Week 3: Logit and probit models

POLI803

Howard Liu

Week 3

University of South Carolina

Outline

- Logit and probit models
- Estimation and interpretation
- Calculating marginal effects

Limited dependent variable models

- So far, we have only considered interval-level dependent variables.

Limited dependent variable models

- So far, we have only considered interval-level dependent variables.
- Yet, there are many interesting political outcomes that are not interval-level, for example,
  - voter turn out, onset of war, etc.

Limited dependent variable models

- So far, we have only considered interval-level dependent variables.

- Yet, there are many interesting political outcomes that are not interval-level, for example,
  - voter turn out, onset of war, etc. [binary]
  - levels of support for legalization of marijuana

Limited dependent variable models

- So far, we have only considered interval-level dependent variables.
- Yet, there are many interesting political outcomes that are not interval-level, for example,
  - voter turn out, onset of war, etc. [binary]
  - levels of support for legalization of marijuana [ordinal]
  - vote choice between Con., Lib. Dem., or Labour

Limited dependent variable models

- So far, we have only considered interval-level dependent variables.
- Yet, there are many interesting political outcomes that are not interval-level, for example,
  - voter turn out, onset of war, etc. [binary]
  - levels of support for legalization of marijuana [ordinal]
  - vote choice between Con., Lib. Dem., or Labour [nominal]
  - number of terrorist attacks

Limited dependent variable models

- So far, we have only considered interval-level dependent variables.
- Yet, there are many interesting political outcomes that are not interval-level, for example,
  - voter turn out, onset of war, etc. [binary]
  - levels of support for legalization of marijuana [ordinal]
  - vote choice between Con., Lib. Dem., or Labour [nominal]
  - number of terrorist attacks [count]
  - $Y >$ threshold (0)

Limited dependent variable models

- So far, we have only considered interval-level dependent variables.
- Yet, there are many interesting political outcomes that are not interval-level, for example,
  - voter turn out, onset of war, etc. [binary]
  - levels of support for legalization of marijuana [ordinal]
  - vote choice between Con., Lib. Dem., or Labour [nominal]
  - number of terrorist attacks [count]
  - $Y >$ threshold (0) [censored]
  - percentage

Limited dependent variable models

- So far, we have only considered interval-level dependent variables.
- Yet, there are many interesting political outcomes that are not interval-level, for example,
  - voter turn out, onset of war, etc. [binary]
  - levels of support for legalization of marijuana [ordinal]
  - vote choice between Con., Lib. Dem., or Labour [nominal]
  - number of terrorist attacks [count]
  - $Y >$ threshold (0) [censored]
  - percentage [0–100]
- These variables are called **limited dependent variables** (categorical/restricted range).

  We cannot & should not use a linear model to analyze limited DV!

Why can't we use LM?

LM (linear regression model) is not suitable because...

Why can't we use LM?

LM (linear regression model) is not suitable because...

- straight lines from LM would be a poor representation of the X-Y relationship when Y is a limited DV;

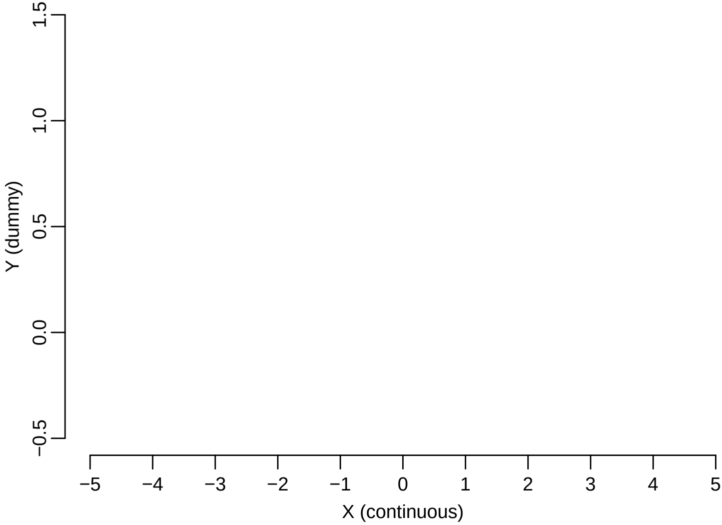
Why can't we use LM?

LM (linear regression model) is not suitable because...

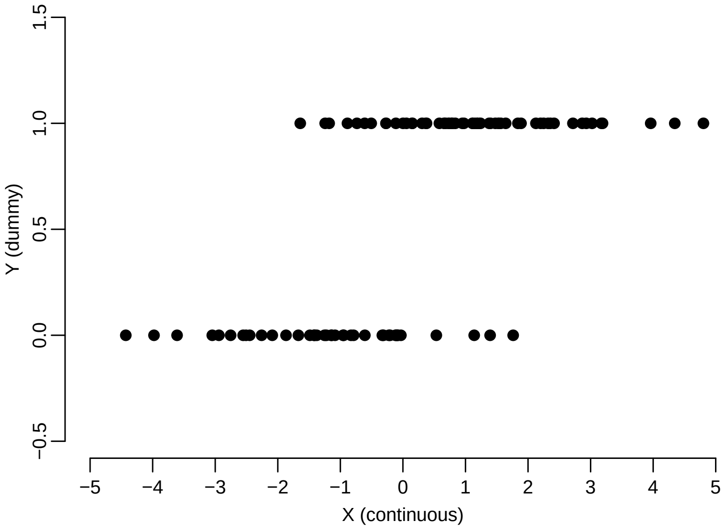
- straight lines from LM would be a poor representation of the X-Y relationship when Y is a limited DV;
  - implausible predicted values
  - marginal effect forced to be constant
- (quadratic / log curves cannot handle this, either)
- We need **Generalized linear models (GLM)**: a flexible generalization that allows for outcome variables to have arbitrary distributions or an arbitrary function of the response variable (the **link function**) to vary linearly with the predictors

We will first consider the case of a binary/dummy DV.

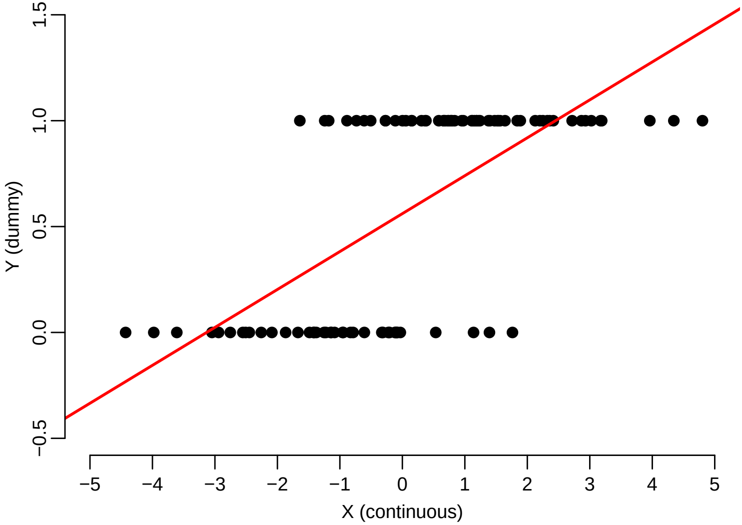
Scatterplot for binary/dummy DV



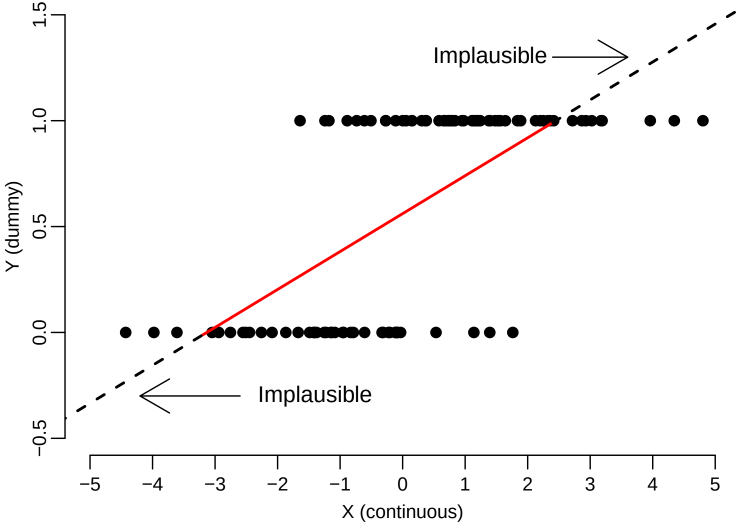
Scatterplot for binary/dummy DV



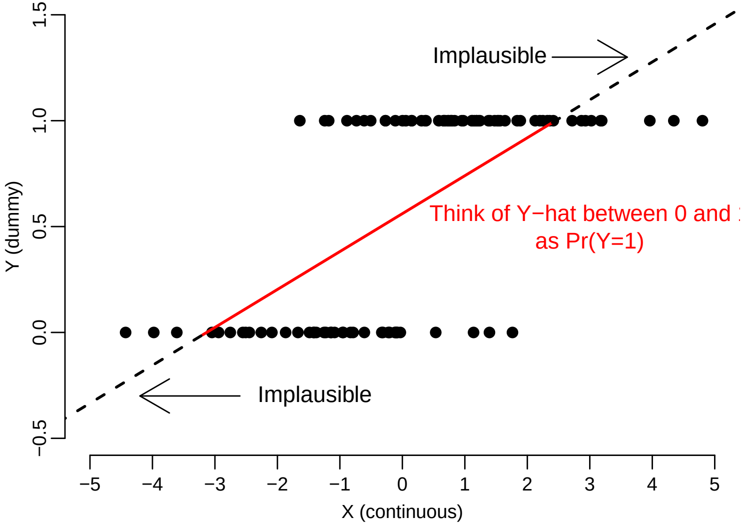
Scatterplot for binary/dummy DV



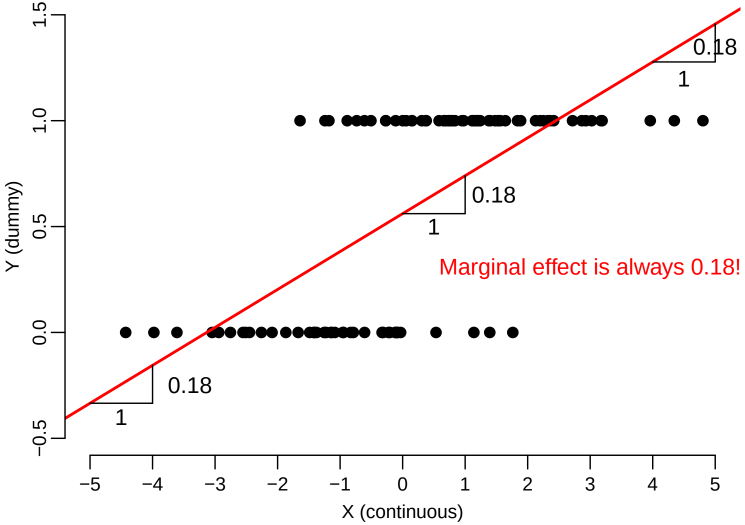
(1) Implausible predictions



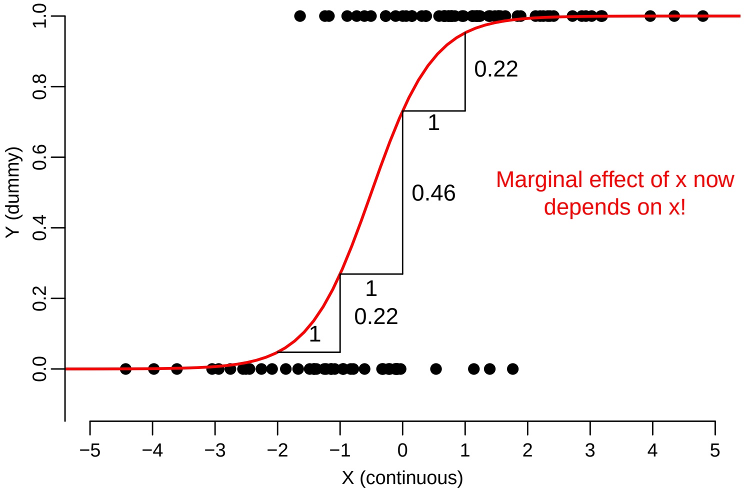
(1) Implausible predictions



(2) Constant marginal effect of X



Fit an S-shaped curve



Fit an S-shaped curve

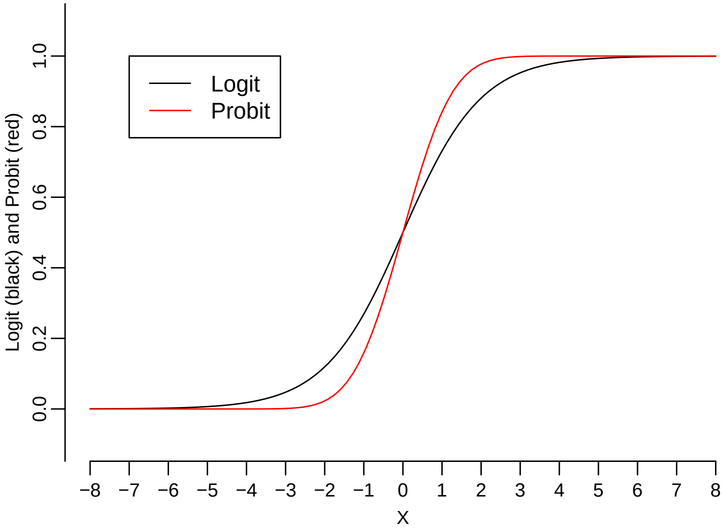
When we fit an S-shaped curve (instead of a line):

- No implausible predicted values;
- Predicted values can be thought of as **latent probabilities** $Pr(Y = 1)$ or $\hat{P}$;
- Marginal effect of X ($\frac{\partial \hat{P}}{\partial X}$) depends on the values of X;
- (Marginal effect of X also depends on the values of the other covariates included in the model!).

Fit an S-shaped curve

- If we run **logit regression** (logistic regression, logit model), we fit a logistic curve.
- If we run **probit regression** (probit model), we fit a probit curve.
- Logit and probit curves are very similar in shape, so you can just run one, not both.

Logist and Probit Curves



Logit regression

Logit regression can be represented as:

$$
\begin{aligned}
Y^* &= \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k \\
\hat{P} &= \Lambda(Y^*)
\end{aligned}
$$

where $\Lambda(x) = \frac{1}{1+\exp(-\beta x)}$ is called the **link function**

- Y^* = latent utility (propensity).
- Y^* can range between $-\infty$ and ∞, but $\hat{P}$ ranges between 0 and 1.
- β_m shows **the marginal effect of X_m on Y^***, but NOT the effect of X_m on $\hat{P}$ itself.
- Yet, we are interested in **the effect of X_m on $\hat{P}$**.

Probit regression

Probit regression can be represented as:

$$
\begin{aligned}
Y^* &= \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k \\
\hat{P} &= \Phi(Y^*)
\end{aligned}
$$

where $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{x^2}{2}\right) dx$

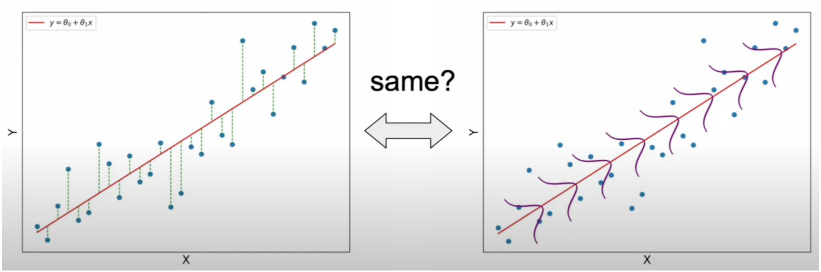
- Logit regression uses the logit link function, $\Lambda()$.
- Probit regression uses the probit link function, $\Phi()$.

The LPM model or Logit model?

- The linear probability model (LPM) fits a linear regression model to a binary response variable, often using OLS.
- OLS supporters:
  - LPM is the wrong but super useful model because changes (marginal effects) can be interpreted in the probability scale
  - OLS not always give nonsensical predictions
  - Causal inference: most causal inference techniques rely on OLS (2SLS and DiD)
- MLE supporters:
  - LPM is the wrong, period
  - If model fit and prediction accuracy are the goals, logit (and other MLE estimators) always win

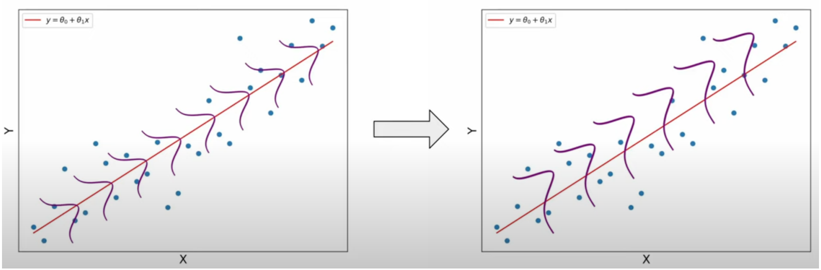
Difference between OLS and MLE

- Our old friend: Ordinary Least Squares (OLS)



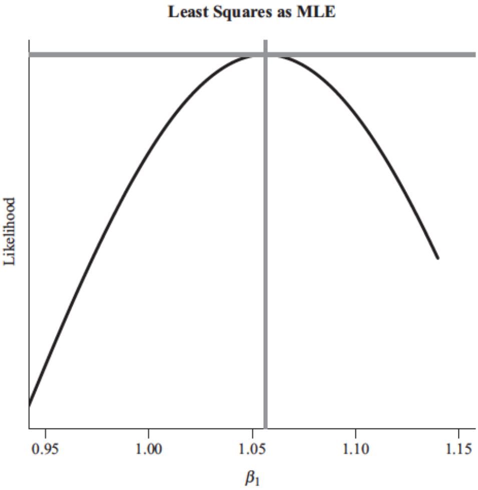
- OLS is a very special case of Maximum Likelihood Estimation that happens when "errors are normally distributed"

What to do if errors are *not* normally distributed?



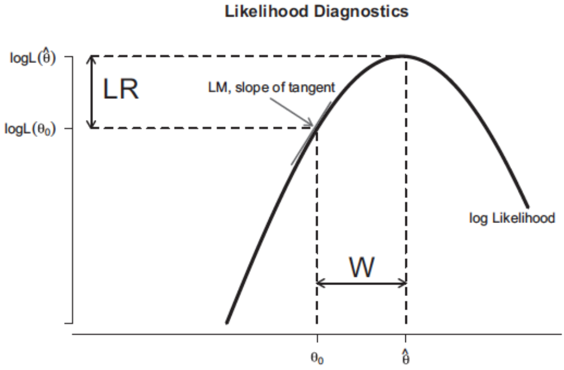
- Least squares method can't work anymore

Our new friend: Maximum Likelihood Estimation (MLE)



Our new friend: Maximum Likelihood Estimation (MLE)

- MLE: maximize the likelihood of observing θ given a probability distribution (e.g., Logit distribution)



MLE Estimator

- The logistic functional form:

$$\theta_i = \texttt{logit}^{-1}(x_i^T\beta) = \frac{1}{1+e^{-x_i^T\beta}} \qquad (1)$$

- Joint probability (the product of all conditional probability) for a Bernoulli random variable

$$\Pr(y \mid \theta) = \prod_{i=1}^{n} \theta_i^{y_i}(1-\theta_i)^{1-y_i} \qquad (2)$$

- Take the log-likelihood

$$\begin{aligned}
\log L(\theta \mid y) &= \sum_{i=1}^{n} [y_i \log \theta_i + (1-y_i)\log(1-\theta_i)] \\
&= \sum_{i=1}^{n} \left[y_i \log \left(\frac{1}{1+e^{-x_i^\top \beta}} \right) + (1-y_i)\log\left(1 - \frac{1}{1+e^{-x_i^\top \beta}}\right)\right] \\
&= \sum_{i=1}^{n} \left[-y_i \log(1+e^{-x_i^\top \beta}) + (1-y_i)\log\left(\frac{e^{-x_i^\top \beta}}{1+e^{-x_i^\top \beta}}\right)\right]
\end{aligned}$$

MLE Estimator

$$
\begin{aligned}
\log L(\theta \mid y) &= \sum_{i=1}^{n} [y_i \log \theta_i + (1 - y_i) \log(1 - \theta_i)] \\
&= \sum_{i=1}^{n} \left[y_i \log \left(\frac{1}{1 + e^{-x_i^\top \beta}} \right) + (1 - y_i) \log \left(1 - \frac{1}{1 + e^{-x_i^\top \beta}} \right) \right] \\
&= \sum_{i=1}^{n} \left[-y_i \log(1 + e^{-x_i^\top \beta}) + (1 - y_i) \log \left(\frac{e^{-x_i^\top \beta}}{1 + e^{-x_i^\top \beta}} \right) \right] \\
&= \sum_{i=1}^{n} \log \left(\frac{e^{-x_i^\top \beta(1-y_i)}}{1 + e^{-x_i^\top \beta}} \right)
\end{aligned}
$$

- How to estimate this θ?

MLE Estimator

$$
\begin{aligned}
\log L(\theta \mid y) &= \sum_{i=1}^{n} [y_i \log \theta_i + (1 - y_i) \log(1 - \theta_i)] \\
&= \sum_{i=1}^{n} \left[y_i \log \left(\frac{1}{1 + e^{-x_i^\top \beta}} \right) + (1 - y_i) \log \left(1 - \frac{1}{1 + e^{-x_i^\top \beta}} \right) \right] \\
&= \sum_{i=1}^{n} \left[-y_i \log(1 + e^{-x_i^\top \beta}) + (1 - y_i) \log \left(\frac{e^{-x_i^\top \beta}}{1 + e^{-x_i^\top \beta}} \right) \right] \\
&= \sum_{i=1}^{n} \log \left(\frac{e^{-x_i^\top \beta(1-y_i)}}{1 + e^{-x_i^\top \beta}} \right)
\end{aligned}
$$

- How to estimate this θ?
- Newton Raphson approximation $\rightarrow$ R will do it for you, fortunately
  - If you're interested the hand derivation in math $\rightarrow$ read

MLE Estimator

$$
\begin{aligned}
\log L(\theta \mid y) &= \sum_{i=1}^{n} [y_i \log \theta_i + (1 - y_i) \log(1 - \theta_i)] \\
&= \sum_{i=1}^{n} \left[y_i \log \left(\frac{1}{1 + e^{-x_i^\top \beta}} \right) + (1 - y_i) \log \left(1 - \frac{1}{1 + e^{-x_i^\top \beta}} \right) \right] \\
&= \sum_{i=1}^{n} \left[-y_i \log(1 + e^{-x_i^\top \beta}) + (1 - y_i) \log \left(\frac{e^{-x_i^\top \beta}}{1 + e^{-x_i^\top \beta}} \right) \right] \\
&= \sum_{i=1}^{n} \log \left(\frac{e^{-x_i^\top \beta (1 - y_i)}}{1 + e^{-x_i^\top \beta}} \right)
\end{aligned}
$$

- How to estimate this θ?
- Newton Raphson approximation $\rightarrow$ R will do it for you, fortunately
  - If you're interested the hand derivation in math $\rightarrow$ read
- Note: MLE estimation is not restricted by gaussian/normal anymore, given that the functional form of logit distribution has been identified by statisticians.

Do you remember? Bayesian estimation of θ

$$\underbrace{\xi(\theta|x)}_{\texttt{posterior dist.}} \propto \underbrace{f(x|\theta)}_{\texttt{data/Likelihood}} \underbrace{\xi(\theta)}_{\texttt{prior dist.}}$$

- Bayes just goes a little further by multiplying the likelihood function with a prior guess

Estimation in R

Running a logit / probit model is quite easy in R.

```
fit <- glm (y ~ x1 + x2 + x3...,
      data = dataset.name,
      family = binomial(link = logit))
```

```
fit <- glm (y ~ x1 + x2 + x3...,
      data = dataset.name,
      family = binomial(link = probit))
```

What's not quite easy is to interpret the results.

Interpretation

In logit/probit models (or in any limited DV models) we cannot interpret the estimated coefficients β as the marginal effect.

Interpretation

In logit/probit models (or in any limited DV models) we cannot interpret the estimated coefficients β as the marginal effect.

- With LM (without interaction terms), we could: $\frac{\partial \hat{Y}}{\partial X_1} = \beta_1$.

Interpretation

In logit/probit models (or in any limited DV models) we cannot interpret the estimated coefficients β as the marginal effect.

- With LM (without interaction terms), we could: $\frac{\partial \hat{Y}}{\partial X_1} = \beta_1$.
- With logit model, β_1 merely shows the marginal effect of X_1 on Y^*, which is not the quantity of interest.

Interpretation

In logit/probit models (or in any limited DV models) we cannot interpret the estimated coefficients β as the marginal effect.

- With LM (without interaction terms), we could: $\frac{\partial \hat{Y}}{\partial X_1} = \beta_1$.
- With logit model, β_1 merely shows the marginal effect of X_1 on Y^*, which is not the quantity of interest.
- With logit model, what we care is: $\frac{\partial \hat{P}}{\partial X_1}$, or the effect of X_1 on the probability $Y = 1$. (We care the $\hat{P}$)

Interpretation

In logit/probit models (or in any limited DV models) we cannot interpret the estimated coefficients β as the marginal effect.

- With LM (without interaction terms), we could: $\frac{\partial \hat{Y}}{\partial X_1} = \beta_1$.
- With logit model, β_1 merely shows the marginal effect of X_1 on Y^*, which is not the quantity of interest.
- With logit model, what we care is: $\frac{\partial \hat{P}}{\partial X_1}$, or the effect of X_1 on the probability $Y = 1$. (We care the $\hat{P}$)
- Moreover, the marginal effect of X_1 on $\hat{P}$ differs depending on the value of X_1 itself as well as other Xs included in the model.

What to do after estimation

Three Steps

1. Produce a regression table using `stargazer`.
   - Identify the "best" model(s)

2. Discuss statistical significance and the **sign** (but not the size) of coefficients.

3. Graphically illustrate the **size** of the marginal effects (and discuss them in the text).
   - Do this for "interesting" and/or "representative" cases in your covariates

How to illustrate the marginal effect of X

$$\hat{P} = \Lambda(\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k)$$

1. Choose one X to focus on. Let's say we are interested in X_1.
2. Set the values of all the other Xs at their "interesting" and/or "representative" values (mean, median, minimum, maximum, etc.).
3. Effect plot: Graphically and numerically show the relationship between X_1 and $\hat{P}$ using the `effect` function.

Example: Titanic passenger survival

```
> head(td)
  survived                                        name pclass     age child
1        1                Allen, Miss. Elisabeth Walton      1 29.0000 Adult
2        1               Allison, Master. Hudson Trevor      1  0.9167 Child
3        0                 Allison, Miss. Helen Loraine      1  2.0000 Child
4        0         Allison, Mr. Hudson Joshua Creighton      1 30.0000 Adult
5        0 Allison, Mrs. Hudson J C (Bessie Waldo Daniels)      1 25.0000 Adult
6        1                          Anderson, Mr. Harry      1 48.0000 Adult
  old female sibsp parch alone     fare cherbourg queenstown southampton
1   0 Female     0     0     1 211.3375         0          0           1
2   0   Male     1     2     0 151.5500         0          0           1
3   0 Female     1     2     0 151.5500         0          0           1
4   0   Male     1     2     0 151.5500         0          0           1
5   0 Female     1     2     0 151.5500         0          0           1
6   0   Male     0     0     1  26.5500         0          0           1
```

- `survived`: 1 (survived) or 0 (not survived)
- `pclass`: passenger class (first, second, third)
- `child`: Adult or Child (under 16 yo)
- `old`: 1 (50+ yo) or 0

Example: Titanic passenger survival

```
> head(td)
  survived                                     name pclass     age child
1        1             Allen, Miss. Elisabeth Walton      1 29.0000 Adult
2        1            Allison, Master. Hudson Trevor      1  0.9167 Child
3        0              Allison, Miss. Helen Loraine      1  2.0000 Child
4        0      Allison, Mr. Hudson Joshua Creighton      1 30.0000 Adult
5        0 Allison, Mrs. Hudson J C (Bessie Waldo Daniels)      1 25.0000 Adult
6        1                       Anderson, Mr. Harry      1 48.0000 Adult
  old female sibsp parch alone     fare cherbourg queenstown southampton
1   0 Female     0     0     1 211.3375         0          0           1
2   0   Male     1     2     0 151.5500         0          0           1
3   0 Female     1     2     0 151.5500         0          0           1
4   0   Male     1     2     0 151.5500         0          0           1
5   0 Female     1     2     0 151.5500         0          0           1
6   0   Male     0     0     1  26.5500         0          0           1
```

- `sibsp`: number of siblings aboard
- `parch`: number of parents / children aboard
- `fare`: Passenger fare (in Pre-1970 British Pounds)
- `cherbourg`, `queenstown`, `southampton`: Embarked at ...

Example: Titanic passenger survival

Let's say we are interested in the following two:

- the effect of fare on survival (i.e., does paying more increase the chance of survival?)
- the effect of `child` and `female` dummies on survival (i.e., was "women and children first" policy implemented?)

	Dependent variable:			
	survived			
	(1)	(2)	(3)	(4)
fare	0.012***	0.012***	0.009***	0.009***
	(0.002)	(0.002)	(0.002)	(0.002)
childChild		0.808***		0.675***
		(0.204)		(0.235)
femaleFemale			2.340***	2.362***
			(0.138)	(0.156)
Constant	-0.882***	-0.893***	-1.718***	-1.722***
	(0.076)	(0.089)	(0.102)	(0.119)
Observations	1,308	1,045	1,308	1,045
Log Likelihood	-827.016	-662.031	-663.249	-528.894
Akaike Inf. Crit.	1,658.032	1,330.061	1,332.498	1,065.788
Note:		*p<0.1; **p<0.05; ***p<0.01		

Model fit

Log likelihood

- **Always negative** (log of "likelihood" = a number between 0 and 1)
- sort of like R^2 (but not really; it doesn't have intuitive interpretation)
- the larger (**smaller in absolute values**), the better

Akaike's Information Criterion (AIC)

- AIC $= -2(L - k)$, where L is the log likelihood and k is the number of coefficients
- sort of like adjusted R^2 (penalizes models with lots of Xs)
- **the smaller, the better**
- Not comparable if n is different

	Dependent variable:			
	survived			
	(1)	(2)	(3)	(4)
fare	0.012***	0.012***	0.009***	0.009***
	(0.002)	(0.002)	(0.002)	(0.002)
childChild		0.808***		0.675***
		(0.204)		(0.235)
femaleFemale			2.380***	2.362***
			(0.155)	(0.156)
Constant	-0.794***	-0.893***	-1.647***	-1.722***
	(0.084)	(0.089)	(0.114)	(0.119)
Observations	1,045	1,045	1,045	1,045
Log Likelihood	-669.970	-662.031	-533.046	-528.894
Akaike Inf. Crit.	1,343.941	1,330.061	1,072.091	1,065.788
Note:		*p<0.1;	**p<0.05;	***p<0.01

Interpretation

1. Regression table
   - Model (4) is fits the data better based on AICs

2. Fare, child dummy, and female dummy are all positive and significant, as expected

In order to see if the effect of independent variables are also **substantively** significant, we need to obtain marginal effect.

Interpretation: marginal effect

$$\hat{P} = \Lambda(-1.722 + 0.009 * fare + 0.675 * child + 2.362 * female)$$

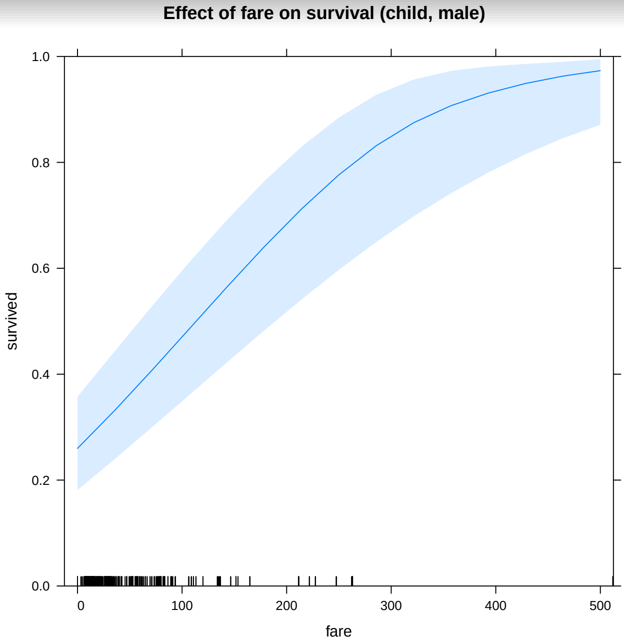
- Let's first calculate and plot the effect of fare on survival.
- To see the relationship between fare and $\hat{P}$, we calculate $\hat{P}$ for several different values of fare, holding constant other variables at some values.
  - The effect function: `effect(term = "fare", mod = fit.4)` sets everything else constant **at its mean value**.
  - But, mean does not make sense for child / female.
- We should do this for the following four cases:
  - Child, Male
  - Child, Female
  - Adult, Male
  - Adult, Female

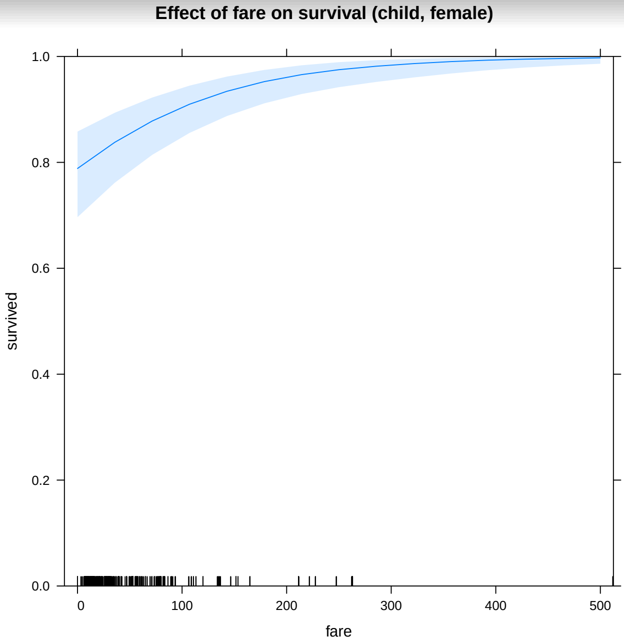
Interpretation: marginal effect

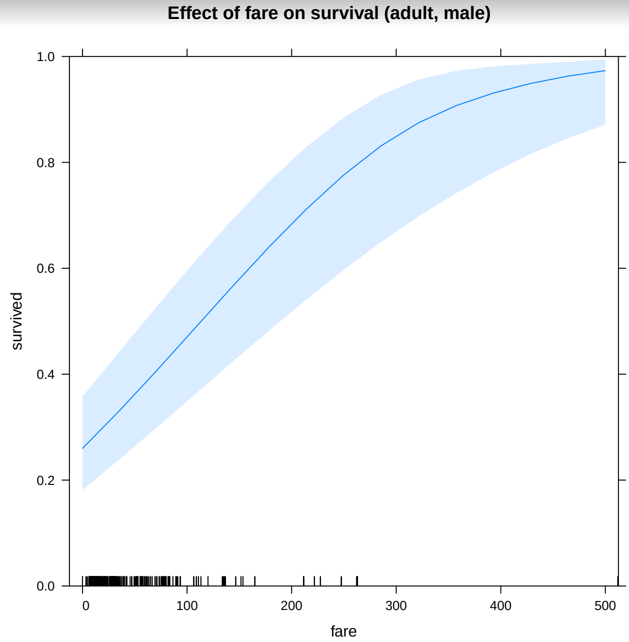
```
> # Child, Male
> effect(term = "fare", mod = fit.4,
+   given.values = c(childChild = 1, femaleFemale = 0) )

 fare effect
fare
        0       100       200       300       400       500
0.2598797 0.4703560 0.6919313 0.8503110 0.9349247 0.9732160
```

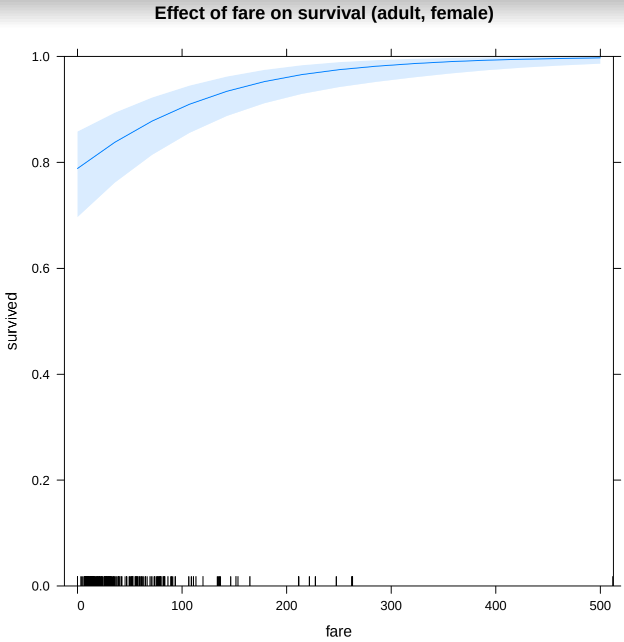
Interpretation: marginal effect

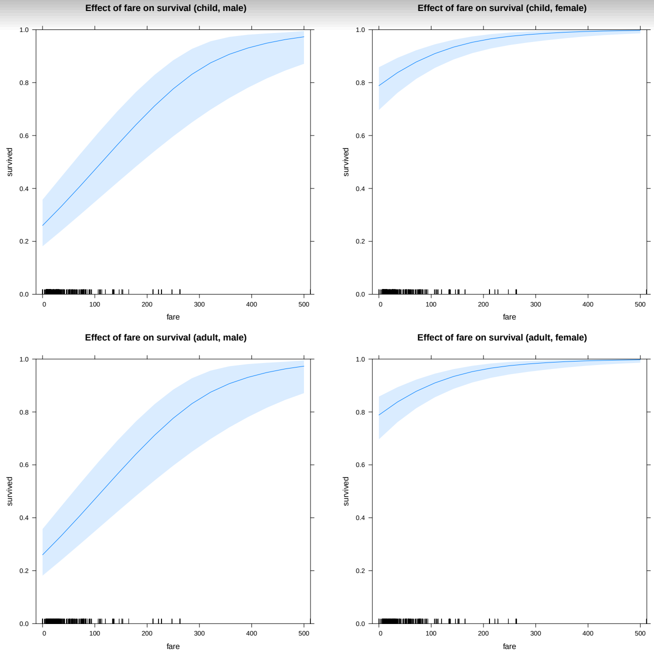
```
> # Child, Male
> effect(term = "fare", mod = fit.4,
+   given.values = c(childChild = 1, femaleFemale = 0) )

 fare effect
fare
        0       100       200       300       400       500 
0.2598797 0.4703560 0.6919313 0.8503110 0.9349247 0.9732160 

> # Child, Female
> effect(term = "fare", mod = fit.4,
+   given.values = c(childChild = 1, femaleFemale = 1) )

 fare effect
fare
        0       100       200       300       400       500 
0.7884491 0.9040867 0.9597422 0.9836853 0.9934850 0.9974139 
```


Effect of fare on survival (child, male)
survived
fare
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500


Effect of fare on survival (child, female)
survived
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
fare


Effect of fare on survival (adult, male)
survived
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
fare


Effect of fare on survival (adult, female)
survived
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
fare


Effect of fare on survival (child, male)
Effect of fare on survival (child, female)
Effect of fare on survival (adult, male)
Effect of fare on survival (adult, female)
survived
fare
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500

Interpretation: marginal effect

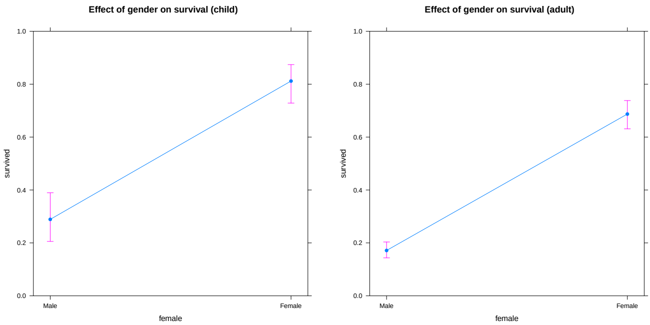
```
> effect(term = "female", mod = fit.4)

 female effect
female
     Male    Female
0.2129694 0.7417487
```

Interpretation: marginal effect

```
> effect(term = "female", mod = fit.4,
+   given.values = c(fare = 15.75, childChild = 1) )

 female effect
female
     Male    Female
0.2889574 0.8117991
```

```
> effect(term = "female", mod = fit.4,
+   given.values = c(fare = 15.75, childChild = 0) )

 female effect
female
     Male    Female
0.1714088 0.6870838
```

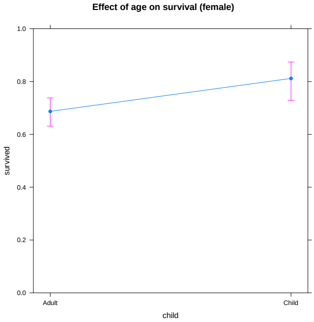
Interpretation: marginal effect (gender)

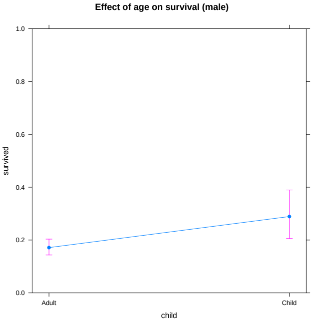


Interpretation: marginal effect

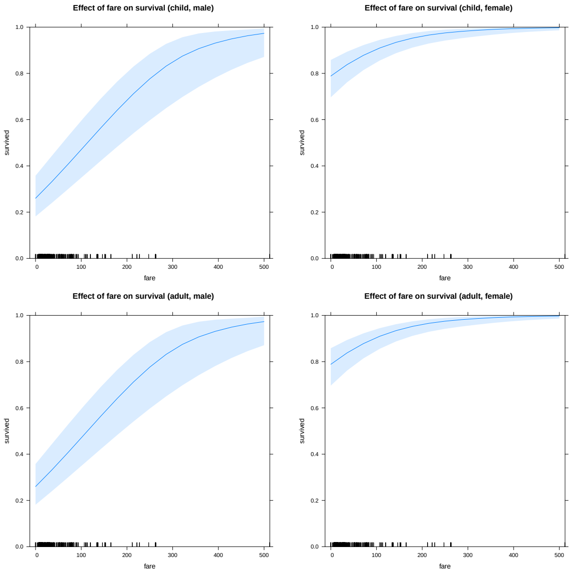
```
> effect(term = "child", mod = fit.4,
+   given.values = c(fare = 15.75, femaleFemale = 1) )

 child effect
child
    Adult     Child
0.6870838 0.8117991

> effect(term = "child", mod = fit.4,
+   given.values = c(fare = 15.75, femaleFemale = 0) )

 child effect
child
    Adult     Child
0.1714088 0.2889574
```

Interpretation: marginal effect (age)





Interpretation: marginal effect (fare)



Summary

Some relationships can only be found by calculating and plotting the marginal effects:

- Passenger's gender has a significant impact on survival probability.
  - For "average" adult passengers (because we held fare at the mean), probability of survival increases from 17% to 69% (= 17% of average adult male passengers survived, whereas 69% of average adult female passengers survived). → *p*.31

Summary

Some relationships can only be found by calculating and plotting the marginal effects:

- Passenger's gender has a significant impact on survival probability.
    - For "average" adult passengers (because we held fare at the mean), probability of survival increases from 17% to 69% (= 17% of average adult male passengers survived, whereas 69% of average adult female passengers survived). → *p*.31
- Passenger's age does have an impact on survival probability, but the effect is much smaller compared with the effect of gender.
    - For "average" male passengers, probability of survival increases from 17% to 29% if he is a child (= 17% of average adult male passengers survived, whereas 29% of average child male passengers survived). → *p*.33
    - For "average" female passenger, probability of survival increases from 69% to 81% if she is a child (= 69% of average adult female passengers survived, whereas 81% of average child female passengers survived).

Summary

Even though we did not include an interaction term, the effect of one variable depends on the values of all the other independent variables.

- Passenger's fare influences the probability of survival, but its effect is much bigger for male passengers.
    - For male child, the probability of survival increases from 26% to 70% when we increase the fare from 0 to 200 GBP (= 44 percentage points increase). → *p*.34
    - For female child, the probability of survival increases from 79% to 96% when we increase the fare from 0 to 200 GBP (= 23 percentage points increase).