STAT021 Statistical Methods II

Lecture 16 MLR Categorical Predictors

Lu Chen
Swarthmore College
11/6/2018

Review

Multiple linear regression

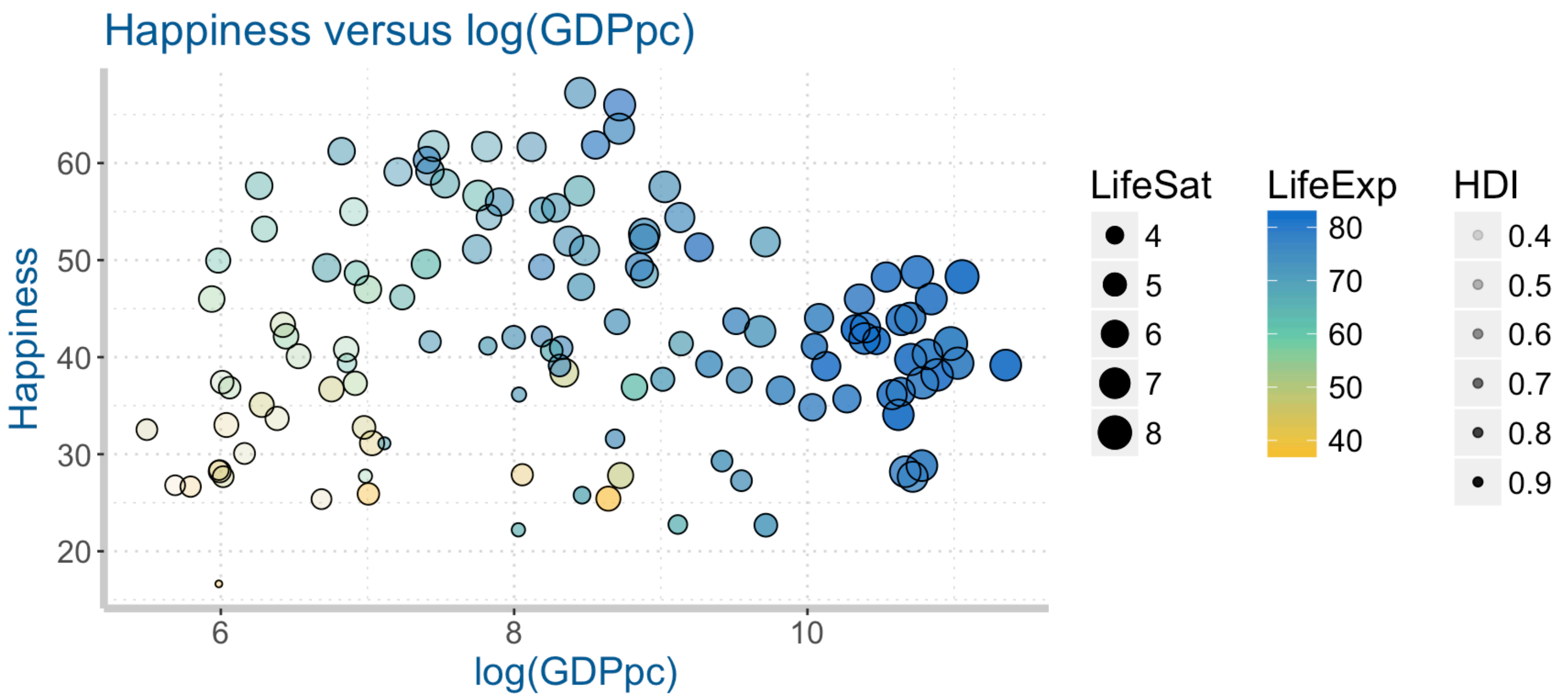
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \epsilon, \text{ where } \epsilon \overset{iid}{\sim} N(0, \sigma)$$

- MLR analysis of variance (ANOVA)
  - F test and R^2
- Three tests
  - t test for the slopes
  - F test for the MLR model
  - t test for the correlations
- Nested F test for a subset of predictors
- Adjusted R^2 for model comparison

Happy Planet Index

```r
head(HappyPlanet, 3)
```

```
##       Country Happiness      GDPpc LifeExp LifeSat       HDI
## 1 Philippines  59.17430  1678.8520    70.4 6.40000 0.6682183
## 2      Rwanda  28.34747   398.2085    43.9 4.43995 0.4832405
## 3     Hungary  37.63759 13842.6055    72.7 5.70000 0.8283505
```



Happy Planet Index

```r
summary(lm(Happiness ~ log(GDPpc)+LifeExp+LifeSat+HDI, data=HappyPlanet))
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.67493    3.26129   3.886 0.000168 ***
## log(GDPpc)  -8.91513    0.94204  -9.464 3.49e-16 ***
## LifeExp      0.73834    0.08806   8.384 1.20e-13 ***
## LifeSat      7.55823    0.58596  12.899  < 2e-16 ***
## HDI         14.09056   11.77288   1.197 0.233738
##
## Residual standard error: 5.412 on 119 degrees of freedom
## Multiple R-squared:  0.7679, Adjusted R-squared:  0.7601
## F-statistic: 98.46 on 4 and 119 DF,  p-value: < 2.2e-16
```

Happy Planet Index

```r
HDI2 <- cut(HappyPlanet$HDI, breaks=c(0, 0.7, 1), labels=c("Low","High"))
HDI4 <- cut(HappyPlanet$HDI, breaks=c(0, 0.55, 0.7, 0.8, 1),
            labels=c("Low","Medium","High","VeryHigh"))
HappyPlanet <- data.frame(HappyPlanet, HDI2, HDI4)
print(head(HappyPlanet), digits=3)
```

```
##                Country Happiness GDPpc LifeExp LifeSat   HDI HDI2     HDI4
## 1          Philippines      59.2  1679    70.4    6.40 0.668  Low   Medium
## 2               Rwanda      28.3   398    43.9    4.44 0.483  Low      Low
## 3              Hungary      37.6 13843    72.7    5.70 0.828 High VeryHigh
## 4               Cyprus      46.0 31387    78.6    6.90 0.850 High VeryHigh
## 5 Trinidad and Tobago      51.9 16530    69.9    6.86 0.772 High     High
## 6             Paraguay      51.1  2312    71.0    6.55 0.679  Low   Medium
```

- The `cut()` function categorizes a quantitative variable based on the provided cutoffs. Each interval is a category and is left-open and right-closed by default.
- To make the intervals left-closed and right-open, add `right=TRUE`.

Categorical predictors and interactions

Response variable

- *Happiness*

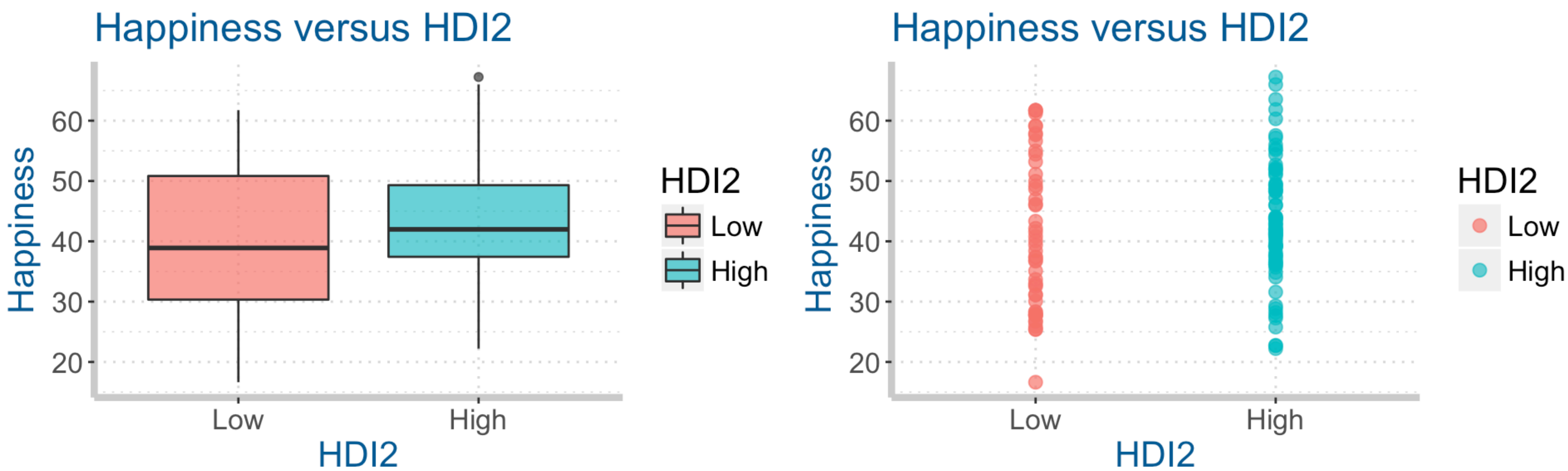
Explanatory variables

- *log(GDPpc)*
- *HDI2*, two categories, or
- *HDI4*, four categories

Relationships to consider:

- *Happiness* ~ *HDI2*
- *Happiness* ~ *HDI4*
- *Happiness* ~ *log(GDPpc)* + *HDI2*
- *Happiness* ~ *log(GDPpc)* + *HDI4*

Happiness ~ HDI2



To evaluate the relationship between a quantitative variable and a binary variable, what method should be used?

- Two-sample t test
- ANOVA
- Simple linear regression

Happiness ~ HDI2

Two-sample t test assuming equal variance

```
t.test(Happiness ~ HDI2, data=HappyPlanet, var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  Happiness by HDI2
## t = -1.0828, df = 122, p-value = 0.281
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.190987  1.812985
## sample estimates:
##  mean in group Low mean in group High
##           40.91051           43.09951
```

- $H_0 : \mu_{Low} = \mu_{High}; t = -1.08, P = 0.281 > 0.05.$
- The mean happiness scores of the two HDI groups are not significantly different.

Happiness ~ HDI2

ANOVA

```r
summary(aov(Happiness ~ HDI2, data=HappyPlanet))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## HDI2          1    143     143   1.172  0.281
## Residuals   122  14878     122
```

- Model: $Y = \mu + \alpha_k + \epsilon$, where $k = Low$ or $High$ and $\epsilon \overset{iid}{\sim} N(0, \sigma)$.
- $H_0 : \alpha_{Low} = \alpha_{High} = 0 \Leftrightarrow \mu_{Low} = \mu_{High}$
- $F = 1.172, P = 0.281$.
- The mean happiness scores of the two HDI groups are not significantly different.
- If the categorical variable has two categories, the ANOVA F test is equivalent to a pooled two-sample t test.

Happiness ~ HDI2

Response: *Happiness* as Y

Explanatory: *HDI2* as X

- Model: $Y = \beta_0 + \beta_1 X + \epsilon$, where $\epsilon \overset{iid}{\sim} N(0, \sigma)$.
- $X = 0$ for *Low* and $X = 1$ for *High*
- Here, X is a **dummy variable**.

> A **dummy variable** (also known as an indicator variable or binary variable) is one that takes the value 0 or 1 to indicate the absence or presence of some categorical effect that may affect the response value.

Happiness ~ HDI2

- Model: $Y = \beta_0 + \beta_1 X + \epsilon$, where $\epsilon \overset{iid}{\sim} N(0, \sigma)$.
- $\mu_Y = \beta_0 + \beta_1 X$
  - $X = 0, \mu_Y = \beta_0$ for the Low group
  - $X = 1, \mu_Y = \beta_0 + \beta_1$ for the $High$ group
  - β_0 is the mean of Y for the $X = 0$ group, which is usually called a reference/baseline category.
  - β_1 is the difference in Y between the two X groups.
  - To test $H_0 : \beta_1 = 0$ is to test whether the two groups means are equal and whether there is a significant relationship between Y and X.

Happiness ~ HDI2

```r
summary(aov(Happiness ~ HDI2, data=HappyPlanet))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## HDI2          1    143     143   1.172  0.281
## Residuals   122  14878     122
```

- $\bar{y}_{Low} = 40.91, \bar{y}_{High} = 43.10$
- $F = 1.172, P = 0.281$

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

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   40.911      1.562  26.196   <2e-16 ***
## HDI2High       2.189      2.022   1.083    0.281
##
## Residual standard error: 11.04 on 122 degrees of freedom
## Multiple R-squared:  0.009519,   Adjusted R-squared:  0.0014
## F-statistic: 1.172 on 1 and 122 DF,  p-value: 0.281
```

- $b_0 = 40.91, b_1 = 2.19$ $(b_0 + b_1 = 43.10)$
- $F = 1.172, P = 0.281$

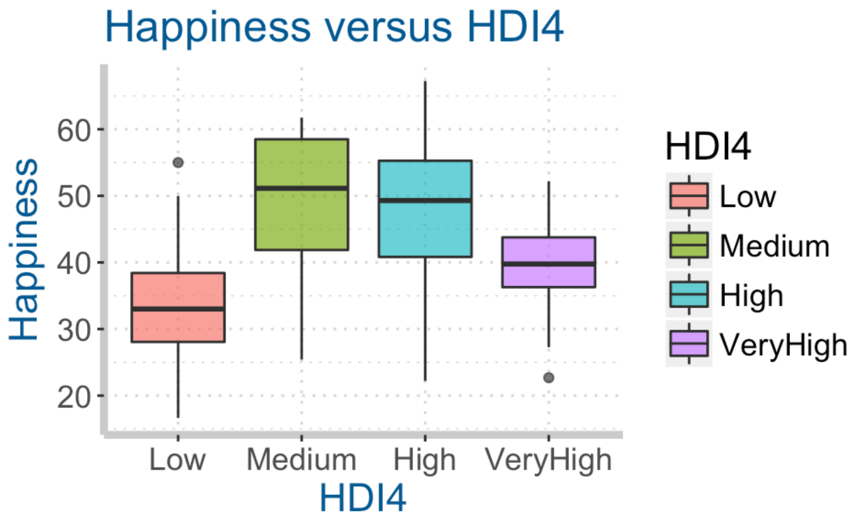
- `HDI2High`: the `lm()` function in R automatically assigns value 0 to the *Low* group (treats it as the baseline category) and value 1 to the *High* group.

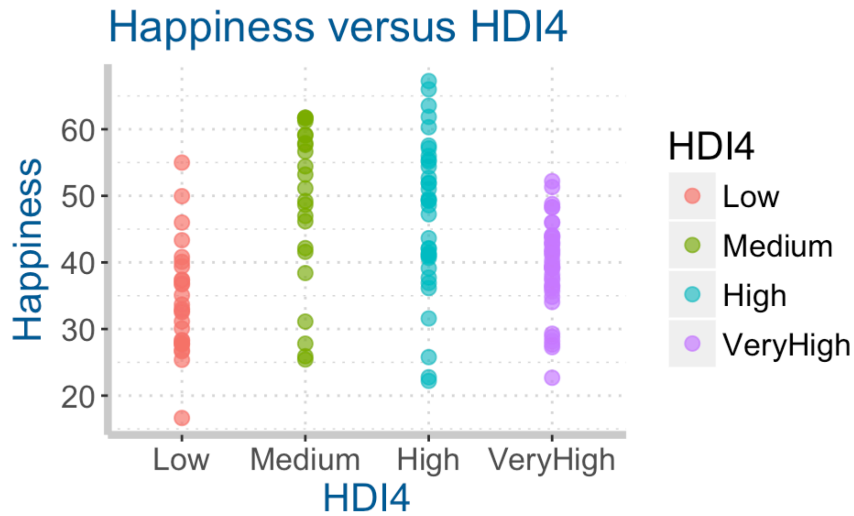
Happiness ~ HDI2

Model	H_0	Test statistic	P-value	Estimates
Pooled 2-sample t test	$\mu_1 = \mu_2$	$t = -1.083$	0.281	$\mu_{Low} = 40.91$, $\mu_{High} = 43.10$
ANOVA F test	$\mu_1 = \mu_2$	$F = 1.172$	0.281	$\mu_{Low} = 40.91$, $\mu_{High} = 43.10$
SLR t test for slope	$\beta_1 = 0$	$t = -1.083$	0.281	$b_0 = 40.91$, $b_1 = 2.19$ $b_0 + b_1 = 43.10$
SLR F test for model	$\beta_1 = 0$	$F = 1.172$	0.281	$b_0 = 40.91$, $b_1 = 2.19$ $b_0 + b_1 = 43.10$

These four tests are all equivalent to each other.

Happiness ~ HDI4






```
aggregate(Happiness ~ HDI4, data=HappyPlanet, FUN=mean)
```

```
##        HDI4 Happiness
## 1       Low  34.31538
## 2    Medium  48.65263
## 3      High  47.15364
## 4  VeryHigh  39.46120
```

Happiness ~ HDI4

ANOVA

```r
summary(aov(Happiness ~ HDI4, data=HappyPlanet))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## HDI4          3   3788  1262.5   13.49 1.21e-07 ***
## Residuals   120  11233    93.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Model: $Y = \mu + \alpha_k + \epsilon$, where $k = Low, Medium, High$ or $VeryHigh$ and $\epsilon \overset{iid}{\sim} N(0, \sigma)$.
- $H_0 : \alpha_L = \alpha_M = \alpha_H = \alpha_V = 0$ or $\mu_L = \mu_M = \mu_H = \mu_V$
- $F = 13.49, P = 1.21 \times 10^{-7} < 0.05$.
- At least one of the four HDI groups has significantly different mean from the others.

Happiness ~ HDI4

How should a linear regression model assume four different means?

- Create several **dummy variables** from one categorical variable.
- Set $HDI4 = Low$ as the baseline category; X_M indicates the $Medium$ group; X_H indicates the $High$ group and X_V indicates the $VeryHigh$ group.
- Model: $Y = \beta_0 + \beta_M X_M + \beta_H X_H + \beta_V X_V + \epsilon$, where $\epsilon \overset{iid}{\sim} N(0, \sigma)$.
- $X_M = 0, X_H = 0, X_V = 0 \iff HDI4 = Low$
- $X_M = 1, X_H = 0, X_V = 0 \iff HDI4 = Medium$
- $X_M = 0, X_H = 1, X_V = 0 \iff HDI4 = High$
- $X_M = 0, X_H = 0, X_V = 1 \iff HDI4 = VeryHigh$

Happiness ~ HDI4

$Y = \beta_0 + \beta_M X_M + \beta_H X_H + \beta_V X_V + \epsilon$, where $\epsilon \overset{iid}{\sim} N(0, \sigma)$

- $X_M = 0, X_H = 0, X_V = 0$, then $\mu_L = \beta_0$
  β_0 is the mean of Y for the baseline (Low) group.
- $X_M = 1, X_H = 0, X_V = 0$, then $\mu_M = \beta_0 + \beta_M$
  β_M is the difference in Y between the $Medium$ group and the baseline (Low) group.
- $X_M = 0, X_H = 1, X_V = 0$, then $\mu_H = \beta_0 + \beta_H$
  β_H is the difference in Y between the $High$ group and the baseline (Low) group.
- $X_M = 0, X_H = 0, X_V = 1$, then $\mu_V = \beta_0 + \beta_V$
  β_V is the difference in Y between the $VeryHigh$ group and the baseline (Low) group.

Happiness ~ HDI4

```r
summary(m2 <- lm(Happiness ~ HDI4, data=HappyPlanet))
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      34.315      1.862  18.429  < 2e-16 ***
## HDI4Medium       14.337      2.745   5.222 7.53e-07 ***
## HDI4High         12.838      2.478   5.180 9.05e-07 ***
## HDI4VeryHigh      5.146      2.422   2.124   0.0357 *
##
## Residual standard error: 9.675 on 120 degrees of freedom
## Multiple R-squared:  0.2522, Adjusted R-squared:  0.2335
## F-statistic: 13.49 on 3 and 120 DF,  p-value: 1.213e-07
```

- $b_0 = \bar{y}_L = 34.3$
  $b_1 = \bar{y}_M - \bar{y}_L = 14.3$
  $b_2 = \bar{y}_H - \bar{y}_L = 12.8$
  $b_3 = \bar{y}_V - \bar{y}_L = 5.1$

- $\bar{y}_L = b_0 = 34.3$, $\bar{y}_M = b_0 + b_1 = 48.7$, $\bar{y}_H = b_0 + b_2 = 47.2$, $\bar{y}_V = b_0 + b_3 = 39.5$
- $F = 13.49$, $P = 1.21 \times 10^{-7} \Rightarrow$ equivalent to the ANOVA F test. The *HDI4* variable is highly significant in explaining *Happiness*.
- The *Medium*, the *High* and the *VeryHigh* group have significantly different *Happiness* score from the *Low* group.

Categorical predictors

- To evaluate the relationship between a quantitative and a categorical variable, the ANOVA F test is equivalent to the linear regression F test. Both test whether the group means are the same.
- To include a categorical predictor in regression, we first transform the categorical variable with m categories to $m-1$ dummy variables and then fit a linear regression model for the response variable on these dummy variables.
  - The intercept b_0 is the mean of the baseline category.
  - The slope of each dummy variable is the difference in mean between the current category and the baseline category.
  - The t test for each slope indicates whether each category is significantly different from the baseline category.

Happiness ~ log(GDPpc) + HDI2

Response: *Happiness* as Y; **Explanatory**: *log(GDPpc)* as X_1 and *HDI2* as X_2.

Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, where $\epsilon \overset{iid}{\sim} N(0, \sigma)$.

```
summary(m3 <- lm(Happiness ~ log(GDPpc) + HDI2, data=HappyPlanet))
```

```
## Coefficients:
##             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    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 11.04 on 121 degrees of freedom
## Multiple R-squared:  0.01828,    Adjusted R-squared:  0.002051 
## F-statistic: 1.126 on 2 and 121 DF,  p-value: 0.3276
```

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

Happiness ~ log(GDPpc) + HDI2

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

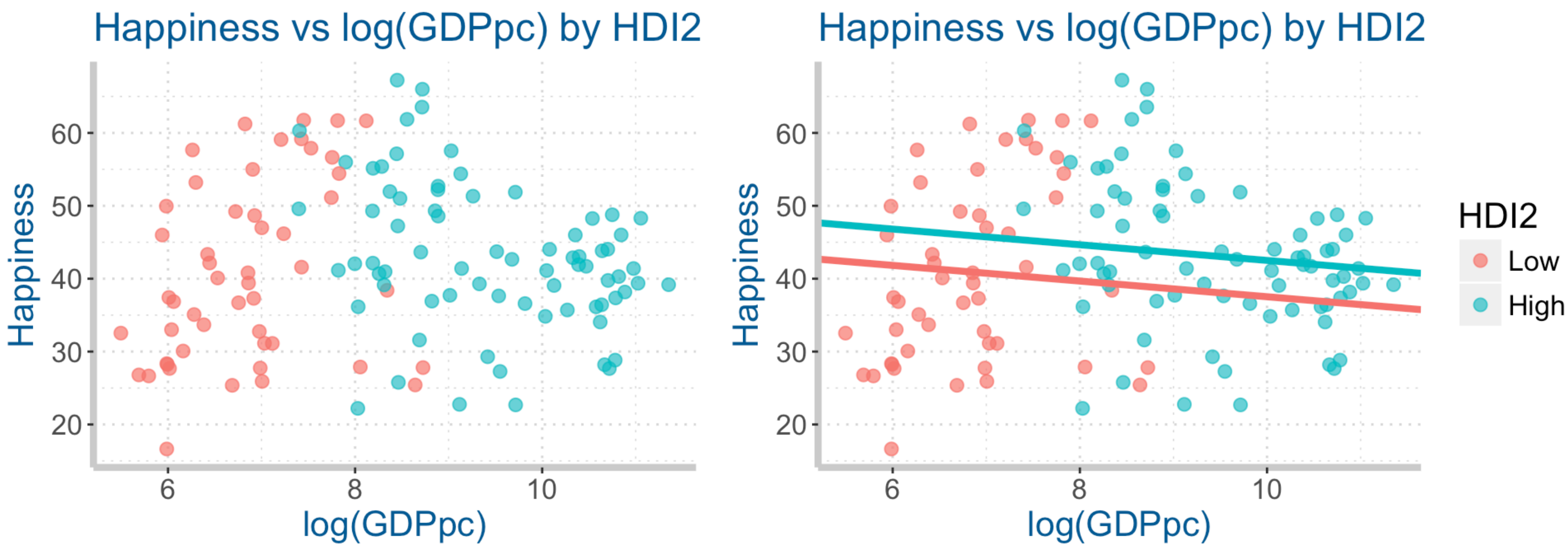
Given that *HDI2* is held constant, *Happiness* decreases 1.1 units as *log(GDPpc)* increases 1 unit.

- When $HDI2 = 0$ (*Low*), $\widehat{Happiness} = 48.3 - 1.1 \times log(GDPpc)$
- When $HDI2 = 1$ (*High*), $\widehat{Happiness} = 48.3 - 1.1 \times log(GDPpc) + 5.0 = 53.3 - 1.1 \times log(GDPpc)$

Given that *log(GDPpc)* is held constant, *Happiness* increases 5.0 units as *HDI2* increases 1 unit.

- For any value of *log(GDPpc)*, the difference in mean *Happiness* between the *Low* and *High* group is 5.0.

Happiness ~ log(GDPpc) + HDI2



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

- $\widehat{Happiness} = 48.3 - 1.1 \times log(GDPpc)$ for the *Low* HDI group.
- $\widehat{Happiness} = 53.3 - 1.1 \times log(GDPpc)$ for the *High* HDI group.

Happiness ~ log(GDPpc) + HDI2

```r
summary(m3)
```

```
## Coefficients:
##             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
##
## Residual standard error: 11.04 on 121 degrees of freedom
## Multiple R-squared:  0.01828,    Adjusted R-squared:  0.002051
## F-statistic: 1.126 on 2 and 121 DF,  p-value: 0.3276
```

- Individual t tests: When both *log(GDPpc)* and *HDI2* are included in the model, none of them is significant in explaining *Happiness.*
- F test: The model is not significant in explaining *Happiness.*
- $R^2 = 0.018$. Only 1.8% of the variablity in *Happiness* is explained by the model.

Happiness ~ log(GDPpc) + HDI4

Response: *Happiness* as Y;

Explanatory: *log(GDPpc)* as X_1 and *HDI4* as X_M, X_H and X_V.

Model: $Y = \beta_0 + \beta_1 X_1 + \beta_M X_M + \beta_H X_H + \beta_V X_V + \epsilon$, where $\epsilon \overset{iid}{\sim} N(0, \sigma)$.

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

```
## Coefficients:
##              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
```

$$\widehat{Happiness} = 40.4 - 1.0 \times log(GDPpc) + 15.3 HDI_M + 14.8 HDI_H + 8.9 HDI_V$$

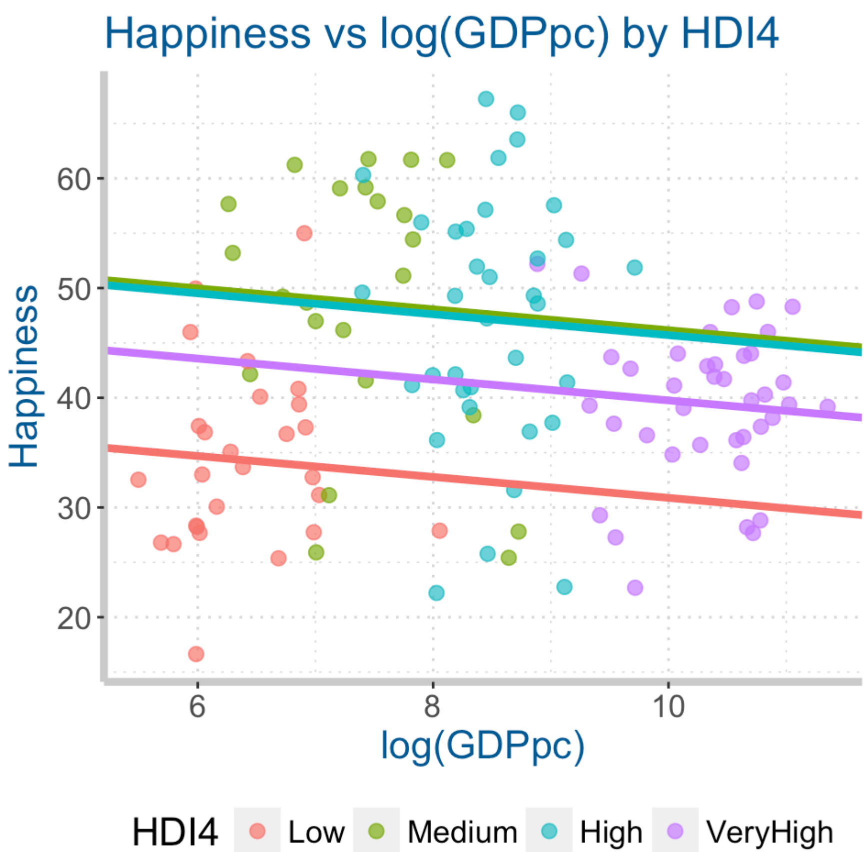
Happiness ~ log(GDPpc) + HDI4

$$\widehat{Happiness} = 40.4 - 1.0 \times log(GDPpc) + 15.3HDI_M + 14.8HDI_H + 8.9HDI_V$$

- Given that *HDI4* is held constant, *Happiness* decreases 1.0 unit as *log(GDPpc)* increases 1 unit.
- Given that *log(GDPpc)* is held constant,
  - the difference in mean *Happiness* between the *Medium* and *Low* group is 15.3.
  - the difference in mean *Happiness* between the *High* and *Low* group is 14.8.
  - the difference in mean *Happiness* between the *VeryHigh* and *Low* group is 8.9.

Happiness ~ log(GDPpc) + HDI4

$$\widehat{Happiness} = 40.4 - 1.0 \times log(GDPpc) + 15.3HDI_M + 14.8HDI_H + 8.9HDI_V$$



$HDI_M = HDI_H = HDI_V = 0$,
$\widehat{Happiness} = 40.4 - 1.0 \times log(GDPpc)$ for the *Low* HDI group.

$HDI_M = 1$,
$\widehat{Happiness} = 55.7 - 1.0 \times log(GDPpc)$ for the *Medium* HDI group.

$HDI_H = 1$,
$\widehat{Happiness} = 55.2 - 1.0 \times log(GDPpc)$ for the *High* HDI group.

$HDI_V = 1$,
$\widehat{Happiness} = 49.3 - 1.0 \times log(GDPpc)$ for the *VeryHigh* HDI group.

Happiness ~ log(GDPpc) + HDI4

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

```
## Coefficients:
##               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
```

- $F = 10.16, P = 4.13 \times 10^{-7}$. The model including both *log(GDPpc)* and *HDI4* is highly significant in explaining *Happiness*.
- $R^2 = 0.2546$. 25.46% of the variability is explained by the model.

Happiness ~ log(GDPpc) + HDI4

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

```
## Coefficients:
##              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
```

- Given that *HDI4* is held constant in the model, *log(GDPpc)* is not significant in explaining *Happiness*.
- Ajusted for *log(GDPpc)*, the *Medium* and the *High* group has significantly different *Happiness* score from the *Low* group. But the difference between the *VeryHigh* and the *Low* group is not significant.
- **Is *HDI4* (the three dummy variables as a whole) significant in explaining *Happiness*?**

Happiness ~ log(GDPpc) + HDI4

Nested F test for the significance of *HDI4* (three dummy variables)

```
m0 <- lm(Happiness ~ log(GDPpc), data=HappyPlanet)
anova(m0, m4)
```

```
## Analysis of Variance Table
##
## Model 1: Happiness ~ log(GDPpc)
## Model 2: Happiness ~ log(GDPpc) + HDI4
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1    122 15014
## 2    119 11197  3    3816.7 13.521 1.19e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $H_0 : \beta_M = \beta_H = \beta_V = 0; F = 13.5$ and $P < 0.05$.
- Given that *log(GDPpc)* is included in the model, $HDI4$ is highly significant in explaining *Happiness*.

HDI vs *HDI2* vs *HDI4*

```r
GLL <- lm(Happiness ~ log(GDPpc)+LifeExp+LifeSat, data=HappyPlanet)
GLL.HDI <- lm(Happiness ~ log(GDPpc)+LifeExp+LifeSat+HDI, data=HappyPlanet)
GLL.HDI2 <- lm(Happiness ~ log(GDPpc)+LifeExp+LifeSat+HDI2, data=HappyPlanet)
GLL.HDI4 <- lm(Happiness ~ log(GDPpc)+LifeExp+LifeSat+HDI4, data=HappyPlanet)
summary(GLL.HDI)
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.67493    3.26129   3.886 0.000168 ***
## log(GDPpc)  -8.91513    0.94204  -9.464 3.49e-16 ***
## LifeExp      0.73834    0.08806   8.384 1.20e-13 ***
## LifeSat      7.55823    0.58596  12.899  < 2e-16 ***
## HDI         14.09056   11.77288   1.197 0.233738
##
## Residual standard error: 5.412 on 119 degrees of freedom
## Multiple R-squared:  0.7679, Adjusted R-squared:  0.7601
## F-statistic: 98.46 on 4 and 119 DF,  p-value: < 2.2e-16
```

- Given that *log(GDPpc)*, *LifeExp* and *LifeSat* are included in the model, *HDI* is not significant in explaining *Happiness*.

HDI vs *HDI2* vs *HDI4*

```r
summary(GLL.HDI2)
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.99569    4.20971   3.325  0.00118 **
## log(GDPpc)  -8.31371    0.63200 -13.155  < 2e-16 ***
## LifeExp      0.77454    0.07479  10.356  < 2e-16 ***
## LifeSat      7.58074    0.60437  12.543  < 2e-16 ***
## HDI2High     1.84528    1.92607   0.958  0.33998
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.424 on 119 degrees of freedom
## Multiple R-squared:  0.767,  Adjusted R-squared:  0.7591
## F-statistic: 97.91 on 4 and 119 DF,  p-value: < 2.2e-16
```

- Given that *log(GDPpc)*, *LifeExp* and *LifeSat* are included in the model, *HDI2* is not significant in explaining *Happiness*.

HDI vs *HDI2* vs *HDI4*

```r
summary(GLL.HDI4)
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.8346     5.7324  -0.320   0.7495
## log(GDPpc)      -5.5587     0.7829  -7.100 1.04e-10 ***
## LifeExp          0.7162     0.0730   9.810  < 2e-16 ***
## LifeSat          7.3325     0.5133  14.284  < 2e-16 ***
## HDI4Medium       3.2891     1.7291   1.902   0.0596 .
## HDI4High         2.4616     2.5077   0.982   0.3283
## HDI4VeryHigh    -6.8255     3.6777  -1.856   0.0660 .
##
## Residual standard error: 4.589 on 117 degrees of freedom
## Multiple R-squared:  0.836,  Adjusted R-squared:  0.8276
## F-statistic: 99.38 on 6 and 117 DF,  p-value: < 2.2e-16
```

- Given that *log(GDPpc)*, *LifeExp* and *LifeSat* are held constant in the model, the *Happiness* value of the *Medium* and the *VeryHigh* group are **marginally** significantly different from the *Low* group.

HDI vs *HDI2* vs *HDI4*

```r
anova(GLL, GLL.HDI4)
```

```
## Analysis of Variance Table
##
## Model 1: Happiness ~ log(GDPpc) + LifeExp + LifeSat
## Model 2: Happiness ~ log(GDPpc) + LifeExp + LifeSat + HDI4
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    120 3527.5
## 2    117 2464.0  3    1063.6 16.834 3.717e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Given that *log(GDPpc)*, *LifeExp* and *LifeSat* are included in the model, *HDI4* is highly significant in explaining *Happiness*.

	GLL	GLL.HDI	GLL.HDI2	GLL.HDI4
R^2	0.7652	0.7679	0.7670	0.8360
R^2_{adj}	0.7593	0.7601	0.7591	0.8276

Summary

- Linear regression of a quantitative response variable on a categorical variable is **equivalent** to the corresponding ANOVA model. Both models evaluate whether the several categories have the the same mean.
- To include a categorical variable with m categories in a linear regression model, it should be first transformed to $m-1$ **dummy variables**.
- The slope(s) of dummy variable(s) in a regression model measures the **difference** in mean response between the current category and the baseline category.
- **Categorizing** a quantitative variable sometimes allows more flexibility in model fitting and can explain more variability in the response variable.