Applied Statistics Qualifying Exams Coaching

Michael Howes*

Summer 2023

Contents

*With lots of content credit given to previous applied quals coaches

0 Deviance and model comparison

0.1 Deviance between two parameter vectors

Let's return to the set-up of Section **??**. That is, suppose we have a one-dimensional exponential family,

$$f_\eta(y) = \exp\left(\eta y - \psi(\eta)\right) f_0(y), \tag{1}$$

Consider now a case where we repeatedly sample from (1), but we let the parameter η vary with each sample. That is we have the model,

$$y_i \overset{\text{ind}}{\sim} f_{\eta_i} \quad 1 \le i \le n, \tag{2}$$

which is parametrized by $\eta \in \mathbb{R}^n$. We will let f_η denote the density for $Y = (y_1, \ldots, y_n)$. The log-likelihood in this model is

$$\ell(\eta; Y) = \log f_\eta(Y) = \sum_{i=1}^n \eta_i y_i - \sum_{i=1}^n \psi(\eta_i) = \eta^\top Y - \sum_{i=1}^n \psi(\eta_i).$$

The *deviance* between two parameter vectors $\eta^{(1)}, \eta^{(2)} \in \mathbb{R}^n$ is given by two times the KL divergence between $f_{\eta^{(1)}}$ and $f_{\eta^{(2)}}$. That is,

$$\begin{aligned} D(\eta^{(1)}, \eta^{(2)}) &= 2\mathbb{E}_{\eta^{(1)}}\left[\log f_{\eta^{(1)}}(Y) - \log f_{\eta^{(2)}}(Y)\right] \\ &= 2\mathbb{E}_{\eta^{(1)}}\left[(\eta^{(1)})^\top Y - \sum_{i=1}^n \psi(\eta_i^{(1)}) - (\eta^{(2)})^\top Y + \sum_{i=1}^n \psi(\eta^{(2)})\right] \\ &= 2\left((\eta^{(1)} - \eta^{(2)})^\top \mu^{(1)} - \sum_{i=1}^n \psi(\eta_i^{(1)}) - \psi(\eta_i^{(2)})\right). \end{aligned}$$

Note that our independence assumption implies that the deviance is additive. That is,

$$D(\eta^{(1)}, \eta^{(2)}) = \sum_{i=1}^n D(\eta_i^{(1)}, \eta_i^{(2)}).$$

If we set $\mu_i = \mathbb{E}_{\eta_i}[y_i]$, then we could parametrize the above model and likelihood by $\mu \in \mathbb{R}^n$. That is

$$f_\mu = f_\eta \quad \text{where } \eta \text{ solves } \mathbb{E}_\eta[Y] = \mu.$$

The deviance between two mean vectors $\mu^{(1)}$ and $\mu^{(2)}$ is then defined to be

$$D(\mu^{(1)}, \mu^{(2)}) = D(\eta^{(1)}, \eta^{(2)}).$$

The model (2) without any constraints on η is called the *saturated model*. The MLE in the model is found by matching the expectation of y_i to the observed value. That is,

$$\hat{\eta}^{\text{sat}} \text{ solves } \mathbb{E}_{\hat{\eta}^{\text{sat}}}[Y] = Y.$$

Or equivalently $\hat{\mu}^{\text{sat}} = Y$. This observation leads to *Hoeffding's formula* which states that for all $\eta \in \mathbb{R}^n$

$$\frac{f_\eta(Y)}{f_{\hat{\eta}^{\text{sat}}}(Y)} = \exp\left(-\frac{1}{2}D(\hat{\eta}^{\text{sat}}, \eta)\right) = \exp\left(-\frac{1}{2}\sum_{i=1}^{n} D(\hat{\eta}_i^{\text{sat}}, \eta_i)\right). \tag{3}$$

If we use the mean parametrization, then Hoeffding's formula says that for all $\mu \in \mathbb{R}^n$,

$$\frac{f_\mu(Y)}{f_Y(Y)} = \exp\left(-\frac{1}{2}D(Y, \mu)\right) = \exp\left(-\frac{1}{2}\sum_{i=1}^{n} D(Y_i, \mu_i)\right). \tag{4}$$

On the log-scale, Hoeffding's formula says that

$$\log f_\eta(Y) = \log f_{\hat{\eta}^{\text{sat}}}(Y) - \frac{1}{2}D(\hat{\eta}^{\text{sat}}, \eta).$$

That is, if move from the MLE $\hat{\eta}^{\text{sat}}$ to the parameter vector η then the log likelihood decreases by half the deviance. Proving Hoeffding's formula follows from the definition of the deviance,

$$\begin{aligned}
-\frac{1}{2}D(\hat{\eta}^{(\text{sat})}, \eta) &= -\left(\hat{\eta}^{(\text{sat})} - \eta\right)^\top \hat{\mu}^{\text{sat}} + \sum_{i=1}^{n} \psi(\hat{\eta}_i^{\text{sat}}) - \psi(\eta_i) \\
&= -\left(\hat{\eta}^{(\text{sat})} - \eta\right)^\top Y + \sum_{i=1}^{n} \psi(\hat{\eta}_i^{\text{sat}}) - \psi(\eta_i) \\
&= \eta^\top Y - \sum_{i=1}^{n} \psi(\eta_i) - \left((\hat{\eta}^{\text{sat}})^\top Y - \sum_{i=1}^{n} \psi(\hat{\eta}_i^{\text{sat}})\right) \\
&= \log f_\eta(Y) - \log f_{\hat{\eta}^{\text{sat}}}(Y).
\end{aligned}$$

0.2 Deviance in GLMs

Now suppose we have features $X_i \in \mathbb{R}^p$ for each observation i. To turn the saturated model (2) into a GLM, we put a linear constraint on the parameter vector η. Specifically, we assume that $\eta_i = X_i^\top \beta$ for some $\beta \in \mathbb{R}^p$. This gives the model

$$Y_i \overset{\text{ind}}{\sim} f_{\eta_i}, \quad \eta = X\beta, \tag{5}$$

where $X \in \mathbb{R}^{n \times p}$ has rows X_i. As discussed previously, the parameters β are fit by maximizing the likelihood. Specifically,

$$\hat{\beta} = \underset{\beta}{\text{argmax}}\, f_{X\beta}(Y).$$

If we let $\mu_i(\beta) = \mathbb{E}_{X_i^\top \beta}[Y_i]$, then the mean parametrization version of Hoeffding's formula gives

$$\begin{aligned}
\hat{\beta} &= \operatorname*{argmax}_\beta f_{\mu(\beta)}(Y) \\
&= \operatorname*{argmax}_\beta \frac{f_{\mu(\beta)}(Y)}{f_Y(Y)} \\
&= \operatorname*{argmax}_\beta \exp\left(-\frac{1}{2} D(Y, \mu(\beta))\right) \\
&= \operatorname*{argmin}_\beta D(Y, \mu(\beta)).
\end{aligned}$$

Thus, maximizing the likelihood is equivalent to minimizing the deviance from the observed data Y to the fitted mean $\mu(\beta)$.

0.3 Difference of deviance

Suppose we have a decomposition of our feature matrix X into $X = [X^{(1)}, X^{(2)}]$ where $X_1 \in \mathbb{R}^{n \times p_1}$ and $X^{(2)} \in \mathbb{R}^{n \times p_2}$. This decomposition gives two nested models

$$\begin{aligned}
&\text{Sub model: } y_i \overset{\text{ind}}{\sim} f_{\eta_i}, \quad \eta = X_1 \beta_1, \\
&\text{Full model: } y_i \overset{\text{ind}}{\sim} f_{\eta_i}, \quad \eta = X_1 \beta_1 + X^{(2)} \beta_2,
\end{aligned}$$

where $\beta_1 \in \mathbb{R}^{p_1}$ and $\beta_2 \in \mathbb{R}^{p_2}$. To see if the sub model is a sufficient explanation of data, we can test the hypothesis

$$\mathcal{H}_0 : \beta_2 = 0.$$

The canonical test statistic for this null hypothesis is the *difference of deviance*,

$$T = D(Y, \mu(\hat{\beta}_{\text{sub}})) - D(Y, \mu(\hat{\beta}_{\text{full}})),$$

where $\hat{\beta}_{\text{sub}}$ and $\hat{\beta}_{\text{full}}$ are the MLE in the two submodels. Since these MLEs are found by minimizing the deviance we have

$$D(Y, \mu(\hat{\beta}_{\text{sub}})) \geq D(Y, \mu(\hat{\beta}_{\text{full}})),$$

and hence $T \geq 0$. Furthermore, maximum likelihood theory gives

$$T \dot{\sim} \chi^2_{p_2},$$

under $\mathcal{H}_0$. Thus, rejecting $\mathcal{H}_0$ when $T \geq \chi^2_{p_2}(1-\alpha)$ is a level α test of $\mathcal{H}_0$. If you have J nested models, then you can perform multiple difference of deviance tests to assess the fit of each model.

0.4 Deviance residuals

The deviance also gives a heuristic that can be used to assess goodness-of-fit in a single GLM. Suppose we have the model

$$Y_i \overset{\text{ind}}{\sim} f_{\eta_i}, \quad \eta = X\beta,$$

where $X \in \mathbb{R}^{n\times p}$. The deviance heuristic is

$$D(Y, \mu(\hat{\beta})) \approx \chi^2_{n-p}.$$

The above approximation may fail, but it is useful to compare the deviance $D(Y, \mu(\hat{\beta}))$ to $n - p = \mathbb{E}[\chi^2_{n-p}]$. If the deviance is much larger than $n - p$, then this suggests your model is missing something important. When this occurs, you can try to diagnose the problem by looking at the *deviance residuals*. These are defined by

$$R_i = \text{Sign}(Y - \mu_i(\hat{\beta}))\sqrt{D(Y_i, \mu_i(\hat{\beta}))},$$

and satisfy

$$\sum_{i=1}^{n} R_i^2 = D(Y, \mu(\hat{\beta})).$$

You can look for trends in R_i or abnormally large values of R_i. Here's a simple example,

Example 0.1 (Quasi-independence model)**.** Consider the below contingency table, taken from [Agresti, 2013],

Table 11.17 Occupational Mobility Data for Exercise 11.11

	Son's Status				
Father's Status	1	2	3	4	5
1	50	45	8	18	8
2	28	174	84	154	55
3	11	78	110	223	96
4	14	150	185	714	447
5	3	42	72	320	411

Source: Reprinted with permission from D. V. Glass (ed.), *Social Mobility in Britain*, Glencoe, IL: Free Press, 1954.

Each observation in this table corresponds to a father-son pair. The row and column variable refer to the father's and son's occupational status. Let Y_{ij} be the count in row i and column j. We could try a Poisson GLM for Y_{ij}. The model below assume independence between father and son status.

$$Y_{ij} \overset{\text{ind}}{\sim} \text{Pois}(\mu_{ij}), \quad \log(\mu_{ij}) = \alpha + \beta_i + \gamma_j. \tag{6}$$

	1	2	3	4	5
1	12.76	5.33	-2.42	-5.54	-5.85
2	2.99	10.55	2.26	-3.53	-8.48
3	-1.25	0.65	4.68	0.78	-4.76
4	-5.51	-4.43	-0.94	3.83	0.39
5	-5.70	-8.11	-3.98	-1.43	9.56

For identifiable, we add the constraints, $\beta_1 = \gamma_1 = 0$. The below table shows the deviance residuals form model (6) for each (i, j) pair, We see that there are big positive residuals along the main diagonal and large negative residuals in the top-right and bottom-left corners. Evidently there is a correlation between father and son status that our model is missing. Here is an alternative model we could try,

$$Y_{ij} \overset{\text{ind}}{\sim} \text{Pois}(\mu_{ij}), \quad \log(\mu_{ij}) = \alpha + \beta_i + \gamma_j + \delta_{i-j}. \tag{7}$$

That is we now have row parameters β_i, column parameters γ_j and "diagonal" parameters δ_{i-j}. To make the model identifiable we will require $\beta_1 = \gamma_1 = 0$ and $\delta_{-4} = \delta_4 = 0$. Here are the resulting deviance residuals, Although some of these residuals are still

	1	2	3	4	5
1	4.36	-0.85	-3.23	-1.28	-0.00
2	-1.82	0.13	-1.00	1.04	0.84
3	-2.17	-1.08	0.50	0.90	0.03
4	-1.26	0.75	1.31	-0.86	0.10
5	0.00	0.85	-0.02	0.19	-0.42

large, we see a substantial reduction and there is less of an obvious trend. Later we will talk about the AIC. The AIC for model (6) is 960 and the AIC for model (7) is 233.

0.5 Model selection

Suppose we have GLM models $M_1, M_2, \ldots, M_J$ for the same data Y. You can think of each M_j as corresponding to a different feature matrix $X^{(j)} \in \mathbb{R}^{n \times p_j}$. Choosing one of the models M_j is *model selection*.

We saw in Section **??** that cross validation can be used for model selection. This is a very flexible and powerful approach and should be your default if an exam questions asks how you would do model selection. However, there are other approaches and an exam question might specifically ask about one of them.

AIC and BIC

The testing based approach required our models to be nested. The AIC and BIC are two model selection methods that are similar but do not require the models to be

nested. Both of them are based on "penalized log-likelihood". Let's write $\ell_j(\beta)$ for the log-likelihood in model j and $\ell_j(\hat{\beta}^{(j)})$ for the maximized log-likelihood in model j. Then, the AIC and BIC are

- **AIC**: The Akaike information criterion for model M_j is

$$\text{AIC}^{(j)} = -2\ell_j(\hat{\beta}^{(j)}) + 2p_j.$$

  The AIC model is the one that minimizes $\text{AIC}^{(j)}$.

- **BIC**: The Bayes information criterion for model M_j is

$$\text{BIC}^{(j)} = -2\ell_j(\hat{\beta}^{(j)}) + \log(n)p_j.$$

  Again the BIC model is the model that minimizes $\text{BIC}^{(j)}$. Note that the BIC places a higher penalty on model complexity and tends to pick models with fewer parameters.

Both the AIC and BIC trade off model fit (measured by the negative-log likelihood) with model complexity (measure by the parameter counts). Both of these are defined for general models. All that's needed is a log-likelihood. The model does not have to be a GLM.

In the special case of GLMs, the AIC and BIC can be written in terms of the deviance. Let $D_j = D(Y, \mu(\hat{\beta}^{(j)}))$ be the deviance for model j. By Hoeffding's formula, we know that

$$D_j = D(Y, \mu(\hat{\beta}^{(j)})) = 2\left(\log f_Y(Y) - \log f_{\mu(\hat{\beta}^{(j)})}(Y)\right).$$

The first term does not depend on j and the second is exactly $-\ell_j(\hat{\beta}^{(j)})$. Thus, in GLMs, we can work with the equivalent criteria

$$\widetilde{\text{AIC}}^{(j)} = D_j + 2p_j,$$
$$\widetilde{\text{BIC}}^{(j)} = D_j + \log(n)p_j.$$

These formulas can only be applied to GLMs. For other models, use the likelihood based formulas above.

Mallow's C_p statistic

Mallow's C_p statistic is a version of the AIC that applies to linear models. If the predictions for model j can be written as $\hat{Y}^{(j)} = H_j Y$ for some matrix $H_j \in \mathbb{R}^{n\times n}$, then the C_p statistic is

$$C = \frac{1}{\sigma^2}\left\|Y - \hat{Y}^{(j)}\right\|_2^2 + 2\,\text{tr}(H_j),$$

where σ^2 is the residual variance in the largest model. To use the C_p statistic, we need an estimate of σ^2. Normally the unbiased estimate from the largest model is used. The term $\text{tr}(H_j)$ is called the *effective degrees of freedom* of model M_j. In OLS regression, the matrix H_j is the orthonormal projection onto a subspace of rank p_j and so $\text{tr}(H_j) = p_j$.

The LASSO

Suppose we have a log-likelihood $\ell(\beta)$ with parameters $\beta \in \mathbb{R}^p$. The LASSO problem with parameter λ is

$$\text{Minimize} -\frac{1}{n}\ell(\beta) + \lambda\|\beta\|_1.$$

If $\lambda = 0$, then we get the maximum likelihood problem. For $\lambda > 0$, we get a regularized maximum likelihood problem. The use of the 1-norm induces sparsity. That is if $\hat{\beta}_\lambda$ is the solution to the LASSO problem, then we would expect $\hat{\beta}_\lambda$ to have many zero entries. A sparse $\hat{\beta}_\lambda$ corresponds to a sub-model since we can drop the columns of X corresponding to zero values of $\hat{\beta}_\lambda$. The parameter λ can be chosen via cross-validation as in Section **??**. Alternatively, λ can be tuned to give a desired level of sparsity. For example, a question may ask for a model which uses 10 features. You could then start with a large value of λ and decrease λ until you have exactly ten non-zero components.

Compared to AIC or BIC, one advantage of the LASSO is that we do not have to do as many model fits. We only have one parameter λ. If we wanted to use AIC or BIC to select the best model sub-model, we would have to do 2^p model fits.

1 Applied 2021: Solution[1]

Key Ideas/ Main Tools: Unmeasured confounders, GLMs, offsets

Problem 1: Modeling association between vaccination and death rates for Covid-19

(a) It is essential to ask the researchers whether they know the population in each county because counties with higher populations will tend to have a higher death count irrespective of the vaccination rates. Also, if rural counties with low populations tend to have low vaccination rates, not accounting for county population can make vaccination seem less effective than it is.

Other questions worth asking the researchers are if there are any confounding variables that are likely to affect both vaccination rates and death rates, and if any of those confounding variables are measured. For example, the quality of the health care system in a county would affect both vaccination rates and deaths, so it would be worth asking if there are any measured variables that reflect the quality of the healthcare system in each county. Another example is the age demographics. The age demographics can influence both the death rate and the vaccination rate in a county, so it would be helpful to know if the researchers have any covariates that reflect age demographics (e.g. the percentage of the population above age 70 in each county). If the researchers mention many confounding variables that they have measurements for, then I would ask them to select the few most important ones based on their domain knowledge. I would mention that they should choose much fewer than 20 variables to control for because there are only 20 samples.

In addition to asking about county population levels and whether there are measured confounder variables not presented in the table, it would be worthwhile to check with the researcher that the deaths were counted after the vaccines were distributed (otherwise any analysis would be unable to say anything about the effect of vaccination rates on death rates).

(b) Suppose the researchers are able to provide you with the population counts $N_1, \ldots, N_{20}$ of the 20 counties, and for each of the 20 they can give you vectors $z_1, \ldots, z_{20}$ of the few most important measured confounder variables for each county (e.g. age demographics and health care system quality metrics). Also suppose that their death counts in each county, indeed only include deaths from a time period after most of the vaccinations were given.

Letting d_i denote the number of deaths in county i, v_i denote the vaccination rate, $x_i \equiv (v_i, z_i)$, I would fit the following Poisson GLM with offsets $\alpha_i \equiv \log(N_i)$

[1] Dan Kluger and M.H.

$$d_i \overset{\text{ind}}{\sim} \text{Poisson}(\mu_i) \quad \log(\mu_i) = \alpha_i + x_i^\top \beta \quad \text{for } i = 1, ..., 20.$$

After fitting the GLM (which can easily be done in R) using the glm function, I would look at the confidence interval for the first estimated coefficient $\hat{\beta}_1$. If the confidence interval only contains negative values, then we can conclude that the data suggests higher vaccination rates are associated with lower death rates when controlling for the confounders encoded in the z_i.

A binomial GLM could also be appropriate

$$d_i \overset{\text{ind}}{\sim} \text{Binom}(N_i, p_i) \quad \text{Logit}(p_i) = \beta_0 + x_i^\top \beta \quad \text{for } i = 1, \ldots, 20.$$

Again, the confidence interval around $\hat{\beta}_1$ will let us do inference on the affect of vaccination controlling for z_i.

Problem 2: Finding an essential subset

Key Ideas/ Main Tools: Group Lasso

Defining the optimal essential subset as a solution to an optimization problem

Observe that one way to obtain an essential subset is to find the matrix $B \in \mathbb{R}^{750 \times 750}$ that minimizes $||R - RB||_F^2$ subject to the constraint that only 25 of the rows of B are allowed to have nonzero entries. More formally, letting $B[i,]$ denote the ith row of the matrix B we could get an essential subset by solving the following optimization problem on B:

$$\boxed{\text{minimize} \quad ||R - RB||_F^2 \quad \text{subject to} \quad \sum_{i=1}^{750} I\{B[i,] \neq \mathbf{0}\} \leq 25, B \in \mathbb{R}^{750 \times 750}}.$$

Let $\tilde{B}$ be the solution to the above optimization problem and define $\mathcal{S}$ to be the set $\{i \in [750] \ : \ \tilde{B}[i,] \neq \mathbf{0}\}$. By definition, $\mathcal{S}$ will give an essential subset, as each portfolio (represented by a column in R) will be reasonably well approximated by a linear combination of the portfolios of at most.

Solving the boxed optimization problem and setting $\mathcal{S}$ to be the nonzero rows of the solution will recover an essential subset of size at most 25, but unfortunately the optimization problem is non-convex (it has an ℓ_0 type constraint).

A tractable approach using the Group Lasso

One way to induce a sparse number of nonzero rows of B is to use the Group Lasso. In particular, for each $\lambda > 0$, the Group Lasso can be used to solve the following convex optimization problem of finding:

$$\boxed{\hat{B}_\lambda \in \underset{B\in\mathbb{R}^{750\times 750}}{\operatorname{argmin}} \Big(\frac{1}{2}||R - RB||_F^2 + \lambda \sum_{i=1}^{750} ||B[i,]||_2\Big).}$$

We can solve this group Lasso problem for many different λ values until we find a solution $\hat{B}_\lambda$ which has exactly 25 rows which have nonzero. In particular, letting $\hat{\mathcal{S}}_\lambda = \{i \in [750] \ : \ \hat{B}_\lambda[i,] \neq \mathbf{0}\}$, we can do a bisection search on λ until we find a λ_* for which $|\hat{\mathcal{S}}_{\lambda_*}| = 25$. Then we can report $\hat{\mathcal{S}}_{\lambda_*}$ to our boss as an essential subset. Note that this may not be the optimal essential subset in the sense of minimizing $||R - RB||_F^2$ subject to 25 nonzero rows of B; however, it will still be an essential subset according to your bosses definition (that any portfolio in not in $\hat{\mathcal{S}}_{\lambda_*}$ can be well approximated by a linear combination portfolios in $\hat{\mathcal{S}}_{\lambda_*}$).

Additional References: In some years, the Group Lasso is covered in the 305 coursework's lecture notes[2], but it is not covered every year. See Obozinski et al. [2011] for a reference on the Group Lasso and some its theoretical guarantees in recovering a sparse set of rows.

Problem 3: Constructing Conformal Prediction Intervals

Key Ideas/ Main Tools: Prediction Intervals, Conformal Inference, Exchangeability

(a) The defining property of the prediction interval is that $\mathbb{P}\big(Y_{n+1} \in [L, U]\big) \geq 1-\alpha$, where $\mathbb{P}$ is the joint distribution of the $n+1$ data points $(X_i, Y_i)_{i=1}^{n+1}$

(b) This procedure is not reasonable because the more you overfit the data, the smaller the predictions intervals will be. Ideally our prediction will not be over-confident about overfit predictions. In an extreme case suppose that $X_1, \ldots, X_n$ are all distinct and that $\hat{\mu}$ is the best fit $n-1$ degree polynomial to the first n datapoints. In this case, $\hat{\mu}(X_i) = y_i$ for all $i \in [n]$ implying that the residuals $r_1, \ldots, r_n$ are all equal to zero, further implying that the proposed prediction interval will have width zero. Clearly, if we overfit the data, the true prediction interval shouldn't have width zero for a new point X_{n+1}.

(c) Let $\mathcal{S} = \{y \in \mathbb{R} \ : \ \pi(y) \leq (1-\alpha)(n+1)/n\}$ and let $L = \inf S$ and $U = \sup S$. To show that this gives a valid prediction interval first note that

$$\begin{aligned}
\mathbb{P}(Y_{n+1} \in [L, U]) &\geq \mathbb{P}(Y_{n+1} \in \mathcal{S}) \\
&= \mathbb{P}(\pi(Y_{n+1}) \leq (1-\alpha)(n+1)/n) \\
&= \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n} I\{R_{Y_{n+1},i} \leq R_{Y_{n+1},n+1}\} \leq (1-\alpha)(n+1)/n\Big)
\end{aligned}$$

[2] https://web.stanford.edu/class/stats305c/notes/Regression/Sparse.html

To simplify the above expression with an exchangeability argument, first define $\tilde{\mu}$ to to the curve fit to the $n+1$ data points $(X_i, Y_i)_{i=1}^{n+1}$. Next, define for $i = 1, \ldots, n+1$, $V_i \equiv |Y_i - \tilde{\mu}(X_i)|$. Observe that since $(X_i, Y_i)_{i=1}^{n+1}$ are IID and $\tilde{\mu}$ is symmetric function of the collection of these $n+1$ data points, $(V_i)_{i=1}^{n+1}$ is an exchangeable sequence of random variables. In addition, since $\tilde{\mu}(\cdot) = \hat{\mu}_{Y_{n+1}}(\cdot)$,

$$V_i \equiv |Y_i - \tilde{\mu}(X_i)| = |Y_i - \hat{\mu}_{Y_{i+1}}(X_i)| = R_{Y_{n+1},i}.$$

Combining this with a previous result and using the exchangeability of $(V_i)_{i=1}^{n+1}$ (and assuming that almost surely $V_i \neq V_j$ for $i \neq j$),

$$\begin{aligned}
\mathbb{P}(Y_{n+1} \in [L, U]) &\geq \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n} I\{R_{Y_{n+1},i} \leq R_{Y_{n+1},n+1}\} \leq (1-\alpha)(n+1)/n\Big) \\
&= \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n} I\{V_i \leq V_{n+1}\} \leq (1-\alpha)(n+1)/n\Big) \\
&= \mathbb{P}\Big(\sum_{i=1}^{n} I\{V_i \leq V_{n+1}\} \leq (n+1)(1-\alpha)\Big) \\
&\geq \mathbb{P}\Big(\sum_{i=1}^{n} I\{V_i \leq V_{n+1}\} \leq \lfloor (n+1)(1-\alpha) \rfloor\Big) \\
&= \mathbb{P}\Big(\text{Unif}\{0, 1, \ldots, n-1, n\} \leq \lfloor (n+1)(1-\alpha) \rfloor\Big) \\
&= \frac{1 + \lfloor (n+1)(1-\alpha) \rfloor}{n+1} \\
&\geq 1 - \alpha.
\end{aligned}$$

Above the step where $\sum_{i=1}^{n} I\{V_i \leq V_{n+1} \sim \text{Unif}\{0, 1, \ldots, n-1, n\}$ follows from exchangeability of $(V_i)_{i=1}^{n+1}$ (and the assumption almost surely $V_i \neq V_j$ for $i \neq j$).

Note you can also cite Lemma's or Theorem's from Lecture 17 in Stats 300C to solve this problem.

(d) If X_{n+1} is far outside the range of the training data, I would be concerned that the assumption that $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ are exchangeable from some distribution P is violated and that the intervals from part (c) are no longer valid. Even if X_{n+1} was technically a draw from P, the collaborator is confusing conditional coverage with marginal coverage. We do not have conditional coverage over all possible values of X_{n+1}. It is very believable that our conditional coverage decreases as X_{n+1} goes to the tails of the distribution of X.

Despite failure to meet the exchangeability assumption, if we went ahead and constructed the prediction intervals defined in (c), we would get prediction intervals with undesirable behavior. In particular, if the curve $\hat{\mu}$ is fit based on

kernel smoothing or local linear regression (and only considers points with similar X values), then it would follow that for all y, $\hat{\mu}_y(X_{n+1}) = y$ implying that $R_{y,n+1} = 0$ for all y, further implying that $\pi(y) = 0$ for all y. Therefore, if $\hat{\mu}$ is fit based on kernel smoothing or local linear regression, the prediction interval would have infinite length. If on the other hand the curve $\hat{\mu}$ is fit based on a global polynomial regression, one would expect $R_{y,n+1}$ to be much larger than $R_{y,i}$ $(i < n)$ for most y values in which case the prediction interval would be very small. However, for a global model, since the global model is unlikely to hold for outliers we would want the prediction intervals to be very large. In summary, if we were to fit a curve with large extrapolation bias, the interval from part (c) would be very small and not reflect the extrapolation bias, but if we were to fit a local smooth model, the intervals from part (c) would be infinite length. The collaborator should therefore expect meaningless intervals if they went ahead and used prediction intervals for their outlier point.

Problem 4: PCA versus k-means

(a) **Explanation for PCA:** Suppose that each of the features is centered such that columns of X each have mean 0. PCA can be thought of in terms of matrix factorizing X. In particular to implement PCA, you take the SVD of X, which is a matrix factorization given by $X = UDV^\top$ where $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{p\times p}$ are orthogonal matrices, and $D_{ij} = 0$ for all $i \neq j$ and $|D_{11}| \geq |D_{22}| \geq \cdots \geq 0$. In PCA, the principal component directions are given by the columns of V, while the principal component scores are given by UD. Since $X = (UD)V^\top$, principal component analysis can be thought of as factorizing X into matrix of the principal component scores (given by UD) and a matrix whose rows are the principal component directions (given by $V^\top$).

Explanation for K-means: K-means can be thought of as an approximate matrix factorization of $X \in \mathbb{R}^{n\times p}$. In particular let

$$\mathcal{C}_K \equiv \Big\{ C \in \{0,1\}^{n\times K} : \sum_{j=1}^{K} C_{ij} = 1 \text{ for } i = 1, \ldots, n \Big\},$$

be the collection of $n \times K$ matrices whose rows contain exactly $K-1$ zeros and 1 one. We can then think of K-means as solving the following approximate matrix factorization problem

$$(\hat{C}, \hat{Z}) = \underset{C\in\mathcal{C}_K, Z\in\mathbb{R}^{K\times p}}{\operatorname{argmin}} ||X - CZ||_F^2.$$

In particular, if you solve the above approximate matrix factorization problem, the K rows of $\hat{Z}$ will give the K cluster centroids for K-means, and the column index of the nonzero entry in each of the n rows of $\hat{C}$ will give the cluster assignment for each of the n points in the K-means algorithm (or put another way, $\hat{C}_{ij}$

is a indicator of whether the ith point is assigned to cluster j in the K-means algorithm).

(b) For PCA, you can determine the matrix V, whose columns give the principal components dierections using X_{train}. Then, you can use the principal component scores on your test set given by $X_{\text{test}}V$ as features (or some subset of the principal component scores as features) for predicting Y_{test}. Note that if your prediction method is linear regression, and only the first k principal component scores are used as features, this approach would be principal components regression.

For K-means, you use the training data X_{train}, to determine the K cluster centroids. Once you have the cluster centroids, you can determine the cluster assignment of each observation in the test set by determining which of the K cluster centroids is closest (in Euclidean distance) to each row of the matrix X_{test}. Once you have the cluster assignments on the test set, you can use the cluster assignments as a categorical feature for prediction Y_{test} (you may want to use other features in addition to the cluster assignment).

You can evaluate the error in predicting Y for your choice of prediction algorithm and your choice features using cross-validation on the set $(X_{\text{test}}, Y_{\text{test}})$.

(c) While my peer isn't wrong that the cluster structure they found is predictive of Y, it is just not such an impressive predictor of Y. In particular my peer's cluster assignments give a statistically significant improvement over just using the grand mean of Y for predicting Y. That being said, looking at the sum of squares in the ANOVA table, using the cluster means to predict Y doesn't give an impressive improvement over using the grand mean to predict Y. In a similar vein, from linear regression on the cluster we also see that the residuals from the linear regression tend to be much larger in absolute value than the estimates of Y for each cluster. So, I agree with my peer that the cluster structure found is predictive of Y, but it doesn't appear to be terribly important structure for predicting Y.

(d) The results don't contradict each other so it does not cause me to doubt my peer's results. It is possible that the first first 5 principle components are simply not that associated (in a linear way) with the outcome variable Y. Perhaps the jth principal component for $j > 5$ is important in predicting Y. Perhaps the 1st principal component is important in explaining Y, but it has U shaped quadratic relationship with Y with a linear coefficient of 0. In either case the 3-means approach that my friend did would give a better prediction of Y than the principal components regression approach (with 5 components) that I took.

While it doesn't sound like my peer did anything wrong, I'm not terribly impressed by his results because the structure he found is quite a weak predictor of Y (admittedly, I am even less impressed by my own results). I would suggest to our supervisor that we continue seeking an alternative to the 3-means approach,

as there likely is a better choice of features out there. Perhaps it would be an 8-means approach or involve more than the first 5 principal components or it would involve interaction terms. Also neither of our approaches leveraged the Y_{train} data, even though we are told that it exists. I would therefore try to convince our supervisor that it is worth digging deeper before sticking to the 3-means approach of my peer.

Problem 5: Testing and inference on a censored Gaussian draw

Key Ideas/ Main Tools: Score test, maximum likelihood estimation, Bayesian inference

(a) Z is distributed as $N(\mu, 1)$ variable constrained to be at least 2. Therefore, letting Φ denote the standard Gaussian CDF and ϕ denote the standard Gaussian pdf the likelihood is given by

$$L(\mu) = \frac{\frac{1}{\sqrt{2\pi}} \exp\big(-\frac{1}{2}(Z-\mu)^2\big)}{\int_2^\infty \frac{1}{\sqrt{2\pi}} \exp\big(-\frac{1}{2}(z-\mu)^2\big) \mathrm{d}z} = \frac{\phi(Z-\mu)}{1-\Phi(2-\mu)} = \frac{\phi(Z-\mu)}{\Phi(\mu-2)}.$$

The log-likelihood is thus given by

$$l(\mu) = \log\big(L(\mu)\big) = \log\big(\frac{1}{\sqrt{2\pi}}\big) - \frac{1}{2}(Z-\mu)^2 - \log\big(\Phi(\mu-2)\big).$$

The score function is therefore given by

$$U(\mu) = l'(\mu) = (Z-\mu) - \frac{\phi(\mu-2)}{\Phi(\mu-2)},$$

and the Fisher information is given by

$$\begin{aligned} I(\mu) &= -\mathbb{E}[l''(\mu) \mid \mu] \\ &= -\mathbb{E}\Big[-1 - \frac{\Phi(\mu-2)\phi'(\mu-2) - [\phi(\mu-2)]^2}{[\Phi(\mu-2)]^2} \mid \mu\Big] \\ &= 1 + \frac{(2-\mu)\Phi(\mu-2)\phi(\mu-2) - [\phi(\mu-2)]^2}{[\Phi(\mu-2)]^2}. \end{aligned}$$

To conduct a score test of the hypothesis $H_0 : \mu = 0$, one would use the test statistic,

$$T = \frac{[U(0)]^2}{I(0)} = \frac{\big(Z - \frac{\phi(-2)}{\Phi(-2)}\big)^2}{1 + \frac{2\Phi(-2)\phi(-2) - [\phi(-2)]^2}{[\Phi(-2)]^2}} = \frac{(Z-r)^2}{1+2r-r^2}$$

where $r = \frac{\phi(-2)}{\Phi(-2)} \approx 2.373$. The score test will reject whenever T exceeds $c_{1-\alpha}$, where $c_{1-\alpha}$ is the $1-\alpha$ quantile of a chi-squared distribution with one degree of

freedom chosen to satisfy $\mathbb{P}(\chi_1^2 \leq c_{1-\alpha}) = 1-\alpha$. (Note $c_{1-\alpha} \approx 3.84$ for $\alpha = 0.05$). Thus the score test rejects whenever $T > c_{1-\alpha}$, or equivalently

$$\frac{(Z-r)^2}{1+2r-r^2} > c_{1-\alpha} \Leftrightarrow Z > r+\sqrt{(1+2r-r^2)c_{1-\alpha}} \text{ or } Z < r-\sqrt{(1+2r-r^2)c_{1-\alpha}}.$$

Since $Z < 2$ cannot be observed, at level $\alpha = 0.05$, noting that $c_{0.95} \approx 3.84$, the score test of H_0 rejects whenever

$$Z > r + \sqrt{(1+2r-r^2)c_{1-\alpha}} \approx 3.036.$$

(b) Since we just have to optimize over a 1-dimensional parameter you can use grid search to find the MLE. While the following argument is probably unneccasary for the applied qual, to double check that grid search can be used we will find $a, b \in \mathbb{R}$ such that we know $\text{argmax}_{\mu\in\mathbb{R}}\, l(\mu) \in [a,b]$. Note that whenever, $\mu > Z$, $l'(\mu) < 0$, so the MLE must be at most Z, and we can set $b = Z$. To find a lower endpoint a for the grid search observe that as a consequence of Theorem 1.2.3 in Durrett, for $x < -1$, $\frac{\phi(x)}{\Phi(x)} \leq (\frac{1}{-x} - \frac{1}{-x^3})^{-1}$

and hence for $\mu < 1$,

$$\begin{aligned}
l'(\mu) &= Z - \mu - \frac{\phi(\mu-2)}{\Phi(\mu-2)} \\
&\geq Z - \mu - \Big(\frac{1}{-(\mu-2)} - \frac{1}{-(\mu-2)^3}\Big)^{-1} \\
&= Z - \mu - \frac{(2-\mu)^3}{(2-\mu)^2 - 1} \\
&= Z - \mu - \frac{(2-\mu)^2}{(3-\mu)(1-\mu)}(2-\mu) \\
&= Z - 2\frac{(2-\mu)^2}{(3-\mu)(1-\mu)} + \frac{\mu(2-\mu)^2 - \mu(3-\mu)(1-\mu)}{(3-\mu)(1-\mu)} \\
&= Z - 2\frac{(2-\mu)^2}{(3-\mu)(1-\mu)} + \frac{\mu}{(3-\mu)(1-\mu)}.
\end{aligned}$$

Since the above inequality holds for any $\mu < 1$, it is easy to see that $\liminf_{\mu\downarrow-\infty} l'(\mu) \geq Z - 2 > 0$. Further we can use the lower bound above to find an a such that for all $\mu < a$,

$$l'(\mu) \geq Z - 2\frac{(2-\mu)^2}{(3-\mu)(1-\mu)} + \frac{\mu}{(3-\mu)(1-\mu)} > 0.$$

Hence we have an interval $[a,b]$ for which $l'(\mu) > 0$ when $\mu < a$ and $l'(\mu) < 0$ when $\mu > b$. It follows that the MLE $\text{argmax}_{\mu\in\mathbb{R}}\, l(\mu)$ must lie in $[a,b]$. Hence we can simply preform grid search over $[a,b]$ to find the MLE.

c) Suppose $\mu \sim \pi$ and an $Z \mid \mu \sim N(\mu, 1)$. We can estimate μ by considering the posterior distribution of μ given Z. In particular, if we observe some $Z > 2$, by Bayes' rule

$$p(\mu \mid Z, Z > 2) = p(\mu \mid Z) \propto \pi(\mu)\frac{1}{\sqrt{2\pi}}\exp\big(-\frac{1}{2}(Z-\mu)^2\big) \propto \pi(\mu)\exp\big(-\frac{1}{2}(Z-\mu)^2\big).$$

Since the posterior distribution of μ is proportional to $\pi(\mu)\exp\big(-\frac{1}{2}(Z-\mu)^2\big)$, we can estimate μ with the MAP estimate given by

$$\hat{\mu}_{\text{MAP}} = \underset{\mu\in\mathbb{R}}{\text{argmax}}\Big\{\pi(\mu)\exp\big(-\frac{1}{2}(Z-\mu)^2\big)\Big\}.$$

As in part (b) this estimator can be found using 1-dimensional grid search. An alternative estimate for μ would be to estimate the posterior mean

$$\hat{\mu} = \mathbb{E}[\mu \mid Z] = \frac{\int_{-\infty}^{\infty}\mu\pi(\mu)\exp\big(-\frac{1}{2}(Z-\mu)^2\big)\mathrm{d}\mu}{\int_{-\infty}^{\infty}\pi(\mu)\exp\big(-\frac{1}{2}(Z-\mu)^2\big)\mathrm{d}\mu}.$$

The numerator and denominator of the above expression can each be approximated numerically using a Gauss–Hermite quadrature. Alternatively, the above posterior mean can be estimated using the Metropolis–Hastings algorithm.

(c) I'm not entirely sure what the question means by "conflict", but the answer in part (c) was a Bayesian approach based on one observation for which $Z > 2$, whereas part (a) and part (b) describe frequentist approaches where Z is drawn until $Z > 2$. The answer in part (c) used a different likelihood than was used in parts (a) and (b). In particular the likelihood in items (a) and (b) was

$$p(Z \mid \mu, Z > 2) = \phi(Z-\mu)/\Phi(\mu-2),$$

whereas the likelihood in part (c) that was used was

$$p(Z \mid \mu) = \phi(Z-\mu).$$

The easiest way to see why it was not a mistake to use a different likelihood in part (c), is to note that we could have used the same likelihood in part (c) when applying Bayes' rule as the likelihood used in (a) and (b), but doing so would have made a more difficult calculation. In particular, conditioning on $Z > 2$, we could have used Bayes' rule as follows

$$p(\mu \mid Z, Z > 2) = \frac{p(\mu \mid Z > 2)p(Z \mid \mu, Z > 2)}{\int_{-\infty}^{\infty}p(\mu \mid Z > 2)p(Z \mid \mu, Z > 2)\mathrm{d}\mu} \propto p(\mu \mid Z > 2)p(Z \mid \mu, Z > 2),$$

and used the likelihood from parts (a) and (b), but $\pi(\mu \mid Z > 2)$ is more difficult to work with than $\pi(\mu)$ and requires applying Bayes' rule. In fact, if we apply Bayes' rule to $\pi(\mu \mid Z > 2)$ in the above expression, we simply recover the approach used in part (c):

$$p(\mu \mid Z, Z > 2) \propto \pi(\mu)p(Z > 2 \mid \mu)p(Z \mid \mu, Z > 2) = \pi(\mu)\Phi(\mu-2)\frac{\phi(Z-\mu)}{\Phi(\mu-2)} = \pi(\mu)\phi(Z-\mu).$$

Problem 6: Cross-validation in the normal linear model

Key Ideas/ Main Tools: Cross validation, properties of the normal linear model.

This problem is based on results from Bates et al. [2022]. Note that the 2021 qual was open internet, so I think that those who were aware of the paper or were able to find the paper found it the problem quite straightforward, but those who didn't found part (b) especially tricky.

(a) Fix any $x_1, x_2, \ldots, x_n, y_1, \ldots, y_n, \kappa$ and u. Now for any subset $S \subset [n]$, let $\hat{\theta}_{S,0}$ denote the OLS estimator trained on the points $\{(x_i, y_i)\}_{i \in S}$ and let $\hat{\theta}_{S,\kappa}$ denote the OLS estimator trained on the shifted points $\{(x_i, y_i + x_i^\top \kappa)\}_{i \in S}$. Also let $\mathcal{X}_S \in \mathbb{R}^{|S| \times p}$ be the design matrix for these OLS regressions whose rows consist of $\{x_i \ : \ i \in S\}$ and letting $\mathcal{Y}_S = (y_i)_{i \in S}$ be the vector of outcomes for the OLS regression to obtain $\hat{\theta}_{S,0}$. Observe that by the formula for an OLS estimator

$$\hat{\theta}_{S,\kappa} = (\mathcal{X}_S^\top \mathcal{X}_S)^{-1} \mathcal{X}_S^\top \big[\mathcal{Y}_S + \mathcal{X}_S \kappa\big] = (\mathcal{X}_S^\top \mathcal{X}_S)^{-1} \mathcal{X}_S^\top \mathcal{Y}_S + \kappa = \hat{\theta}_{S,0} + \kappa.$$

For any $j \notin S$, if we let $\hat{y}_{S,0,j} = x_j^\top \hat{\theta}_{S,0}$ be the OLS prediction at point j when training on the subset S for the raw data (x_i, y_i) and if we let $\hat{y}_{S,\kappa,j} = x_j^\top \hat{\theta}_{S,\kappa}$ be the OLS prediction at point j when training on the subset S of the translated data $(x_i, y_i + x_i^\top \kappa)$, it follows that

$$\begin{aligned}
\ell(\hat{y}_{S,\kappa,j}, y_j + x_j^\top \kappa) &= \ell(x_j^\top \hat{\theta}_{S,\kappa}, y_j + x_j^\top \kappa) \\
&= \ell(x_j^\top \hat{\theta}_{S,0} + x_j^\top \kappa, y_j + x_j^\top \kappa) \\
&= \ell(x_j^\top \hat{\theta}_{S,0}, y_j) \\
&= \ell(\hat{y}_{S,0,j}, y_j),
\end{aligned}$$

where the 2nd last steps holds because ℓ is the squared error loss function. It is clear that above argument holds for any subset S and $j \notin S$.

Letting $S_1(u), \ldots, S_K(u)$ be the K training subsets of $[n]$ that define the cross-validation (which depend on the random draw U which we are fixing to be u), note that applying the previous result,

$$\begin{aligned}
\widehat{\mathrm{Err}}^{(\mathrm{CV})}\Big((x_1, y_1), \ldots, (x_n, y_n), u\Big) &\equiv \frac{1}{n} \sum_{k=1}^{K} \sum_{j \notin S_k(u)} \ell(\hat{y}_{S_k(u),0,j}, y_j) \\
&= \frac{1}{n} \sum_{k=1}^{K} \sum_{j \notin S_k(u)} \ell(\hat{y}_{S_k(u),\kappa,j}, y_j + x_j^\top \kappa) \\
&= \widehat{\mathrm{Err}}^{(\mathrm{CV})}\Big((x_1, y_1 + x_1^\top \kappa), \ldots, (x_n, y_n + x_n^\top \kappa), u\Big).
\end{aligned}$$

Since this argument holds for any fixed $x_1, x_2, \ldots, x_n, y_1, \ldots, y_n, \kappa$ and u, it follows that $\widehat{\mathrm{Err}}^{(\mathrm{CV})}$ is linearly invariant by definition (2).

(b) Recalling from the problem statement that $\hat{\theta}$ is the OLS estimator based on the all of the observed data, and let $(R_1, \ldots, R_n)$ be the residuals for the OLS estimate on the observed data (i.e. $R_i = Y_i - X_i^\top \hat{\theta}$ for all $i \in [n]$). By part (a), if we let $\kappa = -\hat{\theta}$,

$$\begin{aligned}\widehat{\mathrm{Err}}^{(\mathrm{CV})}\Big((X_1, Y_1), \ldots, (X_n, Y_n), U\Big) &= \widehat{\mathrm{Err}}^{(\mathrm{CV})}\Big((X_1, Y_1 - X_1^\top\hat{\theta}), \ldots, (X_n, Y_n - X_n^\top\hat{\theta}), U\Big) \\ &= \widehat{\mathrm{Err}}^{(\mathrm{CV})}\Big((X_1, R_1), \ldots, (X_n, R_n), U\Big).\end{aligned}$$

It follows conditional on $X = (X_1, \ldots, X_n)$, $\widehat{\mathrm{Err}}^{(\mathrm{CV})}$ is a function of only the residuals $(R_1, R_2, \ldots, R_n, U)$. Also observe that conditional on X, Err_{XY} is only a function of $\hat{\theta}$. By a property of linear regression under the homoskedastic linear model, $(R_1, R_2, \ldots, R_n) \perp\!\!\!\perp \hat{\theta}|X$ (to see this, one can check using the hat matrix that conditional on X, the residuals are uncorrelated with the estimator $\hat{\theta}$ and note that for multivariate Gaussian's zero correlation implies independence). Since U is independent of the data, this further implies that

$$(R_1, R_2, \ldots, R_n, U) \perp\!\!\!\perp \hat{\theta}|X.$$

Because conditional on X, we have shown that $\widehat{\mathrm{Err}}^{(\mathrm{CV})}$ is only a function of $(R_1, R_2, \ldots, R_n, U)$ and because conditional on X, Err_{XY} is only a function of $\hat{\theta}$ the conditional independence result displayed above implies that

$$\widehat{\mathrm{Err}}^{(\mathrm{CV})} \perp\!\!\!\perp \mathrm{Err}_{XY}|X.$$

c) The previous result from item (b) does not imply that $\widehat{\mathrm{Err}}^{(\mathrm{CV})}$ is a useless estimate of prediction error. In particular, $\widehat{\mathrm{Err}}^{(\mathrm{CV})}$ is still a good estimate of $\mathrm{Err} \equiv \mathbb{E}[\mathrm{Err}_{XY}]$, which is the expected prediction loss across all training sets (see Chapter 7.12 in Hastie et al. [2009] and Bates et al. [2022]). Even if we are truly interested in estimating Err_{XY} rather than Err, just because $\widehat{\mathrm{Err}}^{(\mathrm{CV})}$ is uncorrelated with Err_{XY}, it does not mean that it is a bad approximation of Err_{XY}: for large n, the random variable Err_{XY} will likely be concentrated closely about its mean $\mathrm{Err} = \mathbb{E}[\mathrm{Err}_{XY}]$.

(c) Sample splitting into a training set and a test set would also give a linearly invariant estimate of the prediction error by a similar argument in part (a) and the same argument in part (b). Another commonly used estimate of prediction error that is linearly invariant, so that (3) holds is Mallow's C_p. See Bates et al. [2022] for discussion about Mallow's C_p and other commonly used linearly invariant estimates of prediction error.

References

Alan Agresti. *Categorical Data Analysis*. John Wiley & Sons, 2013.

Stephen Bates, Trevor Hastie, and Robert Tibshirani. Cross-validation: what does it estimate and how well does it do it? *arXiv preprint arXiv:2104.00673*, 2022.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Guillaume Obozinski, Martin J. Wainwright, and Michael I. Jordan. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39(1):1 – 47, 2011. doi: 10.1214/09-AOS776. URL `https://doi.org/10.1214/09-AOS776`.