discussion week 13 - multi regression

Jeff Shamp

4/21/2020

Multiple Regression with Cars

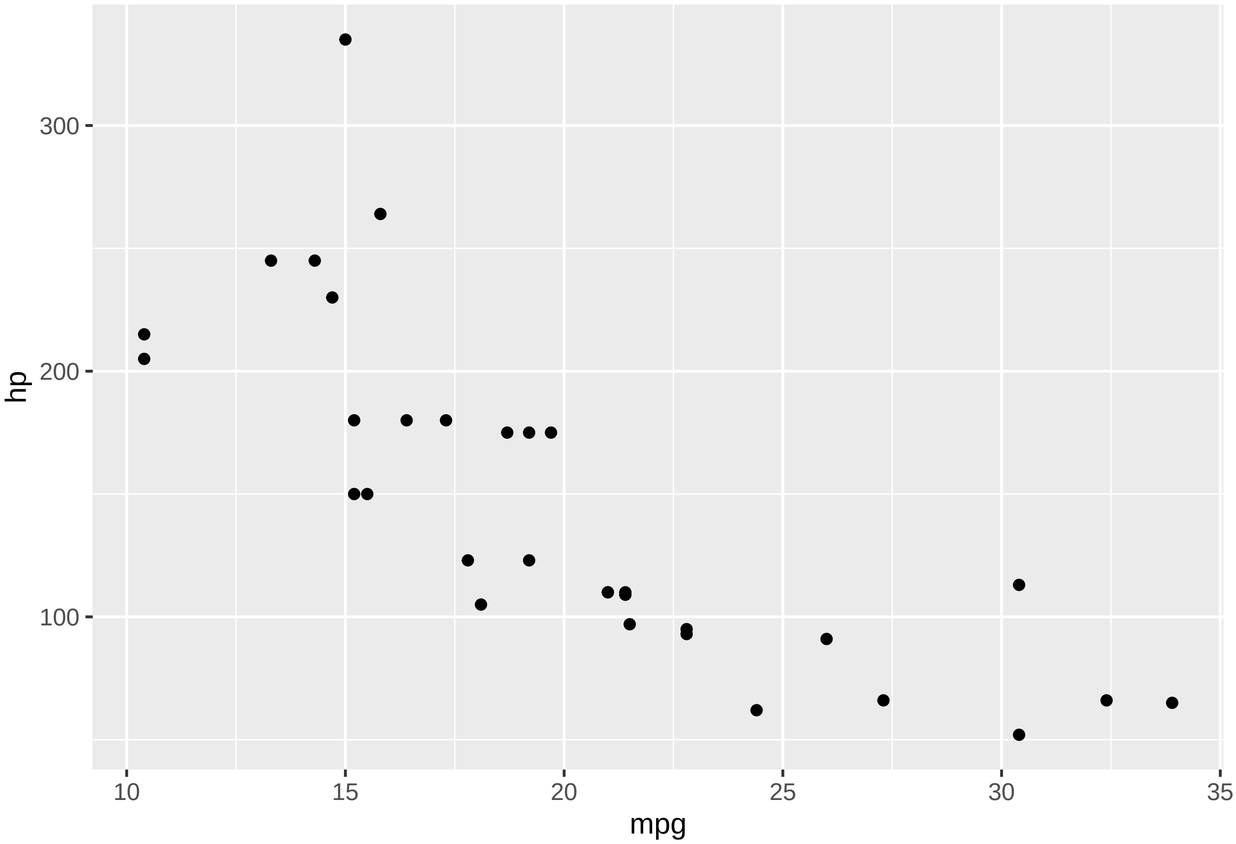
This is the well-used well loved 1974 Motor Trend Cars dataset.

```
data(mtcars)
head(mtcars)
```

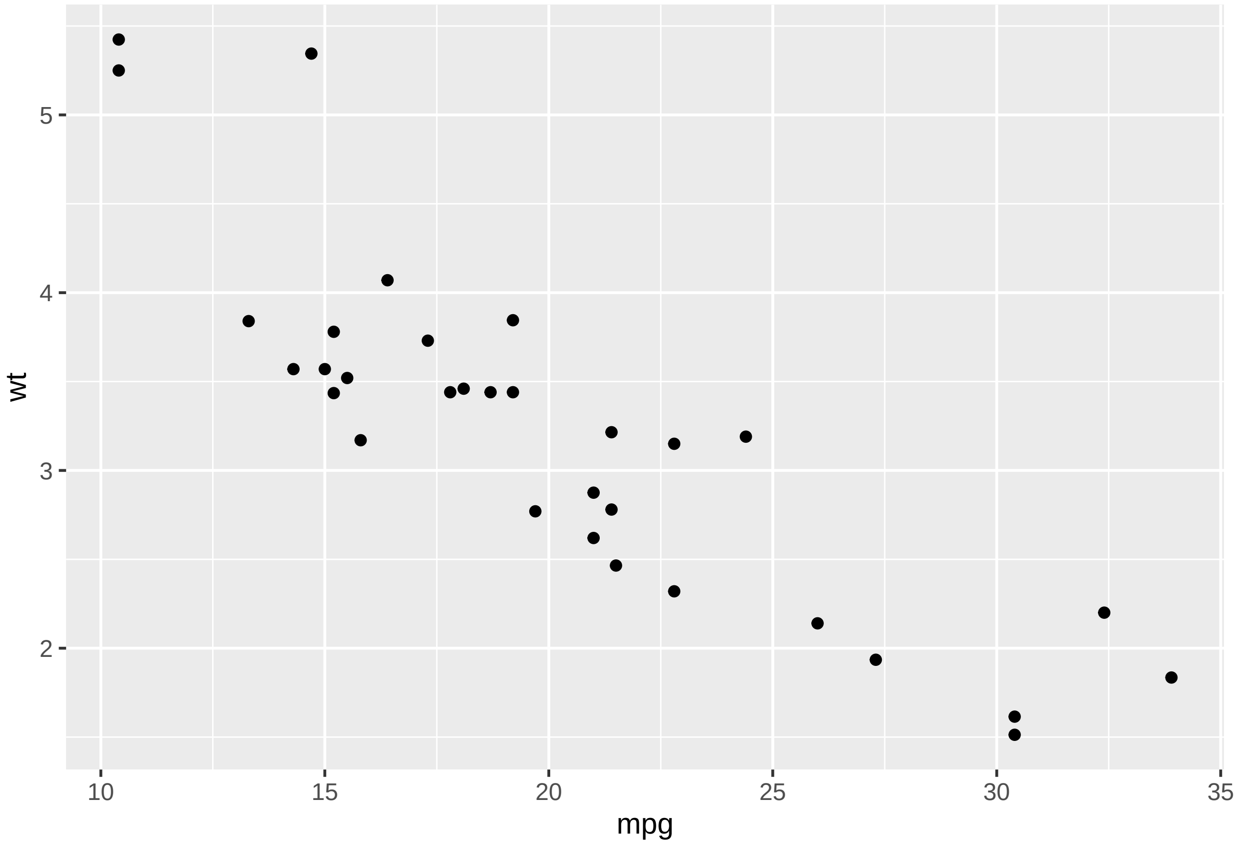
```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

We will plot a few variable and see what looks like a decent candidate for regression.

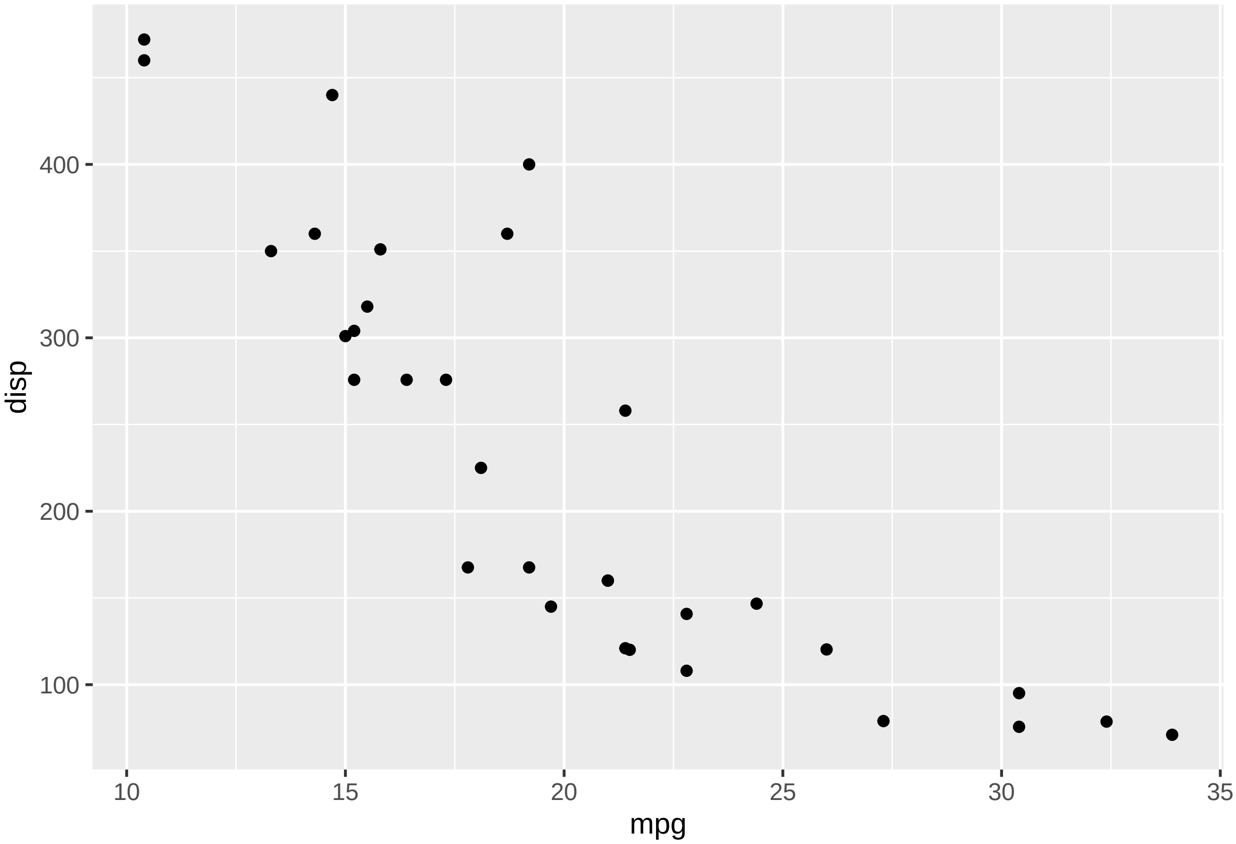
```
mtcars %>%
  ggplot(aes(x=mpg, y=hp)) +
  geom_point()
```



```
mtcars %>%
  ggplot(aes(x=mpg, y=wt)) +
  geom_point()
```



```
mtcars %>%
  ggplot(aes(x=mpg, y=disp)) +
  geom_point()
```



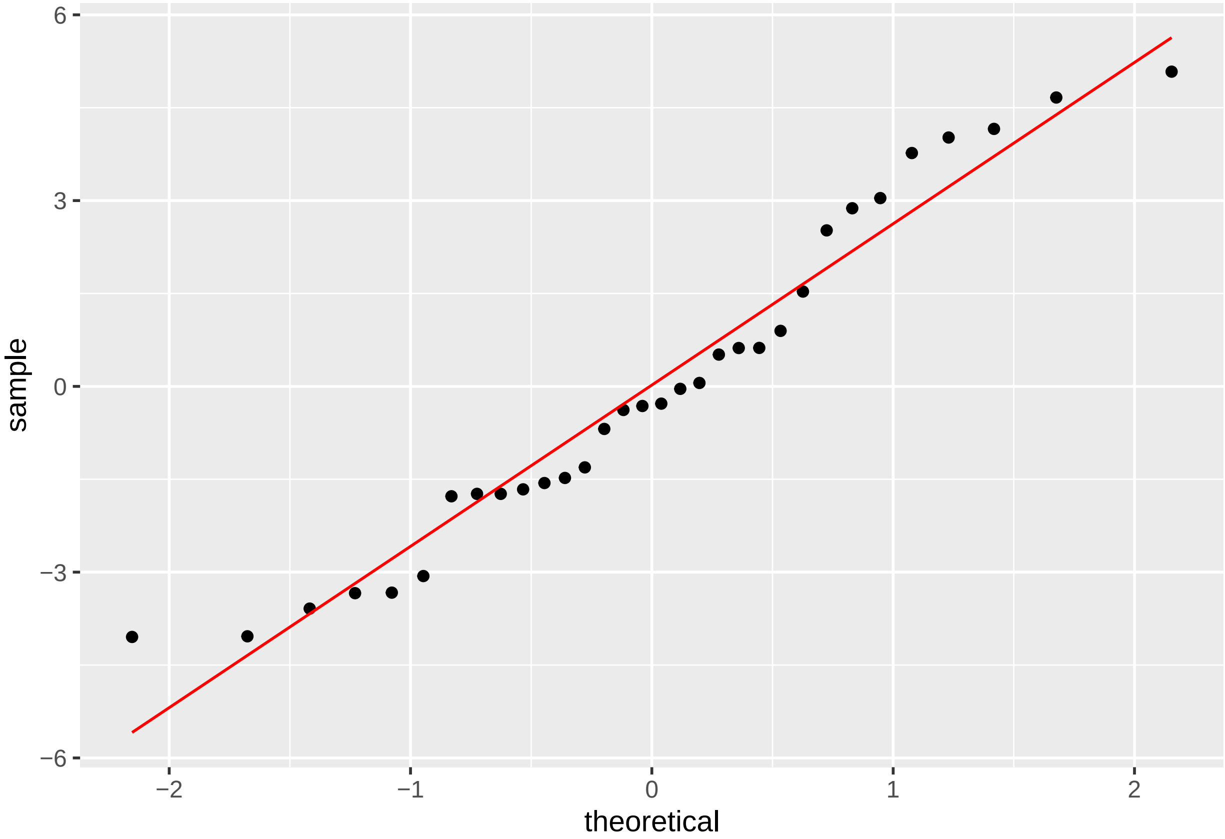
These are all fine options for quadratic regression on the surface. Since *there is no replacement for displacement*, I'll go with that variable.

We will go with Transmission type (am) for the binary variable.

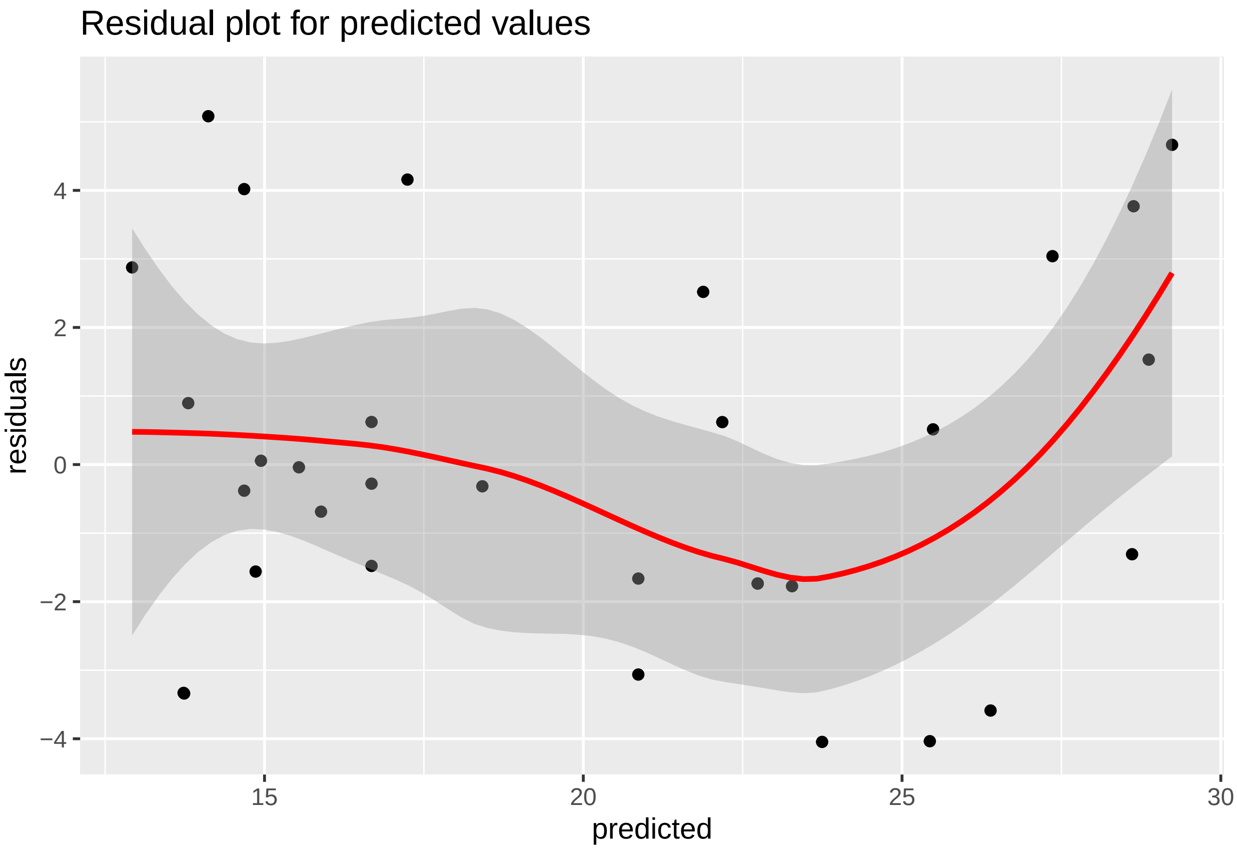
```
lm_1<- lm(data=mtcars,mpg ~ I(disp^2)+am+(disp*am))
summary(lm_1)
```

```
##
## Call:
## lm(formula = mpg ~ I(disp^2) + am + (disp * am), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0462 -1.7356 -0.2978  1.7781  5.0819
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.094e+01  4.410e+00   7.015 1.53e-07 ***
## I(disp^2)    7.768e-05  5.358e-05   1.450   0.1586
## am           4.382e+00  3.360e+00   1.304   0.2032
## disp        -7.312e-02  3.199e-02  -2.285   0.0304 *
## am:disp     -1.795e-02  1.460e-02  -1.230   0.2294
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.851 on 27 degrees of freedom
## Multiple R-squared:  0.8051, Adjusted R-squared:  0.7762
## F-statistic: 27.88 on 4 and 27 DF,  p-value: 3.075e-09
```

```
lm_1 %>%
  ggplot(aes(sample = resid(lm_1))) +
  stat_qq() +
  stat_qq_line(color="red")
```



```
lm_1 %>%
  ggplot(aes(x=fitted(lm_1), y=resid(lm_1))) +
  geom_point() +
  geom_smooth(method = "loess", color="red") +
  labs(x="predicted",
       y="residuals",
       title="Residual plot for predicted values")
```

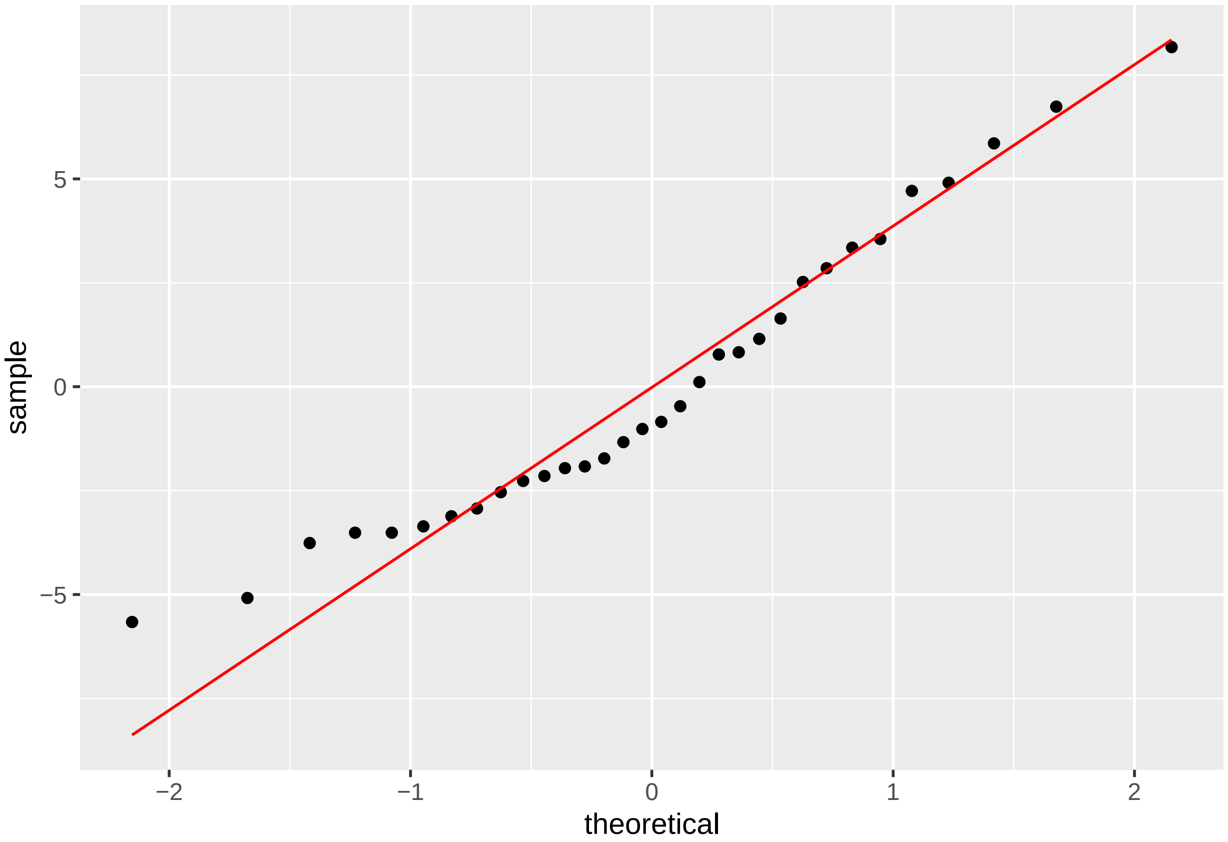


Here the adjusted R square is okay but the error values are suspecious. Many are very close in value to the coefficient, which is not a good sign of good model. The quadratic term confers a positive association with mpg, as does the "manual" transmission categorical. Displacement and the interaction term of `am` and `disp` both have a neagtive association. The p-value for the interaction term is the highest of the three predictors, so that would be elminated if we wanted to refine this model.

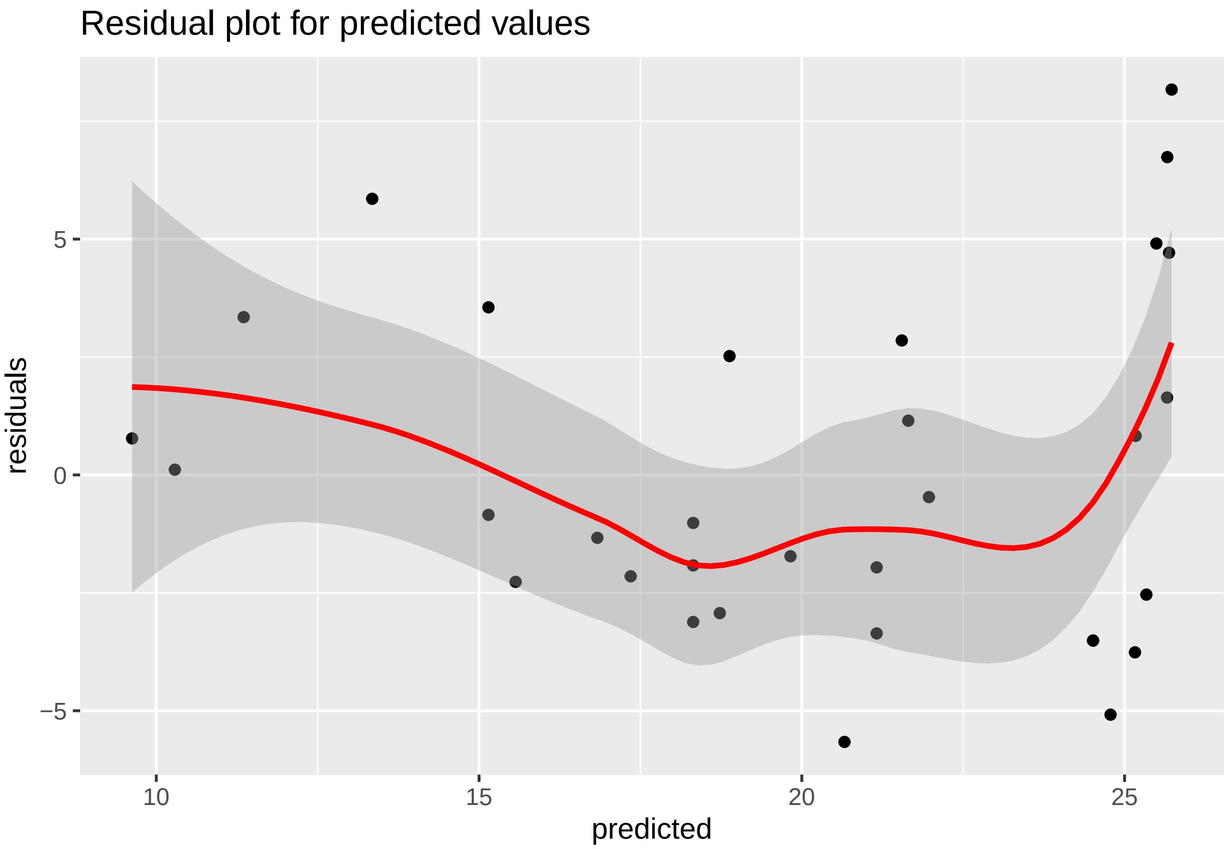
```
lm_2<- lm(data=mtcars,mpg ~ I(disp^2)+am)
summary(lm_2)
```

```
##
## Call:
## lm(formula = mpg ~ I(disp^2) + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6611 -2.6357 -0.9318  2.6023  8.1708
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.282e+01  1.430e+00  15.958 6.73e-16 ***
## I(disp^2)   -5.924e-05  1.205e-05  -4.918 3.19e-05 ***
## am           3.205e+00  1.559e+00   2.055   0.0489 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.682 on 29 degrees of freedom
## Multiple R-squared:  0.6509, Adjusted R-squared:  0.6269
## F-statistic: 27.04 on 2 and 29 DF,  p-value: 2.356e-07
```

```
lm_2 %>%
  ggplot(aes(sample = resid(lm_2))) +
  stat_qq() +
  stat_qq_line(color="red")
```



```
lm_2 %>%
  ggplot(aes(x=fitted(lm_2), y=resid(lm_2))) +
  geom_point() +
  geom_smooth(method = "loess", color="red") +
  labs(x="predicted",
       y="residuals",
       title="Residual plot for predicted values")
```



This appears to be a better model even if the Adjusted R square is not as high. The residuals look better and the p-values for the coefficients being non-zero are much better.