Notes on statistical learning

Dávid Iván

December 18, 2020

Contents

1 Introduction

2 Overview of Supervised Learning

2.1 Linear models and least squares

On page 12 we have that the residual sum of squares:

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N}(y_i - x_i^T\beta)^2 = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) \tag{1}$$

How can we differentiate with respect to β?

$$\mathrm{RSS}(\beta) = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\beta - \beta^T\mathbf{X}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta \tag{2}$$

Using (139, 140), we can differentiate RSS:

$$\frac{d}{d\beta}\mathrm{RSS}(\beta) = 0 - (\mathbf{y}^T\mathbf{X})^T - \mathbf{X}^T\mathbf{y} + (\mathbf{X}^T\mathbf{X} + (\mathbf{X}^T\mathbf{X})^T)\beta \tag{3}$$

$$\frac{d}{d\beta}\mathrm{RSS}(\beta) = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\beta \tag{4}$$

Setting this to zero we get the normal equations:

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\beta \tag{5}$$

2.2 Statistical decision theory

On page 19 we have that

$$\beta = [E(XX^T)]^{-1}E(XY) \tag{6}$$

but how exactly do we get this equation? In general, we have the expected prediction error:

$$\text{EPE}(f) = E(Y - f(X))^2 \tag{7}$$

And we have that the prediction function is linear:

$$f(X) = X^T\beta \tag{8}$$

We seek a β for minimizing the expected prediction error. X and Y are random variables, X being a vector, Y being a scalar.

$$\begin{aligned} \frac{d}{d\beta}\text{EPE} = \frac{d}{d\beta}\text{E}((Y - X^T\beta)^2) = \text{E}\left(\frac{d}{d\beta}(Y - X^T\beta)^2\right) \\ = \text{E}\left(2(Y - X^T\beta)\cdot(-X)\right) = -2\text{E}(YX) + 2\text{E}(X(X^T\beta)) \\ = -2\text{E}(YX) + 2\text{E}((XX^T)\beta) = -2\text{E}(YX) + 2(\text{E}(XX^T))\beta \end{aligned} \tag{9}$$

We used the fact that the expected value is linear, and that β is not random, so we could factor out from the expected value. Setting this to zero we have that:

$$\text{E}(YX) = \text{E}(XX^T)\beta \tag{10}$$

which yields

$$\hat{\beta} = [\text{E}(XX^T)]^{-1}\text{E}(YX) \tag{11}$$

2.2.1 application. Simple linear fit.

Let's see an application for this equation. Let $X = \begin{bmatrix} x \\ 1 \end{bmatrix}$, $\beta = \begin{bmatrix} a \\ b \end{bmatrix}$.

Now $f(x) = a \cdot x + b$

$$XY = \begin{bmatrix} x \cdot y \\ y \end{bmatrix} \tag{12}$$

$$XX^T = \begin{bmatrix} x^2 & x \\ x & 1 \end{bmatrix} \tag{13}$$

If we have N datapoints $\{(x_1, y_1), (x_2, y_2), ...(x_N, y_N)\}$, we can approximate the expectation values.

$$\mathrm{E}(XY) \approx \frac{1}{N} \begin{bmatrix} \sum_i x_i \cdot y_i \\ \Sigma_i y_i \end{bmatrix} \tag{14}$$

$$\mathrm{E}(XX^T) \approx \frac{1}{N} \begin{bmatrix} \sum_i x_i^2 & \Sigma_i x_i \\ \Sigma_i x_i & N \end{bmatrix} \tag{15}$$

Let's denote the followings:

$$\alpha_X = \sum_i x_i \tag{16}$$

$$\alpha_Y = \sum_i y_i \tag{17}$$

$$\alpha_{XY} = \sum_i x_i y_i \tag{18}$$

$$\alpha_{X^2} = \sum_i x_i^2 \tag{19}$$

With these notations:

$$\mathrm{E}(XY) \approx \frac{1}{N} \begin{bmatrix} \alpha_{XY} \\ \alpha_Y \end{bmatrix} \tag{20}$$

$$\mathrm{E}(XX^T) \approx \frac{1}{N} \begin{bmatrix} \alpha_{X^2} & \alpha_X \\ \alpha_X & N \end{bmatrix} \tag{21}$$

Inverting $\mathrm{E}(XX^T)$:

$$[\mathrm{E}(XX^T)]^{-1} \approx \frac{N}{N\alpha_{X^2} - \alpha_X^2} \begin{bmatrix} N & -\alpha_X \\ -\alpha_X & \alpha_{X^2} \end{bmatrix} \tag{22}$$

Plug these in to the equation:

$$\hat{\beta} = [\mathrm{E}(XX^T)]^{-1}\mathrm{E}(YX) \tag{23}$$

$$\hat{\beta} \approx \frac{N}{N\alpha_{X^2} - \alpha_X^2} \begin{bmatrix} N & -\alpha_X \\ -\alpha_X & \alpha_{X^2} \end{bmatrix} \frac{1}{N} \begin{bmatrix} \alpha_{XY} \\ \alpha_Y \end{bmatrix} \tag{24}$$

$$\hat{\beta} \approx \frac{1}{N\alpha_{X^2} - \alpha_X^2} \begin{bmatrix} N\alpha_{XY} - \alpha_X\alpha_Y \\ \alpha_{X^2}\alpha_Y - \alpha_X\alpha_{XY} \end{bmatrix} \tag{25}$$

From here we can get $\hat{a}$ and $\hat{b}$, since $\hat{\beta} = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix}$

2.3 E$|Y-c|$ and the median

On page 20, it asks the question "What happens if we replace the L_2 loss function with the $L_1 : \mathrm{E}|Y - f(X)|$?" Let's investigate this question.

2.3.1 discrete case

We can get rid of the conditional $X = x$, and just ask the question: What c will minimize E$|Y - c|$? Denote this function with g, so $g(c) = \mathrm{E}|Y - c|$. Let's look at two examples.

Example 1. The random variable Y takes 4 possible values with probabilities $\frac{1}{7}$, $\frac{1}{7}$, $\frac{3}{7}$, $\frac{2}{7}$. The figure below shows the probability mass funciton.

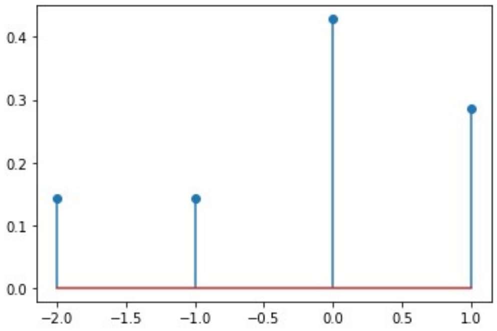

Figure 1: probability mass function of the first example random variable.

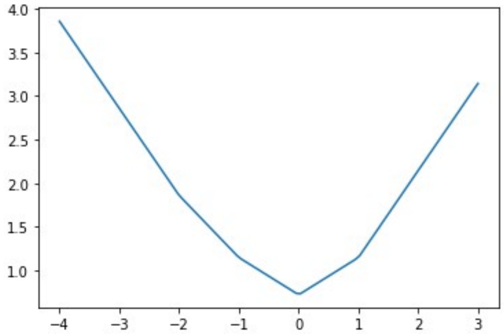

Figure 2: $g(c)$ function. The horizontal axis is c.

Example 2. The random variable Y takes 4 possible values with probabilities 0.1, 0.4, 0.3, 0.2. The figure below shows the probability mass funciton.

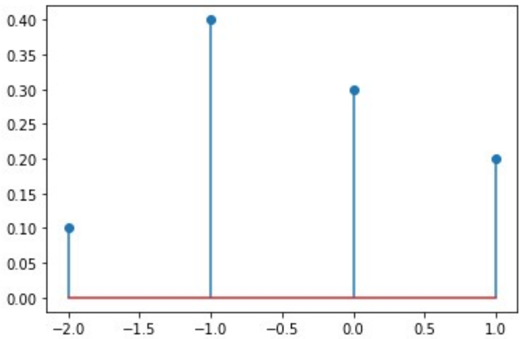

Figure 3: probability mass function of the second example random variable.

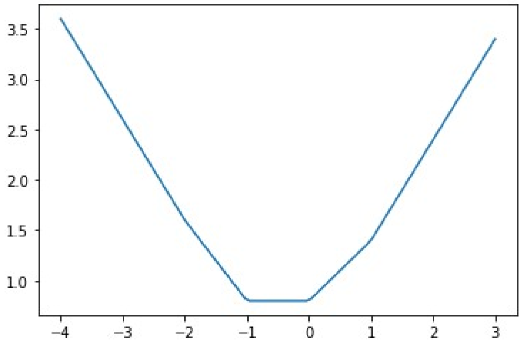

Figure 4: $g(c)$ function. The horizontal axis is c.

We can see that $g(c)$ is a piecewise linear function, and it has minimum, which is a point or a line segment. Let's say we have Y discrete random variable that takes values from $S = \{x_1, x_2, \ldots x_n\}$. The values are ordered: $x_1 < x_2 < ... < x_n$. Y takes these values with corresponding probabilities p_1, p_2, ..., p_n.

Let's calculate the equation of the piecewise linear function. Denote the interval I_k such that $x \in I_k$ if and only if k values from S are smaller than x. So $I_0 = (-\infty, x_1]$, $I_1 = [x_1, x_2]$, ..., $I_n = [x_n, \mathrm{inf})$.

$$g(c) = \mathrm{E}|Y - c| = \Sigma_{i=1}^{n} p_i \cdot |x_i - c| \tag{26}$$

If $c \in I_k$, then

$$g(c) = \mathrm{E}|Y - c| = \Sigma_{i=1}^{k} p_i \cdot (c - x_i) + \Sigma_{i=k+1}^{n} p_i \cdot (x_i - c) \tag{27}$$

$$g(c) = c \cdot (\Sigma_{i=1}^{k} p_i - \Sigma_{i=k+1}^{n} p_i) + (\Sigma_{i=k+1}^{n} p_i x_i - \Sigma_{i=1}^{k} p_i x_i) \tag{28}$$

First we can show that this function is continuous. On one hand ($c \in I_k = [x_k, x_k + 1]$):

$$g_1 = g(x_k) = x_k \cdot (\Sigma_{i=1}^{k} p_i - \Sigma_{i=k+1}^{n} p_i) + (\Sigma_{i=k+1}^{n} p_i x_i - \Sigma_{i=1}^{k} p_i x_i) \tag{29}$$

On the other hand we have ($c \in I_{k-1} = [x_{k-1}, x_k]$):

$$g_2 = g(x_k) = x_k \cdot (\Sigma_{i=1}^{k-1} p_i - \Sigma_{i=k}^{n} p_i) + (\Sigma_{i=k}^{n} p_i x_i - \Sigma_{i=1}^{k-1} p_i x_i) \tag{30}$$

$$g_1 - g_2 = x_k \cdot (p_k + p_k) - p_k x_k - p_k x_k = 0 \tag{31}$$

Now that we showed that this function is continuous, let's find it's minimum. Since it is piecewise linear, its derivative is piecewise constant. Denote the derivative of g on the interval I_k with $g'(I_k)$.

$$\begin{aligned} g'(I_k) &= -1 \\ g'(I_1) &= -1 + 2p_1 \\ g'(I_2) &= -1 + 2p_1 + 2p_2 \\ &\dots \\ g'(I_n) &= -1 + 2p_1 + ... + 2p_n = 1 \end{aligned} \tag{32}$$

So the derivative is increasing from -1 to $+1$. We can distinguish two possibilities. First, assume that the derivative is never zero. In this case, we have a k where $g'(I_{k-1}) < 0$ but $g'(I_k) > 0$, so the minimum is at x_k, the median. The second case is where there is an interval where the derivative is zero. In this case the whole interval is minimum, again, the median.

2.3.2 continuous case

Let's have the following function:

$$f(x) = \int_a^x g(x,t)dt \tag{33}$$

I state without proof that the derivative of this function is as follows:

$$f'(x) = \int_a^x \frac{\partial g(x,t)}{\partial x} dt + g(x,x) \tag{34}$$

Now we have that

$$g(c) = \mathrm{E}(|Y - c||X = x) = \int_{-\infty}^{c} (c - y) f_{Y|X}(y|x) dy + \int_{c}^{\infty} (y - c) f_{Y|X}(y|x) dy \tag{35}$$

$$g'(c) = \int_{-\infty}^{c} f_{Y|X}(y|x) dy + \int_{\infty}^{c} f_{Y|X}(y|x) dy \tag{36}$$

Setting this to zero, we get that

$$\int_{-\infty}^{c} f_{Y|X}(y|x) dy = \int_{c}^{\infty} f_{Y|X}(y|x) dy \tag{37}$$

$$P(Y < c \mid X = x) = P(Y > c \mid X = x) \tag{38}$$

Again, this means the minimum is at the median.

2.4 Local methods in high dimensions

2.4.1 deriving the prediction formula

On page 24 we see an example of a linear data with noise. At first I was confused how it gets $\hat{y}_0 = x_0^T\beta + \sum_{i=1}^{N} l_i(x_0)\epsilon_i$, where $l_i(x_0)$ is the ith element of $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x_0$

In this example we make N experiments, storing the X_i values in the rows of $\mathbf{X}$, and we have also ϵ_i (elements of $\vec{\epsilon}$) and $Y_i = X_i^T\beta + \epsilon_i$ for some fixed β. For approximating β, we use the result:

$$\beta = [\mathrm{E}(XX^T)]^{-1}\mathrm{E}(YX) \tag{39}$$

in this case it will be an approximation, since we have noise (ϵ).

Calculate first $\mathrm{E}(YX)$:

$$\mathrm{E}(YX) = \mathrm{E}((X^T\beta + \epsilon)X) = \mathrm{E}(XX^T)\beta + \mathrm{E}(\epsilon X) \tag{40}$$

Substitute this into the approximation of β:

$$\begin{aligned} \hat{\beta} &= [\mathrm{E}(XX^T)]^{-1}\mathrm{E}(YX) \\ &= [\mathrm{E}(XX^T)]^{-1}(\mathrm{E}(XX^T)\beta + \mathrm{E}(\epsilon X)) \\ &= \beta + [\mathrm{E}(XX^T)]^{-1}\mathrm{E}(\epsilon X) \end{aligned} \tag{41}$$

We do not know of course the exact expectation values, but we have N data samples (training data). So how could we approximate the expectation values? Use the averages:

$$\mathrm{E}(\epsilon X)_i \approx \frac{1}{N}\sum_{k=1}^{N} \mathbf{X}_{ki}\epsilon_k \to \mathrm{E}(\epsilon X) \approx \frac{1}{N}\mathbf{X}^T\vec{\epsilon} \tag{42}$$

similarly,

$$[\mathrm{E}(XX^T)]^{-1} \approx N \cdot (\mathbf{X}^T\mathbf{X})^{-1} \tag{43}$$

putting these all together, we have:

$$\hat{y}_0 = x_0^T\hat{\beta} = x_0^T\beta + x_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{\epsilon} = x_0^T\beta + \vec{\epsilon}^T \cdot \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x_0 \tag{44}$$

and this is the formula that was to be explained.

2.4.2 equation (2.47) on page 37

$$\begin{aligned}\mathrm{EPE}_k(x_0) &= \mathrm{E}\left((Y-\hat{f}_k(x_0))^2|X=x_0\right)\\ &= \mathrm{E}\left((f(x_0)+\epsilon_0-\hat{f}_k(x_0))^2\right)\end{aligned} \tag{45}$$

The data points are fixed: $x_1, x_2, \ldots, x_N$. Denote the closest data point to x_0 as $x_{(1)}$, the second closest $x_{(2)}$, etc. With this notation, the nearest neighbor estimate for $f(x_0)$:

$$\hat{f}_k(x_0) = \frac{1}{k}\sum_{l=1}^{k}(f(x_{(l)})+\epsilon_l) \tag{46}$$

In this equation, $f(x_{(l)})$ is fixed, and epsilons are iid random variables.

$$\begin{aligned}\mathrm{EPE}_k(x_0) &= \mathrm{E}\left(\left(f(x_0)+\epsilon_0-\frac{1}{k}\sum_{l=1}^{k}(f(x_{(l)})+\epsilon_l)\right)^2\right)\\ &= \mathrm{E}\left(\left[\left(f(x_0)-\frac{1}{k}\sum_{l=1}^{k}f(x_{(l)})\right)+\left(\epsilon_0-\frac{1}{k}\sum_{l=1}^{k}\epsilon_l\right)\right]^2\right)\\ = \mathrm{E}\left([\hat{F}+\hat{E}]^2\right) &= \mathrm{E}\left(\hat{F}^2+2\cdot\hat{F}\hat{E}+\hat{E}^2\right) = \hat{F}^2+2\hat{F}\cdot\mathrm{E}(\hat{E})+\mathrm{E}(\hat{E}^2)\end{aligned} \tag{47}$$

where $\hat{F}$ is nonrandom, and $\hat{E}$ is random:

$$\begin{aligned}\hat{F} &\equiv f(x_0)-\frac{1}{k}\sum_{l=1}^{k}f(x_{(l)}),\\ \hat{E} &\equiv \epsilon_0-\frac{1}{k}\sum_{l=1}^{k}\epsilon_l\end{aligned} \tag{48}$$

Let's calculate the expectation of $\hat{E}$:

$$\mathrm{E}(\hat{E}) = \mathrm{E}\left(\epsilon_0-\frac{1}{k}\sum_{l=1}^{k}\epsilon_l\right) = \mathrm{E}(\epsilon_0)-\frac{1}{k}\sum_{l=1}^{k}\mathrm{E}(\epsilon_l) = 0-\frac{1}{k}\sum_{l=1}^{k}0 = 0 \tag{49}$$

The expectation of $\hat{E}^2$:

$$\mathrm{E}(\hat{E}^2) = \mathrm{E}\left(\epsilon_0 - \frac{1}{k}\sum_{l=1}^{k}\epsilon_l\right)^2 = \mathrm{E}\left(\epsilon_0^2 + \sum_{l=1}^{k}\frac{\epsilon_l^2}{k^2} + CrossProducts\right) \tag{50}$$

The expectation of the cross products are zero, since epsilons are independent, so $\mathrm{E}(\epsilon_i\epsilon_j) = \mathrm{E}\epsilon_i \cdot \mathrm{E}\epsilon_j = 0 \cdot 0 = 0$

$$\begin{aligned} \mathrm{E}(\hat{E}^2) = \mathrm{E}\left(\epsilon_0^2 + \sum_{l=1}^{k}\frac{\epsilon_l^2}{k^2}\right) &= \mathrm{E}\epsilon_0^2 + \sum_{l=1}^{k}\frac{\mathrm{E}\epsilon_l^2}{k^2} \\ &= \sigma^2 + \sum_{l=1}^{k}\frac{\sigma^2}{k^2} = \sigma^2 + \frac{\sigma^2}{k} \end{aligned} \tag{51}$$

We used the fact that the error has zero mean, so the variance is $\sigma^2 = \mathrm{Var}(\epsilon) = \mathrm{E}(\epsilon^2) - (\mathrm{E}\epsilon)^2 = \mathrm{E}(\epsilon^2)$. So the final form is:

$$\begin{aligned} \mathrm{E}_k(x_0) &= \hat{F}^2 + 2\hat{F}\cdot\mathrm{E}(\hat{E}) + \mathrm{E}(\hat{E}^2) = \hat{F}^2 + \mathrm{E}(\hat{E}^2) \\ &= \left(f(x_0) - \frac{1}{k}\sum_{l=1}^{k} f(x_{(l)})\right)^2 + \sigma^2 + \frac{\sigma^2}{k} \end{aligned} \tag{52}$$

2.5 Solutions for the Exercises of chapter 2

2.5.1 Ex. 2.2

We have $X \in \mathbb{R}^p$ continuous and G discrete random variables. Assume we have K classes. Each class has its own distribution, let's say that class g has a pdf $f_g(x)$ ($x \in \mathbb{R}^p$). When generating points, we first choose a class with associated probabilities $p_1, p_2, \ldots, p_K$ ($\sum p_i = 1$). When we have chosen the class, we generate a point with the appropriate distribution.

The Bayes classifier classifies each point x to the most probable class. So let's calculate the probability of class g, given the point. It should be noted that when I write $\mathrm{P}(x)$, I mean "the probability that the chosen point is in the infinitesimal neighborhood of x". So I should write $\mathrm{P}(X \in b_{dx}(x))$, i.e., the probability that X is in the dx-volume ball around x. If the pdf was $f(x)$, this probability is $f(x)dx$. But instead, I'll write $\mathrm{P}(x) = f(x)$. Likewise, when I write $\mathrm{P}(g)$, I mean $\mathrm{P}(G = g)$.

$$\mathrm{P}(g|x) = \frac{\mathrm{P}(g \cap x)}{\mathrm{P}(x)} = \frac{\mathrm{P}(g \cap x)}{\mathrm{P}(x)} = \frac{\mathrm{P}(x|g)\mathrm{P}(g)}{\sum_{g'}\mathrm{P}(x|g')\mathrm{P}(g')} \tag{53}$$

The denominator is a normalizing constant, so the chosen class, for which $\mathrm{P}(g|x)$ is maximum:

$$\hat{g}(x) = max_g \mathrm{P}(x|g)\mathrm{P}(g) \tag{54}$$

2.5.2 Ex. 2.3

Given a unit ball in p-dimension. We sample N data points from it uniformly. Let X be the distance from the origin. The pdf must be proportional to x^{p-1}, and integrating it from 0 to 1 gives 1, thus the pdf:

$$f(x) = p \cdot x^{p-1} \tag{55}$$

The probability that a random sample is at least x distant from the origin is:

$$\mathrm{P}(X > x) = \int_x^1 f(x)dx = 1 - x^p \tag{56}$$

The probability that all N sample points are further from origin as x:

$$\mathrm{P}(X_1 > x \cap X_2 > x \cap \cdots \cap X_N > x) = (1 - x^p)^N \tag{57}$$

We seek and x for that this probability is a half (that will give us the median):

$$\begin{aligned} (1 - x^p)^N &= \frac{1}{2} \\ 1 - x^p &= \left(\frac{1}{2}\right)^{1/N} \\ \left[1 - \left(\frac{1}{2}\right)^{1/N}\right]^{1/p} &= x \end{aligned} \tag{58}$$

2.5.3 Ex. 2.4

If we choose a as the first unit base vector ($a = [1, 0, 0, \ldots, 0]^T$), then $a^T \cdot x_i$ is the first coordinate of x_i. It is by definition (standard) normally distributed. Since the distribution is spherically symmetric, we can choose any direction a, $a^T \cdot x_i$ remains standard normal.

I created an experiment on this. Created 1000 sample points in p dimension, and rotated them into the first 2 dimension, so that we can visualize the distances. On the first image below we can see that the points get further and further away from the origin as the dimension increases.

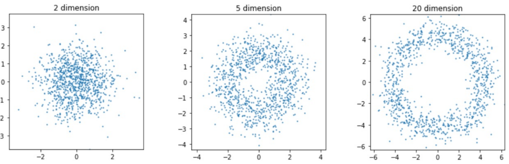

Figure 5: The sample points get further from the origin as we increase the dimension.

But this doesn't mean the points are close to each other. In the following experiment I took the random sample points, chose one of them and set it as the new origin. We can see that still the points are far from a random sample point as we increase the dimension.

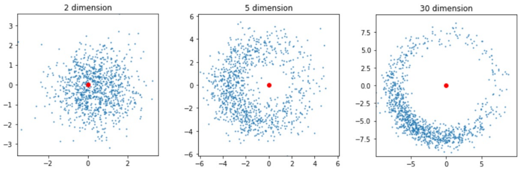

Figure 6: The points get further and further from each other (red dot is a randomly selected sample point) as we increase the dimension.

2.5.4 Ex. 2.5

equation (2.27) on page 26 I won't use indices at the expectation sign, it always confuses me. So this is the expected prediction error:

$$\mathrm{EPE}(x_0) = \mathrm{E}(y_0 - \hat{y_0})^2 \tag{59}$$

Recall, that $y_0 = x_0^T\beta + \epsilon$ is a random variable, since $\epsilon \sim N(0, \sigma^2)$. This is the label (the ground truth) for x_0. The prediction that we make for x_0 is $\hat{y_0} = x_0^T\beta + \vec{\epsilon}^T \cdot \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x_0$. x_0 is a p-vector, β is a p-vector, $\vec{\epsilon}$ is an n-vector, and $\mathbf{X}$ is a n by p matrix (each row is a training sample vector). $\hat{y_0} = x_0^T\beta + \vec{\epsilon}^T \cdot \mathbf{Z}^T x_0$. Here I introduced the p by n matrix $\mathbf{Z}$:

$$\mathbf{Z} \equiv (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{60}$$

In the expression of $\hat{y}_0$, $\vec{\epsilon}$ and $\mathbf{Z}$ are the only random variables. The elements of $\vec{\epsilon}$ are iid RVs. $\vec{\epsilon}$ and $\mathbf{Z}$ are independent. Let's calculate the expectation values of y_0 and $\hat{y}_0$:

$$\mathrm{E}(y_0) = \mathrm{E}(x_0^T\beta + \epsilon) = x_0^T\beta \tag{61}$$

$$\begin{aligned}
\mathrm{E}(\hat{y}_0) =& \mathrm{E}(x_0^T\beta + \vec{\epsilon}^T \cdot \mathbf{Z}^T x_0) \\
=& x_0^T\beta + \mathrm{E}(\vec{\epsilon}^T \cdot \mathbf{Z}^T x_0) \\
=& x_0^T\beta + \mathrm{E}(\vec{\epsilon}^T)\mathrm{E}(\mathbf{Z}^T x_0) \\
=& x_0^T\beta + \vec{0}^T \cdot \mathrm{E}(\mathbf{Z}^T x_0) \\
=& x_0^T\beta
\end{aligned} \tag{62}$$

Here I used that $\vec{\epsilon}$ and $\mathbf{Z}$ are independent, so the expectation value of their product is the product of their expectations. For simplicity, denote $\mu \equiv \mathrm{E}(y_0) = \mathrm{E}(\hat{y}_0) = x_0^T\beta$. The expected prediction error:

$$\begin{aligned}
\mathrm{EPE}(x_0) =& \mathrm{E}(y_0 - \mu + \mu - \hat{y}_0)^2 \\
=& \mathrm{E}(y_0 - \mu)^2 - 2 \cdot \mathrm{E}((y_0 - \mu)(\hat{y}_0 - \mu)) + \mathrm{E}(\hat{y}_0 - \mu)^2 \\
=& \mathrm{E}(y_0 - \mu)^2 + 0 + \mathrm{E}(\hat{y}_0 - \mu)^2 \\
=& \mathrm{Var}(y_0) + \mathrm{Var}(\hat{y}_0)
\end{aligned} \tag{63}$$

Note that y_0 and $\hat{y}_0$ are independent. The epsilon in y_0 is a scalar and is nothing to do with the vector epsilon in $\hat{y}_0$. This is why $\mathrm{E}((y_0 - \mu)(\hat{y}_0 - \mu))$ is zero. Now let's derive the variances:

$$\mathrm{Var}(y_0) = \mathrm{Var}(\mu + \epsilon) = \mathrm{Var}(\epsilon) = \sigma^2 \tag{64}$$

Furthermore, we can write $\mathrm{Cov}(X, X) = \mathrm{E}((X - \mathrm{E}X) \cdot (X - \mathrm{E}X)^T) = \mathrm{E}(X \cdot X^T) - (\mathrm{E}X)(\mathrm{E}X^T)$. Let's apply (146) and (147) to derive the variance of $\hat{y}_0$:

$$\begin{aligned}
\mathrm{Var}(\hat{y}_0) =& \mathrm{Var}(\mu + x_0^T \cdot \mathbf{Z} \cdot \vec{\epsilon}) = \mathrm{Var}(x_0^T \cdot \mathbf{Z} \cdot \vec{\epsilon}) \\
=& x_0^T \cdot \mathrm{Cov}(\mathbf{Z} \cdot \vec{\epsilon}, \mathbf{Z} \cdot \vec{\epsilon}) \cdot x_0 \\
=& x_0^T \cdot \Big(\mathrm{E}(\mathbf{Z}\vec{\epsilon} \cdot \vec{\epsilon}^T \mathbf{Z}^T) - \mathrm{E}(\mathbf{Z}\vec{\epsilon}) \cdot \mathrm{E}(\vec{\epsilon}^T \mathbf{Z}^T)\Big) \cdot x_0 \\
=& x_0^T \cdot \mathrm{E}(\mathbf{Z}\vec{\epsilon} \cdot \vec{\epsilon}^T \mathbf{Z}^T) \cdot x_0
\end{aligned} \tag{65}$$

Here we used the fact that $\mathbf{Z}$ and $\vec{\epsilon}$ are independent, so $\mathrm{E}(\mathbf{Z}\vec{\epsilon}) = \mathrm{E}\mathbf{Z} \cdot \mathrm{E}\vec{\epsilon} = \mathrm{E}\mathbf{Z} \cdot \vec{0} = \vec{0}$, and $\vec{0} \cdot \vec{0}^T = \mathbf{0}$, zero matrix.

$$
\begin{aligned}
\left(\mathbf{Z}\vec{\epsilon}\cdot\vec{\epsilon}^T\mathbf{Z}^T\right)_{i,j} &= \sum_{k,l}(\mathbf{Z})_{i,k}\cdot(\vec{\epsilon}\vec{\epsilon}^T)_{k,l}\cdot(\mathbf{Z}^T)_{l,j} \\
\to \mathrm{E}\left(\mathbf{Z}\vec{\epsilon}\cdot\vec{\epsilon}^T\mathbf{Z}^T\right)_{i,j} &= \sum_{k,l}\mathrm{E}\left(Z_{i,k}\cdot\epsilon_k\cdot\epsilon_l\cdot Z_{j,l}\right) \\
=\sum_{k,l}\mathrm{E}\left(Z_{i,k}\cdot Z_{j,l}\right)\cdot\mathrm{E}\left(\epsilon_k\cdot\epsilon_l\right) &= \sum_{k,l}\mathrm{E}\left(Z_{i,k}\cdot Z_{j,l}\right)\cdot\sigma^2\delta_{k,l} \\
=\sigma^2\cdot\sum_{k}\mathrm{E}\left(Z_{i,k}\cdot Z_{j,k}\right)=\sigma^2\cdot\mathrm{E}\sum_{k}\left(Z_{i,k}\cdot Z_{j,k}\right) &= \sigma^2\cdot\mathrm{E}(\mathbf{Z}\mathbf{Z}^T)_{i,j} \\
\to \mathrm{E}(\mathbf{Z}\vec{\epsilon}\cdot\vec{\epsilon}^T\mathbf{Z}^T) &= \sigma^2\cdot\mathrm{E}(\mathbf{Z}\mathbf{Z}^T)
\end{aligned}
\tag{66}
$$

Substituting this into (65):

$$
\mathrm{Var}(\hat{y}_0) = x_0^T\cdot\sigma^2\mathrm{E}(\mathbf{Z}\mathbf{Z}^T)\cdot x_0 \tag{67}
$$

$$
\mathrm{E}(\mathbf{Z}\mathbf{Z}^T) = \mathrm{E}\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\right) = \mathrm{E}\left((\mathbf{X}^T\mathbf{X})^{-1}\right) \tag{68}
$$

Putting it all together:

$$
\mathrm{EPE}(x_0) = \mathrm{Var}(y_0)+\mathrm{Var}(\hat{y}_0) = \sigma^2+\sigma^2\cdot x_0^T\cdot\mathrm{E}\left((\mathbf{X}^T\mathbf{X})^{-1}\right)\cdot x_0 \tag{69}
$$

And this is what we wanted to derive.

equation (2.28) on page 26

$$
\mathrm{E}\left(x_0^T\mathrm{Cov}(X)^{-1}x_0\right) = \mathrm{E}\sum_{i,j}x_{0,i}\mathrm{Cov}(X)^{-1}_{i,j}x_{0,j} \tag{70}
$$

Assuming that $x_0 \sim X$, i.e., x_0 (the test point) has the same distribution as X (the training data), and the expectation of it is the zero vector, $\mathrm{Cov}(x_0) = \mathrm{Cov}(X) = \mathrm{E}(x_0x_0^T)$.

$$
\begin{aligned}
\mathrm{E}\sum_{i,j}x_{0,i}\mathrm{Cov}(X)^{-1}_{i,j}x_{0,j} &= \mathrm{E}\sum_{i,j}\mathrm{Cov}(X)^{-1}_{i,j}\mathrm{Cov}(x_0)_{i,j} \\
=\mathrm{E}\sum_{i,j}\mathrm{Cov}(X)^{-1}_{i,j}\mathrm{Cov}(X)_{i,j} = \mathrm{E}\sum_{i}&\left(\sum_{j}\mathrm{Cov}(X)^{-1}_{i,j}\mathrm{Cov}(X)_{j,i}\right) \\
=\mathrm{E}\sum_{i}[\mathrm{Cov}(X)^{-1}\mathrm{Cov}(X)]_{i,i} &= \mathrm{E}\left(\mathrm{Trace}(\mathrm{Cov}(X)^{-1}\mathrm{Cov}(X))\right) \\
&= \mathrm{E}\left(\mathrm{Trace}I_{pxp}\right) = p
\end{aligned}
\tag{71}
$$

2.5.5 Ex. 2.6

Assume that we have n identical inputs $x_1 = x_2 = \cdots = x_n \equiv x$ with outputs y_1, y_2, $\ldots$, y_n. The least squares formula:

$$RSS(\theta) = \sum_{i=1}^{n} (y_i - f_\theta(x))^2 \tag{72}$$

The weighted least squares formula:

$$RSS_w(\theta) = n \cdot \left(\frac{\sum_{i=1}^{n} y_i}{n} - f_\theta(x) \right)^2 \tag{73}$$

I claim that the two expressions differ by a constant term that doesn't depend on θ, so both expressions lead to the same solution. This naturally extends to the case when we have groups of equal inputs.

Expanding RSS:

$$RSS(\theta) = \sum_{i=1}^{n} y_i^2 - 2f_\theta(x) \sum_{i=1}^{n} y_i + f_\theta^2(x) \tag{74}$$

Expanding RSS_w:

$$RSS_w(\theta) = \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} - 2f_\theta(x) \sum_{i=1}^{n} y_i + f_\theta^2(x) \tag{75}$$

So the difference of the 2 expressions is a constant that doesn't depend on θ. So when we derive wrt θ, we get the same formulae.

Whenever we have observations with identical values x, we can always refactor the RSS for the groups according to (72) → (73).

2.5.6 Ex. 2.7

Our estimator according to the problem statement:

$$\hat{f}(x_0) = \sum_{i=1}^{N} l_i(x_0; \mathcal{X}) y_i \tag{76}$$

a) For kNN, the weights are:

$$l_i(x_0; \mathcal{X}) = \frac{1}{k}\delta(x_i \in \mathrm{kNN}(x_0)) \tag{77}$$

where $\delta(x_i \in \mathrm{kNN}(x_0))$ is 1 if x_i is in the set of k-nearest neighbors of x_0, and 0 otherwise. So in this case we average the ys of the k-nearest neighbors of x_0.

For linear regression we have

$$\hat{f}(x_0) = x_0^T \beta \tag{78}$$

Where β comes from the following equation (see Section 2.2):

$$\beta = \mathrm{E}(XX^T)^{-1}\mathrm{E}(XY) \tag{79}$$

Now let's calculate this expression. We estimate the expectation values with averages.

$$\begin{aligned} \mathrm{E}(XX^T)^{-1} &\approx \left(\frac{1}{N}\sum_{j=1}^{N} x_j x_j^T\right)^{-1} \\ \mathrm{E}(XY) &\approx \frac{1}{N}\sum_{i=1}^{N} x_i y_i \end{aligned} \tag{80}$$

With these, we can formulate $\hat{f}(x_0)$ as follows:

$$\begin{aligned} \hat{f}(x_0) &= x_0^T \left(\frac{1}{N}\sum_{j=1}^{N} x_j x_j^T\right)^{-1} \cdot \frac{1}{N}\sum_{i=1}^{N} x_i y_i \\ &= \sum_{i=1}^{N} x_0^T \left(\sum_{j=1}^{N} x_j x_j^T\right)^{-1} x_i y_i \\ &\equiv \sum_{i=1}^{N} l_i(x_0; \mathcal{X}) y_i \end{aligned} \tag{81}$$

From this we get the weights:

$$l_i(x_0; \mathcal{X}) = x_0^T \left(\sum_{j=1}^{N} x_j x_j^T\right)^{-1} x_i \tag{82}$$

2.5.7 Ex. 2.9

The short answer is this:

$$\mathrm{E}R_{tr}(\hat{\beta}) \leq \mathrm{E}R_{tr}(\mathrm{E}\hat{\beta}) = \mathrm{E}R_{te}(\mathrm{E}\hat{\beta}) \leq \mathrm{E}R_{te}(\hat{\beta}) \tag{83}$$

Now I explain this in more details.

1. Proving the left inequality. $\hat{\beta}$ comes from the following:

$$\hat{\beta} = \arg\min_{\beta'} R_{tr}(\beta') \tag{84}$$

This implies that for any fix β:

$$R_{tr}(\hat{\beta}) \leq R_{tr}(\beta) \tag{85}$$

Taking the expectation of both sides:

$$\mathrm{E}R_{tr}(\hat{\beta}) \leq \mathrm{E}R_{tr}(\beta) \tag{86}$$

$\hat{\beta}$ is a random variable (which depends on the training data), we can take the expectation, so we get $\mathrm{E}\hat{\beta}$ which is a fix, non-random vector. Substituting into the above inequality we get what we wanted to prove:

$$\mathrm{E}R_{tr}(\hat{\beta}) \leq \mathrm{E}R_{tr}(\mathrm{E}\hat{\beta}) \tag{87}$$

2. Proving the equation in the middle. For any fix β:

$$\mathrm{E}R_{tr}(\beta) = \frac{1}{N}\sum_{i=1}^{N}\mathrm{E}(y_i - \beta^T x_i)^2 = \mathrm{E}(Y - \beta^T X)^2 \tag{88}$$

$$\mathrm{E}R_{te}(\beta) = \frac{1}{M}\sum_{i=1}^{M}\mathrm{E}(\widetilde{y}_i - \beta^T \widetilde{x}_i)^2 = \mathrm{E}(Y - \beta^T X)^2 \tag{89}$$

This is because both the train and the test data come from the same distribution. So for any fix β, $\mathrm{E}R_{tr}(\beta) = \mathrm{E}R_{te}(\beta)$. Since $\mathrm{E}\hat{\beta}$ is a fix vector, we're done with this part.

3. Proving the right inequality. For this we use the fact that the training data and the test data are independent. Thus $\hat{\beta}$ and the test data are also independent. For this part, just forget about the training data. Think of $\hat{\beta}$ as a random vector independent from the (test) data.

$$\mathrm{E}R_{te}(\hat{\beta}) = \mathrm{E}(Y - \hat{\beta}^T X)^2 = \mathrm{EE}\left((Y - \hat{\beta}^T X)^2 | X, Y\right) \tag{90}$$

$$\begin{aligned}\mathrm{E}\left((Y - \hat{\beta}^T X)^2 | X, Y\right) =& \mathrm{E}\left(Y^2 - 2Y\hat{\beta}^T X + (\hat{\beta}^T X)^2 | X, Y\right)\\ =& Y^2 - 2Y\mathrm{E}(\hat{\beta}^T)X + X^T\mathrm{E}(\hat{\beta}\hat{\beta}^T)X\\ =& Y^2 - 2Y\mathrm{E}(\hat{\beta}^T)X + X^T[\mathrm{E}\hat{\beta}\cdot\mathrm{E}\hat{\beta}^T + \mathrm{Cov}(\hat{\beta})]X\\ =& Y^2 - 2Y\mathrm{E}(\hat{\beta}^T)X + (\mathrm{E}\hat{\beta}^T)XX^T(\mathrm{E}\hat{\beta}) + X^T\mathrm{Cov}(\hat{\beta})X\end{aligned} \tag{91}$$

Since the covariance matrix is positive semi-definite, $X^T\mathrm{Cov}(\beta)X \geq 0$

$$\begin{aligned}\mathrm{E}\left((Y - \hat{\beta}^T X)^2 | X, Y\right) \geq& Y^2 - 2Y\mathrm{E}(\hat{\beta}^T)X + (\mathrm{E}\hat{\beta}^T)XX^T(\mathrm{E})\hat{\beta}\\ \mathrm{E}\left((Y - \hat{\beta}^T X)^2 | X, Y\right) \geq& (Y - \mathrm{E}(\hat{\beta}^T)X)^2\\ \mathrm{EE}\left((Y - \hat{\beta}^T X)^2 | X, Y\right) \geq& \mathrm{E}(Y - \mathrm{E}(\hat{\beta}^T)X)^2\\ \mathrm{E}(Y - \hat{\beta}^T X)^2 \geq& \mathrm{E}(Y - \mathrm{E}(\hat{\beta}^T)X)^2\\ \mathrm{E}R_{te}(\hat{\beta}) \geq& \mathrm{E}R_{te}(\mathrm{E}\hat{\beta})\end{aligned} \tag{92}$$

3 Linear Methods for Regression

3.1 equations on page 47 and 48

Variance of beta hat (page 47). We know the formulae for $\hat{\beta}$ (equation 3.6 on page 45):

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{93}$$

Using (146):

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}) \equiv & \mathrm{Cov}(\hat{\beta}) \\
= & \mathrm{Cov}\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\right) \\
= & (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \cdot \mathrm{Cov}\,(\mathbf{y}) \cdot \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
= & (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \cdot \sigma^2\mathbf{I} \cdot \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
= & \sigma^2 \cdot (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
= & \sigma^2 \cdot (\mathbf{X}^T\mathbf{X})^{-1}
\end{aligned} \tag{94}$$

Sigma hat. Deriving the expectation of $\hat{\sigma}^2$. We know that $\mathbf{H}$, that hat matrix is an orthogonal projection onto the column space of $\mathbf{X}$. This implies that $\mathbf{H}^2 = \mathbf{H} = \mathbf{H}^T$, and $\mathrm{Tr}(\mathbf{H}) = p+1$ (the trace of an orthogonal projection is the dimension of the subspace it projects onto, that is, the rank of $\mathbf{X}$). Another thing is that $(\mathbf{I}-\mathbf{H})$ is also an orthogonal projection. It projects to the orthogonal complement of the column space of $\mathbf{X}$. So $(\mathbf{I}-\mathbf{H})^2 = (\mathbf{I}-\mathbf{H}) = (\mathbf{I}-\mathbf{H})^T$, and $\mathrm{Tr}(\mathbf{I}-\mathbf{H}) = N-p-1$.

$$\begin{aligned}
\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = & (\mathbf{y}-\hat{\mathbf{y}})^T(\mathbf{y}-\hat{\mathbf{y}}) \\
= & (\mathbf{y}-\mathbf{H}\mathbf{y})^T(\mathbf{y}-\mathbf{H}\mathbf{y}) \\
= & ((\mathbf{I}-\mathbf{H})\mathbf{y})^T((\mathbf{I}-\mathbf{H})\mathbf{y}) \\
= & \mathbf{y}^T(\mathbf{I}-\mathbf{H})^T(\mathbf{I}-\mathbf{H})\mathbf{y} \\
= & \mathbf{y}^T(\mathbf{I}-\mathbf{H})^2\mathbf{y} = \mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y} \\
= & \mathrm{Tr}(\mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y}) = \mathrm{Tr}((\mathbf{I}-\mathbf{H})\mathbf{y}\mathbf{y}^T)
\end{aligned} \tag{95}$$

We know that $\mathrm{Cov}(\mathbf{y}) = \sigma^2\mathbf{I} = \mathrm{E}(\mathbf{y}\mathbf{y}^T) - (\mathrm{E}\mathbf{y}) \cdot (\mathrm{E}\mathbf{y}^T)$. Taking the expectation:

$$
\begin{aligned}
\mathrm{E}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2 =&\mathrm{E}\left(\mathrm{Tr}((\mathbf{I}-\mathbf{H})\mathbf{y}\mathbf{y}^T)\right)\\
=&\mathrm{Tr}((\mathbf{I}-\mathbf{H})\mathrm{E}(\mathbf{y}\mathbf{y}^T))\\
=&\mathrm{Tr}((\mathbf{I}-\mathbf{H})\cdot(\sigma^2\mathbf{I}+\mathrm{E}\mathbf{y}\cdot\mathrm{E}\mathbf{y}^T))\\
=&\mathrm{Tr}((\mathbf{I}-\mathbf{H})\sigma^2+(\mathbf{I}-\mathbf{H})\cdot\mathrm{E}\mathbf{y}\cdot\mathrm{E}\mathbf{y}^T)\\
=&\mathrm{Tr}((\mathbf{I}-\mathbf{H})\sigma^2)+\mathrm{Tr}((\mathbf{I}-\mathbf{H})\cdot\mathrm{E}\mathbf{y}\cdot\mathrm{E}\mathbf{y}^T)\\
=&\mathrm{Tr}(\mathbf{I}-\mathbf{H})\cdot\sigma^2+\mathrm{Tr}(\mathrm{E}\mathbf{y}^T\cdot(\mathbf{I}-\mathbf{H})\cdot\mathrm{E}\mathbf{y})\\
=&(N-p-1)\cdot\sigma^2+\mathrm{E}\mathbf{y}^T\cdot(\mathbf{I}-\mathbf{H})\cdot\mathrm{E}\mathbf{y}\\
=&(N-p-1)\cdot\sigma^2
\end{aligned}
\tag{96}
$$

At the last step we had to assume that E$\mathbf{y}$ lies in the column space of $\mathbf{X}$, because it means that E$\mathbf{y}$ and $(\mathbf{I}-\mathbf{H})$E$\mathbf{y}$ are perpendicular to each other. This means that the response (y) is linear in its inputs, plus a random variable with zero mean. Now we see that

$$
\frac{1}{N-p-1}\mathrm{E}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2=\sigma^2\rightarrow\mathrm{E}\hat{\sigma}^2=\sigma^2 \tag{97}
$$

Distribution of sigma hat. (3.11) states that $\hat{\sigma}^2$ is proportional to a Chi-square distribution with $N-p-1$ parameters. Now we use the assumption that $\mathbf{y}=\mathbf{X}\beta+\epsilon$, where $\mathbf{X}$ and β are fixed, and ϵ is a vector of iid normal random variables with zero mean and σ^2 variance. According to (95) we can write that:

$$
\sum_{i=1}^{N}(y_i-\hat{y}_i)^2=\|(\mathbf{I}-\mathbf{H})\mathbf{y}\|^2 \tag{98}
$$

Since $\mathbf{H}$ is a projection to the column space of $\mathbf{X}$, $\mathbf{H}\mathbf{X}=\mathbf{X}$, and $(\mathbf{I}-\mathbf{H})\mathbf{X}=\mathbf{X}-\mathbf{H}\mathbf{X}=\mathbf{X}-\mathbf{X}=\mathbf{0}$. So $(\mathbf{I}-\mathbf{H})\mathbf{y}=(\mathbf{I}-\mathbf{H})\cdot(\mathbf{X}\beta+\epsilon)=(\mathbf{I}-\mathbf{H})\cdot\epsilon$.

$$
\sum_{i=1}^{N}(y_i-\hat{y}_i)^2=\|(\mathbf{I}-\mathbf{H})\epsilon\|^2 \tag{99}
$$

Now it is clear that this is $\sigma^2\cdot\chi^2_{N-p-1}$, because we project the spherical normal distribution (ϵ) to a (N-p-1)-dimensional plane (subspace).

The Z-score. According to (3.12) we form the standardized coefficient or Z-score

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}} \tag{100}$$

Why is this a t-distribution under the null hypothesis that $\beta_j = 0$?

$$\hat{\beta} = \beta + \sigma \cdot (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon = \beta + \sigma \cdot (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{H}\epsilon \tag{101}$$

$$\hat{\sigma}^2 = \frac{1}{N-p-1}\left\|(\mathbf{I}-\mathbf{H})\epsilon\right\|^2 \tag{102}$$

Now because $\mathbf{H}\epsilon$ and $(\mathbf{I}-\mathbf{H})\epsilon$ are independent, $\hat{\beta}$ and $\hat{\sigma}^2$ are also independent. Moreover, $\hat{\beta}$ has a covariance $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, $\hat{\beta}_j$ has a variance $\sigma^2 \cdot v_j$, with $v_j = [(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}$. So under the null hypothesis,

$$\frac{\beta_j}{\sigma\sqrt{v_j}} \sim N(0,1) \tag{103}$$

Also,

$$\frac{\hat{\sigma}}{\sigma} \sim \sqrt{\frac{\chi^2_{N-p-1}}{N-p-1}} \tag{104}$$

According to (149), the following has a Student's t-distribution with $N-p-1$ degrees of freedom

$$\frac{\frac{\beta_j}{\sigma\sqrt{v_j}}}{\frac{\hat{\sigma}}{\sigma}} = \frac{\beta_j}{\hat{\sigma}\sqrt{v_j}} = z_j \tag{105}$$

F statistic. According to (3.13) on page 48, we form the following statistic to decide whether we can drop groups of coefficients simultaneously.

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(p_1 - p_0)}{\mathrm{RSS}_1/(N - p_1 - 1)} \tag{106}$$

I will show that under the null-hypothesis, $\mathrm{RSS}_0 - \mathrm{RSS}_1$ is chi-squared with $p_1 - p_0$ degrees of freedom, RSS_1 is chi-squared with $N - p_1 - 1$ degrees of freedom, and are independent. So according to appendix C.2, F has indeed an F-distribution with $(p_1 - p_0), (N - p_1 - 1)$ parameters.

Let X be the $N \times (p_0 + 1)$ data-matrix, while $X_1 = [X|X']$ the extended $N \times (p_1 + 1)$ data-matrix. We assume that the smaller model is true, i.e., $y = X\beta + \epsilon$. We have two different estimates for y. $\hat{y}_0 = H_0 y$ comes from the smaller model, while $\hat{y}_1 = H_1 y$ comes from the bigger model. H_0 and H_1 are projections. H_1 projects onto the column space of X_1, which we denote by W_1. H_0 projects onto the column space of X, which we denote by W_0, this is actually a subspace of W_1. Let W_2 be a subspace in W_1 that is orthogonal to W_0. $H_1 - H_0$ projects onto this subspace. Now calculate the residual sum of squares.

$$\begin{aligned}
\mathrm{RSS}_0 &= \|y - \hat{y}_0\|^2 \\
&= \|y - H_0 y\|^2 \\
&= \|(I - H_0)y\|^2 \\
&= \|(I - H_0)(X\beta + \epsilon)\|^2 \\
&= \|(I - H_0)X\beta + (I - H_0)\epsilon\|^2 \\
&= \|0\beta + (I - H_0)\epsilon\|^2 \\
&= \|(I - H_0)\epsilon\|^2 \\
&= \epsilon^T (I - H_0)^T (I - H_0)\epsilon \\
&= \epsilon^T (I - H_0)(I - H_0)\epsilon \\
&= \epsilon^T (I - H_0)\epsilon
\end{aligned} \tag{107}$$

Similarly,

$$\mathrm{RSS}_1 = \epsilon^T (I - H_1)\epsilon \tag{108}$$

Note that here we used the fact that the columns of X make up the subspace W_0. $I - H_0$ projects onto $W_0^{\perp}$, so $(I - H_0)X = 0$. Similarly, $(I - H_1)X = 0$.

From these, we can calculate the difference of the residual sum of squares.

$$\begin{aligned}
\mathrm{RSS}_0 - \mathrm{RSS}_1 &= \epsilon^T (H_1 - H_0)\epsilon \\
&= \|(H_1 - H_0)\epsilon\|^2
\end{aligned} \tag{109}$$

So $\mathrm{RSS}_0 - \mathrm{RSS}_1$ is a chi-squared random variable with $p_1 - p_0$ degrees of freedom (the dimension of W_2). And RSS_1 is also chi-squared with $N - p_1 - 1$ degrees of freedom (the dimension of $W_1^{\perp}$). $\mathrm{RSS}_0 - \mathrm{RSS}_1$ and RSS_1 are independent, because $I - H_1$ and $H_1 - H_0$ project onto perpendicular subspaces.

3.2 Equation (3.28) on page 54

I'd like to confirm that in general,

$$\hat{\beta}_j = \frac{z_j^T y}{z_j^T z_j} \tag{110}$$

But first some notations and clarifications. $y, z_j \in \mathbb{R}^N$. $x_j \in \mathbb{R}^N$ is the jth column vector of X, the data matrix. W_X is the column space of X, $W_{X(j)}$ is the subspace spanned by all the columns of X except the jth column. $W_{X(j)}^{\perp}$ is a one-dimensional subspace that is orthogonal to $W_{X(j)}$, and

$$W_{X(j)} + W_{X(j)}^{\perp} = W_X \tag{111}$$

$z_j = x_j - P_j x_j$, where P_j projects onto $W_{X(j)}$. We can also express it as

$$z_j = P_j^{\perp} x_j \tag{112}$$

where $P_j^{\perp}$ projects onto $W_{X(j)}^{\perp}$. We know, that

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{113}$$

Denoting the jth column vector of $(X^T X)^{-1}$ by b_j ($(X^T X)^{-1}$ is symmetric, so b_j^T is the jth row vector), we can write:

$$\hat{\beta}_j = b_j^T X^T y \tag{114}$$

From appendix (D.3) we can construct $P_j^{\perp}$:

$$P_j^{\perp} = \frac{X b_j \cdot b_j^T X^T}{v_j} \tag{115}$$

Where v_j is the jth element of b_j ($= [(X^T X)^{-1}]_{j,j}$). With this we can calculate z_j ($A_{:,j}$ denotes the jth column vector of matrix A):

$$\begin{aligned} z_j =& P_j^{\perp} x_j = (P_j^{\perp} X)_{:,j} = \left(\frac{X b_j \cdot b_j^T X^T}{v_j} X \right)_{:,j} \\ =& \left(\frac{X b_j}{v_j} b_j^T X^T X \right)_{:,j} = \left(\frac{X b_j}{v_j} \delta_j^T \right)_{:,j} = \frac{X b_j}{v_j} \end{aligned} \tag{116}$$

Here $\delta_j = I_{:,j}$, the jth column vector of the identity. We can plug this result into (110):

$$\hat{\beta}_j = \frac{\frac{b_j^T X^T}{v_j} y}{\frac{b_j^T X^T X b_j}{v_j^2}} = \frac{b_j^T X^T y}{\frac{b_j^T \delta_j}{v_j}} = \frac{b_j^T X^T y}{\frac{v_j}{v_j}} = b_j^T X^T y \tag{117}$$

It is indeed the same as we got in (114), so the proof is complete.

3.3 Solutions for the Exercises of chapter 3

3.3.1 Ex. 3.1

According to Appendix (C.2), the F-statistics can be written in the form:

$$F \sim \frac{\chi^2_{d_1}/d_1}{\chi^2_{d_2}/d_2} \tag{118}$$

where in our case $d_1 = p_1 - p_0$, $d_2 = N - p_1 - 1$. Dropping a single coefficient means that $p_1 = p_0 + 1 \rightarrow p_1 - p_0 = 1$, so

$$F \sim \frac{\chi^2_1/1}{\chi^2_{d_2}/d_2} \sim \frac{N(0,1)^2}{\chi^2_{d_2}/d_2} \sim \left(\frac{N(0,1)}{\sqrt{\frac{\chi^2_{N-p_1-1}}{(N-p_1-1)}}} \right)^2 \sim t^2_{N-p_1-1} \tag{119}$$

The Z-score is t-distributed with $N - p - 1$ parameters, so the F-statistics for dropping a single coefficient is indeed *distributed* as the square of the corresponding Z-score. Well, this doesn't prove that the square of the calculated Z is equal to the calculated F. So let's prove it. Without loss of generality we can assume that we test for the last coefficient, $j = p + 1$.

$$z_j^2 = \frac{\hat{\beta}_j^{\,2}}{\hat{\sigma}^2 v_j} \tag{120}$$

We have to show that this equals to the following F:

$$F = \frac{(\text{RSS}_0 - \text{RSS}_1)/(p_1 - p_0)}{\text{RSS}_1/(N - p_1 - 1)} \tag{121}$$

Now since $(N - p_1 - 1)\hat{\sigma}^2 = \text{RSS}_1$, and $p_1 - p_0 = 1$, we have to show that

$$\frac{\hat{\beta}_j^{\ 2}}{v_j} = \text{RSS}_0 - \text{RSS}_1 \tag{122}$$

Denote the jth column (=jth row) of $(X^TX)^{-1}$ as b_j. With this notation

$$\hat{\beta}_j = b_j^T X^T y = y^T X b_j \tag{123}$$

$$\frac{\hat{\beta}_j^{\ 2}}{v_j} = y^T \frac{Xb_j \cdot b_j^T X^T}{v_j} y \tag{124}$$

$$\text{RSS}_0 - \text{RSS}_1 = y^T(H_1 - H_0)y \tag{125}$$

Where H_1 projects onto the column space of X, H_0 projects onto W_0. W_0 is the column space of the matrix same as X but dropping the last column. According to Appendix (D.3):

$$H_1 - H_0 = \frac{Xb_j \cdot b_j^T X^T}{v_j} \tag{126}$$

which concludes the proof.

3.3.2 Ex. 3.2

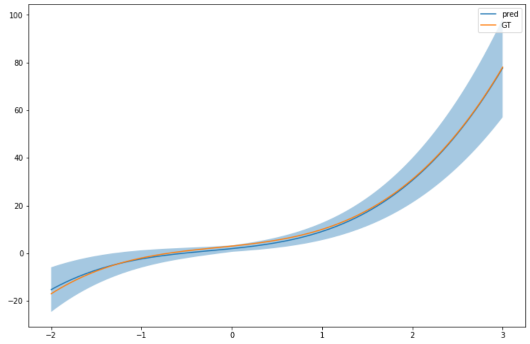

Figure 7: True function, predicted function, with 95% confidence band (pointwise).

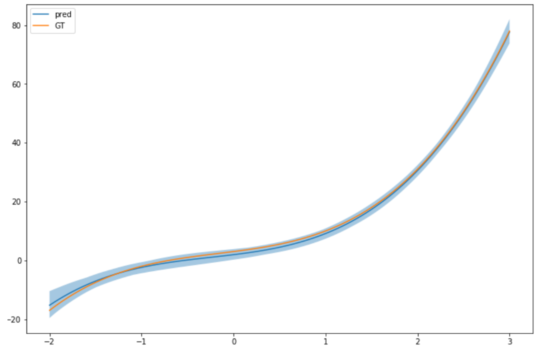

Figure 8: True function, predicted function, with 95% confidence band (from multivariate normal).

3.3.3 Ex. 3.3

a. We formulate an estimate of $a^T\beta$ as c^Ty. For the least squares estimate we have that

$$a^T\hat{\beta} = a^T(X^TX)^{-1}X^Ty \tag{127}$$

From this,

$$c_0 = X(X^TX)^{-1}a \tag{128}$$

Any c can be written as

$$c = c_0 + c_1 = X(X^TX)^{-1}a + c_1 \tag{129}$$

The constraint is that $\mathrm{E}(c^Ty) = a^T\beta$.

$$\mathrm{E}(c^Ty) = \mathrm{E}(c_0^Ty) + \mathrm{E}(c_1^Ty) = a^T\beta \tag{130}$$

Since $\mathrm{E}(c_0^Ty) = a^T\beta$, we have that $\mathrm{E}(c_1^T\beta) = 0$.

$$0 = \mathrm{E}(c_1^Ty) = c_1^TX\beta \tag{131}$$

Because β is unobservable, we conclude that

$$0 = c_1^TX \tag{132}$$

which means

$$c_1^T c_0 = c_1^T X(X^T X)^{-1} a = 0 \tag{133}$$

Now consider the variances.

$$\mathrm{Var}(a^T \hat{\beta}) = a^T \mathrm{Var}(\hat{\beta}) a = \sigma^2 a^T (X^T X)^{-1} a \tag{134}$$

Calculating the variance of a general unbiased estimate, using (133):

$$\begin{aligned}
\mathrm{Var}(c^T y) =&\sigma^2 c^T c \\
=&\sigma^2 (c_0^T + c_1^T)(c_0 + c_1) = \sigma^2 (c_0^T c_0 + 0 + 0 + c_1^T c_1) \\
=&\sigma^2 c_0^T c_0 + \sigma^2 c_1^T c_1 \\
=&\mathrm{Var}(a^T \hat{\beta}) + \sigma^2 c_1^T c_1
\end{aligned} \tag{135}$$

Since $\sigma^2 c_1^T c_1 \geq 0$, we conclude that

$$\mathrm{Var}(c^T y) \geq \mathrm{Var}(a^T \hat{\beta}) \tag{136}$$

b. The solution is basically the same as for the previous one. Here we will use the fact that $A^T A$ is a positive semidefinite matrix for any matrix A. A linear unbiased estimate for β can be expressed as $\widetilde{\beta} = C^T y$, where C is a $N \times (p+1)$ matrix. We can express C as $C = X(X^T X)^{-1} + C_1 = C_0 + C_1$. The estimates are unbiased, so $\mathrm{E}(C^T y) = \beta$. From this we have

$$\mathrm{E}(C^T y) = (C_0 + C_1)^T (X\beta) = \beta + C_1^T X\beta = \beta \rightarrow C_1^T X\beta = 0 \tag{137}$$

Because β is unobservable, we have that

$$C_1^T X = 0 \tag{138}$$

Now consider the variances. $\hat{V} \equiv \mathrm{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$. $\tilde{V} \equiv \mathrm{Var}(\tilde{\beta}) = \mathrm{Var}(C^T y) = \sigma^2 C^T C = \sigma^2 (C_0 + C_1)^T (C_0 + C_1)$. Using (138), we can write that $\tilde{V} = \sigma^2 C_0^T C_0 + \sigma^2 C_1^T C_1 = \hat{V} + \sigma^2 C_1^T C_1$. From this: $\tilde{V} - \hat{V} = \sigma^2 C_1^T C_1$ which is a positive semidefinite matrix. This concludes the proof.

3.3.4 Ex. 3.5

The ridge regression loss in vectorized form:

$$L_{\text{ridge}} = (\mathbf{y} - \beta_0\mathbf{e} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \beta_0\mathbf{e} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$$

$\mathbf{e}$ is a vector of all ones in $\mathbb{R}^N$. $\mathbf{X}$ is $N \times p$ matrix, so it doesn't have $\mathbf{e}$ as its first column. That's why we have a separate β_0, which is a scalar. All other coefficients are in the vector $\boldsymbol{\beta}$.

How to formulate L^c_{ridge}? We can use L_{ridge}, but instead of $\mathbf{X}$, we write $\mathbf{X} - \frac{1}{N}\mathbf{e}\mathbf{e}^T\mathbf{X}$. It's because $\frac{1}{N}\mathbf{e}^T\mathbf{X} = [\overline{x}_1, \overline{x}_2, \dots, \overline{x}_p]$. Let's create the following notation:

$$\mathbf{A} \equiv \frac{\mathbf{e}\mathbf{e}^T}{\mathbf{e}^T\mathbf{e}} = \frac{\mathbf{e}\mathbf{e}^T}{N}$$

This is a projection onto the line defined by $\mathbf{e}$. So to get L^c_{ridge}, we replace $\mathbf{X}$ in L_{ridge} with $(\mathbf{I} - \mathbf{A})\mathbf{X}$. So we only need to solve L_{ridge} for β_0 and $\boldsymbol{\beta}$. Once we have these, we get β_0^c, $\boldsymbol{\beta}^c$ by applying the change $\mathbf{X} \to (\mathbf{I} - \mathbf{A})\mathbf{X}$.

$$\begin{aligned}\frac{\partial L_{\text{ridge}}}{\partial \beta_0} =& -2\mathbf{e}^T(\mathbf{y} - \beta_0\mathbf{e} - \mathbf{X}\boldsymbol{\beta}) = 0 \\ \to& \beta_0 = \overline{y} - \frac{1}{N}\mathbf{e}^T\mathbf{X}\boldsymbol{\beta}\end{aligned}$$

Note that $\mathbf{e}^T\mathbf{e} = N$, and $\mathbf{e}^T\mathbf{y} = N\overline{y}$. Now we can get β_0^c by the above mentioned substitution. Note that $\mathbf{e}^T\mathbf{A} = \frac{1}{N}\mathbf{e}^T\mathbf{e}\mathbf{e}^T = \frac{1}{N}N\mathbf{e}^T = \mathbf{e}^T$.

$$\begin{aligned}\beta_0^c =& \overline{y} - \frac{1}{N}\mathbf{e}^T(\mathbf{I} - \mathbf{A})\mathbf{X}\boldsymbol{\beta}^{\mathbf{c}} \\ =& \overline{y} - \frac{1}{N}(\mathbf{e}^T\mathbf{I} - \mathbf{e}^T\mathbf{A})\mathbf{X}\boldsymbol{\beta}^{\mathbf{c}} \\ =& \overline{y} - \frac{1}{N}(\mathbf{e}^T - \mathbf{e}^T)\mathbf{X}\boldsymbol{\beta}^{\mathbf{c}} \\ =& \overline{y}\end{aligned}$$

So we see that β_0^c does not depend on the features and coefficients, as β_0 does. Now let's move on to $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^c$.

$$\frac{\partial L_{\text{ridge}}}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T(\mathbf{y} - \beta_0\mathbf{e} - \mathbf{X}\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta}$$

Note that

$$\beta_0\mathbf{e} = \mathbf{e}\overline{y} - \frac{1}{N}\mathbf{e}\mathbf{e}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{A}\mathbf{y} - \mathbf{A}\mathbf{X}\boldsymbol{\beta}$$

So we have that

$$\begin{aligned}\frac{\partial L_{\text{ridge}}}{\partial \boldsymbol{\beta}} =& - 2\mathbf{X}^T(\mathbf{y} - \mathbf{A}\mathbf{y} + \mathbf{A}\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta} = 0\\ \rightarrow& \boldsymbol{\beta} = (\mathbf{X}^T(\mathbf{I} - \mathbf{A})\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T(\mathbf{I} - \mathbf{A})\mathbf{y}\end{aligned}$$

From this we can get $\boldsymbol{\beta}^c$ (by the substitution $\mathbf{X} \rightarrow (\mathbf{I} - \mathbf{A})\mathbf{X}$). Note that $\mathbf{I} - \mathbf{A}$ is a projection, so its square is itself, as well as its transpose.

$$\begin{aligned}\boldsymbol{\beta}^c =& (((\mathbf{I} - \mathbf{A})\mathbf{X})^T(\mathbf{I} - \mathbf{A})(\mathbf{I} - \mathbf{A})\mathbf{X} + \lambda\mathbf{I})^{-1}((\mathbf{I} - \mathbf{A})\mathbf{X})^T(\mathbf{I} - \mathbf{A})\mathbf{y}\\ =& (\mathbf{X}^T(\mathbf{I} - \mathbf{A})(\mathbf{I} - \mathbf{A})(\mathbf{I} - \mathbf{A})\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T(\mathbf{I} - \mathbf{A})(\mathbf{I} - \mathbf{A})\mathbf{y}\\ =& (\mathbf{X}^T(\mathbf{I} - \mathbf{A})\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T(\mathbf{I} - \mathbf{A})\mathbf{y}\\ =& \boldsymbol{\beta}\end{aligned}$$

So the conclusion is that by centering the inputs, we get the same result for $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^c$, while β_0 gets replaces by $\beta_0^c = \overline{y}$.

3.3.5 Ex. 3.9 *Forward stepwise regression*

Actually we don't need $\mathbf{Q}$, we only need $\mathbf{H} = \mathbf{Q}\mathbf{Q}^T$, and updating this matrix. Some notations first. W_1 is the column space of $\mathbf{X}_1$. $\mathbf{u}$ is an arbitrary column vector of $\mathbf{X}_2$. $\mathbf{X}_{1u}$ is the matrix $[\mathbf{X_1}|\mathbf{u}]$. W_{1u} is the column space of $\mathbf{X}_{1u}$. $\mathbf{H}_1$ is the projection matrix that projects onto W_1. $\mathbf{H}_{1u}$ projects onto W_{1u}.

We know the residual:

$$\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}_1\mathbf{y}$$

Now the new residual:

$$\mathbf{r}' = \mathbf{y} - \hat{\mathbf{y}}' = \mathbf{y} - \mathbf{H}_{1u}\mathbf{y}$$

Let's calculate $\mathbf{H}_{1u}$ efficiently. It projects onto W_{1u}. We have an orthonormal base in W_1, namely, the columns of $\mathbf{Q}$. We need an orthonormal base in W_{1u} as well. So lets regress $\mathbf{u}$ onto W_1. $\mathbf{u}_p = \mathbf{H}_1\mathbf{u}$ is the projection of $\mathbf{u}$ onto W_1, so $\mathbf{u} - \mathbf{H}_1\mathbf{u}$ is orthogonal to W_1. Let's normalize this vector and append it to $\mathbf{Q}$.

$$\mathbf{q} = \frac{\mathbf{u} - \mathbf{H}_1\mathbf{u}}{||\mathbf{u} - \mathbf{H}_1\mathbf{u}||}$$

With this we can construct $\mathbf{Q}' = [\mathbf{Q}|\mathbf{q}]$, and the columns of this matrix are orthonormal. The projection matrix:

$$\mathbf{H}_{1u} = \mathbf{Q}'\mathbf{Q}'^T = \mathbf{Q}\mathbf{Q}^T + \mathbf{q}\mathbf{q}^T = \mathbf{H}_1 + \mathbf{q}\mathbf{q}^T$$

The new residual:

$$\mathbf{r}' = \mathbf{y} - \mathbf{H}_{1u}\mathbf{y} = \mathbf{y} - \mathbf{H}_1\mathbf{y} - \mathbf{q}\mathbf{q}^T\mathbf{y} = \mathbf{r} - \mathbf{q}\mathbf{q}^T\mathbf{y}$$

From this:

$$\text{RSS}' = \text{RSS} - 2\mathbf{r}^T\mathbf{q}\mathbf{q}^T\mathbf{y} + \mathbf{y}^T\mathbf{q}\mathbf{q}^T\mathbf{y}$$

The drop in RSS:

$$\text{RSS} - \text{RSS}' = (2\mathbf{r} - \mathbf{y})^T\mathbf{q}\mathbf{q}^T\mathbf{y}$$

So we iterate through the columns of $\mathbf{X}_2$ and seek for the largest drop in RSS. Once we have the best column ($\mathbf{u}^*$), we pass $\mathbf{H}_{1u*}$ to the next iteration. Consider the notebook "forward_stepwise.ipynb" where I have implemented this algorithm.

3.3.6 Ex. 3.10 Backward stepwise regression

Now we seek for a column vector $\mathbf{x}$ that we can drop from $\mathbf{X}$. Notations. $\mathbf{x}_i$ is the ith column vector of $\mathbf{X}$. $\mathbf{X}_i$ is the matrix obtained from $\mathbf{X}$ by dropping the ith column. $\mathbf{B} \equiv (\mathbf{X}^T\mathbf{X})^{-1}$. The ith column vector of $\mathbf{B}$ is $\mathbf{b}_i$.

Now, $\mathbf{H} = \mathbf{X}\mathbf{B}\mathbf{X}^T$ projects onto the column space of $\mathbf{X}$. According to Appendix (D.3), $\mathbf{H}_i = \mathbf{H} - \mathbf{P}_i$ projects onto the column space of $\mathbf{X}_i$, where

$$\mathbf{P}_i = \frac{\mathbf{X}\mathbf{b}_i\mathbf{b}_i^T\mathbf{X}^T}{v_i}$$

and v_i is the ith element in the diagonal of $\mathbf{B}$. The original residual:

$$\mathbf{r} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

The new residual (after dropping the ith column in $\mathbf{X}$:

$$\mathbf{r}_i = (\mathbf{I} - \mathbf{H}_i)\mathbf{y} = \mathbf{r} + \mathbf{P}_i\mathbf{y}$$

From this we get the increment in RSS introduced by dropping the feature:

$$\Delta\text{RSS}_i = (2\mathbf{r} + \mathbf{y})^T\mathbf{P}_i\mathbf{y}$$

We seek for a column vector of $\mathbf{X}$ for which ΔRSS_i is minimal. We drop that feature. Consider the notebook "backward_stepwise.ipynb" where I implement this algorithm.

3.3.7 Ex. 3.11

$$\begin{aligned}\text{RSS}(\mathbf{B}) =& \sum_{i=1}^{N}(y_i - f(x_i))^T\mathbf{\Sigma}^{-1}(y_i - f(x_i))\\ =& \text{Tr}\left((\mathbf{Y} - \mathbf{XB})\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{XB})^T\right)\\ =& \text{Tr}(\mathbf{Y\Sigma}^{-1}\mathbf{Y}^T) - \text{Tr}(\mathbf{Y\Sigma}^{-1}\mathbf{B}^T\mathbf{X}^T) - \text{Tr}(\mathbf{XB\Sigma}^{-1}\mathbf{Y}^T) + \text{Tr}(\mathbf{XB\Sigma}^{-1}\mathbf{B}^T\mathbf{X}^T)\end{aligned}$$

Calculating the derivatives:

$$\frac{d}{d\mathbf{B}}\text{Tr}(\mathbf{Y\Sigma}^{-1}\mathbf{Y}^T) = \mathbf{0}$$

$$\frac{d}{d\mathbf{B}}\text{Tr}(\mathbf{Y\Sigma}^{-1}\mathbf{B}^T\mathbf{X}^T) = \frac{d}{d\mathbf{B}}\text{Tr}(\mathbf{X}^T\mathbf{Y\Sigma}^{-1}\mathbf{B}^T) = \mathbf{X}^T\mathbf{Y\Sigma}^{-1}$$

$$\frac{d}{d\mathbf{B}}\text{Tr}(\mathbf{XB\Sigma}^{-1}\mathbf{Y}^T) = \frac{d}{d\mathbf{B}}\text{Tr}(\mathbf{\Sigma}^{-1}\mathbf{Y}^T\mathbf{XB}) = (\mathbf{\Sigma}^{-1}\mathbf{Y}^T\mathbf{X})^T = \mathbf{X}^T\mathbf{Y\Sigma}^{-1}$$

$$\begin{aligned}\frac{d}{d\mathbf{B}}\text{Tr}(\mathbf{XB\Sigma}^{-1}\mathbf{B}^T\mathbf{X}^T) =& \frac{d}{d\mathbf{A}}\text{Tr}(\mathbf{XA\Sigma}^{-1}\mathbf{B}^T\mathbf{X}^T) + \frac{d}{d\mathbf{A}}\text{Tr}(\mathbf{XB\Sigma}^{-1}\mathbf{A}^T\mathbf{X}^T)\\ =& (\mathbf{\Sigma}^{-1}\mathbf{B}^T\mathbf{X}^T\mathbf{X})^T + \mathbf{X}^T\mathbf{XB\Sigma}^{-1}\\ =& 2\mathbf{X}^T\mathbf{XB\Sigma}^{-1}\end{aligned}$$

Now we can substitute:

$$\frac{d}{d\mathbf{B}}\text{RSS}(\mathbf{B}) = -2\mathbf{X}^T\mathbf{Y\Sigma}^{-1} + 2\mathbf{X}^T\mathbf{XB\Sigma}^{-1}$$

Setting this to zero, we get:

$$\mathbf{B} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

What happens if the covariance matrices $\mathbf{\Sigma_i}$ are different for each observation? The Residual Sum of Squares:

$$\mathrm{RSS}(\mathbf{B}) = \sum_{i=1}^{N}(y_i - \mathbf{B}^T x_i)^T \mathbf{\Sigma}_i^{-1}(y_i - \mathbf{B}^T x_i)$$

Derivating this w.r.t $\mathbf{B}$, and setting to zero, we get:

$$\sum_{i=1}^{N} x_i x_i^T \mathbf{B}\mathbf{\Sigma}_i^{-1} = \sum_{i=1}^{N} x_i y_i^T \mathbf{\Sigma}_i^{-1}$$

Appendices

Here I collected the useful mathematical knowledge required to understand some proofs.

A differentiation

A.1 differentiation w.r.t. a vector

1. Let $\mathbf{a} \in \mathbb{R}^n$ be a constant vector, $\mathbf{x} \in \mathbb{R}^n$. Then

$$\frac{d}{d\mathbf{x}}(\mathbf{x}^T\mathbf{a}) = \frac{d}{d\mathbf{x}}(\mathbf{a}^T\mathbf{x}) = \mathbf{a} \tag{139}$$

2. Let $\mathbf{A} \in \mathbb{R}^{nxn}$ be a constant matrix, $\mathbf{x} \in \mathbb{R}^n$ Then

$$\frac{d}{d\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = (\mathbf{A} + \mathbf{A}^T)\mathbf{x} \tag{140}$$

We can derive this as follows:

$$\begin{aligned}
\frac{d}{d\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) =& \frac{d}{d\mathbf{y}}(\mathbf{y}^T\mathbf{A}\mathbf{x}) + \frac{d}{d\mathbf{y}}(\mathbf{x}^T\mathbf{A}\mathbf{y}) \\
=& \frac{d}{d\mathbf{y}}(\mathbf{y}^T\mathbf{A}\mathbf{x}) + \frac{d}{d\mathbf{y}}(\mathbf{y}^T\mathbf{A}^T\mathbf{x}) \\
=& \mathbf{A}\mathbf{x} + \mathbf{A}^T\mathbf{x}
\end{aligned}$$

A.2 differentiation w.r.t. a matrix

1. Let $\mathbf{A} \in \mathbb{R}^{n\times m}$, $\mathbf{X} \in \mathbb{R}^{m\times n}$. Then

$$\frac{d}{d\mathbf{X}}\mathrm{Tr}(\mathbf{A}\mathbf{X}) = \frac{d}{d\mathbf{X}}\mathrm{Tr}(\mathbf{X}\mathbf{A}) = \mathbf{A}^T \tag{141}$$

2. Let $\mathbf{A} \in \mathbb{R}^{n\times m}$, $\mathbf{X} \in \mathbb{R}^{n\times m}$. Then

$$\frac{d}{d\mathbf{X}}\mathrm{Tr}(\mathbf{A}\mathbf{X}^{\mathbf{T}}) = \frac{d}{d\mathbf{X}}\mathrm{Tr}(\mathbf{X}^{\mathbf{T}}\mathbf{A}) = \mathbf{A} \tag{142}$$

3. Let $\mathbf{A} \in \mathbb{R}^{m \times m}$, $\mathbf{X} \in \mathbb{R}^{m \times n}$. Then

$$\frac{d}{d\mathbf{X}} \mathrm{Tr}(\mathbf{X}^T \mathbf{A} \mathbf{X}) = (\mathbf{A} + \mathbf{A}^T)\mathbf{X} \tag{143}$$

We can derive it as follows:

$$\begin{aligned}\frac{d}{d\mathbf{X}} \mathrm{Tr}(\mathbf{X}^T \mathbf{A} \mathbf{X}) =& \frac{d}{d\mathbf{Y}} \mathrm{Tr}(\mathbf{Y}^T \mathbf{A} \mathbf{X}) + \frac{d}{d\mathbf{Y}} \mathrm{Tr}(\mathbf{X}^T \mathbf{A} \mathbf{Y})\\ =& \mathbf{A}\mathbf{X} + (\mathbf{X}^T \mathbf{A})^T\\ =& \mathbf{A}\mathbf{X} + \mathbf{A}^T \mathbf{X} = (\mathbf{A} + \mathbf{A}^T)\mathbf{X}\end{aligned}$$

Example. Consider now this example.

$$f(\mathbf{X}) = \mathrm{Tr}(\mathbf{X}^T \mathbf{A} \mathbf{X} \mathbf{B} \mathbf{X}^T \mathbf{C})$$

where $\mathbf{X} \in \mathbb{R}^{n \times m}$, $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{m \times m}$, $\mathbf{C} \in \mathbb{R}^{n \times m}$.

$$\begin{aligned}\frac{d}{d\mathbf{X}} f(\mathbf{X}) =& \frac{d}{d\mathbf{Y}} \mathrm{Tr}(\mathbf{Y}^T \mathbf{A} \mathbf{X} \mathbf{B} \mathbf{X}^T \mathbf{C})\\ &+ \frac{d}{d\mathbf{Y}} \mathrm{Tr}(\mathbf{X}^T \mathbf{A} \mathbf{Y} \mathbf{B} \mathbf{X}^T \mathbf{C})\\ &+ \frac{d}{d\mathbf{Y}} \mathrm{Tr}(\mathbf{X}^T \mathbf{A} \mathbf{X} \mathbf{B} \mathbf{Y}^T \mathbf{C})\end{aligned}$$

Calculating these:

$$\frac{d}{d\mathbf{Y}} \mathrm{Tr}(\mathbf{Y}^T \mathbf{A} \mathbf{X} \mathbf{B} \mathbf{X}^T \mathbf{C}) = \mathbf{A} \mathbf{X} \mathbf{B} \mathbf{X}^T \mathbf{C}$$

$$\begin{aligned}\frac{d}{d\mathbf{Y}} \mathrm{Tr}(\mathbf{X}^T \mathbf{A} \mathbf{Y} \mathbf{B} \mathbf{X}^T \mathbf{C}) =& \frac{d}{d\mathbf{Y}} \mathrm{Tr}(\mathbf{Y} \mathbf{B} \mathbf{X}^T \mathbf{C} \mathbf{X}^T \mathbf{A})\\ =& (\mathbf{B} \mathbf{X}^T \mathbf{C} \mathbf{X}^T \mathbf{A})^T\\ =& \mathbf{A}^T \mathbf{X} \mathbf{C}^T \mathbf{X} \mathbf{B}^T\end{aligned}$$

$$\frac{d}{d\mathbf{Y}} \mathrm{Tr}(\mathbf{X}^T \mathbf{A} \mathbf{X} \mathbf{B} \mathbf{Y}^T \mathbf{C}) = \frac{d}{d\mathbf{Y}} \mathrm{Tr}(\mathbf{Y}^T \mathbf{C} \mathbf{X}^T \mathbf{A} \mathbf{X} \mathbf{B}) = \mathbf{C} \mathbf{X}^T \mathbf{A} \mathbf{X} \mathbf{B}$$

So the result is:

$$\frac{d}{d\mathbf{X}} \mathrm{Tr}(\mathbf{X}^T \mathbf{A} \mathbf{X} \mathbf{B} \mathbf{X}^T \mathbf{C}) = \mathbf{A} \mathbf{X} \mathbf{B} \mathbf{X}^T \mathbf{C} + \mathbf{A}^T \mathbf{X} \mathbf{C}^T \mathbf{X} \mathbf{B}^T + \mathbf{C} \mathbf{X}^T \mathbf{A} \mathbf{X} \mathbf{B}$$

B variance and covariance properties

B.1 scalar multiple

We know that if $a \in \mathbb{R}$ is a constant scalar, and X is a scalar random variable, then

$$\mathrm{Var}(aX) = a^2 \cdot \mathrm{Var}(X) \tag{144}$$

B.2 vector multiple

Now what if $a \in \mathbb{R}^p$ is a (constant) vector, and $X \in \mathbb{R}^p$ is a random vector, and we take the inner product $a^T \cdot X$? What is the variance $\mathrm{Var}(a^T \cdot X)$?

$$\begin{aligned}
\mathrm{Var}(a^T \cdot X) =& \mathrm{Var}(a_1 \cdot X_1 + a_2 \cdot X_2 + \cdots + a_n \cdot X_n) \\
=& \mathrm{E}\left(\sum_i a_i \cdot (X_i - \mathrm{E}X_i)\right)^2 \\
=& \mathrm{E}\left(\sum_{i,j} a_i \cdot (X_i - \mathrm{E}X_i) \cdot a_j \cdot (X_j - \mathrm{E}X_j)\right) \\
=& \sum_{i,j} a_i \cdot a_j \cdot \mathrm{E}\left((X_i - \mathrm{E}X_i) \cdot (X_j - \mathrm{E}X_j)\right) \\
=& \sum_{i,j} a_i \cdot a_j \cdot \mathrm{Cov}(X_i, X_j) \\
=& a^T \cdot \mathrm{Cov}(X, X) \cdot a = a^T \cdot \Sigma \cdot a
\end{aligned} \tag{145}$$

Here $\Sigma \equiv \mathrm{Cov}(X, X)$ is the covariance matrix, $\mathrm{Cov}(X, X)_{i,j} = \mathrm{Cov}(X_i, X_j)$. Let's state our finding again. $a \in \mathbb{R}^p$ is a constant vector, $X \in \mathbb{R}^p$ is a random vector, then:

$$\mathrm{Var}(a^T \cdot X) = a^T \cdot \mathrm{Cov}(X, X) \cdot a \tag{146}$$

Furthermore, we can write for the covariance matrix:

$$\mathrm{Cov}(X, X) = \mathrm{E}((X - \mathrm{E}X) \cdot (X - \mathrm{E}X)^T) = \mathrm{E}(X \cdot X^T) - (\mathrm{E}X)(\mathrm{E}X^T) \tag{147}$$

B.3 matrix multiple

Let $A \in \mathbb{R}^{nxp}$ a constant matrix, $X \in \mathbb{R}^p$ a random vector. The covariance matrix:

$$\begin{aligned} \mathrm{Cov}(AX) =& \mathrm{E}(AXX^TA^T) - \mathrm{E}(AX)\mathrm{E}(X^TA^T) \\ =& A \cdot \mathrm{E}(XX^T) \cdot A^T - A \cdot \mathrm{E}(X)\mathrm{E}(X^T) \cdot A^T \\ =& A \cdot \big(\mathrm{E}(XX^T) - \mathrm{E}(X)\mathrm{E}(X^T)\big) \cdot A^T \\ =& A \cdot \mathrm{Cov}(X) \cdot A^T \end{aligned} \tag{148}$$

C distributions

C.1 Student's t-distribution

The t-distribution with ν degrees of freedom can be expressed as

$$T = \frac{Z}{\sqrt{V/\nu}} \tag{149}$$

where

- $Z \sim N(0,1)$
- $V \sim \chi^2_\nu$
- Z and V are independent

C.2 F-distribution

A random variate of the F-distribution with parameters d_1 and d_2 arises as the ratio of two appropriately scaled chi-squared variates:

$$X = \frac{U_1/d_1}{U_2/d_2} \tag{150}$$

where

- U_1 and U_2 have chi-squared distributions with d_1 and d_2 degrees of freedom respectively, and
- U_1 and U_2 are independent.

D projections

D.1 sum of projections

Let V be a vector space, $W_1 \subset V$ a subspace, $W_2 \subset V$ a subspace such that $W_1 \perp W_2$.

Let P_1 be an orthogonal projection onto the subspace W_1, P_2 be an orthogonal projection onto the subspace W_2. I claim that $P_1{+}P_2$ is an orthogonal projection onto $W_1 + W_2$.

Proof. Denote $W_\perp$ the orthogonal complement of $W_1 + W_2$.

$$(W_1 + W_2) + W_\perp = V \tag{151}$$

and

$$(W_1 + W_2) \perp W_\perp \tag{152}$$

Any vector $v \in V$ can be decomposed as

$$v = w_\perp + w_1 + w_2 \tag{153}$$

where $w_\perp \in W_\perp$, $w_1 \in W_1$, $w_2 \in W_2$. This decomposition is unique.

$$P_1 v = 0 + w_1 + 0 = w_1 \tag{154}$$

$$P_2 v = 0 + 0 + w_2 = w_2 \tag{155}$$

From these

$$(P_1 + P_2)v = w_1 + w_2 \tag{156}$$

So $P_1 + P_2$ projects onto $W_1 + W_2$.

□

D.2 difference of projections

Let V be a vector space, $W_1 \subset V$ a subspace, $W_2 \subset W_1$ a subspace, $W_3 \subset W_1$ a subspace, such that $W_2 \perp W_3$, and $W_2 + W_3 = W_1$.

Let P_1 be an orthogonal projection onto the subspace W_1, P_2 be an orthogonal projection onto the subspace W_2. I claim that $P_1 - P_2$ is an orthogonal projection onto W_3.

Proof. Denote $W_\perp$ the orthogonal complement of W_1: $W_1 + W_\perp = V$, and $W_1 \perp W_\perp$

Any vector $v \in V$ can be decomposed as

$$v = w_\perp + w_2 + w_3 \tag{157}$$

where $w_\perp \in W_\perp$, $w_2 \in W_2$, $w_3 \in W_3$. This decomposition is unique.

$$P_1 v = 0 + w_2 + w_3 = w_2 + w_3 \tag{158}$$

$$P_2 v = 0 + w_2 + 0 = w_2 \tag{159}$$

From these

$$(P_1 - P_2)v = (w_2 + w_3) - w_2 = w_3 \tag{160}$$

So $P_1 - P_2$ projects onto W_3.

□

D.3 The special vector Xb

(I couldn't find any better name for this subsection, sorry for this...) Let's begin with the $N \times p$ matrix X, where we denote the column vectors by x_i. Assume that X has a full column-rank, so $\mathrm{rank}(X) = p$. Denote the subspace $W = span(x_1, x_2, \ldots, x_{p-1})$, which is the subspace generated by all the columns of X, except the last one. Let b be the last column vector of $(X^T X)^{-1}$. I claim that Xb is a vector that is perpendicular to W. Obviously, Xb is in the column space of X. We have that

$$X^T X (X^T X)^{-1} = I \tag{161}$$

Considering the ith row, jth column (q_j being the jth column vector of $(X^T X)^{-1}$, so $q_p = b$)

$$x_i^T X q_j = \delta_{i,j} \tag{162}$$

Choosing $j = p$

$$x_i^T X b = \delta_{i,p} \tag{163}$$

This means that Xb is perpendicular to x_i $(i \neq p)$, which is what I wanted to prove. Now let's project onto the subspace $W_p = \text{span}(Xb)$:

$$P_p = \frac{Xb \cdot b^T X^T}{b^T X^T X b} = \frac{Xb \cdot b^T X^T}{v} \tag{164}$$

where $v \equiv b_p$ is the last element of the vector b, that is, $v \equiv [(X^T X)^{-1}]_{p,p}$.

Now we can create the same projection according to Appendix D.2. Let P_X be the projection onto the column space of X, and P_W the projection onto W:

$$P_X = X(X^T X)^{-1} X^T \tag{165}$$

$$P_W = X_0(X_0^T X_0)^{-1} X_0^T \tag{166}$$

where we get X_0 from X by dropping the last column. Now we see that

$$X(X^T X)^{-1} X^T - X_0(X_0^T X_0)^{-1} X_0^T = \frac{Xb \cdot b^T X^T}{v} \tag{167}$$