HW02p

[Angel Montero]

March 6, 2018

```
knitr::opts_chunk$set(error = TRUE) #this allows errors to be printed into the PDF
```

Welcome to HW02p where the "p" stands for "practice" meaning you will use R to solve practical problems. This homework is due 11:59 PM Tuesday 3/6/18.

You should have RStudio installed to edit this file. You will write code in places marked "TO-DO" to complete the problems. Some of this will be a pure programming assignment. Sometimes you will have to also write English.

The tools for the solutions to these problems can be found in the class practice lectures. I want you to use the methods I taught you, not for you to google and come up with whatever works. You won't learn that way.

To "hand in" the homework, you should compile or publish this file into a PDF that includes output of your code. To do so, use the knit menu in RStudio. You will need LaTeX installed on your computer. See the email announcement I sent out about this. Once it's done, push the PDF file to your github class repository by the deadline. You can choose to make this respository private.

For this homework, you will need the `testthat` libray.

```
pacman::p_load(testthat)
```

1. Source the simple dataset from lecture 6p:

```
Xy_simple = data.frame(
 response = factor(c(0, 0, 0, 1, 1, 1)), #nominal
 first_feature = c(1, 1, 2, 3, 3, 4),    #continuous
 second_feature = c(1, 2, 1, 3, 4, 3)    #continuous
)
X_simple_feature_matrix = as.matrix(Xy_simple[, 2 : 3])
y_binary = as.numeric(Xy_simple$response == 1)
```

Try your best to write a general perceptron learning algorithm to the following `Roxygen` spec. For inspiration, see the one I wrote in lecture 6.

```
#' This function implements the "perceptron learning algorithm" of Frank Rosenblatt (1957).
#'
#' @param Xinput      The training data features as an n x (p + 1) matrix where the first column is all
#' @param y_binary    The training data responses as a vector of length n consisting of only 0's and 1'
#' @param MAX_ITER    The maximum number of iterations the perceptron algorithm performs. Defaults to 1
#' @param w           A vector of length p + 1 specifying the parameter (weight) starting point. Defaul
#'                    NULL which means the function employs random standard uniform values.
#' @return            The computed final parameter (weight) as a vector of length p + 1
perceptron_learning_algorithm = function(Xinput, y_binary, MAX_ITER = 10, w = NULL){
  #TO-DO

  if (is.null(w)){
    w = runif(ncol(Xinput)) #intialize a p+1-dim vector with random values
  }
  for (i in 1:MAX_ITER) {
    for (d in 1:nrow(Xinput)) {
      x_hat = Xinput[d,]
      y_hat = ifelse(x_hat %*% w > 0, 1, 0)

      diff = as.numeric(y_binary[d] - y_hat)

      delta_w = diff * x_hat

      w = w + delta_w
    }
  }
  w
}
```

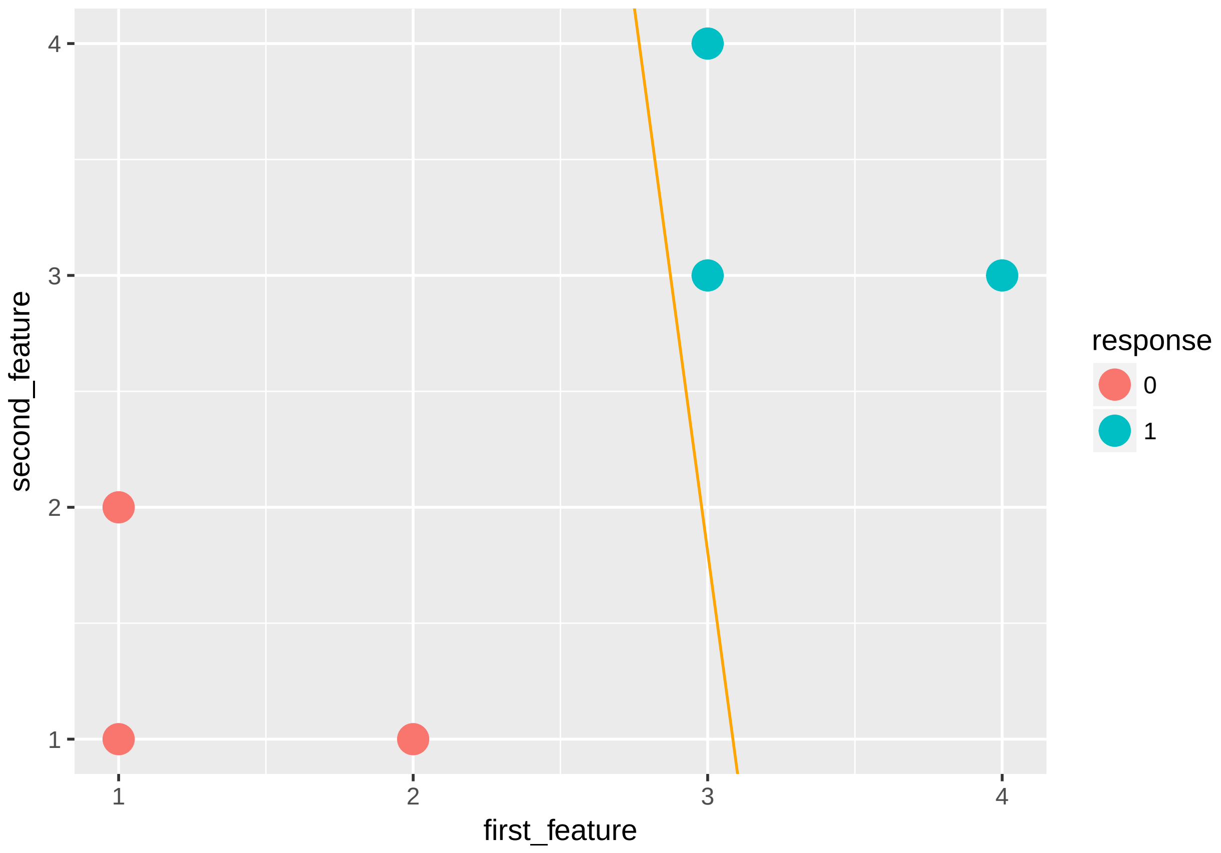
Run the code on the simple dataset above via:

```
w_vec_simple_per = perceptron_learning_algorithm(
  cbind(1, Xy_simple$first_feature, Xy_simple$second_feature),
  as.numeric(Xy_simple$response == 1))
w_vec_simple_per
```

```
## [1] -8.5055865  2.6648289  0.2826384
```

Use the ggplot code to plot the data and the perceptron's g function.

```
pacman::p_load(ggplot2)
simple_viz_obj = ggplot(Xy_simple, aes(x = first_feature, y = second_feature, color = response)) +
  geom_point(size = 5)
simple_perceptron_line = geom_abline(
    intercept = -w_vec_simple_per[1] / w_vec_simple_per[3],
    slope = -w_vec_simple_per[2] / w_vec_simple_per[3],
    color = "orange")
simple_viz_obj + simple_perceptron_line
```



Why is this line of separation not "satisfying" to you?

It its not directly in the middle of the 0's and 1's.

2. Use the `e1071` package to fit an SVM model to `y_binary` using the predictors found in `X_simple_feature_matrix`. Do not specify the λ (i.e. do not specify the `cost` argument).

```
svm_model = svm(X_simple_feature_matrix, y_binary, kernel = "linear", scale = FALSE)
```

```
## Error in svm(X_simple_feature_matrix, y_binary, kernel = "linear", scale = FALSE): could not find fu
```

and then use the following code to visualize the line in purple:

```
w_vec_simple_svm = c(
  svm_model$rho, #the b term
  -t(svm_model$coefs) %*% X_simple_feature_matrix[svm_model$index, ] # the other terms
)
```

```
## Error in eval(expr, envir, enclos): object 'svm_model' not found
```

```
simple_svm_line = geom_abline(
    intercept = -w_vec_simple_svm[1] / w_vec_simple_svm[3],
    slope = -w_vec_simple_svm[2] / w_vec_simple_svm[3],
    color = "purple")
```

```
## Error in data.frame(intercept = intercept, slope = slope): object 'w_vec_simple_svm' not found
```

```
simple_viz_obj + simple_perceptron_line + simple_svm_line
```

```
## Error in eval(expr, envir, enclos): object 'simple_svm_line' not found
```

```
parabola = function(x) {
  x^2 + 15
}
```

```
pacman::p_load(neldermead)
?optim
optim_output = optim(c(2), parabola)
```

```
## Warning in optim(c(2), parabola): one-dimensional optimization by Nelder-Mead is unreliable:
## use "Brent" or optimize() directly
```

```
optim_output
```

```
## $par
## [1] -1.776357e-15
##
## $value
## [1] 15
##
## $counts
## function gradient
##       30       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

```
w_vec = optim_output$par
```

Is this SVM line a better fit than the perceptron?

No is is a worst fit than the perceptron.

TO-DO

3. Now write pseuocode for your own implementation of the linear support vector machine algorithm respecting the following spec making use of the nelder mead `optim` function from lecture 5p. It turns out you do not need to load the package `neldermead` to use this function. You can feel free to define a function within this function if you wish.

Note there are differences between this spec and the perceptron learning algorithm spec in question #1. You should figure out a way to respect the `MAX_ITER` argument value.

For extra credit, write the actual code.

```
getMax = function(a,b){
  max(a,b)
}
SEA = function(wb_vec,x_vec, y_binar, lambda = 0.1) {
  w_vec = tail(wb_vec,-1)
  b = head(wb_vec,1)[1]
  t = 1 - y_binar * (x_vec %*% w_vec - b)
  maxes = sapply(t, getMax, b=0)
  mean(maxes) #+ (lambda * (w_vec%*%w_vec))
}
```

```r
DUAL = function(a,x_vec, y_binar) {
  accumulator = 0;
  for( i in 1:length(a)) {
    for (j in i: length(a)) {
      term = (a[i]*a[j])*(y_binar[i]%*%y_binar[j])*(x_vec[i]%*%x_vec[j])
      accumulator = accumulator + term
    }
  }

  sum(a) - (0.5 * accumulator)
}
```

```r
a = rep(0,nrow(X_simple_feature_matrix))
opt = optim(a,DUAL, y_binar = y_binary, x_vec = X_simple_feature_matrix,)#sap = SEA(w_vec,b)
opt
```

```
## $par
## [1] -1.375516e+16 -1.457093e+16 -1.219953e+16  9.827121e+15  1.124789e+16
## [6]  1.283665e+16
##
## $value
## [1] -4.442731e+33
##
## $counts
## function gradient
##      501       NA
##
## $convergence
## [1] 1
##
## $message
## NULL
```

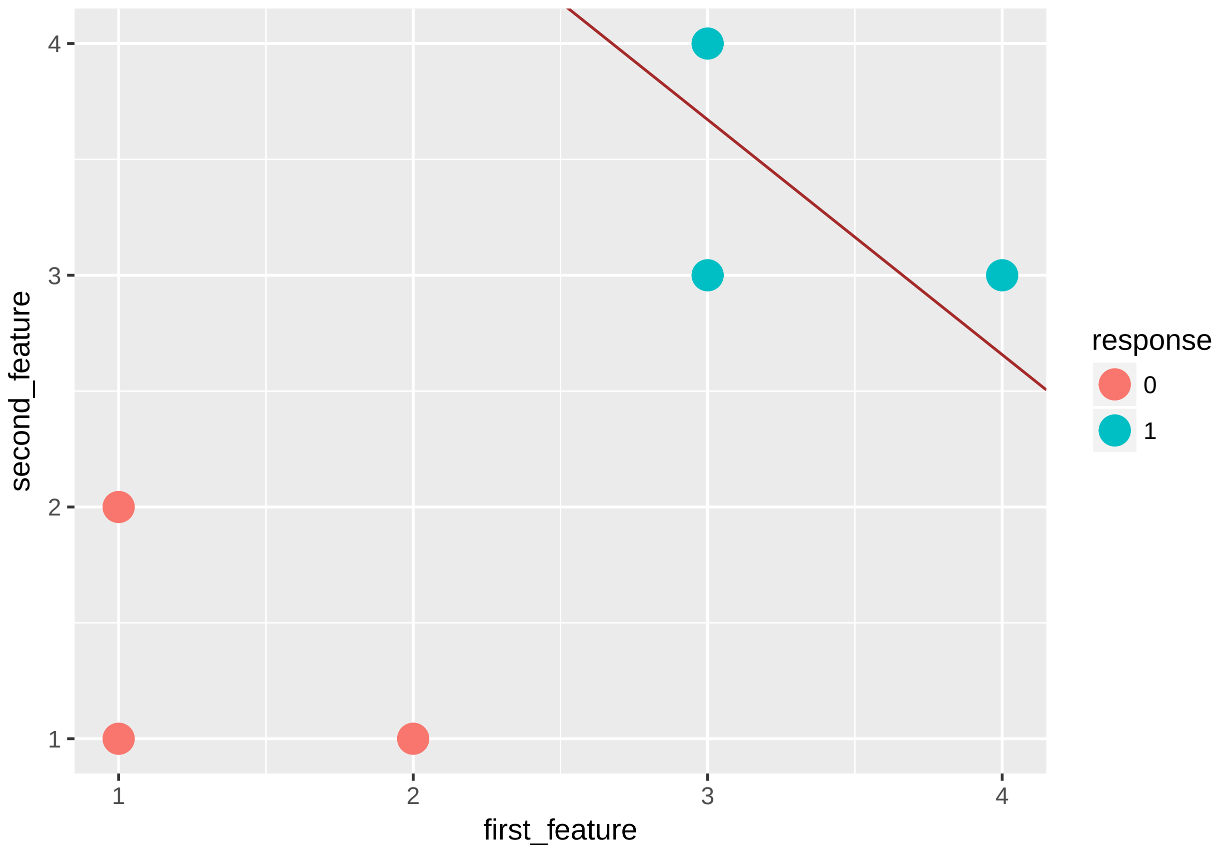
```r
#' This function implements the hinge-loss + maximum margin linear support vector machine algorithm of
#'
#' @param Xinput      The training data features as an n x p matrix.
#' @param y_binary    The training data responses as a vector of length n consisting of only 0's and 1'
#' @param MAX_ITER    The maximum number of iterations the algorithm performs. Defaults to 5000.
#' @param lambda      A scalar hyperparameter trading off margin of the hyperplane versus average hinge
#'                    The default value is 1.
#' @return            The computed final parameter (weight) as a vector of length p + 1
linear_svm_learning_algorithm = function(Xinput, y_binary, MAX_ITER = 5000, lambda = 0.1){
  #TO-DO
  #control=list(fnscale=-1)
  #w_vec = rep(0,ncol(Xinput) + 1)
  a_vec = rep(0,nrow(Xinput))
  #optimal_SEA = optim(w_vec,SEA, y_binar = y_binary, x_vec = Xinput, lambda = 0.1)
  optimal = optim(a_vec,DUAL, y_binar = y_binary, x_vec = Xinput)
  a_vecs = optimal$par
  print(a_vecs)
  a_support_indices = which(a_vecs > 0) #get support vectors

  mat = cbind(Xinput,y_binary,a_vecs)
  S = mat[a_support_indices,]
  bias = numeric(nrow(S))
  for (i in 1:length(bias)) {
    print(S[i,1:2])
    bias[i] = S[i,3] - sum( S[,3] * S[,4] * (S[,1:2] %*% S[i,1:2]) )
  }

  b = mean(bias)
  print (b)
  w_vec = c(b,colSums(a_vecs * y_binary * Xinput))
  w_vec
}
```

If you wrote code (the extra credit), run your function using the defaults and plot it in brown vis-a-vis the previous model's line:

```
svm_model_weights = linear_svm_learning_algorithm(X_simple_feature_matrix, y_binary)
```

```
## [1] -1.375516e+16 -1.457093e+16 -1.219953e+16  9.827121e+15  1.124789e+16
## [6]  1.283665e+16
##  first_feature second_feature
##              3              3
##  first_feature second_feature
##              3              4
##  first_feature second_feature
##              4              3
## [1] -7.585152e+17
```

```
my_svm_line = geom_abline(
    intercept = -svm_model_weights[1] / svm_model_weights[3],#NOTE: negative sign removed from intercep
    slope = -svm_model_weights[2] / svm_model_weights[3],
    color = "brown")
simple_viz_obj  + my_svm_line
```



Is this the same as what the `e1071` implementation returned? Why or why not?

4. Write a $k = 1$ nearest neighbor algorithm using the Euclidean distance function. Respect the spec below:

```r
#' This function implements the nearest neighbor algorithm.
#'
#' @param Xinput      The training data features as an n x p matrix.
#' @param y_binary    The training data responses as a vector of length n consisting of only 0's and 1'
#' @param Xtest       The test data that the algorithm will predict on as a n* x p matrix.
#' @return            The predictions as a n* length vector.
nn_algorithm_predict = function(Xinput, y_binary, Xtest){
  #TO-DO
  max_dist = Inf
  best_index = 0;
  for (i in 1:nrow(Xinput)) {
    test_dist = (Xinput[i,]-Xtest) %*% (Xinput[i,]-Xtest)
    if (test_dist < max_dist){
      max_dist = test_dist
      best_index = i
    }
  }
  y_binary[best_index]
}
```

Write a few tests to ensure it actually works:

```r
#TO-DO
library('testthat')
```

```r
expect_equal(nn_algorithm_predict(Xinput = X_simple_feature_matrix, y_binary = y_binary, Xtest = c(4,3)
expect_equal(nn_algorithm_predict(Xinput = X_simple_feature_matrix, y_binary = y_binary, Xtest = c(3,3)
expect_equal(nn_algorithm_predict(Xinput = X_simple_feature_matrix, y_binary = y_binary, Xtest = c(3,4)

expect_equal(nn_algorithm_predict(Xinput = X_simple_feature_matrix, y_binary = y_binary, Xtest = c(1,1)
expect_equal(nn_algorithm_predict(Xinput = X_simple_feature_matrix, y_binary = y_binary, Xtest = c(1,2)
expect_equal(nn_algorithm_predict(Xinput = X_simple_feature_matrix, y_binary = y_binary, Xtest = c(2,1)

expect_equal(nn_algorithm_predict(Xinput = X_simple_feature_matrix, y_binary = y_binary, Xtest = c(3.5,

expect_equal(nn_algorithm_predict(Xinput = X_simple_feature_matrix, y_binary = y_binary, Xtest = c(5,5)

expect_equal(nn_algorithm_predict(Xinput = X_simple_feature_matrix, y_binary = y_binary, Xtest = c(2.5,
```

For extra credit, add an argument `k` to the `nn_algorithm_predict` function and update the implementation so it performs KNN. In the case of a tie, choose $\hat{y}$ randomly. Set the default `k` to be the square root of the size of $\mathcal{D}$ which is an empirical rule-of-thumb popularized by the "Pattern Classification" book by Duda, Hart and Stork (2007). Also, alter the documentation in the appropriate places.

```r
#not required TO-DO --- only for extra credit




nn_algorithm_predict = function(Xinput, y_binary, Xtest, k = 1){
  #TO-DO
  max_dist = Inf
  best_index = 0;
  dist_ary = as.numeric(nrow(Xinput))
  for (i in 1:nrow(Xinput)) {
    test_dist = (Xinput[i,]-Xtest) %*% (Xinput[i,]-Xtest)
    dist_ary[i] = test_dist
  }

  dist_ary_ind = order(dist_ary)

  chosen_indices = dist_ary_ind[1:k]
  chosen_y = y_binary[chosen_indices]
  binchoose = mean(chosen_y)
  final = ifelse( binchoose == 0.05, rbinom(1,1,0.5), round(binchoose) )
  final
}

#TO-DO
library('testthat')
expect_equal(nn_algorithm_predict(Xinput = X_simple_feature_matrix, y_binary = y_binary, Xtest = c(4,3)

expect_equal(nn_algorithm_predict(Xinput = X_simple_feature_matrix, y_binary = y_binary, Xtest = c(4,3)

expect_equal(nn_algorithm_predict(Xinput = X_simple_feature_matrix, y_binary = y_binary, Xtest = c(4,2)
```

For extra credit, in addition to the argument `k`, add an argument `d` representing any legal distance function to the `nn_algorithm_predict` function. Update the implementation so it performs KNN using that distance function. Set the default function to be the Euclidean distance in the original function. Also, alter the documentation in the appropriate places.

```r
euclid = function(u,v) {
  sum((u-v)^2)
}

nn_algorithm_predict = function(Xinput, y_binary, Xtest, k = 1, d= euclid){
  #TO-DO
  max_dist = Inf
  best_index = 0;
  dist_ary = as.numeric(nrow(Xinput))
  for (i in 1:nrow(Xinput)) {
    test_dist = d(Xinput[i,],Xtest)
    dist_ary[i] = test_dist
  }

  dist_ary_ind = order(dist_ary)

  chosen_indices = dist_ary_ind[1:k]
  chosen_y = y_binary[chosen_indices]
  binchoose = mean(chosen_y)
  final = ifelse( binchoose == 0.05, rbinom(1,1,0.5), round(binchoose) )
  final
}

expect_equal(nn_algorithm_predict(Xinput = X_simple_feature_matrix, y_binary = y_binary, Xtest = c(4,2)
```

5. We move on to simple linear modeling using the ordinary least squares algorithm.

Let's quickly recreate the sample data set from practice lecture 7:

```r
n = 20
x = runif(n)
beta_0 = 3
beta_1 = -2
y = beta_0 + beta_1 * x + rnorm(n, mean = 0, sd = 0.33)
```

Solve for the least squares line by computing b_0 and b_1 *without* using the functions `cor`, `cov`, `var`, `sd` but instead computing it from the x and y quantities manually. See the class notes.

```r
#TO-DO

SSE = function(B,X,Y) {
  sum((B[1] + (B[2] * X) - Y)^2)
}
Bs = c(1,1)

Bs = optim(par=Bs,SSE, X=x, Y=y )$par

b_0 = Bs[1]
b_1 = Bs[2]

Bs
```

```
## [1]  3.058722 -2.303047
```

Verify your computations are correct using the `lm` function in R:

```
lm_mod = lm( y ~ x)
b_vec = coef(lm_mod)
expect_equal(b_0, as.numeric(b_vec[1]), tol = 1e-4) #thanks to Rachel for spotting this bug - the b_vec
expect_equal(b_1, as.numeric(b_vec[2]), tol = 1e-4)
```

6. We are now going to repeat one of the first linear model building exercises in history — that of Sir Francis Galton in 1886. First load up package `HistData`.

```
library('HistData')
```

In it, there is a dataset called `Galton`. Load it using the `data` command:

```
data('Galton')
```

You now should have a data frame in your workspace called `Galton`. Summarize this data frame and write a few sentences about what you see. Make sure you report n, p and a bit about what the columns represent and how the data was measured. See the help file `?Galton`.

```
#TO-DO
```

The dataset is a survey of the average height of parents plotted against the height of their child. The independent variable is average parent height while the dependent variable is the height of the child. n = 928 (children) and p = 2.

Find the average height (include both parents and children in this computation).

```
avg_height = mean(unlist(Galton))
```

Note that in Math 241 you learned that the sample average is an estimate of the "mean", the population expected value of height. We will call the average the "mean" going forward since it is probably correct to the nearest tenth of an inch with this amount of data.

Run a linear model attempting to explain the childrens' height using the parents' height. Use `lm` and use the R formula notation. Compute and report b_0, b_1, RMSE and R^2. Use the correct units to report these quantities.

```
#TO-DO

model = lm ( child ~ parent, data= Galton)
bvals = coef(model)
b_0 = bvals[1]
b_1 = bvals[2]

yhat = b_0 + b_1 * x #this is the g(x^*) function!
e = y - yhat
sse = sum(e^2)
mse = sse / length(y)
rmse = sqrt(mse)
sse
```

```
## [1] 9801.858
```

```
mse
```

```
## [1] 490.0929
```

```
rmse
```

```
## [1] 22.13804
```

```
s_sq_y = var(y)
s_sq_e = var(e)
```

```r
rsq = (s_sq_y - s_sq_e) / s_sq_y
rsq
```

```
## [1] -0.4916982
```

Interpret all four quantities: b_0, b_1, RMSE and R^2.

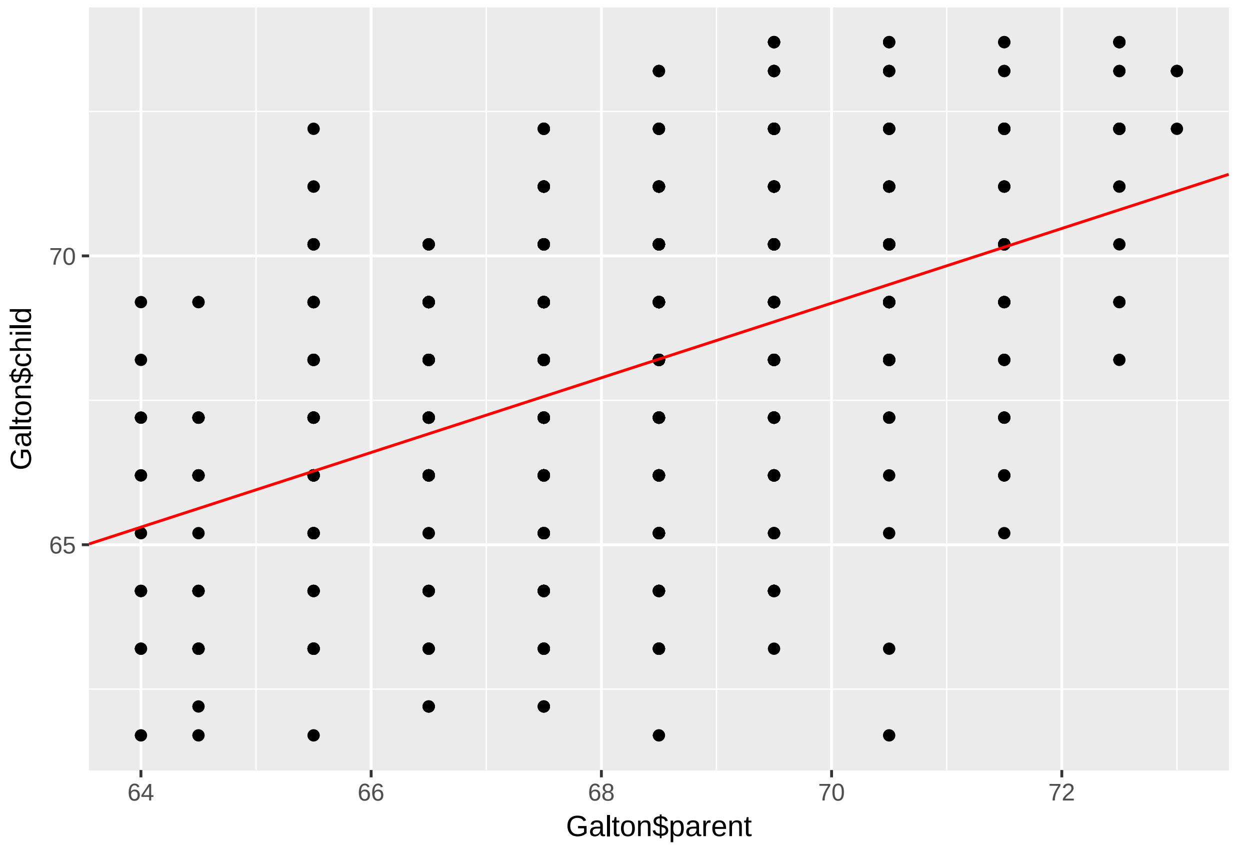
TO-DO

How good is this model? How well does it predict? Discuss.

TO-DO

Now use the code from practice lecture 8 to plot the data and a best fit line using package `ggplot2`. Don't forget to load the library.

```r
simple_viz_obj = ggplot(Galton, aes(x = Galton$parent, y = Galton$child)) + geom_point()
simple_ls_regression_line = geom_abline(intercept = bvals[1], slope = bvals[2], color = "red")
simple_viz_obj + simple_ls_regression_line
```



It is reasonable to assume that parents and their children have the same height. Explain why this is reasonable using basic biology.

Height is a genetic trait which is is information passed from parent to child via the inheritance of genes.

If they were to have the same height and any differences were just random noise with expectation 0, what would the values of β_0 and β_1 be?

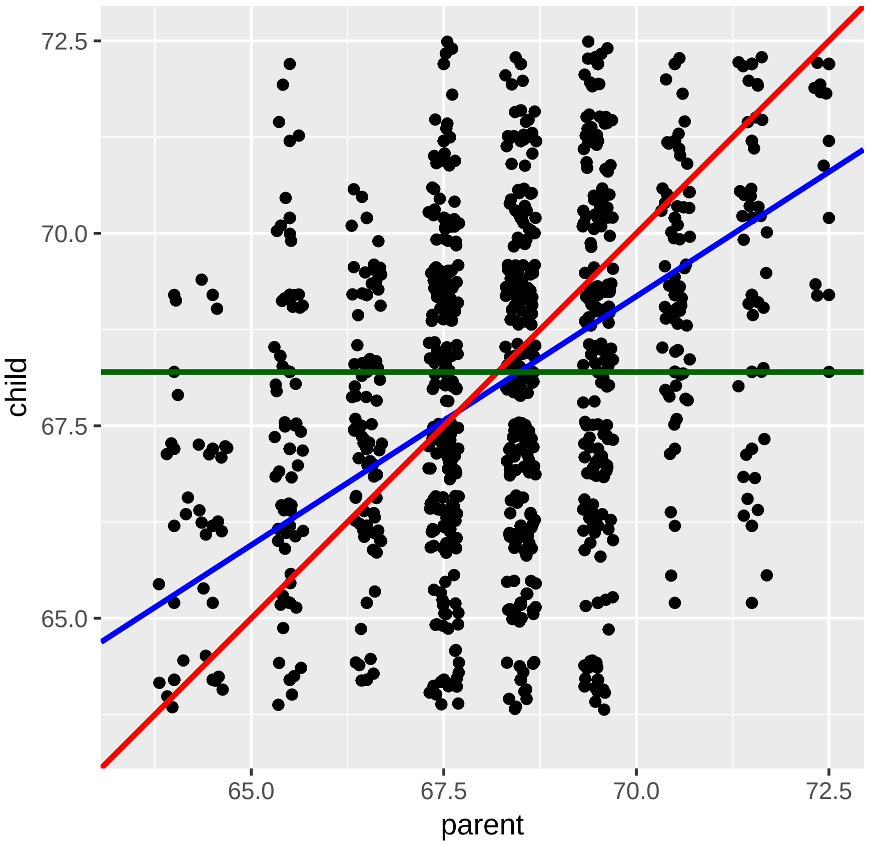
$\beta_0 = 1$ and $\beta_1 = 1$

Let's plot (a) the data in $\mathbb{D}$ as black dots, (b) your least squares line defined by b_0 and b_1 in blue, (c) the theoretical line β_0 and β_1 if the parent-child height equality held in red and (d) the mean height in green.

```r
ggplot(Galton, aes(x = parent, y = child)) +
  geom_point() +
  geom_jitter() +
  geom_abline(intercept = b_0, slope = b_1, color = "blue", size = 1) +
  geom_abline(intercept = 0, slope = 1, color = "red", size = 1) +
  geom_abline(intercept = avg_height, slope = 0, color = "darkgreen", size = 1) +
  xlim(63.5, 72.5) +
  ylim(63.5, 72.5) +
  coord_equal(ratio = 1)
```

```
## Warning: Removed 76 rows containing missing values (geom_point).
```

```
## Warning: Removed 86 rows containing missing values (geom_point).
```



Fill in the following sentence:

TO-DO: Children of short parents became [taller] on average and children of tall parents became [shorter] on average.

Why did Galton call it "Regression towards mediocrity in hereditary stature" which was later shortened to "regression to the mean"?

The height lies on a distribution with a mean, a child is more likely to have a height closer to the mean than one farther form it.

Why should this effect be real?

If it weren't real we would have a runaway effect where tall parents gave birth to taller children and short parents to shorter children, over the generations, we would have a population of giants and dwarfs, but this is not the case obviously.

You now have unlocked the mystery. Why is it that when modeling with y continuous, everyone calls it "regression"? Write a better, more descriptive and appropriate name for building predictive models with y continuous.

To generate the model line, we first picked the true Betas B_0 and B_1 and then added random noise to generate points that lie close to but not exactly on the line.

Regression reverses this process, the purpose is to squish or regress the points (or regress through time) back to the mean value that lies on the line.