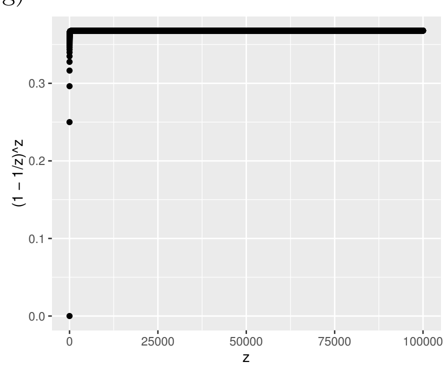
Math 650 HW 3

Mose Wintner

1.

$$\begin{aligned}
\varphi(\alpha) = \mathrm{Var}(\alpha X + (1-\alpha)Y) &= \alpha^2\sigma_X^2 + (1-\alpha)^2\sigma_Y^2 + 2\alpha(1-\alpha)\sigma_{XY} \\
0 = \frac{\partial\varphi}{\partial\alpha} &= 2\alpha\sigma_X^2 - 2(1-\alpha)\sigma_Y^2 + [2(1-\alpha) - 2\alpha)]\sigma_{XY} \\
0 &= \alpha(\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}) - \sigma_Y^2 + \sigma_{XY} \\
\alpha &= \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}
\end{aligned}$$

2. a) It's 1-1/n. Everything is chosen with replacement so a given observation is just as likely to be chosen at any draw.
b) $(1 - 1/n)$.
c) Each of the n draws is taken with replacement, so each draw is independent and equally likely to give the j^{th} observation, so the probability is $(1 - 1/n)^n$.
d) $1 - 0.8^5 \approx .672$
e) $1 - 0.99^{100} \approx .634$
f) $1 - 0.9999^{10000} \approx .632139$
g)



The quantity converges very quickly for $z \to \infty$.
h) It comes out near to .634.

3. a) Fix an integer k and several models, e.g. a vector of possible hyperparameters of a single model. Split the data into k (approximately) equally-sized parts. For each possible model, train it on $k - 1$ of the parts and compute the test error of the model on the left out part. Aggregate these k test errors in the overall MSE and choose the hyperparameter λ for the model that minimized the MSE.
b) k-fold cross validation allows us to use the entire dataset to train our model, which is an advantage

over the validation set approach. The validation set approach is superior to k-fold cross validation in that it's very simple to implement, but the MSE on a validation set as a function of the data has a higher variance than for cross-validation, since the latter is based on bootstrapping. LOOCV is nice since it involves no randomness–its value is always the same on a fixed dataset. However, it is computationally expensive for sophisticated models. Also the variance of the test error yielded by LOOCV is high since the bias-variance tradeoff rate is much heavier for the variance term, meaning that in allowing a small amount of bias, variance drops off significantly, cf. section 5.1.4. in the text.

4. Choose a validation method like, say, k-fold cross validation, so the data is indexed by a partition of $\{1, \ldots, n\}$ into k sets $J_1 \ldots J_k$. Let $\hat{y}_j^{(i)}$ be the estimate of the j^{th} observation computed by the model trained on the entire dataset minus those indexed by J_i. The estimated MSE

$$\frac{1}{k}\sum_{i=1}^{k}\frac{1}{|J_i|}\sum_{j\in J_i}(\hat{y}_j^{(i)} - y_j)^2$$

computed by the validation will approximate the variance of the prediction; its square root will approximate the standard deviation.

5. See attached .r file. The results obtained for each validation set are very similar, with accuracy near .975. Adding the student dummy variable doesn't significantly decrease test error rate.

6-8. See attached .r file.