Math 150 - Methods in Biostatistics - Homework 2

Your Name Here

Due: Wednesday, Feb 1, 2023

Assignment Summary (Goals)

- Run a least square regression model, try different transformations on the explanatory and response variables to find a model for which the technical conditions hold.
- Analyze two different datasets using a simulation method (you will need the **infer** package) as well as Fisher's Exact Test
- For plotting and **infer** code, see the class notes describing the Botox study:
  click here to link for boxplots and click here to link for infer for simulating

Q1. Collaborative Learning Describe one thing you learned from someone (a fellow student or mentor) in our class this week (it could be: content, logistical help, background material, R information, etc.) 1-3 sentences.

Q2. Hippel-Lindau disease Eisenhofer et al. (1999) investigated the use of plasma normetanephrine and metanephrine for detecting pheochromocytoma in patients with von Hippel-Lindau disease and multiple endocrine neoplasia type 2. The data set (vonHippelLindau.csv, posted online) contains data from this study on 26 patients with von Hippel-Lindau disease and nine patients with multiple endocrineneoplasia. The variables in the data set are (problem from Dupont, chp 2.22, PubMed article at [http://www.ncbi.nlm.nih.gov/pubmed/10369850]):

Note: the goal is to model `p_ne` (the response variable) from `tumorvol` (the explanatory variable).

variable	units
disease	0: patient has von Hippel-Lindau disease 1: patient has multiple endocrine neoplasia type 2
p_ne	plasma norepinephrine (pg/ml)
tumorvol	tumor volume (ml)

Note: the data this week is imported from the internet, so everyone can use the same link! The directories below do not go to my own computer, they go to a URL pointing to a dataset in the cloud.

```
tumor <- readr::read_csv("http://pages.pomona.edu/~jsh04747/courses/math150/vonHippelLindau.csv")
head(tumor, 3)
```

```
## # A tibble: 3 x 4
##   disease    id  p_ne tumorvol
##     <dbl> <dbl> <dbl>    <dbl>
## 1       0     2  1845      336
## 2       0     3  1734      216
## 3       0     4   739      128
```

(a) Regress plasma norepinephrine against tumor volume. Draw a scatter plot of norepinephrine against tumor volume together with the estimated linear regression curve. What is the slope estimate for

this regression? What proportion of the total variation in norepinephrine levels is explained by the regression?

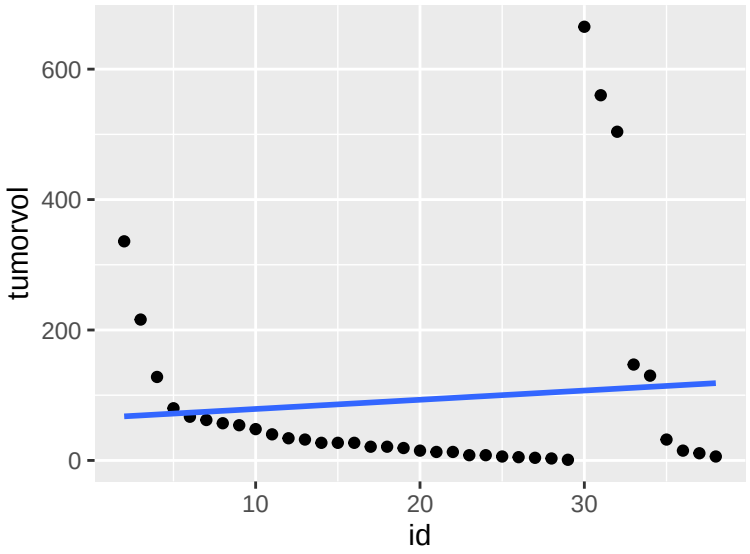
R hints: If the linear model is piped into `tidy()`, the output will be important information on a *per parameter* basis. For example, coefficients, standard errors, etc.

If the linear model is piped into `glance()`, the output will be important information on a *per model* basis. For example, R^2, overall model p-value, model degrees of freedom, etc.

If the linear model is piped into `augment()`, the output will be be important information on a *per observation* basis. For example, residuals (`.resid`), fitted values / predicted values (`.fitted`), etc.

To make a plot in R you want to add a series of layers. The code below is meant as an example, although the variables are totally wrong. Work through the lines of code below and see if you can follow. If you don't follow the lines, ask me!

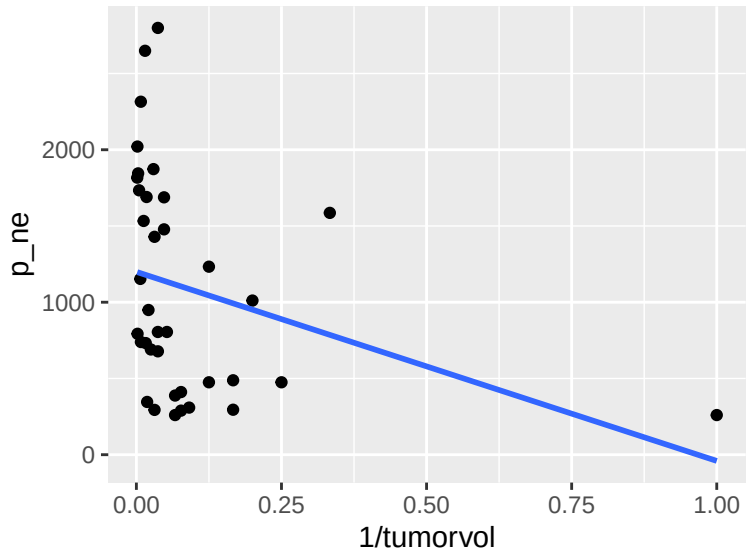
```
tumor %>%                                  # which dataset?
  ggplot(aes(x = id, y = tumorvol)) +      # set up the plot
  geom_point() +                           # add the points
  geom_smooth(method = "lm", se = FALSE)   # add the line a linear model without error bounds
```



(b) Experiment with different transformations of norepinephrine and tumor volume. Find transformations that provide a good fit to a linear model. Report your new linear model. What is your new R^2? Does the R^2 matter in choosing your transformation? Explain.

R hints: First transform one or both of your variables (see pg 49 in your text), then re-plot the data. Below is an example, but it turns out that I made a bad choice of transformation because the plot is terrible. Why (what makes the plot look bad)?

```
tumor %>%
  ggplot(aes(x = 1/tumorvol, y = p_ne)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



Q3. Regression Conditions Which of the following conditions are required to test hypotheses using simple linear regression? If the condition isn't valid, explain why not.

(a) The random variable Y (not conditional on X) is normally distributed.
(b) The variance of Y depends on X.
(c) The random variable Y is normally distributed at each value of X.
(d) The mean of Y (given X) is a linear function of X.
(e) The random variable X is randomly distributed on some scale.

Q4. Chp 6, E1: Cancer and Smoking: Fisher's Exact Test and Simulations Studies Answer the following questions for the data displayed below. Hint: see the class notes for help with the R code. And ask lots of questions!

	lung cancer	healthy	
smoker	41	28	69
non-smoker	19	32	51
	60	60	120

```
smokecancer <- data.frame(act = c(rep("non-smoker", 51), rep("smoker", 69)),
                     outcome = c(rep("lung_cancer", 19), rep("healthy", 32),
                                 rep("lung_cancer", 41), rep("healthy", 28)))
smokecancer %>% table()
```

```
##             outcome
## act          healthy lung_cancer
##   non-smoker      32          19
##   smoker          28          41
```

(a) Was either the explanatory variable (row) or the response (column) variable fixed before the study was conducted?

(b) Is this an example of an experiment or an observational study?

(c) Is this a cross-classification, cohort, or case-control study? Explain.

(d) Created a segmented bar chart for the data.

(e) Create a simulation study to test the one-sided hypothesis that smokers are more likely to have lung cancer. Provide a p-value and state your conclusions.

(f) Use Fisher's exact test to test the one-sided hypothesis that smokers are more likely to have lung cancer. Provide a p-value and state your conclusions.

```
praise()
```

```
## [1] "You are neat!"
```