# Comprehension Check: Regularization

The exercises in Q1-Q8 work with a simulated dataset for 1000 schools. This pre-exercise setup walks you through the code needed to simulate the dataset.

An education expert is advocating for smaller schools. The expert bases this recommendation on the fact that among the best performing schools, many are small schools. Let's simulate a dataset for 1000 schools. First, let's simulate the number of students in each school, using the following code:

```
set.seed(1986) #for R 3.5 or earlier
#if using R 3.6 or later, use `set.seed(1986, sample.kind="Rounding")` instead
n <- round(2^rnorm(1000, 8, 1))
```

Now let's assign a **true** quality for each school that is completely independent from size. This is the parameter we want to estimate in our analysis. The true quality can be assigned using the following code:

```
set.seed(1) #for R 3.5 or earlier
#if using R 3.6 or later, use `set.seed(1, sample.kind="Rounding")` instead
mu <- round(80 + 2*rt(1000, 5))
range(mu)
schools <- data.frame(id = paste("PS",1:1000),
                      size = n,
                      quality = mu,
                      rank = rank(-mu))
```

We can see the top 10 schools using this code:

```
schools %>% top_n(10, quality) %>% arrange(desc(quality))
```

Now let's have the students in the school take a test. There is random variability in test taking, so we will simulate the test scores as normally distributed with the average determined by the school quality with a standard deviation of 30 percentage points. This code will simulate the test scores:

```
set.seed(1) #for R 3.5 or earlier
#if using R 3.6 or later, use `set.seed(1, sample.kind="Rounding")` instead
mu <- round(80 + 2*rt(1000, 5))

scores <- sapply(1:nrow(schools), function(i){
       scores <- rnorm(schools$size[i], schools$quality[i], 30)
       scores
})
schools <- schools %>% mutate(score = sapply(scores, mean))
```

---

## Q1

1.0/1.0 point (graded)

What are the top schools based on the average score? Show just the ID, size, and the average score.

Report the ID of the top school and average score of the 10th school.

What is the ID of the top school?

Note that the school IDs are given in the form "PS x" - where x is a number. Report the **number** only.

567 ✔ **Answer:** 567

567

What is the average score of the 10th school?

87.95731 ✔ **Answer:** 87.95

87.95731

**Explanation**

The ID, size, and average score of the top schools can be identified using this code:

`schools %>% top_n(10, score) %>% arrange(desc(score)) %>% select(id, size, score)`.

Submit You have used 1 of 10 attempts

---

ℹ Answers are displayed within the problem

---

## Q2

1.0/1.0 point (graded)

Compare the median school size to the median school size of the top 10 schools based on the score.

What is the median school size overall?

261 ✔ **Answer:** 261

261

What is the median school size of the of the top 10 schools based on the score?

185.5 ✔ **Answer:** 185.5

185.5

**Explanation**
The median school sizes can be compared using the following code:

```
median(schools$size)
schools %>% top_n(10, score) %>% .$size %>% median()
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

## Q3

1.0/1.0 point (graded)
According to this analysis, it appears that small schools produce better test scores than large schools. Four out of the top 10 schools have 100 or fewer students. But how can this be? We constructed the simulation so that quality and size were independent. Repeat the exercise for the worst 10 schools.

What is the median school size of the bottom 10 schools based on the score?

219 ✔ **Answer:** 219

219

**Explanation**
The median school size for the bottom 10 schools can be found using the following code:

```
median(schools$size)
schools %>% top_n(-10, score) %>% .$size %>% median()
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

## Q4

1/1 point (graded)

From this analysis, we see that the worst schools are also small. Plot the average score versus school size to see what's going on. Highlight the top 10 schools based on the **true** quality.

What do you observe?

- There is no difference in the standard error of the score based on school size; there must be an error in how we generated our data.
- **The standard error of the score has larger variability when the school is smaller, which is why both the best and the worst schools are more likely to be small.** ✔
- The standard error of the score has smaller variability when the school is smaller, which is why both the best and the worst schools are more likely to be small.
- The standard error of the score has larger variability when the school is very small or very large, which is why both the best and the worst schools are more likely to be small.
- The standard error of the score has smaller variability when the school is very small or very large, which is why both the best and the worst schools are more likely to be small.

**Explanation**

You can generate the plot using the following code:

```
schools %>% ggplot(aes(size, score)) +
        geom_point(alpha = 0.5) +
        geom_point(data = filter(schools, rank<=10), col = 2)
```

We can see that the standard error of the score has larger variability when the school is smaller. This is a basic statistical reality we learned in PH125.3x: Data Science: Probability and PH125.4x: Data Science: Inference and Modeling courses! Note also that several of the top 10 schools based on **true** quality are also in the top 10 schools based on the exam score:
`schools %>% top_n(10, score) %>% arrange(desc(score))`.

Submit

You have used 1 of 2 attempts

Answers are displayed within the problem

## Q5

1.0/1.0 point (graded)

Let's use regularization to pick the best schools. Remember regularization **shrinks** deviations from the average towards 0. To apply regularization here, we first need to define the overall average for all schools, using the following code:

```
overall <- mean(sapply(scores, mean))
```

Then, we need to define, for each school, how it deviates from that average.

Write code that estimates the score above the average for each school but dividing by $n + \alpha$ instead of $n$, with $n$ the school size and $\alpha$ a regularization parameter. Try $\alpha = 25$.

What is the ID of the top school with regularization?

Note that the school IDs are given in the form "PS x" - where x is a number. Report the **number** only.

191 ✔ **Answer:** 191

191

What is the regularized score of the 10th school?

87.15 ✔ **Answer:** 87.15

87.15

**Explanation**

The regularization and reporting of scores can be done using the following code:

```
alpha <- 25
score_reg <- sapply(scores, function(x)  overall + sum(x-overall)/(length(x)+alpha))
schools %>% mutate(score_reg = score_reg) %>%
        top_n(10, score_reg) %>% arrange(desc(score_reg))
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

## Q6

1/1 point (graded)
Notice that this improves things a bit. The number of small schools that are not highly ranked is now lower. Is there a better $\alpha$? Using values of $\alpha$ from 10 to 250, find the $\alpha$ that minimizes the RMSE.

$$\text{RMSE} = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(\text{quality} - \text{estimate})^2}$$

What value of $\alpha$ gives the minimum RMSE?

135 ✔ **Answer:** 135

135

**Explanation**
The value of $\alpha$ that minimizes the MSE can be calculated using the following code:

```
alphas <- seq(10,250)
rmse <- sapply(alphas, function(alpha){
        score_reg <- sapply(scores, function(x) overall+sum(x-
overall)/(length(x)+alpha))
        sqrt(mean((score_reg - schools$quality)^2))
})
plot(alphas, rmse)
alphas[which.min(rmse)]
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

## Q7

1.0/1.0 point (graded)
Rank the schools based on the average obtained with the best $\alpha$. Note that no small school is incorrectly included.

What is the ID of the top school now?
Note that the school IDs are given in the form "PS x" - where x is a number. Report the **number** only.

191 ✔ **Answer:** 191

191

What is the regularized average score of the 10th school now?

85.48132 ✔ **Answer:** 85.4

85.48132

**Explanation**
The new ranking can be done using the following code:

```
alpha <- alphas[which.min(rmse)]
score_reg <- sapply(scores, function(x)
        overall+sum(x-overall)/(length(x)+alpha))
schools %>% mutate(score_reg = score_reg) %>%
        top_n(10, score_reg) %>% arrange(desc(score_reg))
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

## Q8

1/1 point (graded)
A common mistake made when using regularization is shrinking values towards 0 that are not centered around 0. For example, if we don't subtract the overall average before shrinking, we actually obtain a very similar result. Confirm this by re-running the code from the exercise in Q6 but without removing the overall mean.

What value of $\alpha$ gives the minimum RMSE here?

10 ✔ **Answer:** 10

10

**Explanation**
The code here is nearly the same as in Q6, but we don't subtract the overall mean. The value of $\alpha$ that minimizes the RMSE can be calculated using the following code:

```
alphas <- seq(10,250)
rmse <- sapply(alphas, function(alpha){
        score_reg <- sapply(scores, function(x) sum(x)/(length(x)+alpha))
        sqrt(mean((score_reg - schools$quality)^2))
})
plot(alphas, rmse)
alphas[which.min(rmse)]
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

Ask your questions or make your comments about Regularization here! **Remember, one of the best ways to reinforce your own learning is by explaining something to someone else, so we encourage you to answer each other's questions (without giving away the answers, of course).**

Some reminders:

- Search the discussion board before posting to see if someone else has asked the same thing before asking a new question.
- Please be specific in the title and body of your post regarding which question you're asking about to facilitate answering your question.
- Posting snippets of code is okay, but posting full code solutions is not.
- If you do post snippets of code, please format it as code for readability. If you're not sure how to do this, there are instructions in a pinned post in the "general" discussion forum.

## Discussion: Regularization

Show Discussion

**Topic:** Section 6: Model fitting and recommendation systems / 6.3.1: Regularization