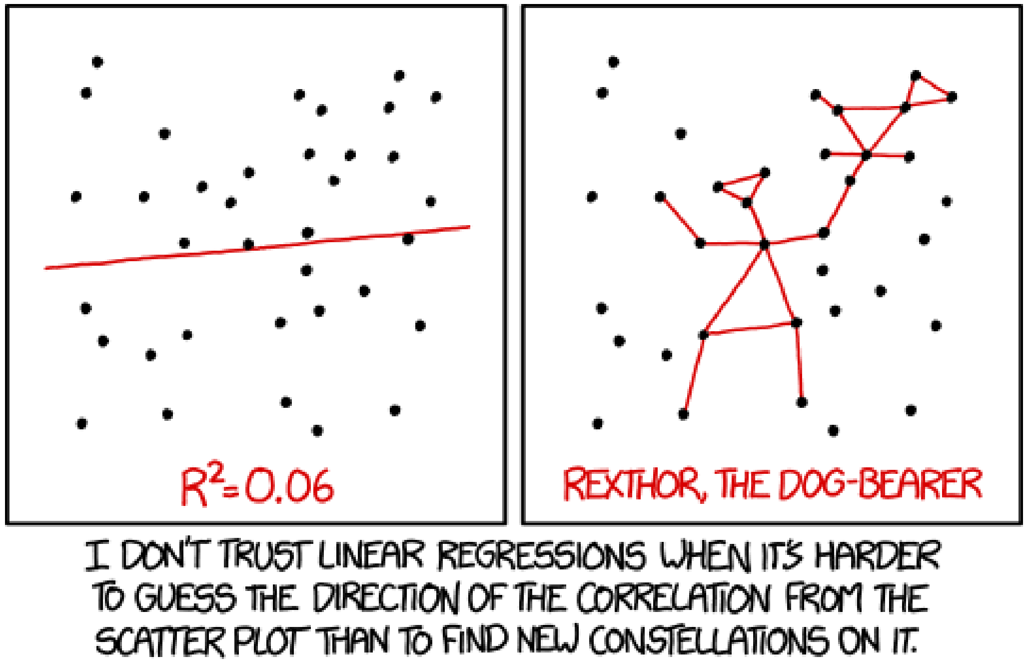
Linear Regression
Module II: Leaves and Trees

Dr. Mark Williamson
DaCCoTA
University of North Dakota

Introduction

- Linear regression is a foundational statistical technique
- Takes on many forms
  - Broad Outline, Predictor Variable, Response Variable, Other Considerations
- Here, we'll look in more details at the underpinning
- Also, we'll go through several examples



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

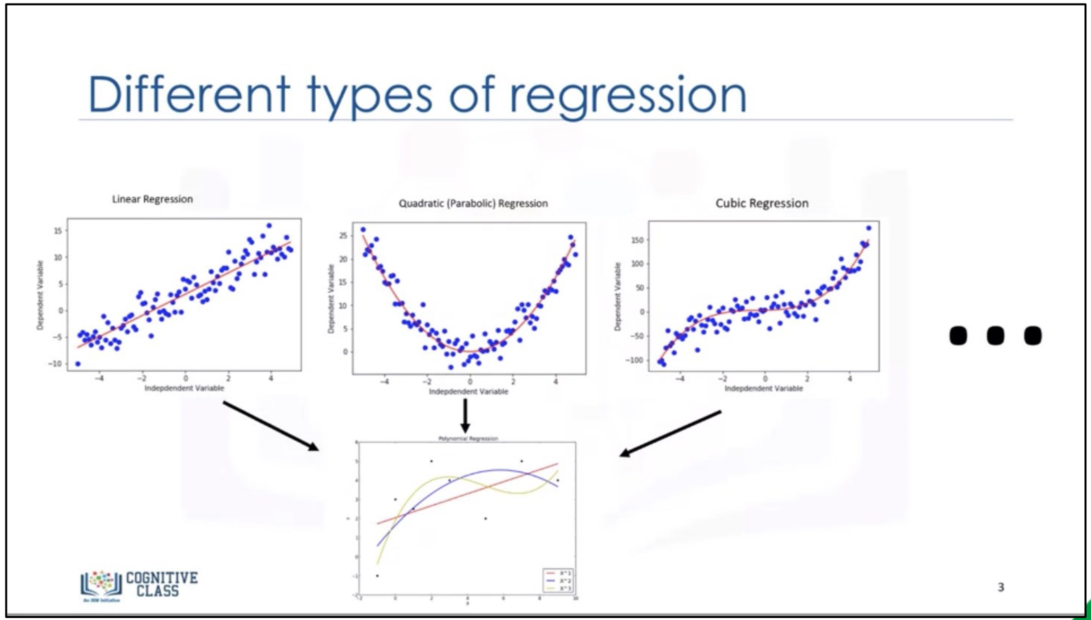
UNIVERSITY OF NORTH DAKOTA

Rationales

When should you use linear regression?

- Want to predict a variable's value
- Want to model the relationship between Y variable and X variable(s)
- Both Y and X and typically numerical
- Expect there to be a linear relationship

X variable(s) → Y variable ↓	Categorical	Numerical	Categorical + Numerical
Categorical	Chi-Square	Regression	Regression
Numerical	ANOVA*	Regression	Regression



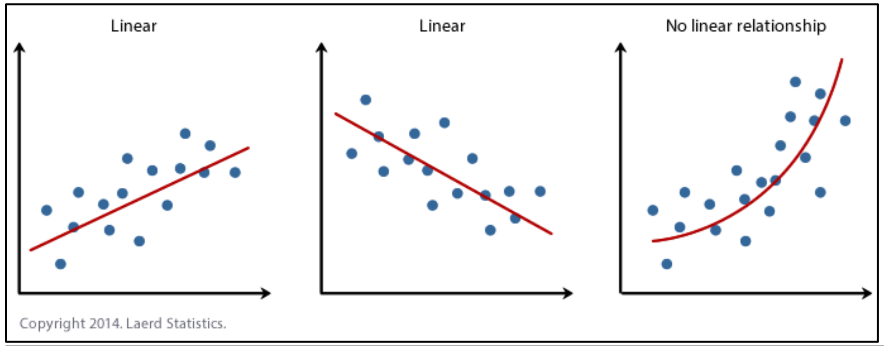
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

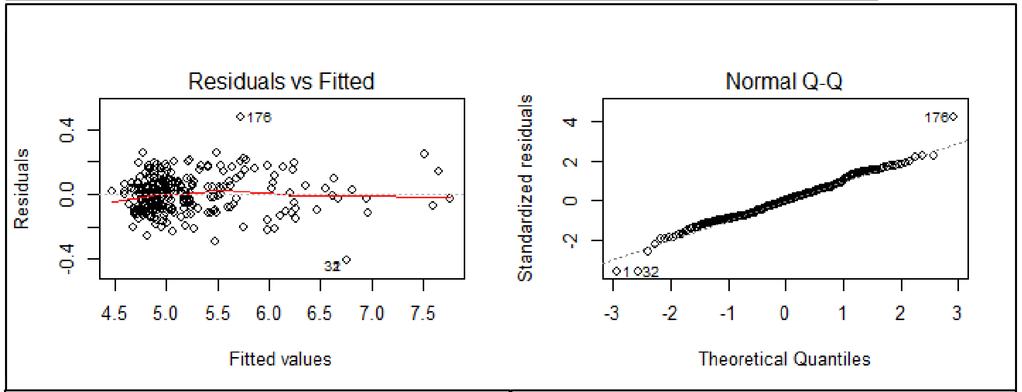
UND
UNIVERSITY OF NORTH DAKOTA

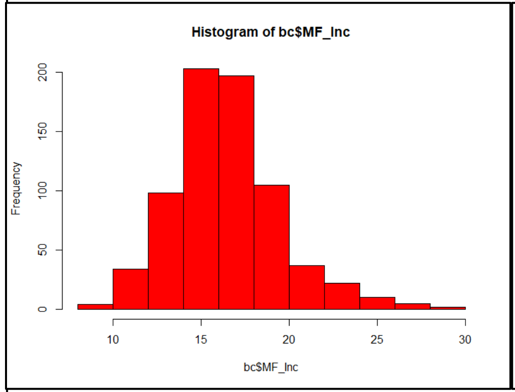
Rationales

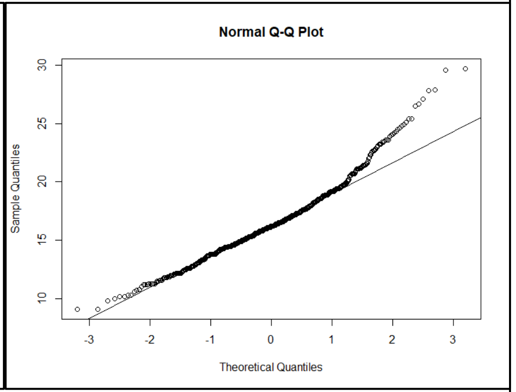
What assumptions are there for basic linear regression?

- **Linearity** – relationship between X and mean of Y is linear
- **Homoscedasticity** – the variance of the residuals is the same for any value of X
- **Independence** – observations do not depend on one another
- **Normality** – for any fixed value of X, Y follows a Gaussian distribution







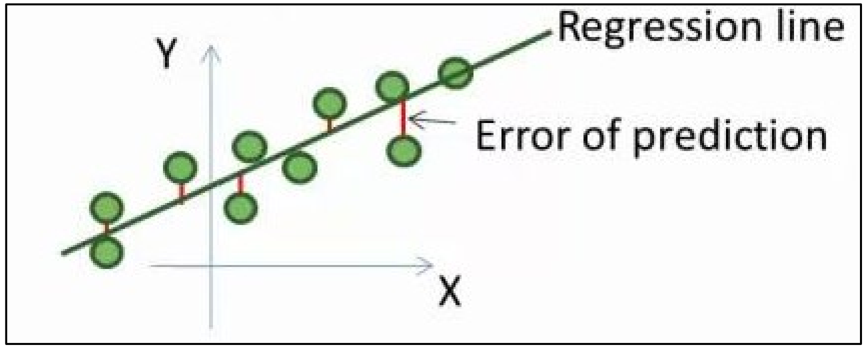


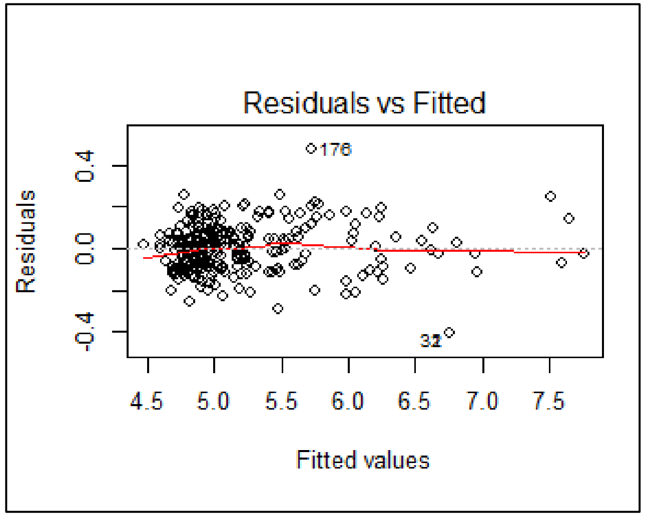
Descriptions

- **Variables:**
  - Y variable -> response, measured, dependent
  - X variable -> predictor, control, independent
- **Fitted/Predicted values ->** values of Y generated by plugging X into model
- **Residuals ->** fitted values minus the actual observed values of Y
- **Ordinary least squares:**
  - Minimizes the squared distance between each Y value and line
  - Creates line of best fit
- **Variable types:**
  - Numerical: discrete, continuous
  - Categorical: ordinal, nominal
  - Fixed and Random
- **Distributions:**
  - Normal/Gaussian
  - Log Normal
  - Binomial
  - Poisson
  - Negative Binomial
  - Beta
  - Etc.

Dependent Variable
Population Y intercept
Population Slope Coefficient
Independent Variable
Random Error term

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

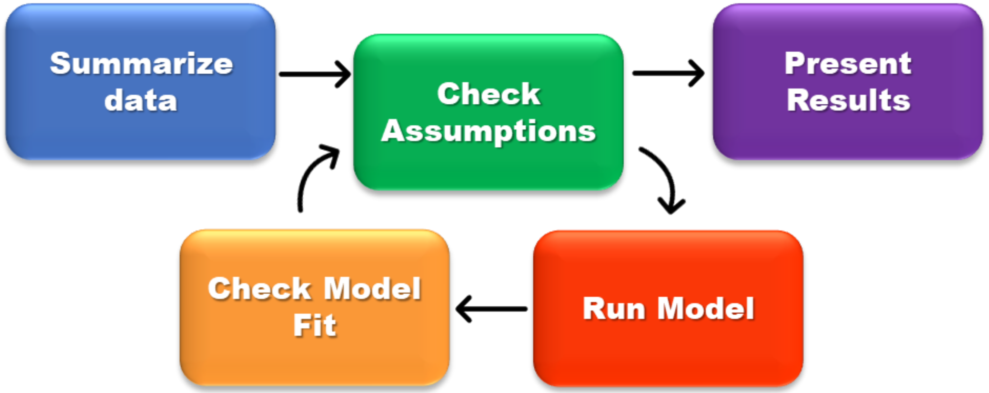




Step-by-step Examples 1

Youth Risk Behaviors Survey

A. Can we predict weight from height?
B. Can we predict weight from height and age?
C. Can we predict weight from height, age, gender, and race/ethnicity?

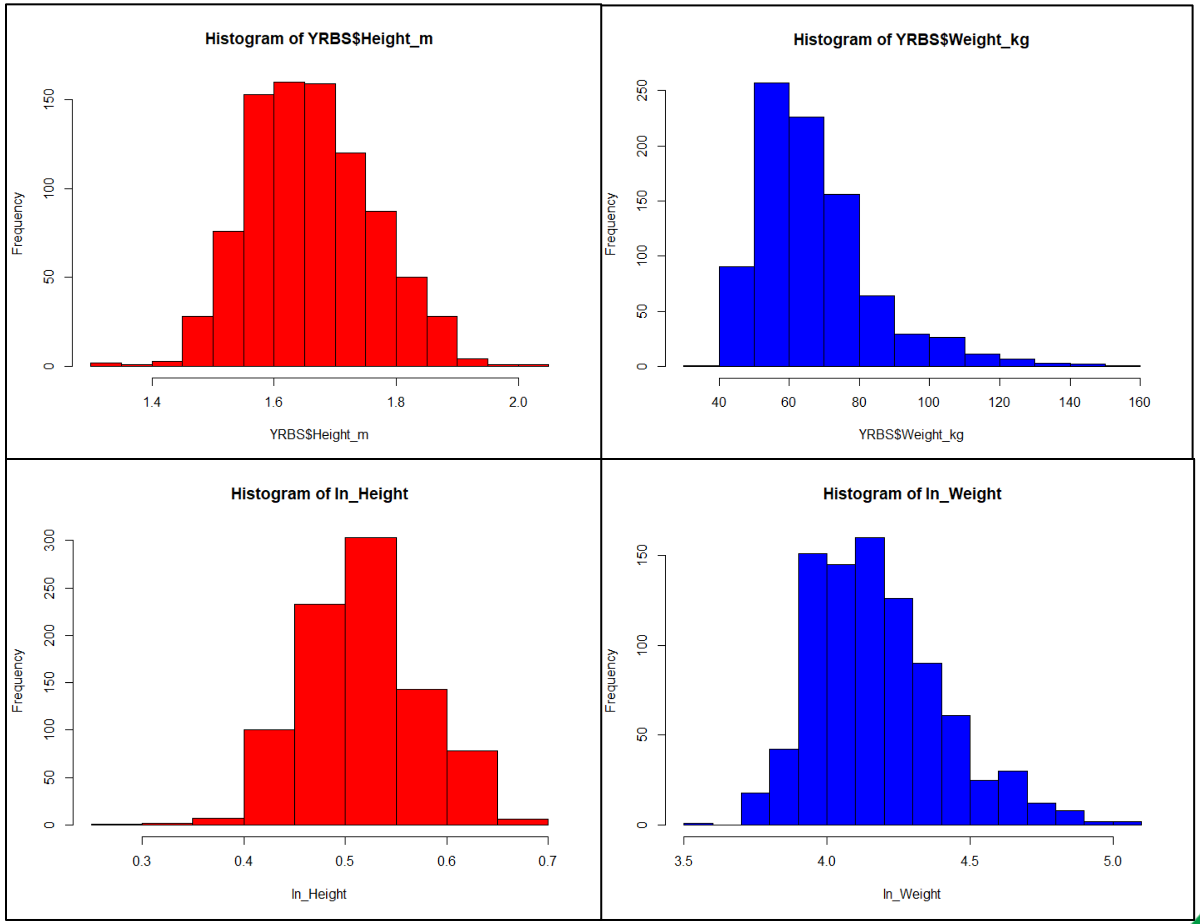


DaCCoTA
DAKOTA CANCER COLLABORATIVE
ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Step-by-step Examples 1

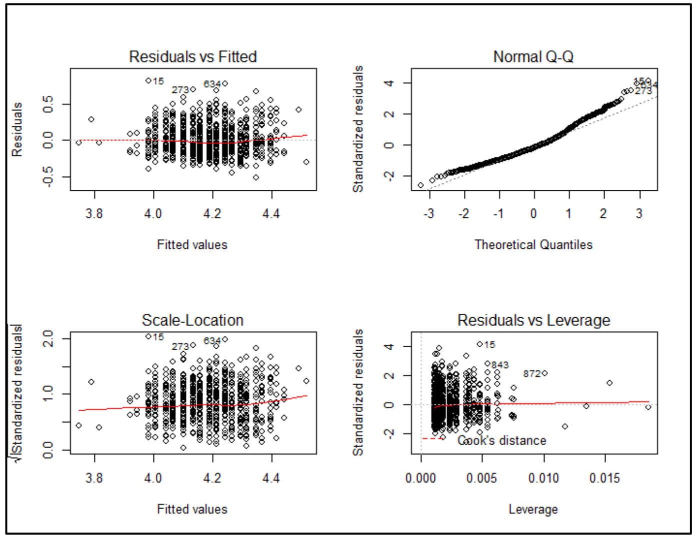
```
>YRBS<-read.csv('YRBS_Example.csv')
>head(YRBS)

>summary(YRBS$Height_m)
>summary(YRBS$Weight_kg)

>hist(YRBS$Height_m, col='red')
>hist(YRBS$Weight_kg, col='blue’)

>ln_Weight<-log(YRBS$Weight_kg)
>ln_Height<-log(YRBS$Height_m)

>hist(ln_Weight, col='red')
>hist(ln_Height, col='blue')
```

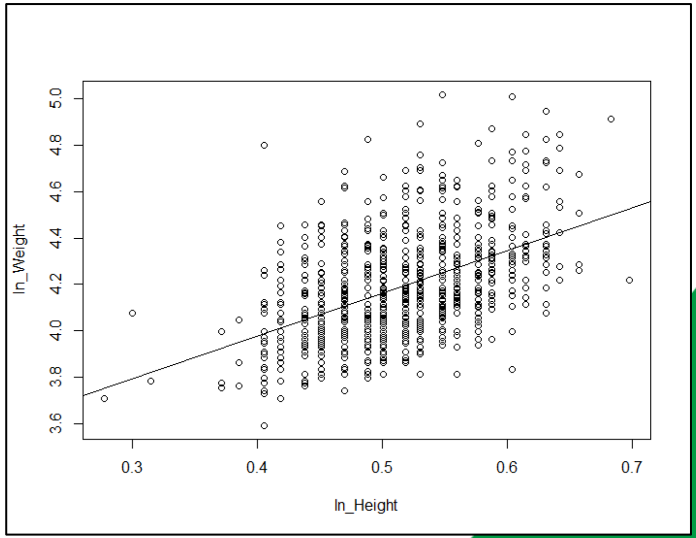


Step-by-step Examples 1

```
#Weight = Height
>lm1 <-lm(ln_Weight~ln_Height)
>summary(lm1)
>par(mfrow=c(2,2))
>plot(lm1)
>par(mfrow=c(1,1))
>plot(ln_Weight~ln_Height)
>abline(3.24,1.84)
```

```
Call:
lm(formula = ln_Weight ~ ln_Height)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.51216 -0.13726 -0.03131  0.10775  0.81904 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.23654    0.05703   56.76   <2e-16 ***
ln_Height    1.83699    0.11025   16.66   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1976 on 871 degrees of freedom
  (114 observations deleted due to missingness)
Multiple R-squared:  0.2417, Adjusted R-squared:  0.2408 
F-statistic: 277.6 on 1 and 871 DF,  p-value: < 2.2e-16
```

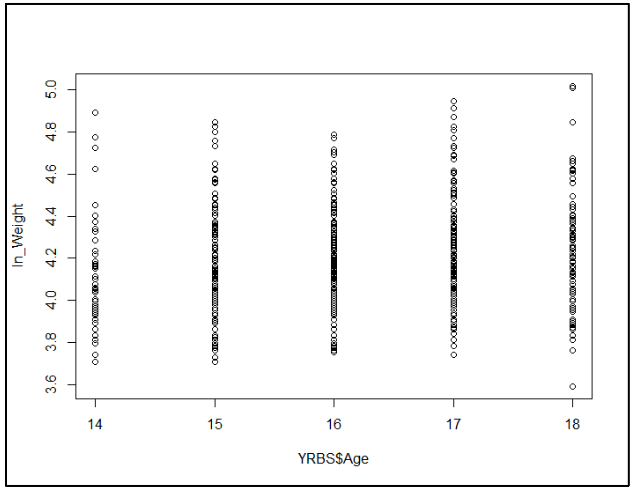


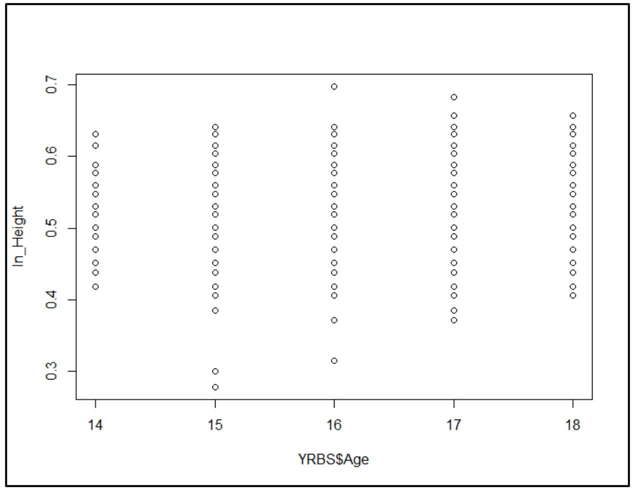


Step-by-step Examples 1

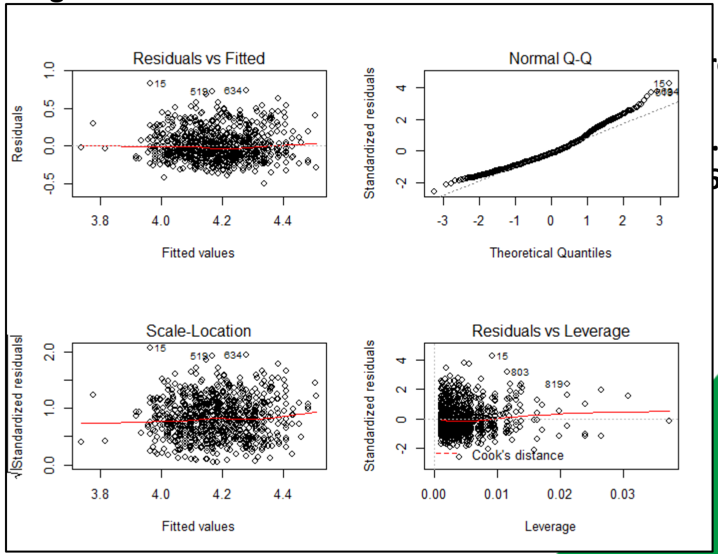
#Weight = Height + Age

```
> plot(ln_Weight~YRBS$Age)
> plot(ln_Height~YRBS$Age)

> lm2<-
  lm(ln_Weight~ln_Height*YRBS$Age)
> summary(lm2)
> par(mfrow=c(2,2))
> plot(lm2)
> par(mfrow=c(1,1))
```





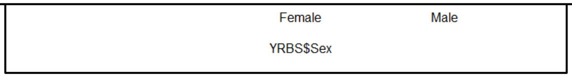

```
Call:
lm(formula = ln_Weight ~ ln_Height * YRBS$Age)

Residuals:
     Min       1Q   Median       3Q      Max
-0.50566 -0.13501 -0.03491  0.10430  0.83447

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.22099    0.77650   4.148 3.68e-05 ***
ln_Height           1.27217    1.51109   0.842     0.40
YRBS$Age            0.00177    0.04774   0.037     0.97
ln_Height:YRBS$Age  0.03331    0.09280   0.359     0.72
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

reedom
.2494
5
```

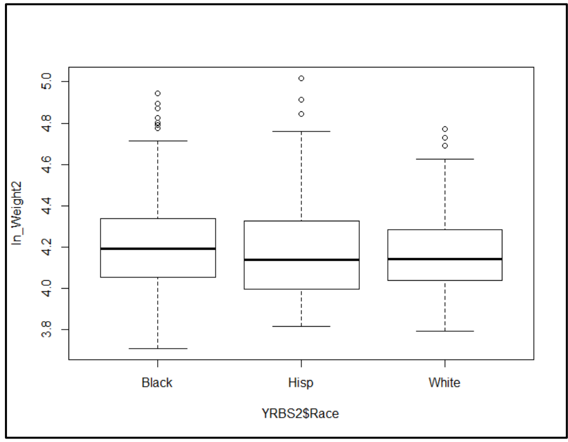


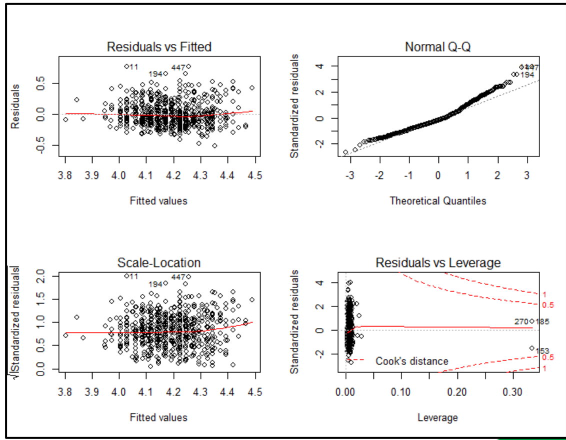
Step-by-step Examples 1

```
#Weight = Height + Age + Gender + Race
>   boxplot(ln_Weight~YRBS$Race)
>   boxplot(ln_Weight~YRBS$Sex)
>   table(YRBS$Race)
>   YRBS2 <- YRBS[ which(YRBS$Race=='Black'|
    YRBS$Race=='White' | YRBS$Race=='Hisp') ,]
>   table(YRBS2$Race)

>   ln_Weight2<-log(YRBS2$Weight_kg)
>   ln_Height2<-log(YRBS2$Height_m)
>   YRBS2$Race<-factor(YRBS2$Race)
>   boxplot(ln_Weight2~YRBS2$Race)

>   lm3<-lm(ln_Weight2~ln_Height2 +
    YRBS2$Sex + YRBS2$Race)
>   summary(lm3)
>   par(mfrow=c(2,2))
>   plot(lm3)
>   par(mfrow=c(1,1))
```

```
Call:
lm(formula = ln_Weight2 ~ ln_Height2 + YRBS2$Sex + YRBS2$Race)

Residuals:
     Min       1Q   Median       3Q      Max
-0.51826 -0.13125 -0.03368  0.09829  0.77320

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.32178    0.06740  49.286  < 2e-16 ***
ln_Height2       1.73981    0.12895  13.492  < 2e-16 ***
YRBS2$SexMale    0.06794    0.11355   0.598    0.550
YRBS2$RaceHisp  -0.08835    0.11330  -0.780    0.436
YRBS2$RaceWhite -0.07789    0.01873  -4.158 3.67e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1949 on 619 degrees of freedom
Multiple R-squared:  0.2372, Adjusted R-squared:  0.2322
F-statistic: 48.11 on 4 and 619 DF,  p-value: < 2.2e-16
```







	AI/AN	Asian	Black	Hisp	Multiple_Hisp	Multiple_NH	NH/PI	White
114	4	21	256	177	198	23	3	191

	AI/AN	Asian	Black	Hisp	Multiple_Hisp	Multiple_NH	NH/PI	White
0	0	0	256	177	0	0	0	191

Step-by-step Examples 1

#Comparison of Models

```
> anova(lm1, lm2)

> lm4 <- lm(ln_Weight2~ln_Height2)
> summary(lm4)

> anova(lm4, lm3)
```

```
Analysis of Variance Table

Model 1: ln_Weight ~ ln_Height
Model 2: ln_Weight ~ ln_Height * YRBS$Age
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    871 33.996
2    869 33.537  2   0.45886 5.9449 0.002727 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Call:
lm(formula = ln_Weight2 ~ ln_Height2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.50305 -0.13383 -0.04251  0.10190  0.79625

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.3244     0.0666   49.91   <2e-16 ***
ln_Height2    1.6765     0.1285   13.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1973 on 622 degrees of freedom
Multiple R-squared:  0.2149, Adjusted R-squared:  0.2137
F-statistic: 170.3 on 1 and 622 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Model 1: ln_Weight2 ~ ln_Height2
Model 2: ln_Weight2 ~ ln_Height2 + YRBS2$Sex + YRBS2$Race
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    622 24.211
2    619 23.525  3    0.6853 6.0105 0.0004864 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

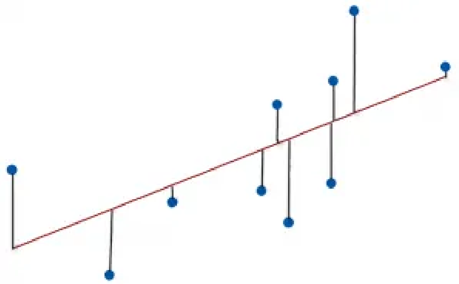
Assessment 1

1. What are the four assumptions of basic linear regression?

2. Can you still run linear regression if your Y variable is not normally distributed? Why or why not?

3. In the figure to the right, the red line represents the __________, the blue dots represent the __________, and the black lines represents the __________.

(choices: residuals, observed values, predicted values)

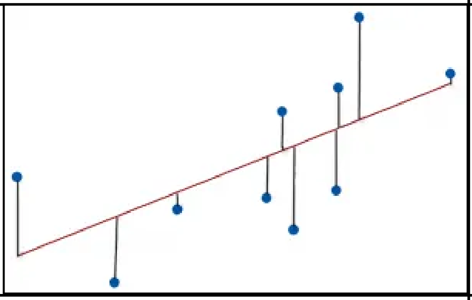


4. To the right is part of the summary table from R for the distance as a function of speed. Is speed significant? If so, why? How much variation does speed explain?

```
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) -17.5791     6.7584  -2.601   0.0123 *
#> speed         3.9324     0.4155   9.464 1.49e-12 ***
#> Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
#> F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

5. Suppose you want to determine if average year income (income) can be predicted by IQ, age, gender, and region. You also want to account for the interaction between gender and region. How would you set up your regression model?
   a) lm(IQ ~ income + age|gender + region)
   b) lm(IQ~ income + age*gender + region)
   c) lm(Age|gender ~ income + IQ + region)
   d) lm(Age*gender ~ income + IQ + region)
   e) lm(income ~ IQ + age:gender + region)
   f) lm(income ~ IQ + age*gender + region)

Assessment 1

1. What are the four assumptions of basic linear regression?	Linearity, Homoscedasticity, Independence, Normality
2. Can you still run linear regression if your Y variable is not normally distributed? Why or why not?	Yes, you can. You can use a generalized linear model, using the appropriate distribution. Examples of common generalized models are logistic (binary data) and Poisson (count) models.
3. In the figure to the right, the red line represents the __________, the blue dots represent the __________, and the black lines represents the __________. (choices: residuals, observed values, predicted values) 	In the figure to the right, the red line represents the **predicted values**, the blue dots represent the **observed values** and the black lines represents the **residuals**.
4. To the right is part of the summary table from R for the distance as a function of speed. Is speed significant? If so, why? How much variation does speed explain? **#> Coefficients:** **#> Estimate Std. Error t value Pr(>|t|)** **#> (Intercept) -17.5791 6.7584 -2.601 0.0123 *** **#> speed 3.9324 0.4155 9.464 1.49e-12 ***** **#> Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438** **#> F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12**	Yes, the p-value (**9.464 1.49e-12)** is significant. Based on the R-squared, speed explains about 65% of the variation in distance.
5. Suppose you want to determine if average year income (income) can be predicted by IQ, age, gender, and region. You also want to account for the interaction between gender and region. How would you set up your regression model? a) lm(IQ ~ income + age|gender + region) b) lm(IQ~ income + age*gender + region) c) lm(Age|gender ~ income + IQ + region) d) lm(Age*gender ~ income + IQ + region) e) lm(income ~ IQ + age:gender + region) f) lm(income ~ IQ + age*gender + region)	f) Lm(income ~ IQ + age*gender + region)

DaCCoTA
DAKOTA CANCER COLLABORATIVE
ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Step-by-step Examples 2

Youth Risk Behaviors Survey 2 (Logistic)

A. Can we predict whether a student got into a fight by weight?
B. Can we predict whether a student got into a fight by sex, age, height, and weight?

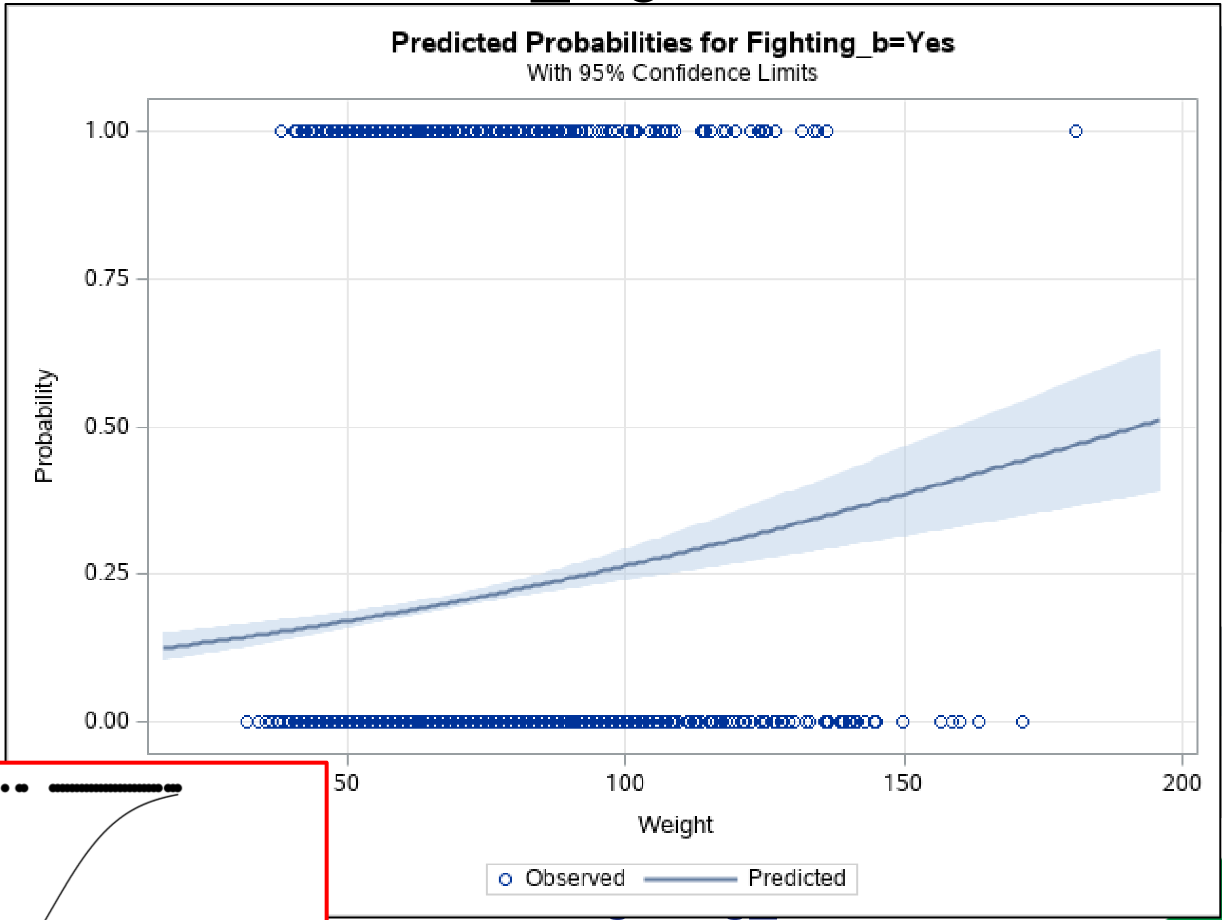
Warp Breaks (Poisson)

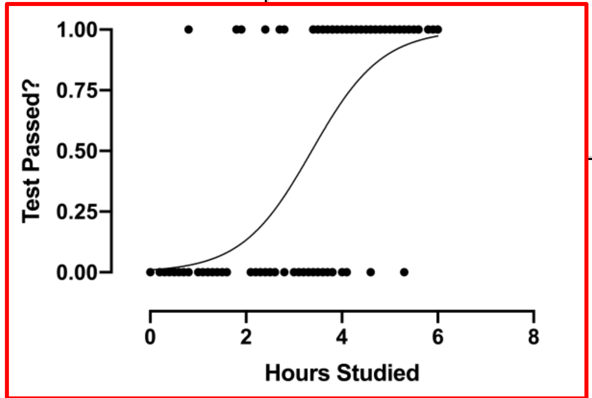
C. Is the number of breaks different across wool type and tension?

Step-by-step Examples 2

Can we predict whether a student got into a fight by weight?

```
PROC IMPORT datafile='/home/…/YRBS_Example_2.csv'
        dbms=csv out=youth replace; getnames=yes;
PROC PRINT data=youth(obs=10);
PROC FREQ data=youth;
        tables Fighting_b;
PROC LOGISTIC data=youth plots=effect;
        where age in (3,4,5,6,7);
        model Fighting_b(event='Yes')=Weight;
```





DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Step-by-step Examples 2

Can we predict whether a student got into a fight by sex, age, height, and weight?

```
PROC FREQ data=youth;
        where age in (3,4,5,6,7);
        tables Fighting_b*Sex;
        tables Fighting_b*Age;
PROC GLIMMIX data=youth;
        where age in (3,4,5,6,7);
        class Sex Age;
        model Fighting_b(event='Yes')=Sex Age Height Weight /dist=binary oddsratio;
```

Odds Ratio Estimates

Sex	Age	Height	Weight	_Sex	_Age	_Height	_Weight	Estimate	DF	95% Confidence Limits	
Female		1.7045	67.522	Male		1.7045	67.522	0.416	4703	0.341	0.508
	3	1.7045	67.522		7	1.7045	67.522	1.722	4703	1.263	2.348
	4	1.7045	67.522		7	1.7045	67.522	1.791	4703	1.381	2.324
	5	1.7045	67.522		7	1.7045	67.522	1.503	4703	1.158	1.951
	6	1.7045	67.522		7	1.7045	67.522	1.121	4703	0.860	1.462
		2.7045	67.522			1.7045	67.522	1.039	4703	0.354	3.046
		1.7045	68.522			1.7045	67.522	1.006	4703	1.001	1.010

Effects of continuous variables are assessed as one unit offsets from the mean. The AT suboption modifies the reference value and the UNIT suboption modifies the offsets.

The GLIMMIX procedure is modeling the probability that Fighting_b='Yes'.

Fit Statistics

-2 Log Likelihood	4552.73
AIC (smaller is better)	4568.73
AICC (smaller is better)	4568.76
BIC (smaller is better)	4620.39
CAIC (smaller is better)	4628.39
HQIC (smaller is better)	4586.89
Pearson Chi-Square	4708.24
Pearson Chi-Square / DF	1.00

Type III Tests of Fixed Effects

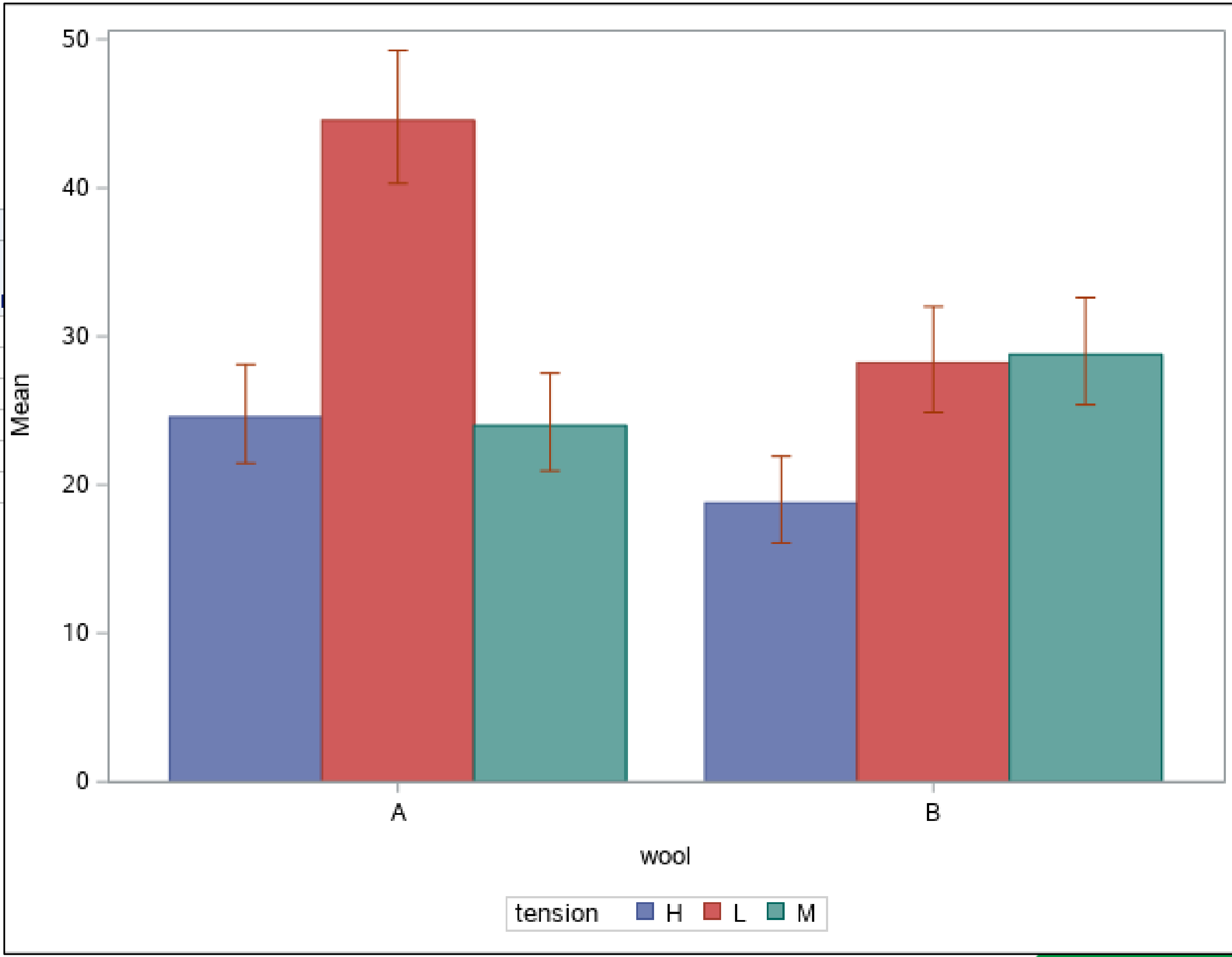
Effect	Num DF	Den DF	F Value	Pr > F
Sex	1	4703	74.17	<.0001
Age	4	4703	8.10	<.0001
Height	1	4703	0.00	0.9443
Weight	1	4703	5.57	0.0183

Step-by-step Examples 2

Is the number of breaks different across wool type and tension?

```
PROC IMPORT datafile='/home/…/warpbreaks.csv'
        dbms=csv out=warp replace;
        getnames=yes;
PROC PRINT data=warp(obs=10);
PROC GLIMMIX data=warp;
        class wool tension;
        model breaks=wool|tension/dist=poisson;
        lsmeans wool*tension /ilink cl;
        ods output LSMeans=warp_lsm;
PROC SGPLOT data=warp_lsm;
        vbarparm category=wool response=Mu/
        group=tension groupdisplay=cluster
        limitupper=UpperMu limitlower=LowerMu;
```

wool	tension
A	H
A	L
A	M
B	H
B	L
B	M



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Assessment 2

1. What type of regression should be used for binary response data (0/1, Yes/No, etc.)?
   What type of regression should be used for count response data?

2. If you want to display the number of observations across groups in SAS, what procedure should you use?
   a) PROC UNIVARIATE
   b) PROC MEANS
   c) PROC FREQ
   d) PROC LOGISITIC

3. Generally speaking, what does the following term mean in SAS?
   Variable1*Variable2

4. To the right are odds ratios from a logistic regression. Do any of the races have significantly higher or lower odds than the reference? How and why? What about for sex?

Odds Ratio Estimates

RACE	SEX	_RACE	_SEX	Estimate	DF	95% Confidence Limits	
2		1		0.866	263E3	0.840	0.893
3		1		0.838	263E3	0.781	0.899
4		1		0.753	263E3	0.700	0.810
	2		1	1.202	263E3	1.183	1.222

5. To the right are the Type III tests of fixed effects from a Poisson regression. Which variables are significant? Why?

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Year	1	30	151.70	<.0001
Region	1	30	0.10	0.7589
Year*Region	1	30	0.09	0.7616

Assessment 2

1. What type of regression should be used for binary response data (0/1, Yes/No, etc.)? What type of regression should be used for count response data?	Logistic Poisson
2. If you want to display the number of observations across groups in SAS, what procedure should you use? a) PROC UNIVARIATE c) PROC FREQ b) PROC MEANS d) PROC LOGISITIC	c) PROC FREQ
3. Generally speaking, what does the following term mean in SAS? Variable1*Variable2	The interaction between Variable1 and Variable2
4. To the right are odds ratios from a logistic regression. Do any of the races have significantly higher or lower odds than the reference? How and why? What about for sex?	Races 2, 3, and 4 all have significantly lower odds than Race 1. The upper confidence limits are all below 1.0. Sex 2 has significantly higher odds than Sex 1. The lower confidence limit is above 1.0.
5. To the right are the Type III tests of fixed effects from a Poisson regression. Which variables are significant? Why?	Year is significant because the p value is <0.05 Neither Region nor the interaction between Year and Region (Year*Region) is significant because the p values are not <0.05.

Odds Ratio Estimates

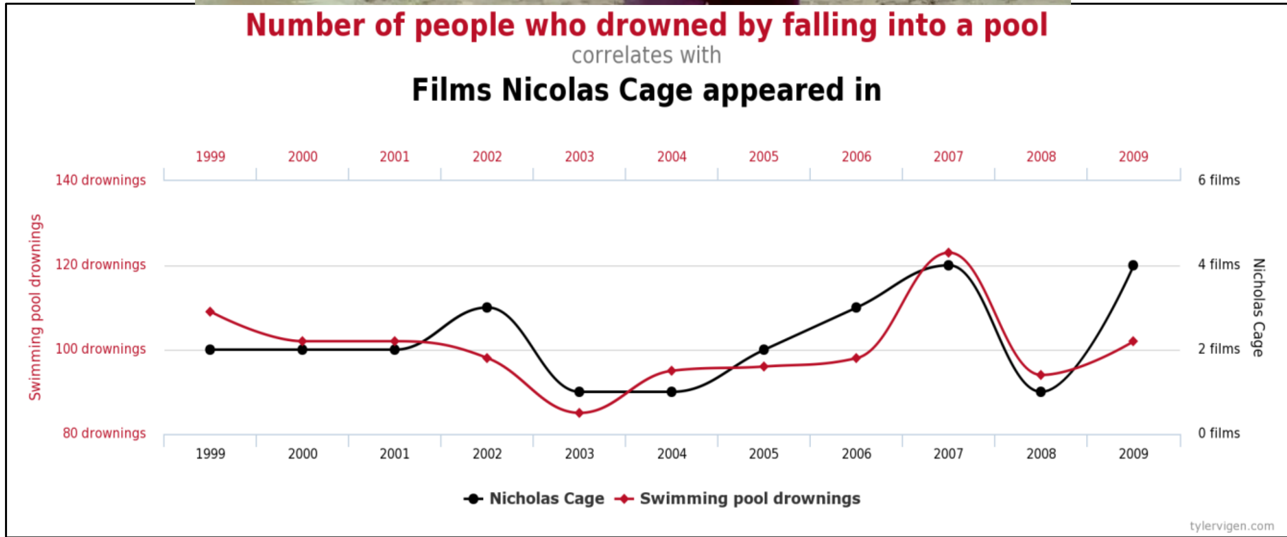
RACE	SEX	_RACE	_SEX	Estimate	DF	95% Confidence Limits	
2		1		0.866	263E3	0.840	0.893
3		1		0.838	263E3	0.781	0.899
4		1		0.753	263E3	0.700	0.810
	2		1	1.202	263E3	1.183	1.222

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Year	1	30	151.70	<.0001
Region	1	30	0.10	0.7589
Year*Region	1	30	0.09	0.7616

Caveats and Concerns

- The art of model building
- Variable inclusion
- Interpretation of complex models
- Distributions
- Model fit:
  - Residuals, Pearson Chi-Square/DF, AICc, etc.
  - Underfitting
  - Overfitting
- Computing considerations
  - Can do same procedures across software systems and functions/procedures
  - Computers are fast but dumb -> you need to be the one with understanding
- Correlation and causation



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Real World Examples

Zuur, A. F., Ieno, E. N., & Freckleton, R. (2016). A protocol for conducting and presenting results of regression-type analyses. *Methods in Ecology and Evolution, 7*(6), 636-645. doi:10.1111/2041-210x.12577

Effective statistical analysis and communication **643**

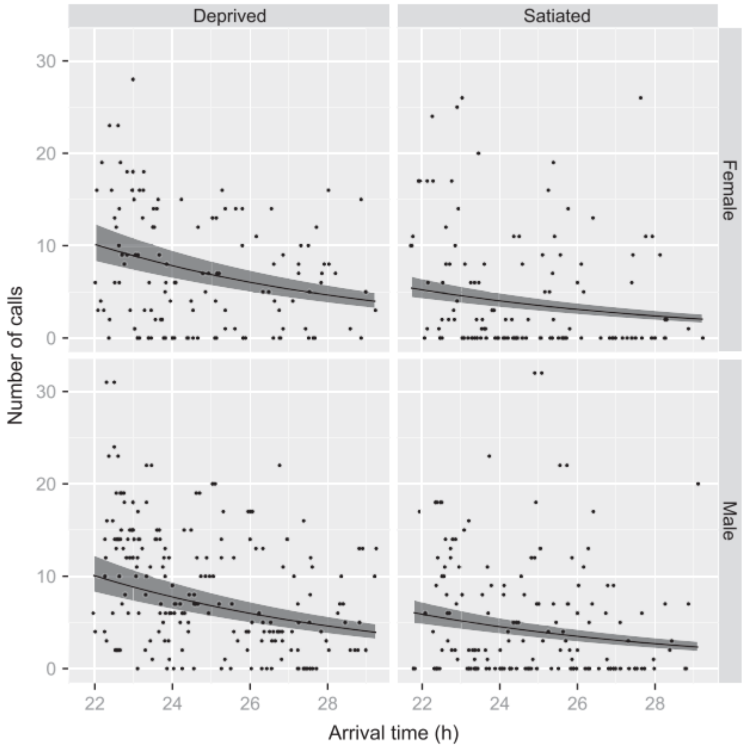


Fig. 6. Fit of the Poisson GLMM in eqn (1) for the owl data.

Protocol for conducting and presenting results of regression-type analyses

1. **State appropriate questions**
2. **Visualize the experimental design**
3. **Conduct data exploration**
4. **Identify the dependency structure in the data**
5. **Present the statistical model**
6. **Fit the model**
7. **Validate the model**
8. **Interpret and present the numerical output of the model**
9. **Create a visual representation of the model**
10. **Simulate from the model**

Table 1. Estimated regression parameters, standard errors, *z*-values and *P*-values for the Poisson GLMM presented in eqn (1). The estimated value for σ_{Nest} is 0·484.

	Estimate	Std. error	*z* value	*P*-value
Intercept	5·169	0·292	17·665	<0·05
FoodTreatmentSatiated	−0·654	0·468	−1·395	0·162
ArrivalTime	−0·129	0·011	−11·472	<0·05
SexParentMale	−0·009	0·045	−0·208	0·834
FoodTreatmentSatiated : SexParentMale	0·129	0·070	1·842	0·065
FoodTreatmentSatiated : ArrivalTime	−0·000	0·019	−0·026	0·979

Real World Examples

Brown, D. R., & Blanton, C. J. (2007). Physical Activity, Sport Participation, and Suicidal Behavior. *Medicine and Science in Sports and Exercise, 39*(12), 2248-2257.

TABLE 1. Percentage (± 95% CI; n=) of suicidal behavior[1] by sociodemographic characteristics and selected health risk behaviors among participants in the National College Health Risk Behavior Survey, 1995.

Characteristic	Total % Suicidal Behavior	Men % Suicidal Behavior	Women % Suicidal Behavior
Gender	11.4 (±1.0;4728)	11.2 (±1.3;1797)	11.7 (±1.4;2906)
Age group (years)			
18–24	13.2 (±1.4;2868)	13.0 (±2.0;1184)	13.5 (±1.9;1684)
≥25	8.4 (±1.3;1766)	7.2 (±2.3; 588)	9.2 (±1.7;1175)
Race/ethnic group			
White non-Hispanic	10.6 (±1.0;2919)	9.7 (±1.4;1108)	11.4 (±1.5;1810)
Black non-Hispanic	11.0 (±2.6; 631)	12.5 (±4.8; 204)	10.1 (±3.0; 427)
Hispanic	12.4 (±3.0; 693)	14.0 (±5.0; 279)	11.3 (±3.5; 413)
Asian/Pacific Islander	16.8 (±5.3; 258)	17.0 (±7.1; 124)	16.5 (±7.9; 133)
Other	17.7 (±6.2; 159)	17.7 (±9.2; 323)	17.8 (±9.4; 88)
Cigarette smoking status			
Never	9.6 (±1.1;3192)	9.5 (±1.5;1223)	9.7 (±1.4;1950)
Former	12.7 (±3.4; 469)	13.5 (±6.0; 172)	12.2 (±3.9; 295)
Current	17.7 (±2.8; 878)	17.2 (±4.2; 323)	18.0 (±3.9; 549)
Previous 30-day heavy episodic alcohol use			
Yes	14.2 (±1.7;1476)	13.0 (±2.3; 744)	15.7 (±2.8; 723)
No	10.0 (±1.2;3132)	9.7 (±2.1;1007)	10.3 (±1.5;2111)
Lifetime ever drug use			
None	8.3 (±1.1;2511)	7.4 (±1.7; 911)	9.1 (±1.6;1584)
1–9 times	12.1 (±2.1; 886)	12.5 (±3.9; 325)	11.9 (±2.4; 557)
10–39 times	16.1 (±3.6; 472)	16.6 (±6.3; 174)	15.8 (±4.4; 297)
≥ 40 times	16.7 (±2.8; 845)	16.3 (±3.9; 381)	16.6 (±3.8; 461)
BMI[2] and Perceived overweight (Yes or No)			
BMI ≥ 25/No	7.4 (±3.2; 329)	7.4 (±3.5; 265)	8.1 (±7.1; 62)
BMI ≥ 25/Yes	13.0 (±1.8;1374)	11.1 (±2.4; 511)	14.4 (±2.7; 856)
BMI < 25/No	10.7 (±1.2;2350)	11.5 (±2.3; 940)	10.0 (±1.6;1400)
BMI < 25/Yes	13.4 (±3.0; 611)	18.7 (±10.8; 69)	12.4 (±2.9; 537)

[1] Suicidal behavior is defined as thoughts about, planning for, or attempting suicide.
[2] Body Mass Index (BMI) cutpoints for overweight are based on National Institutes of Health, and National Heart, Lung, and Blood Institute guidelines.

TABLE 2. Prevalence of suicidal behavior[1] by category of physical activity among college men, National College Health Risk Behavior Survey, 1995.

Category of physical activity	Sample Size	Unadjusted Prevalence of Suicidal Behavior	Unadjusted Prevalence Ratio	Adjusted Odds Ratio (95% CI)[2]
No reported physical activity	335	13.57	1.00	1.00 (referent)
Low active	496	8.23	0.57	0.54 (0.33,0.88)
Moderately active	196	10.85	0.77	0.70 (0.36,1.39)
Vigorously active (3–5 days/week)	573	12.74	0.93	1.12 (0.68,1.83)
Frequently vigorously active (6–7 days/week)	197	10.13	0.72	0.87 (0.47,1.64)

[1] Suicidal behavior is defined as thoughts about, planning for, or attempting suicide.
[2] Odds ratios were adjusted for age, race/ethnicity, sports participation, BMI/weight perception, cigarette smoking status, any episodic heavy alcohol use during the past 30 days, and number of times lifetime ever drug use.

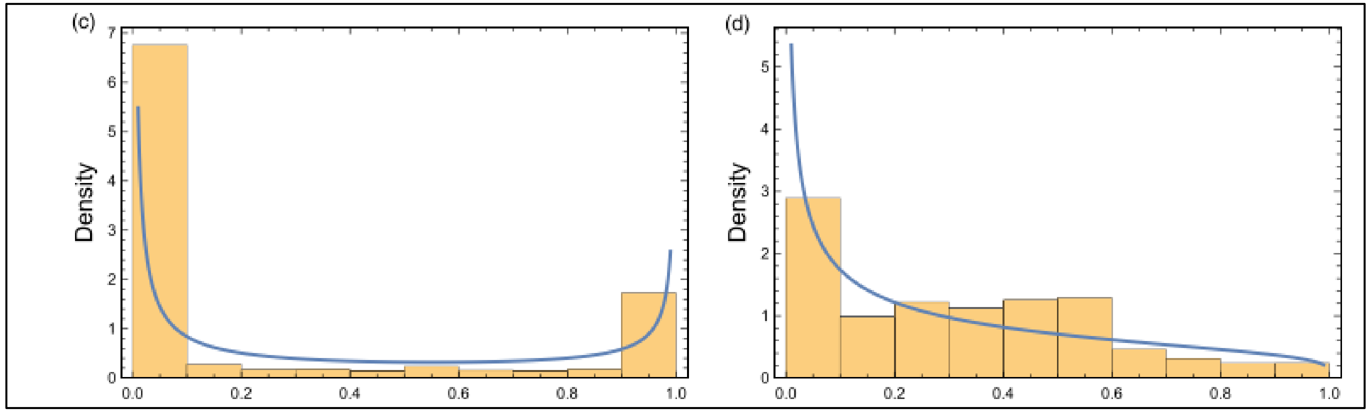
TABLE 4. Prevalence of suicidal behavior[1] by intramural or extramural sports participation among college men, National College Health Risk Behavior Survey, 1995.

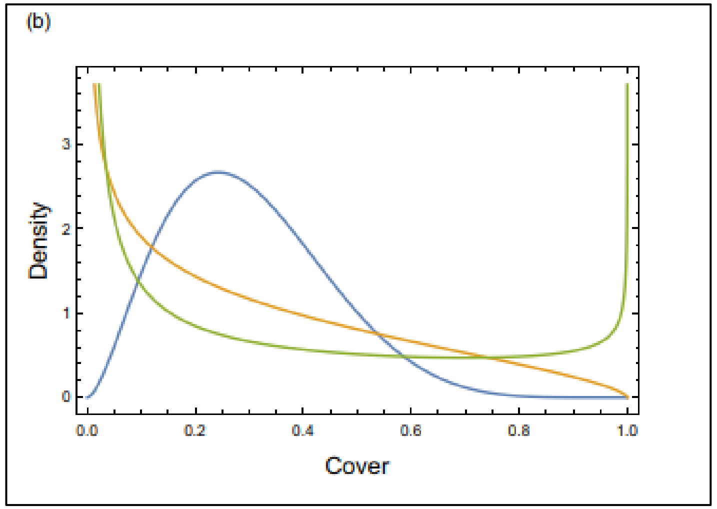
Category of sports participation	Sample Size	Unadjusted Prevalence of Suicidal Behavior	Unadjusted Prevalence Ratio	Adjusted Odds Ratio (95% CI)[2]
Sports participation	428	7.27	1.00	1.00 (referent)
No sports participation	1363	12.54	1.83	2.46 (1.52,3.99)

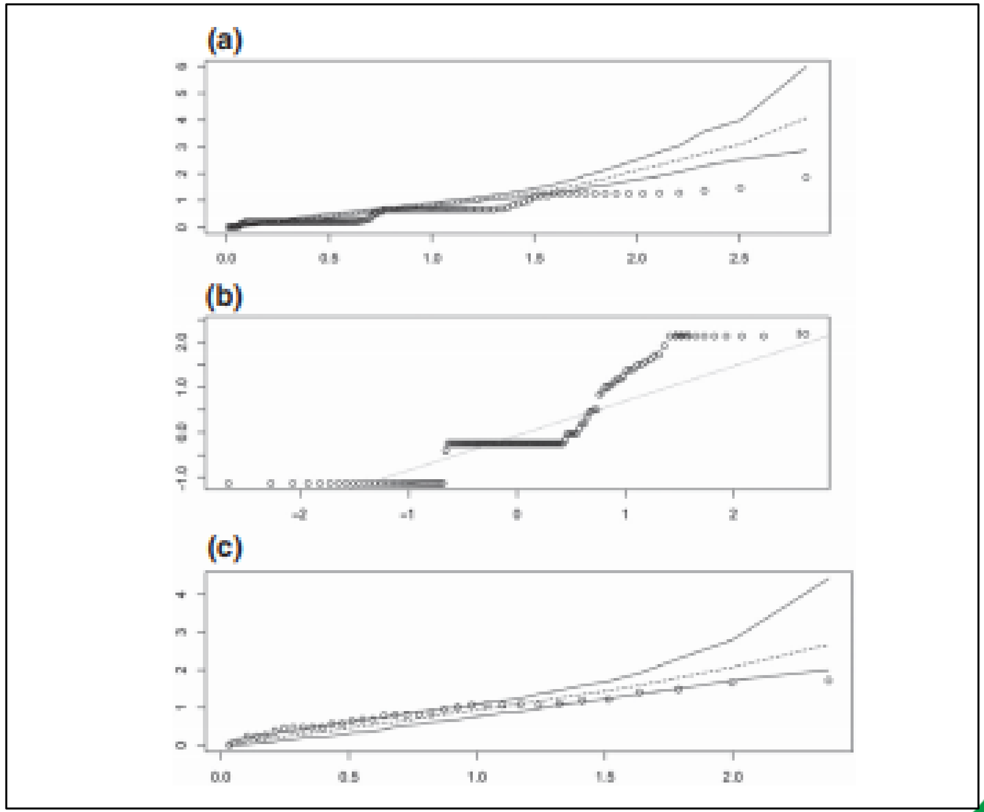
[1] Suicidal behavior is defined as thoughts about, planning for, or attempting suicide.
[2] Odds ratios were adjusted for age, race/ethnicity, category of physical activity, BMI/weight perception, cigarette smoking status, any episodic heavy alcohol use during the past 30 days, and number of times lifetime ever drug use.

Real World Examples

Damgaard, C. F., Irvine, K. M., & Stott, I. (2019). Using the beta distribution to analyse plant cover data. *Journal of Ecology, 107*(6), 2747-2759. doi:10.1111/1365-2745.13200

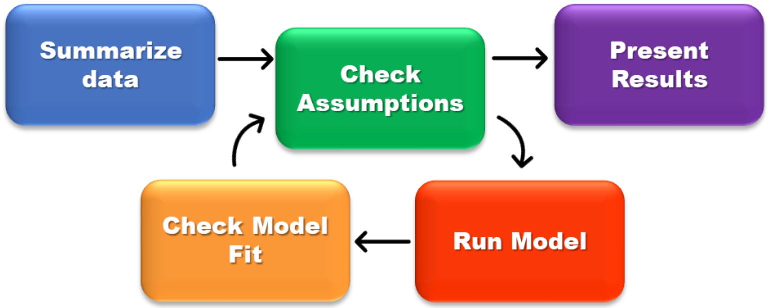


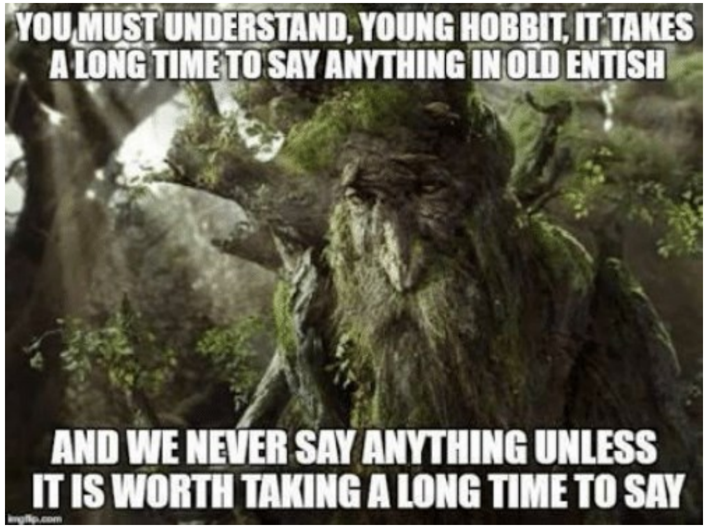




Summary and Conclusion

- Linear regression covers a vast swath of statistical models
- The type of regression depends on your response and predictor variables
- Need to consider assumptions and model fit
- Typically an iterative process
- Take your time





- Tune in next time for a plunge into advanced topics of Linear Regression in Module III: Deep Dive

Acknowledgements

- The DaCCoTA is supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729.
- For the labs that use the Biostatistics, Epidemiology, and Research Design Core in any way, including this Module, please acknowledge us for publications. ***"Research reported in this publication was supported by DaCCoTA (the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729)".***

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY