Linear Regression
Week 7

PH 700A, Spring 2025

Rick Calvo

Table of contents

1 Continuous Data Analysis - Linear Regression

1.1 Outline

- Course Overview
- Assumptions Recap
- Linear Model Development
- Variable Selection
- Manual vs. Stepwise
- Interactions
- Regression Diagnostics

1.2 Overview

- Moving from *Data Preparation* and *Exploration* to *Data Analysis*
- Some analyses will require on-the-fly adjustments
- Analyses are *sequential* and *iterative*
- Descriptive analyses informs multivariable analyses
- Multivariable results drive model development
- Good model development equates to high model validity

1.3 From Last Week

Linear regression assumptions must be met for results to have validity.

1.3.1 Assumptions

- *Linearity* of Association
- *Independence* of Residuals
- *Normality* of Residuals
- *Equality* of Variances
- *Orthogonality* of the Predictors

1.3.2 Violations Require Treatment

- Mathematical Transformation
- Categorization
- Quantile Generation

1.4 Linear Regression Review

Used to evaluate relationship between *independent* variable(s) and a *continuous dependent variable.*

Linear regression models assess trajectory based on principles of geometric lines.

$y = b1x1 + b0 + E$

Where:

- $b1$ is the regression coefficient that is estimated (aka *line slope*)
- $x1$ is the independent variable containing a series of values
- $b0$ is the baseline estimate (aka y-intercept)
- E is the error term (aka the unaccounted variability)
- y is the continuous dependent variable with a series of values

1.5 Crude vs. Adjusted Models

"Crude models" regard one dependent variable and one primary independent variable.

"Adjusted models" regard one dependent variable, a primary independent variable, and at least one other independent variable.

"Adjusted models" == "Multivariable models"

Model development is the process of taking a *crude model* and selecting other independent variables to compile a *multivariable model* that is representative.

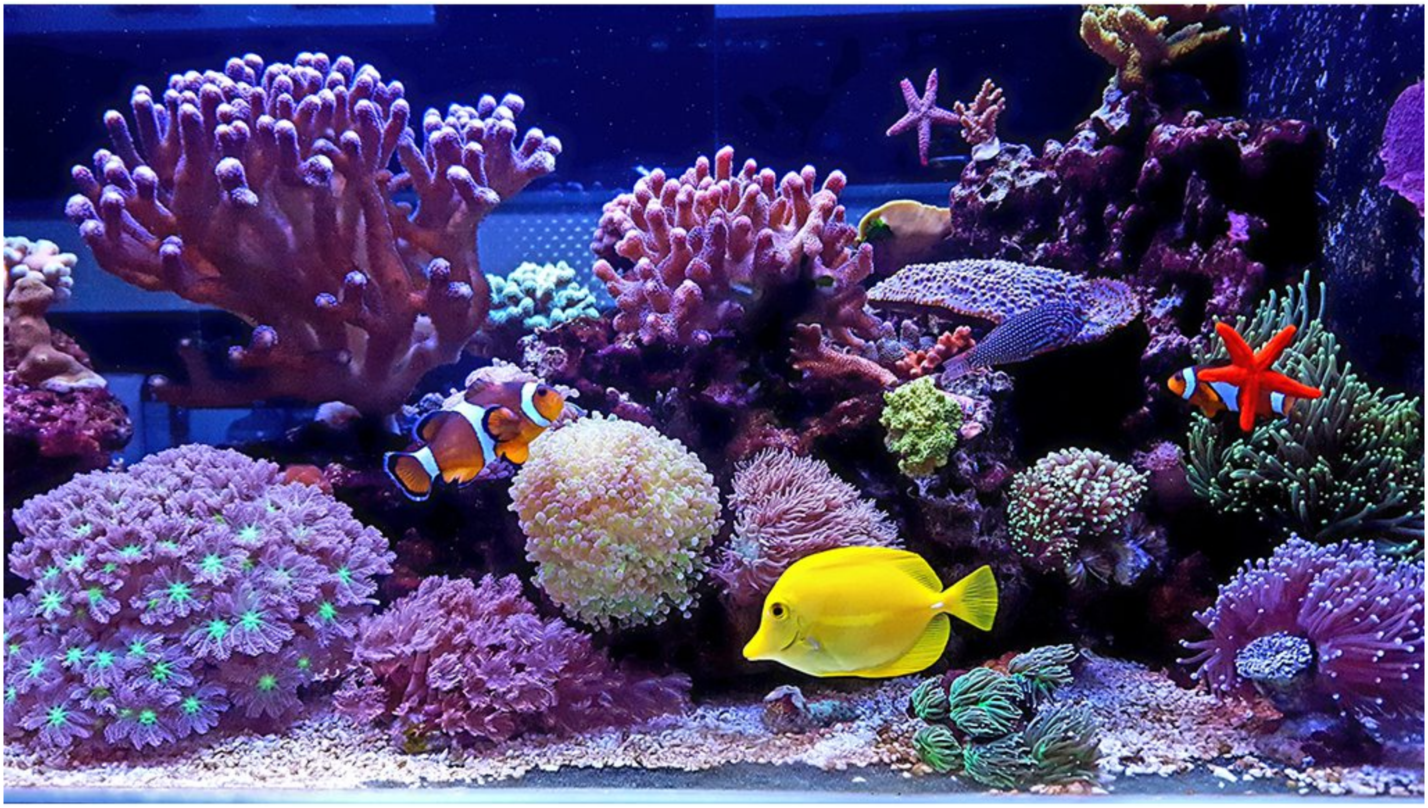
1.6 Statistical Modeling

1.6.1 What is a model?

- A mathematical relationship between measurements across observations
- A numerical representation of a **facet of health**
- A small example that can be used to extrapolate associations and create predictions

1.7 Diorama Analogy

What you're trying to describe:



How it gets described:

1.8 Multiple Linear Regression Details

We can add more independent variables to generate a longer equation to "predict" an average effect of `y` based on these variables

`y = b1x1 + b2x2 + b3x3 + ... + b0 + E`

Linear regression utilizes the *ordinary least squares* method of estimation.

Ideally:

- Every `x` variable is relevant and associated with `y`
- `E` is minimized as much as possible
- Each calculated `b` represents a "weight" applicable to the respective `x` and is applicable to the general population

Objective:

To identify best set of variables that fit our observations to "explain" the outcome.

1.9 Example Linear Regression Code

```
library(stats) # usually pre-loaded

model1 <- lm(formula = y ~ x1 + x2 + ... + xn, data = df)

summary(model1)
```

- `lm()` the primary command
- `formula =` parameter to designate the variables to assess
- `y` is the dependent continuous variable in your data `df`
- `~` symbolizes the equals sign in R formula notation
- `x1 + x2 + ... + xn` is the additive list of all your independent variables to include in the model
- `model1` is the object that will hold the results of the `lm()` procedure

1.10 Example Linear Regression Output

```
table(df$arrival_transport)
```

```

AMBULANCE     OTHER   UNKNOWN   WALK IN 
      133         1        13        75 
```

```
table(df$disposition)
```

```

                   ADMITTED                      ELOPED 
                        150                           1 
                       HOME LEFT AGAINST MEDICAL ADVICE 
                         60                           2 
    LEFT WITHOUT BEING SEEN                       OTHER 
                          2                           2 
                   TRANSFER 
                          5 
```

```
model1 <- lm(formula = as.numeric(hlos) ~ arrival_transport + o2sat, data = df)
summary(model1)
```

```

Call:
lm(formula = as.numeric(hlos) ~ arrival_transport + o2sat, data = df)

Residuals:
   Min     1Q Median     3Q    Max 
-533.0 -240.4 -136.1   60.6 3837.8 
```

```
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)  
(Intercept)              -2103.52    1356.46  -1.551   0.1226  
arrival_transportOTHER    -232.92     527.41  -0.442   0.6593  
arrival_transportUNKNOWN  -338.66     192.09  -1.763   0.0795 .
arrival_transportWALK IN   -77.38      78.26  -0.989   0.3240  
o2sat                       27.26      13.88   1.963   0.0511 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 524.8 on 193 degrees of freedom
  (24 observations deleted due to missingness)
Multiple R-squared:  0.03874,	Adjusted R-squared:  0.01882 
F-statistic: 1.945 on 4 and 193 DF,  p-value: 0.1046
```

Estimates to focus on:

- `adjusted R\^2`: what % of y is being explained by the vars
- `Estimate`: the beta estimates for each of the independent variable levels vs. a reference
- `Pr(\>\|t\|)`: the p-value associated with the independent var
- `Residual standard error`: amount of unallocated error
- `missingness`: number of observations used for the analysis
- `F-statistic + p-value`: fitness of the model vs. a horizontal line

1.11 Model Development Modalities

Model construction can be:

- Simple to Complex
- Complex to Simple
- Manually Developed
- Automated Development

1.12 Variable Selection

How do you select a variable to include?

When it is part of your hypothesis.

When a mechanism dictates its importance.

When you need to 'control' for interference by a factor.

If its presence helps address other issues with your sample.

Retention of a variable is determined after assessing model estimates.

1.13 Major Variable Selection Assumptions

- There is a biological or mechanistic rationale for its use
- They represent a feature that needs to be controlled for
- Continuous variables are normally distributed
- Categorical variables are *approximately* balanced
- Independent variables are not significantly correlated
- "Outliers" are addressed or rationalized
- Missing values are addressed
- The variables **represent what they purport to measure**

1.14 Independent Variables Assessment

Selection of variables must be initially informed by *bivariate analyses*.

Cannot arbitrarily include variables.

Continuous Vars	Categorical Vars
Pearson `cor()`	2-cats `t.test()`
Bivariate `lm()`	>2-cats `anova()`
	Bivariate `lm()`

Make sure your categorical vars are set as `factor()` and your continuous vars are set as `numeric()`!

1.15 Number of Variables "Rule of Thumb"

~ 10 obs per category/variable

Overfitting

Adding too many variables will for an overfit model into your data and predictions lose validity. Model development requires balance between using important variables and superfluous variables.

1.16 Automated Model Building

A model can be automatically fit based on *patterns in the data* and *pre-specified criteria*.

Stepwise methods include:

- Forward selection
- Backward elimination
- Hierarchical selection

Default method is based on optimizing the `Akaike's Information Criterion (AIC)` to select variables.

Many caveats exist.

1.17 Model Optimization

Stepwise methods can be based on a variety of optimized statistics.

The *Akaike information criterion (AIC)* is an estimator of *prediction error* and an estimate of the *relative* quality of a model for a given set of data. **Lower AIC = Better Model**.

Other options include:

- Likelihood Ratio Test ("LRT")
- F-test ("F")
- Bayesian Information Criterion (BIC)

1.18 Forward Stepwise Code

Typically two models must be specified; an "empty" model and a "full" model.

However, it is easier to `select()` ALL the variables you want to assess in a temporary data frame and just tell R to use all the variables in the frame.

Variables will be iteratively added and only retained if they are significantly associated with the dependent variable.

Development will stop once reaches an optimal AIC after exhausting all variables.

```r
tempdf <- df %>% select(y, x1, x2, x3, x4, ... xn)

emptymodel <- lm(y ~ 1, data = tempdf)

fwmodel <- step(emptymodel,
                scope = list(lower = ~ 1, upper = ~ . - 1),
                direction = "forward")

summary(fwModel)
```

This will perform the forward stepwise selection process starting with an empty model and iterate through all variables in the `tempdf` data frame.

- `step()` is the command for stepwise methods
- `emptymodel` is the object containing the intercept-only model
- `scope()` lists the range of models to be considered. Only the right-hand side of the equation needs to be specified
- `lower = ~ 1` represents the empty model
- `upper = ~ . - 1` represents the full model with all predictor variables minus the dependent variable

1.19 Backward Stepwise

Backward stepwise is a process of elimination.

It starts with a full model containing *all* predictor variables and sequentially removes variables that are not significant.

Only one model needs to be specified (the full model).

```
tempdf <- df %>% select(y, x1, x2, x3, x4, ... xn)

fullmodel <- lm(y ~ ., data = tempdf)

bwmodel <- step(fullmodel,
           direction = "backward")

summary(bwmodel)
```

1.20 Hierarchical

```
tempdf <- df %>% select(y, x1, x2, x3, x4, ... xn)

fullmodel <- lm(y ~ ., data = tempdf)

bothModel <- step(fullmodel,
                  direction = "both")
```

Variables will be sequentially removed and re-added one at a time starting with a full model.

The AIC is prioritized over any statistical significance changes among covariates.

1.21 Major Caveats with Stepwise Methods

- Variable selection is based on patterns in the data and AIC optimization
- Biological plausibility, interrelatedness in variables, and mechanisms of action will not matter to it
- Resulting models may not make sense
- Statistical significance may be attributable to data artifacts
- Omission of important variables may occur
- High internal validity at a sacrifice of external validity
- Overfitting risk
- Multicollinearity in retained predictors

2 Regression Diagnostics

2.1 Multicollinearity

A phenomenon where multiple variables are correlated and retained in a model.

Results in overemphasis of a specific facet used to "explain" the outcome.

Also leads to inaccuracies in the estimate for correlated factors.

Use the *Variance Inflation Factor* to assess multicollinearity.

```r
library(car)

vif(model)
```

As collinearity increases, VIF also increases.

A threshold of **VIF $>$ 10** is obvious clear collinearity

VIF $>$ 5 is questionable collinearity requiring variable removal

VIF $>$ 2 requires investigation; potentially coincidental

2.2 Tolerance

Tolerance is the inverse of the VIF

As collinearity increases, tolerance decreases.

A threshold for collinearity is **Tolerance $<$ 0.10**

2.3 Statistical Interaction

Requires stratification or restriction to address

A finding of your study, but presents a major flaw if not identified

Use `anova()` to look at pairwise associations between factors for non-linear effects.

```r
model <- lm(formula = y ~ (x1 + x2 + x3 + x4)^2, data = df)

anova(model)
```

The above code generates every pairwise multiplicative effect between variables `x1` through `x4`.

Although every multiplicative effect is shown with the `anova()` procedure, you should only focus on *suspected interactions*.

Interactions should be assessed at a $p < 0.10$ threshold.

2.4 Residual Analysis

Residuals are the difference between each observation and the sample mean for each continuous variable.

Larger residuals = Larger `E` term

Four main types of residuals:

- Ordinary
- Standardized
- Studentized
- Jackknife

2.5 Ordinary Residual Plot

```r
plot(model$residuals, main = "Residual Plot", ylab = "Residuals")
```

Every object that is output from the `lm()` procedure will have the residuals stored within.

The assumption of normality of the residuals can be visually assessed here.

Allows for viewing the effect of potential outliers.

2.6 Other Residual Plots

```
plot(model, which=1)
```

The `which =` argument refers to the following pre-set plots:

value	Plot Type	Visual Evaluation
1	Tukey-Anscombe Residuals vs. Fitted	Look for monotonicity
2	Standard Residual Q-Q	Evaluate alignment across quantiles
3	Scale-Location	Assess for homoskedasticity around fit line, find high leverage outliers
4	Cook's Distance Thresholds	Identification of outliers at extreme thresholds
5	Residuals vs Leverage	Identification of outliers at extreme thresholds
6	Cook's dist vs Transformed Leverage	Identification of outliers at extreme thresholds

2.7 General Rules in Outlier Treatment

- Investigate outlier for data entry error; correct if flawed and re-run model
- If value(s) is erroneous or biased, exclusion is appropriate
- If value(s) appear correct but simply extreme, they are representative and should be retained
- Outliers are most dangerous in small samples and, generally speaking, are handled more delicately anyways, so exercise caution