Regression

Prabidhik KC

2022-11-22

```
## loading the necessary libraries
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

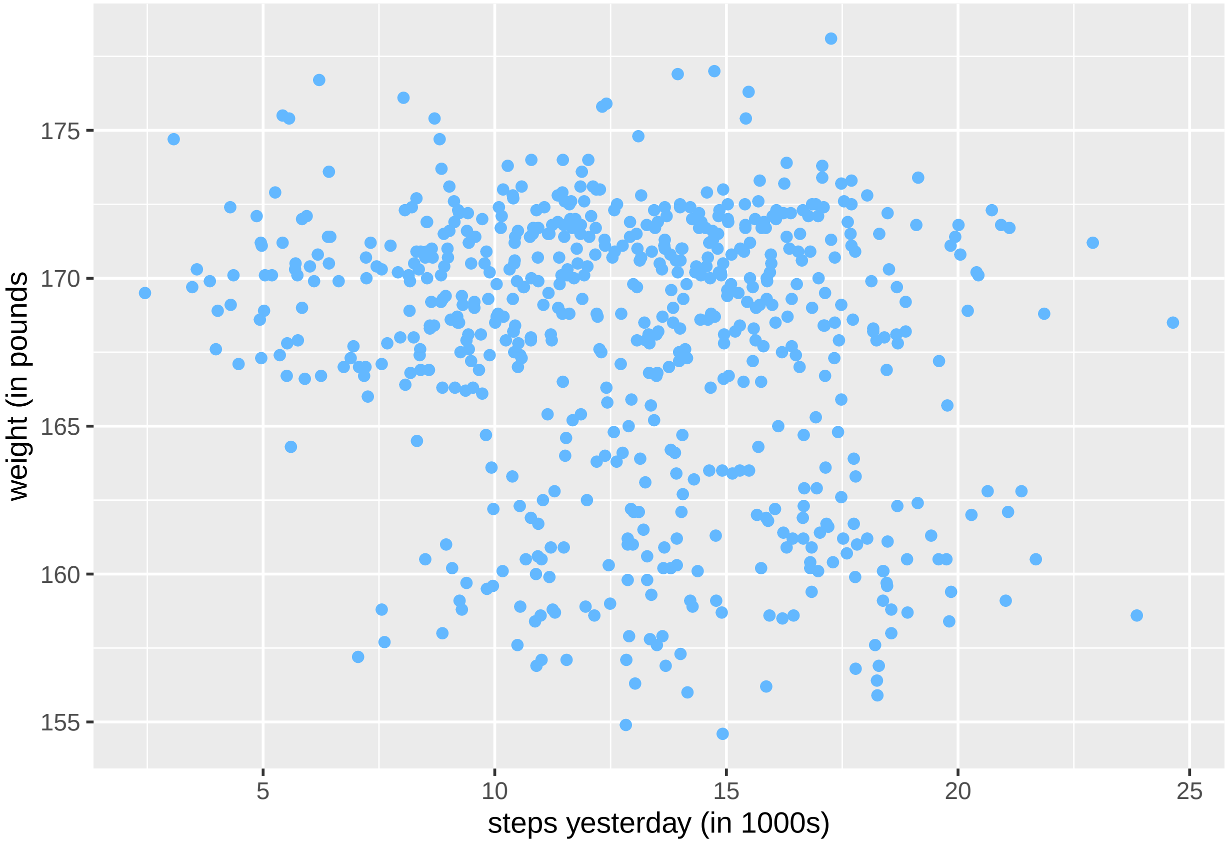
```
library(gov50data)
```

Looking at the data

```
health
```

```
## # A tibble: 644 x 6
##    date       active_calories steps weight steps_lag calorie_lag
##    <date>               <dbl>  <dbl>  <dbl>     <dbl>       <dbl>
##  1 2015-08-09             480  17.5    168      NA           NA
##  2 2015-08-10             996. 18.4    169.     17.5        480
##  3 2015-08-11            1127. 19.6    168      18.4        996.
##  4 2015-08-12             522. 10.4    167.     19.6       1127.
##  5 2015-08-13             844. 18.7    168.     10.4        522.
##  6 2015-08-14             396.  9.14   168.     18.7        844.
##  7 2015-08-15             423.  8.69   166.      9.14       396.
##  8 2015-08-16             958. 13.8    168.      8.69       423.
##  9 2015-08-17             597. 11.9    169      13.8        958.
## 10 2015-08-18            1378. 24.6    169.     11.9        597.
## # ... with 634 more rows
```

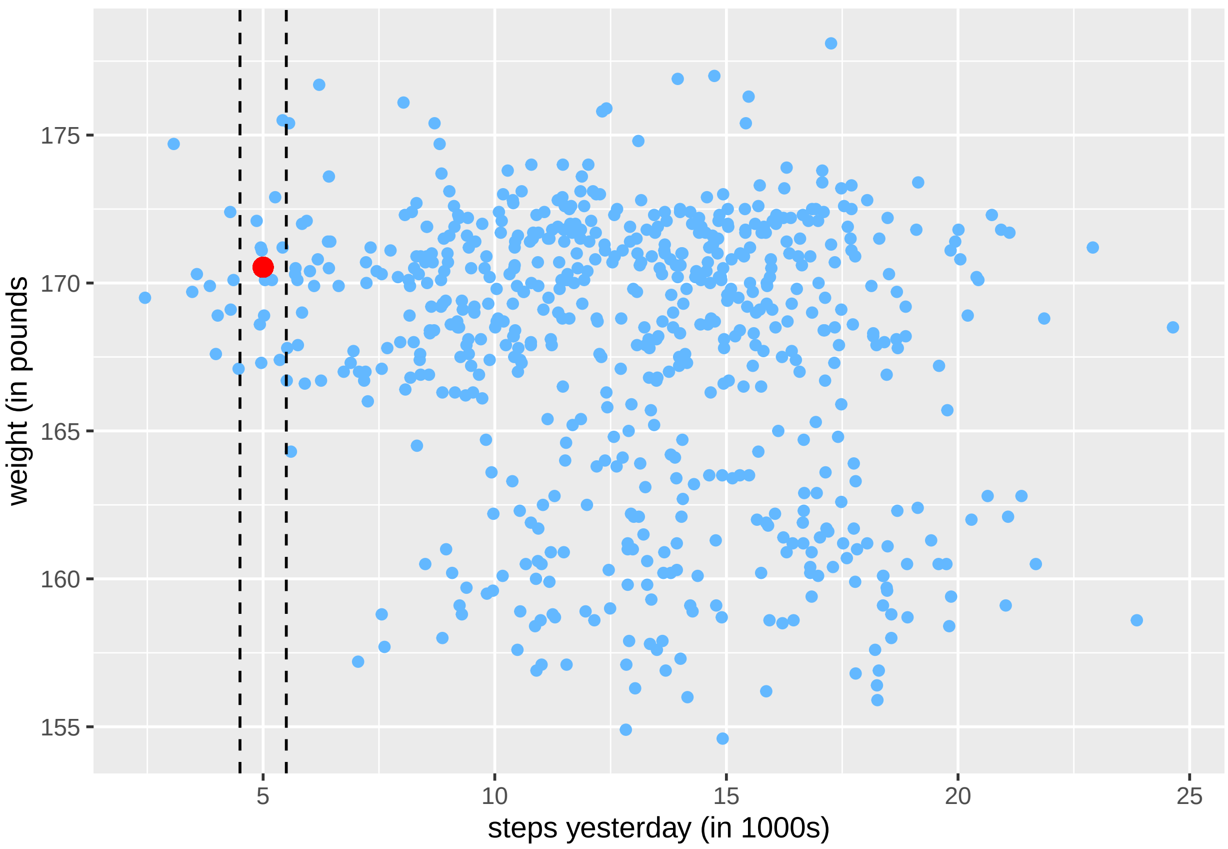
```
health <- health %>%
  drop_na()
health %>%
  ggplot(aes(x = steps_lag, y = weight)) +
  geom_point(color = "steelblue1") +
  labs(x = "steps yesterday (in 1000s)",
       y = "weight (in pounds")
```



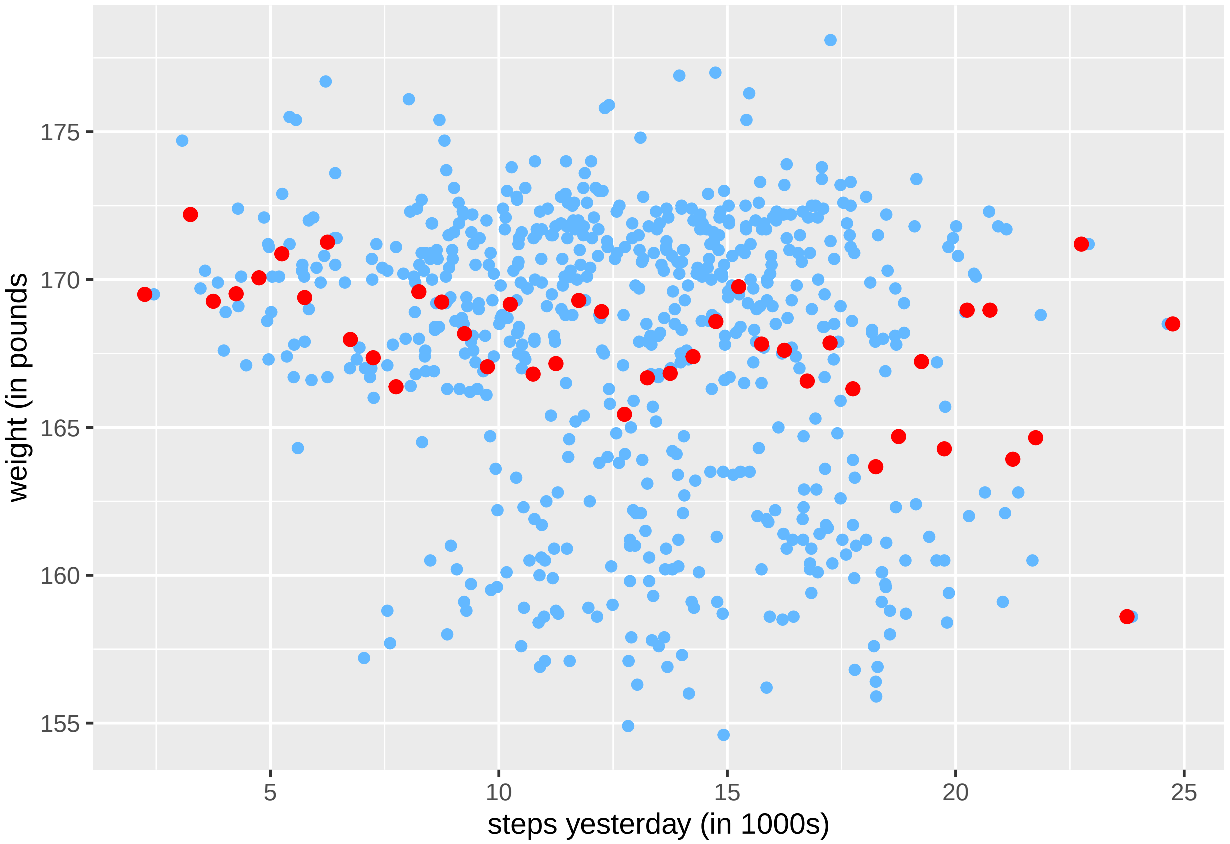
```
mean_wt_5ksteps <- health %>%
  filter(round(steps_lag) == 5) %>%
  summarize(mean(weight)) %>%
  pull()


mean_wt_5ksteps
```

```
## [1] 170.5333
```

```
health %>%
  ggplot(aes(x = steps_lag, y = weight)) +
  geom_point(color = "steelblue1") +
  labs(x = "steps yesterday (in 1000s)",
       y = "weight (in pounds") +
  geom_vline(xintercept = c(4.5, 5.5), linetype = "dashed") +
  geom_point(aes(x = 5, y = mean_wt_5ksteps), size = 3, color = "red")
```

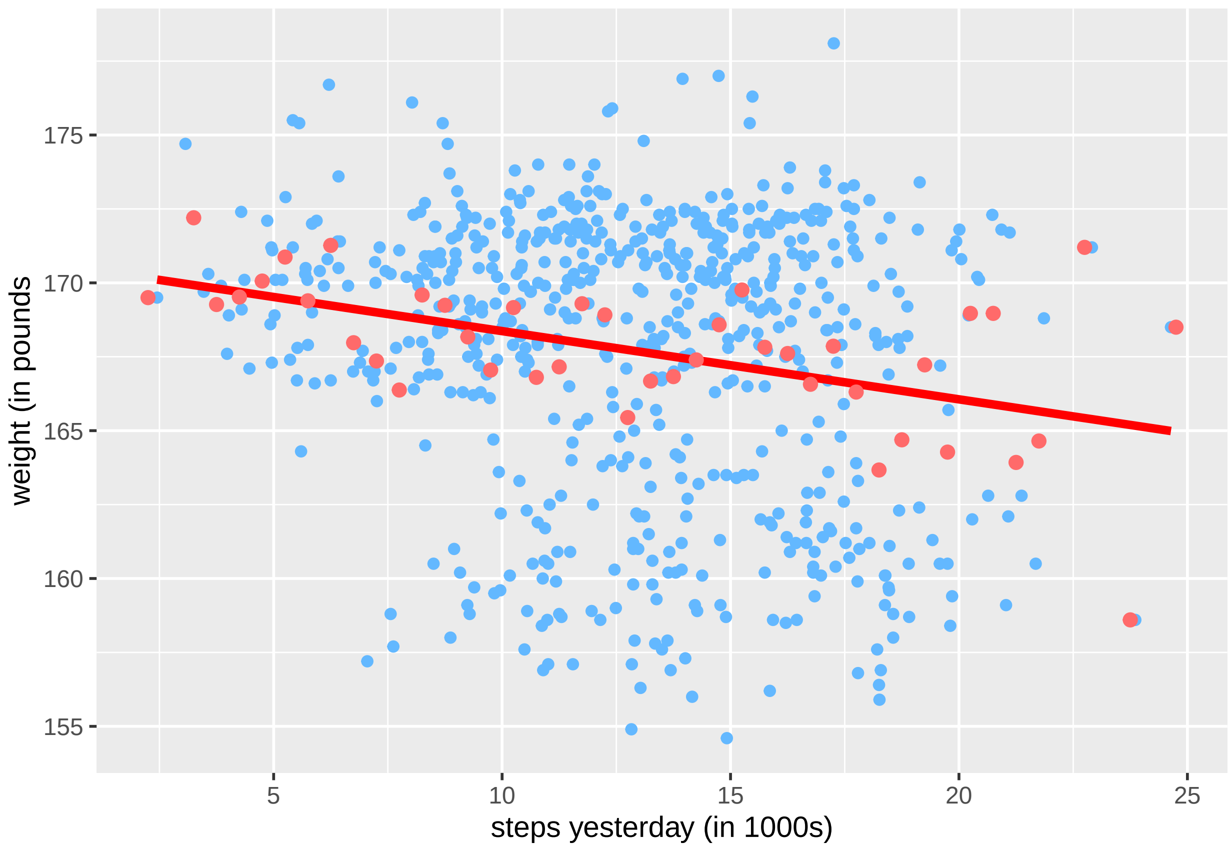


```
health %>%
  ggplot(aes(x = steps_lag, y = weight)) +
  geom_point(color = "steelblue1") +
  labs(x = "steps yesterday (in 1000s)",
       y = "weight (in pounds") +
  stat_summary_bin(fun = "mean", geom = "point", size = 2, color = "red", binwidth = 0.5)
```



```
health %>%
  ggplot(aes(x = steps_lag, y = weight)) +
  geom_point(color = "steelblue1") +
  labs(x = "steps yesterday (in 1000s)",
       y = "weight (in pounds") +
  geom_smooth(method = "lm", se = FALSE, color = "red", size = 1.5) +
  stat_summary_bin(fun = "mean", geom = "point", size = 2, color = "indianred1", binwidth = 0.5)
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```

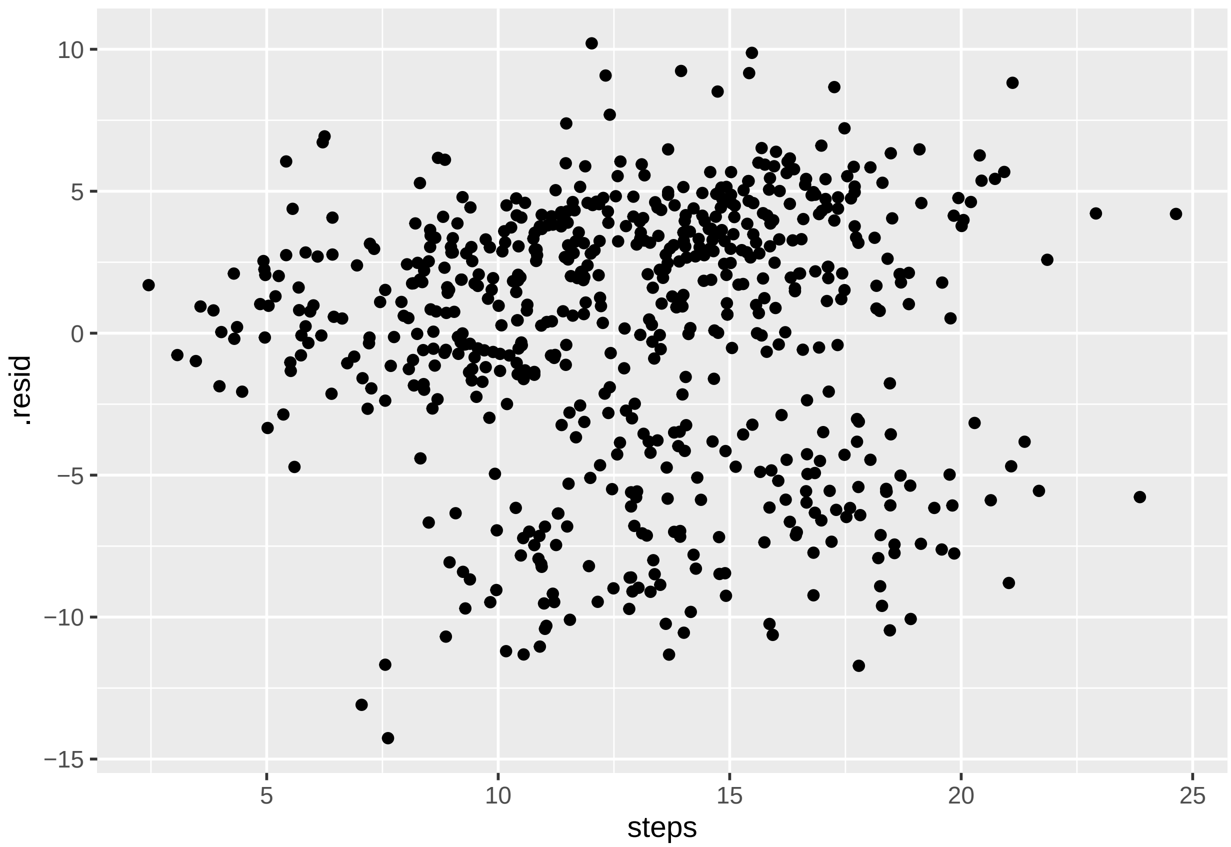


Lienar Models

```
fit <- lm(weight ~ steps, data = health)
fit
```

```
##
## Call:
## lm(formula = weight ~ steps, data = health)
##
## Coefficients:
## (Intercept)        steps
##    170.5493      -0.2212
```

```
library(broom)
augment(fit) %>%
  ggplot(aes(x = steps, y = .resid)) +
  geom_point()
```



```
coef(fit)
```

```
## (Intercept)       steps
## 170.5492866  -0.2211606
```

The coefficient on steps is -0.2211606

```
augment(fit) %>%
  summarize(mean(.resid))
```

```
## # A tibble: 1 x 1
##   `mean(.resid)`
##            <dbl>
## 1      -8.20e-14
```

```
augment(fit) %>%
  summarize(mean(.resid))
```

```
## # A tibble: 1 x 1
##   `mean(.resid)`
##            <dbl>
## 1      -8.20e-14
```

```
library(gov50data)
midterms
```

```
## # A tibble: 20 x 6
##     year president  party approval seat_change rdi_change
##    <dbl> <chr>      <chr>    <dbl>       <dbl>      <dbl>
##  1  1946 Truman     D           33         -55     NA    
##  2  1950 Truman     D           39         -29      8.2  
##  3  1954 Eisenhower R           61          -4      1    
##  4  1958 Eisenhower R           57         -47      1.1  
##  5  1962 Kennedy    D           61          -4      5    
##  6  1966 Johnson    D           44         -47      5.3  
##  7  1970 Nixon      R           58          -8      6.6  
##  8  1974 Ford       R           54         -43      6.4  
##  9  1978 Carter     D           49         -11      7.7  
## 10  1982 Reagan     R           42         -28      4.8  
## 11  1986 Reagan     R           63          -5      5.1  
## 12  1990 H.W. Bush  R           58          -8      5.6  
## 13  1994 Clinton    D           46         -53      3.9  
## 14  1998 Clinton    D           66           5      5.6  
## 15  2002 W. Bush    R           63           6      2.6  
## 16  2006 W. Bush    R           38         -30      5.7  
## 17  2010 Obama      D           45         -63      3.5  
## 18  2014 Obama      D           40         -13      4.6  
## 19  2018 Trump      R           38         -42      4.1  
## 20  2022 Biden      D           42          NA     -0.003
```

```
fit <- lm(seat_change ~ approval, data = midterms)
fit
```

```
##
## Call:
## lm(formula = seat_change ~ approval, data = midterms)
##
## Coefficients:
## (Intercept)     approval
##      -96.58         1.42
```

```
fit_rdi <- lm(seat_change ~ rdi_change, data = midterms)
fit_rdi
```

```
##
## Call:
## lm(formula = seat_change ~ rdi_change, data = midterms)
##
## Coefficients:
## (Intercept)   rdi_change
##     -29.413        1.215
```

```
summary(fit)$r.squared
```

```
## [1] 0.4498696
```

```
summary(fit_rdi)$r.squared
```

```
## [1] 0.01202348
```

```
glance(fit)
```

```
## # A tibble: 1 x 12
##   r.squ~1 adj.r~2 sigma stati~3 p.value    df logLik   AIC   BIC devia~4 df.re~5
##     <dbl>   <dbl> <dbl>   <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>   <int>
## 1   0.450   0.418  16.9    13.9 0.00167     1  -79.6  165.  168.   4852.      17
## # ... with 1 more variable: nobs <int>, and abbreviated variable names
## #   1: r.squared, 2: adj.r.squared, 3: statistic, 4: deviance, 5: df.residual
```

```
glance(fit)$r.squared
```

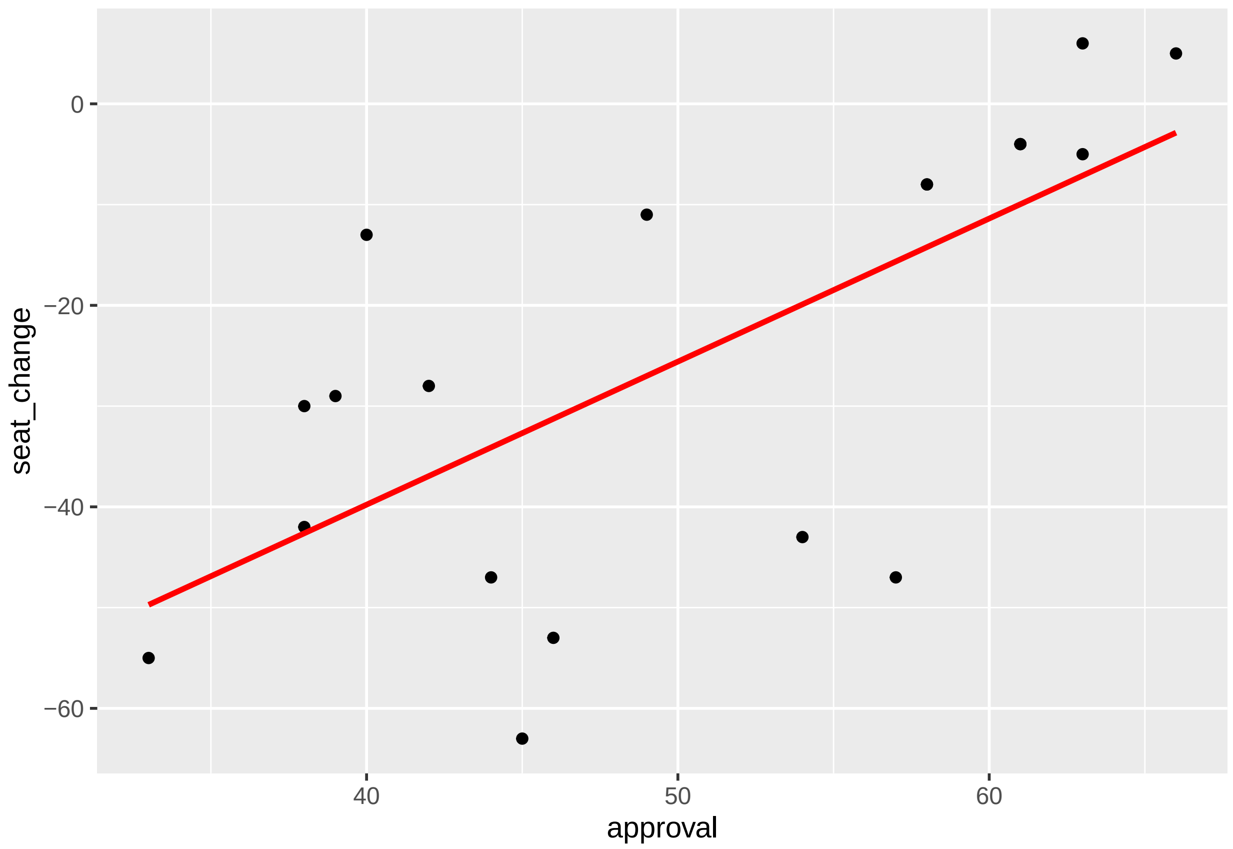
```
## [1] 0.4498696
```

```
midterms %>%
  ggplot(aes(x = approval, y = seat_change)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red")
```

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

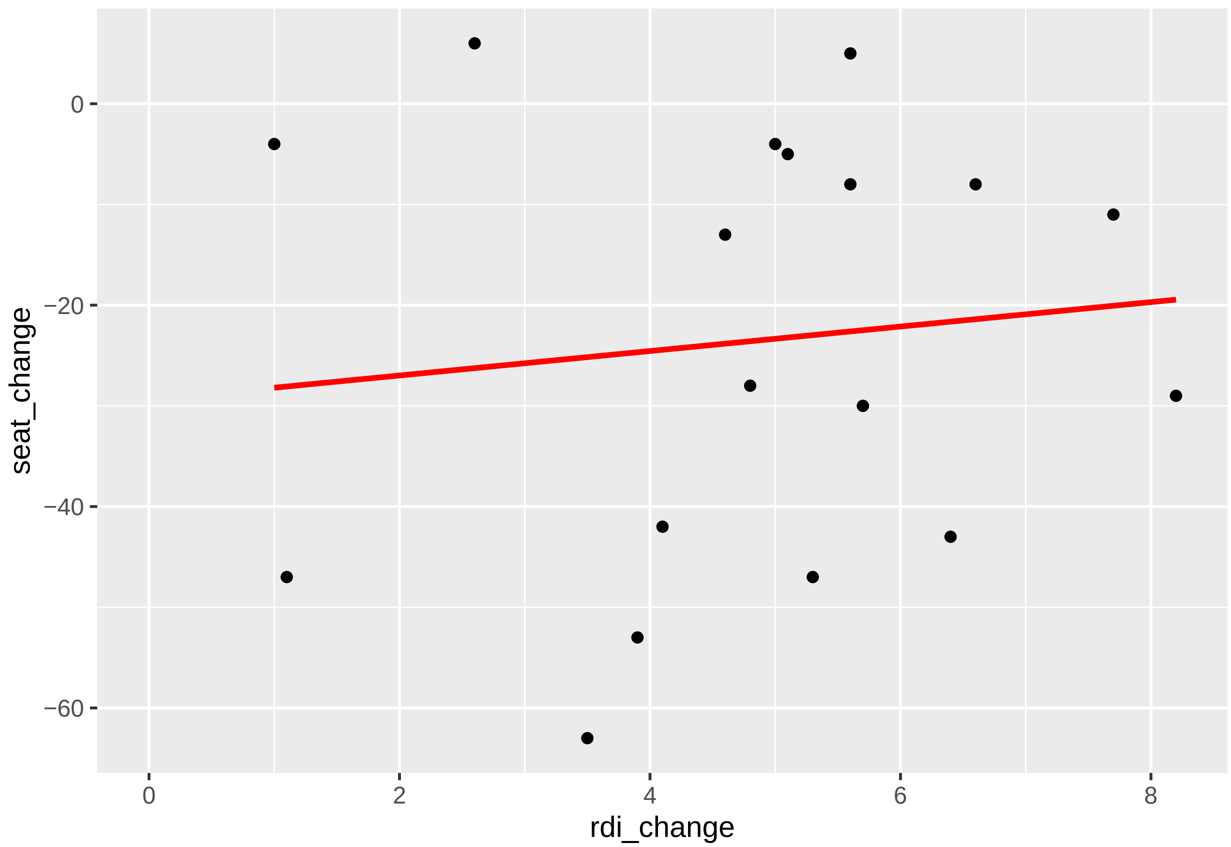
```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



```
midterms %>%
  ggplot(aes(x = rdi_change, y = seat_change)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’

## Warning: Removed 2 rows containing non-finite values (stat_smooth).

## Warning: Removed 2 rows containing missing values (geom_point).
```



Multiple regressions

```
mult.fit <- lm(seat_change ~ approval + rdi_change, data = midterms)
mult.fit
```

```
##
## Call:
## lm(formula = seat_change ~ approval + rdi_change, data = midterms)
##
## Coefficients:
## (Intercept)     approval   rdi_change
##    -117.226        1.526        3.217
```

```
midterms <- midterms %>%
  mutate(
    noise = runif(nrow(midterms))
  )
midterms
```

```
## # A tibble: 20 x 7
##     year president  party approval seat_change rdi_change noise
##    <dbl> <chr>      <chr>    <dbl>       <dbl>      <dbl> <dbl>
##  1  1946 Truman     D           33         -55      NA    0.273
##  2  1950 Truman     D           39          -29      8.2   0.738
##  3  1954 Eisenhower R           61           -4      1     0.846
##  4  1958 Eisenhower R           57          -47      1.1   0.118
##  5  1962 Kennedy    D           61           -4      5     0.294
##  6  1966 Johnson    D           44          -47      5.3   0.625
##  7  1970 Nixon      R           58           -8      6.6   0.221
##  8  1974 Ford       R           54          -43      6.4   0.556
##  9  1978 Carter     D           49          -11      7.7   0.473
## 10  1982 Reagan     R           42          -28      4.8   0.694
## 11  1986 Reagan     R           63           -5      5.1   0.397
## 12  1990 H.W. Bush  R           58           -8      5.6   0.280
## 13  1994 Clinton    D           46          -53      3.9   0.731
## 14  1998 Clinton    D           66            5      5.6   0.629
## 15  2002 W. Bush    R           63            6      2.6   0.956
## 16  2006 W. Bush    R           38          -30      5.7   0.678
## 17  2010 Obama      D           45          -63      3.5   0.909
## 18  2014 Obama      D           40          -13      4.6   0.951
## 19  2018 Trump      R           38          -42      4.1   0.236
## 20  2022 Biden      D           42           NA     -0.003 0.399
```

```
noise.fit <- lm(seat_change ~ approval + rdi_change + noise, data = midterms)

fit
```

```
##
## Call:
## lm(formula = seat_change ~ approval, data = midterms)
##
## Coefficients:
## (Intercept)     approval
##      -96.58         1.42
```

```
mult.fit
```

```
##
## Call:
## lm(formula = seat_change ~ approval + rdi_change, data = midterms)
##
## Coefficients:
## (Intercept)     approval   rdi_change
##    -117.226        1.526        3.217
```

```
noise.fit
```

```
##
## Call:
## lm(formula = seat_change ~ approval + rdi_change + noise, data = midterms)
##
## Coefficients:
## (Intercept)     approval   rdi_change        noise
##    -137.379        1.672        3.735       17.689
```

```
glance(fit)
```

```
## # A tibble: 1 x 12
##   r.squ~1 adj.r~2 sigma stati~3 p.value    df logLik   AIC   BIC devia~4 df.re~5
##     <dbl>   <dbl> <dbl>   <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>   <int>
## 1   0.450   0.418  16.9    13.9 0.00167     1  -79.6  165.  168.   4852.      17
## # ... with 1 more variable: nobs <int>, and abbreviated variable names
## #   1: r.squared, 2: adj.r.squared, 3: statistic, 4: deviance, 5: df.residual
```

```
glance(mult.fit)
```

```
## # A tibble: 1 x 12
##   r.squ~1 adj.r~2 sigma stati~3 p.value    df logLik   AIC   BIC devia~4 df.re~5
##     <dbl>   <dbl> <dbl>   <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>   <int>
## 1   0.468   0.397  16.7    6.59 0.00884     2  -74.6  157.  161.   4196.      15
## # ... with 1 more variable: nobs <int>, and abbreviated variable names
## #   1: r.squared, 2: adj.r.squared, 3: statistic, 4: deviance, 5: df.residual
```

```
glance(noise.fit)
```

```
## # A tibble: 1 x 12
##   r.squ~1 adj.r~2 sigma stati~3 p.value    df logLik   AIC   BIC devia~4 df.re~5
##     <dbl>   <dbl> <dbl>   <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>   <int>
## 1   0.511   0.406  16.6    4.87  0.0160     3  -73.8  158.  162.   3858.      14
## # ... with 1 more variable: nobs <int>, and abbreviated variable names
## #   1: r.squared, 2: adj.r.squared, 3: statistic, 4: deviance, 5: df.residual
```

Predicted values

```
library(modelr)
```

```
##
## Attaching package: 'modelr'
```

```
## The following object is masked from 'package:broom':
##
##     bootstrap
```

```
midterms %>%
  filter(year == 2022) %>%
  add_predictions(mult.fit)
```

```
## # A tibble: 1 x 8
##    year president party approval seat_change rdi_change noise  pred
##   <dbl> <chr>     <chr>    <dbl>       <dbl>      <dbl> <dbl> <dbl>
## 1  2022 Biden     D           42          NA     -0.003 0.399 -53.2
```

```r
midterms %>%
  filter(year == 2022) %>%
  gather_predictions(fit, mult.fit, noise.fit)
```

```
## # A tibble: 3 x 9
##   model      year president party approval seat_change rdi_change noise  pred
##   <chr>     <dbl> <chr>     <chr>    <dbl>       <dbl>      <dbl> <dbl> <dbl>
## 1 fit        2022 Biden     D           42          NA     -0.003 0.399 -36.9
## 2 mult.fit   2022 Biden     D           42          NA     -0.003 0.399 -53.2
## 3 noise.fit  2022 Biden     D           42          NA     -0.003 0.399 -60.1
```

```r
data("progresa", package = "qss")

cct <- as_tibble(progresa) %>%
  select(treatment, pri2000s, t2000)

cct
```

```
## # A tibble: 417 x 3
##    treatment pri2000s t2000
##        <int>    <dbl> <dbl>
##  1         1     40.8  55.8
##  2         1     22.4  31.2
##  3         1     38.9  47.0
##  4         1     31.2  45.0
##  5         0     76.9 100
##  6         0     23.9  37.4
##  7         1     47.3  64.9
##  8         1     21.4  58.1
##  9         1     56.5  71.3
## 10         1     36.6  51.2
## # ... with 407 more rows
```

```r
cct %>%
  group_by(treatment) %>%
  summarize(t2000 = mean(t2000)) %>%
  pivot_wider(names_from = treatment, values_from = t2000) %>%
  mutate(ATE = `1` - `0`)
```

```
## # A tibble: 1 x 3
##     ‘0‘   ‘1‘   ATE
##   <dbl> <dbl> <dbl>
## 1  63.8  68.1  4.27
```

```r
lm(pri2000s ~ treatment, data = cct)
```

```
##
## Call:
## lm(formula = pri2000s ~ treatment, data = cct)
##
## Coefficients:
## (Intercept)    treatment
##      34.489        3.622
```