Multiple Linear Regression and the General Linear Model
Outline

1. Introduction to Multiple Linear Regression
2. Statistical Inference
3. Topics in Regression Modeling
4. Example
5. Variable Selection Methods
6. Regression Diagnostic and Strategy for Building a Model
1. Introduction to Multiple Linear Regression

Extending simple linear regression to two or more regressors
History

- Francis Galton coined the term regression in his biological research

- Karl Pearson and Udny Yule extended Galton’s work to the statistical context

- Legendre and Gauss developed the method of least squares

- Ronald Fisher developed the maximum likelihood method used in the related statistical inference.
History

Francis Galton (1822/2/16 – 1911/1/17).
He is regarded as the founder of Biostatistics.
In his research he found that tall parents usually have shorter children; and vice versa. So the height of human being has the tendency to regress to its mean.
Subsequently, he coined the word “regression” for such phenomenon and problems.

Adrien-Marie Legendre (1752/9/18 – 1833/1/10).
In 1805, he published an article named *Nouvelles méthodes pour la détermination des orbites des comètes*.
In this article, he introduced Method of Least Squares to the world. Yes, he was the first person who published article regarding to method of least squares, which is the earliest form of regression.

He developed the fundamentals of the basis for least-squares analysis in 1795 at the age of eighteen. He published an article called *Theoria Motus Corporum Celestium in Sectionibus Conicis Solem Ambientum* in 1809. In 1821, he published another article about least square analysis with further development, called *Theoria combinationis observationum erroribus minimis obnoxiae*. This article includes Gauss–Markov theorem.

Karl Pearson.
Ronald Aylmer Fisher.
They both developed regression theory after Galton.

Most content in this page comes from Wikipedia.
Probabilistic Model

$Y_i$ is the observed value of the random variable (r.v.) which depends on $k$ fixed predictor values according to the following model:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

Here $\beta_0, \beta_1, \ldots, \beta_k$ are unknown model parameters, and $n$ is the number of observations.

The random error, $\varepsilon_i$, are assumed to be independent r.v.’s with mean 0 and variance $\sigma^2$.

Thus $Y_i$ are independent r.v.’s with mean $\mu_i$ and variance $\sigma^2$, where

$$\mu_i = E(Y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}$$
Fitting the model

- The least squares (LS) method is used to find a line that fits the equation

\[ Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \epsilon_i, \quad i = 1, \ldots, n \]

- Specifically, LS provides estimates of the unknown model parameters, \( \beta_0, \beta_1, \ldots, \beta_k \), which minimizes, \( \Delta \), the sum of squared difference of the observed values, \( y_i \), and the corresponding points on the line with the same \( x \)'s

\[
\Delta = \sum_{i=1}^{n} \left[ Y_i - \left( \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} \right) \right]^2
\]

- The LS can be found by taking partial derivatives of \( Q \) with respect to the unknown parameters \( \beta_0, \beta_1, \ldots, \beta_k \) and setting them equal to 0. The result is a set of simultaneous linear equations.

- The resulting solutions, \( \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k \), are the least squares (LS) estimators of \( \beta_0, \beta_1, \ldots, \beta_k \), respectively.

- Please note the LS method is non-parametric. That is, no probability distribution assumptions on \( Y \) or \( \epsilon \) are needed.
Goodness of Fit of the Model

• To evaluate the goodness of fit of the LS model, we use the residuals defined by
  \[ e_i = y_i - \hat{y}_i \quad (i = 1, 2, \ldots, n) \]
  • \( \hat{y}_i \) are the fitted values:
  \[
  \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}
  \]
  • An overall measure of the goodness of fit is the error sum of squares (SSE)
  \[
  \min Q = SSE = \sum_{i=1}^{n} \hat{e}_i^2
  \]
  • A few other definition similar to those in simple linear regression:
  - total sum of squares (SST):
  \[
  SST = \sum (y_i - \bar{y})^2
  \]
  - regression sum of squares (SSR):
  \[
  SSR = SST - SSE
  \]
• coefficient of determination:

\[ R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \]

• \( 0 \leq R^2 \leq 1 \)
  • values closer to 1 represent better fits
  • adding predictor variables never decreases and generally increases \( R^2 \)

• multiple correlation coefficient (positive square root of \( R^2 \)):

\[ R = +\sqrt{R^2} \]

• only positive square root is used
  • \( R \) is a measure of the strength of the association between the predictors (x’s) and the one response variable Y
Multiple Regression Model in Matrix Notation

* The multiple regression model can be represented in a compact form using matrix notation

Let:

\[
Y = \begin{bmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_n
\end{bmatrix}, \quad y = \begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{bmatrix}, \quad \varepsilon = \begin{bmatrix}
\varepsilon_1 \\
\varepsilon_2 \\
\vdots \\
\varepsilon_n
\end{bmatrix}
\]

be the \( n \times 1 \) vectors of the r.v.'s \( Y_i 's \), their observed values \( y_i 's \), and random errors \( \varepsilon_i 's \), respectively for all \( n \) observations

Let:

\[
X = \begin{bmatrix}
1 & x_{11} & \cdots & x_{k1} \\
1 & x_{12} & \cdots & x_{k2} \\
\vdots & \ddots & \ddots & \vdots \\
1 & x_{1n} & \cdots & x_{kn}
\end{bmatrix}
\]

be the \( n \times (k + 1) \) matrix of the values of the predictor variables for all \( n \) observations

(\text{the first column corresponds to the constant term } \beta_0 )
Let:

\[ \mathbf{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{and} \quad \hat{\mathbf{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} \]

be the \((k + 1) \times 1\) vectors of unknown model parameters and their LS estimates, respectively.

- The model can be rewritten as:
  \[ Y = X \mathbf{\beta} + \mathbf{\varepsilon} \]
- The simultaneous linear equations whose solutions yields the LS estimates:
  \[ X'X\mathbf{\beta} = X'y \]
- If the inverse of the matrix \(X'X\) exists, then the solution is given by:
  \[ \hat{\mathbf{\beta}} = (X'X)^{-1}X'y \]
2. Statistical Inference
Determining the statistical significance of the predictor variables:

- For statistical inferences, we need the assumption that $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ *i.i.d. means independent & identically distributed

- We test the hypotheses: $H_{0j} : \beta_j = 0$ vs. $H_{1j} : \beta_j \neq 0$

- If we can’t reject $H_{0j} : B_j = 0$, then the corresponding regressor $x_j$ is not a significant predictor of $y$.

- It is easily shown that each $\hat{\beta}_j$ is normal with mean $\beta_j$ and variance $\sigma^2 v_{jj}$, where $v_{jj}$ is the jth diagonal entry of the matrix $V = (X'X)^{-1}$
Deriving a **pivotal quantity** for the inference on \( \beta_j \)

- Recall \( \hat{\beta}_j \sim N(\beta_j, \sigma^2 v_{jj}) \)
- The unbiased estimator of the unknown error variance \( \sigma^2 \) is given by
  \[
  S^2 = \frac{SSE}{n-(k+1)} = \frac{\sum e_i^2}{n-(k+1)} = \frac{MSE}{d.o.f.}
  \]
- We also know that \( W = \frac{(n-(k+1))S^2}{\sigma^2} = \frac{SSE}{\sigma^2} \sim \chi^2_{n-(k+1)} \), and that \( S^2 \) and \( \hat{\beta}_j \) are statistically independent.
- With \( Z = \frac{\hat{\beta}_j - \beta_j}{\sigma \sqrt{v_{jj}}} \sim N(0,1) \), and by the definition of the t-distribution,
  we obtain the pivotal quantity for the inference on \( \beta_j \)

\[
T = \frac{Z}{\sqrt{W / n-(k+1)}} = \frac{\hat{\beta}_j - \beta_j}{S \sqrt{v_{jj}}} \sim T_{n-(k+1)}
\]
Derivation of the **Confidence Interval** for $\beta_j$

$$P(-t_{\alpha/2, n-(k+1)} \leq \hat{\beta}_j - \beta_j \leq t_{\alpha/2, n-(k+1)}) = 1 - \alpha$$

$$P(\hat{\beta}_j - t_{\alpha/2, n-(k+1)} s\sqrt{v_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2, n-(k+1)} s\sqrt{v_{jj}}) = 1 - \alpha$$

Thus the 100(1-$\alpha$)% confidence interval for $\beta_j$ is:

$$\hat{\beta}_j \pm SE(\hat{\beta}_j)$$

where $SE(\hat{\beta}_j) = s\sqrt{v_{jj}}$
Derivation of the Hypothesis Test for $\beta_j$

at the significance level $\alpha$

Hypotheses:

$H_0 : \beta_j = 0$

$H_a : \beta_j \neq 0$

The test statistic is:

$$T_0 = \frac{\hat{\beta}_j - 0}{S \sqrt{v_{jj}}} \sim T_{n-(k+1)}$$

The decision rule of the test is derived based on the Type I error rate $\alpha$. That is

$$P(\text{Reject } H_0 \mid H_0 \text{ is true}) = P(|T_0| \geq c) = \alpha$$

$$\Rightarrow c = t_{\frac{\alpha}{2}, n-(k+1)}$$

Therefore, we reject $H_0$ at the significance level $\alpha$ if and only if $|t_0| \geq t_{\frac{\alpha}{2}, n-(k+1)}$, where $t_0$ is the observed value of $T_0$. 
Another Hypothesis Test

Now consider:

\[ H_0 : \beta_j = 0 \quad \text{for all} \quad 1 \leq j \leq k \]
\[ H_a : \beta_j \neq 0 \quad \text{for at least one} \quad 1 \leq j \leq k \]

When \( H_0 \) is true, the test statistics

\[ F_0 = \frac{MSR}{MSE} \sim f_{k,n-(k+1)} \]

• An alternative and equivalent way to make a decision for a statistical test is through the p-value, defined as:

\[ p = P(\text{observe a test statistic value at least as extreme as the one observed} \mid H_0) \]

• At the significance level \( \alpha \), we reject \( H_0 \) if and only if \( p < \alpha \)
The General Hypothesis Test

- Consider the full model:

\[ Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i \quad (i=1,2,\ldots,n) \]

- Now consider a partial model:

\[ Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_{k-m} x_{k-m,i} + \varepsilon_i \quad (i=1,2,\ldots,n) \]

- Hypotheses: \( H_0 : \beta_{k-m+1} = \ldots = \beta_k = 0 \) vs. \( H_a : \beta_j \neq 0 \)

  for at least one \( k-m+1 \leq j \leq k \)

- Test statistic: \( F_0 = \frac{(SSE_{k-m} - SSE_k) / m}{SSE_k / [n-(k+1)]} \sim f_{m,n-(k+1)} \)

- Reject \( H_0 \) when \( F_0 > f_{m,n-(k+1),\alpha} \)
Estimating and Predicting Future Observations

• Let \( x^* = (x_0^*, x_1^*, \ldots, x_k^*)' \) and let

\[
\begin{align*}
    y^* &= \beta_0 + \beta_1 x_1^* + \cdots + \beta_k x_k^* + \varepsilon^* \\
    \mu^* &= \beta_0 + \beta_1 x_1^* + \cdots + \beta_k x_k^*
\end{align*}
\]

• The pivotal quantity for \( \mu^* \) is

\[
T = \frac{\hat{\mu}^* - \mu^*}{s\sqrt{x^* Vx^*}} \sim T_{n-(k+1)}
\]

• Using this pivotal quantity, we can derive a CI for the estimated mean \( \hat{\mu}^* \):

\[
\hat{\mu}^* \pm t_{n-(k+1), \alpha/2} s\sqrt{x^* Vx^*}
\]

• Additionally, we can derive a prediction interval (PI) to predict \( Y^* \):

\[
\hat{Y}^* \pm t_{n-(k+1), \alpha/2} s\sqrt{1 + x^* Vx^*}
\]
Regression in SAS

/* simple linear regression */
proc reg;
  model y = x;
/* multiple regression */
proc reg;
  model y = x1 x2 x3;

Here are some print options for the model phrase:
model y = x / noint;  /* regression with no intercept */
model y = x / ss1;    /* print type I sums of squares */
model y = x / p;      /* print predicted values and residuals */
model y = x / r;      /* option p plus residual diagnostics */
model y = x / clm;    /* option p plus 95% CI for estimated mean */
model y = x / cli;    /* option p plus 95% CI for predicted value */
model y = x / r cli clm; /* options can be combined */
Confidence and Prediction Band in SAS

The CLM option adds confidence limits for the mean predicted values. The CLI option adds confidence limits for the individual predicted values.

```
proc sgplot data=sashelp.class;
eg x=height y=weight / CLM CLI;
run;
```

For information about the SAS Sample Library, see:

http://support.sas.com/documentation/cdl/en/grstatproc/67909/HTML/default/viewer.htm#p0jxq3ea4njtvnn1xnj4a7s37pyz.htm
Regression in SAS

It is possible to let SAS do the predicting of new observations and/or estimating of mean responses. The way to do this is to enter the x values (or x1,x2,x3 for multiple regression) you are interested in during the data input step, but put a period (.) for the unknown y value.

```sas
data new;
  input x y;
  datalines;
  1 0
  2 3
  3 .
  4 3
  5 6
; 
run;
```

```sas
proc reg;
  model y = x / p cli clm;
```

3.

Topics in Regression Modeling
3.1 Multicollinearity

* **Definition.** The predictor variables are linearly dependent.

* This can cause serious numerical and statistical difficulties in fitting the regression model unless “extra” predictor variables are deleted.
How does the multicollinearity cause difficulties?

\[ \hat{\beta} \] is the solution to the equation \((X^T X)\beta = X^T y\), thus \((X^T X)\) must be invertable in order for \(\beta\) to be unique and computable.

If the approximate multicollinearity happens:

1. \(X^T X\) is nearly singular, which makes \(\hat{\beta}\) numerically unstable. This reflected in large changes in their magnitudes with small changes in data.

2. The matrix \(V = (X^T X)^{-1}\) has very large elements. Therefore \(Var(\hat{\beta}) = \sigma^2 v_{jj}\) are large, which makes \(\beta_j\) statistically nonsignificant.
Measures of Multicollinearity

1. The correlation matrix R. Easy but can’t reflect linear relationships between more than two variables.

2. Determinant of R can be used as measurement of singularity of $X^T X$

3. Variance Inflation Factors (VIF): the diagonal elements of $R^{-1}$. VIF > 10 is regarded as unacceptable.
3.2 Polynomial Regression

A special case of a linear model:

\[ y = \beta_0 + \beta_1 x + \ldots + \beta_k x^k + \varepsilon \]

Problems:

1. The powers of \( x \), i.e., \( x, x^2, \ldots, x^k \) tend to be highly correlated.
2. If \( k \) is large, the magnitudes of these powers tend to vary over a rather wide range.

So, set \( k \leq 3 \) if a good idea, and almost never use \( k > 5 \).
Solutions

1. Centering the $x$-variable:

\[ y = \beta_0^* + \beta_1^* (x - \bar{x}) + \ldots + \beta_k^* (x - \bar{x})^k + \epsilon \]

2. Effect: removing the non-essential multicollinearity in the data.

3. Further more, we can standardize the data by dividing the standard deviation $s_x$ of $x$: \[ \left( \frac{x - \bar{x}}{s_x} \right) \]

4. Effect: helping to alleviate the second problem.

5. Using the first few principal components of the original variables instead of the original variables.
3.3 Dummy Predictor Variables & The General Linear Model

How to handle the categorical predictor variables?

1. If we have categories of an ordinal variable, such as the prognosis of a patient (poor, average, good), one can assign numerical scores to the categories (poor=1, average=2, good=3).
2. If we have nominal variable with \( c \geq 2 \) categories. Use \( c-1 \) indicator variables, \( x_1, \ldots, x_{c-1} \), called **Dummy Variables**, to code.

\[
x_i = 1
\]

for the \( i \)th category, \( 1 \leq i \leq c-1 \)

\[
x_1 = \ldots = x_{c-1} = 0
\]

for the \( c \)th category.
Why don’t we just use $c$ indicator variables: $x_1, x_2, \ldots, x_c$?

If we use that, there will be a linear dependency among them:

$$x_1 + x_2 + \ldots + x_c = 1$$

This will cause multicollinearity.
Example of the dummy variables

* For instance, if we have four years of quarterly sale data of a certain brand of soda. How can we model the time trend by fitting a multiple regression equation?

Solution: We use quarter as a predictor variable $x_1$. To model the seasonal trend, we use indicator variables $x_2$, $x_3$, $x_4$, for Winter, Spring and Summer, respectively. For Fall, all three equal zero. That means: Winter-$(1,0,0)$, Spring-$(0,1,0)$, Summer-$(0,0,1)$, Fall-$(0,0,0)$. Then we have the model:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \varepsilon_i, i = 1, \ldots, 16$$
3. Once the dummy variables are included, the resulting regression model is referred to as a “General Linear Model”.

This term must be differentiated from that of the “Generalized Linear Model” which include the “General Linear Model” as a special case with the identity link function:

\[ \mu_i = E(Y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} \]

The generalized linear model will link the model parameters to the predictors through a link function. For another example, we recall the logit link in the logistic regression.
4. Example, Galton

- Here we revisit the classic regression towards Mediocrity in Hereditary Stature by Francis Galton
- He performed a simple regression to predict offspring height based on the average parent height
- [http://www.math.uah.edu/stat/data/Galton.html](http://www.math.uah.edu/stat/data/Galton.html)
- Slope of regression line was less than 1 showing that extremely tall parents had less extremely tall children
- At the time, Galton did not have multiple regression as a tool so he had to use other methods to account for the difference between male and female heights
- We can now perform multiple regression on parent-offspring height and use multiple variables as predictors
Example, Galton

**Our Model:**

\[ Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i \]

- \( Y = \) height of child
- \( x_1 = \) height of father
- \( x_2 = \) height of mother
- \( x_3 = \) gender of child
Example, Galton

In matrix notation:

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

\[ \hat{\beta} = (X'X)^{-1}X'y \]

We find that:
\[ \beta_0 = 15.34476 \]
\[ \beta_1 = 0.405978 \]
\[ \beta_2 = 0.321495 \]
\[ \beta_3 = 5.22595 \]
### Example, Galton

<table>
<thead>
<tr>
<th>Child</th>
</tr>
</thead>
<tbody>
<tr>
<td>73.2</td>
</tr>
<tr>
<td>69.2</td>
</tr>
<tr>
<td>69</td>
</tr>
<tr>
<td>69</td>
</tr>
<tr>
<td>73.5</td>
</tr>
<tr>
<td>72.5</td>
</tr>
<tr>
<td>65.5</td>
</tr>
<tr>
<td>65.5</td>
</tr>
<tr>
<td>65.5</td>
</tr>
<tr>
<td>71</td>
</tr>
<tr>
<td>68</td>
</tr>
</tbody>
</table>

\[
\mathbf{Y} = \ldots \ldots
\]
**Example, Galton**

\[ X = \begin{array}{c|ccc}
\text{Father} & \text{Mother} & \text{Gender} \\
1 & 78.5 & 67 & 1 \\
1 & 78.5 & 67 & 1 \\
1 & 78.5 & 67 & 0 \\
1 & 75.5 & 66.5 & 1 \\
1 & 75.5 & 66.5 & 1 \\
1 & 75.5 & 66.5 & 0 \\
1 & 75.5 & 66.5 & 0 \\
1 & 75 & 64 & 1 \\
1 & 75 & 64 & 0 \\
\cdots & \cdots & \cdots \\
\end{array} \]
Example, Galton

Important calculations

\[ SSE = \sum (y_i - \hat{y}_i)^2 = 4149.162 \]

\[ SST = \sum (y_i - \bar{y})^2 = 11,515.060 \]

\[ r^2 = 1 - \frac{SSE}{SST} = 0.6397 \]

\[ MSE = \frac{SSE}{n - (k + 1)} = 4.64 \]

\[ \hat{y}_i = X_i \hat{\beta} \]

Is the predicted height of each child given a set of predictor variables
Example, Galton

Are these values significantly different than zero?

- $H_0: \beta_j = 0$
- $H_a: \beta_j \neq 0$

Reject $H_0$ if

$$\left| t_j \right| = \left| \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \right| > t_{n-(k+1), \alpha/2} = t_{894,0.025} = 2.245$$

$$V = (X'X)^{-1} = \begin{bmatrix} 1.63 & -0.0118 & -0.0125 & -0.00596 \\ -0.0118 & 0.000184 & -0.0000143 & 0.0000223 \\ -0.0125 & -0.0000143 & 0.000211 & 0.0000327 \\ -0.00596 & 0.0000223 & 0.0000327 & 0.00447 \end{bmatrix}$$

$$SE(\hat{\beta}_j) = \sqrt{MSE \cdot v_{jj}}$$
Example, Galton

<table>
<thead>
<tr>
<th></th>
<th>β-estimate</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>15.3</td>
<td>2.75</td>
<td>5.59*</td>
</tr>
<tr>
<td>Father Height</td>
<td>0.406</td>
<td>0.0292</td>
<td>13.9*</td>
</tr>
<tr>
<td>Mother Height</td>
<td>0.321</td>
<td>0.0313</td>
<td>10.3*</td>
</tr>
<tr>
<td>Gender</td>
<td>5.23</td>
<td>0.144</td>
<td>36.3*</td>
</tr>
</tbody>
</table>

* p<0.05. We conclude that all β are significantly different than zero.
Example, Galton

Testing the model as a whole:

- $H_0: \beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$
- $H_a$: The above is not true.

Reject $H_0$ if

$$F = \frac{r^2[n-(k+1)]}{k(1-r^2)} > f_{k,n-(k+1),\alpha} = f_{3,894,.05} = 2.615$$

$$r^2 = 1 - \frac{SSE}{SST} = 0.6397$$

$$SSE = \sum (y_i - \hat{y}_i)^2 = 4149.162$$

$$SST = \sum (y_i - \bar{y})^2 = 11,515.060$$

Since $F = 529.032 > 2.615$, we reject $H_0$ and conclude that our model predicts height better than by chance.
Example, Galton

Making Predictions

Let’s say **George Clooney (71 inches)** and **Madonna (64 inches)** would have a baby boy.

$$\hat{Y}^* = 15.34 + 0.405978(71) + 0.321495(64) + 5.225951(1) = 69.97$$

95% Prediction interval:

$$\hat{Y}^* \pm 4.84 = (65.13, 74.81)$$
SAS code, data step:

```
http://www.math.uah.edu/stat/data/Galton.html

Data Galton;
Input Family Father Mother Gender $ Height Kids;
Datalines;
1 78.5 67 M 73.2 4
1 78.5 67 F 69.2 4
1 78.5 67 F 69 4
1 78.5 67 F 69 4
2 75.5 66.5 M 73.5 4
2 75.5 66.5 M 72.5 4
2 75.5 66.5 F 65.5 4
2 75.5 66.5 F 65.5 4
...
;
Run;
```
ods graphics on;
data revise;
set Galton;
if Gender = 'F' then sex = 1.0;
else sex = 0.0;
run;

proc reg data=revise;
title "proc reg; Dependence of Child Heights on Parental Heights";
model height = father mother sex / vif collin;
run;
quit;
### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>7365.90034</td>
<td>2455.30011</td>
<td>529.03</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>894</td>
<td>4149.16204</td>
<td>4.64112</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>897</td>
<td>11515</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

|                   |     |                |             |         |        |
| Root MSE          | 2.15433 |  R-Square | 0.6397 |
| Dependent Mean    | 66.76069 | Adj R-Sq | 0.6385 |
| Coeff Var         | 3.22694 |           |        |

### Parameter Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>Variance Inflation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>20.57071</td>
<td>2.74067</td>
<td>7.51</td>
<td>&lt;.0001</td>
<td>0</td>
</tr>
<tr>
<td>Father</td>
<td>1</td>
<td>0.40598</td>
<td>0.02921</td>
<td>13.90</td>
<td>&lt;.0001</td>
<td>1.00607</td>
</tr>
<tr>
<td>Mother</td>
<td>1</td>
<td>0.32150</td>
<td>0.03128</td>
<td>10.28</td>
<td>&lt;.0001</td>
<td>1.00660</td>
</tr>
<tr>
<td>sex</td>
<td>1</td>
<td>-5.22595</td>
<td>0.14401</td>
<td>-36.29</td>
<td>&lt;.0001</td>
<td>1.00188</td>
</tr>
</tbody>
</table>
EXAMPLE, GALTON

Residual by Regressors for Height

Father

Mother

sex
Alternatively, one can use proc GLM procedure that can incorporate the categorical variable (sex) directly via the class statement.

Another added benefit is that SAS will provide an overall significance test for the categorical variable.

```
proc glm data=Galton;
Class gender;
model height = father mother gender;
  run;
quit;
```
### EXAMPLE, Galton

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Father</td>
<td>1</td>
<td>896.716584</td>
<td>896.716584</td>
<td>193.21</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Mother</td>
<td>1</td>
<td>490.217369</td>
<td>490.217369</td>
<td>105.62</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Gender</td>
<td>1</td>
<td>6111.965365</td>
<td>6111.965365</td>
<td>1316.92</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

| Parameter | Estimate  | Standard Error  | t Value | Pr > |t| |
|-----------|-----------|-----------------|---------|------|---|
| Intercept | 20.57071133 | 2.74066703      | 7.51    | <.0001 |
| Father    | 0.40597803  | 0.02920696      | 13.90   | <.0001 |
| Mother    | 0.32149514  | 0.03128178      | 10.28   | <.0001 |
| Gender F  | -5.22595131 | 0.14400791      | -36.29  | <.0001 |
| Gender M  | 0.00000000  | 0.00000000      | 0       |      |
EXAMPLE, GALTON

- Dear Students, did you notice any violation of assumptions in the analysis of the Galton data?
- The answer is that we have violated the independent observations assumption as many kids were from the same families!
- This, however, can be easily resolved with more advanced statistical models that will include ‘Family’ as a random effect (random regressor.)
5. Variables Selection Method

A. Stepwise Regression
Variables selection method

(1) Why do we need to select the variables?

(2) How do we select variables?
   * stepwise regression
   * best subset regression
Stepwise Regression

* (p-1)-variable model:

\[ Y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_{p-1} x_{p-1,i} + \varepsilon_i \]

* P-variable model

\[ Y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_{p-1} x_{p-1,i} + \beta_p x_{p,i} + \varepsilon_i \]
Partial F - test

Hypothesis test

\[ H_{0p}: \beta_p = 0 \]

\[ H_{1p}: \beta_p \neq 0 \]

\[ F_p = \frac{(SSE_{p-1} - SSE_p)}{SSE_p / [n - (p + 1)]} > f_{1,n-(p+1),\alpha} \]

test statistic : \[ t_p = \frac{\beta_p}{SE(\beta_p)} \]

reject \( H_{0p} : |t_p| > t_{n-(p+1),\alpha/2} \)
Partial correlation coefficients

\[ r^2_{yx|\bar{x}_{1\ldots p-1}} = \frac{SSE_{p-1} - SSE_p}{SSE_p} = \frac{SSE(x_{1\ldots p-1}) - SSE(x_{1\ldots p})}{SSE(x_{1\ldots p-1})} \]

\[ F_p = t^2_{p} = \frac{r^2_{yx|\bar{x}_{1\ldots p-1}} [n - (p + 1)]}{1 - r^2_{yx|\bar{x}_{1\ldots p-1}}} \]
5. Variables selection method

A. Stepwise Regression:
   SAS Example
Example 11.5 (T&D pg. 416), 11.9 (T&D pg. 431)

The following table shows data on the heat evolved in calories during the hardening of cement on a per gram basis (y) along with the percentages of four ingredients: tricalcium aluminate (x1), tricalcium silicate (x2), tetracalcium alumino ferrite (x3), and dicalcium silicate (x4).

<table>
<thead>
<tr>
<th>No.</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>26</td>
<td>6</td>
<td>60</td>
<td>78.5</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>29</td>
<td>15</td>
<td>52</td>
<td>74.3</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>56</td>
<td>8</td>
<td>20</td>
<td>104.3</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>31</td>
<td>8</td>
<td>47</td>
<td>87.6</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>52</td>
<td>6</td>
<td>33</td>
<td>95.9</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>55</td>
<td>9</td>
<td>22</td>
<td>109.2</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>71</td>
<td>17</td>
<td>6</td>
<td>102.7</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>31</td>
<td>22</td>
<td>44</td>
<td>72.5</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>54</td>
<td>18</td>
<td>22</td>
<td>93.1</td>
</tr>
<tr>
<td>10</td>
<td>21</td>
<td>47</td>
<td>4</td>
<td>26</td>
<td>1159</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>40</td>
<td>23</td>
<td>34</td>
<td>83.8</td>
</tr>
<tr>
<td>12</td>
<td>11</td>
<td>66</td>
<td>9</td>
<td>12</td>
<td>113.3</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
<td>68</td>
<td>8</td>
<td>12</td>
<td>109.4</td>
</tr>
</tbody>
</table>

Ref: T & D:
SAS Program (stepwise variable selection is used)

data example115;
input x1 x2 x3 x4 y;
datalines;
  7  26  6 60  78.5
  1  29 15 52  74.3
 11  56  8 20 104.3
 11  31  8 47  87.6
  7  52  6 33  95.9
 11  55  9 22 109.2
  3  71 17  6 102.7
  1  31 22 44  72.5
  2  54 18 22  93.1
 21  47  4 26 115.9
  1  40 23 34  83.8
 11  66  9 12 113.3
10  68  8 12 109.4
;  
run;
proc reg data=example115;
  model y = x1 x2 x3 x4 /selection=stepwise;
run;
### Selected SAS output

The REG Procedure  
Model: MODEL1  
Dependent Variable: y

**Stepwise Selection: Step 4**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>Type II SS</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>52.57735</td>
<td>2.28617</td>
<td>3062.60416</td>
<td>528.91</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>x1</td>
<td>1.46831</td>
<td>0.12130</td>
<td>848.43186</td>
<td>146.52</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>x2</td>
<td>0.66225</td>
<td>0.04585</td>
<td>1207.78227</td>
<td>208.58</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Bounds on condition number: 1.0551, 4.2205
SAS Output (cont)

* All variables left in the model are significant at the 0.1500 level.

* No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection

<table>
<thead>
<tr>
<th>Step</th>
<th>Variable Entered</th>
<th>Variable Removed</th>
<th>Number Vars In</th>
<th>Partial R-Square</th>
<th>Model R-Square</th>
<th>C(p)</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>x4</td>
<td></td>
<td>1</td>
<td>0.6745</td>
<td>0.6745</td>
<td>138.731</td>
<td>22.80</td>
<td>0.0006</td>
</tr>
<tr>
<td>2</td>
<td>x1</td>
<td></td>
<td>2</td>
<td>0.2979</td>
<td>0.9725</td>
<td>5.4959</td>
<td>108.22</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>3</td>
<td>x2</td>
<td></td>
<td>3</td>
<td>0.0099</td>
<td>0.9823</td>
<td>3.0182</td>
<td>5.03</td>
<td>0.0517</td>
</tr>
<tr>
<td>4</td>
<td>x4</td>
<td></td>
<td>2</td>
<td>0.0037</td>
<td>0.9787</td>
<td>2.6782</td>
<td>1.86</td>
<td>0.2054</td>
</tr>
</tbody>
</table>
5. Variables selection method

B. Best Subsets Regression
Best Subsets Regression

For the stepwise regression algorithm

* The final model is not guaranteed to be optimal in any specified sense.

In the best subsets regression,

* subset of variables is chosen from the collection of all subsets of k predictor variables that optimizes a well-defined objective criterion.
Best Subsets Regression

In the stepwise regression,
* We get only one single final models.

In the best subsets regression,
* The investor could specify a size for the predictors for the model.
Best Subsets Regression

Optimality Criteria

• $r_p^2$-Criterion:
  \[ r_p^2 = \frac{SSR_p}{SST} = 1 - \frac{SSE_p}{SST} \]

• Adjusted $r_p^2$-Criterion:
  \[ r_{adj,p}^2 = 1 - \frac{MSE_p}{MST} \]

• $C_p$-Criterion (recommended for its ease of computation and its ability to judge the predictive power of a model)

  \[ \Gamma_p = \frac{1}{\sigma^2} \sum_{i=1}^{n} [E[\hat{Y}_{ip}] - E[Y_i]]^2 \]

  The sample estimator, Mallows’ $C_p$-statistic, is given by

  \[ C_p = \frac{SSE_p}{\hat{\sigma}^2} + 2(p + 1) - n \]
Best Subsets Regression

Algorithm

Note that our problem is to find the minimum of a given function.

* Use the stepwise subsets regression algorithm and replace the partial F criterion with other criterion such as $C_p$.
* Enumerate all possible cases and find the minimum of the criterion functions.
* Other possibility?
**Best Subsets Regression & SAS**

```sas
proc reg data=example115;
  model y = x1 x2 x3 x4 /selection=ADJRSQ;
run;
```

For the `selection` option, SAS has implemented 9 methods in total. For best subset method, we have the following options:

- **Maximum $R^2$ Improvement (MAXR)**
- **Minimum $R^2$ (MINR) Improvement**
- **$R^2$ Selection (RSQUARE)**
- **Adjusted $R^2$ Selection (ADJRSQ)**
- **Mallows' $C_p$ Selection (CP)**
6. Building A Multiple Regression Model

Steps and Strategy
Modeling is an iterative process. Several cycles of the steps maybe needed before arriving at the final model.

The basic process consists of seven steps.
Get started and Follow the Steps

Categorization by Usage → Collect the Data

Divide the Data ← Explore the Data

Fit Candidate Models

Select and Evaluate

Select the Final Model
Linear Regression Assumptions

* Mean of Error Is 0
* Variance of Error is Constant
* Probability Distribution of Error is Normal
* Errors are Independent
Residual Plot for Functional Form (Linearity)

Add $X^2$ Term

Correct Specification
Residual Plot for Equal Variance

Unequal Variance

Correct Specification

Fan-shaped.
Standardized residuals used typically (residual divided by standard error of prediction)
Residual Plot for Independence

Not Independent

Correct Specification

SR

X

SR

X
Questions?