Lecture 15: Diagnostics and Inference for Multiple Linear Regression

1 Review

In the multiple linear regression model, we assume that the response Y is a linear function of all the predictors, plus a constant, plus noise:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p + \epsilon. \tag{1}$$

The number of coefficients is $q = p + 1$.

We make no assumptions about the (marginal or joint) distributions of the X_i, but we assume that $\mathbb{E}[\epsilon|X] = 0$, $\mathrm{Var}[\epsilon|X] = \sigma^2$, and that ϵ is uncorrelated across measurements. In matrix form, the model is

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

where $\mathbf{X}$ is an $n \times q$ matrix that includes an initial column of all 1's. Remember that $q = p + 1$. When we add the Gaussian noise assumption, we are making all of the assumptions above, and further assuming that

$$\epsilon \sim MVN(\mathbf{0}, \sigma^2 \mathbf{I}) \tag{3}$$

independently of $\mathbf{X}$.

The least squares estimate of the coefficients is

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{4}$$

Under the Gaussian noise assumption, this is also the maximum likelihood estimate.

The fitted values (i.e., estimates of the conditional means at data points used to estimate the model) are given by the "hat" or "influence" matrix:

$$\widehat{\mathbf{Y}} \equiv \widehat{\mathbf{m}} = \mathbf{X}\widehat{\beta} = \mathbf{H}\mathbf{Y} \tag{5}$$

which is symmetric and idempotent. The residuals are given by

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y} \tag{6}$$

and $\mathbf{I} - \mathbf{H}$ is also symmetric and idempotent.

The expected mean squared error, which is the maximum likelihood estimate of σ^2, has a small negative bias:

$$\mathbb{E}\left[\widehat{\sigma}^2\right] = \mathbb{E}\left[\frac{1}{n}\mathbf{e}^T\mathbf{e}\right] = \sigma^2\frac{n-q}{n} = \sigma^2\left(1 - \frac{q}{n}\right). \tag{7}$$

Since $\mathbf{H}\mathbf{X}\beta = \mathbf{X}\beta$, the residuals can also be written

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\epsilon \tag{8}$$

hence

$$\mathbb{E}[\mathbf{e}] = \mathbf{0} \quad \text{and} \quad \mathrm{Var}[\mathbf{e}] = \sigma^2(\mathbf{I} - \mathbf{H}). \tag{9}$$

Under the Gaussian noise assumption, $\widehat{\beta}$, $\widehat{\mathbf{m}}$ and $\mathbf{e}$ all have Gaussian distributions.

1.1 Point Predictions

Suppose that $\mathbf{X}'$ is the $m \times q$ dimensional matrix storing the values of the predictor variables at m points where we want to make predictions. (These may or may not include points we used to estimate the model, and m may be bigger, smaller or equal to n.) Similarly, let $\mathbf{Y}'$ be the $m \times 1$ matrix of random values of Y at those points. The point predictions we want to make are

$$\mathbb{E}\left[\mathbf{Y}'|\mathbf{X}' = \mathbf{X}'\right] = \mathbf{m}(\mathbf{X}') = \mathbf{X}'\beta \tag{10}$$

and we *estimate* this by

$$\widehat{\mathbf{m}}(\mathbf{X}') = \mathbf{X}'\widehat{\beta} \tag{11}$$

which is to say

$$\widehat{\mathbf{m}}(\mathbf{X}') = \mathbf{X}'(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{12}$$

(It's easy to verify that when $\mathbf{X}' = \mathbf{X}$, this reduces to $\mathbf{HY}$.)

2 Diagnostics for Multiple Linear Regression

Before proceeding to detailed statistical inference, we need to check our modeling assumptions, which means we need diagnostics.

2.1 Plots

All of the plots we learned how to do for simple linear regression remain valuable:

1. *Plot the residuals against the predictors.* This now means p distinct plots, of course. Each of them should show a flat scatter of points around 0 (because $\mathbb{E}[\epsilon|X_i] = 0$), of roughly constant width (because $\mathrm{Var}[\epsilon|X_i] = \sigma^2$). Curvature or steps to this plot is a sign of potential nonlinearity, or of an omitted variable. Changing width is a potential sign of non-constant variance.

2. *Plot the squared residuals against the predictors.* Each of these p plots should show a flat scatter of points around $\widehat{\sigma}^2$.

3. *Plot the residuals against the fitted values.* This is an extra plot, redundant when we only have one predictor (because the fitted values were linear in the predictor).

4. *Plot the squared residuals against the fitted values.*

5. *Plot the residuals against coordinates.* If observations are dated, time-stamped, or spatially located, plot the residuals as functions of time, or make a map. If there is a meaningful order to the observations, plot residuals from successive observations against each other. Because the ϵ_i are uncorrelated, all of these plots should show a lack of structure.

6. *Plot the residuals' distribution against a Gaussian. (qq-plot)*

Out-of-sample predictions, with either random or deliberately selected testing sets, also remain valuable.

2.1.1 Collinearity

A linear dependence between two (or more) columns of the $\mathbf{X}$ matrix is called **collinearity**, and it keeps us from finding a solution by least squares. Computationally, collinearity will show up in the form of the determinant of $\mathbf{X}^T\mathbf{X}$ being zero. Equivalently, the smallest eigenvalue of $\mathbf{X}^T\mathbf{X}$ will be zero. If `lm` is given a collinear set of predictor variables, it will sometimes give an error messages, but more often it will decide not to estimate one of the collinear variables, and return an `NA` for the offending coefficient. We will return to the subject of collinearity in a future lecture.

2.1.2 Interactions

Another possible complication for multiple regression which we didn't have with the simple regression model is that of *interactions* between variables. One of our assumptions is that each variable makes a distinct, additive contribution to the response, and the size of this contribution is completely insensitive to the contributions of other variables. If this is *not* true — if the relationship between Y and X_i changes depending on the value of another predictor, X_j — then there is an **interaction** between them. There are several ways of looking for interactions. We will return to this subject in a future lecture.

2.2 Remedies

All of the remedies for model problems we discussed earlier, for the simple linear model, are still available to us.

Transform the response. We can change the response variable from Y to $g(Y)$, in the hope that the assumptions of the linear-Gaussian model are more nearly satisfied for this new variable. That is, we hope that

$$g(Y) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon,\ \epsilon \sim N(0,\sigma)^2. \tag{13}$$

Transform the predictors. We can also transform each of the predictors, making the model

$$Y = \beta_0 + \beta_1 f_1(X_1) + \ldots \beta_p f_p(X_p) + \epsilon,\ \epsilon \sim N(0,\sigma^2) \tag{14}$$

As the notation suggests, each X_i could be subject to a different transformation. Again, it's just a matter of what we put in the columns of the $\mathbf{X}$ matrix before solving for $\widehat{\beta}$. (In 402 you will see how the functions can be estimated automatically.)

Changing the variables used. One option which is available to us with multiple regression is to add in new variables, or to remove ones we're already using. This should be done carefully, with an eye towards satisfying the model assumptions, rather than blindly increasing some score. We will discuss this extensively later.

Removing Outliers. As always, we can remove outliers, as long as we document the fact that we are doing so.

3 Inference for Multiple Linear Regression

The results in this section presume that all of the modeling assumptions are correct. Also, all distributions stated are conditional on $\mathbf{X}$.

3.1 Sampling Distributions

As in the simple linear model, the sampling distributions are the basis of all inference.

In the simple linear model, because the noise ϵ is Gaussian, and the coefficient estimators were linear in the noise, $\widehat{\beta}_0$ and $\widehat{\beta}_1$ were also Gaussian. This remains true in for Gaussian multiple linear regression models:

$$\begin{aligned} \widehat{\beta} &= (\mathbf{X}^T\mathbf{X})\mathbf{X}^T\mathbf{Y} && (15)\\ &= (\mathbf{X}^T\mathbf{X})\mathbf{X}^T(\mathbf{X}\beta + \epsilon) && (16)\\ &= \beta + (\mathbf{X}^T\mathbf{X})\mathbf{X}^T\epsilon && (17) \end{aligned}$$

Since $(\mathbf{X}^T\mathbf{X})\mathbf{X}^T\epsilon$ is a constant times a Gaussian, it is also a Gaussian; adding on another Gaussian still leaves us with a Gaussian. We saw the expectation and variance last time, so

$$\widehat{\beta} \sim MVN(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}) \tag{18}$$

It follows that

$$\widehat{\beta}_j \sim N\left(\beta_j, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}_{jj}\right). \tag{19}$$

The same logic applies to the estimates of conditional means. In §1.1, we saw that the estimated conditional means at new observations $\mathbf{X}'$ are given by

$$\widehat{\mathbf{m}}(\mathbf{X}') = \mathbf{X}'(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{20}$$

so it follows that

$$\widehat{\mathbf{m}}(\mathbf{X}') \sim MVN(\mathbf{X}'\beta, \sigma^2\mathbf{X}'(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}')^T). \tag{21}$$

Eq. 21 simplifies for the special case of the fitted values, i.e., the estimated conditional means on the original data.

$$\widehat{\mathbf{Y}} \sim MVN(\mathbf{X}\beta, \sigma^2\mathbf{H}). \tag{22}$$

Similarly, the residuals have a Gaussian distribution:

$$\mathbf{e} \sim MVN(0, \sigma^2(\mathbf{I} - \mathbf{H}). \tag{23}$$

The in-sample mean squared error, or training error, or estimate of σ^2, is

$$\widehat{\sigma}^2 = n^{-1}\mathbf{e}^T\mathbf{e}$$

and

$$\frac{n\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-q} \tag{24}$$

where again $q = p + 1$. We will not prove this here.

Constraints on the residuals. The residuals are not all independent of each other. In the case of the simple linear model, the fact that we estimated the model by least squares left us with two constraints, $\sum_i e_i = 0$ and $\sum_i e_i X_i = 0$. If we had only one constraint, that would let us fill in the last residual if we knew the other $n-1$ residuals. Having two constraints meant that knowing any $n-2$ residuals determined the remaining two.

We got those constraints from the normal or estimating equations, which in turn came from setting the derivative of the mean squared error (or of the log-likelihood) to zero. In the multiple regression model, when we set the derivative to zero, we get the matrix equation

$$\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\widehat{\beta}) = \mathbf{0} \tag{25}$$

But the term in parentheses is just $\mathbf{e}$, so the equation is

$$\mathbf{X}^T\mathbf{e} = \mathbf{0} \tag{26}$$

Expanding out the matrix multiplication,

$$\begin{bmatrix} \sum_i e_i \\ \sum_i X_{i1} e_i \\ \vdots \\ \sum_i X_{ip} e_i \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{27}$$

Thus the residuals are subject to $p+1$ linear constraints, and knowing any $n-(p+1)$ of them will fix the rest. The vector of residuals $\mathbf{e}$ is a point in an n-dimensional space. As a random vector, without any constraints it could lie anywhere in that space, as, for instance, ϵ can. The constraints, however, for it to live in a lower-dimensional subspace, specifically, a space of dimension $n-(p+1)$.

Bias of $\widehat{\sigma}^2$. Let's compute the bias of $\widehat{\sigma}^2$. Before we do so, remember from Lecture 13 that if $Q = Z^T C Z$ is a quadratic form, then $\mathbb{E}[Q] = \mu^T C \mu + \mathrm{tr}(C\Sigma)$ where $\mu = \mathbb{E}$ and $\Sigma = \mathrm{Var}(Z)$. Also remember that $\mathbf{H}^T = \mathbf{H}$ and $\mathbf{H}^2 = \mathbf{H}$. So,

$$\begin{aligned}
\mathbb{E}\left[\widehat{\sigma}^2\right] &= \frac{1}{n}\mathbb{E}\left[\mathbf{e}^T\mathbf{e}\right] && (28)\\
&= \frac{1}{n}\mathbb{E}\left[((\mathbf{I}-\mathbf{H})\epsilon)^T((\mathbf{I}-\mathbf{H})\epsilon)\right] && (29)\\
&= \frac{1}{n}\mathbb{E}\left[\epsilon^T(\mathbf{I}^T-\mathbf{H}^T)(\mathbf{I}-\mathbf{H})\epsilon\right] && (30)\\
&= \frac{1}{n}\mathbb{E}\left[\epsilon^T(\mathbf{I}-\mathbf{H}-\mathbf{H}^T+\mathbf{H}^T\mathbf{H})\epsilon\right] && (31)\\
&= \frac{1}{n}\mathbb{E}\left[\epsilon^T(\mathbf{I}-\mathbf{H})\epsilon\right] && (32)\\
&= \frac{1}{n}\,\mathrm{tr}\left((\mathbf{I}-\mathbf{H})\mathrm{Var}\left[\epsilon\right]\right) && (33)\\
&= \frac{1}{n}\,\mathrm{tr}\left((\mathbf{I}-\mathbf{H})\sigma^2\mathbf{I}\right) && (34)\\
&= \frac{\sigma^2}{n}\,\mathrm{tr}\left(\mathbf{I}-\mathbf{H}\right) && (35)\\
&= \frac{\sigma^2}{n}(n-q) && (36)
\end{aligned}$$

since $\mathrm{Var}\left[\epsilon\right] = \sigma^2 I$, $\mathrm{tr}\,\mathbf{I} = n$ and $\mathrm{tr}\,\mathbf{H} = p+1$ (homework).

3.2 t Distributions for Coefficient and Conditional Mean Estimators

From Eq. 19, it follows that

$$\frac{\widehat{\beta}_j - \beta_j}{\mathrm{se}_j} \sim N(0,1) \tag{37}$$

where

$$\mathrm{se}_j = \sqrt{\sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}_{jj}}. \tag{38}$$

The estimate standard error is

$$\widehat{\mathrm{se}}\left[\widehat{\beta}_j\right] = \sqrt{\widehat{\sigma}^2 (\mathbf{X}^T \mathbf{X})^{-1}_{jj}}. \tag{39}$$

We then have that

$$\frac{\widehat{\beta}_j - \beta_j}{\widehat{\mathrm{se}}\left[\widehat{\beta}_j\right]} \sim t_{n-q}. \tag{40}$$

The same applies to the estimated conditional means, and to the distribution of a new Y' around the estimated conditional mean (in a prediction interval). Thus, all the theory we did for parametric and predictive inference in the simple model carries over, just with a different number of degrees of freedom.

As with the simple model, $t_{n-q} \to N(0,1)$, so t statistics approach z statistics as the sample size grows. R uses the t-distribution but you can use the Normal approximation if you like.

3.3 Hypothesis Testing

The `summary` function in R lists a p-value for testing $H_0 : \beta_j = 0$, for every cirefficient j. As usual, we should be skeptical about whether this is useful. As we said earlier, it is probably more important to focus on confidence intervals and prediction.

Common Mistakes. Looking at the hypothesis tests often leads to making some mistakes. Here are some of these common mistakes:

- Saying "β_i wasn't significantly different from zero, so X_i doesn't matter for Y". After all, X_i could still be an important cause of Y, but we don't have enough data, or enough variance on X_i, or enough variance in X_i uncorrelated with other X's, to accurately estimate its slope. All of these would prevent us from saying that β_i was *significantly* different from 0, i.e., distinguishable from 0 with high reliability.

- Saying "β_i was significantly different from zero, so X_i really matters to Y". After all, any β_i which is not *exactly* zero can be made arbitrarily significant by increasing n and/or the sample variance of X_i. That is, its t statistic will go to $\pm\infty$, and the p-value as small as you have patience to make it.

- Deleting all the variables whose coefficients didn't have stars by them, and re-running the regression. After all, since it makes no sense to pretend that the statistically significant variables are the only ones which matter, limiting the regression to the statistically significant variables is even less sensible.

- Saying "all my coefficients are really significant, so the linear-Gaussian model must be right". After all, all the hypothesis tests on individual coefficients *presume* the linear Gaussian model, both in the null and in the alternative. The tests have no power to notice nonlinearities, non-constant noise variance, or non-Gaussian noise.