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:

```r
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 + Treated * 5 + 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 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?
- But suppose that we do not observe wages below a specific threshold (due to the features of a questionnaire, privacy concerns, coding, etc.)
- We need to generate data containing years of education and wages

Censored regression: simulation

```r
# 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(desc(wage_star))
```

```
## # A tibble: 3 × 3
##   education wage_star  wage
##       <dbl>     <dbl> <dbl>
## 1      14.9     4202.  3500
## 2      14.8     4174.  3500
## 3      14.9     4167.  3500
```

```
## # A tibble: 3 × 3
##   education wage_star  wage
##       <dbl>     <dbl> <dbl>
## 1      5.26     1890. 1890.
## 2      5.04     1878. 1878.
## 3      5.06     1837. 1837.
```

Censored regression: OLS

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

```r
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"

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 tobit-model
- The solution provided by the 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 wages: wages above 3500 are coded as 3500

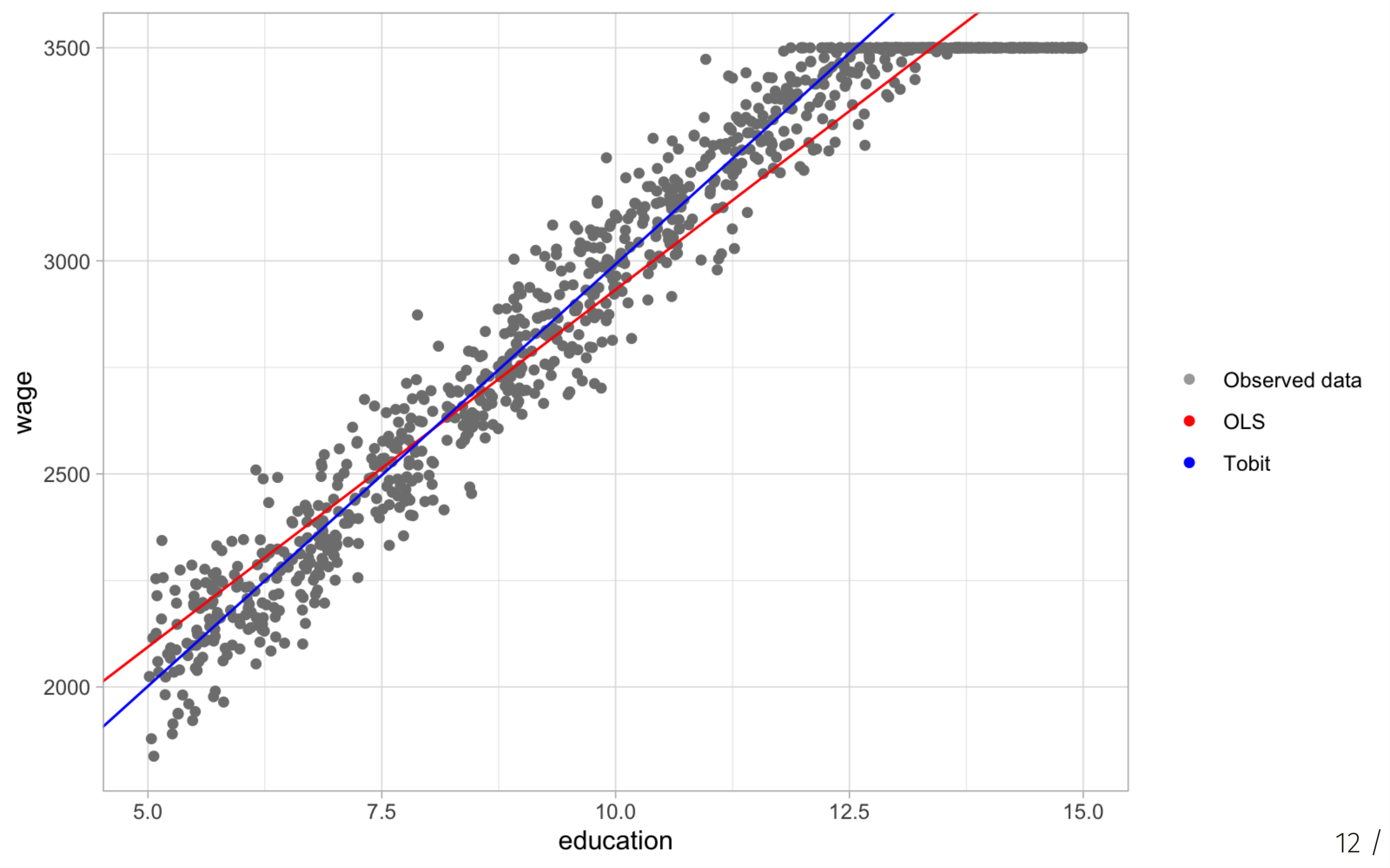
```r
library(censReg)
tobit_model <- censReg(wage ~ education, data = df, right = 3500)
```

	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 the tobit-model is that the unobserved term is normally distributed (which is the case in our simulated dataset)
- 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, we can't use censoring
- Censoring cannot be applied because the availability of data is influenced by the choice of agents (i.e. 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 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 that equal to

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

Sample selection model: assumptions

- To estimate the sample selection model, distributional assumptions on the disturbances terms are made, such as bivariate normality

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

- 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 dataset 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(1, 0.7), c(0.7, 1))),
             i_star = 4 - 5 * z + uv[, 1],
             y_star = 6 - 3 * x + uv[, 2],
             # this is our selection mechanism
             y = ifelse(i_star > 0, y_star, 0))
head(df)
```

```
## # 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  0.865  -0.237 -0.0799   4.51  0
## 2 0.398   0.813  0.0624 -0.256  2.07      3.31  3.31
## 3 0.116   0.278 -0.540   0.229  2.88      5.39  5.39
## 4 0.0697  0.968 -1.74   -1.90   1.91      1.20  1.20
## 5 0.244   0.247 -0.250   0.594  2.53      5.85  5.85
## 6 0.792   0.905 -1.78   -0.733 -1.74      2.55  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)	4.457***	4.635***
	(0.287)	(0.135)
z	-5.647***	
	(0.379)	
x		-2.074***
		(0.237)
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)
- 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 those steps
  - **First stage**: predict treatment X_1 with the instrument Z, with a control variable X_2

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

- **Second stage**: use that equation to predict X_1, getting $\hat{X}_1$. Then, use those 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
  - **Relevance**: the instrument must be a *strong predictor* of the treatment. It can't be trivial or unimportant

$$\mathbf{cov}(X, Z) \neq 0$$

  - **Validity**: the instrument must actually be *exogenous* (or at least exogenous after adding controls)

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

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

$$\frac{\mathbf{cov}(Z, U)}{\mathbf{cov}(X, Z)}$$

- If Z is valid, then in infinite samples $\mathbf{cov}(Z, U) = 0$ and this goes away. But in a non-infinite sample, it will be nonzero by chance, inducing some bias. The smaller the sample, the more likely we are to get a large value by random chance
- The bias is smaller
  - the stronger the relationship between X and Z
  - the smaller the sum of squared errors
  - the bigger the variation in X
- What happens when $\mathbf{cov}(X, Z)$ is small?

IV: weak instrument problem

- If Z has only a trivial effect on X, then it's not **relevant** (even if it's truly **exogenous**)
- And our **small-sample bias** will be big
- Thus, weak instrument problem means that we probably shouldn't be using IV in small samples
- This also means that it's really important that $\mathbf{cov}(X, Z)$ is not small
- There are some rules of thumb for how strong an instrument must be to be counted as "not weak"
- A t-statistic above 3, or an 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 - 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 dataset with instruments

```
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)
head(df)
```

```
## # A tibble: 6 × 4
##       z1     u1     x1        y1
##    <dbl>  <dbl>  <dbl>     <dbl>
## 1  2.29   1.25   6.86   26.8
## 2 -1.20  -0.766 -5.40  -20.0
## 3 -0.694  0.216 -0.358   0.00737
## 4 -0.412 -0.364 -1.91   -7.55
## 5 -0.971 -0.821 -2.72  -12.3
## 6 -0.947  0.582  2.32    9.87
```

Weak IV: simulation

```r
# 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 relevance: $\mathbf{cov}(X_1, Z_1) =$ 0.068, so it's a weak instrument
  - Check validity: $\mathbf{cov}(Z_1, U_1) =$ 0.027, pretty close to zero

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),
             # all coefficients stay the same
             x2 = 0.2*z2 + 4*u2 + rnorm(1000),
             y2 = 3*x2 + 5*u2)
head(df)
```

```
## # A tibble: 6 × 5
##        v      z2     u2     x2     y2
##    <dbl>   <dbl>  <dbl>  <dbl>  <dbl>
## 1  2.29  -1.04    1.64   7.57   30.9
## 2 -1.20   0.431  -2.22  -9.87  -40.7
## 3 -0.694  0.910  -1.15  -5.12  -21.1
## 4 -0.412  0.0479 -0.412 -3.49  -12.5
```

Weak IV: simulation

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

	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 biased OLS estimator than by using IV
- Why? Because of the weak instrument problem
  - Check relevance: $\mathbf{cov}(X_2, Z_2) =$ 0.021
  - Check validity: $\mathbf{cov}(Z_2, U_2) =$ -0.036

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)
head(df)
```

```
## # A tibble: 6 × 5
##        v      z2      u2    x3    y2
##    <dbl>   <dbl>   <dbl> <dbl> <dbl>
## 1  2.29  -1.04    1.64    4.65  22.2
## 2 -1.20   0.431  -2.22   -8.66 -37.1
## 3 -0.694  0.910  -1.15   -2.57 -13.5
## 4 -0.412  0.0479 -0.412  -3.36 -12.1
## 5 -0.971  0.150   0.659   5.27  19.1
## 6 -0.947  1.53    0.0833  5.19  16.0
```

Weak IV: simulation

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

	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 relevance: $\text{cov}(X_3, Z_2) =$ 0.697
  - Check validity: $\text{cov}(Z_2, U_2) =$ -0.036

Weak IV: F-test

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

```r
thef <- fitstat(iv_model3, 'ivf', verbose = FALSE)$`ivf1::x3`$stat
```

- 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 10 and they suggest to use as a rule of thumb the minimum value of the F-test 104.7, which is needed to ensure a test with 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: simulation

- We can do this using `diagnostics = TRUE` in `iv_robust` again

```
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 instrument and the error term is zero

$$\textbf{cov}\left(Z, U\right) = 0$$

- 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: simulation

- 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 the very simple OLS version testing this model using the experimental data:

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

  where Y_i is an outcome variable, D_i is a tretment variable
- Recall that this will estimate our ATE for the treatment
- What is the interpretation of this affect?
  - $E(Y_i(1)) - E(Y_i(0))$, i.e. the expected change in the outcome for a person moving from untreated to treated. That's a useful metric
- 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 factors influencing total medical expenditures for people with low-, medium-, and high- expenditures?
  - What are the effects of a training program on employment opportunities for people with the different number of years of education?
- Quantile regression also solves the problems with
  - Skewed variables – no more worrying about logs or outliers in the outcome variable
  - Censoring – in many datasets, our outcome variables are top-coded or bottom-coded
- 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, quantile regression is the rigth tool for empirical analysis

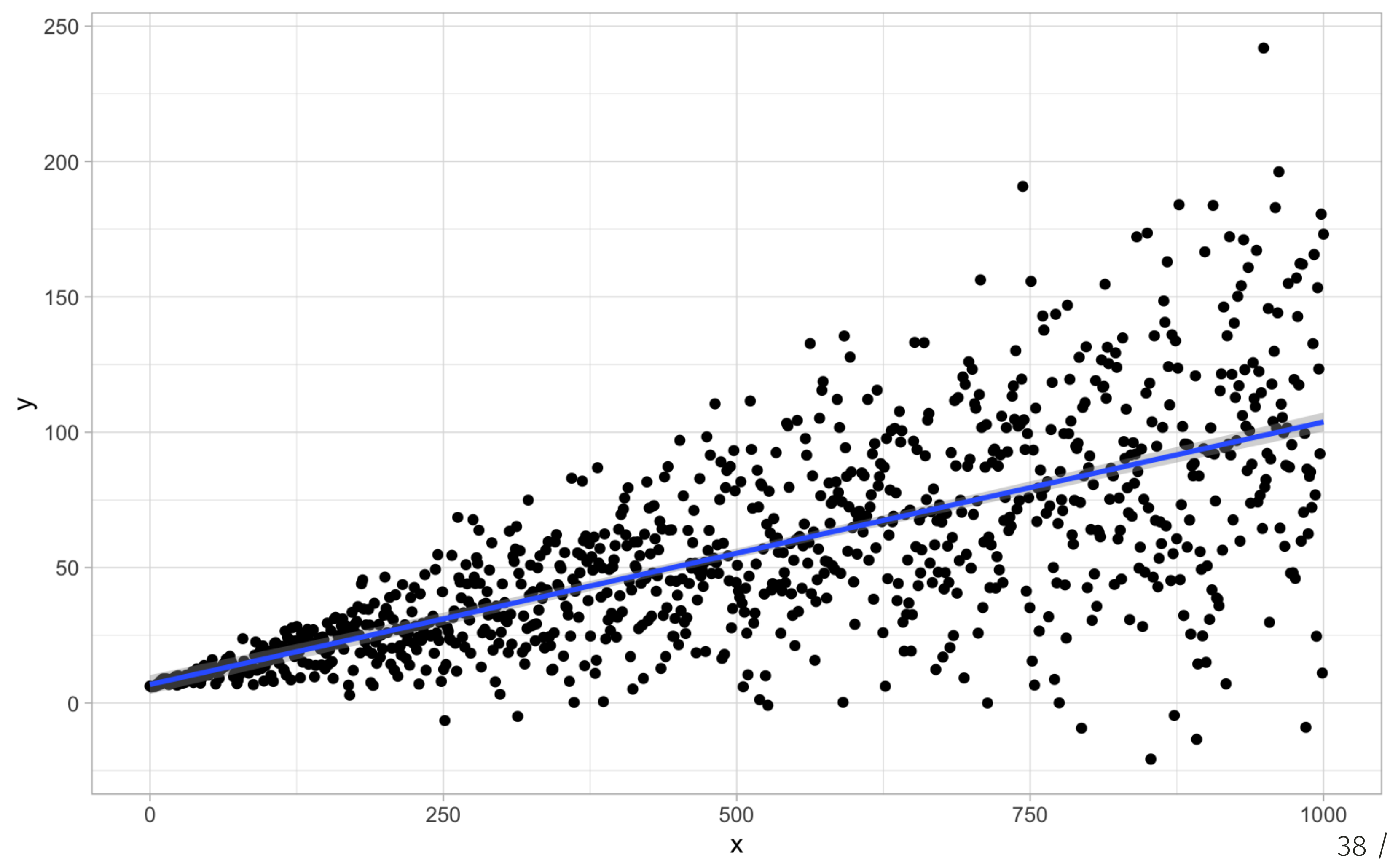
Quantile regression: simulation

- Let's simulate the dataset with normal random error with 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 conditional on X 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 "one unit increase in X is associated with 0.161 increase in the 90^{th} quantile of Y"
- The "lower bd" and "upper bd" values are confidence intervals calculated using the "rank" method

Quantile regression: simulation

- Let's look at different quantiles at once

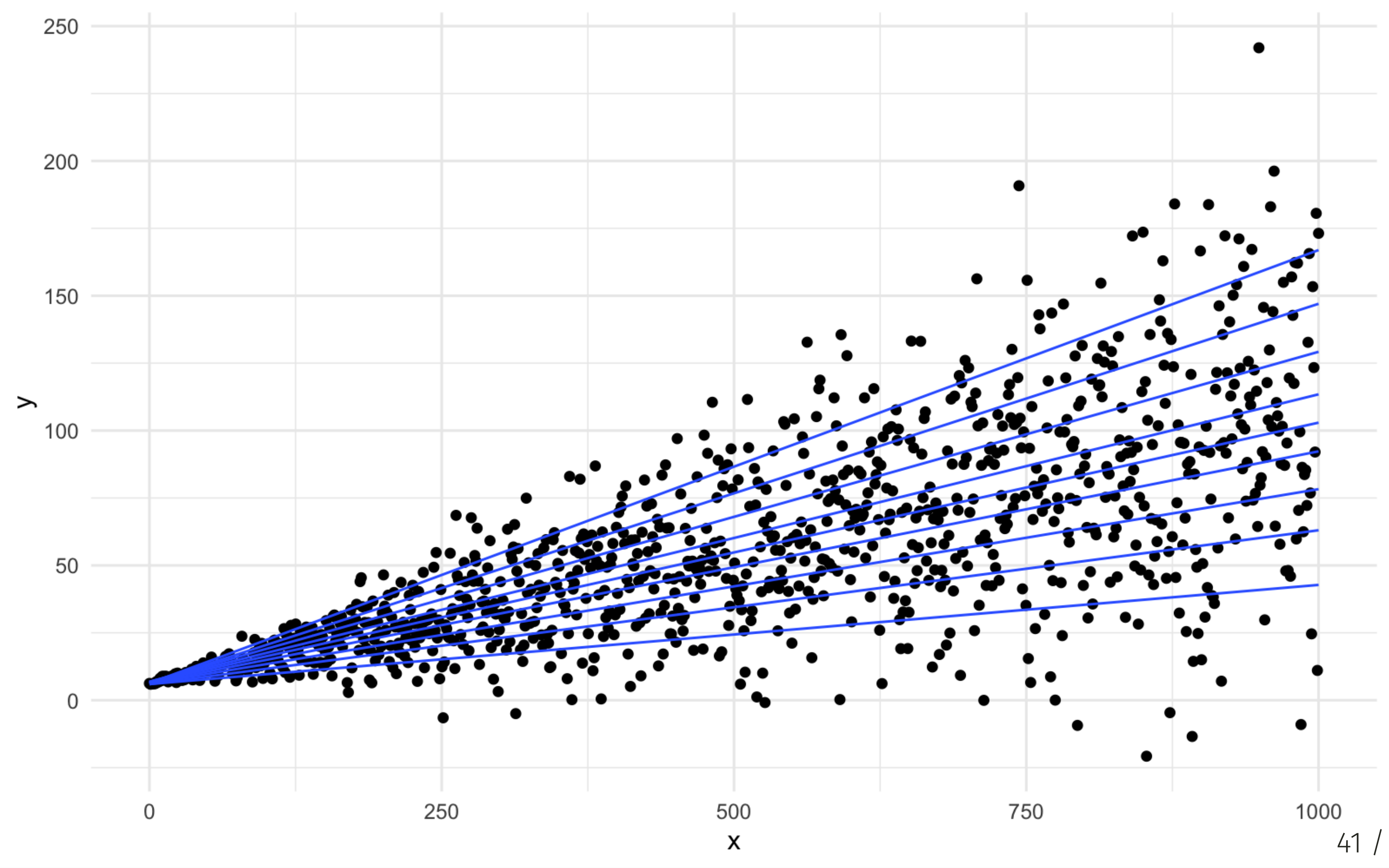
```
qr2 <- rq(y ~ x, data = df, tau = seq(0.1, 0.9, by = 0.1))
coef(qr2)
```

```
##                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 estimate doesn’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 least squares estimate and its confidence interval
- Lower and upper quartiles are well beyond the least squares 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 Slides, 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