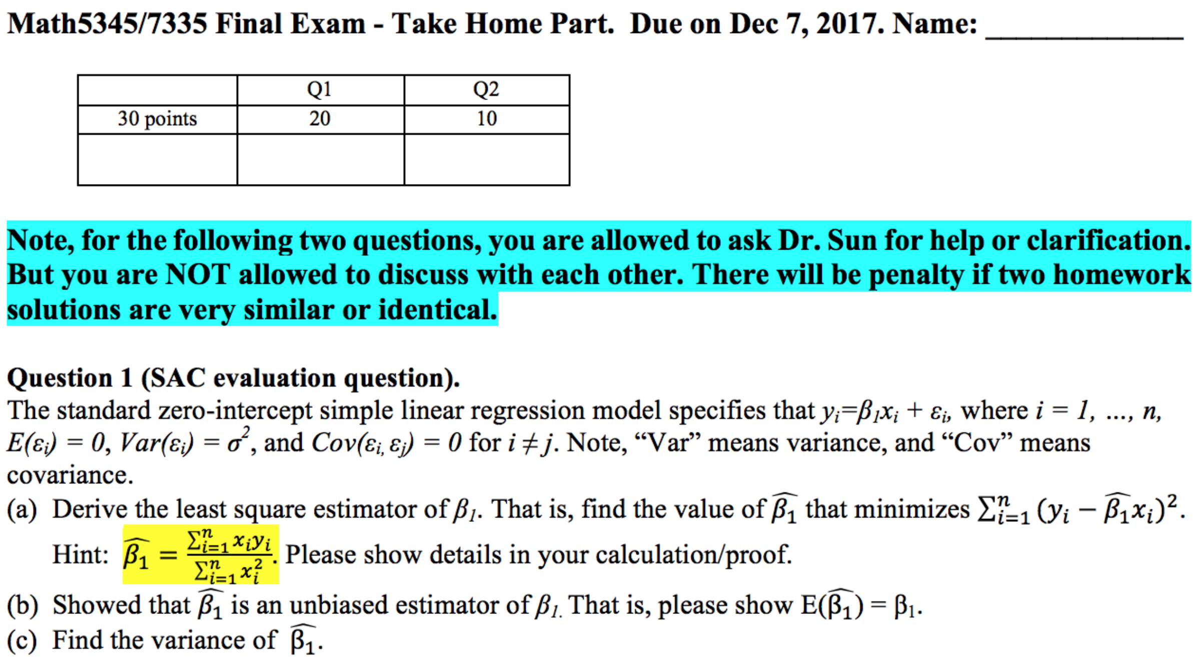


Math5345/7335 Final Exam - Take Home Part. Due on Dec 7, 2017. Name: ______________

	Q1	Q2
30 points	20	10

Note, for the following two questions, you are allowed to ask Dr. Sun for help or clarification. But you are NOT allowed to discuss with each other. There will be penalty if two homework solutions are very similar or identical.

Question 1 (SAC evaluation question).
The standard zero-intercept simple linear regression model specifies that $y_i = \beta_1 x_i + \varepsilon_i$, where $i = 1, \ldots, n$, $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$, and $Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$. Note, "Var" means variance, and "Cov" means covariance.

(a) Derive the least square estimator of β_1. That is, find the value of $\widehat{\beta_1}$ that minimizes $\sum_{i=1}^{n}(y_i - \widehat{\beta_1} x_i)^2$.
Hint: $\widehat{\beta_1} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$. Please show details in your calculation/proof.

(b) Showed that $\widehat{\beta_1}$ is an unbiased estimator of β_1. That is, please show $E(\widehat{\beta_1}) = \beta_1$.

(c) Find the variance of $\widehat{\beta_1}$.

(a)
The goal is to minimize the following equation with respect to $\hat{\beta}_1$.

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_i)^2$$

Since y$_i$ and x$_i$ for i = 1, 2, ... n are just constants from our sample, we can actually just take the derivative of this equation with respect to $\hat{\beta}_1$ and set it to 0 to find our minimum.
We know that the value we find will be a global minimum since the function we're minimizing is a degree 2 polynomial such that the squared term has a positive coefficient.

$$\frac{\mathrm{d}}{\mathrm{d}\hat{\beta}_1}\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_i)^2$$

$$2\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_i) \times (-1 x_i)$$

$$-2\sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_i) x_i$$

Let this derivative equal 0 now.

$$0 = -2\sum_{i=1}^{n}(y_i x_i - \hat{\beta}_1 x_i^2)$$

$$0 = \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \hat{\beta}_1 x_i^2$$

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

(b)

Since we are hypothesizing that our model is linear, we assume that:

$$y_i = \beta_1 x_i + \epsilon_i$$

where

$$\mathrm{E}[y_i] = \beta_1 x_i$$

From **(a)**:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2} \\
\hat{\beta}_1(\sum_{i=1}^n x_i^2) &= \sum_{i=1}^n y_i x_i \\
\mathrm{E}[\hat{\beta}_1(\sum_{i=1}^n x_i^2)] &= \mathrm{E}[\sum_{i=1}^n y_i x_i] \\
&= \sum_{i=1}^n x_i \mathrm{E}[y_i] \\
&= \sum_{i=1}^n x_i \beta_1 x_i \\
&= \beta_1 \sum_{i=1}^n x_i^2
\end{aligned}$$

Also notice that:

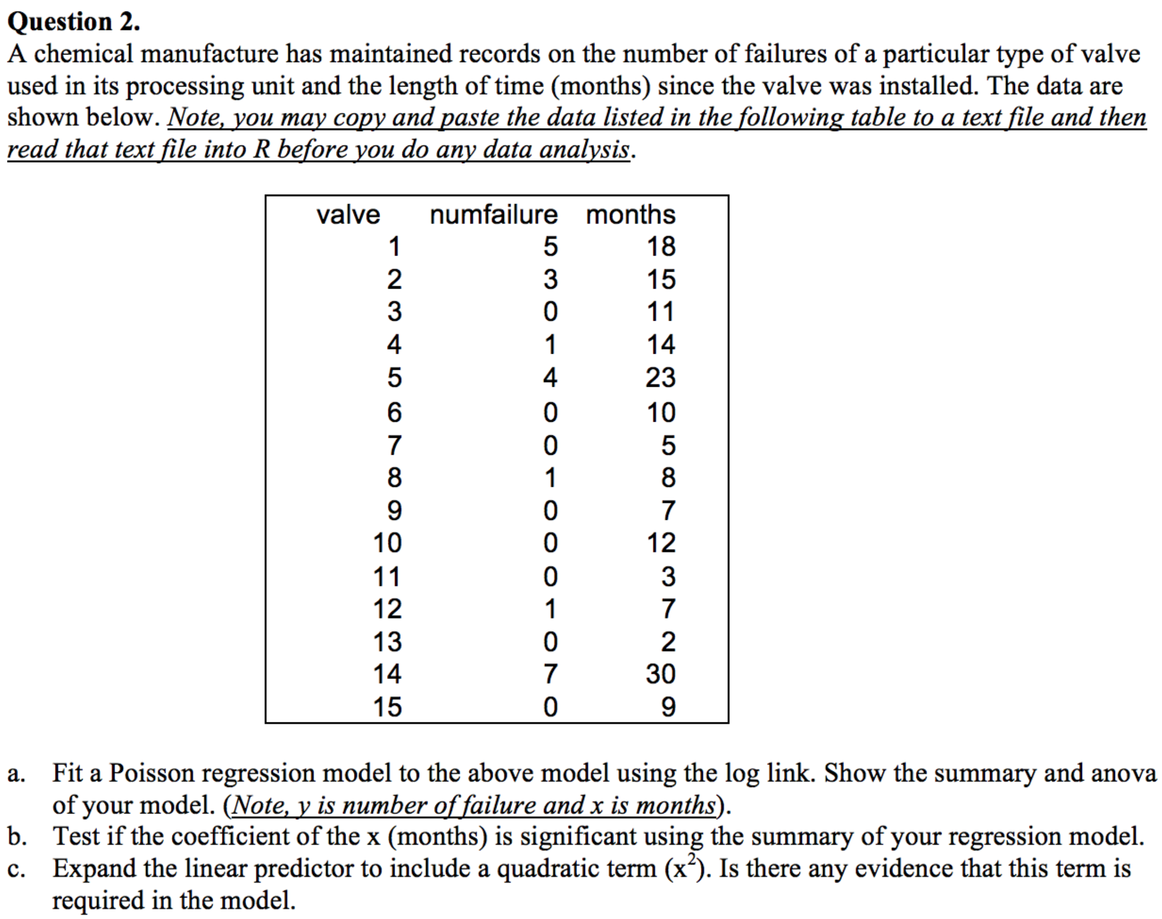
$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2} \\
\mathrm{E}[\hat{\beta}_1] &= \mathrm{E}[\frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2}] \\
\mathrm{E}[\hat{\beta}_1] &= \frac{\mathrm{E}[\sum_{i=1}^n y_i x_i]}{\sum_{i=1}^n x_i^2}
\end{aligned}$$

But we solved for the expected value of $\hat{\beta}_1$'s numerator above. So:

$$\begin{aligned}
\mathrm{E}[\hat{\beta}_1] &= \frac{\mathrm{E}[\sum_{i=1}^n y_i x_i]}{\sum_{i=1}^n x_i^2} \\
\mathrm{E}[\hat{\beta}_1] &= \frac{\beta_1 \sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2} \\
\mathrm{E}[\hat{\beta}_1] &= \beta_1
\end{aligned}$$

(c): Find the variance of $\hat{\beta}_1$: V[$\hat{\beta}_1$]

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2} \\
V[\hat{\beta}_1] &= V[\frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2}] \\
&= \frac{\sum_{i=1}^n x_i^2 V[y_i]}{(\sum_{i=1}^n x_i^2)^2} \\
&= \frac{\sigma^2 \sum_{i=1}^n x_i^2}{(\sum_{i=1}^n x_i^2)^2} \\
V[\hat{\beta}_1] &= \frac{\sigma^2}{\sum_{i=1}^n x_i^2}
\end{aligned}$$

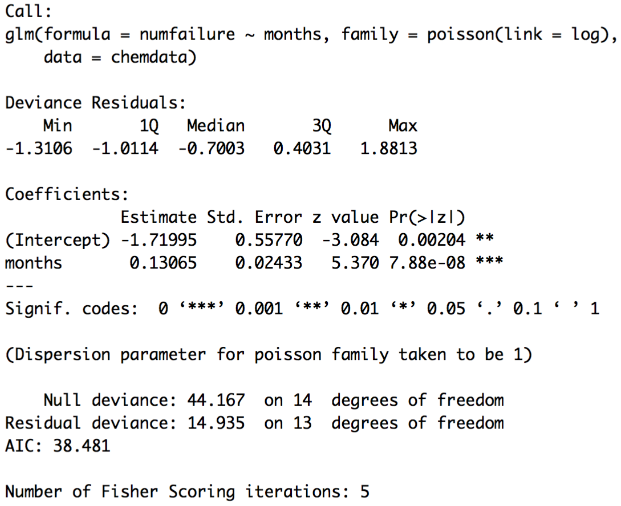


Question 2.
A chemical manufacture has maintained records on the number of failures of a particular type of valve used in its processing unit and the length of time (months) since the valve was installed. The data are shown below. *Note, you may copy and paste the data listed in the following table to a text file and then read that text file into R before you do any data analysis*.

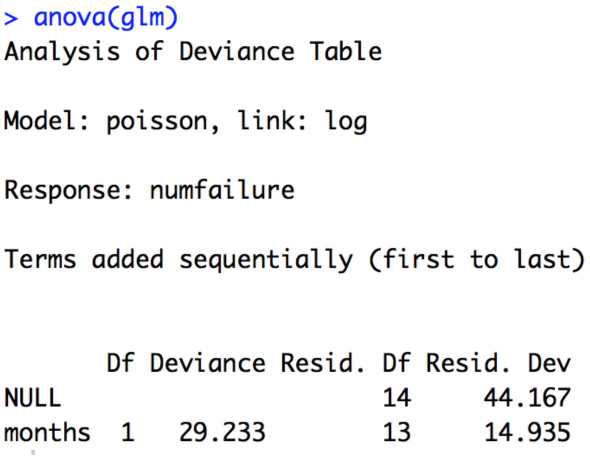
valve	numfailure	months
1	5	18
2	3	15
3	0	11
4	1	14
5	4	23
6	0	10
7	0	5
8	1	8
9	0	7
10	0	12
11	0	3
12	1	7
13	0	2
14	7	30
15	0	9

a. Fit a Poisson regression model to the above model using the log link. Show the summary and anova of your model. (*Note, y is number of failure and x is months*).
b. Test if the coefficient of the x (months) is significant using the summary of your regression model.
c. Expand the linear predictor to include a quadratic term (x^2). Is there any evidence that this term is required in the model.

(a)



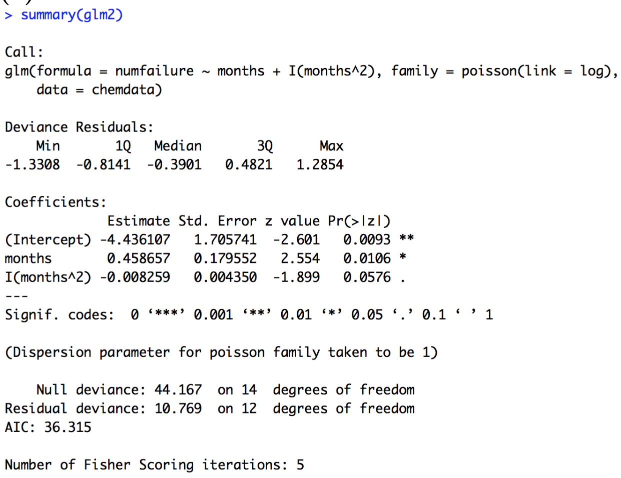
```
Call:
glm(formula = numfailure ~ months, family = poisson(link = log),
    data = chemdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3106  -1.0114  -0.7003   0.4031   1.8813

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.71995    0.55770  -3.084  0.00204 **
months       0.13065    0.02433   5.370 7.88e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 44.167  on 14  degrees of freedom
Residual deviance: 14.935  on 13  degrees of freedom
AIC: 38.481

Number of Fisher Scoring iterations: 5
```



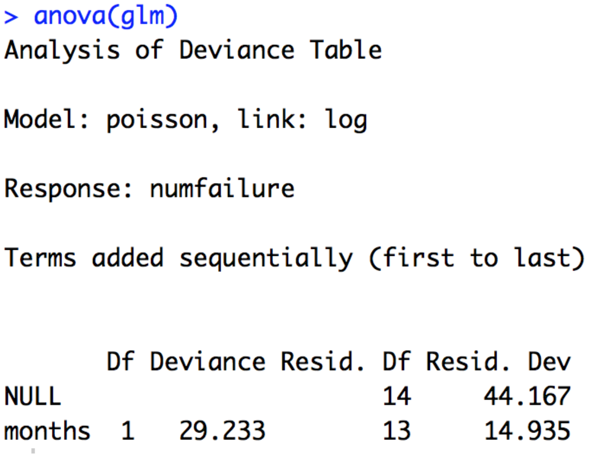
```
> anova(glm)
Analysis of Deviance Table

Model: poisson, link: log

Response: numfailure

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev
NULL                      14     44.167
months  1   29.233        13     14.935
```

(b)
Based on the Wald test statistic, months(***) is statistically significant.

(c)



```
> summary(glm2)

Call:
glm(formula = numfailure ~ months + I(months^2), family = poisson(link = log),
    data = chemdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3308  -0.8141  -0.3901   0.4821   1.2854

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.436107   1.705741  -2.601   0.0093 **
months       0.458657   0.179552   2.554   0.0106 *
I(months^2) -0.008259   0.004350  -1.899   0.0576 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 44.167  on 14  degrees of freedom
Residual deviance: 10.769  on 12  degrees of freedom
AIC: 36.315

Number of Fisher Scoring iterations: 5
```

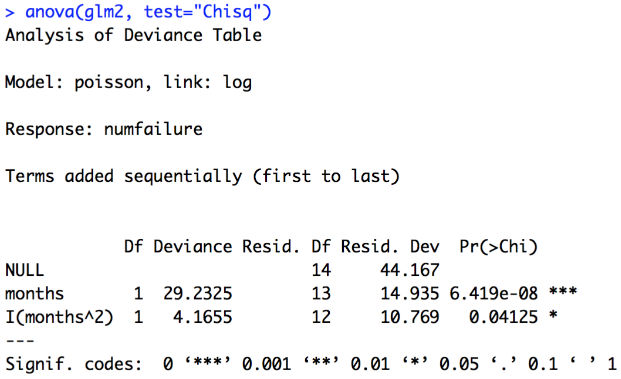


```
> anova(glm)
Analysis of Deviance Table

Model: poisson, link: log

Response: numfailure

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev
NULL                      14     44.167
months  1   29.233        13     14.935
```

Using another Wald test statistic for months^2, it looks as though there is some evidence (.) that the quadratic term is significant.

Using a log likelihood ratio test:

```
anova(glm2, test="Chisq")
```



```
> anova(glm2, test="Chisq")
Analysis of Deviance Table

Model: poisson, link: log

Response: numfailure

Terms added sequentially (first to last)


            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                           14     44.167
months       1  29.2325        13     14.935 6.419e-08 ***
I(months^2)  1   4.1655        12     10.769   0.04125 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

There is slightly more evidence that the quadratic term is significant.
So, overall, yes: there is some evidence to suggest that the quadratic term is statistically significant.