Statistics 601 **November 13, 2018**

Assignment 5 — Due Nov 22, 2018

1. Eight isolates of rose blackspot fungus (from eight different areas) were grown for 20 days at seven different temperatures ranging from 55^oF to 85^oF. We denote by `Temp` the categorical variable that has seven temperature levels, and by `Isolate` the categorical variable that has eight levels. The log weights in mg are reported in the file `fungus.txt`.

   (a) Consider an additive model $Y \sim \texttt{Isolate} + \texttt{Temp}$. Complete the decomposition of the total sum of squares:

Source	SS	df	MS	F	p-value
Isolate		7			0.99
Temp		7			
Resid	3.17	41			

   Here MS denotes the mean square, and F is the relevant F-ratio. Based on the above results, can we drop the covariate `Isolate` from the model?

   (b) Now consider a linear model where the covariates are polynomials of temperature. Specifically, let P_r denote the subspace of polynomials of degree r or less in temperature, i.e. $P_r = \text{Span}(1, t, \ldots, t^r)$, where $t \in \mathbb{R}^n$ is the vector of raw temperatures and $t^2 \in \mathbb{R}^n$ is the vector of squared temperatures, etc.

   Complete the following partial sums of squares in an ANOVA table:

Source	SS	df	MS	F	p-value
$P_1 \vert 1$	0.0117	1	0.0117		
$P_2 \vert P_1$					
$P_3 \vert P_2$					
$P_4 \vert P_3$					
Temp$\vert P_4$					
Temp$\vert P_3$	0.0439	3			

   Hint: Note that P_1 is the one-dimensional subspace of constant function, and $1 \subset P_1 \subset \cdots \subset P_6 = \text{Span}(1, \ldots, t^6)$. Let $\mathcal{X} \subset \mathcal{X}' \subset \mathbb{R}^n$ denote two subspaces and $\boldsymbol{Y} \in \mathbb{R}^n$ the response vector. Then the sum of squares associated with $\mathcal{X}$ is the squared length of the projection, $\|\mathcal{P}_{\mathcal{X}}\boldsymbol{Y}\|^2$, and the sum of squares associated with $\mathcal{X}'/\mathcal{X}$ is the additional squared length $\|P_{\mathcal{X}'}\boldsymbol{Y}\|^2 - \|P_{\mathcal{X}}\boldsymbol{Y}\|^2$.

   (c) What does the preceding table tell you about the effect of temperature on growth? Estimate the temperature at which the growth rate is a maximum.

2. `online submission` A local health clinic sent fliers to its clients to encourage everyone, but especially older persons at high risk of complications, to get a flu shot in time for protection against an expected flu epidemic. In a pilot follow-up study, 159 clients were randomly selected and asked whether they actually received a flu shot. A client who received a flu shot was coded $Y = 1$, and a client who did not receive a flu shot was coded $Y = 0$. In addition, data were collected on their age (X_1) and their health awareness. The latter data were combined into a health awareness index (X_2), for which higher values indicate greater awareness. Also included in the data was client gender, where males were coded $X_3 = 1$ and females were coded $X_3 = 0$. The data are stored in the file `flushot.txt` .

   (a) Fit a multiple logistic regression model with the three explanatory variables by maximum likelihood estimation. Report the summary table from the `R` output. Obtain and interpret the maximum likelihood estimates of $\beta_0, \beta_1, \beta_2$, and β_3. State the fitted logistic response function.

   (b) Obtain $\exp(\beta_1)$, $\exp(\beta_2)$, and $\exp(\beta_3)$ and the respective 95% CI. Interpret the results.

(c) What is the estimated probability that male clients aged 55 with a health awareness index of 60 will receive a flu shot? Provide a 95% CI. Interpret the results.

(d) Perform a Wald test to determine whether X_3, client gender, can be dropped from the regression model.

(e) Perform a likelihood ratio test to determine whether X_3, client gender, can be dropped from the regression model.

(f) Perform stepwise model selection based on AIC. Report the summary table for your final model.

(g) Create an ROC plot for the final model. Interpret the results.

3. In a clinical trial, m subjects are assigned to each of the treatment group and the control group. In the treatment group, y_1 of the m subjects have positive response, while in the control group, y_2 subjects have positive response. We are interested in estimating the treatment effect and providing its confidence interval.

   (a) Present the above problem as a logistic regression problem and give the interpretation of the regression coefficients β.

   (b) Express the likelihood function as a function of β, and derive the MLE $\hat{\beta}$.

   (c) Find the asymptotic variance-covariance matrix of $\hat{\beta}$.

   (d) Suppose that $m = 20$, $y_1 = 12$, and $y_2 = 9$. Give the point estimate and the corresponding 95% confidence interval for the odds ratio of having positive response between the treatment group and the control group. Can you conclude that the treatment is effective?

4. `online submission` Flour beetles *Tribolium castaneum* were sprayed with one of three insecticides in solution at different doses. The number of insects killed after a six-day period is recorded below:

	Deposit of insecticide (mg/10 cm^2)					
Insecticide	2.00	2.64	3.48	4.59	6.06	8.00
DDT	3/50	5/49	19/47	19/38	24/49	35/50
γ-BHC	2/50	14/49	20/50	27/50	41/50	40/50
DDT+γ-BHC	28/50	37/50	46/50	48/50	48/50	50/50

   (a) Investigate graphically the relationship between the dose, either in original units or in log units, and the kill rate.

   (b) On the graph for part (a), plot the linear logistic fitted curve for each of the insecticides plus the combination.

   (c) Consider the two models, one in which the relationship is described by three parallel straight lines in the log dose and and one in which the three lines are straight but not parallel. Assess the evidence against the hypothesis of parallelism.

   (d) Let `chem` denote a 3-level categorical factor, and let `ldose` be the log dose. Explain the relationship between the regression coefficients in the model formulae `chem + ldose` and `chem + ldose - 1`. Explain the relationship between the two covariance matrices.

   (e) On the assumption that three parallel straight lines suffice, estimate the potency of the combination relative to each of the components. Obtain a 90% confidence interval for each of these relative potencies.

   (f) Check to see if one of the alternative link functions probit, c-log log or log log, gives an appreciably better fit. Give the answer to part (e) for the c-log log model.

   (g) Under the linear logistic model, estimate the combination dose required to give a 99% kill rate, and obtain a 90% confidence interval for this dose.

   (h) Give a brief summary of your conclusions regarding the effectiveness of these three insecticides.