Lecture 22: Review for Exam 2

1 Basic Model Assumptions (without Gaussian Noise)

We model one continuous response variable Y, as a linear function of p numerical predictors, plus noise:

$$Y = \beta_0 + \beta_1 X_1 + \ldots \beta_p X_p + \epsilon. \tag{1}$$

Linearity is an assumption, which can be wrong. Further assumptions take the form of restrictions on the noise:

$$\mathbb{E}\left[\epsilon|X\right] = 0, \quad \mathrm{Var}\left[\epsilon|X\right] = \sigma^2.$$

Moreover, we assume ϵ is uncorrelated across observations.

We convert this to matrix form:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{2}$$

$\mathbf{Y}$ is an $n \times 1$ matrix of random variables; $\mathbf{X}$ is an $n \times (p+1)$ matrix, with an extra column of all 1s; ϵ is an $n \times 1$ matrix. Beyond linearity, the assumptions translate to

$$\mathbb{E}\left[\epsilon|\mathbf{X}\right] = \mathbf{0}, \quad \mathrm{Var}\left[\epsilon|\mathbf{X}\right] = \sigma^2\mathbf{I}. \tag{3}$$

We don't know β. If we guess it is $\mathbf{b}$, we will make an $n \times 1$ vector of predictions $\mathbf{Xb}$ and have an $n \times 1$ vector of errors $\mathbf{Y} - \mathbf{Xb}$. The mean squared error, as a function of $\mathbf{b}$, is then

$$MSE(\mathbf{b}) = \frac{1}{n}(\mathbf{Y} - \mathbf{Xb})^T(\mathbf{Y} - \mathbf{Xb}). \tag{4}$$

2 Least Squares Estimation and Its Properties

The least squares estimate of the coefficients is the one which minimizes the MSE:

$$\widehat{\beta} \equiv \operatorname*{argmin}_{\mathbf{b}} MSE(\mathbf{b}). \tag{5}$$

To find this, we need the derivatives:

$$\nabla_{\mathbf{b}} MSE = \frac{2}{n}(\mathbf{X}^T\mathbf{Y} - \mathbf{X}^T\mathbf{Xb}). \tag{6}$$

We set the derivative to zero at the optimum:

$$\frac{1}{n}\mathbf{X}^T\left(\mathbf{Y} - \mathbf{X}\widehat{\beta}\right) = \mathbf{0}. \tag{7}$$

The term in parentheses is the vector of errors when we use the least-squares estimate. This is the vector of residuals,

$$\mathbf{e} \equiv \mathbf{Y} - \mathbf{X}\widehat{\beta} \tag{8}$$

so the have the **normal**, **estimating** or **score** equations,

$$\frac{1}{n}\mathbf{X}^T\mathbf{e} = \mathbf{0}. \tag{9}$$

We say "equations", plural, because this is equivalent to the set of $p+1$ equations

$$\frac{1}{n}\sum_{i=1}^{n} e_i = 0 \tag{10}$$

$$\frac{1}{n}\sum_{i=1}^{n} e_i X_{ij} = 0 \tag{11}$$

(Many people omit the factor of $1/n$.) This tells us that while $\mathbf{e}$ is an n-dimensional vector, it is subject to $p+1$ linear constraints, so it is confined to a linear subspace of dimension $n-p-1$. Thus $n-p-1$ is the number of **residual degrees of freedom**.

The solution to the estimating equations is

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{12}$$

This is one of the two most important equations in the whole subject. It says that the coefficients are a linear function of the response vector $\mathbf{Y}$.

The least squares estimate is a constant plus noise:

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{13}$$

$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\beta+\epsilon) \tag{14}$$

$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon \tag{15}$$

$$= \beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon. \tag{16}$$

The least squares estimate is unbiased:

$$\mathbb{E}\left[\widehat{\beta}\right] = \beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}\left[\epsilon\right] = \beta. \tag{17}$$

Its variance is

$$\mathrm{Var}\left[\widehat{\beta}\right] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}. \tag{18}$$

Since the entries in $\mathbf{X}^T\mathbf{X}$ are usual proportional to n, it can be helpful to write this as

$$\mathrm{Var}\left[\widehat{\beta}\right] = \frac{\sigma^2}{n}\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)^{-1}. \tag{19}$$

The variance of any one coefficient estimator is

$$\mathrm{Var}\left[\widehat{\beta}_i\right] = \frac{\sigma^2}{n}\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)^{-1}_{i+1,i+1}. \tag{20}$$

The vector of fitted means or conditional values is

$$\widehat{\mathbf{Y}} \equiv \mathbf{X}\widehat{\beta}. \tag{21}$$

This is more conveniently expressed in terms of the original matrices:

$$\widehat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}\mathbf{Y}. \tag{22}$$

The fitted values are thus linear in $\mathbf{Y}$: set the responses all to zero and all the fitted values will be zero; double all the responses and all the fitted values will double.

The $n \times n$ **hat matrix** $\mathbf{H} \equiv \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, also called the influence, projection or prediction matrix, controls the fitted values. It is a function of $\mathbf{X}$ alone, ignoring the response variable totally. It is an $n \times n$ matrix with several important properties:

- It is symmetric, $\mathbf{H}^T = \mathbf{H}$.
- It is idempotent, $\mathbf{H}^2 = \mathbf{H}$.
- Its trace $\mathrm{tr}\,\mathbf{H} = \sum_i H_{ii} = p + 1$, the number of degrees of freedom for the fitted values.

The variance-covariance matrix of the fitted values is

$$\mathrm{Var}\left[\widehat{\mathbf{Y}}\right] = \mathbf{H}\sigma^2\mathbf{I}\mathbf{H}^T = \sigma^2\mathbf{H}. \tag{23}$$

To make a prediction at a new point, not in the data used for estimation, we take its predictor coordinates and group them into a $1 \times (p+1)$ matrix $\mathbf{X}_{new}$ (including the 1 for the intercept). The point prediction for Y is then $\mathbf{X}_{new}\widehat{\beta}$. The expected value is $\mathbf{X}_{new}\beta$, and the variance is $\mathrm{Var}\left[\mathbf{X}_{new}\widehat{\beta}\right] = \mathbf{X}_{new}\mathrm{Var}\left[\widehat{\beta}\right]\mathbf{X}_{new}^T = \sigma^2\mathbf{X}_{new}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{new}^T$.

The residuals are also linear in the response:

$$\mathbf{e} \equiv \mathbf{Y} - \widehat{\mathbf{m}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}. \tag{24}$$

The trace of $\mathbf{I} - \mathbf{H}$ is $n - p - 1$. The variance-covariance matrix of the residuals is

$$\mathrm{Var}\left[\mathbf{e}\right] = \sigma^2(\mathbf{I} - \mathbf{H}). \tag{25}$$

The mean squared error (training error) is

$$MSE = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\mathbf{e}^T\mathbf{e}. \tag{26}$$

Its expectation value is slightly below σ^2:

$$\mathbb{E}\left[MSE\right] = \sigma^2\frac{n-p-1}{n}. \tag{27}$$

(This may be proved using the trace of $\mathbf{I} - \mathbf{H}$.) An unbiased estimate of σ^2, which I will call $\widehat{\sigma}^2$ throughout the rest of this, is

$$\widehat{\sigma}^2 \equiv MSE\frac{n}{n-p-1}. \tag{28}$$

The **leverage** of data point i is H_{ii}. This has several interpretations:

1. $\mathrm{Var}\left[\widehat{Y}_i\right] = \sigma^2 H_{ii}$; the leverage controls how much variance there is in the fitted value.
2. $\partial\widehat{Y}_i/\partial Y_i = H_{ii}$; the leverage says how much changing the response value for point i changes the fitted value there.

3. $\text{Cov}\left[\widehat{Y}_i, Y_i\right] = \sigma^2 H_{ii}$; the leverage says how much covariance there is between the i^{th} response and the i^{th} fitted value.

4. $\text{Var}\left[e_i\right] = \sigma^2(1 - H_{ii})$; the leverage controls how big the i^{th} residual is.

The **standardized residual** is

$$r_i = \frac{e_i}{\widehat{\sigma}\sqrt{1 - H_{ii}}}. \tag{29}$$

The only restriction we have to impose on the predictor variables X_i is that $(\mathbf{X}^T\mathbf{X})^{-1}$ needs to exist. This is equivalent to

- $\mathbf{X}$ is not **collinear**: none of its columns is a linear combination of other columns; which is also equivalent to
- The eigenvalues of $\mathbf{X}^T\mathbf{X}$ are all > 0. (If there are zero eigenvalues, the corresponding eigenvectors indicate linearly-dependent combinations of predictor variables.)

Nearly-collinear predictor variables tend to lead to large variances for coefficient estimates, with high levels of correlation among the estimates.

It is perfectly OK for one column of $\mathbf{X}$ to be a function of another, provided it is a nonlinear function. Thus in **polynomial** regression we add extra columns for powers of one or more of the predictor variables. (Any other nonlinear function is however also legitimate.) This complicates the interpretation of coefficients as slopes, just as though we had done a transformation of a column. Estimation and inference for the coefficients on these predictor variables goes exactly like estimation and inference for any other coefficient.

One column of $\mathbf{X}$ could be a (nonlinear) function of two or more of the other columns; this is how we represent **interactions**. Usually the interaction column is just a product of two other columns, for a **product** or **multiplicative** interaction; this also complicates the interpretation of coefficients as slopes. (See the notes on interactions.) Estimation and inference for the coefficients on these predictor variables goes exactly like estimation and inference for any other coefficient.

We can include qualitative predictor variables with k discrete categories or levels by introducing binary indicator variables for $k - 1$ of the levels, and adding them to $\mathbf{X}$. The coefficients on these indicators tell us about amounts that are added (or subtracted) to the response for every individual who is a member of that category or level, compared to what would be predicted for an otherwise-identical individual in the baseline category. Equivalently, every category gets its own intercept. Estimation and inference for the coefficients on these predictor variables goes exactly like estimation and inference for any other coefficient.

Interacting the indicator variables for categories with other variables gives coefficients which say what amount is added to the *slope* used for each member of that category (compared to the slope for members of the baseline level). Equivalently, each category gets its own slope. Estimation and inference for the coefficients on these predictor variables goes exactly like estimation and inference for any other coefficient.

Model selection for prediction aims at picking a model which will predict well on new data drawn from the same distribution as the data we've seen. One way to estimate this out-of-sample performance is to look at what the expected squared error would be on new data with the same $\mathbf{X}$

matrix, but a new, independent realization of $\mathbf{Y}$. In the notes on model selection, we showed that

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}(\mathbf{Y}' - \widehat{m})^T(\mathbf{Y}' - \widehat{m})\right] &= \mathbb{E}\left[\frac{1}{n}(\mathbf{Y} - \widehat{m})^T(\mathbf{Y} - \widehat{m})\right] + 2\frac{1}{n}\sum_{i=1}^{n}\mathrm{Cov}\left[Y_i, \widehat{m}_i\right] && (30)\\
&= \mathbb{E}\left[\frac{1}{n}(\mathbf{Y} - \widehat{m})^T(\mathbf{Y} - \widehat{m})\right] + \frac{2}{n}\sigma^2 \,\mathrm{tr}\, H && (31)\\
&= \mathbb{E}\left[\frac{1}{n}(\mathbf{Y} - \widehat{m})^T(\mathbf{Y} - \widehat{m})\right] + \frac{2}{n}\sigma^2(p+1). && (32)
\end{aligned}$$

Mallow's C_p estimates this by

$$MSE + \frac{2}{n}\widehat{\sigma}^2(p+1) \tag{33}$$

using the $\widehat{\sigma}^2$ from the largest, model being selected among (which includes all the other models as special cases). An alternative is leave-one-out cross-validation, which amounts to

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{e_i}{1-H_{ii}}\right)^2. \tag{34}$$

We also considered K-fold cross-validation, AIC and BIC.

3 Gaussian Noise

The Gaussian noise assumption is added on to the other assumptions already made. It is that $\epsilon_i \sim N(0, \sigma^2)$, independent of the predictor variables and all other ϵ_j. In other words, ϵ has a multivariate Gaussian distribution,

$$\epsilon \sim MVN(\mathbf{0}, \sigma^2\mathbf{I}). \tag{35}$$

Under this assumption, it follows that, since $\widehat{\beta}$ is a linear function of ϵ, it also has a multivariate Gaussian distribution:

$$\widehat{\beta} \sim MVN(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}) \tag{36}$$

and

$$\widehat{\mathbf{Y}} \sim MVN(\mathbf{X}\beta, \sigma^2\mathbf{H}). \tag{37}$$

It follows from this that

$$\widehat{\beta}_i \sim N(\beta_i, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}_{i+1,i+1} \tag{38}$$

and

$$\widehat{Y}_i \sim N(\mathbf{X}_i\beta, \sigma^2 H_{ii}). \tag{39}$$

The sampling distribution of the estimated conditional mean at a new point $\mathbf{X}_{new}$ is

$$N(\mathbf{X}_{new}\beta, \sigma^2\mathbf{X}_{new}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{new}^T).$$

The mean squared error follows a χ^2 distribution:

$$\frac{nMSE}{\sigma^2} \sim \chi^2_{n-p-1}. \tag{40}$$

Moreover, the MSE is statistically independent of $\widehat{\beta}$. We may therefore define

$$\widehat{\text{se}}\left[\widehat{\beta}_i\right] = \widehat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{i+1,i+1}} \tag{41}$$

and

$$\widehat{\text{se}}\left[\widehat{Y}_i\right] = \widehat{\sigma}\sqrt{H_{ii}} \tag{42}$$

and get t distributions:

$$\frac{\widehat{\beta}_i - \beta_i}{\widehat{\text{se}}\left[\widehat{\beta}_i\right]} \sim t_{n-p-1} \approx N(0,1) \tag{43}$$

and

$$\frac{\widehat{Y}_i - m(X_i)}{\widehat{\text{se}}\left[\widehat{m}_i\right]} \sim t_{n-p-1} \approx N(0,1). \tag{44}$$

The **Wald test** for the hypothesis that $\beta_i = \beta_i^*$ therefore forms the test statistic

$$\frac{\widehat{\beta}_i - \beta_i^*}{\widehat{\text{se}}\left[\widehat{\beta}_i\right]} \tag{45}$$

and rejects the hypothesis if it is too large (above or below zero) compared to the quantiles of a t_{n-p-1} distribution. The `summary` function of R runs such a test of the hypothesis that $\beta_i = 0$. There is nothing magic or even especially important about testing for a 0 coefficient, and the same test works for testing whether a slope = 42 (for example).

Important! The null hypothesis being test is

> Y is a linear function of $X_1, \ldots X_p$, and of no other predictor variables, with independent, constant-variance Gaussian noise, and the coefficient $\beta_i = 0$ exactly.

and the alternative hypothesis is

> Y is a linear function of $X_1, \ldots X_p$, and of no other predictor variables, with independent, constant-variance Gaussian noise, and the coefficient $\beta_i \neq 0$.

The Wald test does not test any of the model assumptions (it presumes them all), and it cannot say whether in an absolutely sense X_i matters for Y; adding or removing other predictors can change whether the true $\beta_i = 0$.

Warning! Retaining the null hypothesis $\beta_i = 0$ can happen if *either* the parameter is precisely estimated, and confidently known to be close to zero, or if it is *im*-precisely estimated, and might as well be zero or something huge on either side. Saying "We can ignore this because we can be quite sure it's small" can make sense; saying "We can ignore this because we have no idea what it is" is preposterous.

To test whether several coefficients $(\beta_j : \ j \in S)$ are all simultaneously zero, use an F test. The null hypothesis is

$$H_0 : \beta_j = 0 \ \text{ for all } j \in S$$

and the altermative is

$$H_1 : \beta_j \neq 0 \text{ for at least one } j \in S.$$

The F statistic is

$$F_{stat} = \frac{(\widehat{\sigma}^2_{null} - \widehat{\sigma}^2_{full})/s}{\widehat{\sigma}^2_{full}/(n-p-1)} \tag{46}$$

where s is the number of elements in S. Under that null hypothesis,

$$F_{stat} \sim F_{s,n-p-1} \tag{47}$$

If we are testing a subset of coefficients, we have a "partial" F test. A "full" F test sets $s = p$, i.e., it tests the null hypothesis of an intercept-only model (with independent, constant-variance Gaussian noise) against the alternative of the linear model on $X_1, \ldots X_p$ (and only those variables, with independent, constant-variance Gaussian noise). This is only of interest under very unusual circumstances.

Once again, no F test is capable of checking any modeling assumptions. This is because both the null hypothesis and the alternative hypothesis presume that the all of the modeling assumptions are exactly correct.

A $1-\alpha$ confidence interval for β_i is

$$\widehat{\beta}_i \pm \widehat{\mathrm{se}}\left[\beta_i\right] t_{n-p-1}(\alpha/2) \approx \widehat{\beta}_i \pm \widehat{\mathrm{se}}\left[\beta_i\right] z_{\alpha/2}. \tag{48}$$

We saw how to create a confidence ellipsoid for several coefficients. These make a *simultaneous* guarantee: all the parameters are trapped inside the confidence region with probabiluty $1-\alpha$. A simpler way to get a simultaneous confidence region for all p parameters is to use $1-\alpha/p$ confidence intervals for each one ("Bonferroni correction"). This gives a confidence hyper-rectangle.

A $1-\alpha$ confidence interval for the regression function at a point is

$$\widehat{m}(X_i) \pm \widehat{\mathrm{se}}\left[\widehat{m}(X_i)\right] t_{n-p-1}(\alpha/2). \tag{49}$$

Residuals. The cross-validated or studentized residuals are:

1. Temporarily hold out data point i
2. Re-estimate the coefficients to get $\widehat{\beta}^{(-i)}$ and $\widehat{\sigma}^{(-i)}$.
3. Make a prediction for Y_i, namely, $\widehat{Y}_{i(i)} = \widehat{m}^{(-i)}(X_i)$.
4. Calculate

$$t_i = \frac{Y_i - \widehat{Y}_{i(i)}}{\widehat{\sigma}^{(-i)} + \widehat{\mathrm{se}}\left[\widehat{m}_i^{(-i)}\right]}. \tag{50}$$

This can be done without recourse to actually re-fitting the model:

$$t_i = r_i \sqrt{\frac{n-p-1}{n-p-r_i^2}} \tag{51}$$

(Note that for large n, this is typically extremely close to r_i.) Also,

$$t_i \sim t_{n-p-2} \tag{52}$$

(The -2 is because we're using $n-1$ data points to estimate $p+1$ coefficients.)

Cook's distance for point i is the sum of the (squared) changes to all the fitted values if i was omitted; it is

$$D_i = \frac{1}{p+1} e_i^2 \frac{H_{ii}}{(1-H_{ii})^2}. \tag{53}$$