Exam 2

Data Science for Studying Language & the Mind

Instructions

The exam is worth **111 points**. You have **1 hour 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

(5 points) 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?

Please do not turn the page until the exam begins

1. (22 points) True or false

(a) **(2 points)** Regression is a type of nonlinear classifier.

- ☐ True
- ☐ False

(b) **(2 points)** Model specification involves defining the functional form of the model.

- ☐ True
- ☐ False

(c) **(2 points)** The equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ expresses y as a weighted sum of inputs.

- ☐ True
- ☐ False

(d) **(2 points)** Regression and classification are both supervised learning models.

- ☐ True
- ☐ False

(e) **(2 points)** In gradient descent, we search through all possible parameters in the parameter space.

- ☐ True
- ☐ False

(f) **(2 points)** Gradient descent is an example of an iterative optimization algorithm.

- ☐ True
- ☐ False

(g) **(2 points)** The largest possible R^2 value is 1 (or 100% if expressed as a percentage).

- ☐ True
- ☐ False

(h) **(2 points)** An overfit model performs poorly on the sample, but well on predicting new data.

- ☐ True
- ☐ False

(i) **(2 points)** Model reliability and model accuracy are the same thing by a different name.

- ☐ True
- ☐ False

(j) **(2 points)** Our parameter estimates become more stable as we increase our sample size.

- ☐ True
- ☐ False

(k) **(2 points)** Generalized linear models can be used for classification problems.

- ☐ True
- ☐ False

(l) **(2 points)** In matrix notation, $\mathbf{X}$ is the matrix of explanatory variables.

- ☐ True
- ☐ False

(m) **(2 points)** `optimg` and `lm` return identical parameter estimates.

- ☐ True
- ☐ False

(n) **(2 points)** `infer` and `lm` return identical parameter estimates.

- ☐ True
- ☐ False

2. (14 points) Model specification

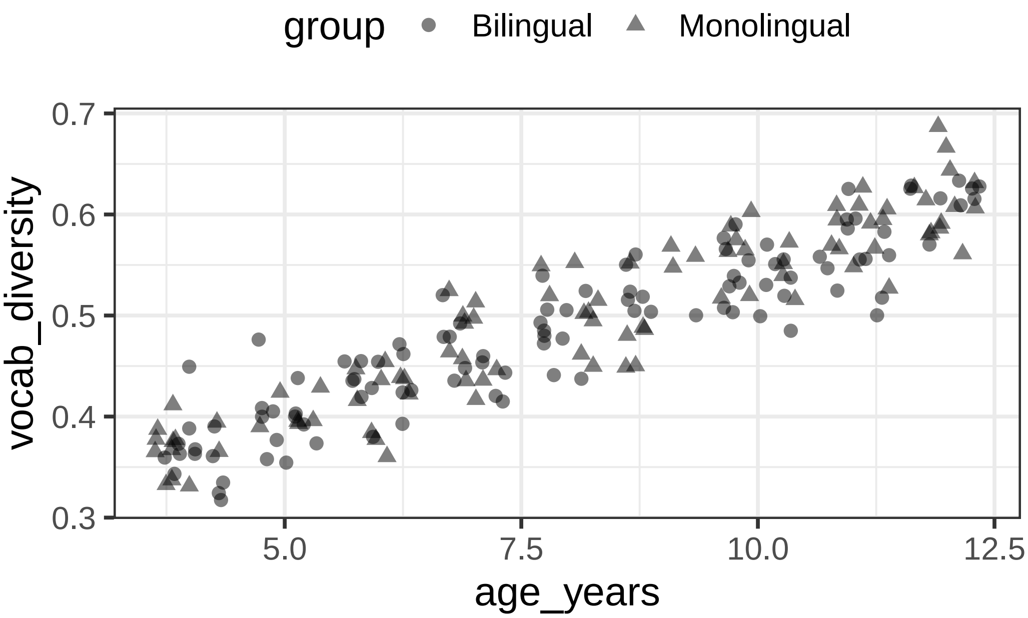
Suppose you are working with a fictional dataset called `narr_prod`, which contains language measures from children who completed a narrative retelling task. Each child viewed the wordless picture book *Good Dog, Carl* and was asked to tell the story in their own words. Researchers transcribed each narrative and coded several features reflecting the child's language. The dataset includes:

- `age_years`: the child's age in years (4–12).
- `group`: the child's language background ("Monolingual" or "Bilingual").
- `num_clauses`: total number of clauses in the child's narrative.
- `vocab_diversity`: a type–token ratio capturing how varied the child's vocabulary was.
- `coherence_rating`: a 1–5 rating of how coherent, organized, and story-like the retelling was.

The first 6 rows of these data and an exploratory plot are printed below for your reference.

```
# A tibble: 6 x 6
  child_id age_years group       num_clauses vocab_divers
     <dbl>     <dbl> <chr>             <dbl>             <d
1        1        10 Monolingual          36             0.
```

```
2            2            7 Monolingual           27              0.4
3            3           11 Bilingual             32              0.6
4            4            8 Monolingual           21              0.5
5            5           10 Monolingual           36              0.5
6            6            6 Bilingual             19              0.4
```



Suppose we specify the following model with `lm`:

```
model <- lm(vocab_diversity ~ age_years, data = narr_prod)
```

(a) **(3 points)** Which of the following is the model's specification as a mathematical expression:

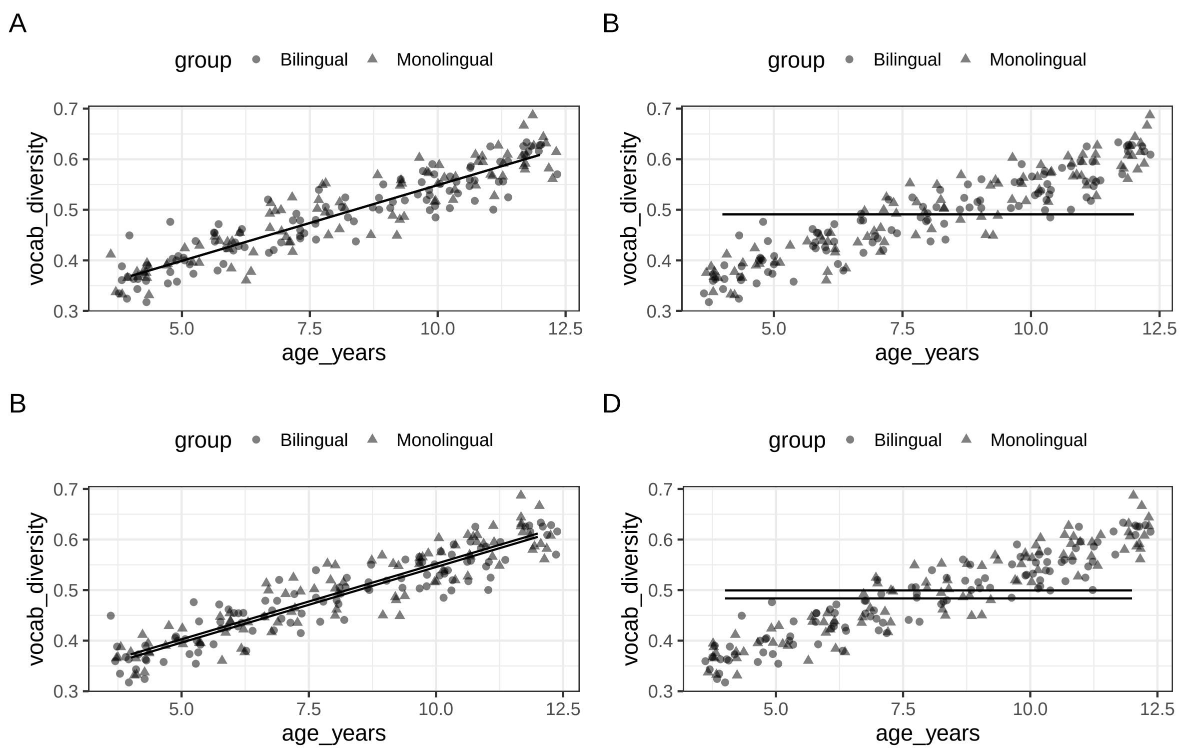
☐ $vocab\_diversity = w_0 + w_1 age\_years$

- ☐ $vocab\_diversity = w_1 age\_years$
- ☐ $age\_years = w_0 + w_1 vocab\_diversity$
- ☐ $age\_years = w_1 vocab\_diversity$

(b) **(3 points)** For each of the following, circle the option that best describes the type of model we fit.

(i) **(1 point)** Supervised or unsupervised
(ii) **(1 point)** Regression or classification
(iii) **(1 point)** Linear or linearlizable nonlinear

(c) **(3 points)** Each of the figures below show a model's predictions for these data plotted with black lines. Circle the figure that is most likely to be the plot of the model spcified to `lm`? Choose one.



(d) **(3 points)** Suppose we also fit the model with `infer`, which returns the parameter estimates below. Which of the following could be the predicted `vocab_diversity` for a 5 year old child?

```
# A tibble: 2 x 2
  term      estimate
  <chr>        <dbl>
1 intercept   0.249
2 age_years   0.0299
```

- ☐ 0.15
- ☐ 0.0299
- ☐ 0.40
- ☐ 0.249
- ☐ Not enough information to determine this

You may show your work here, if you wish:

(e) **(2 points)** True or false, the model's prediction depends on whether the child is bilingual or monolingual.

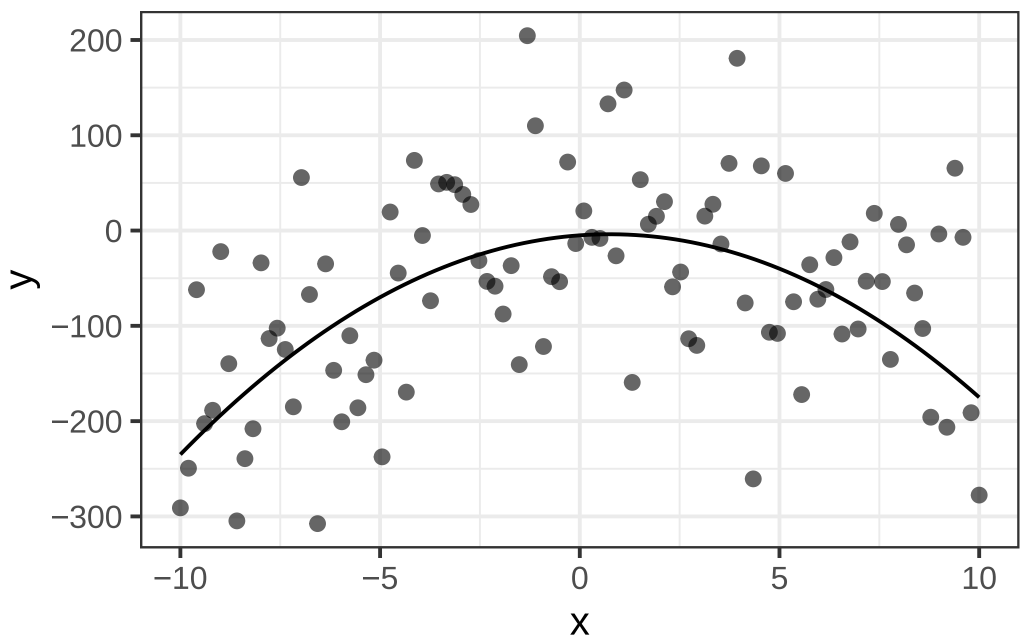
- ☐ True

☐ False

3. (12 points) Applied model specification

Suppose we encounter the following dataset, glimpsed, plotted and fit here.

```
Rows: 100
Columns: 2
$ x <dbl> -10.000000, -9.797980, -9.595960, -9.393939, -9
$ y <dbl> -291.04756, -249.41250, -62.08193, -202.62317, 
```




```
Call:
lm(formula = y ~ poly(x, 2), data = data)
```

```
Coefficients:
(Intercept)  poly(x, 2)1  poly(x, 2)2
     -63.97       247.43      -514.32
```

(a) **(2 points)** What type of polynomial is included in the model specification?

- ☐ Constant
- ☐ Linear
- ☐ Quadratic
- ☐ Cubic
- ☐ Quartic

(b) **(3 points)** Which of the following is the parameter estimate on the quadratic term?

- ☐ -63.97
- ☐ 247.43
- ☐ -514.32
- ☐ Not enough information to determine this

(c) **(2 points)** In class we learned about two ways to linearlize a nonlinear model. Which option best describes what we have done here?

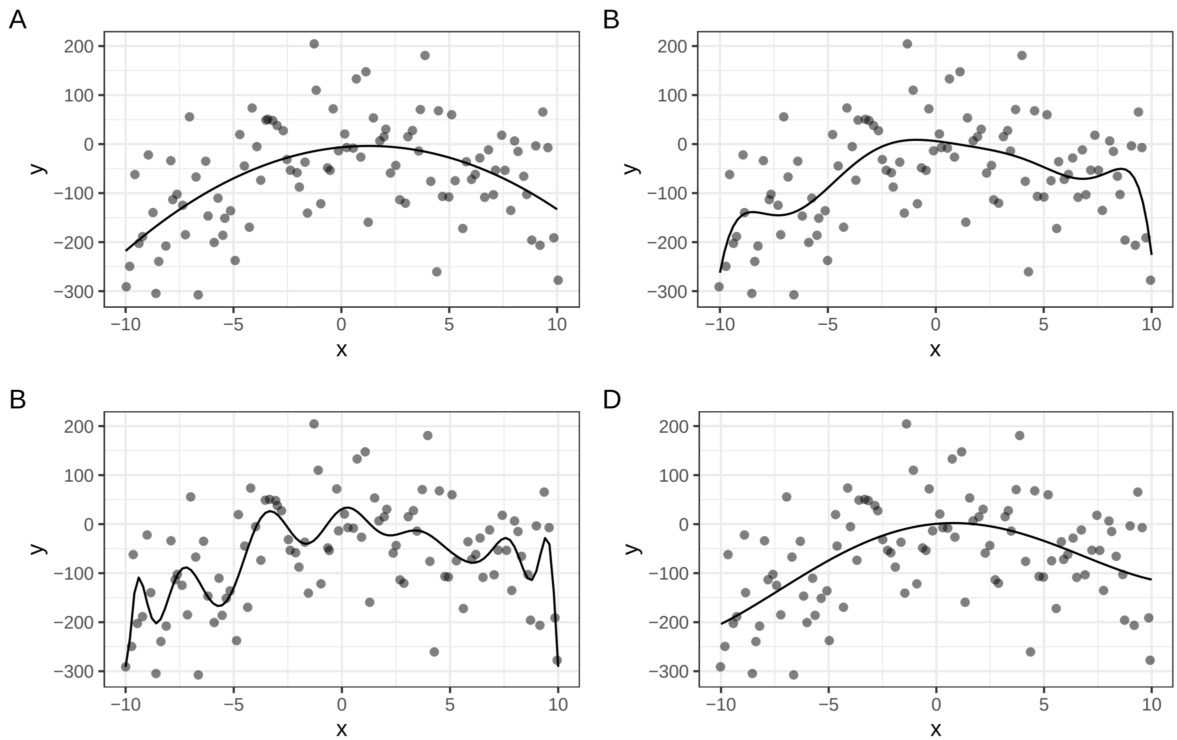
- ☐ Expanding the input space by adding new terms
- ☐ Transforming an existing term

(d) **(2 points)** Given the plot of the fitted model, what does the model predict for the value of y when $x = 0$?

- ☐ Nearly 0
- ☐ Less than -200

☐ Greater than 5
☐ Between 1 and 2

(e) **(3 points)** Below are four fitted polynomial models of these data. Which model uses the highest-degree polynomial?



4. (16 points) Model fitting

Section 4 refers to the **narr_prod** tibble from section 2. We have returned the first 6 rows of the tibble here for your reference.

```
# A tibble: 6 x 6
  child_id age_years group       num_clauses vocab_diversi
     <dbl>     <dbl> <chr>             <dbl>            <dl
1        1        10 Monolingual          36            0.5
2        2         7 Monolingual          27            0.4
3        3        11 Bilingual            32            0.6
4        4         8 Monolingual          21            0.5
5        5        10 Monolingual          36            0.5
6        6         6 Bilingual            19            0.4
```

Suppose we estimate the free parameters with `optimg` and `lm`, which return the following results:

```
optimg(data = narr_prod, par = c(0,0), fn=SSE, method = "S
```

```
$par
[1] 0.24926974 0.02994827

$value
```

```
[1] 0.1962883

$counts
[1] 6

$convergence
[1] 0
```

```r
lm(vocab_diversity ~ 1 +age_years, data = narr_prod)
```

```
Call:
lm(formula = vocab_diversity ~ 1 + age_years, data = narr_

Coefficients:
(Intercept)    age_years
    0.24927      0.02995
```

(a) **(2 points)** Which set of values did we use to initialize the search in `optimg`? Choose one.

- ☐ $0, 0$
- ☐ $0.24926974, 0.02994827$
- ☐ $6, 0$
- ☐ 0.1962883
- ☐ A random set of values

(b) **(2 points)** What is the cost function used by `optimg`? Choose one.

- ☐ SSE
- ☐ STGD
- ☐ Gradient descent
- ☐ R^2
- ☐ Not enough information to determine this

(c) **(2 points)** How many steps did our iterative optimization algorithm take? Choose one.

- ☐ 0
- ☐ 6
- ☐ 10
- ☐ 100
- ☐ Not enough information to determine this.

(d) **(2 points)** What was the sum of squared error of the optimal parameters according to `optimg`? Choose one.

- ☐ 0
- ☐ 6
- ☐ 0.1962883
- ☐ 0.24926974 and 0.2994827
- ☐ Not enough information to determine this.

(e) **(2 points)** Which approach does `lm` use to estimate the free parameters? Choose one.

- ☐ Ordinary least-squares solution
- ☐ Iterative optimization
- ☐ Cross validation
- ☐ Classification

(f) **(6 points)** Given the model specified in the code to `lm`, fill in the missing values for the first 6 rows of the input matrix $\mathbf{X}$ and the output vector y. The first six rows of the dataframe are returned to assist you.

```
# A tibble: 6 x 4
  child_id age_years vocab_diversity num_clauses
     <dbl>     <dbl>           <dbl>       <dbl>
1        1        10           0.517          36
2        2         7           0.435          27
3        3        11           0.625          32
4        4         8           0.502          21
5        5        10           0.540          36
6        6         6           0.455          19
```

$$\begin{bmatrix} \underline{\hspace{2cm}} \\ \underline{\hspace{2cm}} \\ \underline{\hspace{2cm}} \\ \underline{\hspace{2cm}} \\ \underline{\hspace{2cm}} \\ \underline{\hspace{2cm}} \end{bmatrix} = \begin{bmatrix} \underline{\hspace{2cm}} & 10 \\ \underline{\hspace{2cm}} & 7 \\ \underline{\hspace{2cm}} & 11 \\ \underline{\hspace{2cm}} & 8 \\ \underline{\hspace{2cm}} & 10 \\ \underline{\hspace{2cm}} & 6 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \end{bmatrix}$$

5. (11 points) Model accuracy

Suppose we want to determine how accurate our model is for the `narr_prod` dataset. Section 5 refers to the following code and output.

First we specify and fit our model with `lm` and return the model summary.

```
Call:
lm(formula = vocab_diversity ~ age_years, data = narr_pro

Residuals:
      Min        1Q    Median        3Q       Max
-0.078481 -0.022311  0.000626  0.020353  0.080245

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.2492683  0.0071907   34.66   <2e-16 ***
age_years   0.0299484  0.0008473   35.35   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 0.03149 on 198 degrees of freedo
Multiple R-squared:  0.8632,    Adjusted R-squared:  0.86
F-statistic:  1249 on 1 and 198 DF,  p-value: < 2.2e-16
```

Then we perform cross-validation and return the validation metrics with `collect_metrics()`

```
set.seed(2)
splits <- vfold_cv(narr_prod, v = 15)

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

our_workflow <-
  workflow() %>%
  add_model(model_spec) %>%
  add_formula(vocab_diversity ~ 1 + age_years)

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   0.0312    15 0.00170 pre0_mod0_post0
2 rsq     standard   0.854     15 0.0225  pre0_mod0_post0
```

(a) **(2 points)** What is the R^2 estimate for the population? Choose one.

- ☐ 0
- ☐ 0.0312
- ☐ 0.854
- ☐ 0.8632
- ☐ Not enough information to determine this

(b) **(2 points)** What kind of cross-validation did we perform? Choose one.

- ☐ k-fold
- ☐ boostrapping
- ☐ leave-one out
- ☐ Not enough information to determine this

(c) **(2 points)** How many splits of our data does our code generate?

- ☐ 1000
- ☐ 100
- ☐ 10
- ☐ 15
- ☐ Not enough information to determine this

(d) **(3 points)** Which of the following is the equation for R^2?

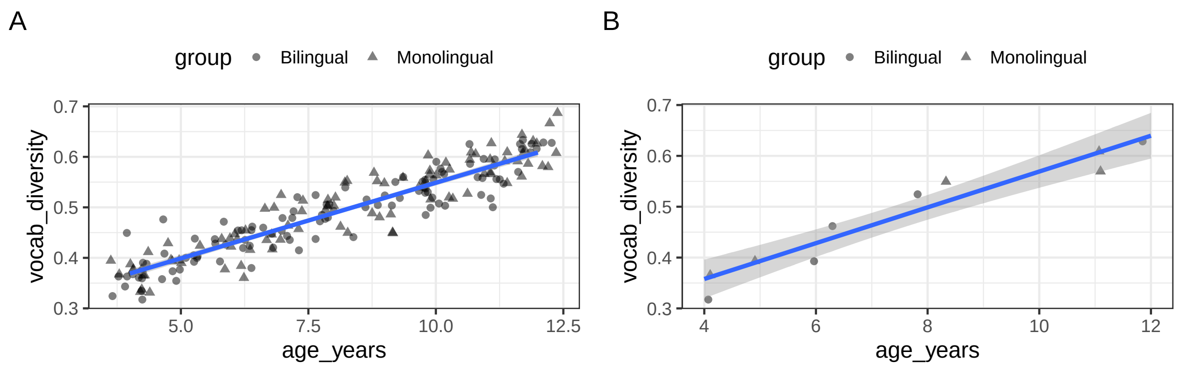
- ☐ $f(a) = \frac{1}{1+e^{-a}}$
- ☐ $\sum_{i=i}^{n}(d_i - m_i)^2$
- ☐ $\sum_{i=1}^{n} w_i x_i$
- ☐ $100 \times (1 - \frac{SSE_{model}}{SSE_{reference}})$

(e) **(2 points)** Suppose we change `add_formula(vocab_diversity ~ 1 + age_years)` to `add_formula(vocab_diversity ~ age_years)`. What will happen to our R^2 value?

- ☐ increases
- ☐ decreases
- ☐ stays the same
- ☐ becomes undefined

6. (12 points) Model reliability

Suppose we replicated our `narr_prod` narrative retelling study a year later with 10 additional children. We will call this `narr_prod_2`. The new and original datasets are visualized below, along with the model summarise.




```
Call:
lm(formula = vocab_diversity ~ age_years, data = narr_pro

Residuals:
      Min         1Q     Median         3Q        Max
-0.078481 -0.022311  0.000626  0.020353  0.080245
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.2492683  0.0071907   34.66   <2e-16 ***
age_years   0.0299484  0.0008473   35.35   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 0.03149 on 198 degrees of freedo
Multiple R-squared:  0.8632,    Adjusted R-squared:  0.86
F-statistic:  1249 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = vocab_diversity ~ age_years, data = narr_proc

Residuals:
      Min        1Q    Median        3Q       Max
-0.040428 -0.028900  0.002339  0.021058  0.050526

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.216917   0.029376   7.384 7.74e-05 ***
age_years   0.035226   0.003663   9.615 1.14e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 0.03287 on 8 degrees of freedom
Multiple R-squared:  0.9204,    Adjusted R-squared:  0.910
F-statistic: 92.46 on 1 and 8 DF,  p-value: 1.137e-05
```

(a) **(2 points)** Which model is more accurate? Choose one.

- ☐ The model fitted to the original data (A)
- ☐ The model fitted to the new data (B)
- ☐ Both models are equally accurate
- ☐ Not enough information to determine this

(b) **(2 points)** Which model is more reliable? Choose one.

☐ The model fitted to the original data (A)
☐ The model fitted to the new data (B)
☐ Both models are equally accurate
☐ Not enough information to determine this

(c) **(2 points)** What is the reliability on the `age_years` parameter estimate for the original data (A)?

☐ 0.0299484
☐ 0.0008473
☐ 35.35
☐ 0.8632
☐ 0.8625

(d) **(3 points)** Suppose we bootstrap a 68% confidence interval for our parameter estimates for the new dataset (B). What would happen if we changed the level of the confidence interval to 95%? Choose one.

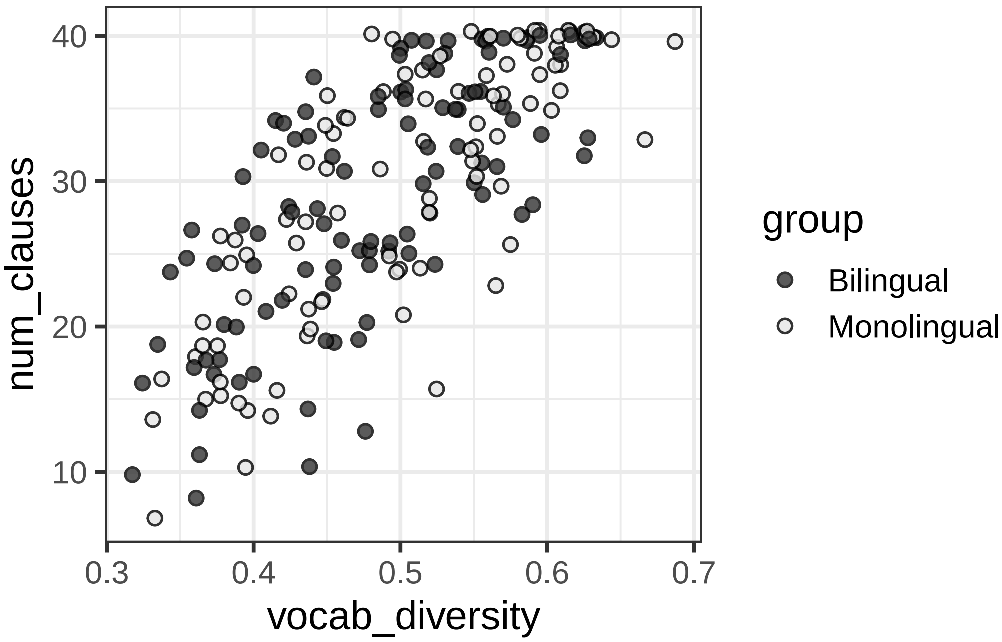
☐ It would get smaller (narrower)
☐ It would get bigger (wider)
☐ It would stay the same

(e) **(3 points)** What does 'model reliability' refer to?

- ☐ How well the model fits the training data
- ☐ How consistent the model's predictions or estimates are across different samples
- ☐ How large the R^2 value is
- ☐ How quickly the model runs

7. (13 points) Classification

Suppose we want to predict whether a kid in our dataset is bilingual or monolingual by their `vocab_diversity` in our dataset. Here is an expoloratory plot and the fitted model.




```
Call:  glm(formula = bilingual ~ vocab_diversity + num_cla
    data = narr_prod)

Coefficients:
    (Intercept)  vocab_diversity      num_clauses
        1.43340         -4.03941          0.02359
```

```
Degrees of Freedom: 199 Total (i.e. Null);  197 Residual
Null Deviance:      276.3
Residual Deviance: 273.7    AIC: 279.7
```

(a) **(3 points)** For each of the following, circle the option that best describes the type of model we fit.

   (i) **(1 point)** Supervised or unsupervised
   (ii) **(1 point)** Regression or classification
   (iii) **(1 point)** Linear or linearlizable nonlinear

(b) **(2 points)** How many free parameters is this model estimating?

- ☐ 1
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ Not enough information to determine this

(c) **(2 points)** Which of the following parsnip specifications could specify and fit this same model?

- ☐ `linear_reg() %>% set_engine("lm")`
- ☐ `logistic_reg() %>% set_engine("glm")`
- ☐ Both work

(d) **(2 points)** Which of the following expresses the link function for the `glm` we fit?

- ☐ $f(a) = \frac{1}{1+e^{-a}}$
- ☐ $\sum_{i=i}^{n}(d_i - m_i)^2$
- ☐ $y = \sum_{i=1}^{n} w_i x_i$

☐ $R^2 = 100 \times (1 - \frac{SSE_{model}}{SSE_{reference}})$

(e) **(2 points)** What do we call the type of classification we performed via our `glm`?

☐ linear regression
☐ logistic regression
☐ nearest-prototype regression
☐ support vector machine

Suppose we run cross validation on a few of our models and return the following outputs.

```
# A tibble: 3 x 6
  .metric     .estimator  mean     n std_err .config
  <chr>       <chr>      <dbl> <int>   <dbl> <chr>
1 accuracy    binary     0.499    15 0.0361  pre0_mod0_pos
2 brier_class binary     0.254    15 0.00560 pre0_mod0_pos
3 roc_auc     binary     0.526    15 0.0437  pre0_mod0_pos
```

```
# A tibble: 2 x 6
  .metric .estimator   mean     n std_err .config
  <chr>   <chr>       <dbl> <int>   <dbl> <chr>
1 rmse    standard   0.0312    15 0.00170 pre0_mod0_post0
2 rsq     standard   0.854     15 0.0225  pre0_mod0_post0
```

(f) **(2 points)** What could be the estimate of model accuracy for our classificaiton model? Choose one.

- ☐ 0.499
- ☐ 0.254
- ☐ 0.526
- ☐ 0.0312
- ☐ 0.854
- ☐ All of the above