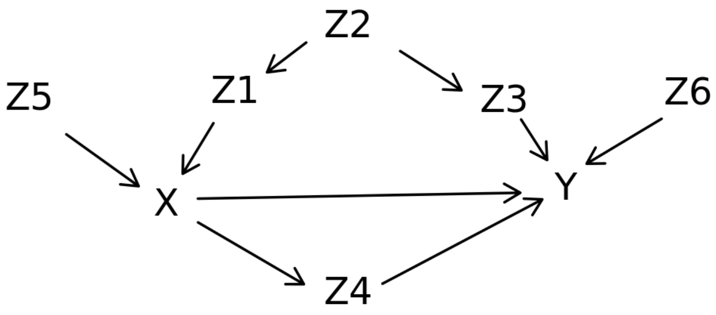

1. Consider the linear DAG above and the following models:
   1. `Y ~ X`
   2. `Y ~ X + Z6`
   3. `Y ~ X + Z1`
   4. `Y ~ X + Z1 + Z4`
   5. `Y ~ X + Z1 + Z3`
   6. `Y ~ X + Z3 + Z6`
   7. `Y ~ X + Z1 + Z5`
   8. `Y ~ X`

a) [10] For each of these models discuss briefly whether fitting the model would produce an unbiased estimate of the causal effect of `X`.

b) [10] Among the models that provide an unbiased estimate of the causal effect of `X`, order them, to the extent possible from the information in the DAG, according to the expected standard deviation of $\hat{\beta}_X$. Briefly state the basis for your ordering.

c) [10] Are there reasons why you might prefer to use a model that the DAG would identify as having a larger standard deviation of $\hat{\beta}_X$?

2. [10] Consider a multiple regression of the form $Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2 I)$ and X_1, X_2 represent blocks of variables such that the matrix $[X_1 X_2]$ is of full column rank.
Prove that the Added Variable Plot for the regression of Y on X_1 has the same vector of least-squares coefficients as the least-squares coefficients for X_1 in the multiple regression.

3. [10] Consider the following statement:
   "In a multiple regression, if you add a predictor whose effect is not significant, the coefficients of the other predictors should not have changed very much, nor should the p-values associated with them."
   Is this a valid statement? If so, discuss why, illustrating your answer with appropriate figures.

4. [10] Are there any situations in which it would be important to drop a term in a model although its coefficient is highly statistically significant. Discuss the circumstances, if any, in which this would be true, and the consequences of including or excluding the variable in question.

5. This question is based on the 'Vocab' data set used in Assignment 3. Recall that it records a vocabulary score for over 30,000 subjects tested over the years between 1974 and 2016. The questions below refer to the following output in R. Assume that modeling the effect of 'year' with a linear term is adequate to describe the relationships in the data.

```
summary(Vocab)
```

```
      year          sex          education       vocabulary
 Min.   :1974   Female:17148   Min.   : 0.00   Min.   : 0.000
 1st Qu.:1987   Male  :13203   1st Qu.:12.00   1st Qu.: 5.000
 Median :1994                  Median :12.00   Median : 6.000
 Mean   :1995                  Mean   :13.03   Mean   : 6.004
 3rd Qu.:2006                  3rd Qu.:15.00   3rd Qu.: 7.000
 Max.   :2016                  Max.   :20.00   Max.   :10.000
```

```
fit <- lm(vocabulary ~ year * sex, Vocab)
summary(fit)
```

```

Call:
lm(formula = vocabulary ~ year * sex, data = Vocab)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.0848 -1.0590 -0.0145  1.0799  4.1114 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)  
(Intercept)   1.7517998  2.5541242   0.686   0.4928  
year          0.0021493  0.0012801   1.679   0.0932 .
sexMale      -2.0788248  3.8406372  -0.541   0.5883  
year:sexMale  0.0009994  0.0019247   0.519   0.6036  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.115 on 30347 degrees of freedom
Multiple R-squared:  0.0006394,	Adjusted R-squared:  0.0005406 
F-statistic: 6.472 on 3 and 30347 DF,  p-value: 0.000225
```

```
fit2 <- lm(vocabulary ~ year + sex, Vocab)
summary(fit2)
```

```

Call:
lm(formula = vocabulary ~ year + sex, data = Vocab)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.0940 -1.0629 -0.0059  1.0735  4.0995 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.8697866  1.9073945   0.456  0.64839    
year         0.0025914  0.0009559   2.711  0.00672 ** 
sexMale     -0.0846035  0.0244833  -3.456  0.00055 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.115 on 30348 degrees of freedom
Multiple R-squared:  0.0006305,    Adjusted R-squared:  0.0005646
F-statistic: 9.573 on 2 and 30348 DF,  p-value: 6.979e-05
```

1. [10] The output above shows the result of fitting two models predicting vocabulary using ‘year’ and ‘sex’. One model does not show evidence (in the sense of achieving p-values less than 0.05) for interactions and for main effects of sex and year. The other does show evidence of effects of sex and year. Explain the apparent discrepancies in the output of these two models. [Hint: If you feel puzzled by this question, you might like to try to complete the next part first.]
2. [10] Using the model with an interaction term, estimate the gender gap in vocabulary in the year 2000. Describe briefly how you would go about testing whether this gap is ‘statistically significant’.

6. [10] Write a function in R that takes a matrix as input and returns the index of the column that has the largest sum of squares.

7. [10] Many researchers who find that a hypothesis test of a particular null hypothesis has achieved a p-value of 0.04 have the impression that there is strong evidence against the null hypothesis and it is 'unlikely to be correct.' Discuss whether this is a correct interpretation of the p-value.

END OF EXAM