Midterm Review Notes

Nickhil Sethi

Ordinary Linear Regression

estimate parameters

In OLS, we estimate the parameters by choosing β_0, β_1 to minimize the sum of squares.

$$\min_{\beta_0,\beta_1} \sum_i (Y_i - (\beta_0 + \beta_1 X_i))^2$$

That is, the sum of squared differences between our prediction and the true values Y_i. For OLS, closed form solutions exist for the parameter estimates:

$$\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}$$
$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X}$$

estimate variance

The variance of the noise ϵ is estimated based on the parameters β_0, β_1 and the data set $\{X_i, Y_i\}$:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$
$$\hat{\sigma}^2 = \frac{\sum_i (\hat{Y}_i - Y_i)^2}{n-2}$$

Notice the denominator is $n-2$ – the "degrees of freedom", i.e. the number of data points minus the number of estimated parameters. This estimator is unbiased, meaning $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$.

least squares vs MLE

Least squares and MLE estimation for linear regression are very similar; under the assumption of gaussian noise, they are nearly equivalent, with the only difference being the estimation of the sample variance.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$
$$\hat{\sigma}^2_{MLE} = \frac{\sum_i (\hat{Y}_i - Y_i)^2}{n}$$

Notice the difference between the MLE estimator of the sample variance and the least squares estimator of the sample variance. The MLE estimator is biased i.e. $\mathbb{E}(\hat{\sigma}^2) \neq \sigma^2$; however, the bias term decays to zero as the number of data points n increases.

residuals and their properties

The residuals are defined as $\varepsilon_i = \hat{y}_i - y_i$. Of course, the loss is defined as the sum of squared residuals:

$$\sum_i \varepsilon_i = 0$$
$$\sum_i \varepsilon_i \cdot X_i = 0$$

From the above properties, it's very simple to show the following additional properties:

$$\sum_i \hat{Y}_i = \sum Y_i$$
$$\sum_i \varepsilon_i \cdot \hat{Y}_i = 0$$

Typically, we assume the noise is gaussian, in which case the residuals should be roughly normally distributed. We can check this via a q-q plot.

Confidence intervals vs Prediction Intervals

A confidence interval estimates a range of plausible values of a parameter of interest, e.g. $\mathrm{E}\left(Y_i|X_i\right)$; a prediction interval, on the other hand, tries to estimate a range of plausible variables for a single out of sample point e.g. the a new Y_i. The difference lies in the fact that a prediction interval has to account for the variance in the distribution of Y_i.

The distribution for an single out of sample $\hat{Y}_i$ would be gaussian with the following mean and variance

$$\begin{aligned}
\mathrm{E}[\hat{Y}_i] &= \beta_0 + \beta_1 X_i \\
\sigma^2[\hat{Y}_i] &= \hat{\sigma}^2[1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_j (X_j - \bar{X})^2}]
\end{aligned}$$

Diagnostics (R squared)

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y} - y_i)^2}$$

- Multiple Variables
- Qualitative

-AIC

Bernoiulli/Binomial Data

- link functions
- predictions
- odds, probability
- risk ratio vs odds ratio
- goodness of fit
- Diagnostics (Pearson Residuals)
- different scoring functions
- confusion matrix and properties (sensitivity, specificity, PPV, accuracy, NPV)

- comparing nested models
- overdispersion
- f statistics
- quasibinomial