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 for β_1 is an interval-valued function of the data; intuitively, a confidence interval is supposed to provide a range of reasonable values for the parameter of interest.

- Here's how we construct a confidence interval for β_1:

$$\hat{\beta}_1 + \hat{\sigma} \cdot t(\frac{1-\alpha}{2}, n-2)$$
$$\hat{\beta}_1 + \hat{\sigma} \cdot t(\frac{1+\alpha}{2}, n-2)$$

- Interpretation: What is meant by "95%"? The model assume the data is distributed like as follows

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Confidence intervals are computed with respect to fixed X_i but varying ε_i. So when we refer to a "95% confidence interval for β_1", we mean over 100 datasets that share the same $\{X_i\}$, roughly 95 of the confidence intervals contain the true parameter β_1.

- Prediction intervals: whereas a confidence interval estimates a range of plausible values of a parameter of interest, e.g. $\mathrm{E}(Y_i|X_i)$; 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:

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

Notice the difference between the above and the *confidence* interval for Y_i:

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

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

If we have more than one predictor, we can still find closed form solutions for all our coefficients. The form is now:

$$X_i = (1, x_{i,1}, x_{i,2}, ...x_{i,k})$$
$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Keep in mind that ading a variable to a regression can change ALL of the regression coefficients.

Qualitative

- Qualitative variables (e.g. species, race) can be represented by indicators i.e. a variable that takes value 1 if a condition is met and zero otherwise.
- Keep in mind R chooses a reference point, meaning if we have k categories in a variable, R will create $k-1$ indicator variables and the kth category will be represented by all the others taking a value of zero.

AIC

- The AIC, short for Akaike-Information Criterion, is a measure of fit of a model that balances goodness of fit (i.e. log likelihood) against the number of variables in the model q

$$L(\beta) = \Pi_i \mathrm{P}(y_i | x_i, \beta)$$
$$\mathrm{AIC(lmod)} = 2 \cdot \log(L(\hat{\beta})) + q$$

- This is a very general and powerful way of comparing models.

Bernoulli/Binomial Data

A word on terminology. Make sure not to confuse bernoulli with binomial! They're closely related but different concepts.

- A bernoulli random variable models a binary outcome, e.g. "yes" or "no", "success" or "failure", "male" or "female" etc. It takes values 0 or 1. A bernoulli random variable has one parameter – the "success probability" p. The distribution is:

$$\mathrm{bernoulli(p)} = p^k (1-p)^{1-k}$$

- A binomial random variable models the number of successes in m independent and identically distributed bernoulli random variables. It has parameters m the total number of trials, p the success probability of any given trial.
- A binomial random variable can take values from 0 to m, inclusive.

$$P(Y = k) = \binom{m}{k} \cdot p^k \cdot (1-p)^{m-k}$$

- Binomial data is data where the response variable is a binomial random variable. For example, the input might be economic data at the county-level, and the response might be the number of people in the county who voted democrat in the most recent election (out of the total number of people in the county).

- Note in binomial regression we need to know both the parameter m_i and the number of successes k_i for each data point i.
- Bernoulli data can be thought of as binomial data where the $m = 1$

link functions

The problem with using basic linear regression to model binomial and bernoulli data is that the linear predictor $\sum_i \beta_i * x_i$ is unbounded – it isn't guaranteed to fall in $0, 1$ or $0, m$.

predictions

- Pass the odds through the inverse-logit function to get the predicted probability p_i

odds, probability

- Odds is a different way of parametrizing of probability. The odds of an event with probability p is $p/1 - p$.
- The odds concept is useful in thinking about bernoulli, binomial regression because the the linear predictor is the log of the odds:

$$\log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

risk ratio vs odds ratio

- An odds ratio is a ratio of the odds of two events $odds(A)/odds(B)$.
- Because the linear predictor models the log odds for each data point, the success odds of data point i are:

$$odds_i \,|X_i = e^{\beta_0} \,\Pi_k \, e^{\beta_1 X_{i,k}}$$

- The odds ratio of success given two levels of predictor $X_{i,k}$, say s, t is

$$\text{odds\_ratio}_i(X_i = s, X_i = t) = e^{\beta_k(s-t)}$$

- A risk ratio is the ratio probabilities given two events

$$\text{risk\_ratio(A, B)} = P(A)/P(B)$$

goodness of fit - Under the assumption that a model is correct, the deviance (i.e. the log-likelihood) is distributed χ^2_{n-q-1}; a low p-value under this distribution indicates a poor-fit.

- We can assess goodness of fit with a calibration curve, which plots $empirical_probability$ against $pred_probability$. Empirical probabilities are determined via binning. Ideally the curve would follow the diagonal $y = x$

Diagnostics (Pearson Residuals)

- The pearson residual is:

$$\begin{aligned} r_i &= \frac{y_i - n_i\hat{p}_i}{\sqrt{\text{var}(\hat{y}_i)}} \\ &= \frac{y_i - n_i\hat{p}_i}{n_i\hat{p}_i * (1 - \hat{p}_i)} \end{aligned}$$

- the log likelihood can be thought of as the sum of the residuals, analagous to least squares

different scoring functions

- likelihood
- brier score:

$$\frac{1}{n}\sum_i (\hat{p}_i - y_i)^2$$

confusion matrix and properties (sensitivity, specificity, PPV, accuracy, NPV)

comparing nested models

- Given two models S and L (for small and large), where the predictor set of S is a subset of the predictor set of L, the deviance is chi squared distributed under the assumption that the smaller is correct with degrees of freedom $l - s$

overdispersion

f statistics

quasibinomial