---

title: 'Lab 3: 2 Numeric EDA & Simple Linear Regression with transformations & Intro to Inference'
date: Fall 2020"
output:

html_document: default

Name: Parth

Skills

- Given a study, distinguish between explanatory and response variables
- Given a set of raw data, make a scatterplot or regression output using R
- Given a scatterplot, identify patterns such as positive and negative associations
- Given the least squares line and a value of x, calculate the predicted value of y
- Given standard regression output, identify and utilize key parts of the output (estimated slope, intercept, R^2^, etc.)
- Given a study, interpret the value of the coefficient of determination (R^2^).
- Be able to describe the primary distinction between regression and correlation analysis.
- If assumptions for linearity fail then use transformations to linearize our data

Optional Skills:

- State and test the assumptions of inference about the regression model
- Given a study objective, significance level, and summary statistics, conduct a formal test of significance on a slope based on ANOVA or the t-distribution by conducting the appropriate steps (including stating hypotheses, calculating test statistics, calculating and interpreting p-values and interpreting the conclusion in context).
- Be able to calculate and interpret the confidence interval for β given output from a linear regression analysis.
- Given standard regression output, interpret the results of the test of hypothesis about the slope.

Please make sure to show all R code and output after each question so that we can see your work. Write a sentence for each numerical value produced describing its meaning **in context with the proper units**. If you don't recall how to do something, you should first refer to your textbook, in-class examples, and the primers from Lab #1 and #2, then reach out to your peers and Abhi or me. Of course the internet is a good resource as well.

Here are some symbols that you might find useful:
H_0 H_1 $\neq$ σ μ μ_0 μ_{word} ρ β_0 β_1 $\hat{y}$ $\hat{y}_{word}$

START HERE: Beloew are the packages you may need to load to calculate statistics, create graphics, calculate power (optional), etc. in the code chunk below. You should start all labs by loading these in as they are the ones we commonly use.
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE, comment = "",warning = FALSE, message = FALSE)
require(rmarkdown, quietly=TRUE)
require(ggformula, quietly=TRUE)
require(pwr)
require(mosaic, quietly)
```

```
*Markdown tip: putting two spaces after a line forces the next line in your markdown document to also be a new line in

*Markdown tip: equations in Mardown will ignore any attempt to include spaces, unless you insert at `~` inside the equa

*Markdown tip: If you want to test out two lines of code, such as for a plot (such as with `qqnorm()` and `qqline()`),

***
```

```
### Scenario 1: Muscle Mass

A person's muscle mass is expected to decrease with age. To explore this relationship in women, a nutritionist randomly


1.a State the statistical research question for your analysis.

> Insert answer here
What linear impact does a woman's age have upon their muscle mass?


1.b Name the response variable of interest and state the type (quantitative or qualitative).  Always include your units

> Insert answer here
Muscle mass (kg), quantitative

1.c Name the explanatory variable of interest and state the type (quantitative or qualitative). Always include your uni

> Insert answer here
Age (yrs), quantitative

1.d TRUE or FALSE? This is an experimental data collection technique (An experiment is where we have randomly assigned
**FALSE**, this is not an experiment. This is an obeservational study with no treatment.


To answer a researcher's questions in the context of the problem (for a specified explanatory and response variable), w

$$Y_{muscle~mass} = \beta_0 + \beta_1 \cdot X_{age} + \epsilon$$
Note:  the $\epsilon$ in the formula is the error associated with the variability of our data, sometimes called the **r

*Statisticians use the above formula but the AP uses and Ti calculators use the formula below*

Note: The AP exam and your Ti calculators use   $$Y_{muscle~mass} = A + B \cdot X_{age}$$



1.e **When doing statistical inference it's very important to always check your assumptions/conditions to make sure the

> Insert your answer here


1.f*(optional)* State the null and alternative hypothesis both symbolically and verbally to test the slope of the linea

**Hint:** The Null hypothesis $H_0$ can be thought of in various ways:  The "status quo", the hypothesis of no change/n
Whereas as the Alternative Hypothesis $H_1$ can be thought of as the hypothesis we are looking to find evidence to supp

> $H_0: \beta_1 = 0$
> The _____ of a linear model between _____ and _____ is zero.
> $H_1: \beta_1 _____ 0$
> The _____ of a linear model between _____ and _____ is _____ zero


1.g Estimate the population regression line using the data (`MuscleMass.csv`). Then, create a scatterplot of the data,

**Hint:**  the function to create a linear model is lm(dataframe$response variable ~ dataframe$explanatory variable)
The function to create a scatterplot is gf_point(dataframe$response variable ~ dataframe$explanatory variable, argument
*remember your units on the axes labels always!*

to pipe use %>%
The function to add a linear model to your scatterplot after using the piping command above is:
gf_lm(response variable ~ explanatory variable, data = dataframe)
```

{r}
muscle<-read.csv("musclemass.csv") # reads in the csv file and names it "muscle" names(muscle) list variables dataframe
mm.age<-***($ ~ $_) # create a linear model of muscle mass as a function of age and assigns it the name MM.age***
***(~ , xlab="", ylab="") 'replace with piping command' # create a scatterplot of muscle mass as a function of age and pipe it to the function below* (~, data =) # add the linear model to the scatterplot**

```
1.h Before drawing any conclusion, we must evaluate the parametric assumptions of linear regression:  **Must always che

#### Assumptions for Linear Regression Inference

*The scatterplot show that using a linear model is reasonable for the data  (your data follows a linear pattern)

*The QQ-Plot shows our values fit the theoretical 1:1 line (the resulting cumulative frequency follows that of the norm

*The Residual plot show no left over pattern and most values are within an even range of the zero line.  Thus our error


Conduct the appropriate testing for the assumptions of regression analysis. Modify the code below to create the additio

Hint:  To make a residual plot put a 1 in the argument after the name of the linear model.  This code extracts the tge
plot(Linear Model, 1)

To make a QQ plot put a 2 in the argument after the name of the linear model.  This code extracts the qqplot to test no
plot(Linear Model, 2)
```

{r}
plot(MM.age, *) # Fitted Values vs. Residuals plot(MM.age,) # QQ plot

```
TRUE or FALSE? The assumption of linearity seems reasonable.
> Plot used:
> Justification:

TRUE or FALSE? The assumption of normally distributed error seems reasonable.
> Plot used:
> Justification:

TRUE or FALSE? The assumption of equal error variance (homoscedacity) seems reasonable.
> Plot used:
> Justification:

TRUE or FALSE? The assumption of the independence of the residuals seems reasonable.
> Plot used:
> Justification:

1.i After checking that we can use a model for the null distribution (i.e.the T-Distributiion to get a t-statistic and

We can use another diagnostic scatterplot of 'Residuals vs Leverage (weight/influence each point holds in the overall l

**Cooks Distance: a measure of each point's importance in determining the regression result.**  `plot(model, 5)`

When looking at the plot, the red contour lines indicate the critical Cook's Distance, which estimates the importance o

Smaller distances from the central cluster of observations means that removing the observations has little effect on th

Distances *larger than one* indicate values that have a high influence on the estimate of the slope and regression anal

Distances between 0.5 and 1 indicate values of concern

We don't want any values to fall past a Cook's distance of 1 (most commonly used value), note:  Cook's distance of 1 wi


Modify the code below to create the plot to check this guideline. Then bold if you think the statement is met or not, l
```

{r}
plot(MM.age, _) # Cook's distance

```
TRUE or FALSE? There are no highly influential points that might alter our analysis and estimate of the slope substanti

> Plot used:
> Justification:
```

```
1.j Modify the code below to create a summary table (the code is summary(Linear Model)), to assess whether there is a s
```

{r}
summary(_)

```
> $\hat{y}_{muscle~mass} = _____ + _____ \cdot x_{age}$


1.k Interpret the slope of the linear model, in context.

> As _____ increases by _____ _____, _____ is expected to _____ by _____ _____.


1.l Interpret the intercept of the linear model, in context. (Provide the mathematical interpretation, whether it makes

> When _____ is zero, we would expect _____ to be _____ _____.


1.m State and then interpret the coefficient of determination.

> R^2^ = _____
> _____ of the variation in _____ is explained by the _____ _____ with _____.


1.n Assess the strength of the relationship using the coefficient of determination. Is age a good predictor of muscle m

> Insert your answer here.


1.o Determine the 95% confidence interval for the slope of the regression line. Modify the code below to determine the

Note:  General Confidence Interval Formula is:  Remember you can hover your cursor over the things within the dollar si
$point estimate \pm standard error$
where $standard error = (critical value) \cdot (margin of error) $

Note: Formula for confidence level using the T-distribution is confint(Linear Model, level = Confidence level in decima

Note:  Our general confidence interval interpretation is this:
**Based on our sample, we are *insert confidence level* confident that the true *insert parameter of interest* is betwe

Note:  A general version specific for Linear Regression is:  **Based on our sample, we are *insert confidence level* co
```

{r}
***(MM.age, level=*)**

```
> Fill in the blanks below for our general confidence interval interpretation for linear regression.

Based on our sample, we are _____ confident that the true _____ of the linear regression between _____ and ____ is betw


1.p Modify the code below to perform a power analysis for your regression model, using an effect size of 0.15 (medium e

# Interpret your power analysis below.

Note:  Power is a good thing, we want as much of it as we can get (because the more power we have the lower our chance

Note:  in the code below leave the blank empty after the 'power=' this tells R to find the power (the thing we're inter

$u = 1$ for simple linear regression

v = residual/error degrees of freedom = sample size n - 2 since we have 2 variables in this case
```

{r}
pwr.f2.test(u=*, v=*, f2=*, sig.level=*, power=_)

> Fill in the blanks below for interpretation of power.

1.q Which of the following statements are true? **Bold** all statements (but not the list-letter) that are true.

a. The individuals were sampled randomly (i.e., a random sample), so it implies an observational study (i.e., cannot de
b. The individuals were not randomly assigned to their value of the explanatory variable, so this is not an experiment
c. The individuals were randomly assigned to the explanatory variable, so this is an experiment, so we can say any chan
d. The individuals were randomly assigned to the explanatory variable, so this is an experiment, but we can only determ
e. The individuals were randomly assigned to their value of the explanatory variable, but we cannot generalize to all i
f. The individuals were randomly selected, so we can generalize to the population.

1.r Can we trust the results of your conclusion? Explain your answer, statistically.

Note: In order to trust our results we must show we've satisfied the conditions so the model (T-distribution in this c

Note:(power analysis not needed (unless using it to determine appropriate sample size) if you reject the null hypothesi

> Insert your answer here.

***

Scenario 2: Digoxin

Digoxin is a commonly used medication for treating domestic cats suffering from congestive heart failure. Its primary b

Researchers conducted a postmortem study of 144 adult domestic cats (over 2kg in weight) that died from symptoms of con

*Data Source: R. A. Fisher (1947) The analysis of covariance method for the relation between a part and the whole, Biom

2.a State the statistical research question of this analysis.

> Insert your answer here

2.b Identify the variable type by completing the table below by filling in the blanks

Variable	Description	Type (Categorical or Numeric)
Bwt	Body Weight (kg)	**___**
Hwt	Heart Weight (g)	**___**
Sex	Sex (F, M)	**___**

The researchers start by completing a linear regression analysis for the heart weight (g) against body weight (kg) in a

2.c State the alternative hypotheses for a regression hypothesis test of body and heart weight, both symbolically and v

> $H_0: \beta_1 = 0$
> The true slope of a linear relationship between body (kg) and heart (g) weight in adult cats is zero.
> $H_1: \beta_1 > 0$

> Insert your symbolic alternative here
> Insert your symbolic alternative here

2.d Modify the code below to perform a power analysis for your regression model, using an effect size of 0.25 (equivale

See the power analysis above we did for more guidance if needed.

{r}
pwr._.test()

> We have enough power (over 80%) to detect a true alternative with an effect size of 0.25.

> We do not have enough power (<80%) to detect a true alternative with an effect size of 0.25.

2.e A useful summary statistic for the relationship between body and heart weight in adult cats would be the correlatio

Note: use this code to calculate correlation coefficient (r), `cor(dataframe$Response , dataframe$explanatory)`

{r}
cats<-read.csv("cats.csv", header="TRUE)" #do not change this code *(, __)*

2.f Suppose the parametric conditions are met for this analysis (usually we will check all these like we did in the fir

Conduct a linear regression analysis on these data. Modify the code below (fill in the blanks/replace) so it will make

{r}
gf_point(_ ~ , data=, xlab="", ylab="*") %>% gf_lm(~ , data=, col=1) replace with piping symbol*
gf_theme(theme_classic())
cats.lm<-lm(_ ~ , *data=cats) summary(__)*

2.g Does the linear model with body weight explain much of the variability in heart weight for adult domestic cats?
How much variability does it explain? Does that seem like a good model to use for predictions? Hint: I'm asking you a

> Insert your answer here.

The researchers notice that the strength of linear relationship may depend on sex of the cat, male and female. To this

{r}

run this code chunk to see the output.

gf_point(Hwt ~ Bwt, data=cats, xlab="Body Weight (kg) of adult domestic cats", ylab="Heart Weight (g)", color= ~ Sex)
%>%
gf_theme(theme_bw())

2.h Write the null and alternative hypotheses for a regression hypothesis on heart weight (g) against body weight (kg)

> $H_{0,~males}$: $\beta_{1,~males} = 0$
> The true slope of a linear relationship between heart weight (g) and body weight for adult male domestic cats is zero
> $H_{1,~male}$:
>

> $H_{0,~female}$:
>
> $H_{1,~female}$:
>

2.i Assess whether the parametric conditions of linear regression have been satisfied, for sex group (males, females);

{r}
#subset your data by sex
m.cats<-subset(cats, Sex == "M")
<-subset(<* em*>, _ == "_") # you try it for the females

condition tests for Males

**male.lm<-lm(_ ~ , *data =) # creates a linear model and assigns it the name male.lm replace with scatterplot*
function(response ~ explanatory, data = dataframe, xlab="", ylab="") %>%
gf_lm(dataframe$response ~ dataframe$explanatory, col=4)
*plot(linear model,) plot(linear model,)***
plot(linear model, _)

condition tests for Females

```
*Male Cats Regression*
Condition/Guideline             | Plot            | Satisfied or Violated?  | Justification
--------------------------------|-----------------|-------------------------|-------------------
 Linearity                      |                 |                         |
 Normality of Residuals         |                 |                         |
 Constant variance of residuals |                 |                         |
 Independence of residuals      |                 |                         |
 No Influential Points with     |                 |                         |
 Cooks Distabce beyond CD=1

Note:  In the future you will need to remember/be able to look up all the conditions (like the five conditions above) f

*Female Cats Regression*
Condition/Guideline             | Plot            | Satisfied or Violated?  | Justification
--------------------------------|-----------------|-------------------------|-------------------
 Linearity                      |                 |                         |
 Normality of Residuals         |                 |                         |
 Constant variance of residuals |                 |                         |
 Independence of residuals      |                 |                         |
 No Influential Points          |                 |                         |
Cooks Distabce beyond CD=1



2.j Create a summary table for the linear regression for each sex (one summary for males and one sumary for females). U
See how we did this in earlier example if needed.

<Insert code chunk here>

> Write linear regression lines here for male and female cats in using math notation $'s!  See earlier problem for exam



2.k Based on our fitted linear model, if we observe a male cat with a body weight of 3.1 kg, what is the cat's estimate


> If a male adult cat's body weight is 3.1 kg, we would expect (or predict) it to have a heart weight of _____ g.

in the future and on the AP you'll be expected to write this sentence out on your own and remember to use the word pred


2.l Which sex has a linear regression with stronger prediction power? Provide and compare relevant statistics (compare

> Insert answer here



2.m TRUE or FALSE? (Bold your answer).  Based on the results of the t-test (specific type of hypothesis test that uses


2.n TRUE or FALSE? (Bold your answer). (Bold your answer). If there is no association between heart and body weight in


2.o Do you trust your results of your regressions? Justify your answer.

> Insert your answer here
```

```
2.p Can we determine if body weight **caused** the heart weight in a cat? Explain why or why not with statistical reaso

> Insert your answer here.


2.q Can we **generalize** our results to all adult cats with congestive heart failure? Explain why or why not with stat

> Insert your answer here.



***

### Scenario 3: Suspended Sediment

Runoff from agricultural fields, urban areas, and construction sites can carry away soil, producing cloudy or muddy wat

You are working as a hydrologist to characterize the relationship between suspended sediment concentrations (mg/L) and
 The data are found in `salinas.csv`.

3.a Test the parametric assumptions and guidelines of linear regression for the simple linear regression of `sediment`
```

{r}
salinas<-read.csv("salinas.csv") # do not change this line of code

```
Condition                | Plot            | Satisfied or Violated?  | Justification
-------------------------|-----------------|-------------------------|-------------------
                         |                 |                         |
                         |                 |                         |
                         |                 |                         |
                         |                 |                         |
                         |                 |                         |

Hint: one or two or the above conditions should not be satisfied (you should be able to tell from your diagnostic plots

Sometimes, the linear model assumptions are satisfied after taking an appropriate transformation of the explanatory var

We will try the most common transformation (Logs) to see if we can transform our data closer to linearity.  More specif


3.b Now assess the parametric assumptions and guidelines of linear regression for the transformed linear regression of

List each assumption, the graph used to assess it, and whether you think the assumption has been satisfied or violated,
Note: I'm calling the transformed linear model sed.log.lm but you could call it whatever you want (it's nice to get a p

Note: Also remember we have to always label the transformed variable on our graphs
```

{r}
sed.log.lm<-lm(dataframe$response~log10(dataframe$explanatory)) # this is our newly transformed linear model

gf_point(dataframe$response~transformation(dataframe$explanatory), xlab=expression("Log of Discharge (ft"^2~"/s / mi"^2~")"), ylab="Sediment Concentrations (mg/L)", title = "Salinas River") %>%
gf_lm(salinas$sediment~log10(salinas$discharge), col=1)
plot(transformed model, *) plot(transformed model,*)
plot(transformed model, _)
```

| Condition | Plot | Satisfied or Violated? | Justification |
|---|---|---|---|
| **Linearity** | | | |
| **Normality of Residuals** | | | |
```

Condition	Plot	Satisfied or Violated?	Justification
Constant variance of residuals			
Independence of residuals			
No Influential Points			
no value beyond CD=1			

Take-home message: Sometimes, the linear model assumptions are satisfied after taking an appropriate transformation of the explanatory variable and/or the response variable. In the data (salinas.csv), the log-base-10-transformation on the explanatory variable led to satisfying linear model assumptions. If you want you could now proceed with your statistical inference using the t-distribution as our probability distribution (density curve) to calculate a p-value