Statistical Models
Examination Solutions
19 December 2012

General instructions: For each problem below, write your solution clearly and thoroughly, using consistent notation. You will receive credit for partial solutions. For computations you may use any non-programmable calculator, provided it is not part of a device that is capable of communication with other devices. The exam grade will equal the total points earned (maximum = 100) divided by 10.

1. A producer of woolen yarn uses nine different looms in her factory. Two types of wool are used in the looms (Type A and Type B), and the looms can be set to three different tension levels (Low, Medium and High). To study the effects of wool type and tension setting on the quality of yarn, a researcher measures the rate of warp breaks per fixed length of yarn at each of the nine looms for each of the six combinations of factor levels, for a total of 54 measurements. Let Y_{ijk} denote the rate of warp breaks for the kth loom using wool type i and tension setting j, where $i = 1, 2$, $j = 1, 2, 3$, and $k = 1, \dots, 9$.

   a. (5 pts.) Let μ denote the general mean, α_i denote the main effect of wool type i, β_j denote the main effect of tension setting j, and γ_{ij} denote the interaction effect of wool type i and tension setting j. Write the appropriate fully-parametrized two-way Analysis of Variance model Ω that the researcher should use to investigate the relationship between Y_{ijk} and these effects, using mathematical notation. Be sure to specify all model assumptions and the constraints needed to make the model identifiable.

   Solution:

$$\Omega : \begin{cases} Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, & i = 1, 2;\ j = 1, 2, 3; \\ \mathcal{E}(e_{ijk}) = 0, & k = 1, \dots, 9 \\ \mathrm{Cov}(e_{ijk}, e_{lmr}) = \begin{cases} \sigma^2, & (i,j,k) = (l,m,r) \\ 0, & (i,j,k) \neq (l,m,r) \end{cases} & \end{cases}$$

$$\sum_{i=1}^{2} \alpha_i = 0; \quad \sum_{j=1}^{3} \beta_j = 0; \quad \sum_{i=1}^{2} \gamma_{ij} = 0 \text{ for } j = 1, 2, 3; \quad \sum_{j=1}^{3} \gamma_{ij} = 0 \text{ for } i = 1, 2.$$

   b. (5 pts.) Under the model you specified in part a, what is the least-squares estimate of the main effect corresponding to wool type B? What is the least-squares estimate of the interaction effect between wool type A and the medium tension setting? Use the information in the table below to perform these computations.

Average rate of warp breaks for each wool type and for each tension level.

	Low Tension	Medium Tension	High Tension	combined
Wool A	44.56	24.00	24.56	31.04
Wool B	28.22	28.78	18.78	25.26
combined	36.39	26.39	21.67	28.15

   Solution:

$$\hat{\alpha}_2 = \overline{Y}_{2\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot} = 25.26 - 28.15 = -2.89$$

$$\hat{\gamma}_{12} = \overline{Y}_{12\cdot} - \overline{Y}_{1\cdot\cdot} - \overline{Y}_{\cdot 2\cdot} + \overline{Y}_{\cdot\cdot\cdot} = 24.00 - 31.04 - 26.39 + 28.15 = -5.28$$

c. (8 pts.) After fitting the ANOVA model to the data, the researcher obtained an ANOVA table. This table is partially presented below. Provide the missing information. Round to two decimal places if necessary.

Source	Sum of Squares	Degrees of Freedom	Mean Square	F statistic
Wool Type	451	1	451	$451/119.69 = 3.77$
Tension Setting	$9233 - 5745 - 1003 - 451 = 2034$	$J - 1 = 2$	1017	$1017/119.7 = 8.50$
Interaction	1003	2	$1003/2 = 501.5$	$501.5/119.69 = 4.19$
Residuals	5745	48	$5745/48 = 119.69$	
Total	9233	$n - 1 = 53$		

d. (7 pts.) Let the significance level $\alpha = 0.05$, and assume the model error components are normally distributed. Based on the completed ANOVA table in part c, and the selected critical values of the $\mathcal{F}$-distribution given below, is the researcher justified in rejecting the full model in favor of the additive model? Support your answer.

$\mathcal{F}_{1,48;0.95} = 4.04$, $\mathcal{F}_{1,48;0.975} = 5.35$, $\mathcal{F}_{2,48;0.95} = 3.19$, $\mathcal{F}_{2,48;0.975} = 3.99$, $\mathcal{F}_{1,54;0.95} = 4.02$, $\mathcal{F}_{1,54;0.975} = 5.32$, $\mathcal{F}_{2,54;0.95} = 3.17$, $\mathcal{F}_{2,54;0.975} = 3.95$.

Solution: The additive model is the model with no interactions. The conclusion of the corresponding hypothesis test is:

Reject $H_0 : \gamma_{11} = \gamma_{12} = \gamma_{21} = \gamma_{22} = \gamma_{31} = \gamma_{32} = 0$, since $4.190 > \mathcal{F}_{2,48;0.95} = 3.19$.

That is, there is an interaction effect, so the researcher is not justified in rejecting the full model in favor of the additive model.

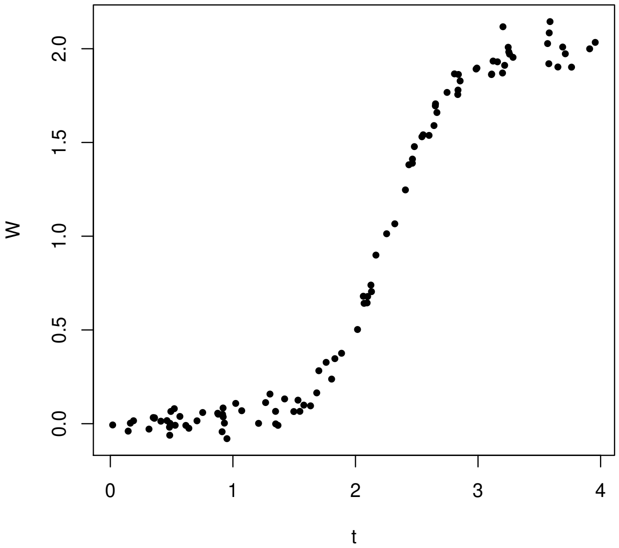
e. (5 pts.) If the researcher fits the additive model, will the values of the F statistics corresponding to “Wool Type” and “Tension Setting” increase, decrease, or remain the same? Support your answer.

Solution: If the researcher fits the additive model, the interaction sum-of-squares, and the degrees of freedom, will be absorbed by the residual sum-of-squares. The mean-square error will increase to $6748/50 = 134.96$. Then the F statistics corresponding to the two effects will **decrease**, since their respective mean squares will be divided by the larger mean-square error.

2. The weight W of an organism at time t is modeled by the function

$$w = f(t;\ \boldsymbol{\theta}) = \frac{\theta_1}{1 + \exp[-(\theta_2 + \theta_3 t)]}\ ,$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$. A researcher obtains data from 100 independent observations $(w_1, t_1), \ldots,$ (w_{100}, t_{100}), which are presented in the following plot:



He wants to estimate $\boldsymbol{\theta}$ by fitting the model

$$W_i = f(t_i;\ \boldsymbol{\theta}) + \varepsilon_i, \qquad i = 1, \ldots, 100$$

to these data, where $\varepsilon_1, \ldots, \varepsilon_{100}$ are independent random errors with common mean zero and common variance σ^2.

a. (5 pts.) In order to compute the least-squares estimate of $\hat{\boldsymbol{\theta}}$ using software, suppose the researcher has chosen a starting value of 1 for θ_1 and 2 for θ_2. Choose an appropriate starting value for θ_3, based on $f(t;\ \boldsymbol{\theta})$ and the approximate value of W at $t = 2$ in the plot.

Solution: At $t = 2$ we have $w = 0.5$. Thus

$$0.5 = \frac{1}{1 + \exp[-(2 + 2\theta_3)]}\ .$$

Solving this equation gives $\theta_3 = -1$ as a starting value.

b. (5 pts.) Using software, the researcher obtains the least-squares estimate

$$\hat{\boldsymbol{\theta}} = (2.002, -8.936, 3.959)\ .$$

Explain how this estimate can be used to construct a consistent estimator of σ^2 (give as much detail as possible).

Solution: A consistent estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{S(\hat{\boldsymbol{\theta}})}{n-p} = \frac{1}{100-3} \sum_{i=1}^{100} \left(w_i - \frac{2.002}{1 + \exp[-(-8.936 + 3.959 t_i)]} \right)^2\ .$$

c. (5 pts.) After fitting the nonlinear model to the data, the researcher obtains the following estimate of $\text{Cov}(\hat{\theta})$:

$$\begin{bmatrix} 0.0002 & 0.0014 & -0.0007 \\ 0.0014 & 0.0528 & -0.0238 \\ -0.0007 & -0.0238 & 0.0110 \end{bmatrix}.$$

Using classical theory based on the asymptotic normality of $\hat{\boldsymbol{\theta}}$, and $t_{97,0.975} = 1.98$, compute a 95% confidence interval for θ_3. Round to three decimal places.

Solution: The 95% confidence interval for θ_3 is $\hat{\theta}_3 \pm t_{97,0.975}\sqrt{\hat{\text{Cov}}(\hat{\theta}_3)}$. From the covariance matrix, we have $\hat{\text{Cov}}(\hat{\theta}_3) = 0.0110$. From part b we know $\hat{\theta}_3 = 3.959$. Thus the 95% confidence interval for θ_3 is (3.751, 4.167).

d. (5 pts.) Based on 1000 samples of the centered residuals, the researcher computes 1000 bootstrap estimates θ_3^* of θ_3. Let $\theta_{3;q}^*$ denote the bootstrap estimate of θ_3 that is in position q when the bootstrap estimates are ordered from smallest to largest. If $\theta_{3;26}^* = 3.755$ and $\theta_{3;976}^* = 4.174$, compute the 95% bootstrap confidence interval for θ_3. Round to three decimal places. Compare the bootstrap confidence interval to the classical confidence interval.

Solution: From part b we know $\hat{\theta}_3 = 3.959$. Thus $2\hat{\theta}_3 = 7.918$, so the lower bound for the 95% bootstrap confidence interval is $2\hat{\theta}_3 - \theta_{3;976}^* = 7.918 - 4.174 = 3.744$ and the upper bound for the 95% bootstrap confidence interval is $2\hat{\theta}_3 - \theta_{3;26}^* = 7.918 - 3.755 = 4.163$. That is, the 95% bootstrap confidence interval is $(3.744, 4.163)$. This is just slightly narrower than the classical confidence interval.

3. a. Data were collected from 77 patients in a urology clinic of Stanford Medical School in the US. Of interest is whether the formation of oxalate crystals (and ultimately kidney stones) is related to various urine characteristics: specific gravity, acidity, osmolarity, conductivity, urea concentration and calcium concentration. Logistic regression was performed in `R` using the logit link. The following Analysis of Deviance table was obtained:

	Df	Deviance	Resid. Df	Resid. Dev
NULL			76	105.168
SpecificGrav	1	14.933	75	90.235
Acidity	1	0.072	74	90.163
Osmolarity	1	9.557	73	80.606
Conductivity	1	0.011	72	80.595
UreaConc	1	1.334	71	79.261
CalciumConc	1	21.701	70	57.560

i. (5 pts.) Which **three** urine characteristics have the most significant relationship to the formation of oxalate crystals, based on this table? Explain.
Solution: Calcium concentration, specific gravity and osmolarity, beause these three characteristics contribute the greatest reduction in deviance in the model ($14.933 + 21.701 + 9.557$).

ii. (5 pts.) Based on the information provided, is there evidence of overdispersion in the fitted model? Explain.
Solution: The overdispersion parameter is estimated by $\hat{\sigma}^2 = D/(n-p-1) = 57.56/(77-6-1) = 57.56/70 < 1$. Since the estimate is less than one, there is no evidence of overdispersion.

iii. (5 pts.) The AIC for the fit above is 71.56. Another logistic regression is performed on the same data, using the complementary log link. The AIC for this new fit is 70.11. Based on the AIC, which model is preferred?
Solution: The second model is preferred, since the AIC is lower.

b. (5 pts.) Suppose Y_i is a random variable having the inverse Gaussian distribution with parameter μ_i and constant σ^2. The probability density function of Y_i can be written in the canonical form of the exponential family as

$$f(y;\mu_i,\sigma^2) = \exp\left[\frac{y(1/\mu_i^2)-(2/\mu_i)}{-2\sigma^2} + \left(\frac{\sigma^2 y\log(2\pi\sigma^2 y^3)+1}{-2\sigma^2 y}\right)\right], \quad \text{for } y \geq 0\,.$$

Set $\theta_i = 1/\mu_i^2$. Based on this canonical form, write the function $b(\theta_i)$ in terms of θ_i, and verify that $b'(\theta_i) = \mu_i$.

Solution: Since $\theta_i = 1/\mu_i^2$, we have $\mu_i = 1/\sqrt{\theta}$. Then

$$\frac{2}{\mu_i} = \frac{2}{1/\sqrt{\theta_i}} = 2\sqrt{\theta_i} = b(\theta_i),$$

so

$$b'(\theta_i) = 2\cdot\frac{1}{2}(\theta_i)^{-1/2} = \frac{1}{\sqrt{\theta_i}} = \mu_i\,.$$

c. (5 pts.) Consider the following five possible link functions:

$$g(\mu_i) = \frac{1}{\mu_i^2}, \qquad g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right), \qquad g(\mu_i) = \mu_i,$$

$$g(\mu_i) = \log(\mu_i), \qquad g(\mu_i) = \Phi^{-1}(\mu_i).$$

For each of the following data descriptions, state which of the above functions is the canonical link that corresponds to the appropriate generalized linear model:

i. $Y_1, \ldots, Y_n$ represent independent normally-distributed random variables with common variance whose association with the covariate is assumed to be linear.
Solution: This is simple linear regression, so the canonical link function is the identity function: $g(\mu_i) = \mu_i$.

ii. $Y_1, \ldots, Y_n$ represent the proportions of workers at n independent factories who are injured within a year.
Solution: For proportions the canonical link is the logit link: $g(\mu_i) = \log(\mu_i/(1 - \mu_i))$.

iii. $Y_1, \ldots, Y_n$ represent independent inverse Gaussian data.
Solution: The canonical link function requires $g(\mu_i) = \theta_i$. From the canonical form of the inverse Gaussian above, $\theta_i = 1/\mu_i^2$, so $g(\mu_i) = 1/\mu_i^2$.

iv. $Y_1, \ldots, Y_n$ represent the numbers of eggs laid by n independent fruit flies within a month.
Solution: For counts the canonical link is the log link: $g(\mu_i) = \log(\mu_i)$.

4. Let $\{Z_t\}$ denote white noise with variance σ^2.

   a. (8 pts.) Show that the time series $\{X_t\}$ given by $X_t = \sum_{i=1}^{t} Z_i$ is *not* weakly stationary.
   Solution: First,

$$\mathcal{E}(X_t) = \mathcal{E}\left(\sum_{i=1}^{t} Z_i\right) = \sum_{i=1}^{t} \mathcal{E}(Z_i) = 0$$

   for all t. Next,

$$\begin{aligned}
\mathcal{E}(X_t X_{t+h}) &= \mathrm{Cov}(X_t, X_{t+h}) + \mathcal{E}(X_t)\mathcal{E}(X_{t+h}) \\
&= \mathrm{Cov}\left(\sum_{i=1}^{t} Z_i, \sum_{j=1}^{t+h} Z_j\right) + 0 \\
&= \sum_{i=1}^{t}\sum_{j=1}^{t+h} \mathrm{Cov}(Z_i, Z_j) \\
&= \sum_{i=1}^{t} \mathrm{Var}(Z_t) \\
&= t\sigma^2
\end{aligned}$$

   for all t. Since $\mathcal{E}(X_t X_{t+h})$ depends on t, $\{X_t\}$ is not weakly stationary.

   b. Let $\{X_t\}$ be the AR(1) time series given by $X_t = -0.3X_{t-1} + Z_t$.

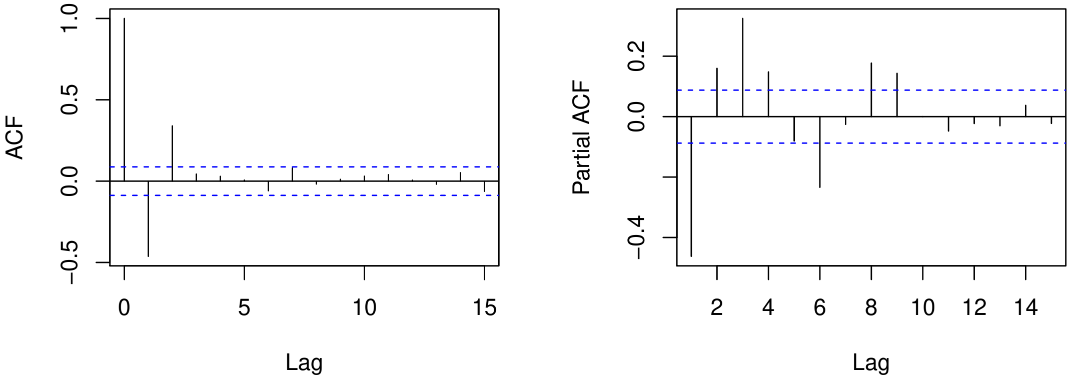
      i. (4 pts.) Find $\rho_X(2)$.
      Solution: $\rho_X(2) = \alpha^{|2|} = (-0.3)^2 = 0.09$.

      ii. (4 pts.) Consider the time series $\{Y_t\}$ given by $Y_t = \nabla_2 X_t$. Show that $\{Y_t\}$ is an ARMA(p, q) time series, and identify the values of the coefficients p and q.
      Solution:

$$\begin{aligned}
Y_t &= \nabla_2 X_t = X_t - X_{t-2} \\
&= -0.3X_{t-1} + Z_t - (-0.3X_{t-3} + Z_{t-2}) \\
&= -0.3X_{t-1} + 0.3X_{t-3} + Z_t - Z_{t-2}
\end{aligned}$$

      which is an ARMA(3,2) time series.

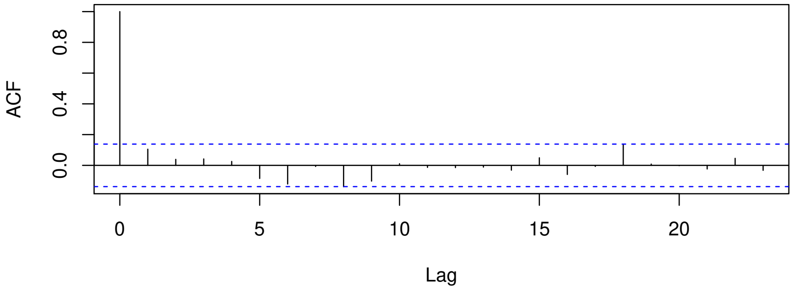
c. (5 pts.) For an observed time series, the correlograms of its sample ACF and its sample PACF are given below.



What ARMA(p, q) model would be most appropriate to model this time series? Why?

Solution: Since the sample autocorrelation function exhibits a dramatic cutoff after lag 2, while the sample partial autocorrelation function does not exhibit a dramatic cutoff, the most appropriate ARMA(p, q) model would be an ARMA(0,2) model, which is an MA(2) model.

d. (4 pts.) After fitting an ARMA(p, q) time series model to a data set, a correlogram of the sample ACF of the residuals is obtained:



What conclusion about the goodness of fit can be made from this plot? Defend your answer.

Solution: The autocorrelation among the residuals is very small, so that the residuals can be viewed as manifestations of white noise. This implies that the fit is good.