Week 3 - Simple and multiple linear regression

PSC 103B

Marwin Carmo

Today's dataset

```
reading <- read.csv("data/Lab3Data.csv", header = TRUE)
```

`ParentChildAct`: A composite measure of how often parents spend time with reading-related activities to their children, measured in minutes.

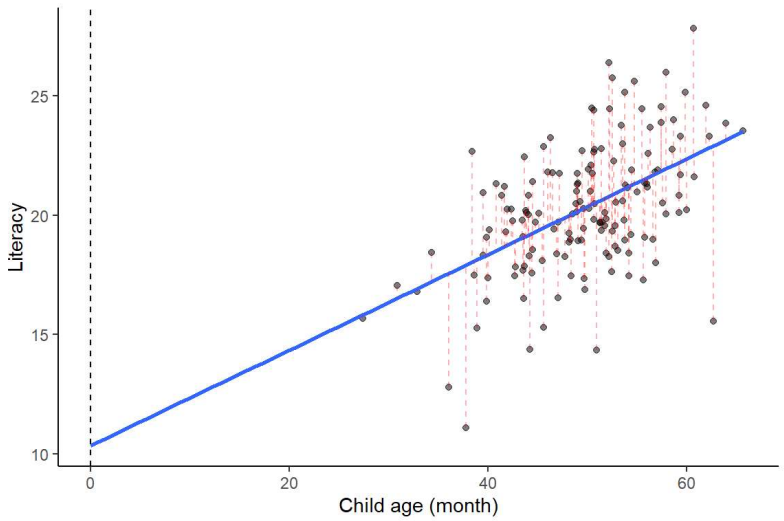
`ChildAge`: The child's age in months.

Five measures of literacy and language skills: `PrintKnowledge`, `ReadingInterest`, `EmergentWriting`, `ExpressLang`, and `ReceptiveLang`.

`OverallLiteracy`: A composite measure averaging the 5 measures of literacy/language.

Simple linear regression

- Regression examines the relation between an outcome variable and a set of q predictor variables.
- In simple linear regression there is only one predictor:
- $y_i = b_0 + b_1 x_1 + \epsilon_i$
  - b_0 = intercept;
  - b_1 = slope;
  - ϵ_i = residuals.



Simple linear regression

- In R we use the `lm()` to estimate a regression model
- `lm(dependent_variable ~ independent_variable, data = data_name)`

Exercise

Fit a linear model with the `lm()` function predicting `OverallLiteracy` from `ChildAge` and save to an object called `literacy_age`.

```
1 literacy_age <- lm(OverallLiteracy ~ ChildAge, data = reading)
```

- What function can we use to see the results of the regression analysis?

The `lm()` output

```
1 summary(literacy_age)
```

```
Call:
lm(formula = OverallLiteracy ~ ChildAge, data = reading)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3463 -1.3542 -0.0035  1.5401  5.5831

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.33375    1.40062   7.378 1.07e-11 ***
ChildAge     0.20033    0.02792   7.175 3.23e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.249 on 148 degrees of freedom
```

Exercise

Now fit a linear model predicting `OverallLiteracy` from `ParentChildAct`.

Q1 Q2 Q3

What is the estimate of the y-intercept for the model, rounded to two decimal places? What does this number mean?

> 13.27. This means that when the parent and child activity is 0 we expect the child's literacy score to be 13.27.

Confidence intervals for the estimates

We can obtain the 95% CI using the `confint()` function:

```r
confint(object = literacy_parent, level = 0.95)
```

```
                   2.5 %     97.5 %
(Intercept)    9.9718817 16.5645063
ParentChildAct 0.1119778  0.3072325
```

- If we were to collect another sample size of 150, measure Parent-child activity and a Literacy score, and estimate b_1 over and over again, 95% of the intervals would cover the true value – and 5% would miss it.

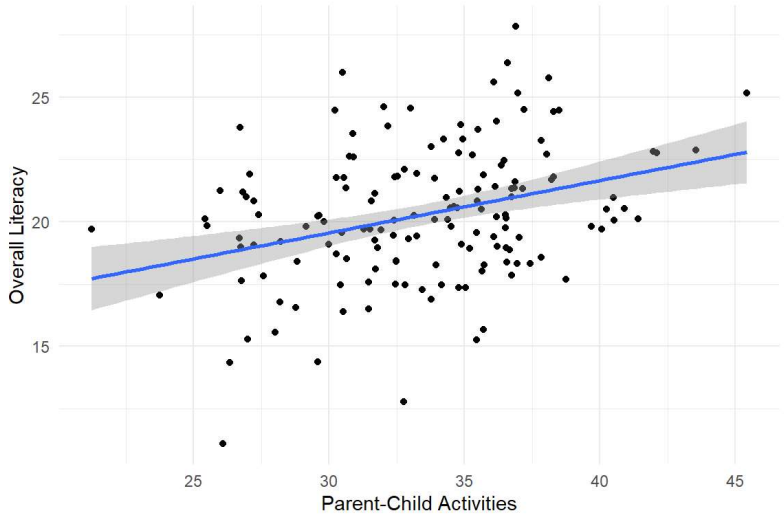
Write-up the results

> We ran a simple linear regression to determine whether Parent-Child activity predicts child reading literacy. The effect of Parent and Child activity is statistically significant and positive ($b = 0.21$, 95% CI [0.11, 0.31], $t(148) = 4.24$, $p < .001$). Parent-Child activity explained 11% of the total variability in child reading literacy.

Confidence Limits on Y

This is analogous to calculating a confidence interval for each of the predicted values $\hat{Y}_i$.

Given our model, where would we expect to see future observations fall?



Confidence Limits on Y

Let's say that we want to calculate the prediction interval for $\widehat{Literacy}_i$ for a child whose parents spend 20 minutes each day reading with them.

$$\hat{Y}_i = 13.27 + 0.21X_i$$

Step 1: calculate $\hat{Y}^*$

```
1 pred_literacy <- 13.27 + 0.21*20
2 pred_literacy
```

```
[1] 17.47
```

Confidence Limits on Y

Step 2: calculate $s_{\hat{Y}^*}$

- $s_{\hat{Y}^*} = s_\epsilon \sqrt{1 + \frac{1}{N} + \frac{(X^* - \bar{X})^2}{SS_X}}$
- s_ϵ:

```r
sqrt( (1/(nrow(reading) - 2)) * sum((reading$OverallLiteracy - predict(literacy_parent))^2) )
```

```
[1] 2.56781
```

```r
# OR
s_epsilon <- summary(literacy_parent)$sigma
```

- SS_X:

```r
ss_x <- sum((reading$ParentChildAct - mean(reading$ParentChildAct))^2)
```

Confidence Limits on Y

Step 2: calculate $s_{\hat{Y}^*}$

$$s_{\hat{Y}*} = s_\epsilon \sqrt{1 + \frac{1}{N} + \frac{(X^* - \bar{X})^2}{SS_X}}$$

```
1 s_y <- s_epsilon * sqrt( 1+(1/nrow(reading)) + (( (20 - mean(reading$OverallLiteracy))^2 ) / ss_x))
2 s_y
```

```
[1] 2.576395
```

Confidence Limits on Y

Step 3: Calculate the 95% P.I. for $\hat{Y}^*$

```
1 t_crit <- qt(.025, nrow(reading) - 2, lower.tail = FALSE)
```

Lower confidence limit:

$$\hat{Y}^* - t_{crit}\left(s_{\hat{Y}^*}\right)$$

```
1 pred_literacy - (t_crit*s_y)
```

```
[1] 12.37873
```

Upper confidence limit:

$$\hat{Y}^* + t_{crit}\left(s_{\hat{Y}^*}\right)$$

```
1 pred_literacy + (t_crit*s_y)
```

```
[1] 22.56127
```

Now you try

Exercise

Go back to the estimates given by the model `literacy_age`. Find the 95% Prediction Interval for `OverallLiteracy` for a child of 36 months of age.

- If we sample several children from the population, and all of them age 36 months, we expect 95% to have a Overall Literacy test score between 12.89 and 22.17.

```
1 predict(object = literacy_age, newdata = data.frame( ChildAge=36), level = 0.95, interval = "predict")
```

```
       fit      lwr      upr
1 17.54548 12.84008 22.25088
```

Multiple Regression

Multiple Regression

- When we use multiple regression, we use more than one predictor.
- $y_i = b_0 + b_1 x_{1_i} + b_2 x_{2_i} + \ldots + b_q x_{q_i} + \epsilon_i$
- Say that we want to use `ParentChildAct` and `ChildAge` simultaneously in the model, we would have:
  - $OverallLiteracy_i = b_0 + b_1 \times ChildAge_i + b_2 \times ParentChildAct_i +$

Multiple Regression - running the model

- To run it in R, we just have to add on the predictor to our model

```
1 literacy_multiple <- lm(OverallLiteracy ~ ChildAge + ParentChildAct, data = reading)
```

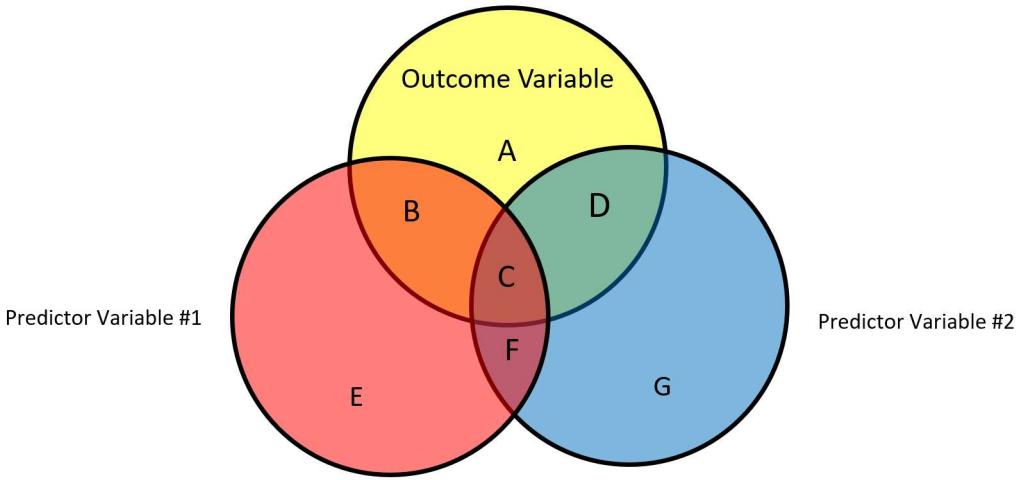
term	estimate	std.error	statistic	p.value
(Intercept)	3.3663152	1.9079381	1.764373	0.0797461
ChildAge	0.1999239	0.0259076	7.716807	0.0000000
ParentChildAct	0.2086014	0.0418195	4.988135	0.0000017

Multiple Regression - coefficients

- Now we have more terms: we have the intercept, the effect of age, and the slope associated with the time parents spend doing reading-related activities.
- Like before, the intercept is the expected value of our outcome when *all* predictors are 0.
- When a child is 0 months old, and a parent spends 0 minutes per day doing reading-related activities with them the expected literacy score is 3.37.

Multiple Regression - coefficients

- The slope is now the expected change in Y for a 1-unit increase in X, **holding the other variable constant**.
- It ensures we **isolate** the effect of the specific independent variable we're interested in, without conflating it with the influences of other variables in the model.



Multiple Regression - coefficients

- The slope for age means that if we keep the amount of time on parent-child activities constant, then increasing child's age by 1 month will increase the literacy score by 0.20.
- Similarly, the slope for parent-child activities means that if we hold the child's age constant, then increasing the time on reading activities by 1 minute is expected to increase the child's score by 0.21 points.
- Are these slopes significant?

Multiple Regression - coefficients

- Multiple regression lets us examine the unique effect of each variable in our model, given all the other variables; therefore, if the slope is still significant, it means that there is something unique about our predictor that is predicting the outcome.
- The slope for `ChildAge` and `ParentChildAct` is the same as in the simple linear regression models. But don't always expect that! That only happens because `ChildAge` and `ParentChildAct` have a correlation of only 0.003.
- There is no overlapping variance between the two, so both have maintained their unique contribution to predict the outcome.

Multiple Regression - R^2

- What is our R^2 now?

```
1 summary(literacy_multiple)$r.squared
```

```
[1] 0.3654792
```

- Compare the previous model and take note of what happened. Can you guess why?

```
1 summary(literacy_age)$r.squared
```

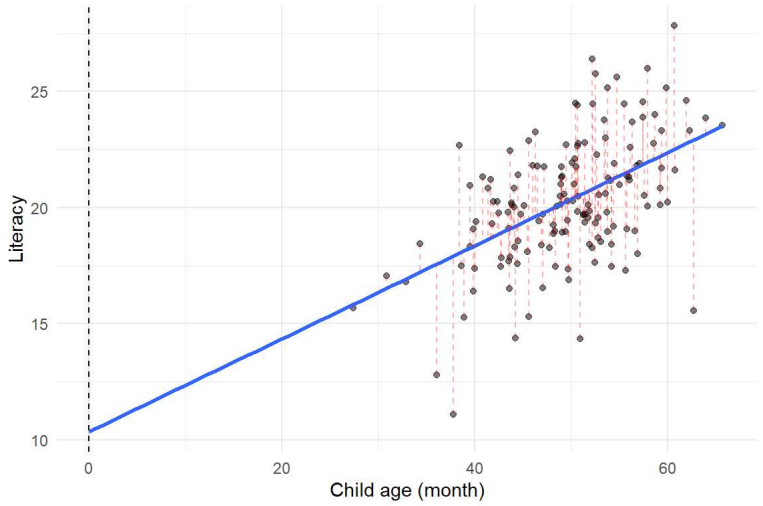
```
[1] 0.258079
```

```
1 summary(literacy_parent)$r.squared
```

```
[1] 0.1084373
```

Centering predictors

- You might have noticed that the intercept doesn't always make sense.
- Are we ever expected to have a child who is 0 months old? No! That child isn't even born yet!



Centering predictors

Centering around the mean shifts all the values of your variable so that the overall distribution stays the same but the mean of that distribution is now 0.

```
reading$ChildAge_c <- reading$ChildAge - mean(reading$ChildAge)
# OR with the scale() function
# reading$ChildAge_c <- scale(reading$ChildAge, scale = FALSE)
```

- Notice what happens

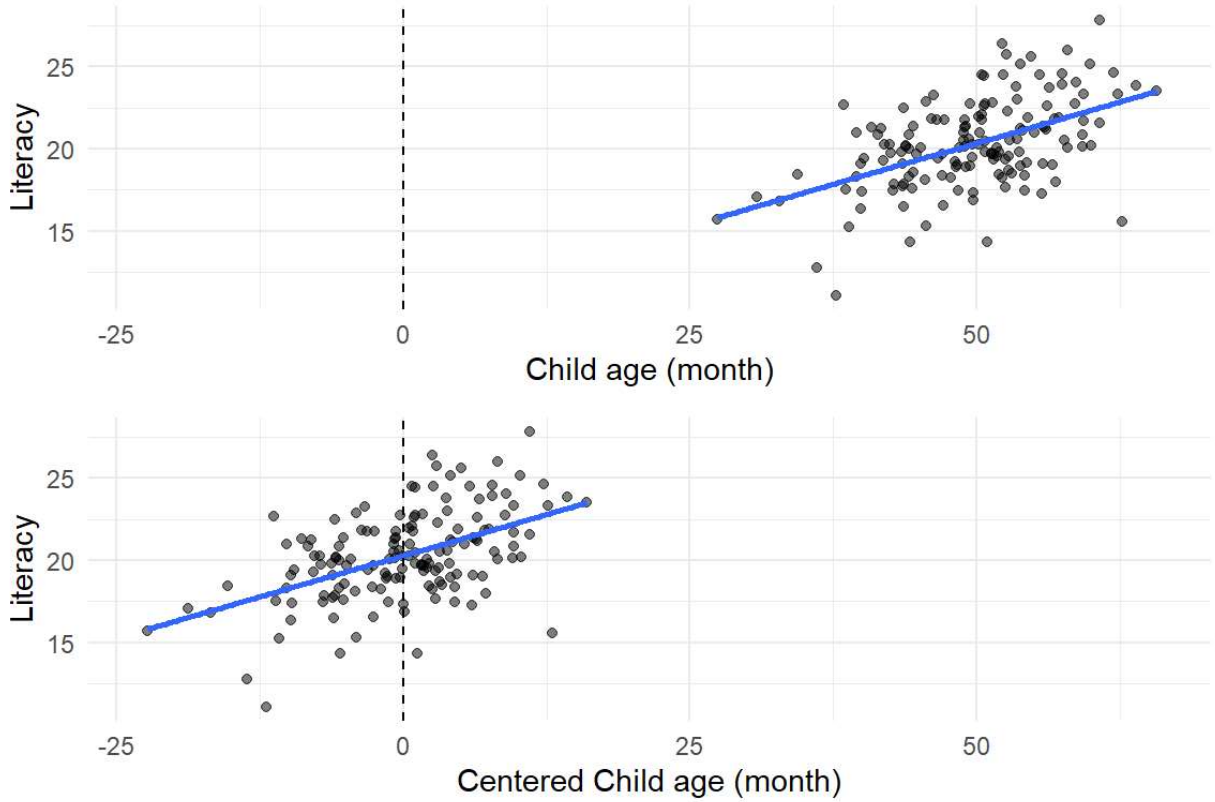
```
head(reading$ChildAge)
```

```
[1] 51.80535 52.90266 52.82730 49.01695 46.30549 56.90872
```

```
head(reading$ChildAge_c)
```

```
[1]  2.1089301  3.2062448  3.1308895 -0.6794638 -3.3909251  7.2123041
```

Centering predictors



Centering predictors

Let's also center our parent-child activities variable:

```
reading$ParentChildAct_c <- reading$ParentChildAct - mean(reading$ParentChildAct)
```

Let's re-run our regression to see what changes

```
literacy_multiple_centered = lm(OverallLiteracy ~ ChildAge_c + ParentChildAct_c,
                                data = reading)
```

term	estimate	std.error	statistic	p.value
(Intercept)	20.2892233	0.1774747	114.321769	0.0e+00
ChildAge_c	0.1999239	0.0259076	7.716807	0.0e+00
ParentChildAct_c	0.2086014	0.0418195	4.988135	1.7e-06

Centering predictors

$$Y_i = b_0 + b_1(X_{1_i} - \bar{X}_1) + b_2(X_{2_i} - \bar{X}_2) + \epsilon_i$$

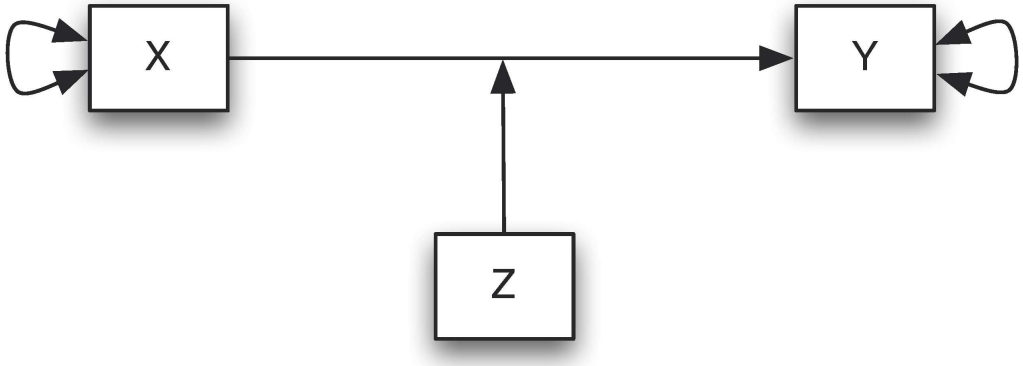
- Notice that our slopes have not changed.
- The intercept still represents the expected academic performance when all our predictors are 0.
- But now our predictors are centered, so the predictors are only 0 when the variables are equal to the average value.

Centering predictors

This makes the intercept more useful: it is the expected literacy when we have a child who is the **average age** and whose parents spend the **average amount of time doing reading related activities** with them.

Interactions

- Interactions are how we can determine whether there is a moderation effect.
- Moderation between 2 predictor variables is like saying “depends on” - the effect of one predictor variable “depends on” what value the other predictor takes.



Interactions

- Perhaps the effect of parent-child activities depends on the child's age.
- Maybe it matters that parents spend time doing these activities with their children when their children are a little bit older, versus when they're younger and parent-child activities has no effect.
- We might want to test if there is an interaction between child's age and parent-child activities.

Interactions

- To test an interaction in R, you just need to multiply your 2 predictors together instead of adding:

```r
interaction_model <- lm(OverallLiteracy ~ ChildAge_c * ParentChildAct_c, data = reading)
```

```
# A tibble: 4 × 5
  term                        estimate std.error statistic   p.value
  <chr>                          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)                  20.3       0.178     114.    1.10e-144
2 ChildAge_c                    0.204     0.0265      7.72  1.65e- 12
3 ParentChildAct_c              0.211     0.0420      5.03  1.42e-  6
4 ChildAge_c:ParentChildAct_c   0.00531   0.00622     0.854 3.95e-  1
```

Interactions

- Is our interaction term significant?
- No. So that means the effect of parent-child activities does not depend on how old the child is – it's equally important at all ages.
- But what would we do if we had a significant interaction?

Interactions

- In that case, we want to better understand what this effect is, and it's usually better to graph the regression line between the outcome and one predictor at different values of the other predictor.
- If your "other predictor" is numeric, this is usually done at 3 values: 1 SD below the mean, 1 SD above the mean, and the mean.

Interactions

```
int_plot <- sjPlot::plot_model(interaction_model, type = "int", jitter = TRUE,
            mdrt.values = "meansd", show.data = TRUE)
```