Homework 2 Solution

STA2101F 2020

Due November 5 2020 11.59 pm

Homework to be submitted through Quercus

Please submit both the .Rmd and the .pdf file for this homework. (It is not possible to annotate .html files in Quercus.) In the .pdf document you should just include properly formatted output, not unedited chunks of ##R output. See the solutions to HW 1 Q 4, for example.

You are welcome to discuss the questions with others, but the solutions and code must be written independently.

1. *Two-way layout with replication.* If we have a balanced two-factor experiment with replication, a suitable linear model might be

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \tag{1}$$

where $i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K$. In this model we separate the main effects of factors A and B from the interaction effect. If the observations were set out in a two-way layout, there would be K observations in each cell.

(a) If we assume that ϵ_{ijk} are all independent and identically distributed $N(0, \sigma^2)$, argue that the variation in a single cell, $\Sigma_k(y_{ijk} - \bar{y}_{ij.})^2$ estimates σ^2 with $K-1$ degrees of freedom. From this conclude that $\Sigma_{ij}\Sigma_k(y_{ijk} - \bar{y}_{ij.})^2$ estimates σ^2 with $IJ(K-1)$ degrees of freedom. Suggest and construct a plot that might be used to check on the assumption of constant variance.

Solution: Since $\epsilon_{ijk} \overset{\text{iid}}{\sim} N(0, \sigma^2)$, for fixed i and j:

$$\frac{\sum_k (y_{ijk} - \bar{y}_{ij\cdot})^2}{\sigma^2} \sim \chi^2_{(K-1)} \quad \text{(using Cochran's theorem)}$$

For $i' \neq i$ and $j' \neq j$, the random variables $\frac{\sum_k (y_{ijk} - \bar{y}_{ij\cdot})^2}{\sigma^2}$ and $\frac{\sum_k (y_{i'j'k} - \bar{y}_{i'j'\cdot})^2}{\sigma^2}$ are independent. The sum of these random variables are the sum of two independent χ^2 random variables with $K-1$ df each, which yields a χ^2 with $2(K-1)$ df. Repeating this by summing over i and j:

$$\sum_{ij} \frac{\sum_k (y_{ijk} - \bar{y}_{ij\cdot})^2}{\sigma^2} \sim \chi^2_{IJ(K-1)}$$

In Figure 1, the log of sample variance for each cell is plotted against the cell mean. A positive pattern can clearly be observed, which indicates a violation of constant variance.

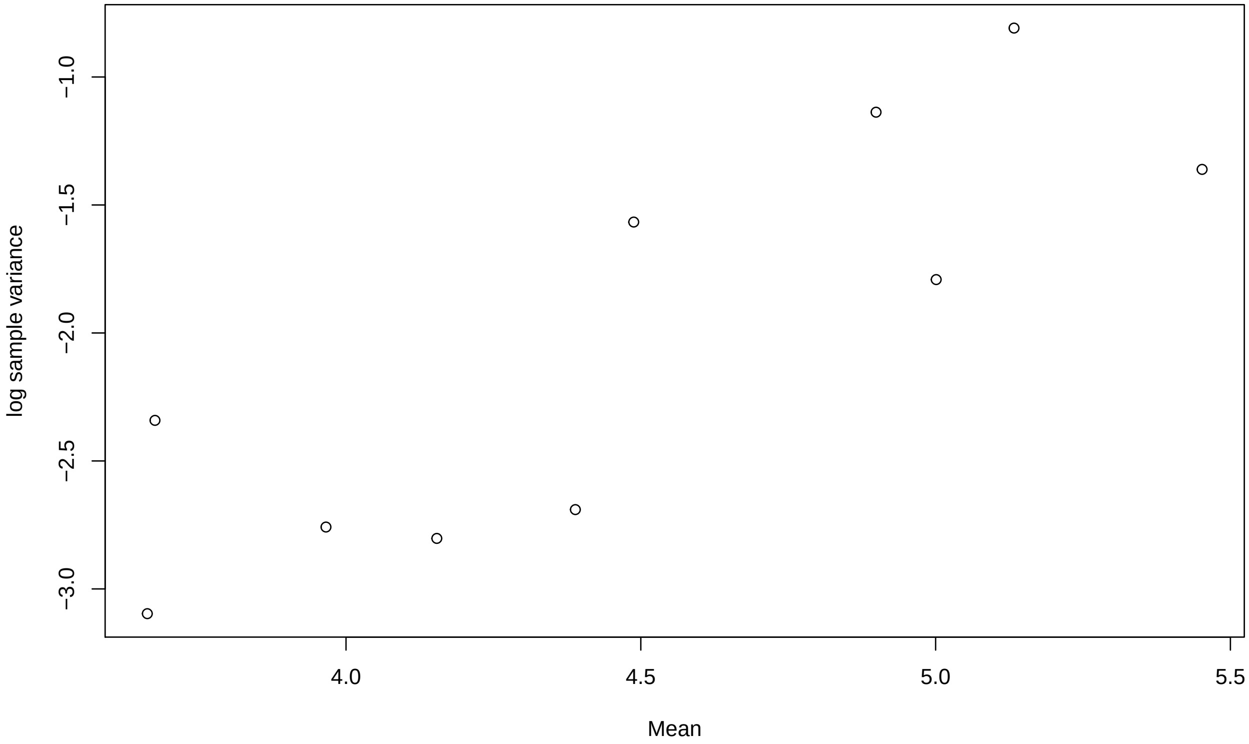


Figure 1: Confidence intervals for the variance estimates

(b) The data set `butterfat` in `library(faraway)` is a two-way layout with 10 observations per cell. Present a two-way table of cell means, showing as well the row and column means. Is there evidence of interaction between the two factors? Explain.

Solution: The following table shows the cell means of the butterfat content for the Breed-Age combinations as well as the row and column means:

	Two Years	Mature	Row Means
Ayrshire	3.966	4.154	4.060
Canadian	4.488	4.389	4.439
Guernsey	4.899	5.001	4.950
Holstein-Fresian	3.663	3.676	3.670
Jersey	5.133	5.452	5.292
Column Means	4.430	4.534	4.482

The cow's age appears to have a positive effect on the average butterfat content for all breeds except the Canadian and Holstein-Fresian cows whose average butterfat content was almost equal or lower. These differential age effects suggests that the breed and age factors potentially interact in their effect on butterfat content of the milk.

(c) Fit the linear model of equation (1) to this data and give the analysis of variance table. Give an algebraic expression for the sum of squares due to interaction in the analysis of variance table. What is the estimate of σ^2? Does it have $IJ(K-1)$ degrees of freedom?[1]

Solution: Model (1) was fitted to the butterfat data. The resulting ANOVA table is show by the following:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(Breed)	4	34.321	8.580	49.565	0
factor(Age)	1	0.274	0.274	1.58	0.212
factor(Breed):factor(Age)	4	0.514	0.128	0.742	0.566
Residuals	90	15.580	0.173		

The total sum of squares can be partitioned as follows:

$$\begin{aligned}\sum_{ijk}(y_{ijk}-\bar{y})^2 &= \sum_{ijk}\left((y_{ijk}-\bar{y}_{ij.})+(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y})+(\bar{y}_{i..}-\bar{y})+(\bar{y}_{.j.}-\bar{y})\right)^2\\ &= \underbrace{\sum_{ijk}(y_{ijk}-\bar{y}_{ij.})^2}_{\text{Residuals}}+\underbrace{K\sum_{ij}(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y})^2}_{\text{Interaction}}+\\ &\underbrace{KJ\sum_{i}(\bar{y}_{i..}-\bar{y})^2}_{\text{Main effect of }\alpha}+\underbrace{KI\sum_{j}(\bar{y}_{.j.}-\bar{y})^2}_{\text{Main effect of }\beta}.\end{aligned}$$

The summary output in R shows that the estimated variance is 0.1731 with 90 df. We can confirm that the df is the same as in (a): $IJ(K-1) = 5\times 2\times(10-1) = 90$.

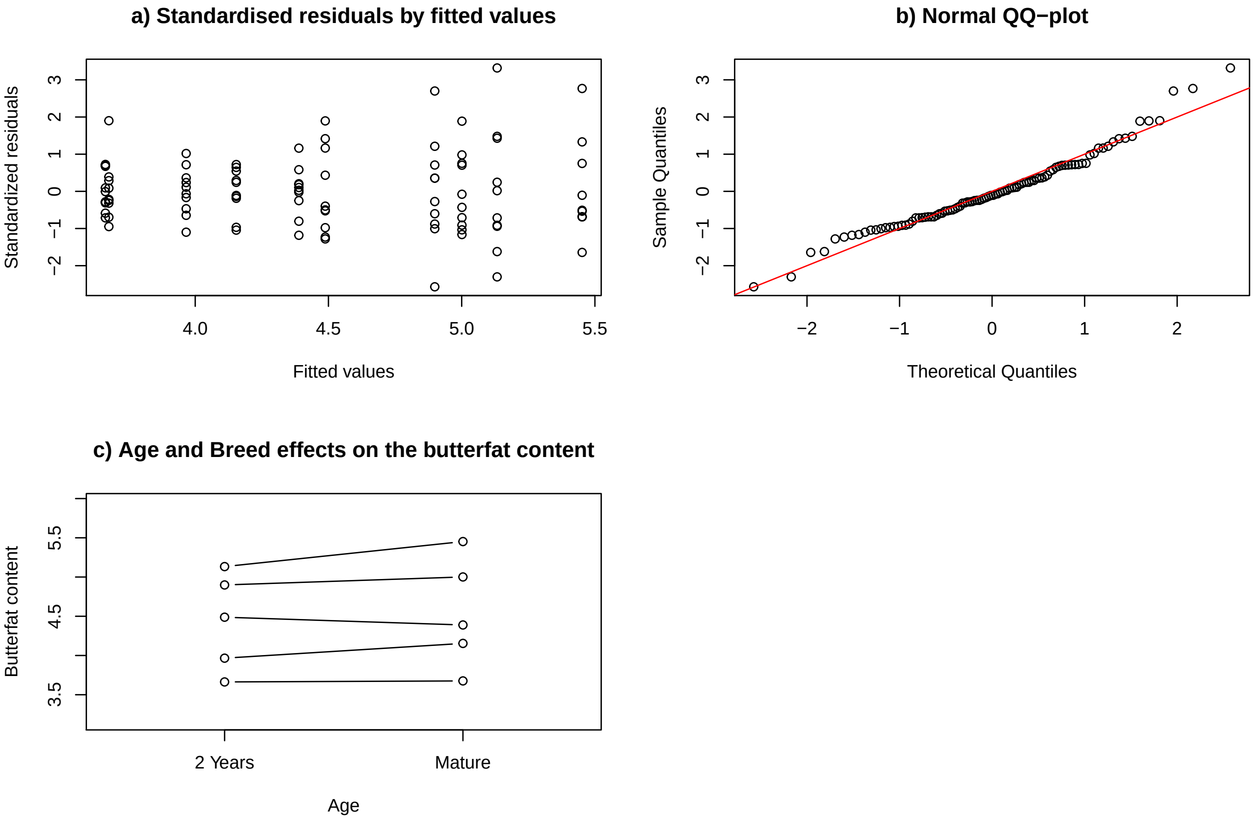
(d) Provide at most three plots of the data and/or fitted model. What information is available from these plots?

Solution: Figure 2 - a) shows the standardized residuals plotted against the fitted values. The spread of the standardized residuals appears to increase over the fitted values. Figure 2 - b) shows the QQ-plot for the standardized residuals. The plots lie approximately on the diagonal line shown in red, which indicates that the normality assumption appears to be appropriate. Figure 2 - c) shows lines joining the average butterfat contents of two-year-old cows and those of mature cows for each breed. Since the lines are not parallel, there is a possibility of a Breed-Age interaction.

(e) Is the best breed in terms of butterfat content clearly superior to the second best breed? Why or why not? Explain your calculations.

Solution: Since the data showed showed a weak evidence for the interaction

[1] if not you've made a mistake somewhere.



a) Standardised residuals by fitted values

b) Normal QQ−plot

c) Age and Breed effects on the butterfat content

Figure 2: Plots of the data and the fitted model

effects, we ignore the interaction effects for this question. The following table shows the sample means $\bar{y}_{i..} = \frac{\sum_j \sum_k y_{ijk}}{JK}$:

i	Breed	Butterfat
1	Ayrshire	4.060
2	Canadian	4.439
3	Guernsey	4.950
4	Holstein-Fresian	3.670
5	Jersey	5.292

The two largest mean butterfat contents came from Jersey ($i = 5$) and Guernsey ($i = 3$) breeds. The null hypothesis is $H_0 : \alpha_5 - \alpha_3 = 0$. Assuming that age and breed don't interact, an unbiased estimator of this difference is $\bar{Y}_{5..} - \bar{Y}_{3..}$. The variance for the difference is $Var(\bar{Y}_{5..} - \bar{Y}_{3..}) = \frac{\sigma^2}{10}$. Therefore, an unbiased variance estimator is $\frac{MSE}{10}$. MSE can be computed as follows:

$$MSE = \frac{SSE}{IJ(K-1)} = \frac{\sum_i \sum_k (y_{ik} - \bar{y}_{i\cdot})}{90}$$

From the ANOVA table in (b), $MSE = 0.173$. Based on this, we can use a T-test to test our hypothesis. The test statistic is:

$$T = \frac{\bar{y}_{5\cdot} - \bar{y}_{3\cdot}}{\sqrt{MSE/10}} = \frac{5.2925 - 4.9500}{\sqrt{0.173/10}} = 2.6040$$

Under H_0, the test statistic is distributed as $T(90)$. The corresponding p-value is 0.0103. We conclude that we have evidence that the two largest mean butterfat contents differ.

2. (*FELM Exercise 2.2*) The dataset `wbca` in the library `faraway` comes from a study of breast cancer in Wisconsin (see `?wbca`). There are 681 cases of potentially cancerous tumours, of which 238 are actually malignant. Malignancy of a tumor is traditionally determined by an invasive surgical procedure. The purpose of this study was to determine whether a new procedure called fine needle aspiration, which draws only a small sample of tissue, could be effective in determining tumor status.

   (a) Fit a binomial regression with `Class` as the response and the other nine variables as predictors. Report the residual deviance and associated degrees of freedom. Can this information be used to determine if this model fits the data? Explain.
   Solution: The residual deviance and associated df after fitting the model is shown in Table 4. The residual deviance is 89.46 with 671 df. Here, the residual deviance does not provide any useful information about the model fit. As we will show in Q3 - b) below, in the case of a binary response, the residual deviance is not a function of the response y but the estimated regression parameters.

	Df (Deviance)	Deviance	Df (Residual Deviance)	Residual Deviance
NULL			680	881.388
Adhes	1	422.2865	679	459.102
BNucl	1	215.4552	678	243.647
Chrom	1	57.8114	677	185.835
Epith	1	21.9961	676	163.839
Mitos	1	18.7107	675	145.128
NNucl	1	18.4231	674	126.705
Thick	1	35.3506	673	91.355
UShap	1	1.8313	672	89.523
USize	1	0.0592	671	89.464

(b) Use the `step` function to choose final model.

Solution: The following table shows the summary of the final model selected from the backward stepwise regression:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	11.033	1.363	8.094	0.000
Adhes	-0.398	0.129	-3.080	0.002
BNucl	-0.419	0.102	-4.111	0.000
Chrom	-0.568	0.184	-3.085	0.002
Mitos	-0.646	0.363	-1.777	0.076
NNucl	-0.292	0.124	-2.358	0.018
Thick	-0.622	0.158	-3.937	0.000
UShap	-0.254	0.178	-1.423	0.155

(c) For a patient with the following measured values of the predictors: `Adhes` = 1, `BNucl` = 1, `Chrom` = 3, `Epith` = 2, `Mitos` = 1, `NNucl` = 1, `Thick` = 4, `UShap` = 1, `USize` = 1, predict the `Class` of the tumor, using the full model and the final model. Give confidence intervals for your predictions, and explain how you obtained them.

Solution: On the logit scale, the prediction for the patient using the model selected from the backward stepwise regression is:

$$\hat{\beta}_{Adhes} \cdot 1 + \hat{\beta}_{BNucl} \cdot 1 + \hat{\beta}_{Chrom} \cdot 3 + \cdots + \hat{\beta}_{UShap} \cdot 1 = 4.834$$

The variance of this prediction is:

$$\begin{aligned}\sigma^2_{\text{pred}} &= Var(\hat{\beta}_{Adhes} \cdot 1 + \hat{\beta}_{BNucl} \cdot 1 + \cdots + \hat{\beta}_{UShap} \cdot 1)\\ &= Var(\hat{\beta}_{Adhes}) + Var(\hat{\beta}_{BNucl}) + \cdots + Var(\hat{\beta}_{USize}) +\\ &\quad 2Cov(\hat{\beta}_{Adhes}, \hat{\beta}_{BNucl}) + 2 \times 3Cov(\hat{\beta}_{Adhes}, \hat{\beta}_{Chrom}) \cdots + 2 \times 4Cov(\hat{\beta}_{Thick}, \hat{\beta}_{UShap})\end{aligned}$$

An estimated variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ can be obtained using an iterative method (See Q5 in Assignment 1). Based on the output from R, $\hat{\sigma^2}_{\text{pred}} = 0.3382$.

Hence, a 95% prediction intervals for this prediction on the logit scale is $4.834 \pm z_{0.975} \times \sqrt{0.3382}$ or $(3.695, 5.974)$. Since the logit function is strictly monotone, we apply its inverse (i.e. logistic function) to obtain the prediction interval on the original scale: $\left(\frac{1}{1+\exp(-3.695)} = 0.9757, \frac{1}{1+\exp(-5.974)} = 0.9975\right)$.
Similarly, the same procedure can be repeated for the full model. The 95% prediction interval is (0.9759, 0.9976).

(d) Suppose that a cancer is classified as benign (`Class` = 1) if $p(\hat{\beta}) > 0.5$, and otherwise is classified as malignant. Give a table showing the misclassification errors if this method is applied to the current data, using the final model from (b). Give a similar table but using the cutoff 0.9 instead.
Solution: The following table shows the misclassification error with a cut-off of 0.5 where the column represents the actual data and the row, the predicted outcome:

	Benign	Malignant
Benign	0.637298	0.016153
Malignant	0.013216	0.333333

The following table shows the misclassification error with a cut-off of 0.9:

	Benign	Malignant
Benign	0.627019	0.001468
Malignant	0.023495	0.348018

(e) The error estimates in (d) are overly optimistic, because they are computed on the same data that was used to fit the model. A better strategy is to fit the model on training data and assess the errors on test data. Choose a random 1/3 of the data to hold out as a test set, and find the best fitting model on the remainder. Construct the tables of misclassification errors for the cutoffs 0.5 and 0.9, and compare to the results in (d). *Note: If you have issues with convergence on the randomly selected training set, you could try following Faraway's suggestion of choosing each 3rd case to construct the test set.*

Solution: With a random test dataset, the following table shows the misclassification error matrix with a cut-off of 0.5 where the column represents the actual data and the row, the predicted outcome:

	Benign	Malignant
Benign	0.612335	0.022026
Malignant	0.022026	0.343612

The misclassification error matrix with a cut-off of 0.9 is shows as follows:

	Benign	Malignant
Benign	0.585903	0.004405
Malignant	0.048458	0.361233

As expected, the error estimates, the off-digonal elements in the matrices above, increase when we use a random test dataset compared with the results in (d).

3. *SM, Exercise 10.4.1* Suppose $y_1, \dots, y_n$ are independent Binomial(n_i, p_i) random variables, with $p_i = \exp(x_i^T\beta)/\{1+\exp(x_i^T\beta)\}, i = 1, \dots, n$.

   (a) Derive expressions for the log-likelihood function, the equations defining the maximum likelihood estimator of β, and the residual deviance.

   Solution: The kernel of the likelihood function for our model is:

$$L(\boldsymbol{\beta}) \propto \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{n_i-y_i}$$

By taking the log, we obtain the log-likelihood:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log p_i + (n_i - y_i)\log(1-p_i)$$

Taking the first derivative and setting it equal to 0, we obtain the score equations:

$$\begin{cases} \frac{\partial l(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} y_i(1-p_i) - (n_i - y_i)p_i = 0 \\ \frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij}y_i(1-p_i) - x_{ij}(n_i - y_i)p_i = 0 \quad (j = 1, \cdots, q) \end{cases}$$

It can be shown that the log-logistic function is concave. Since the log-likelihood is sum of log-logistic functions, it is itself concave. Hence, the score equations have a unique solution which is the MLE.

To obtain the residual deviance D, we derive the log-likelihood for the saturated model:

$$l_S(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log\left(\frac{y_i}{n_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i}\right)$$

The log-likelihood for the fitted model is:

$$l_M(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log \hat{p}_i + (n_i - y_i)\log(1-\hat{p}_i)$$

Then, the residual deviance D is:

$$\begin{aligned} D &= -2(l_M(\boldsymbol{\beta}) - l_S(\boldsymbol{\beta})) \\ &= -2\sum_{i=1}^{n} y_i \log \hat{p}_i + (n_i - y_i)\log(1-\hat{p}_i) - y_i \log\left(\frac{y_i}{n_i}\right) - (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i}\right) \\ &= 2\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{n_i\hat{p}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\hat{p}_i}\right) \end{aligned}$$

(b) Use the approximation $\log(1+x) \simeq x - x^2/2$ to show that the residual deviance is approximately

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)}.$$

Solution: From (a), the residual deviance D is:

$$\begin{aligned} D &= 2\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{n_i\hat{p}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\hat{p}_i}\right) \\ &= 2\sum_{i=1}^{n} y_i \log\left(1 + \frac{y_i - n_i\hat{p}_i}{n_i\hat{p}_i}\right) + (n_i - y_i)\log\left(1 + \frac{n_i\hat{p}_i - y_i}{n_i - n_i\hat{p}_i}\right) \end{aligned}$$

Since $\hat{p}_i \rightarrow p_i$, $\frac{y_i - n_i\hat{p}_i}{n_i\hat{p}_i}$ and $\frac{n_i\hat{p}_i - y_i}{n_i - n_i\hat{p}_i}$ are small. Based on this, we can use the approximation $\log(1+x) \simeq x - x^2/2$ when $x \simeq 0$:

$$\begin{aligned} D \simeq\ & 2\sum_{i=1}^{n}\left[y_i\left(\frac{y_i - n_i\hat{p}_i}{n_i\hat{p}_i}\right) + (n_i - y_i)\left(\frac{n_i\hat{p}_i - y_i}{n_i - n_i\hat{p}_i}\right)\right] - \\ & \sum_{i=1}^{n}\left[y_i\left(\frac{y_i - n_i\hat{p}_i}{n_i\hat{p}_i}\right)^2 + (n_i - y_i)\left(\frac{n_i\hat{p}_i - y_i}{n_i - n_i\hat{p}_i}\right)^2\right] \\ =\ & 2\sum_{i=1}^{n}\frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)} - \\ & \sum_{i=1}^{n}\left[(y_i - n_i\hat{p}_i + n_i\hat{p}_i)\left(\frac{y_i - n_i\hat{p}_i}{n_i\hat{p}_i}\right)^2 + (n_i - y_i + n_i\hat{p}_i - n_i\hat{p}_i)\left(\frac{n_i\hat{p}_i - y_i}{n_i - n_i\hat{p}_i}\right)^2\right] \\ =\ & \sum_{i=1}^{n}\frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)} - \sum_{i=1}^{n}\left[\frac{(y_i - n_i\hat{p}_i)^3}{n_i^2\hat{p}_i^2} + \frac{(n_i\hat{p}_i - y_i)^3}{(n_i - n_i\hat{p}_i)^2}\right]. \end{aligned}$$

Assuming the third order terms $\frac{(y_i - n_i\hat{p}_i)^3}{n_i^2\hat{p}_i^2} + \frac{(n_i\hat{p}_i - y_i)^3}{(n_i - n_i\hat{p}_i)^2}$ are small and $\simeq 0$, the residual deviance is approximately

$$D \simeq X^2 = \sum_{i=1}^{n}\frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)}$$

(c) Show that if $n_i \equiv 1$, that the deviance is a function only of $\hat{\beta}$, and not otherwise a function of y. Show that if $n_i \equiv 1$ and $p_i \equiv p$, that $X^2 = n$. *It is worth looking at Davison's version of this question in SM p.497.*

Solution: With $n_i \equiv 1$, the score equations in (b) becomes the following:

$$\begin{cases} \frac{\partial l(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} y_i(1-p_i) - (1-y_i)p_i = 0 \\ \frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij}y_i(1-p_i) - x_{ij}(1-y_i)p_i = 0 \quad (j = 1, \cdots, p) \end{cases}$$

With some simplification:

$$\begin{cases} \sum_{i=1}^{n} p_i = \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{ij}p_i = \sum_{i=1}^{n} x_{ij}y_i \quad (j = 1, \cdots, p) \end{cases}$$

We will use these conditions for the proof.
The deviance D is:

$$\begin{aligned} D &= -2\sum_{i=1}^{n} y_i \log \hat{p}_i + (1-y_i)\log(1-\hat{p}_i) \\ &= -2\sum_{i=1}^{n} \log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) y_i + \log(1-\hat{p}_i) \\ &= -2\sum_{i=1}^{n} \hat{\beta}_0 y_i + \hat{\beta}_1 x_{i1} y_i + \cdots + \hat{\beta}_p x_{ip} y_i + \log(1-\hat{p}_i) \end{aligned}$$

Using the conditions from the score equations, this becomes:

$$D = -2\sum_{i=1}^{n} \hat{\beta}_0 \hat{p}_i + \hat{\beta}_1 x_{i1} \hat{p}_i + \cdots + \hat{\beta}_p x_{ip} \hat{p}_i + \log(1-\hat{p}_i)$$

This shows that the deviance is no longer a function of y but $\hat{\boldsymbol{\beta}}$.
When $n_i \equiv 1$ and $p_i \equiv p$, the conditions from the score equations lead to $\hat{p} = \bar{y}$.
Using this, the residual deviance is approximately:

$$\begin{aligned} X^2 &= \sum_{i=1}^{n} \frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)} \\ &= \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{\bar{y}(1-\bar{y})} \\ &= \frac{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}{\bar{y}(1-\bar{y})} \\ &= \frac{n\bar{y} - n\bar{y}^2}{\bar{y}(1-\bar{y})} \\ &= n. \end{aligned}$$

(d) Explain why the results in (c) mean that neither the residual deviance nor Pearson's X^2 can be used to assess goodness of fit with binary data.
A goodness of fit requires a comparison of observed outcome and expected outcome. However, for a binary outcome, the observed data do not enter in the equation of residual deviance or Pearson's χ^2 residuals. Therefore, both are not useful when the outcome is binary.

4. This question refers to the article "Clinical characteristics and predictors of outcomes of hospitalized patients with Coronavirus Disease 2019 in a multiethnic London National Health Service Trust: a retrospective cohort study" by Perez-Guzman et al. (2020) in *Clinical Infectious Diseases*. The article and supplementary material are posted on my web page for this class.

(a) What is a "retrospective cohort study", and why is this study described that way?
I was looking for an indication of "back in time". A cohort study records data on a group of individuals. Several definitions online and elsewhere refer to "treatment" being compared to "control", but that is not really the case here. The main point is that they assess vital status (on May 6), and related this to various covariates measured on admission.

(b) What is the "Elixhauser comorbidity score"?
This is an overall measure of health conditions; instead of trying to assess the effect of each one, an index measure like this can be used for data sets that aren't very large.

(c) What was the primary outcome of the study?
Vital status on May 6. (Some people interpreted "outcome" to mean the results of the study. But the authors do use the phrase "primary outcome" in their paper.)

(d) The methods section refers to "unadjusted" and "adjusted" logistic regression. What is the difference between these?
Unadjusted means a simple logistic regression with just one covariate. There are so many covariates, that the authors first did this to try to see which ones to keep in a multiple regression model.

(e) There are 9 models reported in Table S4 of the Supplementary Material. Which of these models ended up being summarized in Table 2 of the main paper?
I could not figure this out, so I gave everyone full marks for this. The descriptions are very confusing, and the numbers don't seem to match very well.

(f) The first 22 or 23 lines of Table 2 show one column of estimates called "unadjusted", and a second column "adjusted for age." As the authors note, "when accounting for age, only diabetes and chronic kidney disease remained statistically significant" (p.6). Why do you think this is?
I think because most of the other measures are confounded with age. Probably diabetes and CKD are also confounded with age, but their effect on COVID survival is so large that it's important even after adjusting for age.

(g) On p.7 the authors state that they performed "unadjusted and adjusted logistic regression analyses to assess the odds of death of BAME groups compared to white patients". In your own words, what did these analyses show?
Once adjusted for age and co-morbidities, death rates are higher for Black patients. From another point of view the Black patients had similar mortality to whites, but they were younger on admission.

(h) In the discussion section, the authors say "our findings .. sugest that important biological differences may be potentially driving differences in COVID-19 hospitalization outcomes by ethnicity, which cannot solely be explained by socioeconomic differences". What is the basis for this claim?
I'm not sure what the basis for this was. It might be their finding that Black patients did seem younger and with fewer co-morbidities, but had worse outcomes. In their introduction they mention that other studies that have attributed worse outcomes in BAME groups to socio-economic status did not have detailed information on their symptoms or underlying conditions when they arrived at the hospital.

(i) Summarize in your own words what you think are the two most important findings in this paper.
That Black, and to some extent Asian, patients have poorer outcomes after adjusting for co-morbidities and age seems very important. There is a huge effect of gender, which they don't discuss, perhaps because they thought it was well known. This paper had the most confusing tables ever, imho.

5. *Bonus Question* The article by Roozenbeek et al. (2020) *RSOS* was discussed in class on October 15. The authors carried out several statistical analyses, as described in Section 2.3: Pearson's correlation coefficent; one-way analysis of variance; ordingary least squares regression; two logistic regressions, and an OLS linear regression.
   (a) Choose two of these analyses and try to reproduce them using the authors' data.[2]
   (b) Try to reproduce one of the plots in the paper or the supplementary document.
   (c) In the supplementary information, Figure S1 shows residual plots from OLS regression. Why do you think there is an apparent linear boundary on the residuals vs fitted plot?
   (d) In the paper (Footnote 10 on p.8), the authors refer to a "robust standard error regression". What method did they use for this? (It might be obvious from their code what `R` package they used, but I am looking for some detail on what statistical method they used.)
   (e) The headline from this paper stated "Poor numerical literacy linked to great susceptibility to Covid-19 fake news". Do you think this headline is supported by the analyses in the paper? Why or why not?

[2] You are welcome to use their code or your own; I found it necessary to read the data using `read.csv`, for example, not `read.xls` as they recommend.