MAST20005/MAST90058: Assignment 2 Solutions

1. (a) The 95% CI for μ is $\left(\bar{x} - c\frac{\sigma}{\sqrt{n}}, \bar{x} + c\frac{\sigma}{\sqrt{n}}\right)$, where $c = \Phi^{-1}(0.975) = 1.96$. Since $\bar{x} = 8$, $\sigma = 0.6$ and $n = 9$, the 95% CI for μ is $(7.61, 8.39)$.

   (b) The margin of error is $\epsilon = 0.2/2 = 0.1$. The required sample size is given by

$$n = \left(\frac{c\sigma}{\epsilon}\right)^2 = \left(\frac{1.96 \times 0.6}{0.1}\right)^2 = 138.3.$$

   Therefore, we need a sample size of at least 139.

   (c) This time the 95% CI for μ is $\left(\bar{x} - c\frac{s}{\sqrt{n}}, \bar{x} + c\frac{s}{\sqrt{n}}\right)$, where c is the 0.975 quantile from a t_8 distribution, which is $c = 2.306$. From the data we obtain $s = 0.652$. Therefore, the 95% CI for μ is $(7.5, 8.5)$. The width of the CI is a little wider than the CI from part (a).

2. (a) Here, we use $\hat{p} = 0.8$ as the worst-case scenario consistent with the given information, together with $c = \Phi^{-1}(0.975) = 1.96$ and $\epsilon = 0.05$,

$$n = \frac{c^2\hat{p}(1-\hat{p})}{\epsilon^2} = \frac{1.96^2 \times 0.8 \times (1-0.8)}{0.05^2} = 245.9.$$

   The sample size required is 246.

   (b) Similar to above, $\hat{p} = 0.8$, $c = \Phi^{-1}(0.975) = 1.96$ and $\epsilon = 0.02$,

$$n = \frac{c^2\hat{p}(1-\hat{p})}{\epsilon^2} = \frac{1.96^2 \times 0.8 \times (1-0.8)}{0.02^2} = 1536.6.$$

   The sample size required is 1537.

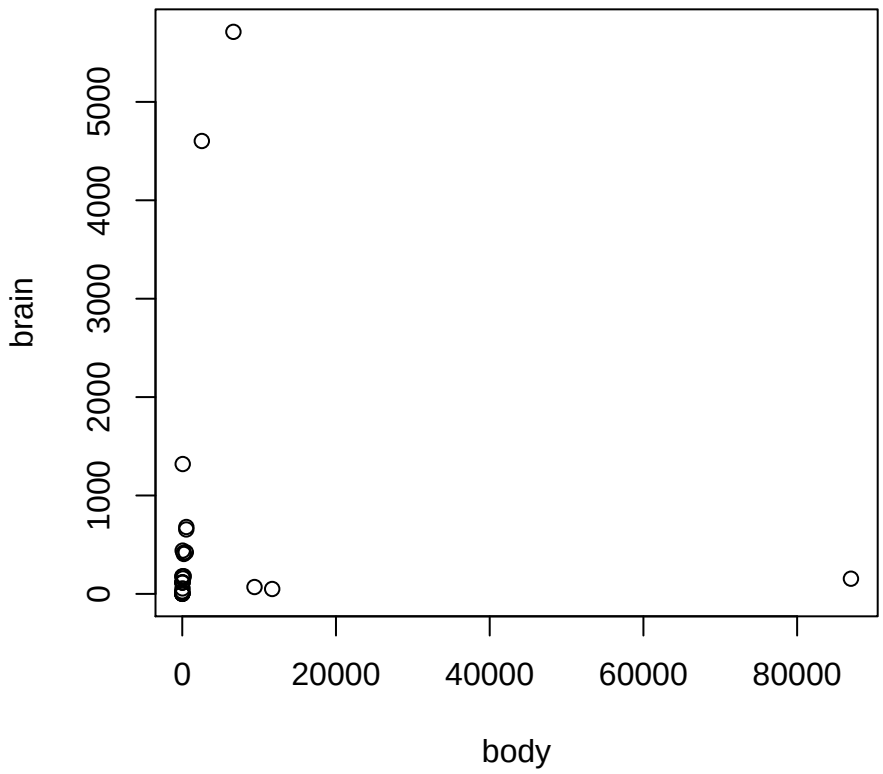
3. (a) Let's look at the data:

```
data(Animals, package = "MASS")
Animals
```

```
##                       body  brain
## Mountain beaver      1.350    8.1
## Cow                465.000  423.0
## Grey wolf           36.330  119.5
## Goat                27.660  115.0
## Guinea pig           1.040    5.5
## Dipliodocus      11700.000   50.0
## Asian elephant    2547.000 4603.0
## Donkey             187.100  419.0
## Horse              521.000  655.0
## Potar monkey        10.000  115.0
## Cat                  3.300   25.6
## Giraffe            529.000  680.0
## Gorilla            207.000  406.0
## Human               62.000 1320.0
## African elephant  6654.000 5712.0
## Triceratops       9400.000   70.0
```

```
## Rhesus monkey        6.800  179.0
## Kangaroo            35.000   56.0
## Golden hamster       0.120    1.0
## Mouse                0.023    0.4
## Rabbit               2.500   12.1
## Sheep               55.500  175.0
## Jaguar             100.000  157.0
## Chimpanzee          52.160  440.0
## Rat                  0.280    1.9
## Brachiosaurus    87000.000  154.5
## Mole                 0.122    3.0
## Pig                192.000  180.0
```

```
plot(Animals)
```

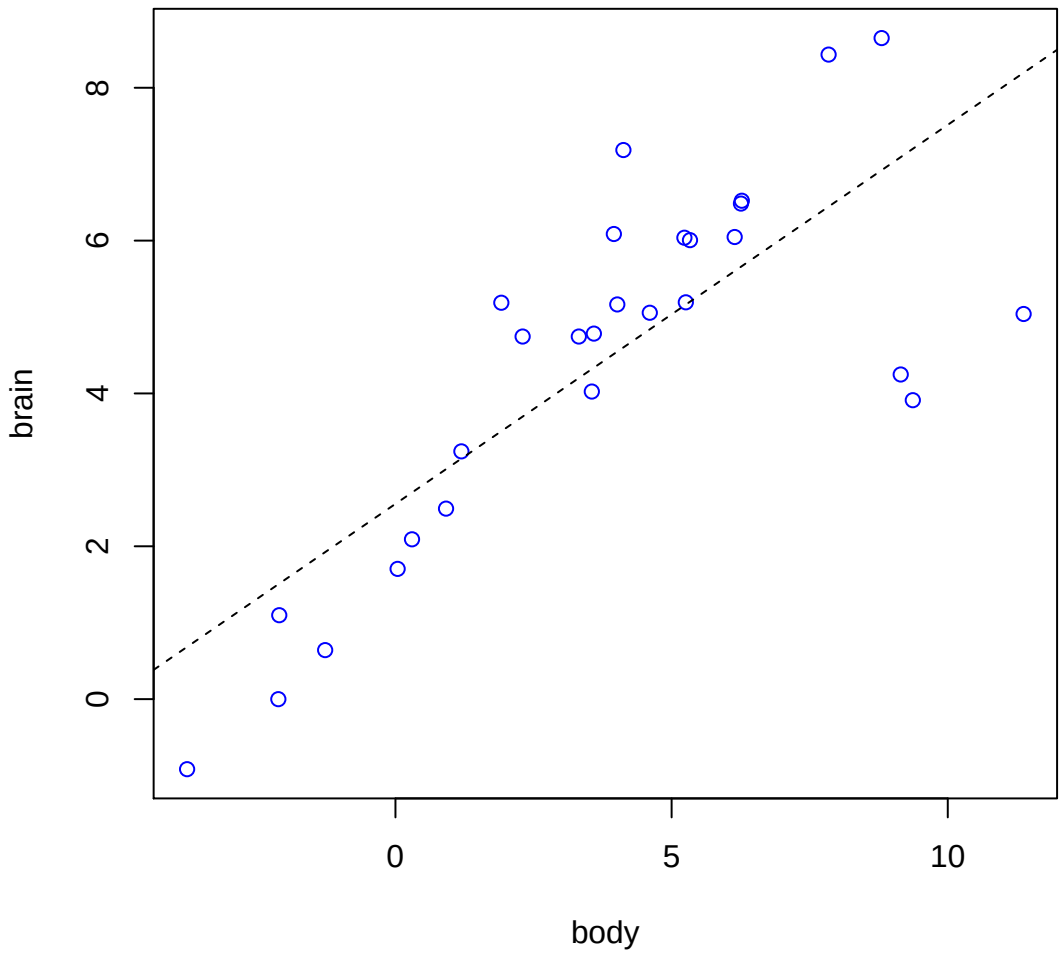


The measurements are highly skewed and clearly do not follow linear relationship. They are also bounded below by zero.

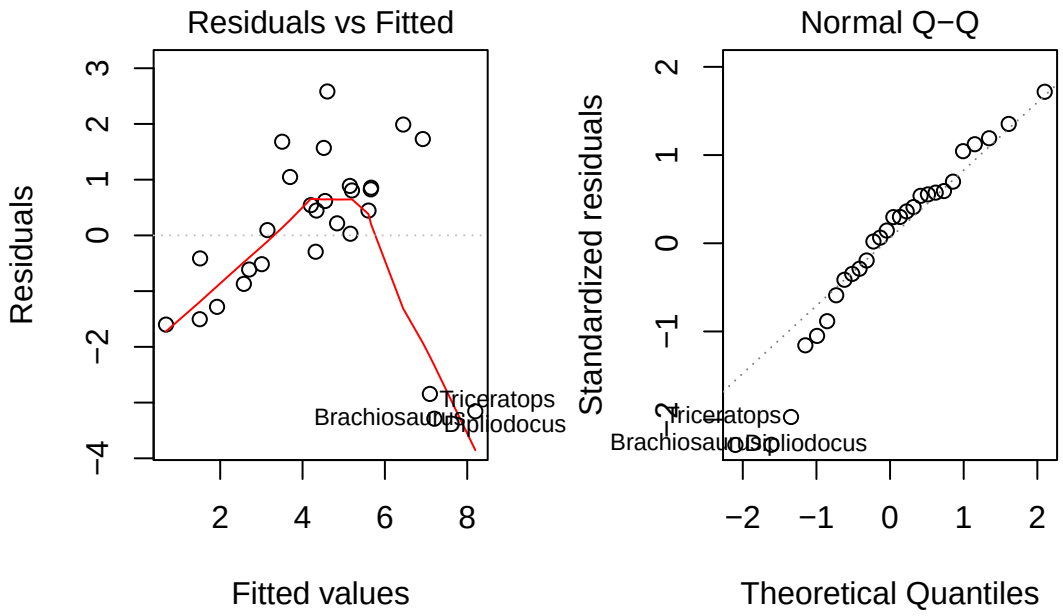
(b)
```
LogAnimals <- log(Animals)
m1 <- lm(brain ~ body, data = LogAnimals)
summary(m1)

##
## Call:
## lm(formula = brain ~ body, data = LogAnimals)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2890 -0.6763  0.3316  0.8646  2.5835
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.55490    0.41314   6.184 1.53e-06 ***
## body         0.49599    0.07817   6.345 1.02e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.532 on 26 degrees of freedom
## Multiple R-squared:  0.6076,Adjusted R-squared:  0.5925
## F-statistic: 40.26 on 1 and 26 DF,  p-value: 1.017e-06
```

(c)
```
plot(LogAnimals, col = 4)
abline(m1, lty = 2)
```



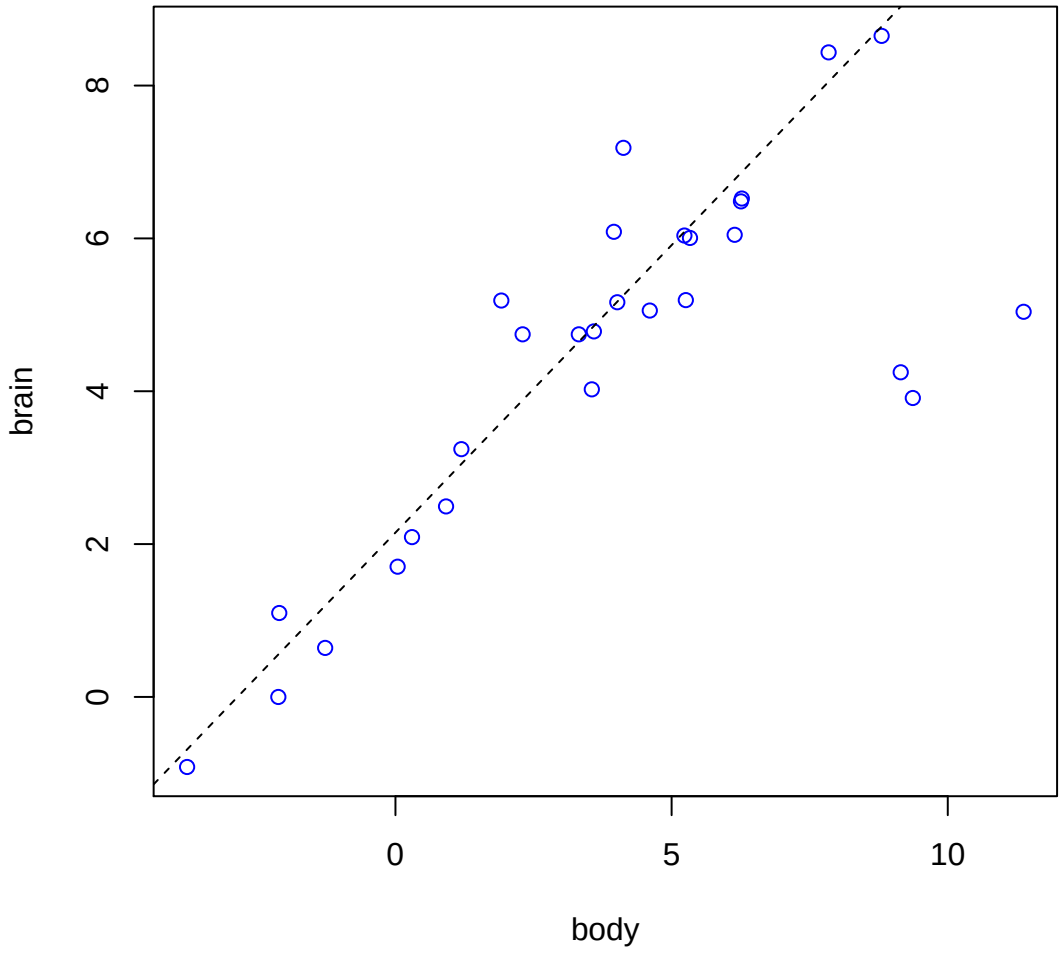
```
par(mfrow = c(1, 2))
plot(m1, 1:2)
```



There is a cluster of three points away from the rest. If you look closely at the data, you will notice that these three are all dinosaurs while the other animals are all mammals. This is a good basis for excluding the three points; the rest of the data will then reflect the brain–body relationship for mammals.

(d)
```
LogAnimals2 <- LogAnimals[-c(6, 16, 26), ]  # omit dinosaurs
m2 <- lm(brain ~ body, data = LogAnimals2)
summary(m2)

##
## Call:
## lm(formula = brain ~ body, data = LogAnimals2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9125 -0.4752 -0.1557  0.1940  1.9303
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.15041    0.20060   10.72 2.03e-10 ***
## body         0.75226    0.04572   16.45 3.24e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7258 on 23 degrees of freedom
## Multiple R-squared:  0.9217,Adjusted R-squared:  0.9183
## F-statistic: 270.7 on 1 and 23 DF,  p-value: 3.243e-14
```

(e)
```
plot(LogAnimals, col = 4)
abline(m2, lty = 2)
```



The new model is now a much better fit.

(f)
```
newdata = data.frame(body = log(500))
camelConfLog <- predict(m2, newdata, interval = "confidence")
camelConfLog

##        fit      lwr      upr
## 1 6.825418 6.399984 7.250851

exp(camelConfLog[-1])  # transform to usual scale (g)

## [1]  601.8355 1409.3038
```

4. We wish to compare the average plant growth between the two conditions. The point estimates of the means suggest that the CO_2-enriched atmosphere increases plant growth but we need to assess the strength of evidence for this.

   The parameter of interest is the difference in population means. We will assume a normal distribution for each of the two groups. Therefore, calculating a 95% confidence interval

for the differnce is one of our 'standard scenarios'. We just need to choose whether or not we should assume the variances are equal for the two groups.

The sample variances seem to differ quite a bit, so it's safer to not assume they are equal. It's also easy enough to do it both ways and see that there's only a small difference in the answers.

We will use our usual notation for the parameters and statistics, and x and y to refer to the enriched and normal conditions 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 $r = 10.3$ and $F^{-1}(0.975) = 2.22$. This results in the following interval:

$$8.21 - 7.36 \pm 2.22\sqrt{1.610^2/8 + 0.956^2/12} = (-0.554, 2.25).$$

Using the pooled variance estimator 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 $n + m - 2 = 18$, $F^{-1}(0.975) = 2.10$ and

$$s_P = \sqrt{\frac{7 \times 1.610^2 + 11 \times 0.956^2}{18}} = 1.252.$$

This results in the following interval:

$$8.21 - 7.36 \pm 2.10 \times 1.252\sqrt{1/8 + 1/12} = (-0.350, 2.05).$$

Either method shows that, while there's suggestive evidence that the CO_2-enriched air increases plant growth, the evidence is not strong enough to be conclusive. It is quite plausible that the effect of the enriched air is small, or even negative. More importantly, the confidence intervals are quite wide here (relative to the measurements), reflecting the fact that there's actually not much information we have to work with here, which is unsurprising given the sample sizes are so small.

Some further notes:

- It appears that there is a big difference in the variance of the measurements between the two conditions. If you compare the sample variances (e.g. using a confidence interval), you will see that we have insufficient evidence to say this reflects a true difference between the groups. More importantly, this assumption is not crticial since both of the methods used above give similar answers.
- If you calculate one-sided intervals rather than two-sided, you get the same conclusions.

5. (a) $H_0 : p_1 = p_2$ versus $H_1 : p_1 \neq p_2$.

   (b)
```
prop.test(x = c(120, 60), n = c(800, 600))
```
```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(120, 60) out of c(800, 600)
## X-squared = 7.2105, df = 1, p-value = 0.007248
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.01406774 0.08593226
## sample estimates:
## prop 1 prop 2
##   0.15   0.10
```

   The p-value is less than our significance level (0.05) so we reject the null hypothesis.

   (c) The conclusion is still the same even with a more stringent significance level of 0.01.

   (d) From the R output, a 95% confidence interval for $p_1 - p_2$ is (0.014, 0.086).

   Note: Similar answers are obtained if you don't use continuity correction:

```
prop.test(x = c(120, 60), n = c(800, 600), correct = FALSE)
```
```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(120, 60) out of c(800, 600)
## X-squared = 7.6503, df = 1, p-value = 0.005676
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.01552608 0.08447392
## sample estimates:
## prop 1 prop 2
##   0.15   0.10
```

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

   (a) $\alpha = \Pr(X \geqslant 4 \mid p = 0.4)$

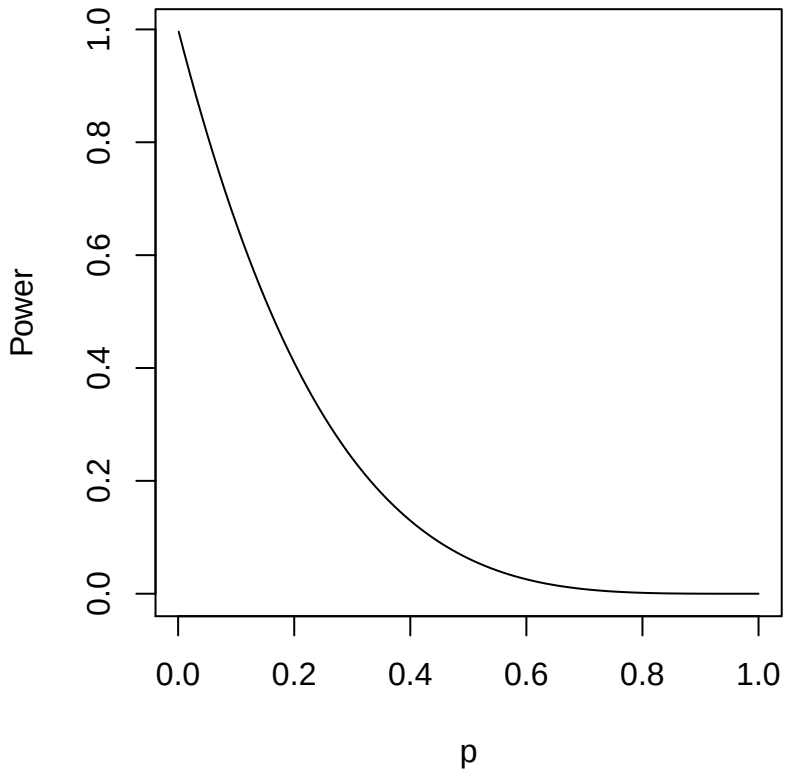
```
1 - pgeom(3, 0.4)
```
```
## [1] 0.1296
```

   (b) $\beta = \Pr(X \leqslant 3 \mid p = 0.2)$

```
pgeom(3, 0.2)
```
```
## [1] 0.5904
```

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

```
# There are various ways to draw this.  Here's a compact way:
curve(1 - pgeom(3, x), 0.001, 1, xlab = "p", ylab = "Power")
```



(d) Solve $0.05 = \Pr(X \geqslant c \mid p = 0.4)$. 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:

```
qgeom(0.95, 0.4)

## [1] 5
```

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

```
1 - pgeom(4:6, 0.4)

## [1] 0.0777600 0.0466560 0.0279936
```

Let's use $c = 6$. This gives a test with rejection region $X \geqslant 6$ and has significance level **0.047**.

(Note that $X \geqslant 6$ corresponds to `1 - pgeom(5, 0.4)` due to the discreteness of the distribution.)