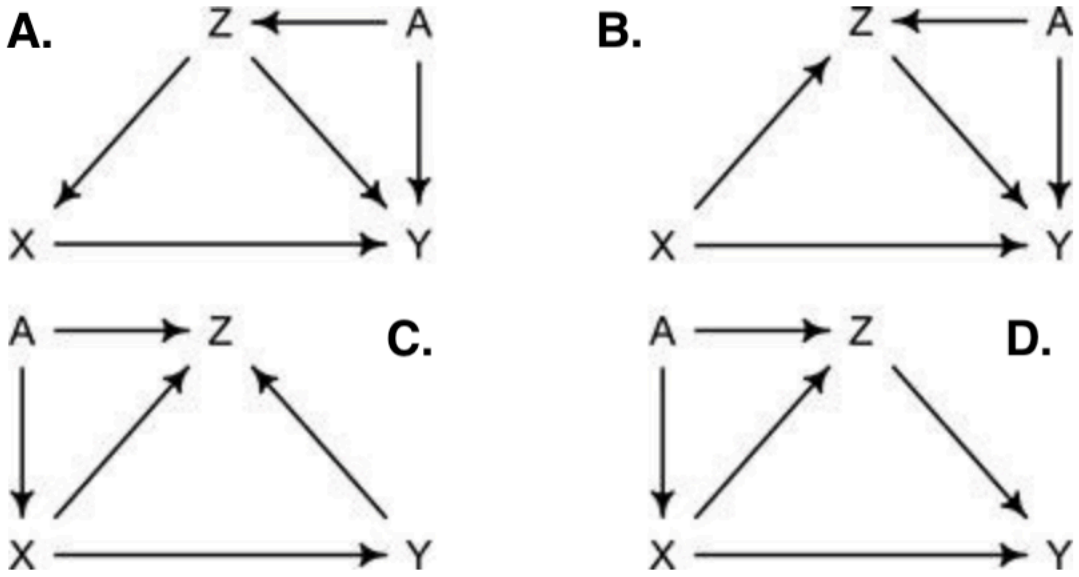
Important Note: *All solutions to the problems below must use the approaches taught so far in the course for answering the questions. There are other approaches for solving these problems that do not require the use of Bayesian approaches, PyMC, quadratic approximation, etc. However, we will soon encounter problems where those tools lack the functionality that is needed to solve the problems. It will be to your benefit to practice using the approach shown in lecture and found in the textbook to begin practicing for what is to come later. In addition, you will not receive full credit on your answers if you do not use the techniques being taught in the course.*

Question 1 (6 points). For each of the four DAGs below, state which variables (**if any**) you must adjust for/condition on/stratify by (all of these terms are equivalent) to estimate the total causal influence of X on Y.

Be sure to **provide justification** for your response to each part based on identification of elementary confounds and using the backdoor criterion. **Limit your responses to each part to no more than 2 sentences.**



Note: Feel free to use https://dagitty.net/dags.html to confirm your adjustment set but understand that you will not receive full credit without justifying your responses.

A)

B)

C)

D)

A) Stratifying by Z would remove all confounds leading into the two target variables X and Y. The backdoor path into X is the arrow from Z to X and stratifying by Z would block that path.

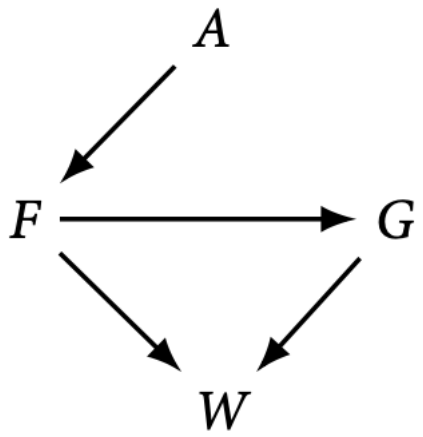
B) There are no paths leading into X, and by backdoor criterion there are no paths to be closed.

C) Stratifying by A would remove the confounds between X and Y, it closes the path leading from A into X. Z is a collider so we shouldn't stratify on it.

D) Stratifying by A closes the path leading from A into X. X -> Z -> Y is a pipe and if we stratify by Z we would not be able to identify the causal effect between X and Y, so ultimately only A is conditioned on.

Question 2 (17 points). The data `foxes.csv` are 116 foxes from 30 different urban groups in England. These fox groups are like street gangs. Group size (`groupsize`) varies from 2 to 8 individuals. Each group maintains its own (almost exclusive) urban territory. Some territories are larger than others. The area variable encodes this information. Some territories also have more average food (`avgfood`) than others. And food influences the weight of each fox.

Assume the causal model defined by this DAG:



where F is `avgfood`, G is `groupsize`, A is `area`, and W is `weight`.

Solve the following problems based on the above data and causal model:

1. Considering area (A) as the treatment and average food (F) as the outcome, use the backdoor criterion to determine the variables that should be included in your model.

   Note: Be sure to explicitly state how the backdoor criterion was applied and the implications for the definition of your statistical model due to your use of the backdoor criterion.

There are no other variables within F and G that impact the paths we have from A to F. There is no path from any other variable besides A and F that lead into A, and by backdoor criterion no other variable should be included in the model besides A and F.

2. Estimate the **total causal effect** of A on F. **Include a prior predictive simulation** to justify your assignment of prior distributions to unobserved intercept and slope parameters in your model. **Limit your justification to no more than 3 sentences.**

   Note:

   - Perform your prior predictive simulation without using the `pymc.sample_prior_predictive()` function.
   - Feel free to assign an Exponential(1) to your standard deviation parameter in your model without justifying this choice.
   - You might want to consider standardizing the variables in your model (but this is not a requirement).

In [90]:
```
df = pd.read_csv(
    "Data/foxes.csv",
    sep=',',
    header=0
)
# df.head()
# prior predictive simulation
a_std = standardize(df.area)
f_std = standardize(df.avgfood)

with pm.Model() as m_AF:
    a = pm.Normal("a", 1, 1)
    b_a = pm.Normal("b_a", 1, 0.5)
    b_f = pm.Normal("b_f", 0.5, 0.5)
    mu = pm.Deterministic("mu", a + b_a * a_std)

    sigma = pm.Exponential("sigma", 1)
    div = pm.Normal("div", mu, sigma, observed=f_std)

    idata_AF, _ = quap([a, b_a, b_f, sigma])

az.summary(idata_AF, kind="stats")
```

```
<IPython.core.display.HTML object>
```

```
<IPython.core.display.HTML object>
```

Out[90]:

	mean	sd	hdi_5.5%	hdi_94.5%
a	0.002	0.044	-0.069	0.071
b_a	0.884	0.043	0.815	0.953
b_f	0.497	0.500	-0.350	1.256
sigma	0.466	0.066	0.359	0.568

Prior distributions are set to a mean and standard deviation of 1 and 1 respectively for “a” which represents area, because the possibility of the mean amount of food being 0 is unlikely – a mean of 1 is reasonable and a mean of 2 would be too high. The slope mean and standard deviation for area is 1 and 0.5 respectively because it would be reasonable to assume an area can range from +/- 1.5. For food, the slope of amounts could have a mean and deviation of 0.5 and 0.5 respectively

Answer the following question based on the results of your model:

What effect would increasing the area of a territory have on the amount of food inside it? **Limit your answer to no more than 2 sentences.**

As the size of the area increases, the amount of food within it also increases. The mean value of food is 0.884 which is between the 89% HPDI, this is also the same for the mean amount of area.

Question 3 (17 points). In this question, you will estimate **the total and direct causal effects** of adding food (F) to a territory on the weight (W) of foxes.

Which model variables do you need to adjust for in each case?

Make sure to explicitly state the adjustment sets needed for properly estimating the effects. *Hint: The backdoor criterion only includes variables in the adjustment set that need to be included in the model to close backdoors into your treatment variable. Variables may need to be included in your adjustment set to properly estimate a causal effect even when the variable is not part of a backdoor path.*

Adjustment set for total causal effect: {A} Adjustment set for direct causal effect: {A, G} For the purposes of measuring F -> W in the situation of total causal effect, we will leave G out of the model. If it were in the model that would mean we are attributing some of the effects of G into W while we are measuring F into W. By leaving G out we are isolating the effects of G from F -> W. By backdoor criterion, A is leading into F and we should stratify by A. For direct causal effect, G will be adjusted in the model to account for influence by G into W. A is also adjusted since it is a backdoor path into F.

Estimate **the total causal effect** of adding food (F) to a territory on the weight (W) of foxes. Communicate this effect using a summary table or visualization/plot of the posterior estimate.

In [91]:
```
# Total causal effect: Need to leave G out of the model, stratify by A because of backdoor cri
df = pd.read_csv(
    "Data/foxes.csv",
    sep=',',
    header=0
)
# df.head()

# prior predictive simulation
f_std = standardize(df.avgfood)
w_std = standardize(df.weight)
a_std = standardize(df.area) # Backdoor criterion "a"
with pm.Model() as m_FWA:
    f = pm.Normal("f", 1, 1)
    b_f = pm.Normal("b_f", 0.5, 0.5)
    b_w = pm.Normal("b_w", 0.5, 0.7)
    b_a = pm.Normal("b_a", 1, 0.5)
    mu = pm.Deterministic("mu", f + b_f * f_std + b_a * a_std)

    sigma = pm.Exponential("sigma", 1)
    div = pm.Normal("div", mu, sigma, observed=w_std)

    idata_FWA, _ = quap([f, b_f, b_w, sigma])

az.summary(idata_FWA, kind="stats")
```

```
<IPython.core.display.HTML object>
```

```
<IPython.core.display.HTML object>
```

Out[91]:

	mean	sd	hdi_5.5%	hdi_94.5%
f	0.010	0.099	-0.144	0.169
b_f	-0.858	0.097	-1.018	-0.710
b_w	0.508	0.704	-0.605	1.635
sigma	1.058	0.065	0.950	1.158

Estimate **the direct causal effect** of adding food (F) to a territory on the weight (W) of foxes. Communicate this effect using a summary table or visualization/plot of the posterior distribution.

In [92]:
```
# Direct causal effect: Stratify by G to measure direct causal effect between F -> W
df = pd.read_csv(
    "Data/foxes.csv",
    sep=',',
    header=0
)
# df.head()

# prior predictive simulation
f_std = standardize(df.avgfood)
w_std = standardize(df.weight)
g_std = standardize(df.groupsize)
with pm.Model() as m_FWG:
    f = pm.Normal("f", 1, 1)
    b_f = pm.Normal("b_f", 0.65, 0.4)
    b_w = pm.Normal("b_w", 0.5, 0.7)
    b_g = pm.Normal("b_g", 0.5, 0.7)
    mu = pm.Deterministic("mu", f + b_f * f_std + b_g * g_std)

    sigma = pm.Exponential("sigma", 1)
    div = pm.Normal("div", mu, sigma, observed=w_std)

    idata_FWG, _ = quap([f, b_f, b_w, sigma])

az.summary(idata_FWG, kind="stats")
```

```
<IPython.core.display.HTML object>
```

```
<IPython.core.display.HTML object>
```

Out[92]:

	mean	sd	hdi_5.5%	hdi_94.5%
f	0.010	0.101	-0.157	0.165
b_f	-0.409	0.097	-0.564	-0.254
b_w	0.502	0.699	-0.613	1.627
sigma	1.080	0.066	0.973	1.183

Question 4 (10 points). To estimate the the causal effect of group size (G) on weight (W), which variables do you need to adjust for? **Explicitly state the adjustment set needed for properly estimating the effect.**

We need to adjust for variable F because it is leading into G when we are trying to estimate G -> W. We don't need to consider A because although A -> F, our adjustment to F will account for influence from A. Adjustment set: {F}

Estimate the causal effect of group size (G) on weight (W). Express this effect using a summary table or plot/visualization of the posterior distribution.

In [93]:
```
# Causal effect: Stratify by F to estimate G -> W
df = pd.read_csv(
    "Data/foxes.csv",
    sep=',',
    header=0
)
# df.head()

# prior predictive simulation
g_std = standardize(df.groupsize)
w_std = standardize(df.weight)
f_std = standardize(df.avgfood) # Backdoor criterion "f"
with pm.Model() as m_GWF:
    g = pm.Normal("g", 1, 1)  # Group size mean set to 1 fox
    b_g = pm.Normal("b_g", 0.5, 0.7)
    b_w = pm.Normal("b_w", 0.5, 0.7)
    b_f = pm.Normal("b_f", 0.5, 0.5)
    mu = pm.Deterministic("mu", g + b_g * g_std + b_f * f_std)

    sigma = pm.Exponential("sigma", 1)
    div = pm.Normal("div", mu, sigma, observed=w_std)

    idata_GWF, _ = quap([g, b_g, b_w, sigma])

az.summary(idata_GWF, kind="stats")
```

```
<IPython.core.display.HTML object>
```

```
<IPython.core.display.HTML object>
```

Out[93]:

	mean	sd	hdi_5.5%	hdi_94.5%
g	0.008	0.087	-0.134	0.144
b_g	-0.594	0.086	-0.732	-0.458
b_w	0.503	0.703	-0.559	1.686
sigma	0.941	0.065	0.835	1.043

As the group size decreases, weights of the foxes are increasing. Likewise, if the size of the group increases, foxes weigh less which could indicate they are eating less food as a group.

In light of your estimates from Questions 2 - 4, what do you think is going on with these foxes? Feel free to speculate — **all that matters is that you justify your speculation. Limit your response to 6 sentences.**

As area size increases, the amount of food increases. However, as the group size increases, the weight of the foxes decrease. This could indicate that the distribution of food these foxes are consuming leave them in a deficit. There could be some competition occurring between foxes within the same group.

```
In [94]: grader.check("q4.3")
```

```
Out[94]: q4.3 results: All test cases passed!
```