# SOSC 5340: Generalized Linear Models

Han Zhang

Feb 23, 2021

# Outline

Logistics

Generalized Linear Models

Multinomial and Ordered Logit

Poisson, Negative Binomial, and Zero-inflated Poisson

Model Selection

Bias-Variance Trade-Off

Today's Review

# Next week

- Presentation
- Exercise 1 Due (before the class, submit your code and results on Canvas)

# Limited Dependent Variable

- Beyond binary outcomes, $Y \in \{0, 1\}$
- Categorical:
  - e.g., major choices;
- Integer (count): $Y \in \{0, 1, 2, \cdots\}$
  - e.g., event counts
- Censored: observed $Y$ is in a certain range, but we know in reality they should not be
  - e.g., US census write anyone who report their age $> 90$ as 90; so in census, age is between [0,90]
- The common problem is that the outcome $Y$ is limited to some regions, not in $(-\infty, \infty)$
  - so economists sometimes call them as limited dependent variable

# Generalized Linear Model

- To model limited dependent variables, we use generalized linear model (GLM)
- GLM looks like:
  - $h(E(Y|X)) = X\beta$
  - or, $E(Y|X) = h^{-1}(X\beta)$
- $h()$ is called link function
- Linear regression is a kind of GLM, where $h(X) = X$
- Logistic regression is a kind of GLM, where $h(X) = logit(X)$
- Other GLM we will learn today choose different $h()$ to model different types of $Y$

# GLM

- In practice, scholars use MLE to make statistical estimation and inference for GLM
- Recall that to use MLE, we need to make assumptions about what $p(Y|X)$ looks like

# Estimation and Inference of MLE

- Steps are standard
  1. write down $P(Y|X)$
  2. write down $\log L$: the log-likelihood function
  3. obtain coefficient estimates that maximize log-likelihood
     - and use Hessian matrix to calculate confidence interval

# Extending Logistic Regression

- Suppose we have categorical outcome with more than two values
- Sometimes, these categories have no intrinsic orders
  - E.g., majors choices between ( Economics = 1, Political Science = 2, Sociology= 3, Public Policy = 4 )
- Other times, these categories are ordinal
  - E.g., a survey ask whether you think religion deters economic growth, on a 1-7 scale.
  - 1 means strongly disagree, and 7 means strongly agree
  - Order gives more information than pure categories
  - Why not use continuous outcome models?
    - Dont want to assume equal distances between levels
    - Say, moving from 1-4 is different from 4-7
    - Assuming continuous $Y$ does not distinguish these two

## Ordered Logit: ordered outcome

- Peter McCullough, *Regression Models for Ordinal Data*, 1980
- Recall that logistic regression assumes a generating process based on latent variables

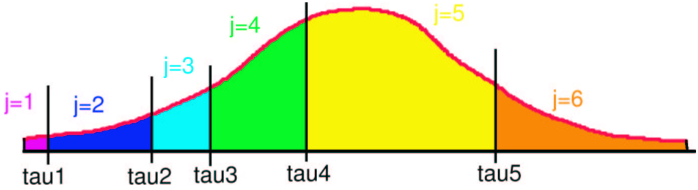
$$
\begin{aligned}
Y^* &= X\beta + \epsilon \\
Y &= \begin{cases} 1 & \text{if } y^* > 0 \\ 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{1}
$$

- $Y^*$ is an unobserved latent variable
- if the latent variable is bigger than a pre-determined cutoff (here 0), $Y = 1$
- Otherwise, $Y = 0$

# Ordered Logit

- We can borrow the same intuition to derive ordered logit regression, with $J > 2$ ordinal categories
- We create $J - 1$ latent cutoffs

$$Y = \begin{cases} 1 & \text{if } Y^* \leq \tau_1 \\ 2 & \text{if } \tau_1 < Y^* \leq \tau_2 \\ 3 & \text{if } \tau_2 < Y^* \leq \tau_3 \\ . & \\ J & \text{if } \tau_{J-1} \leq Y^* \end{cases} \tag{2}$$



## Ordered Logit

- So now the Assumption 3 for ordered logit becomes:

$$Y^* = X\beta + \epsilon$$

$$Y = \begin{cases} 1 & \text{if } Y^* \leq \tau_1 \\ 2 & \text{if } \tau_1 < Y^* \leq \tau_2 \\ 3 & \text{if } \tau_2 < Y^* \leq \tau_3 \\ . & \\ J & \text{if } \tau_{J-1} \leq Y^* \end{cases} \tag{3}$$

- And the error $\epsilon$ follows a standard logistic distribution (the same as logistic regression)

# Ordered Logit vs Linear Regression

- It may be easier to change from "very unlikely" (1) to "unlikely" (2), but it is more difficult to change from "unlikely" to "neutral" (3)
- For linear regression
  - It takes the same amount of changes in $X$ to turn $Y$ from 1 to 2 versus $Y$ from 2 to 3
  - Linear regression does not capture this difference
- For ordered logit
  - $Y$ changing from 1 to 2 means latent $Y^*$ changes from below $\tau_1$ to $(\tau_1, \tau_2)$
  - $Y$ changing from 2 to 3 means latent $Y^*$ changes from $(\tau_1, \tau_2)$ to $(\tau_2, \tau_3)$
  - It often requires a different amount a change in $X$ to move $Y$ from 1 to 2 versus from 2 to 3. That's what we want to capture

## Ordered Logit

- For MLE, we have to explicitly write down $P(Y|X)$

$$
\begin{aligned}
\mathrm{P}(Y=1|X) &= \Pr\left(\beta X + \epsilon \leq \tau_1 | X\right) \\
&= \mathrm{P}\left(\epsilon \leq \tau_1 - \beta X | X\right) \\
&= F\left(\tau_1 - \beta X\right), \text{(definition of cumulative probability } F) \\
&= logit^{-1}(\tau_1 - \beta X)
\end{aligned}
\tag{4}
$$

$$
\begin{aligned}
\mathrm{P}(Y=2|X) &= \Pr\left(\tau_1 < \beta X + \epsilon \leq \tau_2 | X\right) \\
&= \Pr\left(\tau_1 - \beta X < \epsilon \leq \tau_2 - \beta X | X\right) \\
&= F\left(\tau_2 - \beta X\right) - F\left(\tau_1 - \beta X\right) \\
&= logit^{-1}\left(\tau_2 - \beta X\right) - logit^{-1}\left(\tau_1 - \beta X\right)
\end{aligned}
\tag{5}
$$

And so on and so forth, for $j$ up to $J-1$

# Ordered Logit

The last category $J$

$$\begin{aligned} \mathrm{P}(Y = J|X) &= \mathrm{P}\left(\tau_{J-1} \leq \beta X + \epsilon|X\right) \\ &= \mathrm{P}\left(\epsilon \geq \tau_{J-1} - \beta X|X\right) \\ &= 1 - \mathrm{P}\left(\epsilon < \tau_{J-1} - \beta X\right) \\ &= 1 - F\left(\tau_{J-1} - \beta X\right) \\ &= 1 - logit^{-1}\left(\tau_{J-1} - \beta X\right) \end{aligned} \tag{6}$$

- We have written down $P(Y|X)$ for every possible value of $Y$.
- Now we can use MLE to estimate parameters
- Now, there are regression coefficients $\beta$, as well as cutoffs $\tau$
- Statistical software will return estimates for both

# Ordered Logit

- What do the cutoffs $\tau$ mean?
- Recall that $P(Y = 1|X) = logit^{-1}(\tau_1 - \beta X)$
- And $P(Y = 2|X) = logit^{-1}(\tau_2 - \beta X) - logit^{-1}(\tau_1 - \beta X)$
- We add then together:

$$P(Y = 1|X) + P(Y = 2|X) = P(Y \leq 2|X) = logit^{-1}(\tau_2 - \beta X) \quad (7)$$

- And take the logit:

$$logit\,(P(Y \leq 2)) = \tau_2 - \beta X \quad (8)$$

- The rest is similar

$$logit\,(P(Y \leq j)) = \tau_j - \beta X$$

- In this way, $\tau$ looks like intercepts in normal regressions; so some other software (R) call them intercepts

# Multinomial Logit: categorical outcome

- Multinomial logit: for categorical outcomes that have no intrinsic order
- We extend logistic regression in a different way
- $Y$ has $J$ levels, from 0 to $J-1$
- For logistic regression, $P(Y=1|X) = logit^{-1}X\beta = \frac{exp(X\beta)}{1+exp(X\beta)}$
- For multinomial logit, we make similar assumptions about $P(Y=j|X)$

$$P(Y=j|X) = logit^{-1}X\beta_j = \frac{exp(X\beta_j)}{1+\sum_{j=1}^{J} exp(X\beta_j)} \tag{9}$$

- And for reference group, its

$$P(Y=0|X) = logit^{-1}X\beta_j = \frac{1}{1+\sum_{j=1}^{J} exp(X\beta_j)} \tag{10}$$

# Multinomial Logit

- For all levels except the reference group, it has its own regression coefficients
- Say we have 7 categories and 4 predictors (each of them is continuous), then in total we will have 6 * 5 = 30 coefficients
  - 6 = 7 - 1
  - 5 = 4 + 1 (plus intercepts)
- Also because we know what $P(Y = j|X)$ looks like for every possible value of $Y$, we can use MLE to estimate $\beta_j$

# Interpreting multinomial logit

- Based on the assumptions of multinomial, it is easy to see:

$$\frac{P(Y=j|X)}{P(Y=0|X)} = exp(X\beta_j) \tag{11}$$

- Therefore, one unit increase in $X$ leads to $exp(\beta_j)$ increase in odds ratio of $Y=j$ occurring, relative to $Y=0$

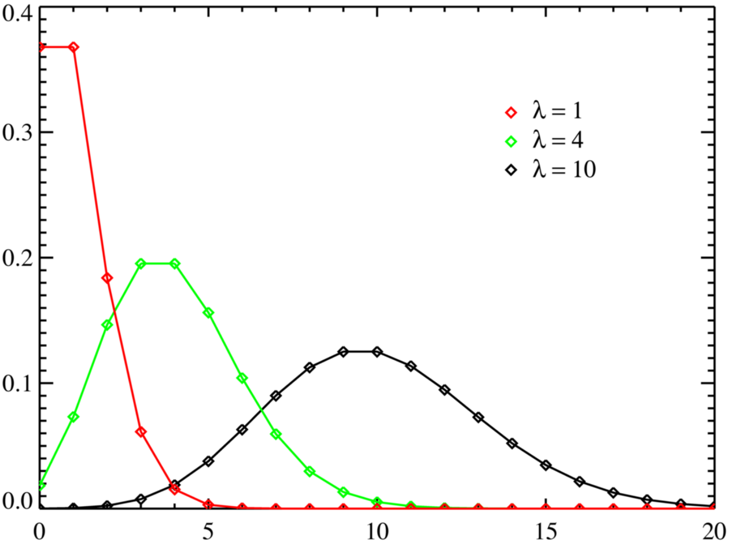
# Poisson Distribution

- Example: $Y$ is event count
  - e.g., number of times each person visit a physician)
  - Number of new born / decease in a country
  - Usually small counts are more likely than large counts
- Key difference: $Y$ are non-negative integers; in linear regression $Y$ is assumed to be continuous variable between $(-\infty, \infty)$
- Event count usually follows Poisson distribution

$$Pr(X = k) = \frac{\tau^k e^{-\tau}}{k!}$$

- $k! = k(k-1)(k-2)\cdots 1$ is factorial
- Property: $E(X) = V(X) = \tau$

# Poisson Distribution



# Poisson Regression

- The conditional probability $P(Y|X)$ is assumed to be distributed according to Poisson:

$$P\left(Y = y|X\right) = \frac{\exp\left(-\tau\right)\tau^{y}}{y!}, \quad y = 0, 1, 2, \ldots \qquad (12)$$
$$\tau = \exp\left(X\beta\right)$$

- And the conditional expectation $E(Y|X)$ is given by:

$$E(Y|X) = \tau = exp(X\beta) \qquad (13)$$

# Poisson Regression (cont'd)

- Why don't we explicitly write $E(\epsilon) = 0$ and $E(\epsilon X) = 0$ as in the Assumption 1 and 2 of linear, logistic and probit regressions?
  - Hint: our assumption of the form of $P(Y|X)$ is very strong
  - It directly gives what $E(Y|X)$ should look like
  - And $E(\epsilon) = 0$ and $E(\epsilon X) = 0$ are essentially the property of $\epsilon = Y - E(Y|X)$
  - So in many textbooks, when introducing generalized linear models, they will omit Assumptions 1 and 2, since it is implied by the assumption of the function form of $P(Y|X)$
- Poisson assumption implies that the data is heteroskedastic:

$$
\begin{aligned}
V(\epsilon|X) &= V(Y - E(Y|X)|X) \\
&= V(Y|X) \\
&= exp(X\beta)
\end{aligned}
\tag{14}
$$

# Poisson and Log-Linear model

- Poisson regressions:

$$E(Y|X) = \tau = exp(X\beta)$$

- An alternative way is to take log at both side of the equation

$$\log E(Y|X) = log(\tau) = X\beta$$

- It means that the link function of Poisson regression is log
- Sociologists and demographers call $\log E(Y|X) = log(\tau) = X\beta$ as log-linear model

# MLE for Poisson regression

1. $P(Y = y|X) = \frac{\exp(-\tau)\tau^y}{y!}$; $\tau = exp(X\beta)$
2. Likelihood is: $L = \prod_{i=1}^{N} \frac{\exp(-\tau_i)\tau_i^y}{y!}$
   - and log-likelihood is :

$$\sum_{i=1}^{n} y_i X_i' \beta - \exp(X_i'\beta) - \log y_i!$$

3. try to maximize by setting the derivative to be 0

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} \left(y_i - \exp(X_i'\beta)\right) X_i = 0$$

- There is no closed-form solution, unfortunately. Numerical optimization is required.

# Interpretation of Poisson Regression

- In log-linear model format:

$$\log E(Y|X) = log(\tau) = X\beta$$

- One unit increase in $X$ leads to $\beta$ increase of the average of $y$ in its log scale
- In Poisson regression format:

$$E(Y|X) = exp(X\beta)$$

- One unit increase in $X$ leads to $exp(\beta) - 1$ increase in $Y$
- One unit increase in $X$ multiplies the mean of $Y$ by a factor $exp(\beta)$
- The ratio between the new $Y$ and old $Y$ is $exp(\beta)$, on average

# Over-dispersion of Count Data

- Poisson regression assumes that $P(Y|X)$ follows a Poisson distribution
- Recall that Poisson distribution assumes that the mean and the variance is the same
- Sometimes we have data whose variance is bigger than mean
- E.g., Long, J. Scott. 1990. *The Origins of Sex Differences in Science*. Social Forces. 68(3):1297-1316.
- The outcome is the number of published articles by a Ph.D. student in biochemistry
- The mean number of articles is 1.69 and the variance is 3.71, a bit more than twice the mean.
- Why? There are always super-starts :) and people who publish nothing : (

# Zero-inflated Poisson Regression

- One common situation of over-dispersion: there are a lot of zeros in the outcome $Y$ and a few big values, which boosts the variance of outcome
- Example: civil war as outcome.
- Zero-inflated Poisson Regression is designed to address this issue
- It assumes that data has two generating processes
  1. With probability $1-\lambda$, the data is generated according to Poisson with mean $\tau$
  2. With probability $\lambda$, we generate excess zeros.
- The final conditional probability is

$$P\left(Y=y|X\right)=\lambda+\left(1-\lambda\right)\frac{\exp\left(-\tau\right)\tau^{y}}{y!}$$

# Zero-inflated Poisson Regression (cont'd)

- With the assumptions in the previous slide
- $E(Y|X) = (1-\lambda)\tau$
- $V(Y|X) = (1-\lambda)\tau(1+\tau\lambda)$
- $V$ is bigger than $E$, of a ratio of $1+\tau\lambda$
- Essentially, zero-inflated Poissin regression is the mix of two regressions:
  - One Poisson regression, with prob $1-\lambda$
  - One logistic regressions (0 and all others), with prob $\lambda$
  - Each regression has its own coefficients
- So it is a more complex model than negative binomial regression, which adds only one additional parameter

# Negative binomial regression

- Another way to deal with over-dispersion: choose a different functional form about $P(Y|X)$

$$P(Y = y|X) = \frac{\Gamma(\alpha + y)}{y!\Gamma(\alpha)(\tau + \alpha)^{\alpha+y}} \tag{15}$$

- And $\tau = exp(X\beta)$
- $\Gamma$ is Gamma function, an extension of factorial
- With this more complex parametric assumption
- $E(Y|X) = \tau$ (similar to Poisson regression)
- $V(Y|X) = \tau(1 + \frac{1}{\alpha}\tau)$
- Positive $\alpha$ ensures that variance is bigger than the mean

# Other count data model

- Zero truncated regressions
  - Say, the outcome of the length of stay in a hospital, which is at least 1 day
  - Zero-truncated Poisson:
    - Remove the probability $P(y = 0)$ because it's not possible)
    - Re-scale the rest of the probability distribution to make it sums to 1

# How do we choose between models?

- Let us use our example of number of published articles by Ph.D. biochemists
- We can choose between three models:
  - Poisson regression
  - Negative binomial regression
  - Zero-inflated Poisson regression
- Decide whether or not to use Poisson regression is relative easier: (Cameron and Trivedi, "Regression-based tests for overdispersion in the Poisson model", *Journal of Econometrics*, 1990)
- Assume $E(Y|X) = \tau$, then
- Null Hypothesis: $V(Y|X) = E(Y|X) = \tau$
- Alternative Hypothesis: $V(Y|X) = \tau + c\tau$
- Cameron and Trivedi's overdispersion test just seeks to examine whether $c = 0$
- (For R users: `dispersiontest` in AER package)

# Use Likelihood for Hypothesis Testing

- But how can we compare negative binomial regression vs zero-inflated Poisson regression?
- We can compare Likelihood among similar models to choose the best one
- Intuition:
  - Likelihood $L$ represents the joint probability that we observe the entire data, given our parameters
  - Assume we have two models
  - A better model should have larger likelihood

# Likelihood Ratio Test

- Define Likelihood Ratio Test Statistics $D$ as:

$$\begin{aligned} D &= -2\log\frac{L_{\text{null}}}{L_{\text{alternative}}} \\ &= 2(\log L_{\text{alternative}} - \log L_{\text{null}}) \end{aligned} \tag{16}$$

- For comparing models, null model is often the simpler model, and alternative model is often the more complex model
- Null Hypothesis: $D = 0$
- Alternative Hypothesis: $D > 0$
- The bigger the $D$, the more evidence for the alternative model

# Likelihood Ratio Test (cont'd)

- Wilk's Theorem (1938): $D$ has an $\chi^2$ -distribution, with degrees of freedom equal to the difference in number of parameters between alternative model and the null model, if the null model is nested within the alternative model
- Nested basically means that the null model can be viewed as a simple case of the alternative model
  - e.g., null is logistic regression with 5 variables; alternative adds another variable
  - null is Poisson; alternative is negative binomial or zero-inflated Poisson
- For non-nested models, Wilk's Theorem does not hold; we need something else (shortly)

# Likelihood Ratio Test (cont'd)

- How do express Wilk's Theorem in the p-value language?
  - Say we get a $D = 12$, and the degree of freedom is 2
  - Definition: the probability of obtaining a test statistics that equals to $D$ or higher is approximately $p \iff$ p-value is $p$
  - $P(D < 12, d.f. = 2) = 0.9975$
    - in R, just type `pchisq(12, 2)`, which is the cumulative probability distribution of $D$
    - It means that the probability of observing a $D$ smaller than 12 is 0.9975
    - So the probability we observe a $D$ equal to or larger than 12 is 1 - 0.9975 = 0.0025, which is our p-value)

# Bias-Variance Trade-Off and Likelihood Ratio Test

- But, a more complex model (adding more parameters) usually can predict more accurately and thus often always have larger likelihood
- AIC: Akaike information criterion (named after Hirotugu Akaike, 1974); reaching balance between predictive power and model complexity
- $k$ is the number of parameters in a model

$$\text{AIC} = 2k - 2\log L \tag{17}$$

# Bias-variance trade-off

- AIC wants to balance predictive power and model complexity
- This is a fundamental idea in machine learning
- The idea is called bias-variance trade-off
- Recall:
  - We use $g(X)$ to predict $Y$;
  - Among all possible $g(X)$, $E(Y|X)$ is the best predictor of $Y$ because it minimizes mean squared error $E[(Y - g(X)^2]$
  - and $\hat{g}(X)$ is estimator of $g(X)$ based on sample data
- Now we compare the mean squared error between $Y$ and its empirical prediction $\hat{g}(X)$, $E[(Y - \hat{g}(X))^2]$

# Bias Variance Decomposition

$$E[(Y-\hat{g}(X))^2] = V\left[Y-E(Y|X)\right] + \left[\hat{g}(X)-E(Y|X)\right]^2 + V(\hat{g}(x))$$

$=$ variance of irreducible error $+$ (prediction bias)$^2+$ prediction variance

- Variance of irreducible error: this only relates to your data;
  - They are irreducible as long as you have selected $X$; it will be large if $X$ has nothing to do with $Y$.
- Prediction bias: relating to your model
  - How current estimator $g(X)$ differs from the best predictor $E(Y|X)$
  - OLS is often bad at approaching $E(Y|X)$
- Prediction Variance: relating to your model
  - It roughly indicates how varied your predictions can be
  - OLS actually has small prediction variance

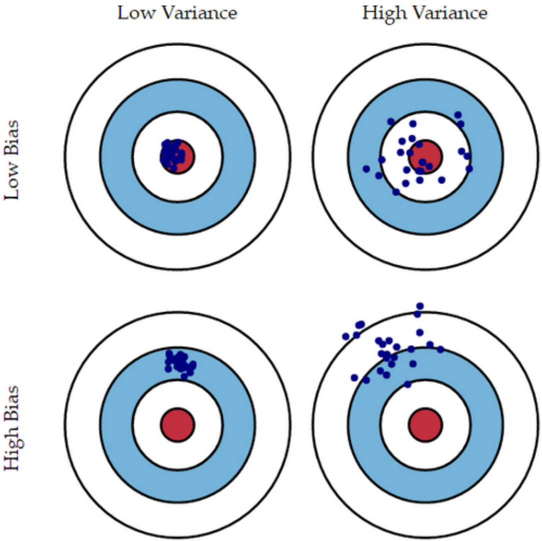
# Bias Variance Decomposition (cont'd)

- Variance of Estimator: $V(\hat{g}(x))$
  - It is not the variance of estimated parameters $\hat{V}(\hat{\beta})$; it's the variance of your predicted values!
  - One intuition: the population has 10000 individuals, and each time you sample 100 individuals, and fit an OLS regression.
    - These OLS fitted lines would not vary a lot.
    - But if you use a very complex model, each time predictions can change a lot; thus prediction variances can be high

# Bias Variance Trade-off

- To reduce irreducible error: find more predictive $X$
- The other two quantities relate to your model (estimator):
  - Simple models (like OLS) have large estimator bias, but small estimator variance
  - Complex models have small estimator bias, but large estimator variance

# Bias vs Variance (illustration)

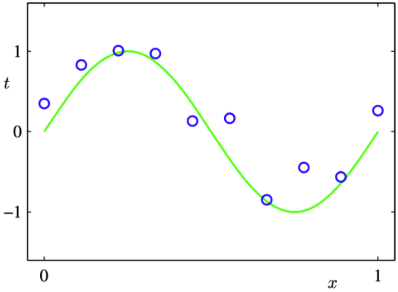


# Bias Variance Trade-off (example)

- We have a linear regression with only one variable $X$, but we add higher order terms

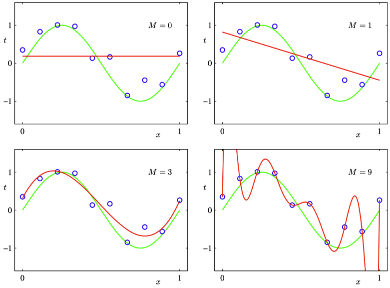
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_M X^M$$

- The true $M$ is 3; simulate some data
- Then we try different OLS models by adding more and more high-order terms



# Bias Variance Trade-off (cont'd)

- $M = 1$, fits the data very bad (high prediction bias)
- $M = 9$, fits the data so well (small prediction bias), but it is highly sensitive to small changes in observations
  - The prediction on new data can be very bad
  - This is known as over-fitting
- $M = 3$, it achieves a good balance between prediction bias and variance
  - And it actually is the correct $M$



# Bias Variance Trade-Off

- Simple model predicts the data very bad (high prediction bias)
- Complex model predicts the data too well (low prediction bias), but it has high estimation variance and is does not generalize well
    - If social science research care about policy implications, generalizability is important.
- Ideal predictive models should balance the prediction bias and variances
- And this principle has been used in many statistics/machine learning applications
    - AIC is one example
    - We will see more next week

# Today's Review

| Type of $Y$ | Regression to use |
|---|---|
| Continuous | linear |
| Binary | logit/probit |
| Categorical | multinomial logit / ordered logit |
| Count (integer) | Poisson, negative binomial and zero-inflated |

# Recommended Readings

- There are many other GLMs (e.g., censored outcome).
- GLM
  - `https://data.princeton.edu/wws509`, Generalized Linear Models course by Germán Rodríguez
  - Powers, Daniel, and Yu Xie. *Statistical methods for categorical data analysis*. Emerald Group Publishing, 2008.
- Machine Learning:
  - Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer
  - `https://web.stanford.edu/~hastie/ElemStatLearn/`
  - Bias-variance decomposition is discussed in Chapter 2