Data provided: Tables of critical values

**MAS367**

The University Of Sheffield.

**SCHOOL OF MATHEMATICS AND STATISTICS** **Autumn Semester 2021–22**

**Linear and Generalized Linear Models** **2 hours 30 minutes**

*This is an open book exam.*

*Answer all questions.*

*You can work on the exam during the 24 hour period starting from 10am (BST), and you must submit your work within 2 hours 30 minutes of accessing the exam paper or by the end of the 24 hour period (whichever is earlier).*

***Late submission will not be considered without extenuating circumstances.*** *Calculations should be performed by hand. A university-approved calculator may be used. The use of any other calculational device, software or service is not permitted. To gain full marks, you will need to show your working.*

*By uploading your solutions you declare that your submission consists entirely of your own work, that any use of sources or tools other than material provided for this module is cited and acknowledged, and that no unfair means have been used.*

*There are 40 marks available on the paper.*

**1** A fruit supplier is investigating the yields of four different breeds of its plants. The yields (in kilograms) of 128 plants were recorded, together with the soil type (clay, coded as `"C"`, or loam, coded as `"L"`), the breed of the plant (1,2,3 or 4) and the sun exposure (in hours per day) of the site.

The data are stored in a data frame `yield.data` in R, with the variables being `yield` for the yield of the plant, `soil` for the soil type, `breed` for the breed and `sun` for the sun exposure.

(a) A model was fitted in R using

```
model1 <- lm(yield~soil+factor(breed)+sun,data=yield.data)
```

(i) Describe a suitable design matrix for this model. You should illustrate your description by giving the rows corresponding to two observations: one with clay soil, breed 2, and 5.18 hours per day sun exposure, and one with loam soil, breed 4, and 5.34 hours per day sun exposure. ***(7 marks)***

(ii) Explain why you might prefer the code above to the alternative

```
model0 <- lm(yield~soil+breed+sun,data=yield.data)
```

***(3 marks)***

(iii) Some of the output from entering `summary(model1)` is given below.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.52480    0.39094  11.574  < 2e-16 ***
soilL            1.48463    0.17582   8.444 7.52e-14 ***
factor(breed)2   2.90102    0.25507  11.373  < 2e-16 ***
factor(breed)3   1.12247    0.25273   4.441 1.98e-05 ***
factor(breed)4   1.15865    0.24831   4.666 7.94e-06 ***
sun              0.61997    0.06645   9.330 5.98e-16 ***
```

Give interpretations of the numbers 1.48463, 2.90102 and 0.61997 which appear in the `Estimate` column. ***(3 marks)***

(b) A second model was fitted in R using

```
model2 <- lm(yield~soil+factor(breed)*sun,data=yield.data)
```

(i) Explain the difference between this model and `model1`. How many columns would there be in the design matrix? ***(3 marks)***

(ii) Some output from the code `anova(model1,model2)` is below.

```
Model 1: yield ~ soil + factor(breed) + sun
Model 2: yield ~ soil + factor(breed) * sun
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    122 120.356
2    119  77.705  3    42.651 21.772 2.615e-11 ***
```

What is the null hypothesis of the test being performed here? What conclusion would you make? ***(4 marks)***

**2** (a) Consider that observations $y_i$ are generated by the exponential family of distributions

$$f(y_i) = \exp\left[w_i \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right], \qquad (1)$$

where $w_i$ are the weights, $\phi$ is the scale parameter, $\theta_i$ is the natural parameter, $b(\theta_i)$ is a known function of $\theta_i$ and $c(y_i, \phi)$ is a known function of $y_i$, $\phi$. Assume that $y_i$ is independent of $y_j$, for $i \neq j$, and that the mean $\mu_i = E(y_i)$ is mapped to the linear predictor $\eta_i = x_i^T\beta$ via the link function $g(\cdot)$, so that $g(\mu_i) = \eta_i$.

(i) Based on a set of data $y = (y_1, y_2, \ldots, y_n)$, write down the log-likelihood function $\ell(\mu; y)$ of $\mu = [\mu_1, \mu_2, \ldots, \mu_n]^T$. ***(1 mark)***

(ii) Using the canonical link, calculate the partial derivatives of $\ell(\mu; y)$, with respect to $\beta_k$, for $k = 1, 2, \ldots, p$. ***(6 marks)***

(iii) Using part (ii) show that the MLE $\hat{\mu}$ of $\mu$ satisfies the matrix equality

$$X^T W y = X^T W \hat{\mu},$$

where $X$ is the design matrix and $W$ is the weight matrix, defined below

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \quad \text{and} \quad W = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix}.$$

***(3 marks)***

(b) Consider the discrete random variable $Y_i = y_i$ (taking values $1, 2, 3\ldots$) generated by the following distribution

$$f(y_i) = \Pr[Y_i = y_i] = (1 - \pi_i)^{y_i - 1}\pi_i, \quad y = 1, 2, \ldots, \quad 0 \leq \pi_i \leq 1,$$

where $i = 1, 2, \ldots, n$.

(i) Write $f(y_i)$ in exponential form (1) and hence determine $\theta_i$, $b(\theta_i)$, $c(y_i, \phi)$, $\phi$ and $w_i$. ***(2 marks)***

(ii) Use part (i) to calculate the mean $E(Y_i)$ and the variance $\mathrm{Var}(Y_i)$ of $Y_i$. ***(3 marks)***

(iii) Show that the canonical link of $f(y_i)$ is

$$g(\mu_i) = \log\frac{\mu_i - 1}{\mu_i},$$

where $\mu_i = E(Y_i)$. ***(3 marks)***

(iv) Show that the $X^2$ statistic is

$$X^2 = \sum_{i=1}^{n} \frac{(y_i\hat{\pi}_i - 1)^2}{1 - \hat{\pi}_i},$$

where $\hat{\pi}_i$ is the estimate of $\pi_i$. ***(2 marks)***

## End of Question Paper

# Tables of Percentage Points (also known as Quantiles or Critical Values) for Three Standard Distributions

The tables contain values of quantiles $q$ such that $P[X \leq q] = p$ for various probabilities $p$ when $X$ has the specified distribution (which may depend on particular degrees of freedom $\nu$). In these tables, $p$ has been expressed as a percentage rather than a decimal. The relevant $R$ commands for generating the $q$ are also shown. For the $N(0,1)$ distribution, the tabulated function is also known as the $\Phi^{-1}$ function.

STANDARD NORMAL DISTRIBUTION PERCENTAGE POINTS

`qnorm(`$p$`)` where $p$ is percentage, e.g. for 95%, $p = 0.95$

| | 60.0% | 66.7% | 75.0% | 80.0% | 87.5% | 90.0% | 95.0% | 97.5% | 99.0% | 99.5% | 99.9% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| `qnorm` | 0.253 | 0.431 | 0.674 | 0.842 | 1.150 | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 3.090 |

CHI-SQUARED PERCENTAGE POINTS

`qchisq(`$p$`,` $\nu$`)` where $p$ is percentage, e.g. for 95%, $p = 0.95$

| $\nu$ | 60.0% | 66.7% | 75.0% | 80.0% | 87.5% | 90.0% | 95.0% | 97.5% | 99.0% | 99.5% | 99.9% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.708 | 0.936 | 1.323 | 1.642 | 2.354 | 2.706 | 3.841 | 5.024 | 6.635 | 7.879 | 10.828 |
| 2 | 1.833 | 2.197 | 2.773 | 3.219 | 4.159 | 4.605 | 5.991 | 7.378 | 9.210 | 10.597 | 13.816 |
| 3 | 2.946 | 3.405 | 4.108 | 4.642 | 5.739 | 6.251 | 7.815 | 9.348 | 11.345 | 12.838 | 16.266 |
| 4 | 4.045 | 4.579 | 5.385 | 5.989 | 7.214 | 7.779 | 9.488 | 11.143 | 13.277 | 14.860 | 18.467 |
| 5 | 5.132 | 5.730 | 6.626 | 7.289 | 8.625 | 9.236 | 11.070 | 12.833 | 15.086 | 16.750 | 20.515 |
| 6 | 6.211 | 6.867 | 7.841 | 8.558 | 9.992 | 10.645 | 12.592 | 14.449 | 16.812 | 18.548 | 22.458 |
| 7 | 7.283 | 7.992 | 9.037 | 9.803 | 11.326 | 12.017 | 14.067 | 16.013 | 18.475 | 20.278 | 24.322 |
| 8 | 8.351 | 9.107 | 10.219 | 11.030 | 12.636 | 13.362 | 15.507 | 17.535 | 20.090 | 21.955 | 26.125 |
| 9 | 9.414 | 10.215 | 11.389 | 12.242 | 13.926 | 14.684 | 16.919 | 19.023 | 21.666 | 23.589 | 27.877 |
| 10 | 10.473 | 11.317 | 12.549 | 13.442 | 15.198 | 15.987 | 18.307 | 20.483 | 23.209 | 25.188 | 29.588 |

STUDENT'S $t$ PERCENTAGE POINTS

`qt(`$p$, $\nu$`)` where $p$ is percentage, e.g. for 95%, $p = 0.95$

| $\nu$ | 60.0% | 66.7% | 75.0% | 80.0% | 87.5% | 90.0% | 95.0% | 97.5% | 99.0% | 99.5% | 99.9% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.325 | 0.577 | 1.000 | 1.376 | 2.414 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 | 318.31 |
| 2 | 0.289 | 0.500 | 0.816 | 1.061 | 1.604 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 22.327 |
| 3 | 0.277 | 0.476 | 0.765 | 0.978 | 1.423 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 10.215 |
| 4 | 0.271 | 0.464 | 0.741 | 0.941 | 1.344 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173 |
| 5 | 0.267 | 0.457 | 0.727 | 0.920 | 1.301 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.893 |
| 6 | 0.265 | 0.453 | 0.718 | 0.906 | 1.273 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208 |
| 7 | 0.263 | 0.449 | 0.711 | 0.896 | 1.254 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.785 |
| 8 | 0.262 | 0.447 | 0.706 | 0.889 | 1.240 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 4.501 |
| 9 | 0.261 | 0.445 | 0.703 | 0.883 | 1.230 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.297 |
| 10 | 0.260 | 0.444 | 0.700 | 0.879 | 1.221 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.144 |
| 11 | 0.260 | 0.443 | 0.697 | 0.876 | 1.214 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.025 |
| 12 | 0.259 | 0.442 | 0.695 | 0.873 | 1.209 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.930 |
| 13 | 0.259 | 0.441 | 0.694 | 0.870 | 1.204 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.852 |
| 14 | 0.258 | 0.440 | 0.692 | 0.868 | 1.200 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.787 |
| 15 | 0.258 | 0.439 | 0.691 | 0.866 | 1.197 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.733 |
| 16 | 0.258 | 0.439 | 0.690 | 0.865 | 1.194 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.686 |
| 17 | 0.257 | 0.438 | 0.689 | 0.863 | 1.191 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.646 |
| 18 | 0.257 | 0.438 | 0.688 | 0.862 | 1.189 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.610 |
| 19 | 0.257 | 0.438 | 0.688 | 0.861 | 1.187 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.579 |
| 20 | 0.257 | 0.437 | 0.687 | 0.860 | 1.185 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.552 |
| 21 | 0.257 | 0.437 | 0.686 | 0.859 | 1.183 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.527 |
| 22 | 0.256 | 0.437 | 0.686 | 0.858 | 1.182 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.505 |
| 23 | 0.256 | 0.436 | 0.685 | 0.858 | 1.180 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.485 |
| 24 | 0.256 | 0.436 | 0.685 | 0.857 | 1.179 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.467 |
| 25 | 0.256 | 0.436 | 0.684 | 0.856 | 1.178 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.450 |
| 26 | 0.256 | 0.436 | 0.684 | 0.856 | 1.177 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.435 |
| 27 | 0.256 | 0.435 | 0.684 | 0.855 | 1.176 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.421 |
| 28 | 0.256 | 0.435 | 0.683 | 0.855 | 1.175 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.408 |
| 29 | 0.256 | 0.435 | 0.683 | 0.854 | 1.174 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.396 |
| 30 | 0.256 | 0.435 | 0.683 | 0.854 | 1.173 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.385 |
| 35 | 0.255 | 0.434 | 0.682 | 0.852 | 1.170 | 1.306 | 1.690 | 2.030 | 2.438 | 2.724 | 3.340 |
| 40 | 0.255 | 0.434 | 0.681 | 0.851 | 1.167 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 | 3.307 |
| 45 | 0.255 | 0.434 | 0.680 | 0.850 | 1.165 | 1.301 | 1.679 | 2.014 | 2.412 | 2.690 | 3.281 |
| 50 | 0.255 | 0.433 | 0.679 | 0.849 | 1.164 | 1.299 | 1.676 | 2.009 | 2.403 | 2.678 | 3.261 |
| 55 | 0.255 | 0.433 | 0.679 | 0.848 | 1.163 | 1.297 | 1.673 | 2.004 | 2.396 | 2.668 | 3.245 |
| 60 | 0.254 | 0.433 | 0.679 | 0.848 | 1.162 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 | 3.232 |
| $\infty$ | 0.253 | 0.431 | 0.674 | 0.842 | 1.150 | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 3.090 |