A Descriptive Title Here: Be Creative

Lando Calrissian, Leia Organa, Luke Skywalker, Han Solo

December 13, 2017

Abstract

This is the abstract. It should probably be at most about five sentences. The abstract should briefly explain what you are doing, why you are doing it, and what you have found. Reading only the abstract, the reader should have a good idea about what to expect from the rest of the document.

Introduction

The **introduction** should discuss and setup a real-world problem. Essentially, you need to motivate why the analysis that you're about to do should be done. Why is a model useful in this situation? What is the goal of this model? The introduction should also provide enough background on the subject area for a reader to understand your analysis. Do not assume your reader knows anything about the subject area that your data comes from. If the reader does not understand your data, there is no way the reader will understand your motivation.

Figure 1: Space advertising?

This document will walk you through some of the necessary steps of formatting your report. Do not mistake the length of this document as an example of the length of a proper report. Length is not important. Communicating your results in a concise but complete manner is important.

Materials and Methods

The **materials and methods** section should discuss how you solved your problem. The material that you are using is the data you have selected. The methods that you are using are those learned in class. This section should contain the bulk of your "work." This section will contain the bulk of the `R` code that is used to generate the results. Your `R` code is not expected to be perfect idiomatic `R`, but it is expected to be understood by a reader without too much effort. The majority of your code should be suppressed from the final report, but consider displaying code that helps illustrate the analysis you performed, for example, training of models.

```
gbm_grid = expand.grid(interaction.depth = c(1, 2, 3),
                       n.trees = (1:30) * 100,
```

```
                        shrinkage = c(0.1, 0.3),
                        n.minobsinnode = 20)
```

```
sim_gbm_mod = train(y ~ ., data = sim_trn, method = "gbm",
                    trControl = trainControl(method = "cv", number = 5),
                    tuneGrid = gbm_grid, verbose = FALSE)
```

Consider adding subsections in this section. One potential set of subsections could be **data** and **models**. The data section would describe your data. What is it? Where did it come from? How will it be useful in answering your problem? What if any preprocessing have you done to it? Provide references to information about the data, but explain enough that your reader does not need to utilize them. The models section would describe the modeling methods that you will consider, as well as strategies for comparison.

`R` Code and `rmarkdown`

An important part of the report is communicating results in a well-formatted manner. This template document should help a lot with that task. Some thoughts on using `R` and `rmarkdown`:

- Chunks are set to not echo by default in this document.
- Include at least one chunk that is echoed, else, this template may break.
- Consider naming your chunks. This will be necessary for referencing chunks that create tables or figures.
- One chunk per table or figure!
- Tables should be created using `knitr::kable()`.
- Consider using `kableExtra()` for better presentation of tables. (Examples in this document.)
- Caption all figures and tables. (Examples in this document.)
- Use the `img/` sub-directory for any external images.
- Use the `data/` sub-directory for any external data.

LaTeX

While you will not directly work with LaTeX, since your final report will be a pdf, you will need to have LaTeX installed.

- MiKTeX (Windows)
- MacTeX (OSX)
- TeX Live (Linux)

If you are interested, some details on working with TeX can be found in this guide by UIUC Mathematics Professor A.J. Hildebrand.

With `rmarkdown`, LaTeX can be used inline, like this, $a^2 + b^2 = c^2$, or using display mode,

$$\mathbb{E}_{X,Y}\left[(Y - f(X))^2\right] = \mathbb{E}_X \mathbb{E}_{Y|X}\left[(Y - f(X))^2 \mid X = x\right]$$

For examples of LaTeX code, you can right click on any equation in `R4SL` to obtain the LaTeX used to generate.

You are not required to use BibTeX for references, but if you are familiar, please consider doing so. Otherwise, you can simply manually cite your references. But with BibTeX, it is extremely easy. For example, we could reference the `rmarkdown` paper (Allaire et al. 2015) or the tidy data paper. (Wickham and others 2014) Some details can be found in the `bookdown` book. Also, hint, Google Scholar makes obtaining BibTeX reference extremely easy.

Because we're using LaTeX to render the final document, you will have no control over the placement of tables and figures. Be OK with this! (If you really need control, see the first image of this document.) Since

Figure 2: A photograph of R. A. Fisher in his younger years.

they'll essentially appear where LaTeX decides to put them, we need to be able to reference them. For example, we could talk about Figure 1, which talks about space advertising. Notice that this is numbered automatically, but internally referenced using the chunk name.

Results

The **results** section should contain numerical or graphical summaries of your results. What are the results of applying your selected methods to your materials? Consider reporting a "final" or "best" model you have chosen. There is not necessarily one, singular correct model, but certainly some methods and models are better than others in certain situations. In this section you should provide evidence that your final choice of

Figure 3: A photograph of R. A. Fisher in his older years. Originally a larger image, but made smaller through the use of chunk options.

Table 1: An example table.

mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	name
18	8	307	130	3504	12.0	70	1	chevrolet chevelle malibu
15	8	350	165	3693	11.5	70	1	buick skylark 320
18	8	318	150	3436	11.0	70	1	plymouth satellite
16	8	304	150	3433	12.0	70	1	amc rebel sst
17	8	302	140	3449	10.5	70	1	ford torino
15	8	429	198	4341	10.0	70	1	ford galaxie 500

model is a good one.

Notice that captioning of tables is done using `kable()` while captioning of figures is done using chunk options.

Discussion

The **discussion** section should contain discussion of your results. This should also frame your results in the context of the data. What do your results mean? Results are often just numbers, here you need to explain what they tell you about the problem you are trying to solve. The results section tells the reader what the results are. The discussion section tells the reader why those results matter. Since the results are created in the results section, but they are being discussed in the discussion section, you will need to reference tables and figures from the results section. For example, we could talk about Table 1, which displays an automobile dataset. Or we could discuss Figure 4, which displays multiple models fit to the same dataset. Figure 5 shows two models of different complexities fit to different simulated datasets.

To demonstrate that you understand course concepts, consider spending some time discussing:

- Was your final model a linear or non-linear method?
- Was your final model a parametric or non-parametric method?
- Was your final model generative or discriminant?

Perhaps compare these details to some other methods you considered.

Conclusion

The **conclusion** should be a brief recap of what you did, why you did it, and what you found. Submission of your report will be considered a submission to the journal *Annals of STAT 432* and the grading process will be considered part of the "peer review" process. If a report is well written, uses thoughtful and throughout analysis, and is sufficiently interesting you may be asked to have your work "published" as an example for future students. All group members will have to agree to publication. You may also be asked to make edits before publication, but you should be sure to **proofread** and **spellcheck** your work before your initial submission.

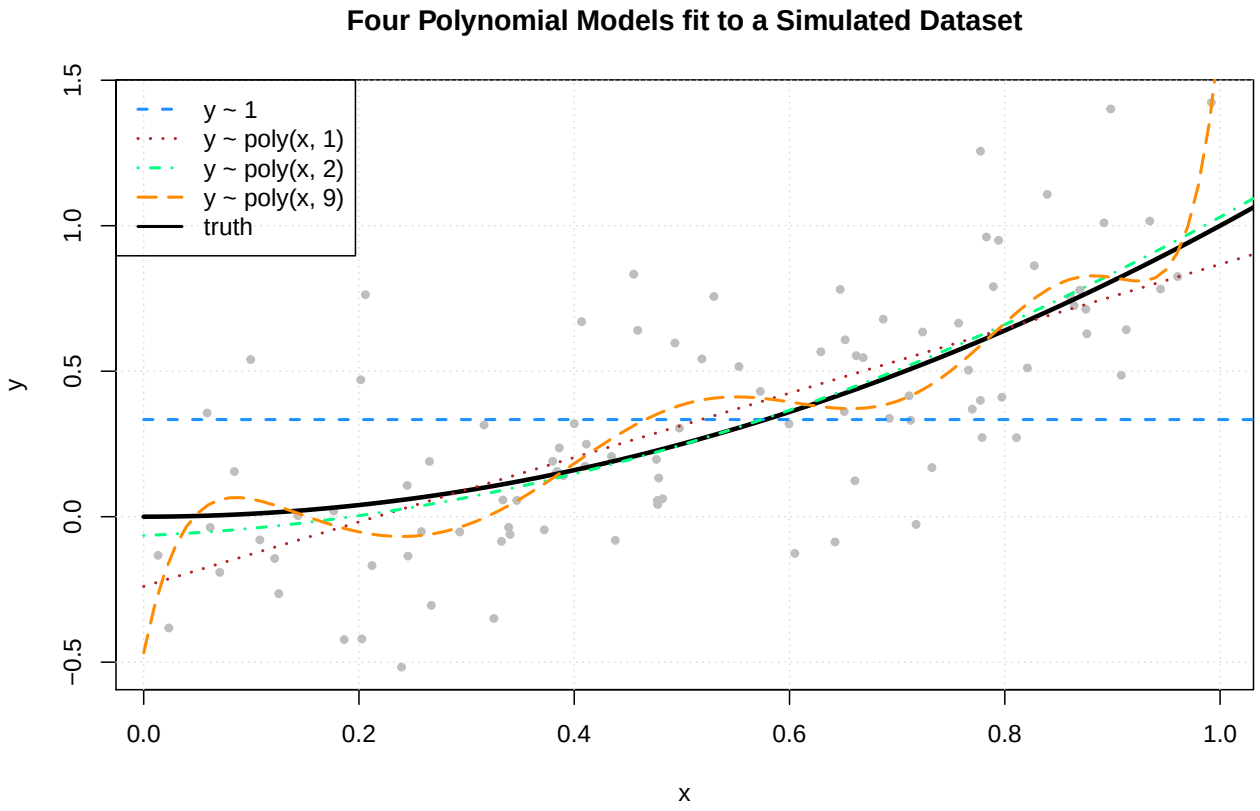

Figure 4: This is a caption.

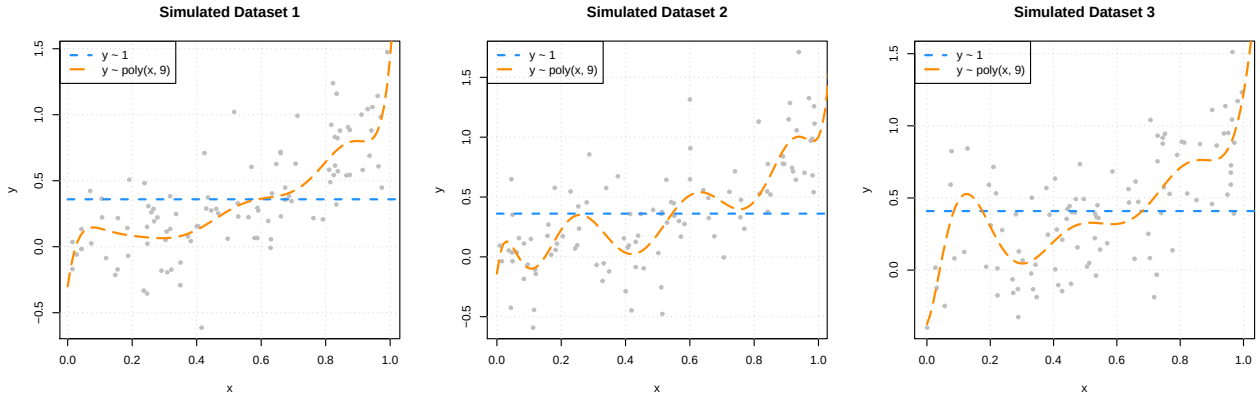

Figure 5: This is a another caption. Your captions should probably be more descriptive.

Table 2: This is an example of a table in the Appendix. Notice that it is way too big, and has way too much information. We use `kableExtra()` to shrink it down, but even then, no one would actually read this table.

	AtBat	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	CHmRun	CRuns	CRBI	CWalks	League	Division	PutOuts	Assists	Errors	Salary	NewLeague
-Andy Allanson	293	66	1	30	29	14	1	293	66	1	30	29	14	A	E	446	33	20	NA	A
-Alan Ashby	315	81	7	24	38	39	14	3449	835	69	321	414	375	N	W	632	43	10	475.000	N
-Alvin Davis	479	130	18	66	72	76	3	1624	457	63	224	266	263	A	W	880	82	14	480.000	A
-Andre Dawson	496	141	20	65	78	37	11	5628	1575	225	828	838	354	N	E	200	11	3	500.000	N
-Andres Galarraga	321	87	10	39	42	30	2	396	101	12	48	46	33	N	E	805	40	4	91.500	N
-Alfredo Griffin	594	169	4	74	51	35	11	4408	1133	19	501	336	194	A	W	282	421	25	750.000	A
-Al Newman	185	37	1	23	8	21	2	214	42	1	30	9	24	N	E	76	127	7	70.000	A
-Argenis Salazar	298	73	0	24	24	7	3	509	108	0	41	37	12	A	W	121	283	9	100.000	A
-Andres Thomas	323	81	6	26	32	8	2	341	86	6	32	34	8	N	W	143	290	19	75.000	N
-Andre Thornton	401	92	17	49	66	65	13	5206	1332	253	784	890	866	A	E	0	0	0	1100.000	A
-Alan Trammell	574	159	21	107	75	59	10	4631	1300	90	702	504	488	A	E	238	445	22	517.143	A
-Alex Trevino	202	53	4	31	26	27	9	1876	467	15	192	186	161	N	W	304	45	11	512.500	N
-Andy VanSlyke	418	113	13	48	61	47	4	1512	392	41	205	204	203	N	E	211	11	7	550.000	N
-Alan Wiggins	239	60	0	30	11	22	6	1941	510	4	309	103	207	A	E	121	151	6	700.000	A
-Bill Almon	196	43	7	29	27	30	13	3231	825	36	376	290	238	N	E	80	45	8	240.000	N
-Billy Beane	183	39	3	20	15	11	3	201	42	3	20	16	11	A	W	118	0	0	NA	A
-Buddy Bell	568	158	20	89	75	73	15	8068	2273	177	1045	993	732	N	W	105	290	10	775.000	N
-Buddy Biancalana	190	46	2	24	8	15	5	479	102	5	65	23	39	A	W	102	177	16	175.000	A
-Bruce Bochte	407	104	6	57	43	65	12	5233	1478	100	643	658	653	A	W	912	88	9	NA	A
-Bruce Bochy	127	32	8	16	22	14	8	727	180	24	67	82	56	N	W	202	22	2	135.000	N

Table 3: This is another example of a ridiculous table. Notice that it is automatically numbered.

	year	age	maritl	race	education	region	jobclass	health	health_ins	logwage	wage
231655	2006	18	1. Never Married	1. White	1. < HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good	2. No	4.318063	75.04315
86582	2004	24	1. Never Married	1. White	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	2. No	4.255273	70.47602
161300	2003	45	2. Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	1. <=Good	1. Yes	4.875061	130.98218
155159	2003	43	2. Married	3. Asian	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	5.041393	154.68529
11443	2005	50	4. Divorced	1. White	2. HS Grad	2. Middle Atlantic	2. Information	1. <=Good	1. Yes	4.318063	75.04315
376662	2008	54	2. Married	1. White	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	4.845098	127.11574
450601	2009	44	2. Married	4. Other	3. Some College	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	5.133021	169.52854
377954	2008	30	1. Never Married	3. Asian	3. Some College	2. Middle Atlantic	2. Information	1. <=Good	1. Yes	4.716003	111.72085
228963	2006	41	1. Never Married	2. Black	3. Some College	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	4.778151	118.88436
81404	2004	52	2. Married	1. White	2. HS Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	4.857333	128.68049
302778	2007	45	4. Divorced	1. White	3. Some College	2. Middle Atlantic	2. Information	1. <=Good	1. Yes	4.763428	117.14682
305706	2007	34	2. Married	1. White	2. HS Grad	2. Middle Atlantic	1. Industrial	2. >=Very Good	2. No	4.397940	81.28325
8690	2005	35	1. Never Married	1. White	2. HS Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	4.494155	89.49248
153561	2003	39	2. Married	1. White	4. College Grad	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	4.903090	134.70538
449654	2009	54	2. Married	1. White	2. HS Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	4.903090	134.70538
447660	2009	51	2. Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	4.505150	90.48191
160191	2003	37	1. Never Married	3. Asian	4. College Grad	2. Middle Atlantic	1. Industrial	2. >=Very Good	2. No	4.414973	82.67964
230312	2006	50	2. Married	1. White	5. Advanced Degree	2. Middle Atlantic	2. Information	2. >=Very Good	2. No	5.360552	212.84235
301585	2007	56	2. Married	1. White	4. College Grad	2. Middle Atlantic	1. Industrial	1. <=Good	1. Yes	4.861026	129.15669
153682	2003	37	1. Never Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	4.591065	98.59934

Appendix

The **appendix** section should contain any additional code, tables, and graphics that are not explicitly referenced in the narrative of the report.

References

Allaire, JJ, Joe Cheng, Yihui Xie, Jonathan McPherson, Winston Chang, Jeff Allen, Hadley Wickham, Aron Atkins, and Rob Hyndman. 2015. "Rmarkdown: Dynamic Documents for R." *R Package Version 0.5*.

Wickham, Hadley, and others. 2014. "Tidy Data." *Journal of Statistical Software* 59 (10). Foundation for Open Access Statistics: 1–23.