STAT 659 Spring 2016
Homework 5 Solution

3.11

(a) Since $\log \mu_A = \alpha, \log \mu_B = \alpha + \beta$, thus $\beta = \log \mu_B - \log \mu_A = \log(\mu_B/\mu_A)$. Then $e^\beta = \mu_B/\mu_A$.

(b) After fitting the data to the poisson model, the fitted prediction equation is $\log \hat{\mu} = 1.609 + 0.588x$, where $\hat{\beta} = 0.588 = \log(\hat{\mu}_B/\hat{\mu}_A)$. We can see $\hat{\beta}$ gives the log of ratio between the number of imperfections of treatment B and the number of imperfections of treatment A.

(c) The standard error of $\hat{\beta} = 0.1764$. So for the Wald test, the test statistic is $\frac{\hat{\beta}-0}{Se_{\hat{\beta}}} = 3.332$ which has approximately a standard normal distribution. The corresponding P-value is $0.00086 < 0.05$, so we should reject the null hypothesis that $\beta = 0$. For the likelihood ratio test, the test statistic is 11.589 which has a chi-square distribution with degree of freedom one. The P-value is $0.00066 < 0.05$, so we reject H_0.

(d) Since $\beta = \log(\mu_B/\mu_A)$, we can first construct a Wald confidence interval for β, which is $0.588 \pm 1.96 * 0.1764 = (0.2423, 0.9337)$. Then the 95 percent confidence interval for μ_B/μ_A is $(e^{0.2423}, e^{0.9337}) = (1.274, 2.544)$.

3.12

First we consider the model with an interaction term xz which is $\log \mu = \alpha+\beta_1 x+\beta_2 z+\beta_3 xz$. The output gives the P-value of the interaction term is 0.4458, so this term is not significant. Then we consider the reduced model $\log \mu = \alpha + \beta x + \gamma z$. The fitted model is $\log \mu = 1.7177 + 0.5878x - 0.2296z$. The standard error for $\hat{\beta}$ is $Se_{\hat{\beta}} = 0.1764$ and the standard error for $\hat{\gamma}$ is $Se_{\hat{\gamma}} = 0.1701$. Then, the test statistic for $\hat{\beta}$ is 3.32 with P-value 0.000861 and the test statistic for $\hat{\gamma}$ is -1.349 with P-value 0.17725. Thus, we can make the conclusion: the effects of treatment type is significant but the effects of thickness of coating is not significant.

3.13

(a) The fitted model is $\log Y = -0.4284 + 0.5893x$. (Notice that here the unit of weight is kg).

(b) We can plug in $x = 2.44$ to obtain the estimated mean of Y which is $\log \hat{Y} = -0.4284 + 0.5893 * 2.44 = 1.0095$. Hence, $\hat{Y} = e^{1.0095} = 2.7442$.

(c) The effect of a 1kg increase in weight is a multiplicative effect of $e^{0.5893} = 1.8027$. The confidence interval for β is $0.5893 \pm 1.96 * 0.06502 = (0.4619, 0.7167)$ while the CI for the ratio of Y is $(e^{0.4619}, e^{0.7167}) = (1.587, 2.048)$.

(d) The null hypothesis is $H_0 : \beta = 0$ and alternative is $\beta \neq 0$. The standard error for $\hat{\beta} = 0.06502$, so the test statistic is $\frac{0.5893-0}{0.06502} = 9.064$ with p-value close to zero. Thus we should reject null hypothesis that Y is independent of weight.

(e) The reduced model is $\log y = 1.0713$. The chi-squared test statistic is 71.925 with p-value close to zero. So we make the same conclusion that Y is not independent of weight.

3.14

(a) The fitted model is $\log \mu = -0.8647 + 0.7603x$, where x is the weight. The standard error of $\hat{\beta}$ is 0.1578. The estimated dispersion parameter is 1.0741 with standard error 0.194. So the 95 percent confidence interval for dispersion parameter is $1.0741 \pm 1.96 * 0.194 = (0.69386, 1.45434)$ which excludes zero. So we can see the estimated dispersion parameter is not zero and the negative binomial log-linear model is a better fit.

(b) The confidence interval for β is $0.7603 \pm 1.96 * 0.1769 = (0.4136, 1.1069)$ which is wider than the CI obtained by poisson log-linear model. This is because $\hat{D} > 0$ and the negative binomial model has greater estimated standard deviation; the poisson log-linear model can not capture this over-dispersion.

3.15

(a) $\log \hat{\mu}_{1j} = -2.38 + 1.733 = -0.647$. So the estimated population mean is $e^{-0.647} = 0.524$ for blacks. $\log \hat{\mu}_{2j} = -2.38$, so the estimated population mean for whites is $e^{-2.38} = 0.0926$.

(b) Since $\hat{\beta} = \log(\hat{\mu}_{1j}/\hat{\mu}_{2j})$, we can first construct the CI for β. The 95 percent CI for β is $1.733 \pm 1.96 * 0.147 = (1.445, 2.021)$, so the 95 percent CI for $\mu_{1j}/\mu_{2j} = (e^{1.445}, e^{2.021}) = (4.24, 7.55)$.

(c) The CI obtained by negative binomial log-linear model is more appropriate. We can see the standard error of $\hat{\beta}$ by negative binomial GLM fit is greater that by poisson GLM fit, also the CI by negative binomial GLM fit is wider than that of the poisson GLM fit. Since poisson model is a special case of the negative binomial model with dispersion parameter equals to zero, so it implies that the over-dispersion is not captured by the poisson model. Also, the ratio of the sample mean to the sample variance seems to be far from 1. This implies that the Poisson model may not be appropriate for this data.

(d) The 95 percent CI of the dispersion parameter is $(4.94 \pm 1.96 * 1) = (2.98, 6.9)$ which excludes zero. So the negative binomial GLM is more appropriate than the poisson model.

3.17

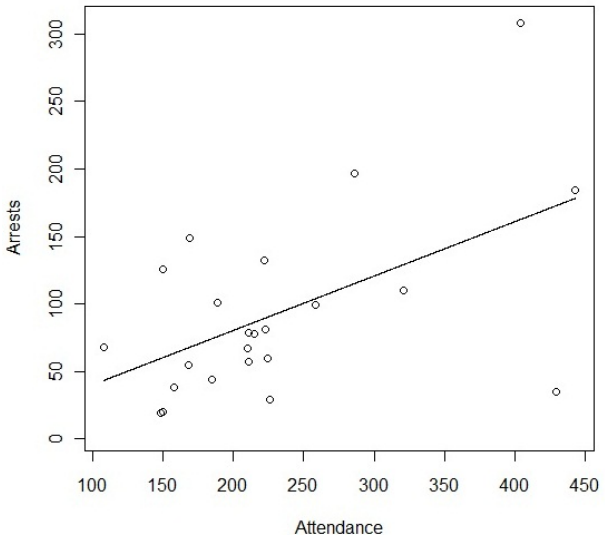
$\log \hat{\mu} - \log(38.7) = 2.549$ and the standard error for the intercept is 0.4495. It indicates that 95 % confidence interval for the log rate is $2.549 \pm 1.96(0.04495)$, which is $(2.46, 2.64)$. Hence, 95 % confidence interval for the true rate is $(11.7, 14.0)$.

3.18

(a) From the data, we can see the number of record of arrests will increase with the number of attendances. Then we can model the rate of the arrests over a mount of attendances. Since $E[Y] = \mu t$, then we have $\log(E[Y]) = \log \mu + \log t \Rightarrow \log(E[Y]/t) = \log \mu = \alpha \Rightarrow \log \mu - \log t = \alpha$ where the offset term is $-\log t$.

(b) The fitted model is $\log \hat{\mu} - \log t = -0.9103$ and the standard error for $\hat{\alpha}$ is 0.0216. So $\frac{\hat{\mu}}{t} = \exp(-0.9103) = 0.4024$, that is $\hat{\mu} = 0.4024t$.

(c) The plot of arrests against attendance is as follows:



The residuals are as follows:

(d) When fitting the data to negative binomial model, the $\hat{\alpha} = -0.9052$ with standard error 0.12 which is much larger than the standard error of $\hat{\alpha}$ obtained by poisson model. The estimated dispersion parameter is 0.3189 with 95 percent Wald confidence interval $(0.2253, 0.4125)$ which excludes zero. Thus the poisson assumption is not appropriate.

3.19

(a) The test statistic is $TS = 35.1 - 23.5 = 11.6 \sim \chi^2_1$ and the corresponding P-value is $0.00066 < 0.05$. So we should reject the null hypothesis that the collision counts are independent Poisson variates with constant rate over 29 years.

Observation	Raw Residual	Pearson Residual	Deviance Residual	Std Deviance Residual	Std Pearson Residual	Likelihood Residual
1	145.42577	11.405531	10.136615	10.545888	11.866039	10.652121
2	81.910324	7.6352007	6.9246837	7.119132	7.8496006	7.1603899
3	5.7317253	0.4292876	0.4270174	0.4460357	0.448407	0.4462341
4	80.992464	9.8212348	8.4702274	8.6083758	9.9814179	8.6554493
5	42.664657	4.5139485	4.2115954	4.3025302	4.6114115	4.3158917
6	65.638282	8.4484377	7.3605021	7.4667607	8.570402	7.5001789
7	-19.17408	-1.687045	-1.731598	-1.786458	-1.740494	-1.783712
8	24.944235	2.8602508	2.7220537	2.7718491	2.9125743	2.7769823
9	-4.822155	-0.473256	-0.476992	-0.489025	-0.485195	-0.48884
10	-8.737755	-0.922385	-0.937996	-0.958343	-0.942393	-0.957678
11	-5.908817	-0.641245	-0.648907	-0.662202	-0.654383	-0.661893
12	-8.518463	-0.915813	-0.931494	-0.950952	-0.934944	-0.950309
13	24.539563	3.722372	3.4354303	3.4709226	3.7608288	3.4770622
14	-17.50641	-1.904374	-1.976552	-2.016851	-1.943201	-2.013989
15	-30.14017	-3.174581	-3.381978	-3.45568	-3.243763	-3.447
16	-27.90882	-3.028761	-3.222669	-3.288697	-3.090816	-3.281058
17	-12.60512	-1.533054	-1.584829	-1.610521	-1.557906	-1.608882
18	-30.44612	-3.528669	-3.822862	-3.891274	-3.591817	-3.881226
19	-25.58101	-3.208145	-3.470251	-3.523082	-3.256985	-3.515452
20	-137.6345	-10.47523	-12.7891	-13.33951	-10.92606	-13.1609
21	-61.94499	-6.495563	-7.589358	-7.756275	-6.638423	-7.711983
22	-40.36172	-5.195039	-6.04471	-6.131974	-5.270037	-6.109286
23	-40.5569	-5.255314	-6.139955	-6.227387	-5.330148	-6.204128

(b) The test statistic for β is $\frac{\hat{\beta}-0}{Se_{\hat{\beta}}} = \frac{-0.0337}{0.013} = -2.5923$ with P-value $0.0096 < 0.05$. So we reject the null hypothesis that $\beta = 0$.

(c) The confidence interval for the multiplicative annual effect on the accident rate is $(e^{-0.06}, e^{-0.008}) = (0.9418, 0.9920)$. It means when the year increased by one, then the increase of accident rate will fall into this interval.

3.20

(a) The death rates per 1000 person years for nonsmokers and smokers and the relative risks of nonsmokers and smokers in each age stage are as follows:

Age	Nonsmokers	Smokers	RelativeRisk
$35-44$	0.1064	0.6106	5.7376
$45-54$	1.1243	2.4047	2.1388
$55-64$	4.904	7.1998	1.4682
$65-74$	10.8317	14.6885	1.3561
$75-84$	21.2038	19.1838	0.9047

We can see the relative risk of coronary death rates per 1000 person years of smokers and nonsmokers are decreasing as the age increases. So it is possible that the coronary death rates also depend on age.

(b) Let x_i be the indicator variable for the first ith age group, where $i = 1, 2, 3, 4$. $x_i = 1$, if the person belongs to ith age group and $x_i = 0$ otherwise. Let z be the indicator variable for the smoking habits. $z = 1$ for smokers and $z = 0$ for nonsmokers. The model is $\log \mu = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4 + \beta z$, so $\log \mu_{\text{smokers}} - \log \mu_{\text{nonsmokers}} = \beta$ for i-th age group. But according to (a), we can see the relative risk of nonsmokers and smokers are different for different age stage. Since the age is ordinal and the ratio of death rate for smokers and nonsmokers at each age period is not a constant, an interaction term of age and smoking status is possible.

(c) We can assign scores to the different age stages. Let the score be i for ith age status, where $i = 1, 2, 3, 4, 5$. Then the model with a quantitative interaction term of age and smoking status is $\log \mu = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4 + \beta z + \gamma z \cdot \text{age}$ So $\log \mu_{\text{smokers}} - \log \mu_{\text{nonsmokers}} = \log(\mu_{\text{smokers}}/\mu_{\text{nonsmokers}}) = \beta + \gamma * \text{age}$ for a given age stage. Thus the log of the ratio of coronary death rates changes linearly with age.

(d) The fitted model for (b) is $\log \mu = -4.219 - 3.7x_1 - 2.216x_2 - 1.073x_3 - 0.350x_4 + 0.355z$. When including the interaction term, the fitted model is $\log \mu = -3.855 - 4.731x_1 - 2.996x_2 - 1.583x_3 - 0.589x_4 + 2.3558z - 0.031 \cdot z\text{age}$. Since the model (b) is nested in model (c), we can see the difference of the deviances of the above two models will have a chi-squared distribution with degree of freedom one under the null hypothesis $\gamma = 0$. The difference of deviances of the two models is $12.1339 - 1.5464 = 10.5875$ with P-value $0.0011 < 0.05$, therefore we should reject the null hypothesis.

Additional Problems

1. SAS output:

Summary Measures for Poisson Regression for the Crab Data

modnum	model_name	G2	DF	loglikhd	numpar	aic	aicc	daic	daicc
1		632.792	172	35.9898	1	634.792	634.800	72.9867	72.8744
2	color	609.139	169	47.8160	4	617.139	617.219	55.3342	55.2939
3	spine	621.161	170	41.8053	3	627.161	627.208	65.3556	65.2833
4	width	567.879	171	68.4463	2	571.879	571.902	10.0736	9.9773
5	weight	560.866	171	71.9524	2	564.866	564.890	3.0615	2.9651
6	color spine	602.193	167	51.2890	6	614.193	614.362	52.3882	52.4366
7	color width	559.345	168	72.7132	5	569.345	569.465	7.5398	7.5398
8	color weight	551.805	168	76.4831	5	561.805	561.925	0.0000	0.0000
9	spine width	566.605	169	69.0831	4	574.605	574.685	12.8000	12.7597
10	spine weight	559.488	169	72.6418	4	567.488	567.568	5.6827	5.6424
11	width weight	559.901	170	72.4353	3	565.901	565.949	4.0957	4.0233
12	color spine width	558.629	166	73.0708	7	572.629	572.855	10.8245	10.9297
13	color spine weight	549.702	166	77.5343	7	563.702	563.928	1.8975	2.0026
14	color width weight	551.380	167	76.6954	6	563.380	563.549	1.5754	1.6238
15	spine width weight	558.826	168	72.9725	5	568.826	568.946	7.0212	7.0212
16	color spine width weight	549.586	165	77.5928	8	565.586	565.876	3.7806	3.9507

When the model is (color , weight), there are the smallest AIC and AICc. So, it can be the first candidate.

2. Zero inflated negative binomial regression model with weight in the zero model has the smallest AIC and AICc.

model	AIC	AICc
poisson	920.1641	920.2347
zero inflated poisson with intercept only	762.0018	762.1438
zero inflated poisson with weight	735.2043	735.4424
negative binomial	754.6437	754.7857
zero inflated negative binomial with intercept only	740.1369	740.3750
zero inflated negative binomial with weight	715.5272	715.8865

Only for students having taken STAT 414, 610, 630, or another mathematical statistics course

3.21

We can see the model is equivalent to $\mu = \alpha t + \beta xt$. So it is a linear model with identity link for μ. The t and xt are covariates and this linear model does not have an intercept.