In-Class Test 2

36–402

April 28, 2022

SOLUTIONS

Exam Instructions:

- We will distribute exams promptly at 8:35am. The exam is due at **9:55am**.
- You may not discuss the contents of this exam with other students **until grades are released**.
- Write your answers in the boxes provided. If your answer doesn't fit, you may be writing too much—check if you can simplify your answer.
- Please write legibly; we will not give credit for what we cannot read.
- Your answers must be in the boxes provided. If you need additional scratch paper, ask us.
- You may use **one** normal-size (letter or A4) piece of paper with notes.
- No calculators, computers, or other electronic devices.
- Definitions and formulas are provided when needed by a specific question.
- Your exam will be returned to you via Gradescope.

Useful Formulas

You may use these results anywhere you feel they are necessary.

Expectation

- When $a \in \mathbb{R}$ is a constant, $\mathbb{E}[aX] = a\mathbb{E}[X]$
- When $a \in \mathbb{R}^p$ is a constant vector and X is a p-dimensional random vector, $\mathbb{E}[a^T X] = a^T \mathbb{E}[X]$
- When $A \in \mathbb{R}^{n \times p}$ is a constant matrix and X is a p-dimensional random vector, $\mathbb{E}[AX] = A\mathbb{E}[X]$
- When X and Y are random variables, $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- When X and Y are independent, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
- $\mathbb{E}[X] = \mathbb{E}\left[\mathbb{E}[X \mid Y]\right]$, where the outer expectation averages over Y

Variance and Covariance

- When $a \in \mathbb{R}$ is a constant, $\operatorname{Var}(aX) = a^2 \operatorname{Var}(X)$
- When $a \in \mathbb{R}^p$ is a constant vector and X is a p-dimensional random vector, $\operatorname{Var}(a^T X) = a^T \operatorname{Var}(X) a$
- When $A \in \mathbb{R}^{n \times p}$ is a constant matrix and X is a p-dimensional random vector, $\operatorname{Var}[AX] = A \operatorname{Var}[X] A^T$
- When a and b are constants, and X and Y are random variables,

  $$\operatorname{Var}(aX + bY) = a^2 \operatorname{Var}(X) + b^2 \operatorname{Var}(Y) + 2ab \operatorname{Cov}(X, Y)$$

- When $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are random variables and $a_1, \ldots, a_m$ and $b_1, \ldots, b_n$ are constants, then

  $$\operatorname{Cov}\left(\sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j\right) = \sum_{i=1}^m \sum_{j=1}^n a_i b_j \operatorname{Cov}(X_i, Y_j)$$

- The sample variance of the observations $x_1, x_2, \ldots, x_n$ is

  $$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2,$$

  where $\bar{x}$ is the sample mean of the observations.

- The sample standard deviation of the observations $x_1, x_2, \ldots, x_n$ is the square root of its sample variance.

Multiple-Choice Questions

Select **only one** answer to these questions. Mark your choice by clearly circling the letter. You do not need to write any derivations or justification, and there is no partial credit.

1. An additive model using p predictor variables:

   A. Generally suffers from high variance but low bias, compared to a fully nonparametric estimator (like kernel regression)

   B. Allows us to estimate a nonparametric regression function using the predictors, using tools from univariate nonparametric regression

   C. Computationally only requires us to compute p univariate nonparametric fits, one for each predictor variable

   D. Both (b) and (c)

2. We observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ and consider fitting different regression models. The effective degrees of freedom of a model $\hat{r}(x)$:

   A. Measures the effective number of parameters used by an estimator

   B. Is random, i.e., it depends on the data

   C. Is always equal to the number of terms/coefficients required to write out the mathematical formula for $\hat{r}(x)$

   D. Cannot be negative

   E. Both (A) and (C) are true

   F. (A), (C), and (D) are all true

   G. (A) through (D) are true

3. Suppose $Y = r(X) + \epsilon$, as usual, where $\text{Var}(\epsilon) > 0$. We obtain training data $(Y_1, x_1), \ldots, (Y_n, x_n)$ and decide to fit a nonparametric model to it, such as a kernel smoother.

   Consider each of the following errors:

   - The training error,

   $$\hat{R}_{\text{train}} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{r}(x_i))^2$$

   - The in-sample test error,

   $$R_{in} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[(Y_i' - \hat{r}(x_i))^2 \right],$$

   where Y_i' is independent of, and has the same distribution as, Y_i, conditional on $X_i = x_i$.

   - The out-of-sample prediction error

   $$R = \mathbb{E}\left[(Y' - \hat{r}(X'))^2 \right],$$

   where (Y', X') is an independent draw from the joint distribution of Y and X.

   Suppose we keep making the model more flexible, so it can fit the training data more and more closely. Which of the following is true?

   A. All three errors will go to zero

   B. Training and in-sample test error will go to zero, but out-of-sample prediction error will not

   C. Training error will go to zero, but the in-sample test and out-of-sample prediction error will not

   D. None of the errors can ever be zero

4. In lecture, we showed that the MSE for a univariate kernel regression, when we choose the optimal bandwidth, is $O(n^{-4/5})$.

   Another simple regression method is called the regressogram. The regressogram divides the range of X into bins, and within each bin, simply takes the mean of Y as its prediction. One can show that the regressogram, when we choose the optimal bin width, has MSE that converges at the rate $O(n^{-2/3})$.

   Suppose you fit a kernel regression and a regressogram to data, using cross-validation to choose the best tuning parameters. There is only one predictor variable. Assuming large sample sizes, which method do you expect will have the smallest MSE?

   A. Kernel regression

   B. The regressogram

   C. They will have about the same MSE after balancing bias and variance

Short Answer Questions

Write your answers inside the boxes provided. For questions that ask for explanation, justify each answer with **only a few sentences**.

5. In one of your homework assignments, you predicted the death rate in Chicago by fitting an additive model for the log of the death counts on various predictors $X_1, \ldots, X_p$, including pollution and temperature. How is this different from a Poisson GAM?

   (Note: For full credit, you should explain what each regression model is, including what the response variable Y is, and what the probability model of Y given $X = (x_1, \ldots, x_p)$ is.)

Solution: If we write both in GAM form (which we can, because linear models are equivalent to Gaussian GAMs with the identity link), we can easily see the difference. The model fit in the homework assignment looked like this:

$$\begin{aligned} \log(Y_i) \mid X_i = x_i &\sim \text{Gaussian}(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + r_1(x_{i1}) + \cdots + r_p(x_{ip}). \end{aligned}$$

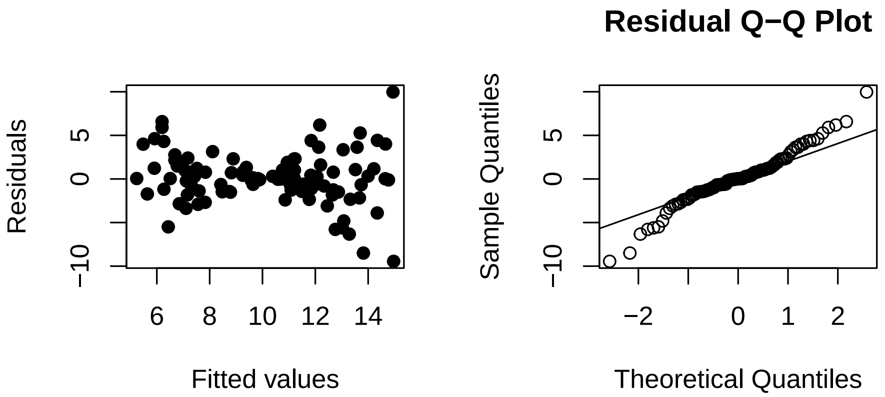
A Poisson-response GAM with the default (log) link, however, is this model:

$$\begin{aligned} Y_i \mid X_i = x_i &\sim \text{Poisson}(\mu_i) \\ \log(\mu_i) &= \beta_0 + r_1(x_{i1}) + \cdots + r_p(x_{ip}) \end{aligned}$$

Both are additive models, but with different functional forms. For example, the original additive model has a σ^2 term for the variance of $\log Y_i$ given X_i, but in the Poisson model, there is no separate variance, since the Poisson's variance is equal to its mean. Each model has a different likelihood function as well.

6. Farmer Brown grows thousands of pumpkins every year. Each pumpkin receives a blend of five different fertilizers and pesticides. She wants to grow the largest pumpkins possible, so she takes a random sample of plants, tracks the amount of each fertilizer and pesticide they receive ($X_1, \ldots, X_5$), and then records their final weight when harvested (Y, in kilograms). She knows many factors have a nonlinear relationship with weight, so she fits an additive model using splines.

   She obtains the following residual plots:



Farmer Brown's goal is to obtain a confidence interval for the mean weight of pumpkins grown a certain way. She intends to estimate $\mathrm{Var}(\hat{r}(x)) = \sigma^2 \ell(x)^T \ell(x)$ and then construct a normal-based confidence interval for $r(x)$. State the assumptions that are necessary for this to be a reasonable approach; based on the residual plots above, state which of these assumptions you can check and whether they appear to be satisfied.

Solution: There was a typo in this question: it should say that $\mathrm{Var}(\hat{r}(x)) = \sigma^2 \ell(x)^T \ell(x)$, the variance estimate for linear smoothers. We announced this correction in class, so students should be aware of it.

Our model is that $Y_i = r(x_i) + \epsilon_i$, where we fit $r(x_i)$ with an additive model. In order to make inference, we assume that

- the model is approximately unbiased
- the errors ϵ_i are normally distributed
- the errors ϵ_i have constant variance.

We check these assumptions by getting the residuals $\hat{\epsilon}_i = Y_i - \hat{r}(x_i)$. We'd expect they look approximately normal (e.g. in a Q-Q plot), and that when plotted against the fitted values or covariates, no obvious trends are seen, the mean appears to be 0 everywhere, and the variance appears constant.

The residuals do not have to be approximately normally distributed if (i) the distribution is not too heavy-tailed and (ii) the sample size is reasonably large. If both of these conditions hold, the CLT will make the sampling distributions approximately normally distributed.

Here, we see no reason to think there's bias: the residuals appear to have mean 0 everywhere. But the residuals do not have constant variance and do not look particularly normally distributed.

7. Under the law of the United Kingdom, the Queen officially owns all swans. Each year, the government conducts the "swan upping," a count of all swans on the River Thames. Let Y be

the number of swans born each year. The Royal Swan Upper uses a Poisson GLM with log link to regress Y on two continuous covariates:

- X_1: average water temperature in the last year (Celsius)
- X_2: average amount of salt in the river water in the last year (grams per kilogram)

The Swan Upping has been conducted for hundreds of years, so there is plenty of data.

(a) The Swan Upper determines that an interaction is necessary between temperature and salinity. Write down the mathematical relationship between $\mathbb{E}[Y \mid X = x]$ (or the logarithm of $\mathbb{E}[Y \mid X = x]$) and the covariates, assuming a Poisson GLM with the covariates and the interaction.

Solution: We have

$$\log(\mathbb{E}[Y \mid X = x]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2,$$

or equivalently,

$$\mathbb{E}[Y \mid X = x] = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2).$$

(b) The Swan Upper fits the model, obtaining the coefficients and their estimated covariance matrix:

$$\hat{\beta} = [\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3] \qquad \mathrm{Var}(\hat{\beta}) = \begin{pmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} & \sigma_{03} \\ \sigma_{10} & \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{20} & \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{30} & \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{pmatrix}$$

Given these, write the estimated change in mean number of swans born per year when the mean temperature is increased by one degree. Assume the amount of salt is fixed at 25 g/kg. Give both the point estimate and a 95% confidence interval. You can leave your answer in terms of the coefficients and covariances shown above.

Hint: First, figure out what the point estimate and 95% CI is in the "link scale" of your Poisson GLM. Then, invert that CI back to the desired "response scale" by applying the inverse link function to the CI.

Solution: When salt is fixed at 25 g/kg, that means $X_2 = 25$. In this case, the relationship is:

$$\mathbb{E}[Y \mid X_1 = x_1] = \exp(\beta_0 + 25\beta_2 + (\beta_1 + 25\beta_3)X_1)$$

When X_1 increases by one degree, the mean is hence **multiplied** by $\exp(\beta_1 + 25\beta_3)$.

We calculate:

$$\begin{aligned} \mathrm{Var}(\hat\beta_1 + 25\hat\beta_3) &= \mathrm{Var}(\hat\beta_1) + 25^2\,\mathrm{Var}(\hat\beta_3) + 2 \times 25\,\mathrm{Cov}(\hat\beta_1, \hat\beta_3) \\ &= \sigma_1^2 + 25^2\sigma_3^2 + 50\sigma_{13} \\ \mathrm{se}(\hat\beta_1 + 25\hat\beta_3) &= \sqrt{\sigma_1^2 + 25^2\sigma_3^2 + 50\sigma_{13}}. \end{aligned}$$

The normal-based confidence interval is hence that the mean is multiplied by between

$$\left[\exp(\hat\beta_1 + 25\hat\beta_3 - 2\,\mathrm{se}), \exp(\hat\beta_1 + 25\hat\beta_3 + 2\,\mathrm{se})\right]$$

where 2 can be replaced with the appropriate Normal quantiles as desired.

8. A study (Koch & Edwards, 1988) considered a new treatment for rheumatoid arthritis. There were 84 patients; they were randomly assigned to receive either the new treatment or a placebo. The data includes four variables:

   Treatment Factor indicating if the patient received treatment or control

   Sex The patient's sex (male or female)

   Age The patient's age (in years)

   Improved Whether the patient's condition improved after treatment (1 if yes, 0 if no)

   To model the data, we run the following R code (output trimmed slightly for length):

```
> logit_fit <- glm(Improved ~ Treatment + Sex + Age, family = binomial,
+                  data = Arthritis)
> summary(logit_fit)
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.10833  -0.91158   0.05362   0.91681   1.84659

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.01546    1.16777  -2.582  0.00982 **
TreatmentTreated  1.75980    0.53650   3.280  0.00104 **
SexMale          -1.48783    0.59477  -2.502  0.01237 *
Age               0.04875    0.02066   2.359  0.01832 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 116.449  on 83  degrees of freedom
Residual deviance:  92.063  on 80  degrees of freedom
AIC: 100.06
> confint(logit_fit)
                        2.5 %      97.5 %
(Intercept)      -5.477304188 -0.84351926
TreatmentTreated  0.750803551  2.87506538
SexMale          -2.729719456 -0.37239953
Age               0.009951561  0.09194283
```

(a) State the mathematical relationship between treatment and your odds of improvement, compared to receiving the placebo. Give specific numbers from the output above.

> **Solution:** Treatment is associated with log-odds of improvement that are 1.76 higher, or odds of improvement that are multiplied by $\exp(1.76)$, when compared to people who receive the placebo. (And when holding sex and age fixed.)

(b) Does this study provide evidence that the treatment **causes** this change in odds of improvement, given the study design? Why or why not?

> **Solution:** Yes. Because the study randomly assigned treatment, there cannot be confounding variables. Though a single study is not definitive proof, it does provide evidence that the treatment causes improved outcomes.
>
> Note that random *sampling*, as opposed to random assignment, has nothing to do with the causal claim here.

(c) Provide an approximate 95% confidence interval (CI) for how age affects the odds of improvement. Make sure that it is clear what population quantity the CI is for; that is, the final answer should have the form "a 95% CI for [fill in] is [fill in]."

> **Solution:** A 95% CI for the population increase in log-odds of improvement associated with each additional year of age is [0.01, 0.09]. (This is read directly from the `confint` output, and rounded.) The CI for the odds is $[\exp(0.01), \exp(0.09)]$.
>
> It's fine if students only give the CI for log-odds, as long as they clearly say it's for log-odds, since the question says they must be clear what the CI is for.

9. Professor E. Fudd is studying the diet of rabbits. He obtains a random sample of $n = 450$ rabbits, offers each of them a carrot, and records whether the rabbit eats the carrot ($Y = 1$) or refuses it ($Y = 0$). He also records various covariates, such as the rabbit's age, the time since it last had a meal, and so on.

(a) Professor Fudd considers two different logistic regression models. Model 1 is a standard logistic regression model with no interactions; Model 2 is a generalized additive model, where every covariate enters the model as a spline smoother.

Consider using a deviance test to compare these models. State the null and alternative hypotheses of this test, and give the assumptions under which the deviance test statistic has a χ^2 distribution.

> **Solution:** The null hypothesis is that Model 1 is correct. The alternative hypothesis is that Model 2 is correct.
>
> The deviance test statistic has an asymptotically χ^2 distribution when the null hypothesis is true and the logistic regression model is correctly specified.

(b) Professor Fudd let `mgcv` use generalized cross-validation to pick the GAM's smoothing parameters, and knows the deviance test needs to take this into account. He decides not to rely on the approximate χ^2 distribution.

Describe how to use the bootstrap to resample the data. In the next part you will explain how to use the bootstrap samples to conduct the test; in this part, you need only describe how to do the bootstrap sampling.

Your steps should be specific: if you randomly draw some values, indicate how many must be drawn and from what distribution; if you repeat some steps several times, indicate how many times they should be repeated. You don't need to elaborate on or explain your steps.

Solution: Here we are conducting a bootstrap to do a hypothesis test, so we should use a parametric bootstrap: one where the null hypothesis is true. We complete the following steps:

1. Using the fitted Model 1, calculate $\hat{p}_i = \text{logit}^{-1}(x_i^T \hat{\beta})$.
2. Draw $Y_i^* \sim \text{Bernoulli}(\hat{p}_i)$ for $i = 1, \dots, n$. (This is important: Y_i^* should not be a deterministic function of x_i, but randomly drawn.)
3. Fit both Model 1 and Model 2 to $(x_1, Y_1^*), \dots, (x_n, Y_n^*)$.
4. Calculate the test statistic T_b^*, the difference in deviance between Model 1 and Model 2.
5. Repeat steps 2 onward many times, e.g. $B = 1000$ times.

Depending on how students interpreted the question, they may have put the test statistic calculation in part (c) instead of here, which is fine.

A nonparametric (resample cases) bootstrap is not appropriate here, since we are trying to get the distribution under the null hypothesis.

(c) Finally, describe how to use these bootstrap samples to conduct a deviance test. State specifically what your test statistic is and how you will obtain a p value based on the bootstrap samples you have obtained.

> **Solution:** In part (b), we calculated a test statistic T_b^* from each bootstrap $b = 1, \dots, B$. That test statistic was the difference in deviance.
>
> We then calculate T, the difference in deviance between the *original* Model 1 and Model 2 fit to the original data.
>
> We let p be the fraction of T_b^* values greater than T.

(d) Professor Fudd conducts the bootstrap you described in (c) and obtains $p = 0.471$. State how to interpret this p value, and what Professor Fudd can conclude about Model 1 and Model 2.

> **Solution:** We fail to reject the null hypothesis. Professor Fudd does not have evidence that Model 2 fits substantially better than Model 1; however, this does not prove that Model 1 is a good model.