HW 03

Group 10

2022-10-13

1(a)

```
auto = read.csv('Auto.csv', na.strings = 'NA')
fit_1 = lm(mpg ~ cylinders + displacement +
             weight + origin + year +
             I(year^2), data=auto)
summary(fit_1)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + weight + origin +
##     year + I(year^2), data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6728 -1.9705 -0.0843  1.7412 13.0999
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.683e+02  7.919e+01   4.651 4.54e-06 ***
## cylinders    -1.822e-01  3.153e-01  -0.578    0.564
## displacement  4.110e-03  6.689e-03   0.615    0.539
## weight       -6.018e-03  5.546e-04 -10.849  < 2e-16 ***
## origin        1.219e+00  2.585e-01   4.717 3.35e-06 ***
## year         -9.441e+00  2.092e+00  -4.512 8.50e-06 ***
## I(year^2)     6.718e-02  1.375e-02   4.885 1.51e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.253 on 390 degrees of freedom
## Multiple R-squared:  0.8299, Adjusted R-squared:  0.8273
## F-statistic: 317.1 on 6 and 390 DF,  p-value: < 2.2e-16
```

Positive relationship: displacement, origin, year^2
Negative relationship: cylinders, weight, year
Weight, origin, year, and year squared are significantly different from 0 $R^2 = 0.8273$

1(b)

```
library(car)
```

```
## Loading required package: carData
```

```
vif(fit_1)
```

```
##    cylinders displacement       weight       origin         year    I(year^2)
##    10.777079    18.244389     8.278553     1.610621  2231.356088  2229.212604
```

Except origin, other variables indicates substantial and even severe multidisciplinary.

1(c)

```
fit_2 = lm(mpg ~ cylinders + displacement +
             origin + year + I(year^2),
           data=auto)
summary(fit_2)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + origin + year +
##     I(year^2), data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.508  -2.073  -0.077   1.999  13.990
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  547.713145  88.244451   6.207 1.38e-09 ***
## cylinders     -0.374380   0.358769  -1.044    0.297
## displacement  -0.038759   0.006149  -6.303 7.89e-10 ***
## origin         1.310805   0.294371   4.453 1.11e-05 ***
## year         -14.344667   2.328014  -6.162 1.79e-09 ***
## I(year^2)      0.099054   0.015309   6.470 2.93e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.706 on 391 degrees of freedom
## Multiple R-squared:  0.7785, Adjusted R-squared:  0.7757
## F-statistic: 274.9 on 5 and 391 DF,  p-value: < 2.2e-16
```

The parameter estimates' absolute value increase, which indicates the variables exert more fundamental influence on mpg.

$R^2 = 0.7757: \ decreases$

1(d)

```
fit_3 = lm(mpg ~ cylinders + origin + year +
             I(year^2), data=auto)
summary(fit_3)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + origin + year + I(year^2), data = auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.7680  -2.2858  -0.2858   2.0398  14.4988
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 468.81177   91.56522   5.120 4.80e-07 ***
## cylinders    -2.45827    0.14601 -16.837  < 2e-16 ***
## origin        1.85753    0.29487   6.300 8.03e-10 ***
## year        -12.23505    2.41495  -5.066 6.26e-07 ***
## I(year^2)     0.08551    0.01589   5.382 1.27e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.885 on 392 degrees of freedom
## Multiple R-squared:  0.756,  Adjusted R-squared:  0.7535
## F-statistic: 303.7 on 4 and 392 DF,  p-value: < 2.2e-16
```

Compared with model fit_1, the parameter estimates' absolute value increase substantially and R^2 decreases, compared with model fit_2, cylinders and origin increase their estimate parameters' absolute value, whereas estimate parameters of year and year^2 decrease. And R^2 decreases.

2(a)

```
set.seed(1)
x1 = runif(100)
x2 = .5*x1 + rnorm(100)/10
y = 2 + 2*x1 + .3*x2 + rnorm(100)
```

$$y = 2 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$
$$Where:$$
$$\beta_1 = 2,\ \beta_2 = 0.3,\ \epsilon \sim N(0,1)$$

2(b)

The correlation between x1 and x2 is 0.8351212.

2(c)

```
fit_4 = lm(y ~ x1 + x2)
summary(fit_4)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8311 -0.7273 -0.0537  0.6338  2.3359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
## x1            1.4396     0.7212   1.996   0.0487 *
## x2            1.0097     1.1337   0.891   0.3754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 97 degrees of freedom
## Multiple R-squared:  0.2088, Adjusted R-squared:  0.1925
## F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05
```

```
confint(fit_4)
```

```
##                    2.5 %   97.5 %
## (Intercept)  1.670278673 2.590721
## x1           0.008213776 2.870897
## x2          -1.240451256 3.259800
```

The parameter estimates for x1 and x2 are 1.4396 and 1.0097. Only x1 is significant at an $\alpha = 0.05$

Both true parameters are covered by the confidence intervals of the slope estimates.

2(d)

```
fit_5 = lm(y ~ x1)
summary(fit_5)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89495 -0.66874 -0.07785  0.59221  2.45560
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1124     0.2307   9.155 8.27e-15 ***
## x1            1.9759     0.3963   4.986 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.055 on 98 degrees of freedom
## Multiple R-squared:  0.2024, Adjusted R-squared:  0.1942
## F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06
```

```
confint(fit_5)
```

```
##                2.5 %   97.5 %
## (Intercept) 1.654488 2.570299
## x1          1.189529 2.762329
```

β_1 significantly different from 0. The true β_1 is still covered by the confidence interval.

2(e)

```
fit_6 = lm(y ~ x2)
summary(fit_6)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62687 -0.75156 -0.03598  0.72383  2.44890
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3899     0.1949   12.26  < 2e-16 ***
## x2            2.8996     0.6330    4.58 1.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.072 on 98 degrees of freedom
## Multiple R-squared:  0.1763, Adjusted R-squared:  0.1679
## F-statistic: 20.98 on 1 and 98 DF,  p-value: 1.366e-05
```

```
confint(fit_6)
```

```
##                2.5 %   97.5 %
## (Intercept) 2.003116 2.776783
## x2          1.643324 4.155846
```

β_2 is significantly different than 0. The true β_2 is not covered by the confidence interval however.

2(f)

Yes, in 2c both confidence intervals are covering the true values. The betas are also both significant. However, in 2e we found that β_2 alone is not significant. This is weird because in one model it is and one model it isn't. Normally, one would expect a significant independent variable to stay significant no matter what other variables are present in the model or not. This is unless there is multicollinearity.

3(a)

This equation represents a fork

3(b)

```
set.seed(1)
w = runif(500, min = 0, max = 5)
d = rnorm(500)
e = rnorm(500)
x = w + d
y = 4 + 2*x - 3*w + e

df <- data.frame(y, x, w)

cor(df)
```

```
##             y          x          w
## y  1.00000000 0.02953706 -0.5359720
## x  0.02953706 1.00000000  0.7968971
## w -0.53597195 0.79689706  1.0000000
```

```
summary(df)
```

```
##        y                 x                w
##  Min.   :-6.7202   Min.   :-2.222   Min.   :0.009184
##  1st Qu.:-0.4741   1st Qu.: 1.208   1st Qu.:1.290643
##  Median : 1.4850   Median : 2.447   Median :2.381348
##  Mean   : 1.4297   Mean   : 2.444   Mean   :2.478275
##  3rd Qu.: 3.3197   3rd Qu.: 3.683   3rd Qu.:3.670729
##  Max.   :10.0874   Max.   : 6.602   Max.   :4.980387
```

```
print("Standard Deviation Below (y, x, w)")
```

```
## [1] "Standard Deviation Below (y, x, w)"
```

```
print(c(sd(y), sd(x), sd(w)))
```

```
## [1] 2.775645 1.748330 1.416604
```

3(c)

```
fit_3c = lm(y ~ x)
summary(fit_3c)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0671 -1.9470  0.0493  1.9231  8.6186
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.31511    0.21364   6.156 1.54e-09 ***
## x            0.04689    0.07111   0.659     0.51
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.777 on 498 degrees of freedom
## Multiple R-squared:  0.0008724,  Adjusted R-squared:  -0.001134
## F-statistic: 0.4349 on 1 and 498 DF,  p-value: 0.5099
```

```
confint(fit_3c)
```

```
##                   2.5 %    97.5 %
## (Intercept)  0.89536188 1.7348570
## x           -0.09282142 0.1866073
```

The 95% CI does not cover the true slope for x and the slope for x is not significant.

3(d)

```
fit_3d = lm(y~x+w)
summary(fit_3d)
```

```
##
## Call:
## lm(formula = y ~ x + w)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1940 -0.7304 -0.0063  0.7367  3.0967
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.01829    0.09430   42.61   <2e-16 ***
## x            1.98649    0.04432   44.82   <2e-16 ***
## w           -3.00389    0.05470  -54.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.046 on 497 degrees of freedom
## Multiple R-squared:  0.8587, Adjusted R-squared:  0.8581
## F-statistic:  1510 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
confint(fit_3d)
```

```
##                 2.5 %    97.5 %
## (Intercept)  3.833009  4.203566
## x            1.899413  2.073563
## w           -3.111353 -2.896423
```

The coefficient of x significant at 0.05 level. the 95% CI covers the true slope for x.

3(e)

```
library(car)
vif(fit_3d)
```

```
##        x        w
## 2.740063 2.740063
```

The vif for x and w is 2.740063.

4(a)

This equation represents a collider.

4(b)

```
set.seed(1)
x = runif(500, min=0, max=5)
delta = rnorm(500)
e = rnorm(500)
y = x + delta
w = 4 + 2*x + 3*y + e

df = data.frame(y, x, w)
cor(df)
```

```
##           y         x         w
## y 1.0000000 0.7968971 0.9659606
## x 0.7968971 1.0000000 0.9031342
## w 0.9659606 0.9031342 1.0000000
```

```
summary(df)
```

```
##        y                x                  w
##  Min.   :-2.222   Min.   :0.009184   Min.   :-2.098
##  1st Qu.: 1.208   1st Qu.:1.290643   1st Qu.:10.283
##  Median : 2.447   Median :2.381348   Median :15.953
##  Mean   : 2.444   Mean   :2.478275   Mean   :16.266
##  3rd Qu.: 3.683   3rd Qu.:3.670729   3rd Qu.:22.377
##  Max.   : 6.602   Max.   :4.980387   Max.   :34.121
```

```
print("Standard Deviation Below (y, x, w)")
```

```
## [1] "Standard Deviation Below (y, x, w)"
```

```
print(c(sd(y), sd(x), sd(w)))
```

```
## [1] 1.748330 1.416604 7.738156
```

4(c)

```
fit_4c = lm(y ~ x)
summary(fit_4c)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9664 -0.6945 -0.0304  0.7220  3.8374
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.007071   0.095349   0.074    0.941
## x           0.983506   0.033410  29.437   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.057 on 498 degrees of freedom
## Multiple R-squared:  0.635,  Adjusted R-squared:  0.6343
## F-statistic: 866.6 on 1 and 498 DF,  p-value: < 2.2e-16
```

```
confint(fit_4c)
```

```
##                  2.5 %    97.5 %
## (Intercept) -0.1802644 0.1944062
## x            0.9178637 1.0491486
```

The coefficient of x is significant at the 0.05 level.

The 95% CI covers the true slope for x.

4(d)

```
fit_4d = lm(y ~ x + w)
summary(fit_4d)
```

```
##
## Call:
## lm(formula = y ~ x + w)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90578 -0.22784 -0.01861  0.22985  1.02035
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.212058   0.035014  -34.62   <2e-16 ***
## x           -0.505421   0.024465  -20.66   <2e-16 ***
## w            0.301809   0.004479   67.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3324 on 497 degrees of freedom
## Multiple R-squared:  0.964,  Adjusted R-squared:  0.9639
## F-statistic:  6654 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
confint(fit_4d)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.2808520 -1.1432648
## x           -0.5534892 -0.4573538
## w            0.2930095  0.3106088
```

The coefficient of x is significant at the 0.05 level.

The 95% CI does not cover the true slope for x.

4(e)

```
vif(fit_4d)
```

```
##        x        w
## 5.424507 5.424507
```

4(f)

The value of R^2 in the first model is 0.635, the value of R^2 in the second model is 0.964. According to R^2, the second model is better. But this may not be the right model because the vif suggests there are high multicollinearity.

5(a)

This is a pipe.

5(b)

```
set.seed(1)
w = runif(500, min = 0, max = 5)
d = rnorm(500)
e = rnorm(500)
w = x + d
y = 2*w + e

df <- data.frame(y, w)
cor(df)
```

```
##           y         w
## y 1.0000000 0.9576021
## w 0.9576021 1.0000000
```

```
summary(df)
```

```
##        y                w
##  Min.   :-4.194   Min.   :-2.222
##  1st Qu.: 2.277   1st Qu.: 1.208
##  Median : 4.964   Median : 2.447
##  Mean   : 4.865   Mean   : 2.444
##  3rd Qu.: 7.613   3rd Qu.: 3.683
##  Max.   :13.975   Max.   : 6.602
```

```
print("Standard Deviation Below (y, w)")
```

```
## [1] "Standard Deviation Below (y, w)"
```

```
print(c(sd(y), sd(w)))
```

```
## [1] 3.622223 1.748330
```

5(c)

```
fit_5c = lm(y ~ x)
summary(fit_5c)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9526 -1.5554 -0.0339  1.5004  7.8134
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.03233    0.21154   0.153    0.879
## x            1.94984    0.07412  26.305   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.346 on 498 degrees of freedom
## Multiple R-squared:  0.5815, Adjusted R-squared:  0.5807
## F-statistic: 691.9 on 1 and 498 DF,  p-value: < 2.2e-16
```

```
confint(fit_5c)
```

```
##                  2.5 %    97.5 %
## (Intercept) -0.3832934 0.4479613
## x            1.8041993 2.0954716
```

The coefficient of x is significant at $\alpha = 0.05$

5(d)

```
fit_5d = lm(y ~ x + w)
summary(fit_5d)
```

```
##
## Call:
## lm(formula = y ~ x + w)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1940 -0.7304 -0.0063  0.7367  3.0967
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.018288   0.094301   0.194    0.846
## x           -0.003888   0.054697  -0.071    0.943
## w            1.986488   0.044319  44.823   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.046 on 497 degrees of freedom
## Multiple R-squared:  0.917,  Adjusted R-squared:  0.9167
## F-statistic:  2746 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
confint(fit_5d)
```

```
##                  2.5 %    97.5 %
## (Intercept) -0.1669907 0.2035660
## x           -0.1113529 0.1035774
## w            1.8994132 2.0735629
```

We can see that x is not significant and w is significant.

Q6 true: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

fit: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \epsilon$

show: $E(\hat{\beta}_1) = \beta_1 + \beta_2 \sum_{i=1}^{n} c_i x_{i2} = \beta_1 + \frac{\beta_2}{S_{11}} \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)$

$= \beta_1 + \beta_2 r \frac{s_2}{s_1}$, where $c_i = (x_{i1} - \bar{x}_1)/S_{11}$, $S_{11} = \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2$

$\hat{\beta}_1$ can be written as: $\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$, where $c_i = (x_{i1} - \bar{x}_1)/S_{11}$,

and c_i satisfies: $\sum_{i=1}^{n} c_i = 0$ and $\sum_{i=1}^{n} c_i^2 = \frac{\sum (x_{i1} - \bar{x}_1)^2}{S_{11}^2} = \frac{1}{S_{11}}$

therefore $E(\hat{\beta}_1) = E(\sum_{i=1}^{n} c_i y_i)$

$$= \sum_{i=1}^{n} c_i E(y_i)$$

$$= \sum_{i=1}^{n} c_i E(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$$

$$= \sum_{i=1}^{n} c_i (\beta_0 + \beta_1 x_1 + \beta_2 x_2)$$

$$= \beta_0 \cancelto{0}{\sum_{i=1}^{n} c_i} + \beta_1 \sum_{i=1}^{n} c_i x_{i1} + \beta_2 \sum_{i=1}^{n} c_i x_{i2}$$

$$= \beta_1 \cancelto{1}{\sum_{i=1}^{n} c_i x_{i1}} + \beta_2 \sum_{i=1}^{n} c_i x_{i2} \quad \text{(see note)}$$

note:

$\sum_{i=1}^{n} c_i x_{i1} = \frac{1}{S_{11}} \sum_{i=1}^{n} (x_{i1} - \bar{x}_1) x_{i1}$

$= \frac{1}{S_{11}} \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i1} - \bar{x}_1)$

$= \frac{1}{S_{11}} \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2$

$= 1$

$$= \beta_1 + \beta_2 \sum_{i=1}^{n} c_i x_{i2}$$

$$\sum_{i=1}^{n} c_i x_{i2} = \frac{1}{S_{11}} \sum_{i=1}^{n} (x_{i1} - \bar{x}_1) x_{i2}$$

$$= \frac{1}{S_{11}} \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)$$

$$\Rightarrow E(\hat{\beta}_1) = \beta_1 + \beta_2 / S_{11} \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)$$

$$r = \frac{\operatorname{cov}(X_{1i}, X_{2i})}{S_1 S_2} = \frac{\sum_{i=1}^{n} (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{S_1 S_2}$$

$$\Rightarrow \sum_{i=1}^{n} (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2) = r \cdot S_1 \cdot S_2$$

$$\Rightarrow E(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{r \cancel{S_1} S_2}{S_1^{\cancel{2}}} = \beta_1 + \beta_2 \cdot r \frac{S_2}{S_1}$$

$\hat{\beta}_1$ is unbiased when $\frac{\beta_2}{S_{11}} \operatorname{Cov}(X_{1i}, X_{2i}) = 0$. Meaning when X_1 and X_2 are not correlated, $\hat{\beta}_1$ is unbiased.

Problem 7.

Given $y = \beta_1 z_1 + \beta_2 z_2 + e$, $Z = \begin{pmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \\ \vdots & \vdots \\ z_{n1} & z_{n2} \end{pmatrix}$ $Z^T = \begin{pmatrix} z_{11} & z_{21} & \cdots & z_{n1} \\ z_{12} & z_{22} & \cdots & z_{n2} \end{pmatrix}$

$$Var(\hat{\beta}) = \sigma^2 (Z^T Z)^{-1}$$

$$= \sigma^2 \begin{pmatrix} \sum_{i=1}^n z_{i1}^2 & \sum_{i=1}^n z_{i1} z_{i2} \\ \sum_{i=1}^n z_{i2} z_{i1} & \sum_{i=1}^n z_{i2}^2 \end{pmatrix}^{-1}$$

$$= \frac{\sigma^2}{\sum_{i=1}^n z_{i1}^2 \sum_{i=1}^n z_{i2}^2 - \sum_{i=1}^n z_{i1} z_{i2} \sum_{i=1}^n z_{i2} z_{i1}} \begin{pmatrix} \sum_{i=1}^n z_{i2}^2 & -\sum_{i=1}^n z_{i1} z_{i2} \\ -\sum_{i=1}^n z_{i2} z_{i1} & \sum_{i=1}^n z_{i1}^2 \end{pmatrix}$$

$\because \frac{1}{n} \sum_{i=1}^n z_{ij} = 0$ $\quad \sum_{i=1}^n z_{ij}^2 = n-1$ $\quad r = \frac{1}{n-1} \sum_{i=1}^n z_{i1} z_{i2} \rightarrow \sum_{i=1}^n z_{i1} z_{i2} = r(n-1)$

Assume z_{i1} and z_{i2} are not correlated, then $r = 0 \rightarrow \sum_{i=1}^n z_{i1} z_{i2} = 0$.

$$Var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^n z_{i1}^2 \sum_{i=1}^n z_{i2}^2} \begin{pmatrix} \sum_{i=1}^n z_{i2}^2 & 0 \\ 0 & \sum_{i=1}^n z_{i1}^2 \end{pmatrix}$$

$$= \frac{\sigma^2}{(n-1)^2} \begin{pmatrix} n-1 & 0 \\ 0 & n-1 \end{pmatrix} = \begin{pmatrix} \frac{\sigma^2}{n-1} & 0 \\ 0 & \frac{\sigma^2}{n-1} \end{pmatrix}$$

If z_{i1} and z_{i2} are correlated, then $r \neq 0$.

$$Var(\hat{\beta}) = \frac{\sigma^2}{(n-1)^2 - [(n-1)r]^2} \begin{bmatrix} n-1 & -(n-1)r \\ -(n-1)r & n-1 \end{bmatrix} = \begin{bmatrix} \frac{\sigma^2}{(n-1)(1-r^2)} & \frac{-\sigma^2}{(n-1)r(1-r^2)} \\ \frac{-\sigma^2}{(n-1)r(1-r^2)} & \frac{\sigma^2}{(n-1)(1-r^2)} \end{bmatrix}$$

$\because 0 \leq r^2 \leq 1$

$\therefore \frac{\sigma^2}{(n-1)(1-r^2)} \geq \frac{\sigma^2}{n-1}$

So variance of $\hat{\beta}$ will be inflated when there's multicollinearity.