- ANOVA
- SLR
- LR.
- Logistic Regression

ANOVA Table

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Between)	depth	2	16.96	8.48	6.13	0.0063
(Within Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Degrees of freedom associated with ANOVA

- groups: $df_B = k - 1$, where k is the number of groups
- total: $df_T = n - 1$, where n is the total sample size
- error: $df_E = df_T - df_B$

n-k

- $df_B = k - 1 = 3 - 1 = 2$
- $df_T = n - 1 = 30 - 1 = 29$
- $df_E = 29 - 2 = 27$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Between)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares between groups, SSB Measures the variability between groups

$$SSB = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

where n_i is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean.

	n	mean
bottom	10	6.04
middepth	10	5.05
surface	10	4.2
overall	30	5.1

$$\begin{aligned} SSB &= \left(10 \times (6.04 - 5.1)^2\right) \\ &+ \left(10 \times (5.05 - 5.1)^2\right) \\ &+ \left(10 \times (4.2 - 5.1)^2\right) \\ &= 16.96 \end{aligned}$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Between)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares total, SST Measures the variability between groups

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$$

where x_{ij} represent each observation in the dataset.

$\frac{SST}{n-1}$ = sample variance of all obs. data

$(x_{11}, \ldots, x_{2j}, \ldots, x_{3,n_3})$

pool all data together

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Between)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares error, SSE Measures the variability within groups:

$$SSE = SST - SSB$$

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

$$SSE = 54.29 - 16.96 = 37.33$$

Mean square error Mean square error is calculated as sum of squares divided by the degrees of freedom.

$$MSB = 16.96/2 = 8.48$$
$$MSE = 37.33/27 = 1.38$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Between)	depth	2	16.96	8.48	6.14	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

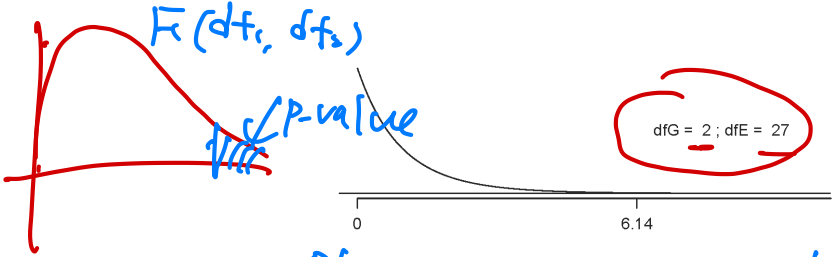
Test statistic, F value As we discussed before, the F statistic is the ratio of the between group and within group variability.

$$F = \frac{MSB}{MSE}$$

$$F = \frac{8.48}{1.38} = 6.14$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Between)	depth	2	16.96	8.48	6.14	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

p-value p-value is the probability of at least as large a ratio between the "between group" and "within group" variability, if in fact the means of all groups are equal. It's calculated as the area under the F curve, with degrees of freedom df_G and df_E, above the observed F statistic.



$P(F(3,5) > c) = \alpha \Leftrightarrow P(F(5,3) < \frac{1}{c}) = 1-\alpha$

Conclusion

$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots$

$H_1:$ not H_0

What is the conclusion of the hypothesis test?

The data provide convincing evidence that the average aldrin concentration

(a) is different for all groups.
(b) on the surface is lower than the other levels.
(c) is different for at least one group.
(d) is the same for all groups.

From Summarized Statistics to ANOVA

An experiment aims to compare the effectiveness of three different fertilizers (A, B, C) on plant yield. The results are summarized as follows:

Fertilizer	Sample Size	Mean	Variance
A	5	10	12
B	6	8	15
C	7	12	20
Overall	18	10.11	17.34

Table: Summary of yield data for different fertilizers.

Source	Df	Sum Sq	Mean Sq	F value
Between				
Residuals				
Total				

Table: Analysis of Variance (ANOVA)

From Summarized Statistics to ANOVA

An experiment aims to compare the effectiveness of three different fertilizers (A, B, C) on plant yield. The results are summarized as follows:

Fertilizer	Sample Size	Mean	Variance
A	5	10	12
B	6	8	15
C	7	12	20
Overall	18	10.11	17.34

Table: Summary of yield data for different fertilizers.

Source	Df	Sum Sq	Mean Sq	F value
Between	2	51.8	25.9	1.6
Residuals	15	243	16.2	
Total	17	294.8		

Table: Analysis of Variance (ANOVA)

Example 2

2-way ANOVA

- An important consideration in deciding which database management system to employ is the mean time required to learn how to use the system. A test was designed involving three systems and four users. Each user took the following amount of time (in hours) in training with each system:

	User			
	1	2	3	4
System 1	20	23	18	17
System 2	20	21	17	16
System 3	28	26	23	22

H0A

H0B: Col's effect

row effect

The study goal is to infer the differences between systems and between users.

- This type of problem can be analyzed by 2-factor ANOVA (or 2-way ANOVA).

	User			
	1	2	3	4
System 1	20	23	18	17
System 2	20	21	17	16
System 3	28	26	23	22

time sys user

20 S1 U1

20 S2 U2

28 S3 U3

23 S1 U4

21 S2 U4

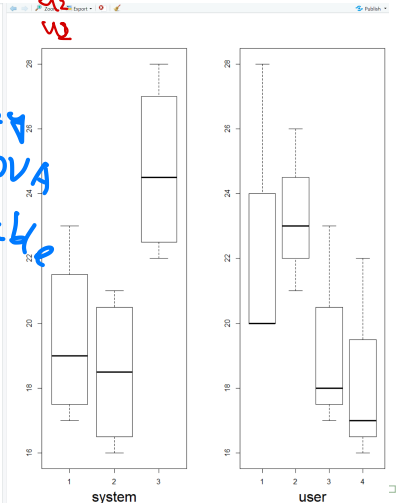
S3 U4

```
> par(mfcol=c(1,2))
> boxplot(time~system, data=x, xlab="system", cex.lab=2)
> boxplot(time~user, data=x, xlab="user", cex.lab=2)
> fit1 = lm(time~system+user, data=x)
> anova(fit1)   #2-way ANOVA table
Analysis of Variance Table

Response: time
          Df Sum Sq Mean Sq F value   Pr(>F)
system     2 90.167  45.083  41.615 0.000304 ***
user       3 54.250  18.083  16.692 0.002570 **
Residuals  6  6.500   1.083
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(fit1)

Call:
lm(formula = time ~ system + user, data = x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.25000 -0.18750  0.08333  0.08333  1.50000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 21.2500     0.7360  28.873 1.14e-07 ***
system2     -1.0000     0.7360  -1.359 0.223089
system3      5.2500     0.7360   7.133 0.000382 ***
user2        0.6667     0.8498   0.784 0.462605
user3       -3.3333     0.8498  -3.922 0.007781 **
user4       -4.3333     0.8498  -5.099 0.002224 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.041 on 6 degrees of freedom
Multiple R-squared:  0.9569,    Adjusted R-squared:  0.921
F-statistic: 26.66 on 5 and 6 DF,  p-value: 0.0004991
```

F(2,6)

F(3,6)

2-way ANOVA Table



One-Way

```
> fit1<-lm(time~System,data=x)#one-way ANOVA table
> summary(fit1)

Call:
lm(formula = time ~ System, data = x)

Residuals:
   Min     1Q Median     3Q    Max
-2.750 -1.938 -0.500  1.750  3.500

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   19.500      1.299  15.011 1.12e-07 ***
System2       -1.000      1.837  -0.544   0.5994
System3        5.250      1.837   2.858   0.0188 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.598 on 9 degrees of freedom
Multiple R-squared:  0.5975,    Adjusted R-squared:  0.508
F-statistic: 6.679 on 2 and 9 DF,  p-value: 0.01666

> anova(fit1)
Analysis of Variance Table

Response: time
          Df Sum Sq Mean Sq F value  Pr(>F)
System     2 90.167  45.083   6.679 0.01666 *
Residuals  9 60.750   6.750
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

54.25+6.5

One-Way

```
> fit2<-lm(time~User,data=x)#one-way ANOVA table
> summary(fit2)

Call:
lm(formula = time ~ User, data = x)

Residuals:
   Min     1Q Median     3Q    Max
-2.667 -2.333 -1.333  2.917  5.333

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.6667     2.0069  11.294  3.4e-06 ***
User2         0.6667     2.8382   0.235    0.820
User3        -3.3333     2.8382  -1.174    0.274
User4        -4.3333     2.8382  -1.527    0.165
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.476 on 8 degrees of freedom
Multiple R-squared:  0.3595,    Adjusted R-squared:  0.1193
F-statistic: 1.497 on 3 and 8 DF,  p-value: 0.2877

> anova(fit2)
Analysis of Variance Table

Response: time
          Df Sum Sq Mean Sq F value Pr(>F)
User       3 54.250  18.083  1.4966 0.2877
Residuals  8 96.667  12.083
```

90.167+6.5

2-Way

```
Response: time
          Df Sum Sq Mean Sq F value   Pr(>F)
system     2 90.167  45.083  41.615 0.000304 ***
user       3 54.250  18.083  16.692 0.002570 **
Residuals  6  6.500   1.083
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Assumptions on ANOVA Models

- 1-factor model: $X_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$.
- 2-factor model: $X_{ij} = \mu_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$.
- Key assumptions:
  - data are independent
  - normal distributed
  - equal variance accross groups
- Can we check these assumptions?

- Residual Plots
- Q-Q plot

Check Normality Assumption

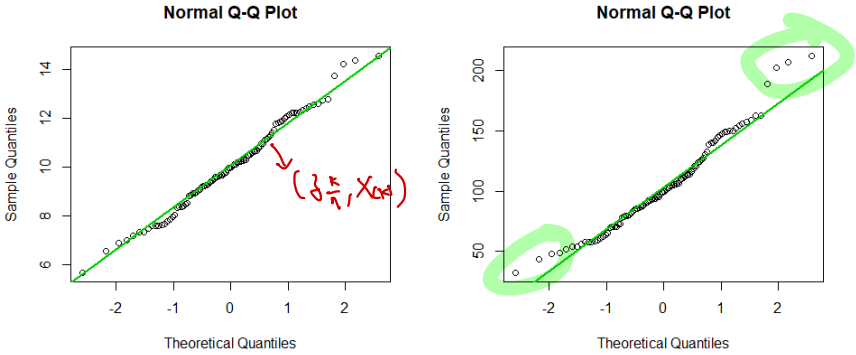
QQ (quantile-quantile) plot: a visualization method

$\frac{x_q - \mu}{\sigma} = z_q$

- q-th quantile of $N(\mu, \sigma^2)$ satisfies $x_q = \mu + \sigma z_q$, where z_q is q-th quantile of N(0,1).
- Under normality, x_q should be a linear function of z_q.
- Ordered data: $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$
- Sample quantile estimated from the ordered data:
  $X_{(k)}$ corresponds to $\frac{k-0.5}{n}$-th sample quantile
- If data X_i's follow a normal distribution, we expect that the plot of $X_{(k)}$ v.s. N(0,1) quantiles $z_{(k-0.5)/n}$ behaves like a straight line.
- View this quantile-quantile plot (sample quantiles vs normal quantiles) to determine if the normality is a reasonable assumption.

QQ Plot for Normality Checking

Q: Which data set is more normal-like?

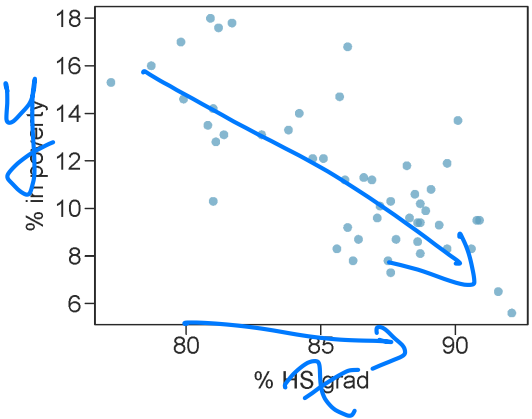


- A straight line indicates normality assumption is reasonable!

Linear Regression

Poverty vs. HS graduate rate

The scatterplot below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Response variable?

% in poverty

Explanatory variable?

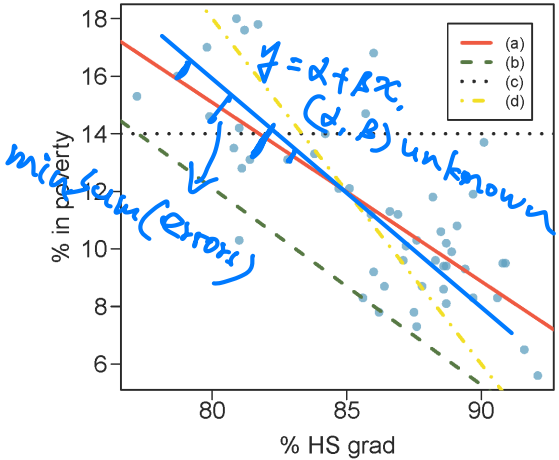
% HS grad

Relationship?

linear, negative, moderately strong

Eyeballing the line

Which of the following appears to be the line that best fits the linear relationship between % in poverty and % HS grad? Choose one. (a)

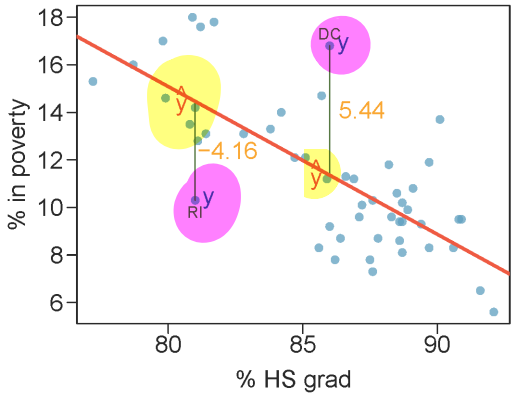


Residuals

Residuals are the leftovers from the model fit: Data = Fit + Residual

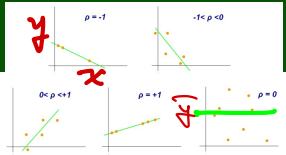
Residual is the difference between the observed (y_i) and predicted $\hat{y}_i$.

$$e_i = y_i - \hat{y}_i$$

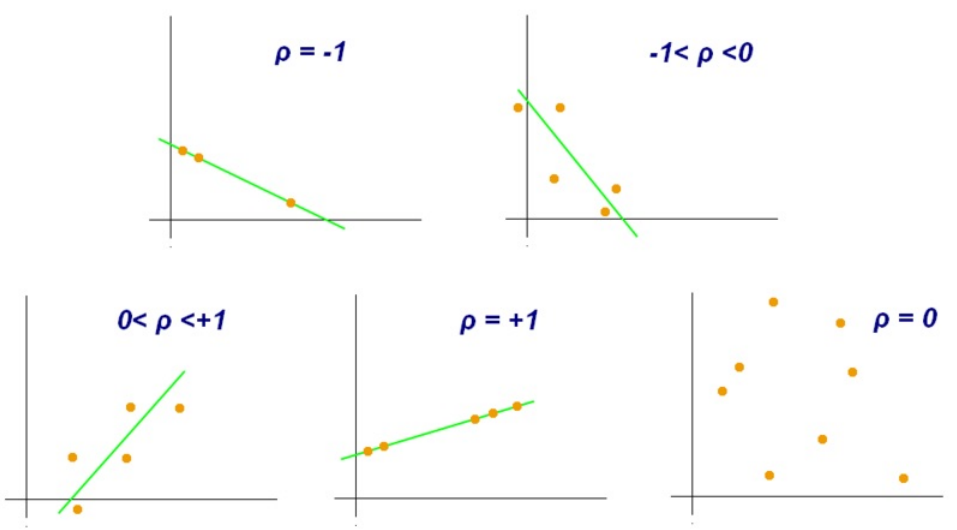


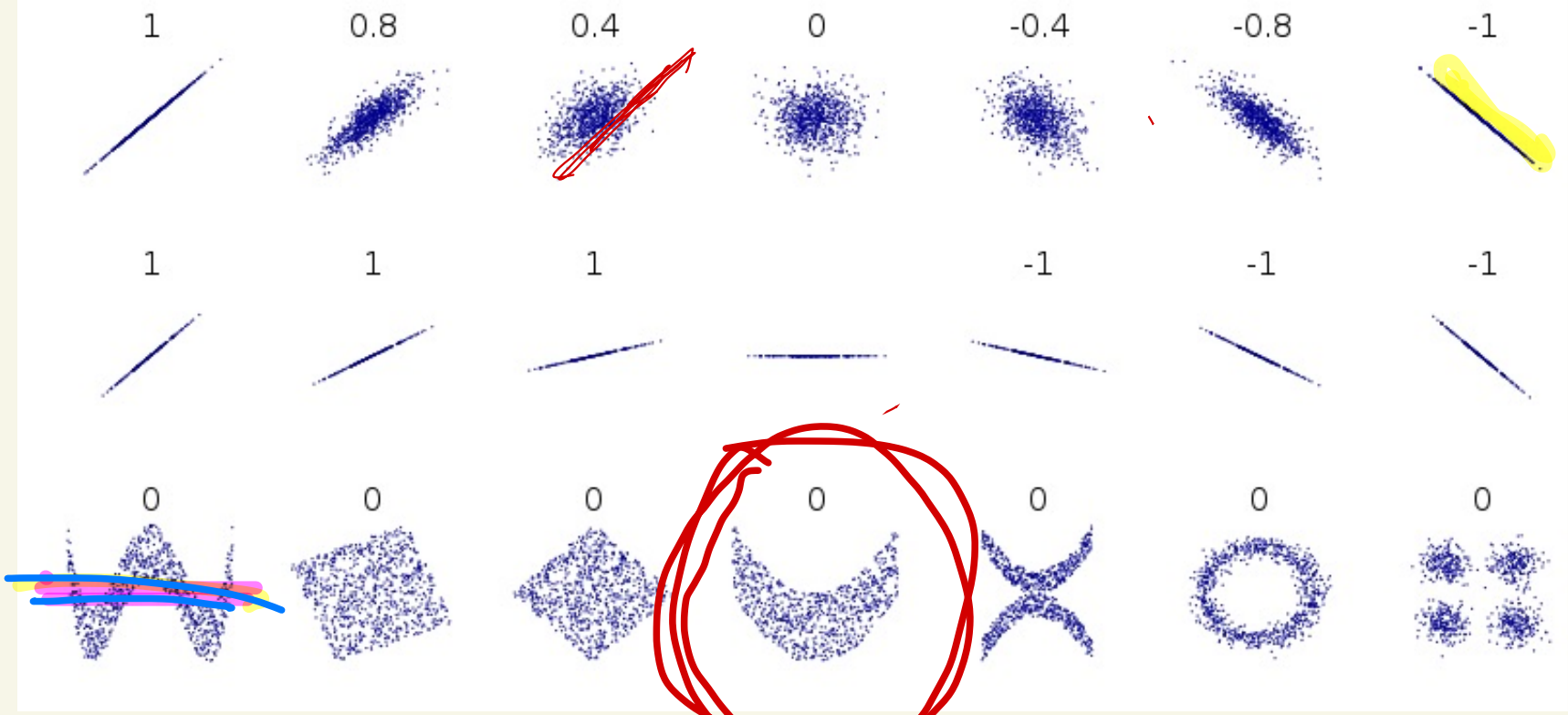
- % living in poverty in DC is 5.44% more than predicted.
- % living in poverty in RI is 4.16% less than predicted.

Quantifying the relationship



- Correlation describes the strength of the linear association between two variables.
- It takes values between -1 (perfect negative) and +1 (perfect positive).
- A value of 0 indicates no linear association.
- $R = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\, s_x s_y}$, where s_x, s_y are the sample standard deviations of x_i and y_i.
- Note that R represents the cosine of the angle between the vectors $\mathbf{a} = (x_1 - \bar{x}, \ldots, x_n - \bar{x})$ and $\mathbf{b} = (y_1 - \bar{y}, \ldots, y_n - \bar{y})$.





R: (linear) correlation coef.

The least squares line

model : $y = \beta_0 + \beta_1 x + \varepsilon$

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

$\beta_0, \beta_1, \sigma^2$ unknown

$E(y|x) = \beta_0 + \beta_1 x$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- $\hat{y}$: Predicted value of the response variable, y
- β_0: Intercept, parameter
  - b_0 (or $\hat{\beta}_0$): Intercept, point estimate
- β_1 : Slope, parameter
  - b_1 (or $\hat{\beta}_1$): Slope, point estimate
- x: Explanatory variable

Conditions for the least squares line

Data: $(x_1, y_1) \cdots (x_n, y_n)$

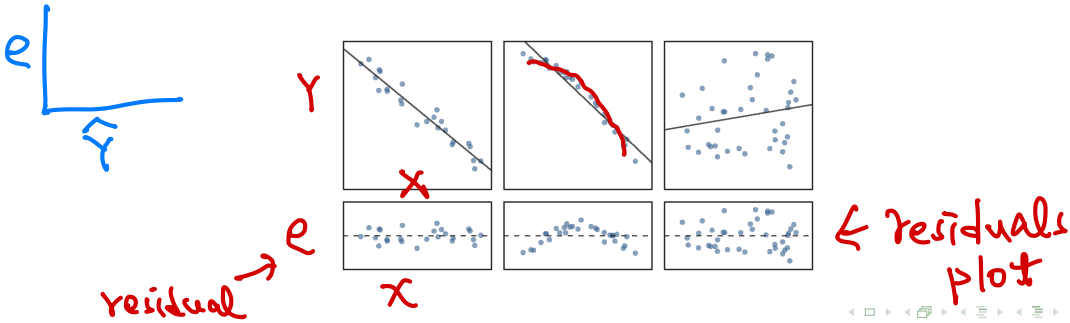
model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, \cdots n$

ε_i 獨立 $N(0, \sigma^2)$

1. Linearity
2. Nearly normal residuals
3. Constant variability

Conditions: (1) Linearity

- The relationship between the explanatory and the response variable should be linear.
- Methods for fitting a model to non-linear relationships exist, but are beyond the scope of this class. If this topic is of interest, an `http://www.openintro.org/download.php?file=os2_extra_nonlinear_relationships&referrer=/stat/textbook.php`Online Extra is available on openintro.org covering new techniques.
- Check using a scatterplot of the data, or a residuals plot.



Conditions: (2) Nearly normal residuals

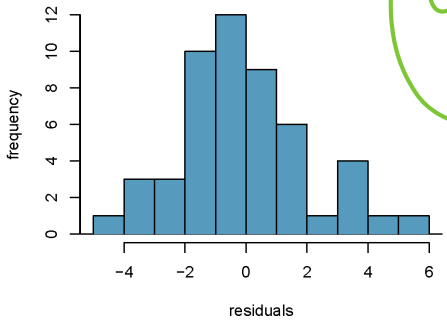
- The residuals should be nearly normal.
- This condition may not be satisfied when there are unusual observations that don't follow the trend of the rest of the data.
- Check using a histogram, Q-Q plot

model: $y = \beta_0 + \beta_1 x_i + \varepsilon_i$

Fitted $\hat{y}_i = \hat{\beta}_0 + \hat{\beta} x_i$

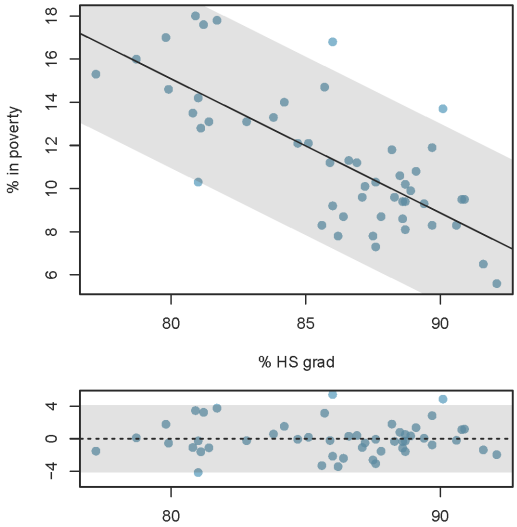
$e_i = y_i - \hat{y}_i$

$e_i \approx \hat{\varepsilon}_i$



Q-Q plot

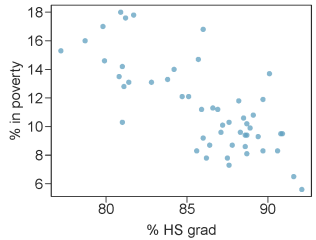
Conditions: (3) Constant variability



- The variability of points around the least squares line should be roughly constant.
- This implies that the variability of residuals around the 0 line should be roughly constant as well.
- Also called homoscedasticity.
- Check using a residuals plot.

← residuals plot

Given...



	% HS grad (x)	% in poverty (y)
mean	$\bar{x} = 86.01$	$\bar{y} = 11.35$
sd	$s_x = 3.73$	$s_y = 3.1$
	correlation	$R = -0.75$

Slope

The slope of the regression can be calculated as

$$b_1 = \frac{s_y}{s_x} R \quad (\hat{\beta}_1)$$

In context...

$$b_1 = \frac{3.1}{3.73} \times -0.75 = -0.62$$

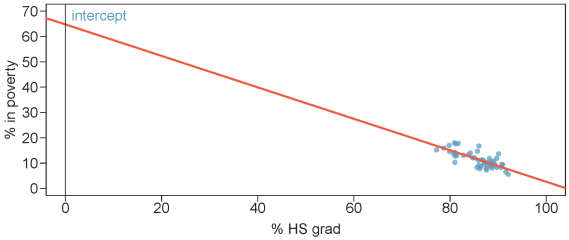
Interpretation

For each additional % point in HS graduate rate, we would expect the % living in poverty to be lower on average by 0.62% points.

Intercept

The intercept is where the regression line intersects the y-axis. The calculation of the intercept uses the fact the a regression line always passes through $(\bar{x}, \bar{y})$.

$$b_0 = \bar{y} - b_1 \bar{x} \quad (\hat{\beta}_0)$$



$$\begin{aligned} b_0 &= 11.35 - (-0.62) \times 86.01 \\ &= 64.68 \end{aligned}$$

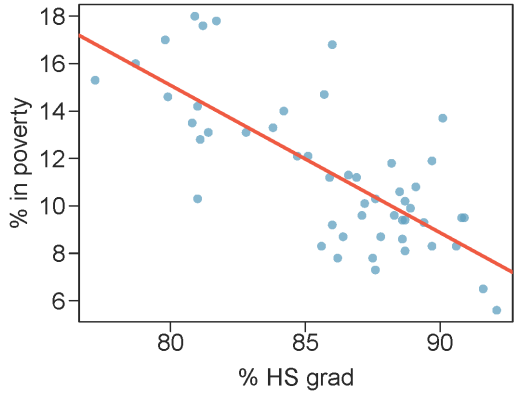
① $\hat{y} = b_0 + b_1 x$, ② $b_1 = \frac{s_y}{s_x} R$

③ $\hat{y} = \bar{y} + b_1 (x - \bar{x})$

Regression line

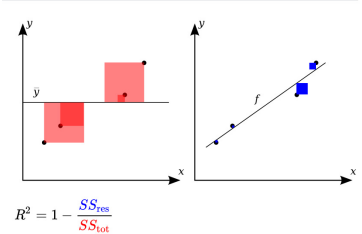
Predictions work best where we have data

$$\widehat{\%\ in\ poverty} = 64.68 - 0.62\ \%\ HS\ grad$$



R^2

- The strength of the fit of a linear model is most commonly evaluated using R^2.
- R^2 is calculated as the square of the correlation coefficient.
- It tells us what percent of variability in the response variable is explained by the model.
- For the model we've been working with, $R^2 = (-0.75)^2 = 0.56$.
- $R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\text{Residuals sum of squares}}{\text{Total sum of squares}}$



Interpretation of R^2

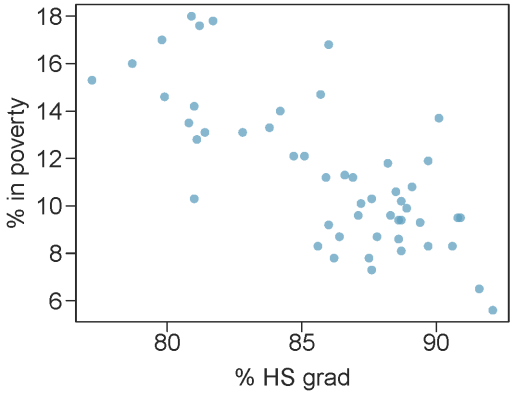
Which of the below is the correct interpretation of $R = -0.75$, $R^2 = 0.56$?

(a) 56% of the variability in the % of HG graduates among the 51 states is explained by the model.

(b) 56% of the variability in the % of residents living in poverty among the 51 states is explained by the model.

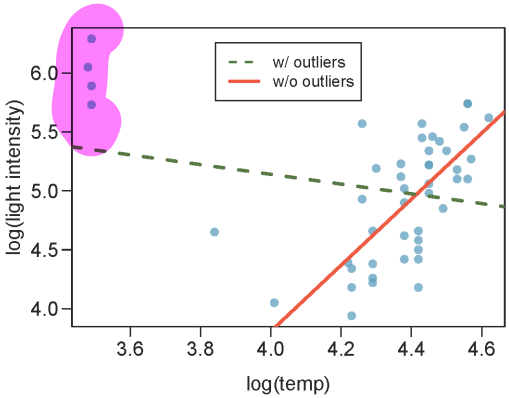
(c) 56% of the time % HS graduates predict % living in poverty correctly.

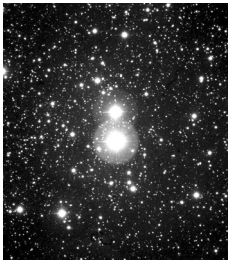
(d) 75% of the variability in the % of residents living in poverty among the 51 states is explained by the model.



Influential points

Data are available on the log of the surface temperature and the log of the light intensity of 47 stars in the star cluster CYG OB1.





Testing for the slope

Assuming that these 27 twins comprise a representative sample of all twins separated at birth, we would like to test if these data provide convincing evidence that the IQ of the biological twin is a significant predictor of IQ of the foster twin. What are the appropriate hypotheses?

(a) H_0: $b_0 = 0$; H_A: $b_0 \neq 0$
(b) H_0: $\beta_0 = 0$; H_A: $\beta_0 \neq 0$
(c) H_0: $b_1 = 0$; H_A: $b_1 \neq 0$
(d) H_0: $\beta_1 = 0$; H_A: $\beta_1 \neq 0$

$y = \beta_0 + \beta_1 x + \varepsilon$

$\beta_0, \beta_1, \varepsilon \sim N(0, \sigma^2)$

unknown

$\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}$

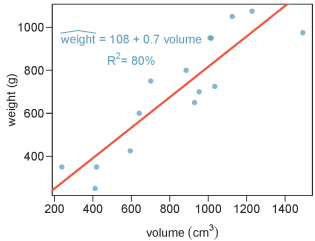
Weights of books

	weight (g)	volume (cm^3)	cover
1	800	885	hc
2	950	1016	hc
3	1050	1125	hc
4	350	239	hc
5	750	701	hc
6	600	641	hc
7	1075	1228	hc
8	250	412	pb
9	700	953	pb
10	650	929	pb
11	975	1492	pb
12	350	419	pb
13	950	1010	pb
14	425	595	pb
15	725	1034	pb

From: Maindonald, J.H. and Braun, W.J. (2nd ed., 2007) "Data Analysis and Graphics Using R"

Weights of books (cont.)

The scatterplot shows the relationship between weights and volumes of books as well as the regression output. Which of the below is correct?



(a) Weights of 80% of the books can be predicted accurately using this model.

(b) Books that are 10 cm^3 over average are expected to weigh 7 g over average.

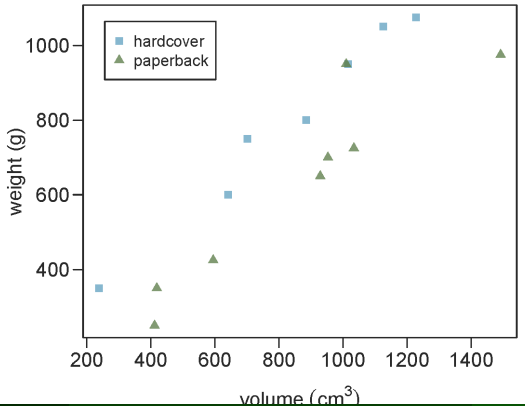
(c) The correlation between weight and volume is $R = 0.80^2 = 0.64$.

(d) The model underestimates the weight of the book with the highest volume.

Weights of hardcover and paperback books

Can you identify a trend in the relationship between volume and weight of hardcover and paperback books?

Paperbacks generally weigh less than hardcover books after controlling for the book?s volume.



Modeling weights of books using volume and cover type

$(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

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

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 197.96284   59.19274   3.344 0.005841 **
volume        0.71795    0.06153  11.669  6.6e-08 ***
cover:pb   -184.04727   40.49420  -4.545 0.000672 ***


Residual standard error: 78.2 on 12 degrees of freedom
Multiple R-squared: 0.9275,Adjusted R-squared: 0.9154
F-statistic: 76.73 on 2 and 12 DF,  p-value: 1.455e-07
```

$\hat{\sigma}$

Linear model

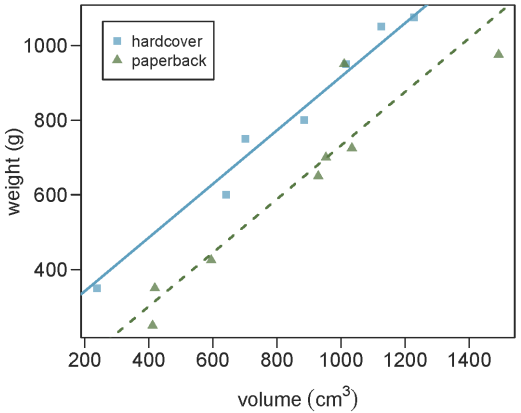
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05\ cover:pb$$

1. For hardcover books: plug in 0 for cover

$$\begin{aligned}\widehat{weight} &= 197.96 + 0.72\ volume - 184.05 \times 0 \\ &= 197.96 + 0.72\ volume\end{aligned}$$

Visualising the linear model



Adjusted R^2

Adjusted R^2

$$R^2_{adj} = 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right)$$

where n is the number of cases and p is the number of predictors (explanatory variables) in the model.

- Because p is never negative, R^2_{adj} will always be smaller than R^2.
- R^2_{adj} applies a penalty for the number of predictors included in the model.
- Therefore, we choose models with higher R^2_{adj} over others.

Full model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.6282	0.1720	26.90	0.00
beauty	0.1080	0.0329	3.28	0.00
gender.male	0.2040	0.0528	3.87	0.00
age	-0.0089	0.0032	-2.75	0.01
formal.yes [1]	0.1511	0.0749	2.02	0.04
lower.yes [2]	0.0582	0.0553	1.05	0.29
native.non english	-0.2158	0.1147	-1.88	0.06
minority.yes	-0.0707	0.0763	-0.93	0.35
students [3]	-0.0004	0.0004	-1.03	0.30
tenure.tenure track [4]	-0.1933	0.0847	-2.28	0.02
tenure.tenured	-0.1574	0.0656	-2.40	0.02

- tenure is a categorical variable : non-tenure, tenure track, tenured.

Assessing significance

Which predictors do not seem to meaningfully contribute to the model, i.e. may not be significant predictors of professor's rating score?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.6282	0.1720	26.90	0.00
beauty	0.1080	0.0329	3.28	0.00
gender.male	0.2040	0.0528	3.87	0.00
age	-0.0089	0.0032	-2.75	0.01
formal.yes	0.1511	0.0749	2.02	0.04
lower.yes	0.0582	0.0553	1.05	0.29
native.non english	-0.2158	0.1147	-1.88	0.06
minority.yes	-0.0707	0.0763	-0.93	0.35
students	-0.0004	0.0004	-1.03	0.30
tenure.tenure track	-0.1933	0.0847	-2.28	0.02
tenure.tenured	-0.1574	0.0656	-2.40	0.02

Backward-elimination

Full	beauty + gender + age + formal + lower + native + minority + students + tenure	0.0839
Step 1	gender + age + formal + lower + native + minority + students + tenure	0.0642
	beauty + age + formal + lower + native + minority + students + tenure	0.0557
	beauty + gender + formal + lower + native + minority + students + tenure	0.0706
	beauty + gender + age + lower + native + minority + students + tenure	0.0777
	beauty + gender + age + formal + native + minority + students + tenure	0.0837
	beauty + gender + age + formal + lower + minority + students + tenure	0.0788
	beauty + gender + age + formal + lower + native + students + tenure	0.0842
	beauty + gender + age + formal + lower + native + minority + tenure	0.0838
	beauty + gender + age + formal + lower + native + minority + students	0.0733
Step 2	gender + age + formal + lower + native + students + tenure	0.0647
	beauty + age + formal + lower + native + students + tenure	0.0543
	beauty + gender + formal + lower + native + students + tenure	0.0708
	beauty + gender + age + lower + native + students + tenure	0.0776
	beauty + gender + age + formal + native + students + tenure	0.0846
	beauty + gender + age + formal + lower + native + tenure	0.0844
	beauty + gender + age + formal + lower + native + students	0.0725
Step 3	gender + age + formal + native + students + tenure	0.0653
	beauty + age + formal + native + students + tenure	0.0534
	beauty + gender + formal + native + students + tenure	0.0707
	beauty + gender + age + native + students + tenure	0.0786
	beauty + gender + age + formal + students + tenure	0.0756
	beauty + gender + age + formal + native + tenure	0.0855
	beauty + gender + age + formal + native + students	0.0713
Step 4	gender + age + formal + native + tenure	0.0667
	beauty + age + formal + native + tenure	0.0553
	beauty + gender + formal + native + tenure	0.0723
	beauty + gender + age + native + tenure	0.0806
	beauty + gender + age + formal + tenure	0.0773
	beauty + gender + age + formal + native	0.0713

Forward selection

1. Start with regressions of response vs. each explanatory variable
2. Pick the model with the highest R^2_{adj}
3. Add the remaining variables one at a time to the existing model, and once again pick the model with the highest R^2_{adj}
4. Repeat until the addition of any of the remanning variables does not result in a higher R^2_{adj}

- Backward elimination with the p-value approach:
  1. Start with the full model
  2. Drop the variable with the highest p-value and refit a smaller model
  3. Repeat until all variables left in the model are significant
- Forward selection with the p-value approach:
  1. Start with regressions of response vs. each explanatory variable
  2. Pick the variable with the lowest significant p-value
  3. Add the remaining variables one at a time to the existing model, and pick the variable with the lowest significant p-value
  4. Repeat until any of the remaining variables does not have a significant p-value

`step` function in R

The `step` function in R does a similar backward elimination process, however it uses a different metric called AIC (Akaike Information Criterion) instead of adjusted R^2 to do the model selection.

$$AIC = n \log\left(\frac{RSS}{n}\right) + 2|B|$$

- $\frac{RSS}{n} = \hat{\sigma}^2$

Modeling conditions

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

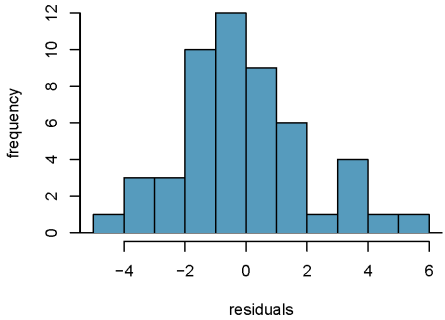
The model depends on the following conditions

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

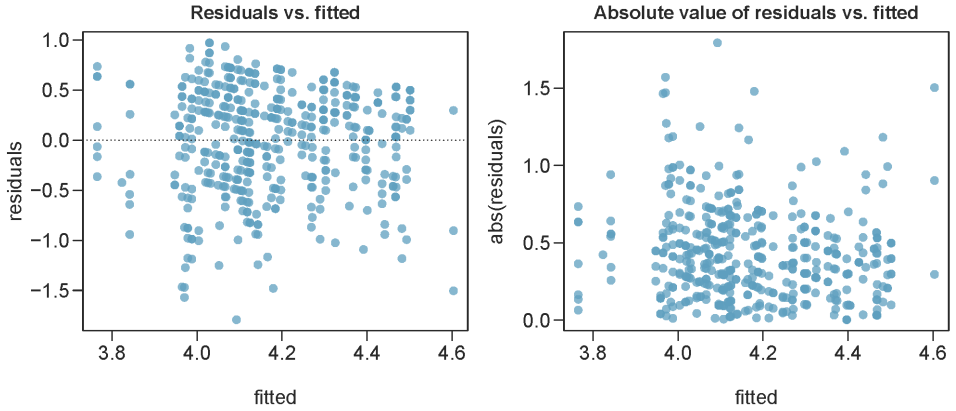
normal probability plot and/or histogram of residuals:



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?

Example - Donner Party

The last 2 pages from Ch 9.1

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 rote 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

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

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

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:

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

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r$$

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

From which we arrive at,

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

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 *
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 56.291  on 43  degrees of freedom
## AIC: 60.291
##
## Number of Fisher Scoring iterations: 4
```

$$\ln\frac{p}{1-p} = \beta_0 + \beta_1 x_i$$

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 (cont.)

Model:

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

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:

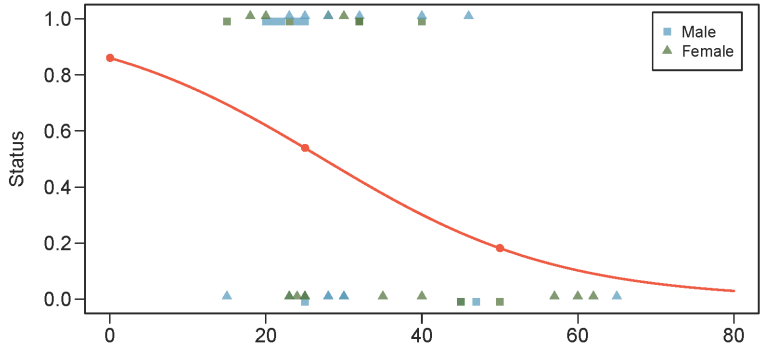
$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 50$$

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

$$p = 0.222/1.222 = 0.181$$

Example - Donner Party - Prediction (cont.)

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



Example - Donner Party - 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 then not we care only about sign and relative magnitude.

Example - Donner Party - Interpretation - Slope

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

$$= 1.8185 - 0.0665x - 0.0665$$

$$\log\left(\frac{p_0}{1-p_0}\right) = 1.8185 - 0.0665x$$

$$\log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_0}{1-p_0}\right) = -0.0665$$

$$\log\left(\frac{p_1}{1-p_1} \Big/ \frac{p_0}{1-p_0}\right) = -0.0665$$

$$\frac{p_1}{1-p_1} \Big/ \frac{p_0}{1-p_0} = \exp(-0.0665) = 0.94$$

odds

odds ratio

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
```

Gender slope: When the other predictors are held constant this is the log odds ratio between the given level (Female) and the reference level (Male).

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:

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

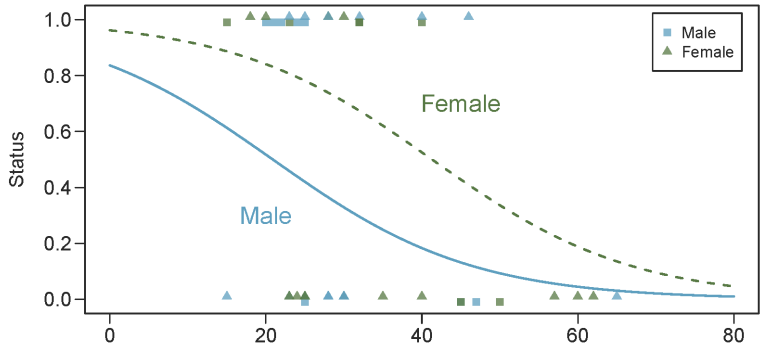
Male model:

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

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
```

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.

95% CI = $\hat{\beta} \pm 2\,SE(\hat{\beta})$

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

all blue color

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

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

$$\begin{aligned} \text{p-value} &= P(|Z| > 2.10) = P(Z > 2.10) + P(Z < -2.10) \\ &= 2 \times 0.0178 = 0.0359 \end{aligned}$$

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:

$$\exp(CI) = (\exp(-0.1513), \exp(-0.0051)) = (0.8597, 0.9949)$$