Transformation of outcome variable in linear regression – why and how

Henrik Støvring

Senior researcher, MSc, PhD, DMSc

Steno Diabetes Center Aarhus - Denmark

hersto@rm.dk

September 14, SDCA

midt regionmidtjylland

Overview

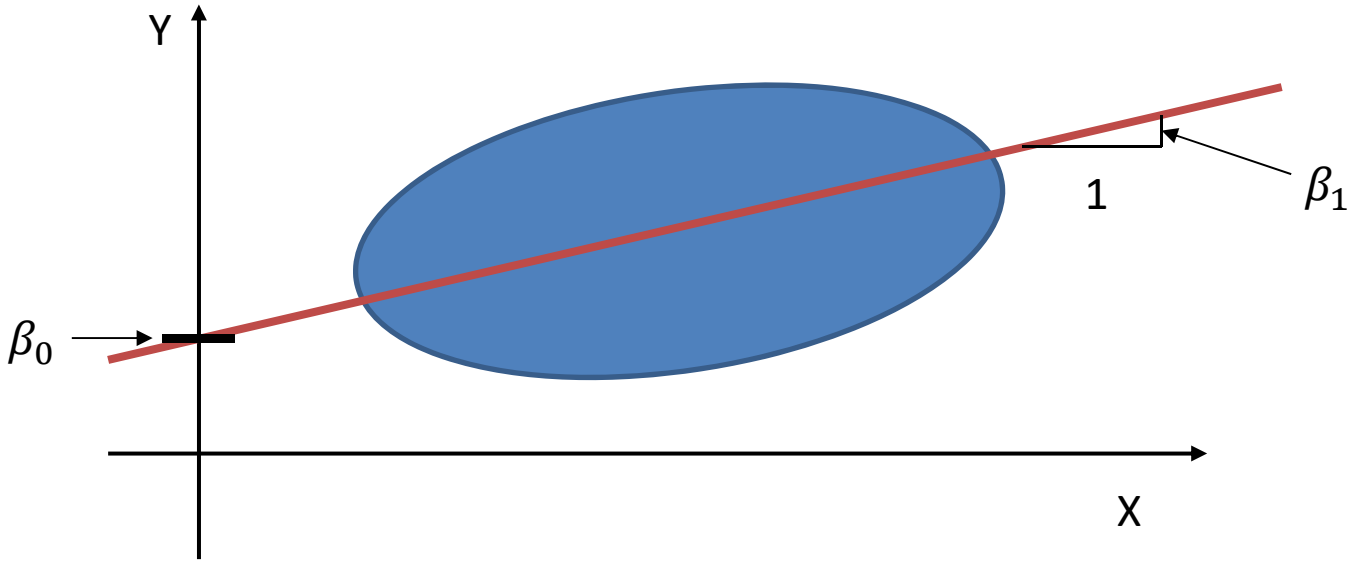
- Why do we need to transform?
- Spotting the need for a transformation
- Interpretation of regression coefficients with log-transformed outcome

Linear regression – normality assumption

- Model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$
$$\varepsilon \sim N(0, \sigma^2)$$

- here σ^2 is the variance (= Standard Deviation squared)



Statistical theory (and rule-of-thumb experience)

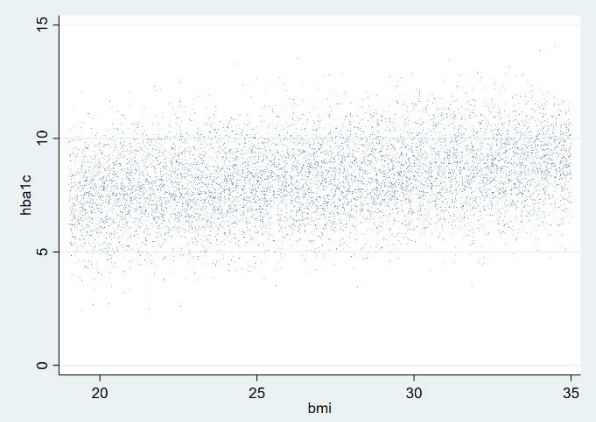
- If residual variation (distribution of ε) is not normal:
  - Standard errors are biased
  - Confidence intervals do not maintain their nominal coverage

    → Less or more than 95% of all 95% CIs will actually contain the true value

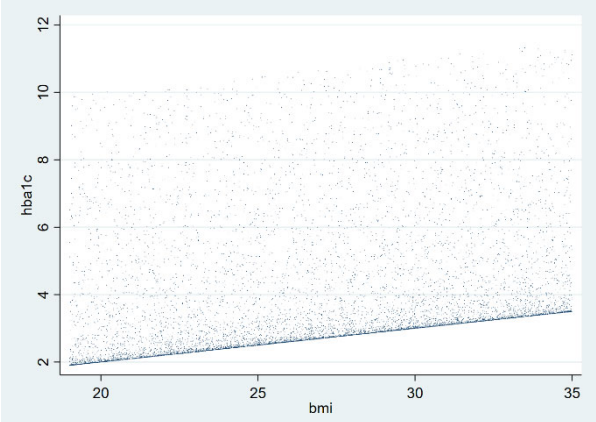
- HOWEVER: All analyses based on linear normal models (t-test, linear regression, ANOVA, ANCOVA) are considered to be rather robust to misspecification of model for residual variation

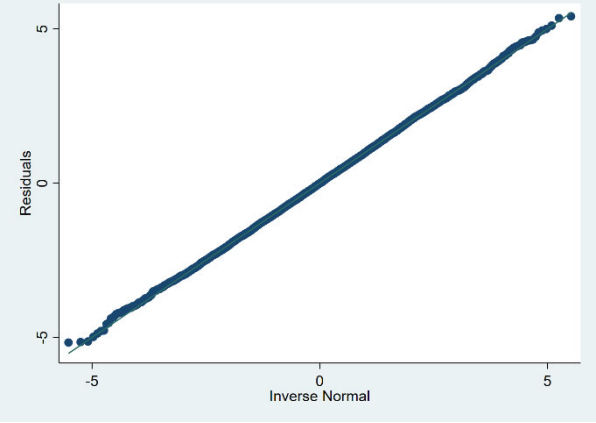
Simulation example

- Linear regression where residual variation is either
  - Normal
  - Uniform^5 (highly skewed)

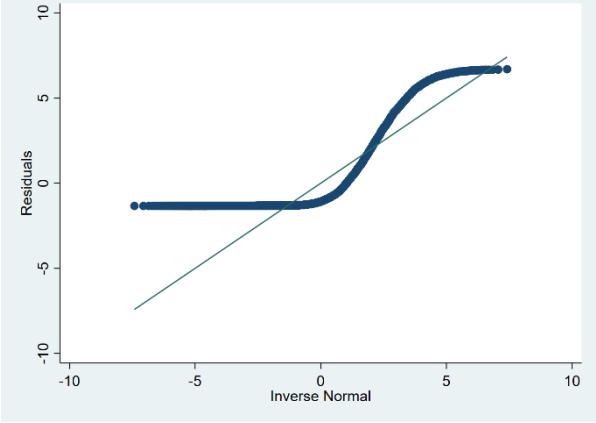


Scatter plot
Y vs X





QQ-plot of
residuals



Simulation setup

- Small datasets: n=10
- 10,000 repetitions
- For each data use linear regression to estimate slope (true value = 0.1 increase in Hba1c per BMI unit)
- Check if estimated confidence interval contains true value

Simulation results

- With **normally** distributed residual variation
  - Bias (w/ 95% CI):

    -0.0010 (-0.0033; 0.0012)
  - Coverage probability of 95% CI:

    95.1% (94.7%; 95.5%)

- With **non-normally** distributed residual variation
  - Bias (w/95% CI):

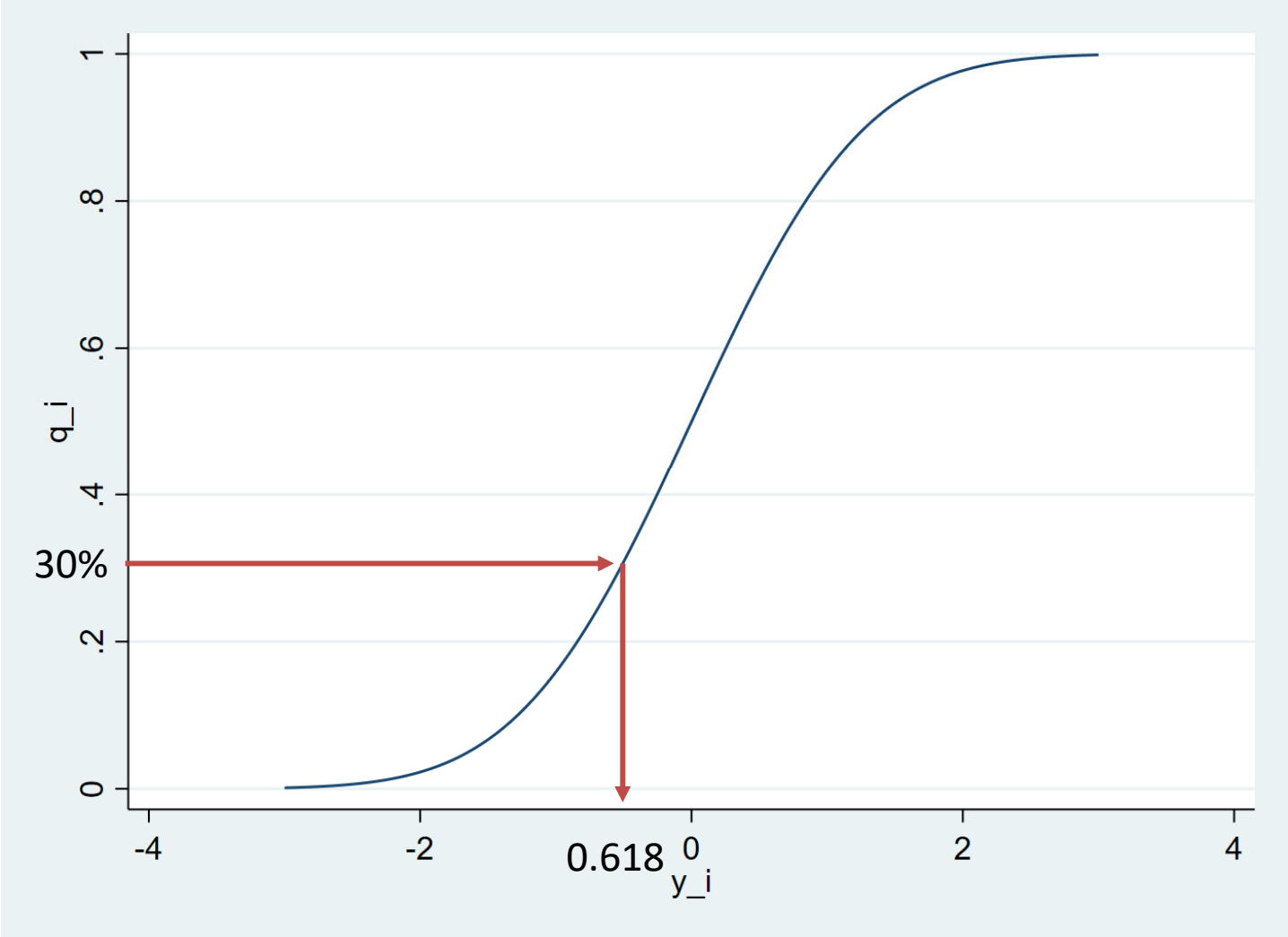
    0.0009 (-0.0021; 0.0040)
  - Coverage probability of 95% CI:

    95.7% (95.3%; 96.1%)

How do we spot a non-normal distribution

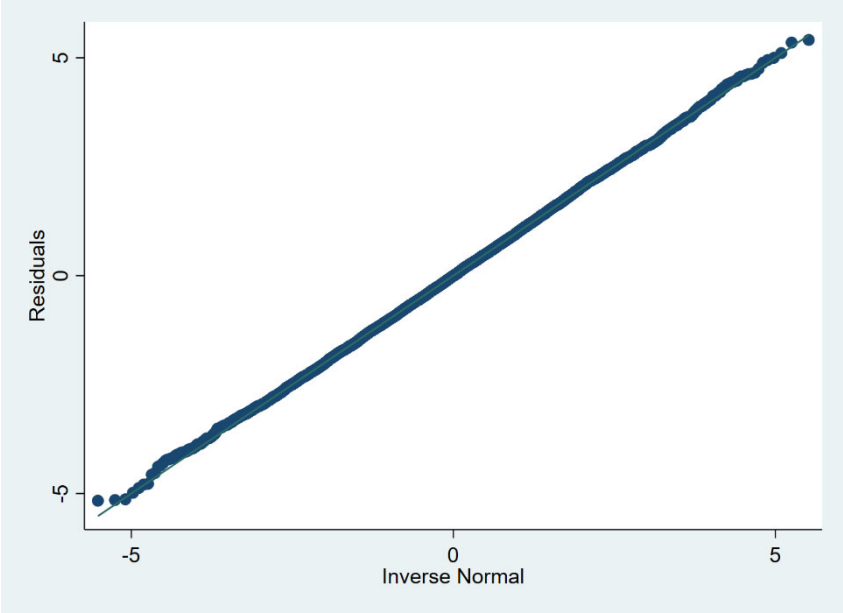
- Q-Q-plot
- Algorithm
  - Order data in ascending order $x_{[1]} < x_{[2]} < \cdots < x_{[n]}$
  - r_i is rank, ie 1, 2, 3,..., n
  - Compute $q_i = \frac{r_i}{n}$, i.e. the cumulative proportion of data points below or equal to data point i
  - Compute the quantile in a standard normal distribution for each q_i to obtain y_i
  - Plot $x_{[i]}$ vs y_i

Normal quantiles

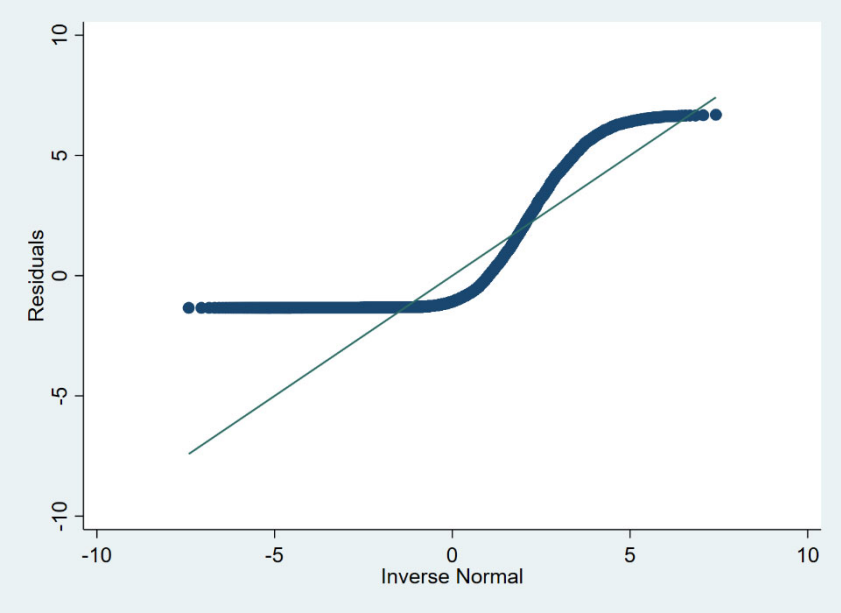


Q-Q-plot

- Rationale is to create a linear function if data are normally distributed



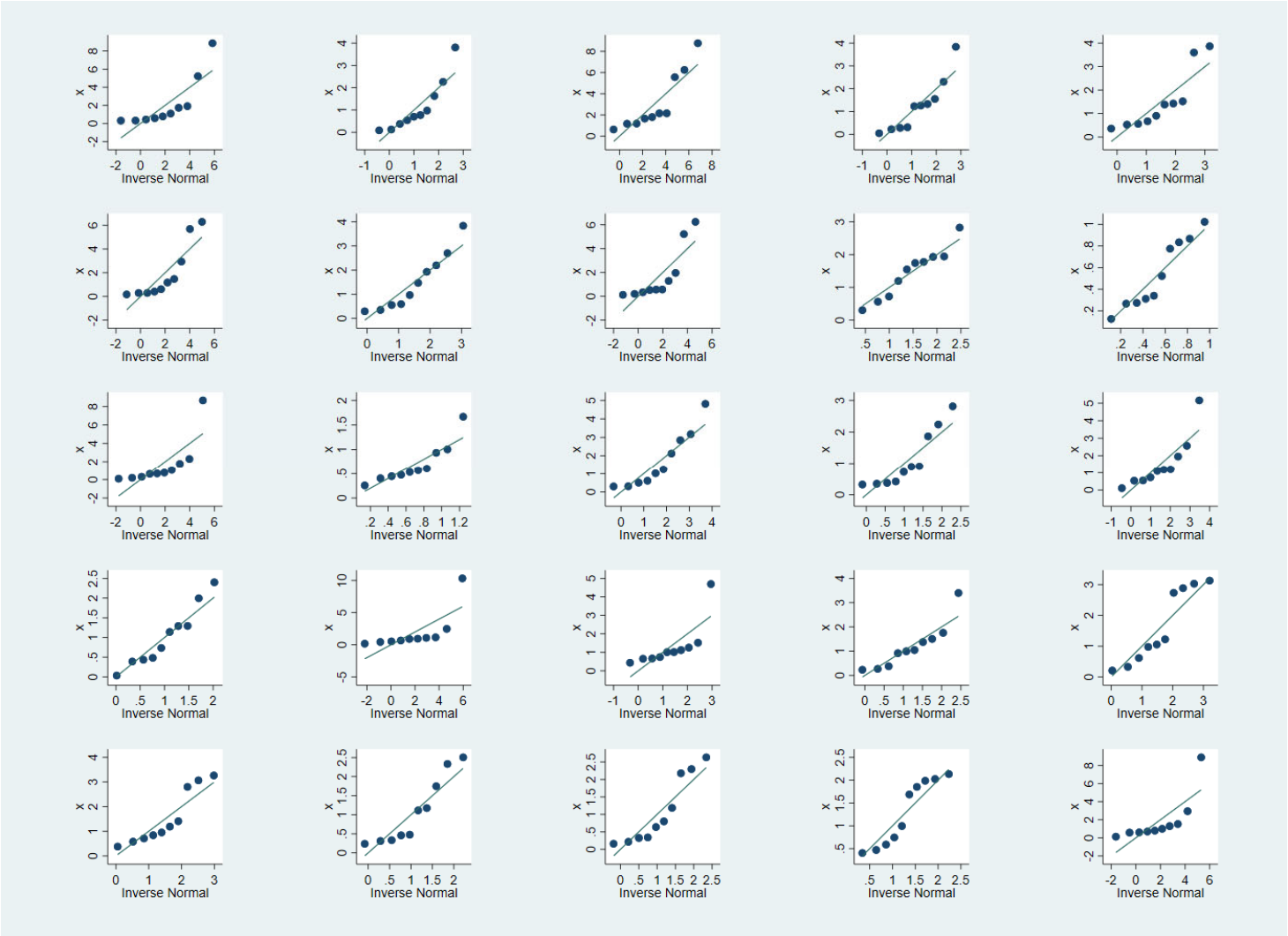
GOOD!



BAD!

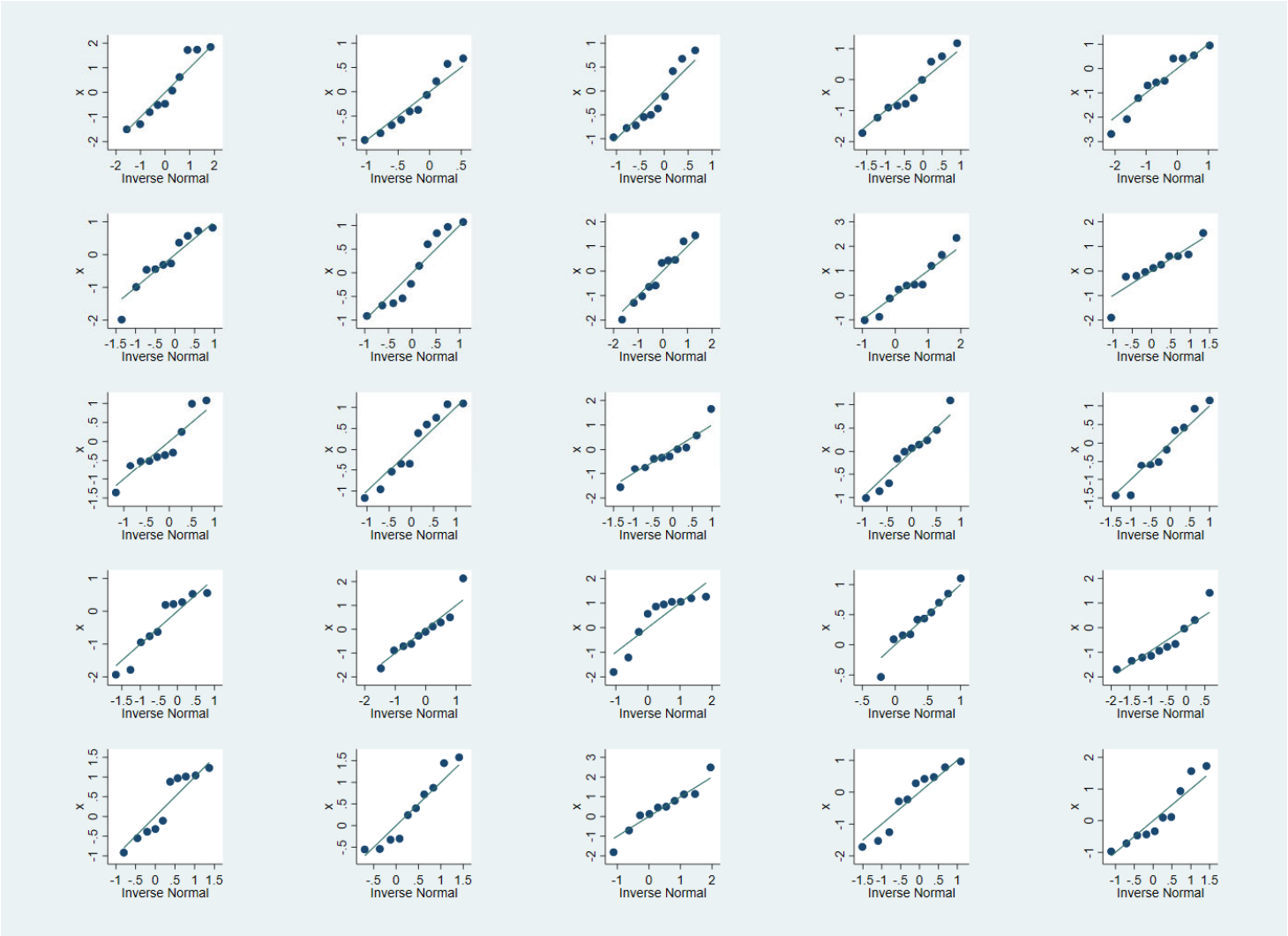
Evaluating Q-Q-plots (I)

- Normal?



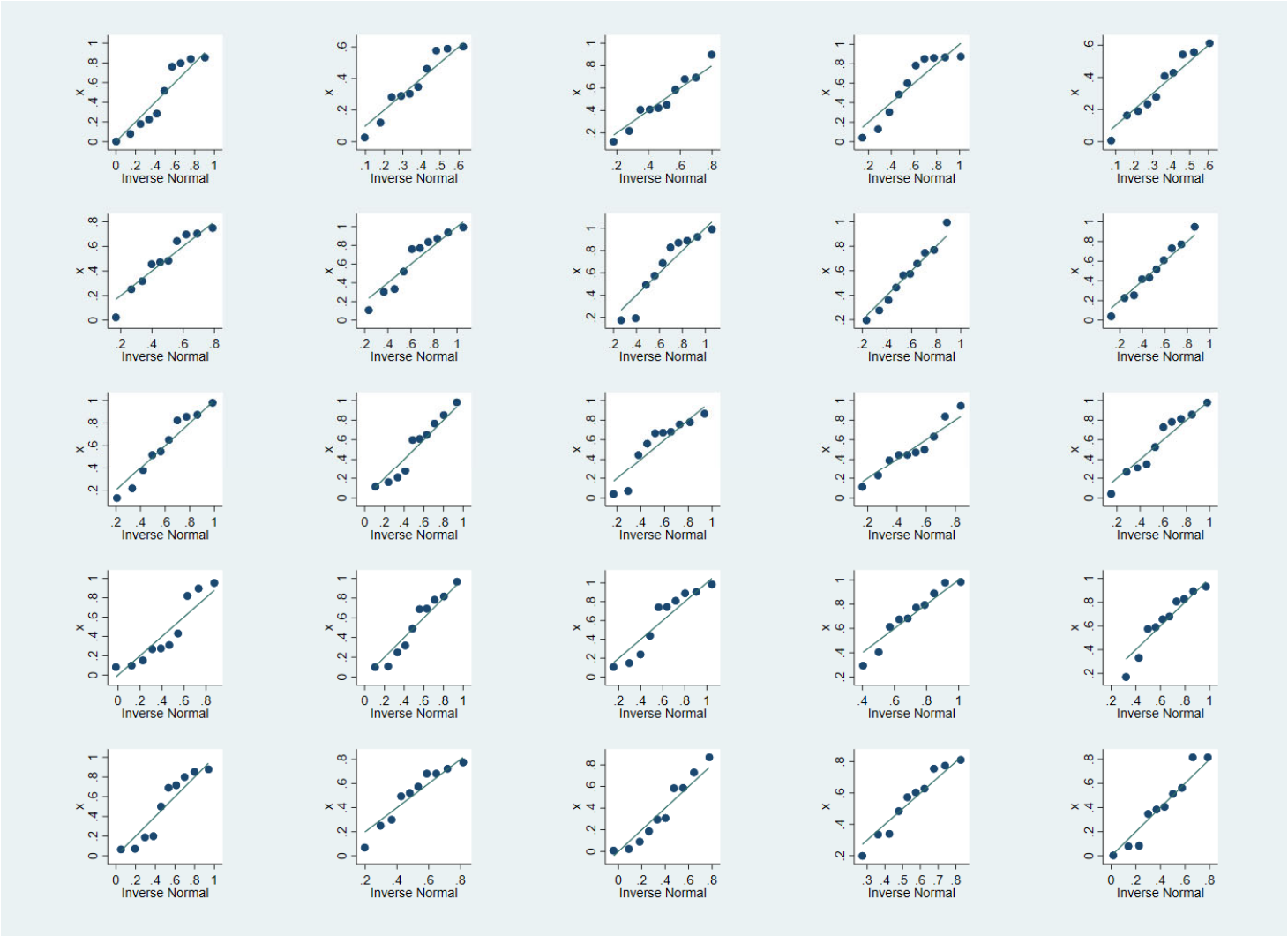
Evaluating Q-Q-plots (II)

- Normal?



Evaluating Q-Q-plots (III)

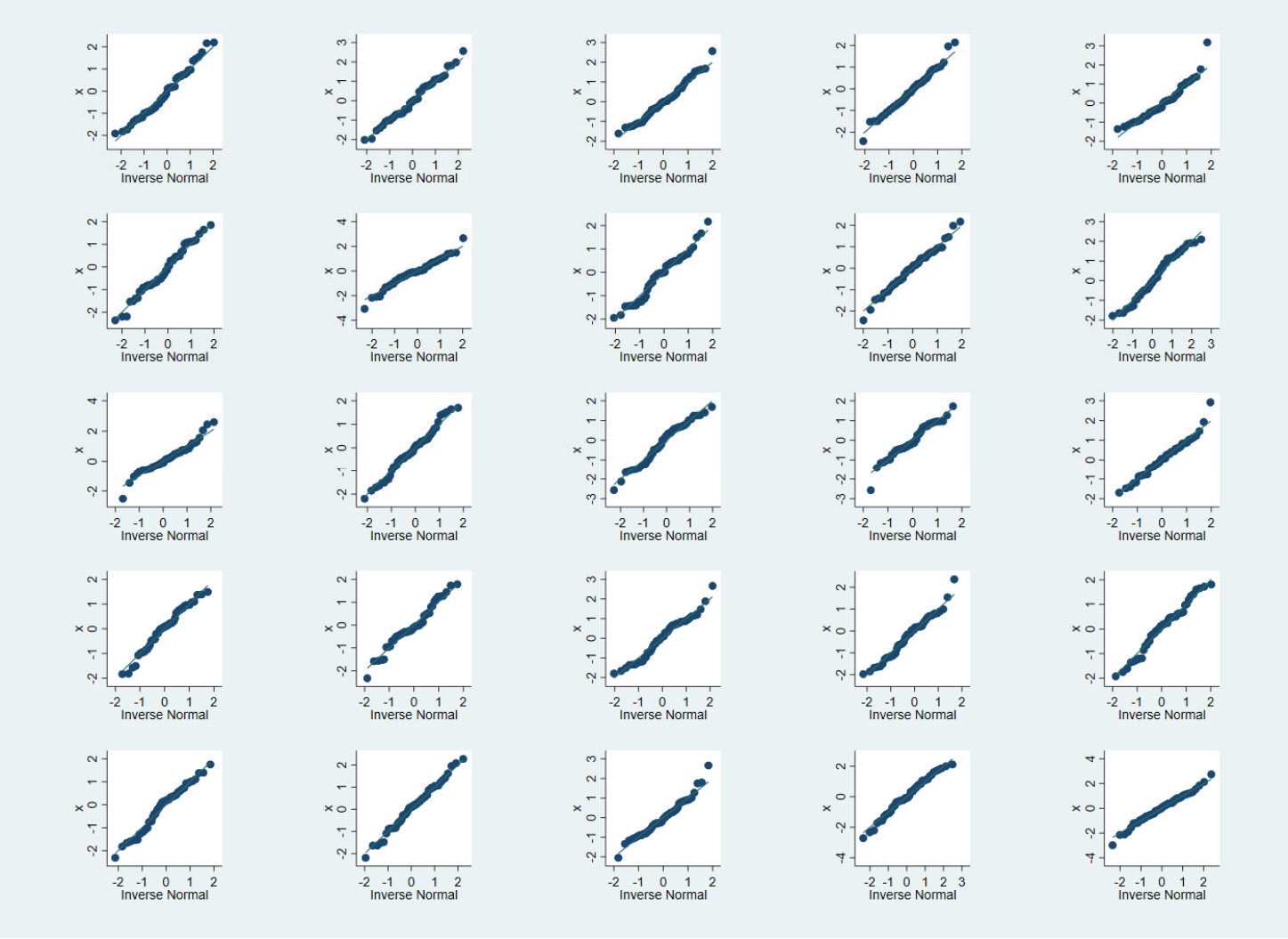
- Normal?



Some guidance

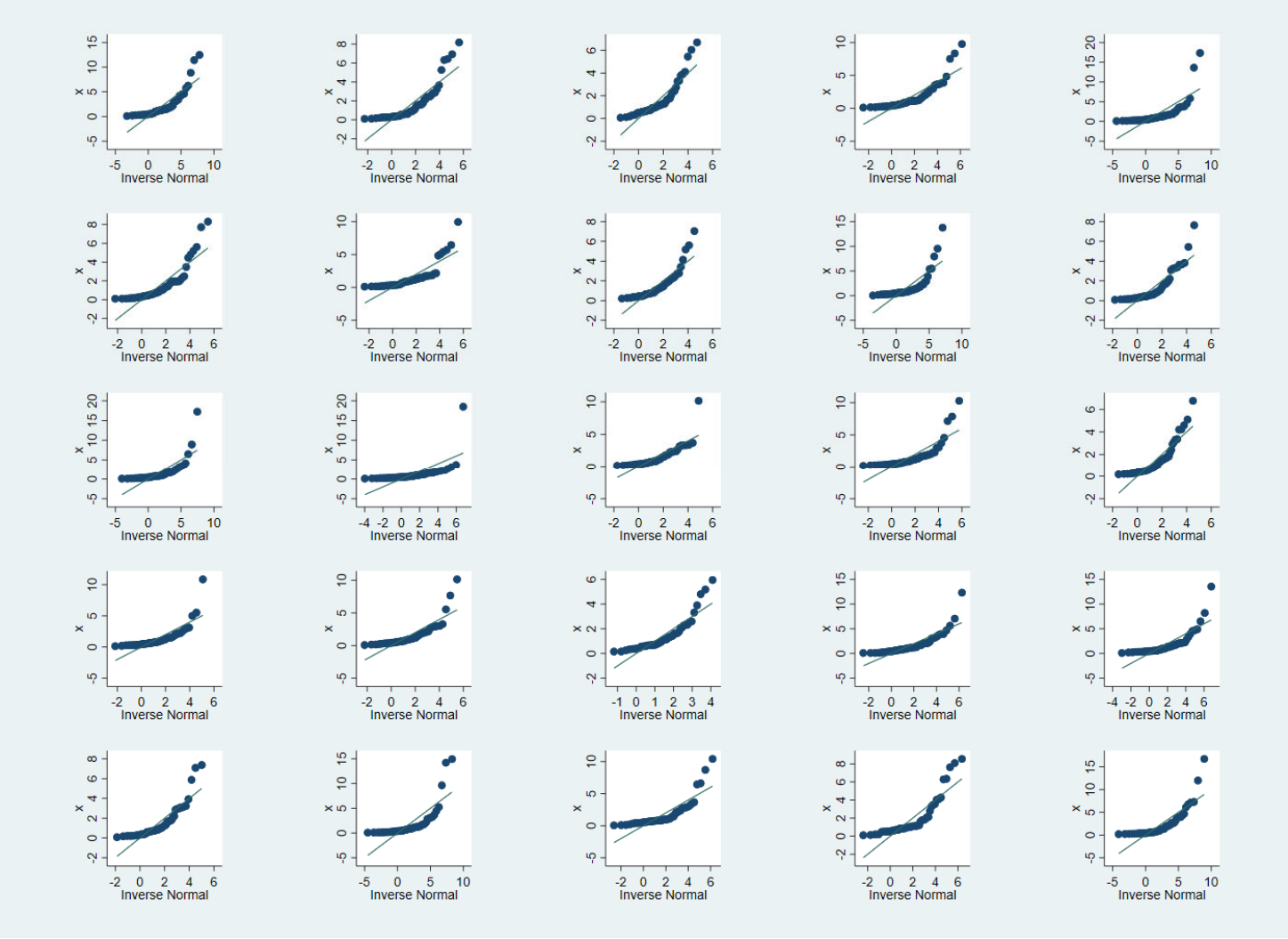
- Look for systematic deviations from linearity
- OK with single points deviating at the tails
- OK with criss-crossing over the line in the middle part
- S or L shapes typically indicates non-normality
- Log-normal is common and resembles an L lying down with the short edge pointing up
- Very difficult in small samples to determine that data are not following a normal distribution

Q-Q-plots for normally distributed data, n = 50

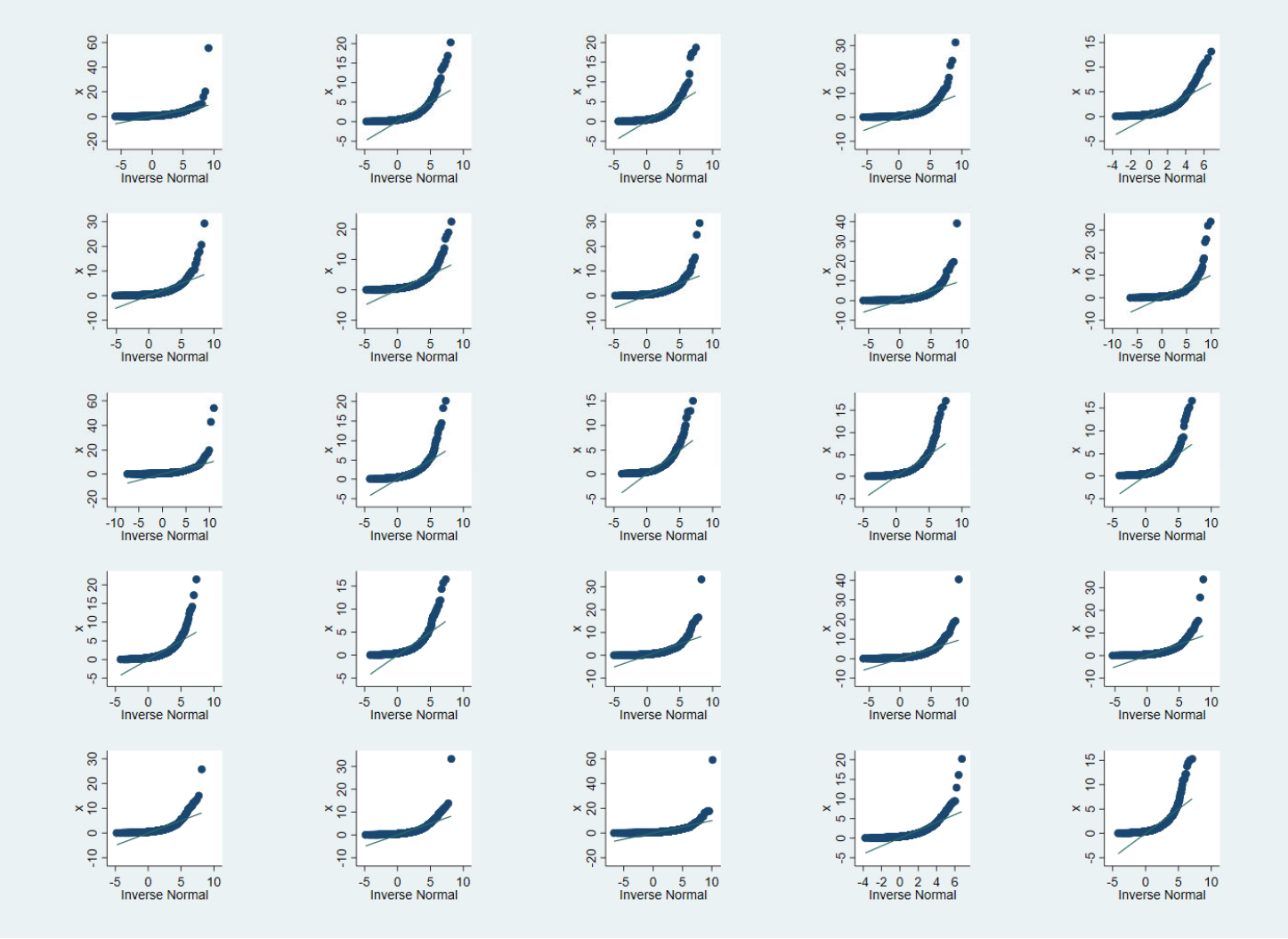


Q-Q-plots for normally distributed data, n = 1,000

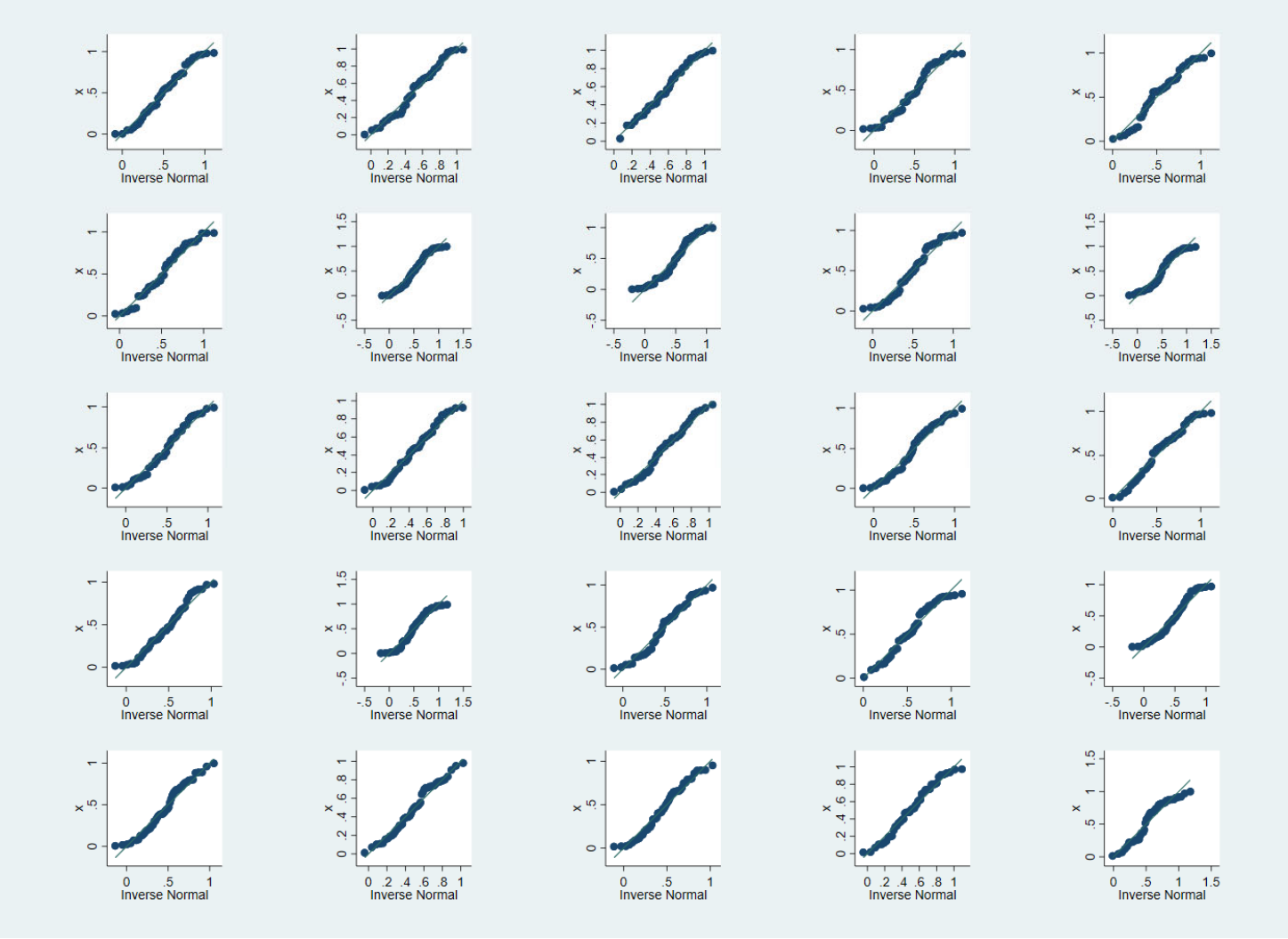
Q-Q-plots for log-normally distributed data, n = 50



Q-Q-plots, log-normally distributed data, n = 1,000



Q-Q-plots for uniformly distributed data, n = 50

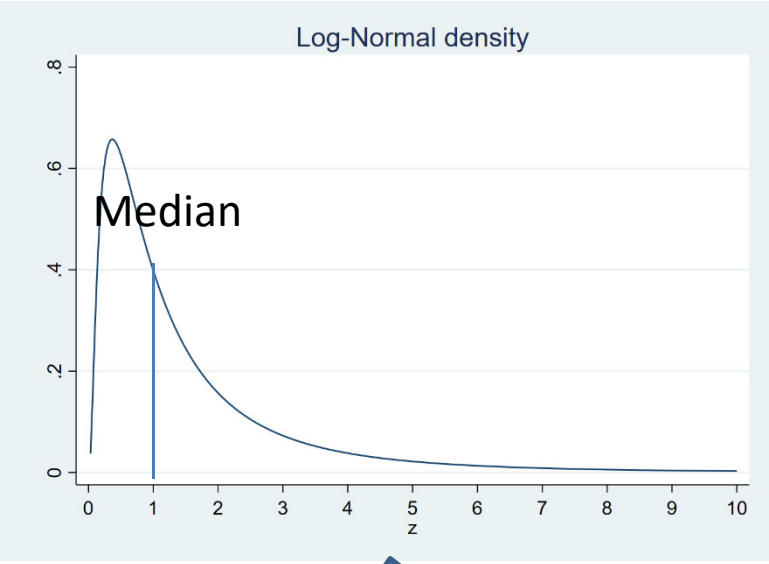


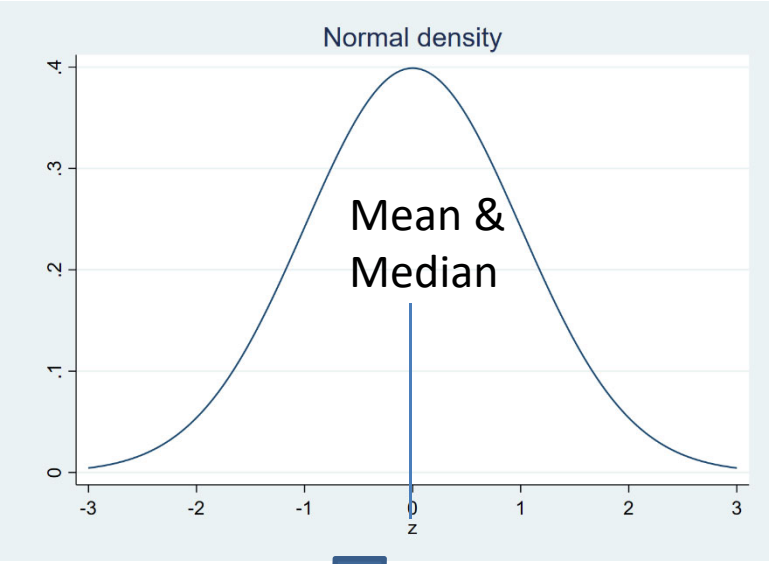
Q-Q-plots, uniformly distributed data, n = 1,000

Relation between Log-normal and normal distribution

Examples:
$\exp(2) = 7.39$
$\exp(0) = 1$
$\exp(-1) = 0.37$

$\log(Y)$





Data scale
Y

$\exp(Mean) \rightarrow Median$

Analysis scale
$\log(Y)$

Interpretation of linear regression with log(Y) outcome

- Assume we have transformed Y to estimate the following

$$\log(Y) = \beta_0 + \beta_1 \cdot x_1 + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

- Using highschool math with exponential on both sides:

$$Y = \exp \beta_0 \cdot \exp(\beta_1)^x \cdot \exp \varepsilon$$

- In other words, the model is **multiplicative**

Interpretation of coefficients

- Intercept:

  $\exp(\beta_0)$: Median of outcome for reference person

- Slope:

  $\exp(\beta_1)$: Factor for increase in median of outcome per unit in explanatory variable

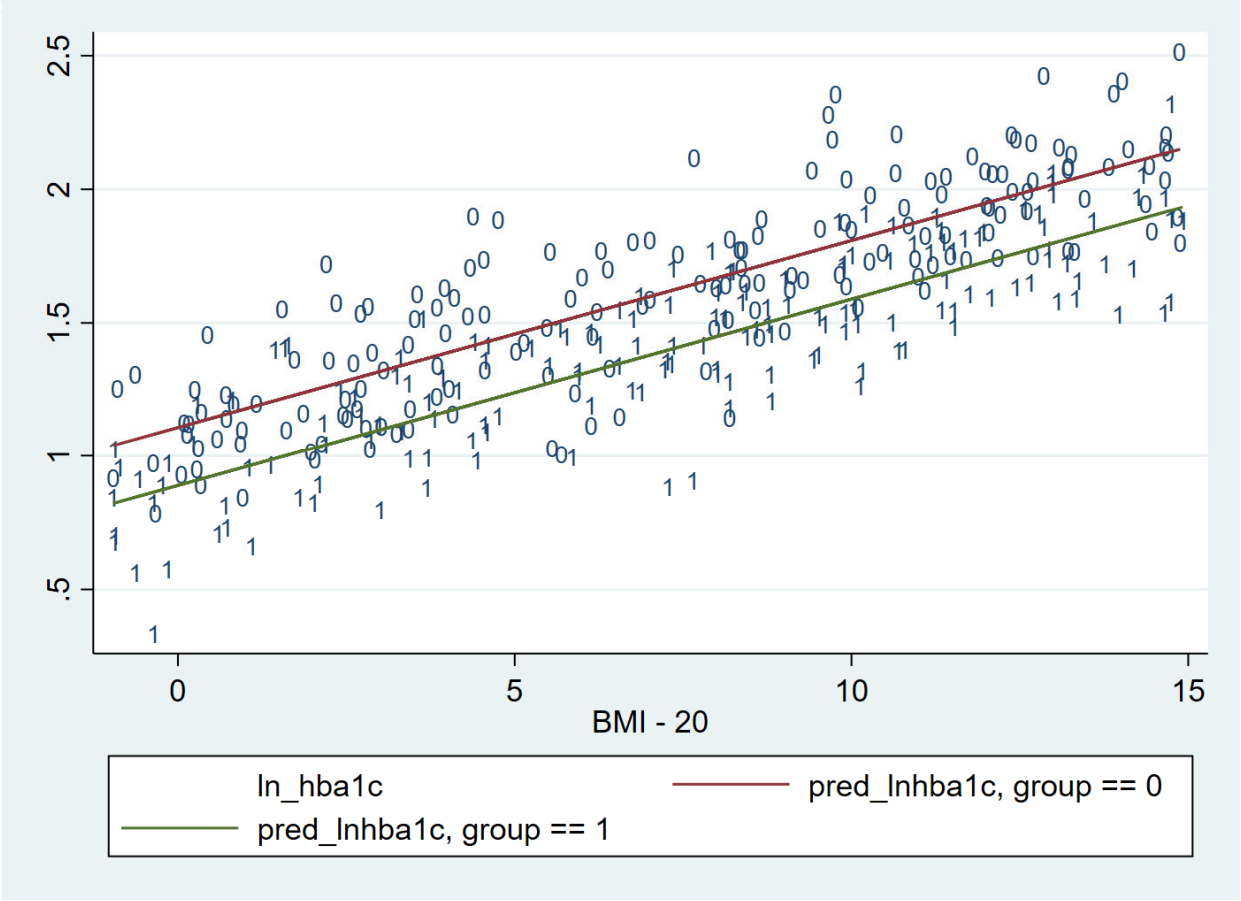
- Note: Non-linear model on original scale

Example output (log-scale)

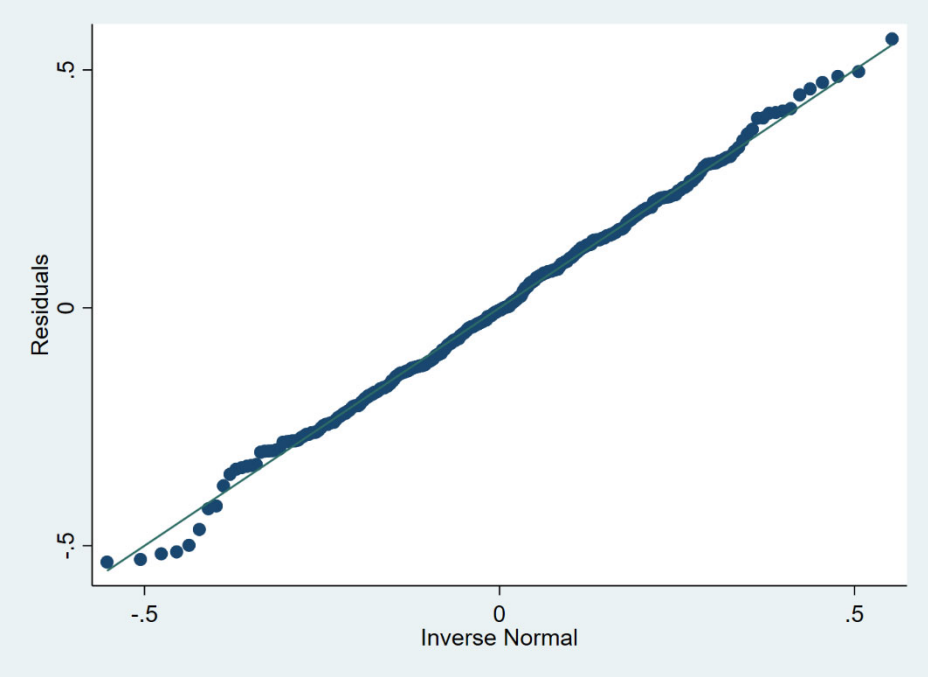
```
. regress ln_hba1c b0.group bmi

      Source |       SS           df       MS      Number of obs   =       342
-------------+----------------------------------   F(2, 339)       =    472.07
       Model |  38.1824433         2  19.0912216   Prob > F        =    0.0000
    Residual |  13.7098077       339   .04044191   R-squared       =    0.7358
-------------+----------------------------------   Adj R-squared   =    0.7342
       Total |  51.8922509       341  .152176689   Root MSE        =     .2011

------------------------------------------------------------------------------
    ln_hba1c | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     1.group |  -.2193906   .0217735   -10.08   0.000    -.2622187   -.1765625
       bmi20 |   .0700393   .0024052    29.12   0.000     .0653084    .0747702
       _cons |   1.108013    .022921    48.34   0.000     1.062928    1.153098
------------------------------------------------------------------------------
```

Example output (log-scale)

```
. regress ln_hba1c b0.group b

      Source |       SS
-------------+---------------
       Model |  38.1824433
    Residual |  13.7098077
-------------+---------------
       Total |  51.8922509

------------------------------
    ln_hba1c | Coefficient  S
-------------+----------------
     1.group |  -.2193906   .
       bmi20 |   .0700393   .
       _cons |   1.108013
------------------------------
```



Normality check (log-scale)



Example output (original scale)

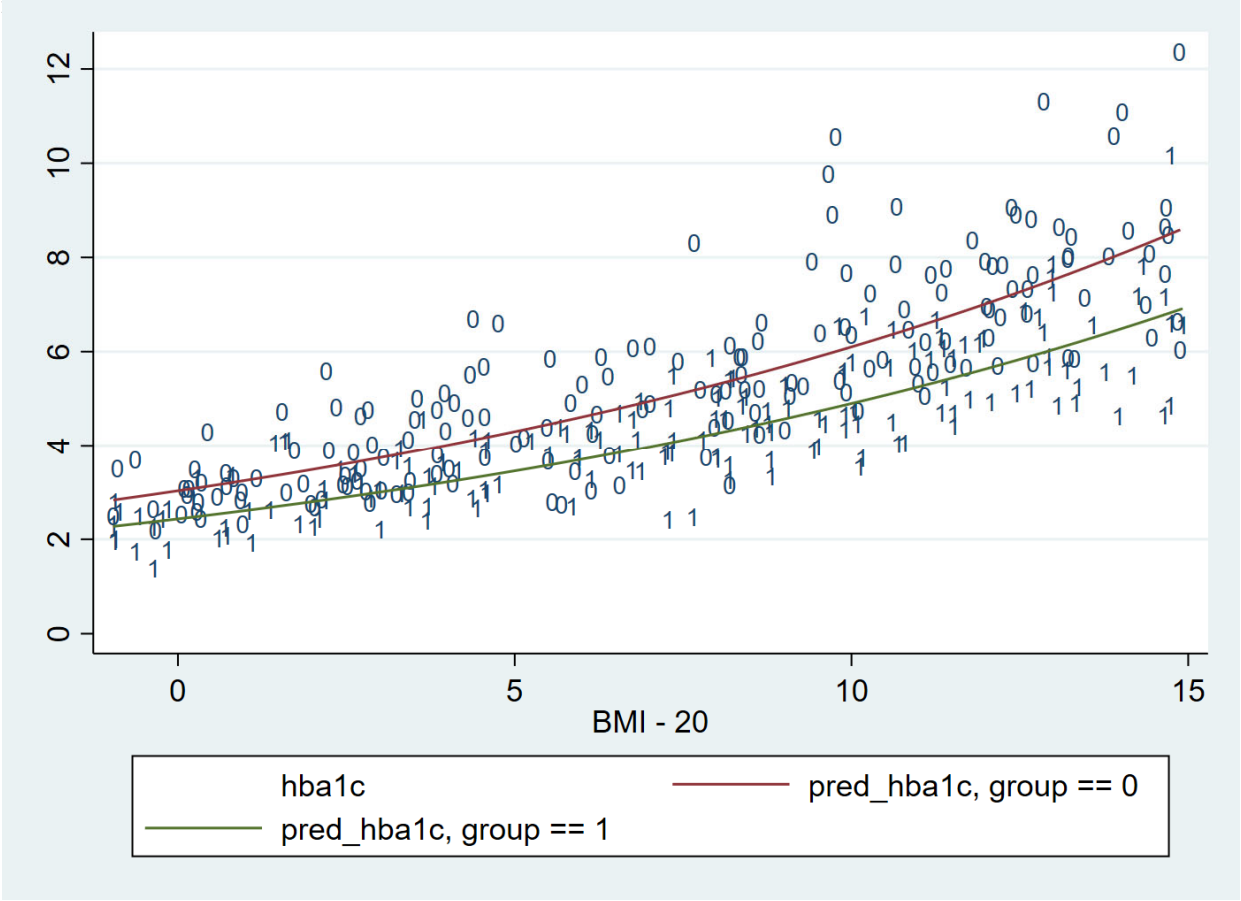
```
. regress ln_hba1c b0.group bmi, eform("MedRat")

      Source |       SS           df       MS      Number of obs   =       342
-------------+----------------------------------   F(2, 339)       =    472.07
       Model |  38.1824433         2  19.0912216   Prob > F        =    0.0000
    Residual |  13.7098077       339   .04044191   R-squared       =    0.7358
-------------+----------------------------------   Adj R-squared   =    0.7342
       Total |  51.8922509       341  .152176689   Root MSE        =     .2011

------------------------------------------------------------------------------
    ln_hba1c |     MedRat   Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     1.group |    .803008   .0174843   -10.08   0.000     .7693428    .8381464
       bmi20 |    1.07255   .0025797    29.12   0.000     1.067488    1.077637
       _cons |   3.028335   .0694126    48.34   0.000     2.894833    3.167993
------------------------------------------------------------------------------
```

Example output (original scale)

```
. regress ln_hba1c b0.group b
```

Source	SS
Model	38.1824433
Residual	13.7098077
Total	51.8922509

ln_hba1c	MedRat	S
1.group	.803008	.
bmi20	1.07255	.
_cons	3.028335	.



Thanks for your attention – questions welcome!

(Djursland, July 2015 – H Støvring)