BST 140.752
Problem Set 4

1 Residuals

1. Consider a linear model $Y = X\beta + \delta\Delta + \epsilon$ where δ is a vector with a 1 at position i_0 and 0 elsewhere. Argue the following.

   A. The i_0 residual is 0 for this model.

   B. The fitted value for β using all of the data and this model is equivalent to that using only the data with the i_0 observation deleted.

   C. Argue that the standardized Press residuals are a test statistic for $\Delta = 0$.

2. Consider the residuals for the ordinary linear model. Derive their mean and variance.

3. Carefully write up the proof that relates the Press residuals to the ordinary residuals. Derive the mean and variance/covariance of the Press residuals.

4. Prove the Sherman/Morrison/Woodburry theorem.

5. Prove that the hat matrix diagonals are between 0 and 1.

6. Why are the studentized residuals not exactly distributed as t statistics?

2 Inference under incorrectly specified models

For all of this section, let Model 1 be $Y = X_1\beta_1 + \epsilon$ and Model 2 be $Y = X_1\beta_1 + X_2\beta_2 + \tilde{\epsilon}$.

1. Suppose that Model 1 is fit while Model 2 represents the actual truth. Give the bias and variance of β_1. Give the expected value of S^2.

2. Suppose that Model 2 is fit while Model 1 is true. Give the bias and variance of the estimted β. Give the expected value of S^2.

3 GLMs

1. Calculate the mean and variance of a random variable from an exponential family using the cumulant generating function.

2. Show that when using a canonical link function, the Fisher scoring and Newton Raphson algorithms for finding glm MLEs are identical.

3. Show that, using the notation from class, for known ϕ and a canonical link, the sufficient statistics for a glm are $X^t y$ where t denotes a transpose.

4. Suppose that y_i is Poisson with $g(\mu_i) = \alpha + \beta x_i$ where g is the link function and $x_i = 1$ for $i = 1, \ldots, n_a$ and $x_i = 0$ for $i = n_a + 1, \ldots, n_a + n_b$. That is, x_i, is a treatment indicator for two groups, A and B. Show that, regardless of the link function, the fitted means equal the two sample means.

5. Consider the class of *binary* glms where the link function satisfies $g\{\mu(x)\} = \Phi^{-1}\{\mu(x)\} = \alpha + \beta x$ where $\Phi(\cdot)$ is a distribution function and $\mu(x)$ is the Bernoulli mean. Let ϕ be the (assumed continuous) associated density. Show that the x at which $\mu(x) = .5$ is $x = -\alpha/\beta$. Further show that the rate of change of $\mu(x)$ at this point is $\beta\phi(0)$. Illustrate that this is $.25\beta$ for the logit link and $\beta/\sqrt{2\pi}$ for the probit link.

4 Coding and data analysis exercises

1. Consider the sleep data from the previous homework.

   A. Consider the model fit from the previous homework. Write a program to grab the hat diagnals as well as use R's lm to obtain them directly. Look at the influence of various data points.

   B. Consider the model fit from the previous homework. Write a program to grab the residuals and Press residuals. Investigate these residuals in the context of this model.

2. Consider the baseball data from the previous exercise.

   A. Consider the model fit from the previous homework. Write a program to grab the hat diagnals as well as use R's lm to obtain them directly. Look at the influence of various data points.

   B. Consider the model fit from the previous homework. Write a program to grab the residuals and Press residuals. Investigate these residuals in the context of this model.

3. Write a function that takes a Y $(n \times 1)$ an X_1 $(n \times 1)$ and an X_2 $(n \times (p-1))$ and produces the partial regression plot of $e_{Y|X_2}$ by $e_{Y|X_2}$.

4. Consider the Challenger O-ring data

5. The table below shows the temperature (Temp in Fahrenheit) and presence (1) or absence of O-ring distress (OD) at the time of flight for the 23 flights before the 1986 Challenger mission disaster.

Temp	OD	Temp	OD	Temp	OD	Temp	OD	Temp	OD
66	0	70	1	69	0	68	0	67	0
72	0	73	0	70	0	57	1	63	1
70	1	78	0	67	0	53	1	67	0
75	0	70	0	81	0	76	0	79	0
75	1	76	0	58	1				

   A. Use logistic regression to model the effect of temperature on the probability of thermal distress. Interpret the results. Plot a figure of the fitted model.

B. Estimate the probability of thermal distress at 31 degrees, which was the temperature at the time of the Challenger flight.
C. Construct a profile likelihood for the effect of temperature on the odds of thermal distress, interpret.
D. Check model fit by comparing this model to a more complex model.