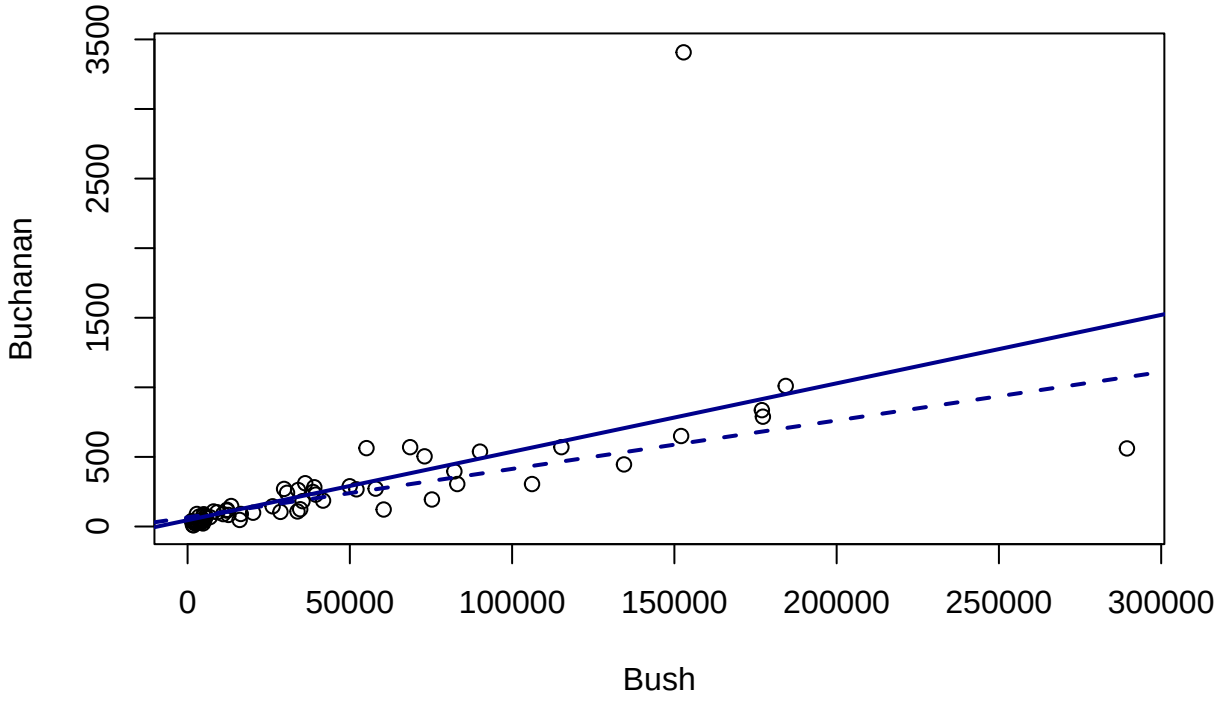
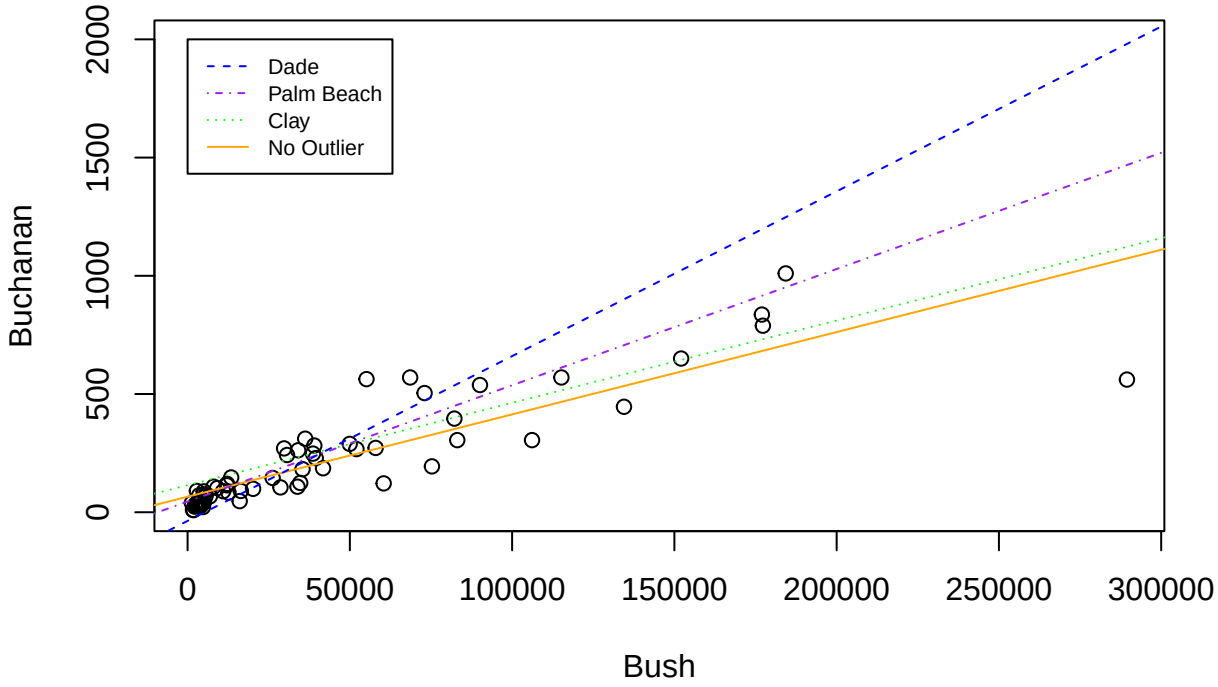


FIGURE 1.28 Regression lines with an outlier of 3407 "moved" to different counties

This one is fairly involved and requires working with up four different dataframes.

```
#make the initial plot without Palm Beach
plot(Buchanan~Bush,data=NoPalmBeach,ylim=c(0,2000))
#plot regression line with no outlier
abline(regnoPB,col="orange")
#plot regession line with the outlier in Palm Beach
abline(regall,lty=4,col="purple")
#put the outlier in Clay (County #10 )
InClay=NoPalmBeach
InClay$Buchanan[10]=3407
regClay=lm(Buchanan~Bush,data=InClay)
#plot regession line with the outlier in Clay
abline(regClay,lty=3,col="green")
#put the outlier in Dade (County #13)
InDade=NoPalmBeach
InDade$Buchanan[13]=3407
regDade=lm(Buchanan~Bush,data=InDade)
#plot regession line with the outlier in Dade
abline(regDade,col="blue",lty=2)
legend(0,2000,legend=c("Dade","Palm Beach", "Clay", "No Outlier"),
       col=c("blue","purple","green","orange"),lty=c(2,4,3,1),cex=0.7)
```


Dade
Palm Beach
Clay
No Outlier
Buchanan
0
500
1000
1500
2000
0
50000
100000
150000
200000
250000
300000
Bush