PM592: Regression Analysis for Data Science Name: Flemming Wu

HW2

Bivariate relationships, simple linear regression

Instructions

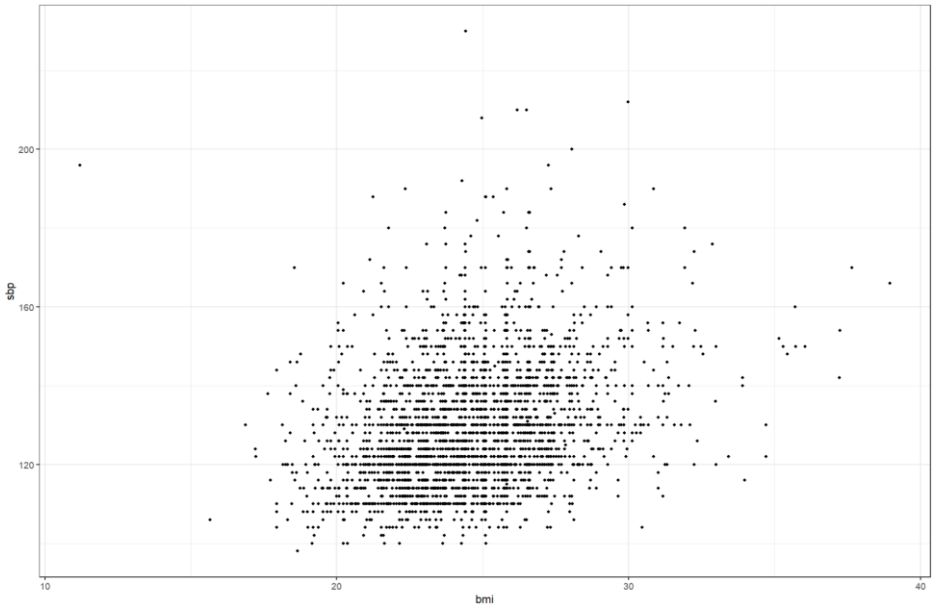
- Answer questions directly within this document.
- Upload to Blackboard by the due date & time.
- Clearly indicate your answers to all questions.
- If a question requires analysis, attach all relevant output to this document in the appropriate area. Do not attach superfluous output.
- There are 4 questions and 30 points possible.

Question 1 [7 points]

Use the WCGS data you saved from HW1.

1a. [1 point]. Provide a scatter plot of the relationship between SBP (Y) and BMI (X). Does anything in this scatter plot give you concern?

```
> ggplot(data=wcgs, aes(x=bmi, y=sbp)) +
+   geom_point(size=1) +
+   theme_bw()
```



There is an outlier in the scatterplot, representing an individual with a BMI of approximately 11. This individual's BMI is much lower than the rest of the sample. Additionally, there are a few individuals with very high SBP levels, above 200.

1b. [1 point] Provide the null and alternative hypotheses that would test whether SBP is related to BMI.

Null hypothesis: the slope of the regression line of BMI and SBP is equal to 0

Alternative hypothesis: the slope of the regression line of BMI and SBP is not equal to 0

1c. [1 point] Provide the linear regression equation for the relationship of SBP regressed on mean-centered BMI.

```
> wcgs <-
+   wcgs %>%
+   mutate(bmi_c = bmi - mean(bmi))
> model1c <- lm(sbp ~ bmi_c, data=wcgs)
> summary(model1c)

Call:
```

```
lm(formula = sbp ~ bmi_c, data = wcgs)

Residuals:
    Min      1Q  Median      3Q     Max
-34.707  -9.782  -2.410   7.382 101.551

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 128.6328     0.2578  498.88   <2e-16 ***
bmi_c         1.6945     0.1004   16.87   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.48 on 3152 degrees of freedom
Multiple R-squared:  0.08282,    Adjusted R-squared:  0.08253
F-statistic: 284.6 on 1 and 3152 DF,  p-value: < 2.2e-16
```

$\hat{Y} = 128.63 + 1.69X$

1d. [2 points] Based on your regression results, what are your decisions for the hypotheses in (1b)? Provide the test statistic and p-value.

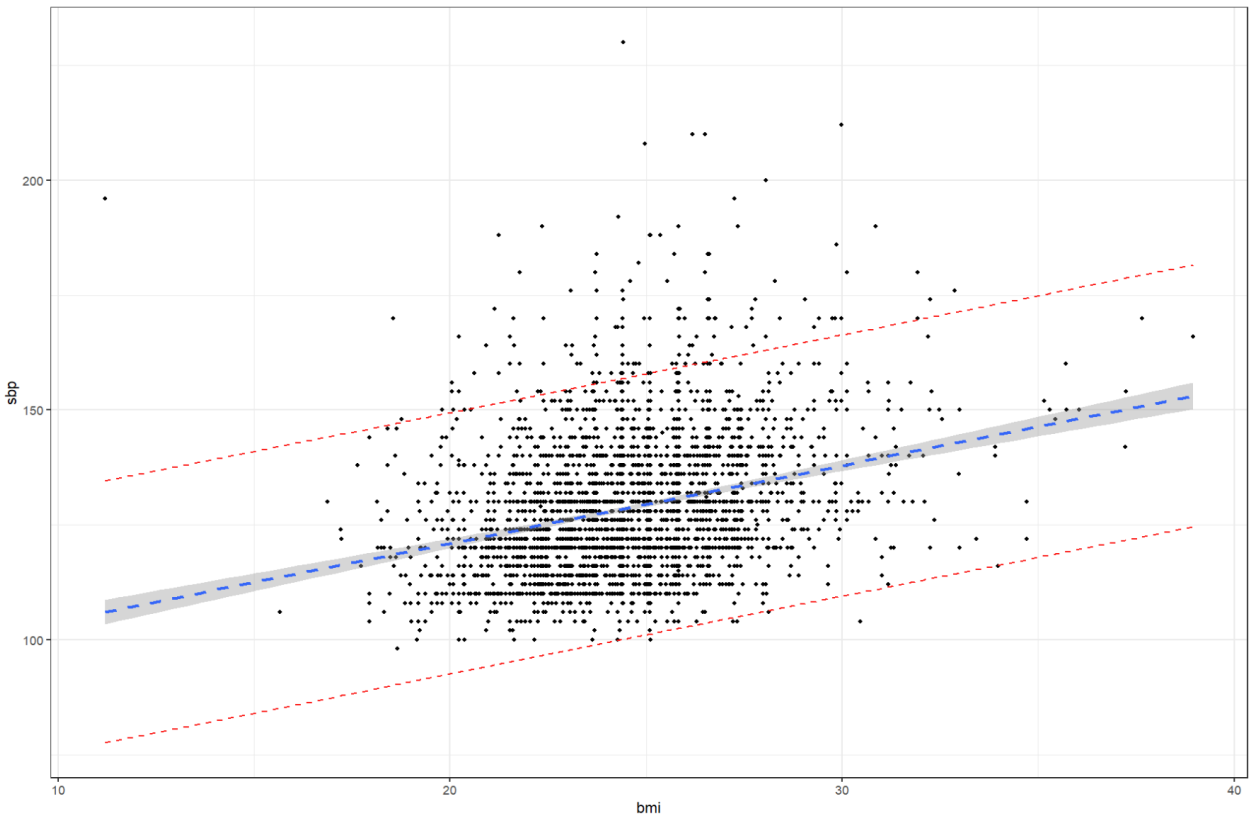
Based on my regression results, there is enough evidence to reject the null hypothesis ($p << 0.001$), and accept the alternative hypothesis that the slope of the line is not equal to 0, or in other words that there is a relationship between BMI and SBP.

1e. [1 point] How much is SBP expected to change if BMI increases by 5 units?

SBP is expected to increase by 5 x 1.69 = 8.45 units if BMI increases by 5 units.

1f. [1 point] Add the confidence interval and prediction interval to your plot in (1a). What is the interpretation of each of these intervals?

```
> cbind(wcgs, predict(model1c, wcgs, interval="prediction")) %>%
+   ggplot(aes(x=bmi, y=sbp)) +
+   geom_point(size=1) +
+   theme_bw() +
+   geom_smooth(formula=y ~ x, method="lm", linetype="dashed") +
+   geom_line(aes(y=lwr), color="red", linetype="dashed") +
+   geom_line(aes(y=upr), color="red", linetype="dashed")
```



The confidence interval, given by the gray region around the blue dotted line, represents some reasonable values for which the regression line of BMI vs SBP could also be. On the other hand, the prediction interval, given by the red lines, represent an interval in which most individual values are expected to be.

Question 2 [8 points]

Continue to use your WCGS file.

2a. [3 points] Fill in the table below by running 3 separate regressions of SBP on NCIGS (# cigarettes smoked per day). For each regression, provide the estimate of the intercept (& 95% CI), the estimate of the slope (& 95% CI), and the R^2 value.

Model	Predictor	Intercept (95% CI)	Slope (95% CI)	R^2
1	NCIGS	128.27 (127.56, 128.95)	0.031 (-0.0051, 0.068)	0.00090
2	(NCIGS – 10)	128.58 (128.05, 129.11)	0.031 (-0.0051, 0.068)	0.00090
3	(NCIGS – mean(NCIGS))	128.63 (128.11, 129.16)	0.031 (-0.0051, 0.068)	0.00090

2b. [2 points] For each of the 3 models, interpret the estimates of the intercept in a meaningful way.

In model 1, the intercept gives the estimated SBP for someone who smokes 0 cigarettes per day, which is about 128.27.

Model 2's intercept gives the estimated SBP for someone who smokes 10 cigarettes per day, which is about 128.58.

Model 3's intercept is centered on the mean of X, so the intercept can be interpreted as the estimated SBP for someone who smokes the mean number of cigarettes per day, which is about 128.63.

2c. [1 point] Comment on the respective widths of the 95% confidence intervals.

The widths of the 95% confidence intervals for the intercepts of models 1,2 and 3 are 1.39, 1.06, 1.05 respectively. The widths of the confidence interval is highest for model 1, and model 2 and 3 are almost the same. However, from just looking at the lower and higher confidence intervals for the intercepts, the widths all look very similar.

2d. [1 point] Comment on the slopes and explain why they are similar or different.

The slopes for all three models are the same. This is because the X value's transformation in models 2 and 3 was a subtraction, which causes the values to shift along the x-axis. The shifting changes the intercept but not the slope of the regression line (the relationship between the dependent and independent variables doesn't change). If the x-values were scaled (i.e. multiplied or divided by a value), that would have caused the slope of the regression line to change.

2e. [1 point] Use the predict() function to obtain the estimate of SBP for somebody who smokes 19 cigarettes per day.

```
> predict(model2a.1, newdata=data.frame(ncigs=19))
       1
128.8638
```

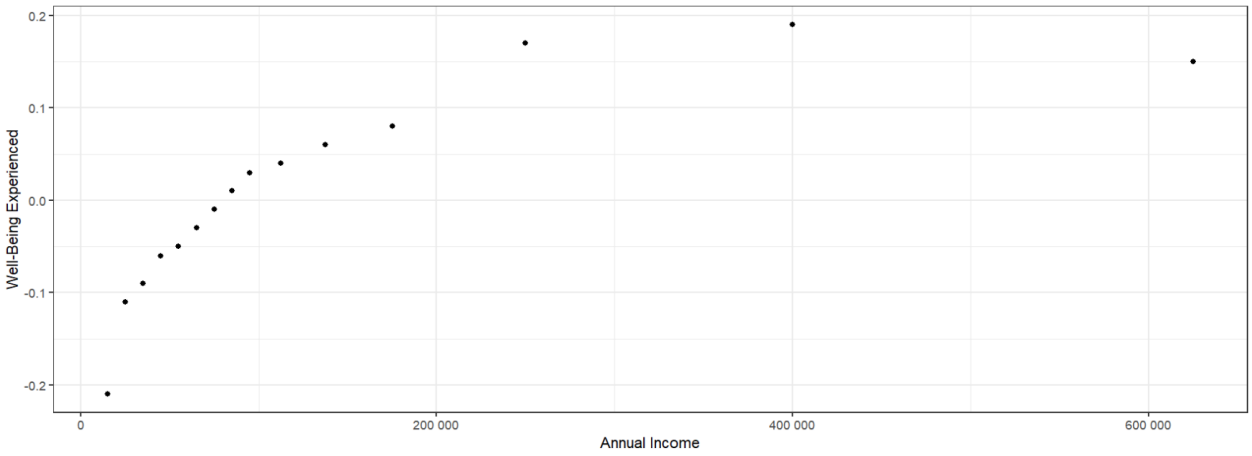
The estimated SBP for somebody who smokes 19 cigarettes per day is 128.86

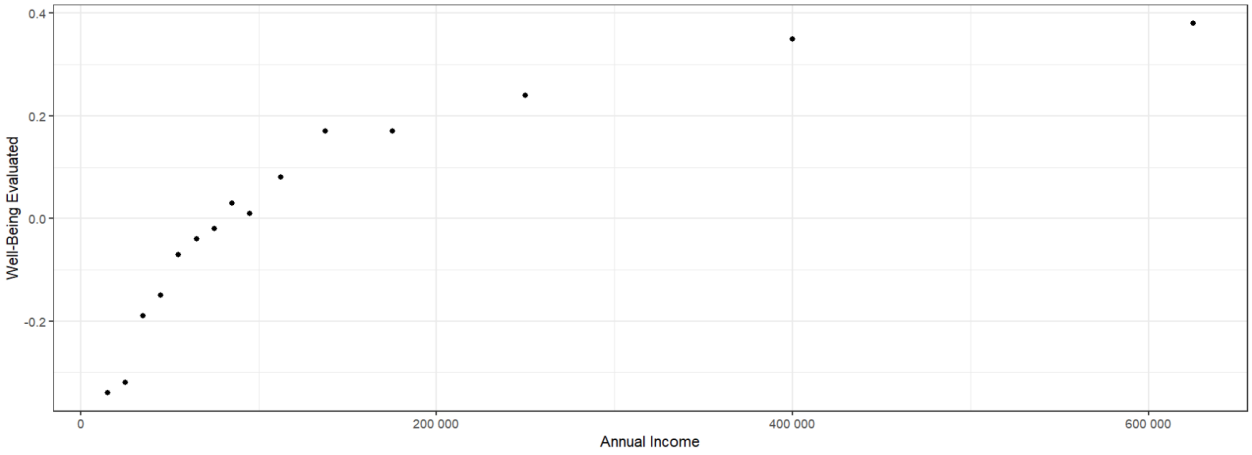
Question 3 [8 points]

Read the following web page:
https://www.visualcapitalist.com/chart-money-can-buy-happiness-after-all/

3a. [2 points] Download or manually enter in the data from the section "The Results". You should have 3 variables: annual income, well-being (experienced), and well-being (evaluative). Produce a scatter plot of each well-being variable vs. annual income.

```
> url <- 'https://www.visualcapitalist.com/chart-money-can-buy-happiness-after-all/'
>
> table_nodes <-
+   read_html(url) %>%
+   html_elements("body") %>%
+   html_nodes("table")
>
> table <-
+   table_nodes[1] %>%
+   html_table() %>%
+   as.data.frame()
>
> names(table) <- c("annual_income", "well_being_experienced",
"well_being_evaluative")
>
> table <-
+   table %>%
+   mutate(annual_income_numeric = as.numeric(gsub("[\\$,]", "", annual_income)))
>
> p1 <- ggplot(table, aes(x=annual_income_numeric, y=well_being_experienced)) +
+   geom_point() +
+   theme_bw() +
+   labs(x="Annual Income", y="Well-Being Experienced") +
+   scale_x_continuous(labels = scales::number_format(scale = 1, accuracy = 1))
> p2 <- ggplot(table, aes(x=annual_income_numeric, y=well_being_evaluative)) +
+   geom_point() +
+   theme_bw() +
+   labs(x="Annual Income", y="Well-Being Evaluated") +
+   scale_x_continuous(labels = scales::number_format(scale = 1, accuracy = 1))
> grid.arrange(p1, p2)
```

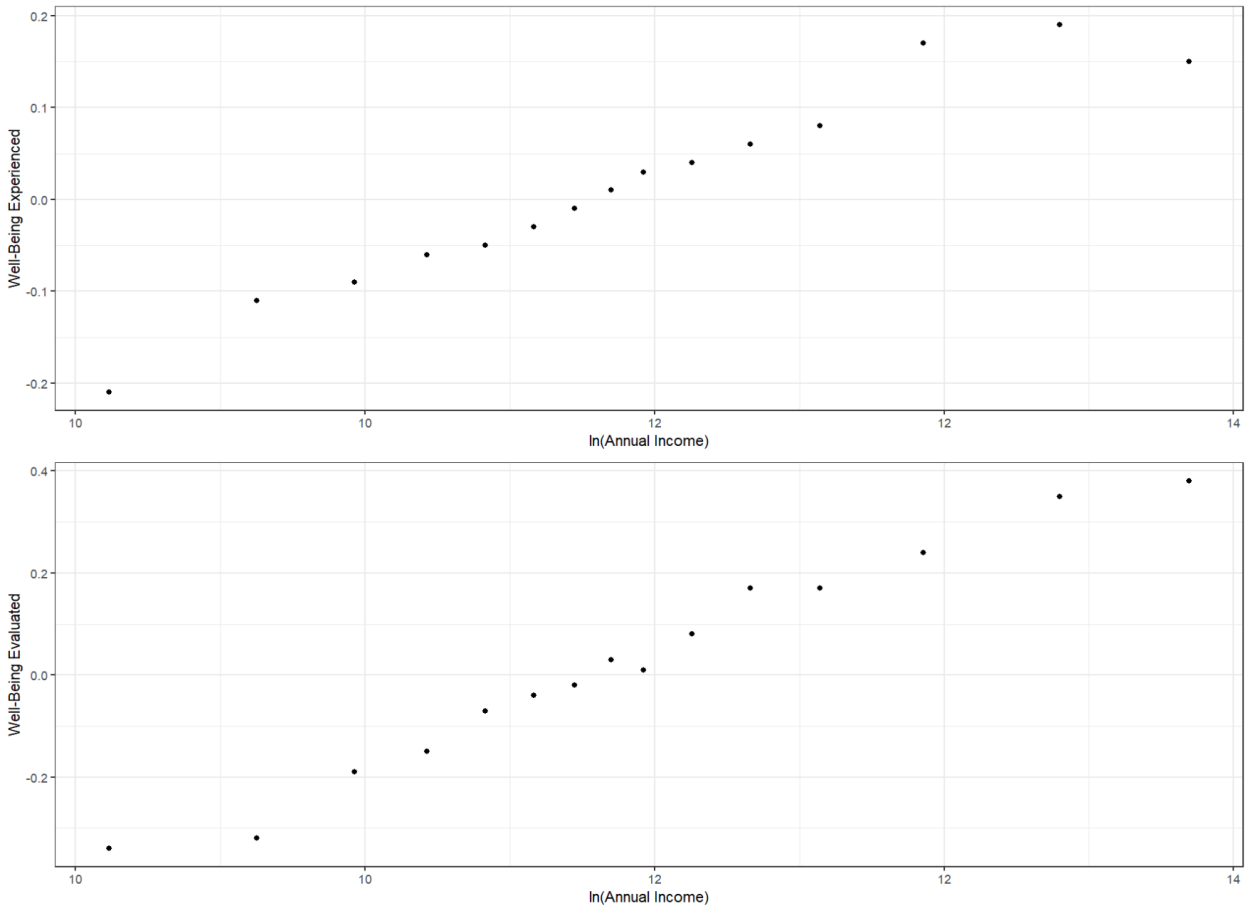




3b. [2 points] Does well-being appear linearly related to income? The website reports "a recent study found that happiness increases linearly with reported income (logarithmic)." Does performing this transformation appear to better satisfy the linearity assumption?

Well-being does not appear to be linearly related to income.

```
> p1 <- ggplot(table, aes(x=annual_income_log, y=well_being_experienced)) +
+   geom_point() +
+   theme_bw() +
+   labs(x="ln(Annual Income)", y="Well-Being Experienced") +
+   scale_x_continuous(labels = scales::number_format(scale = 1, accuracy = 1))
> p2 <- ggplot(table, aes(x=annual_income_log, y=well_being_evaluative)) +
+   geom_point() +
+   theme_bw() +
+   labs(x="ln(Annual Income)", y="Well-Being Evaluated") +
+   scale_x_continuous(labels = scales::number_format(scale = 1, accuracy = 1))
> grid.arrange(p1, p2)
```

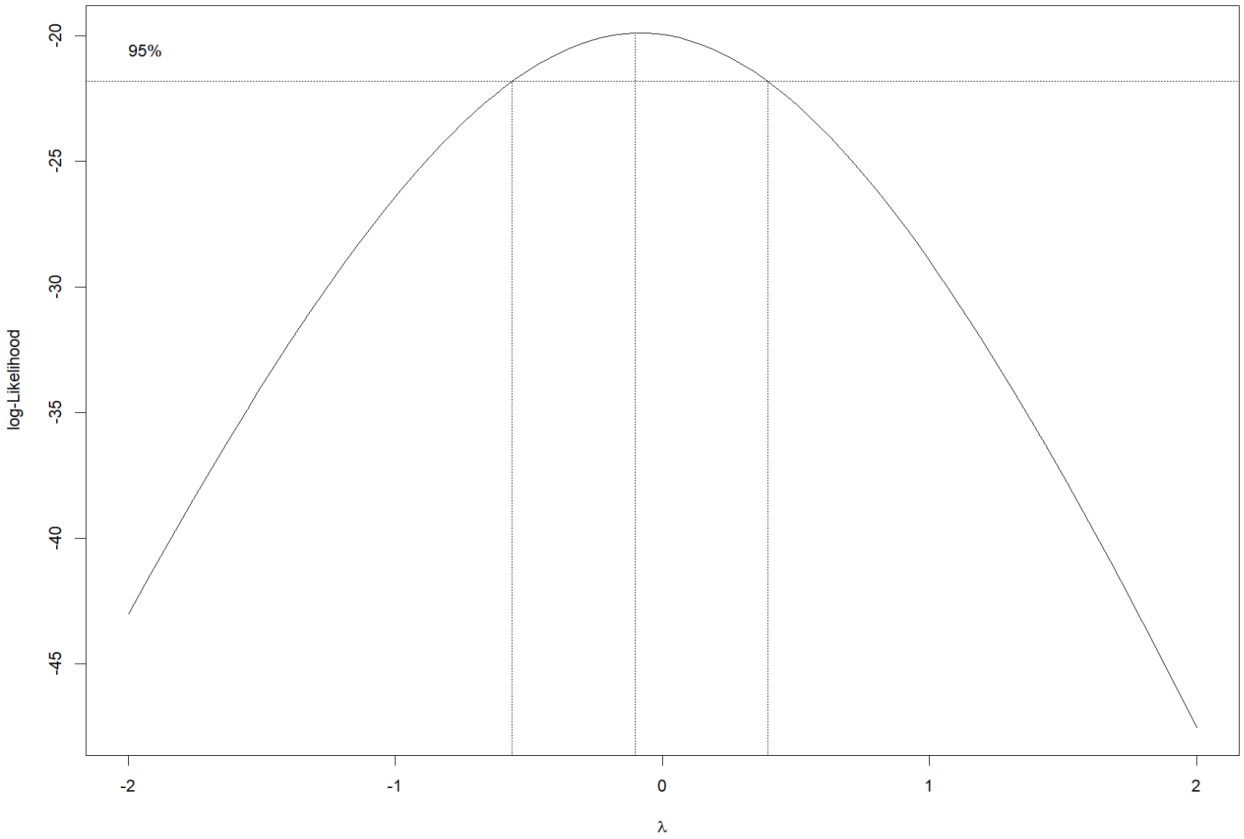


Taking the log of annual income, it appears that the transformation better satisfies the linear assumption.

3c. [2 points] The website reports the well-being variables are "measured in standard deviations from the mean". What is another name for this type of variable? Why do you think the authors did this transformation?

Another name for a type of variable that is measured in standard deviations from the mean is a standardized score (z score). This is likely done to make the well-being variables easier to interpret, which would be difficult without standardization since the scores assigned are arbitrary.

3d. [2 points] Report what you believe is the best regression relationship between each well-being variable and income (however you choose to transform it, if at all).



From the boxcox estimation of the best transformation parameter for annual income, it looks like the peak of the graph is close to lambda = 0, which corresponds to a logarithmic transformation of X. Given this, and the scatterplots above, I believe a log transformation of income vs well-being is the best linear regression relationship.

Question 4 [8 points]

Use the article by Sarafidis et al. on Blackboard to answer the following questions.

4a. [2points] Write a brief summary describing the hypotheses being tested, how the sample was obtained, the sample sizes in each group, and the statistical procedures used to analyze the data.

The article by Sarafidis et al tests two hypotheses: the first being whether Clara cell protein (CC16) levels are associated with changes in alveolar integrity that occur with different ventilary modes (HFOV vs SIMV). The second hypothesis being tested is whether Interleukin-6 (IL)-6 levels are associated with CC16 levels.

The criteria for selecting samples was that they were inborn, preterm neonate (premature baby) with a gestational age of less than 30 weeks and respiratory failure which required artificial ventilation within the first 2 hours of life. After several exclusion criteria were evaluated, there were 24 preterm neonates included in the statistical analysis. 12 were in the SMIV group, and 12 in the HFOV group.

Statistical procedures used to analyze the data include: a KS test to assess the normality of the data, Dunnett's multiple comparisons test to test the effect of time and ventilation mode on IL-6 and CC16 levels, Fisher's exact test to test independence between categorical variables, and Pearson's correlation to describe relationships between continuous variables.

4b. [2 points] It is not explicitly stated, but which statistical test do you think the authors used to obtain the p-values in Table 1?

The p-values in Table 1 are for testing for difference in means between the two groups in the study for a variety of variable such as birth weight, number of c-sections, number of males, etc. I believe the authors used a two-sided t-test between the two groups to test for differences in means, and the p-values represent evidence against the null hypothesis that the means of these variables are not different between the two groups.

4c. [2 points] Write a small R program that can input a mean, standard deviation, and sample size for each of two groups and compute this test comparing the two groups. Can you re-create the p-values obtained by the authors using this method?

```
> my.t.test <- function(mean1, mean2, sd1, sd2, n1, n2) {
+   t <- (mean1 - mean2) / (sqrt((((sd1)**2)/n1) + ((sd2)**2)/n2))
+   return(2*pt(abs(t), df=n1+n2-2, lower.tail = F))
+ }
>
> my.t.test(909, 959, 217, 182, 12, 12)
[1] 0.547098
> my.t.test(27, 27.1, 1.8, 1.5, 12, 12)
[1] 0.8838123
> my.t.test(4.8, 5.1, 1.4, 1.8, 12, 12)
[1] 0.6530486
> my.t.test(7.3, 7.3, 1.2, 1, 12, 12)
[1] 1
```

My R program takes the means, standard deviations, and standard deviations and outputs p-values based on the t-distribution. I used the formula:

$$t = \frac{(x_1 - x_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

to get the t-statistic. I then inputted this into the `pt` function in R to get the p-value and ensured to set lower.tail to false, and multiplied by two to get a two-sided p-value.

I ran the program on the variables in the paper for which mean and standard deviation were provided: birth weight, gestational age, apgar score (1 min), and apgar score (5 min). The p-values I got for the first three comparisons are very similar to the ones provided in the table, the last p-value (1) I got for testing difference in mean apgar score (5 min) was higher than the p-value (0.606) provided in the table

4d. [1 point] Did the authors report whether they tested any of the assumptions of linear regression?

Yes, they reported testing the normality assumption using the KS test.