(Practice) Exam 2

Data Science for Studying Language & the Mind

Instructions

You have **1 hours and 30 minutes** to complete the exam.

- The exam is closed book/note/computer/phone except for the provided reference sheets
- If you need to use the restroom, leave your exam and phone with the TAs
- If you finish early, you may turn in your exam and leave early

Preliminary questions

Please complete these questions *before* the exam begins.

(a) **(1 point)** What is your full name?

(b) **(1 point)** What is your penn ID number?

(c) **(1 point)** What is your lab section TA's name?

(d) **(1 point)** Who is sitting to your left?

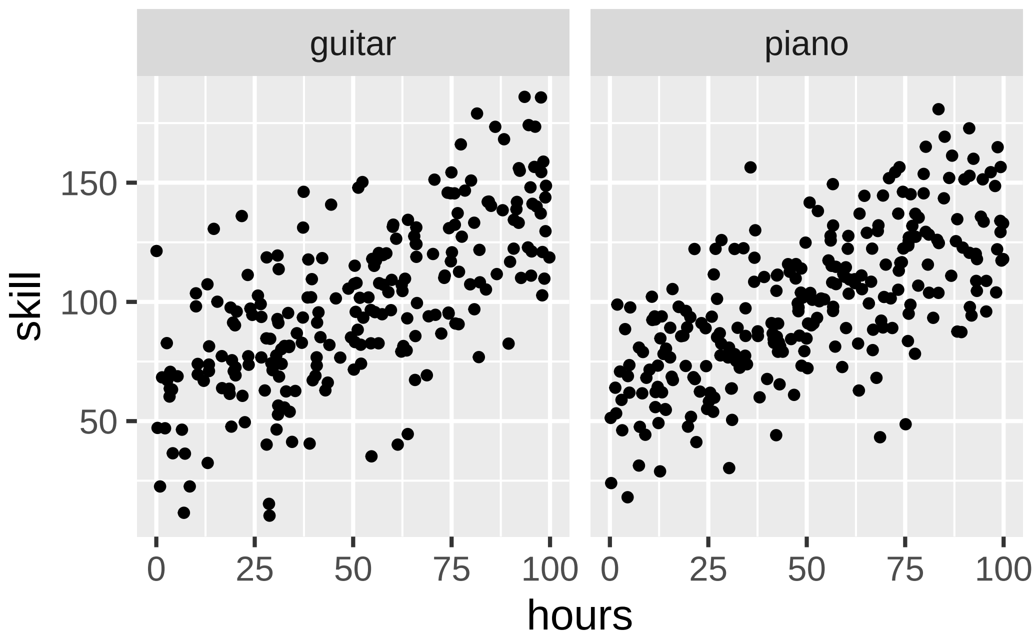
(e) **(1 point)** Who is sitting to your right?

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.

```r
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, 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~
```




```r
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 acheived:

$y = b_0 + b_1 x_1 + b_2 x_2$

(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

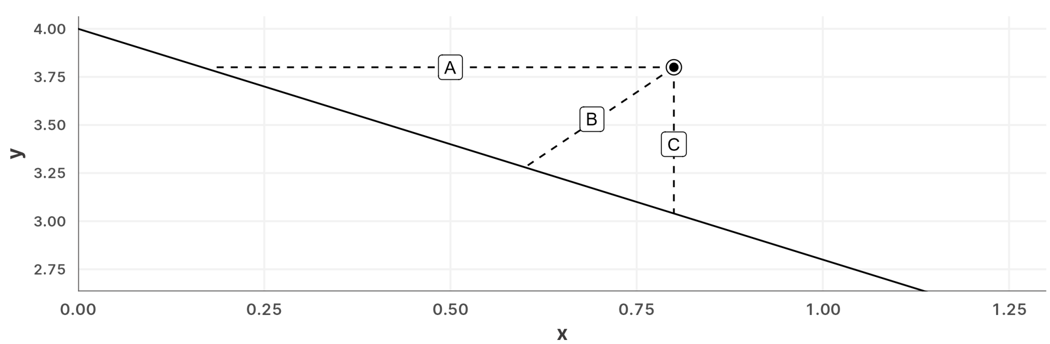
(b) True or false, when performing gradient descent on a **nonlinear** model, we might arrive at a local minimum and miss the global one.

- ☒ True
- ☐ False

(c) True or False, given the model above, gradient descent and ordinary least squares would both converge on approximately the same parameter estimates.

- ☒ True
- ☐ False

(d) The following plots a linear model of the formula `y ~ 1 + x` and one data point. Which dashed line represents the model's **residual** for this point? Circle one.



Line **C**

2 Model Fitting in R

Questions in section 2 refer to the code below.

```
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
optimg(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) 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`

(b) In the code, `SSE()` is a function we have defined to calculate the sum of squared errors. Which of the following correctly describes the steps of calculating SSE? Choose one.

- ☒ 1) calculate the residuals, 2) square each of the residuals, 3) add them up
- ☐ 1) calculate the residuals, 2) add them up, 3) square the sum of residuals
- ☐ 1) calculate the residuals, 2) calculate their standard deviation, 3) square it
- ☐ 1) calculate the residuals, 2) calculate their mean, 3) square it

(c) Using the estimated parameters from `lm()`, fill in the blanks to calculate the model's predicted value of `skill` for a participant who played the **piano** for **20 hours**. You may round to the first decimal place.

```
skill = 58.9 + ( 0.8 * 20 ) + ( 0.7 * 1 )
```

(d) Which of the following is the most likely value of the sum of squared errors when the parameters b_0, b_1, and b_2 are all set to 0? Choose one.

- ☐ exactly 0
- ☐ exactly 286497.6
- ☒ a value higher than 286497.6
- ☐ a value lower than 286497.6

3 Model Accuracy

Questions in section 3 refer to the following `summary()` of the same model from section 2:

(a) Which of the following is a correct interpretation of the model's R^2 value? Choose one.

- ☐ The model has a 46.49% chance of explaining the true pattern in the data.
- ☒ The model explains 46.49% of the variance found in the data.
- ☐ The sample shows 46.49% of the variance found in the population.

(b) Which of the following is true about the model's R^2? Choose all that apply.

- ☒ 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) Which one of the following is true about R^2? Use the below formula as a guide and choose one.

$R^2 = 1 - \frac{unexplained\ variance}{total\ variance}$

- ☐ The unexplained variance refers to the fact that linear model haveh low accuracy.
- ☐ The total variance is about the overall variability of the data in the population.
- ☒ R^2 of 0 means that the model predicts the mean of the data but nothing else.
- ☐ R^2 of 1 means that the model will be perfect at predicting new data points.

(d) Which of the following is a correct statement about estimating R^2 for the *population*? Choose all that apply.

- ☐ We can use OLS
- ☒ We can use bootstrapping
- ☒ We can use cross-validation
- ☐ We must go out and collect more samples from the population

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 secification
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 the R^2 estimate for the population?

- ☐ 23.8
- ☒ 0.468
- ☐ 0.468 + 0.0267

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

- ☒ k-fold cross-valiation
- ☐ leave one out cross-valiation
- ☐ boostrapping

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

- ☐ 10
- ☒ 20
- ☐ 100

(d) You are no longer doing a valid cross-validation if you change (choose all that apply):

- ☐ How many iterations you want to do.
- ☐ How much data you want to use for each part of training vs. testing.
- ☒ Whether models are fitted to the entire sample instead of a part of the sample
- ☒ Whether models are tested on the training data instead of the testing data

5 Model reliability

(a) True or false, as we collect more data, the confidence interval around our parameter estimates gets bigger (wider).

- ☐ True
- ☒ False

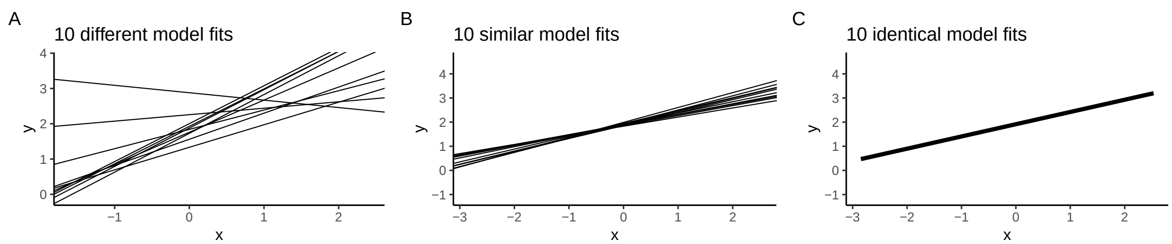
(b) Model reliability asks how certain we can be about our parameter estimates. Why is there uncertainty around our parameter estimates?

Due to sampling error! Our goal is to find the parameters that would best describe the population, but we only have a small sample. If we took a different random sample, we'd get different parameter estimates.

(c) True or false, a model with low reliability also has low accuracy.

☐ True
☒ False

(d) Suppose we conduct an experiment by drawing a random sample from the population. We fit a linear model to these data. Then we repeat our experiment 10 times, fitting the same model each time. Which figure could show the fitted models for the 10 experiments? Choose all that apply.



☒ Figure A
☐ Figure B
☐ Figure C

6 Nonlinear models

(a) Circle the figure below that plots the model represented by the equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$

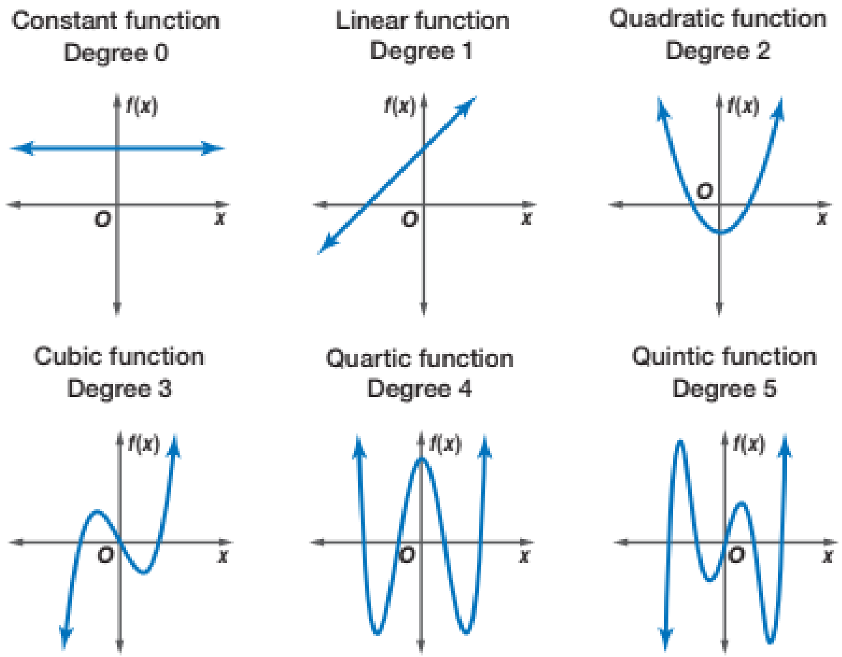


Figure **Quadratic function degree 2**

(b) Which of the model specifications expresses a cubic polynomial model in R?

- ☐ `y ~ poly(x, 1)`
- ☐ `y ~ poly(x, 2)`
- ☒ `y ~ poly(x, 3)`
- ☐ `y ~ poly(x, 4)`

(c) True or false, we can use `lm()` to fit a quadratic polynomial.

- ☒ `True`
- ☐ `False`

(d) Which of the following aspects of model building apply to nonlinear models? Choose all that apply.

- ☒ model specification
- ☒ model fitting
- ☒ model accuracy

☒ model reliability

7 Classification

(a) True or false, logistic regression is a linear classificaiton model.

☒ True
☐ False

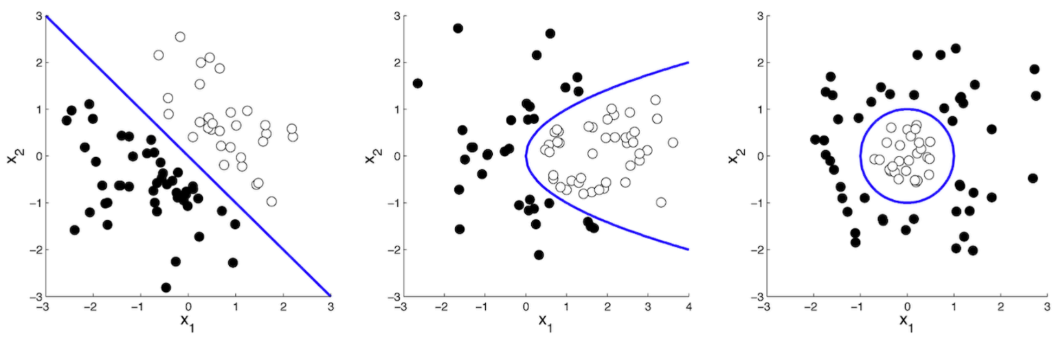
(b) What is the difference between regression and classification?

If y is continuous, we call the problem “regression”; if y is discrete/categorical we call is classification

(c) What accuracy metric(s) have we been applying to classification models? Choose all that apply.

☒ Percent correct
☐ R^2
☐ RMSE
☐ Sum of squared error

(d) True or false, each of the following figures can be modeled with a linear classifier.



☒ True
☐ False

8 Classification in R

(a) Which of the following can be used to fit a logistic regression model in R? Choose all that apply.

☐ `lm()`
☒ `glm()`
☐ `poly()`
☐ `log()`

(b) True or false, the link function in a generalized linear model **must be** the logistic function.

☒ True
☐ False

(c) Which of the following fits a logistic regression model in R? Choose all that apply.

```
# code A
glm(y ~ x, data = data, family = "binomial")

# code B
data %>%
    specify(y ~ x) %>%
    fit()

# code C
linear_reg %>%
    set_engine("lm") %>%
    fit(y ~ x, data = data)
```

☒ Code A
☒ Code B
☐ Code C

(d) What 3 elements do all generalized linear models have?

(1) a paticular distribution for modeling the response variable; (2) a linear model; (3) a link function