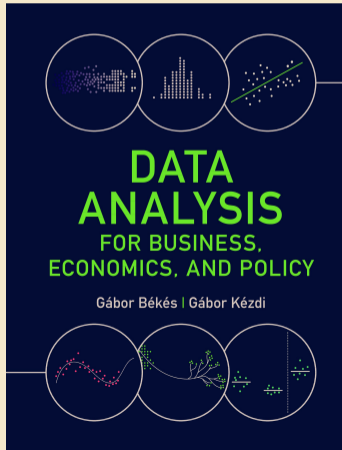


Békés-Kézdi: Data Analysis, Chapter 11: Modelling probabilities



Data Analysis for Business, Economics, and Policy

Gábor Békés (Central European University)
Gábor Kézdi (University of Michigan)

Cambridge University Press, 2021

gabors-data-analysis.com

Central European University

Version: v3.1 License: CC BY-NC 4.0

Any comments or suggestions:
gabors.da.contact@gmail.com

Motivation

- ▶ *What are the health benefits of not smoking? Considering the 50+ population, we can investigate if differences in smoking habits are correlated with differences in health status.*

Binary events

- ▶ Start with binary events: things that either happen or don't happen captured by **binary variable**
- ▶ How can we model these events?
 - ▶ We do not observe 'on average' larger values for y in this case.
- ▶ Solution - model instead the probabilities!

$$E[y] = P[y = 1]$$

- ▶ The average of a 0–1 binary variable is also the probability that it is one.
 - ▶ Frequency (25% of cases) — probability (25% chance)
- ▶ Expected value = average probability of event happening
 - ▶ Use the same tools, but interpretation is changing!

Modelling events: LMP

Linear probability model - LPM

- ▶ Modelling probability – regression with *binary dependent variable*.
- ▶ *Linear Probability Model (LPM)* is a linear regression with a binary dependent variable
- ▶ Differences in average y are also differences in the probability that $y = 1$
- ▶ Linear regressions with binary dependent variables show
 - ▶ differences in expected y by x , is also differences in the probability of $y = 1$ by x .
- ▶ Introduce notation for probability:

$$y^P = P[y = 1|x_1, x_2, \dots]$$

- ▶ Linear probability model (LPM) regression is

$$y^P = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Linear probability model - interpretation

$$y^P = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- ▶ y^P denotes the probability that the dependent variable is one, conditional on the right-hand-side variables of the model.
- ▶ β_0 shows the probability of y if all x are zero.
- ▶ β_1 shows the difference in the probability that $y = 1$ for observations that are different in x_1 but are the same in terms of x_2 .
- ▶ Still true: average difference in y corresponding to differences in x_1 with x_2 being the same.

Linear probability model - modelling

- ▶ Linear probability model (LPM) using OLS.
- ▶ We can use all transformations in x , that we used before:
 - ▶ Log, Polinomials, Splines, dummies, interactions, ect.
- ▶ All formulae and interpretations for standard errors, confidence intervals, hypotheses and p-values of tests are the same.
- ▶ Heteroskedasticity robust error are essential in this case!

Predicted values in LPM

- ▶ Predicted values - \hat{y}^P - may be problematic, calculated the same way, but to be interpreted as probabilities.

$$\hat{y}^P = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

- ▶ Predicted values need to be between 0 and 1 because they are probabilities
- ▶ But in LPM, they may be below 0 and above 1. No formal bounds in the model.
 - ▶ With continuous variables that can take any value (GDP, Population, sales, etc), this could be a serious issue
 - ▶ With binary variables, no problem ('saturated models')
- ▶ Problem if goal is prediction!
- ▶ Not a big issue for inference → uncover patterns of association.
 - ▶ But note in theory it may give biased estimates...

Does smoking pose a health risk?

The question of the case study is whether, and by how much less likely smokers are to stay healthy than non-smokers.

- ▶ focus on people of age 50 to 60 who consider themselves healthy
- ▶ ask them four years later as well

Research question: Does smoking lead to deteriorating health?

Data

- ▶ $y = 1$ if person stayed healthy
- ▶ $y = 0$ if person became unhealthy
- ▶ Data comes from [SHARE](#) (Survey for Health, Aging and Retirement in Europe)
 - ▶ 14 European countries
 - ▶ Demographic information on all individual
 - ▶ 2011 and 2015 participants are used
 - ▶ Being healthy means to report “feeling excellent” or “very good”
 - ▶ $N = 3,109$

LPM

Start with a simple univariate model with being a smoker.

$$stays\ healthy^P = \alpha + \beta smoker$$

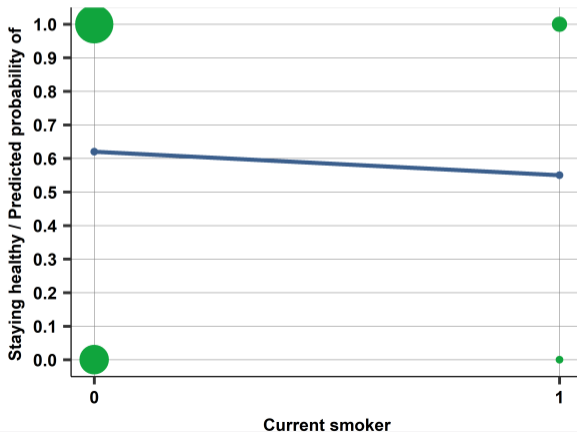
Both dependent and independent models are using only dummy variables.

Estimated β is -0.072

Can we draw a scatterplot?

Scatterplot

Figure: Staying healthy - scatterplot and regression line



LPM Interpretation

- ▶ The coefficient on smokes shows the difference in the probability of staying healthy comparing current smokers and current nonsmokers.
- ▶ Current smokers are *7 percentage points* less likely to stay healthy than those that did not smoke.
- ▶ Can add additional controls to capture if quitting matters.

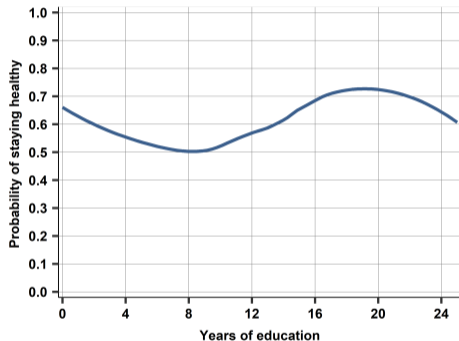
LPM with many regressors I.

- ▶ Multiple regression – closer to causality
 - ▶ compare people who are very similar in many respects but are different in smoking habits
 - ▶ find many confounders that could be correlated with smoking habits and health outcomes
- ▶ Smokers / non-smokers – different in many other behaviors and conditions:
 - ▶ personal traits
 - ▶ behavior such as eating, exercise
 - ▶ socio-economic conditions
 - ▶ background - e.g. country they live in

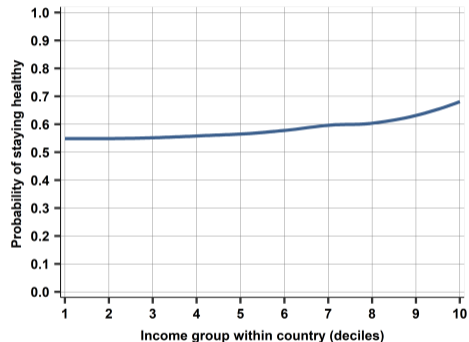
LPM with many regressors II.

- ▶ Pick variables:
 - ▶ gender dummy, age, years of education,
 - ▶ income (measured as in which of the 10 income groups individuals belong within their country),
 - ▶ body mass index (a measure of weight relative to height),
 - ▶ whether the person exercises regularly, the country in which they live.
 - ▶ country - set of binary indicators.
- ▶ Think functional form:
 - ▶ Continuous control variables might have nonlinear relationship with staying healthy
 - ▶ Explore the relationship with nonparametric tools

Functional form selection



Staying healthy and years of education



Staying healthy and income group

Decisions: (1) Include education as a piecewise linear spline with knots at 8 and 18 years; (2) include income in a linear way.

LPM results

Probability of staying healthy - extended model

VARIABLES	Staying healthy	VARIABLES (cnt.)	
Current smoker (Y/N)	-0.061* (0.024)	Income group	0.008* (0.003)
Ever smoked (Y/N)	0.015 (0.020)	BMI (for < 35)	-0.012** (0.003)
Female (Y/N)	0.033 (0.018)	BMI (for >= 35)	0.006 (0.017)
Age	-0.003 (0.003)	Exercises regularly (Y/N)	0.053** (0.017)
Years of education (for < 8)	-0.001 (0.007)	Years of education (for >= 18)	-0.010 (0.012)
Years of education (for >= 8 and < 18)	0.017** (0.003)	Country indicators	YES
Observations	3,109		

Robust standard errors in parentheses. ** p<0.01, * p<0.05

Y/N denotes binary vars. BMI and education entered as spline. Age in years. Income in deciles.

Detour: Regression Tables

- ▶ If need to show many explanatory variables
- ▶ Do not show table 12*2 rows, people will not see it.
- ▶ Either only show selected variables
- ▶ Or may need to create two columns.
- ▶ Make sure you have title, N of observations, footnote on SE, stars.
 - ▶ SE, stars: many different notations. Check carefully.
 - ▶ Default is ***= $p < 0.01$. But often **= $p < 0.01$ (like here)

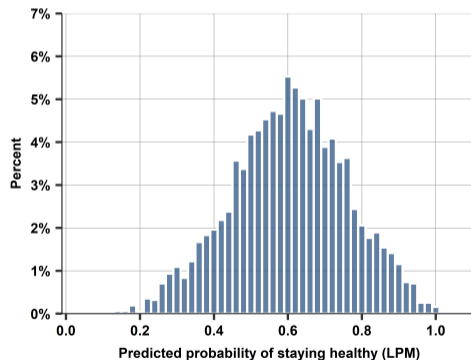
Does smoking pose a health risk?– LPM interpret

- ▶ coefficient on currently smoking is -0.06
 - ▶ The 95% confidence interval is relatively wide $[-0.11, -0.01]$, but it does not contain zero
- ▶ no significant differences in staying healthy when comparing never smokers to those who used to smoke but quit
- ▶ women are 3 percentage points more likely to stay in good health
- ▶ age does not seem to matter in this relatively narrow age range of 50 to 60 years
- ▶ differences in years of education
 - ▶ do not matter if we compare people with less than 8 years or more than 18 years,
 - ▶ matters a lot in-between, with a one-year-difference corresponding to 1.7 percentage point difference in the likelihood of staying healthy
- ▶ income matters somewhat less, maybe non-linear?
- ▶ Regular exercise matters.

LPM's predicted probabilities

- ▶ Predicted probabilities are calculated from the extended linear probability model.
- ▶ Predicted probability of staying healthy from this linear probability model ranges between 0.036 and 1.011
 - ▶ LPM means it can be below 0 or above 1...
 - ▶ Here, only marginally above 1

Histogram of the predicted probabilities



Source: share-health dataset.

Compare predicted probability distribution

► Drill down in distribution:

- Looking at the composition of people: top vs bottom part of probability distribution
- Look at average values of covariates for top and bottom 1% of predicted probabilities!

Top 1% predicted probability:

- no current smokers, women,
- avg 17.3ys of education, higher income
- BMI of 20.7, and 90% of them exercise.

Bottom 1% predicted probability:

- 37.5% current smokers, 63% men
- 7.6 years of education, lower income
- BMI of 30.5, 19% exercise

Modelling events: logit

Probability models: logit and probit

- ▶ Prediction: predicted probability need to be between 0 and 1
- ▶ For prediction, we use non-linear models
- ▶ Relate the probability of the $y = 1$ event to a nonlinear function of the linear combination of the explanatory variables -> 'Link function'
 - ▶ Link function is some $F(\cdot)$, s.t. $F(y)$ may be used in linear models.
- ▶ Two options: Logit and probit – different link function
 - ▶ Resulting probability is always strictly between zero and one.

Link functions I.

The **logit** model has the following form:

$$y^P = \Lambda(\beta_0 + \beta_1 x_1, \beta_2 x_2 + \dots) = \frac{\exp(\beta_0 + \beta_1 x_1, \beta_2 x_2 + \dots)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots)}$$

where the link function $\Lambda(z) = \frac{\exp(z)}{1 + \exp(z)}$ is called the *logistic function*.

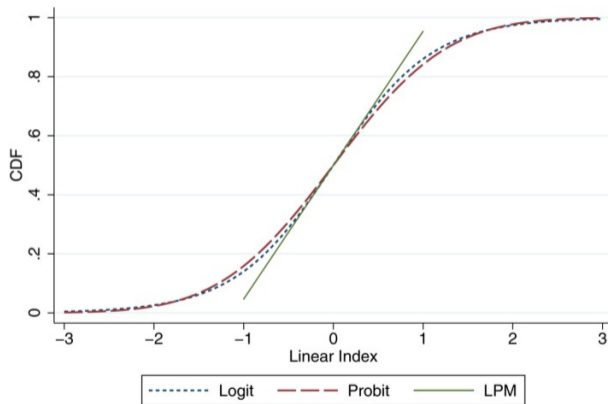
The **probit** model has the following form:

$$y^P = \Phi(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots)$$

where the link function $\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz$, is the cumulative distribution function (CDF) of the standard normal distribution.

Link functions II.

- ▶ Both Λ and Φ are increasing S-shape curves, bounded between 0 and 1. (Y here is $\Lambda(z)$ and $\Phi(z)$)
- ▶ Plotted against their respective "z" values. (Here -3 to 3)
- ▶ Small difference (indistinguishable) - logit less steep close to zero and one = thicker tails than the probit.
- ▶ In our models, 'z' is a linear combination of β coefficients and x-s. The parameter estimates are typically different in probit vs logit.



Logit and probit interpretation

- ▶ Both the probit and the logit transform the $\beta_0 + \beta_1 x_1 + \dots$ linear combination using a link function that shows an S-shaped curve.
- ▶ The slope of this curve keeps changing as we change whatever is inside.
 - ▶ The slope is steepest when $y^P = 0.5$;
 - ▶ it is flatter further away; and it becomes very flat if y^P is close to zero or one.
- ▶ The difference in y^P that corresponds to a unit difference in any explanatory variable is not the same.
 - ▶ You need to take the partial derivatives. It depends on the value of x
- ▶ Important consequence: no direct interpretation of the raw coefficient values!

Marginal differences

- ▶ Link functions makes variation in association between x and y^P – for logit and probit models, we do not interpret raw coefficients!
- ▶ Instead, transform them into ‘marginal differences’ for interpretation purposes
- ▶ The **average marginal difference** for x is the average difference in the probability of $y = 1$, that corresponds to a one unit difference in x .
 - ▶ Software may call them ‘marginal effects’ or ‘average marginal effects (AME)’ or ‘average partial effects’.
- ▶ Average marginal difference has the exact **same interpretation as the coefficient of linear probability models**.

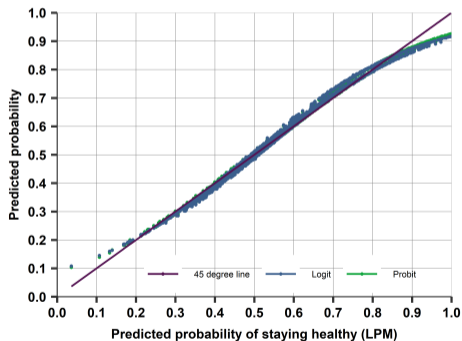
Maximum likelihood estimation

- ▶ When estimating a logit or probit model, we use 'maximum likelihood' estimation.
 - ▶ See 11.U2 for details.
- ▶ Idea for maximum likelihood is another way to get coefficient estimates. Done in steps.
 - ▶ You specify a (conditional) distribution, that you will use during the estimation.
 - ▶ This is logistic for logit and normal for probit model.
 - ▶ You maximize this function w.r.t. your β parameters \rightarrow gives the maximum likelihood for this model.
- ▶ No closed form solution \rightarrow need to use search algorithms.
 - ▶ Search algorithms will play critical role in machine learning as well.
 - ▶ More in DA3.

Predictions for LMP, Logit and Probit I.

- ▶ Compare the three model results
- ▶ Baseline is LPM - extended model.
- ▶ 45 degree line is LPM
- ▶ Predicted probabilities from the logit and the probit shown vs LPM

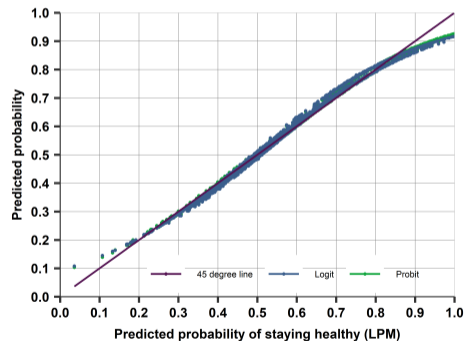
Comparing probabilities from models



Predictions for LMP, Logit and Probit II.

- ▶ Predicted probabilities from the logit and the probit are practically the same
 - ▶ range is between 0.10 and 0.92, which is narrower than the LPM, which ranges from 0.036 to 0.101
- ▶ LPM, logit and probit models produce almost exactly the same predicted probabilities
- ▶ except for the lowest and highest probabilities

Comparing probabilities from models



Coefficient results for logit and probit

Dep.var.: stays healthy	(1) LPM	(2) logit coeffs	(3) logit marginals	(4) probit coeffs	(5) probit marginals
Current smoker	-0.061* (0.024)	-0.284** (0.109)	-0.061** (0.023)	-0.171* (0.066)	-0.060* (0.023)
Ever smoked	0.015 (0.020)	0.078 (0.092)	0.017 (0.020)	0.044 (0.056)	0.016 (0.020)
Female	0.033 (0.018)	0.161* (0.082)	0.034* (0.018)	0.097 (0.050)	0.034 (0.018)
Years of education (if < 8)	-0.001 (0.007)	-0.003 (0.033)	-0.001 (0.007)	-0.002 (0.020)	-0.001 (0.007)
Years of education (if >= 8 and < 18)	0.017** (0.003)	0.079** (0.016)	0.017** (0.003)	0.048** (0.010)	0.017** (0.003)
Years of education (if >= 18)	-0.010 (0.012)	-0.046 (0.055)	-0.010 (0.012)	-0.029 (0.033)	-0.010 (0.012)
Income group	0.008* (0.003)	0.036* (0.015)	0.008* (0.003)	0.022* (0.009)	0.008* (0.003)
Exercises regularly	0.053** (0.017)	0.255** (0.079)	0.055** (0.017)	0.151** (0.048)	0.053** (0.017)
Age, BMI, Country	YES	YES	YES	YES	YES
Observations	3,109	3,109	3,109	3,109	3,109

Does smoking pose a health risk?– logit and probit

- ▶ LPM – interpret the coefficients.
- ▶ Logit, probit - Interpret the *marginal differences*. Basically the same.
 - ▶ Marginal differences are essentially the same across the logit and the probit.
 - ▶ Essentially the same as the corresponding LPM coefficients.
- ▶ Happens often:
 - ▶ We could not know which is the "right model" for inference
 - ▶ Often LPM is good enough for interpretation.
 - ▶ Check if logit/probit very different.
 - ▶ Investigate functional forms if yes.

Goodness of fit measures

- ▶ There is no comprehensively accepted goodness of fit measure...
 - ▶ This is because we do not observe probabilities only 1 and 0...
- ▶ R-squared is not the same meaning as before
 - ▶ Evaluating fit for probability models, we compare predictions that are between zero and one to values that are zero or one.
 - ▶ But predicted probabilities would not fit the zero-one variables, so we'd never get it right.
- ▶ R-squared less natural measure of fit, but we can calculate it as usual.
 - ▶ *But*: R-squared can not be interpreted the same way we did for linear models.

Brier score

- ▶ Brier score

$$Brier = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i^P - y_i)^2$$

- ▶ The Brier score is the average distance (mean squared difference) between predicted probabilities and the actual value of y .
- ▶ Smaller the Brier score, the better.
 - ▶ When comparing two predictions, the one with the smaller Brier score is the better prediction because it produces less (squared) error on average.
- ▶ Related to a main concept in prediction: mean squared error (MSE)

Pseudo R²

- ▶ Pseudo R-squared
 - ▶ Similar to the R-squared – measures the goodness of fit, tailored to binary outcomes.
 - ▶ Many versions of this measure. Most widely used: McFadden's R-squared
 - ▶ Computes the ratio of log-likelihood of the model vs intercept only.
 - ▶ Can be computed for the logit and the probit but not for the linear probability model. (No likelihood function there...)
- ▶ Another alternative is 'Log-loss' measure
 - ▶ Negative number. Better prediction comes with a smaller log-loss in absolute values.

Practical use

- ▶ There are several measures of model fit, they often give the same ranking of models.
- ▶ Do not use: R-squared could be computed for any model, but it no longer has the interpretation we had for linear models with quantitative dependent variable.
- ▶ Only probit vs logit: pseudo R-squared may be used to rank logit and probit models.
- ▶ Use, especially for prediction: Brier score is a metric that can be computed for all models and is used in prediction.

Does smoking pose a health risk?– Goodness of fit

Table: Statistics of goodness of fit for probability predictions models

Statistic	Linear probability	Logit	Probit
R-squared	0.103	0.104	0.104
Brier score	0.215	0.214	0.214
Pseudo R-squared	n.a.	0.080	0.080
Log-loss	-0.621	-0.617	-0.617

Source: `share-health` data. People of age 50 to 60 from 14 European countries who reported to be healthy in 2011. N=3109.

Does smoking pose a health risk?– Goodness of fit

- ▶ Stable ranking – better predictions have a
 - ▶ higher R-squared and pseudo R-squared
 - ▶ and a lower Brier score
 - ▶ a smaller log-loss in absolute values.
- ▶ Logit and the probit are of the same quality.
- ▶ Logit/probit better than the predictions from linear probability model. The differences are small.

Bias of the predictions

- ▶ Post-prediction: we may be interested to study some features of our model
- ▶ One specific goal: evaluating the *bias of the prediction*.
 - ▶ Probability predictions are *unbiased* if they are right on average = the average of predicted probabilities is equal to the actual probability of the outcome.
 - ▶ If the prediction is unbiased, the bias is zero.
- ▶ If, in our data, 20% of observations have $y = 0$ and 80% have $y = 1$, and the average of our prediction is $\sum_{i=1}^N \hat{y}_i / N = 0.8$, then our prediction is unbiased.
- ▶ A large value of bias indicates a greater tendency to underestimate or overestimate the chance of an event.

Calibration

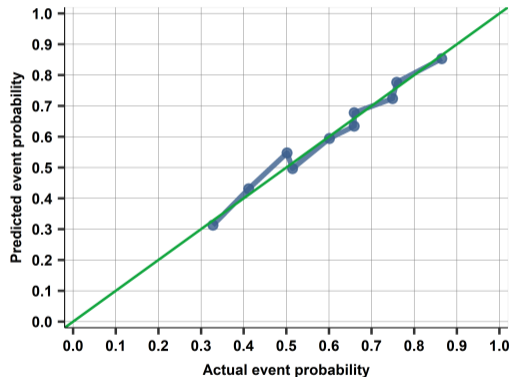
- ▶ Unbiasedness refers to the whole distribution of probability predictions is
- ▶ A finer and stricter concept is *calibration*
 - ▶ A prediction is *well calibrated* if the actual probability of the outcome is equal to the predicted probability for each and every value of the predicted probability.
- ▶ You take predicted probabilities which are around 10% and check the average for the realized outcome. If it is 10%, then the prediction is well calibrated.
- ▶ ‘Calibration curve’ is used to show this.
- ▶ A model may be unbiased (right on average) but not well calibrated
 - ▶ underestimate high probability events and overestimate low probability ones

Calibration curve

- ▶ A *calibration curve*
 - ▶ Horizontal axis shows the values of all predicted probabilities (\hat{y}^P).
 - ▶ Vertical axis shows the fraction of $y = 1$ observations for all observations with the corresponding predicted probability.
- ▶ A well-calibrated case, the calibration curve is close to the 45 degree line.
- ▶ In practice we create bins for predicted probabilities and make comparisons of the actual event's probability.
 - ▶ Use percentiles in general. Some cases equal widths are used (this is a more noisy estimate)

Calibration curve

- ▶ A **calibration curve** for the logit model
- ▶ 10 bins
- ▶ Not only unbiased, but well calibrated!



Probability models summary

- ▶ Find patterns with ease when y is binary - model probability with regressions
- ▶ Linear probability model is mostly good enough, easy inference.
 - ▶ Predicted values could be below 0, above 1
- ▶ Logit (and probit) - better when aim is prediction, predicted values strictly between 0-1
- ▶ Most often, LPM, logit, probit - similar inference
 - ▶ Use marginal (average) differences
- ▶ No trivial goodness of fit. Brier score or pseudo-R-Squared.
- ▶ Calibration is useful diagnostics tool: well-calibrated models will predict a 20% chance for events that tend to happen one out of five cases.