Linear Regression Practical Overview

Ken Mwai

3/19/2021

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(haven)
library(broom)## tidy model reports
```

Data

1. We are going to use a dataset from an HIV study. Read the dataset "hivinfection weight" into R We will be using the variables:

a. weight_kg - child birth weight
b. cd4 – maternal CD4 count at delivery
c. momhb – Maternal Haemoglobin levels at delivery
d. ga_weeks – gestational age at delivery/birth
e. hiv_status – Child's HIV status at birth
f. sex - the sex of the child

2. The primary research question is **"Is child birth weight associated with child HIV status?"**

- Adjust for any confounders and fit a multi-variable model

3. The primary research question is **"Is child birth weight associated with maternal CD4 count at birth?"**

- Adjust for any confounders and fit a multi-variable model

Child birth weight associate with maternal HIV status

```
hiv_df <- read_dta(file = "hivdata.dta")

## structure of the data
glimpse(hiv_df)
```

```
## Rows: 562
## Columns: 6
```

```
## $ cd4        <dbl> 652, 343, 320, 413, 351, 681, 677, 268, 419, 419, 355, 670,~
## $ momhb      <dbl> NA, NA, 134, NA, NA, 86, NA, 78, NA, 104, NA, 136, NA, 82, ~
## $ ga_weeks   <dbl> 39.71, 39.71, 38.29, 40.43, 40.43, 40.43, 40.43, 41.14, 36.~
## $ weight_kg  <dbl> 3.000, 4.050, 2.550, 2.570, 2.965, 4.070, 2.730, 4.200, 1.7~
## $ hiv_status <dbl+lbl> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0~
## $ sex        <dbl> 0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0,~
```

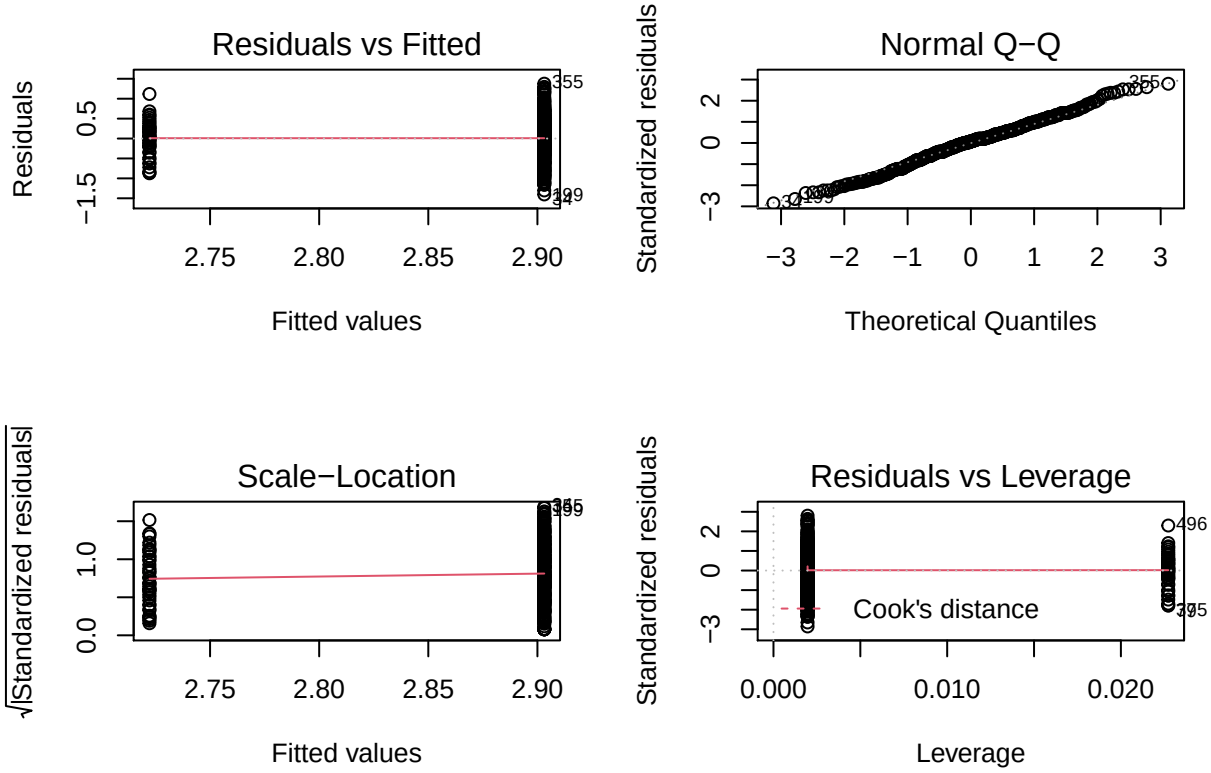
Fitting the model

```
lm1 <- lm(weight_kg ~ hiv_status , data = hiv_df)
summary(lm1)
```

```
##
## Call:
## lm(formula = weight_kg ~ hiv_status, data = hiv_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4031 -0.3031  0.0419  0.2969  1.3769
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)   2.9031     0.0218  133.16 <0.0000000000000002 ***
## hiv_status   -0.1809     0.0773   -2.34                0.02 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.492 on 551 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.00984,    Adjusted R-squared:  0.00805
## F-statistic: 5.48 on 1 and 551 DF,  p-value: 0.0196
```

Diagnostic plots for model 1

```
par(mfrow = c(2, 2))
plot(lm1)
```



Tidy model

```
confint(lm1)
```

```
##                2.5 %   97.5 %
## (Intercept)  2.8603  2.94597
## hiv_status  -0.3327 -0.02909
```

```
tidy(lm1 ,  conf.int = T )
```

```
## # A tibble: 2 x 7
##   term        estimate std.error statistic p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 (Intercept)    2.90     0.0218    133.    0         2.86     2.95
## 2 hiv_status    -0.181    0.0773     -2.34  0.0196   -0.333   -0.0291
```

Extract the AIC

```
glance_lm1 <- glance(lm1)
glance_lm1
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1   0.00984       0.00805 0.492      5.48  0.0196     1  -391.  789.  802.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Report

On average we observe that the HIV positive individuals have a significantly lower birth weight of 0.18kg (95% C.I =(-0.33, -0.03), pval=0.01) compared to the HIV negative individuals.

Adjsut for gestation weeks

```
lm2 <- lm(weight_kg ~ hiv_status + ga_weeks , data = hiv_df)
confint(lm2)
```

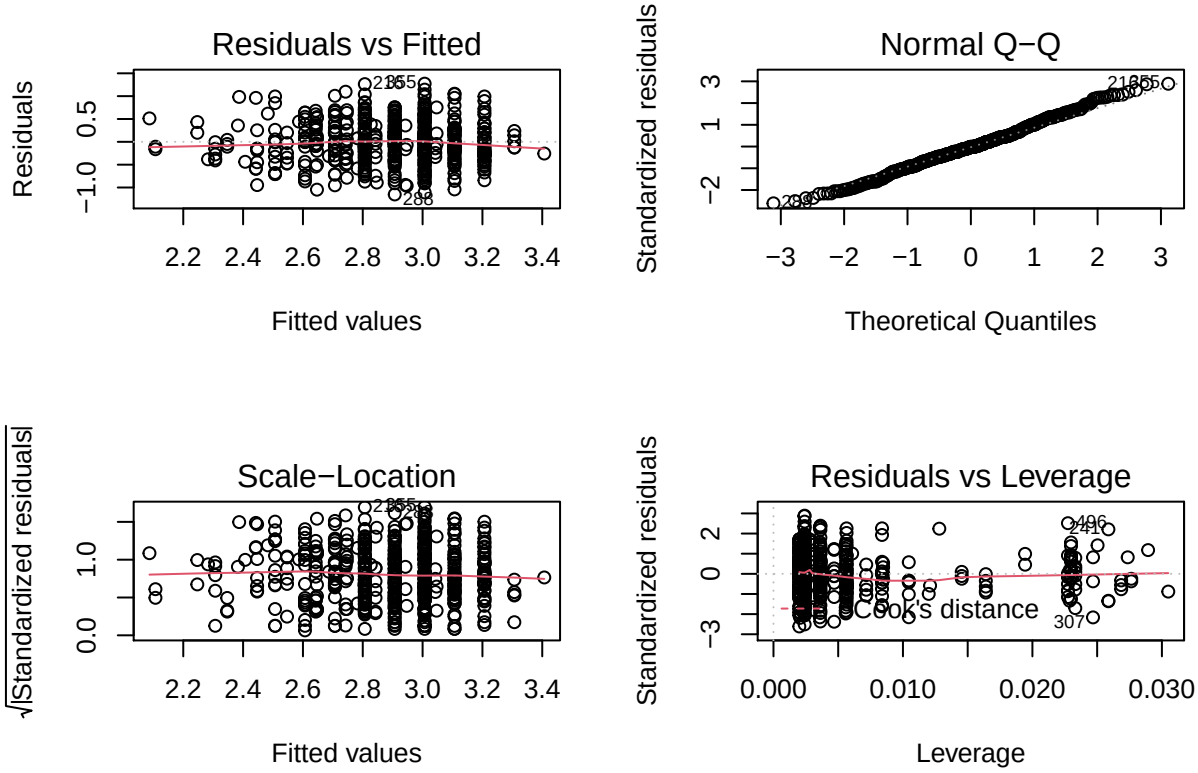
```
##                2.5 %   97.5 %
## (Intercept) -3.5019 -1.60261
## hiv_status  -0.3005 -0.02778
## ga_weeks     0.1156  0.16431
```

```
tidy(lm2 , conf.int = T )
```

```
## # A tibble: 3 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)   -2.55     0.483      -5.28 1.88e- 7   -3.50    -1.60
## 2 hiv_status    -0.164    0.0694     -2.36 1.84e- 2   -0.301   -0.0278
## 3 ga_weeks       0.140    0.0124     11.3  9.97e-27    0.116    0.164
```

Diagnostic plots for model 2

```
par(mfrow = c(2, 2))
plot(lm2)
```




```
glance_lm2 <- glance(lm2)
glance_lm2
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.199         0.196 0.441      67.3 7.72e-27     2  -326.  660.  677.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

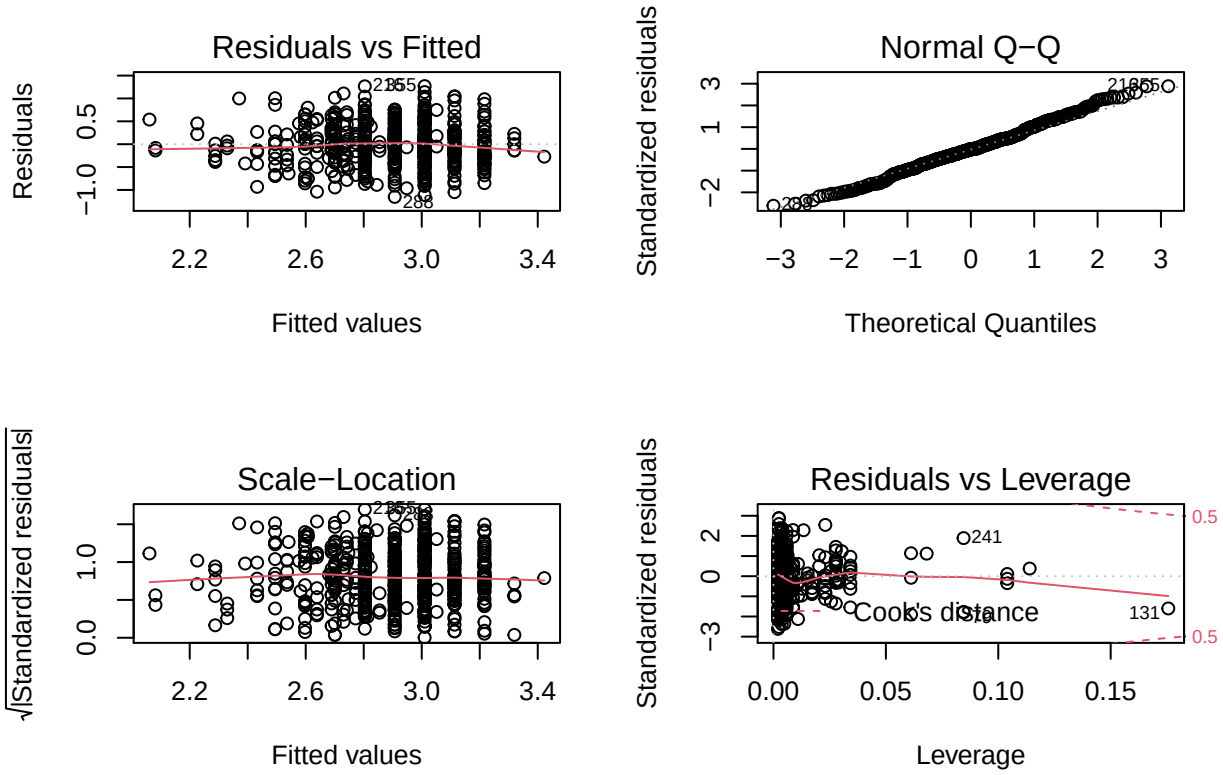
Interaction predictors

```
##https://cran.r-project.org/web/packages/interactions/vignettes/interactions.html
lm3 <- lm(weight_kg ~ hiv_status + ga_weeks + hiv_status*ga_weeks , data = hiv_df)
tidy(lm3 , conf.int = T )
```

```
## # A tibble: 4 x 7
##   term                estimate std.error statistic  p.value conf.low conf.high
##   <chr>                  <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)          -2.72      0.496      -5.50 5.98e- 8   -3.70    -1.75
## 2 hiv_status            3.20      2.19        1.46 1.44e- 1   -1.10     7.51
## 3 ga_weeks              0.144     0.0127     11.4  5.30e-27    0.119    0.169
## 4 hiv_status:ga_weeks  -0.0866    0.0564     -1.54 1.25e- 1   -0.197    0.0241
```

Diagnostic plots for model 3

```
par(mfrow = c(2, 2))
plot(lm3)
```



Child birth weight associate with maternal CD4 count

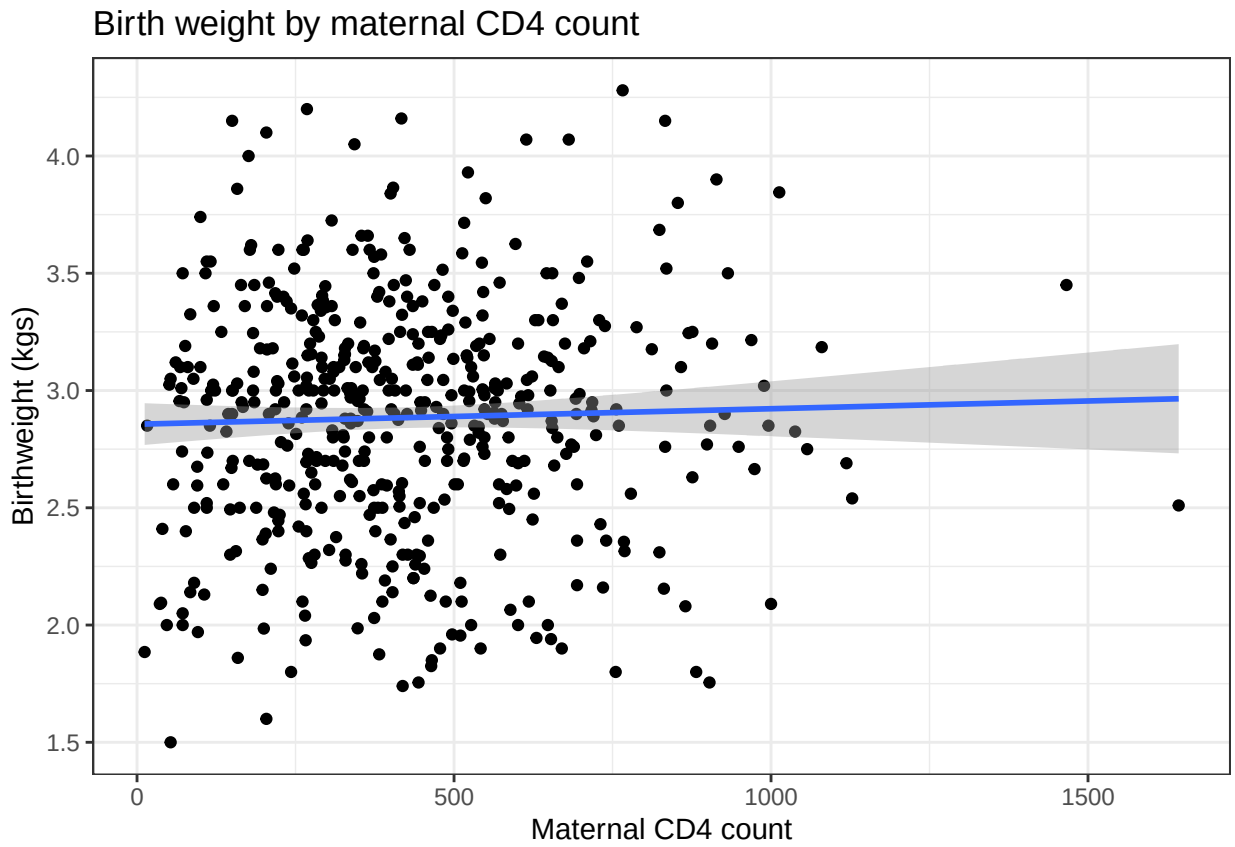
Check if there is a linear association between the outcome and the predictor

```
ggplot(data = hiv_df ,mapping = aes(x = cd4,
                                    y = weight_kg )) +
  #Add a point geom
  geom_point() +
  ##Make it straight
  geom_smooth(method = "lm") +
  ## add labels
  labs(x = "Maternal CD4 count", y = "Birthweight (kgs)",
       title = "Birth weight by maternal CD4 count")+
  theme_bw()
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 71 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 71 rows containing missing values (geom_point).
```




```
lm4 <- lm(weight_kg ~ cd4  , data = hiv_df)
#confint(lm4)
broom::tidy(lm4 , conf.int = T )
```

```
## # A tibble: 2 x 7
##   term         estimate std.error statistic   p.value  conf.low conf.high
##   <chr>           <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept) 2.86      0.0462       61.8   2.78e-233  2.77      2.95
## 2 cd4         0.0000661 0.0000953     0.694 4.88e-  1 -0.000121  0.000253
```

- Diagnostic plots model 4 ?? . We create a CD4 count in 100's

```
hiv_df <- hiv_df %>%
  mutate(cd4_100=cd4/100)
```

```
lm5 <- lm(weight_kg ~ cd4_100 , data = hiv_df)
#confint(lm4)
broom::tidy(lm5 , conf.int = T )
```

```
## # A tibble: 2 x 7
##   term        estimate std.error statistic   p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
## 1 (Intercept)  2.86      0.0462     61.8   2.78e-233   2.77      2.95
## 2 cd4_100      0.00661   0.00953     0.694 4.88e-  1  -0.0121    0.0253
```

Adjust for confounders

```
lm6 <- lm(weight_kg ~ cd4_100 + sex + ga_weeks +momhb , data = hiv_df)
#confint(lm4)
```

- Check how the sample size reduces if you fit the model using a predictor with missing data points. Compare model `lm6` and `lm5`

```
summary(lm6)
```

```
##
## Call:
## lm(formula = weight_kg ~ cd4_100 + sex + ga_weeks + momhb, data = hiv_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1343 -0.3019 -0.0099  0.2616  1.2454
##
## Coefficients:
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept) -4.164882   0.722991   -5.76          0.000000023 ***
## cd4_100     -0.000137   0.011387   -0.01                0.990
## sex          0.131831   0.053881    2.45                0.015 *
## ga_weeks     0.175296   0.017968    9.76 < 0.0000000000000002 ***
## momhb        0.001280   0.001403    0.91                0.362
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.448 on 273 degrees of freedom
##   (284 observations deleted due to missingness)
## Multiple R-squared:  0.277,  Adjusted R-squared:  0.266
## F-statistic: 26.1 on 4 and 273 DF,  p-value: <0.0000000000000002
```

```
summary(lm5)
```

```
##
## Call:
## lm(formula = weight_kg ~ cd4_100, data = hiv_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3597 -0.3142  0.0462  0.3073  1.3732
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  2.85619    0.04619   61.83 <0.0000000000000002 ***
## cd4_100      0.00661    0.00953    0.69                0.49
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5 on 489 degrees of freedom
##   (71 observations deleted due to missingness)
## Multiple R-squared:  0.000983,   Adjusted R-squared:  -0.00106
## F-statistic: 0.481 on 1 and 489 DF,  p-value: 0.488
```