PSTAT 127, Spring 2022, Homework 3 - Due uploaded to Gradescope by 11:59 pm on Friday April 22.

Throughout this homework, you may state and use any results you know from pre-requisite classes (PSTAT 120A, 120B, 126), and 127. Clear working must be shown to receive credit.

1. Consider the following data set on alcohol and tobacco spending from Moore, David S., and George P. McCabe (1989). Introduction to the Practice of Statistics, p. 179. Original source: Family Expenditure Survey, Department of Employment, 1981 (British official statistics)

```
Description: Average weekly household spending, in British pounds, on tobacco
products and alcoholic beverages for each of the 11 regions of Great Britain.

Number of cases:
11

Variable Names:

   1. Region: Region of Great Britain
   2. Alcohol: Average weekly household spending on alcoholic beverages in pounds
   3. Tobacco: Average weekly household spending on tobacco products in pounds

The Data:

Region           Alcohol Tobacco
North            6.47    4.03
Yorkshire        6.13    3.76
Northeast        6.19    3.77
East Midlands    4.89    3.34
West Midlands    5.63    3.47
East Anglia      4.52    2.92
Southeast        5.89    3.20
Southwest        4.79    2.71
Wales            5.27    3.53
Scotland         6.08    4.51
Northern Ireland 4.02    4.56
```

Use R to do the following, and include both your R code and the results within your homework 3 answer file. - in pdf format.

(a) Plot these data.

(b) Fit a simple Gaussian homoskedastic regression of alcohol spending on tobacco spending (include both an intercept and slope).

(c) Calculate the leverage values.

(d) Calculate the Cook's distance measures, and identify which region has the highest Cook's distance value.

(e) What does the Cook's distance indicate in this data set?

(f) Can you think of a geographic reason that might explain what you are seeing, and further research that you would do about geographic variation?

(g) Comment on the sensitivity of your regression coefficients to the point with highest Cook's distance measure. (Fit the model both with and without that point, and comment on how your coefficients and fitted values change.)

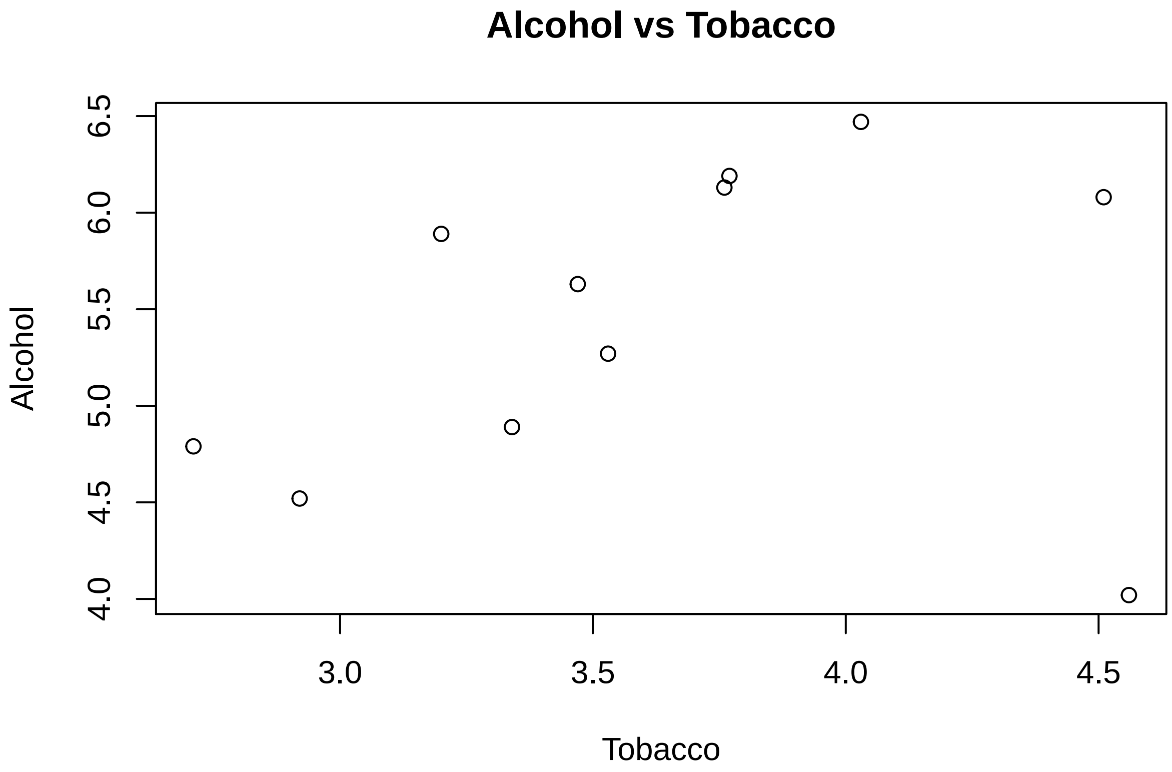
(h) Plot the observations with your fitted line superimposed for each of these fits from the previous part. The plot for each of these fits may be a separate panel (i.e., two panels total), but use R commands to control the axis ranges of the adjacent panels to be the same so that you can visually compare the slopes of your fitted lines for your two fits.

#1

```r
#print dataset
data = read.delim('alcohol x tobacco.txt')
data
```

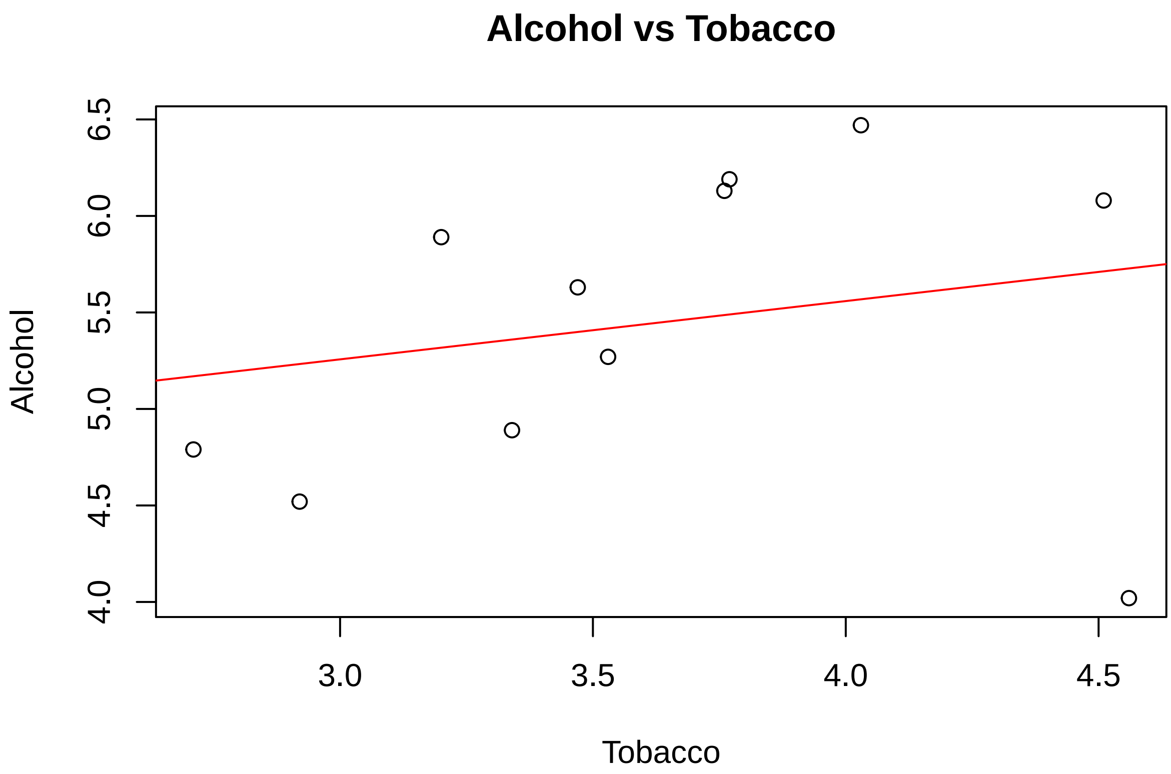
```
##                Region Alcohol Tobacco
## 1               North    6.47    4.03
## 2           Yorkshire    6.13    3.76
## 3           Northeast    6.19    3.77
## 4       East Midlands    4.89    3.34
## 5       West Midlands    5.63    3.47
## 6         East Anglia    4.52    2.92
## 7           Southeast    5.89    3.20
## 8           Southwest    4.79    2.71
## 9               Wales    5.27    3.53
## 10           Scotland    6.08    4.51
## 11 Northern Ireland      4.02    4.56
```

```r
#Part a)
Alcohol = data$Alcohol
Tobacco = data$Tobacco
plot(Tobacco,Alcohol, main = "Alcohol vs Tobacco")
```



```r
#Part b)
fit1 <- lm(Alcohol ~ Tobacco)
```

```
plot(Tobacco,Alcohol, main = "Alcohol vs Tobacco")
abline(fit1$coefficients[1],fit1$coefficients[2], col= 'red')
```



```
summary(fit1)
```

```
##
## Call:
## lm(formula = Alcohol ~ Tobacco)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7080 -0.4245  0.2311  0.6081  0.9020
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.3512     1.6067   2.708   0.0241 *
## Tobacco       0.3019     0.4388   0.688   0.5087
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8196 on 9 degrees of freedom
## Multiple R-squared:  0.04998,    Adjusted R-squared:  -0.05557
## F-statistic: 0.4735 on 1 and 9 DF,  p-value: 0.5087
```

```r
#part c)
hatvalues(fit1)
```

```
##          1          2          3          4          5          6          7 
## 0.13951228 0.09667301 0.09751452 0.11308652 0.09720189 0.23060730 0.14102598 
##          8          9         10         11 
## 0.32728291 0.09313759 0.31884167 0.34511633 
```

There is no extreme leverage values in the data.

```r
#Part d).
cooks.distance(fit1)
```

```
##           1           2           3           4           5           6 
## 0.114101051 0.036517838 0.043728951 0.023600304 0.004740759 0.147326647 
##           7           8           9          10          11 
## 0.046646563 0.077488350 0.001821694 0.068921892 1.747233521 
```

From the results we can see that region 11, which corresponds to North Ireland, has the highest Cook's distance value.

#Part e).

In the dataset, the first 10 regions have decent Cook's distance values, but the last region, North Ireland has an extremely large Cook's distance.This indicates that North Ireland could be a highly influential point in our dataset and our estimated regression could change immensely if we were to compare the linear regression fit with North Ireland included versus a linear regression where North Ireland is not included in the dataset.

#Part f).

Geographically, the first 10 regions are in Great Britain but the last region is in North Ireland. This minor change in the last observation being outside of the Great Britain islands could be assiociated with sustanstially different personal choices among its citizens. Further research that can be done is by considering an explanatory variable like if the observations are collected specific to which subregions they are in.

```r
#Part g).

Alcohol = data$Alcohol
Tobacco = data$Tobacco
fit1 <- lm(Alcohol ~ Tobacco)
summary(fit1)
```

```
## 
## Call:
## lm(formula = Alcohol ~ Tobacco)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1.7080 -0.4245  0.2311  0.6081  0.9020 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept)   4.3512     1.6067   2.708   0.0241 *
## Tobacco       0.3019     0.4388   0.688   0.5087
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8196 on 9 degrees of freedom
## Multiple R-squared:  0.04998,    Adjusted R-squared:  -0.05557
## F-statistic: 0.4735 on 1 and 9 DF,  p-value: 0.5087
```

```
data_alt = data[-c(11),]
alcohol = data_alt$Alcohol
tobacco = data_alt$Tobacco
fit2 <- lm(alcohol ~ tobacco)
summary(fit2)
```

```
##
## Call:
## lm(formula = alcohol ~ tobacco)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51092 -0.42434  0.06056  0.34406  0.62991
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.0412     1.0014   2.038  0.07586 .
## tobacco       1.0059     0.2813   3.576  0.00723 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.446 on 8 degrees of freedom
## Multiple R-squared:  0.6151, Adjusted R-squared:  0.567
## F-statistic: 12.78 on 1 and 8 DF,  p-value: 0.007234
```

fit1, which incorporates region 11 (North Ireland), has an almost 0 Adjusted R-squared value showing insignificance of explanatory variables.

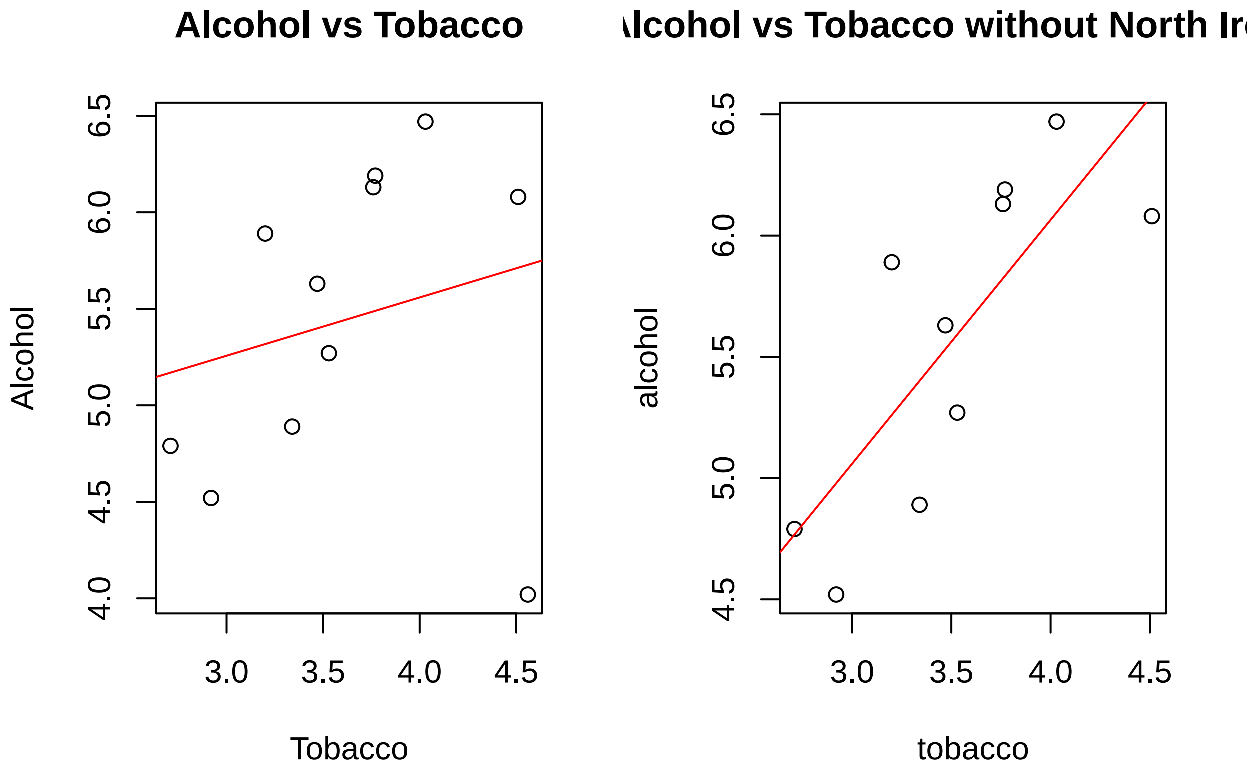
In fit2, which does not incorporate region 11 (North Ireland), has a completely changed Adjusted R-squared value of 0.567 showing a sigificance in an explanatory variable. P-value = 0.007 $<$ 0.05 shows the Tobacco variable is significant in the data.

North Ireland essentially influences whether the data has a signifncant or insignificant explanatory variable and as the data changes completely when it is removed.

```
#Part h).

par(mfrow=c(1,2))
plot(Tobacco,Alcohol, main = "Alcohol vs Tobacco")
abline(fit1$coefficients[1],fit1$coefficients[2], col= 'red')

plot(tobacco,alcohol, main = "Alcohol vs Tobacco without North Ireland")
abline(fit2$coefficients[1],fit2$coefficients[2], col= 'red')
```


Alcohol vs Tobacco
Alcohol
Tobacco
4.0
4.5
5.0
5.5
6.0
6.5
3.0
3.5
4.0
4.5
lcohol vs Tobacco without North Ir
alcohol
tobacco
4.5
5.0
5.5
6.0
6.5
3.0
3.5
4.0
4.5

2. $P(Y=y) = (1-p)^y p$ if $y \in \{0,1,2,3,\dots\}$

a) pmf in natural exponential family has a density of Y_i given by:

always pvt

$$f(y_i \mid \theta_i, \underbrace{\phi}_{\text{constant}}) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{\underbrace{a_i(\phi)}_{\text{constant}}} + c(y_i, \phi) \right\}$$

* we know for a geometric distribution $a_i(\phi) = 1$

Therefore,

$$P(Y=y) = \exp\left[\ln(1-p)^y p\right]$$

$$= \exp\left[\underbrace{y}_{y_i} \cdot \underbrace{\ln(1-p)}_{\theta_i} + \underbrace{\ln(p)}_{b(\theta)}\right]$$

$$\boxed{P(Y=y) = \exp\left[y \cdot \ln(1-p) + \ln(p)\right]}$$

2. Consider a single random variable Y with probability mass function Geom.

$$P(Y = y) \quad = \quad (1-p)^y p\,, \qquad \text{if } y \in \{0, 1, 2, 3, \ldots\}$$

for $p \in (0, 1)$. $Y \sim geom(p)$

(a) Write this pmf in the natural exponential family form used in PSTAT 127, showing clear working.

↪ on scratch paper

(b) Answer the following using your answer to part (2a):

i. Write down the canonical parameter θ in terms of p: $\boxed{\theta = \ln(1-p)}$

ii. Now find p in terms of θ: (i.e., write p as a function of θ) $\underset{e}{\theta} = \underset{e}{\cancel{\ln}}(1-p) \Rightarrow e^{\theta} = 1-p \Rightarrow \boxed{p = 1-e^{\theta}}$

iii. $b(\theta) = -\ln(p) = -\ln(1-e^{\theta})$

iv. $\phi = 1$, always for geometric

v. $a(\phi) = 1$, a constant given $\phi = 1$

vi. $c(y, \phi) = 0$

---

3. Explain why the logistic regression model for $Y_i \overset{indep}{\sim}$ Bernoulli(p_i) for $i \in \{1, \ldots, n\}$ reads

$$logit(p_i) = \boldsymbol{x_i^T \beta}$$

instead of

$$logit(y_i) = \boldsymbol{x_i^T \beta}$$

As part of your answer, explain how the logistic regression model preserves the parameter restrictions that $p_i \in (0, 1)$ if $Y_i \sim$ Bernoulli(p_i) .

Hint: within your answer also explain whether or not the latter can even be calculated when GLM random component is $Y_i \overset{indep}{\sim}$ *Bernoulli*(p_i) *for* $i \in \{1, \ldots, n\}$

---

4. Review hypothesis tests, confidence intervals, and Gaussian Linear Model selection approaches studied in 120B-126. Nothing needs to be handed in for this part.

3) we modeled $Y_i \overset{ind}{\sim}$ Bernoulli(π_i) in lecture 4 of class. In this problem, let $\pi_i = p_i$.

We know, $\pi_i = p_i = \dfrac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}$

our logit function for bernoulli is

$$g(\pi_i) = g(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

where the logit function exists such that $0 < p_i < 1$.

π_i is a probability, so $\log\left(\frac{\pi}{1-\pi}\right)$ is the log(odd).

where odds of an event $A = \dfrac{\Pr(A)}{1-\Pr(A)}$

Also, since $p_i = \dfrac{ex(X_i'\beta)}{1+exp(X_i'\beta)}$

we can say logit function,

$$g(p_i) = log\left(\frac{p_i}{1-p_i}\right) = log\left[\frac{\left(\frac{exp(X_i'\beta)}{1+exp(X_i'\beta)}\right)}{\left(\frac{1}{1+exp(X_i'\beta)}\right)}\right]$$

$$= log[exp(X_i'\beta)] = X_i'\beta$$

Thuse the logistic regression says thet $y_i \sim Bernoulli(p_i)$

where $log\left(\dfrac{p_i}{1-p_i}\right) = X_i'\beta$

↓

The logit links the mean of y_i (i.e. p_i) to the linear predictor $x_i'\beta$

So,

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = x_i'\beta$$

$$\rightarrow \quad \text{logit}(p_i) = x_i'\beta$$