STAT 640: Homework 9

Due **Friday, April 8, 11:59pm MT** on the course Canvas webpage. Please follow the homework guidelines on the syllabus.

Name: Hannah Butler

Problem 1

Consider the two regression models in which response variables Y_i, for $i = 1, \ldots, n$, satisfy

$$Y_i = \beta_0 + X_i\beta_1 + \epsilon_i$$
$$Y_i = \gamma_0 + X_i\gamma_1 + Z_i\gamma_2 + \delta_i$$

You may assume:

- The corresponding design matrices are full rank.
- $\mathrm{E}[\epsilon_i] = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$
- $\mathrm{E}[\delta_i] = 0$ and $\mathrm{Var}(\delta_i) = \tau^2$
- The values of X_i and Z_i are fixed and known.

a. Under what condition(s) are the LSEs $\hat{\beta}_1$ and $\hat{\gamma}_1$ equal?

Answer:

When X and Z are independent/orthogonal, or their covariance is zero.

---

b. Under what condition(s) is $\hat{\beta}_1$ BLUE for γ_1?

Answer:

Because the design matrices for both models are full rank, we can assume that the BLUE for γ_1 is unique. So, similar to part a, the BLUE for β_1 will be the BLUE for γ_1 if the covariance between X and Z is zero.

---

c. Suppose that $\gamma_1 = 0$ and $\gamma_2 \neq 0$. What is the impact of this situation on the distribution of $\hat{\beta}_1$, and how would $\hat{\beta}_1$ compare to $\hat{\gamma}_1$?

Answer:

In this situation, we take the second model to be the true model, in which case, the first model, containing X as the only regressor would be underfit. We have not made any distributional assumptions, so we don't know what the distribution of the LSE would be, but we can see that the expectation of $\hat{\beta}_1$ will be biased since

$$E\hat{\beta}_1 = E((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E\boldsymbol{Y} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z}\gamma_2.$$

Had we fit the latter model, we would have

$$E\hat{\boldsymbol{\gamma}} = (\begin{bmatrix}\boldsymbol{X} & \boldsymbol{Z}\end{bmatrix}^T\begin{bmatrix}\boldsymbol{X} & \boldsymbol{Z}\end{bmatrix})^{-1}\begin{bmatrix}\boldsymbol{X} & \boldsymbol{Z}\end{bmatrix}^T\begin{bmatrix}\boldsymbol{X} & \boldsymbol{Z}\end{bmatrix}\boldsymbol{\gamma} = \boldsymbol{\gamma},$$

So that $\hat{\gamma}_1$ is an unbiased estimator of γ_1. We can similarly compare the variances of the two. The variance of $\hat{\beta}_1$, had we underfit the model, would be

$$\text{Var}(\hat{\beta}_1) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2\boldsymbol{I}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)^T = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

versus fitting model 2,

$$\text{Var}(\hat{\gamma}_1) = \tau^2(\begin{bmatrix}\boldsymbol{X} & \boldsymbol{Z}\end{bmatrix}^T\begin{bmatrix}\boldsymbol{X} & \boldsymbol{Z}\end{bmatrix})^{-1} = \tau^2\begin{bmatrix}\boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{Z}\end{bmatrix}^{-1}$$

So the variance of $\hat{\gamma}_1$ would be

$$\begin{aligned}\text{Var}(\hat{\gamma}_1) &= \left(\tau^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right) \\ &= \tau^2\left((\boldsymbol{X}^T\boldsymbol{X})^{-1} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z}(\boldsymbol{Z}^T(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)\end{aligned}$$

---

d. Repeat (c), but now with $\gamma_1 \neq 0$ and $\gamma_2 = 0$.

Answer:

---

e. A colleague wants to know the impact of X_i on Y_i. Which model would you recommend they fit, and why?

Answer:

Problem 2

For this question, the phrase "p-value of the jth variable in the linear model" refers to the p-value associated with the F-test (or equivalent t-test) for $H_0 : \beta_j = 0$.
Create two sets (i.e. one for (a) and one for (b)) of p variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ and a response vector $\boldsymbol{y}$ of length n such that the following properties hold:

a. The jth variable has a p-value less than 0.05 in the linear model involving $\boldsymbol{y}$ and all p features, but its p-value increases above 0.05 if the kth variable is removed from the model.

b. The jth variable has a p-value above 0.05 in the linear model involving $\boldsymbol{y}$ and all p features, but its p-value decreases below 0.05 if the kth variable is removed from the model.

You can choose any value of $p \geq 2$, and any value of n, in constructing solutions to (a) and (b). For (a) and (b), explain how you constructed $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ and $\boldsymbol{y}$ to achieve the desired property, and why this property holds. In other words, do not just provide an answer, but explain why your answer works.

Answers:

For part a, I visualized the space spanned by columns of X as a flat 2D surface. It can be seen that for Y very close to the plane would be well explained by both vectors, but not as well explained by any single vector.

part b: correlation between X1 and X2 is greater than the correlation between X2 and Y and the correlation between X1 and Y. Here, there is a causal relationship between X1 and X2, but a weaker (or no) causal relationship between either of X1 or X2 and Y.

```
set1 <- data.frame(Y = c(1, 1, .1)
           , X1 = c(1,0,0)
           , X2 = c(0,1,0)
           )
summary(lm(Y ~ 0 + X1 + X2, set1))
```

```
##
## Call:
## lm(formula = Y ~ 0 + X1 + X2, data = set1)
##
## Residuals:
##   1   2   3
## 0.0 0.0 0.1
##
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)
## X1      1.0        0.1      10   0.0635 .
## X2      1.0        0.1      10   0.0635 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1 on 1 degrees of freedom
## Multiple R-squared:  0.995,  Adjusted R-squared:  0.9851
## F-statistic:   100 on 2 and 1 DF,  p-value: 0.07053
```

```r
summary(lm(Y ~ 0 + X1, set1))
```

```
##
## Call:
## lm(formula = Y ~ 0 + X1, data = set1)
##
## Residuals:
##   1   2   3
## 0.0 1.0 0.1
##
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)
## X1   1.0000     0.7106   1.407    0.295
##
## Residual standard error: 0.7106 on 2 degrees of freedom
## Multiple R-squared:  0.4975, Adjusted R-squared:  0.2463
## F-statistic:  1.98 on 1 and 2 DF,  p-value: 0.2947
```

```r
summary(lm(Y ~ 0 + X2, set1))
```

```
##
## Call:
## lm(formula = Y ~ 0 + X2, data = set1)
##
## Residuals:
##   1   2   3
## 1.0 0.0 0.1
##
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)
## X2   1.0000     0.7106   1.407    0.295
##
## Residual standard error: 0.7106 on 2 degrees of freedom
## Multiple R-squared:  0.4975, Adjusted R-squared:  0.2463
## F-statistic:  1.98 on 1 and 2 DF,  p-value: 0.2947
```

```r
set.seed(80085)
X1 <- 1:10
X2 <- X1 + rnorm(10, 0, 1)
Y <- X2 + rnorm(10, 0, 2)
set2 <- data.frame(Y, X1, X2)

set2 %>%
  summarize(X1xX2 = cor(X1, X2)
            , X1xY = cor(X1, Y)
            , X2xY = cor(X2, Y))
```

```
##      X1xX2      X1xY     X2xY
## 1 0.993223 0.7857318 0.771203
```

```r
summary(lm(Y ~ 0 + X1 + X2, set2))
```

```
##
## Call:
## lm(formula = Y ~ 0 + X1 + X2, data = set2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6895 -0.6168 -0.3293  1.1079  4.4205
##
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)
## X1    2.146      1.630   1.316    0.225
## X2   -1.189      1.592  -0.747    0.477
##
## Residual standard error: 1.872 on 8 degrees of freedom
## Multiple R-squared:  0.9229, Adjusted R-squared:  0.9036
## F-statistic: 47.86 on 2 and 8 DF,  p-value: 3.539e-05
```

```r
summary(lm(Y ~ 0 + X1, set2))
```

```
##
## Call:
## lm(formula = Y ~ 0 + X1, data = set2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2694 -0.8289  0.0908  1.0942  4.6625
##
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)
## X1  0.93070    0.09303      10 3.57e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.825 on 9 degrees of freedom
## Multiple R-squared:  0.9175, Adjusted R-squared:  0.9083
## F-statistic: 100.1 on 1 and 9 DF,  p-value: 3.565e-06
```

```r
summary(lm(Y ~ 0 + X2, set2))
```

```
##
## Call:
## lm(formula = Y ~ 0 + X2, data = set2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7329 -0.3310  0.1776  1.2196  4.8537
##
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)
```

```
## X2  0.90352    0.09691   9.323 6.39e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.947 on 9 degrees of freedom
## Multiple R-squared:  0.9062, Adjusted R-squared:  0.8957
## F-statistic: 86.92 on 1 and 9 DF,  p-value: 6.394e-06
```

Problem 3

This question asks you to perform a simulation to compare the variability of slope estimates in simple linear regression under three different data generating models for the relationship between Y and x, and three different study designs.

Simulate 2000 data sets, each with 120 observations, using the same seed for each of the nine combinations of:
3 models:

- Linear model: $Y = x + \epsilon$, where $\epsilon \sim N(0, (0.4)^2)$
- Linear model with heteroscedasticity: $Y = x + \epsilon$, where $\epsilon \sim N(0, (0.4)^2 e^{|x|})$
- Quadratic model: $Y = x^2 + \epsilon$, where $\epsilon \sim N(0, (0.4)^2)$

3 designs for x:

- Observational data, continuous distribution: $x \sim N(0, 1)$
- Observational data, discrete distribution: $x \sim$ a shifted and scaled trinomial distribution, where $\Pr[x = -\sqrt{2}] = .25, \Pr[x = 0] = .5, \Pr[x = \sqrt{2}] = .25$.
- Designed experimental data, where there are
  30 observations at x = $-\sqrt{2}$
  60 observations at x = 0
  30 observations at x = $\sqrt{2}$

Note that for the two discrete distributions, x takes on the same three possible values, and for all three distributions, the variance of x is one.

For each replication in each of the nine settings, fit the linear regression model

$$\mathrm{E}[Y] = \beta_0 + \beta_1 x$$

and obtain: the estimate $\hat{\beta}_1$, a model-based standard error estimate for $\hat{\beta}_1$, and a sandwich standard error estimate for $\hat{\beta}_1$. Note the model being fit is the correct model for 2/3 of the settings, and it is a misspecified model for 1/3 of the settings.

a. Complete the following table summarizing the simulation (round to 3 decimal places):

Model	Design	$\mathrm{E}(\hat{\beta}_1)$	$\mathrm{SD}(\hat{\beta}_1)$	$\widehat{\mathrm{SE}}_{Model}(\hat{\beta}_1)$	$\widehat{\mathrm{SE}}_{Sand}(\hat{\beta}_1)$
Linear	Observational, continuous Observational, categorical Designed experiment				
Linear Heteroscedastic	Observational, continuous Observational, categorical Designed experiment				
Quadratic	Observational, continuous Observational, categorical Designed experiment				

Answer:

```
nsim <- 2000
obs <- 120

set.seed(80085)
e1 <- replicate(2000, rnorm(120, 0, 0.4))
X1 <- replicate(2000, rnorm(120, 0,1))
X1_e2 <- lapply(1:2000, function(x) {rnorm(120, 0, .4*sqrt(exp(abs(X1[,x]))))}) %>% do.call(cbind, .)

X2 <- replicate(2000, sample(c(-sqrt(2), sqrt(2), 0, 0), 120, T))
X2_e2 <- lapply(1:2000, function(x) {rnorm(120, 0, .4*sqrt(exp(abs(X2[,x]))))}) %>% do.call(cbind, .)

X3 <- replicate(2000, sample(rep(c(-sqrt(2), sqrt(2), 0), times = c(30, 30, 60)), 120, F))
X3_e2 <- lapply(1:2000, function(x) {rnorm(120, 0, .4*sqrt(exp(abs(X3[,x]))))}) %>% do.call(cbind, .)

com1 <- lapply(1:2000, function(x) {
  data.frame(X = X1[,x], Y = X1[,x] + e1[,x])
  })
com2 <- lapply(1:2000, function(x) {
  data.frame(X = X2[,x], Y = X2[,x] + e1[,x])
  })
com3 <- lapply(1:2000, function(x) {
  data.frame(X = X3[,x], Y = X3[,x] + e1[,x])
  })
com4 <- lapply(1:2000, function(x) {
  data.frame(X = X1[,x], Y = X1[,x] + X1_e2[,x])
  })
com5 <- lapply(1:2000, function(x) {
  data.frame(X = X2[,x], Y = X2[,x] + X2_e2[,x])
  })
com6 <- lapply(1:2000, function(x) {
  data.frame(X = X3[,x], Y = X3[,x] + X3_e2[,x])
  })
com7 <- lapply(1:2000, function(x) {
  data.frame(X = X1[,x]^2, Y = X1[,x]^2 + e1[,x])
  })
com8 <- lapply(1:2000, function(x) {
  data.frame(X = X2[,x]^2, Y = X2[,x]^2 + e1[,x])
```

```
  })
com9 <- lapply(1:2000, function(x) {
  data.frame(X = X3[,x]^2, Y = X3[,x]^2 + e1[,x])
  })

df_list <- list(com1, com2, com3, com4, com5, com6, com7, com8, com9)
fits <- list()
for(i in 1:9) {
  fit <- lapply(df_list[[i]], function(x) {
    m <- summary(lm(Y ~ X, x))$coefficients[2, 1:2]
    }) %>% do.call(rbind, .)
  fits[[i]] <- fit
}

lapply(fits, colMeans)
```

```
## [[1]]
##   Estimate Std. Error
## 1.00092901 0.03687748
##
## [[2]]
##   Estimate Std. Error
##  1.0013423  0.0367334
##
## [[3]]
##   Estimate Std. Error
## 1.00051261 0.03646507
##
## [[4]]
##   Estimate Std. Error
## 1.00025275 0.06058481
##
## [[5]]
##   Estimate Std. Error
##  0.9993030  0.0582655
##
## [[6]]
##   Estimate Std. Error
## 1.00052348 0.05818191
##
## [[7]]
##   Estimate Std. Error
## 1.00029335 0.02701651
##
## [[8]]
##   Estimate Std. Error
## 1.00084077 0.03661136
##
## [[9]]
##   Estimate Std. Error
## 1.00017101 0.03646364
```

```r
lapply(fits, function(x) {
  apply(x, 2, sd)
})
```

```
## [[1]]
##   Estimate Std. Error
## 0.03618072 0.00342924
##
## [[2]]
##    Estimate  Std. Error
## 0.037781935 0.002967671
##
## [[3]]
##    Estimate  Std. Error
## 0.035527188 0.002393304
##
## [[4]]
##    Estimate  Std. Error
## 0.089948400 0.005362022
##
## [[5]]
##    Estimate  Std. Error
## 0.073563108 0.004754101
##
## [[6]]
##    Estimate  Std. Error
## 0.072833966 0.004504773
##
## [[7]]
##    Estimate  Std. Error
## 0.027381403 0.004934859
##
## [[8]]
##    Estimate  Std. Error
## 0.035728018 0.002415461
##
## [[9]]
##    Estimate  Std. Error
## 0.035722665 0.002390785
```

---

b. Compare the slope standard deviations for each of the nine settings. What trends and differences do you see?

Answer:

---

c. How well does the model-based standard error estimate the true variability of the coefficient estimate in each of these settings, with this sample size?

Answer:

d How well does the sandwich standard error estimate the true variability of the coefficient estimate in each of these settings, with this sample size?

Answer: