Biost/Stat 533
Emerson, Spr 99
Homework #5 Key
June 5, 1999

1. Consider a simple linear regression model $\vec{Y} = \gamma_0 + \gamma_1 \vec{X} + \vec{\epsilon}$ with $\vec{\epsilon} \sim (0, \sigma^2 \mathbf{I}_n)$.

   a. We are interested in a two-sided level α test of the null hypothesis $H_0 : \gamma_1 = c_0$, and we would like to have power β to detect the alternative hypothesis $H_1 : \gamma_1 = c_1$. Using the asymptotic distribution for $\hat{\gamma}_1$, derive a sample size formula for the case where the X_i's will be sampled from a distribution with mean μ_X and variance V_X. Show that this sample size formula reduces to the usual formula for the two sample t test when the X_i's are dichotomous variables.

   Ans: We can derive a sample size formula based on the asymptotic marginal distribution of $\hat{\gamma}_1$

$$\hat{\gamma}_1 \sim \mathcal{N}\left(\gamma_1, \frac{\sigma^2}{S_{XX}}\right)$$

Using the fact that $S_{XX}/(n-1)$ estimates $var(X)$, we have

$$Z = \frac{\sqrt{(n-1)var(X)}(\hat{\gamma}_1 - c_0)}{\sigma} \sim \mathcal{N}(\delta, 1)$$

with $\delta = \sqrt{(n-1)V_X}(\gamma_1 - c_0)/\sigma$. Under $H_0 : \gamma_1 = c_0$ Z has a standard normal distribution. Thus we would reject the null hypothesis whenever $|Z| > z_{1-\alpha_2}$, where z_p is the pth percentile of the standard normal distribution.

To compute the power, we note that

$$Pr_{H_1}(Z > z_{1-\alpha_2}) = Pr_{H_1}(Z - \delta > z_{1-\alpha_2} - \delta) = 1 - \Phi(z_{1-\alpha_2} - \delta)$$

where $\Phi(\cdot)$ is the cumulative distribution function for the standard normal.

If we want power β when $\gamma_1 = c_1$, then

$$\begin{aligned}\Phi^{-1}(1-\beta) = -z_\beta &= z_{1-\alpha/2} - \delta \\ -z_\beta &= z_{1-\alpha/2} - \sqrt{(n-1)V_X}(c_1 - c_0)/\sigma\end{aligned}$$

and solving for n

$$n = 1 + \frac{(z_{1-\alpha/2} + z_\beta)^2 \sigma^2}{V_X (c_1 - c_0)^2}$$

Note that when X is a binary 0-1 variable with equal sample sizes, V_X is just 1/4. In this case $n/2$ is the sample size for a single group and the above formula reduces to

$$\frac{n}{2} = 1 + \frac{(z_{1-\alpha/2} + z_\beta)^2 2\sigma^2}{(c_1 - c_0)^2}$$

Usually, we do not consider the the aspect that we are using the sample variance of our predictors, and thus the leading 1 in the formula drops out, and we obtain the standard sample size formula.

For $c_0 = 0$, an alternative approach can be based on the fact that the test that $H_0 : \gamma_1 = 0$ is exactly equivalent to the test of $H_0 : \rho = 0$, where ρ is the correlation between X and Y

(technically we require X and Y to be bivariate normal random variables). It can be shown that the correlation ρ between X and Y should be approximately

$$\rho = \gamma_1 \sqrt{\frac{var(X)}{\gamma_1^2 var(X) + \sigma^2}}$$

so we can easily translate the problem into finding the sample size having adequate power to test rho. Thus $\rho_1 = c_1\sqrt{V_X/(c_1^2 V_X + \sigma^2)}$ corresponds to the design alternative $H_1 : \gamma_1 = c_1$ for given V_X and σ^2.

In order to find a sample size formula, we need to know the distribution of our test statistic under the null and alternative. The statistic based on r above is not very good for nonzero ρ, so we use the distribution of the Fisher transform, and consider a statistic $Z = \frac{\sqrt{n-3}}{2} \ln\left(\frac{1+r}{1-r}\right) \sim \mathcal{N}(\delta, 1)$ where $\delta = \sqrt{n-3}\ln((1+\rho)/(1-\rho))/2$. Note that under $H_0 : \rho = 0$, Z has a standard normal distribution. Thus we would reject the null hypothesis whenever $|Z| > z_{1-\alpha_2}$, where z_p is the pth percentile of the standard normal distribution.

To compute the power, we note that

$$Pr_{H_1}(Z > z_{1-\alpha_2}) = Pr_{H_1}(Z - \delta > z_{1-\alpha_2} - \delta) = 1 - \Phi(z_{1-\alpha_2} - \delta)$$

where $\Phi(\cdot)$ is the cumulative distribution function for the standard normal.

If we want power P when $\gamma_1 = c_1$, then

$$\begin{aligned} \Phi^{-1}(1-P) = -z_P &= z_{1-\alpha/2} - \delta \\ -z_P &= z_{1-\alpha/2} - \frac{\sqrt{n-3}}{2} \ln\left(\frac{1+\rho_1}{1-\rho_1}\right) \end{aligned}$$

Solving for n, we find

$$n = 3 + \frac{4(z_{1-\alpha/2} + z_P)^2}{\ln^2\left(\frac{1+\rho_1}{1-\rho_1}\right)}$$

We can compare the two equations as a function of the sample size and the distribution of the predictors. I note that neither sample size formula uses the t distribution for the test statistic, so both might be expected to be a little low. The following table presents the sample size estimated by each equation along with the estimated power of those sample sizes from simulation studies. In all cases, I used $c_1 = 2$, $\alpha = .05$, and $V_X = 5$. “Correlation” refers to the second formula, and “Slope” refers to the first. For each method, I consider cases when the predictors are normally distributed, uniformly distributed, a binary variable (50% per group on average), or exponentially distributed. I vary σ^2 and P, and for each case, I present the sample size predicted by the formula (n) and the simulated power estimate corresponding to that sample size (“Sim Pwr”).

From this table you can see that the sample sizes predicted by the “Correlation” formula tend to be 5-6 subjects higher than the “Slope” formula. When small sample sizes are predicted, the “Slope” formula tends to be too optimistic. When the distribution of the predictors is exponential, the “Slope” formula is particularly bad in small samples. As the sample sizes have not changed as I have changed the distribution of the predictors, this merely suggests that the power-sample size relationship is rather steep with exponentially distributed predictors.

			Correlation		Slope	
X Distn	σ^2	P	n	Sim Pwr	n	Sim Pwr
Normal	25	0.50	9	0.569	6	0.326
Normal	25	0.80	15	0.826	11	0.664
Normal	25	0.95	23	0.954	17	0.890
Normal	100	0.50	23	0.498	20	0.455
Normal	100	0.80	45	0.819	40	0.774
Normal	100	0.95	72	0.954	66	0.936
Normal	200	0.50	43	0.502	39	0.464
Normal	200	0.80	84	0.798	79	0.790
Normal	200	0.95	137	0.954	131	0.940
Uniform	25	0.50	9	0.556	6	0.340
Uniform	25	0.80	15	0.842	11	0.685
Uniform	25	0.95	23	0.966	17	0.891
Uniform	100	0.50	23	0.514	20	0.454
Uniform	100	0.80	45	0.820	40	0.766
Uniform	100	0.95	72	0.954	66	0.936
Uniform	200	0.50	43	0.514	39	0.476
Uniform	200	0.80	84	0.808	79	0.789
Uniform	200	0.95	137	0.944	131	0.945
Binary	25	0.50	9	0.576	6	0.431
Binary	25	0.80	15	0.868	11	0.702
Binary	25	0.95	23	0.978	17	0.916
Binary	100	0.50	23	0.516	20	0.463
Binary	100	0.80	45	0.829	40	0.768
Binary	100	0.95	72	0.958	66	0.942
Binary	200	0.50	43	0.518	39	0.491
Binary	200	0.80	84	0.814	79	0.790
Binary	200	0.95	137	0.955	131	0.949
Exponential	25	0.50	9	0.500	6	0.292
Exponential	25	0.80	15	0.746	11	0.590
Exponential	25	0.95	23	0.910	17	0.788
Exponential	100	0.50	23	0.505	20	0.422
Exponential	100	0.80	45	0.757	40	0.732
Exponential	100	0.95	72	0.928	66	0.903
Exponential	200	0.50	43	0.514	39	0.467
Exponential	200	0.80	84	0.780	79	0.766
Exponential	200	0.95	137	0.930	131	0.930

b. We are interested in estimating γ_1 with a two-sided $100(1-\alpha)\%$ confidence interval such that the width of the confidence interval is Δ. Using the asymptotic distribution for $\hat{\gamma}_1$, derive a sample size formula for the case where the X_i's will be sampled from a distribution with mean μ_X and variance V_X. Show that this formula corresponds to the same formula as derived in part (a) when $\Delta = c_1 - c_0$ and $\beta = 1 - \alpha/2$. (Thus, when designing a two-sided level α test having power $\beta = 1 - \alpha/2$ to detect the alternative, a $100(1-\alpha)\%$ confidence interval will with probability 1 contain at most one of the null or alternative hypotheses. In this sense, the experiment will with $100(1-\alpha)\%$ confidence discriminate between the null and the alternative.)

Ans: The width of a confidence interval is $\Delta = 2z_{1-\alpha/2}\widehat{se}(\hat{\gamma}_1)$. As above, the standard error can be approximated by $\sigma/\sqrt{(n-1)V_X}$ and solving for n we obtain

$$n = 1 + \frac{(2z_{1-\alpha/2})^2\sigma^2}{V_X\Delta^2}$$

When $\Delta = c_1 - c_0$ and $\beta = 1 - \alpha/2$, this is the same formula as was given above.

2. Consider a linear regression model $\vec{Y} = \beta_0 + \beta_1 \vec{X} + \beta_2 \vec{Z} + \vec{\epsilon}$ with $\vec{\epsilon} \sim (0, \sigma^2 \mathbf{I}_n)$. We are interested in using simulation to explore the differences between fitting the full model as given above and fitting a reduced model $\vec{Y} = \gamma_0 + \gamma_1 \vec{X} + \vec{\delta}$ under the following design situations:

   A. Replications of an experimental design with the same design matrix each time and in which the sample correlation between $\vec{X}$ and $\vec{Z}$ is zero.

   B. Replications of an experimental design in which the design matrix can vary across replications, but in each case the sample correlation between $\vec{X}$ and $\vec{Z}$ is zero.

   C. Replications of a completely randomized design in which the design matrix varies across replications and $\vec{X}$ and $\vec{Z}$ are sampled independently of each other.

   D. Replications of an observational study in which the design matrix varies across replications and $\vec{X}$ and $\vec{Z}$ are sampled from a distribution in which $corr(X_i, Z_i) = \rho$.

   In each of the above situations, we are interested in examining the mean and standard deviation of the least squares estimates $\hat{\beta}_1$ and $\hat{\gamma}_1$, as well as the mean and standard deviation of the estimated standard errors $\widehat{se}(\hat{\beta}_1)$ and $\widehat{se}(\hat{\gamma}_1)$ across the simulated replications. Ultimately, we wish to compare the degree to which the estimated standard errors from each model accurately predict the true standard deviation of the least squares estimates.

   For the purposes of the simulation, we will choose $\beta_0 = 0$, $\epsilon_i \sim \mathcal{N}(0,1)$, $\rho = .7$, and $n = 200$. Furthermore, for standardization of results we want to sample $\vec{X}$ and $\vec{Z}$ such that $\overline{X} = \overline{Z} = 0$ and $\sum X_i^2 = \sum Z_i^2 = n$. For settings A and B, we will also require that $\sum X_i Z_i = 0$ in each design matrix.

   Simulate 100 replications of each setting for the four cases of

   1. $\beta_1 = \beta_2 = 0$
   2. $\beta_1 = 0$ and $\beta_2 = 1$
   3. $\beta_1 = 1$ and $\beta_2 = 0$
   4. $\beta_1 = \beta_2 = 1$

   For each of the 16 simulations (A1-A4,B1-B4,C1-C4,D1-D4), comment on any observed bias of the least squares estimates $\hat{\beta}_1$ and $\hat{\gamma}_1$, the agreement between the observed standard deviations of those estimates and the average estimated standard errors for those estimates, and estimate the true coverage probability for 95% confidence intervals based on the least squares estimates and their estimated standard errors, using the observed standard deviation of the least squares estimates across the replications as their true standard deviation. Be certain to explain how you estimated that coverage probability.

   Ans: Example simulation results are contained in a separate file, `hw5key.rsl`.