Lecture 15: Categorical Predictors with More Than Two Levels

STAT 432, Spring 2021

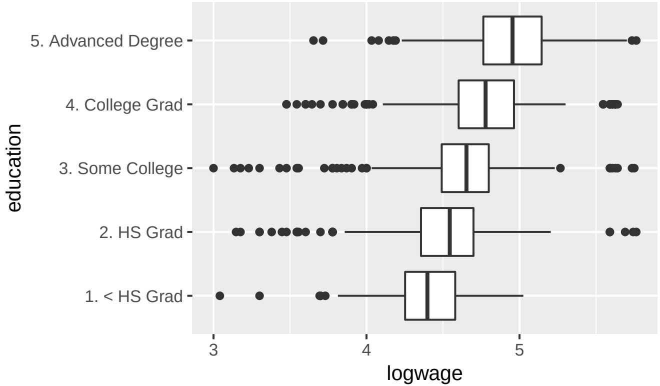
- When a categorical predictor contains more than two levels, we create additional dummy variables.
- For example, consider the `Wage` data set also from the `ISLR` package. The data contain information on 3000 males workers in the Mid-Atlantic region.
- The response variable is `logwage`, the log of the workers wage.
- The predictor `education` is a categorical variable indicating education level with 5 levels: 1. $<$ HS Grad, 2. HS Grad, 3. Some College, 4. College Grad, and 5. Advanced Degree.

We can write the regression equation with 4 dummy variables:

$$\begin{aligned}
\log(\texttt{Wage}) &= \beta_0 + \beta_1 \texttt{HS_Grad} + \beta_2 \texttt{Some_College} \\
&\quad + \beta_3 \texttt{College_Grad} + \beta_4 \texttt{Advanced_Degree} + \epsilon \\
&= \begin{cases}
\beta_0 + \epsilon & \text{if } {<}\texttt{HS_Grad} \text{ (baseline)} \\
\beta_0 + \beta_1 + \epsilon & \text{if } \texttt{HS_Grad} = 1 \\
\beta_0 + \beta_2 + \epsilon & \text{if } \texttt{Some_College} = 1 \\
\beta_0 + \beta_3 + \epsilon & \text{if } \texttt{College_Grad} = 1 \\
\beta_0 + \beta_4 + \epsilon & \text{if } \texttt{Advanced_Degree} = 1
\end{cases}
\end{aligned}$$

In general, if we have a categorical variable with k levels, then the regression equation contains $k - 1$ dummy variables.

```
> library(ISLR)
> library(ggplot2)
> ggplot(Wage, aes(education, logwage)) +
    geom_boxplot() + coord_flip()
```



```
> lm1 <- lm(logwage ~ education, data=Wage)
> summary(lm1)

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  4.39759    0.01891 232.502  < 2e-16 ***
education2. HS Grad          0.12295    0.02137   5.754 9.57e-09 ***
education3. Some College     0.23821    0.02248  10.597  < 2e-16 ***
education4. College Grad     0.37373    0.02231  16.752  < 2e-16 ***
education5. Advanced Degree  0.56036    0.02414  23.212  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3096 on 2995 degrees of freedom
Multiple R-squared:  0.2262,Adjusted R-squared:  0.2251
F-statistic: 218.8 on 4 and 2995 DF,  p-value: < 2.2e-16
```

Using the summary output we can write the fitted regression model as

$$
\begin{aligned}
\widehat{\log(\texttt{Wage})} &= 4.398 + 0.123\texttt{HS\_Grad} + 0.238\texttt{Some\_College} \\
&\quad + 0.374\texttt{College\_Grad} + 0.560\texttt{Advanced\_Degree} \\
&= \begin{cases}
4.398 & \text{if } \texttt{<HS\_Grad} \text{ (baseline)} \\
4.398 + 0.123 = 4.521 & \text{if } \texttt{HS\_Grad} = 1 \\
4.398 + 0.238 = 4.636 & \text{if } \texttt{Some\_College} = 1 \\
4.398 + 0.374 = 4.772 & \text{if } \texttt{College\_Grad} = 1 \\
4.398 + 0.560 = 4.958 & \text{if } \texttt{Advanced\_Degree} = 1
\end{cases}
\end{aligned}
$$

We can also include interaction effects between the categorical predictor `education` and a quantitative variable such as `age` (age of worker). The model can be written out as:

$$
\begin{aligned}
\log(\texttt{Wage}) &= \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{HS\_Grad} + \beta_3 \texttt{Some\_College} \\
&\quad + \beta_4 \texttt{College\_Grad} + \beta_5 \texttt{Advanced\_Degree} \\
&\quad + \beta_6 \texttt{HS\_Grad} \cdot \texttt{age} + \beta_7 \texttt{Some\_College} \cdot \texttt{age} \\
&\quad + \beta_8 \texttt{College\_Grad} \cdot \texttt{age} + \beta_9 \texttt{Advanced\_Degree} \cdot \texttt{age} + \epsilon \\
&= \begin{cases}
\beta_0 + \beta_1 \texttt{age} + e & \text{if } \texttt{<HS\_Grad} \text{ (baseline)} \\
\beta_0 + \beta_2 + (\beta_1 + \beta_6) \texttt{age} + \epsilon & \text{if } \texttt{HS\_Grad} = 1 \\
\beta_0 + \beta_3 + (\beta_1 + \beta_7) \texttt{age} + \epsilon & \text{if } \texttt{Some\_College} = 1 \\
\beta_0 + \beta_4 + (\beta_1 + \beta_8) \texttt{age} + \epsilon & \text{if } \texttt{College\_Grad} = 1 \\
\beta_0 + \beta_5 + (\beta_1 + \beta_9) \texttt{age} + \epsilon & \text{if } \texttt{Advanced\_Degree} = 1
\end{cases}
\end{aligned}
$$

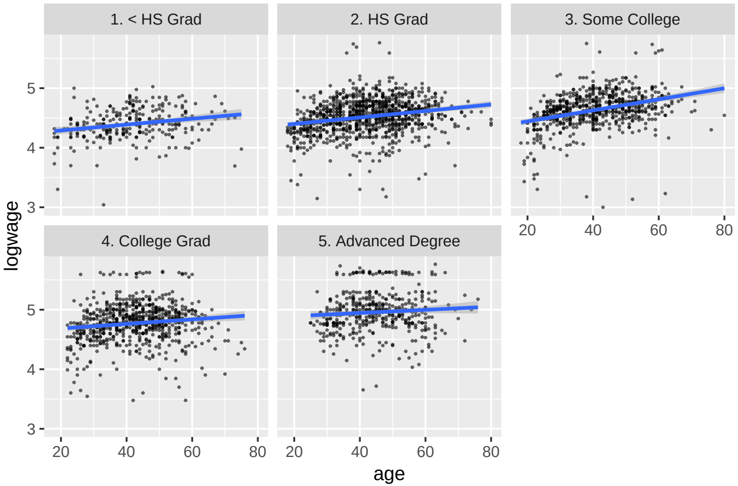
The regression model gives separate regression lines, which have different slopes and intercepts, for each level of the categorical predictor `education`.

```
> lm3 <- lm(logwage ~ age + education + age:education, data=Wage)
> summary(lm3)

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      4.1921197  0.0640086  65.493  < 2e-16 ***
age                              0.0049162  0.0014664   3.353 0.000811 ***
education2. HS Grad              0.0979291  0.0731558   1.339 0.180791
education3. Some College         0.0644316  0.0775180   0.831 0.405937
education4. College Grad         0.4160484  0.0792801   5.248 1.65e-07 ***
education5. Advanced Degree      0.6467308  0.0918866   7.038 2.40e-12 ***
age:education2. HS Grad          0.0005434  0.0016738   0.325 0.745466
age:education3. Some College     0.0043591  0.0017917   2.433 0.015033 *
age:education4. College Grad    -0.0011018  0.0018093  -0.609 0.542593
age:education5. Advanced Degree -0.0022699  0.0020470  -1.109 0.267563
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3022 on 2990 degrees of freedom
Multiple R-squared:  0.2642,Adjusted R-squared:  0.262
F-statistic: 119.3 on 9 and 2990 DF,  p-value: < 2.2e-16
```

```r
ggplot(Wage, aes(age, logwage)) +
  geom_point(size = 0.3, alpha=0.6) + facet_wrap( ~ education) +
  geom_smooth(method='lm')
```



To determine whether the interaction effects are actually meaningful to include we can use a model selection criteria such as adjusted R^2.

```
> lm1 <- lm(logwage ~ education, data=Wage)
> summary(lm1)$adj.r.squared
[1] 0.2251165
> lm2 <- lm(logwage ~ age + education, data=Wage)
> summary(lm2)$adj.r.squared
[1] 0.2580081
> lm3 <- lm(logwage ~ age + education + age:education, data=Wage)
> summary(lm3)$adj.r.squared
[1] 0.261979
```

We see that the model `lm3` with `age`, `education`, and the interaction effects between `age` and `education` is the best fitting model according to the adjusted R^2.

The F-test can also be used to compare the nested models. For example we can test whether or not $H_0 : \beta_6 = \cdots = \beta_9 = 0$ (the coefficients for the interaction terms are all zero).

```
> anova(lm2, lm3)
Analysis of Variance Table

Model 1: logwage ~ age + education
Model 2: logwage ~ age + education + age:education
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   2994 274.87
2   2990 273.03  4    1.8363 5.0273 0.0004885 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Since the p-value< 0.001 we reject H_0, which means that the model with the interactions is superior. This agrees with the adjusted R^2 criteria.