1. Consider the linear model $y_i = \beta_0 + \beta_1 x_{i1} \cdot + \beta_p x_{ip} + e_i$ for $i = 1, \dots, n$ where $e \sim N(0, \sigma^2 I_N)$.

(a) Let $\hat{y}_i$ denote the $i$th fitted value and let $\hat{y}_{i(i)}$ denote the predicted response value for the $i$th subject in the regression. Write the difference $\hat{y}_i - \hat{y}_{i(i)}$ in terms of the $i$th standardized residual and the $i$th leverage.

*Solution.*

$$
\begin{aligned}
\hat{y}_i - \hat{y}_{i(i)} &= (y_i - \hat{y}_{i(i)}) - (y_i - \hat{y}_i) \\
&= \hat{e}_{[i]} - \hat{e}_i \\
&= \frac{\hat{\sigma}}{\sqrt{1 - h_{ii}}} r_i - \hat{e}_i \\
&= \frac{\hat{\sigma}}{\sqrt{1 - h_{ii}}} \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} - \hat{e}_i \\
&= \hat{e}_i \left( \frac{1}{1 - h_{ii}} - 1 \right) \\
&= \hat{e}_i \frac{h_{ii}}{1 - h_{ii}}
\end{aligned}
$$

□

(b) Calculate the distribution of $\hat{y}_i - \hat{y}_{i(i)}$.

*Solution.* Since $\hat{e}_i \sim N(0, \sigma^2(1 - h_{ii})$, this quantity is also normal with mean $\mathbb{E}[\hat{e}_i \frac{h_{ii}}{1 - h_{ii}}] = \frac{h_{ii}}{1 - h_{ii}} \mathbb{E}[\hat{e}_i] = 0$ and variance $\operatorname{Var}(\hat{e}_i \frac{h_{ii}}{1 - h_{ii}}) = (\frac{h_{ii}}{1 - h_{ii}})^2 \operatorname{Var}(\hat{e}_i) = (\frac{h_{ii}}{1 - h_{ii}})^2 \sigma^2 (1 - h_{ii}) = \sigma^2 \frac{h_{ii}^2}{1 - h_{ii}}$. So $\hat{y}_i - \hat{y}_{i(i)} \sim N(0, \sigma^2 \frac{h_{ii}^2}{1 - h_{ii}})$. □

(c) Can you obtain an unbiased estimator for $\sigma^2$ that is independent of $\hat{y}_i - \hat{y}_{i(i)}$? If yes, specify such an unbiased estimator. If no, explain why.

*Proof.* No, this is not possible. Suppose some variable $s^2$ is an unbiased estimator for $\sigma^2$. The farther that $\hat{y}_i - \hat{y}_{i(i)}$ is from 0, the higher we expect $s^2$ to be. Conversely, we have for example that $\mathbb{P}(|\hat{y}_i - \hat{y}_{i(i)}| < \epsilon \mid s^2 = 1) < \mathbb{P}(|\hat{y}_i - \hat{y}_{i(i)}| < \epsilon \mid s^2 = 2)$, for any $\epsilon > 0$. Therefore, these quantities cannot be independent. □

2. I fit a linear model to the usual data $y_1, \dots, y_n$ and $x_{ij}$ for $i = 1, \dots, n$ and $j = 1, \dots, p$. Let $RSS$ denote the residual sum of squares and $\hat{e}$ denote the vector of residuals.

I have been told that the data on an explanatory variable has not been collected. More specifically, the right model here is apparently

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \gamma z_i + e_i$$

where $e_1, \dots, e_n$ are uncorrelated mean zero errors with constant variance $\sigma^2$. Here $z_1, \dots, z_n$ denote the values of a variable that has not been observed unfortunately.

(a) Is $RSS/(n - p - 1)$ an unbiased estimator of $\sigma^2$? If yes, explain with reason. If no, calculate the bias.

*Proof.* No. Let $H$ denote the projection map onto the column space of $X$ and note that $y = X\beta + z\gamma + e$. So

$$\begin{aligned}\hat{e} &= (I-H)y \\ &= (I-H)(X\beta + \gamma z + e) \\ &= (I-H)X\beta + (I-H)\gamma z + (I-H)e \\ &= (I-H)(\gamma z + e)\end{aligned}$$

$$\begin{aligned}\mathbb{E}[RSS] &= \mathbb{E}[\hat{e}^T\hat{e}] \\ &= \mathbb{E}[(\gamma z + e)^T(I-H)^2(\gamma z + e)] \\ &= \mathbb{E}[(\gamma z + e)^T(I-H)(\gamma z + e)] \\ &= \mathbb{E}[(\gamma z)^T(I-H)\gamma z] + \mathbb{E}[e^T(I-H)\gamma z] + \mathbb{E}[(\gamma z)^T(I-H)e] + \mathbb{E}[e^T(I-H)e] \\ &= (\gamma z)^T(I-H)(\gamma z) + 0 + 0 + \sigma^2(n - tr(H)) \\ &= (\gamma z)^T(I-H)(\gamma z) + \sigma^2(n-p-1)\end{aligned}$$

Therefore,

$$\mathbb{E}\left[\frac{RSS}{n-p-1}\right] = \sigma^2 + \frac{(\gamma z)^T(I-H)(\gamma z)}{n-p-1}$$

so the bias is $(\gamma z)^T(I-H)(\gamma z)/(n-p-1)$. □

(b) Is the sum of the residuals $\hat{e}_i$ zero? Answer with reason.

*Proof.* Yes, of course. This is simply a result of the fact that 1) the residuals are the projection of $y$ onto the orthogonal complement of $X$, and 2) the vector $u$ of all ones is in the column space of $X$. Thus we have

$$\sum \hat{e}_i = u^T\hat{e} = u^T(I-H)y = ((I-H)u)^T y = 0^T y = 0.$$

□

(c) What is the expected value of $\hat{e}$?

*Proof.* Recall from part (a) that $\hat{e} = (I-H)(\gamma z + e)$, so

$$\begin{aligned}\mathbb{E}[\hat{e}] &= \mathbb{E}[(I-H)(\gamma z + e)] \\ &= \mathbb{E}[(I-H)\gamma z] + (I-H)\mathbb{E}e \\ &= (I-H)\gamma z\end{aligned}$$

□

3. Consider the usual data on response $y_1, \ldots, y_n$ and explanatory variable data $x_{ij}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Suppose $y_i$ can be modelled as a Poisson random variable with mean $\lambda_i$. Moreover suppose $y_1, \ldots, y_n$ can be assumed to be independent.

(a) Write down the form of the canonical GLM.

*Solution.* The Poisson pdf can be written as

$$f(x) = \frac{1}{x!}\exp(x\log\lambda_i - \lambda_i).$$

Letting $\theta_i = \log\lambda_i$, $b(\theta_i) = e^{\theta_i}$, and $\phi_i = 1$ shows that $f(x)$ takes the proper form. The model, using the canonical link function, is then

$$\log\lambda_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

□

(b) Write down the log-likelihood function of $\beta$.

*Proof.* Since $\lambda_i = \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})$, the log-likelihood function is

$$\begin{aligned} l(\beta) &= \log\prod_{i=1}^{n} e^{-\lambda_i}\frac{\lambda_i^{y_i}}{y_i!} \\ &= \sum_{i=1}^{n}\left(-\lambda_i + y_i\log\lambda_i - \sum_{j=1}^{y_i}\log j\right) \\ &= \sum_{i=1}^{n}\left(-\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) + y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) - \sum_{j=1}^{y_i}\log j\right) \end{aligned}$$

□

4. The three plots in Figure 1 give the three partial regression or added variable plots for this regression.

(a) Can you identify which plot corresponds to which variable? Provide reasoning.

*Solution.* The first plot corresponds to x2, the second to x3, and the third to x1. This is because the slope of a line fit to these plots should be the coefficient of the corresponding variable in the linear model. These choices are the only ones that make the slopes line up with the plots. □

(b) Consider the data in the first added variable plot. Suppose I fit a linear model to the y-variable based on the x-variable. What is the value of the RSS for this regression?

*Solution.*The RSS of this model is equivalent to that of the original model. Since the residual standard error is

$$\hat{\sigma} = 1.827 = \sqrt{\frac{RSS}{n-p-1}} = \sqrt{\frac{RSS}{116}},$$

we know that $RSS = 387.1998$. Therefore, this is also the RSS for the regression of the $y$ variable against the $x$ variable in the first plot. □

5. Determine whether each of the following statements is true or false. Provide reasons in each case.

(a) The magnitude of a predicted residual is never smaller than the magnitude of the corresponding residual.

*Proof.* True. Since $0 \leq h_{ii} \leq 1$, we know that $0 \leq 1 - h_{ii} \leq 1$. So $\hat{e}_{[i]} = \frac{\hat{e}_i}{1-h_{ii}} \geq \hat{e}_i$. □

(b) The leverage of the $i$th subject depends on the value of $y_i$.

*Proof.*False. The leverage of the $i$th subject is an entry in the projection matrix onto the column space of $X$. This projection map depends only on the column space of $X$, thus so does the $i$th leverage. □

(c) Model selection via AIC tends to produce smaller models than BIC.

*Proof.* False. Model selection via $AIC$ attempts to minimize the quantity $n \log\left(\frac{RSS(m)}{n}\right) + 2(1 + p(m))$, while $BIC$ attempts to minimize $n \log\left(\frac{RSS(m)}{n}\right) + (\log n)(1 + p(m))$. Since $(\log n)(1 + p(m))$ is larger than $2(1 + p(m))$ whenever $\log n > 2$, which is whenever $n \geq 8$. This should always be true if we are using this model, since the linear model is not very useful with less than 8 data points. □

(d) The GCV is a computationally simpler model selection technique than PRESS.

*Proof.* True. The expression for the $GCV$ score is $GCV(m) = \left(1 - \frac{1+p(m)}{n}\right)^{-2} RSS(m) \approx RSS(m) + 2(1 + p(m))$, which does not involve any summation. However, $PRESS(m) = \sum_{i=1}^{n} \frac{\hat{e}_i^2(m)}{1-h_{ii}(m))^2}$ involves a summation over $\{1, \ldots, n\}$, thus is more computationally heavy. □

(e) The optimal model selected by Mallow's $C_p$ criterion can have a $C_p$ value that is more that $p + 1$.

*Proof.* False. The full model always has a $C_p$ of $p + 1$, thus would be preferred to any model with a greater $C_p$. So no model with a $C_p$ greater than $p + 1$ would ever be selected. □