Lecture 11
Weighted Least Squares and Review

07 October 2015

Taylor B. Arnold
Yale Statistics
STAT 312/612

Yale

Notes

- Problem Set #3 - Due today
- Midtern I - In class, next Monday

Goals for today

- Notes on weighted least squares and GLS
- Review of the standard linear regression theory

WLS and GLS

On the problem set, you considered a regression model where the covariance marix of the error terms is known to be proportional to some matrix $V(X)$.

The standard way to solve this problem is to decompose the inverse of V as $C^t C$, and to left multiply the regression problem by C:

$$\begin{aligned} y &= X\beta + \epsilon \\ Cy &= CX\beta + C\epsilon \\ \tilde{y} &= \tilde{X}\beta + \tilde{\epsilon} \end{aligned}$$

The standard way to solve this problem is to decompose the inverse of V as C^tC, and to left multiply the regression problem by C:

$$
\begin{aligned}
y &= X\beta + \epsilon \\
Cy &= CX\beta + C\epsilon \\
\tilde{y} &= \tilde{X}\beta + \tilde{\epsilon}
\end{aligned}
$$

Now, we see that the covariance matrix of the transformed error terms are spherical:

$$
\begin{aligned}
\mathbb{V}(\tilde{\epsilon}|X) &= \mathbb{E}(\tilde{\epsilon}\tilde{\epsilon}^t|X) \\
&= \mathbb{E}(C\epsilon\epsilon^t C^t|X) \\
&= C\mathbb{E}(\epsilon\epsilon^t|X)C^t \\
&= \sigma^2 CVC^t \\
&= \sigma^2 \mathbb{I}_n
\end{aligned}
$$

The (very) important thing to notice about this transformation, is that it does not effect β; the regression vector is exactly the same!

The (very) important thing to notice about this transformation, is that it does not effect β; the regression vector is exactly the same! We have only transformed the data for the purpose of applying ordinary least squares.

The (very) important thing to notice about this transformation, is that it does not effect β; the regression vector is exactly the same! We have only transformed the data for the purpose of applying ordinary least squares.

Therefore $\widehat{\beta}$ and s^2 can be taken directly from the model fit on the tilde versions of the variables.

In particular, prediction can be done as follows (only the colored parts are different):

$$y_{new}|X \in X_{new}\widehat{\beta} \pm t \cdot \sqrt{s^2 \mathrm{diag}\left(V_{new}(X_{new}) + X_{new}(X^t V(X) X)^{-1} X_{new}^t\right)}$$

In particular, prediction can be done as follows (only the colored parts are different):

$$y_{new}|X \in X_{new}\widehat{\beta} \pm t \cdot \sqrt{s^2 \text{diag}\left(V_{new}(X_{new}) + X_{new}(X^t V(X) X)^{-1} X_{new}^t\right)}$$

Notice that we only need the diagonal of $V_{new}(X_{new})$. For prediction, we do not care about the covariance between predictions; only the raw variances matter, and they can be completely different than the variance of the data used for fitting the data.

If the matrix $V(X)$ is diagonal, so only homoskedasticity is broken, there is an even simplier way to approach this problem using weighted least squares.

If the variance of is known to follow the equation:

$$\mathbb{E}(\epsilon\epsilon^t|X) = \sigma^2 \text{diag}(w_1, \ldots w_n)$$

Then C is a diagonal matrix with entries equal to $1/\sqrt{w_i}$, and the tranformed model is just a weighted form of the original:

$$\tilde{y_i} = \frac{y_i}{\sqrt{w_i}}$$
$$\tilde{X_{i,j}} = \frac{X_{i,j}}{\sqrt{w_i}}$$

Review

Format of the exam:

- Six question related to an applied problem
- Six short answers based on theoretical concepts
- No proofs
- Only covers up to contrasts; no hierarchical models
- Calculate t-tests, confidence intervals, F-tests from regression tables

Ordinary least squares

We established that the least squares solution to the model:

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

Yields the solution:

$$\widehat{\beta} = (X^t X)^{-1} X^t y$$

As long as the matrix $X^t X$ is invertable.

Projection matricies

From a geometric interpretation of the least squares estimator, we introduce an important matrix P_X called the *projection matrix*.

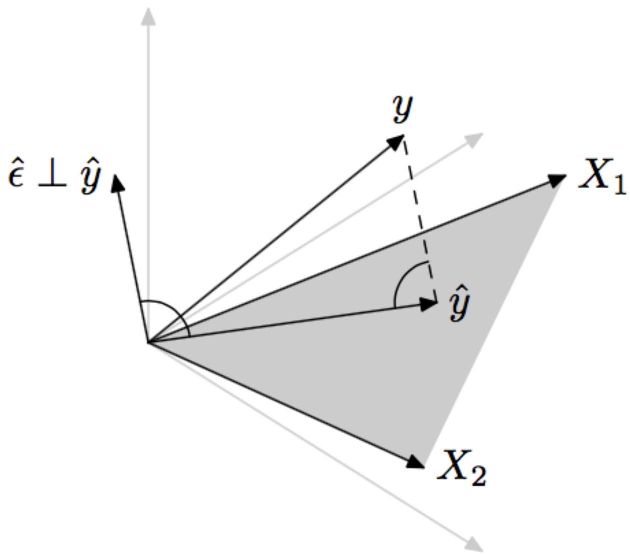
$$P = X(X^t X)^{-1} X^t$$

And the similarly defined annihilator matrix:

$$M = 1 - P$$

We showed the following properties of these matricies:

$$
\begin{aligned}
P^2 &= P^t = P \\
M^2 &= M^t = M \\
PX &= X \\
MX &= 0 \\
Py &= X\beta \\
My &= M\epsilon = r
\end{aligned}
$$



y
$\hat{\epsilon} \perp \hat{y}$
X_1
$\hat{y}$
X_2

Three final definitions

The residuals, estimate of the σ^2 parameter, and sum of squared residuals are given as:

$$
\begin{aligned}
r &= y - X\widehat{\beta} \\
s^2 &= \frac{1}{n-p} r^t r \\
\mathrm{SSR} &= r^t r
\end{aligned}
$$

Classical linear model assumptions

I. Linearity $\quad Y = X\beta + \epsilon$

II. Strict exogeneity $\quad \mathbb{E}\left(\epsilon|X\right) = 0$

III. No multicollinearity $\quad \mathbb{P}\left[\text{rank}(X) = p\right] = 1$

IV. Spherical errors $\quad \mathbb{V}\left(\epsilon|X\right) = \sigma^2 \mathbb{I}_n$

V. Normality $\quad \epsilon|X \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n)$

Finite sample properties

Under assumptions I-III:

(A) $\mathbb{E}(\widehat{\beta}|X) = \beta$

Under assumptions I-IV:

(B) $\mathbb{V}(\widehat{\beta}|X) = \sigma^2(X^tX)^{-1}$

(C) $\widehat{\beta}$ is the best linear unbiased estimator (Gauss-Markov)

(D) $Cov(\widehat{\beta}, r|X) = 0$

(E) $\mathbb{E}(s^2|X) = \sigma^2$

Under assumptions I-V:

(F) $\widehat{\beta}$ achieves the Cramér–Rao lower bound

T-test

Under assumptions $I - V$, to test the hypothesis that $H_0 : \beta = b_j$ we construct the following T-test:

$$
\begin{aligned}
t &= \frac{\widehat{\beta}_j - b_j}{\sqrt{s^2 \left((X^t X)^{-1}_{jj} \right)}} \\
&= \frac{\widehat{\beta}_j - b_j}{\text{S.E.}(\widehat{\beta}_j)} \\
&\sim t_{n-p}
\end{aligned}
$$

There is also a corrisponding confidence interval using the same standard error.

The Hypothesis test $H_0 : D\beta = d$ for a full rank k by p matrix D yields the following **F-test**:

$$F = \frac{(\mathrm{SSR}_R - \mathrm{SSR}_U)/k}{\mathrm{SSR}_U/(n-p)}$$

Where we let SSR_U be the sum of squared residuals of the unrestricted model ($r^t r$) and SSR_R be the sum of squared residuals of the restricted model (where the sum of squares is minimzed subject to $D\beta = d$).

We did a lot of matrix manipulations in the proofs of these two results. The most important 'big picture' results to remember are:

- If B is a symmetric idempotent matrix and $u \sim \mathcal{N}(0, \mathbb{I}_n)$, then $u^t B u \sim \chi^2_{\mathrm{tr(B)}}$.
- If B is a symmetric idempotent matrix, then all of B's eigenvalues are 0 or 1. In terms of the $Q^t \Lambda Q$ eigen-value decomposition, this helps explain why we think of P and M as projection matricies.

```
> out <- lm(Height ~ Father + Gender, data=h)
> summary(out)

Call:
lm(formula = Height ~ Father + Gender, data = h)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3708 -1.4808  0.0192  1.5616  9.4153

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.46113    2.13628   16.13   <2e-16 ***
Father       0.42782    0.03079   13.90   <2e-16 ***
GenderM      5.17604    0.15211   34.03   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.277 on 895 degrees of freedom
Multiple R-squared:  0.5971,  Adjusted R-squared:  0.5962
F-statistic: 663.2 on 2 and 895 DF,  p-value: < 2.2e-16
```

We formally defined leverage as the diagonal elements of the projection matrix:

$$\begin{aligned} l_i &= P_{ii} \\ &= \left[X(X^tX)^{-1}X^t\right]_{ii} \end{aligned}$$

From here, this suggested that we construct the following confidence interval for the mean of y_{new}:

$$\widehat{\mathbb{E}(y_{new}|X)} \in X_{new}\widehat{\beta} \pm t_{n-p,1-\alpha/2} \cdot \sqrt{s^2 X_{new}(X^t X)^{-1} X_{new}^t}$$

Finally, we then constructed the following prediction interval:

$$y_{new}|X \in X_{new}\widehat{\beta} \pm t_{n-p,1-\alpha/2} \cdot \sqrt{s^2 \left[I_k + X_{new}(X^tX)^{-1}X_{new}^t\right]}$$

Which is exactly a factor of s wider than the confidence interval.