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.

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \]

\( signal \ 
\frac{\text{noise}}{\text{noise}} \)
• Single response and single predictor (simple linear regression)
• Single response and multiple predictors (multiple linear regression)
• Multiple responses and predictors (multivariate linear regression)

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

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

• 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?
Structures and Uses

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

**Predictor variable attributes**

<table>
<thead>
<tr>
<th>Level of race</th>
<th>New variable 1 (x1)</th>
<th>New variable 2 (x2)</th>
<th>New variable 3 (x3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Hispanic)</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2 (Asian)</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3 (African American)</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4 (white)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**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

- 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)

**Response variable attributes**

**Linear Regression**
- Predicted $Y$ can exceed 0 and 1 range

**Logistic Regression**
- Predicted $Y$ lies within 0 and 1 range

**Poisson Distribution Formula**

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

where

- $\lambda = 0, 1, 2, 3, ...$
- $\lambda$ = mean number of occurrences in the interval
- $e$ = Euler’s constant $\approx 2.71828$
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)
R

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

summary(model1)
```

```R
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")
```

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

PROC GLIMMIX data=covid2;
  where cases>0;
  class region(ref='West');
  model
    dpc = region
    bed bedut 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;

Fit Statistics
- 2 Log Likelihood - 7750.40
  AIC (smaller is better) - 7714.40
  AICC (smaller is better) - 7714.03
  BIC (smaller is better) - 7614.94
  CAIC (smaller is better) - 7596.94
  HQIC (smaller is better) - 7677.74

Pearson Chi-Square 2432.23
  Pearson Chi-Square / DF 1.32

Parameter Estimates

| Effect       | Estimate | Standard Error | DF  | t Value | Pr > |t| Alpha | Lower  | Upper  | Mean  | Standard Error | Lower Mean | Upper Mean |
|--------------|----------|----------------|-----|---------|-------|------|-------|--------|--------|----------------|------------|------------|
| Intercept    | -5.7165  | 0.7877         | 1838| -7.26   | <.001 | 0.05 | -2.8954| -2.7577| 0.05591| 0.001852    | 0.05238    | 0.05965    |
| region Midwest | -2.8265  | 0.03509        | 1838| -80.54  | <.001 | 0.05 | -2.8954| -2.7577| 0.05591| 0.001852    | 0.05238    | 0.05965    |
| region Northeast | -2.7643  | 0.05079        | 1838| -54.43  | <.001 | 0.05 | -2.8639| -2.6847| 0.05928| 0.002833    | 0.05397    | 0.06509    |
| region Southeast | -3.1417  | 0.03291        | 1838| -95.46  | <.001 | 0.05 | -3.2063| -3.0772| 0.04142| 0.001307    | 0.03893    | 0.04406    |
| region Southwest | -2.7466  | 0.06136        | 1838| -44.76  | <.001 | 0.05 | -2.8670| -2.6263| 0.06028| 0.003476    | 0.05381    | 0.06747    |
| region West    | -2.9191  | 0.06509        | 1838| -44.84  | <.001 | 0.05 | -3.0468| -2.7915| 0.05122| 0.003163    | 0.04536    | 0.05779    |

Type III Tests of Fixed Effects

<table>
<thead>
<tr>
<th>Effect</th>
<th>Num DF</th>
<th>Den DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>region</td>
<td>4</td>
<td>1838</td>
<td>17.35</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>bed</td>
<td>1</td>
<td>1838</td>
<td>7.92</td>
<td>0.0049</td>
</tr>
<tr>
<td>bedut</td>
<td>1</td>
<td>1838</td>
<td>4.96</td>
<td>0.0261</td>
</tr>
<tr>
<td>vent</td>
<td>1</td>
<td>1838</td>
<td>0.01</td>
<td>0.9321</td>
</tr>
<tr>
<td>famsize</td>
<td>1</td>
<td>1838</td>
<td>0.01</td>
<td>0.9212</td>
</tr>
<tr>
<td>perc_white</td>
<td>1</td>
<td>1838</td>
<td>0.35</td>
<td>0.5523</td>
</tr>
<tr>
<td>perc_black</td>
<td>1</td>
<td>1838</td>
<td>0.14</td>
<td>0.7034</td>
</tr>
<tr>
<td>perc_hisp</td>
<td>1</td>
<td>1838</td>
<td>23.02</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>sex_ratio</td>
<td>1</td>
<td>1838</td>
<td>1.26</td>
<td>0.2624</td>
</tr>
<tr>
<td>med_age</td>
<td>1</td>
<td>1838</td>
<td>68.79</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>med_income</td>
<td>1</td>
<td>1838</td>
<td>3.11</td>
<td>0.0779</td>
</tr>
<tr>
<td>perc_insure</td>
<td>1</td>
<td>1838</td>
<td>4.57</td>
<td>0.0327</td>
</tr>
<tr>
<td>perc_poverty</td>
<td>1</td>
<td>1838</td>
<td>22.50</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>
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

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
<table>
<thead>
<tr>
<th></th>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Multivariate linear regression involves what number of response and predictor variables?</td>
<td>More than one response variable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>More than one predictor variable</td>
</tr>
<tr>
<td>2.</td>
<td>Is polynomial regression linear or non-linear? Why?</td>
<td>Still linear, because regression coefficients are still linear</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_h X^h + \epsilon,$</td>
</tr>
<tr>
<td>3.</td>
<td>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.</td>
<td>Logistic, Binomial, Multinomial, Ordinal, Poisson, Negative Binomial, Quasi Poisson, Gamma, Exponential, Cox, etc.)</td>
</tr>
<tr>
<td>4.</td>
<td>What type of model includes both fixed and random effects?</td>
<td>Mixed model</td>
</tr>
<tr>
<td>5.</td>
<td>Fill in the blanks: Linear regression models the relationship between the ______ variable and one or more ______ variables</td>
<td>Response, Predictor</td>
</tr>
<tr>
<td>6.</td>
<td>What type of regression is pictured to the right?</td>
<td>Join-point/piecewise regression</td>
</tr>
</tbody>
</table>
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