Censored regression, selection model, weak IV, and quantile regression

Tutorial 1

Stanislav Avdeev

Tutorials

- 7 TA sessions
  - 6 TA sessions are about lecture material
  - the last session is primarily about exam and remaining questions about the course material (TBA)
- Send me **any** questions you want to discuss before each TA session
  - use Canvas or send me an email (s.avdeev@tinbergen.nl)
  - alternately, leave your questions anonymously here: https://onlinequestions.org (enter the event code 18631)

Assignments

- Due date: 11:59pm on Sundays (the first assignment is an exception: 11:59am on Tuesday)
- Assignments are graded within a week from the deadline
- Solutions will not be shared so if you want to discuss a specific exercise, let me know before the TA session (you submit your solutions on Sunday, thus, we can discuss any questions on the following TA session on Tuesday)

Course objective

- The key objective of the course is **applying** microeconometric techniques rather than **deriving** econometric and statistical properties of estimators
- In other words, there's way less of this

$$\text{plim}\hat{\beta}_{OLS} = \beta + \text{plim}(\frac{1}{N}X'X)^{-1}\text{plim}\frac{1}{N}X'\varepsilon = \beta + Q^{-1}\text{plim}\frac{1}{N}X'\varepsilon$$

- And way more of this

```
library(fixest)

df <- tibble(groups = sort(rep(1:10, 600)),
             time = rep(sort(rep(1:6, 100)), 10),
             Treated = I(groups > 5) * I(time > 3),
             Y = groups + time + 5*Treated + rnorm(6000))
did <- feols(Y ~ Treated | groups + time, data = df)
```

- If you would like to go deeper into the former, take Advanced Econometrics I and II next year

Goal for today's tutorial

1. Use a tobit model to estimate censored data
2. Discuss a sample selection model and implement the selection mechanism
3. Work with strong and weak instrumental variables
4. Test instrumental variables
5. Work with a quantile regression and discuss inference tools

Censored regression

- Censoring occurs when the value of a variable is limited due to some constraint
    - for example, we tend not to see some values of self-declared earnings with discrete categories (if you earn **at least** 3500 euro per month, write 3500)
- In this case OLS estimates are biased
    - a standard method to account for censoring is to combine a probit model with OLS, i.e. **tobit model**

Censored regression: simulation

- The clearest way to understand how a certain estimator works is to generate data yourself so you know the true **data generating process** - DGP
- Let's estimate returns to education: does education increase wages?
- Let's assume the following model

$$Y_i = \alpha + X_i'\beta_1 + U_i$$

  where Y_i are monthly wages, X_i are years of education
- But suppose that we do not observe wages above a specific threshold (due to the features of a questionnaire, privacy concerns, coding, etc.)
  - how can we estimate the model in this case?
- First, we need to generate data containing years of education and wages

Censored regression: simulation

```
# Always set seed so you can replicate your results
set.seed(7)
df <- tibble(education = runif(1000, 5, 15),
             wage_star = 1000 + 200*education + rnorm(1000, 0, 100),
             wage = ifelse(wage_star > 3500, 3500, wage_star))  %>%
  arrange(wage_star)
```

education	wage_star	wage
5.062510	1837.496	1837.496
5.036464	1878.439	1878.439
5.260247	1889.825	1889.825
5.266903	1913.437	1913.437
14.988792	4081.234	3500.000
14.931072	4167.111	3500.000
14.849343	4174.288	3500.000
14.900142	4201.639	3500.000

Censored regression: OLS

- Now let's pretend that we do not know the DGP and simply apply OLS

```
ols_model <- lm(wage ~ education, df)
```

	Model 1
(Intercept)	1255.542***
	(14.018)
education	167.686***
	(1.340)
Num.Obs.	1000
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001	

- Using these OLS estimates, we would wrongly conclude that "an additional year of education is associated with 167.686 increase in monthly wages"
  - if we think that we **causally** identified the effect we'd say "an additional year of education **causes** 167.686 increase in monthly wages"

Censored regression: tobit model

- But these are biased estimates since we know the true effect is 200 (remember DGP)
- Let's try to recover unbiased effects of education on wages by using a **tobit model**
- The solution provided by a tobit model is to
  - use a probit model to account for the censoring
  - estimate OLS on the non-censored data
- Tobit model estimator is easy to implement with `censReg` package

Censored regression: tobit model

- Remember that we have right censored data: wages above 3500 are coded as 3500

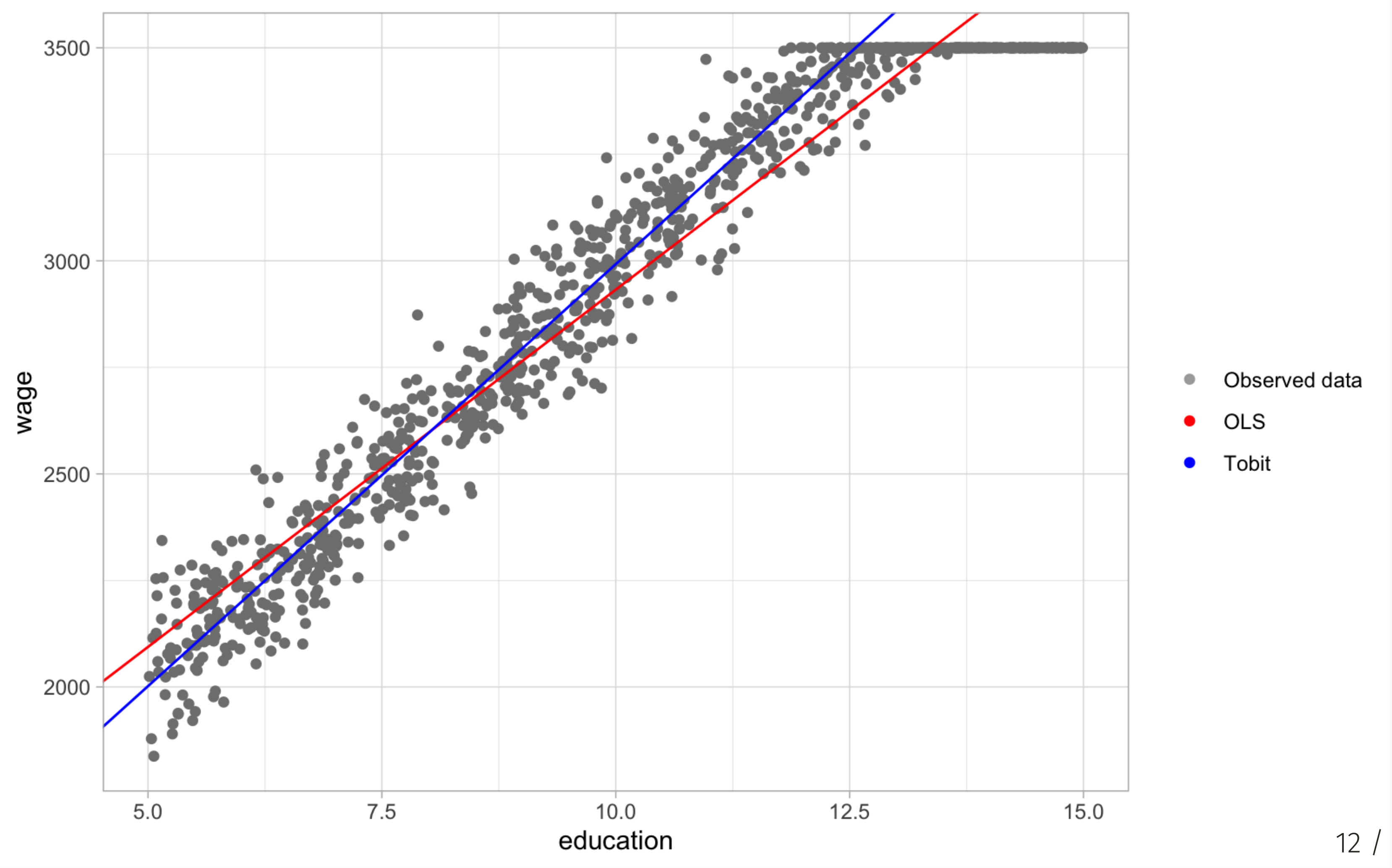
```r
library(censReg)
tobit_model <- censReg(wage ~ education, data = df, right = 3500)
```

	Model 1
(Intercept)	1011.629***
	(13.998)
education	198.053***
	(1.495)
logSigma	4.586***
	(0.026)
Num.Obs.	1000
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001	

- We recovered the **unbiased** estimates of returns to education

Censored regression: graphically

- We will use a lot of graphs since they provide more intuition of what is happening



Censored regression: some remarks

- You can specify both left and right censoring using `censReg` function
- Important assumption of a tobit model is that the unobserved term is normally distributed (which is the case in our simulated data set)
- What if the data is missing not because the outcome variable is **above (or below)** some threshold but because individuals in the data have made a **choice** such that we can't observe their outcome variable?
- In this case censoring cannot be applied because the availability of data is **influenced** by the choice of agents
  - it is called **selection on unobservables**
  - it is a typical **sample selection problem**

Sample selection model

- Let us consider the case of studying female's wages
  - usually, wages are observed for a fraction of women in the sample, whereas the remaining part of women are observed as unemployed or inactive
  - if we run an OLS regression using the observed wages, this would deliver consistent estimations only if working females are a **random sample** of the population
- However, theory of labor supply suggests that this may not be the case, since (typically) female labor supply is sensitive to household decisions
  - that is, female workers **self-select** into employment, and the self-selection is not random
  - this difference may lead us to underestimate the gender wage gap

Sample selection model

- Suppose a female worker decides to work or not based on a latent variable I_i^* (say, utility derived from working), which depends on a set of observed Z_i and unobserved V_i characteristics

$$I_i^* = Z_i'\gamma + V_i$$

- The indicator function (decision to work or not), based on I_i^*, takes two values

$$I_i = \begin{cases} 1 \text{ (working)} & \text{if } I_i^* > 0 \\ 0 \text{ (not working)} & \text{if } I_i^* \leq 0 \end{cases}$$

- Suppose there is a latent outcome Y_i^*, i.e. wages of female workers, which depend on a set of observed X_i and unobserved U_i characteristics

$$Y_i^* = X_i'\beta + U_i$$

- However, we observe wages only for females who decide to work: Y_i are observed wages

$$Y_i = \begin{cases} Y_i^* & \text{if } I_i = 1 \\ \text{missing} & \text{if } I_i = 0 \end{cases}$$

Sample selection model: assumptions

- As always we need to have some assumptions, for example, in an OLS regression we usually assume $U_i \sim \mathcal{N}(0, \sigma^2)$)
- To estimate the sample selection model, we make an assumption that disturbances terms are bivariate normal

$$\begin{bmatrix} U_i \\ V_i \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{bmatrix}\right)$$

- Note that the variance of the normal distribution is not identified in the probit model so it is set to 1

Sample selection model: simulation

- Let's simulate a data set with the selection mechanism

```r
library(mvtnorm) # to simulate bivariate normal random variable
set.seed(7)
df <- tibble(z = runif(1000),
             x = runif(1000),
             uv = rmvnorm(1000, mean = c(0, 0),
                  sigma = rbind(c(2, 0.7),
                                c(0.7, 1))),
             i_star = 4 - 5 * z + uv[, 1],
             y_star = 6 - 3 * x + uv[, 2],
             y = ifelse(i_star > 0, y_star, 0)) # this is a selection mechanism
```

```
## # A tibble: 6 × 6
##         z     x uv[,1]   [,2] i_star y_star     y
##     <dbl> <dbl>  <dbl>  <dbl>  <dbl>  <dbl> <dbl>
## 1 0.989  0.418  1.49  -0.357  0.549   4.39  4.39
## 2 0.398  0.813  0.190 -0.283  2.20    3.28  3.28
## 3 0.116  0.278 -0.960  0.310  2.46    5.47  5.47
## 4 0.0697 0.968 -2.19  -1.84   1.46    1.25  1.25
## 5 0.244  0.247 -0.613  0.670  2.17    5.93  5.93
## 6 0.792  0.905 -2.65  -0.582 -2.61    2.70  0
```

Sample selection model: simulation

- The true effect of Z on I (decision to work) is -5 and the effect X on Y (wages) is -3

```
selection_equation <- glm(I(y > 0) ~ z, df, family = binomial(link = "probit"))
wage_equation <- lm(y ~ x, df)
```

	Model 1	**Model 2**
(Intercept)	3.013***	4.575***
	(0.178)	(0.137)
z	-3.770***	
	(0.251)	
x		-2.111***
		(0.239)
Num.Obs.	1000	1000

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

- Clearly, the estimates are biased since $\mathbf{cov}(U_i, V_i) \neq \mathbf{0}$

Sample selection model: Heckman

- To solve the sample selection problem, one needs to use the **Heckman selection model** (left as an exercise in the 1^{st} assignment)
  - the Heckman estimator is very similar to the Tobit estimator
  - the difference is that this estimator allows for a set of characteristics that **determine** whether or not the outcome variable is censored

IV

- The basic idea of the IV estimator is to
  - use the instrument to predict treatment
  - use that predicted treatment to predict the outcome
- We need a separate equation for each of these steps
  - **First stage**: predict treatment X_1 with the instrument Z and a control variable X_2

$$X_1 = \gamma_0 + \gamma_1 Z + \gamma_2 X_2 + V$$

  - **Second stage**: use these predictions to predict Y

$$Y = \beta_0 + \beta_1 \hat{X}_1 + \beta_2 X_2 + U$$

- Notice that we need to use all exogenous variables in both stages

IV: conditions

- For the IV to work, we need two things to hold
  - **Validity**: the instrument must actually be exogenous (or at least exogenous after adding controls)

$$\text{cov}(Z, U) = 0$$

  - **Relevance**: the instrument must be a strong predictor of the treatment. It can't be trivial or unimportant

$$\text{cov}(X_1, Z) \neq 0$$

IV: small-sample bias

- IV is actually a **biased** estimator
  - the mean of its sampling distribution is not the population parameter
  - it would be the population parameter at infinite sample size, but we don't have that
  - in small samples, the **bias of IV** is

$$\frac{\mathrm{cov}(Z,U)}{\mathrm{cov}(X_1,Z)} = \frac{\mathrm{corr}(Z,U)}{\mathrm{corr}(X_1,Z)} \frac{\sigma_U}{\sigma_{X_1}}$$

- If Z is valid, then in infinite samples $\mathrm{cov}(Z,U) = 0$ and this goes away
- But in a non-infinite sample, it will be nonzero by chance, inducing some bias
- The bias is smaller
  - the stronger the relationship between X_1 and Z
  - the smaller the sum of squared errors
  - the bigger the variation in X_2
  - the bigger the sample
- What happens when $\mathrm{corr}(X_1,Z)$ is small?

Weak IV

- If Z has only a trivial effect on X, then it's not **relevant** (even if it's truly **exogenous**)
  - our **small-sample bias** will be big (remember the formula on the previous slide)
- Thus, **weak IV** means that we probably shouldn't be using IV in small samples
  - this also means that it's really important that $\text{corr}(X_1, Z)$ is not small
- There are rules of thumb for how strong IV must be to be counted as "not weak"
  - t-statistic above 3
  - F-statistic from a joint test of the instruments that is 10 or above
- These rules of thumb aren't great
  - selecting a model on the basis of significance naturally biases your results
  - what you really want is to know the **population effect** of Z on X_1 - you want the F-statistic from that to be bigger than 10. Of course, we don't actually know that

Weak IV: estimation

- There are a bunch of ways to do the IV analysis
  - the classic one is `ivreg` in the `AER` package

- Other functions are more fully-featured, including robust SEs, clustering, and fixed effects

  - `feols` in `fixest`
  - `felm` in `lfe`
  - `tsls` in `sem`
  - `ivpack`

- We'll be using `feols` from `fixest`

Weak IV: simulation

- Let's create a data set with an instrument

```
library(fixest)
set.seed(7)
df <- tibble(z1 = rnorm(1000),
             u1 = rnorm(1000),
             # x1 is endogenous since it correlates with u1 by construction
             x1 = 0.2*z1 + 4*u1 + rnorm(1000),
             y1 = 3*x1 + 5*u1)
```

z1	u1	x1	y1
2.2872472	1.2465863	6.8585432	26.8085610
-1.1967717	-0.7655089	-5.4047210	-20.0417075
-0.6942925	0.2161769	-0.3578382	0.0073698
-0.4122930	-0.3643673	-1.9102661	-7.5526346

Weak IV: simulation

```
# The true effect is 3
library(fixest)
ols_model <- lm(y1 ~ x1, df)
iv_model <- feols(y1 ~ 1 | x1 ~ z1, df, se = 'hetero')
```

	Model 1	Model 2
x1	4.171***	
	(0.009)	
fit_x1		3.479***
		(0.348)
Num.Obs.	1000	1000
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001		

- Z_1 is a pretty effective instrument even if the correlation between Z_1 and X_1 is small
  - check validity: $\mathbf{corr}(Z_1, U_1) = 0.027$, pretty close to zero
  - check relevance: $\mathbf{corr}(X_1, Z_1) = 0.068$, so it's a weak instrument

Weak IV: simulation

- Remember that usually we can't test the **validity** assumption when we have one instrument, but we know the DGP in this case
- Now let's see what happens when there is a small correlation between Z and U
- Imagine there is some additional explanatory variable V which is unobserved but partially explains the instrument

```
set.seed(7)
df <- tibble(v = rnorm(1000),
             z2 = -v + rnorm(1000),
             u2 = 0.1*v + rnorm(1000),
             x2 = 0.2*z2 + 4*u2 + rnorm(1000), # all coefficients stay the same here
             y2 = 3*x2 + 5*u2) # and here
```

v	z2	u2	x2	y2
2.2872472	-1.0406609	1.6434732	7.567896	30.92105
-1.1967717	0.4312628	-2.2230083	-9.868396	-40.72023
-0.6942925	0.9104694	-1.1531165	-5.119197	-21.12317
-0.4122930	0.0479257	-0.4115676	-3.494073	-12.54006

Weak IV: simulation

```
# The true effect is 3
ols_model2 ← lm(y2 ~ x2, df)
iv_model2 ← feols(y2 ~ 1 | x2 ~ z2, data = df, se = 'hetero')
```

	Model 1	Model 2
x2	4.163***	
	(0.009)	
fit_x2		0.944
		(4.365)
Num.Obs.	1000	1000
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001		

- In this case we get a better estimate by using the OLS estimator than by using IV
- Why? Because of the weak instrument problem and the bias
    - check validity: $\textbf{corr}(Z_2, U_2) =$ -0.036
    - check relevance: $\textbf{corr}(X_2, Z_2) =$ 0.021

Weak IV: simulation

- These results are primarily a function of the weakness of Z at explaining X. Let's see what happens if Z has more explanatory power

```
set.seed(7)
df <- tibble(v = rnorm(1000),
             z2 = -v + rnorm(1000),
             u2 = 0.1*v + rnorm(1000),
             # we only change the coefficient for z2 in the equation for x3
             x3 = 3*z2 + 4*u2 + rnorm(1000),
             y2 = 3*x3 + 5*u2)
```

v	z2	u2	x3	y2
2.2872472	-1.0406609	1.6434732	4.654045	22.17950
-1.1967717	0.4312628	-2.2230083	-8.660860	-37.09762
-0.6942925	0.9104694	-1.1531165	-2.569883	-13.47523
-0.4122930	0.0479257	-0.4115676	-3.359881	-12.13748

Weak IV: simulation

```r
# The true effect is 3
ols_model3 ← lm(y2 ~ x3, df)
iv_model3 ← feols(y2 ~ 1 | x3 ~ z2, data = df, se = 'hetero')
```

	Model 1	**Model 2**
x3	3.578***	
	(0.020)	
fit_x3		2.956***
		(0.036)
Num.Obs.	1000	1000
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001		

- Even though the correlation between Z and U is the same as previously, the strength of the instrument in explaining X wins out and gives us a better estimator than OLS
  - check validity: $\mathbf{corr}(Z_2, U_2) =$ -0.036
  - check relevance: $\mathbf{corr}(X_3, Z_2) =$ 0.697

Weak IV: F-test

- Let's look at the F-test from the output of `feols()`

```
## TSLS estimation, Dep. Var.: y2, Endo.: x3, Instr.: z2
## Second stage: Dep. Var.: y2
## Observations: 1,000
## Standard-errors: Heteroskedasticity-robust
##              Estimate Std. Error   t value  Pr(>|t|)
## (Intercept) -0.006571   0.162236 -0.040501    0.9677
## fit_x3       2.955645   0.036370 81.266072 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 5.12118   Adj. R2: 0.939727
## F-test (1st stage), x3: stat =   941.5, p < 2.2e-16, on 1 and 998 DoF.
##             Wu-Hausman: stat = 8,467.5, p < 2.2e-16, on 1 and 997 DoF.
```

- 941.48 is way above 10
- Lee, D., et al. (2021) discuss the potentially severe large-sample distortions from using conventional value of the F-test equal to 10 and they suggest to use as a rule of thumb the minimum value of the F-test equal to 104.7, which is needed to ensure a test with a significance level 0.05

Weak IV: overidentification test

- **Overidentification** just means we have more identifying conditions (**validity assumptions**) than we actually need
  - we only need one instrument, but we have two (or more)
  - so we can compare what we get using each instrument individually
- If we assume that **at least one of them is valid**, and they both produce similar results, then that's evidence that **both** are valid

Weak IV: overidentification test

- We can do this using `fitstat` in `fixest`

```
set.seed(7)
# Create data where z1 is valid and z2 is invalid
df <- tibble(z1 = rnorm(1000),
             z2 = rnorm(1000),
             x = z1 + z2 + rnorm(1000),
             y = 2*x + z2 + rnorm(1000))

iv <- feols(y ~ 1 | x ~ z1 + z2, df, se = 'hetero')
fitstat(iv, 'sargan')
```

```
## Sargan: stat = 248.7, p < 2.2e-16, on 1 DoF.
```

- The null hypothesis of the **Sargan test** is that the covariance between the instruments and the error term is zero

$$\mathrm{corr}(Z, U) = 0$$

- Thus, rejecting the null indicates that at least one of the instruments is not valid
- So we reject the null, indicating that one of the instruments is endogenous (although without seeing the true DGP we couldn't guess if it were Z_1 or Z_2)

Weak IV: overidentification test

- The true effect is 2

```
iv1 ← feols(y ~ 1 | x ~ z1, df, se = 'hetero')
iv2 ← feols(y ~ 1 | x ~ z2, df, se = 'hetero')
```

	Model 1	**Model 2**
(Intercept)	0.029	0.012
	(0.043)	(0.049)
fit_x	2.058***	2.912***
	(0.044)	(0.046)
Num.Obs.	1000	1000
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001		

Quantile regression

- Consider a very simple OLS

$$Y_i = \alpha + D_i' \beta_1 + U_i$$

  where Y_i is an outcome variable, D_i is a treatment variable
- What is the interpretation of the effect of D_i on Y_i?
  - it is the expected change in the outcome for a person moving from untreated to treated
  - in other words, it characterizes the **mean** of our outcome variable

Quantile regression

- What if we care about other things but the mean?
  - what are the effects of a subsidized insurance policy on medical expenditures for people with low-, medium-, and high- expenditures?
  - what are the effects of a training program on employment opportunities for people with different years of education?
- Quantile regression can handle such questions
- Quantile regression also solves problems with
  - skewed variables – no more worrying about logs or outliers in the outcome variable
  - censoring
- But it has its own issues
  - it is noisier
  - it is challenging to interpret in an intuitive way
- If you have underlying theory that has implications for distribution of the effects, the quantile regression is the right tool for empirical analysis

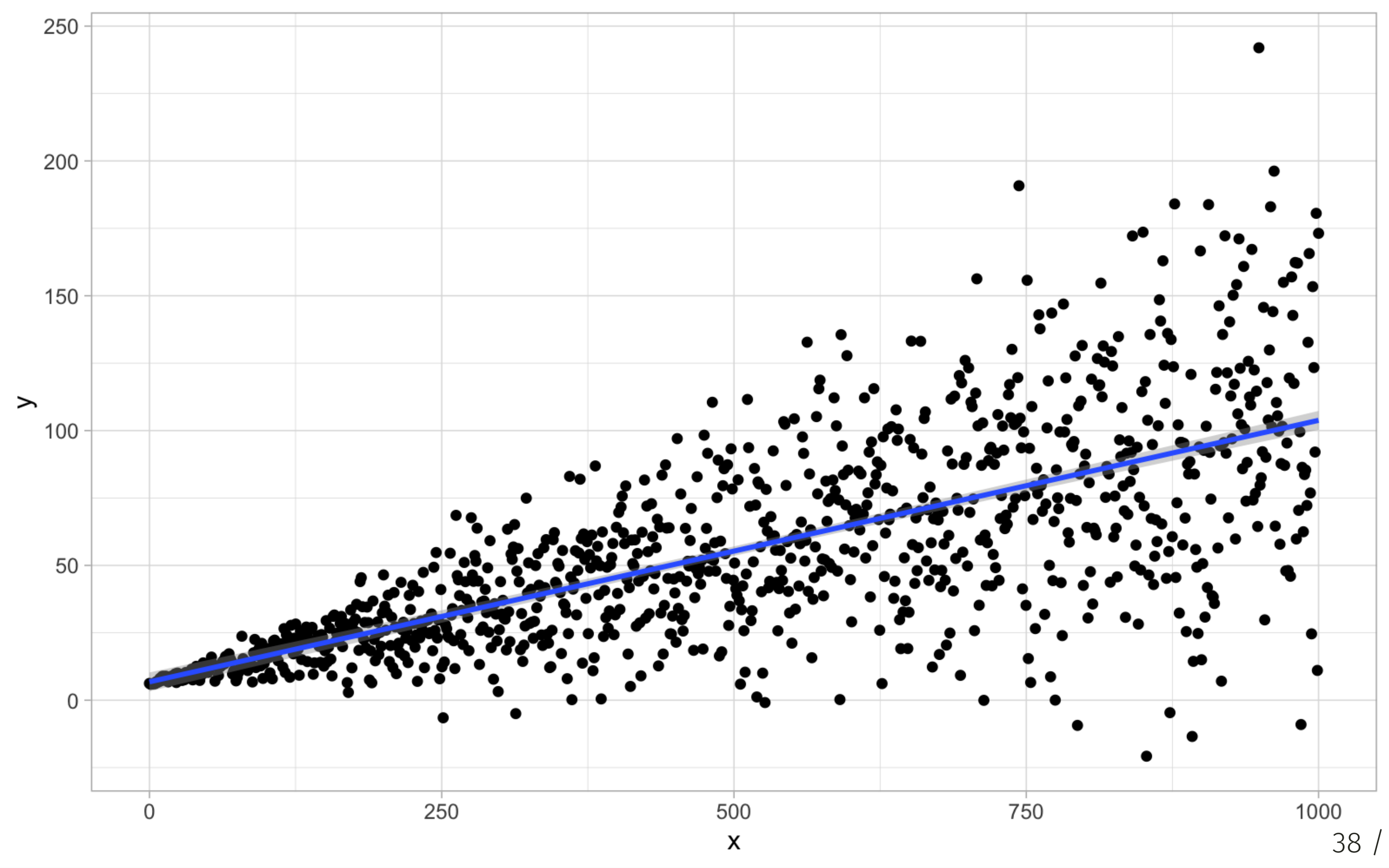
Quantile regression: simulation

- Let's simulate a data set with normal random errors with a non-constant variance

```
set.seed(7)
df <- tibble(x = seq(0, 1000, length.out = 1000),
             # non-constant variance
             sig = 0.1 + 0.05*x,
             y = 6 + 0.1*x + rnorm(1000, mean = 0, sd = sig))
```

Quantile regression: simulation

- We can see the increasing variability: as X gets bigger, Y becomes more variable



Quantile regression: simulation

- The estimated mean of an OLS regression 0.097 is still unbiased
  - but it doesn't tell us much about the relationship between X and Y
  - especially as X gets larger
- To perform quantile regression, use the `quantreg` package and specify $\textbf{tau}$ - a quantile

```
library(quantreg)
qr <- rq(y ~ x, df, tau = 0.9)
```

```
##
## Call: rq(formula = y ~ x, tau = 0.9, data = df)
##
## tau: [1] 0.9
##
## Coefficients:
##             coefficients lower bd upper bd
## (Intercept) 6.28754      5.97387  7.43619
## x           0.16066      0.15581  0.16472
```

- The X coefficient estimate of 0.161 says that "one unit increase in X is associated with 0.161 increase in the 90^{th} quantile of Y"

Quantile regression: simulation

- Let's look at different quantiles at once

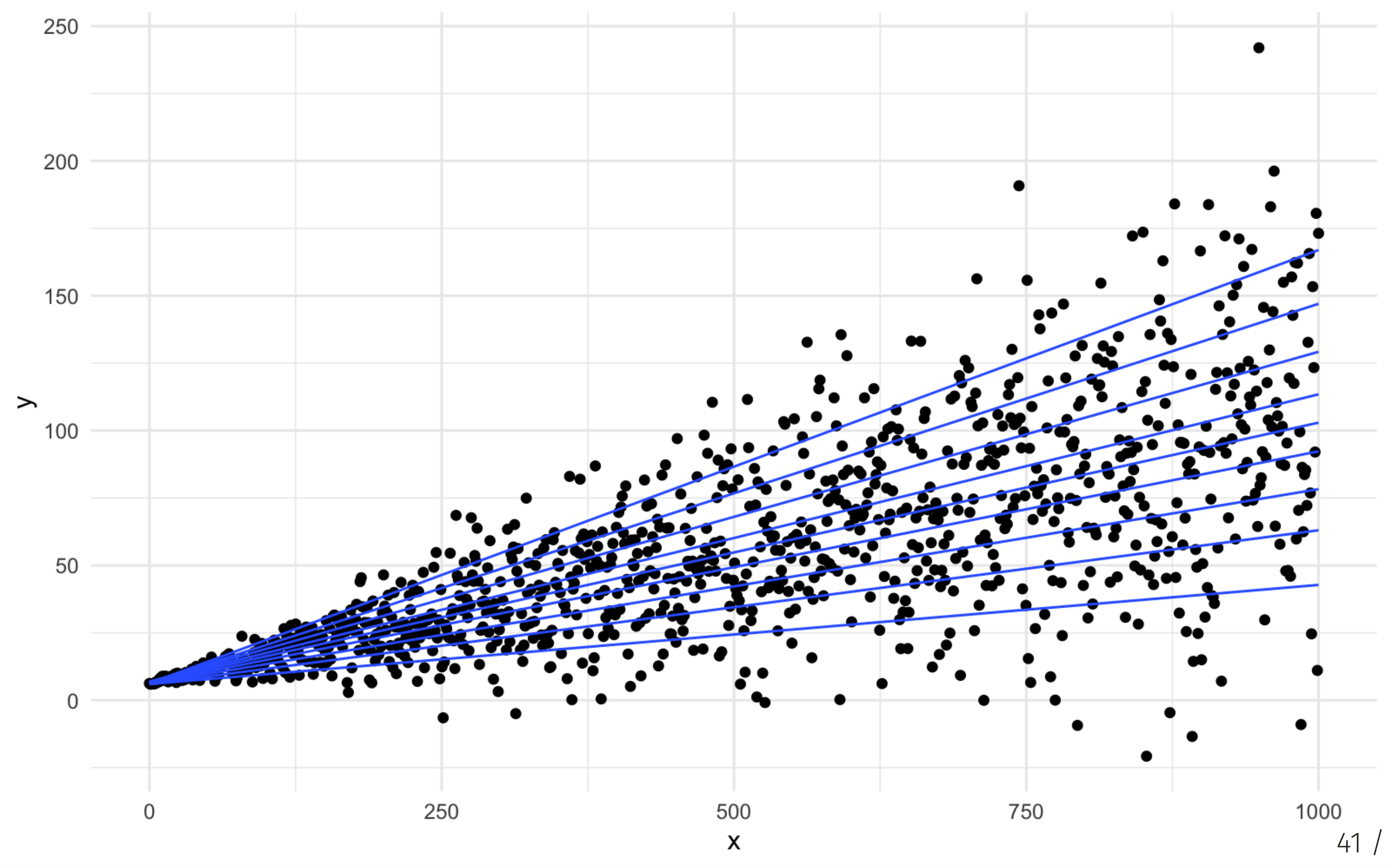
```
qr2 <- rq(y ~ x, data = df, tau = seq(0.1, 0.9, by = 0.1))
coef(qr2)
```

```
##                tau= 0.1   tau= 0.2   tau= 0.3   tau= 0.4   tau= 0.5 tau= 0.6
## (Intercept) 5.96165209 5.94698236 6.18267262 6.35504045 6.66850218 6.883414
## x           0.03680116 0.05708777 0.07209288 0.08594706 0.09627532 0.106561
##               tau= 0.7  tau= 0.8 tau= 0.9
## (Intercept) 6.8040670 6.4478973 6.287539
## x           0.1224144 0.1405742 0.160662
```

- The intercept estimates don’t change much but the slopes steadily increase

Quantile regression: simulation

- Let's plot our quantile estimates



Quantile regression: simulation

- Each black dot is the slope coefficient for the quantile indicated on the x axis
- The red lines are the OLS estimate and its confidence interval
- Lower and upper quantiles are well beyond the OLS estimate

Quantile regression: inference

- There are several alternative methods of conducting inference about quantile regression coefficients
  - rank-inversion confidence intervals: `summary.rq(qr)`
  - more conventional standard errors: `summary.rq(qr, se = "nid")`
  - bootstraped stanard errors: `summary.rq(qr, se = "boot")`
- To read more about calculating confidence intervals, use `?summary.rq`

References

Books

- Huntington-Klein, N. The Effect: An Introduction to Research Design and Causality, Chapter 19: Instrumental Variables
- Cunningham, S. Causal Inference: The Mixtape, Chapter 7: Instrumental Variables
- Adams, C. Learning Microeconometrics with R, Chapter 6: Estimating Selection Models

Slides

- Huntington-Klein, N. Econometrics Course, Week 8: Instrumental Variables
- Goldsmith-Pinkham P. Applied Empirical Methods Course, Week 7: Linear Regression III: Quantile Estimation

Articles

- Lee, D. S., McCrary, J., Moreira, M. J., & Porter, J. R. (2021). Valid t-ratio Inference for IV (No. w29124). National Bureau of Economic Research