HW02p

Moshe Dinowitz

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
#'                    \code{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 = 1000, w = NULL){
    if (is.null(w)){
    w = runif(ncol(Xinput)) #intialize a p+1-dim vector with random values
  }
  for (iter in 1 : MAX_ITER){
    for (i in 1 : nrow(Xinput)){
      x_i = Xinput[i, ]
      yhat_i = ifelse(x_i %*% w > 0, 1, 0)
      w = w + as.numeric(y_binary[i] - yhat_i) * x_i
    }
  }
  w
}
```

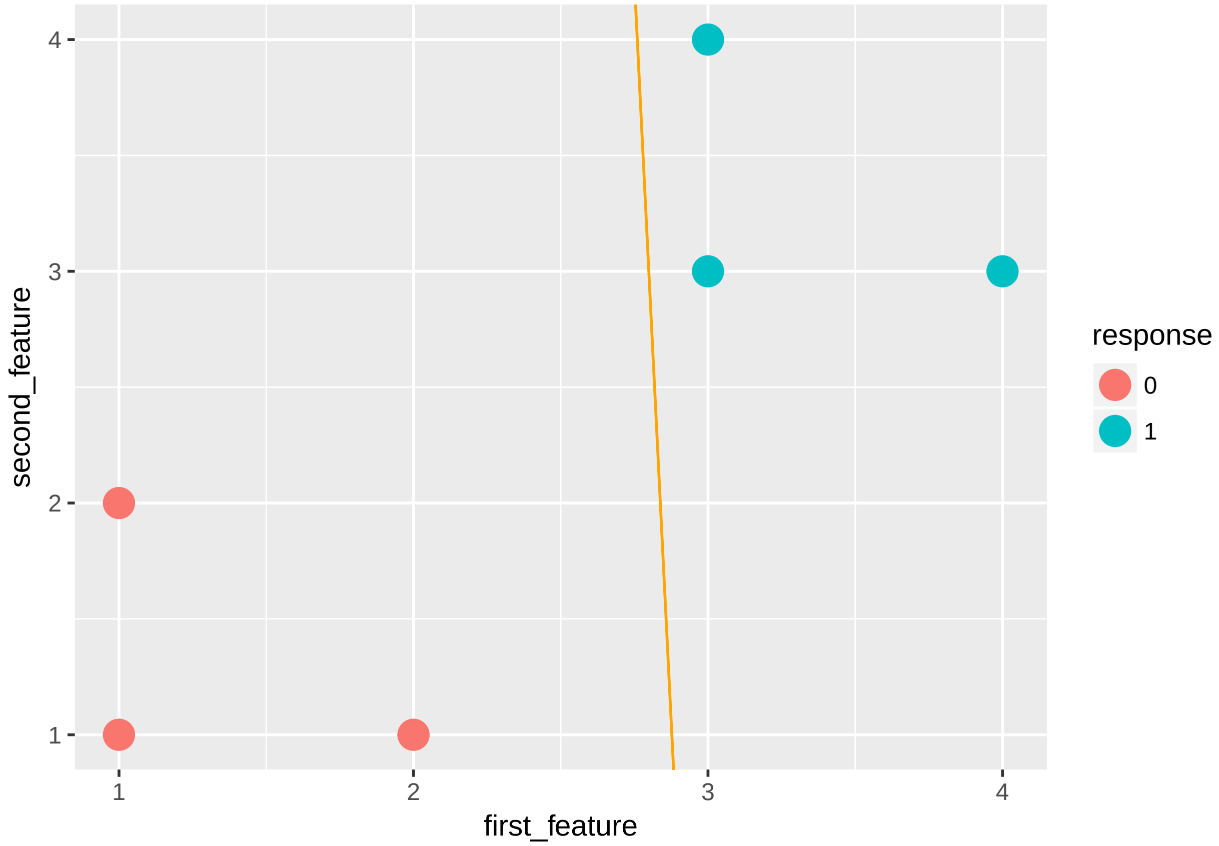
Run the code on the simple dataset above via:

```
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] -8.4063429  2.8825014  0.1126945
```

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?

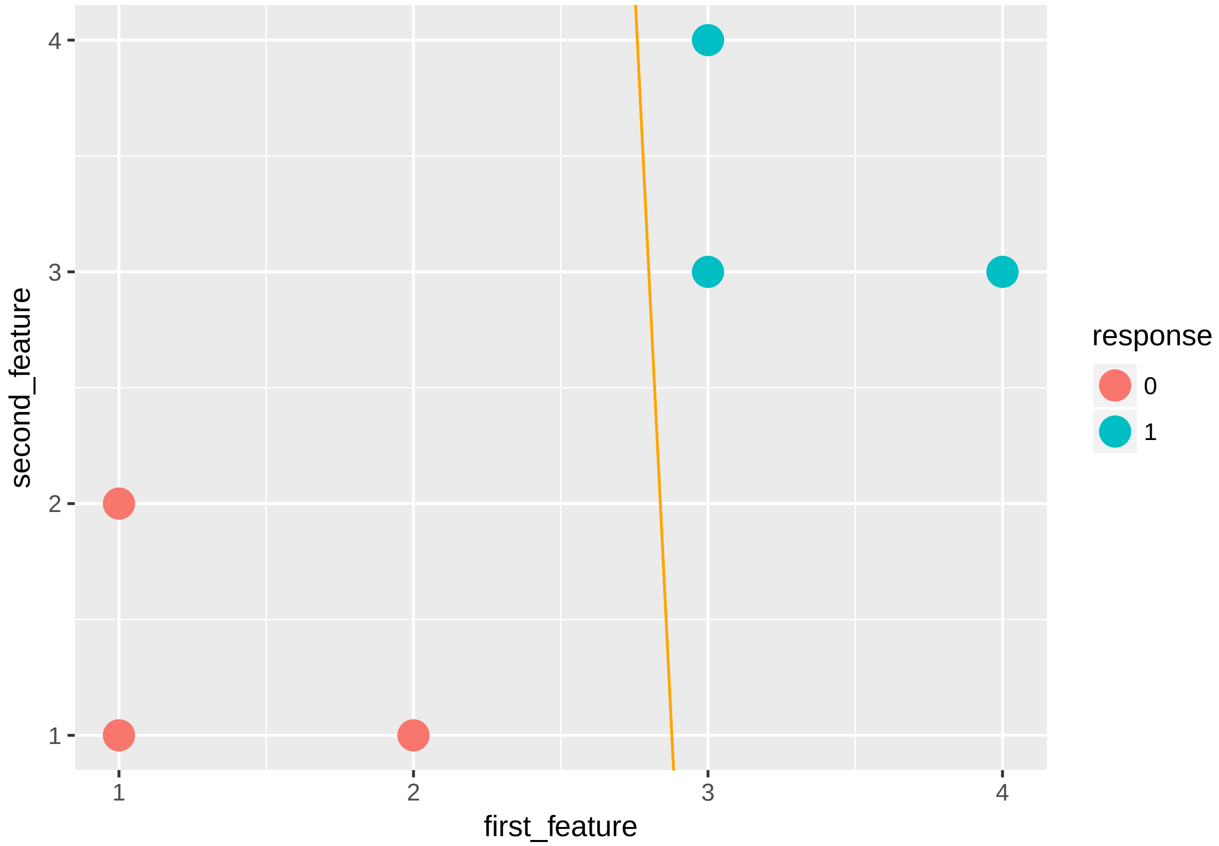
Because it seems very arbitrary

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).

```
pacman::p_load(e1071)
svm_model = svm(X_simple_feature_matrix, y_binary, kernel = "linear")
```

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
)
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")
simple_viz_obj + simple_perceptron_line + simple_svm_line
```



Is this SVM line a better 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.

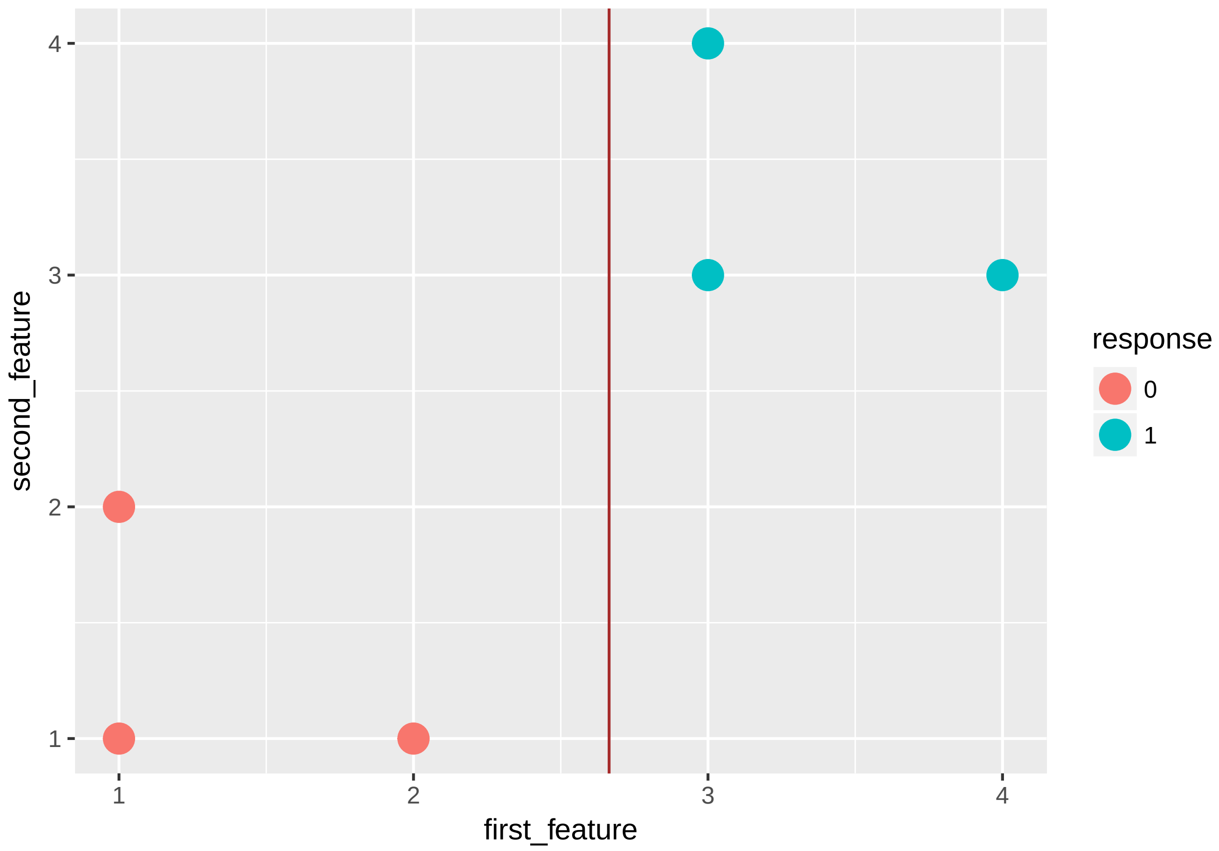
Note there are differences between this spec and the perceptron learning algorithm spec in question #1. You should figure out a way to respect the `MAX_ITER` argument value.

For extra credit, write the actual code.

```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 = 1){
  n = nrow(Xinput)
  p = ncol(Xinput)
  metric = c()
  vap = function(w_vec){
    for(j in 1:n){
      metric[j] = max(0, (.5-(y_binary[j] - .5)*(sum(w_vec[2:3] * Xinput[j, ]) - w_vec[1])))
    }
    sum(metric)/n + (lambda * sum(w_vec[2:p+1]^2))
  }
  V = optim(rep(0, p+1), vap, control = list(maxit = MAX_ITER))
  V$par
}
```

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:

```r
svm_model_weights = linear_svm_learning_algorithm(X_simple_feature_matrix, y_binary)
my_svm_line = geom_abline(
    intercept = svm_model_weights[1] / svm_model_weights[3],#NOTE: negative sign removed from intercept
    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? no, it optimized for lambda as well.

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

```
#' 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){
  predict = c()
  best_sqd_distance = Inf
  i_star = NA
  for(m in 1 : nrow(Xtest)){
    best_sqd_distance = Inf
    for (i in 1 : nrow(Xinput)){
      dsqd = sum((Xinput[i, ] - Xtest[m, ])^2)
      if (dsqd < best_sqd_distance){
        best_sqd_distance = dsqd
        i_star = i
      }
    }
    predict[m] = y_binary[i_star]
  }
```

```r
  predict
}
```

Write a few tests to ensure it actually works:

```r
if((identical(nn_algorithm_predict(X_simple_feature_matrix, y_binary, X_simple_feature_matrix), y_binar
    print("test failed")
}
```

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
#' This function implements the k-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.
#' @param k           The number of neighbors we're searching for
#' @return            The predictions as a n* length vector.
knn_algorithm_predict = function(Xinput, y_binary, Xtest, k = sqrt(nrow(Xinput))){
  Mode <- function(x){
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
  }
  predict = c()
  for(m in 1 : nrow(Xtest)){
    metric = c()
    for (j in 1:nrow(Xinput)){
      metric[j] = sum((Xinput[j, ] - Xtest[m, ])^2)
    }
    o = order(metric)[1:k]
    out = y_binary[o]
    predict[m] = Mode(out)
  }
  predict
}

knn_algorithm_predict(matrix(c(1, 3, 5, 7)), matrix(c(1, 0, 1, 2)), matrix(c(2, 4)), 1)
```

```
## [1] 1 0
```

```r
c(2, 5, 7, 19)[c(1, 3, 2)]
```

```
## [1] 2 7 5
```

```r
Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
Mode(c(1, 2, 1, 2))
```

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

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.

```
#' This function implements the k-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.
#' @param k           The number of neighbors we're searching for
#' @param d           A distance function
#' @return            The predictions as a n* length vector.
knn_algorithm_predict_distance = function(Xinput, y_binary, Xtest, k, d = function(x, y){sum((x-y)^2)})
  Mode <- function(x){
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
  }
  predict = c()
  for(m in 1 : nrow(Xtest)){
    metric = c()
    for (j in 1:nrow(Xinput)){
      metric[j] = d(Xinput[j, ], Xtest[m, ])
    }
    o = order(metric)[1:k]
    out = y_binary[o]
    predict[m] = Mode(out)
  }
  predict
}
```

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:

```
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.

```
b_1 = (sum(y*x)-(sum(x)*sum(y)/n))/(sum(x^2)-((sum(x)^2) / n))
mean(x)*mean(y)*n
```

```
## [1] 14.86829
```

```
b_0 = (sum(y) - (b_1 * sum(x))) / n
```

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`.

```
pacman::p_load(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`.

```
summary(Galton)
```

```
##      parent          child      
##  Min.   :64.00   Min.   :61.70  
##  1st Qu.:67.50   1st Qu.:66.20  
##  Median :68.50   Median :68.20  
##  Mean   :68.31   Mean   :68.09  
##  3rd Qu.:69.50   3rd Qu.:70.20  
##  Max.   :73.00   Max.   :73.70
```

```
head(Galton)
```

```
##   parent child
## 1   70.5  61.7
## 2   68.5  61.7
## 3   65.5  61.7
## 4   64.5  61.7
## 5   64.0  61.7
## 6   67.5  62.2
```

```
n = nrow(Galton)
p = ncol(Galton)
```

TO-DO

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

```
avg_height = (sum(Galton[1]) + sum(Galton[2]))/(2*n)
```

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.

```
lm_G = lm(child ~ parent, Galton)
b_0 = coef(lm_G)[1]
b_1 = coef(lm_G)[2]
summary(lm_G)$r.squared
```

```
## [1] 0.2104629
```

```
summary(lm_G)$sigma
```

```
## [1] 2.238547
```

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

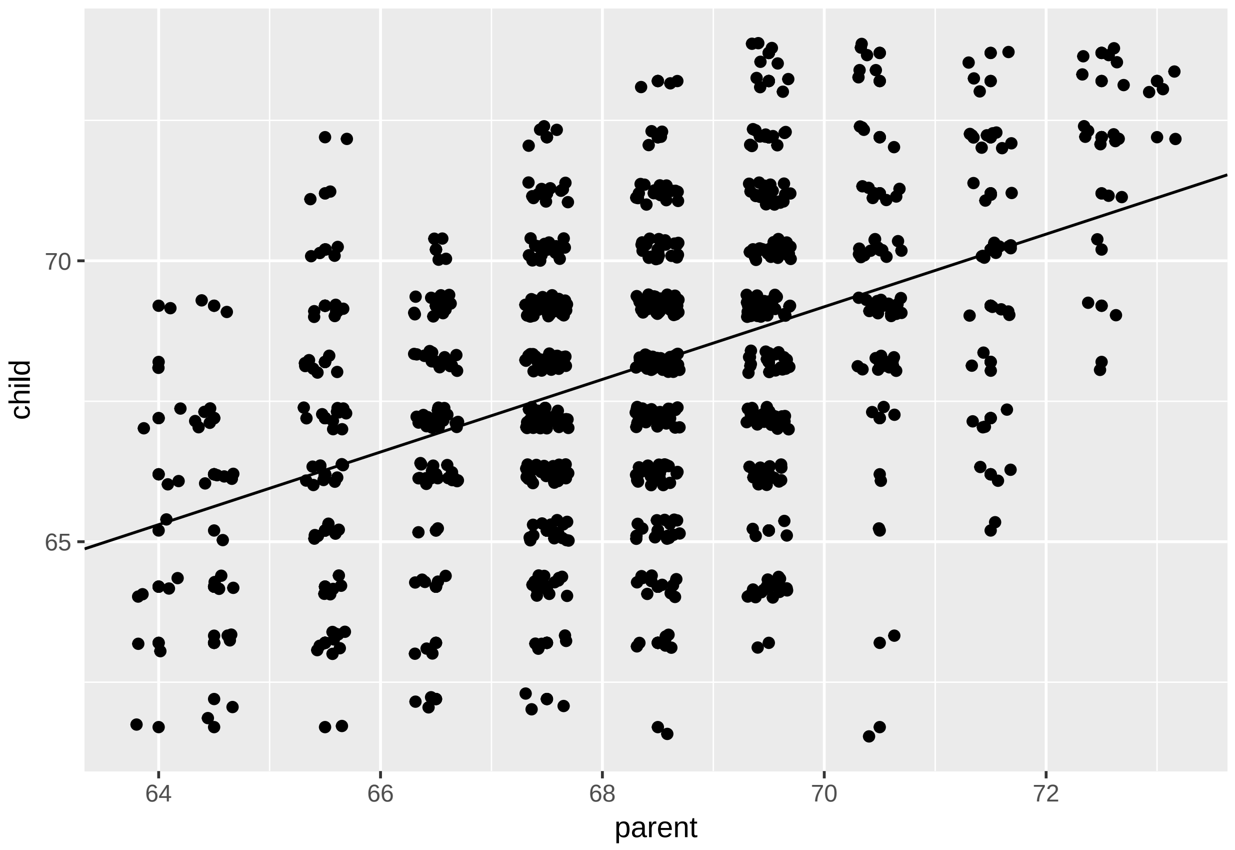
b_0 is the intercept, it means that if the parents had zero height, the child would be predicted to have b_0 height. It's not very interpretable. b_1 is the slope. if you look at two parent's whose height differs by one, their childrens height is predicted to differ by the slope. R squared is telling us that we're doing 21% better than just picking the threshold model. RMSE is telling us that 95% of the heights of children will be plus or minus 4.46 away from the mean.

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

R^2 of 21% is pretty bad. That being said predicting within 4 inches 95% of the time is pretty good. So it's an alright model.

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
pacman::p_load(ggplot2)
ggplot(Galton, aes(x = parent, y = child)) +
  geom_point() + geom_jitter() + geom_abline(intercept = b_0, slope = b_1)
```



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

Acoording to biology approximatly 80% of height is genetic so it's pretty reasonable

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?

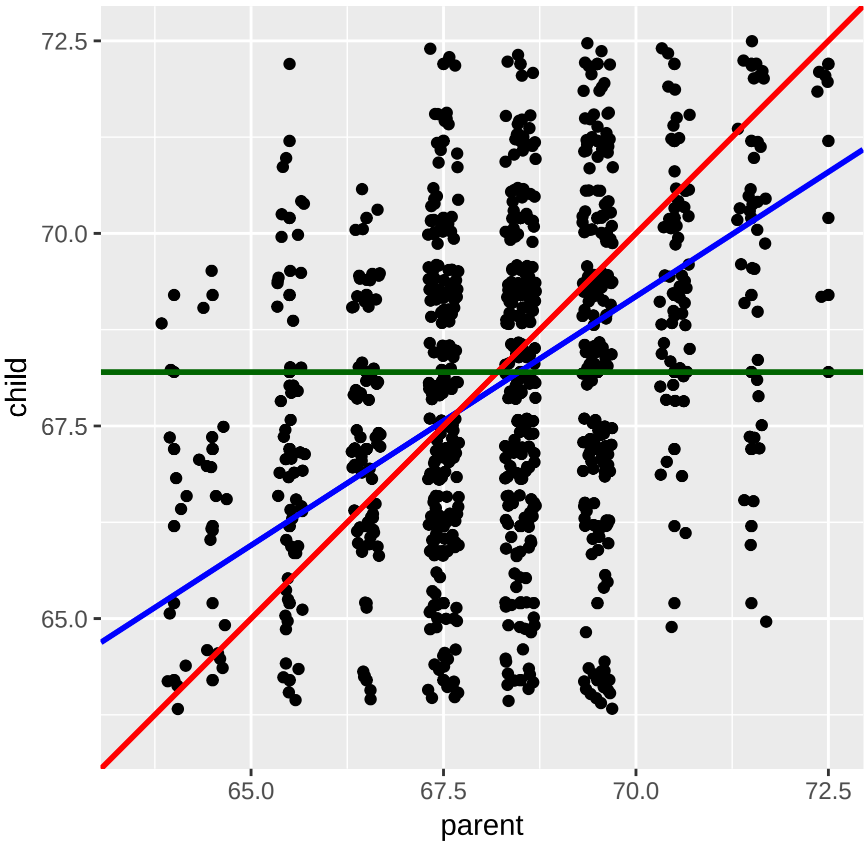
b_0 = 0, b_1 = 1

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.

```
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 88 rows containing missing values (geom_point).
```



Fill in the following sentence:

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”?

Because it looks like data at extremum tends to go towards the mean over time

Why should this effect be real?

If there’s a high chance a tall person’s child will be the same height as his parent’s, a low chance they’ll be taller, and a decent chance they’ll be shorter, then with a lot of data you will see more people tend towards the mean than to either extremum.

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.

Because as you add more data points, your model “regresses” towards the mean. I would call models with y continuous optimizations, because you usually have to optimize a metric in order to build your model.