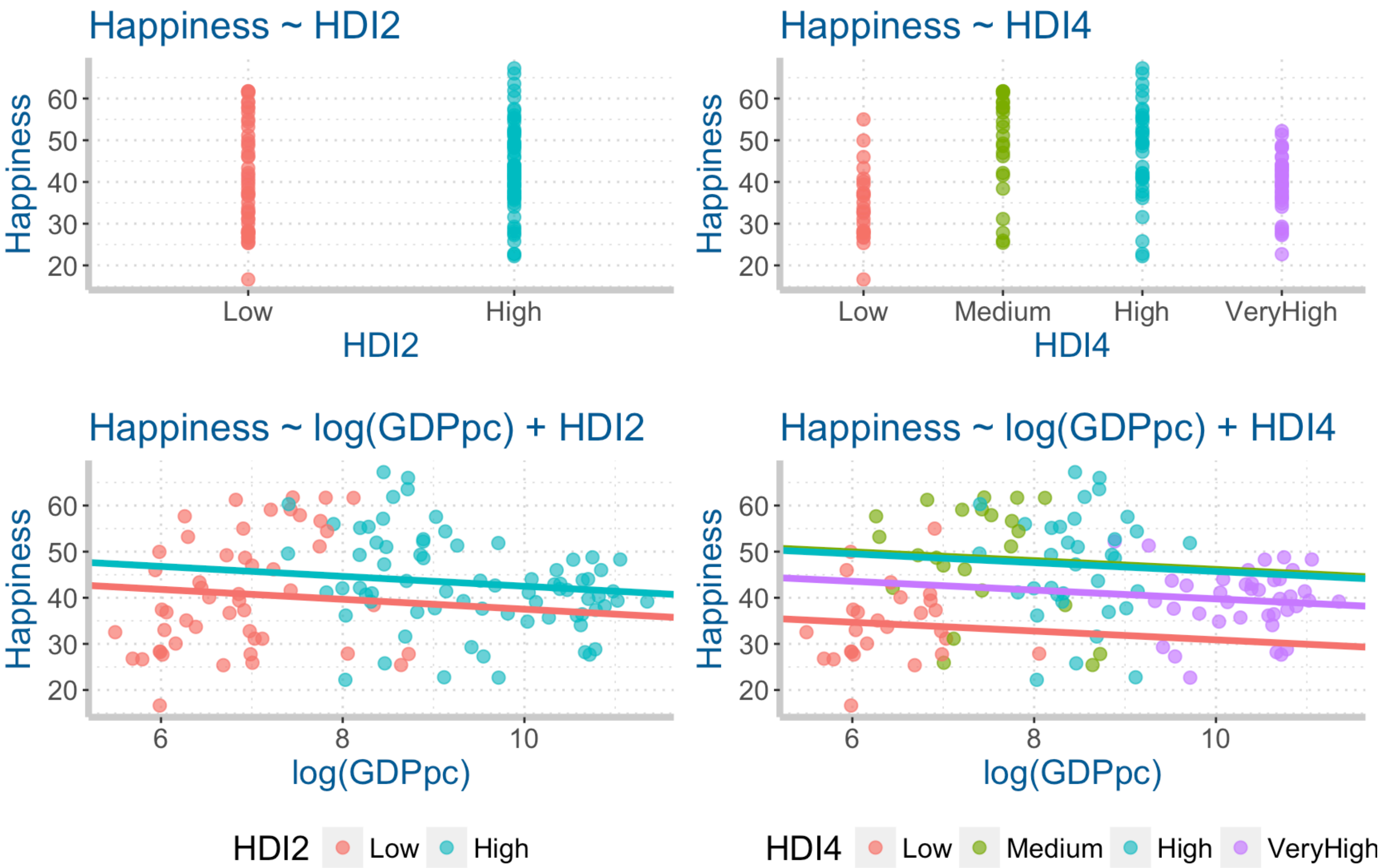
STAT021 Statistical Methods II

Lecture 17 MLR Interaction

Lu Chen
Swarthmore College
11/8/2018

MLR with categorical predictors



Happiness ~ *log*(*GDPpc*) + *HDI2*

```r
m1 <- lm(Happiness ~ log(GDPpc) + HDI2, data=HappyPlanet)
summary(m1)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   48.283      7.265   6.646 9.15e-10 ***
## log(GDPpc)    -1.076      1.035  -1.039    0.301
## HDI2High       4.982      3.363   1.481    0.141
```

$$\widehat{Happiness} = 48.3 - 1.1 \times log(GDPpc) + 5.0 \times HDI2$$

- $HDI2 = 0$ (*Low*) $\Longrightarrow \widehat{Happiness} = 48.3 - 1.1 \times log(GDPpc)$
- $HDI2 = 1$ (*High*) $\Longrightarrow \widehat{Happiness} = 53.3 - 1.1 \times log(GDPpc)$
- The effect of *log(GDPpc)* on *Happiness* ($b_1 = -1.1$) is the same for the two *HDI2* groups.
- The effect of *HDI2* on *Happiness* ($b_2 = 5.0$) is the same for all *log(GDPpc)* values.
- The two regression lines for the *Low* group and the *High* group are parallel.
- **Is this a good fit to the data?**

Happiness ~ *log(GDPpc)*

Fit the model *Happiness* ~ *log(GDPpc)* for the two *HDI2* groups **separately**.

```r
m.low <- lm(Happiness ~ log(GDPpc), subset = (HDI2 == "Low"), data=HappyPlanet)
m.high <- lm(Happiness ~ log(GDPpc), subset = (HDI2 == "High"), data=HappyPlanet)
summary(m.low)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.114     14.928   0.610   0.5444
## log(GDPpc)     4.640      2.164   2.144   0.0371 *
```

- $b_1 = 4.640, P = 0.037$

```r
summary(m.high)
```

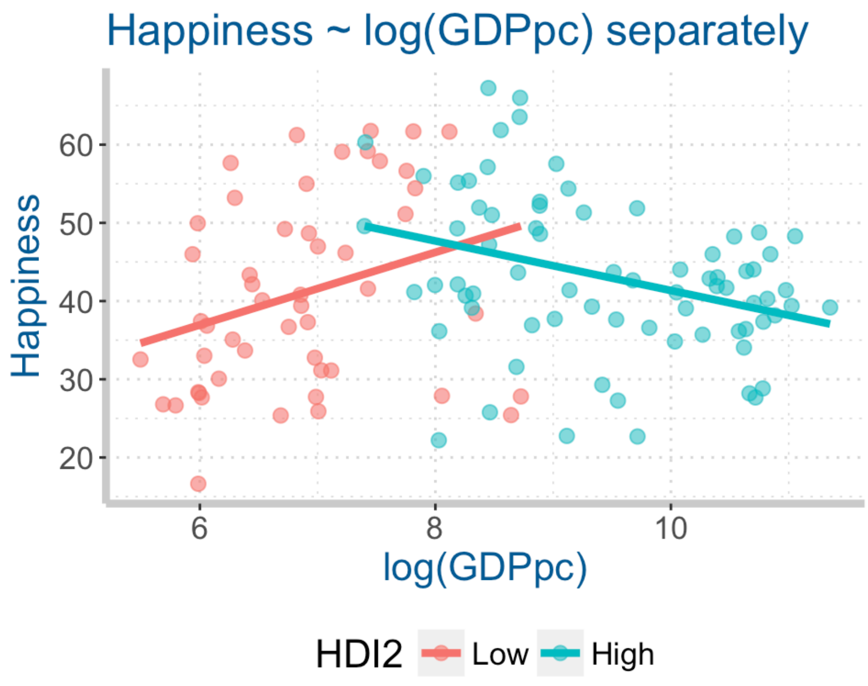
```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   73.037      9.932   7.354 2.45e-10 ***
## log(GDPpc)    -3.168      1.045  -3.033  0.00336 **
```

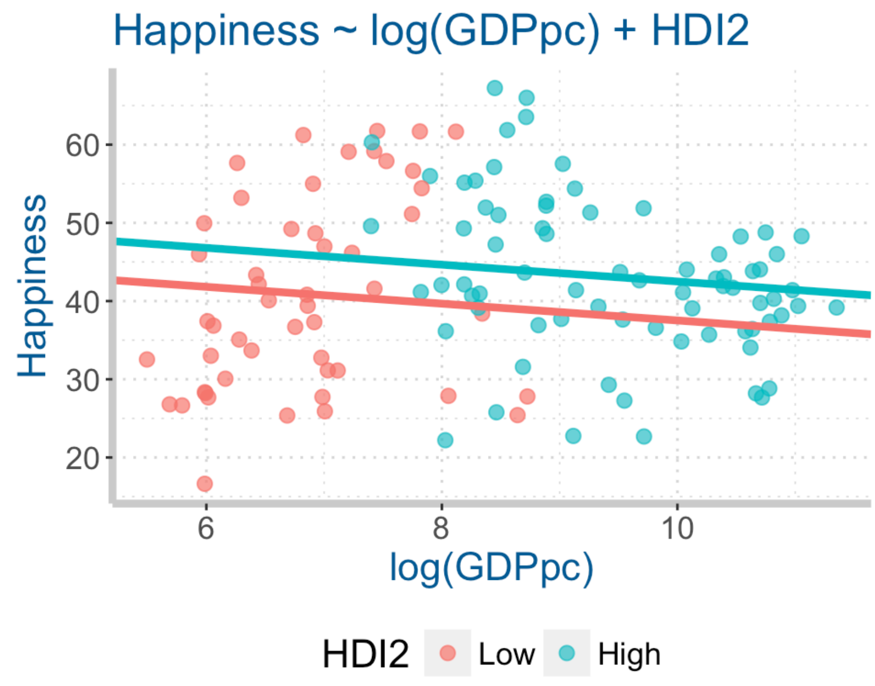
- $b_1 = -3.168, P = 0.003$

- *Low*: Effect of *log(GDPpc)* on *Happiness* is positive and significant.
- *High*: Effect of *log(GDPpc)* on *Happiness* is negative and significant.

Happiness ~ log(GDPpc)

```r
ggplot(HappyPlanet, aes(x=log(GDPpc), y=Happiness, color=HDI2))+
  geom_point(size=2.5, alpha=0.6)+
  geom_smooth(method="lm", se=F, size=1.5)+
  ggtitle("Happiness ~ log(GDPpc) separately")+
  theme(legend.position = 'bottom')
```





*Happiness ~ log(GDPpc) * HDI2*

The model *Happiness ~ log(GDPpc) + HDI2* does NOT fit the data very well. We now consider a model with the **interaction** (product) term of *log(GDPpc)* and *HDI2*.

- **Response variable**: $Happiness\,(Y)$
- **Predictors**: $log(GDPpc)\,(X_1)$ and $HDI2\,(X_2)$
- **Model**:

  $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon,$$

  where $\epsilon \overset{iid}{\sim} N(0, \sigma)$.
- **R**: in R, to specify a model with interaction, we can do either `Happiness ~ log(GDPpc) * HDI2` or `Happiness ~ log(GDPpc) + HDI2 + log(GDPpc):HDI2`.

*Happiness ~ log(GDPpc) * HDI2*

Model without interaction

```r
summary(m1 <- lm(Happiness ~ log(GDPpc) + HDI2, data=HappyPlanet))
```

```
## Multiple R-squared:  0.01828,    Adjusted R-squared:  0.002051
## F-statistic: 1.126 on 2 and 121 DF,  p-value: 0.3276
```

- $F = 1.126, P = 0.3276 > 0.05, R^2 = 0.0183, R^2_{adj} = 0.0021$
- The model is not significant and explains only 1.83% variability.

Model with interaction

```r
summary(m2 <- lm(Happiness ~ log(GDPpc) * HDI2, data=HappyPlanet))
```

```
## Multiple R-squared:  0.1088, Adjusted R-squared:  0.0865
## F-statistic: 4.883 on 3 and 120 DF,  p-value: 0.003077
```

- $F = 4.883, P = 0.003 < 0.05, R^2 = 0.1088, R^2_{adj} = 0.0865$
- The model is significant and explains 10.88% variability.

*Happiness ~ log(GDPpc) * HDI2*

Model without interaction

```r
summary(m1 <- lm(Happiness ~ log(GDPpc) + HDI2, data=HappyPlanet))
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    48.283      7.265   6.646 9.15e-10 ***
## log(GDPpc)     -1.076      1.035  -1.039    0.301
## HDI2High        4.982      3.363   1.481    0.141
```

Model with interaction

```r
summary(m2 <- lm(Happiness ~ log(GDPpc) * HDI2, data=HappyPlanet))
```

```
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)            9.114     13.199   0.690 0.491223
## log(GDPpc)             4.640      1.914   2.425 0.016816 *
## HDI2High              63.923     17.188   3.719 0.000305 ***
## log(GDPpc):HDI2High   -7.808      2.237  -3.491 0.000674 ***
```

- What does a significant interaction term mean?

- Individual t tests for the slopes of *log(GDPpc)* and *HDI2* become significant after adding the interaction term, which is also significant.

Happiness ~ log(GDPpc) * HDI2

$$\widehat{\boldsymbol{Happiness}} = 9.1 + 4.6 \times \boldsymbol{log(GDPpc)} + 63.9 \times HDI2 - 7.8 \times \boldsymbol{log(GDPpc)} \times HDI2$$
$$= [9.1 + 63.9 \times HDI2] + [4.6 - 7.8 \times HDI2] \times \boldsymbol{log(GDPpc)}$$

- This model can be viewed as a regression model of *Happiness* based on *log(GDPpc)*, where
- the intercept is $9.1 + 63.9 \times HDI2$ and the slope is $4.6 - \mathbf{7.8} \times HDI2$
- This model allows the slope of *log(GDPpc)* to vary according to the values of *HDI2*.

  $HDI2 = 0\ (Low) \Longrightarrow \widehat{Happiness} = 9.1 + 4.6 \times log(GDPpc)$

  $HDI2 = 1\ (High) \Longrightarrow \widehat{Happiness} = 73.0 - 3.2 \times log(GDPpc)$
- But the model without interaction does not allow that:

  $HDI2 = 0\ (Low) \Longrightarrow \widehat{Happiness} = 48.3 - 1.1 \times log(GDPpc)$

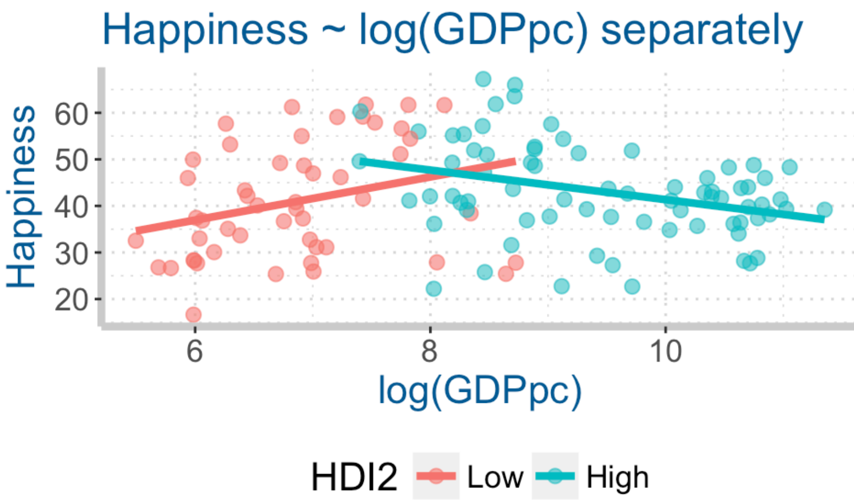
  $HDI2 = 1\ (High) \Longrightarrow \widehat{Happiness} = 53.3 - 1.1 \times log(GDPpc)$

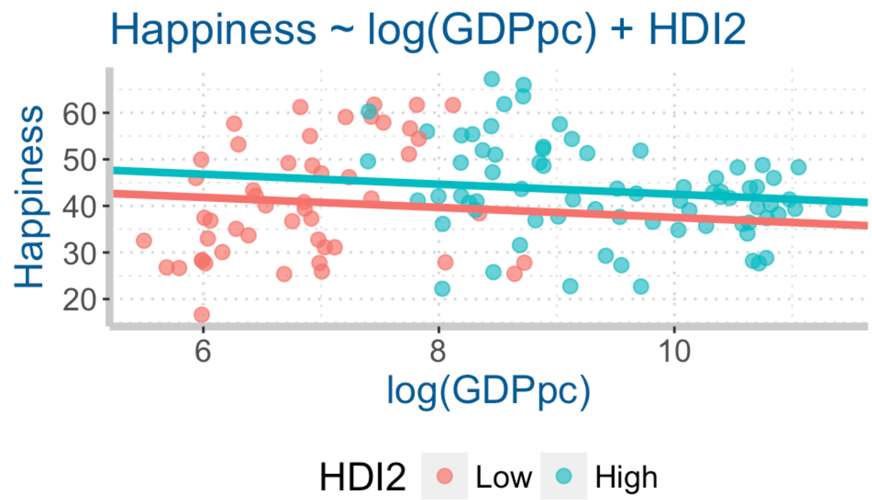
Happiness ~ log(GDPpc) * HDI2

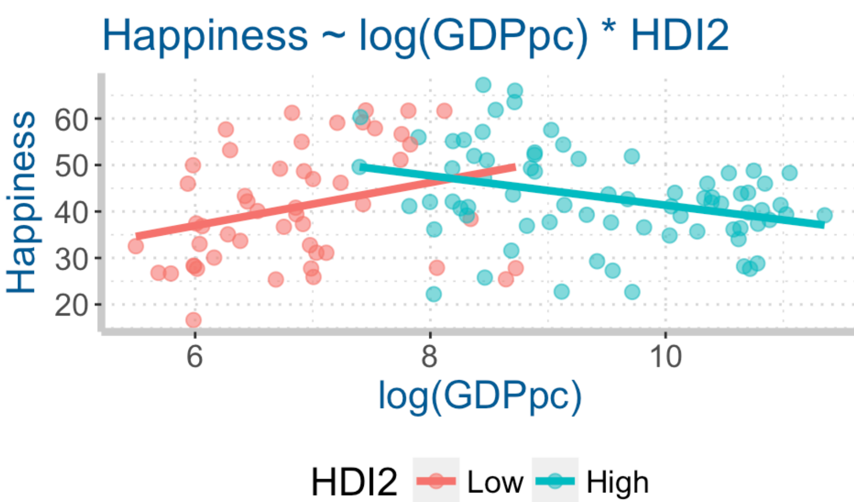
$$\widehat{\boldsymbol{Happiness}} = 9.1 + 4.6 \times \boldsymbol{log(GDPpc)} + 63.9 \times HDI2 - 7.8 \times \boldsymbol{log(GDPpc)} \times HDI2$$
$$= [9.1 + 63.9 \times HDI2] + [4.6 - 7.8 \times HDI2] \times \boldsymbol{log(GDPpc)}$$

- This model can be viewed as a regression model of *Happiness* based on *log(GDPpc)*, where
- the intercept is $9.1 + 63.9 \times HDI2$ and the slope is $4.6 - \mathbf{7.8} \times HDI2$
- This model allows the slope of *log(GDPpc)* to vary according to the values of *HDI2*.
- This model allows the effect of *log(GDPpc)* on *Happiness* to vary according to the values of *HDI2*.
- The interaction slope value -7.8 is the difference of the slope values between the two regression lines and is thus the **difference of differences**.
- The two regression lines calculated based on this model are exactly the same as the two regression lines obtained separately.

Happiness ~ log(GDPpc) * HDI2







- The model with interaction fits the data exactly the same as the separate regression models.
- Q: Then why do we need this complicated model? Why don't we simply run two separate models on the data?

Happiness ~ log(GDPpc) * HDI2

```
summary(m2 <- lm(Happiness ~ log(GDPpc) * HDI2, data=HappyPlanet))
```

```
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)            9.114     13.199   0.690 0.491223
## log(GDPpc)             4.640      1.914   2.425 0.016816 *
## HDI2High              63.923     17.188   3.719 0.000305 ***
## log(GDPpc):HDI2High   -7.808      2.237  -3.491 0.000674 ***
```

- The MLR model with interaction facilitates a test of whether the interaction is significant.

If the interaction is significant,

- the effect of one predictor on the response variable is significantly different for different values of the other predictor.
- the model with it is significantly better than the model without it; we usually keep the interaction term in the model.
- the individual t tests for the main effect terms are usually not interpreted.

Happiness ~ log(GDPpc) * HDI2

```r
summary(m2 <- lm(Happiness ~ log(GDPpc) * HDI2, data=HappyPlanet))
```

```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)              9.114     13.199   0.690 0.491223
## log(GDPpc)               4.640      1.914   2.425 0.016816 *
## HDI2High                63.923     17.188   3.719 0.000305 ***
## log(GDPpc):HDI2High     -7.808      2.237  -3.491 0.000674 ***
```

- The MLR model with interaction facilitates a test of whether the interaction is significant.

If the interaction is NOT significant,

- the effect of one predictor on the response variable does not depend on the values of the other predictor.
- the model with it is NOT better than the model without it; we determine whether to keep it based on other criteria (e.g., adjusted R^2).
- we then check the individual t tests for the main effect terms.

Happiness ~ log(GDPpc) + HDI4

```r
summary(m3 <- lm(Happiness ~ log(GDPpc) + HDI4, data=HappyPlanet))
```

```
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     40.4224    10.0109   4.038 9.59e-05 ***
## log(GDPpc)      -0.9543     1.5369  -0.621 0.535837    
## HDI4Medium      15.2773     3.1413   4.863 3.57e-06 ***
## HDI4High        14.8278     4.0546   3.657 0.000381 ***
## HDI4VeryHigh     8.8817     6.4883   1.369 0.173614    
## 
## Residual standard error: 9.7 on 119 degrees of freedom
## Multiple R-squared:  0.2546, Adjusted R-squared:  0.2295 
## F-statistic: 10.16 on 4 and 119 DF,  p-value: 4.134e-07
```

- $\widehat{Happiness} = 40.4 - 1.0 \times log(GDPpc) + 15.3 \times M + 14.8 \times H + 8.9 \times V$
- Given that *HDI4* is held constant, *log(GDPpc)* is not significant in explaining *Happiness*.
- Ajusted for *log(GDPpc)*, the $Medium$ and the $High$ group have significantly different *Happiness* values from the Low group.

Happiness ~ log(GDPpc) * HDI4

Model without interaction

```r
summary(m3 <- lm(Happiness ~ log(GDPpc) + HDI4, data=HappyPlanet))
```

```
## Multiple R-squared:  0.2546, Adjusted R-squared:  0.2295
## F-statistic: 10.16 on 4 and 119 DF,  p-value: 4.134e-07
```

- $F = 10.16, P = 4.13 \times 10^{-7} < 0.05, R^2 = 0.2546, R^2_{adj} = 0.2295$
- The model is highly significant and explains 25.46% variability.

Model with interaction

```r
summary(m4 <- lm(Happiness ~ log(GDPpc) * HDI4, data=HappyPlanet))
```

```
## Multiple R-squared:  0.268,  Adjusted R-squared:  0.2238
## F-statistic: 6.066 on 7 and 116 DF,  p-value: 4.821e-06
```

- $F = 6.07, P = 4.82 \times 10^{-6} < 0.05, R^2 = 0.2680, R^2_{adj} = 0.2238$
- The model is highly significant and explains 26.80% variability.

*Happiness ~ log(GDPpc) * HDI4*

```r
summary(m4 <- lm(Happiness ~ log(GDPpc) * HDI4, data=HappyPlanet))
```

```
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               24.205     21.856   1.107   0.2704
## log(GDPpc)                 1.580      3.403   0.464   0.6433
## HDI4Medium                58.407     31.425   1.859   0.0656 .
## HDI4High                  25.233     36.079   0.699   0.4857
## HDI4VeryHigh              16.599     35.339   0.470   0.6394
## log(GDPpc):HDI4Medium     -6.178      4.566  -1.353   0.1787
## log(GDPpc):HDI4High       -1.849      4.794  -0.386   0.7004
## log(GDPpc):HDI4VeryHigh   -1.710      4.336  -0.394   0.6940
```

- Since *HDI4* has three dummy variables (M, H and V) in the model, the interaction of $log(GDPpc) \times HDI4$ also has three terms, $log(GDPpc) \times M$, $log(GDPpc) \times H$ and $log(GDPpc) \times V$.
- None of the individual t tests for the interaction terms is significant.

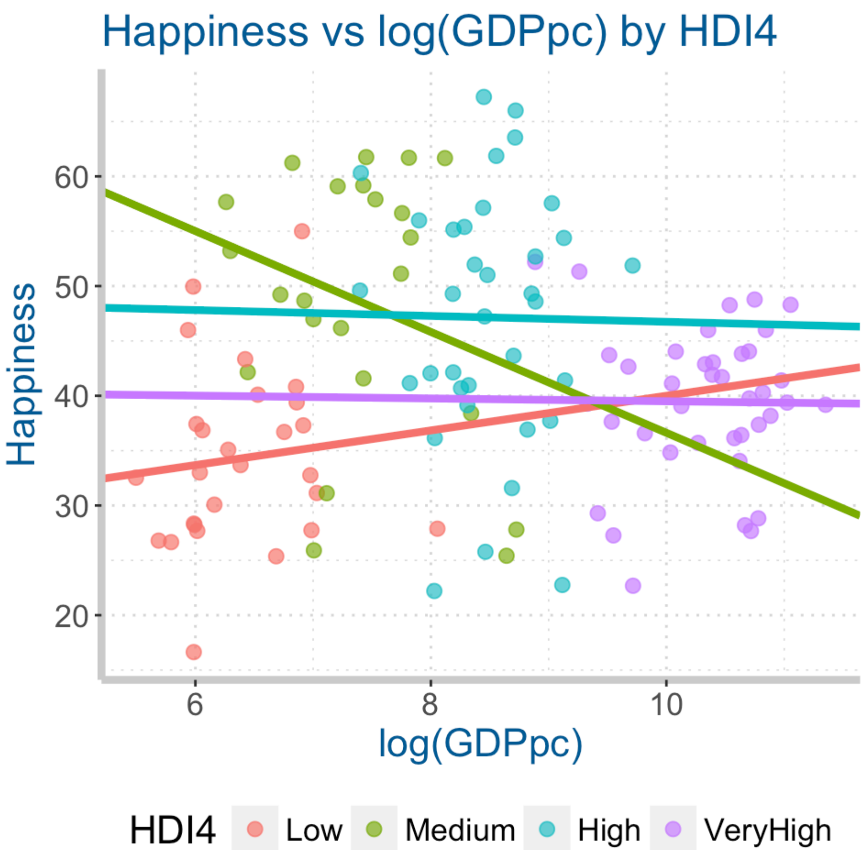
Happiness ~ log(GDPpc) HDI4*

$$\begin{aligned}\widehat{Happiness} &= 24.2 + 1.6 \times log(GDPpc) + 58.4 \times M + 25.2 \times H + 16.6 \times V \\ &\quad - 6.2 \times log(GDPpc) \times M - 1.8 \times log(GDPpc) \times H \\ &\quad - 1.7 \times log(GDPpc) \times V \\ &= [24.2 + 58.4 \times M + 25.2 \times H + 16.6 \times V] \\ &\quad + [1.6 - 6.2 \times M - 1.8 \times H - 1.7 \times V] \times log(GDPpc)\end{aligned}$$

- Intercept: $24.2 + 58.4 \times M + 25.2 \times H + 16.6 \times V$
- Slope: $1.6 - 6.2 \times M - 1.8 \times H - 1.7 \times V$
- This model allows the slope of *log(GDPpc)* (effect of *log(GDPpc)* on *Happiness*) to vary according to the values of *HDI4*.
- -6.2, -1.8 and -1.7 are the difference of the slopes of *log(GDPpc)* between each of the *M*, *H*, *V* groups and the baseline group *L*, respectively.

*Happiness ~ log(GDPpc) * HDI4*

$$\widehat{Happiness} = [24.2 + 58.4 \times M + 25.2 \times H + 16.6 \times V] + [1.6 - 6.2 \times M - 1.8 \times H - 1.7 \times V] \times log(GDPpc)$$



- L: $\widehat{Happiness} = 24.2 + 1.6 \times log(GDPpc)$
- M: $\widehat{Happiness} = 82.6 - 4.6 \times log(GDPpc)$
- H: $\widehat{Happiness} = 49.4 - 0.2 \times log(GDPpc)$
- V: $\widehat{Happiness} = 40.8 - 0.1 \times log(GDPpc)$
- Although these lines are not parallel, statistically their slopes are not that different.
- Shall we remove the interaction term log$(GDPpc) \times HDI4$?
- Nested F test.

Happiness ~ log(GDPpc) * HDI4

```r
anova(m3, m4)
```

```
## Analysis of Variance Table
##
## Model 1: Happiness ~ log(GDPpc) + HDI4
## Model 2: Happiness ~ log(GDPpc) * HDI4
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    119 11197
## 2    116 10996  3    201.14 0.7073 0.5495
```

Nested F test for the significance of $log(GDPpc) \times HDI4$:

- $F = 0.707$ and $P = 0.550$.
- Given that *log(GDPpc)* and *HDI4* are both included in the model, the interaction between them is not significant in explaining *Happiness*.
- The model with $log(GDPpc) \times HDI4$ is not significantly better and has slighly smaller R^2_{adj} than the model without the interaction. We should remove it.

Compare the models

```r
anova(m1, m2, m3, m4)
```

```
## Analysis of Variance Table
##
## Model 1: Happiness ~ log(GDPpc) + HDI2
## Model 2: Happiness ~ log(GDPpc) * HDI2
## Model 3: Happiness ~ log(GDPpc) + HDI4
## Model 4: Happiness ~ log(GDPpc) * HDI4
##   Res.Df   RSS Df Sum of Sq       F    Pr(>F)
## 1    121 14746
## 2    120 13387  1   1359.49 14.3419 0.0002431 ***
## 3    119 11197  1   2189.76 23.1009 4.647e-06 ***
## 4    116 10996  3    201.14  0.7073 0.5495392
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model	R^2	R^2_{adj}
1	0.0183	0.0021
2	0.1088	0.0865
3	0.2546	0.2295
4	0.2680	0.2238

Model 2 is significantly better than Model 1 ($P = 0.0002$). Model 3 is significantly better than Model 2 ($P = 4.647 \times 10^{-6}$). Model 4 is no better than Model 3 ($P = 0.5495$). Model 3 has the highest adjusted R^2 among the four models. Therefore, Model 3 is the best.