Lecture 6:
Dummy Variable Regression
STAT 432, Spring 2021

- So far we have only considered explanatory variables that are quantitative (e.g., a person's height).
- An **indicator** or **dummy** variable is an explanatory variable that only takes on two possible numeric values (e.g., 0 or 1).
- In simple linear regression, dummy variables allow us to compare the means of two groups. The concept can be further generalized in multiple linear regression to problems involving more than two groups.

Example

- Data set from Ebay for the game Mario Kart for the Nintendo Wii.
- Consider a simple linear regression model relating the total price to the dummy variable for game condition (new or used).

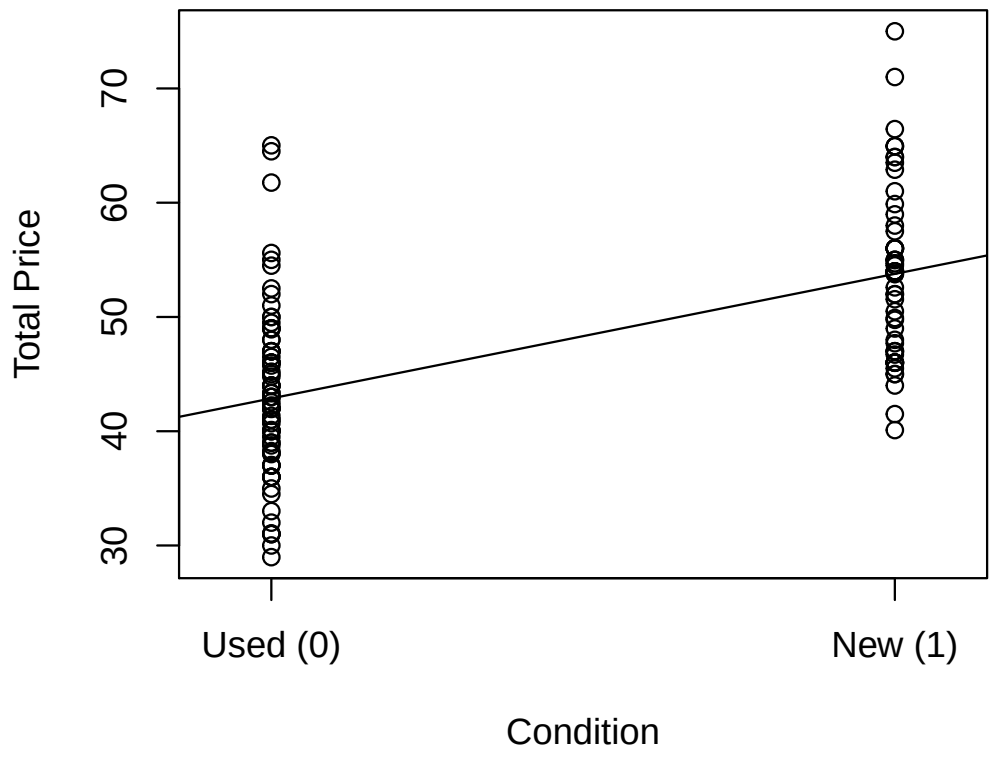
$$x_i = \begin{cases} 1, & \text{if } i^{th} \text{ game is new} \\ 0, & \text{if } i^{th} \text{ game is used} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \text{if } i^{th} \text{ game is new} \\ \beta_0 + \epsilon_i, & \text{if } i^{th} \text{ game is used} \end{cases}$$

```
> library(openintro)

# remove outliers
> ind <- which(mariokart$total_pr > 100)
> mariokart2 <- mariokart[-ind, ]

# create dummy variable
> mariokart2$cond_bin <- ifelse(mariokart2$cond == "new", 1, 0)
```

```
> lm1 <- lm(total_pr ~ cond_bin, data = mariokart2)
> summary(lm1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   42.871      0.814  52.668  < 2e-16 ***
cond_bin      10.900      1.258   8.662 1.06e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.371 on 139 degrees of freedom
Multiple R-squared:  0.3506,Adjusted R-squared:  0.3459
F-statistic: 75.03 on 1 and 139 DF,  p-value: 1.056e-14
```

```
> plot(total_pr ~ cond_bin, data = mariokart2, xaxt = "n",
      xlim = c(-0.1, 1.1), xlab = "Condition", ylab = "Total Price")
> axis(1, at = c(0,1), labels = c("Used (0)", "New (1)"))
> abline(lm1)
```



Interpretation

Equation for the least squares regression line from R summary:

$$\hat{y} = 42.87 + 10.9x$$

- The intercept $\hat{\beta}_0 = 42.87$ is the estimated average price when the game is in used condition (x = 0).
- $\hat{\beta}_0 + \hat{\beta}_1 = 42.87 + 10.9 = 53.77$ is the estimated average price when the game is in new condition (x = 1).
- The slope $\hat{\beta}_1 = 10.9$ indicates that, on average, a new game sells for \$10.9 dollars more than a used game.

We can also interpret a confidence interval for the slope β_1:

```
> confint(lm1)
                 2.5 %   97.5 %
(Intercept) 41.261713 44.48048
cond_bin     8.411621 13.38754
```

We are 95% confident that the average price of a new game is between \$8.41 and \$13.39 more than a used game.