Homework 4

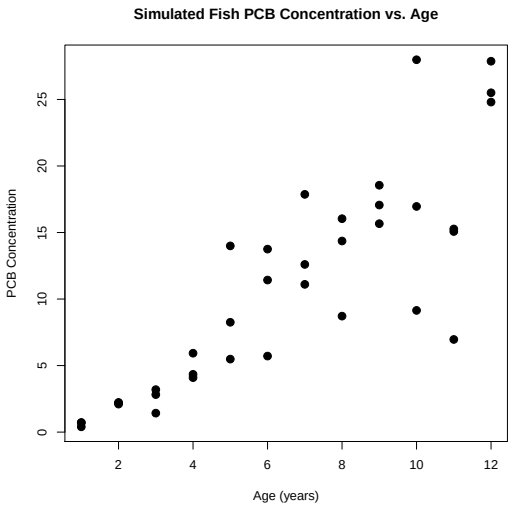
Steve Harms
STAT 520

October 28, 2018

1 Part 1

1.1

Data are generated from the model $Y_{ij} = 0.9x_i^{1.3} + 0.2(0.9x_i^{1.3})^{1.25}\epsilon_{ij}$, where $\epsilon_{ij} \sim N(0,1)$ and $i = 1, ..., 12$ and $j = 1, 2, 3$. Below is a scatterplot of one realization of the 36 values:



1.2

We can write $\hat{\beta_u}'$ as a vector with two components: $\hat{\beta_u}' = [\hat{\beta}_{u,0}, \hat{\beta}_{u,1}]'$, and write each component's MSE in terms of its bias-variance decomposition:

$$MSE(\hat{\beta}_{u,0}) = E[(\hat{\beta}_{u,0} - \beta_{u,0})^2]$$
$$= (E[\hat{\beta}_{u,0} - \beta_{u,0}])^2 + E[(\hat{\beta}_{u,0} - E[\hat{\beta}_{u,0}])^2]$$

and

$$MSE(\hat{\beta}_{u,1}) = E[(\hat{\beta}_{u,1} - \beta_{u,1})^2]$$
$$= (E[\hat{\beta}_{u,1} - \beta_{u,1}])^2 + E[\hat{\beta}_{u,1} - E[\hat{\beta}_{u,1}])^2]$$

1.3

The Monte Carlo estimates for the MSEs of the two parameters above are:

$$MSE_M(\hat{\beta}_{u,0}) = b_M(\hat{\beta}_{u,0})^2 + V_M(\hat{\beta}_{u,0})$$
$$= [M^{-1}\sum_{m=1}^{M}(\hat{\beta}_{u,0}^{(m)} - \beta_{u,0})]^2 + M^{-1}\sum_{m=1}^{M}(\hat{\beta}_{u,0}^{(m)} - \hat{\beta}_{u,0})^2$$

and

$$MSE_M(\hat{\beta}_{u,1}) = b_M(\hat{\beta}_{u,1})^2 + V_M(\hat{\beta}_{u,1})$$
$$= [M^{-1}\sum_{m=1}^{M}(\hat{\beta}_{u,1}^{(m)} - \beta_{u,1})]^2 + M^{-1}\sum_{m=1}^{M}(\hat{\beta}_{u,1}^{(m)} - \hat{\beta}_{u,1})^2$$

where $\hat{\beta}_{u,p}$ is the Monte Carlo estimate of the parameter.

1.4

For Monte Carlo sample sizes of M=500 and M=5000, the following estimates were obtained for the 3 models:

M = 500					
Model	**Coefficient**	**Estimate**	**Bias**	**Variance**	**MSE**
U	$\hat{\beta}_{u,0}$	.9029	.0029	.1051	.105088
G	$\hat{\beta}_{g,0}$	.8983	-0.0017	.0934	.093459
T	$\hat{\beta}_{t,0}$	.8969	-0.0031	.0913	.091292
U	$\hat{\beta}_{u,1}$	1.3003	0.0003	.0921	.092135
G	$\hat{\beta}_{g,1}$	1.3017	0.0017	.0837	.083654
T	$\hat{\beta}_{t,1}$	1.3022	0.0022	.0850	.084979
M = 5000					
Model	**Coefficient**	**Estimate**	**Bias**	**Variance**	**MSE**
U	$\hat{\beta}_{u,0}$	.9102	.0102	.09362	.093725
G	$\hat{\beta}_{g,0}$	.9050	.00496	.08604	.0934587
T	$\hat{\beta}_{t,0}$	.9012	.00121	.086254	.086255
U	$\hat{\beta}_{u,1}$	1.2960	$-.0040$	.086024	.086040
G	$\hat{\beta}_{g,1}$	1.2980	0.0017	.082625	.082696
T	$\hat{\beta}_{t,1}$	1.3000	0.0000179	.0841382	.0841382

1.5

If we define the most efficient estimator as the one that has the smallest MSE, we can see from (4) that for both sample sizes and both coefficients, the MSE is largest for model U, thus the estimators of β_u are the least efficient. This relates to the problem in that we have imposed more structure on the error terms in models G and T, so that we can use the estimation of variance parameters to help in our estimation of the coefficients. In models G and T, we have information about β in both the mean and variance. Thus we can gain efficiency in our estimators of β by assuming that the error structure of the model is also a function of the mean, and estimating them through regression.

2 Part II

First, estimate the model. Here is our response variable as Y_i = OXY concentration in gull i, and our covariate is X_i = body mass (in g) of gull i. We can write the distribution Y_i as a generalized linear model with a gamma distribution and log link as:
$f(y_i|\theta) = exp\{\phi\{y_i\theta_i - b(\theta_i)\} + c(y_i, \phi)\}$,
where
$\phi = \alpha^{-1}, \qquad \theta_i = \frac{-\rho}{\alpha},$
$b(\theta_i) = log\left(\frac{-1}{\phi\theta_i}\right), \qquad c(y, \phi) = -log(\Gamma(\alpha)) - log(y_i)$
and our systematic component is $\mu_i = exp(\beta_0 + \beta_1 x_i)$ for a log link.
We then estimate our model parameters $\hat{\beta}_0, \hat{\beta}_1$, and $\hat{\phi}$:

Model Parameter Estimates	
Coefficient	**Estimate**
$\hat{\beta}_0$	0.5553386
$\hat{\beta}_1$	0.0015587
$\hat{\phi}$	1.427939

With these in hand, we construct a parametric bootstrap confidence interval. The quantity of interest is $P(Y \geq 20|X = 1700)$, which can be written in terms of model parameters as $\hat{p*} = 1 - G(20|\hat{\alpha}, \hat{\rho}_i, X = 1700)$ where G is the cdf of a Gamma distribution with shape α and scale ρ_i. We can get $\hat{\alpha}$ and $\hat{\rho}_i$ from our estimated parameters and model definition.

We then simulate B=1000 datasets from the estimated model. For each simulated dataset, we estimate the model and then a bootstrap estimate of the quantity of interest as $\hat{p*}^b$ for that dataset, using the pgamma() function in R. The bounds for a 95% confidence interval can then be found, using a ratio comparison function, as $[\frac{\hat{p*}^2}{\hat{p*}_{[.975]}}, \frac{\hat{p*}^2}{\hat{p*}_{[.025]}}]$. Note that the ratio comparison function enforces that our estimates must be > 0, but does not necessarily enforce that the bound be < 1. For this B, we get estimated quantities as $\hat{p*} = 0.4055$, $\hat{p*}_{[.025]} = 0.3299$, and $\hat{p*}_{[.975]} = 0.4940$. The resulting bootstrap confidence interval is

$\boxed{[0.3330\ ,\ 0.4985\ ]}$