Scarlet Huang,
Nathan Tran, Alex Li

1. (a)
```
x <- c(2.68, 3.19, 3.87, 4.67, 5.74, 6.74, 8.45, 9.20, 11.07, 10.40)
y <- c(-3.01, 2.71, 4.14, 2.86, 5.86, 5.00, 7.24, 6.99, 9.38, 6.09)
model<-lm(y~x)
ti <- rstudent(model)[1]
```

Output
```
        1
-4.489831
```

The externally studentized residual is given by -4.489831 with $n = 10, p = 2$, and thus degrees of freedom $= 10 - 2 - 1 = 7$. We compute the p-value by $P(|T| \geq 4.489831)$. So on R we run,

```
2*(1-pt(abs(ti), 10-2-1))
```

Output
```
          1
0.002832655
```

Hence the p-value is $0.002832655 < 0.05$ and we reject H_0 under a 5% level.

(b) Same as above. We use the same test statistic and thus arrive at the conclusion that we reject H_0 under a 5% level.

(c) We have the summary output of the fitted model:

```
x <- c(2.68,3.19,3.87,4.67,5.74,6.74,8.45,9.20,11.07,10.40)
y <- c(-3.01,2.71,4.14,2.86,5.86,5.00,7.24,6.99,9.38,6.09)
dummy <- c(1,rep(0,9))
shiftmodel<-lm(y~x+dummy)
summary.lm(shiftmodel)

Call:
lm(formula = y ~ x + dummy)

Residuals:
     Min       1Q   Median       3Q      Max
-1.73692 -0.36891 -0.01861  0.70062  1.13856

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8962     0.9926   0.903  0.39659
x             0.6664     0.1316   5.064  0.00146 **
dummy        -5.6922     1.2678  -4.490  0.00283 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.073 on 7 degrees of freedom
Multiple R-squared:  0.9226,Adjusted R-squared:  0.9005
F-statistic: 41.71 on 2 and 7 DF,  p-value: 0.0001291
```

Here, we can see the estimate of Δ_1 is -5.6922.

(d) We first fit the model without the first observation and obtain the residual standard error and square it. Now, the reciprocal of this will be our weight in all entries except entry 1. In entry 1, our weight will just be the reciprocal of the standard deviation of the additional values plus the original value of -3.01. We do this because the weighted least squares estimates will have smaller standard errors if they are proportional to $1/\sigma^2$. This is the optimal choice in the sense it will give the smallest standard errors. Also this makes intuitive sense as if the sample additional data collected has a high variance, its weighting is small.

The code is given below.

```
x <- c(2.68, 3.19, 3.87, 4.67, 5.74, 6.74, 8.45, 9.20, 11.07, 10.40)
y <- c(-3.01, 2.71, 4.14, 2.86, 5.86, 5.00, 7.24, 6.99, 9.38, 6.09)

x1 <- x[-c(1)]
y1 <- y[-c(1)]
model<-lm(y1~x1)
std_error<-(summary.lm(model)$sigma)^2
var_outlier = var(c(-3.01, 13.40, 5.54, 1.84, -1.83, -18.91, -6.30, 3.82, 11.81))
w <- c(1/var_outlier, rep(1/std_error, 9))
weighted_model <- lm(y~x, weight=w)
summary.lm(weighted_model)

Call:
lm(formula = y ~ x, weights = w)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-1.6260 -0.5206 -0.1524  0.6590  1.0737

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8576     0.9482   0.904   0.3922
x             0.6708     0.1258   5.333   0.0007 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.9575 on 8 degrees of freedom
Multiple R-squared:  0.7805,Adjusted R-squared:  0.753
F-statistic: 28.44 on 1 and 8 DF,  p-value: 0.0006998
```

2. It has been proved in the previous assignment that $e_{j,-j} = \dfrac{e_j}{1 - H_{jj}}$.

We have,

$$
\begin{aligned}
e &= y - \widehat{y} \\
&= y - Hy \\
&= (I - H)y \\
&= (I - H)(X\beta + \varepsilon) \\
&= X\beta + \varepsilon - HX\beta - H\varepsilon
\end{aligned}
$$

Now the hat matrix is given by $H = X(X^T X)^{-1}X^T$ and so so $HX = X(X^T X)^{-1}X^T X = X$. This implies, $e = \varepsilon - H\varepsilon = (I - H)\varepsilon$ and so $e_j = ((I - H)\varepsilon)_j$.

Computing e_j explicitly yields,

$$e_j = (1 - H_{jj})\varepsilon_j - \sum_{k \neq j} H_{jk}\varepsilon_k.$$

Hence,

$$
\begin{aligned}
\mathbb{E}(e_j) &= \mathbb{E}\Big[(1 - H_{jj})\varepsilon_j - \sum_{k \neq j} H_{jk}\varepsilon_k\Big] \\
&= (1 - H_{jj})\mathbb{E}(\varepsilon_j) - \sum_{k \neq j} H_{jk}\mathbb{E}(\varepsilon_k) \\
&= (1 - H_{jj})\mathbb{E}(\varepsilon_j), \text{ since } \mathbb{E}(\varepsilon_k) = 0 \text{ for all } k \neq j \\
&= (1 - H_{jj})\Delta_j.
\end{aligned}
$$

From above, we have,

$$
\begin{aligned}
\mathbb{E}(e_{j,-j}) &= \mathbb{E}\Big[\frac{e_j}{1 - H_{jj}}\Big] \\
&= \frac{\mathbb{E}(e_j)}{1 - H_{jj}} \\
&= \frac{(1 - H_{jj})\Delta_j}{1 - H_{jj}} \\
&= \Delta_j.
\end{aligned}
$$

3. Firstly, input the data sourced from Daniel and Wood, where the response variable is the energy radiated from a carbon filament lamp per cm^2 per second, and the predictor variable is the absolute temperature of the filament in 1000 degrees Kelvin.

```
library(readxl)
filament <- read_excel("~/Desktop/MATH2931/MATH2931_Ass3Q3.xlsx")

> filament
```

```
 Energy Temperature
1  2.138        1.309
2  3.421        1.471
3  3.597        1.490
4  4.340        1.565
5  4.882        1.611
6  5.660        1.680
```

As it is suspected that there is a non-linear relationship between the responses y and the predictors x, that is $y_i \approx \beta_0+\beta_1 x_i^{\alpha}$, we need to apply the Box-Tidwell procedure to estimate the power α.

The original model is: $y_i = \beta_0 + \beta_1 x_i^{\alpha} + \epsilon_i$.

To transform this non-linear model into a linear one, take $f(c) := z^c$ and consider its first order Taylor Expansion around $c = 1$.

$$f(c) \approx f(1) + f'(1)(c-1) + z + zln(z)(c-1) \tag{1}$$

Applying this to the predictor x_i^{α}, we get: $x_i^{\alpha} \approx x_i + x_i \ln(x_i)(\alpha - 1)$.

Substituting this into our original model, our transformed model is: $y_i = \beta_0+\beta_1 x_i+\beta_1 x_i ln(x_i)(\alpha-1) + \epsilon_i$.

According to the Box-Tidwell procedure, we first fit the original model to obtain the value of the estimator b_1, which is the coefficient of the predictor 'Temperature'.

```
lmodel1 <- lm(filament$Energy~filament$Temperature)

> summary.lm(lmodel1)
Call:
lm(formula = filament$Energy ~ filament$Temperature)

Residuals:
      1        2        3        4        5        6
 0.14341 -0.11087 -0.11516 -0.08386  0.02163  0.14486

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -10.4270     0.7201  -14.48 0.000132 ***
filament$Temperature   9.4893     0.4720   20.11 3.61e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1366 on 4 degrees of freedom
Multiple R-squared:  0.9902,Adjusted R-squared:  0.9878
F-statistic: 404.2 on 1 and 4 DF,  p-value: 3.613e-05
```

From the summary output, the estimator $b_1 = 9.4893$.

Next, we fit the transformed model to obtain the value of the estimator $\beta_1(\alpha-1)$, which is the coefficient of the log term $x_i \ln(x_i)$. Let $\eta = \beta_1(\alpha - 1)$.

```
logterm <- filament$Energy*log(filament$Energy)
lmodel2 <- lm(filament$Energy~filament$Temperature + logterm)

> summary.lm(lmodel2)

Call:
lm(formula = filament$Energy ~ filament$Temperature + logterm)

Residuals:
        1         2         3         4         5         6
0.001125 -0.003245 -0.002290 -0.001268  0.012201 -0.006523

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -3.338362   0.221901  -15.04 0.000638 ***
filament$Temperature  3.865049   0.174954   22.09 0.000203 ***
logterm               0.255994   0.007854   32.60 6.35e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.008368 on 3 degrees of freedom
Multiple R-squared:      1,Adjusted R-squared:      1
F-statistic: 5.437e+04 on 2 and 3 DF,  p-value: 1.449e-07
```

From the summary output, $\eta = 0.255994$.

Now we have enough information to estimate the first iteration of the transformation power α.
$\hat{\alpha} = \frac{\hat{\eta}}{b_1} + 1 = 1.026977$.

Therefore, the power $\alpha = 1.026977$.

4. (a) Let e_i be the i-th standard basis in $\mathbb{R}^n$. Now,

$$
\begin{aligned}
\mathrm{Var}(\widehat{y}_i) &= \mathrm{Var}(e_i^T \widehat{y}) \\
&= \mathrm{Var}(e_i^T H_d y) \\
&= e_i^T H_d \mathrm{Var}(y) H_d^T e_i \\
&= \sigma^2 e_i^T H_d H_d^T e_i \\
&= \sigma^2 e_i^T H_d e_i \ \ \text{(since } H_d \text{ is symmetric and idempotent)} \\
&= \sigma^2 H_{d,ii}, \quad \text{(where } H_{d,ii} \text{ is the } i\text{-th diagonal element of } H_d\text{)}.
\end{aligned}
$$

Thus, $\sum_i \mathrm{Var}(\widehat{y}_i) = \sum_i \sigma^2 H_{d,ii} = \sigma^2 \mathrm{tr}(H_d)$ as required.

Now,

$$\begin{aligned}\sum_i \text{Var}(\widehat{y}_i) &= \sigma^2 \text{tr}(H_d) \\ &= \sigma^2 \text{tr}(X_d(X_d^T X_d)^{-1} X_d^T) \\ &= \sigma^2 \text{tr}(X_d^T X_d (X_d^T X_d)^{-1}) \quad \text{(Trace is invariant under cyclic permutations)} \\ &= \sigma^2 \text{tr}(I_{p\times p}) \\ &= \sigma^2 p.\end{aligned}$$

Hence, $\dfrac{\sum_i \text{Var}(\widehat{y}_i)}{\sigma^2} = p.$

(b) Computing the sum of the biases squared we have,

$$\begin{aligned}\sum_i \text{Bias}^2(\widehat{y}_i) &= \big[\mathbb{E}(\widehat{y}) - \mathbb{E}(y)\big]^T \big[\mathbb{E}(\widehat{y}) - \mathbb{E}(y)\big] \\ &= \|\mathbb{E}(\widehat{y}) - \mathbb{E}(y)\|^2 \\ &= \|\mathbb{E}(H_d y) - \mathbb{E}(y)\|^2 \\ &= \|H_d \mathbb{E}(y) - \mathbb{E}(y)\|^2 \\ &= \|(H_d - I)\mathbb{E}(y)\|^2 \\ &= \|(H_d - I)X_t\beta_t\|^2 \\ &= \big[(H_d - I)X_t\beta_t\big]^T \big[(H_d - I)X_t\beta_t\big] \\ &= (X_t\beta_t)^T (H_d - I)^T (H_d - I)X_t\beta_t \\ &= \beta_t^T X_t^T (H_d^T - I)(H_d - I)X_t\beta_t \\ &= \beta_t^T X_t^T \big[H_d^T H_d - H_d^T - H_d + I\big] X_t\beta_t \\ &= \beta_t^T X_t^T \big[H_d - H_d - H_d + I\big] X_t\beta_t \quad \text{(Since } H_d \text{ is symmetric and idempotent)} \\ &= \beta_t^T X_t^T \big[-H_d + I\big] X_t\beta_t\end{aligned}$$

Hence, $\displaystyle\sum_i \text{Bias}^2(\widehat{y}_i) = \beta_t^T X_t^T (I - H_d) X_t\beta_t.$

5. (a) Let X_i be the i-th row of the design matrix X. Consider $S_1 = \displaystyle\sum_{\text{all } j}(y_j - (X\beta)_j)^2$. We can make the split,

$$\begin{aligned}S_1 &= \sum_{\text{all } j}(y_j - (X\beta)_j)^2 \\ &= (\widehat{y}_{i,-i} - (X\beta)_i)^2 + \sum_{j\neq i}(y_j - (X\beta)_j)^2 \\ &\geq \sum_{j\neq i}(y_j - (X\beta)_j)^2\end{aligned}$$

Since b_{-i} is the least squares estimator for y_i, we can deduce that,

$$\begin{aligned}
S_1 &\geq \sum_{j\neq i}(y_j - (Xb_{-i})_j)^2 \\
&= \sum_{j\neq i}(y_j - X_j b_{-i})^2 \quad \text{(Note: } j \text{ is not indexed on the vector } b_{-i} \text{ because its fixed)}
\end{aligned}$$

Let us now consider $S_2 = \sum_{\text{all } j}(y_j - (Xb_{-i})_j)^2$. Again we can split,

$$\begin{aligned}
S_2 &= \sum_{\text{all } j}(y_j - (Xb_{-i})_j)^2 \\
&= (\widehat{y}_{i,-i} - (Xb_{-i})_i)^2 + \sum_{j\neq i}(y_j - (Xb_{-i})_j)^2 \\
&= (\widehat{y}_{i,-i} - \widehat{y}_{i,-i})^2 + \sum_{j\neq i}(y_j - (Xb_{-i})_j)^2 \qquad \text{(I)} \\
&= \sum_{j\neq i}(y_j - (Xb_{-i})_j)^2 \\
&= \sum_{j\neq i}(y_j - X_j b_{-i})^2
\end{aligned}$$

Hence b_{-i} minimizes S_1 and from S_2 we can conclude that $b_* = b_{-i}$. Thus, the linear model obtained from fitting all responses except the i-th is the same as the one obtained from fitting the data y_*.

(b) Since the L^2-norm is just the sum of squares of the entries in the vector, $y_{-i} - X_{-i}b_{-i}$ with the i-th observation removed, we have,

$$\begin{aligned}
\widehat{\sigma}^2_{-i} &= \frac{1}{n-1-p}\left\|y_{-i} - X_{-i}b_{-i}\right\|^2 \\
&= \frac{1}{n-1-p}\sum_{j\neq i}(y_j - (Xb_{-i})_j)^2 \qquad \text{(II)}
\end{aligned}$$

Now from the same calculation as (I) in part (a), noting we used y_i here instead of $\widehat{y}_{i,-i}$; because our response vector is different, we have,

$$\sum_{j\neq i}(y_j - (Xb_{-i})_j)^2 = \sum_{\text{all } j}(y_j - (Xb_{-i})_j)^2 - (y_i - \widehat{y}_{i,-i})^2$$

Consider $\sum_{\text{all } j}(y_j - (Xb_{-i})_j)^2$. We have,

$$\begin{aligned}
\sum_{\text{all } j}(y_j - (Xb_{-i})_j)^2 &= \left\|y - Xb + Xb - Xb_{-i}\right\|^2 \\
&= \left\|y - Xb\right\|^2 + \left\|Xb - Xb_{-i}\right\|^2 \qquad \text{(by orthogonality)}
\end{aligned}$$

Now since $\widehat{\sigma}^2 = \frac{1}{n-p}\left\|y - \widehat{y}\right\|^2 = \frac{1}{n-p}\left\|y - Xb\right\|^2 \iff (n-p)\widehat{\sigma}^2 = \left\|y - Xb\right\|^2$, the above becomes,

$$\sum_{\text{all } j}(y_j - (Xb_{-i})_j)^2 = (n-p)\widehat{\sigma}^2 + \left\|Xb - Xb_{-i}\right\|^2.$$

We now examine the far right, most term,

$$\begin{aligned}
\|Xb - Xb_{-i}\|^2 &= \|Xb - Xb_*\| \quad (\text{part } (a)) \\
&= \left\|X(X^TX)^{-1}X^Ty - X(X^TX)^{-1}X^Ty_*\right\| \\
&= \|Hy - Hy_*\| \\
&= \|Hy - H(y - (y_i - \widehat{y}_{i,-i})e_i)\| \quad (\text{Here}, e_i \text{ is the } i\text{-th standard basis not residual}) \\
&= \|(Hy_i - H\widehat{y}_{i,-i})e_i\| \\
&= \|He_i\|^2 (y_i - \widehat{y}_{i,-i})^2 \\
&= (He_i)^T(He_i)(y_i - \widehat{y}_{i,-i})^2 \\
&= e_i^T H^T He_i(y_i - \widehat{y}_{i,-i})^2 \\
&= e_i^T He_i(y_i - \widehat{y}_{i,-i})^2 \quad (\text{Idempotency of the hat matrix}) \\
&= H_{ii}(y_i - \widehat{y}_{i,-i})^2 \quad (\text{Diagonal extraction of the hat matrix}) \quad (\text{III})
\end{aligned}$$

Hence, $\sum_{\text{all } j}(y_j - (Xb_{-i})_j)^2 = (n-p)\widehat{\sigma}^2 + H_{ii}(y_i - \widehat{y}_{i,-i})^2$.

And finally,

$$\begin{aligned}
\widehat{\sigma}^2_{-i} &= \frac{1}{n-1-p}\{(n-p)\widehat{\sigma}^2 + H_{ii}e^2_{i,-i} - e^2_{i,-i}\} \\
&= \frac{1}{n-1-p}\{(n-p)\widehat{\sigma}^2 - e^2_{i,-i}(1 - H_{ii})\} \\
&= \frac{1}{n-1-p}\{(n-p)\widehat{\sigma}^2 - \frac{e_i^2}{(1-H_{ii})^2}(1 - H_{ii})\} \\
&= \frac{1}{n-1-p}\{(n-p)\widehat{\sigma}^2 - \frac{e_i^2}{(1-H_{ii})}\}
\end{aligned}$$

Thus, $\widehat{\sigma}_{-i} = \sqrt{\dfrac{(n-p)\widehat{\sigma}^2 - e_i^2/(1-H_{ii})}{n-p-1}}$.

(c) Since $D_i := \dfrac{(b-b_{-i})^TX^TX(b-b_{-i})}{p\widehat{\sigma}^2}$ we can rewrite the numerator in terms of a norm as usual,

$$
\begin{aligned}
D_i &= \frac{\|X(b - b_{-i})\|^2}{p\widehat{\sigma}^2} \\
&= \frac{H_{ii}(y_i - \widehat{y}_{i,-i})^2}{p\widehat{\sigma}^2} \quad \text{(by equation (III))} \\
&= \frac{H_{ii}e_i^2/(1 - H_{ii})^2}{p\widehat{\sigma}^2} \\
&= \frac{H_{ii}e_i^2}{p\widehat{\sigma}^2(1 - H_{ii})(1 - H_{ii})} \\
&= (\frac{e_i}{\widehat{\sigma}\sqrt{1 - H_{ii}}})^2 \frac{H_{ii}}{p(1 - H_{ii})} \\
&= \frac{r_i^2}{p} \frac{H_{ii}}{1 - H_{ii}}.
\end{aligned}
$$