Home Work 5

Miles Tweed

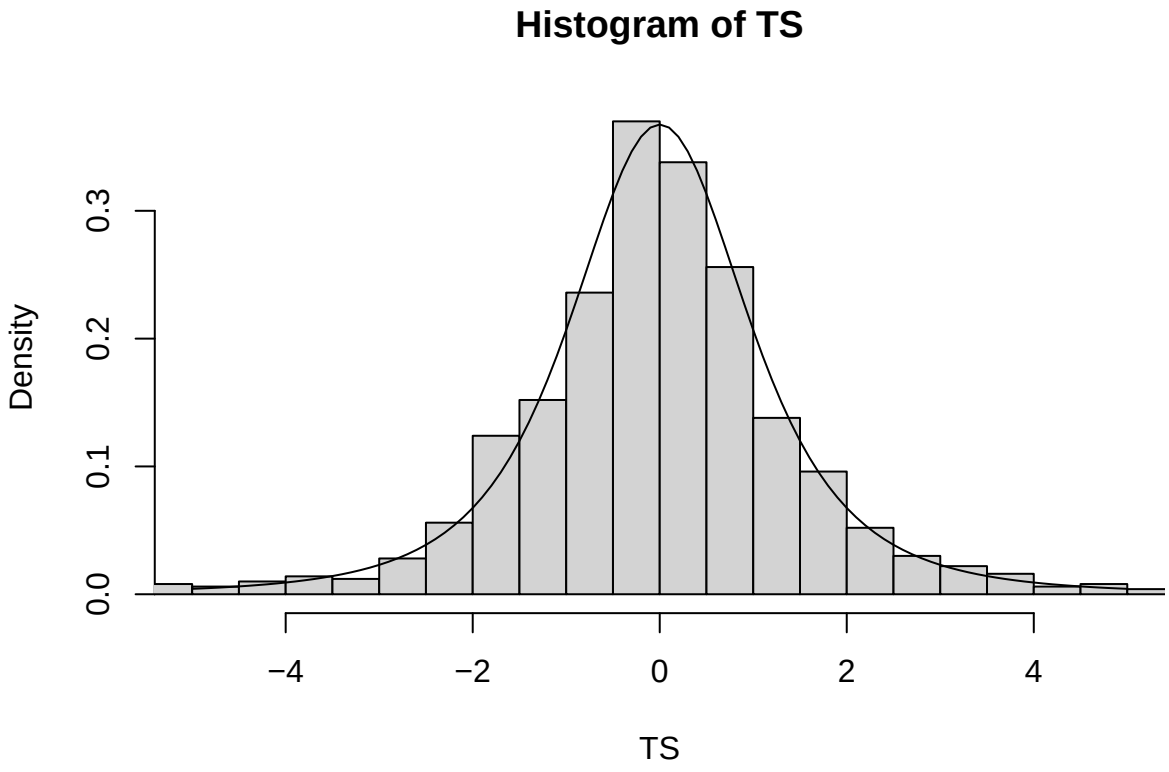
3/14/2021

Problem 1

```
set.seed(123)

X <-  runif(5, -50, 50)

TS <- numeric(1000)

for (i in 1:1000) {
  y.hat <- vector()
  for (x in X){
    y.hat <- append(y.hat, (2 + 3*x + rnorm(1, mean=0, sd=10**2)))
  }
  lm.obj <- lm(y.hat~X)
  b1.hat <- lm.obj$coefficients[2]
  b1.SE <- summary(lm.obj)$coefficients[2,2]

  TS[i] <- (b1.hat - 3)/b1.SE
}

# Mean of TS
mean(TS)
```

```
## [1] -0.1030687
```

```
hist(TS, breaks = 100, xlim=c(-5,5), freq=FALSE)
curve(dt(x,df=3), from = -5, to = 5, add=TRUE)
```



Problem 2

```
set.seed(123)

x = runif(200, min=-50, max=50)


conf_int_b0 <- matrix(0, nrow=1000, ncol=2)
conf_int_b1 <- matrix(0, nrow=1000, ncol=2)

for (i in 1:1000) {
  y.hat <- vector()
  for (x in X){
    y.hat <- append(y.hat, (2 + 3*x + rnorm(1, mean=0, sd=10**2)))
  }
  lm.obj <- lm(y.hat~X)
  conf_int <- confint(lm.obj, level=0.90)
  conf_int_b0[i,] <- conf_int[1,]
  conf_int_b1[i,] <- conf_int[2,]
}


print(mean(conf_int_b0[,1] < 2 & 2 < conf_int_b0[,2]))
```

```
## [1] 0.892
```

```
print(mean(conf_int_b1[,1] < 3 & 3 < conf_int_b1[,2]))
```

```
## [1] 0.904
```

These percentages (89.2% and 90.4%) are what I would expect from a 90% confidence interval. I would expect the true values of β_0 and β_1 to fall within the confidence interval approximately 90% of the time.

Problem 3

Part 1

```
ads <- read.csv('../../Data/Advertising.csv')

rad.lm.obj <- lm(sales~radio, data=ads)
news.lm.obj <- lm(sales~newspaper, data=ads)

# Confidence intervals for sales~radio
confint(rad.lm.obj)
```

```
##                 2.5 %     97.5 %
## (Intercept) 8.2015885 10.4216877
## radio       0.1622443  0.2427472
```

```
# Confidence intervals for sales~newspaper
confint(news.lm.obj)
```

```
##                   2.5 %      97.5 %
## (Intercept) 11.12595560 13.57685854
## newspaper    0.02200549  0.08738071
```

For sales~radio

Intercept: We can be 95% confident that between 8,202 and 10,422 items will be sold on average if $0k is spent on radio advertisement. This is averaged across all markets where $0k is spent on radio advertisement.

Slope: We can be 95% confident that sales will increase by between 162 and 243 items on average for every additional $1000 spent on radio advertisement. This is averaged across all markets where expenditure on radio advertisement differs by $1000.

For sales~newspaper

Intercept: We can be 95% confident that between 11,126 and 13,577 items will be sold on average if $0k is spent on newspaper advertisement. This is averaged across all markets where $0k is spent on newspaper advertisement.

Slope: We can be 95% confident that sales will increase by between 22 and 87 items on average for every additional $1000 spent on newspaper advertisement. This is averaged across all markets where expenditure on newspaper advertisement differs by $1000.

Part 2

```
# Number of items predicted for $20,000 spent on radio advertisement
predict(rad.lm.obj, newdata = data.frame({radio=20})) * 1000
```

```
##        1
## 13361.55
```

```
# Number of items predicted for $20,000 spent on newspaper advertisement
predict(news.lm.obj, newdata = data.frame({newspaper=20})) * 1000
```

```
##        1
## 13445.27
```

Radio prediction

We would expect to sell 13,362 items on average if $20,000 is spent on radio advertisement. This is averaged across all markets where $20,000 is spent on radio advertisement.

Newspaper prediction

We would expect to sell 13,445 items on average if $20,000 is spent on newspaper advertisement. This is averaged across all markets where $20,000 is spent on newspaper advertisement.

Part 3

```
# Confidence bands of prediction for $20,000 spent on radio advertisement
predict(rad.lm.obj, newdata = data.frame({radio=20}), int='c')[,c('lwr','upr')]
```

```
##      lwr      upr
## 12.75114 13.97197
```

```
# Confidence bands of prediction for $20,000 spent on newspaper advertisement
predict(news.lm.obj, newdata = data.frame({newspaper=20}), int='c')[,c('lwr','upr')]
```

```
##      lwr      upr
## 12.65579 14.23474
```

Radio Confidence Band

We are 95% confident that when $20,000 is spent on radio advertisement the average sales will be between 12,751 and 13,972 items.

Newspaper Confidence Band

We are 95% confident that when $20,000 is spent on newspaper advertisement the average sales will be between 12,656 and 14,235 items.

Part 4

```
# Prediction bands of prediction for $20,000 spent on radio advertisement
predict(rad.lm.obj, newdata = data.frame({radio=20}), int='p')[,c('lwr','upr')]
```

```
##       lwr       upr
##  4.909218 21.813889
```

```
# Prediction bands of prediction for $20,000 spent on newspaper advertisement
predict(news.lm.obj, newdata = data.frame({newspaper=20}), int='p')[,c('lwr','upr')]
```

```
##       lwr       upr
##  3.371825 23.518713
```

Radio Prediction Band

When $20,000 is spent on radio advertisement, 95% of all sales will be between 4,909 and 21,814 items.

Newspaper Prediction Band

When $20,000 is spent on newspaper advertisement, 95% of all sales will be between 3,372 and 23,519 items.

Part 5

The prediction band is wider because it is trying to capture 95% of all possible values for a specific expenditure in advertisement whereas the confidence band is trying to capture the average response for a particular advertisement expenditure.

Problem 4

```
crime <- read.csv('../../Data/fl_crime.csv')
names(crime) <- c("county","crime","education","urban","income")
```

Part 1

a)

The full modelling equation is:

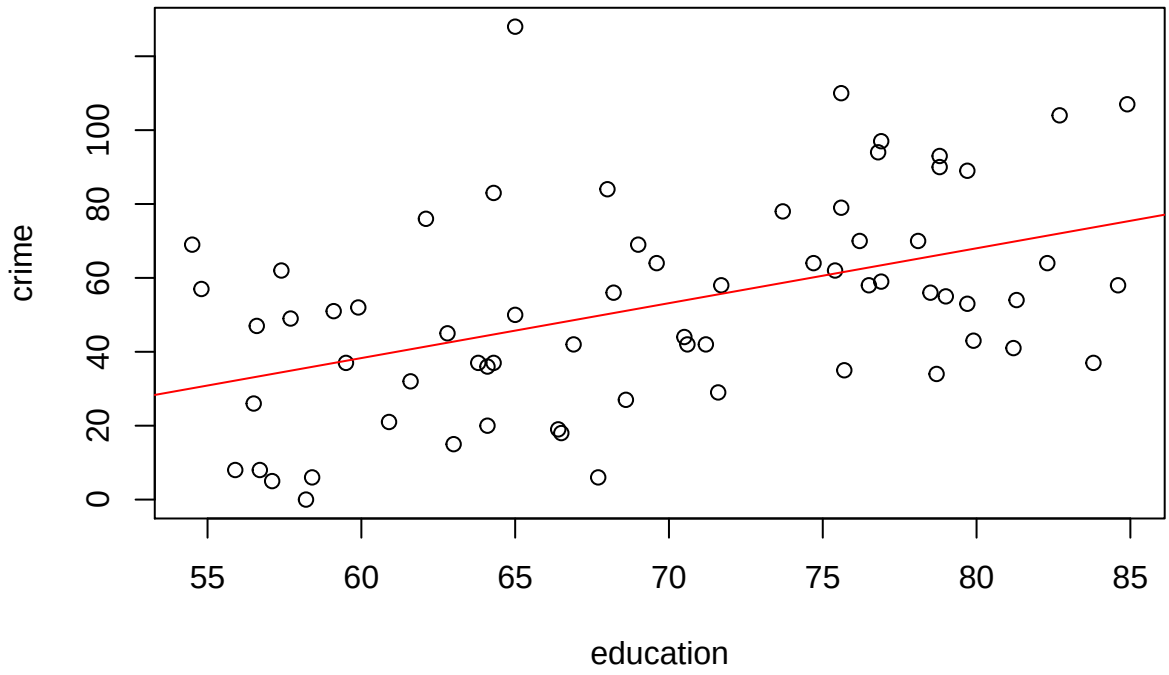
$$\text{crime}_i = \beta_0 + \beta_1 \cdot \text{education}_i + \epsilon_i,\ \epsilon_i \sim_{i.i.d.} \mathbb{N}(0, \sigma^2),\ i = 1, 2, \ldots, n$$

b)

```
ed.lm.obj <- lm(crime ~ education, data=crime)
summary(ed.lm.obj)
```

```
##
## Call:
## lm(formula = crime ~ education, data = crime)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -43.74 -21.36  -4.82  17.42  82.27
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -50.8569    24.4507  -2.080   0.0415 *
## education     1.4860     0.3491   4.257 6.81e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.12 on 65 degrees of freedom
## Multiple R-squared:  0.218,  Adjusted R-squared:  0.206
## F-statistic: 18.12 on 1 and 65 DF,  p-value: 6.806e-05
```

```
with(crime, plot(crime~education))
abline(ed.lm.obj, col='red')
```



c)

The effect of education on crime is statistically significant according to this model (p-value = 6.81e-05). The interpretation would be that for every percent increase of residents aged at least 25 in the county with at least a high school diploma the number of crimes in the past year per 1000 residents will increase by 1.49 on average across all counties where this education metric differs by one.

Part 2

a)

The full modelling equation is:

$$\text{crime}_i = \beta_0 + \beta_1 \cdot \text{education}_i + \beta_2 \cdot \text{urbanization}_i + \epsilon_i,\ \epsilon_i \sim_{i.i.d.} \mathbb{N}(0, \sigma^2),\ i = 1, 2, \ldots, n$$

b)

```
ed.urb.lm.obj <- lm(crime ~ education + urban, data=crime)
summary(ed.urb.lm.obj)
```

```
##
## Call:
## lm(formula = crime ~ education + urban, data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.693 -15.742  -6.226  15.812  50.678
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  59.1181    28.3653   2.084   0.0411 *
## education    -0.5834     0.4725  -1.235   0.2214
## urban         0.6825     0.1232   5.539 6.11e-07 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.82 on 64 degrees of freedom
## Multiple R-squared:  0.4714, Adjusted R-squared:  0.4549
## F-statistic: 28.54 on 2 and 64 DF,  p-value: 1.379e-09
```

```
plot3d(ed.urb.lm.obj, col='red', size=8)
```

***See the crime~edu+urb_plot.png file for the plot of the fitted plane.**

c)

According to this model, for every percent increase of residents aged at least 25 in the county with at least a high school diploma, holding the percentage of residents who live in urban areas constant, the number of crimes in the past year per 1000 residents will **decrease** by 0.58 on average across all counties with the same percentage of residents who live in urban areas where the percentage of residents aged at least 25 in the county with at least a high school diploma differs by one. This is a change from the model that only considered education where a one unit increase in education caused an increase in crime. Additionally, the relationship between education and crime is no longer statistically significant. This is likely because urbanization was a confounding variable that is correlated with education.

d)

In counties where zero percent of residents aged at least 25 have at least a high school diploma and zero percent of residents live in urban areas, the number of crimes in the year per 1000 residents will be 59.1 on average across all such counties. Although I would consider this a very unlikely scenario it is not nonsensical. One could imagine a very rural county in which none of the residents that live there completed high school having a low rate of crime for the year. This intercept for problem 3 would be nonsensical since having a negative crime rate is nonsensical.