Multiple Regression

- The Multiple Linear Regression Model
- Estimating the Regression Parameters
- Checking the Model Assumptions
- Inference for Regression Coefficients
- ANOVA Table for Multiple Regression
- F-tests for Multiple Regression

The Multiple Linear Regression Model

We have already discussed methods for inference when we have a single explanatory variable $x$.

When there are two or more explanatory variables that are used to explain or predict a single response variable, multiple regression is often used. Suppose we have

- a single response variable $y$
- several predictor/explanatory variables $x_1, \ldots, x_p$

Example

<table>
<thead>
<tr>
<th>Individual</th>
<th>Predictors</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i$</td>
<td>$x_1$</td>
<td>$x_2$</td>
</tr>
<tr>
<td>1</td>
<td>$x_{11}$</td>
<td>$x_{12}$</td>
</tr>
<tr>
<td>2</td>
<td>$x_{21}$</td>
<td>$x_{22}$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$n$</td>
<td>$x_{n1}$</td>
<td>$x_{n2}$</td>
</tr>
</tbody>
</table>

Following our principles of data analysis, we look first at each variable separately, then at relationships among the variables using plots and numerical descriptions.

Correlations

<table>
<thead>
<tr>
<th></th>
<th>Yield/Fertilizer</th>
<th>Yield/Rainfall</th>
<th>Fertil./Rainfall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlations</td>
<td>0.92</td>
<td>0.88</td>
<td>0.66</td>
</tr>
</tbody>
</table>

What if the correlation between Fertilizer and Rainfall are very close to 1 or -1?

Two highly correlated variables contain about the same information, so that both together may explain response little better than either alone.
The statistical model for multiple linear regression linear model gives the following relationship between $y$ and $x_1, \ldots, x_p$:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

where

- $\epsilon \sim N(0, \sigma)$ is a random variable
- The $\epsilon_i$’s corresponding to observations $(y_i; x_{i1}, x_{i2}, \cdots, x_{ip})$ on different individuals are independent of each other
- $\beta_j$ is the change in $y$ for each unit change in $x_j$ when holding all other predictors constant

Why $n - p - 1$? The degrees of freedom equal the sample size $n$ minus $p + 1$, the number of $\beta$’s that must be estimated to fit the model.

For simple linear regression, $p = 1$.

Estimating the Regression Parameters

The true population parameters $\beta_0, \beta_1, \ldots, \beta_p$ and $\sigma$ are estimated from the data by the least squares method. That is, we minimize the residual sum of squares

$$SS(\text{Residual}) = \sum_{i=1}^{n} (e_i)^2$$

$$= \sum_{i=1}^{n} (y_i - \hat{y})^2$$

$$= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \cdots - b_p x_{ip})^2$$

The method of least squared gives the values of the $b$’s that make SS(Residual) as small as possible.

The parameter $\sigma^2$ measures the variability of the responses about the population regression equation. Similar to simple linear regression, we estimate $\sigma^2$ by an average of the squared residuals. The estimator of $\sigma^2$ is

$$s^2 = \frac{\sum (e_i)^2}{n - p - 1}$$

where $n - p - 1$ is the number of degrees of freedom.

Checking the Model Assumptions

As with simple linear regression, we need to check that the model assumptions are met:

- The sample is a SRS from the population
  This can’t be checked; this needs to be taken care of when the sample is drawn.
- There is a linear relationship in the population
  Checking this isn’t as straightforward as with simple linear regression, but we should draw a plot of residuals vs. fitted values and check for any patterns.
- The standard deviation of the residuals are constant.
  Using the same plot as above, check for non-uniformity in the spread of residuals around the center line.
- The response varies Normally about the population regression line.
  Check with a Normal quantile plot of the residuals.
Inference for Regression Coefficients

Similar to simple linear regression, we can obtain confidence intervals for each of the regression coefficients $\beta_j$ for multiple linear regression.

A 95% confidence interval for $\beta_j$ is

$$ b_j \pm t^* \text{SE}(b_j) $$

where $t^*$ is the number such that 95% of the area of the $t_{n-p-1}$ distribution falls between $-t^*$ and $t^*$

For multiple regression, the standard errors of the $b$'s have some complicated formulas. Statistical software is generally used to obtain them.

To test the hypothesis

$$ H_0 : \beta_j = 0, \quad \beta_i \neq 0 \quad i \neq j $$

compute the $t$-statistic

$$ T = \frac{b_j}{\text{SE}(b_j)} $$

• the $p$-value for this test statistic is computed from the $t_{n-p-1}$ distribution
  → for $H_0 : \beta_j > 0$, $p$-value is $P(t_{n-p-1} > T)$
  → for $H_0 : \beta_j < 0$, $p$-value is $P(t_{n-p-1} < T)$
  → for $H_0 : \beta_j \neq 0$, $p$-value is $2P(t_{n-p-1} > |T|)$

• if the regression model assumptions are true, testing $H_0 : \beta_j = 0$ corresponds to testing whether or not $x_j$ is a significant predictor of $y$, assuming all the other predictors are already in the model.

Example (cont.)

<table>
<thead>
<tr>
<th>Yield</th>
<th>Fertilizer</th>
<th>Rainfall</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>200</td>
<td>10</td>
</tr>
<tr>
<td>50</td>
<td>200</td>
<td>20</td>
</tr>
<tr>
<td>50</td>
<td>300</td>
<td>10</td>
</tr>
<tr>
<td>70</td>
<td>400</td>
<td>30</td>
</tr>
<tr>
<td>65</td>
<td>500</td>
<td>20</td>
</tr>
<tr>
<td>65</td>
<td>600</td>
<td>20</td>
</tr>
<tr>
<td>80</td>
<td>700</td>
<td>30</td>
</tr>
</tbody>
</table>

. regress yield fert rainfall

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1128.57143</td>
<td>2</td>
<td>564.285714</td>
<td>Prob &gt; F = 0.0003</td>
</tr>
<tr>
<td>Residual</td>
<td>21.4285714</td>
<td>4</td>
<td>5.35714286</td>
<td>R-squared = 0.9814</td>
</tr>
<tr>
<td>Total</td>
<td>1150.00</td>
<td>6</td>
<td>191.666667</td>
<td>Root MSE = 2.3146</td>
</tr>
</tbody>
</table>

| yield | Coef. | Std. Err. | t | P>|t| |
|-------|-------|-----------|---|-----|
| fert  | .0380952 | .0058321 | 6.532 | 0.003 |
| rainfall | .8333333 | .1543033 | 5.401 | 0.006 |
| _cons | 28.09524 | 2.491482 | 11.277 | 0.000 |

<table>
<thead>
<tr>
<th>yield</th>
<th>[95% Conf. Int]</th>
</tr>
</thead>
<tbody>
<tr>
<td>fert</td>
<td>[.0219027 .0542878]</td>
</tr>
<tr>
<td>rainfall</td>
<td>[.4049186 1.261748]</td>
</tr>
<tr>
<td>_cons</td>
<td>21.17777 35.0127</td>
</tr>
</tbody>
</table>

If fertilizer a significant predictor of crop yield?
ANOVA table for multiple regression

The basic ideas of the regression ANOVA table are the same in simple and multiple regression.

ANOVA expresses variation in the form of sums of squares. It breaks the total variation into two parts: SS(Regression) and SS(Residual):

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>$\sum_{i=1}^{n}(y_i - \bar{y})^2$</td>
<td>$p$</td>
</tr>
<tr>
<td>Residual</td>
<td>$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$</td>
<td>$n - p - 1$</td>
</tr>
<tr>
<td>Total</td>
<td>$\sum_{i=1}^{n}(y_i - \hat{y})^2$</td>
<td>$n - 1$</td>
</tr>
</tbody>
</table>

The statistic

$$R^2 = \frac{SS(\text{Regression})}{SS(\text{Total})} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \hat{y})^2}$$

is the proportion of the variation of the response variable $y$ that is explained by the explanatory variables $x_1, x_2, \cdots, x_p$. $R^2$ is called the multiple correlation coefficient.

F-tests for Multiple Regression

To test the hypotheses

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$

$H_a : \exists j \in \{1, \ldots, p\}$ such that $\beta_j \neq 0$

Calculate the $F$ statistic

$$F = \frac{\text{MS(Reg.)}}{\text{MS(Res.)}} = \frac{SS(\text{Reg.})/p}{SS(\text{Res.})/(n - p - 1)}$$

Under $H_0$,

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

The null hypothesis says that none of the explanatory variables are predictors of the response variable when used in the form expressed by the multiple regression equation.

Example (cont.)

Full regression result:

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1128.57143</td>
<td>2</td>
<td>564.285714</td>
</tr>
<tr>
<td>Residual</td>
<td>21.4285714</td>
<td>4</td>
<td>5.35714286</td>
</tr>
<tr>
<td>Total</td>
<td>1150.00</td>
<td>6</td>
<td>191.666667</td>
</tr>
</tbody>
</table>

Number of obs = 7

$F(2, 4) = 105.33$

Prob > $F$ = 0.0003

R-squared = 0.9814

Adj R-squared = 0.9720

Root MSE = 2.3146