Chapter 3 Lecture Notes

Elements of Statistical Learning

1 Means and Variances

Important Properties of Estimators
Let $x_1, \ldots, x_N$ be i.i.d. with mean μ and variance σ^2. Let $\hat{\mu} = E_i[x_i]$ and $\hat{\sigma}^2 = \frac{N}{N-1} E_i(x_i - \hat{\mu})^2$ be the estimated mean and estimated variance respectively. Then $E[\hat{\mu}] = \mu$ and $E[\hat{\sigma}^2] = \sigma^2$.

First, it is easy to show that $E[\hat{\mu}] = \mu$:

$$\begin{aligned} E[\hat{\mu}] &= E[E_i[x_i]] \\ &= E_i[E[x_i]] \\ &= E_i[\mu] \\ &= \mu. \end{aligned} \tag{1}$$

Next, we know that

- $\operatorname{Var}(cx) = c^2 \operatorname{Var}(x)$
- $\operatorname{Var}(\sum x_i) = n \operatorname{Var}(x_i) = n\sigma^2$, since they are i.i.d..
- The above two, give us that $\operatorname{Var}[\hat{\mu}] = \frac{\sigma^2}{n}$. This makes sense, the more data we have the closer we get to the mean, and so the variance becomes smaller.

Using the above property, we can compute the expected estimated variance. First, notice that

$$\begin{aligned} E_i(x_i - \hat{\mu})^2 &= E_i(x_i^2 - 2x_i\hat{\mu} + \hat{\mu}^2) \\ &= E_i(x_i^2) - 2\hat{\mu}E_i(x_i) + \hat{\mu}^2 \\ &= E_i(x_i^2) - \hat{\mu}^2. \end{aligned} \tag{2}$$

Therefore, we compute $E(\hat{\sigma}^2)$ as follows:

$$\begin{aligned}
E(\hat{\sigma}^2) &= \frac{N}{N-1} E(E_i(x_i - \hat{\mu})^2) \\
&= \frac{N}{N-1} E(E_i(x_i^2) - \hat{\mu}^2) \\
&= \frac{N}{N-1} \left[E(E_i(x_i^2)) - E(\hat{\mu}^2)\right] \\
&= \frac{N}{N-1} \left[E_i(E(x_i^2)) - E(\hat{\mu}^2)\right] \\
&= \frac{N}{N-1} \left[E_i(\sigma^2 - \mu^2) - E(\hat{\mu}^2)\right] \\
&= \frac{N}{N-1} \left[\sigma^2 - \mu^2 - \operatorname{Var} \hat{\mu} + (E\hat{\mu})^2\right] \\
&= \frac{N}{N-1} \left[\sigma^2 - \mu^2 - \sigma^2/N + \mu^2\right] \\
&= \frac{N}{N-1} \frac{N-1}{N} \sigma^2 = \sigma^2.
\end{aligned} \tag{3}$$

2 Estimated Values in Linear Regression

Important Properties of Estimators
Suppose that the regression function $E[Y|X] = f(X)$ and that $\operatorname{Var} Y = \sigma^2$. Suppose also that x_i are fixed, not random, and the only randomness is on the y_i. We compute the least squares as $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ and the estimated variance as $\hat{\sigma}^2 = \frac{N}{N-p-1} E_i(y_i - \hat{y}_i)^2$. Then $\operatorname{Var}[\hat{\beta}] = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$ and $E[\hat{\sigma}^2] = \sigma^2$.

We first compute $\operatorname{Var}[\hat{\beta}]$. We have

$$\begin{aligned}
\operatorname{Var}[\hat{\beta}] &= \operatorname{Var}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \operatorname{Var}[\mathbf{y}]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{I}_p\sigma^2\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2.
\end{aligned} \tag{4}$$

3 Important Distributions

3.1 Chi-squared

Let $x_1, \ldots, x_n$ be IID standard normal random variables $x_i \sim \mathcal{N}(0,1)$ and let $\|x\|_2 = nE_i x_i^2 = \sum_i x_i^2$. Then

$$\|x\|_2 \sim \chi_n^2.$$

Of course χ^2_n is always positive and drops similarly to how the normal drops on its right.

Clearly, the same holds no matter whether the vector is free or bound. Hence, if $x \sim \mathcal{N}(\mu, I_n)$, then

$$\|x\|_2 = \sum (x_i - \mu)^2 \sim \chi^2_n$$

again.

Additionally, Cochran's theorem states that $nE_i(x_i - E_i x_i)^2 = \sum_i (x_i - \hat{\mu})^2 \sim \chi^2_{n-1}$. So, if we instead of the real mean we use the estimated mean, we lose one degree of freedom.

Also, $(n-1)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{n-1}$.

3.2 t-distribution

As before, let $x_1, \ldots, x_n \sim \mathcal{N}(\mu, \sigma^2)$. We saw above that the sample variance (a.k.a. unbiased variance estimation) is $\hat{\sigma}^2 = \frac{N}{N-1} E_i(x_i - \hat{\mu})^2$. It holds that

$$\frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1). \qquad \text{(See Wiki)}$$

However, if instead we use the sample variance, we get

$$\frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}. \tag{5}$$

Observe that both of the above quantities have distributions that do not depend neither on μ nor on σ. However, specifically the second one has only one unknown, the mean μ. So we can use this distribution to derive confidence intervals for μ.

For example, we can make the null hypothesis that $\mu = \mu^*$ for some μ^*. If this is the case, then we expect the expression

$$\frac{\hat{\mu} - \mu^*}{\hat{\sigma}\sqrt{n}}$$

to follow the t_{n-1}-distribution which is almost the same as the normal distribution. But note that we know all parameters of this expression so we can easily evaluate it. If we get a value that is close to 0 then we accept the hypothesis, otherwise we reject it.

Hypothesis testing using the t-distribution
As the sample size n increases, the two distributions $\mathcal{N}(0,1)$ and t_{n-1} look very alike and $t_\infty = \mathcal{N}(0,1)$. Especially the difference between their tail quantiles becomes negligible in n. This suggests an easy way to verify the null hypothesis that $\mu = 0$. Simply take the quantity $\frac{\hat{\mu}}{\hat{\sigma}^2/\sqrt{n}}$ and check whether it is far from zero. The further from zero it is, the more likely that the null hypothesis does not hold and should be rejected.

Definition 1 (t-distribution)**.** Let Z, V independent random variables, such that $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_v$. Then

$$T = \frac{Z}{\sqrt{V/v}} \sim t_v$$

Student's t looks like a normal distribution but pushed down.

3.3 F-distribution

My first observation is that the F distribution is to the t-distribution what the χ-distribution is to the normal distribution. The F-distribution has two degrees of freedom d_1, d_2 and we write F_{d_1,d_2}.

Property of χ^2 and F
Suppose X has a Student's t-distribution with degree of freedom v; i.e., $X \sim t_v$. Then $X^2 \sim F_{1,v}$.

Definition 2. Suppose $S_1 \sim \chi^2_{d_1}$ and $S_2 \sim \chi^2_{d_2}$ are two independent random variables. Then

$$X = \frac{S_1/d_1}{S_2/d_2} \sim F_{d_1,d_2}.$$

So it is the ratio of two independent appropriately scaled χ^2 distributions.

Similarly to the chi-squared distribution, the F-distribution is always positive and has a similar shape as the chi-squared.

3.4 Z-score

Let X be a random variable with mean μ and variance σ. Let $x \leftarrow X$, be a sample from this distribution. Then the Z-score of x is the distance of x from the mean, measured in standard deviations:

$$z = \frac{x - \mu}{\sigma}.$$

Notice that if $x \sim \mathcal{N}(\mu, \sigma^2)$, then $z \sim \mathcal{N}(0, 1)$; i.e., it follows the standard normal distribution. This is called *normalization* in general. The goal is to use our samples to arrive to a *pivotal quantity*, meaning a quantity whose distribution is known to us and is independent of the parameters of the real distribution.

Example. Suppose that $\hat{x} = \hat{\mu} = E_i x_i$ is the average (sample mean, mean estimation) of the x_is. We saw earlier that $E[\hat{x}] = \mu$. Moreover, $\hat{\sigma} = \frac{N}{N-1} E_i (x_i - \hat{\mu})^2$ is the sample variance such that $E[\hat{\sigma}^2] = \sigma^2$. Then the Z-score of $\hat{x}$, is

$$z = \frac{\hat{x} - E[\hat{x}]}{\hat{\sigma}/\sqrt{N}}.$$

This follows the distribution t_{N-1} as we saw.

4 Applications to 3.2

We have already seen that

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

Moreover, similarly to how we computed the unbiased variance estimator above for one dimension, here we have $p+1$ dimensions, and hence,

$$\hat{\sigma}^2 = \frac{N}{N-p-1}E_i(\hat{y}_i - y_i)^2.$$

As before, $E[\hat{\sigma}^2] = \sigma^2$.

Now, let's additionally assume that the real model is linear; i.e. $E[Y|X] = X^T\beta$ and $Y = X^T\beta + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

In this case, our beta estimator $\hat{\beta}$ is also normal:

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

Moreover, $(N-p-1)\hat{\sigma}^2 \sim \sigma^2\chi^2_{N-p-1}$ similarly to what we had above but for more dimensions.

Also, $\text{Cov}(\hat{\beta}, \hat{\sigma}^2) = 0$. We define the Z-score of $\hat{\beta}_j$ as

$$z_j = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{v_j}},$$

where v_j is the jth diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$. Notice that in the above expression the only unknown is β_j. So we can possibly plug different values until we find an expression that is close to 0.

For example, if we make the null hypothesis that $\beta_j = 0$, then $z_j \sim t_{N-p-1}$, which is almost the same as $\mathcal{N}(0,1)$.

4.1 Removing more variables

Often, we may have a feeling that there is a group of coefficients that does not affect the dependent variable. In this case, we can apply least squares *with* and *without* this set of variables getting RSS_1 and RSS_0 respectively. In this case, the F-statistic is:

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

where p_1+1 is number of parameters of the bigger model and p_0+1 is the number of parameters of the smaller model. As we guess,

$$F \sim F_{p_1-p_0, N-p_1-1}.$$

Again, the F-distribution is very similar to the $\chi^2_{p_1-p_0}/(p_1-p_0)$ distribution at the tail quantiles. So the further we are from zero, the more likely the larger model to be the correct one.

4.2 Confidence Intervals

Finally, suppose that we have computed the pivotal quantity z_j and got a particular value. Now, since we know that the distribution of z_j is very close to the standard normal, we can get a confidence interval for β_j.

For example, suppose we want a 95% confidence interval. Then, we look for a c such that

$$\Pr[-c \leq z_j \leq c] = 0.95,$$

which amounts to $c = 1.96$.

We have

$$\begin{aligned} -c \leq \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{v_j}} \leq c \implies & -c\hat{\sigma}\sqrt{v_j} \leq \hat{\beta}_j - \beta_j \leq c\hat{\sigma}\sqrt{v_j} \\ \implies & -c\hat{\sigma}\sqrt{v_j} - \hat{\beta}_j \leq -\beta_j \leq c\hat{\sigma}\sqrt{v_j} - \hat{\beta}_j \\ \implies & -c\hat{\sigma}\sqrt{v_j} + \hat{\beta}_j \leq \beta_j \leq c\hat{\sigma}\sqrt{v_j} + \hat{\beta}_j, \end{aligned} \tag{6}$$

where c is the 95% percentile of the standard normal distribution. Hence, with probability 95%,

$$\beta_j \in (\hat{\beta}_j - c\hat{\sigma}\sqrt{v_j}, \hat{\beta}_j + c\hat{\sigma}\sqrt{v_j}),$$

where $\hat{\sigma}\sqrt{v_j} = \text{se}(\hat{\beta}_j)$ is the standard error of $\hat{\beta}_j$. By approximating $c = 2$, we get the standard practice of reporting $\hat{\beta}_j \pm \text{se}(\hat{\beta}_j)$ as the 95% confidence interval.

4.3 Reproducing Correlation Matrix

By running the python script `prostate-data/correlation.py`, we get

Correlation Matrix

	cavol	lweight	age	lbph	svi	lcp	gleason	pgg45	lpsa
lcavol	1.00000	0.30023	0.28632	0.06317	0.59295	0.69204	0.42641	0.48316	0.73316
lweight	0.30023	1.00000	0.31672	0.43704	0.18105	0.15683	0.02356	0.07417	0.48522
age	0.28632	0.31672	1.00000	0.28735	0.12890	0.17295	0.36592	0.27581	0.22764
lbph	0.06317	0.43704	0.28735	1.00000	-0.13915	-0.08853	0.03299	-0.03040	0.26294
svi	0.59295	0.18105	0.12890	-0.13915	1.00000	0.67124	0.30688	0.48136	0.55689
lcp	0.69204	0.15683	0.17295	-0.08853	0.67124	1.00000	0.47644	0.66253	0.48920
gleason	0.42641	0.02356	0.36592	0.03299	0.30688	0.47644	1.00000	0.75706	0.34243
pgg45	0.48316	0.07417	0.27581	-0.03040	0.48136	0.66253	0.75706	1.00000	0.44805
lpsa	0.73316	0.48522	0.22764	0.26294	0.55689	0.48920	0.34243	0.44805	1.00000

OLS Regression Results

Dep. Variable:	lpsa	R-squared:	0.694
Model:	OLS	Adj. R-squared:	0.652
Method:	Least Squares	F-statistic:	16.47
Date:	Fri, 01 Apr 2022	Prob (F-statistic):	2.04e-12
Time:	20:36:00	Log-Likelihood:	-55.359
No. Observations:	67	AIC:	128.7

```
Df Residuals:                       58   BIC:                             148.6
Df Model:                            8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1.527e-16      0.073    2.1e-15      1.000      -0.145       0.145
lcavol         0.5931      0.111      5.366      0.000       0.372       0.814
lweight        0.2423      0.088      2.751      0.008       0.066       0.419
age           -0.1180      0.085     -1.396      0.168      -0.287       0.051
lbph           0.1755      0.085      2.056      0.044       0.005       0.346
svi            0.2563      0.104      2.469      0.017       0.049       0.464
lcp           -0.2393      0.128     -1.867      0.067      -0.496       0.017
gleason       -0.0173      0.118     -0.147      0.884      -0.254       0.219
pgg45          0.2296      0.132      1.738      0.088      -0.035       0.494
==============================================================================
Omnibus:                        0.825   Durbin-Watson:                   1.690
Prob(Omnibus):                  0.662   Jarque-Bera (JB):                0.389
Skew:                          -0.164   Prob(JB):                        0.823
Kurtosis:                       3.178   Cond. No.                         4.44
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

<F test: F=1.6697548846375199, p=0.16933707265225223, df_denom=58, df_num=4>
```

In the above, the z-score is denoted by `t` and measures the effect of dropping the variable from the model. An absolute value greater than 2, means that with high probability this variable is significant.

5 Shrinkage

Dropping predictors is only one way we can potentially reduce the variance at the cost of introducing bias. By dropping some parameters we also get the advantage of having a smaller and more intuitive model. However, the downside of best subset selection is that it is a discrete process so it does not usually behave well.

Ridge regression shrinks the coefficients of the linear regression but in a "homogeneous" way. It does not allow for a huge shrinkage in one coefficient but no shrinkage in another. How can we enforce this?

- Ridge regression. Here, we impose a universal penalty defined via the complexity parameter λ and hence we try to minimize the expression:

$$\overset{\text{ridge}}{\text{RSS}}(\lambda, \beta) = \sum_{i \in [N]} \left(y_i - \beta_0 - \sum_{j \in [p]} x_{i,j} \beta_j \right)^2 + \lambda \sum_{j \in [p]} \beta_j^2$$

- The lasso. Here we use a different norm and we have:

$$\overset{\text{lasso}}{\text{RSS}}(\lambda, \beta) = \sum_{i \in [N]} \left(y_i - \beta_0 - \sum_{j \in [p]} x_{i,j} \beta_j \right)^2 + \lambda \sum_{j \in [p]} |\beta_j|$$

- In general, we could have any norm, and so we would get

$$\overset{q}{\text{RSS}}(\lambda, \beta) = \text{RSS}(\beta) + \lambda \|\beta\|_q,$$

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

Notice that now, one cannot have too "long" coefficients since this would make the RSS considerably large. In other words, these new metrics, enforce picking small coefficients. This is not the case necessarily in a linear regression model since one could have a very large positive coefficient β_i and a very large negative coefficient β_j. These, would cancel each other in RSS and could give us a small error, yet a very large variance. On the other hand, they would not cancel each other in these expanded versions since we add up their absolute values.

Notes

- Ridge regression is not equivariant under scaling of inputs.
- It is normal to first standardize the inputs before applying ridge regression
- We need to omit the intercept β_0 from the procedure. If we don't, then shifting each target y_i by a constant c does not simply shift the predictions by c. In other words, if we include β_0, then the procedure would depend on the origin chosen for Y. I don't see why this is bad. It is not a linear model if the first case.
- Assuming we omit β_0 from the process, we work as follows:
  1. Set $x_{i,j} = x_{i,j} - \overline{x}_j$, $i \in [N], j \in [p]$. These are called the *centered* inputs.
  2. Set $\beta_0 = \overline{y}$
  3. Run ridge regression without intercept, where $\mathbf{X}$ has p columns.

Now we just need to find

$$\arg\min_{\beta} \left[(\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda \beta^T \beta \right],$$

where $\mathbf{X}$ has p columns.

5.1 Ridge Interpretation as the Mode of a Distribution

Suppose that $y \sim \mathcal{N}(\beta_0 + x^T\beta, \sigma^2)$ and $\beta \sim \mathcal{N}(0, \tau^2\mathbf{I})$. Now, given N samples $(x_i, y_i)_{i\in[N]}$, what is the conditional distribution of β? Notice that there is a dependency between the samples and β, therefore, the distribution of β given the samples is not necessarily the same as original distribution. It can be shown that

$$-\log f(\beta|(x_i, y_i)_{i\in[N]}) = \overset{\mathsf{lasso}}{\mathrm{RSS}}(\lambda, \beta),$$

where $\lambda = \sigma^2/\tau^2$ and $f(\beta)$ is the conditional probability density function of β, given the samples.

Now, by definition,

$$\arg\max_{\beta} \overset{\mathsf{lasso}}{\mathrm{RSS}}(\lambda, \beta) = \mathrm{mode}(f).$$

5.2 Solution of Ridge Regression

Minimizing the criterion $\mathrm{RSS}(\lambda, \beta) = \mathrm{RSS}(\beta) + \lambda\|\beta\|_2 = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta$, gives us the solution

$$\hat{\beta}^{\mathsf{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}.$$

Compare this with the least squares solution

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

Observation. Note that in ridge regression, as with least squares, it is very rare that a coefficient will be zero. On the other hand, we would be interested in being able to compare the ridge solution with the subset selection solution. This becomes possible using the effective degrees of freedom:

$$\mathsf{df}(\lambda) = \mathrm{trace}[\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T] = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}.$$

Notice that this value is at most p and if $\lambda = 0$, we get the standard least squares solution and $\mathsf{df}(0) = p$. By setting some of the coefficients to 0 in subset selection, we get a smaller value.

Assuming that the response has higher variance along the directions of inputs with higher variance, Ridge regression performs better.

6 Important Matrix Decompositions

6.1 Eigendecomposition

Let $\mathbf{A}$ be an $N \times N$ matrix with eigenvectors $q_1, \ldots, d_N$ and eigenvalues $\lambda_1, \ldots, \lambda_N$. Then we know that

$$\begin{aligned} & \mathbf{A}q_i = q_i\lambda_i \\ \implies & \mathbf{AQ} = \mathbf{Q\Lambda} \\ \implies & \mathbf{A} = \mathbf{Q\Lambda Q}^{-1} \text{ and } \mathbf{A}^{-1} = \mathbf{Q\Lambda}^{-1}\mathbf{Q}^{-1} \end{aligned} \tag{7}$$

where $\mathbf{Q} = [q_1, \ldots, q_N]$ is an $N \times N$ matrix whose ith column is q_i. So if we know Q and Q^{-1}, we can just invert $\mathbf{\Lambda}$ and since $\text{diag}^{-1}(\lambda_1, \ldots, \lambda_N) = \text{diag}(1/\lambda_1, \ldots, 1/\lambda_N)$, this is easy to calculate.

6.2 Singular Value Decomposition

SVD is the generalization of eigendecomposition for matrices that are not square. If $\mathbf{X}$ is an $N \times p$ matrix, then it can be written as

$$\mathbf{X} = \mathbf{UDV}^T,$$

where

- $\mathbf{U} = [u_1, \ldots, u_p]$ is $N \times p$. Its columns span the columnspace of $\mathbf{X}$. It is orthogonal, meaning $\mathbf{U}^T = \mathbf{U}^{-1}$, and more generally, it is unitary, if complex ($\mathbf{U}^* = \mathbf{U}^{-1}$).
- $\mathbf{V} = [v_1, \ldots, v_p]$ is $p \times p$. Its columns span the rowspace of $\mathbf{X}$. Similarly, it is also unitary: $\mathbf{V}^T = \mathbf{V}^{-1}$.
- $\mathbf{D} = \text{diag}(d_1, \ldots, d_p)$, such that $d_1 \geq d_2 \ldots \geq d_p \geq 0$, called the singular values of $\mathbf{X}$.

We can rewrite $\mathbf{X}$ as

$$\mathbf{X} = \sum_{j \in [p]} d_i u_i v_i^T.$$

which shows that $\mathbf{X}$ can be written as the sum of p rank-1 matrices.

More intuition about SVD can be found here.

We see that

$$\begin{aligned} \mathbf{X}^T\mathbf{X} &= \mathbf{VD}^T\mathbf{U}^T\mathbf{UDV}^T = \mathbf{V}(\mathbf{D}^T\mathbf{D})\mathbf{V}^T \\ \mathbf{XX}^T &= \mathbf{UDV}^T\mathbf{VD}^T\mathbf{U}^T = \mathbf{U}(\mathbf{DD}^T)\mathbf{U}^T \end{aligned}$$

So these are the eigendecompositions of $\mathbf{X}^T\mathbf{X}$ and $\mathbf{XX}^T$ respectively.

6.2.1 Applying SVD to Least Squares and Ridge Regression

Least Squares. The least squares fitted vector can be written as

$$\begin{aligned}\hat{\mathbf{y}}^{\mathsf{ls}} &= \mathbf{X}\hat{\beta}^{\mathsf{ls}} \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{U}\mathbf{U}^T\mathbf{y}.\end{aligned} \tag{8}$$

In other words, $\mathbf{U}\mathbf{U}^T$ is another hat matrix.

Ridge. Now for ridge regression, the fitted vector can be written as

$$\begin{aligned}\hat{\mathbf{y}}^{\mathsf{ridge}} &= \mathbf{X}\hat{\beta}^{\mathsf{ridge}} \\ &= \mathbf{U}\mathbf{D}(\mathbf{D}\mathbf{D}^T + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^T\mathbf{y} \\ &= \sum_{j=1}^{p} \mathbf{u}_j \frac{d_j^2}{d_j^2 + \lambda}\mathbf{u}_j^T\mathbf{y},\end{aligned} \tag{9}$$

where $\mathbf{u}_j$ are the column vectors of $\mathbf{U}$.

We see that ridge regression, like linear regression computes the coordinates of $\mathbf{y}$ with respect to the orthonormal basis $\mathbf{U}$. The difference is that ridge regression also shrinks these coordinates by the fraction above.

6.3 QR Decomposition

The QR decomposition states that any real $N \times p$ matrix $\mathbf{X}$ can be written as

$$\mathbf{X} = \mathbf{Q}\mathbf{R},$$

where

- $\mathbf{Q}$ is an $N \times p$ orthogonal matrix, meaning $\mathbf{Q}^T = \mathbf{Q}^{-1}$.
- $\mathbf{R}$ is a $p \times p$ upper triangular matrix.

We know of three ways to create the QR decomposition: (a) Gram-Schmidt, (b) Householder reflections, which is very similar to Grover's algorithm, and (c) Givens rotations.

7 Intuition of Linear Regression

Linear regression is defined as

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

However, this does not give us any intuition other than it being the argmax of the residual sum of squares. Here we start with the basics, and we first consider the scenario of univariate regression.

Suppose that we have only one variable and N samples, hence $\mathbf{x} = \mathbf{X} \in \mathbb{R}^{N\times 1}$ where $\mathbf{x} = [x_1, \ldots, x_N]^T$. As always, suppose $\mathbf{y} = [y_1, \ldots, y_N]^T$. What is the least squares estimate in this case? Minimizing the RSS, we get

$$\hat{\beta} = \frac{\sum_{i=1}^N x_i y_i}{\sum_{i=1}^N x_i x_i} = \frac{\langle \mathbf{x}, \mathbf{y}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle},$$

called the *univariate estimate*. Since $\text{proj}_{\mathbf{x}}\mathbf{y} = \hat{\beta}\mathbf{x}$, we see that $\hat{\beta}$ corresponds to the length of the projection of $\mathbf{y}$ on $\mathbf{x}$ as a fraction of the length of $\mathbf{x}$.

Moreover, letting $r_i = y_i - x_i\hat{\beta}$, we get $\mathbf{r} = \mathbf{y} - \mathbf{x}\hat{\beta} = \mathbf{y} - \text{proj}_{\mathbf{x}}\mathbf{y}$. In other words, the residuals are just the vectors that need to be added to the projection to get $\mathbf{y}$. Observe that the residual $\mathbf{r}$ from regressing $\mathbf{y}$ on $\mathbf{x}$ is orthogonal to $\mathbf{x}$, i.e., $\langle \mathbf{r}, \mathbf{x}\rangle = 0$.

7.1 Orthogonal Inputs

Suppose now that we have p variables and hence $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$, with $\mathbf{x}_i \in \mathbb{R}^N$. Suppose, moreover, that $\langle \mathbf{x}_j, \mathbf{x}_k\rangle = 0$ for any $j \neq k$. Then we can see that each $\hat{\beta}_j$ depends only on $\mathbf{x}_j$ and $\mathbf{y}$, and in particular,

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

So for each variable, the corresponding estimate is independent of the other variables.

7.2 General Inputs

In general, the inputs are rarely orthogonal and therefore we cannot conclude that $\hat{\beta}_j$ depends only on $\mathbf{x}_j$. However, if we first orthogonalize the inputs, then this idea carries over. We apply this orthogonalization in the next two subsections.

7.2.1 Univariate Regression with Intercept

Consider a single variable again, so $\mathbf{x} = [x_1, \ldots, x_N]^T$ but that we additionally have an intercept $\hat{\beta}_0$ that we want to compute. Denote $\mathbf{x}_0 = [1, \ldots, 1]^T = \mathbf{1}_N$. Then, if we regress $\mathbf{x}$ on $\mathbf{x}_0$, we get coefficient

$$\hat{\beta}_0 = \frac{\langle \mathbf{x}_0, \mathbf{x}\rangle}{\langle \mathbf{x}_0, \mathbf{x}_0\rangle} = \frac{\sum x_i}{N} = E_i x_i = \overline{x}$$

and residual

$$\mathbf{z} = \mathbf{x} - \overline{x}\mathbf{x}_0 = \mathbf{x} - \overline{x}\mathbf{1}.$$

Finally, to compute $\hat{\beta}_1$, we regress $\mathbf{y}$ on $\mathbf{z}$ and we get

$$\hat{\beta}_1 = \frac{\langle \mathbf{y}, \mathbf{z} \rangle}{\langle \mathbf{z}, \mathbf{z} \rangle} = \frac{\langle \mathbf{y}, \mathbf{x} - \overline{x}\mathbf{1} \rangle}{\langle \mathbf{x} - \overline{x}\mathbf{1}, \mathbf{x} - \overline{x}\mathbf{1} \rangle}$$

So what we did was

1. We adjust $\mathbf{x}$ for $\mathbf{1}$, getting a coefficient $\hat{\beta}_0$ and residual $\mathbf{z} = \mathbf{x} - \hat{\beta}_0\mathbf{1}$.
2. We adjust $\mathbf{y}$ for $\mathbf{z}$ getting a coefficient $\hat{\beta}_1$ and residual $\mathbf{z}' = \mathbf{y} - \hat{\beta}_1\mathbf{z}$ (although we are not interested in the final residual).

7.2.2 Many Variables

So, now we can easily generalize.

Algorithm 1 Regression by Succesive orthogonalization

$\mathbf{x}_0 = \mathbf{z}_0 = \mathbf{1}$
for $j = 1, \dots, p$ **do**
Regress $\mathbf{x}_i$ on all *previous* $\mathbf{z}_0, \dots, \mathbf{z}_{j-1}$ to get coefficients $\hat{\gamma}_{0,j}, \dots, \hat{\gamma}_{j-1,j}$ and residual $\mathbf{z}_j = \mathbf{x}_j - \sum \gamma_{\ell,j}\mathbf{z}_\ell$.
Regress $\mathbf{y}$ on $\mathbf{z}_p$ to get coefficient $\hat{\beta}_p$.

Notice that since all $\mathbf{z}_j$ are orthogonal to each other, they form a basis for the column space of $\mathbf{X}$.

So overall, if we want the estimate $\hat{\beta}_j$, all we need to do is adjust $\mathbf{x}_j$ for $\mathbf{x}_1, \dots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \dots, \mathbf{x}_p$ getting residual $\mathbf{z}_j$, and then adjust $\mathbf{y}$ for $\mathbf{z}_j$.

Intuition on coefficients
The coefficient $\hat{\beta}_j$ represents the information of $\mathbf{y}$ that depends of $\mathbf{x}_j$, after $\mathbf{x}_j$ has been adjusted for the remaining predictors.

Before, we saw that the variance-covariance matrix of the estimates is

$$\mathrm{Var}\left[\hat{\beta}\right] = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2.$$

An alternative formular of the estimates variance is

$$\mathrm{Var}\left[\hat{\beta}_j\right] = \frac{\sigma^2}{\langle \mathbf{z}_j, \mathbf{z}_j \rangle}.$$

Notice that if $\mathbf{x}_j$ is very correlated with one or more other parameters, then the corresponding residual $\mathbf{z}_j$ will be very short. In this case, it would be better to use the Z-score to remove it. Moreover, notice that in this case, its variance will be huge.

7.2.3 Another interpretation of QR Decomposition

If we set $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_p]$ and $\mathbf{\Gamma}$ as the matrix with elements $\hat{\gamma}_{\ell,j}$, then we get

$$\mathbf{X} = \mathbf{Z}\mathbf{\Gamma},$$

where $\mathbf{\Gamma}$ is upper triangular. Notice that although the columns of $\mathbf{Z}$ are orthogonal, they are not also orthonormal. By dividing each row with its length $\|\mathbf{z}_i\|$ we can make them orthonormal. This is expressed by introducing the matrix $\mathbf{D} = \text{diag}(\|\mathbf{z}_0\|, \ldots, \|\mathbf{z}_p\|)$ and having

$$\mathbf{Q} = \mathbf{Z}\mathbf{D}^{-1}$$

and

$$\mathbf{R} = \mathbf{D}\mathbf{\Gamma}.$$

Now, $\mathbf{Q}$ is orthonormal, and so, $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$ and $\mathbf{R}$ remains upper triangular.

Notice that now, we can get another least squares solution by setting to zero the first derivative

$$\begin{aligned} \mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) &= \mathbf{R}^T\mathbf{Q}^T(\mathbf{y} - \mathbf{Q}\mathbf{R}\beta) \\ &= \mathbf{R}^T\mathbf{Q}^T\mathbf{y} - \mathbf{R}^T\mathbf{R}\beta. \end{aligned} \tag{10}$$

We get

$$\mathbf{R}^T\mathbf{R}\beta = \mathbf{R}^T\mathbf{Q}^T\mathbf{y} \implies \mathbf{R}\beta = \mathbf{Q}^T\mathbf{y} \implies \hat{\beta} = \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{y}.$$

Of course, this assumes that $\mathbf{R}^T$ (and therefore $\mathbf{R}$) is invertible. Moreover, the hat matrix can now be computed as

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}\mathbf{R}^{-1}\mathbf{Q}^T\mathbf{y} = \mathbf{Q}\mathbf{R}\mathbf{R}^{-1}\mathbf{Q}^T\mathbf{y} = \mathbf{Q}\mathbf{Q}^T\mathbf{y} = \mathbf{H}\mathbf{y}.$$

8 Principal Components

Let $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ be the SVD decomposition of $\mathbf{X}$. The eigendecomposition of $\mathbf{X}^T\mathbf{X}$ is

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^T\mathbf{D}\mathbf{V}^T,$$

so the column vectors of $\mathbf{V}$ are the eigenvectors of $\mathbf{X}^T\mathbf{X}$ and the corresponding eigenvalues are the elements of the diagonal of $\mathbf{D}^T\mathbf{D}$. The columns $\mathbf{v}_j$ of $\mathbf{V}$ are also called the principal components *directions* of $\mathbf{X}$. The corresponding principal components of $\mathbf{X}$ are then $\mathbf{z}_j = \mathbf{X}\mathbf{v}_j = \mathbf{u}_j d_j$, where d_j are the singular values of $\mathbf{X}$. Moreover, $\mathbf{u}_j$ are the normalized principal components.

Notice that up to a scalar N, the above is also the sample covariance matrix. In other words,

$$\mathbf{S} = \frac{\mathbf{X}^T\mathbf{X}}{N}.$$

Takeaways.

The small singular values correspond to directions in the column space of $\mathbf{X}$ *having small variance. Ridge regression shrinks these directions the most.*

Conversely, the eigenvectors with the largest eigenvalues of the (sample) covariance matrix correspond the directions along which $\mathbf{X}$ *has the maximum variance.*

9 A Few Notes on the Lasso

The lasso solution is:

$$\hat{\beta}^{\mathsf{lasso}} = \arg\min_{\beta} \left(\frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

Equivalently, it is the solution to the following quadratic programming problem:

$$\begin{aligned} \hat{\beta}^{\mathsf{lasso}} = & \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \\ & \text{subject to} \sum_{j=1}^{p} |\beta_j| \leq t. \end{aligned} \tag{11}$$

There is a one-to-one correspondence between λ and t.

Contrary to ridge regression, there is no analytical solution. Nonetheless, it can be found algorithmically with around the same complexity as ridge. This is interesting because the ridge solution can be computed by just multiplying some matrices (or maybe by using a modification of Gram-Schmidt?). So my guess is that the ridge solution can be computed in around $O(N^3)$ time.

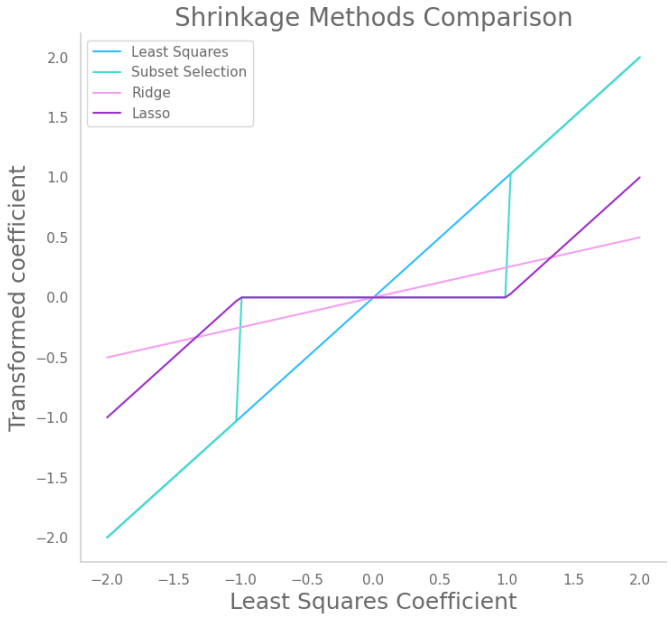
If λ is sufficiently small, some of the coefficients will be 0 but it is not clear why. For an intuition, see here.

Let $\hat{\beta}^{\mathsf{ls}}$ be the least squares coefficients. If $t > \|\hat{\beta}^{\mathsf{ls}}\|_1$, then t is too large and the solution gives again the least squares coefficients. Intuitively, if $t > \|\hat{\beta}^{\mathsf{ls}}\|_1/2$, then the lasso solution is around half the least squares solution; i.e., $\|\hat{\beta}^{\mathsf{lasso}}\|_1 \approx \|\hat{\beta}^{\mathsf{ls}}\|_1/2$. This is interesting. Can we prove this?

10 Subset Selection vs Ridge vs Lasso

Let $\hat{\beta}$ be the least squares solution and assume that the input $\mathbf{X}$ is orthonormal. Then

- M-subset selection keeps the M largest coeefficients and zeroizes all the rest.
- λ-ridge regression, shrinks by a factor of λ: $\hat{\beta}_j^{\text{ridge}} = \hat{\beta}_j/(1+\lambda)$.
- λ-lasso, shortens $\hat{\beta}$ by λ and zeroizes it if it is already short.



We see that the Lasso behaves like subset selection but additionally it shrinks the larger coefficients.

11 Terminology

Likelihood - Prior - Posterior

- The likelihood is the distribution of the data $\mathbf{X}, \mathbf{y}$ given a fixed parameter of the system (for example, β if the system is linear).
- The prior is the distribution of the parameters.
- The posterior is the distribution of the parameters given the data $\mathbf{X}, \mathbf{y}$.

Often it is convenient to work with the (minus) log of these distributions. For example, if the distributions follow some form of normal distribution, we end up working with the (positive) exponents only.