

# Quiz 3

## Data Science for Studying Language & the Mind

 **Estimated time: 30 minutes**

You may need more time if programming is completely new to you, or less if you have some experience already.

### Instructions

- The quiz is closed book/note/computer/phone
- If you need to use the restroom, leave your exam and phone with the TA
- You have 60 minutes to complete the quiz. If you finish early, you may turn in your quiz and leave early

**Name:** \_\_\_\_\_

**PennKey:** \_\_\_\_\_

**Lab section TA:** \_\_\_\_\_

### Score by topic area

Model Fitting	
Model Fitting in R	
Model Accuracy	
Model Accuracy in R	
Total	

## The data

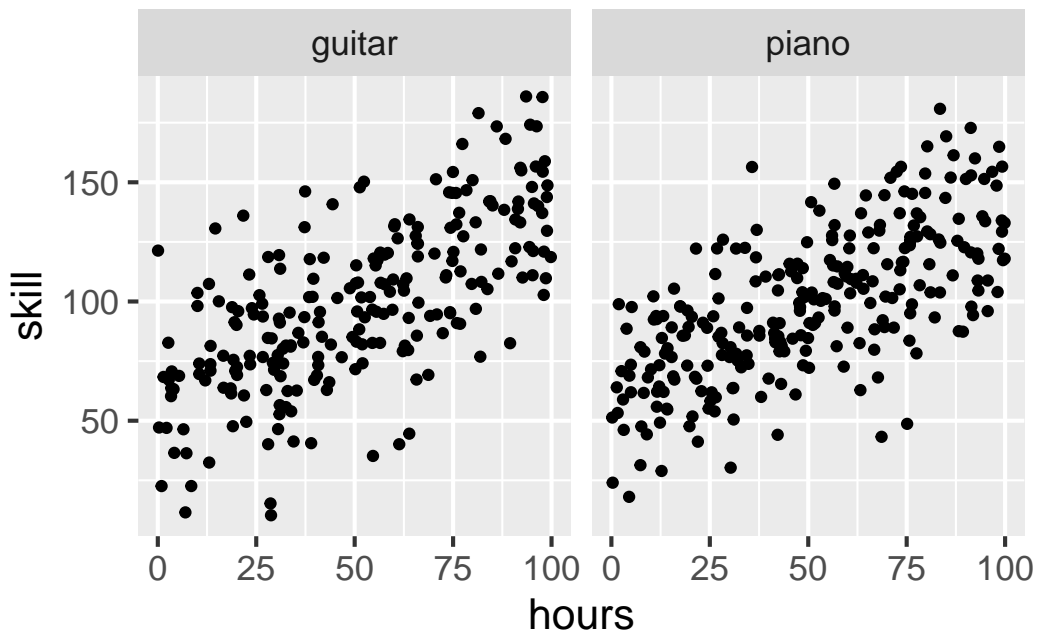
Suppose we want to study the effect hours practicing an instrument has on your ultimate skill level with the instrument. We study 500 participants who are learning to play either piano or guitar. Below we explore these data in a few ways.

```
glimpse(data)
```

Rows: 500

Columns: 4

```
$ hours          <dbl> 11.3703411, 62.2299405, 60.9274733, 62.3379442, 86.~  
$ instrument_recoded <dbl> 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, ~  
$ skill          <dbl> 93.91577, 79.16551, 126.48513, 107.13986, 173.43843~  
$ instrument      <chr> "piano", "guitar", "guitar", "guitar", "guitar", "p~
```



```
data %>%  
  group_by(instrument) %>%  
  summarise(  
    n = n(),  
    mean_skill = mean(skill), sd_skill = sd(skill),  
    mean_hours = mean(hours), sd_hours = sd(hours))
```

```
# A tibble: 2 x 6
  instrument      n mean_skill sd_skill mean_hours sd_hours
  <chr>      <int>    <dbl>    <dbl>    <dbl>    <dbl>
1 guitar     233     99.2     34.8     51.0     28.4
2 piano      267     99.1     30.9     50.1     28.3
```

## 1 Model Fitting

Suppose we fit a model represented by the following equation, where  $x_1$  is the number of hours spent practicing,  $x_2$  is the instrument, and  $y$  is the skill achieved:

$$y = b_0 + b_1x_1 + b_2x_2$$

- (a) Which of the following would work to estimate the free parameters of this model? Choose one.
- ☐ only gradient descent
  - ☐ only ordinary least squares
  - ☐ both gradient descent and ordinary least squares
- (b) True or false, ordinary least squares finds the best fitting free parameters by solving a system of linear equations.
- ☐ True
  - ☐ False
- (c) True or false, when performing gradient descent on a **nonlinear** model, we might arrive at a local minimum and miss the global one.
- ☐ True
  - ☐ False
- (d) True or False, given the model above, gradient descent and ordinary least squares would both converge on approximately the same parameter estimates.
- ☐ True
  - ☐ False

## 2 Model Fitting in R

Questions in section 2 refer to the code below.

```
# fit model with lm  
model
```

Call:

```
lm(formula = skill ~ hours + instrument_recoded, data = data)
```

Coefficients:

(Intercept)	hours	instrument_recoded
58.9493	0.7885	0.6834

```
#fit model with optimg  
optim(data = data, par = c(1,1,1), fn=SSE, method = "STGD")
```

\$par

```
[1] 58.9488377  0.7884514  0.6900582
```

\$value

```
[1] 286497.6
```

\$counts

```
[1] 26
```

\$convergence

```
[1] 0
```

(a) What parameters did the gradient descent algorithm try first?

(b) Which of the following could be the model specification in R? Choose all the apply.

- ☐ skill ~ hours + instrument\_recoded
- ☐ skill ~ hours \* instrument\_recoded
- ☐ skill ~ 1 + hours + instrument\_recoded

- ☐  $y \sim x$   
☐  $y \sim 1 + x$

(c) Given the equation given below, what are the best fitting free parameter for  $b_1$  and  $b_2$ ?

$y = b_0 + b_1x_1 + b_2x_2$  where  $x_1$  is the number of hours spent practicing,  $x_2$  is the instrument, and  $y$  is the skill achieved.

\answerbox

(d) Which of the following could be the value of the sum of squared errors when the parameters  $b_0$ ,  $b_1$ , and  $b_2$  are set to 0?

- ☐ exactly 286497.6  
☐ a value higher than 286497.6  
☐ a value lower than 286497.6  
☐ approximately 26  
☐ approximately 0

### 3 Model Accuracy

Questions in section 3 refer to the following code below.

```
summary(model)
```

Call:

```
lm(formula = skill ~ hours + instrument_recoded, data = data)
```

Residuals:

Min	1Q	Median	3Q	Max
-71.284	-15.388	-0.196	16.230	68.624

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	58.94933	2.49405	23.636	<2e-16 ***
hours	0.78845	0.03795	20.778	<2e-16 ***
instrument_recoded	0.68342	2.15273	0.317	0.751

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.01 on 497 degrees of freedom

Multiple R-squared: 0.4649, Adjusted R-squared: 0.4627  
F-statistic: 215.9 on 2 and 497 DF, p-value: < 2.2e-16

(a) What is the  $R^2$  for this model?

(b) Which of the following is true about this  $R^2$  value?

- ☐ tends to overestimate  $R^2$  on the population
- ☐ tends to underestimate  $R^2$  on the population
- ☐ tends to overestimate  $R^2$  on the sample
- ☐ tends to underestimate  $R^2$  on the sample

(c) Explain why an overfit model would perform well on the sample, but poorly on predicting new values.

(d) True or false, we can use cross-validation **or** bootstrapping to estimate  $R^2$  on the population?

- ☐ True
- ☐ False

## 4 Model Accuracy in R

Questions in section 4 refer to the following code:

```
# we divide the data
set.seed(2)
splits <- vfold_cv(data, v = 20)

# model specification
model_spec <-
  linear_reg() %>%
  set_engine(engine = "lm")

# add a workflow
```

```

our_workflow <-
  workflow() %>%
  add_model(model_spec) %>%
  add_formula(skill ~ hours + instrument_recoded)

# fit models
fitted_models <-
  fit_resamples(
    object = our_workflow,
    resamples = splits
  )

fitted_models %>%
  collect_metrics()

```

```

# A tibble: 2 x 6
  .metric .estimator   mean     n std_err .config
  <chr>   <chr>       <dbl> <int>   <dbl> <chr>
1 rmse    standard    23.8     20  0.762 Preprocessor1_Model1
2 rsq     standard     0.468     20  0.0267 Preprocessor1_Model1

```

(a) In the output above, what is our estimate for  $R^2$  on the population?

(b) In the code above, what method did we use to estimate  $R^2$  on the population? Choose one.

- ☐ k-fold cross-validation
- ☐ leave one out cross-validation
- ☐ bootstrapping
- ☐ workflow()

(c) In the code above, how many models did we fit when calling `fit_resamples()`?

- ☐ 2
- ☐ 20
- ☐ 10
- ☐ 100

(d) True or false, if we estimated the  $R^2$  for the population with another approach, the value would be exactly the same.

☐ True

☐ False