STAT571HW1

Problem 1

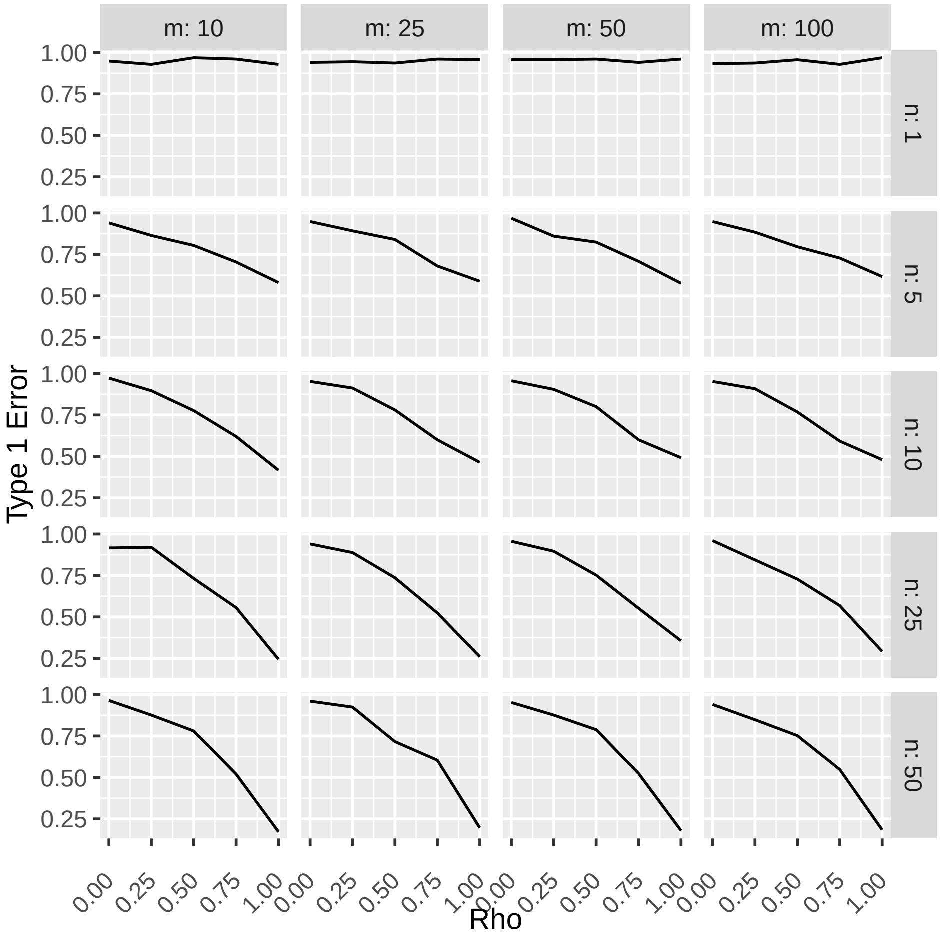
In this simulation we investigate the effect of correlated observations on the performance of linear regression. Specifically, we evaluate how different degrees of correlations effect the type-1 error, bias and confidence interval coverage rate of linear regression. The simulation design is as follows.

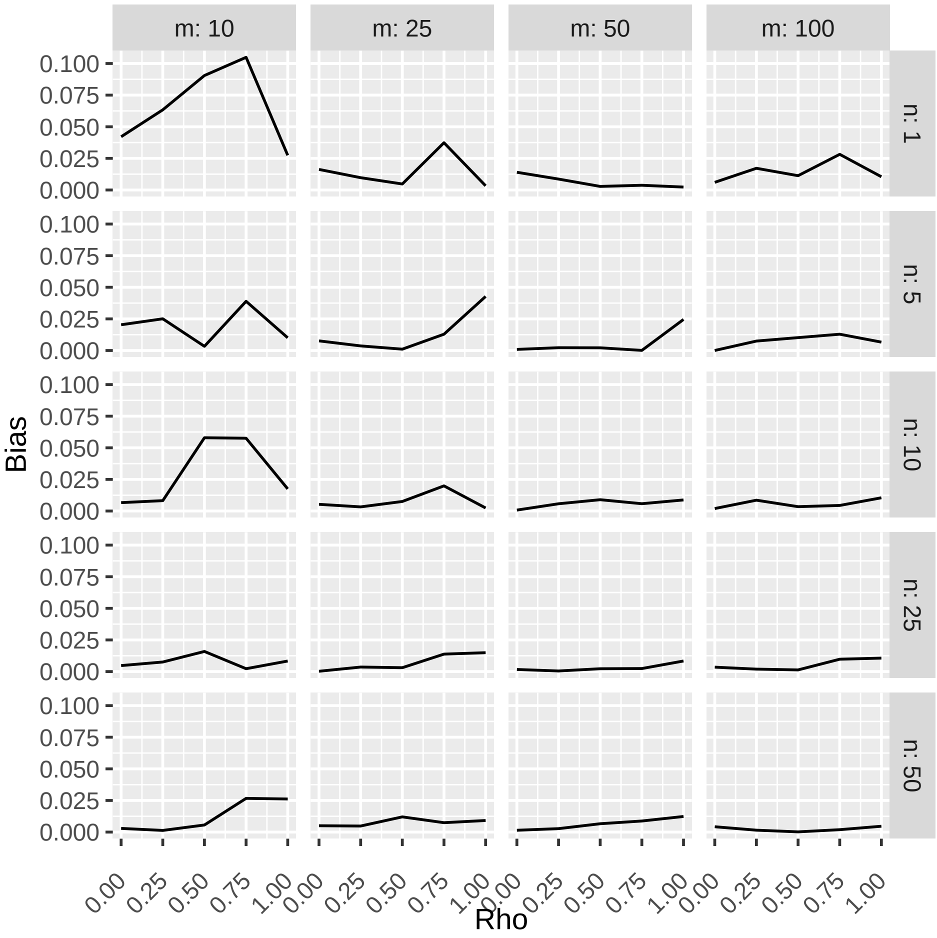
We consider m individuals each with n observations. The individuals are independent of one another and identically distributed; however, for each individual, the observations are correlated. We represent each individual by the data-structure $O_i = (X_i, Y_{i1}, Y_{i2}, \dots, Y_{in})$ where X_i is a baseline covariate that is uniformly distributed in $[-1, 1]$ and $(Y_{i1}, \dots, Y_{in})$ are the n individual-specific outcome observations that are drawn from a multivariate normal distribution with mean $1 + \beta \cdot X$ where β is a to-be-specified coefficient of interest and covariance matrix Σ with $\Sigma_{ij} = 1.5 \cdot \rho^{|i-j|}$ where ρ is a correlation coefficient.

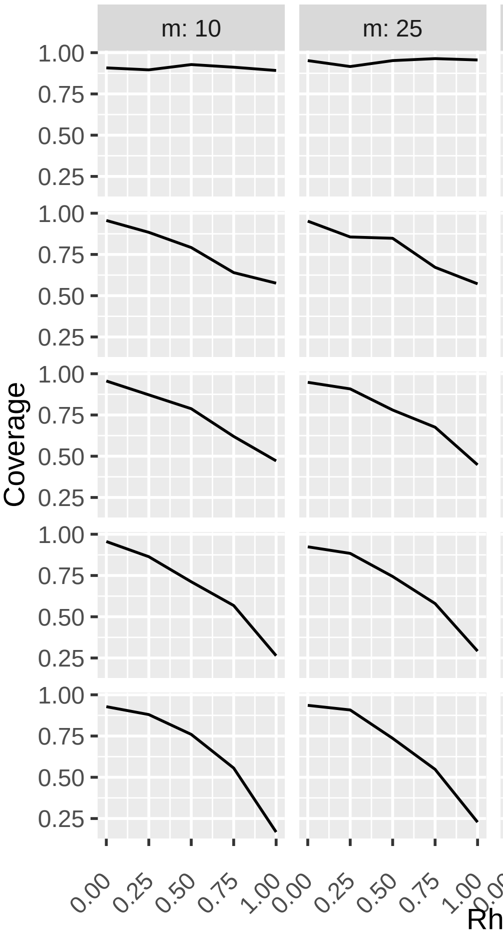
To evaluate the type-1 error, we take $\beta = 0$ (the null model) and do the following. For all combinations of the following values of m, n, ρ, we simulate ($K = 100$ times) m iid realizations $[O_i : i = 1, \dots, m]$ and compute the linear regression of the outcomes Y_{ij} on the X_i using the model $E[Y_{ij} \mid X_i] = \beta_0 + \beta_1 X_i$. For each simulation iteration, we obtain a p-value for the coefficient β_1 and check whether it is below the cutoff $\alpha = 0.05$. We record the proportion of times this occurs, which is an estimate of the type-1 error of the linear regression estimator.

To evaluate confidence interval coverage, we run simulations in the same way as the previous paragraph but with $\beta = 1$. We compute 0.95% confidence intervals for β_1 in the model $E[Y_{ij} \mid X_i] = \beta_0 + \beta_1 X_i$ and record the proportion of times that the confidence interval contains $\beta = 1$ and the average absolute difference between the estimated and true coefficients, which are estimates of the confidence interval coverage and bias of the linear regression estimator.

The results of both simulations are graphically displayed below. As expected, we see that the type-1 error and confidence interval coverage becomes worse as the correlation between individual-specific outcomes increase. We achieve the desired type-1 error and coverage rate when there is no correlation ($\rho = 0$) and achieve terrible performance when there is near-perfect correlation ($\rho \approx 1$). This makes sense because when there is no correlation the outcomes are all independent and the standard theory applies. On the other hand, where there is perfect correlation, the individual-specific outcomes are all identical and provide no additional information relative to observing a single outcome. Thus, we artificially inflate the sample size, which leads to incorrect inference. We see that the bias increases as the correlation increases, which occurs because the effective sample size decreases, but the bias still converges to 0 as the number of individuals increases, which makes sense because linear regression is still consistent when there is correlation.


m: 10
m: 25
m: 50
m: 100
n: 1
n: 5
n: 10
n: 25
n: 50
Type 1 Error
Rho
1.00
0.75
0.50
0.25
0.00
0.25
0.50
0.75
1.00





Problem 2

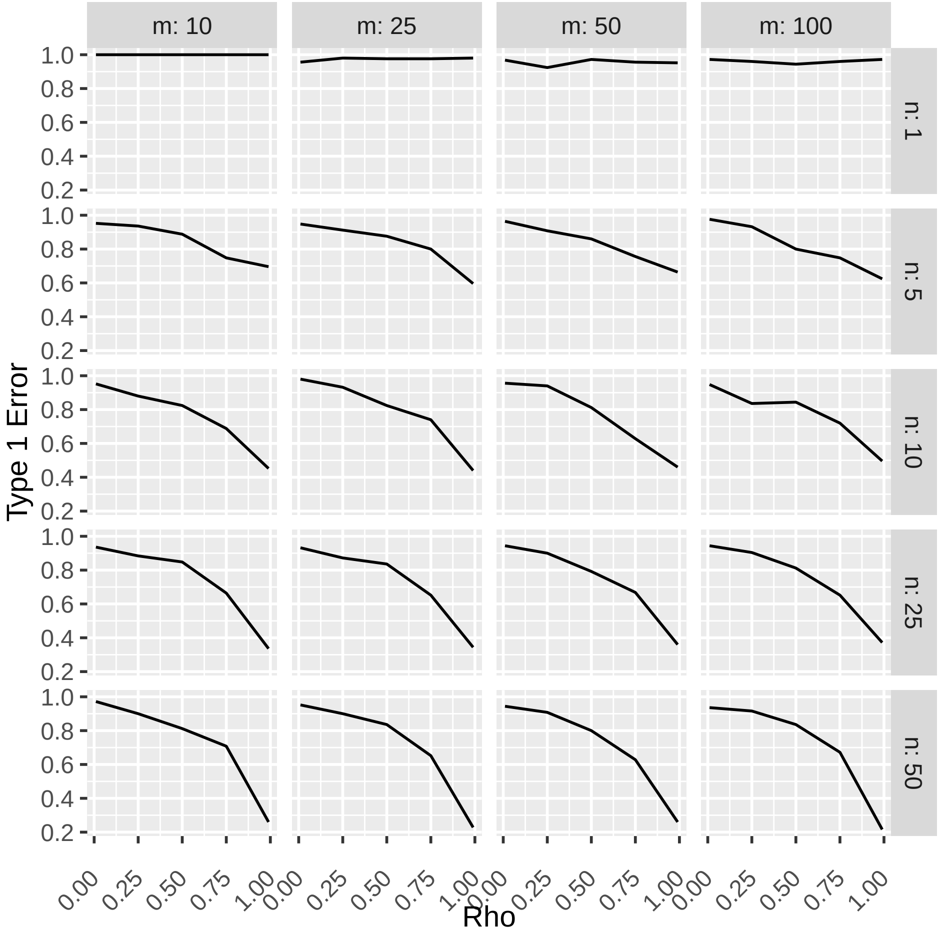
In this simulation we investigate the effect of correlated observations on the performance of logistic regression. Specifically, we evaluate how different degrees of correlations effect the type-1 error, bias and confidence interval coverage rate of linear regression. The simulation design is as follows.

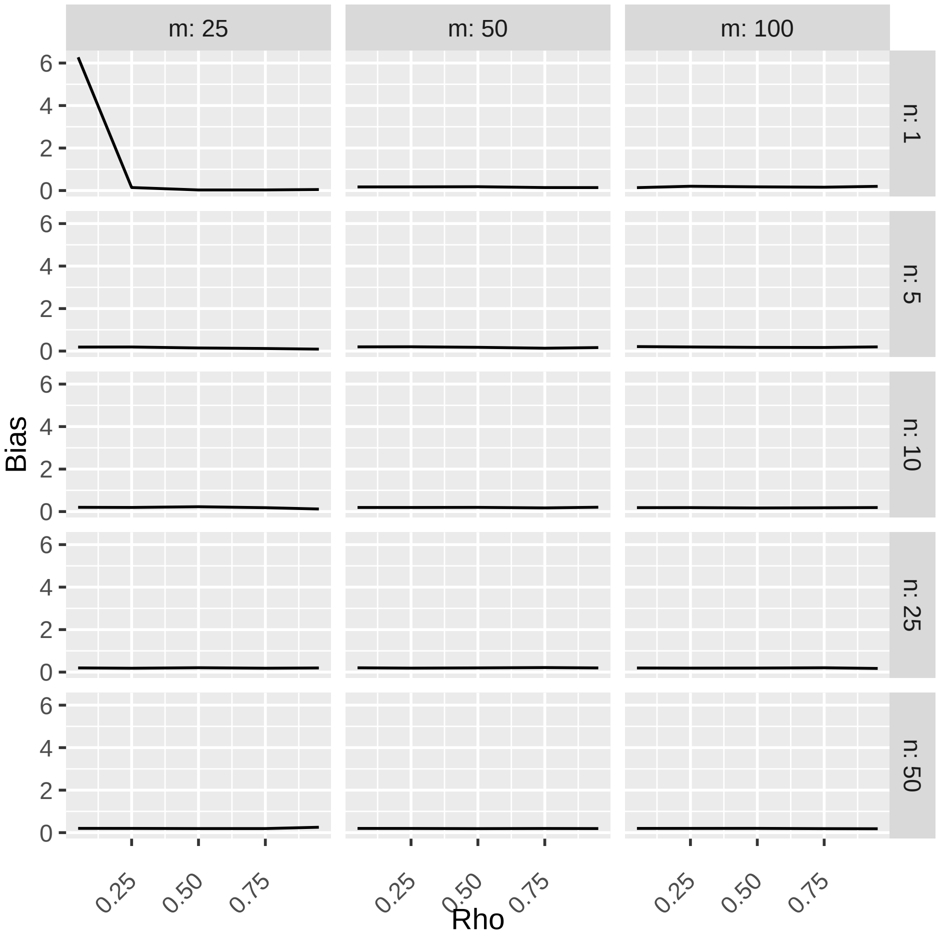
We consider m individuals each with n observations. The individuals are independent of one another and identically distributed; however, for each individual, the observations are correlated. We represent each individual by the data-structure $O_i = (X_i, Y_{i1}, Y_{i2}, \ldots, Y_{in})$ where X_i is a baseline covariate that is uniformly distributed in $[-1, 1]$ and $(Y_{i1}, \ldots, Y_{in})$ are the n individual-specific outcome observations that are drawn from a multivariate binomial distribution (based on thresholding a normal distribution; see the R package bindata and function rmvbin for a description) with marginal probabilities $expit(1 + \beta \cdot X_{ij})$ where β is a to-be-specified coefficient of interest and covariance matrix Σ (for the normal distribution component) with $\Sigma_{ij} = 1.5 \cdot \rho^{|i-j|}$ where ρ is a correlation coefficient.

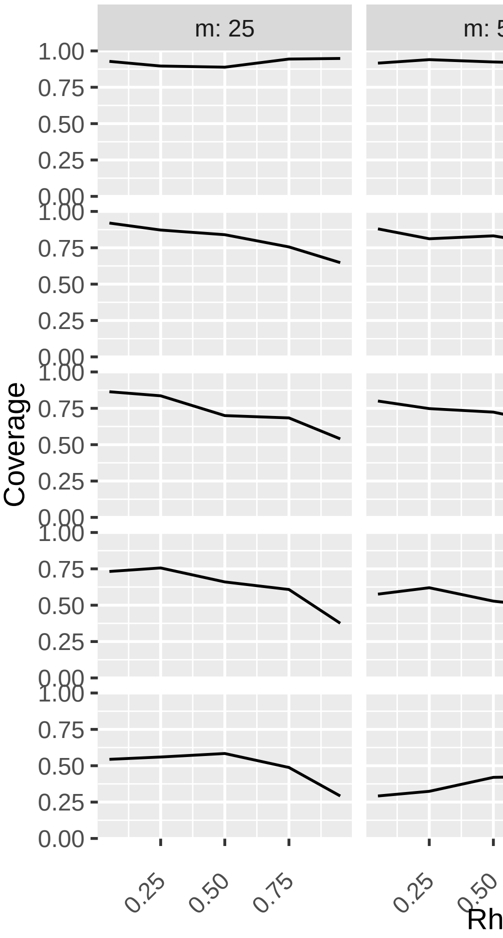
To evaluate the type-1 error, we take $\beta = 0$ (the null model) and do the following. For all combinations of the following values of m, n, ρ, we simulate ($K = 100$ times) m iid realizations $[O_i : i = 1, \ldots, m]$ and compute the logistic regression of the outcomes Y_{ij} on the X_i using the model $E[Y_{ij} \mid X_i] = expit(1 + \beta \cdot X_{ij})$. For each simulation iteration, we obtain a p-value for the coefficient β_1 and check whether it is below the cutoff $\alpha = 0.05$. We record the proportion of times this occurs, which is an estimate of the type-1 error of the logistic regression estimator.

To evaluate confidence interval coverage, we run simulations in the same way as the previous paragraph but with $\beta = 1$. We compute 0.95% confidence intervals for β_1 in the model $E[Y_{ij} \mid X_i] = expit(\beta_0 + \beta_1 X_i)$ and record the proportion of times that the confidence interval contains $\beta = 1$ and the average absolute difference between the estimated and true coefficients, which are estimates of the confidence interval coverage and bias of the logistic regression estimator.

The results of both simulations are graphically displayed below. We see the same results as found in Problem 1. This is expected since there is not much fundamentally different between the binomial and gaussian case.




m: 25
m: 50
m: 100
n: 1
n: 5
n: 10
n: 25
n: 50
Bias
0
2
4
6
0.25
0.50
0.75
Rho


m: 25
m: 5
Coverage
0.00
0.25
0.50
0.75
1.00
0.25
0.50
0.75
Rh