

UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in: STK3100/4100 — Introduction to Generalized Linear Models

Day of examination: Friday December 19th, 2025

Examination hours: 09.00–13.00

This problem set consists of 6 pages.

Appendices: Formulas in STK3100/4100

Permitted aids: Approved calculator

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Problem 1

We assume that the random variable Y follows a gamma distribution with parameters μ, k , written as $Y \sim \text{Gamma}(\mu, k)$, and thus has the probability density function (pdf)

$$f(y; \mu, k) = \frac{1}{\Gamma(k)} \left(\frac{k}{\mu}\right)^k y^{k-1} e^{-\frac{ky}{\mu}}, \quad y > 0. \quad (1)$$

a

Show that the distribution of Y is in the exponential dispersion family, so that its pdf can be written on the form

$$f(y; \theta, \phi) = \exp \left(\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi) \right),$$

with $\theta = -\frac{1}{\mu}$, $\phi = \frac{1}{k}$, $b(\theta) = -\log(-\theta)$ and $a(\phi) = \phi$, and determine $c(y, \phi)$.

b

Use the expressions for $b(\theta)$ and $a(\phi)$ to show that $E(Y) = \mu$ and $\text{Var}(Y) = \phi\mu^2$.

Assume in the rest of the problem that the random variables $Y_1, \dots, Y_n$ are independent with $Y_i \sim \text{Gamma}(\mu_i, k)$, $i = 1, \dots, n$, $k = 1/\phi$. Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ be the vectors of covariate values, with $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$, $i = 1, \dots, n$. Now, consider a generalised linear model (GLM) with linear predictor $\eta_i = \mathbf{x}_i \boldsymbol{\beta} = \sum_{j=1}^p \beta_j x_{ij}$ and link function $g(\mu_i) = \eta_i$.

(Continued on page 2.)

c

- Explain what a canonical link function is, and find the canonical link function for this particular model.
- The most common link function in a gamma GLM is the log-link $\eta_i = \log(\mu_i)$. Show that the likelihood equations for the gamma GLM with the log-link function can be written as

$$\sum_{i=1}^n \frac{y_i - \mu_i}{\mu_i} x_{ij} = 0, \quad j = 1, \dots, p.$$

The Pearson statistic of a GLM is in general given by $X^2 = \sum_{i=1}^n \frac{(Y_i - \hat{\mu}_i)^2}{\nu(\hat{\mu}_i)}$, where $\nu(\mu_i) = \text{Var}(Y_i)$ is the variance function of the GLM.

d

- Show that the Pearson statistic of the gamma GLM is given by

$$X^2 = \frac{1}{\phi} \sum_{i=1}^n \left(\frac{Y_i - \hat{\mu}_i}{\hat{\mu}_i} \right)^2.$$

- Argue that ϕ can be estimated by $\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^n \frac{(Y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^2}$ (*Hint: Recall that X^2 is approximately χ_{n-p}^2 -distributed.*)

Problem 2

The cost of car insurance claims can be affected by several factors, such as the age and gender of the driver and the value of the vehicle. In this problem, we will consider a car insurance data set from one-year vehicle insurance policies taken out in 2004 or 2005, that contains $n = 4624$ records of the following explanatory variables:

- **log_nclaims**: Number of claims from the policy holder during the year
- **veh_value**: Vehicle value in 10,000 USD
- **gender**: Gender of the individual (binary; 1: male, 0: female)
- **agecat**: Age category of the policy holder (categorical; levels: 1, 2, 3, 4, 5 and 6, where higher level means higher age)

The response variable is the cost **claimcst** of all insurance claims from a given policy holder during the year. The data set contains only claim costs when at least one claim was made, so the claim cost is always > 0 .

a

Output from fitting a GLM in R with all the main effects described above, **log_nclaims** being an offset, is given below.

- Describe the model that is used, including all necessary assumptions.
- What is an offset?
- Explain why the offset in this case leads to a model for the expected claim cost per claim., i.e. $g\left(E\left(\frac{Y_i}{n_{\text{claims}_i}}\right)\right) = \mathbf{x}_i\boldsymbol{\beta}$, where g is the link function. Why this is sensible?

(Continued on page 3.)

Call:

```
glm(formula = claimcst ~ offset(log_nclaims) + veh_value + factor(gender)
    + factor(agecat), family = Gamma(link = log), data = car_insurance)
```

```
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.75434      0.09125  84.981 < 2e-16 ***
veh_value     -0.01021      0.02267  -0.450 0.652396
factor(gender)1 0.17921      0.05307   3.377 0.000739 ***
factor(agecat)2 -0.20356      0.09833  -2.070 0.038501 *
factor(agecat)3 -0.29975      0.09556  -3.137 0.001719 **
factor(agecat)4 -0.29358      0.09560  -3.071 0.002147 **
factor(agecat)5 -0.40022      0.10683  -3.746 0.000182 ***
factor(agecat)6 -0.35285      0.12207  -2.890 0.003865 **
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 3.127052)

Null deviance: 7338.6 on 4623 degrees of freedom

Residual deviance: 7241.9 on 4616 degrees of freedom

AIC: 79334

b

Denote $\tilde{\mu}(\mathbf{x})$ and $\tilde{\mu}(\mathbf{x}^*)$ the expected claim costs per claim for policy holders with covariate vector $\mathbf{x}$ and $\mathbf{x}^*$, respectively.

- Show that the claim cost per claim for a policy holder with covariate vector $\mathbf{x}^*$, relative to a policy holder with covariate vector $\mathbf{x}$ is given by $\frac{\tilde{\mu}(\mathbf{x}^*)}{\tilde{\mu}(\mathbf{x})} = e^{(\mathbf{x}^* - \mathbf{x})\boldsymbol{\beta}}$.

- Use the above result to give an interpretation of $e^{\hat{\beta}_2}$ and $e^{\hat{\beta}_3}$, where $\hat{\beta}_2$ and $\hat{\beta}_3$ are the estimated β -coefficients for `veh_value` and `gender`, respectively.

c

- Explain why $e^{\hat{\beta}_j \pm 1.96 se_j}$, where se_j is the standard error of $\hat{\beta}_j$, is an approximate 95% confidence interval for $e^{\hat{\beta}_j}$.

- Calculate this confidence interval for $e^{\hat{\beta}_2}$ and $e^{\hat{\beta}_3}$.

- Use these confidence intervals to determine conclusions to hypothesis tests for $H_{0j} : \beta_j = 0$ versus $H_{Aj} : \beta_j \neq 0$, $j = 2, 3$, with a 5% significance level, and compare with the results from the Wald tests in the R output.

d

As the covariates `veh_value` and `agecat` seem to have the least significant effect on the claim cost per claim (see the R output below), we try to refit the model, first without `veh_value`, and then without `agecat`. Some of the entries have been replaced by question marks.

- Give the missing numbers, explaining how you found them.

- Which of the models do you think fits the data the best, and why?

(Continued on page 4.)

```
drop1(fit.mod1,test="LRT")
```

```
Model:  claimcst ~ offset(log_nclaims)+veh_value+factor(gender)+ factor(agecat)
Df Deviance   AIC scaled dev.  Pr(>Chi)
<none>                7241.9 79334
veh_value      1    7242.6 79332      0.2305 0.6311243
factor(gender) 1    7277.5 79343      11.3795 0.0007426 ***
factor(agecat) 5    7298.6 79342      18.1249 0.0027938 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Analysis of Deviance Table

```
Model 1: claimcst ~ offset(log_nclaims)+factor(gender)
Model 2: claimcst ~ offset(log_nclaims)+factor(gender)+factor(agecat)
Model 3: claimcst ~ offset(log_nclaims)+veh_value+factor(gender)+factor(agecat)
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      4622      7298.7
2      4617        ?      ?   56.021 0.003055 **
3        ?      7241.9      1    ?    0.631124
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Problem 3

We will now consider a study on how the distance from the pituitary to the pterygomaxillary fissure depends on age and gender. The study followed $n = 27$ children (16 males, 11 females) from age 8 till 14, making $d = 4$ measurements of the response **distance** on each child, resulting in 108 measurements in total, including the two explanatory variables

- **age**: Age in years
- **gender**: Gender of the individual (binary; 1: male, 0: female)

For these data, we assume a normal linear mixed model (LMM) with identity link and random intercept. Let Y_{ij} denote the response for subject j in cluster i , $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2})$ be vectors of explanatory variables and $u_i \sim N(0, \sigma_u^2)$, the random intercept, for $i = 1, \dots, 27$, $j = 1, \dots, 4$. The LMM is then given by

$$Y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + u_i + \varepsilon_{ij},$$

with $(\varepsilon_{i1}, \dots, \varepsilon_{i4})^T \sim N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I})$, so that

$$E(Y_{ij}|u_i) = \mathbf{x}_{ij}\boldsymbol{\beta} + u_i.$$

Further, it is assumed that u_i is independent of ε_{ij} . Output from fitting this model to the data is shown below.

(Continued on page 5.)

Linear mixed-effects model fit by maximum likelihood

Data: orthodont

AIC BIC logLik
444.8565 458.2671 -217.4282

Random effects:

Formula: ~1 | subject

(Intercept) ResidualStdDev:

1.730079 1.422728

Fixed effects: distance ~ age + factor(gender)

	Value	Std.Error	DF	t-value	p-value
(Intercept)	15.385690	0.8909087	80	17.269660	0.0000
age	0.660185	0.0620929	80	10.632212	0.0000
factor(gender)1	2.321023	0.7430668	25	3.123572	0.0045

Correlation: (Intr) age

age -0.767

factor(gender)1 -0.494 0.000

Standardized Within-Group Residuals:

Min	Q1	Med	Q3	Max
-3.77682007	-0.55426744	-0.01578248	0.45835495	3.68124620

Number of Observations: 108

Number of Groups: 27

a

- Derive the marginal model of the LMM, i.e. $\mu_{ij} = E(Y_{ij})$.

(Hint: You may use the law of double expectation, i.e. for two random variables V and W , $E(V) = E(E(V|W))$).

- Comment on the link function for this marginal model and on the effects of the explanatory variables compared to the fixed effects of the LMM.

b

- Show that the correlation between two measurements on the same child in the LMM, $\text{Corr}(Y_{ij}, Y_{ik})$, $j \neq k$, is given by $\frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2}$.

- Compute this correlation using $\sigma_u = 1.73$ and $\sigma_\varepsilon = 1.42$ from the R output.

- Results from fitting a normal linear model assuming that observations are independent, also the ones made on the same child, are shown below. Comparing the outputs from fitting the LMM and the normal linear model, what is the main consequence of assuming that measurements made on the same individual are independent, and how does this relate to the results from **a**?

(Continued on page 6.)

Call:

```
lm(formula = distance ~ age + factor(gender), data = orthodont)
```

```
Residuals:  Min      1Q  Median      3Q      Max
-5.9882 -1.4882 -0.0586  1.1916  5.3711
```

Coefficients:

```
Estimate Std. Error t value Pr(>|t|)
(Intercept)    15.38569    1.12857   13.633 < 2e-16 ***
age              0.66019    0.09776    6.753 8.25e-10 ***
factor(gender)1  2.32102    0.44489    5.217 9.20e-07 ***
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 2.272 on 105 degrees of freedom

Multiple R-squared: 0.4095, Adjusted R-squared: 0.3983

F-statistic: 36.41 on 2 and 105 DF, p-value: 9.726e-13

END