# Comments for MEDB 5502, Module 01

# Topics to be covered

- What you will learn
    - Syllabus
    - Linear regression with one continuous variable
    - Linear regression with one binary categorical variable
    - Live demo, part 1
    - Logistic regression with one continuous independent variable
    - Logistic regression with one binary categorical variable
    - Live demo, part 2

# Syllabus

You can find the syllabus for this class on my github site.

# Break #1

- What you have learned
  - Syllabus
- What's coming next
  - Linear regression with one continuous variable

# Linear regression interpretation of a straight line

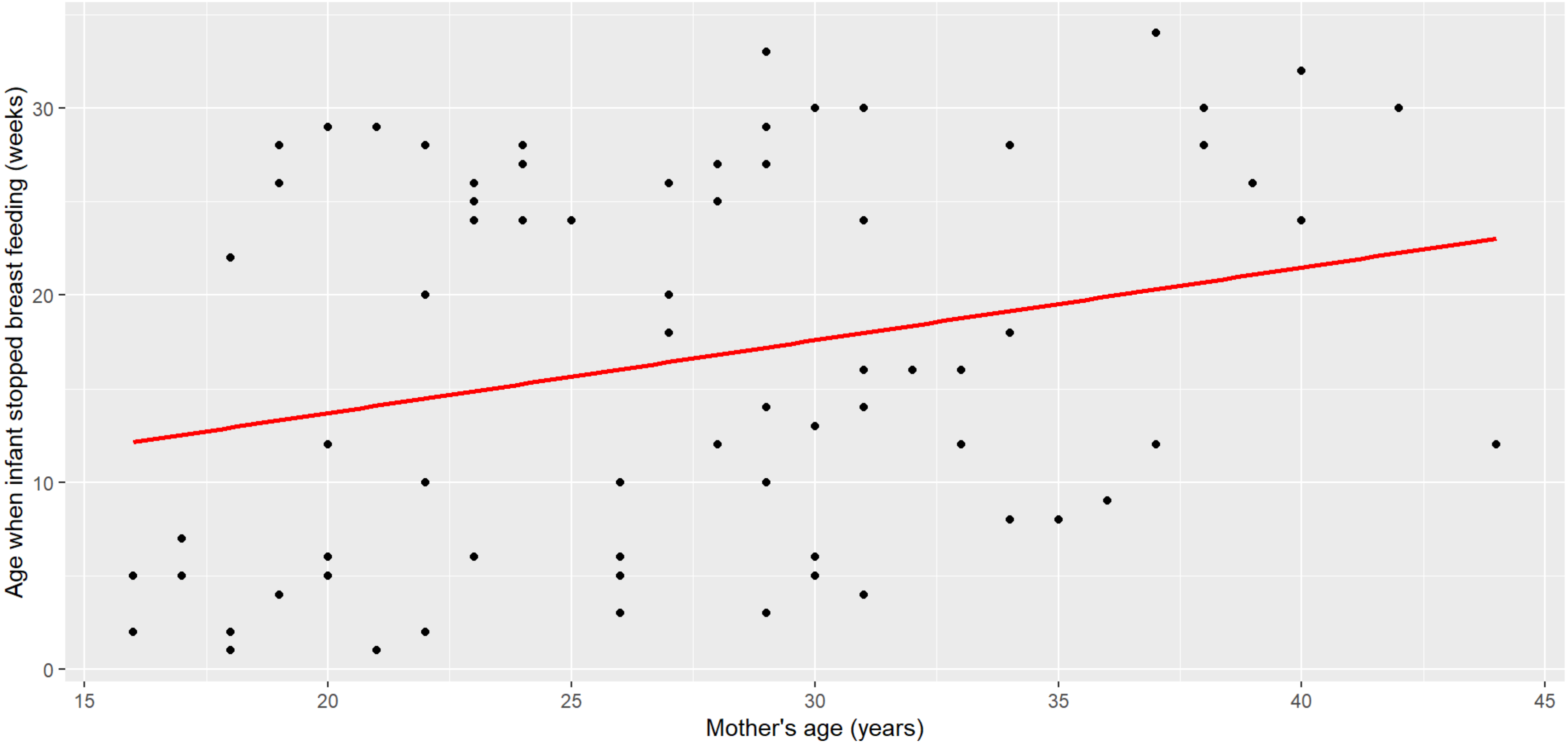
- Regression equation
  - $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- The slope represents the estimated average change in Y when X increases by one unit.
- The intercept represents the estimated average value of Y when X equals zero.
- Terminology
  - X is the independent or predictor variable
  - Y is the dependent or outcome variable

Speaker notes

In linear regression, we use a straight linear to estimate a trend in data. We can’t always draw a straight line that passes through every data point, but we can find a line that “comes close” to most of the data. This line is an estimate, and we interpret the slope and the intercept of this line as follows:

Be cautious with your interpretation of the intercept. Sometimes the value X=0 is impossible, implausible, or represents a dangerous extrapolation outside the range of the data.

# Simple regression example with interpretation, 1



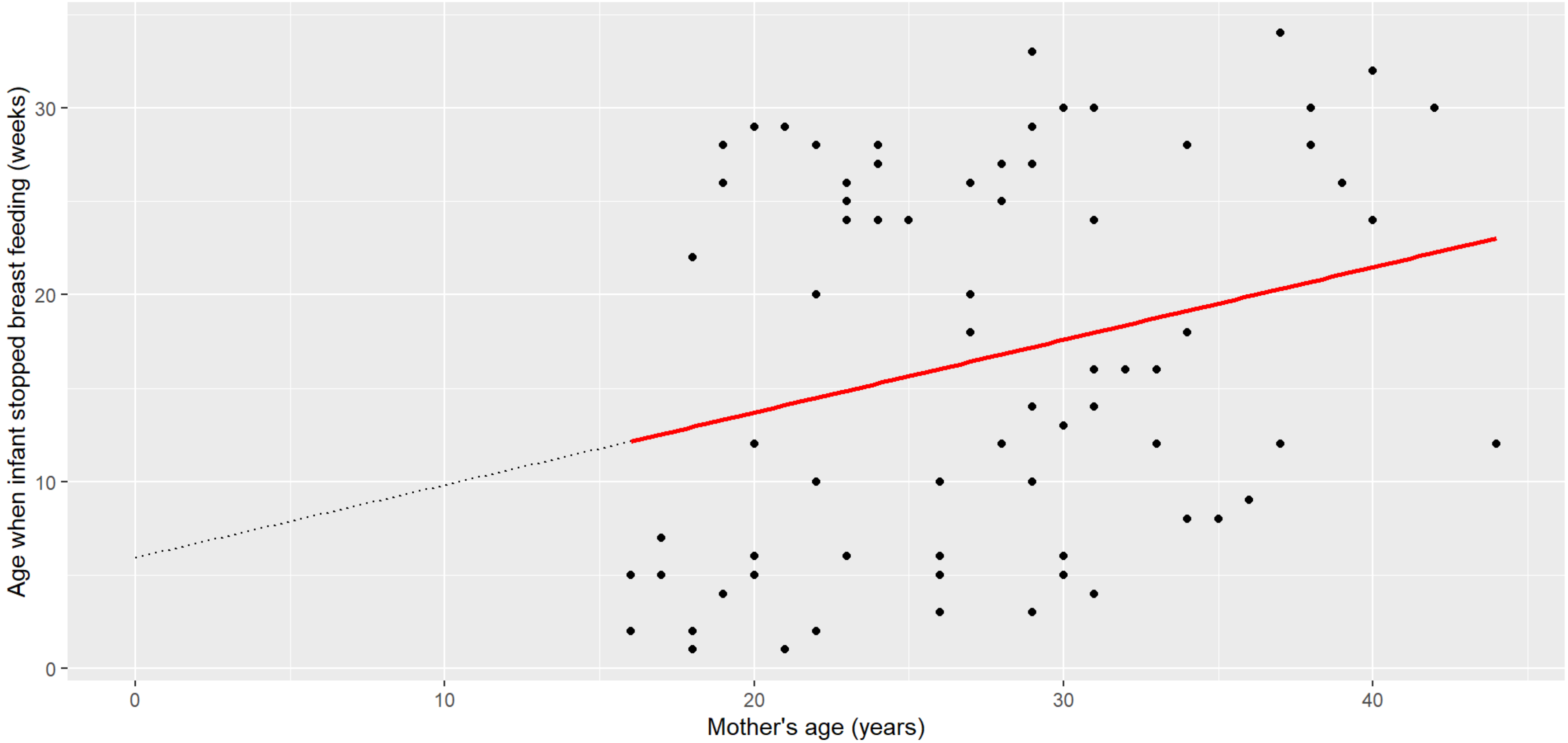
Steve Simon and Suman Sahil, 2025-01-21, CC0

Speaker notes

Here is an illustration of a linear regression model. This data is from a study of breast feeding in pre-term infants. Successful breast feeding is more difficult for a pre-term infant because the mother goes home from the birth hospital before the infant. The independent variable, X, is the mother’s age. The dependent variable is the age at which the infant stopped breast feeding. The goal is to reach at least six months of breast feeding.

Notice a weak trend here. Older mother’s seem to do a bit better than younger mothers, but there are some 20 year old moms who breast feed for quite a long time and some 40 year old moms who stop breast feeding early. Still, there is a tendency for older moms to breast feed longer than younger moms.

# Simple regression example with interpretation, 2



Steve Simon and Suman Sahil, 2025-01-21, CC0

Speaker notes

Here is a modification of the graph that expands the limits of the X axis to include X=0. This graph also extends the line beyond the range of the data all the way down to X=0. I used a dotted line and a different color to emphasize that this is an extrapolation beyond the range of the data.

The graph shown below represents the relationship between mother's age and the duration of breast feeding in a research study on breast feeding in pre-term infants.

The regression coefficients are shown below. The intercept, 6, is represented the estimated average duration of breast feeding for a mother that is zero years old. This is an impossible value, so the interpretation is not useful. What is useful, is the interpretation of the slope, approximately 0.4. The estimated average duration of breast feeding increases by 0.4 weeks for every extra year in the mother's age.

# Simple regression example with interpretation, 3

```
# A tibble: 2 × 5
  term        estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)    5.92      4.58       1.29  0.200 
2 mom_age        0.389     0.162      2.40  0.0188
```

Speaker notes

The actual values are pretty close to the rough estimates that we got from the graph.

# Predicted values, 1

- How long would you expect a 20 year old mom to breast feed?

```
# A tibble: 5 × 2
  mom_age age_stop
    <dbl>    <dbl>
1      20        5
2      20       29
3      20        6
4      20       NA
5      20       12
```

Speaker notes

Easy way out is to find a 20 year old mother in the data. That is actually not such a good idea. But there are five of them, four if you are only counting non-missing values.

In a different dataset maybe you have data on 19 and 21 year olds but no 20 year olds.

If you predict using a linear regression model, you are incorporating information from all mothers into your prediction, not just 20 year old mothers. Unless there are some major problems with the regression model, this is a much better choice.

# Predicted values, 2

- For an existing value in the data, $X_i$
  - $\hat{Y}_i = b_0 + b_1 X_i$
- For a new value of X
  - $\hat{Y}_{new} = b_0 + b_1 X_{new}$
  - Do not predict outside the range of X values

Speaker notes

To make a prediction at an existing value in the dataset, use the first formula. If you want to make a prediction at a value that is not in your dataset, use the same formula, but notice that the subscript is “new” to emphasize that this is a new value not seen before in the data. Be careful here. It is okay to make predictions inside the range of the X values. In the breast feeding example that I have been talking about, it is okay to make predictions for a mother between the ages of 16 and 44, but making predictions for a 14 year old mother or a 50 year old mother is risky. You could be making a dangerous extrapolation.

# Why predict for a value you already have seen?

- Future Y may differ from previous Y
- $\hat{Y}_i$ is more precise
- Comparison of $\hat{Y}_i$ to existing $Y_i$.

# Predicted values, 3

Predicted age_stop = 5.92 + 0.389*20 = 13.7

```
# A tibble: 1 × 2
  mom_age .fitted
    <dbl>   <dbl>
1      20    13.7
```

Speaker notes

Here is the predicted value from R.

# Residuals, 1

- $e_i = Y_i - \hat{Y}_i$
  - Residual = Observed - Predicted
- Very helpful in assessing assumptions

Speaker notes

The residual is also important. It represents the deviation between what was actually observed and what was predicted.

The residual is very helpful in assessing the assumptions needed for linear regression. This is a topic I will reserve for a later module.

# Residuals, 2

```
# A tibble: 4 × 5
  .rownames mom_age age_stop .fitted .resid
  <chr>       <dbl>    <dbl>   <dbl>  <dbl>
1 8              20        5    13.7  -8.70
2 40             20       29    13.7  15.3 
3 44             20        6    13.7  -7.70
4 67             20       12    13.7  -1.70
```

# Break #2

- What you have learned
  - Linear regression with one continuous variable
- What’s coming next
  - Linear regression with one binary categorical variable

# Categorical independent variables, 1

- Regression equation
  - $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- How do you modify this if $X_i$ is categorical?
  - Indicator variables
- Examples
  - Treatment: active drug=1, placebo=0
  - Second hand smoke: exposed=1, not exposed=0
  - Gender: male=1, female=0
- To be discussed later: three of more category levels

Speaker notes

The regression equation expects a numerical value for both X_i and Y_i. What if X_i is a categorical variable like treatment group, second-hand smoke exposure, or gender? You can't plug a category like "active drug" or "placebo" into this equation.

The trick is to convert your categorical variable into an indicator variable. An indicator variable is equal to 1 for a particular category and 0 for the other category.

It is a bit arbitrary which category gets the 1 and which gets the 0. I like to visualize the choice as 0 representing the absence of a quality and 1 representing the presence of a quality. So I always choose 0 for the placebo group and 1 for the active drug. I choose 0 for the unexposed group and 1 for the group with exposure.

So for gender, I always use 0 for females and 1 for males. This represents absence or presence of the y-chromosome.

# Categorical independent variables, 2

- If $X_i$ = 0
  - $Y_i = \beta_0 + \beta_1(0) + \epsilon_i$
  - $Y_i = \beta_0 + \epsilon_i$
- If $X_i$ = 1
  - $Y_i = \beta_0 + \beta_1(1) + \epsilon_i$
  - $Y_i = \beta_0 + \beta_1 + \epsilon_i$

Speaker notes

When X is equal to either zero or one, the equation simplifies. For the “zero category”, Y is just equal to beta0 plus epsilon. For the “one category”, Y is equal to beta0 plus beta1 plus epsilon.

# Categorical independent variables, 3

- Intperetation
  - $b_0$ is the estimated average value of Y when X equals the “zero category”
  - $b_1$ is the estimated average change in Y when X changes from the “zero category” to the “one category.”

Speaker notes

The interpretation changes, but only slightly, when X is an indicator variable.

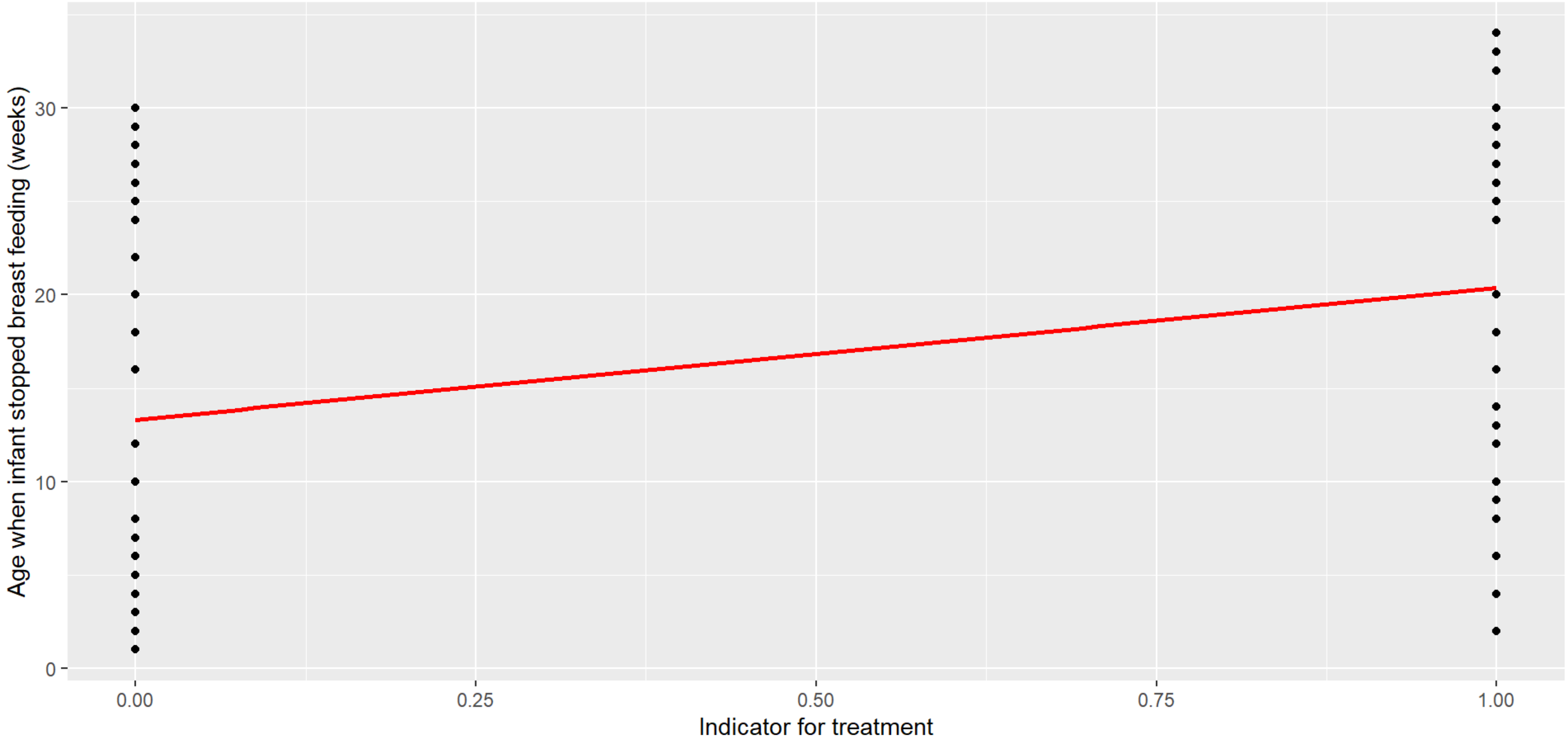
# Creating an indicator variable

```
# A tibble: 6 × 3
  feed_type age_stop i_treatment
  <chr>        <dbl>       <dbl>
1 Treatmen        30           1
2 Treatmen         4           1
3 Control         12           0
4 Treatmen        29           1
5 Control         24           0
6 Control         24           0
```

Speaker notes

Here is a small piece of the breastfeeding dataset with an indicator variable, i_treatment, added.

# Graphical display using the indicator variable



Steve Simon, 2025-01-23, CC0

Speaker notes

It’s a bit hard to read this graph, but it looks like the line is around 13 when X equals zero. That would be the intercept. The line does show an increase . At X equals one, the line appears to be around 20. This is a 7 unit increase in Y for a one unit increase in X.

# Linear regression using the indicator variable

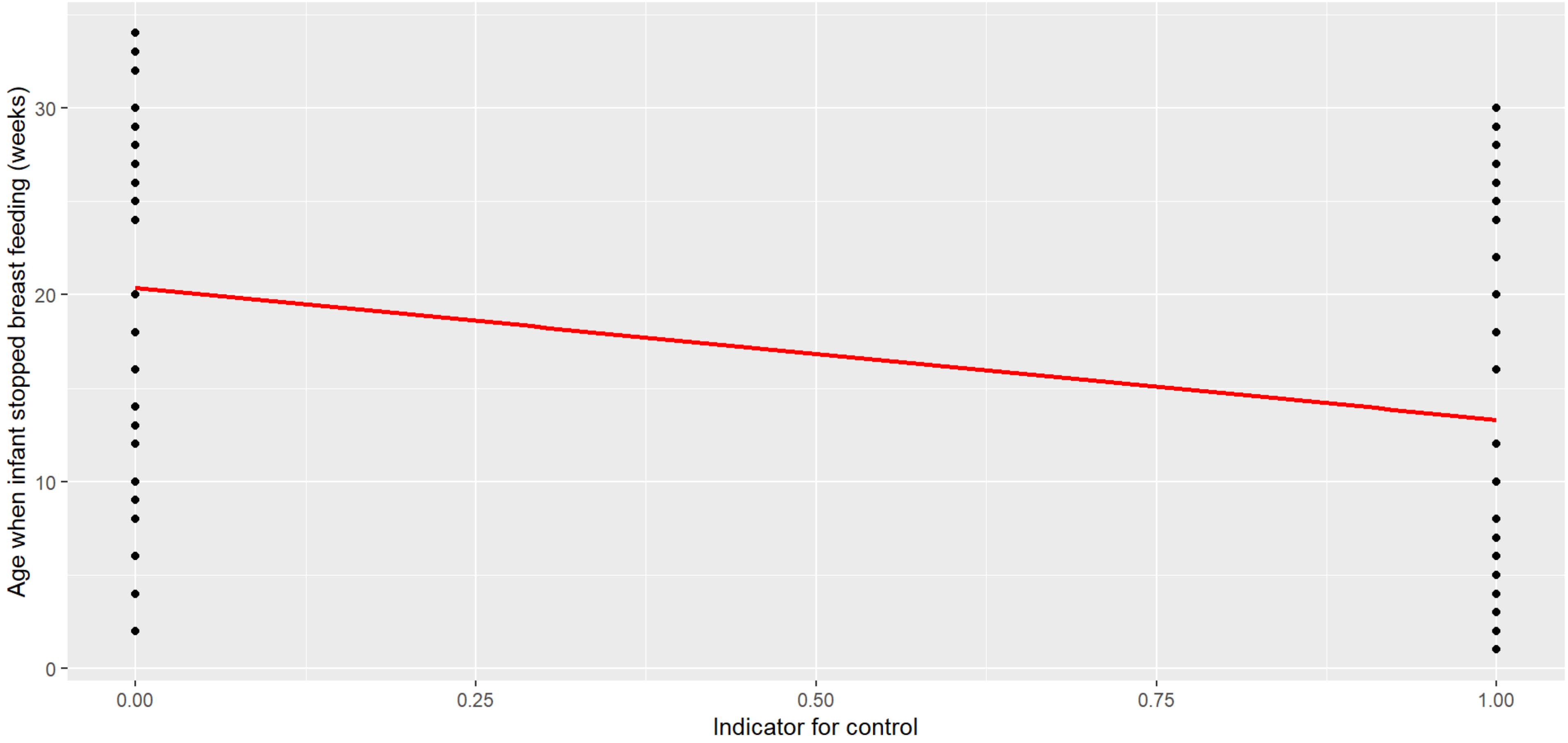
```
# A tibble: 2 × 5
  term                 estimate std.error statistic  p.value
  <chr>                   <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)             13.3       1.46      9.13 4.78e-14
2 feed_typeTreatmen        7.05      2.14      3.29 1.48e- 3
```

The estimated average fev value is 2.45 liters for females. The estimated average fev value is 0.36 liters larger for males.

Speaker notes

The intercept represents the estimated average value of Y when X equals zero. In this case, it represents the estimated average fev for female.s The slope represents the estimated average value of Y when X increases by one unit. In this case, it represents how much larger the estimated average fev is for males compared to females.

# Graphical display using alternate indicator variable



Steve Simon, 2025-01-23, CC0

Speaker notes

The choice of 1 for treatment was arbitrary, and you could have just as easily designated 1 as the control. When you do, the graph flips. The intercept is 20 and the slope is -7.

# Letting your software create the indicator variable

- Different rules for different software
  - SPSS, SAS: first alphabetical category=1, second=0
  - R: second alphabetical category=1, first=0
- Always compare your output to the descriptive statistics

Speaker notes

You don’t have to create the indicator variable yourself. Most statistical software will do it for you. Just be careful because the software has to make an arbitrary choice. SPSS and SAS choose the first category that appears when you put the data in alphabetical order. So they would choose control as 1 because the “control” appears alphabetically before the “treatment”. R does the opposite. If you ask R to create indicator variables automatically, it codes treatment as 1.

It is easy to get confused about this, so you should always orient yourself by looking at the graphs and simple descriptive statistics before trying to interpret the output from a linear regression model.

# Break #3

- What you have learned
  - Linear regression with one binary categorical variable
- What's coming next
  - Live demo, part 1

# Live demo, part 1

Review part 1 of the output from simon-5502-01-demo.

Speaker notes

Add note here.

# Break #4

- What you have learned
  - Live demo, part 1
- What’s coming next
  - Logistic regression with one continuous independent variable

# Probability

- Binary outcome
  - n1 successes
  - n2 failures
- Probability = n1 / (n1+n2)
  - Always between 0 and 1

# Odds

- Odds = n1/n2
  - Always between 0 and +∞
- Log-odds = log(n1/n2)
  - Traditional to use natural logarithm (base e)
  - Always between -∞ and +∞

# Logistic regression

- Linear on a log-odds scale
  - Impossible to produce an invalid response
- Back transform
  - Odds = exp(log-odds)
  - Probability = Odds / (1 + Odds)

# Interpretation of logistic regression coefficient

- Intercept: estimated log odds when X=0
  - Often an extrapolation
- Slope: estimated average change in log-odds
  - When x increases by one unit
  - Equivalent to a log odds ratio

# Untransformed logistic regression coefficients

```
# A tibble: 2 × 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  -2.38      0.636      -3.74 0.000183
2 dc_age        0.0336    0.0151      2.23 0.0256
```

# Transformed logistic regression coefficients

```
# A tibble: 2 × 6
  term        estimate std.error statistic  p.value transformed_estimate
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>                <dbl>
1 (Intercept)  -2.38      0.636      -3.74 0.000183               0.0925
2 dc_age        0.0336    0.0151      2.23 0.0256                 1.03
```

# Prediction from a logistic regression model

- log-odds = -2.38 + 40 $\times$ 0.034 = -1.034
- odds = exp(-1.034) = 0.356
- probability = 0.356 / (1 + 0.356) = 0.262

# Break #5

- What you have learned
  - Logistic regression with one continuous independent variable
- What’s coming next
  - Logistic regression with one binary categorical variable

# Binary independent variable

- Slight change in interpretation
- Intercept: estimated log odds for 0 category
- Slope: estimated log odds ratio
    - comparing 0 category to 1 category

# Untransformed logistic regression coefficients

```
# A tibble: 2 × 5
  term        estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)    0.103     0.788     0.131   0.896
2 del_type      -0.870     0.532    -1.64    0.102
```

# Transformed logistic regression coefficients

```
# A tibble: 2 × 6
  term        estimate std.error statistic p.value transformed_estimate
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>                <dbl>
1 (Intercept)    0.103     0.788     0.131   0.896                 1.11 
2 del_type      -0.870     0.532    -1.64    0.102                 0.419
```

# Calculation of odds ratio from crosstabulation, 1

```
          Continued bf Stopped bf
Vaginal             28         13
C-section           36          7
```

# Calculation of odds ratio from crosstabulation, 1

- Odds_vaginal = 13 / 28 = 0.464
- Odds_csection = 7 / 36 = 0.194
- Odds_ratio = 0.194 / 0.464 = 0.419

# Break #6

- What you have learned
  - Logistic regression with one binary categorical variable
- What’s coming next
  - Live demo, part 2

# Live demo, part 2

Review part 2 of the output from simon-5502-01-demo.

# Summary

- What you have learned
  - Syllabus
  - Linear regression with one continuous variable
  - Linear regression with one binary categorical variable
  - Live demo, part 1
  - Logistic regression with one continuous independent variable
  - Logistic regression with one binary categorical variable
  - Live demo, part 2