Advanced Methods in Biostatistics I

Lecture 3

Martin Lindquist

September 5, 2017

Simple linear regression

- In today's class we will consider simple linear regression.
- But again we begin with some review of linear algebra.

Projections

- In the previous lecture we defined projections of one vector in $\mathbf{V}$ onto another.
- Now we focus on the projection of a vector onto a subspace of $\mathbf{V}$.

Orthogonality to a subspace

Definition

A vector $\mathbf{v} \in \mathbf{V}$ is orthogonal to a subspace $\mathbf{W} \subseteq \mathbf{V}$ if $\langle \mathbf{v}, \mathbf{w} \rangle = 0$ for all $\mathbf{w} \in \mathbf{W}$

Note: If $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k$ form a basis for $\mathbf{W}$, then $\mathbf{v}$ is orthogonal to $\mathbf{W}$ if $\langle \mathbf{v}, \mathbf{w}_i \rangle = 0$ for $i = 1, \ldots k$.

Projections

Definition

The projection of a vector $\mathbf{y}$ on a subspace $\mathbf{W}$ of $\mathbf{V}$ is the vector $\hat{\mathbf{y}} \in \mathbf{V}$ such that $(\mathbf{y} - \hat{\mathbf{y}}) \perp \mathbf{V}$. The vector $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ will be called the residual vector for $\mathbf{y}$ relative to $\mathbf{V}$.

Projections

Theorem

Let $\mathbf{W}$ be a subspace and $\mathbf{y}$ a vector in $\mathbf{V}$. Assume $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_k\}$ is an orthogonal basis for $\mathbf{W}$. Then the vector

$$\hat{\mathbf{y}} = \sum_{i=1}^{k} \frac{\langle \mathbf{y}, \mathbf{x}_i \rangle}{\langle \mathbf{x}_i, \mathbf{x}_i \rangle} \mathbf{x}_i$$

in $\mathbf{W}$ is the projection of $\mathbf{y}$ onto $\mathbf{W}$.

Projections

- Note that this result does NOT hold for a basis that is not an orthogonal set.
- Every subspace contains an orthogonal basis.
- Such a basis can be constructed by using Gram-Schmidt orthogonalization.

Projections

- In general for $\hat{\mathbf{y}} = b_1\mathbf{x}_1 + \cdots + b_k\mathbf{x}_k$ to be the projection of $\mathbf{y}$ on $\mathbf{V} = sp\{\mathbf{x}_1, \ldots \mathbf{x}_k\}$ we need $(\mathbf{y}, \mathbf{x}_i) = (\hat{\mathbf{y}}, \mathbf{x}_i)$ for all i.
- This leads to the so-called normal equations:

$$(\hat{\mathbf{y}}, \mathbf{x}_i) = \sum_{j=1}^{k} b_j(\mathbf{x}_j, \mathbf{x}_i) = (\mathbf{y}, \mathbf{x}_i)$$

for $i = 1, \ldots k$.

Orthogonal complements

Definition

If $\mathbf{W}$ is a set of vectors in $\mathbf{V}$, then the set $\mathbf{W}^{\perp}$ is called the orthogonal complement of $\mathbf{W}$ in $\mathbf{V}$ and is defined as

$$\mathbf{W}^{\perp} = \{\mathbf{x} : \langle \mathbf{x}, \mathbf{y} \rangle = 0; \mathbf{y} \in \mathbf{W}\}.$$

Orthogonal complements

Theorem

If $\mathbf{W}$ is a subspace of $\mathbf{V}$ with $dim(\mathbf{W}) = r$ and $dim(\mathbf{W}) = n$, then $dim(\mathbf{W}^{\perp}) = n - r$

Orthogonal complements

Theorem

If $\mathbf{W}$ is a subspace of $\mathbf{V}$, then the orthogonal complement of $\mathbf{W}$ with respect to $\mathbf{V}$ is a subspace of $\mathbf{V}$, and any $\mathbf{v} \in \mathbf{V}$ can be written uniquely as $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$ where $\mathbf{v}_1 \in \mathbf{W}$, $\mathbf{v}_2 \in \mathbf{W}^\perp$.

Simple linear regression

Let us consider the case of simple linear regression.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \text{for } \; i = 1, \ldots n$$

or

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Simple linear regression

- Consider the scatterplot of points (x_i, y_i).
- The goal is to find the best fitting line of the form $y = \beta_0 + \beta_1 x$ by minimizing the sum of the squared vertical distances between the points and the fitted line.

- That is, we seek to minimize:

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Simple linear regression

- Here β_1 is referred to as the slope, and β_0 as the intercept.
- The slope β_1 has units 'y-units per x-units'.

- The intercept corresponds to the value of y when $x = 0$, and is not always meaningful if the 0 lies outside the range of reasonable values for x.

Simple linear regression

- In matrix formulation we seek to minimize:

$$||\mathbf{y} - (\beta_0 \mathbf{J}_n + \beta_1 \mathbf{x})||^2$$

over β_1 and β_2.

Simple linear regression

- The least squares estimates can be found by differentiating f with respect to β_0 and β_1 and setting the partial derivatives equal to 0, i.e.

$$\begin{aligned}\frac{\partial f(\beta_0, \beta_1)}{\partial \beta_0} &= -2\mathbf{J}_n'(\mathbf{y} - \beta_0\mathbf{J}_n - \beta_1\mathbf{x}) = 0 \\ \frac{\partial f(\beta_0, \beta_1)}{\partial \beta_1} &= -2\mathbf{x}'(\mathbf{y} - \beta_0\mathbf{J}_n - \beta_1\mathbf{x}) = 0.\end{aligned}$$

Normal equations

- The values of $\hat{\beta}_0$ and $\hat{\beta}_0$ that minimize f are given by the solution to the normal equations:

$$\begin{aligned} \mathbf{J}_n'\mathbf{y} &= \beta_0 \mathbf{J}_n'\mathbf{J}_n + \beta_1 \mathbf{J}_n'\mathbf{x} \\ \mathbf{x}'\mathbf{y} &= \beta_0 \mathbf{x}'\mathbf{J}_n + \beta_1 \mathbf{x}'\mathbf{x} \end{aligned}$$

Normal equations

- The normal equations can alternatively be expressed as follows:

$$
\begin{aligned}
\sum_{i=1}^{n} y_i &= n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i-1}^{n} x_i \\
\sum_{i=1}^{n} x_i y_i &= \hat{\beta}_0 \sum_{i-1}^{n} x_i + \hat{\beta}_1 \sum_{i-1}^{n} x_i^2 .
\end{aligned}
$$

Solution to the normal equations

- Solving the normal equations gives us the following estimates:

$$\hat{\beta}_1 = \frac{\sum_{i-1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Alternative formulation

- Note we can also write $\hat{\beta}_1$ as follows:

$$\hat{\beta}_1 = \frac{(\mathbf{x} - \bar{x}\mathbf{J}_n)'(\mathbf{y} - \bar{y}\mathbf{J}_n)}{(\mathbf{x} - \bar{x}\mathbf{J}_n)'(\mathbf{x} - \bar{x}\mathbf{J}_n)}.$$

Hessian

- To check whether $\hat{\beta}_0$ and $\hat{\beta}_1$ correspond to the minimum of f, it suffices to check whether the Hessian matrix is positive definite.
- The Hessian matrix can be expressed as

$$\begin{pmatrix} \frac{\partial f}{\partial \beta_0^2} & \frac{\partial f}{\partial \beta_0 \beta_1} \\ \frac{\partial f}{\partial \beta_0 \beta_1} & \frac{\partial f}{\partial \beta_1^2} \end{pmatrix} = \begin{pmatrix} 2n & 2\sum x_i \\ 2\sum x_i & 2\sum x_i^2 \end{pmatrix}$$

Hessian

- The matrix is positive definite if $n > 0$ and the determinant is > 0.
- This corresponds to $\sum(x_i - \bar{x})^2 > 0$, which holds if not all values of x_i are the same.
- If this does not hold, simple linear regression is not meaningful, so this is a reasonable assumption.

Comments

- Note that we can express the estimated slope as follows:

$$\hat{\beta}_1 = \hat{\rho}_{xy} \frac{\hat{\sigma}_y}{\hat{\sigma}_x}.$$

- Thus, the best fitting line has a slope equal to the correlation times the ratio of the standard deviations.

Comments

- If we reverse the role of **x** and **y**, we simply invert the ratio of the standard deviations.
- If we center and scale our data first so that the resulting vectors have mean 0 and variance 1, our slope is exactly the correlation between the vectors.

R code

Fit a simple linear regression using the `diamond` dataset.

```
> library(UsingR)
> data(diamond)
> x = diamond$carat
> y = diamond$price
> beta1 = cor(x, y) * sd(y) / sd(x)
> beta0 = mean(y) - beta1 * mean(x)
> c(beta0, beta1)
[1] -259.6259 3721.0249
> # versus estimate with lm
> coef(lm(y ~ x))
(Intercept)           x
  -259.6259   3721.0249
```

Fitted values

- The term $y_i = \beta_0 + \beta_1 x_i$, for $i = 1, \ldots n$ is called the fitted value for the i^{th} observation.
- We define $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)'$ to be the vector of fitted values.
- This can be expressed as $\hat{\mathbf{y}} = \hat{\beta}_0 \mathbf{J}_n + \hat{\beta}_1 \mathbf{x}$.

Fitted values

- Whereas $\mathbf{y}$ is in $\mathbb{R}^n$, $\hat{\mathbf{y}}$ is in Γ, the two dimensional linear subspace of $\mathbb{R}^n$ spanned by the two vectors, $\mathbf{J}_n$ and $\mathbf{x}$.
- We can think of our least squares criteria as minimizing

$$||\mathbf{y} - \hat{\mathbf{y}}||$$

  over all $\hat{\mathbf{y}} \in \Gamma$.
- The fitted values are the orthogonal projection of the observed data onto this linear subspace.

R code

Compute the predicted values for the `diamond` data.

```
> yhat = beta0 + beta1*x
> yhat
 [1]  372.9483  335.7381  372.9483  410.1586  670.6303  335.7381  298.5278
 [8]  447.3688  521.7893  298.5278  410.1586  782.2611  335.7381  484.5791
[15]  596.2098  819.4713  186.8971  707.8406  670.6303  745.0508  410.1586
[22]  335.7381  372.9483  335.7381  372.9483  410.1586  372.9483  410.1586
[29]  372.9483  298.5278  372.9483  931.1020  931.1020  298.5278  335.7381
[36]  335.7381  596.2098  596.2098  372.9483  968.3123  670.6303 1042.7328
[43]  410.1586  670.6303  670.6303  298.5278  707.8406  298.5278

> predict(fit)
        1         2         3         4         5         6         7         8
 372.9483  335.7381  372.9483  410.1586  670.6303  335.7381  298.5278  447.3688
        9        10        11        12        13        14        15        16
 521.7893  298.5278  410.1586  782.2611  335.7381  484.5791  596.2098  819.4713
       17        18        19        20        21        22        23        24
 186.8971  707.8406  670.6303  745.0508  410.1586  335.7381  372.9483  335.7381
       25        26        27        28        29        30        31        32
 372.9483  410.1586  372.9483  410.1586  372.9483  298.5278  372.9483  931.1020
       33        34        35        36        37        38        39        40
 931.1020  298.5278  335.7381  335.7381  596.2098  596.2098  372.9483  968.3123
       41        42        43        44        45        46        47        48
 670.6303 1042.7328  410.1586  670.6303  670.6303  298.5278  707.8406  298.5278
```

R code

Compute the predicted value for $x = 0.20$ for the `diamond` data.

```
> beta0 + beta1 * .20
[1] 484.5791
> predict(lm(y ~ x), newdata = data.frame(x = .2))
       1
484.5791
```

Residuals

- A residual, denoted e_i, is the difference between the observed and the predicted value of y_i, i.e., $e_i = y_i - \hat{y}_i$.
- The residuals show how far the individual data points fall from the regression function.
- We define $\mathbf{e} = (e_1, \ldots, e_n)'$ to be the vector of residuals.
- This can be expressed as $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$.

Residuals

- Each residual is the vertical distance between $\mathbf{y}$ and the fitted regression line.
- Geometrically, the residuals are the orthogonal vector pointing to $\mathbf{y}$ from $\hat{\mathbf{y}}$.
- Least squares can be thought of as minimizing the sum of the squared residuals.
- The quantity $||\mathbf{e}||^2$ represents the sum of the squared errors while $\frac{1}{n-2}||\mathbf{e}||^2$ is the mean squared error or the residual variance.

Residuals

The regression line and the residuals have the following properties:

- The sum of the residuals is zero.
- The sum of the squared residuals is a minimum of f.
- $\mathbf{y}'\mathbf{J}_n = \hat{\mathbf{y}}'\mathbf{J}_n$
- $\mathbf{x}'\mathbf{e} = 0$ and $\hat{\mathbf{y}}'\mathbf{e} = 0$
- The regression line always goes through the point $(\bar{x}, \bar{y})$.

R code

Compute the residuals for the `diamond` data set.

```
> y-yhat
 [1] -17.9483176  -7.7380691 -22.9483176 -85.1585661 -28.6303057   6.2619309
 [7]  23.4721795  37.6311854 -38.7893116  24.4721795  51.8414339  40.7389488
[13]   0.2619309  13.4209369  -1.2098087  40.5287002  36.1029250 -44.8405542
[19]  79.3696943 -25.0508027  57.8414339   9.2619309 -20.9483176  -3.7380691
[25] -19.9483176  27.8414339 -54.9483176   8.8414339 -26.9483176  16.4721795
[31] -22.9483176 -13.1020453 -12.1020453  -0.5278205   3.2619309   2.2619309
[37]  -1.2098087 -43.2098087 -27.9483176 -23.3122938 -15.6303057  43.2672091
[43]  32.8414339   7.3696943   4.3696943 -11.5278205 -14.8405542  17.4721795

> resid(fit)
          1           2           3           4           5           6 
-17.9483176  -7.7380691 -22.9483176 -85.1585661 -28.6303057   6.2619309 
          7           8           9          10          11          12 
 23.4721795  37.6311854 -38.7893116  24.4721795  51.8414339  40.7389488 
         13          14          15          16          17          18 
  0.2619309  13.4209369  -1.2098087  40.5287002  36.1029250 -44.8405542 
         19          20          21          22          23          24 
 79.3696943 -25.0508027  57.8414339   9.2619309 -20.9483176  -3.7380691 
         25          26          27          28          29          30 
-19.9483176  27.8414339 -54.9483176   8.8414339 -26.9483176  16.4721795 
         31          32          33          34          35          36 
-22.9483176 -13.1020453 -12.1020453  -0.5278205   3.2619309   2.2619309 
         37          38          39          40          41          42 
 -1.2098087 -43.2098087 -27.9483176 -23.3122938 -15.6303057  43.2672091 
         43          44          45          46          47          48 
 32.8414339   7.3696943   4.3696943 -11.5278205 -14.8405542  17.4721795 
```

Connecting the pieces

- Here is an alternative approach towards estimating the parameters of the simple linear regression model that links back to the single variable regression models.
- Consider fixing β_1 and minimizing the least square criteria

$$||\mathbf{y} - \beta_1 \mathbf{x} - \beta_0 \mathbf{J}_n||^2$$

with respect to β_0.

Connecting the pieces

- Let $\hat{\beta}_0(\beta_1)$ be the least squares minimum for β_0 for a given value of β_1.
- Note β_0 is now a function of β_1.
- Following the results from mean only regression:

$$\hat{\beta}_0(\beta_1) = \frac{1}{n}(\mathbf{y} - \beta_1 \mathbf{x})\mathbf{J}_n = \bar{y} - \beta_1 \bar{x}.$$

Connecting the pieces

- Therefore, plugging this into the least squares equation, we know that

$$
\begin{aligned}
||\mathbf{y} - \beta_0 \mathbf{J}_n - \beta_1 \mathbf{x}||^2 &\geq ||\mathbf{y} - (\bar{y} - \beta_1 \bar{x})\mathbf{J}_n - \beta_1 \mathbf{x})||^2 \\
&= ||\mathbf{y} - \bar{y}\mathbf{J}_n - \beta_1 (\mathbf{x} - \bar{x}\mathbf{J}_n)||^2 \\
&= ||\tilde{\mathbf{y}} - \beta_1 \tilde{\mathbf{x}}||^2, \qquad (1)
\end{aligned}
$$

  where $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{x}}$ are the centered versions of $\mathbf{y}$ and $\mathbf{x}$, respectively.

Connecting the pieces

- We know from previously that (1) is minimized by

$$\hat{\beta}_1 = \frac{< \tilde{\mathbf{x}}, \tilde{\mathbf{y}} >}{< \tilde{\mathbf{x}}, \tilde{\mathbf{x}} >} = \frac{(\mathbf{x} - \bar{x}\mathbf{J}_n)'(\mathbf{y} - \bar{y}\mathbf{J}_n)}{(\mathbf{x} - \bar{x}\mathbf{J}_n)'(\mathbf{x} - \bar{x}\mathbf{J}_n)}.$$

- Plugging this into $\hat{\beta}_0(\hat{\beta}_1)$ we get that

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Connecting the pieces

- Note, the slope estimate when including an intercept is identical to that of regression through the origin after centering the data.
- The intercept simply forces the line through $(\bar{x}, \bar{y})$.

Yet another approach

- Compute an orthogonal basis for Γ, for example $\mathbf{u}_1 = \mathbf{J}_n$ and $\mathbf{u}_2 = \mathbf{x} - \bar{x}\mathbf{J}_n$.
- The projection of $\mathbf{y}$ onto Γ can be expressed as the sum of the individual projections of $\mathbf{y}$ onto $\mathbf{u}_1$ and $\mathbf{y}$ onto $\mathbf{u}_2$, i.e. $\hat{\mathbf{y}} = \hat{\mathbf{y}}_1 + \hat{\mathbf{y}}_2$.

Yet another approach

- The projection of $\mathbf{y}$ onto $\mathbf{u}_1$ can be expressed as $\hat{\mathbf{y}}_1 = \hat{\alpha}_0 \mathbf{J}_n$ where $\hat{\alpha}_0 = \bar{y}$.
- The projection of $\mathbf{y}$ onto $\mathbf{u}_2$ can be expressed as $\hat{\mathbf{y}}_1 = \hat{\alpha}_1(\mathbf{x} - \bar{x}\mathbf{J}_n)$ where

$$\hat{\alpha}_1 = \frac{(\mathbf{x} - \bar{x}\mathbf{J}_n)'\mathbf{y}}{(\mathbf{x} - \bar{x}\mathbf{J}_n)'(\mathbf{x} - \bar{x}\mathbf{J}_n)} = \frac{(\mathbf{x} - \bar{x}\mathbf{J}_n)'(\mathbf{y} - \bar{y}\mathbf{J}_n)}{(\mathbf{x} - \bar{x}\mathbf{J}_n)'(\mathbf{x} - \bar{x}\mathbf{J}_n)}.$$

Yet another approach

- Note that $\hat{\alpha}_1 = \hat{\beta}_1$ from before.
- Thus, we can write

$$\begin{aligned}
\hat{\mathbf{y}} &= \hat{\mathbf{y}}_1 + \hat{\mathbf{y}}_2 \\
&= \bar{y}\mathbf{J}_n + \hat{\beta}_1(\mathbf{x} - \bar{x}\mathbf{J}_n) \\
&= (\bar{y} - \hat{\beta}_1\bar{x})\mathbf{J}_n + \hat{\beta}_1\mathbf{x}
\end{aligned}$$

- Setting $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ provides the familiar solution.