Lecture 6: Multiple Linear Regression II

Module 2: part 2

Spring 2025

Logistics

- Part 2 of Module 2 today
- Assessment for Module 1 is due at 11:59pm tomorrow
- Categorical covariates, interaction terms, and part of transformations

Refresher

The population model we are trying to recover is

$$E(Y_i \mid \mathbf{X_i} = \mathbf{x}) = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$$

Refresher

The population model we are trying to recover is

$$E(Y_i \mid \mathbf{X_i} = \mathbf{x}) = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$$

- b_0 is the expected value of Y_i when **all** observed covariates are 0
- For $j \neq 0$, b_j is the difference in the expected value of Y_1 and Y_2 when $X_{1,j}$ and $X_{2,j}$ differ by 1 unit (i.e., $X_{1,j} - X_{2,j} = 1$), **but** the value of all other observed covariates are the same
- Can also include covariates which are non-linear in other covariates for polynomial regression

Categorical variables

So far, we've considered covariates which are continuous variables, but categorical variables can be included in regressions as well.

Example: The home price data set also include home style

[1]

[1] https://rethority.com/home-styles/

One Categorical Variable

Suppose I want to model housing prices based on house style (bungalow, craftsmen, cottage):

- Pick a reference category
- Create dummy variables (0 or 1) for all other categories

One Categorical Variable

Suppose I want to model housing prices based on house style (bungalow, craftsmen, cottage):

- Pick a reference category
- Create dummy variables (0 or 1) for all other categories

Suppose the style of homes in our data are: "bungalow," "craftsmen," and "cottage."

$$\text{Home Price}_i = b_0 + b_1 X_{i,\text{bungalow}} + b_2 X_{i,\text{craftsmen}}$$

$$X_{i,\text{bungalow}} = \begin{cases} 1 & \text{if House i is a ``bungalow''} \\ 0 & \text{if House i is not a ``bungalow''} \end{cases}$$

$$X_{i,\text{craftsmen}} = \begin{cases} 1 & \text{if House i is a ``craftsmen''} \\ 0 & \text{if House i is not a ``craftsmen''} \end{cases}$$

One Categorical Variable

If House i is a "cottage", then we would expect the price to be:

$$\text{Home Price}_i = b_0 + b_1 X_{i,\text{bungalow}} + b_2 X_{i,\text{craftsmen}} = b_0$$

If House i is a "bungalow", then we would expect the price to be:

$$\text{Home Price}_i = b_0 + b_1 X_{i,\text{bungalow}} + b_2 X_{i,\text{craftsmen}} = b_0 + b_1$$

If House i is a "craftsmen", then we would expect the price to be:

$$\text{Home Price}_i = b_0 + b_1 X_{i,\text{bungalow}} + b_2 X_{i,\text{craftsmen}} = b_0 + b_2$$

One Categorical Variable

If House i is a “cottage”, then we would expect the price to be:

$$\text{Home Price}_i = b_0 + b_1 X_{i,\text{bungalow}} + b_2 X_{i,\text{craftsmen}} = b_0$$

If House i is a “bungalow”, then we would expect the price to be:

$$\text{Home Price}_i = b_0 + b_1 X_{i,\text{bungalow}} + b_2 X_{i,\text{craftsmen}} = b_0 + b_1$$

If House i is a “craftsmen”, then we would expect the price to be:

$$\text{Home Price}_i = b_0 + b_1 X_{i,\text{bungalow}} + b_2 X_{i,\text{craftsmen}} = b_0 + b_2$$

The intercept b_0 indicates the expected price for the **reference category**.
b_1 and b_2 indicate the expected change in price (when compared to the reference category).

Categorical variables

Why do we not need a binary variable for all three styles? What would you use if the variable was two categories: Yes or No?

Categorical variables

Why do we not need a binary variable for all three styles? What would you use if the variable was two categories: Yes or No?

You would probably only include one variable which is either 0 (No) or 1 (Yes). You would not include two separate variables which are:

$$X_{i,\text{yes}} = \begin{cases} 1 & \text{if answer i is ``yes''} \\ 0 & \text{if answer i is not ``yes''} \end{cases}$$

$$X_{i,\text{no}} = \begin{cases} 1 & \text{if answer i is ``no''} \\ 0 & \text{if answer i is not ``no''} \end{cases}$$

Categorical variables

Consider the following model:

$$\text{Home Price}_i = b_0 + b_1 X_{i,\text{bungalow}} + b_2 X_{i,\text{craftsmen}} + b_3 X_{i,\text{cottage}}$$

Then the expected price for House i is:

$$\text{Home Price}_i = \begin{cases} b_0 + b_1 & \text{if House i is a "bungalow"} \\ b_0 + b_2 & \text{if House i is a "craftsman"} \\ b_0 + b_3 & \text{if House i is a "cottage"} \end{cases}$$

Categorical variables

Consider the following model:

$$\text{Home Price}_i = b_0 + b_1 X_{i,\text{bungalow}} + b_2 X_{i,\text{craftsmen}} + b_3 X_{i,\text{cottage}}$$

Then the expected price for House i is:

$$\text{Home Price}_i = \begin{cases} b_0 + b_1 & \text{if House i is a "bungalow"} \\ b_0 + b_2 & \text{if House i is a "craftsman"} \\ b_0 + b_3 & \text{if House i is a "cottage"} \end{cases}$$

If we set,

$$\tilde{b}_0 = b_0 + 5, \qquad \tilde{b}_1 = b_1 - 5, \qquad \tilde{b}_2 = b_2 - 5, \qquad \tilde{b}_3 = b_3 - 5$$

then the following model is:

$$\text{Home Price}_i = \begin{cases} \tilde{b}_0 + \tilde{b}_2 & \text{if House i is a "bungalow"} \\ \tilde{b}_0 + \tilde{b}_3 & \text{if House i is a "craftsman"} \\ \tilde{b}_0 + \tilde{b}_4 & \text{if House i is a "cottage"} \end{cases}$$

Categorical variables

Consider the following model:

$$\text{Home Price}_i = b_0 + b_1 X_{i,\text{bungalow}} + b_2 X_{i,\text{craftsmen}} + b_3 X_{i,\text{cottage}}$$

Then the expected price for House i is:

$$\text{Home Price}_i = \begin{cases} b_0 + b_1 & \text{if House i is a "bungalow"} \\ b_0 + b_2 & \text{if House i is a "craftsman"} \\ b_0 + b_3 & \text{if House i is a "cottage"} \end{cases}$$

If we set,

$$\tilde{b}_0 = b_0 + 5, \qquad \tilde{b}_1 = b_1 - 5, \qquad \tilde{b}_2 = b_2 - 5, \qquad \tilde{b}_3 = b_3 - 5$$

then the following model is:

$$\text{Home Price}_i = \begin{cases} \tilde{b}_0 + \tilde{b}_2 & \text{if House i is a "bungalow"} \\ \tilde{b}_0 + \tilde{b}_3 & \text{if House i is a "craftsman"} \\ \tilde{b}_0 + \tilde{b}_4 & \text{if House i is a "cottage"} \end{cases}$$

Too many parameters cause different parameter values to give the same underlying model

Multiple Categorical Variables

Suppose I want to model housing prices based on house style (bungalow, craftsmen, cottage) and home quality (High, medium, low):

- Pick a reference category **for each variable**: (cottage, low)
- For each variable, create dummy variables (0 or 1) for all other categories

$$\text{Home Price}_i = b_0 + b_1 X_{i,\text{bungalow}} + b_2 X_{i,\text{craftsmen}} + b_3 X_{i,\text{high}} + b_4 X_{i,\text{medium}}$$

Multiple Categorical Variables

$$\text{Home Price}_i = b_0 + b_1 X_{i,\text{bungalow}} + b_2 X_{i,\text{craftsmen}} + b_3 X_{i,\text{high}} + b_4 X_{i,\text{medium}}$$

When the house is a **low quality, cottage**, all dummy variables are 0 so

$$\text{Home Price}_i = b_0$$

When the house is a low quality bungalow:

$$\text{Home Price}_i = b_0 + b_1$$

When the house is a medium quality craftsmen:

$$\text{Home Price}_i = b_0 + b_2 + b_4$$

Multiple Categorical Variables

- b_0 indicates the value of y_i when **all** variables are set to the reference category
- Interpretation of other coefficients are the same: b_1 is the difference in the conditional expectation of the Home price between a home which is a bungalow compared to a home which is a cottage

Interaction Terms

Interaction terms

The price of a home generally decreases as the age increases, but does the rate of decrease depend on the quality of the construction?

Interaction terms

The price of a home generally decreases as the age increases, but does the rate of decrease depend on the quality of the construction?

Consider the following model where $\text{Quality}_i = 1$ if the quality is high and $\text{Quality}_i = 0$ if the quality is low:

$$\text{Home Price}_i = b_0 + b_1\text{Age}_i + b_2\text{Quality}_i + b_3(\text{Age}_i \times \text{Quality}_i)$$

Interaction terms

When House i is low quality, then the expected price is

$$\begin{aligned}\text{Home Price}_i &= b_0 + b_1\text{Age}_i + b_2\text{Quality}_i + b_3(\text{Age}_i \times \text{Quality}_i) \\ &= b_0 + b_1\text{Age}_i + b_2(0) + b_3(\text{Age}_i \times 0) \\ &= b_0 + b_1\text{Age}_i\end{aligned}$$

When House i is high quality, then the expected price is

$$\begin{aligned}\text{Home Price}_i &= b_0 + b_1\text{Age}_i + b_2\text{Quality}_i + \beta_3(\text{Age}_i \times \text{Quality}_i) \\ &= b_0 + b_1\text{Age}_i + b_2(1) + \beta_3(\text{Age}_i \times 1) \\ &= (b_0 + b_2) + (b_1 + b_3)\text{Age}_i\end{aligned}$$

where the intercept is $b_0 + b_2$ and the slope is $b_1 + b_3$

Interaction terms

- Product of two (or more) covariates is an **interaction term**

$$E(Y_i \mid \mathbf{X_i} = \mathbf{x}) = b_0 + b_1 x_1 + b_2 x_2 + b_3(x_1 \times x_2)$$

- This means the slope of a covariate changes depending on the value of another covariate
- For one continuous and one categorical variable, the interaction term corresponds to a different slope for each category
- When $x_2 = 0$ (i.e., the reference category)

$$E(Y_i \mid \mathbf{X_i} = \mathbf{x}) = b_0 + b_1 x_1$$

- When $X_{i,2} = 1$

$$E(Y_i \mid \mathbf{X_i} = \mathbf{x}) = (b_0 + b_2) + (b_1 + b_3) x_1$$

Interaction terms

- Product of two (or more) covariates is an **interaction term**

$$E(Y_i \mid \mathbf{X_i} = \mathbf{x}) = b_0 + b_1 x_1 + b_2 x_2 + b_3 (x_1 \times x_2)$$

- For an interaction of two continuous variable, the interaction term corresponds to a slope which depends on the value of other covariates

Interaction terms

- Product of two (or more) covariates is an **interaction term**

$$E(Y_i \mid \mathbf{X_i} = \mathbf{x}) = b_0 + b_1 x_1 + b_2 x_2 + b_3(x_1 \times x_2)$$

- For an interaction of two continuous variable, the interaction term corresponds to a slope which depends on the value of other covariates

- **Interpretation:** The slope for covariate 1 is $b_1 + b_3 x_2$

Interaction terms

- Product of two (or more) covariates is an **interaction term**

$$E(Y_i \mid \mathbf{X_i} = \mathbf{x}) = b_0 + b_1 x_1 + b_2 x_2 + b_3 (x_1 \times x_2)$$

- For an interaction of two continuous variable, the interaction term corresponds to a slope which depends on the value of other covariates

Interaction terms

- Product of two (or more) covariates is an **interaction term**

$$E(Y_i \mid \mathbf{X_i} = \mathbf{x}) = b_0 + b_1 x_1 + b_2 x_2 + b_3(x_1 \times x_2)$$

- For an interaction of two continuous variable, the interaction term corresponds to a slope which depends on the value of other covariates

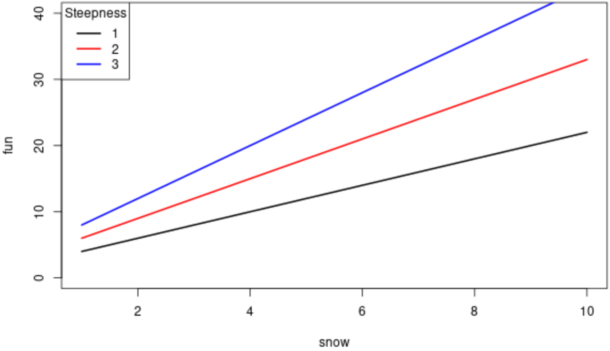
- **Interpretation:**
The difference in $E(Y_i \mid \mathbf{X_i})$ and $E(Y_k \mid \mathbf{X_k})$ if
    - $X_{i,1} - X_{k,1} = 1$ (differ by one unit in covariate 1)
    - $X_{i,2} = X_{k,2}$ (have the same value of covariate 2)

$$\begin{aligned} E(Y_i \mid \mathbf{X_i}) - E(Y_k \mid \mathbf{X_k}) &= [b_0 + b_1 X_{i,1} + b_2 X_{i,2} + b_3(X_{i,1} \times X_{i,2})] \\ &\quad - [b_0 + b_1 X_{k,1} + b_2 X_{k,2} + b_3(X_{k,1} \times X_{k,2})] \\ &= b_1(X_{i,1} - X_{k,1}) + b_3(X_{i,1}X_{i,2} - X_{k,1}X_{k,2}) \\ &= b_1(X_{i,1} - X_{k,1}) + b_3(X_{i,1}X_{i,2} - X_{k,1}X_{i,2}) \\ &= b_1(X_{1,1} - X_{2,1}) + b_3 X_{i,2}(X_{i,1} - X_{k,1}) \\ &= b_1 + b_3 X_{i,2} \end{aligned}$$

Example: interaction term

Suppose I'm interested in how much fun to expect while sledding

$$E(\text{Fun} \mid Snow, Steep) = b_0 + b_1\text{snow} + b_2\text{steep} + b_3\text{snow} \times \text{steep}$$



Interaction terms

$$E(Y_i \mid \mathbf{X_i} = \mathbf{x}) = b_0 + b_1 x_1 + b_2 x_2 + b_3(x_1 \times x_2)$$

- b_1 and b_2 are the **main effects** of $X_{i,1}$ and $X_{i,2}$
- b_3 is the interaction term of $X_{i,1}$ and $X_{i,2}$
- We almost always include the main effect if we include an interaction effect
- Becomes difficult to interpret the interaction term without the main effect

Interaction terms

$$E(Y_i \mid \mathbf{X_i} = \mathbf{x}) = b_0 + b_1 x_1 + b_2 x_2 + b_3(x_1 \times x_2)$$

- b_1 and b_2 are the **main effects** of $X_{i,1}$ and $X_{i,2}$
- b_3 is the interaction term of $X_{i,1}$ and $X_{i,2}$
- We almost always include the main effect if we include an interaction effect
- Becomes difficult to interpret the interaction term without the main effect

Discuss a potential problem of interest where including interaction terms may be useful

Transformations

Transformations

We transformed x by taking a square, but we can use other transformations

- Most common transform is log transform
- Sometimes $1/y$ or $\sqrt{y}$ is also used
- Can transform covariates

$$E(Y \mid X = x) = b_0 + b_1 \log(x)$$

Transformations

We transformed x by taking a square, but we can use other transformations

- Most common transform is log transform
- Sometimes $1/y$ or $\sqrt{y}$ is also used
- Can transform covariates
- Can transform dependent variable

$$E(\log(Y) \mid X = x) = b_0 + b_1 x$$

Transformations

We transformed x by taking a square, but we can use other transformations

- Most common transform is log transform
- Sometimes $1/y$ or $\sqrt{y}$ is also used
- Can transform covariates
- Can transform dependent variable
- Can transform dependent variable and covariates

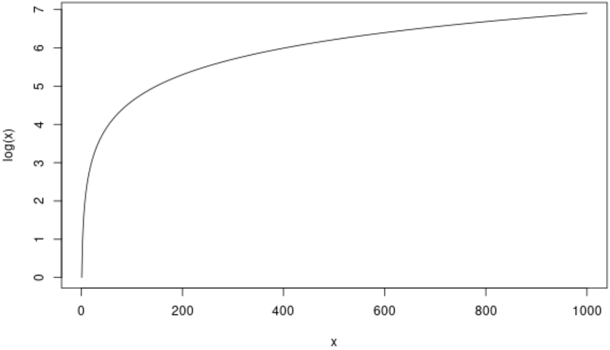
$$E(\log(Y) \mid X = x) = b_0 + b_1 \log(x)$$

Transformations

- Fitting a linear model with transformed data is conceptually the same
- Just “plug-in” transformed data
- Careful about interpretation!

Properties of log

$$e \approx 2.718$$
$$\log_e(x) = a \Leftrightarrow e^a = \exp(a) = x$$
$$\log_e(xy) = \log_e(x) + \log_e(y)$$
$$\log_e(x/y) = \log_e(x) - \log_e(y)$$



Interpretation when transforming covariates

$$Y_i = b_0 + b_1 \log_e(X_i) + \varepsilon_i$$

$$E(Y_i \mid X_i = x) = b_0 + b_1 \log_e(x)$$

- Suppose $X_j = 1.01X_i$ so X_j is 1% larger than X_i

$$\begin{aligned} E(Y_j \mid X_j = x1.01) - E(Y_j \mid X_j = x) &= b_0 + b_1 \log(X_j) - [b_0 + b_1 \log(X_i)] \\ &= b_1 \log(1.01X_i) - b_1 \log(X_i) \\ &= b_1 \log\left(\frac{1.01X_i}{X_i}\right) \\ &= b_1 \log(1.01) \approx b_1 \times .01 \end{aligned}$$

Interpretation when transforming covariates

$$Y_i = b_0 + b_1 \log_e(X_i) + \varepsilon_i$$

$$E(Y_i \mid X_i = x) = b_0 + b_1 \log_e(x)$$

- Suppose $X_j = 1.01X_i$ so X_j is 1% larger than X_i

$$\begin{aligned} E(Y_j \mid X_j = x1.01) - E(Y_j \mid X_j = x) &= b_0 + b_1 \log(X_j) - [b_0 + b_1 \log(X_i)] \\ &= b_1 \log(1.01X_i) - b_1 \log(X_i) \\ &= b_1 \log\left(\frac{1.01X_i}{X_i}\right) \\ &= b_1 \log(1.01) \approx b_1 \times .01 \end{aligned}$$

- Interpretation of b_1: Two observations with x which differ by 1% have expected values of Y which differ by $b_1 \log(1.01)$

Transforming dependent variable

$$\log_e(Y_i) = b_0 + b_1 X_i + \varepsilon_i$$

$$\begin{aligned} e^{\log_e(Y_i)} &= e^{b_0 + b_1 X_i + \varepsilon_i} \\ Y_i &= e^{b_0} \times e^{b_1 X_i} \times e^{\varepsilon_i} \end{aligned}$$

Transforming dependent variable

$$\log_e(Y_i) = b_0 + b_1 X_i + \varepsilon_i$$

$$\begin{aligned} e^{\log_e(Y_i)} &= e^{b_0 + b_1 X_i + \varepsilon_i} \\ Y_i &= e^{b_0} \times e^{b_1 X_i} \times e^{\varepsilon_i} \end{aligned}$$

$$\begin{aligned} E(Y_i \mid X_i = x) &= E(e^{b_0} \times e^{b_1 X_i} \times e^{\varepsilon_i} \mid X_i = x) \\ &= E(e^{b_0} \times e^{b_1 x} \times e^{\varepsilon_i} \mid X_i = x) \\ &= e^{b_0} \times e^{b_1 x} \times E(e^{\varepsilon_i}) \end{aligned}$$

Transforming dependent variable

In general,

$$E(e^X) \neq e^{E(X)}$$

Obs	X	e^X
1	1	2.72
2	2	7.39
3	5	148.41
4	-2	0.14
Avg	1.5	39.66

$$e^{1.5} = 4.48 \neq 39.66$$

Transforming dependent variable

$$\log_e(Y_i) = b_0 + b_1 X_i + \varepsilon_i$$

$$e^{\log_e(Y_i)} = e^{b_0 + b_1 X_i + \varepsilon_i}$$
$$Y_i = e^{b_0} \times e^{b_1 X_i} \times e^{\varepsilon_i}$$

$$\begin{aligned}
E(Y_i \mid X_i = x) &= E(e^{b_0} \times e^{b_1 X_i} \times e^{\varepsilon_i} \mid X_i = x) \\
&= E(e^{b_0} \times e^{b_1 x} \times e^{\varepsilon_i} \mid X_i = x) \\
&= e^{b_0} \times e^{b_1 x} \times E(e^{\varepsilon_i}) \\
&\neq e^{b_0} \times e^{b_1 x} \qquad \text{(because } E(e^{\varepsilon_i}) \neq 1\text{)}
\end{aligned}$$

Transforming dependent variable

$$\log_e(Y_i) = b_0 + b_1 X_i + \varepsilon_i$$

$$e^{\log_e(Y_i)} = e^{b_0 + b_1 X_i + \varepsilon_i}$$

$$Y_i = e^{b_0} \times e^{b_1 X_i} \times e^{\varepsilon_i}$$

$$\begin{aligned}
E(Y_i \mid X_i = x) &= E(e^{b_0} \times e^{b_1 X_i} \times e^{\varepsilon_i} \mid X_i = x) \\
&= E(e^{b_0} \times e^{b_1 x} \times e^{\varepsilon_i} \mid X_i = x) \\
&= e^{b_0} \times e^{b_1 x} \times E(e^{\varepsilon_i}) \\
&\neq e^{b_0} \times e^{b_1 x} \qquad (\text{because } E(e^{\varepsilon_i}) \neq 1)
\end{aligned}$$

The median of e^{ε_i} is 1 if ε_i is symmetric, so we can say that $e^{b_0} \times e^{b_1 x}$ is the median of Y_i conditional on $X_i = x$

Transforming dependent variable

- Suppose $X_j - X_i = 1$
- Comparing $E(Y_i \mid X_i = x)$ and $E(Y_j \mid X_j = x + 1)$, we have

$$\frac{E(Y_j \mid X_j = x+1)}{E(Y_i \mid X_i = x)} = \frac{e^{b_0} \times e^{b_1(x+1)} \times E(e^{\varepsilon_i})}{e^{b_0} \times e^{b_1 x} \times E(e^{\varepsilon_i})}$$
$$= e^{b_1}$$

- Interpretation of b_1: Observations which differ in 1 unit of x have a e^{b_1}-fold difference in the expected value of Y

Transforming dependent variable

- Suppose $X_j - X_i = 1$
- Comparing $E(Y_i \mid X_i = x)$ and $E(Y_j \mid X_j = x + 1)$, we have

$$\begin{aligned}\frac{E(Y_j \mid X_j = x+1)}{E(Y_i \mid X_i = x)} &= \frac{e^{b_0} \times e^{b_1(x+1)} \times E(e^{\varepsilon_i})}{e^{b_0} \times e^{b_1 x} \times E(e^{\varepsilon_i})} \\ &= e^{b_1}\end{aligned}$$

- Interpretation of b_1: Observations which differ in 1 unit of x have a e^{b_1}-fold difference in the expected value of Y

$$100 \times \left(\frac{E(Y_j \mid X_j = x+1)}{E(Y_i \mid X_i = x)} - 1 \right) = 100 \times (e^{b_1} - 1)$$

- Interpretation of b_1: Observations which differ in 1 unit of x have a $100(e^{b_1} - 1)$ percentage difference in the expected value of Y

Transforming both dependent variable and covariates

$$\log(y_i) = b_0 + b_1 \log(X_i) + \varepsilon_i$$

$$e^{\log_e(Y_i)} = e^{b_0 + b_1 \log(X_i) + \varepsilon_i}$$
$$Y_i = e^{b_0} \times e^{b_1 \log(X_i)} \times e^{\varepsilon_i}$$

- Suppose $X_j = 1.01X_i$ so X_j is 1% larger than X_i
- Comparing $E(Y_i \mid X_i = x)$ and $E(Y_j \mid X_j = 1.01x)$, we have

$$\begin{aligned}
\frac{E(Y_j \mid X_j = 1.01x)}{E(Y_i \mid X_i = x)} &= \frac{e^{b_0} \times e^{b_1 \log(1.01x)} \times E(e^{\varepsilon_i})}{e^{b_0} \times e^{b_1 \log(x)} \times E(e^{\varepsilon_i})} \\
&= \frac{e^{b_1 \log(1.01x)}}{e^{b_1 \log(x)}} = e^{b_1 \log(1.01x/x)} \\
&= \left(e^{\log(1.01)}\right)^{b_1} = 1.01^{b_1}
\end{aligned}$$

Transforming both dependent variable and covariates

$$\log(Y_i) = b_0 + b_1 \log(X_i) + \varepsilon_i$$

- Comparing $E(Y_i \mid X_i = x)$ and $E(Y_j \mid X_j = 1.01x)$, we have

$$\frac{E(Y_j \mid X_j = 1.01x)}{E(Y_i \mid X_i = x)} = 1.01^{b_1}$$

Transforming both dependent variable and covariates

$$\log(Y_i) = b_0 + b_1 \log(X_i) + \varepsilon_i$$

- Comparing $E(Y_i \mid X_i = x)$ and $E(Y_j \mid X_j = 1.01x)$, we have

$$\frac{E(Y_j \mid X_j = 1.01x)}{E(Y_i \mid X_i = x)} = 1.01^{b_1}$$

- Interpretation of b_1: Observations which differ in x by 1% have a 1.01^{b_1} **fold** difference in the expected value of Y
- Interpretation of b_1: Observations which differ in x by 1% have a $100(1.01^{b_1} - 1)$ percentage difference in the expected value of Y