Lecture 13:
Categorical Predictors and Interactions
STAT 432, Spring 2021

Introduction

- Predictors in a multiple linear regression model can either be *quantitative* (e.g, weight, age) or *qualitative* (e.g., gender, education level). Qualitative predictors are also called *categorical* or *factors*.
- A categorical predictor with two levels (0 or 1) is called a *dummy* or *indicator* variable.
- Sometimes the effect that a quantitative predictor has on the response changes depending on the level of categorical predictor. For example, perhaps the effect age has on salary depends on the education status of the person. This is called an *interaction* effect.

Parallel Regression Lines

Let x be a quantitative variable, and d a dummy variable.

$$y = \beta_0 + \beta_1 x + \beta_2 d + \epsilon = \begin{cases} \beta_0 + \beta_1 x + \epsilon, & \text{if d=0} \\ (\beta_0 + \beta_2) + \beta_1 x + \epsilon, & \text{if d=1} \end{cases}$$

- This model gives two separate regression lines that have the same slope but different intercepts.
- The parameter β_2 represents the vertical distance between the two lines.

Unrelated Regression Lines

Let x be a quantitative variable, and d a dummy variable.

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x + \beta_2 d + \beta_3 d \cdot x + \epsilon \\
&= \begin{cases} \beta_0 + \beta_1 x + \epsilon, & \text{if d=0} \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x + \epsilon, & \text{if d=1} \end{cases}
\end{aligned}
$$

- This model gives two separate regression lines that have different slopes, and different intercepts.
- β_3 is the coefficient for the *interaction* between the dummy variable, d, and the quantitative variable, x.

Example: Credit Card Data Set

- We consider the `Credit` data set from the `ISLR` package. Type `help(Credit)` to read about this data set in the help menu.
- The response variable is `Balance`, the average credit card balance in dollars.
- The predictors of interest are `Income` (in thousands of dollars) and `Student`, a dummy variable indicating student status (No = 0 or Yes = 1).

```
> library(ISLR)
> library(ggplot2)
> head(Credit, n=5)
  ID  Income Limit Rating Cards Age Education Gender Student Married Ethnicity Balance
1  1  14.891  3606    283     2  34        11   Male      No     Yes Caucasian     333
2  2 106.025  6645    483     3  82        15 Female     Yes     Yes     Asian     903
3  3 104.593  7075    514     4  71        11   Male      No      No     Asian     580
4  4 148.924  9504    681     3  36        11 Female      No      No     Asian     964
5  5  55.882  4897    357     2  68        16   Male      No     Yes Caucasian     331
```

```
> lm1 <- lm(Balance ~ Income + Student, data=Credit)

# shows coding R uses for the dummy variable
> contrasts(Credit$Student)
    Yes
No    0
Yes   1

> summary(lm1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 211.1430    32.4572   6.505 2.34e-10 ***
Income        5.9843     0.5566  10.751  < 2e-16 ***
StudentYes  382.6705    65.3108   5.859 9.78e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 391.8 on 397 degrees of freedom
Multiple R-squared:  0.2775,Adjusted R-squared:  0.2738
F-statistic: 76.22 on 2 and 397 DF,  p-value: < 2.2e-16
```

We can write the regression equation for the fit:

$$\widehat{\texttt{Balance}} = \hat{\beta}_0 + \hat{\beta}_1 \texttt{income} + \hat{\beta}_2 \texttt{student} =$$

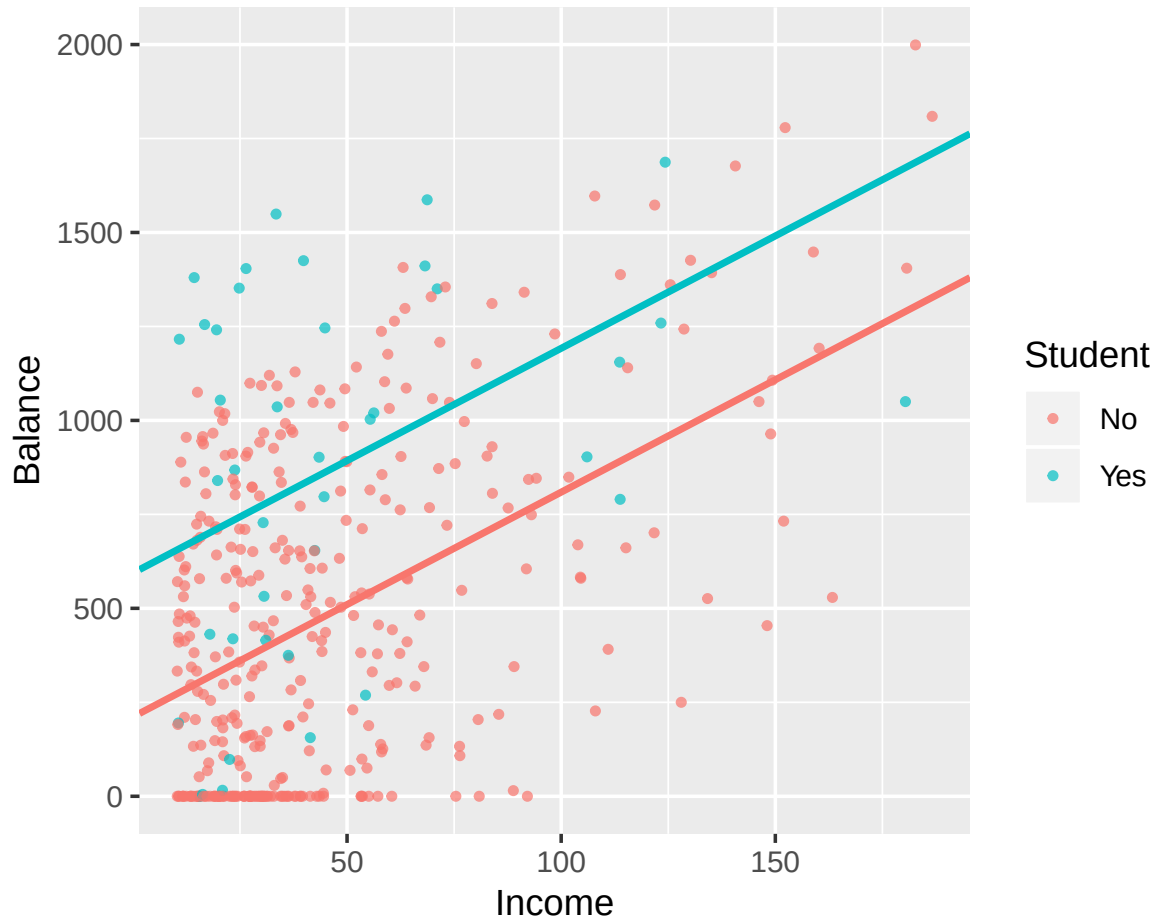
$$= \begin{cases} \hat{\beta}_0 + \hat{\beta}_1 \texttt{income}, & \text{if } \texttt{student}{=}0 \text{ (No)} \\ (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 \texttt{income}, & \text{if } \texttt{student}{=}1 \text{ (Yes)} \end{cases}$$

Plugging in the coefficients from the regression summary gives:

$$\widehat{\texttt{Balance}} = 211.14 + 5.98 \texttt{income} + 382.67 \texttt{student} =$$

$$= \begin{cases} 211.14 + 5.98 \texttt{income}, & \text{if } \texttt{student}{=}0 \text{ (No)} \\ 593.81 + 5.98 \texttt{income}, & \text{if } \texttt{student}{=}1 \text{ (Yes)} \end{cases}$$

```
ggplot(Credit, aes(Income, Balance, color = Student)) +
  geom_point(alpha=0.7) +
  geom_abline(intercept = 211.1, slope = 5.98, color = "#F8766D", size = 1) +
  geom_abline(intercept = 593.8, slope = 5.98, color = "#00BFC4", size = 1)
```



```
> lm2 <- lm(Balance ~ Income + Student + Income:Student, data=Credit)
> summary(lm2)

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       200.6232    33.6984   5.953 5.79e-09 ***
Income              6.2182     0.5921  10.502  < 2e-16 ***
StudentYes        476.6758   104.3512   4.568 6.59e-06 ***
Income:StudentYes  -1.9992     1.7313  -1.155    0.249
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 391.6 on 396 degrees of freedom
Multiple R-squared:  0.2799,Adjusted R-squared:  0.2744
F-statistic:  51.3 on 3 and 396 DF,  p-value: < 2.2e-16
```

We can write the regression equation for the fit:

$$\widehat{\texttt{Balance}} = \hat{\beta}_0 + \hat{\beta}_1 \texttt{income} + \hat{\beta}_2 \texttt{student} + \hat{\beta}_3 \texttt{student} \cdot \texttt{income}$$

$$= \begin{cases} \hat{\beta}_0 + \hat{\beta}_1 \texttt{income}, & \text{if } \texttt{student}{=}0 \text{ (No)} \\ (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3)\texttt{income}, & \text{if } \texttt{student}{=}1 \text{ (Yes)} \end{cases}$$

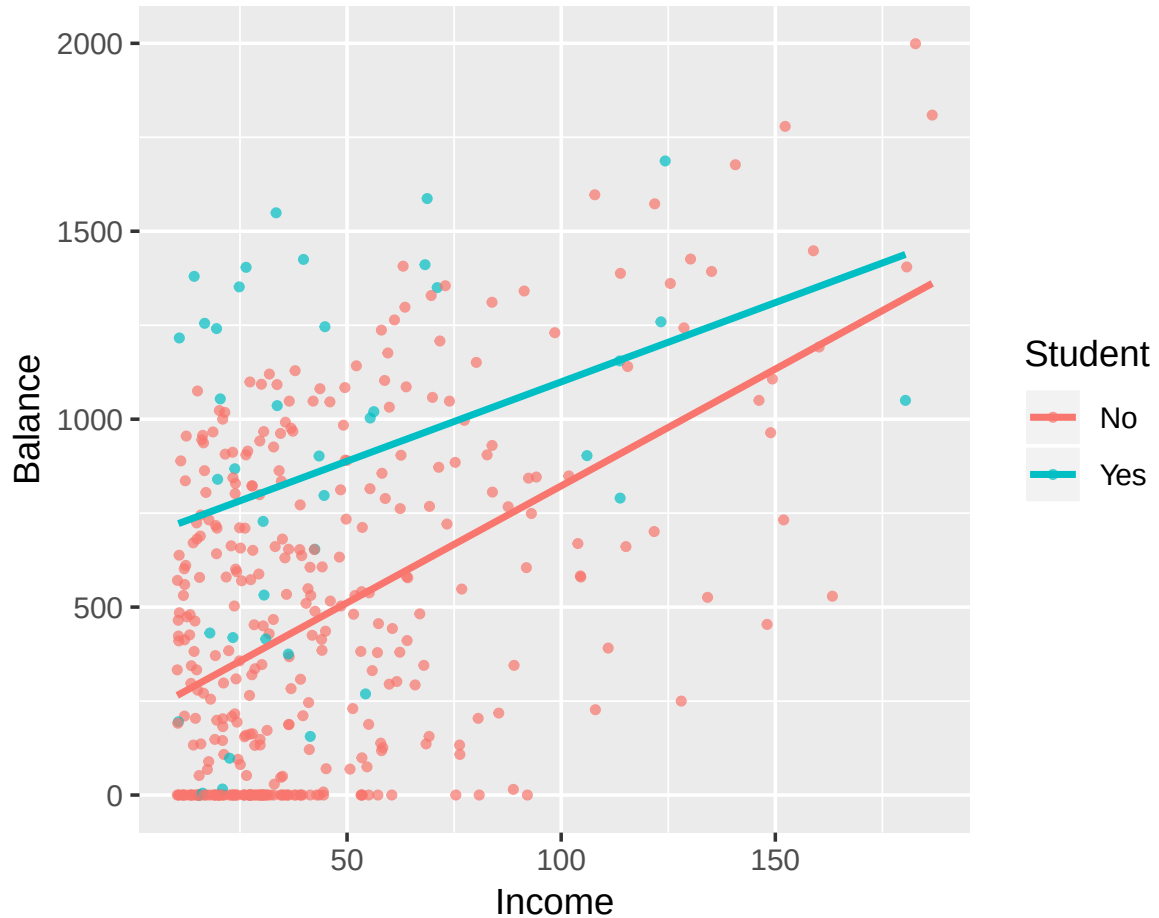
Plugging in the coefficients from the regression summary gives:

$$\widehat{\texttt{Balance}} = 200.62 + 6.22 \texttt{income} + 476.68 \texttt{student} - 2.00 \texttt{student} \cdot \texttt{income}$$

$$= \begin{cases} 200.62 + 6.22 \texttt{income}, & \text{if } \texttt{student}{=}0 \text{ (No)} \\ 677.3 + 4.22 \texttt{income}, & \text{if } \texttt{student}{=}1 \text{ (Yes)} \end{cases}$$

Note that the coefficient for the interaction, β_3, is not significant (p-value$= 0.249$), so we do not necessarily need to include the interaction term.

```
ggplot(Credit, aes(Income, Balance, color = Student)) +
  geom_point(alpha=0.7) +
  geom_smooth(method="lm", se=FALSE)
```



```
ggplot(Credit, aes(Income, Balance, color = Student)) +
  geom_point(alpha=0.7, size=0.9) +
  facet_wrap(~ Student) +
  geom_smooth(method="lm", se=FALSE, size=1)
```