Kumar - SDS 291 Homework 4

HOMEWORK 4: MULTIPLE REGRESSION MODELS

Load libraries

```
library(Stat2Data)
library(ggplot2)
library(equatiomatic)
library(performance)
tests = c("qq", "linearity")
```

EXERCISE 1

To test this theory, the biologists run an experiment where the oxygen consumption rate of shore crabs is measured while the crabs are either exposed to the noise of a cargo ship engine for 7.5 minutes, or experience 7.5 minutes of ambient ocean noise.

Load data

```
data("CrabShip")
head(CrabShip)
```

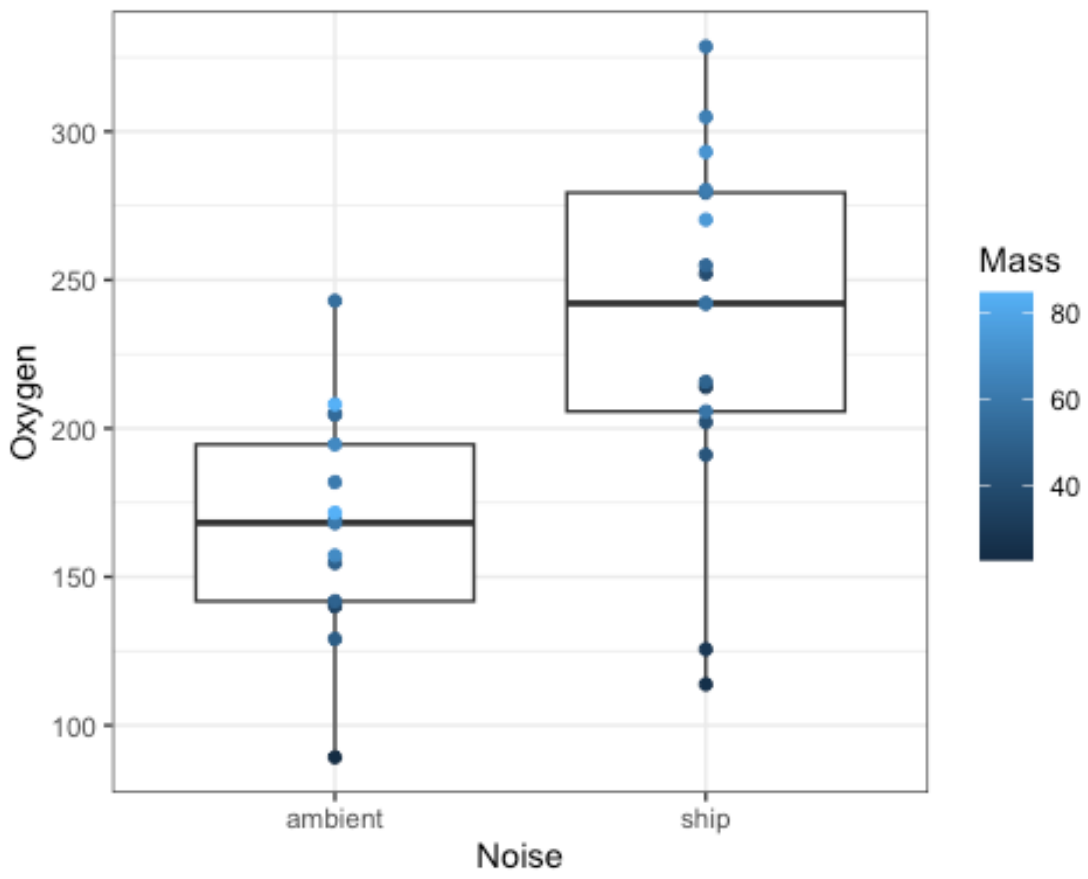
```
##   Mass Oxygen   Noise
## 1 22.7   89.2 ambient
## 2 34.6  141.1 ambient
## 3 36.0  140.1 ambient
## 4 40.1  204.9 ambient
## 5 47.5  129.1 ambient
## 6 49.6  154.6 ambient
```

EXERCISE 1A:

The researchers hypothesized that exposure to ship noise would increase oxygen consumption relative to ambient noise conditions. Furthermore, the researchers believe that the magnitude of this increase would be the same for large crabs and small crabs.

Visualize the model that instantiates the researcher's expectations

```
ggplot(data = CrabShip, mapping = aes(y = Oxygen,
                                      x = Noise)) +
  geom_boxplot() +
  geom_point(aes(color = Mass)) +
  # geom_smooth(method = lm, formula = y ~ factor(x), se = F) +
  theme_bw()
```



Fit the regression model + report regression table

```
crab_noise_model = lm(Oxygen ~ Noise, data = CrabShip)
summary(crab_noise_model)

##
## Call:
## lm(formula = Oxygen ~ Noise, data = CrabShip)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -122.418  -27.396    1.679   35.543   92.382
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   168.67      11.78  14.322 1.81e-15 ***
## Noiseship      67.55      16.66   4.055    3e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.56 on 32 degrees of freedom
## Multiple R-squared:  0.3395, Adjusted R-squared:  0.3188
## F-statistic: 16.45 on 1 and 32 DF,  p-value: 0.0002996
```

Interpret the coefficients

```
extract_eq(crab_noise_model, use_coefs = T)
```

$$\widehat{\text{Oxygen}} = 168.67 + 67.55(\text{Noise}_{\text{ship}})$$

- `(Intercept)` = 168.67: the oxygen consumption level if the noise level is 0, aka ambient.
- `Noise_{ship}` = 67.55: the average increase in oxygen consumption for a crab from the original level when the noise rises to ship level, above ambient.

Does this model evidence the researcher's hypothesis?

The researcher's hypothesis pt 1: that exposure to ship noise would increase oxygen consumption relative to ambient noise conditions

There is evidence that there is some impact in oxygen consumption increases when the noise level is above ambient. The $Noise_{ship}$ value is a measure of how much the oxygen level changes when the noise level rises from ambient to ship level. This value being positive means that the average change in oxygen consumption is an increase (by 67.55), and the p-value = $0.0003 < 0.05$ indicates that this increase is statistically significant.

Hypothesis pt 2: the magnitude of this increase would be the same for large crabs and small crabs.

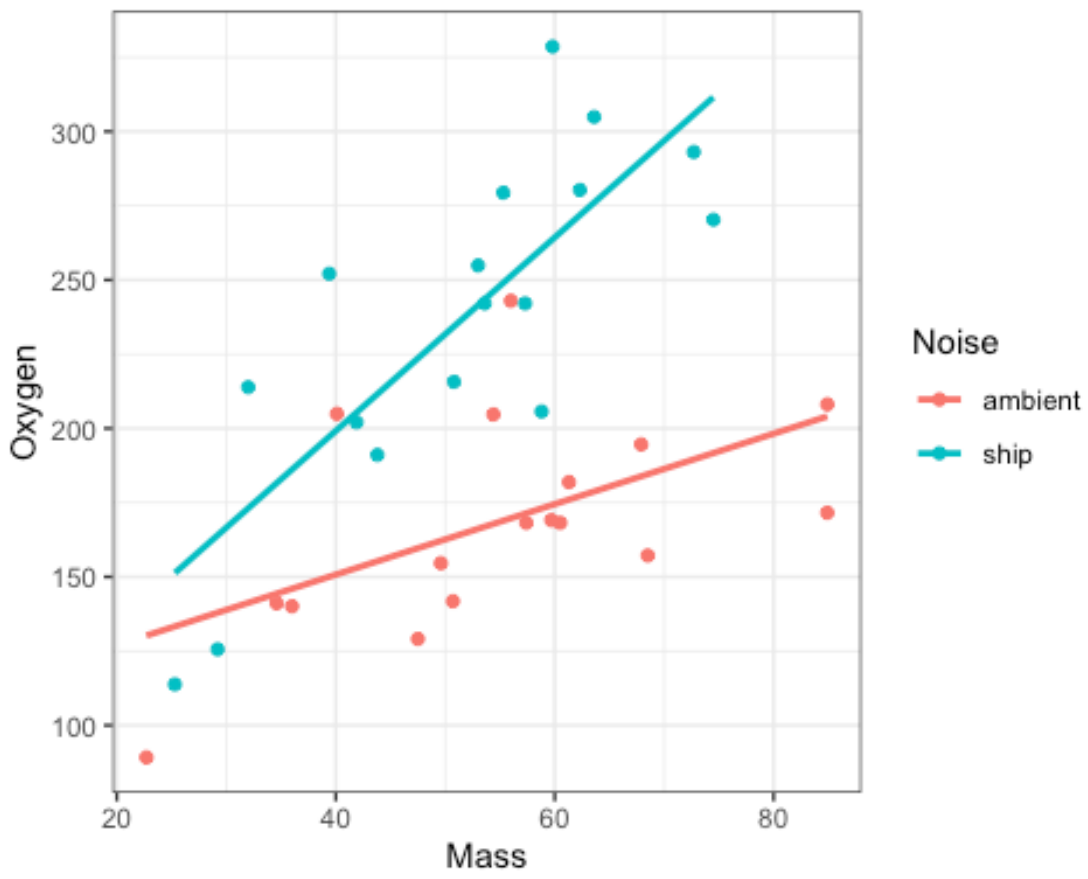
This model doesn't take into account the impact of mass, so we cannot conclude whether the effect of the change in noise level on oxygen consumption is the same across crab masses.

EXERCISE 1B:

Another research group argues that the model presented by the first group of marine biologists is actually mis-specified, and that the effect of ship noise on oxygen consumption actually depends on a crab's mass.

Visualize a new model to assess this possibility

```
ggplot(data = CrabShip, mapping = aes(x = Mass,
                                      y = Oxygen,
                                      color = Noise)) +
  geom_point() +
  geom_smooth(method = lm, se = F, formula = y ~ x) +
  theme_bw()
```



Fit the regression model

```
crab_mass_model = lm(Oxygen ~ Noise * Mass, data = CrabShip)
summary(crab_mass_model)

##
## Call:
## lm(formula = Oxygen ~ Noise * Mass, data = CrabShip)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.719 -21.350  -4.149  12.715  73.261
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    103.2703    29.3894   3.514  0.00142 **
## Noiseship      -34.3904    43.0782  -0.798  0.43096
## Mass             1.1869     0.5121   2.318  0.02746 *
## Noiseship:Mass   2.0705     0.7826   2.646  0.01286 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.9 on 30 degrees of freedom
```

```
## Multiple R-squared:  0.6982, Adjusted R-squared:  0.6681
## F-statistic: 23.14 on 3 and 30 DF,  p-value: 5.942e-08
```

Interpret the coefficients

```
extract_eq(crab_mass_model, use_coefs = T)
```

$$\widehat{\text{Oxygen}} = 103.27 - 34.39\left(\text{Noise}_{\text{ship}}\right) + 1.19(\text{Mass}) + 2.07\left(\text{Noise}_{\text{ship}} \times \text{Mass}\right)$$

- `(Intercept)` $= 103.27$: the model's projected oxygen consumption for a crab with mass 0 in an ambient noise condition.
- `Noise_{ship}` $= -34.39$: the average change in the oxygen consumption for the ship noise group relative to the ambient noise group.
- `Mass` $= 1.19$: the average increase in oxygen consumption for each increase in mass by 1 for the ambient noise condition.
- `Noise_{ship} * Mass` $= 2.07$: increase in average oxygen consumption for each increase in mass by one for all crabs in the ship noise group.

Does the model evidence the alternative hypothesis?

Alternative hypothesis: the effect of ship noise on oxygen consumption depends on a crab's mass

These researchers posit that the effect of the noise level on oxygen consumption also depends on a crab's mass. We can see from the visualization above as a crab's mass increases so too does their oxygen consumption, across both noise levels.

The `Mass` coefficient represents this change, showing that the average oxygen consumption for a crab increases by 2.07 for each increase in mass by 1 for both noise levels. The p-value $0.0000285 < 0.05$ clearly indicates that this difference is significant.

EXERCISE 2

Load data:

```
data("Pollster08")
head(Pollster08)
```

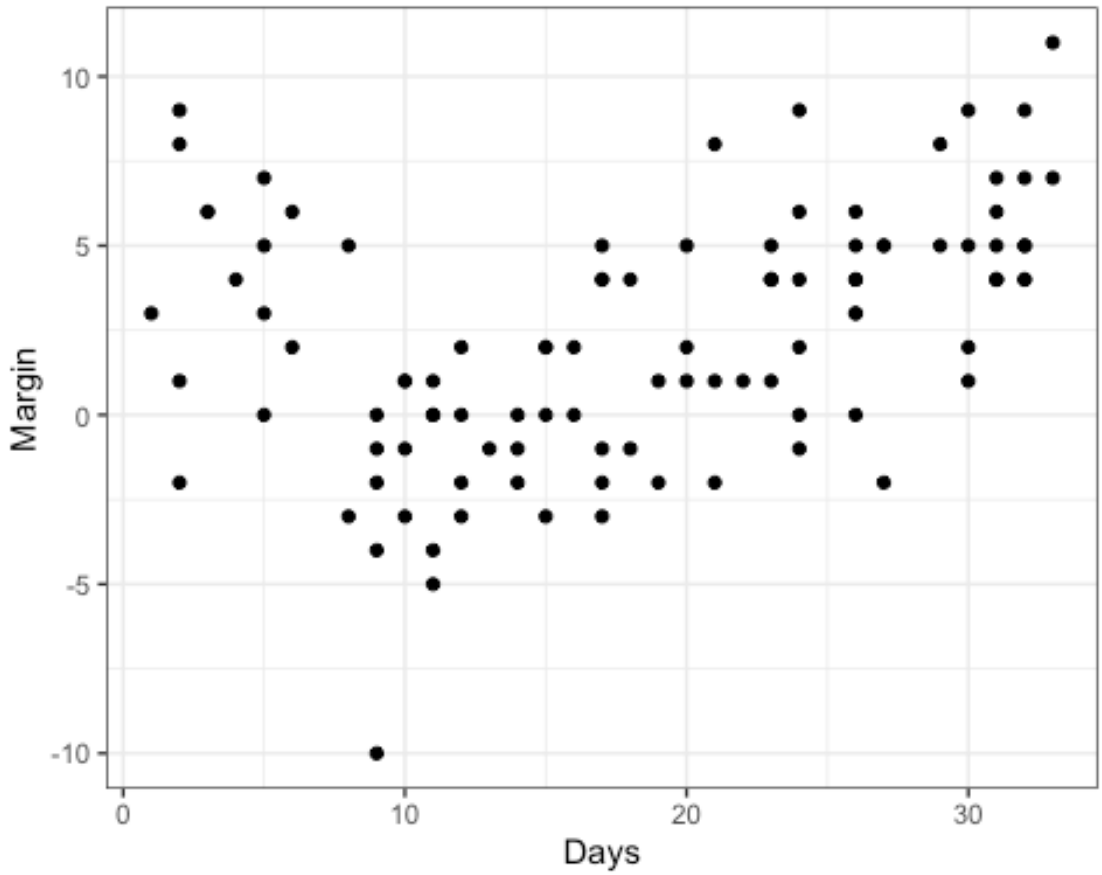
```
##        PollTaker   PollDates MidDate Days    n Pop McCain Obama Margin
Charlie
## 1      Rasmussen  8/28-30/08    8/29    1 3000  LV     46    49      3
0
## 2          Zogby  8/29-30/08    8/30    2 2020  LV     47    45     -2
0
## 3 Diageo/Hotline  8/29-31/08    8/30    2  805  RV     39    48      9
0
## 4            CBS  8/29-31/08    8/30    2  781  RV     40    48      8
0
## 5            CNN  8/29-31/08    8/30    2  927  RV     48    49      1
0
## 6      Rasmussen 8/30-9/1/08    8/31    3 3000  LV     45    51      6
```

```
0
##   Meltdown
## 1        0
## 2        0
## 3        0
## 4        0
## 5        0
## 6        0
```

EXERCISE 2A:

Create a scatterplot of the relationship between Obama's margin in the polls and the time when the poll was taken

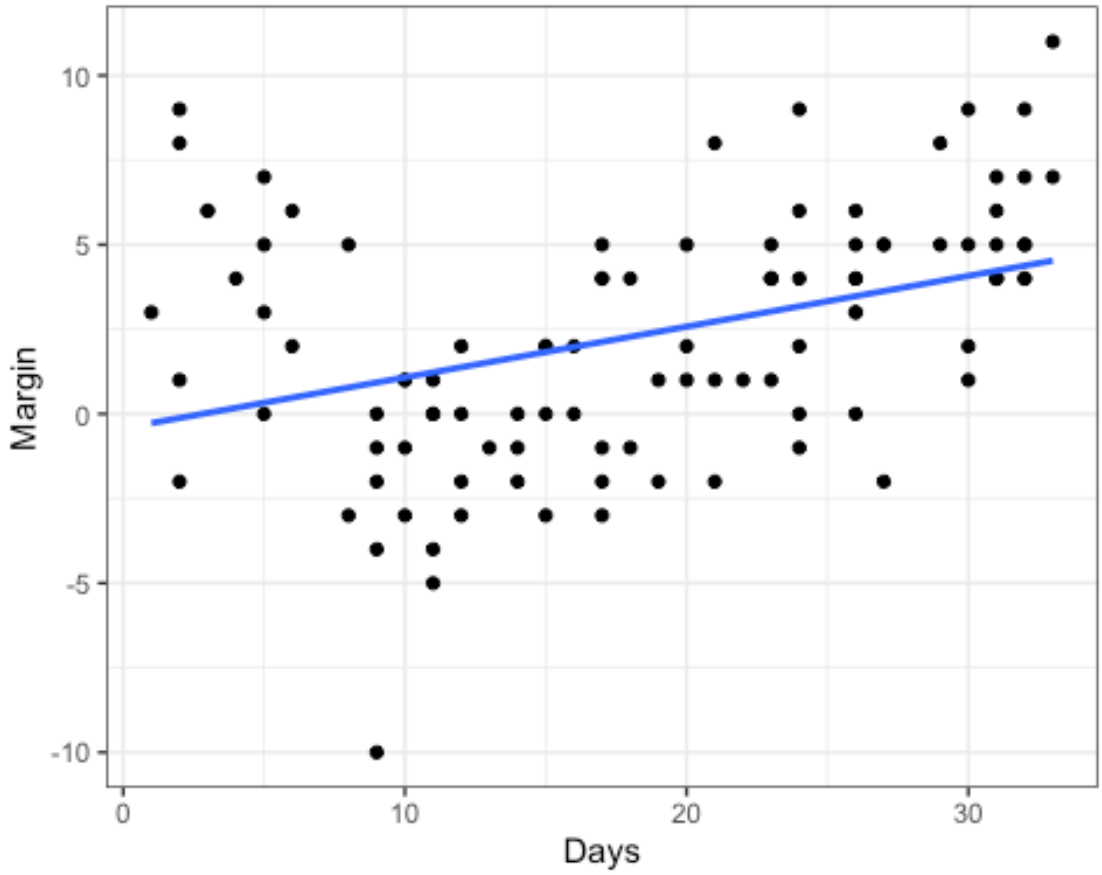
```
ggplot(data = Pollster08, mapping = aes(x = Days,
                                        y = Margin)) +
  theme_bw() +
  geom_point()
```



EXERCISE 2B:

Model Obama's margin in the polls as a function of time using an appropriate regression model.

Linear model



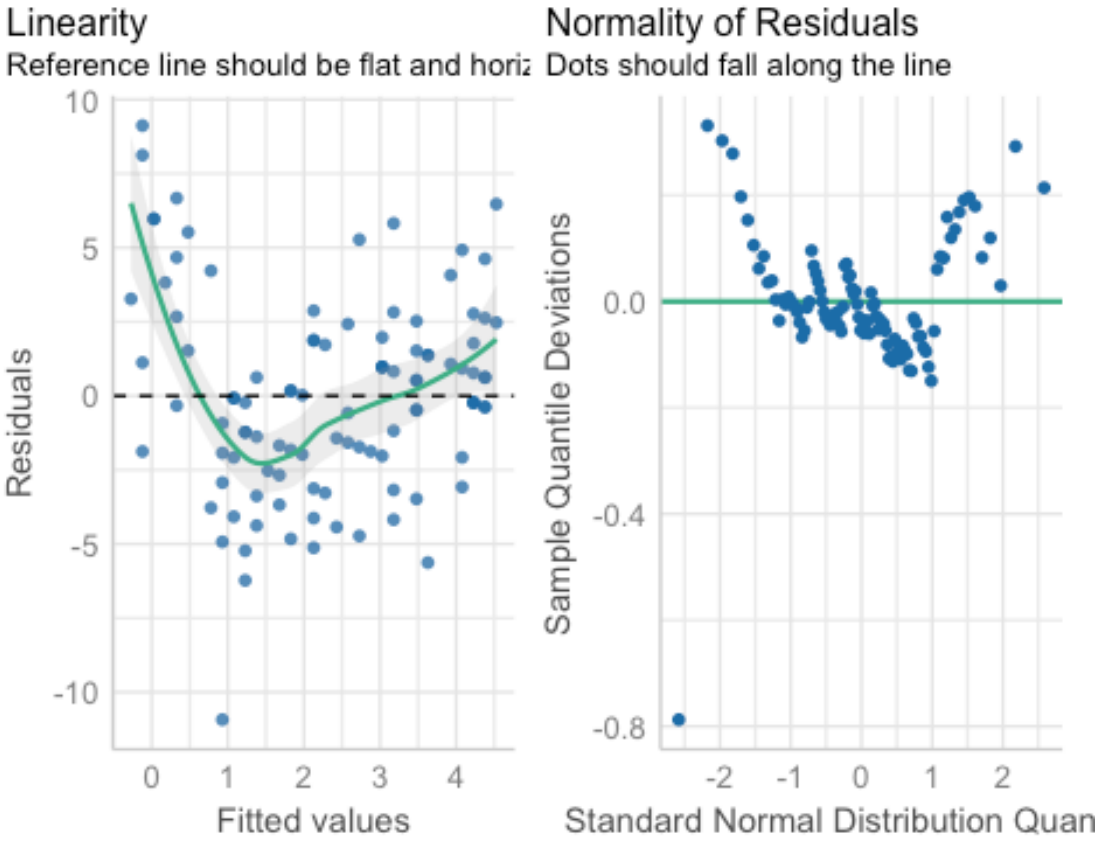

```
obama_linear = lm(Margin ~ Days, data = Pollster08)
summary(obama_linear)$r.squared

## [1] 0.1437693
```

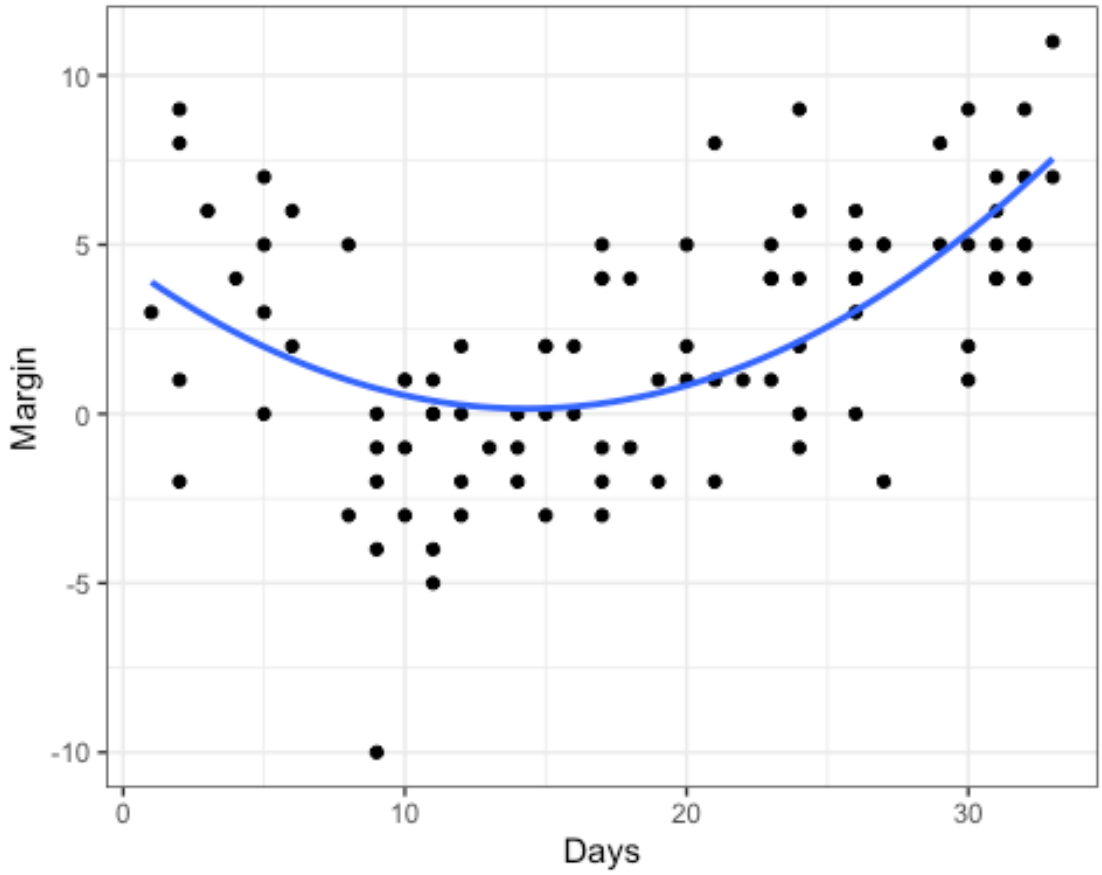
Linear residuals

These plots clearly show that this model is not a good fit for this data (no normal residuals, and no linearity). We should try another model.

```
check_model(obama_linear, check = tests)
```


Linearity
Reference line should be flat and horiz
Residuals
10
5
0
-5
-10
0
1
2
3
4
Fitted values
Normality of Residuals
Dots should fall along the line
Sample Quantile Deviations
0.0
-0.4
-0.8
-2
-1
0
1
2
Standard Normal Distribution Quan

Polynomial model



This version of the model explains slightly more:

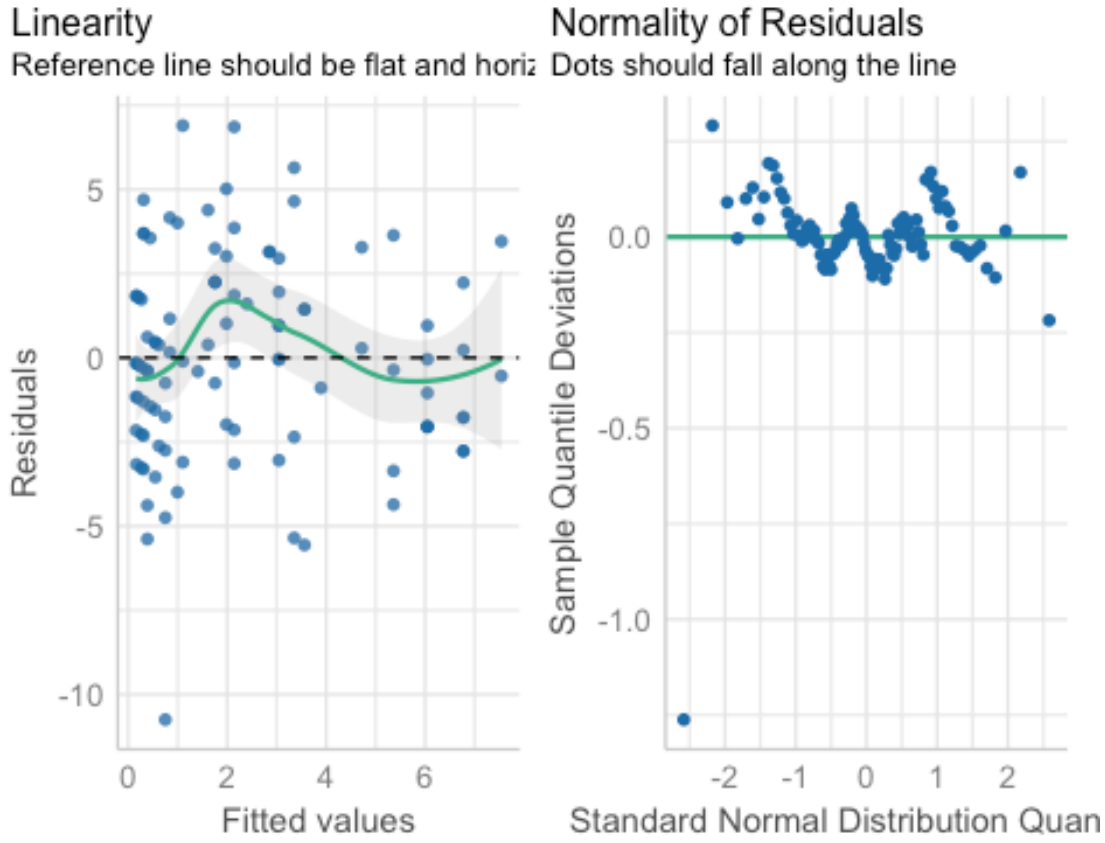
```
obama_polynomial = lm(Margin ~ I(Days^2) + Days, data = Pollster08)
summary(obama_polynomial)$r.squared

## [1] 0.3494697
```

Polynominal residuals

These residuals also have a pattern, but are much closer to what we expect – this violations are only mild. Thus, I believe we should use this model.

```
check_model(obama_polynomial, check = tests)
```



EXERCISE 2C:

If you used the model you fit in Exercise 2b on election night to predict the margin of the election, how accurate would your prediction have been?

```
predict(obama_polynomial, data.frame(Days = c(37)))
```

```
##        1
## 11.04033
```

Our predicted value was 11.04 percentage points, which, compared to the actual margin of 7.28 percentage points, was an overestimation by

```
## [1] 3.76
```