Lab 4

Vincent Miceli

11:59PM March 9, 2019

Note: the content of this lab is on the midterm exam (March 5) even though the lab itself is due after the midterm exam.

We now 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 `mean`, `cor`, `cov`, `var`, `sd` but instead computing it from the x and y quantities manually using base function such as `sum` and other basic operators. See the class notes.

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

```
## [1] -1.887044
```

```
b_0 = ybar - b_1 * xbar
b_0
```

```
## [1] 2.877937
```

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

```
lm_mod = lm(y ~ x)
b_vec = coef(lm_mod)
pacman::p_load(testthat)
expect_equal(b_0, as.numeric(b_vec[1]), tol = 1e-4)
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`.

```
if (!require("pacman")){install.packages("pacman")}
```

```
## Loading required package: pacman
```

```
pacman::p_load(HistData)
```

In it, there is a dataset called `Galton`. Load it up.

```
#HistData::Galton
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
```

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`.

It is a dataset of parents heights and their child's height. The child height seems to match the parents height closely.

```
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
```

```
n = 928
p = 2
?Galton
```

```
## starting httpd help server ... done
```

TO-DO

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

```
avg_height = (sum(Galton$parent) * 2 + sum(Galton$child)) / (3*n)
avg_height
```

```
## [1] 68.23495
```

If you were to use the null model, what would the RMSE be of this model be?

```
null = mean(Galton$child)
y = Galton$child
e = y - null

SSE = sum(e^2)

1/(n-2) * SSE -> MSE

RMSE = sqrt(MSE)
RMSE
```

```
## [1] 2.519301
```

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.

```r
model = lm(Galton$child ~ Galton$parent)
model
```

```
##
## Call:
## lm(formula = Galton$child ~ Galton$parent)
##
## Coefficients:
##   (Intercept)  Galton$parent
##       23.9415         0.6463
```

```r
b_vec = coef(model)
b0 = b_vec[1]
b1 = b_vec[2]

b0
```

```
## (Intercept)
##    23.94153
```

```r
b1
```

```
## Galton$parent
##     0.6462906
```

```r
SSE = sum((model$residuals)^2)

1/(n-2) * SSE -> MSE

RMSE = sqrt(MSE)
RMSE
```

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

```r
summary(model)$r.squared
```

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

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

b_0, the intercept of our line, is 23.94. b_1, the weight given to our feature parent height, is .646. The RMSE and R2 of our model are 2.24 and .21, so our model is only slightly better than the null model.

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

The model is not good because it is only slightly better than the null model; it has a similar RMSE and a low R2 value. It does not predict well.

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

It is reasonable to assume that children with taller parents will generally be taller due to having the same genes as their parents. The problem is children have 2 parents, so looking at just the average of both parents height does not create a good model. A child with one short parent and one extremely tall parent could take after either parent.

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?

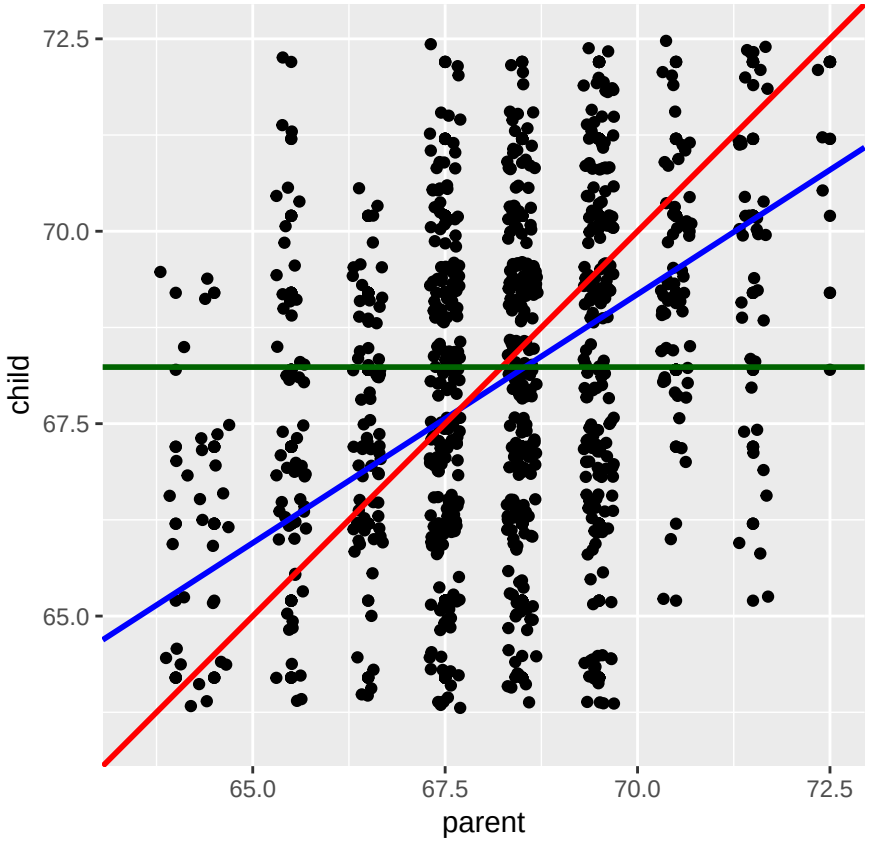
beta_0 would be 0, and beta_1 would be 1.

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.

```
pacman::p_load(ggplot2)
ggplot(Galton, aes(x = parent, y = child)) +
  geom_point() +
  geom_jitter() +
  geom_abline(intercept = b0, slope = b1, 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 90 rows containing missing values (geom_point).
```



Fill in the following sentence:

Children of short parents became short on average and children of tall parents became tall on average.

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

As heights of the parents deviated from the average height, their children tended to be less extreme in height. That is, the heights of the children regressed to the average height of an adult.

Why should this effect be real?

This effect should be real in this case because there are two parents, so children with one outlier parent will likely be closer to the mean in height than the average of their parents' heights.

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.

It is called regression because of Galton's discovery of the tendency of extreme data values to regress toward the mean. A more appropriate name would be continuous modelling.

Create a dataset $\mathbb{D}$ which we call `Xy` such that the linear model as R^2 about 50% and RMSE approximately 1.

```
x = 1:50
y = x + rnorm(50, mean = 25, sd = 15)
Xy = data.frame(x = x, y = y)

model = lm(y ~ x)
summary(model)$r.squared
```

```
## [1] 0.5734766
```

Create a dataset $\mathbb{D}$ which we call `Xy` such that the linear model as R^2 about 0% but x, y are clearly associated.

```
x = 1:50
y = x^75
Xy = data.frame(x = x, y = y)

model = lm(y ~ x)
summary(model)$r.squared
```

```
## [1] 0.0905152
```

Load up the famous iris dataset and drop the data for Species "virginica".

```
data(iris)
new = iris[iris$Species != "virginica", ]
head(new)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

If the only input x is Species and you are trying to predict y which is Petal.Length, what would a reasonable, naive prediction be under both Species? Hint: it's what we did in class.

```
mean(new[iris$Species == "setosa", "Petal.Length"])
```

```
## [1] 1.462
```

```
mean(new[iris$Species == "versicolor", "Petal.Length"])
```

```
## [1] 4.26
```

Prove that this is the OLS model by fitting an appropriate `lm` and then using the predict function to verify you get the same answers as you wrote previously.

```
model = lm(new$Petal.Length ~ new$Species)
predict(model, new)
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12
## 1.462 1.462 1.462 1.462 1.462 1.462 1.462 1.462 1.462 1.462 1.462 1.462
##    13    14    15    16    17    18    19    20    21    22    23    24
## 1.462 1.462 1.462 1.462 1.462 1.462 1.462 1.462 1.462 1.462 1.462 1.462
##    25    26    27    28    29    30    31    32    33    34    35    36
## 1.462 1.462 1.462 1.462 1.462 1.462 1.462 1.462 1.462 1.462 1.462 1.462
##    37    38    39    40    41    42    43    44    45    46    47    48
## 1.462 1.462 1.462 1.462 1.462 1.462 1.462 1.462 1.462 1.462 1.462 1.462
##    49    50    51    52    53    54    55    56    57    58    59    60
## 1.462 1.462 4.260 4.260 4.260 4.260 4.260 4.260 4.260 4.260 4.260 4.260
##    61    62    63    64    65    66    67    68    69    70    71    72
## 4.260 4.260 4.260 4.260 4.260 4.260 4.260 4.260 4.260 4.260 4.260 4.260
##    73    74    75    76    77    78    79    80    81    82    83    84
## 4.260 4.260 4.260 4.260 4.260 4.260 4.260 4.260 4.260 4.260 4.260 4.260
##    85    86    87    88    89    90    91    92    93    94    95    96
## 4.260 4.260 4.260 4.260 4.260 4.260 4.260 4.260 4.260 4.260 4.260 4.260
##    97    98    99   100
## 4.260 4.260 4.260 4.260
```