Elements of Statistical Learning

book by Hastie, Tibshirani, and Friedmann

Chapter 1

LRS - Important stuff

1.1 Questions

- When he writes f in ch2, is he referring specifically to the regression function $f(x) = E(Y|X = x)$ - or just any function? Is there a difference here, since the regression function, without any details on the joint distribution function is a just any function.
- What is the real content behind the "additive error assumption"?
  - I.e., when he says, assume:
    $$Y = f(x) + \epsilon$$
    where $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$.
  - Is f here the *regression function* $f(x) = E(Y|X = x)$ by assumption? (it seems to be a consequence at the very least)
  - What is the Expectation in $E(\epsilon) = 0$ taken over?
  - What else is hidden in this assumtpion? how general or restrictive is it?
  - How is the joint probability distribution $Pr(X, Y)$ information encoded in here? And should ϵ be considered as another random variable.
- I like the idea of many variables $x_1, \cdots x_v$ with independent prob densities, and Y exactly determined by the vars. But then not measuring many variables, and hiding all their effects on Y in a single ϵ r.v. - so really we have (X, ϵ) as random variables.

  develop this further....
- excercise 2.6 - what does *reduced weighted least squares* mean?

1.2 TODO

- sovle problem 2.5 (after reading first few sections of chapter 3)

Chapter 2

Overview of Supervised Learning

2.0 LRS Summary of Main Ideas

- have observables (X, Y)
- want to predict Y based on observations of X
- given an observation of x, best predictor for Y (where best is defined via squared error loss function) is $f(x) = E(Y|X = x)$
- I assume this conditional expectation is defined using generally unkown joint probability distribution $Pr(X, Y)$
- ultimately we are trying to find a useful approximation for f
- The class of nearest neighbor methods can be viewed as a direct approximation to this conidtional expectation

  suffers from the curse of dimensionality
- talks about linear regression, a different class of model

  no curse of dimensionality - but potentially high bias
- generally most pairs (X, Y) will not have a deterministic relationship like $Y = f(X)$

  other unmeasured variables that also contribute to Y, including measurement error.
- often assume $Y = f(X) + \epsilon$ where $E(\epsilon) = 0$, there errors are independent of X, and are indentically distributed

  I am still uncertain as to its significance and import of this assumption

  I guess generally all you have for sure is a conditioanl distribution $P(Y|X = x)$

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

  where f_X is the marginal density

$$f_X(x) = \int_y f(x, y)dy$$

- discusses two paradigms for obtaining said approximation $\hat{f}$ to f, given some data; the algorithmic gray box view, vs the geometric functional approximation view

  I believe it is fruitfull to label the approximation to $f(x)$ and $\hat{f}(x; D)$; It acknowledges that you need both a procedure to get the approximation, and some data.
- He then talks about the **complexity** of models and the bias variance tradeof.

  important point in all these proofs is to consider a single prediction point x_0 and a set τ of different (all possible?) training data sets D.

  in the additive error model this encompasses both the observed x's and the not directly observed ϵ, which he typically just uses $Y = f + \epsilon$ assumption to separate the sources of error.

  I this part should be done more clearly.

2.1 Notation

- Input variable typically denoted by X.
- if X is vector, its components accessed by subscript X_j
- upper case refers to generic/abstract variable. Observed values are written with lowercase: i.e. ith observed value is x_i (which again can be scalar or vector)
- Matrices represented by bold upper case $\mathbf{X}$

  example: a set of N input p-vectors x_i where $i = 1..N$, is represented as the $N \times p$ matrix X
- In general vectors are not bolded, except when they have N components. This distinguishes a p-vector of inputs x_i for the ith observation from the N-vector $\mathbf{x_j}$ consisting of all observations of variable X_j.
- All vectors are assumed to be column vectors. Hence the ith row of $\mathbf{X}$ is x_i^T.

2.2 Two Simple Approaches to Prediction: Least Squares and Nearest Neighbors

2.2.1 Linear Model

- Given a vector of inputs $X^T = (X_1, \cdots, X_p)$, we predict the output Y via

$$\hat{Y} = \beta_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j$$

- often convenient to include the **bias** (aka intercept) term β_0 into X by including a constant variable 1 in X. Then letting $\hat{\beta}$ be the vector of coefficients including the bias, we can write the linear model in vector form as the inner product

$$\hat{Y} = X^T \hat{\beta}$$

- in general, $\hat{Y}$ could be a k-vector (i.e. the output is not scalar valued), in which case $\hat{\beta}$ would be a $p \times K$ matrix of coefficients.
- most popular approach to fit the linear model is the method of **least squares**: Pick the coefficients $\hat{\beta}$ to minimize the residual sum of squares:

$$RSS(\beta) := \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

  in matrix notation

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

  for nonsingular $X^T X$ solution is

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

2.3 Nearest-Neighbor Methods

- The k-nearest neighbor fit for $\hat{Y}$ is defined as follows:

$$\hat{Y} = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

  where $N_k(x)$ is the neighborhood of x defined by the k closest points x_i in the training sample.

  requires metric to define closest - typically euclidean

  No training required! That is no parameters are fit.
- while k-NN appears to have a single parameter k, in truth it has N/k effective parameters, where N is data size.

  this is generally much larger than linear model parameters and thus requires much more data.

  heuristic: if you have nonoverlapping clusters ok k points, then you have N/k such nhoods, and thus need N/K parms (the means) to describe result/fit.
- cannot use RSS on training to pick k, because it would always pick $k = 1$, which leads to zero training error.

2.4 Statistical Decision Theory

- let $X \in \mathbb{R}^p$ be real valued random input vector, $Y \in \mathbb{R}$ be real valued random ouput, with joint distribution $Pr(X, Y)$
- we seek function $f(X) : \mathbb{R}^p \to \mathbb{R}$ for prediting Y value
- for this we need a **loss function** $L(Y, f(X))$ for penalizing errors in prediction
- most common loss function is: **squared error loss**

$$L(Y, f(X)) = (Y - f(X))^2$$

- criterion for choosing f: minimize the expected prediction error (EPE)

$$EPE(f) = E(Y - f(X))^2 = \int [y - f(x)]^2 Pr(dx, dy)$$

- He presents some short argument that via conditioning on X and pointwise minimization, you can arrive at the solution

$$f(x) = E(Y|X = x),$$

that is, the *conditional expecation*, aka the **regression function**.

- I am interested in how to rigorously solve this - perhaps a constrained variational calculus problem?
- KNN is an approximation to this, where instead of relying only on observations at x exactly, a neighborhood of x is use to obtain the expected value of y.
- for linear regression - we propose ansatz $f(x) \sim x^T\beta$ - so don't search over all functions

  this is a model based approach

  theoretical solutions is:

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

- Using L1 instead of L2 as loss function leads to $\hat{f}(x) = median(Y|X = x)$
- For categorical output, loss function is matrix $K \times K$ matrix L, where K is number of categories. Zero in the diagonals, and nonnegative elsewhere

  typical is **zero-one loss function** - where all off diagonals are 1.

- The $EPE = E[L(G, \hat{G}(x)]$ where expectation is taken over $Pr(G, x)$
- using same conditioning argumetn and pointwise minimization, and zero one loss, leads to solution known as the **Bayes classifier**

$$\hat{G} = g_k \quad \text{if} \quad Pr(g_k|X = x) = max_{g \in G} Pr(g|X = x)$$

that is, classify as the most probable class, using discrete conditional distribution $Pr(G|X)$.

Error rate of the bayes classifier is often called the **Bayes rate**.

2.5 Local Methods in High Dimension

- Fitting/prediciton methods which rely on local approximations (like KNN), struggle as dimensions get high - **the curse of dimentionality**
- example: consider p-dimensional unit hypercube with uniformly distributed inputs. If we place a sub hypercube about the origin, such that a fraction r of the data is contained there, the linear dimension of this hypercube must be $r^{1/p}$. For 10 dimensions, the expected edge length for $r = 0.01$ is 0.63.

  so to capture one percent of the data, you must go %63 of the distance avaialble in each dimension! That is not local.

- another manifestation comes from looking at unit sphere with uniform distribution - median value for closest to origin is about 1/2 radius. So every point close to edge.

  this is a problem because for points on the edge, prediction is often extrapolation instead of interpolation. Which is much shakier.

- another manifestation of this curse is that the sampling density is proportional to $N^{1/}$ where N is sample size and p is dimension. So if in 1-D $N = 100$ represents a dense sample size, the equivalent density in 10 D is 100^{10}.

  so in high D, all feasable samples are sparse.

- **Bias-variance tradeoff** - nice little "proof" - setup: fix a point in domain x_0, get a large *set of training samples* τ - study the expected error in prediction at x_0 that comes from the sampling.

  notation: $\hat{y}_0$ is prediction using some model. y_0 is true value.

$$\begin{aligned} MSE(x_0) &= E_\tau[(y_0 - \hat{y}_0)^2] && (2.1)\\ &= E_\tau[(y_0 - E_\tau[\hat{y}_0] + E_\tau[\hat{y}_0] - \hat{y}_0)^2] && (2.2)\\ &= E_\tau[(y_0 - E_\tau[\hat{y}_0])^2] + 2E_\tau[(y_0 - E_\tau[\hat{y}_0])(E_\tau[\hat{y}_0] - \hat{y}_0)] + E_\tau[(\hat{y}_0 - E_\tau[\hat{y}_0])^2] && (2.3)\end{aligned}$$

  The term in blue is the variance of the prediction. The term in orange: $E_\tau[(y_0 - E_\tau[\hat{y}_0])^2] = (y_0 - E_\tau[\hat{y}_0])^2$ which is just the Bias of prediction squared. The middle term (in black) is zero (can factor out first piece, and distribute expetation over subtraction). So

$$MSE(x_0) = Bias_\tau(\hat{y}_0)^2 + Var_\tau(\hat{y}_0) \quad (2.4)$$

  This relationship ends up being very generic. More on this at the end of the chapter.

2.6 Statistical Models, Supervised Learning and Functional Approximation

- goal is to fint approximation $\hat{f}$ to the function f that undeliers the predictive relationship between inputs and outputs
- Sum of squared errors loss leads to optimal f being $f(x) = E(Y|X = x)$ - called the **regression function**.
- The class of nearest neighbor methods can be viewed as a direct approximation to this conidtional expectation
- More generally, we seek to approximate conditional probability $Pr(Y|X)$
- one common model: **additive error model**

  assumes:

$$Y = f(X) + \epsilon$$

  where f is the regression function, ϵ is a random error independent of X with some distribution such $E(\epsilon) = 0$

  in what sense is error dependent or independent of Y? - clearly for a given x, error has perfect correlation to Y - excess y is the error. But what about in general? Is that even a sensible question to ask?

  McElreath would say this is terrible way to think about it - better to just say something like $Y \sim N(\mu_x, \sigma)$ and $\mu_x = f(x)$ - this generalized better to non additive models. The additive nature in this case evident from normal distro.

- MAIN POINT: This is a specific claim about the conditional probability distribution $Pr(Y|X)$, namely that Y is distributed like ϵ plus a value determined by X.

  moreover, note that X only comes in through the conditional mean $f(x)$, and it does not come into the variance of Y (though that can and is often relaxed)

- For quantitative responses, this is typically not the assumption, but rather that of some bernoulli process (for binary var) with p of outcomes determined by X, that is $p(X)$. This binds both the conditional expectation and the variance to x.

  McElreath would say $Y \sim Bern(p_x)$ and $p_x = P(X)$.

- His claim is that there are two main ways to think of this endeavor to find a good approximation for f
  - Supervised learning - there is some algorithm that can take an input x_i and map it to an output $\hat{f}(x_i)$, which can also adjust $\hat{f}$ based on the difference between predicted value and observed y_i. This algorithm should produce a map that can be used for predictions.
  - Function Approximation - x_i, y_i are viewed as points in a (p+1)-dimensional Euclidean space. The idea is that the data satisfies some relationship $y_i = f(x_i) + \epsilon_i$, and goal is to obtain a useful approximation to f thatn is valid for all points in some region.

    This paradigm encourages mathematical concepts of geometry and probability, so they prefer it.

- often the approximations are restricted to some parameterized fammily of functions, and the challenge is to find the best parameter

  ex: linear model - $f(x) = x^T\beta$ (params are β), or more generally a **linear basis expansion**

$$f_\theta(x) = \sum_{k=1}^{K} h_k(x)\theta_k$$

  where h_k are a suitable set of functions or transformatiosn of the input vector.

- often fit by minimizing the residual sum-of-squares (this is just least squares error).
- A more general criteria is **maximum likelihood estimation** aka MLE
  - given random sample $y_i, i = 1 \cdots N$ from a density $Pr_\theta(y)$,
  - log-probability of data is
$$L(\theta) = \sum_i log(Pr_\theta(y_i))$$
  - the pricniple of MLE says most reasonable θ is that which maximizes $L(\theta)$
  - Least squares with additive error model $Y = f_\theta(x) + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$ is equivalent to MLE using conditional likelihood $Pr(Y|X) = N(f_\theta(X), \sigma^2)$
  - Another example: Assume multinomial qualitative output G, with regression function $Pr(G|X)$. Suppose we have model $Pr(G = g_k|X = x) = p_{k,\theta}(x)$, then the loglikehood is
$$L(\theta) = \sum_{i=1}^{N} \log p_{g_i,\theta}(x_i)$$
    which is also referred to as the **cross-entropy**
- From entropy, $\int_x p_x log(p_x)$ to cross entropy $\int_x u_x log p_x$, then to observed cross entropy (the $\int_x u_x$ become the sum ovber observed values) $\sum_i logp(x_i)$

2.7 Structured Regression Models

- infinite many solutions to min RSS (they just have to interpolate between data somehow)
- must impose restrictions on family of potential solutions - they are controlling the *complexity* of solutions in one way or another

  often impose some regularity on small neighborhoods

  size of neighborhood dicatates strength of complexity reduction (in k-NN, k controls that)
- main point: any method that constraints local variation in small isotropic neighborhoods will suffer curse of dimensionality; any method that overcomes the curse has some way of measuring neighborhoods which does not allow them to be small in all directions.

2.8 Classes of Restricted Estimators

- main approaches listed below:
- Roughness penalty
$$PRSS(f; \lambda) = RSS(f) + \lambda J(f)$$
where J is some functional that will large for rapidly changing functions over small reguins.

  one example $J(f) = \int [f''(x)]^2 ds$. $\lambda = \infty$ only allows linear functions.

  these are also called **regularization** methods
- Kernel methods

  estimate the regression function by specyfing nature of local neighborhoods

  use a **kernel function** $K_\lambda(x_0, x)$ which assigns weights to points x in a region around x_0

  example, Gaussian density function
$$K_\lambda(x_0, x) = \frac{1}{\lambda} \exp\left(-\frac{||x - x_0||^2}{2\lambda}\right)$$

  one example is the Nadaraya-Watson weighted average, where $\hat{f}(x)$ is the weighted sum of all y_i observations times kernels $K_\lambda(x, x_i)$.

  In general we can define a *local regression estimate* of $f(x_0)$ as $f_{\hat{\theta}}(x_0)$, where $\hat{\theta}$ minimizes
$$RSS(f_\theta, x_0) = \sum_{i=1}^{N} K_\lambda(x_0, x_1)(y_i - f_\theta(x_i))^2$$
and f_θ is some parameterized function, like a low order poly

ex: $f_\theta(x) = \theta_0$, results in Ndaraya-Watson

or linear $f_\theta(x) = \theta_0 + \theta_1 x$ — results in popular **local linear regression model**

notice that RSS here depends on both f_θ and x_0

- Basis functions

  includes linear and polynomial expansions (and much more)

  postulate a structure

$$f_\theta(x) = \sum_{m=1}^{M} \theta_m h_m(x)$$

  note that it is linear in the θs.

2.9 Model Selection and the Bias Variance tradeoff

- all of the models above has a **smoothing** or **complexity parameter**
- cannot use RSS on training data to determine these parameters

  such method ends up picking a solution that interpolates between data and hence has zero residuals, but is wild in between and not good at predicting future data.

- he goes through another - weirdly artifical - bias variance decomposition that show the dependency of the piece on the complexity parameter (for knn)
- main idea: as model complexity increases, squared bias decreases but variance increases.
- ideally, chose complexity parameter that leads to minimum test error.
- obvious estimate of test error is train error.
- unfortunately test error does not properly account for error that comes from model complexity.

2.10 Excercises

- 2.1. In the context of classification problem, suppose each of the K classes as associated with a vector t_k whose components $(t_k)_i = \delta_{k,i}$. Moreover the prediction $\hat{y}$ is a vector of probabilities (with components y_j), which sum to 1. Show that classifying via choosing the biggest probability y_g is equivalent to choosing the k via $min_k||t_k - \hat{y}||$.

  SOLUTION: assume L2 metric. Then must show $\sum_i(\delta_{g,i} - y_i)^2$ is less than $\sum_i(\delta_{j,i} - y_i)$ for all $j \neq g$. Just expand sum, cancel terms and it boils down to $y_g > y_j$.

- 2.2 - Each class has a probability distribution that is the sum of 10 normals. The bayes clasifier is all the set of points where these two distributions are equal.
- 2.3 - derive equation 2.24: consider N data points uniformly distributed in a p-dimensional sphere centered at the origin. Show that the median distance from the origin to the closes data point is given by

$$d(p, N) = \left(1 - \frac{1}{2}^{1/N}\right)^{1/p}$$

  SOLUTION: first, recall radial probaility density if pr^{p-1}, so CDF is $F_R(r) := P(R \leq r) = r^p$. Since he asks about the distribution of the min of N points, we are dealing with order statistics. In this case we want the probability that the min value is less than or equal to some r_0. The probability that the min is less than r_0 is equal to the probability that all points are greater than r_0. For a single point, the probability is $1 - r_0^p$. For N points it is $(1 - r_0^p)^N$. So the probability that the min is less then r_0 is equal to 1 minus that expression. To get the median, simply set the result equal to 1/2 and then solve for r_0. QED.

- 2.4. Another example of edge effect.
  - start with spherical multinormal: $X \sim N(0, I_p)$. distribution breaks up into product of p independent $N(0, 1)$.
  - expected distance from origin is just $E(x_1^2 + \cdots + x_n^2)$ (this is distribute according to χ_p^2 by definition).
  - by linearity of expectation and the fact that x_i is $N(0, 1)$, the epected distance is just p (each one is 1).
  - now fix a point x_0 (expected ditance is p). and define $\hat{a} = x_0/|x_0|$ (that is unit vector in that direction)
  - for each new sample x_i, let $z_i := \hat{a} \cdot x_i$.
  - by expanding the sum, can show hat $E(z_i) = 0$ (taking the x_0 values as fixed constants)

– similarly

$$E(z_i^2) = \sum_{j=1}^{p} E\left(\frac{(x_0)_j}{|x_0|}(x_i)_j\right)^2 + 0 \tag{2.5}$$

$$= \sum \frac{(x_0)_j^2}{|x_0|^2} = 1 \tag{2.6}$$

in line one we already expanded square of sum, and dropped all terms linear in the coordinates, since they lead to zero expectations.

– This tells you that given a point x_0, its expected distance from origin if p. All other data points projection onto that direction have expected distance 1. So as p grows, each point sees itself as lying at the edge of the training data.

Chapter 3

Linear Methods for Regression

3.0 LRS Summary of Main Ideas

- To understand the different perspectives, I think it is necessary to start from a very meta probability space. If we observe variables X and Y N times, we really have the randome meta-vector $(X_1, Y_1, \cdots X_N, Y_N)$, and this has some joint probability distribution.
- from this we need certain assumptions to simplify to a single (X, Y) probability distribution

  presumably some spacetime-translation invariance or something like that.
- but at a moments notice we must be ready to switch between then single/abstract (X, Y) perspective, and the each one is its own random variable perspective
- indeed if we fix the X's (and assume independence of Y's), it is useful to talk about N conditional $Y|_{X=x_i}$

Now on to the main ideas in chapter

-

3.2 Linear Regression Models and Least Squares

- Start with a random vector $X^T = (X_1, \cdots X_p)$ and random variable Y.
- linear model assumes that the regression function $Y = f(X)$ where $f(x) = E(Y|X = x)$ is (at least approximately) of the form $f(x) = \sum_{j=1}^{p} \beta_j X_j$

  the β_j's are uknown parameters
- let $\mathbf{X}$ be the $n \times p$ matrix of observations of X, and y be the observations of Y and $\beta = (\beta_0, \beta_1, ..\beta_n)^T$
- minimizng squared error loss function yield the following estimate $\hat{\beta}$

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

  minimum squared error criterion is a statistically reasonable criterion if the observations (x_i, y_i) represent a random draw from their population.

  even if x_i's not randomly drawn, it is still valid so long as y_i's are conditionally independent given the inputs x_i. What does this mean?

  perhaps this means, think not of (X,Y), but rather just Y, whose distribution is a function of x, and drawing from that family of distributions. As opposed to some meta distribution where y drawings depend on previous y drawings
- this process of minimizing squared errors, can be interpreted geometrically as finding the best p-dim hyperplane that approximates the data in $(p + 1)$-dim space.
- fitted values at the training inputs are:

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

  the matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^\mathbf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathbf{T}$ is called the **hat matrix** - because it puts hat on y.
- Hat matrix view shows a different geometrical interpretation:

- consider $\mathbb{R}^N$ this time (N number of observations)
- the column vectors of X span a subspace of $\mathbb{R}^N$ - call it Span($\mathbf{X}$)
- minimzing
$$RSS(\beta) = ||\mathbf{y} - \mathbf{X}\beta||^2$$
leads to an estimate $\hat{y}$ that lies in Span($\mathbf{X}$).
this can be seen by actually taking a derivative wrt β, set to 0, and solving - leads to $\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$
- hence $\hat{\mathbf{y}}$ is the *orthgonal projection* of y, and $\mathbf{H}$ is the projection matrix
can check it satisfies $HH = H$
- rank deficiencies leads to non-uniqueness. Typically taken care of by dropping some features.

- now "In order to pin down the sampling properties of $\hat{\beta}$, we now assume that the observations y_i are uncorrelated and have constant variance σ^2, and that the [observations] x_i are fixed (non-random)"

I read this: consider each y_i as a random variable. That is don't marginalize y but rather consider conditional distribution $y = f_{Y|X=x_i}(y)$. Assume this distribution has variance σ^2. And that these random variables are pairwise uncorrelated.

- first, from solution for $\hat{\beta}$ we can prove that
$$VAR(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$$
This is relatively straight forward. Can prove $VAR(AZ) = AVAR(Z)A^T$ for constant matrix A and random vector Z.
and uncorrelated, constant variance implies $VAR(Y) = \mathbf{I_n}\sigma^2$.

- second: the typical estimate of σ^2 is
$$\hat{\sigma}^2 = \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2.$$
This estimator is unbiased - that is: $E(\hat{\sigma}^2) = \sigma^2$.

- basic idea of proof: Not in ESL
  - recall, $\mathbf{X}$ values are fixed (thus so is $\mathbf{H}$). Consider random vector $y = (y_1, \cdots, y_n)$ (where each is Y conditional on fixed x_i). Further consider the random vector
  $$\hat{y} := \mathbf{H}y.$$
  recall what is still random is the as-yet to be observed $y = (y_1, \cdots y_n)$, which implies the $\hat{\beta}$ are random variables, and thus $\hat{y}$ are too.
  - now
  $$y - \hat{y} = (1 - \mathbf{H})y$$
  substituting $y = \mathbf{X}\beta + \epsilon$ and using $(1 - \mathbf{H})\mathbf{X} = 0$ yields
  $$y - \hat{y} = (1 - \mathbf{H})\epsilon$$
  note that $y - X\beta = \epsilon$, but $y - \hat{y} = (1 - H)\epsilon$, that is due to the difference in β and $\hat{\beta}$.
  - $1 - \mathbf{H}$ is also a projection matrix, and it projects onto space perpendicular so Span(X). Call that space Span$(X)^{\perp}$.
  - there exits an $N \times (N-p-1)$ matrix U such that $1-\mathbf{H}$ can be written as UU^T (by nature of being a projection operator)
  basically U has as columns an orthnormal basis for the space Span$(X)^{\perp}$ (which is (N-p-1) dimensional)
  - so we have
  $$\begin{aligned}
  \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 &= ||y - \hat{y}||^2 && (3.1)\\
  &= ||UU^T\epsilon||^2 && (3.2)\\
  &= \epsilon^T UU^T UU^T \epsilon && (3.3)\\
  &= \epsilon^T UU^T \epsilon && (3.4)\\
  &= ||U^T\epsilon||^2 && (3.5)
  \end{aligned}$$
  - Now, the assumption is that $\epsilon = (\epsilon_1, \cdots \epsilon_N) \sim N(0, I_N\sigma^2)$, so $U^T\epsilon \sim N(0, U^TU\sigma^2)$ and $U^TU = I_{N-p-1}$

- so $E||U^T\epsilon||^2 = (N-p-1)\sigma^2$
- more specifically $||U^T\epsilon|| \sim \sigma^2\chi^2_{N-p-1}$
- comment: clearly, for this proof we already added the more strict assumption referred to in the next lines

- Then "To draw inferences about the parameters and the model, additional assumptions are needed" - asssume:
  - the conditional expectation of Y really is linear in the X
  - the deviations of Y around its expectation are additive and gaussian, hence

$$Y = E(Y|X_1, \cdots X_p) + \epsilon \tag{3.6}$$

  where $\epsilon \sim N(0, \sigma^2)$

Now it is easy to show

$$\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$$

to do this, think of each observation y_i as a random variable Y_i, which by assumption is $Y_i = E(Y|X = x_i) + \epsilon_i$, where the ϵ_i are i.i.d $N(0, \sigma^2)$. Plugging this in, yields said result.

- in summary

$$\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2) \tag{3.7}$$

$$\hat{\sigma^2} \sim \frac{\sigma^2}{N-p-1}\chi^2_{N-p-1} \tag{3.8}$$

- moreover they are statistically independent. (How do we know this? they seem very tied together?)
- now we create hypothesis tests and confidence intervals
  - for a specific coefficient β_j we the *s*tandardized coefficient or **Z-score** is defined as

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

  where v_j is the j-th diagonal element of $(\mathbf{X^T X})^{-1}$
  - first note $\hat{\beta}_j \sim N(\beta_j, \sigma^2 v_j)$ - this is just marginalizing the multivariate distribution. so $\hat{\beta}_j/\sqrt{v_j} \sim N(\beta_j, \sigma^2)$.
  - so z_j has a ration distribution between a normal and a chi distribution. Using the ration of distributions theorem, and a lot math, you can show that z_j has a student-t distribution with $N-p-1$ dof, assuming that $\beta_j = 0$ (the null hypothesis).
    hence large absolute values of z_j lead to rejection of the null hypothesis.
  - if testing whether a group of coefficients have an impact together, we use the **F-statistic**

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(N-p_1-1)}$$

  where subscript 1 is for the bigger model with $p_1 + 1$ parameters, and 0 is for the nested smaller model, having $p_1 - p_0$ parameters constrained to be zero.
    the normalization factor is an estimate of σ^2.

3.2.1 Gauss Markov

- Gauss Markov theorem
  - VIP result!
  - context: assume linear model $y = X\beta + \epsilon$
  - assume we are trying to estimate any linear combination of the β parameters - say $\theta = a^T\beta$
    one such example is prediction $f(x_0) = x_0^T\beta$
  - claim: least squares estimate

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

  has the smallest variance among all *unbiased*, *linear estimates*.
    this is linear in the following sense: considering $\mathbf{X}$ to be fixed, then $\hat{\theta}$ is a linear function $c_0^T y$ of the response vector y.
  - in other words, given any other estimate $\tilde{\theta} = c^T y$, such that $E(c^T y) = a^T\beta$, then

$$Var(a^T\beta) \leq Var(c^T y)$$

- Proof of this and slight generalization is excercise 3.3

- Why do we care?
  - MSE for an estimator $\tilde{\theta}$, estimating θ, can be decomposed (bias-var tradeoff)

$$MSE(\tilde{\theta}) = E(\tilde{(\theta)} - \theta)^2 = Var(\tilde{\theta}) + \Big(E(\tilde{\theta} - \theta\Big)^2$$

  - so amongst all unbiased estimators, the lowest error is the lowest variance
  - moreover, expected prediction error, which is what we ultimately care about, is intimately tied to MSE
    given prediction $Y_0 = f(x_0) + \epsilon_0$ for new input x_0, the EPE for an estimate $\tilde{f}(x_0)$ is

$$E(Y_0 - \tilde{f}(x_0)) = \sigma^2 + MSE(\tilde{f}(x_0))$$

- However, if we lift unbiased restriction, we can often get an estimator with even lower MSE. Will see next section.
  All models are wrong (here read biased), but some models are useful.

3.2.2 From Univariate to MultiVariate regression

- start with no-intercept, univarate linear regression

$$Y = X\beta + \epsilon$$

- letting $\mathbf{y} = (y_1, \cdots y_N)^T$ and $\mathbf{x} = (x_1, \cdots x_N)^T$, and using $\langle a, b\rangle$ to denote inner prod, then we can write

$$\hat{\beta} = \frac{\langle \mathbf{x}, \mathbf{y}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle} \tag{3.9}$$

$$\mathbf{r} = \mathbf{y} - \mathbf{x}\hat{\beta} \tag{3.10}$$

  where $\mathbf{r}$ is the resisual vector

- Now, consider the *multiple linear regression model* where inputs are all orthgonal - that is $\langle \mathbf{x}_i, \mathbf{x}_j\rangle = 0$ for $i \neq j$
  this means that $\mathbf{X}^T\mathbf{X}$ is diagonal with values $\langle \mathbf{x}_i, \mathbf{x}_i\rangle$.
  this leads to

$$\hat{\beta}_j = \frac{\langle \mathbf{x}_j, \mathbf{y}\rangle}{\langle \mathbf{x}_j, \mathbf{x}_j\rangle}$$

  which is just the regression coefficient for each one independently

- So the difference between man univariate regression coeffieicents and a general multivariate regression solution, is the extent to which the input variables are not orthgonal to each other.
  NOTE THEY ARE NOT SAYING UNCORRELATED - BUT ORTHGONAL!

- now consider *univariate regression* with an intercept. Can show that the least squares coefficient of $\mathbf{x}$ has the form

$$\frac{\langle \mathbf{x} - \bar{x}\mathbf{1}, \mathbf{y}\rangle}{\langle \mathbf{x} - \bar{x}\mathbf{1}, \mathbf{x} - \bar{x}\mathbf{1}\rangle}$$

  this can be arrived at by 1) regressing $\mathbf{x}$ on $\mathbf{1}$ and getting the residual $\mathbf{z} = \mathbf{x} - \bar{x}\mathbf{1}$ and then 2) reressing $\mathbf{y}$ on residual $\mathbf{x}$

- This can be generalized to p variables, into what is called **Regression by Succesive orthgonalization** aka **Gram-Schmidt procedure for multiple regression**

- process: choose a variable whose coeffient you will calulate, say $\mathbf{x}_k$, then
  - initialize $\mathbf{z}_0 = \mathbf{x}_0 = 1$
  - for all $j \neq k$ regress x_j on all the $\mathbf{z}$s produce so far. They are orthogonal, so coefficient is univariate regression coefficient.
    then get the residual $\mathbf{z}_j$.
  - finally regress $\mathbf{x}_k$ on all the $p-1$ $\mathbf{z}$s, and obtain its residual $\mathbf{z}_k$.

$$\hat{\beta}_p = \frac{\langle \mathbf{z}_k, \mathbf{y}\rangle}{\langle \mathbf{z}_k, \mathbf{z}_k\rangle} \tag{3.11}$$

  - Then regress $\mathbf{y}$ on $\mathbf{z}_k$ to give $\hat{\beta}_k$.

- repeat this, from scratch for each p
- since all the $\mathbf{z}$s are orthogonal, they form a basis space for Span $\mathbf{X}$, therefore the least squares projection onto this space is $\hat{y}$. Since $\mathbf{z}_k$ alone involves $\mathbf{x}_k$, with coefficient 1, that means that $\mathbf{z}_k$ coefficient of $\hat{y}$ is indeed the $\mathbf{x}_k$ coefficient.
- in conclusion: "The multiple regression coefficient $\hat{\beta}_j$ represents the additional contribution of $\mathbf{x}_j$ on $\mathbf{y}$, after $\mathbf{x}_j$ has been adjusted for $\mathbf{x}_0, \cdots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \cdots, \mathbf{x}_p$.
- from (3.11) can obtain $Var(\hat{\beta}_k) = \sigma^2/||\mathbf{z}_k||$.

  this says that the precision with which we can estimate that coefficient depends on the length of the residual - that is, how much new info is in that
- One pass with **QR decomposition** via Gram schmidt
  - after one pass of gram-schmidt, can write
    $$\mathbf{X} = \mathbf{Z\Gamma}$$
    where $\mathbf{Z}$ has as columns the $\mathbf{z}_j$ (in order), and the $\mathbf{\Gamma}$ is the upper triangular matrix that converts it to $\mathbf{X}$ - basically the regression coefficients obtained during gram-schmidt
  - Introduce the diagonal matrix $\mathbf{D}$, with jth diagonal entry $D_jj = ||\mathbf{z}_j||$
    $$\begin{aligned} \mathbf{X} &= \mathbf{ZD^{-1}D\Gamma} && (3.12) \\ &= \mathbf{QR} && (3.13) \end{aligned}$$
    this is the *QR decomposition* of $\mathbf{X}$. $\mathbf{Q}$ satisfies $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$ and $\mathbf{R}$ is upper triangular.
  - can show that
    $$\begin{aligned} \mathbf{R}\hat{\beta} &= \mathbf{Q}^T\mathbf{y} && (3.14) \\ \hat{\mathbf{y}} &= \mathbf{QQ}^T\mathbf{y} && (3.15) \end{aligned}$$
    The top equation is easy to solve because R is upper triangular.
- can easily generalize to multiple outputs.

  Basically write each component as its own regression

  if erros uncorrelated, just sum all dimensions - least square solution has exactly same form

  if there is reason to believe errors are correlated, then loss function is different, but optimal estimate is still the same.

3.3 Subset selection

- Main idea: use training data to produce sequence of models of varying complexity, indexed by a single parameter,

  Use CV or AIC (will talk about later) to choose value of complexity parameter

  CV is performed using only training data since it is being used to train the metaparameter.
- four approaches using choice of predictors
  - best-subset selection

    for each $N \in \{0, \cdots, p\}$, choose best choice of predictors.

    feasible only for $p \sim\leq 40$
  - forward step-wise selection

    start with just intercept, then one at a time add predictor that most improves fit

    greedy algorithm

    much faster computationally and much more constrained, so less variance but more bias
  - backwards stepwise selection

    similar but start with full set and take one out at a time (drop var with smallest z-score)
  - forward stagewise selection

    start with $y = \bar{y} + 0x_1 + \cdots$ with xs centered

    compute residual; pick var most correlated to residual; add regression to residual

    this is slow, but good in high dimensions

    eventually converges to OLS
- be careful - must account for multiple testing problem of overfitting

- also must consider variables that naturally come in groups (like dummy variables)
- question: in the ten-fold CV procedure, say in best subset selection, for a given number of predictors k, each fold might yield a different set of predictors. So in choosing the best model, we cannot really be choosing the best k predictors, but rather choosing the parameter k. Correct? And then fit it again using the whole training data?
- The CV procedure can be used to provide standard error bands
- in the end they use the **one standard error rule** - pick the most parsimonious model within one standard error of the minimum.

3.4 Shrinkage Methods

- Ridge regression - impose L2 penalty on coefficients

$$\hat{\beta}_{ridge} = argmin_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{3.16}$$

  here λ id s complexity parameter

  there is an equivalent formulation in using regular OLS with constraint that sum of squares of betas are less than some fixed value (which is tunable - and related to λ above).

- VIP: notice that the intercept β_0 has been left out of the penalty term. Including it would make procedure dependent on origin of Y.
- VIP: The ridge solutions are not equivalent under scaling of the inputs. So one normally standardizes inputs before solving.
- this procedure is very useful when there are many correlated variables
- Note: if we start with centered X - namely use $\tilde{x}_{ij} := x_{ij} - \bar{x}_j$ - and obtain ridge regression coefficients $\hat{\beta}_c$, simple substitution into ridge definition (3.16), yields that centered and uncentered solutions are related via

$$\beta^0 = \beta_c^0 + \sum_{j=1}^{p} \bar{x}_j \beta_c^j \tag{3.17}$$

$$\beta^j = \beta_c^j \tag{3.18}$$

- Furthermore, the β_0 appears only in the first piece, and its minimum value can be easily shown to be (in the centerd case)

$$\beta_0 = \bar{y}$$

- The remaining problem is now reduced to a regression without intercept, with centered X and Y. Henceforth we can assume that this is the starting point of ridge regression.
  - we can rewrite ridge criterion as
$$RSS(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$
  - solution is
$$\hat{\beta}^{ridge} = \left(\mathbf{X}^T\mathbf{X} + \lambda I \right)^{-1} \mathbf{X}^T\mathbf{y}$$

3.5 Excercises

excercise 3.1

- Uses known relationship where student-t squared is f-distribution (for the appropriate parameters)

excercise 3.3

- The typical proof for this is all over the web - it starts by letting $c = a + d$, deriving a constraint on d based on the unbiased nature and the using that constraint in tha variance expansion. The generalization to vector/matrix is straightforwadr (part b).
- I attempted a slitghly different proof. Here is the gist:
  - OLS estimator $\theta = a \cdot \hat{\beta}$ where a is a constant and $[a] = p \times 1$.

- other linear estimator $\tilde{\theta} = c \cdot y$ which is also unbiased
  thus $c \cdot \bar{y} = a \cdot \beta$
- cauchy schwarz says
$$c \cdot c \geq \frac{(c \cdot \bar{y})^2}{\bar{y} \cdot \bar{y}}$$
- now $c \cdot \bar{y} = a \cdot \beta$ so
$$\begin{aligned} (c.\bar{y})^2 &= (a \cdot \beta)^2 && (3.19)\\ &= (a \cdot E(\hat{\beta})^2 && (3.20)\\ &= (a \cdot E\left((X^T X)^{-1} X^T y\right))^2 && (3.21)\\ &= (a \cdot (X^T X)^{-1} X^T \bar{y})^2 && (3.22) \end{aligned}$$
- now the last line can be rewritten as
$$\cdots = (a^T (X^T X)^{-1} X^T \bar{y})(\bar{y}^T X (X^T X)^{-1T} a^T) \quad (3.23)$$
- now, I don't see how to do it right now, but if we could group the middle $\bar{y}$ together and somehow factor it out, then the expression would simplify to
$$a^T (X^T X)^{-1} a (\bar{y}^T \bar{y}).$$
- then put back into cauchy-schwarz above would yield
$$c \cdot c \geq a^T (X^T X)^{-1} a$$
  which is what was to be proven. but I must justify that leap somehow. Perhaps not a strict equality, but using an inequality. Will continue down this path later.

excercise 3.4

- see companion notebook: Ch3_02_GramSchmidtRegressionAndQRDecomp

excercise 3.5

- Done in situ, see equation (3.17).