STAT0028 Week 2 Exercises

Andrew Wu

2024-10-08

Question 1

a)

$$E[Z] = AE[Y] + c$$

$$V(Z) = AV(Y)A^\top$$

b)

$$\begin{aligned} E[\hat{\beta}] &= (X^\top V^{-1} X)^{-1} X^\top V^{-1} E[Y] \\ &= (X^\top V^{-1} X)^{-1} X^\top V^{-1} X\beta + (X^\top V^{-1} X)^{-1} X^\top V^{-1} E[\epsilon] \\ &= \beta \end{aligned}$$

$$\begin{aligned} V(\hat{\beta}) &= (X^\top V^{-1} X)^{-1} X^\top V^{-1} V(X\beta + \epsilon) \left(X^\top V^{-1} X)^{-1} X^\top V^{-1}\right)^\top \\ &= (X^\top V^{-1} X)^{-1} X^\top V^{-1} V V^{-1} X (X^\top V^{-1} X)^{-1} \\ &= (X^\top V^{-1} X)^{-1} \end{aligned}$$

In the OLS case, we have $V = \sigma^2 I$ and $V(\hat{\beta}) = \sigma^2 (X^\top X)^{-1}$.

Question 2

a)

Define the design matrix X as usual but with second column being $(x_i - \bar{x})] \in \mathbb{R}^N$. Then,

$$\begin{aligned} \alpha &= (X^\top X)^{-1} X^\top y \\ &= (\bar{y}, \frac{C_{xy}}{C_{xx}})^\top \end{aligned}$$

b)

$$\begin{aligned} V(\alpha) &= \sigma^2 (X^\top X)^{-1} \\ &= \frac{1}{n\sum_i (x_i - \bar{x})^2 - (\sum_i x_i - \bar{x})^2} \begin{bmatrix} \sum_i (x_i - \bar{x})^2 & \sum_i x_i - \bar{x} \\ \sum_i x_i - \bar{x} & n \end{bmatrix} \\ &= \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^2}{C_{xx}} \end{bmatrix}. \end{aligned}$$

c)

$$\begin{aligned}
\text{RSS} &= \sum_i (y_i - (\alpha_0 + \alpha_1(x_i - \bar{x}))^2 \\
&= \sum_i (y_i - \bar{y} + \frac{C_{xy}}{C_{xx}}(x_i - \bar{x}))^2 \\
&= \sum_i (y_i - \bar{y})^2 - \frac{C_{xy}^2}{C_{xx}} \sum_i \left(\frac{2(y_i - \bar{y})(x_i - \bar{x})}{C_{xy}} - \frac{(x_i - \bar{x})^2}{C_{xx}} \right) \\
&= C_{yy} - \frac{C_{xy}^2}{C_{xx}}.
\end{aligned}$$

d)

The model in the R output is the same as the model described earlier, except that the predictor (log depth) has been demeaned. Therefore, we can expect the same OLS estimate for the slope, while the intercepts should differ by slope $\times \bar{x}$. The fitted values and hence predictions provided by both models should also be identical. Indeed, if we denote α_i, β_i, $i = 1, 2$, for the coefficients of the model above and the R model respectively we will see the same estimates:

```
xbar <- -0.88686
ybar <- 0.21475
Cxx <- 1.1482
Cxy <- 3.1738
Cyy <- 9.3516

alpha1 <- Cxy/Cxx
alpha1
```

```
## [1] 2.764153
```

```
alpha0 <- ybar
alpha0
```

```
## [1] 0.21475
```

```
beta0 <- 2.6661
beta0 + alpha1*xbar
```

```
## [1] 0.2146836
```

Checking the residual standard error:

```
RSS <- Cyy - Cxy^2/Cxx
n <- 10
p <- 2
sigmahat <- sqrt(RSS/(n-p))
sigmahat
```

```
## [1] 0.2689639
```

Checking the covariance matrix of parameter estimates:

```
varAlpha1 <- sigmahat^2/Cxx
covAlpha <- c(sigmahat^2/n, -xbar*varAlpha1, -xbar*varAlpha1, varAlpha1)
dim(covAlpha) <- c(2,2)
covAlpha
```

```
##              [,1]       [,2]
## [1,] 0.007234157 0.05587601
## [2,] 0.055876015 0.06300432
```

Question 3

a)

Let $v_{N+1} = x_{N+1}^\top (X^\top X)^{-1} x_{N+1}$ and $t_{N-p,\alpha/2}$ be the $1-\alpha/2$ quantile of the t-distribution with $N-p$ degrees of freedom.

$$\left[x_{N+1}^\top \hat{\beta} - t_{N-p,\alpha/2}\sqrt{\hat{\sigma}^2 v_{N+1}}, \quad x_{N+1}^\top \hat{\beta} + t_{N-p,\alpha/2}\sqrt{\hat{\sigma}^2 v_{N+1}}\right]$$

b)

Observe that

$$x_{N+1}^\top \hat{\beta} - x_{N+1}^\top \beta \sim \mathcal{N}(0, \sigma^2 v)$$

and since $e_{N+1} \sim \mathcal{N}(0,\sigma^2)$ is independent of of $x_{N+1}^\top \hat{\beta} - x_{N+1}^\top \beta$, we have

$$\frac{x_{N+1}^\top \hat{\beta} - x_{N+1}^\top \beta - e_{N+1}}{\sqrt{\sigma^2(v+1)}} \sim \mathcal{N}(0,1).$$

Since $\hat{\sigma}^2/\sigma^2 = \frac{\text{RSS}}{\sigma^2} \cdot \frac{1}{N-p}$ and $\frac{\text{RSS}}{\sigma^2} \sim \chi^2_{N-p}$, we have

$$\frac{x_{N+1}^\top \hat{\beta} - x_{N+1}^\top \beta - e_{N+1}}{\sqrt{\hat{\sigma}^2(v+1)}} \sim t_{N-p}.$$

c)

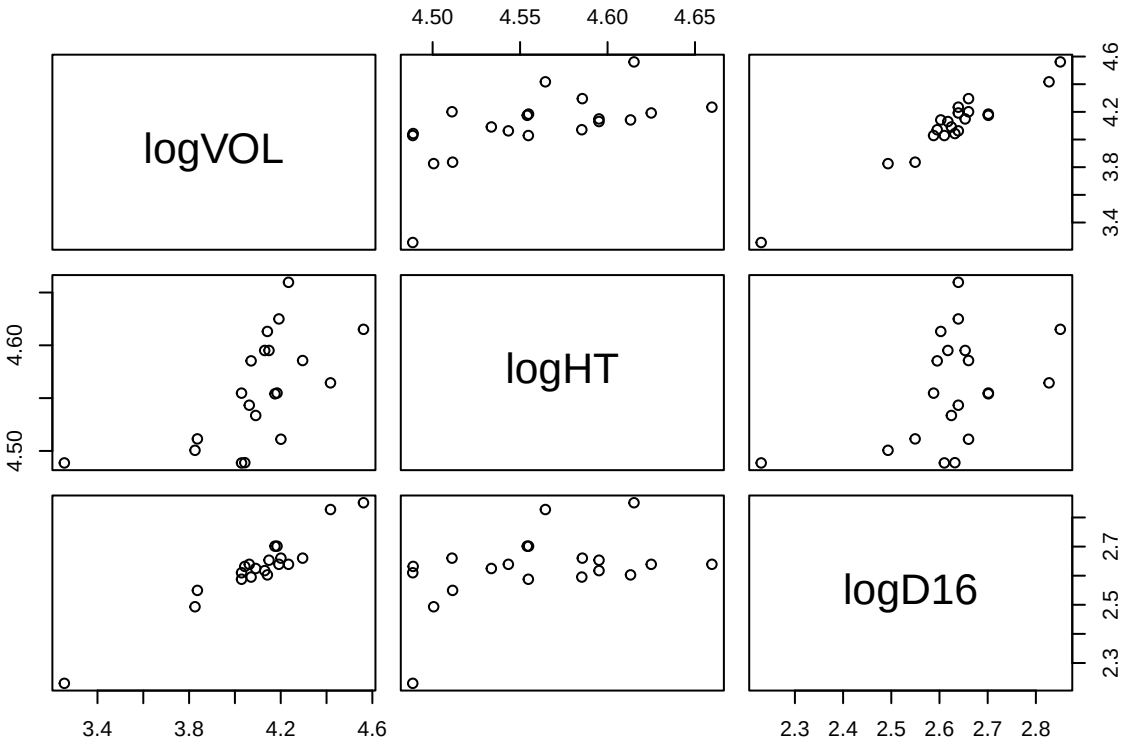
$$\left[x_{N+1}^\top \hat{\beta} - t_{N-p,\alpha/2}, \sqrt{\hat{\sigma}^2(v+1)}, \quad x_{N+1}^\top \hat{\beta} + t_{N-p,\alpha/2}\sqrt{\hat{\sigma}^2(v+1)}\right]$$

Question 4

In our comparison to the handout, we only consider the linear model with two predictors, as the model has a highly correlated predictor that is not particularly useful.

The log-transformed responses and predictors appear linearly related, in particular there is a strong linear relationship between `logVOL` and `logD16`. The relation between `logVOL` and `logHT` is weaker, and there is a clear outlier tree that had a low height and very low volume. The correlation between the two predictors is positive as one would expect, but not strong enough to indicate multicollinearity. Based on our cylindrical assumption, our log-transformed model should exhibit stronger linear relationships between the volume and the predictors, but this is not strongly evidenced by the scatterplots.

```
treevol <- read.table("treevol.txt",header=TRUE)
logHT <- log(treevol$HT)
logD16 <- log(treevol$D16)
logVOL <- log(treevol$VOL)
pairs(cbind(logVOL,logHT,logD16))
```



Upon fitting the log-linear model, we obtain a slightly higher `R-squared` of `0.9671` compared to the linear model and again a statistically significant `F-statistic` (at 5% power level) suggesting at least one predictor is useful. The p-values are both predictors are also statistically significant.

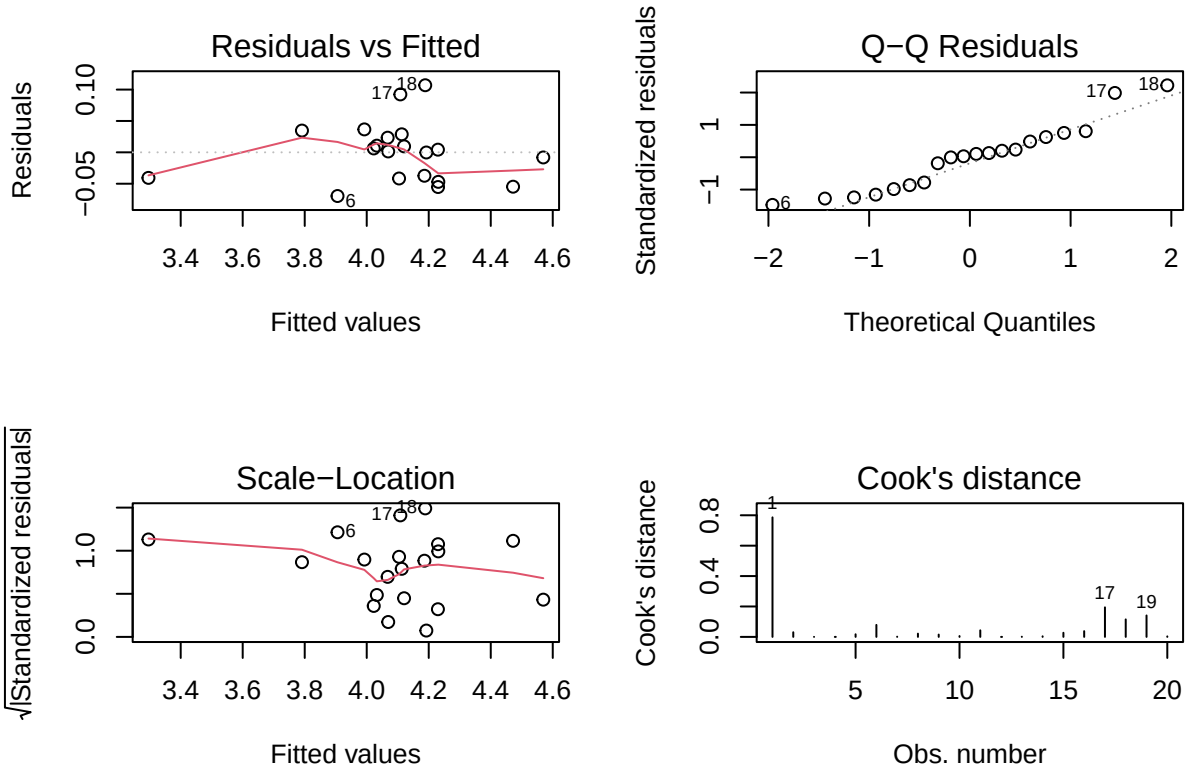
```
lmtreelog <- lm(logVOL ~ logHT + logD16)
summary(lmtreelog)
```

```
##
## Call:
## lm(formula = logVOL ~ logHT + logD16)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.069602 -0.040922  0.002851  0.024690  0.106917
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.6243     1.0605  -5.304 5.83e-05 ***
## logHT         1.0771     0.2532   4.255 0.000534 ***
## logD16        1.8321     0.1034  17.713 2.16e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.04976 on 17 degrees of freedom
## Multiple R-squared:  0.9671, Adjusted R-squared:  0.9632
## F-statistic: 249.9 on 2 and 17 DF,  p-value: 2.485e-13
```

There is no indication of heteroskadesticity from the (standardised) residuals vs fitted values plot and the normality assumption appears reasonable from the qqplot on the residuals. Interestingly, the linear model appears to have more moderately influential data points based on the Cook's distance plot, though both clearly identify observation one as a clear outlier and this is supported by the earlier scatterplots.

```
par(mfrow=c(2,2))
plot(lmtreelog, which=1:4)
```



Now, we compute the prediction interval for our log linear model and compare it to our linear model. Note that since our transform (log) is monotone, it suffices to take the inverse to compute our intervals.

```
newtree <- data.frame(HT=100,D16=10)
lognewtree <- data.frame(logHT=log(newtree$HT),logD16=log(newtree$D16))
exp(predict(lmtreelog,lognewtree,interval=c("confidence"),level=0.95))
```

```
##        fit      lwr      upr
## 1 34.97218 32.02958 38.18513
```

```
exp(predict(lmtreelog,lognewtree,interval=c("prediction"),level=0.95))
```

```
##        fit      lwr      upr
## 1 34.97218 30.49736 40.10359
```

```r
lmtree <- lm(VOL ~ D16 + HT, data=treevol)
predict(lmtree,newtree,interval=c("prediction"),level=0.95)
```

```
##        fit      lwr      upr
## 1 35.88012 27.23681 44.52344
```

We see that the prediction interval is much wider for the linear model than our log-linear model suggesting a lot more uncertainty in the estimate.

In conclusion, while both the log-linear and linear models offer strong predictive power for tree volume, and do not obviously violate the assumptions of the linear regression, there are a few reasons we might prefer the log-linear model:

- The log-linear model is strictly positive, and this makes sense for tree volume.
- The log-linear model is affected by a fewer number of influential observations
- The log-linear model offers more confidence when making predictions

Some further considerations beyond the scope of this report:

- Consider refitting the models without the outliers (eg. observation 1, and potentially others). Then, perform another comparison.
- Under the assumption that the log-linear model has normal errors, the inverse-transformed response is log-normally distributed. The mean response is given by

$$E[Y|x] = e^{x^T\beta}e^{\sigma^2/2}.$$

and therefore, when computing our predictions, we should multiply by a bias correction factor of $e^{\hat{\sigma}^2}/2$.