The intuition behind quantile regression

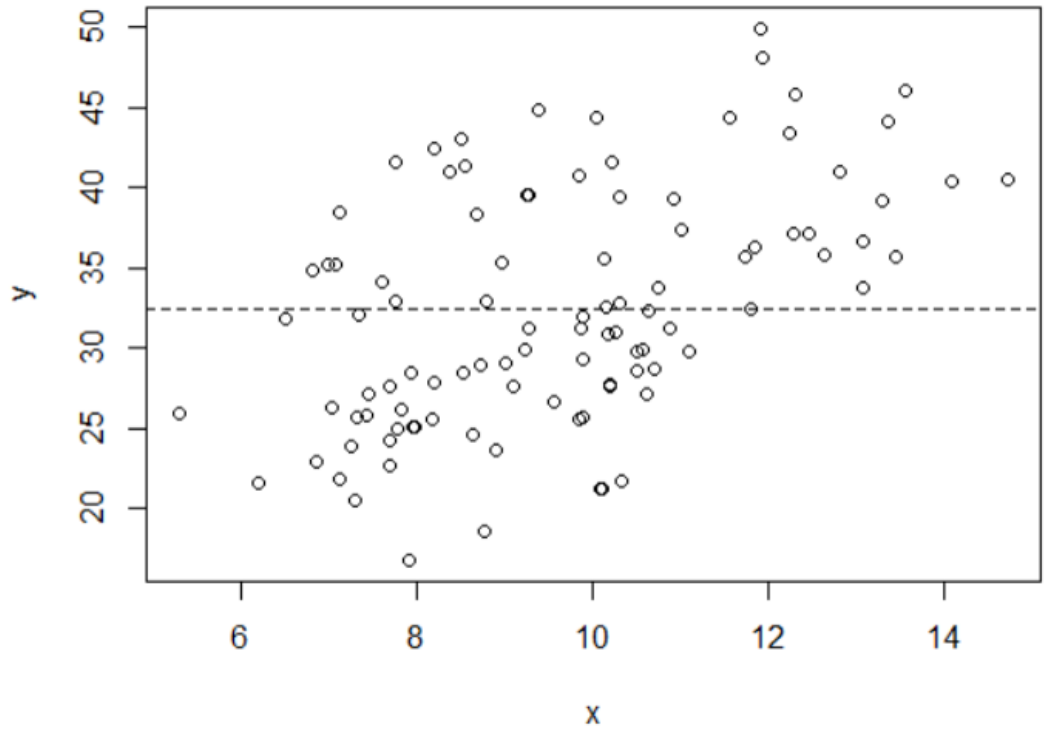
Prof. Alex COAD

$$y_i = a + bx_i + \varepsilon_i$$

- If we don't know anything about x, what is our best estimate of y?
- The mean of y: $\bar{y}$
  - Known as the unconditional mean
- If we do know the value of x, what is our best estimate of y?
  - The conditional mean of y (i.e. conditional on x): $E(y|x)$

The unconditional mean is our best estimate of y, if we don't know x

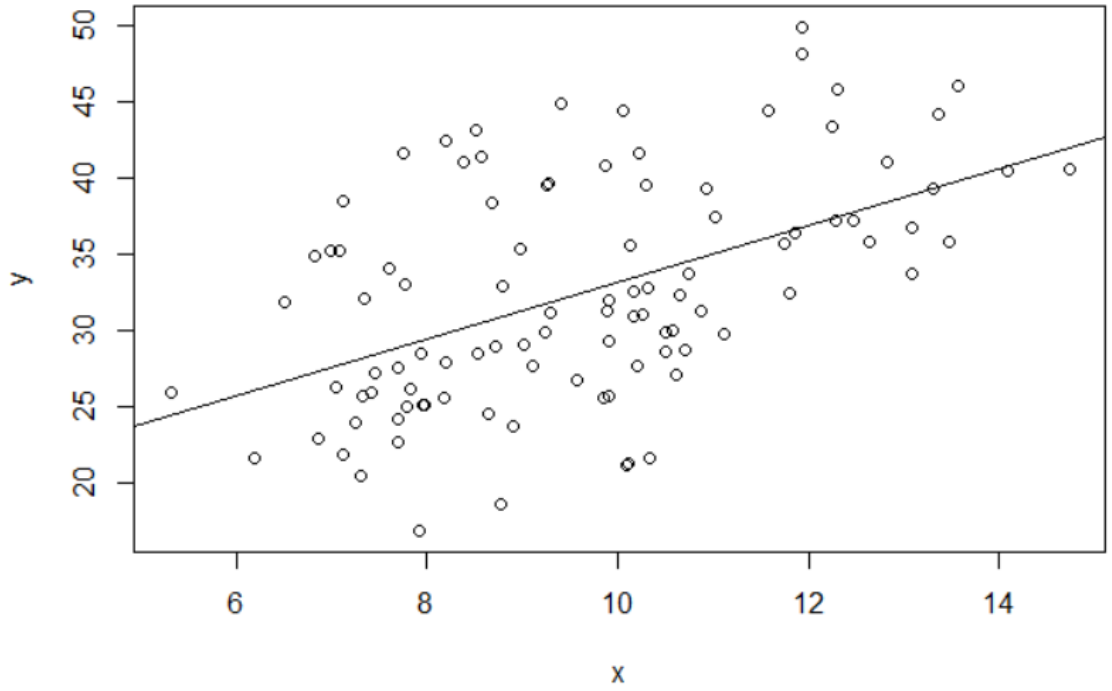
```
>
> # the mean of y
> mean(mydata$y)
[1] 32.42825
> plot(y~x, data=mydata)
> abline(h=mean(mydata$y), lty=2)
>
```



But if we do know x...

- our best estimate of y is the conditional mean
- $E(y|x)$

```
>
> # the CONDITIONAL mean of y (conditional on x)
> summary(lm(y~x, data=mydata))

Call:
lm(formula = y ~ x, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-12.402  -4.215  -1.378   4.881  13.185

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.4855     3.1417   4.611 1.21e-05 ***
x             1.8621     0.3192   5.833 7.01e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.392 on 98 degrees of freedom
Multiple R-squared:  0.2577,    Adjusted R-squared:  0.2501
F-statistic: 34.03 on 1 and 98 DF,  p-value: 7.009e-08

> plot(y~x, data=mydata)
> abline(lm(y~x, data=mydata))
>
```



Unconditional quantiles

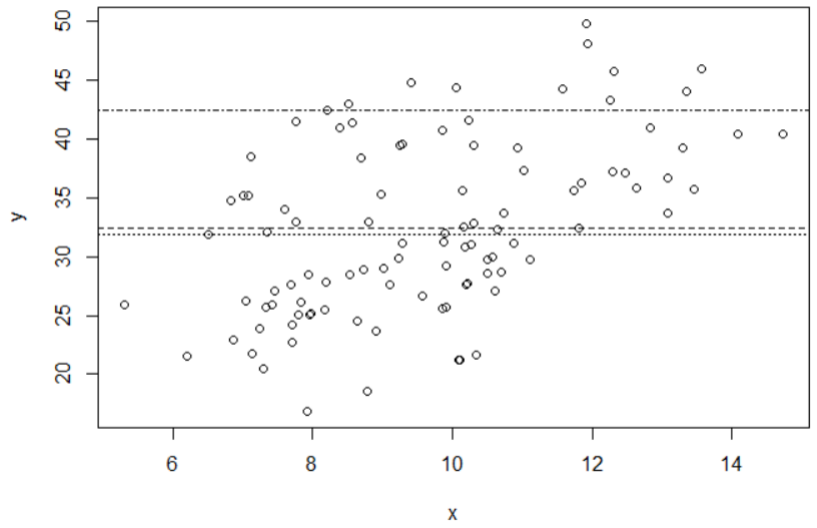
Unconditional median

- 50% of observations are above the median

Unconditional 90% quantile

- 10% of observations are above the line, if τ=0.9

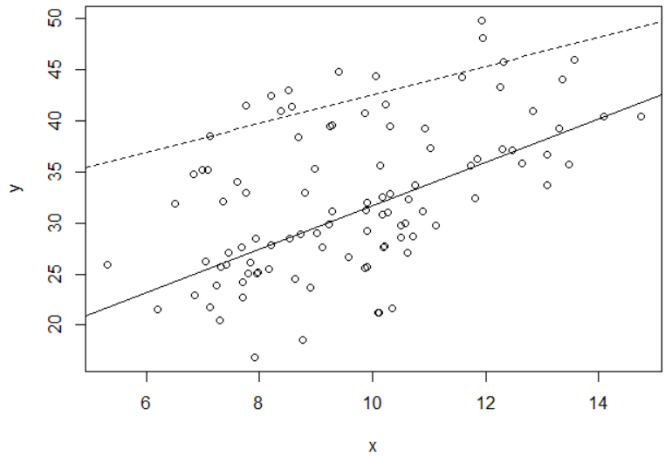
```
>
> # the median of y
> median(mydata$y)
[1] 31.93987
> plot(y~x, data=mydata)
> abline(h=median(mydata$y), lty=3)
> # note: in this case, the median is not equal to the mean
> abline(h=mean(mydata$y), lty=2)
>
>
> # the 90% quantile of y
> quantile(mydata$y, probs=0.9)
     90%
42.49014
> abline(h=quantile(mydata$y, probs=0.9),lty=4)
>
```



OLS gives us the conditional mean

Median regression

- Conditional median
- 50% of observations are above the median

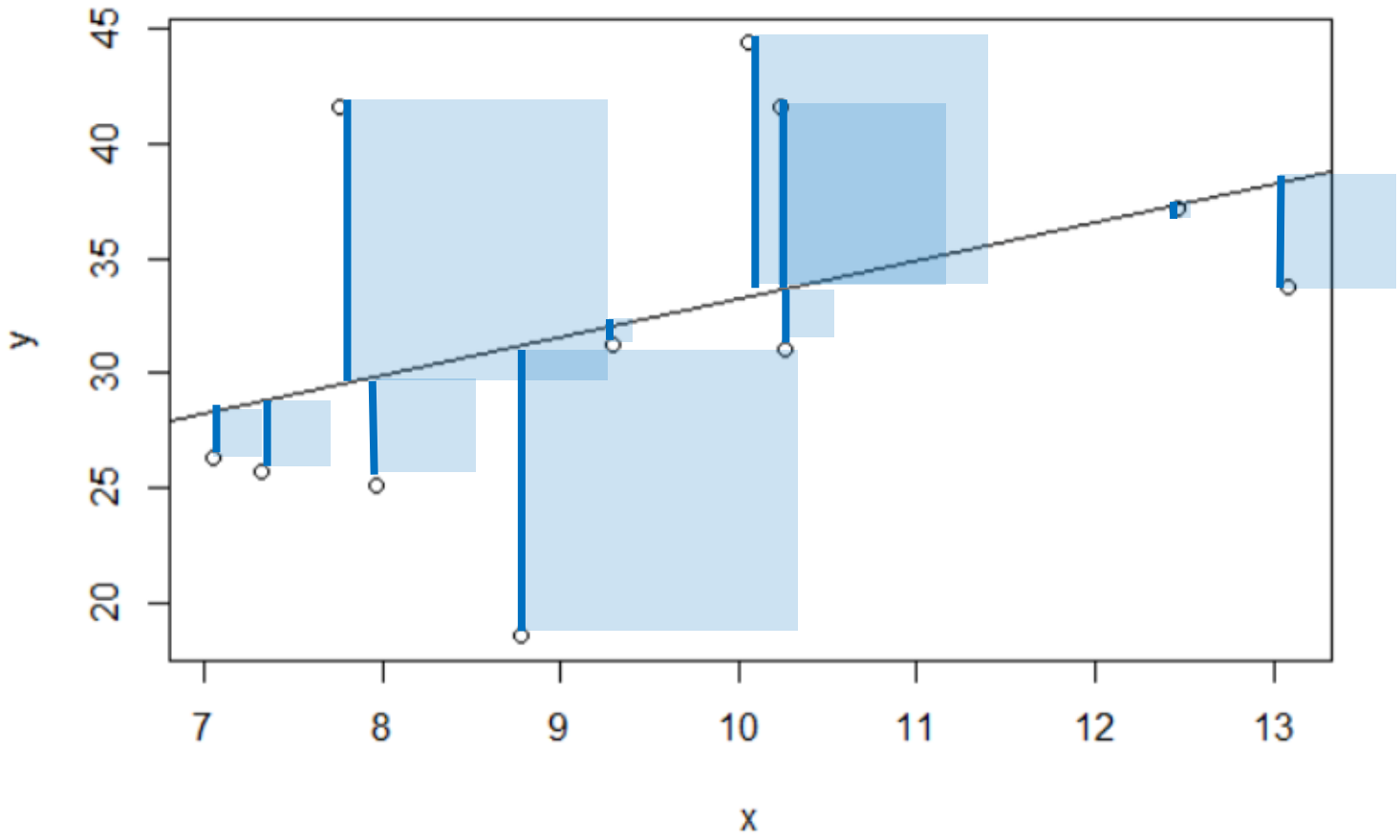


Quantile regression

- Conditional quantiles
  - Median regression as a special case, where $\tau=0.5$
  - 10% of observations are above the line, if $\tau=0.9$

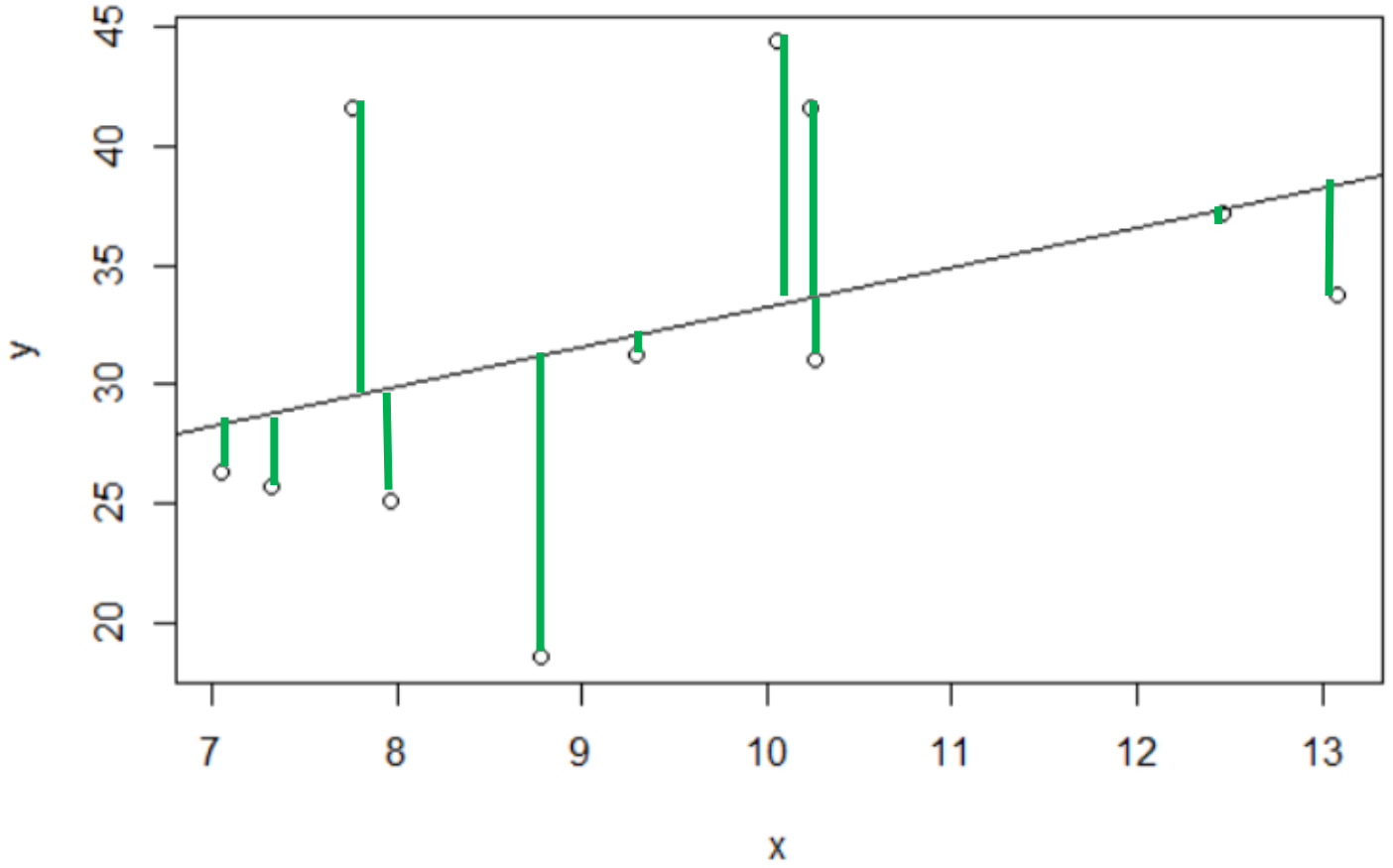
Remember the OLS line of best fit

- OLS: Minimize the sum of SQUARES of residuals



Remember the OLS line of best fit

- OLS: Minimize the sum of SQUARES of residuals
- Median regression: Minimize the sum of absolute values of residuals
- Median (unlike mean): 50% of observations must be above, 50% below
- Median regression (unlike OLS): 50% of observations must be above the line of best fit, 50% must be below



Quantile regression: background

- OLS regressions model the (conditional) mean of the dependent variable
- We could calculate from the fitted regression line the value that y would take for any values of the explanatory variables
  - "y conditional on x"
- But this would be an extrapolation of the behaviour of the relationship between y and x at the mean to the remainder of the data
  - "the average effect for the average firm" (if you use firm-level data)
- Quantile regression models the entire conditional distribution of y (conditional on the explanatory variables)

Quantile regression: assumptions

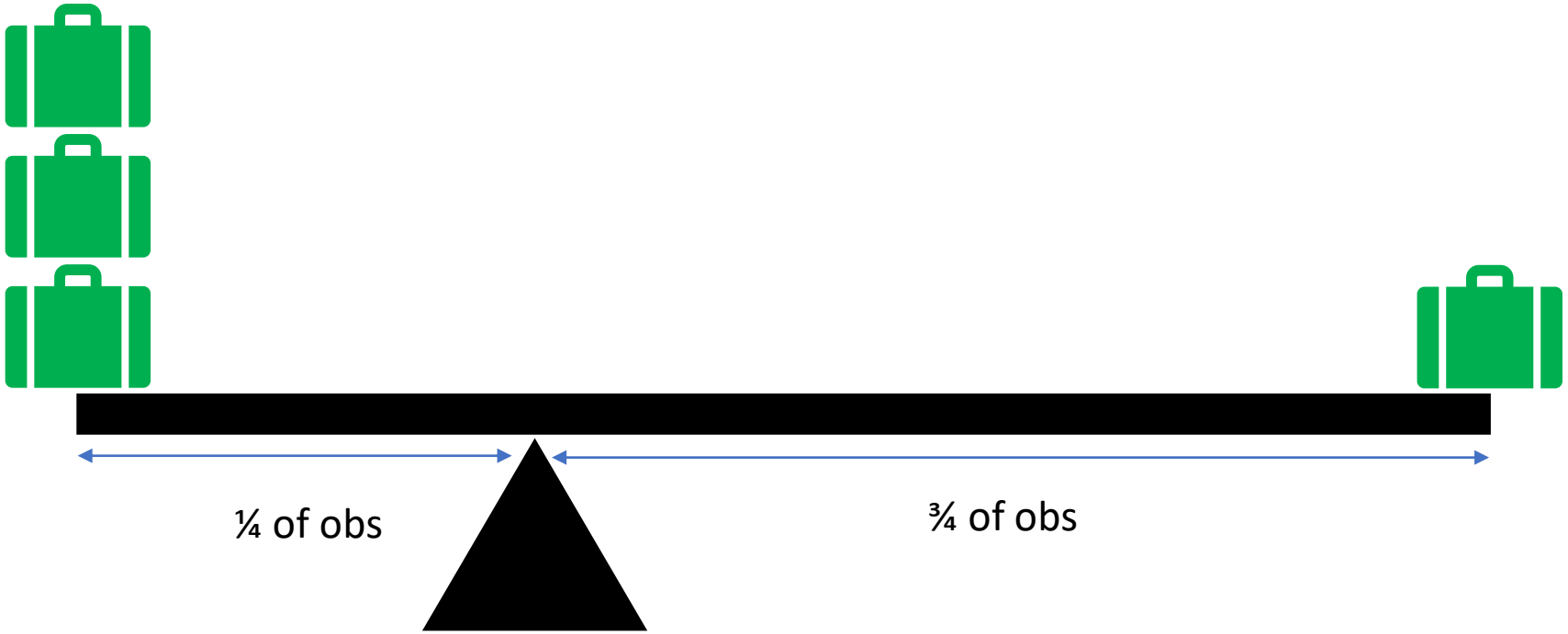
- Quantile regression assumes that the error terms are independently distributed and homoscedastic
  - Errors need not be normally distributed
    - Quantile regression & median regression are robust to outliers on the DV that tend to ±∞
  - Quantile regression is a "semi-parametric" technique since no distributional assumptions are required
  - Semi-parametric, not non-parametric, because we assume the regression model is linear

Quantiles: τ

- Quantiles must lie between 0 and 100%
- Common choices for results tables are: 10%, 25%, 50%, 75%, 90%

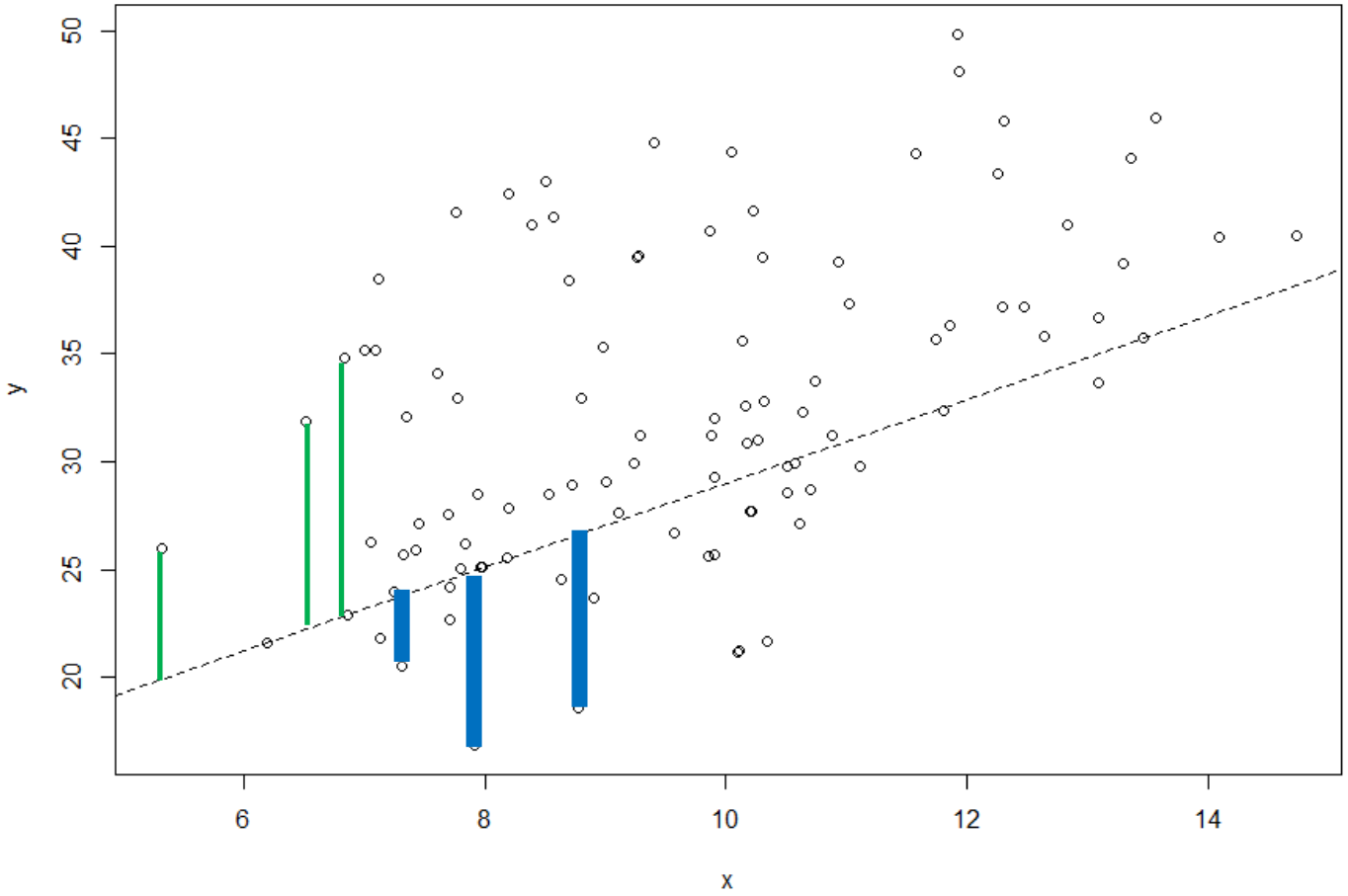
- $\tau = 0.5$: the median
- $\tau = 0.99$? Keep in mind the size of your dataset!
  - $\tau = 0.99$ means that only 1% of your observations are above the line of best fit

τ as the “fulcrum”: here τ = 25%



- QR uses all observations
  - But different weights to datapoints above & below the line of best fit
- We can’t choose $\tau = 0\%$ and $\tau = 100\%$!

τ as the “fulcrum”: here τ = 25%



- When $\tau = 25\%$, points **below** the best-fit line get 3x higher weights than those above

Analysis of quantiles: how not to do it

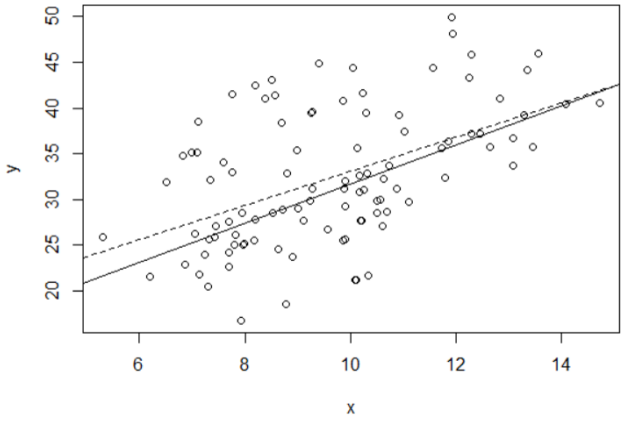
- Possible alternative to quantile regression?
- Partition the data into subsamples, and run separate regressions on each
  - E.g., drop the bottom 90% of observations on (conditional) y, and the corresponding data points for the x's, and run an OLS regression on the remainder

- However, this could lead to severe sample selection biases
- Also, you throw away your observations
  - smaller dataset, less statistical power

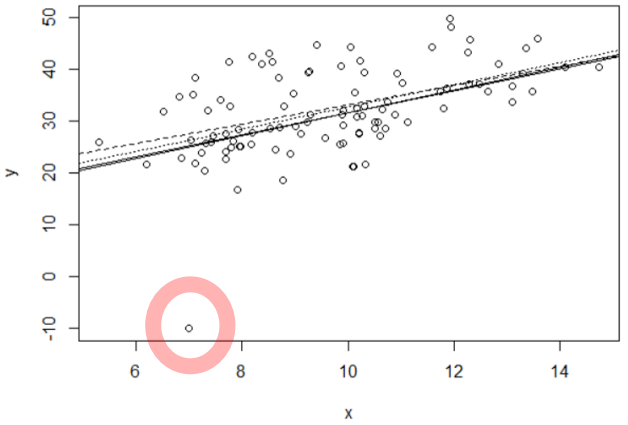
- Quantile regression does not partition the data
- All observations are used in the estimation of the parameters for every quantile
  - but given different weights

Quantile regression is less sensitive than OLS to outliers on y

```
>
> # let's manually introduce an outlier in row 5
> mydata$y[5] <- -10
> head(mydata)
```



- The median regression solution changes a bit...
  - Model rq.1: $\hat{b} = 2.132$
  - Model rq.2: $\hat{b} = 2.208$
- … but not as much as OLS
  - Model ols.1: $\hat{b} = 1.862$
  - Model ols.2: $\hat{b} = 2.159$



References

- Koenker R., Hallock K.F. (2001). Quantile Regression. Journal of Economic Perspectives, 15 (4), 143-156.