2018 First Year Exam – Methods
Statistics
210-211-212
June 25, 2018
9:00 – 12:00

Instructions

- There are 4 questions on the examination, each with multiple parts. Select any 3 of them to solve.
- Your solutions to each of the 3 problems you solve should be written on separate sheets of paper. Label *each sheet* with your student id number, the problem number, and the page for that problem written in the upper right hand corner. For example, the labeling on a page might be:

  ID# 912346378
  Problem 2, page 3
- You have 3 hours to complete your solution. Please be prepared to turn in your exam at 12:00 noon.

METHODS 210-211-212 (2018), Problem 1

[Note: Students may leave any numerical computations unevaluated in expression form.]

Percentage yields from a chemical reaction for changing temperature (x_1) and agitation speed (x_2) are as follows

Average Yields (%)		x_2: Agitation Speed		Marginal mean
		Fast (1)	Slow (-1)	
x_1: Temperature	High (1)	$\bar{y}_{HF}$ = 80	$\bar{y}_{HS}$ = 74	$\bar{y}_{H\cdot}$ = 77
	Low (-1)	$\bar{y}_{LF}$ = 78	$\bar{y}_{LS}$ = 70	$\bar{y}_{L\cdot}$ = 74
Marginal mean		$\bar{y}_{\cdot F}$ = 79	$\bar{y}_{\cdot S}$ = 72	$\bar{y}_{\cdot\cdot}$ = 75.5

The factors are defined as

$$x_1 = 1\ if\ temperature\ is\ high\ and\ -1\ if\ low$$

$$x_2 = 1\ if\ agitation\ speed\ is\ fast\ and\ -1\ if\ slow$$

Each listed yield is actually the average of five (5) individual independent experiments. The variance of individual measurements can be estimated from the five replications in each cell. It is found that

$$s^2 = \frac{\sum_{i=L}^{H}\sum_{j=S}^{F}\sum_{k=1}^{5}(y_{ijk} - \bar{y}_{ij})^2}{4(5-1)} = 12.5$$

a) The 20 data points are to be fitted with the following multiple linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where the error terms are iid N(0, σ^2) with unknown σ^2. Write down the model matrix **X** associated with the data set.

b) Find the least squares estimators for the regression coefficients and express them as functions of average yields $\bar{y}_{ij}{}'s$. Calculate the estimate values based on the data table provided above.

c) Complete the following ANOVA table for the no-interaction model in a):

Source	d.f.	SS	MS
x_1: Temperature			
x_2: Agitation Speed			
Residual			
Total (corrected)			---

d) Based on the ANOVA table, perform hypothesis testing at 5% significance level individually for each of the 2 hypotheses below. Define the critical values and state the decision rules for each testing clearly.
   i. $H_0: \beta_1 = \beta_2 = 0$ vs. H_a: either $\beta_1 \neq 0$ or $\beta_2 \neq 0$
   ii. $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$

e) Provide a 95% 2-sided confidence interval for the difference of mean yield percentages between fast agitation speed and slow agitation speed. (Write in the needed distributional percentile in notation form with a clear definition.)

f) What is the 95% 2-sided prediction interval for a new observation at the normal temperature (i.e., $x_1 = 0$) and with the average agitation speed (i.e., $x_2 = 0$)? (Write in the needed distributional percentile in notation form with a clear definition.)

g) A criticism was made toward the above model and analyses because a potential interaction between x_1 and x_2 was neglected. How would you respond to the criticism?

METHODS 210-211-212 (2018), Problem 2

[Note: Students may leave any numerical computations unevaluated in expression form.]

Data of response y_i and a continuous covariate x_3 are collected for three groups A, B, and C. A one-way ANCOVA model without interaction terms is to be fitted as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

where the error terms are independently distributed as N(0, σ^2) with unknown σ^2. The groups are coded as

$$x_{i1} = \begin{cases} 1, if\ group = A \\ 0, otherwise \end{cases}$$

$$x_{i2} = \begin{cases} 1, if\ group = B \\ 0, otherwise \end{cases}$$

Sample sizes for the three groups are denoted as $n_A,\ n_B,\ and\ n_C$, respectively.

a) Write down the vector **y** and the model matrix **X**, in which x_{i3} is to be represented by its notation and x_{i1} and x_{i2} by their values.

b) Interpret the meaning of each regression coefficient in the model.

Group	$\bar{y}$	$\bar{x}_3$	n	S_{33}	S_{3y}
A	$\bar{y}_A = 5$	$\bar{x}_{3A} = 8$	$n_A = 15$	$\sum_{i=1}^{n_A}(x_{i3} - \bar{x}_{3A})^2 = 10$	$\sum_{i=1}^{n_A}(x_{i3} - \bar{x}_{3A})(y_i - \bar{y}_A) = 6$
B	$\bar{y}_B = 9$	$\bar{x}_{3B} = 6$	$n_B = 16$	$\sum_{i=n_A+1}^{n_A+n_B}(x_{i3} - \bar{x}_{3B})^2 = 12$	$\sum_{i=n_A+1}^{n_A+n_B}(x_{i3} - \bar{x}_{3B})(y_i - \bar{y}_B) = 5$
C	$\bar{y}_C = 12$	$\bar{x}_{3C} = 4$	$n_C = 14$	$\sum_{i=n_A+n_B+1}^{n_A+n_B+n_C}(x_{i3} - \bar{x}_{3C})^2 = 14$	$\sum_{i=n_A+n_B+1}^{n_A+n_B+n_C}(x_{i3} - \bar{x}_{3C})(y_i - \bar{y}_C) = 7$
Notations: For group A, $\bar{y}_A$ is the group sample mean of response y and $\bar{x}_{3A}$ is the group sample mean of the covariate x_3. For groups B and C, $\bar{y}_B$ and $\bar{y}_C$ as well as $\bar{x}_{3B}$ and $\bar{x}_{3C}$ are similarly defined.					

c) Using the statistics in the above table, calculate the LS estimates for the four regression coefficients.

Note: The remainder of the question is from an alternate set of data. The data set is exactly the same size (n_A = 15, n_B = 16, n_C = 14) and the same model was applied to the data. A summary of the regression output for the data and selected summary statistics are provided below. (The table includes an additional column relative to the table above. To make the column fit the summation limits have been deleted. They are the same as in the table above.) Use these data for parts (d), (e), (f), (g).

```
Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
     Min       1Q   Median       3Q      Max
-11.9077  -2.3848   0.4552   2.1183  11.6258

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.6879     2.8160  -0.599    0.552
X1           24.7967     1.9289  12.856 5.69e-16 ***
X2            9.8757     1.7505   5.642 1.40e-06 ***
X3            1.8173     0.1787  10.168 8.99e-13 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Group	$\bar{y}$	$\bar{x}_3$	n	S_{33}	S_{3y}	S_{yy}
A	$\bar{y}_A$ = 57.0	$\bar{x}_{3A}$ = 18.6	n_A = 15	$\sum_i (x_{i3} - \bar{x}_{3A})^2$ = 194.5	$\sum_i (x_{i3} - \bar{x}_{3A})(y_i - \bar{y}_A)$ = 602.8	$\sum_i (y_i - \bar{y}_A)^2$ = 1944.8
B	$\bar{y}_B$ = 37.0	$\bar{x}_{3B}$ = 15.8	n_B = 16	$\sum_i (x_{i3} - \bar{x}_{3B})^2$ = 143.9	$\sum_i (x_{i3} - \bar{x}_{3B})(y_i - \bar{y}_B)$ = 342.7	$\sum_i (y_i - \bar{y}_B)^2$ = 917.7
C	$\bar{y}_C$ = 23.9	$\bar{x}_{3C}$ = 14.1	n_C = 14	$\sum_i (x_{i3} - \bar{x}_{3C})^2$ = 355.3	$\sum_i (x_{i3} - \bar{x}_{3C})(y_i - \bar{y}_C)$ =315.0	$\sum_i (y_i - \bar{y}_C)^2$ = 336.6

Notations: For group A, $\bar{y}_A$ is the group sample mean of response y and $\bar{x}_{3A}$ is the group sample mean of the covariate x_3. For groups B and C, $\bar{y}_B$ and $\bar{y}_C$ as well as $\bar{x}_{3B}$ and $\bar{x}_{3C}$ are similarly defined.

```
Residual standard error: 4.707 on 41 degrees of freedom
Multiple R-squared:  0.9195,   Adjusted R-squared:  0.9136
F-statistic: 156.1 on 3 and 41 DF,  p-value: < 2.2e-16
```

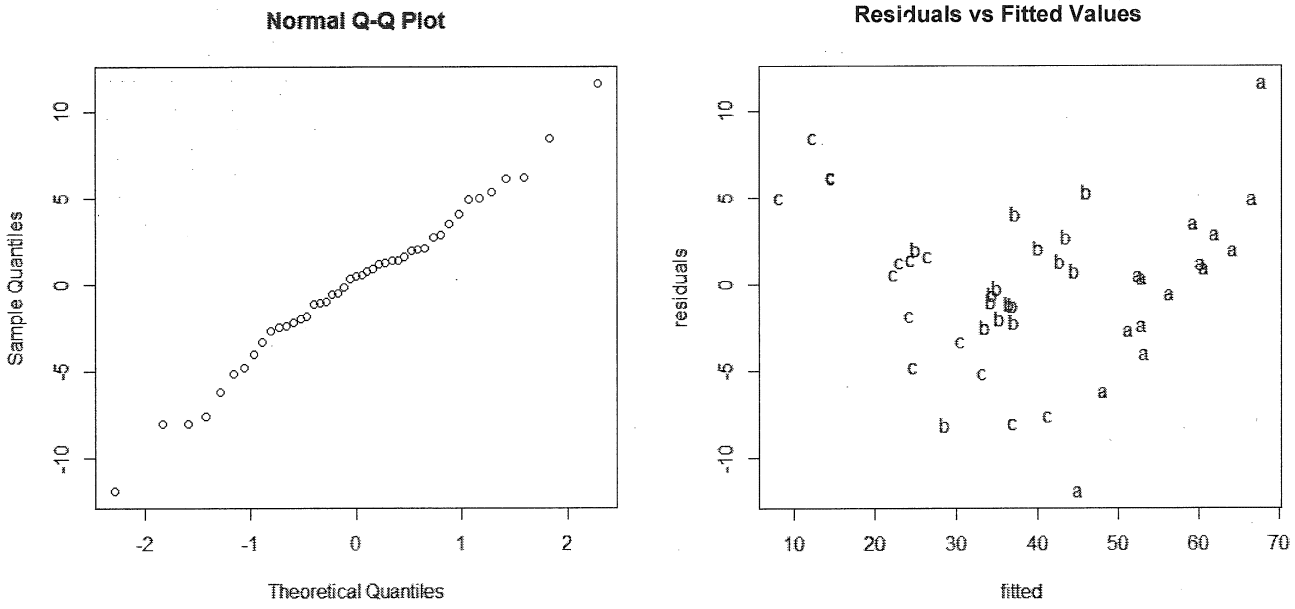
d) Calculate the Adjusted Group Means for Groups A, B, and C, respectively.

e) Assume that the (corrected) Total Sum of Squares for response y = 11284.2. Complete the following ANOVA table.

Source	d.f.	SS	MS
$SS_R(\beta_3\|\beta_0)$			
$SS_R(\beta_1,\beta_2\|\beta_3,\beta_0)$			
Residual			
Total (corrected)		11284.2	----

f) Use the ANOVA table to perform a hypothesis test at the 5% significance level for H_o: $\beta_1 = \beta_2 = 0$ vs the alternative that at least one of these coefficients is non-zero. Define the test statistic, obtain the critical value, and state your conclusion clearly.

g) Two residual plots are provided below. In the right-hand plot the observations are identified by plotting the group to which they belong. Based on these plots, comment on the appropriateness of the modeling assumptions. If you identify any possible problems with the model, explain how you would address these weaknesses.



METHODS 210-211-212 (2018), Problem 3

3. The Progabide clinical trial designed to assess whether treatment with progabide can lower the rate of seizures among epileptic patients (relative to placebo). In this study patients were observed for 8 weeks and the number of seizures were recorded (denote this as Y_{i0}, $i = 1, \ldots, 59$). At that point, patients were randomized to receive progabide ($\texttt{trt}_i$=1) or placebo ($\texttt{trt}_i = 0$), and the number of seizures were observed every two weeks for eight additional weeks. Let Y_{ij} denote the number of seizures on patient i during period j, $j = 1, 2, 3, 4$, and let T_{ij} denote the length of the observation period during period j, $j = 0, 1, 2, 3, 4$. So for each patient, $T_{i0} = 8$ weeks and $T_{ij} = 2$ weeks for $j = 1, ..., 4$. For this problem we will ignore response measures at visits 1, 2, and 3, and only focus on week 4.

   (a) We will start by considering a regression model of the following form:

   $$\log(\mu_{i4}) = \gamma_0 + \gamma_1 \texttt{trt}_i,$$

   where it is assumed that $Y_{i4} \sim \text{Poisson}(\mu_{i4})$, $i = 1, \ldots, 59$. Provide a precise interpretation of γ_0 and γ_1 (or some suitable transformation of these parameters) in words that can be understood by a statistical layman.

   (b) Note that the model in (a) does not contain an offset term. Should one be included? Explain.

   (c) Under the model formulation in (a), derive the:
   i. log-likelihood function
   ii. score function

   (d) There are three common (frequentist) statistical tests that could be used to test $H_0 : \gamma_1 = 0$. Name and define each of them, then draw a picture to illustrate how each is related to the log-likelihood function.

   (e) The output in APPENDIX - PROBLEM 3 provides two separate fits for the linear predictor provided in part (a). The first assumes $Y_{i4} \sim \text{Poisson}(\mu_{i4})$. The second is a quasi-Poisson fit. Write down in notation how the scale parameter (call it ϕ) in the quasi-Poisson fit is computed and use the output provided to give an estimate of ϕ. Based upon your findings, does the assumption that $Y_{i4} \sim \text{Poisson}(\mu_{i4})$ seems reasonable?

   (f) The trial included both male and female subjects. Suppose that in truth the following model holds:

   $$\log(\mu_{i4}) = \delta_0 + \delta_1 \texttt{trt}_i + \delta_2 \texttt{female}_i,$$

   where $Y_{i4} \sim \text{Poisson}(\mu_{i4})$ and $\texttt{female}_i$ is an indicator of female sex for patient i, $i = 1, \ldots, 59$, and $\delta_2 \neq 0$. Further, assume that the proportion of epileptic patients that are female is given by π. Find the mean and variance of $Y_i | \texttt{trt}_i$ (ie. marginalized over sex). From this, would inference based upon the model assumptions given in part (a) yield asymptotically valid inference for the effect of progabide on the rate of seizures? Explain?

   (g) If the model in (f) were true, write down the asymptotic distribution of $\widehat{\vec{\gamma}}$, the estimator obtained by setting the score function you wrote down in (c) to zero and solving. Each element of the asymptotic distribution should be fully defined.

   (h) Again suppose that the model in (f) were true. Would the quasi-likelihood model provided in APPENDIX - PROBLEM 3 provide (asymptocially) valid inference? If yes, explain why. If not, provide (in detail) how you would compute the variance for $\widehat{\vec{\gamma}}$ so that (asymptocially) valid inference for γ_1 could be obtained.

(i) Now suppose that it was suggested by a colleague that you should try to incorporate the baseline number of seizures as a predictor into the model in (a). Specifically, it was suggested that you fit a model of the form:

$$\log(\mu_{i4}) = \beta_0 + \beta_1 \texttt{trt}_i + \beta_2 \log(Y_{i0}).$$

Provide a precise interpretation of β_1 (or some suitable transformation of this parameter) in words that can be understood by a statistical layman. Explain why you would or would not a priori choose model the systematic model in (a) or (i).

(j) Finally, suppose you wanted to investigate whether the functional form of Y_{i0} (ie. $\log(Y_{i0})$) in the model given in (i) is correctly specified. Explain in detail how you would do this. A complete answer will

i. Name and mathematically define any residual(s) you would consider;

ii. Specify what (if any) plot(s) you would consider;

iii. Specify what you would expect to see in the plot if the functional form of Y_{i0} were correctly specified. (If you would consider one.)

METHODS 210-211-212 (2018) , Problem 4

4. In this problem we will consider data from a randomized trial seeking to investigate whether or not a new experimental therapy can slow cognitive decline in elderly adults. $N = 516$ subjects were randomized in a 1:1 fashion to either receive the new therapy (delivered in pill form) or a placebo that visually matched the therapy. The outcome of interest in the trial is known as the modified mini-mental status exam (3MSE) score. This exam is a broad measure of cognition and a patient's score on the exam can range from 0 to 100. Patients were measured at baseline (before treatment), then once a year for up to 8 years.

   (a) A common approach to analyzing these data would be to consider a marginal mean model of the form

   $$\mathrm{E}[Y_{ij}] = \beta_0 + \beta_2 \texttt{trt}_i + \beta_3 \texttt{year}_{ij} + \beta_3 \texttt{trt}_i \times \texttt{year}_{ij},$$

   where Y_{ij} is the observed 3MSE score for subject i at visit j, $\texttt{trt}_i$ is an indicator of whether patient i was randomized to treatment, and $\texttt{year}_{ij}$ is the year of the j-th visit (1, ..., 9; 1 denoting baseline). Provide precise interpretations of each parameter in the above model (or appropriate transformations of the parameters) and identify which parameter is of primary interest given the goals of the study.

   (b) You suggest to your colleagues that one way to estimate the model parameters in (a) is to use the general linear model for correlated data. In this case, you explain, that it is necessary to estimate the covariance of $\vec{Y}_i = (Y_{i1}, \ldots, Y_{i9})$, $i = 1, \ldots, N$. Let $\mathbf{V}_i$ denote the assumed covariance for $\vec{Y}_i$. Write down the estimating equation for the general linear model in this case. Each component of your estimating equation should be fully defined.

   (c) Your colleague says that they have been told that unbiased estimates of $\vec{\beta} = (\beta_0, \ldots, \beta_3)$ can be obtained by using simple ordinary least squares estimation. Is your colleague correct? Justify your answer.

   (d) You explain to your colleague that using the general linear model for correlated data yields "better" estimates, particularly if you have correctly specified the structure of the covariance of $\vec{Y}_i$, $i = 1, \ldots, N$. Precisely state what is meant by "better" in this case by precisely stating the Gauss-Markov theorem and showing that the Gauss-Markov theorem applies to the general linear model for correlated data case when the structure of the covariance of $\vec{Y}_i$, $i = 1, \ldots, N$ is correctly specified.

   (e) In order to propose a reasonable structure for the covariance of $\vec{Y}_i$, you tell your colleague that an empirical variogram plot can be helpful as it seeks to decompose the covariance structure into *trait*, *state*, and *measurement error* components. Figure 1 displays the variogram resulting from these data. Break the y-axis of the plot into segments highlighting the different components above, clearly labeling each, and specify the form of covariance structure that you might assume based upon this diagnostic.

   (f) Letting $\mathbf{V}_i$, $i = 1, \ldots, N$ denote the assumed covariance structure you specified in (e), provide (in detail) an iterative algorithm that can be used to obtain an estimate of $\vec{\beta}$ by solving the estimating equation you wrote down in (b).

   (g) Suppose that the true covariance of $\vec{Y}_i$ is $\mathbf{V}_i^* \neq \mathbf{V}_i$, $i = 1, \ldots, N$. What is the asymptotic distribution of $\widehat{\vec{\beta}}$, the estimator obtained from your iterative algorithm in (f)? Each element of the distribution should be clearly defined. Explain in detail how you would obtain a consistent estimate of $\mathrm{Var}[\widehat{\vec{\beta}}]$.

   (h) Your supervisor later comes to you and tells you that there is company interest in determining if there is heterogeneity in the rates of cognitive decline among patients (conditional upon treatment). Write down a linear mixed effects model (defining each component) that can be used to address your supervisor's question and state explicitly how you would go about testing whether or not there was heterogeneity in the rates of cognitive decline among patients.

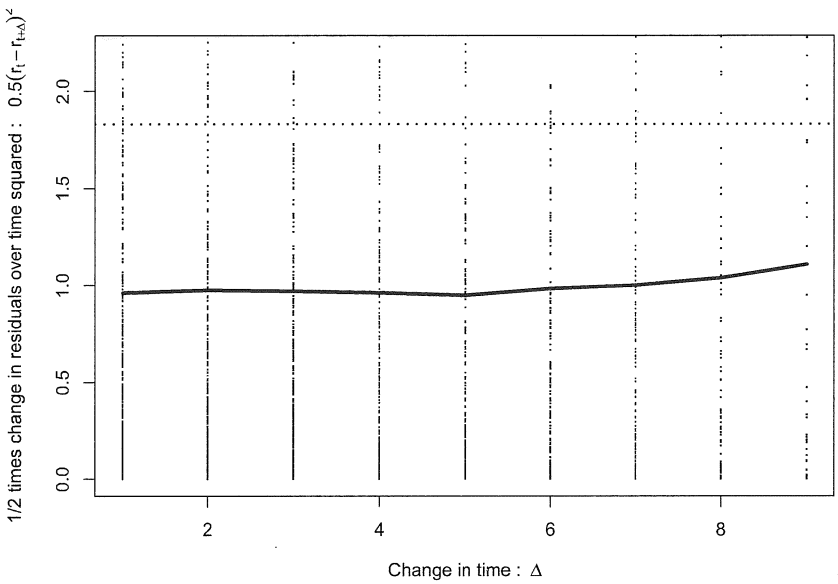


Figure 1: Plot of variogram over time. The horizontal dashed line represents total estimated variation.

(i) Finally, you learn that a 3MSE score below 80 is a clinical threshold of importance and so you create indicator variables $Z_{ij} \equiv I_{Y_{ij}<80}$, $i = 1, \ldots, N, j = 1, \ldots, 9$. You then specify a marginal mean model of the form

$$\text{logit}(\mu_{ij}) = \gamma_0 + \gamma_1 \texttt{trt}_i + \gamma_2 \texttt{year}_{ij} + \gamma_3 \texttt{trt}_i \times \texttt{year}_{ij},$$

where $\mu_{ij} = \Pr[Z_{ij} = 1]$. Write down the form of the estimating equation that would be used to obtain estimates of $\vec{\gamma} = (\gamma_0, \ldots, \gamma_3)$ in a generalized estimating equations (GEE) framework. Each element of your estimating equation should be fully defined.

(j) Suppose you assume an exchangeable covariance structure for $\vec{Z}_i$. Provide (in detail) the iterative algorithm utilized in GEE to obtain an estimate of $\vec{\gamma}$.

(k) What is the asymptotic distribution of $\widehat{\vec{\gamma}}$, the estimator from your algorithm in (j) if the assumption of an exchangeable covariance structure is incorrect? Each element of the distribution should be clearly defined.

APPENDIX - PROBLEM 3

```
##
##### Poisson Fit
##
> fit1 <- glm( seiz ~ tx, family=poisson, data=seizure, subset=time==4 )
> summary(fit1)

Call:
glm(formula = seiz ~ tx, family = poisson, data = seizure, subset = time ==
    4)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-3.991  -2.018  -1.132   0.421  13.023

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.0750     0.0670   30.99   <2e-16 ***
tx           -0.1714     0.0964   -1.78    0.075 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 476.25  on 58  degrees of freedom
Residual deviance: 473.08  on 57  degrees of freedom
AIC: 664.9

##
##### Quasi-Poisson Fit
##
> fit2 <- glm( seiz ~ tx, family=quasipoisson, data=seizure, subset=time==4 )
> summary(fit2)

Call:
glm(formula = seiz ~ tx, family = quasipoisson, data = seizure,
    subset = time == 4)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-3.991  -2.018  -1.132   0.421  13.023

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.075      0.245    8.46  1.2e-11 ***
tx            -0.171      0.353   -0.49     0.63
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for quasipoisson family taken to be ??????)

    Null deviance: 476.25  on 58  degrees of freedom
Residual deviance: 473.08  on 57  degrees of freedom
AIC: NA
```