Regression 2: More perspectives, shortcomings

Ryan Tibshirani
Data Mining: 36-462/36-662

March 5 2013

Optional reading: ISL 3.2.3, 2.2.2; ESL 3.2, 7.3

Reminder: explicit formula for regression coefficients

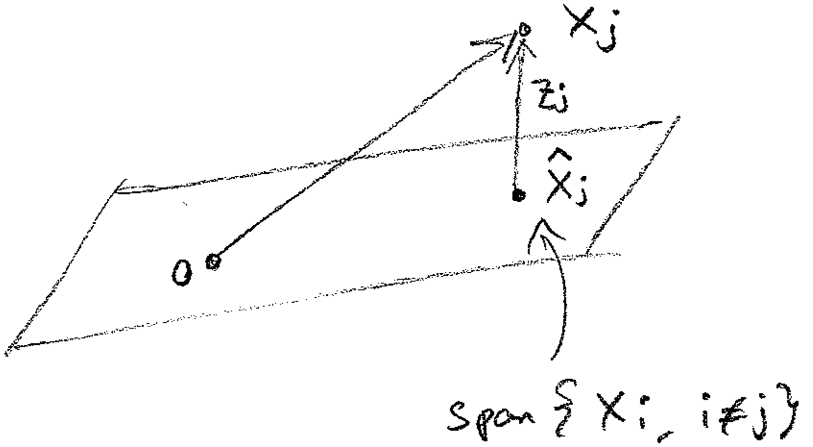
Last time we proved that for the multiple regression of $y \in \mathbb{R}^n$ on predictors $X_1, \ldots X_p \in \mathbb{R}^n$, the jth coefficient can be written as

$$\hat{\beta}_j = \frac{\langle z_j, y \rangle}{\|z_j\|_2^2}$$

This is the coefficient of the univariate regression of y on z_j

Here $z_j \in \mathbb{R}^n$ is the residual from the regression of X_j on all other predictors X_j, $i \neq j$. This is called orthogonalizing X_j with respect to the other predictors, because $\langle z_j, X_i \rangle = 0$ for all $i \neq j$ (Why?)

$$< X_j - \hat{X}_j , X_i > = 0.$$



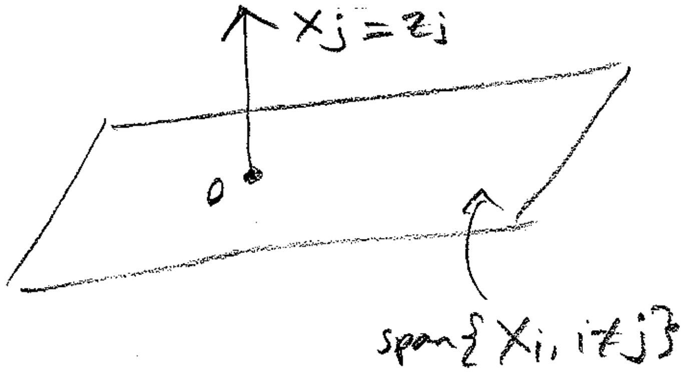
You can think of this as removing the components of X_i, $i \neq j$ from the jth predictor X_j

Orthogonal predictor variables

If $X_1, \ldots X_p$ are orthogonal, then we claimed last time that the jth multiple regression coefficient of y on $X_1, \ldots X_p$ is equal to the univariate regression coefficient of y on X_j. We can easily verify this fact with our new formula

Note that z_j is the residual from regressing X_j onto X_i, $i \neq j$. Remember that the regression fit of X_j onto X_i, $i \neq j$ is really just the projection of X_j onto the linear subspace $\mathrm{span}\{X_i : i \neq j\}$

If X_j is orthogonal to the rest, then this fit is exactly 0, so the residual z_j is simply $z_j = X_j - 0 = X_j$

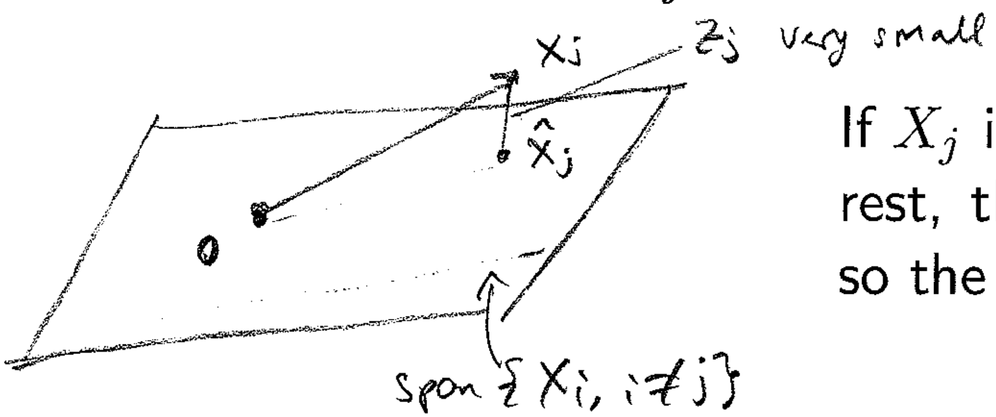


Therefore $\hat{\beta}_j = \dfrac{\langle z_j, y\rangle}{\|z_j\|_2^2} = \dfrac{\langle X_j, y\rangle}{\|X_j\|_2^2}$ is just the univariate regression coefficient of y on X_j

Correlated predictor variables

If $X_1, \ldots X_p$ are correlated, then this new formula gives some insight into what happens to the multiple regression coefficients

Note that z_j is the residual from regressing X_j onto X_i, $i \neq j$. Remember that the regression fit of X_j onto X_i, $i \neq j$ is really just the projection of X_j onto the linear subspace $\mathrm{span}\{X_i : i \neq j\}$



If X_j is highly correlated with the rest, then this fit is close to X_j, so the residual z_j is close to 0

This makes the regression coefficient $\hat{\beta}_j = \dfrac{\langle z_j, y \rangle}{\|z_j\|_2^2}$ unstable, as the denominator is very small, but the numerator can be too

$z_j^T y$ inner product

Variance inflation

From this formula we can explicitly compute the variance of the jth multiple regression coefficient:

$\text{Var}(\langle z_j, y\rangle) = z_j^T \text{Var}(y) z_j$

$\hat{\beta}_j = \frac{\langle z_j, y\rangle}{\|z_j\|_2^2}$

$$\underline{\text{Var}(\hat{\beta}_j)} = \frac{\text{Var}(\langle z_j, y\rangle)}{\|z_j\|_2^4} = \frac{\cancel{\|z_j\|_2^2}\sigma^2}{\|z_j\|_2^{\cancel{4}}} = \boxed{\frac{\sigma^2}{\|z_j\|_2^2}}$$

We can see that having correlated predictors inflates the variance of multiple regression coefficients. Remember that the Z-statistic for the jth regression coefficient is

$$Z_j = \frac{\hat{\beta}_j}{\sqrt{\text{Var}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j}{\sigma} \cdot \|z_j\|_2$$

so if X_j is highly correlated with the other predictors, its regression coefficient will likely be not significant (according to Z_j)

Dropping predictor variables

Now suppose that X_j and X_k both contribute in explaining y, but are highly correlated with each other. Then from what we said on the last slide, neither $|Z_j|$ nor $|Z_k|$ will be very large, so they won't be significant

Now what happens if we remove one of them—say, X_k—from the model, and recompute the regression coefficients? The term $\|z_j\|_2^2$ will be much larger (assuming that X_j is not highly correlated with other predictors than X_k). Hence it's variance will decrease, and Z_j will likely increase

This is why we can't remove two (or more) supposedly insignificant predictors at a time—in short: because significance depends on what other predictors are in the model!

Shortcomings of regression

Two main themes:

1. Predictive ability: the linear regression fit often does not predict well, especially when p (the number of predictors) is large

   (Important to note that is not even necessarily due to nonlinearity in the data! Can still predict poorly even when a linear model could fit well)

2. Interpretative ability: linear regression "freely" assigns a coefficient to each predictor variable. When p is large, we may sometimes seek, for the sake of interpretation, a smaller set of important variables

   Hence we want to "encourage" our fitting procedure to make only a subset of the coefficients large, and others small or even better, zero

Prediction accuracy and mean-squared error

Suppose we observe data of the form

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots n$$

Here $f : \mathbb{R}^p \to \mathbb{R}$ is some true function, $x_i = (x_{i1}, \ldots x_{ip}) \in \mathbb{R}^p$ are fixed predictor measurements, and $\epsilon_i \in \mathbb{R}^n$ are random errors with $\mathrm{E}[\epsilon_i] = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$, and $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$

Consider one more point of the form

$$y_0 = f(x_0) + \epsilon_0$$

independent $y_1, \cdots y_n$

and suppose that we want to predict y_0 at the fixed point $x_0 \in \mathbb{R}^p$, from the observed pairs $(y_1, x_1), \ldots (y_n, x_n)$

Think of, e.g., the typical linear regression model: here we have $f(x_i) = \underline{x_i^T \beta^*}$, for some true regression coefficients β^*

$$= \beta_1^* x_{i1} + \ldots \beta_p^* x_{ip}$$

Suppose that we use $\hat{f}$ to predict f (again, think of regression: $\hat{f}(x_i) = x_i^T\hat{\beta}$). In particular, we predict y_0 via $\hat{f}(x_0)$

$= x_{i1}\hat{\beta}_1 + \cdots x_{ip}\hat{\beta}_p.$

Question: how does prediction error (PE) relate to to mean squared error (MSE)?

$$
\begin{aligned}
\underline{\mathrm{PE}\big(\hat{f}(x_0)\big)} &= \mathrm{E}\big[\big(y_0 - \hat{f}(x_0)\big)^2\big] \\
&= E[(y_0 - f(x_0) + f(x_0) - \hat{f}(x_0))^2] \\
&= \underline{E[(y_0 - f(x_0))^2]} + \underline{E[(f(x_0) - \hat{f}(x_0))^2]} \\
&\quad + 2E\underline{[(y_0 - f(x_0))(f(x_0) - \hat{f}(x_0))]} \\
&= \sigma^2 + \mathrm{MSE}\big(\hat{f}(x_0)\big)
\end{aligned}
$$

$$\shortparallel$$

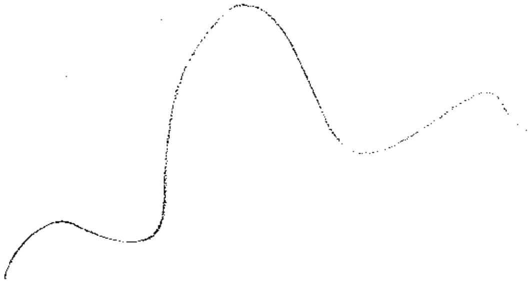
$$2\underbrace{E(y_0 - f(x_0))}_{0} \cdot E(f(x_0) - \hat{f}(x_0))$$

So PE and MSE are essentially the same thing ... in the sense that doing well in terms of one is the same as in terms of the other

Bias and variance

Now let's focus on MSE. Again we used $\hat{f}$ to predict f

Question: what kind of quantities does MSE depend on?



$$
\begin{aligned}
\mathrm{MSE}\big(\hat{f}(x_0)\big) &= \mathrm{E}\Big[\big(f(x_0) - \hat{f}(x_0)\big)^2\Big] \\
&= E\Big[\big(f(x_0) - E[\hat{f}(x_0)] + E[\hat{f}(x_0)] - \hat{f}(x_0)\big)^2\Big] \\
&= \big(E(\hat{f}(x_0)) - f(x_0)\big)^2 + E\Big[\big(\hat{f}(x_0) - E[\hat{f}(x_0)]\big)^2\Big] \\
&= \Big[\mathrm{Bias}\big(\hat{f}(x_0)\big)\Big]^2 + \mathrm{Var}\big(\hat{f}(x_0)\big) + \underbrace{2E\Big[\big(f(x_0) - E[\hat{f}(x_0)]\big)\cdot\big(E(\hat{f}(x_0)) - \hat{f}(x_0)\big)\Big]}_{0}
\end{aligned}
$$

This is called the bias-variance tradeoff (or decomposition)

Think: what kinds of estimators will have high bias? not flexible What kinds will have high variance? flexible

think of $f(x_i) = x_i^T \beta^*$ $\quad E[x_i^T \hat{\beta}] = x_i^T \beta^*$

$\hat{f}(x_i) = x_i^T \hat{\beta}$ $\quad$ Zero bias (unbiased)

Back to the linear model

Here we assume that we have observations of the form

$$y_i = x_i^T \beta^* + \epsilon_i, \quad i = 1, \ldots n$$

i.e., $f(x_i) = x_i^T \beta^*$. Now what about the least squares prediction $\hat{f}^{\mathrm{LS}}(x_i) = x_i^T \hat{\beta}$, where $\hat{\beta}$ are the estimated regression coefficients?

Recall the Gauss-Markov theorem said that this estimator is the BLUE: best linear unbiased estimator. I.e., for a fixed input point x_0, if $\hat{f}(x_0)$ is any other linear, unbiased estimator of $x_0^T \beta^*$, then

$$\mathrm{MSE}\big(\hat{f}(x_0)\big) \geq \mathrm{MSE}\big(\hat{f}^{\mathrm{LS}}(x_0)\big) = \mathrm{MSE}(x_0^T \hat{\beta})$$

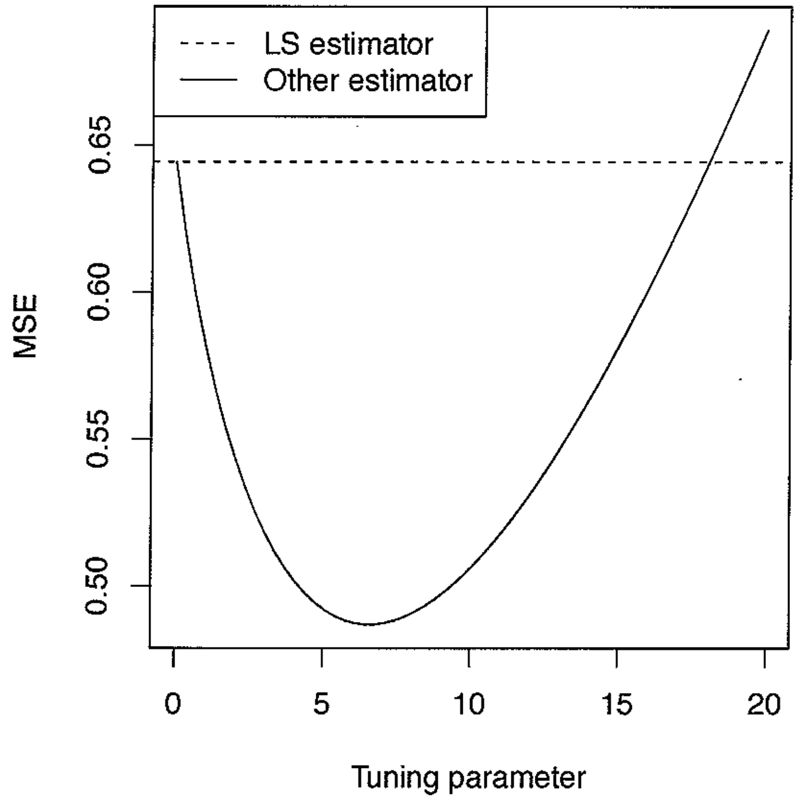
- Unbiased: this means that $\underline{\mathrm{E}[\hat{f}(x_0)] = x_0^T \beta^*}$
- Linear: this means linear in $y = (y_1, \ldots y_n)$, i.e., $\underline{\hat{f}(x_0) = c^T y}$ for some c

Therefore, for any unbiased $\hat{f}(x_0)$

$$\begin{aligned}\mathrm{MSE}\big(\hat{f}(x_0)\big) &= \big[\mathrm{Bias}\big(\hat{f}(x_0)\big)\big]^2 + \mathrm{Var}\big(\hat{f}(x_0)\big)\\ &= 0 + \mathrm{Var}\big(\hat{f}(x_0)\big)\end{aligned}$$

The Gauss-Markov theorem says that among unbiased and linear estimates, $\hat{f}^{\mathrm{LS}}$ has the smallest MSE, i.e., the smallest variance

But wait ... I claim to know another linear estimator of $x_0^T \beta^*$ that has a smaller MSE then the least squares estimate! How can this be?



Averaging over all inputs

It is helpful to look at the average PE or MSE across all the input points $x_1, \ldots x_n$ can look at one fixed input x_0

$$\mathrm{PE}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{PE}\big(\hat{f}(x_i)\big), \quad \mathrm{MSE}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{MSE}\big(\hat{f}(x_i)\big)$$

Note the same relationships hold:

$$\begin{aligned}\mathrm{PE}(\hat{f}) &= \sigma^2 + \mathrm{MSE}(\hat{f}) \\ &= \sigma^2 + \frac{1}{n}\sum_{i=1}^{n} \big[\mathrm{Bias}\big(\hat{f}(x_i)\big)\big]^2 + \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}\big(\hat{f}(x_i)\big)\end{aligned}$$

What does this look like for regression, $\hat{f}^{\mathrm{LS}}(x_i) = x_i^T\hat{\beta}$?

$$\begin{aligned}\mathrm{PE}(\hat{f}^{\mathrm{LS}}) &= \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}\left[\mathrm{Bias}(x_i^T\hat{\beta})\right]^2 + \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(x_i^T\hat{\beta}) \\ &= \sigma^2 + 0 + \frac{p\sigma^2}{n} \quad \checkmark\end{aligned}$$

Why this last expression for the variance?

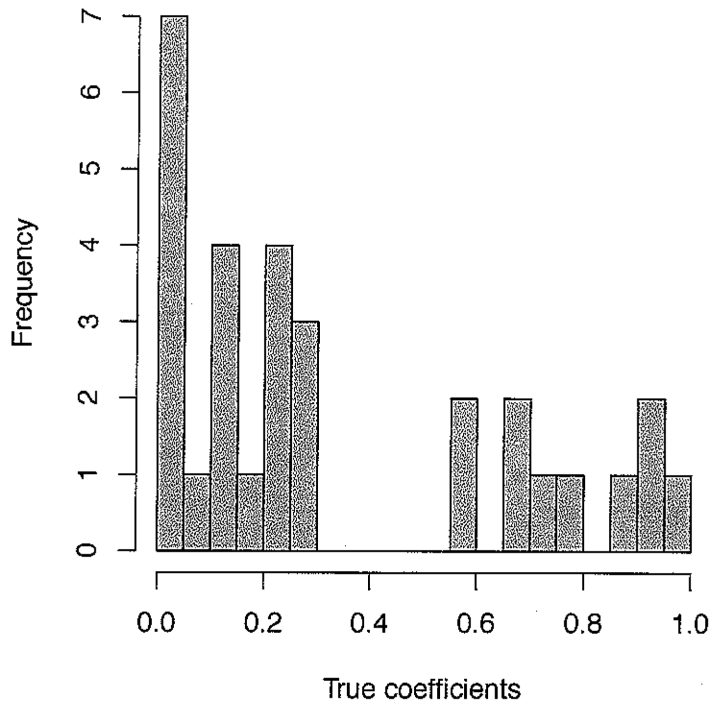
$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(x_i^T\hat{\beta}) &= \frac{1}{n}\mathrm{trace}\big(\overbrace{\mathrm{Var}(X\hat{\beta})}^{n\times n \text{ matrix}}\big) \\ &= \frac{1}{n}\mathrm{trace}(\mathrm{Var}(\underbrace{X(X^TX)^{-1}X^T}_{H}y)) \\ &= \frac{1}{n}(\mathrm{trace}(H\sigma^2 I H)) \\ &= \frac{\sigma^2}{n}\mathrm{trace}(H) = \frac{\sigma^2}{n}\mathrm{trace}(X(X^TX)^{-1}X^T) \\ &= \frac{\sigma^2}{n}\mathrm{trace}(\underbrace{X^TX(X^TX)^{-1}}_{I_{p\times p}}) \\ &= \frac{\sigma^2}{n}p\end{aligned}$$

This scales linearly with the number of predictors p

Example: small regression coefficients

Example: simulation with $n = 50$ and $p = 30$. The entries of the predictor matrix $X \in \mathbb{R}^{50\times 30}$ are all i.i.d. $N(0,1)$, so overall the variables have low correlation

Histogram of the true regression coefficients $\beta^* \in \mathbb{R}^{30}$:



Here 10 coefficients are large (between 0.5 and 1) and 20 coefficients are small (between 0 and 0.3)

The response $y \in \mathbb{R}^{50}$ is drawn from the model $y = X\beta^* + \epsilon$, where the entries of $\epsilon \in \mathbb{R}^{50}$ are all i.i.d $N(0,1)$ (hence the noise variance is $\sigma^2 = 1$)

We repeated the following 100 times:

- Generate a response vector y
- Compute the linear regression fit $X\hat{\beta}$
- Generate a new response y'
- Record the error $1/n \sum_{i=1}^n (y_i' - x_i^T \hat{\beta})^2$

We averaged this observed error over the 100 repetitions to get an estimate of the the prediction error

We also estimated the squared bias and variance of the fits $X\hat{\beta}$ over the 100 repetitions. Recall that it should be true that prediction error $= 1 +$ squared bias $+$ variance

Results:

```
> bias
[1] 0.00647163
> var
[1] 0.6273129
> p/n
[1] 0.6
> 1 + bias + var
[1] 1.633785
> prederr
[1] 1.644363
```

30/50

This is a good check for our formulas!

How can we do better?

For linear regression, its prediction error is just $\sigma^2 + p/n \cdot \sigma^2$, the second term being the variance $1/n \sum_{i=1}^n \mathrm{Var}(x_i^T \hat{\beta})$

What can we see from this? Each additional predictor variable will add the same amount of variance σ^2/n, regardless of whether its true coefficient is large or small (or zero)

$$y = X\beta^* + \varepsilon + X^{more} \cdot 0$$

In the previous example, we were "spending" variance in trying to fit truly small coefficients—there were 20 of them, out of 30 total

So can we do better by shrinking small coefficients towards zero, incurring some bias, so as to reduce the variance? You can think of this as trying to ignore some "small details" in order to get a more stable "big picture"

The answer, as we'll see next time, is yes

Recap: more perspectives, shortcomings

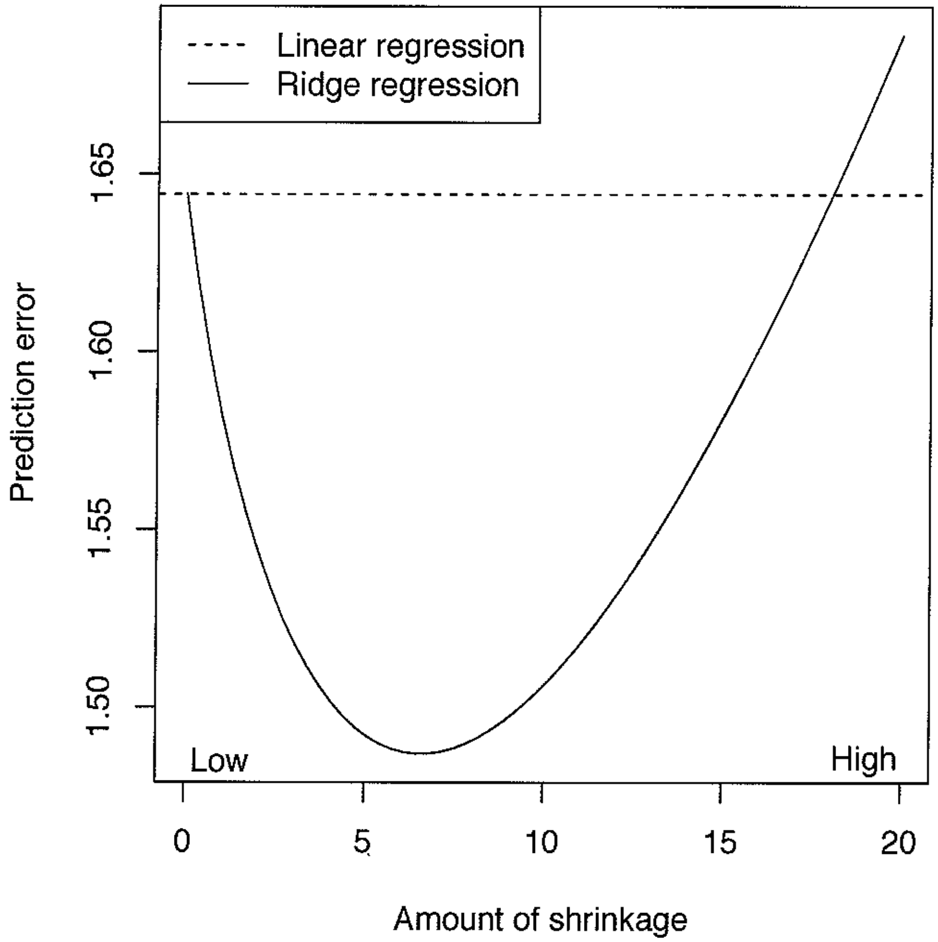
In this lecture, we investigated in a little more detail the explicit formula for multiple regression coefficients. We convinced ourselves that for orthogonal predictors, the multiple regression coefficients are just the univariate ones. For correlated predictors, these can be very different.

The variance of the jth regression coefficient also has an explicit formula in terms of the residual regressing the jth predictor onto all of the others. This shows that the variance is inflated by the presence of correlated variables; hence the significance is degraded

We discussed two shortcomings of linear regression: its predictive ability and its interpretative ability. We looked at the former in more detail in terms of the bias-variance decomposition. We argued that it may help the overall prediction accuracy to decrease the variance at the expense of slightly increasing the bias

Next time: moving into modern regression

Ridge regression can outperform linear regression in terms of prediction error


- - - - Linear regression
—— Ridge regression
Prediction error
Low
High
Amount of shrinkage