PSTAT127 | Fall 2024 | Homework 1

Billy Dang

2024-10-07

```
library(dplyr)
library(tinytex)
library(tidyverse)
library(ggplot2)
library(ggthemes)
```

Question 1

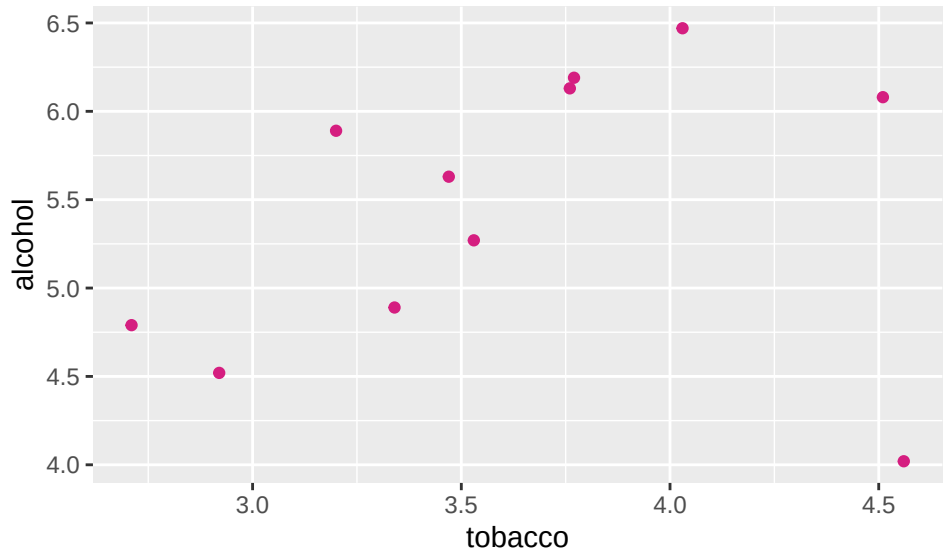
Part 1

Plot the data with alcohol on the vertical axis, and tobacco on the horizontal axis (scatterplot). Present your code here:

```
#Create data frame to prep for plotting

spending_data <- data.frame(
  region = c("North", "Yorkshire", "Northeast", "East Midlands", "West Midlands", "East Anglia", "South
  alcohol = c(6.47, 6.13, 6.19, 4.89, 5.63, 4.52, 5.89, 4.79, 5.27, 6.08, 4.02),
  tobacco = c(4.03, 3.76, 3.77, 3.34, 3.47, 2.92, 3.20, 2.71, 3.53, 4.51, 4.56)
)

#Create scatterplot
ggplot(spending_data,
       aes(x = tobacco, y = alcohol)) + geom_point(color = "#D51E80")
```



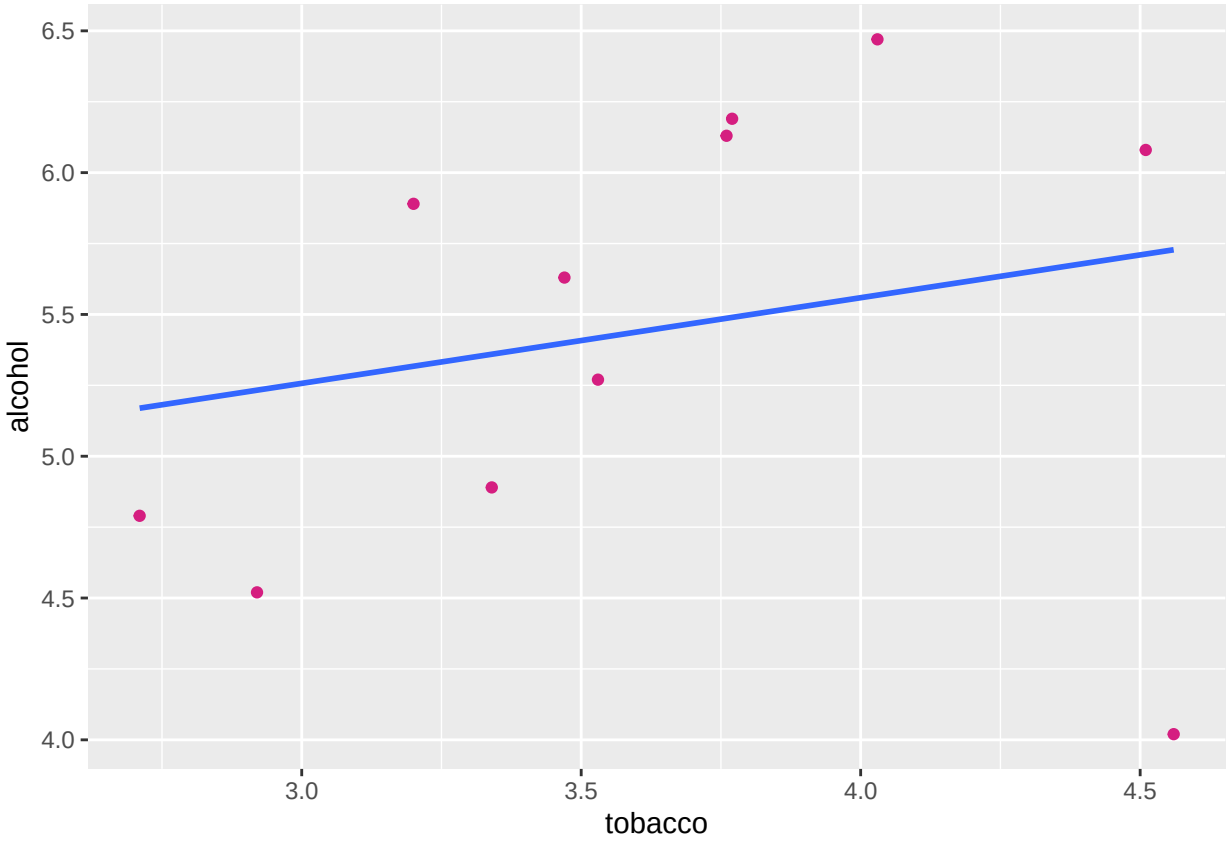
Part B

Fit a simple Gaussian homoskedastic linear regression of alcohol spending on tobacco spending (include both an intercept & slope) and add the fitted line to the scatter plot produced above.

```
#Fitting the model and examining results
linear_model <- lm(alcohol ~ tobacco, data = spending_data)
```

```
ggplot(spending_data,
       aes(x = tobacco, y = alcohol)) +
  geom_point(color = "#D51E80") +
  geom_smooth(method = 'lm', se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Part C

Write down corresponding model and assumptions, clearly defining notation. You may write this model in the way you did in PSTA 126 - using additive error with assumptions on the random variables. (**The underlying model & assumptions not the resulting numeric equation for the fitted values**)

$H_0 : \beta_1 = 0$ - Tobacco spending does not have a significant effect on alcohol spending.

$H_1 : \beta_1 \neq 0$ - Tobacco spending does have a significant effect on alcohol spending.

Y_i is the dependent (response) variable, in this case, the "Alcohol" spending for observations i.

X_i is the independent (predictor) variable, which is the "Tobacco" spending for observations i.

β_0 is the intercept, which represents the expected tobacco spending when alcohol spending (X) is 0

β_1 is the slope coefficient, representing the change in Y for a one-unit change in X (Alcohol spending)

ϵ is the error term for observation i

The formula of our simple linear regression model is shown below:

$$Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i, i = 1, 2, ..., n$$

Assumptions: - e_i are uncorrelated - e_i aka the errors are independent of each other - $Var(e_i) = \sigma^2$ - $E[e_i] = 0$

Part D

```
summary(linear_model)
```

```
##
## Call:
## lm(formula = alcohol ~ tobacco, data = spending_data)
##
## 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
```

$H_0 : \beta_1 = 0$ - Tobacco spending does not have a significant effect on alcohol spending.

$H_1 : \beta_1 \neq 0$ - Tobacco spending does have a significant effect on alcohol spending.

t = .688, t(.01,2,9) = 3.249 Reject H_0 if |t| > 3.2498 p = .5087 > alpha

1. We fail to reject the null hypothesis in favor of the idea that the effect of tobacco spend on alcohol is not significant

$H_0 : \beta_1 = 0$ - The model is not significant

$H_1 : \beta_1 \neq 0$ - Not the null hypothesis

F = .4735 F(.01,1,9) = 11.26 Reject H_0 if F > 11.26 p = .5027 > alpha

2. We fail to reject the null hypothesis as there is not enough evidence to suggest the model is significant

Part E

Obtain the leverage values using any approach you studied in PSTAT 126 (Make sure you review the formulae for the future, and review how leverage values are used in linear model diagnostics)

```
#Where values greater than (2p/n) are considered leverage values
leverage <- hatvalues(linear_model)
print(leverage)
```

```
##          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
```

Part F

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

```
cooks_list = cooks.distance(linear_model)
print(cooks_list)
```

```
##           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
```

The regions with the highest Cook's distance is *Northern Ireland* with a Cook's distance of ~2.20865.

Part G

What do the Cook's distances indicate in this data set, i.e. do any points seem potentially influential requiring further investigation?

The Cook's distances indicate that there is a potentially influential point within the dataset, particularly the 11th one where alcohol spending was only ~4 when tobacco spending was ~4.5

Part H

Can you think of a brief geographic reason that might explain what you are seeing, and any further research that you would do about geographic variation in smoking and drinking practices?

There may be significant variation due to Northern Ireland being separated from the rest of the UK, leading to difference in smoking and drinking practices (and presumably others as well). Other influences on this variation can be further explored by examining socioeconomic data / lifestyle studies.

Part I

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, fitted values, and hypothesis test decisions change.)

```
#remove influential point
spending_data_2 <- data.frame(
  region = c("North", "Yorkshire", "Northeast", "East Midlands", "West Midlands", "East Anglia", "South
  alcohol = c(6.47, 6.13, 6.19, 4.89, 5.63, 4.52, 5.89, 4.79, 5.27, 6.08),
  tobacco = c(4.03, 3.76, 3.77, 3.34, 3.47, 2.92, 3.20, 2.71, 3.53, 4.51)
)

linear_model_2 <- lm(alcohol ~ tobacco, data = spending_data_2)
```

```
summary(linear_model)
```

```
##
## Call:
## lm(formula = alcohol ~ tobacco, data = spending_data)
```

```
##
## 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
```

```
summary(linear_model_2)
```

```
##
## Call:
## lm(formula = alcohol ~ tobacco, data = spending_data_2)
##
## 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
```

Our model was extremely sensitive to the point with the highest Cook's distance measure as the intercept dropped by 2 units after removing Northern Ireland, meanwhile the slope increase by ~.7 as well. In the full model with Northern Ireland, the smaller slope implies that tobacco spending has a weaker effect on alcohol spending. Additionally, the removal of Northern Ireland changes the decision of failing to reject the null hypothesis (no relationship) to rejecting the null hypothesis.

Part J

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 or on the same panel. If using two panels, specify R options to control the axis ranges of the adjacent panels to be the same so you can compare the lines.

```
ggplot(spending_data, aes(x = tobacco, y = alcohol)) +
  geom_point() +
  geom_smooth(method = 'lm', color = 'blue', se = FALSE) +
  geom_smooth(data = spending_data_2, method = 'lm', color = 'red', se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```