Simulations of fold number selection in cross-validation of linear and LASSO regression

Angelos Vasilopoulos Gregory J. Matthews

Abstract

k-fold cross-validation is a popular method of error estimation for 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 a growing body of 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

1 Introduction

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

$$\hat{\theta} = CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\widehat{\text{PE}}_i$$

is meant to estimate the true model error θ, i.e., the error of the model tested on the population (Bates, Hastie, and Tibshirani (2022)).

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 $Bias(\hat{\theta}) = \text{E}(\hat{\theta}) - \theta$ and variance $Var(\hat{\theta}) = \text{E}(\hat{\theta}^2) - \text{E}(\hat{\theta})^2$ of k-fold error estimation decrease and increase, respectively. This idea appears in early

literature (Efron (1983); Kohavi (2001)) but also in modern, widely used textbooks (James et al. (2021)).

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 (Burman (1989)) and that bias and variance of error estimation are uniformly low (Breiman and Spector (1992)).

More recently have been proposed ways to quantify the variance reduction achieved by cross-validation when the true prediction error (PE) is not known, e.g., as mean-square stability (Kale, Kumar, and Vassilvitskii (2011)) or as loss stability (Kumar et al. (2013)). 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 (Bengio and Grandvalet (2004)).

In addition to theoretical analyses of variance reduction by cross-validation, there are some simulation results showing this phenomenon. However, simulations currently in the literature provide limited insight into the dependence of optimal fold number k_{optimal} on sample size n (Zhang and Yang (2015)) or involve biased variance calculations (Marcot and Hanea (2021)). 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 of a known population.

2 Simulation

In machine learning tasks, selection of k for k-fold cross validation is largely an arbitrary decision between, e.g., 5 and 10. With the results of the following simulation, we make a less arbitrary recommendation of k with the aim of improving model selection accuracy.

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 mean squared error (MSE) is

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

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

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

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 $k \in \{2, 10, 20, 30, \ldots, n\}$ for each $n \in \{100, 200, 300, \ldots, 1000\}$. We perform this simulation 1000 times and then calculate simulation-wise 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 $k \in \{2, 10, 20, 30, \ldots, 100, n\}$ for each $n \in \{250, 500, 750, 1000\}$. To avoid data leakage, we introduce an inner 5-fold cross-validation loop for the selection of parameter λ as in the minimization of

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

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 $\widehat{\text{PE}}$ is more likely to be selected. 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 MSE($\hat{\theta}$) such that a competing model $\hat{f} \in F$ has a lower $\widehat{\text{PE}}$ than the true model f. This false model may have the lowest $\widehat{\text{PE}}$ 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 $\widehat{\text{PE}}$.

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}.

To estimate $k_{\text{optimal}}*$ for each n, we first fit to the dataset (k, A) for each sample size n a model of the form

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

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

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

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

$$ax + by + c = 0$$

is

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

so

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

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

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

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

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

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

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

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

3 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 (Kohavi (2001)).

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).

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 (Zhang and Yang (2015)), it is important to distinguish between possible goals of cross-validation that previously have been conflated (Kohavi (2001)): estimation of PE, in which case $k_{\text{optimal}} = \arg\min_k(\text{PE} - \widehat{\text{PE}})$, and model selection, in which case

$$k_{\text{optimal}} = \arg\max_k(A)$$

, 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 (Yang (2006)): 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).

4 Conclusion

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.

5 Appendix

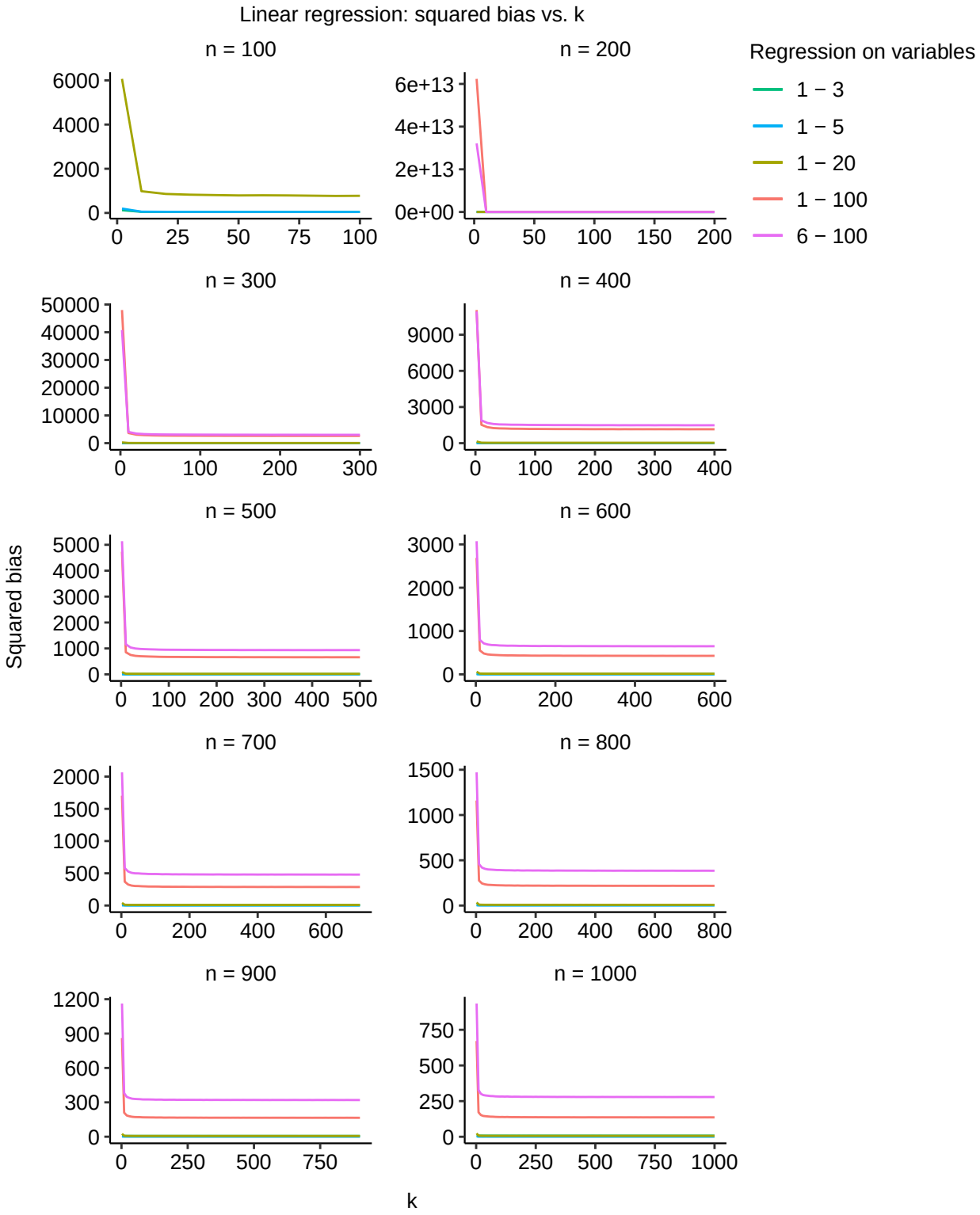

Figure 1. Linear regression squared bias vs. fold number for various sample sizes. As fold number increases, bias decreases initially before leveling off.

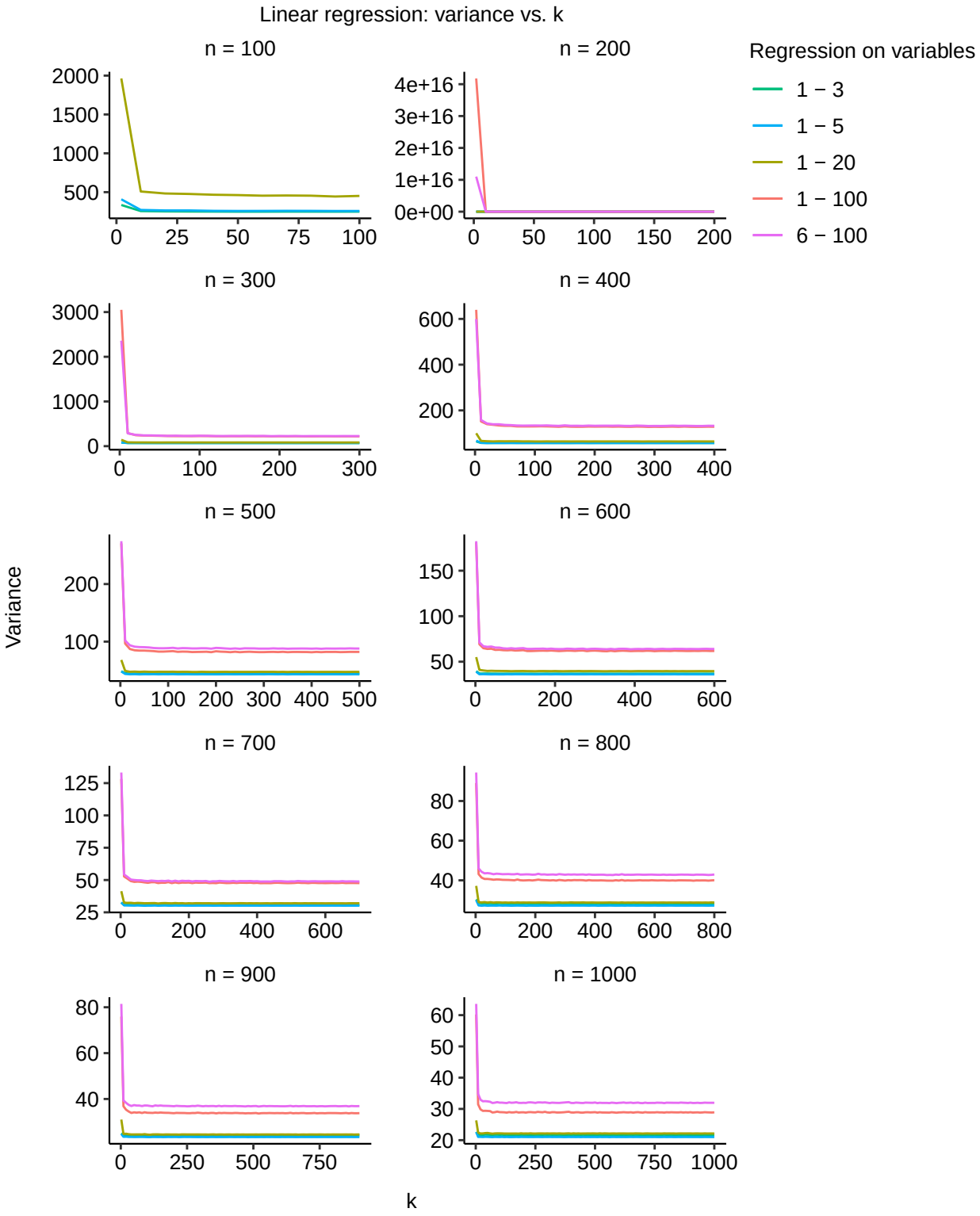

Figure 2. Linear regression variance vs. fold number for various sample sizes. As fold number increases, variance decreases initially before leveling off.

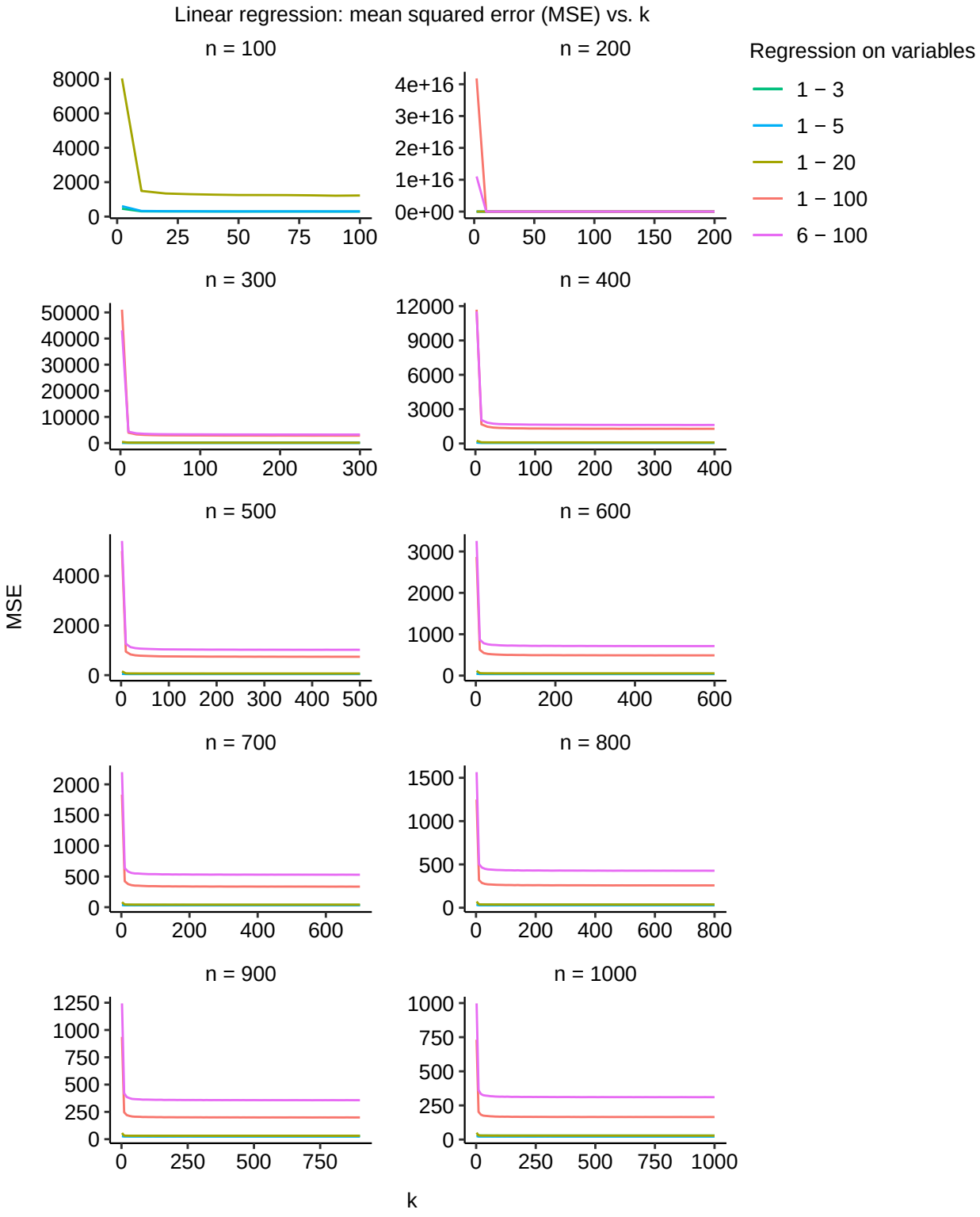

Figure 3. Linear regression mean squared error (MSE) vs. fold number for various sample sizes. As fold number increases, MSE decreases initially before leveling off.

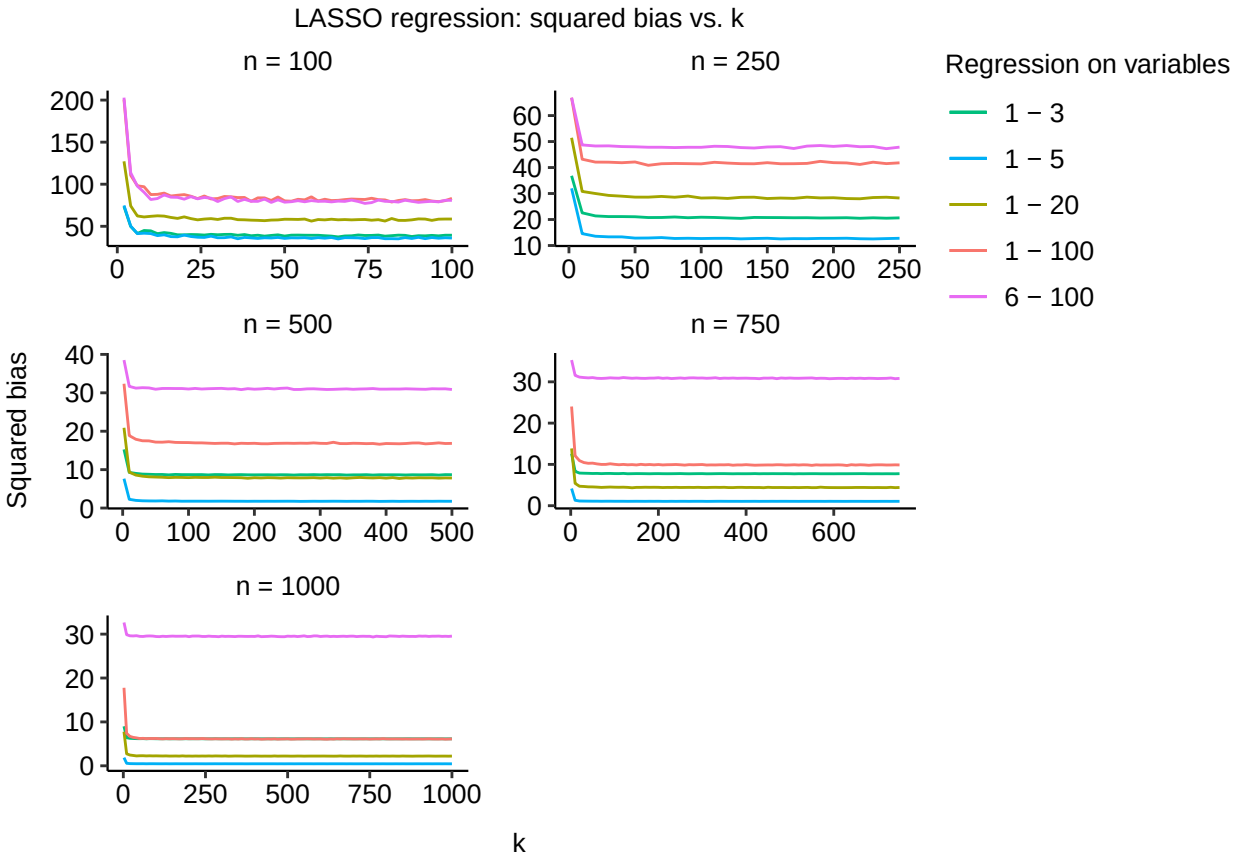

Figure 4. LASSO regression squared bias vs. fold number for various sample sizes. As fold number increases, bias decreases initially before leveling off.

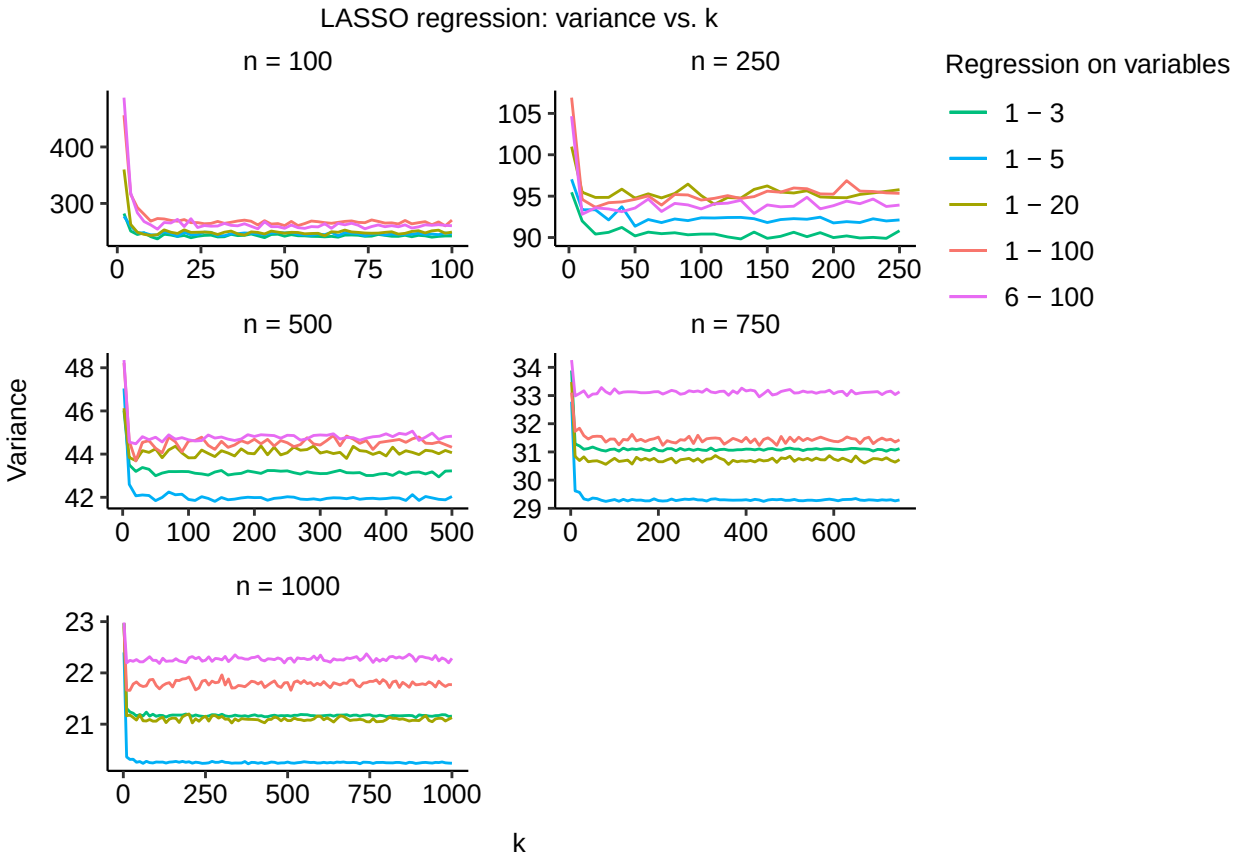

Figure 5. LASSO regression variance vs. fold number for various sample sizes. As fold number increases, variance decreases initially before leveling off.

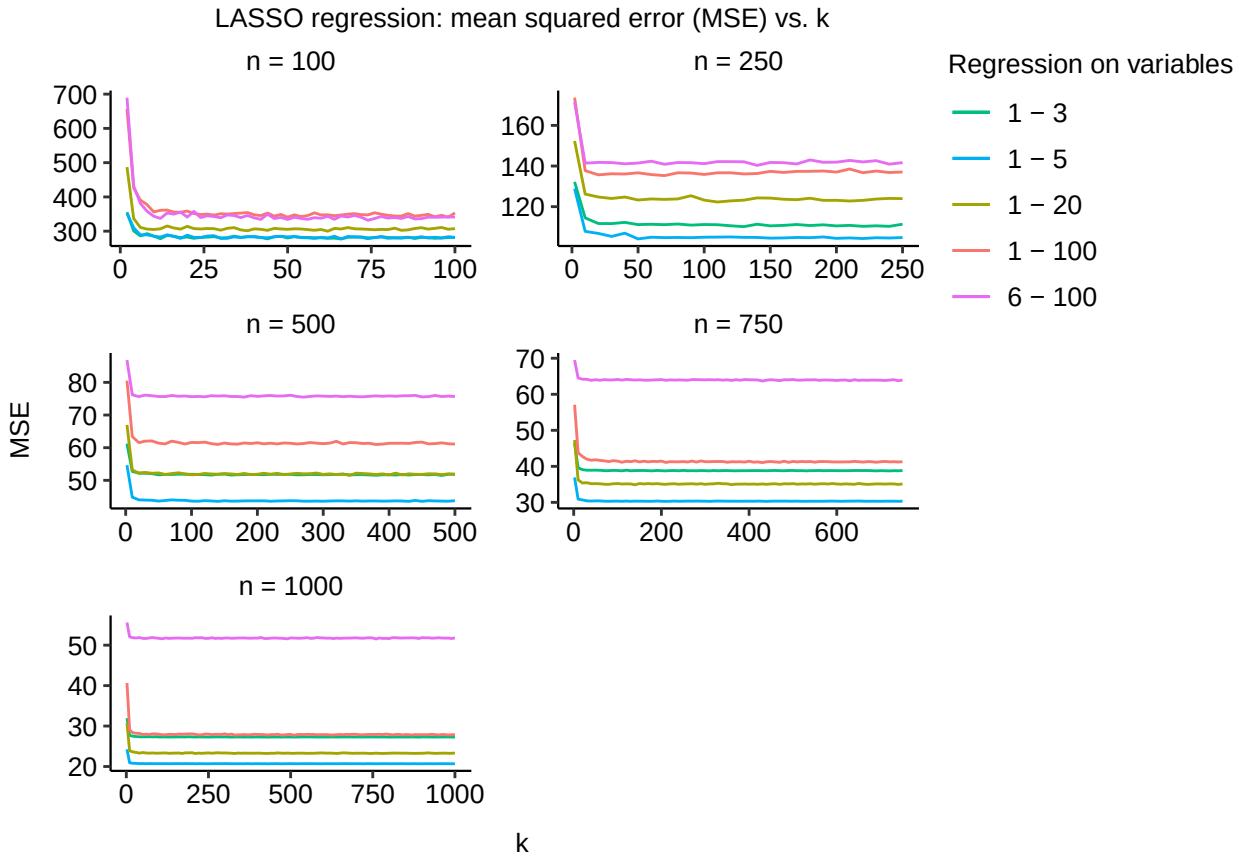

Figure 6. LASSO regression mean squared error (MSE) vs. fold number for various sample sizes. As fold number increases, MSE decreases initially before leveling off.

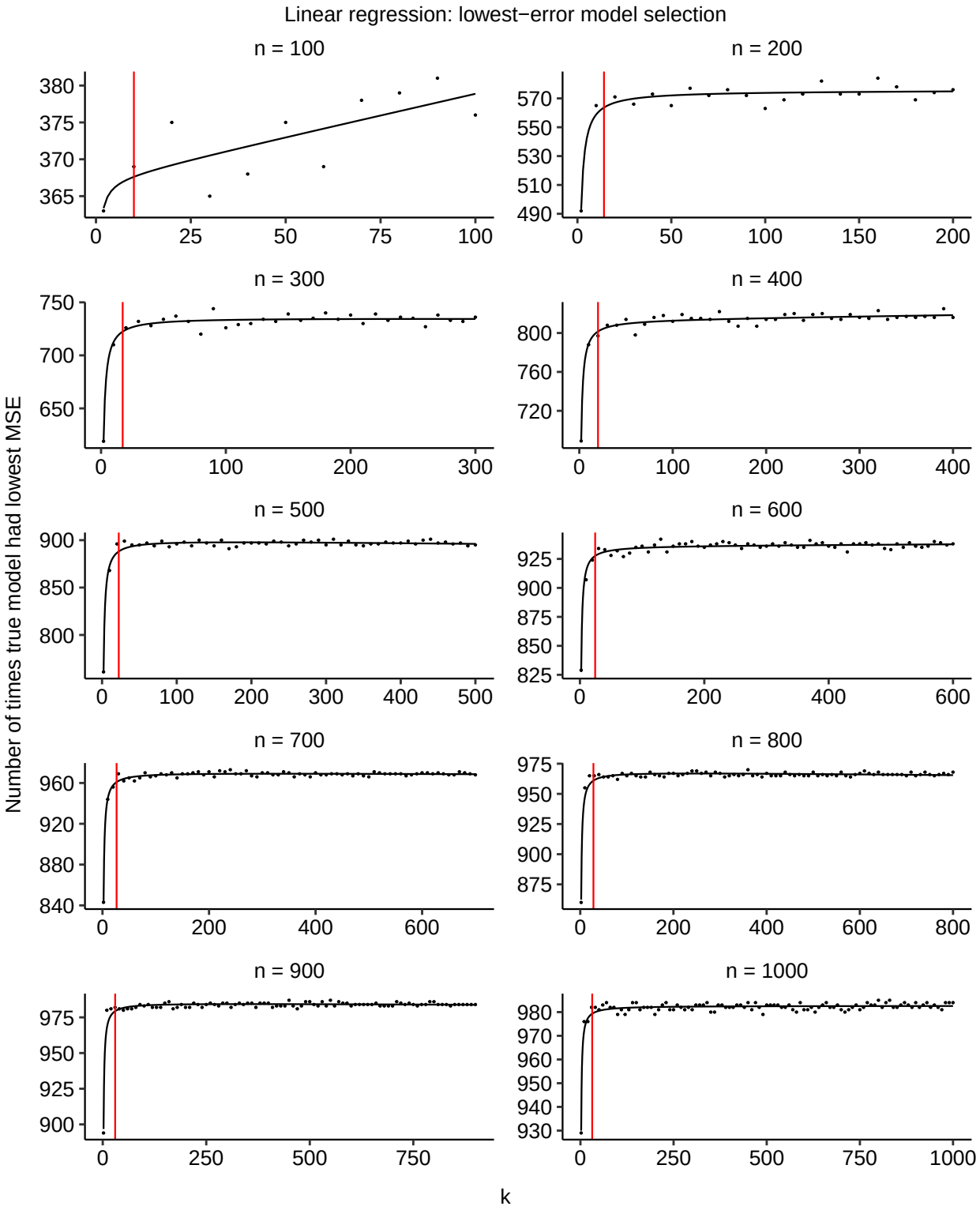

Figure 7. Count of true linear regression model selection by the lowest-error criterion vs. fold number for various sample sizes. Red line indicates optimal value of k.

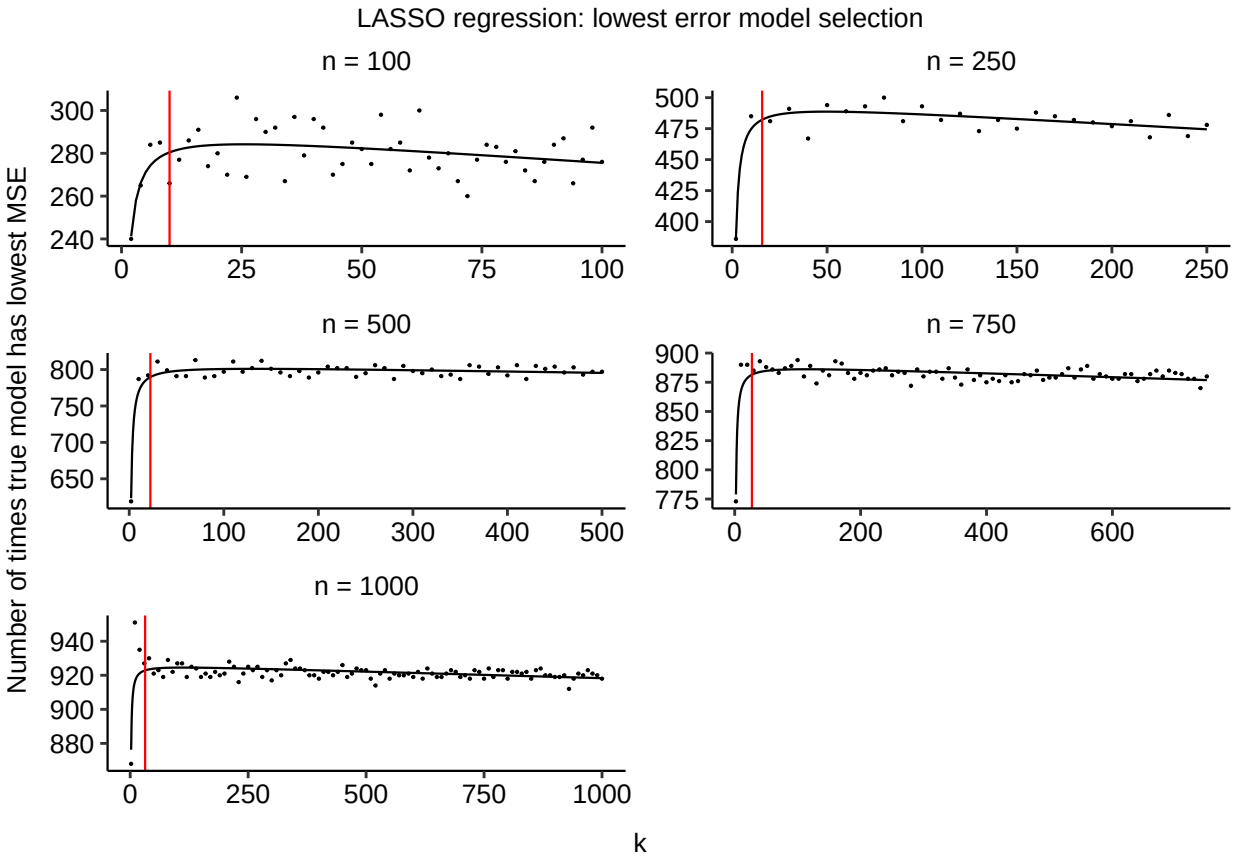

Figure 8. Count of true LASSO regression model selection by the lowest-error criterion vs. fold number for various sample sizes. Red line indicates optimal value of k.

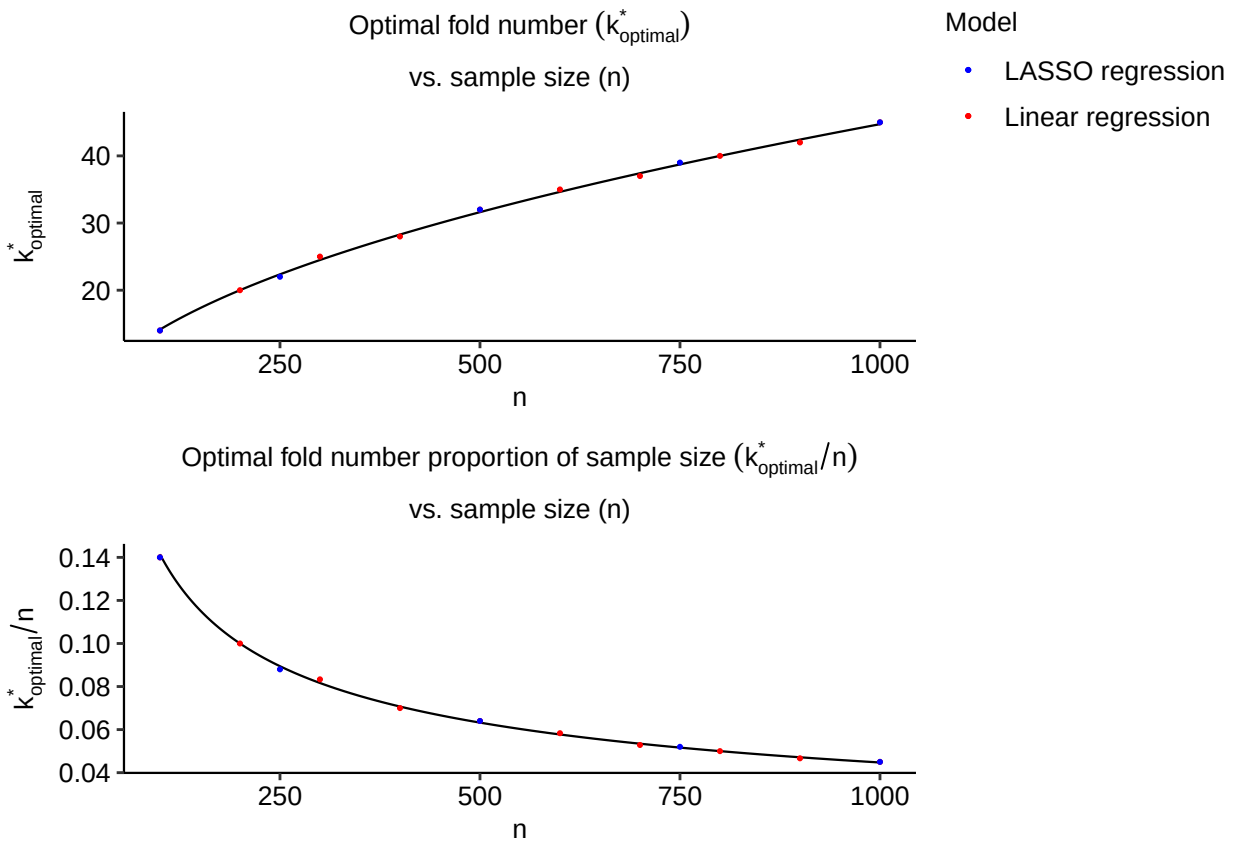

Figure 9. Raw optimal fold number vs. sample size (top); fold number proportion of sample size vs. sample size (bottom).

Acknowledgements

Dr. Gregory Matthews's Fall 2022 Predictive Analytics class.

Supplementary Material

All code for reproducing the analyses in this paper is publicly available at https://github.com/gjm112/optimal_k.

References

Bates, Stephen, Trevor Hastie, and Robert Tibshirani. 2022. "Cross-Validation: What Does It Estimate and How Well Does It Do It?" https://arxiv.org/abs/2104.00673.

Bengio, Yoshua, and Yves Grandvalet. 2004. "No Unbiased Estimator of the Variance of k-Fold Cross-Validation." *Journal of Machine Learning Research* 5 (June): 1089–1105.

Breiman, Leo, and Philip Spector. 1992. "Submodel Selection and Evaluation in Regression. The x-Random Case." *International Statistical Review / Revue Internationale de Statistique* 60 (December): 291.

Burman, Prabir. 1989. "A Comparative Study of Ordinary Cross-Validation, v-Fold Cross-Validation and the Repeated Learning-Testing Methods." *Biometrika* 76 (September): 503–14.

Efron, Bradley. 1983. "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation." *Journal of the American Statistical Association* 78 (March): 316–31.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning*. 2nd ed. Springer.

Kale, Satyen, Ravi Kumar, and Sergei Vassilvitskii. 2011. "Cross-Validation and Mean-Square Stability." In *International Conference on Supercomputing*, 487–95.

Kohavi, Ron. 2001. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." *International Joint Conference on Artificial Intelligence* 14 (March).

Kumar, Ravi, Daniel Lokshtanov, Sergei Vassilvitskii, and Andrea Vattani. 2013. "Near-Optimal Bounds for Cross-Validation via Loss Stability." In *Proceedings of the 30th International Conference on Machine Learning*, 28:27–35. Proceedings of Machine Learning Research. PMLR.

Marcot, Bruce, and Anca Hanea. 2021. "What Is an Optimal Value of k in k-Fold Cross-Validation in Discrete Bayesian Network Analysis?" *Computational Statistics* 36 (September): 2009–31.

Yang, Yuhong. 2006. "Comparing Learning Methods for Classification." *Statistica Sinica* 16 (April): 635–57.

Zhang, Yongli, and Yuhong Yang. 2015. "Cross-Validation for Selecting a Model Selection Procedure." *Journal of Econometrics* 187 (February): 95–112.