qbs121_hw2_gibran

Gibran Erlangga

1/16/2022

Problems (Bonus)

2. Show that minimizing mean square error is the same as maximzing R^2 is the same as minimizing the sum of squares.

4. Suppose you add to your model the interactions of two categorical variables, and that the number of categories of these two categorical variables are r and s respectively. How many degrees of freedom are used by the interaction?

6. a. Suppose E[log(Y)|X1, X2] = b0 + b1 log(X1) + b2X2. How does a k fold increase in X1

affect the expected value of Y holding X2 constant?

Data Analyses

2.1 Analysis of the FEV Data

Load the data.

```
FEV.Data <- read.delim("http://jse.amstat.org/datasets/fev.dat.txt", sep="", header=FALSE)
names(FEV.Data) <- c("Age","FEV","Height","Male","Smoker")
attach(FEV.Data)
```

1. Effect of Smoking: Report the effect of smoking on FEV, using a univariable model (unadjusted) and multivariable model adjusting for age, height and gender.

```
summary(lm(FEV ~ Smoker))
```

```
##
## Call:
## lm(formula = FEV ~ Smoker)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7751 -0.6339 -0.1021  0.4804  3.2269
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.56614    0.03466  74.037  < 2e-16 ***
## Smoker       0.71072    0.10994   6.464 1.99e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8412 on 652 degrees of freedom
## Multiple R-squared:  0.06023,    Adjusted R-squared:  0.05879
## F-statistic: 41.79 on 1 and 652 DF,  p-value: 1.993e-10
```

```
summary(lm(FEV ~ Smoker + Age + Height + Male))
```

```
##
## Call:
## lm(formula = FEV ~ Smoker + Age + Height + Male)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.37656 -0.25033  0.00894  0.25588  1.92047
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.456974   0.222839 -20.001  < 2e-16 ***
## Smoker      -0.087246   0.059254  -1.472    0.141
## Age          0.065509   0.009489   6.904 1.21e-11 ***
## Height       0.104199   0.004758  21.901  < 2e-16 ***
## Male         0.157103   0.033207   4.731 2.74e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4122 on 649 degrees of freedom
## Multiple R-squared:  0.7754, Adjusted R-squared:  0.774
## F-statistic:   560 on 4 and 649 DF,  p-value: < 2.2e-16
```

2. Effect of Age and Gender: Test if the effect of age on FEV is different in males and females. If so, do subgroup analyses reporting the effect of age in males and females separately.

```
# overall
summary(lm(FEV ~ Age + Male))
```

```
##
## Call:
## lm(formula = FEV ~ Age + Male)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.41495 -0.35175 -0.03717  0.31756  1.97394
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 0.281378   0.077300   3.640 0.000294 ***
## Age         0.220445   0.007215  30.553  < 2e-16 ***
## Male        0.323335   0.042609   7.588 1.13e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5444 on 651 degrees of freedom
## Multiple R-squared:  0.607,  Adjusted R-squared:  0.6058
## F-statistic: 502.7 on 2 and 651 DF,  p-value: < 2.2e-16
```

```
# female
summary(lm(FEV ~ Age, subset=Male==0))
```

```
##
## Call:
## lm(formula = FEV ~ Age, subset = Male == 0)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.09240 -0.28991 -0.03762  0.28749  1.13451
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.849467   0.085695   9.913   <2e-16 ***
## Age         0.162729   0.008345  19.500   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4357 on 316 degrees of freedom
## Multiple R-squared:  0.5461, Adjusted R-squared:  0.5447
## F-statistic: 380.3 on 1 and 316 DF,  p-value: < 2.2e-16
```

```
# male
summary(lm(FEV ~ Age, subset=Male==1))
```

```
##
## Call:
## lm(formula = FEV ~ Age, subset = Male == 1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.64072 -0.37752 -0.05318  0.36893  1.86867
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.0736     0.1128   0.653    0.514
## Age           0.2735     0.0108  25.329   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5881 on 334 degrees of freedom
## Multiple R-squared:  0.6576, Adjusted R-squared:  0.6566
## F-statistic: 641.6 on 1 and 334 DF,  p-value: < 2.2e-16
```

3. Effect of Height and Gender: Test if the effect of height on FEV is different in males and females. If so, do subgroup analyses reporting the effect of height in males and females separately.

```r
# overall
summary(lm(FEV ~ Height + Male))
```

```
##
## Call:
## lm(formula = FEV ~ Height + Male)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6763 -0.2505  0.0001  0.2347  2.0722
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.390263   0.180082 -29.932  < 2e-16 ***
## Height       0.130231   0.002964  43.933  < 2e-16 ***
## Male         0.125123   0.033801   3.702 0.000232 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4265 on 651 degrees of freedom
## Multiple R-squared:  0.7587, Adjusted R-squared:  0.758
## F-statistic:  1024 on 2 and 651 DF,  p-value: < 2.2e-16
```

```r
# female
summary(lm(FEV ~ Height, subset=Male==0))
```

```
##
## Call:
## lm(formula = FEV ~ Height, subset = Male == 0)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.54654 -0.20323  0.01498  0.22968  1.02038
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.318219   0.252449   -17.1   <2e-16 ***
## Height       0.112426   0.004179    26.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3566 on 316 degrees of freedom
## Multiple R-squared:  0.696,  Adjusted R-squared:  0.6951
## F-statistic: 723.6 on 1 and 316 DF,  p-value: < 2.2e-16
```

```r
# male
summary(lm(FEV ~ Height, subset=Male==1))
```

```
##
```

```
## Call:
## lm(formula = FEV ~ Height, subset = Male == 1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.13438 -0.30820 -0.00568  0.30821  2.00491
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.863848   0.254470  -23.04   <2e-16 ***
## Height       0.139883   0.004082   34.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4729 on 334 degrees of freedom
## Multiple R-squared:  0.7786, Adjusted R-squared:  0.7779
## F-statistic:  1175 on 1 and 334 DF,  p-value: < 2.2e-16
```

2.1 Analysis of HSB Data

Download the following dataset and install the library multcomp.

```r
hsb2 <- read.csv("https://stats.idre.ucla.edu/stat/data/hsb2.csv")
library(multcomp)
```

```
## Warning: package 'multcomp' was built under R version 4.1.1

## Loading required package: mvtnorm

## Warning: package 'mvtnorm' was built under R version 4.1.1

## Loading required package: survival

## Loading required package: TH.data

## Warning: package 'TH.data' was built under R version 4.1.1

## Loading required package: MASS

##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##     geyser
```

1. Model "read" in terms of female, schtyp and ses (as a factor);

```r
model <- lm(read ~ female + schtyp + factor(ses), hsb2)
summary(model)
```

```
##
## Call:
## lm(formula = read ~ female + schtyp + factor(ses), data = hsb2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.450  -6.663  -1.066   7.013  21.484
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   47.0432     2.6310  17.880  < 2e-16 ***
## female        -0.4628     1.4201  -0.326   0.7449
## schtyp         1.4852     1.9377   0.766   0.4443
## factor(ses)2   2.9873     1.8067   1.653   0.0998 .
## factor(ses)3   7.9212     1.9776   4.006  8.8e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.877 on 195 degrees of freedom
## Multiple R-squared:  0.09073,    Adjusted R-squared:  0.07208
## F-statistic: 4.864 on 4 and 195 DF,  p-value: 0.0009242
```

2. Show the first few rows of the design matrix.

```r
head(X <- model.matrix(~female + schtyp + factor(ses), hsb2), 10)
```

```
##    (Intercept) female schtyp factor(ses)2 factor(ses)3
## 1            1      0      1            0            0
## 2            1      1      1            1            0
## 3            1      0      1            0            1
## 4            1      0      1            0            1
## 5            1      0      1            1            0
## 6            1      0      1            1            0
## 7            1      0      1            1            0
## 8            1      0      1            1            0
## 9            1      0      1            1            0
## 10           1      0      1            1            0
```

3. Calculate formula where Y is the reading score and X is the design matrix.

```r
Y <- hsb2$read

solve(t(X) %*% X) %*% t(X) %*% Y
```

```
##                   [,1]
## (Intercept)  47.043230
## female       -0.462757
## schtyp        1.485233
## factor(ses)2  2.987252
## factor(ses)3  7.921233
```

4. Compare the value computed in the previous step to the coefficients from the lm. Are they the same?

Yes, they are close enough.

```
summary(model$coefficients)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.4628  1.4852  2.9872 11.7948  7.9212 47.0432
```

5. Use the "waldtest" function of the library "lmtest" to test the null hypothesis that the factor ses explains no variation in reading scores.

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
model_null <- lm(read ~ female + schtyp, hsb2)
waldtest(model_null, model)
```

```
## Wald test
##
## Model 1: read ~ female + schtyp
## Model 2: read ~ female + schtyp + factor(ses)
##   Res.Df Df      F    Pr(>F)
## 1    197
## 2    195  2 8.6147 0.0002599 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6. Repeat the last step manually using syntax like t(coef(o)[4:5]) %% *solve(vcov(o))[4:5] %%* coef(o)[4:5] to create the test statistic and using an F-test.

```
t_statistic <- t(coef(model)[4:5]) %*% solve(vcov(model))[4:5] %*% coef(model)[4:5]

p_val <- 1 - pf(t_statistic, df1=2, df2=model$df.residual)
p_val
```

```
##              [,1] [,2]
## [1,] 1.306028e-09    0
```

2.3 Smoothing

1. Locate the dataset "ryegrass" in the CRAN library "drc".

```r
library(drc)
```

```
##
## 'drc' has been loaded.

## Please cite R and 'drc' if used for a publication,

## for references type 'citation()' and 'citation('drc')'.

##
## Attaching package: 'drc'

## The following objects are masked from 'package:stats':
##
##     gaussian, getInitial
```

```r
data <- ryegrass
attach(data)
```

2. Fit a straightline to the data and superimpose it on the scatterplot of rootl versus conc.

```r
plot(rootl, conc, data=data)
```

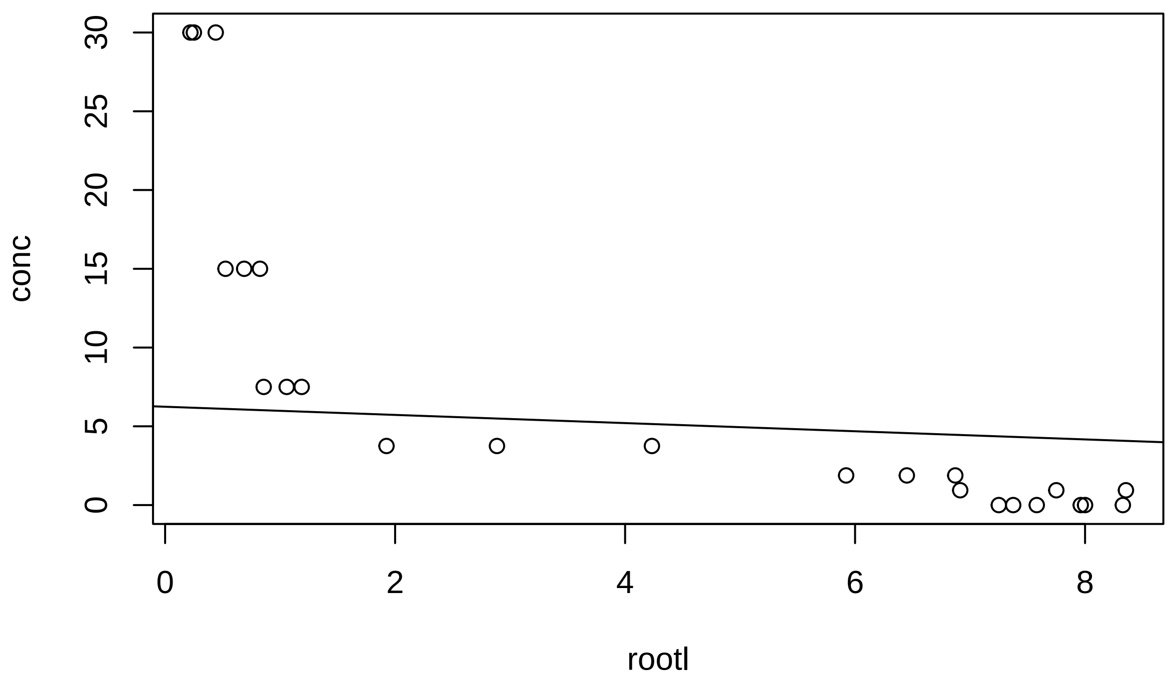
```
## Warning in plot.window(...): "data" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "data" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not a
## graphical parameter

## Warning in box(...): "data" is not a graphical parameter

## Warning in title(...): "data" is not a graphical parameter
```

```r
summary(o <- lm(rootl ~ conc, data=data))
```

```
##
## Call:
## lm(formula = rootl ~ conc, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4399 -1.7055  0.9623  1.7532  2.3575
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  6.24176    0.52973  11.783 5.64e-11 ***
## conc        -0.25929    0.04326  -5.994 4.94e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.07 on 22 degrees of freedom
## Multiple R-squared:  0.6202, Adjusted R-squared:  0.603
## F-statistic: 35.93 on 1 and 22 DF,  p-value: 4.939e-06
```

```
abline(o)
```



3. Using different colors add a quadratic fit of rootl versus conc.

```
#create a new variable for conc2
data$conc2 <- data$conc^2

#fit quadratic regression model
quadraticModel <- lm(rootl ~ conc + conc2, data=data)

#view model summary
summary(quadraticModel)
```

```
##
## Call:
```

```
## lm(formula = rootl ~ conc + conc2, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68138 -0.34397 -0.04228  0.78019  1.58094
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.574810   0.338120  22.403 3.84e-16 ***
## conc        -0.871233   0.086095 -10.119 1.57e-09 ***
## conc2        0.021240   0.002876   7.384 2.90e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.117 on 21 degrees of freedom
## Multiple R-squared:  0.8944, Adjusted R-squared:  0.8844
## F-statistic: 88.94 on 2 and 21 DF,  p-value: 5.597e-11
```

```r
plot(rootl, conc, data=data)
```

```
## Warning in plot.window(...): "data" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "data" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not a
## graphical parameter

## Warning in box(...): "data" is not a graphical parameter

## Warning in title(...): "data" is not a graphical parameter
```

```r
summary(o <- lm(rootl ~ conc, data=data))
```

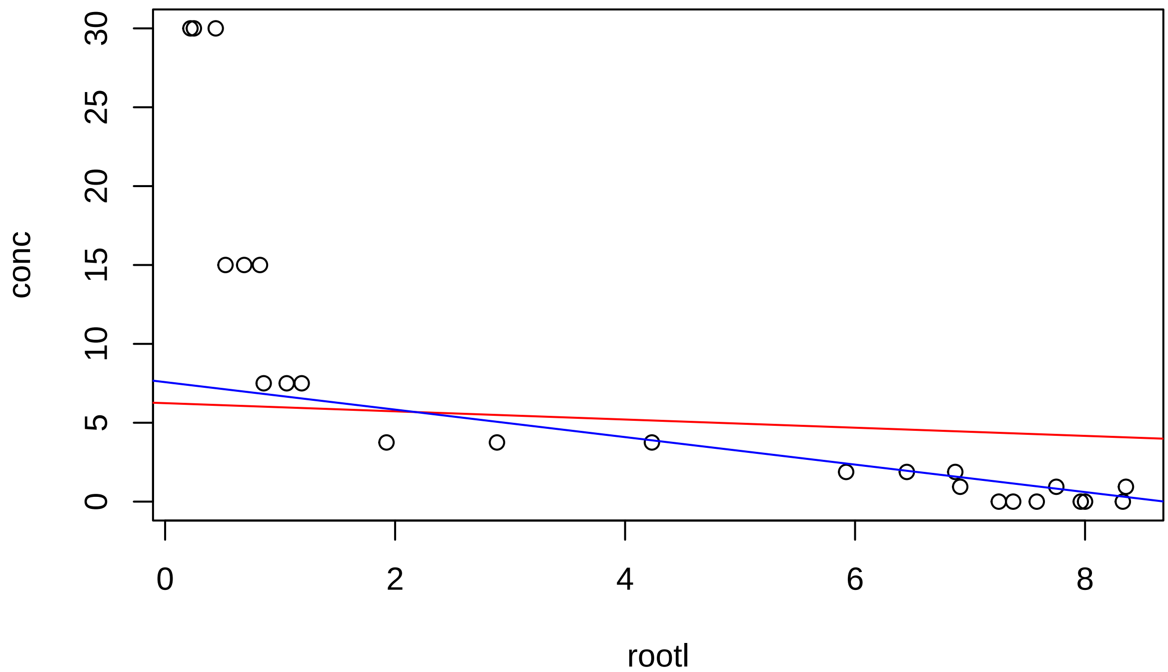
```
##
## Call:
## lm(formula = rootl ~ conc, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4399 -1.7055  0.9623  1.7532  2.3575
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.24176    0.52973  11.783 5.64e-11 ***
## conc        -0.25929    0.04326  -5.994 4.94e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.07 on 22 degrees of freedom
## Multiple R-squared:  0.6202, Adjusted R-squared:  0.603
## F-statistic: 35.93 on 1 and 22 DF,  p-value: 4.939e-06
```

```r
summary(q <- quadraticModel)
```

```
##
## Call:
## lm(formula = rootl ~ conc + conc2, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68138 -0.34397 -0.04228  0.78019  1.58094
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.574810   0.338120  22.403 3.84e-16 ***
## conc        -0.871233   0.086095 -10.119 1.57e-09 ***
## conc2        0.021240   0.002876   7.384 2.90e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.117 on 21 degrees of freedom
## Multiple R-squared:  0.8944, Adjusted R-squared:  0.8844
## F-statistic: 88.94 on 2 and 21 DF,  p-value: 5.597e-11
```

```r
abline(o, col=c("red"))
abline(q, col=c("blue"))
```

```
## Warning in abline(q, col = c("blue")): only using the first two of 3 regression
## coefficients
```



4. Use the gam function from the gam library to fit a smooth curve.

```
library(gam)
```

```
## Loading required package: splines
```

```
## Loading required package: foreach
```

```
## Loaded gam 1.20
```

```
library(mgcv)
```

```
## Warning: package 'mgcv' was built under R version 4.1.1
```

```
## Loading required package: nlme
```

```
## Warning: package 'nlme' was built under R version 4.1.1
```

```
## This is mgcv 1.8-38. For overview type 'help("mgcv-package")'.
```

```
##
## Attaching package: 'mgcv'
```

```
## The following objects are masked from 'package:gam':
##
##     gam, gam.control, gam.fit, s
```

```
gam_model <- gam(rootl ~ s(conc, k=7), data=data)

xvals <- data.frame(seq(0, 30, 0.1))
colnames(xvals) <- "conc"

gam_pred <- predict.gam(gam_model, xvals)

plot(rootl, conc, data=data)
```

```
## Warning in plot.window(...): "data" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "data" is not a graphical parameter
```

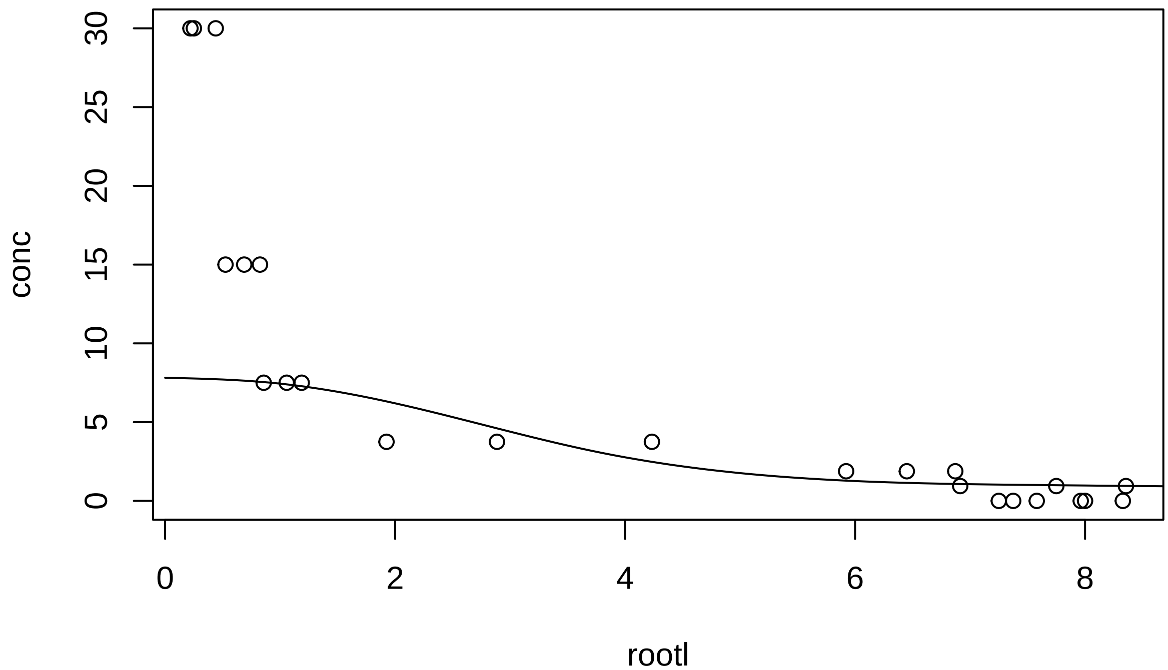
```
## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not a
## graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not a
## graphical parameter
```

```
## Warning in box(...): "data" is not a graphical parameter
```

```
## Warning in title(...): "data" is not a graphical parameter
```

```
lines(xvals$conc, gam_pred)
```



2.5 Simulate and Analyze 2

1. Generate the following data consisting of a dependent variable Y an exposure of interest X and a covariate Z.

```
set.seed(121)

n <- 300
Z <- runif(n) < 0.5
X <- rnorm(n) + ifelse(Z, 1.5, -1.5)
Y<- ifelse(Z, X-2.5, X+2.5) + rnorm(n)
summary(lm(Y~X)) $coef
```

```
##                 Estimate Std. Error    t value    Pr(>|t|)
## (Intercept)  0.01187338 0.09961291  0.1191952 0.905201028
## X           -0.14117129 0.05417648 -2.6057672 0.009627311
```

2. Interpret the results of the linear regression, and conclude if Y increases, decreases or has no association with X.

Based on the linear regression result above, for every one unit increase of X, the value of Y decreases by 0.14 point.

3. Now consider the covariate Z. Is it associated with Y?

```r
summary(lm(Y ~ Z))$coef
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  0.9208342  0.1143456   8.053083 1.951882e-14
## ZTRUE       -1.9189495  0.1679858 -11.423286 2.619649e-25
```

4. Run and interpret the following analyses.

```r
summary(lm(Y~X, subset=Z))$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -2.646120 0.15648675 -16.90954 2.410378e-35
## X            1.077654 0.08525334  12.64061 9.282773e-25
```

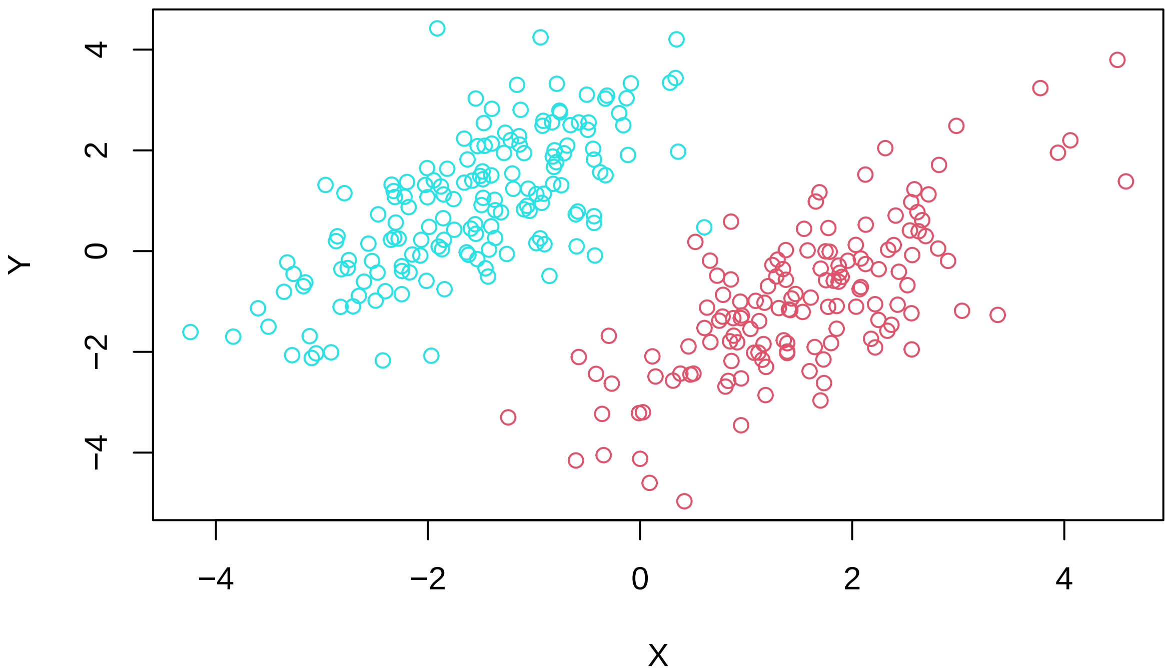
```r
summary(lm(Y~X, subset=!Z))$coef
```

```
##             Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 2.567983 0.15624946 16.43515 4.178613e-36
## X           1.041012 0.08485509 12.26811 9.060048e-25
```

```r
summary(lm(Y~X + Z))$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  2.597016 0.12421192  20.90795 2.586792e-60
## X            1.059361 0.06003702  17.64513 3.687683e-48
## ZTRUE       -5.215161 0.22072505 -23.62741 3.509442e-70
```

```r
plot(X, Y, col=ifelse (Z,2 ,5))
```



5. Comment on the disparity in results for the association of Y and X. Based on the regression results above, Z has a significant effect on Y given the p-value score shown above.

6. Test if there is an interaction of X and Z.

```
summary(lm(Y ~ X*Z))$coef
```

```
##                Estimate Std. Error     t value     Pr(>|t|)
## (Intercept)  2.56798304 0.15670032  16.3878607 2.106307e-43
## X            1.04101165 0.08509994  12.2328125 3.993268e-28
## ZTRUE       -5.21410268 0.22109018 -23.5836016 6.164178e-70
## X:ZTRUE      0.03664284 0.12025799   0.3047019 7.608073e-01
```