Week 2 Homework

Ben Arancibia

June 20, 2015

HA 4.1 Electricity consumption was recorded for a small town on 12 randomly chosen days. The following maximum temperatures (degrees Celsius) and consumption (megawatt-hours) were recorded for each day.

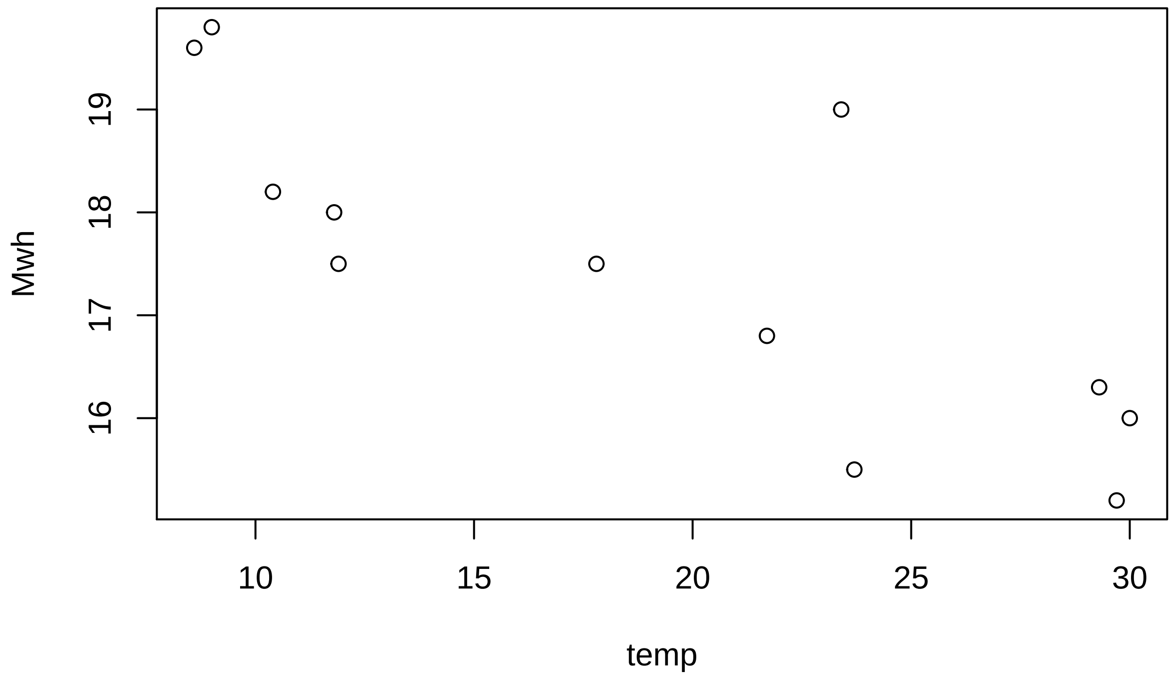
```
library(fpp)
```

```
knitr::kable(econsumption)
```

Mwh	temp
16.3	29.3
16.8	21.7
15.5	23.7
18.2	10.4
15.2	29.7
17.5	11.9
19.8	9.0
19.0	23.4
17.5	17.8
16.0	30.0
19.6	8.6
18.0	11.8

a) Plot the data and find the regression model for Mwh with temperature as an explanatory variable. Why is there a negative relationship?

```
plot(Mwh ~ temp, data=econsumption)
```



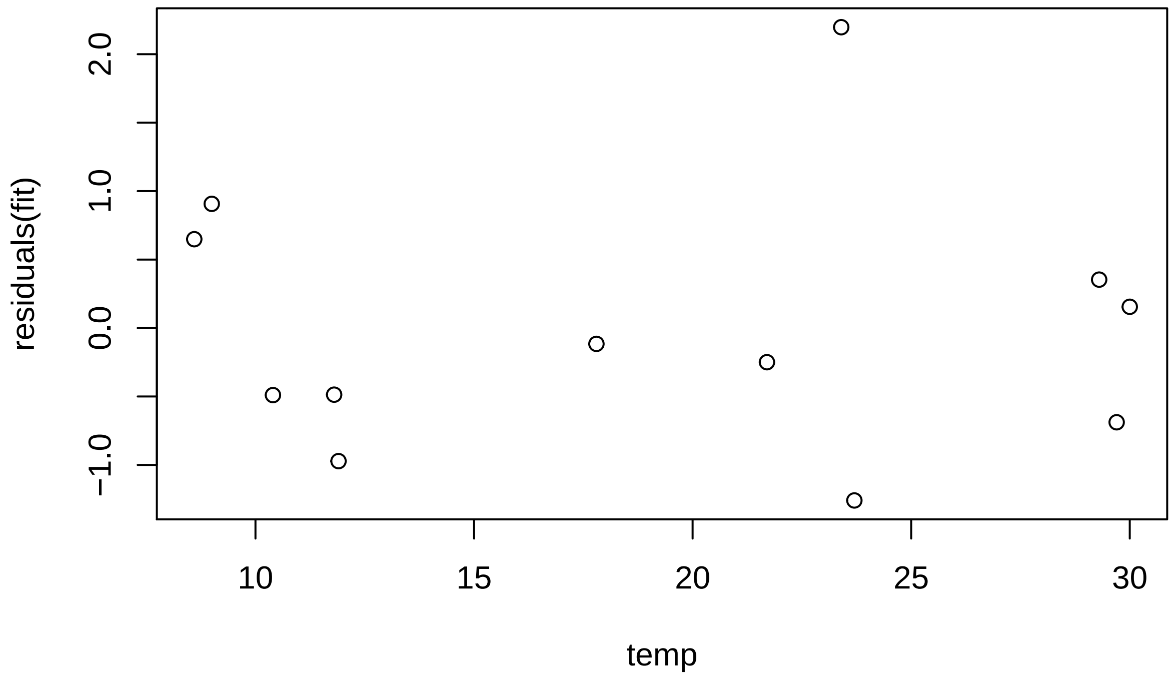

```r
fit  <- lm(Mwh ~ temp, data=econsumption)
summary(fit)
```

```
##
## Call:
## lm(formula = Mwh ~ temp, data = econsumption)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2593 -0.5395 -0.1827  0.4274  2.1972
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.19952    0.73040   27.66 8.86e-11 ***
## temp        -0.14516    0.03549   -4.09  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9888 on 10 degrees of freedom
## Multiple R-squared:  0.6258, Adjusted R-squared:  0.5884
## F-statistic: 16.73 on 1 and 10 DF,  p-value: 0.00218
```

It seems that there is a negative relationship because a simple linear model is not appropriate. A non-linear model will be necessary for the data.

b) Produce a residual plot. Is the model adequate? Are there any outliers or influential observations?

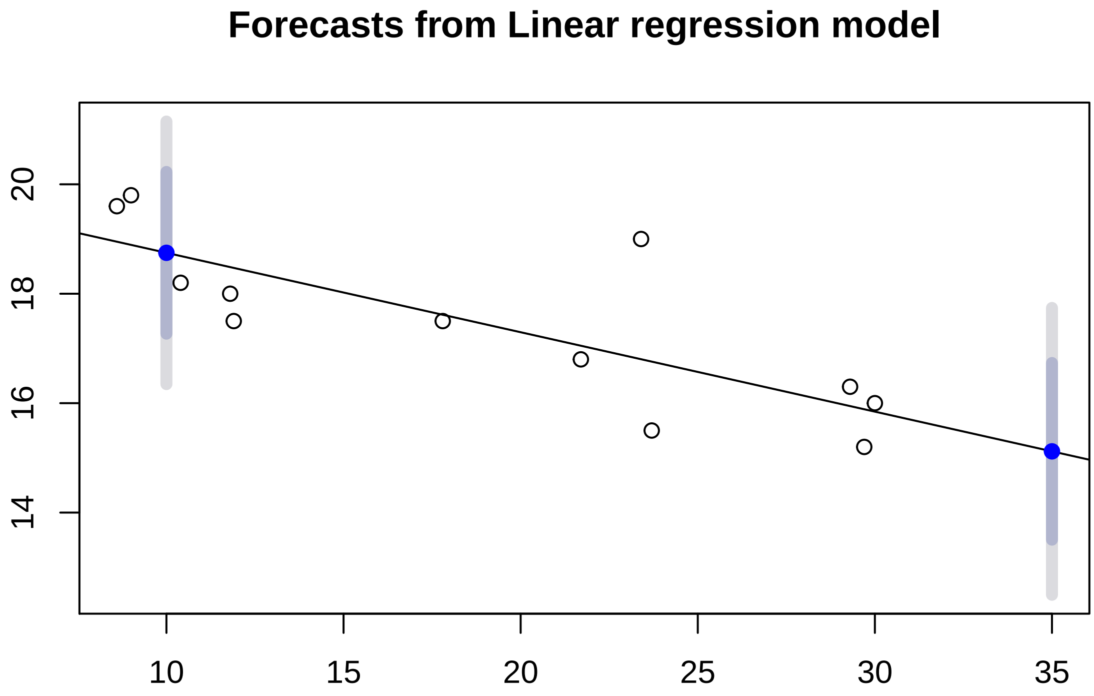
```r
plot(residuals(fit)~temp, data=econsumption)
```



The model appears to be adequate, there is no apparent pattern. The only thing that might cause an issues is 1 outlier at temperature 23.4.

c) Use the model to predict the electricity consumption that you would expect for a day with maximum temperature 10 and a day with maximum temperature 35. Do you believe these predictions?

```
fcast <- forecast(fit, newdata=data.frame(temp=c(10,35)))
plot(fcast)
```



Based on the predictive intervals and the nearby data points, the predictions are believable.

d) Give prediction intervals for your forecasts. The following R code will get you started:

```
summary(fit)
```

```
##
## Call:
## lm(formula = Mwh ~ temp, data = econsumption)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2593 -0.5395 -0.1827  0.4274  2.1972
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.19952    0.73040   27.66 8.86e-11 ***
## temp        -0.14516    0.03549   -4.09  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9888 on 10 degrees of freedom
## Multiple R-squared:  0.6258, Adjusted R-squared:  0.5884
## F-statistic: 16.73 on 1 and 10 DF,  p-value: 0.00218
```

```
forecast(fit, newdata=data.frame(temp=c(60)))
```

```
##   Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 1       11.49008 9.041979 13.93819 7.514874 15.46529
```

HA 5.2 The data below (data set texasgas) shows the demand for natural gas and the price of natural gas for 20 towns in Texas in 1969.

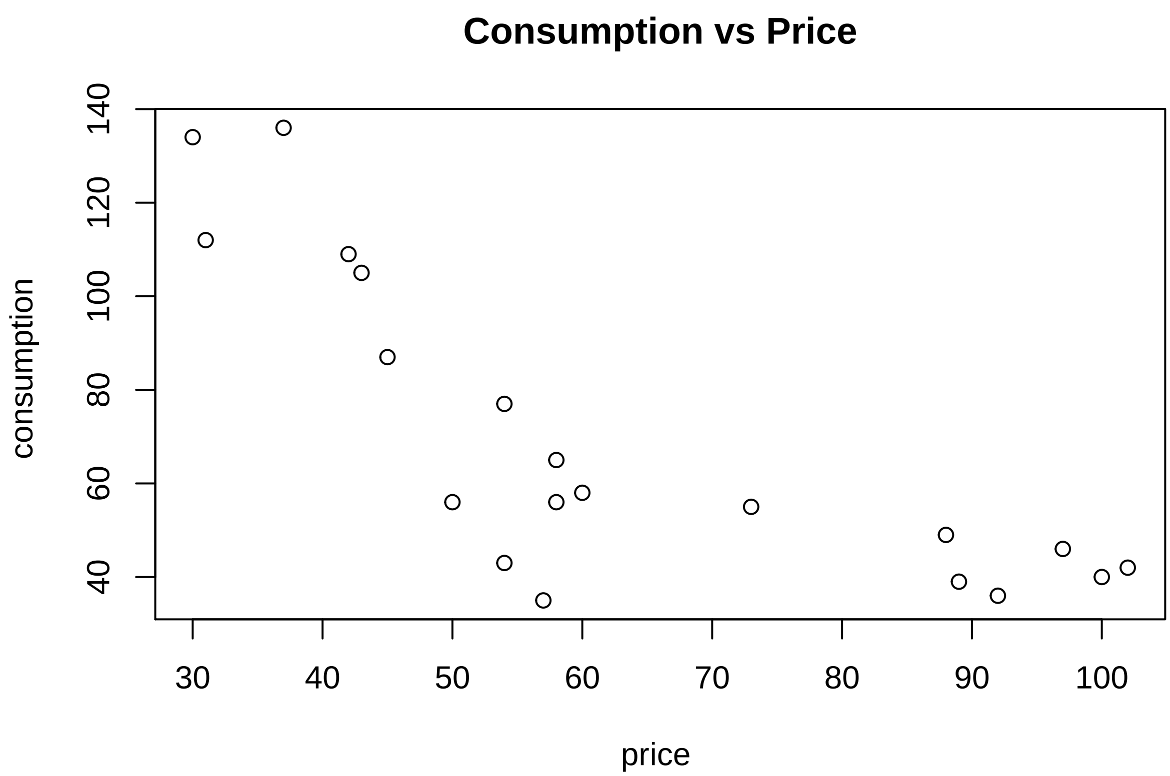
```
knitr::kable(texasgas)
```

price	consumption
30	134
31	112
37	136
42	109
43	105
45	87
50	56
54	43
54	77
57	35
58	65
58	56

price	consumption
60	58
73	55
88	49
89	39
92	36
97	46
100	40
102	42

a) Do a scatterplot of consumption against price.

```
plot(consumption ~ price, data=texasgas,main="Consumption vs Price", xlab="price ", ylab="consumption ")
```

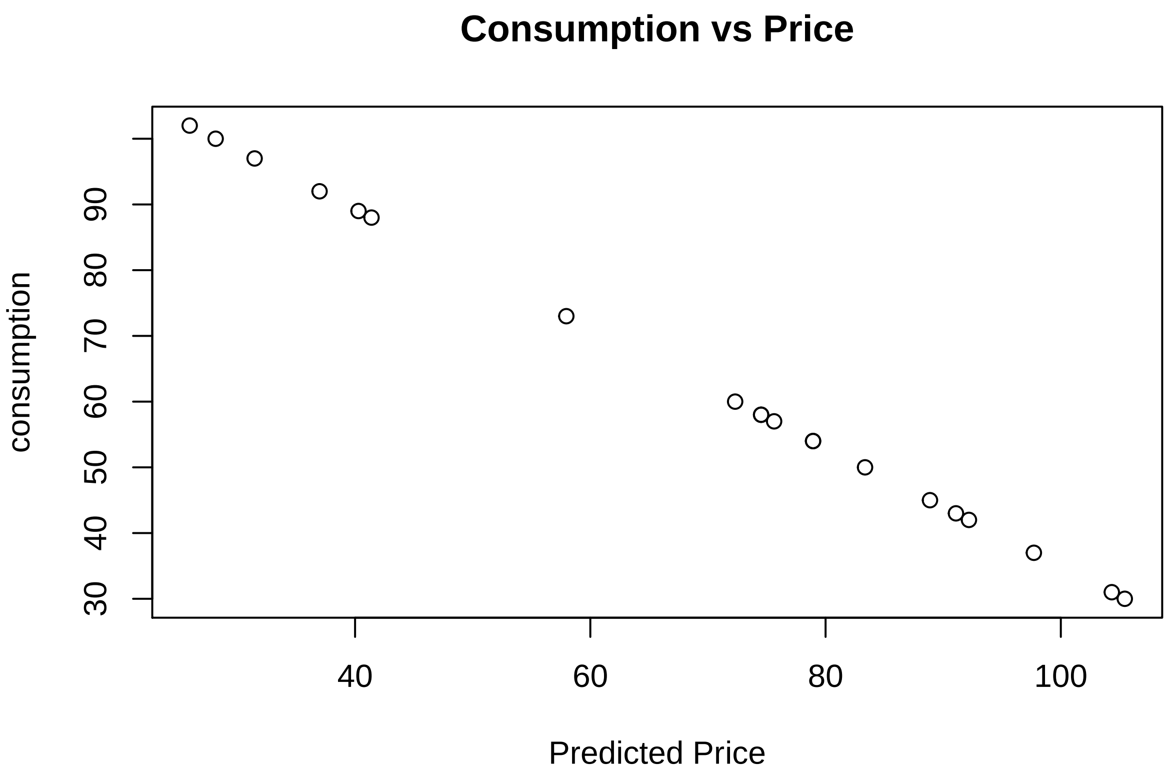


b) The slope of the fitted line should change with P, because this is not a linear relationship. It is a non-linear relationship since it is a negatative relationship.

c) Fit the three models and find the coefficients, and residual variance in each case.

Model 1 - Log

```r
fit1 <- lm(consumption~price, data=texasgas)
plot(fitted(fit1), texasgas$price,  main="Consumption vs Price",ylab="consumption", xlab="Predicted Pri
```

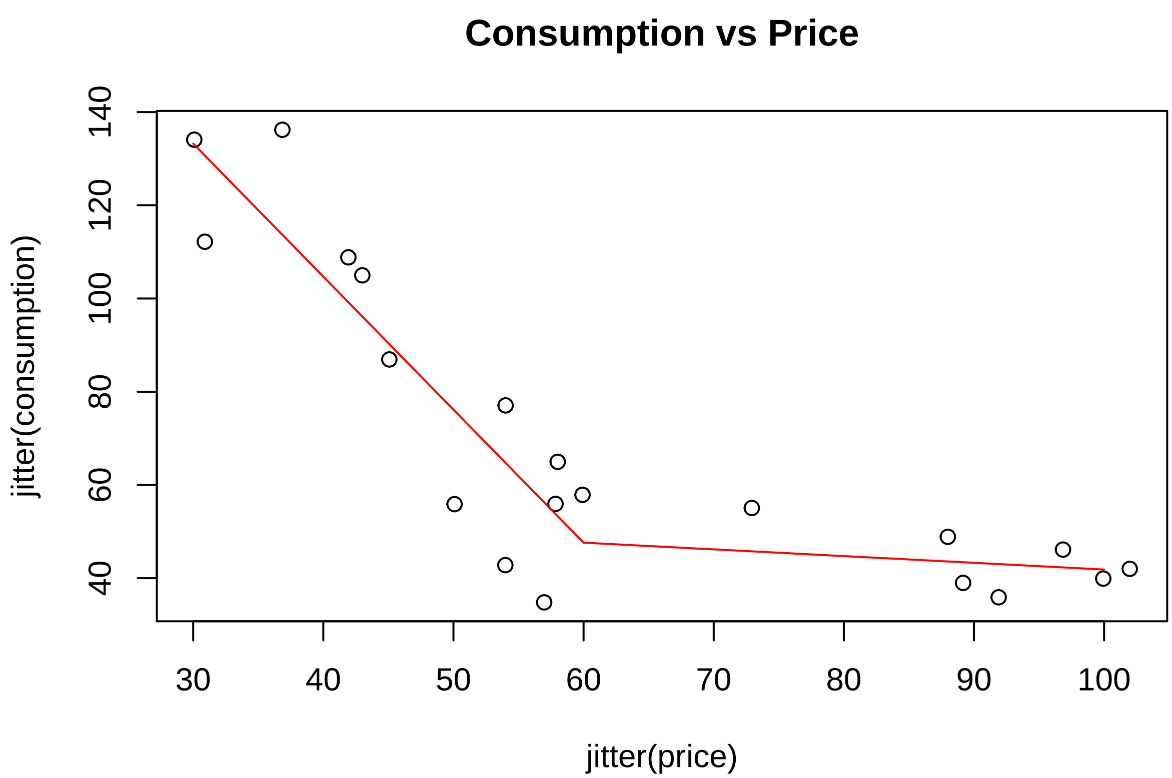



```r
summary(fit1)
```

```
##
## Call:
## lm(formula = consumption ~ price, data = texasgas)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.625 -10.719  -1.136  14.073  38.292
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  138.561     13.552  10.225 6.34e-09 ***
## price         -1.104      0.202  -5.467 3.42e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.86 on 18 degrees of freedom
## Multiple R-squared:  0.6241, Adjusted R-squared:  0.6033
## F-statistic: 29.89 on 1 and 18 DF,  p-value: 3.417e-05
```

Model 2 - Piecewise Linear

```r
pricep <- pmax(texasgas$price-60,0)
fit2 <- lm(consumption~price+pricep,data=texasgas)
x <- 30:100; z <- pmax(x-60,0)
fcast2 <- forecast(fit2, newdata=data.frame(price=x,pricep=z))
plot(jitter(consumption)~jitter(price), data=texasgas, main="Consumption vs Price" )
lines(x, fcast2$mean,col="red")
```

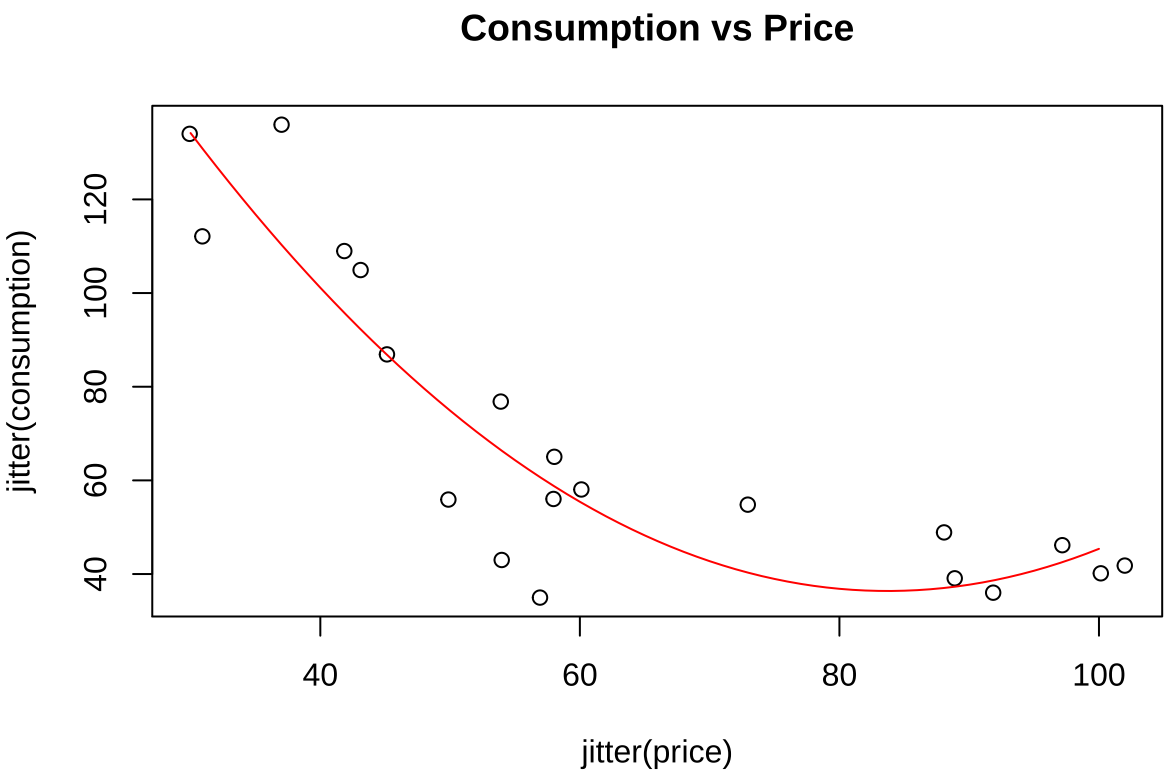



```r
summary(fit2)
```

```
##
## Call:
## lm(formula = consumption ~ price + pricep, data = texasgas)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.744  -5.084   1.722   9.442  22.749
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 218.8263    17.4986  12.505 5.34e-10 ***
## price        -2.8534     0.3560  -8.015 3.56e-07 ***
## pricep        2.7092     0.5144   5.266 6.30e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.23 on 17 degrees of freedom
## Multiple R-squared:  0.8572, Adjusted R-squared:  0.8404
## F-statistic: 51.01 on 2 and 17 DF,  p-value: 6.55e-08
```

Model 3 - Polynomial nonlinear regression

```r
fit3 <- lm(consumption ~ price + I(price^2), texasgas)
fcast3 <- forecast(fit3, newdata=data.frame(price=x,pricep=z))
plot(jitter(consumption)~jitter(price), data=texasgas, main="Consumption vs Price" )
lines(x, fcast3$mean,col="red")
```




```r
summary(fit3)
```

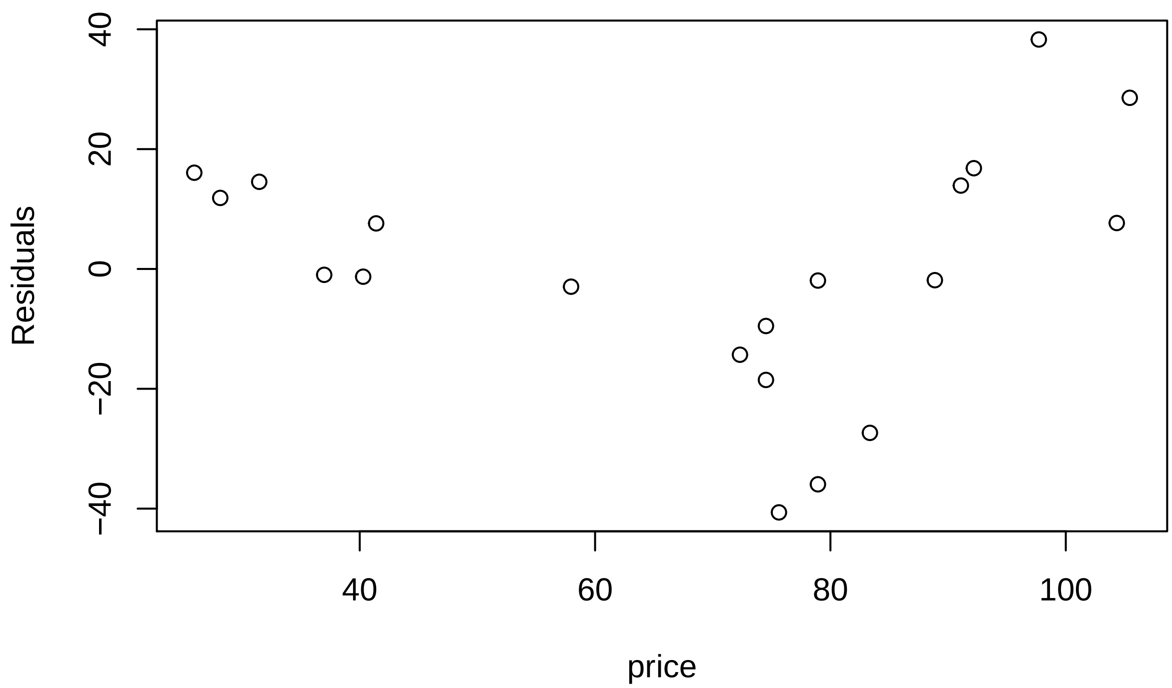
```
##
## Call:
## lm(formula = consumption ~ price + I(price^2), data = texasgas)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.5601  -5.4693   0.7502  11.0252  25.6619
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 273.930628  31.031614   8.827 9.32e-08 ***
## price        -5.675863   1.009086  -5.625 3.03e-05 ***
## I(price^2)    0.033904   0.007412   4.574 0.000269 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.37 on 17 degrees of freedom
## Multiple R-squared:  0.8315, Adjusted R-squared:  0.8117
## F-statistic: 41.95 on 2 and 17 DF,  p-value: 2.666e-07
```

d) For each model, find the value of R2 and AIC, and produce a residual plot. Comment on the adequacy of the three models.

```r
CV(fit1)
```

```
##          CV         AIC        AICc         BIC       AdjR2
## 475.7952381 125.4055763 126.9055763 128.3927731   0.6032525
```

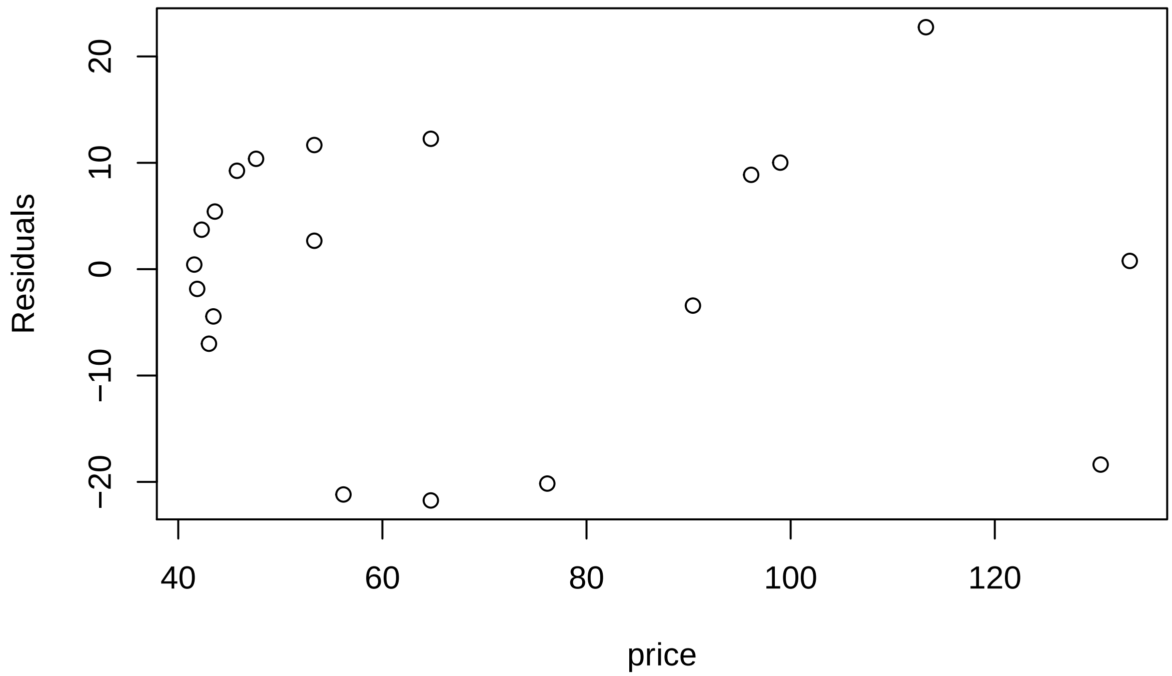
```r
plot(fitted(fit1), residuals(fit1), xlab="price", ylab="Residuals")
```




```r
CV(fit2)
```

```
##          CV         AIC        AICc         BIC       AdjR2
## 204.3115937 108.0556267 110.7222933 112.0385558   0.8403537
```

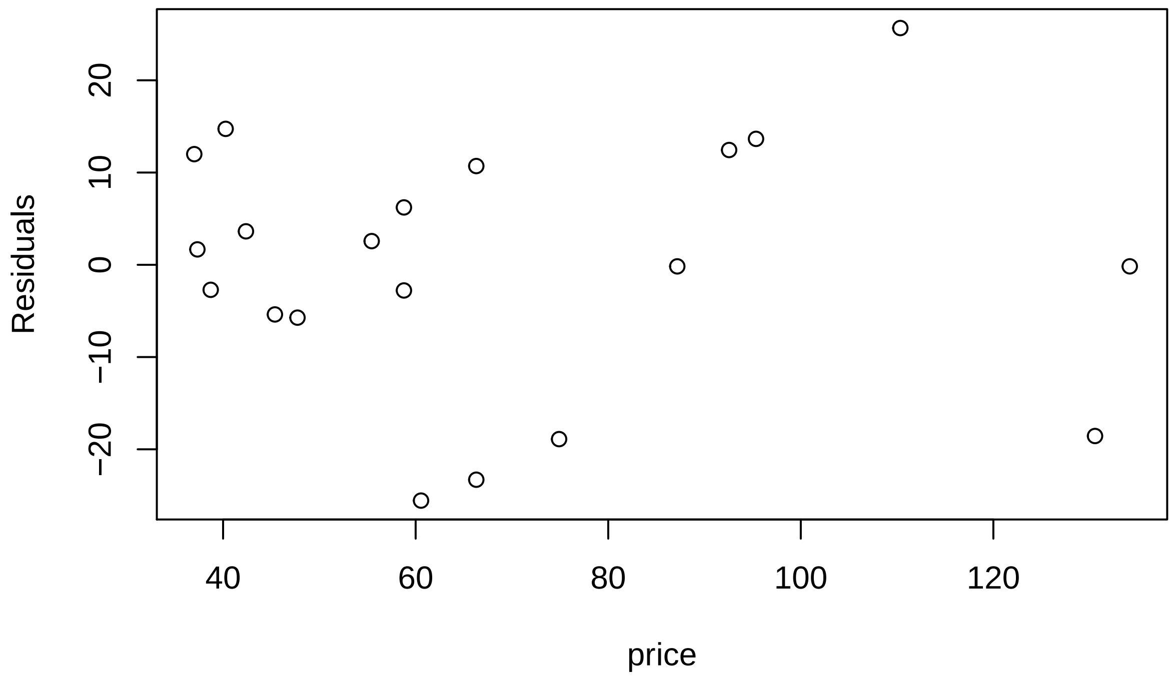
```r
plot(fitted(fit2), residuals(fit2), xlab="price", ylab="Residuals")
```




```
CV(fit3)
```

```
##         CV        AIC       AICc        BIC      AdjR2
## 238.565929 111.358304 114.024971 115.341233   0.811689
```

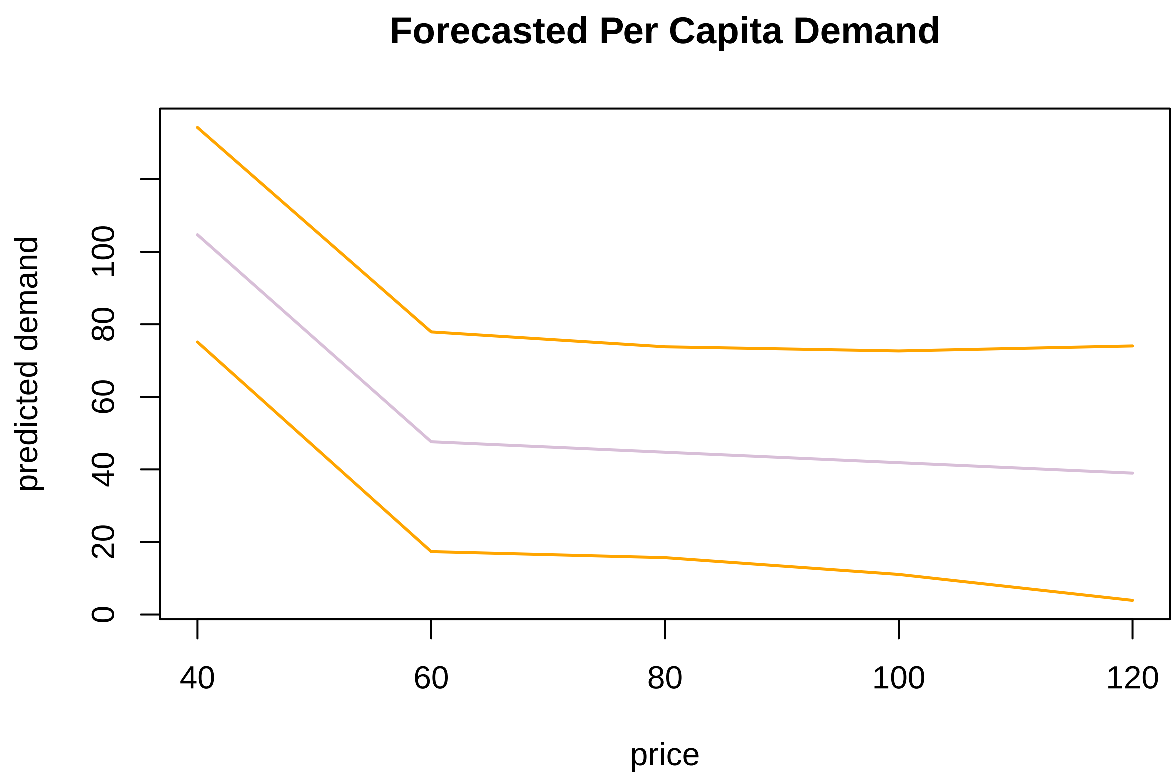
```
plot(fitted(fit3), residuals(fit3), xlab="price", ylab="Residuals")
```



The piecewise linear and polynomial nonlinear regression models have the "best" adjusted R-square values and AIC. It seems that the piecewise linear model does the best explaining the relation of the two variables.

e) For prices 40, 60, 80, 100, and 120 cents per 1,000 cubic feet, compute the forecasted per capita demand using the best model of the three above.

```
x <- seq(40,120, 20); z <- pmax(x-60,0)
predicted <- predict(fit2,newdata=data.frame(price=x,pricep=z),interval="prediction")
matplot(x,predicted,type="l",lty=1,lwd=1.5,col=c("thistle","orange","orange"), xlab="price", ylab="predi
```



f) Compute 95% prediction intervals. Make a graph of these prediction intervals and discuss their interpretation. ####Model 1

```
confint(fit1,level=0.95)
```

```
##                   2.5 %      97.5 %
## (Intercept) 110.090281 167.0317658
## price        -1.528447  -0.6798396
```

```
confint(fit2,level=0.95)
```

Model 2

```
##                   2.5 %     97.5 %
## (Intercept) 181.907467 255.745183
## price        -3.604495  -2.102259
## pricep        1.623794   3.794565
```

```r
confint(fit3,level=0.95)
```

Model 3

```
##                   2.5 %      97.5 %
## (Intercept) 208.45964560 339.4016112
## price        -7.80484861  -3.5468767
## I(price^2)    0.01826611   0.0495416
```

g) What is the correlation between P and P2? Does this suggest any general problem to be considered in dealing with polynomial regressions—especially of higher orders?

```r
cor(texasgas$price,I(texasgas$price^2))
```

```
## [1] 0.9904481
```

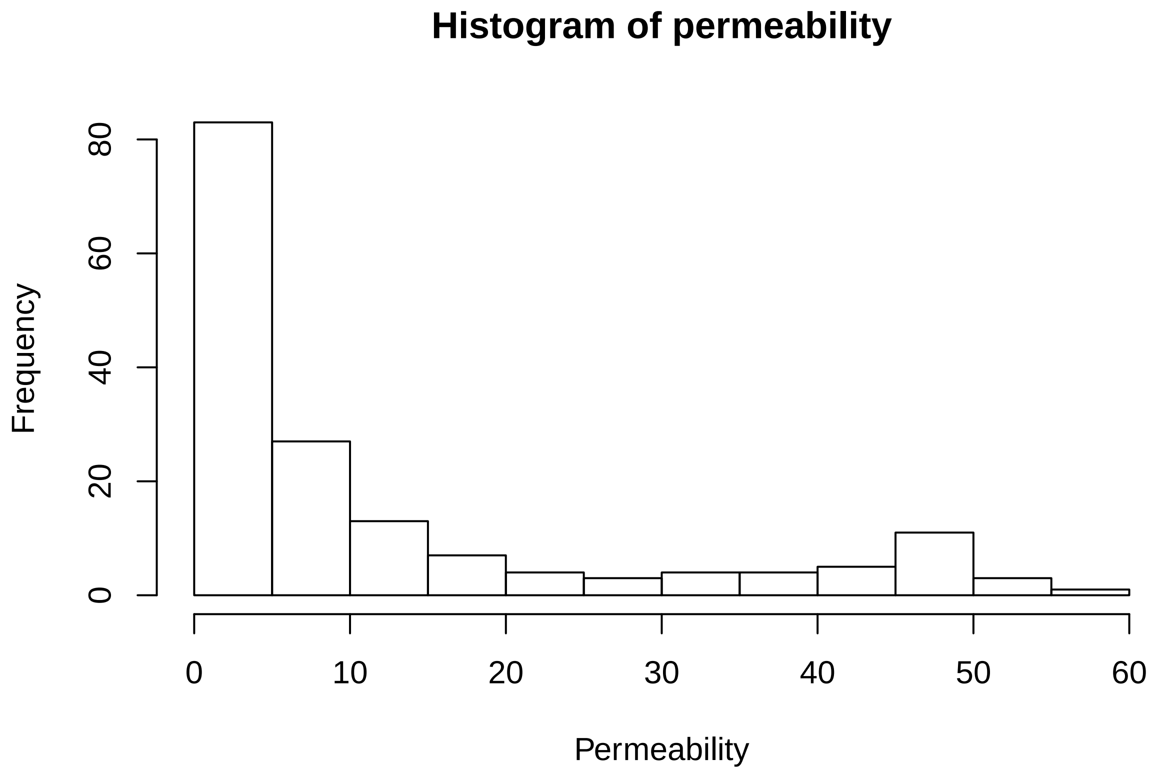
Since it is almost 1, this suggests an issue of multicollinearity. This means it can be difficult to estimate the regression model, uncertainty associated with individual regression coeffiecients will be large, and forecasts will be unreliable if the values of the future predictors are outside the range of the historical values of the predictors.

KJ 6.2

a) Load the necessary data and take a look at it

```r
library(AppliedPredictiveModeling)
data(permeability)

h1 <- hist(permeability, xlab="Permeability")
```




```
h1
```

```
## $breaks
##  [1]  0  5 10 15 20 25 30 35 40 45 50 55 60
##
## $counts
##  [1] 83 27 13  7  4  3  4  4  5 11  3  1
##
## $density
##  [1] 0.100606061 0.032727273 0.015757576 0.008484848 0.004848485
##  [6] 0.003636364 0.004848485 0.004848485 0.006060606 0.013333333
## [11] 0.003636364 0.001212121
##
## $mids
##  [1]  2.5  7.5 12.5 17.5 22.5 27.5 32.5 37.5 42.5 47.5 52.5 57.5
##
## $xname
## [1] "permeability"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

b)