Lab8

Zeren Li

10/31/2019

Roadmap

- Dummy variable & Categortical variable
- Interaction effect
- Marginal Effect

Teaching Rating Dataset

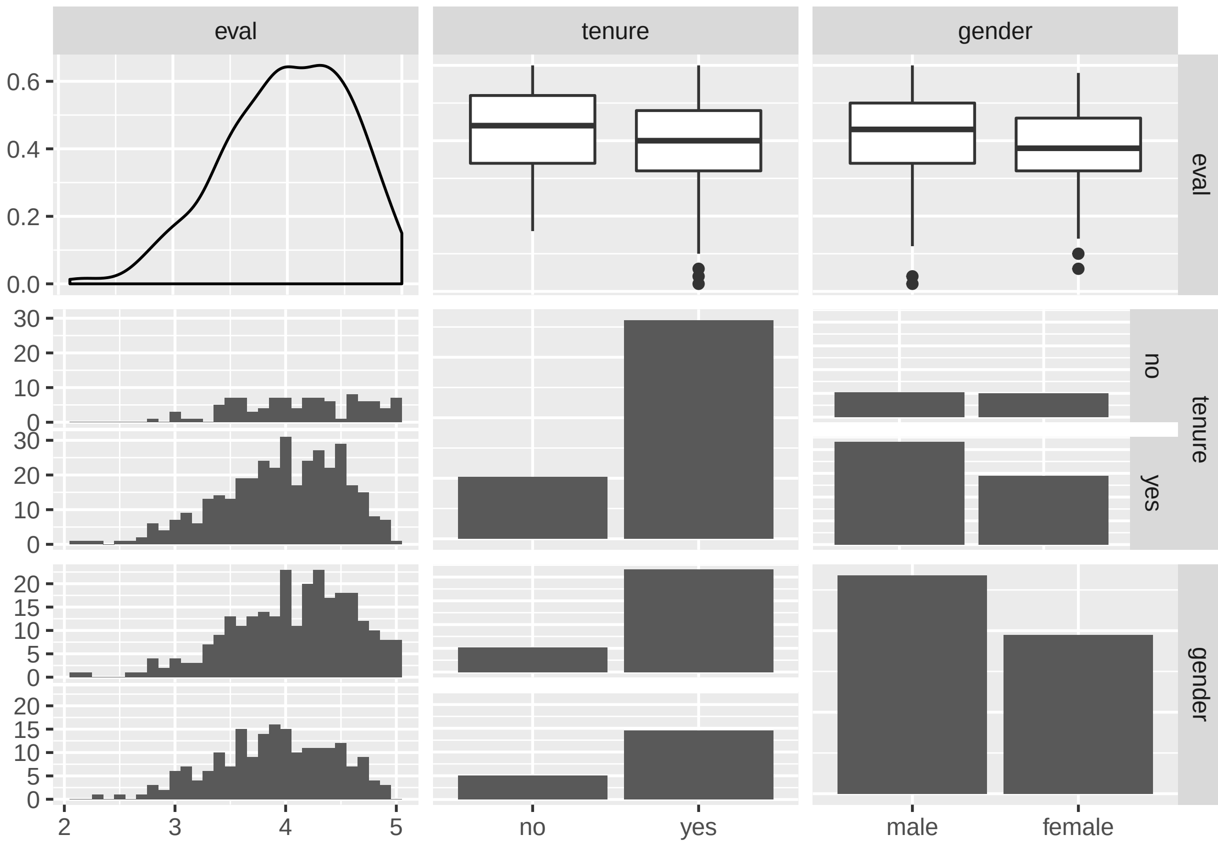
This is a dataset on course evaluations, course characteristics, and professor characteristics for 463 courses for the academic years 2000–2002 at the University of Texas at Austin.

```
data(data = "TeachingRatings", package = "AER")

tr1 <- TeachingRatings %>%
  mutate( female = if_else(gender == "female",1,0),
          male = if_else(gender == "male",1,0) ) %>%
  mutate_at(.vars = c("gender", "minority", "tenure"), .funs = function(x){as.factor(x)})
```

Summary Statistics

```
tr1 %>%
  dplyr::select(eval, tenure, gender) %>%
  as.data.frame() %>%
  ggpairs()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```r
stargazer::stargazer(tr1, header = F)
```

Table 1:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	463	48.365	9.803	29	42	57	73
beauty	463	0.00000	0.789	−1.450	−0.656	0.546	1.970
eval	463	3.998	0.555	2.100	3.600	4.400	5.000
students	463	36.624	45.018	5	15	40	380
allstudents	463	55.177	75.073	8	19	60	581
female	463	0.421	0.494	0	0	1	1
male	463	0.579	0.494	0	0	1	1

Fit a simple model

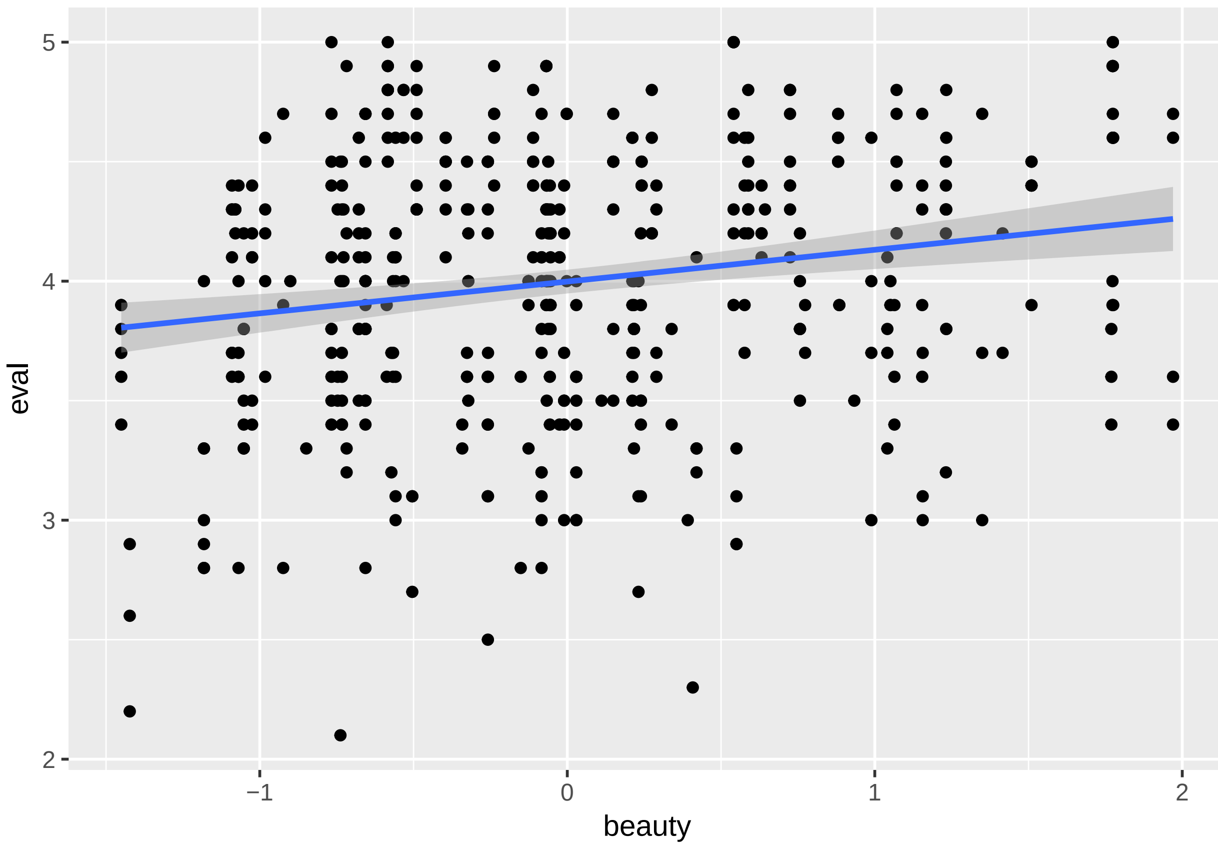
$$eval = \beta_0 + \beta_1 beauty + e$$

```r
m1 <- lm(eval ~ beauty  , tr1 )

summary(m1)
```

```
##
## Call:
## lm(formula = eval ~ beauty, data = tr1)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.80015 -0.36304  0.07254  0.40207  1.10373
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.99827    0.02535 157.727  < 2e-16 ***
## beauty       0.13300    0.03218   4.133 4.25e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5455 on 461 degrees of freedom
## Multiple R-squared:  0.03574,    Adjusted R-squared:  0.03364
## F-statistic: 17.08 on 1 and 461 DF,  p-value: 4.247e-05
```

```
y_hat <- predict(m1)

ggplot(tr1, aes(x = beauty, y = eval)) +
  geom_point() +
  geom_smooth(method = "lm" )
```



Adding dummy variable(s)

$$eval = \beta_0 + \beta_1 beauty + \delta gender + e$$

We can interpret the coefficients as follows:

β_0: the intercept, or the predicted outcome when beauty=0 and gender=0.

β_1: the slope (or effect) of beauty; for a one-unit change in beauty, the predicted change in eval, all else being equal.

δ: the slope (or effect) of gender; for a one-unit change in gender, the predicted change in eval, all else being equal.

Think the substantial meaning of it:

Difference between the predicted value of `eval` for females and that for males, all else being equal.

Dummies here are best thought of as shifts in the constant. Thus, the new intercept for the female line is the coefficient on female plus the original intercept.

$$E(eval) = \begin{cases} (\beta_0 + \delta) + \beta_1 beauty., & \text{gender} = 1 \\ \beta_0 + \beta_1 beauty., & \text{gender} = 0 \end{cases}$$

```r
m2 <- lm(eval ~ beauty + gender , tr1 )

# you cannot include both feamle and male!!!
lm(eval ~ beauty + female , tr1 )
```

```
##
## Call:
## lm(formula = eval ~ beauty + female, data = tr1)
##
## Coefficients:
## (Intercept)       beauty       female
##      4.0816       0.1486      -0.1978
```

```r
lm(eval ~ beauty + male , tr1 )
```

```
##
## Call:
## lm(formula = eval ~ beauty + male, data = tr1)
##
## Coefficients:
## (Intercept)       beauty         male
##      3.8838       0.1486       0.1978
```

```r
lm(eval ~ beauty + female + male , tr1 )
```

```
##
## Call:
## lm(formula = eval ~ beauty + female + male, data = tr1)
##
## Coefficients:
## (Intercept)       beauty       female         male
##      4.0816       0.1486      -0.1978           NA
```

```r
# coefficients
m2$coefficients
```

```
##  (Intercept)       beauty genderfemale
##    4.0815829    0.1485876   -0.1978096
```

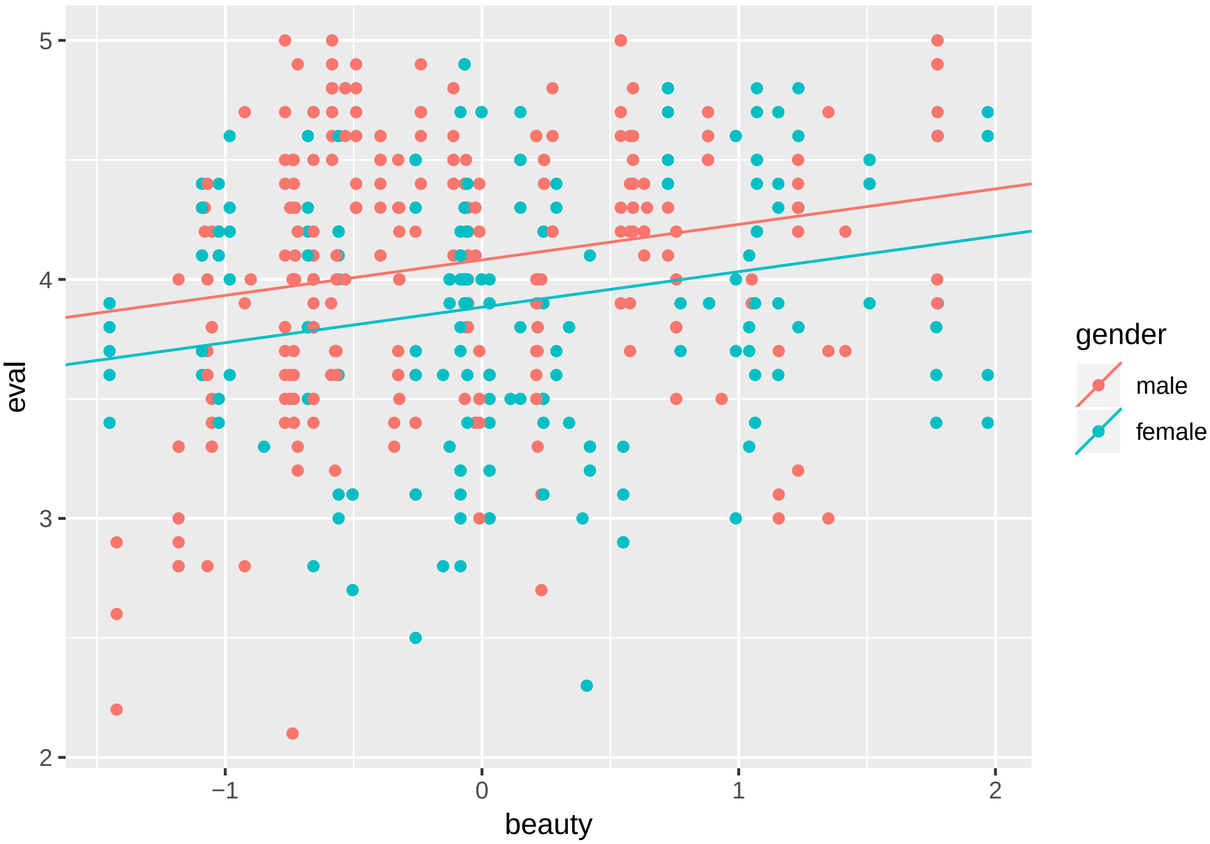
```r
# male intercept
coef(m2)["(Intercept)"]
```

```
## (Intercept)
##    4.081583
```

```r
# female intercept
coef(m2)["(Intercept)"] + coef(m2)["genderfemale"]
```

```
## (Intercept)
##    3.883773
```

Visulization



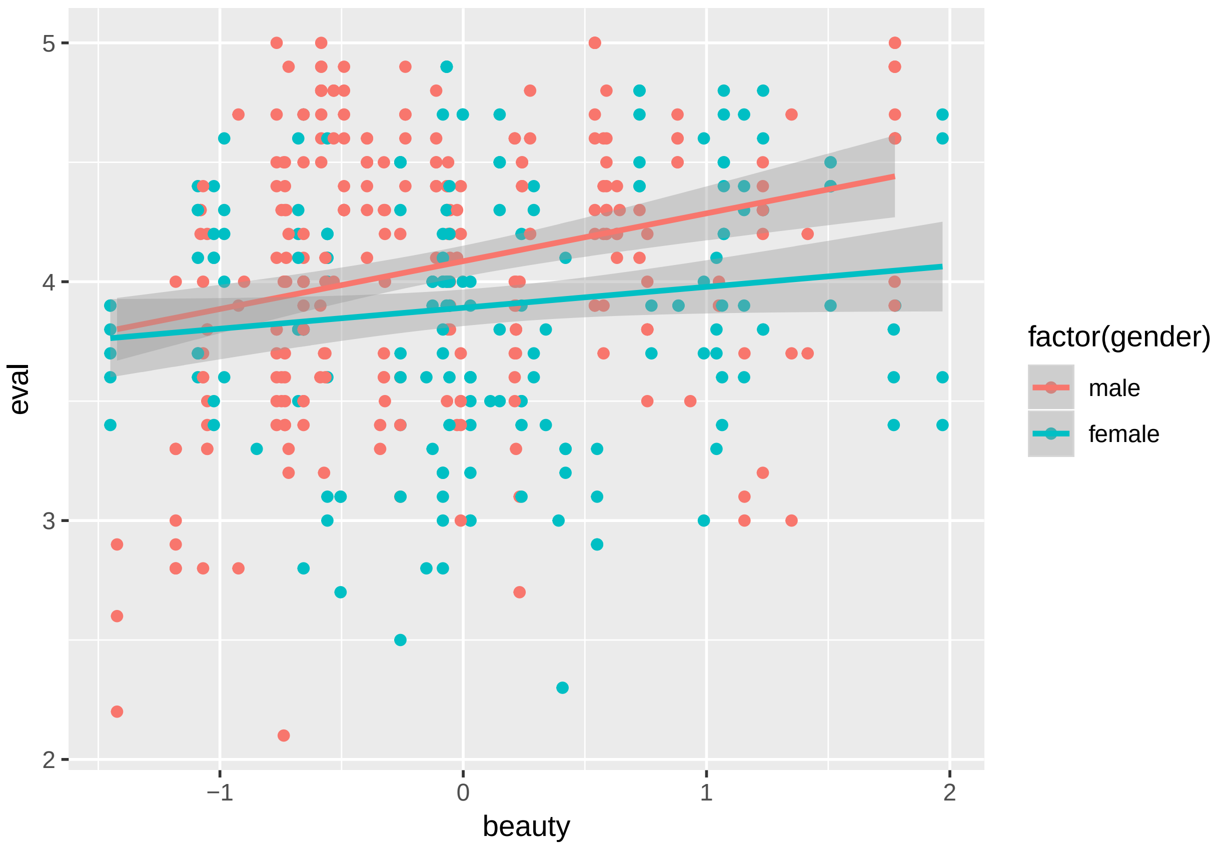
Notice the different intercepts we predicted, but the slopes are the **same**.

Two sub-sample visualization

The visualization using two subsample analyses is shown as follows. BUT WE DON'T RECOMMEND THIS! If we are interested in how the beauty effect on course evaluation changes across different gender groups, we should use interaction analysis.

```
##
## Call:
## lm(formula = eval ~ beauty, data = .)
##
## Coefficients:
## (Intercept)       beauty
##     3.89085      0.08762
```

```
##
## Call:
## lm(formula = eval ~ beauty, data = .)
##
## Coefficients:
## (Intercept)       beauty
##      4.0859       0.2003
```



Two dummies

Now, what if I had two dummies in this regression model. Let's add another dummy variable: `tenure` that is coded 1 if tenured and 0 if untenured.

```
m3 <- lm(eval ~ beauty + gender + tenure, tr1)
coef(m3)
```

```
##  (Intercept)       beauty genderfemale    tenureyes
##    4.2317159    0.1476121   -0.2092403   -0.1863785
```

Intercepts

```
# For male without tenure (female = 0, tenure = 0)
coef(m3)["(Intercept)"]
```

```
## (Intercept)
##    4.231716
```

```r
#For male with tenure (female = 0, tenure = 1)
coef(m3)["(Intercept)"] + coef(m3)["tenureyes"]
```

```
## (Intercept)
##    4.045337
```

```r
#For female without tenure (female = 1, tenure = 0)
coef(m3)["(Intercept)"] + coef(m3)["genderfemale"]
```

```
## (Intercept)
##    4.022476
```

```r
# For female with tenure (female = 1, tenure = 1)
coef(m3)["(Intercept)"] + coef(m3)["genderfemale"]
```

```
## (Intercept)
##    4.022476
```

Multiple categorical variables

Create age group

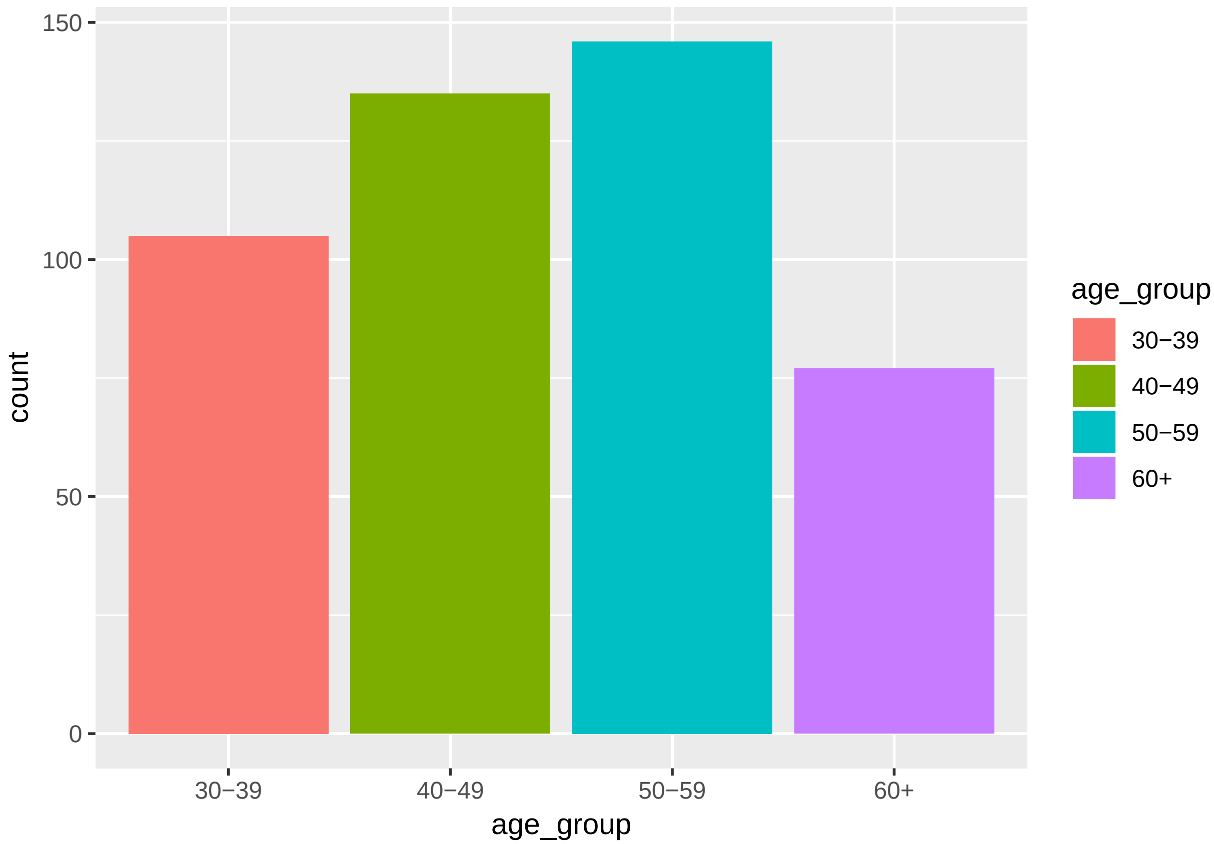
Here, we create a categorical variable named `age_group`. If a professor is in 30-39, then we code `age_group` as 1; If 40-49, we code `age_group` as 2, If 50-59, we code `age_group` as 3; If 60 and above, then 4.

```r
tr1 <- tr1 %>%
  mutate(age_group  = if_else(age <40 ,1 ,
                      if_else(age <50 ,2,
                      if_else(age <60 ,3, 4))) %>% as.factor())
levels(tr1$age_group) <- c("30-39", "40-49", "50-59","60+" )
# alternative way of constructing categorical variable
tr1 %>%
  mutate(age_group40s = if_else(age <= 40 & age < 50,1,0 ),
         age_group50s = if_else(age <= 50 & age < 60,1,0 ),
         age_group60plus = if_else(age > 60,1,0 ))  %>%
  dplyr::select(age_group40s,age_group50s, age_group60plus)  %>%
  head()
```

```
##   age_group40s age_group50s age_group60plus
## 1            1            1               0
## 2            0            0               0
## 3            0            0               0
## 4            1            1               0
## 5            1            1               0
## 6            0            0               1
```

Visualizing the distribution

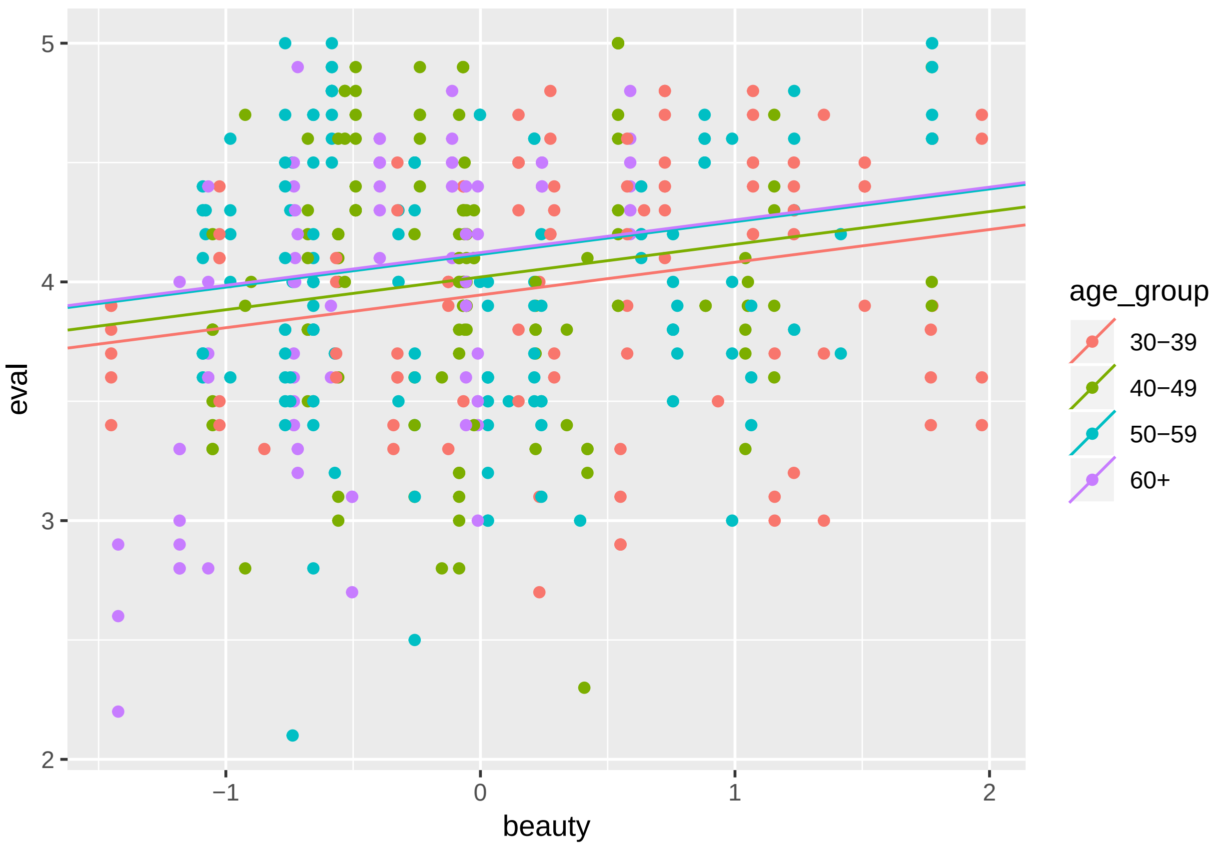
```r
ggplot(tr1, aes(x = age_group, fill = age_group )) +
  geom_bar()
```




```r
m_cat <- lm(eval ~ age_group + beauty, tr1)
summary(m_cat)
```

```
##
## Call:
## lm(formula = eval ~ age_group + beauty, data = tr1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83865 -0.36622  0.05647  0.41053  1.06535
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.945220   0.055096  71.606   <2e-16 ***
## age_group40-49 0.075308   0.072650   1.037    0.300
## age_group50-59 0.094569   0.071528   1.322    0.187
## age_group60+   0.007653   0.086724   0.088    0.930
## beauty          0.137096   0.034162   4.013    7e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5457 on 458 degrees of freedom
## Multiple R-squared:  0.04113,    Adjusted R-squared:  0.03276
## F-statistic: 4.912 on 4 and 458 DF,  p-value: 0.0006924
```

Visualization



Interaction Analysis

Marginal Effect

Marginal effects tell us how a dependent variable (outcome) changes when a specific independent variable (explanatory variable) changes. Other covariates are assumed to be held constant. Marginal effects are often calculated when analyzing regression analysis results.

$$\frac{\partial \hat{y}}{\partial x_k} = \frac{\partial \beta_1 x_1 + \beta_2 x_2 + ...\beta_i x_i}{\partial x_k}, k \in [1, i]$$

For example, we have this model:

$$\hat{eval}_i = \hat{\beta}_0 + \hat{\beta}_1 beauty_i + \hat{\beta}_2 age + u$$

The marginal effect of `beauty` and `age` on `eval`?

$$\frac{\partial \hat{eval}}{\partial beauty} = \hat{\beta}_1$$

$$\frac{\partial \hat{eval}}{\partial age} = \hat{\beta}_2$$

Interpretation: One unit increase in `beauty` is associated with $\hat{\beta}_1$ units change in `eval`, **holding all else being equal**.

Interacting a dummy variable with a dummy variable

$$\hat{eval}_i = \hat{\beta}_0 + \hat{\beta}_1 female_i + \hat{\beta}_2 tenure_i + \hat{\beta}_3 female_i * tenure_i + u$$

```r
# two equaviliant way of interaction
m_inter1 <- lm(eval ~ gender*tenure,  tr1)
summary(m_inter1)
```

```
##
## Call:
## lm(formula = eval ~ gender * tenure, data = tr1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.89028 -0.36000  0.00972  0.40972  1.00972
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               4.39615    0.07438  59.107  < 2e-16 ***
## genderfemale             -0.53615    0.10623  -5.047 6.48e-07 ***
## tenureyes                -0.40588    0.08285  -4.899 1.34e-06 ***
## genderfemale:tenureyes  0.46105    0.12083   3.816 0.000154 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5363 on 459 degrees of freedom
## Multiple R-squared:  0.07173,    Adjusted R-squared:  0.06567
## F-statistic: 11.82 on 3 and 459 DF,  p-value: 1.795e-07
```

```r
m_inter2 <- lm(eval ~  gender + tenure + gender:tenure ,  tr1)
summary(m_inter2)
```

```
##
## Call:
## lm(formula = eval ~ gender + tenure + gender:tenure, data = tr1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.89028 -0.36000  0.00972  0.40972  1.00972
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               4.39615    0.07438  59.107  < 2e-16 ***
## genderfemale             -0.53615    0.10623  -5.047 6.48e-07 ***
## tenureyes                -0.40588    0.08285  -4.899 1.34e-06 ***
## genderfemale:tenureyes  0.46105    0.12083   3.816 0.000154 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5363 on 459 degrees of freedom
## Multiple R-squared:  0.07173,    Adjusted R-squared:  0.06567
```

```
## F-statistic: 11.82 on 3 and 459 DF,  p-value: 1.795e-07
```

$$\hat{eval}_i = \hat{\beta}_0(4.396) + \hat{\beta}_1(-0.536)female_i + \hat{\beta}_2(-0.406)tenure_i + \hat{\beta}_3(0.461)female_i \times tenure_i + u$$

Marginal Effect of `female`

$$\frac{\partial \hat{eval}}{\partial female} = \hat{\beta}_1 + \hat{\beta}_3 \times tenure = -0.536 + 0.461 \times tenure$$

Question: What's marginal Effect of `feamle` on `eval` when `tenure` $= 1$? Hint: Insert one into the above equation.

Marginal Effect of `tenure`

$$\frac{\partial \hat{eval}}{\partial tenure} = \hat{\beta}_2 + \hat{\beta}_3 \times tenure = -0.406 + 0.461 \times gender$$

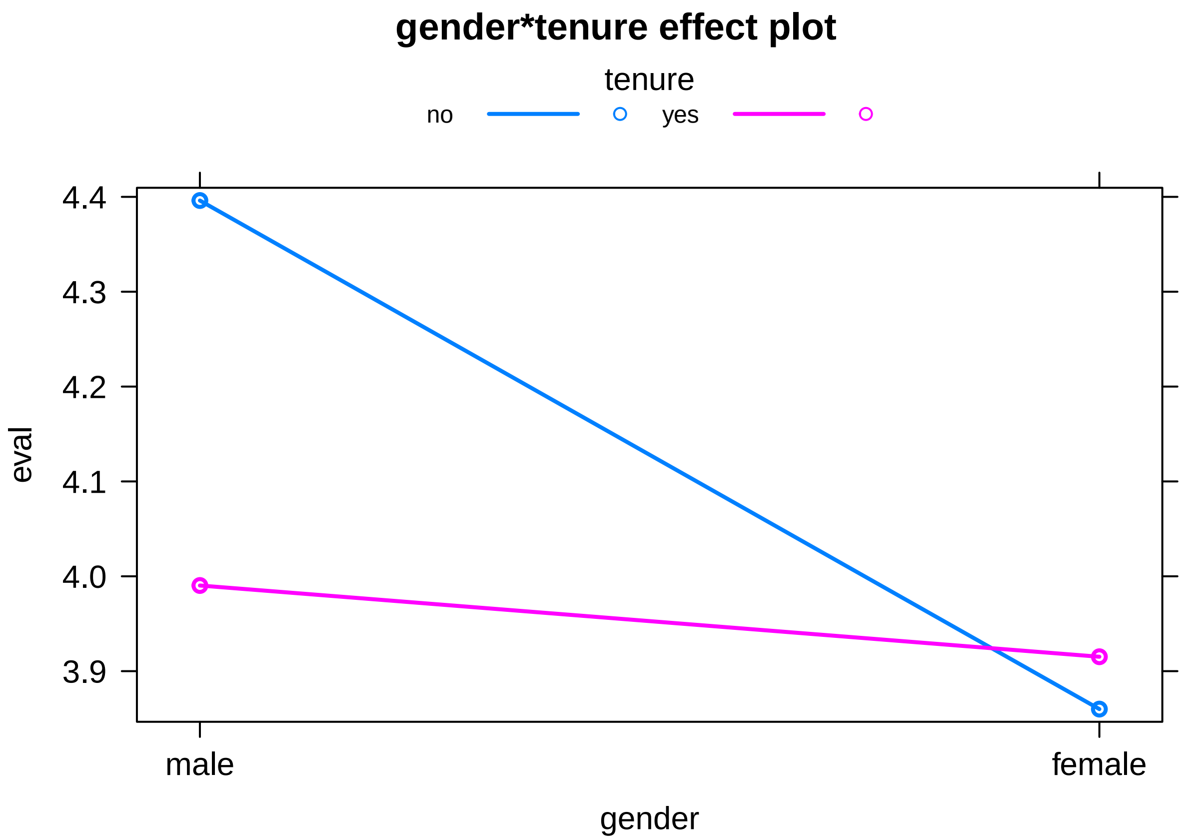
Substantive Effects of Dummy Interactions (Predicted Value)

	untenured	tenured
male	$\hat{\beta}_0 = 4.40$	$\hat{\beta}_0 + \hat{\beta}_2 = 4.40 + (-0.41) = 3.99$
female	$\hat{\beta}_0 + \hat{\beta}_1 = 4.40 + (-0.54) = 3.86$	$\hat{\beta}_0 + \hat{\beta}_1\hat{\beta}_2 + \hat{\beta}_3 = 4.40 + (-0.5) + (-0.4) + (0.46) = 3.91$

```
(eff = Effect(c("gender", "tenure"), mod=m_inter2))
```

```
##
##  gender*tenure effect
##        tenure
## gender        no      yes
##   male   4.396154 3.990278
##   female 3.860000 3.915172
```

```
plot(eff,multiline=TRUE)
```



Interacting a dummy variable with a continuous variable

$$eval_i = \beta_0 + \beta_1 beauty_i + \beta_2 gender + \beta_3 gender_i \times beauty_i + u$$

Fit the model:

```
m_int1 <- lm(eval ~  beauty + gender + gender:beauty ,  tr1)
summary(m_int1)
```

```
##
## Call:
## lm(formula = eval ~ beauty + gender + gender:beauty, data = tr1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83820 -0.37387  0.04551  0.39876  1.06764
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.08595    0.03295 123.999  < 2e-16 ***
## beauty                0.20027    0.04333   4.622 4.95e-06 ***
## genderfemale         -0.19510    0.05089  -3.834 0.000144 ***
## beauty:genderfemale -0.11266    0.06398  -1.761 0.078910 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5361 on 459 degrees of freedom
```

```
## Multiple R-squared:  0.07256,    Adjusted R-squared:  0.0665
## F-statistic: 11.97 on 3 and 459 DF,  p-value: 1.47e-07
```

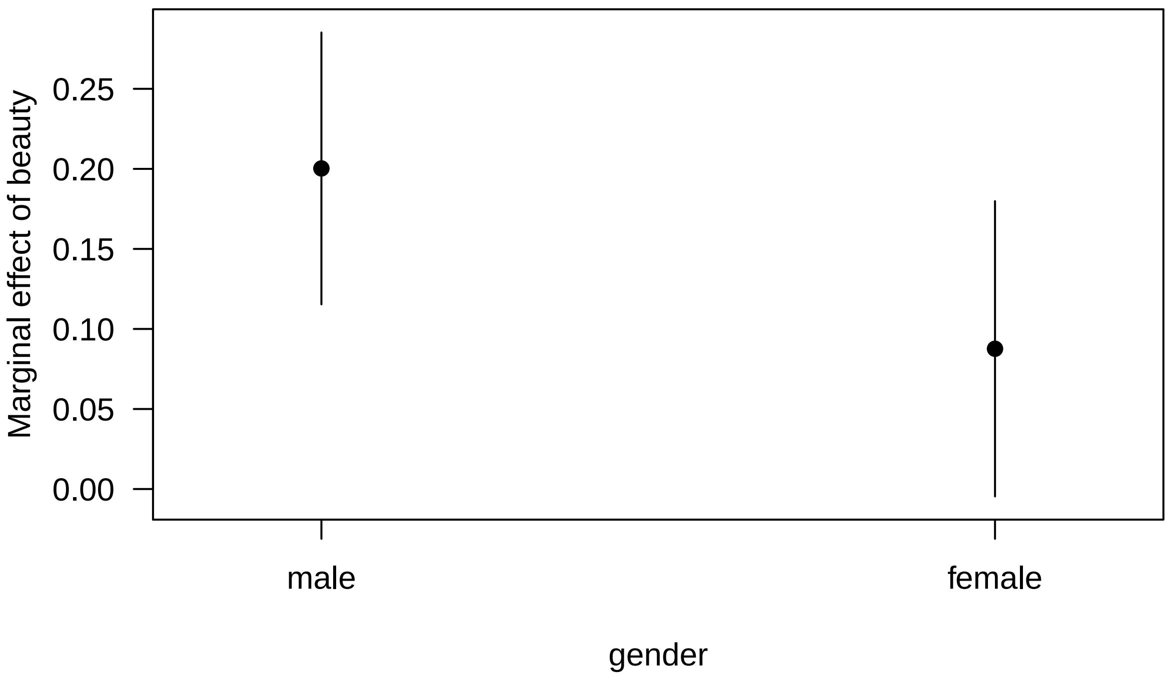
$$\hat{eval_i} = \hat{\beta}_0(4.08595) + \hat{\beta}_1(0.20027)beauty_i + \hat{\beta}_2(-0.19510)gender + \hat{\beta}_3(-0.11266)gender_i \times beauty_i + \hat{u}$$

The marginal effect of `beauty` on `eval`

$$\frac{\partial \hat{eval}}{\partial beauty} = \hat{\beta}_1 + \hat{\beta}_3 \times gender = 0.20027 - 0.11266 \times gender$$

Visualization

```r
# marginal effect
cplot(m_int1, x= "gender", dx = "beauty", what = "effect")
```

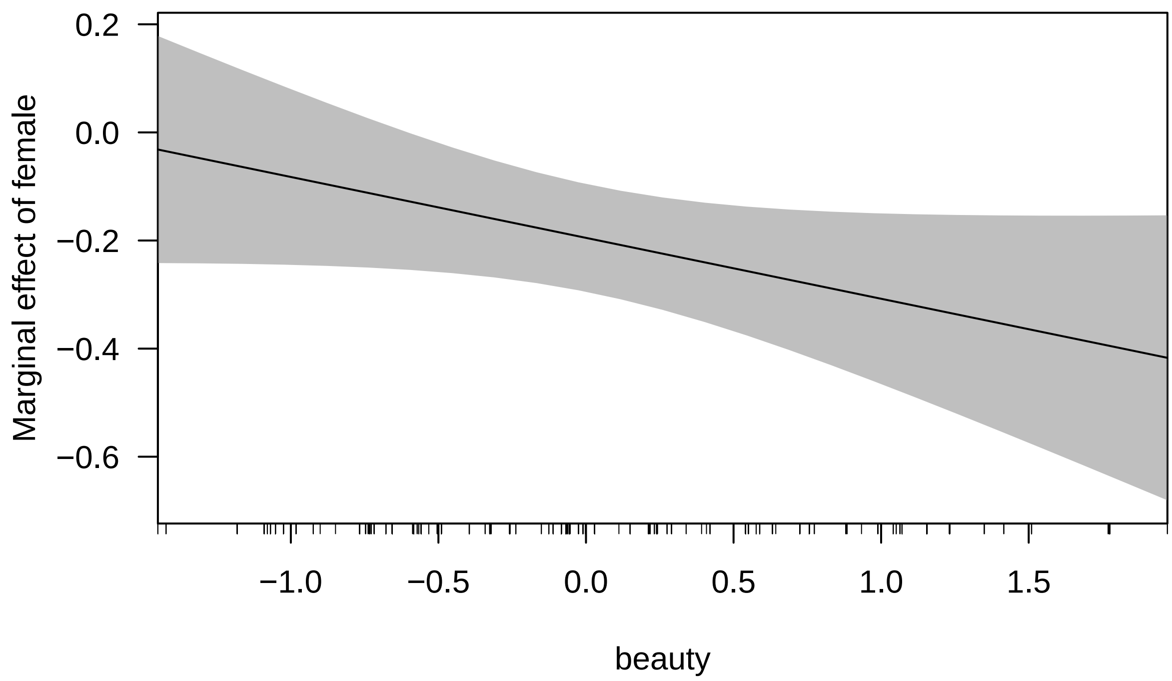


Answer: One unit increase in `beauty` is associated with $\hat{\beta}_3 + \hat{\beta}_5 \times gender$ units increase of `eval`.

Exercise: the marginal effect of `gender` on `eval`

$$\frac{\partial \hat{eval}}{\partial gender} = \hat{\beta}_2 + \hat{\beta}_3 \times beauty$$

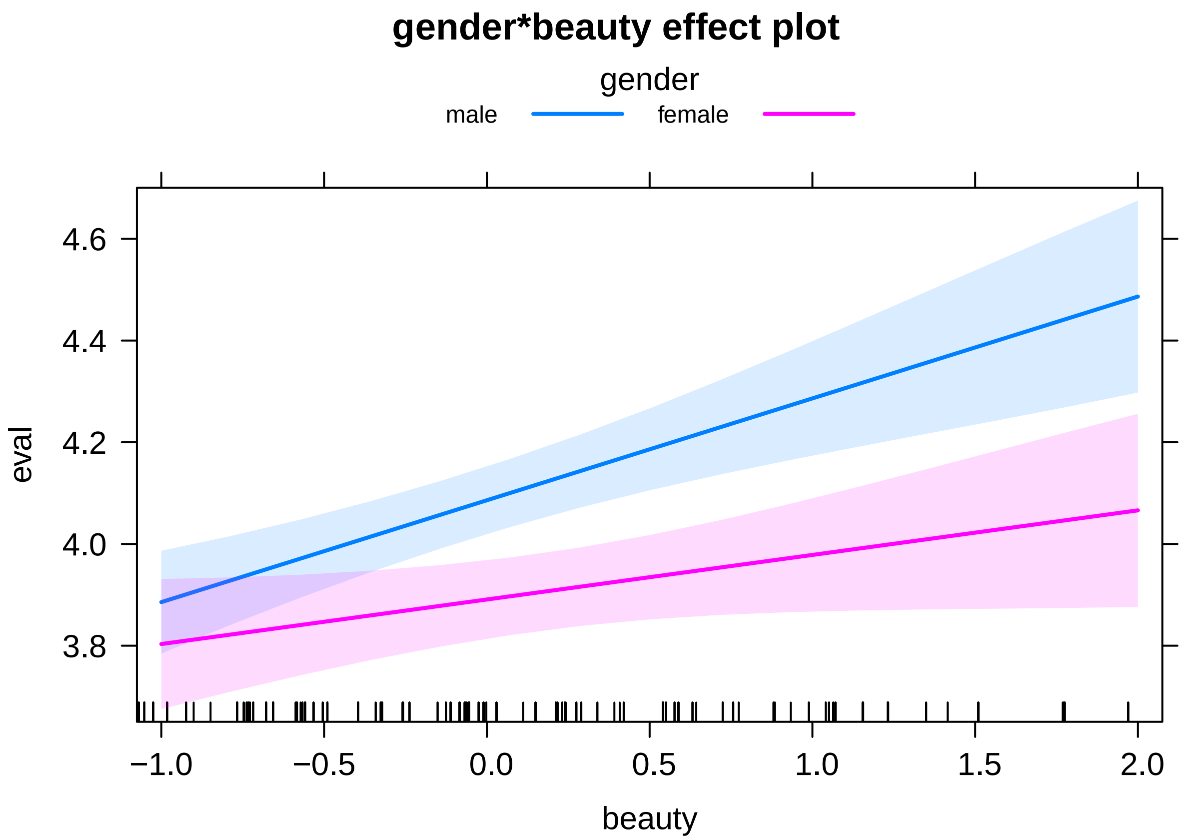
```r
x1 =  lm(eval ~ beauty*female, tr1)

cplot(x1, x = "beauty" ,dx = "female", what = "effect")
```



Predicted Value

The marginal effect of `female` on `eval` across different values of `beauty` is basically the difference between the female fitted line and male fitted line.

```
plot(Effect(c("gender", "beauty"), mod=m_int1, se=TRUE),
x.var = "beauty",
multiline=TRUE, ci.style = 'bands')
```



Interacting a continuous variable with a continuous variable

$$eval_i = \beta_0 + \beta_1 beauty_i + \beta_2 age + \beta_3 beauty_i \times age_i + u$$

The marginal effect of `beauty` on `eval`

```
m_int2 <- lm(eval ~  beauty + age + beauty:age ,  tr1)
summary(m_int2)
```

```
##
## Call:
## lm(formula = eval ~ beauty + age + beauty:age, data = tr1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74828 -0.36705  0.03469  0.41307  1.15642
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.9904777  0.1323781  30.145  < 2e-16 ***
## beauty      -0.3391625  0.1495750  -2.268  0.02382 *
## age          0.0006434  0.0026893   0.239  0.81102
## beauty:age   0.0101498  0.0031271   3.246  0.00126 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.5405 on 459 degrees of freedom
## Multiple R-squared:  0.05739,    Adjusted R-squared:  0.05123
## F-statistic: 9.316 on 3 and 459 DF,  p-value: 5.451e-06
```

Our fitted model is:

$$eval_i = \beta_0(3.9904777) + \beta_1(-0.3391625)beauty_i + \beta_2(0.0006434)age + \beta_3(0.0101498)beauty_i \times age_i + u$$

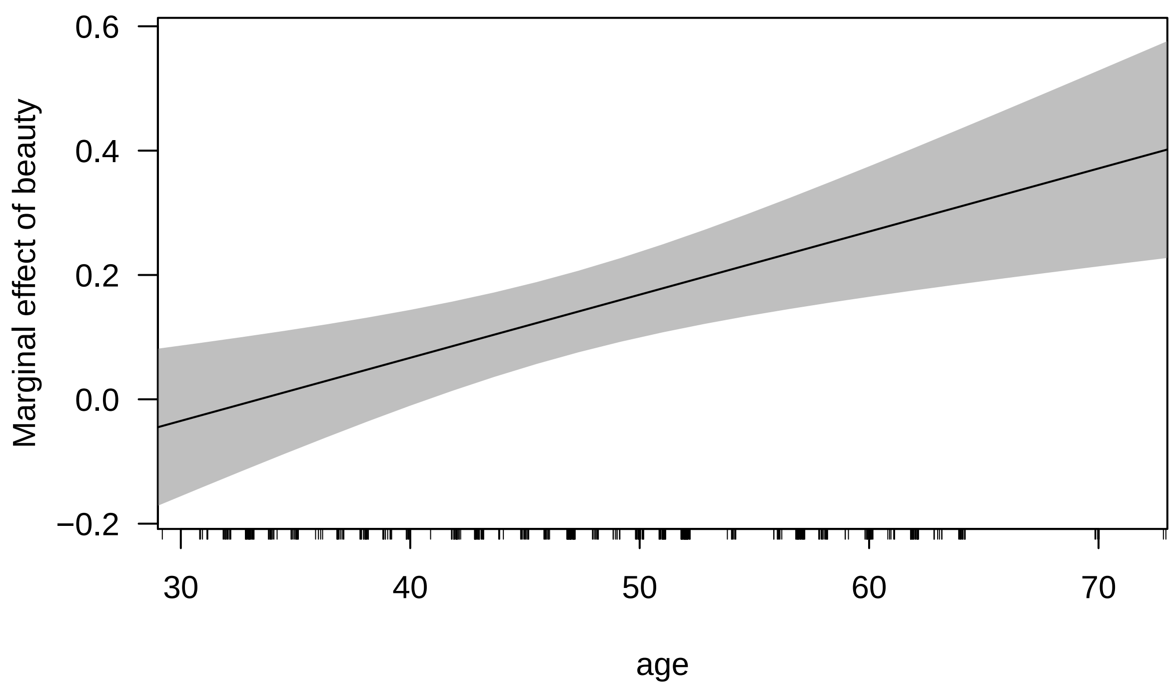
$$\frac{\partial \hat{eval}}{\partial beauty} = \hat{\beta}_1 + \hat{\beta}_3 \times age = -0.3391625 + 0.0101498 \times age$$

Exercise: What's the marginal effect of age

$$\frac{\partial \hat{eval}}{\partial age} =$$

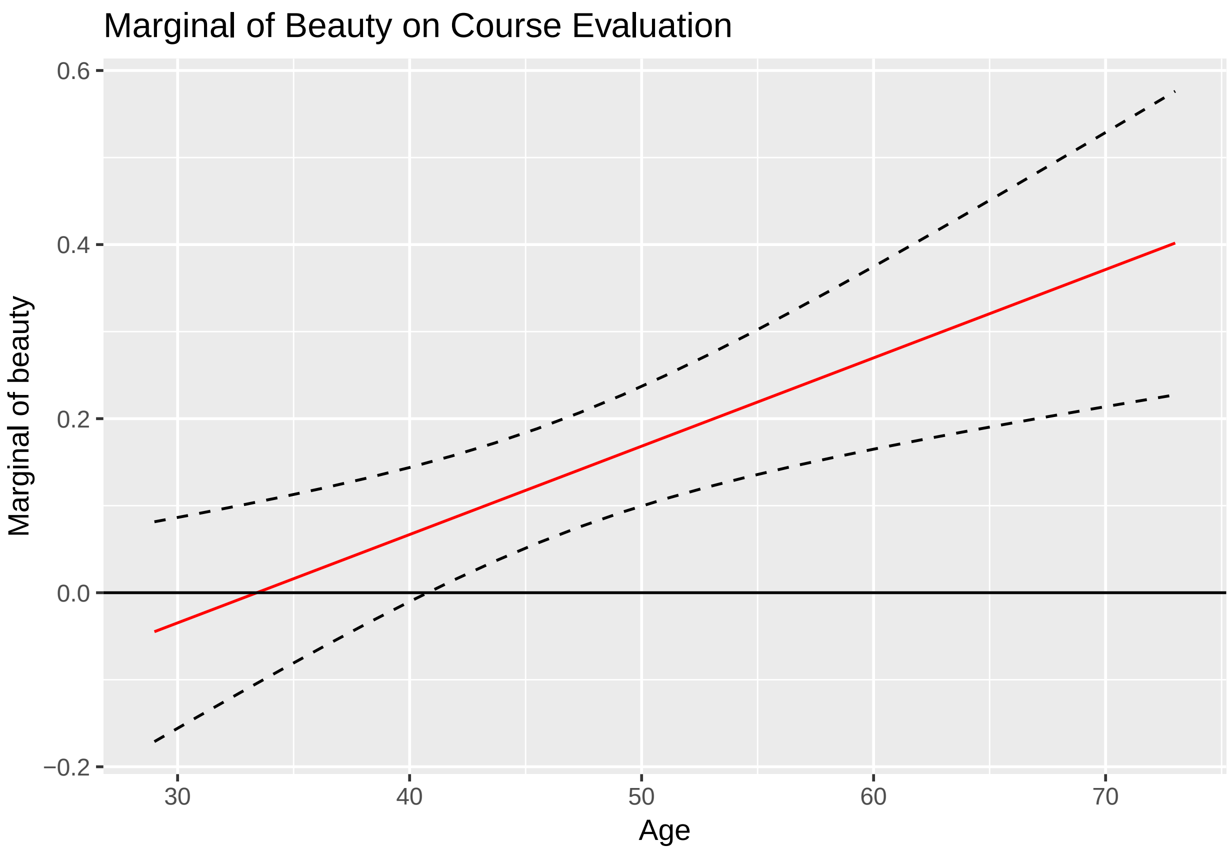
Visualization

```
m_int2 <- lm(eval ~  beauty + age + beauty:age ,  tr1)

# use margins package's cplot function
cplot(m_int2, "age", dx = "beauty", what = "effect")
```




```
# you could add/change feactures using ggplot()
cdat <- cplot(m_int2, "age", dx = "beauty", what = "effect",  draw = FALSE)

ggplot(cdat, aes(x = xvals)) +
  geom_line(aes(y = yvals), color = "red") +
```

```r
  geom_line(aes(y = upper), linetype = 2) +
  geom_line(aes(y = lower), linetype = 2) +
  geom_hline(yintercept = 0) +
  ggtitle("Marginal of Beauty on Course Evaluation") +
  xlab("Age") +
  ylab("Marginal of beauty")
```




```r
ggplot(cdat, aes(x = xvals)) +
  geom_line(aes(y = yvals), color = "red") +
  geom_point(aes(y = yvals)) +
  geom_errorbar(aes(ymin = lower, ymax = upper)) +
  geom_hline(yintercept = 0, type = "dash") +
  ggtitle("Marginal effect of Beauty on  Course Evaluation") +
  xlab("Age") +
  ylab("Marginal effect of beauty")
```

```
## Warning: Ignoring unknown parameters: type
```

If you are interested in how to compute mariginal effect by hand..

Check this: https://rpubs.com/milesdwilliams15/326345