COLUMBIA UNIVERSITY
Statistical Inference and Modeling
Quiz 5

Prof. Marco Avella
TA: Ian Kinsella
November 26, 2019

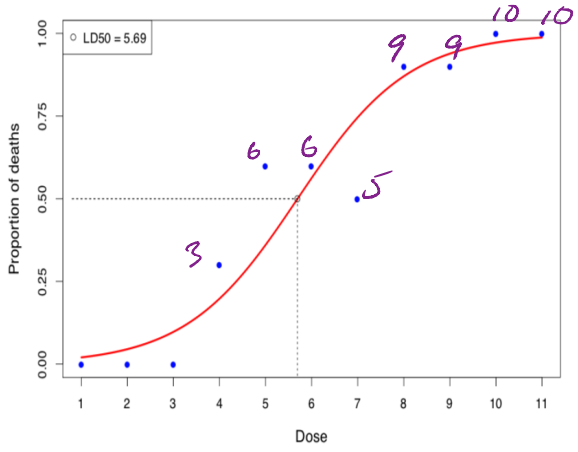
Name:

UNI:

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

Question 1 (4 points)

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.



Answer the following questions based on Figure 1.

1. Give the estimated dose for %50 mortality rate. $\approx 5,8$
2. How many mice died in this experiment? 48
3. Let Y_i be a binary variable indicating whether mouse i survived and D_i the dose of the drug applied to it. Write down the likelihood of the model.
4. Give a formula for the mean and variance $Y_i|D_i$

4) $\mathbb{E}[Y_i \mid D_i] = \dfrac{e^{\beta_0+\beta_1 D_i}}{1+e^{\beta_0+\beta_1 D_i}} = p_i$, $\mathrm{Var}[Y_i \mid D_i] = p_i(1-p_i) = \dfrac{e^{\beta_0+\beta_1 D_i}}{(1+e^{\beta_0+\beta_1 D_i})^2}$

3) $L_n(\beta_0,\beta_1) = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{Y_i}$

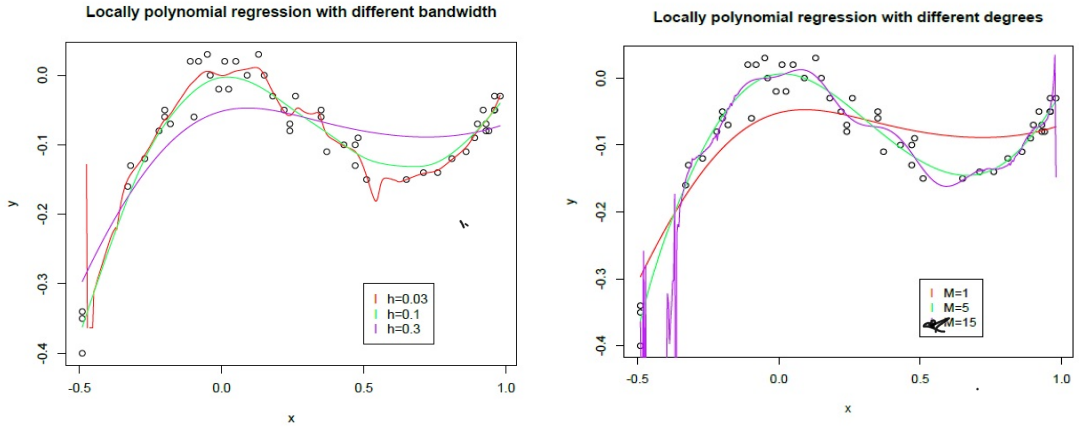
Question 2 (3 points)

In this problem we would like to assess the performance of the locally polynomial regression in terms of M (degree of the polynomial) and the bandwidth h. We downloaded a dataset fitted a curve to the data using the locally polynomial regression in the following two cases:

1. We set the degree of the polynomial to 1 and fit a curve using bandwidth $= 0.03, 0.1, 0.3$.

2. We set the bandwidth to 0.3 and considered three different degree for the polynomial 1, 5, 15.

Answer the following questions based on the two plots below: which curve gives the best fitt? Which one suffers from the highest variance? Which one suffers from the highest bias?



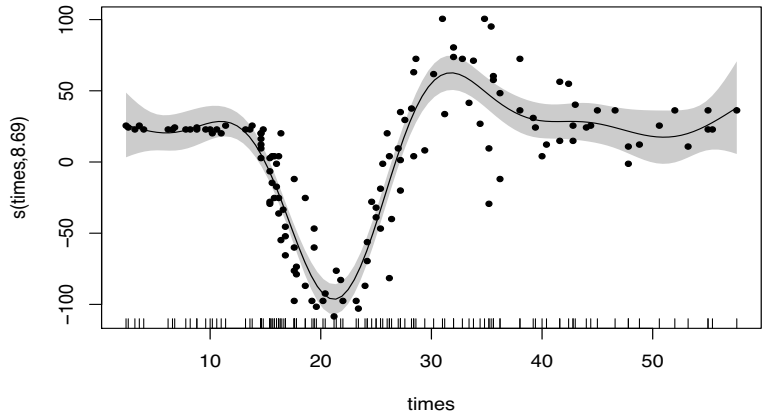
Best fit : local polynomial of degree 5 as the fit is smooth as has a small bias

Highest variance : local polynomial of degree 15, in particular it exhibits large variance close to the boundaries

Highest bias : the local polynomial with degree 1 and $h=0,03$ is clearly oversmoothing (underfitting) the data

Question 3 (3 points)

We fiited a smooth curve to the motocycle data with the `mgcv` R package and obtained the following output:




```
summary(fit0)
fit0=gam(accel~s(times),data=mcycle)
plot(fit0,select=1,residuals=T)

Family: gaussian
Link function: identity

Formula:
accel ~ s(times)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -25.546      1.951   -13.1   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
           edf Ref.df     F p-value
s(times) 8.693  8.972 53.52  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.783   Deviance explained = 79.8%
GCV = 545.78  Scale est. = 506       n = 133
```

(zero mean function
i.e. $\int f_0(x)dx = 0$

Answer in 2-3 sentences the following questions:

1. Give some intuition for the varying widths of the shadowed areas around the fitted curve.
2. Explain the meaning of the intercept reported above
3. What can you say about the complexity of the fitted curve?

1) Narrower shadowed areas reflect both smaller variability of the response and more observations in the "time" neighborhood

2) Denoting the fitted curve by $\hat{f}(x)$, the intercept reported is also $\int \hat{f}(x)dx = \hat{\beta}_0 = -25,546$

3) The equivalent degrees of freedom reported indicates that the complexity of the fitted curve is roughly that of a polynomial of order 8-9.