The General Linear Model

Thus far, we have discussed measures of uncertainty for the estimated parameters ($\hat{\beta}_0, \hat{\beta}_1$) and responses ($\hat{y}_i$) from the simple linear regression model. We would like to extend these ideas to any linear statistical model.

Definition: The **general linear model (GLM)** is any statistical model of the following form:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_{p-1} x_{pi-1} + \epsilon_i, \ i = 1, \ldots, n,$$

where:

- $n$ = the sample size,
- $y_i$ = the $i^{th}$ value of the response variable,
- $x_{ki}$ = the $i^{th}$ value of the $k^{th}$ explanatory variable,
- $\beta_k$ = the parameter corresponding to the $k^{th}$ variable,
- $\epsilon_i$ = the $i^{th}$ error and we usually assume $\epsilon_i \overset{iid}{\sim} (0, \sigma^2)$.

- Let $y = (y_1, \ldots, y_n)'$ and $X$ be a matrix of the values observed for the X-variables (defined below). To use least squares to find the parameter estimates, we can construct the squared error loss function as before:
  $$Q(\beta_0, \beta_1, \ldots, \beta_{p-1}|y, X) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_{p-1} x_{pi-1})^2,$$

  differentiate $Q$ with respect to the $p$ parameters $\beta_0, \beta_1, \ldots, \beta_{p-1}$, set each to zero and solve the resulting linear system of $p$ equations in $p$ unknowns. YIKES! It’s time for matrix algebra!

Matrix Formulation of the GLM: In the GLM $y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_{p-1} x_{pi-1} + \epsilon_i$, let:

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1}, \quad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{p-1,1} \\ 1 & x_{12} & \cdots & x_{p-1,2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{p-1,n} \end{bmatrix}_{n \times p}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}_{p \times 1}, \quad \text{and} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}_{n \times 1}.$$

Then the GLM can be written in matrix notation as:

$$y = X\beta + \epsilon.$$

- If we assume that $\epsilon \sim (0, \sigma^2 I)$, this is known as the **Gauss-Markov model**.

- For temporally or spatially correlated data, we might assume that $\epsilon \sim (0, \sigma^2 V)$.

- If we assume that $y = g(Xb) + \epsilon$, this is known as the **generalized linear model**, where the inverse function of $g(\cdot)$ is known as the link function.
• If we assume that \( \text{Var}(\epsilon) = \Sigma(\theta) \), where the elements of \( \Sigma \) are functions of a parameter vector \( \theta \) with parameter space \((\beta, \theta)\), this is known as the **general linear mixed model**.

• Under the GLM \( y = X\beta + \epsilon \), the least squares loss function can then be written as:

\[
Q(\beta|X, y) = \sum_{i=1}^{n} \epsilon_i^2 = \epsilon^\prime \epsilon = (y - X\beta)'(y - X\beta) = y'y - \beta'X'y - \beta'X'X\beta + \beta'X'y + \beta'X'X\beta.
\]

Applying rules of vector differentiation with respect to \( \beta \) on our least squares function \( Q \), and setting to 0 produces the least squares estimates for the GLM parameters as follows:

\[
\frac{\partial Q}{\partial \beta} \bigg|_{\beta = \hat{\beta}} = \frac{\partial}{\partial \beta} \left[ y'y - 2\beta'X'y + \beta'X'X\beta \right]_{\beta = \hat{\beta}} = -2X'y + 2X'X\beta \bigg|_{\beta = \hat{\beta}} = 0
\]

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

which exists if and only if (iff) \((X'X)^{-1}\) exists \iff \(X_{n \times p}\) is full column rank.

• Note also that since the parameter estimates \( \beta = [\beta_0, \beta_1, \ldots, \beta_{p-1}]' \) are linear functions of the \( y_i \)'s, then if we assume a normal error structure \((\epsilon \sim N(0, \Sigma))\), the \( \hat{\beta}_i \)'s will also follow normal distributions. Thus we can use t-based confidence intervals to characterize uncertainty in their values and t-based tests of significance to test significance of the parameters using p-values.

• From a practical point of view, this means that as long as there are no dependencies among the variables in \( X \), the least squares solution exists. The more correlated sets of \( X \)-variables are, the less stable the inverse \((X'X)^{-1}\) becomes, and the more variability the parameter estimates have.

Example: A study was conducted to determine whether infection surveillance and control programs have reduced the rates of hospital-acquired infection in US hospitals. Specifically, data from a random sample of 28 hospitals on the following variables were collected:

\[
\begin{align*}
y &= \text{(risk)} = \text{the average estimated probability of acquiring infection in hospital}, \\
x_1 &= \text{(stay)} = \text{the average length of stay (days) of all patients in the hospital}, \\
x_2 &= \text{(age)} = \text{the average age of patients (in years)}, \\
x_3 &= \text{(rcr)} = \text{the ratio of the number of cultures performed to number of patients without signs or symptoms of hospital-acquired infection (times 100)}, \\
x_4 &= \text{(school)} = \text{indicator variable for medical school affiliation (1=yes, 0=no)}. 
\end{align*}
\]

For these data, researchers wanted simply to assess whether risk could be explained by the other four explanatory variables. Before fitting a model, the schematic at the top of the next page represents a strategy for data analysis whenever any type of initial model is being developed. The diagram is taken from *The Statistical Sleuth*.
A Strategy for Data Analysis using Statistical Models

**Preliminaries:** Define the questions of interest. Review the design of the study (for thinking about model assumptions). Correct errors in the data.

1. **Explore** the data. Look for initial answers to questions and for potential models.

2. **Formulate** an inferential model.

3. **Check** the model.
   - If appropriate, fit a richer model (with interactions or curvature, for example).
   - Examine residuals.
   - See if extra terms can be dropped.

- **Model not OK**

- **Model OK**

4. **Infer** the answers to the questions of interest using appropriate inferential tools.

**Presentation:** Communicate the results to the intended audience.

- Use graphical tools; consider transformations; check outliers.

- Word the questions of interest in terms of model parameters.

- Check for nonconstant variance; assess outliers. Test whether extra terms in the rich model can be dropped.

- Confidence intervals, tests, prediction intervals, calibration intervals (as needed).

- Answer questions (as much as possible in subject matter language - not statistical language). Make inferential statements compatible with study design.

In trying to adhere to this strategy, a matrix of scatterplots for each of these variables is shown below. What do you notice?
A multiple linear regression model was fit to these data in the form:

\[ y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_4 x_{4i} + \epsilon_i, \quad i = 1, \ldots, 28. \]

In matrix form, this can be written as:

\[
\begin{bmatrix}
4.1 \\
1.6 \\
\vdots \\
3.9
\end{bmatrix}
= 
\begin{bmatrix}
1 & 7.13 & 55.7 & 9.0 & 0 \\
1 & 11.18 & 45.7 & 40.5 & 1 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
1 & 11.15 & 56.5 & 7.7 & 0
\end{bmatrix}
\begin{bmatrix}
\beta_0 \\
\beta_1 \\
\vdots \\
\beta_4
\end{bmatrix}
+ 
\begin{bmatrix}
\epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_{28}
\end{bmatrix}
\]

Using the least squares solution derived earlier, namely \( \hat{\beta} = (X'X)^{-1}X'y \), MatLab was used to calculate the parameter estimates for this model, with the output following:

```matlab
load hosp.mat; % Loads hospital data
vars = [hosp.risk hosp.stay hosp.rcr hosp.age hosp.school]; % Puts all 5 variables into a 28x5 matrix
varnames = {'Infection Risk';'Average Stay';'Culture Ratio';'Patient Age';'Medical School'}; % Assigns text names to all variables for use in the matrix of scatterplots
gplotmatrix(vars); % Matrix of scatterplots
text([.02 .22 .43 .64 .82], repmat(-.1,1,5), varnames, 'FontSize',8); % Places text labels on x-axis of plots
text(repmat(-.08,1,5), [.83 .62 .41 .23 0], varnames, 'FontSize',8, 'Rotation',90); % Places text labels on y-axis of plots
X1 = [ones(28,1) vars(:,2:5)]; % Creates design matrix X1
y = hosp.risk; % Creates response vector y
betahat = inv(X1'*X1)*X1'*y; % Beta parameter estimates

These estimates can be more easily obtained in MatLab using the `regstats` function:

```matlab
X = vars(:,2:5); % Matrix of X-variables only
out = regstats(y,X); % Runs a multiple regression of y on X (GLM)
out.beta % Prints the parameter estimates
```

• Following the development of tools to quantify uncertainty in the simple linear regression model, the analogous results for GLM’s are established in matrix form and illustrated with the hospital example in the pages to follow.

• But first, within this extension to the multiple regression setting, the use of the \( R^2 \)-statistic and the F-statistic to assess the quality of model fit will be introduced.

1. **Interpretation of \( R^2 \)**: In the simple linear regression model, the coefficient of determination \( R^2 \) is the percentage of variation in \( y \) explained by the linear regression of \( y \) on \( x \). This
$R^2$ is the percentage of variation in $y$ explained by the linear regression of $y$ on $x_1, \ldots, x_k$ simultaneously. This value is often denoted as: $R^2_{y,x_1,\ldots,x_k}$ to indicate precisely which variables were used in the model.

How is $R^2_{y,x_1,\ldots,x_k}$ computed? As with SLR, the value $R^2$ is computed as:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{Regression Sum of Squares}}{\text{Total Sum of Squares}},$$

where:

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \quad \text{(total variation in } y),$$

$$\text{SSE} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ki})^2,$$

$$\text{SSR} = \text{SST} - \text{SSE}.$$

- In this way, SST is the total variation in the $y$’s, SSR is that part of the total variation explained by the model, and hence SSE is that part of the total variation remaining unexplained.

Relationship Between $R^2$ and correlation $r$: In simple linear regression (SLR), the statistic $r$ represented the correlation between $y$ and $x$. In multiple linear regression, we are interested in the individual correlations between $y$ and each of the $x_i$’s, $i = 1, \ldots, k$. For this, some new notation is required:

$$r_{yi} = \text{sample correlation between } y \text{ and } x_i, \quad i = 1, \ldots, k.$$

In SLR, it can be shown mathematically that $R^2 = r^2$. Does a similar relationship hold for multiple regression? Two cases:

(a) If $x_1, \ldots, x_k$ are uncorrelated, then: $R^2 = r_{y1}^2 + \cdots + r_{yk}^2$.

(b) If $x_1, \ldots, x_k$ are correlated, then: $R^2 < r_{y1}^2 + \cdots + r_{yk}^2$.

Implications:

- If two variables $X_1$ & $X_2$ are uncorrelated, and the $R^2$ value is 0.4 (insignificant) for a SLR of $y$ on $x_1$ and $x_2$ individually, then the multiple linear regression model with both $x_1$ and $x_2$ included in the model has $R^2 = 0.8$, a good fit.

- So although individual variables may not explain much of the variation in $y$ alone, the group as a whole may provide a good model for explaining the variation in $y$.

- The less correlated the explanatory variables are, the more information toward explaining $y$ can be extracted from them. I.e.: Two highly correlated explanatory variables should never be included together in a model as they provide essentially the same information regarding $y$. 

39
The problems caused by highly correlated explanatory variables is known as the **multi-collinearity** problem, and is investigated through methods such as principal components analysis. From a computational perspective, collinearity among the explanatory variables can lead to numerical instability in the resulting model.

Simple correlations between pairs of variables can be investigated by computing all pairwise correlations among the variables. In MatLab, this can be done easily by typing `corr(X)` to give a matrix of correlations as shown below.

<table>
<thead>
<tr>
<th></th>
<th>stay</th>
<th>rcr</th>
<th>age</th>
<th>school</th>
</tr>
</thead>
<tbody>
<tr>
<td>stay</td>
<td>1.0000</td>
<td>0.5509</td>
<td>0.0762</td>
<td>0.5584</td>
</tr>
<tr>
<td>rcr</td>
<td>0.5509</td>
<td>1.0000</td>
<td>-0.4131</td>
<td>0.6612</td>
</tr>
<tr>
<td>age</td>
<td>0.0762</td>
<td>-0.4131</td>
<td>1.0000</td>
<td>-0.2691</td>
</tr>
<tr>
<td>school</td>
<td>0.5584</td>
<td>0.6612</td>
<td>-0.2691</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

**2. F-Test of Model Significance**: The most common test performed when a GLM is fit is the F-test of model significance. This test tests the hypotheses:

\[ H_0 : \text{The regression model is not significant vs.} \]
\[ H_a : \text{The regression model is significant.} \]

The test statistic for this test is given by:

\[ F = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}/(p - 1)}{\text{SSE}/(n - p)}, \]

where \( p - 1 \) is the number of variables and \( n - p \) is the model degrees of freedom. Under the null hypothesis \( (H_0) \), this statistic \( F \) can be shown to follow an F-distribution with \( p - 1 \) & \( n - p \) degrees of freedom. The larger the value of \( F \), the more evidence there is against \( H_0 \) and thus the more evidence there is of a significant model fit. We will use the p-value from this test to assess whether or not the model is significant (i.e.: whether any of the variables are significant predictors).

**3. Distribution of \( y \)**: Under the assumption that the \( \epsilon_i \) iid \( \sim (0, \sigma^2) \), we have as before that \( y_i \) indep \( \sim (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1}, \sigma^2) \). In matrix form, this can be expressed as:

\[ y \sim (X\beta, \sigma^2 I), \]

where:

\[ \text{Var}(\mathbf{y}) = \begin{bmatrix} \text{Var}(y_1) & \text{Cov}(y_1,y_2) & \cdots & \text{Cov}(y_1,y_n) \\ \text{Cov}(y_1,y_2) & \text{Var}(y_2) & \cdots & \text{Cov}(y_2,y_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(y_1,y_n) & \text{Cov}(y_2,y_n) & \cdots & \text{Var}(y_n) \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma^2 \end{bmatrix} = \sigma^2 I. \]

**4. Estimate of error variance \( \sigma^2 \)**: The error variance is estimated as before using the MSE:

\[ \hat{\sigma}^2 = \frac{\text{MSE}}{n-p} = \frac{\text{SSE}}{n-p} = \frac{\sum \hat{\epsilon}_i^2}{n-p} \]

where the only modification occurs with the degrees of freedom = \( n - p \).
5. Estimate of Var(\( \hat{\beta}_i \)): To compute the variance matrix \( \text{Var}(\hat{\beta}) \), we need the following results. Let \( \mathbf{a} \) be an \( n \times 1 \) vector of constants, let \( \mathbf{A} \) be a \( p \times n \) matrix of constants, and let \( \mathbf{y} \) be an \( n \times 1 \) random vector. Then:

(a) \( \text{Var}(\mathbf{A} \mathbf{y}) = \mathbf{A} \text{Var}(\mathbf{y}) \mathbf{A}' \).
(b) \( \text{Var}(\mathbf{a}' \mathbf{y}) = \mathbf{a}' \text{Var}(\mathbf{y}) \mathbf{a} \).

Viewing the \( x_{pi} \)'s as fixed quantities:

\[
\text{Var}(\hat{\beta}_i) = \text{Var}[X'X]^{-1}X'y = \sigma^2 X'X^{-1}X'y = \sigma^2 X'X^{-1}X'y = \sigma^2 (X'X)^{-1}X'X^{-1} = \sigma^2 \frac{(X'X)^{-1}X'y}{\mathbf{I}}
\]

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

6. Estimate of Var(\( \hat{y}_h \)) for \( \hat{y}_h \) an estimate of \( E(y_h) \): Since \( \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_{p-1,i}, \ i = 1, \ldots, n \), then \( \hat{\mathbf{y}} = [\hat{y}_1, \ldots, \hat{y}_n]' = \mathbf{X} \hat{\beta} \). Computing:

\[
\text{Var}(\hat{y}_h) = \text{Var}(\mathbf{X} \hat{\beta}) = \mathbf{X} \cdot \text{Var}(\hat{\beta}) \mathbf{X}' = \sigma^2 \frac{X(X'X)^{-1}X'}{\mathbf{P}} = \sigma^2 \frac{X(X'X)^{-1}X'}{\mathbf{P}}
\]

7. Confidence intervals for \( \beta_i \)'s & \( E(y_h) \): As before, individual 95\% confidence intervals for \( \beta_i \) and \( \hat{y}_h \) are given by:

\[
\hat{\beta}_i \pm t_{.975} \cdot \text{SE}(\hat{\beta}_i) \quad \text{and} \quad \hat{y}_h \pm t_{.975} \cdot \text{SE}(\hat{y}_h),
\]

where the degrees of freedom associated with the t-critical value is \( n - p \).

- If you wish to have 95\% confidence for a number of intervals simultaneously, the \( \alpha = 0.05 \)-level needs to be adjusted. The most common adjustment is the Bonferroni correction where if we want 95\% confidence (\( \alpha = 0.05 \)) for \( L \) intervals, we find the t-critical value for \( .05/L \) instead of \( .05 \).

Back to the Hospital Example: We can compute the quantities discussed on the previous page using matrix algebra, but the relevant information can be acquired in MatLab from the \texttt{regstats} function output. Some of the more relevant output summaries are shown below. What does it all mean?
out = regstats(y,X); % Regresses y on X (GLM)

out.tstat % Requests the t-test statistics as shown below

% Value Std. Error t-value p-value
Intercept -2.8922 2.9262 -0.9884 0.3333
stay 0.3577 0.1609 2.2233 0.0363
rcr 0.0685 0.0235 2.9188 0.0077
age 0.0545 0.0563 0.9696 0.3424
school 0.0341 0.6094 0.0559 0.9559

out.rsquare 0.6558 % Multiple R-Squared Value
out.mse 0.9734 % Mean Squared Error (MSE)
out.fstat.f 10.9552 % Model F-Statistic
out.fstat.dfr 4 % F-statistic numerator df
out.fstat.dfe 23 % F-statistic denominator df
out.fstat.pval 4.0264e-005 % F-statistic p-value

out.covb % Requests covariance matrix for % parameter estimates

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>8.5629</td>
<td>-0.0104</td>
<td>-0.0250</td>
<td>-0.1469</td>
</tr>
<tr>
<td>stay</td>
<td>-0.0104</td>
<td>0.0259</td>
<td>-0.0016</td>
<td>-0.0038</td>
</tr>
<tr>
<td>rcr</td>
<td>-0.0250</td>
<td>-0.0016</td>
<td>0.0006</td>
<td>0.0006</td>
</tr>
<tr>
<td>age</td>
<td>-0.1469</td>
<td>-0.0038</td>
<td>0.0006</td>
<td>0.0032</td>
</tr>
<tr>
<td>school</td>
<td>0.0823</td>
<td>-0.0329</td>
<td>-0.0056</td>
<td>0.0046</td>
</tr>
</tbody>
</table>

Since school added nothing significant above and beyond the other three explanatory variables (Why?), it was dropped and the model refit. The resulting table of parameter estimates was produced:

out2 = regstats(y,X(:,1:3)); % Reruns the model without "school"

% Value Std. Error t-value p-value
Intercept -2.8997 2.8618 -1.0133 0.3210
stay 0.3607 0.1484 2.4312 0.0229
rcr 0.0690 0.0211 3.2666 0.0033
age 0.0541 0.0546 0.9917 0.3312

Mean squared error: 0.9330
Model F-statistic: 15.2389 on 3 and 24 degrees of freedom
F-Test p-value: 9.2180e-006
Multiple R-Squared: 0.6557
Finally, the age variable was dropped to give the final model below, where the average length of stay (stay) and ratio of number of cultures performed to number of patients without signs or symptoms of hospital-acquired infection times 100 (rcr) together explained 64.2% of the variation in risk. The resulting output is given below.

\[
\text{out3} = \text{regstats}(y,X(:,1:2)); \quad \% \text{Reruns the model w/o "school" & "age"}
\]

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>-0.3234</td>
<td>1.2001</td>
<td>-0.2695</td>
<td>0.7898</td>
</tr>
<tr>
<td>stay</td>
<td>0.4195</td>
<td>0.1360</td>
<td>3.0854</td>
<td>0.0049</td>
</tr>
<tr>
<td>rcr</td>
<td>0.0576</td>
<td>0.0177</td>
<td>3.2549</td>
<td>0.0032</td>
</tr>
</tbody>
</table>

Mean squared error: 0.9324  
Model F-statistic: 22.3813 on 2 and 25 degrees of freedom  
F-Test p-value: 2.6851e-006  
Multiple R-Squared: 0.6416

- It should be mentioned that there are model selection techniques or tools for looking at several competing models and deciding which of them agrees most closely with the data. The use of the $R^2$-statistic is one way of comparing competing models but must be used carefully.

To extract confidence intervals for the parameter estimates from MatLab, we can issue the following commands:

\[
\begin{align*}
\text{n} &= \text{length(hosp.risk)}; \quad \% \text{Sample size} \\
\text{p} &= 3; \quad \% \text{Defines # parameters} \\
\text{bhat} &= \text{out3.beta}; \quad \% \text{Vector of parameter estimates} \\
\text{seb} &= \text{sqrt(diag(out3.covb))}; \quad \% \text{Vector of standard errors} \\
\text{tbonf} &= \text{tinv((1-.025/p),n-p)}; \quad \% \text{Bonferroni-corrected t*-value} \\
\text{ci_b} &= \left[\text{bhat}-\text{tbonf}*\text{seb}, \ldots \right. \\
&\left.\text{bhat}+\text{tbonf}*\text{seb}\right]; \quad \% \text{in 2 columns (lower,upper)} \\
\text{ci_b} &= \% \text{Prints confidence intervals}
\end{align*}
\]

\[
\begin{array}{lcc}
\text{Lower} & \text{Upper} \\
\text{Intercept} & -3.40290499 & 2.7561003 \\
\text{stay} & 0.07062473 & 0.7683569 \\
\text{rcr} & 0.01218240 & 0.1029346 \\
\end{array}
\]

Hence, the simultaneous 95% confidence intervals for $\beta_0, \beta_1,$ and $\beta_2$ are given respectively as:

\[
\beta_0 : (-3.40, 2.76), \quad \beta_1 : (0.07, 0.77), \quad \beta_2 : (0.012, 0.103).
\]
A Second Example: Meadowfoam (*Limnanthes alba*) is a small plant which grows in moist meadows in the Pacific Northwest. A nongreasy and highly stable vegetable oil can be extracted from the meadowfoam seed. To study how to maximize meadowfoam production to a profitable crop, researchers set up the following controlled experiment. Using 6 different light intensities at levels of 150, 300, 450, 600, 750, and 900 $\mu$mol/m$^2$/sec, and 2 different times before the onset of the light treatment (at photoperiodic floral induction (PFI) or 24 days before PFI), the average number of flowers per meadowfoam plant from 10 seedlings in each of the 12 light intensity-light onset treatments was recorded. The entire experiment was repeated twice to give some replication, and produced the following data [M. Seddigh and G.D. Joliff, “Light Intensity Effects on Meadowfoam Growth and Flowering,” *Crop Science* 34 (1994): 497-503.]:

<table>
<thead>
<tr>
<th>Light Intensity ($\mu$mol/m$^2$/sec)</th>
<th>150</th>
<th>300</th>
<th>450</th>
<th>600</th>
<th>750</th>
<th>900</th>
</tr>
</thead>
<tbody>
<tr>
<td>at PFI</td>
<td>62.3</td>
<td>55.3</td>
<td>49.6</td>
<td>39.4</td>
<td>31.3</td>
<td>36.8</td>
</tr>
<tr>
<td>24 days (1)</td>
<td>77.4</td>
<td>54.2</td>
<td>61.9</td>
<td>45.7</td>
<td>44.9</td>
<td>41.9</td>
</tr>
<tr>
<td>before PFI</td>
<td>75.6</td>
<td>78.0</td>
<td>71.1</td>
<td>52.2</td>
<td>45.6</td>
<td>44.4</td>
</tr>
</tbody>
</table>

- This experiment has the following variables of interest:

  \[ y = \text{the average number of flowers on a meadowfoam plant} \]

  \[ x_1 = \text{the light intensity used on the plant ($\mu$mol/m$^2$/sec)} \]

  \[ x_2 = \text{the timing of the light treatment (at PFI, 24 days before PFI)} \]

- Some questions sought by the researchers are:

  1. What are the effects of the different light intensity levels?
  2. What is the effect of timing?
  3. Does the effect of intensity depend on timing?

The first step taken toward analysis of the meadowfoam data is to make a simple scatterplot of our response variable \( y \) vs. \( x_1 \), at the two levels of \( x_2 \). We are treating \( x_1 \) (light intensity) as a quantitative variable, and \( x_2 \) (timing) as a categorical variable at 2 levels. Under this viewpoint, the following scatterplot with fitted regression lines depicts the relationship between these three variables. This plot was generated in MatLab using the gscatter function as illustrated in the file *meadowfoam.m*.
What does this plot tell you? Can you qualitatively answer the three posed questions:

1. What are the effects of the different light intensity levels?
2. What is the effect of timing?
3. Does the effect of intensity depend on timing?

Based on this plot then, what type of model might you suggest to fit? Surely, no quadratic terms are needed, and although there does not appear to be an interaction between intensity and timing, it might be worth testing for this lack of interaction formally. Letting \( x_1 \) = intensity, and \( x_2 = \begin{cases} 1 & \text{if the timing is “at PFI”} \\ 0 & \text{otherwise} \end{cases} \), one suggested model might be:

\[
y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon_i, \; i = 1, \ldots, n.
\]

- This model enables separate regression fits for each of the two timings. How is this accomplished? Consider what happens for timings of 0 and 1:

Model for Timing 0:

Model for Timing 1:

- In this way, the four model parameters have the following meanings:

  \[ \beta_0 = \text{the intercept when timing} = 0, \]
  \[ \beta_1 = \text{the slope when timing} = 0, \]
  \[ \beta_2 = \text{the increase in the intercept for timing} = 1, \]
  \[ \beta_3 = \text{the increase in the slope for timing} = 1. \]

- To include the interaction term, we need to create a variable \( x_1 x_2 \). To do this in MatLab, simply type \( x1x2 = \text{timing.*intensity} \); and add this term to the design matrix \( X \) as follows:

\[
x1x2 = \text{timimg.*intensity}; \quad \% \text{Computes interaction between x1 & x2}
X = [\text{intensity timing x1x2}]; \quad \% \text{Sets up design matrix with interaction}
reg2 = \text{regstats(numflow,X)}; \quad \% \text{Fits regression model with interaction}
\]

The multiple regression model with an interaction term, given as:

\[
y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon_i, \; i = 1, \ldots, n,
\]

was fit in MatLab using the regstats function as shown above. With \( y = \text{avg. number of flowers} \) as the response variable, \( x_1 = \text{intensity} \), \( x_2 = \text{timing (“at PFI”} = 1, \text{“24 days before PFI”} = 0) \), and the interaction \( x_1 x_2 \), the following output was obtained.

45
ANOVA Table

<table>
<thead>
<tr>
<th>Source of Variance</th>
<th>Sums of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>3467.276</td>
<td>3</td>
<td>1155.759</td>
<td>26.549</td>
<td>.000</td>
</tr>
<tr>
<td>Residual</td>
<td>870.660</td>
<td>20</td>
<td>43.533</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4337.936</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ R^2 = 0.799 \]

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Beta</th>
<th>Std. Error</th>
<th>t</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>83.147</td>
<td>4.343</td>
<td>19.144</td>
<td>.000</td>
</tr>
<tr>
<td>Light Intensity</td>
<td>-.040</td>
<td>.007</td>
<td>-5.362</td>
<td>.000</td>
</tr>
<tr>
<td>Timing</td>
<td>-11.523</td>
<td>6.142</td>
<td>-1.876</td>
<td>.075</td>
</tr>
<tr>
<td>Intensity × Timing</td>
<td>-1.210E-03</td>
<td>.011</td>
<td>-.115</td>
<td>.910</td>
</tr>
</tbody>
</table>

- Is the regression model significant?

- Is there a significant interaction between light intensity and timing? (i.e.: Does the effect of intensity on the average number of flowers per plant depend on timing?)

- Is timing significant?

Since the interaction term is not significant, it was removed from the model, and the multiple linear regression model below was fit:

\[ y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i, \ i = 1, \ldots, n. \]

Time permitting later in the course, we will discuss methods of model selection with the goal of finding the “best model” in some sense. Probably the most common method of the last 5 years is the Akaike’s Information Criterion (AIC), although there are many other tools available.

The output from the first-order model with both \( x_1 \) (intensity) and \( x_2 \) (timing) appears on the next page.
**ANOVA Table**

<table>
<thead>
<tr>
<th>Source of Variance</th>
<th>Sums of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>3466.700</td>
<td>2</td>
<td>1733.350</td>
<td>41.780</td>
<td>.000</td>
</tr>
<tr>
<td>Residual</td>
<td>871.236</td>
<td>21</td>
<td>41.487</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4337.936</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*R^2 = 0.799*

**Coefficients**

<table>
<thead>
<tr>
<th></th>
<th>Beta</th>
<th>Std. Error</th>
<th>t</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>83.464</td>
<td>3.274</td>
<td>25.495</td>
<td>.000</td>
</tr>
<tr>
<td>Light Intensity</td>
<td>-.040</td>
<td>.005</td>
<td>-7.886</td>
<td>.000</td>
</tr>
<tr>
<td>Timing</td>
<td>-12.158</td>
<td>2.630</td>
<td>-4.624</td>
<td>.000</td>
</tr>
</tbody>
</table>

- Is the model significant?
- What are the effects of the different light intensity levels?
- What is the effect of timing?
- Find 95% confidence intervals for β₀, β₁, & β₂.
- How much did the *R^2* change when the interaction term was dropped from the model? What does this mean?
- How did the SSE, SSR, and SST change between the two models? What do these changes mean?
- Can the inferences made above be trusted? How can we tell?
- What assumptions are we making in drawing these inferences?
If these inferences are valid, can we establish a causal relationship between the average number of flowers per plant and the intensity and timing of the light treatment? Why or why not?

To investigate the assumptions underlying the inferences made above, a residual plot and normal quantile (QQ) plot were created via MatLab using the code below.

```matlab
X = [intensity timing]; % Defines design matrix X
reg3 = regstats(numflow,X); % Regresses numflow on X-matrix variables
figure(1) % 1st Figure
plot(reg3.yhat,reg3.r,'ko'); % Plots residuals vs. predicted y-values
xlabel('Predicted Values','fontsize',14,'fontweight','b');
ylabel('Residuals','fontsize',14,'fontweight','b');
title('Residual Plot','fontsize',14,'fontweight','b');
figure(2); % 2nd Figure
qqplot(reg3.r); % Normal quantile plot of residuals
xlabel('Standard Normal Quantiles','fontsize',14,'fontweight','b');
ylabel('Residuals','fontsize',14,'fontweight','b');
title('Normal Quantile Plot','fontsize',14,'fontweight','b');
```

- In viewing the residual plot, there does not appear to be any departure from the homogeneous variance assumption on the errors.
- There is some S-shaped pattern in the normal quantile plot, but nothing to get too excited about with just 24 data values. The Lilliefors test statistic is \( W' = 0.1268 \), with a large p-value, indicating no significant departure from normality in the residuals (In MatLab, type: \([h p lstat cv]\) = lillietest(reg3.r)). The p-value is reported in \( p \) and the test statistic in \( lstat \).