The Geometry of Linear Models

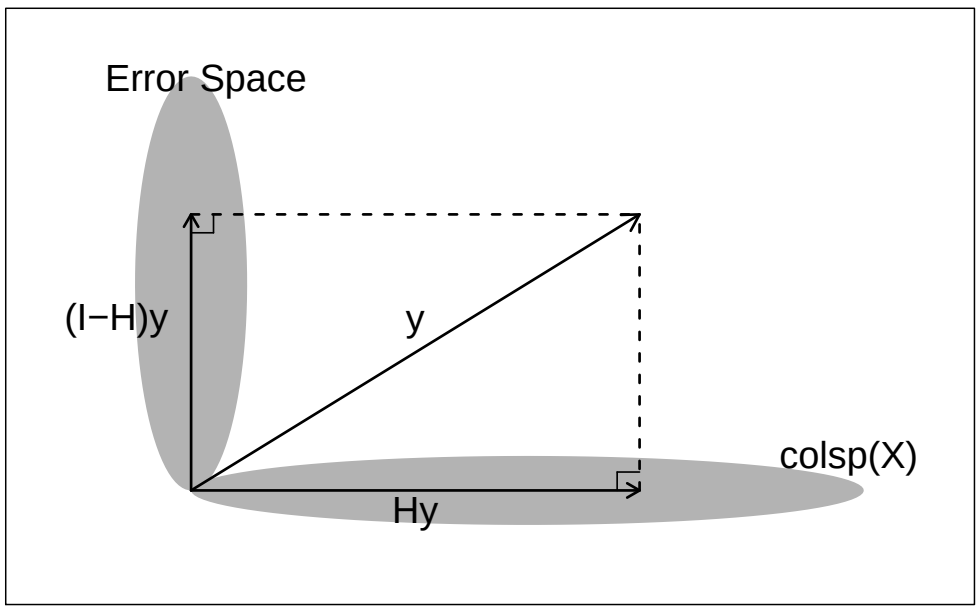
STATS 762 – Lecture Notes 3

Arden Miller

Department of Statistics, University of Auckland

Geometric Interpretation

$\hat{\boldsymbol{\mu}}_{\mathbf{Y}}$ is the orthogonal projection of $\mathbf{y}$ onto the colsp($\mathbf{X}$) and $\mathbf{r}$ is the orthogonal projection of $\mathbf{y}$ onto the orthogonal compliment of colsp($\mathbf{X}$).



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

$$\mathbf{r} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

Pythagoras: $\|\mathbf{y}\|^2 = \|\hat{\boldsymbol{\mu}}_{\mathbf{Y}}\|^2 + \|\mathbf{r}\|^2 \quad \rightarrow \quad \mathbf{y}^t\mathbf{y} = \hat{\boldsymbol{\mu}}_{\mathbf{Y}}^t\hat{\boldsymbol{\mu}}_{\mathbf{Y}} + \mathbf{r}^t\mathbf{r}.$

The Catheter Example: *R* output

If we fit a regression model for the catheter data using *R*, then the summary command gives the following output:

```
> catheter.lm<-lm(ca~.,data=catheter.df)
> summary(catheter.lm)
Coefficients:
             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
```

What We Know So Far

We know that the line

```
Residual standard error: 3.778 on 9 degrees of freedom
```

is our estimate for σ and comes from:

$$\hat{\sigma} = \sqrt{\frac{\mathbf{r}^t\mathbf{r}}{n-k-1}} = \sqrt{\frac{\mathbf{y}^t(\mathbf{I}-\mathbf{H})\mathbf{y}}{n-k-1}}$$

- This estimate is based on projecting the response vector onto the error space.
- The degrees of freedom indicate the dimension of the error space.

More of What We Know So Far

The estimated coefficients and their standard errors:

	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	

come from $\hat{\boldsymbol{\beta}} = \left[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t \right] \mathbf{y}$ and $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = \hat{\sigma}^2 (\mathbf{X}^t\mathbf{X})^{-1}$.

- For each coefficient there is also a t-test for H_0: $\beta_i = 0$.

$$\texttt{t value} = \frac{\hat{\beta}_i}{\text{se}(\hat{\beta}_i)} \qquad \texttt{Pr(>|t|)} = 2 \times \Pr(t_{n-k-1} \geq |\text{t-stat}|)$$

F-Tests

We can do a lot of useful inference based on using F-tests that compare projections of **y** onto different subspaces of R^n.

- We need a test statistic that has an F-distribution under the null hypothesis – under the alternative hypothesis the distribution should be shifted to higher values.
- An F-distribution results from the (scaled) ratio of two independent random variables that have χ^2-distributions.

$$W \sim \mathrm{F}_{\nu_1,\nu_2} \quad \text{if} \quad W = \frac{V_1/\nu_1}{V_2/\nu_2} \quad \text{where} \quad V_1 \sim \chi^2_{\nu_1} \quad V_2 \sim \chi^2_{\nu_2}$$

- The sum of squared independent N(0, 1) random variables has a χ^2-distribution. If the random variables are N(0, σ^2) the distribution is scaled (multiplied) by σ^2.

What We Need

To have the ingredients for an F-Test, we need two projections of $\mathbf{Y}$, say $\mathbf{P_1 Y}$ and $\mathbf{P_2 Y}$, such that:

- $\mathbf{P_1 Y}$ and $\mathbf{P_2 Y}$ are independent random vectors.
- Under H_0 the squared length of each projection has a (scaled) χ^2-distribution.
- Under H_1 the distribution of $\mathbf{P_1 Y}$ should be shifted to higher values.

What We Have

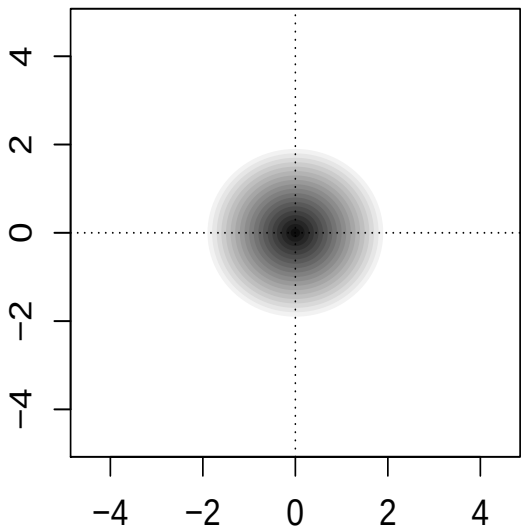
We already know that if projecting $\mathbf{Y}$ onto a particular subspace is equivalent to projecting $\boldsymbol{\epsilon}$ onto that subspace then the squared length of the image has a (scaled) χ^2-distribution.

- Our estimate of $\hat{\sigma}^2$ is based on the orthogonal projection of $\mathbf{y}$ onto the error space and the fact that the squared length of this projection has a (scaled) χ^2-distribution.
- Projecting $\mathbf{y}$ onto the error space is equivalent to projecting $\boldsymbol{\epsilon}$ onto the error space because $\boldsymbol{\mu}_{\mathbf{Y}}$ is orthogonal to the error space.

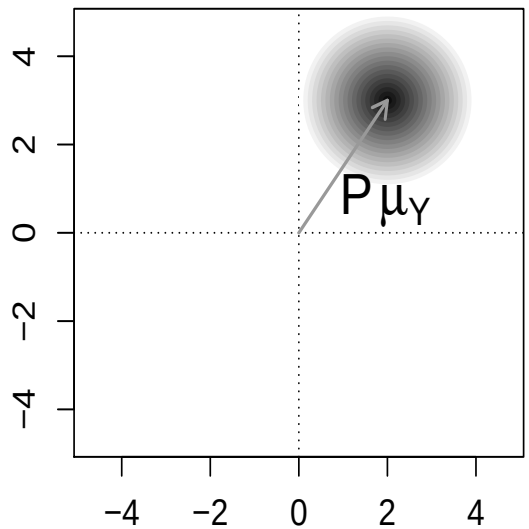
What If

What if we project $\mathbf{Y}$ onto a subspace that is not orthogonal to $\boldsymbol{\mu}_{\mathbf{Y}}$?

$\mathbf{PY} = \mathbf{P}\boldsymbol{\epsilon}$



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



Conceptually it is easy to see that we expect the squared length of the projection should be larger.

What We Want

Therefore, to construct an F-test:

- For the numerator we want to project $\mathbf{y}$ onto a subspace such that this is equivalent to projecting ϵ if the null hypothesis is true but not if the alternative hypothesis is true.
- For the denominator, ideally we want to project $\mathbf{y}$ onto a subspace such that this is equivalent to projecting ϵ under both the null and the alternative hypotheses.
- These two projections must result in independent random vectors.

We now need to consider under what conditions two projections of $\mathbf{Y}$ will result in independent random vectors.

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 $\boldsymbol{\Sigma}_{\mathbf{UV}} = \mathbf{0}$.

Recollections of Orthogonal Vector Spaces

Recall the following about orthogonal subspaces:

- Subspaces S_1 and S_2 are orthogonal if every vector in S_1 is orthogonal to every vector in S_2.
- We can show subspaces S_1 and S_2 are orthogonal ($S_1 \perp S_2$) by showing that $\mathbf{P_1 P_2} = \mathbf{0}$ where $\mathbf{P_1}$ and $\mathbf{P_2}$ are the projection matrices for S_1 and S_2.

Useful Properties of Projection Matrices

All orthogonal projection matrices possess two properties that are often very useful in mathematical derivations.

Any projection matrix $\mathbf{P}$ is:

1. Idempotent: $\mathbf{PP} = \mathbf{P}$.
2. Symmetric: $\mathbf{P}^t = \mathbf{P}$.

A Foundational Result

Our construction of F-tests, will depend explicitly on the following result:

Let $\mathbf{V}$ be a Normal random vector with covariance matrix $\mathbf{\Sigma_V} = \sigma^2\mathbf{I}$ and let S_1 and S_2 be two orthogonal subspaces ($S_1 \perp S_2$).

If we project $\mathbf{V}$ onto S_1 and onto S_2, the resulting random vectors, $\mathbf{P_1V}$ and $\mathbf{P_2V}$, are independent Normal random vectors.

Derivation of our Foundational Result

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

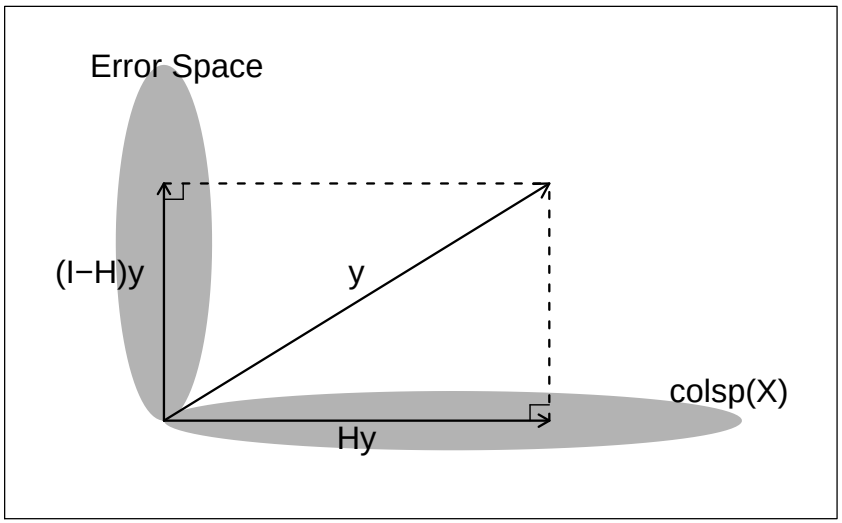
$$\text{Thus:} \quad \mathbf{\Sigma_W} = \begin{bmatrix} \mathbf{P_1} \\ \mathbf{P_2} \end{bmatrix} \sigma^2 \mathbf{I} \begin{bmatrix} \mathbf{P_1^t} & \mathbf{P_2^t} \end{bmatrix} = \sigma^2 \begin{bmatrix} \mathbf{P_1 P_1^t} & \mathbf{P_1 P_2^t} \\ \mathbf{P_2 P_1^t} & \mathbf{P_2 P_2^t} \end{bmatrix}$$

$$\text{Simplifying:} \quad \mathbf{\Sigma_W} = \sigma^2 \begin{bmatrix} \mathbf{P_1 P_1} & \mathbf{P_1 P_2} \\ \mathbf{P_2 P_1} & \mathbf{P_2 P_2} \end{bmatrix} = \sigma^2 \begin{bmatrix} \mathbf{P_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{P_2} \end{bmatrix}$$

Thus $\mathbf{P_1 V}$ and $\mathbf{P_2 V}$ are independent Normal random vectors.

An F-Test

We can construct an F-test based on:



$$\hat{\mu}_{\mathbf{Y}} = \mathbf{HY}$$

$$\mathbf{r} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

- $\mathbf{H}$ is the projection matrix for the colsp($\mathbf{X}$) and $(\mathbf{I} - \mathbf{H})$ is the projection matrix onto the orthogonal compliment of colsp($\mathbf{X}$). As a result $\hat{\mu}_{\mathbf{Y}} = \mathbf{HY}$ and $\mathbf{r} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$ are independent random vectors.

An F-Test

Previously we argued that

$$\|\mathbf{r}\|^2/\sigma^2 \sim \chi^2_{n-k-1} \quad \longrightarrow \quad \|\mathbf{r}\|^2 \sim \sigma^2 \times \chi^2_{n-k-1}$$

based on $\mathbf{r}$ being the same as the orthogonal projection of $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0},\ \sigma^2\mathbf{I})$ onto a subspace of dimension $n-k-1$.

$\hat{\boldsymbol{\mu}}_{\mathbf{Y}}$ is a projection of $\mathbf{Y} = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\epsilon}$ onto a subspace of dimension $k+1$. If $\boldsymbol{\mu}_{\mathbf{Y}} = \mathbf{0}$ then this equates to a projection of $\boldsymbol{\epsilon}$. In this case:

$$\|\hat{\boldsymbol{\mu}}_{\mathbf{Y}}\|^2/\sigma^2 \sim \sigma^2 \times \chi^2_{k+1}$$

An F-Test

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

$$\text{If } \boldsymbol{\mu}_{\mathbf{Y}} = \mathbf{0}: \quad \frac{\|\hat{\boldsymbol{\mu}}_{\mathbf{Y}}\|^2/(k+1)}{\|\mathbf{r}\|^2/(n-k-1)} \sim \mathsf{F}_{k+1,\, n-k-1}$$

If $\boldsymbol{\mu}_{\mathbf{Y}} \neq \mathbf{0}$: then we expect $\|\hat{\boldsymbol{\mu}}_{\mathbf{Y}}\|^2$ to be larger than if $\boldsymbol{\mu}_{\mathbf{Y}} = \mathbf{0}$.

An F-Test

So we can test H_0: $\boldsymbol{\mu}_{\mathbf{Y}} = \mathbf{0}$ versus H_1: $\boldsymbol{\mu}_{\mathbf{Y}} \neq \mathbf{0}$, using:

$$\text{F-stat} = \frac{\|\hat{\boldsymbol{\mu}}_{\mathbf{Y}}\|^2/(k+1)}{\|\mathbf{r}\|^2/(n-k-1)}$$

$$\text{p-value} = \Pr(F_{k+1,\, n-k-1} \geq \text{F-stat})$$

- Note that this is equivalent to testing H_0: $\boldsymbol{\beta} = \mathbf{0}$ versus H_1: $\boldsymbol{\beta} \neq \mathbf{0}$.

Catheter Example

So for the catheter example

```
> num<- (t(H%*%y)%*%(H%*%y))/3
> denom<-(t(res)%*%res)/(12-2-1)
> fstat<-num/denom
> fstat
         [,1]
[1,] 380.6896
> pval<-1-pf(fstat,3,9)
> pval
              [,1]
[1,] 8.521609e-10
```

Catheter Example

Thus we have extremely strong evidence against H_o: $\boldsymbol{\mu_Y} = \mathbf{0}$.
Since $\boldsymbol{\mu_Y} = \mathbf{X}\boldsymbol{\beta}$ this is equivalent to testing H_o: $\boldsymbol{\beta} = \mathbf{0}$.

BUT

This test is pointless ... we know that catheter length cannot possibly be zero. What we really want to test is whether the two explanatory variables help predict the response. So we want to test:

$$H_o: \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} c \\ 0 \\ 0 \end{bmatrix}$$

where c is **not** specified.

The Added Variable F-Test

This brings us to the added variable F-test. In our case we would like to test whether the model that includes the two explanatory variables does a better job of predicting the response than the null (intercept only) model.
In terms of geometry we are testing:

$$H_o: \boldsymbol{\mu}_{\mathbf{Y}} \in \text{colsp}(\mathbf{1}) \quad \text{vs} \quad H_1: \boldsymbol{\mu}_{\mathbf{Y}} \in \text{colsp}(\mathbf{X})$$

where $\text{colsp}(\mathbf{1}) \subset \text{colsp}(\mathbf{X})$.

- The procedure would be the same for any situation where the vector space in H_o is a subspace of the one in H_1

The Projection Matrix onto colsp($\mathbf{1}$)

To do this we need to divide $\hat{\mu}_{\mathbf{Y}}$ into two orthogonal components: the projection of $\hat{\mu}_{\mathbf{Y}}$ onto the colsp($\mathbf{1}$) and the projection onto the orthogonal compliment of colsp($\mathbf{1}$) in colsp($\mathbf{X}$).

Let $\mathbf{P_0}$ be the projection matrix onto colsp($\mathbf{1}$):

$$\mathbf{P_0} = \mathbf{1}\left(\mathbf{1}^t\mathbf{1}\right)^{-1}\mathbf{1}^t = \frac{1}{n}\mathbf{1}\mathbf{1}^t = \begin{bmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & & \vdots \\ \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix}$$

- This matrix produces a vector where each entry is the mean of the entries for the vector on which it operates.

The Projection of $\hat{\mu}_{\mathbf{Y}}$ onto colsp($\mathbf{1}$)

Projecting $\hat{\mu}_{\mathbf{Y}}$ onto colsp($\mathbf{1}$) is the same as projecting $\mathbf{y}$ onto colsp($\mathbf{1}$):

$$\mathbf{P_0 y} = \mathbf{P_0}\left(\hat{\mu}_{\mathbf{Y}} + \mathbf{r}\right) = \mathbf{P_0}\hat{\mu}_{\mathbf{Y}} + \mathbf{P_0 r} = \mathbf{P_0}\hat{\mu}_{\mathbf{Y}}.$$

- $\mathbf{P_0 r} = 0$ since $\mathbf{r} \perp$ colsp($\mathbf{X}$) and therefore $\mathbf{r}$ is perpendicular to every vector in colsp($\mathbf{X}$) which includes $\mathbf{1}$.

Decomposing $\mathbf{y}$

Previously we decomposed $\mathbf{y}$ into $\hat{\mu}_{\mathbf{Y}}$ and $\mathbf{r}$ where $\hat{\mu}_{\mathbf{Y}} \perp \mathbf{r}$:

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

Now we further decompose $\hat{\mu}_{\mathbf{Y}} = \mathbf{H}\mathbf{y}$ into the projections onto colsp($\mathbf{1}$) and onto its orthogonal compliment in colsp($\mathbf{X}$):

$$\mathbf{y} = \mathbf{H}\mathbf{y} + (\mathbf{I} - \mathbf{H})\,\mathbf{y} = \mathbf{P_0}\mathbf{y} + (\mathbf{H} - \mathbf{P_0})\,\mathbf{y} + (\mathbf{I} - \mathbf{H})\,\mathbf{y}$$

Components of $\mathbf{y}$

Suppose the true model has the form: $\mathbf{Y} = \mathbf{1}\beta_0 + \boldsymbol{\epsilon}$. Consider each of the components:

$$\begin{aligned}
\mathbf{P_0 Y} &= \mathbf{P_0}\left(\mathbf{1}\beta_0 + \boldsymbol{\epsilon}\right) = \mathbf{P_0}(\mathbf{1}\beta_0) + \mathbf{P_0}\boldsymbol{\epsilon} \\
&= \mathbf{1}\beta_0 + \boldsymbol{\epsilon}
\end{aligned}$$

$$\begin{aligned}
\left(\mathbf{H} - \mathbf{P_0}\right)\mathbf{Y} &= \left(\mathbf{H} - \mathbf{P_0}\right)\left(\mathbf{1}\beta_0 + \boldsymbol{\epsilon}\right) = \left(\mathbf{H} - \mathbf{P_0}\right)\mathbf{1}\beta_0 + \left(\mathbf{H} - \mathbf{P_0}\right)\boldsymbol{\epsilon} \\
&= \left(\mathbf{1}\beta_0 - \mathbf{1}\beta_0\right) + \left(\mathbf{H} - \mathbf{P_0}\right)\boldsymbol{\epsilon} \\
&= \left(\mathbf{H} - \mathbf{P_0}\right)\boldsymbol{\epsilon}
\end{aligned}$$

$$\begin{aligned}
\left(\mathbf{I} - \mathbf{H}\right)\mathbf{Y} &= \left(\mathbf{I} - \mathbf{H}\right)\left(\mathbf{1}\beta_0 + \boldsymbol{\epsilon}\right) = \left(\mathbf{I} - \mathbf{H}\right)\mathbf{1}\beta_0 + \left(\mathbf{I} - \mathbf{H}\right)\boldsymbol{\epsilon} \\
&= \left(\mathbf{1}\beta_0 - \mathbf{1}\beta_0\right) + \left(\mathbf{I} - \mathbf{H}\right)\boldsymbol{\epsilon} \\
&= \left(\mathbf{I} - \mathbf{H}\right)\boldsymbol{\epsilon}
\end{aligned}$$

The Ingredients of an F-Test

If the true model is $\mathbf{Y} = \mathbf{1}\beta_0 + \boldsymbol{\epsilon}$, each of $(\mathbf{H} - \mathbf{P_0})\,\mathbf{y}$ and $(\mathbf{I} - \mathbf{H})\,\mathbf{y}$ are equivalent to projecting $\boldsymbol{\epsilon}$ onto a vector space.

- Therefore the squared length of each of these projections divided by σ^2 has a χ^2-distribution.
- $(\mathbf{H} - \mathbf{P_0})$ projects onto a vector space of dimension k and $(\mathbf{I} - \mathbf{H})$ projects onto a vector space of dimension $n - k - 1$.
- The projections are independent as the vector spaces are orthogonal.

If the true model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ (where $\mathbf{X}$ contains columns for the two explanatory variables as well as the intercept) then:

- $(\mathbf{H} - \mathbf{P_0})\,\mathbf{y}$ is no longer equivalent to projecting $\boldsymbol{\epsilon}$ onto a vector space since $\mathbf{H}\,(\mathbf{X}\boldsymbol{\beta})$ will not equal $\mathbf{P_0}\,(\mathbf{X}\boldsymbol{\beta})$.

Our F-Test

We can now construct an F-test for

$$\mathrm{H}_0\colon\ \boldsymbol{\mu}_{\mathbf{Y}} = \mathbf{1}\beta_0 \qquad \text{versus} \qquad \mathrm{H}_1\colon\ \boldsymbol{\mu}_{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta}.$$

$$\begin{aligned} \text{F-stat} &= \frac{\|\left(\mathbf{H} - \mathbf{P_0}\right)\mathbf{Y}\|^2/(k)}{\|\left(\mathbf{I} - \mathbf{H}\right)\mathbf{Y}\|^2/(n-k-1)} \\ \text{p-value} &= \Pr(\mathrm{F}_{k,\,n-k-1} \geq \text{F-stat}) \end{aligned}$$

- This tests whether the regression model is better for predicting the response than simply using the mean.

Catheter Example

So for the catheter data:

```
> t(y)%*%(H-Po)%*%y
          [,1]
[1,] 607.1878
> num<-(t(y)%*%(H-Po)%*%y)/2
> fstat<-num/denom
> fstat
          [,1]
[1,] 21.26688
> pval<-1-pf(fstat,2,9)
> pval
               [,1]
[1,] 0.0003887566
```

Conclusion

Thus there is strong evidence (p-value $= .00039$) that the regression model is better for predicting the response than using the mean response.

- Note that this test is included in the *R* output (see slide 3) from `summary(catheter.lm)`:

```
F-statistic: 21.27 on 2 and 9 DF,  p-value: 0.0003888
```

Wrap Up

We have developed the machinery for the "added variables F-Test".

- In our application we tested the null model (intercept only) versus the full model (all of the regressors).
- We can easily adapt the procedure to test any two models as long as one of the models is a submodel of the other.