SUMMER 2025

TEXAS
The University of Texas at Austin

INTRODUCTION TO STATISTICAL MODELING

Center for Biomedical Research Support

LAYLA GUYOT
Assistant Professor of Instruction, Ph.D.
Department of Statistics and Data Sciences
The University of Texas at Austin

Access materials

laylaguyot

Overview | Repositories 8 | Projects

https://github.com/laylaguyot/CBRS_Intro_Statistical_Modeling

Layla Guyot
laylaguyot

Statistics and Data Science enthusiast: teacher and researcher in education, focusing on bridging the gap between academia and industry.

README

Introduction to Statistical Modeling

Materials

Download the slides and corresponding .Rmd files each day

Access RStudio

Go to https://gsafcomp02.ccbb.utexas.edu and select RStudio

Tentative Schedule

Day 1 Exploring Data

- Study design and variables
- Descriptive statistics and visualizations
- Introduction to hypothesis testing

Day 2 Making Inferences

- Probability, random variables, and common probability distributions
- Sampling distributions and Central Limit Theorem
- Confidence intervals, t-tests, ANOVA, and Chi-square tests

Day 3 Linear Regression

- Simple Linear Regression
- Multiple Regression with different types of predictors
- Model assumptions, evaluation, and comparisons

Day 4 Logistic Regression

- Odds
- Logistic Regression
- Model evaluation with ROC curves or confusion matrix

Day 5 Model Building

- Underfitting, overfitting, and cross-validation
- Common pitfalls: multicollinearity, transformations
- Missing data

Model Building

Find a model that is simple yet useful and provides (1 or more):

- summary of trend in response
- good predictions of the response
- good estimates of the coefficients

Model Building

Four possible outcomes for the regression model:

- correctly specified
- underfitted
- with some extraneous predictors
- overfitted

Model Building

Recommended steps:

- ➢ define goal (with Research Question)
- ➢ identify all possible candidate predictors
- ➢ use variable selection procedures (stepwise, best subsets)
- ➢ refine model (interactions, higher order, transformations, …)

Model Building

Variable selection procedure:

- stepwise regression

Procedure Forward:

1. Define an alpha to enter, α_E, and an alpha to remove, α_R, a predictor (typically both are 0.15).
2. Compare t-test p-values of SLR between each predictor and the response. Add predictor with smallest p-value, less than α_E, to the model.
3. Compare t-test p-values of MLR between each pair of predictors (but all including predictor from 2.) and the response. Add predictor with smallest p-value, less than α_E, to the model. Check that the predictor from 2. still has a p-value greater than α_R, remove the predictor.
4. Continue the process until no additional predictor has a p-value less than α_E.

Model Building

Variable selection procedure:

➢ stepwise regression

- The final model is not guaranteed to be optimal
- Stepwise regression does not account for researchers' knowledge about the predictors. It may be necessary to force the procedure to include important predictors.
- We should not over-interpret the order in which predictors are entered into the model.
- We cannot conclude that all the important predictor variables for predicting Y have been identified.

Model Building

Variable selection procedure:

- stepwise regression

Example: Are a person's brain size and body size predictive of his or her intelligence?

On a sample of 38 college students, the following variables were collected:

Y : Performance IQ scores (PIQ) from the revised Wechsler Adult Intelligence Scale.

X_1 : Brain size based on the count obtained from MRI scans (given as count/10,000).

X_2 : Height (in inches).

X_3 : Weight (in pounds).

Model Building

Variable selection procedure:

- stepwise regression

<u>Example</u>: Are a person's brain size and body size predictive of his or her intelligence?

```
> model1A <- lm(PIQ~Brain,iqsize)
> summary(model1A)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.6519    43.7118   0.106   0.9158
Brain         1.1766     0.4806   2.448   0.0194 *
> model1B <- lm(PIQ~Height,iqsize)
> summary(model1B)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 147.4067    64.3498   2.291   0.0279 *
Height       -0.5271     0.9389  -0.561   0.5780
> model1C <- lm(PIQ~Weight,iqsize)
> summary(model1C)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.110e+02  2.451e+01   4.527 6.31e-05 ***
Weight      2.418e-03  1.604e-01   0.015    0.988
---
```

The variable Brain has the smallest *p*-value, also smaller than the alpha enter of 0.15.

Model Building

Variable selection procedure:

- stepwise regression

<u>Example</u>: Are a person's brain size and body size predictive of his or her intelligence?

```
> model2A <- lm(PIQ~Brain+Height,iqsize)
> summary(model2A)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 111.2757    55.8673   1.992 0.054243 .
Brain         2.0606     0.5466   3.770 0.000604 ***
Height       -2.7299     0.9932  -2.749 0.009399 **
```

```
> model2B <- lm(PIQ~Brain+Weight,iqsize)
> summary(model2B)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.7520    43.0250   0.110  0.91269
Brain         1.5925     0.5512   2.889  0.00659 **
Weight       -0.2503     0.1704  -1.469  0.15071
```

The variable Height has the smallest *p*-value, also smaller than the alpha enter of 0.15 and the *p*-value of Brain is still less than 0.15.

Model Building

Variable selection procedure:

- stepwise regression

<u>Example</u>: Are a person's brain size and body size predictive of his or her intelligence?

```
> model3 <- lm(PIQ~Brain+Height+Weight,iqsize)
> summary(model3)

Call:
lm(formula = PIQ ~ Brain + Height + Weight, data = iqsize)

Residuals:
   Min     1Q Median     3Q    Max
-32.74 -12.09  -3.84  14.17  51.69

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.114e+02  6.297e+01   1.768 0.085979 .
Brain        2.060e+00  5.634e-01   3.657 0.000856 ***
Height      -2.732e+00  1.229e+00  -2.222 0.033034 *
Weight       5.599e-04  1.971e-01   0.003 0.997750
```

The final model only contains two predictors, Brain and Height (model 2A).

Model Building

Variable selection procedure:

- stepwise regression algorithm

```
> FitStart <- lm(PIQ ~ 1, iqsize)
> FitAll <- lm(PIQ~Brain+Height+Weight,iqsize)
> step(FitStart,direction="forward", scope = formula(FitAll))
Start:  AIC=237.94
PIQ ~ 1

         Df Sum of Sq   RSS    AIC
+ Brain   1   2697.09 16198 234.09
<none>                18895 237.94
+ Height  1    163.97 18731 239.61
+ Weight  1      0.12 18894 239.94

Step:  AIC=234.09
PIQ ~ Brain

         Df Sum of Sq   RSS    AIC
+ Height  1   2875.65 13322 228.66
+ Weight  1    940.94 15256 233.82
<none>                16198 234.09

Step:  AIC=228.66
PIQ ~ Brain + Height

         Df Sum of Sq   RSS    AIC
<none>                13322 228.66
+ Weight  1 0.0031633 13322 230.66

Call:
lm(formula = PIQ ~ Brain + Height, data = iqsize)

Coefficients:
(Intercept)        Brain       Height
    111.276        2.061       -2.730
```

Model Building

Variable selection procedure:

- stepwise regression
- best subsets regression

Procedure:

1. Identify all possible models.
2. Define criteria to consider.
3. Further evaluate and refine some models (diagnostics, interaction, …)

Model Building

Criteria for model selection:

- R^2 and R^2_{adj}
- Mallow C_p
- Bayesian Information Criterion BIC)

Model Building

Criteria for model selection:

- R^2 and R^2_{adj}

The best regression model has the smallest SS_{error} and/or MS_{error}, but the more predictors are added, the higher R^2 is so we usually compare adjusted R^2 instead.

maximize $R^2 = 1 - \frac{SS_{error}}{SS_{total}}$ and/or $R^2_{adj} = 1 - \frac{MS_{error}}{\frac{SS_{total}}{n-1}}$

Model Building

Criteria for model selection:

- R^2 and R^2_{adj}
- Mallow C_p — estimates the size of the bias that is introduced into the predicted responses by having an underspecified model (if C_p is near p, the bias is small)

minimize $$C_p = \frac{SS_{error}}{MS_{error}} - (n - 2p)$$

Model Building

Criteria for model selection:

- R^2 and R^2_{adj}
- Mallow C_p
- Bayesian Information Criterion (BIC)

combines information about the SS_{error}, number of parameters in the model, and the sample size.

minimize $$BIC = n \ln SS_{error} - n \ln n + p \ln n$$

Model Building

Variable selection procedure:

<u>Example</u>: Are a person's brain size and body size predictive of his or her intelligence?

On a sample of 38 college students, the following variables were collected:

Y : Performance IQ scores (PIQ) from the revised Wechsler Adult Intelligence Scale.

X_1 : Brain size based on the count obtained from MRI scans (given as count/10,000).

X_2 : Height (in inches).

X_3 : Weight (in pounds).

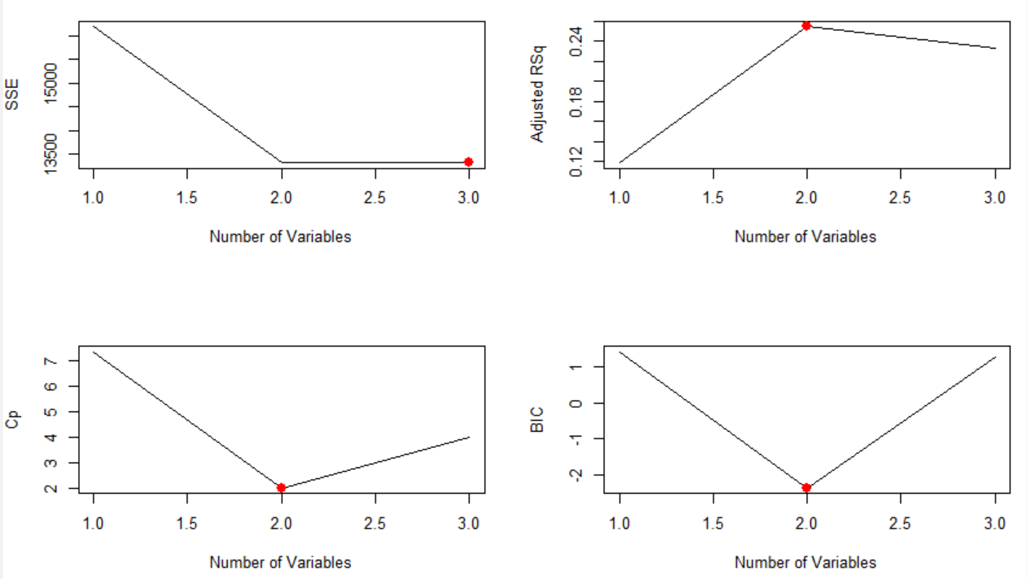
How many models are possible?

Model Building

Variable selection procedure:

<u>Example</u>: Are a person's brain size and body size predictive of his or her intelligence?

```
> model <- regsubsets(PIQ~Brain+Height+Weight,iqsize, nvmax = 3)
> summary(model)
Subset selection object
Call: regsubsets.formula(PIQ ~ Brain + Height + Weight, iqsize, nvmax = 3)
3 Variables  (and intercept)
       Forced in Forced out
Brain      FALSE      FALSE
Height     FALSE      FALSE
Weight     FALSE      FALSE
1 subsets of each size up to 3
Selection Algorithm: exhaustive
         Brain Height Weight
1  ( 1 ) "*"   " "    " "
2  ( 1 ) "*"   "*"    " "
3  ( 1 ) "*"   "*"    "*"
```



Best models with each number of predictors

Model Building

Model validation:

- collect new data
- compare to theoretical expectations, earlier results
- use holdout sample: cross-validation

Split data into training and test datasets

K-fold cross-validation

Model Building

Strategy for model building in 7 steps:

1. Decide on the goal: predictive, inferential, data summary
2. Decide which predictors and response
3. Explore data: univariate and bivariate analysis
4. Divide the data into a training and test set
5. Identify candidate models: stepwise or best subsets regression
6. Select and evaluate a few models, using some criteria
7. Select the final model: there is not necessarily only one good model for a given dataset

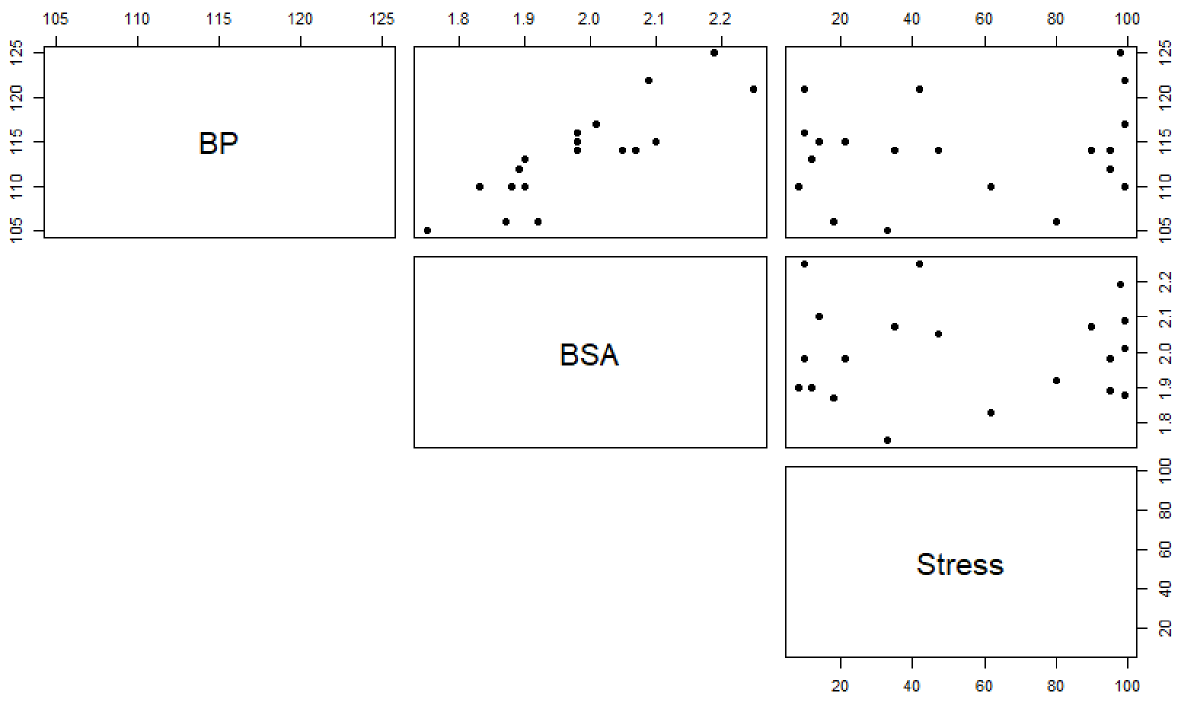
Multicollinearity

- Multicollinearity exists when two or more of the predictors in a regression model are moderately or highly correlated.

- It is a problem because individual coefficients and t-tests can be unreliable.

Multicollinearity

If the predictors are nearly uncorrelated:




```
> cor(data)
              BP        BSA     Stress
BP     1.0000000 0.86587887 0.16390139
BSA    0.8658789 1.00000000 0.01844634
Stress 0.1639014 0.01844634 1.00000000
```

y versus x1 and x2

```
> reg <- lm(BP~BSA+Stress, data)
> summary(reg)

Call:
lm(formula = BP ~ BSA + Stress, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.8992 -1.6483 -0.1643  1.7790  3.8524 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 44.24452    9.26104   4.777 0.000175 ***
BSA         34.33423    4.61110   7.446 9.56e-07 ***
Stress       0.02166    0.01697   1.277 0.218924    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.743 on 17 degrees of freedom
Multiple R-squared:  0.7716,	Adjusted R-squared:  0.7448 
F-statistic: 28.72 on 2 and 17 DF,  p-value: 3.534e-06
```

y versus x2 and x1

```
> reg21 <- lm(BP~Stress+BSA, data)
> summary(reg21)

Call:
lm(formula = BP ~ Stress + BSA, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.8992 -1.6483 -0.1643  1.7790  3.8524 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 44.24452    9.26104   4.777 0.000175 ***
Stress       0.02166    0.01697   1.277 0.218924    
BSA         34.33423    4.61110   7.446 9.56e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.743 on 17 degrees of freedom
Multiple R-squared:  0.7716,	Adjusted R-squared:  0.7448 
F-statistic: 28.72 on 2 and 17 DF,  p-value: 3.534e-06
```

y versus x1

```
> reg1 <- lm(BP~BSA, data)
> summary(reg1)

Call:
lm(formula = BP ~ BSA, data = data)

Residuals:
   Min     1Q Median     3Q    Max 
-5.314 -1.963 -0.197  1.934  4.831 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   45.183      9.392   4.811  0.00014 ***
BSA           34.443      4.690   7.343 8.11e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.79 on 18 degrees of freedom
Multiple R-squared:  0.7497,	Adjusted R-squared:  0.7358 
F-statistic: 53.93 on 1 and 18 DF,  p-value: 8.114e-07
```

y versus x2

```
> reg2 <- lm(BP~Stress, data)
> summary(reg2)

Call:
lm(formula = BP ~ Stress, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.6394 -3.3014  0.0722  2.2181  9.9287 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 112.71997    2.19345  51.389   <2e-16 ***
Stress        0.02399    0.03404   0.705     0.49    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.502 on 18 degrees of freedom
Multiple R-squared:  0.02686,	Adjusted R-squared:  -0.0272 
F-statistic: 0.4969 on 1 and 18 DF,  p-value: 0.4899
```

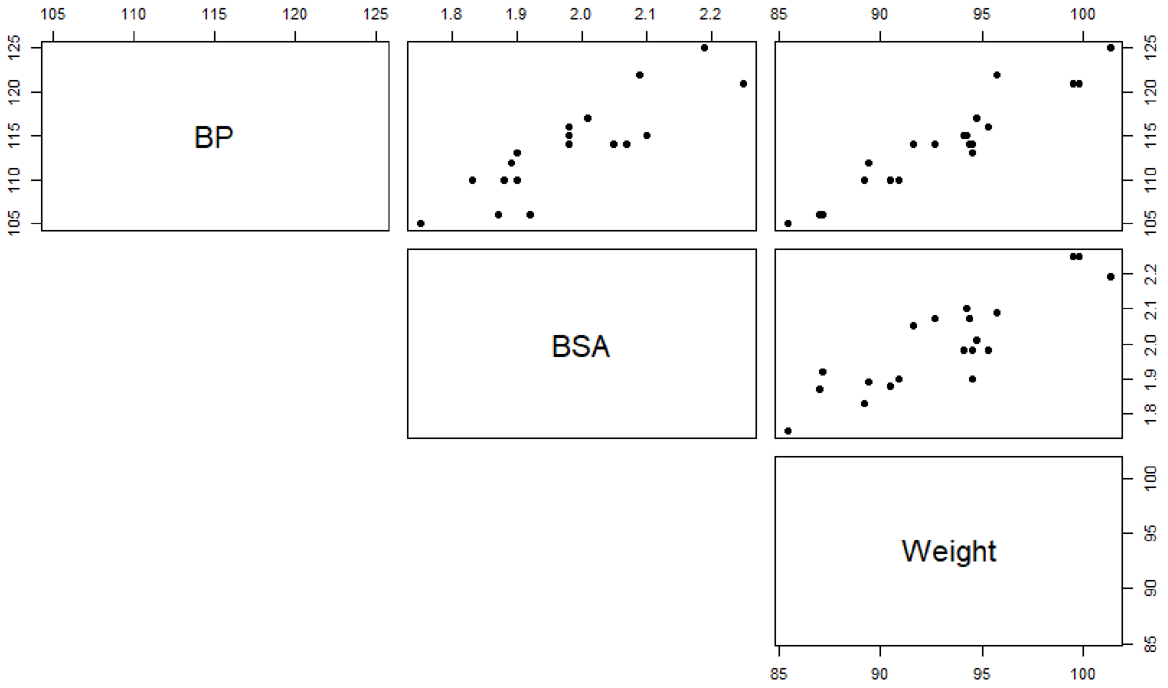
Multicollinearity

If the predictors are nearly uncorrelated:

- The coefficients b_1 and b_2 are very similar between SLR and MLR
- The standard errors of the coefficients b_1 and b_2 are very similar between SLR and MLR
- The sum of squares are very similar between SLR and MLR

Multicollinearity

If the predictors are highly correlated:




```
> cor(data)
              BP       BSA    Weight
BP     1.0000000 0.8658789 0.9500677
BSA    0.8658789 1.0000000 0.8753048
Weight 0.9500677 0.8753048 1.0000000
```

y versus x1 and x2

```
> reg <- lm(BP~BSA+Weight, data)
> summary(reg)

Call:
lm(formula = BP ~ BSA + Weight, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8932 -1.1961 -0.4061  1.0764  4.7524 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   5.6534     9.3925   0.602    0.555    
BSA           5.8313     6.0627   0.962    0.350    
Weight        1.0387     0.1927   5.392 4.87e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.744 on 17 degrees of freedom
Multiple R-squared:  0.9077,	Adjusted R-squared:  0.8968 
F-statistic: 83.54 on 2 and 17 DF,  p-value: 1.607e-09
```

y versus x2 and x1

```
> reg21 <- lm(BP~Weight+BSA, data)
> summary(reg21)

Call:
lm(formula = BP ~ Weight + BSA, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8932 -1.1961 -0.4061  1.0764  4.7524 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   5.6534     9.3925   0.602    0.555    
Weight        1.0387     0.1927   5.392 4.87e-05 ***
BSA           5.8313     6.0627   0.962    0.350    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.744 on 17 degrees of freedom
Multiple R-squared:  0.9077,	Adjusted R-squared:  0.8968 
F-statistic: 83.54 on 2 and 17 DF,  p-value: 1.607e-09
```

y versus x1

```
> reg1 <- lm(BP~BSA, data)
> summary(reg1)

Call:
lm(formula = BP ~ BSA, data = data)

Residuals:
   Min     1Q Median     3Q    Max 
-5.314 -1.963 -0.197  1.934  4.831 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   45.183      9.392   4.811  0.00014 ***
BSA           34.443      4.690   7.343 8.11e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.79 on 18 degrees of freedom
Multiple R-squared:  0.7497,	Adjusted R-squared:  0.7358 
F-statistic: 53.93 on 1 and 18 DF,  p-value: 8.114e-07
```

y versus x2

```
> reg2 <- lm(BP~Weight, data)
> summary(reg2)

Call:
lm(formula = BP ~ Weight, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.6933 -0.9318 -0.4935  0.7703  4.8656 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.20531    8.66333   0.255    0.802    
Weight       1.20093    0.09297  12.917 1.53e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.74 on 18 degrees of freedom
Multiple R-squared:  0.9026,	Adjusted R-squared:  0.8972 
F-statistic: 166.9 on 1 and 18 DF,  p-value: 1.528e-10
```

Multicollinearity

If the predictors are highly correlated:

- The coefficients b_1 and b_2 change drastically between SLR and MLR
- The standard errors of the coefficients b_1 and b_2 increase for MLR
- The sum of squares decrease for MLR models

Multicollinearity

How to detect multicollinearity:

- ➢ Look at the correlation matrix of the predictors
- ➢ Compute the *V*ariance *I*nflation *F*actor (*VIF*) for each predictor

$$VIF_i = \frac{1}{1 - R_i^2}$$

Coefficient of determination for predicting X_i using the other predictors

- If $VIF = 1$, no issue
- If $VIF > 5$, investigate carefully
- If $VIF > 10$, some serious issues

Multicollinearity

How to handle multicollinearity:

1) Choose a better set of predictors
2) Eliminate some of the redundant predictors
3) Combine predictors into a scale
4) “Ignore” the individual coefficients and tests

USING R AND RSTUDIO

Addressing Potential Issues

In which cases should we consider data transformations?

- Nonlinearity → Predictor transformation
- Lack of normality
- Unequal variance

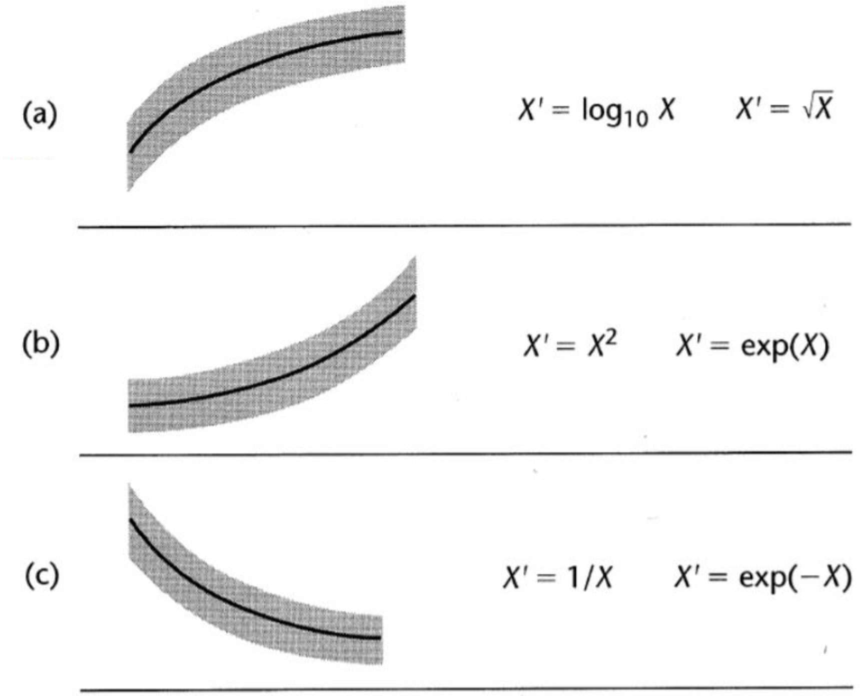
Response transformation

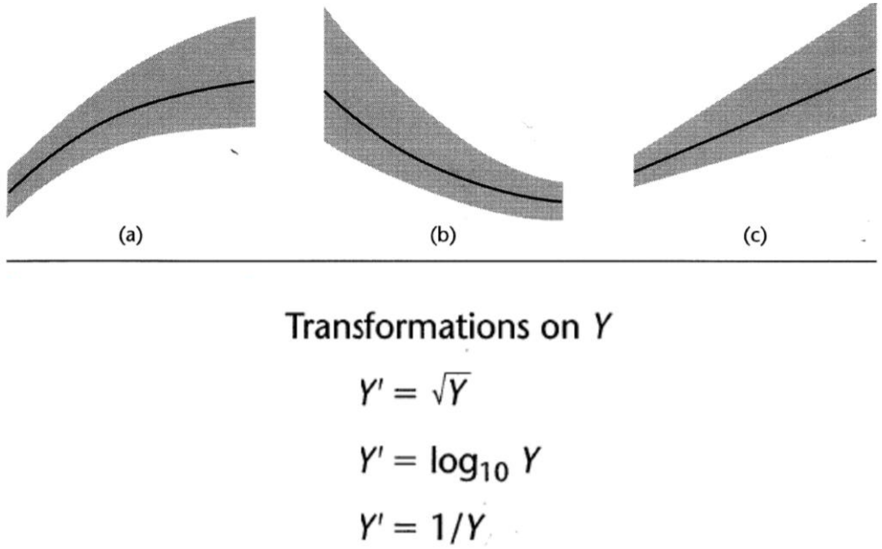
- Influential points

Addressing Potential Issues

Common transformations

- Logarithm
- Square root
- Exponential
- Power function
- Reciprocal





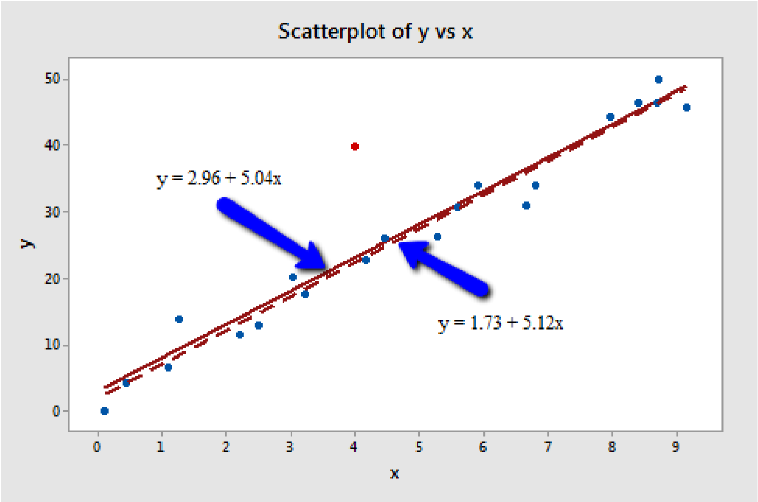
Addressing Potential Issues

How should we identify influential points?

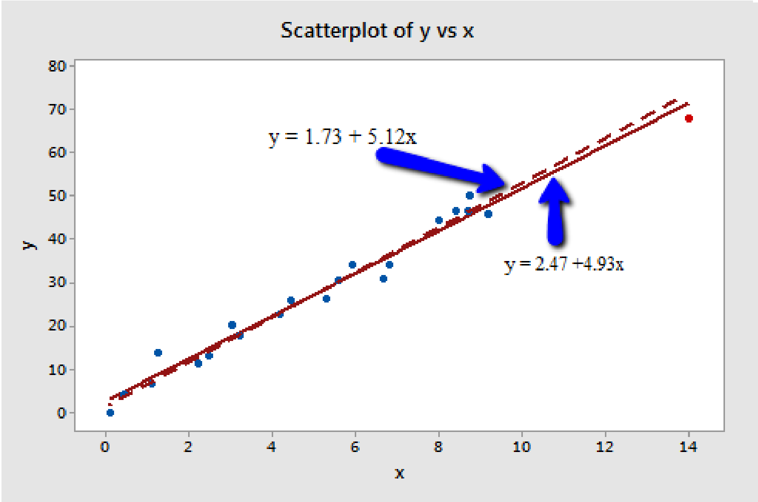
- An outlier is a point whose response Y does not follow general trend
- A data point with high leverage has an extreme X predictor value
- A data point is influential if it influences any part of the regression analysis (slope coefficients, predicted responses, …)

Addressing Potential Issues

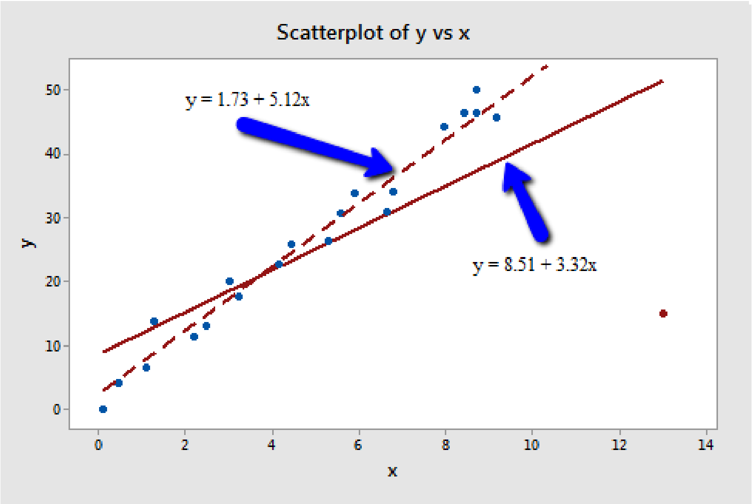
How should we identify influential points?



Outlier, no leverage



Not outlier, high leverage



Outlier, high leverage

Addressing Potential Issues

How should we identify influential points?

- An outlier is a point whose response Y does not follow general trend
- A data point with high leverage has an extreme X predictor value
- A data point is influential if it influences any part of the regression analysis (slope coefficients, predicted responses, …)
- Plot residuals or use Cook's distance

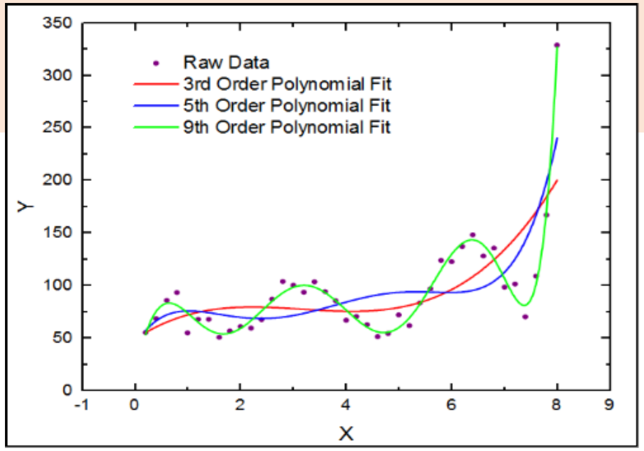
$$D_i = \frac{\sum\left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{p \cdot MS_{error}}$$

A data point having a large D_i indicates that the data point strongly influences the fitted values

BREAK TIME

BACK AT …

Polynomial Regression



- Include higher orders of one or more predictors:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + \varepsilon_i$$

with $x_{i1} = X_{i1} - \bar{X}$

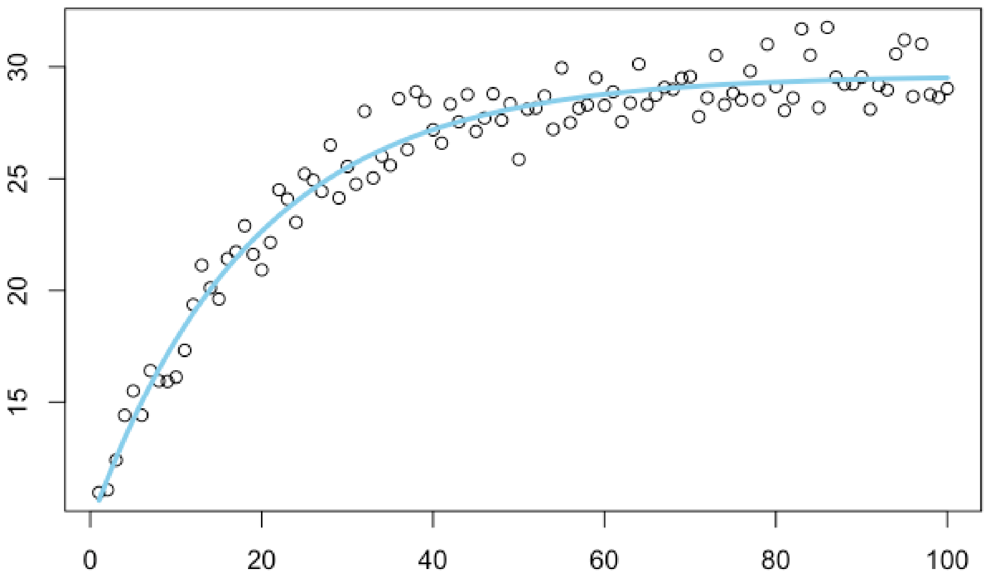
$$x_{i2} = X_{i2} - \bar{X}$$

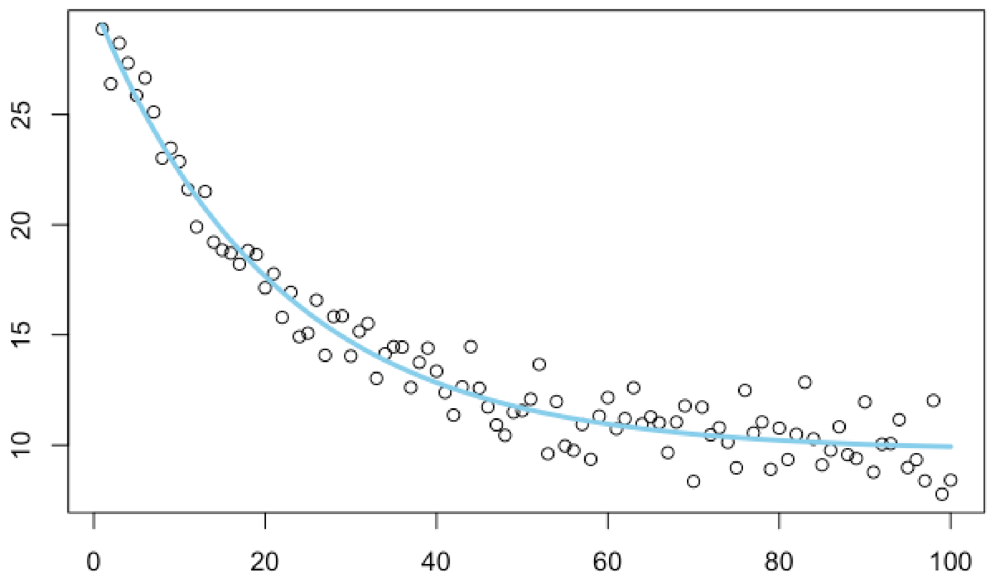
- We center the variables to reduce multicollinearity

Nonlinear Regression

Some example of nonlinear models:

Exponential model: $Y_i = \gamma_0 + \gamma_1 e^{\gamma_2 X_i} + \varepsilon_i$

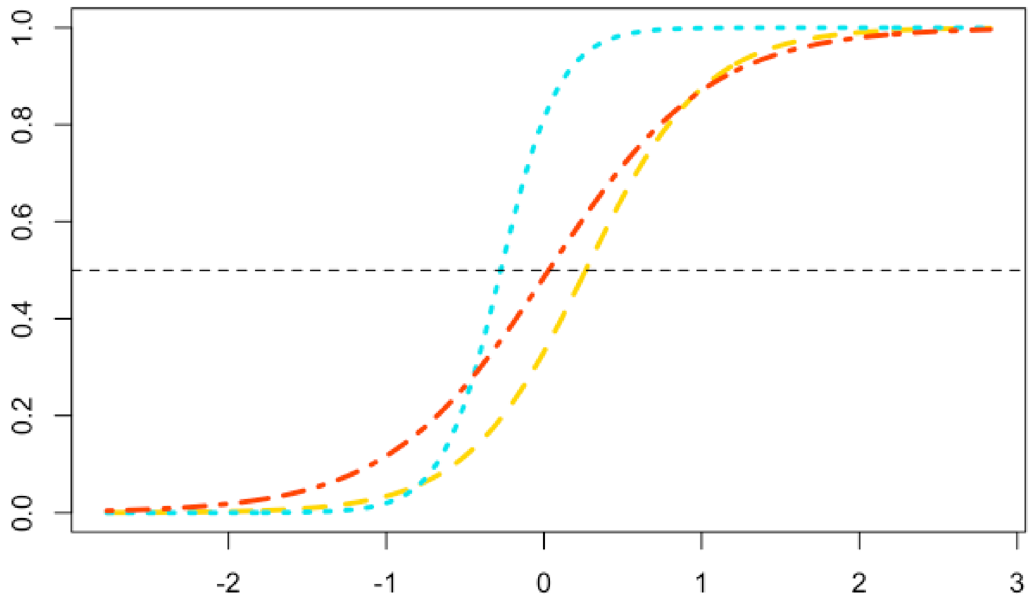




Nonlinear Regression

Some example of nonlinear models:

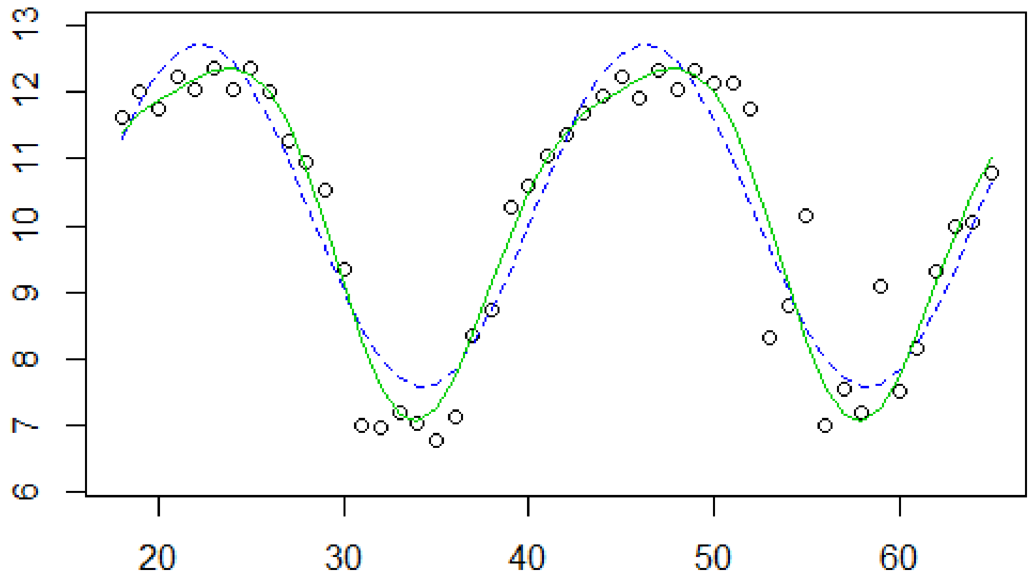
Logistic model: $Y_i = \frac{\gamma_0}{1+\gamma_1 e^{\gamma_2 X_i}} + \varepsilon_i$



Nonlinear Regression

Some example of nonlinear models:

Harmonic model: $Y_i = \gamma_0 + \gamma_1 \cos(\gamma_2 X_i + \gamma_3) + \varepsilon_i$



Nonlinear Regression

Similarities / Differences with linear regression:

- same definitions of sums of squares: $SS_{error} = \sum\left(Y_i - f(X_i, g)\right)^2$

But $SS_{error} + SS_{\mathrm{reg}}$ does not necessarily add up to SS_{total}
R^2 does not have a meaningful interpretation

- same method to estimate the coefficients: minimize sum of squares error (SSE)

But calculations differ (derivatives, Taylor series, Gauss-Newton's method…)

- same assumptions about the errors: normal, equal variance, independent

But residuals do not necessarily add up to 0, normality might be problematic
And the most important assumption: the model represents the data well (estimate parameters)

Nonlinear Regression

Similarities / Differences with linear regression:

- same diagnostics: residuals vs fitted values plot, normal probability plot

But the assumptions are rarely met, especially normality

- different methods for inferences

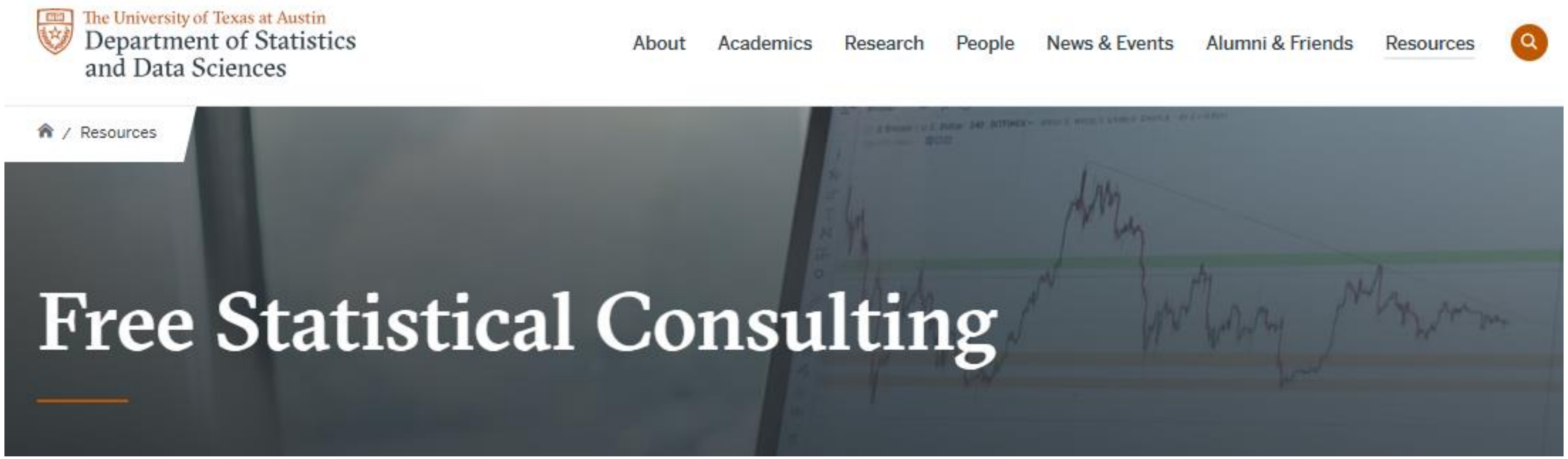
Inferences are difficult because the assumption of normality is not often met, and the coefficients may be biased. But:

  - larger sample size
  - bootstrapping

- different number of parameters vs predictors

We can have p parameters for q predictors $(p > q)$

Next



Please note that our consulting services are in high demand and reserved on a first-come, first-served basis.

All UT Austin graduate students, faculty and staff are eligible to sign up for a free 30-minute appointment to speak with a faculty member in the Department of Statistics and Data Sciences (SDS) for a brief consultation. An additional follow-up appointment may be arranged depending on appointment availability. All appointments will take place on Zoom.

To schedule an appointment, please email stat.admin@austin.utexas.edu and provide the following information:

- Full Name
- Title (e.g., graduate student, faculty member)
- Department/Program Affiliation
- Email address (should be a UT email)
- 1-2 paragraph summary of the issue you hope to discuss with the consultant
- Whether you have met previously with an SDS consultant

Next

If you have a dataset you'd like to explore, now is a great time to pull it up!

We’re happy to help you:

- ✓ Clean or organize your data
- ✓ Fit a model (e.g., linear, logistic, or multiple regression)

We can come by and take a look, but please note that we can’t guarantee we will be able to answer all questions.

Let’s see what we can discover!

Next

Please complete this quick survey so we can better understand future interest and expectations for this workshop.

https://forms.gle/caS8whnoA3S9ow4m6