Q1

(i). We have : $X_1, \cdots, X_n \sim \text{Unif}[0, \theta]$.

$$L(X_1, \cdots, X_n) = \prod_{j=1}^{n} \frac{1}{\theta} \mathbf{1}_{X_j \leq \theta} = \frac{1}{\theta^n} \mathbf{1}_{\max\{X_j\} \leq \theta}$$
$$\mathbf{E}[X] = \frac{\theta}{2}$$

So, $\hat{\theta} = \max\{X_j\}$ and $\widetilde{\theta} = 2\overline{X}$.

(ii). Clearly, for $\mu_2 = \mathbf{E}[\widetilde{\theta}]$ and $\sigma_2^2 = \text{Var}\ [\widetilde{\theta}]$

$$\hat{F}(x) = \mathbb{P}\left\{\max\{X_j\} \leq x\right\} = \prod_{j=1}^{n} \mathbb{P}\left\{X_j \leq x\right\}$$
$$= \left[\frac{x}{\theta}\right]^n \mathbf{1}_{\{0<x<\theta\}} + \mathbf{1}_{\{x \geq \theta\}}$$

And since, $\mu_2 = \theta$ and $\sigma_2^2 = \frac{1}{n^2}\left[n\frac{\theta^2}{12}\right] = \frac{\theta^2}{12n}$, we can use the Central Limit Theorem to conclude that for large n,

$$\widetilde{F}(x) = \mathbb{P}\left\{2\overline{X} \leq x\right\} = \mathbb{P}\left\{\frac{\overline{X} - \mu_2}{\sigma_2} \leq \sqrt{3}\frac{x-\theta}{\theta}\right\}$$
$$\simeq \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\sqrt{3n}\left(\frac{x-\theta}{\theta}\right)} e^{-x^2/2}\ dx\right] \mathbf{1}_{\{0<x<\theta\}} + \mathbf{1}_{\{x \geq \theta\}}$$

(iii). Clearly, for $\mu_1 = \mathbf{E}[\hat{\theta}]$ and $\sigma_1^2 = \text{Var}\ [\hat{\theta}]$

$$\mu_1 = \int_0^M x \cdot \frac{n}{\theta}\left[\frac{x}{\theta}\right]^{n-1}\ dx = \frac{n}{n+1}\theta$$
$$\mu_1^2 + \sigma_1^2 = \int_0^\theta x^2 \cdot \frac{n}{\theta}\left[\frac{x}{\theta}\right]^{n-1}\ dx = \frac{n}{n+2}\theta^2$$
$$\sigma_1^2 = \frac{n}{(n+1)^2(n+2)}\theta^2$$

So, the bias is $\frac{\theta}{n+1}$ and the variance is as above.

(iv). Let $\breve{\theta} = \left(1 + \frac{1}{n}\right)\max\{X_j\}$. Then $\mathbf{E}[\breve{\theta}] = \theta$ and $\text{Var}[\breve{\theta}] = \frac{1}{n(n+2)}\theta^2$.

(v). We calculate the Fisher information as follows :

$$\mathcal{I}_n(\theta) = \mathbf{E}\left[\left(\frac{\partial}{\partial\theta}\log\left[\frac{nx^{n-1}}{\theta^n}\right]\right)^2 \middle| \theta\right] = \mathbf{E}\left[\frac{n}{\theta^2}\middle|\theta\right] = \frac{n}{\theta^2}$$

So,

$$\sqrt{n}(\hat{\theta} - \theta) \to \mathcal{N}(0, \theta^2)$$

By delta method we get,

$$\sqrt{n}(\breve{\theta} - \theta) \to \mathcal{N}(0, \theta^2)$$

Q2

(a)

$$\mathbb{E}\,\widehat{\rho} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\,X_iY_i = \mathbb{E}\,XY = \rho\,,$$

so we have unbiasedness. For consistency, define $Z_i := X_iY_i$. Then $\widehat{\rho} = \frac{1}{n}\sum_{i=1}^{n} Z_i$, thus by SLLN for the sample $Z_1, \ldots, Z_n$ we have $\widehat{\rho} \to \mathbb{E}\,Z_1 = \rho$, where the convergence is almost sure, which gives us consistency.

(b) This depends on which formula for CLT students have in their cheat-sheet. I think it will be easy for them to read the formula wrong. For example,

$$\frac{\widehat{\rho}-\rho}{\sqrt{\text{var}(\widehat{\rho})}} \to \mathcal{N}(0,1)\,.$$

Using the hint, $\text{var}(\widehat{\rho}) = \frac{1}{n}\text{var}(X_1, Y_1) = \frac{1}{n}((1+2\rho^2)-\rho^2) = \frac{1}{n}(1+\rho^2)$, leading to

$$\sqrt{n}\frac{\widehat{\rho}-\rho}{\sqrt{1+\rho^2}} \to \mathcal{N}(0,1)\,.$$

(c) Using the Slutzsky's theorem, we can replace ρ in the denominator by $\widehat{\rho}$, using the fact that $\widehat{\rho}$ is consistent (this is minimally sufficient justification). Then it is standard algebra to get the answer :

$$P(-q_{0.975} < \sqrt{n}\frac{\widehat{\rho}-\rho}{\sqrt{1+\widehat{\rho}^2}} < -q_{0.975}) = 0.95$$

... $CI_{0.95}(\rho) = (\widehat{\rho} \pm q_{0.975}\frac{\sqrt{1+\widehat{\rho}^2}}{\sqrt{n}})$

Q3

(a) Write $F_n(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{[X_i\le x]}$. Then

$$\widehat{f}(x) = \frac{1}{2nh}\sum_{i=1}^{n}\mathbf{1}_{\left[X_i\in(x-h,x+h]\right]} = \frac{1}{2nh}\sum_{i=1}^{n}\mathbf{1}_{\left[1\le\frac{x-X_i}{h}<h\right]} = \frac{1}{nh}\sum_{i=1}^{n}K\Big(\frac{x-X_i}{h}\Big)\,,$$

where $K(y) = \frac{1}{2}\mathbf{1}_{[|y|\le 1]}$. So the kernel corresponds to $U[-1,1]$ distribution.

(b)

$$\mathbb{E}\,\widehat{f}(x) = \frac{1}{2h}\Big[\mathbb{E}\,F_n(x+h) - \mathbb{E}\,F_n(x-h)\Big] = \frac{1}{2h}\Big[F(x+h)-F(x-h)\Big]$$

If $h \to 0$, then the previous expression is the definition for derivative of F, thus $\mathbb{E}\,\widehat{f}(x) \to f(x)$ for $h \to 0$. This leads to $\text{bias}(\widehat{f}(x)) \to 0$.

Note that students will probably start by writing down the definition of bias. This leads to the same conclusion.

Q4

(a) As explained on slide 342 of the course, the uncentered coefficient of determination measures the proportion of the squared norm of $\boldsymbol{Y}$ explained by the fitted values $\hat{\boldsymbol{Y}}$. As a ratio of two positive quantities, $R0^2$ is necessarily positive. Moreover, we have :

$$\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y} + (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y} = \hat{\boldsymbol{Y}} + (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}$$

where the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is an orthogonal projection operator. We can hence apply the Pythagorean theorem to obtain

$$\|\boldsymbol{Y}\|^2 = \|\hat{\boldsymbol{Y}}\|^2 + \|(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}\|^2,$$

and hence $\|\boldsymbol{Y}\|^2 \geq \|\hat{\boldsymbol{Y}}\|^2 \Rightarrow R_0^2 \leq 1$.
Alternative solution : *Use spectral norm of* $\boldsymbol{H}$ *:*

$$\|\hat{\boldsymbol{Y}}\|^2 = \|\boldsymbol{H}\boldsymbol{Y}\|^2 \leq \|\boldsymbol{H}\|_2^2 \|\boldsymbol{Y}\|^2.$$

Since $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ *is an orthogonal projection it has maximal eigenvalue 1, yielding* $\|\boldsymbol{H}\|_2^2 = 1$ *and finally* $R_0^2 \leq 1$.
The inequality is saturated when $p = n$ indeed, is this case $\boldsymbol{X}$ is square and hence invertible from the full-rank assumption, yielding $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{X}\boldsymbol{X}^{-1}\boldsymbol{X}^{-T}\boldsymbol{X}^T = I$ and trivially $\hat{\boldsymbol{Y}} = \boldsymbol{Y} \Rightarrow R_0^2 = 1$.

(b) The new fitted values are given by

$$\begin{aligned}
\tilde{\boldsymbol{Y}} &= \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{Y} \\
&= (\boldsymbol{X}, \boldsymbol{x}_{p+1}) \begin{pmatrix} (\boldsymbol{X}^T\boldsymbol{X})^{-1} & \boldsymbol{0} \\ 0 & 1/\|\boldsymbol{x}_{p+1}\|^2 \end{pmatrix} \begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{x}_{p+1}^T \end{pmatrix} \boldsymbol{Y} \\
&= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \;+\; \frac{\boldsymbol{x}_{p+1}^T\boldsymbol{Y}}{\|\boldsymbol{x}_{p+1}\|^2}\boldsymbol{x}_{p+1} \\
&= \hat{\boldsymbol{Y}} \;+\; \frac{\boldsymbol{x}_{p+1}^T\boldsymbol{Y}}{\|\boldsymbol{x}_{p+1}\|^2}\boldsymbol{x}_{p+1},
\end{aligned}$$

since $\boldsymbol{X}^T\boldsymbol{x}_{p+1} = 0$ by assumption. From Pythagore again we get

$$\|\tilde{\boldsymbol{Y}}\|^2 = \|\hat{\boldsymbol{Y}}\|^2 + |\boldsymbol{x}_{p+1}^T\boldsymbol{Y}|^2,$$

which yields :

$$\frac{\|\hat{\boldsymbol{Y}}\|^2}{\|\boldsymbol{Y}\|^2} \leq \frac{\|\tilde{\boldsymbol{Y}}\|^2}{\|\boldsymbol{Y}\|^2}.$$

A higher uncentered coefficient of determination R_0^2 is not necessarily desirable. Indeed, when we add too many covariates to the model, we explain more and more of the variations in the data, including after a certain threshold the noise variations, which is of course undesirable. This phenomenon is called *overfitting*.

(c) We have $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ and

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n),$$

by assumption of the Gaussian linear model. From properties of Gaussian random vector, we know that $\hat{\boldsymbol{\beta}}$ is again a Gaussian random vector (as linear transformation of a Gaussian rv), with exact distribution given by

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}).$$

Similarly the distribution of $\boldsymbol{c}^T\hat{\boldsymbol{\beta}}$ for some arbitrary vector $\boldsymbol{c} \in \mathbb{R}^p$ is Gaussian and given by

$$\boldsymbol{c}^T\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{c}^T\boldsymbol{\beta}, \sigma^2\boldsymbol{c}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c}).$$

(d) Choosing $\boldsymbol{c} = \boldsymbol{e}_i$ where $\boldsymbol{e}_i$ is the ith canonical vector of $\mathbb{R}^p$, we get

$$\hat{\beta}_i = \boldsymbol{e}_i^T\hat{\boldsymbol{\beta}}.$$

From the previous question and the symmetry of the standard normal distribution we hence get that

$$\mathbb{P}\left\{ \frac{\left|\hat{\beta}_i - \beta_i\right|}{\sqrt{\sigma^2(\boldsymbol{X}^T\boldsymbol{X})_{ii}^{-1}}} \leq \Phi_{\alpha/2} \right\} = 1 - \alpha,$$

where $\Phi_{\alpha/2}$ denotes the quantile $\alpha/2$ of the standard normal distribution. We get hence the following confidence intervals for each $i = 1, \ldots, p$,

$$IC_{1-\alpha}(\beta_i) = \left[\hat{\beta}_i - \Phi_{\alpha/2}\sqrt{\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{ii}}, \hat{\beta}_i + \Phi_{\alpha/2}\sqrt{\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{ii}}\right].$$

(e) The width of the confidence interval is $2\Phi_{\alpha/2}\sqrt{\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{ii}}$. Notice that it is proportional to $\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{ii}}$. In the event of multicollinearity, the matrix $(\boldsymbol{X}^T\boldsymbol{X})$ will be very ill-conditioned and hence the numerical inversion very unstable, yielding inaccurate and potentially very large coefficients $\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{ii}}$, which may blow up the width of the confidence interval.

Q5

(a) By plugging the given value for μ_i into the Poisson density function one has :

$$L(y_i|\alpha, \beta) = c.\exp(-e^{\alpha+\beta x_i})\exp(\alpha y_i + \beta x_i y_i)$$

where c is a constant.
Therefore the loglikelihood is

$$l(\alpha, \beta) = -\sum_{i=1}^{n} e^{\alpha+\beta x_i} + \sum_{i=1}^{n} \alpha y_i + \beta x_i y_i \ + constant$$

(b) By factorization theorem $t_1 = \sum y_i$ and $t_2 = \sum x_i y_i$ are sufficient statistics for α and β.

(c) Since

$$l(\alpha, \beta) = -\sum_{i=1}^{n} e^{\alpha+\beta x_i} + \alpha t_1 + \beta t_2 + constant$$

thus

$$\frac{\partial l}{\partial \alpha} = 0 \to t_1 = \sum_{i=1}^{n} e^{\alpha+\beta x_i}$$

and

$$\frac{\partial l}{\partial \beta} = 0 \to t_2 = \sum_{i=1}^{n} x_i e^{\alpha+\beta x_i}$$

(d) if $\beta = 0$, $l(\alpha, \beta) = -\sum_{i=1}^{n} e^{\alpha} + \alpha t_1$. This is maximized with respect to α with $\alpha^* = \log(t_1/n)$

(e) One could refer to

$$2\{l(\hat{\alpha}, \hat{\beta}) - l(\alpha^*, 0)\}$$

having χ^2_1 distribution by Wilk's theorem.

Bonus Q Let $Y_n \sim \text{Bernoulli}(1/n)$ and define $X_n = nY_n$. Let $\varepsilon > 0$. Then

$$\mathbb{P}[|X_n| > \varepsilon] = \mathbb{P}[|Y_n| > 0] = \frac{1}{n}.$$

for all $n > \varepsilon$. It follows that

$$X_n \xrightarrow{p} 0.$$

On the other hand, for all n,

$$\mathbb{E}[X_n] = n \times \frac{1}{n} + 0 \times \left(1 - \frac{1}{n}\right) = 1.$$