COLUMBIA UNIVERSITY
Statistical Inference and Modeling
Quiz 5

Prof. Marco Avella
TA: Ian Kinsella, Shun Xu
April 18, 2019

Name:

UNI:

You have 20 minutes to answer the following 10 questions. Good luck!

Question 1

A generalized linear model fit in R gave the following output.

```
Call:
glm(formula = Claim ~ ., family = Gamma(link = log), data = IndustryAuto)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.39708  -0.16712   0.07158   0.13790   0.21437

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -68.21431   24.78548  -2.752  0.00813 **
Incurral.Year      0.03923    0.01239   3.165  0.00259 **
Development.Year   0.10228    0.01239   8.253 5.08e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.03801091)

    Null deviance: 4.5020  on 54  degrees of freedom
Residual deviance: 2.2085  on 52  degrees of freedom
AIC: 1148.6

Number of Fisher Scoring iterations: 6
```

A generalized linear model fit in `R` gave the above output. Which of the following statements is NOT correct?

(a) The estimated dispersion parameter suggests that assuming a gamma distribution for `Claims` is more appropriate than assuming a exponential distribution.

(b) The estimated Fisher information corresponding to the slope parameter of `Incurral.Year` is $1/0.01239^2$.

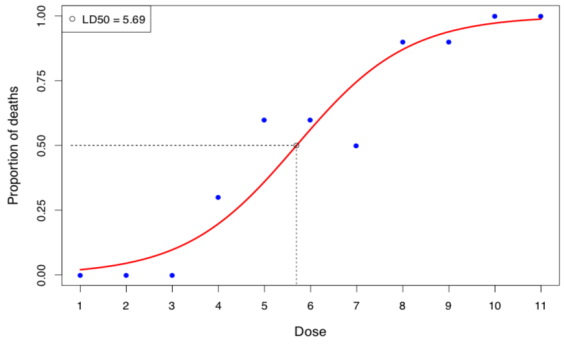
(c) The fitted mean model is roughly

$$\mathbb{E}[\texttt{Claim}_i] = \exp(-68.2 + 0.04\texttt{Incurral}_i + 0.10\texttt{Development}_i).$$

(d) The model postulates that the deviance residuals should be symmetric.

Question 2

Figure 1: Logistic regression fit of dose-response study. Groups of 10 mice were exposed to increasing doses of experimental drug. *The points are the observed proportions that died in each group.* The fitted curve is the maximum-likelihood estimate of the logistic regression model.



Based on Figure 1, which of the following statements is NOT correct?

(a) The open circle on the curve is the estimated dose for 50% mortality.

(b) The fitted model postulates that $\mathbb{E}[Y_i] = e^{\alpha+\beta D_i}/(1+e^{\alpha+\beta D_i})$, where Y_i is a binary variable indicating whether mouse i survived and D_i the dose of the drug applied to it.

(c) We cannot figure out how many mice died in this experiment.

(d) The data suggests that doses larger than 3 units can be lethal.

Question 3

Which of the following statements is NOT correct?

(a) Local polynomial estimators can estimate both unknown smooth functions f and their derivatives simultaneously.

(b) Local polynomial estimators can be computed using a weighted least squares algorithm.

(c) Local linear estimators mitigate boundary effects relative to local constant estimators.

(d) Local polynomial estimators are computationally cheaper to tune than spline estimators.

Question 4

Consider the regression model

$$Y_i = f(i/n) + \varepsilon_i,\ i = 1, \ldots, n \text{ where } \varepsilon \overset{iid}{\sim} N(0, \sigma^2)$$

and a local linear estimator of the form $\hat{\mathbf{f}} = \mathbf{S}_h \mathbf{y}$, where h is the bandwidth. Which of the following statements is correct?

(a) The bias of $\hat{f}(x)$ increases when $h \to 0$.

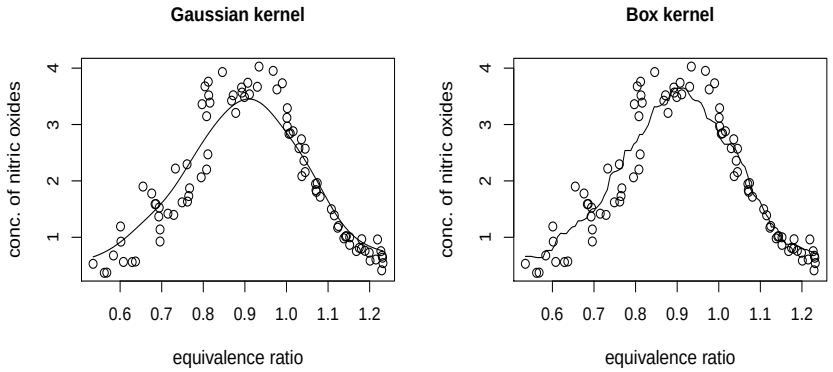
(b) $\hat{\mathbf{f}} \sim N(\mathbf{S}_h \mathbf{f}, \sigma^2 \mathbf{S}_h \mathbf{S}_h^T)$

(c) The variance of $\hat{f}(x)$ is larger for small values of h.

(d) $\hat{\mathbf{f}}$ is a Nadaraya-Watson estimator.

Question 5

Figure 2: Two Nadaraya-Watson estimates for the Ethatnol data.



How do you explain the differences of the two fits in Figure 2? Are the differences likely to be explained by the use of a different bandwidth?

- The uniform kernel induces some clear discontinuities every time a new observation drops or enters the local averaging window. The transitions are smoother for the Gaussian kernel
- Both fits are obtained with similar bandwidths and similar pictures will be obtained for different bandwidths

Question 6

```
Call:
glm(formula = Class ~ Bare.nuclei, family = binomial(link = probit),
    na.action = na.omit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2302  -0.3652  -0.3652   0.1043   2.3413

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.96960    0.10934  -18.01   <2e-16 ***
Bare.nuclei  0.45172    0.03177   14.22   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 884.35  on 682  degrees of freedom
Residual deviance: 342.94  on 681  degrees of freedom
  (16 observations deleted due to missingness)
AIC: 346.94

Number of Fisher Scoring iterations: 7
```

A generalized linear model fit in R gave the above output. Which of the following statements is correct?

1. Was the canonical link function used to fit this model?
2. Give the null hypothesis used to compute the p-value associated with the covariate `Bare.nuclei`.
3. Give a confidence interval for the estimated slope parameter of `Bare.nuclei`.
4. Write the log-likelihood of the postulated model
5. Give a formula for the estimated variance of $Y_i|X_i = x_i$

1) Probit link is not canonical, $\mathbb{E}[Y_i|x_i] = \Phi(x_i^T\beta) = p_i$

2) $H_0: \beta_{\text{Bare.nuclei}} = 0$

3) $CI_{1-\alpha}(\hat{\beta}_{\text{Bare.nuclei}}) = (0{,}45 \pm z_{1-\alpha/2}\ 0{,}03)$

4) $\sum_{i=1}^{n} \log p_i^{y_i}(1-p_i)^{1-y_i} = \sum_{i=1}^{n}\{y_i \log \Phi(x_i^T\beta) + (1-y_i)\log(1-\Phi(x_i^T\beta))\}$

5) $p_i(1-p_i) = \Phi(x_i^T\beta)(1-\Phi(x_i^T\beta))$