Lab 13

14.9.37

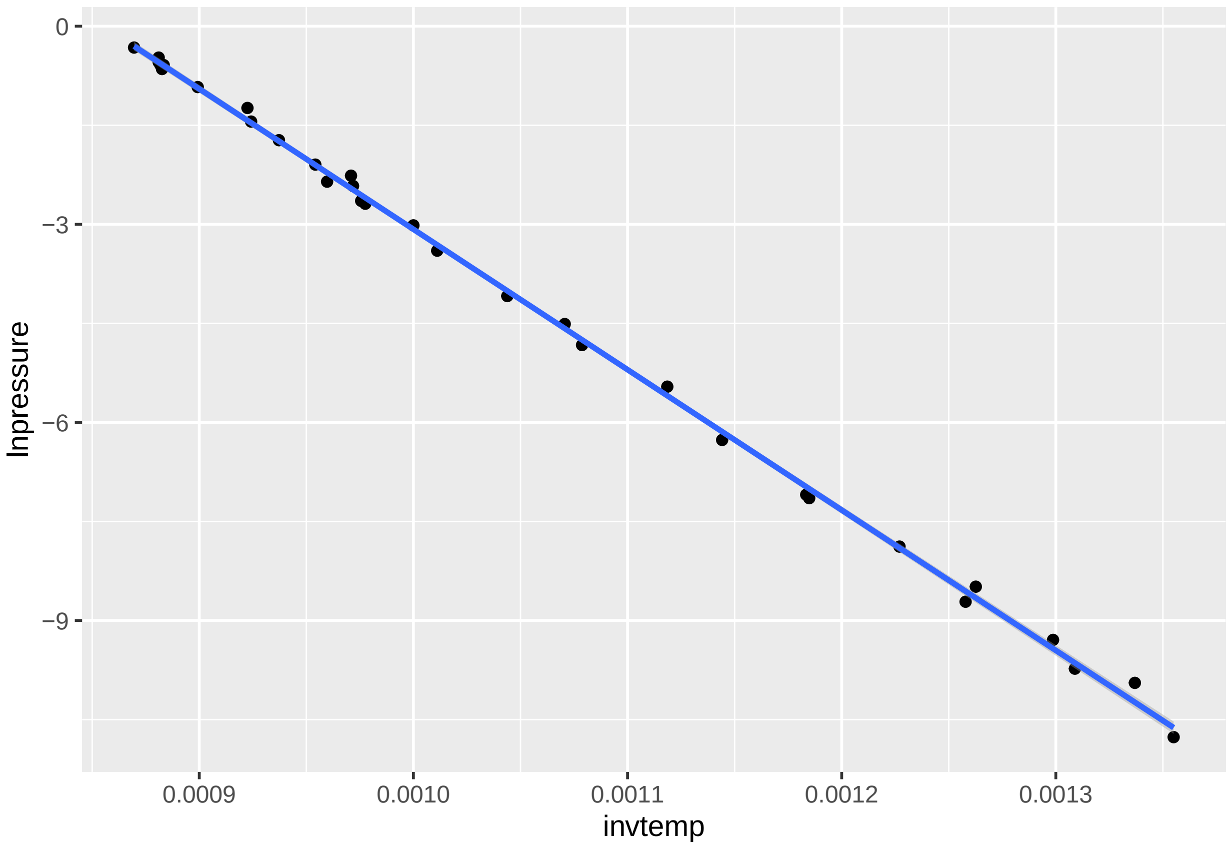
Part (a): lm

```
barium <- read.table("barium.txt", header = T)
names(barium) <- c("temp", "pressure")
```

Transforming the variables

```
barium <- barium %>% mutate(lnpressure = log(pressure)) %>% mutate(invtemp = temp^(-1))
ggplot(barium, aes(x = invtemp, y = lnpressure)) + geom_point() + geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Estimating A (intercept) and B (slope) by hand.

```
B = cov(barium$lnpressure, barium$invtemp)/var(barium$invtemp) #This is the slope from linear regressio
A = mean(barium$lnpressure) - mean(barium$invtemp)*B
paste("y =",B, "x +", A )
```

```
## [1] "y = -21259.6919282566 x + 18.1847029680274"
```

So the regression equation is

$$\hat{\text{lnpressre}} = A + B * \text{invtemp}$$

We can also use lm() in R

```
reg <- lm(lnpressure ~ invtemp, data = barium)
summary(reg)
```

```
##
## Call:
## lm(formula = lnpressure ~ invtemp, data = barium)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.158596 -0.087968  0.004495  0.061145  0.293036
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.818e+01  1.419e-01   128.2   <2e-16 ***
## invtemp     -2.126e+04  1.335e+02  -159.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1164 on 30 degrees of freedom
## Multiple R-squared:  0.9988, Adjusted R-squared:  0.9988
## F-statistic: 2.538e+04 on 1 and 30 DF,  p-value: < 2.2e-16
```

Now we create a 95% CI for A and B.

$$Var(\hat{A}) = \sigma^2(\frac{1}{n} + \frac{\bar{X}}{\sum(X_i - \bar{X})^2})$$

and

$$Var(\hat{B}) = \frac{\sigma^2}{\sum(X_i - \bar{X})^2}$$

First we need to estimate σ^2. We calculate RSS

```
n <- nrow(barium)
RSS <- sum(reg$residuals**2)
s2 <- RSS/(n-2)
s <- sqrt(s2)
varA <- s2*(1/n + mean(barium$invtemp)^2/((n-1)*var(barium$invtemp)))
varB <- s2/((n-1)*var(barium$invtemp))
```

Under CLT, for n > 20, both the coefficients will behave like a normal distribution. Since we are estimating σ^2, we will use a t-distribution instead.

```
t <- qt(0.975, df = n-2)
CI_A <- c(A - t*sqrt(varA), A + t*sqrt(varA))
CI_B <- c(B - t*sqrt(varB), B + t*sqrt(varB))
CI_A; CI_B
```

```
## [1] 17.89500 18.47441
```

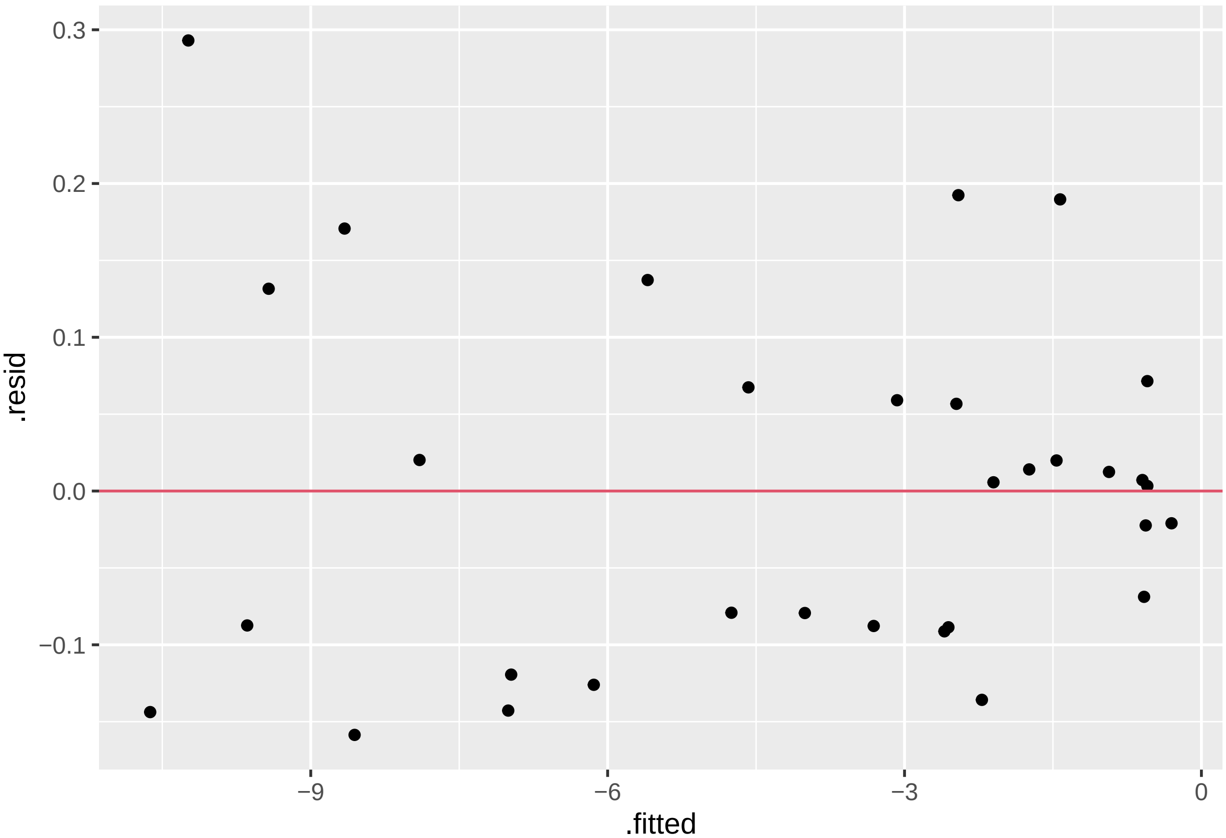
```
## [1] -21532.24 -20987.14
```

```
#A function built in R to help you check your work :)
confint(reg)
```

```
##                    2.5 %       97.5 %
## (Intercept)      17.895      18.47441
## invtemp     -21532.241 -20987.14257
```

Plotting the Residuals

```
ggplot(reg) + geom_point(aes(x=.fitted,y=.resid)) + geom_hline(aes(yintercept = 0), col = 2)
```



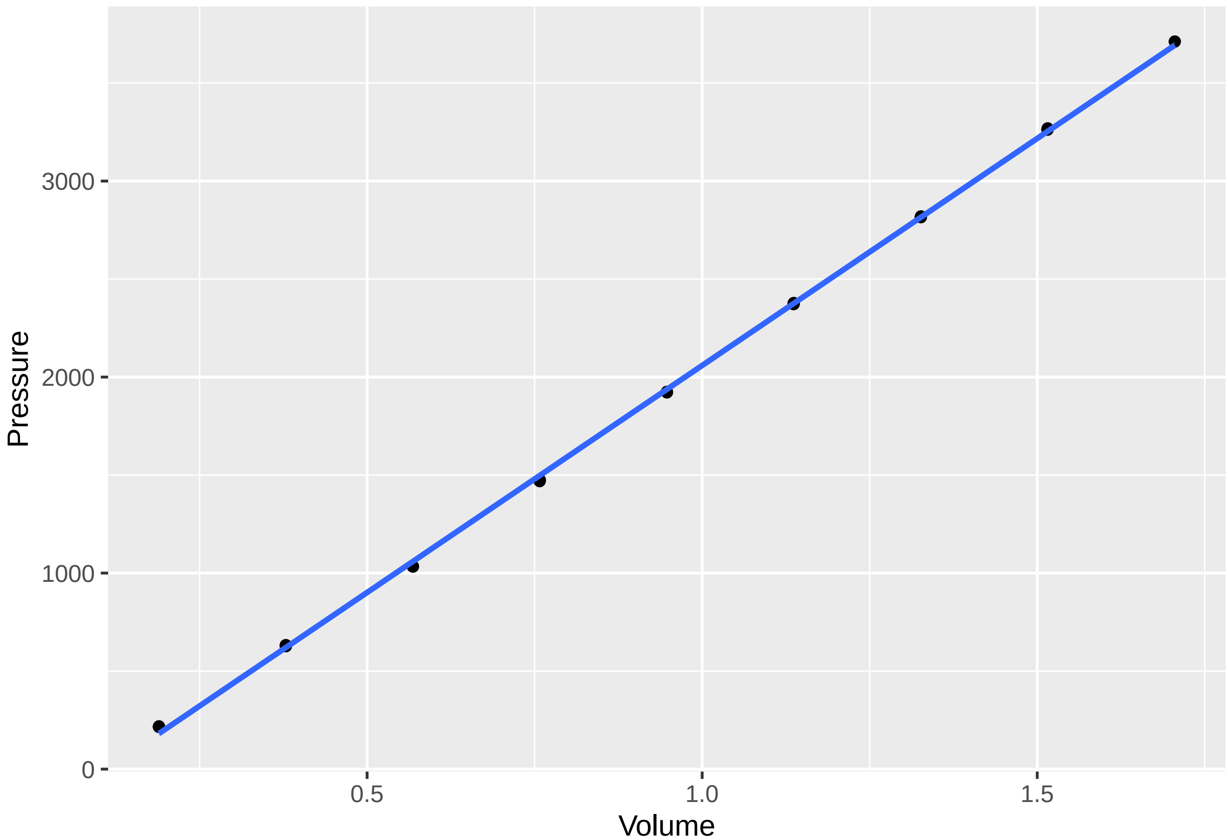
It seems that the positive residuals have a bigger variablity than the negative ones, but overall the The residual plots look pretty random. We should expect this from looking at the plot of the actual data.

14.9.39

```
tankvolume <- read.table("tankvolume.txt", head = T)
```

```
ggplot(tankvolume, aes(x = Volume, y = Pressure)) + geom_point() + geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

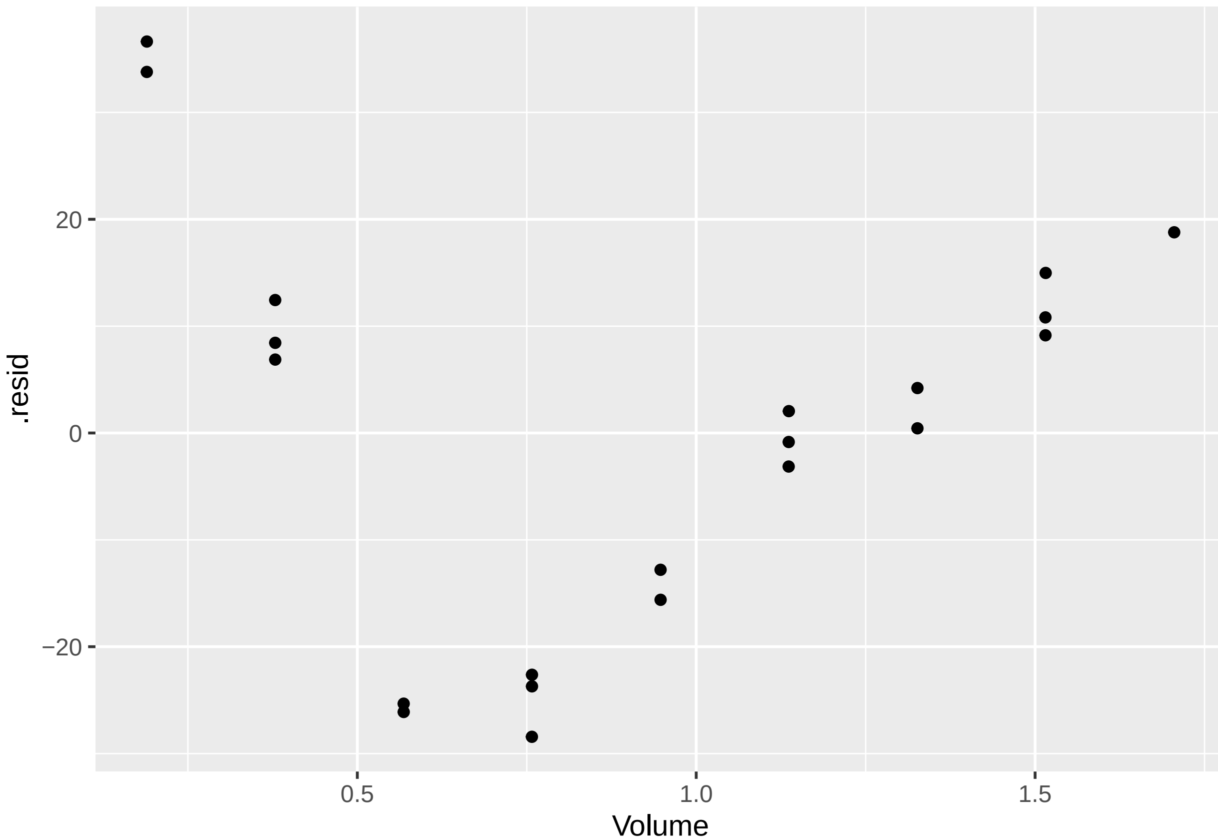


From eyeballing, the result looks almost perfectly linear.

```
fit <- lm(Pressure ~ Volume, data = tankvolume)
summary(fit)
```

```
##
## Call:
## lm(formula = Pressure ~ Volume, data = tankvolume)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.429 -15.610   2.047  10.819  36.634
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -257.301      9.430  -27.29   <2e-16 ***
## Volume      2316.469      9.243  250.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.44 on 19 degrees of freedom
## Multiple R-squared:  0.9997, Adjusted R-squared:  0.9997
## F-statistic: 6.28e+04 on 1 and 19 DF,  p-value: < 2.2e-16
```

```
ggplot(fit) + geom_point(aes(x = Volume, y = .resid))
```



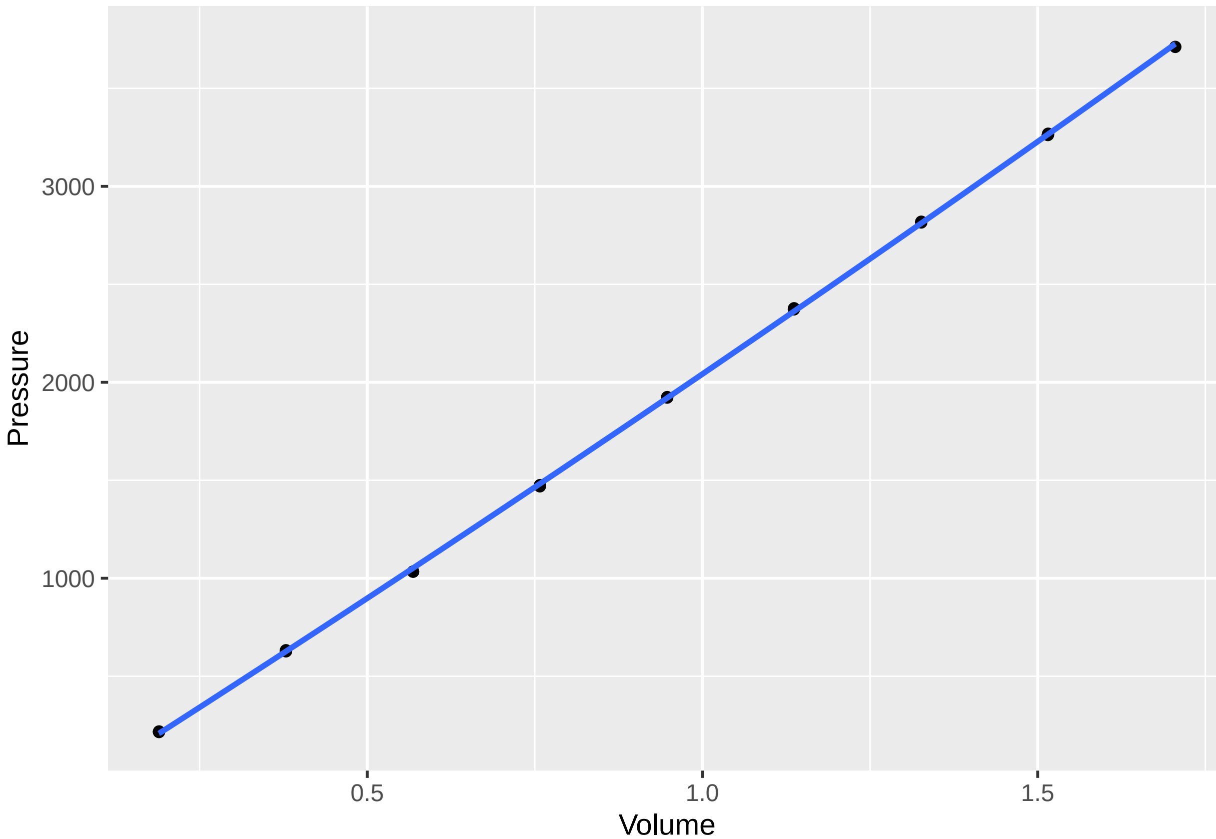
The residuals appear to have a really obvious quadratic pattern. This suggests that there might be better models than linear regression.

```r
tankvolume$Volume2 <- (tankvolume$Volume)**2
fit2 <- lm(Pressure ~ Volume + Volume2, data = tankvolume)
summary(fit2)
```

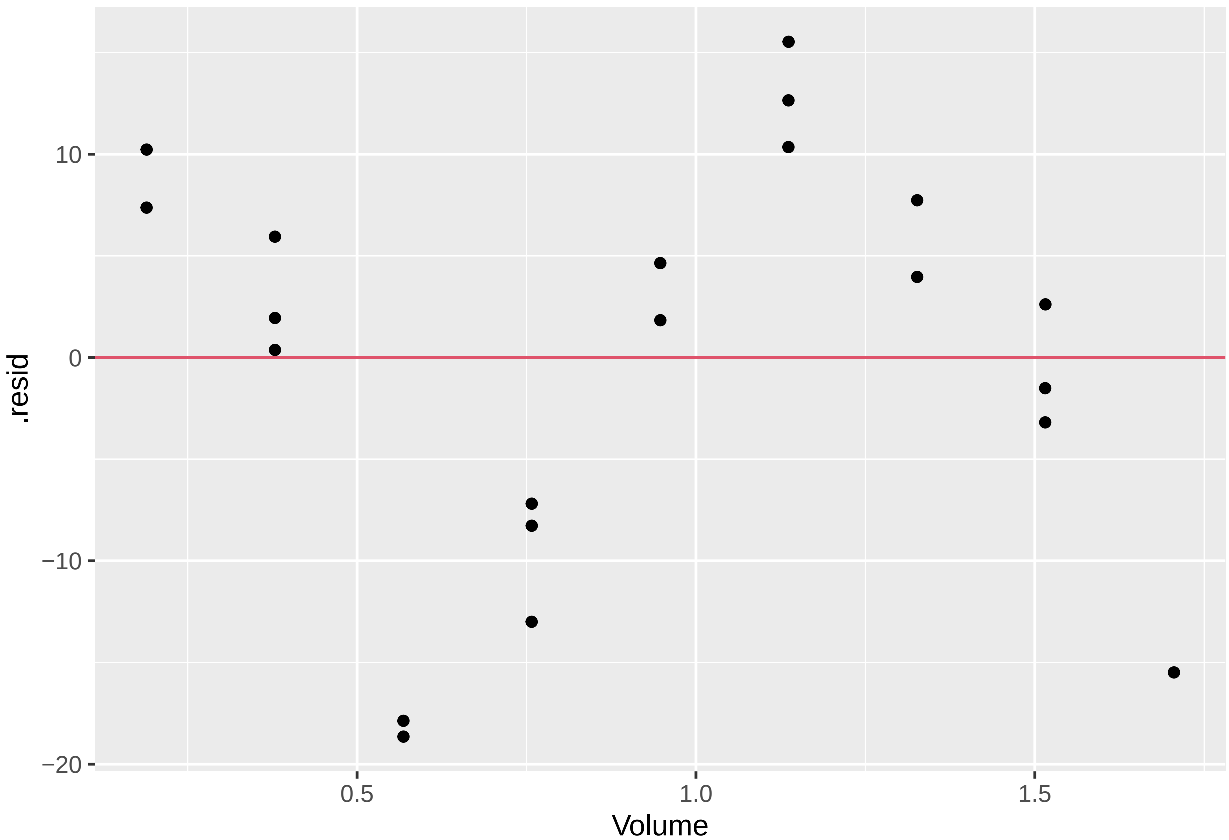
```
##
## Call:
## lm(formula = Pressure ~ Volume + Volume2, data = tankvolume)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.645  -7.189   1.944   7.371  15.528
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -204.995      9.274 -22.104 1.70e-14 ***
## Volume      2164.032     23.052  93.877  < 2e-16 ***
## Volume2       83.191     12.276   6.777 2.39e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.6 on 18 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999
## F-statistic: 1.057e+05 on 2 and 18 DF,  p-value: < 2.2e-16
```

We see that the adjusted R squared decreased (suggesting a better model). Let's look at some plots:

```r
ggplot(tankvolume, aes(x = Volume, y = Pressure)) + geom_point() + geom_smooth(method = "lm", formula =
```




```r
ggplot(fit2) + geom_point(aes(x = Volume, y = .resid)) + geom_hline(aes(yintercept = 0), col = 2)
```



Although the residuals still seem to follow some higher order curve, they look much more random than before.