# Lecture 14: Multiple Linear Regression

## 1 Review of Simple Linear Regression in Matrix Form

We have $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ and an $n \times 2$ matrix $\mathbf{X}$ whose first column is all 1's. The model is $\mathbf{Y} = \mathbf{X}\beta + \epsilon$. The error (mse) is $n^{-1}(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)$. The derivative of the MSE with respect to $\beta$ is

$$\frac{2}{n}(-\mathbf{X}^T\mathbf{Y} + \mathbf{X}^T\mathbf{X}\beta) \tag{1}$$

Setting this to zero at the optimum coefficient vector $\widehat{\beta}$ gives the (matrix) estimating equation

$$-\mathbf{X}^T\mathbf{Y} + \mathbf{X}^T\mathbf{X}\widehat{\beta} = 0 \tag{2}$$

whose solution is

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{3}$$

The fitted values are

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

where $\mathbf{H}$ is the hat matrix. Geometrically, this means that we find the fitted values by taking the vector of observed responses $\mathbf{Y}$ and projecting it onto the column space of $\mathbf{X}$.

## 2 Multiple Linear Regression

We are now ready to go from the *simple* linear regression model, with one predictor variable, to *multiple* linear regression models, with more than one predictor variable.

In the basic form of the **multiple linear regression model**,

1. There are $p$ quantitative predictor variables, $X_1, X_2, \ldots X_p$. We make no assumptions about their distribution; in particular, they may or may not be dependent. $X$ without a subscript will refer to the vector of all of these taken together. Thus, $X = (X_1, \ldots, X_p)$.
2. There is a single response variable $Y$.
3. $Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \epsilon$, for some constants (coefficients) $\beta_0, \beta_1, \ldots \beta_p$.
4. The noise variable $\epsilon$ has $\mathbb{E}[\epsilon|X = x] = 0$ (mean zero), $\mathrm{Var}[\epsilon|X = x] = \sigma^2$ (constant variance), and is uncorrelated across observations.

In matrix form, when we have $n$ observations,

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{4}$$

where $\mathbf{X}$ is a $n \times (p+1)$ matrix of random variables whose first column is all 1's. We assume that $\mathbb{E}[\epsilon|\mathbf{X}] = 0$ and $\mathrm{Var}[\epsilon|\mathbf{X}] = \sigma^2\mathbf{I}$.

Sometimes we further assume that $\epsilon \sim MVN(\mathbf{0}, \sigma^2\mathbf{I})$, independent of $\mathbf{X}$. From these assumptions, it follows that, conditional on $\mathbf{X}$, $\mathbf{Y}$ has a multivariate Gaussian distribution,

$$\mathbf{Y}|\mathbf{X} \sim MVN(\mathbf{X}\beta, \sigma^2\mathbf{I}). \tag{5}$$

# 3 Derivation of the Least Squares Estimator

We now wish to estimate the model by least squares. Fortunately, we did essentially all of the necessary work last time.

The MSE is

$$\frac{1}{n}(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) \tag{6}$$

with gradient

$$\nabla_\beta MSE(\beta) = \frac{2}{n}\left(-\mathbf{X}^T\mathbf{Y} + \mathbf{X}^T\mathbf{X}\beta\right). \tag{7}$$

The estimating equation is

$$-\mathbf{X}^T\mathbf{Y} + \mathbf{X}^T\mathbf{X}\widehat{\beta} = 0 \tag{8}$$

and the solution, the **ordinary least squares** (OLS) estimator, is

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{9}$$

## 3.1 Why Multiple Regression Isn't Just a Bunch of Simple Regressions

When we do multiple regression, the slopes we get for each variable aren't the same as the ones we'd get if we just did $p$ separate simple regressions. Why not?

Suppose the real model is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$. (Nothing turns on $p = 2$, it just keeps things short.) What would happen if we did a simple regression of $Y$ on just $X_1$? We know that the optimal (population) slope on $X_1$ is

$$\frac{\mathrm{Cov}\left[X_1, Y\right]}{\mathrm{Var}\left[X_1\right]} \tag{10}$$

Let's substitute in the model equation for $Y$:

$$\begin{aligned}
\frac{\mathrm{Cov}\left[X_1, Y\right]}{\mathrm{Var}\left[X_1\right]} &= \frac{\mathrm{Cov}\left[X_1, \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon\right]}{\mathrm{Var}\left[X_1\right]} && (11)\\
&= \frac{\beta_1 \mathrm{Var}\left[X_1\right] + \beta_2 \mathrm{Cov}\left[X_1, X_2\right] + \mathrm{Cov}\left[X_1, \epsilon\right]}{\mathrm{Var}\left[X_1\right]} && (12)\\
&= \beta_1 + \frac{\beta_2 \mathrm{Cov}\left[X_1, X_2\right] + 0}{\mathrm{Var}\left[X_1\right]} && (13)\\
&= \beta_1 + \beta_2 \frac{\mathrm{Cov}\left[X_1, X_2\right]}{\mathrm{Var}\left[X_1\right]} && (14)
\end{aligned}$$

The total covariance between $X_1$ and $Y$ includes $X_1$'s direct contribution to $Y$, plus the indirect contribution through correlation with $X_2$, and $X_2$'s contribution to $Y$.

### 3.2 Point Predictions and Fitted Values

Just as with simple regression, the vector of fitted values $\widehat{\mathbf{Y}}$ is linear in $\mathbf{Y}$, and given by the hat matrix:

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}\mathbf{Y}. \tag{15}$$

All of the interpretations given of the hat matrix in the previous lecture still apply. In particular, $\mathbf{H}$ projects $\mathbf{Y}$ onto the column space of $\mathbf{X}$.

## 4 Properties of the Estimates

As usual, we will treat $\mathbf{X}$ as fixed. Now

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{16}$$

and

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{17}$$

and so

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon = \beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon. \tag{18}$$

### 4.1 Bias

This is straight-forward:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\beta}\right] &= \mathbb{E}\left[\beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon\right] && (19)\\
&= \beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}\left[\epsilon\right] && (20)\\
&= \beta && (21)
\end{aligned}$$

so the least squares estimate is unbiased.

### 4.2 Variance and Standard Errors

This needs a little more work. We have

$$\begin{aligned}
\mathrm{Var}\left[\widehat{\beta}\right] &= \mathrm{Var}\left[\beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon\right] && (22)\\
&= \mathrm{Var}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon\right] && (23)\\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{Var}\left[\epsilon\right]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} && (24)\\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} && (25)\\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} && (26)\\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} && (27)
\end{aligned}$$

To understand this a little better, let's re-write it slightly:

$$\text{Var}\left[\widehat{\beta}\right] = \frac{\sigma^2}{n}\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)^{-1}. \tag{28}$$

The first term, $\sigma^2/n$, is what we're familiar with from the simple linear model. As $n$ grows, we expect the entries in $\mathbf{X}^T\mathbf{X}$ to be increasing in magnitude, since they're sums over all $n$ data points; dividing all entries in the matrix by $n$ compensates for this. If the sample covariances between all the predictor variables were 0, when we took the inverse we'd get $1/s^2_{X_i}$ down the diagonal (except for the top of the diagonal), just as we got $1/s^2_X$ in the simple linear model.

# 5 Collinearity

We have been silently assuming that $(\mathbf{X}^T\mathbf{X})^{-1}$ exists, in other words, that $\mathbf{X}^T\mathbf{X}$ is "invertible" or "non-singular". There are a number of equivalent conditions for a matrix to be invertible:

1. Its determinant is non-zero.
2. It is of "full column rank", meaning all of its columns are linearly independent[1].
3. It is of "full row rank", meaning all of its rows are linearly independent.

The equivalence of these conditions are mathematical facts, proved in linear algebra.

What does this amount to in terms of our data? It means that the variables must be linearly independent *in our sample*. That is, there must not be any set of constants $a_0, a_1, \ldots a_p$ where, for *all* rows $i$,

$$a_0 + \sum_{j=1}^{p} a_j x_{ij} = 0 \tag{29}$$

This, in other words, means that $\mathbf{X}$ must be of full column rank.

To understand why linearly dependence among variables is a problem, take an easy case, where two predictors, say $X_1$ and $X_2$, are exactly equal to each other. It's then not surprising that we don't have any way of estimating their coefficients. If we get one set of predictions with coefficients $\beta_1, \beta_2$, we'd get exactly the same predictions from $\beta_1 + \gamma$, $\beta_2 - \gamma$, no matter what $\gamma$ might be. If there are other exact linear relations among two variables, we can similarly trade off their coefficients against each other, without any change in anything we can observe. If there are exact linear relationships among more than two variables, all of their coefficients become ill-defined.

We will come back in a few lectures to what to do when faced with collinearity. For now, we'll just mention a few clear situations:

[1] Recall that a set of vectors is linearly independent if no linear combination of them is exactly zero.

- If $n < p + 1$, the data are collinear.
- If one of the predictor variables is constant, the data are collinear.
- If two of the predictor variables are proportional to each other, the data are collinear.
- If two of the predictor variables are otherwise linearly related, the data are collinear.

# 6 R

```
>
> pdf("plots.pdf")
>
> n  = 100
> x1 = runif(n)
> x2 = runif(n)
> x3 = runif(n)
> y  = 5 + 2*x1 + 3*x2 + 7*x3 + rnorm(n)
>
> Z = cbind(x1,x2,x3,y)
>
> pairs(Z,pch=20)
>
>
>
> out = lm(y ~ x1 + x2 + x3)
>
> print(out)

Call:
lm(formula = y ~ x1 + x2 + x3)

Coefficients:
(Intercept)           x1           x2           x3
      4.619        2.840        2.607        7.286

>
> coefficients(out)
(Intercept)          x1          x2          x3
   4.618816    2.840239    2.607443    7.285716
>
> confint(out)
               2.5 %   97.5 %
```

```
(Intercept) 4.017390 5.220243
x1          2.257652 3.422826
x2          2.004926 3.209960
x3          6.659326 7.912105
>
> head(fitted(out))
        1         2         3         4         5         6
11.984243 11.300009 12.006982 11.556004  9.205792 11.400073
>
> head(residuals(out))
         1          2          3          4          5          6
-0.3574136  0.0609240 -1.4612416 -0.3427516 -0.1730116  0.3899873
>
> summary(out)

Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
     Min       1Q   Median       3Q      Max
-1.91902 -0.59934  0.00622  0.65931  1.81582

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.6188     0.3030  15.244  < 2e-16 ***
x1            2.8402     0.2935   9.677 7.34e-16 ***
x2            2.6074     0.3035   8.590 1.58e-13 ***
x3            7.2857     0.3156  23.088  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.8557 on 96 degrees of freedom
Multiple R-squared:  0.8656,Adjusted R-squared:  0.8614
F-statistic: 206.1 on 3 and 96 DF,  p-value: < 2.2e-16

>
> newx = data.frame(x1 = .2, x2 = .3, x3 = .7)
>
> predict(out,newdata = newx)
      1
11.0691
>
```

```
> dev.off()
```