Weight Loss Dataset

This data was collected from a crowd sourced experiment looking at the effects of various parameters on weight loss. Data was gathered through the use of self-tracking apps (such as activity trackers like FitBit™ or through questionnares. The study lasted 3 months and the values present here have been aggregaged and patients with missing data have been dropped along with any potentially identifiable information. Our overall goal is to find which parameters are most associated with wieght loss.

Dataset Description

The data is given in a sqlite datbase named **wightdata.sqlite**, with a single table, **patient**.

- `age` Patient's age at the start of the study
- `gender` self-identified gender
- `height` the patient's height (in inches)
- `initweight` the weight of the patient at the start of the study (in pounds)
- `calintake` the average number of calories consumed by day
- `jobstatus` the self-identified job status, either Active or Inactive
- `timecardio` the average minutes of cardio (per week)
- `timeresist` the average minutes of resistance training (per week)
- `sleep` the average amount of sleep (hours) per day
- `steps` the average number of steps per day (counted using phone accelerometers)
- `deltaweight`the change in weight after the 3-month study (in pounds)

Variable names

- `data` All patient data, represented as a MATLAB table object

Objectives

- Understand the dataset
- Be able to manipulate the database tables in the DB Browser for SQlite
- Connection to the dataset using MATLAB functions, native connection, and JDBC.
- Use SQL query in MATLAB; import data to a cell array, convert a cell array to a table, and convert a cell array to a matrix
- Be able to manipulate the data in a cell array, table, and matrix
- Logical Index Array
- `corrcoef()` https://www.mathworks.com/help/matlab/ref/corrcoef.html#bunkanr
- Two-sample t-test; MATLAB *ttest2* `[h,p,ci,stats] = ttest2(___)`
- Combine variables to test a new model
- Multivariate linear regression, `mvregress()`
- `Z-scaling`

Connect to the Database

```
% Establish connection to JDBC data source
conn = database('hw5_weightdata', '', '');
```

Data Statistics Using SQL

Use SQL queries to answer each of the following questions.

```
% Establish sql query of 'SELECT * FROM patient'
% Use fetch to import data to dbdata (as a cell array)
% Use cell2table() to convert the cell array dbdata to a table as data
% assigning the following column names (VariableNames)
% 'id' 'age' 'gender' 'height' 'initweight' 'calintake' 'jobstatus'
% 'timecardio' 'timeresist' 'sleep' 'steps' 'deltaweight'
query = 'SELECT * FROM patient';
dbdata = fetch(conn, query, 'DataReturnFormat', 'cellarray')
```

dbdata = 250×12 cell

...

	1	2	3	4	5	6	7	8
1	1	19	'Female'	59.3734	137.1087	1.9678e+03	'Inactive'	39.1656
2	2	19	'Female'	82.6719	154.9608	1.8850e+03	'Inactive'	37.6153
3	3	37	'Female'	61.0149	132.0130	1500	'Active'	38.3226
4	4	36	'Female'	62.4880	129.7659	1.7951e+03	'Inactive'	36.2942
5	5	32	'Female'	59.1823	132.2316	1500	'Active'	36.5790
6	6	38	'Male'	62.5520	187.9075	1.7481e+03	'Inactive'	15.2110
7	7	21	'Female'	60.6551	153.3522	1500	'Active'	0
8	8	24	'Female'	69.7559	131.2006	1500	'Active'	46.0211
9	9	44	'Male'	75.5218	222.2269	1.8750e+03	'Inactive'	14.6797
10	10	24	'Female'	62.5327	144.5377	1.9022e+03	'Inactive'	48.1395
11	11	44	'Male'	73.5189	179.1951	1.7726e+03	'Inactive'	17.2165
12	12	38	'Male'	55.3979	213.1399	1.7168e+03	'Inactive'	8.4671
13	13	29	'Female'	61.6995	143.1719	1500	'Active'	40.2879
14	14	23	'Male'	65.4305	185.0747	1.8586e+03	'Inactive'	17.2012
15	15	22	'Female'	65.4774	134.0369	1.9647e+03	'Inactive'	32.4872
16	16	39	'Female'	57.7803	133.4423	1500	'Active'	0
17	17	22	'Female'	69.5728	139.8936	1500	'Active'	41.1586
18	18	24	'Male'	69.1721	177.2989	1500	'Active'	0
19	19	41	'Male'	77.9292	174.8174	1500	'Active'	0
20	20	22	'Female'	61.9492	140.2975	1500	'Active'	36.5182
21	21	22	'Male'	73.5419	197.3122	1500	'Active'	14.1205

	1	2	3	4	5	6	7	8
22	22	27	'Female'	63.2412	141.8073	1.7054e+03	'Inactive'	31.8667
23	23	34	'Male'	85.2456	202.2675	2.0625e+03	'Inactive'	10.5634
24	24	25	'Female'	67.2219	137.1239	1500	'Active'	39.1190
25	25	39	'Male'	62.1298	197.8195	1500	'Active'	19.0210
26	26	30	'Female'	75.8700	126.4978	2.0191e+03	'Inactive'	39.9069
27	27	29	'Male'	71.3347	180.5776	2.0153e+03	'Inactive'	12.2731
28	28	41	'Male'	72.4046	196.5634	1500	'Active'	0
29	29	24	'Female'	66.8003	125.0450	1500	'Active'	41.6673
30	30	33	'Female'	58.8487	163.8223	1.8111e+03	'Inactive'	43.8693
31	31	24	'Female'	60.9871	169.6279	1500	'Active'	34.9298
32	32	22	'Female'	67.1167	139.3932	1.9243e+03	'Inactive'	35.2446
33	33	47	'Female'	66.2543	161.1063	2.0790e+03	'Inactive'	0
34	34	41	'Female'	59.1782	141.5105	1.8624e+03	'Inactive'	45.0549
35	35	39	'Female'	62.4401	131.4985	1.8279e+03	'Inactive'	46.9436
36	36	46	'Female'	69.2459	151.9187	1500	'Active'	32.9758
37	37	26	'Male'	52.4626	159.0762	1500	'Active'	9.6720
38	38	36	'Male'	71.0037	188.0001	1500	'Active'	0
39	39	18	'Male'	65.3908	202.1782	1.9541e+03	'Inactive'	25.8097
40	40	21	'Male'	66.5177	220.0976	1.7408e+03	'Inactive'	13.6092
41	41	43	'Female'	60.8930	120.0183	1500	'Active'	0
42	42	36	'Female'	73.1585	139.3641	1.8599e+03	'Inactive'	40.4361
43	43	23	'Female'	56.2075	147.1721	1.6190e+03	'Inactive'	0
44	44	48	'Female'	59.1720	119.0975	1.7044e+03	'Inactive'	39.7961
45	45	22	'Female'	65.7152	132.9015	2.0385e+03	'Inactive'	49.8571
46	46	29	'Female'	68.5144	131.2718	2.0528e+03	'Inactive'	35.2459
47	47	36	'Male'	63.7203	160.1505	1500	'Active'	15.7062
48	48	21	'Male'	63.8213	185.5053	1.7352e+03	'Inactive'	9.5861
49	49	34	'Female'	65.7873	133.1923	1.9241e+03	'Inactive'	33.5687
50	50	22	'Female'	66.9509	132.7390	1500	'Active'	0
51	51	38	'Female'	64.7826	137.1880	1500	'Active'	36.8852
52	52	30	'Female'	59.1492	123.1797	1500	'Active'	39.1745
53	53	21	'Male'	67.7952	194.1069	1.7509e+03	'Inactive'	16.1817
54	54	20	'Female'	57.2832	153.9002	2.0870e+03	'Inactive'	33.8361

	1	2	3	4	5	6	7	8
55	55	43	'Male'	83.7546	186.2477	1500	'Active'	22.5823
56	56	39	'Male'	69.2562	178.1039	1.9263e+03	'Inactive'	21.0596
57	57	45	'Female'	70.9013	135.8652	1.9161e+03	'Inactive'	49.0730
58	58	36	'Female'	68.5303	154.2571	1500	'Active'	41.7313
59	59	45	'Male'	78.7450	219.5455	1500	'Active'	22.9854
60	60	23	'Male'	88.5646	176.1487	1500	'Active'	13.4440
61	61	24	'Male'	50.3550	216.5725	1.7155e+03	'Inactive'	12.2776
62	62	40	'Female'	58.6235	151.4327	1500	'Active'	38.5048
63	63	29	'Male'	69.7131	172.9651	1500	'Active'	12.7610
64	64	31	'Female'	62.0872	146.0219	1.8256e+03	'Inactive'	40.2592
65	65	44	'Male'	71.9858	144.1522	1500	'Active'	17.4377
66	66	24	'Female'	74.9814	129.3345	1500	'Active'	42.8447
67	67	21	'Female'	68.9830	155.6516	1500	'Active'	0
68	68	26	'Female'	65.1914	147.3916	1500	'Active'	42.6067
69	69	43	'Male'	64.7788	199.2751	1.9540e+03	'Inactive'	10.7451
70	70	28	'Male'	77.8480	162.5924	1500	'Active'	4.4518
71	71	29	'Male'	78.1216	215.8337	1.8717e+03	'Inactive'	17.2140
72	72	33	'Male'	68.0995	179.5369	1.9410e+03	'Inactive'	23.3751
73	73	37	'Male'	57.8204	147.9261	1500	'Active'	6.1969
74	74	38	'Male'	53.1875	226.8471	2.0768e+03	'Inactive'	0
75	75	39	'Female'	67.9117	135.5261	2.0115e+03	'Inactive'	41.8512
76	76	32	'Male'	46.4906	191.8603	1.9402e+03	'Inactive'	19.2644
77	77	39	'Male'	65.3603	193.2385	1.6569e+03	'Inactive'	0
78	78	20	'Female'	65.8864	151.8960	2.0488e+03	'Inactive'	42.6029
79	79	29	'Male'	79.6971	208.9313	1.6445e+03	'Inactive'	1.2603
80	80	44	'Male'	69.3296	179.0395	1.8644e+03	'Inactive'	0
81	81	36	'Female'	64.4863	140.2773	1.6088e+03	'Inactive'	0
82	82	31	'Male'	72.1059	192.7995	1500	'Active'	16.4761
83	83	20	'Female'	58.7973	136.7891	1.8251e+03	'Inactive'	0
84	84	27	'Female'	69.5145	122.6887	1500	'Active'	34.1839
85	85	24	'Female'	66.8786	138.3935	1500	'Active'	36.3763
86	86	32	'Male'	74.4864	187.8576	1500	'Active'	19.8843
87	87	32	'Male'	67.9866	202.6006	1.9907e+03	'Inactive'	12.9473

	1	2	3	4	5	6	7	8
88	88	33	'Male'	69.7979	219.0734	1.6630e+03	'Inactive'	9.0759
89	89	31	'Female'	70.1493	122.9622	1.7778e+03	'Inactive'	41.6324
90	90	32	'Male'	63.8500	174.6370	1.8546e+03	'Inactive'	13.9026
91	91	20	'Female'	63.7745	122.3978	1500	'Active'	0
92	92	47	'Female'	55.2011	136.2520	1500	'Active'	0
93	93	47	'Male'	52.8228	178.3886	1500	'Active'	10.6547
94	94	38	'Female'	58.4764	131.4541	1500	'Active'	0
95	95	30	'Female'	71.3329	122.7571	1.6219e+03	'Inactive'	39.5402
96	96	39	'Female'	66.5982	136.9215	1500	'Active'	32.5852
97	97	39	'Female'	69.0131	133.8118	1.6230e+03	'Inactive'	34.6741
98	98	46	'Female'	64.7141	138.5840	1500	'Active'	34.8480
99	99	37	'Female'	64.1979	116.7309	1500	'Active'	50.1447
100	100	46	'Male'	60.1285	175.9285	1500	'Active'	13.7283

⋮

```
% Convert data to a table with specific column headers
cols = {'id', 'age', 'gender', 'height', 'initweight', 'calintake', 'jobstatus', 'timecardio',
        'timeresist', 'sleep', 'steps', 'deltaweight'};
data = cell2table(dbdata, "VariableNames", cols)
```

```
data = 250×12 table
```

...

	id	age	gender	height	initweight	calintake	jobstatus
1	1	19	'Female'	59.3734	137.1087	1.9678e+03	'Inactive'
2	2	19	'Female'	82.6719	154.9608	1.8850e+03	'Inactive'
3	3	37	'Female'	61.0149	132.0130	1500	'Active'
4	4	36	'Female'	62.4880	129.7659	1.7951e+03	'Inactive'
5	5	32	'Female'	59.1823	132.2316	1500	'Active'
6	6	38	'Male'	62.5520	187.9075	1.7481e+03	'Inactive'
7	7	21	'Female'	60.6551	153.3522	1500	'Active'
8	8	24	'Female'	69.7559	131.2006	1500	'Active'
9	9	44	'Male'	75.5218	222.2269	1.8750e+03	'Inactive'
10	10	24	'Female'	62.5327	144.5377	1.9022e+03	'Inactive'
11	11	44	'Male'	73.5189	179.1951	1.7726e+03	'Inactive'
12	12	38	'Male'	55.3979	213.1399	1.7168e+03	'Inactive'
13	13	29	'Female'	61.6995	143.1719	1500	'Active'

	id	age	gender	height	initweight	calintake	jobstatus
14	14	23	'Male'	65.4305	185.0747	1.8586e+03	'Inactive'
15	15	22	'Female'	65.4774	134.0369	1.9647e+03	'Inactive'
16	16	39	'Female'	57.7803	133.4423	1500	'Active'
17	17	22	'Female'	69.5728	139.8936	1500	'Active'
18	18	24	'Male'	69.1721	177.2989	1500	'Active'
19	19	41	'Male'	77.9292	174.8174	1500	'Active'
20	20	22	'Female'	61.9492	140.2975	1500	'Active'
21	21	22	'Male'	73.5419	197.3122	1500	'Active'
22	22	27	'Female'	63.2412	141.8073	1.7054e+03	'Inactive'
23	23	34	'Male'	85.2456	202.2675	2.0625e+03	'Inactive'
24	24	25	'Female'	67.2219	137.1239	1500	'Active'
25	25	39	'Male'	62.1298	197.8195	1500	'Active'
26	26	30	'Female'	75.8700	126.4978	2.0191e+03	'Inactive'
27	27	29	'Male'	71.3347	180.5776	2.0153e+03	'Inactive'
28	28	41	'Male'	72.4046	196.5634	1500	'Active'
29	29	24	'Female'	66.8003	125.0450	1500	'Active'
30	30	33	'Female'	58.8487	163.8223	1.8111e+03	'Inactive'
31	31	24	'Female'	60.9871	169.6279	1500	'Active'
32	32	22	'Female'	67.1167	139.3932	1.9243e+03	'Inactive'
33	33	47	'Female'	66.2543	161.1063	2.0790e+03	'Inactive'
34	34	41	'Female'	59.1782	141.5105	1.8624e+03	'Inactive'
35	35	39	'Female'	62.4401	131.4985	1.8279e+03	'Inactive'
36	36	46	'Female'	69.2459	151.9187	1500	'Active'
37	37	26	'Male'	52.4626	159.0762	1500	'Active'
38	38	36	'Male'	71.0037	188.0001	1500	'Active'
39	39	18	'Male'	65.3908	202.1782	1.9541e+03	'Inactive'
40	40	21	'Male'	66.5177	220.0976	1.7408e+03	'Inactive'
41	41	43	'Female'	60.8930	120.0183	1500	'Active'
42	42	36	'Female'	73.1585	139.3641	1.8599e+03	'Inactive'
43	43	23	'Female'	56.2075	147.1721	1.6190e+03	'Inactive'
44	44	48	'Female'	59.1720	119.0975	1.7044e+03	'Inactive'
45	45	22	'Female'	65.7152	132.9015	2.0385e+03	'Inactive'
46	46	29	'Female'	68.5144	131.2718	2.0528e+03	'Inactive'

	id	age	gender	height	initweight	calintake	jobstatus
47	47	36	'Male'	63.7203	160.1505	1500	'Active'
48	48	21	'Male'	63.8213	185.5053	1.7352e+03	'Inactive'
49	49	34	'Female'	65.7873	133.1923	1.9241e+03	'Inactive'
50	50	22	'Female'	66.9509	132.7390	1500	'Active'
51	51	38	'Female'	64.7826	137.1880	1500	'Active'
52	52	30	'Female'	59.1492	123.1797	1500	'Active'
53	53	21	'Male'	67.7952	194.1069	1.7509e+03	'Inactive'
54	54	20	'Female'	57.2832	153.9002	2.0870e+03	'Inactive'
55	55	43	'Male'	83.7546	186.2477	1500	'Active'
56	56	39	'Male'	69.2562	178.1039	1.9263e+03	'Inactive'
57	57	45	'Female'	70.9013	135.8652	1.9161e+03	'Inactive'
58	58	36	'Female'	68.5303	154.2571	1500	'Active'
59	59	45	'Male'	78.7450	219.5455	1500	'Active'
60	60	23	'Male'	88.5646	176.1487	1500	'Active'
61	61	24	'Male'	50.3550	216.5725	1.7155e+03	'Inactive'
62	62	40	'Female'	58.6235	151.4327	1500	'Active'
63	63	29	'Male'	69.7131	172.9651	1500	'Active'
64	64	31	'Female'	62.0872	146.0219	1.8256e+03	'Inactive'
65	65	44	'Male'	71.9858	144.1522	1500	'Active'
66	66	24	'Female'	74.9814	129.3345	1500	'Active'
67	67	21	'Female'	68.9830	155.6516	1500	'Active'
68	68	26	'Female'	65.1914	147.3916	1500	'Active'
69	69	43	'Male'	64.7788	199.2751	1.9540e+03	'Inactive'
70	70	28	'Male'	77.8480	162.5924	1500	'Active'
71	71	29	'Male'	78.1216	215.8337	1.8717e+03	'Inactive'
72	72	33	'Male'	68.0995	179.5369	1.9410e+03	'Inactive'
73	73	37	'Male'	57.8204	147.9261	1500	'Active'
74	74	38	'Male'	53.1875	226.8471	2.0768e+03	'Inactive'
75	75	39	'Female'	67.9117	135.5261	2.0115e+03	'Inactive'
76	76	32	'Male'	46.4906	191.8603	1.9402e+03	'Inactive'
77	77	39	'Male'	65.3603	193.2385	1.6569e+03	'Inactive'
78	78	20	'Female'	65.8864	151.8960	2.0488e+03	'Inactive'
79	79	29	'Male'	79.6971	208.9313	1.6445e+03	'Inactive'

	id	age	gender	height	initweight	calintake	jobstatus
80	80	44	'Male'	69.3296	179.0395	1.8644e+03	'Inactive'
81	81	36	'Female'	64.4863	140.2773	1.6088e+03	'Inactive'
82	82	31	'Male'	72.1059	192.7995	1500	'Active'
83	83	20	'Female'	58.7973	136.7891	1.8251e+03	'Inactive'
84	84	27	'Female'	69.5145	122.6887	1500	'Active'
85	85	24	'Female'	66.8786	138.3935	1500	'Active'
86	86	32	'Male'	74.4864	187.8576	1500	'Active'
87	87	32	'Male'	67.9866	202.6006	1.9907e+03	'Inactive'
88	88	33	'Male'	69.7979	219.0734	1.6630e+03	'Inactive'
89	89	31	'Female'	70.1493	122.9622	1.7778e+03	'Inactive'
90	90	32	'Male'	63.8500	174.6370	1.8546e+03	'Inactive'
91	91	20	'Female'	63.7745	122.3978	1500	'Active'
92	92	47	'Female'	55.2011	136.2520	1500	'Active'
93	93	47	'Male'	52.8228	178.3886	1500	'Active'
94	94	38	'Female'	58.4764	131.4541	1500	'Active'
95	95	30	'Female'	71.3329	122.7571	1.6219e+03	'Inactive'
96	96	39	'Female'	66.5982	136.9215	1500	'Active'
97	97	39	'Female'	69.0131	133.8118	1.6230e+03	'Inactive'
98	98	46	'Female'	64.7141	138.5840	1500	'Active'
99	99	37	'Female'	64.1979	116.7309	1500	'Active'
100	100	46	'Male'	60.1285	175.9285	1500	'Active'

⋮

Active Versus Inactive & Average Age Difference between Active and Inactive

How many patients have "Active" jobstatus, and how many are "Inactive"?

```
% Use SQL query to get the number of 'Active' and 'Inactive' patients
query = 'SELECT jobstatus, COUNT(jobstatus) FROM patient GROUP BY jobstatus';
num_jobstatus = fetch(conn, query)
```

num_jobstatus = 2×2 table

	jobstatus	COUNT_jobstatus_
1	'Active'	129
2	'Inactive'	121

Average Age of Patients with Active & Inactive `jobstatus`

What is the average age of the patients with Active & Inactive jobstatus

```
% Use SQL query to get the average age of patients in the different groups
query = 'SELECT jobstatus, AVG(age) FROM patient GROUP BY jobstatus';
fetch(conn, query)
```

ans = 2×2 table

	jobstatus	AVG_age_
1	'Active'	33.0233
2	'Inactive'	32.8843

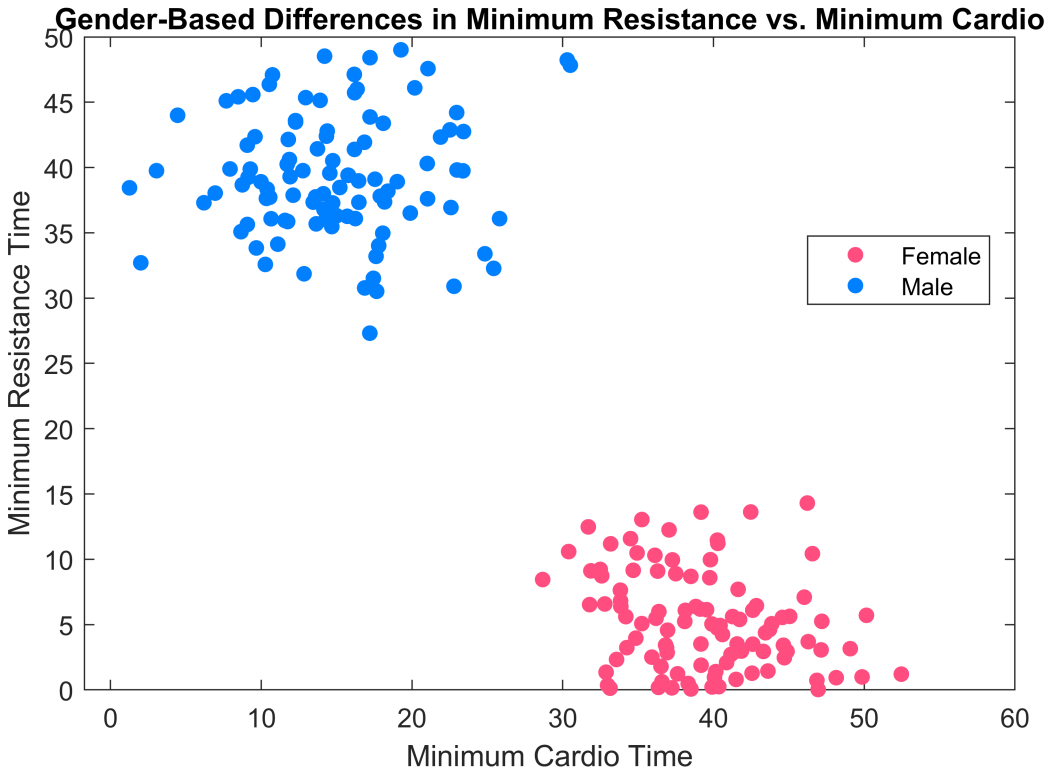
```
% Alternatively, use a cell array to compute the means
means = splitapply(@mean, [dbdata{:,strcmpi(cols, 'age')}]', findgroups(dbdata(:,strcmp(cols,
fprintf(['The mean age of Active patients was %.2f years.\n' ...
            'The mean age of Inactive patients was %.2f years.'], means)
```

```
The mean age of Active patients was 33.02 years.
The mean age of Inactive patients was 32.88 years.
```

Gender-Specific Differences in Exercise Preference

Use the data to answer whether there ae gender-specific differences in exercise preference (resistance vs cardio) and whether the subject does ANY exercise. Remember to not include patiemts with missing exercise information in the analysis.

```
% Create mask for patients with no exercise
I = (data.timecardio == 0) | (data.timeresist == 0);

% Initiate figure
figure('Position', [0 0 600 400]);
% Create a scatter plot comparing timecardio and timeresist for patients
% who exercised grouped by gender
gscatter(data.timecardio(~I), data.timeresist(~I), data.gender(~I), [1 .3 .5; 0 .5 1], [], [20
xlabel('Minimum Cardio Time');
ylabel('Minimum Resistance Time');
title('Gender-Based Differences in Minimum Resistance vs. Minimum Cardio')
```



Gender Figure Caption:

This figure shows the minimum time spent resistance training plotted against the minimum time spent doing cardio for all the patients involved in the study. The pink data points are female patients and the blue data points are male.

Gender Figure Conclusion:

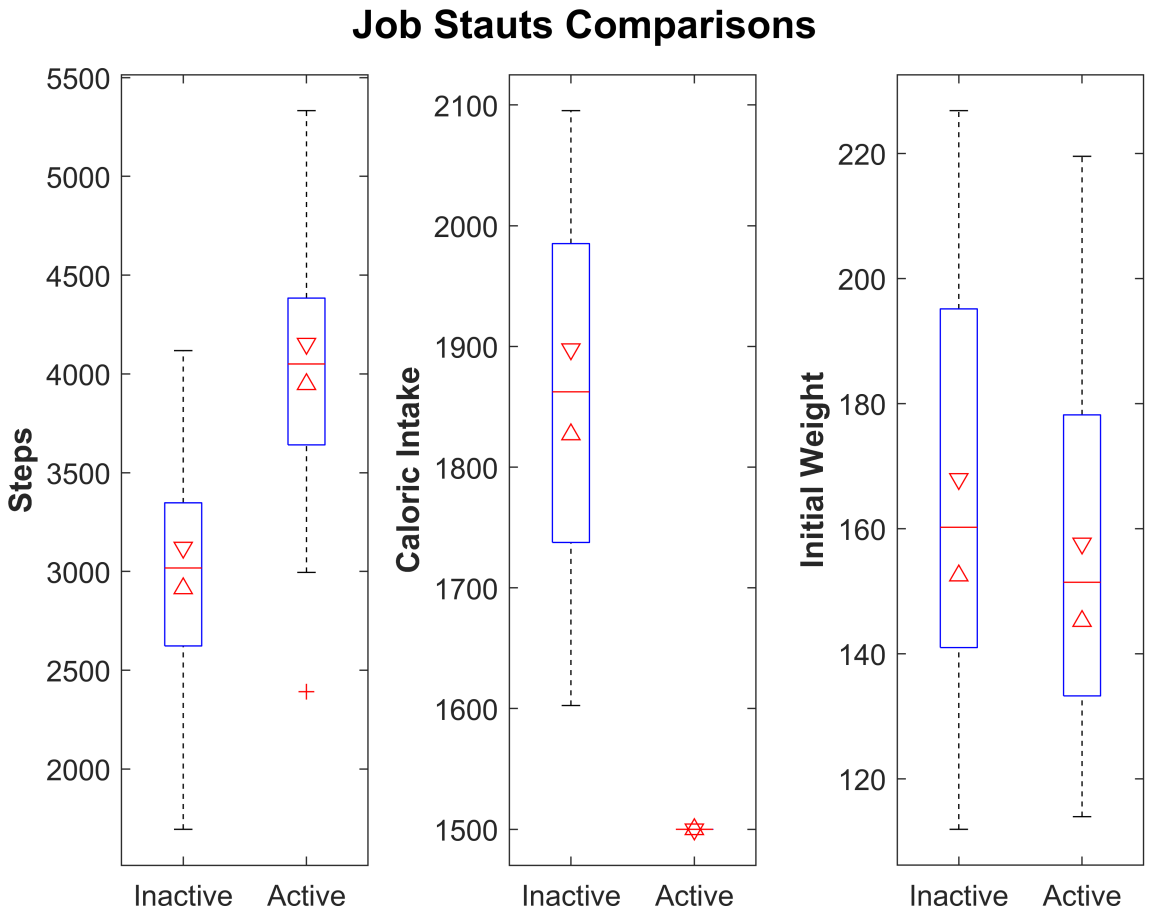
The figure shows clear separation between the male and female patients. Female patients spend significantly more time on cardio training that they do on resistance training. The reverse is true for male patients. Therefore, we can conclude that female patients had a preference for cardio while male patients had a preference for resistance training.

Differences between Job Status

Use the data (box plots) to answer whether the job status influences the average number of steps a subject takes per day and the average number of calories consumed. Does job status affect the initial weight?

```
% Create the Job Status figure

% Create plots in 1-by-3 subplot grid
figure('Position', [0 0  700 500]);
sgtitle('Job Stauts Comparisons', 'FontWeight', 'bold')

% Create Boxplot: steps
subplot(1,3,1);
boxplot(data.steps, data.jobstatus, 'Notch', 'marker');
set(gca, 'FontSize', 11);
```

```
ylabel('Steps', 'FontWeight', 'bold', 'FontSize', 12);

% Create Boxplot: caloric intake
subplot(1,3,2);
boxplot(data.calintake, data.jobstatus, 'Notch', 'marker');
set(gca, 'FontSize', 11);
ylabel('Caloric Intake', 'FontWeight', 'bold', 'FontSize', 12);

% Create Boxplot: Initial Weigh
subplot(1,3,3);
boxplot(data.initweight, data.jobstatus, 'Notch', 'marker');
set(gca, 'FontSize', 11);
ylabel('Initial Weight', 'FontWeight', 'bold', 'FontSize', 12);
```



Jobstatus Figure Caption:

The figure shows boxplots comparing *Steps* (first from left), *Caloric Intake* (second from left), and *Initial Weight* (last from left) in active and inactive patients. The red lines indicate the medians of each group, the triangular notches indicate the comparison interval at the 5% significance level; two groups are significantly different if their comparison intervals do not overlap.

Jobstatus Figure Conclusion:

We can conclude that there is a significant difference in *Steps* and *Caloric Intake* between active and inactive patients. Active patients had a higher step count, and a remarably lower caloric intake, while incative patients had a lower step count, and a higher caloric intake. There was no significant difference between the initial weight of active and inactive patients.

Effects of Sleep

Plenty of research has shown that sleep is important for weight loss. Use this data to explore this question. Is it correlated with the weight change? Is it different between genders or job statuses (active or inactive)? Is it correlated with anything else in this dataset?

What Variables are Significantly Correlated with Sleep, and Why?

```
% Select columns to test for correlation
cols = {'sleep', 'age', 'height', 'initweight', 'deltaweight',...
          'calintake', 'timecardio', 'timeresist', 'steps'};
% Compute correlation coefficients and p-values
[R, p] = corrcoef(data{:, cols});
% Show variables that correlated significantly with sleep (alpha = 0.01)
sig = find(p(1, :) < .01);
for i = 1:numel(sig)
    fprintf( '%s is correlated significantly with sleep (R^2 = %.4f, p = %.2e) \n',...
                cols{sig(i)}, R(1, sig(i))^2, p(1, sig(i)) )
end
```

```
deltaweight is correlated significantly with sleep (R^2 = 0.0438, p = 8.74e-04)
steps is correlated significantly with sleep (R^2 = 0.0278, p = 8.21e-03)
```

Sleep seems to be correlated significantly with weight loss (`deltaweight`) and `steps`.

Does Job Status Affect Sleep, and Why?

```
% Create mask for active patients
I = strcmp(data.jobstatus, 'Active');

% Run T-Tests for sleep in active & inactive patients
myttest(data.sleep, I, 'amount of sleep gotten', 'active and inactive');
```

```
The amount of sleep gotten by active and inactive patients is not significantly different (p = 2.3634e-01).
```

The patient job status does not affect sleep since there is no significant difference between the amount of sleep gotten by the groups.

Does Gender Affect Sleep, and Why?

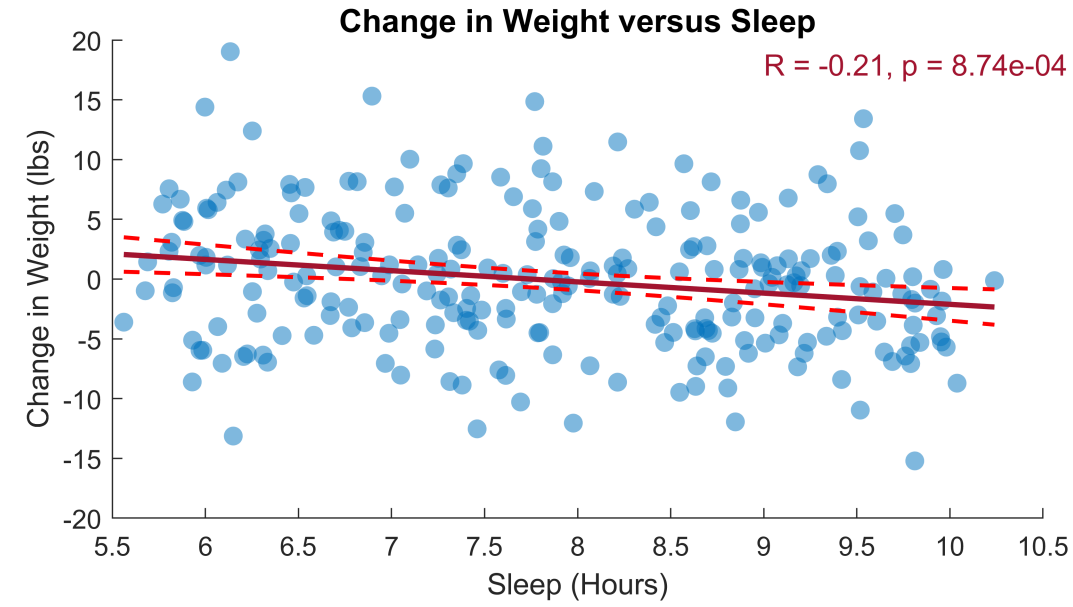
```
% Create mask for male patients
I = strcmp(data.gender, 'Male');

% Run T-Tests for sleep in male & female patients
myttest(data.sleep, I, 'amount of sleep gotten', 'male and female');
```

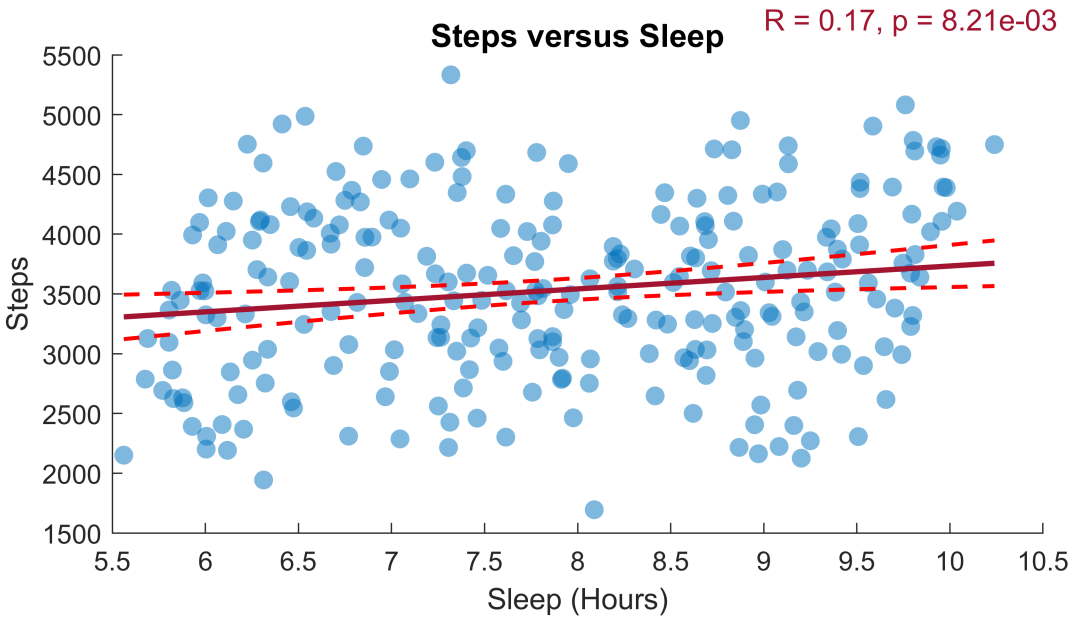
```
The amount of sleep gotten by male and female patients is not significantly different (p = 2.9769e-01).
```

The patient gender does not affect sleep since there is no significant difference between the amount of sleep gotten by the groups.

A Scatter Plot Comparing Sleep with deltaWeight and Steps

```
figure('Position', [0 0 600 700]);

% A scatter plot between sleep and deltaweight
subplot(2,1,1);
hold on;
scatter(data.sleep, data.deltaweight, 50, 'filled',...
          'MarkerFaceColor', '#0072BD', 'MarkerFaceAlpha', 0.5);
xlabel('Sleep (Hours)', 'FontSize', 11);
ylabel('Change in Weight (lbs)', 'FontSize', 11);
title('Change in Weight versus Sleep', 'FontSize', 12, 'FontWeight', 'bold');
% Fit data to a linear regression model
mdl = fitlm(data.sleep, data.deltaweight);
% Compute confidence intervals and trendline
[trend, ci] = predict(mdl, sort(data.sleep));
% Add trendline and confidence intervals to figure
plot(sort(data.sleep), trend, '-', 'Color', '#A2142F', 'LineWidth', 2);
plot(sort(data.sleep), ci, '--r', 'LineWidth', 1.5);
hold off;
% Show R^2 and p values on figure
[R, p] = corrcoef(data.sleep, data.deltaweight);
text(9, 18, sprintf('R = %.2f, p = %.2e', R(2), p(2)),...
       'Color', '#A2142F', 'FontSize', 11)

% A scatter plot between sleep and steps
subplot(2,1,2); hold on;
scatter(data.sleep, data.steps, 50, 'filled',...
          'MarkerFaceColor', '#0072BD', 'MarkerFaceAlpha', 0.5);
xlabel('Sleep (Hours)', 'FontSize', 11);
ylabel('Steps', 'FontSize', 11);
title('Steps versus Sleep', 'FontSize', 12, 'FontWeight', 'bold');
% Fit data to a linear regression model
mdl = fitlm(data.sleep, data.steps);
% Compute confidence intervals and trendline
[trend, ci] = predict(mdl, sort(data.sleep));
% Add trendline and confidence intervals to figure
plot(sort(data.sleep), trend, '-', 'Color', '#A2142F', 'LineWidth', 2);
plot(sort(data.sleep), ci, '--r', 'LineWidth', 1.5);
hold off;
% Show R^2 and p values on figure
[R, p] = corrcoef(data.sleep, data.steps);
text(9, 5800, sprintf('R = %.2f, p = %.2e', R(2), p(2)),...
       'Color', '#A2142F', 'FontSize', 11)
```





Sleep Figure Caption

This figure shows scaterplots showing the relationship between the amount of sleep gotten by patients (`sleep`) and the weight loss experienced (`deltaweight`) and the amount of steps they take (`steps`). The data is represented by the blue points. The red solid lines are trendlines fitted by least squares regression, and the red dashed lines are confidence intervals.The scatter plots are overlayed with the correlation coefficient, R, and $p - value$ of the correlations.

Sleep Figure Conclusions

There is a weak, statistically significant negative correlation between weight loss and sleep, indicating that more sleep tends to lead to greater weight loss (-ve change in weight). The is also a weak, statistically significant positive correlation between steps taken and amount of sleep; indicating that more steps is correlated with more sleep.

Combining Values to Calculate Net Calories and Compare with Weight Loss

We talked previously about inter-correlations between features when building linear models. We know that gender has effects on exercise preference and jobtype affects the number of steps. One way to alleviate those issues is to use domain knowledge to create summary values. We know from biology that what's really important for weight loss is creating a calorie deficit. We know from the data how many calories they consumed but now how much they expended.

Basal calorie and execise calorie

basal - calories burned when not active

execise - calories burned when active

```
% Equation for Calculation Basal Metabolic Rate (BMR)
% BMR as a function of parameters we measured.

% BMR calculation for men (metric)
% BMR = 66.47 + ( 13.75 x weight in kg ) + ( 5.003 x height in cm ) - ( 6.755 x age in years )
% BMR = 66.47 + ( 13.75 * 0.4536 * weight in lbs ) +
%               ( 5.003 * 2.54 * height in inches ) -
%               ( 6.755 x age in years )

% BMR calculation for women (metric)
% BMR = 655.1 + ( 9.563 x weight in kg ) + ( 1.850 x height in cm ) - ( 4.676 x age in years )
% BMR = 655.1 + ( 9.563 * 0.4536 * weight in lbs ) +
%               ( 1.850 * 2.54 * height in inches ) -
%               ( 4.676 x age in years )

% Reference: Harris JA, Benedict FG. A biometric study of human basal metabolism.
%            Proc Natl Acad Sci USA 1918;4(12):370-3.
% Website: http://www.globalrph.com/harris-benedict-equation.htm

% Define lambda functions for computing BMR for each gender
mBMR = @(w,h,a) 66.47 + 13.75*.4536*w + 5.003*2.54*h - 6.755*a;
fBMR = @(w,h,a) 655.1 + 9.563*.4536*w + 1.850*2.54*h - 4.676*a;

% Preallocate BMR column
data.BMR = zeros(height(data), 1);
% Calculate the BMR for male patients
I = strcmp(data.gender, 'Male');
data.BMR(I) = mBMR(data.initweight(I), data.height(I), data.age(I));
% Calculate the BMR for female patients
I = strcmp(data.gender, 'Female');
data.BMR(I) = fBMR(data.initweight(I), data.height(I), data.age(I));
```

```
% Google around and find estimates for the amount of calories burned during
% aerobic exercise and resistance exercise.

% average age used to determine calories burned per hour
fprintf('The average weight of the patients is %.2flbs', mean(data.initweight))
```

```
The average weight of the patients is 161.47lbs
```

```
%   Resistance Training: 422 cal/hour
%   High Impact Aerobics: 493 cal/hour
% Reference: Calories burned during exercise, activities, sports and workouts. (n.d.).
%            Retrieved October 31, 2021, from https://www.nutristrategy.com/caloriesburned.htm

% Compute the total exercise calories (resistance calories + cardio calories)
data.calexercise = (data.timeresist * 422/60) + (data.timecardio * 493/60);

% Calculate the net calorie change
% Net Cal = Cal Intake - Exercise Cal - Basal Cal
data.netcal = data.calintake - data.calexercise - data.BMR;

% Create Change in Weight vs Net Calories Figure
figure;
gscatter(data.netcal, data.deltaweight, data.gender, [1 .3 .5; 0 .5 1], [], [20 20]);
xlabel('Net Calories');
ylabel('Change in Weight');
title('Change in Weight vs Net Calories')
% Fit data to a linear regression model
mdl = fitlm(data.netcal, data.deltaweight);
% Compute trendline
trend = predict(mdl, sort(data.netcal));
% Add trendline and confidence intervals to figure
hold on; plot(sort(data.netcal), trend, '-k', 'LineWidth', 3); hold off;
% Show R^2 and p values on figure
[R, p] = corrcoef(data.netcal, data.deltaweight);
text(-1100, 15, sprintf('R = %.2f,\np = %.2e', R(2), p(2)),...
       'Color', '#A2142F', 'FontSize', 11)
legend({'Female', 'Male', 'Trend Line'}, 'Location', 'southeast')
```

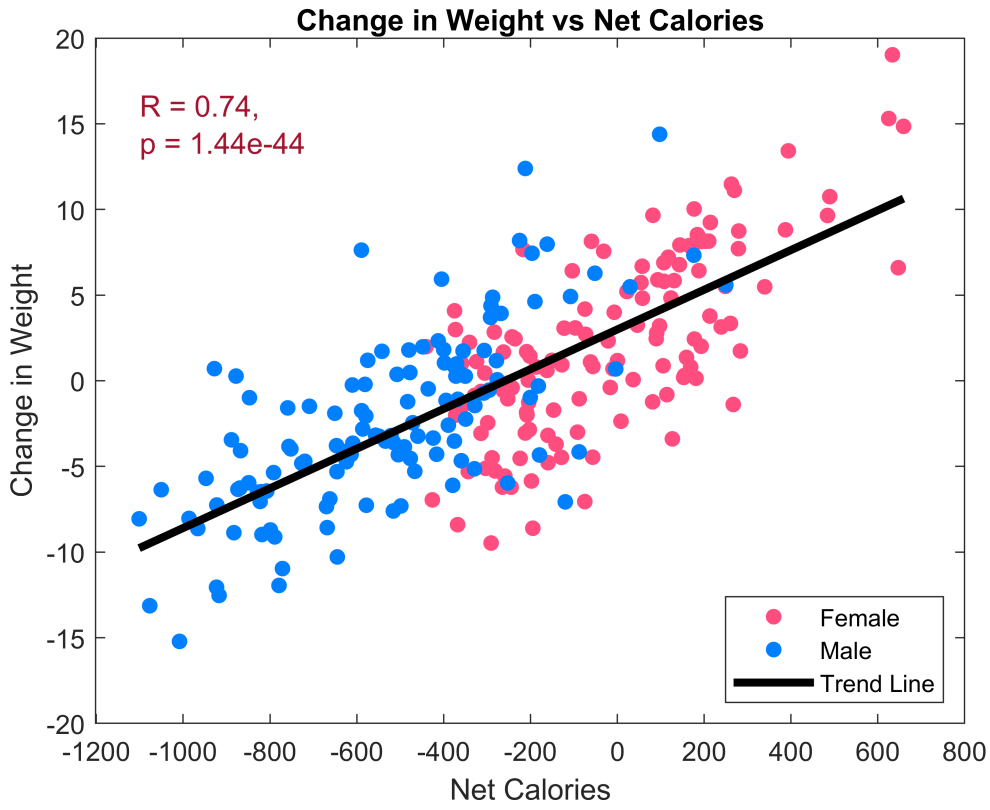


Figure Caption

This figure is a scatter plot showing the change in weight plotted against the net calories burned by the patients. The pink data points are female patients, and the blue data points are male patients. The trend line shows the correlatiion between the variables. The figure is also overlayed with the R and p-values.

Conclusions

There is a strong, statistically significant positive correlation between net calories and weight loss. This indicates that calorie loss is strongly correlated with, and is a fairly good predictor of weight loss $(R^2 = 0.55)$. Also, male patients tended to exhibit more weight loss (and net calorie change) than female patients as evidenced by the point separation in the figure.

Build a Model

Use all of this information to build the best model to predict the change in weight for each subject. Plot your predictions versus reality. Calculate the R^2 and discuss what it implies. Discuss the effect sizes and their relevance to weight loss.

Practice `corrcoef()` and `ttest2()`

```
% Determine what variables are significantly correlated with change in
% weight.
% Select columns to test for correlation
cols = {'deltaweight', 'netcal', 'sleep', 'steps', 'age', 'height', 'initweight',...
          'calintake', 'timecardio', 'timeresist'};
```

```
% Compute correlation coefficients and p-values
[R, p] = corrcoef(data{:, cols});
% Show variables that correlated significantly with sleep (alpha = 0.01)
sig = find(p(1, :) < .01);
for i = 1:numel(sig)
    fprintf('%s is correlated significantly with weight loss (R^2 = %.4f, p = %.2e) \n',...
                cols{sig(i)}, R(1, sig(i))^2, p(1, sig(i)) )
end
```

```
netcal is correlated significantly with weight loss (R^2 = 0.5473, p = 1.44e-44)
sleep is correlated significantly with weight loss (R^2 = 0.0438, p = 8.74e-04)
height is correlated significantly with weight loss (R^2 = 0.0921, p = 1.01e-06)
initweight is correlated significantly with weight loss (R^2 = 0.0557, p = 1.65e-04)
calintake is correlated significantly with weight loss (R^2 = 0.1787, p = 2.96e-12)
timeresist is correlated significantly with weight loss (R^2 = 0.2903, p = 3.18e-20)
```

Does Job Status Affect Weight Loss?

```
% Create mask for active patients
active = strcmp(data.jobstatus, 'Active');

% Run T-Tests for sleep in active & inactive patients
myttest(data.deltaweight, active, 'weight loss', 'active and inactive');
```

```
The weight loss by active and inactive patients is significantly different (p = 1.8546e-08).
```

Does Gender Affect Weight Loss?

```
% Create mask for male patients
male = strcmp(data.gender, 'Male');

% Run T-Tests for sleep in active & inactive patients
myttest(data.deltaweight, male, 'weight loss', 'male and female');
```

```
The weight loss by male and female patients is significantly different (p = 1.7937e-10).
```

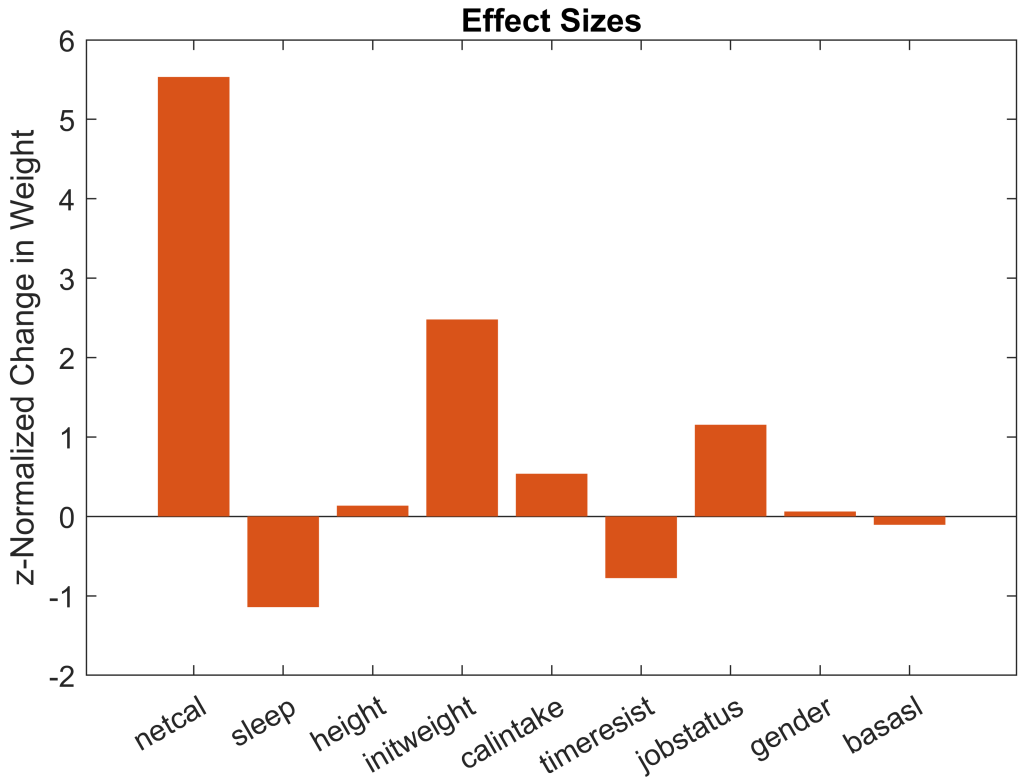
Of the variables tested, 6 showed a significant correlation with weight loss, and 2 (job status and gender) had a significant impact on weight loss. These eight variables will be included in the multivariate regression model.

Multivariate Normal Regression

`mvregress(X,Y)` returns the estimated coefficients for a multivariate normal regression of the *d*-dimensional responses in Y on the design matrices in X.

```
% Initialize the X variable to have 9 columns: netcal, sleep, height,
% initweight, calintake, timeresist, jobstatus, gender, and a ninth column
% containing just ones (Basal).
X = [data{:, {'netcal', 'sleep', 'height', 'initweight', 'calintake', 'timeresist'}},...
       active, male, ones(height(data), 1)];
% Normalize X data for more accurate model fitting
X = normalize(X);
X(:, end) = 1;

% Initialize the Y variable to be change in weight
Y = data.deltaweight;
```

```
% Perform multivariate regression on the data
beta = mvregress(X, Y);

% Compute predictions
YPred = X * beta;

% Discuss the effect sizes and their relecance to weight loss
% Compare effect sizes
figure('Position', [0 0 600 400]);
bar(beta, 'FaceColor', '#D95319', 'LineStyle', 'none');
ylabel('z-Normalized Change in Weight', 'FontSize', 12);
title('Effect Sizes', 'FontSize', 14);
set(gca, 'xticklabel', {'netcal', 'sleep', 'height', 'initweight', 'calintake',...
                        'timeresist', 'jobstatus', 'gender', 'basasl'}, 'FontSize', 11);
```



Effect Size Figure Caption

This figure is a barplot showing the effect size for the 9 different components of our multivariate regression model. The data was normalized beforehand so the bar heights represent the z-normalized change in weight for each of the different components.

Effect Size Figure Conclusion

The net calorie change appears to have lagest (positive) effect on change in weight, followed by the initial weight, amount of sleep gotten, job status, and time spent on resistance training.

```
% Create scatter plot comparing observations and predictions
figure;
scatter(Y, YPred, 50, 'filled', 'MarkerFaceColor', '#0072BD', 'MarkerFaceAlpha', 0.4);
```

```
xlabel('Observed {\Delta}weight', 'FontSize', 12);
ylabel('Predicted {\Delta}weight', 'FontSize', 12);
title('Multivariate Regression Results', 'FontSize', 13);
% Fit a trend line to the data
p = polyfit(Y, YPred, 1);
% Add trend line to data
hold on;
plot(-20:20, (-20:20)*p(1) + p(2), '-', 'LineWidth', 2, 'Color', '#A2142F');
plot(-20:20, -20:20, '--k');
hold off;
% Compute correlation between observed and predicted data
% Show R^2 and p values on figure
[R, p] = corrcoef(Y, YPred);
text(-18, 15, sprintf('R^2 = %.2f,\np = %.2e', R(2)^2, p(2)),...
       'Color', '#A2142F', 'FontSize', 11);
% Add legend
legend({'Data', 'Trendline', 'Ideal'}, 'Location', 'southeast', 'FontSize', 10);
```

Model Figure Caption

This figure is a scatterplot showing the observed change in weight plotted against the change in weight predcted by our model. The data is represented by the blue dots. The black dashed line represents an ideal model - one that perfectly predicts the observed data. The red solid line is a trend line indicating the correlation between the predictions and observations.

Model Figure Conclusion

With $R^2 = 0.69$, our model appears to be a fairly good predictor of the actual weight loss experineced by patients in this study.

```
function myttest(x, I, ttxt, gtxt)
% This function runs ttests and prints out a statement about the results in
% the command line.
  % Input Arguments
    %   x -> numeric vector containing data to perform ttest on
    %   I -> logical vector used to select the two different groups
    %   gtxt -> string specifying groups (for printing in fprinf)
    [h,p] = ttest2(x(I), x(~I), 'Alpha', 0.01);
    if h
        fprintf(['The ' ttxt ' by ' gtxt ' patients is significantly different (p = %.4e).\n'],
    else
        fprintf(['The ' ttxt ' by ' gtxt ' patients is not significantly different (p = %.4e).\
    end
end
```