Predictive Model Degrees of Freedom in Linear Regression

Yoonkyung Lee*
Department of Statistics
The Ohio State University
*joint work with Bo Luan and Yunzhang Zhu

December 9, 2022
The Symposium on Statistics and Risk Management 2022
Department of Statistics
The Chinese University of Hong Kong
Collaborators

Bo Luan  
@ Google

Yunzhang Zhu  
@ Statistics, OSU
Regression model:
\[ y_i = \mu(x_i; \beta) + \varepsilon_i, \quad i = 1, \ldots, n \]

Data: \( X = (x_1, \ldots, x_n)^\top \) and \( y = (y_1, \ldots, y_n)^\top \)

How to define a measure of model complexity for a fitted model \( \hat{\mu} \)?
Review: Model DF

- Model degrees of freedom

  - Least squares linear regression with \( p \) predictors (including intercept):
    \[
    \text{df} = p
    \]

  - Ridge regression or smoothing spline with penalty parameter \( \lambda \):
    \[
    \text{df} = \text{tr}(H_\lambda)
    \]
    with hat matrix \( H_\lambda \) (i.e., \( \hat{\mu} = H_\lambda y \))
Review: Model DF

- Covariance penalty representation of model DF:

\[
df(\hat{\mu}) = \sum_{i=1}^{n} \frac{\text{Cov}(y_i, \hat{\mu}_i|X)}{\sigma^2_{\varepsilon}}
\]

- Classical optimism theory (Efron 1986, 2004):

  - Training error of \( \hat{\mu} \):
    \[
    \text{Err}_{T_{X,Y}} = \frac{1}{n} \| y - \hat{\mu} \|^2
    \]
  
  - In-sample prediction error of \( \hat{\mu} \):
    \[
    \text{Err}_{F_{X,Y}} = \frac{1}{n} \mathbb{E}_{\tilde{y}}(\| \tilde{y} - \hat{\mu} \|^2|X, y)
    \]
  
  - Expected optimism of \( \hat{\mu} \):
    \[
    \text{Opt}_{F_X} = \mathbb{E}_y(\text{Err}_{F_{X,Y}}) - \mathbb{E}_y(\text{Err}_{T_{X,Y}}) = \frac{2}{n} \sum_{i=1}^{n} \text{Cov}(y_i, \hat{\mu}_i|X)
    \]
Ye (1998) defined the generalized model degrees of freedom of $\hat{\mu}$ as

$$\text{GDF}(\hat{\mu}) = \sum_{i=1}^{n} \frac{\partial E(\hat{\mu}_i|X)}{\partial \mu_i}$$
Interpolating Models

- Modeling with high dimensional data in an overparameterized regime
- Deep learning with gradient descent
- Benign overfitting (Belkin et al. 2019, Bartlett et al. 2022, Hastie et al. 2019)

source: www.ibm.com
Double Descent Phenomenon

**Figure:** Risk curves for training and test data as a function of model capacity. Figure is from Belkin et al. (2019), *Reconciling modern machine-learning practice and the classical bias–variance trade-off.* PNAS.
Double Descent Phenomenon

Figure: Prediction risk for (min norm) least squares fit when sample size $n = 20$ and the number of variables $d = 100$. 
Interpolating models produce $\hat{\mu}_i = y_i$

According to the classical model DF, interpolating models have

$$\text{df}_F = n$$

As $p$ increases, the nominal model size increases while the minimum norm restriction regularizes the model space.
Figure: The blue line is for the classical model df, and the red line is for the predictive model df.
Consider out-of-sample prediction for difference among models

Extend the optimism theory for out-of-sample prediction to define a new notion of model complexity

For a new test point $x_*$, let $\mu_* = \mu(x_*; \beta)$, and $y_* = \mu_* + \varepsilon_*$. Let $\hat{\mu}_*$ be the prediction of $y_*$ with a fitted model $\hat{\mu}$.

The out-of-sample prediction error of model $\hat{\mu}$:

$$\text{ErrR}_{x,y} = E(x_*,y_*)[(y_* - \hat{\mu}_*)^2|x, y]$$
Expected Optimism for Out-Of-Sample Prediction

- Letting $\text{ErrR}_\mathbf{X} = E_y(\text{ErrR}_{\mathbf{X}, y})$,

  $\text{OptR}_\mathbf{X} = \text{ErrR}_\mathbf{X} - \text{ErrT}_\mathbf{X}$

- For a linear procedure with hat matrix $\mathbf{H}$ such that $\hat{\mu} = \mathbf{H}y$, define the hat vector $\mathbf{h}_*$ at $\mathbf{x}_*$:

  $\hat{\mu}_* = \mathbf{h}_*^\top \mathbf{y}$

- Bias-variance decomposition of squared error:

  $\text{ErrR}_\mathbf{X} = \sigma_\varepsilon^2 + E[(\mu_* - \mathbf{h}_*^\top \mu)^2 | \mathbf{X}] + \sigma_\varepsilon^2 E(\|\mathbf{h}_*\|^2 | \mathbf{X})$

  $\text{ErrT}_\mathbf{X} = \sigma_\varepsilon^2 + \frac{1}{n} \|\mu - \mathbf{H}\mu\|^2 + \frac{1}{n} \sigma_\varepsilon^2 \text{tr}(\mathbf{H}^\top \mathbf{H} - 2\mathbf{H})$
Predictive Model DF

- Out-of-sample expected optimism:
  \[ \text{OptR}_X = \Delta B_X + \frac{2}{n} \sigma^2 \varepsilon \left[ \text{tr}(H) + \frac{n}{2} \left( \mathbb{E}(\|h^*\|^2 | X) - \frac{1}{n} \text{tr}(H^\top H) \right) \right], \]

  where \( \Delta B_X = \mathbb{E}[(\mu^* - h^\top \mu)^2 | X] - \frac{1}{n} \|\mu - H\mu\|^2 \)

  (excess bias)

- Predictive model degrees of freedom:
  \[ \text{df}_R = \text{tr}(H) + \frac{n}{2} \left( \mathbb{E}(\|h^*\|^2 | X) - \frac{1}{n} \text{tr}(H^\top H) \right) \]

  \[ \text{df}_F \]

  out-of-sample adjustment
Representations of $\text{df}_R - \text{df}_F$

- Covariance penalty representation of $\text{df}_R - \text{df}_F$:

$$\frac{n}{2} \sum_{i=1}^{n} \left[ E \left( \frac{\text{Cov}^2(y_i, \hat{\mu}_* | x_*, X)}{(\sigma^2\varepsilon)^2} \right) \left| X \right) - \frac{1}{n} \sum_{j=1}^{n} \text{Cov}^2(y_i, \hat{\mu}_j | X) \right]$$

- GDF representation of $\text{df}_R - \text{df}_F$:

$$\frac{n}{2} \sum_{i=1}^{n} \left[ E \left( \left( \frac{\partial E(\hat{\mu}_* | x_*, X)}{\partial \mu_i} \right)^2 \right) \left| X \right) - \frac{1}{n} \sum_{j=1}^{n} \left( \frac{\partial E(\hat{\mu}_j | X)}{\partial \mu_i} \right)^2 \right]$$
Least Squares Method in Subset Regression

- $S$: a subset of variable indices $\subset \{1, \ldots, d\}$ with $|S| = p$

- $X_S = (x_{(j)})_{j \in S} \in \mathbb{R}^{n \times p}$: the submatrix of $X$ for $j \in S$

- $\Sigma_S = \text{Var}(x_{i,S}) \in \mathbb{R}^{p \times p}$: the submatrix of $\Sigma$ for $j \in S$

- Estimate $\beta$ in subset regression by least squares method when $p \leq n$ and by solving

$$
\min_{b \in \mathbb{R}^p} \|b\|_2^2, \quad \text{subject to } y = X_S b
$$

when $p > n$:

$$
\hat{\beta}(S) = \begin{cases} 
(X_S^T X_S)^{-1} X_S^T y, & p \leq n \\
X_S^T (X_S X_S^T)^{-1} y, & p > n
\end{cases}
$$
Predictive DF in Subset Regression

- For a test point $x_\ast \in \mathbb{R}^d$, 
  $$\hat{\mu}_\ast = x_\ast^\top S \hat{\beta}(S) = h_\ast^\top y$$
  with the hat vector 
  $$h_\ast = \begin{cases} 
  x_S (X_S^\top X_S)^{-1} x_\ast, & p \leq n \\
  (X_S X_S^\top)^{-1} X_S x_\ast, & p > n 
  \end{cases}$$

- The predictive model df for subset regression:
  $$df_R(S) = \begin{cases} 
  \frac{p}{2} + \frac{n}{2} \text{tr}[(X_S^\top X_S)^{-1} \Sigma_S], & p \leq n \\
  \frac{n}{2} + \frac{n}{2} \text{tr}[X_S^\top (X_S X_S^\top)^{-2} X_S \Sigma_S], & p > n 
  \end{cases}$$
Normal Covariates (Underparameterized)

- If \( x_i \sim \mathcal{N}(0, \Sigma) \), for \( S \) with \(|S| = p < n - 1\)

\[
E(\text{tr}[(X_S^\top X_S)^{-1}\Sigma_S]) = \frac{p}{n-p-1}
\]

- \( E[\text{df}_R(S)] = \frac{p}{2} + \frac{n}{2}E(\text{tr}[(X_S^\top X_S)^{-1}\Sigma_S]) = \frac{p}{2} \left(1 + \frac{n}{n-p-1}\right)\)

- The increment in \( \text{df}_R \):

\[
E[\text{df}_R(p + 1)] - E[\text{df}_R(p)] = \frac{1}{2} + \frac{n(n-1)}{2(n-p-1)(n-p-2)},
\]

which is strictly increasing in \( p \)
Normal Covariates (Overparameterized)

- If \( x_i \sim \mathcal{N}(0, I_p) \), for \( p > n + 1 \)

\[
E[\text{tr}(XX^\top)^{-1})] = \frac{n}{p - n - 1}
\]

- \( E(\text{df}_R) = \frac{n}{2} + \frac{n}{2} \cdot E[\text{tr}(XX^\top)^{-1})] = \frac{n(p - 1)}{2(p - n - 1)}, \)

which is decreasing in \( p \)
Example: Normal Covariates

Figure: The out-of-sample prediction error as a function of $p$ when $n = 20$, $d = 100$, $x_i \sim \mathcal{N}(0, I_d)$ and $\beta_j \propto \frac{1}{j}$
Example: Normal Covariates

Figure: The out-of-sample prediction error as a function of $\log(df_R)$ when $n = 20$, $d = 100$, $x_i \sim \mathcal{N}(0, I_d)$ and $\beta_j \propto \frac{1}{j}$.
Asymptotic Approximation of $df_R$

From Hastie et al. (2019), under some conditions on the spectrum of $\Sigma$ as $n$ and $p \to \infty$ and $\frac{p}{n} \to \gamma$,

- For $\gamma < 1$,
  $$\operatorname{tr}[(X^\top X)^{-1} \Sigma] \xrightarrow{\text{a.s.}} \frac{\gamma}{1 - \gamma}$$

- When $p < n$, substituting $\frac{p}{n}$ for $\gamma$ in $df_R$ yields
  $$df_R \approx \frac{p}{2} + \frac{n}{2} \cdot \frac{p/n}{1 - p/n} = \frac{p}{2} \left(1 + \frac{n}{n - p}\right)$$
Asymptotic Approximation of $df_R$

From Hastie et al. (2019), under some conditions on the spectrum of $\Sigma$ as $n$ and $p \to \infty$ and $\frac{p}{n} \to \gamma$,

- For $\gamma > 1$, it is not easy to express the limit of $\text{tr}[X^T (XX^T)^{-2} X \Sigma]$ explicitly in general.

- When $\Sigma = (1 - \rho)I_d + \rho 11^\top$ with $0 \leq \rho < 1$,

\[
\text{tr}[X^T (XX^T)^{-2} X \Sigma] \xrightarrow{\text{a.s.}} \frac{1}{\gamma - 1}
\]

- Replacing $\gamma$ with $\frac{p}{n}$,

\[
df_R \approx \frac{n}{2} + \frac{n}{2} \cdot \frac{1}{p/n - 1} = \frac{n}{2} \left(1 + \frac{n}{p - n}\right)
\]
Figure: Predictive model df $df_R$ versus the number of variables $p$ under the equal correlation setting of $\Sigma = \frac{1}{2} I_d + \frac{1}{2} 11^\top$. $x_i \overset{iid}{\sim} \mathcal{N}(0, \Sigma)$ with $n = 20, \ d = 100$. The gray lines are $df_R$ based on 100 randomly ordered variable sequences, and the red line is the approximate df.
We can develop out-of-sample prediction error estimators for the least squares method using the predictive model df:

\[
\text{Err}_{R_X} = \text{Err}_{T_X} + \Delta B_X + \frac{2}{n} \sigma^2 \epsilon_{df_R}
\]

Using the LOOCV identity for estimation of \(\Delta B_X\),

\[
\widehat{\text{Err}}_{R_X} = \text{Err}_{T_{X,y}} + \Delta B_X + \frac{2}{n} \sigma^2 \epsilon_{df_R}
\]

\[
= \widehat{\text{Err}}_{R_{\text{loocv}}} + \frac{1}{n} \sigma^2 \epsilon(2 \, df_R - \text{tr}(A)),
\]

where \(A = (I_n - H)^\top D(I_n - H)\) and \(D = \text{diag} \left( \frac{1}{(1-h_{ii})^2} - 1 \right)\).
Figure: Comparison of $\widehat{\text{Err}}_{\text{loocv}}$ and $\widehat{\text{Err}}_{+}$ under two different mean functions. $n = 50$, $d = 120$, $x_i \sim \mathcal{N}(0, I_d)$ and $\epsilon_i \sim \mathcal{N}(0, 1)$. The gray lines are the estimates for 500 random replicates of $(X, y)$. The black line is the average true prediction error.
Comparison of Prediction Error Estimates

The relative mean squared error of \( \hat{\text{ErrR}}_+ \) to \( \hat{\text{ErrR}}_{\text{loocv}} \):

\[
\Pi(p) = \frac{\sum_m [\hat{\text{ErrR}}_+(m)(p) - \text{ErrR}(m)(p)]^2}{\sum_m [\hat{\text{ErrR}}_{\text{loocv}}(m)(p) - \text{ErrR}(m)(p)]^2}.
\]

**Figure:** The relative mean squared error of \( \hat{\text{ErrR}}_+ \) to \( \hat{\text{ErrR}}_{\text{loocv}} \)
Comparison of Selected Model Size

**Figure**: Comparison of $\hat{\text{Err}}_{\text{loocv}}$, 5-fold cross validation and $\hat{\text{Err}}_{\text{R+}}$ in model selection. $p_*$ is the minimizer of the true prediction error $\text{Err}R$ for each replicate.
Concluding Remarks

- Defined the new concept of predictive model df for linear procedures

- Provided insights into the “double descent” phenomenon

- How to estimate the predictive model df using the training data efficiently?

- How to generalize the current framework for heteroscedastic data?

- How to extend the framework to generalized linear models with a general loss function?
Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler.
Benign overfitting in linear regression. 

Mikhail Belkin, Daniel Hsu, and Ji Xu.
Two models of double descent for weak features.

Bradley Efron.
How biased is the apparent error rate of a prediction rule?
References II

Bradley Efron.
The estimation of prediction error: Covariance penalties and cross-validation.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani.
Surprises in high-dimensional ridgeless least squares interpolation.

Bo Luan, Yoonkyung Lee, and Yunzhang Zhu.
Predictive model degrees of freedom in linear regression.
Bo Luan, Yoonkyung Lee, and Yunzhang Zhu. 
On measuring model complexity in heteroscedastic linear regression.  

Jianming Ye. 
On measuring and correcting the effects of data mining and model selection.  