---

title: 'Data 501: Homework Assignment 2'

author: "Student's Name"

date: "Due date = 10/10/2020"

output:

word\_document: default

pdf\_document:

includes:

in\_header: header.tex

bibliography: AppliedStat.bib

---

```{r setup, include=FALSE}

knitr::opts\_chunk$set(echo = TRUE)

```

\*\*Please combine all your answers, the computer code and the figures into one PDF file, and submit a copy to gradescope.\*\*

\*\*Grading scheme: $\left\lbrace 0, 1, 2\right\rbrace$ points per question, total of 80.\*\*

\rc Due date: 11:59 PM October 10, 2020 (Saturday evening).\bc

# Question 1

In a recent, exciting, but also controversial Science article, [\blc Tomasetti and Vogelstein\bc](http://science.sciencemag.org/content/347/6217/78.full) attempt to explain why cancer incidence varies drastically across tissues (e.g. why one is much more likely to develop lung cancer rather than pelvic bone cancer). The authors show that a higher average lifetime risk for a cancer in a given tissue correlates with the rate of replication of stem cells in that tissue. The main inferential tool for their statistical analysis was a simple linear regression, which we will replicate here.

You can download the dataset as follows:

```{R}

tomasetti = read.csv("http://people.math.binghamton.edu/qiao/data501/data/Tomasetti.csv")

head(tomasetti)

```

The dataset contains information about 31 tumour types. The `Lscd` (Lifetime stem cell divisions) column refers to the total number of stem cell divisions during the average lifetime, while `Risk` refers to the lifetime risk for cancer of that tissue type.

1. Fit a simple linear regression model to the data with `log(Risk)` as the response variable and `log(Lscd)` as the predictor variable.

```{r}

tomasetti.lm <- lm(...~..., data = tomasetti)

```

2. Plot the estimated regression line and the data.

```{r}

plot(..., log(tomasetti$Risk) )

```

3. Add upper and lower 95% \*\*prediction\*\* bands to predict the response given a range of covariates on the plot, using `predict()`. That is, produce one line for the upper limit of each interval over a sequence of densities, and one line for the lower limits of the intervals.

4. Interpret the above bands at a `Lscd` = $10^{10}$.

5. Add upper and lower 95% \*\*confidence\*\* bands for the conditional mean response on the plot, using `predict()`. That is, produce one line for the upper limit of each interval over a sequence of densities, and one line for the lower limits of the intervals.

6. Interpret the above bands at a `Lscd` = $10^{10}$.

7. Test whether the slope in this regression is equal to 0 at level $\alpha=0.05$. State the null hypothesis, the alternative hypothesis, the conclusion, and the $p$-value.

```{r}

summary(tomasetti.lm)

```

```{r}

names(summary(tomasetti.lm))

```

8. What are assumptions you made for question (7) above.

9. Give a 95% confidence interval for the slope of the regression line.

10. Interpret your interval in (9).

11. Report the $R^2$ of the model.

12. Report the adjusted $R^2$ of the model.

13. Report an estimate of the variance of the errors in the model.

14. Provide an interpretation of the $R^2$ you calculated above, ideally to your neighbor who does not know much about statistics.

15. According to a [Reuters article](http://www.reuters.com/article/health-cancer-luck-idUSL1N0UE0VF20150101) "Plain old bad luck plays a major role in determining who gets cancer and who does not, according to researchers who found that two-thirds of cancer incidence of various types can be blamed on random mutations and not heredity or risky habits like smoking." Is this a correct interpretation of $R^2$?

# Question 2

From our textbook \*\*CH\*\* page 51, Exercie 2.9.

Let $Y$ and $X$ denote the labor force participation rate of women in 1972 and 1968, respectively, in each of 19 cities in the United States. The regression output for this data set is shown in the following table. It was also found that $\text{SSR} = .0358$ and $\text{SSE} = .0544$. Suppose that the model $Y = \beta\_{0} + \beta\_{1}X + \epsilon$ satifies the ususal regression assumptions.

| Variable |Coefficient|s.e|t-Test|p-value|

|--|--|--|--|--|

|Constant |.203311|.0976|2.08|.0526|

|X|.656040|.1961|3.35|$<.0038$|

|--|--|--|--|--|

|n = 19|$R^{2} = .397$|$R^{2}\_{a} = .362$|$\hat{\sigma} = .0566$|df = 17|

In this table \*\*s.e\*\* is the standard error of the estimate, \*\*t-Test\*\* is the value of the test statistics under the null hypothesis, \*\*p-value\*\* is the p-value of the test.

1. Compute $\widehat{\text{Var}}\left(Y\right)$ and $\widehat{\text{Cov}}\left(Y, X\right)$. Hint: for $\widehat{\text{Var}}\left(Y\right)$, check and compare the definitions of SST and sample variance. For $\widehat{\text{Cov}}\left(Y, X\right)$, (1) note that the correlation between $X$ and $Y$ can be computed from $R^2$; (2) compare the formulae for the sample correlation and the OLS estimate $\widehat\beta$; (3) note $\widehat{\text{Var}}\left(Y\right)$ can be computed.

2. Suppose participation rate of women in 1968 in a given city is $x=45\%$. What is the estimated participation rate of women in 1972 $y$ for the same city?

3. Suppose further that the mean and variance of the participantion rate of women in 1968 (i.e., the sample mean and variance of the $x$ values) are 0.5 and 0.005, respectively. Construct the 95\% \*\*confidence\*\* interval for the estimate in (2). Hint: you may use either equation (2.37) or (2.40) in the textbook to calculate the standard error, which is needed in this confidence interval. First determine which formula to use.

4. Construct the 95\% confidence interval for the slope of the true regression line $\beta\_{1}$. Hint: you may use (2.25) in the textbook to compute the standard error of the slope estimator $\widehat\beta\_1$. Alternatively, it should have been reported in the above table.

5. Test the hypothesis: $\text{H}\_{0}: \beta\_{1} = 1$ versus $\text{H}\_{a}: \beta\_{1} > 1$ at the 5\% significance level. Hint: note that you must not use the T-test reported in the above table since it is based on the null hypothesis that $\beta\_1=0$, not $\beta\_0=0$.

6. Compute the $R^{2}$ for this simple linear regression from the values of SSR and SSE.

7. If $X$ and $Y$ were reversed in the above regression, what would you expect $R^{2}$ to be?

# Question 3

This question is from our textbook \*\*CH\*\* Exercises 2.1, Page 53.

In order to investigate the feasibility of starting a Sunday edition for a large metropolitan newspaper, information was obtained from a sample of 34 newspapers concerning their daily and Sunday circulations (in thousands) \*(Source: Gale Directory of Publications, 1994)\*. The data can be read from the book's Website: [\blc http://www1.aucegypt.edu/faculty/hadi/RABE5/Data5/P054.txt\bc](http://www1.aucegypt.edu/faculty/hadi/RABE5/Data5/P054.txt).

1. Read the data using `read.table` (separator of column is `tab` and the data frame has variable names).

2. Construct a scatter plot of Sunday circulation versus daily circulation.

3. Does the plot suggest a linear relationship between daily and Sunday circulation?

4. Fit a regression line predicting Sunday circulation from daily circulation (Use `lm()`).

5. Is there a significant relationship between Sunday circulation and daily circulation? Justify your answer by a statistical test (Use F test in `anova()`).

6. Indicate what hypothesis your are testing and your conclusion for the test in part (5).

7. Using the `anova` table produced in part (5), compute the proportion of the variability in Sunday circulation is accounted for by daily circulation.

# Question 4

Let $Y$ and $X$ denote variables in a simple linear regression of median home prices versus median income in state in the US. Suppose that the model

$$

Y = \beta\_0 + \beta\_1 X + \epsilon

$$

satisfies the usual regression assumptions.

The table below is a table similar to the output of `anova` when passed a simple linear regression model.

Response: Y

Df Sum Sq Mean Sq F value Pr(>F)

X 1 NA 5291 NA NA

Residuals 48 181289 NA

1. Compute the missing values in the above table.

2. Test the null hypothesis $H\_0 : \beta\_1 = 0$ at level $\alpha = 0.05$ using the above table.

3. Can you test the hypothesis $H\_0 : \beta\_1 < 0$ using the above table? (You may need to use the relationship between the T statistic and the F statistic. Feel free to discuss this problem on Piazza.)

3. What proportion of the variability in $Y$ is accounted for by $X$?

4. If $Y$ and $X$ were reversed in the above regression, what would you expect $R^2$ to be?

# Question 5

In this problem, we will investigate what happens when the assumptions of the simple linear regression model do not hold. When generating data below, set $X$ to be equally spaced between 0 and 1 (i.e. `X = seq(0, 1, by=0.01)`) and use the regression function

$$

Y = 1 + \beta\_1 \cdot X + \epsilon

$$

1. Write a \*\*function\*\* (call the function as `generateTstat`) to generate data from the simple linear regression model with regression function as above and normally distributed errors $\epsilon \sim \text{N}\left(0, \sigma^{2}\right)$ (can use $\sigma^{2} = 1$), returning the $T$-statistic for testing whether the slop of the regression line is equal to 2.

[That is, testing $\text{H}\_{0}: \beta\_{1} = 2$ versus $\text{H}\_{a}: \beta\_{1} \neq 2$.]

The function arguments should be values for $X$ and the true slope of the regression line `beta1`.

```{r}

generateTstat = function(X, beta1){

#Y = write Y as a function of X and error

#fit = fit regression line using lm

# beta1hat = compute least squares estimate of slope using summary(fit)$coefficient[2,1]

# se\_beta1hat = compute standard error of slope estimate using summary(fit)$coefficients[2, 2]

# Tstat = Compute the T-statistic using the appropriate formula (for testing H0: beta\_1 = 2)

# return(Tstat)

Y = 1 + beta1 \* X + rnorm(length(X))

fit = lm(Y~X)

beta1hat = ...

se\_beta1hat = ...

Tstat = (beta1hat - 2)/se\_beta1hat

return(Tstat)

}

```

(I) Using your function, run a simulation with 5000 repetitions and with `beta1=2` to see if the $T$-statistic has distribution close to a $T$ distribution. How many degrees of freedom should it have (consider the length of $X$ to answer this question)? Note that it will take a while to complete the repetitions. Your answer to this question should not take more than 2 lines of code, as below.

```{r}

# X = use seq() to generate 101 'X' values between 0 and 1.

# t\_stat\_vec = use replicate() to compute 5000 T-statistic values

X = seq(0, 1, by=0.01)

t\_stat\_vec = replicate(n = 5000, ...)

```

```{r}

# Plot the distribution of the T-statistic and

# the T distribution that we thought the T-statistic would follow

# if the null hypothesis that beta\_1 = 2 were true.

df = length(X)-2

plot(density(...),col = 1,main='Estimated density of statistic from simulated data and true density')

x\_grid = seq(-5,5,length.out = 100)

lines(x\_grid,dt(x\_grid,df),col = 2) ## col = 2 means red

```

(II) In part (I), how often is your $T$ statistic larger than the usual 5\% threshold?

```{r}

#threshold = find the threshold for testing H0: beta\_1 = 2 versus Ha: beta\_1 not equal to 2

#(degrees of freedom depends on the length of X)

# Find how many of absolute t\_stat\_vec is greater than the threshold

cutoff = qt(... , ...)

mean(abs(t\_stat\_vec)>cutoff)

```

2. Write a new function, say `generateTstat\_t\_error`, with the same regression function but errors that are t-distributed using, say, `rt` with 5 degrees of freedom to generate errors. Repeat (I) and (II) in part (1). Does the $T$-statistic still have close to a $T$ distribution? How often is your $T$ statistic larger than the usual 5% threshold?

```{r}

generateTstat\_t\_error = function(X, beta1){

Y = 1 + beta1 \* X + rt(length(X),5)

...

...

}

t\_stat\_vec = replicate(n = 5000, generateTstat\_t\_error(X,2))

plot(density(t\_stat\_vec),col = 1,main='Estimated density of statistic from simulated data and true density')

x\_grid = seq(-5,5,length.out = 100)

lines(x\_grid,dt(x\_grid,df),col = 2) ## col = 2 means red

cutoff = qt(0.975,df)

mean(abs(t\_stat\_vec)>cutoff)

```

3. Write a new function, say, `generateTstat\_unequal\_var1`, with same regression function but errors that do not have the same variance though they are normally distributed. Construct errors such that the variance of the $i$-th error is `1+X[i]` (recall that `X` is equally spaced over interval 0 to 1 and you can specify in `rnorm` what is the standard deviation of error). Plot the variance of error as a function of `X`. Repeat (I) and (II) in part (1). Does the $T$-statistic still have close to a $T$ distribution? How often are your $T$ statistics larger than the usual 5% threshold?

```{r}

## as an example

## suppose we have var\_example = 1:10

var\_example = 1:10

## then you can generate 10 random variables from normal distribution with mean = 0 and different variance given by these 10 numbers

rnorm(10,0,var\_example)

## likewise, you can do something like

rnorm(10,0,var\_example+1)

rnorm(10,0,exp(var\_example+1))

```

```{r}

generateTstat\_unequal\_var1 = function(X, beta1){

Y = 1 + beta1 \* X + ...

...

...

}

t\_stat\_vec = replicate(n = 5000, generateTstat\_unequal\_var1(X,2))

plot(density(t\_stat\_vec),col = 1,main='Estimated density of statistic from simulated data and true density')

x\_grid = seq(-5,5,length.out = 100)

lines(x\_grid,dt(x\_grid,df),col = 2) ## col = 2 means red

cutoff = qt(0.975,df)

mean(abs(t\_stat\_vec)>cutoff)

```

4. Write a new function, say, `generateTstat\_unequal\_var2`, with same regression function but errors that do not have the same variance though they are normally distributed. Exaggerate the effect of non-constant variance by making the variance of errors `exp(1 + 5 \* X[i])`. Plot the variance as a function of `X`. Repeat (I) and (II) in part (1). Does the $T$-statistic still have close to a $T$ distribution? How often are your $T$ statistics larger than the usual 5% threshold?

```{r}

generateTstat\_unequal\_var2 = function(X, beta1){

Y = 1 + beta1 \* X + ...

...

...

}

t\_stat\_vec = replicate(n = 5000, generateTstat\_unequal\_var2(X,2))

plot(density(t\_stat\_vec),col = 1,main='Estimated density of statistic from simulated data and true density')

x\_grid = seq(-5,5,length.out = 100)

lines(x\_grid,dt(x\_grid,df),col = 2) ## col = 2 means red

cutoff = qt(0.975,df)

mean(abs(t\_stat\_vec)>cutoff)

```

5. Write a new function, say, `generateTstat\_dep\_error`, with same regression function but errors that are not independent. Do this by first generating a vector of errors `error` and then returning a new vector whose first entry is `error[1]` but for $i>1$ the $i$-th entry is `error[i-1] + error[i]`. Repeat (I) and (II) in part (1). Does the $T$-statistic still have close to a $T$ distribution? How often is your $T$ statistic larger than the usual 5% threshold?

```{r}

generateTstat\_dep\_error = function(X, beta1){

error = rnorm(length(X))

error[2:length(X)] = ... # your only need to complete this line.

Y = 1 + beta1 \* X + error

fit = lm(Y~X)

beta1hat = summary(fit)$coefficient[2,1]

se\_beta1hat = summary(fit)$coefficient[2,2]

Tstat = (beta1hat - 2)/se\_beta1hat

return(Tstat)

}

t\_stat\_vec = replicate(n = 5000, generateTstat\_dep\_error(X,2))

plot(density(t\_stat\_vec),col = 1,main='Estimated density of statistic from simulated data and true density')

x\_grid = seq(-5,5,length.out = 100)

lines(x\_grid,dt(x\_grid,df),col = 2) ## col = 2 means red

cutoff = qt(0.975,df)

mean(abs(t\_stat\_vec)>cutoff)

```

6. Summarize your findings in questions 1-5. Which kinds of departures from the assumptions for the error term of the simple linear regression model seem important (or detrimental)?