# IS624 - Assignment 1

*James Quacinella*

*06/19/2015*

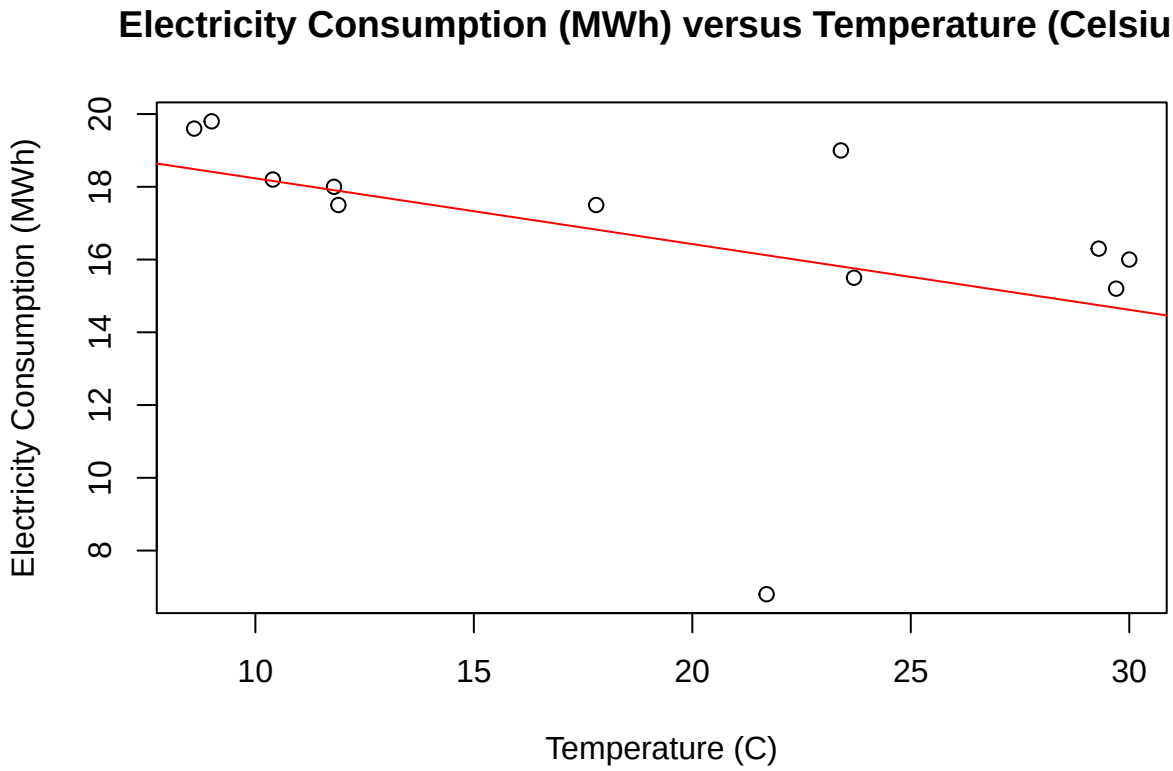
## Question 4.1

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.

```
#Day   1 2  3   4   5   6   7   8   9   10  11  12
data.mwh <- c(16.3, 6.8, 15.5, 18.2, 15.2, 17.5, 19.8, 19.0, 17.5, 16.0, 19.6, 18.0)
data.temp <- c(29.3, 21.7, 23.7, 10.4, 29.7, 11.9, 9.0, 23.4, 17.8, 30.0, 8.6, 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?

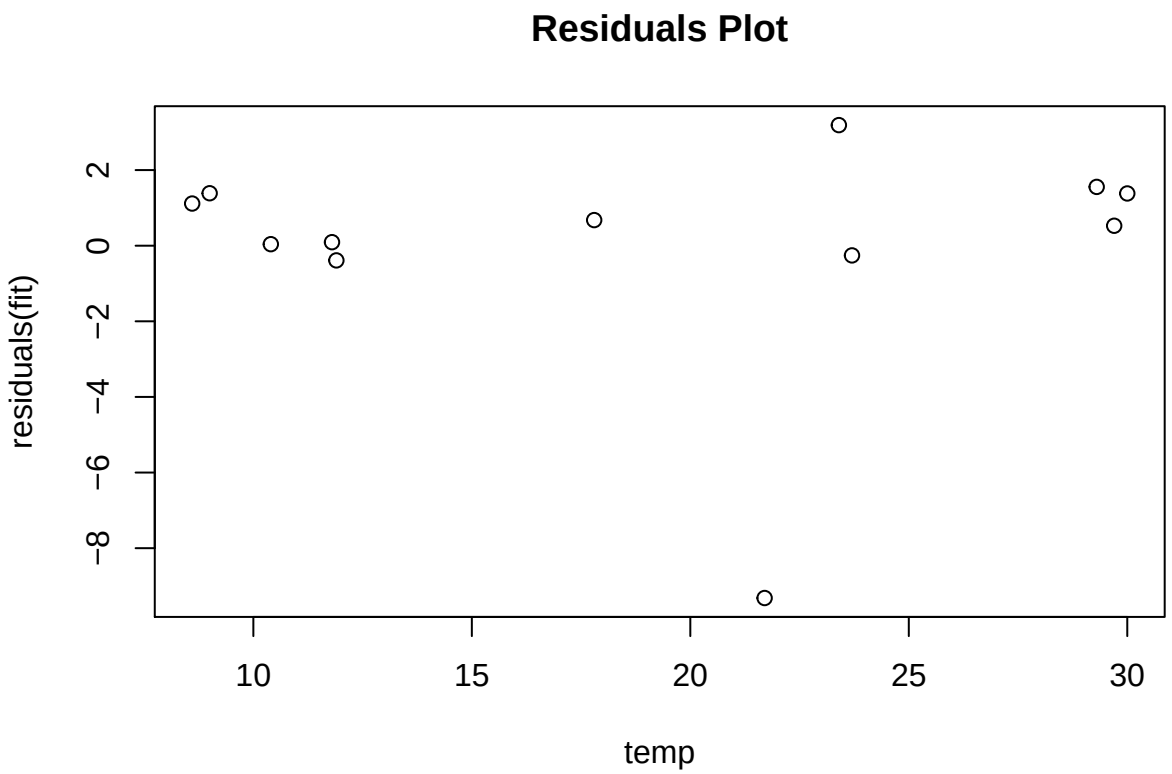
```
data4.1 <- data.frame(Mwh=data.mwh, temp=data.temp)
plot(Mwh ~ temp, data=data4.1, main="Electricity Consumption (MWh) versus Temperature (Celsius)",  xlab=
fit  <- lm(Mwh ~ temp, data=data4.1)
abline(fit, col='red')
```




**Answer:** As we can see on the above plot, there is a negative relationship between consumption of energy and temperature.

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

```r
plot(residuals(fit) ~ temp, data=data4.1, main="Residuals Plot")
```




TODO: look for influential obervations; there is an outlier for sure but otherwise looks like no systemic patterns so the model should be adequate.

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

(and)

d) Give prediction intervals for your forecasts.

```r
forecast(fit, newdata=data.frame(temp=c(10,35)))
```

```
##   Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
## 1       18.23241 13.378919 23.08591 10.351355 26.11347
## 2       13.71495  8.413211 19.01668  5.106039 22.32385
```

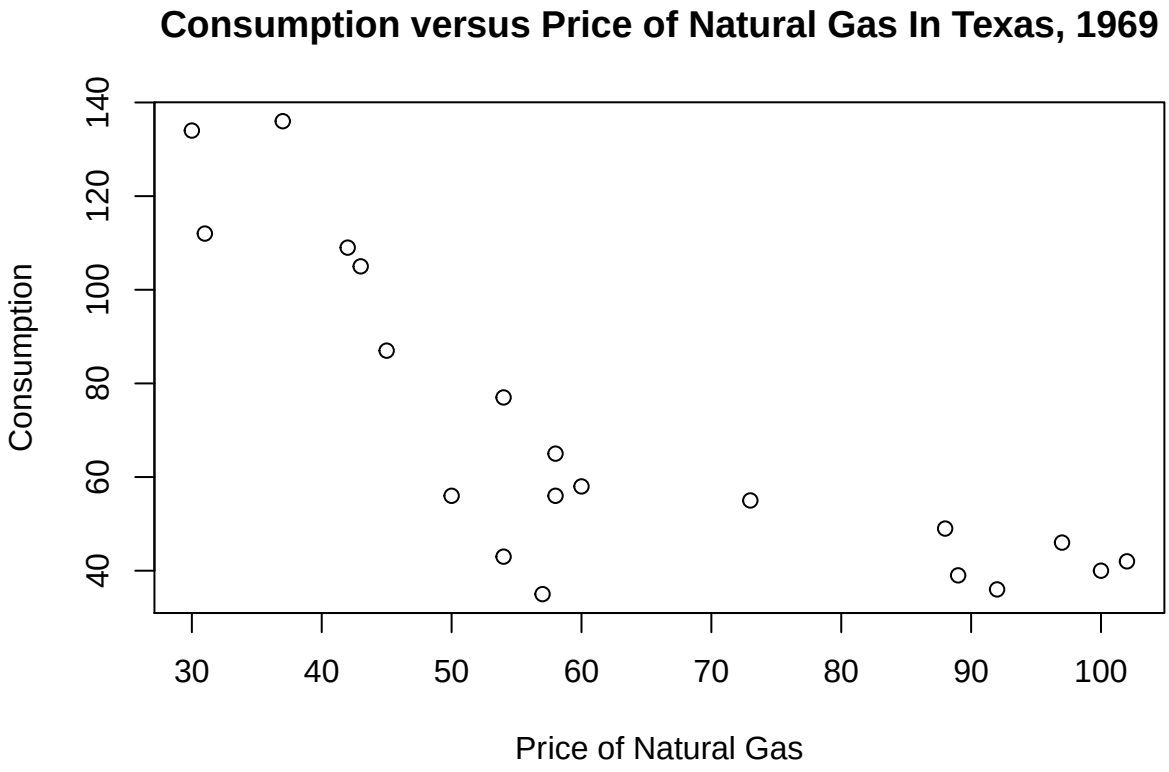
TODO: explain

## Question 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.

a) Do a scatterplot of consumption against price. The data are clearly not linear. Three possible nonlinear models for the data are given below; The second model divides the data into two sections, depending on whether the price is above or below 60 cents per 1,000 cubic feet.

```
plot(consumption ~ price, data = texasgas, main = "Consumption versus Price of Natural Gas In Texas, 19
    xlab = "Price of Natural Gas", ylab = "Consumption")
```



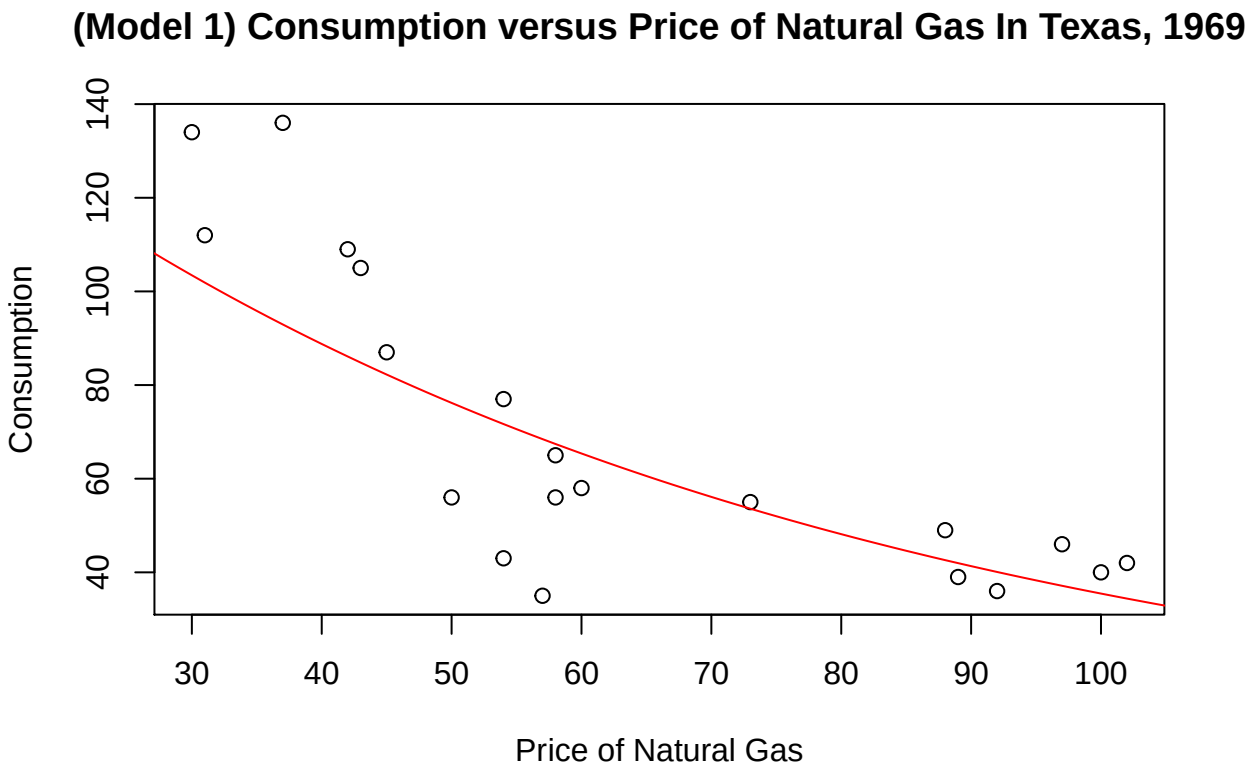

b) Can you explain why the slope of the fitted line should change with P?

**Answer:** The derivative of this graph is the rate of consumtpion, which I do not think would be constant. This means the consumption curve of our model cannot be a simple line, because its derivative should not be constant. Why? Well consumption of natural gas is probably higher when prices are low due to over consumption, and lower after hitting a threshold price where people would rather do without natural gas than pay a high price. Generally speaking, there is a relationship between a good's price and the rate at which it is consumed.

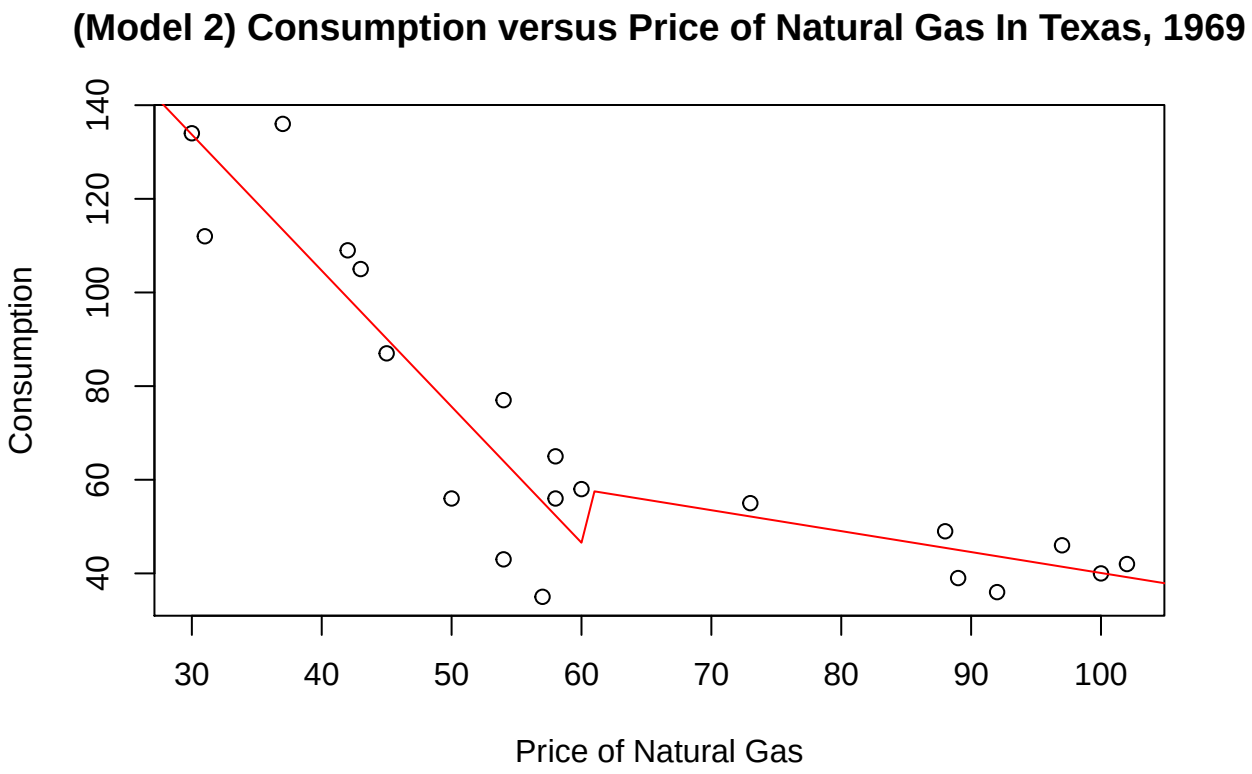
c) Fit the three models and find the coefficients, and residual variance in each case. For the second model, the parameters a1, a2, b1, b2 can be estimated by simply fitting a regression with four regressors but no constant: (i) a dummy taking value 1 when P<=60 and 0 otherwise; (ii) P1=P when P<=60 and 0 otherwise; (iii) a dummy taking value 0 when P<=60 and 1 otherwise; (iv) P2=P when P>60 and 0 otherwise.

**Answer:** For each model, I fit them against the data (or constructed predictors) and plot the model, in red, versus the real data, in black.

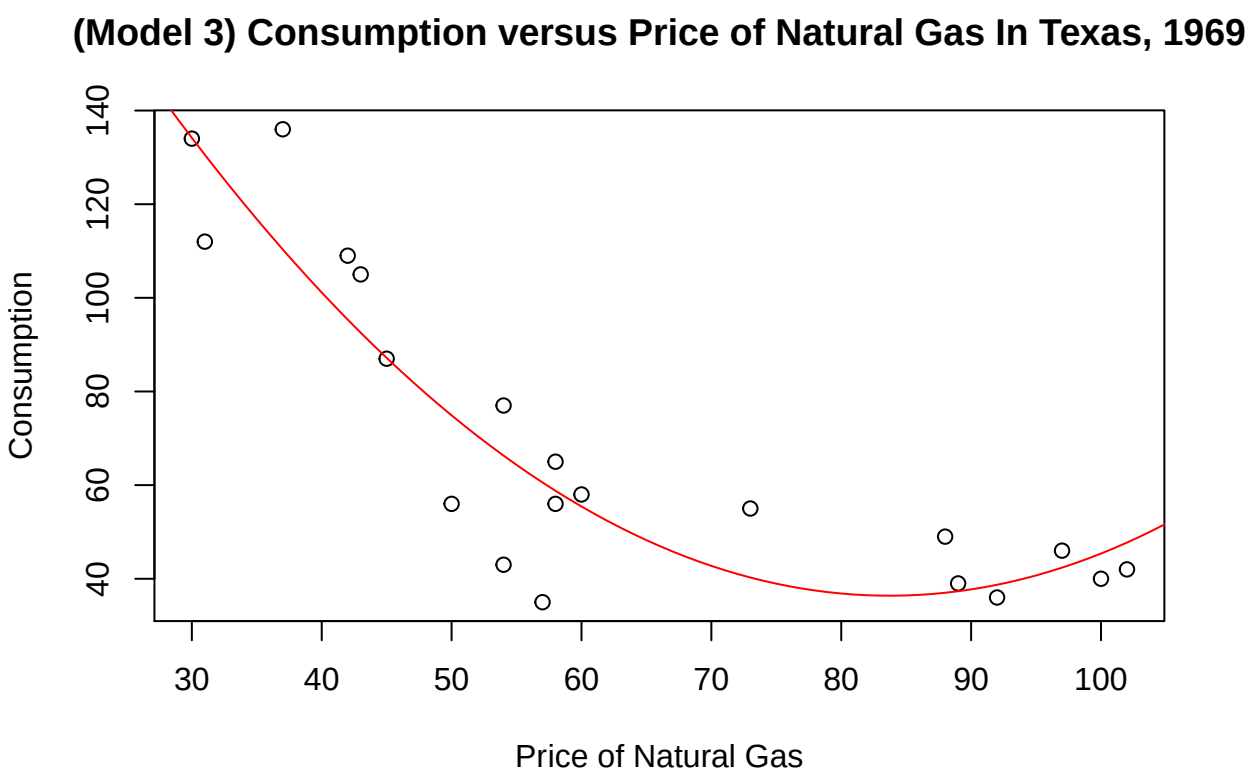
```
prices <- seq(20, 110, by=1)

# Model 1
model1 <- lm(log(consumption) ~ price, data=texasgas)
model1.predict <- function(input) {
  return(exp(model1$coef["price"] * input + model1$coef["(Intercept)"]))
}

# Plot Model1 and Data
plot(consumption ~ price, data=texasgas, main="(Model 1) Consumption versus Price of Natural Gas In Texa
lines(prices, model1.predict(prices), col='red')
```



```r
# Need some predictors for model 2
texasgas$priceLess60 <- ifelse(texasgas$price <= 60, texasgas$price, 0)
texasgas$dummy1 <- as.numeric(texasgas$priceLess60 > 0)
texasgas$priceGreater60 <- ifelse(texasgas$price > 60, texasgas$price, 0)
texasgas$dummy2 <- as.numeric(texasgas$priceGreater60 > 0)

# Model 2
model2 <- lm(consumption ~ 0 + priceLess60 + dummy1 + priceGreater60 + dummy2, data=texasgas)
model2.predict <- function(input) {
  return(ifelse(input <= 60, model2$coef["priceLess60"] * input + model2$coef["dummy1"], model2$coef["p
}

# Plot model2 and data
plot(consumption ~ price, data=texasgas, main="(Model 2) Consumption versus Price of Natural Gas In Texa
lines(prices, model2.predict(prices), col='red')
```



```r
# We need a non-linear predictor for model3
texasgas$price_squared <- texasgas$price ^ 2

# Model 3
model3 <- lm(consumption ~ price + price_squared, data=texasgas)
model3.predict <- function(input) {
  return(model3$coef["price"] * input + model3$coef["price_squared"] * input^2 + model3$coef["(Intercep
}

# Plot Model and data
plot(consumption ~ price, data=texasgas, main="(Model 3) Consumption versus Price of Natural Gas In Texa
lines(prices, model3.predict(prices), col='red')
```

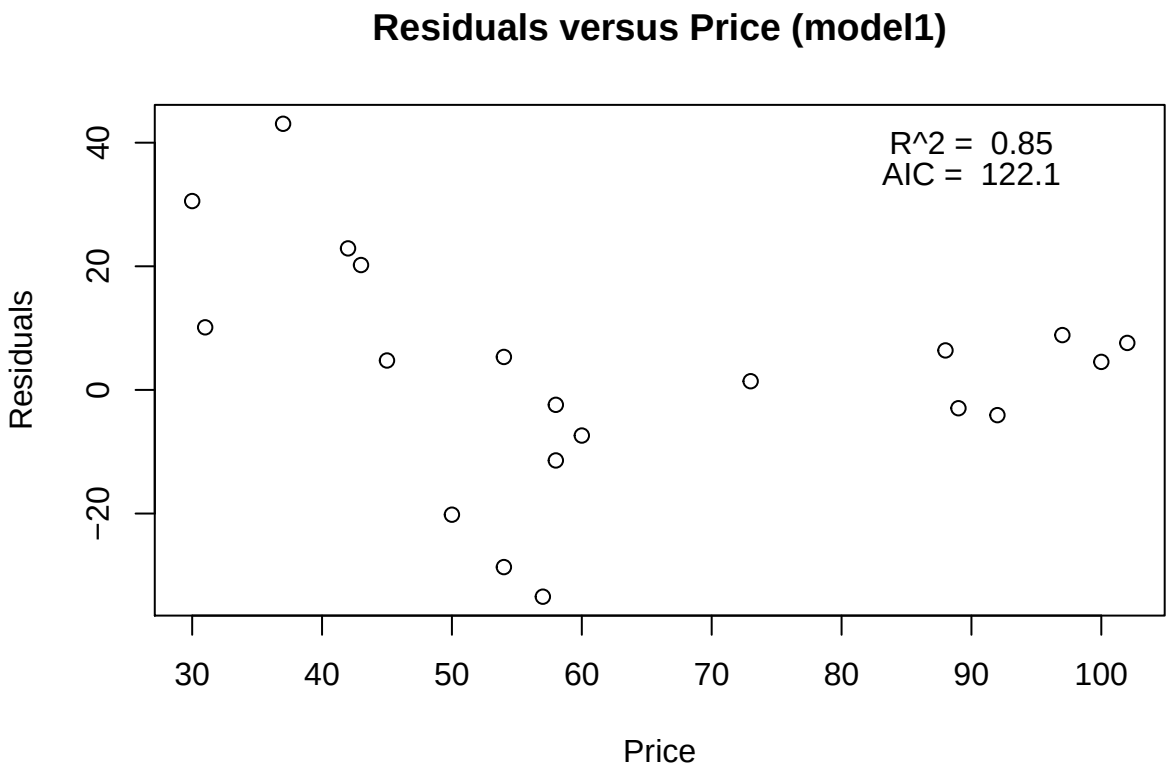


d) For each model, find the value of R2 and AIC, and produce a residual plot. Comment on the adequacy of the three models.

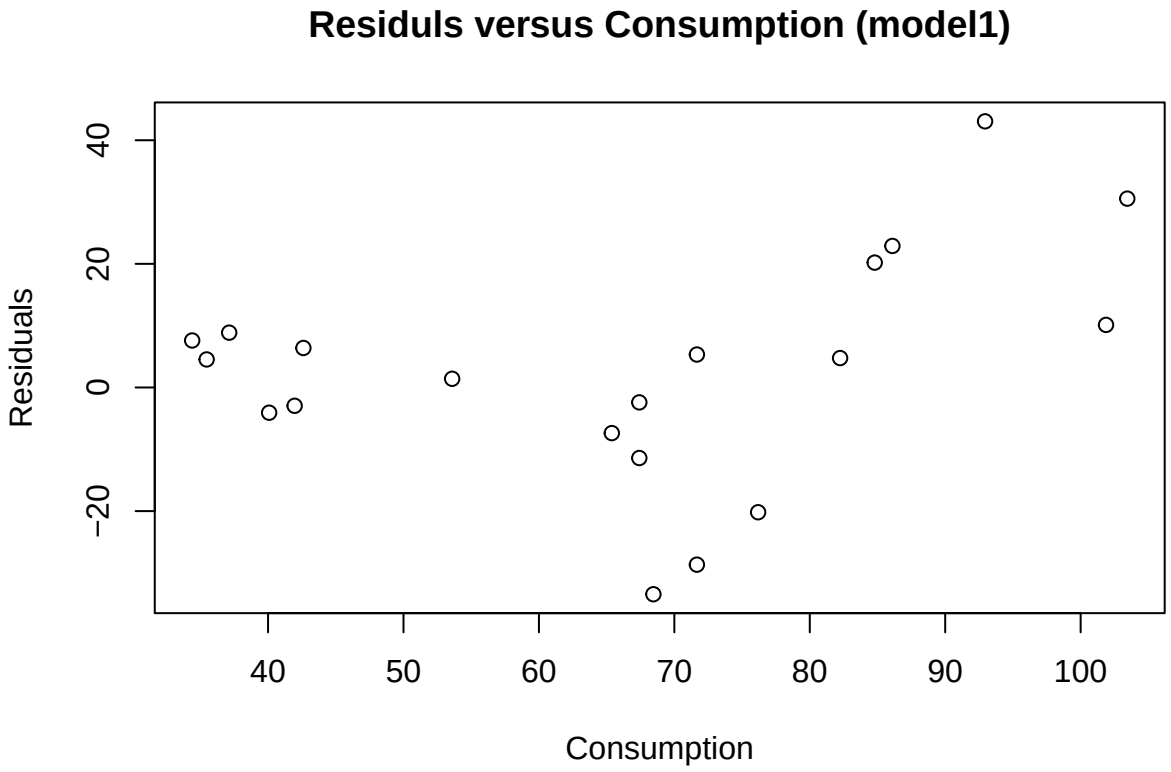
**Answer**: I am a bit surprised: From the R squared and AIC measures, Model 2 would be considered the best. Eyeballing it, I probably would have picked Model 3, which goes to show you why you don't eyeball these matters.

Model2 is interesting because it is jagged at the inflextion point of P=60. Predictions around here are going to take a discontinuous jump around this point.

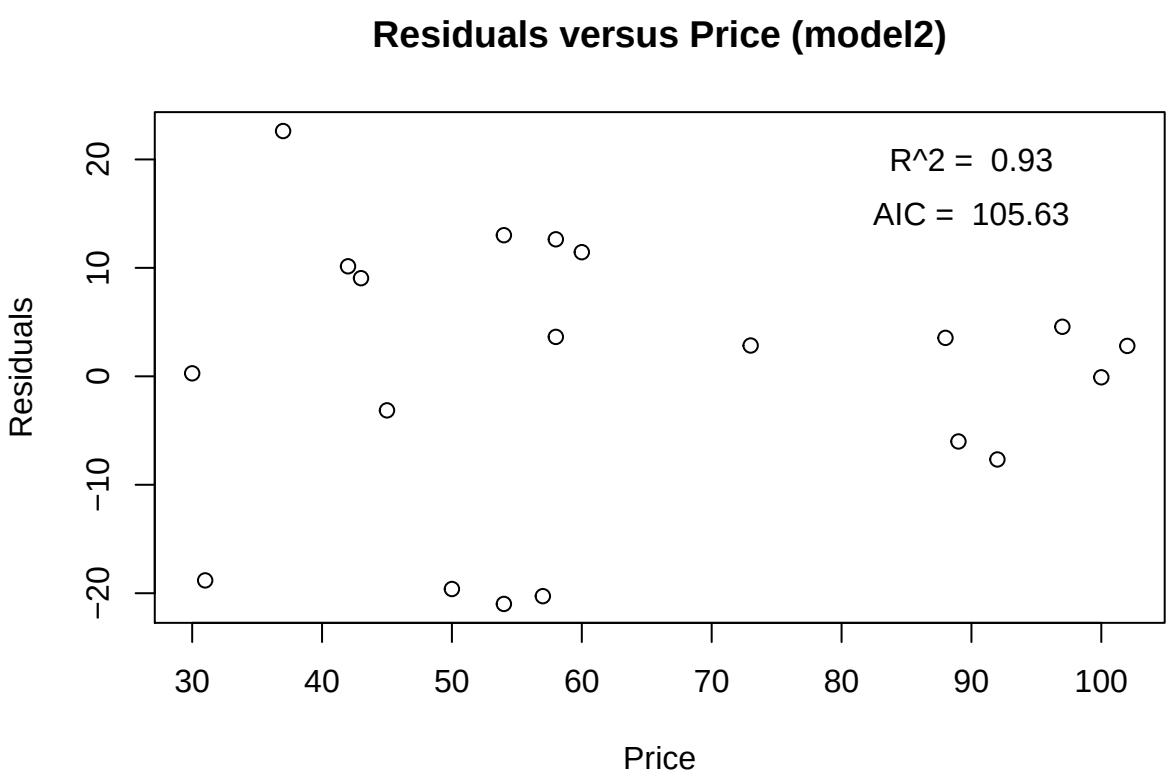
```
N <- nrow(texasgas)
model1.residuals <- texasgas$consumption - model1.predict(texasgas$price)
model1.rsquared <- cor(texasgas$consumption, model1.predict(texasgas$price))
model1.SSE <- sum(model1.residuals^2)
model1.k <- 1
model1.AIC <- N * log(model1.SSE / N) + 2 * (model1.k + 2)

# Plot residuals versus predictor(s)
plot(texasgas$price, model1.residuals, main="Residuals versus Price (model1)", xlab="Price", ylab="Resi
text(90, 40, labels=c(paste("R^2 = ",  round(model1.rsquared, digits=2))))
text(90, 35, labels=c(paste("AIC = ",  round(model1.AIC, digits=2))))
```



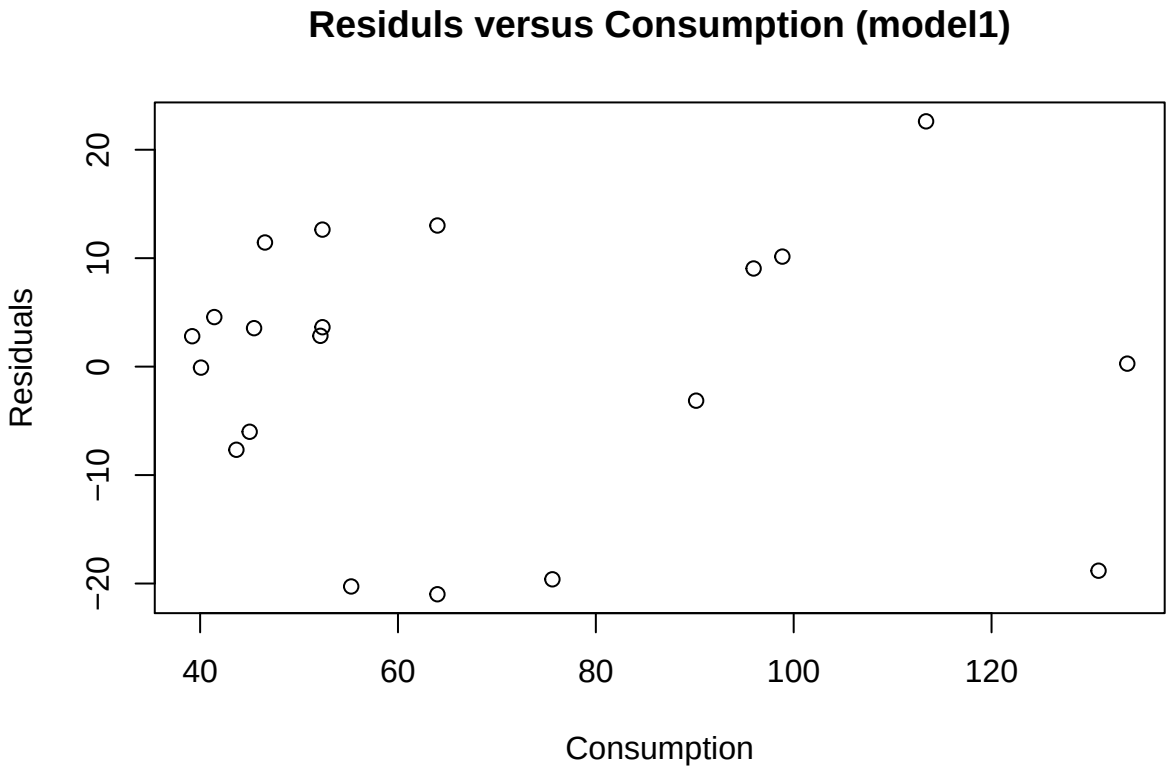

```
# Plot residuals versus predicted consumption
plot(model1.predict(texasgas$price), model1.residuals, main="Residuls versus Consumption (model1)", xlab
```


Residuls versus Consumption (model1)
Residuals
−20
0
20
40
40
50
60
70
80
90
100
Consumption

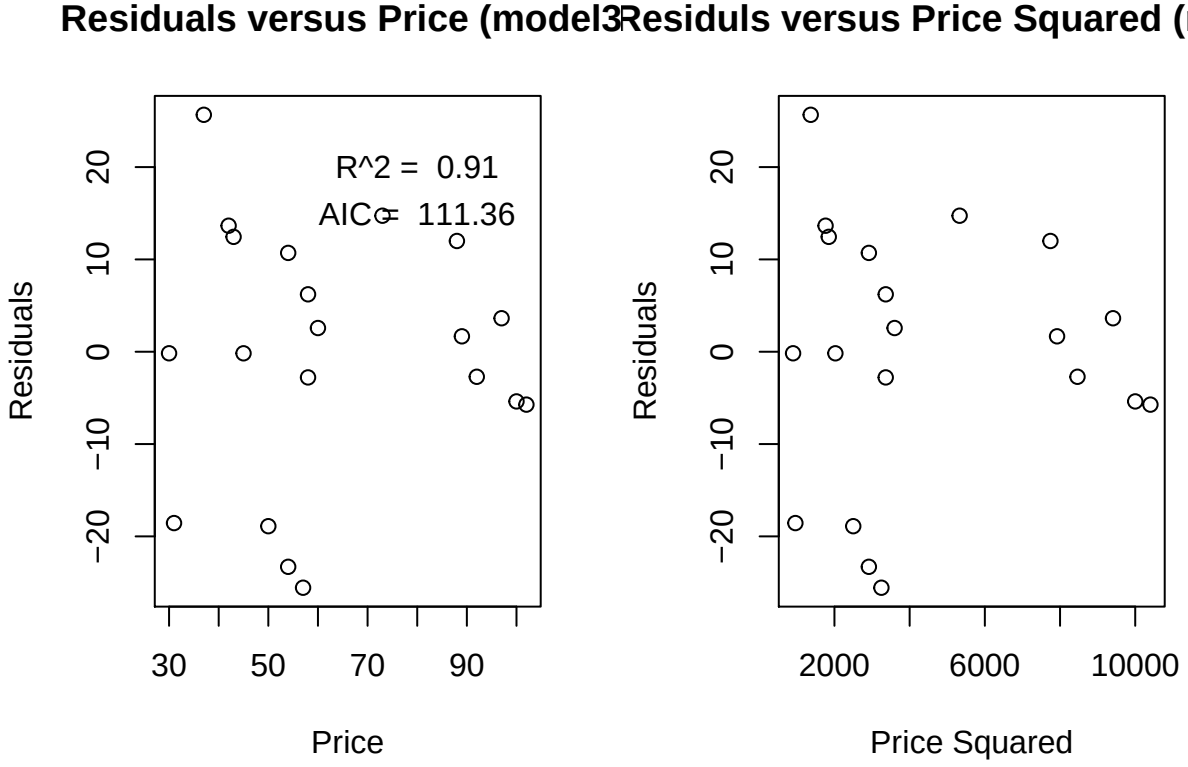
```r
model2.residuals <- resid(model2)
model2.rsquared <- cor(texasgas$consumption, model2.predict(texasgas$price))
model2.SSE <- sum(model2.residuals^2)
model2.k <- 1
model2.AIC <- N * log(model2.SSE / N) + 2 * (model2.k + 2)

# Plot residuals versus predictor(s)
plot(texasgas$price, model2.residuals, main="Residuals versus Price (model2)", xlab="Price", ylab="Resi
text(90, 20, labels=c(paste("R^2 = ",  round(model2.rsquared, digits=2))))
text(90, 15, labels=c(paste("AIC = ",  round(model2.AIC, digits=2))))
```



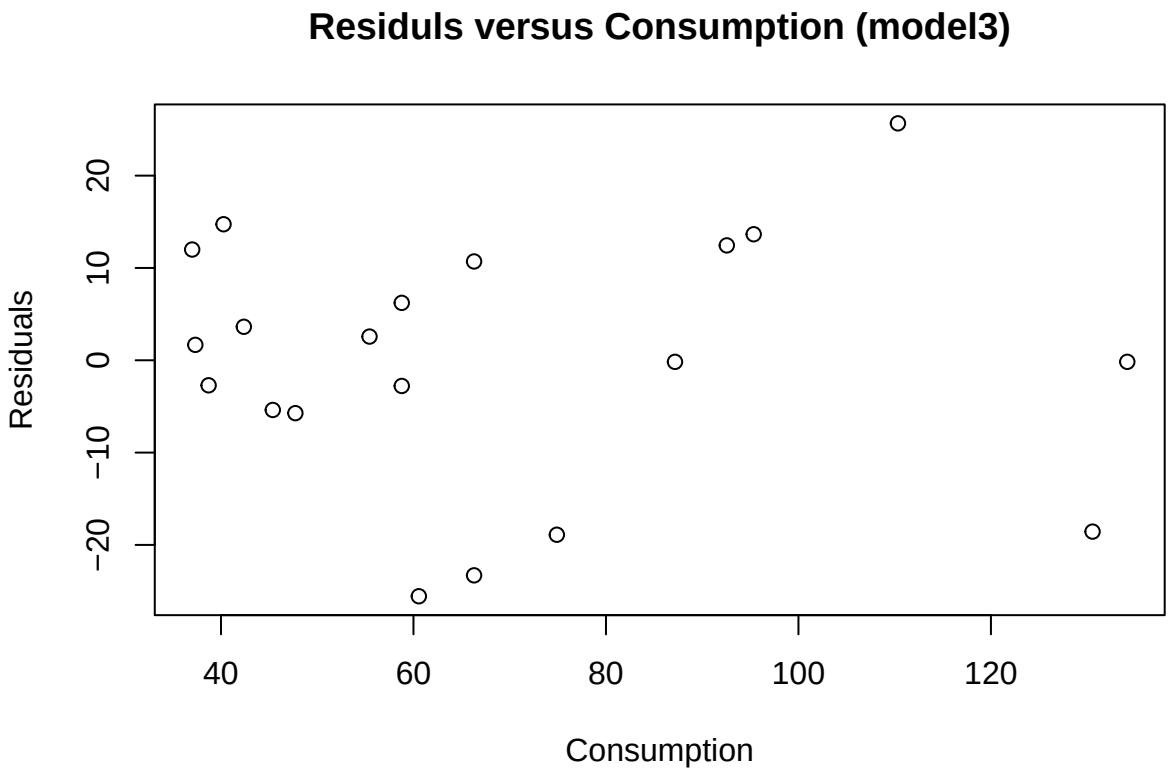

```r
# Plot residuals versus predicted consumption
plot(model2.predict(texasgas$price), model2.residuals, main="Residuls versus Consumption (model1)", xla
```


Residuls versus Consumption (model1)
Residuals
20
10
0
−10
−20
40
60
80
100
120
Consumption

```r
model3.residuals <- resid(model3)
model3.SSE <- sum(model3.residuals^2)
model3.k <- 2   # Two predictors
model3.rsquared <- cor(texasgas$consumption, model3.predict(texasgas$price))
model3.AIC <- N * log(model3.SSE / N) + 2 * (model3.k + 2)

# Plot residuals versus predictor(s)
par(mfrow=c(1,2))
plot(texasgas$price, model3.residuals, main="Residuals versus Price (model3)", xlab="Price", ylab="Resi
text(80, 20, labels=c(paste("R^2 = ",  round(model3.rsquared, digits=2))))
text(80, 15, labels=c(paste("AIC = ",  round(model3.AIC, digits=2))))
plot(texasgas$price_squared, model3.residuals, main="Residuls versus Price Squared (model3)", xlab="Pri
```




```r
# Plot residuals versus predicted consumption
par(mfrow=c(1,1))
plot(model3.predict(texasgas$price), model3.residuals, main="Residuls versus Consumption (model3)", xla
```




f) 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.

```
input_prices <- c(40, 60, 80, 100, 120)
predictions <- model2.predict(input_prices)
predictions
```

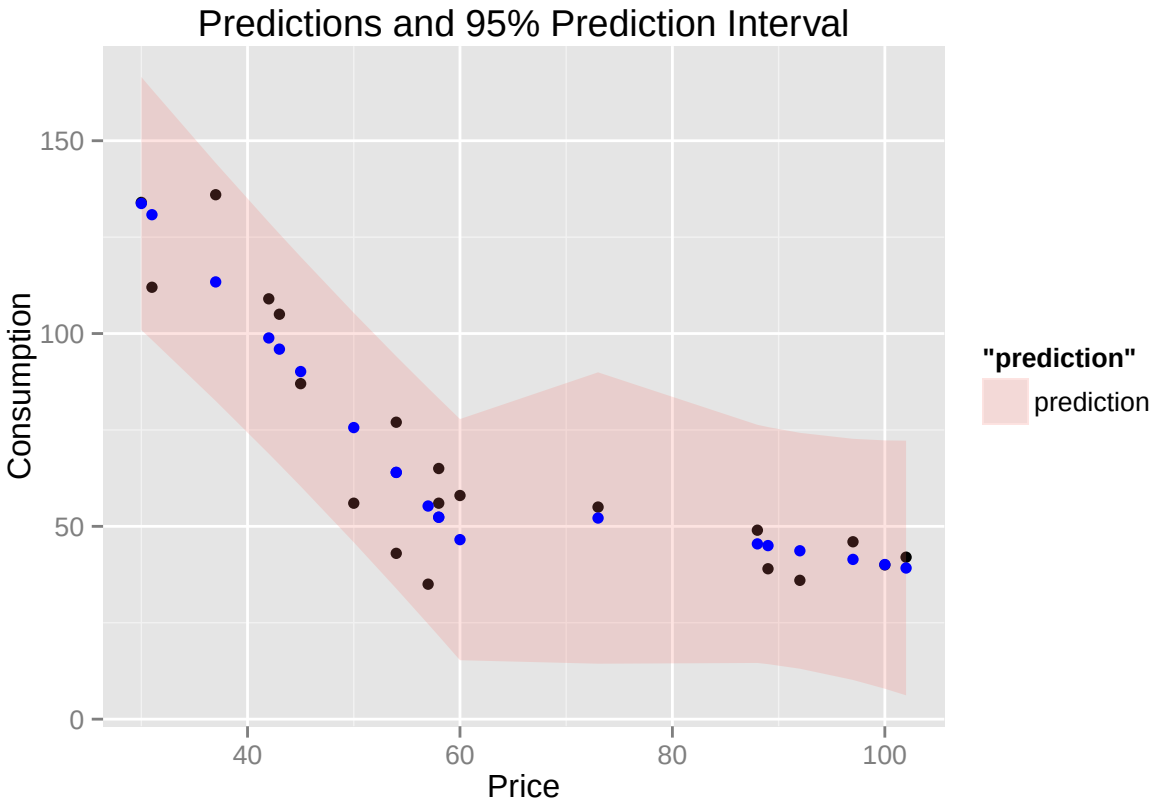
```
## [1] 104.66623  46.55289  49.02913  40.08989  31.15065
```

g) Compute 95% prediction intervals. Make a graph of these prediction intervals and discuss their interpretation.

```
# Calculate intervals
#for(price in prices) {
#  print(ifelse(price <= 60, predict(model2, data.frame(priceLess60=c(price), dummy1=c(1), #dummy2=c(0)
#}

# Print intervals w/ predictions
texasgas_predict <- data.frame(texasgas, predict(model2, interval = 'prediction'))
```

```
## Warning in predict.lm(model2, interval = "prediction"): predictions on current data refer to _future
```

```
ggplot(texasgas_predict, aes(x=price, y=consumption)) + geom_point()  +  geom_ribbon(aes(y = fit, ymin =
```




h) 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?