Math 150 - Methods in Biostatistics - Homework 2

Ethan Ashby

Due: Friday, Feb 12, 2021

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. PodQ Describe one thing you learned from someone in your pod this week (it could be: content, logistical help, background material, R information, etc.) 1-3 sentences.

Annika got a callback from a musical (LEAD ROLE)! YGG! Apparently, they record the singing and stuff prior using special mics, making the performances super unique/special!

Lian just celebrated Chinese New Year with family and lots of good food!

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]):

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)

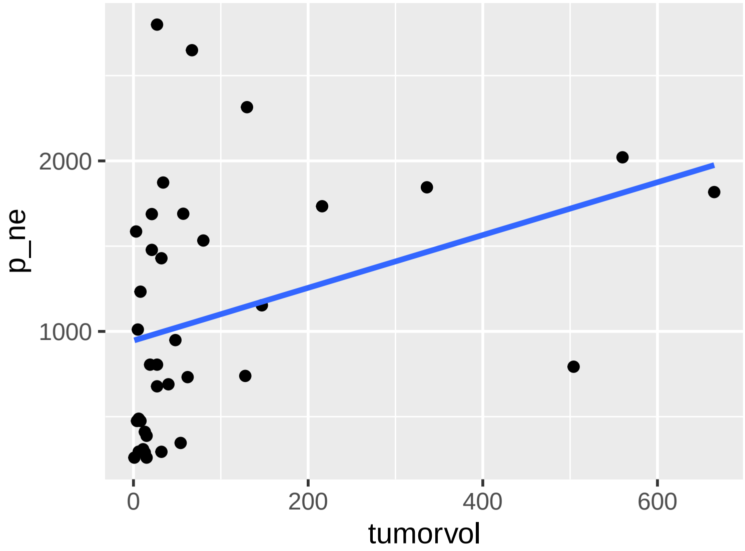
Note: the data this week is imported from the internet, so everyone can use the same link! The directories below do not go to my own computer, they go to a URL pointing to a dataset in the cloud.

```r
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?

```
mod1<-lm(p_ne~tumorvol, data=tumor)
tumor %>%
  ggplot(aes(x = tumorvol, y = p_ne)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



```
#output model params
mod1 %>% tidy()
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic      p.value
##   <chr>          <dbl>     <dbl>     <dbl>        <dbl>
## 1 (Intercept)   946.     130.         7.25 0.0000000181
## 2 tumorvol        1.55     0.708      2.19 0.0356
```

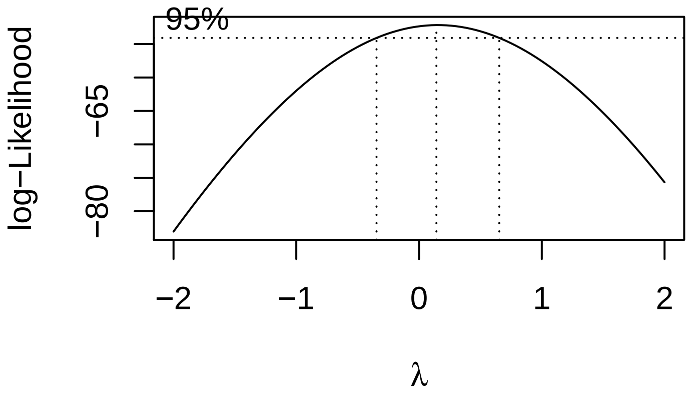
```
#output model R^2
mod1 %>% glance()
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.120        0.0950  685.      4.78  0.0356     1  -293.  592.  597.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

The SLR line is shown in blue amongst the data points. The slope estimate for the regression is 1.55. The variability explained by the regression is quantified using R^2. In this case, the $R^2 = 0.120$, meaning 12% of the variation in the data is explained by the regression.

(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.

```
library(MASS)
MASS::boxcox(mod1)
```



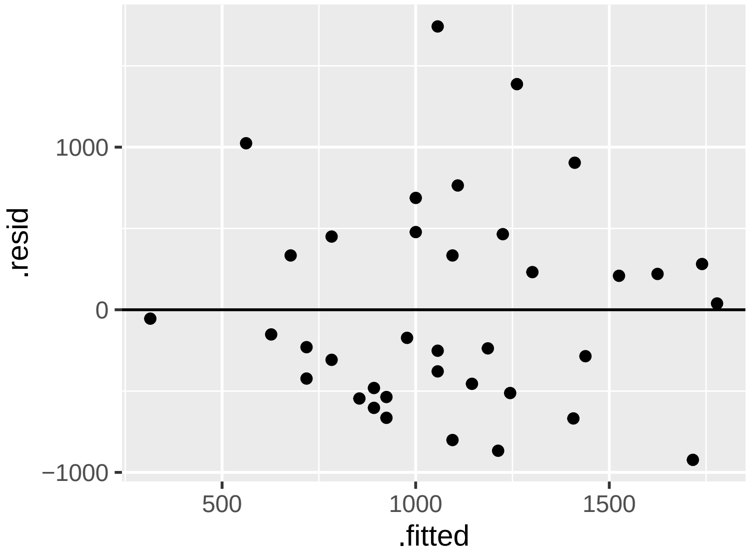
```
#box cox method lambda=0 maximizes log likelihood so we transform the response variable using logarithm

tumor <- tumor %>%
  mutate(logtv = log(tumorvol))

newlm <- lm(p_ne ~ logtv, data = tumor)

newlm %>% tidy()
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)     315.     266.       1.18 0.245
## 2 logtv           225.      70.9      3.18 0.00309
```

```
newlm %>% glance()
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.224         0.202  643.      10.1 0.00309     1  -291.  587.  592.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
newlm %>%
  augment() %>%
  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



The model fit to the log tumor volume is a better fit, evidenced by an increase in R^2 (0.224). The R^2 is useful in choosing our transformation, as it explains how much of the variability in our data is explainable by the model; therefore a model with higher R^2 (and comparable complexity) is going to better describe the data. I'll note that the regression coefficient associated with log tumor volume is 225. The residual plot also shows approximate homoskedasticity and normality of variability about the regression line, two theoretical tenets required for SLR.

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.

NOT required! SLR requires that the random variable is normally distributed **conditional on X**, i.e. the residuals must be normally distributed *about the regression line.* There is no requirement that the data be normally distributed.

(b) The variance of Y depends on X.

FALSE! Variance of Y depending on X is the definition of heteroskedasticity, which is a violation of SLR's requirement of homoskedasticity (or approximately equal variances in Y between different values of X).

(c) The random variable Y is normally distributed at each value of X.

TRUE! This meets the requirement in SLR that the ϵ (noise) terms be normally distributed, with constant variance, and centered at 0.

(d) The mean of Y (given X) is a linear function of X.

TRUE! Per the class notes, SLR requires that the average of the response variable (Y) is a linear function of the explanatory variable (X).

(e) The random variable X is randomly distributed on some scale.

FALSE! In an experimental study, the values of X can be controlled by the researcher! For example, if you were interested in regressing cancer growth in response to quantity of drug administered, the X variable (quantity of drug) is being strictly controlled by the researcher.

Q4. Chp 6, E1: Cancer and Smoking: Fisher's Exact Test and Simulations Studies Answer the following questions for the data displayed below.

	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)))
```

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

The response variables were fixed prior to the study. Richard Doll collected 60 healthy people and 60 lung cancer patients, and observed their smoking behavior.

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

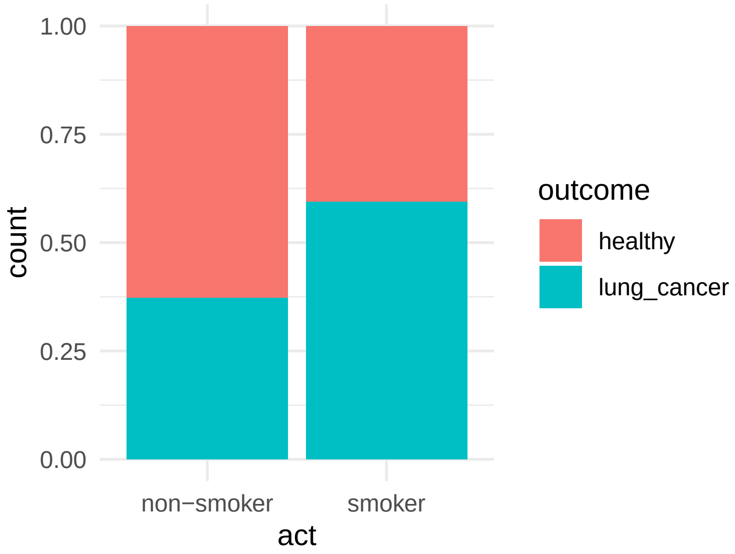
This was an example of an observational study, because the author (Richard Doll) did not choose to allocate patients to treatment (smoking) and control (no smoking) groups.

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

This is a case-control study, because Doll selected people for his study according to their disease status (lung cancer/healthy), i.e. his response variable of interest. On the other hand, a cohort study selects people based on exposure status, and a cross-classification study is a survey, often using a random sampling strategy.

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

```
ggplot(smokecancer)+geom_bar(aes(x=act, fill=outcome), position = "fill")+theme_minimal()
```



The bars are grouped according to the response variable (sampling criterion), and the shaded region of the bars indicate non-smokers while the light bars represent the number of smokers.

(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.

```
set.seed(47)
tab_true<-table(smokecancer)
true_OR<-(tab_true[2,2]/tab_true[2,1])/(tab_true[1,2]/tab_true[1,1])

null_OR<-vector(length=1000)
for(i in 1:1000){
  permsmokecancer<-data.frame("act"=sample(smokecancer$act, replace=FALSE), "outcome"=smokecancer$outcom
  tab_null<-table(permsmokecancer)
  null_OR[i]<-(tab_null[2,2]/tab_null[2,1])/(tab_null[1,2]/tab_null[1,1])
}

#p-value is which null odds ratios are as or more extreme

paste("p-val:", length(which(null_OR>=true_OR))/1000)
```

```
## [1] "p-val: 0.013"
```

This simulation permutes the smoking label (smoker, non-smoker) among the disease labels, creating a null condition of independence between smoking status and cancer status. Then odds ratios are calculated and compared to the null distribution. Out of 1000 null odds ratios, only 13 exceeded our true value, producing a p-value of 0.013. We conclude that the odds ratio is significant, and your odds of getting lung cancer if you smoke are significantly higher than if you don't.

(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.

```
smokecancer %>%
  table() %>%
  fisher.test()
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  .
## p-value = 0.02625
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.101151 5.560219
## sample estimates:
## odds ratio
##   2.447101
```

The p-value of 0.026 indicates that the true odds ratio bewteen smoking and lung cancer is significantly different from 1, meaning that true odds of getting cancer if you smoke are significantly greater than your odds of cancer if you don't smoke.

```r
praise()
```

```
## [1] "You are cat’s meow!"
```