Announcements
Chapter 5: Linear Regression in Matrix Form

The SLR Model in Scalar Form

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim iid \ N(0, \sigma^2) \]
Consider now writing an equation for each observation:

\[ Y_1 = \beta_0 + \beta_1 X_1 + \epsilon_1 \]
\[ Y_2 = \beta_0 + \beta_1 X_2 + \epsilon_2 \]
\[ \vdots \]
\[ Y_n = \beta_0 + \beta_1 X_n + \epsilon_n \]
The SLR Model in Matrix Form

\[
\begin{bmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_n
\end{bmatrix}
= \begin{bmatrix}
\beta_0 + \beta_1 X_1 \\
\beta_0 + \beta_1 X_2 \\
\vdots \\
\beta_0 + \beta_1 X_n
\end{bmatrix} + \begin{bmatrix}
\epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_n
\end{bmatrix}
\]

\[
\begin{bmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_n
\end{bmatrix}
= \begin{bmatrix}
1 & X_1 \\
1 & X_2 \\
\vdots & \vdots \\
1 & X_n
\end{bmatrix} \begin{bmatrix}
\beta_0 \\
\beta_1
\end{bmatrix} + \begin{bmatrix}
\epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_n
\end{bmatrix}
\]
(I will try to use **bold** symbols for matrices. At first, I will also indicate the dimensions as a subscript to the symbol.)

- $X$ is called the **design matrix**.
- $\beta$ is the vector of parameters
- $\epsilon$ is the error vector
- $Y$ is the response vector
The Design Matrix

\[ X_{n \times 2} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \]

Vector of Parameters

\[ \beta_{2 \times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \]
Vector of Error Terms

\[ \epsilon_{n \times 1} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \]
Vector of Responses

$Y_{n \times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$
Thus,

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

\[ Y_{n \times 1} = X_{n \times 2} \beta_{2 \times 1} + \epsilon_{n \times 1} \]
Variance-Covariance Matrix

In general, for any set of variables $U_1, U_2, \ldots, U_n$, their variance-covariance matrix is defined to be

$$\sigma^2\{U\} = \begin{bmatrix}
\sigma^2\{U_1\} & \sigma\{U_1, U_2\} & \cdots & \sigma\{U_1, U_n\} \\
\sigma\{U_2, U_1\} & \sigma^2\{U_2\} & \ddots & \vdots \\
\vdots & \ddots & \ddots & \sigma\{U_{n-1}, U_n\} \\
\sigma\{U_n, U_1\} & \cdots & \sigma\{U_n, U_{n-1}\} & \sigma^2\{U_n\}
\end{bmatrix}$$

where $\sigma^2\{U_i\}$ is the variance of $U_i$, and $\sigma\{U_i, U_j\}$ is the covariance of $U_i$ and $U_j$. 
When variables are uncorrelated, that means their covariance is 0. The variance-covariance matrix of uncorrelated variables will be a **diagonal** matrix, since all the covariances are 0.

**Note:** Variables that are independent will also be uncorrelated. So when variables are correlated, they are automatically dependent. However, it is possible to have variances that are dependent but uncorrelated, since correlation only measures **linear** dependence. A nice thing about normally distributed RV’s is that they are a convenient special case: if they are uncorrelated, they are also independent.
Covariance Matrix of $\epsilon$

\[
\sigma^2 \{\epsilon\}_{n \times n} = \text{Cov} \begin{bmatrix} 
\epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_n 
\end{bmatrix} = \sigma^2 I_{n \times n} = 
\begin{bmatrix}
\sigma^2 & 0 & \cdots & 0 \\
0 & \sigma^2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \sigma^2 
\end{bmatrix}
\]
Covariance Matrix of $Y$

$$\sigma^2 \{Y\}_{n \times n} = Cov \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \sigma^2 I_{n \times n}$$
Distributional Assumptions in Matrix Form

\[ \epsilon \sim N(0, \sigma^2 I) \]

\( I \) is an \( n \times n \) identity matrix.

- Ones in the diagonal elements specify that the variance of each \( \epsilon_i \) is \( 1 \) times \( \sigma^2 \).

- Zeros in the off-diagonal elements specify that the covariance between different \( \epsilon_i \) is zero.

- This implies that the correlations are zero.
Parameter Estimation

Least Squares

Residuals are $\epsilon = Y - X\beta$. Want to minimize sum of squared residuals.

$$\sum \epsilon_i^2 = [\epsilon_1 \epsilon_2 \cdots \epsilon_n] \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} = \epsilon'\epsilon$$
We want to minimize $\epsilon' \epsilon = (Y - X\beta)'(Y - X\beta)$, where the “prime” ($'$) denotes the transpose of the matrix (exchange the rows and columns).

We take the derivative with respect to the vector $\beta$. This is like a quadratic function: think “$(Y - X\beta)^2$”.

The derivative works out to 2 times the derivative of $(Y - X\beta)'$ with respect to $\beta$.

That is, $\frac{d}{d\beta} ((Y - X\beta)'(Y - X\beta)) = -2X'(Y - X\beta)$. We set this equal to 0 (a vector of zeros), and solve for $\beta$.

So, $-2X'(Y - X\beta) = 0$. Or, $X'Y = X'X\beta$ (the “normal” equations).
Normal Equations

\[ X'Y = (X'X)\beta \]

Solving this equation for \( \beta \) gives the least squares solution for

\[ b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}. \]

Multiply on the left by the inverse of the matrix \( X'X \). (Notice that the matrix \( X'X \) is a 2 \( \times \) 2 square matrix for SLR.)
\[ b = \left( X'X \right)^{-1} X'Y \]

REMEMBER THIS.
Reality Break:

This is just to convince you that we have done nothing new nor magic – all we are doing is writing the same old formulas for $b_0$ and $b_1$ in matrix format. Do NOT worry if you cannot reproduce the following algebra, but you SHOULD try to follow it so that you believe me that this is really not a new formula.
Recall in Topic 1, we had

\[
b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \equiv \frac{SS_{XY}}{SS_X}
\]

\[
b_0 = \bar{Y} - b_1 \bar{X}
\]
Now let’s look at the pieces of the new formula:

\[
\begin{align*}
X'X &= \begin{bmatrix}
1 & 1 & \cdots & 1 \\
X_1 & X_2 & \cdots & X_n
\end{bmatrix}
\begin{bmatrix}
1 & X_1 \\
1 & X_2 \\
\vdots & \vdots \\
1 & X_n
\end{bmatrix}
= \begin{bmatrix}
n & \sum X_i \\
\sum X_i & \sum X_i^2
\end{bmatrix} \\
(X'X)^{-1} &= \frac{1}{n \sum X_i^2 - (\sum X_i)^2}
\begin{bmatrix}
\sum X_i^2 & -\sum X_i \\
-\sum X_i & n
\end{bmatrix}
= \frac{1}{nSS_X}
\begin{bmatrix}
\sum X_i^2 \\
-\sum X
\end{bmatrix} \\
X'Y &= \begin{bmatrix}
1 & 1 & \cdots & 1 \\
X_1 & X_2 & \cdots & X_n
\end{bmatrix}
\begin{bmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_n
\end{bmatrix}
= \begin{bmatrix}
\sum Y_i \\
\sum X_i Y_i
\end{bmatrix}
\end{align*}
\]
Plug these into the equation for $b$:

$$b = (X'X)^{-1}X'Y = \frac{1}{nSS_X} \left[ \begin{array}{c} \sum X_i^2 \\ - \sum X_i \end{array} \right] \left[ \begin{array}{c} \sum Y_i \\ n \end{array} \right]$$

$$= \frac{1}{nSS_X} \left[ \begin{array}{c} (\sum X_i^2)(\sum Y_i) - (\sum X_i)(\sum X_i Y_i) \\ - (\sum X_i)(\sum Y_i) + n \sum X_i Y_i \end{array} \right]$$

$$= \frac{1}{SS_X} \left[ \begin{array}{c} \bar{Y}(\sum X_i^2) - \bar{X} \sum X_i Y_i \\ \bar{X} \sum X_i Y_i - n\bar{X}\bar{Y} \end{array} \right]$$

$$= \frac{1}{SS_X} \left[ \begin{array}{c} \bar{Y}(\sum X_i^2) - \bar{Y}(n\bar{X}^2) + \bar{X}(n\bar{X}\bar{Y}) - \bar{X} \sum X_i Y_i \\ SP_{XY} \end{array} \right]$$

$$= \frac{1}{SS_X} \left[ \begin{array}{c} \bar{Y}SS_X - SP_{XY} \bar{X} \\ SP_{XY} \end{array} \right] = \left[ \begin{array}{c} \bar{Y} - \frac{SP_{XY}}{SS_X} \bar{X} \\ \frac{SP_{XY}}{SS_X} \end{array} \right] = \left[ \begin{array}{c} b_0 \\ b_1 \end{array} \right] ,$$
where

\[
SS_X = \sum X_i^2 - n\bar{X}^2 = \sum (X_i - \bar{X})^2
\]

\[
SP_{XY} = \sum X_i Y_i - n\bar{X}\bar{Y} = \sum (X_i - \bar{X})(Y_i - \bar{Y})
\]

All we have done is to write the same old formulas for \(b_0\) and \(b_1\) in a fancy new format. See KNNL page 197 for details. Why have we bothered to do this? The cool part is that the same approach works for multiple regression. All we do is make \(X\) and \(b\) into bigger matrices, and use exactly the same formula.
Other Quantities in Matrix Form

Fitted Values

\[
\hat{Y} = \begin{bmatrix}
\hat{Y}_1 \\
\hat{Y}_2 \\
\vdots \\
\hat{Y}_n
\end{bmatrix} = \begin{bmatrix}
b_0 + b_1 X_1 \\
b_0 + b_1 X_2 \\
\vdots \\
b_0 + b_1 X_n
\end{bmatrix} = \begin{bmatrix}
1 & X_1 \\
1 & X_2 \\
\vdots & \vdots \\
1 & X_n
\end{bmatrix} \begin{bmatrix}
b_0 \\
b_1
\end{bmatrix} = Xb
\]
Hat Matrix

\[ \hat{Y} = Xb \]
\[ \hat{Y} = X(X'X)^{-1}X'Y \]
\[ \hat{Y} = HY \]

where \( H = X(X'X)^{-1}X' \).

We call this the “hat matrix” because it turns \( Y \)’s into \( \hat{Y} \)’s.
**Estimated Covariance Matrix of** $\hat{b}$

This matrix $\hat{b}$ is a linear combination of the elements of $\mathbf{Y}$.

These estimates are normal if $\mathbf{Y}$ is normal.

These estimates will be approximately normal in general.
The vector \( \mathbf{b} \) is normally distributed with mean

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

and covariance matrix

\[
\sigma^2 \left( (X'X)^{-1}X \right) I \left( (X'X)^{-1}X' \right)' = \sigma^2 (X'X)^{-1}(X'X) \left( (X'X)^{-1} \right)'
\]

\[
= \sigma^2 (X'X)^{-1}
\]

using the fact that both \( X'X \) and its inverse are symmetric, so \( \left( (X'X)^{-1} \right)' = (X'X)^{-1} \).
Multiple Regression

Data for Multiple Regression

- $Y_i$ is the response variable (as usual)
- $X_{i,1}, X_{i,2}, \ldots, X_{i,p-1}$ are the $p - 1$ explanatory variables for cases $i = 1$ to $n$.
- Example – In Homework #1 you considered modeling GPA as a function of entrance exam score. But we could also consider intelligence test scores and high
school GPA as potential predictors. This would be 3 variables, so \( p = 4 \).

- Potential problem to remember!!! These predictor variables are likely to be themselves correlated. We always want to be careful of using variables that are themselves strongly correlated as predictors together in the same model.
The Multiple Regression Model

\[ Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_{p-1} X_{i,p-1} + \epsilon_i \text{ for } i = 1, 2, \ldots, n \]

where

- \( Y_i \) is the value of the response variable for the \( i \)th case.
- \( \epsilon_i \sim iid N(0, \sigma^2) \) (exactly as before!)
- \( \beta_0 \) is the intercept (think multidimensionally).
- \( \beta_1, \beta_2, \ldots, \beta_{p-1} \) are the regression coefficients for the explanatory variables.
• $X_{i,k}$ is the value of the $k$th explanatory variable for the $i$th case.

• Parameters as usual include all of the $\beta$'s as well as $\sigma^2$. These need to be estimated from the data.
Interesting Special Cases

- Polynomial model:

\[ Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_{p-1} X_i^{p-1} + \epsilon_i \]

- \( X \)'s can be indicator or dummy variables with \( X = 0 \) or \( 1 \) (or any other two distinct numbers) as possible values (e.g. ANOVA model). Interactions between explanatory variables are then expressed as a product of the \( X \)'s:

\[ Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,1} X_{i,2} + \epsilon_i \]
Model in Matrix Form

\[ Y_{n \times 1} = X_{n \times p} \beta_{p \times 1} + \epsilon_{n \times 1} \]

\[ \epsilon \sim N(0, \sigma^2 I_{n \times n}) \]

\[ Y \sim N(X \beta, \sigma^2 I) \]
Design Matrix $X$:

$$X = \begin{bmatrix}
1 & X_{1,1} & X_{1,2} & \cdots & X_{1,p-1} \\
1 & X_{2,1} & X_{2,2} & \cdots & X_{2,p-1} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & X_{n,1} & X_{n,2} & \cdots & X_{n,p-1}
\end{bmatrix}$$
Coefficient matrix $\beta$: 

$$
\beta = \begin{bmatrix}
\beta_0 \\
\beta_1 \\
\vdots \\
\beta_{p-1}
\end{bmatrix}
$$
Parameter Estimation

Least Squares

Find $b$ to minimize $SSE = (Y - Xb)'(Y - Xb)$

Obtain normal equations as before: $X'Xb = X'Y$

Least Squares Solution

$$b = (X'X)^{-1}X'Y$$
Fitted (predicted) values for the mean of $Y$ are

$$\hat{Y} = Xb = X(X'X)^{-1}X'Y = HY,$$

where $H = X(X'X)^{-1}X'$. 
Residuals

\[ e = Y - \hat{Y} = Y - HY = (I - H)Y \]

Notice that the matrices \( H \) and \( (I - H) \) have two special properties. They are

- **Symmetric:** \( H = H' \) and \( (I - H)' = (I - H) \).
- **Idempotent:** \( H^2 = H \) and
  \[ (I - H)(I - H) = (I - H) \]
Covariance Matrix of Residuals

\[
Cov(e) = \sigma^2(I - H)(I - H)' = \sigma^2(I - H)
\]

\[
Var(e_i) = \sigma^2(1 - h_{i,i}),
\]

where \( h_{i,i} \) is the \( i \)th diagonal element of \( H \).

**Note:** \( h_{i,i} = X_i'(X'X)^{-1}X_i \) where \( X_i' = [1 \ X_{i,1} \cdots \ X_{i,p-1}] \).

Residuals \( e_i \) are usually somewhat correlated:

\[
\text{cov}(e_i, e_j) = -\sigma^2 h_{i,j}; \text{ this is not unexpected, since they sum to 0.}
\]
Estimation of $\sigma$

Since we have estimated $p$ parameters, $SSE = e'e$ has $df_E = n - p$. The estimate for $\sigma^2$ is the usual estimate:

$$s^2 = \frac{e'e}{n - p} = \frac{(Y - Xb)'(Y - Xb)}{n - p} = \frac{SSE}{df_E} = MSE$$

$$s = \sqrt{s^2} = \text{Root MSE}$$
**Distribution of \(b\)**

We know that \(b = (X'X)^{-1}X'Y\). The only RV involved is \(Y\), so the distribution of \(b\) is based on the distribution of \(Y\).

Since \(Y \sim N(X\beta, \sigma^2I)\), and using the multivariate theorem from earlier (if you like, go through the details...
on your own), we have

\[ E(b) = ((X'X)^{-1}X')X\beta = \beta \]

\[ \sigma^2 \{b\} = \text{Cov}(b) = \sigma^2(X'X)^{-1} \]

Since \( \sigma^2 \) is estimated by the MSE \( s^2 \), \( \sigma^2 \{b\} \) is estimated by \( s^2(X'X)^{-1} \).
ANOVA Table

Sources of variation are

- Model (SAS) or Regression (KNNL)
- Error (Residual)
- Total

$SS$ and $df$ add as before

$$SS_M + SSE = SST$$

$$df_M + df_E = df_{Total}$$

but their values are different from SLR.
Sum of Squares

\[
SSM = \sum (\hat{Y}_i - \bar{Y})^2
\]

\[
SSE = \sum (Y_i - \hat{Y}_i)^2
\]

\[
SSTO = \sum (Y_i - \bar{Y})^2
\]
Degrees of Freedom

\[ df_M = p - 1 \]
\[ df_E = n - p \]
\[ df_{Total} = n - 1 \]

The total degrees have not changed from SLR, but the model \( df \) has increased from 1 to \( p - 1 \), i.e., the number of \( X \) variables. Correspondingly, the error \( df \) has decreased from \( n - 2 \) to \( n - p \).
Mean Squares

\[ MSM = \frac{SSM}{df_M} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{p - 1} \]

\[ MSE = \frac{SSE}{df_E} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n - p} \]

\[ MST = \frac{SSTO}{df_{Total}} = \frac{\sum (Y_i - \bar{Y})^2}{n - 1} \]
ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MSE</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>$df_M = p - 1$</td>
<td>$SS_M$</td>
<td>$MS_M$</td>
<td>$\frac{MS_M}{MSE}$</td>
</tr>
<tr>
<td>Error</td>
<td>$df_E = n - p$</td>
<td>$SSE$</td>
<td>$MSE$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$df_T = n - 1$</td>
<td>$SST$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**F-test**

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$ (all regression coefficients are zero)  

$H_A : \beta_k \neq 0$, for at least one $k = 1, \ldots, p - 1$; at least one of the $\beta$’s is non-zero (or, not all the $\beta$’s are zero).

$F = \frac{MSM}{MSE}$

Under $H_0$, $F \sim F_{p-1,n-p}$

Reject $H_0$ if $F$ is larger than critical value; if using SAS, reject $H_0$ if $p$-value $< \alpha = 0.05$. 
The squared multiple regression correlation ($R^2$) gives the proportion of variation in the response variable explained by the explanatory variables. It is sometimes called the *coefficient of multiple determination* (KNNL, page 236).

\[ R^2 = \frac{SSM}{SST} \]  (the proportion of variation explained by the model)

\[ R^2 = 1 - \left( \frac{SSE}{SST} \right) \]  (1 − the proportion not explained by the model)
$F$ and $R^2$ are related:

$$F = \frac{R^2 / (p - 1)}{(1 - R^2) / (n - p)}$$
Inference for Individual Regression Coefficients

Confidence Interval for $\beta_k$

We know that $\mathbf{b} \sim N(\beta, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1})$

Define

$$s^2\{\mathbf{b}\}_{p \times p} = MSE \times (\mathbf{X}'\mathbf{X})^{-1}$$

$$s^2\{b_k\} = [s^2\{\mathbf{b}\}]_{k,k}, \text{ the } k\text{th diagonal element}$$

CI for $\beta_k$: $b_k \pm t^c s\{b_k\}$, where $t^c = t_{n-p}(0.975)$. 
Significance Test for $\beta_k$

$H_0 : \beta_k = 0$

Same test statistic $t^* = b_k / s\{b_k\}$

Still use $df_E$ which now is equal to $n - p$

$p$-value computed from $t_{n-p}$ distribution.

This tests the significance of a variable given that the other variables are already in the model (i.e., fitted last).

Unlike in SLR, the $t$-tests for $\beta$ are different from the $F$-test.
Example: Study of CS Students

Problem: Computer science majors at Purdue have a large drop-out rate.

Potential Solution: Can we find predictors of success? Predictors must be available at time of entry into program.
Data Available

Grade point average (GPA) after three semesters ($Y_i$, the response variable)

Five potential predictors ($p = 6$)

- $X_1 =$ High school math grades (HSM)
- $X_2 =$ High school science grades (HSS)
- $X_3 =$ High school English grades (HSE)
- $X_4 =$ SAT Math (SATM)
- $X_5 =$ SAT Verbal (SATV)
• Gender (1 = male, 2 = female) (we will ignore this one right now, since it is not a continuous variable).

We have \( n = 224 \) observations, so if all five variables are included, the design matrix \( X \) is \( 224 \times 6 \). The SAS program used to generate output for this is \texttt{cs.sas}.
Look at the individual variables

Our first goal should be to take a look at the variables to see...

- Is there anything that sticks out as unusual for any of the variables?

- How are these variables related to each other (pairwise)? *If two predictor variables are strongly correlated, we wouldn’t want to use them in the same model!*
We do this by looking at statistics and plots.

data cs;
  infile 'H:\System\Desktop\csdata.dat';
  input id gpa hsm hss hse satm satv genderm1;
Descriptive Statistics: \texttt{proc means}

\begin{verbatim}
proc means data=cs maxdec=2;
   var gpa hsm hss hse satm satv;
\end{verbatim}

The option \texttt{maxdec = 2} sets the number of decimal places in the output to 2 (just showing you how).
## Output from `proc means`

The `MEANS` Procedure

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>gpa</td>
<td>224</td>
<td>2.64</td>
<td>0.78</td>
<td>0.12</td>
<td>4.00</td>
</tr>
<tr>
<td>hsm</td>
<td>224</td>
<td>8.32</td>
<td>1.64</td>
<td>2.00</td>
<td>10.00</td>
</tr>
<tr>
<td>hss</td>
<td>224</td>
<td>8.09</td>
<td>1.70</td>
<td>3.00</td>
<td>10.00</td>
</tr>
<tr>
<td>hse</td>
<td>224</td>
<td>8.09</td>
<td>1.51</td>
<td>3.00</td>
<td>10.00</td>
</tr>
<tr>
<td>satm</td>
<td>224</td>
<td>595.29</td>
<td>86.40</td>
<td>300.00</td>
<td>800.00</td>
</tr>
<tr>
<td>satv</td>
<td>224</td>
<td>504.55</td>
<td>92.61</td>
<td>285.00</td>
<td>760.00</td>
</tr>
</tbody>
</table>
Descriptive Statistics

Note that `proc univariate` also provides lots of other information, not shown.

```
proc univariate data=cs noprint;
   var gpa hsm hss hse satm satv;
   histogram gpa hsm hss hse satm satv /normal;
```
SAT Math and SAT Verbal
NOTE: If you want the plots (e.g., histogram, qqplot) and not the copious output from proc univariate, use a noprint statement

```
proc univariate data = cs noprint;
   histogram gpa / normal;
```

Getting a Scatter Plot Matrix
Correlations

SAS will give us the $r$ (correlation) value between pairs of random variables in a data set using `proc corr`.

```sas
proc corr data=cs;
  var hsm hss hse;
```

<table>
<thead>
<tr>
<th></th>
<th>hsm</th>
<th>hss</th>
<th>hse</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsm</td>
<td>1.00000</td>
<td>0.57569</td>
<td>0.44689</td>
</tr>
<tr>
<td></td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>hss</td>
<td>0.57569</td>
<td>1.00000</td>
<td>0.57937</td>
</tr>
<tr>
<td></td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>hse</td>
<td>0.44689</td>
<td>0.57937</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td></td>
</tr>
</tbody>
</table>
To get rid of those $p$-values, which make it difficult to read, use a `noprob` statement. Here are also a few different ways to call `proc corr`:

```plaintext
proc corr data=cs noprob;
  var satm satv;
```

<table>
<thead>
<tr>
<th></th>
<th>satm</th>
<th>satv</th>
</tr>
</thead>
<tbody>
<tr>
<td>satm</td>
<td>1.00000</td>
<td>0.46394</td>
</tr>
<tr>
<td>satv</td>
<td>0.46394</td>
<td>1.00000</td>
</tr>
</tbody>
</table>
proc corr data=cs noprob;
  var hsm hss hse;
  with satm satv;
.
.
hsm  hss  hse
satm  0.45351  0.24048  0.10828
satv  0.22112  0.26170  0.24371

proc corr data=cs noprob;
  var hsm hss hse satm satv;
  with gpa;
.
.
hsm  hss  hse  satm  satv
gpa  0.43650  0.32943  0.28900  0.25171  0.11449
Notice that not only do the $X$’s correlate with $Y$ (this is good), but the $X$’s correlate with *each other* (this is bad). That means that some of the $X$’s may be *redundant* in predicting $Y$.

**Use High School Grades to Predict GPA**

```
proc reg data=cs;
  model gpa=hsm hss hse;
```
### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>27.71233</td>
<td>9.23744</td>
</tr>
<tr>
<td>Error</td>
<td>220</td>
<td>107.75046</td>
<td>0.48977</td>
</tr>
<tr>
<td>Corrected Total</td>
<td>223</td>
<td>135.46279</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>18.86</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

- Root MSE: 0.69984
- R-Square: 0.2046
- Dependent Mean: 2.63522
- Adj R-Sq: 0.1937
- Coeff Var: 26.55711
### Parameter Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>0.58988</td>
<td>0.29424</td>
</tr>
<tr>
<td>hsm</td>
<td>1</td>
<td>0.16857</td>
<td>0.03549</td>
</tr>
<tr>
<td>hss</td>
<td>1</td>
<td>0.03432</td>
<td>0.03756</td>
</tr>
<tr>
<td>hse</td>
<td>1</td>
<td>0.04510</td>
<td>0.03870</td>
</tr>
</tbody>
</table>

| Variable | t Value | Pr > |t| |
|----------|---------|------|---|
| Intercept| 2.00    | 0.0462|
| hsm      | 4.75    | <.0001|
| hss      | 0.91    | 0.3619|
| hse      | 1.17    | 0.2451|
### Remove HSS

```plaintext
proc reg data=cs;
    model gpa=hsm hse;
```

R-Square 0.2016  Adj R-Sq 0.1943

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>0.62423</td>
<td>0.29172</td>
</tr>
<tr>
<td>hsm</td>
<td>1</td>
<td>0.18265</td>
<td>0.03196</td>
</tr>
<tr>
<td>hse</td>
<td>1</td>
<td>0.06067</td>
<td>0.03473</td>
</tr>
</tbody>
</table>

| Variable | t Value | Pr > |t| |
|----------|---------|------|------|
| Intercept| 2.14    | 0.0335|
| hsm      | 5.72    | <.0001|
| hse      | 1.75    | 0.0820|
Rerun with HSM only

```plaintext
proc reg data=cs;
    model gpa=hsm;
```

R-Square 0.1905  Adj R-Sq 0.1869

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>0.90768</td>
<td>0.24355</td>
</tr>
<tr>
<td>hsm</td>
<td>1</td>
<td>0.20760</td>
<td>0.02872</td>
</tr>
</tbody>
</table>

| Variable  | t Value | Pr > |t| |
|-----------|---------|------|----|
| Intercept | 3.73    | 0.0002 |
| hsm       | 7.23    | <.0001 |
The last two models (HSM only and HSM, HSE) appear to be pretty good. Notice that \( R^2 \) go down a little with the deletion of HSE, so HSE does provide a little information. This is a judgment call: do you prefer a slightly better-fitting model, or a simpler one?
Now look at SAT scores

```plaintext
proc reg data=cs;
    model gpa=satm satv;
```

R-Square 0.0634    Adj R-Sq 0.0549

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Estimate</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>1.28868</td>
<td>0.37604</td>
</tr>
<tr>
<td>satm</td>
<td>1</td>
<td>0.00228</td>
<td>0.00066291</td>
</tr>
<tr>
<td>satv</td>
<td>1</td>
<td>-0.00002456</td>
<td>0.00061847</td>
</tr>
</tbody>
</table>

| Variable    | t   | Pr > |t|   |
|-------------|-----|------|-----|
| Intercept   | 3.43| 0.0007|
| satm        | 3.44| 0.0007|
| satv        | -0.04| 0.9684|
Here we see that SATM is a significant explanatory variable, but the overall fit of the model is poor. SATV does not appear to be useful at all.

Now try our two most promising candidates: HSM and SATM

```
proc reg data=cs;
   model gpa=hsm satm;
```
| Variable | DF  | Estimate | Standard Error | t Value | Pr > |t| |
|----------|-----|----------|----------------|---------|------|---|
| Intercept| 1   | 0.66574  | 0.34349        | 1.94    | 0.053| |
| hsm      | 1   | 0.19300  | 0.03222        | 5.99    | <.000| |
| satm     | 1   | 0.00061047 | 0.00061117    | 1.00    | 0.319| |

In the presence of HSM, SATM does not appear to be at all useful. Note that adjusted $R^2$ is the same as with HSM only, and that the $p$-value for SATM is now no longer significant.
General Linear Test Approach: HS and SAT’s

proc reg data=cs;
  model gpa=satm satv hsm hss hse;
  *Do general linear test;
  * test H0: beta1 = beta2 = 0;
  sat: test satm, satv;
  * test H0: beta3=beta4=beta5=0;
  hs: test hsm, hss, hse;
The first test statement tests the full vs reduced models ($H_0$ : no satm or satv, $H_a$ : full model)
### Test sat Results for Dependent Variable gpa

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerator</td>
<td>2</td>
<td>0.46566</td>
<td>0.95</td>
<td>0.3882</td>
</tr>
<tr>
<td>Denominator</td>
<td>218</td>
<td>0.49000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We do not reject the $H_0$ that $\beta_1 = \beta_2 = 0$. Probably okay to throw out SAT scores.
The second test statement tests the full vs reduced models ($H_0 : \text{no hsm, hss, or hse, } H_a : \text{all three in the model}$)

Test hs Results for Dependent Variable gpa

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerator</td>
<td>3</td>
<td>6.68660</td>
<td>13.65</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Denominator</td>
<td>218</td>
<td>0.49000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We reject the $H_0$ that $\beta_3 = \beta_4 = \beta_5 = 0$. CANNOT throw out all high school grades.
Can use the `test` statement to test any set of coefficients equal to zero. Related to extra sums of squares (later).

**Best Model?**

Likely the one with just HSM. Could argue that HSM and HSE is marginally better. We’ll discuss comparison methods in Chapters 7 and 8.
Other things that should be considered

- *Diagnostics*: Do the models meet the assumptions?
  You might take one of our final two potential models and check these out on your own.

- Confidence intervals/bands, Prediction intervals/bands?
Example II (KNNL page 236)

Dwaine Studios, Inc. operates portrait studios in \( n = 21 \) cities of medium size.

Program used to generate output for confidence intervals for means and prediction intervals is \texttt{knnl236.sas}.

- \( Y_i \) is sales in city \( i \)
- \( X_1 \) : population aged 16 and under
- \( X_2 \) : per capita disposable income

\[
Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i
\]
Read in the data

data a1;
    infile 'H:\System\Desktop\CH06FI05.DAT';
    input young income sales;
proc print data=a1;

Obs  young  income  sales
  1  68.5    16.7    174.4
  2  45.2    16.8    164.4
  3  91.3    18.2    244.2
  4  47.8    16.3    154.6
...

```plaintext
proc reg data=a1;
model sales=young income;

<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>2</td>
<td>24015</td>
<td>12008</td>
<td>99.10</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>18</td>
<td>2180.92741</td>
<td>121.16263</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>20</td>
<td>26196</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE: 11.00739
R-Square: 0.9167
Adj R-Sq: 0.9075
Coeff Var: 6.05118

| Variable    | DF | Estimate | Error       | t Value | Pr > |t| |
|-------------|----|----------|-------------|---------|------|
| Intercept   | 1  | -68.85707| 60.01695    | -1.15   | 0.2663|
| young       | 1  | 1.45456  | 0.21178     | 6.87    | <.0001|
| income      | 1  | 9.36550  | 4.06396     | 2.30    | 0.0333|
```
**clb option:** Confidence Intervals for the $\beta$'s

```
proc reg data=a1;
   model sales=young income/clb;
```

95% Confidence Limits

<table>
<thead>
<tr>
<th></th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-194.94801</td>
<td>57.23387</td>
</tr>
<tr>
<td>young</td>
<td>1.00962</td>
<td>1.89950</td>
</tr>
<tr>
<td>income</td>
<td>0.82744</td>
<td>17.90356</td>
</tr>
</tbody>
</table>
Estimation of $E(Y_h)$

$X_h$ is now a vector of values. ($Y_h$ is still just a number.)

$\left(1, X_{h,1}, X_{h,2}, \ldots, X_{h,p-1}\right)' = X'_h$: this is row $h$ of the design matrix. We want a point estimate and a confidence interval for the subpopulation mean corresponding to the set of explanatory variables $X_h$. 

Slide 86 of 106-1
Theory for $E(Y)_h$

$$E(Y_h) = \mu_j = X'_h \beta$$

$$\hat{\mu}_h = X'_h b$$

$$s^2 \{\hat{\mu}_h\} = X'_h s^2 \{b\} X_h = s^2 X'_h (X'X)^{-1} X_h$$

95% CI: $\hat{\mu}_h \pm s\{\hat{\mu}_h\} t_{n-p}(0.975)$
**clm** option: Confidence Intervals for the Mean

```plaintext
proc reg data=a1;
   model sales=young income/clm;
   id young income;
```

<table>
<thead>
<tr>
<th>Obs</th>
<th>young</th>
<th>income</th>
<th>sales</th>
<th>Value</th>
<th>Mean</th>
<th>Predict</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>68.5</td>
<td>16.7</td>
<td>174.400</td>
<td>187.1841</td>
<td>3.8409</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>45.2</td>
<td>16.8</td>
<td>164.400</td>
<td>154.2294</td>
<td>3.5558</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>91.3</td>
<td>18.2</td>
<td>244.200</td>
<td>234.3963</td>
<td>4.5882</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>47.8</td>
<td>16.3</td>
<td>154.600</td>
<td>153.3285</td>
<td>3.2331</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs</th>
<th>95% CL Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>179.1146</td>
</tr>
<tr>
<td>2</td>
<td>146.7591</td>
</tr>
<tr>
<td>3</td>
<td>224.7569</td>
</tr>
<tr>
<td>4</td>
<td>146.5361</td>
</tr>
</tbody>
</table>
```

Latest update February 9, 2016
**Prediction of** $Y_h(new)$

Predict a new observation $Y_h$ with $X$ values equal to $X_h$.

We want a prediction of $Y_h$ based on a set of predictor values with an interval that expresses the uncertainty in our prediction. As in SLR, this interval is centered at $\hat{Y}_h$ and is wider than the interval for the mean.
Theory for $Y_h$

\[ Y_h = X_h' \beta + \epsilon \]
\[ \hat{Y}_h = \hat{\mu}_h = X_h' \beta \]
\[ s^2 \{\text{pred}\} = \text{Var}(\hat{Y}_h + \epsilon) = \text{Var}(\hat{Y}_h) + \text{Var}(\epsilon) \]
\[ = s^2 \left( 1 + X_h' (X'X)^{-1} X_h \right) \]

CI for $Y_{h\ (new)}$:
\[ \hat{Y}_h \pm s\{\text{pred}\} t_{n-p}(0.975) \]
**cli option: Confidence Interval for an Individual Observation**

```
proc reg data=a1;
   model sales=young income/cli;
   id young income;
```

<table>
<thead>
<tr>
<th>Obs</th>
<th>young</th>
<th>income</th>
<th>sales</th>
<th>Value</th>
<th>Mean</th>
<th>Predict</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>68.5</td>
<td>16.7</td>
<td>174.4000</td>
<td>187.1841</td>
<td>3.8409</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>45.2</td>
<td>16.8</td>
<td>164.4000</td>
<td>154.2294</td>
<td>3.5558</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>91.3</td>
<td>18.2</td>
<td>244.2000</td>
<td>234.3963</td>
<td>4.5882</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>47.8</td>
<td>16.3</td>
<td>154.6000</td>
<td>153.3285</td>
<td>3.2331</td>
<td></td>
</tr>
</tbody>
</table>

**95% CL Predict**

<table>
<thead>
<tr>
<th>Obs</th>
<th>95% CL Predict</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>162.6910</td>
<td>211.6772</td>
</tr>
<tr>
<td>2</td>
<td>129.9271</td>
<td>178.5317</td>
</tr>
<tr>
<td>3</td>
<td>209.3421</td>
<td>259.4506</td>
</tr>
<tr>
<td>4</td>
<td>129.2260</td>
<td>177.4311</td>
</tr>
</tbody>
</table>
Diagnostics

Look at the distribution of each variable to gain insight.

Look at the relationship between pairs of variables.

*proc corr* is useful here. BUT note that relationships between variables can be more complicated than just pairwise: **correlations are NOT the entire story.**

- Plot the residuals vs...
  - the predicted/fitted values
  - each explanatory variable
- time (if available)

- Are the relationships linear?
  - Look at $Y$ vs each $X_i$
  - May have to transform some $X$’s.

- Are the residuals approximately normal?
  - Look at a histogram
  - Normal quantile plot

- Is the variance constant?
  - Check all the residual plots.
**Remedies**

Similar remedies to simple regression (but more complicated to decide, though).

Additionally may eliminate some of the $X$’s (this is called *variable selection*).

Transformations such as Box-Cox

Analyze with/without outliers

More detail in KNNL Ch 10 and 11