EBA35002 Fall 2022 Mock

Written exam

Jonas Moss

All subexercises are equally weighted.

1 Mathematical questions

In this exercise X_i are n iid observations from a normal distribution with unknown mean μ and unknown standard deviation σ^2 .

1.a.

What is the asymptotic distribution of $\sqrt{n}(\overline{X}-\mu)$, where $\overline{x}$ denotes the mean of $x_1, x_2, \ldots, x_n$?

1.b.

Show that the maximum likelihood estimator of μ is $\overline{x}$.

1.c.

What is the Fisher information of μ, i.e., $I(\mu)$? (*Hint:* There are two ways to calculate this; one is considerably faster than the other.)

1.d.

Show that the maximum likelihood estimator of σ^2 is $\widehat{\sigma^2} = \sum_{i=1}^{n}(X_i - \overline{X})^2/n$. Is this an unbiased estimator of σ^2?

1.e.

Show that the Fisher information of σ^2 is $I(\sigma^2) = 1/(2\sigma^4)$.

1.f

Let θ be a k-dimensional vector parameter and $g : \mathbb{R}^k \to \mathbb{R}$ be a continuously differentiable function. Moreover, suppose $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Sigma)$. What is the asymptotic distribution of $\sqrt{n}(g(\hat{\theta}) - g(\theta))$?

1.g

Let $g(x, y) = x/\sqrt{y}$. Find the partial derivatives of g, i.e.,

$$\frac{\partial g}{\partial x}, \quad \frac{\partial g}{\partial y}.$$

1.h

We want to do inference on μ/σ, sometimes called the *effect size*. What is the asymptotic distribution of $\overline{x}/\sqrt{\widehat{\sigma^2}}$? You may use that the Fisher information of (μ, σ) is a diagonal matrix with zero off-diagonal entries, i.e.,

$$\begin{bmatrix} a & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$$

where a is the Fisher information you found in exercise 1.c.

1.i

Use the information in the previous exercise to construct an approximate 95% confidence interval for μ/σ. If you weren't able to solve the previous exercise, explain how you would do it.

1.j

Using the results in the previous exercise, construct a confidence interval for μ/σ when $\overline{X} = 1$ and $\overline{X^2} = 2$.

2 Regression questions

2.a

Suppose we have a regression model with a continuous response and one continuous covariate. What is the relationship between the R^2 and the correlation coefficient?

2.b

```python
import statsmodels.api as sm
import statsmodels.formula.api as smf
import numpy as np
cpssw8 = sm.datasets.get_rdataset("CPSSW8", "AER").data
model = smf.ols("np.log(earnings) ~ age - 1", data = cpssw8).fit()
```

Below we calculate the R^2:

```python
y = np.log(cpssw8.earnings)
y_mean = y.mean()
1 - np.mean((y - model.predict())**2) / np.mean((y - y_mean)**2)
```

With output `-1.084`! How is this possible? How can you guarantee this never happens? Change the formula to make sure that the R^2 is non-negative.

2.c

Suppose you have 3 categorical covariates `a,b,c` with 3, 7 and 13 levels each. How many regression coefficients are there in the model `y~a*b*c`?

2.e

Mention three reasonable distance functions in linear regression. Which on is most popular? Name three reasons why it is the most popular.

2.f

Alice is in big trouble! Her boss wants her to do a linear regression `satisfaction ~ age + gender`, where `satisfaction` is a number in $\{-3, -2, -1, 0, 1, 2, 3\}$ encoding customer satisfaction. But Alice has somehow managed to throw away her `satisfaction` data, replacing it with the variable `satisfied = 1 * (satisfaction >= 0)` instead. What should Alice do to fulfill her boss's wish, and why would it work?

2.g

Alice shows Bob some limited output for four models she fitted:

Model Number	R squared	Adjusted R squared
1	0.017	0.007
2	0.018	-0.002
3	0.047	-0.026
4	0.161	0.035

Which of these models would you prefer, and why?

2.h

Tjodleif is a social scientist studying the happiness. He has surveyed 200 Bergensere about their happiness level and age. Now he claims that happiness and age is inverse U-shaped, i.e., happiness tends to decrease with age up to a certain point (e.g. 40) and then increase! He based his conclusion on