Medical Statistics – Answers lab 8

Part 1: Building prediction models using backward elimination

Step 1: Fit the initial linear regression model

Create an initial model for hospital length of stay (`los`) using the following predictors: `age`, `gender`, `hr`, `sysbp`, `diasbp`, `bmi`, `cvd`, `sho`. Run/summarize the model to inspect coefficients and p-values.

```
Call:
lm(formula = los ~ age + gender + hr + sysbp + diasbp + bmi + 
    cvd + sho, data = whas500)

Residuals:
   Min     1Q Median     3Q    Max 
-8.335 -2.653 -1.071  1.200 40.073 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.2648144  2.1394896   1.526 0.127659    
age           0.0031994  0.0168850   0.189 0.849792    
genderfemale  0.8575246  0.4489334   1.910 0.056698 .  
hr            0.0190577  0.0090828   2.098 0.036396 *  
sysbp        -0.0008358  0.0085656  -0.098 0.922304    
diasbp        0.0141799  0.0128884   1.100 0.271780    
bmi          -0.0304532  0.0427613  -0.712 0.476700    
cvdyes        0.3903925  0.4981468   0.784 0.433600    
shoyes        3.5265170  1.0329265   3.414 0.000693 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.632 on 491 degrees of freedom
Multiple R-squared:  0.04991,   Adjusted R-squared:  0.03443
F-statistic: 3.224 on 8 and 491 DF,  p-value: 0.001388
```

Step 2: Eliminate the least significant predictor

```
Anova Table (Type III tests)

Response: los
             Sum Sq  Df F value    Pr(>F)
(Intercept)    50.0   1  2.3286 0.1276592
age             0.8   1  0.0359 0.8497924
gender         78.3   1  3.6486 0.0566978 .
hr             94.5   1  4.4025 0.0363962 *
sysbp           0.2   1  0.0095 0.9223042
diasbp         26.0   1  1.2105 0.2717799
bmi            10.9   1  0.5072 0.4766997
cvd            13.2   1  0.6142 0.4336001
sho           250.1   1 11.6561 0.0006929 ***
Residuals   10535.8 491
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table shows that the predictor with the highest p-value is `sysbp` ($p = 0.92$). Systolic blood pressure is the least significant predictor and should be removed from the model.

Step 3: Repeat the steps

The following variables are sequentially removed from the model (after initially removing `sysbp`):

1. `age`: p-value = 0.86
2. `cvd`: p-value = 0.41
3. `bmi`: p-value = 0.44
4. `diasbp`: p-value = 0.22

Step 4: Final model

```r
final_model <- lm(los ~ gender + hr + sho, data = whas500)
summary(final_model)
```

```

Call:
lm(formula = los ~ gender + hr + sho, data = whas500)

Residuals:
   Min     1Q Median     3Q    Max 
-8.663 -2.661 -1.064  1.136 40.826 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.796075   0.799047   4.751 2.66e-06 ***
genderfemale 0.910388   0.424827   2.143 0.032602 *  
hr           0.020680   0.008843   2.339 0.019751 *  
shoyes       3.550448   1.009081   3.518 0.000474 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.622 on 496 degrees of freedom
Multiple R-squared:  0.04443,	Adjusted R-squared:  0.03865 
F-statistic: 7.687 on 3 and 496 DF,  p-value: 4.976e-05
```

```r
# 95% CIs for regression coefficients
confint(final_model)
```

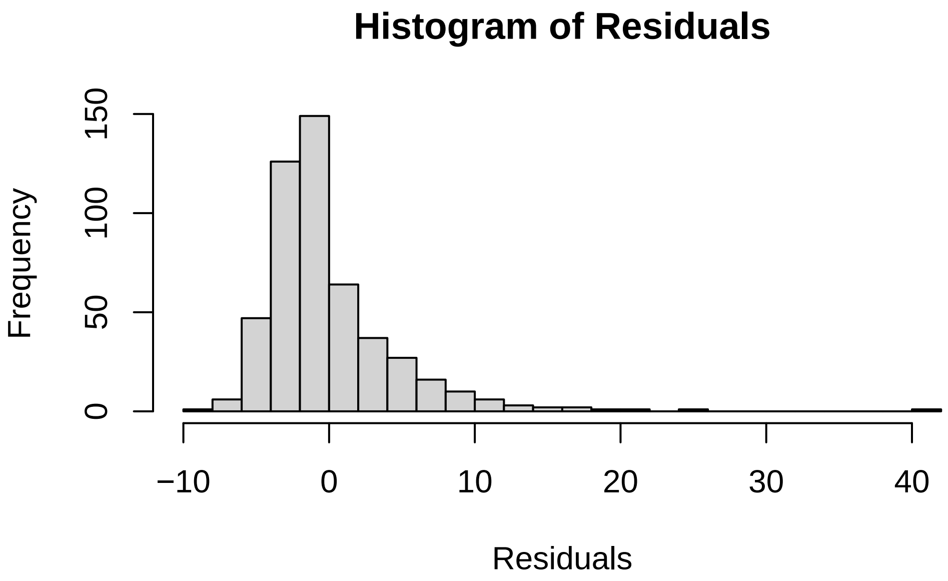
```
                   2.5 %     97.5 %
(Intercept)  2.226140295 5.36601012
genderfemale 0.075705855 1.74507053
hr           0.003306032 0.03805437
shoyes       1.567847029 5.53304829
```

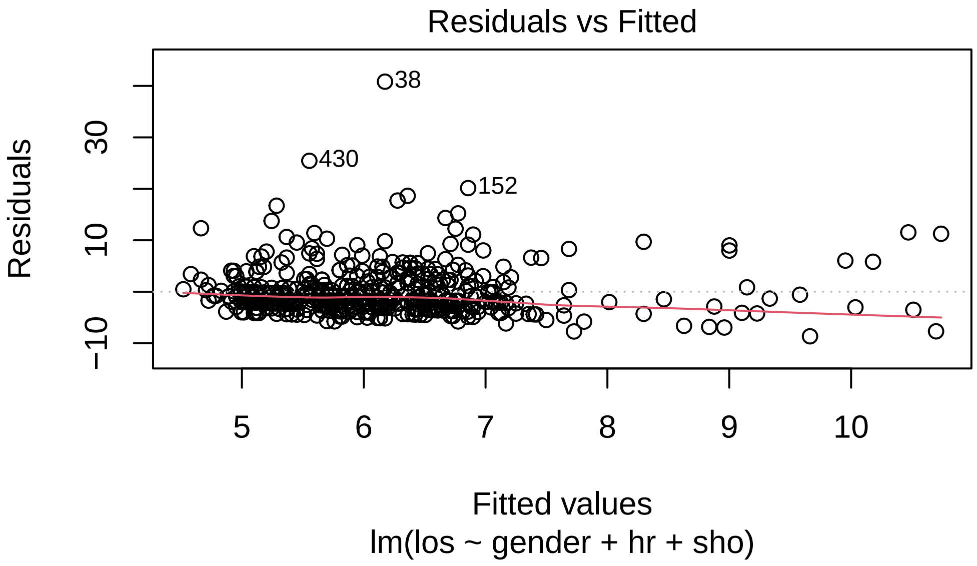
The final model for hospital length of stay includes `gender`, `hr`, and `sho`. All predictors have $p < 0.10$.

Predictor	Coefficient	95% CI	p-value
(Intercept)	3.80	[2.23, 5.37]	< 0.001
Gender (Female)	0.91	[0.08, 1.75]	0.033

Predictor	Coefficient	95% CI	p-value
Heart rate	0.02	[0.003, 0.038]	0.020
Cardiogenic shock	3.55	[1.57, 5.53]	< 0.001

Residual plots





The overall fit of the model appears reasonable, as the residuals are generally centered around zero with no major patterns suggesting severe violations of linearity. However, there are some outliers with very long lengths of stay (LOS) that are not adequately captured by the model. These outliers lead to a right skew in the residual distribution, as seen in the histogram, influencing model fit. While the current model seems to work reasonably well for most observations, further steps (e.g., transformations or robust regression techniques) could be considered to better account for these extreme cases.

Part 2: Automated procedures for building prediction models (logistic regression)

In this part, we explore automated procedures for predictor selection in **logistic regression** prediction models, focusing on predicting in-hospital death (`dstat`).

R

```r
# Create a 0/1 outcome (1 = dead)
whas500 <- whas500 |>
  mutate(dstat01 = ifelse(dstat == "dead", 1, 0))

fit_full <- glm(
  dstat01 ~ age + gender + hr + sysbp + diasbp + bmi + cvd + sho,
  family = binomial,
  data = whas500
)

fit_step <- stepAIC(fit_full, direction = "backward", trace = FALSE)
summary(fit_step)
```

```

Call:
glm(formula = dstat01 ~ age + hr + sysbp + sho, family = binomial,
    data = whas500)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -6.186164   1.687679  -3.665 0.000247 ***
age          0.059272   0.016933   3.500 0.000464 ***
hr           0.013961   0.007497   1.862 0.062572 .  
sysbp       -0.017333   0.006212  -2.790 0.005271 ** 
shoyes       3.061710   0.525719   5.824 5.75e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 273.86  on 499  degrees of freedom
Residual deviance: 204.09  on 495  degrees of freedom
AIC: 214.09

Number of Fisher Scoring iterations: 6
```

```r
# Odds ratios and 95% CIs
exp(cbind(OR = coef(fit_step), confint.default(fit_step)))
```

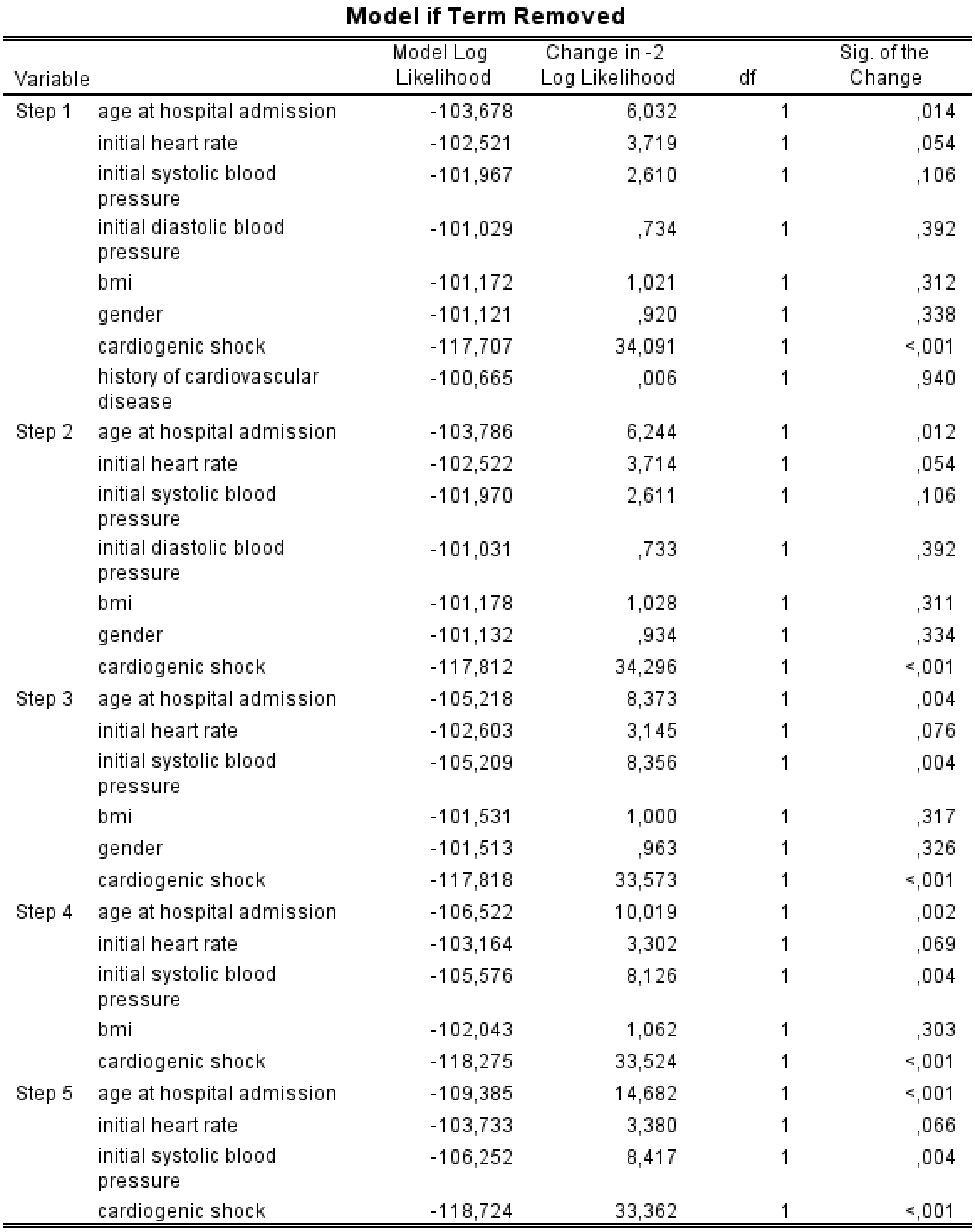
```
                      OR        2.5 %      97.5 %
(Intercept)  0.002057705 7.530576e-05  0.05622609
age          1.061063592 1.026428e+00  1.09686811
hr           1.014058710 9.992675e-01  1.02906886
sysbp        0.982816468 9.709221e-01  0.99485659
shoyes      21.364054186 7.624145e+00 59.86543899
```

Question: Which predictors are retained in the prediction model?

Answer:
The final logistic regression model obtained via backward elimination using AIC includes the following predictors: **age**, **hr** (heart rate), **sysbp** (systolic blood pressure), and **sho** (cardiogenic shock).

SPSS

In SPSS, using the **Backward: LR** method (standard settings), the procedure also yields a final model with these four predictors.



Model if Term Removed

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	age at hospital admission	-103,678	6,032	1	,014
	initial heart rate	-102,521	3,719	1	,054
	initial systolic blood pressure	-101,967	2,610	1	,106
	initial diastolic blood pressure	-101,029	,734	1	,392
	bmi	-101,172	1,021	1	,312
	gender	-101,121	,920	1	,338
	cardiogenic shock	-117,707	34,091	1	<,001
	history of cardiovascular disease	-100,665	,006	1	,940
Step 2	age at hospital admission	-103,786	6,244	1	,012
	initial heart rate	-102,522	3,714	1	,054
	initial systolic blood pressure	-101,970	2,611	1	,106
	initial diastolic blood pressure	-101,031	,733	1	,392
	bmi	-101,178	1,028	1	,311
	gender	-101,132	,934	1	,334
	cardiogenic shock	-117,812	34,296	1	<,001
Step 3	age at hospital admission	-105,218	8,373	1	,004
	initial heart rate	-102,603	3,145	1	,076
	initial systolic blood pressure	-105,209	8,356	1	,004
	bmi	-101,531	1,000	1	,317
	gender	-101,513	,963	1	,326
	cardiogenic shock	-117,818	33,573	1	<,001
Step 4	age at hospital admission	-106,522	10,019	1	,002
	initial heart rate	-103,164	3,302	1	,069
	initial systolic blood pressure	-105,576	8,126	1	,004
	bmi	-102,043	1,062	1	,303
	cardiogenic shock	-118,275	33,524	1	<,001
Step 5	age at hospital admission	-109,385	14,682	1	<,001
	initial heart rate	-103,733	3,380	1	,066
	initial systolic blood pressure	-106,252	8,417	1	,004
	cardiogenic shock	-118,724	33,362	1	<,001

Figure 1: SPSS results of the backward selection for logistic regression

Part 3: Causal diagrams

For each of the exercises below:

- Try solving the diagrams by hand by using the recipe from the lecture (see lecture slides on Brightspace)
- Check your answer using the DAGitty webtool

Exercise 1

In the graph depicted below, for which variables do you need to adjust to assess the unconfounded effect of E on O (there may be several possibilities)?

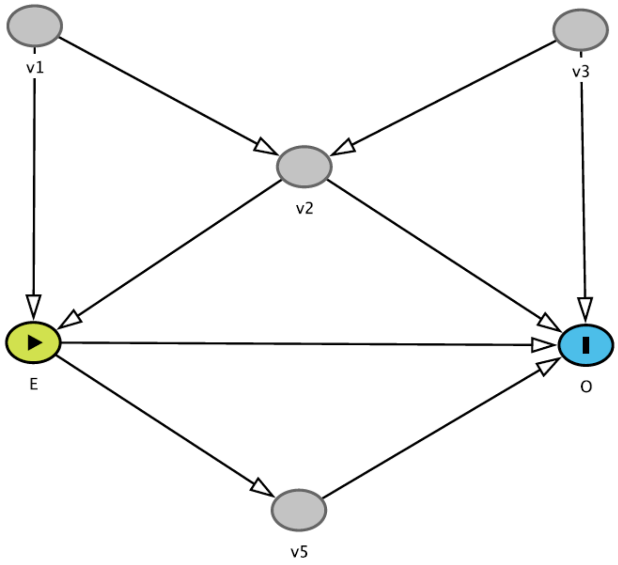


Figure 2: DAG exercise 1

Answer:
Following the recipe: after removing all arrows leaving E, there are several unblocked paths leading from E to O. Just like in the lecture, adjusting for v2 opens a backdoor path (E – v1 – v3 – O) This newly opened backdoor path needs to be closed by also conditioning on v1 or v3, or both. Hence, there are 3 options: (v1, v2, v3) ; (v1, v2) ; and finally, (v2, v3).

Exercise 2

In the graph depicted below, what happens when you additionally adjust for **v5**?

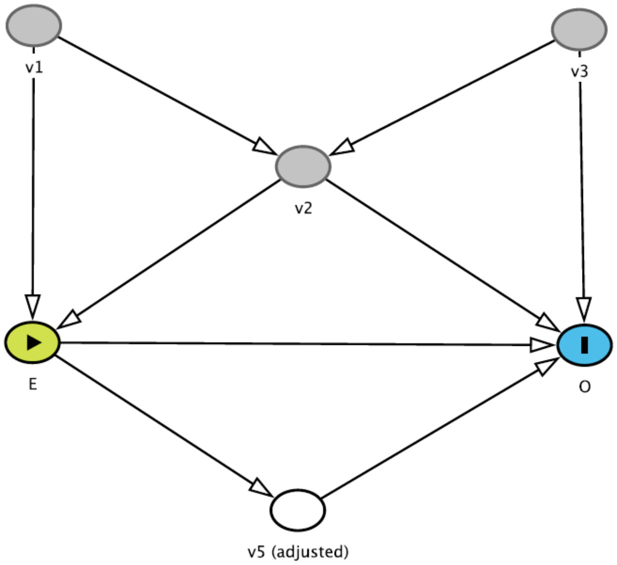


Figure 3: DAG exercise 2

Answer: When adjusting for v5, we are blocking the effect through this indirect path from E to O (v5 is a mediator between E and O). Instead of the total effect of E on O, we will be estimating the direct effect.

In DAGitty, when you set v5 to 'adjusted', the algorithm will say the following: "The total effect cannot be estimated due to adjustment for an intermediate or a descendant of an intermediate."

Exercise 3

This diagram is slightly different: **v1** now is the exposure. For which variables do you need toadjust to assess the unconfounded effect of **v1** on **O**?

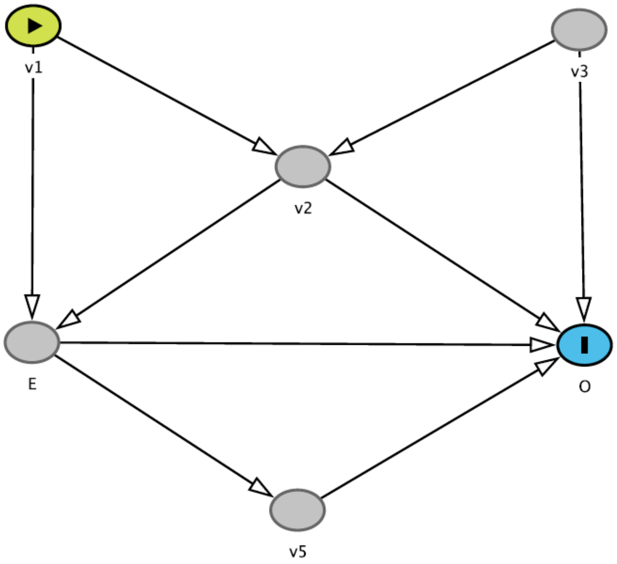


Figure 4: DAG exercise 3

Answer: No adjustment is needed: there are no backdoor paths (removing all arrows leaving v1 reveals no remaining unblocked path from v1 to O).

Exercise 4

Now, **v2** is the exposure. For which variables do you need to adjust to assess the total unconfounded effect of **v2** on **O**?

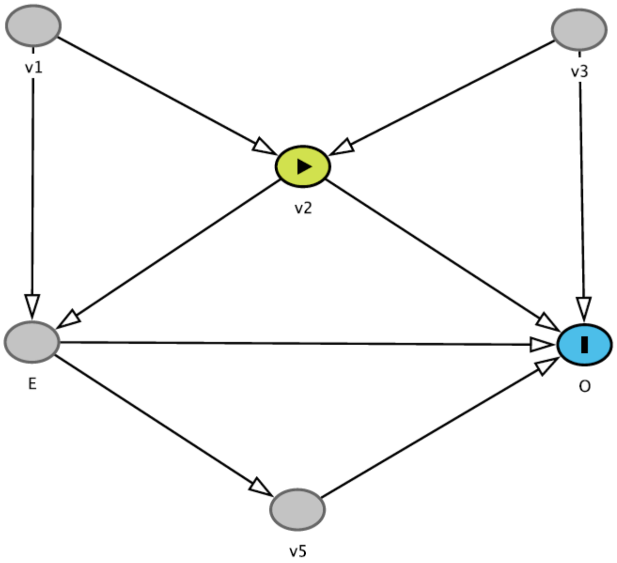


Figure 5: DAG exercise 4

Answer: Following the recipe, there are three unblocked paths left after removing the arrows leaving v2:

a) v2 – v3 – O and
b) v2 – v1 – E – O
c) v2 - v1 – E -v5 - O

Backdoor path a) can be closed by conditioning on v3.

Backdoor path b) can be closed by conditioning on v1 (but not by conditioning on E, as you would no longer be estimating the total effect by blocking the paths from v2 to O mediated by E).

In this case, you should therefore condition on v1 and v3.

Exercise 5

Back to the first DAG. However, **v2** is now unmeasured. Can we still obtain an unconfounded estimate of the effect of **E** on **O**?

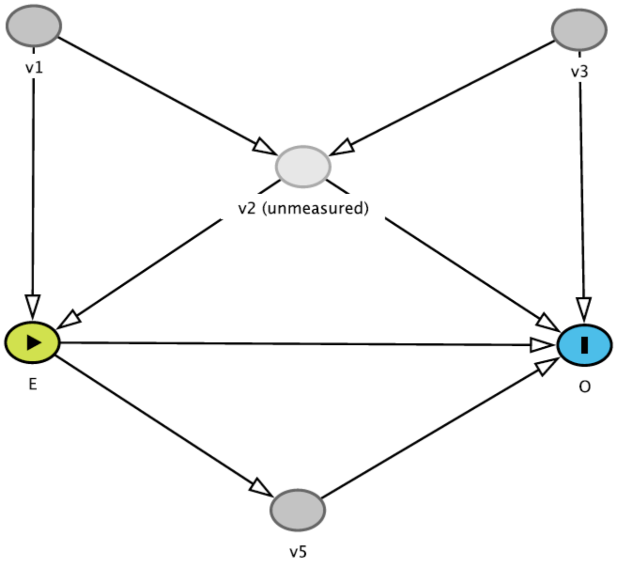


Figure 6: DAG exercise 5

Answer: No, we cannot close the backdoor path between E and O since v2 is unmeasured and cannot be corrected for.

Exercise 6

See the DAG below: you adjusted for **v5**. What would be the consequence of this action?

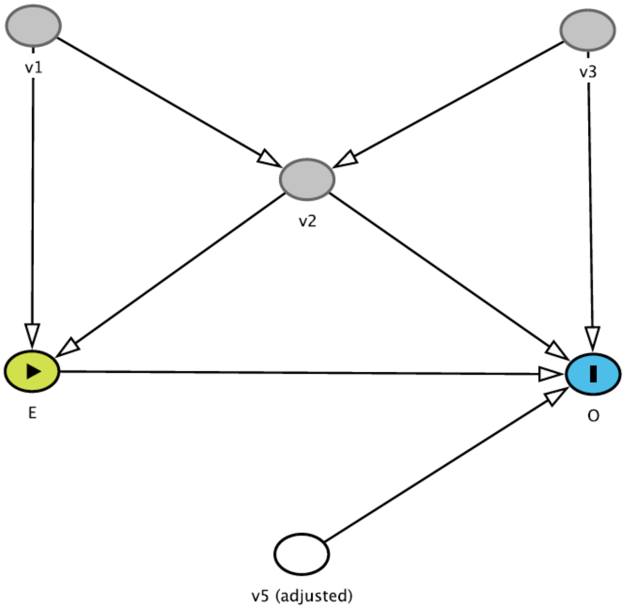


Figure 7: DAG exercise 6

Answer: There is no consequence: conditioning on v5 cannot alter any of the estimated effects in the DAG (it is neither a confounder, collider, nor a mediator in the E-O relationship).