$y = x$

$x_1 = x$
$x_2 = x$

$y \sim x_1 + x_2$

$y = \frac{1}{2}x_1 + \frac{1}{2}x_2$
$y = 2x_1 - x_2$
$y = 100x_1 - 99x_2$

Practical Issues*

old | new

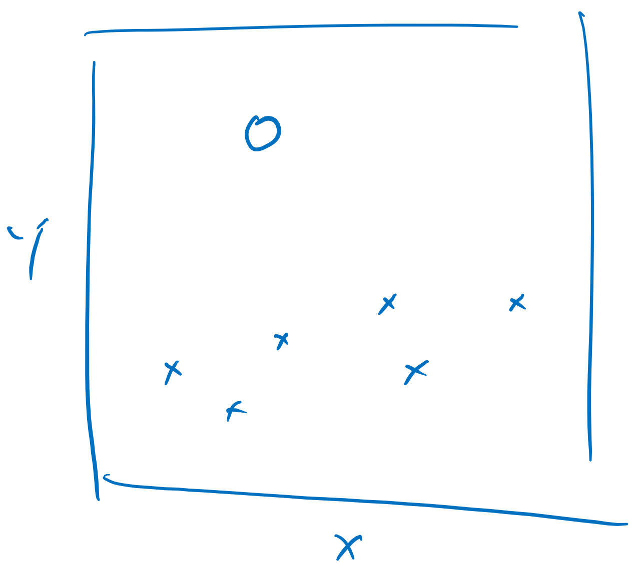
- Multicollinearity — $COR(x_1, x_2) > 0.8 \Rightarrow$ remove x_1 or x_2 — auto selection
- Non-constant variance of error terms (heteroscedasticity)
- Dependence of the error terms

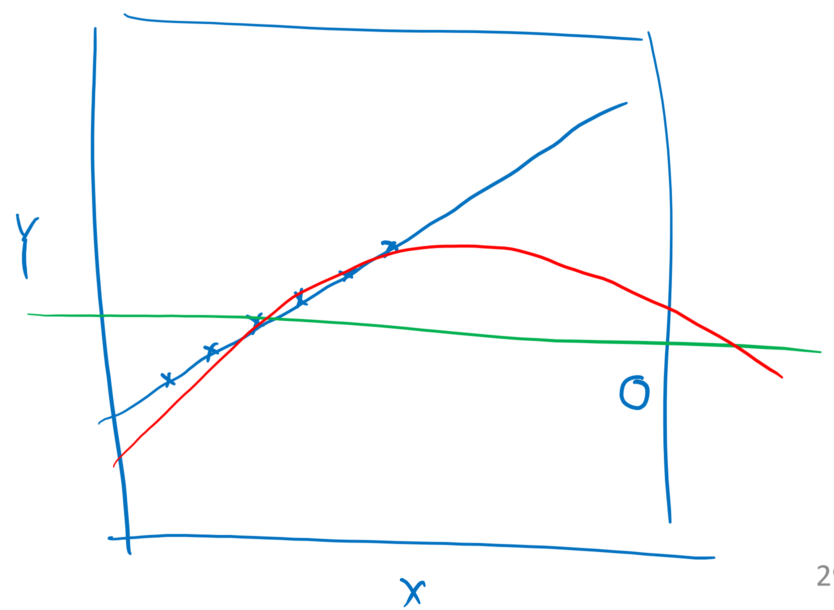
stat tests

~~p-value~~

- Outliers
- High leverage points

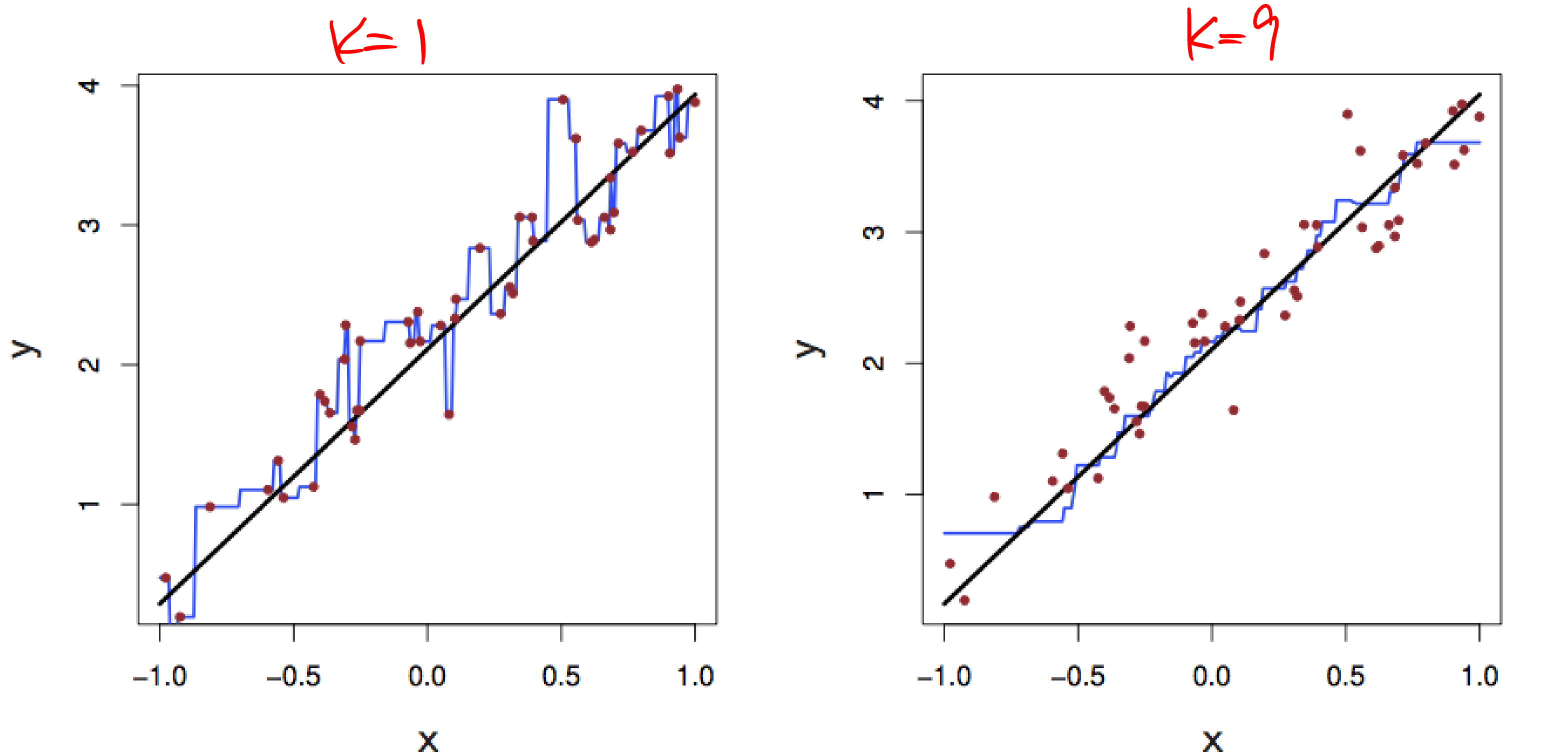
subjective





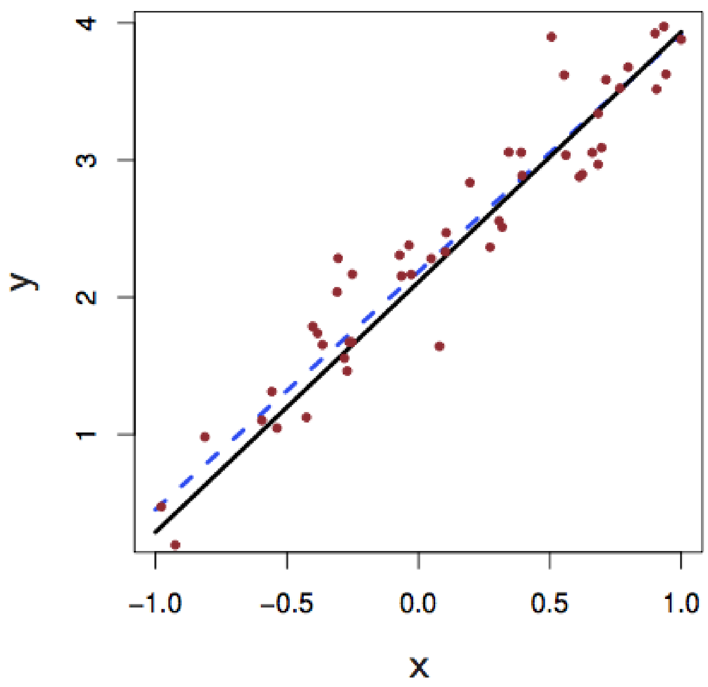
Comparison with KNN: linear data

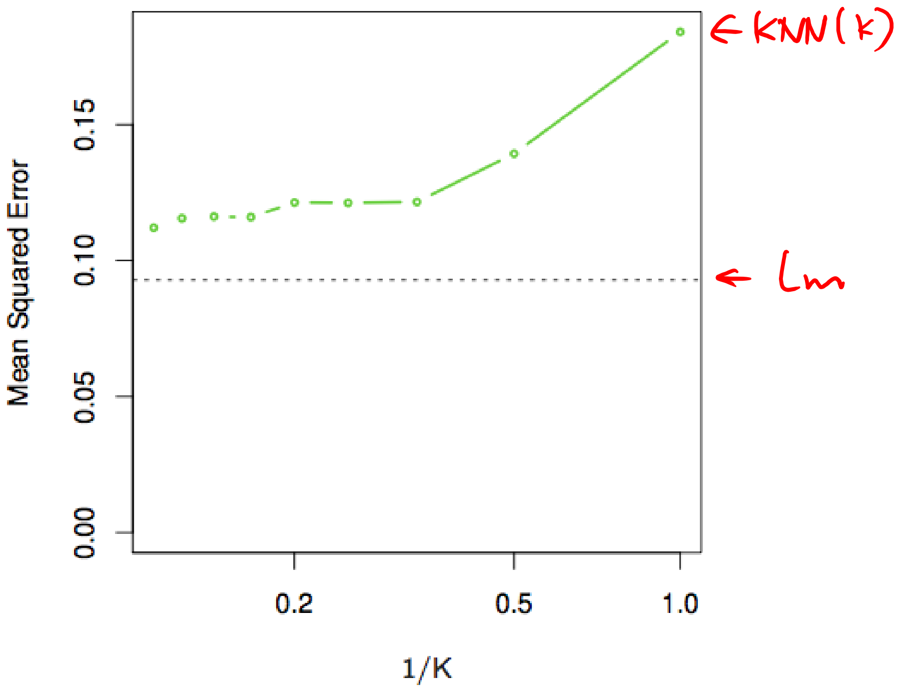
- KNN Fits in one dimension (k =1 and k = 9)



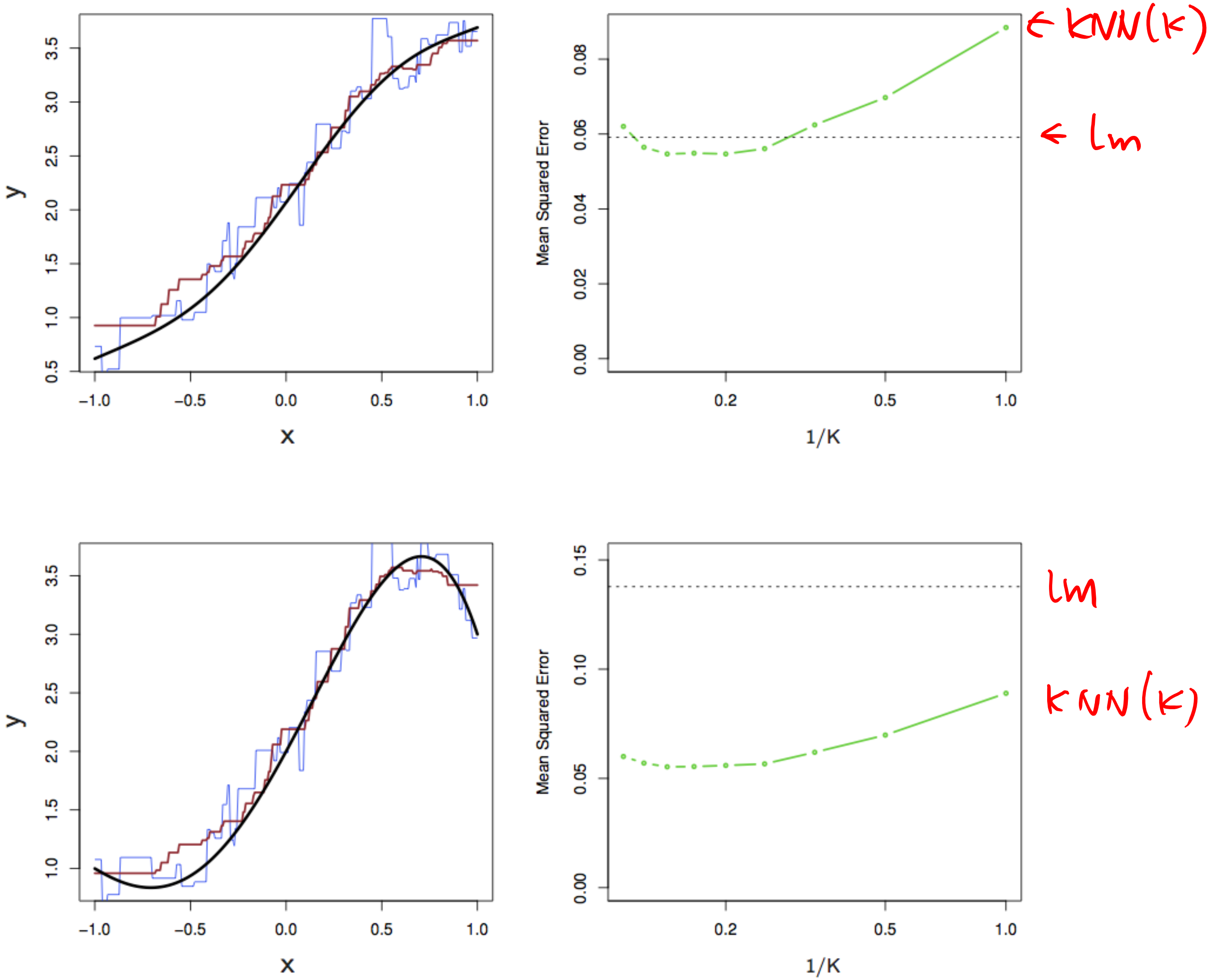
Comparison with KNN: linear data

- Linear Regression Fit

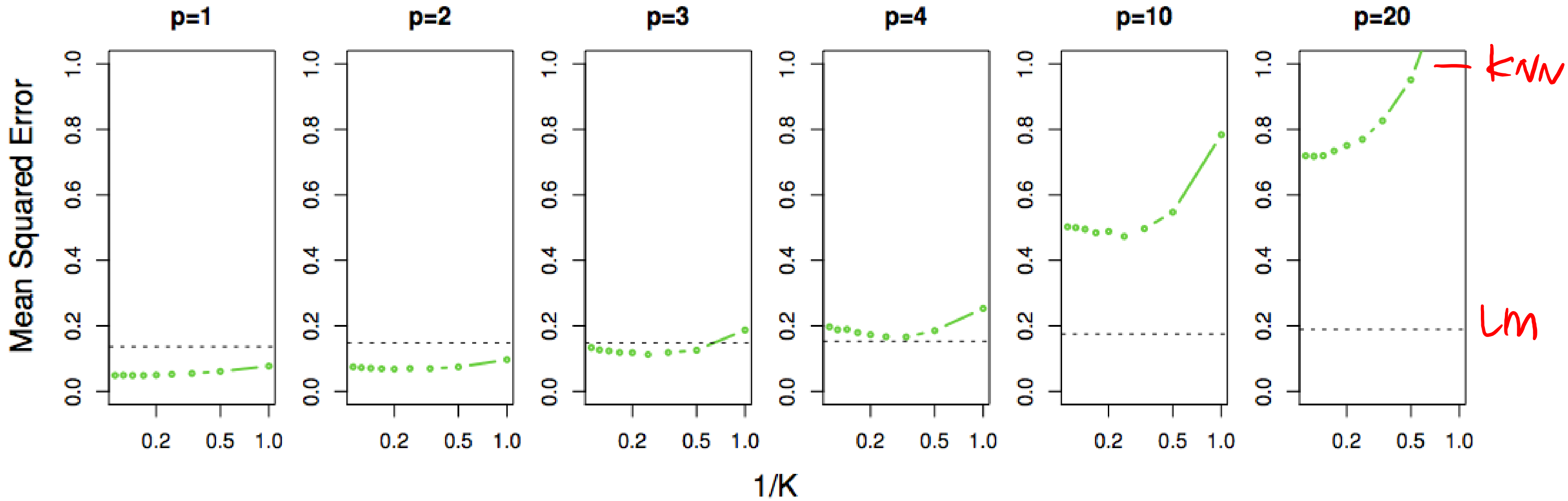




Comparison with KNN: more nonlinearity

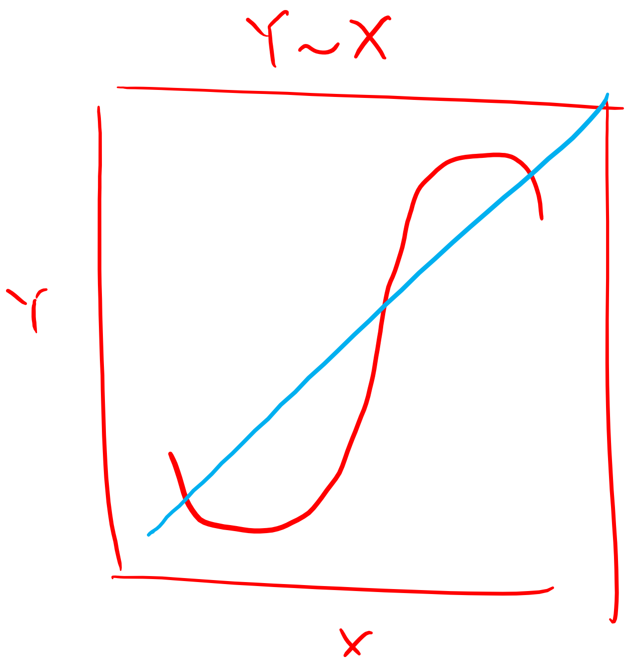


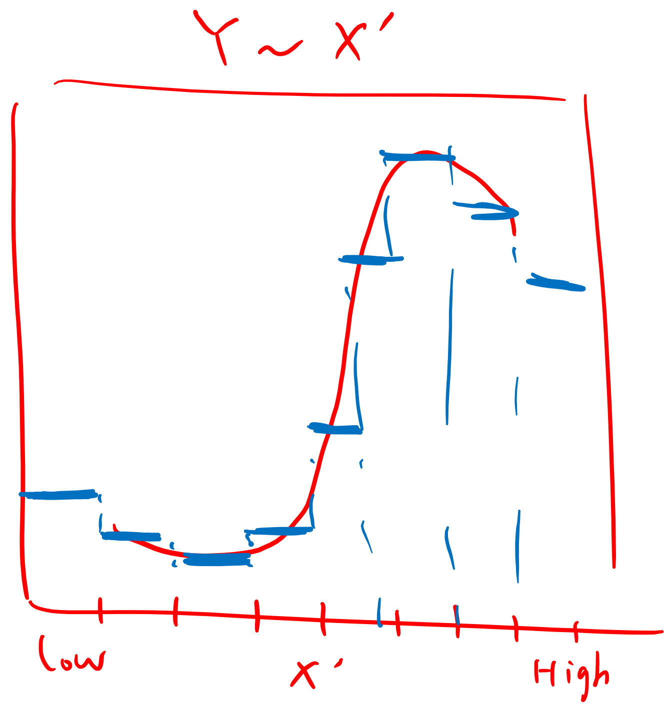
Comparison with KNN: higher dimensionality



Capture More Nonlinearity in Linear Regression

- Discretize X!





$Y \sim X + X'$

DSC5103 Statistics

Session 4. Generalized Linear Models

Review of last session

- Linear Regression
  - Simple and multiple linear regression model
  - Least squares estimation
  - Model assessment
  - ~~Model selection***~~
- Other Considerations in Regression Model
  - Qualitative predictors
  - Introducing nonlinearity: interaction terms, polynomial terms, log transformation
- Practical Issues
  - Multicollinearity
  - Heteroscedasticity
  - Outliers and high leverage points

Plan for today

$Y = \{\text{Default, No Default}\}$

- Binary Logistic Regression

$X \rightarrow \text{Prob}(Y=1) \rightarrow Y : \text{"0"}/\text{"1"}$

other

- Classification in general
  - Model evaluation and comparisons

- Other Generalized Linear Models
  - Poisson Regression
  - Survival Analysis

Generalized Linear Models

$Y = f(x) + \varepsilon \ , \ f(x) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

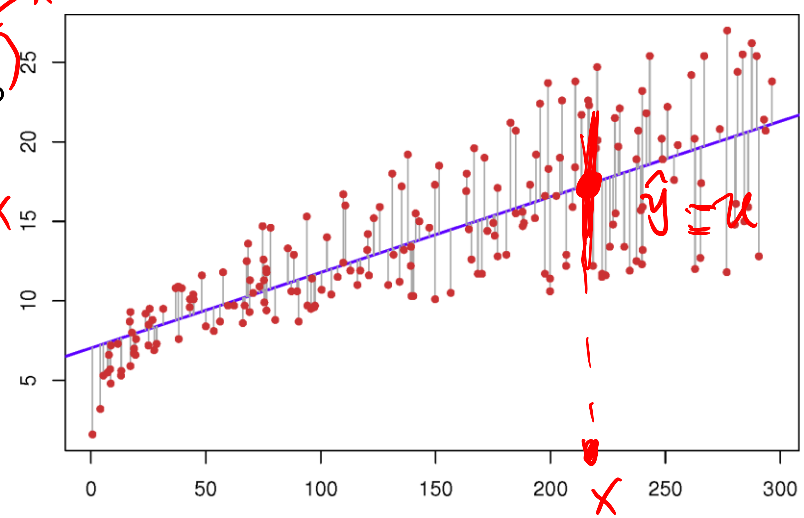
- Linear Regression: $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$
- Given the assumptions made on ε, Y is normally distributed and from -∞ to +∞

$Y \sim \text{Normal}(\mu, \sigma^2) \Rightarrow y \in (-\infty, \infty)$

$\mu = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

- What about other types of Y?
  - Categorical (binary or multinomial)? $Y \sim \text{Bernoulli}(p) \leftarrow x$
    - default={Yes, No}, email={Spam, Ham}, eyecolor={brown, blue, green}
  - Ordinal? Ordered Logit
    - ranking={1, 2, 3, 4}
  - Discrete counts (non-negative integers)? $Y \sim \text{Poisson}(\lambda) \leftarrow x$
    - HomeGoal={0, 1, 2, 3, …}
  - Survival time (non-negative real)? $Y \sim \text{Exp}(\lambda) \leftarrow x$
    - ResidualLife in {0, ∞}

"Time until"



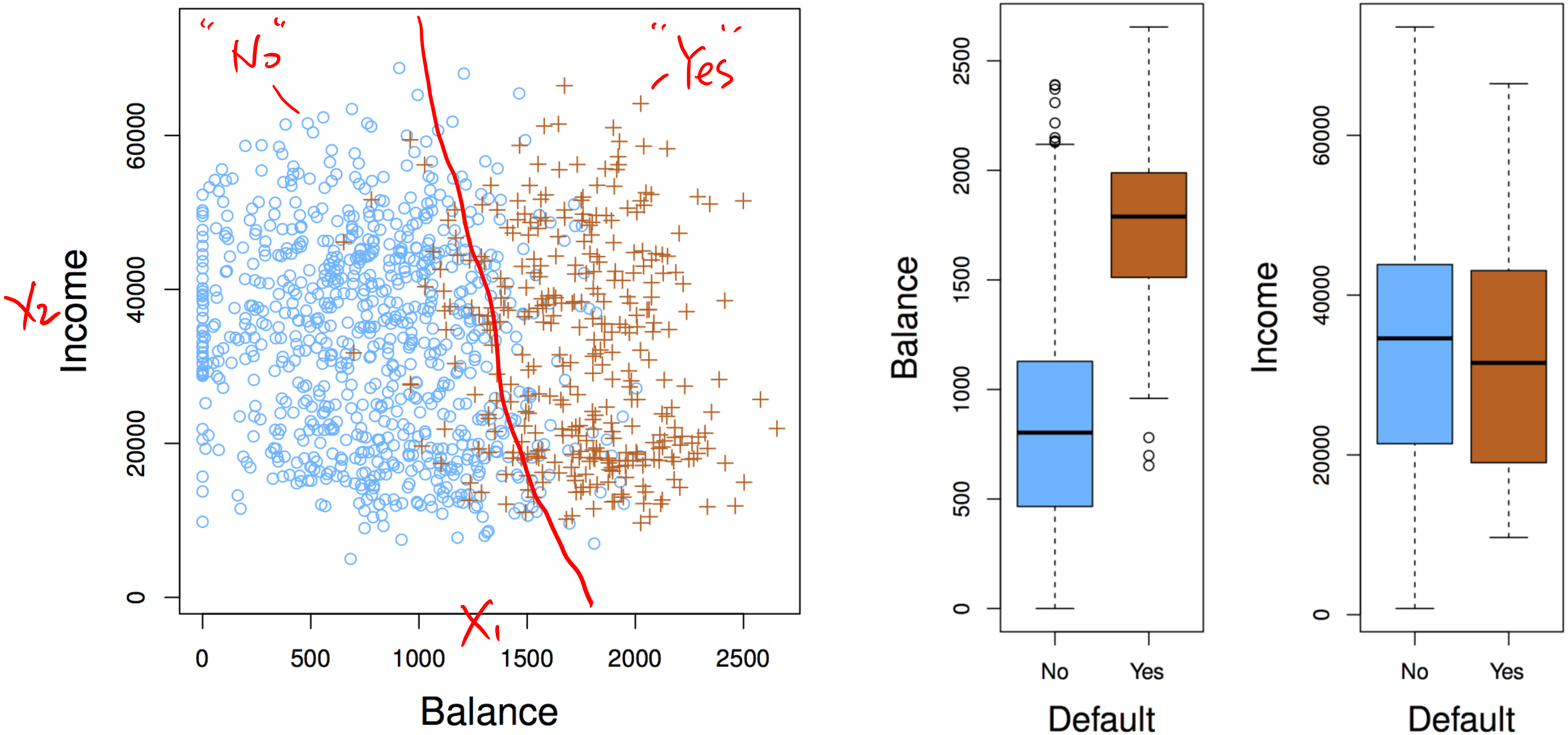
Example: Credit Card Default

- Task: to predict customers default
- The Y variable (Default) is categorical: Yes or No
- Possible X variables are:
  - Annual income
  - Monthly credit card balance
  - Student status

X $\Rightarrow$ Prob (Yes) $\Rightarrow$ Default Yes/No

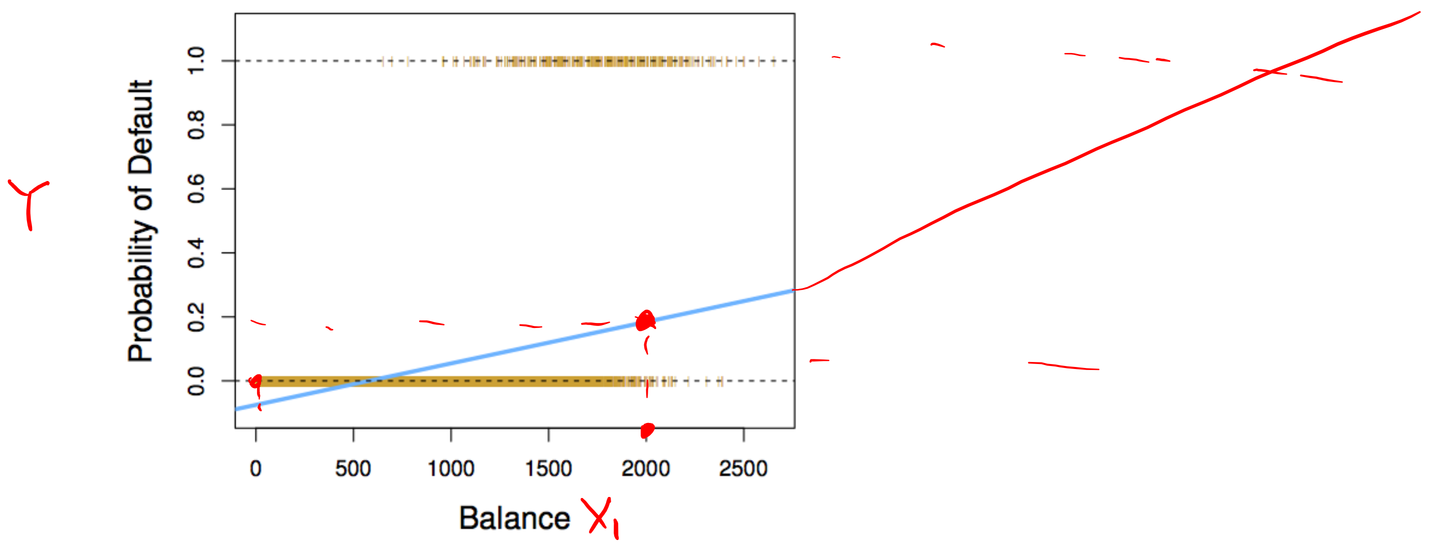
- How do we check the relationship between Y and X?

Credit Card Default – visualization



Why not Linear Regression?

- If we convert the categorical variable Default={Yes, No} into a numerical {1, 0}
- Linear regression works here (for predicting **probability of default**)
- but it is awkward



- What about more than two categories?!

Logistic Regression for Binary Classification

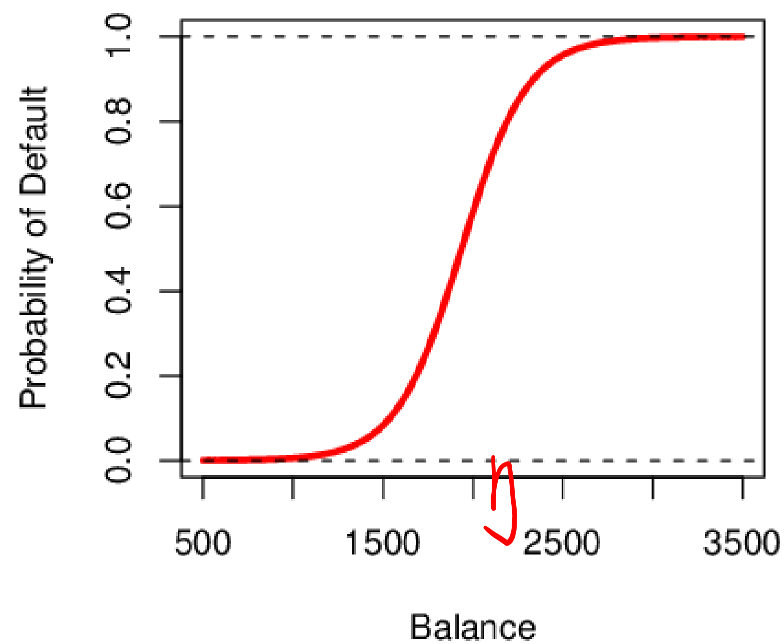
η: signal

- The key issue is that Prob[Default] is in [0, 1], while $\boxed{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p}$ can take any value in (-∞, +∞)
- Need to a proper mapping from (-∞, +∞) to [0, 1]

- The Logistic function

$\in [0,1]$

$$\mathrm{Prob}(Y=1) = \frac{\exp(\overbrace{\boxed{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}^{\eta})}{1+\exp(\underbrace{\boxed{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}_{\eta})}$$



- A.k.a. the “logit” choice model!

	P	→ odds	→ log-odds
coin:	0.5	1	0
dice	1/6	1/5	log 1/5
range:	[0,1]	[0,∞)	(-∞,∞)

Logistic and Logit Function

- Let η = $\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$ be the *linear predictor*

Forward: prediction

Backward: Estimation

Logistic function (η)

(-∞,∞) [0, 1]

η → Prob

$$p = \frac{e^{\eta}}{1+e^{\eta}}$$

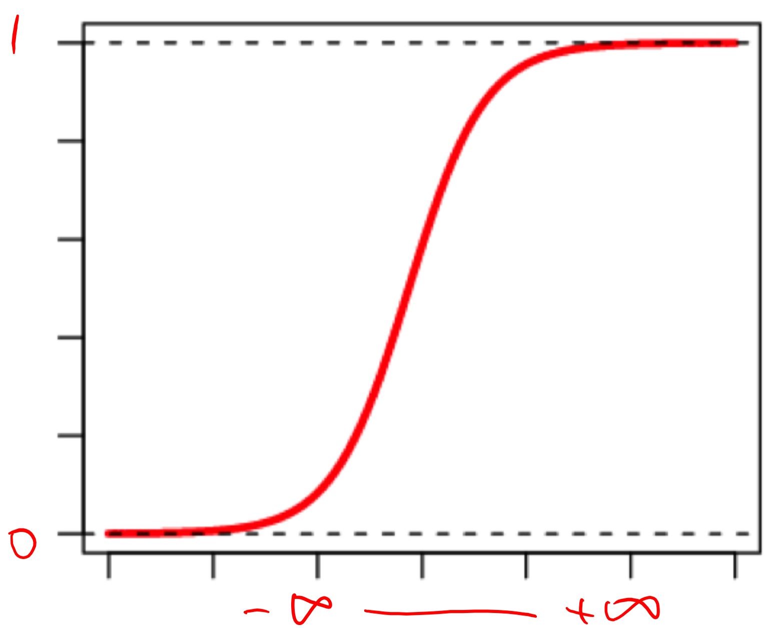
inverse

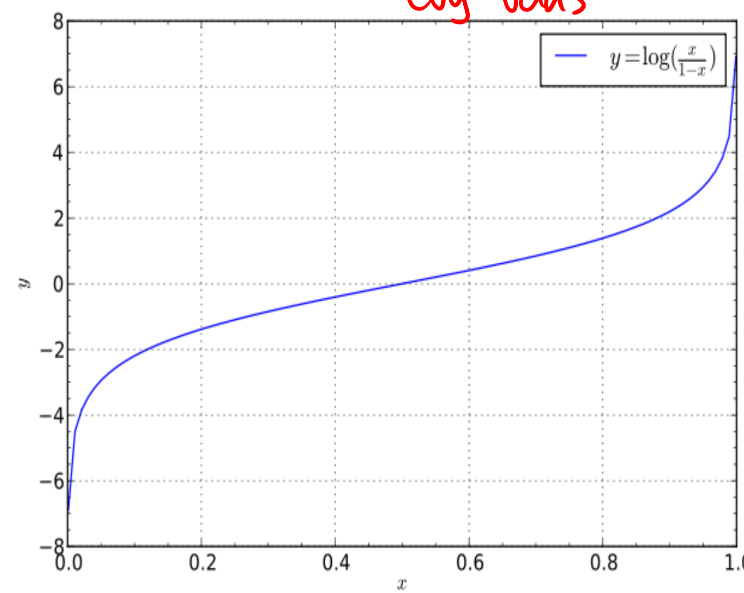
Logit function (p)

$$\eta = \underbrace{\log\left(\boxed{\frac{p}{1-p}}\right)}_{\text{log-odds}}$$

odds

Prob → η





Credit Card Default – fit logistic regression

- Formulation as a logistic regression

prediction : $X, \hat{\beta} \mapsto \hat{p}$

$$\text{Prob(Default)} = \frac{\exp(\beta_0 + \beta_1 \text{Balance})}{1 + \exp(\beta_0 + \beta_1 \text{Balance})}$$

- Estimation can be done by Maximum Likelihood Estimation (MLE)

Estimation

$X, Y \longrightarrow \hat{\beta}$

"most likely"

Credit Card Default – use glm()

default family : family = gaussian ( ) → lm)

- Logistic regression in R using glm()

Binary Classification

Y X

```
Call:
glm(formula = default ~ balance, family = binomial(link = "logit"),
    data = Default)
```

Logistic

lm Residual

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2697  -0.1465  -0.0589  -0.0221   3.7589
```

$\hat{\beta}$ SE($\hat{\beta}$) z-stat P-value Significance level

β_0 β_1

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

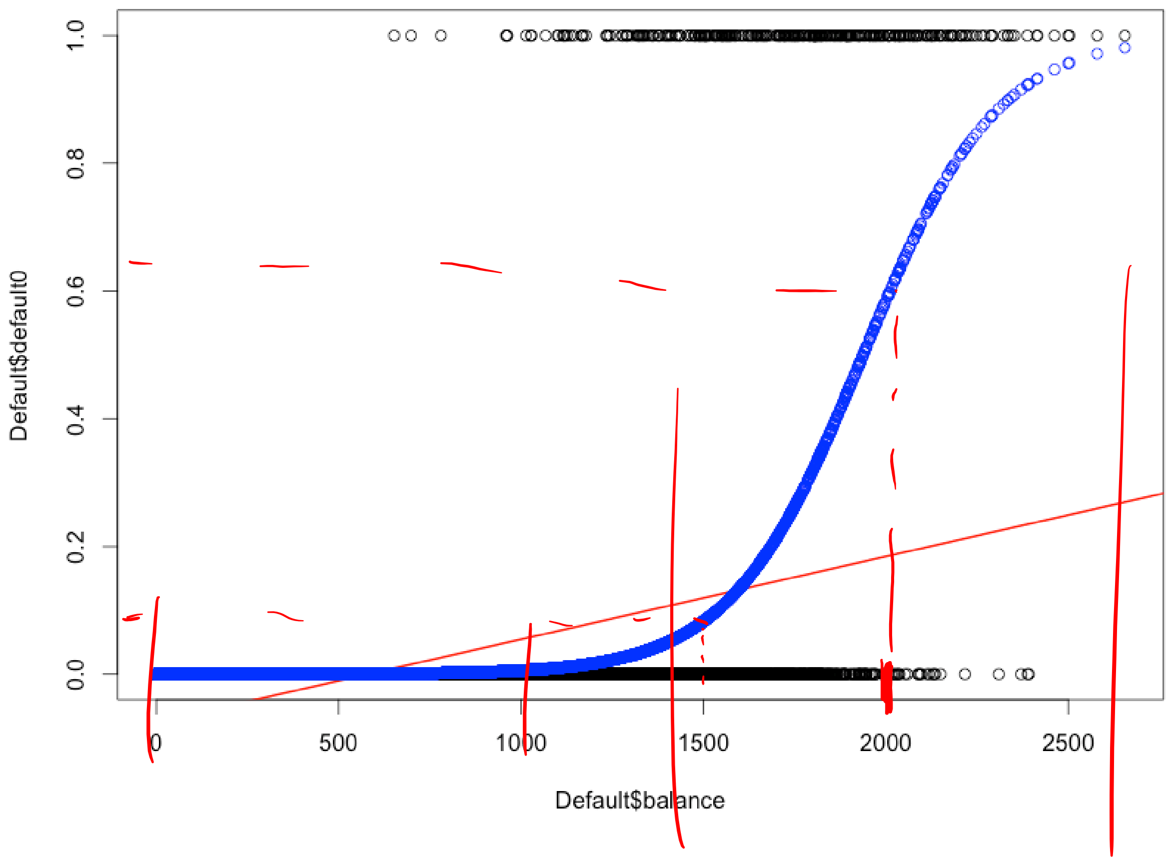
TSS RSS ~ Adj. R^2

```
    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1596.5  on 9998  degrees of freedom
AIC: 1600.5

Number of Fisher Scoring iterations: 8
```

→ Null model: $Y \sim 1$

Credit Card Default – fitted values



Credit Card Default – interpretation of β

insignificant

- If $\beta_1 = 0$, this means that there is no relationship between Y and X.
- If $\beta_1 > 0$, this means that when X gets larger so does the probability that Y = 1.
- If $\beta_1 < 0$, this means that when X gets larger, the probability that Y = 1 gets smaller.

directional

- But how much bigger or smaller depends on where we are on the slope, it is not linear

Credit Card Default – prediction

- Suppose an individual has a balance of $1000. What is the probability of default?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{\overset{\hat{\beta}_0}{-10.6513}+\overset{\hat{\beta}_1}{0.0055}\times \overset{X}{1000}}}{1+e^{-10.6513+0.0055\times 1000}} = 0.00576$$

- The predicted probability of default for an individual with a balance of $1000 is less than 1%.

- For a balance of $2000, the probability is much higher, and equals to 0.586 (58.6%).

Qualitative Predictors in Logistic Regression

- default ~ student

```
Call:
glm(formula = default ~ student, family = binomial(link = "logit"),
    data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2970  -0.2970  -0.2434  -0.2434   2.6585

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.50413    0.07071  -49.55  < 2e-16 ***
studentYes   0.40489    0.11502    3.52 0.000431 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 2908.7  on 9998  degrees of freedom
AIC: 2912.7

Number of Fisher Scoring iterations: 6
```

- β_1 is positive: students tend to have higher default probabilities than non-students

Qualitative Predictors in Logistic Regression

- Prediction

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+\overset{\beta_{student}}{0.4049}\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431,$$

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292.$$

Multiple Logistic Regression

- Default ~ Balance + Income + Student

```
Call:
glm(formula = default ~ balance + income + student, family = binomial(link = "logit"),
    data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4691  -0.1418  -0.0557  -0.0203   3.7383

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1571.5  on 9996  degrees of freedom
AIC: 1579.5

Number of Fisher Scoring iterations: 8
```

Multiple Logistic Regression

- Prediction: a student with a credit card balance of \$1,500 and an income of \$40,000 has an estimated probability of default

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1500+0.003\times 40-0.6468\times 1}}{1+e^{-10.869+0.00574\times 1500+0.003\times 40-0.6468\times 1}} = 0.058.$$

Logistic Regression – Model Assessment

Null : Y ~ 1
reg : Y ~ X
Satured : perfectly fitted

goodness-of-fit

- Deviance
  - = - 2 * (log-likelihood of the fitted model - log-likelihood of the saturated model)

    → 0
  - Measures the badness-of-fit of the model, the smaller the better
  - Corresponds to residuals in linear regression

- AIC (Akaike Information Criterion)
  - = - 2 * (log-likelihood of the fitted model – p)
  - Badness-of-fit penalized for the number of predictors p, the smaller the better
  - Similar to the idea of adjusted R^2 (but not in terms of %)

Logistic Regression – Two Data Formats

- Every individual instance is one data point (Bernoulli distribution)
  - E.g. Credit Card Default data

	Y	X_3	X_1	X_2
	default	student	balance	income
1	No	No	729.5265	44361.625
2	No	Yes	817.1804	12106.135
3	No	No	1073.5492	31767.139
4	No	No	529.2506	35704.494
5	No	No	785.6559	38463.496
6	No	Yes	919.5885	7491.559

X

$Y \sim \text{Bernoulli}(p) \; \{0, 1\}$

- Grouped instances with counts of 0 and 1 (Binomial distribution)
  - E.g. Titanic data

	X_1	X_2	X_3	Y	
	Class	Sex	Age	No	Yes
1	1st	Male	Child	0	5
2	1st	Male	Adult	118	57
3	1st	Female	Child	0	1
4	1st	Female	Adult	4	140
5	2nd	Male	Child	0	11
6	2nd	Male	Adult	154	14
7	2nd	Female	Child	0	13
8	2nd	Female	Adult	13	80
9	3rd	Male	Child	35	13
10	3rd	Male	Adult	387	75
11	3rd	Female	Child	17	14

X

$Y \sim \text{Binomial}(n, p)$

Case-Control Sampling

97% No
3% Yes

- When Yes/No are very unbalanced in the data (typically *p* is small, much more No than Yes)

1 : 4

3% 12%

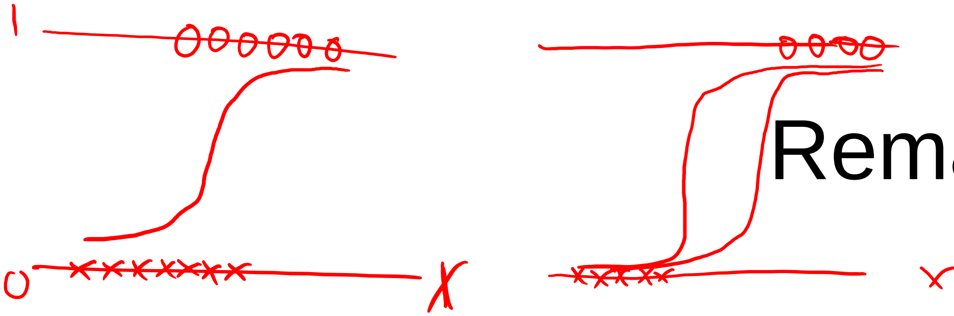
- Case-Control sampling: keep all the rare Yes (case), up to 5 times of No (control)
  - MLE of β_j for j = 1, …, p are not affected
  - Only the MLE of β_0 needs to be adjusted
    - Suppose π is the true proportion of Yes, π_tilde is the proportion in data

3% 20%

$$\hat{\beta}_0^* = \underbrace{\hat{\beta}_0}_{\text{Logistic regression}} + \log \frac{\pi}{1-\pi} - \log \frac{\tilde{\pi}}{1-\tilde{\pi}}$$

True

- Implication in experiment design



Remarks

- Issue when Yes/No are very separable by some predictors $\approx$ multicollinearity
  - "Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred"

Prob("1") < 5%

- Estimation bias in very unbalanced data
  - MLE for logistic regression is biased
  - Bias increases in the unbalanceness

- Multinomial logistic regression (the "glmnet" R package)

$$(R, G, B) \Longrightarrow \begin{cases} (R, 0) \\ (G, 0) \\ (B, 0) \end{cases}$$

$\hat{p} \rightarrow \hat{y}$

cutoff

$\left.\begin{matrix}\hat{y}\\ y\end{matrix}\right\} \Rightarrow Err(y, \hat{y})$

Classification

- After obtaining the predicted probabilities, we still need to make categorical predictions (map probabilities to categories)
- The naïve way: Y = Yes if Prob(Y = Yes) >= 0.5 cutoff $\in [0,1]$
- Evaluation of a (binary) classification method
  - Misclassification rate
  - Confusion matrix

Actual Class

Predicted Class	Y=1	Y=0
Ŷ=1	True Positive ✓	False Positive — Type-I
Ŷ=0	False Negative — Type-II	True Negative ✓

Classification

$Err(y, \hat{y}) \leftrightarrow Err(TP, TN, FP, FN)$

- Other performance measures
  - True positive rate (sensitivity, recall): P(Ŷ = 1 | Y = 1) = TP / (TP + FN)
  - False positive rate (fall-out): P(Ŷ = 1 | Y = 0) = FP / (FP + TN)
  - True negative rate (specificity): P(Ŷ = 0 | Y = 0) = TN / (FP + TN)
  - False negative rate (miss): P(Ŷ = 0 | Y = 1) = FN / (TP + FN)

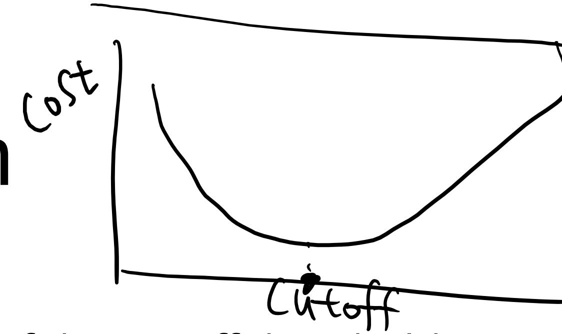
  - Precision: P(Y = 1 | Ŷ = 1) = TP / (TP + FP)
  - Accuracy: P(Ŷ = Y) = (TP + TN) / Total
  - Misclassification rate: 1 - Accuracy
  - Lift: P(Ŷ = 1 | Y = 1) / P(Ŷ = 1)

  - Cost: cost.fp * FP + cost.fn * FN

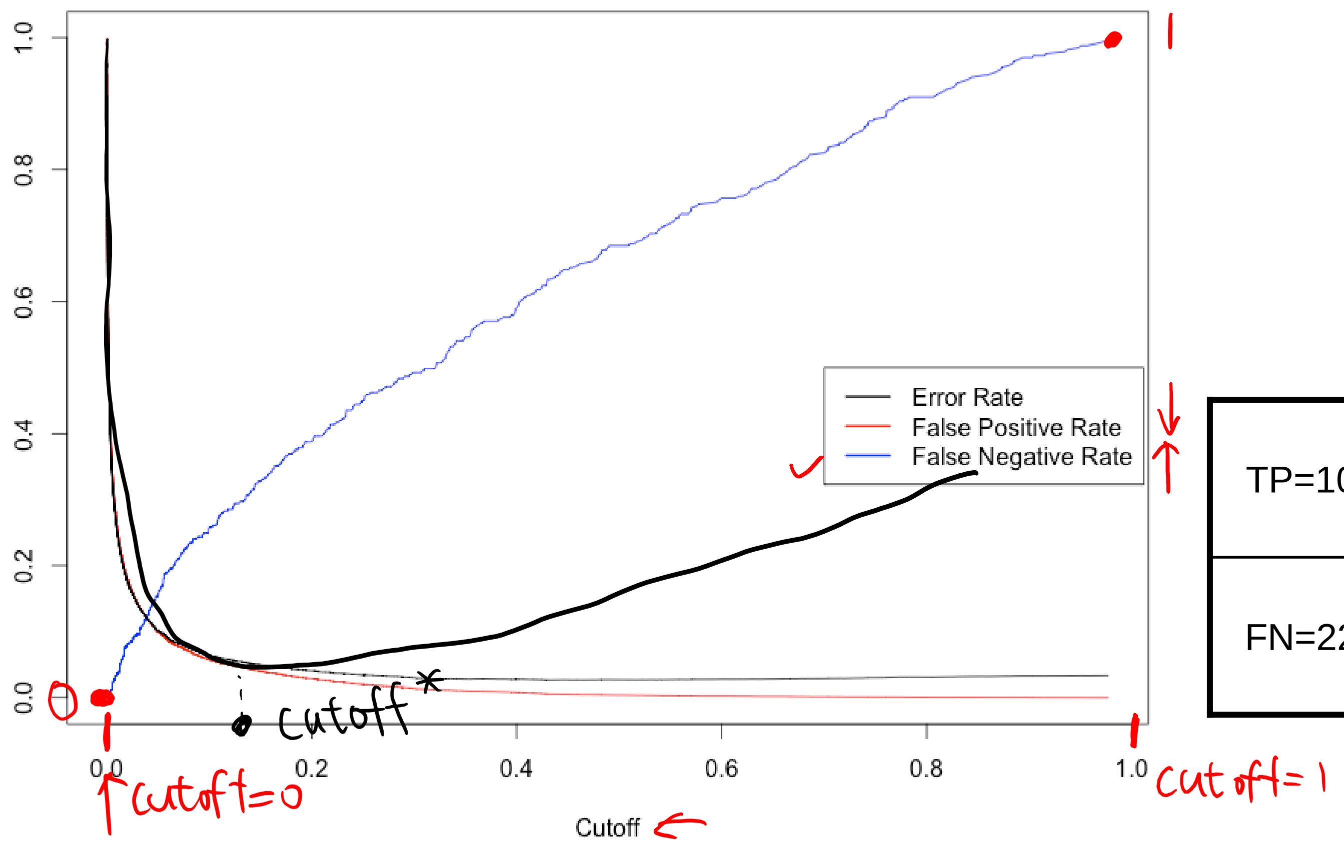
	Y=1	Y=0
Ŷ=1	TP	FP
Ŷ=0	FN	TN

- The measures are determined by the magic cutoff threshold 0.5!

Classification



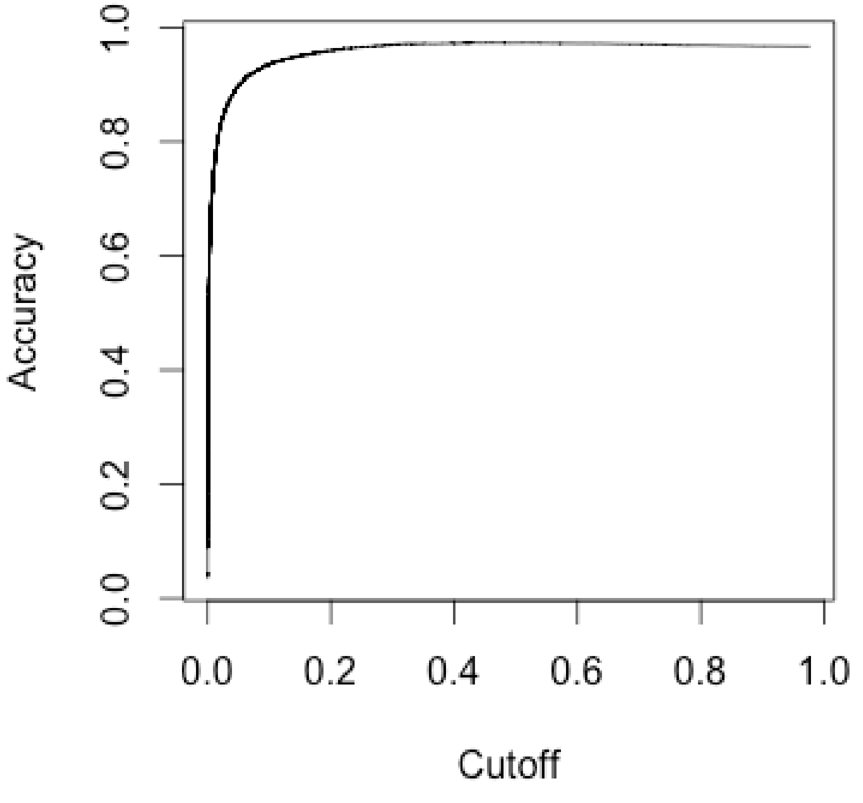
- A better way is to plot the measures as functions of the cutoff threshold
- E.g.: Misclassification rate, false positive rate, false negative rate vs. cutoff
  - False positive and false negative may incur very different costs!

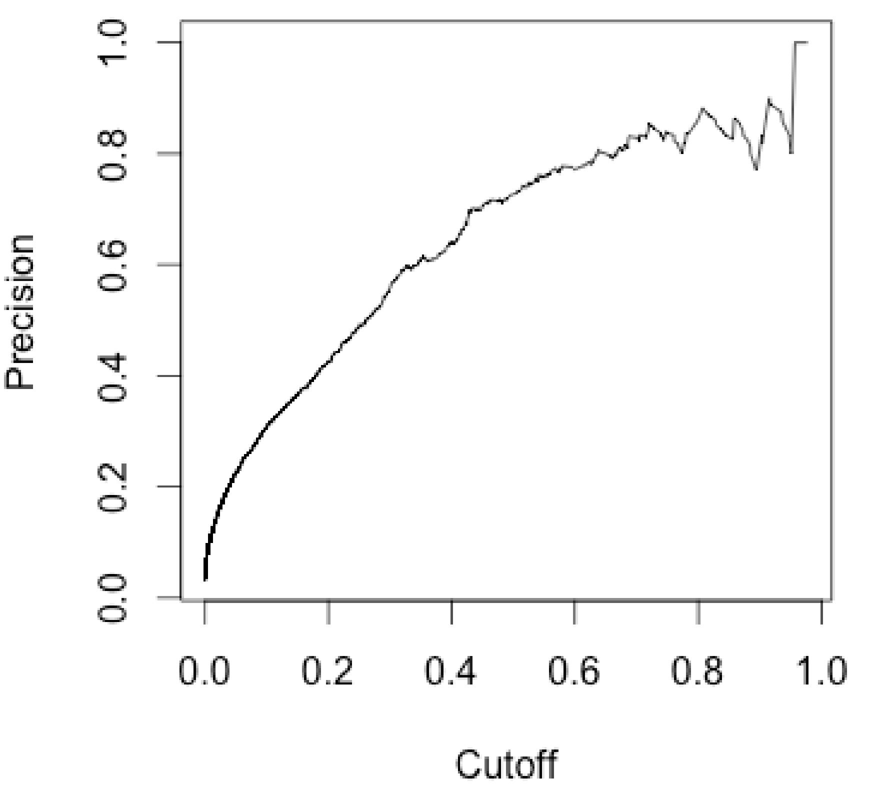


TP=105	FP=39
FN=228	TN=9628

Classification

- Accuracy vs. cutoff and precision vs. cutoff plots

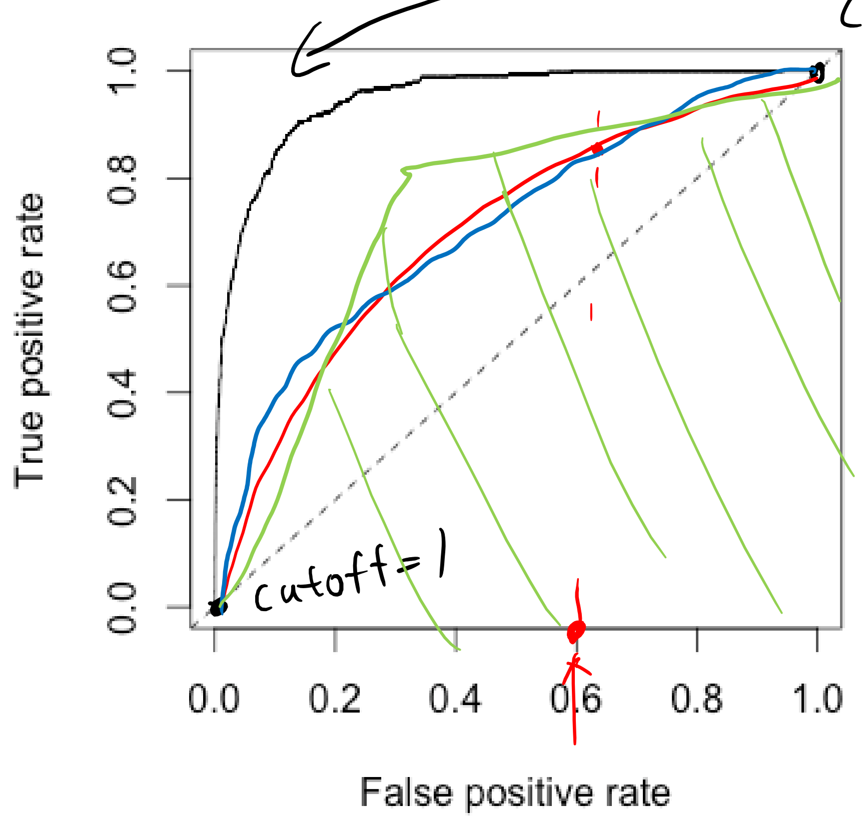




Classification

- We can even plot the trade-off between measures as functions of the cutoff
- ROC (Receiver Operating Characteristic)
- AUC (Area Under ROC): overall performance measure, the higher the better

cutoff=0



$E[Y] = p \cdot 1 + (1-p) \cdot 0 = p$

Generalized Linear Models

- 1-1 Correspondence between linear regression and logistic regression

family

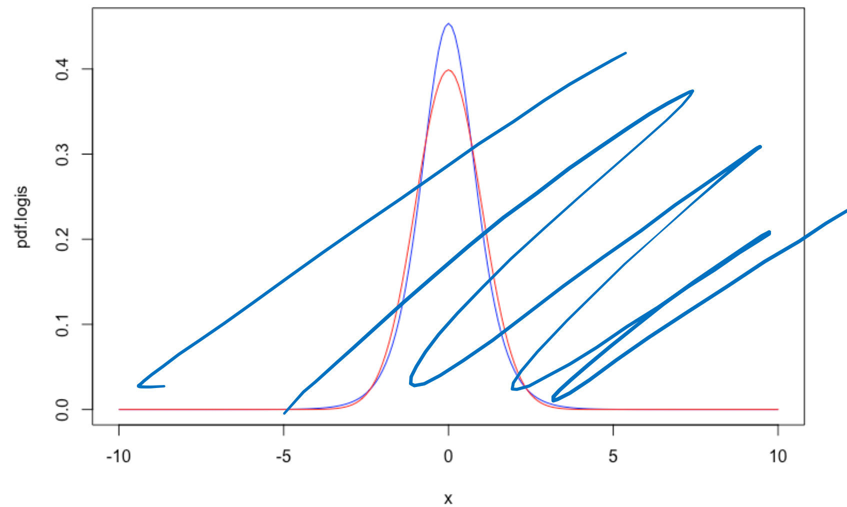
Model	Y	E[Y]	Predictors
Linear Regression (Normal)	$Y = \mu + \varepsilon$ $Y \sim \text{Normal}(\mu, \sigma^2)$	$\boxed{E[Y] = \mu} = \eta$	$\eta = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$
Logistic Regression (Bernoulli)	$Y \sim \text{Bernoulli}(p)$	$\boxed{E[Y] = p} = \dfrac{e^\eta}{1+e^\eta}$	$\eta = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$

Link

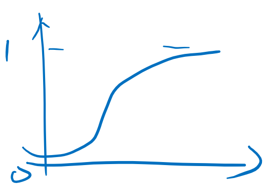
Generalized Linear Models

- Generalize the "link"
  - Logit

$$p = \frac{e^{\eta}}{1 + e^{\eta}}$$



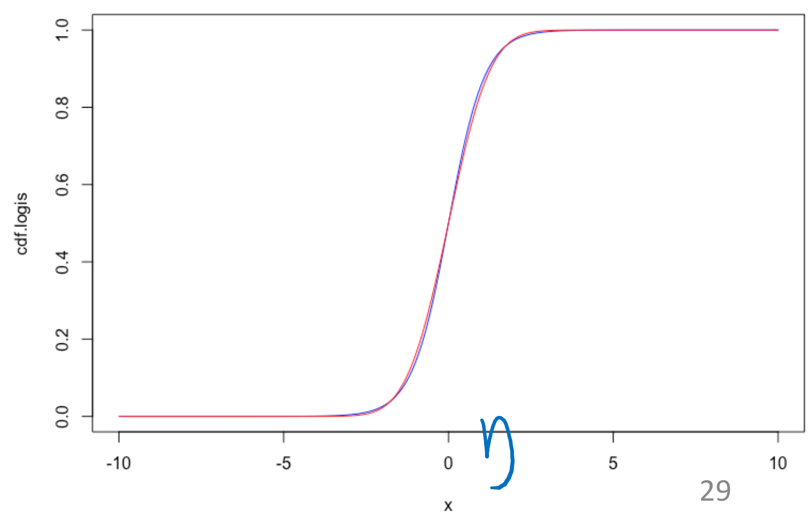
  - Probit



$$p = \Phi(\eta)$$

↳ CDF of normal

P



Generalized Linear Models

- Logit vs. Probit link on the credit default dataset

$$P = \frac{\exp(\eta)}{1+\exp(\eta)}$$

```
Call:
glm(formula = default ~ balance + student, family = binomial(link = "logit"),
    data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4578  -0.1422  -0.0559  -0.0203   3.7435

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.075e+01  3.692e-01 -29.116  < 2e-16 ***
balance      5.738e-03  2.318e-04  24.750  < 2e-16 ***
studentYes  -7.149e-01  1.475e-01  -4.846 1.26e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1571.7  on 9997  degrees of freedom
AIC: 1577.7

Number of Fisher Scoring iterations: 8
```

```
Call:
glm(formula = default ~ balance + student, family = binomial(link = "probit"),
    data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2056  -0.1353  -0.0322  -0.0044   4.1374

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.3918818  0.1728730 -31.190  < 2e-16 ***
balance      0.0028215  0.0001138  24.784  < 2e-16 ***
studentYes  -0.3429201  0.0743964  -4.609 4.04e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1583.5  on 9997  degrees of freedom
AIC: 1589.5

Number of Fisher Scoring iterations: 8
```

Generalized Linear Models

*

- More generally

Family	Support	Typical Use	Link	Link Inverse
Gaussian (Normal)	(-∞, +∞)	Linear response	Identity: $\eta = \mu$	$\mu = \eta$
Bernoulli Binomial	{0, 1} 0, 1, …, N	Count of categories	Logit: $\eta = \log\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\eta)}{1+\exp(\eta)}$
			Probit: $\eta = \Phi^{-1}(\mu)$	$\mu = \Phi(\eta)$
Poisson NegBinomial	0, 1, 2, …	Count of occurrences	Log: $\eta = \log(\mu)$	μ $[0, \infty)$ $\mu = \exp(\eta)$ η $(-\infty, \infty)$
Exponential Gamma	(0, +∞)	Time until an event occurs	Inverse: $\eta = -1/\mu$	$\mu = -1/\eta$

Poisson Regression

- Count data
  - Discrete, skewed distribution
  - High proportion of zeros
  - Nonnegative

- Linear regression won't work
  - Nonnegativity
  - Heteroskedasticity (recall that $E[Y] = \lambda$ and $Var[Y] = \lambda$ for Poisson variable Y)
  - Actually, large Poisson is approximately normal, but not for small ones

- Link: log()
  - $E[Y] = \lambda = \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)$

Poisson Regression

- Goals in English Premier League
  - Number of goals in 150 matches, aggregated by $1^{st}/2^{nd}$ half and Home/Away

```
> head(data, n=10)
   match_id   half home goal
1      2530  First Home    1
2      2530  First Away    0
3      2530 Second Home    0
4      2530 Second Away    0
5      2533  First Home    0
6      2533  First Away    0
7      2533 Second Home    2
8      2533 Second Away    1
9      2534  First Home    0
10     2534  First Away    0
```

```
> summary(data)
    match_id        half         home          goal
 2530   :  4   First :300   Home:300   Min.   :0.0000
 2533   :  4   Second:300   Away:300   1st Qu.:0.0000
 2534   :  4                           Median :0.0000
 2656   :  4                           Mean   :0.6017
 2664   :  4                           3rd Qu.:1.0000
 2665   :  4                           Max.   :3.0000
 (Other):576
```

Poisson Regression

- Effects on goals

```
Call:         Y       X1     X2
glm(formula = goal ~ half + home, family = poisson(), data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1881  -1.0075  -0.9991   0.6095   2.3998

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.67823    0.09668  -7.015 2.29e-12 ***
halfSecond   0.32983    0.10670   3.091  0.00199 **
homeAway    -0.01662    0.10527  -0.158  0.87454
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 609.50  on 599  degrees of freedom
Residual deviance: 599.79  on 597  degrees of freedom
AIC: 1209.9

Number of Fisher Scoring iterations: 5
```

Poisson Regression

$E[Y] = \exp(\beta_0 + \beta_{2nd} \cdot \text{2nd half})$

- Interpreting the coefficients
  - Significant coefficients: $\beta_0 = -0.678$, $\beta[2^{nd}\text{ half}] = 0.329$
  - So $E[Y \mid 1st] = \exp(\beta_0)$, and $E[Y \mid 2nd] = \exp(\beta_0 + 0.329)$
  - => $E[Y \mid 2nd] / E[Y \mid 1st] = \exp(0.329) = 1.390$
  - => on average, there are 39% more goals in the 2^{nd} half than in the 1^{st} half
  - β[Away] is not significant => no difference in average goals between home and away teams

Poisson Regression

- Model assessment
  - AIC

- Also need to check model assumption
  - Residual deviance / degree of freedom is approximately 1

- If not, it means overdispersion
  - Use family=quasipoisson() instead
  - Use Negative Binomial regression

Survival Analysis

censoring

$\Downarrow$

id	time	failure
1	2	1
2	4	1
3	10	0
4	5	0

- Model of time to an event (typically with censoring)
  - Time to death for patients
  - Time to failure for machines

- Exponential/Gamma regression

- Check out the "survival" package in R