[HANDOUT 15: Using Regression Models for Prediction]

The purpose of Regression Models is to (1) Explain relationships between variables, and (2) make predictions for new observations. Prediction performance is measured based on how well it predicts for new observations.

Ofcourse, the performance measurement of a prediction will be invalid if the new observation being predicted is not representative of the observations used to fit the model.

Correlation Coefficients

1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0	-0.1	-0.2	-0.3	-0.4	-0.5	-0.6	-0.7	-0.8	-1.9	-1.0

0: No correlation
0.3 / -0.3 = weak correlation
0.5 / -0.5 = moderate correlation
0.8 / -0.8 = moderate-strong correlation
1.0 / -1.0 = very strong correlation

We use a Correlation Matrix to see correlations between the independent variables.

```
round(cor(evap.df[, -11]), 2)

##       MaxST MinST  AvST MaxAT MinAT  AvAT  MaxH  MinH  AvH  Wind
## MaxST  1.00  0.85  0.95  0.91  0.47  0.83 -0.21 -0.67 -0.76 -0.08
## MinST  0.85  1.00  0.93  0.84  0.67  0.82 -0.16 -0.34 -0.47  0.02
## AvST   0.95  0.93  1.00  0.91  0.59  0.87 -0.19 -0.53 -0.68 -0.08
## MaxAT  0.91  0.84  0.91  1.00  0.57  0.87 -0.11 -0.53 -0.66 -0.07
## MinAT  0.47  0.67  0.59  0.57  1.00  0.78 -0.12  0.19 -0.06  0.43
## AvAT   0.83  0.82  0.87  0.87  0.78  1.00 -0.05 -0.31 -0.54  0.15
## MaxH  -0.21 -0.16 -0.19 -0.11 -0.12 -0.05  1.00  0.16  0.29 -0.13
## MinH  -0.67 -0.34 -0.53 -0.53  0.19 -0.31  0.16  1.00  0.91  0.34
## AvH   -0.76 -0.47 -0.68 -0.66 -0.06 -0.54  0.29  0.91  1.00  0.22
## Wind  -0.08  0.02 -0.08 -0.07  0.43  0.15 -0.13  0.34  0.22  1.00
```

```
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -79.244834  1.221e+02 -0.6492 0.5204303
## MaxST         2.355252  9.350e-01  2.5189 0.0164926
## MinST        -0.137321  1.007e+00 -0.1364 0.8923023
## AvST         -0.676199  3.060e-01 -2.2096 0.0337662
## MaxAT         0.473272  5.572e-01  0.8494 0.4014051
## MinAT         0.451802  7.661e-01  0.5898 0.5591279
## AvAT          0.024874  2.150e-01  0.1157 0.9085497
## MaxH          1.562590  1.124e+00  1.3900 0.1732994
## MinH          0.866240  4.676e-01  1.8526 0.0723914
## AvH          -0.597152  1.586e-01 -3.7640 0.0006143
## Wind          0.009267  8.716e-03  1.0633 0.2949265
```

Few estimated coefficients are insignificant as there exists high correlations between some pairs of regressors. This is because correlation inflates standard errors, which increases their P-Value when compared to the response. High correlations between independent variables mean they are not providing unique, independent information to the model.

VIF: Variation Inflation Factors: Diagonal inverse matrix

Provide a measurement of multicollinearity existence. A VIF=1 for a variable, for example "Temperature" means temperature explains the response well on its own and is not affected by other regressors. A VIF > 10 indicates strong multicollinearity. For example "Height" has a high VIF, with another predictor, let's say "Weight", meaning both weight and height work together to explain the response non-independently.

MSPE: Mean Square Prediction Error

Is used for continuous responses where MSPE = $E(Y - \hat{Y})^2$ which will quantify how well the model predicts the response for new observations. We calculate the prediction error for each new observation that is predicted, find the mean of these errors, and square it. We cannot use new samples in real-life applications as it is time consuming, costly, and expensive, so instead, we split the data into a training and test set.

Example 1:

CV: Cross-Validation

For small datasets, we use cross-validation to split the data. Observations: 46 (If we split data in 10 parts; 10-fold CV)
[01 02 03 04][05 06 07 08][09 10 11 12][13 14 15 16][17 18 19 20][21 22 23 24] [25 26 27 28][29 30 31 32][33 34 35 36][37 38 39 40][41 42 43 44 45 46]

Calculate prediction error using a different subset each time

Test set: [01 02 03 04]	Training set: [05 06 07 08]	Prediction Error: e_1
Test set: [01 02 03 04]	Training set: [09 10 11 12]	Prediction Error: e_2
Test set: [01 02 03 04]	Training set: [13 14 15 16]	Prediction Error: e_3
Test set: [01 02 03 04]	Training set: [17 18 19 20]	Prediction Error: e_4
Test set: [01 02 03 04]	Training set: [21 22 23 24]	Prediction Error: e_5
Test set: [01 02 03 04]	Training set: [25 26 27 28]	Prediction Error: e_6
Test set: [01 02 03 04]	Training set: [29 30 31 32]	Prediction Error: e_7
Test set: [01 02 03 04]	Training set: [33 34 35 36]	Prediction Error: e_8
Test set: [01 02 03 04]	Training set: [37 38 39 40]	Prediction Error: e_9
Test set: [01 02 03 04]	Training set: [41 42 43 44 45 46]	Prediction Error: e_{10}

In the example above, we see that the training set only consists of 90% of the data, leading to an overestimate of the MSPE. A model trained on a smaller subset has less information to learn, which leads to slightly worse predictive performance on the test set.

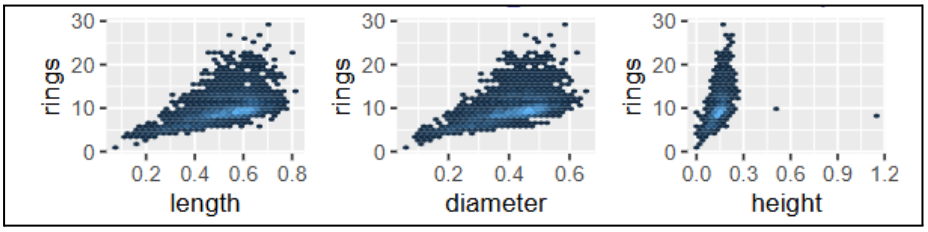
AIC(c) / BIC to shortlist submodel candidates
In a multiple explanatory regression, say we have 10 explanatory variables, then we would have $2^{10} = 1024$ different subset models to choose from. It would be far too computationally extensive to calculate the MSPE for every 1024 submodel, so we use AIC(c)/BIC to shortlist candidate models, and only calculate MSPE for the chosen best candidates.

AICc models [168, 040, 247, 552, 695, 183, 680, 184, 104, 520]
BIC models [040, 168, 516, 008, 552, 520, 183, 247, 772, 004]

Take the top 5 models of both AICc and BIC
Top models = [168, 040, 247, 552, 695, ~~040~~, ~~168~~, 516, 008, ~~552~~]
Top models = [168, 040, 247, 552, 695, 516, 008]
(Overlap in model 168, 040 and 552, so we end up with 7

MSPE for the seven submodels: [51.08 51.36 51.23 51.72 51.80 **50.02** 52.97]
MSPE for the full model: [65.1]
Hence, the winner is model 06 with the lowest MSPE = 50.02; lm(formula = Evap ~ AvAT + AvH + Wind + 1)
*Still an under-estiamte

Example 2 (Poisson):
Another example of choosing the best model for prediction using MSPE (Continuous response), but now we have a dataset with a big sample size. We have 4177 observations in the Abalone data set.



First we inspect the data and notice that:

1) There are two unusually large values for height and there is a minimum value for height of 0.
    - Observation 1258 and 3997 have a height of 0
2) There is an unusual combination of length and diameter
    - ## length diameter
    - ## 1211 0.185 0.375
    - One point has length < diameter which should not happen given the way the measurements were defined.

After removing observations 1358, 3997, and 1221 we have completed data cleaning and can move on to measuring the performance of prediction. (We also saw the the full model had a residual deviance < DF suggesting under-dispersion hence we fit a quasi-poisson to replace the Poisson)

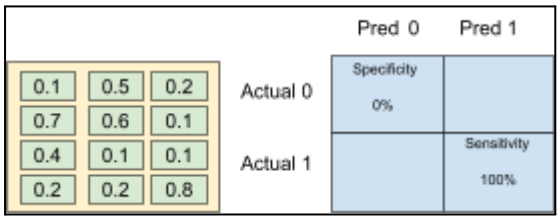
We run AIC and BIC, deriving 6 submodels and then fit **GAM's** to each regressor to explore ways of improving the model. The output indicates that the full model is giving the lowest estimated MSPE.

MSPE for Poisson
It becomes increasingly difficult to make precise predictions for Poisson Models as the number of counts increases because the mean increases as the variance increases.

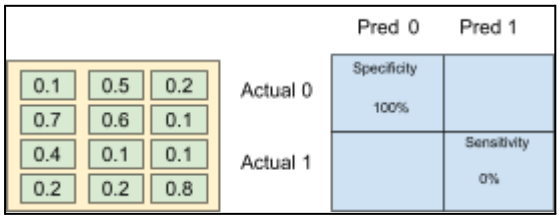
Example 3 (Logistic):
For a logistic regression, measuring performance of prediction is difference because instead of measuring the difference between prediction for observation$_i$ and what observation$_i$ actually was, we classify prediction error as the probability of a wrong classification.

True -ve	False +ve
False -ve	True +ve

We predict a case (malignant) if the estimated probability is ≥ 0.5 and severity = 0 (benign) otherwise. Hence, generally speaking, we predict a malignant case if p ≥ c for some constant c where 0 ≥ c ≥ 1.



c = 0
If c = 0, every observation will be predicted as (1) and we will have a sensitivity=1 and specificity=0. This is because everything that is TRULY a 1 will be predicted a 1, but everything that is truly a 0, will also be predicted as 1, hence there are all true positives, but no true negatives.



c = 1
If c = 1, every observation will be predicted as (0). This means everything that is TRULY 0, will be predicted as 0, but everything that is truly 1, will also be predicted as 0, so NO true positives are captured. So we will have a sensitivity=0 and specificity=1

As c varies from 0 to 1, sensitivity goes from 1 to 0, and specificity goes from 0 to 1.

Say we are creating a Pregnancy test. It would be dangerous to have a bigger number of false negatives, so we would adjust the cut-off, c, such that the model predicts less false negatives, and more false positives. Hence we would rather have more false positives than false negatives. This means, more predictions will be 1 when they are truly 0. Hence, a higher cut-off is needed to achieve this.

[HANDOUT 16: Using Regression Models for Explanation]

Regression models are used for Prediction (1), and Explanation (2), explanatory models can be thought of as either Descriptive modeling, or Causal modeling.

Descriptive Modeling
Perching Birds Example

- Length Mean body length (cm).
- NestType: Type of nest built.
- OorC Is the nest open or closed?
- Location: Location of the nest. Note that decide means that the nest is in a deciduous tree and conif means that it is in a coniferous tree (the remaining levels are self explanatory).
- Eggs Average number of eggs.
- Marking 1 indicates eggs have markings and 0 indicates eggs have no markings.
- Incubate Mean length of time (in days) the eggs are incubated.
- Nestling Mean length of time (in days) the babies are cared for in the nest.
- TotalCare Total care time = Incubate + Nestling

Possible Descriptive modeling Questions: Key work: 'Related' do not state causality

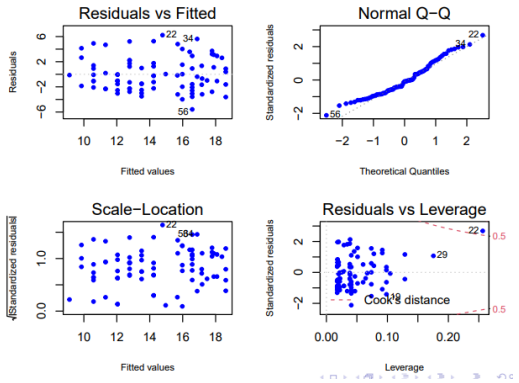
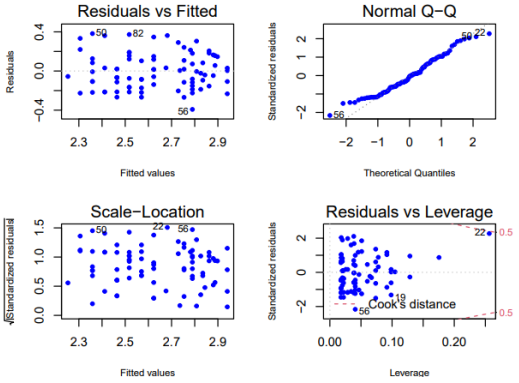
- Does the length of birds differ per nest type?
- Does the nest type relate to the average number of eggs found in each?
- Does the mean nestling time relate to the average number of eggs?

Effect Modification: Be careful about one variable affecting another, fix using an interaction.

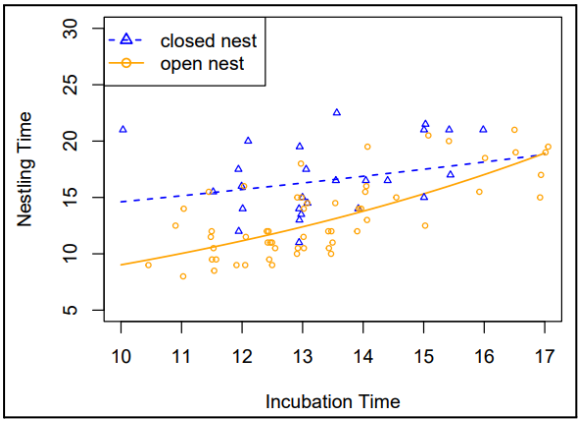
Answering descriptive modeling questions
Data exploration

- Scatter plot (Positive/Negative, outliers, scatter variability, etc)
- Conditional plot for interactions, splitting data into groups to assess difference in pattern according to a specific condition such as the Nestling time for open vs closed nest types. If there is a difference, we explore an interaction between Nestling time and open/close.
- Box-Cox plots: explore transformations of the response

$\lambda = -2$	$\lambda = -1$	$\lambda = -0.5$	$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 2$
$(response)^{-2}$	$(response)^{-1}$	$1 / \sqrt{response}$	log(response)	$\sqrt{response}$	nothing	$response^2$

Model without Log-Transformation	Model with Log-Transformation
## Coefficients: ## Estimate Std. Error t value Pr(>|t|) ## (Intercept) 8.762 5.038 1.74 0.086 . ## Incubate 0.601 0.375 1.60 0.113 ## OorCopen -14.970 5.776 -2.59 0.011 * ## Incubate:OorCopen 0.859 0.430 2.00 0.049 *	## Coefficients: ## Estimate Std. Error t value Pr(>|t|) ## (Intercept) 2.3211 0.3464 6.70 2.8e-09 *** ## Incubate 0.0361 0.0258 1.40 0.1662 ## OorCopen -1.1798 0.3971 -2.97 0.0039 ** ## Incubate:OorCopen 0.0698 0.0296 2.36 0.0208 *
	
Linearity: Based on the Residuals vs Fitted plot, there is some skewness as more data saturates on the RHS, indicating a need for a log transformation. **Independence**: As far as we can tell from the context provided, it appears that the birds are independent, though there may be issues if birds from the same location aren't independent of each other. **Normality**: I am somewhat concerned about the normality assumption due to the clear departure from the line in the QQ plot. That said, linear regression is robust to departures from normality. **EOV**: Based on the Scale-Location plot, it seems like the variance may be non-constant, specifically, it seems to be increasing as the fitted values increase.	**Linearity**: Based on the Residuals vs Fitted plot, there is less skewness. **Independence**: Birds are independent, though there may be issues if birds from the same location aren't independent of each other. **Normality**: Departure from line has decreased very slightly. **EOV**: Variance is still non-constant, but better than the first fit. We can see that the model with the logged-response is a slightly better model.

Findings



(1) We conclude the relationship between incubation time is different for closed and open nest birds. For open nests, incubation time increases as nestling time increases, but for closed nests, there is no clear indication of an increase.

(2) Open nests have a lower nestling time compared to closed ones but as incubation time increases, the difference decreases.

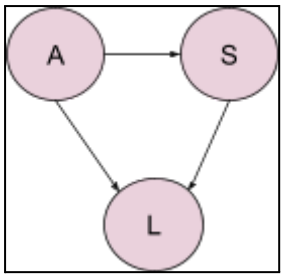
(3) There is quite a bit of scatter around trend lines for both open nest and closed nest species.

Causal Modeling

Example 1: Smoking and Lung Cancer

If we want to investigate the hypothesis that smoking has a causal effect on the occurrence of lung cancer…

- We assume Lung cancer does not cause smoking, smoking does not cause Allele.



This is the most complicated causal model, where Allele might cause Lung Cancer and smoking. Smoking may cause cancer, and Lung Cancer does not cause either.

S-L Connection

L ~ S	Direct Effect, ignores impact of A.
L ~ S + A	Confounding Effect of A on L.

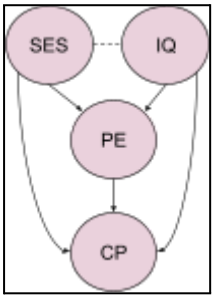
A-L Connection

L ~ A	Direct Effect
L ~ S + A	Indirect Effect of S is removed as S is fixed and we explore A on L.

A-S Connection

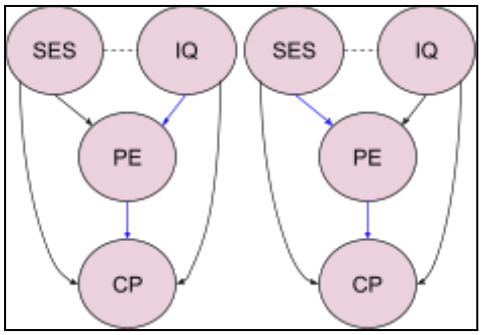
S ~ A	Direct Effect
S ~ S + A + L	Colliding Effect

Example 2: College Plans



Direct Causal Effects
seniors.glm = glm(cbind(CPyes, CPNo) ~ PE + IQ + SES, binomial)

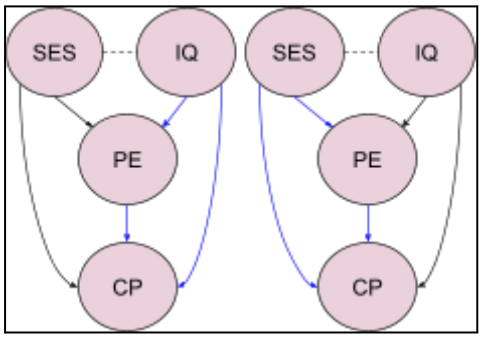
Fit the logistic regression model where log odds of CP=yes is the response and SES, IQ and PE are explanatory variables. This model closes all the other pathways to CP in the causal diagram.



Indirect Causal Pathways

- The effect of IQ on PE
- The effect of SES on Pe
- The direct effect of PE on CP

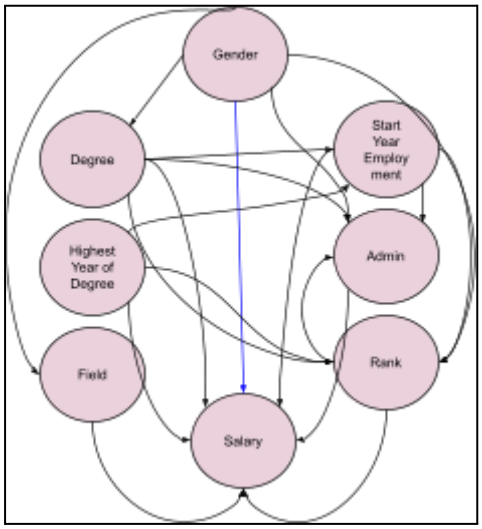
pe.glm = glm(cbind(**PE**high, **PE**low) ~ IQ + SES, binomial)
We find moderate evidence of an interaction between IQ and SES. The only interaction coefficient that is showing up as having a smallish p-value (0.065) is IQH:SESH



Total Effects
seniors3.glm = glm(cbind(**CP**yes, **CP**No) ~ IQ + SES)

To explore the total causal effects of SES and IQ on CP, fit the logistic regression model where log odds of CP=yes is
The response and SES and IQ are explanatory variables. This leaves the indirect causal pathways from IQ to CP and from SES to CP open.
The coefficient for IQ estimates its total effect on CP and similarly the SES coefficient estimates its total effect on CP

Example 3: Discrimination Data



The context of this example means that we are focused on assessing the causal relationship between gender and salary.

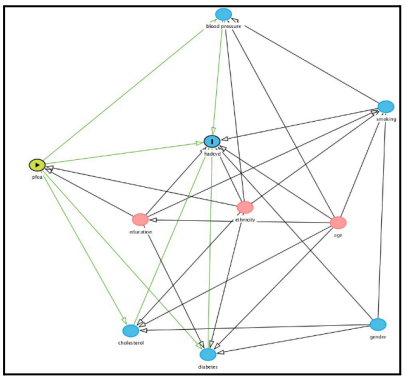
Direct Causal Effect
Blue blue: Direct effect of Gender on Salary, that is, when all other variables are fixed, is there a difference between Females and Males in terms of the Salary.

lm(formula = salary ~ gender + deg + field + startyr + yrdeg + ank + admin, data = salary.df)

Total Causal Effect
The combined effect of all causal paths that lead from gender to Salary. Consider two people, a Female and Male who are of identical ability and drive. Is there a difference in their expected salaries?

Thus we should not include any explanatory variable that lies on an indirect causal effect path from gender to salary in our model. Each of the other variables lie on at least one indirect causal pathway from gender to salary and thus they all must be excluded from the model.
lm(formula = (1/salary) ~ gender, data = salary.df)

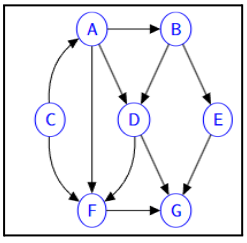
2022 S2 Causal Diagram Question



What variables do you need to include to estimate the effect of pfoa on heart disease. Give brief reasons why you would include or exclude each variable.

To estimate the effect of pfoa on heart disease, we explore the direct effect of pfoa on heart disease. That is, what will be the relationship between a person's PFOA exposure on CHD if all other variables related are fixed.
We need to block the effects of Education, Ethnicity, age, and smoking as they are confounding paths. We should include gender as it is a predictor of the outcome. And we will NOT include the intermediate causal variables about blood pressure and diabetes.

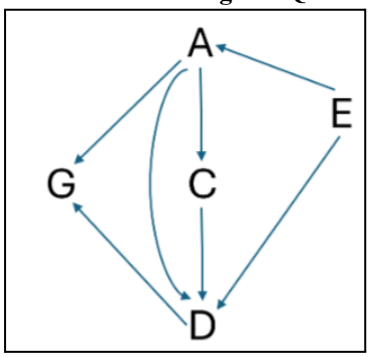
2021 S1 Causal Diagram Question



(a) List all of the variables that have a direct causal effect on G.
(b) Suppose we wish to estimate the direct causal effect that A has on F. List all of the explanatory variables that should be included in the model.
(c) Suppose we wish to estimate the total causal effect that A has on F. List all of the explanatory variables that should be included in the model.
(d) Consider the direct effect of D on F. (i) List all of the variables that are confounders for this effect. and (i) List all of the variables that are colliders for this effect.

a. F, D, E
b. A, C, D
c. A, C
d. A, C
e. G

2024 S1 Causal Diagram Question



a) Suppose we want to estimate the direct effect of A on D. List the variables that should be included as explanatory variables in the model we use for this purpose. A, C, E
b) Suppose we want to estimate the total effect of A on D. List the variables that should be included as explanatory variables in the model we use for this purpose. A, E

Suppose that a variable F is missing in the above causal diagram. The variable F affects D directly and does not affect any of the present variables.

c) How does the inclusion of F as an explanatory variable in the model from Question 2(a) affect the direct effect estimation of A on D? Inclusion of F does not affect the direct estimation of A on D.
d) How does the inclusion of F as an explanatory variable in the model from Question 2(a) affect the model predictability? It will improve predictability as F will capture variability that is not explained by the other variables.

2019 S1 Causal Diagram Question

a) An effect modifier is a variable in a model, as an explanatory variable that affects the effect between two other variables. For example, If the effect of x and y differ depending on z, then z is an effect modifier.
b) Casual relationships can be explored by exploring the direct causal effects, or total causal effects.
c) Direct: A and B: my.glm ← (Y ~ A + B, family=Poisson, data = my.df)
d) Indirect: my.glm ← (B ~ A, family=Poisson, data = my.df)
e) Total: my.glm ← (Y ~ A, family=Poisson, data = my.df)

Direct Effects	Include factors that could confuse the relationship but are part of the direct link. For example, if exercise → energy levels → health, then energy levels are in the direct line. Include energy but ignore any outside factors that don't connect directly on that path. If we think diet affects both exercise and health separately, it's not in that direct line and can be ignored for this estimate.
Total Effects	To get the total effect, include any factor that affects both exercise and health in any way, even if it's not directly in the middle. Example: Diet affects both exercise and health separately, so include it to see the bigger picture.
Confounders	These are variables that affect both the cause and the effect independently, creating an additional connection between them. (directly or indirectly)
Colliders	These are variables that are caused by both the starting point and the endpoint variables.
Effect Modifier	If the effect between two variables, say exercise and health, varies depending on a third variable, that third variable is likely an effect modifier.