

Homework 7

Due Monday, March 7 before 5:00pm

Use [Live Editor > Save > Export to PDF] to prepare your submission for Gradescope.

This assignment uses data from the MAT file HW7_data.mat. Download this file and run

```
load HW7_data.mat
```

to load variable `blood` into the workspace.

Blood Metabolite Diagnostic for Fungal Infections

You are asked to design a diagnostic for bloodborne fungal infections. Ideally, you would measure the number of colony forming units (CFUs) of fungus per ml of blood. However, the fungus is slow growing outside the body, so accurate CFU counts take weeks. Instead, you hope to use standard measurements from a blood metabolic panel to predict the CFUs/ml in a sample.

The Matlab table `blood` contains data from a 250-patient clinical trial. Each datapoint has values for all 14 standard blood metabolite readings:

Metabolite	Variable Name	Units
albumin	albumin	g/dL
alkaline phosphatase	alk_phos	IU/L
alanine aminotransferase	ALT	IU/L
aspartate aminotransferase	AST	IU/L
blood urea nitrogen	BUN	mg/dL
calcium	Ca	mg/dL
chloride	Cl	mmol/L
carbon dioxide	CO2	mmol/L
creatinine	creatinine	mg/dL
glucose	glucose	mg/dL
potassium	K	mEq/L
sodium	Na	mEq/L
total bilirubin	bilirubin	mg/dL
total protein	protein	g/dL

The `blood` table also contains the $\log(\text{CFU})$ counts for each sample. (Note that we use $\log(\text{CFU})$ since CFU counts vary exponentially.)

a.) Using linear regression, build a model that predicts $\log(\text{CFU})$ counts with blood metabolite readings. You do not need to include interactions in your model.

```
% place your code here
fitlm(blood, 'logCFU ~ albumin + alk_phos + ALT + AST + BUN + Ca + Cl + CO2 + creatinine
```

```
ans =
```

```
Linear regression model:
```

```
logCFU ~ 1 + albumin + alk_phos + ALT + AST + BUN + Ca + Cl + CO2 + creatinine + glucose + K + Na + b
```

```
Estimated Coefficients:
```

Estimate	SE	tStat	pValue
----------	----	-------	--------

(Intercept)	-7.1892	9.8616	-0.72901	0.46672
albumin	0.61654	0.64125	0.96147	0.3373
alk_phos	-0.00096736	0.023028	-0.042007	0.96653
ALT	0.062953	0.0908	0.69331	0.4888
AST	-0.031801	0.099961	-0.31813	0.75067
BUN	0.73146	0.1547	4.7283	3.905e-06
Ca	-0.014919	0.031124	-0.47934	0.63214
Cl	-0.797	0.31327	-2.5441	0.011596
CO2	0.00067067	0.11774	0.0056964	0.99546
creatinine	1.5406	2.4793	0.62137	0.53496
glucose	-0.0091501	0.033522	-0.27296	0.78513
K	0.29381	0.6809	0.43149	0.66651
Na	0.010902	0.022427	0.48613	0.62733
bilirubin	3.278	1.6991	1.9292	0.054906
protein	0.027454	0.4327	0.063448	0.94946

Number of observations: 250, Error degrees of freedom: 235
 Root Mean Squared Error: 5.08
 R-squared: 0.127, Adjusted R-Squared 0.0751
 F-statistic vs. constant model: 2.45, p-value = 0.00312

For this example, we will assess the statistical significance of the coefficients by considering only those with $p < 0.05$. Which metabolite readings are significantly predictive of the CFU counts? Do these metabolite levels increase or decrease as the fungus count increases?

Significant metabolites are:

BUN: increases with increasing CFU count

Cl: decreases with increasing CFU count

b.) What is the RMSE for your model? What are its units?

RMSE = 5.08 log(CFUs)

c.) Build another model using only the significant predictors. Does the RMSE change?

```
fitlm(blood, 'logCFU ~ BUN + Cl')
```

ans =

Linear regression model:
 logCFU ~ 1 + BUN + Cl

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	2.0587	3.6099	0.5703	0.56899
BUN	0.69208	0.14922	4.638	5.7082e-06
Cl	-0.76006	0.30062	-2.5283	0.012087

Number of observations: 250, Error degrees of freedom: 247
 Root Mean Squared Error: 5.02
 R-squared: 0.102, Adjusted R-Squared 0.0951
 F-statistic vs. constant model: 14.1, p-value = 1.61e-06

The RMSE decreases slightly to 5.02 log(CFUs).

d.) During sepsis, the number of fungal cells in the blood increases by 100 fold. Would your original model be able to predict this level of change using metabolites? Why or why not?

No. The RMSE of both models is greater than $5 \log(\text{CFUs})$. Predicting a 100-fold change in CFUs would require an RMSE below $\log(100) \log(\text{CFUs})$.