Predictive Analytics Lecture 4

Adam Kapelner

Stat 422/722
at The Wharton School of the University of Pennsylvania

February 7 & 8, 2017

Review: Generalizing the Classification Rule

Recall the classification rule $\hat{y} = \mathbb{1}_{\hat{p} \geq 0.5}$. Using 0.5 is a principled default but we can use any rule $p_0 \in (0, 1)$:

$$\hat{y} = \mathbb{1}_{\hat{p} \geq p_0} := \begin{cases} 1 & \text{if} \quad \hat{p} \geq p_0 \\ 0 & \text{if} \quad \hat{p} < p_0 \end{cases}$$

What happens when we change the p_0 threshold? If $p_0 \uparrow \quad \Rightarrow \quad \hat{P} \downarrow$ and $\hat{N} \uparrow$. If $p_0 \downarrow \quad \Rightarrow \quad \hat{P} \uparrow$ and $\hat{N} \downarrow$. Changing p_0 changes the column totals and obviously creates a whole new confusion matrix.

So now it's simple, vary p_0 and pick the best model according to your cost / error / loss function (the *ME* at the moment). Let's just do every p_0!

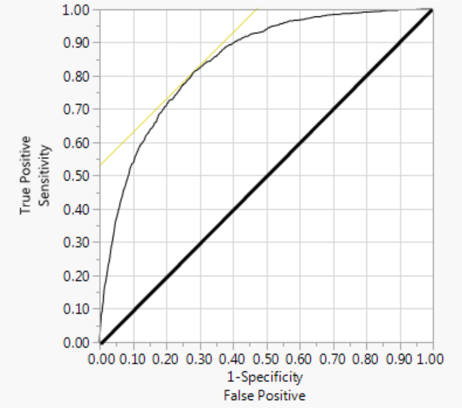
All Possible Confusion Matrices

ROC Table

Prob	1-Specificity	Sensitivity	Sens-(1-Spec)	True Pos	True Neg	False Pos	False Neg
.	0.0000	0.0000	0.0000	0	5163	0	1869
0.8117	0.0000	0.0005	0.0005	1	5163	0	1868
0.8104	0.0000	0.0011	0.0011	2	5163	0	1867
0.8093	0.0000	0.0016	0.0016	3	5163	0	1866
0.8092	0.0000	0.0021	0.0021	4	5163	0	1865
0.8090	0.0000	0.0027	0.0027	5	5163	0	1864
0.8085	0.0000	0.0032	0.0032	6	5163	0	1863
0.8083	0.0000	0.0037	0.0037	7	5163	0	1862
0.8082	0.0000	0.0043	0.0043	8	5163	0	1861
0.8079	0.0000	0.0048	0.0048	9	5163	0	1860
0.8079	0.0000	0.0054	0.0054	10	5163	0	1859
0.8077	0.0000	0.0059	0.0059	11	5163	0	1858
0.8076	0.0002	0.0059	0.0057	11	5162	1	1858
0.8072	0.0002	0.0064	0.0062	12	5162	1	1857
0.8065	0.0002	0.0070	0.0068	13	5162	1	1856
0.8064	0.0002	0.0075	0.0073	14	5162	1	1855
0.8061	0.0002	0.0080	0.0078	15	5162	1	1854

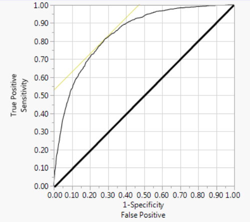
Here, `Prob` is what we denoted p_0.

Receiver-Operator Characteristic Curve



The **ROC Curve**. Each dot represents the sensitivity-specificity tradeoff for each p_0. The starred row of maximum sensitivity + specificity is indicated here by a yellow tangent line.

Area Under the Curve (AUC) Metric



If you built a model by chance the “area under the curve” (or to the right of the curve) on the graph would be ... 0.5 since the graph is a unit square. Under the ROC curve itself (or to its right) is an area ... greater than 0.5. Here, it's 0.844. This metric is called AUC and is widely used as a metric to assess performance of all possible classifiers in this set of models together, it is a composite metric unlike *ME* or anything derived from an individual confusion table.

AUC is nice to evaluate overall performance of all possible models... but at the end of the day... you ship **ONE** model! So we still need a means of evaluating our one model from one confusion table.

Churn Example Where $p_0 = 0.10$

Churn Example Where $p_0 = 0.10$

$p_0 = 0.5$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1012$	$FN = 857$	$P = 1869$	$FNR = 45.9\%$
	0	$FP = 531$	$TN = 4632$	$N = 5163$	$FPR = 10.2\%$
	Totals	$\hat{P} = 1543$	$\hat{N} = 5489$	$n = 7032$	
	Use errors	$FDR = 34.3\%$	$FOR = 15.6\%$		$\boxed{ME = 19.7\%}$

Churn Example Where $p_0 = 0.10$

$p_0 = 0.5$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1012$	$FN = 857$	$P = 1869$	$FNR = 45.9\%$
	0	$FP = 531$	$TN = 4632$	$N = 5163$	$FPR = 10.2\%$
	Totals	$\hat{P} = 1543$	$\hat{N} = 5489$	$n = 7032$	
	Use errors	$FDR = 34.3\%$	$FOR = 15.6\%$		$\boxed{ME = 19.7\%}$

$p_0 = 0.1$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	FP $= 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Churn Example Where $p_0 = 0.10$

$p_0 = 0.5$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1012$	$FN = 857$	$P = 1869$	$FNR = 45.9\%$
	0	$FP = 531$	$TN = 4632$	$N = 5163$	$FPR = 10.2\%$
	Totals	$\hat{P} = 1543$	$\hat{N} = 5489$	$n = 7032$	
	Use errors	$FDR = 34.3\%$	$FOR = 15.6\%$		$\boxed{ME = 19.7\%}$

$p_0 = 0.1$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	$FP = 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Which numbers did not change?

Churn Example Where $p_0 = 0.10$

$p_0 = 0.5$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1012$	$FN = 857$	$P = 1869$	$FNR = 45.9\%$
	0	$FP = 531$	$TN = 4632$	$N = 5163$	$FPR = 10.2\%$
	Totals	$\hat{P} = 1543$	$\hat{N} = 5489$	$n = 7032$	
	Use errors	$FDR = 34.3\%$	$FOR = 15.6\%$		$\boxed{ME = 19.7\%}$

$p_0 = 0.1$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	FP $= 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Which numbers did not change? n, P and N. Why?

Churn Example Where $p_0 = 0.10$

$p_0 = 0.5$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1012$	$FN = 857$	$P = 1869$	$FNR = 45.9\%$
	0	$FP = 531$	$TN = 4632$	$N = 5163$	$FPR = 10.2\%$
	Totals	$\hat{P} = 1543$	$\hat{N} = 5489$	$n = 7032$	
	Use errors	$FDR = 34.3\%$	$FOR = 15.6\%$		$\boxed{ME = 19.7\%}$

$p_0 = 0.1$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	FP $= 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Which numbers did not change? n, P and N. Why? These are fixed according to the dataframe. All other numbers changed! What happend to our first means of evaluation, the Misclassification Error?

Churn Example Where $p_0 = 0.10$

$p_0 = 0.5$		$\hat{y}$: 1	$\hat{y}$: 0	Totals	Model Errors
y	1	$TP = 1012$	$FN = 857$	$P = 1869$	$FNR = 45.9\%$
	0	$FP = 531$	$TN = 4632$	$N = 5163$	$FPR = 10.2\%$
Totals		$\hat{P} = 1543$	$\hat{N} = 5489$	$n = 7032$	
Use errors		$FDR = 34.3\%$	$FOR = 15.6\%$		$\boxed{ME = 19.7\%}$

$p_0 = 0.1$		$\hat{y}$: 1	$\hat{y}$: 0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	FP $= 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
Totals		$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
Use errors		$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Which numbers did not change? n, P and N. Why? These are fixed according to the dataframe. All other numbers changed! What happend to our first means of evaluation, the Misclassification Error? It increased from 19.7% $\rightarrow$ 39.3%. So isn't this a worse model??

Churn Example Where $p_0 = 0.10$

$p_0 = 0.5$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1012$	$FN = 857$	$P = 1869$	$FNR = 45.9\%$
	0	$FP = 531$	$TN = 4632$	$N = 5163$	$FPR = 10.2\%$
	Totals	$\hat{P} = 1543$	$\hat{N} = 5489$	$n = 7032$	
	Use errors	$FDR = 34.3\%$	$FOR = 15.6\%$		$\boxed{ME = 19.7\%}$

$p_0 = 0.1$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	FP $= 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Which numbers did not change? n, P and N. Why? These are fixed according to the dataframe. All other numbers changed! What happend to our first means of evaluation, the Misclassification Error? It increased from 19.7% $\rightarrow$ 39.3%. So isn't this a worse model??

Not necessarily... It depends on what your goal is!

Asymmetric Costs in a Classifier

These are always two types of errors but the costs are not always the same.

$p_0 = 0.1$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	FP $= 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Imagine we really are the Telecom business manager. It costs 5-10x more to acquire a new customer than to engage a customer who is likely to churn. So you give an incentive package to those who are predicted to churn.

Asymmetric Costs in a Classifier

These are always two types of errors but the costs are not always the same.

$p_0 = 0.1$		$\hat{y}$ = 1	$\hat{y}$ = 0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	$FP = 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Imagine we really are the Telecom business manager. It costs 5-10x more to acquire a new customer than to engage a customer who is likely to churn. So you give an incentive package to those who are predicted to churn. Which of the two types of errors specifically is *very* costly?

Asymmetric Costs in a Classifier

These are always two types of errors but the costs are not always the same.

$p_0 = 0.1$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	FP $= 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Imagine we really are the Telecom business manager. It costs 5-10x more to acquire a new customer than to engage a customer who is likely to churn. So you give an incentive package to those who are predicted to churn. Which of the two types of errors specifically is *very* costly? The *FN*. Who are they?

Asymmetric Costs in a Classifier

These are always two types of errors but the costs are not always the same.

$p_0 = 0.1$		$\hat{y}$: 1	$\hat{y}$: 0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	$FP = 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Imagine we really are the Telecom business manager. It costs 5-10x more to acquire a new customer than to engage a customer who is likely to churn. So you give an incentive package to those who are predicted to churn. Which of the two types of errors specifically is *very* costly? The *FN*. Who are they? These are those who you said were not going to churn *and they did*! Cost?

Asymmetric Costs in a Classifier

These are always two types of errors but the costs are not always the same.

	$p_0 = 0.1$	$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	$FP = 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Imagine we really are the Telecom business manager. It costs 5-10x more to acquire a new customer than to engage a customer who is likely to churn. So you give an incentive package to those who are predicted to churn. Which of the two types of errors specifically is *very* costly? The *FN*. Who are they? These are those who you said were not going to churn *and they did*! Cost? You need to acquire a new customer! The other type of error is less costly, the *FP*. Who are they?

Asymmetric Costs in a Classifier

These are always two types of errors but the costs are not always the same.

$p_0 = 0.1$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	$FP = 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Imagine we really are the Telecom business manager. It costs 5-10x more to acquire a new customer than to engage a customer who is likely to churn. So you give an incentive package to those who are predicted to churn. Which of the two types of errors specifically is *very* costly? The *FN*. Who are they? These are those who you said were not going to churn *and they did*! Cost? You need to acquire a new customer! The other type of error is less costly, the *FP*. Who are they? These are the people you thought were going to churn and did not. Cost?

Asymmetric Costs in a Classifier

These are always two types of errors but the costs are not always the same.

$p_0 = 0.1$		$\hat{y}$ 1	0	Totals	Model Errors
y	1	$TP = 1772$	$FN = 97$	$P = 1869$	$FNR = 5.1\%$
	0	$FP = 2669$	$TN = 2494$	$N = 5163$	$FPR = 51.6\%$
	Totals	$\hat{P} = 4441$	$\hat{N} = 2591$	$n = 7032$	
	Use errors	$FDR = 60.1\%$	$FOR = 3.7\%$		$\boxed{ME = 39.3\%}$

Imagine we really are the Telecom business manager. It costs 5-10x more to acquire a new customer than to engage a customer who is likely to churn. So you give an incentive package to those who are predicted to churn. Which of the two types of errors specifically is *very* costly? The *FN*. Who are they? These are those who you said were not going to churn *and they did*! Cost? You need to acquire a new customer! The other type of error is less costly, the *FP*. Who are they? These are the people you thought were going to churn and did not. Cost? Whatever the incentive package is.

Weighted Misclassification Error

We now define two costs: (1)

Weighted Misclassification Error

We now define two costs: (1) the cost of the *FP* denoted c_{FP} and

Weighted Misclassification Error

We now define two costs: (1) the cost of the *FP* denoted c_{FP} and (2) the cost of the *FN* denoted c_{FN}.

Weighted Misclassification Error

We now define two costs: (1) the cost of the *FP* denoted c_{FP} and (2) the cost of the *FN* denoted c_{FN}. We then define the weighted misclassification error evaluation metric:

$$ME_w := \frac{1}{n}\sum_{i=1}^{n} c_{FP}\mathbb{1}_{y_i=0\&\hat{y}=1} + c_{FN}\mathbb{1}_{y_i=1\&\hat{y}_i=0}$$

Weighted Misclassification Error

We now define two costs: (1) the cost of the *FP* denoted c_{FP} and (2) the cost of the *FN* denoted c_{FN}. We then define the weighted misclassification error evaluation metric:

$$ME_w := \frac{1}{n}\sum_{i=1}^{n} c_{FP}\mathbb{1}_{y_i=0\&\hat{y}=1} + c_{FN}\mathbb{1}_{y_i=1\&\hat{y}_i=0}$$

We now vary p_0 to locate the model that optimizes this error to be minimum.

Minimum Weighted Misclassification Error

Let's assume that $c_{FN} = \$1000$ and $c_{FP} = \$100$ just for the example's sake. Note: this is a **cost ratio** of 10:1 (only the ratio matters for the optimal p_0 solution).

	Prob	TP	TN	FP	FN	COST
1	0.8117	1	5163	0	1868	1868000
2	0.8104	2	5163	0	1867	1867000
3	0.8093	3	5163	0	1866	1866000
4	0.8092	4	5163	0	1865	1865000
5	0.8090	5	5163	0	1864	1864000
6	0.8085	6	5163	0	1863	1863000
7	0.8083	7	5163	0	1862	1862000
8	0.8082	8	5163	0	1861	1861000
9	0.8079	9	5163	0	1860	1860000

We now calculate the cost and find the minimum model (i.e. the p_0 to ship). [JMP] Beyond scope: some people select the model with the closest $\#FP/\#FN \approx 10:1$ to match the stakeholder preference of the desired cost ratio. I'm not entirely clear on why this fitness function is used. [JMP ratios sheet]

Expected Value Calculation

You can also imagine assignment of both costs *and* benefits:

$p_0 = 0.1$		$\hat{y}$	
		1	0
y	1	b_{TP}	c_{FN}
	0	c_{FP}	b_{TN}

Expected Value Calculation

You can also imagine assignment of both costs *and* benefits:

	$p_0 = 0.1$	$\hat{y}$ 1	$\hat{y}$ 0
y	1	b_{TP}	c_{FN}
	0	c_{FP}	b_{TN}

and then use the confusion matrix to estimate probabilities:

	$p_0 = 0.1$	$\hat{y}$ 1	$\hat{y}$ 0
y	1	25.1%	1.3%
	0	40.0%	35.5%

Expected Value Calculation

You can also imagine assignment of both costs *and* benefits:

$p_0 = 0.1$		$\hat{y}$ = 1	$\hat{y}$ = 0
y	1	b_{TP}	c_{FN}
	0	c_{FP}	b_{TN}

and then use the confusion matrix to estimate probabilities:

$p_0 = 0.1$		$\hat{y}$ = 1	$\hat{y}$ = 0
y	1	25.1%	1.3%
	0	40.0%	35.5%

The expected value would be?

$$\begin{aligned}\mathbb{E}[T] &= p_{TP} \times b_{TP} + p_{TN} \times b_{TN} + p_{FP} \times c_{FP} + p_{FN} \times c_{FN} \\ &\approx \hat{p}_{TP} \times b_{TP} + \hat{p}_{TN} \times b_{TN} + \hat{p}_{FP} \times c_{FP} + \hat{p}_{FN} \times c_{FN}\end{aligned}$$

Highest expected value model is shipped (ex. from Provost & Fawcett, 2013).

New Type of Response Metric: TTL

What if your response was time? For example:

New Type of Response Metric: TTL

What if your response was time? For example:

- Time for a patient to live (typical in clinical trials)?

New Type of Response Metric: TTL

What if your response was time? For example:

- Time for a patient to live (typical in clinical trials)?
- How long will a customer be a customer?

New Type of Response Metric: TTL

What if your response was time? For example:

- Time for a patient to live (typical in clinical trials)?
- How long will a customer be a customer?
- How long will a car engine last?

New Type of Response Metric: TTL

What if your response was time? For example:

- Time for a patient to live (typical in clinical trials)?
- How long will a customer be a customer?
- How long will a car engine last?

What kind of data type is the response?

New Type of Response Metric: TTL

What if your response was time? For example:

- Time for a patient to live (typical in clinical trials)?
- How long will a customer be a customer?
- How long will a car engine last?

What kind of data type is the response? Continuous.

New Type of Response Metric: TTL

What if your response was time? For example:

- Time for a patient to live (typical in clinical trials)?
- How long will a customer be a customer?
- How long will a car engine last?

What kind of data type is the response? Continuous. But what does response look like at the time of sampling?

For example, recall the Telecom churn dataset with one feature (tenure). If the observation has ...

Tenure Time	Churn?	Total Time as a Customer
2	Yes	

New Type of Response Metric: TTL

What if your response was time? For example:

- Time for a patient to live (typical in clinical trials)?
- How long will a customer be a customer?
- How long will a car engine last?

What kind of data type is the response? Continuous. But what does response look like at the time of sampling?

For example, recall the Telecom churn dataset with one feature (tenure). If the observation has ...

Tenure Time	Churn?	Total Time as a Customer
2	Yes	2
8	Yes	

New Type of Response Metric: TTL

What if your response was time? For example:

- Time for a patient to live (typical in clinical trials)?
- How long will a customer be a customer?
- How long will a car engine last?

What kind of data type is the response? Continuous. But what does response look like at the time of sampling?

For example, recall the Telecom churn dataset with one feature (tenure). If the observation has ...

Tenure Time	Churn?	Total Time as a Customer
2	Yes	2
8	Yes	8
45	No	

New Type of Response Metric: TTL

What if your response was time? For example:

- Time for a patient to live (typical in clinical trials)?
- How long will a customer be a customer?
- How long will a car engine last?

What kind of data type is the response? Continuous. But what does response look like at the time of sampling?

For example, recall the Telecom churn dataset with one feature (tenure). If the observation has ...

Tenure Time	Churn?	Total Time as a Customer
2	Yes	2
8	Yes	8
45	No	unknown (but known to be > 45)

That third observation's response is **censored**.

New Type of Response Metric: TTL

What if your response was time? For example:

- Time for a patient to live (typical in clinical trials)?
- How long will a customer be a customer?
- How long will a car engine last?

What kind of data type is the response? Continuous. But what does response look like at the time of sampling?

For example, recall the Telecom churn dataset with one feature (tenure). If the observation has ...

Tenure Time	Churn?	Total Time as a Customer
2	Yes	2
8	Yes	8
45	No	unknown (but known to be > 45)

That third observation's response is **censored**. What are we supposed to do??

Dealing with censoring

One option is to disregard all censored observations. Why is this a bad idea?

Dealing with censoring

One option is to disregard all censored observations. Why is this a bad idea? Selection bias. Our results will only apply to people who have churned. Those people may be different that the general population.

Dealing with censoring

One option is to disregard all censored observations. Why is this a bad idea? Selection bias. Our results will only apply to people who have churned. Those people may be different that the general population. Another option is to use **survival modeling**.

Dealing with censoring

One option is to disregard all censored observations. Why is this a bad idea? Selection bias. Our results will only apply to people who have churned. Those people may be different that the general population. Another option is to use **survival modeling**.

This is a very well-studied field with many possible models!

The Exponential Model

Assume Y now is time. Time must be positive!

$$Y = f(x_1, \ldots, x_p) + \mathcal{E}$$

We have to be careful to make f positive and $\mathcal{E}$ negative but not too negative to make $Y < 0$.

The Exponential Model

Assume Y now is time. Time must be positive!

$$Y = f(x_1, \ldots, x_p) + \mathcal{E}$$

We have to be careful to make f positive and $\mathcal{E}$ negative but not too negative to make $Y < 0$. One such model is the exponential model with conditional mean $f(x_1, \ldots, x_p)$.

$$Y \sim Exp(f(x_1, \ldots, x_p))$$

Let's review the exponential r.v.

The Exponential Model

Assume Y now is time. Time must be positive!

$$Y = f(x_1, \ldots, x_p) + \mathcal{E}$$

We have to be careful to make f positive and $\mathcal{E}$ negative but not too negative to make $Y < 0$. One such model is the exponential model with conditional mean $f(x_1, \ldots, x_p)$.

$$Y \sim Exp(f(x_1, \ldots, x_p))$$

Let's review the exponential r.v. If $Y \sim Exp(\mu)$, then its density function and cumulative density functions are

$$p(y) = \frac{1}{\mu} e^{-\frac{1}{\mu} y} \quad \text{and} \quad F(y) = 1 - e^{-\frac{1}{\mu} y}$$

with mean $\mathbb{E}[Y] = \mu$ where $\mu > 0$.

The Exponential Model

Assume Y now is time. Time must be positive!

$$Y = f(x_1, \ldots, x_p) + \mathcal{E}$$

We have to be careful to make f positive and $\mathcal{E}$ negative but not too negative to make $Y < 0$. One such model is the exponential model with conditional mean $f(x_1, \ldots, x_p)$.

$$Y \sim Exp(f(x_1, \ldots, x_p))$$

Let's review the exponential r.v. If $Y \sim Exp(\mu)$, then its density function and cumulative density functions are

$$p(y) = \frac{1}{\mu} e^{-\frac{1}{\mu} y} \quad \text{and} \quad F(y) = 1 - e^{-\frac{1}{\mu} y}$$

with mean $\mathbb{E}[Y] = \mu$ where $\mu > 0$. So if $Y \sim Exp(17)$, you expect the observation to be ...

The Exponential Model

Assume Y now is time. Time must be positive!

$$Y = f(x_1, \dots, x_p) + \mathcal{E}$$

We have to be careful to make f positive and $\mathcal{E}$ negative but not too negative to make $Y < 0$. One such model is the exponential model with conditional mean $f(x_1, \dots, x_p)$.

$$Y \sim Exp(f(x_1, \dots, x_p))$$

Let's review the exponential r.v. If $Y \sim Exp(\mu)$, then its density function and cumulative density functions are

$$p(y) = \frac{1}{\mu} e^{-\frac{1}{\mu} y} \quad \text{and} \quad F(y) = 1 - e^{-\frac{1}{\mu} y}$$

with mean $\mathbb{E}[Y] = \mu$ where $\mu > 0$. So if $Y \sim Exp(17)$, you expect the observation to be ... 17 on average.

The Exponential Log-Linear Model (ELLM)

Let's say we want to use our linear model $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ for the conditional mean. Problem?

The Exponential Log-Linear Model (ELLM)

Let's say we want to use our linear model $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ for the conditional mean. Problem? Yes. The mean can only be a number greater than zero. Hence we need the λ link function again! How can we convert $s_{\mathbb{R}} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ to something between 0 and $+\infty$?

The Exponential Log-Linear Model (ELLM)

Let's say we want to use our linear model $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ for the conditional mean. Problem? Yes. The mean can only be a number greater than zero. Hence we need the λ link function again! How can we convert $s_{\mathbb{R}} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ to something between 0 and $+\infty$? Without going into many different link function, let's just use the natural exponential:

$$s = \lambda(s_{\mathbb{R}}) =$$

The Exponential Log-Linear Model (ELLM)

Let's say we want to use our linear model $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ for the conditional mean. Problem? Yes. The mean can only be a number greater than zero. Hence we need the λ link function again! How can we convert $s_{\mathbb{R}} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ to something between 0 and $+\infty$? Without going into many different link function, let's just use the natural exponential:

$$s = \lambda(s_{\mathbb{R}}) = e^{s_{\mathbb{R}}} \quad \Rightarrow$$

The Exponential Log-Linear Model (ELLM)

Let's say we want to use our linear model $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ for the conditional mean. Problem? Yes. The mean can only be a number greater than zero. Hence we need the λ link function again! How can we convert $s_{\mathbb{R}} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ to something between 0 and $+\infty$? Without going into many different link function, let's just use the natural exponential:

$$s = \lambda(s_{\mathbb{R}}) = e^{s_{\mathbb{R}}} \quad \Rightarrow \quad s = e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}$$

The Exponential Log-Linear Model (ELLM)

Let's say we want to use our linear model $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ for the conditional mean. Problem? Yes. The mean can only be a number greater than zero. Hence we need the λ link function again! How can we convert $s_{\mathbb{R}} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ to something between 0 and $+\infty$? Without going into many different link function, let's just use the natural exponential:

$$s = \lambda(s_{\mathbb{R}}) = e^{s_{\mathbb{R}}} \quad \Rightarrow \quad s = e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}$$

And voila we have our survival model:

$$Y \sim Exp(e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p})$$

Interpretation of a unit change in x_1?

The Exponential Log-Linear Model (ELLM)

Let's say we want to use our linear model $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ for the conditional mean. Problem? Yes. The mean can only be a number greater than zero. Hence we need the λ link function again! How can we convert $s_{\mathbb{R}} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ to something between 0 and $+\infty$? Without going into many different link function, let's just use the natural exponential:

$$s = \lambda(s_{\mathbb{R}}) = e^{s_{\mathbb{R}}} \quad \Rightarrow \quad s = e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}$$

And voila we have our survival model:

$$Y \sim Exp(e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p})$$

Interpretation of a unit change in x_1? With all other variables kept constant, a unit change in x_1 will multiply the expected survival by e^{β_1} in a naturally observed new object.

Assumptions of an ELLM

Next up in the recipe... we need to get estimates of the true parameters, we have been denoting these $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$. How to do so?

Assumptions of an ELLM

Next up in the recipe... we need to get estimates of the true parameters, we have been denoting these $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$. How to do so? Maximum Likelihood (just like linear regression and logistic regression). We will first need the "ELLM assumptions"

1. Independence among observations. Thus,

Assumptions of an ELLM

Next up in the recipe... we need to get estimates of the true parameters, we have been denoting these $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$. How to do so? Maximum Likelihood (just like linear regression and logistic regression). We will first need the "ELLM assumptions"

1. Independence among observations. Thus,

$$\begin{aligned} & \mathbb{P}\left(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n \mid \boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X}_2 = \boldsymbol{x}_2, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n\right) \\ = & \prod_{i=1}^{n} \mathbb{P}\left(Y_i = y_i \mid \boldsymbol{X}_1 = \boldsymbol{x}_i\right) \end{aligned}$$

2. Exponential Model. Thus,

Assumptions of an ELLM

Next up in the recipe... we need to get estimates of the true parameters, we have been denoting these $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$. How to do so? Maximum Likelihood (just like linear regression and logistic regression). We will first need the "ELLM assumptions"

1. Independence among observations. Thus,

$$\begin{aligned} & \mathbb{P}\left(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n \mid \boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X}_2 = \boldsymbol{x}_2, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n\right) \\ = \; & \prod_{i=1}^{n} \mathbb{P}\left(Y_i = y_i \mid \boldsymbol{X}_1 = \boldsymbol{x}_i\right) \end{aligned}$$

2. Exponential Model. Thus,

$$= \prod_{i=1}^{n} \frac{1}{\mu} e^{-\frac{1}{\mu} y_i}$$

Assumptions of an ELLM

Next up in the recipe... we need to get estimates of the true parameters, we have been denoting these $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$. How to do so? Maximum Likelihood (just like linear regression and logistic regression). We will first need the "ELLM assumptions"

1. Independence among observations. Thus,

$$\begin{aligned} &\mathbb{P}\left(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n \mid \boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X}_2 = \boldsymbol{x}_2, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n\right) \\ = &\prod_{i=1}^{n} \mathbb{P}\left(Y_i = y_i \mid \boldsymbol{X}_1 = \boldsymbol{x}_i\right) \end{aligned}$$

2. Exponential Model. Thus,

$$= \prod_{i=1}^{n} \frac{1}{\mu} e^{-\frac{1}{\mu} y_i}$$

3. Log-Linear conditional expectation. Thus,

Fitting an ELLM

$$= \prod_{i=1}^{n} \frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} e^{-\frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} y_i}$$

Fitting an ELLM

$$= \prod_{i=1}^{n} \frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} e^{-\frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} y_i}$$

Can we just maximize the above over all values of the β's like before?

Fitting an ELLM

$$= \prod_{i=1}^{n} \frac{1}{e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}}} e^{-\frac{1}{e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}}} y_i}$$

Can we just maximize the above over all values of the β's like before? We are missing one thing. What if survival time is a censored value (denoted y_i')? All we know about that y_i is that it's greater than the last value recorded!

Fitting an ELLM

$$= \prod_{i=1}^{n} \frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} e^{-\frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} y_i}$$

Can we just maximize the above over all values of the β's like before? We are missing one thing. What if survival time is a censored value (denoted y_i')? All we know about that y_i is that it's greater than the last value recorded! Recall $\mathbb{P}(Y > y) = 1 - F(y) = e^{-\frac{1}{\mu} y}$

Let c_i indicate censorship: 1 if censored and 0 if not. Our likelihood is now in two pieces:

$$= \prod_{i=1}^{n} \left(\frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} e^{-\frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} y_i} \right)^{1-c_i}$$

Fitting an ELLM

$$= \prod_{i=1}^{n} \frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} e^{-\frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} y_i}$$

Can we just maximize the above over all values of the β's like before? We are missing one thing. What if survival time is a censored value (denoted y_i')? All we know about that y_i is that it's greater than the last value recorded! Recall $\mathbb{P}(Y > y) = 1 - F(y) = e^{-\frac{1}{\mu} y}$

Let c_i indicate censorship: 1 if censored and 0 if not. Our likelihood is now in two pieces:

$$= \prod_{i=1}^{n} \left(\frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} e^{-\frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} y_i} \right)^{1-c_i} \mathbb{P}\left(Y > y_i'\right)^{c_i}$$

Fitting an ELLM

$$= \prod_{i=1}^{n} \frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} e^{-\frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} y_i}$$

Can we just maximize the above over all values of the β's like before? We are missing one thing. What if survival time is a censored value (denoted y_i')? All we know about that y_i is that it's greater than the last value recorded! Recall $\mathbb{P}(Y > y) = 1 - F(y) = e^{-\frac{1}{\mu} y}$

Let c_i indicate censorship: 1 if censored and 0 if not. Our likelihood is now in two pieces:

$$= \prod_{i=1}^{n} \left(\frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} e^{-\frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} y_i} \right)^{1-c_i} \mathbb{P}\left(Y > y_i'\right)^{c_i}$$

$$= \prod_{i=1}^{n} \left(\frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} e^{-\frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} y_i} \right)^{1-c_i} \left(e^{-\frac{1}{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}} y_i'} \right)^{c_i}$$

Fitting an ELLM

$$= \prod_{i=1}^{n} \frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} e^{-\frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} y_i}$$

Can we just maximize the above over all values of the β's like before? We are missing one thing. What if survival time is a censored value (denoted y_i')? All we know about that y_i is that it's greater than the last value recorded! Recall $\mathbb{P}(Y > y) = 1 - F(y) = e^{-\frac{1}{\mu} y}$

Let c_i indicate censorship: 1 if censored and 0 if not. Our likelihood is now in two pieces:

$$\begin{aligned} &= \prod_{i=1}^{n} \left(\frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} e^{-\frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} y_i} \right)^{1-c_i} \mathbb{P}\left(Y > y_i'\right)^{c_i} \\ &= \prod_{i=1}^{n} \left(\frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} e^{-\frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} y_i} \right)^{1-c_i} \left(e^{-\frac{1}{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}} y_i'} \right)^{c_i} \end{aligned}$$

Now the computer crunches away (similar to a logistic regression fit) and we get values of $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$ back in a split second.

Fitting an ELLM

$$= \prod_{i=1}^{n} \frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} e^{-\frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} y_i}$$

Can we just maximize the above over all values of the β's like before? We are missing one thing. What if survival time is a censored value (denoted y_i')? All we know about that y_i is that it's greater than the last value recorded! Recall $\mathbb{P}(Y > y) = 1 - F(y) = e^{-\frac{1}{\mu} y}$

Let c_i indicate censorship: 1 if censored and 0 if not. Our likelihood is now in two pieces:

$$= \prod_{i=1}^{n} \left(\frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} e^{-\frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} y_i} \right)^{1-c_i} \mathbb{P}\left(Y > y_i'\right)^{c_i}$$

$$= \prod_{i=1}^{n} \left(\frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} e^{-\frac{1}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}} y_i} \right)^{1-c_i} \left(e^{-\frac{1}{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}} y_i'} \right)^{c_i}$$

Now the computer crunches away (similar to a logistic regression fit) and we get values of $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$ back in a split second. Now for inference...

Global Test in Survival ELLM Regression

Just like in logistic regression, we can make use of the ...

Global Test in Survival ELLM Regression

Just like in logistic regression, we can make use of the ... likelihood ratio test.

Global Test in Survival ELLM Regression

Just like in logistic regression, we can make use of the ... likelihood ratio test. Recall:

$$LR := \max_{\theta \in \Theta} \mathcal{L}(\theta; x) \,/\, \max_{\theta \in \Theta_R} \mathcal{L}(\theta; x)$$

Let's now do a "whole model" / "global" / "omnibus" test:

Global Test in Survival ELLM Regression

Just like in logistic regression, we can make use of the ... likelihood ratio test. Recall:

$$LR := \max_{\theta \in \Theta} \mathcal{L}(\theta; x) \,/\, \max_{\theta \in \Theta_R} \mathcal{L}(\theta; x)$$

Let's now do a "whole model" / "global" / "omnibus" test:

$$H_0 : \beta_1 = 0, \beta_2 = 0, \ldots, \beta_p = 0, \quad H_a : \text{at least one is non-zero}$$

So Θ would be the space of all $\beta_0, \beta_1, \ldots, \beta_p$ and Θ_R will restrict the space to only β_0 with zeroes for all other "slope" parameters.

$$LR = \frac{\max\limits_{\beta_0, \beta_1, \ldots, \beta_p} \mathcal{L}(\beta_0, \beta_1, \ldots, \beta_p; y_1, \ldots, y_n, c_1, \ldots, c_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)}{\max\limits_{\beta_0} \mathcal{L}(\beta_0, \beta_1 = 0, \ldots, \beta_p = 0; y_1, \ldots, y_n, y_n, c_1, \ldots, c_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)}$$

Global Test in Survival ELLM Regression

Just like in logistic regression, we can make use of the ... likelihood ratio test. Recall:

$$LR := \max_{\theta \in \Theta} \mathcal{L}(\theta; x) \,/\, \max_{\theta \in \Theta_R} \mathcal{L}(\theta; x)$$

Let's now do a "whole model" / "global" / "omnibus" test:

$$H_0 : \beta_1 = 0, \beta_2 = 0, \ldots, \beta_p = 0, \quad H_a : \text{at least one is non-zero}$$

So Θ would be the space of all $\beta_0, \beta_1, \ldots, \beta_p$ and Θ_R will restrict the space to only β_0 with zeroes for all other "slope" parameters.

$$LR = \frac{\max\limits_{\beta_0, \beta_1, \ldots, \beta_p} \mathcal{L}(\beta_0, \beta_1, \ldots, \beta_p; y_1, \ldots, y_n, c_1, \ldots, c_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)}{\max\limits_{\beta_0} \mathcal{L}(\beta_0, \beta_1 = 0, \ldots, \beta_p = 0; y_1, \ldots, y_n, y_n, c_1, \ldots, c_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)}$$

So in the numerator the computer iterates to find $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$, plugs it in and computes the likelihood and in the denominator the computer independently iterates to find $\left\{\hat{\beta}_0\right\}$, plugs it in and computes the likelihood, then together, the LR.

Partial Tests in Survival Regression

We then look at $Q = 2\ln(LR)$ and compare it to the appropriate χ^2 distribution. Here, since we've dropped p parameters / degrees of freedom, we look at the critical $\chi^2_{p,\alpha}$ value.

Partial Tests in Survival Regression

We then look at $Q = 2\ln(LR)$ and compare it to the appropriate χ^2 distribution. Here, since we've dropped p parameters / degrees of freedom, we look at the critical $\chi^2_{p,\alpha}$ value.

Let's say we want to test something like:

$$H_0 : \beta_1 = 0 \;\&\; \beta_2 = 0, \quad H_a : \text{at least one is non-zero}$$

We can again use the likelihood ratio test:

$$LR = \frac{\max\limits_{\beta_0,\beta_1,\ldots,\beta_p} \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_p; y_1, \ldots, y_n, c_1, \ldots, c_n, x_1, \ldots, x_n\right)}{\max\limits_{\beta_0,\beta_3,\ldots,\beta_p} \mathcal{L}\left(\beta_0, \beta_1 = 0, \beta_2 = 0, \beta_3, \ldots, \beta_p = 0; y_1, \ldots, y_n, c_1, \ldots, c_n, x_1, \ldots, x_n\right)}$$

Partial Tests in Survival Regression

We then look at $Q = 2\ln(LR)$ and compare it to the appropriate χ^2 distribution. Here, since we've dropped p parameters / degrees of freedom, we look at the critical $\chi^2_{p,\alpha}$ value.

Let's say we want to test something like:

$$H_0 : \beta_1 = 0 \ \& \ \beta_2 = 0, \quad H_a : \text{at least one is non-zero}$$

We can again use the likelihood ratio test:

$$LR = \frac{\max_{\beta_0, \beta_1, \ldots, \beta_p} \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_p; y_1, \ldots, y_n, c_1, \ldots, c_n, \mathbf{x}_1, \ldots, \mathbf{x}_n\right)}{\max_{\beta_0, \beta_3, \ldots, \beta_p} \mathcal{L}\left(\beta_0, \beta_1 = 0, \beta_2 = 0, \beta_3, \ldots, \beta_p = 0; y_1, \ldots, y_n, c_1, \ldots, c_n, \mathbf{x}_1, \ldots, \mathbf{x}_n\right)}$$

We then look at $Q = 2\ln(LR)$ and compare it to the appropriate χ^2 distribution. Here, since we've dropped 2 parameters / degrees of freedom, we look at the critical $\chi^2_{2,\alpha}$ value.

Individual Tests in Survival Regression

Let's say we want to test an individual slope coefficient:

$$H_0 : \beta_j = 0, \quad H_a : \beta_j \neq 0$$

(a la the "partial-F test"). We can again use the likelihood ratio test:

$$LR = \frac{\max_{\beta_0,\beta_1,\dots,\beta_p} \mathcal{L}\left(\beta_0, \beta_1, \dots, \beta_p; y_1, \dots, y_n, c_1, \dots, c_n, \mathbf{x}_1, \dots, \mathbf{x}_n\right)}{\max_{\beta_0,\beta_1,\dots,\beta_{j-1},\beta_{j+1},\dots\beta_p} \mathcal{L}\left(\beta_0, \beta_1, \dots, \beta_{j-1}, \beta_j = 0, \beta_{j+1}, \dots \beta_p; y_1, \dots, y_n, c_1, \dots, c_n, \mathbf{x}_1, \dots, \mathbf{x}_n\right)}$$

Individual Tests in Survival Regression

Let's say we want to test an individual slope coefficient:

$$H_0 : \beta_j = 0, \quad H_a : \beta_j \neq 0$$

(a la the "partial-F test"). We can again use the likelihood ratio test:

$$LR = \frac{\max_{\beta_0, \beta_1, \ldots, \beta_p} \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_p; y_1, \ldots, y_n, c_1, \ldots, c_n, \mathbf{x}_1, \ldots, \mathbf{x}_n\right)}{\max_{\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots \beta_p} \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_j = 0, \beta_{j+1}, \ldots \beta_p; y_1, \ldots, y_n, c_1, \ldots, c_n, \mathbf{x}_1, \ldots, \mathbf{x}_n\right)}$$

We then look at $Q = 2\ln(LR)$ and compare it to the appropriate χ^2 distribution. Here, since we've dropped 1 parameter / degrees of freedom; thus we look at the critical $\chi^2_{1,\alpha}$ value.

Individual Tests in Survival Regression

Let's say we want to test an individual slope coefficient:

$$H_0 : \beta_j = 0, \quad H_a : \beta_j \neq 0$$

(a la the "partial-F test"). We can again use the likelihood ratio test:

$$LR = \frac{\max\limits_{\beta_0, \beta_1, \ldots, \beta_p} \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_p; y_1, \ldots, y_n, c_1, \ldots, c_n, \mathbf{x}_1, \ldots, \mathbf{x}_n\right)}{\max\limits_{\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots \beta_p} \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_j = 0, \beta_{j+1}, \ldots \beta_p; y_1, \ldots, y_n, c_1, \ldots, c_n, \mathbf{x}_1, \ldots, \mathbf{x}_n\right)}$$

We then look at $Q = 2\ln(LR)$ and compare it to the appropriate χ^2 distribution. Here, since we've dropped 1 parameter / degrees of freedom; thus we look at the critical $\chi^2_{1,\alpha}$ value.

And again: a χ^2 r.v. with one degree of freedom has the following cool property: $Q \sim \chi^2_1 \Rightarrow$

Individual Tests in Survival Regression

Let's say we want to test an individual slope coefficient:

$$H_0 : \beta_j = 0, \quad H_a : \beta_j \neq 0$$

(a la the "partial-F test"). We can again use the likelihood ratio test:

$$LR = \frac{\max\limits_{\beta_0, \beta_1, \ldots, \beta_p} \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_p; y_1, \ldots, y_n, c_1, \ldots, c_n, \mathbf{x}_1, \ldots, \mathbf{x}_n\right)}{\max\limits_{\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots \beta_p} \mathcal{L}\left(\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_j = 0, \beta_{j+1}, \ldots \beta_p; y_1, \ldots, y_n, c_1, \ldots, c_n, \mathbf{x}_1, \ldots, \mathbf{x}_n\right)}$$

We then look at $Q = 2\ln(LR)$ and compare it to the appropriate χ^2 distribution. Here, since we've dropped 1 parameter / degrees of freedom; thus we look at the critical $\chi^2_{1,\alpha}$ value.

And again: a χ^2 r.v. with one degree of freedom has the following cool property: $Q \sim \chi^2_1 \Rightarrow \sqrt{Q} \sim \mathcal{N}(0, 1)$ i.e. a "z-score". This is how JMP produces standard errors for survival regression coefficients.

Simple Survival Regression and Predictions

How do we predict?

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) =$$

Simple Survival Regression and Predictions

How do we predict?

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \ldots + \hat{\beta}_p x_p^*}$$

What are we predicting?

Simple Survival Regression and Predictions

How do we predict?

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \ldots + \hat{\beta}_p x_p^*}$$

What are we predicting? Average time to survive.

Let's return to the Telecom example and look at just "`SeniorCitizen`" on customer lifetime. [JMP]

Simple Survival Regression and Predictions

How do we predict?

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \ldots + \hat{\beta}_p x_p^*}$$

What are we predicting? Average time to survive.

Let's return to the Telecom example and look at just "`SeniorCitizen`" on customer lifetime. [JMP] Is this variable significant?

Simple Survival Regression and Predictions

How do we predict?

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \ldots + \hat{\beta}_p x_p^*}$$

What are we predicting? Average time to survive.

Let's return to the Telecom example and look at just "SeniorCitizen" on customer lifetime. [JMP] Is this variable significant? Yes.

Simple Survival Regression and Predictions

How do we predict?

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \ldots + \hat{\beta}_p x_p^*}$$

What are we predicting? Average time to survive.

Let's return to the Telecom example and look at just "`SeniorCitizen`" on customer lifetime. [JMP] Is this variable significant? Yes. Now let's predict the time for a new non-senior citizen:

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) =$$

Simple Survival Regression and Predictions

How do we predict?

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \ldots + \hat{\beta}_p x_p^*}$$

What are we predicting? Average time to survive.

Let's return to the Telecom example and look at just "`SeniorCitizen`" on customer lifetime. [JMP] Is this variable significant? Yes. Now let's predict the time for a new non-senior citizen:

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1 (0)} =$$

Simple Survival Regression and Predictions

How do we predict?

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \ldots + \hat{\beta}_p x_p^*}$$

What are we predicting? Average time to survive.

Let's return to the Telecom example and look at just "`SeniorCitizen`" on customer lifetime. [JMP] Is this variable significant? Yes. Now let's predict the time for a new non-senior citizen:

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1(0)} = e^{4.915} = 136.6$$

Now let's predict the time for a new senior citizen:

Simple Survival Regression and Predictions

How do we predict?

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \ldots + \hat{\beta}_p x_p^*}$$

What are we predicting? Average time to survive.

Let's return to the Telecom example and look at just "`SeniorCitizen`" on customer lifetime. [JMP] Is this variable significant? Yes. Now let's predict the time for a new non-senior citizen:

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1(0)} = e^{4.915} = 136.6$$

Now let's predict the time for a new senior citizen:

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) =$$

Simple Survival Regression and Predictions

How do we predict?

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \ldots + \hat{\beta}_p x_p^*}$$

What are we predicting? Average time to survive.

Let's return to the Telecom example and look at just "`SeniorCitizen`" on customer lifetime. [JMP] Is this variable significant? Yes. Now let's predict the time for a new non-senior citizen:

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1 (0)} = e^{4.915} = 136.6$$

Now let's predict the time for a new senior citizen:

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1 (1)} =$$

Simple Survival Regression and Predictions

How do we predict?

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \ldots + \hat{\beta}_p x_p^*}$$

What are we predicting? Average time to survive.

Let's return to the Telecom example and look at just "`SeniorCitizen`" on customer lifetime. [JMP] Is this variable significant? Yes. Now let's predict the time for a new non-senior citizen:

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1 (0)} = e^{4.915} = 136.6$$

Now let's predict the time for a new senior citizen:

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1 (1)} = e^{4.915 + -0.535} = 79.84$$

Is the reduction expected?

Simple Survival Regression and Predictions

How do we predict?

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \ldots + \hat{\beta}_p x_p^*}$$

What are we predicting? Average time to survive.

Let's return to the Telecom example and look at just "`SeniorCitizen`" on customer lifetime. [JMP] Is this variable significant? Yes. Now let's predict the time for a new non-senior citizen:

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1 (0)} = e^{4.915} = 136.6$$

Now let's predict the time for a new senior citizen:

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1 (1)} = e^{4.915 + -0.535} = 79.84$$

Is the reduction expected? Yes. Why are these numbers so large?

Simple Survival Regression and Predictions

How do we predict?

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \ldots + \hat{\beta}_p x_p^*}$$

What are we predicting? Average time to survive.

Let's return to the Telecom example and look at just "`SeniorCitizen`" on customer lifetime. [JMP] Is this variable significant? Yes. Now let's predict the time for a new non-senior citizen:

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1 (0)} = e^{4.915} = 136.6$$

Now let's predict the time for a new senior citizen:

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1 (1)} = e^{4.915 + -0.535} = 79.84$$

Is the reduction expected? Yes. Why are these numbers so large? (1) Likely the exponential model is not a great fit here since the tail is too long and the memorylessness property is not realistic. Also,

Simple Survival Regression and Predictions

How do we predict?

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \ldots + \hat{\beta}_p x_p^*}$$

What are we predicting? Average time to survive.

Let's return to the Telecom example and look at just "`SeniorCitizen`" on customer lifetime. [JMP] Is this variable significant? Yes. Now let's predict the time for a new non-senior citizen:

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1 (0)} = e^{4.915} = 136.6$$

Now let's predict the time for a new senior citizen:

$$\hat{y} = \hat{y}(\boldsymbol{x}^*) = e^{\hat{\beta}_0 + \hat{\beta}_1 (1)} = e^{4.915 + -0.535} = 79.84$$

Is the reduction expected? Yes. Why are these numbers so large? (1) Likely the exponential model is not a great fit here since the tail is too long and the memorylessness property is not realistic. Also, (2) Dataset is not a good sample... was not **designed** for survival.

Multivariable Survival Regression

[JMP]

- Do these coefficients make sense?

Multivariable Survival Regression

[JMP]

- Do these coefficients make sense?
- What's wrong with this dataset??

Multivariable Survival Regression

[JMP]

- Do these coefficients make sense?
- What's wrong with this dataset?? The max survival is 72mo and there's thousands of cases that are censored.

Multivariable Survival Regression

[JMP]

- Do these coefficients make sense?
- What's wrong with this dataset?? The max survival is 72mo and there's thousands of cases that are censored.
- Evaluating the fitness (i.e. $R^2, RMSE, ME_w$, etc) of survival models is complicated... not covered.

Simple Test of Linearity

Here's the "Medicorp" dataset. The response is sales (in \$1000's) and the features are advertising (in \$1000's), American region and bonus for the sales team (in \$1000's).

Simple Test of Linearity

Here's the "Medicorp" dataset. The response is sales (in \$1000's) and the features are advertising (in \$1000's), American region and bonus for the sales team (in \$1000's).

Let's look at a simple regression of sales on bonus.

Simple Test of Linearity

Here's the "Medicorp" dataset. The response is sales (in \$1000's) and the features are advertising (in \$1000's), American region and bonus for the sales team (in \$1000's).

Let's look at a simple regression of sales on bonus. Are we sure this is linear? Or are there diminishing returns? How to test diminishing returns?

Simple Test of Linearity

Here's the "Medicorp" dataset. The response is sales (in \$1000's) and the features are advertising (in \$1000's), American region and bonus for the sales team (in \$1000's).

Let's look at a simple regression of sales on bonus. Are we sure this is linear? Or are there diminishing returns? How to test diminishing returns? Quadratic fit is one way. We should see what kind of sign on the squared term?

Simple Test of Linearity

Here's the "Medicorp" dataset. The response is sales (in $1000's) and the features are advertising (in $1000's), American region and bonus for the sales team (in $1000's).

Let's look at a simple regression of sales on bonus. Are we sure this is linear? Or are there diminishing returns? How to test diminishing returns? Quadratic fit is one way. We should see what kind of sign on the squared term? Negative. And we do. And it's pretty significant... could you make a case of dredging here?

Simple Test of Linearity

Here's the "Medicorp" dataset. The response is sales (in \$1000's) and the features are advertising (in \$1000's), American region and bonus for the sales team (in \$1000's).

Let's look at a simple regression of sales on bonus. Are we sure this is linear? Or are there diminishing returns? How to test diminishing returns? Quadratic fit is one way. We should see what kind of sign on the squared term? Negative. And we do. And it's pretty significant... could you make a case of dredging here? Probably not. Are we sure it's quadratic diminishing returns?

Simple Test of Linearity

Here's the "Medicorp" dataset. The response is sales (in \$1000's) and the features are advertising (in \$1000's), American region and bonus for the sales team (in \$1000's).

Let's look at a simple regression of sales on bonus. Are we sure this is linear? Or are there diminishing returns? How to test diminishing returns? Quadratic fit is one way. We should see what kind of sign on the squared term? Negative. And we do. And it's pretty significant... could you make a case of dredging here? Probably not. Are we sure it's quadratic diminishing returns? Nope... that's much harder to test... beyond scope of course... and probably not all that interesting.

Simple Test of Linearity

Here's the "Medicorp" dataset. The response is sales (in \$1000's) and the features are advertising (in \$1000's), American region and bonus for the sales team (in \$1000's).

Let's look at a simple regression of sales on bonus. Are we sure this is linear? Or are there diminishing returns? How to test diminishing returns? Quadratic fit is one way. We should see what kind of sign on the squared term? Negative. And we do. And it's pretty significant... could you make a case of dredging here? Probably not. Are we sure it's quadratic diminishing returns? Nope... that's much harder to test... beyond scope of course... and probably not all that interesting. What's the takeaway message here?

Simple Test of Linearity

Here's the "Medicorp" dataset. The response is sales (in \$1000's) and the features are advertising (in \$1000's), American region and bonus for the sales team (in \$1000's).

Let's look at a simple regression of sales on bonus. Are we sure this is linear? Or are there diminishing returns? How to test diminishing returns? Quadratic fit is one way. We should see what kind of sign on the squared term? Negative. And we do. And it's pretty significant... could you make a case of dredging here? Probably not. Are we sure it's quadratic diminishing returns? Nope... that's much harder to test... beyond scope of course... and probably not all that interesting. What's the takeaway message here? You should give bonuses but don't make them "too" big.

Interpretation of unit change in x : bonus? Depends on the value where you start from (we are moving away from simple interpretations).

Polynomial Regression

Let's try to fit a better curve to bonus — a 4-degree polynomial. Seems to fit better than a quadratic [see LRT in R]. What's the interpretation of a 4-degree polynomial model??

Polynomial Regression

Let's try to fit a better curve to bonus — a 4-degree polynomial. Seems to fit better than a quadratic [see LRT in R]. What's the interpretation of a 4-degree polynomial model?? Not so intuitive unfortunately. Rarely do we see parametric pre-designed models with more than a quadratic term.

Continuous : Categorical Interactions

It's possible that bonuses may have differential effects in the different American "regions". The way to test this is to allow for a differential slope for each region. [JMP]

Continuous : Categorical Interactions

It's possible that bonuses may have differential effects in the different American "regions". The way to test this is to allow for a differential slope for each region. [JMP]

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1154.5111	116.733	9.89	<.0001*
BONUS	0.6557937	0.408992	1.60	0.1124
REGION[Midwest]	212.82646	16.25271	13.09	<.0001*
REGION[North]	180.20263	19.96834	9.02	<.0001*
REGION[South]	-250.9858	16.3109	-15.39	<.0001*
(BONUS-279.521)*REGION[Midwest]	-3.171339	0.639193	-4.96	<.0001*
(BONUS-279.521)*REGION[North]	-0.765767	0.81714	-0.94	0.3513
(BONUS-279.521)*REGION[South]	1.8218915	0.574401	3.17	0.0021*

How can we interpret this?

Continuous : Categorical Interactions

It's possible that bonuses may have differential effects in the different American "regions". The way to test this is to allow for a differential slope for each region. [JMP]

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1154.5111	116.733	9.89	<.0001*
BONUS	0.6557937	0.408992	1.60	0.1124
REGION[Midwest]	212.82646	16.25271	13.09	<.0001*
REGION[North]	180.20263	19.96834	9.02	<.0001*
REGION[South]	-250.9858	16.3109	-15.39	<.0001*
(BONUS-279.521)*REGION[Midwest]	-3.171339	0.639193	-4.96	<.0001*
(BONUS-279.521)*REGION[North]	-0.765767	0.81714	-0.94	0.3513
(BONUS-279.521)*REGION[South]	1.8218915	0.574401	3.17	0.0021*

How can we interpret this? Could we also interact two continuous features? Two categorical features? Could we interact features with others' polynomials? Yes, yes, yes...

Theories versus Prediction

Above we tested the predictive power of non-linearities in two ways

Theories versus Prediction

Above we tested the predictive power of non-linearities in two ways (1) creating polynomial extensions to given features and finding curvilinear patterns and

Theories versus Prediction

Above we tested the predictive power of non-linearities in two ways (1) creating polynomial extensions to given features and finding curvilinear patterns and (2) creating differential slopes based of one variable when isolating based on the value of another variable.

However, we had specific theories to test in mind

Theories versus Prediction

Above we tested the predictive power of non-linearities in two ways (1) creating polynomial extensions to given features and finding curvilinear patterns and (2) creating differential slopes based of one variable when isolating based on the value of another variable.

However, we had specific theories to test in mind (1) diminishing returns and (2) effects of incentivizing bonuses in different parts of the country.

What if we had no theories to test, but we wanted to fit the data as best as possible (i.e. non-parametric)?

Theories versus Prediction

Above we tested the predictive power of non-linearities in two ways (1) creating polynomial extensions to given features and finding curvilinear patterns and (2) creating differential slopes based of one variable when isolating based on the value of another variable.

However, we had specific theories to test in mind (1) diminishing returns and (2) effects of incentivizing bonuses in different parts of the country.

What if we had no theories to test, but we wanted to fit the data as best as possible (i.e. non-parametric)? Try everything... all polynomial terms up to 5, all interactions with up to the squared terms. [JMP medicorp_exp]

Theories versus Prediction

Above we tested the predictive power of non-linearities in two ways (1) creating polynomial extensions to given features and finding curvilinear patterns and (2) creating differential slopes based of one variable when isolating based on the value of another variable.

However, we had specific theories to test in mind (1) diminishing returns and (2) effects of incentivizing bonuses in different parts of the country.

What if we had no theories to test, but we wanted to fit the data as best as possible (i.e. non-parametric)? Try everything... all polynomial terms up to 5, all interactions with up to the squared terms. [JMP medicorp_exp] We got $R^2 = 98\%$ which rocks!! But none of our variables are significant.... why?

Theories versus Prediction

Above we tested the predictive power of non-linearities in two ways (1) creating polynomial extensions to given features and finding curvilinear patterns and (2) creating differential slopes based of one variable when isolating based on the value of another variable.

However, we had specific theories to test in mind (1) diminishing returns and (2) effects of incentivizing bonuses in different parts of the country.

What if we had no theories to test, but we wanted to fit the data as best as possible (i.e. non-parametric)? Try everything... all polynomial terms up to 5, all interactions with up to the squared terms. [JMP medicorp_exp] We got $R^2 = 98\%$ which rocks!! But none of our variables are significant.... why? Massive collinearity all over the place! do we give a hoot if we only care about predictive accuracy?

Theories versus Prediction

Above we tested the predictive power of non-linearities in two ways (1) creating polynomial extensions to given features and finding curvilinear patterns and (2) creating differential slopes based of one variable when isolating based on the value of another variable.

However, we had specific theories to test in mind (1) diminishing returns and (2) effects of incentivizing bonuses in different parts of the country.

What if we had no theories to test, but we wanted to fit the data as best as possible (i.e. non-parametric)? Try everything... all polynomial terms up to 5, all interactions with up to the squared terms. [JMP medicorp_exp] We got $R^2 = 98\%$ which rocks!! But none of our variables are significant.... why? Massive collinearity all over the place! do we give a hoot if we only care about predictive accuracy? NO.

Theories versus Prediction

Above we tested the predictive power of non-linearities in two ways (1) creating polynomial extensions to given features and finding curvilinear patterns and (2) creating differential slopes based of one variable when isolating based on the value of another variable.

However, we had specific theories to test in mind (1) diminishing returns and (2) effects of incentivizing bonuses in different parts of the country.

What if we had no theories to test, but we wanted to fit the data as best as possible (i.e. non-parametric)? Try everything... all polynomial terms up to 5, all interactions with up to the squared terms. [JMP medicorp_exp] We got $R^2 = 98\%$ which rocks!! But none of our variables are significant.... why? Massive collinearity all over the place! do we give a hoot if we only care about predictive accuracy? NO. But... is the $R^2 = 98\%$ real?

Theories versus Prediction

Above we tested the predictive power of non-linearities in two ways (1) creating polynomial extensions to given features and finding curvilinear patterns and (2) creating differential slopes based of one variable when isolating based on the value of another variable.

However, we had specific theories to test in mind (1) diminishing returns and (2) effects of incentivizing bonuses in different parts of the country.

What if we had no theories to test, but we wanted to fit the data as best as possible (i.e. non-parametric)? Try everything... all polynomial terms up to 5, all interactions with up to the squared terms. [JMP medicorp_exp] We got $R^2 = 98\%$ which rocks!! But none of our variables are significant.... why? Massive collinearity all over the place! do we give a hoot if we only care about predictive accuracy? NO. But... is the $R^2 = 98\%$ real? NO...

The First Clue of Trouble in Paradise...

The First Clue of Trouble in Paradise...

was that F-test demo from Lecture 2 where

$$
\begin{aligned}
x &= \{0, 0.1, 0.2, \ldots, 10.0\} \\
Y &\sim x + \mathcal{E} \\
\mathcal{E} &\sim \mathcal{N}\left(0,\ 5^2\right)
\end{aligned}
$$

[repeat in R].

The First Clue of Trouble in Paradise...

was that F-test demo from Lecture 2 where

$$\begin{aligned} x &= \{0, 0.1, 0.2, \ldots, 10.0\} \\ Y &\sim x + \mathcal{E} \\ \mathcal{E} &\sim \mathcal{N}\left(0,\, 5^2\right) \end{aligned}$$

[repeat in R]. Here, the R^2 goes from 22% up to

- 99.5% all on completely random data which you know is fake!!

The First Clue of Trouble in Paradise...

was that F-test demo from Lecture 2 where

$$
\begin{aligned}
x &= \{0, 0.1, 0.2, \ldots, 10.0\} \\
Y &\sim x + \mathcal{E} \\
\mathcal{E} &\sim \mathcal{N}\left(0,\, 5^2\right)
\end{aligned}
$$

[repeat in R]. Here, the R^2 goes from 22% up to

- 99.5% all on completely random data which you know is fake!!
- 99.9% on splines (polynomials on steroids) which may not be fake??

How do we know?

Recall the Modeling Basics

In lecture 1 we spoke about how

$$Y = f(x_1, \ldots, x_p) + \mathcal{E}$$

where $x_1, \ldots, x_p$ denotes

Recall the Modeling Basics

In lecture 1 we spoke about how

$$Y = f(x_1, \ldots, x_p) + \mathcal{E}$$

where $x_1, \ldots, x_p$ denotes predictors available (including all polynomials and interactions and whatnot!) and $\mathcal{E}$ denotes

Recall the Modeling Basics

In lecture 1 we spoke about how

$$Y = f(x_1, \ldots, x_p) + \mathcal{E}$$

where $x_1, \ldots, x_p$ denotes predictors available (including all polynomials and interactions and whatnot!) and $\mathcal{E}$ denotes **irreducible error** due to information not available (and thus independent of $x_1, \ldots, x_p$), the inaccessible information.

Recall the Modeling Basics

In lecture 1 we spoke about how

$$Y = f(x_1, \ldots, x_p) + \mathcal{E}$$

where $x_1, \ldots, x_p$ denotes predictors available (including all polynomials and interactions and whatnot!) and $\mathcal{E}$ denotes **irreducible error** due to information not available (and thus independent of $x_1, \ldots, x_p$), the inaccessible information.

By including all sorts of polynomials and interactions, we become more nonparametric thereby losing the benefits of the parametric worldview of s...

Recall the Modeling Basics

In lecture 1 we spoke about how

$$Y = f(x_1, \ldots, x_p) + \mathcal{E}$$

where $x_1, \ldots, x_p$ denotes predictors available (including all polynomials and interactions and whatnot!) and $\mathcal{E}$ denotes **irreducible error** due to information not available (and thus independent of $x_1, \ldots, x_p$), the inaccessible information.

By including all sorts of polynomials and interactions, we become more nonparametric thereby losing the benefits of the parametric worldview of s... (i.e. parsimony,

Recall the Modeling Basics

In lecture 1 we spoke about how

$$Y = f(x_1, \ldots, x_p) + \mathcal{E}$$

where $x_1, \ldots, x_p$ denotes predictors available (including all polynomials and interactions and whatnot!) and $\mathcal{E}$ denotes **irreducible error** due to information not available (and thus independent of $x_1, \ldots, x_p$), the inaccessible information.

By including all sorts of polynomials and interactions, we become more nonparametric thereby losing the benefits of the parametric worldview of s... (i.e. parsimony, interpretability

Recall the Modeling Basics

In lecture 1 we spoke about how

$$Y = f(x_1, \ldots, x_p) + \mathcal{E}$$

where $x_1, \ldots, x_p$ denotes predictors available (including all polynomials and interactions and whatnot!) and $\mathcal{E}$ denotes **irreducible error** due to information not available (and thus independent of $x_1, \ldots, x_p$), the inaccessible information.

By including all sorts of polynomials and interactions, we become more nonparametric thereby losing the benefits of the parametric worldview of s... (i.e. parsimony, interpretability and inference) but gaining a closer fit of the true f. But what could go wrong if we take this liberty?

Non-germane footnote: recall that we fit $\hat{s}$ which generally speaking fails to estimate s correctly (model error) and s generally speaking fails to represent f correctly since its a parametric model which lacks flexibility.

Recall the Modeling Basics

In lecture 1 we spoke about how

$$Y = f(x_1, \ldots, x_p) + \mathcal{E}$$

where $x_1, \ldots, x_p$ denotes predictors available (including all polynomials and interactions and whatnot!) and $\mathcal{E}$ denotes **irreducible error** due to information not available (and thus independent of $x_1, \ldots, x_p$), the inaccessible information.

By including all sorts of polynomials and interactions, we become more nonparametric thereby losing the benefits of the parametric worldview of s... (i.e. parsimony, interpretability and inference) but gaining a closer fit of the true f. But what could go wrong if we take this liberty?

Non-germane footnote: recall that we fit $\hat{s}$ which generally speaking fails to estimate s correctly (model error) and s generally speaking fails to represent f correctly since its a parametric model which lacks flexibility.

Going Too Far

$$Y = f(x_1, \ldots, x_p) + \mathcal{E}$$

It's possible our $\hat{f}$ can estimate f (which is good) but...

Going Too Far

$$Y = f(x_1, \ldots, x_p) + \mathcal{E}$$

It's possible our $\hat{f}$ can estimate f (which is good) but... it can encroach on and start fitting and optimizing the $\mathcal{E}$. Why is this bad?

Going Too Far

$$Y = f(x_1, \ldots, x_p) + \mathcal{E}$$

It's possible our $\hat{f}$ can estimate f (which is good) but... it can encroach on and start fitting and optimizing the $\mathcal{E}$. Why is this bad? Since $\mathcal{E}$ is independent of $x_1, \ldots, x_p$, it is creating a random fit.

Going Too Far

$$Y = f(x_1, \ldots, x_p) + \mathcal{E}$$

It's possible our $\hat{f}$ can estimate f (which is good) but... it can encroach on and start fitting and optimizing the $\mathcal{E}$. Why is this bad? Since $\mathcal{E}$ is independent of $x_1, \ldots, x_p$, it is creating a random fit. Random fits are akin to "making up a model" and it is the opposite of the "data-driven approach".

Going Too Far

$$Y = f(x_1, \ldots, x_p) + \mathcal{E}$$

It's possible our $\hat{f}$ can estimate f (which is good) but... it can encroach on and start fitting and optimizing the $\mathcal{E}$. Why is this bad? Since $\mathcal{E}$ is independent of $x_1, \ldots, x_p$, it is creating a random fit. Random fits are akin to "making up a model" and it is the opposite of the "data-driven approach".

Any value different from f is not **generalizable** as the conditional mean minimizes squared error.

Going Too Far

$$Y = f(x_1, \ldots, x_p) + \mathcal{E}$$

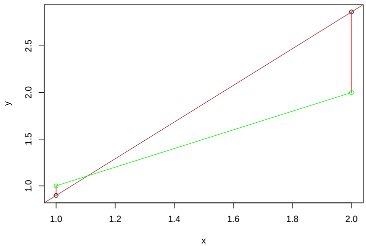
It's possible our $\hat{f}$ can estimate f (which is good) but... it can encroach on and start fitting and optimizing the $\mathcal{E}$. Why is this bad? Since $\mathcal{E}$ is independent of $x_1, \ldots, x_p$, it is creating a random fit. Random fits are akin to "making up a model" and it is the opposite of the "data-driven approach".

Any value different from f is not **generalizable** as the conditional mean minimizes squared error.

But the BIG problem is: we don't know what the form of f is and we don't know the individual values of $\mathcal{E}$. Thus, we have NO WAY to know if we've overfit (as of now)!

When Does This Happen?

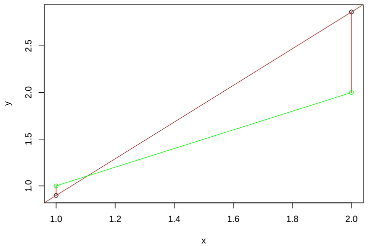
Essentially, when p gets closer to n. Here's the linear model case with $n = 2$ and there's one slope so $p = 1$ (+1 for the intercept) so really, the number of predictors is 2 since there is two degrees of freedom. [R demo]



The green is the true conditional expectation function $f(x)$ and the brown is the fitted model and the red are the true $\mathcal{E}_1$ and $\mathcal{E}_2$ values. Where are e_1 and e_2?

When Does This Happen?

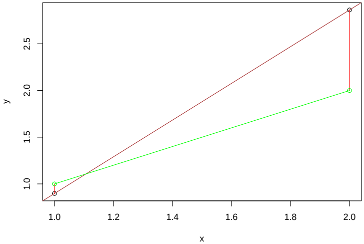
Essentially, when p gets closer to n. Here's the linear model case with $n = 2$ and there's one slope so $p = 1$ (+1 for the intercept) so really, the number of predictors is 2 since there is two degrees of freedom. [R demo]



The green is the true conditional expectation function $f(x)$ and the brown is the fitted model and the red are the true $\mathcal{E}_1$ and $\mathcal{E}_2$ values. Where are e_1 and e_2? They are zero (and thus not pictured). The fitted model has $R^2 = 100\%$. Recall from middle school... when they asked you to draw a line between two points — the line perfectly goes through two points.

When Does This Happen?

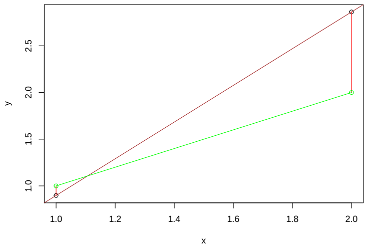
Essentially, when p gets closer to n. Here's the linear model case with $n = 2$ and there's one slope so $p = 1$ (+1 for the intercept) so really, the number of predictors is 2 since there is two degrees of freedom. [R demo]



The green is the true conditional expectation function $f(x)$ and the brown is the fitted model and the red are the true $\mathcal{E}_1$ and $\mathcal{E}_2$ values. Where are e_1 and e_2? They are zero (and thus not pictured). The fitted model has $R^2 = 100\%$. Recall from middle school... when they asked you to draw a line between two points — the line perfectly goes through two points. Why are the $\mathbb{SE}\left[\hat{\beta}_j\right]$'s `NA`?

When Does This Happen?

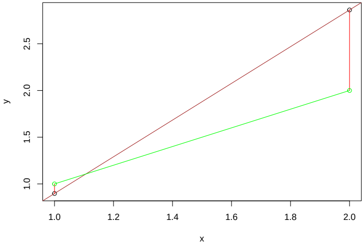
Essentially, when p gets closer to n. Here's the linear model case with $n = 2$ and there's one slope so $p = 1$ (+1 for the intercept) so really, the number of predictors is 2 since there is two degrees of freedom. [R demo]



The green is the true conditional expectation function $f(x)$ and the brown is the fitted model and the red are the true $\mathcal{E}_1$ and $\mathcal{E}_2$ values. Where are e_1 and e_2? They are zero (and thus not pictured). The fitted model has $R^2 = 100\%$. Recall from middle school... when they asked you to draw a line between two points — the line perfectly goes through two points. Why are the $\mathbb{S}\text{E}\left[\hat{\beta}_j\right]$'s `NA`? Division by zero.

When Does This Happen?

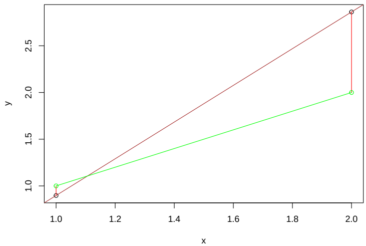
Essentially, when p gets closer to n. Here's the linear model case with $n = 2$ and there's one slope so $p = 1$ (+1 for the intercept) so really, the number of predictors is 2 since there is two degrees of freedom. [R demo]



The green is the true conditional expectation function $f(x)$ and the brown is the fitted model and the red are the true $\mathcal{E}_1$ and $\mathcal{E}_2$ values. Where are e_1 and e_2? They are zero (and thus not pictured). The fitted model has $R^2 = 100\%$. Recall from middle school... when they asked you to draw a line between two points — the line perfectly goes through two points. Why are the $\mathbb{S}\text{E}\left[\hat{\beta}_j\right]$'s `NA`? Division by zero. Can you imagine two predictors and three points and a plane?

When Does This Happen?

Essentially, when p gets closer to n. Here's the linear model case with $n = 2$ and there's one slope so $p = 1$ (+1 for the intercept) so really, the number of predictors is 2 since there is two degrees of freedom. [R demo]



The green is the true conditional expectation function $f(x)$ and the brown is the fitted model and the red are the true $\mathcal{E}_1$ and $\mathcal{E}_2$ values. Where are e_1 and e_2? They are zero (and thus not pictured). The fitted model has $R^2 = 100\%$. Recall from middle school... when they asked you to draw a line between two points — the line perfectly goes through two points. Why are the $\mathbb{S}\mathrm{E}\left[\hat{\beta}_j\right]$'s `NA`? Division by zero. Can you imagine two predictors and three points and a plane? What about in a logistic regression? [Whiteboard demo]

Assessing Overfitting and its Cost to You

Overfitting comes from over-optimizing a sample (i.e. fitting $\mathcal{E}$) and thus having poor generalizability and thus poor predictive performance in the future!

Assessing Overfitting and its Cost to You

Overfitting comes from over-optimizing a sample (i.e. fitting $\mathcal{E}$) and thus having poor generalizability and thus poor predictive performance in the future!

Let's return to the [R demo] to witness the cost of overfitting.

Assessing Overfitting and its Cost to You

Overfitting comes from over-optimizing a sample (i.e. fitting $\mathcal{E}$) and thus having poor generalizability and thus poor predictive performance in the future!

Let's return to the [R demo] to witness the cost of overfitting. How did we demonstrate overfittedness?

Assessing Overfitting and its Cost to You

Overfitting comes from over-optimizing a sample (i.e. fitting $\mathcal{E}$) and thus having poor generalizability and thus poor predictive performance in the future!

Let's return to the [R demo] to witness the cost of overfitting. How did we demonstrate overfittedness? We used "new data" not in the dataframe generated from the same realization process as the historical dataframe (our sample). Hence this new data is called **out of sample (oos)** data. And then we calculated familiar metrics such as SSE, RMSE, R^2 but since these are done oos, we call them oosSSE, oosRMSE, oosR^2 and they are our **out of sample statistics**. Everything we spoke about previously we will now call **in-sample statistics**.

	In Sample	Out of Sample
RMSE	2.0	84.7
R^2	99.9%	9.5%

Overfitting can get arbitrarily bad and this is an extreme example.

Assessments in the Real World

It is easy to assess if you have access to the data-generating process — you can just generate more data and see how you do. But in the real world, you only have a limited set of historical data from the data-generating process. What to do?

Assessments in the Real World

It is easy to assess if you have access to the data-generating process — you can just generate more data and see how you do. But in the real world, you only have a limited set of historical data from the data-generating process. What to do?

Why not *imagine* some of your historical data is future data. That is, split your dataframe into two pieces:

Assessments in the Real World

It is easy to assess if you have access to the data-generating process — you can just generate more data and see how you do. But in the real world, you only have a limited set of historical data from the data-generating process. What to do?

Why not *imagine* some of your historical data is future data. That is, split your dataframe into two pieces:

1. What we've been calling the historical dataframe but now will calling the "training set". We use this to build the model, as we've done.

Assessments in the Real World

It is easy to assess if you have access to the data-generating process — you can just generate more data and see how you do. But in the real world, you only have a limited set of historical data from the data-generating process. What to do?

Why not *imagine* some of your historical data is future data. That is, split your dataframe into two pieces:

1. What we've been calling the historical dataframe but now will calling the "training set". We use this to build the model, as we've done.
2. The "test set" that is the piece you are imagining to come in the future. his gives you a means to evaluate your model you built from the training set.

Assessments in the Real World

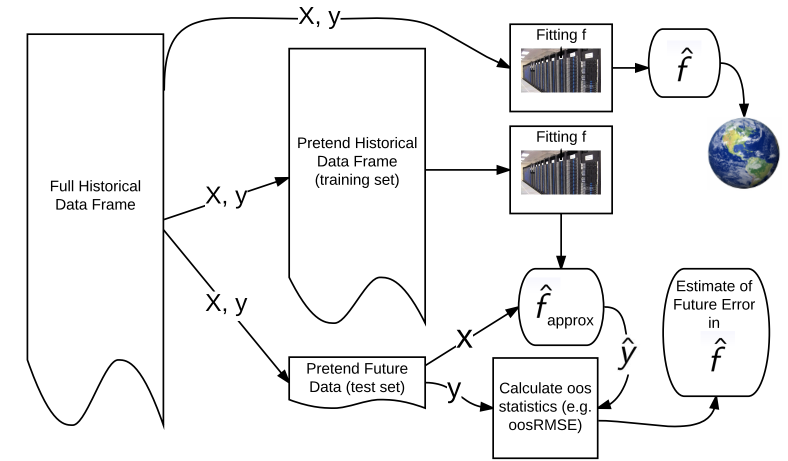
It is easy to assess if you have access to the data-generating process — you can just generate more data and see how you do. But in the real world, you only have a limited set of historical data from the data-generating process. What to do?

Why not *imagine* some of your historical data is future data. That is, split your dataframe into two pieces:

1. What we've been calling the historical dataframe but now will calling the "training set". We use this to build the model, as we've done.
2. The "test set" that is the piece you are imagining to come in the future. his gives you a means to evaluate your model you built from the training set.

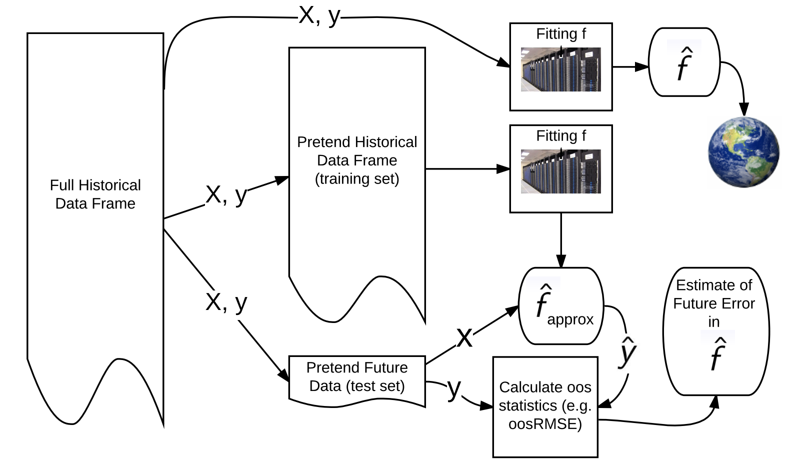
Building the model on the training set and predicting on the test set and comparing these predictions to the real, known values of the response in the test set constitutes *out of sample validation*. Why is it called that?

Model Fitting with OOS Validation



Can oos metrics be better than in-sample metrics (on average)?

Model Fitting with OOS Validation



Can oos metrics be better than in-sample metrics (on average)? No...

A Possible Spin on Validation

Procedure outlined above:

1. Split dataframe into training and test.

A Possible Spin on Validation

Procedure outlined above:

1. Split dataframe into training and test.
2. Build model on training.

A Possible Spin on Validation

Procedure outlined above:

1. Split dataframe into training and test.
2. Build model on training.
3. Predict using the test set.

A Possible Spin on Validation

Procedure outlined above:

1. Split dataframe into training and test.
2. Build model on training.
3. Predict using the test set.
4. Calculate estimate of future generalization error.

Does the following procedure also seem reasonable?

1. Split dataframe into training and test.
2. Build model A on training.
3. Predict using the test set.
4. Calculate estimate of future generalization error of model A.

A Possible Spin on Validation

Procedure outlined above:

1. Split dataframe into training and test.
2. Build model on training.
3. Predict using the test set.
4. Calculate estimate of future generalization error.

Does the following procedure also seem reasonable?

1. Split dataframe into training and test.
2. Build model A on training.
3. Predict using the test set.
4. Calculate estimate of future generalization error of model A.
5. Build a different model B on training.

A Possible Spin on Validation

Procedure outlined above:

1. Split dataframe into training and test.
2. Build model on training.
3. Predict using the test set.
4. Calculate estimate of future generalization error.

Does the following procedure also seem reasonable?

1. Split dataframe into training and test.
2. Build model A on training.
3. Predict using the test set.
4. Calculate estimate of future generalization error of model A.
5. Build a different model B on training.
6. Predict using the test set.

A Possible Spin on Validation

Procedure outlined above:

1. Split dataframe into training and test.
2. Build model on training.
3. Predict using the test set.
4. Calculate estimate of future generalization error.

Does the following procedure also seem reasonable?

1. Split dataframe into training and test.
2. Build model A on training.
3. Predict using the test set.
4. Calculate estimate of future generalization error of model A.
5. Build a different model B on training.
6. Predict using the test set.
7. Calculate estimate of future generalization error of model B.

A Possible Spin on Validation

Procedure outlined above:

1. Split dataframe into training and test.
2. Build model on training.
3. Predict using the test set.
4. Calculate estimate of future generalization error.

Does the following procedure also seem reasonable?

1. Split dataframe into training and test.
2. Build model A on training.
3. Predict using the test set.
4. Calculate estimate of future generalization error of model A.
5. Build a different model B on training.
6. Predict using the test set.
7. Calculate estimate of future generalization error of model B.
8. Pick whichever model has better generalization error.

Valid Validation

What's wrong?

Valid Validation

What's wrong? You snooped the test set...

Valid Validation

What's wrong? You snooped the test set... analagous to looking into the future and

Valid Validation

What's wrong? You snooped the test set... analagous to looking into the future and seeing results and saying "I don't like em",

Valid Validation

What's wrong? You snooped the test set... analagous to looking into the future and seeing results and saying "I don't like em", then returning to the past and trying again.

Valid Validation

What's wrong? You snooped the test set... analagous to looking into the future and seeing results and saying "I don't like em", then returning to the past and trying again. This can lead to very optimistic results — it is essentially overfitting and you've tricked yourself into thinking you are honest.

Valid Validation

What's wrong? You snooped the test set... analagous to looking into the future and seeing results and saying "I don't like em", then returning to the past and trying again. This can lead to very optimistic results — it is essentially overfitting and you've tricked yourself into thinking you are honest.

The oos validation is only valid if...

you treat the test set as a lockbox. Once you open it up, that's it!

Training-Test Splitting

We have a choice to split our dataframe into two pieces. Assuming each data point is independent (the running assumption), you should do this completely randomly. When would this assumption not be true?

Training-Test Splitting

We have a choice to split our dataframe into two pieces. Assuming each data point is independent (the running assumption), you should do this completely randomly. When would this assumption not be true? For example, a time series.

Training-Test Splitting

We have a choice to split our dataframe into two pieces. Assuming each data point is independent (the running assumption), you should do this completely randomly. When would this assumption not be true? For example, a time series.

How large should the test set be? Usual sizes are 10-30%. What's the tradeoff? If the test set is larger, then ...

Training-Test Splitting

We have a choice to split our dataframe into two pieces. Assuming each data point is independent (the running assumption), you should do this completely randomly. When would this assumption not be true? For example, a time series.

How large should the test set be? Usual sizes are 10-30%. What's the tradeoff? If the test set is larger, then ...

1. the more accurate the assessment of generalization error would be (less variance) and

Training-Test Splitting

We have a choice to split our dataframe into two pieces. Assuming each data point is independent (the running assumption), you should do this completely randomly. When would this assumption not be true? For example, a time series.

How large should the test set be? Usual sizes are 10-30%. What's the tradeoff? If the test set is larger, then ...

1. the more accurate the assessment of generalization error would be (less variance) and
2. the less accurate the model will be since it's fitting with less data (more bias)

Training-Test Splitting

We have a choice to split our dataframe into two pieces. Assuming each data point is independent (the running assumption), you should do this completely randomly. When would this assumption not be true? For example, a time series.

How large should the test set be? Usual sizes are 10-30%. What's the tradeoff? If the test set is larger, then ...

1. the more accurate the assessment of generalization error would be (less variance) and
2. the less accurate the model will be since it's fitting with less data (more bias)

If the test set is smaller then vice versa:

1. the less accurate the assessment of generalization error would be (more variance) and

Training-Test Splitting

We have a choice to split our dataframe into two pieces. Assuming each data point is independent (the running assumption), you should do this completely randomly. When would this assumption not be true? For example, a time series.

How large should the test set be? Usual sizes are 10-30%. What's the tradeoff? If the test set is larger, then ...

1. the more accurate the assessment of generalization error would be (less variance) and
2. the less accurate the model will be since it's fitting with less data (more bias)

If the test set is smaller then vice versa:

1. the less accurate the assessment of generalization error would be (more variance) and
2. the more accurate the model will be since it's fitting with less data (less bias)

Training-Test Splitting

We have a choice to split our dataframe into two pieces. Assuming each data point is independent (the running assumption), you should do this completely randomly. When would this assumption not be true? For example, a time series.

How large should the test set be? Usual sizes are 10-30%. What's the tradeoff? If the test set is larger, then ...

1. the more accurate the assessment of generalization error would be (less variance) and
2. the less accurate the model will be since it's fitting with less data (more bias)

If the test set is smaller then vice versa:

1. the less accurate the assessment of generalization error would be (more variance) and
2. the more accurate the model will be since it's fitting with less data (less bias)

Note: the in-sample and oos statistics are statistics! Thus, they are random!

Doing oos Validation in JMP

Let's take a look at the medicorp data and validate it. Recall, the in-sample $R^2 \approx 98\%$ and the in-sample RMSE is about 48.9.

Doing oos Validation in JMP

Let's take a look at the medicorp data and validate it. Recall, the in-sample $R^2 \approx 98\%$ and the in-sample RMSE is about 48.9.

[JMP Cols...Modeling Utilities...Make Validation Col...fit model... validation option is the validation col...crossvalidation tab] Note: "RASE" = root average squared error = root mean squared error = oosRMSE. "Validation" = "test". "Freq" is the sample sizes in training and test. Looks like we were overoptimistic by 6x the standard error on predictions! Substantial overfitting.

Validating Multiple Models

Let's look at three models for the White Wine data. Here the response is wine quality as measured by professional raters and features are 11 features (e.g. acidity, sugar, pH and alcohol content).

A plain linear model

Validating Multiple Models

Let's look at three models for the White Wine data. Here the response is wine quality as measured by professional raters and features are 11 features (e.g. acidity, sugar, pH and alcohol content).

A plain linear model

B six-degree polynomials for all features

Validating Multiple Models

Let's look at three models for the White Wine data. Here the response is wine quality as measured by professional raters and features are 11 features (e.g. acidity, sugar, pH and alcohol content).

A plain linear model

B six-degree polynomials for all features

C six-degree polynomials and all first-order interactions

Validating Multiple Models

Let's look at three models for the White Wine data. Here the response is wine quality as measured by professional raters and features are 11 features (e.g. acidity, sugar, pH and alcohol content).

A plain linear model

B six-degree polynomials for all features

C six-degree polynomials and all first-order interactions

D six-degree polynomials and all interactions up to 11th order

Validating Multiple Models

Let's look at three models for the White Wine data. Here the response is wine quality as measured by professional raters and features are 11 features (e.g. acidity, sugar, pH and alcohol content).

- A plain linear model
- B six-degree polynomials for all features
- C six-degree polynomials and all first-order interactions
- D six-degree polynomials and all interactions up to 11th order

[JMP col validation... fit all models with validation ... save prediction formula cols... analyze model... model comparison] Conclusions?

Validating Multiple Models

Let's look at three models for the White Wine data. Here the response is wine quality as measured by professional raters and features are 11 features (e.g. acidity, sugar, pH and alcohol content).

- A plain linear model
- B six-degree polynomials for all features
- C six-degree polynomials and all first-order interactions
- D six-degree polynomials and all interactions up to 11th order

[JMP col validation... fit all models with validation ... save prediction formula cols... analyze model... model comparison]
Conclusions? Model C looks the best. Where to go from here?

What did I do that wasn't legal? Remember a few slides ago? I looked at the test set four times! We need to solve this problem...