Kate Lassiter

Problem 1 (written) – 25 points

Imagine you have a sequence of N observations $(x_1, \ldots, x_N)$, where each $x_i \in \{0, 1, 2, \ldots, \infty\}$. You model this sequence as i.i.d. from a Poisson distribution with unknown parameter $\lambda \in \mathbb{R}_+$, where

$$p(X|\lambda) = \frac{\lambda^X}{X!}\mathrm{e}^{-\lambda}$$

(a) What is the joint likelihood of the data $(x_1, \ldots, x_N)$?

$x_i \sim_{i.i.d} P(\lambda)$

$P(X|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda} = f(x) \rightarrow Poisson\ r.v$

Recall:

By independence: $f(x_1|x_2) = f(x_1)$

$f(x_1,x_2)/f(x_2) = f(x_1)$

$f(x_1,x_2) = f(x_1)f(x_2)$

Joint likelihood = likelihood function x prior density = joint probability density function of data x and λ

Thus:

$f(x_1,\ldots,x_n) = \frac{\lambda^{x_1}}{x_1!}e^{-\lambda} \times \frac{\lambda^{x_2}}{x_2!}e^{-\lambda} \times \ldots \times \frac{\lambda^{x_n}}{x_n!}e^{-\lambda}$

$$\boxed{f(x_1,\ldots,x_n) = \frac{e^{-\lambda n}\lambda^{\sum_{i=1}^{n} x_i}}{x_1! \times x_2! \ldots \times x_{n-1}! \times x_n!}}$$

(b) Derive the maximum likelihood estimate λ_{ML} for λ.

Recall:

log(ab) = loga + logb

log(a/b) =loga - logb

$\log a^b = b\log a$

d/dx(logx)=1/x

Thus:

$\log(f(x_1,\ldots,x_n|\lambda)) = -\lambda n + \sum_{i=1}^{n} x_i log\lambda - logC$

$\frac{d}{d\lambda}\log(f(x_1,\ldots,x_n|\lambda)) = -n + \frac{\sum_{i=1}^{n} x_i}{\lambda}$

$\frac{d}{d\lambda}\log(f(x_1,\ldots,x_n|\lambda)) = 0$

$0 = -n + \frac{\sum_{i=1}^{n} x_i}{\lambda}$

$n = \frac{\sum_{i=1}^{n} x_i}{\lambda}$

$\lambda n = \sum_{i=1}^{n} x_i$

$$\boxed{\lambda = \frac{\sum_{i=1}^{n} x_i}{n}}$$

Kate Lassiter

To help learn λ, you use a prior distribution. You select the distribution $p(\lambda) = \text{gamma}(a, b)$.

(c) Derive the maximum a posteriori (MAP) estimate λ_{MAP} for λ?

$P(\lambda) = Gamma\,(a,b) \rightarrow Prior$

$P(\lambda) = \frac{\lambda^{a-1}e^{-b\lambda}b^{a}}{\Gamma(a)} = f(x) \rightarrow Gamma\,r.v$

Recall:

$\Gamma(a) = (a-1)! = Gamma\,function$

$\text{MAP} = \max(P(\lambda|y,x))$

$P(\lambda|y,x) = \frac{P(y|\lambda,x)P(\lambda)}{P(y|x)}$

$= max(log(\frac{P(y|\lambda,x)P(\lambda)}{P(y|x)}))$

$= max(log(P(y|\lambda,x)) + log(P(\lambda)) - log(P(y|x)))$

Thus:

$P(y|\lambda,x)P(\lambda) = (\frac{\lambda^{x}}{x!}e^{-\lambda}) \times (\frac{\lambda^{a-1}e^{-b\lambda}b^{a}}{\Gamma(a)})$

$P(y|\lambda,x)P(\lambda) = \frac{b^{a}}{\Gamma(a)x!}[(\lambda^{x}e^{-\lambda}) \times (\lambda^{a-1}e^{-b\lambda})]$

$P(y|\lambda,x_1,\dots,x_n)P(\lambda) = \frac{b^{a}}{\Gamma(a)x_1,\dots,x_n!}[(\lambda^{\sum_{i=1}^{n}x_i}e^{-\lambda n}) \times (\lambda^{a-1}e^{-b\lambda})]$

$P(y|\lambda,x_1,\dots,x_n)P(\lambda) \propto [(\lambda^{\sum_{i=1}^{n}x_i}e^{-\lambda n}) \times (\lambda^{a-1}e^{-b\lambda})]$

$P(y|\lambda,x_1,\dots,x_n)P(\lambda) \propto \lambda^{\sum_{i=1}^{n}x_i+a-1}e^{-\lambda n-b\lambda}$

$\log(P(y|\lambda,x_1,\dots,x_n)P(\lambda)) \propto (\sum_{i=1}^{n}x_i + a - 1)log(\lambda) - \lambda n - b\lambda$

$\frac{d}{d\lambda}\log(P(y|\lambda,x_1,\dots,x_n)P(\lambda)) \propto \frac{\sum_{i=1}^{n}x_i+a-1}{\lambda} - n - b$

$\frac{d}{d\lambda}\log(P(y|\lambda,x_1,\dots,x_n)P(\lambda)) = 0$

$0 = \frac{\sum_{i=1}^{n}x_i+a-1}{\lambda} - n - b$

$n + b = \frac{\sum_{i=1}^{n}x_i+a-1}{\lambda}$

$\frac{n+b}{\sum_{i=1}^{n}x_i+a-1} = \frac{1}{\lambda}$

$$\boxed{\frac{\sum_{i=1}^{n}x_i+a-1}{n+b} = \lambda}$$

(d) Use Bayes rule to derive the posterior distribution of λ and identify the name of this distribution.

$P(\lambda|x) = \frac{P(x|\lambda)P(\lambda)}{P(X)}$

$P(x|\lambda)P(\lambda) = (\frac{\lambda^{\sum_{i=1}^{n}x_i}}{x_1!\dots x_n!}e^{-\lambda n}) \times (\frac{\lambda^{a-1}e^{-b\lambda}b^{a}}{\Gamma(a)})$

$P(x|\lambda)P(\lambda) = \frac{b^{a}}{\Gamma(a)x_1!\dots x_n!}(\lambda^{\sum_{i=1}^{n}x_i+a-1}e^{-\lambda n-b\lambda})$

$P(x) = \int_{0}^{\infty}\frac{b^{a}}{\Gamma(a)x_1!\dots x_n!}(\lambda^{\sum_{i=1}^{n}x_i+a-1}e^{-\lambda n-b\lambda})\,d\lambda$

Kate Lassiter

$$P(x) = \frac{b^a}{\Gamma(a)x_1!\ldots x_n!}\int_0^\infty (\lambda^{\sum_{i=1}^n x_i + a - 1} e^{-\lambda n - b\lambda})\, d\lambda$$

$$P(x|\lambda)P(\lambda) = \frac{\frac{b^a}{\Gamma(a)x_1!\ldots x_n!}(\lambda^{\sum_{i=1}^n x_i + a - 1} e^{-\lambda n - b\lambda})}{\frac{b^a}{\Gamma(a)x_1!\ldots x_n!}\int_0^\infty (\lambda^{\sum_{i=1}^n x_i + a - 1} e^{-\lambda n - b\lambda})\, d\lambda}$$

$$P(x|\lambda)P(\lambda) = \frac{(\lambda^{\sum_{i=1}^n x_i + a - 1} e^{-\lambda n - b\lambda})}{\int_0^\infty (\lambda^{\sum_{i=1}^n x_i + a - 1} e^{-\lambda n - b\lambda})\, d\lambda}$$

$$P(x|\lambda)P(\lambda) \propto \lambda^{\sum_{i=1}^n x_i + a - 1} e^{-\lambda n - b\lambda}$$

$$P(\lambda|x) \propto \lambda^{(\sum_{i=1}^n x_i + a) - 1} e^{-\lambda(b+n)}$$

$$P(\lambda|x) \propto \lambda^{(\sum_{i=1}^n x_i + a) - 1} e^{-\lambda(b+n)}$$

$$P(\lambda|x) \propto \text{Gamma}(\sum_{i=1}^n x_i + a,\ n + b)$$

(e) What is the mean and variance of λ under this posterior? Discuss how it relates to λ_{ML} and λ_{MAP}.

Recall:

$\mu = \frac{a}{b} \rightarrow Gamma\ r.v$

$$E[\mu] = \frac{\sum_{i=1}^n x_i + a}{n+b}$$

Decomposing:

$$E[\mu] = \frac{\sum_{i=1}^n x_i}{n+b} + \frac{a}{n+b}$$

$$E[\mu] = (\frac{\sum_{i=1}^n x_i}{n})(\frac{n}{n+b}) + \frac{a}{n+b}$$

$$E[\mu] = (\frac{\sum_{i=1}^n x_i}{n})(\frac{n}{n+b}) + (\frac{a}{b})(\frac{b}{n+b})$$

$\uparrow\lambda_{MLE}$ $\uparrow$Prior mean

The result is a weighted average of the Poisson λ_{MLE} given the data and the prior distribution, Gamma's, mean. Also, the mean = λ_{MAP} without one subtracted from the numerator. As n goes to infinity, it converges to λ_{MLE}.

Recall:

$\Sigma = \frac{a}{b^2} \rightarrow Gamma\ r.v$

$$Var[\lambda] = \frac{\sum_{i=1}^n x_i + a}{(n+b^2)}$$

Variance converges to 0 as n goes to infinity.

Kate Lassiter

Problem 2 (written) – 20 points

(a) You have data (x_i, y_i) for $i = 1, \ldots, n$, where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. You model this as $y_i \overset{iid}{\sim} N(x_i^T w, \sigma^2)$. You use the data you have to approximate w with $w_{\text{RR}} = (\lambda I + X^T X)^{-1} X^T y$, where X and y are defined as in the lectures. Derive the results for $\mathbb{E}[w_{\text{RR}}]$ and $\mathbb{V}[w_{\text{RR}}]$ given in the slides.

$E[w_{RIDGE}]$

$y_i \sim_{i.i.d} N(x_i^T w, \sigma^2) \rightarrow Normal\ r.v$

$w_{RIDGE} = (\lambda I + X^T X)^{-1} X^T y$

$E[w_{RIDGE}] = \int [(\lambda I + X^T X)^{-1} X^T y] p(y|X,w) dy$

$E[w_{RIDGE}] = [(\lambda I + X^T X)^{-1} X^T y] \int p(y|X,w) dy$

$E[w_{RIDGE}] = [(\lambda I + X^T X)^{-1} X^T y][1]$

$E[w_{RIDGE}] = E[(\lambda I + X^T X)^{-1} X^T y]$

Recall:

$(\lambda I + X^T X)^{-1} X^T$ is a constant

E[y] = Xw

Thus:

$E[w_{RIDGE}] = (\lambda I + X^T X)^{-1} X^T E[y]$

$$E[w_{RIDGE}] = (\lambda I + X^T X)^{-1} X^T X w$$

$Var[w_{RIDGE}]$

Recall:

$Var[w_{OLS}] = \sigma^2 (X^T X)^{-1}$

$w_{OLS} = (X^T X)^{-1} X^T y$

$E[y] = X w_{OLS}$

Thus:

$w_{RIDGE} = (\lambda I + X^T X)^{-1} X^T y$

$w_{RIDGE} = (\lambda I + X^T X)^{-1} X^T X w_{OLS}$

$Var[w_{RIDGE}] = Var[(\lambda I + X^T X)^{-1} X^T X w_{OLS}]$

Recall:

$Covar[kX] = k Covar[X] k^T$

Thus:

$Var[w_{RIDGE}] = Var[(\lambda I + X^T X)^{-1} X^T X w_{OLS}]$

$Var[w_{RIDGE}] = [(\lambda I + X^T X)^{-1} X^T X][Var[w_{OLS}]][(\lambda I + X^T X)^{-1} X^T X]^T$

Kate Lassiter

$$\text{Var}[w_{RIDGE}] = [(\lambda I + X^T X)^{-1} X^T X][\sigma^2 (X^T X)^{-1}][(\lambda I + X^T X)^{-1} X^T X]^T$$

$$\text{Var}[w_{RIDGE}] = [(\lambda I + X^T X)^{-1} X^T X][\sigma^2 (X^T X)^{-1}][X^T X(\lambda I + X^T X)^{-1}]$$

$$\text{Var}[w_{RIDGE}] = [(\lambda I + X^T X)^{-1} X^T X][\sigma^2][(\lambda I + X^T X)^{-1}]$$

$$\text{Var}[w_{RIDGE}] = (\lambda I + X^T X)^{-1} X^T X \sigma^2 (\lambda I + X^T X)^{-1}$$

$$\boxed{\text{Var}[w_{RIDGE}] = \sigma^2 (\lambda I + X^T X)^{-1} X^T X (\lambda I + X^T X)^{-1}}$$

(b) If w_{RR} is the ridge regression solution and w_{LS} is the least squares solution for the above problem, derive an equation for writing w_{RR} as a function of w_{LS} and the singular values and right singular vectors of feature matrix X. Recall that the singular value decomposition of $X = USV^T$.

Recall:

$X = USV^T$

$X^T X = (USV^T)^T (USV^T)$

$X^T X = (VS^T U^T)(USV^T)$

$X^T X = VS^T U^T USV^T$

$X^T X = VS^T SV^T$

$X^T X = VS^2 V^T$

Derive function of OLS:

$w_{RIDGE} = (\lambda I + X^T X)^{-1} X^T y$

Recall:

$w_{OLS} = (X^T X)^{-1} X^T y$

E[y] = Xw_{OLS}

Thus:

$w_{RIDGE} = (\lambda I + X^T X)^{-1} X^T (X^T X)^{-1} X^T y$

$w_{RIDGE} = (\lambda I + X^T X)^{-1} X^T X (X^T X)^{-1} X^T y$

$w_{RIDGE} = (\lambda I + X^T X)^{-1} X^T X w_{OLS}$

Recall:

* Factor out $X^T X$

$(AB)^{-1} = A^{-1} B^{-1}$

Thus:

$w_{RIDGE} = [(X^T X)(\lambda (X^T X)^{-1} + I)]^{-1} X^T X w_{OLS}$

$w_{RIDGE} = \lambda (X^T X)^{-1} + I)^{-1} (X^T X)^{-1} X^T X w_{OLS}$

$w_{RIDGE} = \lambda (X^T X)^{-1} + I)^{-1} w_{OLS}$

$w_{RIDGE} = \lambda (X^T X)^{-1} + I)^{-1} w_{OLS}$

Kate Lassiter

SVD:

$(X^TX)^{-1} = VS^{-2}V^T$

$w_{RIDGE} = \lambda(X^TX)^{-1} + I)^{-1} w_{OLS}$

$w_{RIDGE} = \lambda(VS^{-2}V^T) + I)^{-1} w_{OLS}$

Recall:

$VV^T = I$

$(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$

* If orthogonal, $A^{-1} = A^T$

Thus:

$w_{RIDGE} = \lambda(VS^{-2}V^T) + VV^TI)^{-1} w_{OLS}$

$w_{RIDGE} = (\lambda VS^{-2}V^T + VV^TI)^{-1} w_{OLS}$

$w_{RIDGE} = V(\lambda S^{-2} + I)V^T)^{-1} w_{OLS}$

$w_{RIDGE} = (V^T)^{-1}(\lambda S^{-2} + I)^{-1}(V^{-1}) w_{OLS}$

$w_{RIDGE} = V(\lambda S^{-2} + I)^{-1}(V^{-1}) w_{OLS}$

$w_{RIDGE} = V(\lambda S^{-2} + I)^{-1}(V^{-1}) w_{OLS} VV^T \rightarrow Multiply\ by\ identity$

$w_{RIDGE} = V(\lambda S^{-2} + I)^{-1} w_{OLS} V^T$

$w_{RIDGE} = V(\lambda S^{-2} + I)^{-1} V^T w_{OLS}$

$w_{RIDGE} = (V(\lambda S^{-2} + I)^{-1} V^T w_{OLS})\frac{S^2}{S^2} \rightarrow Scale\ by\ singular\ values$

$$w_{RIDGE} = (V(\frac{S^2}{\lambda + S^2})V^T w_{OLS}$$

Problem 3 (coding) – 30 points

In this problem you will analyze data using the linear regression techniques we have discussed. The goal of the problem is to predict the miles per gallon a car will get using six quantities (features) about that car. The zip file containing the data can be found on Courseworks.[1] The data is broken into training and testing sets. Each row in both "X" files contain six features for a single car (plus a 1 in the 7th dimension) and the same row in the corresponding "y" file contains the miles per gallon for that car.

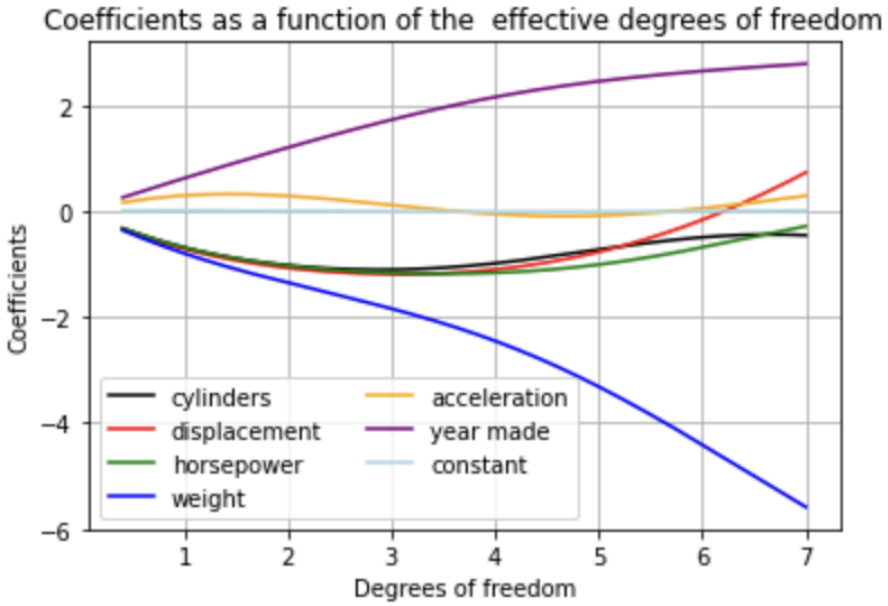
Remember to submit all original source code with your homework. Put everything you are asked to show below in the PDF file.

Part 1. Using the training data only, write code to solve the ridge regression problem

$$\mathcal{L} = \lambda\|w\|^2 + \sum_{i=1}^{350} \|y_i - x_i^T w\|^2.$$

(a) For $\lambda = 0, 1, 2, 3, \ldots, 5000$, solve for w_{RR}. (Notice that when $\lambda = 0$, $w_{\text{RR}} = w_{\text{LS}}$.) In one figure, plot the 7 values in w_{RR} as a function of $df(\lambda)$. You will need to call a built in SVD function to do this as discussed in the slides. Be sure to label your 7 curves by their dimension in x.[2]

Kate Lassiter

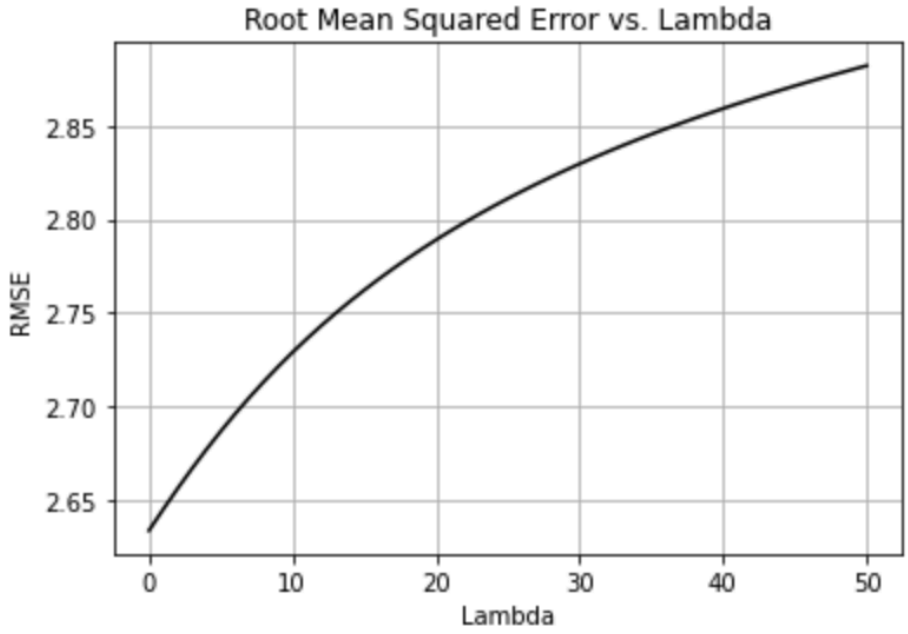


(b) Two dimensions clearly stand out over the others. Which ones are they and what information can we get from this?

The variables weight and year made both have the largest impact on the response variable, miles per gallon. When lambda equals zero and the ridge regression solution is equivalent to the least squares solution, these two variables have larger coefficients than the other variables. Given they are standardized, this means that the average response of an additional unit of one of these variables will have a greater impact on the mean miles per gallon predicted compared to other variables. As the model becomes more constrained with higher values of lambda, the effective degrees of freedom shrink as more and more of the free parameters are driven to zero. The coefficients of these two dimensions shrink to zero slower than the others, suggesting they have the greatest impact on the miles per gallon even in a very constrained model. Directions in the column space of the data matrix X having the smallest variance also correspond to those with the smallest singular values, thus ridge regression greatly reduces these directions/dimensions. Thus, those variables whose singular values are the highest will be kept up until very high values of lambda, like weight and year made. Even as the penalty grows larger and larger for adding an additional parameter to the model, these predictor variables are still added with non-zero coefficients. This means these are particularly important variables for predicting miles per gallon.

(c) For $\lambda = 0, \ldots, 50$, predict all 42 test cases. Plot the root mean squared error (RMSE)[3] on the test set as a function of λ—*not* as a function of $df(\lambda)$. What does this figure tell you when choosing λ for this problem (and when choosing between ridge regression and least squares)?

Kate Lassiter

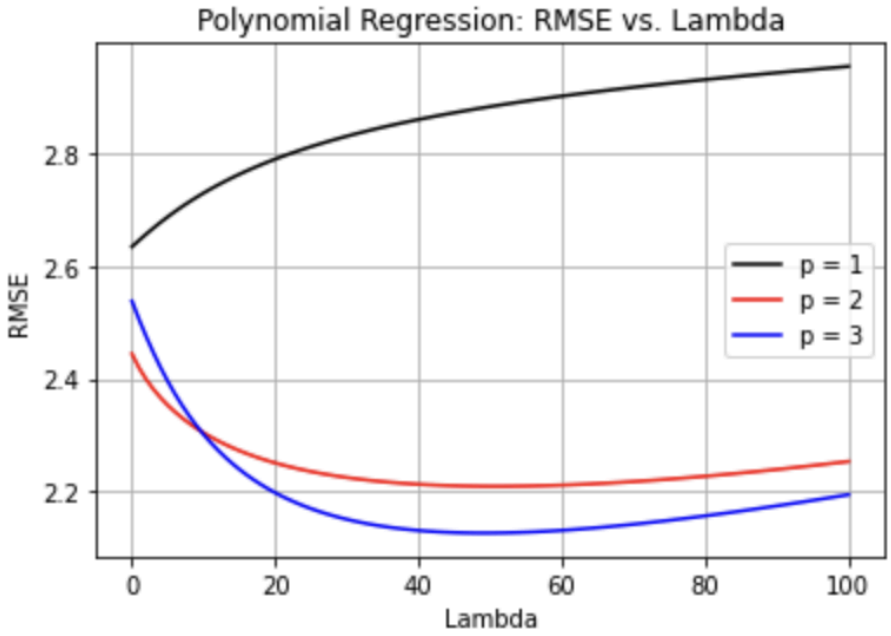


For this data set, as lambda increases, the RMSE increases as well. This is likely because important predictors are being omitted as the model becomes too highly constrained with this high lambda penalty parameter. Because the model does not have enough useful predictors, predictions for the response variable suffer. It was shown previously that lambda equals 50 results in effective degrees of freedom of 4.50027. This means that 2-3 predictors are being omitted from the model due to the regularization, and this could likely contribute to a lower root mean squared error when there is not enough information for the model to correctly predict the response. For this data set, lambda = 0 has the lowest RMSE and thus linear regression (choosing lambda =0) outperforms ridge regression on the test set.

Part 2. Modify your code to learn a pth-order polynomial regression model for $p = 1, 2, 3$. (You've already done $p = 1$ above.) For this implementation use the method discussed in the slides. Also, be sure to standardize each additional dimension of your data.

(d) In one figure, plot the test RMSE as a function of $\lambda = 0, \ldots, 100$ for $p = 1, 2, 3$. Based on this plot, which value of p should you choose and why? How does your assessment of the ideal value of λ change for this problem?

Kate Lassiter



It is clear that there are nonlinear dynamics taking place in this data. When quadratic and cubic features are added to the data set, there is a large improvement in test RMSE over regression on just the original features, regardless of the value chosen for lambda. At first, the data set with cubic features performs worse than the quadratic because there are so many parameters added to the model and many of them are unnecessary. The quadratic is able to capture some of the nonlinear dynamics without having quite as many useless predictors added to the model as the cubic. However, as we increase the value of lambda/degree of regularization, the cubic model quickly outperforms the others and achieves the lowest RMSE of all the models. By including the cubic features and using a large degree of regularization, the very important nonlinear dynamics are still captured while useless parameters are effectively forced to zero. A value of p=3 should be chosen with a high value for lambda, and this assessment has changed greatly from the model fit to just the original data set in which linear regression (lambda=0) beat out ridge regression.