# Workshop #3: More on Regression Models

TRiCAM 2017

W. Pan

# Lecture Outline

Review: Statistical Regression Models

Review: Evaluating Models

Regularization and Bayesian Models

Application: Land Use Regression

# Review: Statistical Regression Models

# Statistical Models: Summary

Recall that a statistical model for an observation $y$, called the ***response variable***, based on ***predictor variables*** $x_1, \ldots, x_J$, posits that:

1. a general mathematical relationship, $f(x_1, \ldots, x_J)$, between $y$ and $x_1, \ldots, x_J$,
2. the observed values of $y$ differ from $f(x_1, \ldots, x_J)$ by ***random noise***.

That is, a statistical model for $y$ using the predictors $x_1, \ldots, x_J$ is

$$y = f(x_1, \ldots, x_J) + \epsilon,$$

where $\epsilon$ is a ***random variable***.

In ***inference***, we fine the parameters of $f$ that minimizes a choice of ***loss function***. Learning the parameters of the function $f$ is called ***'fitting the model'***.

# Simple Linear Regression

A ***simple linear regression*** model is a statistical model where $f$ is a linear function

$$y = a_1 \cdot x + a_0 + \epsilon.$$

Suppose that $\epsilon \sim \mathcal{N}(0, \sigma^2)$, then

$$y | a_1, a_0, x, \epsilon \sim \mathcal{N}(a_1 \cdot x + a_0, \sigma^2).$$

The normal pdf $p(y|a_1, a_0, x)$ is called the ***likelihood function***, and measures how likely the observed data is under the model $a_1 \cdot x + a_0$.

For multiple observations, $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, the likelihood function is

$$\mathcal{L}(a_1, a_0) = \prod_{i=1}^{n} p(y_i | a_1, a_0, x_i) = \prod_{i=1}^{n} \mathcal{N}(y_i; a_1 \cdot x_i + a_0, \sigma^2).$$

# Maximum Likelihood Estimate Model

Given a set of observations, $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, there are good reasons to find $a_1^{MLE}$ and $a_0^{MLE}$ so that the likelihood function is maximized:

$$a_1^{MLE}, a_0^{MLE} = \underset{a_1, a_0}{\text{argmax}}\ \mathcal{L}(a_1, a_0) = \prod_{i=1}^{n} p(y_i | a_1, a_0, x_i).$$

The parameters $a_1^{MLE}$ and $a_0^{MLE}$ are called ***maximum likelihood estimates***.

When $\epsilon \sim \mathcal{N}(0, \sigma^2)$, maximizing likelihood is equivalent to minimizing Mean Square Error (MSE)

$$a_1^{MLE}, a_0^{MLE} = \underset{a_1, a_0}{\text{argmin}} \sum_{i=1}^{n} |y_i - (a_1 x_i + a_0)|^2$$

# Linear Regression in Multiple Variables

A ***multiple linear regression model*** is a linear statistical model relating the response $y$ to multiple predictors $x_1, \ldots, x_J$:

$$y = a_0 + a_1 x_1 + \ldots + x_J x_J + \epsilon.$$

If we take $\epsilon \sim \mathcal{N}(0, \sigma^2)$, then $y|$model is a normally distributed random variable,
$y|$model $\sim \mathcal{N}(a_0 + a_1 x_1 + \ldots + x_J x_J, \sigma^2)$.

The MLE model parameters for observations $\{(x_{1,1}, \ldots, x_{1,J}), \ldots, (x_{n,1}, \ldots, x_{n,J})\}$ are given by

$$a_0^{MLE}, \ldots, a_J^{MLE} = \underset{a_0,\ldots,a_J}{\text{argmin}} \sum_{i=1}^{n} \left| y_i - (a_0 + a_1 x_{i,1} + \ldots + x_J x_{i,J}) \right|^2$$

Fitting a ***multiple linear regression model*** is same as finding the best fitting ***plane*** for the data.

# Polynomial Regression

A ***polynomial regression model of degree*** $m$, relating $y$ to a single predictor $x$, is the statistical model

$$y = a_0 + a_1 x + a_2 x^2 + \ldots + a_m x^m + \epsilon$$

A polynomial model is secretly a multi-linear model with $m$ number of predictors, $\{x, x^2, \ldots, x^m\}$. So, if $\epsilon \sim \mathcal{N}(0, \sigma^2)$, then

$$y|\text{model} \sim \mathcal{N}(a_0 + a_1 x + a_2 x^2 + \ldots + a_m x^m, \sigma^2)$$

Polynomial regression models can also be applied to multiple predictors

$$y = a_0 + (a_{1,1} x_1 + \ldots + a_{m,1} x_1^m) + \ldots + (a_{1,J} x_J + \ldots + a_{m,J} x_J^m) + \epsilon.$$

## Review: Evaluating Models

# MSE and $R^2$

For a set of observations

$$\{(x_{1,1}, \ldots, x_{1,J}, y_1), \ldots, (x_{n,1}, \ldots, x_{n,J}, y_n)\}$$

we typically measure the 'fitness' of a model $\hat{y} = \hat{f}(x_{1,1}, \ldots, x_{1,J})$ we've learned by aggregating the prediction errors

$$MSE = \sum_{i=1}^{n} |y_i - \hat{y}_i|^2 = \sum_{i=1}^{n} \left| y_i - \hat{f}(x_{1,1}, \ldots, x_{1,J}) \right|^2$$

Alternatively, we can also compute the ***explained variance***, the ratio of the variation of the model and the variation in the data. For $\epsilon \sim \mathcal{N}(0, \sigma^2)$, explained variance is

$$0 \leq R^2 = 1 - \frac{\sum_{i=1}^{n} |y_i - \overline{y}_i|^2}{\sum_{i=1}^{n} |\hat{y}_i - \overline{y}_i|^2} \leq 1$$

# Train vs Test Error

By definition, our MLE model minimizes the MSE on the observed data. But the minimized MSE can still be very large (relatively) because the data is not linear or the observed data is skewed!

We need to also measure the fitness of our learned model on new data. To do this, we

1. split the data into a training set and a testing set prior to modeling
2. fit the model using training set; find the MSE and $R^2$, this is called ***training error***
3. use the fitted model to find the MSE and $R^2$ on the testing set, this is called ***testing error***

Which type of error do you expect to be lower? Why?

# Bootstrapping and Standard Error

Typically, our datasets are larger than we can process all at once. We may choose to sample a smaller random set of observations from the data to fit our model.

But this means that our estimates will vary depending on the samples we draw! So just how ‘confident’ can we be in our fitted model?

To gauge the variations of our fitted model depending on the sample data, we ***bootstrap***:

1. we sample multiple sets of observations, for each we fit a model
2. using our multiple fitted models, we compute the mean and variance of each estimated model parameter

The variance of an estimated parameter is called its ***standard error***.

# Significance of Variables

Suppose your fitted linear model looks like:

$$y = 1 + 2.1x_1 + 10x_2 + 0.001x_3$$

Which variable is more ‘important’? Why? Is your definition of ‘important’ fair?

We can also gauge the ***statistical significance*** of each estimated coefficient using ***p-values*** (can be estimated using bootstrap). Smaller p-values indicate stronger evidence against that the variable is completely negligible.

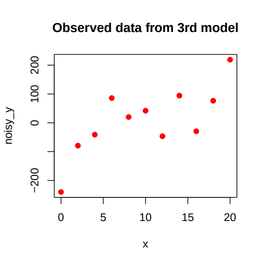
Analyzing the significance our estimated parameters after fitting the model is a necessary and standard step.

# Regularization and Bayesian Models

# Overfitting

One thing we notice immediately was that the error on the training set decreases as the degree of the polynomial increases.

So maybe we generally want to pick very high degree polynomials to model our data?

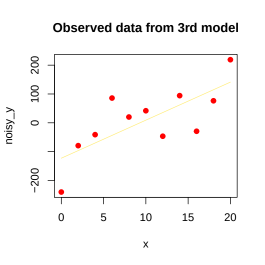


What is happening to our model as the degree increases?

# Overfitting

One thing we notice immediately was that the error on the training set decreases as the degree of the polynomial increases.

So maybe we generally want to pick very high degree polynomials to model our data?

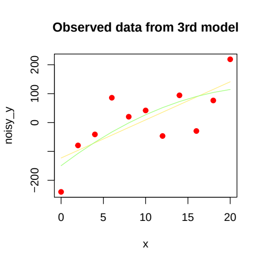


What is happening to our model as the degree increases?

# Overfitting

One thing we notice immediately was that the error on the training set decreases as the degree of the polynomial increases.

So maybe we generally want to pick very high degree polynomials to model our data?

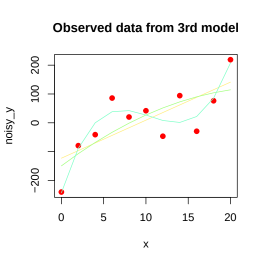


What is happening to our model as the degree increases?

# Overfitting

One thing we notice immediately was that the error on the training set decreases as the degree of the polynomial increases.

So maybe we generally want to pick very high degree polynomials to model our data?

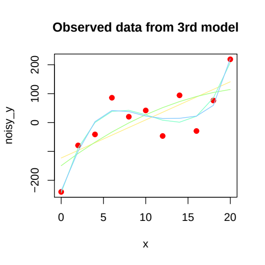


What is happening to our model as the degree increases?

# Overfitting

One thing we notice immediately was that the error on the training set decreases as the degree of the polynomial increases.

So maybe we generally want to pick very high degree polynomials to model our data?

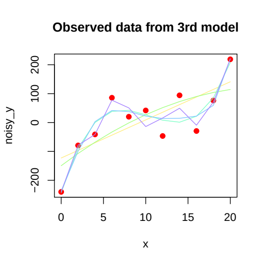


What is happening to our model as the degree increases?

# Overfitting

One thing we notice immediately was that the error on the training set decreases as the degree of the polynomial increases.

So maybe we generally want to pick very high degree polynomials to model our data?

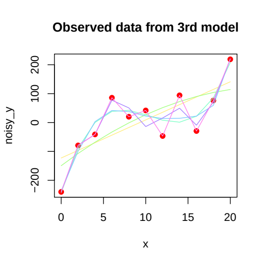


What is happening to our model as the degree increases?

# Overfitting

One thing we notice immediately was that the error on the training set decreases as the degree of the polynomial increases.

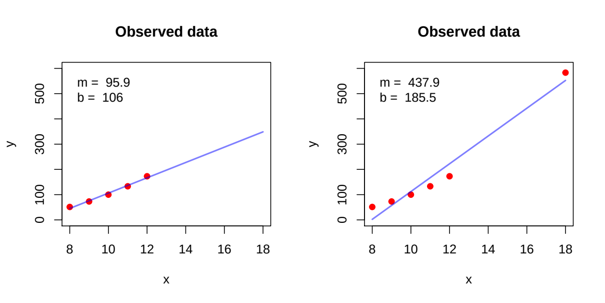
So maybe we generally want to pick very high degree polynomials to model our data?



What is happening to our model as the degree increases?

# Overfitting

Overfitting can happen with linear regression too!



In multiple linear regression, what happens when we have $N$ number of observations and $N$ number of explanatory variables?

# Overfitting

Overfitting happens when we learn parameters or rules that are too specific to the training set, so much that our model is not useful in explaining new data (we do great on train data but poorly on test).

Overfitting can happen when we have too few observations compared to the number of variables in our model with which we try to explain the observations.

We’ll see that overfitting can be curbed by regularization and variable selection.

# Encoding Our Implicit Beliefs/Biases

Before even seeing the California housing data, we already have a set of **beliefs** about the possible values for the parameters in our model,

$$P = mA + b + \epsilon. \tag{1}$$

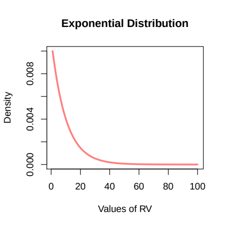
1. $m$ can't be negative
2. $b$ can be negative, but will probably be positive
3. values of $m$ and $b$ can't be too large (to prevent overfitting)

If we want the parameters we learn to reflect these beliefs, we must ***incorporate them into our model***.

# Encoding Our Implicit Beliefs/Biases

We can translated these beliefs about the values of $m$ and $b$ into math

1. $m$ can't be negative and can't be too large



$$m \sim \text{Exp}(0.1), \quad p(m) = 0.1e^{-0.1m}$$

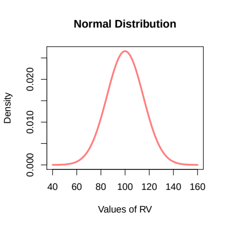
# Encoding Our Implicit Beliefs/Biases

We can translated these beliefs about the values of $m$ and $b$ into math

1. $m$ can't be negative and can't be too large

$$m \sim \text{Exp}(0.1), \quad p(m) = 0.1e^{-0.1m}$$

2. $b$ will probably be positive and can't be too large



$$b \sim \mathcal{N}(100, 15), \quad p(b) = \frac{1}{\sqrt{2\pi \cdot 15^2}} e^{-\frac{1}{2}\left(\frac{b-100}{15}\right)^2}$$

# Encoding Our Implicit Beliefs/Biases

We can translated these beliefs about the values of $m$ and $b$ into math

1. $m$ can't be negative and can't be too large

$$m \sim \text{Exp}(0.1), \quad p(m) = 0.1e^{-0.1m}$$

2. $b$ will probably be positive and can't be too large

$$b \sim \mathcal{N}(100, 15), \quad p(b) = \frac{1}{\sqrt{2\pi \cdot 15^2}} e^{-\frac{1}{2}\left(\frac{b-100}{15}\right)^2}$$

The distributions encoding our ***prior beliefs*** about the parameters $m$ and $b$ are called **priors**.

# Bayesian Statistical Models

A **Bayesian statistical model** is a statistical model that incorporates ***priors*** beliefs about the parameters of the model.

## A Bayesian Price Model

Recall our linear model for housing prices:

$$P = mA + b + \epsilon$$

$$\begin{cases} \epsilon \sim \mathcal{N}(\mu_1, \sigma_1) \end{cases} \qquad \text{(Distribution of Noise)}$$

$$\begin{cases} b \sim \mathcal{N}(\mu_2, \sigma_2) \\ m \sim \text{Exp}(\lambda) \end{cases} \qquad \text{(Priors on Parameters)}$$

where we picked $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 100$, $\sigma_2 = 15$, $\lambda = 0.1$.

# The Posterior Distribution

Learning values for the parameters that takes into account the priors as well as the data is called ***Bayesian inference***.

But just how do we learn while accounting for priors?

# The Posterior Distribution

Consider the distribution of the model parameters ***given*** the data and the model

$$p(m, b|P, A, \epsilon); \tag{1}$$

this is called the **posterior distribution**.

Using Baye's Theorem, we computed the posterior to be

$$\begin{aligned} p(m, b|P, A, \epsilon) &= \frac{p(P|m, A, b, \epsilon)p(m)p(b)}{p(P)} \\ &\propto \underbrace{p(P|m, A, b, \epsilon)}_{Likelihood} \underbrace{p(m)p(b)}_{Priors} \end{aligned}$$

# Bayesian Inference

Working with the ***posterior distribution*** accounts for prior beliefs.

1. (Non-Bayesian) Learn parameters, $m$ and $b$, to maximize the likelihood, i.e. the 'probability' of ***data given the parameters***,

   $$m^{MLE}, b^{MLE} = \underset{m,b}{\text{argmax}}\; p(P|m, A, b, \epsilon)$$

   $m^{MLE}, b^{MLE}$ are called ***maximum likelihood estimates***.

2. (Bayesian) Learn parameters, $m$ and $b$, to maximize the posterior, i.e. the 'probability' of ***parameters given the data and the priors***,

   $$m^{MAP}, b^{MAP} = \underset{m,b}{\text{argmax}}\; p(m, b|P, A, \epsilon) \propto \underbrace{p(P|m, A, b, \epsilon)}_{Likelihood}\underbrace{p(m)p(b)}_{Priors}$$

   $m^{MAP}, b^{MAP}$ are called **maximum a posteriori estimates**.

# Priors and Regularization

Again, maximizing the posterior distribution can be written as minimizing a loss function that penalizes certain kinds of model parameters.

1. ($\ell_1$ regularization) If the parameters have ***Laplace priors*** then

$$m^{MAP}, b^{MAP} = \underset{a_1, a_0}{\text{argmin}} \sum_{i=1}^{n} |y_i - (a_1 x_i + a_0)|^2 + (|a_1| + |a_0|)$$

2. ($\ell_2$ regularization) If the parameters have ***normal priors*** then

$$m^{MAP}, b^{MAP} = \underset{a_1, a_0}{\text{argmin}} \sum_{i=1}^{n} |y_i - (a_1 x_i + a_0)|^2 + \sqrt{a_1^2 + a_0^2}$$

# Priors and Regularization

1. ($\ell_1$ regularization) With $\ell_1$ regularization,

$$m^{MAP}, b^{MAP} = \underset{a_1, a_0}{\text{argmin}} \sum_{i=1}^{n} |y_i - (a_1 x_i + a_0)|^2 + (|a_1| + |a_0|)$$

   we are biased towards models with fewer predictors (fewer non-zero parameters).

2. ($\ell_2$ regularization) With $\ell_2$ regularization,

$$m^{MAP}, b^{MAP} = \underset{a_1, a_0}{\text{argmin}} \sum_{i=1}^{n} |y_i - (a_1 x_i + a_0)|^2 + \sqrt{a_1^2 + a_0^2}$$

   we are biased towards models with smaller parameters.

## Application: Land Use Regression

# Overview of Air Quality Models

***Air quality models*** characterizes the distribution of concentrations of atmospheric pollutants over time and or space. There are two major categories of air quality models:

1. **Deterministic:** these are typically diffusion models that model the movement of atmospheric particles systems, incorporating physical forces and chemical processes
2. **Statistical:** these are models whose corrlate the concentration level of the pollutant with a set of spatio-temporal predicators, incorporating random noise.

***Land Use Regression model*** is an example of a statistical air quality model.

# Land Use Regression Models: Spatial

**Motivation:** Typically monitoring stations are few in number and statically located. There is not enough observed data to understand fine-grained variation of pollution across a region.

Given a set of average readings at $n$ number of sites, $\{p_1, \ldots, p_n\}$, we fit a linear model correlating the land use characteristic, $\{L_1, \ldots, L_J\}$, around each site with the reading at the site:

$$p = a_0 + \sum_{i=1}^{J} a_i L_i \qquad (\textbf{Land Use Regression Model})$$

Typically, we grid the region of interest and extract the same set of land use characteristics $\{L_1, \ldots, L_J\}$ for each grid cell.

# Land Use Regression Models: Spatial

**Question:** How to we use the model

$$p = a_0 + \sum_{i=1}^{J} a_i L_i \qquad (\textbf{Land Use Regression Model})$$

fitted on the readings from the stations, $\{p_1, \ldots, p_n\}$, to fill in the missing observations in other areas?