STAT 135 Lab 13 (Review) Linear Regression, Multivariate Random Variables, Prediction, Logistic Regression and the δ-Method.

Rebecca Barter

May 5, 2015

Linear Regression Review

Linear Regression Review

$$y = \beta_0 + \beta_1 x + \epsilon$$

Suppose that there is some global linear relationship between height and weight in the population. For example

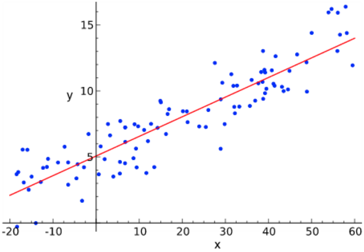
$$\text{Weight} = \beta_0 + \beta_1 \times \text{Height} + \epsilon$$

Obviously height and weight don't fall on a perfectly straight line:

- there is some random noise/deviation from the line (ϵ is a random variable with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$).

Moreover, we don't observe everyone in the population, and so there is no way we can figure out what β_0 and β_1 are, but what we can do is get estimates from a sample.

Linear Regression Review



We observe a sample (blue points) and fit the best line through the sample. The red fitted line corresponds to

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimated from our sample using least squares (finding the β values that minimize the distance from the observed points to the line)

Linear Regression Review

We showed that we can write our linear regression model in matrix form:

$$Y = X\beta + \epsilon$$

and the least squares estimate of β is given by

$$\hat{\beta} = \arg\ \min_{\beta} \|Y - X\beta\|_2^2 = (X^T X)^{-1} X^T Y$$

We further showed that $\hat{\beta}$ was unbiased:

$$E(\hat{\beta}) = \beta$$

and that its variance is given by

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

Linear Regression Review

The **residuals** are then defined to be the observed difference between the true y (blue dot) and the fitted $\hat{y}$ (corresponding point on the red line):

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

which is different from the unobservable error

$$\epsilon_i = y_i - \beta_0 - \beta_1 x_i$$

(this corresponds to the deviation from the observed y_i to the true population line, which we don't know!)

Multivariate Random Variables

Multivariate Random Variables

Suppose we have n random variables: $Y_1, ..., Y_n$, such that $E(Y_i) = \mu$ and $Cov(Y_i, Y_j) = \sigma_{ij}$. Suppose that we want to consider the Y_i's jointly as a single multivariate random variable, $\mathbf{Y}$

$$\mathbf{Y} = (Y_1, ..., Y_n)$$

Then this multivariate random variable $\mathbf{Y}$ has mean *vector* and *covariance matrix*:

mean vector: $E(\mathbf{Y}) = \mu = (\mu_1, ..., \mu_n)$

covariance matrix: $Cov(\mathbf{Y}) = \Sigma_{Y,Y} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \dots & \sigma_{1,n} \\ \sigma_{2,1} & \sigma_2^2 & \dots & \sigma_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{n,1} & \sigma_{n,2} & \dots & \sigma_n^2 \end{bmatrix}$

Multivariate Random Variables

$$E(\mathbf{Y}) = \mu \qquad Cov(\mathbf{Y}) = \Sigma$$

Suppose we have the following **linear transformation** of $\mathbf{Y}$:

$$\mathbf{Z} = b + A\mathbf{Y}$$

where b is a fixed vector and A a fixed matrix.

What is the mean vector and covariance matrix of $\mathbf{Z}$?

$$\boxed{E(\mathbf{Z}) = E(b + A\mathbf{Y}) = b + AE(\mathbf{Y}) = b + A\mu}$$

$$\boxed{Cov(\mathbf{Z}) = Cov(b + A\mathbf{Y}) = Cov(A\mathbf{Y}) = ACov(\mathbf{Y})A^T = A\Sigma A^T}$$

Multivariate Random Variables

Suppose that $\mathbf{Y}$ is a multivariate random vector:
$\mathbf{Y} = (Y_1, ..., Y_n)$.

Recall that for scalars, a **quadratic transformation** of y is given by $x = ay^2 \in \mathbb{R}$. The equivalent form for vectors is

$$X = \mathbf{Y}^T A \mathbf{Y} \in \mathbb{R}$$

This is referred to as a *random quadratic form* in $\mathbf{Y}$

Multivariate Random Variables

$$X = \mathbf{Y}^T A \mathbf{Y} \in \mathbb{R}$$

is a *random quadratic form* ia $\mathbf{Y}$

How can we calculate the expected value of the quadratic form of a random variable?

We can use the following facts

1. Trace and expectation are both linear operators (and so can be interchanged).
2. $X = \mathbf{Y}^T A \mathbf{Y} \in \mathbb{R}$ is a real number (not a matrix or vector), and if $x \in \mathbb{R}$ then $tr(x) = x$.
3. If A and B are matrices of appropriate dimension, then $tr(AB) = tr(BA)$

where the *trace* of a matrix is the sum of its diagonal entries

Multivariate Random Variables

1. Trace and expectation are both linear operators (and so can be interchanged).
2. If $x \in \mathbb{R}$ then $tr(x) = x$.
3. If A and B are matrices, then $tr(AB) = tr(BA)$

So

$$\begin{aligned}
E\left[\mathbf{Y}^T A \mathbf{Y}\right] &= E\left[tr(\mathbf{Y}^T A \mathbf{Y})\right] && \text{fact (2)} \\
&= E\left[tr(A\mathbf{Y}\mathbf{Y}^T)\right] && \text{fact (3)} \\
&= tr\left(E\left[A\mathbf{Y}\mathbf{Y}^T\right]\right) && \text{fact (1)} \\
&= tr\left(AE\left[\mathbf{Y}\mathbf{Y}^T\right]\right) && \text{fact (1)} \\
&= tr\left(A\left(\Sigma + \mu\mu^T\right)\right) && \text{calculating expectation} \\
&= tr(A\Sigma) + tr(A\mu\mu^T) && \text{fact (1)} \\
&= tr(A\Sigma) + tr(\mu^T A\mu) && \text{fact (3)} \\
&= tr(A\Sigma) + \mu^T A\mu && \text{fact (2)}
\end{aligned}$$

Example: Residual sum of squares

The expected value of a quadratic form of a random variable is given by:

$$E\left[\mathbf{Y}^T A \mathbf{Y}\right] = tr(A\Sigma) + \mu^T A \mu$$

We can use this to show that an unbiased estimate for σ^2 is given by

$$\hat{\sigma}^2 = \frac{RSS}{n-p}$$

where

$$RSS = \|e\|_2^2 = e^T e$$

is in quadratic form

Example: Residual sum of squares

Recall that the residuals are defined by

$$\begin{aligned} e &= Y - \hat{Y} \\ &= Y - X\hat{\beta} \\ &= Y - X(X^TX)^{-1}X^TY \\ &= (I - X(X^TX)^{-1}X^T)Y \\ &= P_{X^\perp}Y \end{aligned}$$

Where $P_{X^\perp} := I - X(X^TX)^{-1}X^T$ is a projection matrix onto the space orthogonal to X (check: $P_{X^\perp}X = 0$). Note that a **projection matrix** is a matrix, A, that satisfies

- $A^T = A$
- $AA = A^2 = A$

Example: Residual sum of squares

$$e = (I - X(X^TX)^{-1}X^T)Y = P_{X^\perp}^T Y$$

Recall that the **residual sum of squares** was given by

$$\begin{aligned} RSS &= \|e\|_2^2 = e^Te \\ &= (P_{X^\perp}Y)^T(P_{X^\perp}Y) \\ &= Y^TP_{X^\perp}^TP_{X^\perp}Y \\ &= Y^TP_{X^\perp}Y \end{aligned}$$

Thus, since $E\left[\mathbf{Y}^TA\mathbf{Y}\right] = tr(A\Sigma) + \mu^TA\mu$, the **expected RSS** is given by

$$\begin{aligned} E(RSS) &= E(Y^TP_{X^\perp}Y) \\ &= \sigma^2 tr(P_{X^\perp}) + E(Y)^TP_{X^\perp}E(Y) \qquad (\Sigma = \sigma^2 I_n) \\ &= \sigma^2(n-p) \end{aligned}$$

How did we get that last equality!?

Example: Residual sum of squares

$$E(RSS) = \sigma^2 tr(P_{X^\perp}) + E(Y)^T P_{X^\perp} E(Y)$$
$$= \sigma^2 (n-p)$$

To get the last equality, we use our newly acquired knowledge of projection matrices and the trace operator:

- $P_{X^\perp}$ is a projection matrix orthogonal to X, thus $P_{X^\perp} E(Y) = P_{X^\perp} X\beta = 0$, so
$$E(Y)^T P_{X^\perp} E(Y) = 0$$
- The trace of an $m \times m$ identity matrix is m, so
$$\begin{aligned} tr(P_{X^\perp}) &= tr(I_n - X(X^TX)^{-1}X^T) \\ &= n - tr(X(X^TX)^{-1}X^T) && \text{(linearity of trace)} \\ &= n - tr((X^TX)^{-1}X^TX) && (tr(AB) = tr(BA)) \\ &= n - tr(I_p) \\ &= n - p \end{aligned}$$

Example: Residual sum of squares

Note that we have shown that

$$E(RSS) = \sigma^2(n-p)$$

which tells us that

$$\hat{\sigma}^2 = \frac{RSS}{n-p}$$

is an unbiased estimate of σ^2.

Prediction

Prediction

Suppose we want to predict/fit the responses, $Y_1, ..., Y_n$ corresponding to the observed predictors, $x_1, ..., x_n$ (each x_i is a $1 \times p$ vector), which form the rows of the design matrix X.

Then our **predicted** Y_i's (note that these are the same Y's used to define the model) are given by

$$\hat{Y}_i = x_i\hat{\beta} = x_i(X^T X)^{-1}X^T Y$$

Since these $\hat{Y}_i$'s are random variables, we might be interested in calculating the variance of these predictions.

Prediction

Our **predicted** Y_i's are given by

$$\hat{Y}_i = x_i\hat{\beta} = x_i(X^TX)^{-1}X^TY$$

The variance of the $\hat{Y}_i$'s are given by

$$\begin{aligned}
Var(\hat{Y}_i) &= Var(x_i(X^TX)^{-1}X^TY) \\
&= x_i(X^TX)^{-1}X^T\ Var(Y)\ X(X^TX)^{-1}x_i^T \\
&= \sigma^2 x_i(X^TX)^{-1}X^TX(X^TX)^{-1}x_i^T \\
&= \sigma^2 x_i(X^TX)^{-1}x_i^T
\end{aligned}$$

Prediction

$$Var(\hat{Y}_i) = \sigma^2 x_i (X^T X)^{-1} x_i^T$$

Thus the average variance for across the n observations is:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} Var(\hat{Y}_i) &= \frac{\sigma^2}{n}\sum_{i=1}^{n} x_i (X^T X)^{-1} x_i^T \\
&= \frac{\sigma^2}{n} tr(X(X^T X)^{-1} X^T) \\
&= \frac{\sigma^2}{n} tr((X^T X)^{-1} X^T X) \\
&= \frac{\sigma^2}{n} tr(I_p) \\
&= \sigma^2 \frac{p}{n}
\end{aligned}$$

So the more variables we have in our model, the more variable our fitted/predicted values will be!

Logistic Regression

Logistic Regression

For **linear regression**, we have

$$Y = X\beta + \epsilon$$

and Y is continuous, such that $Y_i | x_i \sim N(\beta^T x_i, \sigma^2)$

- x_i is the ith row/observation in X

We want to use **logistic regression** if Y doesn't take continuous values, but is instead **binary**, i.e. $Y_i \in \{0, 1\}$.

The above linear model no longer makes sense, since a linear combination of our (continuous) x's is very unlikely to be equal to 0 or 1.

First step: think of a distribution for $Y_i | x_i$ that makes more sense for binary Y...

Logistic Regression

Since Y_i is either 0 or 1 we can model it as a Bernoulli random variable

$$\boxed{Y_i|x_i \sim Bernoulli(p(x_i, \beta))}$$

where the probability that $Y_i = 1$ is given by some function of our predictors, $p(x_i, \beta)$.

We need our probability, $p(x_i, \beta)$ to be in $[0, 1]$ and for it to depend on the linear combination of our predictors, $\beta^T x_i$, in some way. We choose:

$$\boxed{p(x_i, \beta) = \frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}}}$$

This is a nice choice since

- $\beta^T x_i \geq 0$ implies that $p(x_i, \beta) \geq 0.5$ so $Y_i = 1$ is most likely
- $\beta^T x_i \leq 0$ implies that $p(x_i, \beta) \leq 0.5$ so $Y_i = 0$ is most likely

Logistic Regression

We have the following logistic regression model:

$$Y_i|x_i \sim Bernoulli\left(\frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}\right)$$

- If we had a new sample x and we knew the β vector, we could calculate the probability that the corresponding $Y = 1$:
$$P(Y=1|x) = p(x,\beta) = \frac{e^{\beta^T x}}{1+e^{\beta^T x}}$$
- If this probability is greater than 0.5, we would set $\hat{Y} = 1$.
- Otherwise, we would set $\hat{Y} = 0$.

But we don't know β: we need to **estimate** the unknown β vector (just like with linear regression).

Logistic Regression

Let's try to calculate the MLE for β. The likelihood function is given by

$$\begin{aligned} lik(\beta) &= \prod_{i=1}^{n} P(y_i | X_i = x_i, \beta) \\ &= \prod_{i=1}^{n} \left(\frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}} \right)^{y_i} \left(\frac{1}{1 + e^{\beta^T x_i}} \right)^{1-y_i} \\ &= \prod_{i=1}^{n} \left(\frac{e^{\beta^T x_i y_i}}{1 + e^{\beta^T x_i}} \right) \end{aligned}$$

So that the log-likelihood is given by

$$\ell(\beta) = \sum_{i=1}^{n} \beta^T x_i y_i - \log\left(1 + e^{\beta^T x_i}\right)$$

Logistic Regression

The log-likelihood is given by

$$\ell(\beta) = \sum_{i=1}^{n} \beta^T x_i y_i - \log\left(1 + e^{\beta^T x_i}\right)$$

Differentiating with respect to β gives

$$\begin{aligned}
\nabla \ell(\beta) &= \sum_{i=1}^{n} x_i y_i - \frac{x_i e^{\beta^T x_i}}{1 + e^{\beta^T x_i}} \\
&= \sum_{i=1}^{n} x_i (y_i - p(x_i, \beta)) \\
&= X^T (y - p)
\end{aligned}$$

where $p = (p(x_1, \beta), p(x_2, \beta), ..., p(x_n, \beta))$.
Setting $\nabla \ell(\beta) = 0$ cannot yield a closed form solution for β, so we need to use an iterative approach such as Newton's method.

Logistic Regression

Newton's method: Suppose that we want to find x such that $f(x) = 0$. Then we could use the iterative approximation given by

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

and the more iterations we conduct, the closer we get to the root.

We need to find the vector β such that $\nabla \ell(\beta) = 0$. To do this, we need to conduct a vector version of Newton's method:

$$\beta^{n+1} = \beta^n - \left(\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} \right)^{-1} \nabla \ell(\beta)$$

Logistic Regression

Newton's method for scalars: Find x such that $f(x) = 0$ by

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Vector version of Newton's method for β:

$$\beta^{n+1} = \beta^n - \left(\frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta^T}\right)^{-1} \nabla\ell(\beta)$$

- $f(x) = \nabla\ell(x)$ is the vector-version of the first derivative
- $f'(x) = \frac{\partial^2 \ell(x)}{\partial x \partial x^T}$ is the vector-version of the second derivative

Logistic Regression

The first derivative (gradient vector) of $\ell(\beta)$ with respect to the vector β:

$$\nabla \ell(\beta) = \begin{bmatrix} \frac{\partial \ell}{\partial \beta_0} \\ \frac{\partial \ell}{\partial \beta_1} \\ \vdots \\ \frac{\partial \ell}{\partial \beta_p} \end{bmatrix}$$

The second derivative (Hessian matrix) of $\ell(\beta)$ with respect to the vector β:

$$\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} = \begin{bmatrix} \frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_0} & \frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1} & \cdots & \frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_p} \\ \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_1} & \cdots & \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_p} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 \ell}{\partial \beta_p \partial \beta_0} & \frac{\partial^2 \ell}{\partial \beta_p \partial \beta_1} & \cdots & \frac{\partial^2 \ell}{\partial \beta_p \partial \beta_p} \end{bmatrix}$$

Logistic Regression

Recall that we were trying to find the MLE for β in the logistic regression using Newton's method

$$\beta^{n+1} = \beta^n - \left(\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} \right)^{-1} \nabla \ell(\beta)$$

We already showed that

$$\nabla(\beta) = X^T(y - p)$$

and one can show that

$$\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} = -X^T W X$$

where W is a diagonal matrix whose diagonal entries are given by $W_{ii} = Var(y_i) = p(x_i, \beta)(1 - p(x_i, \beta))$

Logistic Regression

So our iterative procedure becomes

$$\begin{aligned}\beta^{n+1} &= \beta^n + (X^T W X)^{-1} X^T (y - p) \\ &= (X^T W X)^{-1} X^T W (X\beta^n + W^{-1}(y - p)) \\ &= (X^T W X)^{-1} X^T W z\end{aligned}$$

where $z = X\beta^n + W^{-1}(y - p)$

Recall the least squares linear regression β estimator

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

and our iterative logistic regression estimator

$$\beta^{n+1} = (X^T W X)^{-1} X^T W z$$

We say that the estimate has been *reweighted* by W, and so this estimator is referred to as *iteratively reweighted least squares*

The δ-Method

The δ-Method

The δ-method tells us that if we have a sequence of random variables, X_n, such that

$$\sqrt{n}(X_n - \theta) \to N(0, \sigma^2) \tag{1}$$

then, for any differentiable and non-zero function, $g(\cdot)$, we have that

$$\sqrt{n}(g(X_n) - g(\theta)) \to N(0, \sigma^2(g'(\theta))^2) \tag{2}$$

The most common reason we use the δ-method is to identify the **variance** of a function of a random variable.

Example: The δ-Method

Suppose that $X_n \sim Binom(n, p)$.

Then since $\frac{X_n}{n}$ is the mean of n iid Bernoulli(p) random variables, the cental limit theorem tells us that

$$\sqrt{n}\left(\frac{X_n}{n} - p\right) \to N(0, p(1-p))$$

and that $Var(\frac{X_n}{n}) = \frac{p(1-p)}{n}$

δ-method: for any non-zero, differentiable function $g(\cdot)$, we have

$$\sqrt{n}\left(g\left(\frac{X_n}{n}\right) - g(p)\right) \to N(0, p(1-p)(g'(p))^2)$$

For example, we can use this to calculate $Var\left(\log\left(\frac{X_n}{n}\right)\right)$.

Example: The δ-Method

$$\sqrt{n}\left(g\left(\frac{X_n}{n}\right) - g(p)\right) \to N(0, p(1-p)(g'(p))^2)$$

For example, we can use this to calculate $Var\left(\log\left(\frac{X_n}{n}\right)\right)$.

Select

$$g(x) = \log(x)$$

So differentiating gives

$$g'(x) = \frac{1}{x}$$

and the δ-method says:

$$\sqrt{n}\left(\log\left(\frac{X_n}{n}\right) - \log(p)\right) \to N\left(0, \frac{p(1-p)}{p^2}\right)$$

Example: The δ-Method

The δ-method says:

$$\sqrt{n}\left(\log\left(\frac{X_n}{n}\right) - \log(p)\right) \to N\left(0, \frac{p(1-p)}{p^2}\right)$$

so we can see that the variance of $\log\left(\frac{X_n}{n}\right)$ is

$$Var\left(\log\left(\frac{X_n}{n}\right)\right) = \frac{1-p}{pn}$$