Statistical Modeling 1
Problem set 4 2024

1. Consider the following air pollution study with the dataset ozone.txt:

```
> data<-read.table("ozone.txt", header=TRUE, sep="\t", dec=".")
> data
   rad temp wind ozone
1  190   67  7.4    41
2  118   72  8.0    36
3  149   74 12.6    12
4  313   62 11.5    18
5  299   65  8.6    23
6   99   59 13.8    19
7   19   61 20.1     8
8  256   69  9.7    16
9  290   66  9.2    11
10 274   68 10.9    14
.
.
The dataset is gathered during the air pollution study.
The response variable is ozone. The problem is to find out,
how is ozone concentration related to wind speed, air temperature
 and intensity of solar radiation.
```

Denote the variables as following

$$Y = \text{ozone}, \quad X_1 = \text{rad}, \quad X_2 = \text{temp}, \quad X_3 = \text{wind}.$$

Note that the response variable Y = ozone is continuous random variable where measurement accuracy happens to be in integer level.

(a) Let us assume $Y_i \sim N(\mu_i, \sigma^2)$. Consider the models

$$\begin{aligned}
\mathcal{M}_{\text{identity}}: \quad \mu_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}, \\
\mathcal{M}_{\text{log}}: \quad \log(\mu_i) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}, \\
\mathcal{M}_{\text{inverse}}: \quad \frac{1}{\mu_i} &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}, \\
\mathcal{M}_{\text{exponential}}: \quad \log(\mu_i) &= \beta_0 + \beta_1 \log(x_{i1}) + \beta_2 \log(x_{i2}) + \beta_3 \log(x_{i3}).
\end{aligned}$$

Which model fits the best to the data if the choice of model is done based on the AIC value?

i. $\mathcal{M}_{\text{identity}}$,

ii. $\mathcal{M}_{\text{inverse}}$,

iii. $\mathcal{M}_{\text{log}}$,

iv. $\mathcal{M}_{\text{exponential}}$.

(1 point)

(b) Choose the model based on your solution to (a). Study under different distributional assumptions how Pearson's residuals are behaving. That is, consider the following linear models for Pearson residuals o_i

$$o_i^2 = \alpha_0 + \alpha_1 \hat{\mu}_i + \varepsilon_i$$

in case of normal, Gamma, and Inverse Gaussian distribution. For each distribution, test the null hypothesis $H_0 : \alpha_1 = 0$. Based on these Pearson's residuals testing results, which distributional assumption is the most suitable one?

i. $Y_i \sim N(\mu_i, \sigma^2)$,
ii. $Y_i \sim Gamma(\mu_i, \phi)$,
iii. $Y_i \sim IG(\mu_i, \phi)$.

(2 points)

(c) Regardless of your solutions to (a) and (b), let us assume $Y_i \sim Gamma(\mu_i, \phi)$. Test at 5% significance level, is the explanatory variable $X_3 =$ wind statistically significant variable in the model

$$\mathcal{M} : \quad \log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

Select appropriate test statistic to test the significance of the variable $X_3 =$ wind. Calculate the value of the test statistic, and return it as your answer to the question.

(1 point)

(d) Regardless of your solutions to (a) and (b), let us assume $Y_i \sim Gamma(\mu_i, \phi)$ and

$$\mathcal{M} : \quad \log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

Test the hypotheses

$$H_0 : \beta_2 + \beta_3 = 0,$$
$$H_1 : \beta_2 + \beta_3 \neq 0.$$

Select appropriate test statistic to test the above hypotheses. Calculate the value of the test statistic, and return it as your answer to the question.

(2 points)

2. Consider the data set in the file denim.txt:

```
   Laundry Denim Abrasion
1        0     1   3.2218
2        0     1   3.3547
3        0     1   3.1334
4        0     1   2.6289
5        0     1   3.8816
.
.
89      25     3   2.1734
90      25     3   2.9636


Effects of Laundering Cycles and denim treatment on edge abrasion of denim jeans

Laundry - laundry cycles
Denim -Three types of denim treatments (1 = pre-washed, 2 = stone-washed, 3 = enzyme washed)
Abrasion - abrasion score (lower score means higher damage)

Card, A., Moore, M.A. and Ankeny, M. (2006) Garment washed jeans:
Impact of launderings on physical properties.
Int. J. Clothing Sc. Tech., 18, pp.43-52.
```

Denote variables as following

$$Y = \textsf{Abrasion},\ X_1 = \textsf{Laundry},\ X_2 = \textsf{Denim}.$$

(a) Let us assume $Y_i \sim N(\mu_i, \sigma^2)$. Consider the models

$$\begin{aligned}
\mathcal{M}_{1|2_{\textsf{identity}}} :&\quad \mu_i = \beta_0 + \beta_1 x_{i1} + \alpha_j,\\
\mathcal{M}_{12_{\textsf{identity}}} :&\quad \mu_i = \beta_0 + \beta_1 x_{i1} + \alpha_j + \gamma_j x_{i1},\\
\mathcal{M}_{1|2_{\log}} :&\quad \log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \alpha_j,\\
\mathcal{M}_{12_{\log}} :&\quad \log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \alpha_j + \gamma_j x_{i1},\\
\mathcal{M}_{1|2_{\textsf{inverse}}} :&\quad \frac{1}{\mu_i} = \beta_0 + \beta_1 x_{i1} + \alpha_j,\\
\mathcal{M}_{12_{\textsf{inverse}}} :&\quad \frac{1}{\mu_i} = \beta_0 + \beta_1 x_{i1} + \alpha_j + \gamma_j x_{i1}.
\end{aligned}$$

Which model you feel is fitting the best to the data?

i. $\mathcal{M}_{1|2_{\textsf{identity}}}$,
ii. $\mathcal{M}_{12_{\textsf{identity}}}$,
iii. $\mathcal{M}_{1|2_{\log}}$,
iv. $\mathcal{M}_{12_{\log}}$,
v. $\mathcal{M}_{1|2_{\textsf{inverse}}}$,
vi. $\mathcal{M}_{12_{\textsf{inverse}}}$.

(2 points)

(b) Choose the model based on your solution to (a). Which distributional assumption you feel is the most suitable one?

i. $Y_i \sim N(\mu_i, \sigma^2)$,
ii. $Y_i \sim Gamma(\mu_i, \phi)$,
iii. $Y_i \sim IG(\mu_i, \phi)$.

(1 point)

(c) Choose the model and the distributional assumption based on your solution to (a) and (b). After you have chosen your model, calculate the d-value for the predictive effect size difference $Y_{2f} - Y_{1f}$ when explanatory variables are changed from the values

$$x_{1f1} = 0, \qquad x_{1f2} = 1 = \textsf{pre-washed}$$

to the values

$$x_{2f1} = 25, \qquad x_{2f2} = 3 = \textsf{enzyme washed}.$$

(2 points)

(d) Choose the model and the distributional assumption based on your solution to (a) and (b). Test at 5% significance level, is the explanatory variable $X_2 = \textsf{Denim}$ statistically significant variable. Select appropriate test statistic to test the significance of $X_2 = \textsf{Denim}$. Calculate the value of the test statistic, and return it as your answer to the question.

(1 point)

3. (a) Let us assume $Y_i \sim IG(\mu_i, \phi)$. Consider the model

$$\log(\mu_i) = \beta_0 + \beta_1 \log(x_i).$$

Let the estimates of the parameters β_0, β_1, ϕ be as $\hat{\beta}_0 = 1, \hat{\beta}_1 = 0.5, \tilde{\phi} = 0.05$, when

$$\mathbf{X} = \begin{pmatrix} 1 & \log(3) \\ 1 & \log(3) \\ 1 & \log(3) \\ 1 & \log(6) \\ 1 & \log(6) \\ 1 & \log(6) \\ 1 & \log(9) \\ 1 & \log(9) \\ 1 & \log(9) \end{pmatrix}.$$

Calculate the estimated covariance matrix $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X})^{-1}$.

(2 points)

(b) Let $Y_i \sim Poi(\mu_i)$. Then the probability density function of the random variable Y_i is

$$f(y_i|\mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}.$$

Show first that Y_i belongs to the exponential family of distributions, and then show that

$$\mathrm{E}(Y_i) = \mu_i,$$
$$\mathrm{Var}(Y_i) = \mu_i.$$

Hint! There is no dispersion parameter ϕ in Poisson distribution and hence you may consider it as $\phi = 1$.

(2 points)

(c) Consider the simple Gamma model with

$$Y_i \sim Gamma(\mu_i, \phi),$$
$$\mu_i = \eta_i = \beta_0.$$

Construct the $100(1-\alpha)\%$ prediction interval for the new observation Y_f.

(2 points)