Report 1

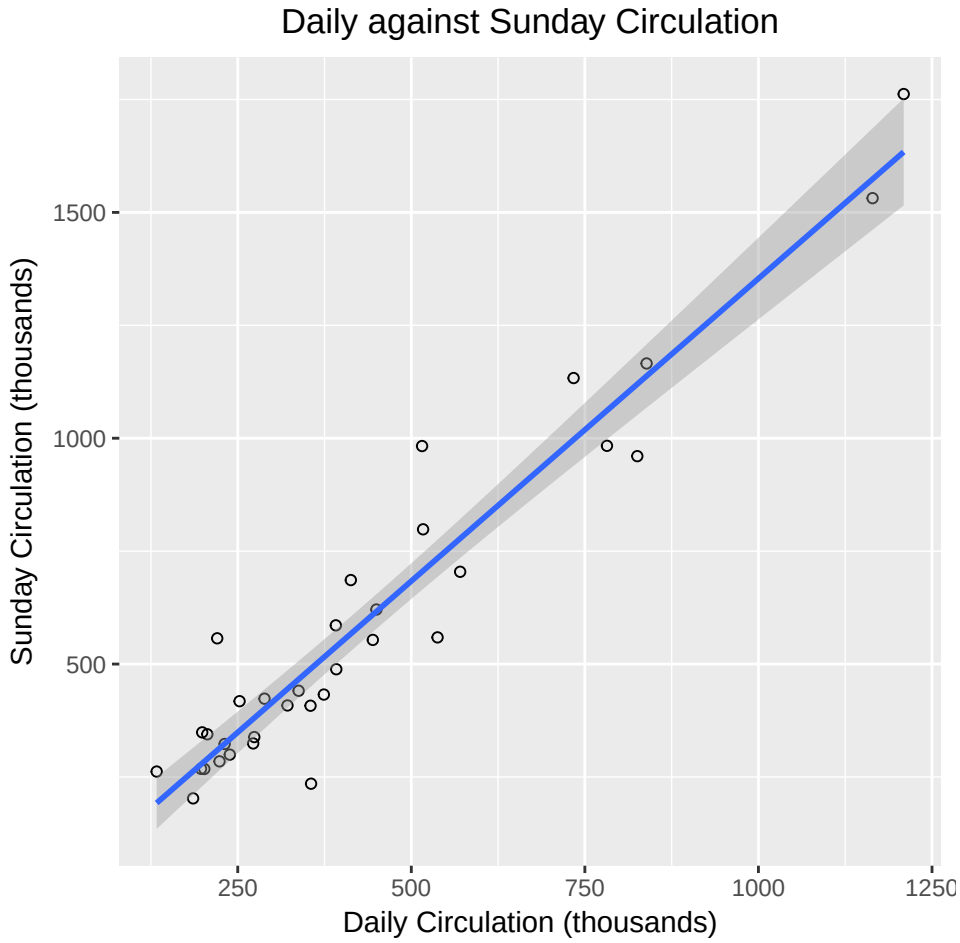
Mufitcan Atalay

4/17/2017

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

Question 1 ~ Exercise 2.12

a)



By observation, we can see that there is a positive linear relationship between Daily and Sunday circulation. This follows logic as papers which have high daily circulation should have high Sunday circulation since they have established readers and audience.

b)

```
fit1<-lm(Sunday ~ Daily, data=P054)
kable(coefficients(summary(fit1)))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	13.835630	35.804006	0.3864269	0.7017382
Daily	1.339715	0.070754	18.9348403	0.0000000

Similarly, we see that that there is a significant correlation between the two variables, further substantiating our claim in **a)**

c) Since $n = 34$, the critical value for a 95% confidence interval is: $t_{(32,\alpha/2)} = 2.0369333$

95% Confidence Interval for β_0:

$13.8356299 \pm 2.0369333 \times 35.8040058 = 13.8356299 \pm 72.9303732$=(-59.09474, 86.766)

95% Confidence Interval for β_1:

‘$1.3397148 \pm 2.0369333 \times 0.070754 = 1.3397148 \pm 0.1441211$= (1.1955939, 1.4838361)

d) $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$

Testing the $\alpha = 5\%$ significance level, we use the same critical value from part **c)** which is equal to 2.0369333

t statistic: $t_1 = \frac{1.4}{0.071} = 18.9348403$

$18.9348403 > 2.0369333$

Therefore, we are able to reject the null hypothesis conclude that there is a linear relationship between Sunday circulation and Daily circulation.

e) The variability in the Sunday circulation that is accounted by Daily circulation can be measured by R^2. This is equivalent to 0.9180597

f) We need to calculate the standard error of the average: $s.e.(\hat{\mu}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$, where σ is the residual standard error and X = Daily circulation.

For this case: $\sigma = \sqrt{\frac{383136}{32}} = 109.42$, $\bar{x} = 430.9624706$ and Var(X) = 72474.82

Therefore, we can say that $s.e.(\hat{\mu}_{500,000}) = 287.8922568$

$\hat{\mu} = 13.835 + 1.4 \times 500,000 = 669871.2$

We can use the same critical value from part **c)** to obtain the 95% confidence interval: 669,871.2 ± 586.4168141

Our confidence interval is given by $(644.1951, 723.191)$

g In this situation, we use the standard error for $\hat{Y}$ which is given as $\hat{\sigma}\sqrt{1 + \frac{1}{34} + \frac{(500-\bar{x})^2}{\sum(x-\bar{x})^2}} = 111.02$

Aside from this change, we have the same process as **f)**, so we have our 95% confidence interval as 669,871.2 ± 226.1403

Our confidence interval is given by $(457.69, 909.98)$

Given that we have a higher standard error, the interval for 95% is larger, as expected.

h We use the same formula from part **g)** for the standard error of $\hat{Y}$ with $x_0 = 2,000,000$ instead of 500,000, we calculate:

```
alpha=0.05
n = nrow(P054)
t_crit = qt(1-alpha/2,n-2)
x_02 = 2000
y_hat_02 = c(1,x_02) %*% summary(fit1)$coeff[,1]
```

```r
sigma_hat = summary(fit1)$sigma #
xbar = mean(P054$Daily)
se_yhat2 = sigma_hat * sqrt(1+(1/n)+((x_02-xbar)^2)/sum((P054$Daily-xbar)^2))
#Lower bound confidence interval
y_hat_02-t_crit*se_yhat2
```

```
##          [,1]
## [1,] 2373.463
```

```r
#Upper bound confidence interval
y_hat_02+t_crit*se_yhat2
```

```
##          [,1]
## [1,] 3013.068
```

This gives us a confidence interval of $(2373.463, 3013.068)$

This interval is much wider than the interval in part (g) in absolute terms. This is a result of a larger standard error from an increase in x. A potential problem with this estimation is that we don't actually have any data points near our confidence interval. If we accept that as the paper scales up the trend continues linearly, then we can conclude that our estimation is accurate.

Question 2 ~ 3.12

a) To test if men are paid more than equally qualified women, we can test a null hypothesis that coefficient for gender is zero when the response variable is salary, therefore H_0: β_2=0, therefore we calculate the t-statistic: $\frac{0.224337}{0.4681}$= 0.4793, we obtain that the p-value for this t-statistic is 0.6329 which is very high, as a result we fail to reject the null hypothesis that men are paid more than equally qualified women

b) To test if men are less qualified than equally paid women, we can test a null hypothesis that the coefficient forgender coefficient is zero when the response variable is qualification, therefore H_0: β_1=0, we calculate the t-statistic is $\frac{0.850979}{0.4349}$= 1.96, and the p-value for this t-statistic is 0.0532. We can therefore reject the null hypothesis, therefore implying that men are less qualified than equally paid women

c) When we compare both of our test we can see the inconsistency. The two models give us different conclusions about how men and women are paid.

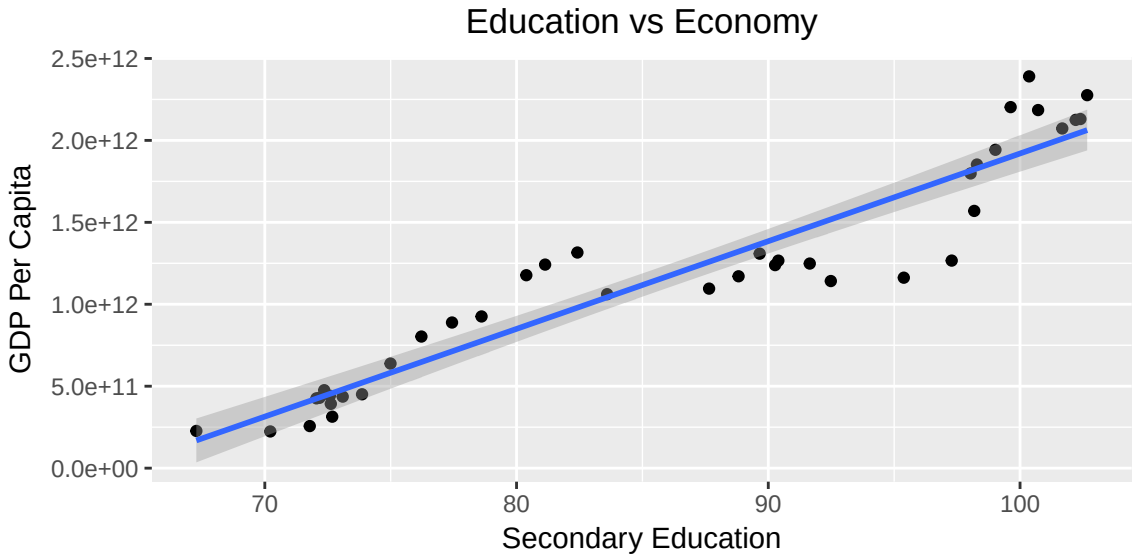
d) I think I would use the first model because the qualification being a predictor for salary seem more intuitive. This would lead us to conclude that men are overpaid in comparison to equally qualified women, which also makes intuitive sense.

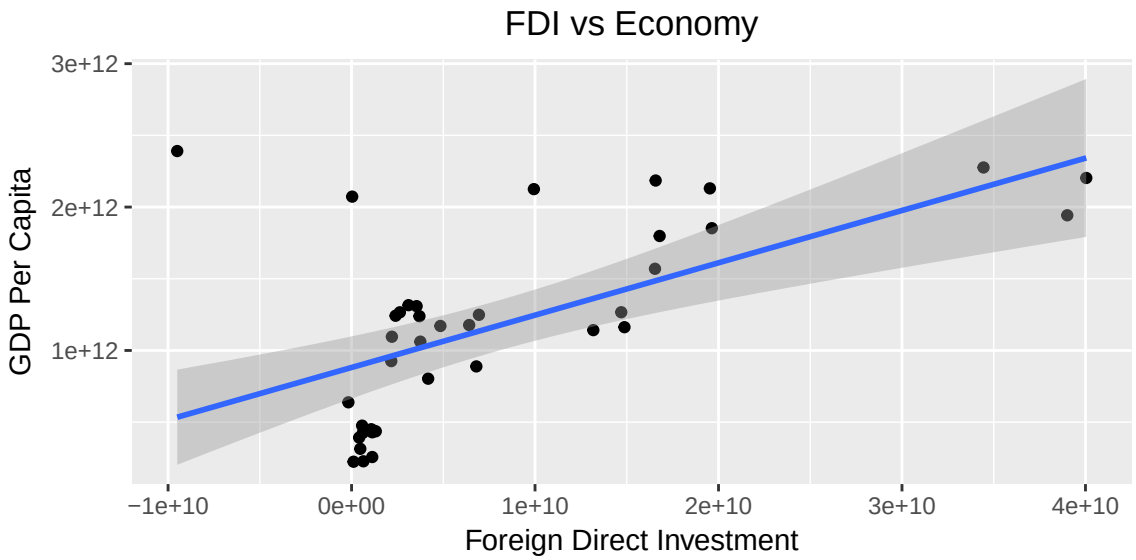
Question 3 ~ 4.2

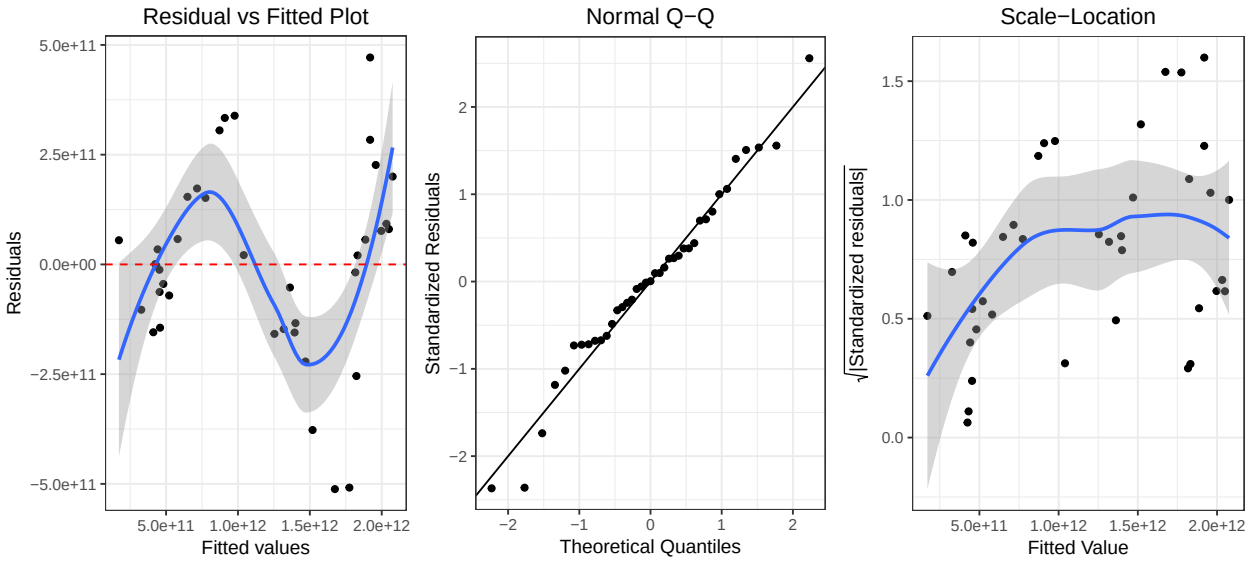
a) In this question I will use statistics from the World Bank on Italy and its GDP per capita. For the predictor values I pick gross secondary education enrollement rate and foreign direct investment, two seemingly unrelated variables that would tradtionally be thought to contribute to GDP per capita.

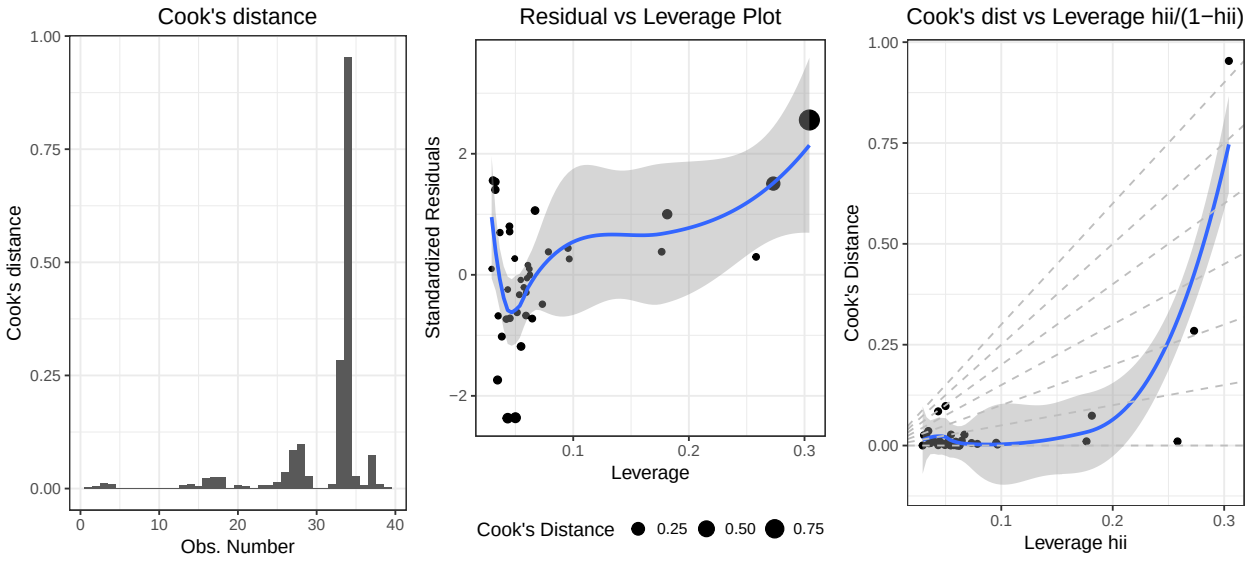
```r
df<-WDI(country = "all",
         indicator=c("NY.GDP.MKTP.CD",
                  "SE.SEC.ENRR",
                  "BX.KLT.DINV.CD.WD"),
         start=1975,
         end=2015)
colnames(df) <- c("ISO","Country","Year","GPD.per.Capita", "Secondary.Education","Foreign.Direct.Investm
```

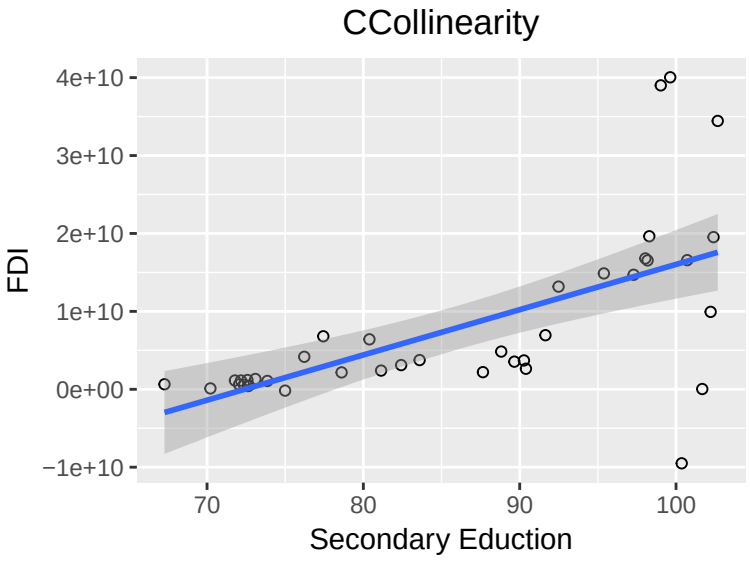
```r
#Analyze data for Italy
d.f <- df %>%
            filter(ISO == "IT", Year<2014)
```






Residual vs Fitted Plot
Residuals
Fitted values
5.0e+11
2.5e+11
0.0e+00
−2.5e+11
−5.0e+11
5.0e+11
1.0e+12
1.5e+12
2.0e+12
Normal Q−Q
Standardized Residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted Value
1.5
1.0
0.5
0.0


Cook's distance
Cook's distance
Obs. Number
1.00
0.75
0.50
0.25
0.00
0
10
20
30
40
Residual vs Leverage Plot
Standardized Residuals
Leverage
Cook's Distance
0.25
0.50
0.75
Cook's dist vs Leverage hii/(1−hii)
Cook's Distance
Leverage hii
0.1
0.2
0.3


CCollinearity
FDI
Secondary Eduction
4e+10
3e+10
2e+10
1e+10
0e+00
−1e+10
70
80
90
100

1. Assumptions about the form: we can see from our residuals vs. fitted values graph that there is a pattern and that the residuals are not roughly equally spread out, therefore we cannot conclude that our linearity assumption holds.
2. Assumptions about errors: we can see that the standardized residuals are weakly normally distributed. This also means that the mean of the errors is probably equal to zero. We can see from the scale-location graph that there is a weak pattern, so we cannot conlude that errors have same variance. Similarly, most of the residuals are not close to each other so we can also conclude that the errors arenot independent
3. Assumptions about the predictors: The predictors are given to us with no errors, but since these are measurments they are going to have some uncertainties.The predictions come from the World Bank, and they use methodology that estimates and extrapolates empty values within a timeseries. However it seems as though this is not really affecting our regression. When we look at the plots of the predictors against each other we see that there is no linear correlation the predictor variables, so we can conclude that the predictors are independent.
4. Assumptions about the observations: We can clearly see that one of our later observations has a Cook's distance very close to 1. This means that it is most likely a point of high leverage. Therefore we are unable to conclude that our observations are mostly equally reliable.

b) We want to test which variable better describes the model. We do this by testing two reduced models and compare that to the full model.

```
attach(d.f)
fullMod    = lm(GPD.per.Capita~ Secondary.Education + Foreign.Direct.Investment)
reducedMod = lm(GPD.per.Capita~ Secondary.Education)

kable(anova(reducedMod, fullMod))
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
37	1.759233e+24	NA	NA	NA	NA
36	1.757531e+24	1	1.701906e+21	0.0348606	0.8529362

```
#Model 2
reducedF = lm(GPD.per.Capita~ Foreign.Direct.Investment)
kable(anova(reducedF, fullMod))
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
37	1.079355e+25	NA	NA	NA	NA
36	1.757531e+24	1	9.036016e+24	185.0873	0

```
summary(fullMod)$r.squared
```

```
## [1] 0.8957263
```

```
summary(reducedMod)$r.squared
```

```
## [1] 0.8956253
```

```
summary(reducedF)$r.squared
```

```
## [1] 0.3596226
```

Clearly the reduced model between GDP per capita and gross secondary education enrollment rate is equally as good of a predictor as the full model. This is a little bit suprising considering that foreign direct investment,

a measure of international trade of a country, is seen as an important factor in the growth of an economy. Maybe countries should invest more in education rather than attracting investors.

Question 4 ~ 3.15

a) If the variable female is not needed, then our coefficent $\beta_{Female} = 0$. Our null hypothesis would test the reduced model, or the regression equation for sales, with $\beta_{Female} = 0$.

$$H_0 : Y = \beta_0 + \beta_{Age}X_1 + \beta_{HS}X_2 + \beta_{Income}X_3 + \beta_{Black}X_4 + \beta_{Price}X_5$$

$$H_A : Y = \beta_0 + \beta_{Age}X_1 + \beta_{HS}X_2 + \beta_{Income}X_3 + \beta_{Black}X_4 + \beta_{Price}X_5 + \beta_{Female}X_5$$

To test this hypothesis, use Anova package to calcualte F-test

```
attach(P088)
fullMod    = lm(Sales~ Age+Black+HS+Income+Price+Female)
reducedMod = lm(Sales~ Age+Black+HS+Income+Price)
kable(anova(reducedMod, fullMod))
```

Res.Df	RSS	Df	Sum of Sq	F	
45	34954.42	NA	NA	NA	
44	34925.97	1	28.45307	0.0358454	
Since we	have a p-le	vel o	f 0.8507, we	fail to rej	ect the null hypothesis. We can therefore say that the variable F

b)

```
fullMod = lm(Sales~ Age+Black+HS+Income+Price+Female)
reducedMod = lm(Sales~ Age+Black+HS+Income+Price)
kable(anova(reducedMod, fullMod))
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
46	34959.77	NA	NA	NA	NA
44	34925.97	2	33.79856	0.0212898	0.9789453

Again our p-level of 0.9789 is larger than 0.05 therefore we fail to reject the null hypothesis, which means that removing variables Female and HS adequately describes the FM.

c) We can use the formula below to compute a confidence interval

$$\hat{\beta}_i \pm t_{(\alpha/2, n-p-1)} \times \text{s.e.}(\hat{\beta}_i)$$

But we could also use the model coefficients and the *confit argument* to onbtain a midpoint of 0.01894645 with a 95% confidence interval of $(0.001642517, 0.03953542)$ **d)**

```
fullModnew <- lm(Sales~ Age+Black+HS+Income+Price+Female)
reducedModnew <- lm(Sales~ Age+Black+HS+Price+Female)
summary(fullModnew)$r.squared
```

```
## [1] 0.3208426
```

```
summary(reducedModnew)$r.squared
```

```
## [1] 0.2677526
```

We see that the reduced model when Income is removed from the above regression table accounts for 26.78% variation in Sales, whereas the full model accounts for 32.08% of variation in Sales.

e) We do the same operation as we did in d), so we create restricted models that only consist of the three variables.

```
reducedModprice <- lm(Sales~ Age+Black+HS+Income+Female)
reducedModage <- lm(Sales~ Black+HS+Income+Price+Female)
reducedModincome <- lm(Sales~ Age+Black+HS+Price+Female)
summary(fullModnew)$r.squared
```

```
## [1] 0.3208426
```

```
summary(reducedModprice)$r.squared
```

```
## [1] 0.16712
```

```
summary(reducedModage)$r.squared
```

```
## [1] 0.2904175
```

```
summary(reducedModincome)$r.squared
```

```
## [1] 0.2677526
```

```
#Percentage accounted by Price
(summary(fullModnew)$r.squared -summary(reducedModprice)$r.squared)*100
```

```
## [1] 15.37226
```

```
#Percentage accounted by Age
(summary(fullModnew)$r.squared -summary(reducedModage)$r.squared)*100
```

```
## [1] 3.042512
```

```
#Percentage accounted by Income
(summary(fullModnew)$r.squared -summary(reducedModincome)$r.squared)*100
```

```
## [1] 5.309002
```

(f)

```
onlyincomereducedMod <-lm(Sales~ Income)
summary(onlyincomereducedMod)$r.squared
```

```
## [1] 0.1063203
```

Income, by itself, seems to account for variations in Sales by 10%.

Question 5 ~ Exercise 4.7 (modified)

a) With Age, we would expect to see a positive relationship, given that people tend to try drugs as they get older and legal restrictions on age limit children's ability to purchase cigarettes

With HS, we would expect to see a negative relationship because people with basic education should have taken a health class and therefore should be aware of the harms of smoking. More education should lead to less smoke.

With Income, we would expect a positive relationship since more income means more ability to consume, which may or may not be in the form of cigarettes. Although it is possible that it might have a negative relationship since cigarattes are pretty inexpensive.

With Black, we would not expect any relationship because smoking is a global habit and therefore should be largely equal among different races.

With Female, we do not expect any relationship because smoking is neither a particularly masculine activity nor a particularly feminine activity.

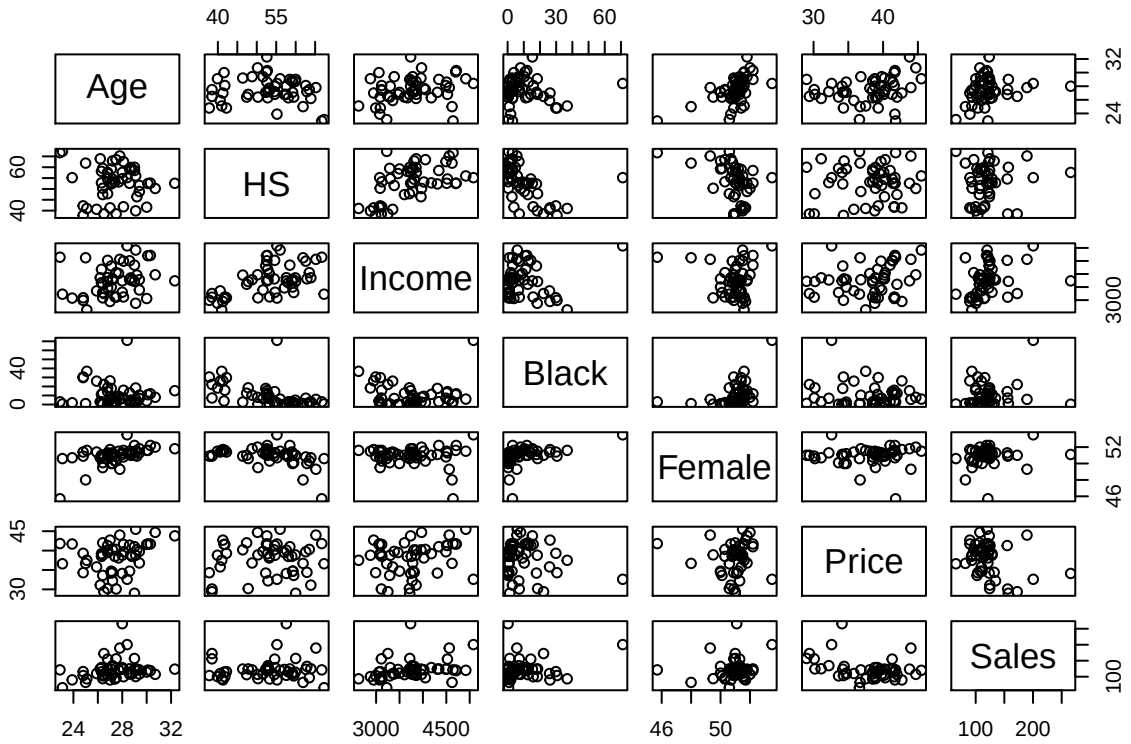
With Price, we would expect a negative relationship because a higher price leads consumers to consume less, considering that cigarettes are not a luxury good.

b)

```
df = data.frame(P088)
cor(df)
```

```
##                Age          HS      Income       Black      Female
## Age     1.00000000 -0.09891626  0.25658098 -0.04033021  0.55303189
## HS     -0.09891626  1.00000000  0.53400534 -0.50171191 -0.41737794
## Income  0.25658098  0.53400534  1.00000000  0.01728756 -0.06882666
## Black  -0.04033021 -0.50171191  0.01728756  1.00000000  0.45089974
## Female  0.55303189 -0.41737794 -0.06882666  0.45089974  1.00000000
## Price   0.24775673  0.05697473  0.21455717 -0.14777619  0.02247351
## Sales   0.22655492  0.06669476  0.32606789  0.18959037  0.14622124
##              Price       Sales
## Age     0.24775673  0.22655492
## HS      0.05697473  0.06669476
## Income  0.21455717  0.32606789
## Black  -0.14777619  0.18959037
## Female  0.02247351  0.14622124
## Price   1.00000000 -0.30062263
## Sales  -0.30062263  1.00000000
```

```
plot(df)
```



Main present differences that were not expected was the relationship between cigarette consumption with gender or race.

A matrix of the coefficients would have been sufficient.

c)

```
income = lm(Sales ~ Income)
age = lm(Sales ~ Age)
hs = lm(Sales ~ HS)
black = lm(Sales ~ Black)
female = lm(Sales ~ Female)
price = lm(Sales ~ Price)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	55.3624541	27.7430820	1.995541	0.0515590
Income	0.0175834	0.0072826	2.414433	0.0195393

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	15.219199	65.448795	0.232536	0.8170903
Age	3.870946	2.377408	1.628221	0.1098918

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	107.3330520	30.7008785	3.4960906	0.0010131
HS	0.2673262	0.5713255	0.4679052	0.6419269

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	116.7377984	5.6980409	20.487357	0.0000000
Black	0.4807148	0.3556512	1.351647	0.1826953

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-93.426013	207.812598	-0.4495686	0.6550040
Female	4.219098	4.077726	1.0346694	0.3059029

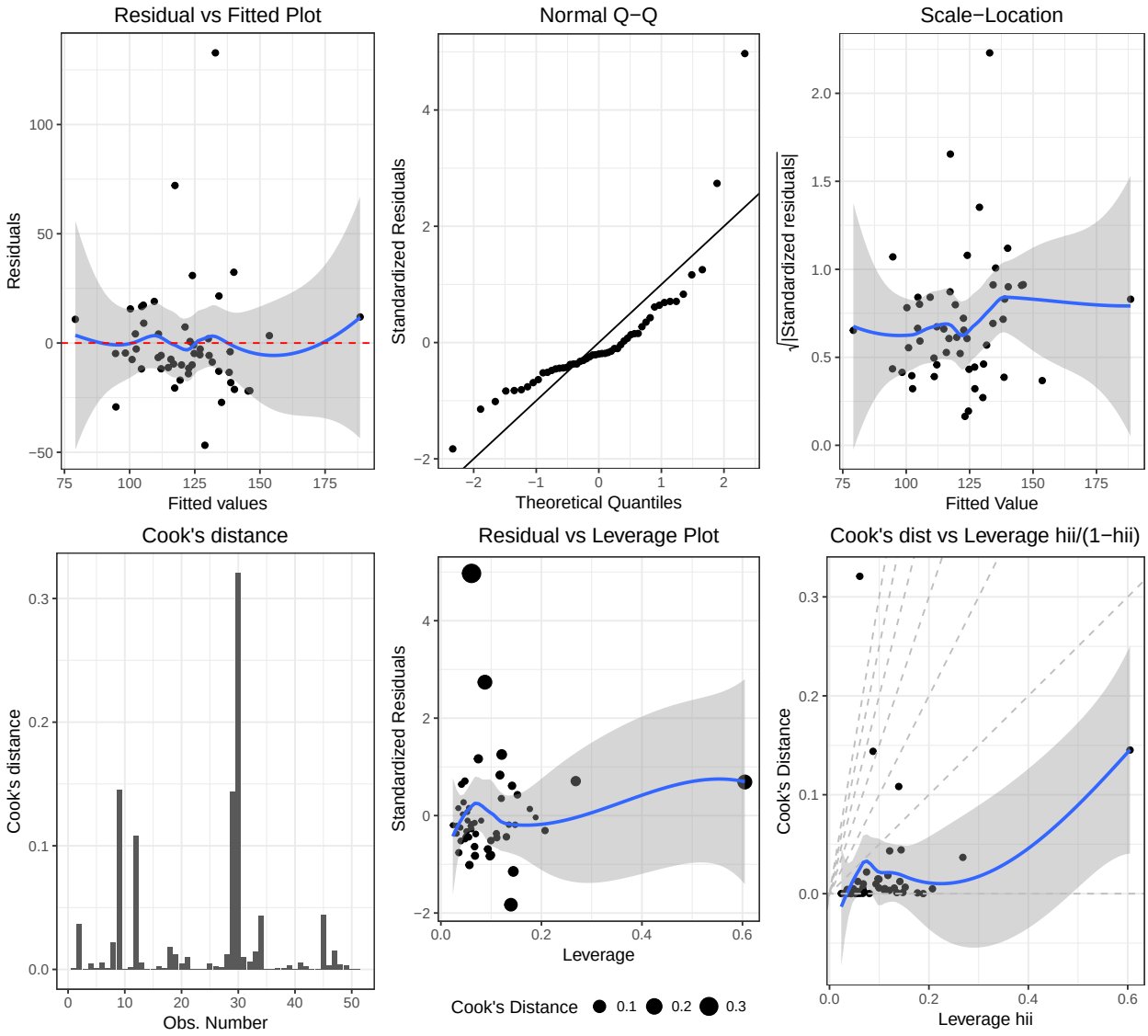
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	210.453051	40.528480	5.192720	0.0000040
Price	-2.335207	1.058369	-2.206421	0.0320724

All the coefficients are positive, except for price. This was largely expected except for the coefficients for race and gender.

The signs of the correlation coefficient are identical to the signs of the regression coefficients however the magnitiudes are different. The magnitude of the correlation coefficient describes the strength of the linear relationship. The magnitude of the regression coefficients describes how Sales will change with every added unit of the respective predictor variables. Therefore, we can say that the correlation coefficient is not sensitive to changes in units while the regression coefficients are. Hence, they are supposed to have similar signs as the relationship effects how the response changes due to a change in a predictor. The magnitudes being different is expected.

d)

```
fit1 <- lm(Sales ~ Age + Income + Black + Price, P088)
```



Linearity: Weak pattern. Assumption is fulfilled.

Variance: Constant varaince in residuals. Homoskedasticity is also avoided.

Residuals: QQ plot appears normal, excluding some outliers

Outliers: No problems here as well. The leverages are within a range of 0.5, which is not very worrisome. All the assumption are met.