

# ST308 - Lent term

## Bayesian Inference

Kostas Kalogeropoulos

Bayesian Linear Regression

# Outline

**Topics covered:** Interpretation of coefficients, linear Basis functions, Least Squares / Maximum Likelihood estimator, Overfit, Ridge and Lasso Regression, Prediction, Model Choice.

- 1 Linear Models for Regression
- 2 Bayesian Linear Regression
- 3 Optional: MLE derivation

# Outline

- 1 Linear Models for Regression
- 2 Bayesian Linear Regression
- 3 Optional: MLE derivation

# Motivating Example: Prostate Cancer

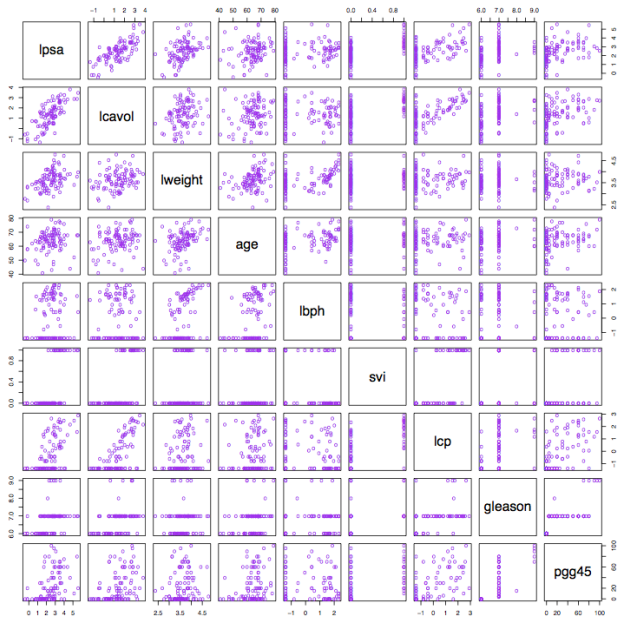
Data from the following study on prostate cancer

*Stamey, T., et al. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II radical prostatectomy treated patients, Journal of Urology 16: 1076-1083.*

Examines association between the level of **prostate-specific antigen** (PSA) and a number of **clinical measures** in men who were about to receive a radical prostatectomy.

The variables are cancer volume (lcavol), prostate weight (lweight), age, amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), capsular penetration (lcp), Gleason score (gleason), and percent of Gleason scores 4 or 5 (pgg45).

# Motivating Example: Data



# Motivating Example: Aims of the analysis

- Determine the level of PSA on a **future** patient based on the clinical measurements. Otherwise detailed histological and morphometric analysis is required.
- How is each of these variables **associated** with PSA? Is the association **linear**?
- Are any of these variables **redundant** in the presence of the others? Which are the **most important**?
- Are there any **synergies** between these variables?

## Data setup

Data consist of measurements on all these variables on several individuals.

We typically denote value of the **response** variable, in this case log-PSA, on the individual  $i$  with  $Y_i$ . The vector  $Y = (Y_1, \dots, Y_n)$  is assumed to be a  $n$ -dimensional random variable.

The remaining variables  $X_1, \dots, X_p$  contain the clinical measurements.  $X_{ji}$  refers to the value of the clinical measurement  $j$  of the individual  $i$ .

The  $X$ 's are **not assumed to be random** they are treated as fixed inputs, with  $Y$  being regarded as the **output**.

# Linear Regression

The linear regression model is written as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \epsilon_i, \quad i = 1, \dots, n.$$

It is more convenient to use matrix algebra. Define  $y = (Y_1, \dots, Y_n)$ ,  $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$  the **error terms**,  $\beta = (\beta_0, \dots, \beta_p)^T$  denoting the **regression coefficients** and the **design matrix**

$$X = \begin{pmatrix} 1 & X_{11} & \cdots & X_{p1} \\ 1 & X_{12} & \cdots & X_{p2} \\ \vdots & \vdots & & \vdots \\ 1 & X_{1n} & \cdots & X_{pn} \end{pmatrix}.$$

Then we can rewrite the model in matrix notation as

$$y = X\beta + \epsilon,$$



# Interpretation of coefficients for continuous inputs

Consider the model  $Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$ . Typically  $E[\epsilon_i] = 0$ , hence

$$E[Y_i | X_{1i} = x] = \beta_0 + \beta_1 x$$

Note that  $E[Y_i | X_{1i} = 0] = \beta_0$ . Hence  $\beta_0$  is the **expected value of the variable  $Y$  for  $X_1 = 0$** .

Consider  $X_{1i} = x$  and  $X_{1i} = x + 1$ . Note that

$$E[Y_i | X_{1i} = x + 1] - E[Y_i | X_{1i} = x] = \beta_0 + \beta_1 x + \beta_1 - \beta_0 - \beta_1 x = \beta_1$$

Hence  $\beta_1$  is the **expected change in  $Y$  for a unit change in  $X$** .

Now consider the model  $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$ . Do the interpretation of  $\beta_0$  and  $\beta_1$  change?

## Example: Prostate Cancer Regression Coefficients

Test your understanding in the table below:

Term	Coefficient	Std. Error	Z Score
Intercept	2.46	0.09	27.60
lcavol	0.68	0.13	5.37
lweight	0.26	0.10	2.75
age	-0.14	0.10	-1.40
lbph	0.21	0.10	2.06
svi	0.31	0.12	2.47
lcp	-0.29	0.15	-1.87
gleason	-0.02	0.15	-0.15
pgg45	0.27	0.15	1.74

# Qualitative inputs

Suppose that we have a **categorical** variable  $G$  as input, taking the values,  $\{A, B, C\}$ .

We can regress the response variable  $Y$  into  $G$  using **dummy variables**. This is done by the following procedure:

- Choosing a reference category -say-  $A$ .
- For the **remaining categories**,  $B$  and  $C$ , create indicator (dummy) variables.
- For example in the case of category  $B$  form the variable  $X_{Bi}$  taking the value 1 if the individual  $i$  is in category  $B$  and 0 otherwise.
- Use the regression model

$$Y_i = \beta_0 + \beta_B X_{Bi} + \beta_C X_{Ci} + \epsilon_i$$

The model above is also known as **ANOVA**.

## Qualitative inputs

In the model above note that  $E[Y_i|X_{Bi} = 0, X_{Ci} = 0]$ . This is equal to  $\beta_0$  that correspond to the **mean of Y in A**.

If we take  $E[Y_i|X_{Bi} = 1, X_{Ci} = 0] - E[Y_i|X_{Bi} = 0, X_{Ci} = 0]$  we will get

$$\beta_0 + \beta_B - \beta_0 = \beta_B$$

Hence  $\beta_B$  is the **mean difference of Y between B and A**.

The parameter  $\beta_C$  is interpreted in a similar manner.

Selecting a different reference category (e.g. C) will give an **equivalent** model with slightly different interpretation.

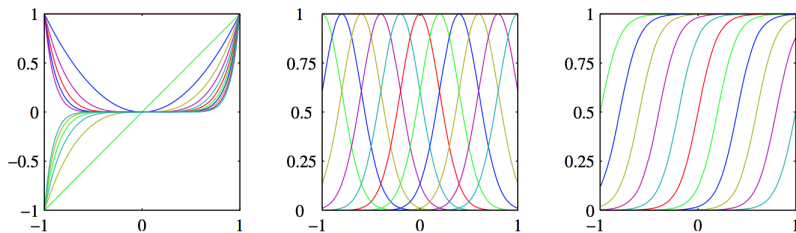
# Linear Basis Functions

The model is linear its parameters  $\beta$  not  $X$  so we can replace each  $X_i$  with a  $h(X_i)$ . We can then write the model as

$$Y = h(X)\beta + \epsilon \text{ or else}$$

$$Y_i = \beta_0 + \beta_1 h_1(X)_i + \cdots + \beta_m h_m(X)_i + \epsilon_i$$

Examples include polynomial terms, Gaussian kernels, sigmoid functions etc.



**Figure 3.1** Examples of basis functions, showing polynomials on the left, Gaussians of the form (3.4) in the centre, and sigmoidal of the form (3.5) on the right.

# MLE of Linear Regression

We assume that the errors  $\epsilon$  are distributed as  $N(0_n, \sigma^2 I_n)$  where  $0_n$  is the vector of  $n$  zeros and  $I_n$  is the identity matrix of dimension  $n$ .

Hence  $y \sim N(X\beta, \sigma^2 I_n)$ . The likelihood function can be written as

$$\begin{aligned} f(y|X, \beta, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta) \right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - X_i\beta)^2 \right\}, \end{aligned}$$

where  $X_i$  denotes the  $i$ -th row of the matrix  $X$ .

Maximising the likelihood wrt  $\beta$  is the same as minimising the blue terms wrt  $\beta$ . Hence MLE is the same as the least squares estimator.

# MLE of Linear Regression

The **MLE** and the **least squares** estimators can be shown to be:

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

The **variance** of the MLE is given by

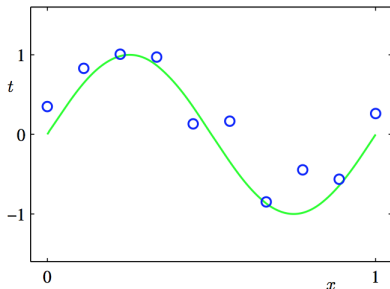
$$\text{Var}(\hat{\beta}) = \hat{\sigma}^2 (X^T X)^{-1},$$

where  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - X_i \hat{\beta})^2$  or  $\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^n (Y_i - X_i \hat{\beta})^2$ .

The distribution of the MLE is the **t-distribution** with  $n - p$  degrees of freedom.

## Example: Polynomial Curve Fitting

Let's consider the following example on simulated data. The generating process is  $\sin(2\pi x)$  and we observe this function on 10 different points in  $[0, 1]$  with independent Gaussian error.



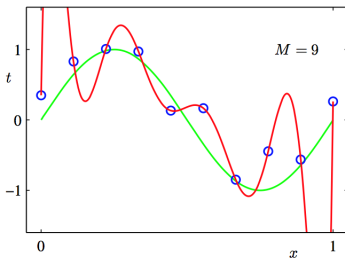
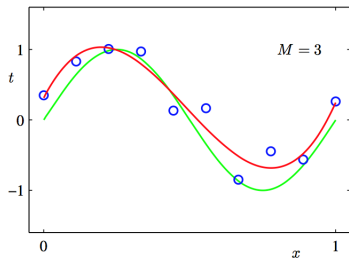
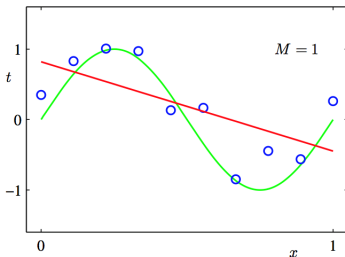
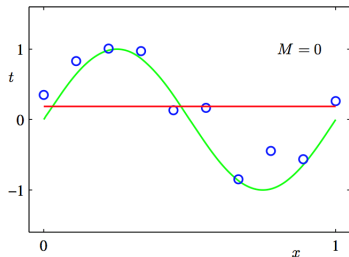
We fit a linear regression model with polynomial basis functions

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_M x_i^M + \epsilon_i$$



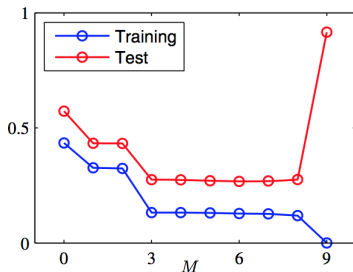
## Example: Fit of different polynomials

For each order of polynomial we find the MLE and plot the corresponding function to assess its fit



# Training and Test Error - Overfitting

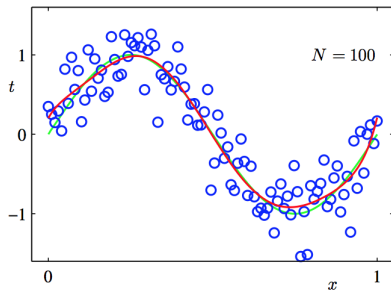
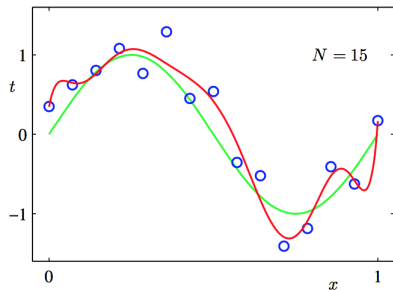
In addition to the training data set of 10 observations we simulate 100 more points in the same way and assess the training and test error.



The test error decreases until  $M = 6$  and increase afterwards. The training error keeps decreasing and drops to 0 for  $M = 9$ . MLE leads to **overfit**.

# Model Complexity

In order to identify the data generating process our model should not be too **complex** compared to the data we are training it. Increasing the data improves learning for the model with  $M = 9$ .



# Parameter Estimates

One way to reduce **model complexity** is to reduce the number of predictors. Not necessarily the best way.

More insight is obtained by looking at the parameter estimates for polynomials of different order.

$M = 0$	$M = 1$	$M = 6$	$M = 9$
0.19	0.82	0.31	0.35
	-1.27	7.99	232.37
		-25.43	-5321.83
		17.37	48568.31
			-231639.30
			640042.26
			-1061800.52
			1042400.18
			-557682.99
			125201.43

# Regularisation

Instead of removing predictors we could instead **restrict** them closer to 0. In the least squares criterion

$$\sum_{i=1}^n (y_i - \beta_0 + \beta_1 x_i + \cdots + \beta_M x_i^M)^2,$$

we add a **penalty term**

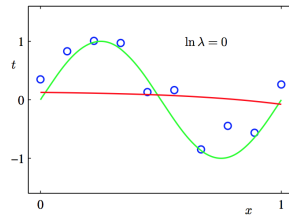
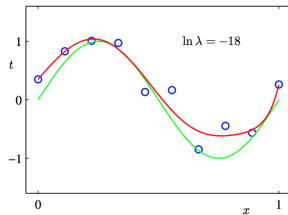
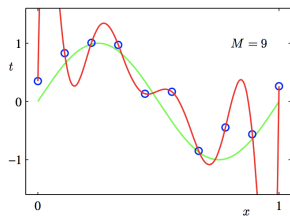
$$\sum_{i=1}^n (y_i - \beta_0 + \beta_1 x_i + \cdots + \beta_M x_i^M)^2 + \lambda \sum_{i=1}^M \beta_i^2.$$

In the general case of a linear regression model  $y = X\beta + \epsilon$  the above is minimised at the point

$$\hat{\beta}^\lambda = (X^T X + \lambda^2 I_p)^{-1} X^T y$$

# Output From Regularised Approach

$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
0.35	0.35	0.13
232.37	4.74	-0.05
-5321.83	-0.77	-0.06
48568.31	-31.97	-0.05
-231639.30	-3.89	-0.03
640042.26	55.28	-0.02
-1061800.52	41.32	-0.01
1042400.18	-45.95	-0.00
-557682.99	-91.53	0.00
125201.43	72.68	0.01



# Outline

- 1 Linear Models for Regression
- 2 Bayesian Linear Regression
- 3 Optional: MLE derivation

# Bayesian Linear Regression

Techniques like this come under the names **shrinkage**, **ridge regression** or **weight decay** in the context of neural networks.

But how can it be justified from a Statistical Inference point of view?

Adjusting for overfit is automatic under the Bayesian framework and comes in a natural way.

It turns out that this is a special case of **Bayesian Linear Regression**.



## Case of known $\sigma^2$

We will first assume that  $\sigma^2$  is **known**.

The **likelihood**  $f(y|\beta)$  is the  $N(X\beta, \sigma^2 I_n)$ , where  $I_n$  is the identity matrix of dimension  $n$ .

The **prior** on  $\beta$  can be set to  $N(\mu_0, \sigma^2 \Omega_0)$

The **posterior** is then the  $N(\mu_n, \sigma^2 \Omega_n)$ , where

$$\begin{aligned}\Omega_n &= (X^T X + \Omega_0^{-1})^{-1}, \\ \mu_n &= (X^T X + \Omega_0^{-1})^{-1}(\Omega_0^{-1} \mu_0 + X^T y),\end{aligned}$$

If we set  $\mu_0 = 0$  and  $\Omega_0 = g^2 I_p$  the **Bayes Estimator** coincides with the **ridge regression estimator** with  $g = 1/\lambda$ .

# Notes on Bayesian Linear Regression

The Bayes estimate is a **weighted average** between the prior mean and the MLE.

The prior  $N(\mu_0, \sigma^2 \Omega_0)$  **shrinks** the parameters to  $\mu_0$ . This can be interpreted as **prior information**.

The amount of force is determined by  $\Omega_0$ . Prior can be viewed as a **tuning parameter** in the Machine Learning context.

The Bayes/ridge regression estimator is **biased** but has smaller variance due to the **shrinking effect**.

# Bayesian Lasso

If we use the **Laplace** prior  $\text{La}(0, 1/\gamma)$  for  $\beta$  the Bayes estimator corresponding to the **posterior mode** is the Lasso Regression estimator.

The Laplace prior can also be written as a hierarchical Normal-Exponential prior:

$$\beta_i \sim N\left(0, \sigma^2 \tau_i^2\right), \quad \tau_i^2 \sim \text{Exponential}\left(\frac{\gamma^2}{2}\right),$$

where  $\lambda = \gamma/\sigma$ .

Note however that the posterior mean and median provide **different Bayes estimators**.

## Example: Results

Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
Intercept	2.465	2.477	2.452	2.468	2.497	2.452
lcavol	0.680	0.740	0.420	0.533	0.543	0.419
lweight	0.263	0.316	0.238	0.169	0.289	0.344
age	-0.141		-0.046		-0.152	-0.026
lbph	0.210		0.162	0.002	0.214	0.220
svi	0.305		0.227	0.094	0.315	0.243
lcp	-0.288		0.000		-0.051	0.079
gleason	-0.021		0.040		0.232	0.011
pgg45	0.267		0.133		-0.056	0.084
Test Error	0.521	0.492	0.492	0.479	0.449	0.528
Std Error	0.179	0.143	0.165	0.164	0.105	0.152

# Unit information prior for Linear Regression

In the linear regression model, the **Fisher information** for  $\beta$  based on  $n$  observations is

$$I(\beta) = -E \left[ -\frac{1}{\sigma^2} X^T X \right] = \frac{X^T X}{\sigma^2}$$

**Unit information** takes the average over  $n$  observations so the variance is set to  $n\sigma^2(X^T X)^{-1}$ . This implies setting  $\Omega_0 = n(X^T X)^{-1}$ , so  $(X^T X)^{-1}$  instead of  $I_p$  and with  $g = n$ .

It is also possible to let  $g$  be an unknown parameter and estimated by the data. This is known as the **Zelner's  $g$  prior**.

## Bayesian Linear Regression model with unknown $\sigma^2$

Assume  $y$  is  $N(X\beta, \sigma^2 I_n)$ . Assign  $N(\mu_0, \sigma^2 \Omega_0)$  as **prior** for  $\beta$  (given  $\sigma^2$ ) and the  $\text{IGamma}(\alpha_0, \beta_0)$ .

The **posterior** for  $\pi(\beta, \sigma^2 | y, X)$  is then the product of the  $\text{IGamma}(\alpha_n, \beta_n)$  and the  $N(\mu_n, \sigma^2 \Omega_n^2)$  where

$$\mu_n = (X^T X + \Omega_0^{-1})^{-1} (\Omega_0^{-1} \mu_0 + X^T y)$$

$$\Omega_n = (X^T X + \Omega_0^{-1})^{-1},$$

$$\alpha_n = \alpha_0 + \frac{n}{2},$$

$$\beta_n = \beta_0 + \frac{1}{2} (y^T y + \mu_0^T \Omega_0^{-1} \mu_0 + \mu_n^T \Omega_n^{-1} \mu_n).$$

The marginal posterior for  $\beta$ ,  $\pi(\beta | y, X)$  is the **multivariate  $t$**  distribution with  $2\alpha_n$  degrees of freedom, location  $\mu_n$  and scale  $\frac{\beta_n}{\alpha_n} \Omega_n$

## Marginal Posterior of $\beta$ and Predictive distribution

To obtain **credible intervals** for  $\beta$  we could use the  $t$  distribution. Alternatively, Monte Carlo can be used; more general choice.

We draw  $N$  samples from  $\pi(\beta|y)$  and use **Monte Carlo** to calculate credible intervals, density plots etc.

Samples from  $\pi(\beta|y)$  can be drawn in the following way:

- 1 Generating samples  $\sigma_i^2$  from **IGamma**( $\alpha_n, \beta_n$ ),  $i = 1, \dots, N$ ,
- 2 Draw  $\beta_i$  sample based on each  $\sigma_i^2$  from **N**( $\mu_n, \sigma_i^2 \Omega_n$ ),

For the **predictive distribution** for a new observation  $y_*$  based on covariates  $X_*$  we can use the additional step of drawing  $y_{*i}$  from **N**( $X_*\beta_i, \sigma_i^2$ ) for each  $(\beta_i, \sigma_i^2)$ .

# Bayesian Model Selection

To compare models we will need to compute the **model evidence** for each model.

We can then use the model with the **highest marginal likelihood**. The use of **unit information** prior is the default option to guard against the Jeffreys-Lindley paradox.

Computing the **model evidence** is generally a very difficult task but here we can use the following **trick**. We can write

$$\begin{aligned}\pi(\beta, \sigma^2 | y, X) &= \frac{\pi(y | \beta, \sigma^2, X) \pi(\beta, \sigma^2)}{\pi(y | X)}, \text{ or else} \\ \pi(y | X) &= \frac{\pi(y | \beta, \sigma^2, X) \pi(\beta, \sigma^2)}{\pi(\beta, \sigma^2 | y, X)}, \text{ for all } \beta, \sigma^2.\end{aligned}$$

The expression above contains **known** Normal and Inverse Gamma pdfs so we can just evaluate for -say- the posterior mean of  $\beta, \sigma^2$ .



# Summary on Bayesian Linear Regression

Bayesian linear regression can be used to

- Incorporate **prior information** the regression coefficients.
- Guard against **overfitting** by shrinkage.
- Select **parsimonious** models via the model evidence.
- Address the **multicollinearity** issue.

# Reading

Gelman et al:

Chapter 14

# Outline

- 1 Linear Models for Regression
- 2 Bayesian Linear Regression
- 3 Optional: MLE derivation

## Some results from matrix algebra/calculus

Let  $A, B$  be arbitrary matrices,  $I$  be the identity matrix  $v$  is a vector and  $c$  is a constant.

The **transpose** of a matrix  $A$  is denoted by  $A^T$ : so that rows and columns are flipped. We also have

- $(A^T)^T = A$ ,  $(A + B)^T = A^T + B^T$ ,  $(AB)^T = B^T A^T$ .
- For symmetric matrices only  $A^T = A$ .

**Check:** Let's expand  $(y - X\beta)^T(y - X\beta)$ :

$$\begin{aligned}(y - X\beta)^T(y - X\beta) &= (y^T - (X\beta)^T)(y - X\beta) \\ &= (y^T - \beta^T X^T)(y - X\beta) \\ &= y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta\end{aligned}$$

## Some results from matrix algebra/calculus (cont'd)

The **inverse** of a matrix  $A$  is denoted with  $A^{-1}$ : such that  $A^{-1}A = AA^{-1} = I$ .

We also have  $(AB)^{-1} = B^{-1}A^{-1}$ .

**Derivatives wrt vector  $v$ .**

$$\frac{\partial}{\partial v} v^T A = A$$

$$\frac{\partial}{\partial v} A^T v = A \text{ or else } \frac{\partial}{\partial v} A v = A^T$$

$$\frac{\partial}{\partial v} v^T v = 2v$$

$$\frac{\partial}{\partial v} v^T A v = 2A v$$

Now we are ready to derive the MLE/Least squares estimator of  $\beta$ .

# MLE of Linear Regression

The **log-likelihood** function can be written as ( $c_1, c_2$  are constants)

$$\begin{aligned}\log f(y|X, \beta, \sigma^2) &= c_1 - \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) \\ &= c_2 - \frac{1}{2\sigma^2}(-y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta)\end{aligned}$$

**Differentiating with respect to  $\beta$**  gives

$$\begin{aligned}\frac{\partial}{\partial \beta} \log f(y|X, \beta, \sigma^2) &= -\frac{1}{2\sigma^2}(-X^T y - X^T y + 2X^T X\beta) \\ &= \frac{1}{\sigma^2}(X^T y - X^T X\beta)\end{aligned}$$

Get  $\hat{\beta}$  from setting  $\frac{\partial}{\partial \hat{\beta}} \log f(y|X, \hat{\beta}, \sigma^2) = 0$ :

$$X^T y - X^T X\hat{\beta} = 0 \quad \text{or else} \quad \hat{\beta} = (X^T X)^{-1} X^T y$$

# MLE of Linear Regression - Least Squares Estimator

Similar (tedious) calculations yield that

$$\text{Var}(\hat{\beta}) = \hat{\sigma}^2 (X^T X)^{-1}$$

The MLE for the variance is  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - X_i \hat{\beta})^2$  although sometimes its unbiased version is preferred where we divide by  $n - p - 1$  rather than  $n$ .

This is how tables like the one below are obtained:

Term	Coefficient	Std. Error	Z Score
Intercept	2.46	0.09	27.60
lcavol	0.68	0.13	5.37
lweight	0.26	0.10	2.75
age	-0.14	0.10	-1.40
lbph	0.21	0.10	2.06
svi	0.31	0.12	2.47
lcp	-0.29	0.15	-1.87
gleason	-0.02	0.15	-0.15
pgg45	0.27	0.15	1.74