

- 4.1) a) $E(Y_i) = n_i p$, $V(Y_i) = n_i p(1-p)$.
 b) $y_i = n_i p + \varepsilon_i$ for $i = 1, \dots, N$, where ε_i are independently distributed with $E(\varepsilon_i) = 0$, and $\text{Var}(\varepsilon_i) = n_i p(1-p)$, the value of p is unknown.
 c) If we write $p(1-p) = \sigma^2$, then the variance-covariance matrix of the noise vector is $\sigma^2 V$, where $V = \text{diag}\{n_1, n_2, \dots, n_N\}$. Ignoring the fact that σ^2 is a function of p , the BLUE is the GLS estimator, which is the solution to $X'V^{-1}Xp = X'V^{-1}y$, where $X' = [n_1, \dots, n_N]$. Now, $X'V^{-1}X = \sum n_i = N$ (say) and $X'V^{-1}y = \sum y_i$. Thus the BLUE for p is $\hat{p} = \sum y_i / N$. Its variance equals $p(1-p) / N$.

[Note: We do not need the distributional assumptions for BLUE. The error terms here are not independent, identically distributed. Of course, the BLUE in this problem is also the MLE and MVUE, once we use the binomial distribution. The OLS for this problem is

$$\tilde{p} = \frac{X'Y}{X'X} = \frac{\sum_i n_i Y_i}{\sum_i n_i^2}. \text{ It is an unbiased estimator. Convince yourself or prove that its}$$

variance is more than $p(1-p) / N$.]

- 4.2) Since $\lambda'b$ is estimable, there exists a vector \mathbf{a} such that $\lambda = \mathbf{X}'\mathbf{a}$. Now $\lambda'\hat{b} = \mathbf{a}'X(X'X)^-X'Y = \mathbf{a}'PY$, where P is the projection matrix on the column space $C(X)$. Since P is unique, thus it does not depend on the choice of the generalized-inverse. Clearly, its variance $\text{Var}(\lambda'\hat{b}) = \mathbf{a}'P[\text{Var}(Y)]Pa = \sigma^2 \mathbf{a}'Pa$, does not depend on the generalized inverse.

- 4.4) The heteroskedasticity model is $y_i = \beta x_i + \varepsilon_i$ where $V(\varepsilon_i) = \sigma^2 x_i \Rightarrow V(\varepsilon_i/\sqrt{x_i}) = \sigma^2$

$$\text{Using the GLS Model, } V = \begin{bmatrix} x_1 & & 0 \\ & \ddots & \\ 0 & & x_n \end{bmatrix}, \text{ so } \hat{\beta}_{\text{GLS}} = \frac{\sum \sqrt{x_i} y_i}{\sum x_i}$$

$$V(\hat{\beta}_{\text{GLS}}) = V\left(\frac{\sum \sqrt{x_i} y_i}{\sum x_i}\right) = \sigma^2 / \sum x_i$$

In the OLS model, let $z_i = y_i/\sqrt{x_i}$ and $u_i = x_i/\sqrt{x_i}$, then the model is $z_i = \beta u_i + \varepsilon_i$

$$\hat{\beta}_{\text{OLS}} = \frac{\sum x_i y_i}{\sum x_i} \text{ and } V(\hat{\beta}_{\text{OLS}}) = V\left(\frac{\sum x_i y_i}{\sum x_i}\right) = \sigma^2 \frac{\sum x_i^3}{(\sum x_i^2)^2}$$

By Aitken's theorem, the GLS estimator is the BLUE, so $V(\hat{\beta}_{\text{GLS}}) \leq V(\hat{\beta}_{\text{OLS}})$

- 4.17) a) Note the trig identity relationship where $d_1 \cos(\omega t) + d_2 \sin(\omega t) = A \sin(\omega t + \theta)$ for some constant d_1 and d_2 . Thus, there is a value of d_1 and d_2 such that $s_4 = d_1 c_1 + d_2 s_1$, $s_5 = d_1 c_2 + d_2 s_2$, and $s_6 = d_1 c_3 + d_2 s_3$.

Likewise, there is a value of d_1' and d_2' such that $c_4 = d_1'c_1 + d_2's_1$, $c_5 = d_1'c_2 + d_2's_2$, and $c_6 = d_1'c_3 + d_2's_3$.

So $(c_4, c_5, c_6, s_4, s_5, s_6)$ and $(c_1, c_2, c_3, s_1, s_2, s_3)$ are linearly dependent.

b) Answers will vary depending on the value of N and the response variable used in the regression. For parts b-d, $N = 50$ and 100 will be used and s_6 is the response and Minitab is used for the exercises. The dependencies will be detected when the variance inflation factors (VIF) are high for all but one predictor.

b) Using 3.1416 as the value for π :
(N = 50) Regression Analysis: sin6 versus cos1, cos2, ...

The regression equation is
 $\text{sin6} = 0.000012 - 0.050 \text{ cos1} - 0.063 \text{ cos2} - 0.028 \text{ cos3} + 0.028 \text{ cos4} + 0.063 \text{ cos5}$
 $+ 0.051 \text{ cos6} - 1.00 \text{ sin1} - 0.014 \text{ sin2} - 0.058 \text{ sin3} - 0.058 \text{ sin4}$
 $- 0.014 \text{ sin5}$

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.00001241	0.00002392	0.52	0.607	
cos1	-0.0503	0.1601	-0.31	0.755	2.24757E+07
cos2	-0.0633	0.1571	-0.40	0.689	2.13389E+07
cos3	-0.0281	0.1638	-0.17	0.865	2.39238E+07
cos4	0.0282	0.1638	0.17	0.864	2.39228E+07
cos5	0.0633	0.1571	0.40	0.689	2.13383E+07
cos6	0.0507	0.1601	0.32	0.753	2.24760E+07
sin1	-1.00002	0.00007	-14729.76	0.000	4.076
sin2	-0.0145	0.1650	-0.09	0.931	2.44059E+07
sin3	-0.0585	0.1583	-0.37	0.714	2.18001E+07
sin4	-0.0585	0.1583	-0.37	0.714	2.18011E+07
sin5	-0.0144	0.1650	-0.09	0.931	2.44064E+07

S = 0.000168477 R-Sq = 100.0% R-Sq(adj) = 100.0%

(N = 100) Regression Analysis: sin6 versus cos1, cos2, ...

The regression equation is
 $\text{sin6} = 0.000011 - 0.015 \text{ cos1} - 0.039 \text{ cos2} - 0.024 \text{ cos3} + 0.024 \text{ cos4} + 0.039 \text{ cos5}$
 $+ 0.016 \text{ cos6} - 1.00 \text{ sin1} + 0.003 \text{ sin2} - 0.039 \text{ sin3} - 0.039 \text{ sin4}$
 $+ 0.003 \text{ sin5}$

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.00001071	0.00003134	0.34	0.733	
cos1	-0.0152	0.1030	-0.15	0.883	5349500.155
cos2	-0.0390	0.1044	-0.37	0.710	5545868.648
cos3	-0.0241	0.1055	-0.23	0.820	5700481.642
cos4	0.0242	0.1055	0.23	0.819	5699907.640
cos5	0.0390	0.1044	0.37	0.710	5545965.513
cos6	0.0160	0.1030	0.16	0.877	5349531.317
sin1	-1.00001	0.00009	-11328.72	0.000	4.016
sin2	0.0026	0.1053	0.03	0.980	5666326.510
sin3	-0.0386	0.1043	-0.37	0.712	5522554.651
sin4	-0.0386	0.1043	-0.37	0.712	5523121.136
sin5	0.0027	0.1053	0.03	0.980	5666221.831

With the VIF for all predictors but s_1 , the regression procedure can detect dependency.

c) Repeating with 3.14 as the value for π :

(N = 50) Regression Analysis: sin6 versus cos1, cos2, ...

The regression equation is

$$\begin{aligned} \text{sin6} = & -0.00279 - 0.132 \text{ cos1} - 0.059 \text{ cos2} - 0.028 \text{ cos3} + 0.022 \text{ cos4} \\ & + 0.055 \text{ cos5} + 0.050 \text{ cos6} - 0.992 \text{ sin1} - 0.011 \text{ sin2} - 0.052 \text{ sin3} \\ & - 0.054 \text{ sin4} - 0.018 \text{ sin5} \end{aligned}$$

Predictor	Coef	SE Coef	T	P	VIF
Constant	-0.002790	0.005162	-0.54	0.592	
cos1	-0.1318	0.1583	-0.83	0.410	471.554
cos2	-0.0592	0.1565	-0.38	0.707	454.027
cos3	-0.0277	0.1628	-0.17	0.866	506.722
cos4	0.0219	0.1636	0.13	0.894	511.367
cos5	0.0553	0.1568	0.35	0.726	456.903
cos6	0.0495	0.1589	0.31	0.757	473.436
sin1	-0.99189	0.01460	-67.95	0.000	4.041
sin2	-0.0109	0.1644	-0.07	0.947	520.063
sin3	-0.0516	0.1582	-0.33	0.746	466.884
sin4	-0.0537	0.1573	-0.34	0.735	462.369
sin5	-0.0175	0.1642	-0.11	0.916	517.243

Regression Analysis: sin6 versus cos1, cos2, ...

The regression equation is

$$\begin{aligned} \text{sin6} = & -0.00227 - 0.165 \text{ cos1} - 0.038 \text{ cos2} - 0.031 \text{ cos3} + 0.024 \text{ cos4} \\ & + 0.037 \text{ cos5} + 0.006 \text{ cos6} - 0.982 \text{ sin1} + 0.014 \text{ sin2} - 0.036 \text{ sin3} \\ & - 0.040 \text{ sin4} + 0.007 \text{ sin5} \end{aligned}$$

Predictor	Coef	SE Coef	T	P	VIF
Constant	-0.002272	0.006773	-0.34	0.738	
cos1	-0.1655	0.1015	-1.63	0.106	111.103
cos2	-0.0384	0.1043	-0.37	0.714	118.382
cos3	-0.0309	0.1050	-0.29	0.769	120.487
cos4	0.0240	0.1059	0.23	0.821	123.056
cos5	0.0366	0.1042	0.35	0.726	117.889
cos6	0.0062	0.1027	0.06	0.952	113.924
sin1	-0.98156	0.01896	-51.77	0.000	3.965
sin2	0.0144	0.1052	0.14	0.891	120.931
sin3	-0.0364	0.1045	-0.35	0.728	119.055
sin4	-0.0401	0.1036	-0.39	0.700	116.521
sin5	0.0069	0.1052	0.07	0.948	121.464

Using 3.14, the linear dependence is still detected, but it is weaker with the VIF being substantially lower than with a more precise estimate of π .

c) Repeating with 3.1416 as the value for π and replacing 7 with 4:
(N=50) Regression Analysis: sin6 versus cos1, cos2, ...

The regression equation is

$$\begin{aligned} \text{sin6} = & -0.000022 - 0.000000 \text{ cos1} - 0.000562 \text{ cos2} + 0.000000 \text{ cos3} \\ & + 0.000022 \text{ cos4} + 0.000000 \text{ cos5} + 0.000562 \text{ cos6} - 0.000000 \text{ sin1} \\ & + 3.00 \text{ sin2} - 0.000000 \text{ sin3} + 0.000000 \text{ sin4} + 0.000000 \text{ sin5} \end{aligned}$$

Predictor	Coef	SE Coef	T	P	VIF
Constant	-0.00002201	0.00003021	-0.73	0.471	
cos1	-0.00000003	0.00004272	-0.00	0.999	2.61461E+15
cos2	-0.00056204	0.00001511	-37.21	0.000	6.54177E+14
cos3	0.00000002	0.00002136	0.00	0.999	6.53654E+14
cos4	0.00002201	0.00003021	0.73	0.471	17.570
cos5	0.00000001	0.00002136	0.00	0.999	6.53653E+14
cos6	0.00056204	0.00001511	37.21	0.000	6.54177E+14
sin1	-0.00000019	0.00004272	-0.00	0.996	2.61462E+15
sin2	3.00000	0.00000	1.28602E+08	0.000	72.269
sin3	-0.00000009	0.00002136	-0.00	0.997	6.53653E+14
sin4	0.00000001	0.00000001	0.71	0.484	17.536
sin5	0.00000010	0.00002136	0.00	0.996	6.53654E+14

(N=100) Regression Analysis: sin6 versus cos1, cos2, ...

The regression equation is

$$\begin{aligned} \text{sin6} = & -0.000022 - 0.000044 \text{ cos1} - 0.00111 \text{ cos2} + 0.000022 \text{ cos3} + 0.000022 \text{ cos4} \\ & + 0.000022 \text{ cos5} + 0.00111 \text{ cos6} - 0.000044 \text{ sin1} + 3.00 \text{ sin2} \\ & - 0.000022 \text{ sin3} + 0.000000 \text{ sin4} + 0.000022 \text{ sin5} \end{aligned}$$

Predictor	Coef	SE Coef	T	P	VIF
Constant	-0.00002204	0.00003942	-0.56	0.578	
cos1	-0.00004406	0.00005575	-0.79	0.431	1.55738E+14
cos2	-0.00111299	0.00001971	-56.46	0.000	3.89347E+13
cos3	0.00002205	0.00002788	0.79	0.431	3.89346E+13
cos4	0.00002204	0.00003942	0.56	0.578	16.431
cos5	0.00002202	0.00002788	0.79	0.432	3.89346E+13
cos6	0.00111299	0.00001971	56.46	0.000	3.89346E+13
sin1	-0.00004406	0.00005575	-0.79	0.431	1.55738E+14
sin2	3.00000	0.00000	49680577.48	0.000	66.720
sin3	-0.00002201	0.00002788	-0.79	0.432	3.89346E+13
sin4	0.00000002	0.00000003	0.54	0.589	16.432
sin5	0.00002205	0.00002788	0.79	0.431	3.89346E+13

The regression procedure detects the linear dependence between most variables with high values of VIF. However, some of the dependencies with the other variables are not detected.

4.19) a) The model is
$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}, \text{ or } \mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon}.$$

Solving for β using OLS, $\hat{\beta} = \mathbf{t}'\mathbf{y}$ where $\mathbf{t}' = [-0.2, -0.1, 0, 0.1, 0.2]$

b) $E(\hat{\beta}) = \sum t_i E(y_i) = \sum t_i(\alpha + i\beta) = -0.2(\alpha + \beta) - 0.1(\alpha + 2\beta) + 0.1(\alpha + 4\beta) + 0.2(\alpha + 5\beta) = (-0.2 - 0.2 + 0.4 + 1)\beta = \beta$, thus is unbiased.

$V(\hat{\beta}) = \sum t_i^2 V(y_i) = (0.2^2 + 0.1^2 + 0.1^2 + 0.2^2) = \sigma^2/10$

c) $E(\hat{\gamma}) = E[(y_4 - y_2)/2] = (1/2)[E(y_4) - E(y_2)] = (1/2)[\alpha + 4\beta - \alpha - 2\beta] = \beta$ (unbiased)

$V(\hat{\gamma}) = (1/4)[V(y_4) + V(y_2)] = \sigma^2/2$

$E(\hat{\eta}) = E[(y_5 - y_1)/4] = (1/4)[E(y_5) - E(y_1)] = (1/4)[\alpha + 5\beta - \alpha - \beta] = \beta$ (unbiased)

$V(\hat{\eta}) = (1/16)[V(y_5) + V(y_1)] = \sigma^2/8$

d) $V(\hat{\delta}) = c^2 V(\hat{\gamma}) + (1-c)^2 V(\hat{\eta}) = c^2 \sigma^2/8 + (1-c)^2 \sigma^2/2$

e) $\frac{dV(\hat{\delta})}{dc} = \frac{2c\sigma^2}{8} - \frac{2(1-c)\sigma^2}{2} = 0 \Rightarrow c/4 = 1 - c \Rightarrow c = 1/5$

f) If $\alpha = 0$, then $X' = [1, 2, 3, 4, 5]$

Solving for β using OLS, $\hat{\beta}_{\text{new}} = \sum y_i/55 \neq \hat{\beta}$

So the value of $\hat{\beta}$ is not the BLUE.

4.26) a) $E(\tilde{\mathbf{b}}) = (\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}'E(\mathbf{y}) = (\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{X}\mathbf{b}$

b) $E(\lambda'\tilde{\mathbf{b}}) = \lambda'(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{X}\mathbf{b} \neq \lambda'\mathbf{b}$, so it is not unbiased.

c) $\text{Cov}(\tilde{\mathbf{b}}) = (\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}'\text{Cov}(\mathbf{y},\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1} = \sigma^2(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1}$

d) $\text{MSE} = E(\|\tilde{\mathbf{b}} - \mathbf{b}\|) = E[(\tilde{\mathbf{b}} - \mathbf{b})'(\tilde{\mathbf{b}} - \mathbf{b})]$

$= E[(\hat{\mathbf{b}} - \mathbf{b})'(\hat{\mathbf{b}} - \mathbf{b})] + \text{trace}(\text{cov}(\tilde{\mathbf{b}} - \mathbf{b}))$

Let $Z = (\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1}$

$= \text{trace}\{E[(\hat{\mathbf{b}} - \mathbf{b})(\hat{\mathbf{b}} - \mathbf{b})']\} + \text{trace}(\sigma^2\mathbf{Z}\mathbf{X}'\mathbf{X}\mathbf{Z})$

$= (\mathbf{Z}\mathbf{X}'\mathbf{X}\mathbf{b} - \mathbf{b})'(\mathbf{Z}\mathbf{X}'\mathbf{X}\mathbf{b} - \mathbf{b}) + \sigma^2\text{trace}[(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1}]$

e) $\text{Cov}(\hat{\mathbf{b}}) = \text{Cov}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{cov}(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

$= \sigma^2 \begin{bmatrix} 1/10 & 0 & 0 \\ 0 & 5/9 & -4/9 \\ 0 & -4/9 & 5/9 \end{bmatrix}$

f) One such value is $k = 1$.

$\text{cov}(\tilde{\mathbf{b}}|k=1) = \sigma^2(\mathbf{X}'\mathbf{X} + \mathbf{I}_p)^{-1} = \sigma^2 \begin{bmatrix} 1/11 & 0 & 0 \\ 0 & 3/10 & -1/5 \\ 0 & -1/5 & 3/10 \end{bmatrix}$

g) The Gauss-Markov Theorem is not violated because $\tilde{\mathbf{b}}$ is not an unbiased estimator. The theorem only applies to unbiased estimators.

4.28) This is a GLS model where $V = \text{diag}(n_i^{-1})$, and need to find R such that $RVR' = I$

So $R = \text{diag}(\sqrt{n_i})$ and thus, $U = RX$

Solving for β using GLS, $\tilde{\beta}_1 = \frac{-(\sum n_i \bar{x}_i)(\sum n_i \bar{y}_i) + \sum n_i (\sum n_i \bar{x}_i \bar{y}_i)}{(\sum n_i)(\sum n_i \bar{x}_i^2) - (\sum n_i \bar{x}_i)^2}$

So $V(\tilde{\beta}_1) = [\sigma^2(U'U)^{-1}]_{22} = \frac{\sigma^2(\sum n_i)}{(\sum n_i)(\sum n_i \bar{x}_i) - (\sum n_i \bar{x}_i)^2} > \sigma^2 / (\sum \sum (x_{ij} - \bar{x}_{..})^2)$

7) Minimizing the error function $E = (Y - X\beta)'M(Y - X\beta)$
 $= Y'MY - Y'MX\beta - \beta'X'MY + \beta'X'MX\beta$

$$\frac{\partial E}{\partial \beta} = -Y'MX - X'MY + 2X'MX\beta = 0$$

$$\Rightarrow 2X'MX\beta = 2X'MY$$

$$\Rightarrow \tilde{\beta}_M = (X'MX)^- X'MY$$

(Note that since full rank is not assumed for X that you cannot assume $X'MX$ is invertible, so a generalized inverse is placed instead.)

$$i) E(I'\tilde{\beta}_M) = E[I'(X'MX)^- X'MY] = I'(X'MX)^- X'MX\beta$$

Since $I'\beta$ is estimable, then for some t , $I' = t'X$

$$= t'X(X'MX)^- X'MX\beta$$

$$= t'X\beta$$

$$= I'\beta$$

This expectation does not depend on M.

$$\text{Thus, } E(I'\tilde{\beta}_M) = E(I'\tilde{\beta}_{GLS})$$

ii) Both estimators are unbiased and we know that $I'\tilde{\beta}_{GLS}$ is the BLUE.