## 1. Text8.1

### (a)

H0: $\rho$ =0; Ha= $\rho$ >0

Durbin Watson statistic is d=0.6208403. By checking table A.6, dlow=1.29 and dup=1.45. Since d< dlow, so we reject H0 and conclude that the $\rho$ >0 and there is first-order correlation.

```
> reg<-lm(H~P,data=text8.1)
> library(car)
> dwt(reg)
 lag Autocorrelation D-W Statistic p-value
   1       0.6511468     0.6208403       0
 Alternative hypothesis: rho != 0
```

### (b)

H0: there is no autocorrelation ; Ha: there is positive autocorrelation

n1 = # of positive residuals = 14; n2 = # of negative residuals = 11

E(R) = (2*n1*n2)/(n1+n2) + 1 = 13.32;

VAR(R) = (2*n1*n2*(2*n1*n2-n1-n2))/((n1+n2-1)*(n1+n2)^2) = 5.8109

R = # of runs = 6

Z statistics = (R-E(R))/sqrt(var_R) = -3.0366 ~ Z(0.05) = -1.645

Since Z statistics < Z(0.05), we reject H0 and conclude there is a positive autocorrelation.

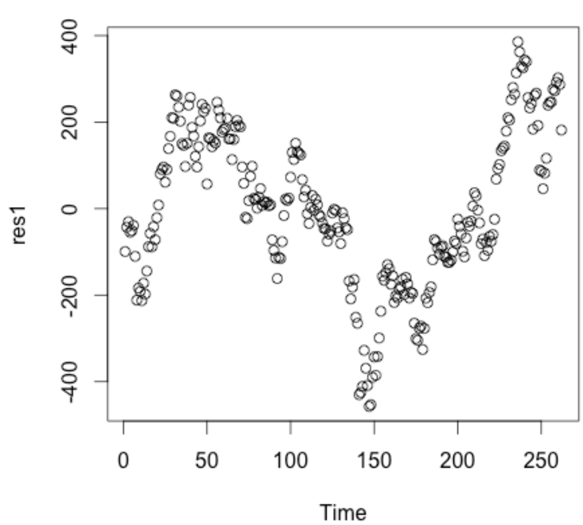
```
> res = resid(reg)
> n1<-0
> n2<-0
> for (i in (1:25)) {if (res[i] > 0) {n1=n1+1} else {n2=n2+1}}
> E_R <- (2*n1*n2)/(n1+n2) + 1
> E_R
[1] 13.32
> var_R <-(2*n1*n2*(2*n1*n2-n1-n2))/((n1+n2-1)*(n1+n2)^2)
> var_R
[1] 5.810933
> R<-6
> Z<- (R-E_R)/sqrt(var_R)
> Z
[1] -3.036604
```

## 2. Text8.4

### (a)

```
> reg1<-lm(DJIA~Time, data=text8.4)
> res1<-resid(reg1)
> plot(Time,res1)
```

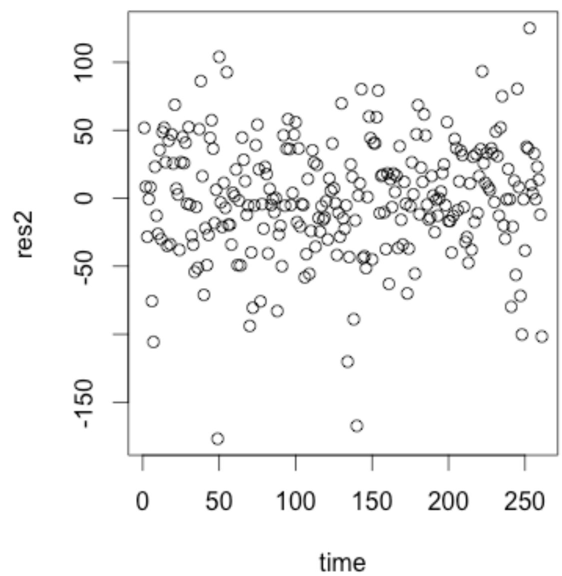
From the plot of Residuals vs. Time, I can know the distribution of resides for time are not random and there is two clear cycles. So, the residuals for time are time dependent.



**(b)**

By considering lag =1, we use $DJIA_{p-1}$ to predict $DJIA_p$. Since there is no such $DJIA_0$ corresponding to $DJIA_1$ and no $DJIA_{p+1}$ corresponding to $DJIA_p$, So I create a new djia variable without $DJIA_1$ and new djia_lag variable without $DJIA_p$.

```
> djia<-text8.4$DJIA
> djia<-djia[-1]
> djia_lag<-text8.4$DJIA
> djia_lag<-djia_lag[-262]
> reg2<-lm(djia~djia_lag)
> summary(reg2)

Call:
lm(formula = djia ~ djia_lag)

Residuals:
     Min       1Q   Median       3Q      Max 
-176.878  -22.397   -0.641   26.478  125.139 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 36.898384  42.989100   0.858    0.392    
djia_lag     0.994459   0.007477 133.002   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 42.37 on 259 degrees of freedom
Multiple R-squared:  0.9856,	Adjusted R-squared:  0.9855 
F-statistic: 1.769e+04 on 1 and 259 DF,  p-value: < 2.2e-16
```



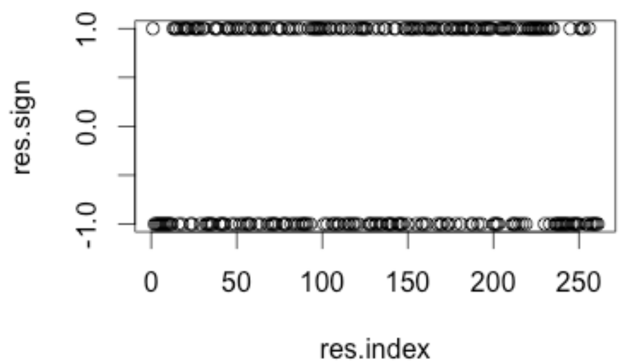

I run the regression model and see $DJIA_{p-1}$ is highly significant including in the model. Therefore, it is an adequate model. Then I check if there is any evidence of autocorrelation in residuals by plot residuals vs. time. The graph shows there is not an autocorrelation. And I did Durbin Watson test to check.

```
> dwt(reg2)
 lag Autocorrelation D-W Statistic p-value
   1       0.1066715      1.758642   0.034
 Alternative hypothesis: rho != 0
```

DW statistics = 1.758642 which is close to 2. So, there is no autocorrelation.

**(c)**

```
> log_djia <- log(djia)
> reg3<-lm(log_djia~djia_lag)
> summary(reg3)

Call:
lm(formula = log_djia ~ djia_lag)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0304723 -0.0040038  0.0002664  0.0047480  0.0187148 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 7.678e+00  7.793e-03   985.2   <2e-16 ***
djia_lag    1.701e-04  1.355e-06   125.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.007681 on 259 degrees of freedom
Multiple R-squared:  0.9838,	Adjusted R-squared:  0.9838 
F-statistic: 1.576e+04 on 1 and 259 DF,  p-value: < 2.2e-16

> library(car)
> dwt(reg3)
 lag Autocorrelation D-W Statistic p-value
   1       0.1762719      1.619276   0.002
 Alternative hypothesis: rho != 0
```




Used log(djia) to run the regression and also see $DJIA_{p-1}$ is highly significant including in the model. Then I check if there is any evidence of autocorrelation in residuals by plot above. The graph shows there is no autocorrelation. So, I did Durbin Watson test to check. DW statistics = 1.619276 which is close to 2. So, there is no autocorrelation.

**3. Text8.5**

(a)

The adequate model is lm(djia~djia_lag) in Problem (b).

```
> reg4<-lm(djia~djia_lag, data = data_130)
> summary(reg4)

Call:
lm(formula = djia ~ djia_lag, data = data_130)

Residuals:
     Min       1Q   Median       3Q      Max 
-170.440  -23.477    0.806   25.819  104.042 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 239.96956  108.59402    2.21   0.0289 *  
djia_lag      0.95732    0.01964   48.74   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.23 on 127 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.9492,	Adjusted R-squared:  0.9488 
F-statistic:  2375 on 1 and 127 DF,  p-value: < 2.2e-16
```

```
> res4<-resid(reg4)
> anova(reg4) #get residual mean square
Analysis of Variance Table

Response: djia
           Df  Sum Sq Mean Sq F value    Pr(>F)    
djia_lag    1 4037704 4037704  2375.2 < 2.2e-16 ***
Residuals 127  215891    1700
```

The residual mean square is **1700**.

(b)

After predict first 15 days of July, we get predicts and also the predict error (the difference between predict and actual DJIA).

```
> data_15<-subset(text8.4,Time<=145)
> data_15<-subset(data_15,Time>=131)
> for (i in 1:15 ) {data_15$predict[i] = 239.96956 + 0.95732*data_15$djia_lag[i]}
> #c
> for (i in 1:15) {data_15$pre_error[i]<-data_15$djia[i]-data_15$predict[i]}
> SSE =0
> for (i in 1:15) {SSE<-SSE+(data_15$pre_error[i])^2}
> SSE
[1] 63856.89
> ave_SSE<-SSE/15
> ave_SSE
[1] 4257.126
```

| | row.names | Date | djia | Time | djia_lag | predict | pre_error |
|---|---|---|---|---|---|---|---|
| 1 | 131 | 7/1/96 | 5729.98 | 131 | 5654.63 | 5653.260 | 76.720048 |
| 2 | 132 | 7/2/96 | 5720.38 | 132 | 5729.98 | 5725.394 | -5.014014 |
| 3 | 133 | 7/3/96 | 5703.02 | 133 | 5720.38 | 5716.204 | -13.183742 |
| 4 | 134 | 7/4/96 | 5703.02 | 134 | 5703.02 | 5699.585 | 3.435334 |
| 5 | 135 | 7/5/96 | 5588.14 | 135 | 5703.02 | 5699.585 | -111.444666 |
| 6 | 136 | 7/8/96 | 5550.83 | 136 | 5588.14 | 5589.608 | -38.777745 |
| 7 | 137 | 7/9/96 | 5581.86 | 137 | 5550.83 | 5553.890 | 27.969864 |
| 8 | 138 | 7/10/96 | 5603.65 | 138 | 5581.86 | 5583.596 | 20.054225 |
| 9 | 139 | 7/11/96 | 5520.50 | 139 | 5603.65 | 5604.456 | -83.955778 |
| 10 | 140 | 7/12/96 | 5510.56 | 140 | 5520.50 | 5524.855 | -14.294620 |
| 11 | 141 | 7/15/96 | 5349.51 | 141 | 5510.56 | 5515.339 | -165.828859 |
| 12 | 142 | 7/16/96 | 5358.76 | 142 | 5349.51 | 5361.162 | -2.402473 |
| 13 | 143 | 7/17/96 | 5376.88 | 143 | 5358.76 | 5370.018 | 6.862317 |
| 14 | 144 | 7/18/96 | 5464.18 | 144 | 5376.88 | 5387.364 | 76.815678 |
| 15 | 145 | 7/19/96 | 5426.82 | 145 | 5464.18 | 5470.938 | -44.118358 |

(c)

From the code in (b), we get the SSE = **63856.89**. Average of squared predict error = SSE/(n) = **4257.126**. Residual mean square =1700 is much less than Average of squared predict error.

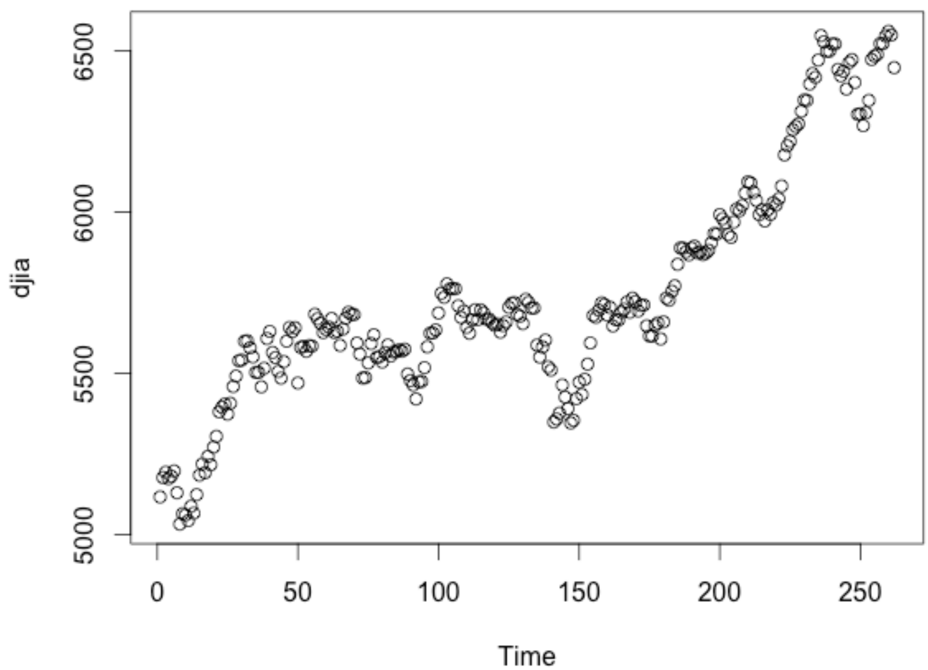
(d)

```
> data_132<-subset(text8.4,Time>130)
> for (i in 1:132 ) {data_132$predict[i] = 239.96956 + 0.95732*data_132$djia_lag[i]}
> for (i in 1:132) {data_132$pre_error[i]<-data_132$djia[i]-data_132$predict[i]}
> SSE =0
> for (i in 1:132) {SSE<-SSE+(data_132$pre_error[i])^2}
> SSE
[1] 319809.7
> ave_SSE<-SSE/132
> ave_SSE
[1] 2422.801
```

After predict second half of the year, we get predicts and also the predict error (the difference between predict and actual DJIA).

We get the SSE = **319809.7**. Average of squared predict error = SSE/(n) = **2422.801**.

Residual mean square =1700 is still much less than Average of squared predict error.
(e)




The scatter plot of Time vs. DJIA shows there is a significant drop of DJIA with first 15 days of July and the increasing trend is very smooth for DJIA in second half of the year. This is the reason the Average of squared predict error of DJIA within first 15 days of July **is much larger** than Average of squared predict error of DJIA within second half of the year.

### 4. Text11.5

**(a)**

No. Since all predictive variables, except X1, are not significant adding the model. Also, adjust R^2 = 75.55% is not good enough.

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9,
    data = text11.5)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7729 -1.9801 -0.0868  1.6615  4.2618

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.31044    5.96093   2.568   0.0223 *
X1           1.95413    1.03833   1.882   0.0808 .
X2           6.84552    4.33529   1.579   0.1367
X3           0.13761    0.49436   0.278   0.7848
X4           2.78143    4.39482   0.633   0.5370
X5           2.05076    1.38457   1.481   0.1607
X6          -0.55590    2.39791  -0.232   0.8200
X7          -1.24516    3.42293  -0.364   0.7215
X8          -0.03800    0.06726  -0.565   0.5810
X9           1.70446    1.95317   0.873   0.3976
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.973 on 14 degrees of freedom
Multiple R-squared:  0.8512,    Adjusted R-squared:  0.7555
```

**(b)**

```
Call:
lm(formula = Y ~ X1 + X6 + X8, data = text11.5)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.7486 -2.4082 -0.3594  2.1378  6.5353 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 14.796013   4.971105   2.976 0.007462 ** 
X1           3.489464   0.729368   4.784 0.000113 ***
X6          -0.415515   1.182262  -0.351 0.728921    
X8           0.004923   0.063597   0.077 0.939062    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.123 on 20 degrees of freedom
Multiple R-squared:  0.7655,	Adjusted R-squared:  0.7303 
F-statistic: 21.76 on 3 and 20 DF,  p-value: 1.653e-06
```

No. Since X6 (Number of rooms) and X8 (Age of the home) are highly insignificant in the model since their P-values are very large; and adjusted R^2 is low. So the model does not adequately describe the sale price.

**(c)**

No. After tried several model, I found a model works better than just adding Taxes as a predictive variable. Left side is the better model; all variables are significant including in the model with adjusted R^2 80.23%. Right side is the model with only Taxes as predictive variable, which has a lower adjusted R^2 75.3%. So, the new model works better.

```
> reg2<-lm(Y~X1+X2+X5+X2:X7, data = text11.5)
> summary(reg2)

Call:
lm(formula = Y ~ X1 + X2 + X5 + X2:X7, data = text11.5)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.2039 -2.4101 -0.0615  2.0829  3.8468 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   6.0241     3.4729   1.735 0.099007 .  
X1            2.4923     0.5125   4.863 0.000108 ***
X2           14.9721     5.4718   2.736 0.013121 *  
X5            2.0956     1.1930   1.757 0.095089 .  
X2:X7        -2.0201     1.0706  -1.887 0.074562 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.674 on 19 degrees of freedom
Multiple R-squared:  0.8367,	Adjusted R-squared:  0.8023 
F-statistic: 24.33 on 4 and 19 DF,  p-value: 2.993e-07
```

```
Call:
lm(formula = Y ~ X1)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.8445 -2.3340 -0.3841  1.9689  6.3005 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  13.3553     2.5955   5.146 3.71e-05 ***
X1            3.3215     0.3939   8.433 2.44e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.988 on 22 degrees of freedom
Multiple R-squared:  0.7637,	Adjusted R-squared:  0.753 
F-statistic: 71.11 on 1 and 22 DF,  p-value: 2.435e-08
```

**5. Text11.6**
**(a)**
No. As the result show of the regression model including all predictive variable, only X5 is significant in the model. Adjust R^2 is not good. So, I would not include all the variables to predict the gasoline consumption of the cars.

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 +
    X10 + factor(X11), data = text9.3)

Residuals:
    Min      1Q  Median      3Q     Max
-5.3498 -1.6236 -0.6002  1.5155  5.2815

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.773204  30.508775   0.583   0.5674
X1           -0.077946   0.058607  -1.330   0.2001
X2           -0.073399   0.088924  -0.825   0.4199
X3            0.121115   0.091353   1.326   0.2015
X4            1.329034   3.099535   0.429   0.6732
X5            5.975989   3.158647   1.892   0.0747 .
X6            0.304178   1.289094   0.236   0.8161
X7           -3.198576   3.105435  -1.030   0.3167
X8            0.185362   0.129252   1.434   0.1687
X9           -0.399146   0.323812  -1.233   0.2336
X10          -0.005193   0.005893  -0.881   0.3898
factor(X11)1  0.598655   3.020681   0.198   0.8451
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.226 on 18 degrees of freedom
Multiple R-squared:  0.8353,	Adjusted R-squared:  0.7346
F-statistic: 8.297 on 11 and 18 DF,  p-value: 5.287e-05
```

**(b)**
Among these regression models, I would choose regression of Y on X8, X5 and X10. All the three variables are significant including in the model and the model has highest adjusted R^2 78.08% among these given models.

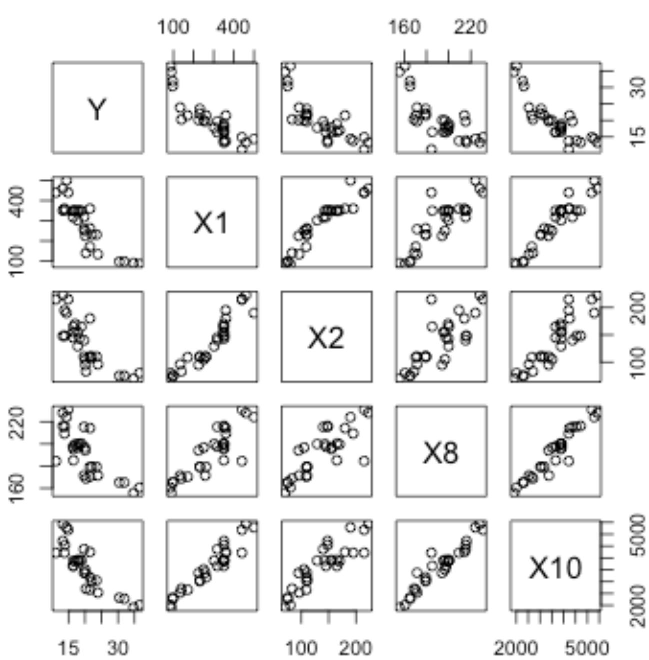
```
Call:
lm(formula = Y ~ X8 + X5 + X10)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5930 -1.9674 -0.6438  2.0314  5.8823

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.494972  11.764747   0.382   0.7055
X8           0.218119   0.087764   2.485   0.0197 *
X5           2.607338   1.263791   2.063   0.0492 *
X10         -0.009482   0.001993  -4.759 6.35e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.932 on 26 degrees of freedom
Multiple R-squared:  0.8035,	Adjusted R-squared:  0.7808
F-statistic: 35.43 on 3 and 26 DF,  p-value: 2.47e-09
```

I found a better model, which shows below. Two interaction terms added in the model and all variable s are significant in the model. The adjust R^2 88.15% is higher.

```
Call:
lm(formula = Y ~ X8 + X5 + X10 + X1:X3 + X1:X7)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.4577 -0.9720  0.1114  1.3833  3.3444 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.124e+01  1.007e+01   2.110 0.045487 *  
X8           2.104e-01  7.673e-02   2.742 0.011357 *  
X5           2.091e+00  9.465e-01   2.209 0.036942 *  
X10         -8.925e-03  2.762e-03  -3.232 0.003554 ** 
X1:X3        1.982e-04  4.123e-05   4.808 6.76e-05 ***
X1:X7       -3.337e-02  7.741e-03  -4.311 0.000239 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.156 on 24 degrees of freedom
Multiple R-squared:  0.9019,    Adjusted R-squared:  0.8815 
F-statistic: 44.15 on 5 and 24 DF,  p-value: 2.486e-11
```

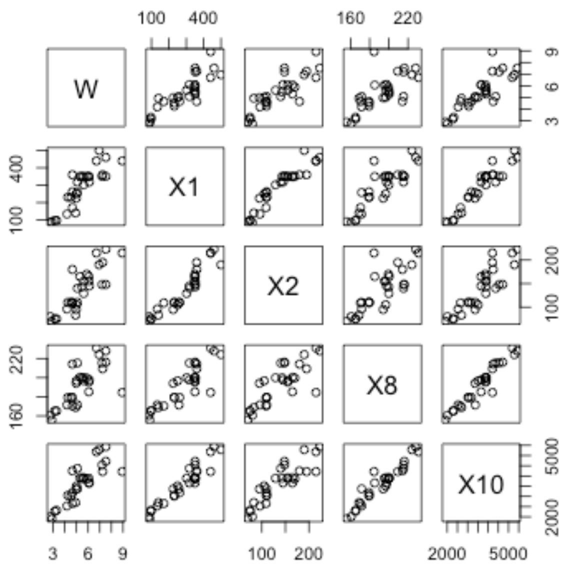
**(c)**

Yes. Form the plot of Y vs. X1, X2, X8, X10 (first row of scatter matrix), we can see there is a exponential trend in each plot. So, Y vs X1, Y vs. X2, Y vs. X8 and Y vs. X10 have a exponential relationship, not a linear relationship.



**(d)**

Yes. Form the plot of W vs. X1, X2, X8, X10 (first row of scatter matrix), we can see there is a linear trend in each plot. So, the relationship between W and the 11 predictor variables is more linear than that between Y and the 11 predictor variables.




**(e)**

After replacing Y by W in these regression model, X2 becomes significant in the model W~X2+X10. But, X5 becomes insignificant in the model W~X5+X8+X10. So, the relationship between W and the 11 predictor variable is more linear, but **not** a linear relationship of W vs. each variables, like W vs. X5.

```
lm(formula = W ~ X2 + X10)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.82734 -0.39360 -0.05088  0.31576  1.95887 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept) 0.8397046  0.5651506   1.486   0.1489  
X2          0.0147369  0.0067923   2.170   0.0390 *
X10         0.0007026  0.0003223   2.180   0.0381 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7514 on 27 degrees of freedom
Multiple R-squared:  0.7576,	Adjusted R-squared:  0.7396 
F-statistic: 42.19 on 2 and 27 DF,  p-value: 4.919e-09
```

```
Call:
lm(formula = Y ~ X2 + X10)

Residuals:
   Min     1Q Median     3Q    Max 
-5.389 -2.155 -1.266  3.044  7.575 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 40.203322   2.516464  15.976 2.77e-15 ***
X2          -0.026227   0.030244  -0.867  0.39349    
X10         -0.004569   0.001435  -3.184  0.00364 ** 
---
```

```
Call:
lm(formula = W ~ X8 + X5 + X10)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.1439 -0.3867 -0.0915  0.4313  1.3642 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.8185512  2.6878166   3.281 0.002946 ** 
X8          -0.0743329  0.0200508  -3.707 0.000998 ***
X5           0.0778394  0.2887303   0.270 0.789602    
X10          0.0029363  0.0004552   6.450 7.79e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6699 on 26 degrees of freedom
Multiple R-squared:  0.8144,    Adjusted R-squared:  0.793 
F-statistic: 38.04 on 3 and 26 DF,  p-value: 1.178e-09
```

```
Call:
lm(formula = Y ~ X8 + X5 + X10)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5930 -1.9674 -0.6438  2.0314  5.8823 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.494972  11.764747   0.382   0.7055    
X8           0.218119   0.087764   2.485   0.0197 *  
X5           2.607338   1.263791   2.063   0.0492 *  
X10         -0.009482   0.001993  -4.759 6.35e-05 ***
```

So, among given regression by the question, I think lm(W ~ X8+W10) is the model. Since all predict variables are significant and adjust R^2 is 80.01%, which is the highest.

```
lm(formula = W ~ X8 + X10)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.1512 -0.3578 -0.1174  0.4314  1.3869 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.1675393  2.3147462   3.960 0.000491 ***
X8          -0.0744971  0.0196944  -3.783 0.000784 ***
X10          0.0029144  0.0004402   6.621 4.19e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6583 on 27 degrees of freedom
Multiple R-squared:  0.8139,    Adjusted R-squared:  0.8001 
F-statistic: 59.05 on 2 and 27 DF,  p-value: 1.384e-10
```

After tried many different models, I think lm(W ~ X8+W10) is the best among them. I didn't find a better model than lm(W ~ X8+W10).

**(f)**

Regressed Y on X13, we can see X3 is highly significant in the model and with very high adjusted R^2 88.82%. So we definitely should include the X13 into the final model to predict the gasoline consumption of the cars.

```
Call:
lm(formula = Y ~ X13)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.7134 -1.2457 -0.0227  1.4211  4.3458 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -11.37       2.10  -5.415 8.93e-06 ***
X13           566.00      37.21  15.213 4.59e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.094 on 28 degrees of freedom
Multiple R-squared:  0.8921,    Adjusted R-squared:  0.8882 
F-statistic: 231.4 on 1 and 28 DF,  p-value: 4.587e-15
```