Multiple Regression

Dan Frey
Associate Professor of Mechanical Engineering and Engineering Systems
Plan for Today

• Multiple Regression
  – Estimation of the parameters
  – Hypothesis testing
  – Regression diagnostics
  – Testing lack of fit

• Case study

• Next steps
The Model Equation

For a single variable

\[ Y = \alpha + \beta x + \varepsilon \]

For multiple variables

\[ y = X\beta + \varepsilon \]

\[ y_1 \]
\[ y_2 \]
\[ \vdots \]
\[ y_n \]

\[ X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \]

\[ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \]

\[ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \]

These 1's allow \( \beta_0 \) to enter the equation without being mult by \( x \)'s.

\( \alpha \) is renamed \( \beta_0 \)

\( p = k + 1 \)
The Model Equation \( y = X\beta + \varepsilon \)

Each row of \( X \) is paired with an observation

\[
\begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{bmatrix}
= \begin{bmatrix}
1 & x_{11} & x_{12} & \cdots & x_{1k} \\
1 & x_{21} & x_{22} & \cdots & x_{2k} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & x_{n1} & x_{n2} & \cdots & x_{nk}
\end{bmatrix}
\begin{bmatrix}
\beta_0 \\
\beta_1 \\
\vdots \\
\beta_k
\end{bmatrix}
+ \begin{bmatrix}
\varepsilon_1 \\
\varepsilon_2 \\
\vdots \\
\varepsilon_n
\end{bmatrix}
\]

There are \( n \) observations of the response

Each column of \( X \) is paired with a coefficient

There are \( k \) coefficients

Each observation is affected by an independent homoscedastic normal variates

\( E(\varepsilon_i) = 0 \)

\( Var(\varepsilon_i) = \sigma^2 \)
Accounting for Indices

\[ \mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon} \]

\[ \begin{align*}
y = & \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \\
X = & \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \\
\beta = & \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \\
\varepsilon = & \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \end{align*} \]
Concept Question

Which of these is a valid matrix?

\[
X = \begin{bmatrix}
1 & 5.0m & 0.3 \text{sec} \\
1 & 7.1m & 0.2 \text{sec} \\
1 & 3.2m & 0.7 \text{sec} \\
1 & 5.4m & 0.4 \text{sec}
\end{bmatrix} \quad X = \begin{bmatrix}
1 & 5.0m & 0.3m \\
1 & 7.1V & 0.2V \\
1 & 3.2 \text{sec} & 0.7 \text{sec} \\
1 & 5.4A & 0.4A
\end{bmatrix} \quad X = \begin{bmatrix}
1 & 5.0m & 0.1 \text{sec} \\
1 & 7.1m & 0.3 \text{sec}
\end{bmatrix}
\]

1) A only  
2) B only  
3) C only  
4) A and B  
5) B and C  
6) A and C  
7) A, B, & C  
8) None  
9) I don’t know
Adding h.o.t. to the Model Equation

Each row of $X$ is paired with an observation

There are $n$ observations of the response

You can add interactions

You can add curvature

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{11}x_{12} & x_{11}^2 \\ 1 & x_{21} & x_{22} & x_{21}x_{22} & x_{21}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n1}x_{n2} & x_{n1}^2 \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_{12} \\ \beta_{11} \end{bmatrix}$$
Estimation of the Parameters $\beta$

Assume the model equation

$$y = X\beta + \varepsilon$$

We wish to minimize the sum squared error

$$L = \varepsilon^T \varepsilon = (y - X\beta)^T (y - X\beta)$$

To minimize, we take the derivative and set it equal to zero

$$\frac{\partial L}{\partial \beta} = -2X^Ty + 2X^TX\hat{\beta}$$

The solution is

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$

And we define the fitted model

$$\hat{y} = X\hat{\beta}$$
MathCad Demo
Montgomery Example 10-1

Breakdown of Sum Squares

"Grand Total Sum of Squares"

\[ \text{GTSS} = \sum_{i=1}^{n} y_i^2 \]

\[ SS_{\text{due to mean}} = ny \bar{y}^2 \]

\[ SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 \]

\[ SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \]

\[ SS_E = \sum_{i=1}^{n} e_i^2 \]

\[ SS_{PE} \]

\[ SS_{LOF} \]
Breakdown of DOF

\[ n \]
number of \( y \) values

\[ n - 1 \]
total sum of squares

\[ k \]
for the regression

\[ n - k - 1 \]
for error

1
due to the mean
Estimation of the Error Variance $\sigma^2$

Remember the model equation $y = X\beta + \epsilon$

If assumptions of the model equation hold, then $E(SS_E/(n-k-1)) = \sigma^2$

So an unbiased estimate of $\sigma^2$ is $\hat{\sigma}^2 = SS_E/(n-k-1)$
a.k.a. “coefficient of multiple determination”

$R^2$ and Adjusted $R^2$

What fraction of the total sum of squares ($SS_T$) is accounted for jointly by all the parameters in the fitted model?

$$R^2 \equiv \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} \quad R^2 \text{ can only rise as parameters are added}$$

$$R^2_{adj} \equiv 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} = 1 - \left( \frac{n-1}{n-p} \right)(1 - R^2) \quad R^2_{adj} \text{ can rise or drop as parameters are added}$$
Why Hypothesis Testing is Important in Multiple Regression

- Say there are 10 regressor variables
- Then there are 11 coefficients in a linear model
- To make a fully 2nd order model requires
  - 10 curvature terms in each variable
  - 10 choose 2 = 45 interactions
- You’d need 68 samples just to get the matrix $X^TX$ to be invertible
- You need a way to discard insignificant terms
Test for Significance of Regression

The hypotheses are

\[ H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \]

\[ H_1 : \beta_j \neq 0 \text{ for at least one } j \]

The test statistic is

\[ F_0 = \frac{SS_R/k}{SS_E/(n-k-1)} \]

Reject \( H_0 \) if \( F_0 > F_{\alpha,k,n-k-1} \)
Test for Significance
Individual Coefficients

The hypotheses are

\[ H_0 : \beta_j = 0 \]

\[ H_1 : \beta_j \neq 0 \]

The test statistic is

\[ t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \]

Reject \( H_0 \) if \( |t_0| > t_{\alpha/2, n-k-1} \)

\[ C = (X^T X)^{-1} \]

Standard error

\[ \sqrt{\hat{\sigma}^2 C_{jj}} \]
Test for Significance of Groups of Coefficients

Partition the coefficients into two groups \( \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \)

to be removed
to remain

Reduced model \( y = X_2 \beta_2 + \epsilon \)

\[
X = \begin{bmatrix}
1 & x_{11} & x_{12} & \cdots & x_{1k} \\
1 & x_{21} & x_{22} & \cdots & x_{2k} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & x_{n1} & x_{n2} & \cdots & x_{nk}
\end{bmatrix}
\]

\[ \rightarrow \]

Basically, you form \( X_2 \) by removing the columns associated with the coefficients you are testing for significance.

\[
H_0 : \beta_1 = 0 \\
H_1 : \beta_1 \neq 0
\]
Test for Significance
Groups of Coefficients

Reduced model \( y = X_2 \beta_2 + \varepsilon \)

The regression sum of squares for the reduced model is

\[
SS_R(\beta_2) = y^T H_2 y - n\bar{y}^2
\]

Define the sum squares of the removed set given the other coefficients are in the model

\[
SS_R(\beta_1 | \beta_2) \equiv SS_R(\beta) - SS_R(\beta_2)
\]

The partial \( F \) test

\[
F_0 = \frac{SS_R(\beta_1 | \beta_2)/r}{SS_E/(n-p)}
\]

Reject \( H_0 \) if \( F_0 > F_{\alpha, r, n-p} \)
Excel Demo -- Montgomery Ex10-2

Exhaustive search of the space of discrete 2-level factors is the full factorial $2^3$ experimental design.
Adding Center Points

Center points allow an experimenter to check for curvature and, if replicated, allow for an estimate of pure experimental error.
Plan for Today

• Mud cards
• Multiple Regression
  – Estimation of the parameters
  – Hypothesis testing
  Regression diagnostics
  – Testing lack of fit
• Case study
• Next steps
The “Hat” Matrix

Since
\[ \hat{\beta} = (X^TX)^{-1}X^T y \]

and
\[ \hat{y} = X\hat{\beta} \]

therefore
\[ \hat{y} = X(X^TX)^{-1}X^T y \]

So we define
\[ H \equiv X(X^TX)^{-1}X^T \]

Which maps from observations \( y \) to predictions \( \hat{y} \)
\[ \hat{y} = Hy \]
Influence Diagnostics

• The relative disposition of points in $x$ space determines their effect on the coefficients
• The hat matrix $\mathbf{H}$ gives us an ability to check for leverage points
• $h_{ij}$ is the amount of leverage exerted by point $y_j$ on $\hat{y}_i$
• Usually the diagonal elements $\sim p/n$ and it is good to check whether the diagonal elements within 2X of that
MathCad Demo on Distribution of Samples and Its Effect on Regression
Standardized Residuals

The residuals are defined as

\[ e = y - \hat{y} \]

So an unbiased estimate of \( \sigma^2 \) is

\[ \hat{\sigma}^2 = \frac{SS_E}{(n - p)} \]

The standardized residuals are defined as

\[ d = \frac{e}{\hat{\sigma}} \]

If these elements were \( z \)-scores then with probability 99.7%

\[ -3 < d_i < 3 \]
Studentized Residuals

The residuals are defined as

\[ e = y - \hat{y} \]

to therefore

\[ e = y - \mathbf{H}y = (\mathbf{I} - \mathbf{H})y \]

So the covariance matrix of the residuals is

\[ \text{Cov}(e) = \sigma^2 \text{Cov}(\mathbf{I} - \mathbf{H}) \]

The studentized residuals are defined as

\[ r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2 (1 - h_{ii})}} \]

If these elements were z-scores then with probability 99.7%

\[ -3 < r_i < 3 \]
Testing for Lack of Fit
(Assuming a Central Composite Design)

• Compute the standard deviation of the center points and assume that represents the $MS_{PE}$

$$MS_{PE} = \frac{\sum (y_i - \bar{y})}{n_C - 1}$$

$$SS_{PE} = (n - 1)MS_{PE}$$

$$SS_{PE} + SS_{LOF} = SS_E$$

$$MS_{LOF} = \frac{SS_{LOF}}{p}$$

$$F_0 = \frac{MS_{LOF}}{MS_{PE}}$$
You perform a linear regression of 100 data points \( n=100 \). There are two independent variables \( x_1 \) and \( x_2 \). The regression \( R^2 \) is 0.72. Both \( \beta_1 \) and \( \beta_2 \) pass a \( t \) test for significance. You decide to add the interaction \( x_1 x_2 \) to the model. Select all the things that cannot happen:

1) Absolute value of \( \beta_1 \) decreases
2) \( \beta_1 \) changes sign
3) \( R^2 \) decreases
4) \( \beta_1 \) fails the \( t \) test for significance
Plan for Today

• Mud cards
• Multiple Regression
  – Estimation of the parameters
  – Hypothesis testing
  – Regression diagnostics
  – Testing lack of fit

Case study

• Next steps
Scenario

- The FAA and EPA are interested in reducing CO2 emissions
- Some parameters of airline operations are thought to effect CO2 (e.g., Speed, Altitude, Temperature, Weight)
- Imagine flights have been made with special equipment that allowed CO2 emission to be measured (data provided)
- You will report to the FAA and EPA on your analysis of the data and make some recommendations
Phase One

- Open a Matlab window
- Load the data (load FAAcase3.mat)
- Explore the data
Phase Two

- Do the regression
- Examine the betas and their intervals
- Plot the residuals

\[
y = \frac{\text{CO2.}}{\text{ground.speed}};
\]

\[
\text{ones}(1:3538) = 1;
\]

\[
X = [\text{ones'} \ TAS \ alt \ temp \ weight];
\]

\[
[b, bint, r, rint, stats] = \text{regress}(y, X, 0.05);
\]

\[
yhat = X * b;
\]

\[
\text{plot}(yhat, r, '+')
\]
This code will remove the points at which the aircraft is climbing or descending.
Try The Regression Again on Cruise Only Portions

• What were the effects on the residuals?
• What were the effects on the betas?

```matlab
hold off
[b,bint,r,rint,stats] = regress(y,X,0.05);
yhat=X*b;
plot(yhat,r,'+')
```
See What Happens if We Remove Variables

- Remove weight & temp
- Do the regression (CO2 vs TAS & alt)
- Examine the betas and their intervals

```
[b,bint,r,rint,stats] = regress(y,X(:,1:3),0.05);
```
Phase Three

- Try different data (flight34.mat)
- Do the regression
- Examine the betas and their intervals
- Plot the residuals

```matlab
y=[fuel_burn];
one(1:34)=1;
X=[ones' TAS alt temp];
[b,bint,r,rint,stats] = regress(y,X,0.05);
yhat=X*b;
plot(yhat,r,'+')
```
Adding Interactions

This line will add a interaction

\[ X(:,5) = X(:,2) \times X(:,3); \]

What’s the effect on the regression?
Case Wrap-Up

• What were the recommendations?
• What other analysis might be done?
• What were the key lessons?
Next Steps

• Wednesday 25 April
  – Design of Experiments
  – Please read "Statistics as a Catalyst to Learning"

• Friday 27 April
  – Recitation to support the term project

• Monday 30 April
  – Design of Experiments

• Wednesday 2 May
  – Design of Computer Experiments

• Friday 4 May?? Exam review??

• Monday 7 May – Frey at NSF

• Wednesday 9 May – Exam #2