STAT 540, Fall 2015

Due on **12:00 A.M., Friday, Dec 11th**

Homework 12

Homework format for all STAT 540 homework this term: Please label all problems clearly and turn in an organized homework assignment. You don't need to spend hours producing beautifully typeset homework, but you won't get credit if we can't find or read your answer. Unless noted otherwise, turn in the following (as appropriate for the problem).

- Theoretical derivation (when asked for).
- Numerical results **with an explanation of your solution**, written in complete sentences. If computer code is absolutely necessary to provide context here, then include it–nicely formatted–within the solution (otherwise, see below).
- Appropriate graphics. Use informative labels, including titles and axis labels. Try to put multiple plots on the page by using, for example, the R command `par(mfrow=c(2,2))`.
- **Only as necessary**: Final clean computer code used to answer the problem **attached to the end of your homework**. Only include the rare code excerpts without which we wouldn't be able to figure out what you did. Annotate your code. Number and order the code in order of the problems. When in doubt, leave it out; consider that we will probably never read it.
- Some problems will be relatively open-ended, such as "Here are some data. Analyze them and write a report." I will provide further instructions about reports later. They should be self-contained, with suitable EDA, graphs, numerical results, and **scientific interpretation**. No computer code should be included. The report should be concise: "no longer than necessary".

(1) The file `temperature.txt` (from HW10) contains data on worldwide average temperature from 1880 to 1987. We wish to estimate the linear slope, i.e. the average increase in temperature per year.

(a) Fit the model $\text{Temp}_i = \beta_0 + \beta_1 \text{year}_i + \epsilon_i$ where $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$. What are the estimated slope, $\widehat{\beta}_1$ and its standard error?

(b) For these data, the assumption of independent errors is probably not reasonable. One reasonable alternative is that errors follow an AR(1) model. For that model, the first

few rows and columns of the variance-covariance matrix of ϵ can be written as

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

I have not written out the full 108×108 data matrix, for obvious reasons, but I hope you can see the pattern. Various lines of evidence suggest $\rho = 0.5$. What are the estimated slope and it standard error, if you assume $\rho = 0.5$?

(c) Based on what you know about consequences of misspecifying the variance-covariance matrix, are these results surprising? Explain why or why not.

(2) Suppose that $\mathbb{E}(\mathbf{y}|bX) = \mathbf{X}\boldsymbol{\beta}$ and $\text{Var}(\mathbf{y}|\mathbf{X}) = \boldsymbol{\Omega}$, and the usual assumption on $\mathbf{X}$ and $\boldsymbol{\Omega}$ apply. Now suppose the researcher is mistaken and believes that $\text{Var}(\mathbf{y}|\mathbf{X}) = \boldsymbol{\Sigma}$ instead, where $\boldsymbol{\Sigma} \neq \boldsymbol{\Omega}$.

(a) Is the GLS estimator using the wrong variance structure $\boldsymbol{\Sigma}$ unbiased?

(b) What is the variance of the GLS estimator based on $\boldsymbol{\Sigma}$?

(3) If $y \sim \text{Pois}(\lambda)$, then the pmf of y is $f(y) = \lambda^y e^{-\lambda}/y!$. Recall the definition of exponential-scale family from of a distribution from the lecture note, write the pmf of y in exponential-scale form and identify $\theta, \eta(\theta), T(y), a(\phi)$ and $b(\theta)$.

(4) The table below contains information on 23 (out of 24) pre-Challenger space shuttle flights. (On one flight, the solid rocket motors were lost at sea and so no data are available.) Provided are launch temperatures, t (in $^\circ F$), and a $0-1$ response, y , indicating whether there was post-launch evidence of a field joint primary O-ring incident. (O-ring failure was apparently responsible for the tragedy.) $y = 1$ indicates that at least one of the 6 primary O-rings showed evidence of erosion.

Temperature	O-ring Incident	Temperature	O-ring Incident
66	0	67	0
70	1	53	1
69	0	67	0
68	0	75	0
67	0	70	0
72	0	81	0
73	0	76	0
70	0	79	0
57	1	75	1
63	1	76	0
70	1	58	1
78	0		

Treat the response variables, y_i, as Bernoulli distributed and independent launch to launch. Note that $\mu = \mathbb{E}(y) = p$ here. We can model

$$h(\mu) = \log\left\{\frac{\mu/n}{1-\mu/n}\right\} = \log\left\{\frac{p}{1-p}\right\} = \beta_0 + \beta_1 t$$

which is a generalized linear model, with binomial (Bernoulli) response, and the "logit" link, that is also the logistic regression model. We may use `R` function `glm` to analyze them.

(a) Fit a logistic regression and estimate β_0, β_1. What is the implication of the estimated β_1 regarding the Challenger tragedy? NASA managers ordered the launch after arguing that these and other data data showed no relationship between temperature and O-ring failure. Was their claim correct? Explain.

(b) *glm* will provide estimated mean responses (and corresponding standard errors) for values of the explanatory variable(s) in the original data set. To see estimated means $\widehat{\mu}_i = \widehat{p}_i$ and corresponding standard errors, you can type

```
> shuttle.fits<-predict.glm(shuttle.out,type="response",se.fit=TRUE)
> shuttle.fits$fit
> shuttle.fits$se.fit
```

Plot estimated means versus t. Connect those with line segments to get a rough plot of the estimated relationship between t and p. Plot "2 standard error" bands around that response function as a rough indication of the precision with which the relationship between t and p could be known from the pre-Challenger data.

The temperature at Cape Canaveral for the last Challenger launch was 31 $^\circ F$. Of course, hind-sight is always perfect, but what does your analysis here say might have been expected in terms of O-ring performance at that temperature? You can get an estimated $31^\circ F$ mean and corresponding standard error by typing

```
> predict.glm(shuttle.out,data.frame(temp=31),se.fit=TRUE, type="response")
```

(5) An engineering researcher group works on a project aimed at reducing jams on a large collating machine. They ran the machine at 3 "Air Pressure" settings and 2 "Bar Tightness" conditions and observed y = the number of machine jams experienced in k seconds of machine run time. (Run time does not include the machine "down" time required to fix the jams.) Their results are below.

Air Pressure	Bar Tightness	y Jams	k Run Time
1(low)	1 (tight)	27	295
2(medium)	1	21	416
3 (high)	1	33	308
1	2 (loose)	15	474
2	2	6	540
3	2	11	498

Motivated perhaps by a model that says times between jams under a given machine set-up are independent and exponentially distributed, we will consider an analysis of these data based on a model that says the jam counts are independent Poisson variables. For $\mu =$ the mean count at air pressure i and bar tightness j, suppose that

$$\log(\mu_{ij}) = \mu + \alpha_i + \beta_j + \log(k_{ij}).$$

Notice that this says

$$\mu_{ij} = k_{ij} \exp(\mu + \alpha_i + \beta_j)$$

(If waiting times between jams are independent exponential random variables, the mean number of jams in a period should be a multiple of the length of the period, hence the multiplication here by k_{ij} is completely sensible.) Notice that the model equation is a special case of the relationship

$$\log(\mu_{ij}) = \mu + \alpha_i + \beta_j + \gamma \log(k_{ij})$$

which is in the form of a generalized linear model with link function $h(\mu) = \log(\mu)$. As it turns out, `glm` will fit a relationship like the proposed model for a Poisson mean that includes an "offset" term ($\log k_{ij}$ here). Enter the data for this problem and set things up by typing

```
> A<-c(1,2,3,1,2,3)
> B<-c(1,1,1,2,2,2)
> y<-c(27,21,33,15,6,11)
> k<-c(295,416,308,474,540,498)
> AA<-as.factor(A)
> BB<-as.factor(B)
> options(contrasts=c("contr.sum","contr.sum"))
```

(a) Fit and view some summaries for the Poisson generalized linear model (with log link and offset) by typing

```
> collator.out<-glm(y~AA+BB,family=poisson,offset=log(k))
> summary(collator.out)
```

The log link is the default for Poisson observations, so one doesn't have to specify it in the function call. Does it appear that there are statistically detectable Air Pressure and Bar Tightness effects in these data? Explain. If one wants small numbers of jams, which levels of Air Pressure and Bar Tightness does one want?

(b) What is that estimated "per second jam rates" in the model parameters? Give estimates of all 6 of these rates based on the fitted model.

(c) One can get `R` to find estimated means corresponding to the 6 combinations of Air Pressure and Bar Tightness for the corresponding values of k. This can either be done on the scale of the observations or on the log scale. To see these first of these, type

```
> collator.fits<-predict.glm(collator.out,type="response", se.fit=TRUE)
> collator.fits$fit
> collator.fits$se
```

How are the "fitted values" related to your values from b)? To see estimated/fitted log means and standard errors for those, type

```
> lcollator.fits<-predict.glm(collator.out,se.fit=TRUE)
> lcollator.fits$fit
> lcollator.fits$se
```

(6) Textbook problems:

(a) # 11.1 Unequal variance interpretation.

(b) # 11.6 Computer-assisted learning.

(c) # 11.18 Variance-covariance matrix of WLS estimates.

(d) # 14.4 Logistic regression.