Homework 5

Please submit the solution in the form of R Markdown report, knitted into either of the available formats (HTML, pdf or Word). Provide all relevant code and output. Goal of this homework is to have you familiarized with 1) statistical inference for both simple and multiple linear regression (confidence intervals, t-test, F-test, prediction/confidence bands); 2) practice your R coding.

Problem #1

```
df <- readr::read_csv(sprintf("https://docs.google.com/uc?id=%s&export=download", "1UJIu7Ku3rRWTnFJpK4u

lm.obj <- lm(sales ~ radio, data = df)
```

1) Interpret the hypothesis test results. Is there a statistically significant relationship between predictor and response? Why?

```
summary( lm.obj )$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 9.3116381 0.56290050 16.542245 3.561071e-39
## radio       0.2024958 0.02041131  9.920765 4.354966e-19
```

There is statistically significant linear relationship between the predictor and the response value because the p-value ($4.354966 \times 10^{-19} \approx 0$) is less than most common significance levels ($\alpha = 0.001, 0.01, 0.05$).

2) Interpret the confidence intervals for both the intercept and the slope.

```
confint(lm.obj)
```

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

β_0 :We are 95% confident that for markets with \$0 spent on radio advertisements, the sales will be between 8201 and 10421 items, on average

β_1 :We are 95% confident that for \$1000 increase in radio advertisement, the sales will increase by between 162 and 242 items, on average

3) For a 20, 000\$ investment into radio ads, provide and interpret

a. A single prediction

```
pred <- predict( lm.obj, data.frame(radio=20) )
pred
```

```
##        1
## 13.36155
```

Markets with $20K spent on radio advertisement, the sales will be 13361 items, on average

b. 95% confidence bands

```
pred <- predict( lm.obj, data.frame(radio=20), interval = 'c' )
pred
```

```
##        fit      lwr      upr
## 1 13.36155 12.75114 13.97197
```

We are 95% confident that for markets with $20K spent on radio advertisement, sales will be between 12751 and 13971 items, on average.

c. 95% predictions bands.

```
pred <- predict( lm.obj, data.frame(radio=20), interval = 'p' )
pred
```

```
##        fit      lwr      upr
## 1 13.36155 4.909218 21.81389
```

We are 95% confident that for any single market with $20K spent on radio advertisement, sales will be between 4909 and 21813 items.

Which ones are wider - confidence or prediction? Why do you think that is?
Prediction bands are wider because it captures where the individual response value is likely to land whereas the confidence band captures the *average* response value.

Problem #2

```
df <- readr::read_csv("https://img1.wsimg.com/blobby/go/bbca5dba-4947-4587-b40a-db346c01b1b3/downloads/
```

1) For simple linear regression $crime \sim education$,

```
lm.obj <- lm(`crime rate (per 1000)` ~ `education (%)`, data = df)
```

a. Write down the **full modeling equation**, with all **error assumptions**
   $crime = \beta_0 + \beta_1 education + \epsilon, \;\; \epsilon \sim N(0, \sigma^2)$

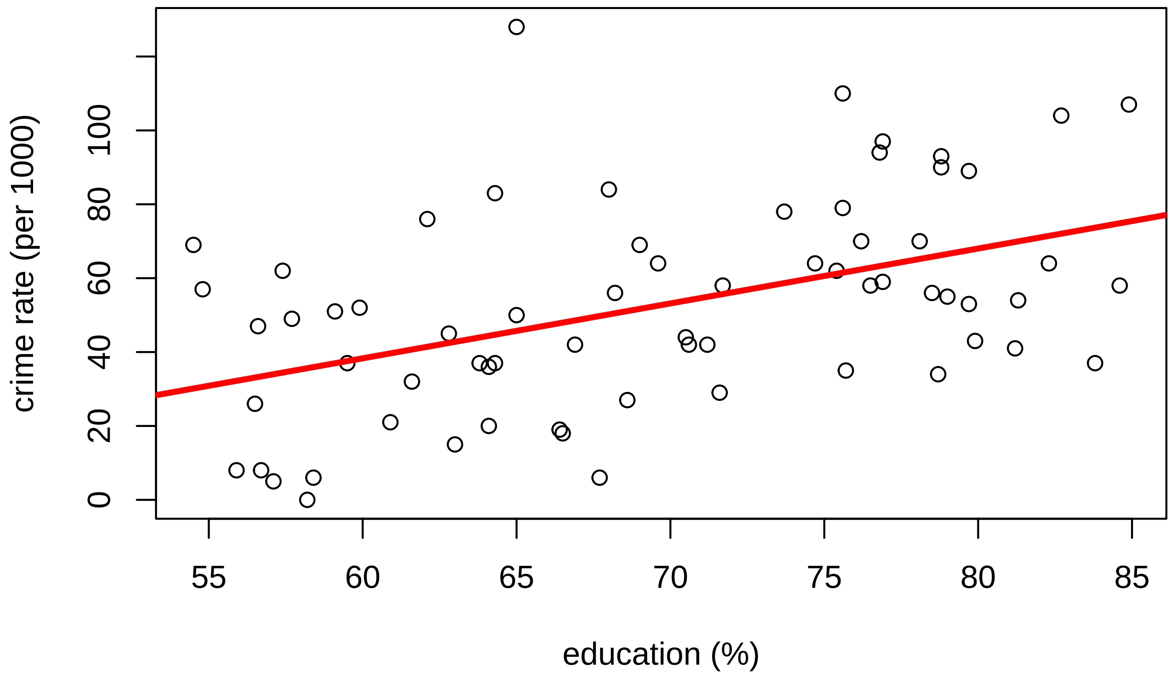
b. Fit the model, provide the **fitted equation**. Provide a plot of the fitted line.

```
lm.obj$coefficients
```

```
##     (Intercept) ‘education (%)‘
##      -50.856902        1.485977
```

$\hat{crime}$ = -50.8569018+ 1.4859772$education$

```
plot(`crime rate (per 1000)` ~ `education (%)`, data = df)
abline(lm.obj, col = 'red', lwd = 3)
```



c. Is there a statistically significant relationship? If yes, interpret the effect of education on crime.
$H_0 : \beta_1 = 0$
$H_a : \beta_1 \neq 0$

```
summary( lm.obj )$coefficients
```

```
##                   Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)     -50.856902 24.4506518 -2.079981 4.147477e-02
## ‘education (%)‘   1.485977  0.3490777  4.256867 6.805641e-05
```

There is statistically significant linear relationship between education (the predictor) and crime (the response value) because the p-value (6.8056411×10^{-5}≈ 0) is less than all common significance levels ($\alpha = 0.001, 0.01, 0.05$).

2) For multiple linear regression $crime \sim education + urbanization$,

```
lm.obj <- lm(`crime rate (per 1000)` ~ `education (%)` + `urbanization (%)`, data = df)
```

a. Write down the **full modeling equation**, with all **error assumptions**.
   $crime = \beta_0 + \beta_1 education + \beta_2 urbanization + \epsilon,\ \epsilon \sim N(0, \sigma^2)$

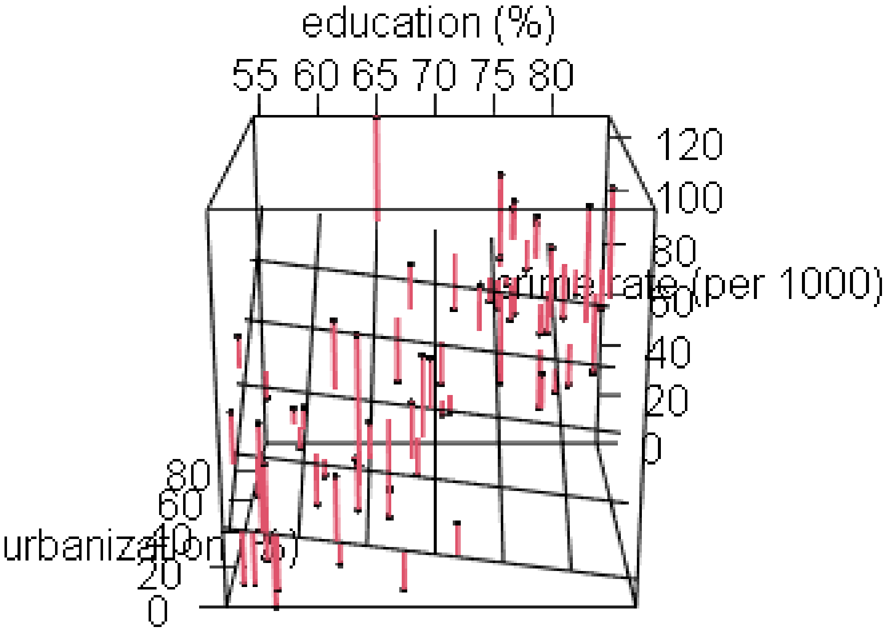
b. Fit the model, provide the **fitted equation**. Provide a plot of the **fitted plane**

```
lm.obj$coefficients
```

```
##         (Intercept)     ‘education (%)‘ ‘urbanization (%)‘
##          59.1180677          -0.5833773           0.6825014
```

$\hat{crime}$ = 59.1180677+ -0.5833773$education$+ 0.6825014$urbanization$

```
rgl::plot3d(lm.obj)
rgl::segments3d(rep(df$`education (%)`, each=2),
           rep(df$`urbanization (%)`, each=2),
           z=matrix(t(cbind(df$`crime rate (per 1000)`,predict(lm.obj))), nc=1),
           add=T,
           lwd=2,
           col=2)
```



c. Provide interpretation of education effect on crime. Did it change compared to part 1? Why?

```
summary( lm.obj )$coefficients
```

```
##                       Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)         59.1180677 28.3653106  2.084168 4.114165e-02
## ‘education (%)‘     -0.5833773  0.4724591 -1.234768 2.214311e-01
## ‘urbanization (%)‘   0.6825014  0.1232126  5.539218 6.110801e-07
```

It is plausible that there is no linear relationship between education (a predictor) and crime (the response value) because the p-value (0.2214311) is higher than most common significance levels ($\alpha = 0.001, 0.01, 0.05$).

```
cor(df[,2:4])
```

```
##                       crime rate (per 1000) education (%) urbanization (%)
## crime rate (per 1000)             1.0000000     0.4669119        0.6773678
## education (%)                     0.4669119     1.0000000        0.7907190
## urbanization (%)                  0.6773678     0.7907190        1.0000000
```

Notice how the education effect on crime changed after including urbanization as one of our predictors. This is due to an issue of colliniarity - two (or more) predictors are strongly correlated and thus one acts as a proxy for another when influencing the response value. This effect is not captured in simple linear regression because it assumes all effect on the response value (crime) is due to a single predictor (education). In this case, education and urbanization are highly correlated with each other, with urbanization being more strongly correlated to crime than education.

d. Provide interpretation of the intercept. By the sound of it, does it make sense to interpret it here? Counties with 0% education and 0% urbanization, the crime rate will be 59.1180677 per 1000 people, on average. Depending on the definition of education, a 0% value could be impossible or highly improbable, as basic education (e.g. counting, language learning, etc.) is reasonably still taught in the smallest/unstructured social communities (e.g. hunter-gatherer). In effect, it could be extrapolation.

3) For multiple linear regression $crime \sim education + urbanization + income$

```
lm.obj <- lm(`crime rate (per 1000)` ~ `education (%)` + `urbanization (%)` + `income (median, in 1000)
```

a. Write down the **full modeling equation**, with all **error assumptions**.
   $crime = \beta_0 + \beta_1 education + \beta_2 urbanization + \beta_3 income + \epsilon,\ \epsilon \sim N(0, \sigma^2)$

b. Having fitted the model from part 3(a), provide the **fitted equation**.

```
lm.obj$coefficients
```

```
##                (Intercept)            ‘education (%)‘
##                 59.7147293                 -0.4672860
##         ‘urbanization (%)‘ ‘income (median, in 1000)‘
##                  0.6971509                 -0.3830938
```

$\widehat{crime}$ = 59.7147293+ -0.467286$education$+ 0.6971509$urbanization$+ -0.3830938$income$

c. Write down the hypotheses (in terms of parameters of the model in part 3(a)) and make conclusions for t-tests on significance of each individual predictor.

```
summary(lm.obj)$coefficients
```

```
##                                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)                  59.7147293 28.5895274  2.0886924 4.078185e-02
## ‘education (%)‘              -0.4672860  0.5544348 -0.8428150 4.025204e-01
## ‘urbanization (%)‘            0.6971509  0.1291330  5.3987034 1.084630e-06
## ‘income (median, in 1000)‘ -0.3830938  0.9405262 -0.4073186 6.851547e-01
```

education: $H_0 : \beta_1 = 0$ $H_a : \beta_1 \neq 0$
it is plausible that education has no linear relationship with crime, holding all other predictors constant
urbanization: $H_0 : \beta_2 = 0$ $H_a : \beta_2 \neq 0$
urbanization has a linear relationship with crime
income: $H_0 : \beta_3 = 0$ $H_a : \beta_3 \neq 0$
it is plausible that income has no linear relationship with crime, holding all other predictors constant

d. Interpret the effect of the only statistically significant predictor from part 3(c).
   Per 1% increase in urbanization, holding education and income constant, crime rate will increase by 0.6971509 per 1000 people, on average

e. Formulate the hypotheses (in terms of parameters of the model in part 3(a) for testing the overall model significance. Provide the conclusion of the respective test
   $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
   $H_a : \{at\ least\ one\ \beta_j \neq 0\, j = 1, 2, 3\}$

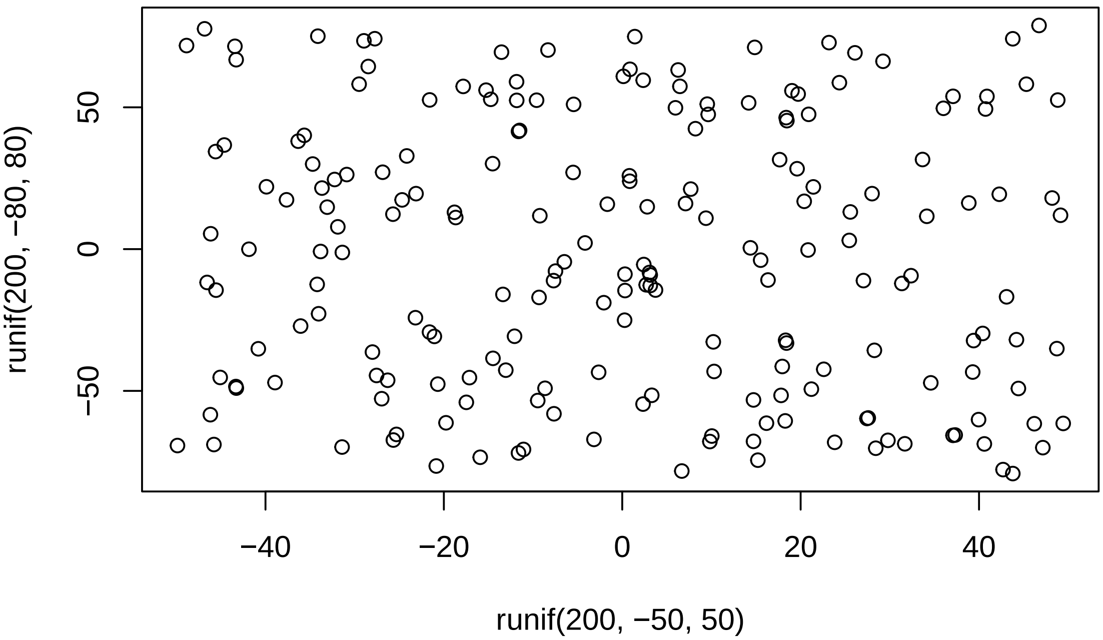
```
fstat <- c(as.list(summary(lm.obj)$fstatistic), FALSE)
names(fstat) <- c("q", "df1", "df2", "lower.tail")
do.call(pf, fstat)
```

```
## [1] 7.822851e-09
```

The linear model is statistically significant because the p-value ($7.8228513 \times 10^{-9} \approx 0$) is less than most significance levels ($\alpha = 0.001, 0.01, 0.05$). Therefore, at least one predictor has a linear relationship with crime (the response variable).

Problem #3 (Why need F-statistic?)

1. Generate a data example where you have a response variable Y and a predictor variable X that are unrelated to each other (make sure to use a random generation mechanism). How would you do that? How would you demonstrate that they're unrelated (think of basic visualizations)?
   To randomly generate unrelated variables, I'd randomly sample values from a range where each value has an equal probability of being selected (uniformly distributed). The scatterplot of the y X would have data points spread across the whole plot, more or less equally distributed across the X and Y value ranges. In other words, a plot with data points everywhere with no particular pattern or clumping to the location of the data points. See below for an example.



2. Having settled on a method of generating such unrelated variables in part 1, proceed to:

a. Generate response variable vector Y (e.g. of length 200)
b. Generate response variable vector Y (e.g. of length 200); **Record them**.

```
set.seed(2)
y <- runif(200, -50,50)
X <- matrix(runif(200*50, -50,50), ncol=50)
lm.obj <- lm(y ~ X)
```

3. Fit a multiple linear regression model, regressing response Y from part 2(a) on all 50 X's you've generated in part 2(b).

```
summary(lm.obj)
```

```
##
## Call:
## lm(formula = y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.336 -20.700  -1.316  23.544  56.266
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.786924   2.518937  -0.312   0.7552
## X1          -0.109339   0.084861  -1.288   0.1996
## X2          -0.016085   0.082447  -0.195   0.8456
## X3          -0.046293   0.084178  -0.550   0.5832
## X4           0.136741   0.085582   1.598   0.1122
## X5           0.124773   0.082694   1.509   0.1335
## X6           0.016506   0.078698   0.210   0.8342
## X7           0.089782   0.094223   0.953   0.3422
## X8           0.110020   0.084544   1.301   0.1952
## X9          -0.003394   0.086654  -0.039   0.9688
## X10         -0.199494   0.086328  -2.311   0.0222 *
## X11          0.018885   0.084166   0.224   0.8228
## X12          0.064814   0.087697   0.739   0.4610
## X13          0.125067   0.080839   1.547   0.1240
## X14         -0.098039   0.095277  -1.029   0.3052
## X15          0.007750   0.084309   0.092   0.9269
## X16         -0.090731   0.092159  -0.984   0.3265
## X17          0.183470   0.088601   2.071   0.0401 *
## X18          0.048759   0.082108   0.594   0.5535
## X19          0.063232   0.088266   0.716   0.4749
## X20         -0.095693   0.088116  -1.086   0.2792
## X21         -0.162614   0.086842  -1.873   0.0631 .
## X22         -0.001544   0.089307  -0.017   0.9862
## X23          0.046503   0.084245   0.552   0.5818
## X24          0.041564   0.086361   0.481   0.6310
## X25         -0.052631   0.087482  -0.602   0.5483
## X26         -0.054702   0.085638  -0.639   0.5240
## X27          0.105381   0.094161   1.119   0.2649
## X28         -0.132203   0.087801  -1.506   0.1343
## X29         -0.078481   0.085541  -0.917   0.3604
## X30          0.090634   0.085008   1.066   0.2881
## X31         -0.074172   0.089461  -0.829   0.4084
## X32         -0.023952   0.081883  -0.293   0.7703
## X33         -0.105899   0.083963  -1.261   0.2092
## X34          0.141428   0.090432   1.564   0.1200
## X35          0.125056   0.088249   1.417   0.1585
## X36          0.029250   0.084344   0.347   0.7292
## X37         -0.006711   0.084209  -0.080   0.9366
## X38          0.069250   0.084860   0.816   0.4158
## X39          0.025403   0.088959   0.286   0.7756
## X40          0.029945   0.087006   0.344   0.7312
## X41          0.024307   0.085794   0.283   0.7773
## X42         -0.061781   0.083850  -0.737   0.4624
## X43         -0.028337   0.091288  -0.310   0.7567
## X44          0.042250   0.098302   0.430   0.6680
## X45         -0.012533   0.085763  -0.146   0.8840
## X46          0.017430   0.089331   0.195   0.8456
## X47         -0.040076   0.086734  -0.462   0.6447
## X48          0.015451   0.088730   0.174   0.8620
## X49         -0.061313   0.084422  -0.726   0.4688
## X50         -0.137590   0.086910  -1.583   0.1155
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 30.79 on 149 degrees of freedom
## Multiple R-squared:  0.2285, Adjusted R-squared:  -0.03042
## F-statistic: 0.8825 on 50 and 149 DF,  p-value: 0.6898
```

a. Report the # of individual t-tests that resulted into a significant p-value, hence rejecting H0 at $\alpha = 0.05$ level. Were those rejections correct decisions or not? Why? (Hint: Remember what's the true relationship between X and Y , given how you generated the data.) If not, what type of error do they correspond to? Why?
   There were 2 significant p-values at the $\alpha = 0.05$ level. These rejections were the incorrect decision because we DESIGNED X to be unrealated to Y. Therefore, they represent Type I Error because we reject the null hypothesis (no linear relationship) when the null hypothesis is true (variables are unrelated).

b. Given that the individual significant t-test aren't necessarily indicative of at least one predictor having a true relationship with the response, what would be the appropriate testing procedure to address that question? Conduct that test, report its p-value, interpret its result.
   F-test would be an appropriate testing procedure to address whether at least one predictor has a true linear relationship with the response value.
   $H_0 : \beta_1 = ... = \beta_{50} = 0$
   $H_a : \{at\ least\ one\ \beta_j \neq 0,\ j = 1, ..., 50\}$

```r
summary(lm.obj)$fstatistic
```

```
##      value      numdf      dendf
##   0.882507  50.000000 149.000000
```

```r
fstat <- c(as.list(summary(lm.obj)$fstatistic), FALSE)
names(fstat) <- c("q", "df1", "df2", "lower.tail")
do.call(pf, fstat)
```

```
## [1] 0.6897567
```

The p-value (0.6897567) is greater than the significance level ($\alpha = 0.05$), therefore we fail to reject the null hypothesis. It is plausible that none of the predictors have a linear relationship with the response; plausible that the linear model is useless.