Public Policy 529
Linear Regression
Part 3

Jonathan Hanson

Gerald R. Ford School of Public Policy
University of Michigan

November 27, 2023

Outline

1. Dichotomous Independent Variables

2. OLS Assumptions

Outline

Using Dichotomous (i.e. dummy) Variables

- Normally, regression involves interval-level variables, but dichotomous variables (0,1) can be used to represent a category.
- A value of 1 represents membership in the category of interest. Every other observation is coded as 0 for this variable.
- These variables are commonly-known as "dummy" variables because they stand in for the category.
- e.g. program participant (1), non-participant (0); Buddhist (1), non-Buddhist (0); retired (1), not-retired (0).

Interpreting Dummy Variables

$$y_i = \beta_0 + \beta_1 \text{Dummy}_i + u_i$$

- The intercept, β_0 is used to calculate the value of y for all members of the sample.
- Mathematically, when the dummy variable is 1, β_1 is relevant. When the dummy variable is 0, β_1 is cancelled out.
- Accordingly, β_1 represents the effect of membership in the category on the value of y.
- On average, members of the category would differ on y by the amount β_1. It is similar to a difference of means.

Example: Religion and Obama Thermometer

- The American National Election Study asks respondents whether religion is important to them (yes/no).
- Those who answer yes are coded 1; those who answer no are coded 0.
- We can use this variable as an independent variable.
- The dependent variable will be the Obama feeling thermometer.

Example: Religion and Obama Thermometer

$$\text{ObamaTherm}_i = \beta_0 + \beta_1 \text{ReligionImpt}_i + \epsilon_i$$

- The estimated coefficients can be used to construct two predicted Obama thermometer scores.
- Those who say religion is not important (ReligionImpt=0):

$$\begin{aligned} \widehat{\text{ObamaTherm}} &= \hat{\beta}_0 + \hat{\beta}_1(0) \\ &= \hat{\beta}_0 \end{aligned}$$

- Those who say religion is important (ReligionImpt=1):

$$\begin{aligned} \widehat{\text{ObamaTherm}} &= \hat{\beta}_0 + \hat{\beta}_1(1) \\ &= \hat{\beta}_0 + \hat{\beta}_1 \end{aligned}$$

Regression Output

```
. reg ObamaTherm ReligImpt
```

Source	SS	df	MS		
Model	25661.6165	1	25661.6165	Number of obs	= 5,464
Residual	6535017.09	5,462	1196.45132	F(1, 5462)	= 21.45
				Prob > F	= 0.0000
				R-squared	= 0.0039
				Adj R-squared	= 0.0037
Total	6560678.7	5,463	1200.92965	Root MSE	= 34.59

ObamaTherm	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
ReligImpt	-4.708207	1.016626	-4.63	0.000	-6.701199	-2.715216
_cons	63.92853	.8476965	75.41	0.000	62.26671	65.59035

The coefficient on the importance of religion variable is -4.7.

Interpretation: those who say religion is important view Obama less favorably on average.

Example: Predicted Effects

$$\widehat{\text{ObamaTherm}} = 63.9 - 4.7(\text{ReligionImpt})$$

Those who say religion is not important (ReligionImpt=0):

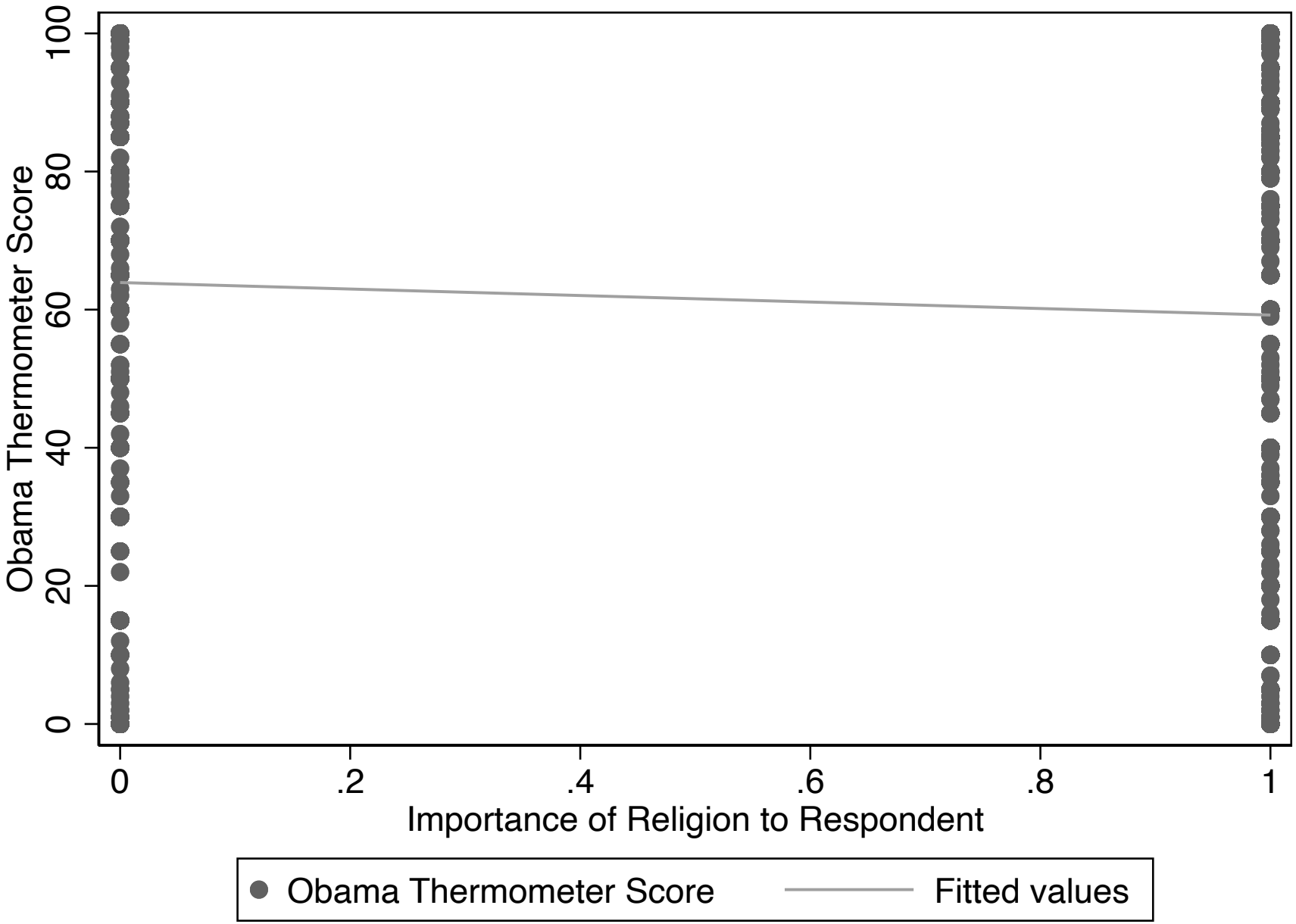
$$\begin{aligned} \widehat{\text{ObamaTherm}} &= 63.9 - 4.7(0) \\ &= 63.9 \end{aligned}$$

Those who say religion is important (ReligionImpt=1):

$$\begin{aligned} \widehat{\text{ObamaTherm}} &= 63.9 - 4.7(1) \\ &= 59.2 \end{aligned}$$

This is like a comparison of means. We can look at the statistical significance of $\hat{\beta}_1$ as a test of the difference.

Illustration



Comparison to a t-test

```
. ttest ObamaTherm, by(ReligImpt)

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
Not Impo	1,665	63.92853	.7623569	31.10753	62.43325	65.42381
Importan	3,799	59.22032	.5842278	36.00948	58.07489	60.36575
Combined	5,464	60.65501	.4688171	34.65443	59.73595	61.57408
diff		4.708207	1.016626		2.715216	6.701199

```
    diff = mean(Not Impo) - mean(Importan)                    t =   4.6312
H0: diff = 0                                     Degrees of freedom =     5462

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Note that you can find the coefficients for the bivariate regression model on this table, as well as the t-statistic for the coefficient on the religion variable.

Changing the Base Category

- With any dichotomous variable, the category coded as 0 serves as the base category in a regression.
- The estimated coefficient on the variable indicates the difference versus that base.
- If we switch which category is the base, the coefficient just flips its sign (+/-).
- Let's make the "religion is important" category be the base (0) and "religion is not important" be coded as 1.

Changing the Base Category

```
. reg ObamaTherm ReligNotImpt
```

Source	SS	df	MS	Number of obs	=	5,464
				F(1, 5462)	=	21.45
Model	25661.6165	1	25661.6165	Prob > F	=	0.0000
Residual	6535017.09	5,462	1196.45132	R-squared	=	0.0039
				Adj R-squared	=	0.0037
Total	6560678.7	5,463	1200.92965	Root MSE	=	34.59

ObamaTherm	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
ReligNotImpt	4.708207	1.016626	4.63	0.000	2.715216	6.701199
_cons	59.22032	.5611938	105.53	0.000	58.12016	60.32048

The coefficient on the importance of religion variable is now +4.7.

The substance of the interpretation does not change. Predicted values stay the same for each category.

Summary: When x_i is Binary

For $x_i = 0$ (the omitted group):

$$\begin{aligned} y_i &= \beta_0 + \beta_1(0) + u_i \\ y_i &= \beta_0 + u_i \\ \mathrm{E}[y_i|x_i = 0] &= \beta_0 \end{aligned}$$

For $x_i = 1$:

$$\begin{aligned} y_i &= \beta_0 + \beta_1(1) + u_i \\ y_i &= \beta_0 + \beta_1 + u_i \\ \mathrm{E}[y_i|x_i = 1] &= \beta_0 + \beta_1 \end{aligned}$$

Therefore, $\beta_1 = \mathrm{E}[y_i|x_i = 1] - \mathrm{E}[y_i|x_i = 0]$.

Adding Additional Variables

- Eventually, when you learn multiple regression, you can have several independent variables.
- By adding control variables, we can account for other factors that may produce a difference of means between two categories.
- Thus, regression can give us a better estimate of the difference of means.
- We can use multiple dummy variables to represent different values of a categorical variable.

Outline

1. Dichotomous Independent Variables

2. OLS Assumptions

Key OLS Assumptions

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- If certain assumptions hold, then Ordinary Least Squares is BLUE, the best linear unbiased estimator.
- There are other possible estimators, such as minimizing sum of absolute deviations between y_i and $\hat{y}_i$.
- The OLS estimator has the minimum variance among these estimators when the assumptions hold.

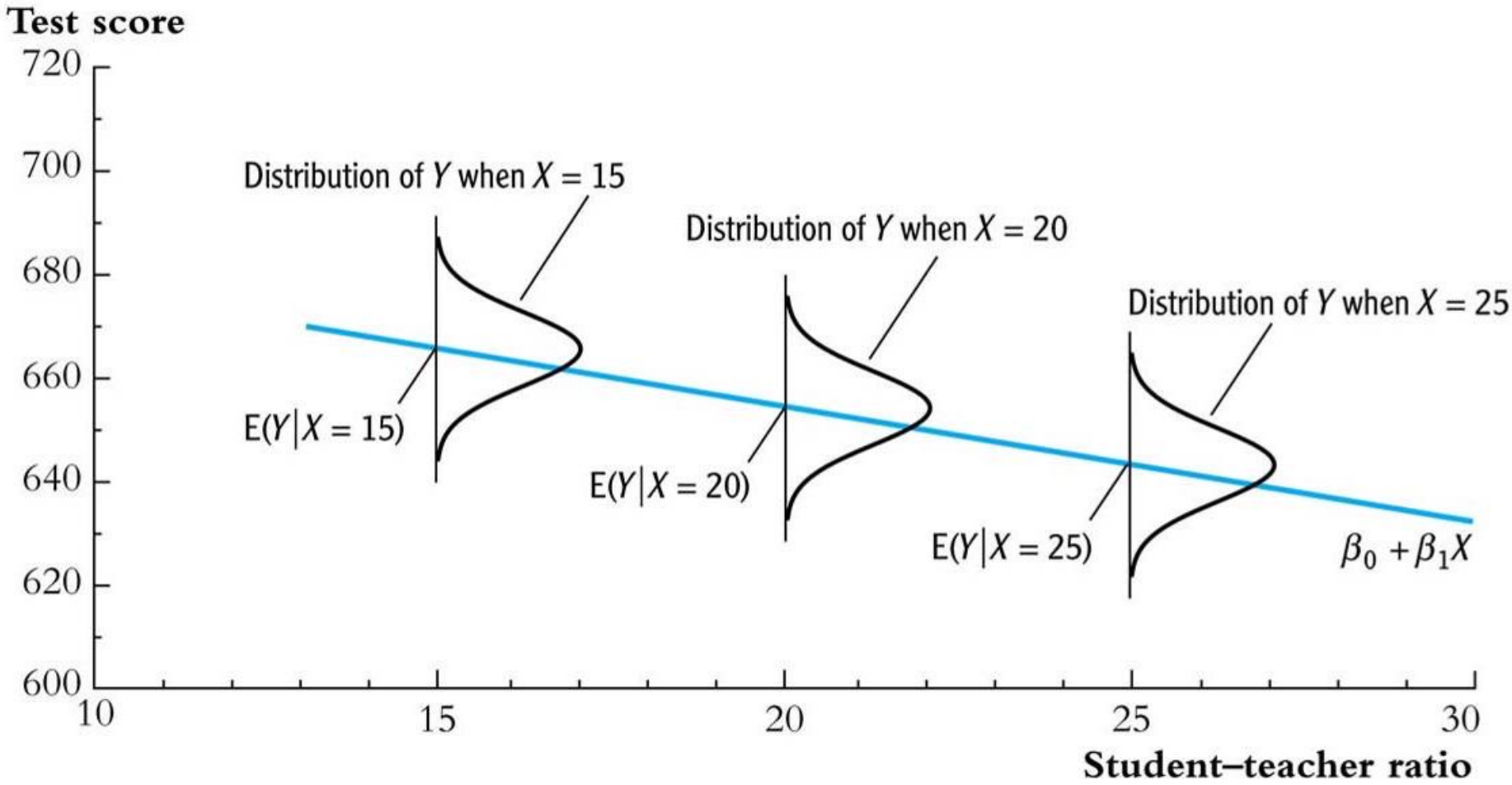
Assumption 1: Conditional Mean of u is 0

$$\mathrm{E}[u_i|x_i] = 0$$

- Given the value of x, u is uncorrelated with y. It is stochastic and contains no systematic information.
- In other words, there are no variables omitted from our model.
- If this assumption does not hold, our $\hat{\beta}_1$ likely is biased (and will be if x_i is correlated with the omitted variable).
- There is not an easy solution.

Illustration

u is uncorrelated with y given x



The value of y_i is $\beta_0 + \beta_1 x_i$ plus u_i, which is drawn from a normal distribution with mean 0.

Assumption 2: Data are i.i.d.

The data are independent and identically distributed (i.i.d.).

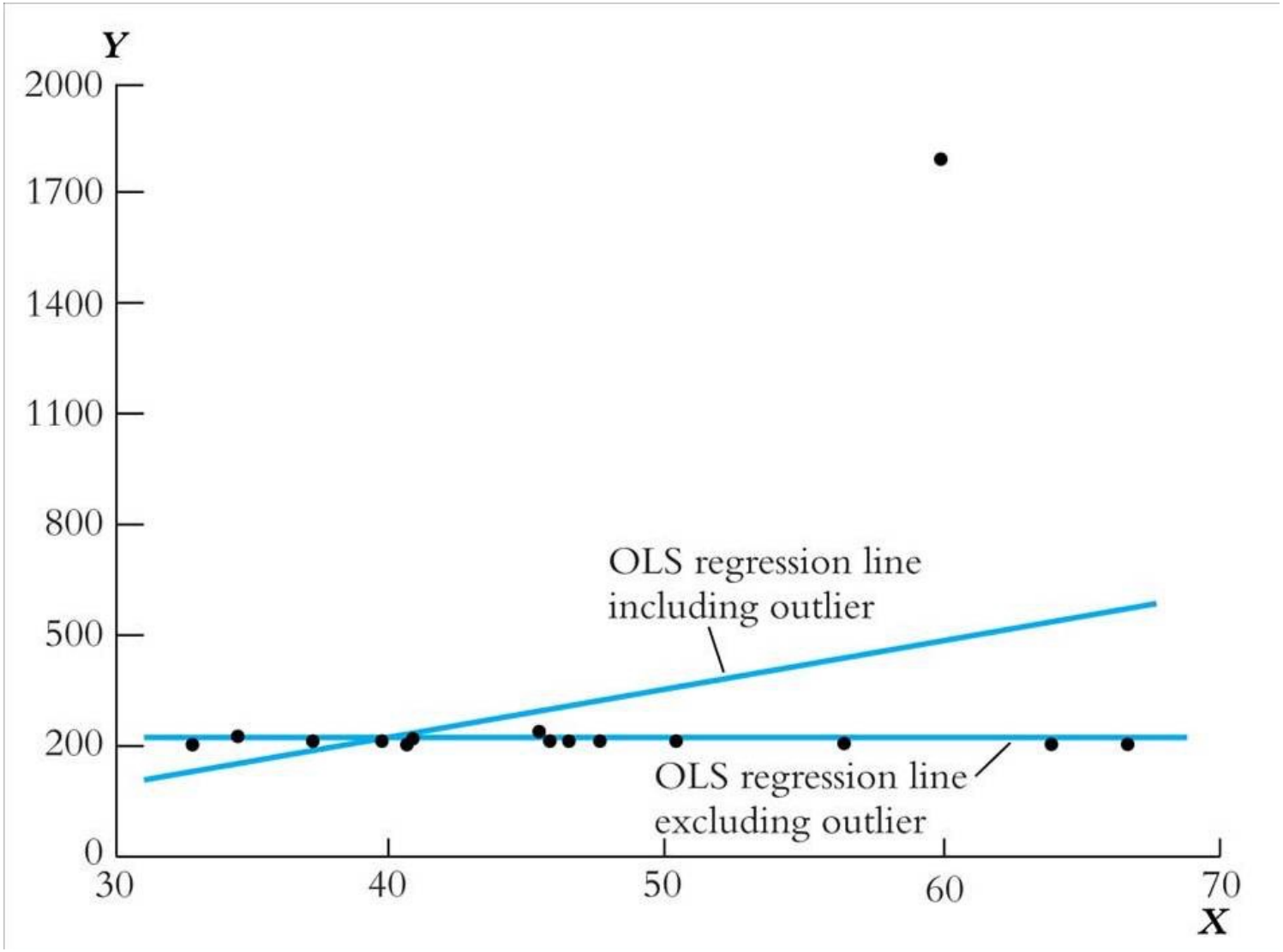
- Each observation is independent of all the others and they come from the same probability distribution.
- This is violated when there are connections between members of the sample: e.g. students in the same school, individuals in the same household.
- Time series data (such as the unemployment rate each month) also violate this assumption.
- It is also violated if, for example, each x_i does not come from a distribution with the same mean and variance.

Consequences of Violating Assumption 2

- Violations of the i.i.d. assumption can cause the standard errors for regression coefficients to be biased.
- There are straightforward solutions, however. We can use different formulas for the standard errors.
- e.g. formulas for "robust" or "clustered" standard errors, or corrections for serial correlation (over time) in the errors.
- We will learn about robust standard errors in the next lecture. Clustered standard errors will come in the next course.

Assumption 3: Large Outliers Unlikely

Values of x_i, y_i far outside the usual range of the data can produce misleading results.

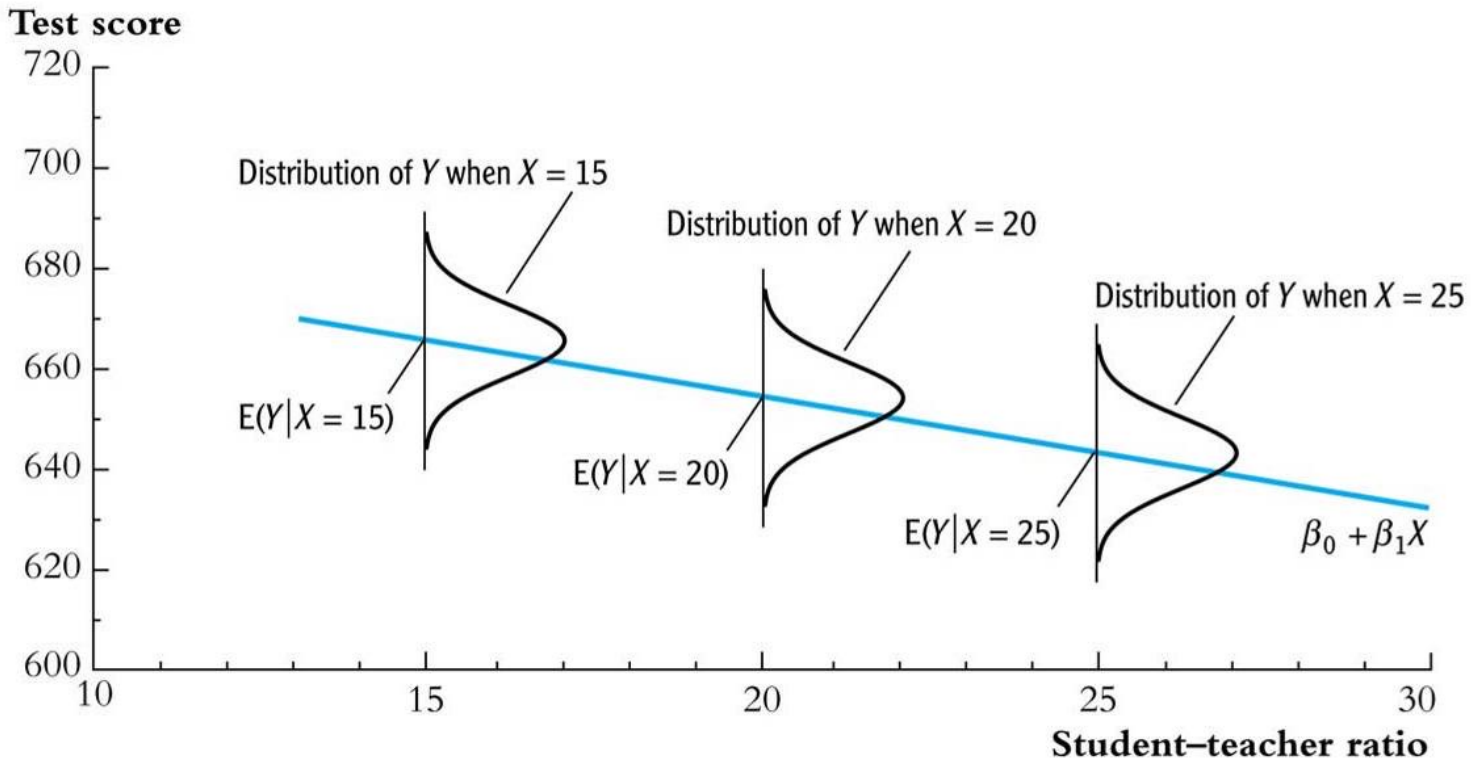


Dealing with Outliers

- Outliers are most problematic when y_i is unusual given x_i and x_i is away from its mean.
- To know whether it makes a difference for $\hat{\beta}_1$, we can drop the outlier and re-estimate the model.
- It may be tempting to just eliminate outliers from our sample, but we need good reason to do that.
- e.g. it may be an error in recording the data or the case does not belong in the sample.

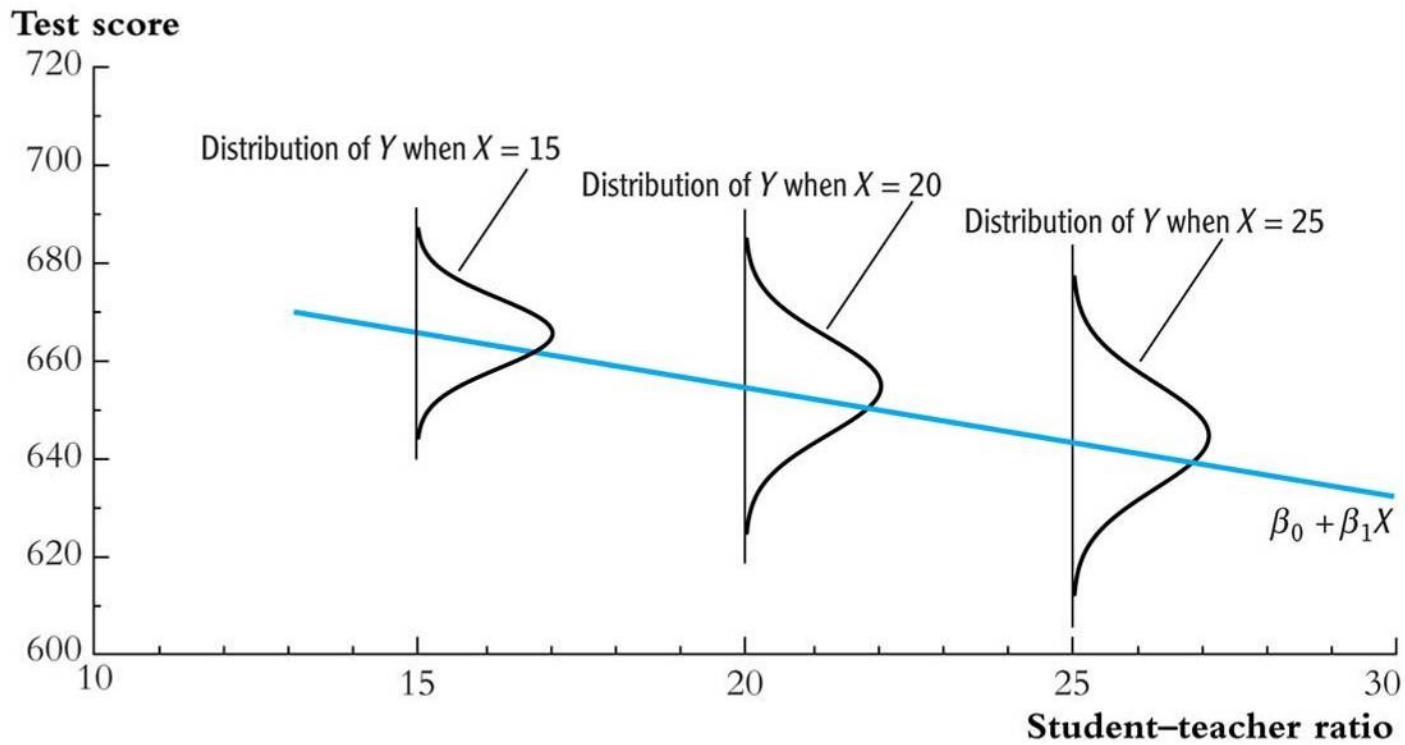
Assumption 4: The Error Term u is Homoskedastic

Homoskedasticity: the variance of u does not change with, or depend on, x.



Heteroskedasticity: A Violation of Assumption 4

Heteroskedasticity: the variance of u depends on x.



Here, the variance of u is greater when x is larger.

Consequences of Heteroskedasticity

- The estimator for $\hat{\beta}_1$ is still unbiased, but OLS is no longer efficient.
- The usual formula for the standard errors is incorrect. We get biased estimates of the standard errors.
- To adjust for this, we can use a formula for "robust" standard errors that are consistent despite heteroskedasticity.
- There is no harm to using robust standard errors. It is generally a good practice even if there is not heteroskedasticity.

Robust Standard Errors

- There are several formulas for robust standard errors. Stata and R use different formulas by default.
- Stata defaults to a formula called "HC1," while R defaults to a formula called "HC3."
- The HC3 standard errors are more conservative and may be preferable, especially when $n < 250$.
- There is little difference once $n \geq 500$.

Robust Standard Errors in Stata

- To use the default HC1 version of robust standard errors:

```
reg y x, robust
```

- To use the HC3 version of robust standard errors:

```
reg y x, vce(hc3)
```

- The second version will match the default robust standard errors in R.

Robust Standard Errors in R

- Install the `lmtest` and `sandwich` libraries if they are not present (this only needs to happen once):

```
install.packages(c("lmtest", "sandwich"))
```

- Once installed, these libraries need to be loaded:

```
library(lmtest)
library(sandwich)
```

- We can then apply robust standard errors to a model object:

```
coeftest(model, vcov = vcovHC(model))
```

- To match Stata's default of HC1 standard errors:

```
coeftest(model, vcov = vcovHC(model, type = "HC1"))
```

Example: Infant Mortality and Water Quality

```
. reg InfMort Water
```

Source	SS	df	MS
Model	152188.787	1	152188.787
Residual	60057.948	170	353.282047
Total	212246.735	171	1241.20897

Number of obs =	172
F(1, 170) =	430.79
Prob > F =	0.0000
R-squared =	0.7170
Adj R-squared =	0.7154
Root MSE =	18.796

InfMort	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Water	-1.537692	.0740864	-20.76	0.000	-1.68394	-1.391444
_cons	167.306	6.222146	26.89	0.000	155.0234	179.5886

```
. reg InfMort Water, vce(hc3)
```

Linear regression

Number of obs =	172
F(1, 170) =	247.85
Prob > F =	0.0000
R-squared =	0.7170
Root MSE =	18.796

InfMort	Coefficient	Robust HC3 std. err.	t	P>|t|	[95% conf.	interval]
Water	-1.537692	.0976724	-15.74	0.000	-1.730499	-1.344885
_cons	167.306	8.855522	18.89	0.000	149.8251	184.787

Example: Infant Mortality and Water Quality

```
> model <- lm(InfMort ~ Water, data = infmort_data)
> summary(model)

Call:
lm(formula = InfMort ~ Water, data = infmort_data)

Residuals:
    Min      1Q  Median      3Q     Max
-59.871  -9.143  -3.956   8.461  57.449

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 167.30601    6.22215   26.89   <2e-16 ***
Water        -1.53769    0.07409  -20.75   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> coeftest(model, vcov = vcovHC(model))

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 167.306006   8.855522  18.893 < 2.2e-16 ***
Water        -1.537692   0.097672 -15.743 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Summary: Key OLS Assumptions

1. Conditional mean of u is zero: $\mathrm{E}[u_i|x_i] = 0$.
2. Data are independent and identically distributed.
3. Large outliers are unlikely.
4. The error term (u) is homoskedastic.

It is important to understand these assumptions, the consequences when they are violated, and what can be done to address the violations.