STATS 500 HW3

Minxuan Chen

2023-09-24

Table of contents

Github repo: https://github.com/PKUniiiiice/STATS_500

Problem 1

(a)

```r
library(faraway)
data(cheddar)
model1 <- lm(taste ~ ., data=cheddar)
summary(model1)
```

```

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
```

From the result, `H2S` and `Lactic` are statistically significant at the 5% level.

(b)

```r
model2 <- lm(taste ~ exp(Acetic) + exp(H2S) + Lactic, data=cheddar)
summary(model2)
```

```

Call:
lm(formula = taste ~ exp(Acetic) + exp(H2S) + Lactic, data = cheddar)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.209  -7.266  -1.651   7.385  26.335 
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.897e+01  1.127e+01  -1.684   0.1042
exp(Acetic)  1.891e-02  1.562e-02   1.210   0.2371
exp(H2S)     7.668e-04  4.188e-04   1.831   0.0786 .
Lactic       2.501e+01  9.062e+00   2.760   0.0105 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.19 on 26 degrees of freedom
Multiple R-squared:  0.5754,    Adjusted R-squared:  0.5264
F-statistic: 11.75 on 3 and 26 DF,  p-value: 4.746e-05
```

From the result, only `Lactic` is statistically significant at the 5% level.

(c)

No, because the models are not nested. In other words, if we set one model corresponding to the NULL hypothesis, we cannot apply that hypothesis to the other model (the full model) to obtain a reduced (null) model.

Regarding the fit to the data, we compare the R^2 values. Model (a) has an R^2 of 0.6518, while model (b) has an R^2 of 0.5754. Therefore, model (a), which includes predictors on a log scale, provides a better fit to the data.

(d)

If all other variables are held constant, and `H2S` is increased 0.01, by $\hat{\beta}_{\mathrm{H2S}}$, `taste` would increase 0.039118.

(e)

We have

$$\Delta_{\log \mathrm{H2S}} = 0.01 = (\log \mathrm{H2S})_n - (\log \mathrm{H2S})_o$$

thus

$$\frac{\mathrm{H2S}_n}{\mathrm{H2S}_o} = \exp 0.01 = 101.005\%$$

(f)

```r
# 95% CI
conf <- confint(model1, level=.95);conf
```

```
                 2.5 %    97.5 %
(Intercept) -69.443503 11.689964
Acetic       -8.839420  9.494902
H2S           1.345656  6.478026
Lactic        1.933267 37.407820
```

```r
# 99% CI
confint(model1, level=.99)
```

```
                  0.5 %    99.5 %
(Intercept) -83.7158634 25.962324
Acetic      -12.0646491 12.720132
H2S           0.4428097  7.380872
Lactic       -4.3071367 43.648223
```

(g)

To construct a 95% confidence region for $(\beta_{H2S}, \beta_{Lactic})$, we need to use

$$\frac{n-p-1}{2}(\hat{\beta}-\beta)^T \left[\widehat{\text{Cov}(\hat{\beta})}\right]^{-1} (\hat{\beta}-\beta) \sim F_{2,n-p-1}$$

where $\beta = (\beta_{H2S}, \beta_{Lactic})^T = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \beta_{tot}, \quad \beta_{tot} = (1, \beta_{Acetic}, \beta_{H2S}, \beta_{Lactic})^T$ and $\widehat{\text{Cov}(\hat{\beta})}$ is the estimated covariance matrix of $\hat{\beta}$.

Therefore, the confidence region is defined by

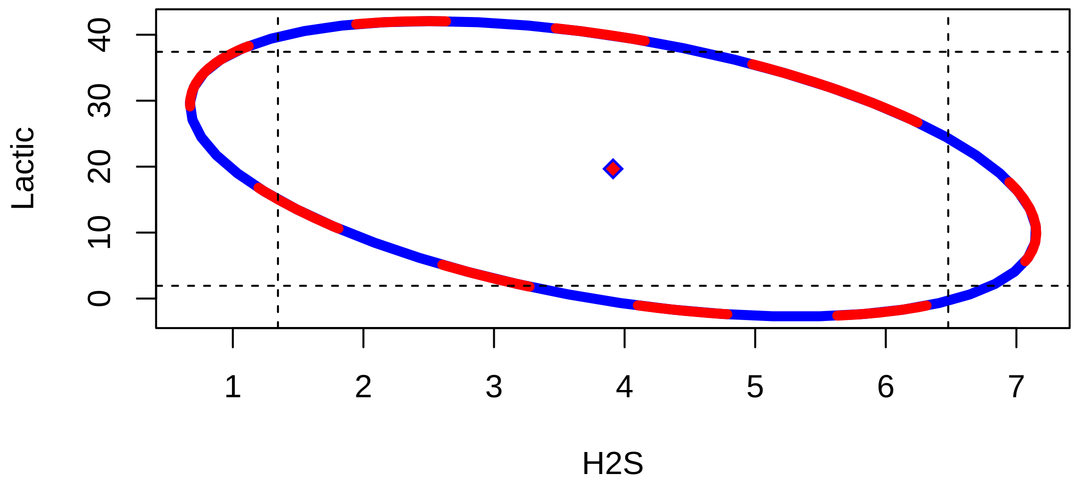
$$\frac{1}{2}(\hat{\beta}-\beta)^T \left[\widehat{\text{Cov}(\hat{\beta})}\right]^{-1} (\hat{\beta}-\beta) \leq F_{2,n-p-1}(\alpha)$$

Note that this expression defines an ellipse. To plot it, we can utilize the `car::ellipse` function. Furthermore, We can verify that this manually implemented plotting yields the same result as calling `ellipse::ellipse()` does. (For a more rigorous check, you need to refer to the source code of `ellipse.lm`, where you will find that it performs the same operation as described above.)

```r
Sys.setLanguage("en")
library(car)
```

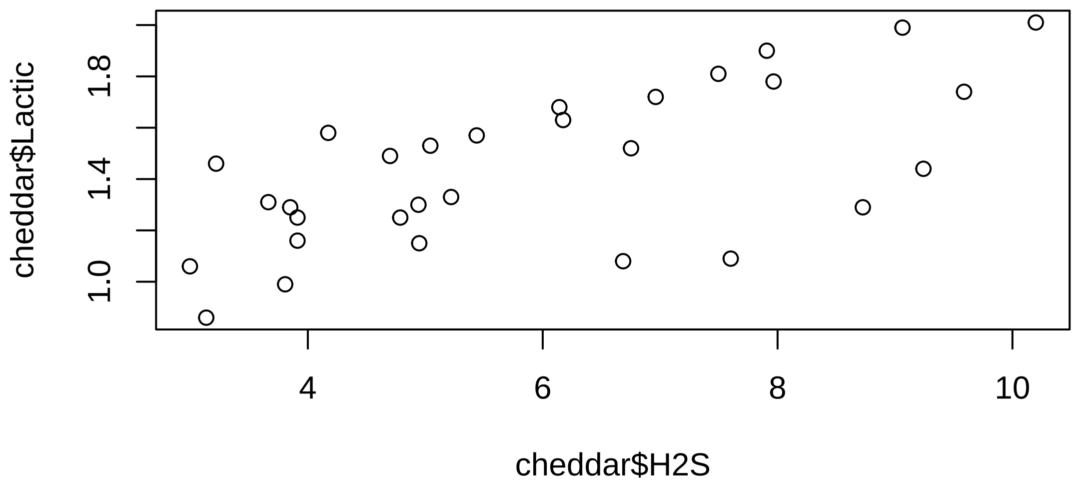
```
Loading required package: carData

Attaching package: 'car'

The following objects are masked from 'package:faraway':

    logit, vif
```

```r
center <- model1$coefficients[c(3,4)]
A <- vcov(model1)[c(3,4), c(3,4)]
const <- 2*qf(0.95, 2, model1$df.residual)
e1 <- car::ellipse(
  shape = A,
  center = center,
  radius = sqrt(const),
  draw = TRUE,
  add = FALSE,
  lwd = 5,
  lty= 1,
  col= 'blue',
  center.pch=18,
  grid=FALSE,
  xlab='H2S',
  ylab='Lactic'
)

lines(e2 <- ellipse::ellipse(model1, c('H2S','Lactic')),
     xlim=c(0,8), lwd=5, lty='aa', col='red')
points(model1$coefficients['H2S'], model1$coefficients['Lactic'],
       pch=18, col='red')
points(0, 0, pch=1)
abline(v=conf['H2S', ], lty=2)
abline(h=conf['Lactic', ], lty=2)
```



(h)

```r
plot(x=cheddar$H2S, y=cheddar$Lactic)
```




```r
cor(cheddar$H2S, cheddar$Lactic)
```

```
[1] 0.6448123
```

We can observe that the correlation coefficient between `H2S` and `Lactic` is positive, while the slope of the major axis of the confidence region (an ellipse) is negative. This opposing sign is not coincidental. In fact, the confidence ellipse is a rescaled clockwise rotation of the data ellipse (the Lactic-H2S plot we draw). A detailed explanation can be found in the book A Mathematical Primer for Social Statistics on page 217. The primary reason for this behavior is the covariance matrices corresponding to these two ellipses are "inversely" related to each other.

Problem 2

Before we begin the proof, we present a lemma at first.

Lemma: If $X \sim N(\mu, \Sigma)$, then $E(X^T AX) = \mathrm{tr}(A\Sigma) + \mu^T A\mu$.

Proof

$$\begin{aligned}
E\left(X^T AX\right) &= E(X-\mu)^T A(X-\mu) + \mu^T A\mu \\
&= \mathrm{tr}\left\{E\left[A(X-\mu)(X-\mu)^T\right]\right\} + \mu^T A\mu \\
&= \mathrm{tr}\left\{AE\left[(X-\mu)(X-\mu)^T\right]\right\} + \mu^T A\mu \\
&= \mathrm{tr}(A\Sigma) + \mu^T A\mu
\end{aligned}$$

(a)

$$\begin{aligned}
RSS &= \sum_{i=1}^{n}(y_i - \hat{y}_i) = Y^T(I-H)Y \\
&= (X\beta+\epsilon)^T(I-H)(X\beta+\epsilon)
\end{aligned}$$

Note that $E(\epsilon) = 0$, so, when we take expectation of RSS, any first-order term involving ϵ will disappear, allowing us to write

$$\begin{aligned}
E(RSS) &= E(\epsilon^T(I-H)\epsilon + (X\beta)^T(I-H)X\beta) \\
&= E(\epsilon^T(I-H)\epsilon) + \beta^T X^T X\beta - \beta^T X^T(X(X^TX)^{-1}X^T)X\beta \\
&= E(\epsilon^T(I-H)\epsilon)
\end{aligned}$$

Note that $\epsilon \sim N(0, \sigma^2 I)$, so we can use the lemma

$$\begin{aligned}
E(RSS) &= \mathrm{tr}((I-H)\sigma^2 I) + 0 \\
&= \sigma^2\,\mathrm{tr}(I-H) \\
&= \sigma^2(\mathrm{tr}(I) - \mathrm{tr}(X(X^TX)^{-1}X^T)) \\
&= \sigma^2(\mathrm{tr}(I_n) - \mathrm{tr}((X^TX)^{-1}X^TX)) \\
&= \sigma^2(\mathrm{tr}(I_n) - \mathrm{tr}(I_{p+1})) \\
&= (n-(p+1))\sigma^2
\end{aligned}$$

(b)

Use similar methods as in part (a), we have

$$\begin{aligned} E(RSS_o) &= E(\epsilon^T(I-H_o)\epsilon + (X\beta)^T(I-H_o)X\beta) \\ &= (n-p)\sigma^2 + \beta^T X^T (I-H_o) X\beta \end{aligned}$$

Next, we explicitly write out X and H_o. For simplicity, we use the term $X_{-p} = X^{(o)}$. We have

$$\begin{aligned} & X = [X_{-p} \quad X_p] \\ & H_o = X_{-p}(X_{-p}^T X_{-p})^{-1} X_{-p}^T \\ & \beta = \begin{bmatrix} \beta_{-p} \\ \beta_p \end{bmatrix} \end{aligned}$$

Thus

$$\begin{aligned} \beta^T X^T (I-H_o) X\beta &= \begin{bmatrix} \beta_{-p}^T & \beta_p^T \end{bmatrix} \begin{bmatrix} X_{-p}^T \\ X_p^T \end{bmatrix} (I-H_o) \begin{bmatrix} X_{-p} & X_p \end{bmatrix} \begin{bmatrix} \beta_{-p} \\ \beta_p \end{bmatrix} \\ &= \Big(\beta_{-p}^T X_{-p}^T + \beta_p^T X_p^T\Big)(I-H_o)\Big(X_{-p}\beta_{-p} + X_p\beta_p\Big) \\ &= \beta_{-p}^T X_{-p}^T (I-H_o) X_{-p}\beta_{-p} + \beta_{-p}^T X_{-p}^T (I-H_o) X_p \beta_p \\ &\quad + \beta_p^T X_p^T (I-H_o) X_{-p}\beta_{-p} + \beta_p^T X_p^T (I-H_o) X_p\beta_p \end{aligned}$$

Note that

$$H_o = X_{-p}(X_{-p}^T X_{-p})^{-1} X_{-p}^T$$

thus the first three terms will all be zero. So

$$E(RSS_o) = (n-p)\sigma^2 + \beta_p^T X_p^T (I-H_o) X_p \beta_p$$

since β_p is a number, we have

$$E(RSS_o) = \beta_p^2 \cdot X_p^T (I-H_o) X_p + (n-p)\sigma^2$$

(c)

When $E(RSS_o) = (n-p)\sigma^2$, we have $\beta_p^2 \cdot X_p^T(I-H_o)X_p = 0$. Let's delve deeper into this equation. Note that $I - H_o$ is both idempotent and symmetric.

$$\beta_p^2 \cdot X_p^T(I-H_o)X_p = \beta_p^2 \cdot ((I-H_o)X_p)^T(I-H_o)X_p = \beta_p^2 \cdot \tilde{X_p}^T \tilde{X_p} = 0$$

Since $\tilde{X_p}$ is a column vector, this equation is equal to zero if

$$\beta_p = 0 \text{ or } \tilde{X_p} = (I-H_o)X_p = 0$$

The first condition $\beta_p = 0$ means that the true model ($y = X\beta + \epsilon$) doesn't need the inclusion of X_p, i.e. it is redundant.

As for the second condition, we consider a non-trivial case, where $I - H_o \neq 0$. If we treat X_p as response and $1, X_1, \cdots, X_{p-1}$ as regressors, then $(I-H_o)X_p$ gives the residuals of the regression

$$X_p \sim 1 + X_1 + \cdots + X_{p-1}$$

Therefore, $(I-H_o)X_p = 0$ means that X_p can be perfectly fitted by $1, X_1, \cdots, X_{p-1}$. In other words, in the dataset, X_p is a linear combination of $1, X_1, \cdots, X_{p-1}$.