Math 343 / 643 Spring 2025
Midterm Examination Two

Professor Adam Kapelner

April 10, 2025

Full Name ______________________________

Code of Academic Integrity

Since the college is an academic community, its fundamental purpose is the pursuit of knowledge. Essential to the success of this educational mission is a commitment to the principles of academic integrity. Every member of the college community is responsible for upholding the highest standards of honesty at all times. Students, as members of the community, are also responsible for adhering to the principles and spirit of the following Code of Academic Integrity.

Activities that have the effect or intention of interfering with education, pursuit of knowledge, or fair evaluation of a student's performance are prohibited. Examples of such activities include but are not limited to the following definitions:

Cheating Using or attempting to use unauthorized assistance, material, or study aids in examinations or other academic work or preventing, or attempting to prevent, another from using authorized assistance, material, or study aids. Example: using an unauthorized cheat sheet in a quiz or exam, altering a graded exam and resubmitting it for a better grade, etc.

I acknowledge and agree to uphold this Code of Academic Integrity.

______________________________ ______________
signature date

Instructions

This exam is 75 minutes (variable time per question) and closed-book. You are allowed **one** page (front and back) of a "cheat sheet", blank scrap paper (provided by the proctor) and a graphing calculator (which is not your smartphone). Please read the questions carefully. Within each problem, I recommend considering the questions that are easy first and then circling back to evaluate the harder ones. No food is allowed, only drinks.

Problem 1 Consider the following full-rank design matrix:

$$\boldsymbol{X} := [\mathbf{1}_n \mid \boldsymbol{x}_{\cdot 1} \mid \ \ldots \ \mid \boldsymbol{x}_{\cdot p}] = \begin{bmatrix} \boldsymbol{x}_{1\cdot} \\ \vdots \\ \boldsymbol{x}_{n\cdot} \end{bmatrix}$$

with column indices $0, 1, \ldots, p$ and row indices $1, 2, \ldots, n$. And let $\boldsymbol{H}$ be the orthogonal projection matrix onto the column space of $\boldsymbol{X}$. We assume also a continuous (real-valued) response model which is linear in these measurements, i.e. $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\mathcal{E}}$. For the error term, we assume the "core assumption",

$$\boldsymbol{\mathcal{E}} \sim \mathcal{N}_n\left(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n\right) \quad \text{where } \sigma^2 > 0.$$

Consider the following estimator for $\boldsymbol{\beta}$: $\boldsymbol{B} := \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ and let $\hat{\boldsymbol{Y}} := \boldsymbol{X}\boldsymbol{B}$ and $\boldsymbol{E} := \boldsymbol{Y} - \hat{\boldsymbol{Y}}$.

(a) [5 pt / 5 pts] Circle all of the following which are non-degenerate random variables.

$$n, \quad p, \quad \boldsymbol{X}, \quad \boldsymbol{x}_{\cdot 1}, \quad \boldsymbol{x}_{n\cdot}, \quad \boldsymbol{H}, \quad \boldsymbol{Y}, \quad \boldsymbol{\beta}, \quad \boldsymbol{\mathcal{E}}, \quad \sigma^2, \quad \boldsymbol{I}_n, \quad \boldsymbol{B}, \quad \hat{\boldsymbol{Y}}, \quad \boldsymbol{E}$$

(b) [3 pt / 8 pts] Of the random variables in the previous question, which two are independent of each other? No need to prove this.

$$\boldsymbol{B}, \boldsymbol{E}$$

(c) [5 pt / 13 pts] Derive the distribution of $\boldsymbol{B}$ with only what is in the problem header, the fact about multivariate normal distributions from 340 and linear algebra manipulations. Show each step.

$$\begin{aligned}
\boldsymbol{B} &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\mathcal{E}}\right) = \boldsymbol{\beta} + \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{\mathcal{E}} \\
&\sim \mathcal{N}_{p+1}\left(\boldsymbol{\beta}, \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\sigma^2\boldsymbol{I}_n\left(\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\right)^\top\right) \\
&= \mathcal{N}_{p+1}\left(\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{X}\left(\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)^\top\right) \\
&= \mathcal{N}_{p+1}\left(\boldsymbol{\beta}, \sigma^2\left(\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)^\top\right) \\
&= \mathcal{N}_{p+1}\left(\boldsymbol{\beta}, \sigma^2\left(\left(\boldsymbol{X}^T\boldsymbol{X}\right)^\top\right)^{-1}\right) \\
&= \mathcal{N}_{p+1}\left(\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)
\end{aligned}$$

(d) [6 pt / 19 pts] Prove estimation error vanishes as $n \to \infty$.

Estimation error is $g(\boldsymbol{x}) - h^*(\boldsymbol{x})$ where $h^*(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta}$ by assumption and $g(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{b}$ because we are using OLS thus estimation error is $\boldsymbol{x}(\boldsymbol{\beta} - \boldsymbol{b})$. Over the whole dataset the estimation errors are $\boldsymbol{x}_{1\cdot}(\boldsymbol{\beta} - \boldsymbol{b}), \ldots, \boldsymbol{x}_{n\cdot}(\boldsymbol{\beta} - \boldsymbol{b})$. One way to holistically measure all estimation errors is to sum their squares, i.e, $||\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{b})||^2$ as seen on the homework.

Now we prove this holistic measure goes to zero as n increases. First note that $\boldsymbol{b}$ is a realization of $\boldsymbol{B}$, the OLS estimator which is also the MLE for $\boldsymbol{\beta}$. From 341, the monster theorem stated that MLE's are consistent. Thus in our setting, $\boldsymbol{B} \overset{p}{\to} \boldsymbol{\beta}$ and thus by the multivariate CMT from 340, $\boldsymbol{B} - \boldsymbol{\beta} \overset{p}{\to} \boldsymbol{0}_{p+1}$. Now we apply this result to our holistic measure. By the multivariate CMT, $||\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{b})||^2 \overset{p}{\to} ||\boldsymbol{X}\boldsymbol{0}_{p+1}||^2 = 0$.

Problem 2 Consider the Boston Housing Data which has $n = 506$ and response `medv` with $\bar{y} = 22.53$ and $s_y = 9.20$. We consider modeling `medv` using OLS on `zn + rm + nox + dis + lstat`, all continuous (non-categorical) features. Below is the $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$ where $\boldsymbol{X}$ is the design matrix:

	(Intercept)	zn	rm	nox	dis	lstat
(Intercept)	0.58000	4.4e-04	-5.1e-02	-2.7e-01	-1.8e-02	-3.0e-03
zn	0.00044	6.9e-06	-4.5e-05	-7.1e-05	-5.1e-05	-2.0e-07
rm	-0.05100	-4.5e-05	6.9e-03	7.5e-04	6.3e-04	4.4e-04
nox	-0.27000	-7.1e-05	7.5e-04	4.2e-01	1.5e-02	-1.9e-03
dis	-0.01800	-5.1e-05	6.3e-04	1.5e-02	1.5e-03	4.3e-05
lstat	-0.00300	-2.0e-07	4.4e-04	-1.9e-03	4.3e-05	8.9e-05

The RMSE for this regression is 5.289 and here are the slope estimates:

(Intercept)	zn	rm	nox	dis	lstat
16.14	0.06	4.44	-15.20	-1.44	-0.66

Assume the "core assumption" (see Problem 1 for its definition) except in (e,f,l,m) which make explicit a new assumption.

(a) [2 pt / 21 pts] Consider creating a $\hat{\hat{CI}}_{\beta_{\text{nox}},95\%}$, the confidence interval for the true slope parameter of the variable `nox`. Which degrees of freedom value would you use to lookup the appropriate t value's quantile?

$$df_{\text{error}} := n - (p+1) = 506 - 6 = 500$$

(b) [5 pt / 26 pts] Compute $\hat{\hat{CI}}_{\beta_{\texttt{nox}},95\%}$ to the nearest two digits. Regardless of the truly appropriate t value, use 1.96 as the t value.

$$
\begin{aligned}
\hat{\hat{CI}}_{\beta_{\texttt{nox}},1-\alpha} &= \left[b_{\texttt{nox}} \pm t_{df_{\text{error}},1-\alpha/2} \cdot s_e \sqrt{\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}_{\texttt{nox},\texttt{nox}}}\right] \\
\hat{\hat{CI}}_{\beta_3,95\%} &= \left[b_3 \pm 1.96 \cdot s_e \sqrt{\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}_{4,4}}\right] \\
&= \left[-15.20 \pm 1.96 \cdot 5.289\sqrt{0.42}\right] = [-21.92, -8.48]
\end{aligned}
$$

(c) [1 pt / 27 pts] The confidence interval in the previous question is... circle one:
exact / approximate

(d) [1 pt / 28 pts] Based on your confidence interval from the previous question, the null hypothesis that $\beta_{\texttt{nox}} = 0$ would be ... circle one:
rejected / retained

(e) [5 pt / 33 pts] Assume the errors are independent, mean centered and homoskedastic but now assume they are *not* normally distributed. Create a $\hat{\hat{CI}}_{\beta,95\%}$ for the variable `nox` to the nearest two digits.

$$[-21.92, -8.48]$$

(f) [1 pt / 34 pts] The confidence interval in the previous question is... circle one:
exact / approximate

(g) [5 pt / 39 pts] Justify and record your decision for the test of $H_0 : \beta_{\texttt{rm}} = 3$, a test on the slope parameter for the variable `rm`. Regardless of the truly appropriate t value, use 1.96 as the t value.

$$\hat{\hat{t}} := \frac{b_{\texttt{rm}} - \beta_{\texttt{rm}}}{s_e \cdot \sqrt{\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}_{\texttt{rm},\texttt{rm}}}} = \frac{4.44 - 3}{5.289 \cdot \sqrt{0.0069}} = 3.277 > 1.96 \quad \Rightarrow \quad \text{Reject } H_0$$

(h) [5 pt / 44 pts] Compute R^2_{adj} to the nearest two digits using the following calculations:

$$s_e \quad := \quad \sqrt{\frac{SSE}{df_{\text{error}}}} \quad \Rightarrow \quad SSE = df_{\text{error}} \cdot s_e^2 = 500 \cdot 5.289^2 = 13986.76$$

$$SST \quad := \quad \sum_{i=1}^{n} (y_i - \bar{y})^2 = (n-1) \cdot s_y^2 = 505 \cdot 9.20^2 = 42743.2$$

$$R^2_{adj} \quad := \quad 1 - \frac{n-1}{df_{\text{error}}} \frac{SSE}{SST} = \frac{505}{500} \frac{13986.76}{42743.2} = 0.33$$

Below is the first six rows and six columns of the $\boldsymbol{H}$ matrix. There are rownames and colnames displayed to help with finding entries (e.g., $\boldsymbol{H}_{2,4} = 0.0076$).

	1	2	3	4	5	6
1	0.0053	0.0020	0.0035	0.0039	0.0024	0.0036
2	0.0020	0.0058	0.0065	0.0076	0.0076	0.0072
3	0.0035	0.0065	0.0100	0.0110	0.0110	0.0085
4	0.0039	0.0076	0.0110	0.0130	0.0130	0.0110
5	0.0024	0.0076	0.0110	0.0130	0.0140	0.0110
6	0.0036	0.0072	0.0085	0.0110	0.0110	0.0110

(i) [5 pt / 49 pts] Estimate the probability the residual for the fourth observation in the boston housing dataset will be greater than 5 as best as you can.

$$E_4 \sim \mathcal{N}\left(0, \sigma^2(1-h_{4,4})\right) \quad \Rightarrow \quad \frac{E_4}{\sigma\sqrt{1-h_{4,4}}} \sim \mathcal{N}(0, 1) \quad \Rightarrow \quad \frac{E_4}{s_e\sqrt{1-h_{4,4}}} \dot{\sim} \mathcal{N}(0, 1)$$

$$\mathbb{P}(E_4 > 5) = \mathbb{P}\left(\frac{E_4}{s_e\sqrt{1-h_{4,4}}} > \frac{5}{s_e\sqrt{1-h_{4,4}}}\right) \approx \mathbb{P}\left(Z > \frac{5}{5.289\sqrt{1-0.0130}}\right)$$

$$= \quad \mathbb{P}(Z > 0.95) \approx 16\%$$

(j) [6 pt / 55 pts] The predicted value for the first observation is $\hat{y}_1 = 29.15$. Find a $\hat{\hat{CI}}_{y_1,95\%}$ where y_1 is the response value for a new census tract with the same measurements as $\boldsymbol{x}_{1\cdot}$ to the nearest two digits. Regardless of the truly appropriate t value, use 1.96 as the t value.

$$\begin{aligned}
\hat{\hat{CI}}_{y_1,95\%} &= \left[\hat{y}_1 \pm t_{1-\alpha/2,n-(p+1)} \cdot s_e \sqrt{1 + \boldsymbol{x}_{1\cdot}^\top \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{x}_{1\cdot}}\right] \\
&= \left[\hat{y}_1 \pm t_{1-\alpha/2,n-(p+1)} \cdot s_e \sqrt{1 + h_{1,1}}\right] \\
&= \left[29.15 \pm 1.96 \cdot 5.289\sqrt{1 + 0.0053}\right] = [18.76, 39.54]
\end{aligned}$$

We now model `medv` using `rm + lstat` via an OLS. The RMSE for this regression is 5.540 and here are the slope estimates:

(Intercept)	rm	lstat
-1.36	5.09	-0.64

(k) [7 pt / 62 pts] Calculate the F-statistic for $H_0 : \beta_{\texttt{zn}} = \beta_{\texttt{nox}} = \beta_{\texttt{dis}} = 0$ to the nearest two digits.

$$\hat{\hat{f}} := \frac{\frac{SSE_A - SSE}{k}}{\frac{SSE}{df_{error}}} = \frac{\frac{15437.87 - 13986.76}{3}}{\frac{13986.76}{500}} = 17.29$$

The value of k is the number of features we are setting to zero in H_0 which is 3. The value of SSE we take from part (h). To obtain the value of SSE_A, see the calculation in part (h). $SSE_A = df_A s_{e_A}^2$ where the quantities on the rhs are now applicable to the reduced model with features in set A. Following part (a), $df_A = n - ((p - k) + 1) = 506 - ((5 - 3) + 1) = 503$. And $s_{e_A}^2$ can be found in the text above this question. Thus, $SSE_A = 503 \cdot 5.54^2 = 15437.87$.

Below is $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\hat{\hat{\boldsymbol{D}}}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$, a matrix where $\boldsymbol{X}$ is the design matrix and $\hat{\hat{\boldsymbol{D}}}$ is the diagonal matrix with the residuals squared along its diagonal.

```
              (Intercept)    rm lstat
(Intercept)         29.20 -4.14 -0.26
rm                  -4.14  0.59  0.03
lstat               -0.26  0.03  0.00
```

(l) [5 pt / 67 pts] Assume the errors are independent, mean centered but neither homoskedastic nor normally distributed. Create a $\hat{\hat{CI}}_{\beta_{\mathtt{rm}},95\%}$, the confidence interval for the true slope parameter of the variable `rm` to the nearest two digits.

$$\hat{\hat{CI}}_{\beta_{\mathtt{rm}},1-\alpha} = \left[b_{\mathtt{rm}} \pm z_{1-\alpha/2}\sqrt{\left(\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\hat{\hat{\boldsymbol{D}}}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)_{\mathtt{rm},\mathtt{rm}}}\right]$$

$$\hat{\hat{CI}}_{\beta_1,95\%} = \left[b_1 \pm 1.96\sqrt{\left(\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\hat{\hat{\boldsymbol{D}}}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)_{2,2}}\right]$$

$$= \left[5.09 \pm 1.96\sqrt{0.59}\right] = [3.58, 6.60]$$

(m) [1 pt / 68 pts] The confidence interval in the previous question is... circle one:
exact / approximate

Problem 3 Consider a subset of the vocab data in the `carData` package. The response is a person's score on a vocabulary test. This score ranges in $\{0, 1, 2, \ldots, 10\}$ and features: `gender` (categorical: male/female), `nativeBorn` (categorical: yes/no), `age` (continuous: measured in years) and `educ` (continuous: measured in years). We will use a negative binomial glm with the standard exponential link-to-linear function for its mean. Below is the output:

```
               Estimate   Std. Error z value Pr(>|z|)
(Intercept)     0.7761238  0.0165403 46.923  < 2e-16 ***
gendermale     -0.0267524  0.0049960 -5.355 8.57e-08 ***
nativeBornyes   0.1603976  0.0094713 16.935  < 2e-16 ***
age             0.0021438  0.0001438 14.907  < 2e-16 ***
educ            0.0582323  0.0008373 69.548  < 2e-16 ***

              Theta:  172454
          Std. Err.:  143423

 2 x log-likelihood:  -115304.3
```

(a) [5 pt / 73 pts] Is there any reason why we should not model this response metric using the negative binomial model with mean log-linear in the covariates?

The response metric doesn't have the support of a true count model as it only ranges from $0, 1, \ldots, 10$ instead of $0, 1, \ldots$ and this means the model may give nonsensical predictions (i.e., vocabulary scores > 10). Thus, the inference will also be suspect.

Despite what you wrote in (a), we will ignore any concerns about the appropriateness of this model going forward.

(b) [5 pt / 78 pts] Considering all other covariate values the same, what would be the predicted *percent* difference in mean score of a male versus a female to the nearest two digits?

Since $\hat{y} = e^{b_0} e^{b_1 x_1} \cdot \ldots \cdot e^{b_p x_p}$, the prediction is affected by not an addition, but a multiple, $e^{b_j x_j}$, thus the percent effect is $(e^{b_j x_j} - 1) \times 100$. For the male-vs-female coefficient, we then get

$$(e^{-0.0267524} - 1) \times 100 = -2.64\%$$

(c) [6 pt / 84 pts] Compute $\hat{\hat{CI}}_{\beta_{\texttt{educ}}, 95\%}$, the confidence interval for the slope parameter within the link function for the variable `educ` to the nearest four digits.

$$\begin{aligned}
\hat{\hat{CI}}_{\beta_{\texttt{educ}}, 1-\alpha} &= \left[b_{\texttt{educ}} \pm z_{1-\alpha/2} \cdot s_{b_{\texttt{educ}}}\right] \\
\hat{\hat{CI}}_{\beta_4, 95\%} &= \left[b_4 \pm 1.96 \cdot s_{b_4}\right] \\
&= [0.0582323 \pm 1.96 \cdot 0.0008373] = [0.0566, 0.0599]
\end{aligned}$$

(d) [1 pt / 85 pts] The confidence interval in the previous question is... circle one:
exact / approximate

(e) [6 pt / 91 pts] Predict the vocabulary score of a female, foreign-born, age 25 with 17yr of education. Round the score to the nearest whole number.

$$
\begin{aligned}
\hat{y} &= \text{round}\left(e^{b_0} e^{b_1 x_1} \cdot \ldots \cdot e^{b_p x_p}\right) \\
&= \text{round}\left(e^{b_0} e^{b_1(0)} e^{b_2(0)} e^{b_3(25)} e^{b_4(17)}\right) \\
&= \text{round}\left(e^{0.7761238} e^{0.0021438(25)} e^{0.0582323(17)}\right) \\
&= \text{round}(6.169809) = 6
\end{aligned}
$$

We now run the same model but this time omitting features `gender` and `nativeBorn`. Below is the output

```
             Estimate   Std. Error  z value Pr(>|z|)
(Intercept) 0.9130409  0.0139108   65.64   <2e-16 ***
age         0.0022436  0.0001438   15.61   <2e-16 ***
educ        0.0578375  0.0008323   69.49   <2e-16 ***

               Theta:  173175
           Std. Err.:  146404

 2 x log-likelihood:  -115635.4
```

Here are some values of the inverse CDF of the χ^2_{df} distribution:

```
               Probability less than the critical value
  df           0.90      0.95     0.975      0.99     0.999
------------------------------------------------------------
  1           2.706     3.841     5.024     6.635    10.828
  2           4.605     5.991     7.378     9.210    13.816
  3           6.251     7.815     9.348    11.345    16.266
  4           7.779     9.488    11.143    13.277    18.467
  5           9.236    11.070    12.833    15.086    20.515
  6          10.645    12.592    14.449    16.812    22.458
  7          12.017    14.067    16.013    18.475    24.322
```

(f) [2 pt / 93 pts] For the test of $H_0 : \beta_{\texttt{gender}} = \beta_{\texttt{nativeBorn}} = 0$ at $\alpha = 1\%$, would would be the critical value of the likelihood ratio test that the test statistic is compared to?

The degrees of freedom of this test is 2 because we are knocking out 2 features. Since $\alpha = 1\%$, we are looking for the 99%ile of the χ^2_2 which according we find in the table above on the second row and the fourth column: 9.210.

(g) [7 pt / 100 pts] Run the test of $H_0 : \beta_{\texttt{gender}} = \beta_{\texttt{nativeBorn}} = 0$ at $\alpha = 1\%$ and record your decision and write one sentence that interprets the result of the decision.

Since $\hat{\hat{\Lambda}} = 2\ln(\mathcal{L}_{\text{full}}) - 2\ln(\mathcal{L}_{\text{reduced}}) = -115304.3 - -115635.4 = 331.1 > \chi^2_{2,99\%} = 9.210$, we reject H_0. We can conclude that `gender` and `nativeBorn` are important in predicting vocabulary score in the context of the other variables and assuming the negative binomial model log-linear in the covariates.