Lecture 09

Prediction and Leverage with ASA Flight Data

30 September 2015

Taylor B. Arnold
Yale Statistics
STAT 312/612

Goals for today

- Quantify leverage of an observation
- Formulate prediction and confidence intervals for multivariate regression
- Apply to ASA airline dataset

Leverage

Recall that *leverage* was generally defined as the amount of influence a point has the estimation of $\widehat{\beta}$.

Recall that *leverage* was generally defined as the amount of influence a point has the estimation of $\widehat{\beta}$.

We can formally define leverage as the diagonal elements of the projection matrix:

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

Notice that l_i will be a number between 0 and 1; because P is idempotent and symmetric:

$$l_i = \sum_j p_{i,j} p_{j,i}$$

Notice that l_i will be a number between 0 and 1; because P is idempotent and symmetric:

$$\begin{aligned} l_i &= \sum_j p_{i,j} p_{j,i} \\ &= \sum_j p_{i,j}^2 \end{aligned}$$

Notice that l_i will be a number between 0 and 1; because P is idempotent and symmetric:

$$\begin{aligned} l_i &= \sum_j p_{i,j} p_{j,i} \\ &= \sum_j p_{i,j}^2 \\ &= p_{i,i}^2 + \sum_{j \neq i} p_{i,j}^2 \end{aligned}$$

Notice that l_i will be a number between 0 and 1; because P is idempotent and symmetric:

$$\begin{aligned} l_i &= \sum_j p_{i,j} p_{j,i} \\ &= \sum_j p_{i,j}^2 \\ &= p_{i,i}^2 + \sum_{j \neq i} p_{i,j}^2 \\ &= l_i^2 + \sum_{j \neq i} p_{i,j}^2 \end{aligned}$$

Notice that l_i will be a number between 0 and 1; because P is idempotent and symmetric:

$$\begin{aligned} l_i &= \sum_j p_{i,j} p_{j,i} \\ &= \sum_j p_{i,j}^2 \\ &= p_{i,i}^2 + \sum_{j \neq i} p_{i,j}^2 \\ &= l_i^2 + \sum_{j \neq i} p_{i,j}^2 \end{aligned}$$

So then $l_i \geq l_i^2$, which shows the bounds on the leverage values.

To see why we use the diagonal of the projection matrix, look at the variance of the residuals:

$$\mathbb{V}(r|X) = \mathbb{E}(rr^t|X)$$

To see why we use the diagonal of the projection matrix, look at the variance of the residuals:

$$\begin{aligned}\mathbb{V}(r|X) &= \mathbb{E}(rr^t|X) \\ &= \mathbb{E}(M\epsilon\epsilon^t M^t|X)\end{aligned}$$

To see why we use the diagonal of the projection matrix, look at the variance of the residuals:

$$\begin{aligned}
\mathbb{V}(r|X) &= \mathbb{E}(rr^t|X) \\
&= \mathbb{E}(M\epsilon\epsilon^t M^t|X) \\
&= M \cdot \mathbb{E}(\epsilon\epsilon^t|X) \cdot M^t
\end{aligned}$$

To see why we use the diagonal of the projection matrix, look at the variance of the residuals:

$$\begin{aligned}
\mathbb{V}(r|X) &= \mathbb{E}(rr^t|X) \\
&= \mathbb{E}(M\epsilon\epsilon^t M^t|X) \\
&= M \cdot \mathbb{E}(\epsilon\epsilon^t|X) \cdot M^t \\
&= \sigma^2 M M^t
\end{aligned}$$

To see why we use the diagonal of the projection matrix, look at the variance of the residuals:

$$\begin{aligned}
\mathbb{V}(r|X) &= \mathbb{E}(rr^t|X) \\
&= \mathbb{E}(M\epsilon\epsilon^t M^t|X) \\
&= M \cdot \mathbb{E}(\epsilon\epsilon^t|X) \cdot M^t \\
&= \sigma^2 MM^t \\
&= \sigma^2 \left[\mathbb{I}_n - P\right]
\end{aligned}$$

To see why we use the diagonal of the projection matrix, look at the variance of the residuals:

$$
\begin{aligned}
\mathbb{V}(r|X) &= \mathbb{E}(rr^t|X) \\
&= \mathbb{E}(M\epsilon\epsilon^t M^t|X) \\
&= M \cdot \mathbb{E}(\epsilon\epsilon^t|X) \cdot M^t \\
&= \sigma^2 MM^t \\
&= \sigma^2 \left[\mathbb{I}_n - P\right]
\end{aligned}
$$

So the variance of an individual residual is $(1 - l_i)$, so for a leverage close to 1 the regression line will generally pass very close to the point i.

The individual variance of the i'th residual suggests that we could standardize each residual as such:

$$\frac{r_i}{\sqrt{s^2(1 - l_i)}}$$

The individual variance of the i'th residual suggests that we could standardize each residual as such:

$$\frac{r_i}{\sqrt{s^2(1 - l_i)}}$$

This is known as the Studentized residual. If s^2 is modifed to be calculated without the point i, externally studentized, then this quantity follows a t distribution with $n - p$ degrees of freedom. (Note: This is actually very easy to prove given our already established results.)

Confidence and Prediction Intervals

Now, we consider the case where we observe a new set of observations that were not used in the estimation of $\widehat{\beta}$:

$$y_{new} = X_{new}\beta + \epsilon_{new}$$

Often we do not actually observe the new values of y, but wish to estimate them from the estimate of β and the new data points.

An obvious guess of $\hat{y}_{new}$ is $X_{new}\widehat{\beta}$.

An obvious guess of $\hat{y}_{new}$ is $X_{new}\widehat{\beta}$. We see easily that:

$$\begin{aligned} \mathbb{E}(\hat{y}_{new}|X) &= \mathbb{E}(X_{new}\widehat{\beta}|X) \\ &= X_{new}\beta \end{aligned}$$

Where the conditional on X is with respect to the original data and the new data matrix X_{new}.

What would we do if we needed a confidence interval for where the values y_{new} should be located?

What would we do if we needed a confidence interval for where the values y_{new} should be located? We need to calculate the variance of our estimator:

$$\begin{aligned}
\mathbb{V}(\hat{y}_{new}|X) &= \mathbb{V}(X_{new}\widehat{\beta}|X) \\
&= X_{new}\mathbb{V}(\widehat{\beta}|X)X_{new}^t \\
&= \sigma^2 X_{new}(X^t X)^{-1} X_{new}^t
\end{aligned}$$

Notice that in the special case that row j of X_{new} is equal to row i of X, we have:

$$\begin{aligned}\mathbb{V}([\hat{y}_{new}]_j \,|X) &= \sigma^2 P_{ii} \\ &= \sigma^2 l_i\end{aligned}$$

So points with high leverage are points where predictions are particularly variable.

Notice that in the special case that row j of X_{new} is equal to row i of X, we have:

$$\begin{aligned} \mathbb{V}([\hat{y}_{new}]_j | X) &= \sigma^2 P_{ii} \\ &= \sigma^2 l_i \end{aligned}$$

So points with high leverage are points where predictions are particularly variable. This may seem counterintuitive.

This suggests that we construct the following confidence interval for the mean of y_{new}:

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

Typically, we are interested in an interval for the actually observations y_{new} rather than the mean of y_{new}.

Typically, we are interested in an interval for the actually observations y_{new} rather than the mean of y_{new}.

To calculate the variance of the actual prediction, see that (k is the number of row of X_{new})

$$\begin{aligned}
\mathbb{V}(y_{new} - \hat{y}_n ew|X) &= \mathbb{V}(y_{new}|X) + \mathbb{V}(\hat{y}_{new}|X) \\
&= \sigma^2 I_k + \sigma^2 X_{new}(X^t X)^{-1} X_{new}^t \\
&= \sigma^2 \left[I_k + \sigma^2 X_{new}(X^t X)^{-1} X_{new}^t \right]
\end{aligned}$$

From here, we now have 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.

Application to ASA Data