Stat 230. HW 6.

2. Lecture 13, Q7.

7 An equivalent form of ridge coefficient Using the Woodbury formula to show that

$$(X^{\mathrm{T}}X + \lambda I_p)^{-1}X^{\mathrm{T}}Y = X^{\mathrm{T}}(XX^{\mathrm{T}} + \lambda I_n)^{-1}Y.$$

The left-hand side involves inverting a $p \times p$ matrix, and it is more useful when $p < n$; the right-hand side involves inverting an $n \times n$ matrix, so it is more useful when $p > n$.

By Woodbury formula:

$$(\lambda I_p + X^T X)^{-1} = \frac{1}{\lambda}\cdot I_p - \frac{1}{\lambda} I_p \cdot X^T (I_n + X \cdot \frac{1}{\lambda} I_p \cdot X^T)^{-1} X \cdot \frac{1}{\lambda} I_p$$

$$= \frac{1}{\lambda} I_p - \frac{1}{\lambda^2} \cdot X^T (I_n + \frac{1}{\lambda} X \cdot X^T)^{-1} X$$

$$\text{Left} = \frac{1}{\lambda} X^T Y - \frac{1}{\lambda^2} X^T (I_n + \frac{1}{\lambda} X X^T)^{-1} X X^T Y$$

$$= X^T \left(\frac{1}{\lambda} I_n - \frac{1}{\lambda} (X X^T + \lambda I_n)^{-1} X X^T \right) Y$$

By Woodbury formula again. (from right to left)

$$\frac{1}{\lambda} I_n - \frac{1}{\lambda} (XX^T + \lambda I_n)^{-1} XX^T$$

$$= (\lambda I_n + I_n \cdot \frac{1}{\lambda} I_n \cdot \lambda \cdot XX^T)^{-1}$$

$$= (\lambda I_n + XX^T)^{-1}$$

Hence $\text{Left} = X^T (\lambda I_n + XX^T)^{-1} Y = \text{Right}$.

3. Lecture 14. Q1. Prove Lemma 1.

Lemma 1. Given b_0 and λ,

$$\arg\min_{b\in R}\frac{1}{2}(b-b_0)^2+\lambda|b| = \text{sign}(b_0)\left(|b_0|-\lambda\right)_+$$
$$= \begin{cases} b_0-\lambda, & \text{if } b_0\geq\lambda, \\ 0 & \text{if } -\lambda\leq b_0\leq\lambda, \\ b_0+\lambda & \text{if } b_0\leq-\lambda, \end{cases}$$

where sign$(\cdot)$ is the sign of a real number and $(\cdot)_+=\max(\cdot,0)$ is the positive part of a real number.

Let $f(b)=\frac{1}{2}(b-b_0)^2+\lambda|b|$

$$= \begin{cases} \frac{1}{2}b^2+(-b_0+\lambda)b+\frac{1}{2}b_0 & \text{if } b\geqslant 0 \\ \frac{1}{2}b^2+(-b_0-\lambda)b+\frac{1}{2}b_0 & \text{if } b<0 \end{cases}$$

$$\frac{\partial f(b)}{\partial b} = \begin{cases} b-b_0+\lambda & \text{if } b\geqslant 0 \\ b-b_0-\lambda & \text{if } b<0. \end{cases}$$

$$\frac{\partial^2 f(b)}{\partial b^2}=1>0$$

Hence when $b\geqslant 0$. and $b_0-\lambda\geqslant 0$. $b=b_0-\lambda$ minimize $f(b)$

when $b<0$ and $b_0+\lambda<0$ $b=b_0+\lambda$ minimize $f(b)$

when $b_0-\lambda\leqslant 0$ and $b_0+\lambda\geqslant 0$. $b=0$ minimize $f(b)$

That is, $\arg\min_{b\in R}\frac{1}{2}(b-b_0)^2+\lambda|b|=\text{sign}(b_0)\left(|b_0|-\lambda\right)_+$

5. Lecture 15. Q1. Prove (1)

With two binary covariates $F_1, F_2 \in \{0,1\}^n$, we can fit an OLS:

$$Y = \hat\beta_0 + \hat\beta_1 F_1 + \hat\beta_2 F_2 + \hat\beta_{12} F_1 \circ F_2 + \hat\varepsilon,$$

where $F_1 \circ F_2$ denotes the component-wise product between the vectors F_1 and F_2. We can show that

$$\hat\beta_{12} = (\bar y_{11} - \bar y_{10}) - (\bar y_{01} - \bar y_{00}), \tag{1}$$

Suppose $f_{1i} = f_{2i} = 1$ and there are n_{11} observations

$y_{11,i} = \hat\beta_0 + \hat\beta_1 + \hat\beta_2 + \hat\beta_{12} + \hat\varepsilon_{11,i}$.

$$\overline{y_{11}} = \left(\textstyle\sum_{i=1}^{n_{11}} y_{11i}\right)/n_{11} = \hat\beta_0 + \hat\beta_1 + \hat\beta_2 + \hat\beta_{12} + \left(\textstyle\sum_{i=1}^{n_{11}} \hat\varepsilon_{11i}\right)/n_{11}$$

$$= \hat\beta_0 + \hat\beta_1 + \hat\beta_2 + \hat\beta_{12}$$

In the same way, we have

$$\overline{y_{10}} = \hat\beta_0 + \hat\beta_1$$

$$\overline{y_{01}} = \hat\beta_0 + \hat\beta_2$$

$$\overline{y_{00}} = \hat\beta_0$$

$$\Rightarrow (\overline{y_{11}} - \overline{y_{10}}) - (\overline{y_{01}} - \overline{y_{00}}) = \hat\beta_{12}$$

6. Lecture 16, Q6.

6 FWL Theorem in WLS Consider the WLS with weights w_i's. Show that $\hat{\beta}_{w,2}$ in the long WLS fit

$$Y = X_1\hat{\beta}_{w,1} + X_2\hat{\beta}_{w,2} + \hat{\varepsilon}_w$$

equals the coefficient of $\tilde{X}_{w,2}$ in the WLS fit of $\tilde{Y}_w$ on $\tilde{X}_{w,2}$, where $\tilde{X}_{w,2}$ are the residual vectors from the column-wise WLS of X_2 on X_1, and $\tilde{Y}_w$ is the residual vector from the WLS of Y on X_1.

Suppose $Y = (X_1, X_2)\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} + \varepsilon$, $\quad Cov(\varepsilon) = \sigma^2 W^{-1} = \sigma^2 \cdot diag(w_1^{-1}, \cdots w_n^{-1})$

Let $Z_1 = W^{\frac{1}{2}} X_1$, $Z_2 = W^{\frac{1}{2}} X_2$, $\quad Z_3 = W^{\frac{1}{2}} Y$.

Hence $Z_3 = (Z_1, Z_2)\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} + \varepsilon_z$, and $Cov(\varepsilon_z) = \sigma^2 I_n$

And $\hat{\beta}^{OLS}$ of this new model equals $\hat{\beta}^{WLS}$ of the original model.

By FWL Theorem. $\hat{\beta}_{w,2} = (\tilde{Z}_2^T \tilde{Z}_2)^{-1} \tilde{Z}_2^T \tilde{Z}_3$

where $\tilde{Z}_2 = (I_n - H_1) Z_2$, $\tilde{Z}_3 = (I_n - H_1) Z_3$,
$H_1 = Z_1 (Z_1^T Z_1)^{-1} Z_1^T$.

Hence $\tilde{Z}_2 = W^{\frac{1}{2}} \cdot \tilde{X}_{w,2}$ $\quad \tilde{Z}_3 = W^{\frac{1}{2}} \cdot \tilde{Y}_w$.

Hence run WLS fit of $\tilde{Y}_w$ on $\tilde{X}_{w,2}$, the coefficient is:

$$\begin{aligned}\hat{\beta} &= (\tilde{Z}_2^T W^{-\frac{1}{2}} W W^{-\frac{1}{2}} \tilde{Z}_2)^{-1} \cdot \tilde{Z}_2^T W^{-\frac{1}{2}} W W^{-\frac{1}{2}} \tilde{Z}_3 \\ &= (\tilde{Z}_2^T \tilde{Z}_2)^{-1} (\tilde{Z}_2^T \tilde{Z}_3) \\ &= \hat{\beta}_{w2}\end{aligned}$$

7. Lecture 16 Q7.

7 EHW robust covariance estimator in WLS We have shown in Section 1 that the coefficients from WLS are identical to those from OLS with transformed variables. Further show that the HC0 version of EHW covariance estimators are also identical.

Suppose $Y = X\beta + \varepsilon$, $\quad Cov(\varepsilon) = \sigma^2 \cdot W^{-1} = \sigma^2 \cdot diag(w_1^{-1}, \cdots, w_n^{-1})$

Let $\tilde{Y} = W^{\frac{1}{2}} Y$, $\tilde{X} = W^{\frac{1}{2}} X$.

$\tilde{Y} = \tilde{X}\beta + \tilde{\varepsilon}$, $\quad Cov(\tilde{\varepsilon}) = \sigma^2 I_n \quad \tilde{\varepsilon} = W^{\frac{1}{2}}\varepsilon$.

Because $Cov(\tilde{\varepsilon}) = W^{\frac{1}{2}} Cov(\varepsilon) W^{\frac{1}{2}}$

$$
\begin{aligned}
V &= (\tilde{X}^T\tilde{X})^{-1}(\tilde{X}^T \Omega \tilde{X})(\tilde{X}^T\tilde{X})^{-1} \\
&= (X^TWX)^{-1}(X^TW^{\frac{1}{2}} \cdot Cov(\tilde{\varepsilon}) \cdot W^{\frac{1}{2}} X)(X^TWX)^{-1} \\
&= (X^TWX)^{-1}(X^TW\, Cov(\varepsilon) WX)(X^TWX)^{-1} \\
&= \sigma^2 (X^TWX)^{-1}(X^TWX)(X^TWX)^{-1} \\
&= \sigma^2 (X^TWX)^{-1} \\
&= Cov(\hat{\beta}_w)
\end{aligned}
$$

Under HC0 version. $\hat{\Omega} = diag(\hat{\tilde{\varepsilon}}_i^2) = W^{\frac{1}{2}} diag(\hat{\varepsilon}_i^2) W^{\frac{1}{2}}$

Hence
$$
\begin{aligned}
\hat{V} &= (X^TWX)^{-1}(X^TW\, diag(\hat{\varepsilon}_i^2) WX)(X^TWX)^{-1} \\
&= n^{-1}(n^{-1}\textstyle\sum_{i=1}^n w_i x_i x_i^T)^{-1}(n^{-1}\sum_{i=1}^n w_i^2 \hat{\varepsilon}_i^2 x_i x_i^T)(n^{-1}\sum_{i=1}^n w_i x_i x_i^T)^{-1} \\
&= \hat{V}_w \qquad \text{still identical.}
\end{aligned}
$$

8. Lecture 17. Q5. Prove Theorem 1.

Assume that

$$y_i \sim \text{Bernoulli}(q), \tag{4}$$

and

$$x_i \mid y_i = 1 \sim \text{N}(\mu_1, \Sigma), \quad x_i \mid y_i = 0 \sim \text{N}(\mu_2, \Sigma). \tag{5}$$

We can verify that $y_i \mid x_i$ follows a logit model as shown in the theorem below.

Theorem 1. Under (4) and (5), we have $\text{logit}\{\text{pr}(y_i = 1 \mid x_i)\} = \alpha + x_i^{\mathrm{T}}\beta$, where

$$\alpha = \log \frac{q}{1-q} - \frac{1}{2}\left(\mu_1^{\mathrm{T}}\Sigma^{-1}\mu_1 - \mu_0^{\mathrm{T}}\Sigma^{-1}\mu_0\right), \quad \beta = \Sigma^{-1}(\mu_1 - \mu_0).$$

By Bayes Theorem.

$$P(y_i=1 \mid x_i) = \frac{f_{x_i \mid y_i=1}(x_i) \cdot P(y_i=1)}{f_{x_i}(x_i)}$$

$$= \frac{q \cdot \cancel{\frac{1}{\sqrt{(2\pi)^k|\Sigma|}}} \exp\left\{-\frac{1}{2}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)\right\}}{q \cdot \cancel{\frac{1}{\sqrt{(2\pi)^k|\Sigma|}}} \exp\left\{-\frac{1}{2}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)\right\} + (1-q)\cancel{\frac{1}{\sqrt{(2\pi)^k|\Sigma|}}} \exp\left\{-\frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\right\}}$$

$$= \frac{1}{1 + \frac{1-q}{q} \cdot \exp\left\{-\frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1 - \mu_1^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2 + \mu_2^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu_2)\right\}}$$

$$= \frac{1}{1 + \exp(-(\alpha + x_i^T\beta))}$$

with $\alpha = \log\frac{q}{1-q} - \frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2)$

$\beta = \Sigma^{-1}(\mu_1 - \mu_2)$

$\Rightarrow \text{logit}(P(y_i=1 \mid x_i)) = \alpha + x_i^T\beta$