International Journal of Research and Development in Engineering Sciences
ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/
Regular Issue, Vol. 8, Issue. 6 (November – December) , 2026 , Pages: 68 – 77
Received: 01 March 2026 ; *Accepted:* 30 October 2026 ; *Published:* 17 November 2026
https://doi.org/10.63328/IJRDES-V8RI6P7
https://www.youtube.com/watch?v=Ij4jyECAb-o
https://archive.org/details/jsp-ijrdes-v8ri6p7
Original Paper : Applied Research – Experimental Research Paper

Learning-Curve Extrapolation Under Limited Data: Bias and Uncertainty in Predicting Training-Set Requirements

Uma Maheswari Krishnamoorthy

Department of Computer Science, University of Illinois at Springfield, Illinois,, USA

* *Corresponding Author : Uma Maheswari Krishnamoorthy ; umakrishnamoorthy27@gmail.com*

Abstract: A common way to plan data collection is to train a model on a few small subsets, fit a power law to the learning curve, and extrapolate to the size at which a target error is reached. The forecast is quoted without an error bar, and the extrapolation is far outside the range of the data. We analyse this procedure. For a log-log least-squares fit, the standard deviation of the logarithm of the predicted size is $(\tau/\beta)\sqrt{(1/m + (\log n^* - x\bar{}\,)^2/S_{xx})}$, where τ is the relative noise of the pilot errors, β the slope, m the number of pilots and $x\bar{}$, S_{xx} the mean and spread of their logarithms; the formula matched simulation within 1% and gave intervals with coverage between 94.0% and 95.2%. With five pilots between 20 and 160 cases and 5% noise, a forecast of 10,000 cases has a 95% interval of a factor of 1.7. Spreading the pilots more widely, or placing half at each end, shrinks the uncertainty. The log-log fit is also biased when the true curve has an irreducible error: because the curve flattens, the extrapolation is optimistic, and with an asymptote of 0.03 it predicted 38% of the true size for a target of 1,280 cases and 13% for 5,120. A three-parameter fit removes the bias but failed in 27% to 40% of the cases. The test set contributes noise of $\sqrt{((1 - e)/(eN))}$, which is 25% for a 3% error on 500 test cases. On the digits and breast cancer datasets the median forecast was between 43% and 112% of the true size. Synthetic results are simulations; the real data are three small learning curves.

Keywords: Learning Curves, Sample-Size Planning, Power Law, Extrapolation, Delta Method, Pilot Study.

1. Introduction

How much data is enough? Before collecting or labelling more, a team wants to know how many cases are needed to reach a target error. The usual answer is empirical: train on a few small subsets, measure the error, fit a power law to the learning curve, and read off the size at which the fitted curve reaches the target [1, 2, 7, 8]. The same method underlies scaling-law forecasts for large models [3, 4].

The forecast is an extrapolation by one or two orders of magnitude from a handful of noisy points, and it is usually quoted as a single number. Two questions follow. How uncertain is it, as a function of the number and spread of the pilot sizes and of the noise in the measured errors? And is it biased when the true curve flattens at an irreducible error, as real learning curves do [6, 9]? We answer both in a model where everything is explicit, and then test the answers on three real learning curves.

We make five contributions.

- A closed-form variance for the predicted size from a log-log fit, which gives an interval and shows how the uncertainty depends on the pilot design (Section 4).
- A result on pilot design: the uncertainty depends on how far the target is from the pilots relative to their spread, and a two-point design halves it.
- A bias result: if the curve has an asymptote, the log-log extrapolation is optimistic at every target, with measured size.
- The noise that a finite test set adds to every pilot, and the test-set size needed to make it small.
- A real-data check with three learning curves of logistic regression and nearest neighbours on the digits and breast cancer datasets.

All synthetic experiments are simulations from known curves. The real-data experiment uses small benchmark datasets, so the "true" curves are themselves estimates.

2. Background and Related Work

Learning curves. The error of a learner as a function of training size is called its learning curve, and a power law with an asymptote is its standard model [1, 9]. Cortes and colleagues studied the asymptotic value and rate of convergence [1], and Viering

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P7

and Loog review the shapes that occur [9]. Perlich and colleagues compared models by their learning curves [6].

Sample-size prediction. Mukherjee and colleagues fitted learning curves to estimate dataset size requirements for microarray classifiers and gave confidence bands by bootstrap [2]. Figueroa and colleagues and Beleites and colleagues used similar fits for clinical and spectroscopy data [7, 8]. For deep learning, scaling laws are fitted and extrapolated over many orders of magnitude [3, 4, 5, 17], and Domhan and colleagues extrapolate learning curves in training time to stop poor runs early [10].

What we add. The fit is a regression and its prediction variance is standard [13]; the inverse prediction of a size at a given error is a delta-method problem [11, 12]. We contribute the closed form for this problem, the consequences for design, the quantified optimism when an asymptote exists, and the role of the test set.

3. Research Methodology and Gap

3.1. Learning Curves and Pilots

Let $e(n)$ be the expected test error of a model trained on n cases. We use the form $e(n) = c + a\, n^{-b}$, with irreducible error $c \geq 0$. We run pilots at sizes $n_1, \ldots, n_m$ and measure errors $\hat{e}_i$. The relative noise is τ: $\log \hat{e}_i = \log e(n_i) + \eta_i$ with η_i independent $N(0, \tau^2)$. The goal is the target size n^* at which $e(n^*) = \varepsilon$.

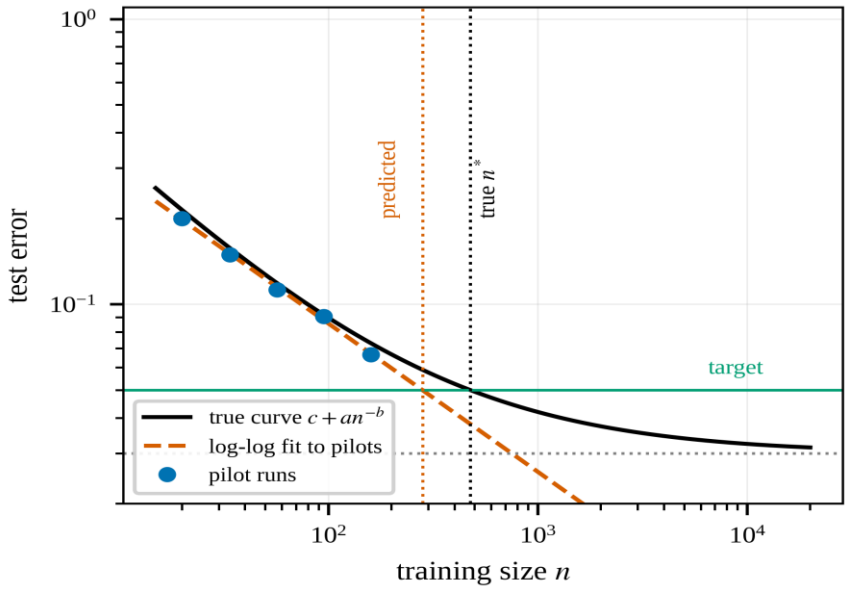


Figure. 1 A learning curve with an asymptote ($c = 0.03$) and a log-log extrapolation from five pilot runs (20 to 160 cases). The fitted line keeps falling, so it reaches the target error at a size well below the true one.

3.2. Estimators

Log-log fit (2 parameters): ordinary least squares of $\log \hat{e}_i$ on $\log n_i$, giving $\log \hat{e} = \alpha\text{◎} - \beta\text{◎} \log n$. The prediction is

$$n\text{◡}^* = \exp(\,(\alpha\text{◎} - \log \varepsilon)\,/\,\beta\text{◎}\,). \qquad (1)$$

This assumes $c = 0$. **Three-parameter fit:** nonlinear least squares of $\hat{e}_i$ on $c + a\, n^{-b}$ with c between 0 and the smallest pilot error, b between 0.05 and 3; if the estimated asymptote is at or above the target, the fit is counted as failed.

4. Proposed System

Proposition 1 (uncertainty of the forecast). Suppose $e(n) = a\, n^{-\beta}$ ($c = 0$) and the pilot noise is as above. Let $x_i = \log n_i$, $x\text{◎}$ its mean, $S_{xx} = \Sigma(x_i - x\text{◎})^2$, and $x^* = \log n^*$. Then, to first order,

$$\mathrm{Var}(\log n\text{◡}^*) \approx (\tau^2/\beta^2)\,[\,1/m + (x^* - x\text{◡})^2 / S_{xx}\,]. \qquad (2)$$

Proof. The estimate is $\log n\text{◡}^* = (\alpha\text{◎} - \log \varepsilon)/\beta\text{◎}$. Its partial derivatives are $1/\beta$ with respect to $\alpha\text{◎}$ and $-x^*/\beta$ with respect to $\beta\text{◎}$. The OLS variances are $\tau^2(1/m + x\text{◡}^2/S_{xx})$ for $\alpha\text{◎}$ and τ^2/S_{xx} for $\beta\text{◎}$, and their covariance is $\tau^2 x\text{◡}/S_{xx}$. Combining them gives Eq. (2).

This is the prediction variance of a regression line at x^*, divided by β^2: the uncertainty grows with the noise, shrinks with a steeper curve, and increases with the distance of the target from the centre of the pilots, measured in units of their spread. An approximate 95% interval for n^* is

$$n\text{◡}^* \cdot \exp(\,\pm\, 1.96 \cdot \mathrm{sd}\,), \qquad (3)$$

with a t quantile with $m - 2$ degrees of freedom in place of 1.96 when τ is estimated.

Corollary 1 (design). *For m pilots over a range $\Delta = \log(n_{\max}/n_{\min})$, the spread is $S_{xx} = m\,\Delta^2(m+1)/(12(m-1))$ for geometrically spaced sizes and $S_{xx} = m\,\Delta^2/4$ when half the pilots are at each end.*

$$S_{xx} = m\,\Delta^2\,(m+1)\,/\,(12(m-1)) \text{ (geometric)}, \quad S_{xx} = m\,\Delta^2\,/\,4 \text{ (two-point)}. \qquad (4)$$

The two-point design has a larger S_{xx} by the factor $3(m-1)/(m+1)$, which is 2 for $m = 5$. It cannot detect curvature,

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P7

so it is useful only if the power law is trusted.

Proposition 2 (bias from an asymptote). If $c > 0$, the log-log slope of the true curve, $b\ a\ n^{-b}/(c + a\ n^{-b})$, decreases in magnitude as n grows. A line fitted over the pilot range has a steeper slope than the curve beyond it, so the two-parameter prediction satisfies $n^{\circledcirc *} < n^*$ for every target beyond the pilots, even without noise.

The proposition follows because the log-log curve is convex when $c > 0$: a chord over the pilot range lies above the curve outside it and therefore falls faster than the curve. The size of the bias depends on the asymptote and the extrapolation distance and is computed numerically in Section 6.3.

Proposition 3 (test-set noise). *A test set of N_{test} cases estimates an error e with relative standard deviation*

$$\tau_{test} = \sqrt{((1 - e) / (e\, N_{test}))}. \quad (5)$$

For the noise to be below τ, the test set must have at least $(1 - e)/(e\tau^2)$ cases. For small errors this is large: $e = 0.03$ and $\tau = 0.05$ need about 13,000 cases. If the same test set is used for every pilot, this noise is shared and does not average out over pilot repeats.

5. Experimental Setup and Implementation

The first study draws log-noise around a curve with $a = 2$, $b = 0.6$ and $c = 0$, with five geometrically spaced pilots from 20 to 160 cases, $\tau \in \{0.02, 0.05, 0.1\}$ and targets $n^* \in \{500, 2{,}000, 10{,}000, 50{,}000\}$, with 20,000 repetitions. The design study varies $m \in \{3, 4, 5, 8, 12\}$ and the range $n_{max}/n_{min} \in \{2, 4, 8, 16, 32\}$, with $n_{min} = 20$, $n^* = 5{,}000$ and 8,000 repetitions. The asymptote study uses $(c, a, b) \in \{(0.02, 1.2, 0.6), (0.03, 1.5, 0.7), (0.05, 1.5, 0.7)\}$, targets 320, 1,280 and 5,120, $\tau = 0.03$ and 1,000 repetitions. The test-noise check draws binomial errors with 20,000 repetitions.

Real data are the digits set (1,797 cases, ten classes) [15] and the Wisconsin breast cancer set (569 cases) [16], from scikit-learn [14]. A fixed random test set (500 and 169 cases) is held out, and the rest is the training pool. The "true" curve is the mean test error over 15 random training subsets of each size (one fit at the full pool). Models are logistic regression on standardised features, five-nearest-neighbours and, for breast cancer, logistic regression. The pilot procedure uses five geometric sizes (20 to 160 for digits, 10 to 80 for breast cancer), the mean of three repeats at each size, a log-log fit and the *t*-based interval of Section 4 with τ estimated from the residuals. It is repeated 150 times. The targets are the true errors at the three largest sizes below the pool size, so the true required sizes are known by construction.

6. Results and Analysis

6.1. How Uncertain Is The Forecast?

Table 1 and Figure 2 check Eq. (2). The standard deviation of log $n^{\circledcirc *}$ matches the simulation within 1% for $\tau \le 0.05$; at $\tau = 0.1$ it is lower than simulated by up to 5%, where the first-order approximation begins to fail. The 95% intervals of the form $n^{\circledcirc *} \exp(\pm 1.96\ sd)$ cover the true size 94.0% to 95.2% of the time. The uncertainty is larger than most plans allow for. With five pilots between 20 and 160 cases and $\tau = 0.05$, a forecast of 10,000 cases is uncertain by a factor of 1.7 in either direction (95%); a forecast of 2,000 by a factor of 1.4.

Table. 1 Forecast uncertainty (five pilots from 20 to 160, $b = 0.6$, $c = 0$): SD of log $n^{\circledcirc *}$ from Eq. (2) and simulation, coverage of the 95% interval, and its multiplicative half-width.

τ	n^*	Eq. (2)	Simulated	Coverage (%)	Factor (95%)
0.02	500	0.047	0.047	94.8	1.10
0.02	2,000	0.074	0.074	95.1	1.16
0.02	10,000	0.106	0.105	95.2	1.23
0.02	50,000	0.139	0.139	95.0	1.31
0.05	500	0.117	0.118	94.8	1.26
0.05	2,000	0.185	0.187	94.8	1.44
0.05	10,000	0.265	0.269	94.7	1.68
0.05	50,000	0.347	0.352	94.6	1.97
0.1	500	0.233	0.245	94.2	1.58
0.1	2,000	0.370	0.390	94.0	2.06
0.1	10,000	0.531	0.556	94.2	2.83
0.1	50,000	0.693	0.721	94.2	3.89

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P7

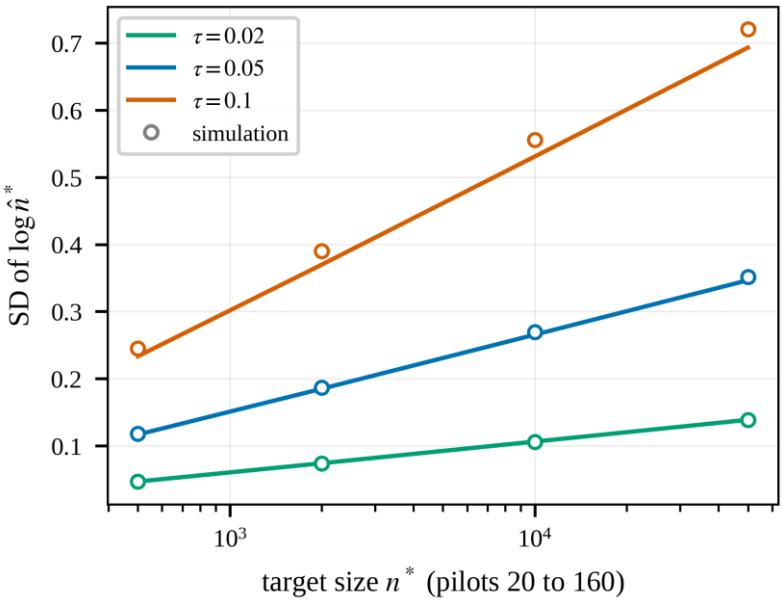


Figure. 2 Standard deviation of log n^* against the target size for three noise levels τ (five pilots from 20 to 160, true curve without asymptote). Lines: Eq. (2). Circles: simulation (20,000 repetitions).

6.2. Pilot Design

Figure 3 and Table 2 show the effect of the pilot design on the uncertainty of a forecast of 5,000 cases with τ = 0.05. Widening the pilot range from a factor of 2 to a factor of 32, with five pilots, lowers the standard deviation of log n^* from 0.88 to 0.12. Adding pilots helps less than widening them: at a range factor of 8, going from 3 to 12 pilots lowers it from 0.26 to 0.17, while at m = 5, going from a range factor of 4 to 16 lowers it from 0.37 to 0.16. The two-point design with half the pilots at each end gives 0.16 for m = 5 and a factor of 8, against 0.24 for geometric spacing, a reduction of 32%. The formula matches the simulation at all settings except the narrowest ranges, where it understates the noise by up to 17%.

Table. 2 SD of log n^* for n^* = 5,000, τ = 0.05: simulation (formula in brackets) by number of pilots and range.

m	Range 2	Range 4	Range 8	Range 16	Range 32
3	1.03 (0.88)	0.43 (0.41)	0.26 (0.26)	0.18 (0.18)	0.14 (0.14)
4	0.95 (0.84)	0.40 (0.39)	0.25 (0.24)	0.17 (0.17)	0.13 (0.13)
5	0.88 (0.79)	0.37 (0.37)	0.24 (0.23)	0.16 (0.16)	0.12 (0.12)
8	0.72 (0.67)	0.32 (0.31)	0.20 (0.20)	0.14 (0.14)	0.10 (0.10)
12	0.60 (0.57)	0.27 (0.27)	0.17 (0.17)	0.12 (0.12)	0.09 (0.09)

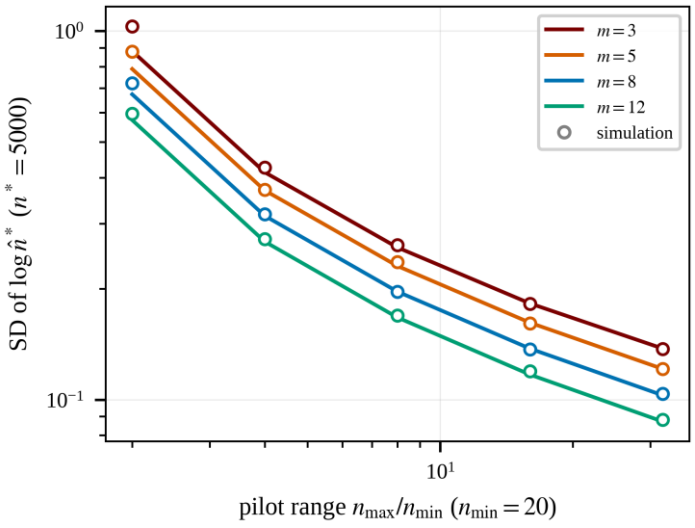


Figure. 3 Standard deviation of log n^* for target n^* = 5,000 against the pilot range n_{max}/n_{min}, for m pilot sizes (τ = 0.05). Lines: Eq. (2). Circles: simulation (8,000 repetitions).

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P7

6.3. The effect of an asymptote

Figure 4 and Table 3 show Proposition 2. With an irreducible error of 0.03, the noise-free two-parameter prediction for a target of 320 cases is 78% of the true size, for 1,280 cases it is 38%, and for 5,120 cases it is 13%. The fitted slope over the pilot range is 0.74 of the true exponent. With larger asymptotes the bias is worse (31% at 1,280 for c = 0.05). The median of the noisy estimates follows the noise-free curve. A forecast that ignores the asymptote therefore says that a target will be reached sooner than it will be, and the error grows with the extrapolation distance.

Table. 3 Predicted over true required size (%) with the asymptote ignored, and with a three-parameter fit (median, IQR, share of failed fits). Pilots 20 to 160, τ = 0.03.

c	n^*	Noise-free 2-par	Median 2-par	Median 3-par	IQR 3-par	Failed (%)
0.02	320	87	86	100	90–121	1
0.02	1,280	54	54	71	58–130	21
0.02	5,120	24	25	33	26–77	40
0.03	320	78	78	98	86–124	1
0.03	1,280	38	38	73	48–142	27
0.03	5,120	13	13	29	16–66	40
0.05	320	72	73	98	83–130	2
0.05	1,280	31	31	64	42–122	28
0.05	5,120	10	10	25	15–55	41

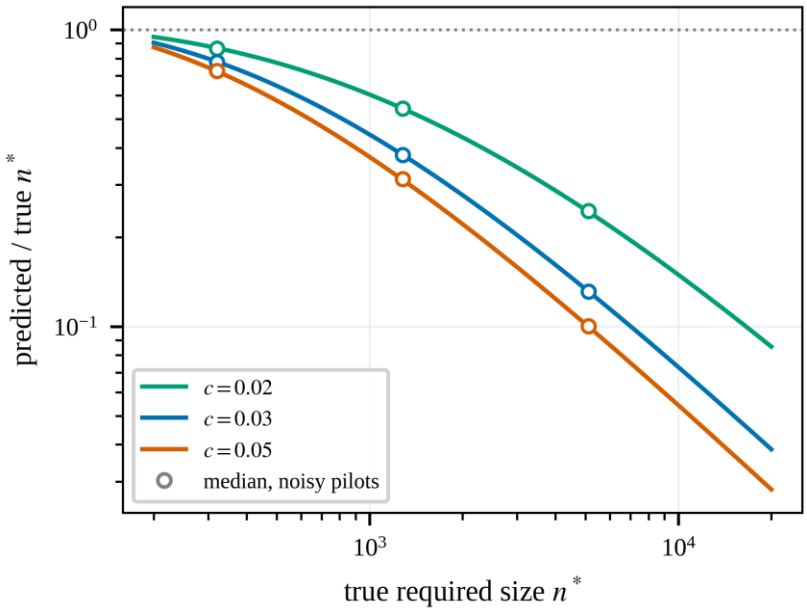


Figure. 4 Predicted over true required size when the curve has an asymptote c but the fit does not (five pilots from 20 to 160). Lines: noise-free prediction. Circles: median with τ = 0.03 noise (1,000 repetitions).

A three-parameter fit that includes the asymptote removes the bias at small extrapolation (median 98% at 320 cases) but is unreliable beyond it (Figure 5). At 1,280 cases the median is 73% with an interquartile range of 48% to 142%, and 27% of the fits fail because the estimated asymptote is above the target. At 5,120 cases the failure rate is 40%. Five noisy points do not identify three parameters.

6.4. Noise from the test set

Figure 6 and Table 4 give the noise that the test set adds to each pilot error, from Eq. (5). The formula agrees with simulation to the digits shown. For an error of 30% a test set of 500 cases gives relative noise 7%; for 10% it gives 13% and for 3% 25%.

Learning curves that reach low errors therefore need large test sets: for relative noise 0.05 the test set must have 933 cases at an error of 30%, 3,600 at 10% and 12,933 at 3%. This noise is shared across pilots if the same test set is used, and repeating pilots with fresh training subsets does not remove it.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P7

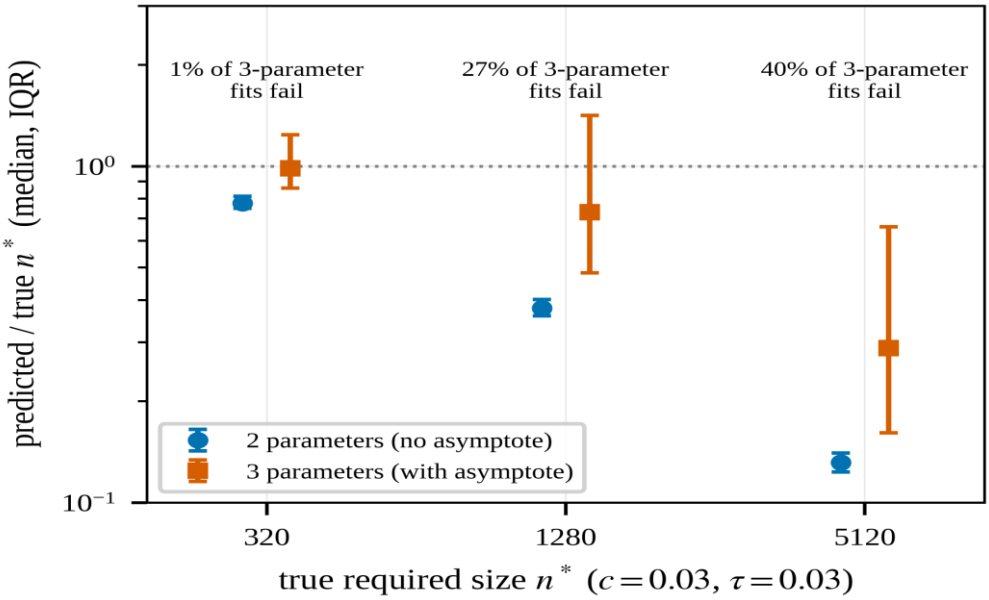


Figure. 5 Two-parameter against three-parameter fits for $c = 0.03$ and $\tau = 0.03$: median and interquartile range of predicted over true required size. The three-parameter fit fails when the estimated asymptote exceeds the target.

Table. 4 Relative noise τ_{test} of an error estimate (Eq. 5, simulation in brackets) and the test-set size needed for $\tau_{test} = 0.05$.

Error	$N = 100$	$N = 500$	$N = 2{,}000$	$N = 10{,}000$	N for 0.05
0.3	0.153 (0.153)	0.068 (0.068)	0.034 (0.034)	0.015 (0.015)	933
0.1	0.300 (0.299)	0.134 (0.133)	0.067 (0.067)	0.030 (0.030)	3,600
0.03	0.569 (0.570)	0.254 (0.254)	0.127 (0.127)	0.057 (0.057)	12,933

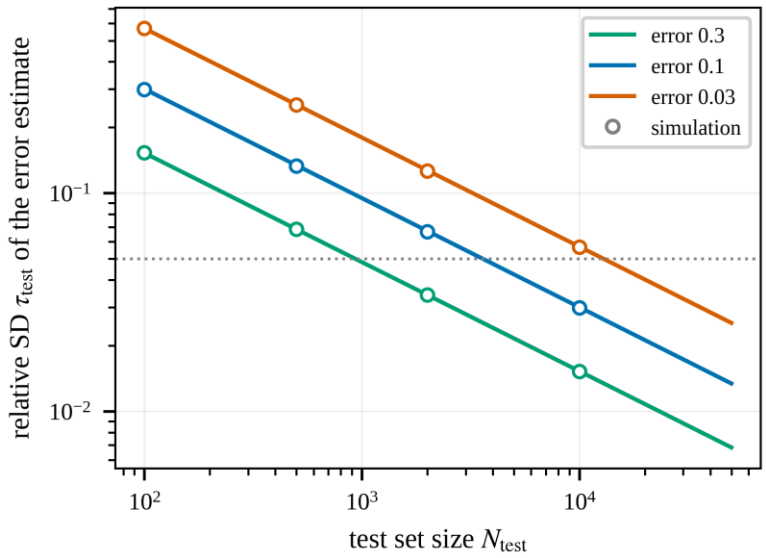


Figure. 6 Relative standard deviation of an error estimate from a test set of N_{test} cases. Lines: Eq. (5). Circles: simulation. Dotted: $\tau = 0.05$.

6.5. Real data

Table 5 and Figures 7 and 8 give the real-data results. The learning curves flatten (Figure 7): for logistic regression on digits the log-log slope is 0.87 over the pilot range and 0.52 beyond it, a decrease of 40%. The forecasts are optimistic, as Proposition 2 predicts. For logistic regression on digits the median forecast is 83%, 52% and 43% of the true size for targets of 320, 640 and 1,297 cases, and the 95% interval covers the true size 30% and 27% of the time for the two larger targets. For the five-nearest-neighbour model, whose curve is closer to a power law over this range, the medians are 91%, 83% and 88% and the coverage 86% to 93%. For breast cancer the medians are 81%, 58% and 112%, with a spread that is wide because the errors are small and the test set has 169 cases (relative noise about 44% at a 3% error).

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P7

Table. 5 Real data: true error at each target, median predicted over true required size (%), interquartile range and coverage (%) of the 95% interval (150 repetitions).

Model	n^*	Target Error (%)	Median (%)	IQR (%)	Coverage (%)
digits logistic	320	5.0	83	76–90	87
digits logistic	640	4.1	52	47–57	30
digits logistic	1297	2.6	43	38–48	27
digits 5-NN	320	5.9	91	83–98	92
digits 5-NN	640	3.5	83	75–92	86
digits 5-NN	1297	1.8	88	76–100	93
breast LR	160	3.1	81	61–118	89
breast LR	320	2.6	58	40–88	79
breast LR	400	1.8	112	67–205	96

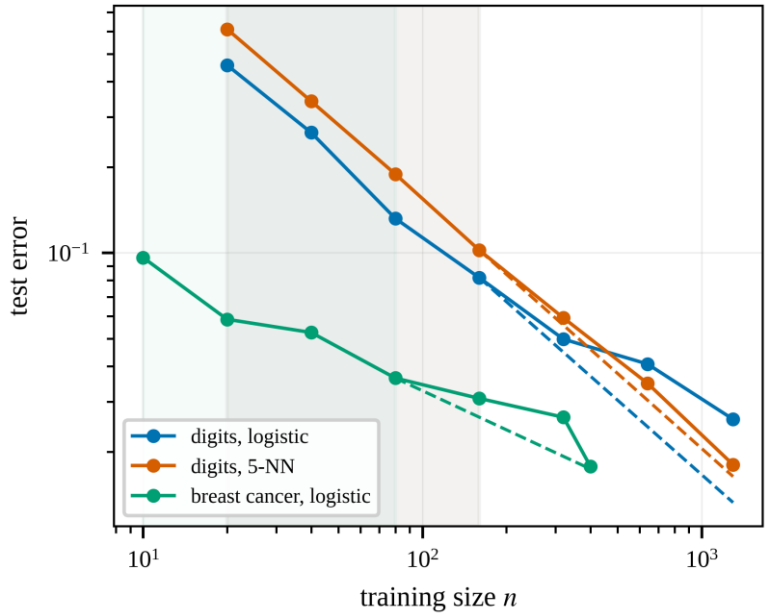


Figure. 7 Real learning curves (mean test error over 15 draws; one draw at the largest size). The shaded band is the pilot range; dashed lines extend the pilot slope from the last pilot size.

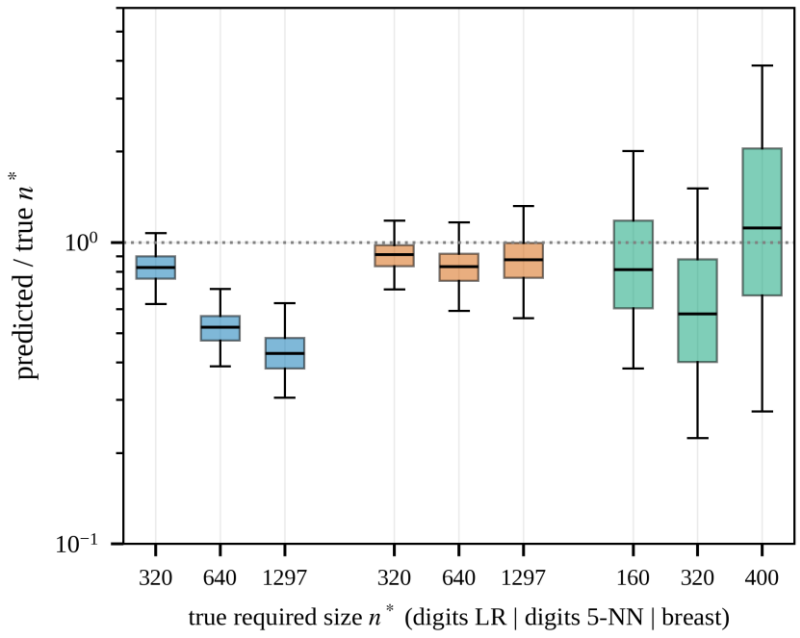


Figure. 8 Predicted over true required size on real data (150 repetitions of the pilot procedure). Boxes: interquartile range; line: median.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P7

6.6. Planning table

Table 6 evaluates Eq. (2) for a design of five pilots from 20 to 160 cases and exponent β = 0.6. A forecast 4 times the largest pilot is uncertain by a factor of 1.29 at τ = 0.05; at 256 times it is 1.93. These are lower bounds on the uncertainty because they ignore the bias of Section 6.3.

Table. 6 Multiplicative 95% half-width of the forecast (Eq. 2) for five pilots from 20 to 160 and β = 0.6, by pilot noise τ and target relative to the largest pilot.

τ	4×	16×	64×	256×
0.02	1.11	1.17	1.23	1.30
0.05	1.29	1.47	1.69	1.93
0.1	1.66	2.17	2.84	3.74
0.2	2.74	4.69	8.09	13.98

6.7. Can the asymptote be detected?

The recipe suggests comparing the log-log slope over the lower and upper halves of the pilot sizes. Figure 9 and Table 7 give the detection rate in simulation for a curve with $a = 1.5$, $b = 0.7$ and $\tau = 0.03$. With five pilots from 20 to 160 cases the slope falls from 0.57 to 0.47 when $c = 0.03$, but the difference is detected in only 50% of pilot sets, and in 25% when $c = 0.01$, close to the false-positive rate of 8% with no asymptote. The test is nominally 5% but the halves share the middle pilot, so the false-positive rate is higher. With seven pilots from 20 to 640 cases, detection is 94% at $c = 0.01$ and above 100% for $c \geq 0.02$. A wider range of pilots therefore does two things: it reduces the uncertainty of the forecast (Section 6.2) and it makes the asymptote visible.

Table. 7 Detection of an asymptote by the half-range slope comparison (%): nominal false-positive rate 5%.

c	5 pilots 20–160	7 pilots 20–640
0	8	9
0.01	25	94
0.02	40	100
0.03	50	100
0.05	62	100

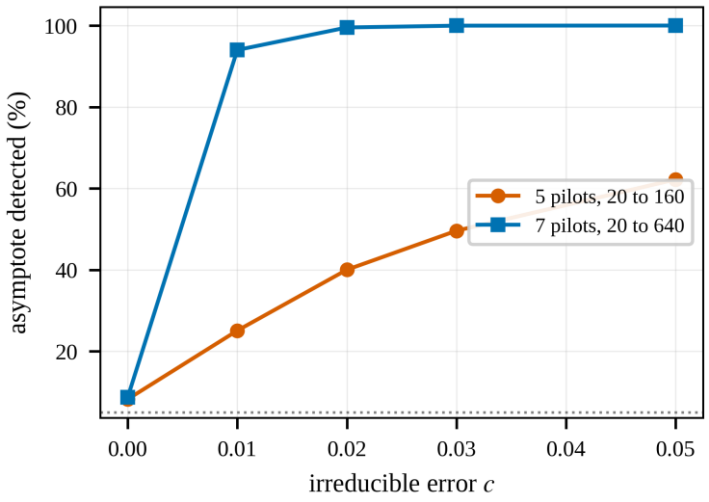


Figure. 9 Share of simulated pilot sets in which the slope over the upper half of the pilots is significantly smaller than over the lower half (one-sided, nominal 5%), by irreducible error ($a = 1.5$, $b = 0.7$, $\tau = 0.03$).

6.8. A Practical Recipe

The results suggest a short procedure. First, quote every forecast with an interval: use Eq. (2) with the residual noise of the fit, and expect factors of two to five at one to two orders of magnitude of extrapolation. Second, spread the pilots: widen the range before adding sizes, and use the endpoints if the power law is trusted. Third, size the test set for the error level: relative noise of 5% at a 3% error needs about 13,000 test cases. Fourth,

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P7

treat a two-parameter forecast as optimistic, and check by fitting the same curve on the lower and upper halves of the pilot sizes: if the slope falls, the asymptote matters. Fifth, if a three-parameter fit is used, report the share of bootstrap fits that fail, and do not extrapolate beyond about ten times the largest pilot.

Discussion and Limitations

What we think carries over. Three points seem general. The uncertainty of a size forecast is governed by the noise-to-slope ratio τ/β and by how far the target is from the pilots measured in units of their spread. Fitting a curve without an asymptote is optimistic by construction, and the bias increases with the distance of the extrapolation. And a test set that is too small adds noise to every pilot, which is not reduced by repeating training subsets.

Synthetic setting. The noise model is log-normal with constant τ, which is a simplification; real noise depends on the error level and on the test set. Eq. (2) is first-order and assumes the power law with no asymptote. The three-parameter fit was a single nonlinear least squares with bounds; other fitting methods, priors or multi-start fits would change its failure rate.

Real data. We used three small learning curves, so the true curve at the largest sizes is estimated from one fit at the full pool and a few draws. The targets are inside the data, and the extrapolation distance (four to eight times the largest pilot) is smaller than in most practical planning. We did not study deep models, other metrics, or forecasts that combine data size with model size.

7. Conclusion

A log-log extrapolation of a learning curve has a closed-form uncertainty that depends on the noise, the slope and the distance of the target from the pilots, and the formula matched simulation within 1%. For five pilots between 20 and 160 cases and 5% noise, a forecast of 10,000 cases is uncertain by a factor of 1.7. A curve with an irreducible error makes the forecast optimistic, by 9962% at a target of 1,280 cases with an asymptote of 0.03, and a three-parameter fit that removes the bias is unstable with five pilots. The next steps are to test the formulas on deep models and larger scales, to combine pilots with prior information on the exponent, and to design the pilots adaptively.

References

[1]. C. Cortes, L. D. Jackel, S. A. Solla, V. Vapnik, J. S. Denker. Learning curves: asymptotic values and rate of convergence. In *NIPS*, 1994.

[2]. S. Mukherjee, P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell, T. R. Golub, J. P. Mesirov. Estimating dataset size requirements for classifying DNA microarray data. *J. Computational Biology*, 10(2):119–142, 2003.

[3]. J. Hestness et al. Deep learning scaling is predictable, empirically. arXiv:1712.00409, 2017.

[4]. J. Kaplan et al. Scaling laws for neural language models. arXiv:2001.08361, 2020.

[5]. J. S. Rosenfeld, A. Rosenfeld, Y. Belinkov, N. Shavit. A constructive prediction of the generalization error across scales. In *ICLR*, 2020.

[6]. C. Perlich, F. Provost, J. S. Simonoff. Tree induction vs. logistic regression: a learning-curve analysis. *J. Mach. Learn. Res.*, 4:211–255, 2003.

[7]. R. L. Figueroa, Q. Zeng-Treitler, S. Kandula, L. H. Ngo. Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, 12:8, 2012.

[8]. C. Beleites, U. Neugebauer, T. Bocklitz, C. Krafft, J. Popp. Sample size planning for classification models. *Analytica Chimica Acta*, 760:25–33, 2013.

[9]. T. Viering, M. Loog. The shape of learning curves: a review. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 45(6):7799–7819, 2023.

[10]. T. Domhan, J. T. Springenberg, F. Hutter. Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. In *IJCAI*, 2015.

[11]. G. A. F. Seber, C. J. Wild. *Nonlinear Regression*. Wiley, 1989.

[12]. A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[13]. N. R. Draper, H. Smith. *Applied Regression Analysis*, 3rd ed. Wiley, 1998.

[14]. F. Pedregosa et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

[15]. E. Alpaydin, C. Kaynak. Optical recognition of handwritten digits. UCI Machine Learning Repository, 1998.

[16]. W. N. Street, W. H. Wolberg, O. L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *SPIE*, vol. 1905, 1993.

[17]. C. Sun, A. Shrivastava, S. Singh, A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, 2017.

A Derivations

Eq. (2). With $y_i = \alpha - \beta x_i + \eta_i$, OLS gives $\mathrm{Var}(\hat{\alpha}) = \tau^2(1/m + \bar{x}^2/S_{xx})$, $\mathrm{Var}(\hat{\beta}) = \tau^2/S_{xx}$ and $\mathrm{Cov}(\hat{\alpha}, \hat{\beta}) = \tau^2\bar{x}/S_{xx}$. The delta method applied to $x^* = (\hat{\alpha} - \log \varepsilon)/\hat{\beta}$ gives $\mathrm{Var} \approx (1/\beta^2)[\mathrm{Var}(\hat{\alpha}) + x^{*2}\mathrm{Var}(\hat{\beta}) - 2x^*\mathrm{Cov}(\hat{\alpha}, \hat{\beta})]$, and the bracket equals $\tau^2[1/m + (x^* - \bar{x})^2/S_{xx}]$.

Corollary 1. For x_i equally spaced over Δ, $\mathrm{Var}(\bar{x}) = \Delta^2(m + 1)/(12(m - 1))$, so $S_{xx} = m\,\mathrm{Var}(\bar{x})$. With half the points at each end, $\mathrm{Var}(\bar{x}) = \Delta^2/4$.

Proposition 2. The log-log curve is $g(x) = \log(c + a\,e^{-b\,x})$, with $g'(x) = -b\,a\,e^{-b\,x}/(c + a\,e^{-b\,x})$, whose magnitude decreases in x when $c > 0$. So g is convex, and the least-squares line through points on g over $[x_1, x_m]$ lies above g outside this

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P7

interval and has a steeper slope than g beyond it.

Eq. (5). The error count is binomial with mean eN and variance $e(1 - e)N$, so the relative standard deviation of the error estimate is $\sqrt{(e(1 - e)N)/(eN)} = \sqrt{((1 - e)/(eN))}$.

B Parameters

Synthetic: $e(n) = c + a\, n^{-b}$; log-normal pilot noise with sd τ. Forecast study: $a = 2$, $b = 0.6$, $c = 0$, pilots 20, 32, 50, 80, 127 (five geometric sizes from 20 to 160, rounded), 20,000 repetitions. Design study: $m \in \{3, 4, 5, 8, 12\}$, range factor $\in \{2, 4, 8, 16, 32\}$, 8,000 repetitions. Asymptote study: three curves, three targets, $\tau = 0.03$, 1,000 repetitions; three-parameter fit by bounded nonlinear least squares. Real data: digits (500 test cases, pool 1,297) and breast cancer (169 test cases, pool 400); logistic regression (digits $C = 0.1$, breast $C = 1$, standardised features) and 5-NN; truth from 15 subsets per size; pilots five geometric sizes, three repeats each, 150 repetitions; interval with t quantile (3 degrees of freedom). All figures were produced by code at 600 dpi and are printed at their stored width. Code, results and figures are in the supplementary archive.

How To Cite:

Uma Maheswari Krishnamoorthy , " Learning-Curve Extrapolation Under Limited Data: Bias and Uncertainty in Predicting Training-Set Requirements ", International Journal of Research and Development in Engineering Sciences (IJRDES), Volume 8 , Issue 6 , pp 68 – 77, November 2026.
CrossRef DOI: https://doi.org/10.63328/ijrdes-v8ri6p7

Declaration

Conflicts of Interest: The authors declare no conflict of interest.

Author Contribution: All authors wrote the main manuscript text and also consent to the submission.

Ethical approval: Not applicable.

Consent to Participate: All authors consent to participate.

Funding: Not applicable, and No funding was received

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Personal Statement: We declare with our best of knowledge that this research work is purely Original Work and No third party material used in this article drafting. If any such kind material found in further online publication, we are responsible only for any judicial and copyright issues.

Generative AI Statement: The authors declare that generative AI was not used in the preparation or creation of this manuscript. The alternative text (alt text) accompanying the figures in this article was generated by Frontiers with the assistance of artificial intelligence. Reasonable efforts have been made to ensure the accuracy and appropriateness of the generated alt text, including review by the authors wherever possible. If you identify any inaccuracies or issues, please contact the editorial team.

Publisher's Note: The statements, opinions, and claims presented in this article are exclusively those of the authors and do not necessarily reflect the views of their affiliated institutions, the publisher, editors, or reviewers. Any products discussed or evaluated, as well as any claims made by their manufacturers, are not warranted, approved, or endorsed by the publisher.

Acknowledgements

We thank everyone who inspired our work

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com