Homework 3

Due Friday, March 5 before 5:00pm

For Part 1, submit your handwritten work using Gradescope. For Parts 2-4, use [Live Editor > Save > Export to PDF] to prepare your submission for Gradescope.

Part 1: Deriving an estimator for $y = \beta_1 x$

In class we derive least-squares estimators for the linear models $y = \beta_0$ and $y = \beta_0 + \beta_1 x$. For this exercise you will derive a formula to fit a single parameter model $y = \beta_1 x$ to a set of n datapoints.

a.) Begin with the total quadratic loss $\sum_{i=1}^{n} (y^{\text{pred}} - y^{\text{true}})^2$. Write this expression after substituting the linear model for y^{pred}.

$$\sum_{i=1}^{n} (\beta_1 x_i^{\text{true}} - y_i^{\text{true}})^2$$

b.) Our goal is to minimize the total loss, which can be found when the partial derivative of the loss with respect to the parameter β_1 is zero. Find and expression for β_1.

$$\frac{d}{d\beta_1} \sum_{i=1}^{n} (\beta_1 x_i^{\text{true}} - y_i^{\text{true}})^2 = 0$$

$$\sum_{i=1}^{n} (\frac{d}{d\beta_1} (\beta_1 x_i^{\text{true}} - y_i^{\text{true}}))^2 = 0$$

$$2 \sum_{i=1}^{n} (\beta_1 x_i^{\text{true}} - y_i^{\text{true}}) x_i^{\text{true}} = 0$$

$$\beta_1 \sum_{i=1}^{n} (x_i^{\text{true}})^2 - \sum_{i=1}^{n} (y_i^{\text{true}} x_i^{\text{true}}) = 0$$

$$\beta_1 = \sum_{i=1}^{n} (y_i^{\text{true}} x_i^{\text{true}}) / \sum_{i=1}^{n} (x_i^{\text{true}})^2$$

c.) Using your formula for β_1, fit the model $y = \beta_1 x$ to the five data points in the table on page 58 of the textbook. Compare your value of β_1 to the value found when fitting the data to the model $y = \beta_0 + \beta_1 x$ in section 8.3.2.

$$\beta_1 = \frac{-0.05 * 0.07 + 0.40 * 0.16 + 0.66 * 0.48 + 0.65 * .68 + 1.12 * 0.83}{0.07^2 + 0.16^2 + 0.48^2 + 0.68^2 + 0.83^2} = 1.238$$

$$PercentDifference = \frac{|1.238 - 1.21|}{\frac{1.238 + 1.21}{2}} * 100 = 2.29\%$$ Very close!

d.) We want to be sure that we're minimizing, not maximizing the sum squared error. Using a second derivative test, show that your estimate for β_1 is a minimum. Be sure to explain your reasoning.

$$\frac{d^2}{(d\beta_1)^2}\sum_{i=1}^{n}(\beta_1 x_i^{\text{true}} - y_i^{\text{true}})^2$$

If β_1 is a minimum, then the above equation should solve for a positive value.

$$\frac{d}{d\beta_1}2\sum_{i=1}^{n}(\beta_1(x_i^{\text{true}})^2 - y_i^{\text{true}}x_i^{\text{true}})$$

$$\frac{d}{d\beta_1}2(\sum_{i=1}^{n}(\beta_1(x_i^{\text{true}})^2) - \sum_{i=1}^{n}(y_i^{\text{true}}x_i^{\text{true}}))$$

$= 2\sum_{i=1}^{n}(x_i^{\text{true}})^2$ Which is always positive.

Parts 2-4 use data from the MAT file HW3_data.mat. Download this file and run

```
clear
close all
clc
load HW3_data.mat
```

to load variables `x`, `y`, `blood`, and `ecm` into the workspace.

Part 2: Polynomial Fitting

Variables `x` and `y` contain 12 values from an unknown cubic polynomial, i.e.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

Using the values `x` and `y`, compute estimates for parameters $\beta_0, \ldots, \beta_3$ using linear regression. **For this problem, you are not allowed to use `fitlm`, `regress`, `polyfit`, or any other linear regression or curve fitting tools.** You must construct the design matrix and calculate parameter estimates via pseudoinversion.

```
X2 = [ones(1,12);x';(x.^2)';(x.^3)']' %set up design matrix
```

```
X2 = 12×4
    1.0000   -2.0000    4.0000   -8.0000
    1.0000   -1.6364    2.6777   -4.3817
    1.0000   -1.2727    1.6198   -2.0616
    1.0000   -0.9091    0.8264   -0.7513
    1.0000   -0.5455    0.2975   -0.1623
    1.0000   -0.1818    0.0331   -0.0060
    1.0000    0.1818    0.0331    0.0060
    1.0000    0.5455    0.2975    0.1623
    1.0000    0.9091    0.8264    0.7513
    1.0000    1.2727    1.6198    2.0616
     ⋮
```

```
X_p = pinv(X2) %set up pseudoinverse
```

```
X_p = 4×12
```

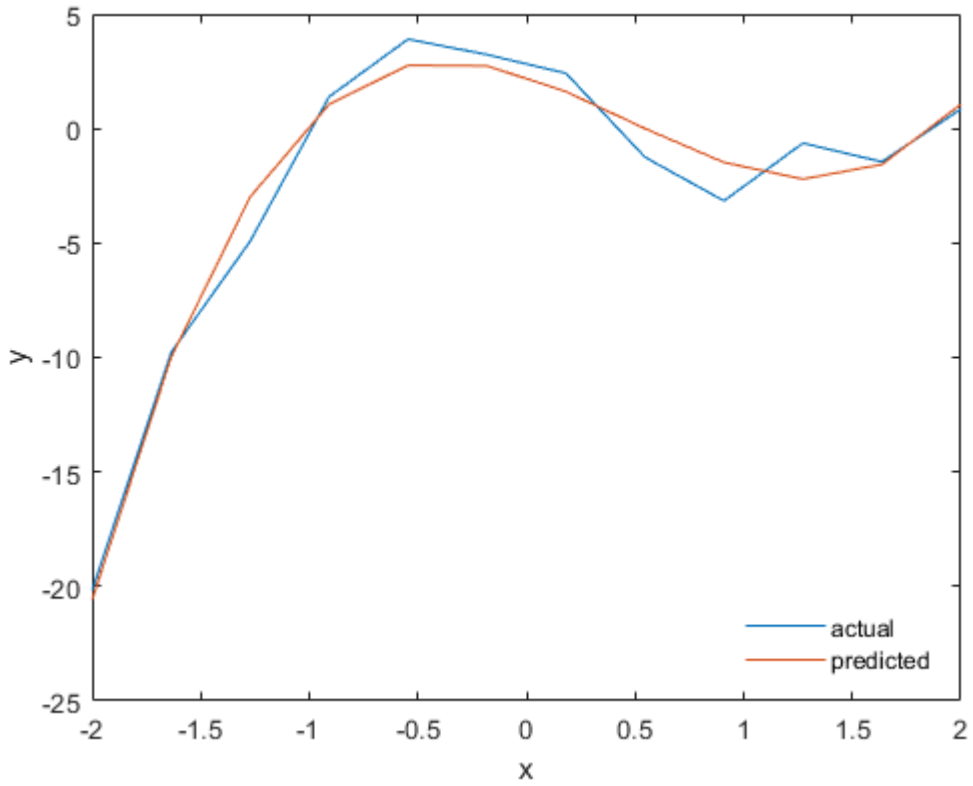
```
   -0.0804    0.0089    0.0804    0.1339    0.1696    0.1875    0.1875    0.1696 ···
    0.1440   -0.1092   -0.2262   -0.2373   -0.1726   -0.0626    0.0626    0.1726
    0.1039    0.0472    0.0019   -0.0321   -0.0548   -0.0661   -0.0661   -0.0548
   -0.0889    0.0081    0.0566    0.0673    0.0512    0.0189   -0.0189   -0.0512
```

```
B = X_p*y %calculate betas 0-3
```

```
B = 4×1
    2.2753
   -3.1669
   -3.0092
    2.1444
```

Using your parameter estimates, plot the points in variables `x` and `y` and a line corresponding to the best fit polynomial. Both the points and the line should be on the same plot.

```
plot(x,y) %actual data
hold on
plot(x, X2*B) %predicted
xlabel('x');
ylabel('y');
legend('actual', 'predicted', 'Location', 'southeast');
legend('boxoff')
hold off
```

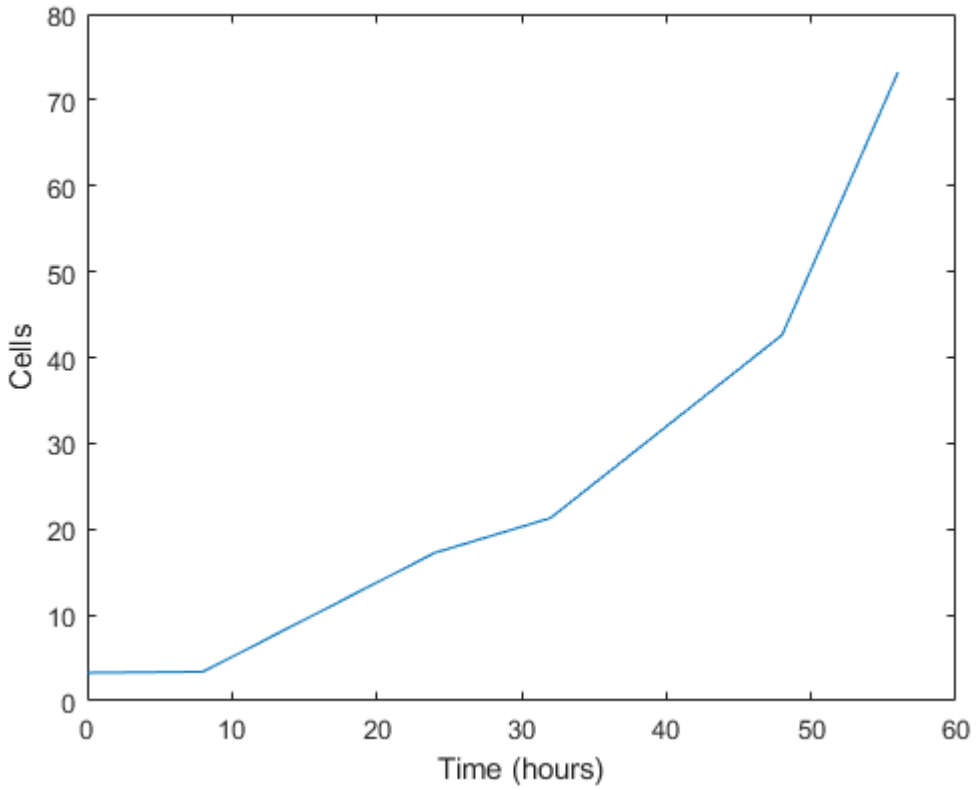


Part 3: Cell Growth

Variables `t` and `cells` contain six cell counts for dividing mammalian cells in a culture dish. (The times in `t` are in hours.) Your task is to find the exponential growth rate of the cells using linear regression. **For this problem, you are not allowed to use `fitlm`, `regress`, `polyfit`, or any other linear regression or curve fitting tools.**

a.) Plot the number of cells over time.

```
plot(t, cells)
xlabel('Time (hours)');
ylabel('Cells');
```



b.) Set up a design matrix for the linearized exponential growth equation from section 9.4.

```
X3 = [ones(1,6); t']'
```

```
X3 = 6×2
     1     0
     1     8
     1    24
     1    32
     1    48
     1    56
```

c.) Calculate the pseudoinverse of the design matrix and use it to fit your model.

```
X3_p = pinv(X3)
```

```
X3_p = 2×6
```

```
    0.4933    0.4000    0.2133    0.1200   -0.0667   -0.1600
   -0.0117   -0.0083   -0.0017    0.0017    0.0083    0.0117
```

```
cells3 = log(cells)
```

```
cells3 = 6×1
    1.1856
    1.2199
    2.8472
    3.0584
    3.7525
    4.2936
```

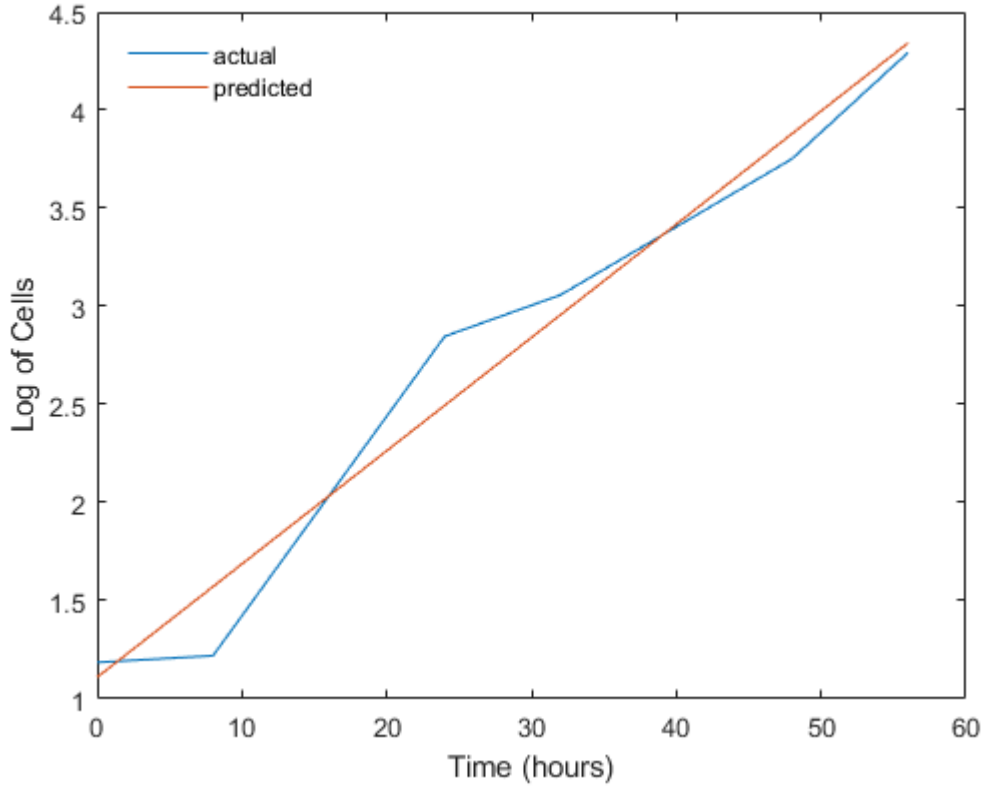
```
B3 = X3_p*(cells3)
```

```
B3 = 2×1
    1.1101
    0.0577
```

```
plot(t, cells3) %actual data
hold on
y3 = X3*B3
```

```
y3 = 6×1
    1.1101
    1.5719
    2.4953
    2.9571
    3.8805
    4.3423
```

```
plot(t, y3) %predicted
xlabel('Time (hours)');
ylabel('Log of Cells');
legend('actual', 'predicted', 'Location', 'northwest');
legend('boxoff')
hold off
```



d.) Calculate the exponential growth rate of the cells. What are the units?

$\mu = 0.0577 hour^{-1}$

e.) Use the fitted parameters to find the inital number of cells. How does this value compare with the number of cells at t = 0 h in your data?

$exp(ln(N_0) + \mu * t) = exp(1.1101) = 3.035 cells$ This estimate is 7.54% difference.

Part 4: Blood Metabolite Diagnostic for Fungal Infections

You are tasked with diagnosing a bloodborne fungal infection. Ideally, you would measure the number of colony forming units (CFUs) 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(CFU) counts for each sample. (*Note that we use log(CFU) since CFU counts vary exponentially. This is unrelated to the logit function or logistic regression.*)

a.) Using linear regression, build a model that predicts log(CFU) counts with blood metabolite readings.

```
X4 = table2array(blood(:,1:14));
y4 = table2array(blood(:,15));
fitlm(X4,y4)
```

```
ans =
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 + x12 + x13 + x14

Estimated Coefficients:
                    Estimate        SE         tStat       pValue
                   ___________    _________    _________    _________

    (Intercept)        -7.1892       9.8616     -0.72901      0.46672
    x1                 0.61654      0.64125      0.96147       0.3373
    x2             -0.00096736     0.023028    -0.042007      0.96653
    x3                0.062953       0.0908      0.69331       0.4888
    x4               -0.031801     0.099961     -0.31813      0.75067
    x5                 0.73146       0.1547       4.7283    3.905e-06
    x6               -0.014919     0.031124     -0.47934      0.63214
    x7                  -0.797      0.31327      -2.5441     0.011596
    x8              0.00067067      0.11774    0.0056964      0.99546
    x9                  1.5406       2.4793      0.62137      0.53496
    x10             -0.0091501     0.033522     -0.27296      0.78513
    x11                0.29381       0.6809      0.43149      0.66651
    x12               0.010902     0.022427      0.48613      0.62733
    x13                  3.278       1.6991       1.9292     0.054906
    x14               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
```

Which metabolite readings are significantly predictive of the CFU counts? Do these metabolite levels increase or decrease as the fungus count increases?

Significant (p < 0.05) metabolites inculde:

x5: blood urea nitrgen (Coeff = 0.731) --> Increase with increased fungus CFUs

x7: cloride (Coeff = -0.797) --> Decreases with increased fungus CFUs

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

No. The uncertainty in the model's predictions (as given by the RMSE), is +/- exp(5.08) >> 100. A 100-fold change is too small to be detected since it falls within the 95% CI of the predictions, so we cannot be sure the change in prediction is due to anything other than chance.