Epidemiology, Biostatistics
and Prevention Institute,
Epidemiology Department
+41 44 634 46 34
www.ebpi.uzh.ch

Interpreting Interactions

Sarah R Haile (sarah.haile@uzh.ch)

Version 1.0 of June 22, 2017

Contents

1 Introduction

Regression models are often used to explore associations between different variances, sometimes including interactions. Unfortunately, interactions are sometimes hard to interpret. Here we explain the interpretation of three different kinds of interactions

1. nominal (sometimes called categorical, or binary if there are only two categories) by nominal;
2. nominal by continuous; and
3. continuous by continuous.

Code examples in STATA and R using the birthweight dataset are provided.

2 Birthweight data

The birthweight data set `birthwt` can be found in the package `MASS` in R.

```
library(MASS)
data(birthwt)
birthwt$smoke <- factor(birthwt$smoke, 0:1, c("non-smoker", "smoker"))
birthwt$race <- factor(birthwt$race, 1:3, c("white", "black", "other"))
birthwt$nonwhite <- birthwt$race != "white"
birthwt$nonwhite <- factor(as.numeric(birthwt$nonwhite), 0:1, c("white", "nonwhite"))

head(birthwt[, c("bwt", "low", "smoke", "nonwhite", "age", "lwt")])

##     bwt low      smoke nonwhite age lwt
## 85 2523   0 non-smoker nonwhite  19 182
## 86 2551   0 non-smoker nonwhite  33 155
## 87 2557   0     smoker    white  20 105
## 88 2594   0     smoker    white  21 108
## 89 2600   0     smoker    white  18 107
## 91 2622   0 non-smoker nonwhite  21 124
```

In STATA, the dataset `lbw` can be loaded from the web directly.

```
. webuse lbw
(Hosmer & Lemeshow data)

. gen nonwhite = race != 1

. list bwt low smoke nonwhite age lwt in 1/5

     +------------------------------------------------+
     |  bwt   low       smoke   nonwhite   age   lwt |
     |------------------------------------------------|
  1. | 2523     0   nonsmoker          1    19   182 |
  2. | 2551     0   nonsmoker          1    33   155 |
  3. | 2557     0      smoker          0    20   105 |
  4. | 2594     0      smoker          0    21   108 |
  5. | 2600     0      smoker          0    18   107 |
     +------------------------------------------------+
```

It should be noted that the R and STATA versions of the dataset are not exactly the same, and therefore the results shown below are slightly different. See Appendix A for a comparison.

3 Linear regression

3.1 Nominal by nominal

Without interaction With only main effects, we assume that the mean difference between categories of one variable is the same, regardless of the value of the 2nd variable, and vice versa.

With interaction Including an interaction term, we assume that the mean difference between categories of one variable differs according to the 2nd variable, and vice versa.

Interpretation of Interaction Coefficient The interaction term gives additional difference in means for non-reference levels of the two categorical variables.

Interpretation The reference category for `smoke` is non-smoking mothers, and for `nonwhite` is white mothers. Babies of smokers have on average -601.9g lower birthweights than non-smokers. Babies of non-white mothers have -604.2g lower birthweights than those of whites. However, the association with birthweight is not as strong as expected in non-white smokers, as they have on average 419.5g higher birthweights than would be expected considering the main effects only.

Interpretation for each group

Non-smoking, white mothers This is the reference group, with an average birthweight given by the intercept: 3428.7g.

Smoking, white mothers White mothers who smoke have babies with on average -601.9g lower birthweights than white mothers who do not smoke.

Non-smoking, non-white mothers Non-white mothers who do not smoke have babies with on average -604.2g lower birthweights than white mothers who do not smoke.

Smoking, non-white mothers Non-white mothers who do smoke have babies with on average $-601.6 + -604.2 + 419.5 = -786.3$g lower birthweights than white mothers who do not smoke.

```
m1 <- lm(bwt ~ smoke * nonwhite, data = birthwt)
summary(m1)

##
## Call:
## lm(formula = bwt ~ smoke * nonwhite, data = birthwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2407.75  -416.85    31.25   483.25  1561.25
##
## Coefficients:
##                                Estimate Std. Error t value   Pr(>|t|)
## (Intercept)                      3428.7      102.7  33.378    < 2e-16 ***
## smokesmoker                      -601.9      139.6  -4.312 0.00002624 ***
## nonwhitenonwhite                 -604.2      130.7  -4.622 0.00000712 ***
## smokesmoker:nonwhitenonwhite      419.5      217.1   1.932     0.0548 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 681.4 on 185 degrees of freedom
## Multiple R-squared:  0.1408,Adjusted R-squared:  0.1268
## F-statistic:  10.1 on 3 and 185 DF,  p-value: 0.000003393
```

```
. regress bwt i.smoke##i.nonwhite, noheader
------------------------------------------------------------------------------
         bwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |
      smoker |  -601.3654   139.5211    -4.31   0.000    -876.6224   -326.1084
  1.nonwhite |  -605.4401    130.685    -4.63   0.000    -863.2646   -347.6156
             |
      smoke#|
    nonwhite |
    smoker#1 |   420.1464   216.9997     1.94   0.054    -7.965728    848.2586
             |
       _cons |    3428.75   102.6848    33.39   0.000     3226.166    3631.334
------------------------------------------------------------------------------
```

3.2 Nominal by continuous

Without interaction With only main effects, we assume that the slope of y over the continuous variable, x is the same regardless of the category of the nominal variable, $z = 0$ or $z = 1$. That is, the regression lines for each group in z are parallel.

With interaction Including an interaction term, we assume that the slope of y over x differs according to $z = 0$ or $z = 1$. The regression lines for each group in z no longer are assumed to be parallel.

Interpretation of Interaction Coefficient The interaction term gives additional change in slope of y over x for the non-reference level of the nominal variable, $z = 1$. The slopes are given by:

$z = 0$: $\hat{\beta}_x$

$z = 1$: $\hat{\beta}_x + \hat{\beta}_{x:z}$

Interpretation For non-smokers, average birthweight increases by 27.7g per year of age of the mother. For smokers, the average birthweight actually decreases by -18.8g ($27.73 + -46.57$) per year of age of the mother.

```
m2 <- lm(bwt ~ smoke * age, data = birthwt)
summary(m2)

##
## Call:
## lm(formula = bwt ~ smoke * age, data = birthwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2189.27  -458.46    51.46   527.26  1521.39
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2406.06     292.19   8.235 3.18e-14 ***
## smokesmoker       798.17     484.34   1.648   0.1011
## age                27.73      12.15   2.283   0.0236 *
## smokesmoker:age   -46.57      20.45  -2.278   0.0239 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 709.3 on 185 degrees of freedom
## Multiple R-squared:  0.06909,Adjusted R-squared:  0.054
## F-statistic: 4.577 on 3 and 185 DF,  p-value: 0.004068
```

```
 . regress bwt i.smoke##c.age, noheader
------------------------------------------------------------------------------
         bwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |
     smoker  |   797.9369   484.3249     1.65   0.101    -157.5731    1753.447
         age |   27.60058   12.14868     2.27   0.024     3.632806    51.56835
             |
 smoke#c.age |
     smoker  |  -46.51558   20.44641    -2.28   0.024    -86.85368   -6.177479
             |
       _cons |   2408.383   292.1796     8.24   0.000     1831.951    2984.815
------------------------------------------------------------------------------
```

Tip Note that the main effect of smoking here gives the mean difference between smokers and non-smokers *for age = 0*. It may be easier to interpret models with nominal by continuous interactions if you first center the continuous variable (at mean, median or other relevant value).

```
median(birthwt$age)

## [1] 23

birthwt$agec <- birthwt$age - 23
m2c <- lm(bwt ~ smoke * agec, data = birthwt)
summary(m2c)$coef

##                     Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)       3043.87967   66.34054 45.882648 5.136217e-103
## smokesmoker       -272.97916  105.82868 -2.579444  1.067228e-02
## agec                27.73138   12.14910  2.282587  2.359245e-02
## smokesmoker:agec   -46.57191   20.44711 -2.277677  2.388962e-02
```

```
 . centile age

                                                       -- Binom. Interp. --
    Variable |       Obs  Percentile      Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
         age |       189         50            23        21.50878          24

. gen agec = age - 23

. regress bwt i.smoke##c.agec, noheader
------------------------------------------------------------------------------
         bwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |
```

```
      smoker  |  -271.9214    105.825    -2.57   0.011    -480.7004   -63.14238
        agec  |   27.60058   12.14868     2.27   0.024     3.632806    51.56835
              |
smoke#c.agec  |
      smoker  |  -46.51558   20.44641    -2.28   0.024    -86.85368   -6.177479
              |
       _cons  |   3043.196   66.33825    45.87   0.000      2912.32    3174.073
------------------------------------------------------------------------------
```

3.3 Continuous by continuous

Without interaction With only main effects, we assume that the slope of y over the continuous variable $x1$ is the same regardless of $x2$ and vice versa.

With interaction Including an interaction term, we assume that the slope of y over the continuous variable $x1$ differs with respect to $x2$, and vice versa.

Interpretation of Interaction Coefficient The interaction term gives the change in slope of y over $x1$ for each unit of $x2$, and the change in slope of y over $x2$ for each unit of $x1$. The actual slopes are given by:

slope over $x1$: $\hat{\beta}_{x1} + x_2\hat{\beta}_{x1:x2}$

slope over $x2$: $\hat{\beta}_{x2} + x_1\hat{\beta}_{x1:x2}$

Interpretation Average birthweight increases by on average 11.7g for every year of the mother's age, and 4.4g for each pound of the mother's weight. Increasing age and weight of the mother make these associations slight less pronounced (-0.3g per year of age and pound).

Tip Unless $x1 = 0$ and $x2 = 0$ are meaningful in your dataset, you may end up with strange values for the intercept or other main effect estimates. If this happens, try centering continuous variables. Don't forget that this will change the calculation of the predicted values:

$$\hat{y} = \hat{\beta}_{(Intercept)} + \hat{\beta}_{agec}(age - 23) + \hat{\beta}_{lwtc}(lwt - 121) + \hat{\beta}_{agec:lwtc}(age - 23)(lwt - 121)$$

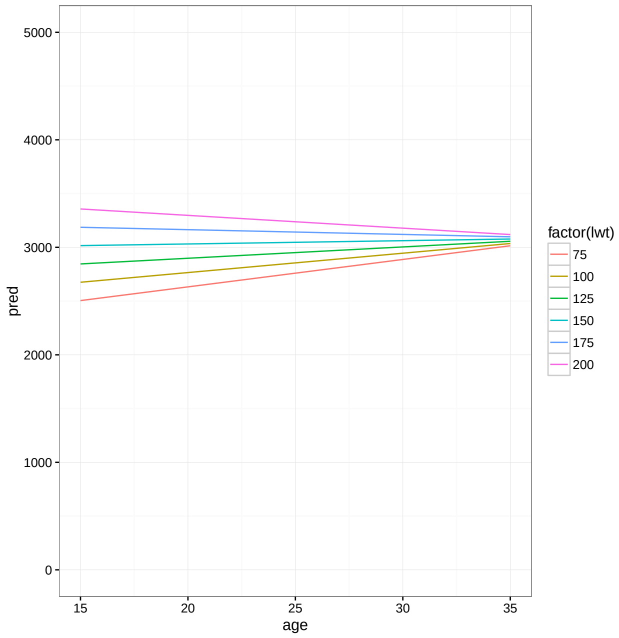
```
median(birthwt$lwt)

## [1] 121

birthwt$lwtc <- birthwt$lwt - 121
m3 <- lm(bwt ~ agec * lwtc, data = birthwt)
summary(m3)

##
## Call:
## lm(formula = bwt ~ agec * lwtc, data = birthwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2258.87  -477.29    16.28   512.40  1824.01
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2912.1115    54.8888  53.055   <2e-16 ***
## agec          11.7363    10.8076   1.086    0.279
```

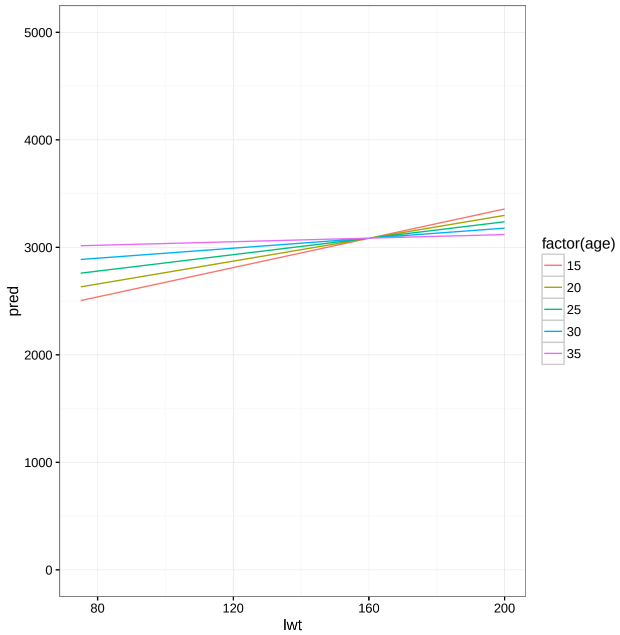
```
## lwtc          4.4237     1.7645   2.507    0.013 *
## agec:lwtc    -0.2992     0.3227  -0.927    0.355
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 719.4 on 185 degrees of freedom
## Multiple R-squared:  0.04229,Adjusted R-squared:  0.02676
## F-statistic: 2.723 on 3 and 185 DF,  p-value: 0.04569
```

```
 . centile lwt

                                                       -- Binom. Interp. --
    Variable |     Obs  Percentile      Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
         lwt |     189         50           121             120         128

. gen lwtc = lwt - 121

. regress bwt c.agec##c.lwtc, noheader
------------------------------------------------------------------------------
         bwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        agec |   11.57163   10.80576     1.07   0.286    -9.746736    32.88999
        lwtc |   4.425356   1.764444     2.51   0.013     .9443383    7.906374
             |
      c.agec#|
      c.lwtc |  -.2953255   .3226567    -0.92   0.361    -.9318852    .3412342
             |
       _cons |   2911.685   54.88113    53.05   0.000     2803.411    3019.958
------------------------------------------------------------------------------
```

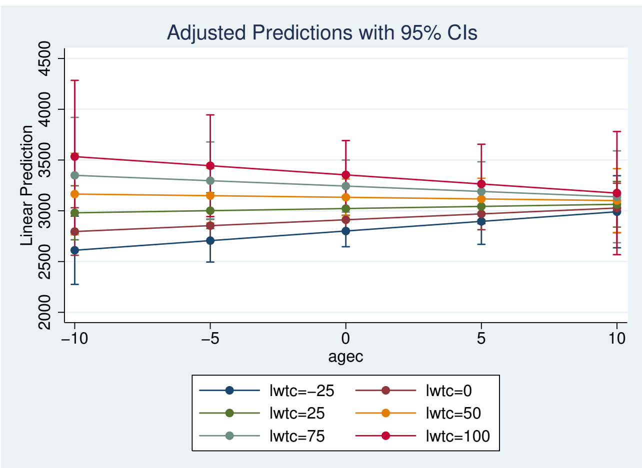
Tip Graph the predicted values in order to make sense of continuous by continuous interactions.

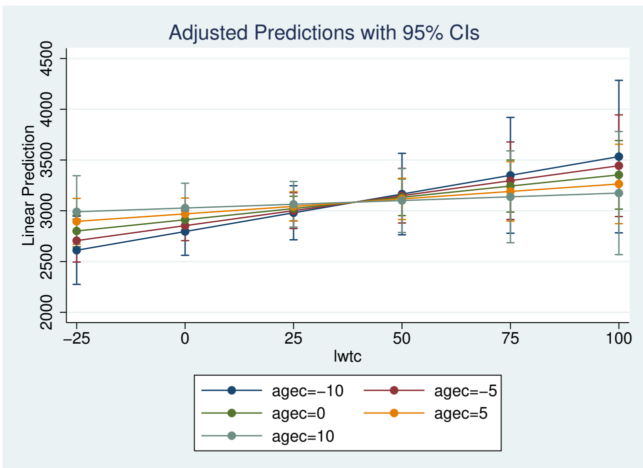
```
nd <- expand.grid(agec = seq(15, 35, 5) - 23, lwtc = seq(75, 200, 25) - 121)
nd$pred <- predict(m3, newdata = nd)
nd$age <- nd$agec + 23
nd$lwt <- nd$lwtc + 121
qplot(age, pred, data = nd, color = factor(lwt), geom = "line") + ylim(0, 5000)
qplot(lwt, pred, data = nd, color = factor(age), geom = "line") + ylim(0, 5000)
```






```
. quietly: margins, at(agec = (-10(5)10) lwtc = (-25(25)100))

. marginsplot

  Variables that uniquely identify margins: agec lwtc

. marginsplot, xdim(lwtc)

  Variables that uniquely identify margins: agec lwtc
```





4 Logistic regression

The interpretations given in this section apply equally to

- logistic regression for binary outcomes ($e^{\hat{\beta}}$ = odds ratio (OR)),
- poisson regression for count outcomes ($e^{\hat{\beta}}$ = incidence rate ratio (IRR)),
- Cox proportional hazards regression for survival outcomes ($e^{\hat{\beta}}$ = hazard ratio (HR)),
- and other regression models where relevant coefficients are intepreted as $e^{\hat{\beta}}$, not $\hat{\beta}$.

4.1 Nominal by nominal

Without interaction With only main effects, we assume that the odds ratio comparing categories of one variable is the same, regardless of the value of the 2nd variable, and vice versa.

With interaction Including an interaction term, we assume that the odds ratio comparing categories of one variable differs according to the 2nd variable, and vice versa. An OR < 1 for the interaction, indicates the association is less strong than expected when considering only the main effects, while OR > 1 indicates the association is stronger than expected.

Interpretation of Interaction Coefficient The interaction term gives multiplicative effect of non-reference levels of the two categorical variables.

For nominal by nominal interactions, we examine the effects of two covariates simultaneously by multiplying the odds ratios. To see the effect of covariates $x1$ and $x2$, we multiply $e^{\hat{\beta}_{x1}}$ with $e^{\hat{\beta}_{x2}}$ to get $e^{\hat{\beta}_{x1}}e^{\hat{\beta}_{x2}} = e^{\hat{\beta}_{x1}+\hat{\beta}_{x2}}$. (Note that we can either a) first add the coefficients and then exponentiate, or b) first exponentiate to get odds ratios, and then multiply.) With interaction, we calculate the odds ratio as follows:

$$OR_{x1,x2} = e^{\hat{\beta}_{x1}}e^{\hat{\beta}_{x2}}e^{\hat{\beta}_{x1:x2}}.$$

```
m4 <- glm(low ~ smoke * nonwhite, data = birthwt, family = binomial)
cbind("OR" = exp(coef(m4)), exp(confint(m4)))

## Waiting for profiling to be done...

##                                   OR       2.5 %     97.5 %
## (Intercept)                 0.100000 0.03001716  0.2479494
## smokesmoker                 5.757575 1.93948909 21.3664428
## nonwhitenonwhite            5.434782 1.91145601 19.6316716
## smokesmoker:nonwhitenonwhite 0.319579 0.06421315  1.3942648
```

Interpretation In this example, we look at the odds of having birthweight less than 2.5kg. Smokers have 5.76 higher odds of having a baby with low birthweight compared to non-smokers. Similarly, nonwhite mothers have a 5.43 higher odds of having a baby with low birthweight compared to white mothers. Nonwhite mothers who smoke however have a 10 times higher odds of having a baby with low birthweight than white mothers who do not smoke.

```
exp(coef(m4)["smokesmoker"]) * exp(coef(m4)["nonwhitenonwhite"]) *
      exp(coef(m4)["smokesmoker:nonwhitenonwhite"])

## smokesmoker
##    9.999999
```

STATA Tip Note the use of the `coeflegend` option to find out what the coefficients are called, in case you want to use them in calculations.

```
. logistic low i.smoke##i.nonwhite

Logistic regression                             Number of obs     =        189
                                                LR chi2(3)        =      16.97
                                                Prob > chi2       =     0.0007
Log likelihood = -108.84968                     Pseudo R2         =     0.0723

------------------------------------------------------------------------------
         low | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |
     smoker  |   5.757576   3.444621     2.93   0.003     1.782321    18.59916
  1.nonwhite |   5.434783   3.153762     2.92   0.004     1.742756    16.94837
             |
       smoke#|
    nonwhite |
   smoker#1  |   .3195789   .2478524    -1.47   0.141      .069891    1.461286
             |
       _cons |         .1   .0524404    -4.39   0.000     .0357788    .2794949
------------------------------------------------------------------------------

. logistic, coeflegend

Logistic regression                             Number of obs     =        189
                                                LR chi2(3)        =      16.97
                                                Prob > chi2       =     0.0007
Log likelihood = -108.84968                     Pseudo R2         =     0.0723

------------------------------------------------------------------------------
         low | Odds Ratio  Legend
-------------+----------------------------------------------------------------
       smoke |
     smoker  |   5.757576  _b[1.smoke]
  1.nonwhite |   5.434783  _b[1.nonwhite]
             |
       smoke#|
    nonwhite |
   smoker#1  |   .3195789  _b[1.smoke#1.nonwhite]
             |
       _cons |         .1  _b[_cons]
------------------------------------------------------------------------------

. di exp(_b[1.smoke]) * exp(_b[1.nonwhite]) * exp(_b[1.smoke#1.nonwhite])
10

. * or equivalently:
.
. di exp(_b[1.smoke] + _b[1.nonwhite] + _b[1.smoke#1.nonwhite])
10
```

4.2 Nominal by continuous

Without interaction With only main effects, we assume that the odds ratio increases the same amount per unit of the continuous variable, x, is the same regardless of the category of the nominal variable, $z = 0$ or $z = 1$.

With interaction Including an interaction term, we assume that the change in odds ratio over the continuous variable differs according the value of z

Interpretation of Interaction Coefficient The interaction term gives additional change in odds for the non-reference level of the nominal variable, $z = 1$. The ORs are given by:

$z = 0$: e^{β_x}

$z = 1$: $e^{\beta_x} e^{\beta_{x:z}}$

Interpretation In this example, the odds of having a baby with low birthweight decreases by a factor of 0.92 per every year of the mother's age if the mother doesn't smoke, and by a factor of $0.92 * 1.08 = 0.99$ for every year if she does smoke.

```
m5 <- glm(low ~ smoke * agec, data = birthwt, family = binomial)
cbind("OR" = exp(coef(m5)), exp(confint(m5)))

## Waiting for profiling to be done...

##                          OR     2.5 %    97.5 %
## (Intercept)       0.3324575 0.2115932 0.5048132
## smokesmoker       2.0492797 1.0889112 3.8894061
## agec              0.9204617 0.8382765 1.0009474
## smokesmoker:agec 1.0758199 0.9474313 1.2256759
```

```
. logistic low i.smoke##c.agec, coef

Logistic regression                               Number of obs   =        189
                                                  LR chi2(3)      =       8.66
                                                  Prob > chi2     =     0.0342
Log likelihood = -113.00535                       Pseudo R2       =     0.0369

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |
      smoker |   .7174884   .3237495     2.22   0.027     .0829511    1.352026
        agec |  -.0828798   .0449925    -1.84   0.065    -.1710635    .0053039
             |
smoke#c.agec |
      smoker |   .0730831   .0653439     1.12   0.263    -.0549886    .2011548
             |
       _cons |  -1.101243   .2206746    -4.99   0.000    -1.533758    -.668729
------------------------------------------------------------------------------
```

4.3 Continuous by continuous

Without interaction With only main effects, we assume that the change in OR over the continuous variable $x1$ is the same regardless of $x2$ and vice versa.

With interaction Including an interaction term, we assume that the change in OR over the continuous variable $x1$ differs with respect to $x2$, and vice versa.

Interpretation of Interaction Coefficient The interaction term gives the change in OR over $x1$ for each unit of $x2$, and the change in slope of y over $x2$ for each unit of $x1$. The actual slopes are given by:

slope over $x1$: $e^{\beta_{x1}+x_2\beta_{x1:x2}}$

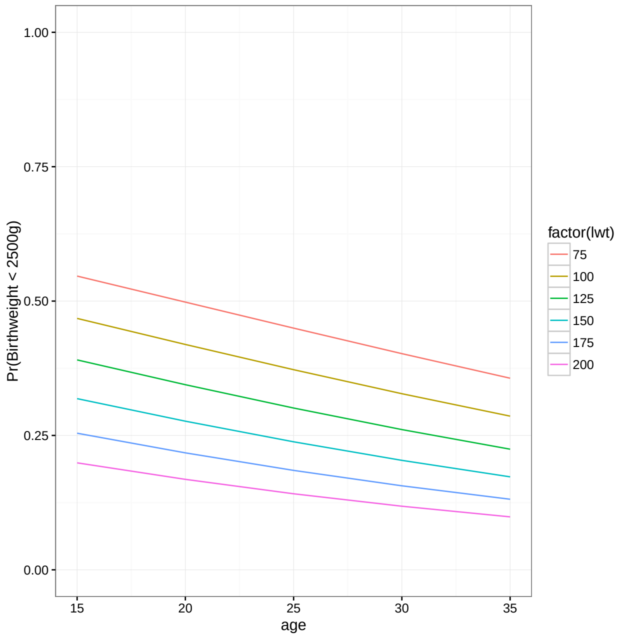
slope over $x2$: $e^{\beta_{x2}+x_1\beta_{x1:x2}}$

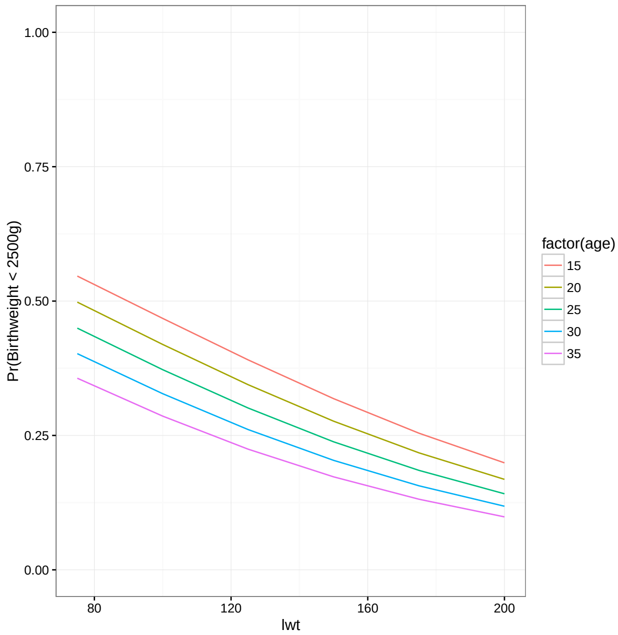
Interpretation The odds ratios considering an interaction between age and weight are *very* slightly lower (99.9% of the odds ratio considering only main effects [99.7 - 1.002%] per year of age and pound in weight), but this difference is not statistically significant.

Tip Plotting predicted odds ratios or probabilities for these models will make the models easier to understand.

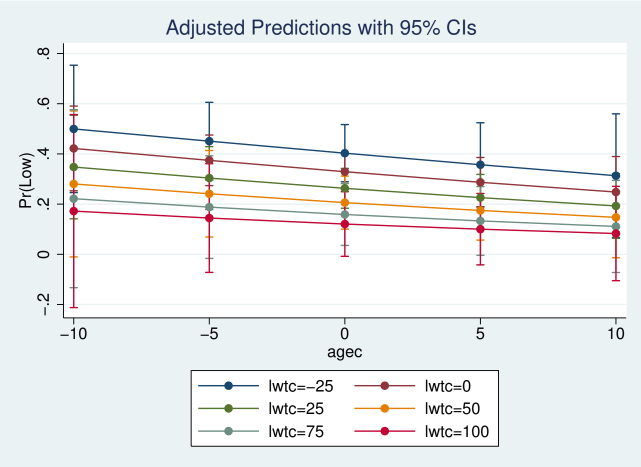
```
m6 <- glm(low ~ agec * lwtc, data = birthwt, family = binomial)
cbind("OR" = exp(coef(m6)), exp(confint(m6)))

## Waiting for profiling to be done...

##                    OR     2.5 %    97.5 %
## (Intercept) 0.4907450 0.3535745 0.6731492
## agec        0.9611041 0.8986580 1.0242901
## lwtc        0.9873092 0.9745634 0.9987438
## agec:lwtc   0.9999823 0.9974573 1.0022256
```

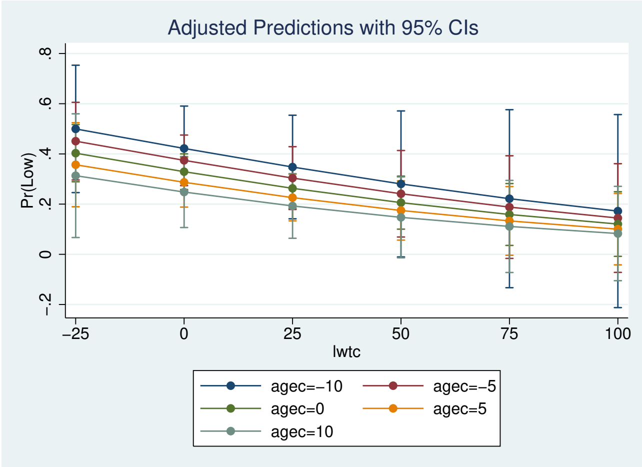
```
nd <- expand.grid(agec = seq(15, 35, 5) - 23, lwtc = seq(75, 200, 25) - 121)
nd$pred <- predict(m6, newdata = nd, type = "response")
nd$age <- nd$agec + 23
nd$lwt <- nd$lwtc + 121
qplot(age, pred, data = nd, color = factor(lwt), geom = "line") +
  ylim(0, 1) + ylab("Pr(Birthweight < 2500g)")
qplot(lwt, pred, data = nd, color = factor(age), geom = "line") +
  ylim(0, 1) + ylab("Pr(Birthweight < 2500g)")
```






```
. logistic low c.agec##c.lwtc, coef

Logistic regression                             Number of obs     =        189
                                                LR chi2(3)        =       7.53
                                                Prob > chi2       =     0.0567
Log likelihood = -113.56918                     Pseudo R2         =     0.0321

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        agec |  -.0396806    .033211    -1.19   0.232     -.104773    .0254118
        lwtc |  -.0127505   .0062141    -2.05   0.040    -.0249298   -.0005711
             |
      c.agec#|
      c.lwtc |  -.0000202   .0011979    -0.02   0.987     -.002368    .0023275
             |
       _cons |  -.7117894   .1638335    -4.34   0.000    -1.032897   -.3906816
------------------------------------------------------------------------------

. quietly: margins, at(agec = (-10(5)10) lwtc = (-25(25)100))

. marginsplot

  Variables that uniquely identify margins: agec lwtc

. marginsplot, xdim(lwtc)

  Variables that uniquely identify margins: agec lwtc
```





Versions

1.0 Original version

R version and packages used to generate this report

R version: R version 3.4.0 (2017-04-21)
Base packages: stats, graphics, grDevices, utils, datasets, methods, base
Other packages: MASS, ggplot2, knitr
This document was generated on 2017-06-22 at 13:35.

A Comparison of R and STATA Datasets

There are a few small differences in `lwt` and `bwt` between the two versions of the dataset we use here, which led to slight differences in the model results.

```
data.r <- birthwt[, c("age", "lwt", "bwt", "race", "smoke")]
data.stata <- read.csv("lbw_stata.csv")

summary(data.r$age - data.stata$age)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0       0       0       0

summary(data.r$lwt - data.stata$lwt)

##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -1.000000  0.000000  0.000000 -0.005291  0.000000  0.000000

data.r[data.r$lwt - data.stata$lwt != 0, ]

##    age lwt  bwt  race      smoke
## 76  20 105 2450 other non-smoker

data.stata[data.r$lwt - data.stata$lwt != 0, ]

##     age lwt  bwt  race     smoke
## 182  20 106 2450 other nonsmoker

summary(data.r$bwt - data.stata$bwt)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -14.0000   0.0000   0.0000   0.3016   0.0000  69.0000

data.r[data.r$bwt - data.stata$bwt != 0, "bwt"]

## [1] 2751 3062 3062 3544 2410

data.stata[data.r$bwt - data.stata$bwt != 0, "bwt"]

## [1] 2750 3076 3076 3475 2395

table(data.r$race, data.stata$race, useNA = "ifany")

##
##         black other white
##   white     0     0    96
##   black    26     0     0
##   other     0    67     0

table(data.r$smoke, data.stata$smoke, useNA = "ifany")

##
##              nonsmoker smoker
##   non-smoker       115      0
##   smoker             0     74
```