L.5: Regression

Econometrics 1: ver. 2024 Fall Semester

Naoki Awaya

Regression Analysis: Introduction

Regression Analysis: Introduction

- `state.x77`: an R built-in data on economic and social characteristics of 50 states in the US as of 1977[1]

```
>    data <- as.data.frame(state.x77)
>    head(data, 4)
```

```
##            Population Income Illiteracy Life Exp Murder HS Grad Fr
## Alabama          3615   3624        2.1    69.05   15.1    41.3
## Alaska            365   6315        1.5    69.31   11.3    66.7
## Arizona          2212   4530        1.8    70.55    7.8    58.1
## Arkansas         2110   3378        1.9    70.66   10.1    39.9
```

```
>    dim(data)
```

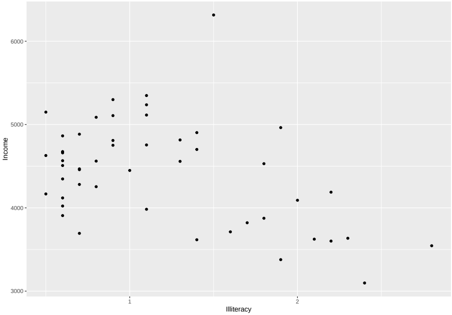
```
## [1] 50  8
```

[1] Source: U.S. Department of Commerce, Bureau of the Census (1977)

Regression Analysis: Introduction

- Suppose we would like to know how illiteracy affects income (note: `Income` is a per-capita income).
- Firstly, we visually check the relationship between these variables:

```
> library(tidyverse)
> ggplot(data, aes(x = Illiteracy, y = Income)) + geom_point()
```



Regression Analysis: Introduction

- From this figure, we can observe that there is a negative relationship between illiteracy and income, as expected.
- Indeed, the (sample) correlation coefficient is

```
>   with(data, cor(Illiteracy, Income))
```

```
## [1] -0.4370752
```

- The above code is equivalent to

`cor(data$Illiteracy, data$Income).`

Using `with()` function, we can omit the "`data$`" part.

Regression Analysis: Introduction

How much does an additional 1 point increase in illiteracy rate decrease the state's income level, on average?

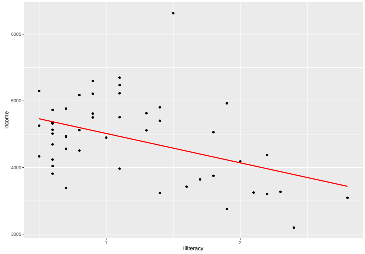
- Here, we assume a "model" that depicts the relationship between the variables (i.e., a regression model). For example,

$$\text{Income} = \beta_0 + \beta_1 \text{Illiteracy} + \epsilon$$

- Under this model, `Illiteracy` and `Income` have a linear relationship.
- Since it is generally impossible that all observations are exactly on the straight line, we need an "error" term ϵ for adjustment.
- We can answer to the above question by estimating β_1 from the data: $\text{Income} + \beta_1 = \beta_0 + \beta_1(\text{Illiteracy} + 1) + \epsilon$

Regression Analysis: Introduction

- We can estimate β_1 by finding the best-fitting line to the data:



- The estimated regression model (the red line in the figure) is

$$\text{Income} = 4951.3 - 440.6\text{Illiteracy} + \text{error}$$

- Thus, we can conclude that, on average

1% increase in illiteracy $\approx$ \$440.6 decrease in per-capita income.

Formal Definition of Regression

Formal Definition of Regression

- Outcome variable of interest dependent variable
- Variables that determine the value of the dependent variable explanatory variables (also referred to as "independent variables" or simply "regressors")
- Let Y denote a dependent variable and $\mathbf{X} = (X_1, \ldots, X_k)$ denote a set of explanatory variables.
- The purpose of regression analysis is

> to estimate a function $g(\cdot)$ of $\mathbf{X}$ that predicts the value of Y.

The function

$$g(\cdot) : \mathbf{X} \rightarrow \text{predicted value of } Y$$

is called the regression function.

Formal Definition of Regression

Simple linear regression model

- Linear regression model with a single explanatory variable:

$$Y = \beta_0 + X\beta_1 + \varepsilon$$

- Y: dependent variable, X: explanatory variable, and ε: error term.
- β_0: intercept, and β_1: regression coefficient (slope parameter) of X. These are parameters of interest to be estimated.

Multiple linear regression model

- Linear regression model with multiple explanatory variables:

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_k\beta_k + \varepsilon$$

- β_0: intercept, and $(\beta_1, \ldots, \beta_k)$: coefficients.

Formal Definition of Regression

Example 1.

- A linear regression model of annual income:

$$\text{Income} = \beta_0 + \text{Experience}\beta_1 + \text{Hours}\beta_2 + \text{Education}\beta_3 + \varepsilon$$

- For example, coefficient β_1 tells us

how much an additional year of working experience increases income

- More formally,

$$\frac{\partial \text{Income}}{\partial \text{Experience}} = \beta_1$$

Thus, β_1 corresponds to the marginal effect of Experience variable on Income.

Formal Definition of Regression

Example 2.

- A randomized experiment with a binary treatment:

$$\text{Outcome} = \beta_0 + X\beta_1 + \varepsilon,$$

  where $X \in \{0, 1\}$.

- Assume that $\mathbb{E}[\varepsilon|X] = 0$. <= Randomly assigning the treatment ensures this assumption.
- Then, the average treatment effect (ATE) is

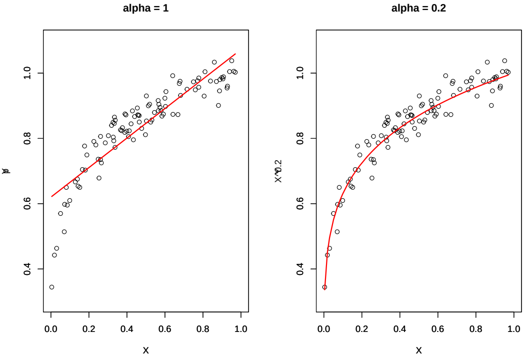
$$\begin{aligned}\mathbb{E}[\text{Outcome}|X = 1] - \mathbb{E}[\text{Outcome}|X = 0] &= (\beta_0 + \beta_1) - \beta_0 \\ &= \beta_1.\end{aligned}$$

- "Linear" regression is a regression analysis based on a linear regression function:

$$g(\mathbf{X}) = \beta_0 + X_1\beta_1 + \cdots + X_k\beta_k.$$

- One may consider a more general "nonlinear" regression function: e.g.,

$$g(\mathbf{X}) = (\beta_0 + X_1\beta_1 + \cdots + X_k\beta_k)^{\alpha}.$$



Conditional Expectation Function

Conditional Expectation Function

What is the theoretically best choice for the regression function?
$\Rightarrow$ conditional expectation function

- Let Y be a dependent variable, and $\mathbf{X} = (X_1, \ldots, X_k)$ be a set of explanatory variables.
- Let $f_{Y|\mathbf{X}}(y|\mathbf{X} = \mathbf{x})$ be the conditional probability density function of Y given $\mathbf{X} = \mathbf{x}$.
- Then, the conditional expectation of Y given $\mathbf{X} = \mathbf{x}$ is

$$\mathbb{E}[Y|\mathbf{X} = \mathbf{x}] = \int_{-\infty}^{\infty} y f_{Y|\mathbf{X}}(y|\mathbf{X} = \mathbf{x}) dy$$

This is the expected value of Y for those satisfying $\mathbf{X} = \mathbf{x}$.

Conditional Expectation Function

- The value of $\mathbb{E}[Y|\mathbf{X}=\mathbf{x}]$ can vary with the value of $\mathbf{x}$.
  $\Longrightarrow \mathbb{E}[Y|\mathbf{X}=\mathbf{x}]$ is the value obtained by plugging $\mathbf{x}$ into $\mathbb{E}[Y|\mathbf{X}]$.

Conditional Expectation Function

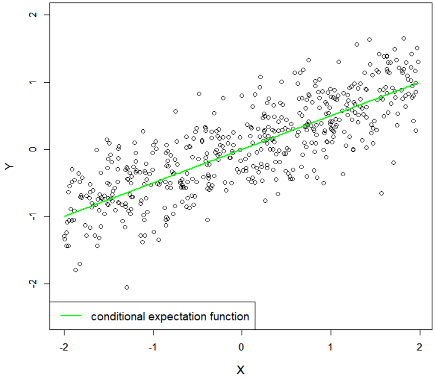
The function $m(\cdot)$ defined as follows

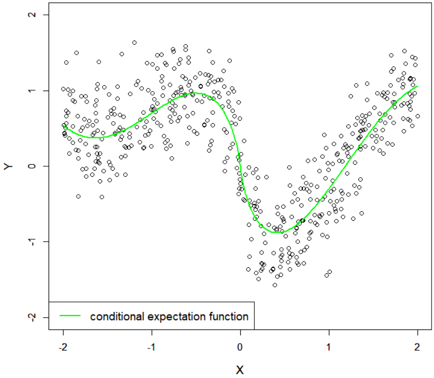
$$m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X}=\mathbf{x}]$$

is called the conditional expectation function (also referred to as conditional mean function) of Y.

Recall that if Y and $\mathbf{X}$ are independent, $\mathbb{E}[Y|\mathbf{X}] = \mathbb{E}[Y]$ holds. Thus, $m(\cdot)$ is a constant function, meaning that $\mathbf{X}$ does not affect Y on average.

Conditional Expectation Function





Law of Iterated Expectations

- Note that since $m(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}]$ is a function of random variable $\mathbf{X}$, $m(\mathbf{X})$ is a random variable, and we can consider its expectation.

Law of Iterated Expectations

The following result is known as the law of iterated expectations:

$$\begin{aligned}\mathbb{E}[Y] &= \mathbb{E}[m(\mathbf{X})] \\ &= \mathbb{E}[\mathbb{E}(Y|\mathbf{X})]\end{aligned}$$

- For the right-hand side term, the "inner expectation" is the conditional expectation of Y given $\mathbf{X}$, and the "outer expectation" is the expectation with respect to $\mathbf{X}$.

Law of Iterated Expectations

- When $\mathbf{X}$ is a set of discrete random variables, the LIE can be restated as
$$\mathbb{E}[Y] = \sum_{\ell=1}^{L} \mathbb{E}[Y|\mathbf{X} = \mathbf{x}_\ell] \Pr(\mathbf{X} = \mathbf{x}_\ell)$$
- That is, $\mathbb{E}[Y]$ is a weighted average of the group-wise means $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}_\ell]$, where each weight is the ratio of the group $\mathbf{x}_\ell$.
- For example
$$\mathbb{E}[\text{height}] = \mathbb{E}[\text{height}|\text{male}] \Pr(\text{male}) + \mathbb{E}[\text{height}|\text{female}] \Pr(\text{female})$$
- For the proof of the LIE for continuous variables, see the appendix.

Best Regression Function

Best Regression Function

- Let $g(\cdot)$ be any candidate regression function, and $e(\mathbf{X})$ be the corresponding prediction error of Y at $\mathbf{X}$; namely

$$e(\mathbf{x}) = Y - g(\mathbf{x}).$$

- It is natural to think that the expectation of prediction error should be zero for an ideal regression function:

$$\mathbb{E}[e(\mathbf{X})] = 0. \quad (\text{Unbiasedness})$$

- When we set $g(\cdot)$ to the conditional expectation function $m(\cdot)$, by LIE

$$\begin{aligned}\mathbb{E}[e(\mathbf{X})] = \mathbb{E}[Y - m(\mathbf{X})] &= \mathbb{E}[Y] - \mathbb{E}[\mathbb{E}(Y|\mathbf{X})] \\ &= \mathbb{E}[Y] - \mathbb{E}[Y] = 0.\end{aligned}$$

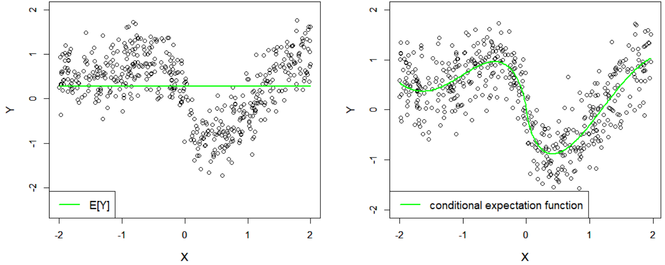
  Thus, the conditional expectation function meets this criterion.

Best Regression Function

- However, the conditional expectation function is not the only function that satisfies $\mathbb{E}[e(\mathbf{X})] = 0$.
- For example, a constant function $g(\mathbf{X}) = \mathbb{E}[Y]$ satisfies this:

$$e(\mathbf{X}) = \mathbb{E}[Y - \mathbb{E}(Y)] = 0$$

- However, using a constant regression function is not appropriate for the purpose of regression analysis.



Best Regression Function

- The "best" regression function should be not only unbiased but also have the smallest variance (variance $=$ the risk of wrong prediction).
- That is, we consider minimizing $\mathbb{E}[e^2(\mathbf{X})]$, the so-called MSE (mean squared error).

Best Regression Function

- For any candidate regression function $g(\cdot)$ and the conditional expectation function $m(\cdot)$, observe that

$$
\begin{aligned}
\text{MSE } \mathbb{E}(e^2(\mathbf{X})) &= \mathbb{E}[(Y - g(\mathbf{X}))^2] \\
&= \mathbb{E}[(Y - m(\mathbf{X}) + m(\mathbf{X}) - g(\mathbf{X}))^2] \\
&= \mathbb{E}[\underbrace{\{Y - m(\mathbf{X})\}^2}_{=I_1}] + 2\mathbb{E}[\underbrace{\{Y - m(\mathbf{X})\}\{m(\mathbf{X}) - g(\mathbf{X})\}}_{=I_2}] \\
&\quad + \mathbb{E}[\underbrace{\{m(\mathbf{X}) - g(\mathbf{X})\}^2}_{=I_3}] \\
&= \mathbb{E}(I_1) + 2\mathbb{E}(I_2) + \mathbb{E}(I_3)
\end{aligned}
$$

- $\mathbb{E}(I_1)$ is independent on the choice of $g(\cdot)$, and thus can be ignored.

Best Regression Function

(cont.)

- In addition, note that $\mathbb{E}(I_2) = 0$, because

$$\begin{aligned}\mathbb{E}[I_2|\mathbf{X}] &= \mathbb{E}[\{Y - m(\mathbf{X})\}\{m(\mathbf{X}) - g(\mathbf{X})\}|\mathbf{X}] \\ &= \underbrace{\{\mathbb{E}[Y|\mathbf{X}] - m(\mathbf{X})\}}_{=0}\{m(\mathbf{X}) - g(\mathbf{X})\} = 0\end{aligned}$$

  and by LIE

$$\mathbb{E}(I_2) = \mathbb{E}[\mathbb{E}(I_2|\mathbf{X})] = \mathbb{E}[0] = 0$$

- Thus, the first two components $\mathbb{E}(I_1)$ and $2\mathbb{E}(I_2)$ of the MSE cannot be made smaller by manipulating the form of $g(\cdot)$.

Best Regression Function

(cont.)

- Consequently, the minimizer of the MSE is a function $g(\cdot)$ that minimizes
$$\mathbb{E}(I_3) = \mathbb{E}[\{m(\mathbf{X}) - g(\mathbf{X})\}^2]$$
- Clearly, it is only when $g(\cdot) = m(\cdot)$ that $\mathbb{E}(I_3)$ is minimized.

$\Longrightarrow$

The best regression function is the conditional expectation function

(in terms of MSE minimization).

Model (Mis)specification

Linear Regression Function

- Among many possible regression models, the linear regression is the most commonly employed in both theoretical and applied research.
- The linear regression model is based on the assumption that the conditional expectation function is linear:

$$\mathbb{E}[Y|\mathbf{X} = \mathbf{x}] = \beta_0 + x_1\beta_1 + \cdots + x_k\beta_k$$

- However, this assumption is very restrictive in general.

> *"All models are wrong, but some are useful".* George E.P. Box.

Model Misspecification

- Without loss of generality, any regression model can be expressed as

$$Y = m(\mathbf{X}) + \epsilon$$

where $m(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}]$, and ϵ is an error term.

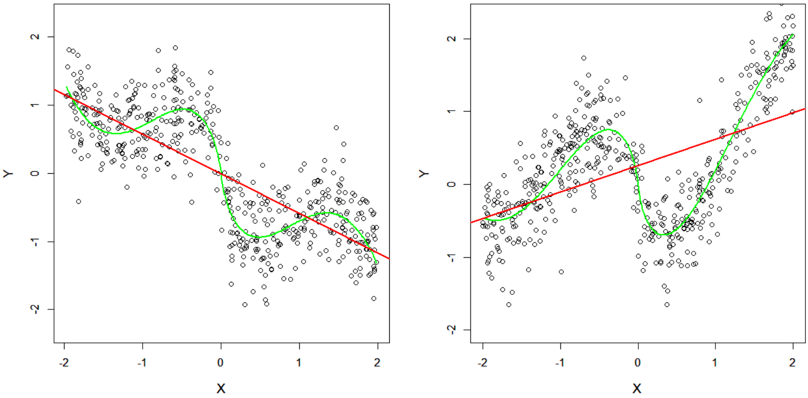
- Suppose that the true regression model is nonlinear, but a linear regression function is (wrongly) employed:

$$\begin{aligned} Y &= \beta_0 + X_1\beta_1 + \cdots + X_k\beta_k + \eta \\ \eta &= \underbrace{m(\mathbf{X}) - (\beta_0 + X_1\beta_1 + \cdots + X_k\beta_k)}_{\text{model mis-specification error}} + \epsilon \end{aligned}$$

- Then, is this model still meaningful?

Model Misspecification

- True conditional expectation function $m(x)$
- Linear regression function $\beta_0 + x\beta_1$



Model Misspecification

- Even when the linear model is wrong, the linear regression provides us with the best linear "approximation" to the true regression function.
- In reality, the assumption that the data follow a perfectly linear function is rarely (or maybe never) met.
- Note that the linear approximation is not always informative:

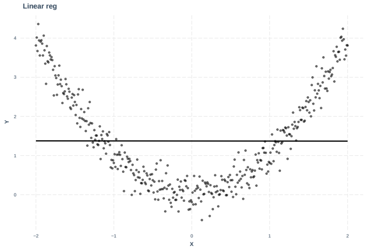
```
>  X <- -200:200/100
>  Y <- X^2 + 0.3*rnorm(100)
>  print(lm(Y ~ X)$coef)
```

```
##  (Intercept)            X
##  1.372116080 -0.001660432
```

Model Misspecification

- Although the regression coefficient of X is almost zero, there is a clear nonlinear relationship between X and Y.

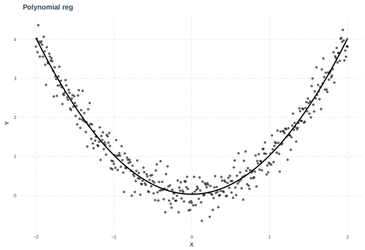
```
> library(jtools)
> reg <- lm(Y ~ X)
> effect_plot(reg, pred = X, main = "Linear reg", plot.points = TRUE)
```



Model Misspecification

- Such nonlinearity can be addressed by adding X^2 as an additional regressor: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 +$ error.

```
> reg <- lm(Y ~ poly(X,2)) # poly(X, k) = X + X^2 + ... + X^k
> effect_plot(reg, pred = X, main = "Polynomial reg", plot.points = TRUE)
```



- It is always a good idea to draw a scatter plot of the data before performing a regression analysis.

Mincer equation

- In some cases, *micro economic theory* can give us a hint for the specification of regression model.
- **Mincer equation**: a classical labor economics model that describes how one's income is determined by his/her education and working experience.

$$\log \text{wage} = \beta_0 + \text{educ}\beta_1 + \text{exp}\beta_2 + \text{exp}^2\beta_3 + \text{error}$$

Mincer equation

```r
>   library(ISLR)
>   library(tidyverse)
>   data(Wage)
```

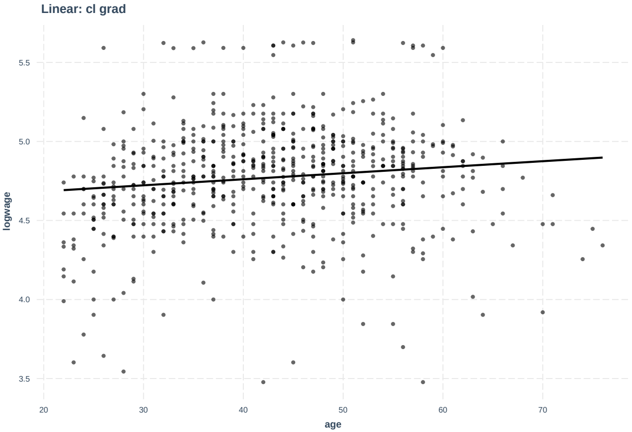
```r
>   head(Wage, 2)
```

```
##        year age            maritl     race       education             region
## 231655 2006  18 1. Never Married 1. White    1. < HS Grad 2. Middle Atlantic
## 86582  2004  24 1. Never Married 1. White 4. College Grad 2. Middle Atlantic
##              jobclass         health health_ins  logwage     wage
## 231655  1. Industrial      1. <=Good      2. No 4.318063 75.04315
## 86582  2. Information 2. >=Very Good      2. No 4.255273 70.47602
```

```r
>   data <- filter(Wage, education == "4. College Grad")
```

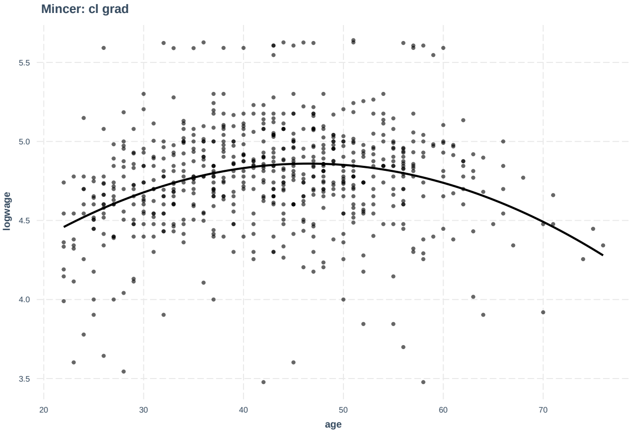
Mincer equation

```
> result <- lm(logwage ~ age, data)
> effect_plot(result, pred = age, main = "Linear: cl grad", plot.points = TRUE)
```



Mincer equation

```
> result <- lm(logwage ~ poly(age, 2), data)
> effect_plot(result, pred = age, main = "Mincer: cl grad", plot.points = TRUE)
```



Summary

- "Regression of Y on $\mathbf{X}$" $=$ Finding a function of $\mathbf{X}$ that predicts the value of Y.
- Regression function: a function of $\mathbf{X}$ that gives the predicted value of Y.
- In terms of MSE, the best regression function is $\mathbb{E}[Y|\mathbf{X}]$.
- When $\mathbb{E}[Y|\mathbf{X}]$ is not actually a linear function of $\mathbf{X}$, the linear regression can give a linear approximation of $\mathbb{E}[Y|\mathbf{X}]$.
- In addition, by adding polynomials of $\mathbf{X}$ as regressors, nonlinearity can be accommodated.
- Economic theory is also useful to find a better regression model.

Appendix: Proof of LIE

Proof of LIE (continuous case)

$$\begin{aligned}
\mathbb{E}[\mathbb{E}(Y|\mathbf{X})] &= \int \left(\int y f_{Y|\mathbf{X}}(y|\mathbf{X}=\mathbf{x}) dy \right) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\
&= \int y \left(\int f_{Y|\mathbf{X}}(y|\mathbf{X}=\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \right) dy \\
&\overset{\text{(i)}}{=} \int y \left(\int f_{Y,\mathbf{X}}(y,\mathbf{x}) d\mathbf{x} \right) dy \overset{\text{(ii)}}{=} \int y f_Y(y) dy = \mathbb{E}[Y],
\end{aligned}$$

where

$$\text{(i)} \quad f_{Y|\mathbf{X}}(y|\mathbf{X}=\mathbf{x}) = \frac{f_{Y,\mathbf{X}}(y,\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}$$

$$\text{(ii)} \quad \int f_{Y,\mathbf{X}}(y,\mathbf{x}) d\mathbf{x} = f_Y(y) \quad : \text{ marginalization}$$

■