8.3 Multiple Regression

Recall that, given data $(x_1, y_1), \cdots, (x_n, y_n)$, 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 _______ independent variables $x_1, \cdots, x_p$ via the *multiple regression* model

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

Notation:

- $(x_{1i}, \cdots, x_{pi}, y_i)$: 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 β_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 _______ of a single variable x:

$$y_i = \beta_0 + \beta_1\_\_\_\_\_\_\_ + \beta_2\_\_\_\_\_\_\_ + \cdots + \beta_p\_\_\_\_\_\_\_ + \varepsilon_i$$

- 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 _____________ $\beta_0, \cdots, \beta_p$, not in the _____________.)

Estimating the Coefficients $\hat{\beta}_0, \cdots, \hat{\beta}_p$

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 _______ linear equations in the _______ unknowns $\hat{\beta}_0, \cdots, \hat{\beta}_p$. A page of matrix calculus gives an elegant solution,

$$\hat{\beta}_i = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

For each estimated coefficient $\hat{\beta}_i$, the estimated standard deviation is $s_{\hat{\beta}_i} =$ __________________________.

Use _____________ to get these numbers.

The Secret Coefficients $\hat{\beta}_0, \cdots, \hat{\beta}_p$ (_______________)

Use the matrix notation

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

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

$$\vec{\hat{y}} = \vec{\hat{\beta}} X$$

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):

$$\begin{aligned} \frac{\partial}{\partial \hat{\beta}_i}(SSE) &= \sum_{j=1}^{n} 2 \left(y_j - \sum_{k=0}^{p} \hat{\beta}_k x_{kj} \right) (-x_{ij}) = 0 \\ \implies \sum_{j=1}^{n} y_j x_{ij} &= \sum_{j=1}^{n} \sum_{k=0}^{p} \hat{\beta}_k x_{kj} x_{ij} \\ \implies \sum_{j=1}^{n} y_j X^T_{ji} &= \sum_{k=0}^{p} \hat{\beta}_k \left(\sum_{j=1}^{n} x_{kj} [X^T]_{ji} \right) \\ \implies [\vec{y} X^T]_i &= \sum_{k=0}^{p} \hat{\beta}_k (XX^T)_{ki} \\ &= [\vec{\hat{\beta}}(XX^T)]_i \end{aligned}$$

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

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

Sums of Squares

Analysis of multple 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 $\varepsilon_1, \cdots, \varepsilon_n$ are __________________________, all having mean _______ and the same variance _______, and are all _____________ distributed: $\varepsilon_i \sim N(0, \sigma^2)$.

Then $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i \sim N($____, _______$)$

βj is the change in _______ caused by a change of _______ in x_j, with the other variables _________________.

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

- $s^2 = \dfrac{\text{SSE}}{n-(p+1)}$, an estimate of σ^2

  Divide by $n-(p+1)$ instead of n because _____________ degrees of freedom are lost in estimating $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ from data.

  s^2 enters into a complicated expression for $s^2_{\hat{\beta}_j}$, 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: $\boxed{\hat{\beta}_j \pm t_{n-(p+1),\alpha/2} s_{\hat{\beta}_j}}$ ($s_{\hat{\beta}_j}$ is from software)
  - Test: $\boxed{\frac{\hat{\beta}_j - \beta_{j0}}{s_{\hat{\beta}_j}} \sim t_{n-(p+1)} \text{ tests } H_0 : \beta_j = \beta_{j0}}$

- 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":

$$R^2 = \frac{\sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = \frac{\text{SST} - \text{SSE}}{\text{SST}} = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$
$$= \text{proportion of variance in } y \text{ explained by regression}$$
$$\in [0, 1]$$

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

  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. Procede with regression only if ____________________.

Example

e.g. (p. 360 #16) An experiment studying the relationship between the speed of a cutting tool (x, in m/s) and the tool lifetime (y, 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 $y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$.

x	1	1.5	2	2.5	3	
y	99	96	88	76	66	$\bar{y} = 85, s_y = 13.86$

calculations:

x^2	1		4	6.25	9	
$\hat{y}_i$		94.89	87.57	77.69	65.23	
$e_i = y_i - \hat{y}_i$		1.11	.43	-1.69	.77	
e_i^2		1.23	.18	2.86	.59	$SSE = \sum e_i^2 =$ ______

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.

   $\hat{y}_1 =$ ______________________________

   $\implies e_1 =$ __________ $-$ __________ $=$ __________

b. Find the error sum of squares SSE $=$ __________ and the total sum of squares SST $=$ ______________________________.

c. Find the error variance estimate s^2 $=$ ______________________________ ($p =$ ______)

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

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

   Degrees of freedom $=$ ______________________________

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

See formula sheet for an "ANOVA table" presentation of the F test.