8.3 Multiple Regression

Recall that, given data \((x_1, y_1), \ldots, (x_n, y_n)\), we saw in §2.2-3 and §8.1-2 how to use linear regression to fit a line to describe how the dependent variable \(y\) changes as the independent variable \(x\) changes.

§8.3 extends this idea to the case of \(y\) depending on \(p\) independent variables \(x_1, \ldots, x_p\) via the multiple regression model

\[
y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i
\]

Notation:

• \((x_{1i}, \ldots, x_{pi}, y_i)\): \(i^{th}\) data point, for \(i = 1\) to \(n\)
• \(\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_p x_{pi}\): fitted regression equation, where \(\hat{\beta}_i\) estimates the unknown \(\beta_i\)
• \(e_i = y_i - \hat{y}_i\): residual, the difference between observed \(y_i\) and predicted \(\hat{y}_i\)

Special cases of multiple regression include

• polynomial regression, in which the \(p\) independent variables are \(\ldots\) of a single variable \(x\):
  \[
y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i
  \]

• a quadratic model in variables \(x_1\) and \(x_2\):
  \[
y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \beta_4 x_{1i}^2 + \beta_5 x_{2i}^2 + \varepsilon_i
  \]

(These are models are linear in the \(\beta_0, \ldots, \beta_p\), not in the \(\beta_i\).

Estimating the Coefficients \(\hat{\beta}_0, \ldots, \hat{\beta}_p\)

As before, minimize the error sum of squares

\[
SSE = \sum_{i=1}^{n} e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_p x_{pi})]^2
\]

Do this by setting \(\frac{\partial S}{\partial \hat{\beta}_i} = 0\) for \(i = 0\) through \(p\), to get a system of \(p\) linear equations in the \(p\) unknowns \(\hat{\beta}_0, \ldots, \hat{\beta}_p\). A page of matrix calculus gives an elegant solution,

\[
\hat{\beta}_i = \text{expression}
\]

For each estimated coefficient \(\hat{\beta}_i\), the estimated standard deviation is \(s_{\hat{\beta}_i} = \text{expression}\). Use \(\text{expression}\) to get these numbers.
The Secret Coefficients $\hat{\beta}_0, \ldots, \hat{\beta}_p$

Use the matrix notation

\[
\vec{y} = [y_1 \cdots y_j \cdots y_n]_{1 \times n}, \quad \vec{\beta} = [\hat{\beta}_0 \cdots \hat{\beta}_p]_{1 \times (p+1)}, \quad \text{and} \quad X = \begin{bmatrix}
1 & \cdots & 1 & \cdots & 1 \\
x_{11} & x_{1j} & \cdots & x_{1n} \\
\vdots & \vdots & \ddots & \vdots \\
x_{p1} & x_{pj} & \cdots & x_{pn}
\end{bmatrix}_{(p+1) \times n}
\]

for $n$ data points (_________ of $\vec{y}$ and $X$) and $p$ independent variables plus 1 constant term (_________ of $X$, where we understand $x_{0i} \equiv 1$).

The fitted system of equations, \( \hat{y}_j = \hat{\beta}_0 x_{0j} + \cdots + \hat{\beta}_p x_{pj} = \sum_{k=0}^p \hat{\beta}_k x_{kj} \) (for $j = 1$ to $n$), is

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

Minimize the sum of the squares of the residuals

\[
SSE = \sum_{j=1}^n \left( y_j - \sum_{k=0}^p \hat{\beta}_k x_{kj} \right)^2
\]

by differentiating with respect to $\hat{\beta}_i$ (for $i = 0$ to $p$):

\[
\frac{\partial}{\partial \hat{\beta}_i} (SSE) = \sum_{j=1}^n 2 \left( y_j - \sum_{k=0}^p \hat{\beta}_k x_{kj} \right) (-x_{ij}) = 0
\]

\[
\implies \sum_{j=1}^n y_j x_{ij} = \sum_{j=1}^n \sum_{k=0}^p \hat{\beta}_k x_{kj} x_{ij}
\]

\[
\implies \sum_{j=1}^n y_j X^T_{ji} = \sum_{k=0}^p \hat{\beta}_k \left( \sum_{j=1}^n x_{kj} [X^T]_{ji} \right)
\]

\[
\implies [\vec{\hat{y}} X^T]_i = \sum_{k=0}^p \hat{\beta}_k (X X^T)_{ki}
\]

\[
= [\hat{\beta}(X X^T)]_i
\]

This is true for all $i$, so we can write the matrix equation $\vec{y} X^T = \hat{\beta} (X X^T)$. To solve it, multiply both sides on the right by $(X X^T)^{-1}$:

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

e.g. Find the regression line for the points (1, 1), (2, 3) (3, 2) (draw).
Sums of Squares

Analysis of multiple regression relies on three sums of squares:

- **Regression** sum of squares, $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$: measures spread of predictions around ____________
- **Error** sum of squares, $SSE = \sum(y_i - \hat{y}_i)^2$: measures errors, spread of $y_i$'s around ____________
- **Total** sum of squares, $SST = \sum(y_i - \bar{y})^2$: measures spread of $y_i$'s around ____________

It can be shown that $SST = SSR + SSE$ (the *analysis of variance identity*).

As before, assume the errors $\varepsilon_1, \ldots, \varepsilon_n$ are ____________, all having mean ______ and the same variance ________, and are all ____________ distributed: $\varepsilon_i \sim N(0, \sigma^2)$.

Then $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{pi} + \varepsilon_i \sim N(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{pi}, \sigma^2)$

$\beta_j$ is the change in ______ caused by a change of ______ in $x_j$, with the other variables ____________.

The Statistics $s^2$, $R^2$, and $F$

- $s^2 = \frac{SSE}{n - (p + 1)}$, an estimate of $\sigma^2$
  
  Divide by $n - (p + 1)$ instead of $n$ because ____________ degrees of freedom are lost in estimating $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ from data.

  $s^2$ enters into a complicated expression for $\frac{s^2}{\hat{\beta}_j}$, which we’ll get from software. Then $\frac{\hat{\beta}_j - \beta_j}{s_{\hat{\beta}_j}} \sim t_{n-(p+1)}$,

  which we can use for inference:

  - Interval: $\begin{pmatrix} \hat{\beta}_j + t_{n-(p+1), \alpha/2}s_{\hat{\beta}_j} \end{pmatrix} (s_{\hat{\beta}_j}$ is from software)$
  
  - Test: $\frac{\hat{\beta}_j - \beta_j}{s_{\hat{\beta}_j}} \sim t_{n-(p+1)}$ tests $H_0 : \beta_j = \beta_{j0}$

- $R^2$: Recall (§2.3) that the *coefficient of determination*, $r^2$, is a measure of the ____________ of the model to the data. For multiple regression, capitalize the “r”:

  \[ R^2 = \frac{\sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = \frac{SSR - SSE}{SST} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \]

  = proportion of variance in $y$ explained by regression

  $\in [0, 1]$

- $F = \frac{SSR/p}{SSE/[n - (p + 1)]} = \frac{SSR/p}{s^2} \sim F_{p, n-(p+1)}$

  Use $F$ to test $H_0 : \beta_1 = \cdots = \beta_p = 0$ (a strong generalization of the one-variable test, "$H_0 : \beta_1 = 0"$), which says $y$ has ____________ with any of the $x_j$’s. Proceed with regression only if ____________.
Example

e.g. (p. 360 #16) An experiment studying the relationship between the speed of a cutting tool $(x, \text{ in m/s})$ and the tool lifetime $(y, \text{ in hours})$ yielded the data below. The residual plot for the $y = \hat{\beta}_0 + \hat{\beta}_1 x$ shows curvature. Use multiple regression to find the best model $y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$.

<table>
<thead>
<tr>
<th>$x$</th>
<th>1</th>
<th>1.5</th>
<th>2</th>
<th>2.5</th>
<th>3</th>
<th>\bar{y} = 85, s_y = 13.86</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>99</td>
<td>96</td>
<td>88</td>
<td>76</td>
<td>66</td>
<td></td>
</tr>
</tbody>
</table>

calculations:

<table>
<thead>
<tr>
<th>$x^2$</th>
<th>1</th>
<th>4</th>
<th>6.25</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{y}_i$</td>
<td>94.89</td>
<td>87.57</td>
<td>77.69</td>
<td>65.23</td>
</tr>
<tr>
<td>$e_i = y_i - \hat{y}_i$</td>
<td>1.11</td>
<td>.43</td>
<td>-1.69</td>
<td>.77</td>
</tr>
<tr>
<td>$e_i^2$</td>
<td>1.23</td>
<td>.18</td>
<td>2.86</td>
<td>.59</td>
</tr>
</tbody>
</table>

$SSE = \sum e_i^2 = \ldots$

The R guide gives the (polynomial regression) model $\hat{y} = 101.4 + 3.37 x - 5.14 x^2$.

a. Using this equation, find the residuals.

$\hat{y}_1 = \ldots$

$\Rightarrow e_1 = \ldots = \ldots$

b. Find the error sum of squares $SSE = \ldots$ and the total sum of squares $SST = \ldots$.

c. Find the error variance estimate $s^2 = \ldots (p = \ldots)$

d. Find the coefficient of determination $R^2 = \ldots$.

e. For $H_0: \beta_1 = \beta_2 = 0$, find $F = \ldots$.

Degrees of freedom = \ldots

f. Can $H_0$ be rejected at the 5% level? Explain.

Checking Assumptions in Multiple Regression

Check assumptions as in simple linear regression ($§8.2$):

- Make a residual plot ($§8.2$) of residuals vs. fitted values, \ldots
- Make a \ldots probability plot ($§4.7$) of the residuals, \{\epsilon_i\}
- Plot residuals \ldots in which the observations were made, \{(i, \epsilon_i)\}
- Plot residuals vs. each independent variable, \{(x_{ji}, \epsilon_i)\} for $j = 1$ to $p$

See formula sheet for an “ANOVA table” presentation of the $F$ test.