Stat 230 HW 2

Name: Victor Huang

worked with: None

Homework 2 is due **by noon, Tues 9/28**. Please complete the assignment in this Markdown document, filling in your answers and R code below. I didn't create answer and R chunk fields like I did with homework 1, but please fill in your answers and R code in the same manner as hw 1. Make sure to follow the homework guidelines when writing up this assignment (handout is located on the right side of moodle page).

Tips for using Markdown with homework sets:

- Work through a problem by putting your R code into R chunks in this .Rmd. Run the R code to make sure it works, then knit the .Rmd to verify they work in that environment.
- Make sure you load your data in the .Rmd and include any needed `library` commands.
- Feel free to edit or delete questions, instructions, or code provided in this file when producing your homework solution.
- For your final document, you can change the output type from `html_document` to `word_document` or `pdf_document`. These two to output types are better formatted for printing.
- Keep the hw problems in **problem order** I give in this doc. You can attach hand-written work for a problem (if needed) but make it clear in this doc when you are answering a problem using work attached to your printed pdf/word doc.

Problem 1: Big Bang: ch. 7 exercise 13

Comment: Complete this problem "by hand" using the info in Display 7.9 (i.e. *don't* load the data and fit a `lm`). Use the `qt` command in R to get your mutliplier t^* for the CI calculation.

```
qt(.975,22)
## [1] 2.073873
```

Problem 2: Agstrat data revisited

Recall the `agstrat.csv` data used for homework 1. This was a stratified random sample of US counties. We will consider the regression of farm acreage in 1992 (y=`acres92`) on farm acreage in 1982 (x=`acres82`).

(2a) Use the `ggplot2` package to create a scatterplot of `acres92` against `acres82` that includes the fitted regression line. Describe the direction, form and strenth of the relationship shown in this graph.

(2b) Fit the regression of `acres92` on `acres82`. Interpret the slope in context for this problem.

(2c) Compute the test statistic and p-value to test $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$ using the estimated slope and SE from part (b). Your answers should match the values given in part (b), but I want to see your *work* showing how these values are calculated in part (c). Then *interpret* the test stat and p-value in context.

(2d) Are the four regression model assumptions satistified for your model in part (b) (e.g. linearity, constant variance, normality and independence). Make sure to show and explain all graphs used to assess these assumptions.

Problem 3: Meat Processing: ch. 7 exercises 17 (a-b only), 18, 20

- For 17(a), the data for this example is called `case0702` which is in the `Sleuth3` package. Your R chunk should look like

```
library(Sleuth3)
head(case0702)
##   Time   pH
## 1    1 7.02
## 2    1 6.93
## 3    2 6.42
## 4    2 6.51
## 5    4 6.07
## 6    4 5.99
```

- Make sure that you use `log(TIME)` as the predictor in your model: `lm(pH ~ log(Time), data=case0702)`
- You don't need to do 17(b) by hand, just use the `predict` command in R. Important note: If you log a predictor in your `lm` command, e.g. `lm(y ~ log(x))`, then you give the `predict` command the value of `x` on the **original** (unlogged) scale when entering a value for `newdata`.

Problem 4: Biological Pest Control: ch. 8 exercise 17

- Data is in the `Sleuth3` package (`ex0817`).
- For part (a), limit yourself to exploring logarithm and square root transformation for one or both variables. How can you make these graphs?
  - The `ggplot2` package let's you easily do this **without** applying those functions to your variables, instead you add another layer that tells R how to scale a particular axis. This method of visualization is nice because your numerical labels on the x/y axis will still be measured in the original units of the variables. If you want to, say, look at the scatterplot of `sqrt(y)` against `log10(x)` (base-10 log) you would add the layers `scale_y_sqrt()` and `scale_x_log10()` to your scatterplot of `y` against `x`. For this example, that would look like:

```
ggplot(pest, aes(x= Load, y = Mass)) +
  geom_point() +
  scale_y_sqrt() +
  scale_x_log10()
```

- For part (b), fit the model and give the fitted regression equation.

Problem 5: Pollen Removal ch. 8 exercise 19

- The data for this problem is `ex0327`

- For part (b) scatterplot, use the `scale` layers described in Problem 4 hints.
- For part (c), use the `filter` command from `dplyr` to make a version of the data set that only contains cases where `DurationOfVisit < 31`

Problem 6: Normality Assumption

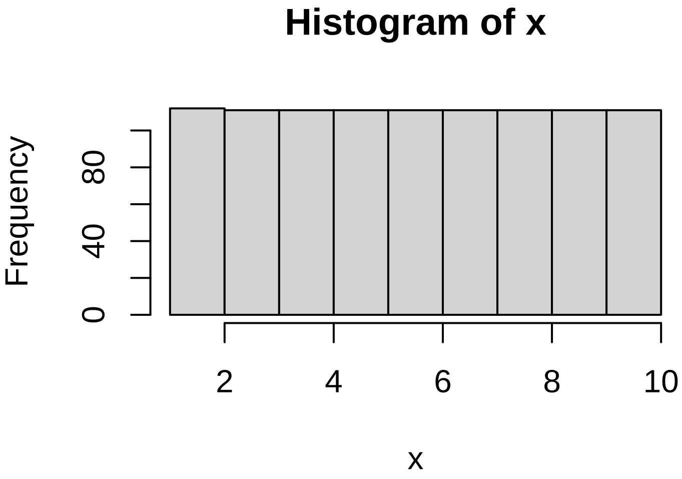
The (hidden) R chunk below defines a function named `reg.sim2` that samples responses from a SLR model given a vector of explanatory values x. You will use the following SLR model to generate your set of responses:

$$Y_i = 20 + 1x_i + \epsilon_i \qquad \epsilon_i \sim N(0, 2)$$

so that $\beta_0 = 20, \beta_1 = 1$ and $\sigma = 2$.

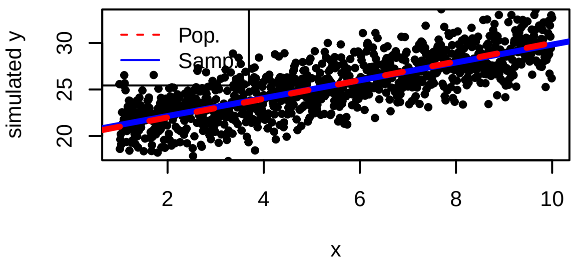
6a We start by generating a sample of 1000 explanatory variable values. Suppose the explanatory variable x is equally (uniformly) distribution between 1 and 10. Generate x and view its distribution (use whatever seed value you like):

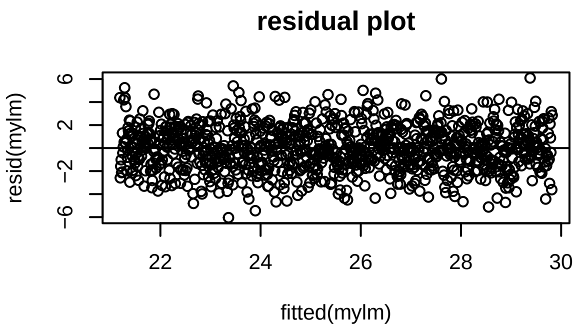
```
set.seed(7)
x <- seq(from=1,to=10,length=1000)
hist(x)
```

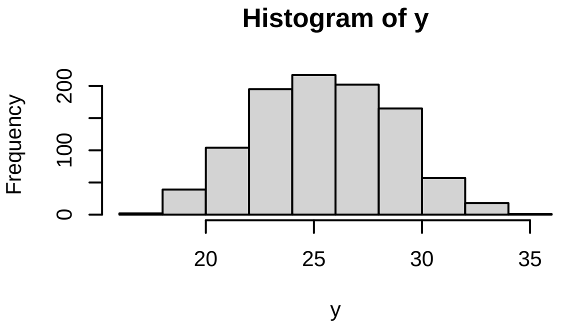


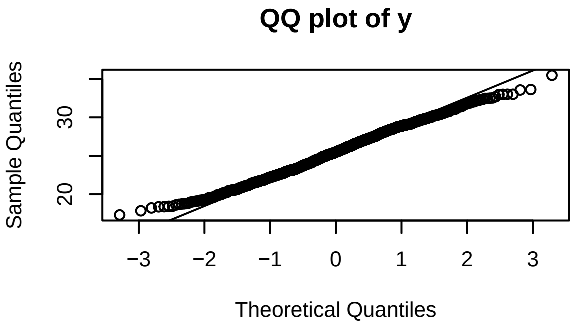
Next, for each of the 1000 x_i's that you just created, use the `reg.sim2` function to generate 1000 responses y_i from the population model described at the start of this problem:

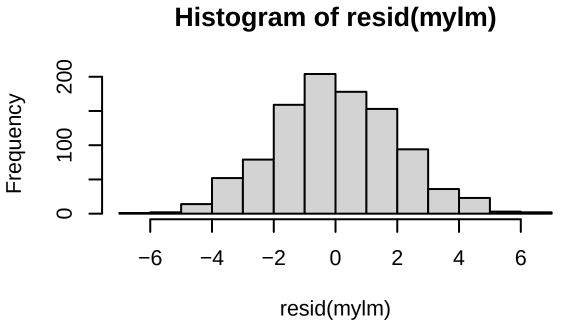
```
reg.sim2(x, beta0=20, beta1=1, sigma=2, grph=T)
```

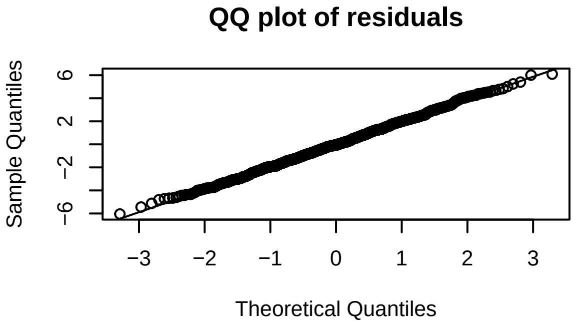



residual plot


Histogram of y


QQ plot of y


Histogram of resid(mylm)


QQ plot of residuals

```
## $b0
## [1] 20.23249
##
## $b1
## [1] 0.9588374
##
## $SE.b0
## [1] 0.1451941
##
## $SE.b1
## [1] 0.0238654
```

The R output gives the estimated slope and intercept $(\hat{\beta}_1, \hat{\beta}_0)$, a scatterplot of the data (along with the sample and population regression lines), and histograms/QQnormal plots for the responses y_i and the residuals r_i. Use the **histograms** and **QQ normal plots** to answer the following questions:

Are the sampled ys normally distributed? If not, describe the general shape of their distribution.

Are the residuals normally distributed? If not, describe the general shape of their distribution.

6b Repeat part (a), but this time generate a sample of 1000 x's that are skewed right using the command `rgamma(1000, 1, 1/2)`.

6c Repeat part (a), but this time generate a sample of 1000 x's that are normally distributed using the command `rnorm(1000, 10, 2)`.

6d Use your simulation results from (a)-(c) to explain how the distribution of the explanatory variable x can affect the distribution of the response.

6e (not a question!) Moral: All the data generated for this problem satisfy the SLR model assumptions. Your take away from Q7 should be to see that *neither* your response nor explanatory variables need to be *normally distributed* for the SLR model assumptions to hold. Rather, the SLR model says that the model errors (variation around the line) are normally distributed. Assessing the SLR "normality" assumption should focus on checking the distribution of the *residuals* rather than the distribution of the responses.