Regression Methods

Linear, Logistic, Survival and Poisson

Samuel Alan Stewart, PhD

Medical Informatics
Department of Community Health & Epidemiology
Dalhousie University, Halifax, Canada
sam.stewart@dal.ca
@medInfProf

July 11, 2017

- Start with a simple review of regression:
  - Simple and multiple linear regression
  - Logistic regression
  - Cox PH regression
- Poisson regression

Modeling Process

1. Develop the model
   - This is where the type and components are defined
2. Estimate the coefficients
3. Evaluate the model fit
4. Test the regression coefficients
5. Test the regression assumptions
   - This isn't always done last in practice, but for these lectures we'll present it there

Linear Regression

1. Developing the Model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Y is the response variable, a continuous value (i.e a real number)
- X is the predictor variable
- β_0 is the intercept term
- β_1 is the regression coefficient (or the slope)
- ϵ is the error term

2. Estimating Regression Coefficients

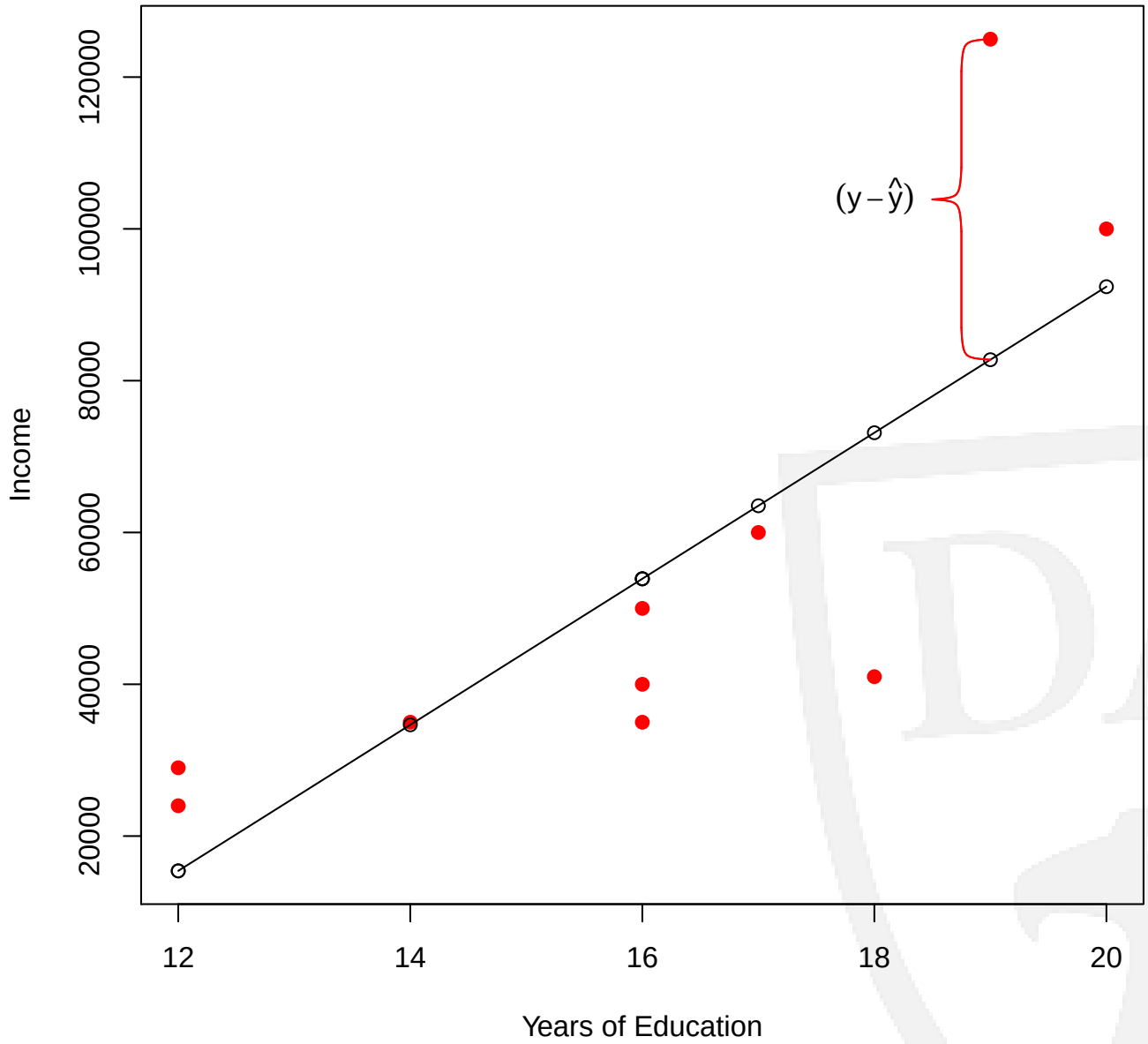
- We need to come up with estimates of β_0 and β_1 that are as close as possible to the observed values

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- We're going to minimize the difference between $\hat{Y}$ and Y by minimizing the following equation

$$S(Y) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

- This is called *Least Squares Regression*
- We can come up with equations for $\hat{\beta}_1$ and $\hat{\beta}_0$


Income
20000
40000
60000
80000
100000
120000
12
14
16
18
20
Years of Education
$(y - \hat{y})$

Using Least Squares

$$
\begin{aligned}
\frac{\delta S(Y)}{\delta \hat{\beta}_1} &= \sum_{i=1}^{n} 2(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(-X_i) \\
0 &= \sum Y_i X_i - \sum \hat{\beta}_0 X_i - \sum \hat{\beta}_1 X_i^2 \\
\hat{\beta}_1 \sum X_i^2 &= \sum Y_i X_i - \hat{\beta}_0 \sum X_i \\
\hat{\beta}_1 \sum X_i^2 &= \sum Y_i X_i - (\overline{Y} - \hat{\beta}_1 \overline{X}) \sum X_i \\
\hat{\beta}_1 \sum X_i^2 &= \sum Y_i X_i - n\overline{YX} - \hat{\beta}_1 n\overline{X}^2 \\
\hat{\beta}_1 \left(\sum X_i^2 + n\overline{X}^2 \right) &= \sum Y_i X_i - n\overline{YX} \\
\hat{\beta}_1 &= \frac{\sum Y_i X_i - n\overline{YX}}{\sum X_i^2 + n\overline{X}^2} \\
\hat{\beta}_1 &= \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}
\end{aligned}
$$

Using Least Squares

$$
\begin{aligned}
S(Y) &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \\
S(Y) &= \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \\
\frac{\delta S(Y)}{\delta \hat{\beta}_0} &= \sum_{i=1}^{n} 2(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(-1) \\
0 &= \sum Y_i - \hat{\beta}_0 n - \hat{\beta}_1 \sum X_i \\
\hat{\beta}_0 n &= \sum Y_i - \hat{\beta}_1 \sum X_i \\
\hat{\beta}_0 &= \overline{Y} - \hat{\beta}_1 \overline{X}
\end{aligned}
$$

Sum of Squares

- These values will re-occur later, so we define them explicitly

$$\begin{aligned} S_{xx} &= \sum (x_i - \overline{x})^2 \\ &= (n-1)V(X) \end{aligned}$$

$$\begin{aligned} S_{yy} &= \sum (y_i - \overline{y})^2 \\ &= (n-1)V(y) \end{aligned}$$

$$\begin{aligned} S_{xy} &= \sum (x_i - \overline{x})(y_i - \overline{y}) \\ &= (n-1)Cov(X, Y) \end{aligned}$$

Sum of Squares

$$\begin{aligned}\hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \\ &= r\sqrt{\frac{V(Y)}{V(X)}} \\ &= \frac{Cov(X,Y)}{V(X)}\end{aligned}$$

Regression Coefficients

$$\hat{\beta}_1 = \frac{Cov(X,Y)}{V(X)} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Model Sources of Error

- There are three measures of deviation to calculate in a regression
  - $(y_i - \hat{y}_i)$ is the residual (difference between observed and estimated value)
  - $(y_i - \overline{y})$ are the deviations of the observations from a non-predictive model (The *null* model)
    - ⋆ i.e., from a model that only uses the intercept term
  - $(\hat{y}_i - \overline{y})$ is the difference between the no regression and regression model

3. Evaluate the Model Fit

- We're going to calculate these three values as sums of squares

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2 = S_{yy}$$

- We're going to test to see if the regression sum of squares is more that the error sum of squares

Regression Table (Analaysis of Variance (ANOVA) table)

Source of Variation	Sum of Squares (SS)	df	Mean Squares	F
Model	$SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Error	$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$	
Total	$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2$	$n-1$		

- We want to know if the variability accounted for by the regression is a significant portion of the overall variation
- Note that $SST = SSR + SSE$

4. Testing the Regression Coefficients

- We can test the regression coefficients individually using a **Wald test**
- We test β_1 using a standard t-test
  - The test statistic is

$$t = \frac{\hat{\beta}_1}{se_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\sqrt{MSE/S_{xx}}}$$

  - Compare that to a t-distribution with (n-2) degrees of freedom
- We can develop a similar statistic for β_0 (the details will wait until we do multiple regression)

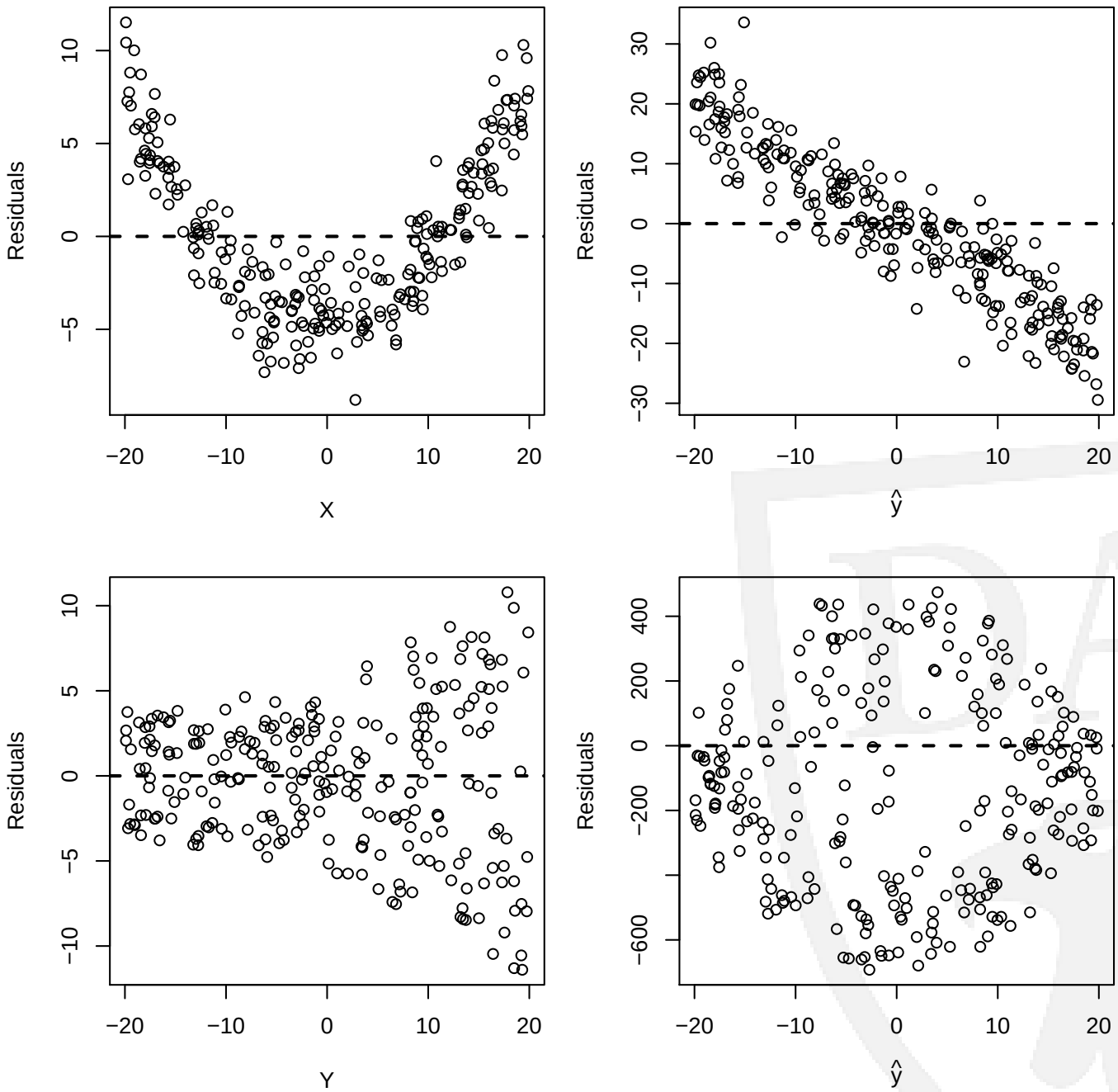
5. Test the Regression Assumptions

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- The observations are independent of one another
- $\epsilon \sim N(0, \sigma^2)$
- There are four separate assumptions to check
  1. The variance of the residuals is σ^2 for all observations
  2. The residuals are independent
  3. The residuals are approximately normally distributed
  4. There are no outliers biasing the regression

Analyzing the Residuals

- Let $r_i = y_i - \hat{y}_i$ (r_i is the residual for the i^{th} observation)
- We'll start by looking at a scatter plot of the residuals
- We're looking for patterns in the residual plot, for extreme residual values, and for the residuals to be centred tightly around 0
- We plot the residuals, r_i on the y-axis and one of the following on the x-axis: $\hat{y}_i$, x, y_i
- We're hoping to see no patterns within the plot


Residuals
X
Residuals
ŷ
Residuals
Y
Residuals
ŷ

Evaluating the Residuals Plots

- Problems with the residual plots: it means there's a problem with the model specification
- We're fitting the model $Y = \beta_0 + \beta_1 X + \epsilon$
  - What if the **true** model is $y = X^2 + \epsilon$
- We need to transform one or more of the variables in order to improve the modelling process
- We can also recognize outliers with this method

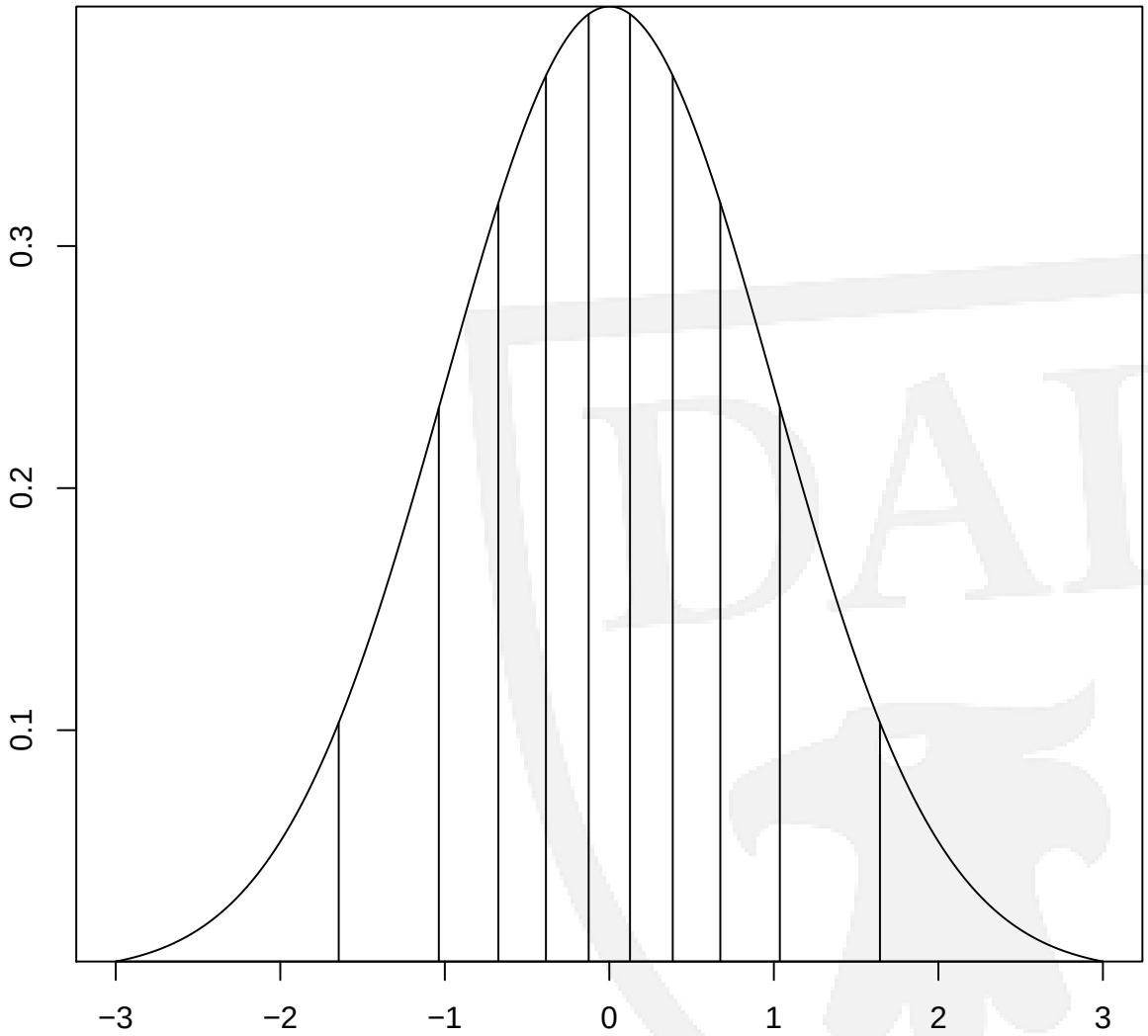
Q-Q Plots

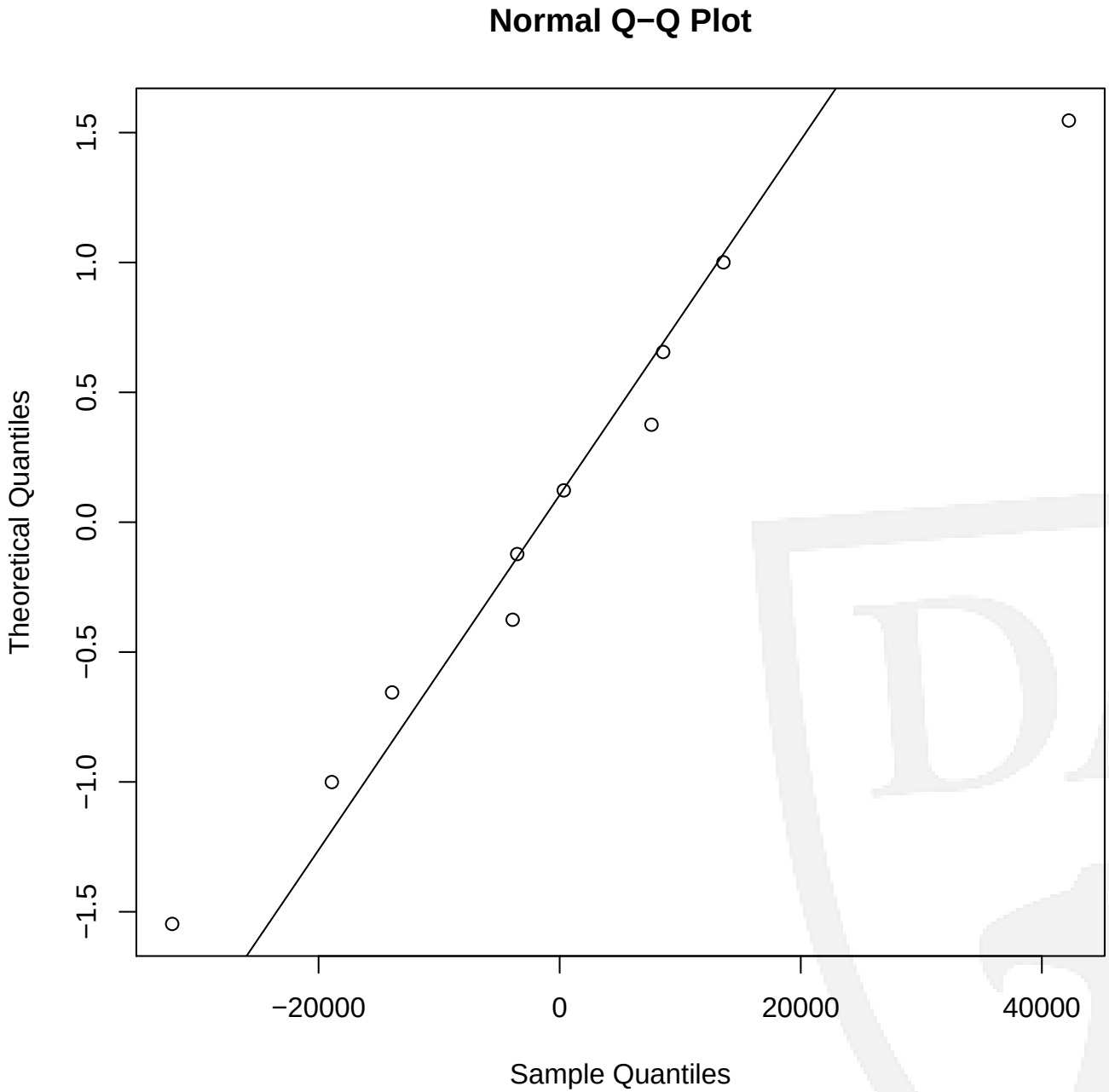
- *Quantile-Quantile Plots* are ways to evaluate if a variable is normal
- We plot the variable we are testing on the x-axis
  - In this case, r_i
- On the y-axis we plot the corresponding values of the normal distribution
  - This is dependent on the number of observed values
- The results should be a set of points that fall on a straight line

Quantile Values

For 10 quantiles

Quantile	Value
5%	-1.64
15%	-1.04
25%	-0.67
35%	-0.39
45%	-0.13
55%	0.13
65%	0.39
75%	0.67
85%	1.04
95%	1.64




Normal Q−Q Plot
Theoretical Quantiles
Sample Quantiles
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
−20000
0
20000
40000

Q-Q Plots

- The idea behind Q-Q plots is to make sure the residuals could have come from a normal distribution
- Let $r_{(1)}$ be the smallest residuals and $r_{(n)}$ be the largest
- We plot the sorted residuals against the expected probability of the smallest to largest values (the quantiles)
- When we inspect the Q-Q plot, we pay particular attention to the middle section (not the extreme values)
  - Between Q1 and Q3 of the residuals

Outliers

- Outliers are observations that don't fit with the rest of the data
- Outliers can bias the regression, resulting in insignificant findings
- It is important to *understand the nature of the outlier:* Is the observation a mistake, or a legitimate value that doesn't fit the rest of the data
- We can't just remove observations because they don't fit the data: we need a legitimate reason

Detecting Outliers

- There are a couple of different methods for detecting outliers
  - The simplest is to just plot y vs. x and look for the outlying points
  - Fit regression lines with and without the outlier in the dataset, compare the results
  - There are formal tests developed to detect the influence an individual observation has on a regression. These include Cook's Distance and the Barnett and Lewis test.
    - ⋆ We will not explore these in detail

Multiple Linear Regression

1. Develop the Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \epsilon$$

- Y is the continuous response variable
- X_1 is the first predictor variable
- β_0 is the intercept term
- β_1 is the first regression coefficient
- ϵ is the error term

$$\epsilon \sim N(0, \sigma^2)$$

2. Estimate the Coefficients

Solving for $\hat{\beta}_i$

- We're going to try and minimize the difference between $\hat{y}$ and y

$$\begin{aligned} S(B) &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 X_1 - \beta_2 X_2 - \ldots - \beta_k X_k)^2 \end{aligned}$$

- Unlike the simple linear system, we can't come up with an explicit equation for $\hat{\beta}_i$
- Instead we're going to need matrix multiplication

Matrix Theory (FOR NOTES ONLY)

Matrix Theory

- Matrices are $r \times c$ arrays of numbers with r rows and c columns, represented by capital letters
- A^T or A' is a re-positioning of a matrix where the rows and columns switch

$$A = \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} \quad A^T = \begin{bmatrix} a & d \\ b & e \\ c & f \end{bmatrix}$$

- Addition and subtraction are done using pairwise comparisons

$$\begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} - \begin{bmatrix} u & v & w \\ x & y & z \end{bmatrix} = \begin{bmatrix} a-u & b-v & c-w \\ d-x & e-y & f-z \end{bmatrix}$$

- Note that they can only be done on matrices of the same size

Matrix Multiplication

- Matrix multiplication is much more complicated
- Is done in a set order: $A \times B \neq B \times A$
- The number of columns in the first matrix must equal the number of rows in the second

$$A = \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} \qquad C = \begin{bmatrix} u & v \\ w & x \\ y & z \end{bmatrix}$$

$$A \times C = \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} \begin{bmatrix} u & v \\ w & x \\ y & z \end{bmatrix}$$

Matrix Multiplication

$$\begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} \begin{bmatrix} u & v \\ w & x \\ y & z \end{bmatrix} = \begin{bmatrix} au + bw + cy & av + bx + cz \\ du + ew + fy & dv + ex + fz \end{bmatrix}$$

- Before you try and perform matrix multiplication, check the dimensions of the two starting matrices, and the resulting matrix

$$\begin{array}{ccccc} A & \times & C & = & D \\ (2 \times 3) & \times & (3 \times 2) & = & (2 \times 2) \end{array}$$

- The final step is understanding inverse matrices

Inverse Matrices

- In scalar math, inverting is simple

$$\begin{aligned} x^{-1} &= \frac{1}{x} \\ x^{-1}x &= 1 \end{aligned}$$

- We define the inverse of a matrix such that $XX^{-1} = X^{-1}X = \mathrm{I}$, where I is the *identity matrix*

$$\mathrm{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- Because we require the order of the multiplication to not matter, *we can only invert square matrices*

Inverse Matrices

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

$$A^{-1} = \begin{bmatrix} d & -c \\ -b & a \end{bmatrix} \frac{1}{|A|}$$

- Where $|A|$ is the *determinant* of A.

$$|A| = ad - bc$$

- Note that inverting is possible for *almost* every square matrix, but the equations are far too complex for matrices bigger than (2×2)

Matrix Math

- Matrix math is hard, but incredibly valuable
- If you don't get it and want to, there's a Ted Ed lecture that might help:
  `https://www.youtube.com/watch?v=kqWCwwyeE6k`
- We will use matrix math (and matrix algebra) to get our estimates of $\hat{\beta}$

Multiple Linear Regression in Matrices

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \epsilon$$

- We're going to re-write the equation in matrix form:

Regression Equation in Matrix Form

$$Y = X\beta + \epsilon$$

$$Y = \underset{(n \times 1)}{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}} \quad X = \underset{(n \times (k+1))}{\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & \cdots & x_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}} \quad \beta = \underset{((k+1) \times 1)}{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}} \quad \epsilon = \underset{(n \times 1)}{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}}$$

Matrix Regression

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} (\beta_0)1 + (\beta_1)x_{11} + (\beta_2)x_{12} + \ldots + (\beta_k)x_{1k} \\ (\beta_0)1 + (\beta_1)x_{21} + (\beta_2)x_{22} + \ldots + (\beta_k)x_{2k} \\ \vdots \\ (\beta_0)1 + (\beta_1)x_{n1} + (\beta_2)x_{n2} + \ldots + (\beta_k)x_{nk} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_k x_{1k} + \epsilon_1 \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_k x_{2k} + \epsilon_2 \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_k x_{nk} + \epsilon_n \end{bmatrix}$$

Estimating β

- The least squares equation is now defined as

$$S(B) = (y - \hat{y})'(y - \hat{y})$$

- We can use matrix algebra procedures to solve this for an estimate of β

$$\hat{\beta} = (X'X)^{-1}X'y$$

Estimating β

$$
\begin{aligned}
S(B) &= (y - \hat{y})'(y - \hat{y}) \\
&= (y' - \beta' X')(y - X\beta) \\
&= y'y - \beta' X' y - y' X \beta + \beta' X' X \beta \\
&= y'y - 2\beta' X' y + \beta' X' X \beta \\
\frac{\partial S(B)}{\partial \hat{\beta}} &= 0 - 2X'y + 2X'X\beta \\
0 &= -X'y + X'X\hat{\beta} \\
X'y &= X'X\hat{\beta} \\
(X'X)^{-1}X'y &= (X'X)^{-1}X'X\hat{\beta} \\
\hat{\beta} &= (X'X)^{-1}X'y
\end{aligned}
$$

3. Evaluate the Model Fit

- Recall the simple linear regression estimates:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2 = S_{yy}$$

- These values are all based of y, which hasn't changed, so they still hold for multiple linear regression

Using the Matrix Values

- $SST = (n-1)V(y) = y'y$
- $SSR = \hat{\beta}' X' y$
- $SSE = SST - SSR$
- The regression table is the same as before

Source of Variation	Sum of Squares (SS)	df	Mean Squares	F
Model	SSR	k	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$
Error	SSE	$n-k-1$	$MSE = \frac{SSE}{n-k-1}$	
Total	SST	$n-1$		

4. Evaluate the Model Fit

Splitting SSR

- We can evaluate each independent variable individually using sums of squares
- There are several different ways to do this
  - SAS calls them Type I, II, III, and IV sum of squares
- We're going to use type III most of the time
- The general idea is to fit the model without the variable included, then with the variable included, and compare their SSR values

Types of sum of squares

- Let's define a function $R()$
- $R(x_1|x_2, x_3)$ is the improvement in the model with x_1, x_2, x_3 vs. x_2, x_3
- Let SSR_1 be the SSR for the reduced model (without x_1)

$$y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

- And let SSR_2 be the SSR for the full model (with x_1)

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

- Then $R(x_1|x_2, x_3) = SSR_2 - SSR_1$. The MSR value is found by dividing $R(x_1|x_2, x_3)$ by the number of coefficients for x_1
- We will divide MSR by the MSE for the *full model* to create an F-statistic.

Type III Sum of Squares

- Type III is *marginal sum of squares*. Order is not important for this method.

Variable	Test Statistic
x_1	$R(x_1 \mid x_2, \ldots, x_k)$
x_2	$R(x_2 \mid x_1, x_3, \ldots, x_k)$
x_3	$R(x_3 \mid x_1, x_2, x_4, \ldots, x_k)$
$\vdots$	$\vdots$
x_k	$R(x_k \mid x_1, x_2, \ldots, x_{k-1})$

- One key difference: for Type I SS the individual $R()$ values will sum to the model SSR value, and they do not for Type III.

Type III Table

Source of Variation	SS	df	Mean Square	F
X_1	$R(X_1 \mid X_2, X_3)$	1	$MSR = \frac{R(X_1 \mid X_2, X_3)}{1}$	$F_1 = \frac{MSR_1}{MSE}$
X_2	$R(X_2 \mid X_1, X_3)$	1	$MSR = \frac{R(X_2 \mid X_1, X_3)}{1}$	$F_2 = \frac{MSR_2}{MSE}$
X_3	$R(X_3 \mid X_1, X_2)$	1	$MSR = \frac{R(X_3 \mid X_1, X_2)}{1}$	$F_3 = \frac{MSR_3}{MSE}$
Error	SSE	$n - k - 1$	$MSE = \frac{SSE}{n-k-1}$	
Total	SST	$n - 1$		

- **This is the new ANOVA table**

4. Test the Regression Coefficients ($\hat{\beta}_i$)

- Transforming the Wald Statistic to work for multiple linear regression
- Uses the same structure, needs the $(X'X)^{-1}$ matrix
- Let x_i be an $n \times 1$ matrix (or a *column vector*) of the values for the i^{th} independent variable
- The X matrix, therefore, is

$$X = [1 \; x_1 \; x_2 \; \ldots \; x_k]$$

Wald Statistics

- The test statistic is for the simple case is

$$t = \frac{\hat{\beta}_1}{\sqrt{MSE/S_{XX}}}$$

- It's tested against the t-distribution with $n - 2$ degrees of freedom
- For the multivariate case we need $S_{X_i X_i}$, which is in C

$$t = \frac{\hat{\beta}_j}{\sqrt{C_{(j+1)(j+1)} MSE}}$$

- And this is tested against $n - k - 1$ degrees of freedom

5. Test the Regression Assumptions

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- The observations are independent of one another
- $\epsilon \sim N(0, \sigma^2)$
- The assumptions and assumption checking are the same as for linear regression

Logistic Regression

1. Develop the Model

Modeling Binary Outcomes

- Let Y be a categorical variable with two levels (binary variable)
- We can't use regular linear regression methods: The assumption that $\epsilon \sim N(0, \sigma^2)$ cannot be satisfied
- Going to arbitrarily assign "success" to one of the two possible outcomes. We'll code the successes as 1 and the failures as 0
- Define p_i as the theoretical probability for a "success"

y_1	Probability
1	$P(y_i = 1) = p_i$
0	$P(y_i = 0) = 1 - p_i$

Linear Transformations

- We need to investigate transformations of the data such that the assumptions are not violated
- The objective is to transform p_i such that it is not bounded
- The **logit transformation** is the most common way to model binary data

$$p_i = \frac{exp(x'\beta)}{1 + exp(x'\beta)}$$

- We can adjust the transformation to obtain a linear model

Creating A Linear Model

$$
\begin{aligned}
p_i &= \frac{exp(x'\beta)}{1 + exp(x'\beta)} \\
p_i(1 + exp(x'\beta)) &= exp(x'\beta) \\
p_i + p_i exp(x'\beta) &= exp(x'\beta) \\
p_i &= exp(x'\beta) - p_i exp(x'\beta) \\
p_i &= (1 - p_i) exp(x'\beta) \\
\frac{p_i}{1 - p_i} &= exp(x'\beta) \\
ln\left[\frac{p_i}{1 - p_i}\right] &= x'\beta
\end{aligned}
$$

The Logit Transformation

$$ln\left[\frac{p_i}{1-p_i}\right] = x'\beta$$

$$logit(p_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$

- The is the of the odds, the log-odds or the *logit* of p_i
- The relationship between $logit(p_i)$ and x is now linear
- p_i is still bound by 0 and 1, but $logit(p_i)$ is unbounded, $logit(p_i) \in [-\infty, \infty]$
- Note that we are modeling p_i rather than y_i

2. Estimate the Coefficients

Maximum Likelihood Estimates

- The probability of observing event y_i is given as follows

$$f(y_i) = p_i^{y_i}(1 - p_i)^{1-y_i}$$

- Since the observations are independent, the cumulative probability can be defined as follows

Likelihood Function

$$L(y_1, y_2, \ldots y_n; \beta) = \prod_{i=1}^{n} f(y_i)$$

- We want this to be as large as possible
- It is more convenient to work with the *log-likelihood function*

Log-Likelihood Function

$$
\begin{aligned}
L(y_1, y_2, \ldots y_n; \beta) &= \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \\
ln\left(L(y_1, y_2, \ldots y_n; \beta)\right) &= ln\left(\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}\right) \\
l(y_1, y_2, \ldots y_n; \beta) &= \sum_{i=1}^{n} y_i ln(p_i) + (1-y_i) ln(1-p_i) \\
l(y_1, y_2, \ldots y_n; \beta) &= \sum y_i \left[ln(p_i) - ln(1-p_i)\right] + ln(1-p_i) \\
l(y_1, y_2, \ldots y_n; \beta) &= \sum y_i \left[ln\left(\frac{p_i}{1-p_i}\right)\right] + \sum ln(1-p_i) \\
l(y_1, y_2, \ldots y_n; \beta) &= \sum y_i x'\beta + \sum ln(1+e^{x'\beta})
\end{aligned}
$$

Log-Likelihood Function

$$l(y_1, y_2, \ldots y_n; \beta) = \sum y_i x' \beta + \sum ln(1 + e^{x' \beta})$$

- We want to find $\hat{\beta}$ that makes this value as large as possible
  - We maximize the function using numeric methods, R uses Iteratively Reweighted Least Squares if you're curious
- Results in maximum likelihood estimates: $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 & \hat{\beta}_1 & \ldots & \hat{\beta}_k \end{bmatrix}'$
- Using these estimates we can estimate p or L

Odds and Odds Ratios

- We can predict the odds of an event directly from the logit equation
  - Recall that logit(p) is the log-odds of the event

$$\hat{o} = e^{\beta' X}$$

- And to compare the odds of two different concentrations, represented by $\hat{o}_A$ and $\hat{o}_B$

$$\begin{aligned} OR &= \frac{\hat{o}_A}{\hat{o}_B} \\ &= \frac{e^{\beta_0+\beta_1 X_A}}{e^{\beta_0+\beta_1 X_B}} \\ &= e^{\beta_0+\beta_1 X_A-(\beta_0+\beta_1 X_B)} \\ &= e^{\beta_1(X_A-X_B)} \end{aligned}$$

Odds and Odds Ratios

- Let $X_A - X_B = 1$. This is true for a binary predictor, or for a unit increase in a continuous predictor
  - What is the increase in odds of tumour growth for a unit increase in toxic concentration?

$$OR = e^{\beta_1}$$

- **This is the key mathematical property of logistic regression**. It states that the effect of predictor x_i is dependent only on the coefficient β_i, and not the other values of X.

OR For Multivariate Logistic Regression

$$
\begin{aligned}
OR &= \frac{\hat{o}_A}{\hat{o}_B} \\
&= \frac{e^{\beta_0+\beta_1 X_A+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4}}{e^{\beta_0+\beta_1 X_B+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4}} \\
&= e^{\beta_0+\beta_1 X_A+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4-(\beta_0+\beta_1 X_B+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4)} \\
&= e^{\beta_1(X_A-X_B)+\beta_2(X_2-X_2)+\beta_3(X_3-X_3)+\beta_4(X_4-X_4)} \\
&= e^{\beta_1(X_A-X_B)}
\end{aligned}
$$

3. Evaluate the Model Fit

- Recall the likelihood equation

$$L(y_1, y_2, \ldots y_n; \beta) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i}$$

- If we have fit the model *perfectly*, then our estimates of p_i will be 1 when $y_i = 1$ and 0 when $y_i = 0$
- This means that the perfect likelihood equation would be 1
  - the perfect log-likelihood equation would be log(1)=0
- An imperfect likelihood equation would be somewhere in the range $[0, 1]$
- This would put an imperfect log-likelihood in the range $[-\infty, 0]$

Null Model vs. the Full Model

- Recall the $p_i = exp(x\beta)/(1 + exp(x\beta))$
- The null model is the same idea as linear regression: is a model with no predictors better than a model with the current set of predictors
- Null Model:

$$p_i = \frac{exp(\beta_0)}{1 + exp(\beta_0)}$$

- Observed Model:

$$p_i = \frac{exp(\beta_0 + \beta_1 x_1 + \ldots)}{1 + exp(\beta_0 + \beta_1 x_1 + \ldots)}$$

Model Deviance

- We will evaluate the fit of our model by calculating the difference between the null model and the observed model

$$\begin{aligned} R(\hat{\beta}) &= 2\left[l(\textit{null model}) - l(\hat{\beta})\right] \\ &= 2\left[l(\beta_0) - l(\hat{\beta})\right] \end{aligned}$$

- Called the **likelihood ratio test**
- The *-2 log-likelihood* is a measure of the fit of the model
- Is tested against the chi-square distribution with k degrees of freedom
- If the difference between the two models is greater than $\chi^2_{1-\alpha,k}$, we can conclude that the fitted model is better

4. Test the Regression Coefficients

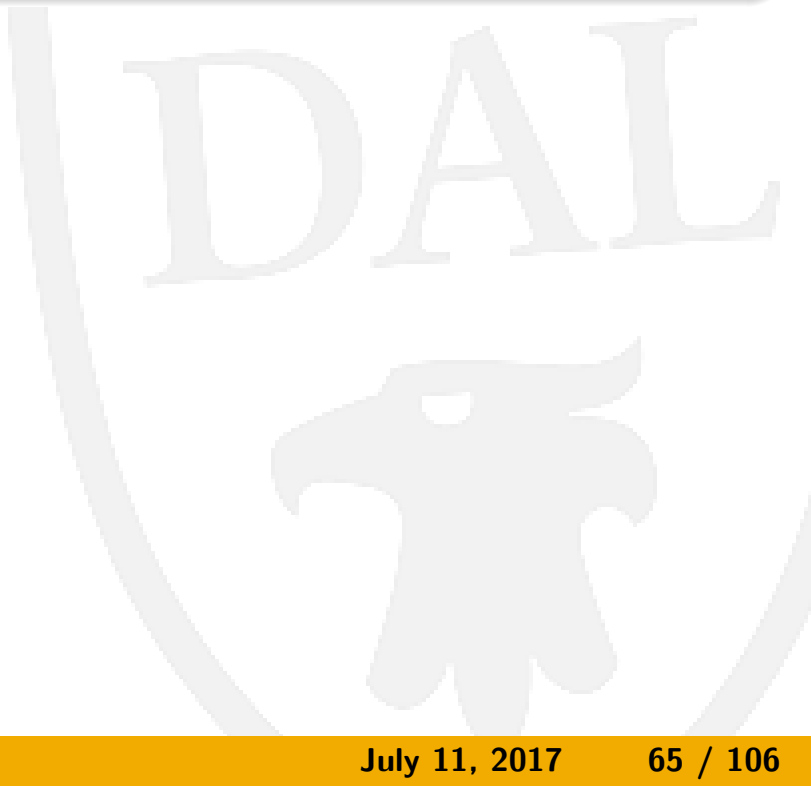
- As with linear regression, we can test the significance of the individual regression coefficients
- Without going into the details, we can assume that the regression coefficients are approximately normally distributed
- Their standard errors can be derived from the second derivative of the log-likelihood equation
- The result is a regression table similar to what we had in multiple linear regression

	Estimate	SE	z-value	p-value
β_0	$\hat{\beta}_0$	SE_0	z_0	$2P(\lvert z_0\rvert > Z)$
β_1	$\hat{\beta}_1$	SE_1	z_1	$2P(\lvert z_1\rvert > Z)$

5. Test the Model Assumptions

- There are far fewer assumptions for logistic regression that there are for linear regression, and little checking is normally done
- You need to make sure that your observations are independent
- You should still check for outliers, mostly amongst your predictor variables
- Your outcome variable needs a good balance of both events and non-events: if you don't have between 10% and 90% events then you will probably run into modeling problems
- You will see/hear discussion of the linearity assumption between continuous predictors and the log-odds of the outcome, and something called a Box-Tidwell test. Don't worry too much about that, I've never had to use it

Survival Analysis



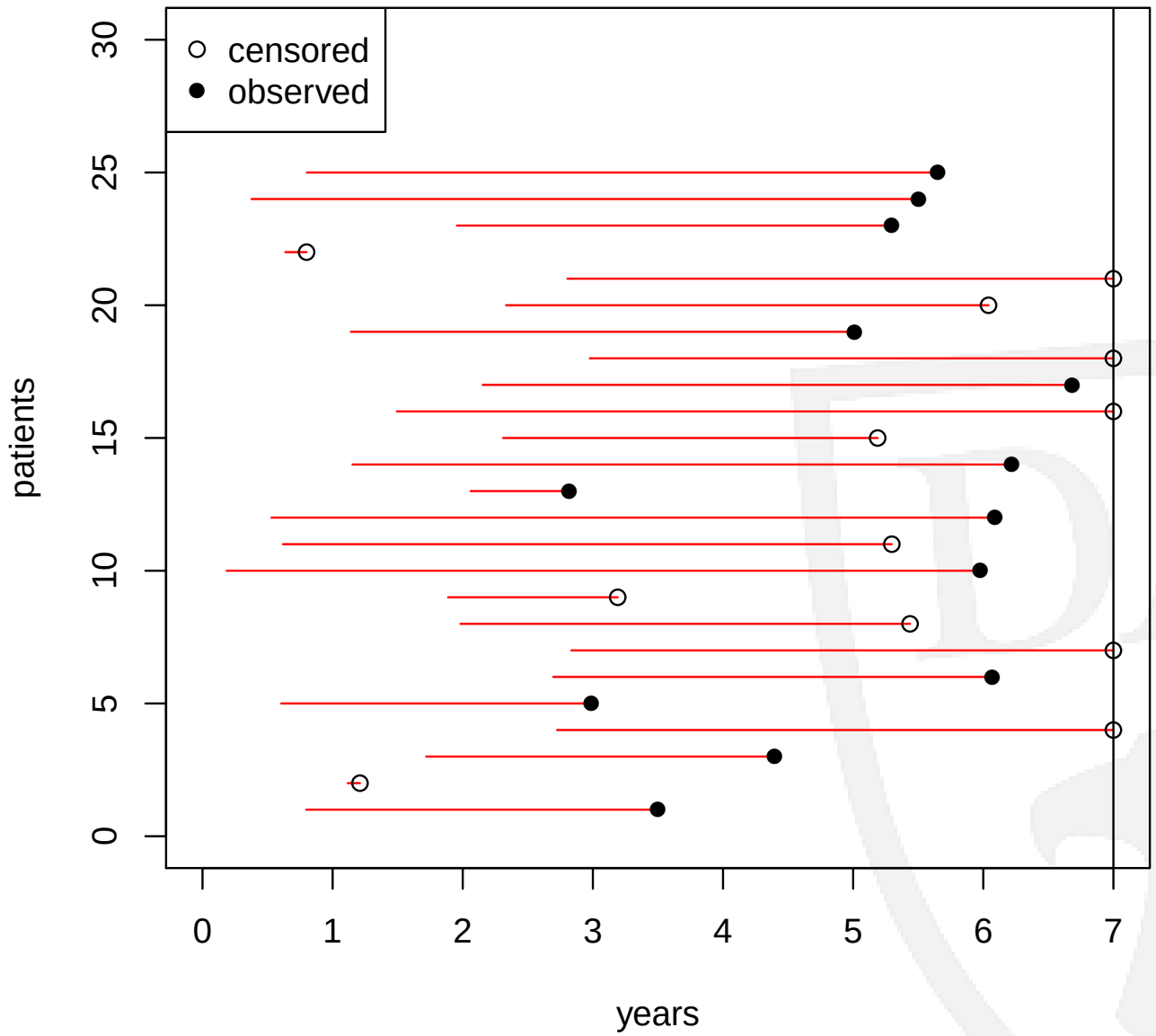
Survival Analysis

- Survival analysis is the analysis of time-to-event data
- We are concerned with measuring the length of time between some initial event and some event of interest (typically death)
  - Time between surgery and infection
- The key difference between survival data and regular data is the incorporation of *censored* observations
  - Not all surgery patients experience infection

Data

- There are three data values required for each subject in a survival data set:
  - Starting point (time/date)
  - Ending point (time/date)
  - Presence of the event (1/0)
- For a regression model we also need some patient characteristics to regress on

Survival Data



1. Develop the Model

Cox Proportional Hazards Regression

$$h(t_i) = h_0(t_i)exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon)$$

$$log\left(\frac{h(t_i)}{h_0(t_i)}\right) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$

- The model is dependent on the idea of the *hazard function*.
- The hazard function, $h(t_i)$, is the probability of experiencing the event at time point t_i, given that you haven't already experienced it.
- It is often referred to as the *instantaneous failure rate*.
- The rate itself is not of particular interest, but comparing rates between two individuals produces interesting results.

2. Estimate the Coefficients

- $h_0(t_i)$ is a baseline hazard rate that we're not interested in
- β_k represents the change in the hazard due to variable k
- We estimate the regression coefficients using maximum likelihood estimates
  - I won't go into the details here, the wikipedia page summarizes it well. Censoring makes it complex.

3\4. Estimate the Model Fit\Test the Regression Coefficients

- We can estimate the fit of the model using a likelihood ratio test as we did with logistic regression
- We get estimates and standard errors for our regression coefficients, allowing us to do Wald tests and confidence intervals
- Since Cox PH use Maximum Likelihood Estimation (as logistic regression does) the methods are largely the same

5. Test the Regression Assumptions

Proportionality Assumption

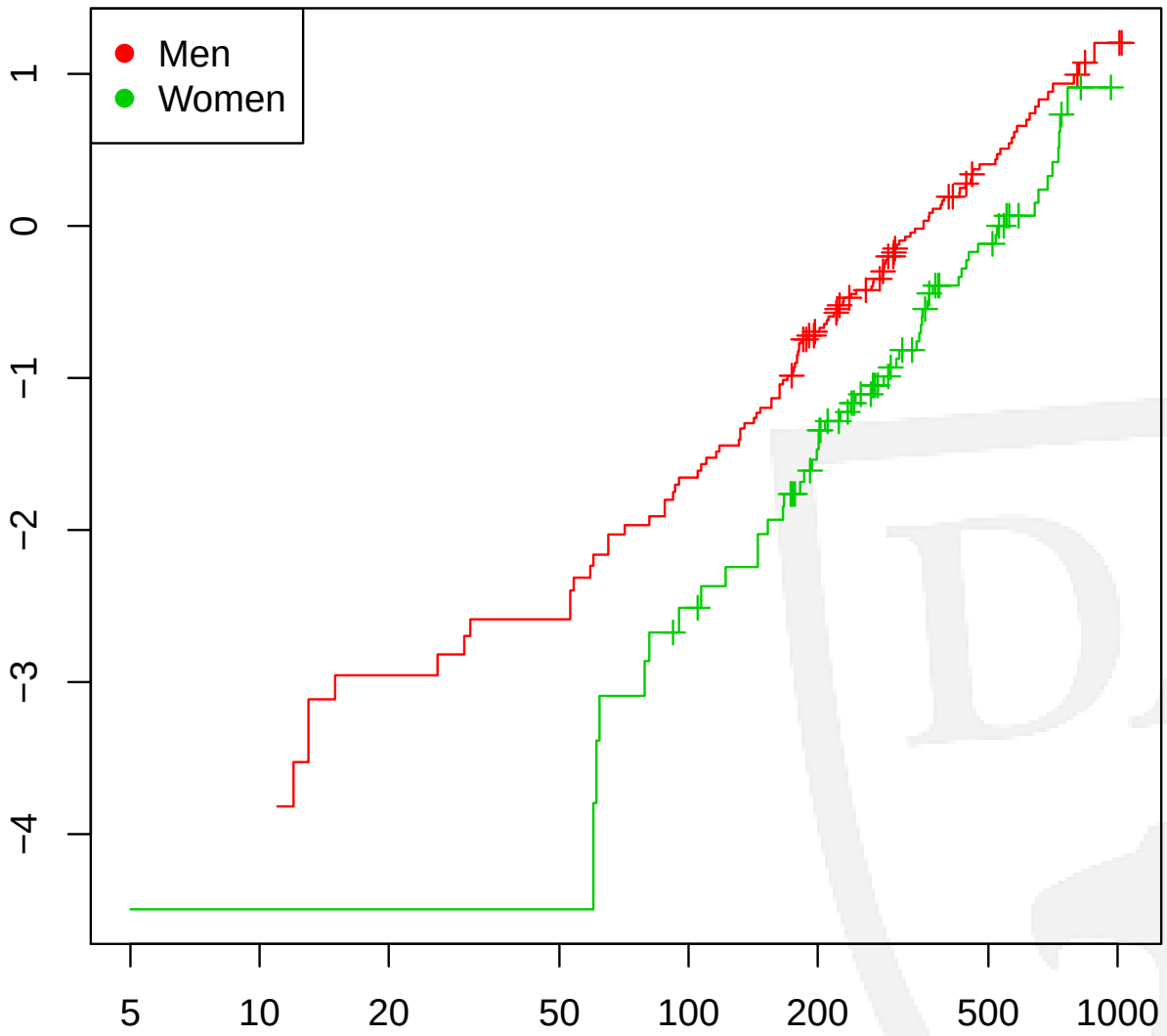
- As mentioned before, there is an assumption of proportionality of the hazard rates for Cox PH regression
- This needs to be checked to ensure the regression results are accurate
- The simplest way to check is to perform a simple analysis of the individual variables using KM curves
- We perform what is called a *log-log plot*, in which we transform the axes such that the lines on the KM plot should be parallel
  - Is based on the idea that the your chance of survival is a cumulative function of the hazard rates

Proportionality Assumption

$$
\begin{aligned}
S(t) &= exp\left(-\int_0^t h_0(u)du\right)^{exp(X'\beta)} \\
log(S(t)) &= \left(-\int_0^t h_0(u)du\right) exp(X'\beta) \\
-log(S(t)) &= \left(\int_0^t h_0(u)du\right) exp(X'\beta) \\
log(-log(S(t))) &= log\left(\int_0^t h_0(u)du\right) + exp(X'\beta)
\end{aligned}
$$

- This is a log-function of t
- Plotting $log(-log(S(t)))$ vs $log(t)$ should result in a linear relationship

Proportionality Example



Modeling

- We looked at three types of models: linear, logistic, survival
- For all three we go through the same process
  1. Develop the model
  2. Estimate the coefficients
  3. Evaluate the model fit
  4. Test the regression coefficients
  5. Test the regression assumptions
- We're now going to look at Poisson regression, then move onto the more complex longitudinal models

General Linear Modeling (GLM)

$$g(E(y)) = g(\mu) = X\beta + \epsilon$$

- $g(\mu)$ is some transformation of the expected value (i.e the mean) of the outcome such that the relationship is linear
- We have looked at three different types of modelling

Linear	regression is used when y is a continuous, somewhat normally distributed variable
Logistic	regression is used when Y is binary
Survival	regression is used when Y is a time to event

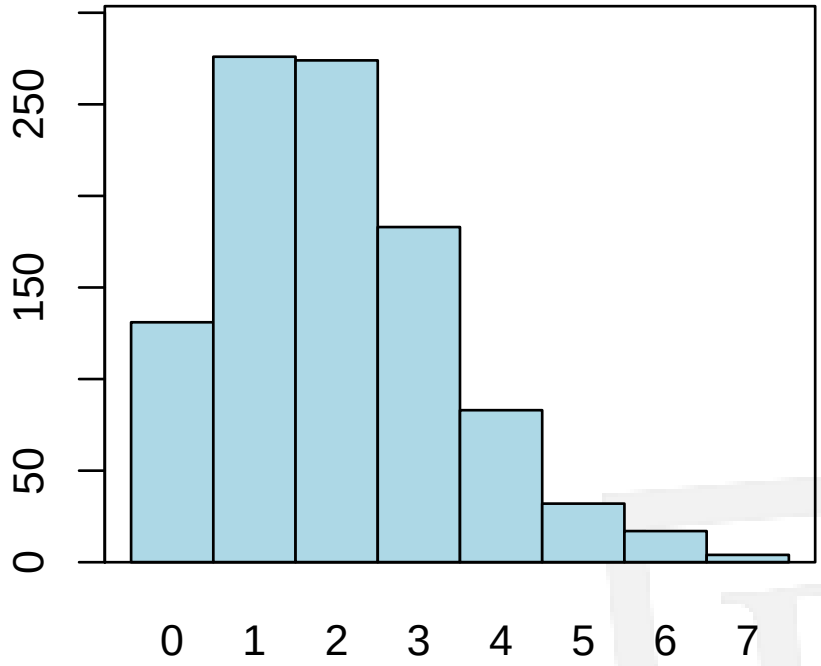
- There are many other forms of regression
- The choice of regression is dependent on the nature of the *outcome* variable

Poisson Regression

1. Develop the Model

- Poisson data typically represents counting data
- Poisson data is "memoryless", i.e., the probability of $y = k$ is independent of the probability of $y = k - 1$
  - If we're modeling the number of people waiting in the ER, the probability of someone visiting the ER is not dependent on the number of people in the ER
- The "exposure" variable is unique to Poisson regression

Poisson Data



$$P(y) = \frac{\lambda^y e^{-\lambda}}{y!} \qquad E(y) = \lambda \qquad V(y) = \lambda$$

$$log(\lambda) = X\beta + \epsilon + log(exposure) \qquad \lambda = e^{X\beta} \times (exp)$$

Exposure and Rates

- When we're modeling counts we're really modeling rates, as it is the number of events per some unit of time, or of some geographic area
  - Number of trees per forest
  - Number of deaths per county
  - Number of ER visits during a shift
- The exposure is often referred to as the **offset**
- If your data always has the same unit of time/area then you can just ignore the offset (or technically set it to 1)

Mathematical Support for Offset

We want to model the rate, y/t instead of the count y

$$
\begin{aligned}
log(\lambda/t) &= X\beta + \epsilon \\
log(\lambda) - log(t) &= X\beta + \epsilon \\
log(\lambda) &= X\beta + \epsilon + log(t)
\end{aligned}
$$

2. Estimate the Coefficients

- Poisson regression, like logistic regression, is a subclass of the *general linear models*

$$g(\lambda) = X\beta + \epsilon + log(exposure)$$

- $g(\lambda)$ is called the **link function**: it is how we transform the outcome so that the model is linear
  - For Poisson regression the link is the log function
  - For logistic regression the link is the logit function
- All GLMs are fit using Maximum Likelihood Estimation to get the coefficient estimates

Maximum Likelihood Estimates

Recap From Before

- The probability of observing event y_i is given as follows

$$f(y_i) = \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}$$

- Since the observations are independent, the cumulative probability can be defined as follows

Likelihood Function

$$L(y_1, y_2, \ldots y_n; \beta) = \prod_{i=1}^{n} f(y_i)$$

- We want this to be as large as possible
- As before we solve this using the log-transform and numeric methods (see here for the derivation)

3. Evaluate the Model Fit

- For all general linear models the model fit is evaluated using the likelihood ratio test
- Recall that we define the model deviance as $-2\times$log-likelihood
- The difference between the model deviance and null deviance is chi-square distributed

4. Test the Regression Coefficients

- The variables can be tested using the likelihood ratios (Type III testing)
- The coefficients can be evaluated using Wald Statistics
- Poisson regression produces *Risk Ratios* (or relative risks, or hazard ratios, they all mean the same thing)
  - This is in contrast to logistic regression producing Odds Ratios

Proof of RR

$$
\begin{aligned}
log(\lambda_1) &= X_1\beta + log(t_1) \\
\lambda_1 &= e^{X_1\beta} t_1 \\
\frac{\lambda_1}{\lambda_2} &= \frac{e^{X_1\beta} t_1}{e^{X_2\beta} t_2} \\
\frac{\lambda_1}{\lambda_2} &= \frac{e^{X_1\beta}}{e^{X_2\beta}} \frac{t_1}{t_2} \\
\frac{\lambda_1}{\lambda_2} \frac{t_2}{t_1} &= e^{X_1\beta - X_2\beta} \\
\frac{\lambda_1 / t_1}{\lambda_2 / t_2} &= e^{(X_1 - X_2)\beta} \\
RR &= e^{(X_1 - X_2)\beta}
\end{aligned}
$$

5. Test the Regression Assumptions

- The most important assumption in Poisson regression is that *the variance must equal the mean*
- If the variance estimates of the data are found to be larger than the mean then the data is over-dispersed
  - The residual deviance should be approximately the same as its degrees of freedom (DF), otherwise we might have overdispersion
  - This can sometimes be fixed by using *quasi-Poisson regression* to estimate the exposure variable
  - Negative binomial regression can also provide an alternative regression approach

Other Poisson Regression Assumptions

- Another problem that arises with Poisson regression is excess zeros: this often occurs when your data is a combination of binomial (Yes/No) and count data
  - Number of cigarettes smoked, or some disease severity score like CDAI for IBD patients
  - Can be addressed using a method called zero-inflated Poisson regression
- The regular issues of independence and outliers also need to be checked

Quasi-Poisson Regression

$$E(Y) = \lambda \qquad V(Y) = \phi\lambda$$

- For cases where the variance and the mean of the distribution are not equal
- We change our assumption about the model *slightly*, and estimate the scale parameter, ϕ
- The coefficient estimates don't change (since our likelihood equation doesn't change)
- The way we estimate the variance and standard errors does
- The models are built and evaluated the same

Negative Binomial Distribution

- The binomial distribution is defined as the number of successes when trying the same experiment n times if each trial has p probability of success
- The negative binomial distribution is the number of successes of a trial before k failures, if each trial has probability p
  - The binomial answers "what's the probability of getting 3 heads when I flip a coin 10 times?"
  - The negative binomial answers "what's the probability of getting 3 heads before 3 tails?"
- It is like the Poisson distribution in that it "counts" events
- It is difference in that, rather than defining the scope by *exposure* it defines it compared to *failures*

Negative Binomial Distribution and Poisson

$$E(Y) = \left(\frac{p}{1-p}\right)k \qquad V(Y) = \frac{pk}{(1-p)^2}$$

- If we let $\lambda = \frac{p}{1-p}k$ then we get the following

$$E(Y) = \lambda \qquad V(Y) = \lambda + \frac{\lambda^2}{k}$$

- NB has same mean as the Poisson
- NB has a variance that is the Poisson variance+a function of k, which we call *the dispersion parameter*
- As k gets very big (or the number of failures we allow gets very large) then the NB approximates the Poisson

Negative Binomial Regression

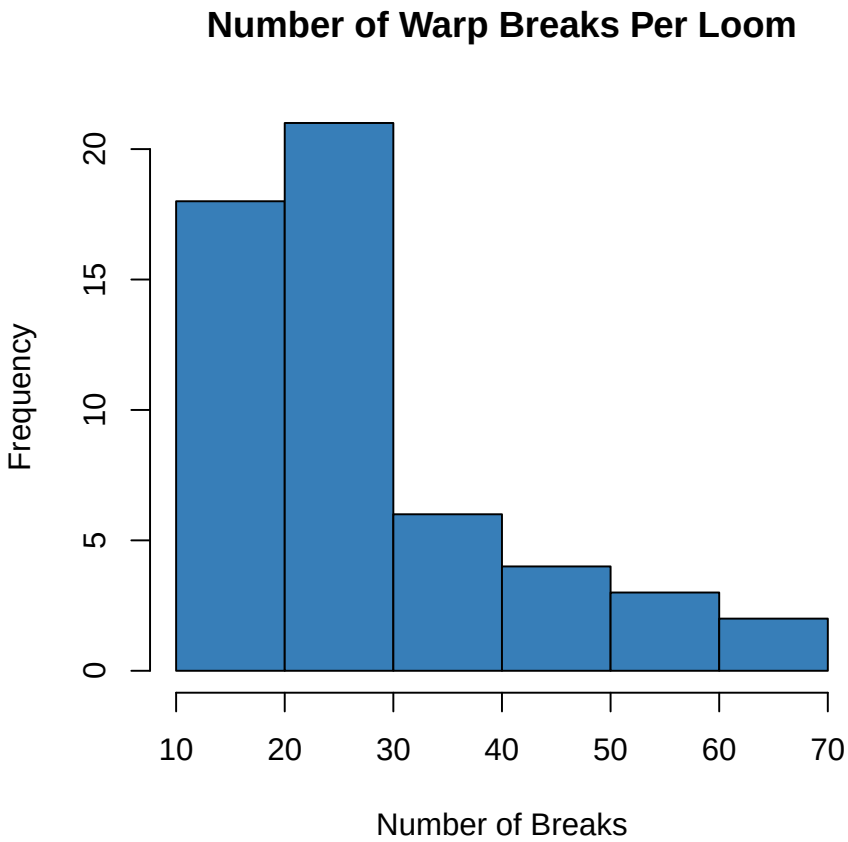
1. Develop the Model: $log(\mu) = X\beta + \epsilon + log(exposure)$
   - Same link function as Poisson, same process for including the exposure
2. Estimate the coefficients: It's another MLE, this page provides a detailed breakdown of the estimations if you're interested
3. Evaluate the Model Fit: Likelihood Ratio Test
4. Test the Regression Coefficients: Type III tests/Wald Statistics
   - NB Regression produces *Incidence Rate Ratios*, which are much like Hazard Ratios, which are much like Relative Risks. The difference between these is better left to the epidemiologists in the department

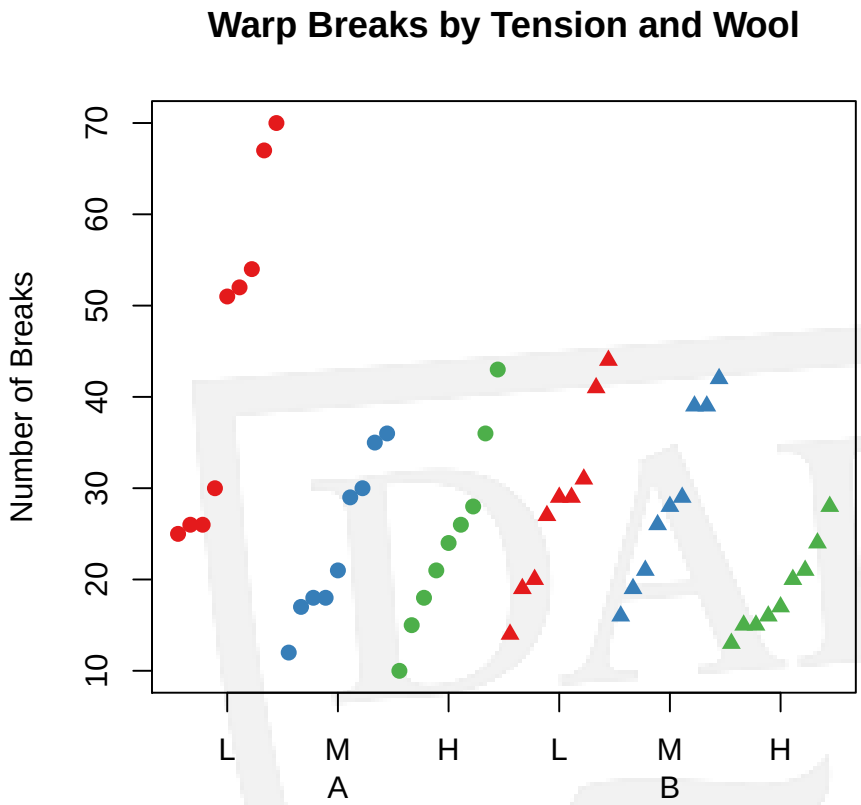
5. Test the Regression Assumptions

- Independence and outliers
- Appropriateness of the NB
  - It's harder to justify the NB theoretically
  - The over-dispersion of the Poisson model might be due to zero-inflation, in which case NB isn't a good solution

Poisson Regression Example

- This is a dataset from the R library called `warpbreaks`, counting the number of "warpbreaks" on a loom based on the tension (Low/Medium/High) and the type of wool (A/B)
- We'll explore the data, then model it, then check our assumptions


Number of Warp Breaks Per Loom
Frequency
0
5
10
15
20
10
20
30
40
50
60
70
Number of Breaks


Warp Breaks by Tension and Wool
Number of Breaks
10
20
30
40
50
60
70
L
M
H
L
M
H
A
B

Model

Model	Deviance	DF	p-value
Null	297.37	53	
Residual	210.39	50	
Difference	86.98	3	< 0.0001

	LR	DF	p-value
wool	16.04	1	0.0001
tension	70.94	2	< 0.0001

	Estimate	Std. Error	z value	p-value
(Intercept)	3.6920	0.0454	81.30	< 0.0001
woolB	-0.2060	0.0516	-3.99	0.0001
tensionM	-0.3213	0.0603	-5.33	< 0.0001
tensionH	-0.5185	0.0640	-8.11	< 0.0001

Problem with Over-Dispersion?

- It looks like the scale parameter in this case is $\frac{210}{50} = 4.2$, or much different from 1
- We'll investigate a quasi-Poisson distribution instead
- We'll also look at the negative binomial results

Quasi-Poisson Results

Model	Deviance	DF	p-value
Null	297.37	53	
Residual	210.39	50	
Difference	86.98	3	< 0.0001

	LR Chisq	Df	Pr(>Chisq)
wool	3.76	1	0.0524
tension	16.65	2	0.0002

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.6920	0.0937	39.38	0.0000
woolB	-0.2060	0.1065	-1.93	0.0587
tensionM	-0.3213	0.1244	-2.58	0.0128
tensionH	-0.5185	0.1320	-3.93	0.0003

Negative Binomial Results

Model	Deviance	DF	p-value
Null	75.46	53	
Residual	53.7	50	
Difference	21.7	3	< 0.0001

	LR Chisq	Df	Pr(>Chisq)
wool	3.37	1	0.0665
tension	17.54	2	0.0002

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.6734	0.0979	37.52	0.0000
woolB	-0.1862	0.1010	-1.84	0.0651
tensionM	-0.2992	0.1217	-2.46	0.0140
tensionH	-0.5114	0.1237	-4.13	0.0000

Comparing Estimates

	Poisson	Quasi-Poisson	Negative-Binomial
wool: B vs A	0.81 [0.74, 0.9]	0.81 [0.66, 1]	0.83 [0.68, 1.01]
tension: M vs L	0.73 [0.64, 0.82]	0.73 [0.57, 0.93]	0.74 [0.58, 0.94]
tension: H vs L	0.6 [0.53, 0.67]	0.6 [0.46, 0.77]	0.6 [0.47, 0.76]

- Effect estimates are roughly the same
- Quasi-Poisson and NB are roughly the same
- Poisson regresison makes mistake of suggesting more confidence in the effect than exists
  - This is the main RISK of poisson regression, underestimates of variance

R Code


```
1  library(MASS)
2  data("warpbreaks")
3  dat = warpbreaks
4
5  mod01 =  glm(breaks~wool+tension,data=dat,family=
       poisson)
6  mod01a = glm(breaks~wool*tension,data=dat,family=
       poisson)
7  mod02 =  glm(breaks~wool+tension,data=dat,family=
       quasipoisson)
8  mod02a = glm(breaks~wool*tension,data=dat,family=
       quasipoisson)
9  mod03 =  glm.nb(breaks~wool+tension,data=dat)
10 mod03a = glm.nb(breaks~wool*tension,data=dat)
```

Regression Methods in STATA

- `poisson` is the command to conduct a Poisson regression
  - `glm` can also fit the Poisson regression
- Either can be used to fit the quasi model
  - In `poisson` we use the option `vce(robust)`
  - In `glm` we using the option `scale(x2)` to do the same thing
- `nbreg` is the command to fit a negative binomial regression
  - `glm` doesn't work with negative binomial regression because the function doesn't estimate the k part of the model correctly

Building a Model in STATA


```
1  import delimited "C:\Users\sstewar2\Documents\Teaching
       \Grad Students\RegressionMethodsCHE\warpbreaks.csv
       "
2
3  //convert the strings to factors
4  encode wool, generate(woolFactor)
5  encode tension, generate(tensionFactor)
6
7  //exploring the data
8  sum breaks
9  tab wool tension
10 tab wool, sum(breaks)
11 tab tension, sum(breaks)
12 tab wool tension, sum(breaks)
```

Building a Model in STATA

```
1 //poisson modeling
2 poisson breaks i.woolFactor ib2.tensionFactor, irr
3 glm breaks i.woolFactor ib2.tensionFactor, family(
    poisson) eform
4 //adapted poisson
5 poisson breaks i.woolFactor ib2.tensionFactor, irr vce
    (robust)
6 glm breaks i.woolFactor ib2.tensionFactor, family(
    poisson) eform scale(x2)
7 //negative binomial
8 nbreg breaks i.woolFactor ib2.tensionFactor, irr
```

Another Example

https://stats.idre.ucla.edu/stata/dae/poisson-regression/

- Awards earned by students at one high school
- Predictors: type of program (vocational, general or academic) and final math score

Try the Analysis Yourselves


```
1 //read the data in
2 use https://stats.idre.ucla.edu/stat/stata/dae/
    poisson_sim, clear
```

- Perform basic exploration on the data (to better understand the variables)
- Build a simple Poisson regression model predicting the number of awards a student might win
  - You choose the predictors that you think are important