MAST20005/MAST90058: Assignment 2 Solutions

1. (a) The 95% CI for μ is $(\bar{x} - c\frac{\sigma}{\sqrt{n}}, \bar{x} + c\frac{\sigma}{\sqrt{n}})$, where $c = \Phi^{-1}(0.975) = 1.96$. Since $\bar{x} = 6.0$, $\sigma = 0.6$ and $n = 9$, the 95% CI for μ is $(5.61, 6.39)$.

   (b) The sample size is

$$n = \left(\frac{c\sigma}{\epsilon}\right)^2 = \left(\frac{1.96 \times 0.6}{0.2}\right)^2 = 34.57.$$

   Therefore, we need a sample size of 35.

   (c)
```
x <- c(6.0, 5.7, 5.8, 6.5, 7.0, 6.3, 5.6, 6.1, 5.0)
t.test(x, conf.level = 0.95)

##
##  One Sample t-test
##
## data:  x
## t = 31.334, df = 8, p-value = 1.171e-09
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  5.558434 6.441566
## sample estimates:
## mean of x
##         6
```

   Therefore, the 95% CI for μ is $(5.56, 6.44)$. The width of CI is slightly wider than the CI in part (a).

2. (a) Here, we use $\hat{p} = 0.7$ as the worst-case scenario consistent with the given information, together with $c = \Phi^{-1}(0.975) = 1.96$ and $\epsilon = 0.05$,

$$n = \frac{c^2\hat{p}(1-\hat{p})}{\epsilon^2} = \frac{1.96^2 \times 0.7(1-0.7)}{0.05^2} = 322.69.$$

   The sample size required is 323.

   (b) Similar to above, $\hat{p} = 0.7$, $c = \Phi^{-1}(0.975) = 1.96$ and $\epsilon = 0.02$,

$$n = \frac{c^2\hat{p}(1-\hat{p})}{\epsilon^2} = \frac{1.96^2 \times 0.7(1-0.7)}{0.02^2} = 2016.84.$$

   The sample size required is 2017.

3.
```
library(datasets)
pres <- pressure$pressure
temp <- pressure$temperature
y <- log10(pres)
x <- 1 / (temp + 273.15 - 10)
Q3data <- data.frame(y = y, x = x)
```

   (a) $y = \log_{10}(\texttt{pressure})$ and $x = 1/(\texttt{temperature} + 273.15 - 10)$

   (b) The following code fits the model, $y_i = \alpha + \beta x_i + \varepsilon_i$:

```
m1 = lm(y ~ x, data = Q3data)
summary(m1)
```

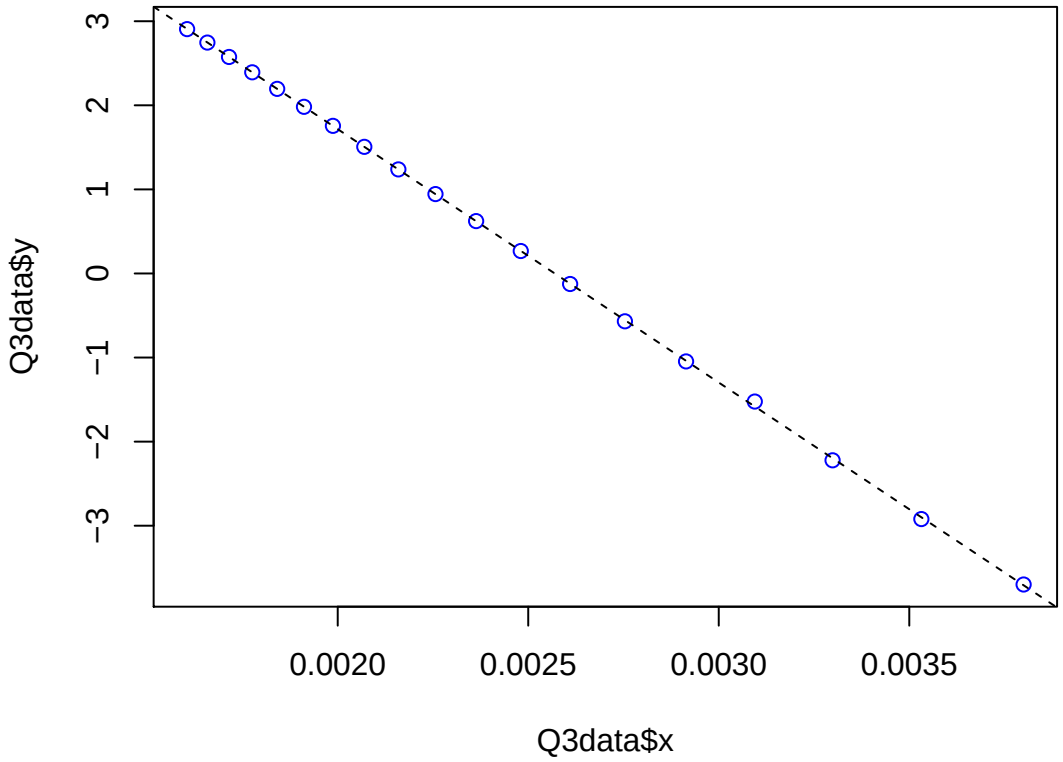
```
##
## Call:
## lm(formula = y ~ x, data = Q3data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.024465 -0.002310 -0.000690  0.001145  0.059228
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.744e+00  1.508e-02   513.4   <2e-16 ***
## x           -3.014e+03  6.041e+00  -498.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.017 on 17 degrees of freedom
## Multiple R-squared:  0.9999,Adjusted R-squared:  0.9999
## F-statistic: 2.489e+05 on 1 and 17 DF,  p-value: < 2.2e-16
```

Estimates $\hat{\alpha}$ and $\hat{\beta}$ are shown above, and can also be accessed via:

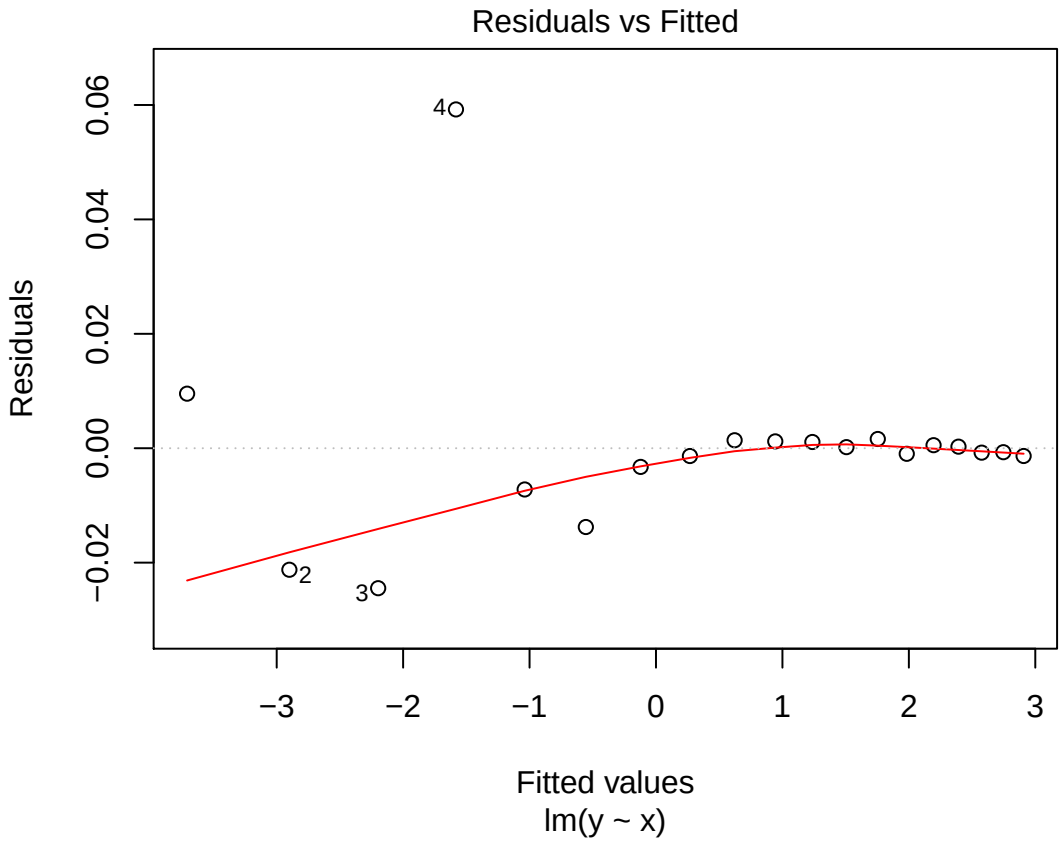
```
coef(m1)
```

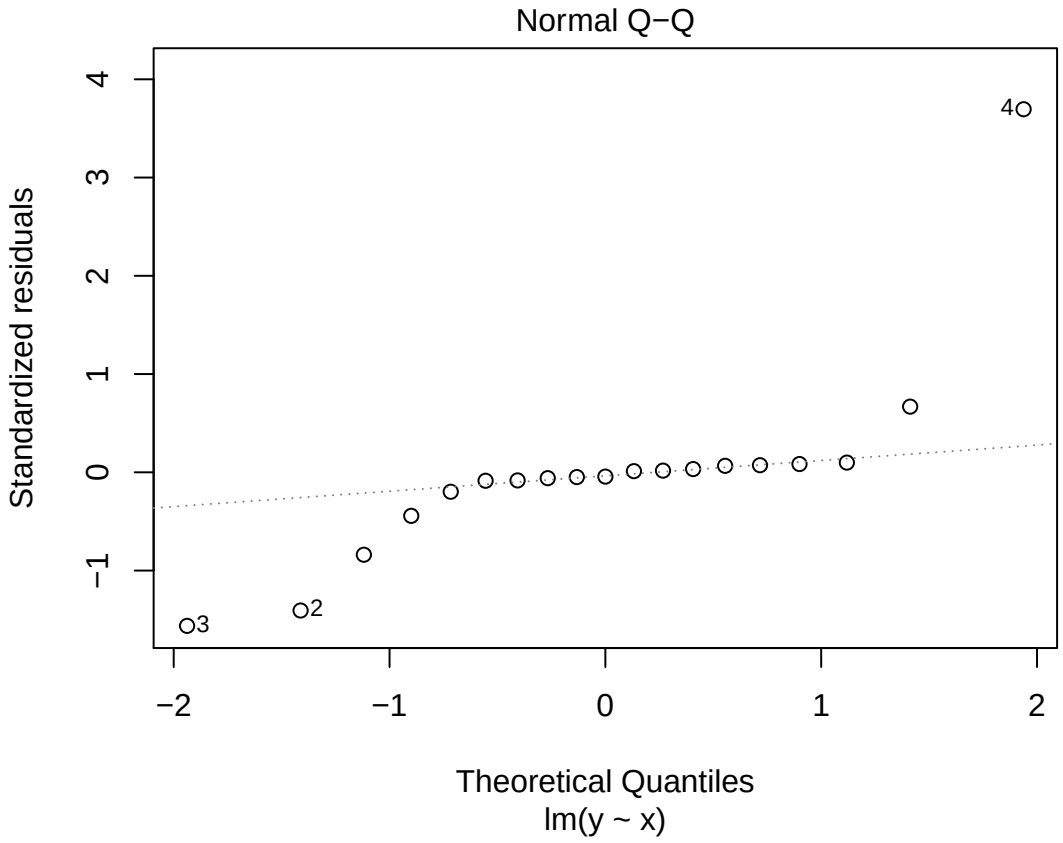
```
##  (Intercept)            x
##     7.743951 -3013.715440
```

(c)
```
plot(Q3data$x, Q3data$y, col = 4)
abline(m1, lty = 2)
```




```
plot(m1, 1:2)
```


Residuals vs Fitted
Residuals
0.06
0.04
0.02
0.00
−0.02
4
2
3
−3
−2
−1
0
1
2
3
Fitted values
lm(y ~ x)


Normal Q−Q
Standardized residuals
4
3
2
1
0
−1
4
3
2
−2
−1
0
1
2
Theoretical Quantiles
lm(y ~ x)

From the above plots, the linear model seems appropriate. The residuals are tiny compared to scale of the data. While there seems to be a bit of heteroskedasticity present, it's relatively minor and shouldn't be of great concern. (Note that we do not expect perfect compliance with the model assumptions.)

(d)
```
confint(m1)
```

```
##                    2.5 %       97.5 %
## (Intercept)     7.712128     7.775774
## x           -3026.460978 -3000.969902
```

Since $\alpha = 4.86$ is outside the confidence interval, the model does not support the claim $\alpha = 4.86$. On the other hand, $\beta = -3007$ is within the confidence interval, and the model supports the claim $\beta = -3007$.

(e)
```
xnew <- 1 / (70 + 273.15 - 10)
newdata <- data.frame(x = xnew)
CIvapor <- predict(m1, newdata, interval = "confidence")
10^CIvapor[-1]
```

```
## [1] 0.04860524 0.05116522
```

(f)
```
PIvapor <- predict(m1, newdata, interval = "prediction")
10^PIvapor[-1]
```

```
## [1] 0.04573684 0.05437405
```

4. We wish to compare the average speed between the two types of trains. The point estimates of the means suggest that Type A train is faster, we need to assess the strength of evidence for this.

   There are a number of valid approaches here, with differing assumptions and methods, including the choice of whether to use an estimation or a hypothesis testing approach. The following outlines one possible solution: a confidence interval for the difference in population means.

   We will use our usual notation for the parameters and statistics, and x and y to refer to Type A and Type B trains respectively.

   Using the Welch approximation gives a 95% confidence interval for $\mu_X - \mu_Y$ of the form:

$$\bar{x} - \bar{y} \pm F^{-1}(0.975)\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}$$

   where $F^{-1}(p)$ is the inverse cdf of t_r and the value of r is given by the Welch approximation formula (see lecture notes). For these data, we obtain $r = 19.6$ and $F^{-1}(0.975) = 2.09$. This results in the following interval:

$$500 - 496 \pm 2.09\sqrt{1.1^2/10 + 1.2^2/20} = (3.08, 4.92).$$

   We could also use the pooled variance estimator instead. This gives an interval of the form:

$$\bar{x} - \bar{y} \pm F^{-1}(0.975)s_P\sqrt{\frac{1}{n} + \frac{1}{m}}$$

where $F^{-1}(p)$ is the inverse cdf of t_{n+m-2} and s_P^2 is the pooled variance estimator. For these data, we have $n+m-2 = 28$, $F^{-1}(0.975) = 2.05$ and

$$s_P = \sqrt{\frac{9 \times 1.1^2 + 19 \times 1.2^2}{28}} = 1.169.$$

This results in the following interval:

$$500 - 496 \pm 2.05 \times 1.169\sqrt{1/10 + 1/20} = (3.07, 4.93).$$

Both methods show that Type A train is faster than Type B train.

Some further notes:

- If you calculate one-sided intervals rather than two-sided, you get the same conclusions.
- Hypothesis testing can also be used here, and will reach the same conclusion.

5. (a) $H_0 : p = 0.8$ versus $H_1 : p \neq 0.8$.

   (b) Here, $Y = 146$, and the test statistic $z = \frac{146-160}{\sqrt{160(1-0.8)}} = -2.47$. The p-value $= 2 \times \Phi(-2.47) = 0.0135$ is less than our significance level (0.05) so we reject the null hypothesis in this case. We have enough evidence to suggest that the proportion of users that are male differs from 0.8.

   (c) With a more stringent significance level of 0.01, we can no longer reject the null hypothesis.

   (d) Here $\hat{p} = 146/200 = 0.73$, and 95% confidence interval of p is $0.73 \pm 1.96\sqrt{\frac{0.73 \times 0.27}{200}} = (0.668, 0.792)$.

6. The cdf of X can be calculated in R using `ppois()`.

   (a) $\alpha = \Pr(X \geqslant 4 \mid \lambda = 2)$

```
1 - ppois(4, lambda = 2)
```

```
## [1] 0.05265302
```

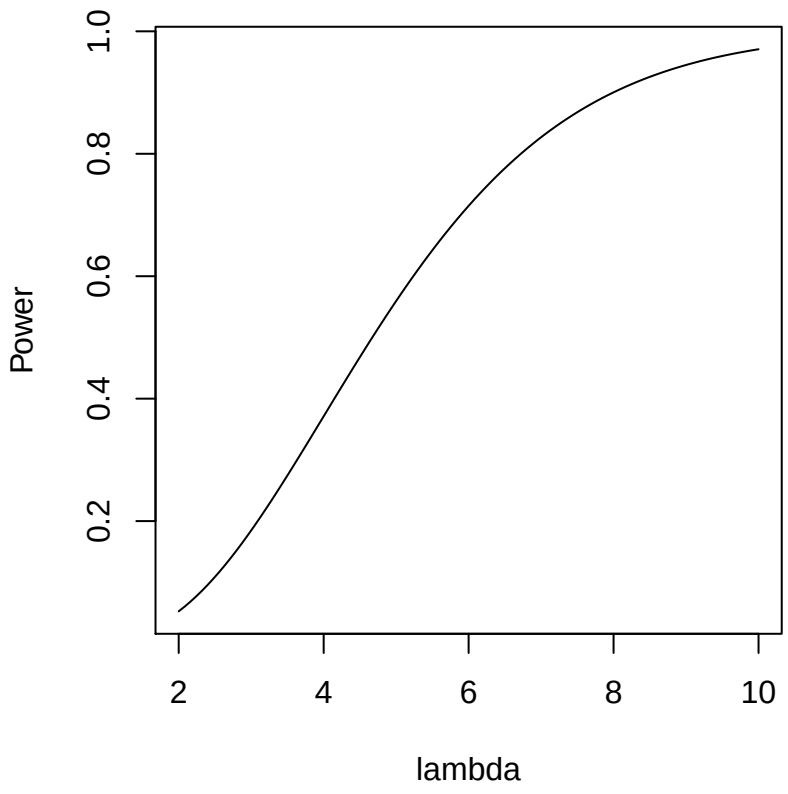
   (b) $\beta = \Pr(X \leqslant 4 \mid \lambda = 5)$

```
ppois(4, lambda = 5)
```

```
## [1] 0.4404933
```

   (c) $\text{Power}(p) = \Pr(X \geqslant 4 \mid \lambda)$

```
# There are various ways to draw this. This way is the most compact:
curve(1 - ppois(4, x), 2, 10, xlab = "lambda", ylab = "Power")
```



(d) Solve $0.05 = \Pr(X \geqslant c \mid \lambda = 2)$. This cannot be solved exactly due to the discreteness of X, but we can find the closest match. First, use the quantile function to find an approximate value:

```
qpois(0.95, 2)

## [1] 5
```

We can then check the actual significance level for various nearby options for c:

```
1 - ppois(4:6, 2)

## [1] 0.052653017 0.016563608 0.004533806
```

Let's use $c = 5$. This gives a test with rejection region $X \geqslant 5$ and has significance level **0.017**.
Alternatively, we can also choose $c = 4$, which gives a test with rejection region $X \geqslant 4$ and has significance level **0.053**.

Challenge: Can you do this question without using R?