Math 583 Exam 3, Fall 2020 Name_____

1) The table below shows simulation results for bootstrapping OLS (reg) and forward selection (vs) with C_p when $\boldsymbol{\beta} = (1, 1, 0, 0)^T$. The β_i columns give coverage = the proportion of CIs that contained β_i and the average length of the CI. The test is for $H_0 : (\beta_3, \beta_4)^T = \mathbf{0}$ and H_0 is true. The "coverage" is the proportion of times the prediction region method bootstrap test failed to reject H_0. Since 1000 runs were used, a cov in [0.93,0.97] is reasonable for a nominal value of 0.95. Output is given for three different error distributions. If the coverage for both methods ≥ 0.93, the method with the shorter average CI length was more precise. (If one method had coverage ≥ 0.93 and the other had coverage < 0.93, we will say the method with coverage ≥ 0.93 was more precise.)

a) For β_2, β_3, and β_4, which method, forward selection or the OLS full model, was more precise?

Table 1: Bootstrapping Forward Selection, $n = 100, p = 4, \psi = 0.9, B = 1000$

		β_1	β_2	β_3	β_4	test
reg	cov	0.93	0.95	0.95	0.94	0.95
	len	1.266	10.703	10.666	10.650	2.547
vs	cov	0.95	0.93	0.997	0.995	0.989
	len	1.260	8.901	8.986	8.977	2.759
reg	cov	0.94	0.93	0.95	0.94	0.95
	len	0.393	3.285	3.266	3.279	2.475
vs	cov	0.94	0.97	0.998	0.997	0.995
	len	0.394	2.773	2.721	2.733	2.703
reg	cov	0.95	0.94	0.95	0.95	0.95
	len	0.656	5.493	5.465	5.427	2.493
vs	cov	0.93	0.95	0.998	0.998	0.977
	len	0.657	4.599	4.655	4.642	2.783

b) The test "length" is the average length of the interval $[0, D_{(U_B)}] = D_{(U_B)}$ where the test fails to reject H_0 if $D_{\mathbf{0}} \leq D_{(U_B)}$. The OLS full model is asymptotically normal, and hence for large enough n and B the reg len row for the test column should be near $\sqrt{\chi^2_{2,0.95}} = 2.477$.

Were the three values in the test column for reg within 0.1 of 2.477?

```
Poisson Regression   Response = y = number of inedible kernels
Terms         = (oil temp time)
Coefficient Estimates
Label      Estimate        Std. Error     Est/SE    p-value
Constant   8.24392         0.530985       15.526     0.0000
oil       -0.0885108       0.0595464      -1.486     0.1372
temp      -0.346247        0.0604555      -5.727     0.0000
time      -0.0267420       0.00405347     -6.597     0.0000
```

2) The above output has Y = the number of inedible popcorn kernels in a batch of popcorn. The output is from a Poisson regression. Predict $\hat{\mu}(\boldsymbol{x})$ if $oil = x_2 = 3.0$, $temp = x_3 = 6.0$ and $time = x_4 = 90.0$.

```
Logistic Regression Output for Reduced Model,
Coefficient Estimates
Label      Estimate        Std. Error     Est/SE    p-value
Constant  -19.7762         3.73243        -5.298     0.0000
circum     0.0244688       0.0111243       2.200     0.0278
length     0.0371472       0.0340610       1.091     0.2754

Number of cases:  267 Degrees of freedom: 264
Deviance:                 318.052
```

3) Consider estimating the proportion of males by measuring the circumference and the length of the head with the above output. Predict $\hat{\rho}(\boldsymbol{x})$ if circumference $= x_2 = 550.0$ and length $= x_3 = 200.0$.

4) This problem used the OLS full model residual bootstrap on a model using x_2, x_3, x_4, and a constant x_1.

	Estimate	Std.Err	95% shorth CI
Intercept	1.0060	0.0102	[0.9856, 1.0277]
x2	1.0159	0.0838	[0.8320, 1.1753]
x3	0.9109	0.0835	[0.7412, 1.0680]
x4	0.0701	0.0824	[-0.1232,0.2299]

a) Give the shorth 95% CI for β_4.

b) Compute the standard 95% CI for β_4.

c) Which variables, if any, are needed in the MLR model given that the other variables are in the model?

d) Find $\hat{Y} = ESP$ if $x_2 = 0.95, x_3 = -1.05$, and $x_4 = 0.13$.

5) The GAM analog to the binary logistic regression model is $Y_1, ..., Y_n$ are independent with

$$Y|AP \sim \text{binomial}(1, \rho(\text{AP})) \text{ where } \text{P(success|AP)} = \rho(\text{AP}) = \frac{\exp(\text{AP})}{1 + \exp(\text{AP})}.$$

Then

$$\hat{\rho}(\boldsymbol{x}) = \hat{\rho}(AP) = \rho(EAP) = \frac{\exp(EAP)}{1 + \exp(EAP)}.$$

For the GAM logistic regression response plot i) what is $\hat{\rho}(\boldsymbol{x})$ if $EAP = 0$?

ii) What is $\hat{\rho}(\boldsymbol{x})$ if $EAP = 5$?

6) The GAM analog to the Poisson regression model is $Y_1, ..., Y_n$ are independent with

$$Y|AP \sim \text{Poisson}(\exp(\text{AP})).$$

Then $\hat{Y} = E(Y|\boldsymbol{x}) = E(Y|AP) = \mu(\boldsymbol{x}) = \exp(AP)$ and

$$\hat{\mu}(\boldsymbol{x}) = \hat{\mu}(AP) = \exp(EAP).$$

If $\boldsymbol{x}$ is such that EAP = 0, find $\hat{\mu}(\boldsymbol{x})$.

7) Consider choosing $\hat{\boldsymbol{\eta}}$ to minimize the criterion

$$Q(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p-1}|\eta_i|^j$$

where $\lambda_{1,n} \geq 0$, $a > 0$, and $j > 0$ are known constants. Consider the regression methods OLS, forward selection, lasso, ridge regression, and lasso variable selection.
a) Which method corresponds to $j = 1$?

b) Which method corresponds to $j = 2$?

c) Which method corresponds to $\lambda_{1,n} = 0$?

8) The response plots for good multiple linear regression estimators (such as ordinary least squares, lasso, forward selection, backward elimination, ridge regression and elastic net) look like the response plot for the additive error regression model $Y = m(\boldsymbol{x}) + e$. Sketch the response plot.

9) The output is for a data set on 82 mussels sampled off the coast of New Zealand. Let the response variable be the logarithm $\log(M)$ of the *muscle mass*, and the predictors are the *length* L and *height* H of the shell in mm, the logarithm $\log(W)$ of the *shell width* W, the logarithm $\log(S)$ of the *shell mass* S, and a constant.

```
      large sample full model inference
      Est.    SE  t   Pr(>|t|)   nparboot      resboot
int -1.249 0.838 -1.49 0.14 [-2.93,-0.093][-3.045,0.473]
L   -0.001 0.002 -0.28 0.78 [-0.005,0.003][-0.005,0.004]
logW 0.130 0.374  0.35 0.73 [-0.457,0.829][-0.703,0.890]
H    0.008 0.005  1.50 0.14 [-0.002,0.018][-0.003,0.016]
logS 0.640 0.169  3.80 0.00 [ 0.244,1.040][ 0.336,1.012]
output and shorth intervals for the min Cp submodel FS
      Est.     SE     95% shorth CI   95% shorth CI
int   -0.9573  0.1519 [-3.294, 0.495] [-2.769, 0.460]
L      0              [-0.005, 0.004] [-0.004, 0.004]
logW   0              [ 0.000, 1.024] [-0.595, 0.869]
H      0.0072  0.0047 [ 0.000, 0.016] [ 0.000, 0.016]
logS   0.6530  0.1160 [ 0.322, 0.901] [ 0.324, 0.913]
               for forward selection for all subsets
```

The minimum C_p model from all subsets variable selection and forward selection both used a constant, H, and $\log(S)$. The shorth(c) nominal 95% confidence intervals for β_i using the residual bootstrap are shown. What is $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0}$?