STAT 425 - Homework #3

PROBLEM 1

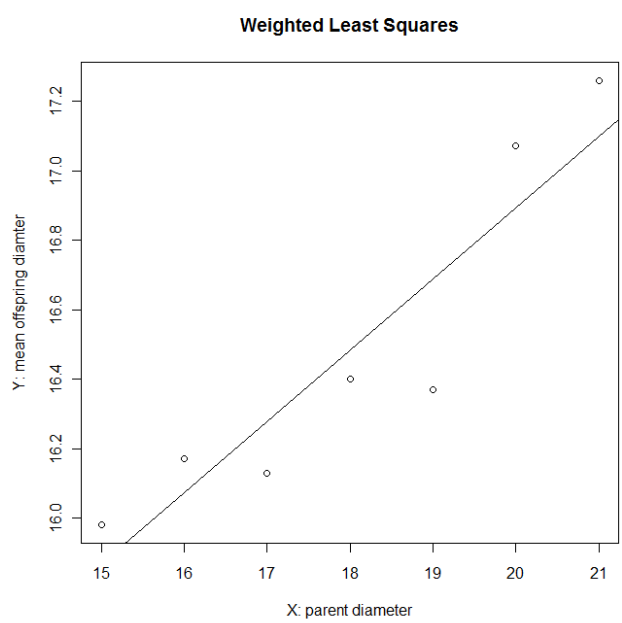
a) Weighted Regression of Y on X

- Weighted Regression

```
> X=c(21,20,19,18,17,16,15)
> Y=c(17.26,17.07,16.37,16.40,16.13,16.17,15.98)
> SD=c(1.988,1.938,1.896,2.037,1.654,1.594,1.763)
>
> g= lm(Y ~ X, weights=SD^-2)
> plot(Y~X,xlab="X: parent diameter",
+          ylab="Y: mean offspring diamter")
> abline(g)
> summary(g)

Call:
lm(formula = Y ~ X, weights = SD^-2)

Residuals:
       1        2        3        4        5        6        7
 0.08187  0.09162 -0.16753 -0.04067 -0.08950  0.06071  0.06328

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.79642    0.68112  18.787 7.87e-06 ***
X            0.20480    0.03815   5.368  0.00302 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.11 on 5 degrees of freedom
Multiple R-squared: 0.8521,     Adjusted R-squared: 0.8225
F-statistic: 28.81 on 1 and 5 DF,  p-value: 0.003021
```

  - Use Weights: $w_i = 1/\sigma_i^2$
  - Coefficients of weighted regression shown in output

- Plot Y vs X and fitted line



b) Hypothesis Testing

- $H_0: \beta_1 = 1$
- $H_1: \beta_1 \neq 1$
- Test statistic: $t_1 = \frac{\hat{\beta}_1 - 1}{se(\hat{\beta}_1)} = -20.844$
- Distribution test statistic: $\sim t_5$ (follows a t-distribution with 5 degrees of freedom)
- P-value$\approx$ 0< 0.05 → reject null : β_1 is statistically significantly different from 1
- Alternative use F-test (shown in R code)

```
> beta1=summary(g)$coeff[2]
> SE_beta1=0.03815
> t_stat1=(beta1-1)/SE_beta1
> t_stat1
[1] -20.84401
> pvalue=2*pt(t_stat1,5)
> pvalue
[1] 4.707033e-06
>
> gr=lm(Y~offset(X), weights=SD^-2)
> anova(gr,g)
Analysis of Variance Table

Model 1: Y ~ offset(X)
Model 2: Y ~ X
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1      6 5.3179
2      5 0.0605  1    5.2573 434.36 4.71e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c) Lack of fit test of the straight-line model

- H_0: model fits well
- H_1: lack of fit
- Test statistic:
  - General form: $\chi^2_{n-p} = \frac{(n-p)\hat{\sigma}^2}{\sigma^2}$
  - n = 7, p = 2, $\hat{\sigma} = 0.11$ (Residual standard error from R summary), $\sigma^2 = 1$ (because of the weights were defined) since standard deviations are assumed to be exactly correct
  - Test statistic: $\chi^2_5 = \frac{(n-p)\hat{\sigma}^2}{\sigma^2} = 0.0605$
- Distribution test statistic: : $\sim \chi^2_5$ (follows a Chi-square distribution with 5 degrees of freedom)
- P-value $\approx$ 0.9999> 0.05 →cannot reject null : **<u>no</u>** evidence of lack of fit of the model

```
> sigma=summary(g)$sigma
> chi_stat=sigma^2*(length(X)-2)
> pvalue=1-pchisq(chi_stat,length(X)-2)
>
> sigma
[1] 0.1100162
> chi_stat
[1] 0.06051786
> pvalue
[1] 0.999953
```

PROBLEM 2

Linear model with colonies as the response and log (dose +1) as the predictor.

```
> X=log(dose+1)
> fit0=lm(colonies~X)
> summary(fit0)

Call:
lm(formula = colonies ~ X)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.376  -6.882  -1.509   5.400  29.119 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   19.823      5.064   3.915  0.00123 **
X              2.396      1.128   2.125  0.04955 * 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.84 on 16 degrees of freedom
Multiple R-squared: 0.2201,     Adjusted R-squared: 0.1713 
F-statistic: 4.514 on 1 and 16 DF,  p-value: 0.04955 

>
> fit1=lm(colonies~factor(X))
> summary(fit1)

Call:
lm(formula = colonies ~ factor(X))

Residuals:
    Min      1Q  Median      3Q     Max 
-15.667  -3.917  -0.500   3.417  17.333 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)   
(Intercept)               21.667      5.506   3.935  0.00198 **
factor(X)2.39789527279837 -3.333      7.787  -0.428  0.67617   
factor(X)3.52636052461616  3.333      7.787   0.428  0.67617   
factor(X)4.61512051684126 21.000      7.787   2.697  0.01942 * 
factor(X)5.8111409929767  15.667      7.787   2.012  0.06722 . 
factor(X)6.90875477931522  8.000      7.787   1.027  0.32449   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.536 on 12 degrees of freedom
Multiple R-squared: 0.5475,     Adjusted R-squared: 0.359 
F-statistic: 2.904 on 5 and 12 DF,  p-value: 0.06047 

>
> Rsq = summary(fit0)$r.sq
> PureError = summary(fit1)$sigma^2
> Rsq
[1] 0.2200512
> PureError
[1] 90.94444
```

a) $\boldsymbol{R^2 = 0.220512}$(fit0)

b) Pure Error estimate of $\boldsymbol{\sigma^2 = 90.944}$($\boldsymbol{\sigma = 9.536}$ from fit1)

c) F-statistic

- $F = \frac{(RSS_0 - RSS_1)/(df_0 - df_1)}{RSS_1/df_1}$
- Note that fit1 (full) is the more complex model since by using factor the design matrix has more columns. Thus, it is expected to have a lower RSS than that of the simpler model fit0 (reduced) because of more variables in the model.
- $RSS_0 = 1881, RSS_1 = 1091, df_0$=16,$df_1 = 12$
- $F = 2.1709$

d) Distribution of F-statistics under the null hypothesis

- F statistic follows a F-distribution with v1=4 and v2=12 degrees of freedom
- $F \sim F_{4,12}$
- Null: reduced model is better (fit0)
- Alternative: full model is better (fit1)

e) Lack of fit

- P-value of F-test = 0.134
- P-value > 0.05 : Can't reject null (Ho: X=log(dose+1)) → **no** evidence of lack of fit
- We can see that the regression standard error (10.84) is not substantially greater than the pure error (9.536), indicating **no** evidence of lack of fit. If the regression standard error were substantially greater, we could conclude a poor fit of the model.

```
> rss.fit0=sum(fit0$res^2)
> rss.fit1=sum(fit1$res^2)
> myF=(rss.fit0-rss.fit1)/(16-12)/(rss.fit1/12)
> pvalue= 1-pf(myF, 4, 12)
> rss.fit0
[1] 1881.063
> rss.fit1
[1] 1091.333
> myF
[1] 2.170913
> pvalue
[1] 0.1341968
> anova(fit0, fit1)
Analysis of Variance Table

Model 1: colonies ~ X
Model 2: colonies ~ factor(X)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     16 1881.1
2     12 1091.3  4    789.73 2.1709 0.1342
```

PROBLEM 3

Using divusadata,fit a regression model with divorce as the response and unemployed, femlab, marriage, birth and military as predictors.

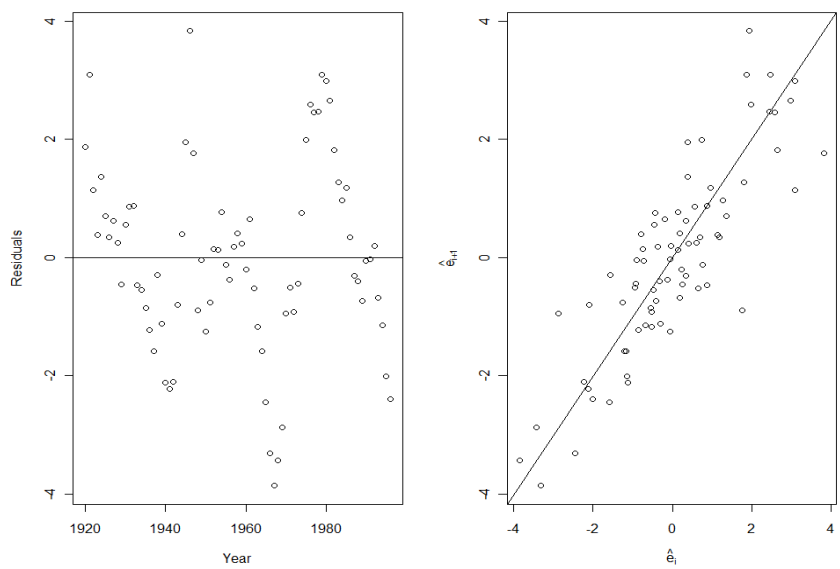
```
> #Question 3
>
> help(divusa)
> data(divusa)
>
> # divusa data
> g=lm(divorce ~ unemployed + femlab + marriage + birth + military, divusa)
> n=nrow(divusa)  # sample size
> cor(g$res[-1], g$res[-n])
[1] 0.8469792
```

- Note the high correlation

a) Correlated Errors

- Graphical Checks

```
> # two graphical displays for correlated errors
> # (1) plot res vs. time
> # (2) plot res[i] vs res[i+1]
>
> par(mfrow=c(1,2))
> plot(divusa$year, g$res, xlab="Year", ylab="Residuals"); abline(h=0)
> plot(g$res[-n], g$res[-1], xlab=expression(hat(e)[i]), ylab=expression(hat(e)[i+1])); abline(0,1)
```



- If there was serial correlation, we would see either longer runs of residuals above or below the line (positive correlation) or greater than normal fluctuation (negative correlation).
- From the graph residual vs time, we see longer runs above and below, indicating possible positive correlation.
- The graph of successive residuals also confirms the high positive correlation between consecutive residuals, suggesting a strong serial correlation.

- Durbin-Watson test

```
> # Durbin-Watson test
> library(lmtest)
> dwtest(divorce ~ unemployed + femlab + marriage + birth + military, data=divusa)

        Durbin-Watson test

data:  divorce ~ unemployed + femlab + marriage + birth + military
DW = 0.2999, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

  - Statistic: $DW = \frac{\sum_{i=2}^{n}(\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1})^2}{\sum_{i=1}^{n}\hat{\varepsilon}_i^{\,2}}$
  - H_0: errors are uncorrelated (assumption of uncorrelated errors follows a linear combination of χ^2)
  - H_1: errors are correlated and follow an AR(1) process

- Conclusion:
  - DW test p-value ≈ 0 <0.05, reject null hypothesis and accept alternative. There is evidence of correlation. Correlation is significantly different from zero.

b) Speculation on correlation in the errors

- In data collected over time such as this, successive errors could be correlated.
- This is because errors related to a given time period carry over into future time periods.
- For instance, it is likely that an underestimate of the divorce rate in 1920 leads to an underestimate in the next years.

c) Serial correlation with an AR(1) model for the errors

- Estimated Correlation. Significant?

```
> g2=gls(divorce ~ unemployed + femlab + marriage + birth + military, correlation = corAR1(form=~year), method="ML", data=divusa)
> # Now you should see "Generalized least squares fit by maximum likelihood..."
> intervals(g2)  # you can use the CI for rho to do hypothesis testing.
Approximate 95% confidence intervals

 Coefficients:
                  lower        est.        upper
(Intercept) -18.12047041 -7.05968161  4.001107187
unemployed    0.01609101  0.10764313  0.199195251
femlab        0.12235846  0.31208493  0.501811411
marriage      0.11867101  0.16432630  0.209981587
birth        -0.09380023 -0.04990919 -0.006018159
military     -0.01050915  0.01794640  0.046401945
attr(,"label")
[1] "Coefficients:"

 Correlation structure:
       lower      est.     upper
Phi 0.6528156 0.9715486 0.9980191
attr(,"label")
[1] "Correlation structure:"

 Residual standard error:
    lower       est.      upper
 0.7974459  2.9076645 10.6019896
```

  - From the R-output, the estimated correlation (phi) $= 0.9715$
  - If we check the confidence interval (0.6529,0.9981), we see it does not include zero (reject null), which means the estimated correlation is significantly different from zero. So there is evidence of serial correlation.

- GLS model change variables to be significant

```
> summary(g2)
Generalized least squares fit by maximum likelihood
  Model: divorce ~ unemployed + femlab + marriage + birth + military
  Data: divusa
       AIC      BIC    logLik
  179.9523 198.7027 -81.97613

Correlation Structure: AR(1)
 Formula: ~year
 Parameter estimate(s):
      Phi
0.9715486

Coefficients:
                Value Std.Error   t-value p-value
(Intercept) -7.059682  5.547193 -1.272658  0.2073
unemployed   0.107643  0.045915  2.344395  0.0219
femlab       0.312085  0.095151  3.279878  0.0016
marriage     0.164326  0.022897  7.176766  0.0000
birth       -0.049909  0.022012 -2.267345  0.0264
military     0.017946  0.014271  1.257544  0.2127

 Correlation:
           (Intr) unmply femlab marrig birth
unemployed -0.420
femlab     -0.802  0.240
marriage   -0.516  0.607  0.307
birth      -0.379  0.041  0.066 -0.094
military   -0.036  0.436 -0.311  0.530  0.128

Standardized residuals:
       Min         Q1        Med         Q3        Max
-1.4509327 -0.9760939 -0.6164694  1.1375377  2.1593261

Residual standard error: 2.907665
Degrees of freedom: 77 total; 71 residual
> summary(g)

Call:
lm(formula = divorce ~ unemployed + femlab + marriage + birth +
    military, data = divusa)

Residuals:
     Min       1Q   Median       3Q      Max
-3.86106 -0.89160 -0.04958  0.86497  3.82998

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.48784    3.39378   0.733   0.4659
unemployed  -0.11125    0.05592  -1.989   0.0505 .
femlab       0.38365    0.03059  12.543  < 2e-16 ***
marriage     0.11867    0.02441   4.861 6.77e-06 ***
birth       -0.12996    0.01560  -8.333 4.03e-12 ***
military    -0.02673    0.01425  -1.876   0.0647 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.65 on 71 degrees of freedom
Multiple R-squared: 0.9208,     Adjusted R-squared: 0.9152
F-statistic: 165.1 on 5 and 71 DF,  p-value: < 2.2e-16
```

  - From the first model (g), the variables found to be significant (at 0.05 level) were femlab, marriage and birth (p-value <0.05). From the GLS model (g2), the variables found to be significant (at 0.05 level) were unemployed, femlab, marriage and birth (p-value <0.05). Thus, the GLS model changed the variable unemployed to be significant compare to the first regression model (although the variable unemployed had a p-value of 0.0505 in the first model, so it was almost significant at 0.05 level; in other words, the GLS changed variable unemployed from borderline significance at 0.05 level to significant at a 0.05 level)

PROBLEM 4

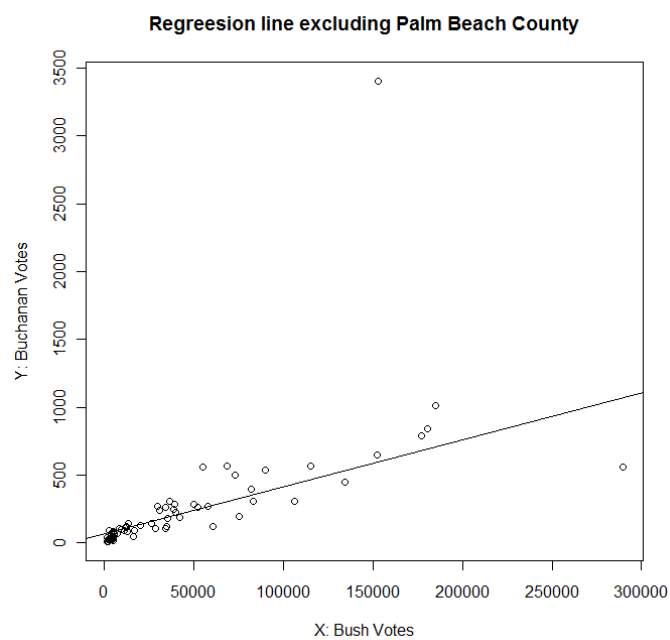
a) Simple linear regression (Y vs X), excluding the sample from the Palm Beach

- Regression

```
> data=read.table(file="E:/R/elec2000.csv", sep=",",header=TRUE)
> Y=data$BUCHANAN
> X=data$BUSH
> fit0=lm(Y[-50]~X[-50])
> summary(fit0)

Call:
lm(formula = Y[-50] ~ X[-50])

Residuals:
    Min      1Q  Median      3Q     Max 
-511.85  -48.34  -13.92   41.51  305.12 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.609e+01  1.728e+01   3.825 0.000299 ***
X[-50]      3.478e-03  2.488e-04  13.982  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 112.2 on 64 degrees of freedom
Multiple R-squared: 0.7534,     Adjusted R-squared: 0.7495 
F-statistic: 195.5 on 1 and 64 DF,  p-value: < 2.2e-16
```

- Plot data and fitted regression line

```
> plot(Y~X,xlab="X: Bush Votes",ylab="Y: Buchanan Votes",
+      main="Regreesion line excluding Palm Beach County")
> abline(fit0)
```

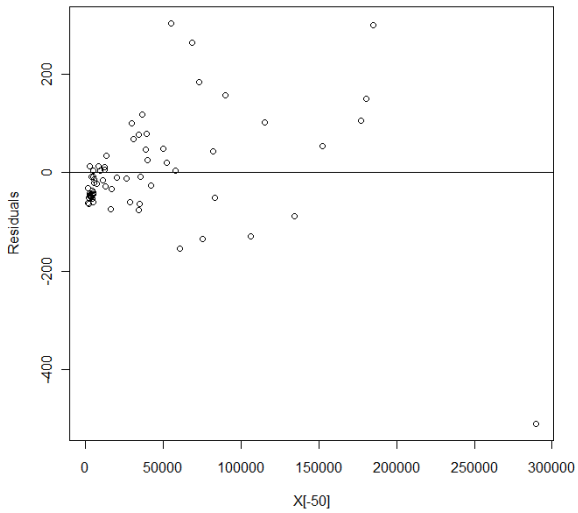


  - Note: the point (Palm Beach County) excluded in the regression model does not seem to fit the regression line

b) Plot residuals versus X to check for non-constant variance.

- Plot

```
> plot(X[-50], fit0$res,xlab="X[-50]",ylab="Residuals")
> abline(h=0)
```



- Conclusion: plot shows non-constant variance since the residuals (absolute value) seem to increase with increasing X. Try transformation to deal with non-constant variance.

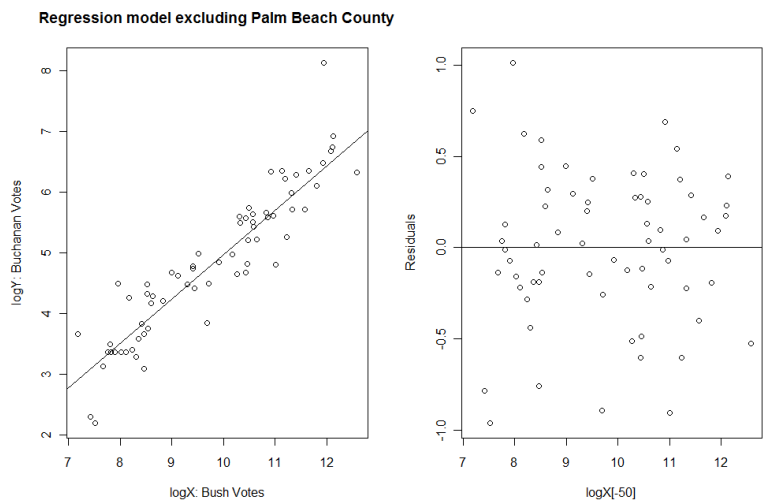
c) Simple linear regression log(Y)vs log(X), excluding the sample from the Palm Beach

- Regression

```
> logY=log(Y)
> logX = log(X)
> fit1=lm(logY[-50]~logX[-50])
> summary(fit1)

Call:
lm(formula = logY[-50] ~ logX[-50])

Residuals:
     Min       1Q   Median       3Q      Max 
-0.96259 -0.20866  0.02995  0.27682  1.01461 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.32980    0.35284  -6.603  9.2e-09 ***
logX[-50]    0.73024    0.03581  20.395  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4181 on 64 degrees of freedom
Multiple R-squared: 0.8667,     Adjusted R-squared: 0.8646 
F-statistic: 415.9 on 1 and 64 DF,  p-value: < 2.2e-16
```

- Plot data and fitted regression line and Plot residuals vs log X

```
> par(mfrow=c(1,2))
> plot(logX,logY,xlab="logX: Bush Votes",ylab="logY: Buchanan Votes",
+      main="Regression model excluding Palm Beach County")
> abline(fit1)
> plot(logX[-50], fit1$res,xlab="logX[-50]",ylab="Residuals")
> abline(h=0)
```



- Compare models
  - The log model seems more reasonable
  - By comparing the data plots, the log regression seems to fit the data better.
  - By comparing the residual plots, the non-constant variance from the first regression seem to be resolved using the log transformation. By looking at the residual plot of the log transformation, we do not see any evidence indicating non constant variance.
  - In addition, thehigher R^2 of the log model (R^2= 0.8667) compared to the original model (R^2= 0.7534) indicates that the log model does a better job explaining the data.

d) Form a 95% prediction interval for Buchanan's vote in Palm Beach.

```
> lx=logX[-50]
> ly=logY[-50]
> fitx=lm(ly~lx)
> lxPalm=data.frame(lx=logX[50])
> logPI=predict.lm(fitx,lxPalm,level=0.95,interval="prediction")
> PI=exp(logPI)
> PI
       fit      lwr      upr
1 594.2119 252.4666 1398.553
> Y[50]
[1] 3407
```

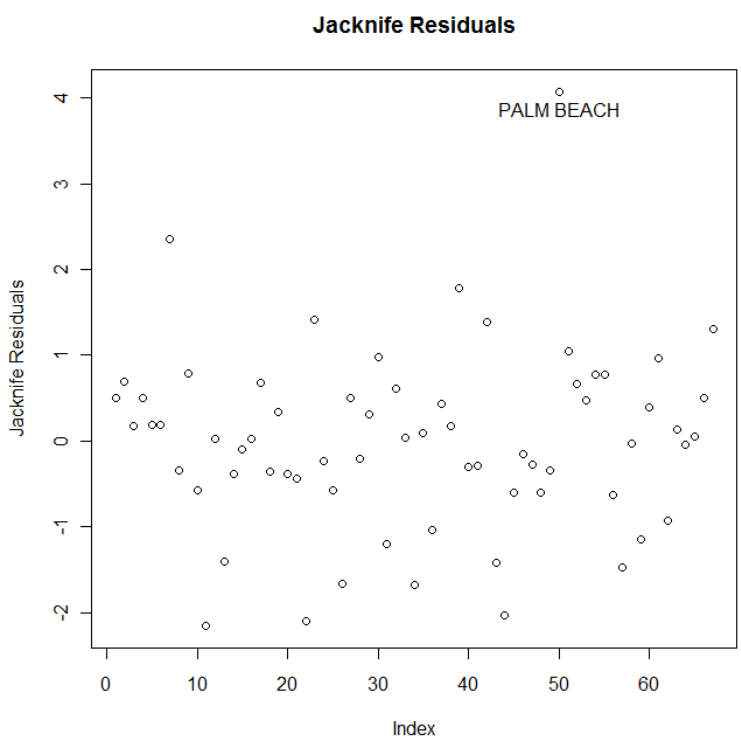
- Use linear model from part c (log transformation excluding Palm Beach County)
- X[50] : number of votes for Bush in Palm Beach
- Y[50] : the observed value for the number of votes for Buchanan in Palm Beach
- The 95% prediction interval for the predicted value of X[50] is: (252.47,1398.56)
- The observed value was Y[50] = 3407. Thus, it does not belong to prediction interval (it is more than twice the upper bound), indicating that Buchanan's votes (3407) in Palm Beach is <u>not</u> a statistically plausible outcome, assuming the linear model is appropriate.

e) Outlier-test on each county with Bonferroni correction.

- Fit a regression model for all samples using the log transformation

```
> fit3=lm(logY~logX)
> jack = rstudent(fit3)
> names(jack)=data$COUNTY
> plot(jack,ylab="Jacknife Residuals",main="Jacknife Residuals")
> identify(1:nrow(data),jack,data$COUNTY)

>
> n=nrow(data)
> n
[1] 67
> p=2
> alpha=0.05
> tcrit=qt(alpha/(2*n),n-p-1)
> outliers=jack[abs(jack)>abs(tcrit)]
> tcrit
[1] -3.542302
> outliers
PALM BEACH
  4.075833
```

  - Compute t_i, the jackknife (externally studentized) residual for all samples
  - t_i is distributed t_{n-p-1} (the i index denotes regression model without sample i)
  - Use Bonferroni correction: use α/n in each test
  - Test t_i against $t_{crit} = t_{64,0.025/67}$
  - If $|t_i|>|t_{crit}|$, then conclude that sample i is an outlier
- Plot



- Performing the outlier test (with Bonferroni correction) for all counties leads to the conclusion that only Palm Beach is an outlier (t-stat =4.08> t-crit=3.54)

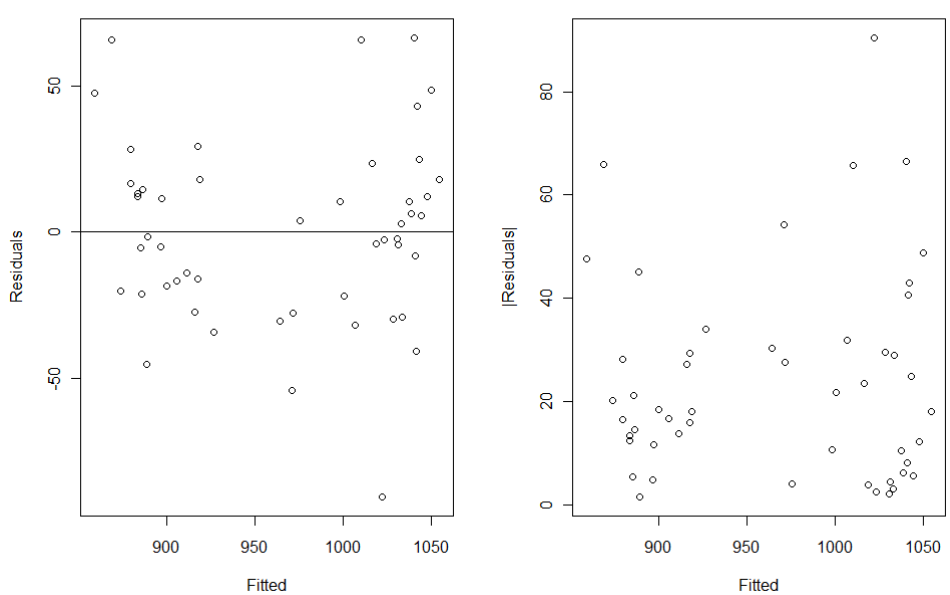
PROBLEM 5

Fit a model with the total SAT score as the response and expend, salary, ratio and takers as predictors.

a) Check the constant variance assumption for the errors

- Under assumption: residuals should vary randomly around zero and the spread of the residuals should be about the same throughout the plot (no systematic patterns, scatter should be symmetric vertically about zero)
- Plot: Residual vs Fitted and l Residual l vs Fitted to increase resolution

```
> library(faraway)
> data(sat)
>
> g = lm(total~expend + salary + ratio + takers, data=sat)
> summary(g)

Call:
lm(formula = total ~ expend + salary + ratio + takers, data = sat)

Residuals:
    Min      1Q  Median      3Q     Max
-90.531 -20.855  -1.746  15.979  66.571

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1045.9715    52.8698  19.784  < 2e-16 ***
expend         4.4626    10.5465   0.423    0.674
salary         1.6379     2.3872   0.686    0.496
ratio         -3.6242     3.2154  -1.127    0.266
takers        -2.9045     0.2313 -12.559 2.61e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 32.7 on 45 degrees of freedom
Multiple R-squared: 0.8246,     Adjusted R-squared: 0.809
F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16

>
> ### part (a) Check the constant variance assumption for the errors
>
> #plot res vs fitted
> par(mfrow=c(1,2))
> plot(g$fit,g$res,xlab="Fitted",ylab="Residuals")
> abline(h=0)
> plot(g$fit,abs(g$res),xlab="Fitted",ylab="|Residuals|")
```



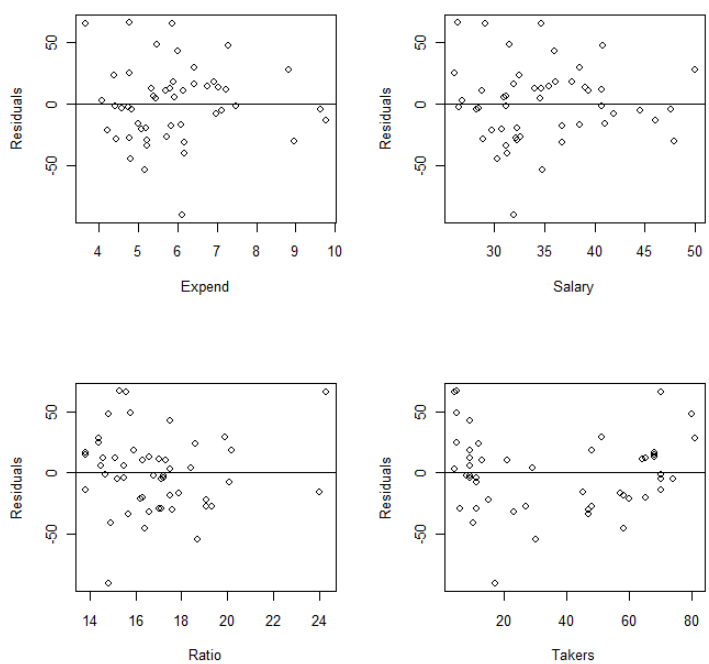
- Comment: there seems to be non-constant variance (high residuals on the edges and low residuals in the center).

- Quick test: regression of absolute residual vs fitted

```
> #regression res vs fit
> summary(lm(abs(g$res) ~ g$fit))

Call:
lm(formula = abs(g$res) ~ g$fit)

Residuals:
    Min      1Q  Median      3Q     Max 
-22.827 -14.323  -5.001   6.204  65.754 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.60384   40.82315   0.260    0.796
g$fit        0.01386    0.04216   0.329    0.744

Residual standard error: 20.05 on 48 degrees of freedom
Multiple R-squared: 0.002247,	Adjusted R-squared: -0.01854 
F-statistic: 0.1081 on 1 and 48 DF,  p-value: 0.7438
```

- The p-value shows the regression is not significant, so there doesn't seem to be a problem with a variance increasing or decreasing relationship (note: this test is not quite right as some weighting should be used and the degrees of freedom should be adjusted).
- Plot: residuals vs predictors

```
> par(mfrow=c(2,2))
> plot(sat$expend,g$res,xlab="Expend",ylab="Residuals")
> abline(h=0)
> plot(sat$salary,g$res,xlab="Salary",ylab="Residuals")
> abline(h=0)
> plot(sat$ratio,g$res,xlab="Ratio",ylab="Residuals")
> abline(h=0)
> plot(sat$takers,g$res,xlab="Takers",ylab="Residuals")
> abline(h=0)
```



- Comment: only the plot with takers as the predictor shows some pattern, so investigate this case. It seems that there two different groups (less than 40 and greater than 40). Make a comparison and test for equal variances using an F-test (using the test statistic of the ratio of the two variance). The null distribution is a F with degrees of freedom given by the two samples.

```
> var.test(residuals(g)[sat$takers>40],residuals(g)[sat$takers<40])

        F test to compare two variances

data:  residuals(g)[sat$takers > 40] and residuals(g)[sat$takers < 40]
F = 0.6079, num df = 22, denom df = 26, p-value = 0.2392
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2708575 1.4034851
sample estimates:
ratio of variances
          0.607878
```

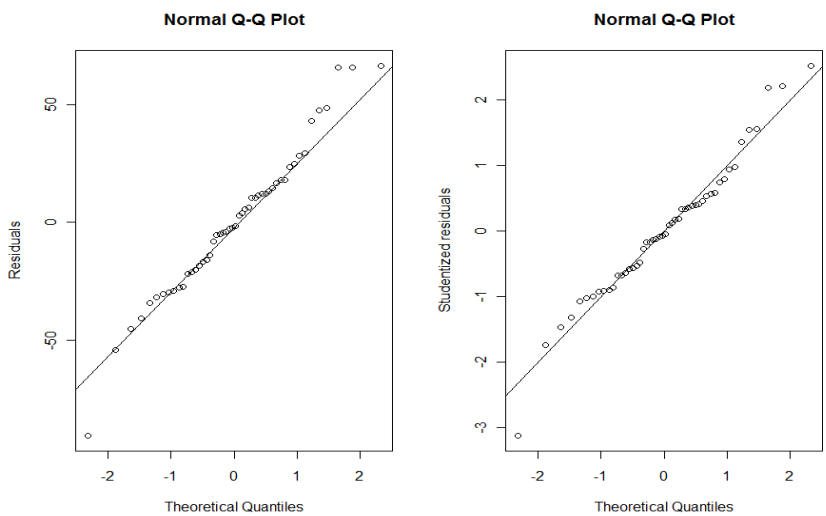
- No significant difference between groups (p-value > 0.05 : no evidence to reject null that variances are equal)
- However, as noted before, the residuals vs takers show non-constant variance (high residuals on the edges and low residuals in the center). This might suggest a problem of non-linearity (investigated in part f)

- ***Conclusion*** *: Non-constant variance detected, so assumption of constant variance is not reasonable. Might need to transform the variable takers to address this issue.*

b) Check the normality assumption

- The residuals can be assessed for normality using a Q–Q plot. This compares the residuals to "ideal" normal observation. Normal residuals should follow the line approximately
- Q-Q Plot (Residuals and Studentized residuals)

```
> par(mfrow=c(1,2))
> qqnorm (residuals (g), ylab="Residuals")
> qqline (residuals (g))
> qqnorm(rstudent(g),ylab="Studentized residuals")
> abline(0,1)
```



  - The residuals look normal since the residuals don't deviate much from a straight line (for the studentized residuals, they should lie along a 45 degree line since they have been normalized.
- The Shapiro-Wilk test is a formal test for normality

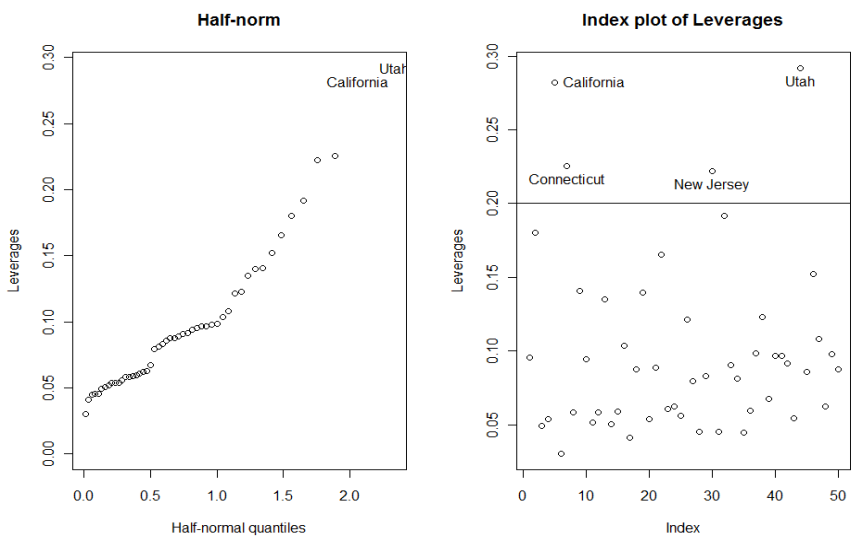
```
> shapiro.test (residuals (g))

        Shapiro-Wilk normality test

data:  residuals(g)
W = 0.9769, p-value = 0.4304
```

  - Since the p-value is large (greater than 0.05), we do not reject the null (the residuals are normal)
- ***Conclusion*** *: assumption of normality (observed samples come from normal distribution) is reasonable*

c) Check for large leverage points

- A leverage point: "is unusual point in the predictor space—it has the potential to influence the fit" (Identifies with unusual or outlying x-vales)
- Leverages greater than 2p/n should be examined
- Half-normal: "for outliers which will be apparent as points that diverge substantially from the rest of the data"
- Plot scatter and half-normal of leverages

```
> par(mfrow=c(1,2))
> states = row.names(sat)
> halfnorm(influence(g)$hat,labs=states, ylab="Leverages", main="Half-norm")
> plot(ginf$hat,ylab="Leverages",main="Index plot of Leverages")
> abline(h=2*5/nrow(sat))
> identify(1:50,lev,states)
[1]  5  7 30 44
>
> lev = ginf$hat
> lev[lev > 2*5/nrow(sat)]
 California Connecticut  New Jersey        Utah
  0.2821179   0.2254519   0.2220978   0.2921128
```

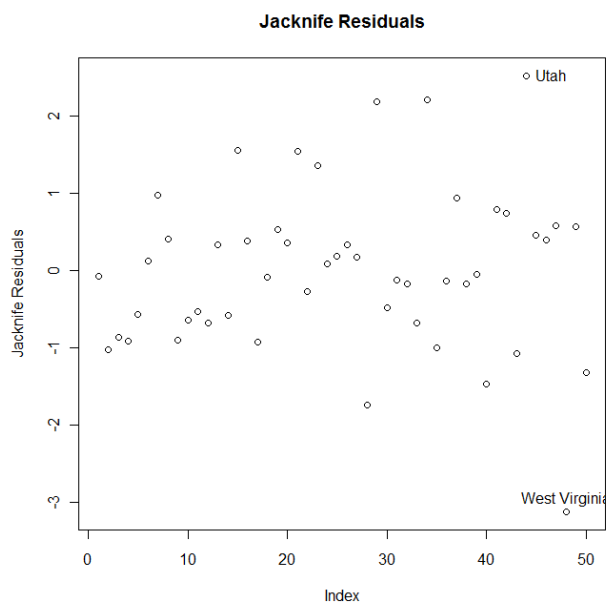


- ***Conclusion:*** High leverage points: California and Utah (Connecticut and New Jersey to a lesser degree). Note: "do not necessarily expect a positive normal distribution (straight line) for quantities like the leverages . We are looking for outliers, which will be apparent as points that diverge substantially from the rest of the data".

d) Check for outliers

- Outlier: a really unusual observation that does not fit model("An observation that deviates from the overall pattern of relationship between Y and X. Typically has a residual that is large in absolute value"
- Test outlier:
  - Compute t_i, the jackknife (externally studentized) residual for all samples
  - t_i is distributed t_{n-p-1} (the i index denotes regression model without sample i)
  - Use Bonferroni correction: use α/n in each test
  - Test t_i against $t_{crit} = t_{44,0.025/50}$
  - If $|t_i|>|t_{crit}|$, then conclude that sample i is an outlier
- Plot jackknife residuals

```
> jack = rstudent(g)
> names(jack)=states
> plot(jack,ylab="Jacknife Residuals",main="Jacknife Residuals")
> identify(1:nrow(sat),jack,states)
[1] 44 48
>
> n=nrow(sat)
> n
[1] 50
> p=5
> alpha=0.05
> tcrit=qt(alpha/(2*n),n-p-1)
> outliers=jack[abs(jack)>abs(tcrit)]
> max(abs(jack))
[1] 3.124428
> abs(tcrit)
[1] 3.525801
> outliers
named numeric(0)
```

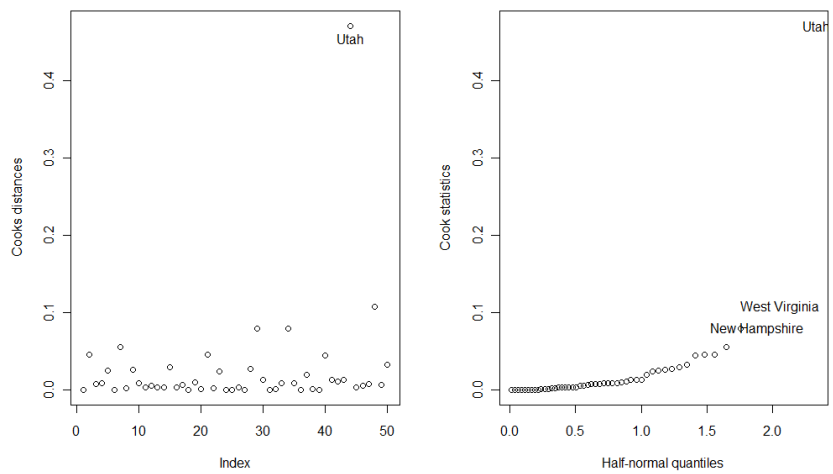


- ***Conclusion***: Performing the outlier test (with Bonferroni correction) for all states leads to the conclusion that there are no outliers (e.g. max abs(jack) = 3.12 < abs(tcrit) =3.53)

e) Check for influential points

- Influential point:"one whose removal from the dataset would cause a large change in the fit"
- Compute cook distance
- Plot scatter and half-norm of cook distance
- Half-normal: "for outliers which will be apparent as points that diverge substantially from the rest of the data"

```
> cook =cooks.distance(g)
> par(mfrow=c(1,2))
> plot(cook,ylab="Cooks distances")
> identify(1:50,cook,states)
[1] 44
> halfnorm(cooks.distance(g), labs=states, nlab=3, ylab="Cook statistics")
```



- Comment: it seems that Utah is the highest influential point (West Virginia and New Hampshire to a lesser degree)

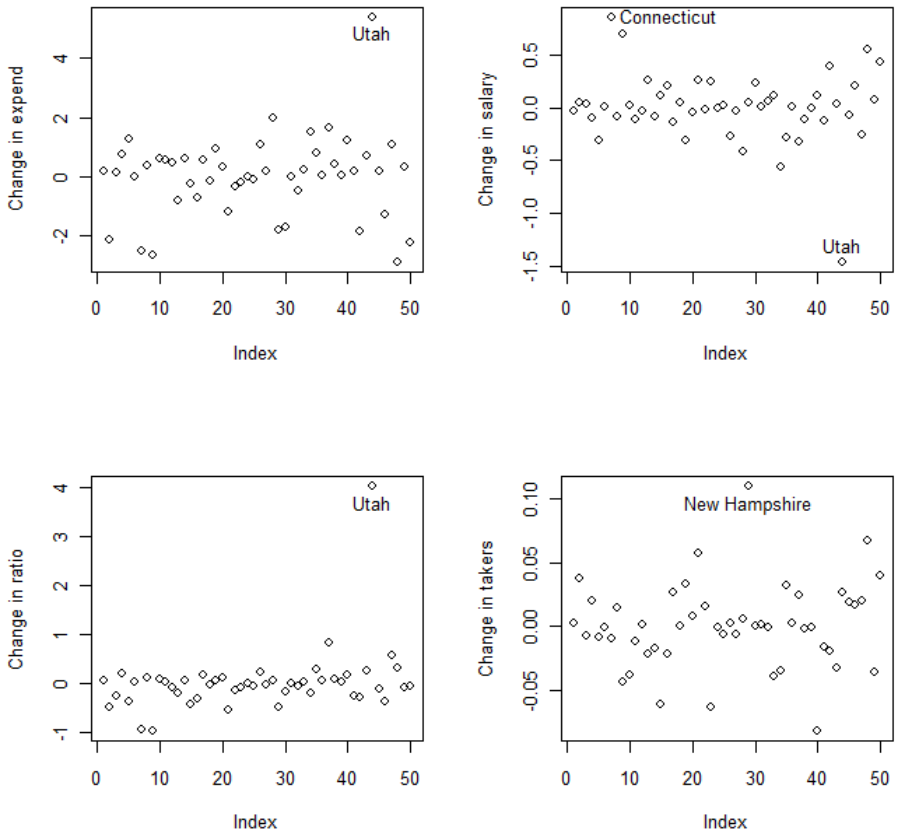
- Linear Regression without the highest cook distance data point

```
> g1 = lm(total~expend + salary + ratio + takers,subset=(cook < max(cook)))
> summary(g1)

Call:
lm(formula = total ~ expend + salary + ratio + takers, subset = (cook <
    max(cook)))

Residuals:
    Min      1Q  Median      3Q     Max
-92.118 -18.402   1.808  14.890  67.669

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1093.8460    53.4226  20.475   <2e-16 ***
expend        -0.9427    10.1922  -0.092    0.927
salary         3.0964     2.3283   1.330    0.190
ratio         -7.6391     3.4279  -2.229    0.031 *
takers        -2.9308     0.2188 -13.397   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30.9 on 44 degrees of freedom
Multiple R-squared: 0.8396,     Adjusted R-squared: 0.825
F-statistic: 57.58 on 4 and 44 DF,  p-value: < 2.2e-16

> summary(g)

Call:
lm(formula = total ~ expend + salary + ratio + takers, data = sat)

Residuals:
    Min      1Q  Median      3Q     Max
-90.531 -20.855  -1.746  15.979  66.571

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1045.9715    52.8698  19.784  < 2e-16 ***
expend         4.4626    10.5465   0.423    0.674
salary         1.6379     2.3872   0.686    0.496
ratio         -3.6242     3.2154  -1.127    0.266
takers        -2.9045     0.2313 -12.559 2.61e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 32.7 on 45 degrees of freedom
Multiple R-squared: 0.8246,     Adjusted R-squared: 0.809
F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16
```

  - Comment: linear regression excluding the high influential point (Utah) changes the predictor 'ratio' to significant. Also the coefficients of expend, ratio, and salary seemed to have changed significantly

- Plot the change of each coefficient when a particular case (e.g. state) is excluded from the model (leave out-one differences in the coefficients)

```
> ginf = influence (g)
>
> par(mfrow=c(2,2))
> plot(ginf$coef [,2], ylab="Change in expend")
> identify (1:50, ginf$coef [, 2], states)
[1] 44
> plot(ginf$coef [,3], ylab="Change in salary")
> identify (1:50, ginf$coef [, 3], states)
[1]  7 44
> plot(ginf$coef [,4], ylab="Change in ratio")
> identify (1:50, ginf$coef [, 4], states)
[1] 44
> plot(ginf$coef [,5], ylab="Change in takers")
> identify (1:50, ginf$coef [, 5], states)
[1] 29
```



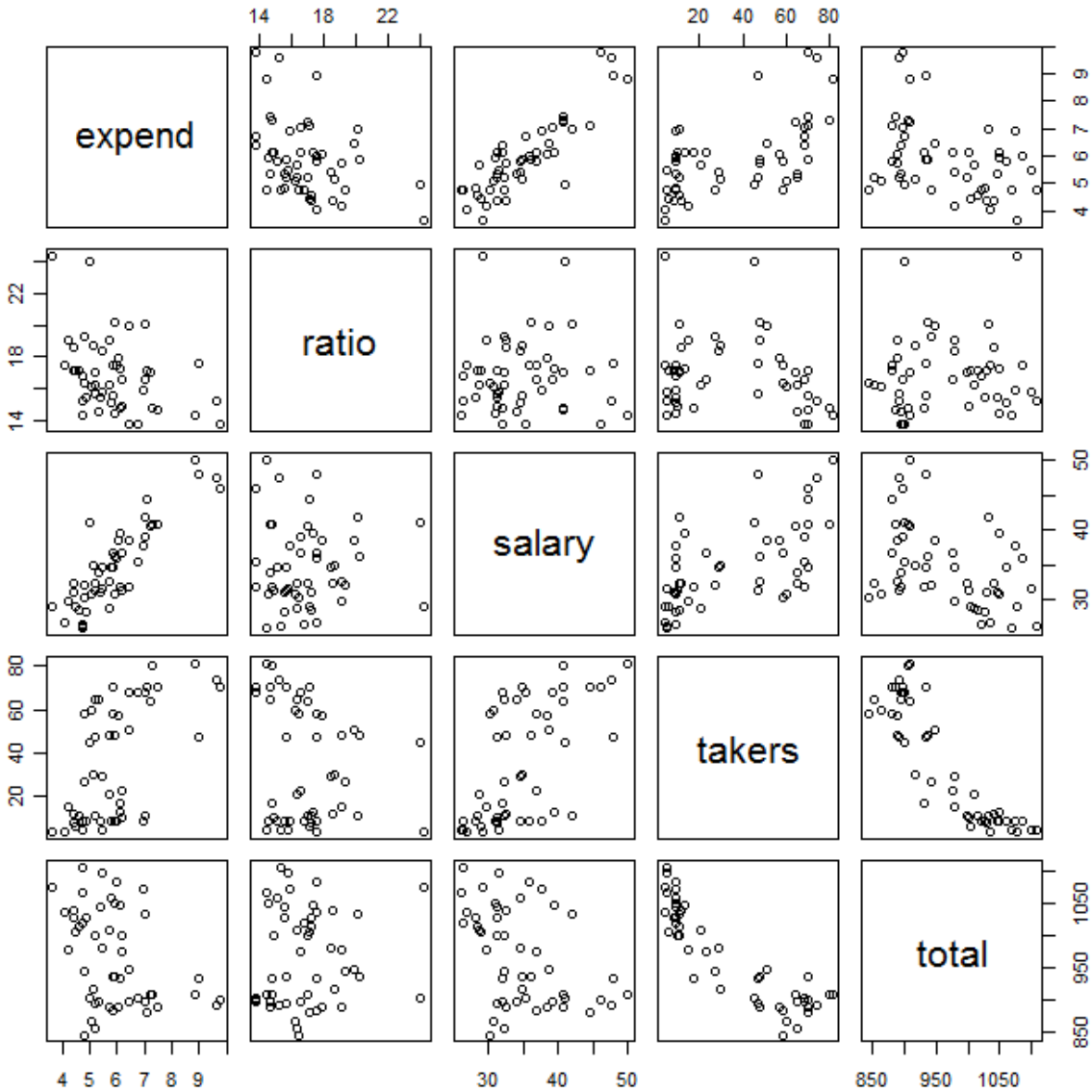
- *Comment:* the Utah point stands out as it is the one that changes the most the coefficients in expend, salary and ratio.

- ***Conclusion:*** from the analysis, it seems that Utah is the most influential point, changing the coefficients of some predictors the greatest and making one coefficient significant when Utah is left out of the model.

f) Check the structure of the relationship between the predictors and the response

- Plot y vs xi

```
> pairs(~expend+ratio+salary+takers+total)
```



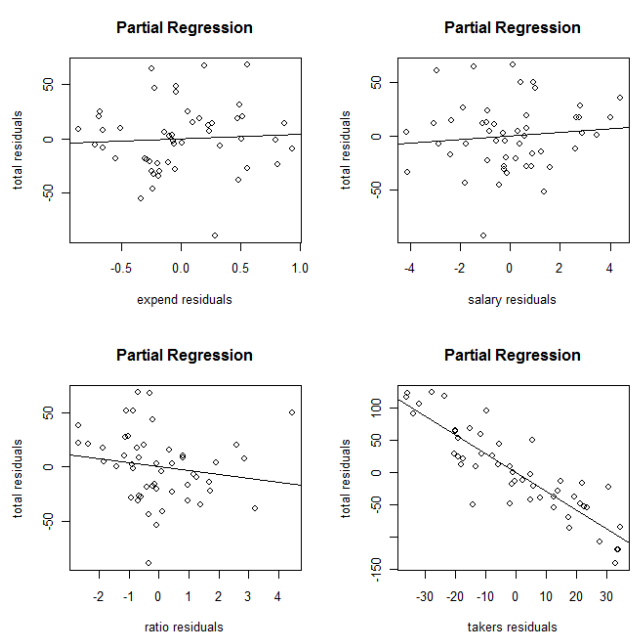
  - Comment: only one scatter plot stands out: total vs takers. We see a possible non-linear relationship between the response total and the variable takers

- Plot residuals vs xi variables (in and out of model)
  - This is plotted in part a)
  - No non-linear trends where spotted, except that there seems to be two groups for the predictor takers
- Plot residuals vs predicted
  - This is plotted in part a)
  - No non-linear trends where spotted,

- The drawback to these plots is that the other predictors impact the relationship.
- Partial Regression or Added Variable plots
  - Can help isolate the effect of xi on y (marginal relationship between the response and the predictor after the effect of the other predictors has been removed)
  - Use: outlier/influential diagnosis
  - Procedure:

    1. Regress y on all x except xi, get residuals d. This represents y with the other X-effect taken out.
    2. Regress xi on all x except xi, get residuals m. This represents xi with the other X-effect taken out
    3. Plot d vs m

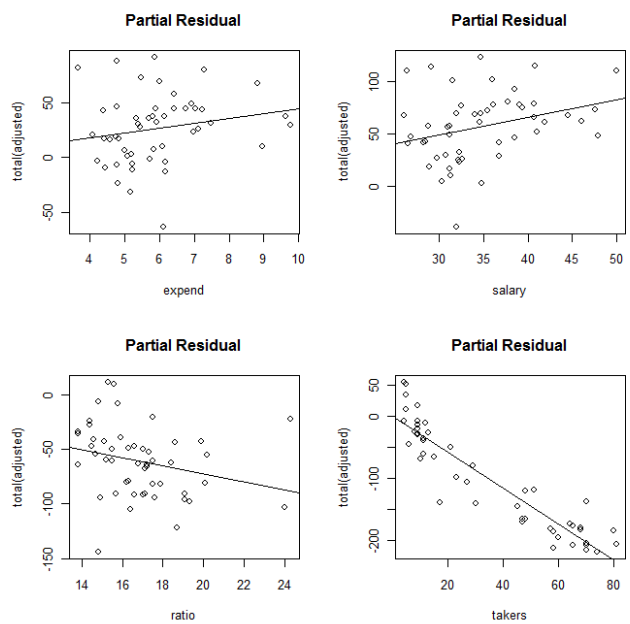
```
> d_e = lm(total~ salary + ratio + takers ,sat)$res
> m_e = lm(expend~ salary + ratio + takers ,sat)$res
> gexpend=lm(d_e~m_e )
>
> d_s = lm(total~ expend + ratio + takers ,sat)$res
> m_s = lm(salary~ expend + ratio + takers ,sat)$res
> gsalary=lm(d_s~m_s)
>
> d_r = lm(total~ expend + salary + takers ,sat)$res
> m_r = lm(ratio~ expend + salary + takers ,sat)$res
> gratio=lm(d_r~m_r)
>
> d_t = lm(total~expend + salary + ratio ,sat)$res
> m_t = lm(takers~expend + salary + ratio ,sat)$res
> gtakers=lm(d_t~m_t)
> par(mfrow=c(2,2))
> plot(m_e,d_e,xlab="expend residuals",ylab="total residuals",
+ main="Partial Regression")
> abline(gexpend)
>
> plot(m_s,d_s,xlab="salary residuals",ylab="total residuals",
+ main="Partial Regression")
> abline(gsalary)
>
> plot(m_r,d_r,xlab="ratio residuals",ylab="total residuals",
+ main="Partial Regression")
> abline(gratio)
>
> plot(m_t,d_t,xlab="takers residuals",ylab="total residuals",
+ main="Partial Regression")
> abline(gtakers)
```



- Comment: after accounting for the rest of the predictors, it seems that there is a positive association between the total sat score and the predictor takers.

- Partial Residual plots
  - These plot $\hat{\varepsilon} + \hat{\beta}_i x_i$ against x_i
  - The benefit is that it can reveal the effect of the predictor xi after accounting for the other predictors
  - Use: non-linearity diagnosis
  - "Most commonly used to identify the nature of the relationship between Y and X_i (given the effect of the other independent variables in the model)"
  - "If the partial residual plot shows a straight line, it is an indication that the true relationship between the response variable and the i-th explanatory variable is straight-line, when all other variables are taken into account. If the plot shows a non-linear relationship, it is an indication that xi affects the response variable in a non-linear fashion"

```
> par(mfrow=c(2,2))
> prplot(g,1)
> prplot(g,2)
> prplot(g,3)
> prplot(g,4)
```



- Comment: we see a non-linearity for the relationship between total and takers. We see the two groups in the plot. It suggests that there may be a different relationship in the two groups. Examine this.

```
> g1 = lm(total~expend + salary + ratio + takers,subset=(takers > 40))
> g2 = lm(total~expend + salary + ratio + takers,subset=(takers < 40))
> summary(g1)

Call:
lm(formula = total ~ expend + salary + ratio + takers, subset = (takers >
    40))

Residuals:
    Min      1Q  Median      3Q     Max
-38.434 -15.938  -2.509  11.444  50.289

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 801.4329   105.6773   7.584  5.2e-07 ***
expend       11.1444    10.8359   1.028    0.317
salary       -0.6354     2.7190  -0.234    0.818
ratio         3.9147     4.8627   0.805    0.431
takers       -0.3003     0.8869  -0.339    0.739
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.74 on 18 degrees of freedom
Multiple R-squared: 0.2627,     Adjusted R-squared: 0.09882
F-statistic: 1.603 on 4 and 18 DF,  p-value: 0.2167

> summary(g2)

Call:
lm(formula = total ~ expend + salary + ratio + takers, subset = (takers <
    40))

Residuals:
    Min      1Q  Median      3Q     Max
-69.155 -16.672   1.747  15.654  54.956

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 993.7178    84.5010  11.760 5.86e-11 ***
expend        7.7581    16.4329   0.472    0.641
salary        1.0293     3.3058   0.311    0.758
ratio         1.4251     4.6111   0.309    0.760
takers       -5.5242     0.8706  -6.345 2.19e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30.95 on 22 degrees of freedom
Multiple R-squared: 0.656,      Adjusted R-squared: 0.5935
F-statistic: 10.49 on 4 and 22 DF,  p-value: 6.549e-05
```

- Comments: in the first regression on the subset (takers >40), no relation between the predictors and the response (p-value>0.05). In contrast, there is a strong relationship in the second subset (takers <40), with p-value < 0.05. This indicates a non-linear relationship of total and all takers data points.

- ***Conclusion:*** percent takers seems to have a nonlinear relationship with the total sat score. It is suggested to use a transformation (log(takers)) to address this issue.

Note: the analysis was based on the requested (SAT score as the response and expend, salary, ratio and takers as predictors). If appropriate transformations for the variables are made, different results for the diagnosis are expected.