Advanced Statistics
DS2003 (BDS-4A)
Lecture 25

Instructor: Dr. Syed Mohammad Irteza

Assistant Professor, Department of Computer Science, FAST

24 May, 2022

Previous Lecture

- Interpretation of regression analysis (example: child's test score, i.e., the response variable, modeled with the mother's age, IQ, work, etc., i.e., the predictor variables)
- An alternative way to calculate ***R^2*** = (explained var. in Y/total var. in Y)
- Calculating ***Adjusted R^2*** (a better choice than R^2 for deciding if adding a new predictor variable is improving our model or not)
- ***Model Selection***: Professor Evaluation (response variable) and their individual features (beauty, gender, age, minority, tenure track, etc.)
  - Interpreting the p-values (what is the null hypothesis for each predictor?)
  - Backward Elimination (using p-values or using adjusted R^2)
  - Forward Selection (using p-values or using adjusted R^2)

Checking model conditions using graphs

Modeling conditions

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

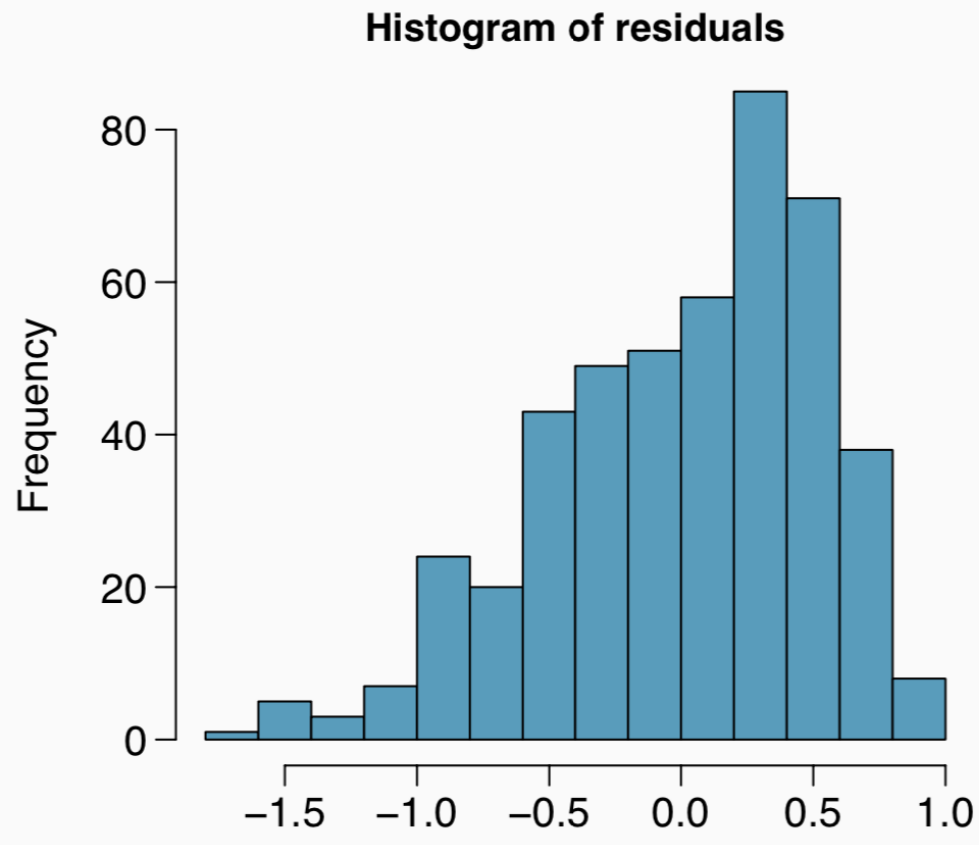
The model depends on the following conditions

1. residuals are nearly normal (less important for larger data sets)
2. residuals have constant variability
3. residuals are independent
4. each variable is linearly related to the outcome

We often use graphical methods to check the validity of these conditions, which we will go through in detail in the following slides.

(1) nearly normal residuals

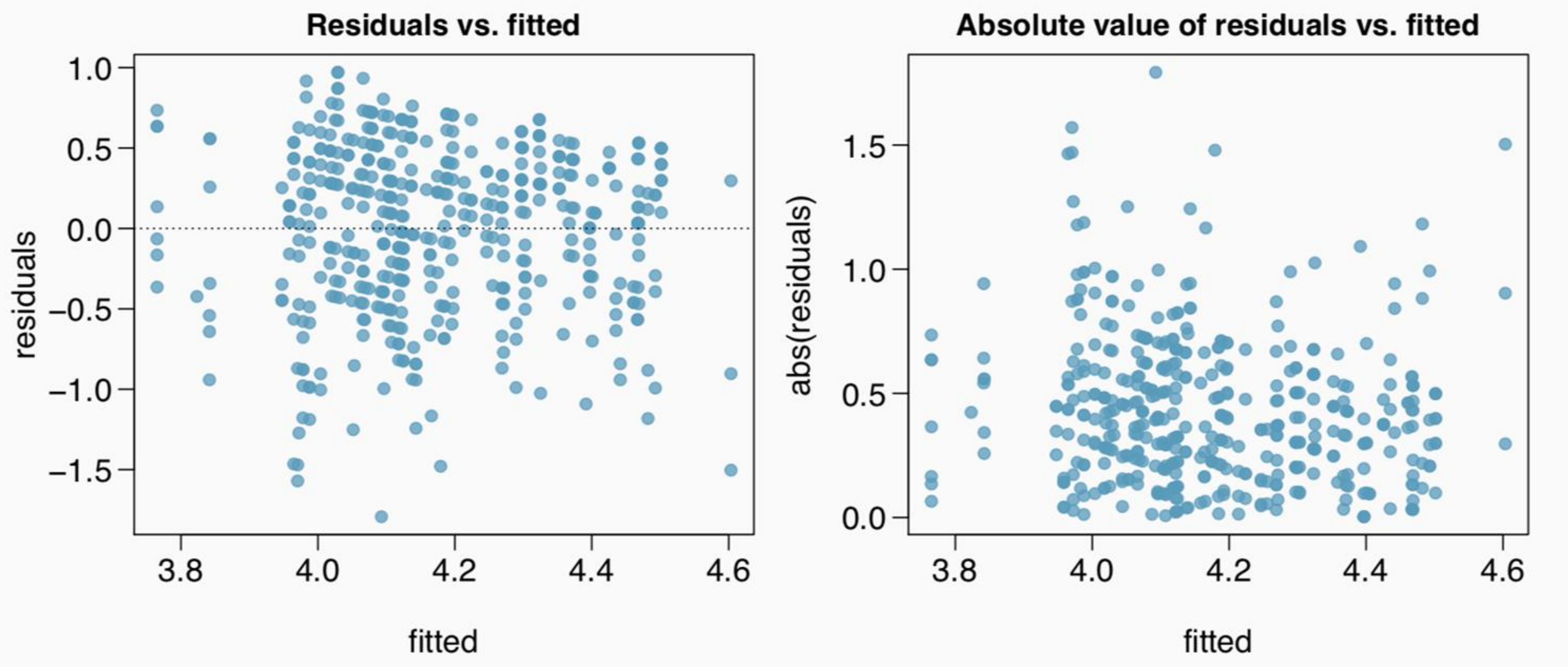
Histogram of the residuals.



Does this condition appear to be satisfied?

(2) constant variability in residuals

Scatterplot of residuals and/or absolute value of residuals vs. fitted (predicted).



Does this condition appear to be satisfied?

Checking constant variance - recap

When we did simple linear regression (one explanatory variable) we checked the constant variance condition using a plot of *residuals vs. x*.

With multiple linear regression (2+ explanatory variables) we checked the constant variance condition using a plot of *residuals vs. fitted*.

Why are we using different plots?

Checking constant variance - recap

When we did simple linear regression (one explanatory variable) we checked the constant variance condition using a plot of *residuals vs. x*.

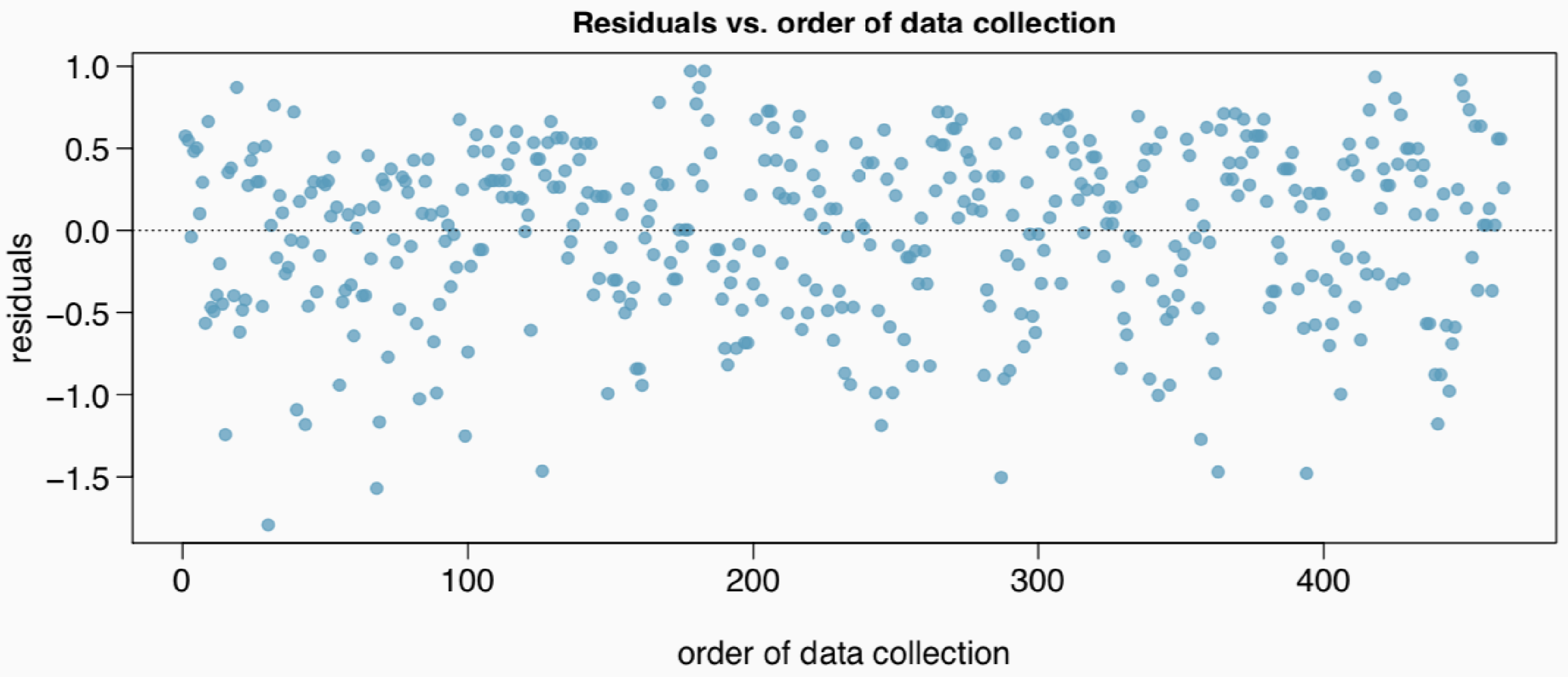
With multiple linear regression (2+ explanatory variables) we checked the constant variance condition using a plot of *residuals vs. fitted*.

Why are we using different plots?

In multiple linear regression there are many explanatory variables, so a plot of residuals vs. one of them wouldn't give us the complete picture.

(3) independent residuals

Scatterplot of residuals vs. order of data collection.



Does this condition appear to be satisfied?

More on the condition of independent residuals

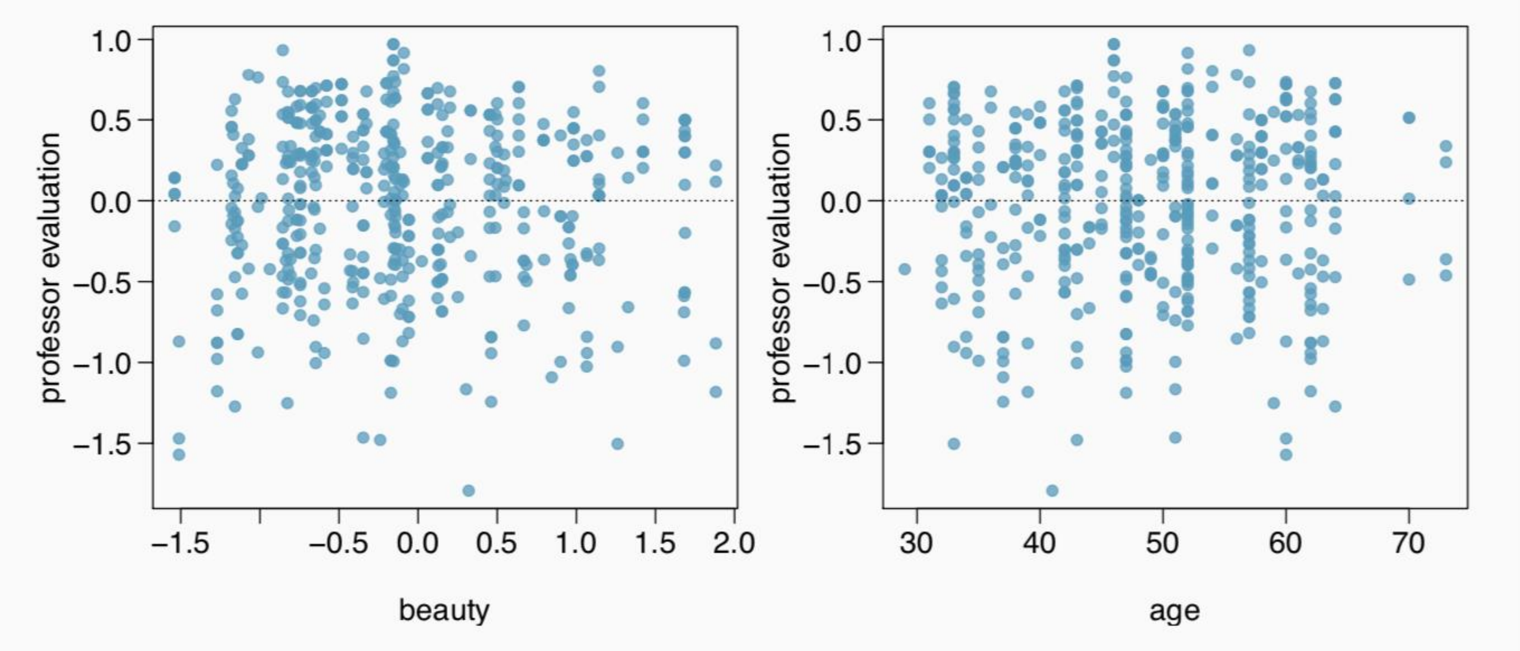
Checking for independent residuals allows us to indirectly check for independent observations.

If observations and residuals are independent, we would not expect to see an increasing or decreasing trend in the scatterplot of residuals vs. order of data collection.

This condition is often violated when we have time series data. Such data require more advanced time series regression techniques for proper analysis.

(4) linear relationships

Scatterplot of residuals vs. each (numerical) explanatory variable.



Does this condition appear to be satisfied?

Note: We use residuals instead of the predictors on the y-axis so that we can still check for linearity without worrying about other possible violations like collinearity between the predictors.

Several options for improving a model

Transforming variables

Seeking out additional variables to fill model gaps

Using more advanced methods that would account for challenges around inconsistent variability or nonlinear relationships between predictors and the outcome

Transformations

If the concern with the model is non-linear relationships between the explanatory variable(s) and the response variable, transforming the response variable can be helpful.

- Log transformation (log y)
- Square root transformation (sqrt(y))
- Inverse transformation (1/y)
- Truncation (cap the max value possible)

It is also possible to apply transformations to the explanatory variable(s), however such transformations tend to make the model coefficients even harder to interpret.

Models can be wrong, but useful

All models are wrong, but some are useful.

- George Box

No model is perfect, but even imperfect models can be useful, as long as we are clear and report the model's shortcomings.

If conditions are grossly violated, we should not report the model results, but instead consider a new model, even if it means learning more statistical methods or hiring someone who can help.

Logistic regression

Regression so far ...

At this point we have covered:

Regression so far ...

At this point we have covered:

Simple linear regression

- Relationship between numerical response and a numerical or categorical predictor

Regression so far ...

At this point we have covered:

Simple linear regression

- Relationship between numerical response and a numerical or categorical predictor

Multiple regression

- Relationship between numerical response and multiple numerical and/or categorical predictors

Regression so far ...

At this point we have covered:

Simple linear regression

- Relationship between numerical response and a numerical or categorical predictor

Multiple regression

- Relationship between numerical response and multiple numerical and/or categorical predictors

What we haven't seen is what to do when the predictors are weird (nonlinear, complicated dependence structure, etc.) or when the response is weird (categorical, count data, etc.)

Odds

Odds are another way of quantifying the probability of an event, commonly used in gambling (and logistic regression).

For some event E,

$$\text{odds}(E) = \frac{P(E)}{P(E^c)} = \frac{P(E)}{1 - P(E)}$$

Similarly, if we are told the odds of E are x to y then

$$\text{odds}(E) = \frac{x}{y} = \frac{x/(x+y)}{y/(x+y)}$$

which implies

$$P(E) = x/(x+y), \quad P(E^c) = y/(x+y)$$

Example - Donner Party

In 1846 the Donner and Reed families left Springfield, Illinois, for California by covered wagon. In July, the Donner Party, as it became known, reached Fort Bridger, Wyoming. There its leaders decided to attempt a new and untested route to the Sacramento Valley. Having reached its full size of 87 people and 20 wagons, the party was delayed by a difficult crossing of the Wasatch Range and again in the crossing of the desert west of the Great Salt Lake. The group became stranded in the eastern Sierra Nevada mountains when the region was hit by heavy snows in late October. By the time the last survivor was rescued on April 21, 1847, 40 of the 87 members had died from famine and exposure to extreme cold.

From *Ramsey, F.L. and Schafer, D.W. (2002). The Statistical Sleuth: A Course in Methods of Data Analysis (2nd ed)*

Example - Donner Party - Data

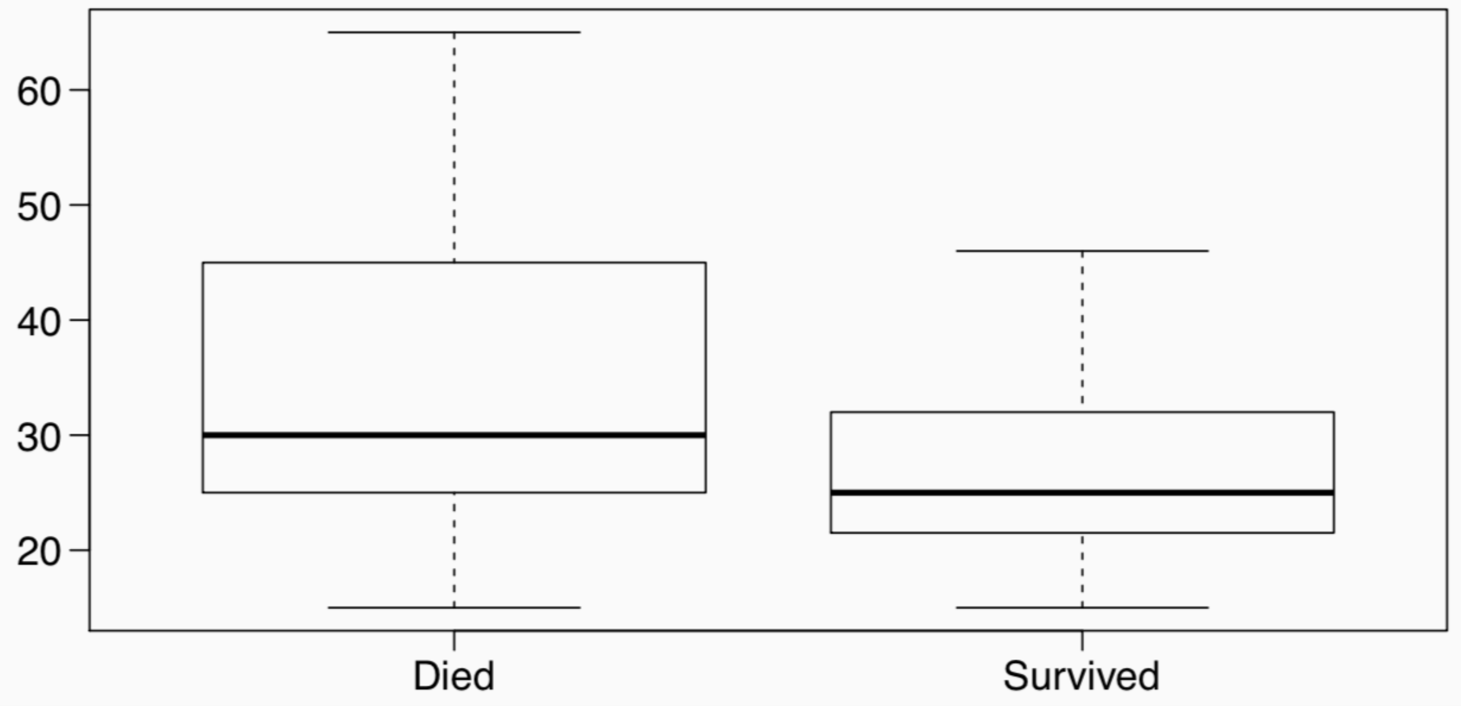
	Age	Sex	Status
1	23.00	Male	Died
2	40.00	Female	Survived
3	40.00	Male	Survived
4	30.00	Male	Died
5	28.00	Male	Died
⋮	⋮	⋮	⋮
43	23.00	Male	Survived
44	24.00	Male	Died
45	25.00	Female	Survived

Example - Donner Party - EDA

Status vs Gender

	Male	Female
Died	20	5
Survived	10	10

Status vs Age



Example - Donner Party

It seems clear that both age and gender have an effect on someone's survival, how do we come up with a model that will let us explore this relationship?

Even if we set Died to 0 and Survived to 1, this isn't something we can transform our way out of - we need something more.

One way to think about the problem - we can treat Survived and Died as successes and failures arising from a binomial distribution where the probability of a success is given by a transformation of a linear model of the predictors.

Generalized linear models

It turns out that this is a very general way of addressing this type of problem in regression, and the resulting models are called generalized linear models (GLMs). Logistic regression is just one example of this type of model.

Generalized linear models

It turns out that this is a very general way of addressing this type of problem in regression, and the resulting models are called generalized linear models (GLMs). Logistic regression is just one example of this type of model.

All generalized linear models have the following three characteristics:

1. A probability distribution describing the outcome variable
2. A linear model: $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$.
3. A link function that relates the linear model to the parameter of the outcome distribution: $g(p) = \eta$ or $p = g^{-1}(\eta)$.

Logistic Regression

Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical predictors.

We assume a binomial distribution produced the outcome variable and we therefore want to model p the probability of success for a given set of predictors.

Logistic Regression

Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical predictors.

We assume a binomial distribution produced the outcome variable and we therefore want to model p the probability of success for a given set of predictors.

To finish specifying the Logistic model we just need to establish a reasonable link function that connects η to p. There are a variety of options but the most commonly used is the logit function.

Logit function

$$logit(p) = \log\left(\frac{p}{1-p}\right), \text{ for } 0 \le p \le 1$$

Properties of the Logit

The logit function takes a value between 0 and 1 and maps it to a value between $-\infty$ and ∞.

Inverse logit (logistic) function

$$g^{-1}(x) = \frac{\exp(x)}{1 + \exp(x)} = \frac{1}{1 + \exp(-x)}$$

The inverse logit function takes a value between $-\infty$ and ∞ and maps it to a value between 0 and 1.

This formulation also has some use when it comes to interpreting the model as logit can be interpreted as the log odds of a success, more on this later.

The logistic regression model

The three GLM criteria give us:

$$y_i \sim \text{Binom}(p_i)$$

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

$$\text{logit}(p) = \eta$$

From which we arrive at,

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i})}{1 + \exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i})}$$

Example - Donner Party - Model

In **R** we fit a GLM in the same was as a linear model except using **glm** instead of **lm** and we must also specify the type of GLM to fit using the **family** argument.

```
summary(glm(Status ~ Age, data=donner,
family=binomial))
## Call:
## glm(formula = Status  ˜Age, family = binomial, data
= donner)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.81852    0.99937   1.820   0.0688 .
## Age         -0.06647    0.03222  -2.063   0.0391 *
##
```

Example - Donner Party - Prediction

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a newborn (Age=0):

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 0$$

$$\frac{p}{1-p} = \exp(1.8185) = 6.16$$

$$p = 6.16/7.16 = 0.86$$

Example - Donner Party - Prediction

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 25$$

$$\frac{p}{1-p} = \exp(0.156) = 1.17$$

$$p = 1.17/2.17 = 0.539$$

Odds / Probability of survival for a 50 year old:

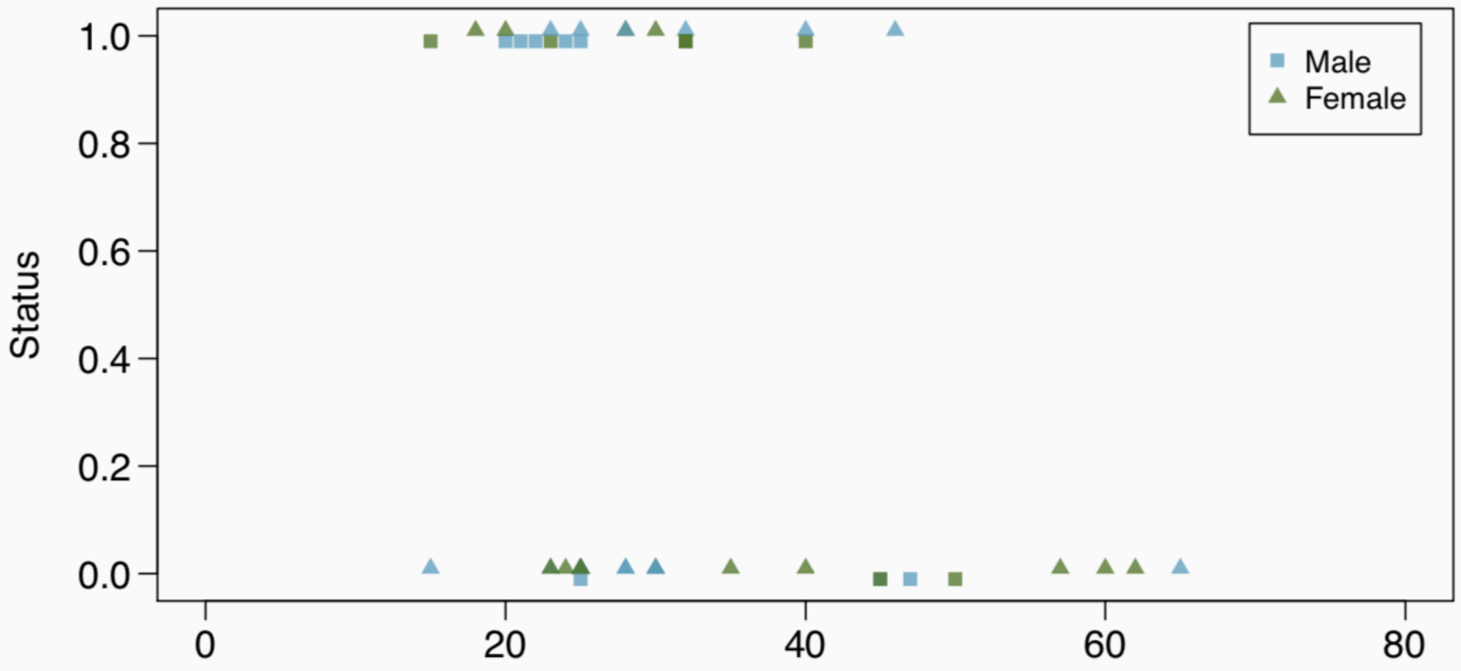
$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 50$$

$$\frac{p}{1-p} = \exp(-1.5065) = 0.222$$

$$p = 0.222/1.222 = 0.181$$

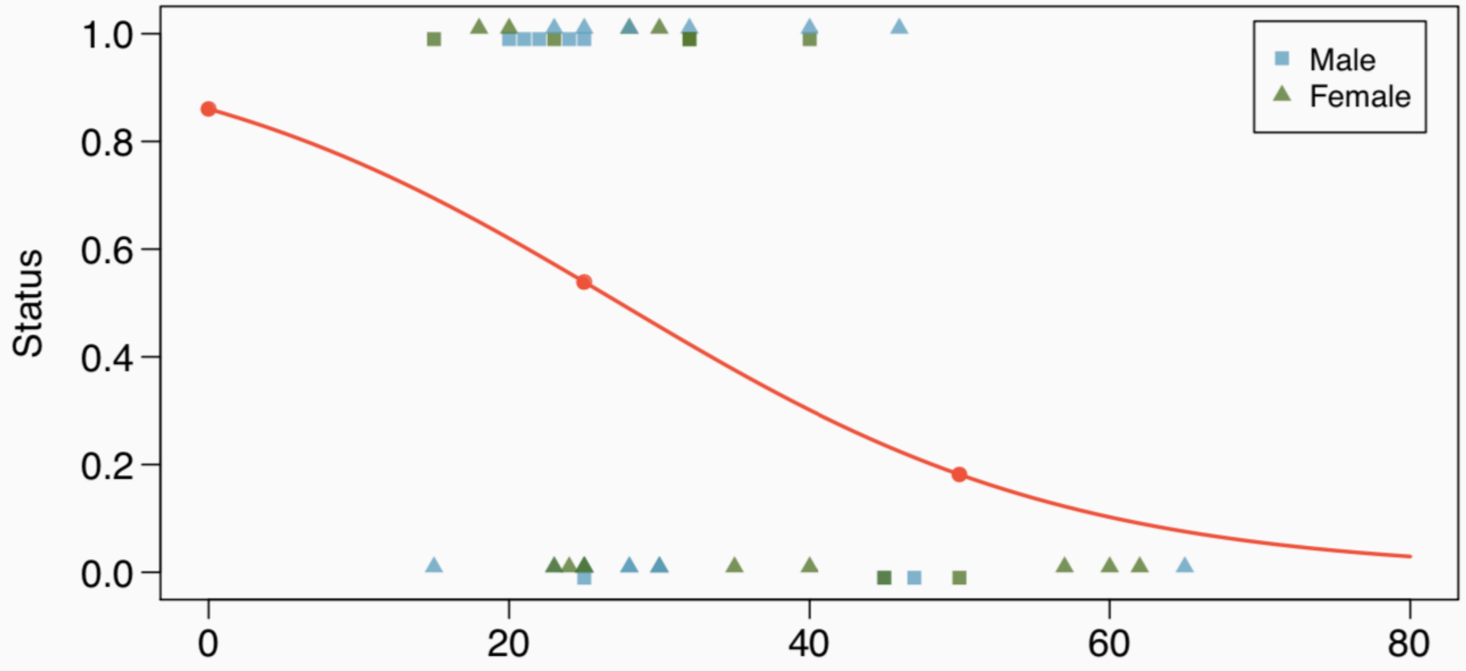
Example - Donner Party - Prediction (cont.)

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$



Example - Donner Party - Prediction (cont.)

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$



Example - Donner Party - Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Simple interpretation is only possible in terms of *log odds* and *log odds ratios* for intercept and slope terms.

Intercept: The *log odds* of survival for a party member with an age of 0. From this we can calculate the odds or probability, but additional calculations are necessary.

Slope: For a unit increase in age (being 1 year older) how much will the *log odds ratio* change, not particularly intuitive. More often than not we care only about sign and relative magnitude.

Example - Donner Party - Interpretation

$$\log\left(\frac{p_1}{1-p_1}\right) = 1.8185 - 0.0665(x+1)$$

$$= 1.8185 - 0.0665x - 0.0665$$

$$\log\left(\frac{p_2}{1-p_2}\right) = 1.8185 - 0.0665x$$

$$\log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_2}{1-p_2}\right) = -0.0665$$

$$\log\left(\frac{p_1}{1-p_1}\Big/\frac{p_2}{1-p_2}\right) = -0.0665$$

$$\frac{p_1}{1-p_1}\Big/\frac{p_2}{1-p_2} = \exp(-0.0665) = 0.94$$

Example - Donner Party - Age and Gender

```
summary(glm(Status  ~Age + Sex, data=donner, family=binomial))
## Call:
## glm(formula = Status  ~Age + Sex, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.63312    1.11018   1.471   0.1413
## Age         -0.07820    0.03728  -2.097   0.0359 *
## SexFemale    1.59729    0.75547   2.114   0.0345 *
## ---
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

Example - Donner Party - Gender Models

Just like MLR we can plug in gender to arrive at two status vs age models for men and women respectively.

General model:

$$\log\left(\frac{p_1}{1-p_1}\right) = 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times \text{Sex}$$

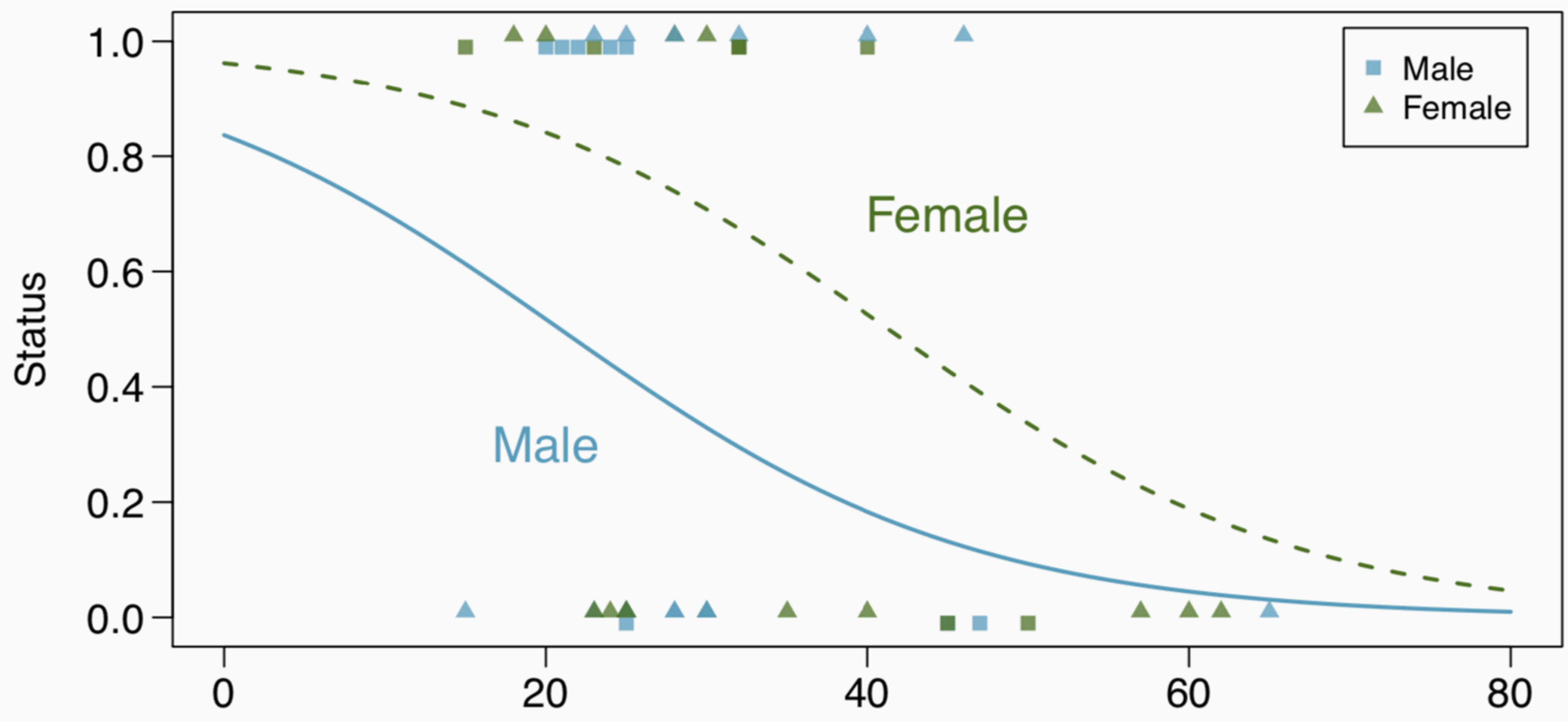
Male model:

$$\begin{aligned}\log\left(\frac{p_1}{1-p_1}\right) &= 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times 0 \\ &= 1.63312 + -0.07820 \times \text{Age}\end{aligned}$$

Female model:

$$\begin{aligned}\log\left(\frac{p_1}{1-p_1}\right) &= 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times 1 \\ &= 3.23041 + -0.07820 \times \text{Age}\end{aligned}$$

Example - Donner Party - Gender Models (cont.)



Hypothesis test for the whole model

```
summary(glm(Status ~ Age + Sex, data=donner, family=binomial))

## Call:
## glm(formula = Status ~ Age + Sex, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.63312    1.11018   1.471   0.1413
## Age         -0.07820    0.03728  -2.097   0.0359 *
## SexFemale    1.59729    0.75547   2.114   0.0345 *
## ---
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

Hypothesis test for the whole model

```
summary(glm(Status ~ Age + Sex, data=donner, family=binomial))

## Call:
## glm(formula = Status ~ Age + Sex, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.63312    1.11018   1.471   0.1413
## Age         -0.07820    0.03728  -2.097   0.0359 *
## SexFemale    1.59729    0.75547   2.114   0.0345 *
## ---
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

Note: The model output does not include any F-statistic, as a general rule there are not single model hypothesis tests for GLM models.

Hypothesis tests for a coefficient

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

We are however still able to perform inference on individual coefficients, the basic setup is exactly the same as what we've seen before except we use a Z-test.

Note: The only tricky bit, which is way beyond the scope of this course, is how the standard error is calculated.

Testing for the slope of Age

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

$$H_0 : \beta_{age} = 0$$

$$H_A : \beta_{age} \neq 0$$

$$Z = \frac{\hat{\beta_{age}} - \beta_{age}}{SE_{age}} = \frac{-0.0782 - 0}{0.0373} = -2.10$$

$$\text{p-value} = P(|Z| > 2.10) = P(Z > 2.10) + P(Z < -2.10)$$

$$= 2 \times 0.0178 = 0.0359$$

Confidence interval for age slope coefficient

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

Remember, the interpretation for a slope is the change in log odds ratio per unit change in the predictor.

Log odds ratio:

$$CI = PE \pm CV \times SE = -0.0782 \pm 1.96 \times 0.0373 = (-0.1513, -0.0051)$$

Odds ratio:

$$\exp(CI) = (\exp -0.1513, \exp -0.0051) = (0.85960.9949)$$

Useful Links & Resources

- **Reference**:
  - openintro.org/os (Chapter 9, Section 9.3, 9.5)
  - https://online.stat.psu.edu/stat462/node/117/
  - http://www2.stat.duke.edu/~cr173/Sta102_Fa15/Lec/Lec22.pdf