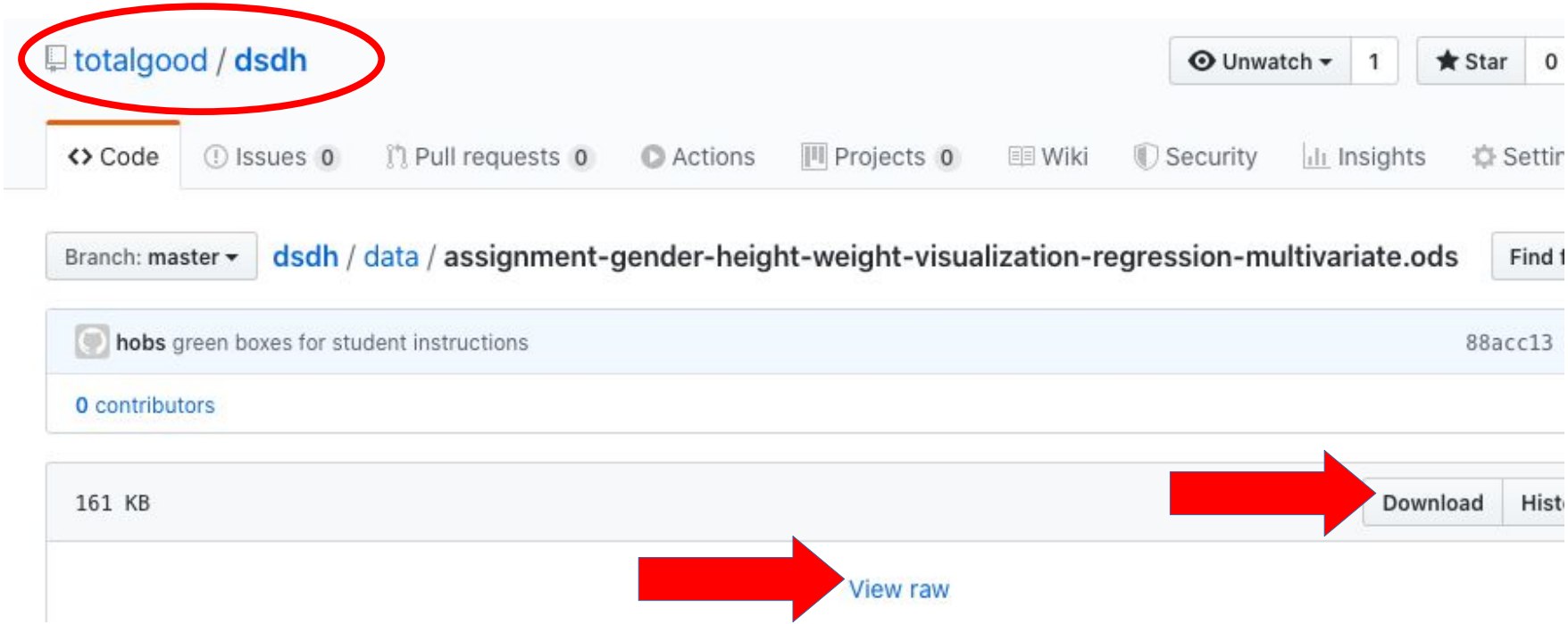
L3: Assignment Spreadsheet Linear Modeling

Hobson Lane, UC San Diego
Instructor

UC San Diego EXTENSION

UC San Diego SCHOOL OF MEDICINE

Dataset Spreadsheet: bit.ly/ucsd-spreadsheet



Dataset Spreadsheet: bit.ly/ucsd-spreadsheet

A: Patient ID (integer, but not usable, unique for each)

B: Gender (“Male’/”Female”), text feature

C: Height (in), continuous numerical feature

D: Weight (lbs), continuous numerical target (regression)

Assignment Goal

- Descriptive Statistics
  - Min, Max, Mean, Standard Deviation
- Scatter Plot
- Linear Regression (by manual trial and error)
- Multivariate Linear Regression

Descriptive Statistics and Scatter Plot

=AVERAGE(C:C)

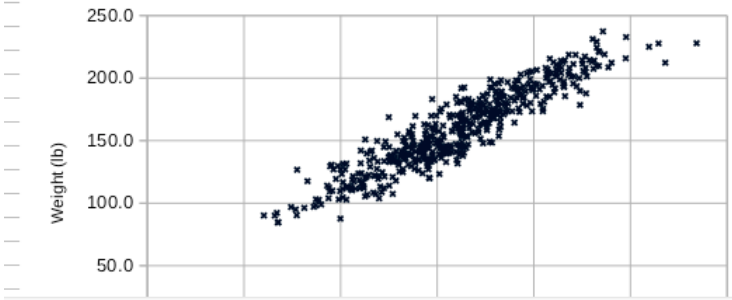
entID	Gender (text)	Gender (0/1)	Height (in)	Weight (lb)
0	Male	0	67.0	167.9
1	Female	1	63.1	124.8
2	Male	0	64.4	162.5
3	Female	1	66.4	162.0
4	Female	1	64.8	152.8
5	Female	1	66.2	140.5
6	Male	0	66.3	156.6
7	Male	0	72.4	199.4
8	Female	1	60.2	104.0
9	Female	1	56.1	89.6
10	Female	1	63.4	141.9
11	Female	1	62.1	112.7
12	Male	0	64.9	158.4
13	Female	1	68.3	164.2
14	Female	1	66.0	133.9

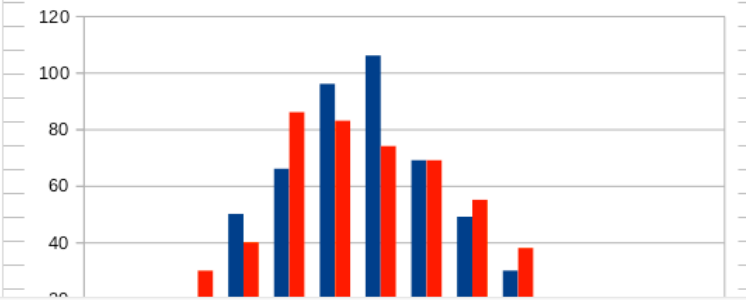
Descriptive Statistics

	Gender (0/1)	Height (in)	Weight (lb)
min	0	56.1	84.0
max	1	78.5	236.8
sum	271	33,062.2	79,593.4
mean	0.542	66.1	159.2
std	0.498732	3.9	32.0

Histogram (Binning and Counting)

bins:	10	10	10		
interval:	1	2	15		
start:	0	56	83		
Bin Value (gender)	**Count (Freq.)**	**Bin Value (height)**	**Count (Freq.)**	**Bin Value (weight)**	**Count (Freq.)**
0	229	56	0	83	0
1	271	58	9	98	12
2	0	60	18	113	30
3	0	62	50	128	40
4	0	64	66	143	86
5	0	66	96	158	83
6	0	68	106	173	74
7	0	70	69	188	69
8	0	72	49	203	55
9	0	74	30	218	38
10	0	76	4	233	12
11	0	78	2	248	1





Example Linear Regression

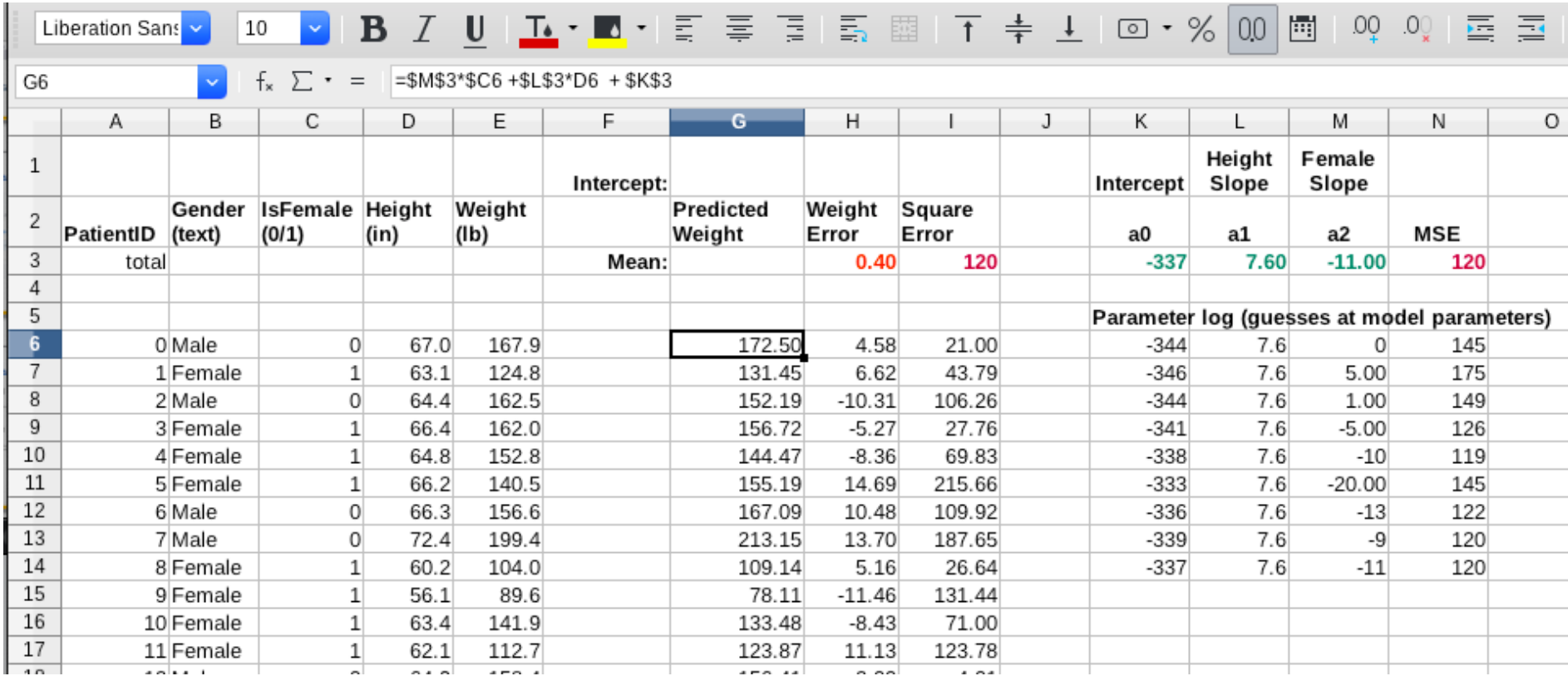
D6 | f_x | $\sum$ | = | =A6*$J$3 + $I$3

	A	B	C	D	E	F	G	H	I	J	K
1			Intercept:						Intercept	Slope	
2	Height (in)	Weight (lb)		Predicted Weight	Weight Error	Absolute Error	Square Error		a0	a1	MSE
3			Mean:	158.55	-0.64	9.65			-344	7.60	145
4											
5									Guesses		
6	67.0	167.9		165.50	-2.42	2.42	5.84		-503	10	233
7	63.1	124.8		135.45	10.62	10.62	112.73		-435	9	175
8	64.4	162.5		145.19	-17.31	17.31	299.58		-368	8	150
9	66.4	162.0		160.72	-1.27	1.27	1.61		-301	7	155
10	64.8	152.8		148.47	-4.36	4.36	18.98		-337	7.50	144
11	66.2	140.5		159.19	18.69	18.69	349.14		-320	7.25	146
12	66.3	156.6		160.09	3.48	3.48	12.14		-353	7.75	145
13	72.4	199.4		206.15	6.70	6.70	44.87		-344	7.6	145
14	60.2	104.0		113.14	9.16	9.16	83.93				
15	56.1	89.6		82.11	-7.46	7.46	55.72				

Linear Regression Template

	A	B	C	D	E	F	G	H	I	J	K
1			Intercept:						Intercept	Slope	
2	Height (in)	Weight (lb)		Predicted Weight	Weight Error	Absolute Error	Square Error		a0	a1	MSE
3			Mean:	193.49	34.31	35.27			-600	12.00	1,617
4											
5									Parameter log (guesses at mod		
6	67.0	167.9		204.47	36.55	36.55	1,336.16		Guess at several values for a0 and a1 (above) until, recording your guesses and their MSE here. You should be able to achieve less than 145 lbs^2 MSE and less than 1 lb weight error (E3) [delete this box]		
7	63.1	124.8		157.02	32.19	32.19	1,036.33				
8	64.4	162.5		172.40	9.91	9.91	98.11				
9	66.4	162.0		196.93	34.94	34.94	1,220.77				
10	64.8	152.8		177.59	24.76	24.76	613.00				
11	66.2	140.5		194.51	54.01	54.01	2,916.76				
12	66.3	156.6		195.93	39.33	39.33	1,546.56				
13	72.4	199.4		268.65	69.20	69.20	4,789.18				
14	60.2	104.0		121.80	17.82	17.82	317.60				
15	56.1	89.6		72.80	-16.77	16.77	281.29				
16	63.4	141.9		160.23	18.33	18.33	335.84				
17	62.1	112.7		145.06	32.32	32.32	1,044.29				
18	64.9	158.4		179.07	20.66	20.66	426.69				

Example Multivariate Linear Regression

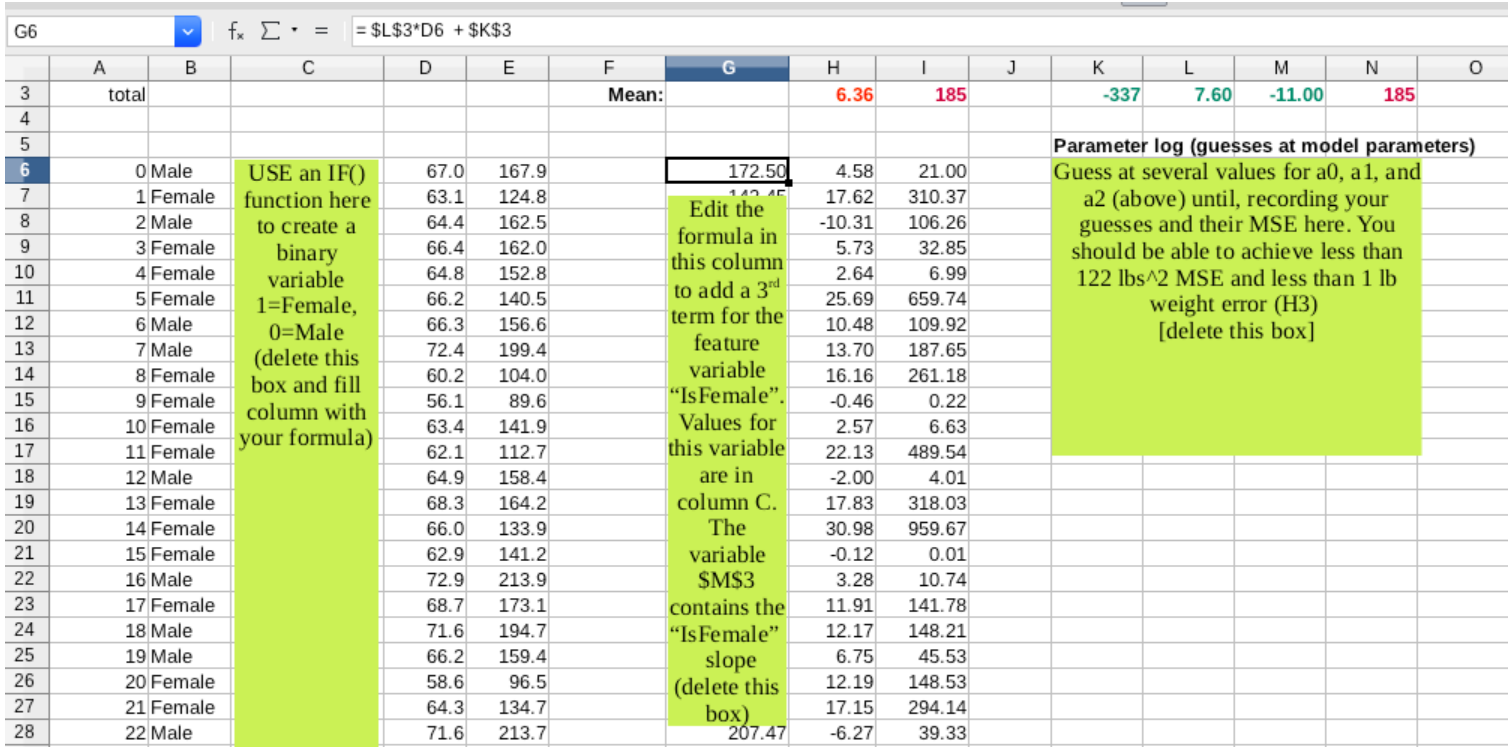


Liberation Sans | 10

G6 | =$M$3*$C6 +$L$3*D6 + $K$3

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1						Intercept:					Intercept	Height Slope	Female Slope		
2	PatientID	Gender (text)	IsFemale (0/1)	Height (in)	Weight (lb)		Predicted Weight	Weight Error	Square Error		a0	a1	a2	MSE	
3	total					Mean:		0.40	120		-337	7.60	-11.00	120	
4															
5											Parameter log (guesses at model parameters)				
6	0	Male	0	67.0	167.9		172.50	4.58	21.00		-344	7.6	0	145	
7	1	Female	1	63.1	124.8		131.45	6.62	43.79		-346	7.6	5.00	175	
8	2	Male	0	64.4	162.5		152.19	-10.31	106.26		-344	7.6	1.00	149	
9	3	Female	1	66.4	162.0		156.72	-5.27	27.76		-341	7.6	-5.00	126	
10	4	Female	1	64.8	152.8		144.47	-8.36	69.83		-338	7.6	-10	119	
11	5	Female	1	66.2	140.5		155.19	14.69	215.66		-333	7.6	-20.00	145	
12	6	Male	0	66.3	156.6		167.09	10.48	109.92		-336	7.6	-13	122	
13	7	Male	0	72.4	199.4		213.15	13.70	187.65		-339	7.6	-9	120	
14	8	Female	1	60.2	104.0		109.14	5.16	26.64		-337	7.6	-11	120	
15	9	Female	1	56.1	89.6		78.11	-11.46	131.44						
16	10	Female	1	63.4	141.9		133.48	-8.43	71.00						
17	11	Female	1	62.1	112.7		123.87	11.13	123.78						

Multivariate Linear Regression Template



G6 | = $L$3*D6 + $K$3

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
3	total					Mean:		6.36	185		-337	7.60	-11.00	185	
4															
5											Parameter log (guesses at model parameters)				
6	0	Male	USE an IF() function here to create a binary variable 1=Female, 0=Male (delete this box and fill column with your formula)	67.0	167.9		172.50	4.58	21.00		Guess at several values for a0, a1, and a2 (above) until, recording your guesses and their MSE here. You should be able to achieve less than 122 lbs^2 MSE and less than 1 lb weight error (H3) [delete this box]				
7	1	Female		63.1	124.8		143.45	17.62	310.37						
8	2	Male		64.4	162.5			-10.31	106.26						
9	3	Female		66.4	162.0			5.73	32.85						
10	4	Female		64.8	152.8			2.64	6.99						
11	5	Female		66.2	140.5			25.69	659.74						
12	6	Male		66.3	156.6			10.48	109.92						
13	7	Male		72.4	199.4			13.70	187.65						
14	8	Female		60.2	104.0			16.16	261.18						
15	9	Female		56.1	89.6			-0.46	0.22						
16	10	Female		63.4	141.9			2.57	6.63						
17	11	Female		62.1	112.7			22.13	489.54						
18	12	Male		64.9	158.4			-2.00	4.01						
19	13	Female		68.3	164.2			17.83	318.03						
20	14	Female		66.0	133.9			30.98	959.67						
21	15	Female		62.9	141.2			-0.12	0.01						
22	16	Male		72.9	213.9			3.28	10.74						
23	17	Female		68.7	173.1			11.91	141.78						
24	18	Male		71.6	194.7			12.17	148.21						
25	19	Male		66.2	159.4			6.75	45.53						
26	20	Female		58.6	96.5			12.19	148.53						
27	21	Female		64.3	134.7			17.15	294.14						
28	22	Male		71.6	213.7		207.47	-6.27	39.33						

Edit the formula in this column to add a 3rd term for the feature variable "IsFemale". Values for this variable are in column C. The variable $M$3 contains the "IsFemale" slope (delete this box)