Stat 21 Homework 4

Solutions sketched

Due: Sunday, Feb 27th by midnight

Contents

Use this file as the template for your submission. Do not delete anything from this template unless you are prompted to do so (e.g. where to write your name above, where to write your solutions or code below). Make sure you have installed the following packages in your version of RStudio: `tidyverse`, `knitr` **before** you attempt to knit this document.

Your completed assignment should be submitted as a single **PDF** using the link under Week 3 titled "Submit HW 4 to Gradescope". You must use R markdown to write up your solutions. For any homework problems that involve coding in R, you must provide **both** the code and a written answer interpreting the output within the context of the problem. You are allowed to work with your classmates on this homework assignment but you are expected to write up your own solutions. Every answer must be supported by a written statement unless otherwise specified. *A good rule of thumb is to make sure your answer is understandable to someone who hasn't read the problem question (or code output associated with it).*

Additionally, make sure that when you upload your solutions to Gradescope, you select which pages go correspond with which questions. Also, check to make sure that your knitted homework document is not uploaded as an extra-long single page document. Failure to do these things will result in a penalty on your homework grade. Finally, I strongly recommend that you address and resolve any knitting or R coding issues before Saturday as solutions to any R-coding questions that are not knitted properly will not receive any credit.

Part I: Non-R Problems

1. Suppose we conduct a t-test for the slope of a linear regression model and fail to reject the null hypothesis that $H_0 : \beta_1 = 0$. In this case, the data supports modeling the response variable as

$$Y = \beta_0 + \epsilon.$$

In this case, what is β_0? Show your work and/or explain your reasoning.

Solution: $\hat{\beta}_0 = \bar{y}$

The next two problems are meant to clarify a mistake I made in class when discussing how to interpret the coefficients of a regression model for a transformed response. Work through these mathematical problems carefully and compare your answers to your notes from class to identify what the mistake was and how to correct it.

For problems 2 and 3, recall the SLR model we fit to model the average number of surviving bacteria in a canned food product and the minutes of exposure to 300 degree Fahrenheit heat. (Note: Do not copy/paste Greek letters into this document of you will not be able to knit your solutions. You may write out mathematical equations by hand and attach an image of your solutions to your final PDF or you may simply write out equations with regular text. For example: Y_hat = 3.3 - 0.98X)

Problem 2

(a) Write out the estimated regression equation for the original data where the number of surviving bacteria is the response and the minutes of exposure, x, is the predictor.

(b) Now, suppose we add one minute of exposure, $x + 1$, and write out the estimated regression equation for the original data with the number of surviving bacteria as the response and the minutes of exposure is the predictor.

(c) What is the difference between the two equations in (a) and (b) and how do we interpret this difference? (Hint: Take the equation from part (b) and subtract the equation in part (a).)

Solution:

(a) $\hat{y} = b + ax$

(b) $\hat{y} = b + a(x + 1)$

(c) $b + a(x + 1) - b + ax = a$ represents the average change in Y per unit increase in x

Problem 3

(a) Write out the estimated regression equation for the (natural) logarithm of the number of surviving bacteria as the response and the minutes of exposure, x, as the predictor in the original units of the response. (E.g if we modeled $log(y) = b + ax$ then this is equivalent to modeling $y = e^{log(y)} = e^{b+ax}$.)

(b) Now, suppose we add one minute of exposure, $x + 1$, and write out the estimated regression equation for the (natural) logarithm of the number of surviving bacteria as the response and the minutes of exposure as the predictor.

(c) What is the ratio between the two equations in (a) and (b) and how do we interpret this? (Hint: Take the equation from part (b) and divide by the equation in part (a).)

Solution:

(a) $\hat{y} = e^{\hat{log(y)}} = e^{b+ax}$

(b) $\hat{y} = e^{\hat{log(y)}} = e^{b+a(x+1)}$

(c) $\frac{e^{b+a(x+1)}}{e^{b+ax}} = \frac{e^b e^{ax} e^a}{e^b e^{ax}} = e^a$ represents the average multiplicative (e.g. percent growth/decrease) change in Y per unit increase in x

Part II: R Problems

Problem 4

The following chunks of code outlines a simulation study regarding confidence intervals for a SLR model. In this problem, we will consider a true (not estimated) regression equation of

$$y = 50 + 10x + \epsilon, \quad \text{where } \epsilon \sim N(0, 16)$$

and we will generate 500 repeated samples of $n = 20$ observations drawing one observation for each level of $x = 1, 1.5, 2, \ldots, 10$ for each sample.

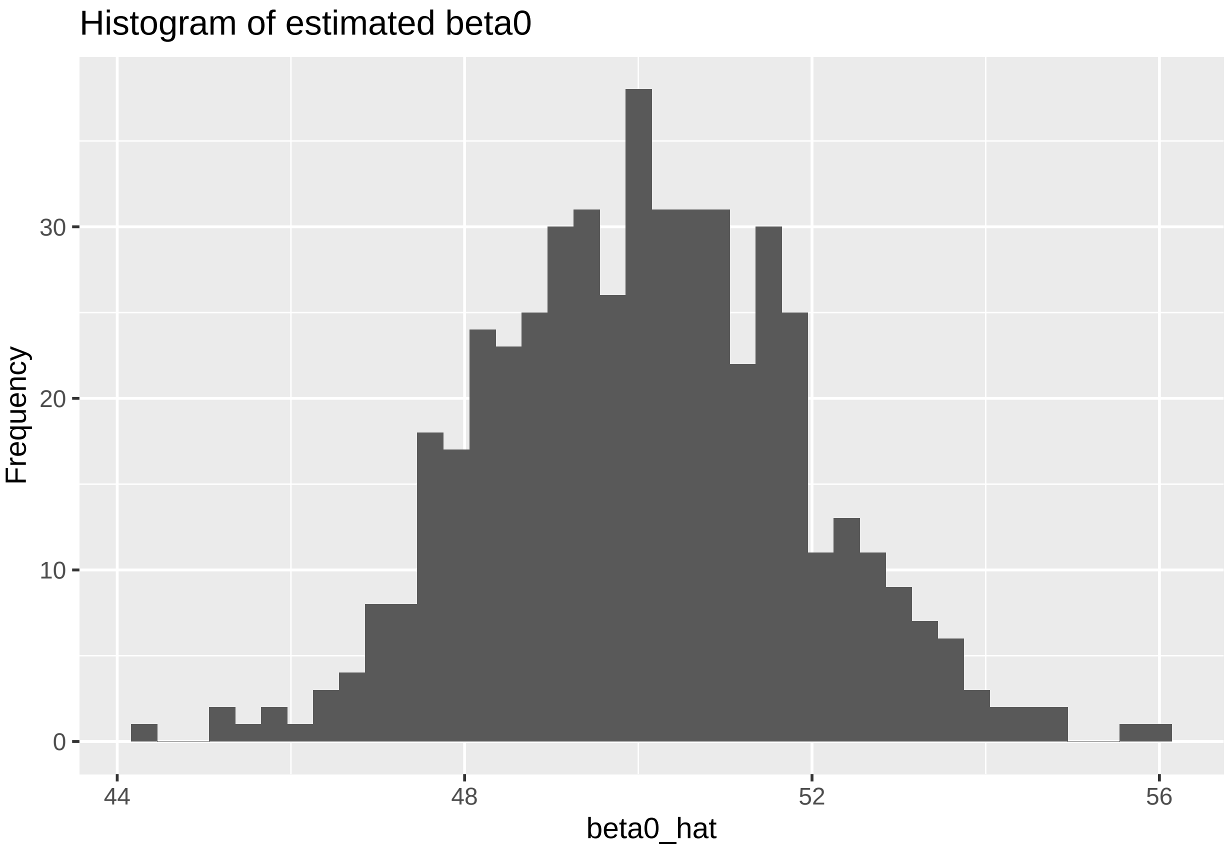
Part 1: For each sample, the code below computes the least-squares estimates of the slope and intercept.

```r
## Initialization
set.seed(1001)
remove(list=ls())
x <- seq(0.5, 10, by=0.5)

## Regression equation
generate.response <- function(x){
  y <- 50 + 10*x + rnorm(20, mean=0, sd=4)
  return(y)
}
y_mtx <- matrix(rep(x, 500), ncol=500, byrow=FALSE)
y_mtx <- apply(y_mtx, 2, generate.response)

## Least squares estimates
estimate.parameters <- function(response, predictor=x){
  mod = lm(response ~ predictor)
  beta0_hat = mod$coefficients[1]
  beta1_hat = mod$coefficients[2]
  return(c(beta0_hat, beta1_hat))
}

estimated_betas <- matrix(rep(NA,500*2), ncol=2)
colnames(estimated_betas) <- c("beta0_hat", "beta1_hat")
for(k in 1:500){
  estimated_betas[k,] = estimate.parameters(y_mtx[ ,k])
}
estimated_betas <- data.frame(estimated_betas)
#estimated_betas ## Please make sure this line is commented out before your knit your solutions to hand
names(estimated_betas)
```

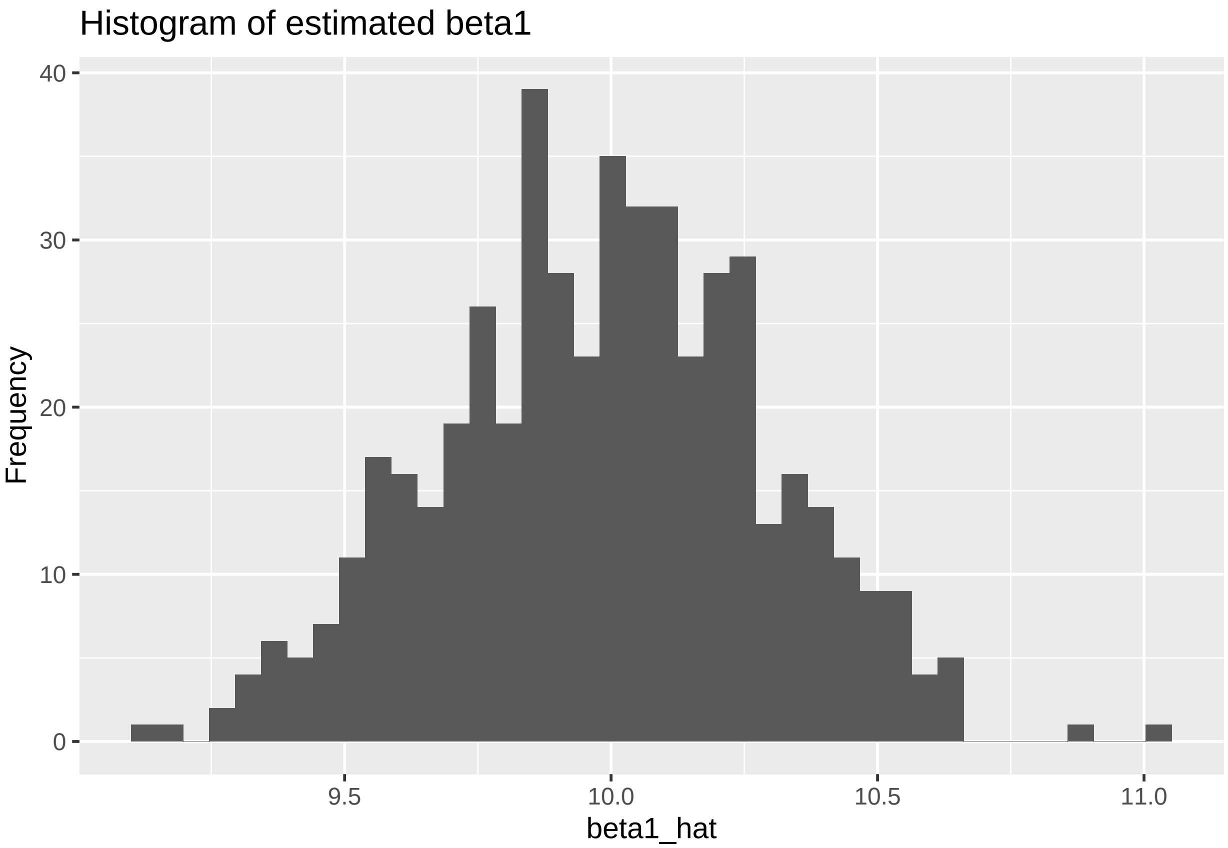
```
## [1] "beta0_hat" "beta1_hat"
```

(a) Construct histograms of the sample vlaues of $\hat{\beta}_0$ and $\hat{\beta}_1$ (the estimated_betas) by uncommenting the code below and filling in the question marks. Discuss the shape of these histograms.

```r
ggplot(estimated_betas, aes(x=beta0_hat)) +
  geom_histogram(bins=40) +
  labs(title= "Histogram of estimated beta0", x = "beta0_hat", y = "Frequency")
```



```r
ggplot(estimated_betas, aes(x=beta1_hat)) +
  geom_histogram(bins=40) +
  labs(title= "Histogram of estimated beta1", x = "beta1_hat", y = "Frequency")
```



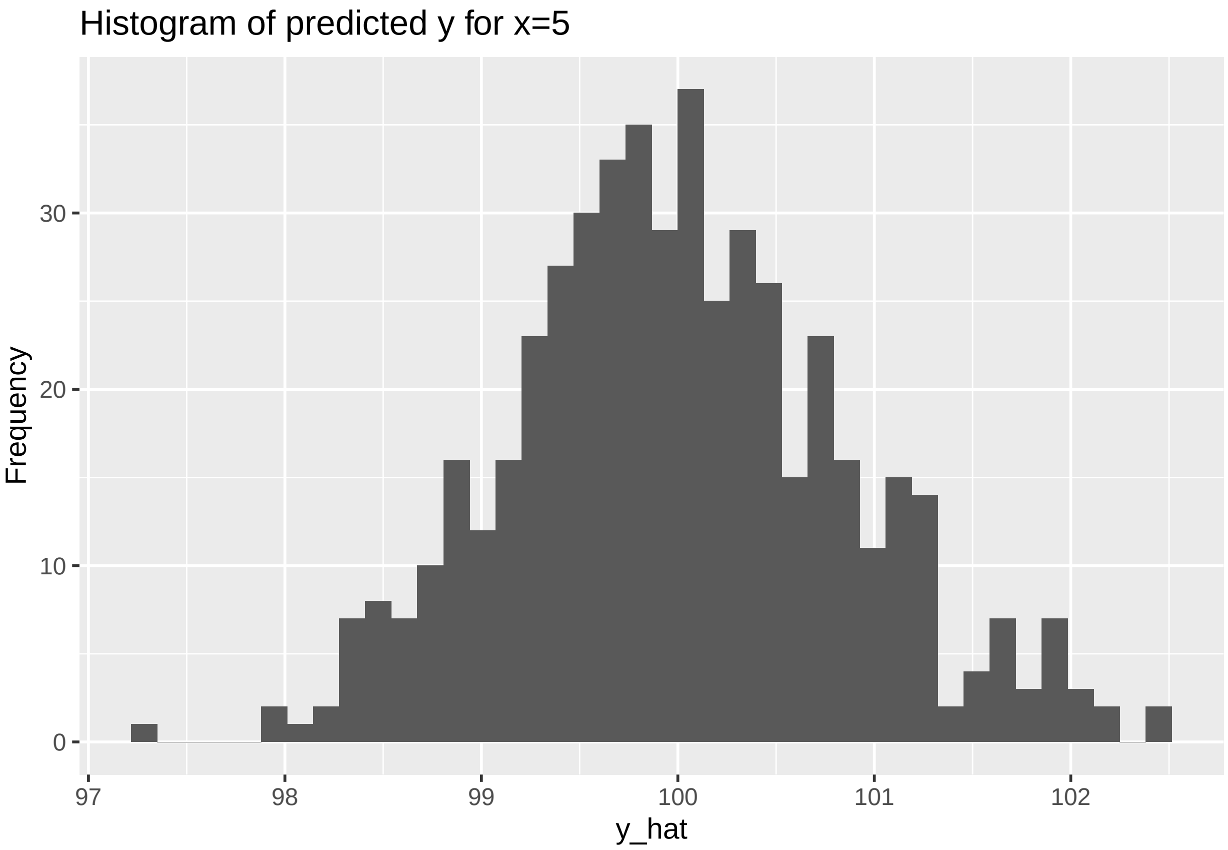
Solution: Both histograms look Normal (symmetric, bell shaped, unimodal)

Part 2: For each sample, the code below computes an estimate of $\hat{y}$ for $x = 5$.

```r
## Predicting Y values
estimation.observation <- function(response,predictor=x){
  mod <- lm(response ~ predictor)
  y_hat <- mod$coefficients[1] + mod$coefficients[2]*5
  return(y_hat=y_hat)
}

y_hat <- rep(NA,500)
for(k in 1:500){
  y_hat[k] = estimate.observation(y_mtx[ ,k])
}
y_hats <- data.frame(y_hat)
#y_hats ## Please make sure this line is commented out before your knit your solutions to hand in!
```

(b) Construct a histogram of the estimates you obtained (the 500 y_hat values). Discuss the shape of the histogram

```r
ggplot(y_hats, aes(x=y_hat)) +
  geom_histogram(bins=40) +
  labs(title= "Histogram of predicted y for x=5", x = "y_hat", y = "Frequency")
```



Solution: Again, looks normal.

Part 3: For each sample, the code below compute a 95% CI on the slope, β_1

```
## CI for slope of predictor
CI.slope <- function(response,predictor=x){
  mod <- lm(response ~ predictor)
  t_crit <- qt(0.05/2, df=20-2, lower.tail=FALSE)
  se_beta1 <- summary(mod)$coefficients[2,2] ##?
  beta1_hat <- mod$coefficients[2]
  LB <- beta1_hat - t_crit*se_beta1
  UB <- beta1_hat + t_crit*se_beta1
  return(c(LB, UB))
}

LB_slope <- rep(NA,500)
UB_slope <- rep(NA, 500)
for(k in 1:500){
  conf_int = CI.slope(y_mtx[ ,k])
  LB_slope[k] = conf_int[1]
  UB_slope[k] = conf_int[2]
}
CI_slope <- tibble(LB_slope,UB_slope)
#CI_slope ## Please make sure this line is commented out before your knit your solutions to hand in!
```

(c) Use the next two lines of code to determine how many of these intervals contain the true value $\beta_1 = 10$? Is this what you would expect?

```
((CI_slope$LB_slope<=10)&(CI_slope$UB_slope>=10))  ## Please make sure this line is commented out befor
```

```
##   [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [13]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [25]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [37]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
##  [49]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE
##  [61]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [73]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [85]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [97]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [109]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [121]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [133]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [145]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
## [157]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [169]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [181]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [193]  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [205]  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [217]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [229]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
## [241]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [253]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [265]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [277]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
## [289]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [301]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
## [313]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
## [325]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
## [337]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [349]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
## [361]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [373]  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [385]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [397]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [409]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [421]  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [433]  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [445]  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [457]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [469]  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [481]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [493]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

```
sum((CI_slope$LB_slope<=10)&(CI_slope$UB_slope>=10))
```

```
## [1] 481
```

Solution: proportion of intervals that cover $\beta_0 = 10$ is approximately the confidence level $= 481/500 = 0.962$

Part 4: For each estimate of $\hat{y}$ for $x = 5$ in part 2, the code below computes the 95% CI for the mean response.

```
## CI for mean response
CI.mean.response <- function(response,predictor=x){
  dat <- data.frame(response, predictor)
  remove(list=c("response","predictor"))
  mod <- lm(response ~ predictor, dat)
  y_hat <- data.frame(response = mod$coefficients[1] + mod$coefficients[2]*5, predictor=5)
  PI <- predict(mod, y_hat, interval="confidence", level = 0.95)
  return(c(PI[2], PI[3]))
}

LB_mean_response <- rep(NA,500)
UB_mean_response <- rep(NA, 500)
for(k in 1:500){
  conf_int = CI.mean.response(y_mtx[ ,k])
  LB_mean_response[k] = conf_int[1]
  UB_mean_response[k] = conf_int[2]
}
CI_mean_y = tibble(LB_mean_response,UB_mean_response)
#CI_mean_y ## Please make sure this line is commented out before your knit your solutions to hand in!
```

(d) Use the next two lines of code to determine how many of these intervals contain the true value of $E[Y \mid x = 5] = 100$? Is this what you would expect?

```
((CI_mean_y$LB_mean_response<=100)&(CI_mean_y$UB_mean_response>=100)) ## Please make sure this line is
```

```
##   [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [13]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
##  [25]  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
##  [37]  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
##  [49]  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [61]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [73]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
##  [85]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [97]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [109]  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [121]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [133]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
## [145]  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [157]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [169]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [181]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
## [193]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [205]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [217]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [229]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE
## [241]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [253]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [265]  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [277]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
## [289]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
## [301]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
## [313]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
## [325]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [337]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [349]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
## [361]  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [373]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [385]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [397]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [409]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [421]  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [433]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [445]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
## [457]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [469]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [481]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [493]  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
```

```
sum((CI_mean_y$LB_mean_response<=100)&(CI_mean_y$UB_mean_response>=100))
```

```
## [1] 476
```

Solution: proportion of intervals that cover $\hat{y} = 100$ is approximately the confidence level $= 481/500 = 0.952$

The following data set was collected in 2018-2019 and recorded different attributes of skyscrapers in NYC. In Problems 5-10 we are going to investigate how the height (in meters) (Y) of a skyscraper depends on the number of stories (i.e. floors) it has (x).

```
skyscraper_data <- read.delim("~/Downloads/skyscraper_data.txt")
head(skyscraper_data)
```

```
##   ID          Building_name height_meters height_ft floors year       material
## 1  1         30 Hudson Yards         386.6     1,268     73 2019 concrete_steel
## 2  2    3 World Trade Center         328.9     1,079     69 2018      composite
## 3  3         35 Hudson Yards         307.8     1,010     71 2019       concrete
## 4  4 220 Central Park South         290.2       952     70 2019       concrete
## 5  5         15 Hudson Yards         278.6       914     70 2019       concrete
## 6  6            The Centrale         244.8       803     64 2019       concrete
##       purpose
## 1      office
## 2       office
## 3 residential
## 4 residential
## 5 residential
## 6 residential
```

Problem 5

Make a scatter plot of the observed predictor and response variables and report

(a) The estimated regression equation.

(b) The value of the standard deviation of height.

(c) The value of R-squared.

```
regmod <- lm(height_meters ~ floors, skyscraper_data)
regmod %>% summary
```

```
##
## Call:
## lm(formula = height_meters ~ floors, data = skyscraper_data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.035 -18.292  -6.147   9.587  91.617
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -19.4884    11.1383   -1.75    0.087 .
## floors        4.3078     0.2434   17.70   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.8 on 45 degrees of freedom
## Multiple R-squared:  0.8743, Adjusted R-squared:  0.8716
## F-statistic: 313.1 on 1 and 45 DF,  p-value: < 2.2e-16
```

Solution:

(a) $\hat{y} = -19.49 + 4.31x$ where $x =$ number of floors and $y =$ height in meters

(b) The estimate for the standard deviation of the random variable Y is the same as the estimate for $sd(\epsilon)$: 25.8 meters

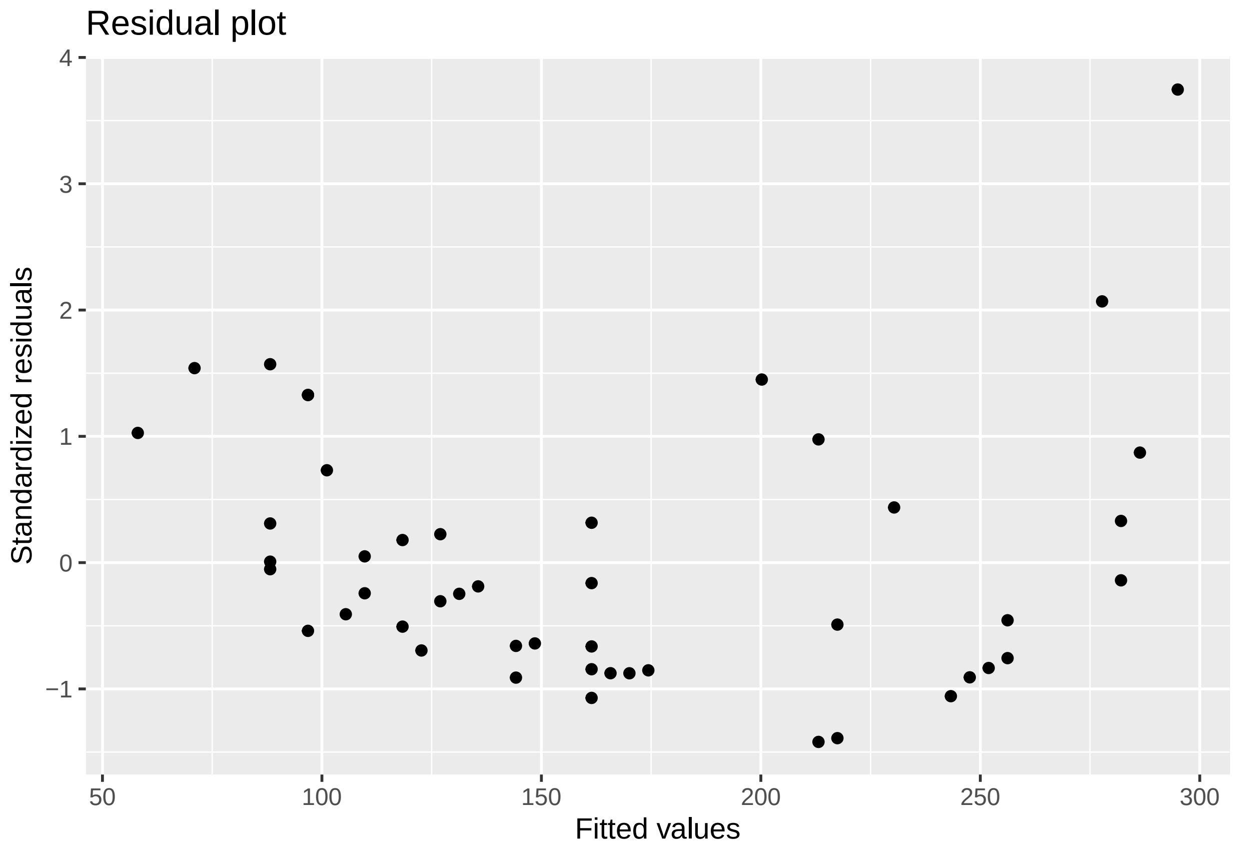
(c) 0.87343 or 0.8716

Problem 6

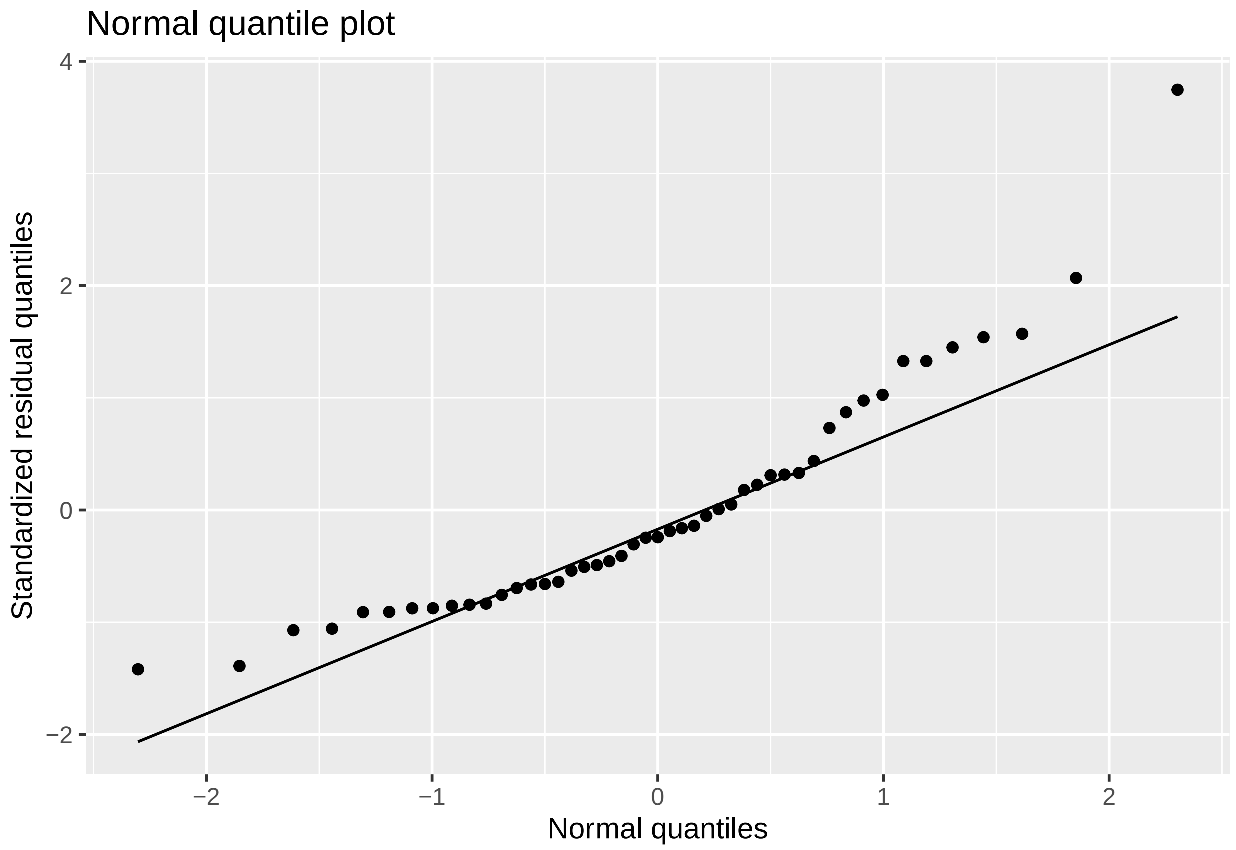
(a) Make a residual plot for the regression model you fit in Problem 5. Are there any apparent violations of the regression model assumptions? Explain your answer in four sentences or less.

(b) Make a Normal probability plot of the standardized residuals to determine if the residuals look like they come from a Normal distribution. Interpret the plot in three sentences or less.

```
sky_data_all <- skyscraper_data %>% mutate(resids = regmod$residuals,
                                           sresids = rstandard(regmod),
                                           fits = regmod$fitted.values)

ggplot(sky_data_all, aes(x = fits, y = sresids)) +
  geom_point() +
  labs(title="Residual plot", x="Fitted values", y="Standardized residuals")
```



```
ggplot(sky_data_all, aes(sample = sresids)) +
  geom_qq() +
  geom_qq_line() +
  labs(title="Normal quantile plot", x="Normal quantiles", y="Standardized residual quantiles")
```



Solution:

(a) there is a distinct curved pattern in the residuals in violation of the linearity assumption. also, there is some change in the vertical spread of the residuals as well as the spread is smaller for low fitted values of Y but larger for higher fitted values, in violation of the constant error variance assumption

(b) both the smallest and largest residuals are larger than we'd expect if the errors were normally distributed indicative of a positive (right) skew to the distribution of the errors. this asymmetry hints that the errors may not be normally distributed.

Problem 7

(a) Calculate a 95% confidence interval for the model parameter β_1, the slope of the variable `floors`. How would you explain the meaning of this confidence interval, in the specific context of this data set, to an architect who has never taken a statistics class?

(b) Test the hypothesis that $\beta_1 = 0$ at an $\alpha = 0.05$ significance level. State your null and alternative hypotheses and report the test statistic and p-value. Interpret, in the context of the problem, the results of this test in two sentences or less.

```
regmod %>% confint
```

```
##                  2.5 %   97.5 %
## (Intercept) -41.922113 2.945363
## floors        3.817518 4.798148
```

Solution:

(a) "Ordinarily I would say that I am 95% confident that each additional floor of a building results in an average increase of somewhere between 3.82 meters and 4.80 meters. However, I am not confident that the linear model which produced these estimates is appropriate for this data. Therefore, although the procedure I used to calculate these estimates is going to be correct 95% of the time we sample skyscrapers such as these, this is not necessarily the case here and I would like to explore other model options to find a more appropriate one."

(b) the test $H_0 : \beta_1 = 0, \quad H_A : \beta_1 \neq 0$ has a test statistic of 17.7 and p-value $< 2.2e - 16 \approx 0$ therefore the data provides evidence against H_0 in favor of H_1. This indicates that the number of floors a building has is related to/predictive of the height of the building (as we'd expect). This does NOT indicate that the relationship is linear however as assessed in Problem 6.

Problem 8

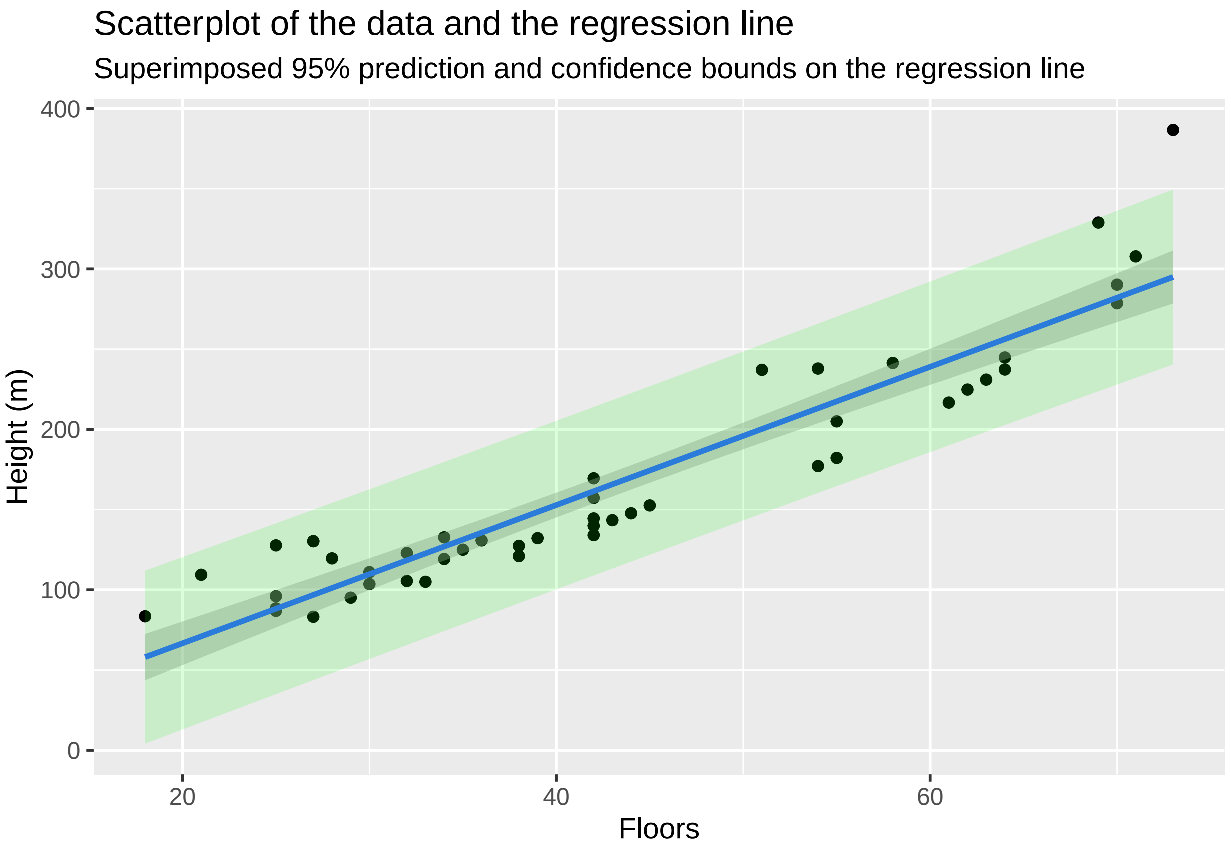
Suppose a developer is working on a new building that has taken the 15 years to get the go-ahead. Suppose they are cheekily designing the building to have 15 floors, one for each year of struggle to get the building approved. If the architect needs to know how tall this building may be, would you report to them a prediction interval or a confidence interval? Justify your answer in three sentences or less.

Solution: Here the architect would probably prefer a prediction interval because a building with 15 floors is very different from the number of floors in the buildings that build this regression model

Problem 9

Create a scatter plot of the observed data and overlay on this plot the estimated regression line and the confidence and prediction bands. Write 1-2 sentences describing what you see in this plot.

```
sky_data_pred_int <- tibble(cbind(sky_data_all, predict(regmod, skyscraper_data, interval="prediction")

sky_data_pred_int %>% ggplot() +
   geom_point(aes(x=floors, y=height_meters)) +
   geom_smooth(aes(x=floors, y=height_meters),method="lm",se=TRUE) +
   geom_ribbon(aes(x=floors, ymin=lwr, ymax=upr), fill="green",alpha=0.15 ) +
   labs(title="Scatterplot of the data and the regression line",
        subtitle="Superimposed 95% prediction and confidence bounds on the regression line",
        x="Floors", y="Height (m)")
```



Solution:

Despite the linearity assumption not being met, the prediction bounds cover just about every single data point. However, in the upper right hand corner, we see a point outside these bounds. This is because of the curved/non-linear relationship that our model fails to incorporate. We could try to remedy this by transforming the response or including a quadratic predictor term perhaps. The confidence bounds are also somewhat misleading because we failed to capture the nonlinear relationship between these variables. (At extreme values of the predictor the confidence bounds over-estimate the response. In the middle however, they underestimate it.)

Problem 10

Which assumptions are required to answer each of the previous four questions? Based on your analysis thus far, which assumptions seem most reasonable and are there any that you suspect are not strongly supported in this example? (Note: This is not really an R-coding problem.)

Solution: Since the linearity assumption is in question, we really can't use this model confidently for either estimation/predition or inference. The previous problems and be identified in the following manner:

Problem 5 - estimation

Problem 6 - checking conditions for both estimation and inference

Problem 7 - inference

Problem 8 - inference

Problem 9 - inference