



Linear Regression and Power Revisited



Håkon K. Gjessing

Professor/Principal Investigator

⇒ Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo
– a Norwegian Research Council Centre of Excellence

⇒ Department of Global Public Health and Primary Care, University of Bergen

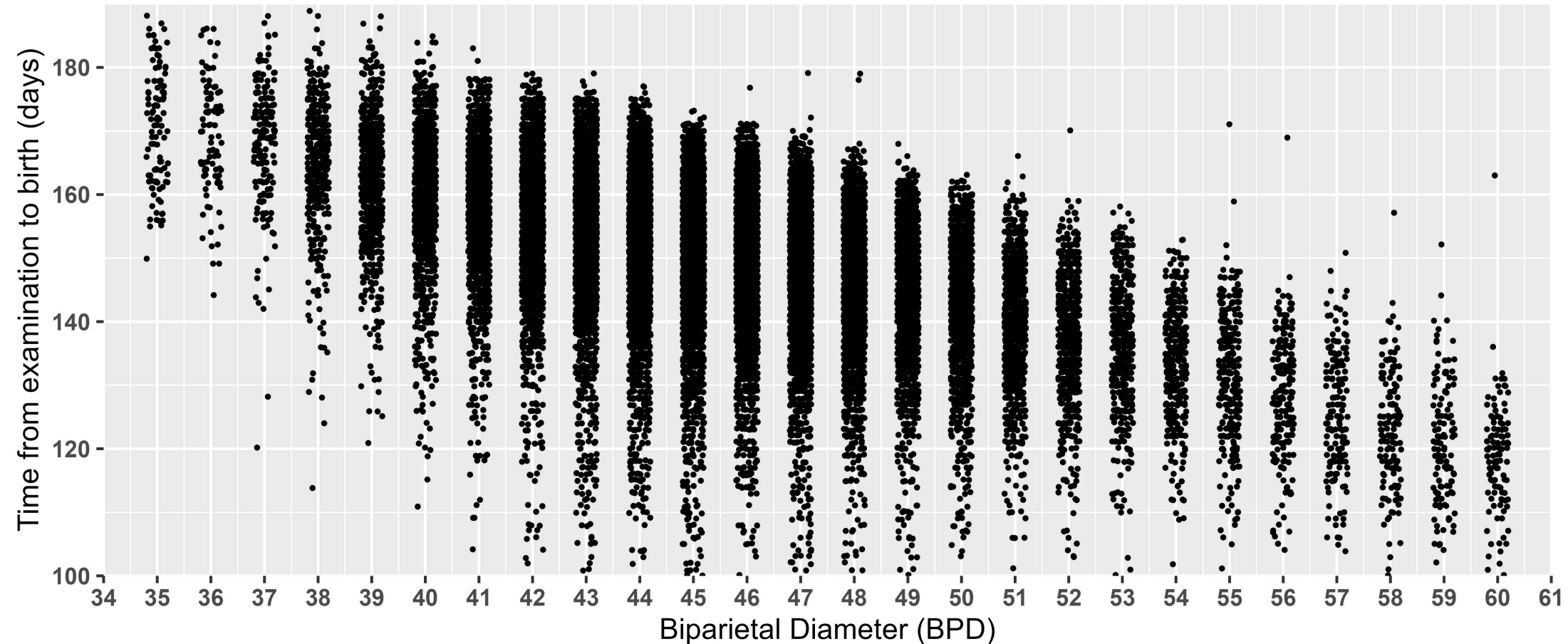


Kathmandu
Monday 5 October 2026



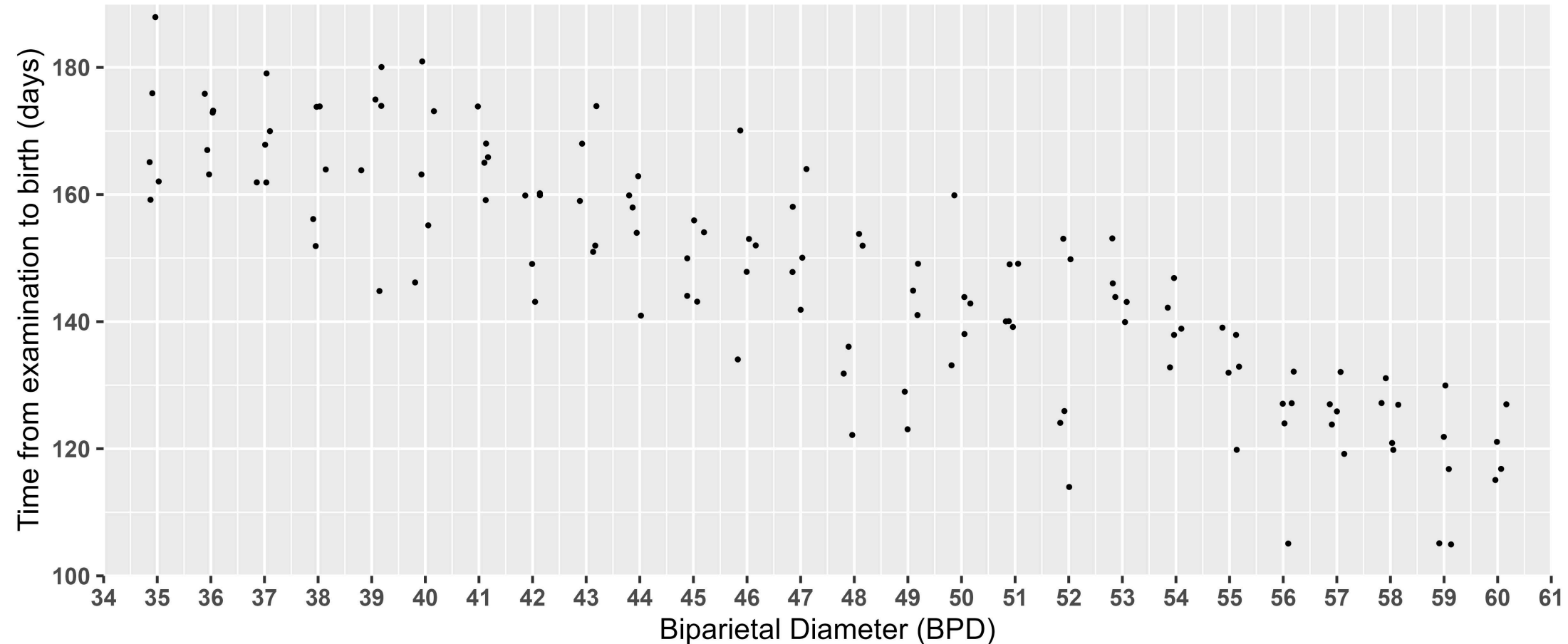
Ultrasound at 18 weeks of pregnancy

All data, 38490 data points in range 35-60 mm



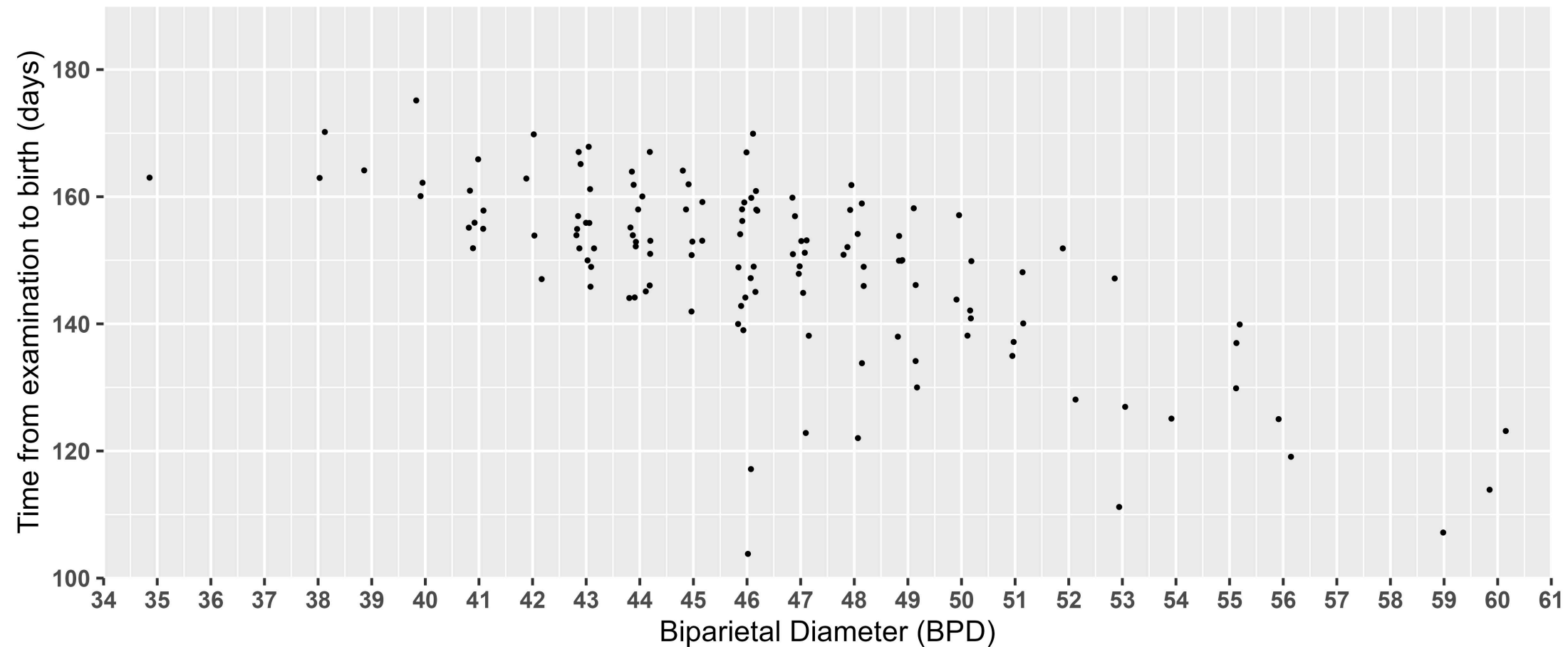
Ultrasound at 18 weeks of pregnancy

Uniform (even) sample, 129 data points in range 35-60 mm



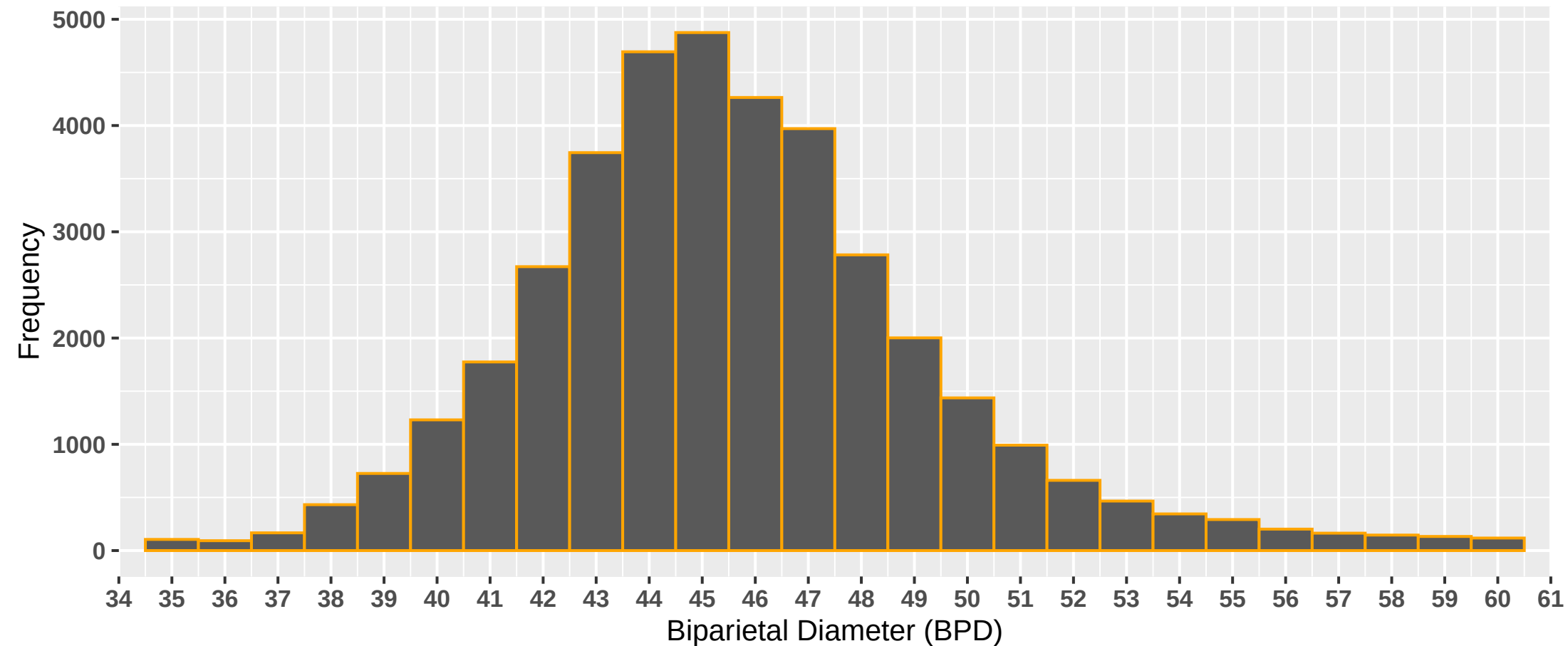
Ultrasound at 18 weeks of pregnancy

Random sample, 129 data points in range 35-60 mm



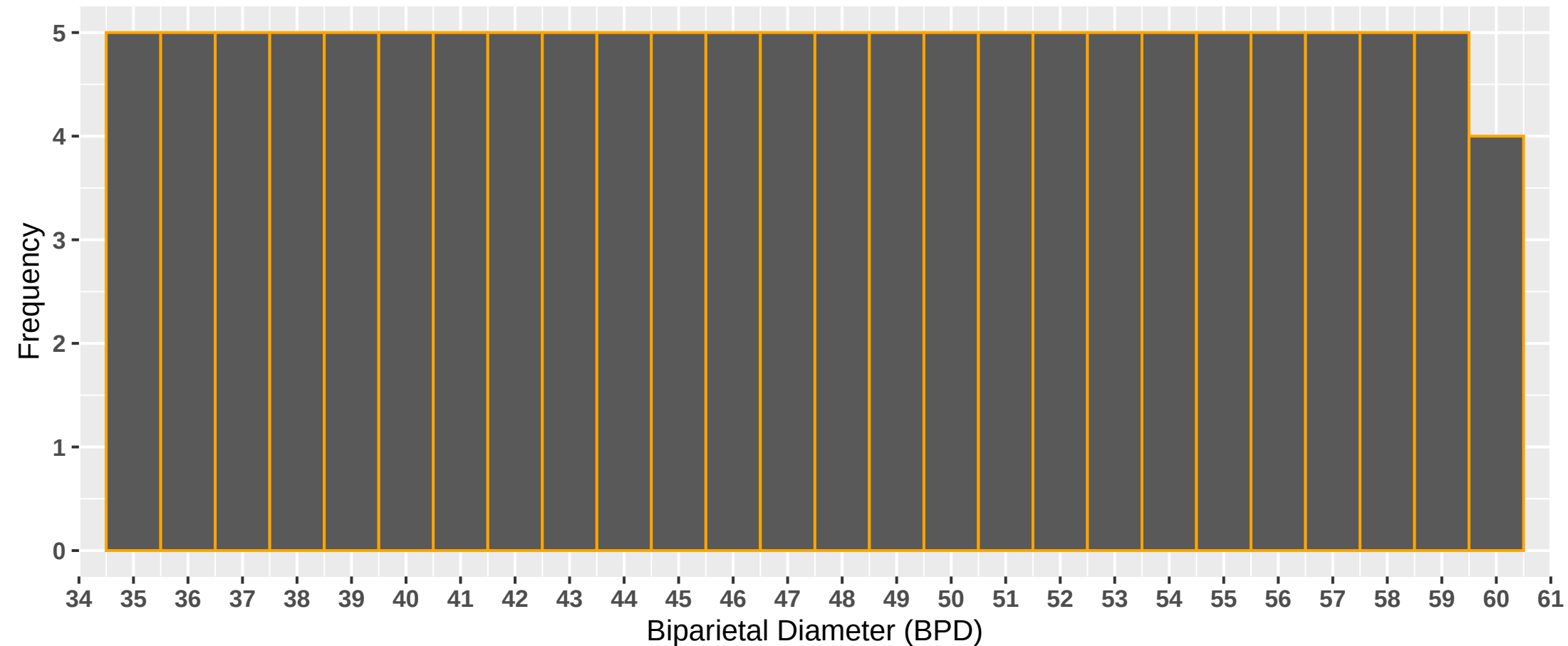
Ultrasound at 18 weeks of pregnancy

All data, 38490 data points in range 35-60 mm



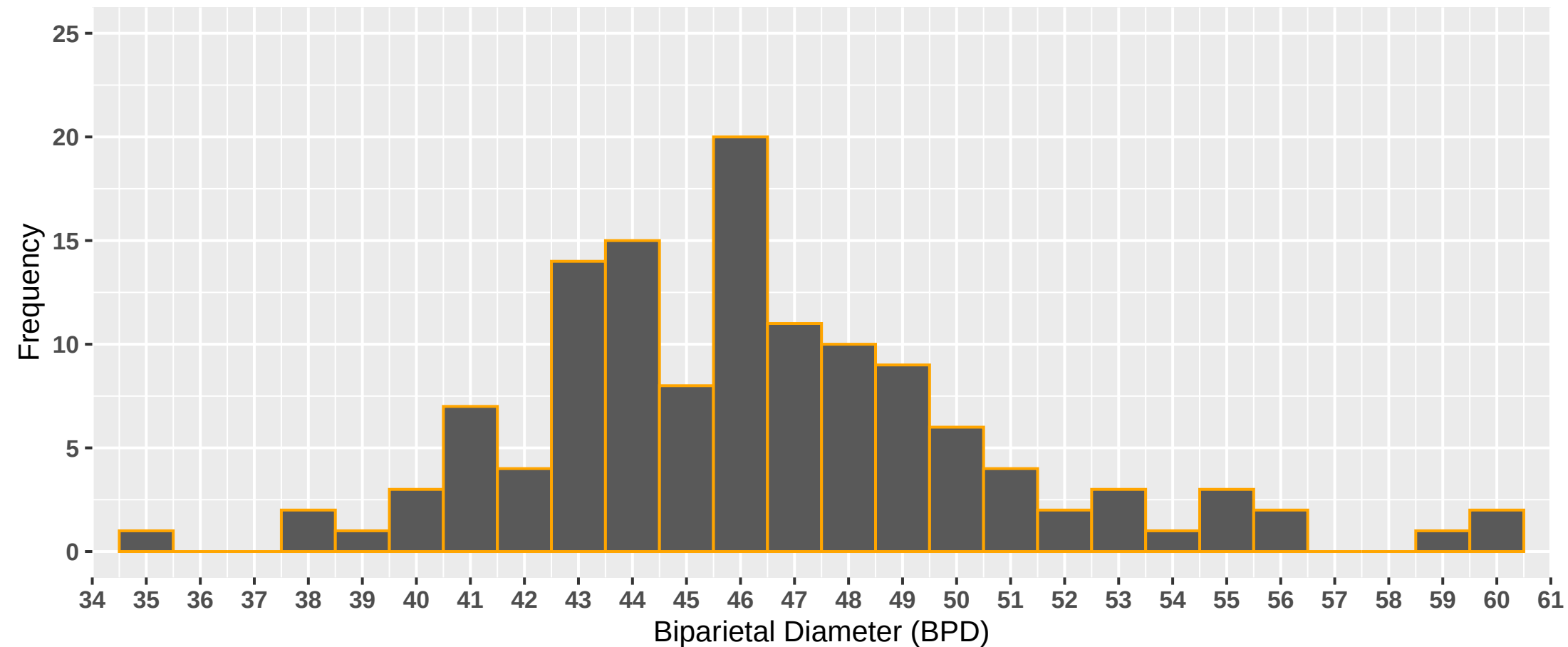
Ultrasound at 18 weeks of pregnancy

Uniform (even) sample, 129 data points in range 35-60 mm



Ultrasound at 18 weeks of pregnancy

Random sample, 129 data points in range 35-60 mm



Sample size: The easy (and best?) way

More general

Pilot study:

$$\text{SE}(\widehat{M}_1) = \frac{\gamma}{\sqrt{n_1}}$$

Your study:

$$\text{SE}(\widehat{M}_2) = \frac{\gamma}{\sqrt{n_2}}.$$

We have

$$\frac{n_2}{n_1} = \left(\frac{\text{SE}(\widehat{M}_1)}{\text{SE}(\widehat{M}_2)} \right)^2$$

γ is gone! (Good riddance!)

Sample size: The easy (and best?) way

More general

$$\frac{n_2}{n_1} = \left(\frac{\text{SE}(\widehat{M}_1)}{\text{SE}(\widehat{M}_2)} \right)^2$$

That is,

The ratio of the sample sizes is the square of the (inverse) ratio of precisions

or equivalently

The ratio of the sample sizes is the square of the (inverse) ratio of CI lengths

Regression results (uniform data sample):

Coefficients:				
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	247.9103	5.4915	45.14	<2e-16
bpd	-2.1278	0.1144	-18.59	<2e-16
Confidence intervals:				
	2.5 %	97.5 %		
(Intercept)	237.043633	258.777061		
bpd	-2.354252	-1.901327		

Regression results (full data):

Coefficients:				
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	236.21837	0.74393	317.5	<2e-16
bpd	-1.90307	0.01627	-117.0	<2e-16
Confidence intervals:				
	2.5 %	97.5 %		
(Intercept)	234.760243	237.676504		
bpd	-1.934961	-1.871181		

Trying the formula

$$\frac{n_2}{n_1} = \left(\frac{SE(\widehat{M}_1)}{SE(\widehat{M}_2)} \right)^2$$

$$\frac{n_2}{n_1} = \frac{38490}{129} = 298$$

but

$$\left(\frac{SE(\widehat{M}_1)}{SE(\widehat{M}_2)} \right)^2 = \left(\frac{0.1144}{0.01627} \right)^2 \approx 49$$

... which is off by a factor of about 6!

Three things combine to create this issue

Sample size: The easy (and best?) way

General formula for standard error:

$$\text{SE}(\widehat{M}_1) = \frac{\gamma}{\sqrt{n_1}}$$

With the formula

$$\frac{n_2}{n_1} = \left(\frac{\text{SE}(\widehat{M}_1)}{\text{SE}(\widehat{M}_2)} \right)^2$$

we don't really need to know γ .

However, this requires the two data sets to be sampled in the same manner.

A few examples of γ

Mean in a Normal distribution:

$$\gamma = \sigma$$

p in a binomial distribution:

$$\gamma = \sqrt{p(1-p)}$$

β_1 in a linear regression:

$$\gamma = \frac{\sigma_{\varepsilon}}{\sigma_X}$$

where σ_{ε} is the residual standard deviation (i.e. vertical variability around regression line),

and σ_X is the standard deviation of X (i.e. spread of data along X-axis).

NOTE the opportunity: Increasing σ_X when you sample gives a more precise estimate of β_1

What is γ for our three data sets?

	n	β_1	σ_ε	σ_X	$\gamma = \sigma_\varepsilon / \sigma_X$	$SE(\beta_1) = \gamma / \sqrt{n}$
Full data	38 490	-1.9	12.2	3.8	3.2	0.016
Uniform sample	129	-2.1	9.7	7.5	1.3	0.11
Random sample	129	-2.1	10.0	4.4	2.3	0.20

Under the standard assumptions for linear regression:

- β_1 is unbiased for all sampling on X
- σ_ε is unbiased for all sampling on X
- σ_X can be “manipulated” by appropriate sampling
- Spread samples give better precision for β_1 , i.e. lower $SE(\hat{\beta}_1)$

What is γ for our three data sets?

	n	β_1	σ_ε	σ_X	$\gamma = \sigma_\varepsilon / \sigma_X$	$SE(\beta_1) = \gamma / \sqrt{n}$
Full data	38 490	-1.9	12.2	3.8	3.2	0.016
Uniform sample	129	-2.1	9.7	7.5	1.3	0.11
Random sample	129	-2.1	10.0	4.4	2.3	0.20

Thus note that:

- For the random sample, everything should be the same as in the full data
...except n and consequently $SE(\beta_1)$
- BUT in our case we have (by chance) gotten a γ that's smaller in the random sample, $3.2/2.3 \approx 1.4$
... which means that the SE of 0.11 is somewhat optimistic
...which will (in this example) impact any sample size calculations by a factor of $1.4^2 \approx 2$
- Note that this again is our “little big problem” when we square something
- This happens even with a “pilot study” size of $n = 129$, and happens to all types of power/sample size calculations, not only this type.
- It thus explains a factor of 2 out of the factor of 6

What is γ for our three data sets?

	n	β_1	σ_ε	σ_X	$\gamma = \sigma_\varepsilon / \sigma_X$	$SE(\beta_1) = \gamma / \sqrt{n}$
Full data	38 490	-1.9	12.2	3.8	3.2	0.016
Uniform sample	129	-2.1	9.7	7.5	1.3	0.11
Random sample	129	-2.1	10.0	4.4	2.3	0.20

- On the good side, at least uniform sampling leads to better precision (0.11 vs. 0.20)
- And for the same reason,

$$(2.3/1.3)^2 = 1.77^2 = 3.1$$

explains the remaining factor of 3.

Summary & Considerations

- The sample size formula is extremely convenient
- But it – and most other sample size calculations – require **a *fixed*** sampling strategy

-
- By spreading the values along the X-axis one can increase σ_X and thus decrease $SE(\beta_1)$
 - This can improve precision considerably

One caveat:

- In our example we want optimal prediction in the ***mid range*** of the data since that's where most women had their ultrasound examination (around week 18 in these data)
- Oversampling at the two ends of the scale and assuming a pure, linear regression may reduce prediction quality in the middle

And one warning:

- Sampling on Y (rather than X) is quite dangerous
- Should only be done in very special cases, such as case-control studies