Homework 2

Hoang Chu

2023-09-14

2.

```
data("uswages")
?uswages
summary(uswages)
```

```
      wage              educ           exper            race      
 Min.   :  50.39   Min.   : 0.00   Min.   :-2.00   Min.   :0.000  
 1st Qu.: 308.64   1st Qu.:12.00   1st Qu.: 8.00   1st Qu.:0.000  
 Median : 522.32   Median :12.00   Median :15.00   Median :0.000  
 Mean   : 608.12   Mean   :13.11   Mean   :18.41   Mean   :0.078  
 3rd Qu.: 783.48   3rd Qu.:16.00   3rd Qu.:27.00   3rd Qu.:0.000  
 Max.   :7716.05   Max.   :18.00   Max.   :59.00   Max.   :1.000  
      smsa             ne              mw               so        
 Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :1.000   Median :0.000   Median :0.0000   Median :0.0000  
 Mean   :0.756   Mean   :0.229   Mean   :0.2485   Mean   :0.3125  
 3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:1.0000  
 Max.   :1.000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000  
       we             pt        
 Min.   :0.00   Min.   :0.0000  
 1st Qu.:0.00   1st Qu.:0.0000  
 Median :0.00   Median :0.0000  
 Mean   :0.21   Mean   :0.0925  
 3rd Qu.:0.00   3rd Qu.:0.0000  
 Max.   :1.00   Max.   :1.0000  
```

```
m <- lm(wage ~ educ + exper, uswages)
summary(m)
```

```

Call:
lm(formula = wage ~ educ + exper, data = uswages)

Residuals:
    Min      1Q  Median      3Q     Max 
-1018.2  -237.9   -50.9   149.9  7228.6 
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -242.7994    50.6816  -4.791 1.78e-06 ***
educ          51.1753     3.3419  15.313  < 2e-16 ***
exper          9.7748     0.7506  13.023  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 427.9 on 1997 degrees of freedom
Multiple R-squared:  0.1351,    Adjusted R-squared:  0.1343 
F-statistic:   156 on 2 and 1997 DF,  p-value: < 2.2e-16
```

For every year of education the model estimates a $51 per week increase in wage. For every year of experience the model esimates only $10.

```r
update(m, log(wage, 10) ~ .) %>% summary()
```

```
Call:
lm(formula = log(wage, 10) ~ educ + exper, data = uswages)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.19572 -0.15180  0.04639  0.19025  1.55037 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.019608   0.034029   59.35   <2e-16 ***
educ        0.039306   0.002244   17.52   <2e-16 ***
exper       0.007851   0.000504   15.58   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.2873 on 1997 degrees of freedom
Multiple R-squared:  0.1749,    Adjusted R-squared:  0.174 
F-statistic: 211.6 on 2 and 1997 DF,  p-value: < 2.2e-16
```

Log transformation of the response, means we interpret the coefficients as magitude (multiplicative) changes. So for each year of education the model expects a 4% increase in wage. For each year of experience < 1%.

3.

```r
x <- 1:20
y <- x + rnorm(20)

m <- lm(y ~ I(x^2))
summary(m)
```

```
Call:
```

```
lm(formula = y ~ I(x^2))

Residuals:
    Min      1Q  Median      3Q     Max 
-2.7803 -0.5854 -0.0946  0.9263  2.1197 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.424344   0.491408   9.003 4.38e-08 ***
I(x^2)      0.043815   0.002585  16.949 1.64e-12 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 1.441 on 18 degrees of freedom
Multiple R-squared:  0.941, Adjusted R-squared:  0.9378 
F-statistic: 287.3 on 1 and 18 DF,  p-value: 1.643e-12
```

```
direct_calc <- function(x, y, degree = 2) {
  x_mat <- model.matrix(~I(x^degree))
  solve(crossprod(x_mat), crossprod(x_mat, y))
}

direct_calc(x, y)
```

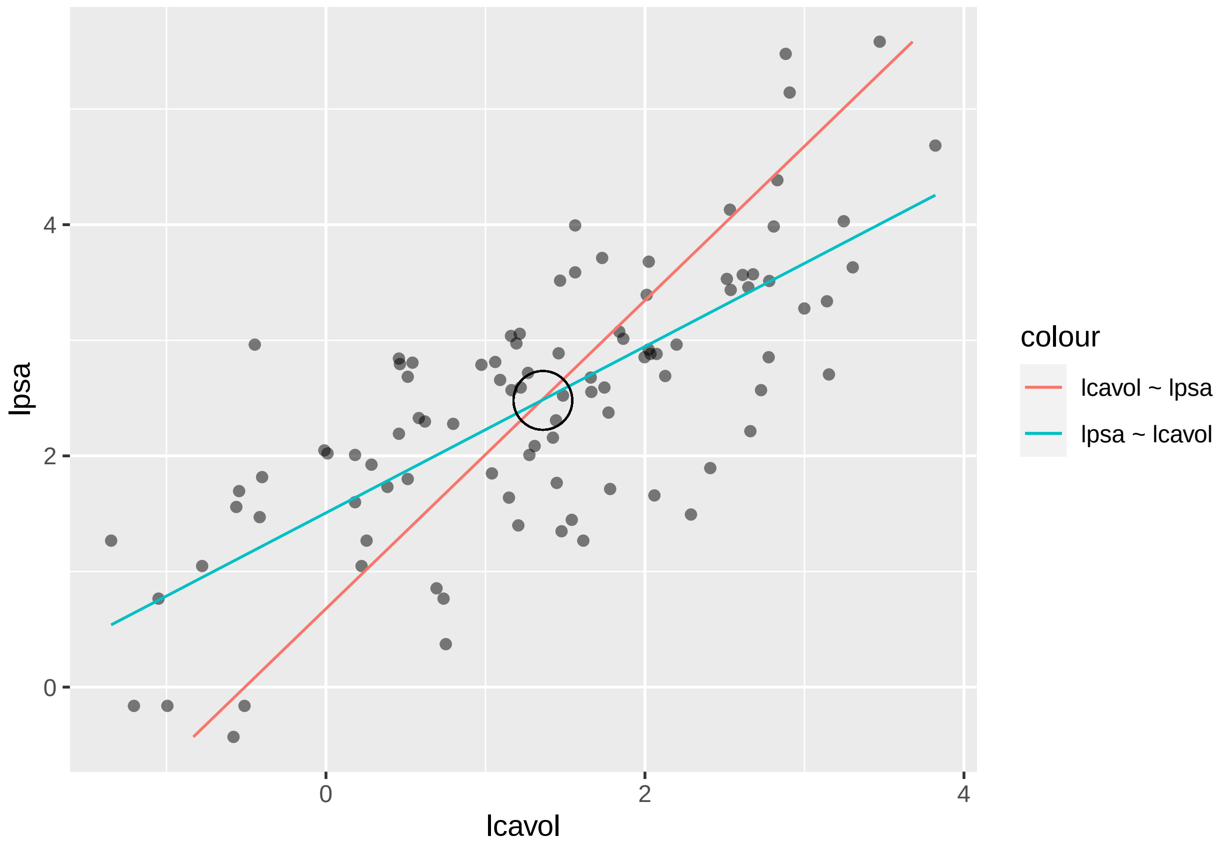
```
                  [,1]
(Intercept) 4.42434396
I(x^degree) 0.04381472
```

```
map(set_names(3:7), possibly(~ direct_calc(x, y, degree = .), "Error"))
```

```
$‘3‘
                   [,1]
(Intercept) 6.004637250
I(x^degree) 0.002134748

$‘4‘
                    [,1]
(Intercept) 6.9137561565
I(x^degree) 0.0001051108

$‘5‘
                    [,1]
(Intercept) 7.515077e+00
I(x^degree) 5.183818e-06

$‘6‘
                    [,1]
(Intercept) 7.945381e+00
I(x^degree) 2.556065e-07

$‘7‘
[1] "Error"
```

5.

```
m <- lm(lcavol ~ lpsa, prostate)
m2 <- lm(lpsa ~ lcavol, prostate)

ggplot(prostate, aes(lcavol, lpsa)) +
  geom_point(alpha = .5) +
  geom_line(aes(x = predict(m), color = "lcavol ~ lpsa")) +
  geom_line(aes(y = predict(m2), color = "lpsa ~ lcavol")) +
  geom_point(aes(y = 2.48, x = 1.36), shape = 1, size = 10)
```



Algebra not showm but you can calculate the intesection if you solve for either `lcavol` or `lpsa` in the system of equations represented by the model coefficients.

6.

```
data("cheddar")
summary(cheddar)
```

```
     taste           Acetic           H2S             Lactic     
 Min.   : 0.70   Min.   :4.477   Min.   : 2.996   Min.   :0.860  
 1st Qu.:13.55   1st Qu.:5.237   1st Qu.: 3.978   1st Qu.:1.250  
 Median :20.95   Median :5.425   Median : 5.329   Median :1.450  
 Mean   :24.53   Mean   :5.498   Mean   : 5.942   Mean   :1.442  
 3rd Qu.:36.70   3rd Qu.:5.883   3rd Qu.: 7.575   3rd Qu.:1.667  
 Max.   :57.20   Max.   :6.458   Max.   :10.199   Max.   :2.010  
```

a.

```r
m <- lm(taste ~ ., cheddar)
summary(m)
```

```
Call:
lm(formula = taste ~ ., data = cheddar)

Residuals:
    Min      1Q  Median      3Q     Max 
-17.390  -6.612  -1.009   4.908  25.449 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) -28.8768    19.7354  -1.463  0.15540   
Acetic        0.3277     4.4598   0.073  0.94198   
H2S           3.9118     1.2484   3.133  0.00425 **
Lactic       19.6705     8.6291   2.280  0.03108 * 
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 10.13 on 26 degrees of freedom
Multiple R-squared:  0.6518,    Adjusted R-squared:  0.6116 
F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

b.

```r
cor(m$fitted.values, cheddar$taste)^2
```

```
[1] 0.6517747
```

R-Squared, the coeffecient of determination or percentage of variance explained

c.

```r
m2 <- update(m, . ~ -1 + .)
cor(m2$fitted.values, cheddar$taste)^2
```

```
[1] 0.6244075
```

d.

```r
m_mat <- model.matrix(m)

qr_decomp <- qr(m_mat)

backsolve(
  qr.R(qr_decomp), # upper-right
  t(qr.Q(qr_decomp)) %*% cheddar$taste
)
```

```
            [,1]
[1,] -28.8767696
[2,]   0.3277413
[3,]   3.9118411
[4,]  19.6705434
```