Report 2

Mufitcan Atalay

5/7/2017

```
#We load the data for the report.
load("/Users/mufitcan/Downloads/Stat_224.Rdata")
```

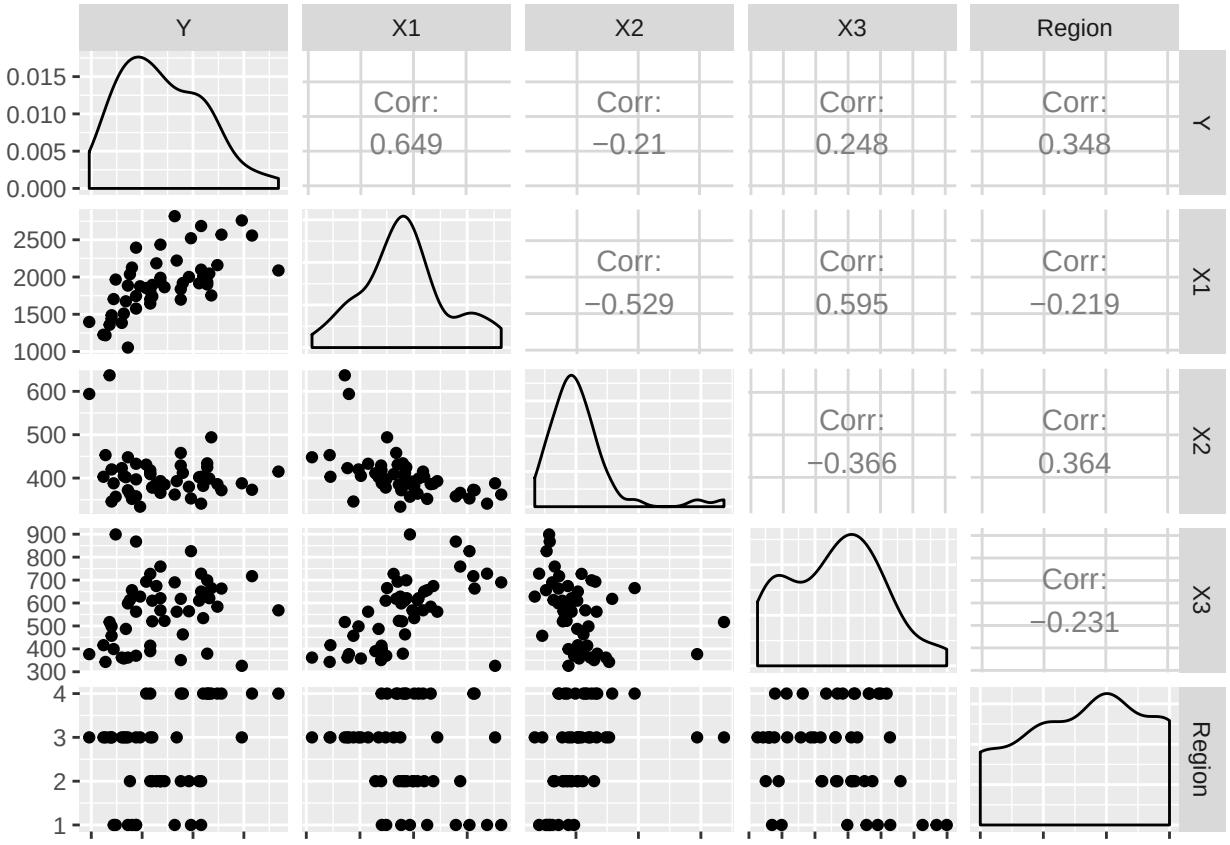
Question 1 ~ Problem 5.4 (modified)

a)

```
#We check model assumptions for 1960
fit1 <- lm(Y ~ X1 + X2 + X3 , P151)
```

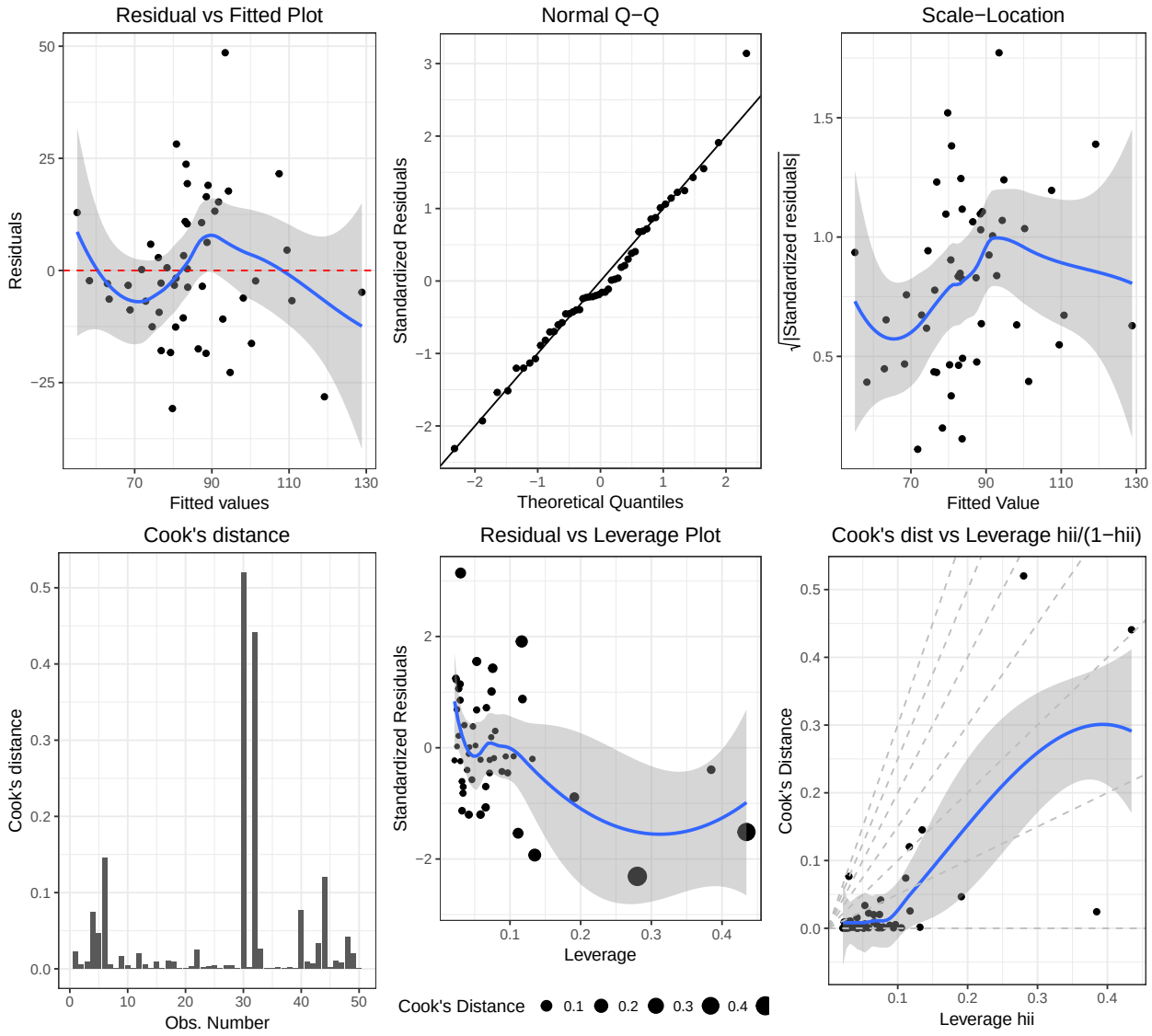
Checking collinearity:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-11.4046270	28.9761647	-0.3935865	0.6957048
X1	0.0449327	0.0076673	5.8602912	0.0000005
X2	0.0662226	0.0488342	1.3560686	0.1816970
X3	-0.0289536	0.0192932	-1.5007151	0.1402625



Variance Inflation Factor

```
##       X1       X2       X3
## 1.874118 1.397080 1.557394
```



Model Form: Plot of Residuals vs Fitted indicates that the form is indeed linear.

Errors: The points do not deviate from the straight line in the normal Q-Q plot. Normality holds. One observation may be an outlier, but scale-location plot makes sure that the model is adequate and random.

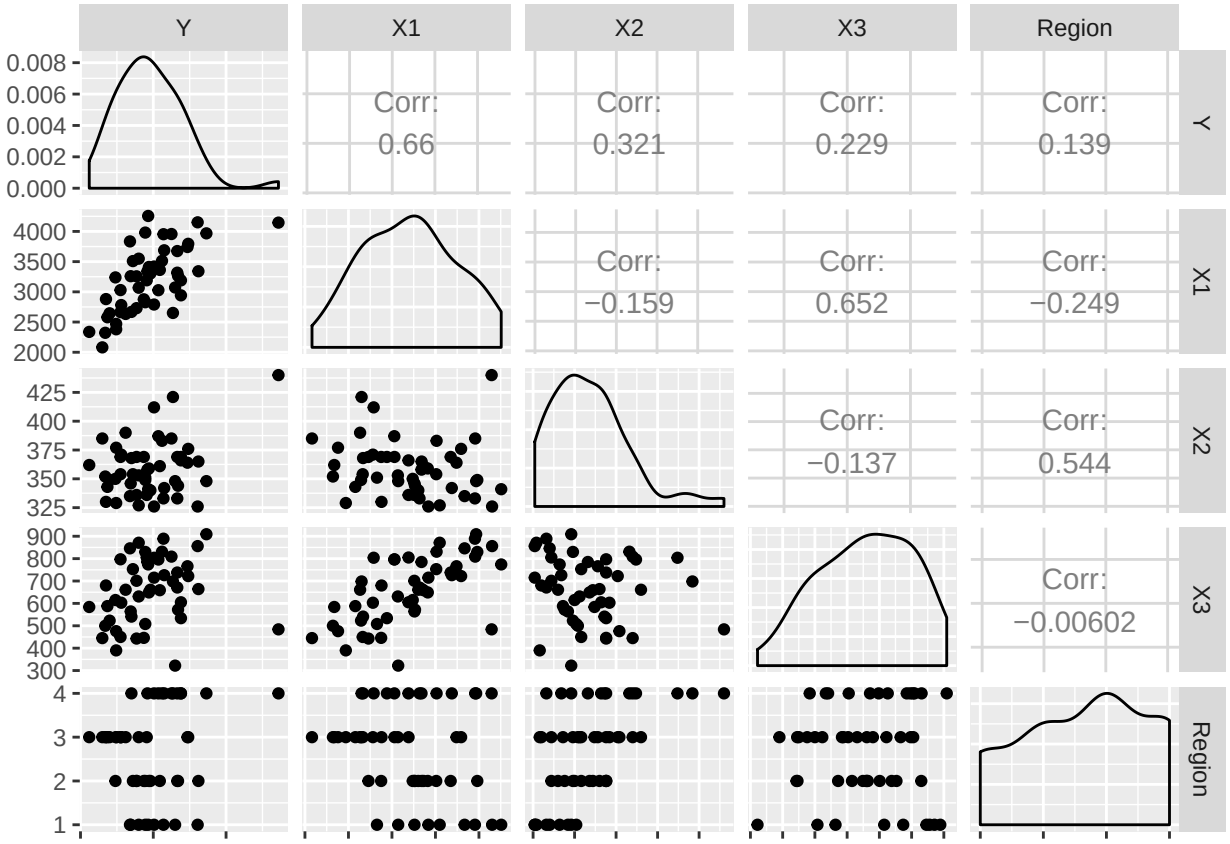
Predictor Independence: X1 seems to correlate slightly with both X3 and X2, but we can look at both the graphs and check VIF to see that this isn't a problem.

Observations: There are no observation with a Cook's Distance close to 1. There are two slightly more influential points, but they are close to 0.5, which is not worrying.

```
#We check model assumptions for 1970
fit2 <- lm(Y ~ X1 + X2 + X3 , P152)
```

Checking collinearity:

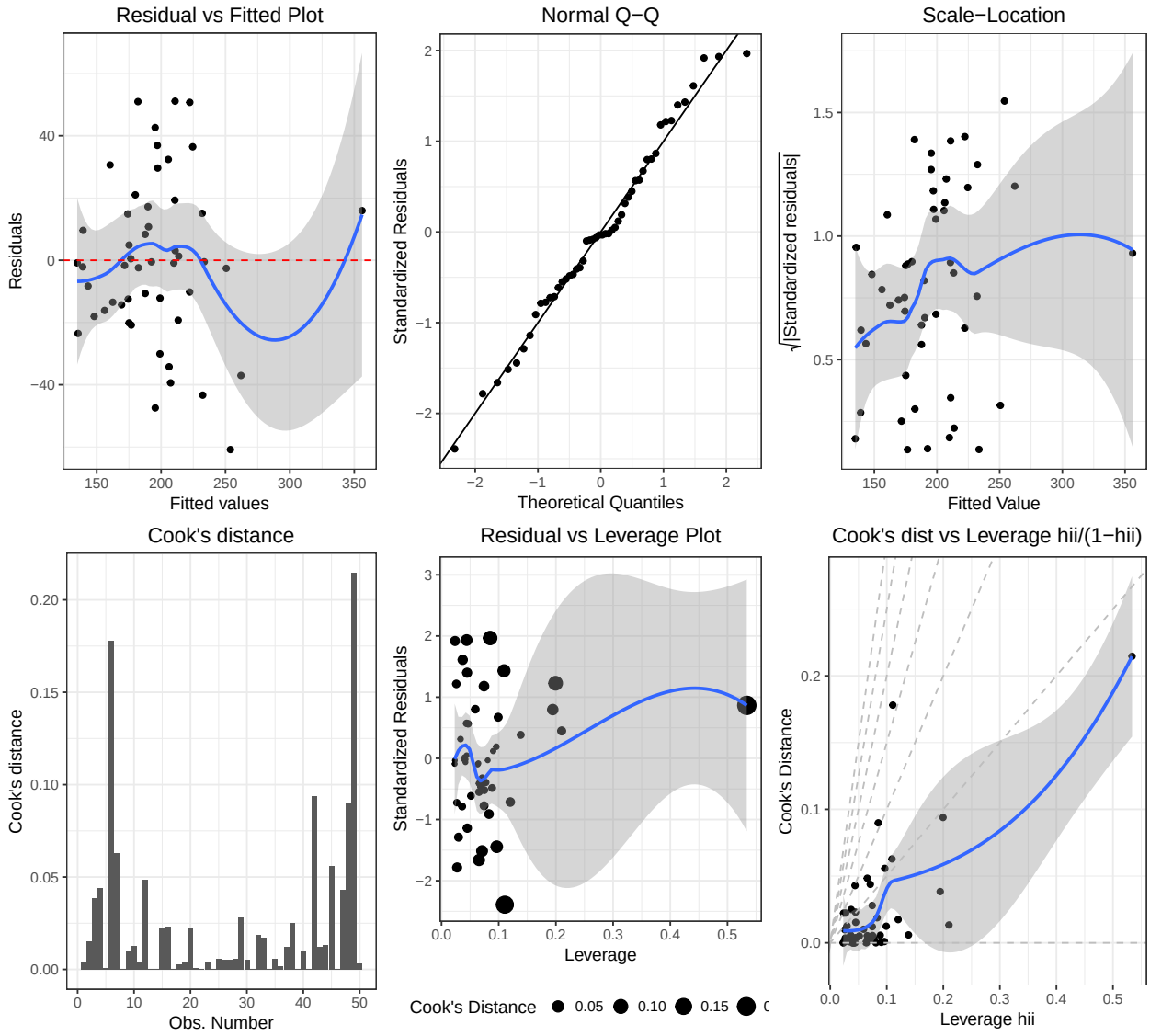
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-289.1792536	66.1695625	-4.370276	0.0000701
X1	0.0808861	0.0094739	8.537785	0.0000000
X2	0.8184112	0.1616285	5.063534	0.0000071
X3	-0.1037663	0.0350621	-2.959501	0.0048565



Variance Inflation Factor

```
vif(fit2)
```

```
##       X1       X2       X3
## 1.753440 1.028002 1.741542
```



Model Form: No significant pattern present in the residuals vs. fitted. Model is adequatly linear.

Errors: The points do not deviate from the straight line in the normal Q-Q plot and scale-location plot does not have a pattern, making sure that the model is adequate and random. Normality holds.

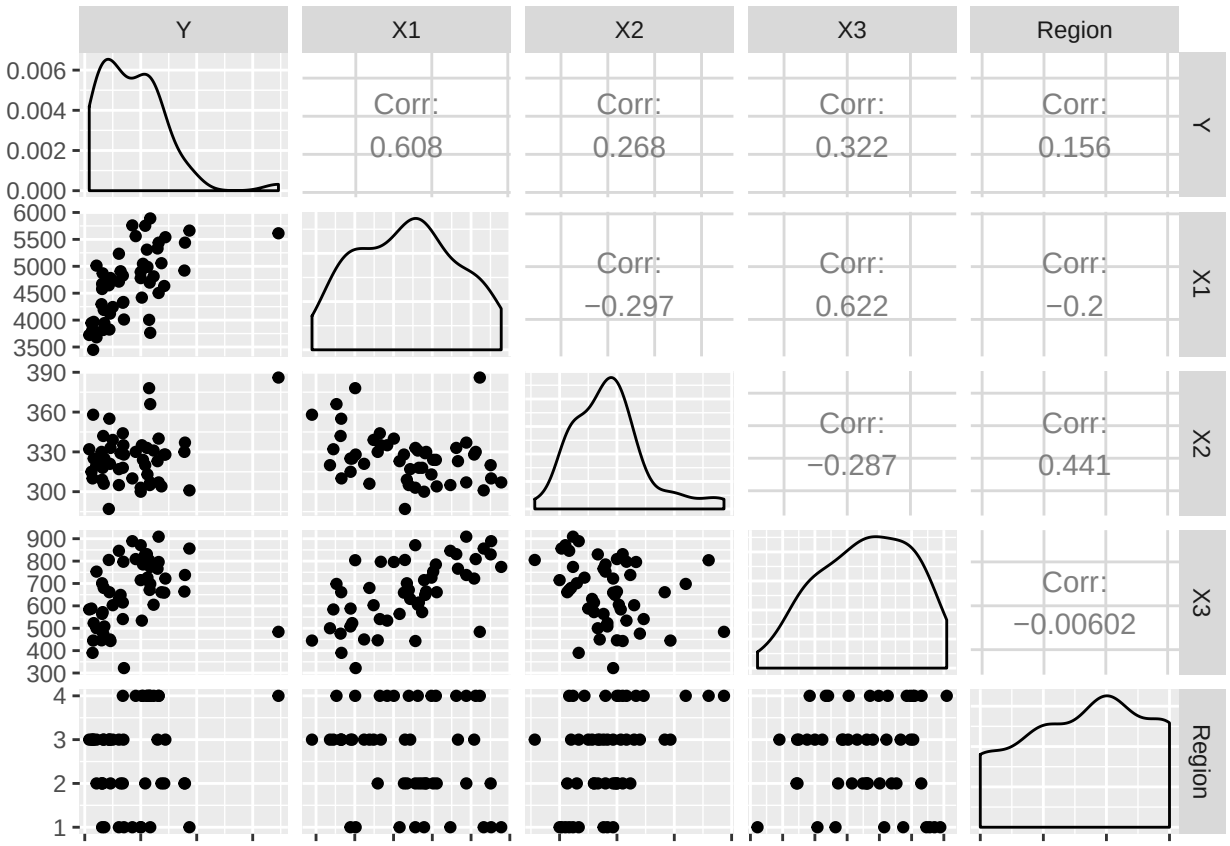
Predictor Independence: X1 and X3 has a worryingly high correlation of 0.652. We cannot count these two predictors as independent.

Observations: There are no observation with a Cook's Distance close to 1. Equality of influence holds.

```
#We check model assumptions for 1975
fit3 <- lm(Y ~ X1 + X2 + X3 , P153)
```

Checking collinearity:

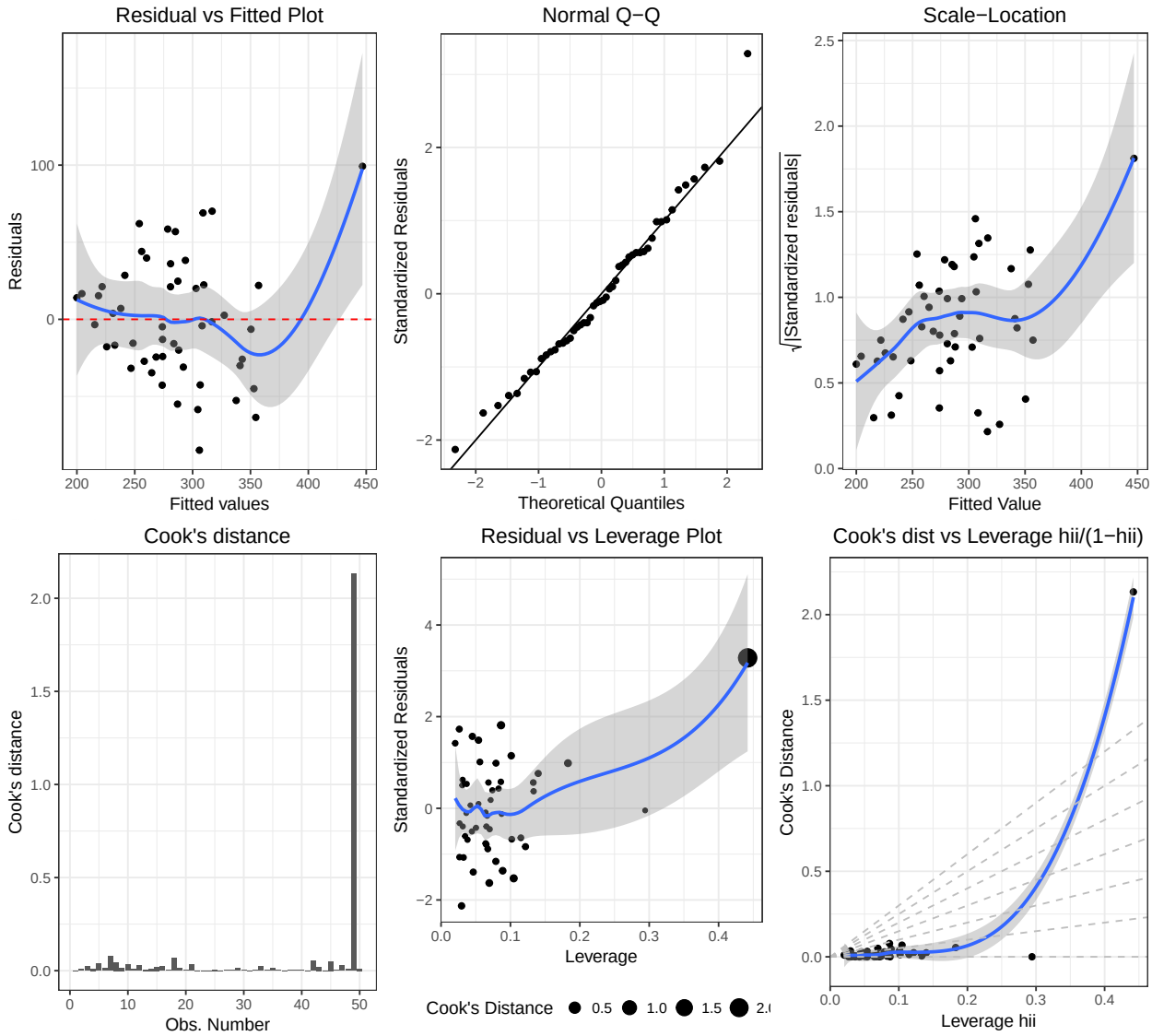
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-556.5680446	123.1952504	-4.5177719	0.0000434
X1	0.0723853	0.0116024	6.2387998	0.0000001
X2	1.5520545	0.3146716	4.9322989	0.0000110
X3	-0.0042690	0.0513929	-0.0830669	0.9341588



Variance Inflation Factor

```
vif(fit3)
```

```
##       X1       X2       X3
## 1.672775 1.117475 1.661592
```



Model Form: No significant patter, but clearly a point with a high leverage makes the residuals vs fitted plot look skewed. Assumption still satisfied. Ted.

Errors: The points do not deviate from the straight line in the normal Q-Q plot and scale-location plot does not have a pattern, making sure that the model is adequate and random. Normality holds.

Predictor Independence: No indication of corellation between predictors from either the corelation matrix or VIF. Assumption holds.

Observations: There is an observation with a Cook's distance clearly above 1. This is obseravtion 49. We might want to remove this point considering that all other points have very low Cook's distance.

b) Testing affects on Y over time

We combine all data and create indicator variables so that $T_1 = 1$ if observation comes from 1960, and $T_2 = 1$ if observatin comes from 1970.

```
#Creating a combined model
fm <- lm(Y~X1*T1+X1*T2+X2*T1+X2*T2+X3*T1+X3*T2, data = data)

kable(coef(summary(fm)))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-556.5680446	89.8137456	-6.1969139	0.0000000
X1	0.0723853	0.0084586	8.5576044	0.0000000
T1	545.1634177	105.0409324	5.1900093	0.0000007
T2	267.3887911	115.3550229	2.3179640	0.0219216
X2	1.5520545	0.2294069	6.7655101	0.0000000
X3	-0.0042690	0.0374672	-0.1139408	0.9094504
X1:T1	-0.0274526	0.0167121	-1.6426760	0.1027267
X1:T2	0.0085008	0.0133780	0.6354309	0.5261997
T1:X2	-1.4858319	0.2470931	-6.0132478	0.0000000
T2:X2	-0.7336433	0.2896438	-2.5329151	0.0124306
T1:X3	-0.0246845	0.0521459	-0.4733749	0.6366937
T2:X3	-0.0994972	0.0536203	-1.8555904	0.0656452

We now do a hypothesis test to check whether the coefficients of all indicatorand their interaction variables are 0.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-141.8916057	40.5925225	-3.495511	0.0006269
X1	0.0802080	0.0034482	23.261056	0.0000000
X2	0.2978954	0.0843104	3.533317	0.0005499
X3	-0.0633980	0.0220906	-2.869909	0.0047162

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
146	185764.3	NA	NA	NA	NA
138	120140.2	8	65624.15	9.422465	0

We see from the anova table that the F-statistic of our RM and FM is 9.42.

We have $n = 150$, $p = 12$, and $k = 4$, thus we need the critical value $F_{9,137;0.05} = 1.9488449$

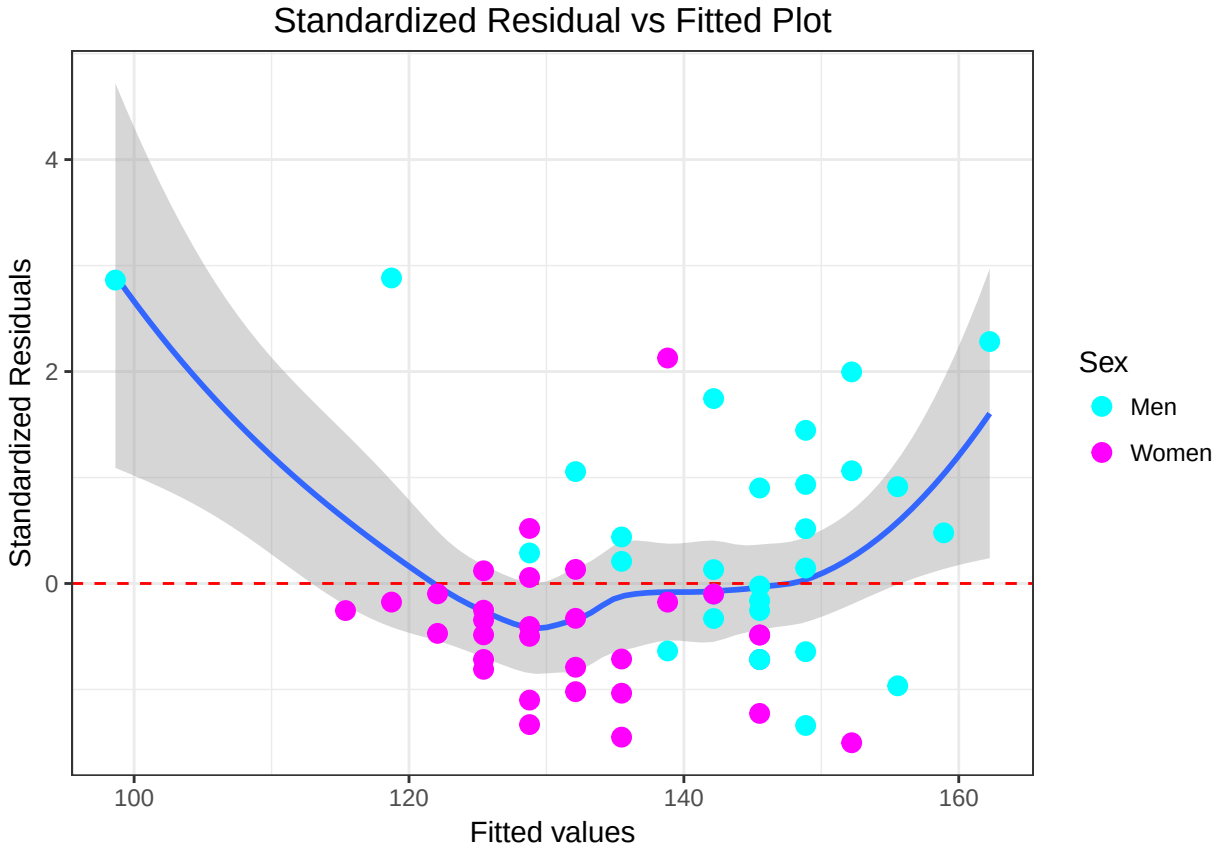
$$9.42 > 1.9488$$

Therefore we can reject the null hypothesis that our reduced model is adequate at a 5% significance level. Therefore we can say that the regression coefficients are not stable over time. More specifically, it can be said that variance in the three data sets is not equivalent. From a practical point of view this is a little bit suprising because one would expect the effect of per capita personal income, number of resident per thousand under 18 years of age and the number of people per thousand residing in urban areas to stay roughly constant over time. It seems as though the effect of technology was greater than expected.

Question 2 ~ Problem 5.8 (modified)

a) I think that picking weight as the response variable and height as the predictor variable is an adequate choice. It is often the case that taller people weigh more than shorter people. Taller people are just larger (volumewise) and mass of humans roughly stays the same, so one would expect a relatively strong positive relationship between height and weight. In terms of our interest, the validation or invalidation of the relationship described would tell us information about the presence of obesity and other such diseases.

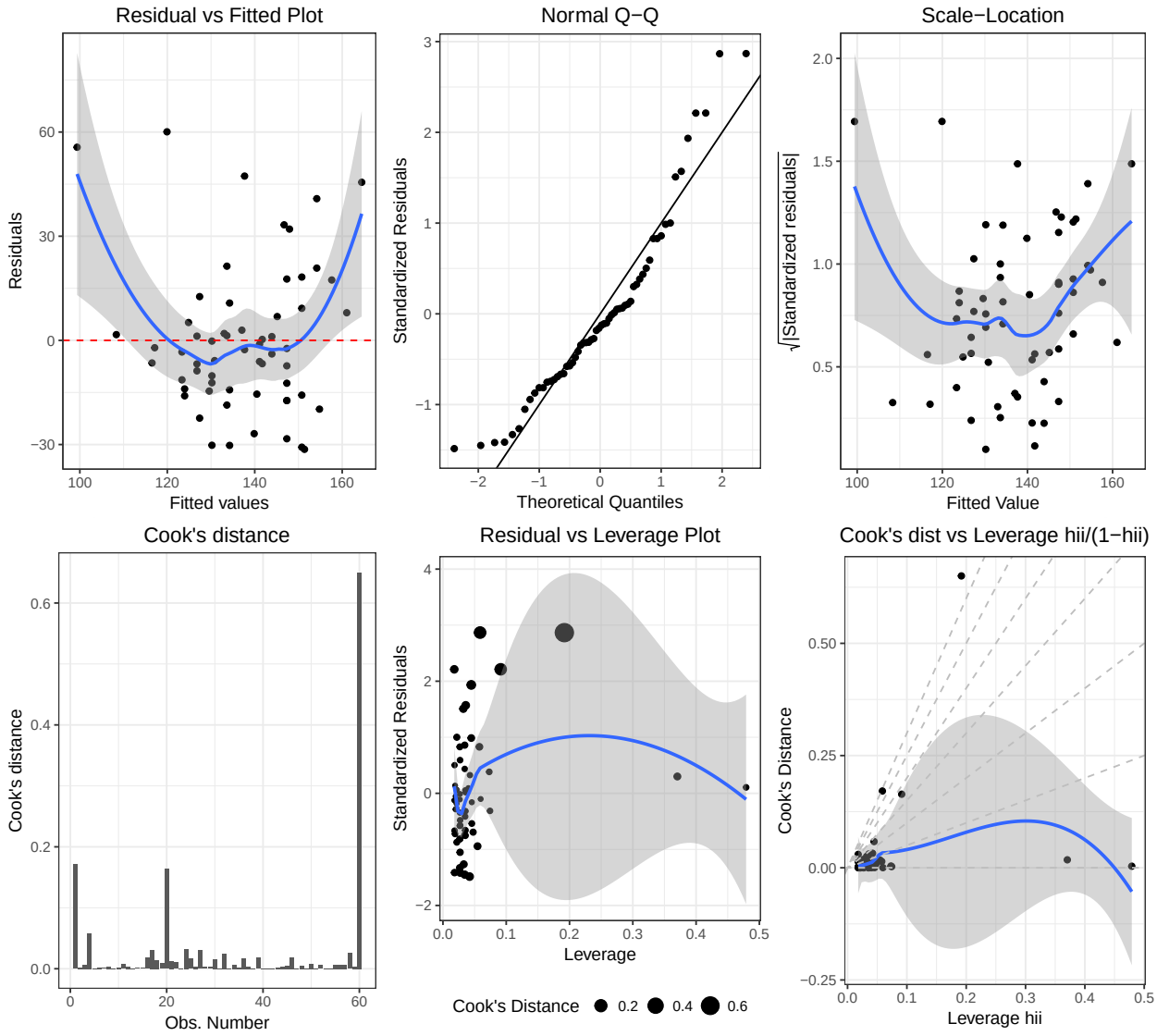
b) To create a pooled linear regression between the height and weight, we just ignore the indicator variable for gender and age variable.



We see that there is a clear pattern and that therefore we cannot conclude that a simple linear regression is adequate to describe the relationship between height and weight for the two groups of students together. It also seems that data points which belong to Men are the points distorting the model.

c)

```
#We first create models using age as an additional variable
fitage <- lm(Weight ~ Height + Age , P159)
```



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-35.657166	60.2146335	-0.5921678	0.5560800
Height	3.425939	0.7783423	4.4015836	0.0000477
Age	-2.810253	1.7284551	-1.6258759	0.1094929

We can clearly see that adding age does not help the model better describe the relationship between height and weight as the Scale-Location plot of this model doesn't make any significant modifications to our answer to ***b)***. We can also check this by doing an F-Test.

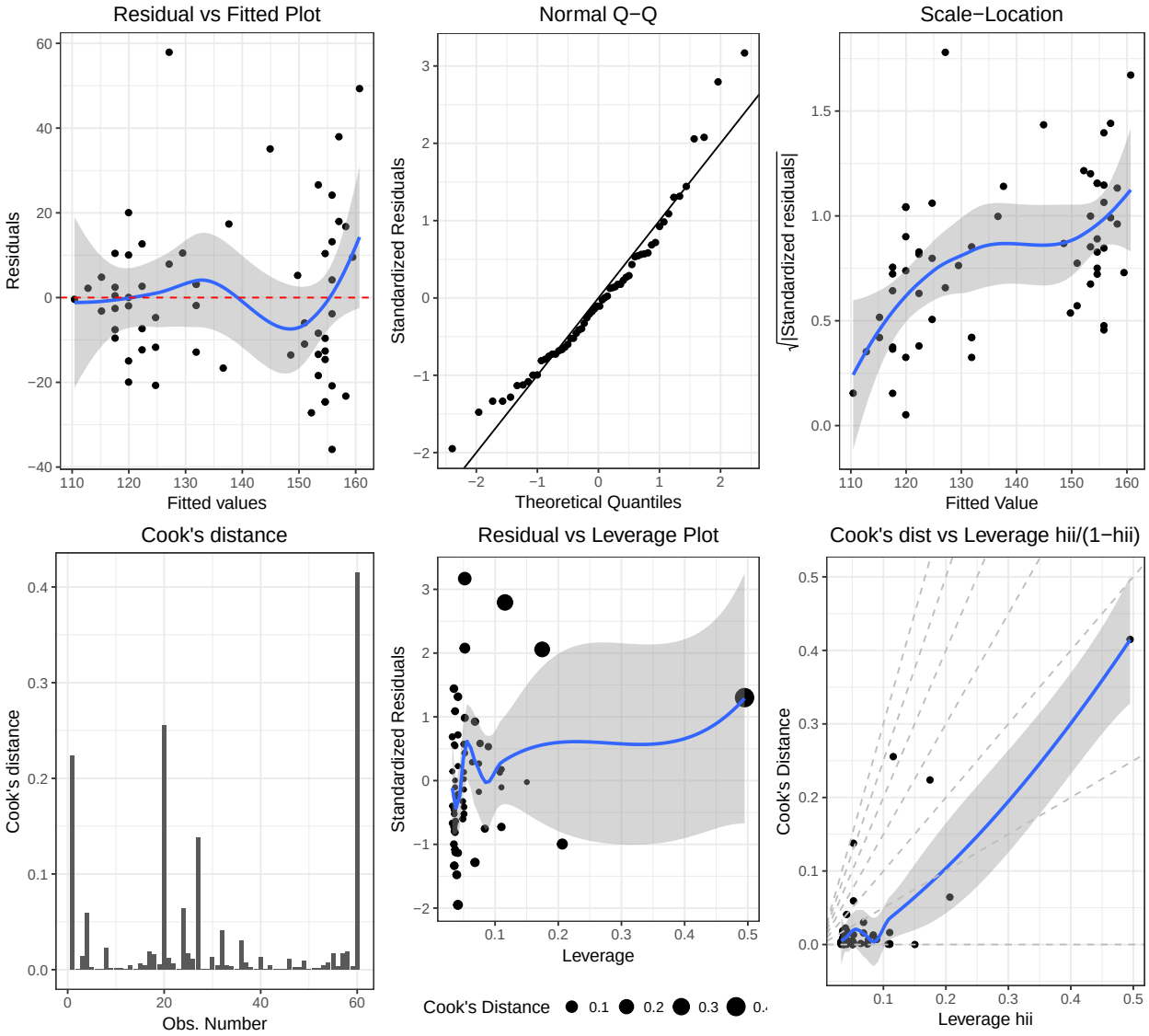
```
anova(fit4,fitage)
```

```
## Analysis of Variance Table
##
## Model 1: Weight ~ Height
## Model 2: Weight ~ Height + Age
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     58 27796
```

```
## 2     57 26564  1    1231.9 2.6435 0.1095
```

We can see from the table above that we have an insignificant p-value and therefore we can conclude that the Reduced model with just height is adequate and that age is not important enough as a variable to be included in the model.

```
#Now we create model using our indicator variable for sex
fitsex <- lm(Weight ~ Height*Sex + Height*(-Sex+1) , P159)
```



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	70.940937	65.1761323	1.0884496	0.2810589
Height	1.212664	0.9517921	1.2740851	0.2078955
Sex	-103.590520	106.6489858	-0.9713221	0.3355648
Height:Sex	1.171714	1.6098252	0.7278516	0.4697369

We can clearly see that using an indicator variable to differentiate between the two sexes does help the model better describe the relationship between height and weight as the Scale-Location plot of this new model does not have as clear of a pattern as ***b)***. We can also check this by doing an F-Test.

```r
anova(fit4,fitsex)
```

```
## Analysis of Variance Table
##
## Model 1: Weight ~ Height
## Model 2: Weight ~ Height * Sex + Height * (-Sex + 1)
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     58 27796
## 2     56 19726  2    8069.9 11.455 6.754e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see from the table above that including Sex as a variable in the model significantly improves it from the p-value. Therefore we can conclude that the full model with Sex as a variable is better than the reduced model and is also the best model. From a practical standpoint this makes a lot sense since men often weigh more than women by virtue of size.

Question 3 ~ Problem 5.9 (modified)

a) We have the following variables:

Variable	Definition
Year	Election Year
V	Democratic share of the two-party presidential vote
I	Indicator Variable Incumbent (1 for Democratic incumbent at the time of the election, -1 for Republican)
D	Categorical variable, 1 for Dem. incumbent running for election, -1 for Rep. 0, otherwise.
W	Indicator variable (1 for the elections of 1920, 1944, and 1948, otherwise)
G	The growth rate of real per capita GDP in the first three quarters of the election year
P	The absolute value of the growth rate of the GDP deflator in the first 15 quarters of the administration
N	Number of quarters in the first 15 quarters of the admin where growth of GDP is greater than 3.2%

```r
#For D = -1, i.e. Republicans
rep<-subset(P160, D==-1)
fitrepublican <- lm(rep$V ~ rep$I + rep$D + rep$W + rep$G*rep$I + rep$P + rep$N)
kable(coef(summary(fitrepublican)))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.4915479	0.0699803	7.0240884	0.0196723
rep$G	-0.0070726	0.0030434	-2.3239067	0.1457462
rep$P	0.0098535	0.0140221	0.7027151	0.5550126
rep$N	-0.0165744	0.0070560	-2.3489722	0.1432853

```r
#For D = 0, i.e. Independent
indep<-subset(P160, D==0)
fitindependent <- lm(indep$V ~ indep$I + indep$D + indep$W + indep$G*indep$I + indep$P + indep$N)
kable(coef(summary(fitindependent)))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6185746	NaN	NaN	NaN

	Estimate	Std. Error	t value	Pr(>|t|)
indep$I	-0.0198706	NaN	NaN	NaN
indep$W	0.0913852	NaN	NaN	NaN
indep$G	0.0056221	NaN	NaN	NaN
indep$P	0.0011520	NaN	NaN	NaN
indep$N	-0.0277949	NaN	NaN	NaN
indep$I$: $indep$G	0.0126071	NaN	NaN	NaN

```
#For D = 1, i.e. Democrats
dem<-subset(P160, D==0)
fitdemocrat <- lm(dem$V ~ dem$I + dem$D + dem$W + dem$G*dem$I + dem$P + dem$N)
kable(coef(summary(fitdemocrat)))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6185746	NaN	NaN	NaN
dem$I	-0.0198706	NaN	NaN	NaN
dem$W	0.0913852	NaN	NaN	NaN
dem$G	0.0056221	NaN	NaN	NaN
dem$P	0.0011520	NaN	NaN	NaN
dem$N	-0.0277949	NaN	NaN	NaN
dem$I$: demG	0.0126071	NaN	NaN	NaN

The coeffificents for D is always non-existent or 0, because it acts like an additiona to β_0 since it is constant regardless of time given that it is an indicator variable. In this sense we can also say that it has no relationship with V because it is a constant when we use varaible D to separate the dataset.

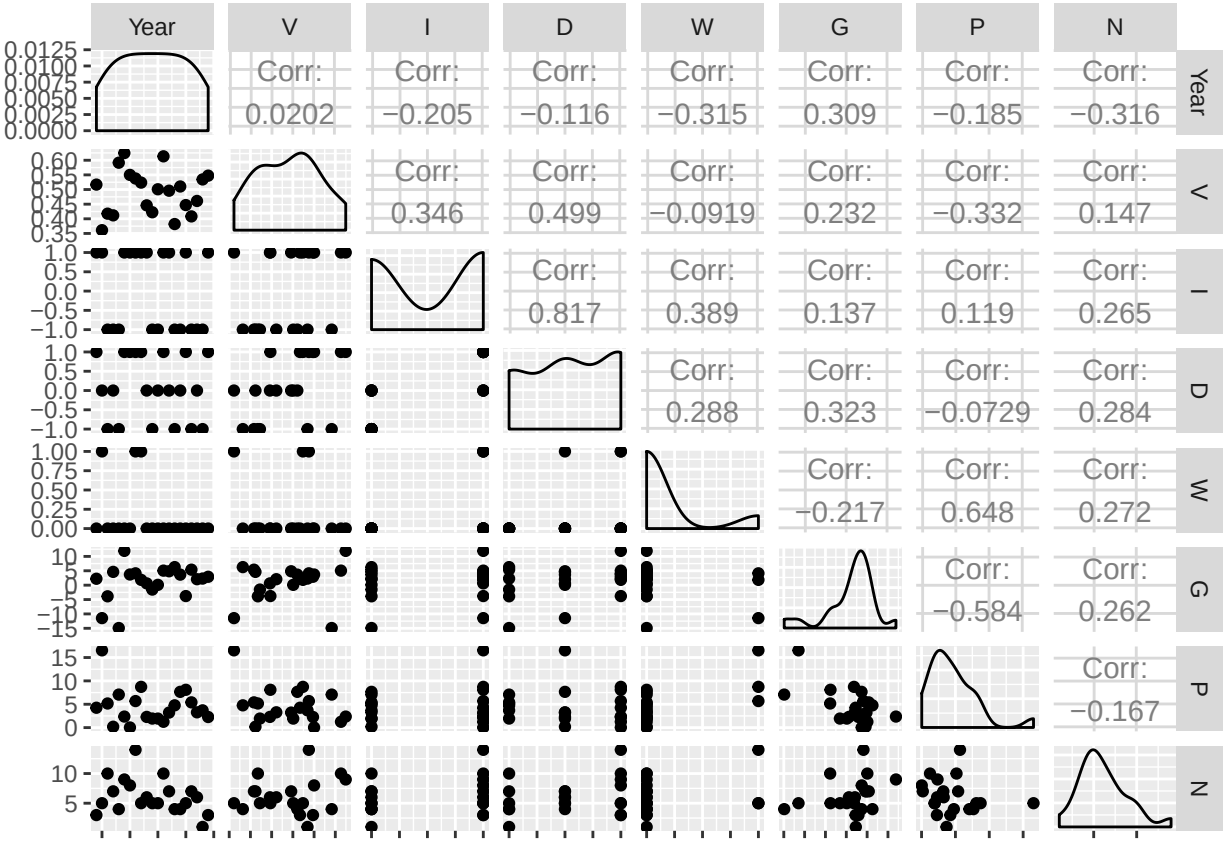
b) The model below is the model given to us in the book

```
fitall <- lm(V~I+W+G*I+N+P+D, data=P160)
vif(fitall)
```

```
##        I        W        G        N        P        D      I:G
## 3.536271 2.720611 2.102990 1.520512 3.648695 3.622490 1.586498
```

We can see that the model has an R-squared value of 0.8148. This is a pretty good fit. We know from the hint in the questions that our aim to decrease the number of predictor variables as much as possible while reamining and adequate fit.

We will proceed to examine the correlation matrix of our dataset. This will tel us which predictors correlate the least with V. We will then attempt to drop these predictors and check whether these new models are adequate using an F-TEST.



It seems as though W has the least correlation with V, so we drop it from the model. Practically this also makes sense because whether the elections were held after or during periods of war is a rather arbitrary decision to make (was invasion of Afghanistan a war? wars don't happen often etc.)

```
#Creating a reduced model without W
rm <- lm(V~I+G*I+N+P+D, data=P160)
anova(rm, fitall)
```

```
## Analysis of Variance Table
##
## Model 1: V ~ I + G * I + N + P + D
## Model 2: V ~ I + W + G * I + N + P + D
##   Res.Df      RSS Df  Sum of Sq      F Pr(>F)
## 1     14 0.020980
## 2     13 0.020868  1 0.00011217 0.0699 0.7957
```

We observe a P-value above 0.05 and therefore we fail to reject the null hypothesis that the reduced model is adequate. We can now officially remove W from the model.

We also observe that variable P has the highest P-value in the initial model. We now proceed to create a new model without W and P.

```
#Creating a new reduced model without W and P
rm <- lm(V~I+G*I+N+D, data=P160)
anova(rm, fitall)
```

```
## Analysis of Variance Table
##
## Model 1: V ~ I + G * I + N + D
## Model 2: V ~ I + W + G * I + N + P + D
```

```
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1     15 0.022269
## 2     13 0.020868  2 0.0014011 0.4364 0.6555
```

We observe a P-value above 0.05 and therefore we fail to reject the null hypothesis that the reduced model is adequate. We can now officially remove W and P from the model.

We have a new reduced model as follows.

```
summary(rm)
```

```
##
## Call:
## lm(formula = V ~ I + G * I + N + D, data = P160)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.03962 -0.02254 -0.01176  0.01639  0.08163
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.509418   0.020830  24.456 1.68e-13 ***
## I           -0.015585   0.015128  -1.030   0.3192
## G            0.001677   0.001620   1.035   0.3169
## N           -0.005407   0.003252  -1.663   0.1172
## D            0.048248   0.019422   2.484   0.0253 *
## I:G          0.009736   0.001522   6.398 1.20e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03853 on 15 degrees of freedom
## Multiple R-squared:  0.8024, Adjusted R-squared:  0.7365
## F-statistic: 12.18 on 5 and 15 DF,  p-value: 7.632e-05
```

```
vif(rm)
```

```
##        I        G        N        D      I:G
## 3.229876 1.244913 1.224226 3.508817 1.125949
```

We observe that D has a VIF value, but practically we know that it is significant. We check its interractions with N.

```
#Creating a new reduced model with N*D
rm <- lm(V~G*I+N*D, data=P160)
summary(rm)
```

```
##
## Call:
## lm(formula = V ~ G * I + N * D, data = P160)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.04337 -0.01590 -0.00098  0.01467  0.05056
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.5209513  0.0163300  31.901 1.79e-14 ***
## G            0.0010013  0.0012579   0.796  0.43932
```

```
## I           -0.0167426  0.0116051  -1.443  0.17110
## N           -0.0081191  0.0026189  -3.100  0.00783 **
## D           -0.0004716  0.0206881  -0.023  0.98213
## G:I          0.0092127  0.0011771   7.826 1.77e-06 ***
## N:D          0.0088669  0.0026134   3.393  0.00437 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02954 on 14 degrees of freedom
## Multiple R-squared:  0.8916, Adjusted R-squared:  0.8451
## F-statistic: 19.18 on 6 and 14 DF,  p-value: 5.201e-06
```

We see a noticable improvement in the value of R-squared. This is good. We now proceed to check whether the data is stable over time by adding year as a variable.

```
summary(lm(V ~ D *N + G * I + Year, data = P160))
```

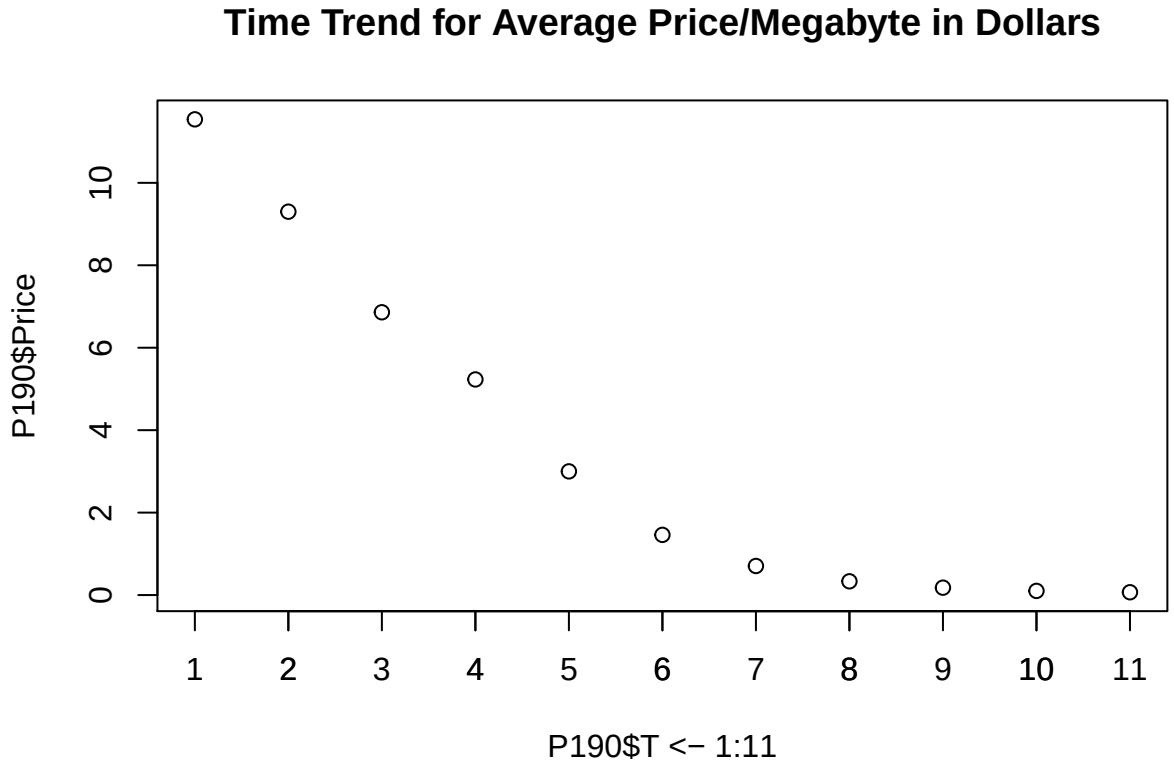
```
##
## Call:
## lm(formula = V ~ D * N + G * I + Year, data = P160)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.042411 -0.018446 -0.000374  0.019064  0.045075
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.1745079  0.6363686  -0.274  0.78822
## D            0.0029010  0.0207758   0.140  0.89109
## N           -0.0069027  0.0028289  -2.440  0.02976 *
## G            0.0004086  0.0013618   0.300  0.76889
## I           -0.0156936  0.0115649  -1.357  0.19788
## Year         0.0003522  0.0003222   1.093  0.29415
## D:N          0.0083252  0.0026423   3.151  0.00766 **
## G:I          0.0094210  0.0011844   7.954 2.38e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02934 on 13 degrees of freedom
## Multiple R-squared:  0.9007, Adjusted R-squared:  0.8472
## F-statistic: 16.84 on 7 and 13 DF,  p-value: 1.409e-05
```

We obtain the highest adjusted R-squared value. This is good enough to conclude this as our final model.

Question 4 ~ Problem 6.8

_a)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.413527	1.0600421	9.823692	4.10e-06
T	-1.148073	0.1562947	-7.345565	4.35e-05



We can see from the plot above that the relationship between the two variables is linear up until the 7th year in the period, at which point price starts to stagnate. Additionally, we can calculate R^2 statistic to check the fit.

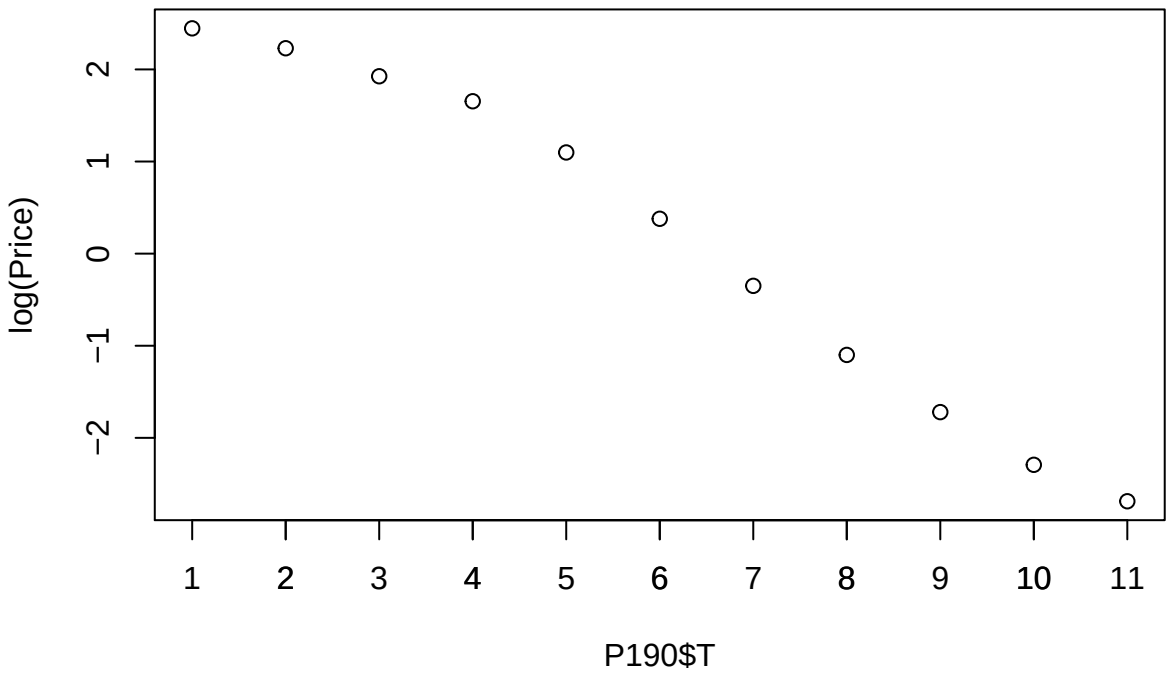
```
cor(Year,Price)^2
```

```
## [1] 0.857046
```

We have a R^2= 0.8570, thus we can say that our data can be represented by a linear trend but it is clear that there seems to be some sort of exponential trend due to the stagnation of price in the last years.

_b)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.5068915	0.1852664	18.92892	0
T	-0.5605034	0.0273160	-20.51921	0



We can see from the plot above that the relationship between the two variables is completely linear for all time. This verifies our claim that there seems to be an exponential trend.Additionally, we can calculate R^2 statistic to check the fit.

```
cor(log(Price),Year)^2
```

```
## [1] 0.9790716
```

We have a R^2= 0.9791, thus we can say that our data is better reprsented with this transfromed model than the linear trend without a transformation.

c) We create a new data set with an indicator variable for years after 1991 and refit to the model with a logarithmic transformation.

```
P190$I=c(0,0,0,0,1,1,1,1,1,1,1)
P190$Cross<-P190$I*P190$T
fit7<-lm(log(Price)~ T + I + Cross, data = P190)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.7336208	0.1444527	18.923983	0.0000003
T	-0.2678531	0.0527467	-5.078104	0.0014341
I	1.4769070	0.2337747	6.317652	0.0003974
Cross	-0.3776303	0.0572629	-6.594680	0.0003058

To compare the two models from ***b)*** and ***c)***, we calculate each models' adjusted R-squared values.

```
summary(fit6)$adj.r.squared
```

```
## [1] 0.9767462
```

```r
summary(fit7)$adj.r.squared
```

```
## [1] 0.9960588
```

Clearly the our new model with indicator variables is better since the adjusted its adjusted R-squared value is larger than the adjusted R-squared value of the model withouth indicator variables. The coefficients indicate that as time increases price decreases, but this relationship is much more pronounced in the years after after 1991 (-0.2678531 > -0.3776303). The coefficient for I being positive is an indication of the slowdown in the decrease in the price as time goes one, which is also why the untouched data looked exponential.

Question 5 ~ Problem 7.4

First we create a full and non-transformed model:

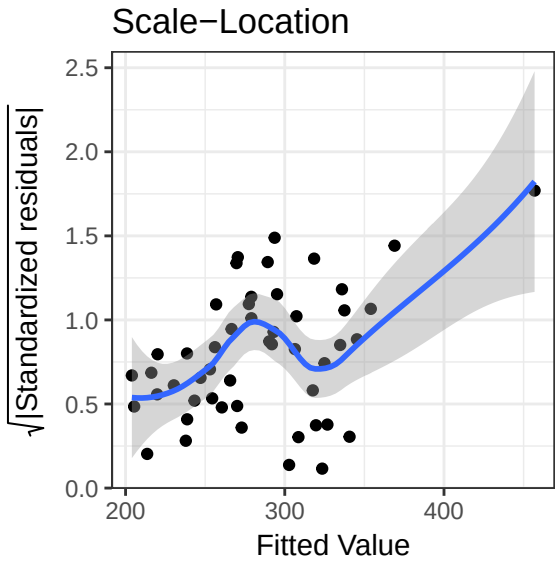
```
Call:
lm(formula = Y ~ X1 + X2 + X3 + factor(Region), data = P198)

Residuals:
    Min      1Q  Median      3Q     Max
-77.963 -25.499  -2.214  17.618  89.106

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -451.67542  139.53852  -3.237 0.002329 **
X1                 0.07204    0.01305   5.520 1.82e-06 ***
X2                 1.30146    0.35717   3.644 0.000719 ***
X3                -0.03456    0.05319  -0.650 0.519325
factor(Region)2  -15.72741   18.16260  -0.866 0.391338
factor(Region)3   -8.63998   18.53938  -0.466 0.643543
factor(Region)4   18.59675   19.68837   0.945 0.350163
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.88 on 43 degrees of freedom
Multiple R-squared:  0.6292,    Adjusted R-squared:  0.5774
F-statistic: 12.16 on 6 and 43 DF,  p-value: 6.025e-08
```

To compare this to the WLS model, we should analyze the homoscedasticity assumption first:



We can see that the current model fails homogeneity since there is a visible curve in the graph above.

When we compare the model above to the WLS model in section 7.4, it is clear to see that the WLS method procues a smaller adjusted R-squared value, but the purpose of WLS is to transform our model such that heteroscedasticy is adressed. Hence, we can say that despite the smaller adjusted R-squared value, the estimates with the WLS model are more accurate. We know this because the standard errors of the estimate are much more smaller with the WLS model.

To test equality of regressions across regions, we perform an f-test based on the model including regions and model without regions.

```
#Defining the two models
fm1 <- lm(Y~X1+X2+X3+factor(Region), data=P198)
rm1 <- lm(Y~X1+X2+X3, data=P198)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1 + X2 + X3 + factor(Region)
## Model 2: Y ~ X1 + X2 + X3
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     43 68371
## 2     46 75348 -3   -6976.5 1.4626 0.2381
```

We see from the table above that the p-value is above 0.05, and is therefore insignificant. Therefore we fail to reject the null hypothesis and conclude that the reduced model is accurate. This means that there is equality of regression across regions since an indictor value for each region is deemed unnessary.