# Question 6

```
data <- read.csv('BodyFatPercentage.csv')
data <- filter(data,BODYFAT!=0)
(lm <- summary(lm(BODYFAT~WRIST,data=data)))
```

```
##
## Call:
## lm(formula = BODYFAT ~ WRIST, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.8443  -5.6083   0.1517   5.0264  25.6330
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -31.563      9.003  -3.506  0.00054 ***
## WRIST          2.773      0.493   5.625 4.97e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.242 on 249 degrees of freedom
## Multiple R-squared:  0.1128, Adjusted R-squared:  0.1092
## F-statistic: 31.64 on 1 and 249 DF,  p-value: 4.97e-08
```

95% confidence intervals for $B_0$ is (-49.2945713,-13.8317969)

95% confidence intervals for $B_1$ is (1.8023379,3.7444175)

$R^2$ is 0.1127505, which means 11.3% of variance in bodyfact could be explained by the variance of wrist circumference.

# Question 7

```
new = data.frame(ABDOMEN=60)
(summary(lm(BODYFAT~ABDOMEN,data=data)))
```

```
##
## Call:
## lm(formula = BODYFAT ~ ABDOMEN, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.4044  -3.5186  -0.0367   3.1052  11.9594
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -34.76949    2.48522  -13.99   <2e-16 ***
## ABDOMEN       0.58051    0.02665   21.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.51 on 249 degrees of freedom
## Multiple R-squared:  0.6559, Adjusted R-squared:  0.6545
## F-statistic: 474.6 on 1 and 249 DF,  p-value: < 2.2e-16
```

Predicted body fat percentage of an individual with an abdominal circumference of 60 centimeters is 0.0613%

# Question 8

## (a) Read data and calculate weeks returns for spot and future.

```
future <- read.csv('WeeklyCrudeOilSpotFutures.csv')
future$FutureRt <-  c(NA,(future$CRUDEFRONT[2:nrow(future)]/future$CRUDEFRONT[1:(nrow
(future)-1)]-1)*100)
future$SpotRt <- c(NA,(future$CRUDESPOT[2:nrow(future)]/future$CRUDESPOT[1:(nrow(futu
re)-1)]-1)*100)
```

## (b)

Regress spot return on future return, the result is:

```
(lm <- summary(lm(SpotRt~FutureRt,data=future)))
```

```
##
## Call:
## lm(formula = SpotRt ~ FutureRt, data = future)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.106  -2.238   0.066   2.054  25.953
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.1627     0.2017   0.806    0.421
## FutureRt      0.3323     0.0372   8.931   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.845 on 362 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.1805, Adjusted R-squared:  0.1783
## F-statistic: 79.75 on 1 and 362 DF,  p-value: < 2.2e-16
```

So in order to hedge the 50,000 barrels of crude oil position, I need to short sell 16.6128671 future contracts

## (c)

Regress spot price on future price:

```
(lm <- summary(lm(CRUDESPOT~CRUDEFRONT,data=future)))
```

```
##
## Call:
## lm(formula = CRUDESPOT ~ CRUDEFRONT, data = future)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.922  -4.639  -1.980   3.540  29.175
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.23789    1.70910   -3.65 0.000301 ***
## CRUDEFRONT   1.11513    0.02073   53.80  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.96 on 363 degrees of freedom
## Multiple R-squared:  0.8886, Adjusted R-squared:  0.8883
## F-statistic:  2894 on 1 and 363 DF,  p-value: < 2.2e-16
```

For hypothesis $H_0 : \beta = 1$, the p-value is 5.391843810^{-8}, which is less than 0.001, so we can rejected the null hypothesis at 0.001 significance level, so for sure we can also reject the null hypothesis at 0.01 and 0.05 significance levels as well.

# Question 9

## (a)Read data and calculate returns

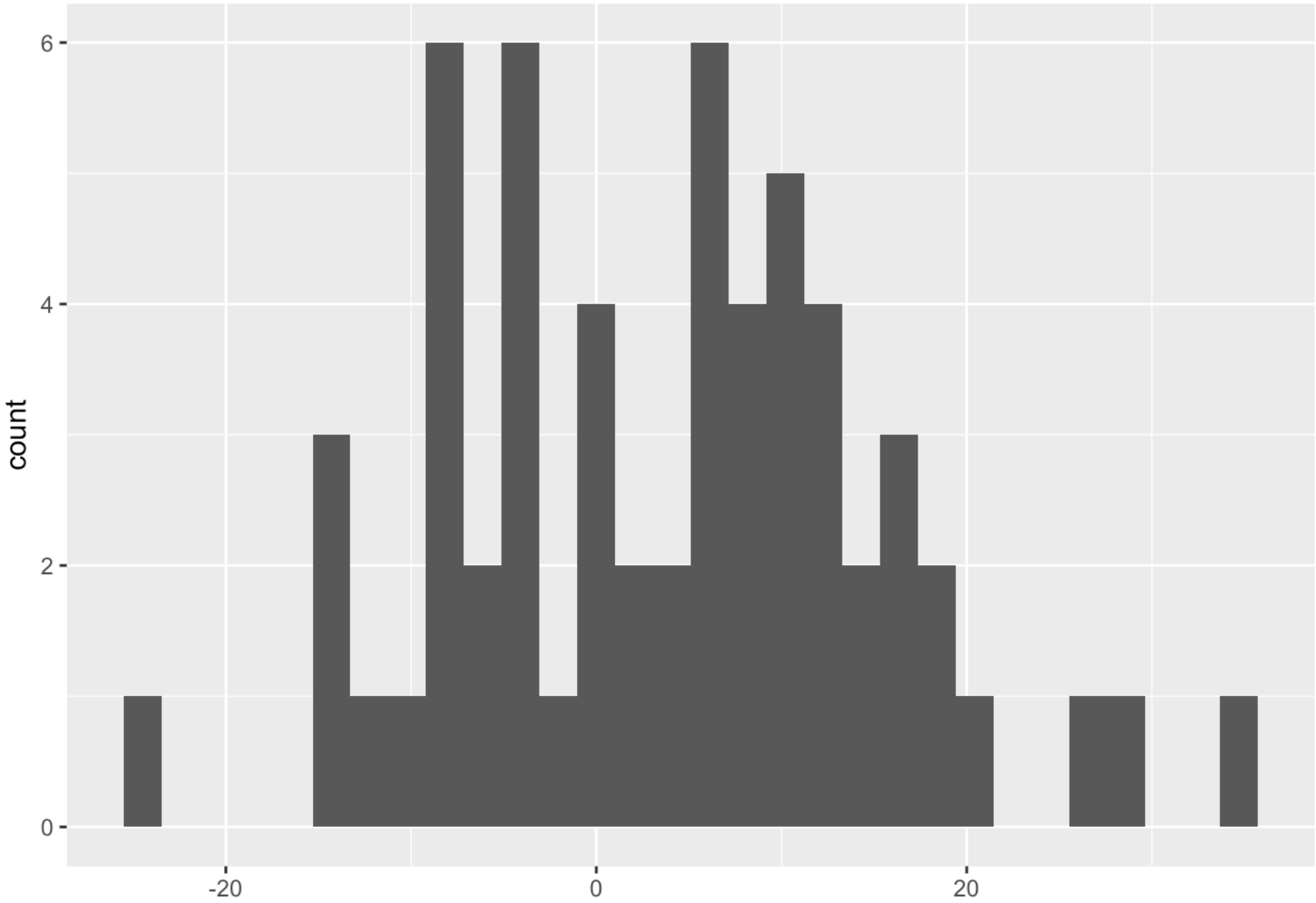
```
CAPM1 <- read.csv('AAPLSP50020022006.csv')
CAPM2 <- read.csv("AAPLSP50020072011.csv")

CAPM1$AAPLrt <- c(NA,100*(CAPM1$AAPL[2:nrow(CAPM1)]/CAPM1$AAPL[1:(nrow(CAPM1)-1)]-1))
CAPM1$SPrt <- c(NA,100*(CAPM1$SP500[2:nrow(CAPM1)]/CAPM1$SP500[1:(nrow(CAPM1)-1)]-1))

CAPM2$AAPLrt <- c(NA,100*(CAPM2$AAPL[2:nrow(CAPM2)]/CAPM2$AAPL[1:(nrow(CAPM2)-1)]-1))
CAPM2$SPrt <- c(NA,100*(CAPM2$SP500[2:nrow(CAPM2)]/CAPM2$SP500[1:(nrow(CAPM2)-1)]-1))
```

Plot the histogram of the Apple returns.

```
ggplot(CAPM1)+
    geom_histogram(aes(x=AAPLrt))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

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

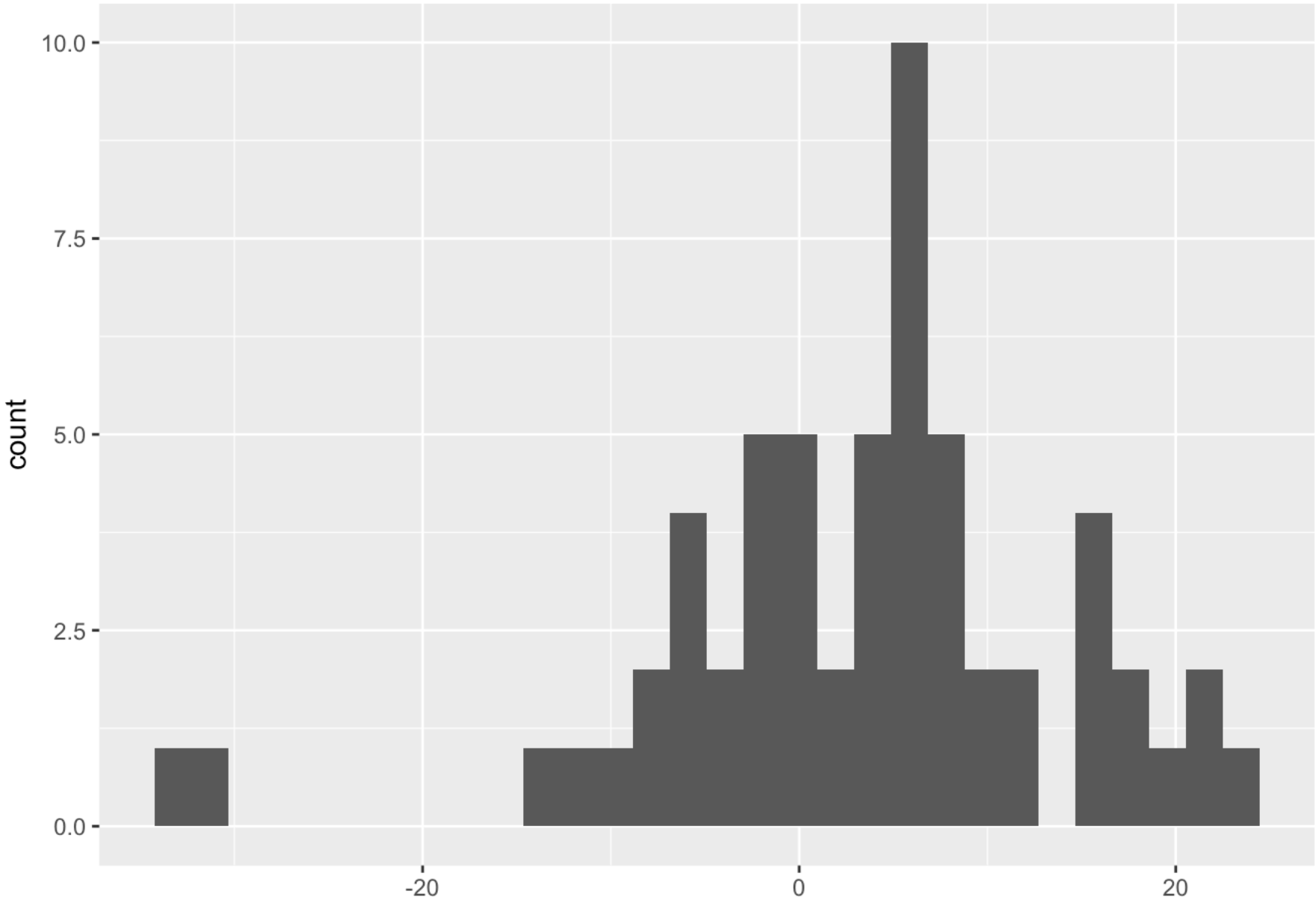



AAPLrt

```
ggplot(CAPM2)+
    geom_histogram(aes(x=AAPLrt))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

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




## (b)

Transforming annulized risk free rate to monthly risk free rate and subtract it from monthly returns of SP500 and Apple

For data set 1

```
CAPM1$month_rf <- ((1+CAPM1$rf/100)^(1/12)-1)*100
CAPM1$AAPL_ExcessRt <- CAPM1$AAPLrt-CAPM1$month_rf
CAPM1$SP500_ExcessRt <- CAPM1$SPrt-CAPM1$month_rf
summary(lm(AAPL_ExcessRt~SP500_ExcessRt,data=CAPM1))
```

```
##
## Call:
## lm(formula = AAPL_ExcessRt ~ SP500_ExcessRt, data = CAPM1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.6940  -8.0804   0.9244   7.4091  29.8925
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.3819     1.3515   2.502 0.015232 *
## SP500_ExcessRt   1.4616     0.3792   3.855 0.000297 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.36 on 57 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.2068, Adjusted R-squared:  0.1929
## F-statistic: 14.86 on 1 and 57 DF,  p-value: 0.0002968
```

For data set 2

```
CAPM2$month_rf <- ((1+CAPM2$rf/100)^(1/12)-1)*100
CAPM2$AAPL_ExcessRt <- CAPM2$AAPLrt-CAPM2$month_rf
CAPM2$SP500_ExcessRt <- CAPM2$SPrt-CAPM2$month_rf
summary(lm(AAPL_ExcessRt~SP500_ExcessRt,data=CAPM2))
```

```
##
## Call:
## lm(formula = AAPL_ExcessRt ~ SP500_ExcessRt, data = CAPM2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.5678  -4.4109   0.2214   4.3756  18.6555
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.3828     1.1028   3.067   0.0033 **
## SP500_ExcessRt   1.2172     0.2019   6.030 1.29e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.467 on 57 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3894, Adjusted R-squared:  0.3787
## F-statistic: 36.36 on 1 and 57 DF,  p-value: 1.286e-07
```

The slopes are not similar, $\beta$ in the first period is 1.4616 at 0.001 significance level, which means Apples excess return would increase 1.4616%, on average, when market excess return is inceasing 1%. $\beta$ is 1.2172 in the second time period at 0.001 significance level, which is lower than first period, which means Apples excess return would increase 1.2172%, on average, when market excess return is inceasing 1%. The reason why $\beta$ decrease in the second horizon could be financial crisis during 2007-2008 that leads to sharply decrease for the whole market but Apples stock price performed relatively strong in that period but did not decrease so much as market did. Another reason could be as Apple's market value grows, it's stock price would likely be less volatile as before.

# Question 10

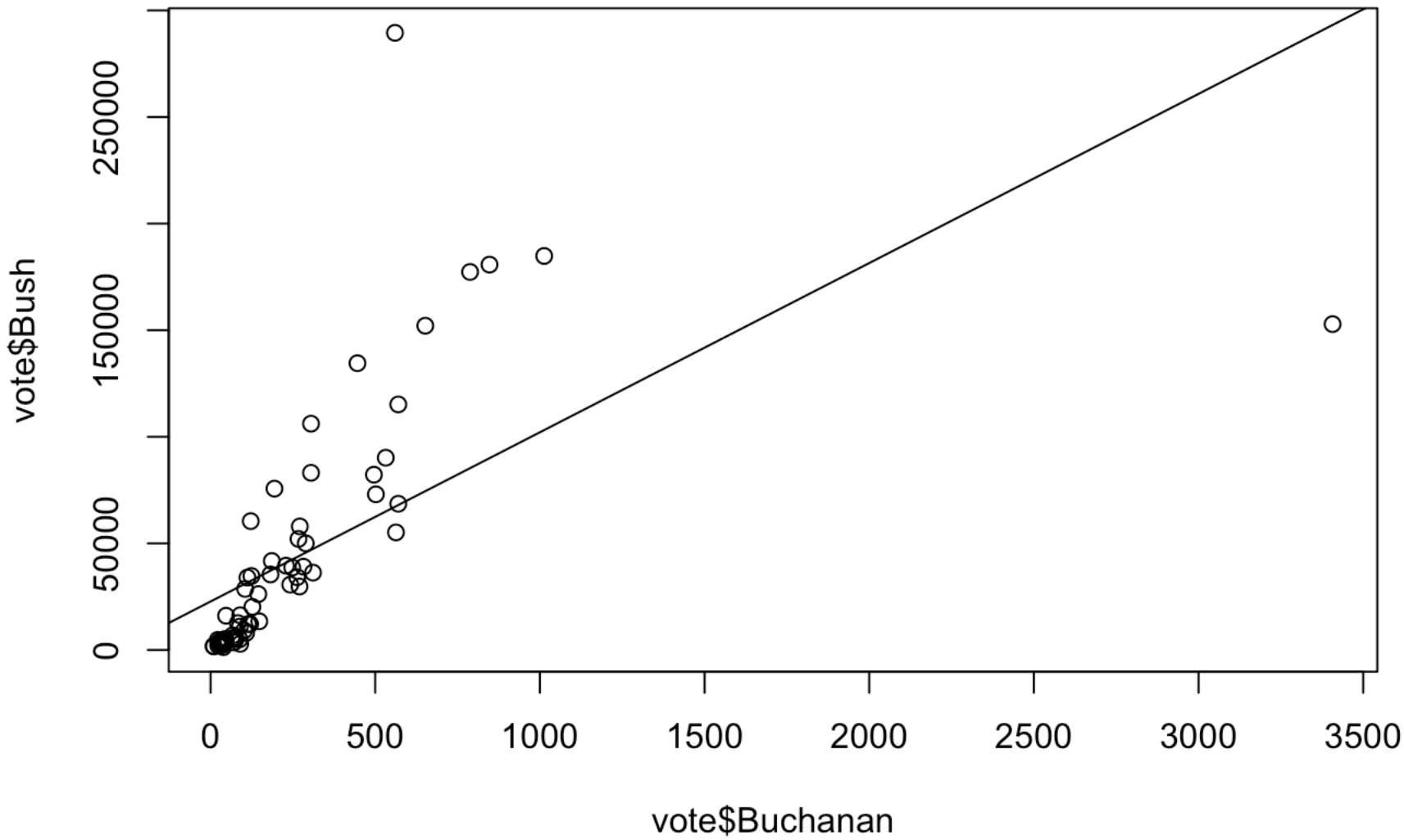
Read the data and check the regression result

```
vote <- read.csv("VoteTotalsFloridaCountiesElection2000.csv")
summary(lm(vote$Bush~vote$Buchanan))
```

```
##
## Call:
## lm(formula = vote$Bush ~ vote$Buchanan)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -140389  -21801  -13604    4279  222293
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   22738.15    6359.78   3.575 0.000666 ***
## vote$Buchanan    79.39      12.30   6.455 1.57e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44960 on 65 degrees of freedom
## Multiple R-squared:  0.3906, Adjusted R-squared:  0.3813
## F-statistic: 41.67 on 1 and 65 DF,  p-value: 1.574e-08
```

Draw the graph to find the outliers

```
plot(vote$Bush~vote$Buchanan)
abline(lm(vote$Bush~vote$Buchanan))
```




From the picture above, we can observe there are two obvious outliers, one has the maximum votes for Buchanan while another has maximum votes for Bush, we then find out those two counties:

```
filter(vote,Bush==max(Bush)|Buchanan==max(Buchanan)) %>%
    select(County)
```

```
##      County
## 1      DADE
## 2 PALMBEACH
```

So the outliers are DADE and PALMBEACH. We then filter out those two counties to estimate the model.

```
new_vote <- filter(vote,Bush!=max(Bush)&Buchanan!=max(Buchanan))
summary(lm(new_vote$Bush~new_vote$Buchanan))
```

```
##
## Call:
## lm(formula = new_vote$Bush ~ new_vote$Buchanan)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -52358  -8151   -958   2750  49913
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -2670.561   2954.897  -0.904     0.37
## new_vote$Buchanan   195.683      9.732  20.106   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17380 on 63 degrees of freedom
## Multiple R-squared:  0.8652, Adjusted R-squared:  0.863
## F-statistic: 404.3 on 1 and 63 DF,  p-value: < 2.2e-16
```

Before deleting outliers, the slope is 79.39 while after deleting outliers, the slope is 195.683, which more than doubled, I can say in this model, $\hat{\beta}_1$ is sensitive to outliers. But in general, as the data set grows, the sensitivity to outliers should decrease.

# Question 11

Read the data

```
Kelley <- read.csv("KelleyBlueBookData.csv")
```

Test $H_0 : \rho = 0$:

```
cor <- cor.test(Kelley$Price,Kelley$Mileage,alternative = 'two.sided')
cor
```

```
##
##  Pearson's product-moment correlation
##
## data:  Kelley$Price and Kelley$Mileage
## t = -4.0932, df = 802, p-value = 4.685e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.21011376 -0.07464739
## sample estimates:
##        cor
## -0.1430505
```

So at 5% significance we can reject the null hopothesis that $H_0 : \rho = 0$.

Test $H_0 : \beta_1 = 0$:

```
summary(lm(Price~Mileage,data=Kelley))
```

```
##
## Call:
## lm(formula = Price ~ Mileage, data = Kelley)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13905  -7254  -3520   5188  46091
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.476e+04  9.044e+02  27.383  < 2e-16 ***
## Mileage     -1.725e-01  4.215e-02  -4.093 4.68e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9789 on 802 degrees of freedom
## Multiple R-squared:  0.02046,    Adjusted R-squared:  0.01924
## F-statistic: 16.75 on 1 and 802 DF,  p-value: 4.685e-05
```

Since the p-value for hypothesis test is 4.68e-05, We can reject the null hypothesis that $H_0 : \beta_1 = 0$ at 0.001 significance level.

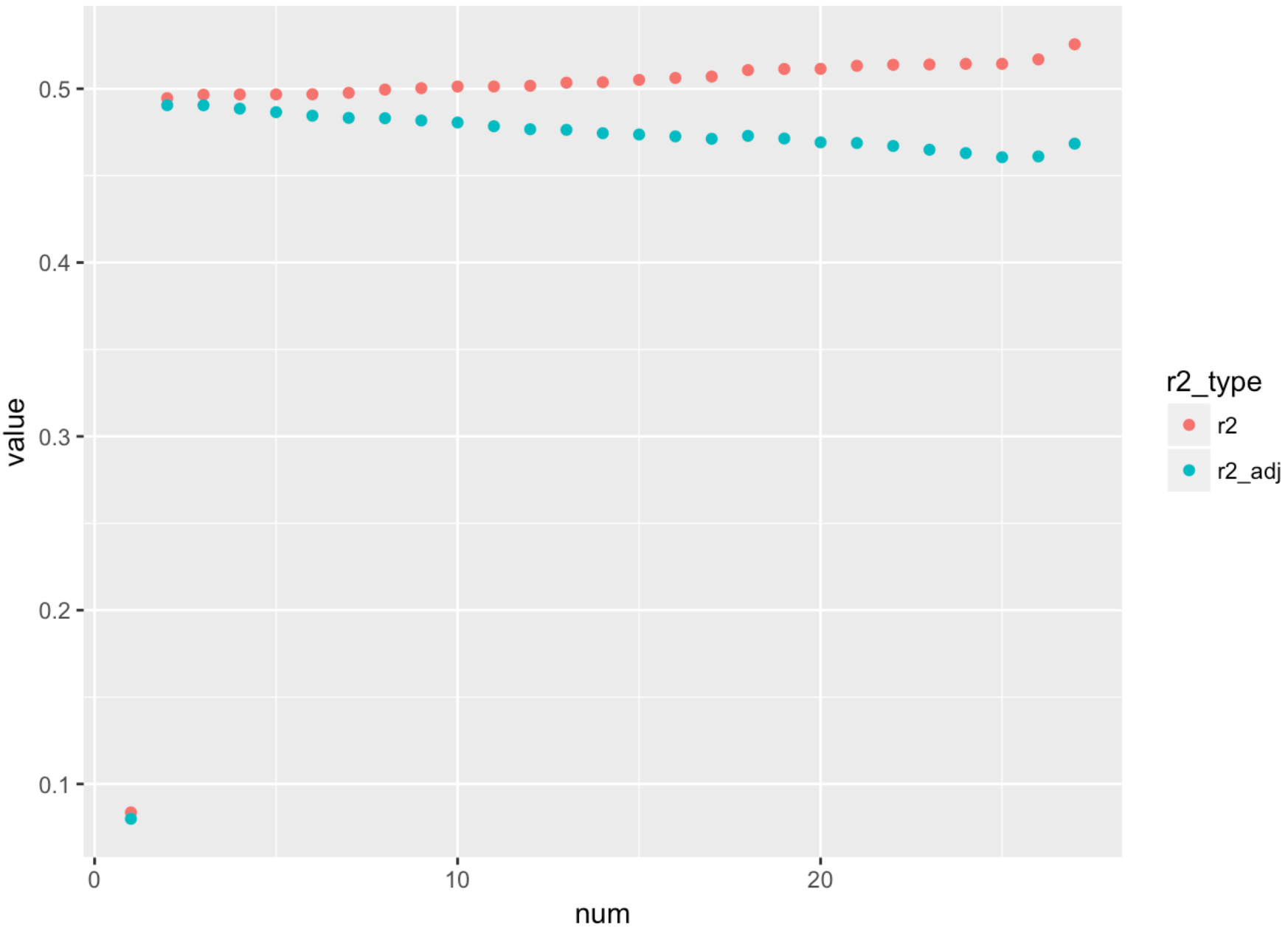
To test null hypothesis $H_0$ : The regression relationship is not significant we can observe the result of F-statistics above, to p-value of F-statistics is 4.685e-05, so we can reject the null hypothesis at significance level at 0.001.

The R-squared is 0.02046, which means 2.046% variance of Price could be explained by the variance of Mileagge.

# Question 12

Read the data, and get the corresponding $R^2$ and Adjusted $R^2$.

```
r <- read.csv('ExtraColumnsOfRandomData.csv')
r2 <- NULL
r2_adj <- NULL
num <- 1:27
for (i in num){
    r2[i] <- summary(lm(BODYFAT~.,data=r[,1:(i+1)]))$r.squared
    r2_adj[i] <- summary(lm(BODYFAT~.,data=r[,1:(i+1)]))$adj.r.squared
}
graph <- data.frame(num,r2,r2_adj) %>%
    gather(r2_type,value,r2,r2_adj)
ggplot(graph)+
    geom_point(aes(x=num,y=value,color=r2_type))
```




From this graph we can see that, after taking the variable Hip as independent variable, both $R^2$ and Adjusted $R^2$ get increased a lot, with more random columns added into the model, $R^2$ keeps increasing while $R^2$ has a decreasing trend.