Exercises for Linear Regression

Jordan Boyd-Graber
Digging into Data

February 17, 2014

1 Linear Regression Formula

Prediction $\hat{y} = f(\mathbf{x})$ given observation $\mathbf{x}$.

$$f(\mathbf{x}) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \tag{1}$$

Distribution of y given $\mathbf{x}$:

$$p(y \mid x) = y \sim N\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j; \sigma^2\right) \tag{2}$$

2 Example Weights

dimension	weight
β_0	1
β_1	2.0
β_2	-1.0
σ	1.0

3 Example Inputs

What are the predictions for these inputs?

1. $\mathbf{x}_1 = \{0.0, 0.0\}$
2. $\mathbf{x}_2 = \{1.0, 1.0\}$
3. $\mathbf{x}_3 = \{.5, 2\}$

4 Probability of Observations given Inputs

1. $\mathbf{x}_1 = \{0.0, 0.0\}$; $y_1 = 1$
2. $\mathbf{x}_2 = \{1.0, 1.0\}$; $y_2 = 3$
3. $\mathbf{x}_3 = \{.5, 2\}$; $y_3 = -1$

5 Newspaper Prediction

The dataset is about the daily and Sunday newspaper circulations of fifty newspapers in the united states (although the data are a couple of years old): `http://goo.gl/RwzqjY`. To be more manageable, the numbers have been scaled down by 1000 (that is, if the readership of a paper is one million, four hundred thousand, that will appear as 1400 in this file).

Often, newspapers in smaller market will have daily editions but no Sunday editions. One common problem is attempting to predict, given an existing readership, what the Sunday readership will be. So well model this as a regression, predicting the Sunday readership from the daily readership.

1. Plot the sunday readership against the daily readership. Does a regression make sense?
2. What is the relationship between daily readership and sunday readership?
3. Write the equation between daily readership (x) and Sunday readership (y).
4. What do your errors look like? Plot the difference between the sunday circulation and the predicted Sunday circulation. What's the variance of this distribution?
5. Imagine that youre an analyst for a newspaper with a daily circulation of 800 thousand. Predict the Sunday circulation of this hypothetical newspaper and give the range that would encompass one standard deviation of the normal distribution induced by this prediction (use the variance estimated from the previous question).

6 Regularized Regression

Create a dataset with two nuisance variables:

```
 data(mtcars)
 mtcars <- cbind(runif(nrow(mtcars)), runif(nrow(mtcars)), mtcars)
colnames(mtcars)[1:2] <- c("dummy1", "dummy2")
```

Fit regularized L1 and L2 regression:

```
 library(glmnet)
reg.l2 <- glmnet(features, target, alpha=0)
reg.l1 <- glmnet(features, target, alpha=1)
```

Plot L_1 coefficients vs. λ (do the same thing for L_2):

```
library(ggplot2)
models <- data.frame(t(rbind(matrix(reg.l1$lambda, nrow=1), as.matrix(reg.l1$beta))))
colnames(models)[1] <- "lambda"
models <- melt(models, c("lambda"))
ggplot(models) + aes(x=log(lambda), y=value, color=variable) + geom_line()
```

Use cross validation to determine the best λ for L_1 (do the same thing for L_2):

```
cv.l1 <- cv.glmnet(features, target, alpha=1)
plot(cv.l1, main = "L1 n-fold cross validation error")
```

Normal Density Table

To use:

1. Let's say that you want to look up the value for $a.bc$
2. Find the row that corresponds to $a.b$
3. **In that row**, find the column for $0.0c$. That entry is your answer for $p(y = a.bc \,|\, \mu = 0.0, \sigma = 1.0)$.
4. For example, the probability of observing $y = 0.12$ is 0.3961
5. Negative number? Just look up the positive value (the distribution is symmetric)
6. Non-zero mean (but $\sigma = 1$)? Just look up $y - \mu$.

Table 1. **Standard normal density function**

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.3989	0.3989	0.3989	0.3988	0.3986	0.3984	0.3982	0.3980	0.3977	0.3973
0.1	0.3970	0.3965	0.3961	0.3956	0.3951	0.3945	0.3939	0.3932	0.3925	0.3918
0.2	0.3910	0.3902	0.3894	0.3885	0.3876	0.3867	0.3857	0.3847	0.3836	0.3825
0.3	0.3814	0.3802	0.3790	0.3778	0.3765	0.3752	0.3739	0.3725	0.3712	0.3697
0.4	0.3683	0.3668	0.3653	0.3637	0.3621	0.3605	0.3589	0.3572	0.3555	0.3538
0.5	0.3521	0.3503	0.3485	0.3467	0.3448	0.3429	0.3410	0.3391	0.3372	0.3352
0.6	0.3332	0.3312	0.3292	0.3271	0.3251	0.3230	0.3209	0.3187	0.3166	0.3144
0.7	0.3123	0.3101	0.3079	0.3056	0.3034	0.3011	0.2989	0.2966	0.2943	0.2920
0.8	0.2897	0.2874	0.2850	0.2827	0.2803	0.2780	0.2756	0.2732	0.2709	0.2685
0.9	0.2661	0.2637	0.2613	0.2589	0.2565	0.2541	0.2516	0.2492	0.2468	0.2444
1.0	0.2420	0.2396	0.2371	0.2347	0.2323	0.2299	0.2275	0.2251	0.2227	0.2203
1.1	0.2179	0.2155	0.2131	0.2107	0.2083	0.2059	0.2036	0.2012	0.1989	0.1965
1.2	0.1942	0.1919	0.1895	0.1872	0.1849	0.1826	0.1804	0.1781	0.1758	0.1736
1.3	0.1714	0.1691	0.1669	0.1647	0.1626	0.1604	0.1582	0.1561	0.1539	0.1518
1.4	0.1497	0.1476	0.1456	0.1435	0.1415	0.1394	0.1374	0.1354	0.1334	0.1315
1.5	0.1295	0.1276	0.1257	0.1238	0.1219	0.1200	0.1182	0.1163	0.1145	0.1127
1.6	0.1109	0.1092	0.1074	0.1057	0.1040	0.1023	0.1006	0.0989	0.0973	0.0957
1.7	0.0940	0.0925	0.0909	0.0893	0.0878	0.0863	0.0848	0.0833	0.0818	0.0804
1.8	0.0790	0.0775	0.0761	0.0748	0.0734	0.0721	0.0707	0.0694	0.0681	0.0669
1.9	0.0656	0.0644	0.0632	0.0620	0.0608	0.0596	0.0584	0.0573	0.0562	0.0551
2.0	0.0540	0.0529	0.0519	0.0508	0.0498	0.0488	0.0478	0.0468	0.0459	0.0449
2.1	0.0440	0.0431	0.0422	0.0413	0.0404	0.0396	0.0387	0.0379	0.0371	0.0363
2.2	0.0355	0.0347	0.0339	0.0332	0.0325	0.0317	0.0310	0.0303	0.0297	0.0290
2.3	0.0283	0.0277	0.0270	0.0264	0.0258	0.0252	0.0246	0.0241	0.0235	0.0229
2.4	0.0224	0.0219	0.0213	0.0208	0.0203	0.0198	0.0194	0.0189	0.0184	0.0180
2.5	0.0175	0.0171	0.0167	0.0163	0.0158	0.0154	0.0151	0.0147	0.0143	0.0139
2.6	0.0136	0.0132	0.0129	0.0126	0.0122	0.0119	0.0116	0.0113	0.0110	0.0107
2.7	0.0104	0.0101	0.0099	0.0096	0.0093	0.0091	0.0088	0.0086	0.0084	0.0081
2.8	0.0079	0.0077	0.0075	0.0073	0.0071	0.0069	0.0067	0.0065	0.0063	0.0061
2.9	0.0060	0.0058	0.0056	0.0055	0.0053	0.0051	0.0050	0.0048	0.0047	0.0046
3.0	0.0044	0.0043	0.0042	0.0040	0.0039	0.0038	0.0037	0.0036	0.0035	0.0034
3.1	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026	0.0025	0.0025
3.2	0.0024	0.0023	0.0022	0.0022	0.0021	0.0020	0.0020	0.0019	0.0018	0.0018
3.3	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014	0.0013	0.0013
3.4	0.0012	0.0012	0.0012	0.0011	0.0011	0.0010	0.0010	0.0010	0.0009	0.0009
3.5	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007	0.0007	0.0007	0.0006
3.6	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0004
3.7	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003	0.0003	0.0003	0.0003
3.8	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002	0.0002	0.0002	0.0002	0.0002
3.9	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001
4.0	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001