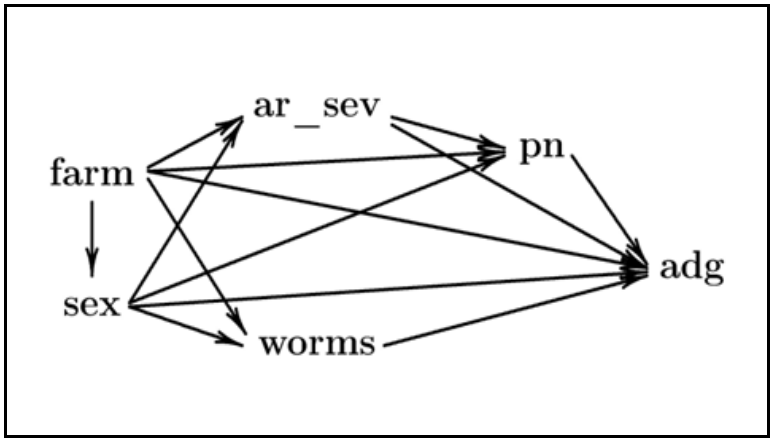
# Model Building Exercise (VER 15)
# Solution (with revised causal diagram)

**1. Identify the primary outcome of interest and main predictors of interest.**

A: The primary outcome of interest is -adg- and the primary predictors of interest are _worms-, -pn-, and -ar_sev-.

**2. Draw the causal diagram for your full causal model.**




**3. Identify which variables are potential confounders.**

For the -pn- –> -adg- relationship, -ar_sev-, -farm- and -sex- are potential confounders.
For the -ar_sev- –> -adg- relationship, -farm- and -sex- are potential confounders.
For the -worms- –> -adg- relationship, -farm- and -sex- are potential confounders.

Since both -sex- and -farm- likely have very strong relationships with -adg-, we will make the apriori decision to force them into all regression models.

**4. Identify any intervening (intermediate) variables.**

A: -pn- is an intervening variable for the -ar_sev- –> -adg- relationship.

**5. Identify any exposure-independent variables.**

A: -worms- is an exposure-independent variable for the -pn- and -ar_sev- effects.

**6. Go through the exercise of computing descriptive statistics for each variable, evaluate unconditional associations and carry out a pairwise correlation/association analysis among all predictors. If you were looking to eliminate potential predictors at this stage, are there any likely candidates?**

Outcome variable:

```
adg   Average daily weight gain (gms)
-------------------------------------------------------------------------------
                  type:  numeric (int)

                 range:  [317,707]                    units:  1
         unique values:  194                      missing .:  0/341

                  mean:   519.399
             std. dev:   70.9674

           percentiles:        10%       25%       50%       75%       90%
                               427       469       519       568       610
```

There are no missing observations for -adg-. The values are well spread out with some pigs gaining over twice as much as other pigs in the dataset. The mean and the median are very close which indicates that the data are roughly symmetrically distributed. The histogram confirms that they have an approximate normal distribution.

**Main predictors:**

```
pn   Pneumonia (lu>0)
-------------------------------------------------------------------------------
                  type:  numeric (byte)
                 label:  yn

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  0/341

            tabulation:  Freq.   Numeric  Label
                           155         0  no
                           186         1  yes.
```

The pigs are roughly split between those with and without pneumonia. This will make the analysis of the effects of -pn- reasonably powerful. There are no missing data.

```
ar_sev   (unlabeled)
-------------------------------------------------------------------------------
                  type:  numeric (float)

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  0/341

            tabulation:  Freq.  Value
                           307  0
                            34  1
```

There are only 34 pigs with severe atrophic rhinitis. Consequently, the study will not have much power to detect an effect of this disease (i.e., we will only find an effect if it is relatively large). There are no missing data.

```
worms   Count of nematodes in small intestine at time of slaughter
-------------------------------------------------------------------------------
                  type:  numeric (byte)

                 range:  [0,72]                       units:  1
         unique values:  31                       missing .:  0/341

                  mean:   3.36657
             std. dev:   9.88183

           percentiles:        10%       25%       50%       75%       90%
                                 0         0         0         2         8
```

Worm counts range from 0 to 72 but over 50% of pigs have no worms at all (distribution is highly skewed). A few pigs with very large worm burdens will probably have a profound influence on the estimate of the effect of worms (if -worms- is modelled as a continuous variable). There are no missing values.

**Potential confounders:**

```
farm   Farm identification number
-------------------------------------------------------------------------------
                  type:  numeric (byte)

                 range:  [1,15]                       units:  1
         unique values:  15                       missing .:  0/341

                  mean:   7.77419
             std. dev:   4.36888

           percentiles:        10%       25%       50%       75%       90%
                                 2         4         8        11        14
```

```
. tab farm

       Farm |
identificat |
 ion number |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         26        7.62        7.62
          2 |         21        6.16       13.78
          3 |         28        8.21       21.99
          4 |         26        7.62       29.62
          5 |         19        5.57       35.19
          6 |         24        7.04       42.23
          7 |         21        6.16       48.39
          8 |         28        8.21       56.60
          9 |         25        7.33       63.93
         10 |         17        4.99       68.91
         11 |         21        6.16       75.07
         12 |         14        4.11       79.18
         13 |         25        7.33       86.51
         14 |         24        7.04       93.55
         15 |         22        6.45      100.00
------------+-----------------------------------
      Total |        341      100.00
```

The pigs were relatively evenly distributed across the 15 farms. There were no missing values.

```
sex   Sex of the pig
-------------------------------------------------------------------------------
                  type:  numeric (byte)
                 label:  sex

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  0/341

            tabulation:  Freq.   Numeric  Label
                           172         0  female
                           169         1  castrate
```

The pigs are roughly equally split between the two sexes. There are no missing values.

**Unconditional associations:**
All of the following associations were evaluated by fitting simple (unconditional) linear regression models (data not shown):

- -farm- is very strongly associated with -adg-.
- -sex- is significantly (P=0.001) associated with -adg-. Females gained an average of 506 gms/day, while the weight gain was 26.5 gms/day higher for castrated males.
- -pn- (pneumonia) is significantly (P=0.001) associated with -adg-, with -pn- positive pigs gaining 24.5 gms/day less per day than -pn- negative pigs (533 gms/day).
- -ar_sev- (severe atrophic rhinitis) seems to have a profound (and strongly significant) impact on -adg-.; on average, affected pigs gained 68 gms/day less than unaffected pigs.
- -worms- (worm burden) seems to be significantly associated with -adg-, but surprisingly, the -adg- goes

up as the worm burden goes up (we would have expected an effect in the opposite direction).

**Correlations:**

```
. pwcorr sex adg pn ar_sev worms

             |      sex      adg       pn   ar_sev    worms
-------------+---------------------------------------------
         sex |   1.0000
         adg |   0.1868   1.0000
          pn |  -0.0021  -0.1723   1.0000
      ar_sev |  -0.0166  -0.2892   0.0482   1.0000
       worms |   0.0066   0.2472  -0.0646  -0.0074   1.0000
```

There are no correlations among predictors that are sufficiently large that I would worry about collinearity problems. (Note: many of the predictors are dichotomous, and correlations are not a good measure of association among dichotomous variables, but the very low correlations observed give us hope that collinearity will not be a serious problem).

**7. Decide what 2-way interactions you want to examine.**

A: Based on knowledge of the two diseases (and to keep this teaching exercise reasonably short), we will apriori decide to look only at the possible interaction between -pn- and -ar_sev-. The model including this interaction only is shown below.

```
. regress adg pn##ar_sev

      Source |       SS       df       MS              Number of obs =     341
-------------+------------------------------           F(  3,   337) =   14.62
       Model |  197166.325     3  65722.1084           Prob > F      =  0.0000
    Residual |  1515201.43   337  4496.14669           R-squared     =  0.1151
-------------+------------------------------           Adj R-squared =  0.1073
       Total |  1712367.76   340  5036.37576           Root MSE      =  67.053

------------------------------------------------------------------------------
         adg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          pn |
         yes |  -18.89445    7.67544    -2.46   0.014    -33.99226   -3.796644
    1.ar_sev |   -42.9117   19.42989    -2.21   0.028    -81.13083   -4.692567
             |
   pn#ar_sev |
      yes#1  |  -38.80518   24.87714    -1.56   0.120    -87.73922    10.12885
             |
       _cons |   536.3732   5.626989    95.32   0.000     525.3048    547.4417
------------------------------------------------------------------------------
```

The interaction is not quite significant, but perhaps strong enough to be included in the model building.

**8. Use forward selection, backward elimination and stepwise selection procedures to identify potential models for further investigation.**

A: All 3 procedures produced exactly the same model. This is unusual and is a function of the fact that we have a small number of predictors and they either have relatively strong associations with the outcome or non at all (ie no borderline significant associations). For the sake of simplicity, only the results for the stepwise backwards elimination procedure are shown.

```
. xi:stepwise, lockterm1 pe(0.05) pr(0.051): regress adg (i.farm sex) worms (i.pn*ar_sev)
i.farm            _Ifarm_1-15         (naturally coded; _Ifarm_1 omitted)
i.pn              _Ipn_0-1            (naturally coded; _Ipn_0 omitted)
i.pn*ar_sev       _IpnXar_se_#        (coded as above)
```

```
                     begin with full model
p = 0.6953 >= 0.0510  removing worms

      Source |       SS       df       MS              Number of obs =     341
-------------+------------------------------           F( 18,   322) =   32.06
       Model |   1099087.8    18  61060.4335           Prob > F      =  0.0000
    Residual |  613279.957   322  1904.59614           R-squared     =  0.6419
-------------+------------------------------           Adj R-squared =  0.6218
       Total |  1712367.76   340  5036.37576           Root MSE      =  43.642

------------------------------------------------------------------------------
         adg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Ifarm_2 |   19.81175   12.83143     1.54   0.124    -5.432275    45.05577
    _Ifarm_3 |   23.34785   12.44097     1.88   0.061    -1.127998    47.82369
    _Ifarm_4 |  -113.5638   12.27253    -9.25   0.000    -137.7082    -89.4193
    _Ifarm_5 |  -24.56906   13.21677    -1.86   0.064    -50.57119    1.433069
    _Ifarm_6 |  -69.86855   12.38692    -5.64   0.000    -94.23807   -45.49903
    _Ifarm_7 |  -14.67511   12.90451    -1.14   0.256    -40.06291    10.71269
    _Ifarm_8 |  -57.57176   13.10607    -4.39   0.000    -83.35611   -31.78742
    _Ifarm_9 |    -62.704   12.39326    -5.06   0.000    -87.08599   -38.32202
   _Ifarm_10 |  -84.08496   13.76267    -6.11   0.000    -111.1611   -57.00886
   _Ifarm_11 |  -97.94475    12.9311    -7.57   0.000    -123.3849   -72.50463
   _Ifarm_12 |  -123.4743   14.57479    -8.47   0.000    -152.1481   -94.80047
   _Ifarm_13 |  -115.3472   12.74396    -9.05   0.000    -140.4192   -90.27531
   _Ifarm_14 |  -60.59613    12.6888    -4.78   0.000    -85.55955   -35.63271
   _Ifarm_15 |   23.50103    13.0893     1.80   0.074    -2.250319    49.25238
         sex |   29.16879    4.77938     6.10   0.000     19.76604    38.57155
       _Ipn_1 |  -17.46428    5.59491    -3.12   0.002    -28.47147    -6.45708
      ar_sev |   -30.9291   13.11464    -2.36   0.019     -56.7303   -5.127898
_IpnXar_se_1 |  -38.84491   16.73416    -2.32   0.021    -71.76701   -5.922819
       _cons |   568.4791   10.17541    55.87   0.000     548.4604    588.4978
------------------------------------------------------------------------------
```

**9. Evaluate potential confounding effects by forcing all removed predictors that may be confounders, back into the model. Do any of them need to be kept, even though not statistically significant, because they appear to exert a confounding effect?**

A: There really is nothing to do here since, based on our causal diagram, worms can not be a confounder for the effects of either -pn- or -ar_sev- on -adg-. However, if we do force it back in the model it has virtually no effect on the coefficients for the other disease variables (results not shown, see do-file for details).

**10. Identify the model which best evaluates the effect of -ar_sev- on -adg-.**

A: This is the model shown above.

**11. Evaluate the reliability of the model. Compute also the PRESS statistic, and explain what it tells you about the fitted model.**

A: With the model built using 60% of the data and then tested on the other 40%, the correlations between the predicted and observed values of -adg- for the first and 2nd subsets were (for this particular random split of the data) as follows:

```
. pwcorr adg pv if rand <0.6

             |      adg       pv
-------------+------------------
         adg |   1.0000
          pv |   0.8332   1.0000
```

This means that R^2=0.8332^2=0.69 for the estimated model.

```
. pwcorr adg pv if rand >=0.6

             |      adg       pv
-------------+------------------
         adg |   1.0000
          pv |   0.7304   1.0000
```

This means that R^2=0.7304^2=0.53 for the predictions, hence a substantial shrinkage. With other random splits of the data we would however achieve different values, so one should at least try this several times to make sure it wasn't just bad luck.

If we instead use the leave-one-out cross-validation implicit in the PRESS statistic, we will compare the R^2=0.64 of the full model (above) to the predictive R^2 value:

```
. predict resid, res
. predict lev, leverage
. gen eq1=(resid/(1-lev))^2
. summ eq1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         eq1 |       341    2033.087    2998.274   .0189977   17353.27

. di "PRESS =" r(sum)
PRESS =693282.69

. di "R2(pred) = "1-(r(sum)/ (`e(mss)' + `e(rss)'))
R2(pred) = .59513213
```

The PRESS value increased to 693,283 (from 613,280 - the residual sum of squares of the full data model), and the corresponding R^2 dropped to 0.595. This shrinkage in predictive ability does however not give rise to serious concern.