Cross-validation Optimal Fold-Number for Model Selection

Angelos Vasilopoulos[a], Gregory J. Matthews[b]*

[a] *Stritch School of Medicine, Loyola University Chicago, Chicago, IL*
[b] *Department of Mathematics and Statistics, Loyola University Chicago, Chicago, IL*

https://doi.org/10.33697/ajur.2024.123

Students: avasilopoulos1@luc.edu
*Mentor: gmatthews1@luc.edu**

ABSTRACT

The resampling method of k-fold cross-validation is popular for error estimation and model selection in computational research. However, there is limited focus in the literature on the question of what fold number k is appropriate for various dataset dimensions. Here we review relevant literature and present a simulation of linear and least absolute shrinkage and selection operator (LASSO) regression prediction error estimation at various values of k and sample size n. In agreement with current literature, we find that contrary to a persisting understanding, there is no bias-variance trade-off in selection of k. Instead, with increasing k both bias and variance decrease, perhaps asymptotically. Our results also suggest a predictable relationship between optimal values of k and n.

KEYWORDS

Cross-validation; Optimization; Fold number

INTRODUCTION

Cross-validation, also known as k-fold cross-validation, is a popular method of error estimation for model selection in computational research. In this method, n observations are first divided into k groups. In a first iteration, $k-1$ groups are used as a training set. The remaining group is used as a test set to calculate estimated prediction error ($\widehat{\mathrm{PE}}$). This process is repeated i times, with a different test set in each iteration. The average of k estimates of PE

$$\hat{\theta} = CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\widehat{\mathrm{PE}}_i \qquad \textbf{Equation } 1.$$

is meant to estimate the true model error θ, i.e., the error of the model tested on the population.[1] When the objective is to choose a best-performing model of a set, $CV_(k)$ is calculated for each candidate model and the model with the lowest CV_k is usually selected.

Despite the popularity of cross-validation, there is limited focus in the literature on the question of what fold number k is appropriate for cross-validation with a dataset of a given size. One popular idea is that selection of k comes with a bias-variance trade-off—specifically, that as k increases, the bias $Bias(\hat{\theta}) = \mathrm{E}(\hat{\theta}) - \theta$ and variance $Var(\hat{\theta}) = \mathrm{E}(\hat{\theta}^2) - \mathrm{E}(\hat{\theta})^2$ of k-fold error estimation decrease and increase, respectively. This idea appears in early literature but also in modern, widely used textbooks.[4, 5, 7] If the idea that the choice of k is associated with bias and variance, via a bias-variance trade-off or differently, is true, then the selection of k can drastically influence model performance and selection.

Subsequent, albeit limited, literature argues differently. In the case of leave-one-out cross-validation (LOOCV), i.e., with $k = n$, some authors suggest that asymptotically both bias and variance of error estimation decrease as k increases and that bias and variance of error estimation are uniformly low.[9, 10]

More recently, different ways have been proposed to quantify the variance reduction achieved by cross-validation when the true prediction error (PE) is not known, e.g., as mean-square stability or as loss stability.[11, 12] However, it has also been demonstrated that, due to overlap between training and test sets in cross-validation, there is no universal (i.e., valid under all distributions) unbiased estimator of the variance of k-fold cross-validation.[13]

In addition to theoretical analyses, there are some simulation results showing the phenomenon of variance reduction by cross-validation. However, simulations currently in the literature provide limited insight into the dependence of optimal fold number k_{optimal} on sample size n or involve biased variance calculations.[14, 15] Here we present a simulation of linear regression and least absolute shrinkage and selection operator (LASSO) regression to observe the relationship of cross-validation fold number k to model selection accuracy for various samples of size n under the assumptions of linear regression.[19] Our focus is a known population with a feature Y of the form $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_5 X_5 + \epsilon$ where $\beta_1 = \cdots = \beta_5 = 1$, $\epsilon \sim N(0, 10)$, and mean squared error (MSE) is the loss function.

METHODS AND PROCEDURES

In machine learning tasks, the values of 5 and 10 traditionally have been used as the values of k in k-fold cross validation and here we evaluate these choices and alternatives.[20] With the results of the following simulation, we make a different recommendation of k with the aim of improving model selection accuracy.

Population

Consider a population of size N = 500,000 with five of $P = 100$ features $X_1 ... X_{100} \sim N(0, 1)$ a linear combination of feature $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_5 X_5 + \epsilon$ where $\beta_1 = \cdots = \beta_5 = 1$ and $\epsilon \sim N(0, 10)$. The true model of Y is $f \in F$, a set of competing models. $f = \mathrm{E}(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_5 X_5$ and its MSE is

$$\theta = \mathrm{MSE}(f) = \frac{\sum_{i=1}^{N} [Y_i - \mathrm{E}(Y_i)]^2}{N}. \qquad \textbf{Equation 2.}$$

Sampling and regression

From the population we take a sample of size n. We estimate Y as $\hat{Y}$ by regression on a subset of $X_1 ... X_{100}$, regression coefficients estimated by the least squares method, and compute $\mathrm{MSE}(\hat{Y})$ by k-fold cross-validation as

$$\hat{\theta} = \mathrm{MSE}(\hat{Y}) = \frac{1}{k} \sum_{j=1}^{k} \frac{\sum_{i=1}^{n/k} (Y_{j_i} - \hat{Y_{j_i}})^2}{n/k} \qquad \textbf{Equation 3.}$$

where j_i is the index of the i^{th} element of the j^{th} fold.

In addition to regression with $X_1 ... X_5$, f, we under-fit with $X_1 ... X_3$, over-fit with $X_1 ... X_{20}$ and $X_1 ... X_{100}$, and do regression with noise $X_6 ... X_{100}$ only, for each $k \in$ 2, 10, 20, 30, ..., n for each $n \in$ 100, 200, 300, ..., 1000.

Simulation

We perform this simulation 1000 times and then calculate simulation-wise $\mathrm{MSE}(\hat{\theta})$ for each k, for each n.

We perform an additional 1000 simulations predicting Y as $\hat{Y}$ by LASSO regression instead of linear regression for each $k \in$ 2, 10, 20, 30, ..., 100, n for each $n \in$ 250, 500, 750, 1000. To avoid data leakage, in which testing data is used in model training[2, 3], we introduce an inner 5-fold cross-validation loop for the selection of parameter λ as in the minimization of

$$\sum_{i=j}^{p} (Y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \qquad \textbf{Equation 4.}$$

For each k, for each n, we count the number of times A that the true model is selected and consider as optimal fold number for each n the value k_{optimal}, at which A is the highest.

This is because in practice the model with the lowest $\widehat{\mathrm{PE}}$ is more likely to be selected.[16] We refer to this as lowest-error model selection. As our simulation results show, however, it is possible for k-fold cross-validation to result in calculations of $\mathrm{MSE}(\hat{\theta})$ such that a competing model $\hat{f} \in F$ has a lower $\widehat{\mathrm{PE}}$ than the true model f. This false model may have the lowest $\widehat{\mathrm{PE}}$ but its generalization error will be higher in the long-run (i.e., when tested on a large part of the population) than the generalization error of the true model. Thus, it is preferable for the value of k selected to result in f having the lowest $\widehat{\mathrm{PE}}$.

Fold number as a function of sample size

Having considered the effect of k on model selection accuracy for a given n, we may also consider the relationship of k_{optimal} and n. Our results suggest that after a certain value of k, changes in A become negligible, in the case of linear regression, or negative, in the case of LASSO regression, and resource-expensive increases of fold number become undesirable. We refer to this "point of diminishing returns" as k^*_{optimal} using the "elbow method" commonly used in cluster analysis.[17]

To estimate k^*_{optimal} for each n, we first fit to the dataset (k, A) for each sample size n a hypoethhical model of the form

$$\hat{A}(k) = Bk + C + \frac{D}{k}$$

Equation 5.

with constants B to D specific to each n. Certain more complicated models (e.g., with higher-order terms) may provide better approximation of A but make optimization by the following scheme non-trivial. We draw a line L through $(k_1, \hat{A}_1)$ and $(k_n, \hat{A}_n)$ and maximize the perpendicular distance d of each point on $\hat{A}$ from L, so that

$$\begin{aligned}
k^*_{\mathrm{optimal}} &= \arg\max{}_k d[\hat{A}(k), L] \\
&= \arg\max{}_k d[Bk + C + \frac{D}{k}, y - (Bn + C + \frac{D}{n}) = m(x-n)] \\
&= \arg\max{}_k d[Bk + C + \frac{D}{k}, y - (Bn + C + \frac{D}{n}) = \frac{(Bn + C + \frac{D}{n}) - (2B + C + \frac{D}{2})}{n-2}(x-n)] \\
&= \arg\max{}_k d[Bk + C + \frac{D}{k}, y - (Bn + C + \frac{D}{n}) = (B - \frac{D}{2n})(x-n)] \\
&= \arg\max{}_k d[Bk + C + \frac{D}{k}, (B - \frac{D}{2n})x - y + C + \frac{D}{2} + \frac{D}{n} = 0]
\end{aligned}$$

Equation 6.

It is known that the distance d between a point $(k, \hat{A}(k))$ and a line of the form

$$ax + by + c = 0$$

Equation 7.

is

$$d = \frac{|ak + b\hat{A}(k) + c|}{\sqrt{a^2 + b^2}}$$

Equation 8.

so

$$\begin{aligned}
k^*_{\mathrm{optimal}} &= \arg\max{}_k d[Bk + C + \frac{D}{k}, (B - \frac{D}{2n})x - y + C + \frac{D}{2} + \frac{D}{n} = 0] \\
&= \arg\max{}_k \frac{|(B - \frac{D}{2n})k - (Bk + C + \frac{D}{k}) + (C + \frac{D}{2} + \frac{D}{n})|}{\sqrt{(B - \frac{D}{2n})^2 + (-1)^2}}.
\end{aligned}$$

Equation 9.

Maximizing d, knowing that n and k are always positive, we find that

$$\begin{aligned}\frac{d}{dk}d[\hat{A}(k), L] &= \frac{d}{dk}\left(\frac{|(B-\frac{D}{2n})k-(Bk+C+\frac{D}{k})+(C+\frac{D}{2}+\frac{D}{n})|}{\sqrt{(B-\frac{D}{2n})^2+(-1)^2}}\right)\\ &= \frac{|n||k|(2nD-Dk^2)(-Dk^2-2nD+nDk+2Dk)}{nk^3|-Dk^2-2nD+nDk+2Dk|\sqrt{D^2+4n^2}} = 0\\ &\Rightarrow k = 0, 2, \sqrt{2n}, n.\end{aligned}$$

Equation 10.

Evaluating d at these four values of k, we find that

$$\begin{aligned}d(0) &= \text{DNE}\\ d(2) &= 0\\ d(\sqrt{2n}) &= \frac{|nD+2D-\sqrt{2n}D-D|}{\sqrt{(B-\frac{D}{2n})^2+(-1)^2}}\\ d(n) &= 0.\end{aligned}$$

Equation 11.

Since $d(\sqrt{2n}) > 0$,

$$k^*_{\text{optimal}} = \sqrt{2n}$$

Equation 12.

which fits the relationship of k^*_{optimal} $vs.$ n (Figure 9).

A note on LASSO

It may seem that the choice of $k = 5$ for the selection of λ by cross-validation is presumptive for a study that seeks to determine the best value of k for cross-validation. If the choice of k influences the value of λ, it influences the subset of models that will be competing with f. Does k^*_{optimal} change with the subset of models competing with f? Replications of this simulation varying k only in the selection of λ may elucidate.

RESULTS AND DISCUSSION

Interpretations of early literature have resulted in lasting misconceptions about the use of cross-validation. Such misconceptions include the idea that there is a bias-variance trade-off $Bias^2(\hat{\theta}) \propto 1/Var(\hat{\theta})$ associated with selection of k and that $k = 10$ is the best value to use in k-fold cross-validation.[5–8]

In agreement with a small but growing body of literature, our simulation results suggest that neither of these ideas are necessarily correct. Instead, in the context of linear and LASSO regression with standard normal data and certain random error, we find that for various n both bias and variance decrease as k increases (Figures 1 - 6), i.e., $Bias^2(\hat{\theta}) \propto Var(\hat{\theta})$, and although in the case of LASSO, 10-fold cross-validation seems to be a near-optimal choice for large n, for smaller samples and linear regression other values of k appear to be optimal for model selection (Figures 7 and 8; Table 1).

Table 1. Optimal values of k for sample sizes n in linear regression.

n	100	200	300	400	500	600	700	800	900	1000
Optimal k	14	20	25	28	32	35	37	40	42	45

Table 2. Optimal values of k for sample sizes n in LASSO regression.

n	100	250	500	750	1000
Optimal k	14	22	32	39	45

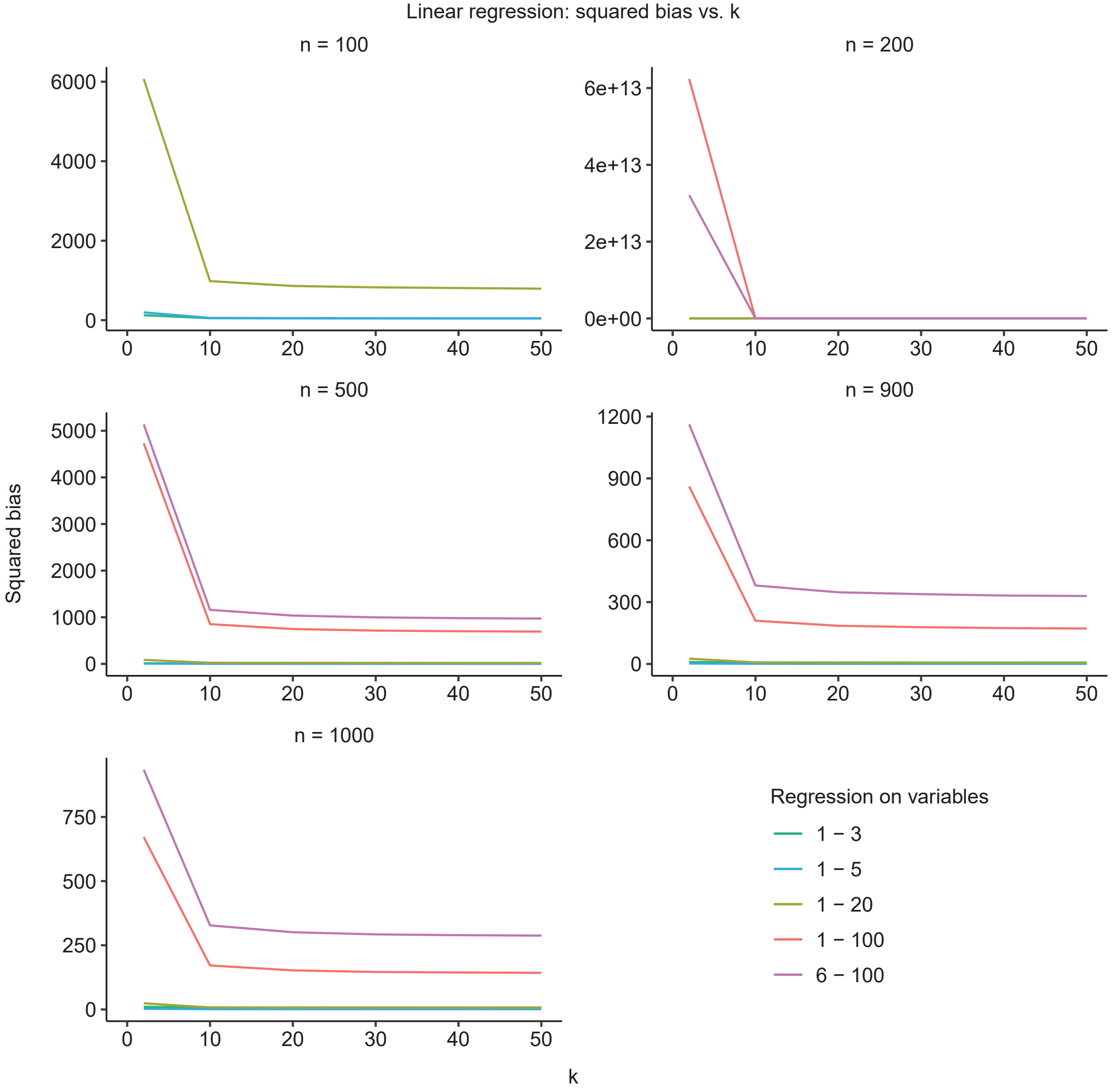

Figure 1. Linear regression squared bias vs. fold number for various sample sizes. As fold number increases, bias decreases initially before leveling off. See Figure 10 in appendix for more sample sizes and fold numbers over 50.

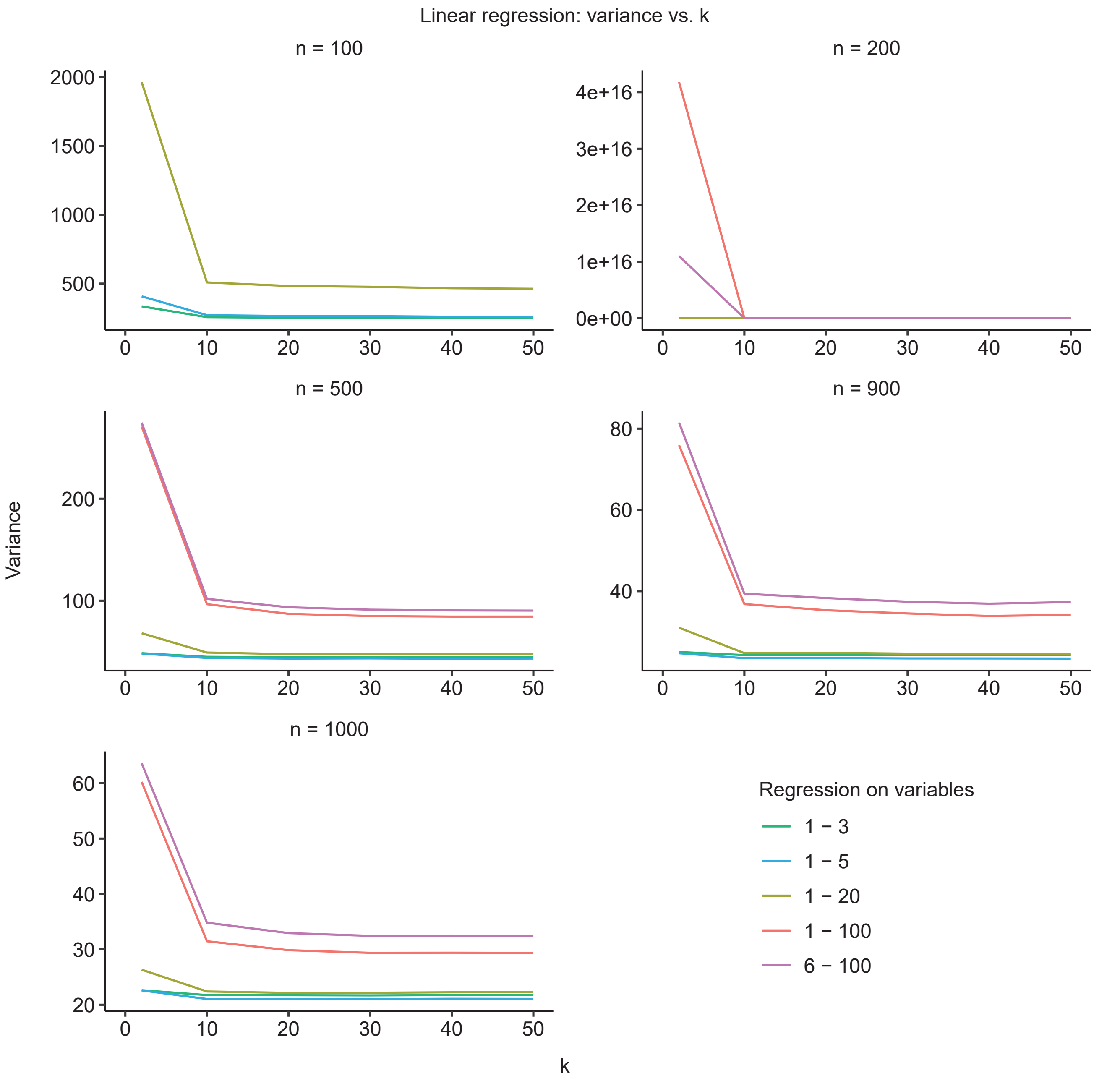

Figure 2. Linear regression variance *vs*. fold number for various sample sizes. As fold number increases, variance decreases initially before leveling off. See Figure 11 in appendix for more sample sizes and fold numbers over 50.

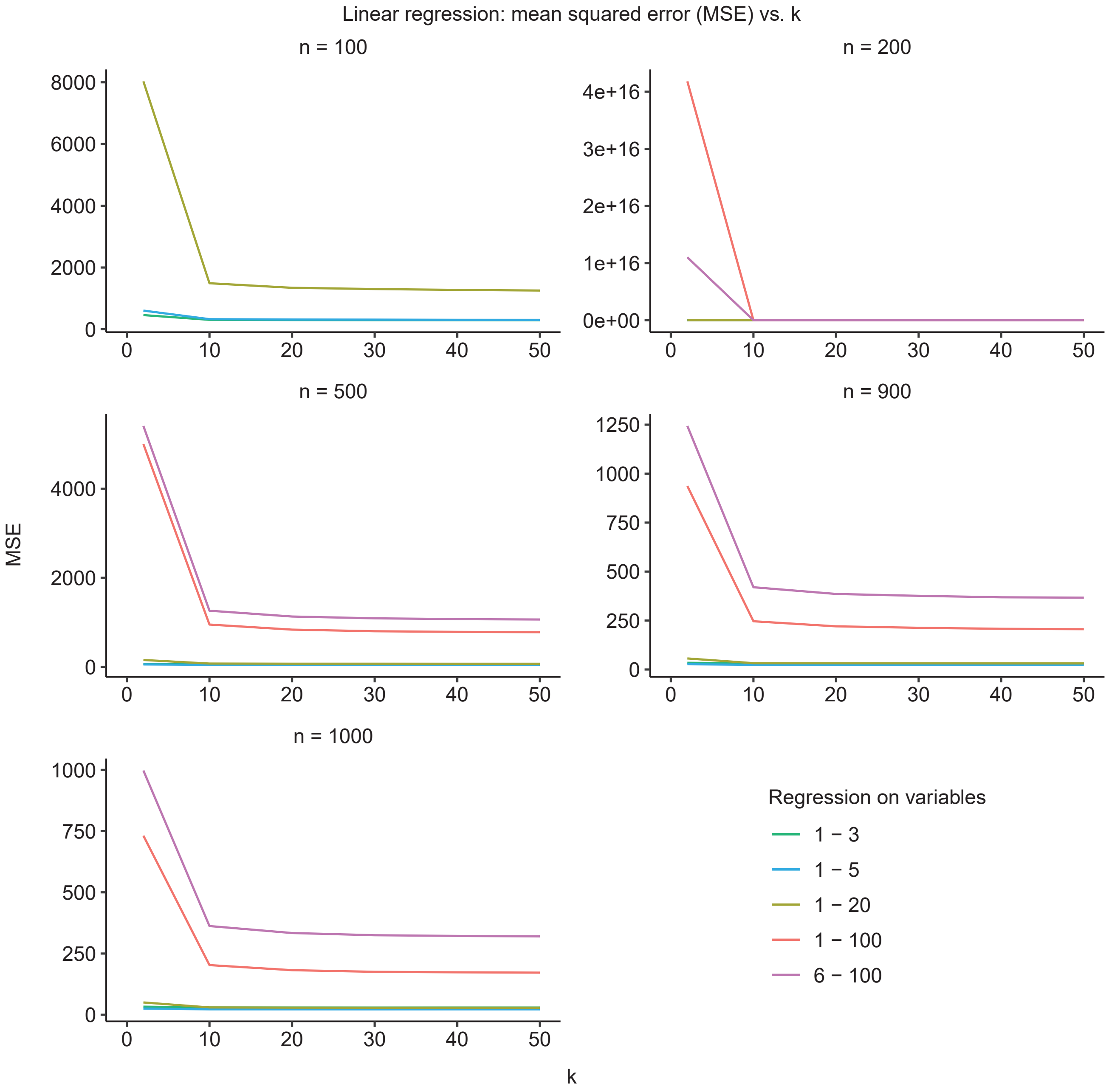

Figure 3. Linear regression mean squared error $vs.$ fold number for various sample sizes. As fold number increases, mean squared error decreases initially before leveling off. See Figure 12 in appendix for more sample sizes and fold numbers over 50.

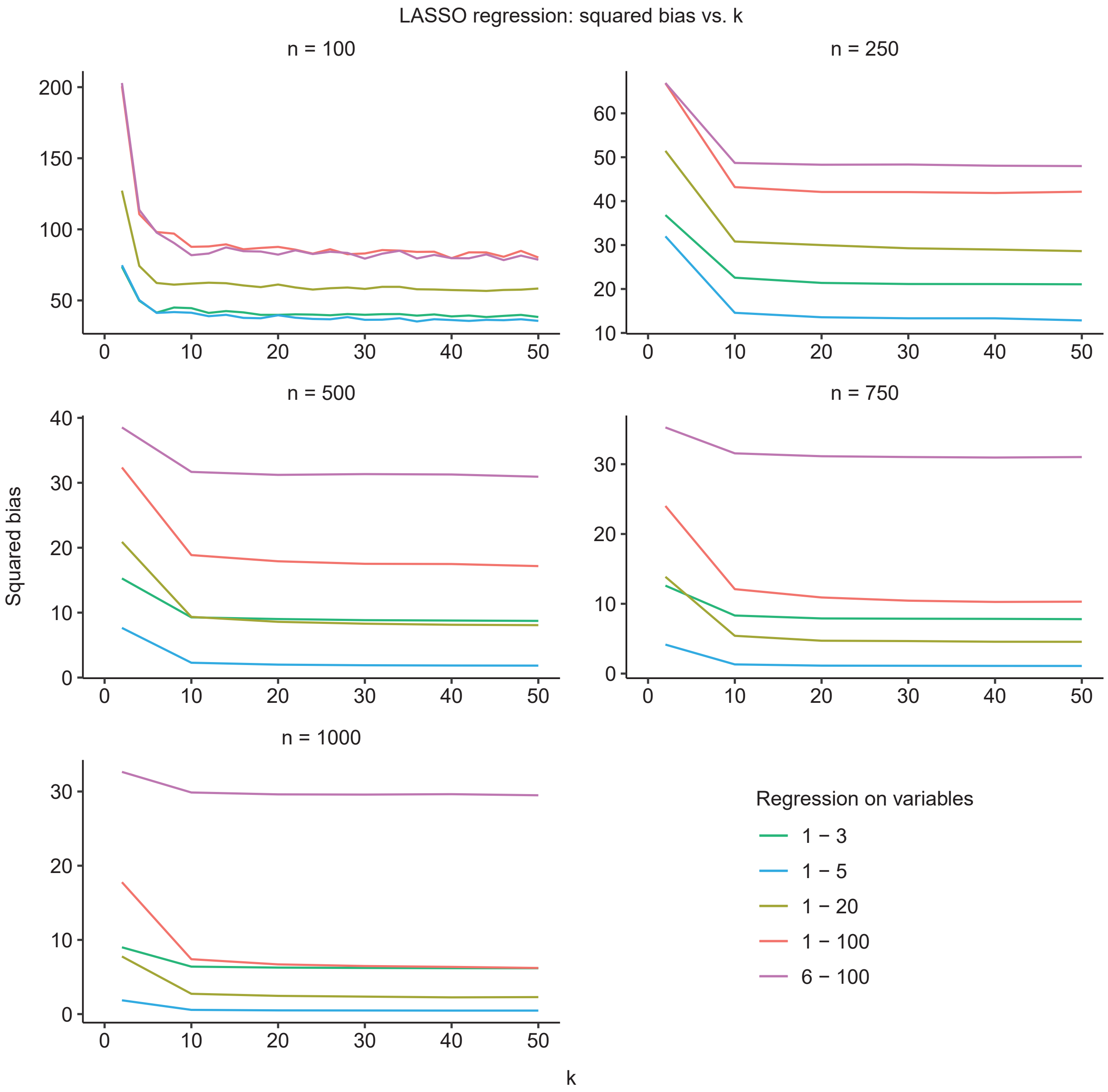

Figure 4. Linear regression squared bias $vs.$ fold number for various sample sizes. As fold number increases, bias decreases initially before leveling off. See Figure 13 in appendix for fold numbers over 50.

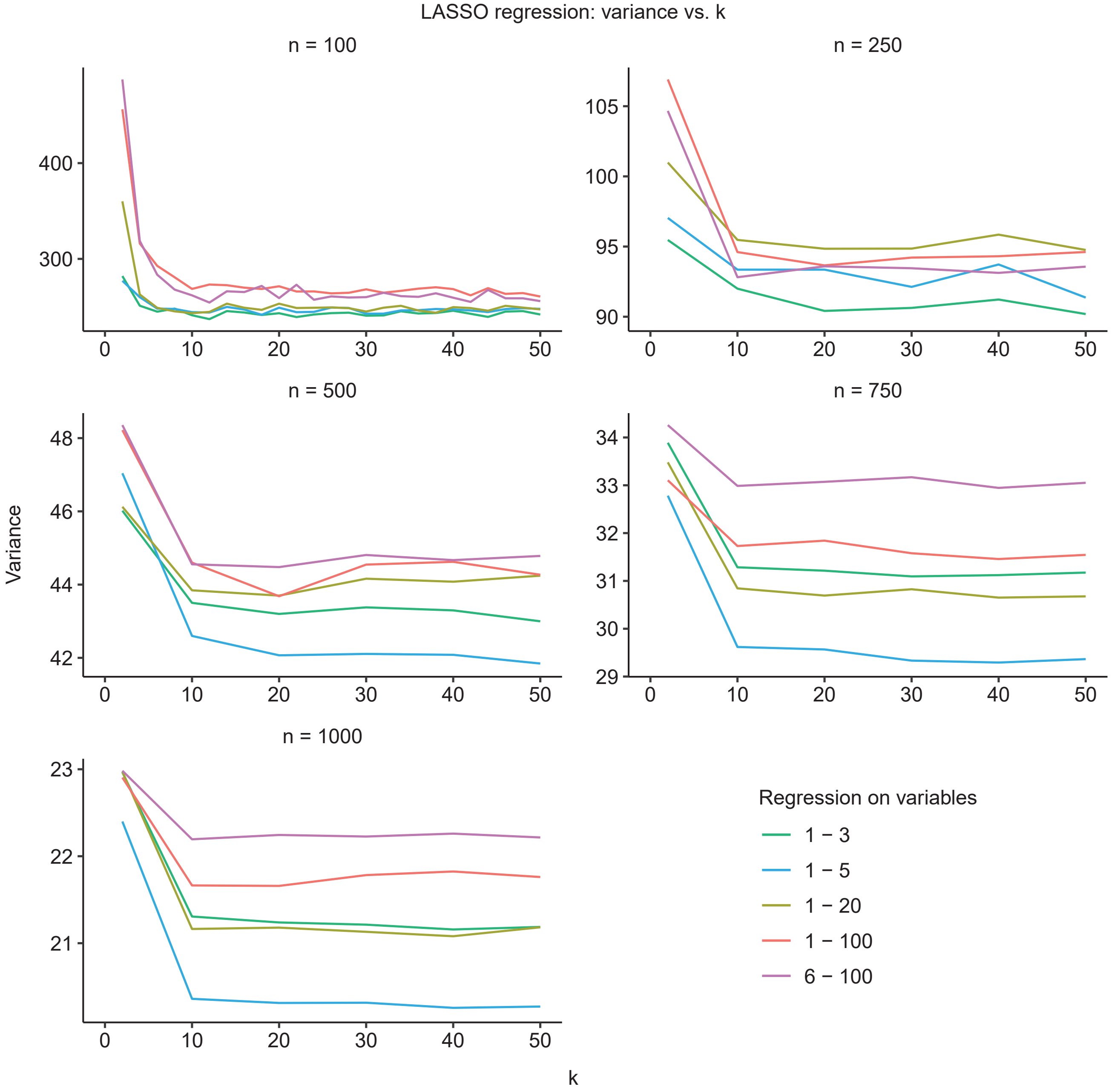

Figure 5. LASSO regression variance *vs*. fold number for various sample sizes. As fold number increases, variance decreases initially before leveling off. See Figure 14 in appendix for fold numbers over 50.

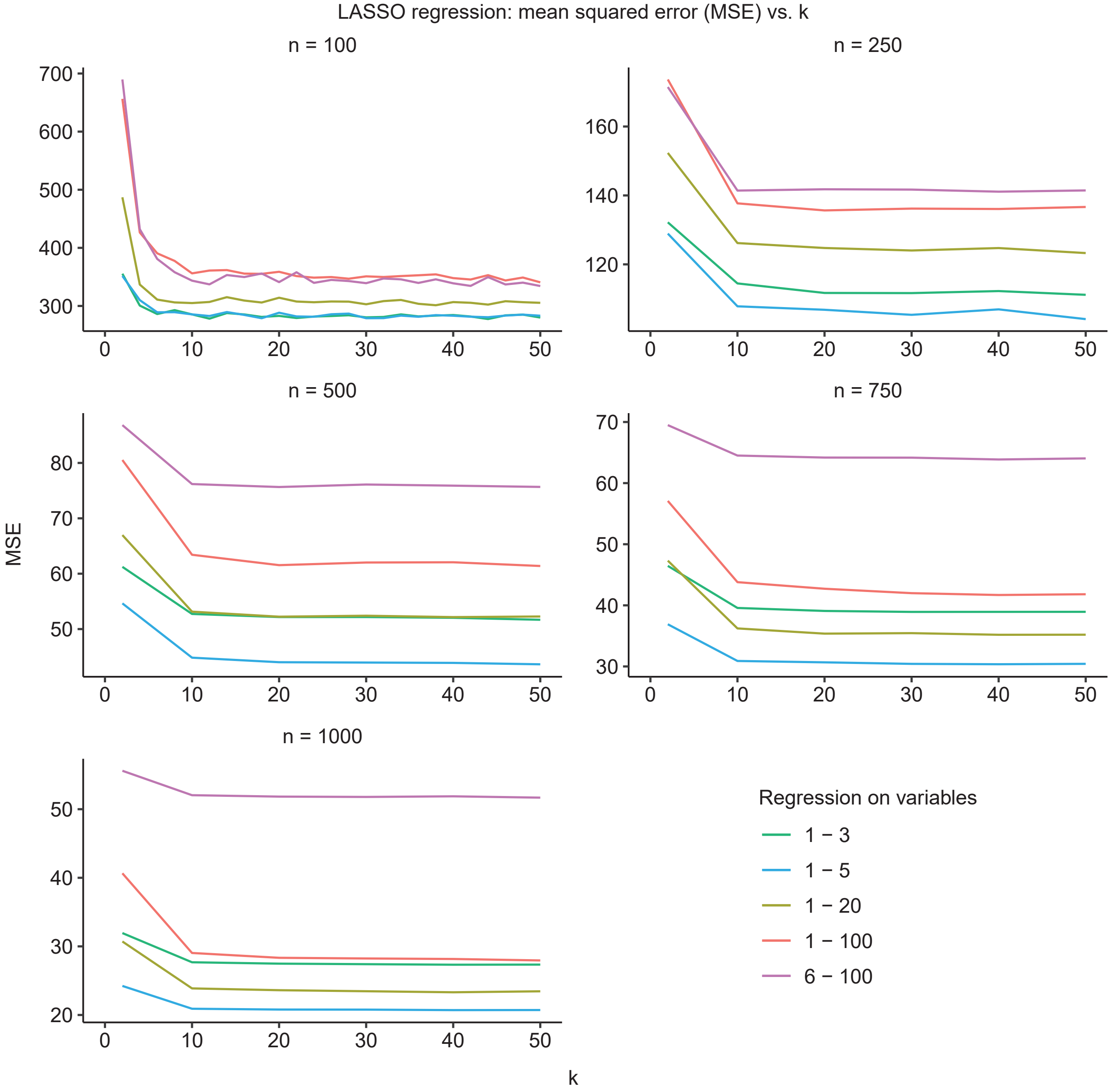

Figure 6. LASSO regression mean squared error *vs*. fold number for various sample sizes. As fold number increases, mean squared error decreases initially before leveling off. See Figure 15 in appendix for fold numbers over 50.

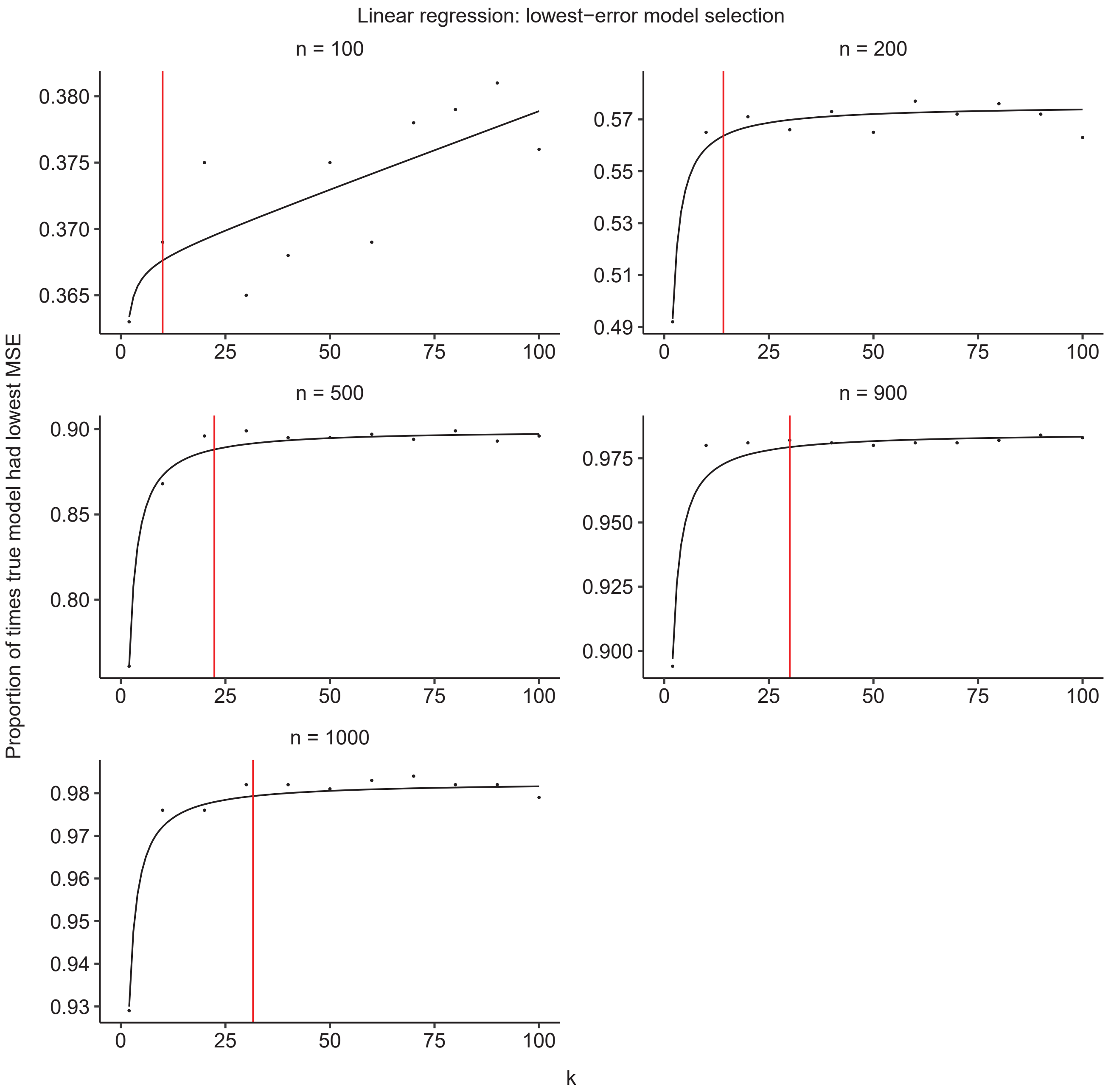

Figure 7. For linear regression, the proportion of times the true model has the lowest mean squared error $vs.$ fold number for various sample sizes. As fold number increases, the proportion of times that the true model is selected increases, but the rate of this increase declines. See Figure 16 in appendix for fold numbers over 100 and more sample sizes.

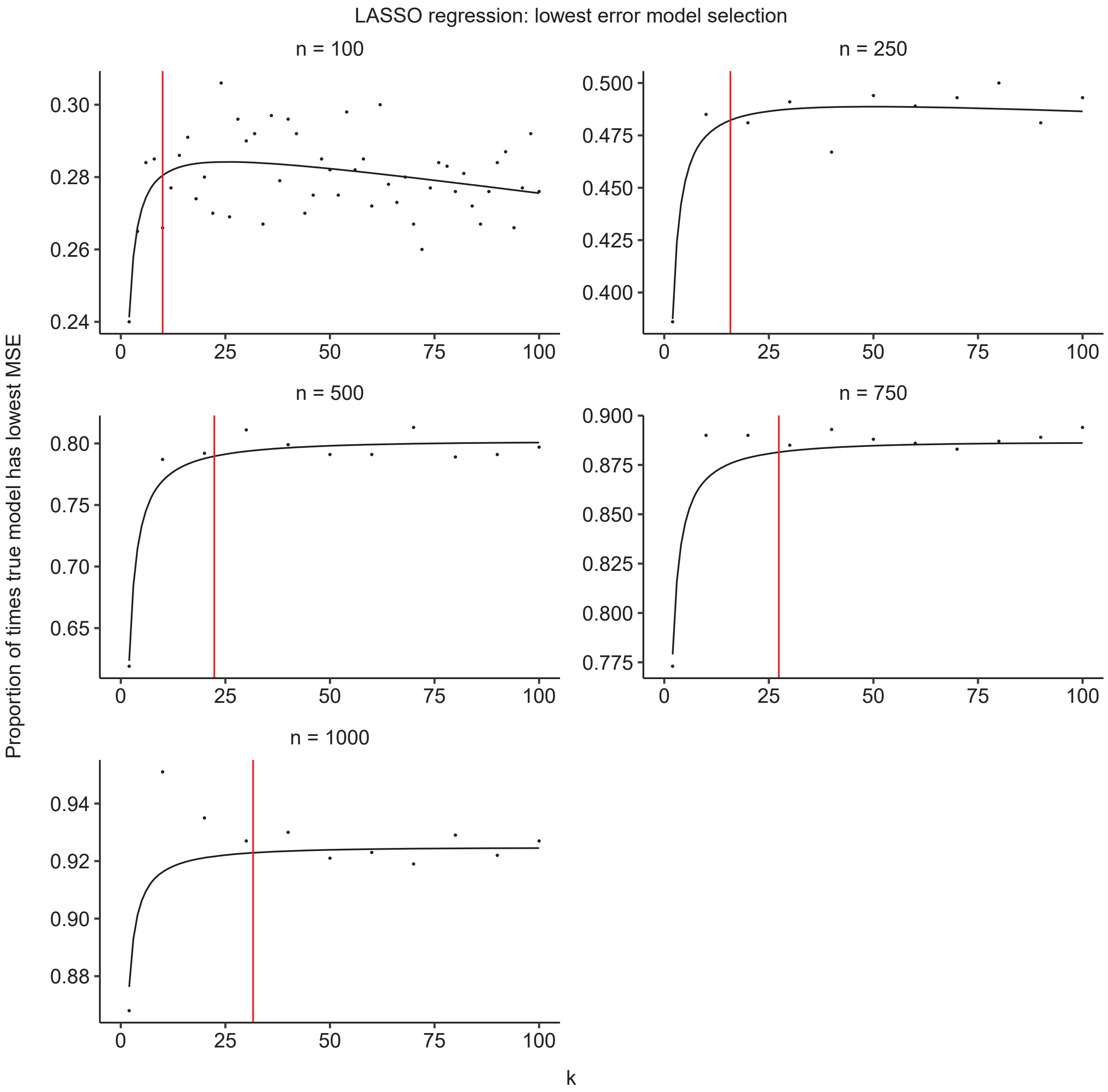

Figure 8. For LASSO regression, the proportion of times the true model has the lowest mean squared error *vs*. fold number for various sample sizes. As fold number increases, the proportion of times that the true model is selected increases, but the rate of this increase declines. See Figure 17 in appendix for fold numbers over 100 and more sample sizes.

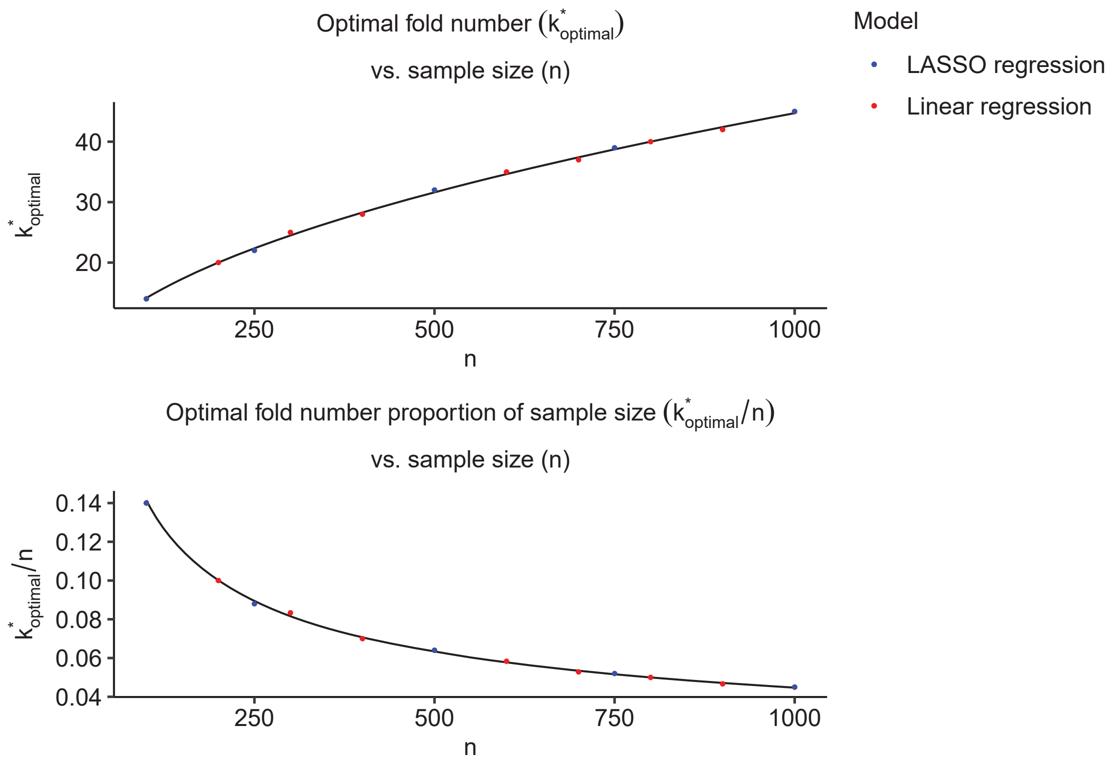

Figure 9. Sample size vs. fold number. As sample size increases, optimal fold number increases at a diminishing rate.

Note that the case of linear regression with $n = 100$ and number of features $p = 100$ is not displayed because there are $n - p - 1 = -1$ degrees of freedom, which is not a sufficient number for regression. The case of $n = 100$ and $p = 95$ degrees of freedom is similarly problematic and is not displayed because the high level of error in bias, variance, and MSE estimation associated with $n - p - 1 = 5$ degrees of freedom obscures important patterns in the rest of the data.

As previous authors have noted,[14] it is important to distinguish between possible goals of cross-validation that previously have been conflated[5]: estimation of PE, in which case $k_{\text{optimal}} = \arg\min_k(\text{PE} - \widehat{\text{PE}})$, and model selection, in which case

$$k_{\text{optimal}} = \arg\max_k(A) \quad \textbf{Equation } 13.$$

which is related to the definition of k^*_{optimal} used in this study.

We find that in the cases of both linear and LASSO regression, as n increases, k^*_{optimal} increases but at a lower rate than n, such that k^*_{optimal}/n decreases, perhaps asymptotically. For LASSO regression, we also find that as k increases past a certain value, lowest-error model selection becomes less reliable, i.e., $\hat{A}$ decreases, i.e., the true model has the lowest MSE less frequently (Figure 8). The rate of this reduction increases with n.

The reason for this is related to what is known as the cross-validation paradox[18]: greater quantity of data results in more accurate estimation of PE; occasionally, this makes model-wise differences in $\widehat{\text{PE}} = CV_{(k)}$ less exaggerated, making it more difficult to distinguish between models and resulting in less accurate model selection.

However, we do not observe the cross-validation paradox in our linear regression simulations. Although cross-validation results in similar bias-variance reductions in both cases, in the case of linear regression $\hat{A}$ increases with increasing k initially before leveling off. However, this is not surprising, as our linear regression simulation does not involve variable selection, so that the differences in models are more pronounced. To observe the cross-validation paradox in linear regression would require comparison of more similar linear models predisposed to similarity in $\widehat{\text{PE}}$, as in the simulation of LASSO regression.

In the cases of both linear and LASSO regression, k^*_{optimal}/n seems to change predictably with n. Specifically, it may be possible to model the relationship with some asymptotically decreasing function, while the relationship between k^*_{optimal} and n seems to follow a positive pattern (Figures 9 and 10).

CONCLUSIONS

Early literature suggests that increasing cross-validation fold number is related to decreasing bias and increasing variance of error estimation. However, more recent work suggests that this is not the case. Instead, increasing k results in bias and variance reduction. This phenomenon is observable in our simulation results, which suggest that bias and variance decrease asymptotically with increasing k.

Our results also indicate a predictable relationship between k^*_{optimal} and sample size n. Although further data and analysis are needed to draw any reliable conclusions, modeling the relationship between n and k^*_{optimal} would have practical utility, potentially improving the selection of k from a largely arbitrary decision between 5 and 10.

Future research may also focus on error estimation of other models, including models capturing non-linear relationships or involving tuning of multiple hyper-parameters, e.g., random forest or gradient boosting. It may also be interesting to study the value of k in repeated cross-validation or nested cross-validation, with the value of k variable in the inner loop, outer loop, or both, as the use of $k = 5$ for the selection of λ in this simulation is a notable limitation of the study.

Comparisons of k-fold cross-validation and other model selection tools like the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and forward or backward selection are also of interest.

ACKNOWLEDGEMENTS

The authors thank Dr. Gregory Matthews's Fall 2022 Predictive Analytics class.

REFERENCES

1. Bates, T.H., Tibshirani, R. (2023) Cross-validation: What does it estimate and how well does it do it? *Journal of the American Statistical Association* 0(0), 1–12. *https://doi.org/10.48550/arXiv.2104.00673*
2. Montano, I.H., Aranda, J.J.G., Diaz, R.J. et al. (2022) Survey of Techniques on Data Leakage Protection and Methods to address the Insider threat. *Cluster Computing* 25, 4289–4302. *https://doi.org/10.1007/s10586-022-03668-2*
3. Kaufman, S., Rosset, S., Perlich, C., Stitelman, O. (2012) Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data* 6, 1–21. *https://doi.org/10.1145/2020408.2020496*
4. Efron, B. (1983) Estimating the error rate of a prediction rule: Improvement on cross-validation, *Journal of the American Statistical Association* 78, 316–331. *https://doi.org/10.1080/01621459.1983.10477973*
5. Kohavi, R. (2001) Near-optimal bounds for cross-validation via loss stability, *Proceedings of Machine Learning Research* 27–35.
6. Hastie, T., Tibshirani, R., Friedman, J. (2009) *The Elements of Statistical Learning* 2nd ed., Springer, New York.
7. James, G., Witten, D., Hastie, T., Tibshirani, R. (2021) *An Introduction to Statistical Learning* 2nd ed., Springer, New York.
8. Barber, D. (2012) *Bayesian Reasoning and Machine Learning* 1st ed., Cambridge University Press, Cambridge.
9. Burman, P. (1989) A comparative study of ordinary cross-validati9on, v-fold cross-validation and repeated learning-testing methods, *Biometrika* 76, 503–514. *https://doi.org/10.1093/biomet/76.3.503*
10. Breiman, L., Spector, P. (1992) Submodel selection and evaluation in regression: the X-random case, *International Statistical Review* 60, 291. *https://doi.org/10.2307/1403680*
11. Kale, S., Kumar, R., Vassilvitskii, S. (2011) Cross-validation and mean-square stability, *International Conference on Supercomputing* 487–495.

12. Kumar, R., Lokshtanov, D., Vassilvitski, S., Vattani, A. (2013) Near-optimal bounds for cross-validation via loss stability, *Proceedings of Machine Learning Research* 27–35.

13. Bengio, Y., Grandvalet, Y. (2004) No unbiased estimator of the variance of k-fold cross-validation, *Journal of Machine Learning Research* 5, 1089–1105.

14. Zhang, Y., Yang, Y. (2015) Cross-validation for selecting a model selection procedure, *Journal of Econometrics* 187, 95–112. *https://doi.org/10.1016/j.jeconom.2015.02.006*

15. Marcot, B., Hanea, A. (2021) What is an optimal value of k in k-fold cross-validation in discrete bayesian network analysis? *Computational Statistics* 36, 2009–2031. *https://doi.org/10.1007/s00180-020-00999-9*

16. Jung, Y., Hu, J. (2015) A K-fold Averaging Cross-validation Procedure. *J Nonparametr Stat* 27(2), 167–179. *https://doi.org/10.1080/10485252.2015.1010532*

17. Satopaa, V.A., Albrecht, J.R., Irwin, D.E., Raghavan, B. (2011) Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior, *31st International Conference on Distributed Computing Systems Workshops* 31, 166–171. *https://doi.org/10.1109/ICDCSW.2011.20*

18. Yang, Y. (2006) Comparing learning methods for classification, *Statistica Sinica* 16, 635–657.

19. Kutner, M.H., Nachtsheim, C.J., Neter, J., Li, W. (2004) *Applied Linear Statistical Models* 5th ed., McGraw-Hill Higher Education, Yew York.
An Introduction to Statistical Learning 2nd ed., Springer, New York.

20. Yates, L.A., Aandahl, Z., Richards, S.A., Brook, B.W. (2023) Cross validation for model selection: A review with examples from ecology, *Ecological Monographs* 93, e1557. *https://doi.org/10.1002/ecm.1557*

ABOUT THE STUDENT AUTHOR

Angelos Vasilopoulos graduated from Loyola University Chicago in 2023 with a degree in statistics and is currently a medical student at the Stritch School of Medicine.

PRESS SUMMARY

In the field of data science, k-fold cross-validation is a popular method for estimating model error and selecting optimal models. It involves splitting a dataset into k parts, or folds, training a model on each of those parts, and averaging the models' respective errors as an estimate of "true" model error. What has received limited attention in the literature is the following question: what is the optimal number of parts to split a dataset into for the purposes of error estimation and model selection? Here we explore this question and simulate the results of different fold number selections in two different model settings. We suggest that there may be a predictable relationship between optimal values of k and n.