# B201A HW3

Lillian Chen

12/1/2021

1. Consider further the data from Homework 2, Problem 3.

```
valiumdata <- data.frame(S = c(32,42,52,61,62,65,66),
                         C = c(6.6,7.4,8.8,9.7,10.5,11.8,10.7))
kbl(valiumdata, booktabs = T,
    col.names = c("Sedation Score S", "Cortisol C, ug/dl")) %>%
  kable_styling(position = "center",
                latex_options = c("basic", "hold_position"))
```

| Sedation Score S | Cortisol C, ug/dl |
|---|---|
| 32 | 6.6 |
| 42 | 7.4 |
| 52 | 8.8 |
| 61 | 9.7 |
| 62 | 10.5 |
| 65 | 11.8 |
| 66 | 10.7 |

1.(a) For the regression of $S$ on $C$, calculate the least-squares estimate of the intercept, slope, and residual variance.

Solution: The least-squares estimates of the intercept and slope, respectively, are $\beta_0 = -8.19$ and $\beta_1 = 6.68$. The residual variance is the square of the residual standard error, which is $\sigma^2 = (3.903)^2 = 15.2$

```
lm <- lm(S ~ C, data=valiumdata)
summary(lm)
```

```
##
## Call:
## lm(formula = S ~ C, data = valiumdata)
##
## Residuals:
##        1        2        3        4        5        6        7
## -3.87780  0.78105  1.43402  4.42522  0.08406 -5.59532  2.74877
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -8.1867     8.1161  -1.009 0.359410
## C             6.6764     0.8529   7.828 0.000546 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.903 on 5 degrees of freedom
## Multiple R-squared:  0.9246, Adjusted R-squared:  0.9095
## F-statistic: 61.27 on 1 and 5 DF,  p-value: 0.0005458
```

```
(residual_sigma <-summary(lm)$sigma)
```

```
## [1] 3.902736
```

```
(residual_var<- residual_sigma^2)
```

```
## [1] 15.23135
```

1.(b) Calculate the standard error of the slope and the standard error of the intercept, and comment on whether each of these parameters is significantly different from 0.

Solution: The standard error of the intercept is 8.12, and the standard error of the slope is 0.853. We test whether each of these standard errors is significantly different from 0 by running a two-sided t-test ($t = \beta_j / SE(\beta_j)$ for $j = 0, 1$) The intercept for the regression of S on C is not significantly different from 0 at the $\alpha = 0.05$ level ($p = .359$). The slope for the regression of S on C is significantly different from 0 at the $\alpha = 0.05$ level ($p < .001$).

```
summary(lm)
```

```
##
## Call:
## lm(formula = S ~ C, data = valiumdata)
##
## Residuals:
##        1        2        3        4        5        6        7
## -3.87780  0.78105  1.43402  4.42522  0.08406 -5.59532  2.74877
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -8.1867     8.1161  -1.009 0.359410
## C             6.6764     0.8529   7.828 0.000546 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.903 on 5 degrees of freedom
## Multiple R-squared:  0.9246, Adjusted R-squared:  0.9095
## F-statistic: 61.27 on 1 and 5 DF,  p-value: 0.0005458
```

1.(c) Comment on what the findings from 1.(a) and 1.(b) tell you about whether the correlation is significantly different from 0.

Solution: The standard error of the slope in part b) has a significance at the $\alpha = 0.05$ level, indicating that the slope for the regression of S on C is significantly different from 0 and that there is a positive relationship for every 1-unit change in the predictor $S$. This implies that the correlation $\rho$ should also be positive and significantly different from 0.

1.(d) What is the proportion of variation in $S$ explained by $C$?

Solution: The proportion of variation in $S$ explained by $C$ is equivalent to the explained variance, which is also known as $R^2$. From the regression of S on C, we see that $R^2 = 0.925$, meaning that 92.5% of the variance in S is explained by C.

```
summary(lm)$r.squared
```

```
## [1] 0.9245546
```

1.(e) Provide a 95% confidence interval for the expected sedation score associated with a cortisol level of 10.0.

Solution: The 95% confidence interval for an individual who has a cortisol level of 10.0 is [54.5, 62.6].

```
newCe <- data.frame(C=10.0)

# calculate confidence interval
(predinte <- predict(lm, newdata = newCe, interval = 'confidence', level = 0.95))
```

```
##        fit      lwr      upr
## 1 58.57772 54.53238 62.62305
```

```
# width of confidence interval
predinte[3]-predinte[2]
```

```
## [1] 8.090677
```

1.(f) Provide a 95% confidence interval for the expected sedation score associated with a cortisol level of 5.0, and explain in a sentence why the width of the interval is different from the width of the interval in part 1.(e).

Solution: The 95% confidence interval for an individual who has a cortisol level of 5.0 is [14.9, 35.5], with a width of 20.6. The width of this confidence interval is larger than the width of the 95% confidence interval for a cortisol level of 10.0 (8.1) as calculated in part 1.(e) because the data point of 5.0 falls far from and outside the range of observed cortisol levels, which leads to wider confidence intervals; the data point of 10.0 is much closer to the mean of the observed cortisol levels, leading to a narrower confidence interval.

```
newCf <- data.frame(C=5.0)

# calculate confidence interval
(predintf <- predict(lm, newdata = newCf, interval = 'confidence', level = 0.95))
```

```
##        fit      lwr     upr
## 1 25.19548 14.91737 35.4736
```

```
# width of confidence interval
predintf[3]-predintf[2]
```

```
## [1] 20.55622
```

1.(g) Provide a 95% prediction interval (i.e., a 95% confidence interval for a new individual observation) for an individual who has a cortisol level of 10.0, and explain in a sentence why the width of the interval is different from the width of the interval in part 1.(e).

Solution: The 95% prediction interval for an individual who has a cortisol level of 10.0 is [47.7, 69.4]. The width of this prediction interval is 21.6, as compared to the confidence interval of width 8.1 for a cortisol level of 10.0. The difference in interval width (where the prediction interval is larger than the confidence interval) can be attributed to the additional variability associated with variation of individual observations around the regression line for a prediction interval (while already including uncertainty of the predicted mean), whereas a confidence interval does not account for this variation and merely accounts for uncertainty in the predicted mean.

```
newCg <- data.frame(C=10.0)

# calculate prediction interval
(predintg <- predict(lm, newdata = newCg, interval = 'prediction', level = 0.95))
```

```
##        fit      lwr      upr
## 1 58.57772 47.76051 69.39492
```

```
# width of prediction interval
predintg[3]-predintg[2]
```

```
## [1] 21.6344
```

1.(h) Consider two new observations, one with a cortisol level of 5.0 and the other with a cortisol level of 10.0, and consider the corresponding prediction intervals obtained from fitting a linear regression model to the data above. Would you regard the intervals as equally likely to cover the observed sedation scores, or would you consider one of the intervals as more likely to cover the corresponding observed sedation score than the other? Explain your reasoning in a sentence.

Solution: The prediction interval corresponding to the cortisol level of 5.0 is more likely to cover the corresponding observed sedation score as opposed to the other option, because it is extremely wide due to the value of cortisol level falling outside the range of observed cortisol levels – intervals get wider the farther they are from the mean of the predictor.

```
newCh <- data.frame(C=c(5.0,10.0))
predict(lm, newdata = newCh, interval = 'prediction', level = 0.95)
```

```
##        fit      lwr      upr
## 1 25.19548 10.83280 39.55817
## 2 58.57772 47.76051 69.39492
```

2. Suppose you are interested to compare whether there is a significant difference in the average number of hours of sleep per week night obtained by male and female UCLA undergraduates. Suppose the standard deviation of the number of hours of sleep per week night is 1 hour in both groups, and you would regard the result as scientifically interesting if there was a difference on average of as much as half an hour per week night. Assume that plans call for a study with the same number of male and female undergraduates and that it is desired to have at least 80% power for a test at the $\alpha = 0.05$ level. Drawing on tabled information in your course reader or another similar source, how many subjects per group would be needed?

Solution:

$H_0 : \mu_{male} - \mu_{female} < 0.5$ and $H_A : \mu_{male} - \mu_{female} \geq 0.5$

Using the two-sample t test power calculation, I anticipate a minimum of 64 subjects per group would be necessary to achieve 80% power at the $\alpha = 0.05$ level for this test.

```
power.t.test(n = NULL, delta = 0.5, sd = 1, sig.level = 0.05, power = 0.80,
             type = "two.sample", alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 63.76576
##           delta = 0.5
##              sd = 1
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

3. Suppose you are asked to compare the proportions of male and female UCLA undergraduates who are able to continue for as long as 15 minutes on a treadmill stress test following a protocol where the slope of the incline and the speed of the treadmill increase in stages every 3 minutes. Suppose also you are told it would be scientifically interesting if the difference in the proportions was as much as 10%. Assume further that you do not have a good estimate in advance of the underlying proportions in the two groups but that regardless of the values of the underlying proportions, you would like to have at least 80% power to find a difference of as much as 10% in the proportions using a test at the $\alpha = 0.05$. Drawing on tabled information in your course reader or another similar source, what would be a good total number of subjects to include in the study? Explain your reasoning in a sentence or two.

$H_0 : p_{male} - p_{female} < 0.10$ and $H_A : p_{male} - p_{female} \geq 0.10$

Solution: We want to the most conservative number for total number of subjects to include in the study given that we want an expected difference of at least 10% at 80% power. The most conservative number of subjects will come from proportion values close to 0.5 since as we approach p = 0.5, sample proportions will maximize their variance $\hat{p}(1-\hat{p})$. Upon doing a search for sample size, we see that using the power.prop.test function we want a total of 784 subjects to include in the study, and using the unconditional approach for a two-sided two-proportion test we want 778 subjects to include in the study for a conservative estimate.

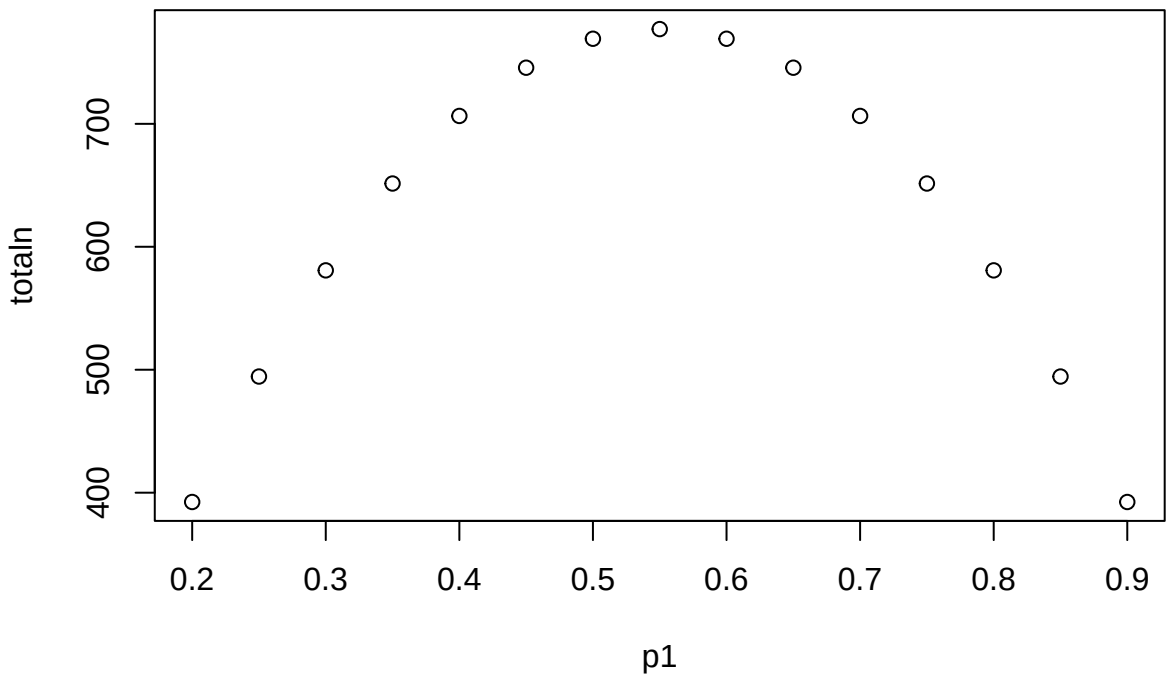
```
power.prop.test(n = NULL, p1 =0.55, p2 = 0.45, sig.level = 0.05, power = 0.80,
                alternative = "two.sided")
```

```
##
##      Two-sample comparison of proportions power calculation
##
##               n = 391.263
##              p1 = 0.55
##              p2 = 0.45
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
p1 <- seq(0.2, 0.9, 0.05)
p2 <- p1-0.1
d <- 0
totaln <- ( 2*(qnorm(0.80)+qnorm(0.975))^2 * (p1*(1-p1) + p2*(1-p2)) )/(p1 - p2 -d)^2
round(max(totaln),2)
```

```
## [1] 777.04
```

```
plot(p1, totaln)
```




4. When patients are unable to eat for long periods, they must be given intravenous nutrients, a process called parenteral nutrition. Unfortunately, patients on parenteral nutrition show increased calcium loss via their urine, sometimes losing more calcium than they are given in their intravenous fluids. Such a calcium loss might contribute to bone loss as the body pulls calcium out of bones to try to keep the calcium level in the blood within the normal range. In order to better understand the mechanisms of the calcium loss in urine, Lipkin and coworkers (American Journal of Clinical Nutrition, 1988; 47:515-523) measured urinary calcium UCa and related it to dietary calcium DCa , glomerular filtration rate Gfr (a measure of kidney function), urinary sodium UNa , and dietary protein level DP . The data are available for download from the web site for our text book: go to

https://people.vetmed.wsu.edu/slinkerb/appliedregression/

follow the link to "File downloads related to the book", scroll down to "Data for Problems (Appendix D)", and click on Table D-5 (where the order of the columns are as presented above: UCa, DCa, Gfr, UNa, and DP).

```
q4data <- read_csv(file = "hw3q4data.csv", show_col_types = F)
```

4.(a) Using a computer to facilitate your analysis, fit simple regression models of UCa on each of the other four variables. Is there evidence of significant correlation between UCa and any of the other variables? What can you tell about the relative importance of these four variables in determining UCa ?

Solution: There is evidence of significant correlation between UCa and each of the other four variables at the $\alpha = 0.05$ level. The p-values of the coefficients from each SLR are as follows: DCa ($p < .001$), Gfr ($p = .034$), UNa ($p = .009$), and DP ($p < .001$).

To determine the relative importance of these four variables in determining UCa, we compare the magnitudes of standardized coefficients so that we can compare coefficients on the scale of standard deviations (for every 1 standard deviation change in X, there is an associated change in Y of $\beta_{st.}$ standard deviations). Based on the standardized coefficients, the relative importance of the four variables from most important to least important is as follows: DCa ($\beta_{st.} = 0.76$), UNa ($\beta_{st.} = 0.63$), DP ($\beta_{st.} = 0.49$), and Gfr ($\beta_{st.} = 0.41$).

```
ucadca <- lm(UCa~DCa, data = q4data)
ucagfr <- lm(UCa~Gfr, data = q4data)
ucauna <- lm(UCa~UNa, data = q4data)
ucadp <- lm(UCa~DP, data = q4data)

summary(ucadca)
```

```
##
## Call:
## lm(formula = UCa ~ DCa, data = q4data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -51.79 -35.24  -9.75  37.93  70.29
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.74954   11.60949   2.132    0.043 *
## DCa          0.28848    0.04959   5.817 4.58e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.63 on 25 degrees of freedom
## Multiple R-squared:  0.5751, Adjusted R-squared:  0.5581
## F-statistic: 33.84 on 1 and 25 DF,  p-value: 4.581e-06
```

```
summary(ucagfr)
```

```
##
## Call:
## lm(formula = UCa ~ Gfr, data = q4data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -82.07 -30.91 -23.95  25.19 127.19
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.3530    33.5397   0.100   0.9212
## Gfr           1.4200     0.6313   2.249   0.0335 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 56.85 on 25 degrees of freedom
## Multiple R-squared:  0.1683, Adjusted R-squared:  0.135
## F-statistic: 5.059 on 1 and 25 DF,  p-value: 0.03355
```

```
summary(ucauna)
```

```
##
## Call:
## lm(formula = UCa ~ UNa, data = q4data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -76.157 -37.336  -8.164  27.809 131.379
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  46.4098    14.4264   3.217  0.00356 **
## UNa           0.7817     0.2756   2.836  0.00891 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.22 on 25 degrees of freedom
## Multiple R-squared:  0.2435, Adjusted R-squared:  0.2132
## F-statistic: 8.045 on 1 and 25 DF,  p-value: 0.008913
```

```
summary(ucadp)
```

```
##
## Call:
## lm(formula = UCa ~ DP, data = q4data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.246 -36.596  -6.052  36.051  92.553
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.0518    14.8557   1.821  0.08060 .
## DP            1.2999     0.3168   4.103  0.00038 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.19 on 25 degrees of freedom
## Multiple R-squared:  0.4024, Adjusted R-squared:  0.3785
## F-statistic: 16.83 on 1 and 25 DF,  p-value: 0.0003803
```

```
sy <- sd(q4data$UCa)
sx1 <- sd(q4data$DCa)
sx2 <- sd(q4data$Gfr)
sx3 <- sd(q4data$UNa)
sx4 <- sd(q4data$DP)
```

```
std_b1 <- summary(ucadca)$coefficients[2, 1] * sx1 / sy
std_b2 <- summary(ucagfr)$coefficients[2, 1] * sx2 / sy
std_b4 <- summary(ucauna)$coefficients[2, 1] * sx3 / sy
std_b3 <- summary(ucadp)$coefficients[2, 1] * sx4 / sy

stbetas <- data.frame(variable = c("DCa", "Gfr", "UNa", "DP"),
            beta_st = c(std_b1, std_b2, std_b3, std_b4))

kbl(stbetas, booktabs = T) %>%
  kable_styling(position = "center",
                latex_options = c("basic", "hold_position"))
```

| variable | beta_st |
|---|---|
| DCa | 0.7583619 |
| Gfr | 0.4102536 |
| UNa | 0.6343372 |
| DP | 0.4934168 |

4.(b) Carry out a multiple regression analysis of UCa on all four of the other variables. Comment on which variables seem to be important determinants of UCa. Also comment on similarities and differences with the findings from part 4.(a), and on whether you regard the separate simple regression results or the multiple regression result as more useful.

Solution: Upon fitting a multiple regression of UCa on all four other variables, we see that DCa remains statistically significant ($p < .001$) and UNa remains statistically significant ($p = 0.03$) at the $\alpha = 0.05$ level. The other two variables have p-values of a much larger magnitude and can be considered both statistically and practically insignificant.

Obvious differences in the findings here are that the coefficients for DP and Gfr are no longer statistically significant, which makes sense because they are included in a model with more variables that may better explain the variance in the outcome. Additionally, the coefficient for DP has changed signs from positive to negative. The magnitudes for coefficients of Gfr and UNa decreased from the results in 4a, while the magnitude for the coefficient of DCa increased from the results in 4a.

The order of coefficients arranged by ascending p-value aligns with the order of relative importance arranged by descending standardized coefficient magnitude, which is a similarity between the findings in 4b with 4a.

The multiple regression results are more useful because we can use this to both better inform our model and to remove unnecessary predictors that do not further explain the variance in the model. The multiple regression also has a higher $R^2$ attributed to the extra information included in the model so we gain knowledge using multiple covariates at a time.

There is high collinearity between DP and DCa, so ideally we would leave out DP since the coefficient for DP was not statistically significant. Additionally, there is some collinearity observed between Gfr and UNa. Refitting the model to have predictors DCa and UNa, we see that we get a model with both coefficients remaining statistically significant and maintain a similar $R^2$ value as the full multiple regression model.

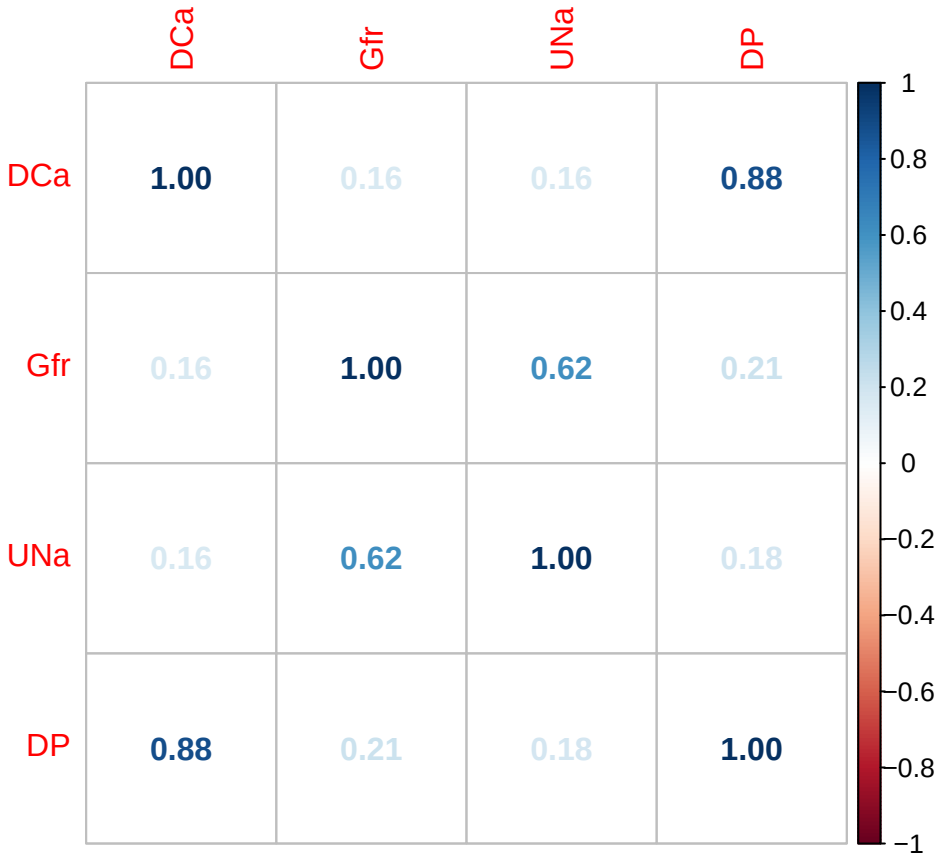
```
ucaall <- lm(UCa ~ ., data = q4data)

summary(ucaall)
```

```
##
## Call:
## lm(formula = UCa ~ ., data = q4data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.553 -25.722  -4.973  13.928  73.640
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.7316    21.8887  -0.262 0.795872
## DCa           0.3450     0.0889   3.881 0.000806 ***
## Gfr           0.4265     0.4872   0.876 0.390740
## UNa           0.4980     0.2213   2.250 0.034753 *
## DP           -0.5113     0.4831  -1.058 0.301391
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.19 on 22 degrees of freedom
## Multiple R-squared:  0.7352, Adjusted R-squared:  0.6871
## F-statistic: 15.27 on 4 and 22 DF,  p-value: 4.071e-06
```

```
corrplot::corrplot(cor(q4data[,2:5]), method = 'number')
```




```
ucaedit <- lm(UCa ~ DCa + UNa, data = q4data)
summary(ucaedit)
```

```
##
```

```
## Call:
## lm(formula = UCa ~ DCa + UNa, data = q4data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.012 -21.440  -4.127  13.281  74.461
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.12714   10.97747   0.649  0.52234
## DCa          0.26481    0.04202   6.302 1.63e-06 ***
## UNa          0.60078    0.17500   3.433  0.00217 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.96 on 24 degrees of freedom
## Multiple R-squared:  0.7151, Adjusted R-squared:  0.6913
## F-statistic: 30.11 on 2 and 24 DF,  p-value: 2.866e-07
```

5. When antibiotics are given to fight infections, they must be administered in such a way that the blood level of the antibiotic is high enough to kill the bacteria responsible for the infection. Because antibiotics are usually given periodically, the blood levels change over time, rising after an injection, then falling back down until the next injection. The interval between injections of recently introduced antibiotics has been determined by extrapolating from studies of older antibiotics. To update knowledge pertaining to dosing schedules, Vogelman and coworkers (Journal of Infectious Disease, 1988; 158: 831-847) studied the effect of different dosing intervals on the effectiveness of several newer antibiotics against a variety of bacteria in mice. One trial was the effectiveness of gentamicin against the bacterium Escherichia coli. As part of their assessment of the drug, they evaluated the effectiveness of gentamicin in killing E. coli as a function of the percentage of time the blood level of the drug remained above the effective level (the so-called mean inhibitory concentration M). Effectiveness was evaluated in terms of the number of bacterial colonies C that could be grown from the infected mice after treatment with a given dosing schedule (known as "colony-forming units", or CFU), where the lower the value of C, the more efficacious the antibiotic. The data are available for download from the web site for our text book: go to

https://people.vetmed.wsu.edu/slinkerb/appliedregression/

following the link to "File downloads related to the book", scrolling down to "Data for Problems (Appendix D)", and clicking on Table D-3. The first column, which can be labeled Y = "log CFU difference" and which can be treated as the outcome variable, is the difference between log (base 10) of the number of colony-forming units recorded after a period of antibiotic treatment and log (base 10) of the number of colony-forming units recorded before the period of antibiotic treatment. The second column, which can be labeled X1 = "Percentage of 24-hour period above the mean inhibitory concentration M". The third column can be labeled X2 = "Dose code" (where "0" refers to 1-4 hour intervals and "1" refers to 6-12 hour intervals)".

```
q5data <- read_csv(file = "hw3q5data.csv", show_col_types = F)
```

5.(a) Perform the regression of Y on X1 for the cases where Dose code = 0, and comment on whether X1 is a significant predictor of Y.

Solution: The regression of Y on X1 for the cases where Dose Code = 0 yields a coefficient estimate for X1 of -0.04, statistically significant at the $\alpha = 0.05$ level ($p < .001$). It appears that X1 is a significant predictor of Y.

```
yx1 <- lm(log_CFU~X1, data = q5data[q5data$dose_code == 0,])
summary(yx1)
```

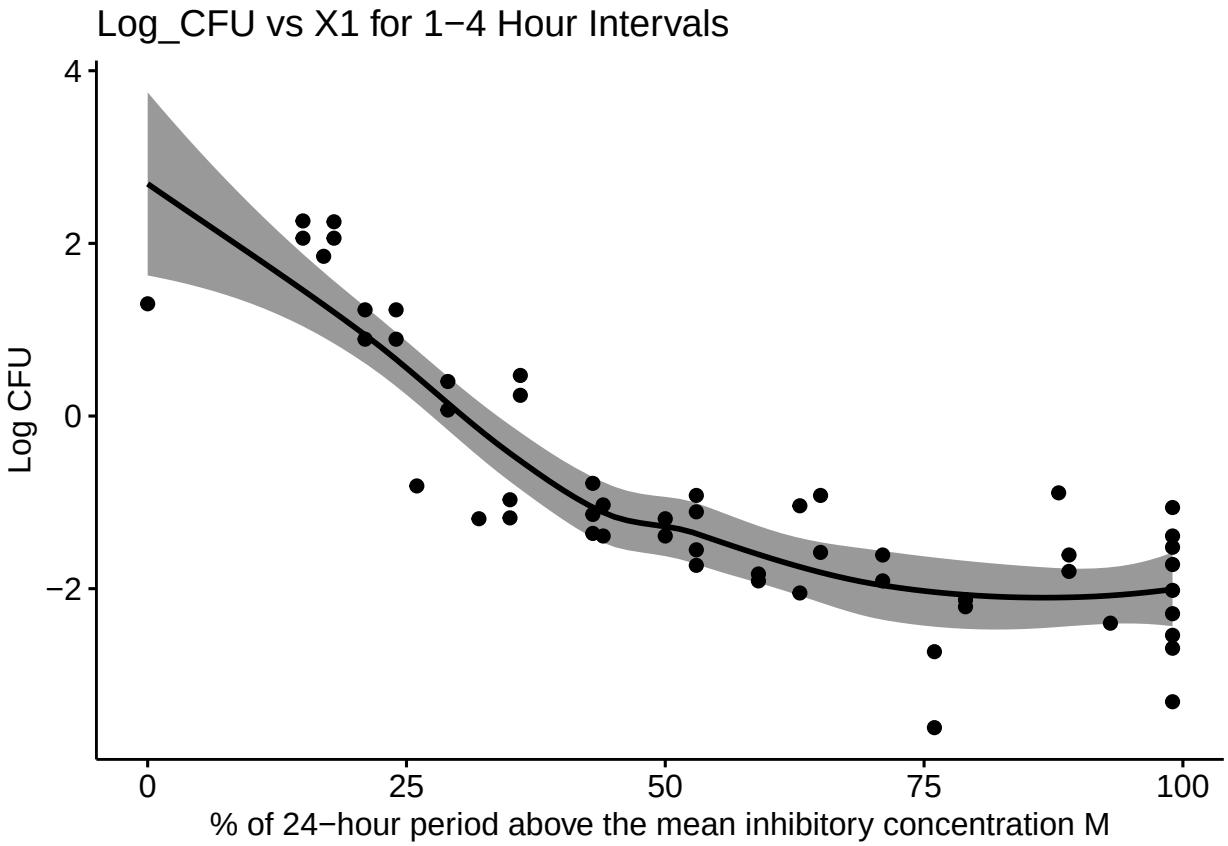
```
##
## Call:
## lm(formula = log_CFU ~ X1, data = q5data[q5data$dose_code ==
##     0, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.89878 -0.75228 -0.07804  0.60628  1.62080
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.355541   0.259993   5.214 3.25e-06 ***
## X1          -0.040352   0.004116  -9.803 2.06e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8705 on 52 degrees of freedom
## Multiple R-squared:  0.6489, Adjusted R-squared:  0.6421
## F-statistic: 96.09 on 1 and 52 DF,  p-value: 2.06e-13
```

5.(b) For the cases where "Dose code" is equal to 0, produce a scatter plot of Y versus X1 , and comment on whether the plot looks linear.

Solution: The plot looks linear at lower % values of X1, but gradually seems less linear and levels off at higher % values of X1. The loess curve superimposed on the scatter plot can illustrate a general negative slope trend that becomes less linear at higher % values of X1.

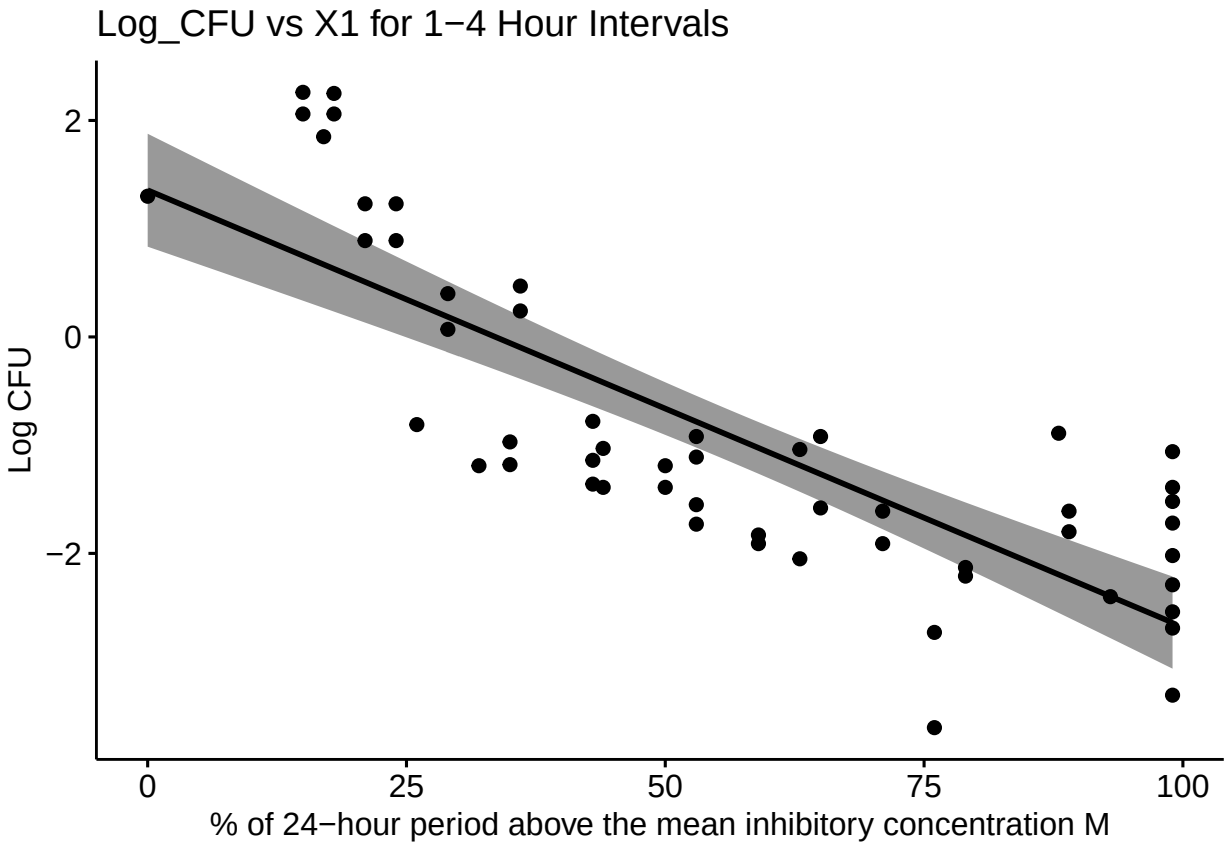
```
ggscatter(data = q5data[q5data$dose_code == 0,], x = "X1", y = "log_CFU",
          title = "Log_CFU vs X1 for 1-4 Hour Intervals",
          xlab = "% of 24-hour period above the mean inhibitory concentration M",
          ylab = "Log CFU", conf.int = T, add = "loess")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
ggscatter(data = q5data[q5data$dose_code == 0,], x = "X1", y = "log_CFU",
          title = "Log_CFU vs X1 for 1-4 Hour Intervals",
          xlab = "% of 24-hour period above the mean inhibitory concentration M",
          ylab = "Log CFU", conf.int = T, add = "reg.line")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




5.(c) Based on the information presented to you, does the direction of the coefficient of X1 suggest that a greater percentage of time above the mean inhibitory concentration M is beneficial, or that a lower percentage of time above M is beneficial? Explain your reasoning in a sentence.

Solution: As stated in the prompt, the lower the value of C, the more efficacious the antibiotic. In other words, the antibiotic is more efficacious when log CFU is low. Based on the information presented, the negative direction of the coefficient of X1 suggests that a greater percentage of time above the mean inhibitory concentration M is beneficial since it lowers the number of CFU observed, showing greater efficacy of the antibiotic.

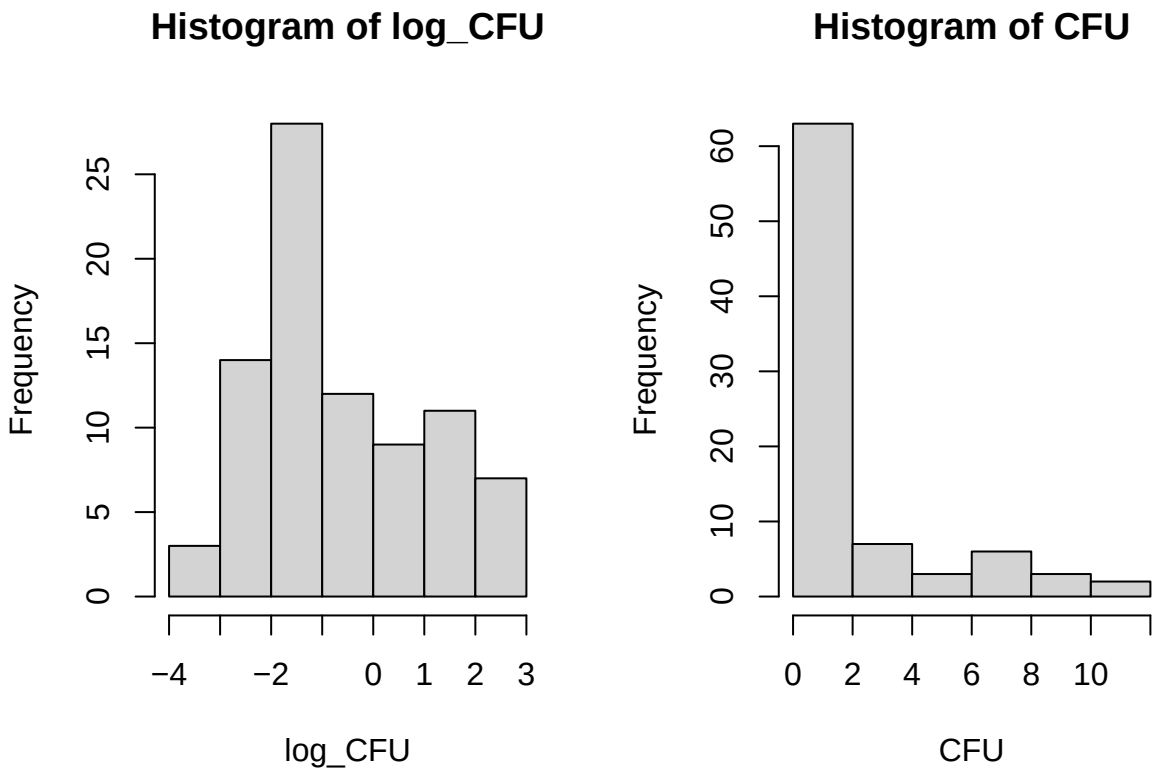
5.(d) A web search on "colony-forming unit" yielded the following description (from Wikipedia):

In microbiology, colony-forming unit (CFU or cfu) is a measure of viable bacterial or fungal numbers. Unlike direct microscopic counts where all cells, dead and living, are counted, CFU measures viable cells. For convenience the results are given as CFU/mL (colony-forming units per milliliter) for liquids, and CFU/g (colony-forming units per gram) for solids. A dilution made with bacteria and peptoned water is placed in an Agar plate ... and spread over the plate.... The theory behind the technique of CFU is to establish that a single bacterium can grow and become a colony via binary fission. ... [The] technique allows the determination of the number of CFU per mL in the sample, and thus the microbiological load and the magnitude of the infection in humans or animals, or the degree of contamination in samples of water, vegetables, soil or fruits, and in industrial products and equipment.

Why do you think investigators analyzing data from a dilution-based laboratory procedure might decide to perform regression analysis on log-transformed values rather than on the original CFU values?

Solution: Examining the log-transformed values of CFU, we see that it is not quite a bell-shaped curve. However, transforming the log_CFU values back to the original scale we see that there is an extremely zero-inflated and right skewed distribution, indicating that analysis is best performed on the log CFU values to better meet model assumptions related to normality that are required in regression analysis.

```
par(mfrow = c(1,2))
hist(q5data$log_CFU, main = NULL, xlab = "log_CFU")
title("Histogram of log_CFU")
hist(exp(q5data$log_CFU), main = NULL, xlab = "CFU")
title("Histogram of CFU")
```




5.(e) Define a new variable (say "X1squared") where X1squared = (X1)2 , and perform a multiple regression analysis of Y versus X1 and X1squared. Produce predicted values for each of the following possible values of X1:

(i) X1 = 20 (ii) X1 = 40 (iii) X1 = 60 (iv) X1 = 80.

Also comment on whether there is a statistically significant departure from linearity in the data used to fit the model.

Solution: The predicted values for the four values of X1 (20, 40, 60, 80) are as follows: i) 0.969, ii) -0.627, iii) -1.653, and iv) -2.109. There is a statistically significant departure from linearity in the data used to fit the model since the quadratic term is statistically significant in the multiple regression model, and this is further corroborated by the shape of the loess line in the residuals vs. fitted values plot below.

```
q5data$X1squared <- (q5data$X1)^2
logCFUmlr <- lm(log_CFU ~ X1 + X1squared, data = q5data[q5data$dose_code == 0,])
summary(logCFUmlr)
```

```
##
## Call:
## lm(formula = log_CFU ~ X1 + X1squared, data = q5data[q5data$dose_code ==
##     0, ])
```

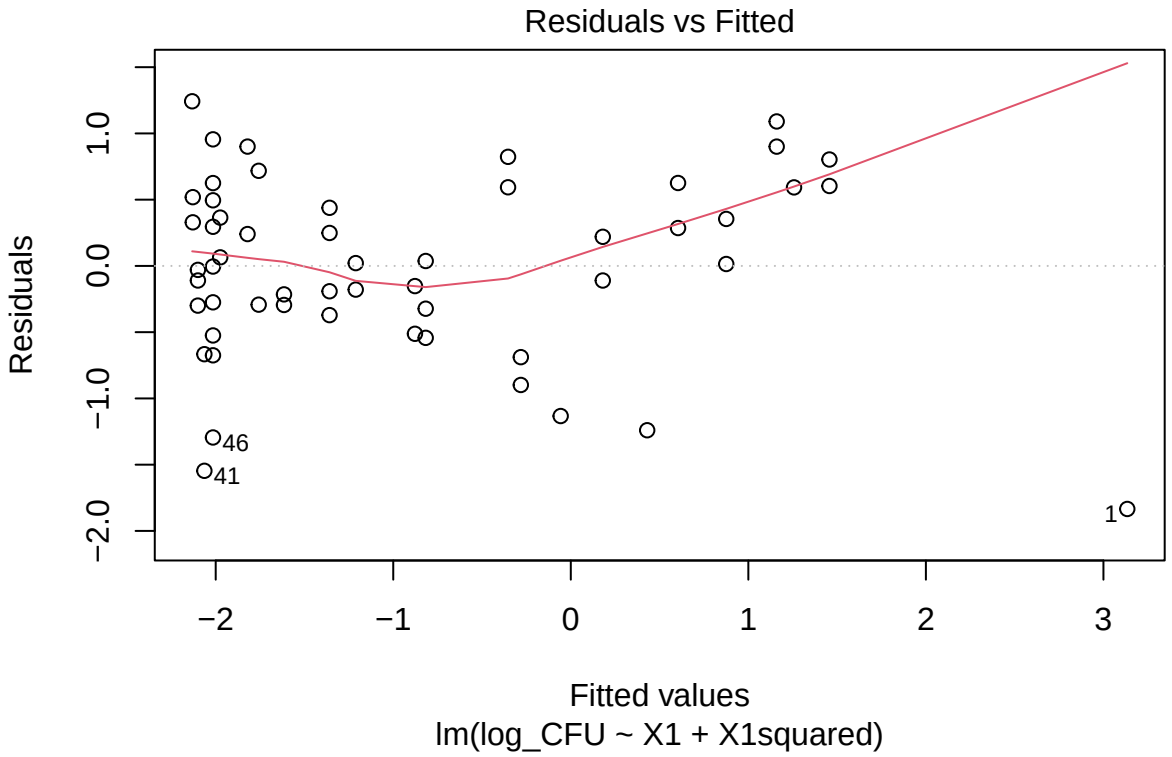
```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83425 -0.31695  0.01788  0.51303  1.24247
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.1342481  0.3785740   8.279 5.33e-11 ***
## X1          -0.1225153  0.0150168  -8.159 8.22e-11 ***
## X1squared    0.0007121  0.0001270   5.606 8.44e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6914 on 51 degrees of freedom
## Multiple R-squared:  0.7827, Adjusted R-squared:  0.7742
## F-statistic: 91.87 on 2 and 51 DF,  p-value: < 2.2e-16
```

```r
intercept <- summary(logCFUmlr)$coefficients[1, 1]
x1_coef <- summary(logCFUmlr)$coefficients[2, 1]
x1sq_coef <- summary(logCFUmlr)$coefficients[3, 1]

yhat <- function(x){
  y <- intercept + (x1_coef * x) + (x1sq_coef * (x^2))
  return(y)
}
X1_20 <- yhat(20) # prediction for X1 = 20
X1_40 <- yhat(40) # prediction for X1 = 40
X1_60 <- yhat(60) # prediction for X1 = 60
X1_80 <- yhat(80) # prediction for X1 = 80
X1est <- data.frame(X1 = c(20,40,60,80),
                    predicted = round(c(X1_20, X1_40, X1_60, X1_80),2))
kbl(X1est, booktabs = T) %>%
  kable_styling(position = "center",
                latex_options = c("basic", "hold_position"))
```

| X1 | predicted |
|---|---|
| 20 | 0.97 |
| 40 | -0.63 |
| 60 | -1.65 |
| 80 | -2.11 |

```r
#residual vs fitted plot
plot(logCFUmlr, which = 1)
```

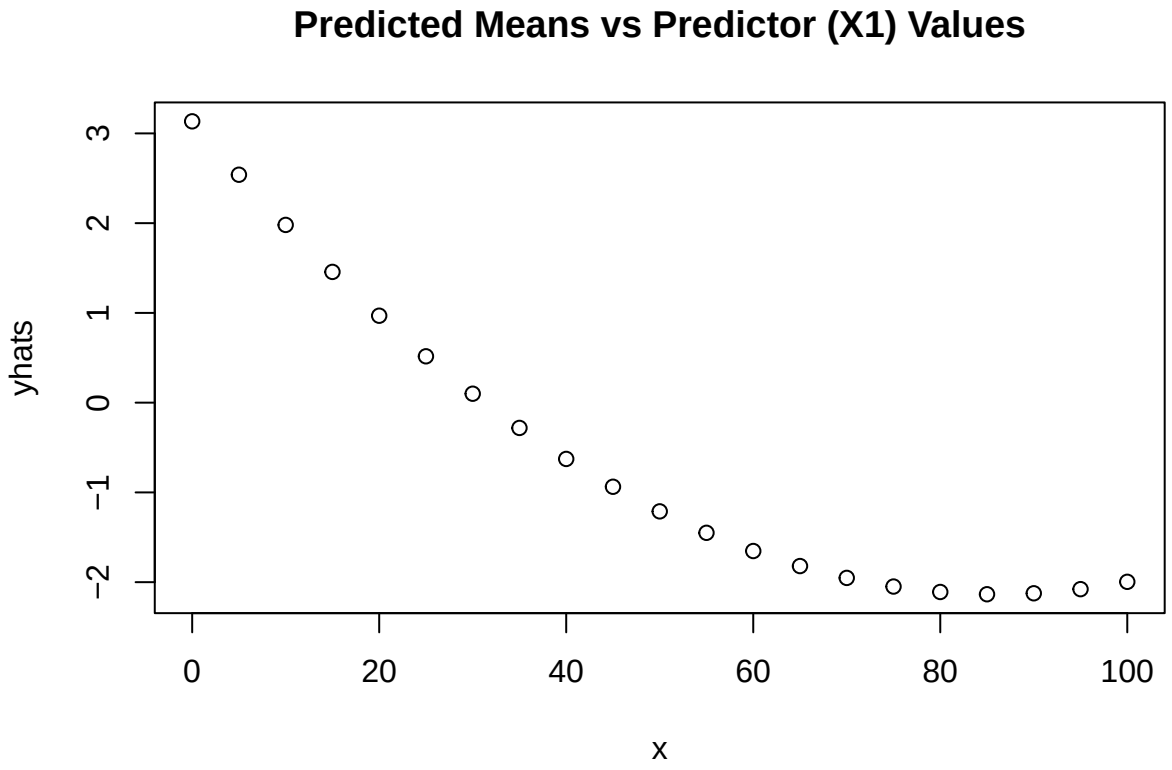



5.(f) In a setting where there is reason to think that the outcome variable would have a "monotone" (i.e., always increasing or always decreasing) pattern across the range of possible values of the predictor, one possible concern about including a "quadratic" predictor (reflecting the squared value of another predictor variable) is that a quadratic function would in general have the shape of a parabola, which informally speaking might be described as having either a U-shape or an inverted U-shape. That is, the scientific context might imply that there should be a monotone pattern across the range of possible values of the predictor, but the mathematical form of the regression model would allow for a non-monotone pattern, with the predicted means going down and then up across the range of predictor values (or going up and then down across the range of predictor values). Suppose it can be taken as a given in the context of the antibiotic study that the outcome should display a monotone pattern across the range of possible predictor values. Is the fitted model from part (e) in conflict with the notion that the mean outcome follows a monotone pattern across the range of predictor values? Support your answer with illustrative predicted values from the fitted regression model.

Solution: The fitted regression model from part (e) may be in conflict with the notion that the predicted mean outcome follows a monotone pattern across the range of predictor values. We can see that the predicted values from the model generally follow a monotonic pattern, but the X1 values from 85-100 begin increasing slightly instead of following the same monotonic trend for all values before that.

```
x <- seq(0,100,5)
yhats <- yhat(x)

plot(x, yhats)
title("Predicted Means vs Predictor (X1) Values")
```




6. The summaries below and on the following page are taken from an analysis of deaths in London during the month of December, 1952, and their possible relationship to air pollution. The 15 observations represent one observation for each day from December 1-15, 1952. The variable "death" is a count of the number of deaths in London Administrative County on that day, the variable "smoke" is the mean atmospheric concentration of smoke particles at County Hall on the given day, and the variable "sulfur" is the mean atmospheric concentration of sulfur dioxide at County Hall on the given day.

6.(a) The data point for December 6, 1952 featured smoke = 3.45, sulfur = .86, and death = 294. For each regression model summarized below, obtain the predicted value and residual corresponding to this data point.

Solution: See table below.

```
# data point values
smoke <- 3.45
sulfur <- 0.86
death <- 294

# Model 1: death = b0 + b1*smoke + e
mod1_estimate <- 171.81881 + 63.76092 * smoke

# Model 2: death = b0 + b1*sulfur + e
mod2_estimate <- 144.11078 + 256.23556 * sulfur

# Model 3: death = b0 + b1*smoke + b2*sulfur + e
mod3_estimate <- 89.51080 - 220.32438 * smoke + 1051.81646 * sulfur

# calculating residuals for each model estimate
mod1_rsd <- death - mod1_estimate
```

```
mod2_rsd <- death - mod2_estimate
mod3_rsd <- death - mod3_estimate

tbl6a <- data.frame(model = c("Model 1", "Model 2", "Model 3"),
                    predicted = round(c(mod1_estimate, mod2_estimate, mod3_estimate),2),
                    residual = round(c(mod1_rsd, mod2_rsd, mod3_rsd),2))

# summary of predicted values and residuals for all models
kbl(tbl6a, booktabs = T) %>%
  kable_styling(position = "center",
                latex_options = c("basic", "hold_position"))
```

| model | predicted | residual |
|---|---|---|
| Model 1 | 391.79 | -97.79 |
| Model 2 | 364.47 | -70.47 |
| Model 3 | 233.95 | 60.05 |

6.(b) Characterize the correlation between "smoke" and "death" according to whether it is positive or negative and according to whether it is significantly different from 0 or not. Point to the evidence that supports your conclusions.

Solution: The correlation between an outcome and predictor can be determined by using the formula $\beta = r\frac{s_y}{s_x}$. Solving for r, we get r = 0.756. Examining values of Pearson's R associated with 13 (df = 15-2) degrees of freedom, the critical value for a two-tailed test at the $\alpha = 0.05$ level is 0.514 (from table of critical values of Pearson's $\rho$. Since our calculated value is greater than the critical value, we reject the null and conclude that the correlation between "smoke" and "death" is significantly different from 0 ($p < .001$), and we also know that it has a positive moderately strong correlation based on its sign and magnitude.

```
s_x <- 1.548344 # sample sd from q6 data summary
s_y <- 130.5898 # sample sd from q6 data summary
beta <- 63.76092 # coef estimate from q6 model 1 summary
r <- beta*(s_x/s_y)
r
```

```
## [1] 0.7559843
```

```
# calculate t for calculated r
tstar <- r*sqrt((15-2)/(1-r^2))
# p value associated with calculated t
pt(tstar, df = 15-2, lower.tail = F)
```

```
## [1] 0.0005558381
```

6.(c) In the regression of death on smoke, consider the p-value for the test of whether the coefficient of smoke is equal to zero versus the alternative that the coefficient of smoke is not equal to zero. Is this p-value greater than 0.001, less than 0.001, precisely equal to 0.001, or is there not enough information to say? Support your answer in a sentence.

Solution: In the regression of death on smoke as seen in the results from Model 1, the t-value is calculated to be 4.16 with an associated p-value of 0.001. There is enough information to say so, given that the regression coefficient and standard error are all available in the calculation. We also know that the model is a fairly

good fit (p-value of F = .0011) so we can trust the estimates of the intercept and coefficient given in the output.

6.(d) What is the correlation between smoke and sulfur in this data set?

Solution: To examine the correlation between smoke and sulfur, we look to the Variance Inflation Factor (VIF), which can be calculated by $VIF = \frac{1}{2-r_k^2}$, with $r_k^2$ representing the coefficient of determination between two covariates (in this case, between smoke and sulfur). The model output for Model 3 already gives us the VIF = 40.458. The VIF is significantly greater than 10 (rule of thumb for VIF), indicating that there issues with collinearity that may affect the precision of coefficient estimates.

6.(e) Suppose you work in the public health department and are told that in addition to the type of coal that is most widely used in London factories (Type A), there are two alternative types of coal that cost the same as the most widely used type of coal. One alternative form of coal (Type B) produces less sulfur dioxide but more smoke. The other alternative form of coal (Type C) produces both less sulfur dioxide and less smoke. Based on everything you know, including the results of these regression analyses, which type of coal would you recommend for use in coal-burning factories? Support your answer in a few sentences.

Solution: Due to the VIF results in the multiple regression, it is better to model smoke and sulfur as predictors in separate models (Model 1 and Model 2). The coefficient estimates for smoke and sulfur in Model 1 and Model 2 are both positive, indicating positive slopes for the best fit line of smoke and death and sulfur and death. Based on these relationships, it is natural to recommend the Type C alternative form of coal, since it produces both less sulfur and less smoke and will minimize the estimate of death in both of these models.