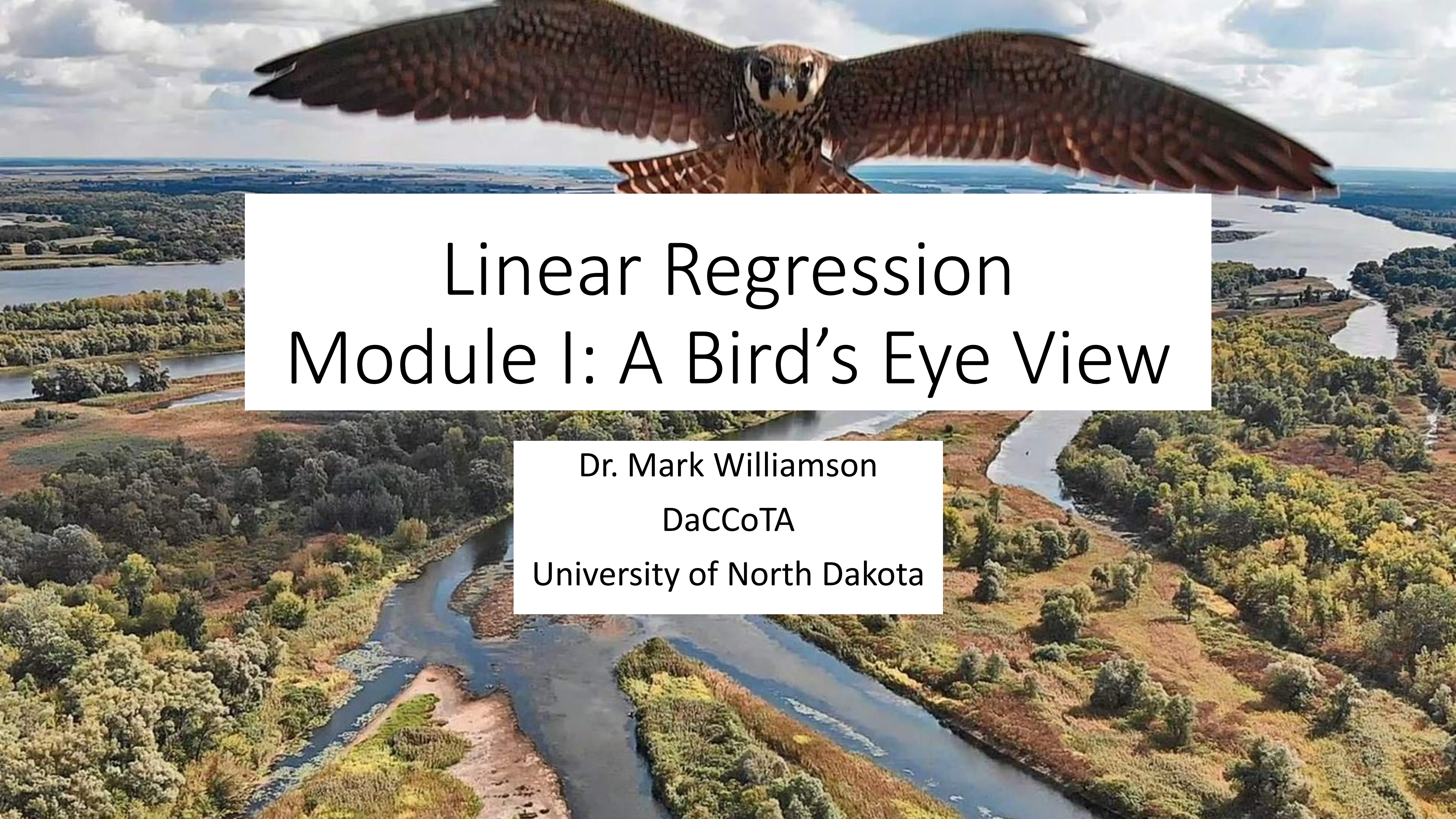


Linear Regression
Module I: A Bird’s Eye View

Dr. Mark Williamson
DaCCoTA
University of North Dakota

Introduction

Linear regression models the relationship between a response variable and one or more predictor variables

Dependent Variable

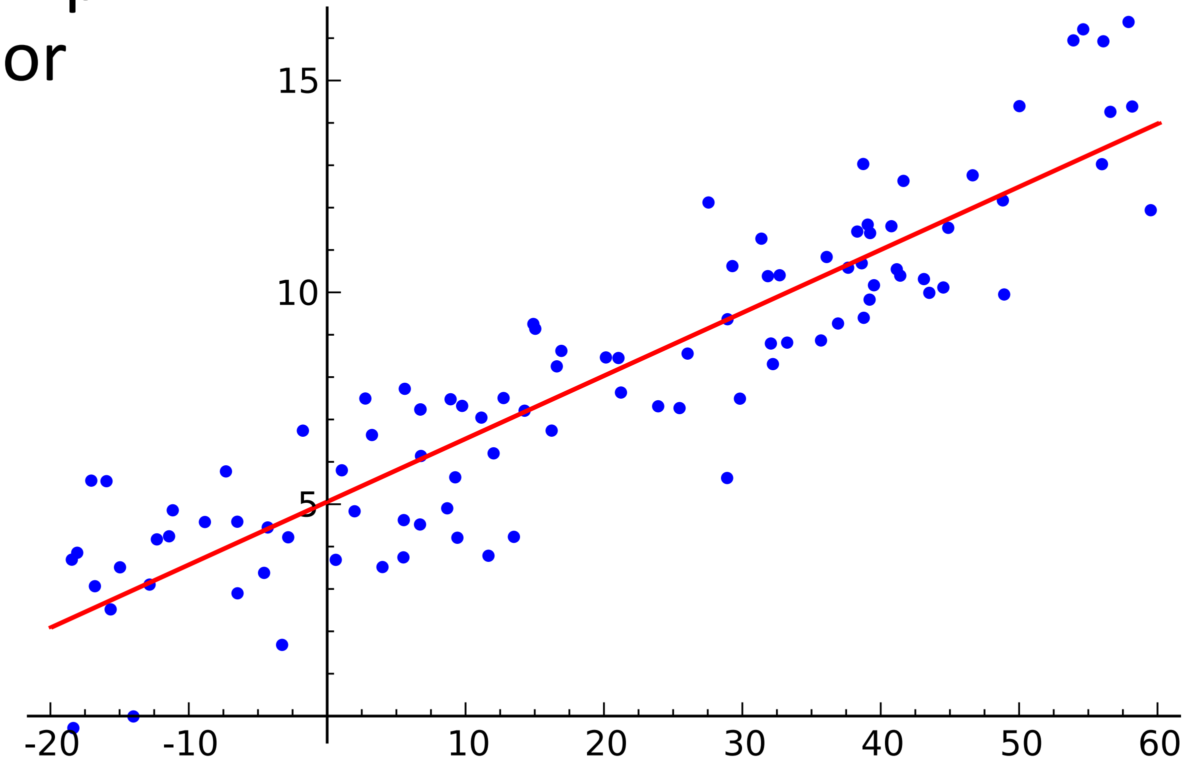
Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$



$$\frac{signal}{noise}$$

Landscape

Linear Regression

Broad Structure

- Single response and single predictor (simple linear regression)
- Single response and multiple predictors (multiple linear regression)
- Multiple responses and predictors (multivariate linear regression)

Predictor variable attributes

- Numerical/categorical (recoding)
- Higher order terms (polynomial regression)
- Fixed and random predictor variables (mixed model)
- Nested predictor variables (hierarchical model)

Response variable attributes

- Normally distributed (Gaussian regression)
- Categorical response (Logistic or Ordinal regression)
- Count data (Poisson, Negative Binomial, or Quasi Poisson regression)
- Time to event (Cox regression)

Other Considerations

- Non-mean (Quantile regression)
- Censoring (Tobit regression)
- Collinearity or overfitting issues (Ridge, Lasso, Elastic net, Principle Components, or Partial Least Squares regression)
- Time trends or similar gradients (Piecewise, Join-point regression)

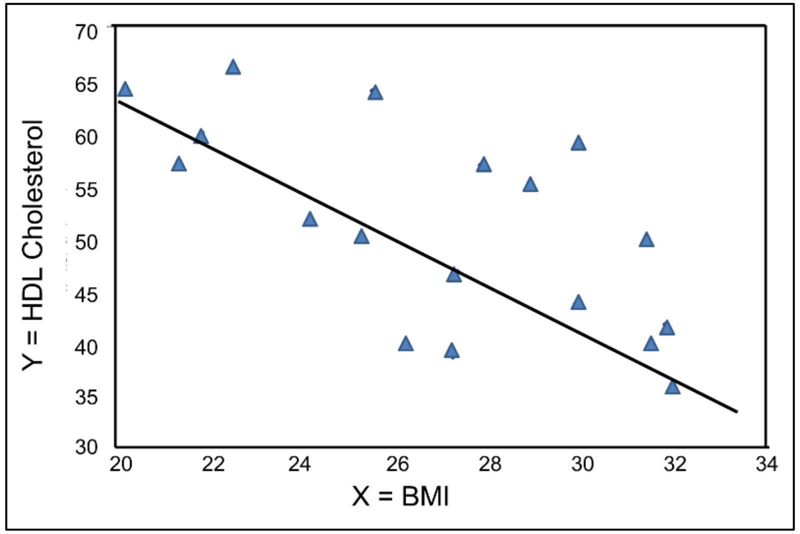
Structures and Uses

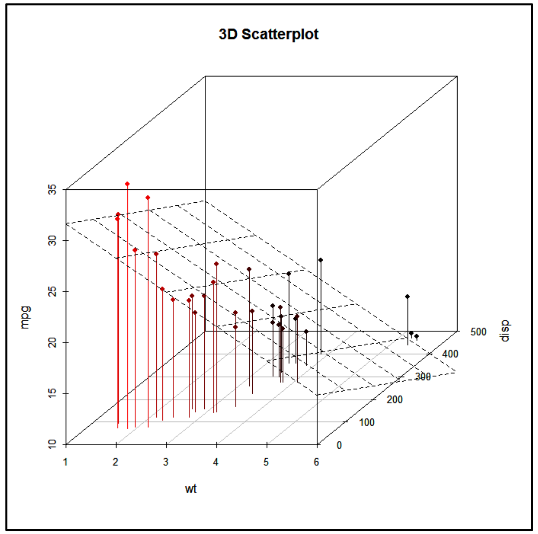
Broad Structure

- Single response and single predictor (simple linear regression)
- Single response and multiple predictors (multiple linear regression)
- Multiple responses and predictors (multivariate linear regression)

Can Y be predicted by X?
Can Y be predicted by X1, X2, X3…?
Can Y1, Y2, Y3… be predicted by X1, X2, X3…?

Can Weight be predicted by Height?
Can Cancer Risk be predicted by Smoking Rate, BMI, and Age?
Can Ice Cream, Canned Food, and Hotdog sales be predicted by Temperature, Storm Chance, and Gas Price?





DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND UNIVERSITY OF NORTH DAKOTA

Structures and Uses

Predictor variable attributes

- Numerical/categorical (recoding)
- Higher order terms (polynomial regression)
- Fixed and random predictor variables (mixed model)
- Nested predictor variables (hierarchical model)

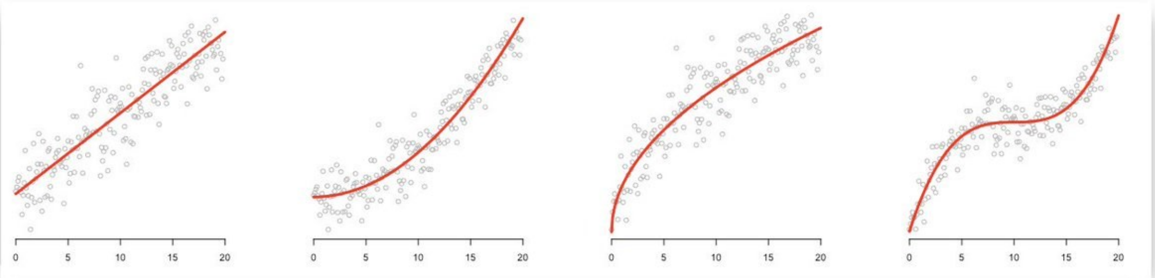
Level of race	New variable 1 (x1)	New variable 2 (x2)	New variable 3 (x3)
1 (Hispanic)	1	0	0
2 (Asian)	0	1	0
3 (African American)	0	0	1
4 (white)	0	0	0

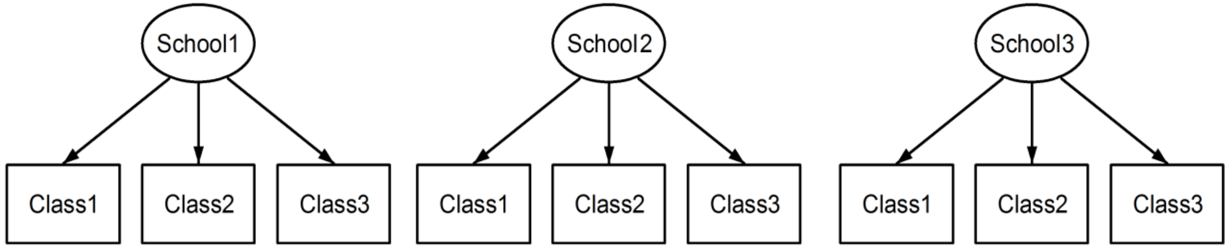
Fixed effects:

- All categories {of interest} are present in the model
- **Example:** 4 treatment groups

Random effects:

- Categories present in the model are subset of the total number of categories
- **Example:** 4 state hospitals

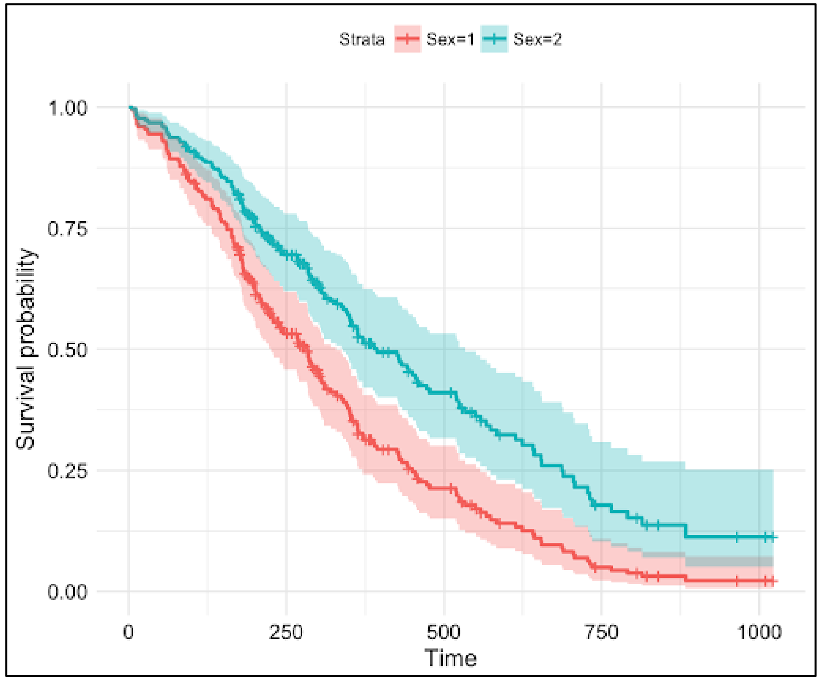


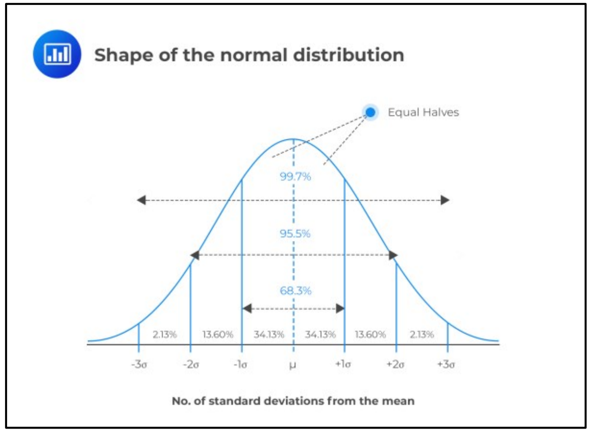


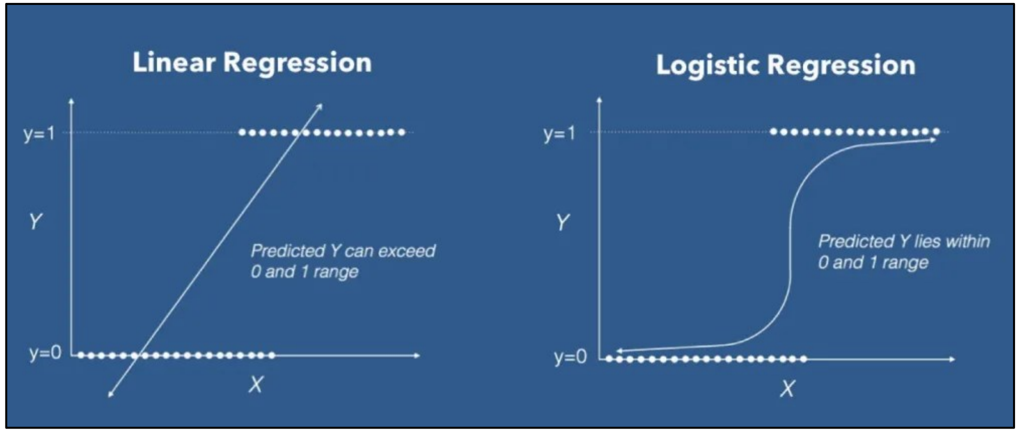
Structures and Uses

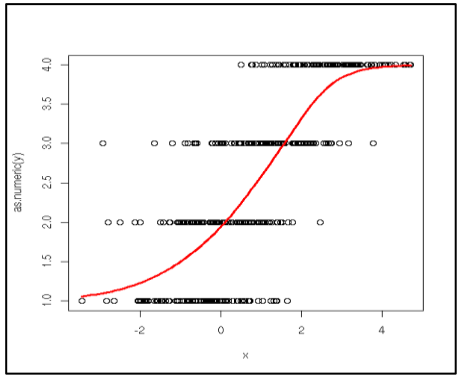
Response variable attributes

- Normally distributed (Gaussian regression)
- Categorical response (Binomial, Logistic or Ordinal regression)
- Count data (Poisson, Negative Binomial, or Quasi Poisson regression)
- Time to event (Cox regression)









Poisson Distribution Formula

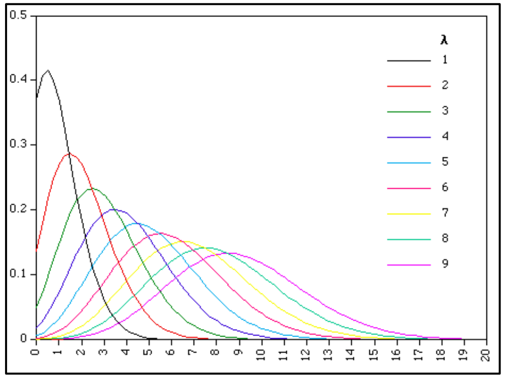
$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where

x = 0, 1, 2, 3, …

λ = mean number of occurrences in the interval

e = Euler's constant $\approx$ 2.71828



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

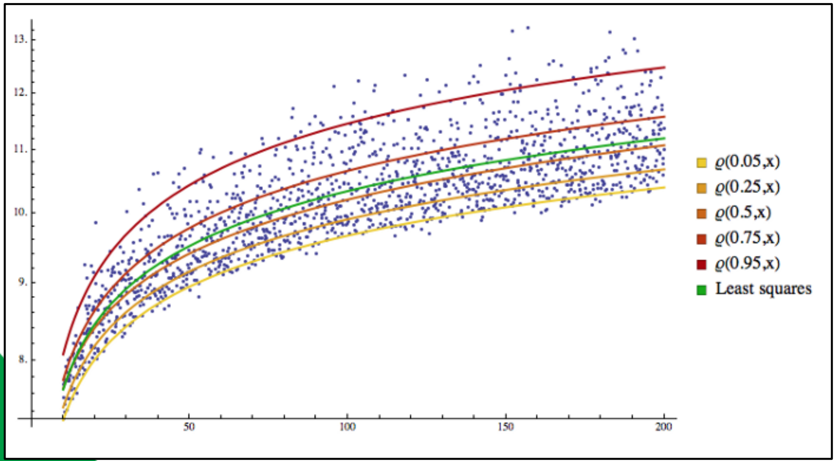
UNIVERSITY OF NORTH DAKOTA

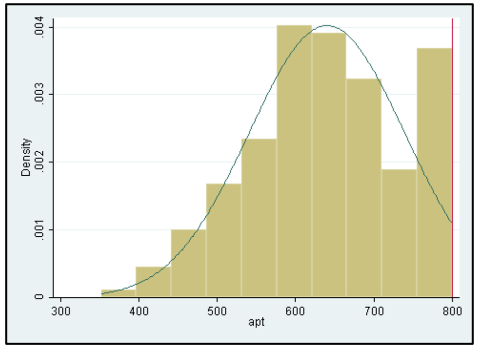
Structures and Uses

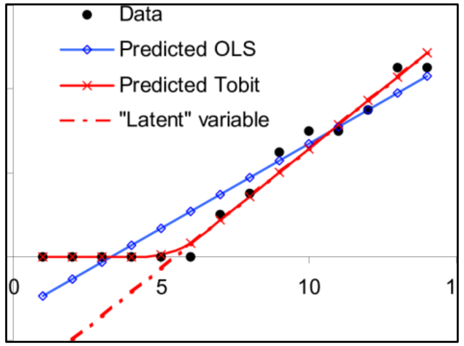
Other Considerations

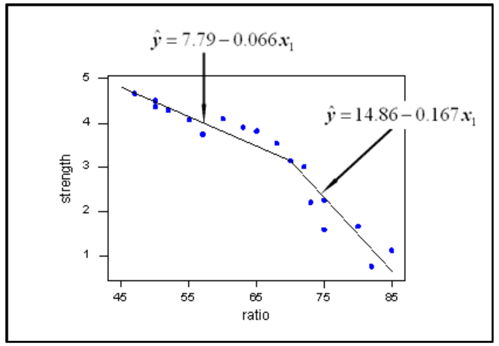
- Non-mean (Quantile regression)
- Censoring exists (Tobit regression)
- Collinearity or overfitting issues (Ridge, Lasso, Elastic net, Principle Components, or Partial Least Squares regression)
- Time trends or similar gradients (Piecewise, Join-point regression)

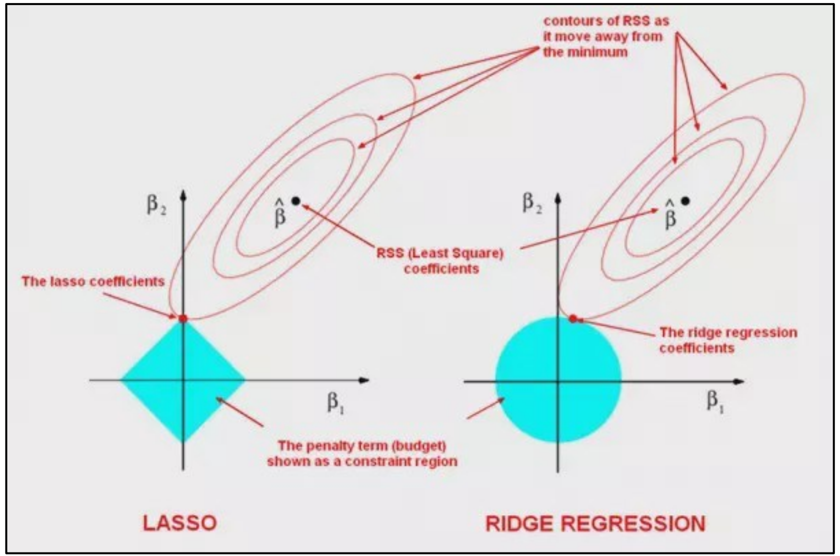
Quantile regression is an extension of linear **regression** that is used when the conditions of linear **regression** are not met (i.e., linearity, homoscedasticity, independence, or normality

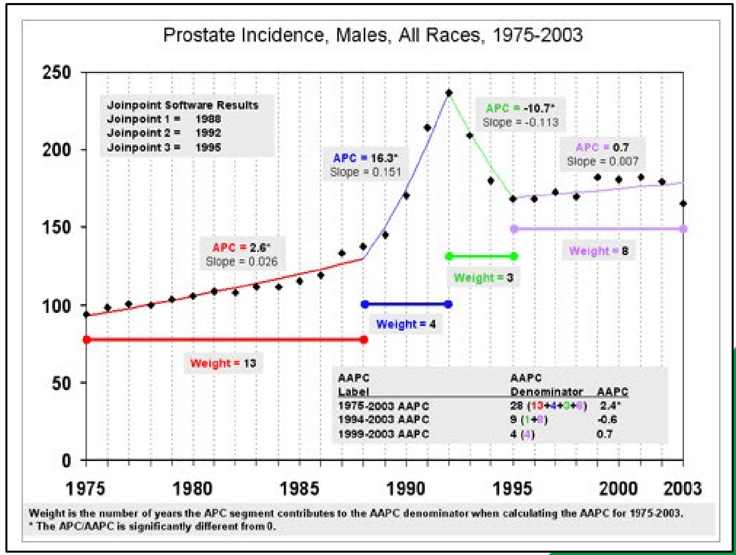












Examples

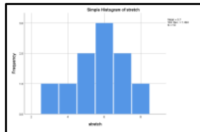
SPSS

Heathip.sav [DataSe

File Edit View Data Transform Analyze Graphs Utilities Extensions Window

15 : discomf

	height	stretch	heatstre	discomf	heatdisc	var
1	1.52	7	11	7	6	
2	1.72	3	5	8	7	
3	1.67	8	8	3	3	
4	1.61	6	12	6	5	
5	1.66	4	15	8	3	
6	1.78	5	11	6	4	
7	1.70	5	8	7	6	
8	1.57	6	12	5	4	
9	1.61	6	11	8	5	
10	1.60	7	14	7	5	
11						



Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Height in metres[b]	.	Enter

a. Dependent Variable: stretch

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.548[a]	.300	.213	1.326

a. Predictors: (Constant), Height in metres

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6.037	1	6.037	3.434	.101[b]
	Residual	14.063	8	1.758		
	Total	20.100	9			

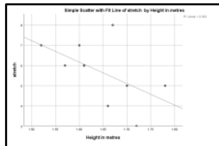
a. Dependent Variable: stretch

b. Predictors: (Constant), Height in metres

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	23.174	9.438		2.455	.040
	Height in metres	-10.629	5.735	-.548	-1.853	.101

a. Dependent Variable: stretch



Examples

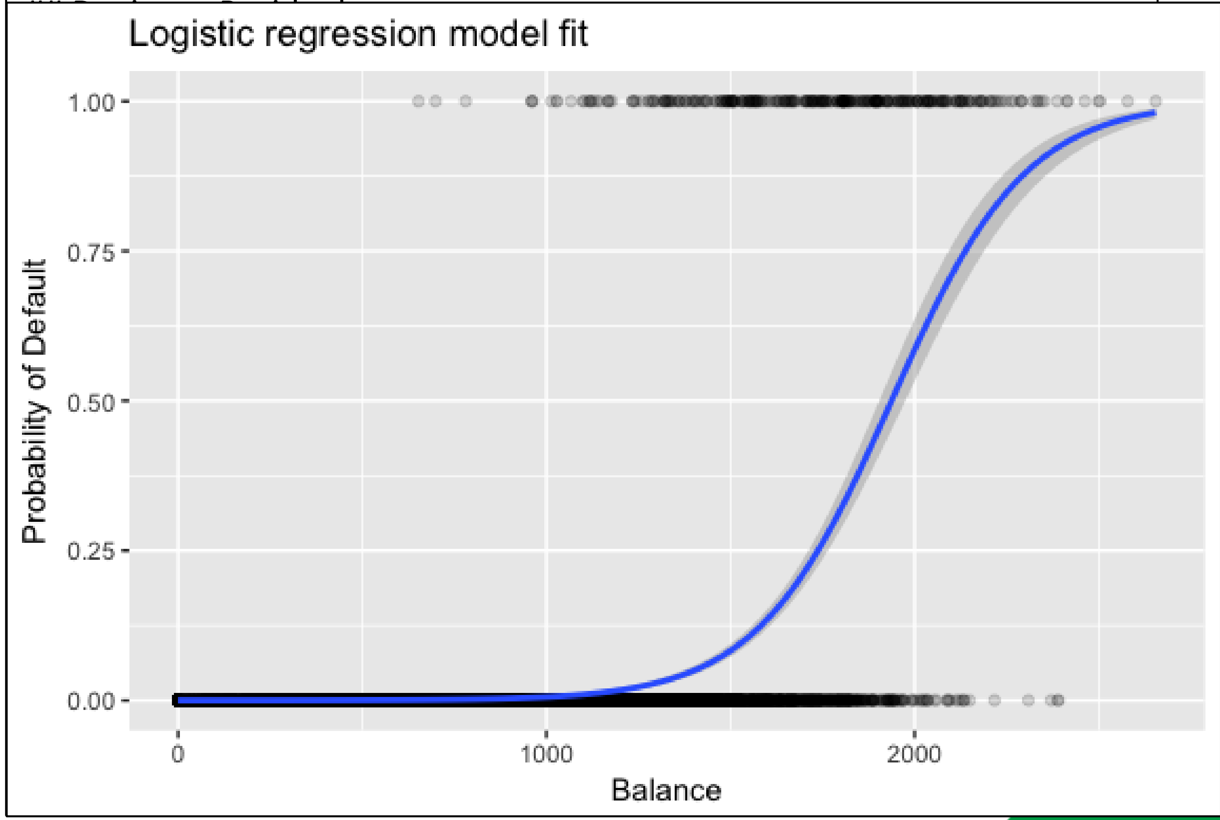
R

```
model1 <- glm(default ~ balance, family = "binomial", data = train)
```

```
summary(model1)
```

```
default %>%
  mutate(prob = ifelse(default == "Yes", 1, 0)) %>%
  ggplot(aes(balance, prob)) +
  geom_point(alpha = .15) +
  geom_smooth(method = "glm", method.args = list(family = "binomial")) +
  ggtitle("Logistic regression model fit") +
  xlab("Balance") +
  ylab("Probability of Default")
```

```
##
## Call:
## glm(formula = default ~ balance, family = "binomial", data = train)
##
```



https://uc-r.github.io/logistic_regression

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Examples

SAS

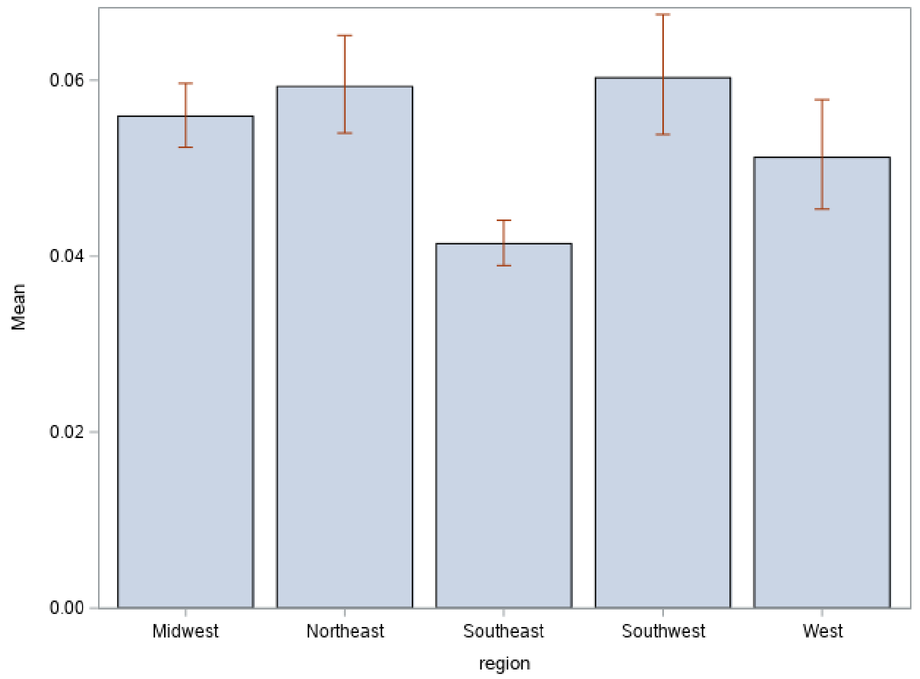
```
PROC GLIMMIX data=covid2;
        where cases>0;
        class region(ref='West');
        model dpc=region
        bed bedut vent famsize perc_white perc_black perc_hisp
        sex_ratio med_age med_income perc_insure perc_poverty
        /solution distribution=beta;
        lsmeans region/ ilink cl;
        ods output LSMeans=lsm8;
PROC SGPLOT data=lsm8;
        vbarparm category=region response=Mu/
        limitupper=UpperMu limitlower=LowerMu;
```

region Least Squares Means

region	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Mean	Standard Error Mean	Lower Mean	Upper Mean
Midwest	-2.8265	0.03509	1838	-80.54	<.0001	0.05	-2.8954	-2.7577	0.05591	0.001852	0.05238	0.05965
Northeast	-2.7643	0.05079	1838	-54.43	<.0001	0.05	-2.8639	-2.6647	0.05928	0.002833	0.05397	0.06509
Southeast	-3.1417	0.03291	1838	-95.46	<.0001	0.05	-3.2063	-3.0772	0.04142	0.001307	0.03893	0.04406
Southwest	-2.7466	0.06136	1838	-44.76	<.0001	0.05	-2.8670	-2.6263	0.06028	0.003476	0.05381	0.06747
West	-2.9191	0.06509	1838	-44.84	<.0001	0.05	-3.0468	-2.7915	0.05122	0.003163	0.04536	0.05779

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
region	4	1838	17.35	<.0001
bed	1	1838	7.92	0.0049
bedut	1	1838	4.96	0.0261
vent	1	1838	0.01	0.9321
famsize	1	1838	0.01	0.9212
perc_white	1	1838	0.35	0.5523
perc_black	1	1838	0.14	0.7034
perc_hisp	1	1838	23.02	<.0001
sex_ratio	1	1838	1.26	0.2624
med_age	1	1838	68.79	<.0001
med_income	1	1838	3.11	0.0779
perc_insure	1	1838	4.57	0.0327
perc_poverty	1	1838	22.50	<.0001



Effect	Pr > |t|
Intercept	<.0001
region	0.2067
region	0.0565
region	0.0041
region	0.0330
region	.
bed	0.0049
bedut	0.0261
vent	0.9321
famsize	0.9212
perc_white	0.5523
perc_black	0.7034
perc_hisp	<.0001
sex_ratio	0.2624
med_age	<.0001
med_income	0.0779
perc_insure	0.0327
perc_poverty	<.0001
Scale	.

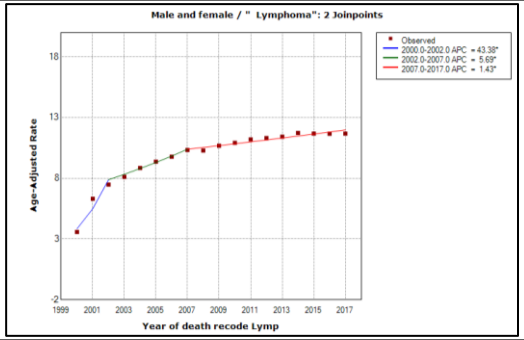
UNIVERSITY OF NORTH DAKOTA

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

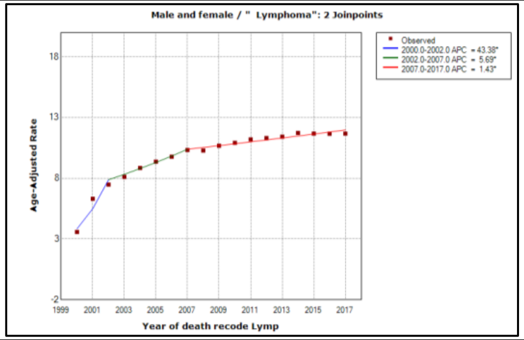
UNIVERSITY OF NORTH DAKOTA

Quick Assessment

1. Multivariate linear regression involves what number of response and predictor variables?
2. Is polynomial regression linear or non-linear? Why?
3. What types of regression can be used if the response variable is not normal (doesn’t follow a Gaussian distribution)? List at least two examples.
4. What type of model includes both fixed and random effects?
5. Fill in the blanks: Linear regression models the relationship between the _______variable and one or more ______ variables
6. What type of regression is pictured to the right?



Quick Assessment

1. Multivariate linear regression involves what number of response and predictor variables?	More than one response variable More than one predictor variable
2. Is polynomial regression linear or non-linear? Why?	Still linear, because regression coefficients are still linear $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_h X^h + \epsilon,$
3. What types of regression can be used if the response variable is not normal (doesn't follow a Gaussian distribution)? List at least two examples.	Logistic, Binomial, Multinomial, Ordinal, Poisson, Negative Binomial, Quasi Poisson, Gamma, Exponential, Cox, etc.)
4. What type of model includes both fixed and random effects?	Mixed model
5. Fill in the blanks: Linear regression models the relationship between the _______variable and one or more ______ variables	Response, Predictor
6. What type of regression is pictured to the right? 	Join-point/piecewise regression

DaCCoTA
DAKOTA CANCER COLLABORATIVE
ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA

Summary and Conclusion

- Linear regression is a fundamental tool in statistical analysis
- Ranges from very basic to very sophisticated
- Understand your data to understand what the best approach is