Econometric Models

The Linear Model, Binary Choice, and a Tour of the Toolbox

What Is an Econometric Model?

An econometric model links an **outcome** Y to **explanatory variables** X and to everything else we do not observe, u:

$$Y = f(X) + u$$

Most of the time we model the **conditional mean** $E[Y \mid X]$ — the average outcome among people who share the same X.

Describe

How do wages differ across education levels?

Predict

How likely is this applicant to default?

Explain

What does one more year of school **do** to a wage?

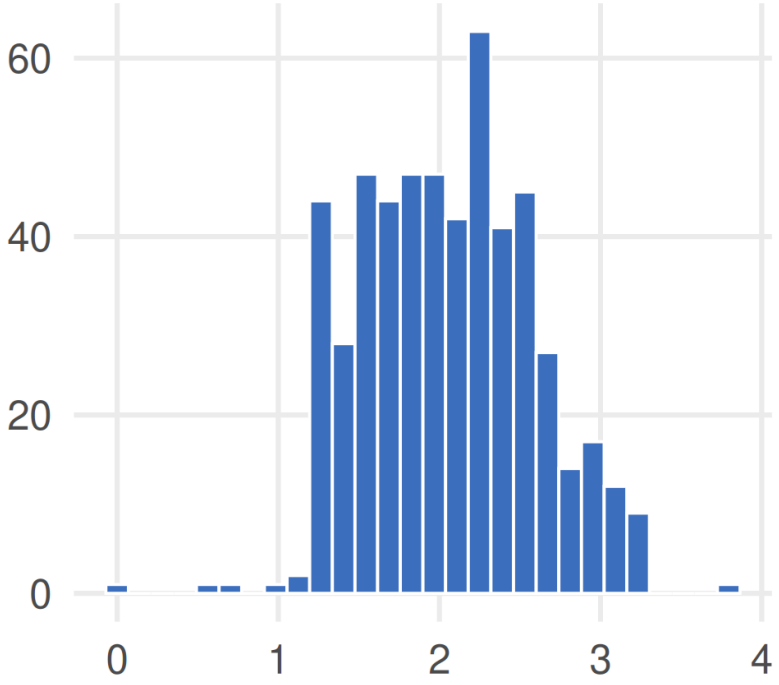
The same regression serves all three — but the third needs the most assumptions.

Three Outcomes, Three Shapes

The **type of outcome** decides which model fits. This deck walks through them in order.

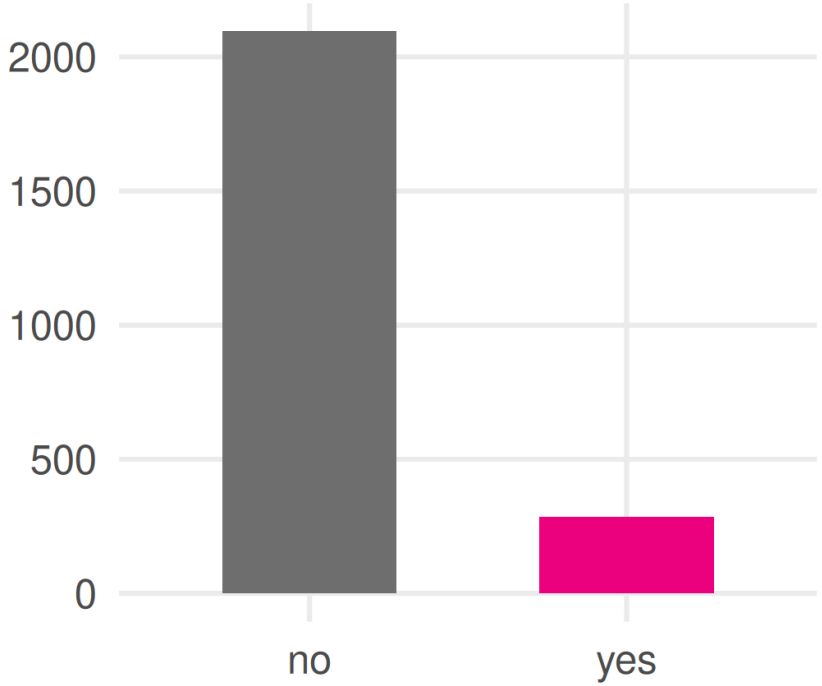
1 · Continuous

log hourly wage, CPS 1985



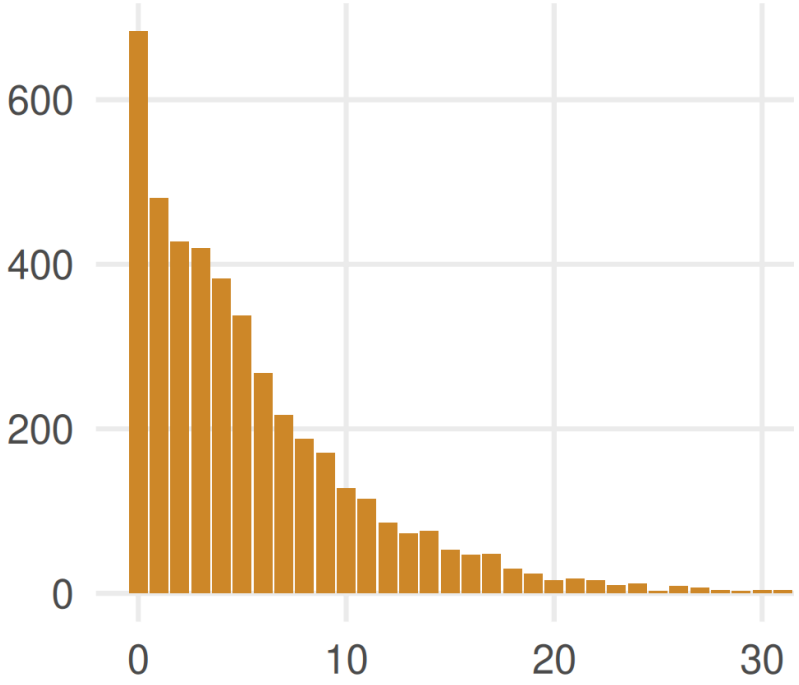
2 · Binary

mortgage denied? Boston 1990



3 · Count (and more)

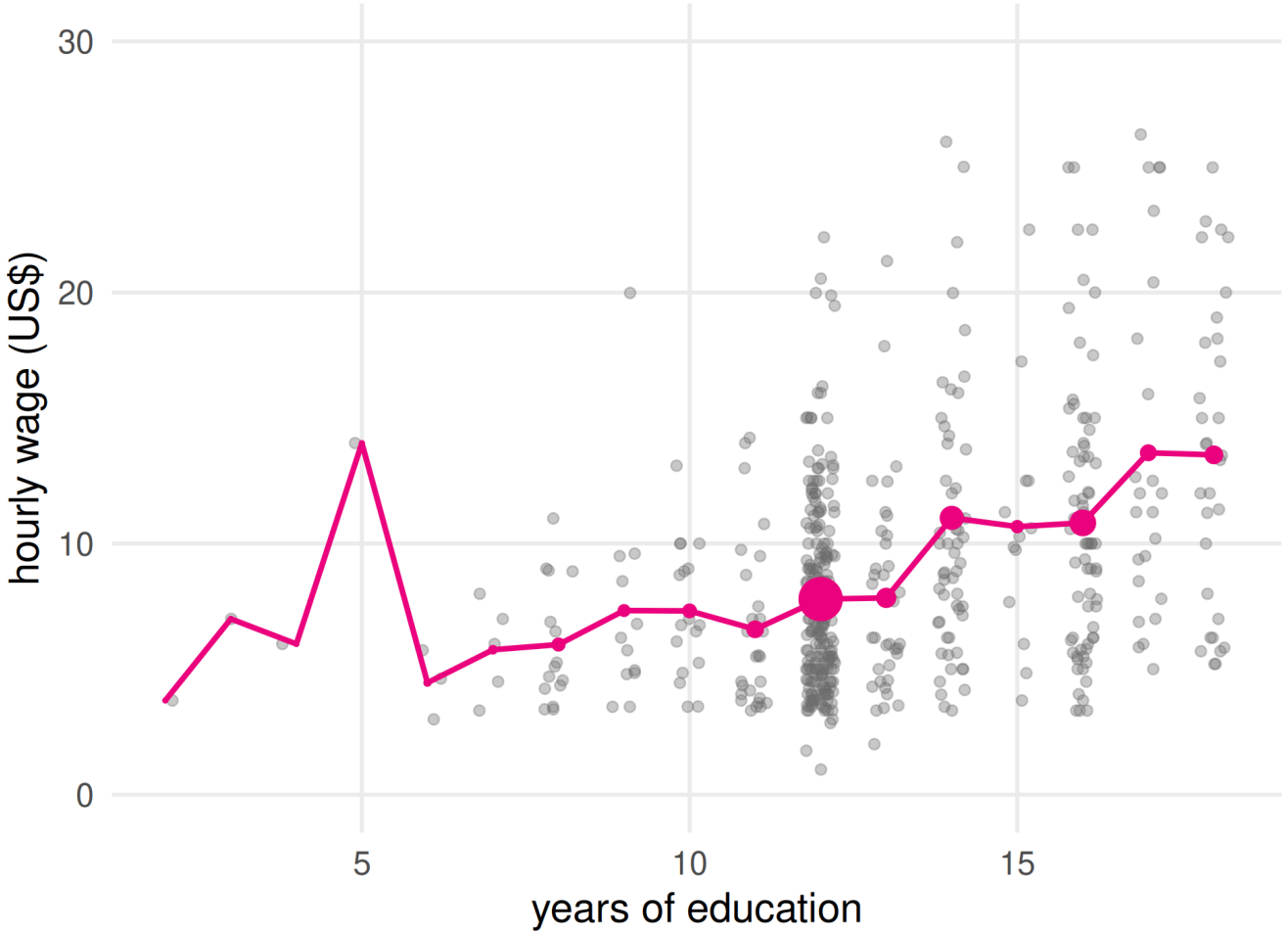
doctor visits per year, NMES 1988



Linear model for the first, **logit / probit** for the second, and a **family of specialised models** for the rest.

1 · The Linear Model

Does Education Pay?



Data: `CPS1985` from the `AER` package — 534 US workers.

- Each grey dot is **one worker**
- Each pink dot is the **average wage** at that education level: $E[\text{wage} \mid \text{educ}]$
- Dot size = number of workers

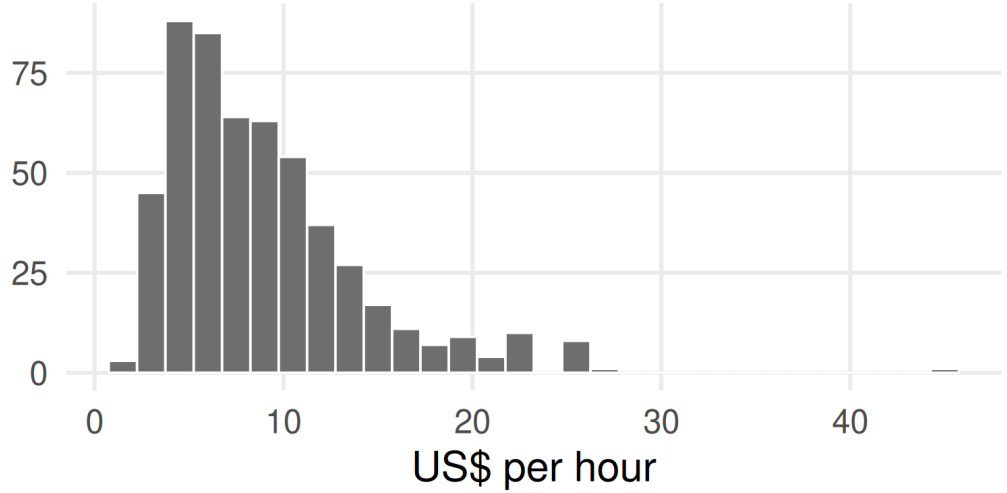
The idea of regression

Summarise the pink dots with a simple formula — a **line**.

Why Log Wages?

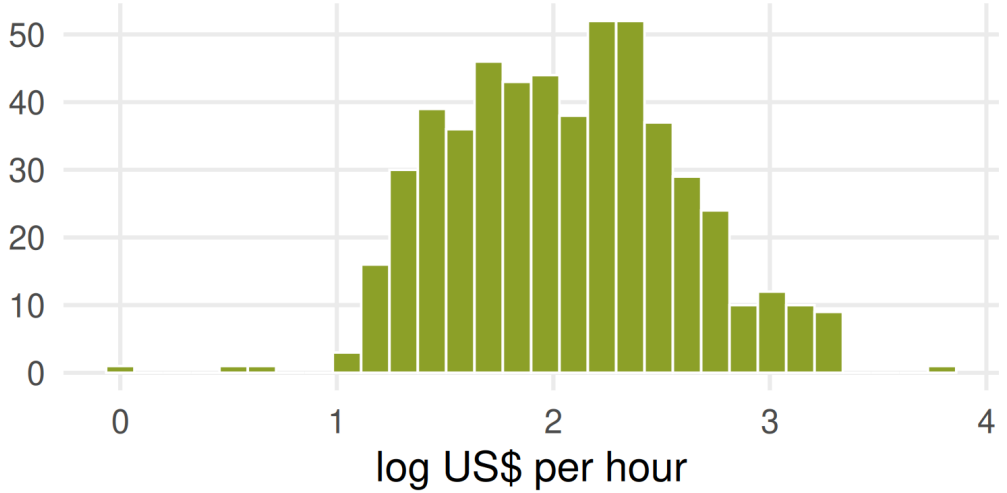
wage

long right tail: a few very high earners



log(wage)

roughly symmetric



Statistical reason

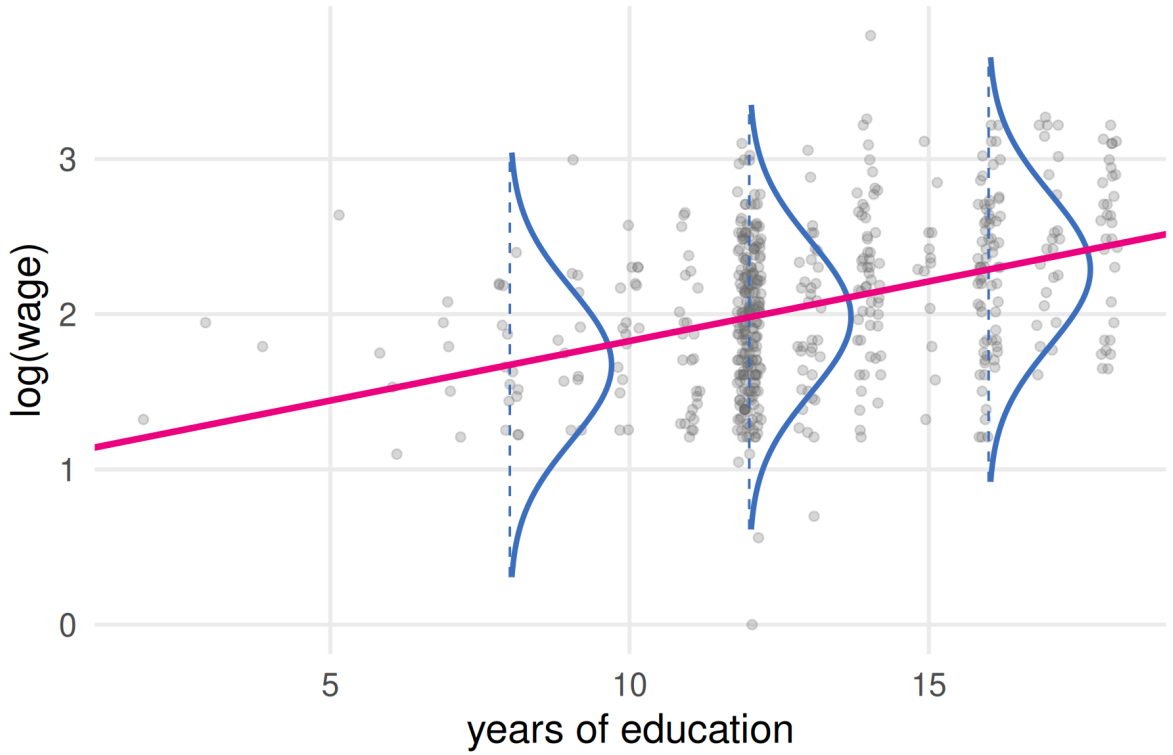
Wages are **skewed**; the log pulls in the tail, so a few extreme earners do not dominate the fit.

Economic reason

Education plausibly raises wages by a **percentage**, not by a fixed dollar amount. In logs, coefficients read as **% changes**.

The Linear Model

$$\log(\text{wage}_i) = \underbrace{\beta_0 + \beta_1\,\text{educ}_i}_{E[Y|X]\text{: systematic part}} + \underbrace{u_i}_{\text{everything else}}$$



- The **pink line** is $E[Y \mid X]$: where the average sits at each education level
- The **blue curves** are the spread of u around it: many people above, many below

The key assumption

$E[u \mid \text{educ}] = 0$: whatever is left in u averages to **zero at every education level**. Then the line really is the conditional mean.

Least Squares, Live

Choose the line yourself. Least squares picks the one with the **smallest sum of squared residuals** (SSR).

slope b_1

0.01

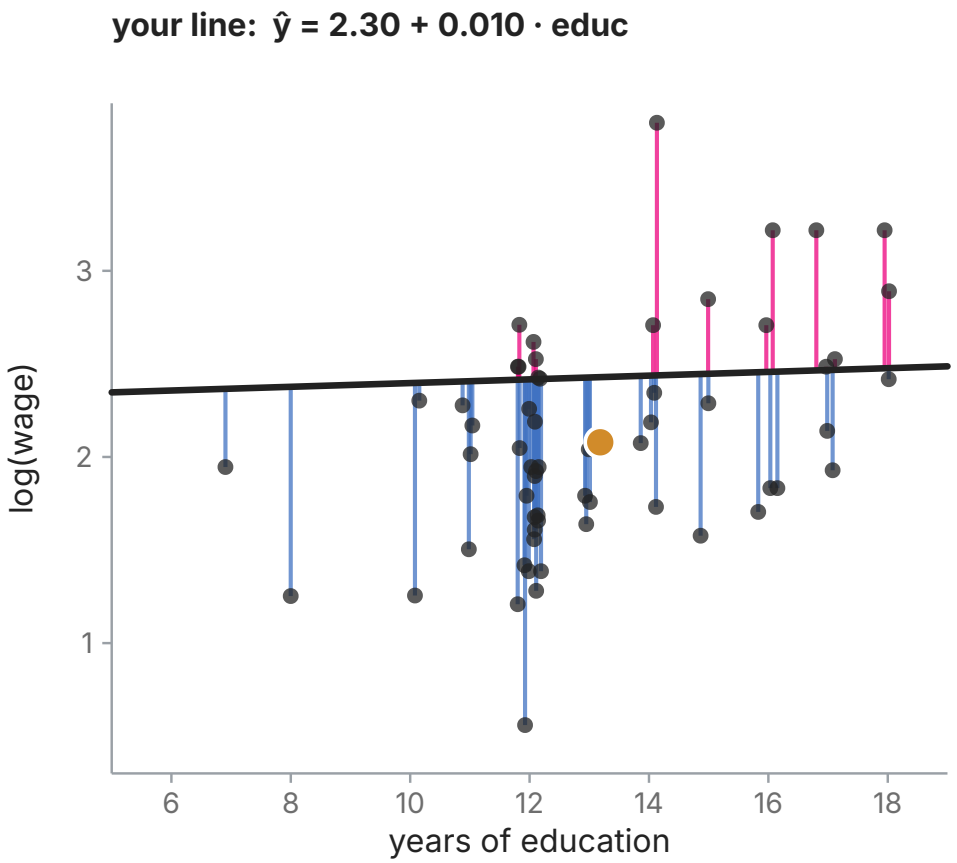
shift up / down

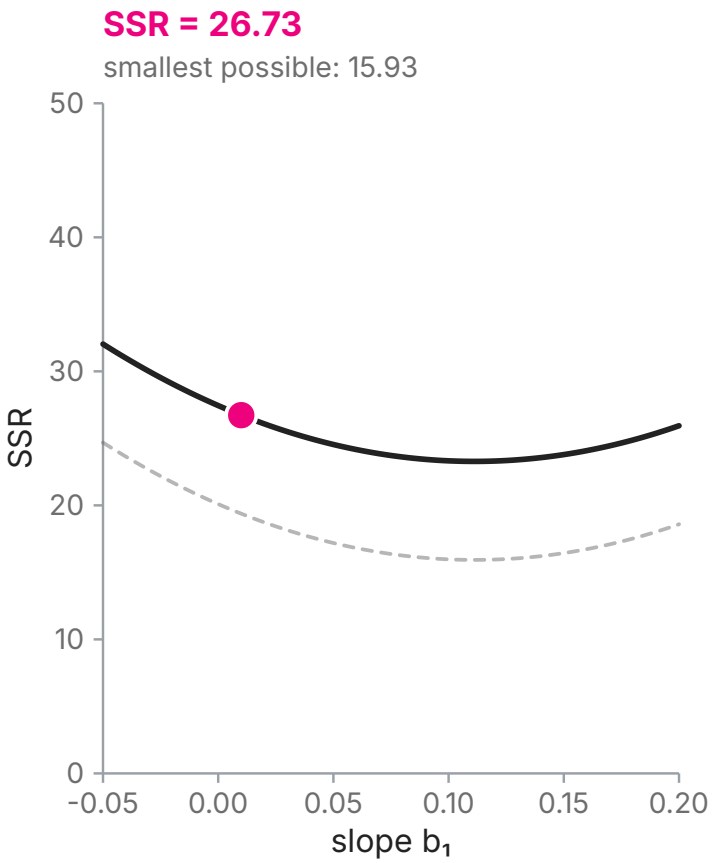
0.35

show the OLS line ☐

60 workers drawn from `CPS1985`. Each vertical segment is a **residual** $y_i - \hat{y}_i$: pink above the line, blue below.

The orange dot is the point of means $(\bar{x}, \bar{y})$. The least-squares line always passes through it, so **shift = 0** is always part of the answer.





Running It in R

```r
m1 <- lm(log(wage) ~ education, data = CPS1985)
summary(m1)
```

```
Call:
lm(formula = log(wage) ~ education, data = CPS1985)

Residuals:
     Min       1Q   Median       3Q      Max
-1.98099 -0.37155  0.03391  0.34975  1.66098

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.059890   0.107432   9.866   <2e-16 ***
education   0.076759   0.008091   9.487   <2e-16 *** (1) (2) (3)
---

Residual standard error: 0.4885 on 532 degrees of freedom (4)
Multiple R-squared:  0.1447,    Adjusted R-squared:  0.1431  (5)
F-statistic: 90.01 on 1 and 532 DF,  p-value: < 2.2e-16
```

(1) **Estimate.** One more year of education goes with 0.077 higher log wage — about **+8%** ($e^{0.077} - 1$).

(2) **Std. Error.** How much the estimate would move from sample to sample.

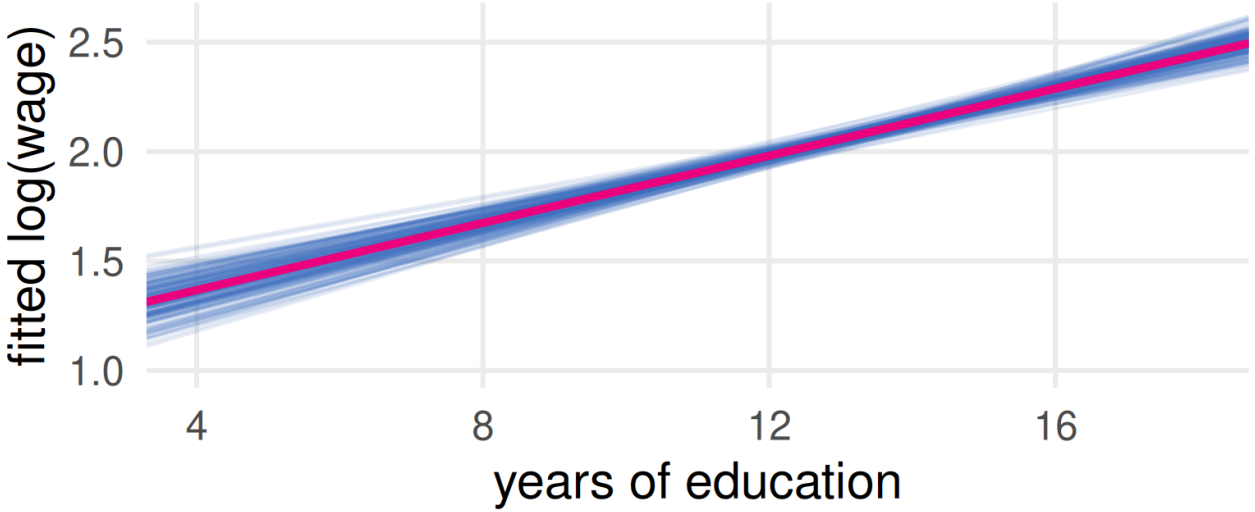
(3) **t value** = estimate / SE, and its **p-value** for $H_0 : \beta_1 = 0$. Here: no doubt that $\beta_1 \neq 0$.

(4) **Residual SE.** The typical size of u: about 0.49 log points.

(5) **R^2.** Education accounts for **14%** of the variation in log wages. The F-statistic tests all slopes $= 0$ at once.

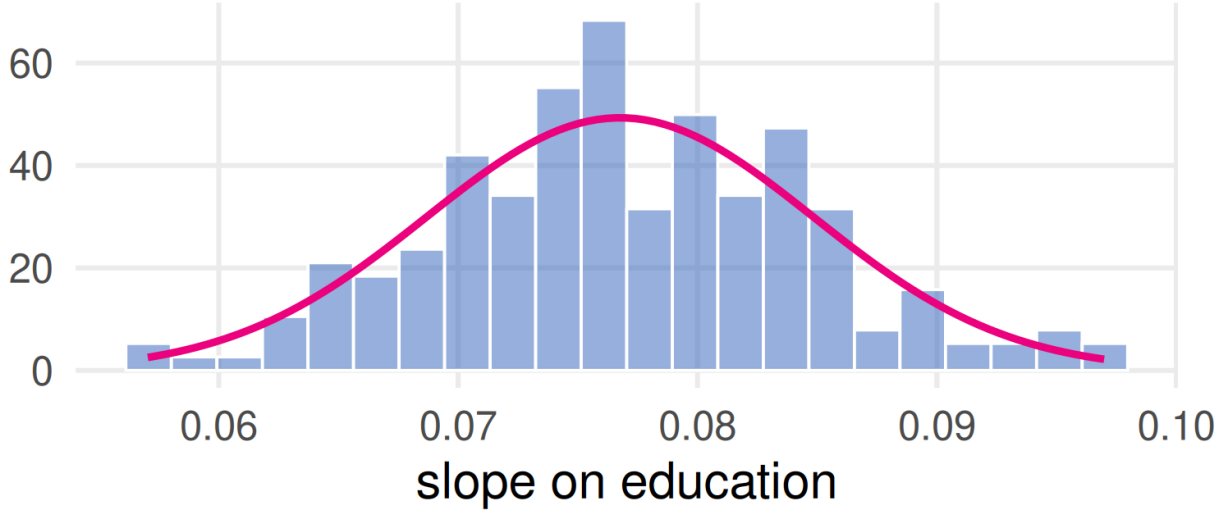
Standard Errors Are Sampling Variability

200 resampled data sets, 200 regression lines



their slopes

pink: normal curve with sd = the reported Std. Error




```
1  confint(m1)
```

```
                2.5 %     97.5 %
(Intercept) 0.84884618 1.27093355
education   0.06086475 0.09265238
```

The **Std. Error** is the width of that histogram — no resampling needed, R computes it from one sample. The 95% interval is roughly **estimate ± 2 SE**: a return of **6% to 9%** per year.

Reading Coefficients: Units Matter

Model	Formula in R	Interpretation of β_1
level–level	`lm(wage ~ educ)`	+1 year $\Rightarrow$ wage changes by β_1 **dollars**
log–level	`lm(log(wage) ~ educ)`	+1 year $\Rightarrow$ wage changes by about $100 \cdot \beta_1$ **%**
level–log	`lm(wage ~ log(sales))`	+1% sales $\Rightarrow$ wage changes by $\beta_1/100$ **dollars**
log–log	`lm(log(q) ~ log(p))`	+1% price $\Rightarrow$ quantity changes by β_1 **%** — an **elasticity**

Exact, not approximate

In log–level models the exact change is $e^{\beta_1} - 1$. For $\beta_1 = 0.077$: 7.98%. For small β the two agree; for $\beta = -0.26$ they do not (-22.7%).

Always state the units

"The coefficient is 0.077" means nothing on its own. "One more year of schooling goes with about 8% higher hourly wages" is an answer.

More Than One Regressor

```r
m2 <- lm(log(wage) ~ education + experience + I(experience^2) +
           gender, data = CPS1985)
summary(m2)
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.6007445  0.1194927   5.027 6.81e-07 ***
education        0.0912936  0.0080049  11.405  < 2e-16 ***  (1)
experience       0.0360522  0.0054352   6.633 8.14e-11 ***  (2)
I(experience^2) -0.0005412  0.0001197  -4.520 7.64e-06 ***  (3)
genderfemale    -0.2570355  0.0387066  -6.641 7.77e-11 ***  (4)

Residual standard error: 0.4442 on 529 degrees of freedom
Multiple R-squared:  0.2968,    Adjusted R-squared:  0.2915   (5)
F-statistic: 55.81 on 4 and 529 DF,  p-value: < 2.2e-16
```

Each coefficient is now an effect **holding the other regressors fixed** — *ceteris paribus*.

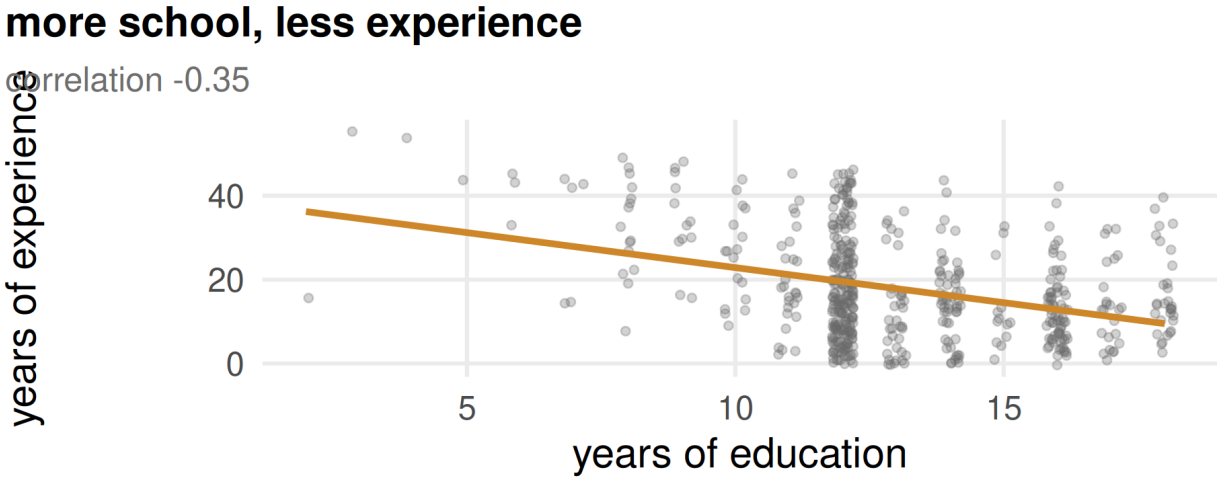
(1) Return to education **rises** from 0.077 to 0.091. Why? Next slide.

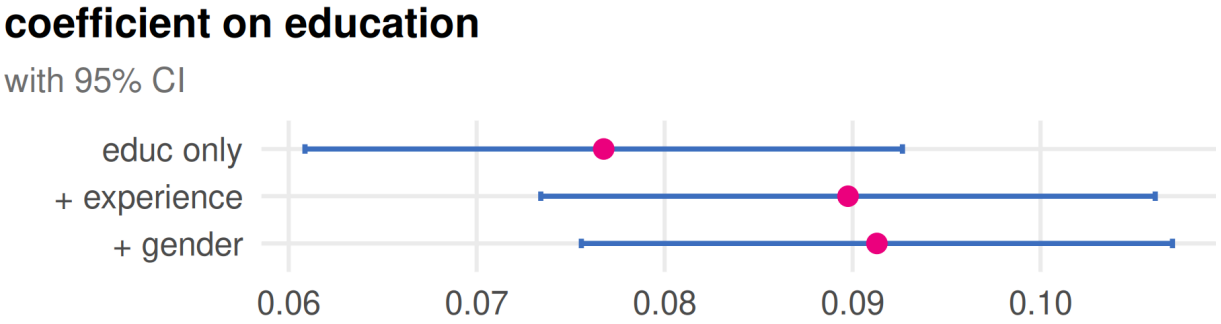
(2) (3) Experience raises wages, but by **less and less** — the square lets the profile bend.

(4) Women earn about **23% less** ($e^{-0.257} - 1$) at the same education and experience.

(5) Adjusted R^2 penalises extra regressors; it doubled.

Why Did the Education Coefficient Move?





Leaving **experience** out of the model puts it into u. It is **not** harmless there:

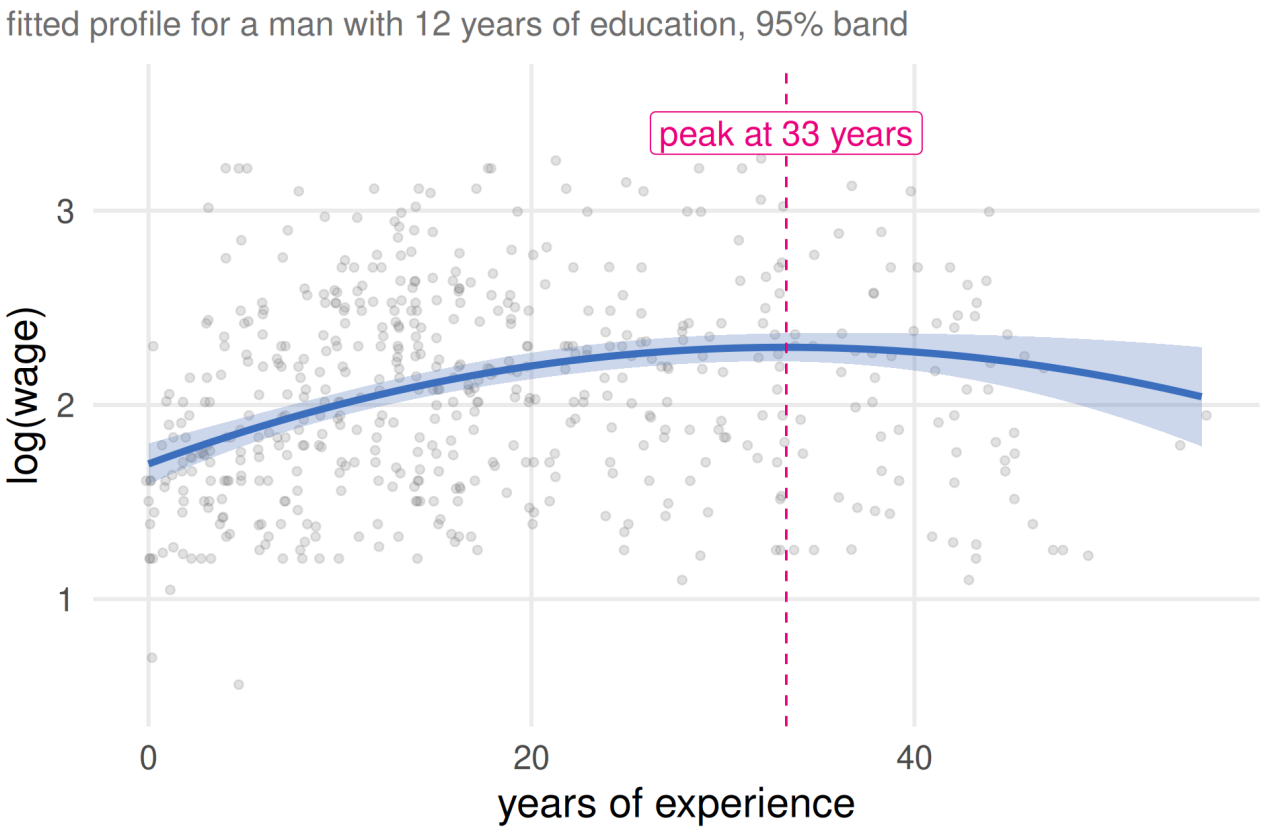
- experience **raises** wages $(+)$
- experience is **lower** for the more educated $(-)$

Omitted variable bias

bias $=$ (effect of the omitted variable on Y) $\times$ (how it moves with X)

Here $(+) \times (-)$: the short regression **understates** the return to education.

Curves Are Still "Linear" Models



$$\ldots + \beta_2\,\mathrm{exper} + \beta_3\,\mathrm{exper}^2$$

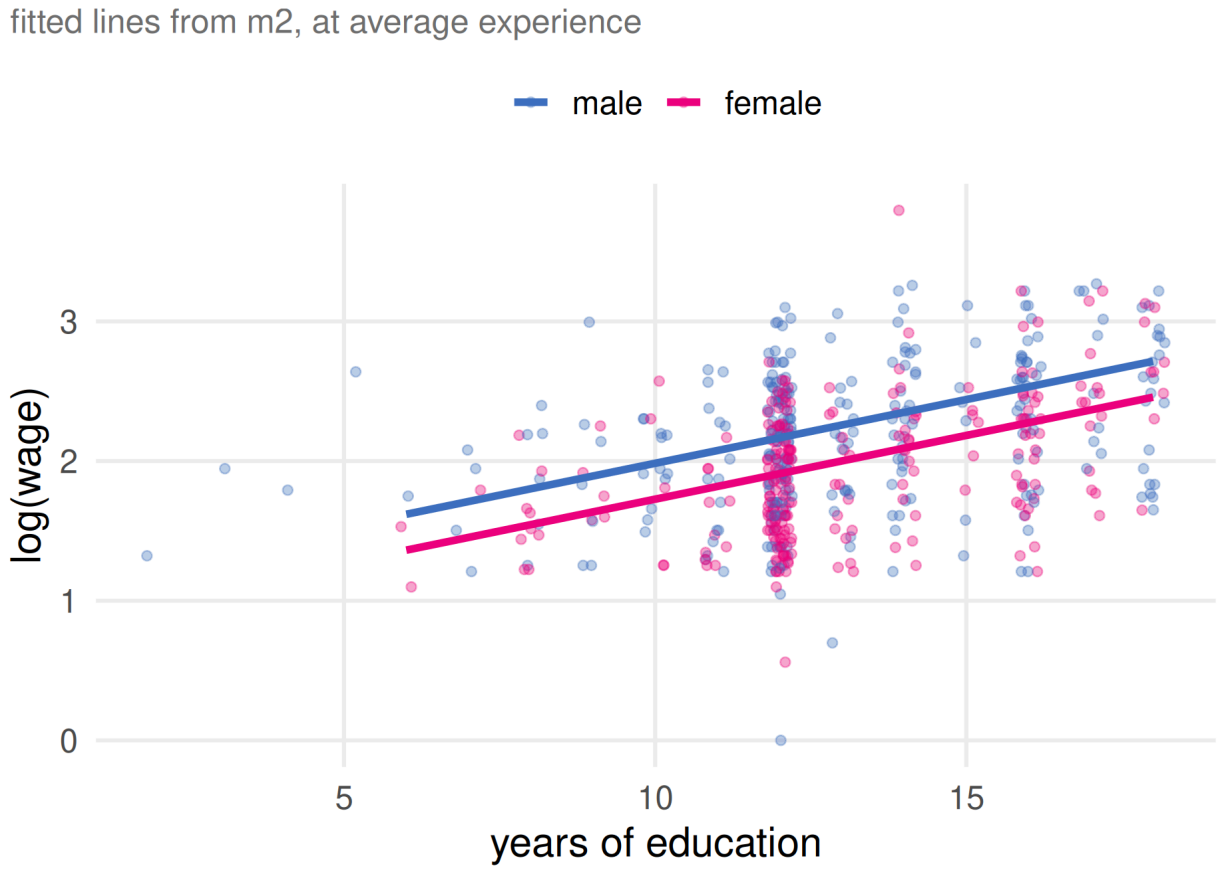
"Linear" means linear in the **coefficients** β, not in X. Squares, logs and interactions all work in `lm()`.

One more year of experience now adds $\beta_2 + 2\beta_3\,\mathrm{exper}$: a lot early in a career, **zero at the peak** $-\hat{\beta}_2/(2\hat{\beta}_3) \approx 33$ years.

In R

Write `I(experience^2)` — inside a formula, a bare `^` means something else.

Dummy Variables Shift the Line



`gender` is a **factor**; R turns it into a 0/1 dummy `genderfemale` automatically. The first level (`male`) is the **reference group**.

The dummy moves the **intercept**:

- men: $\beta_0 + \ldots$
- women: $(\beta_0 - 0.257) + \ldots$

Parallel by construction

The gap is the **same at every education level** — because we assumed it, not because the data said so. To let it differ, interact.

Interactions Let the Slopes Differ

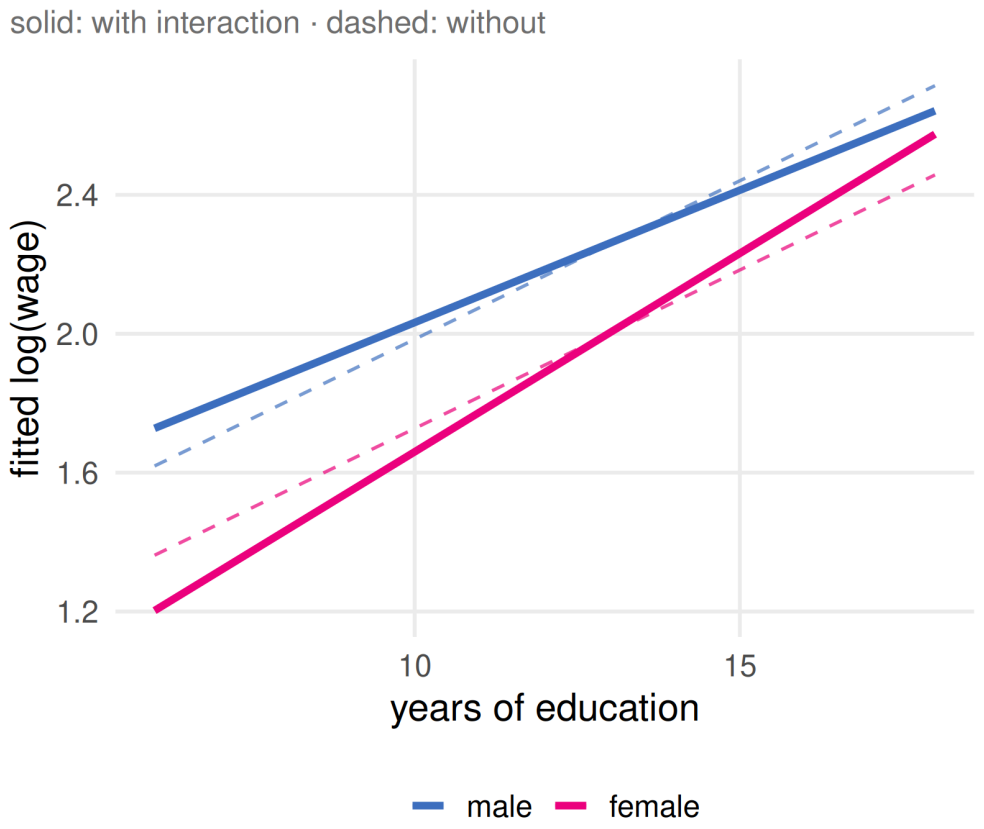
```r
m3 <- lm(log(wage) ~ education * gender + experience +
           I(experience^2), data = CPS1985)
```

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.7896897  0.1402019   5.633 2.89e-08 ***
education                0.0762191  0.0099290   7.676 7.97e-14 ***  (1)
genderfemale            -0.7539256  0.1992145  -3.784 0.000172 ***  (2)
experience               0.0369215  0.0054182   6.814 2.59e-11 ***
I(experience^2)         -0.0005585  0.0001193  -4.681 3.64e-06 ***
education:genderfemale   0.0381424  0.0150037   2.542 0.011300 *    (3)
```

(1) Return to education for **men**: 7.6%.

(2) Gender gap at **zero** education — an extrapolation, do not read it alone.

(3) **Extra** return for women: $0.076 + 0.038 = 0.114$ per year.



The gender gap **narrows** with education.

Testing Several Coefficients at Once

The t-test in `summary()` checks **one** coefficient. To ask "do experience and gender matter **at all**?" compare the two models with an **F-test**:

```
anova(m1, m2)
```

```
Analysis of Variance Table

Model 1: log(wage) ~ education
Model 2: log(wage) ~ education + experience + I(experience^2) + gender
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    532 126.97                                 (1)
2    529 104.39  3    22.576 38.134 < 2.2e-16 *** (2) (3)
```

(1) SSR of the **restricted** model — experience and gender set to zero.

(2) SSR of the **full** model: 3 extra coefficients, 22.6 less squared error.

(3) Is that drop larger than chance would produce? $F = 38$, $p \approx 0$: **yes**.

The logic from the testing deck

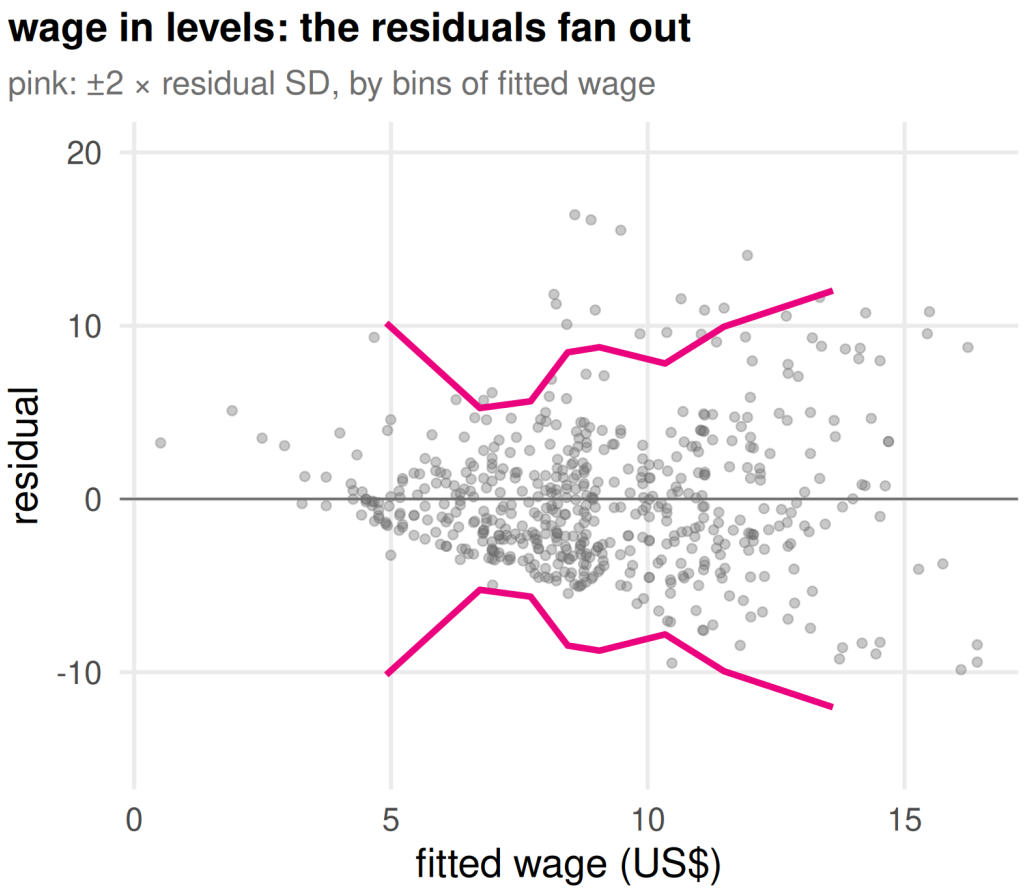
Same recipe as before: a test statistic, its distribution under H_0, a p-value. `car::linearHypothesis()` tests any other linear restriction, e.g. $\beta_{\text{exper}} = \beta_{\text{educ}}$.

The Assumptions, in Plain Words

Assumption	In words	What fails without it
Linear in parameters	Y is a sum of $\beta\times$ (something built from X)	The fitted line is the wrong shape
Random sample	Observations are independent draws from the population	Standard errors — see clustering
No perfect collinearity	No regressor is an exact combination of others	R cannot separate them: `NA` coefficient
Exogeneity $E[u \mid X] = 0$	The omitted stuff is unrelated to X	**The estimates themselves are biased**
Homoskedasticity	The spread of u is the same for every X	Only the usual standard errors
Normal errors	u is normally distributed	Nothing, once n is moderate (CLT)

> **Two very different kinds of failure**
>
> The first four give you the **right number**. The last two only give you the **right uncertainty** — and are easy to fix. Exogeneity is the one that keeps econometricians busy.

Heteroskedasticity: Look at the Residuals




```
 1  m0 <- lm(wage ~ education + experience +
 2             I(experience^2) + gender, data = CPS1985)
 3  round(cbind(usual  = sqrt(diag(vcov(m0))),
 4              robust = sqrt(diag(vcovHC(m0, type = "HC1"
```

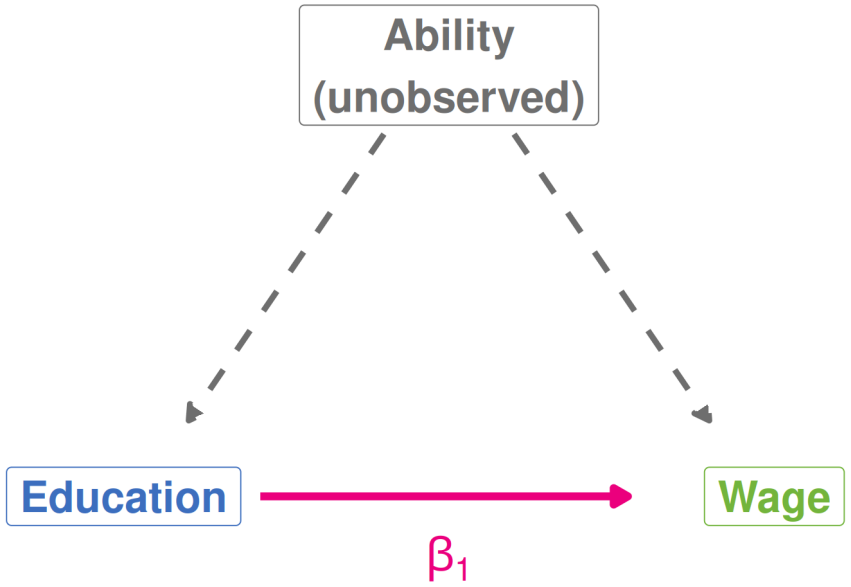
```
                  usual robust
(Intercept)      1.1888 1.2493
education        0.0796 0.0880
experience       0.0541 0.0631
I(experience^2) 0.0012 0.0013
genderfemale     0.3851 0.3936
```

- The **estimates do not change** — only the standard errors
- Robust ("sandwich", HC1) SEs are valid **whether or not** the spread is constant
- In a table of results: `coeftest(m0, vcov = vcovHC(m0, type = "HC1"))`

Rule of thumb

Use robust standard errors **by default**.

Regression Is Not Yet Causation



Ability raises **both** schooling and wages. It sits in u, so $E[u \mid \text{educ}] \neq 0$ and $\hat{\beta}_1$ mixes the effect of school with the effect of ability.

What $\hat{\beta}_1 = 0.077$ is

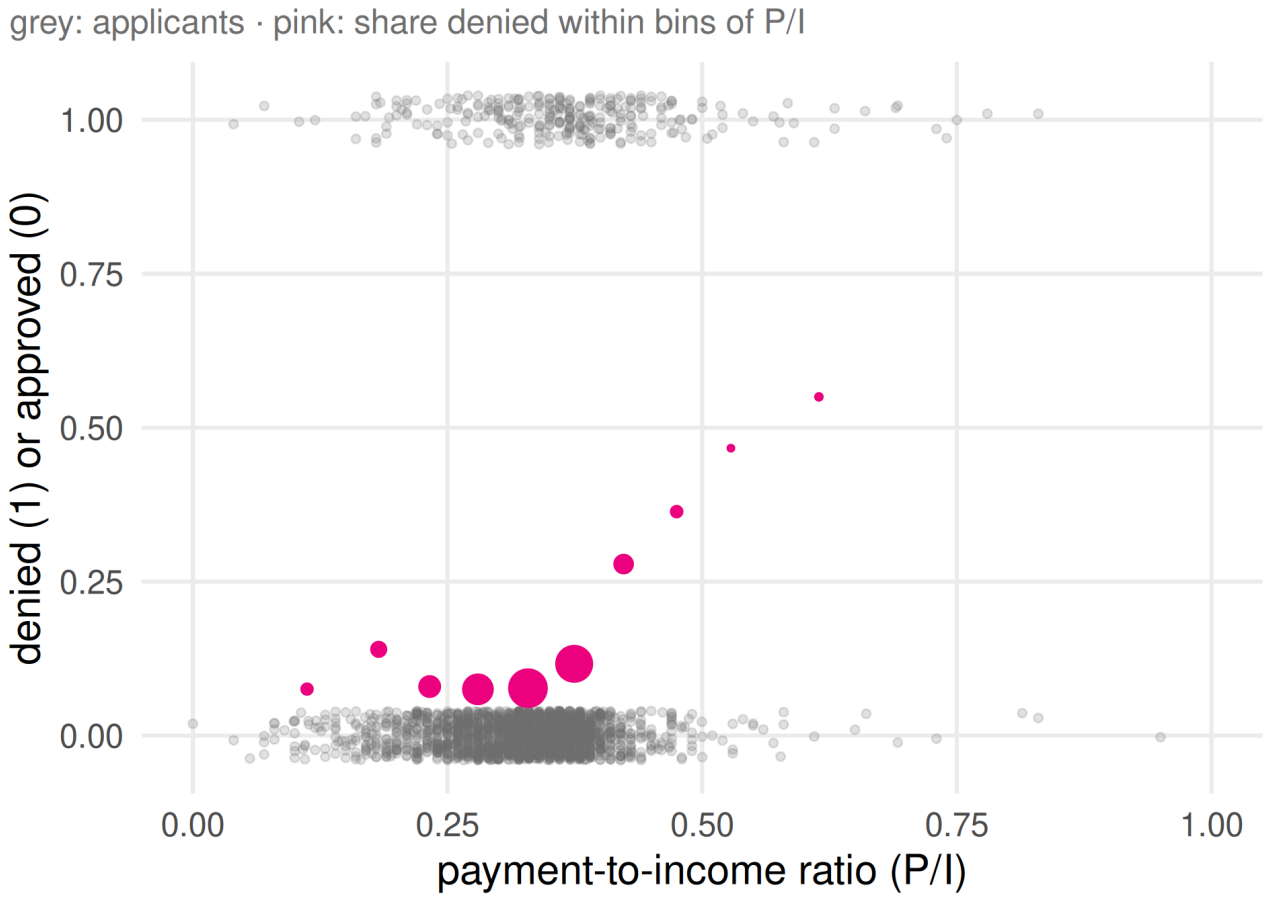
A well-estimated **conditional mean difference**: how wages differ between people with different schooling.

What it is not (yet)

The effect of **sending** someone to school for one more year. That needs a design — **panel data, IV, DiD** in part 3.

2 · Binary Choice

When the Outcome Is Yes or No



Labour-force participation, loan default, buying a car, migrating, voting: many outcomes are **0 or 1**.

Data: `HMDA` — 2,380 Boston mortgage applications, **12% denied**. The key regressor is the ratio of monthly debt payments to income.

The conditional mean is a probability

For $Y \in \{0, 1\}$:

$$E[Y \mid X] = P(Y = 1 \mid X)$$

So a model for $E[Y \mid X]$ is a model for the **probability** of a yes.

The Linear Probability Model

Just run OLS on the 0/1 outcome — a **linear probability model** (LPM):

```r
lpm <- lm(I(deny == "yes") ~ pirat, data = HMDA)
coeftest(lpm, vcov = vcovHC(lpm, type = "HC1"))
```

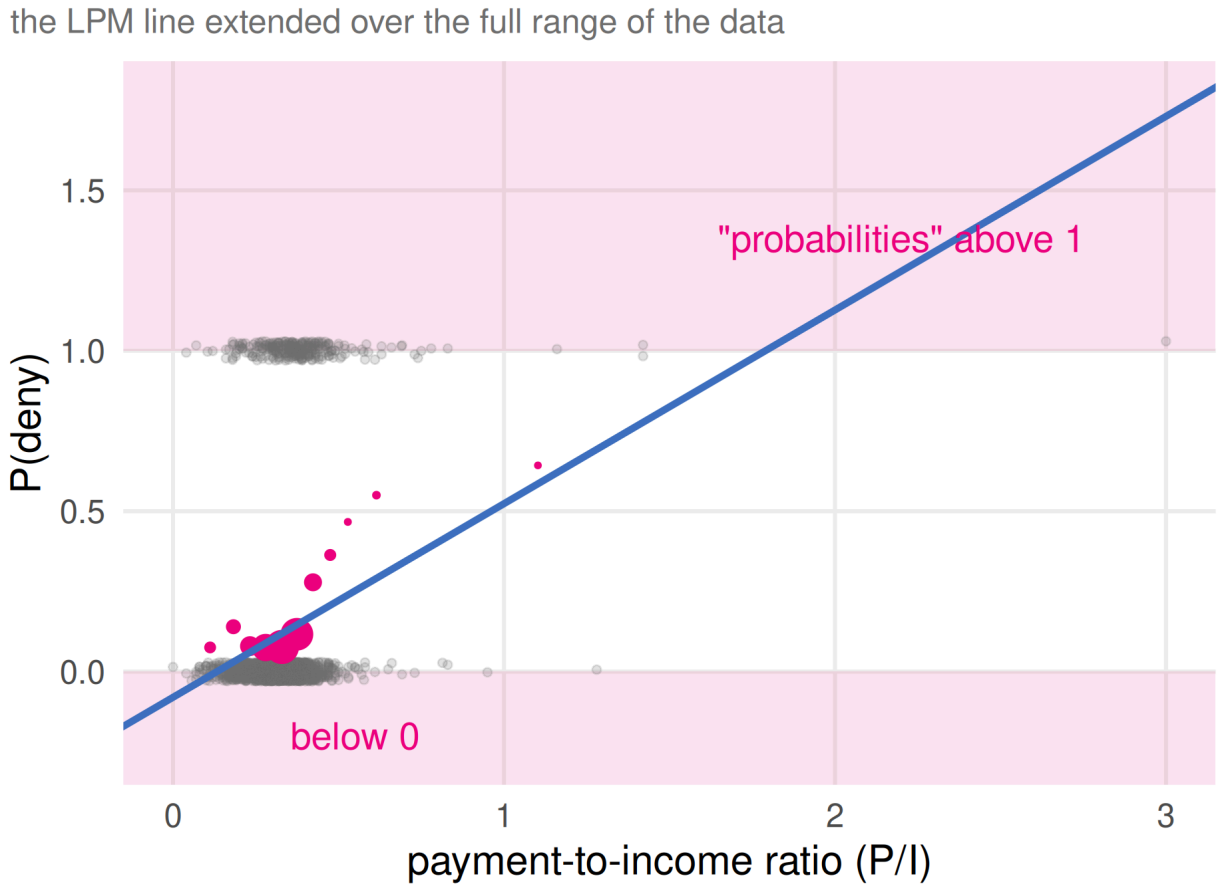
```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -0.079910   0.031967 -2.4998   0.01249 *    3
pirat        0.603535   0.098483  6.1283 1.036e-09 *** 1 2
```

1. Coefficients are **changes in probability**: $+0.1$ in P/I raises the denial probability by 0.06 — **6 percentage points**.
2. Robust SEs are **mandatory** here: with a 0/1 outcome, $\mathbf{Var}(u \mid X) = p(1-p)$ changes with X by construction.
3. The intercept is a **negative probability**. That is a warning sign.

Why people like it

Easy to fit, easy to read, and its slopes are usually close to the average effects of the fancier models. A sensible first look.

Where the Straight Line Breaks



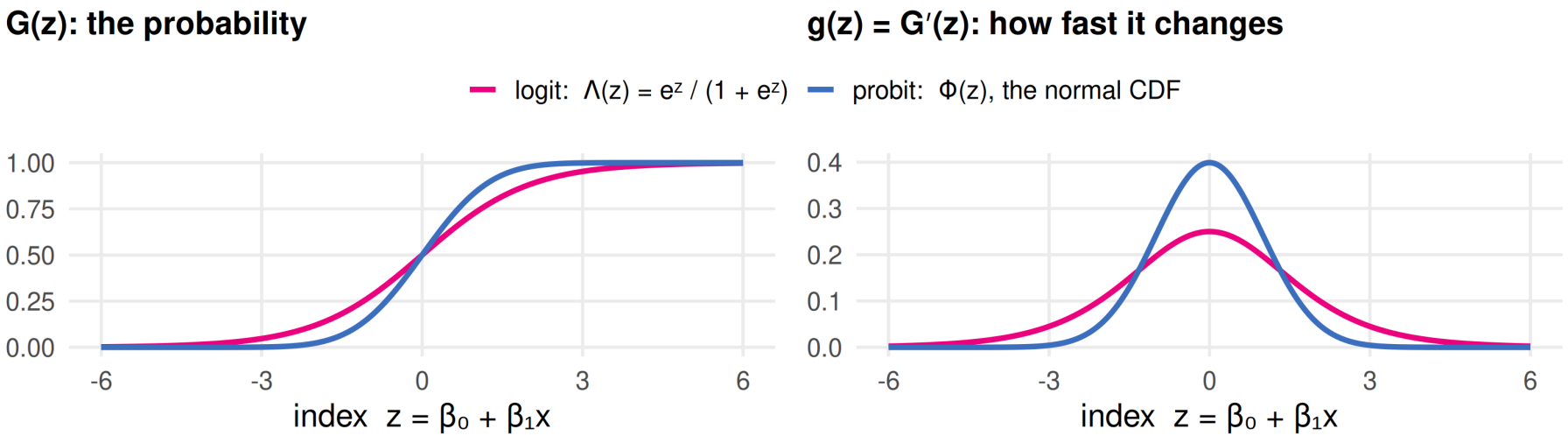
- **Impossible predictions.** At P/I = 3 (it happens) the LPM predicts a denial probability of 1.73
- **Constant effects.** +0.1 in P/I adds 6 points whether you start at 1% or at 95% — but near 0 and 1 a probability **cannot** keep moving at the same pace
- **Built-in heteroskedasticity** — fixed by robust SEs

What we want

A curve that stays inside $[0, 1]$ and **flattens** at both ends.

Bend the Line: The S-Curve

$$P(Y = 1 \mid X) = G(\beta_0 + \beta_1 X), \qquad G : \mathbb{R} \to [0, 1] \text{ increasing}$$



Two choices of G

Logit uses the logistic CDF, **probit** the normal CDF. Same shape; the logistic has fatter tails.

Different scales

The logistic is more spread out, so logit coefficients are about **1.6×** probit coefficients. Only the **probabilities** are comparable.

Where the S-Curve Comes From, Live

Each applicant has a latent "denial score" $y^* = \beta_0 + \beta_1 x + \varepsilon$. The bank denies if $y^* > 0$. So $P(\text{deny}) = P(\varepsilon > -\beta_0 - \beta_1 x) = G(\beta_0 + \beta_1 x)$.

P/I ratio x

0.33

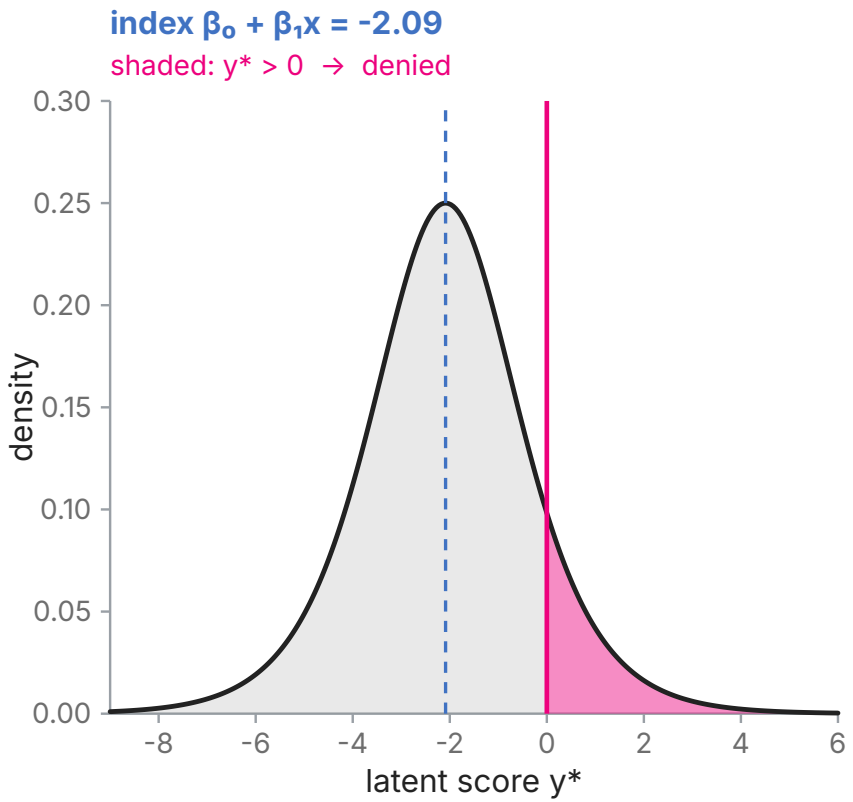
error ε

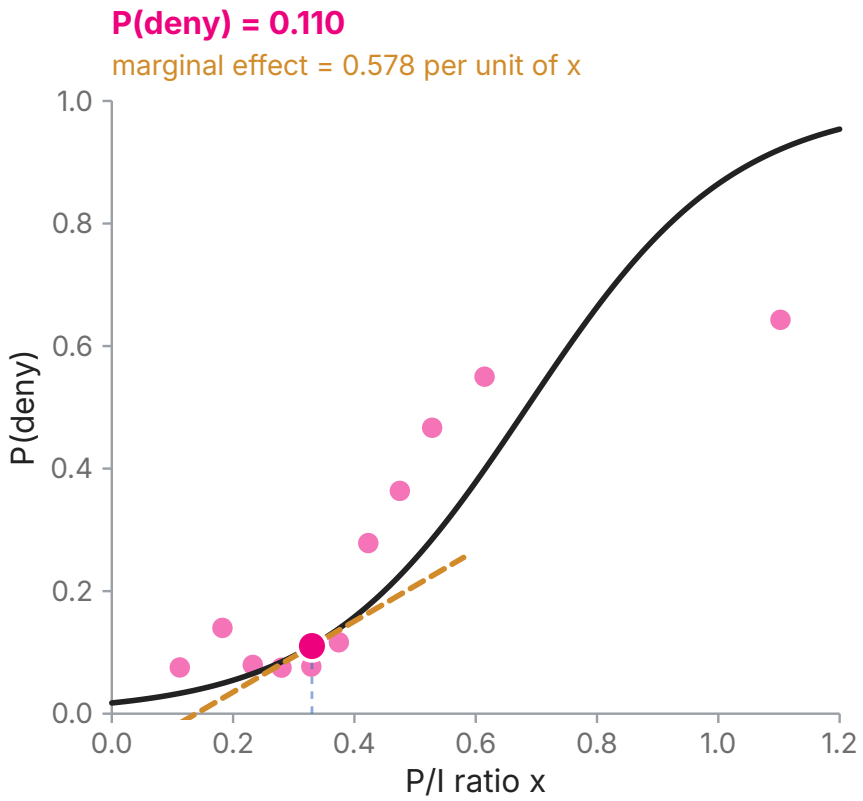
◉ logit ○ probit

Coefficients are the ones fitted to `HMDA` on the next slides.

Left: the distribution of y^* for applicants with this x. The shaded area beyond 0 is the probability of denial.

Right: that area, traced out over x. The dashed tangent is the **marginal effect** at this x: density $\times \beta_1$.





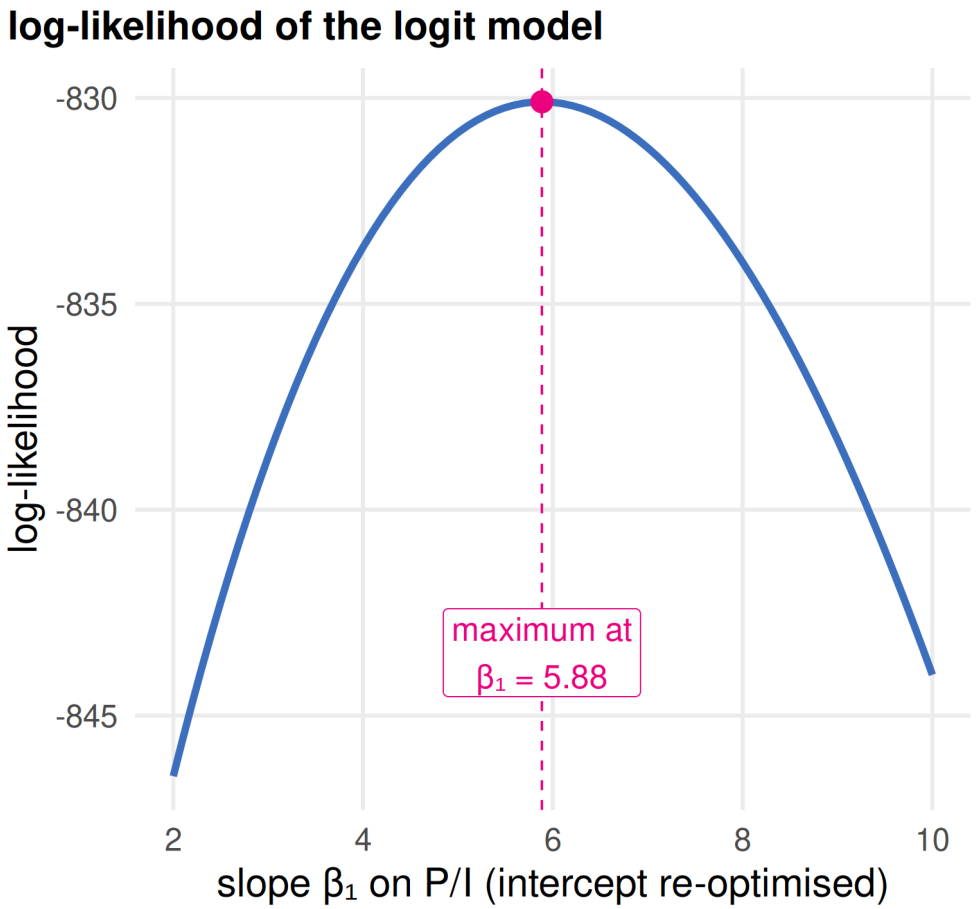
Estimation: Maximum Likelihood

There is no "residual" to square. Instead ask: which β makes the **observed** denials and approvals most probable?

- a denied applicant contributes $p_i = G(\beta_0 + \beta_1 x_i)$
- an approved one contributes $1 - p_i$

$$\ell(\beta) = \sum_i \Big[y_i \log p_i + (1 - y_i) \log(1 - p_i) \Big]$$

R climbs this function numerically — no formula like OLS.



From the estimation deck

This is the MLE recipe exactly. Every likelihood is a loss; OLS was the special case with normal errors.

Logit in R

```r
lg <- glm(deny ~ pirat, family = binomial(link = "logit"),
          data = HMDA)
summary(lg)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.0284     0.2686 -14.999  < 2e-16 ***
pirat         5.8845     0.7336   8.021 1.05e-15 ***  (1) (2)

    Null deviance: 1744.2  on 2379  degrees of freedom
Residual deviance: 1660.2  on 2378  degrees of freedom  (3)
AIC: 1664.2  (4)

Number of Fisher Scoring iterations: 5  (5)
```

(1) The coefficient is on the **index** $\beta_0 + \beta_1 x$, i.e. on the **log-odds**. Its **sign** and **significance** are readable; its size is not a probability.

(2) **z value**, not t: MLE standard errors rely on large samples.

(3) **Deviance** $= -2\times$ log-likelihood. The drop from the null deviance (1744) is a likelihood-ratio test of the slope.

(4) **AIC**: deviance + a penalty for parameters; smaller is better when comparing models.

(5) Steps of the numerical optimiser.

Making the Coefficient Speak

Odds ratios — logit only:

```
exp(coef(lg))
```

```
(Intercept)       pirat
 0.01780222 359.42220584
```

```
exp(0.1 * coef(lg)["pirat"])   # per 0.1 of P/I
```

```
   pirat
1.801194
```

Raising P/I by 0.1 multiplies the **odds** of denial, $p/(1-p)$, by 1.8. Common in medicine; many find odds hard to think in.

Predicted probabilities — always work:

```
predict(lg, newdata = data.frame(pirat = c(0.3, 0
        type = "response")
```

```
         1          2
0.09422692 0.25233766
```

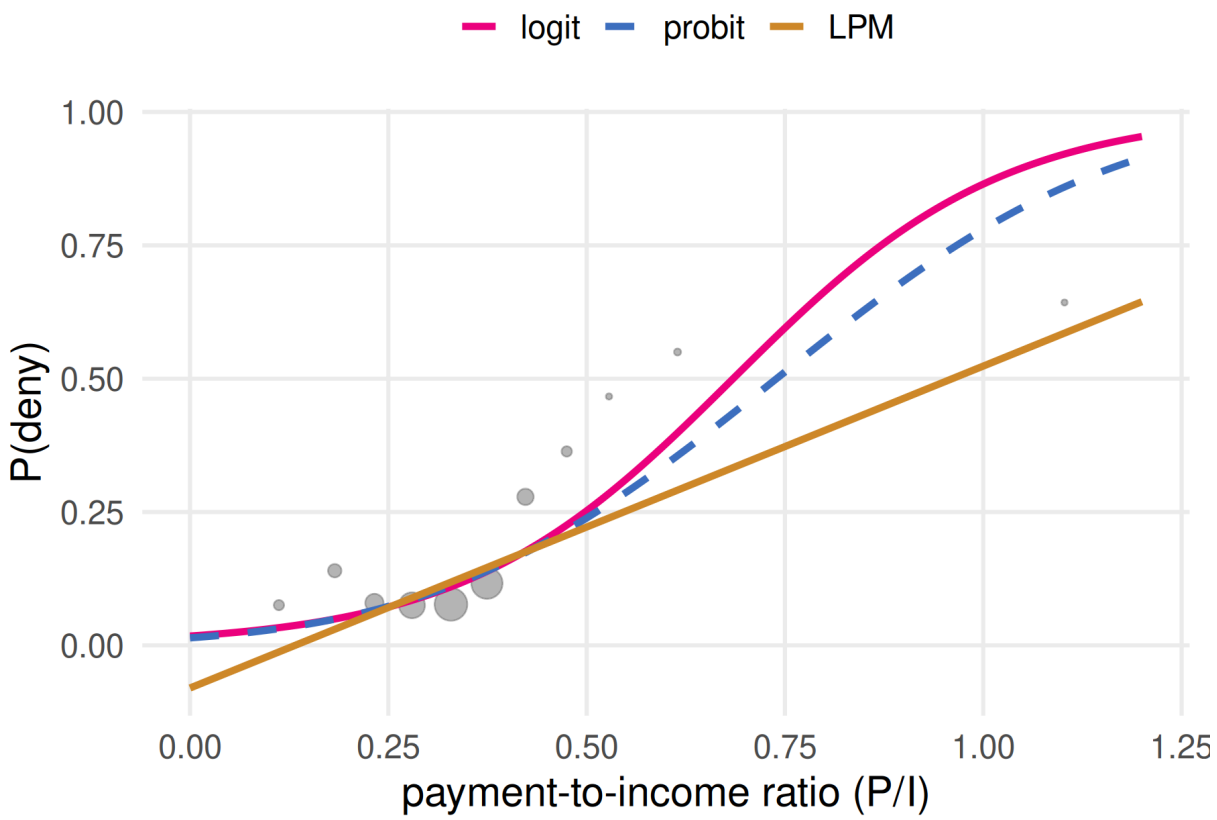
Moving from P/I 0.3 to 0.5 raises the denial probability from **9% to 25%**. `type = "response"` returns $G(\hat{\beta}_0 + \hat{\beta}_1 x)$; the default `"link"` returns the index.

What you can read straight from the logit table

Sign (direction), **significance** (whether), and **ratios** of two coefficients (relative importance). Not: “how many percentage points”.

Logit, Probit, LPM: Same Story

grey: share denied in bins of P/I



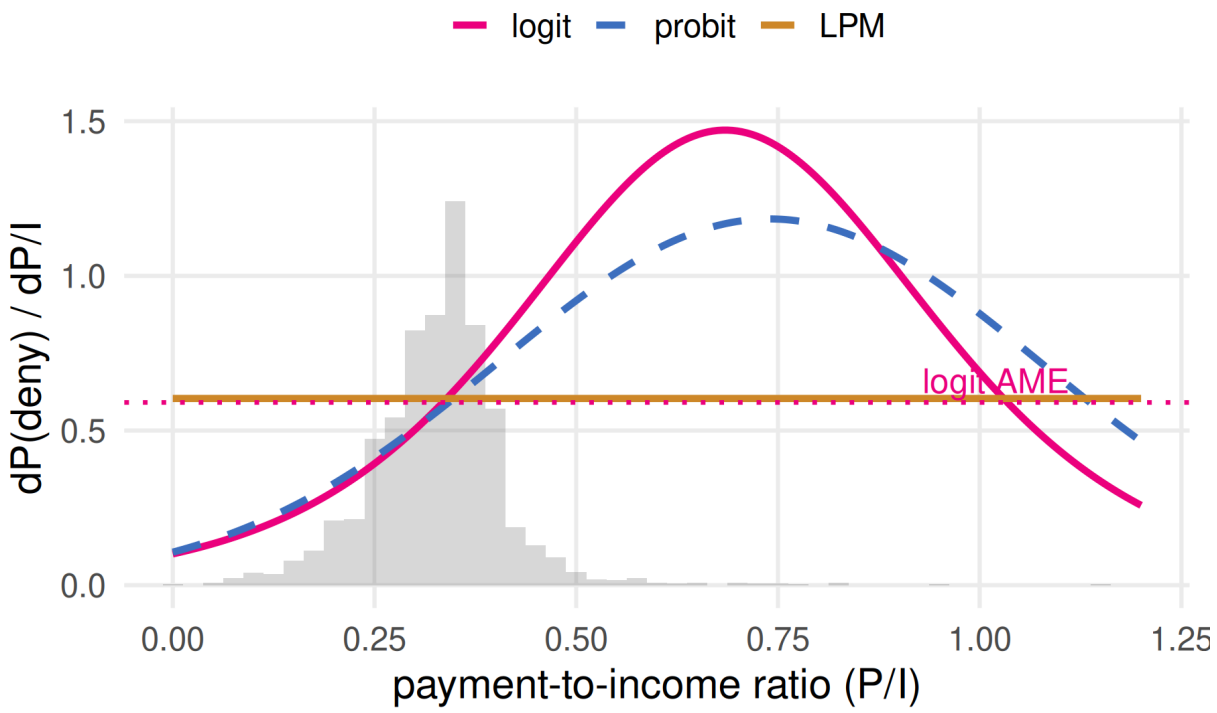

```
1  rbind(logit  = coef(lg),
2        probit = coef(pb),
3        LPM    = coef(lpm))
```

```
        (Intercept)     pirat
logit   -4.02843220 5.8844978
probit  -2.19414501 2.9678666
LPM     -0.07990965 0.6035349
```

- **Coefficients** differ a lot: 5.9 vs 3.0 vs 0.6 — different scales, not different findings
- **Predicted probabilities** are nearly identical where the data are
- Logit vs probit is rarely a substantive choice. Pick one and report **probabilities** or **marginal effects**

Marginal Effects Vary Along the Curve

grey: where the applicants are



In logit/probit the effect of x on the probability is

$$\frac{\partial P}{\partial x} = g(\beta_0 + \beta_1 x) \cdot \beta_1$$

— it **depends on where you stand**. Two ways to boil it down to one number:

Two one-number summaries

AME (average marginal effect): the effect for **every** person, then averaged — the standard choice.

MEM (effect at the mean): plug in the average person, who may not exist.

Average Marginal Effects in R

```r
eta <- predict(lg, type = "link")       # index, per perso
mean(dlogis(eta)) * coef(lg)["pirat"]  # logit AME
```

```
    pirat
0.5907417
```

```r
eta_p <- predict(pb, type = "link")
mean(dnorm(eta_p)) * coef(pb)["pirat"] # probit AME
```

```
    pirat
0.5664755
```

```r
coef(lpm)["pirat"]                      # LPM slope
```

```
    pirat
0.6035349
```

All three models now say the same: **+0.1 in P/I ≈ +6 percentage points** in the denial probability, on average.

In practice

The package `marginaleffects` does this for any model, with standard errors:

`avg_slopes(lg)`

For a dummy it averages the change in predicted probability from switching 0 → 1.

A Richer Model

```r
lg3 <- glm(deny ~ pirat + lvrat + phist + insurance + afam,
           family = binomial(link = "logit"), data = HMDA)
summary(lg3)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -5.7309     0.4705 -12.181  < 2e-16 ***
pirat          4.8496     0.7811   6.209 5.34e-10 ***  (1)
lvrat          1.9120     0.4815   3.971 7.16e-05 ***
phistyes       1.6702     0.1898   8.801  < 2e-16 ***  (2)
insuranceyes   4.3560     0.5450   7.993 1.32e-15 ***  (3)
afamyes        0.9622     0.1663   5.787 7.17e-09 ***  (4)
```

Controls: loan-to-value ratio `lvrat`, public bad-credit record `phist`, denied mortgage insurance `insurance`, and whether the applicant is African American `afam`.

(1) P/I still matters, holding the rest fixed.

(2) (3) Bad credit and denied insurance are **strong** predictors — insurance denial almost settles it.

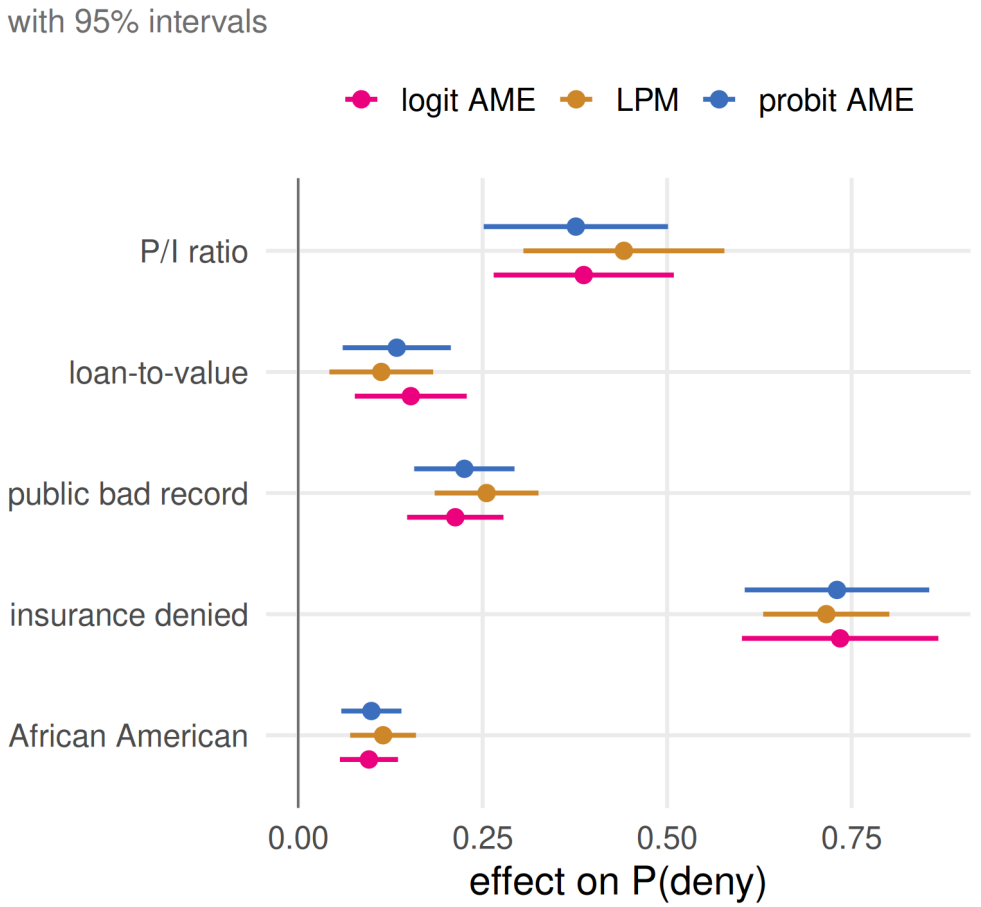
(4) The **afam** coefficient is positive and clearly significant: a race gap **after** these controls — the question the original study asked.

From Log-Odds to Percentage Points

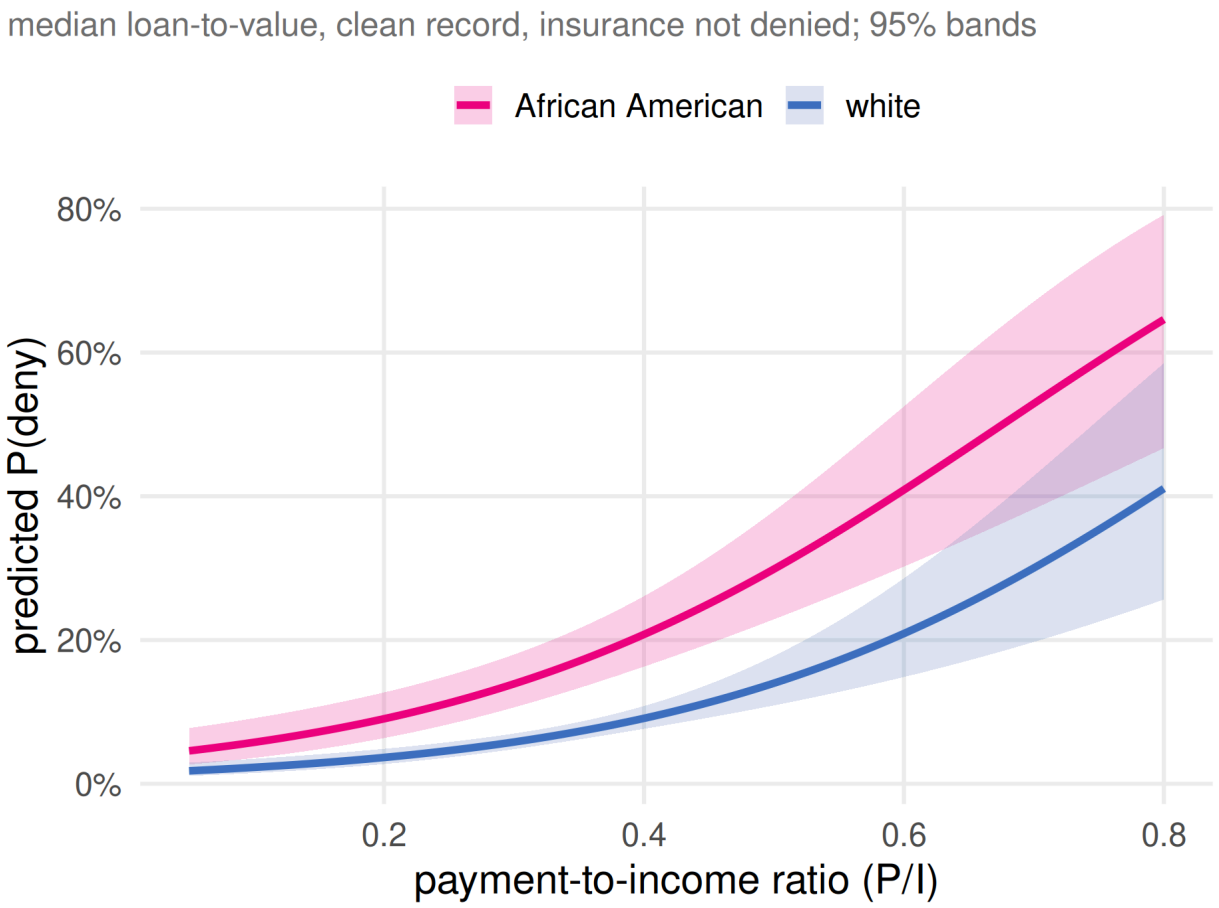
```
1 # helper in this deck's source; the same numbers
2 # as marginaleffects::avg_slopes(lg3)
3 ame(lg3)
```

```
               AME Std. Error z value Pr(>|z|)
pirat        0.3870     0.0623    6.21 5.28e-10
lvrat        0.1526     0.0387    3.95 7.89e-05
phistyes     0.2129     0.0333    6.40 1.59e-10
insuranceyes 0.7345     0.0679   10.82 2.75e-27
afamyes      0.0959     0.0201    4.78 1.71e-06  (1)
```

(1) At the same P/I, loan-to-value, credit history and insurance status, African American applicants were on average **9.6 percentage points** more likely to be denied (base rate 12%).



Predicted Probabilities Tell the Story



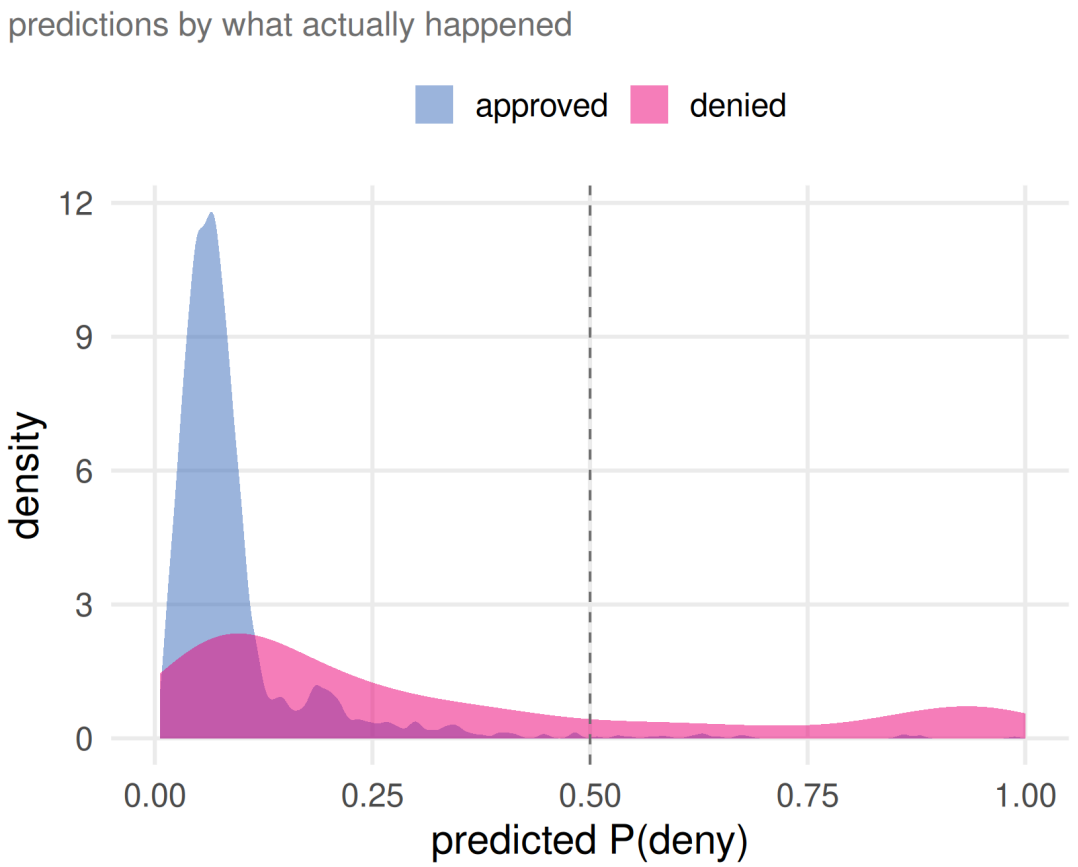
Fix the controls at a **typical applicant** and vary one thing at a time. `predict(..., type = "response")` on a grid of values; build bands on the **link** scale, then transform.

The gap **widens** as P/I rises: in logit and probit the effect of a dummy in **percentage points** depends on the other regressors, even without an interaction term.

Good practice

A plot of predicted probabilities for a few meaningful profiles usually communicates better than any table.

How Well Does It Fit?




```
1  # McFadden pseudo-R²
2  1 - lg3$deviance / lg3$null.deviance
```

```
[1] 0.2246673
```

```
1  table(predicted = fitted(lg3) > 0.5, actual = HMDA$
```

```
         actual
predicted   no  yes
    FALSE 2073  212
    TRUE    22   73
```

- **Pseudo-R^2** compares log-likelihoods; values of 0.2–0.4 are already good. Not comparable to OLS R^2
- **Accuracy** is 90% — but saying "approve" to everyone scores 88%. With rare outcomes, look at the **distributions**, not the hit rate

Binary Choice: Practical Advice

Choosing a model

- **LPM** with robust SEs: fine for average effects, easy with fixed effects and IV
- **Logit / probit**: when predictions must stay in $[0, 1]$ or you care about effects away from the average
- Report **AMEs** or **predicted probabilities**, not raw coefficients

Traps

- **Perfect separation**: a regressor predicts Y perfectly; R warns *fitted probabilities numerically 0 or 1* and the coefficient runs off to infinity
- Do not compare raw coefficients **across** models or samples — the scale is not fixed
- **Interactions** in logit/probit: the coefficient on the product is **not** the interaction effect on probabilities

3 · A Tour of Other Models

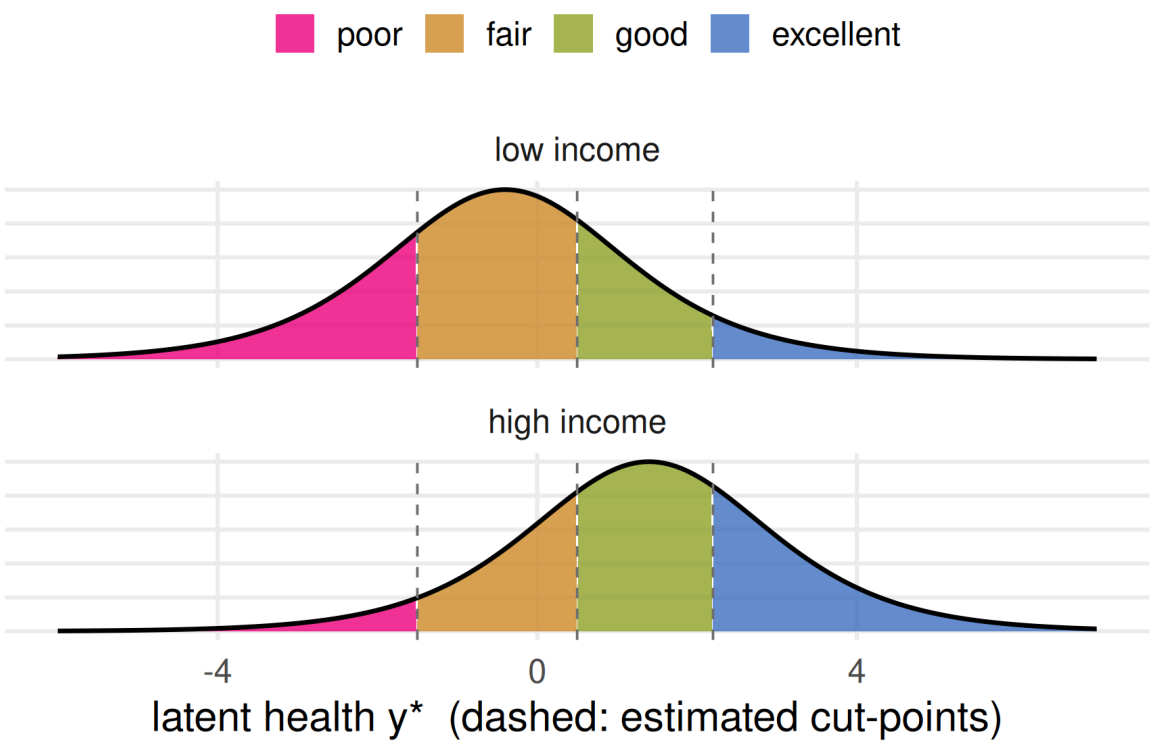
Match the Model to the Outcome

Outcome	Example	Model	R
Continuous	log wage	**linear regression**	`lm()`
Binary	mortgage denied	**logit / probit**	`glm(family = binomial)`
Ordered categories	health: poor … excellent	**ordered logit / probit**	`MASS::polr()`
Unordered categories	car, bus, bike	**multinomial logit**	`nnet::multinom()`
Count	doctor visits	**Poisson / negative binomial**	`glm(family = poisson)`, `MASS::glm.nb()`
Many zeros, then continuous	hours worked	**Tobit**	`AER::tobit()`
Observed only for a selected group	wages of workers	**Heckman selection**	`sampleSelection::selection()`
Time until an event	unemployment spell	**duration / survival**	`survival::coxph()`

And by data structure or design

The same unit observed many times → **panel / fixed effects**. A policy that hits some units at some date → **difference-in-differences**. An endogenous regressor with an outside source of variation → **instrumental variables**. Effects that differ across the distribution → **quantile regression**.

Ordered and Unordered Choices



Ordered: one latent scale, several cut-points

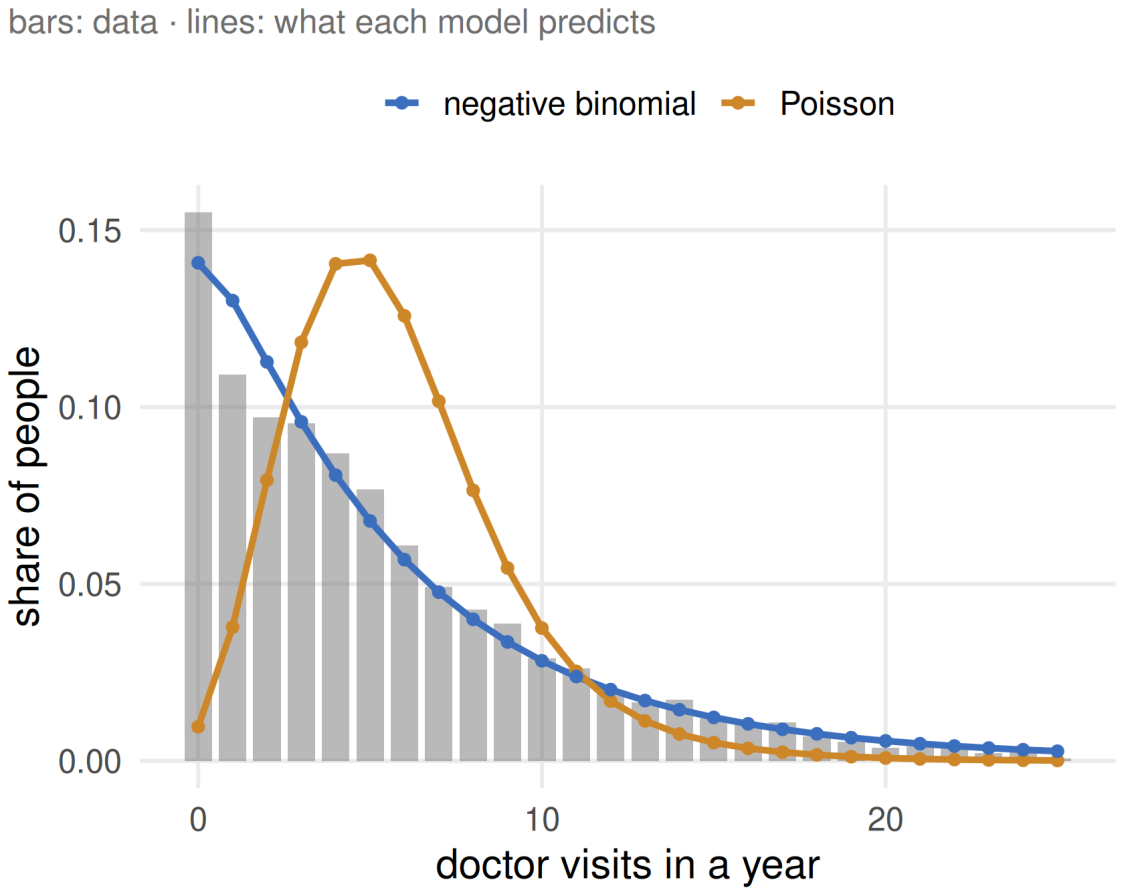
The binary latent model with **more thresholds**. X shifts the distribution; the areas between cut-points are the probabilities.

`MASS::polr()`

Unordered: one equation per alternative

Car, bus or bike: no natural order. **Multinomial logit** models each option relative to a base option. `nnet::multinom()`

Count Data: Poisson and Negative Binomial



Counts are $0, 1, 2, \ldots$ and skewed. Model the **log of the mean**: $E[Y \mid X] = e^{X\beta}$, so $e^{\beta} - 1$ is a **% change**.

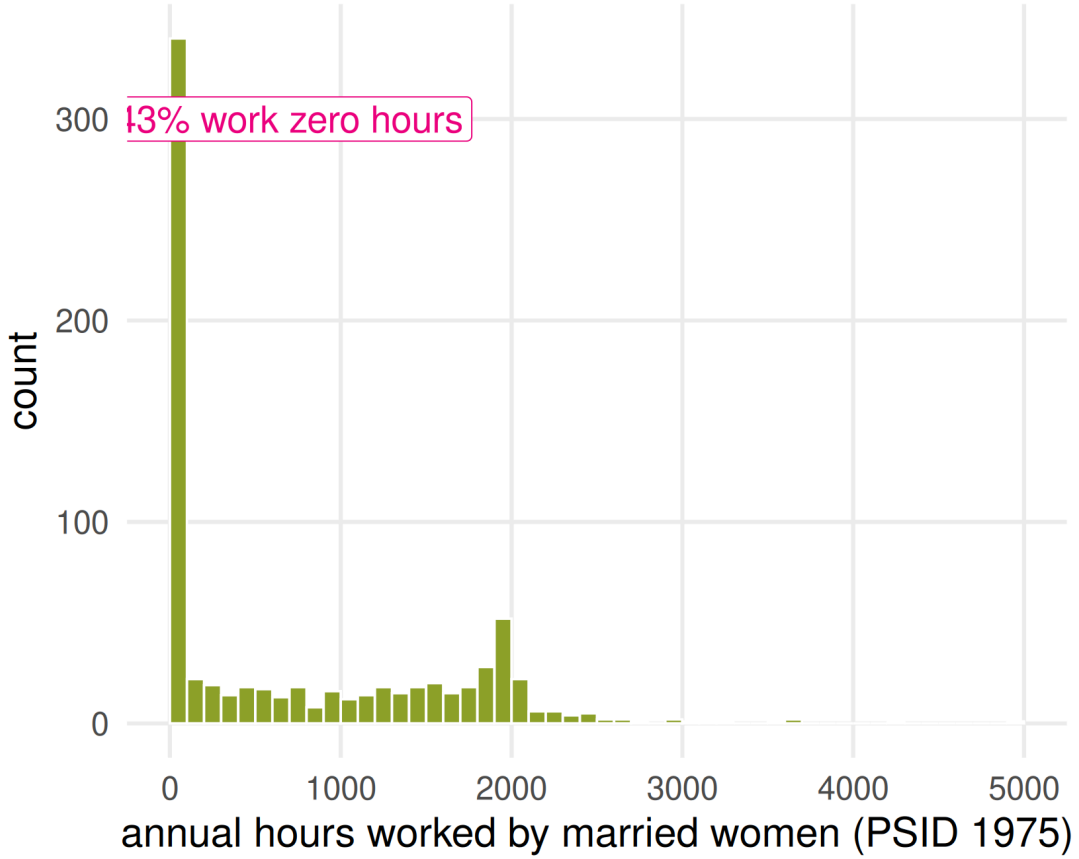
Poisson forces variance = mean. Here the mean is 5.8 and the variance 45.7: **overdispersion**. The negative binomial adds a parameter for it — and matches the zeros.

```
nb <- glm.nb(visits ~ health + chronic + age +
               gender + insurance, data = NMES198
round(exp(coef(nb)[c("chronic", "insuranceyes")]))
```

```
  chronic insuranceyes
    1.216        1.388
```

Each chronic condition: **+22%** visits; insurance: **+39%**.

Corner Solutions: Tobit



A pile-up at zero: many women **choose** not to work. A straight line through this cloud is pulled toward the zeros.

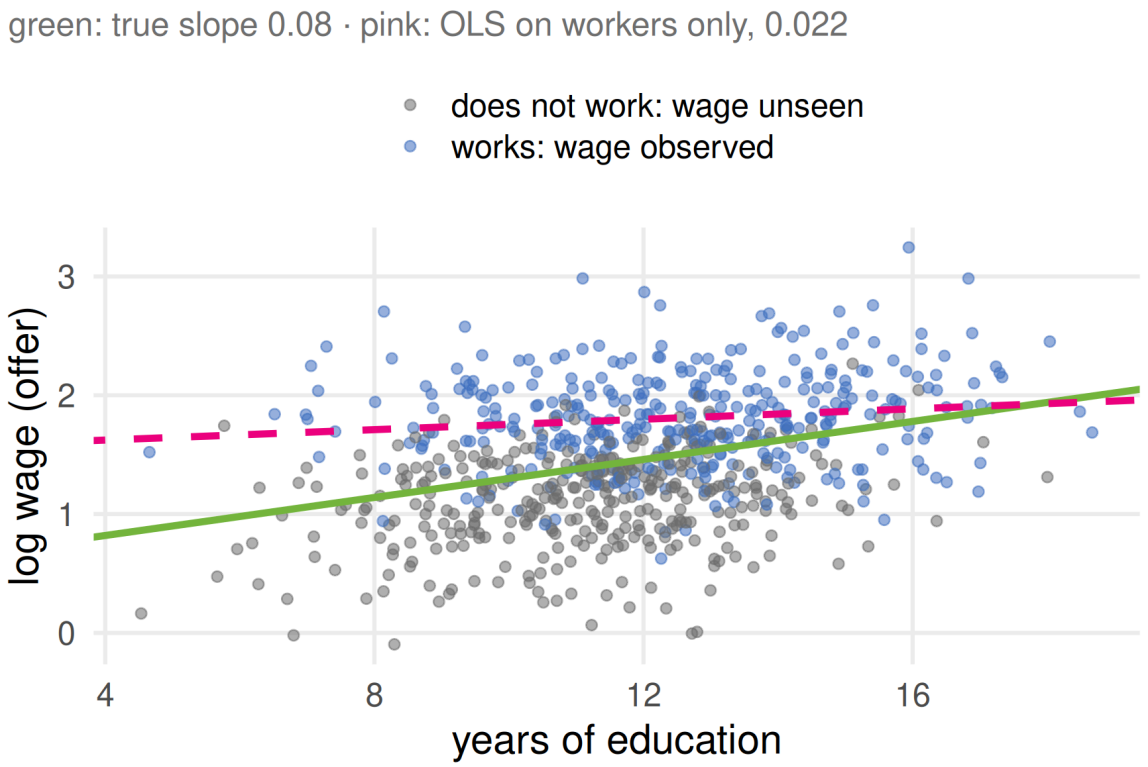
Tobit: latent desired hours $y^* = X\beta + u$; we see $y = \max(0, y^*)$. Estimated by maximum likelihood, like probit for the zeros and OLS for the rest.

```
PSID1976$nwinc <- with(PSID1976,
                       (fincome - hours * wage) / 1(
f  <- hours ~ nwinc + education + experience +
              age + youngkids
tb <- tobit(f, data = PSID1976)
round(cbind(OLS   = coef(lm(f, data = PSID1976)),
            Tobit = coef(tb)), 1)
```

```
                OLS  Tobit
(Intercept) 1306.9 1287.8
nwinc         -3.5   -8.5
education     32.7   86.1
experience    46.8   77.4
age          -30.3  -58.3
youngkids   -444.6 -919.9
```

Tobit coefficients refer to **desired** hours y^* and are larger than OLS. For effects on actual hours, compute marginal effects — as with probit.

Sample Selection



We see wages only for people who **work**. The low-educated work mainly if their offer is **unusually high**, so the slope comes out too **flat**.

Heckman's two steps

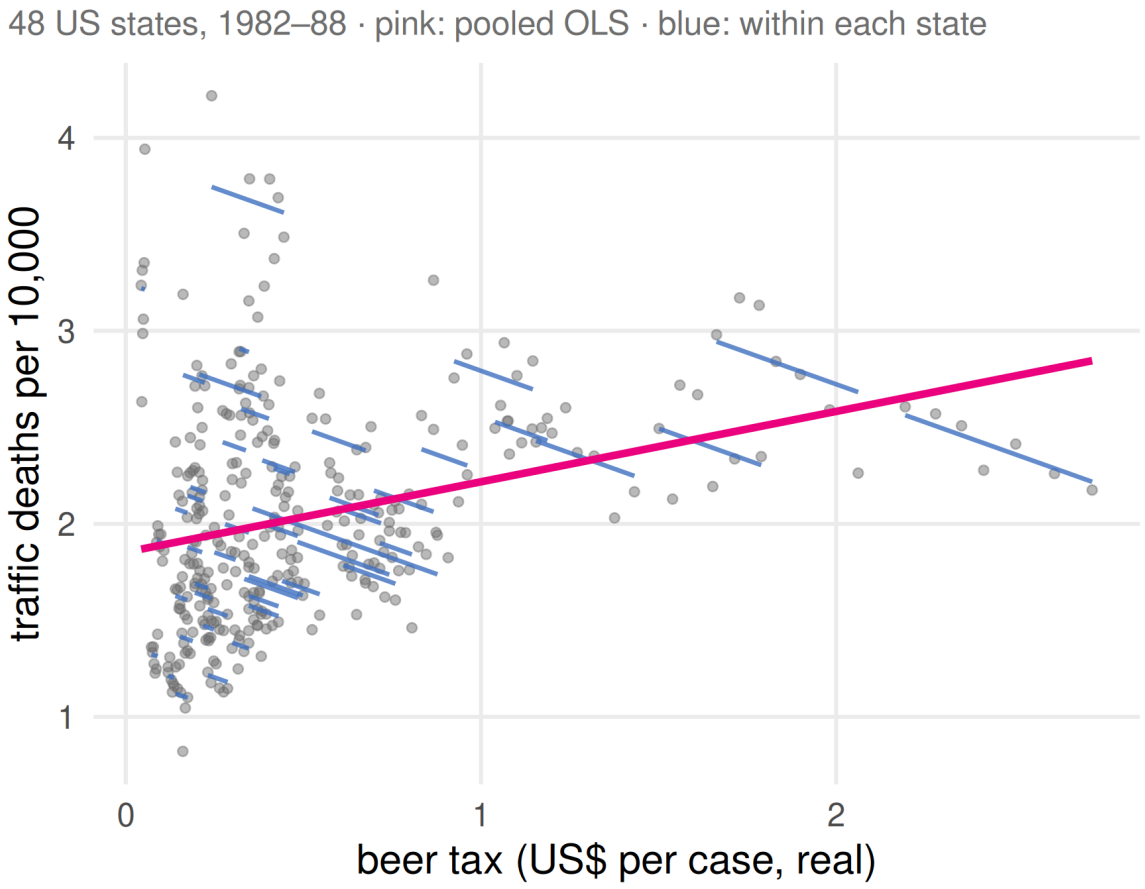
(1) a **probit** for who works

(2) the wage regression plus a correction term from step 1

Needs a variable that shifts **working but not wages** (e.g. young children).

Simulated data. In R: `sampleSelection::selection()`.

Panel Data: Fixed Effects



Across states, **higher beer taxes go with more deaths** — because rural, southern states have both. That is omitted variable bias again, from things that differ **between states** but barely change over time.

Fixed effects

Give every state its own intercept. The slope is then estimated only from changes **within** a state over time — everything constant about a state drops out, **observed or not**.

The within-state slopes point **down**.

Fixed Effects in R

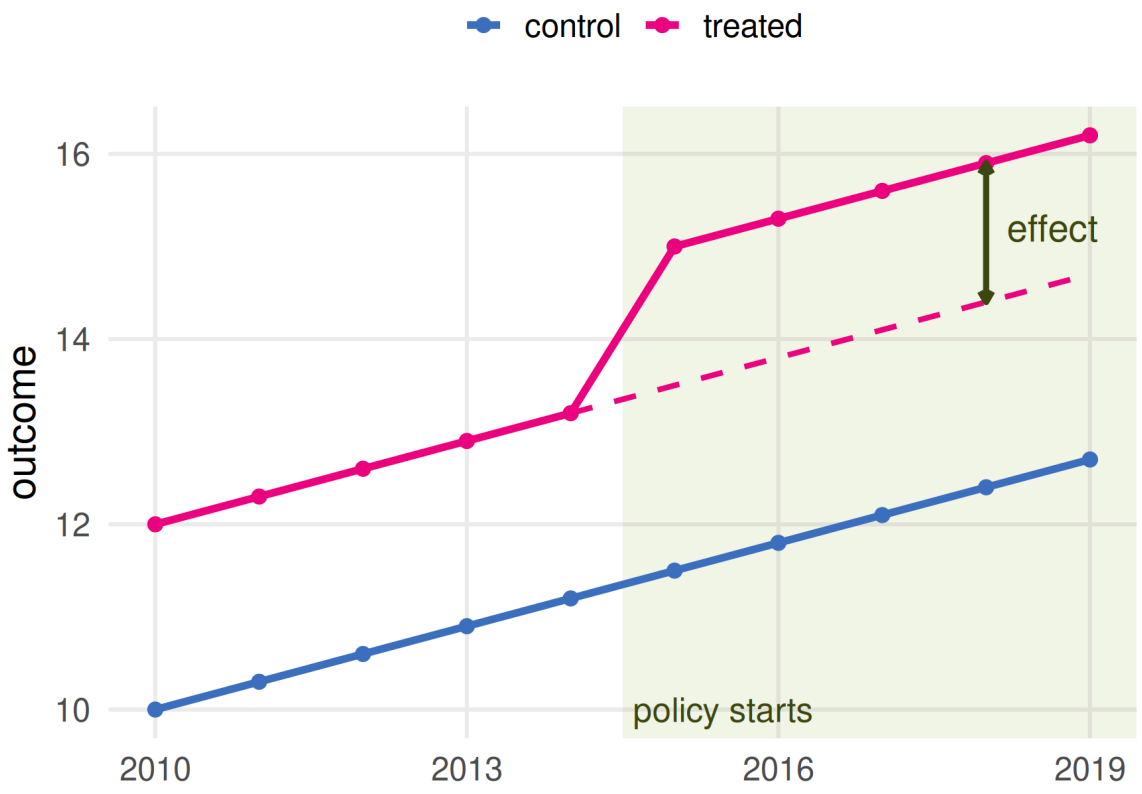
```r
library(fixest)
pooled <- feols(frate ~ beertax, Fatalities, cluster = ~state)
fe_s   <- feols(frate ~ beertax | state, Fatalities, cluster = ~state)
fe_st  <- feols(frate ~ beertax | state + year, Fatalities,
                cluster = ~state)
etable(pooled, fe_s, fe_st)
```

```
                          pooled              fe_s              fe_st
Dependent Var.:             frate             frate             frate

Constant         1.853*** (0.1185)
beertax          0.3646** (0.1197) -0.6559* (0.2919) -0.6400. (0.3571)
Fixed-Effects:   ----------------- ----------------- -----------------
state                          No               Yes               Yes
year                           No                No               Yes
_________________ _________________ _________________ _________________
S.E.: Clustered          by: state         by: state         by: state
Observations                   336               336               336
R2                         0.09336           0.90501           0.90893
Within R2                       --           0.04075           0.03606
```

1. Pooled OLS: taxes **raise** deaths?
2. State fixed effects: the sign **flips**. A $1 tax increase goes with 0.66 fewer deaths per 10,000.
3. Year effects also remove nationwide shocks (recessions, safety laws) — the **two-way** FE model.
4. **Clustered** SEs: the 7 years of one state are not independent draws.

Difference-in-Differences

dashed: the treated group without the policy



A policy hits **some** units from **some date** on. Compare the change in the treated group with the change in the control group:

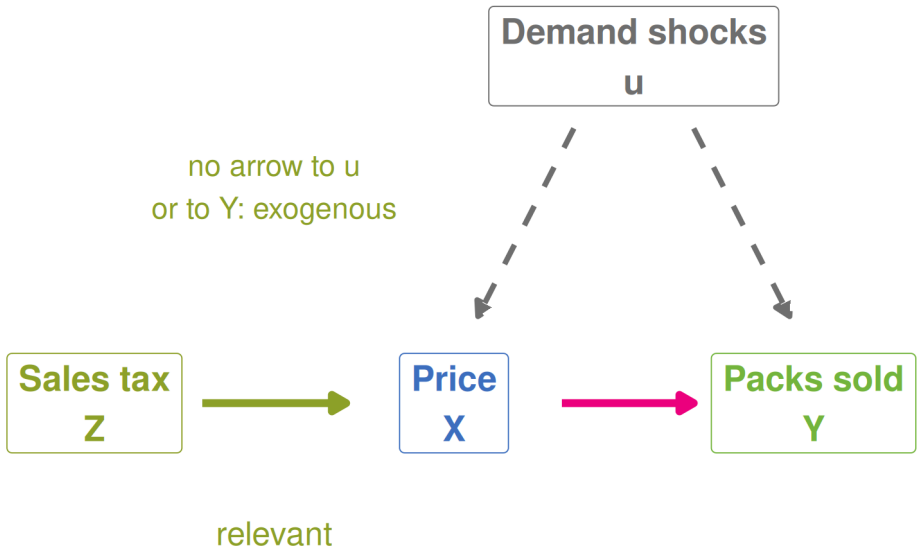
$$\hat{\delta} = (\bar{y}^{T}_{\text{after}} - \bar{y}^{T}_{\text{before}}) - (\bar{y}^{C}_{\text{after}} - \bar{y}^{C}_{\text{before}})$$

Parallel trends

Without the policy, both groups would have moved **in parallel**. Check the pre-period; that is where the credibility comes from.

In R: `feols(y ~ treated:post | unit + year, data)` — a two-way fixed-effects regression.

Instrumental Variables



Prices and quantities are set **together** by supply and demand, so price is endogenous in a demand equation.

An **instrument** Z moves X but has no other route to Y. Use **only the part of** X **that** Z **moves**:

(1) regress X on Z (**first stage**) and keep the fitted $\hat{X}$

(2) regress Y on $\hat{X}$ (**second stage**)

This is **two-stage least squares** (2SLS).

> **Two conditions**
>
> **Relevance** — Z really moves X: testable.
>
> **Exogeneity** — Z affects Y only through X: an argument, not a test.

IV in R: Cigarette Demand

```r
c95 <- subset(CigarettesSW, year == "1995")
c95 <- transform(c95, rprice = price / cpi, salestax = (taxs - tax) / cp
                      rincome = income / population / cpi)
iv <- ivreg(log(packs) ~ log(rprice) + log(rincome) | log(rincome) + sal
            data = c95)
summary(iv, diagnostics = TRUE)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.4307     1.3584   6.943 1.24e-08 ***
log(rprice)   -1.1434     0.3595  -3.181  0.00266 **   (1)
log(rincome)   0.2145     0.2686   0.799  0.42867

Diagnostic tests:
                 df1 df2 statistic  p-value
Weak instruments   1  45    45.158 2.65e-08 *** (2)
Wu-Hausman         1  44     1.102       0.3     (3)

Residual standard error: 0.1896 on 45 degrees of freedom
Multiple R-Squared: 0.4189, Adjusted R-squared: 0.3931
Wald test: 6.534 on 2 and 45 DF,  p-value: 0.003227
```

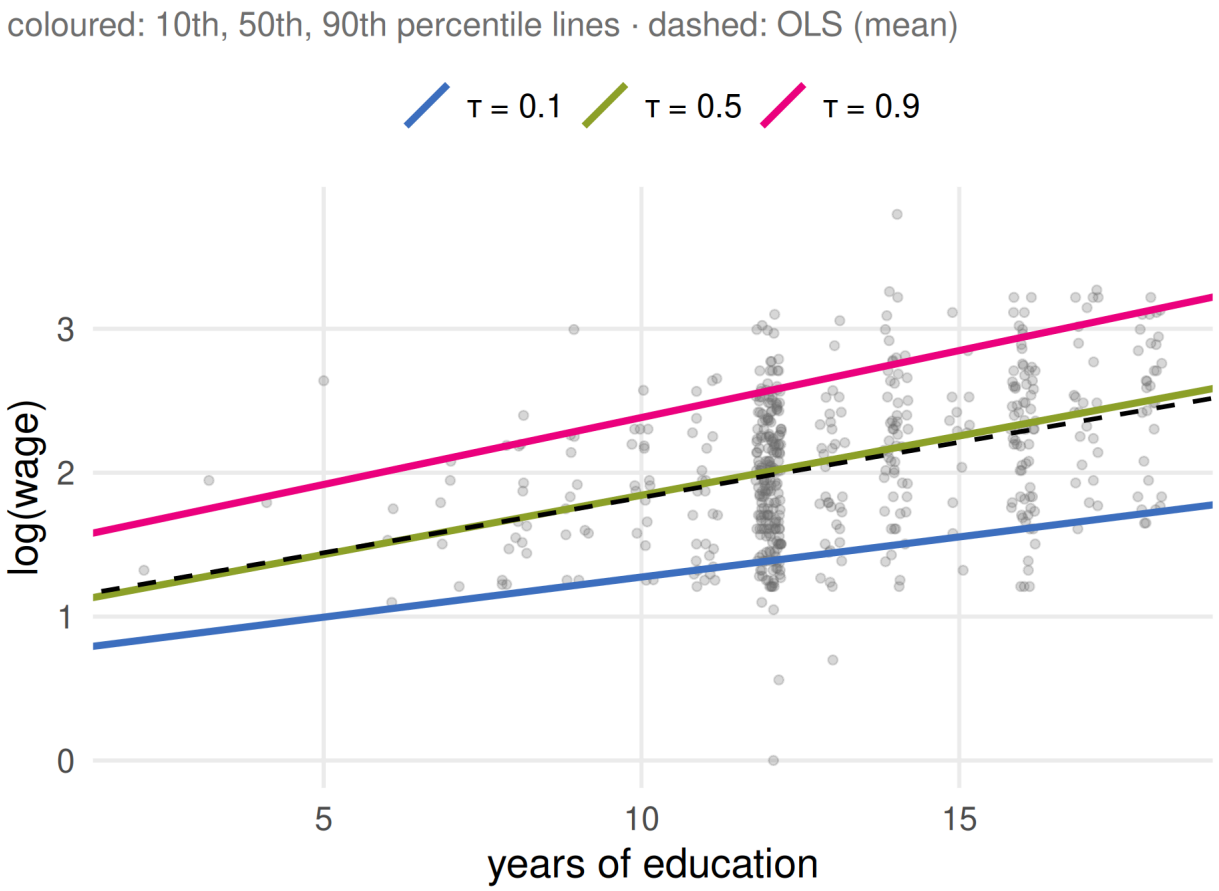
After the `|`: the **instruments** — exogenous regressors (income) plus the excluded instrument (sales tax).

1. Price **elasticity** of demand: a 1% price rise cuts packs sold by 1.14%. OLS gives -1.41.

2. **Weak-instrument** test = first-stage F. Far above the rule-of-thumb 10: the tax is **relevant**.

3. **Wu–Hausman**: OLS and IV do not differ significantly here. With one instrument, **exogeneity** cannot be tested.

Quantile Regression



OLS models the **mean**. Quantile regression models a **percentile** of Y given X — e.g. the wage at the bottom 10% of workers with 12 years of school.

```
qr <- rq(log(wage) ~ education,
         tau = c(0.1, 0.5, 0.9), data = CPS
round(coef(qr), 3)
```

```
             tau= 0.1 tau= 0.5 tau= 0.9
(Intercept)     0.717    1.019    1.453
education       0.056    0.082    0.093
```

Education pays **more at the top**: 5.6% per year at the 10th percentile, 9.3% at the 90th. It also **spreads** wages out. The asymmetric loss from the estimation deck, put to work.

Also Worth Knowing

Time series

ARIMA for one series, **VAR** for several that move together. Beware trends: two unrelated trending series look correlated.

Regression discontinuity

Treatment switches on at a **cut-off** in a running variable (a test score, an age). Compare units just below and just above. `rdrobust`.

Duration models

Time until an event — finding a job, a firm exiting. Handle spells still running at the end of the data (**censoring**). `survival::coxph()`.

Machine learning meets econometrics

Lasso and **random forests** for prediction and for choosing controls; **double machine learning** for causal effects with many controls.

The common thread

Each is a **conditional mean** with a new shape (fitted by least squares or ML), or a **design** that makes $E[u \mid X] = 0$ believable.

Summary

Key Takeaways

(1) A regression models the **conditional mean** $E[Y \mid X]$. Read every coefficient **in units**

(2) `summary()` output: estimate, standard error, t / z, p-value, fit. Use **robust** standard errors by default

(3) "Holding other things fixed" only holds for what is **in** the model — omitted variables bias the rest

(4) For 0/1 outcomes: LPM, logit or probit. Report **marginal effects** or **predicted probabilities**, not raw logit coefficients

(5) Let the **outcome** choose the model — and the **design** decide whether the answer is causal

The one thing to remember

Every model in this deck answers: **how does the distribution of Y change with X?**

The models differ in the **shape** they allow. Whether the answer is **causal** depends on $E[u \mid X] = 0$ — and no choice of model can promise that.