Final Exam

Name:

1. (4 points) If you believe that only a small number of variables are important, which of the following regression approaches are appropriate ways to select the important variables:

   i. Ridge regression
   ii. Lasso regression
   iii. p-values.
   iv. Forward stepwise regression

2. (4 points) The R^2 in the linear regression $Y = \beta_0 + \beta_1 X + \epsilon$ from data $(y_1, x_1), \ldots, (y_n, x_n)$ represents which of the following:

   i. the fraction of variance Y explained by $\beta_0 + \beta_1 X$
   ii. the expected prediction error for new data $(y_{n+1}, x_{n+1}), \ldots, (y_{2n}, x_{2n})$
   iii. the square of the correlation between Y and $\hat{\beta}_0 + \hat{\beta}_1 X$

3. (4 points) Identify and describe at least one problem or confounding factor in the following experimental design.
   Arabadopsis thaliana is a flowering plant native to Europe. Strains in northern Europe tend to have an earlier flowering time than strains in southern Europe. To identify epigenetic factors that may contribute to the difference, researchers took samples from Tunisia and Norway. The samples were processed for bisulfite sequencing on site and then sent to a central sequencing center for processing and analysis.

4. (8 points) Suppose we have data on a response variable Y and three predictor variables X_1 and X_2. We fit the models on all possible subsets. Circle the model that will be selected according the following criteria.

criterion	X_1	X_2	X_1, X_2
AIC	126.86	106.57	106.59
BIC	129.85	109.56	110.57
adj R^2	-0.023	0.63	0.64
10-fold CV	402.56	131.34	151.99

5. Using a subsample of the Structural Survey of Wages for Spain in 2006, the following model is estimated to explain wages (in tens of thousands of dollars) as a function of education (in years:

Table 2: Fitting linear model: wages ~ education

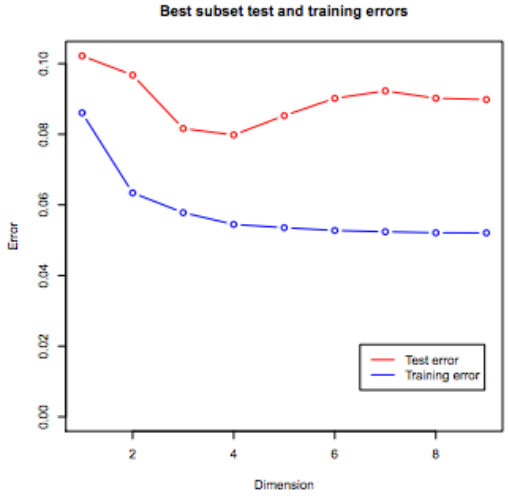
	Estimate	Std. Error	t value	Pr(>|t|)
education	0.0527	0.0121	5.762	1.33e-05
(Intercept)	1.918	0.7155	5.268	7.35e-03
R^2	0.2445			

a. (4 points) What is the interpretation of the education coefficient?

b. (4 points) Assuming the linear model is true, what is the expected wages for someone with 16 years of education?

c. (4 points) What is percentage of variance in wages explained by education?

6. Suppose we have 9 predictor variables and 1 response variable. We run best subset regression and obtain the following estimates of training and test error for the number of predictor variables ranging from 0 to 9 (labeled dimension).



a. (4 points) What is the optimal number of predictors according to the training error?

b. (4 points) What is the optimal number of predictors according to the test error?

7. In species sampling experiments, the number of individuals captured from each species is thought to be Poisson distributed with pmf $f(x) = \Pr(X = x) = \lambda^x e^{-\lambda}/x!$. Unfortunately, species with no captured individuals are not observed. In this case, the number of individuals captured for each observed species (those with at least one observed individual) follows a zero-truncated Poisson distribution with pmf $f(x) = \Pr(X = x|X > 0)$. Suppose we observed 10 species with counts $x_1, \ldots, x_{10}$.

   a. (4 points) If $\lambda = 2$, what is the probability $X = 2$?

   b. (4 points) Write out the log-likelihood function.

   c. (4 points) Set up the maximum likelihood equations. Do not solve.

   d. (4 points) Calculate the expectation and variance of a zero-truncted Poisson distribution with parameter λ.

   e. (4 points) Set up the Method of Moments equations. Do not solve.

   f. (3 points) Suppose we fit a zero-truncated Poisson and a zero truncated Negative Binomial distribution and obtain the following AIC values: 123.2 and 127.5953, respectively. Which model would be chosen by the AIC criterion?

8. A new internet provider claims that using their system you can get transmission speeds of 1 gigabits per second. To test that consumers are getting at least the advertised speed, regulators visit 100 houses and measure speeds of 0.956 gb/s with an estimated standard deviation 0.213 gb/s.

   a. (3 points) State the null and alternative hypotheses.

   b. (5 points) Construct a 95% confidence interval for the average transmission speed.

   c. (5 points) Can you reject the null hypothesis at the 99% confidence level? Compute the p-value or explain why or why not you can reject the null.

   d. (5 points) Suppose the true speed is Normal with average speed of 0.96 gb/s and standard deviation of 0.2 gb/s. What is the approximate power of the test assuming a null distribution of $N(1, 0.2)$ at the 95% confidence level and $n = 100$?

9. To bid for government contracts, businesses submit bids and the government selects the lowest bid that satisfies the criteria. Suppose that bids are exponentially distributed with average bid equal to $ 10 million.

   a. (4 points) What is the probability the first bid is below $ 3 million?

   b. (5 points) If there are 10 bids, what is the expected value of the lowest bid?

   c. (5 points) If the government decides to take bids one by one until one is below $ 3 million. How many bids will the government have to wait on average?

   d. (5 points) If you want to enter the auction (for 11 bids total, 1 being yours and 10 other bids), what bid would you have to submit to ensure a 90% probability of having the lowest bid?

Binomial(n, p):
pmf: $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$
expectation: np
variance: $np(1-p)$

Poisson(λ):
pmf: $f(x) = \lambda^x e^{-\lambda}/x!$
expectation: λ
variance: λ

Geometric(p):
pmf: $f(x) = (1-p)^x p$
expectation: $\frac{1}{p} - 1$
variance: $\frac{1-p}{p^2}$

Negative Binomial(r, p):
pmf: $f(x) = \binom{x+r-1}{x} p^x (1-p)^r$
expectation: $\frac{pr}{1-p}$
variance: $\frac{pr}{(1-p)^2}$

Hypergeometric(N_1, N_2, n):
pmf: $f(x) = \frac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N_1+N_2}{n}}$
expecation: $nN_1/(N_1+N_2)$
variance: $n\frac{N_1}{N_1+N_2}\frac{N_2}{N_1+N_2}\frac{N_1+N_2-n}{N_1+N_2-1}$

Uniform(a, b):
pdf: $f(x) = 1/(b-a)$
cdf: $F(x) = \begin{cases} 0 & x \le a \\ (x-a)/(b-a) & a < x < b \\ 1 & x > b \end{cases}$
expectation: $(a+b)/2$
variance: $(b-a)^2/12$

Exponential(λ):
pdf: $f(x) = \lambda e^{-\lambda x}$, $x > 0$
cdf: $F(x) = 1 - e^{-\lambda x}$, $x > 0$
expectation: λ^{-1}
variance: λ^{-2}

Gamma(k, θ):
pdf: $f(x) = x^{k-1} e^{-x/\theta} / \big(\Gamma(k)\theta^k\big)$, $x > 0$
expectation: $k\theta$
variance: $k\theta^2$

Beta(α, β):
pdf: $f(x) = x^{\alpha-1}(1-x)^{\beta-1}/\mathcal{B}(\alpha, \beta)$, $0 < x < 1$
expectation: $\alpha/(\alpha+\beta)$
variance: $\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$

Normal(μ, σ^2):
pdf: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/\sigma^2}$
expectation: μ
variance: σ^2

Standard normal table for z_α (or $z_{1-\alpha}$ depending on the notation)

```
##   alpha         z
## 1 0.200 0.8416212
## 2 0.100 1.2815516
## 3 0.050 1.6448536
## 4 0.025 1.9599640
## 5 0.010 2.3263479
## 6 0.005 2.5758293
```