Regularization

ST552 Lecture 24

Charlotte Wickham

2019-03-06

What to expect on the final?

- Study guide posted along with ~~last year's~~ a final.

heavy on conceptual

Three questions worth roughly equal amounts

1. **Calculations, inference and interpretation.** Like Q2 and Q3 on midterm: t-tests, F-tests, confidence intervals, prediction intervals etc. Know your formulas, how to use them and how to interpret the results.
2. **Assumptions and diagnostics.** Examine plots, identify problems, discuss the consequences of the problem, suggest remedies, suggest ways to verify if the suggestions worked. Or suggest ways to diagnose certain problems.
3. Everything else, at a conceptual level.

Lab in Week 10

Some options:

- Revisit a topic (or more depth on a topic). Which topic?
- Exam review session → Concrete task
- No lab

"think about what to do/cover"

#1

#2

#3

"Previous final question that isn't on the practice"

Today

- The Bias-Variance tradeoff
- Regularized regression: lasso and ridge

Bias-Variance tradeoff

ESTIMATION

For estimates, the mean squared error of an estimate can be broken down into bias and variance terms:

$$MSE(\hat{\theta}) = \mathrm{E}\left((\hat{\theta}-\theta)^2\right) = \mathrm{E}\left(\left(\hat{\theta}-\mathrm{E}\left(\hat{\theta}\right)\right)^2\right) + \left(\mathrm{E}\left(\hat{\theta}\right)-\theta\right)^2$$

$$= \mathrm{Var}\left(\underline{\hat{\theta}}\right) + \mathrm{Bias}\left(\underline{\hat{\theta}}\right)^2$$

Often in statistics, we focus on estimates that are unbiased (so the second term is zero), and focus on minimising the variance term.

You might argue that you are willing to introduce a little bias, if it reduces the variance enough to reduce the overall mean squared error in the estimate.

There is a similar breakdown for the mean square error in prediction. Let $\boxed{\hat{f}(X)}$ indicate the regression model for predicting an observation with explanantory values X.

$\downarrow \hat{\theta}$ $\qquad \to \theta$

If the true data is generated according to $\boxed{Y = f(X) + \epsilon,}$ where $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$, then the MSE for a point x_0:

$$MSE(\hat{f}(X)|X = x_0) =$$

$$\mathrm{E}\left(\boxed{(Y - \hat{f}(X))^2}|\underline{X = x_0}\right)$$

$$= \mathrm{E}\left(\left(\hat{f}(x_0) - \mathrm{E}\left(\hat{f}(x_0)\right)\right)^2\right) + \mathrm{E}\left(\left(Y - \mathrm{E}\left(\hat{f}(x_0)\right)\right)\right)^2$$

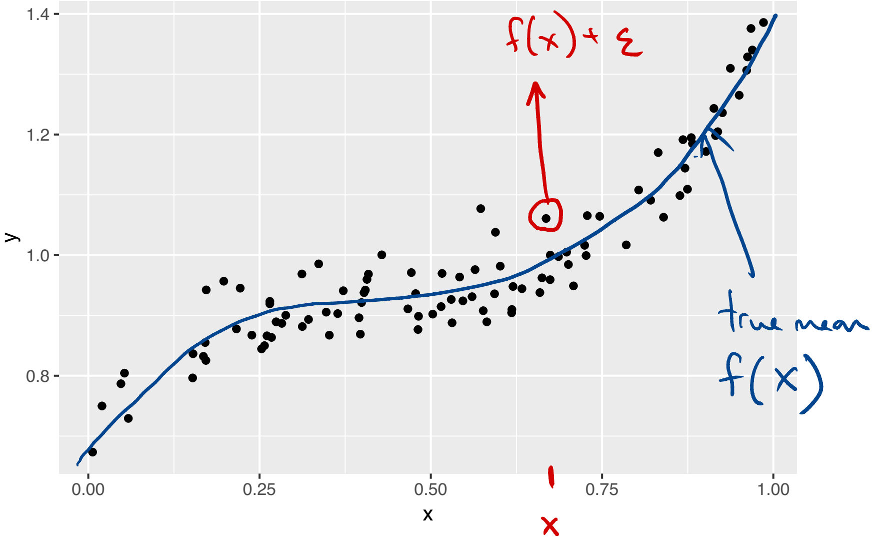
$$= \underline{\mathrm{Var}\left(\hat{f}(x_0)\right)} + \underline{\mathrm{Bias}\left(\hat{f}(x_0)^{\cancel{2}}\right)^2} + \underline{\underline{\sigma^2}}$$

how close is our estimated function to true function on average

Bias captures how far our predictions are from the true mean on average (over repeated samples).

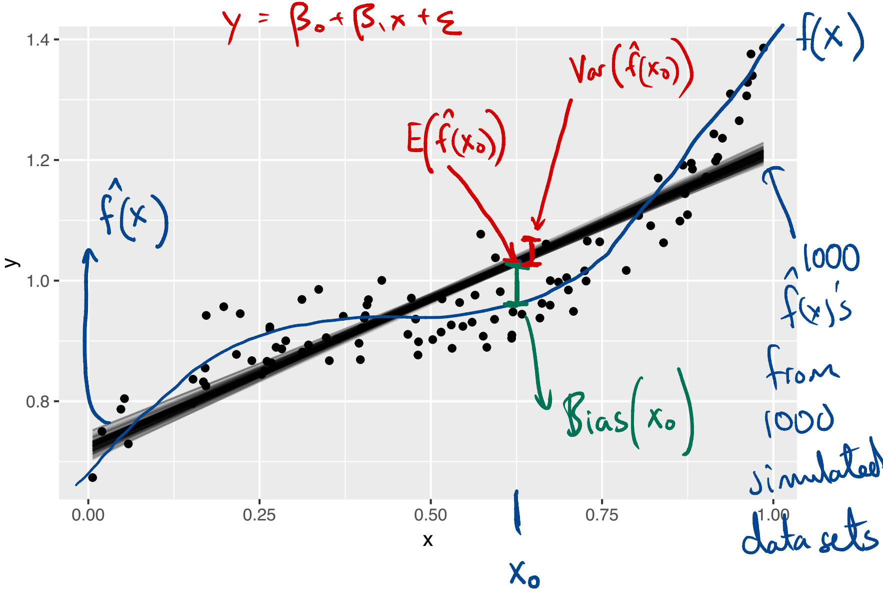
Variance captures how much our predictions vary (over repeated samples).

Simulated Example



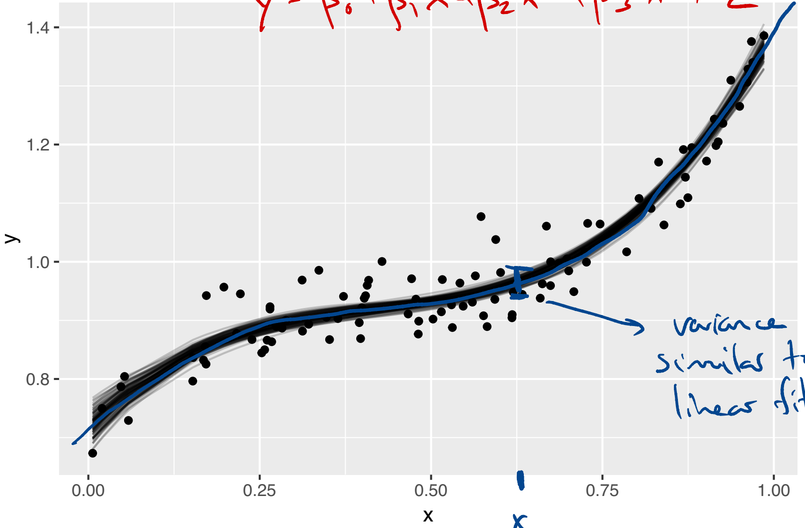
High bias, low variance → linear fit

$y = \beta_0 + \beta_1 x + \varepsilon$



Low bias, low variance → cubic

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

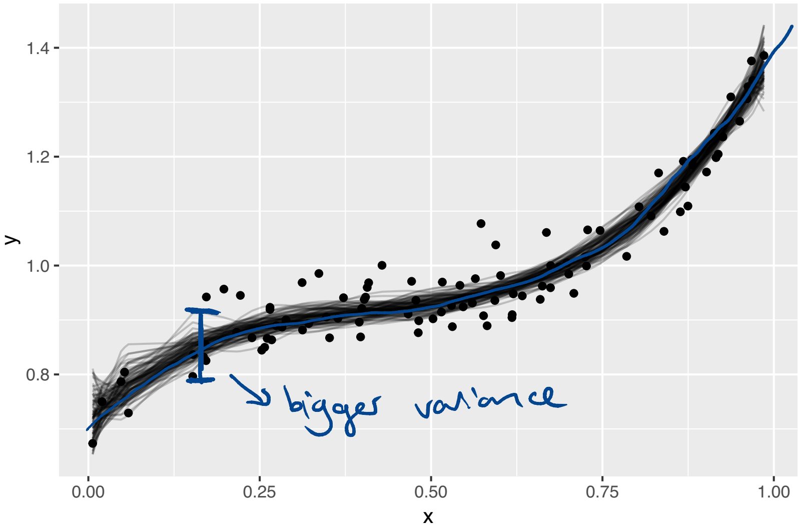


$f(x)$

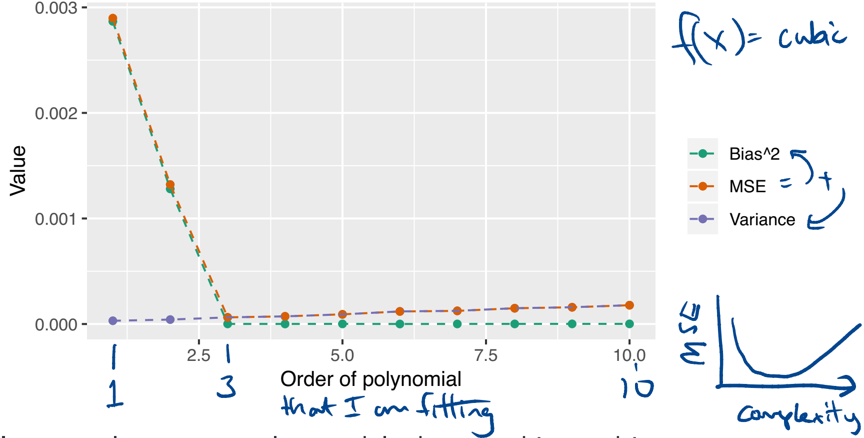
variance similar to linear fits

x_0

Low bias, high variance → 10th degree polynomials



Bias Variance tradeoff



In general, more complex models decrease bias and increase variance. We hunt for the sweet spot where MSE is minimized.

There aren't nice partitions into bias and variance for other metrics, but the pattern is usually the same. Increasing complexity only improves performance to a point, then it decreases performance.

Regularized regression

One approach that introduces bias into the coefficient estimates, is regularized (a.k.a. penalized) regression. Instead of minimising

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

"tuning" parameter

minimise

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \boxed{\boxed{\lambda} \sum_{j=1}^{p} f(\beta_j)}$$

penalty

big coefficients => big penalty

When $f(\beta_j) = \beta_j^2$, the method is called **ridge regression**, and when $f(\beta_j) = |\beta_j|$ the method is called **lasso**.

The general idea: the first term rewards good fit to the data, the second penalizses for large values on the coefficients.

The result: estimates shrink toward zero (introducing bias) and have smaller variance.

Alternative view

You can also view ridge and lasso as constrained minimisation, where we minimise

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

subject to the constraints

$$\sum_{j=1}^{p} \beta_j^2 \leq t \quad \text{for ridge}$$

$$\sum_{j=1}^{p} |\beta_j| \leq s \quad \text{for lasso}$$

'tuning parameters

↓

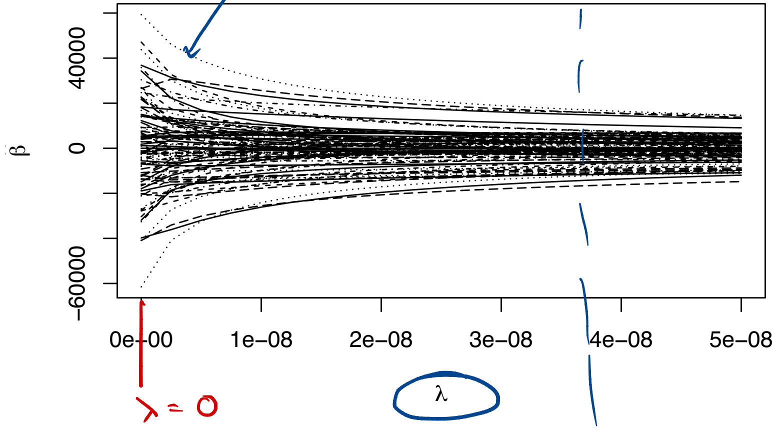
instead of

λ

Ridge estimates

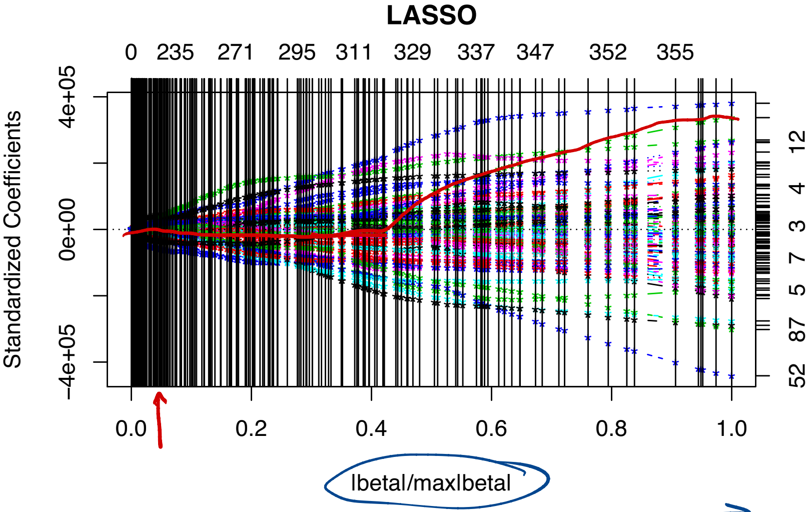
a particular $\hat{\beta}_j$

optimal λ
smallest MSE



$\lambda = 0$
least squares
regression

Lasso estimates



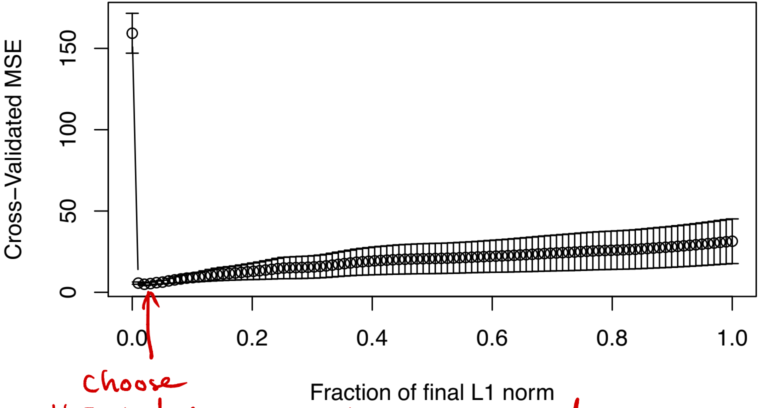
less complex

least squares

Finding the tuning parameter

```
cvout <- cv.lars(as.matrix(trainmeat[ , -101]), trainmeat$fat)
```

Lasso



Choose this value of tuning parameter

tuning parameter

Tuning parameter

```
(best_s <- cvout$index[which.min(cvout$cv)])
```

```
## [1] 0.02020202
```

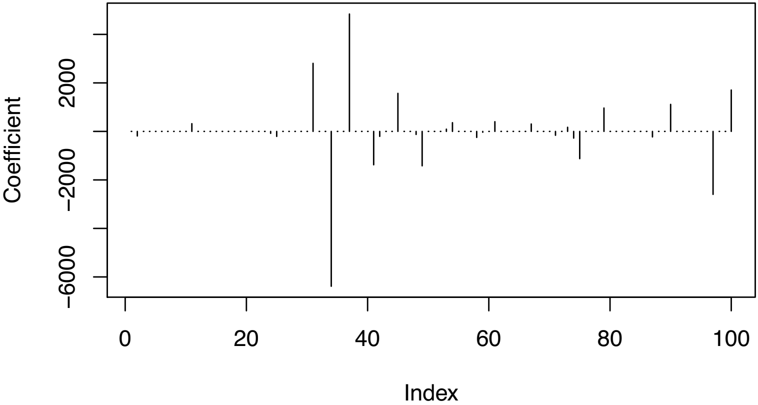
RMSE for this value:

```
testx <- as.matrix(testmeat[,-101])

predlars <- predict(fit_lasso, testx, s=best_s,
  mode="fraction")
sqrt(mean((testmeat$fat - predlars$fit)^2))
```

```
## [1] 2.062124
```

Estimated coefficicents LASSO



```
## [1] 27
```

~~#~~ # $\hat{\beta}_j$'s are not zero

Key points

- Regularized/Penalized regression models have a tuning parameter that controls the degree of penalization/shrinkage
- The tuning parameter may be chosen to optimize some kind of criterion
- Lasso estimates can be exactly zero (so it performs model selection as well)