Linear Regression Models

STATS 762 – Lecture Slides 2

March 11, 2019

The Model Space

For any GLM, the vector space defined by $\mathbf{X}\boldsymbol{\beta}$ defines the restrictions placed on the mean vector by the model through the link function:

Ordinary Regression: $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$

Logistic Regression: $\log(\mathbf{p}/(\mathbf{1} - \mathbf{p})) = \mathbf{X}\boldsymbol{\beta}$

Poisson Regression: $\log(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$

Consequences

Therefore, in each case the model space defines the essence of the model (it defines the restrictions placed on $\boldsymbol{\mu}$).

- Two different model matrices $\mathbf{X}$ and $\mathbf{X}^*$ that generate the same model space represent different *parameterizations* of the same model.
- Since there are infinitely many bases for the column space of $\mathbf{X}$, they are infinitely many ways of parameterizing any model.

Alternative Model Matrices

Consider the following ways of adapting the catheter model matrix:

1. convert height (X_1) from inches to centimetres and weight (X_2) from pounds to kilograms.
2. centre (subtract the mean) both X_1 and X_2.
3. log both X_1 and X_2.

Which of these represent different parameterizations of our original model and which represent a different model?

Reparameterizing a Model

In order to find a different parameterization of a given model we need to find a different basis for the model space.

- for the catheter example we need a set of three linearly independent vectors that are linear combinations of the column vectors of $\mathbf{X}$.

Alternatively, we can say that the new model matrix $\mathbf{X}^*$ is a reparameterisation of the old model matrix $\mathbf{X}$ if (and only if) we can find a *non-singular* 3×3 matrix $\mathbf{A}$ such that

$$\mathbf{X}^* = \mathbf{XA}$$

Reparameterizing a Model (cont.)

For our three proposed parameterizations, we can identify such a matrix for the first two scenarios:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2.54 & 0 \\ 0 & 0 & 0.454 \end{bmatrix} \qquad \mathbf{A} = \begin{bmatrix} 1 & -40.36 & -38.13 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

but not for the third scenario.

Same Projection Matrix

Consider fitting a alternative linear model which has model matrix as $\mathbf{W} = \mathbf{XA}$ where $\mathbf{A}$ is a nonsingular $(k+1) \times (k+1)$ matrix.

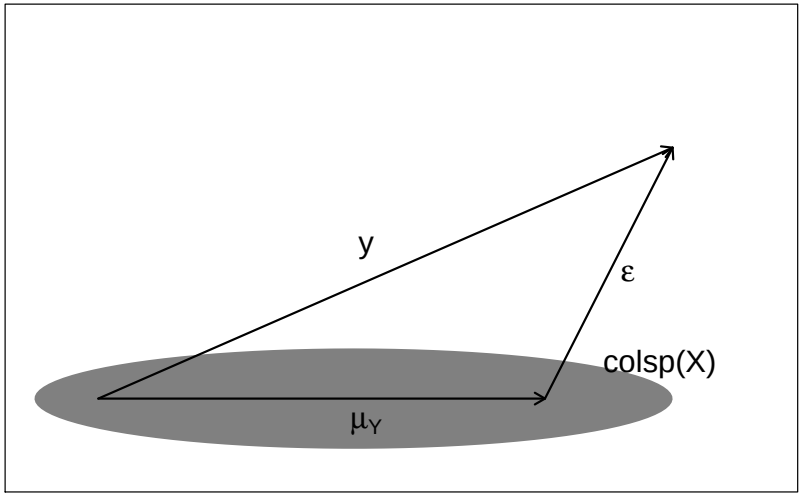
- the vector of fitted values for this alternative model will be identical to that for the original model since the projection matrices for the two models will be identical.

$$\mathbf{W}\left(\mathbf{W}^t\mathbf{W}\right)^{-1}\mathbf{W}^t = \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t$$

Fitting the Ordinary Regression Model

Linear Model

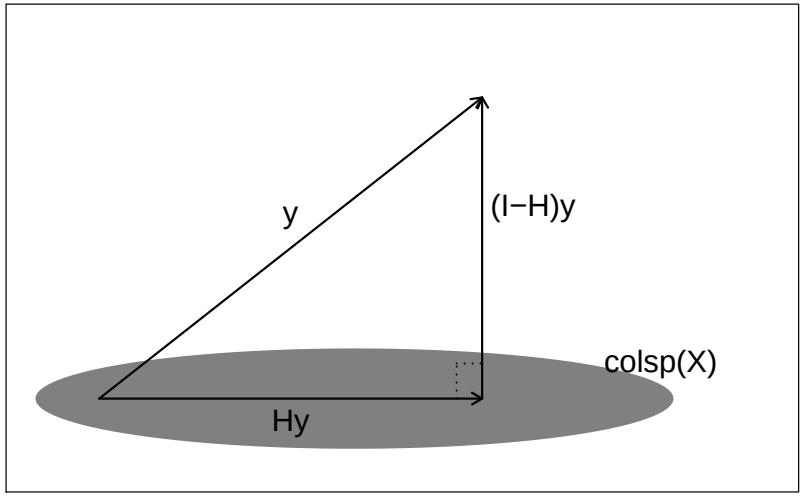
$$\mathbf{y} = \boldsymbol{\mu}_\mathbf{Y} + \boldsymbol{\epsilon}$$



$$\boldsymbol{\mu}_\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$$
$$\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}_2)$$

Fitted Model

$$\mathbf{y} = \hat{\boldsymbol{\mu}}_\mathbf{Y} + \mathbf{r}$$



$$\hat{\boldsymbol{\mu}}_\mathbf{Y} = \mathbf{H}\mathbf{Y}$$
$$\mathbf{r} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t$$

Parameter Estimates

The column vectors of $\mathbf{X}$ form a basis for colsp($\mathbf{X}$). Thus there is a unique linear combination of the columns of $\mathbf{X}$ that produce $\hat{\boldsymbol{\mu}}_{\mathbf{Y}}$. Putting the coefficients for this relation in a vector $\hat{\boldsymbol{\beta}}$ gives $\hat{\boldsymbol{\mu}}_{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.

$$\begin{aligned} \text{Combining:} \quad & \hat{\boldsymbol{\mu}}_{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \\ \text{with:} \quad & \hat{\boldsymbol{\mu}}_{\mathbf{Y}} = \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y} \\ \text{gives:} \quad & \hat{\boldsymbol{\beta}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y} \end{aligned}$$

Parameter Estimates for Catheter Data

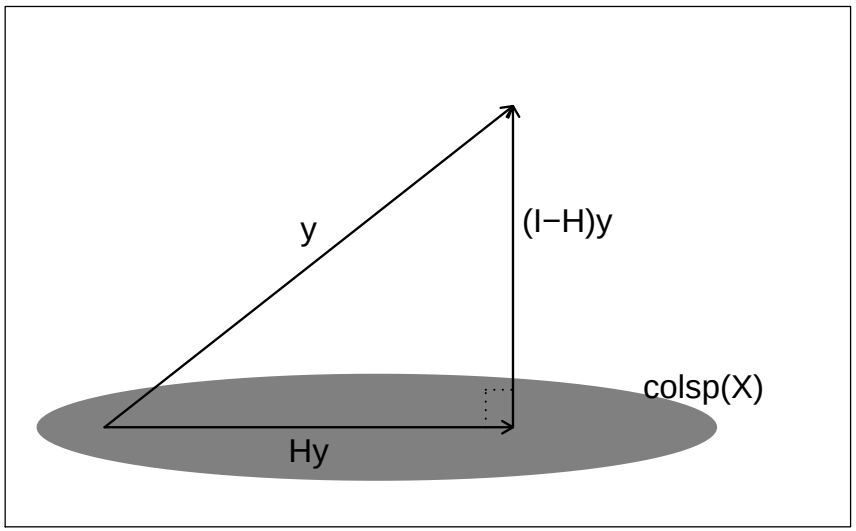
To get $\hat{\beta}$ for our catheter data:

```
> solve(t(X)%*%X)%*%t(X)%*%y
          [,1]
   20.3757645
x1  0.2107473
x2  0.1910949
```

Fitting the Linear Model

Linear Model: $\mathbf{Y} = \boldsymbol{\mu}_\mathbf{Y} + \boldsymbol{\epsilon}$ where $\boldsymbol{\mu}_\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$.

Fitted Model: $\mathbf{Y} = \hat{\boldsymbol{\mu}}_\mathbf{Y} + \mathbf{r}$ where $\hat{\boldsymbol{\mu}}_\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}}$.



$$\begin{aligned} \hat{\boldsymbol{\mu}}_\mathbf{Y} &= \mathbf{H}\mathbf{Y} \\ \hat{\boldsymbol{\beta}} &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y} \\ \mathbf{r} &= (\mathbf{I} - \mathbf{H})\mathbf{Y} \end{aligned}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$

Distributions of $\hat{\mu}_{\mathbf{Y}}$, $\hat{\beta}$ and $\mathbf{r}$

$\hat{\mu}_{\mathbf{Y}}$, $\hat{\beta}$ and $\mathbf{r}$ are all linear transformations of $\mathbf{Y}$ – i.e. they can be written as a matrix times $\mathbf{Y}$:

$$\hat{\mu}_{\mathbf{Y}} = \left[\mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\right]\mathbf{Y}$$

$$\hat{\beta} = \left[\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\right]\mathbf{Y}$$

$$\mathbf{r} = [\mathbf{I} - \mathbf{H}]\,\mathbf{Y}$$

As a result we can derive each of their distributions from the distribution of $\mathbf{Y}$.

Linear Transformations of Random Vectors

The following properties of linear transformations of random vectors are extremely useful. Consider:

$$\mathbf{U} = \mathbf{M}\mathbf{V} + \mathbf{C}$$

where $\mathbf{M}$ and $\mathbf{C}$ are fixed and $\mathbf{V}$ is a random vector with mean vector $\boldsymbol{\mu}_{\mathbf{V}}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{V}}$.

$$\underbrace{\begin{bmatrix} U_1 \\ \vdots \\ U_p \end{bmatrix}}_{\mathbf{U}} = \underbrace{\begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1k} \\ \vdots & \vdots & & \vdots \\ m_{p1} & m_{p2} & \cdots & m_{pk} \end{bmatrix}}_{\mathbf{M}} \underbrace{\begin{bmatrix} V_1 \\ \vdots \\ V_k \end{bmatrix}}_{\mathbf{V}} + \underbrace{\begin{bmatrix} c_1 \\ \vdots \\ c_p \end{bmatrix}}_{\mathbf{C}}$$

Then $\boldsymbol{\mu}_{\mathbf{U}} = \mathbf{M}\boldsymbol{\mu}_{\mathbf{V}} + \mathbf{C}$ and $\boldsymbol{\Sigma}_{\mathbf{U}} = \mathbf{M}\boldsymbol{\Sigma}_{\mathbf{V}}\mathbf{M}^t$.

- Further, if $\mathbf{V}$ has a Normal distribution so does $\mathbf{U}$.

Distributions of $\hat{\boldsymbol{\mu}}_{\mathbf{Y}}$, $\hat{\boldsymbol{\beta}}$ and $\mathbf{r}$

Given the previous slide and

$$\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}_{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta},\ \boldsymbol{\Sigma}_{\mathbf{Y}} = \sigma^2\mathbf{I})$$

we can deduce the distributions of $\hat{\boldsymbol{\mu}}_{\mathbf{Y}}$, $\hat{\boldsymbol{\beta}}$ and $\mathbf{r}$:

$\hat{\boldsymbol{\mu}}_{\mathbf{Y}} \sim$

$\hat{\boldsymbol{\beta}} \sim$

$\mathbf{r} \sim$

Inference for $\hat{\boldsymbol{\beta}}$

We will use $\hat{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{\beta},\ \sigma^2 (\mathbf{X}^t\mathbf{X})^{-1})$ as the basis for inference about the regression coefficients.

- The diagonal elements of $\sigma^2 (\mathbf{X}^t\mathbf{X})^{-1}$ are the variances of the $\hat{\beta}_i$'s and the off-diagonal elements are covariances between pairs of $\hat{\beta}_i$'s. All of these depend on σ^2.
- If σ^2 is not known (which is almost always the case) then we need to estimate σ^2 before we can proceed.

Estimating σ^2

The standard way of estimating σ^2 is to divide the residual sum of squares by the residual degrees of freedom:

$$\hat{\sigma}^2 = \frac{\mathbf{r}^t\mathbf{r}}{n-k-1}$$

- To justify this estimator we will deduce the distribution of $\mathbf{r}^t\mathbf{r} = \|\mathbf{r}\|^2$.
- We start by considering the distribution of $\|\boldsymbol{\epsilon}\|^2$.
- Then we look at the squared length of projections of $\boldsymbol{\epsilon}$.

The Distribution of $\|\boldsymbol{\epsilon}\|^2$

Since the ϵ_i's are independent N(0, σ^2) random variables,

$$\frac{\|\boldsymbol{\epsilon}\|^2}{\sigma^2} = \frac{\boldsymbol{\epsilon}^t\boldsymbol{\epsilon}}{\sigma^2} = \left(\frac{\epsilon_1}{\sigma}\right)^2 + \ldots + \left(\frac{\epsilon_n}{\sigma}\right)^2$$

must have a χ^2_n distribution.

- We say that $\|\boldsymbol{\epsilon}\|^2$ has a scaled χ^2_n distribution and use the notation

$$\|\boldsymbol{\epsilon}\|^2 \sim \sigma^2 \times \chi^2_n$$

.

Projections of ϵ

Consider the projection $\mathbf{P}\epsilon$ onto a subspace S.

- Given that the distribution of ϵ is completely symmetrical around the origin, the distribution of $\mathbf{P}\epsilon$ – and therefore the distribution of $\|\mathbf{P}\epsilon\|^2$ – depends only on the dimension of S.
- So if we can deduce the distribution of $\|\mathbf{P}\epsilon\|^2$ for a particular subspace of dimension d, the distribution is the same for any other subspace of that dimension.

Projections of ϵ (cont.)

Consider the projection of $\boldsymbol{\epsilon}$ onto a subspace of dimension 2 which is spanned by:

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \text{ and } \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \quad \text{Then} \quad \mathbf{P}\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Thus $\|\mathbf{P}\boldsymbol{\epsilon}\|^2 = \epsilon_1^2 + \epsilon_2^2$ and has a $\sigma^2 \times \chi_2^2$ distribution.

- This is also true for any other subspace of dimension 2.

Some Key Results

1. If $\mathbf{P}\epsilon$ is a projection onto a subspace of dimension d then

$$\frac{\|\mathbf{P}\epsilon\|^2}{\sigma^2} \sim \chi^2_d$$

2. The residual vector $\mathbf{r} = (\mathbf{I} - \mathbf{H})\mathbf{y}$ is the orthogonal projection of $\mathbf{y}$ onto the error space. This is equivalent to projecting ϵ onto the error space (why?).

3. The expected value of a χ^2_d distribution is d.

Estimating σ^2

As a consequence of the results on the previous page we can conclude:

- $\|\mathbf{r}\|^2/\sigma^2$ has a χ^2_{n-k-1} distribution.
- $\mathrm{E}(\|\mathbf{r}\|^2) = (n-k-1)\sigma^2$

Thus $\hat{\sigma}^2 = \dfrac{\mathbf{r}^t\mathbf{r}}{n-k-1}$ is an unbiased estimate of σ^2.

The Catheter Example: $\hat{\beta}$

For the catheter example we can calculate $\hat{\beta}$:

```
> x1<-c(42.8,63.5,37.5,39.5,45.5,38.5,
+       43.0,22.5,37.0,23.5,33.0,58.0)
> x2<-c(40.0,93.5,35.5,30.0,52.0,17.0,
+       38.5, 8.5,33.0, 9.5,21.0,79.0)
> X<-cbind(1,x1,x2)
> y<-matrix(c(37,50,34,36,43,28,37,20,34,30,38,47),12,1)
> BETAhat<- solve(t(X)%*%X)%*%t(X)%*%y
> BETAhat
          [,1]
   20.3757645
x1  0.2107473
x2  0.1910949
```

The Catheter Example: $(\mathbf{X}^t\mathbf{X})^{-1}$

To get $(\mathbf{X}^t\mathbf{X})^{-1}$:

```
> XtXinv<-solve(t(X)%*%X)
> XtXinv
                            x1           x2
    4.9262358 -0.197172444  0.081695704
x1 -0.1971724  0.008363701 -0.003681904
x2  0.0816957 -0.003681904  0.001754749
```

- We need to multiply $(\mathbf{X}^t\mathbf{X})^{-1}$ by $\hat{\sigma}^2$ to get our estimated covariance matrix for $\hat{\boldsymbol{\beta}}$.

The Catheter Example: Residuals

For the catheter example we can calculate the residuals from:

```
> res<-(diag(12)-H)%*%y
> res
              [,1]
 [1,] -0.03954422
 [2,] -1.62559020
 [3,] -1.06265657
 [4,]  1.56687083
 [5,]  3.09829930
 [6,] -3.73814817
 [7,]  0.20494867
 [8,] -6.74188499
 [9,] -0.47954567
[10,]  2.85627283
[11,]  6.65658228
[12,] -0.69560408
```

The Catheter Example: $\hat{\sigma}^2$

To get $\hat{\sigma}^2$:

```
> sig2hat<-(t(res)%*%res)/(12-2-1)
> sig2hat
          [,1]
[1,] 14.27543
```

Thus to get the estimated covariance matrix for $\hat{\boldsymbol{\beta}}$

```
> as.numeric(sig2hat)*XtXinv
                          x1           x2
   70.324134 -2.81472140  1.16624129
x1 -2.814721  0.11939543 -0.05256076
x2  1.166241 -0.05256076  0.02504980
```

The Catheter Example: Inference for $\hat{\boldsymbol{\beta}}$

Given we now have $\hat{\boldsymbol{\beta}}$ and an estimated covariance matrix for $\hat{\boldsymbol{\beta}}$, we can do the standard types of statistical inference (find confidence intervals and do hypothesis tests.

- Since we had to estimate σ^2, we must use a t_{n-k-1} reference distribution.

To get a $(1-\alpha)100\%$ confidence interval for β_i:

$$\hat{\beta}_i \pm t_{n-k-1}(1-\alpha/2) \times \text{se}(\hat{\beta}_i)$$

To test H_0: $\beta_i = \text{constant}$, calculate

$$\text{t-stat} = \frac{\hat{\beta}_i - \text{constant}}{\text{se}(\hat{\beta}_i)}$$

$$\text{p-value} = 2 \times \Pr(t_{n-k-1} \geq |\text{t-stat}|)$$

The Catheter Example: Inference for $\hat{\beta}$

For example, to get a 95% confidence interval for β_1:

$$\begin{aligned} \hat{\beta}_1 &\pm t_9(.975) \times \text{se}(\hat{\beta}_1) \\ 0.211 &\pm 2.26\sqrt{.119} \\ 0.211 &\pm 0.782 \end{aligned}$$

Or to test H_0:$\beta_1 = 0$

$$\text{t-stat} = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)} = \frac{0.211}{\sqrt{.119}} = 0.611$$

$$\text{p-value} = 2 \times \Pr(t_9 \geq |t - stat|) = 2 \times \Pr(t_9 \geq 0.611) = 0.56$$

The Catheter Example: Predicted Values

Suppose we wanted to predict the catheter length for a child that was 44 inches tall (X_1) and weighed 35 pounds (X_2).

Using the estimated covariance matrix from slide 115:

$$\text{estimate} = \begin{bmatrix} 1 & 44 & 35 \end{bmatrix} \begin{bmatrix} 20.376 \\ 0.211 \\ 0.191 \end{bmatrix} = 36.34$$

$$\text{est. var.} = \begin{bmatrix} 1 & 44 & 35 \end{bmatrix} \begin{bmatrix} 70.324 & -2.815 & 1.166 \\ -2.815 & 0.119 & -0.053 \\ 1.166 & -0.053 & 0.025 \end{bmatrix} \begin{bmatrix} 1 \\ 44 \\ 35 \end{bmatrix} = 4.21$$

The $(\mathbf{X}^t\mathbf{X})^{-1}$ Matrix

Clearly the $(\mathbf{X}^t\mathbf{X})^{-1}$ matrix has a big impact on any inference related to the fitted coefficients.

The "ideal" situation is for $(\mathbf{X}^t\mathbf{X})^{-1}$ to be a diagonal matrix.

- What would this imply about the $\widehat{\beta}$'s?
- What does this imply about the columns of $\mathbf{X}$?

ML Estimation for Ordinary Regression

For ordinary regression, we could equate maximum likelihood estimation to the orthogonal projection of $\mathbf{Y}$ onto the model space which led to the explicit expression

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}.$$

Alternatively, these expressions can be deduced by (i) writing down the log likelihood function, (ii) taking the partial derivatives with respect to the β's and setting these equal to zero and (iii) solving for the β's. Note that step (ii) results in a set equations that are linear in the β's and thus an explicit solution is obtained.

MLE Estimation for GLM's

For most GLM's (including logistic and Poisson regressions) the equations produced by step (ii) are not linear and thus explicit solutions cannot be found.

- Solutions can be found using the iteratively re-weighted least squares (IRLS) algorithm.
- IRLS usually converges to the correct solution and usually does so quite quickly.
- Used for the `glm` function in *R*.

Brief Outline of IRLS

IRLS involves a series of projection onto the model space. However, it doesn't make sense to project $\mathbf{Y}$ onto the model space since $\boldsymbol{\mu}_{\mathbf{Y}}$ is not an element of the model space (the link function gets in the way). So conceptually, the idea is to:

- Create a vector $\mathbf{Y}^*$ such that $\boldsymbol{\mu}_{\mathbf{Y}^*} = \mathbf{X}\boldsymbol{\beta}$ and $\mathbf{Y}^* - \boldsymbol{\mu}_{\mathbf{Y}^*}$ is "equivalent" to $\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}}$.
- Project $\mathbf{Y}^*$ onto the model space in order to estimate $\boldsymbol{\mu}_{\mathbf{Y}^*}$ and thus $\boldsymbol{\beta}$. However we need to use a weighted projection since the elements of $\mathbf{Y}^*$ don't all have the same variance.

A Bit on Weighted Least Squares

For Normal response regression suppose that the variances are not all equal but $\text{var}(Y_i) = \sigma_i^2$. Let $w_i = 1/\sigma_i^2$ and create a diagonal matrix $\mathbf{W}$ with diagonal entries w_1 to w_n.

The weighted least squares estimates for $\boldsymbol{\mu}_{\mathbf{Y}}$ and for $\boldsymbol{\beta}$ are:

$$\begin{aligned} \hat{\boldsymbol{\mu}}_{\mathbf{Y}} &= \mathbf{X}\left(\mathbf{X}^t\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{W}\mathbf{Y} \\ \hat{\boldsymbol{\beta}} &= \left(\mathbf{X}^t\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{W}\mathbf{Y} \end{aligned}$$

Back to IRLS

The problem is that we need to know $\boldsymbol{\beta}$ to find $\mathbf{Y}^*$ and the corresponding weights. This is where the iterative part comes in:

1. Set the entries of $\boldsymbol{\beta}$ to some starting values.
2. Based on these values find the corresponding $\mathbf{Y}^*$ and their weights.
3. Use weighted least squares to estimate $\boldsymbol{\beta}$.
4. Repeat 2 and 3 until estimates converge.

Sampling Distribution of $\widehat{\boldsymbol{\beta}}$

Once the algorithm has converged, the sampling distribution for $\widehat{\boldsymbol{\beta}}$ can be approximated by

$$\mathrm{E}\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta} \qquad\qquad \widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}} = (\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}$$

where $\mathbf{W}$ contains the weights for the last iteration of IRLS.

- This sampling distribution is asymptotically Normal (we need to invoke the central limit theorem).
- We should treat any inference based on this sampling distribution with caution.

The CHD Logistic Regression Example

Recall the fitted logistic regression model for the CHD data from our first set of lecture slides:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.27844    1.13054  -4.669 3.03e-06 ***
age          0.11032    0.02402   4.593 4.37e-06 ***
```

We can get out the covariance matrix for $\widehat{\boldsymbol{\beta}}$:

```
> summary(chd.glm)$cov.scaled
             (Intercept)           age
(Intercept)  1.27811454 -0.0265567468
age         -0.02655675  0.0005768822
```

Predictions

To get the estimated value of the linear predictor ($\beta_0 + \beta_1 \text{age}$) at age 25:

```
> new.df<-data.frame(age=25)
> predict(chd.glm,new.df,type="link",se.fit = TRUE)
$fit
-2.520425
$se.fit
[1] 0.55752
```

The standard error comes from the square root of:

$$\text{est. var.} = \begin{bmatrix} 1 & 25 \end{bmatrix} \begin{bmatrix} 1.278114 & -0.026556 \\ -0.026556 & 0.0005768 \end{bmatrix} \begin{bmatrix} 1 \\ 25 \end{bmatrix} = 0.31082$$

Predictions (cont.)

To get predictions for the probability (rather than the linear predictor):

```
> predict(chd.glm,new.df,type="response",se.fit = TRUE)
$fit
0.07443868
$se.fit
0.03841177
```

The fitted value comes from applying the logistic function to the estimate for the linear predictor:

```
> exp(-2.520425)/(1+exp(-2.520425))
[1] 0.07443866
```

Predictions (cont.)

The standard error is estimated using the delta method:

$$\text{var}\left(f(v)\right) \approx f'(v)^2 \times \text{var}(v)$$

.

For logistic regression:

$$f(v) = \frac{e^v}{1+e^v} \quad \text{and} \quad f'(v) = \frac{e^v}{(1+e^v)^2}$$

.

Plugging in our values we get:

```
> (exp(-2.520425)/((1+exp(-2.520425))^2))^2* 0.3108286
[1] 0.001475463
> sqrt(0.001475463)
[1] 0.03841176
```

Comments

The `predict` function allows you to get estimates of either the linear predictor (default) or the mean of the response – it will also supply standard errors.

If you want to create a confidence interval for the mean of the response it is almost always better to first create an interval for the linear predictor and then transform.

- the Normal distribution approximation is usually better for the sampling distribution of the estimated linear predictor than it is for the sampling distribution of the estimated response.

The Added Variable F-Test

Returning to ordinary regression, one of the most useful tools for doing statistical inference is the added variable F-test.

- Compares a submodel to a full model – the explanatory variables in the submodel must be a subset of those in the full model.
- Evaluates the evidence that the extra variables in the full model contribute to explaining the response. The null hypothesis is that they don't contribute.

In what follows a subscript S denotes the submodel and a subscript F denotes the full model.

The Added Variable F-Test (cont.)

Consider dividing the response vector into three components:

$$\begin{aligned} \mathbf{y} &= (\mathbf{H_S} + \mathbf{H_F} - \mathbf{H_S} + \mathbf{I} - \mathbf{H_F})\mathbf{y} \\ &= \mathbf{H_S y} + (\mathbf{H_F} - \mathbf{H_S})\mathbf{y} + (\mathbf{I} - \mathbf{H_F})\mathbf{y} \end{aligned}$$

- $\mathbf{H_S y}$ is the projection onto the model space for the submodel.
- $(\mathbf{I} - \mathbf{H_F})\mathbf{y}$ is the projection onto the error space for the full model.
- $(\mathbf{H_F} - \mathbf{H_S})\mathbf{y}$ is what is left over – under the full model it goes into the model space and under the submodel it goes into the error space.

The Added Variable F-Test (cont.)

$$\mathbf{Y} = \mathbf{H_S Y} + (\mathbf{H_F} - \mathbf{H_S})\mathbf{Y} + (\mathbf{I} - \mathbf{H_F})\mathbf{Y}$$

represents the decomposition of $\mathbf{Y}$ into three orthogonal components. That is $\mathbf{Y}$ has been projected onto 3 mutually orthogonal subspaces.

- An important consequence of this is that $\mathbf{H_S Y}$, $(\mathbf{H_F} - \mathbf{H_S})\mathbf{Y}$ and $(\mathbf{I} - \mathbf{H_F})\mathbf{Y}$ are mutually independent random vectors.

Excursion: Independent Random Vectors

Suppose we have random vectors $\mathbf{V}$ and $\mathbf{U}$:

$$\mathbf{V} = \begin{bmatrix} V_1 \\ \vdots \\ V_p \end{bmatrix} \qquad \mathbf{U} = \begin{bmatrix} U_1 \\ \vdots \\ U_q \end{bmatrix}.$$

$\mathbf{V}$ and $\mathbf{U}$ are independent if each V_i is independent of each U_j.

- If all V_i and U_j have Normal distributions then this independence requirement is the same as having $\text{cov}(V_i, U_j) = 0$ for all i and j.

Independent Normal Random Vectors

Suppose that $\mathbf{V}$ and $\mathbf{U}$ are Normal random vectors. Combine $\mathbf{V}$ and $\mathbf{U}$ into a single vector $\mathbf{W}$ and consider the partitioned covariance matrix for $\mathbf{W}$:

$$\mathbf{W} = \begin{bmatrix} \mathbf{V} \\ \mathbf{U} \end{bmatrix} \quad \boldsymbol{\Sigma}_{\mathbf{W}} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{V}} & \boldsymbol{\Sigma}_{\mathbf{VU}} \\ \boldsymbol{\Sigma}_{\mathbf{UV}} & \boldsymbol{\Sigma}_{\mathbf{U}} \end{bmatrix} \text{ where } \boldsymbol{\Sigma}_{\mathbf{UV}} = \boldsymbol{\Sigma}_{\mathbf{VU}}^{t}$$

- $\mathbf{V}$ and $\mathbf{U}$ are independent if $\boldsymbol{\Sigma}_{\mathbf{VU}} = \mathbf{0}$ which also implies that $\boldsymbol{\Sigma}_{\mathbf{UV}} = \mathbf{0}$.

Independent Normal Random Vectors (cont.)

$$\text{Let:} \quad \mathbf{W} = \begin{bmatrix} \mathbf{P_1 Y} \\ \mathbf{P_2 Y} \end{bmatrix} = \begin{bmatrix} \mathbf{P_1} \\ \mathbf{P_2} \end{bmatrix} \mathbf{Y} \qquad \text{where} \qquad \mathbf{\Sigma_Y} = \sigma^2 \mathbf{I}$$

$$\text{Thus:} \quad \mathbf{\Sigma_W} = \begin{bmatrix} \mathbf{P_1} \\ \mathbf{P_2} \end{bmatrix} \sigma^2 \mathbf{I} \begin{bmatrix} \mathbf{P_1^t} & \mathbf{P_2^t} \end{bmatrix} = \sigma^2 \begin{bmatrix} \mathbf{P_1 P_1^t} & \mathbf{P_1 P_2^t} \\ \mathbf{P_2 P_1^t} & \mathbf{P_2 P_2^t} \end{bmatrix}$$

$$= \sigma^2 \begin{bmatrix} \mathbf{P_1} & \mathbf{P_1 P_2} \\ \mathbf{P_2 P_1} & \mathbf{P_2} \end{bmatrix}$$

It follows that $\mathbf{P_1 V}$ and $\mathbf{P_2 V}$ are independent Normal random vectors if (and only if) $\mathbf{P_1 P_2} = \mathbf{0}$.

Back to the Added Variable F-Test

An F-distribution is the ratio of two independent χ^2-distributions divided by their degrees of freedom.

As $(\mathbf{I} - \mathbf{H_F})\mathbf{Y}$ and $(\mathbf{H_F} - \mathbf{H_S})\mathbf{Y}$ are independent, it follows that $\|(\mathbf{I} - \mathbf{H_F})\mathbf{y}\|^2$ and $\|(\mathbf{H_F} - \mathbf{H_S})\mathbf{y}\|^2/$ are independent.

- $\|(\mathbf{I} - \mathbf{H_F})\mathbf{y}\|^2/\sigma^2$ has a $\chi^2_{n-k_F-1}$ distribution under both the submodel and the full model.
- $\|(\mathbf{H_F} - \mathbf{H_S})\mathbf{y}\|^2/\sigma^2$ has a $\chi^2_{k_F-k_S}$ distribution under the submodel but not the full model (why?).

Back to the Added Variable F-Test (cont.)

Therefore under the submodel:

$$f_o = \frac{\|(\mathbf{H_F} - \mathbf{H_S})\mathbf{y}\|^2/(k_F - k_S)}{\|(\mathbf{I} - \mathbf{H_F})\mathbf{y}\|^2/(n - k_F - 1)} \sim \mathrm{F}_{k_F - k_S, n - k_F - 1}$$

$$\text{p-value} = \Pr(\mathrm{F}_{k_F - k_S, n - k_F - 1} \geq f_o)$$

Under the full model $(\mathbf{H_F} - \mathbf{H_S})\mathbf{y}$ is not equivalent to $(\mathbf{H_F} - \mathbf{H_S})\epsilon$.

- $\|(\mathbf{H_F} - \mathbf{H_S})\mathbf{y}\|^2/\sigma^2$ tends to be larger than it would be under the submodel.

A More Familiar Form

In most regression books, you will find the F-statistic for the added variables procedure defined in terms of the residual sums of squares (RSS) for the two models:

$$f_o = \frac{(\text{RSS}_S - \text{RSS}_F)/(k_F - k_S)}{\text{RSS}_F/(n - k_F - 1)}$$

It is not difficult to show that this is equivalent to the definition on the previous page.

Example of an Added Variable F-Test

The bottom line of the output from `summary` is an added variable F-test:

```
> summary(catheter.lm)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.3758     8.3859   2.430    0.038 *
ht            0.2107     0.3455   0.610    0.557
wt            0.1911     0.1583   1.207    0.258

Residual standard error: 3.778 on 9 degrees of freedom
Multiple R-squared: 0.8254,     Adjusted R-squared: 0.7865
F-statistic: 21.27 on 2 and 9 DF,  p-value: 0.0003888
```

- The submodel just contains the intercept and the full model contains all of the regressors.

The anova Command Output

The output from anova applied to an lm object, produces results for a sequence of F-tests.

```
> anova(catheter.lm)
Response: ca
          Df Sum Sq Mean Sq F value    Pr(>F)
ht         1 586.38  586.38 41.0760 0.0001239 ***
wt         1  20.81   20.81  1.4578 0.2580548
Residuals  9 128.48   14.28
```

- The ht line tests adding ht to the null model.
- The wt line tests adding wt to the model that contains ht.

The anova Command Output II

There is a smallish wrinkle: the anova command always uses the projection on to the error space for the model that contains all of the variables as the basis for its denominator. Note that this is not the same as using the "full model" error space.

- Suppose $\mathbf{P_1}$ is the projection matrix for the model that just contains ht and $\mathbf{P_2}$ is the projection matrix for the model that contains both ht and wt. Consider testing whether ht should be added to the null model (projection matrix $\mathbf{P_0}$).

$$\text{anova test:} \quad \text{F-stat} = \frac{\|(\mathbf{P_1}-\mathbf{P_0})\mathbf{y}\|^2/1}{\|(\mathbf{I}-\mathbf{P_2})\mathbf{y}\|^2/9} \quad \text{p-value} = \mathrm{P}\left(\mathrm{F}_{1,9} \geq \text{F-stat}\right)$$

$$\text{added variable:} \quad \text{F-stat} = \frac{\|(\mathbf{P_1}-\mathbf{P_0})\mathbf{y}\|^2/1}{\|(\mathbf{I}-\mathbf{P_1})\mathbf{y}\|^2/10} \quad \text{p-value} = \mathrm{P}\left(\mathrm{F}_{1,10} \geq \text{F-stat}\right)$$