Final Exam Lab Portion Solutions - R Studio

Dwight Wynne

December 20, 2018

Problem 1

If you're as bad at chem lab as I was, you will have a lot of measurement error affecting your results. Suppose that I am terrible at pipetting and, when I attempt to pipette 10 mL of water, the actual amount I pipette is normally distributed with mean 10 mL and standard deviation 0.3 mL.

Part (a)

What is the probability that I pipette less than 9.9 mL of water?

```
pnorm(9.9, mean = 10, sd = 0.3)
```

```
## [1] 0.3694413
```

The probability that I pipette less than 9.9 mL of water is 0.369.

Part (b)

What is the probability that, over 10 independent attempts, I average less than 9.9 mL of water in the pipette?

```
sd_xbar <- 0.3/sqrt(10)  # sd of sampling distribution
pnorm(9.9, mean = 10, sd = sd_xbar)
```

```
## [1] 0.1459203
```

The probability that I average less than 9.9 mL of water in the pipette over 10 independent attempts is 0.146.

Problem 2

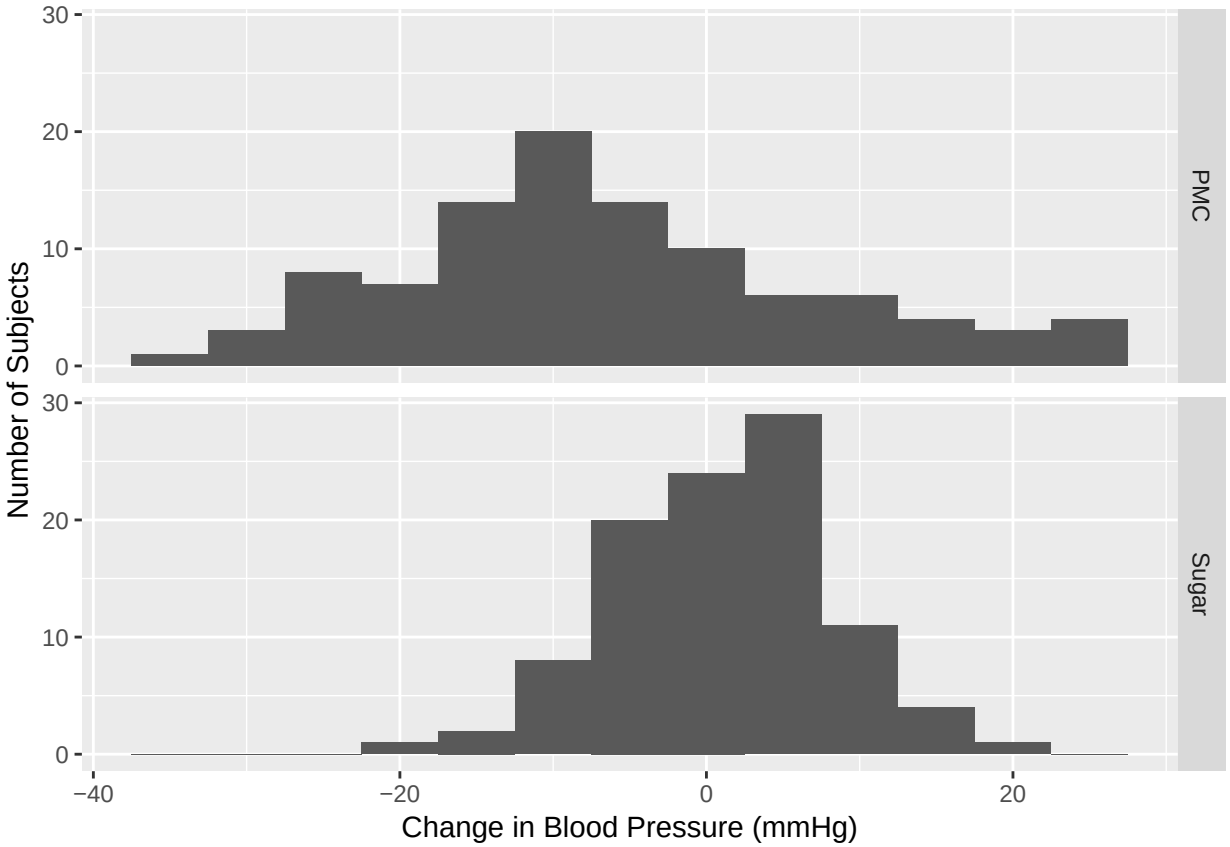
Having prolonged high blood pressure can have many negative effects on general health and can potentially lead to stroke and/or death. There are many drugs available that are known to help lower blood pressure, however, these drugs often come with negative side effects. A clinical trial was set up to explore a more natural treatment of taking a potassium, magnesium, and calcium (PMC) complex tablet. These minerals are known to counteract the blood pressure raising effects of high sodium intake.

A random sample of 200 adult males (29-49 years of age) currently diagnosed with hypertension (high blood pressure) were obtained from a pool of subjects who volunteered to be part of the study. The subjects were then randomly assigned to one of two groups: one group received the PMC complex tablet and the other received a sugar tablet.

Part (a)

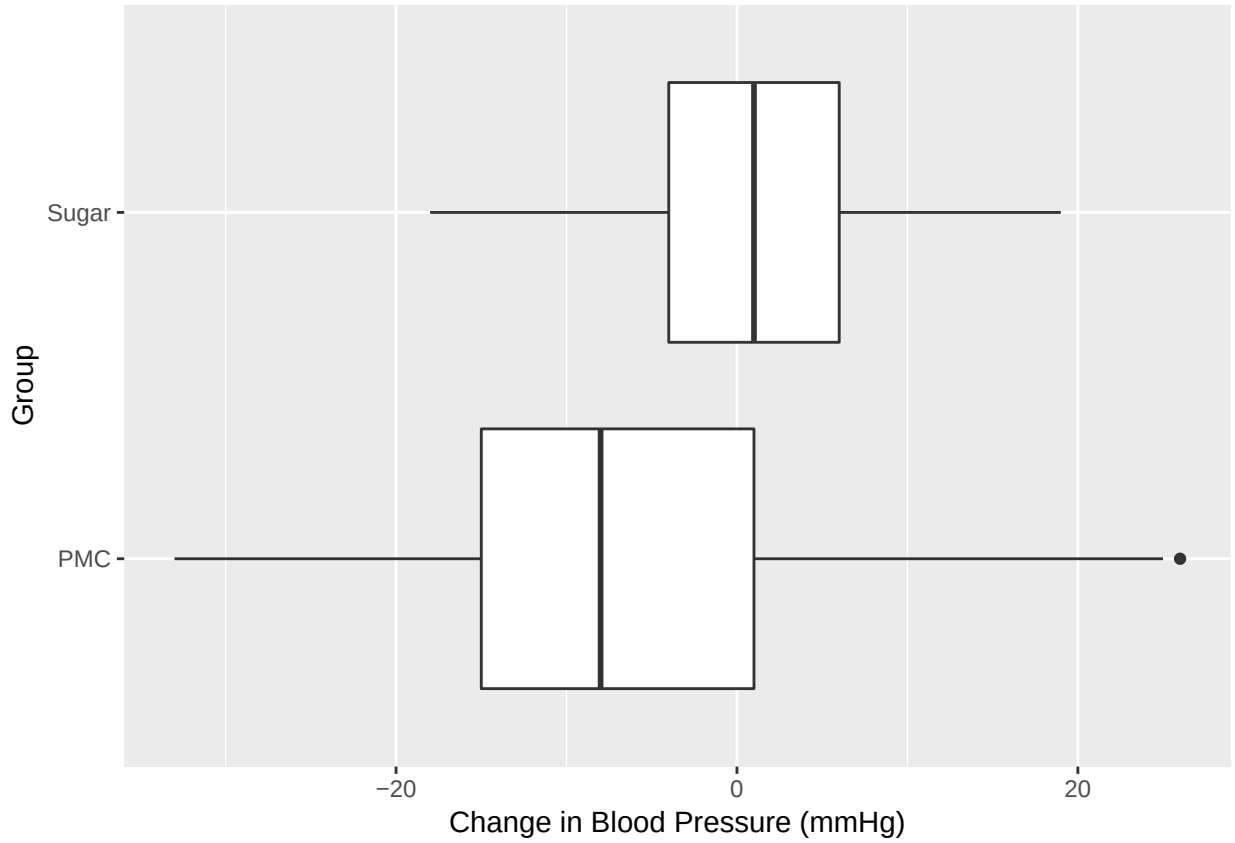
Perform exploratory analysis on this dataset - describe the distribution of the change in blood pressure in each group in the observed sample of 200 adult males and compare the two distributions.

```r
library (ggplot2)
PMC_hist <- ggplot(PMC, mapping = aes(x = Change)) +
  geom_histogram(center = 5, binwidth = 5)
PMC_hist_labeled <- PMC_hist +
  labs(x = "Change in Blood Pressure (mmHg)", y = "Number of Subjects") +
  facet_grid(Group~.)  # histograms one on top of the other
print(PMC_hist_labeled)
```



According to the histograms above, the distribution of the change in the PMC group is skewed right with a peak around -10, while the distribution of the chnge in the Sugar group is slightly skewed left with a peak around 5. The variability in the PMC group is considerably larger than the variability in the Sugar group.

```r
library (ggplot2)
PMC_box <- ggplot(PMC, mapping = aes(x = Group, y = Change)) +
  geom_boxplot()
PMC_box_labeled <- PMC_box +
  labs(x = "Group", y = "Change in Blood Pressure (mmHg)") + coord_flip()
print(PMC_box_labeled)
```



The boxplots confirm the difference in variability, with one outlier identified in the PMC group, and the difference in center, with the median in the PMC group being negative and the median in the Sugar group being positive. The boxplots confirm that the PMC group is skewed right, but suggest that the Sugar group may look roughly symmetric with a different choice of bins.

Part (b)

Perform an appropriate statistical hypothesis test to draw conclusions about the effectiveness of the PMC complex. Write a short paragraph describing your conclusions.

Based on our background analysis and the exploratory analysis in part (a), we have two independent samples with quite different variation; although there is an outlier in one of the groups, we have a reasonably large sample size (100 in each group), we should be okay to do a two-sample t-test.

H_0: the two groups have the same population mean, $\mu_{PMC} = \mu_{Sugar}$; or, the PMC tablet has no effect on the mean change in blood pressure (compared to placebo)

H_a: the PMC complex is effective at reducing blood pressure, $\mu_{PMC} < \mu_{Sugar}$

Appropriate alternative H_a: the PMC tablet has an effect on the mean change in blood pressure (compared to placebo), $\mu_{PMC} \neq \mu_{Sugar}$

Significance level: $\alpha = 0.05$

We perform a two-sample t-test:

```
t.test(Change ~ Group, data = PMC, alternative = "less")
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  Change by Group
## t = -4.8834, df = 151.79, p-value = 1.309e-06
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -4.91864
## sample estimates:
##   mean in group PMC mean in group Sugar
##               -6.53                0.91
```

Because the p-value of $0.0000013 < 0.05$, we reject the null hypothesis.

If you used a two-sided alternative

```
t.test(Change ~ Group, data = PMC, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  Change by Group
## t = -4.8834, df = 151.79, p-value = 2.618e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.450048  -4.429952
## sample estimates:
##   mean in group PMC mean in group Sugar
##               -6.53                0.91
```

Because the p-value of $0.0000026 < 0.05$, we reject the null hypothesis.

Conclusion: we found evidence that the PMC tablet does significantly decrease blood pressure (as compared to placebo); or, if you used a two-sided alternative, we found evidence of a significantly different effect of the PMC tablet on blood pressure as compared to placebo.

Problem 3

A 2018 study investigated whether macaque monkeys can perform statistical reasoning. In the study, 11 macaques watched a researcher select food from two buckets. On each trial, the researcher drew a piece of food from a bucket containing 80% grapes with one hand and drew from a bucket containing only 20% grapes with the other hand. After looking at the drawing, the monkeys indicated which hand they wanted food from. Each monkey selected a hand on 12 trials.

Estimate with 95% confidence the proportion of the time that Sophie would make the "correct" choice if presented with very many trials. Do you believe that Sophie is picking a hand entirely at random, or do you think that she is thinking about which hand is more likely to have the grapes?

```
head(monkey)
```

```
## # A tibble: 6 x 2
##   Individual Score
##   <chr>      <int>
## 1 Paul           6
## 2 Sophie         7
## 3 Max            6
## 4 Ilia           6
## 5 Lenny          6
```

```
## 6 Lord            7
```

Sophie is in row 2; she got 7 out of 12 correct.

We would like to perform a one-sample confidence interval for a population proportion here. The B and N assumptions of BINS are definitely met: "success" is when Sophie picks the "correct" hand and "failure" is when she does not, and there are 12 trials.

If you believe that Sophie is learning during these trials, then the I and S assumptions of BINS are not met. If you do not believe that she is learning, then the I assumption of BINS is met (the trials are independent) and it's reasonable to assume the S is too (she has the same chance of picking the "correct" hand on each trial).

Either way, we only have 7 successes and 5 failures, so we should do a binomial exact confidence interval:

```
binom.test(x = 7, n = 12, conf.level = 0.95)
```

```
##
##  Exact binomial test
##
## data:  7 and 12
## number of successes = 7, number of trials = 12, p-value = 0.7744
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.2766697 0.8483478
## sample estimates:
## probability of success
##              0.5833333
```

If you believe that all four of the BINS assumptions are met, you can interpret as: We are 95% confident that Sophie will get between 27.7% and 84.8% correct in the long run. Since this interval includes 50%, it is certainly plausible that Sophie is picking a hand entirely at random.

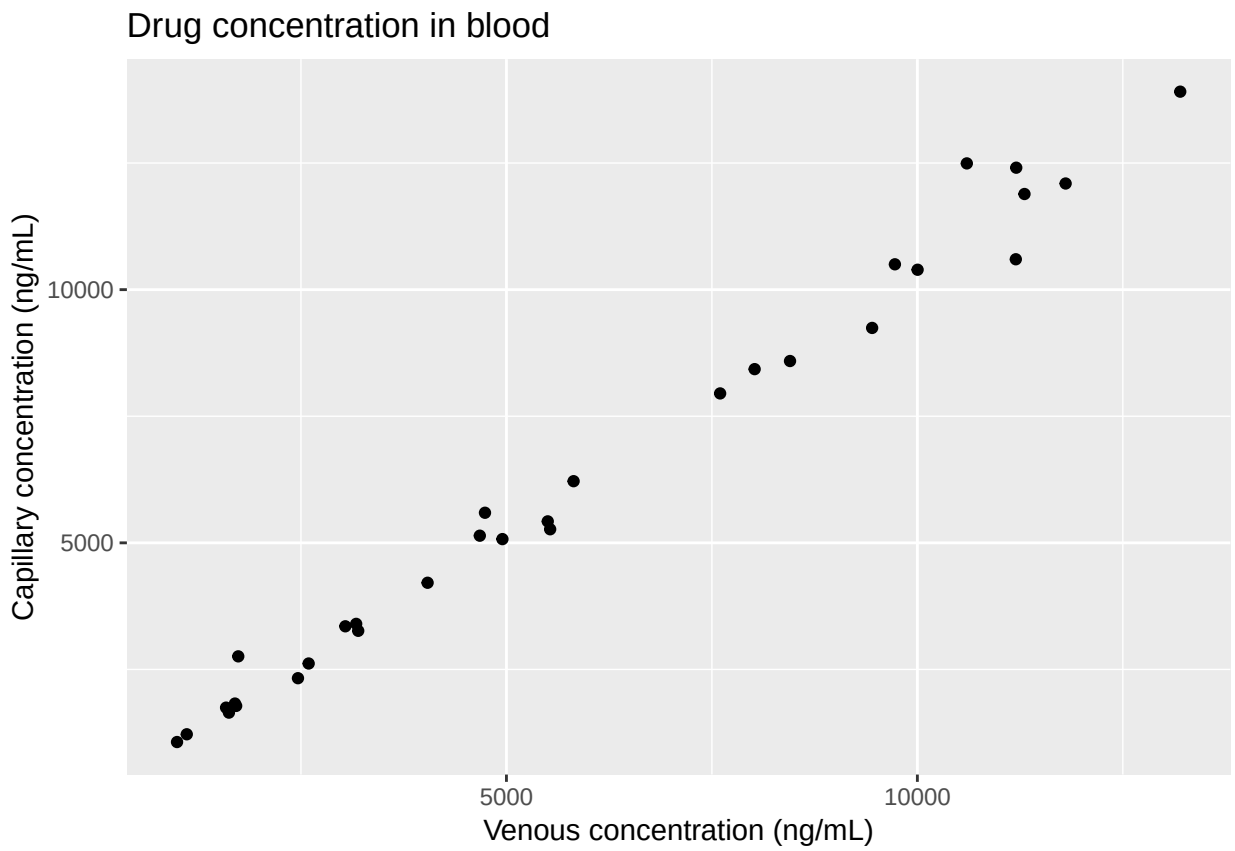
If you believe that only two of the BINS assumptions are met, you can interpret as: The confidence interval we obtained for the proportion of the time Sophie is correct in the long run is (0.277, 0.848). However, because the assumptions necessary for the inference to be valid are not met, this interval has no real-world meaning. Sophie may be learning to not pick a hand entirely at random, and we do not have an inferential framework to account for this.

Problem 4

Researchers would like to predict the concentration of lumefantrine in capillary plasma (response variable) from the concentration of lumefantrine in venous plasma (predictor variable). When researchers make a scatterplot of the data for the 31 women, they notice that the points have an approximately linear relationship and that the slope of the regression line is approximately 1.

Suppose that a new pregnant woman is given a dose of lumefantrine and, two hours later, the concentration in her venous plasma is measured to be 1700 ng/mL. Predict with 95% confidence the concentration of lumefantrine in her capillary plasma. Do you have faith in this method of prediction, or are assumptions necessary for this prediction violated?

```
lf_scatter <- ggplot(lumefantrine, mapping = aes(x = Vc_2h, y = Cc_2h)) +
  geom_jitter()  # I prefer jittered scatterplots but geom_point() works too
lf_scatter_labeled <- lf_scatter +
  labs(x = "Venous concentration (ng/mL)", y = "Capillary concentration (ng/mL)",
       title = "Drug concentration in blood")
print(lf_scatter_labeled)
```



A linear model certainly seems appropriate based on the scatterplot and the information in the problem.

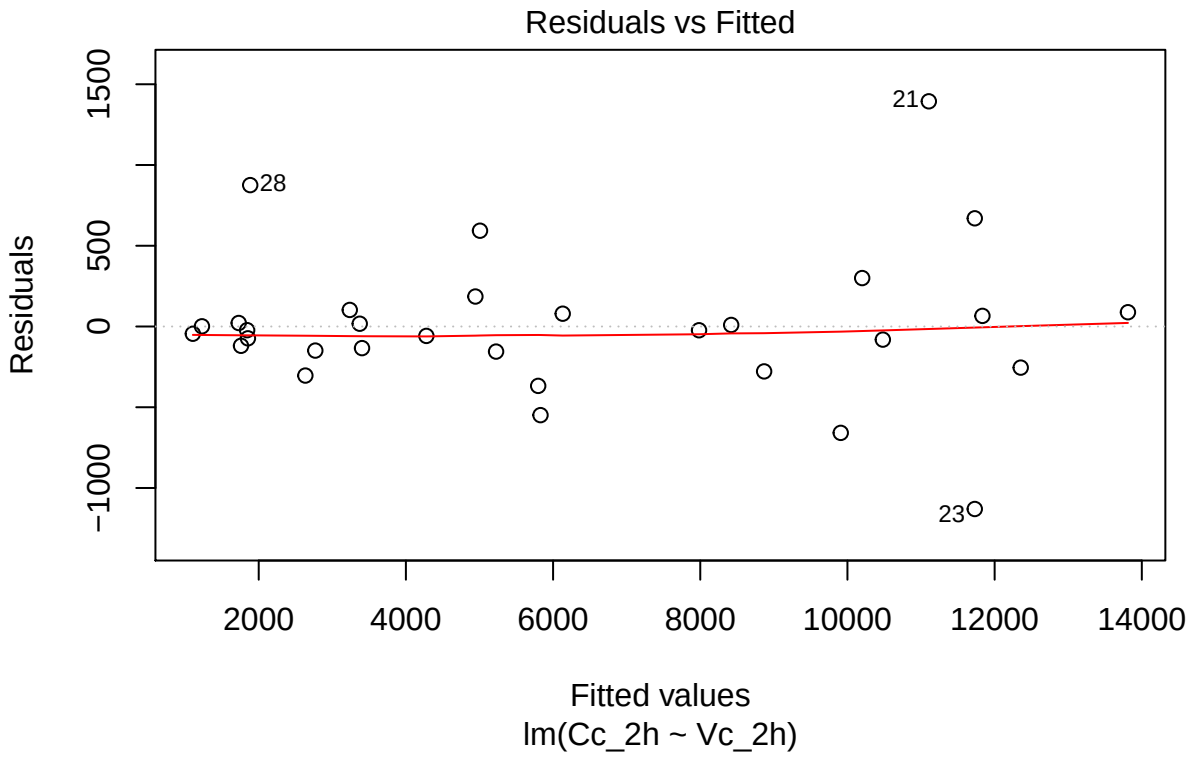
We fit the model:

```
lm_lf <- lm(Cc_2h ~ Vc_2h, data = lumefantrine)
summary(lm_lf)
```

```
##
## Call:
## lm(formula = Cc_2h ~ Vc_2h, data = lumefantrine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1130.12  -152.23   -23.36    83.77  1394.34
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  73.54202  153.12937    0.48    0.635
## Vc_2h         1.04077    0.02187   47.60   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 460.3 on 29 degrees of freedom
## Multiple R-squared:  0.9874, Adjusted R-squared:  0.9869
## F-statistic:  2265 on 1 and 29 DF,  p-value: < 2.2e-16
```

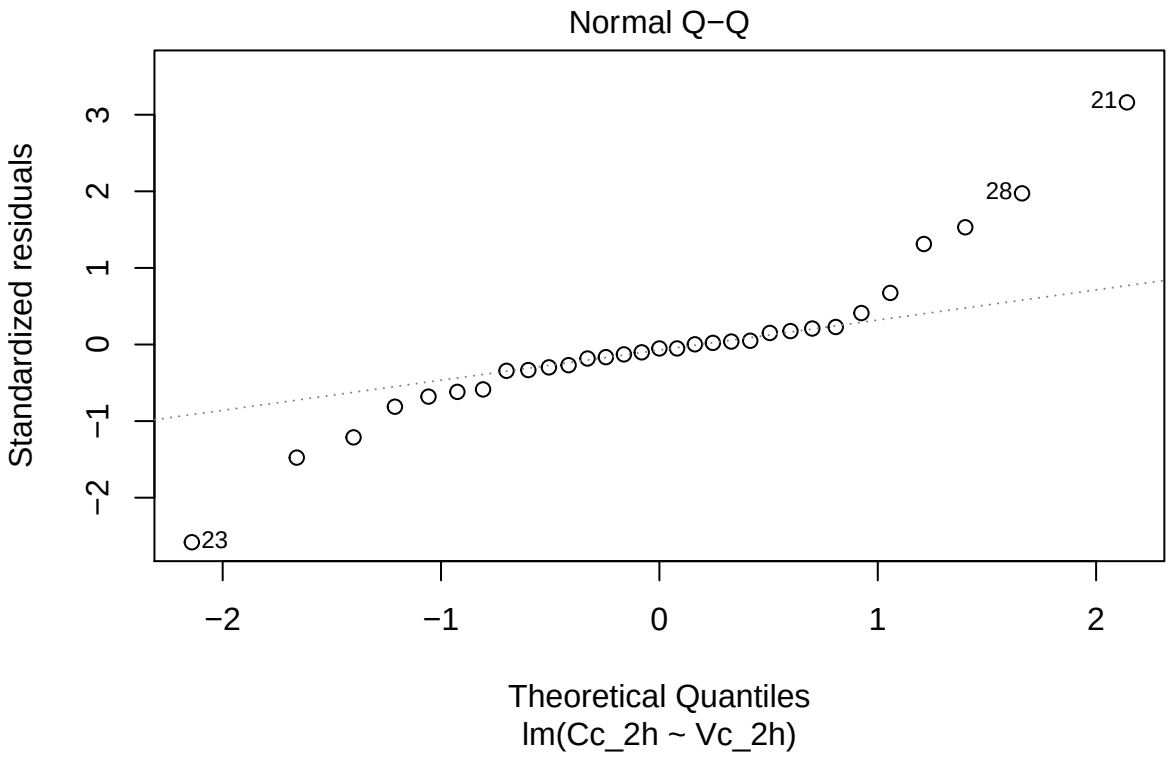
We have a very good R^2 value of 0.9874 and the residuals have a median of -23.36, which is fairly close to 0 considering the scale we are dealing with. Let's see if we see any issues with constant variance or normality:

```
plot(lm_lf, which = 1)
```



Constant variance seems to be a bit of a problem; the residuals look like they tend to be larger at high fitted values.

```
plot(lm_lf, which = 2)
```



Normality is definitely an issue.

The problem asks us to get a 95% prediction interval anyway:

```
lf1 <- data.frame(Vc_2h = 1700)
predict(lm_lf, newdata = lf1, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 1842.844 868.1114 2817.577
```

If the assumptions were anywhere close to met, we would claim with 95% confidence that the concentration of lumefantrine in the woman's capillaries was between 868 and 2818 ng/mL. However, our assumptions of constant variance and normality of the residuals are not met, and so we have little faith in our prediction.