1 Simple Linear Regression

1.1 Least Squares Estimation

Given a dataset of n observations, say $(y_1, x_1), \ldots, (y_n, x_n)$, with the sample regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, the sum of squares (between the observed response y_i and the straight line) is:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Setting $\frac{\partial S}{\partial \beta_0}$ and $\frac{\partial S}{\partial \beta_1}$ to 0, obtain:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad (1); \quad \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \quad (2)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = (n-1)\text{Var}(X)$

$S_{xy} = \sum_{i=1}^{n} y_i x_i - \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n} = \sum_{i=1}^{n} y_i (x_i - \bar{x})$
$= (n-1)\text{Cov}(x, y), \qquad e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \beta_1 x_i$

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\text{Cov}(x,y)}{\text{Var}(x)} = \frac{\text{Cor}(x,y)\sqrt{\text{Var}(x)}\sqrt{\text{Var}(y)}}{\text{Var}(x)}$

1.2 Properties of Least Squares Estimators

- The OLS estimator of the slope β_1 is a linear combination of the observations y_i.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} c_i y_i$$

  where $c_i = (x_i - \bar{x}/S_{xx})$, then $\sum_{i=1}^{n} c_i x_i = 1$ and $\sum_{i=1}^{n} c_i^2 = 1/S_{xx}$.

- They are unbiased estimators of their respective parameter:

$$\mathbf{E}[\hat{\beta}_1] = \beta_1, \quad \mathbf{E}[\hat{\beta}_0] = \beta_0$$

- $\text{Var}(\hat{\beta}_1) = \text{Var}\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i^2 \text{Var}(y_i) = \frac{\sigma^2}{S_{xx}}$
- $\text{Var}(\hat{\beta}_0) = \text{Var}(\bar{y} - \hat{\beta}_1 \bar{x}) = Var(\bar{y}) + \bar{x}^2 \text{Var}(\hat{\beta}_1) - 2\bar{x}\text{Cov}(\bar{y}, \hat{\beta}_1) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$
- $\sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0, \qquad \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$
- $\sum_{i=1}^{n} x_i e_i = 0, \qquad \sum_{i=1}^{n} \hat{y}_i e_i = 0$
- The least-squares regression line always passes through the point $(\bar{x}, \bar{y})$ of the data.

1.3 Estimation of σ^2

The estimate of σ^2 is obtaineed from the residual sum of squares:

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

Since $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, we can have

$$SS_{Res} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \hat{\beta}_1 S_{xx} = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \hat{\beta}_1 S_{xx}$$

Denote $SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$, the corrected sum of squares of the response observations, then $SS_{Res} = SS_T - \hat{\beta}_1 S_{xy}$. $\mathbf{E}(SS_{Res}) = (n-2)\sigma^2$, so an unbiased estimator of σ^2 is (RSE) $\hat{\sigma}^2 = SS_{Res}/(n-2) = MS_{Res}$. SS_{Res} has $(n-2)$ degrees of freedom (due to estimate of $\hat{\beta}_0$ and $\hat{\beta}_1$).

1.4 Hypothesis Testing

- Assumed relationship between x and y is linear, errors are uncorrelated with mean 0 and constant variance σ^2
- Now we must assume ϵ_i normally distributed, iid $N(0, \sigma^2)$, so: there is a normally distributed sub-population of responses for each value of the explanatory variable, each sub-population has same variance

Test the hypothesis that the slope is a constant, β_{10}:

$$H_0 : \beta_1 = \beta_{10} \quad vs \quad H_1 : \beta_1 \neq \beta_{10}$$

$$Z_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\text{SD}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\sigma^2/S_{xx})}} \sim N(0,1)$$

Typically σ^2 is unknown, so we estimate test statistic by replacing σ^2 by its unbiased estimator $\hat{\sigma}^2 = SS_{Res}/(n-2) = MS_{Res}$. Z_0 estimated by:

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}} = \frac{\hat{\beta}_1 - \beta_{10}}{\text{SE}(\hat{\beta}_1)} \sim t_{n-2}, \quad \text{where SE}(\hat{\beta}_1) = \sqrt{\frac{MS_{Res}}{S_{xx}}}$$

Test the hypothesis that the intercept is a constant, β_{00}: Similarly,

$$t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\text{SE}(\hat{\beta}_0)} \sim t_{n-2}, \quad \text{where SE}(\hat{\beta}_0) = \sqrt{MS_{Res}(1/n + \bar{x}^2/S_{xx})}$$

Testing the significance of regression:

$$H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$$

We can either use t-test with test statistic $t_0 = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)}$, or use ANOVA.

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$SS_T = SS_R + SS_{Res}$$

SS_T has $df = n - 1$ due to the constraint $\sum_{i=1}^{n} (y_i - \bar{y}) = 0$. $SS_R = \hat{\beta}_1 \times S_{xy}$ has $df = 1$, SS_{Res} has $df = n - 2$.

Under $H_0 : \beta_1 = 0$,

$$F_0 = \frac{SS_R/1}{SS_{Res}/(n-2)} = \frac{MS_R}{MS_{Res}} \sim F_{1,n-2}$$

1.5 Interval Estimation

- If the errors are iid $N(0, \sigma^2)$ then sampling distribution of both $(\hat{\beta}_1 - \beta_1)/\text{SE}(\hat{\beta}_1)$ and $(\hat{\beta}_0 - \beta_0)/\text{SE}(\hat{\beta}_0)$ are both t_{n-2}.
- $100(1-\alpha)\%$ CI for slope, intercept, variance σ^2:
  - $\hat{\beta}_1 - t_{n-2}(\alpha/2) \times \text{SE}(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{n-2}(\alpha/2) \times \text{SE}(\hat{\beta}_1)$
  - $\hat{\beta}_0 - t_{n-2}(\alpha/2) \times \text{SE}(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t_{n-2}(\alpha/2) \times \text{SE}(\hat{\beta}_0)$
  - $\frac{(n-2)MS_{Res}}{\chi^2_{n-2}(\alpha/2)} \leq \sigma^2 \leq \frac{(n-2)MS_{Res}}{\chi^2_{n-2}(1-\alpha/2)}$

Let x_0 be the level of the regressor for which we want to estimate the mean response $\mathbf{E}(y|x_0)$. An unbiased point estimator of $\mathbf{E}(y|x_0)$ is:

$$\widehat{\mathbf{E}(y|x_0)} = \hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$\hat{\mu}_{y|x_0}$ is normally distributed with the variance:

$$\text{Var}(\hat{\mu}_{y|x_0}) = \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \text{Var}[\bar{y} + \hat{\beta}_1 (x_0 - \bar{x})]$$
$$= \frac{\sigma^2}{n} + \frac{\sigma^2 (x_0 - \bar{x})^2}{S_{xx}} = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

$$\frac{\hat{\mu}_{y|x_0} - \mathbf{E}(y|x_0)}{\sqrt{MS_{Res}[1/n + (x_0 - \bar{x})^2/S_{xx}]}} \sim t_{n-2}$$

$$\hat{\mu}_{y|x_0} - t_{n-2}(\alpha/2)\sqrt{MS_{Res}[1/n + (x_0 - \bar{x})^2/S_{xx}]} \leq \mathbf{E}(y|x_0) \leq$$
$$\hat{\mu}_{y|x_0} + t_{n-2}(\alpha/2)\sqrt{MS_{Res}[1/n + (x_0 - \bar{x})^2/S_{xx}]}$$

Interval width minimised at $x_0 = \bar{x}$ and widens as $|x_0 - \bar{x}|$ increases. The SE for a CI for mean response takes into account sampling uncertainty. The prediction interval for future observation y_0 can also be obtained. With $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ for certain x_0, consider the r.v. $\psi = y_0 - \hat{y}_0$, which is normally distributed with mean 0 and $\text{Var}(\psi) = \text{Var}(y_0 - \hat{y}_0)$. Since y_0 is independent of $\hat{y}_0$:

$$\text{Var}(\psi) = \text{Var}(y_0) - 2\text{Cov}(y_0, \hat{y}_0) + \text{Var}(\hat{y}_0)$$
$$= \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right] = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

$$\hat{y}_0 - t_{n-2}(\alpha/2)\sqrt{MS_{Res}[1 + 1/n + (x_0 - \bar{x})^2/S_{xx}]} \leq y_0 \leq$$
$$\hat{y}_0 + t_{n-2}(\alpha/2)\sqrt{MS_{Res}[1 + 1/n + (x_0 - \bar{x})^2/S_{xx}]}$$

The SE for PI of future observation takes into account sampling uncertainty, as well as variability of the individuals around the predicted mean.

1.6 Coefficient of Determination R^2

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T} = \text{Cor}(y, \hat{y})^2 = \text{Cor}(x, y)^2$$

R^2 is the proportion of variation explained by the regressor x.

1.7 No-Intercept Regression Model

Model: $y = \beta_1 x + \epsilon$; LS function: $S(\beta_1) = \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2$
$\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$; unbiased estimator of slope: $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$
Estimator of σ^2 is $\hat{\sigma}^2 = MS_{Res} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - \hat{\beta}_1 \sum_{i=1}^{n} y_i x_i}{n-1}$

1.8 Estimation by Maximum Likelihood

$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. Differentiate L wrt $\beta_0, \beta_1, \sigma^2$ and set to 0.
$L(y_i, x_i, \beta_0, \beta_1) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left[\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right]$
The estimates of $\tilde{\beta}_0$ and $\tilde{\beta}_1$ are the same as OLS method, but (biased) $\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2 = [(n-1)/n]\hat{\sigma}^2$.
In general, MLE have better statistical properties than LS estimators: they are unbiased ($\tilde{\sigma}^2$ asymptotically unbiased), have minimum variance when compared to all other unbiased estimators, are consistent, and are a set of sufficient statistics, but require full distributional assumption.

2 Multiple Linear Regression

2.1 Least Squares Estimation

Model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$; $\quad \mathbf{y} = \mathbf{X}\beta + \epsilon$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, provided $(\mathbf{X}'\mathbf{X})^{-1}$ exists (which is always the case if the regressors are linearly independent). The fitted model is:

$$\hat{y} = \mathbf{x}'\hat{\beta} = \hat{\beta_0} + \sum_{i=1}^{k} \hat{\beta_j} x_j = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

2.2 Properties of Least Squares Estimators

$\mathbf{E}(\hat{\beta}) = \beta$ (unbiased)
$\mathrm{Cov}(\hat{\beta}) = \mathrm{Var}(\hat{\beta}) = \mathrm{Var}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right]$
$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{Var}(\mathbf{y})\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]'$
$= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

Denote $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$, then $\mathrm{Var}(\hat{\beta}_j) = \sigma^2 C_{jj}$, and $\mathrm{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 C_{ij}$

2.3 Estimation of σ^2

Estimator of σ^2 (model dependent) is $\hat{\sigma}^2 = MS_{Res}$.

$$SS_{Res} = \sum_{i=1}^{n}(y_i - \hat{y_i})^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e} = (y - \mathbf{X}\hat{\beta})'(y - \mathbf{X}\hat{\beta}) = \mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{y}$$

SS_{Res} has $n-p$ df. $MS_{Res} = \frac{SS_{Res}}{n-p}$; $\quad \mathbf{E}(MS_{Res}) = \sigma^2$ (unbiased).

2.4 Hypothesis Testing

Test for significance of regression:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad vs \quad H_1 : \beta_j \neq 0, \quad \text{for at least one } j$$

$$SS_R = \hat{\beta}'\mathbf{X}'\mathbf{y} - \tfrac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2, \qquad SS_{Res} = \mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{y},$$
$$SS_T = \mathbf{y}'\mathbf{y} - \tfrac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2, \qquad SS_T = SS_R + SS_{Res}$$

$$F_0 = \frac{SS_R/k}{SS_{Res}/(n-k-1)} = \frac{MS_R}{MS_{Res}} \sim F_{k,n-k-1}$$

R^2_{Adj} penalizes for added terms to the model that are not significant: $R^2_{Adj} = 1 - \dfrac{SS_{Res}/(n-p)}{SS_T/(n-1)}$

Test significance of individual β_j (contribution of x_j given all other regressors in model): $\quad H_0 : \beta_j = 0 \quad vs \quad H_1 : \beta_j \neq 0$

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)} \sim t_{n-p} = t_{n-k-1}$$

where C_{jj} is the diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ corresponding to $\hat{\beta}_j$

Test significance of a group of variables:

Denote $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = (\underbrace{\beta_0 \quad \beta_1 \quad \cdots}_{\beta_1} \quad \underbrace{\beta_{k-r-1} \quad \cdots \quad \beta_k}_{\beta_2 \text{ with } r \text{ coefficients}})'$

Full model: $\mathbf{y} = \mathbf{X}\beta + \epsilon = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon$

$H_0 : \beta_2 = 0 \quad vs \quad H_1 : \beta_2 \neq 0$ $\quad$ Reduced model: $\mathbf{y} = \mathbf{X}_1\beta_1 + \epsilon$

LSE of β_1 in reduced model: $\hat{\beta} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$, with $SS_R(\beta_1)$

$SS_R(\beta_2|\beta_1) = SS_R(\beta) - SS_R(\beta_1)$ (extra sum-of-squares due to β_2)

$$F_0 = \frac{SS_R(\beta_2|\beta_1)/r}{MS_{Res}} \sim F_{r,n-p}$$

2.5 Interval Estimation

Assume errors ϵ_i normally distrubuted, mean 0, variance σ^2, hence: $y_i \sim N(\beta_0 + \sum_{i=1}^{k}\beta_j x_{ij}, \sigma^2)$, with $\hat{\beta} \sim N\left(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)$. Thus the marginal distribution of $\hat{\beta}_j$ is $N\left(\beta_j, \sigma^2 C_{jj}\right)$ where C_{jj} is the jth diagonal element of the matrix $(\mathbf{X}'\mathbf{X})^{-1}$.

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \sim t_{n-p}, \qquad j = 0, 1, \cdots, k$$

$100(1-\alpha)\%$ CI for β_j (where $\mathrm{SE}(\hat{\beta}_j = \sqrt{\hat{\sigma}^2 C_{jj}})$:

$$\hat{\beta}_j - t_{n-p}(\alpha/2)\sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{n-p}(\alpha/2)\sqrt{\hat{\sigma}^2 C_{jj}}$$

CI on mean response: Define $\mathbf{x}_0 = (1 \quad x_{01} \quad \cdots \quad x_{0k})'$. Fit $\hat{y_0} = \mathbf{x}_0'\hat{\beta}$. This is an unbiased estimator of $\mathbf{E}(y|\mathbf{x}_0)$, since $\mathbf{E}(\hat{y_0}) = \mathbf{x}_0'\beta = \mathbf{E}(y|\mathbf{x}_0)$.
$\mathrm{Var}(\hat{y_0}) = \sigma^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$.
$100(1-\alpha)\%$ CI on mean response at $x_{01}, x_{02}, \cdots, x_{0k}$ is:

$$\left(\hat{y_0} - t_{n-p}(\alpha/2)\sqrt{\hat{\sigma}^2 x_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}; \hat{y_0} + t_{n-p}(\alpha/2)\sqrt{\hat{\sigma}^2 x_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}\right)$$

It can be shown that $\frac{(\hat{\beta}-\beta)'\mathbf{X}'\mathbf{X}(\hat{\beta}-\beta)}{pMS_{Res}} \sim F_{p,n-p}$.
Hence $100(1-\alpha)\%$ joint CI for all parameters in β is: $\quad \dfrac{(\hat{\beta}-\beta)'\mathbf{X}'\mathbf{X}(\hat{\beta}-\beta)}{pMS_{Res}} \leq F_{p,n-p}$

Another general approach for obtaining simultaneous CI of parameters is constructing $\hat{\beta_j} \pm \Delta\mathrm{SE}(\hat{\beta_j}), \quad j = 1, \cdots, k$. In Bonferroni method, set $\Delta = t_{n-p}(\alpha/2p)$. With this method, the probability is at least $(1-\alpha)$ that all intervals are correct. For each interval, the confidence level is $(1-\alpha/p)$.

At a particular point $x_{01}, x_{02}, \cdots, x_{0k}$, $100(1-\alpha)\%$ prediction inverval for future observation is:

$$\hat{y} - t_{n-p}(\alpha/2)\sqrt{\hat{\sigma}^2\left(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\right)} \leq y_0 \leq \hat{y} + t_{n-p}(\alpha/2)\sqrt{\hat{\sigma}^2\left(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\right)}$$

2.6 Interpretation of Regression Coefficients

First interpretation: consider regressor x_j, keeping all other regressors constant, when x_j increases by 1 unit, mean response increases by β_j units. But if model has (at least) an interaction term that involves x_j then this interpretation may not be correct.

Second interpretation: the contribution of x_j to y after both y and x_j have been linearly adjusted for all other regressors. Consider $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ to interpret the effect of x_2 on y.

Step 1: Model 1 = model y on x_1 (linearly adjust y on x_1):
$\hat{y} = \hat{\alpha_0} + \hat{\alpha_1} x_1$, residual: $y - \hat{y} = e_{y.x_1}$
Step 2: Model 2 = model x_2 on x_1 (linearly adjust x_2 on x_1):
$\hat{x_2} = \hat{\gamma_0} + \hat{\gamma_1} x_1$, residual: $x_2 - \hat{x_2} = e_{x_2.x_1}$
Step 3: Model 3 = model $e_{y.x_1}$ on $e_{x_2.x_1}$ (the effect of x_2 after y and x_2 are linearly adjusted for x_1):
$\hat{e}_{y.x_1} = \hat{\lambda_0} + \hat{\lambda_1} x_1$, residual: $e_{y.x_1} - \hat{e}_{y.x_1}$
Now with Model 2, $y = \beta_0 + \beta_1 x_1 + \beta_2(\gamma_0 + \gamma_1 x_1 + e_{x_2.x_1}) = (\beta_0 + \beta_2\gamma_0) + (\beta_1 + \beta_2\gamma_1)x_1 + (\beta_2 e_{x_2.x_1} + \epsilon)$. With Model 1 and 3, we have:

$$\alpha_0 = \beta_0 + \beta_2\gamma_0, \quad \alpha_1 = \beta_1 + \beta_2\gamma_1, \quad e_{y.x_1} = \beta_2 e_{x_2.x_1} + \epsilon$$

Similarly,

$$e_{y.x_2} = \beta_1 e_{x_1.x_2} + \epsilon$$

In general, for a multiple linear regression model:

$$e_{y.x_2x_3\cdots x_k} = \beta_1 e_{x_1.x_2x_3\cdots x_k} + \epsilon$$
$$e_{y.x_1x_3\cdots x_k} = \beta_2 e_{x_2.x_1x_3\cdots x_k} + \epsilon$$
$$\cdots$$
$$e_{y.x_1x_2\cdots x_{k-1}} = \beta_k e_{x_k.x_1x_2\cdots x_{k-1}} + \epsilon$$

2.7 Hidden Extrapolation in Multiple Regression

Define the smallest convex set containing all of original n data points $(x_{i1}, x_{i2}, \cdots, x_{ik}), i = 1, \cdots, n$ as the regressor variable hull (RVH). If a point $x_{01}, x_{02}, \cdots, x_{0k}$ lies inside or on the boundary of RVH, interpolation; else extrapolation. The diagonal elements h_{ii} of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ are useful in detecting hidden extrapolation. Denote the largest h_{ii} as h_{max}. The set of points x that satisfy $\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x} \leq h_{max}$ is an ellipsoid enclosing all points inside the RVH. To check for the point $\mathbf{x}_0' = [1, x_{01}, x_{02}, \cdots, x_{0k}]$, then $h_{00} = \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$. If $h_{00} > h_{max}$, then extrapolation and vice versa.

2.8 Standardized Regression Coefficients

Assume response y with observations $y_1, \cdots, y_n$, k regressors x_j, each has n observations: $x_{ij}, i = 1, \cdots, n, j = 1, \cdots, k$. Unit normal scaling:
Regressors: $\quad z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad \bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}$
$\quad s_j^2 = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}{n-1}$
Response: $\quad y_i^* = \frac{y_i - \bar{y}}{s_y}, \quad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$
$\quad s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$
Now the model becomes $\hat{y}^* = b_1 z_1 + b_2 z_2 + \cdots + b_k z_k$. The model using scaled response and regressors has no intercept (all centered at 0).
Unit length scaling:
Regressors: $\quad w_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{S_{jj}}}, \quad S_{jj} = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$
Response: $\quad y_i^0 = \frac{y_i - \bar{y}}{\sqrt{SS_T}}$
S_{jj} is the corrected sum of squares for x_j. The regression model is now:
Now each w_j has mean $\bar{w}_j = 0$ and length $\sqrt{\sum_{i=1}^{n}(w_{ij} - \bar{w}_j)^2} = 1$ $\quad \hat{y^0} = b_1 w_1 + b_2 w_2 + \cdots + b_k w_k$
$\mathbf{Z}'\mathbf{Z} = (n-1)\mathbf{W}'\mathbf{W}$ so estimates of regression coefficients from these two scaling methods are identical.

2.9 Indicator Variables

Levels in categorical variables induce changes in intercept (slope unchanged and identical). The slope will also change if there are any interaction terms. For categorical variables with a levels, we would need $a-1$ indicator variables. Assume a fitted model $\hat{y} = \hat{\beta_0} + \hat{\beta_1} x_1 + \hat{\beta_2} x_2$, with x_2 as an indicator variable. The $(100-\alpha)\%$ CI on β_2 is: $\hat{\beta_2} \pm t_{n-p}(\alpha/2)\mathrm{SE}(\hat{\beta_2})$.

Suppose we add interaction term $\beta_3 x_1 x_2$ to our model. Test for the significance of the interaction term: $H_0 : \beta_3 = 0 \quad vs \quad H_1 : \beta_3 \neq 0$. This can be done by t-test or ANOVA:

$$F_0 = \frac{SS_R(\beta_3|\beta_2, \beta_1, \beta_0)}{MS_{Res}} \sim F_{1,n-p}$$

Test if the 2 regression lines for Type A and Type B are identical (or test the significance of the variable x_2):
$H_0 : \beta_2 = \beta_3 = 0 \quad vs \quad H_1 : \beta_2 \quad \text{and/or} \quad \beta_3 \neq 0$

$$F_0 = \frac{SS_R(\beta_2, \beta_3|\beta_1, \beta_0)/2}{MS_{Res}} \sim F_{2,n-p}$$

where $SS_R(\beta_2, \beta_3|\beta_1, \beta_0) = SS_R(\beta_2|\beta_1, \beta_0) + SS_R(\beta_3|\beta_2, \beta_1, \beta_0)$ or $SS_R(\beta_2, \beta_3|\beta_1, \beta_0) = SS_R(\beta_3, \beta_2, \beta_1|\beta_0) - SS_R(\beta_1|\beta_0)$.

3 Checking for Model Adequacy

- Linearity assumption: relationship between the response y and the regressors is linear (at least approximately). We check the scatter plot of y vs x, but linearity in the multiple model is more difficult due to dimensionality of the data.
- Assume errors $\epsilon_1, \epsilon_2, \cdots, \epsilon_n \overset{\text{iid}}{\sim} \text{N}(0,\sigma^2)$. This implies the normality, constant variance, and independent errors assumptions.

Plotting one-dimensional graphs (e.g. histogram, stem-and-leaf, scatter plot, boxplot) indicate the distribution (symmetric or skewed), gives an idea as to whether we should work with the original or transformed variables. It can also point out presence of outliers in the variables. Two-dimensional graphs present scatter plots of pairwise variables to observe correlation between variables.

3.1 Residual Analysis

The residuals $e_i = y_i - \hat{y}_i$ have zero mean and their approximate variance can be estimated by

$$\frac{\sum_{i=1}^n (e_i - \bar{e})^2}{n-p} = \frac{\sum_{i=1}^n e_i^2}{n-p} = \frac{SS_{Res}}{n-p} = MS_{Res}$$

The residuals are not independent (because of assumption $\sum_{i=1}^n e_i = 0$), but when $p << n$ the nonindependence has little effect on their use for model adequacy checking.

We defined hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and have $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$. Then $\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \cdots + h_{in}y_n$ for $i = 1, \cdots, n$ which means $\hat{y}_i$ is a weighted sum of all the given observations. $h_i i$ is the leverage value for the ith observation, the weight given to y_i in determining the ith fitted value $\hat{y}_i$. The residual then can be written as $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$. Substituting $\mathbf{y} = \mathbf{x}\beta + \epsilon$, we have $\mathbf{e} = (\mathbf{I} - \mathbf{H})(\mathbf{X}\beta + \epsilon) = \mathbf{X}\beta - \mathbf{H}\mathbf{X}\beta + (\mathbf{I} - \mathbf{H})\epsilon = \mathbf{X}\beta - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{I} - \mathbf{H})\epsilon = (\mathbf{I} - \mathbf{H})\epsilon$. Furthermore $\text{Var}(\epsilon) = \sigma^2\mathbf{I}$ and $\mathbf{I} - \mathbf{H}$ is symmetric and idempotent, so $\text{Var}(\epsilon) = (\mathbf{I} - \mathbf{H})\text{Var}(\epsilon)(\mathbf{I} - \mathbf{H})' = \sigma^2(\mathbf{I} - \mathbf{H})$. Thus $\text{Var}(e_i) = \sigma^2(1 - h_{ii})$, $\text{Cov}(e_i, e_j) = -\sigma^2 h_{ij}$.

Standardized residuals is defined as $\frac{e_i}{\sigma\sqrt{(1-h_{ii})}}$

$$\hat{\sigma}^2_{(i)} = \frac{SS_{Res}}{n-k=2} = \frac{SS_{Res(i)}}{n-p-1}$$

and $SS_{Res}(i)$ is the sum of squared residuals when we fit the model to the $(n-1)$ observations by omitting the ith observation.

Since σ is unknown, estimate it by $\sqrt{MS_{Res}}$ or $\hat{\sigma}_{(i)}$, where

Both MS_{Res} and $\hat{\sigma}^2_{(i)}$ are unbiased estimator of σ^2.

Internally studentized residuals: Substitute $\sqrt{MS_{Res}}$ into the standardized residuals: $r_i = \frac{e_i}{\sqrt{MS_{Res}}\sqrt{(1-h_{ii})}}$ (`rstandard()` in R)

Externally studentized residuals: Substitute $\hat{\sigma}_{(i)}$: $r_i^* = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{(1-h_{ii})}}$

The two forms of the residuals are related by $r_i^* = r_i\sqrt{\frac{n-p-1}{n-p-r_i^2}}$.

Normal probability plot: ordered S.R. (x-axis) vs cumulative probability or normal scores (y-axis). Normal scores are what we expect to obtain when we take a sample of size n from $N(0,1)$. If the residuals are normally distributed, ordered residuals should be the same as ordered scores. Expect a (nearly) straight line with intercept 0 and slope 1.

Scatter plot of S.R. vs fitted values: expect points to be scattered randomly. Funnel/double-bow shape indicate nonconstant variance; apply transformation to regressor/response, or use WLS. Curved shape indicated nonlinearity; transform or add more variables.

3.2 Detection and Treatment of Outliers

An outlier is an extreme observation considerably different from the majority of the data. Residuals considerably larger in absolute value than the others (say 3 or 4 sd from the mean) indicate potential y-space outliers.

3.3 Lack of Fit of the Regression Model

Lack of fit test (more useful for simple model): Suppose x has m levels $x_1, \cdots, x_m$, with n_i observations on the response at the ith level. Let $y_{i,j}$ denote the jth observation on the response at x_i. There are $n = \sum_{i=1}^m n_i$ total observations. The ijth residual is $y_{ij} - \hat{y}_i = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \hat{y}_i)$, where $\bar{y}_i$ is the average of the n_i observations at x_i. Squaring both sides and summing over i and j, we have

$$\sum_{i=1}^m\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_i)^2 = \sum_{i=1}^m\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^m n_i(\bar{y}_i - \hat{y}_i)^2$$

$$SS_{Res} = SS_{PE} + SS_{LOF}$$

where SS_{PE} is the sum of squares due to pure error and SS_{LOF} is the sum of squares due to lack of fit. SS_{LOF} is a weighted sum of squared deviations between mean response $\bar{y}_i$ at each x level and the corresponding fitted values. If $\hat{y}_i$ are close to $\bar{y}_i$, there is a strong indication that the regression function is linear. The test statistic for lack of fit is

$$F_0 = \frac{SS_{LOF}/(m-2)}{SS_{PE}/(n-m)} = \frac{MS_{LOF}}{MS_{PE}} \sim F_{m-2,n-m}$$

For simple model, use $H_0 : \beta_1 \neq 0 \equiv$ "model is linear". Reject this H_0 if $F_0 > F_{m-2,n-m}(\alpha)$ and conclude that regression model is not linear.

3.4 Leverage and Influence

The hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ determines the variances and covariances of $\hat{\mathbf{y}}$ and $\mathbf{e}$ since $\text{Var}(\hat{\mathbf{y}}) = \sigma^2\mathbf{H}$ and $\text{Var}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$. The elements h_{ij} from $\mathbf{H}$ can be interpreted as the amount of leverage exerted by the ith observation y_i on the jth fitted value $\hat{y_j}$.

Diagonal elements of $\mathbf{H}$, $h_{ii} = \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'$ (where $\mathbf{x}_i'$ is the ith row of the $\mathbf{X}$ matrix) is a standardized measure of the distance of the ith observation from the center of the x space. Large h_{ii} reveals observations that are potentially influential.

Since $\sum h_{ii} = \text{rank}(\mathbf{H}) = \text{rank}(X) = p$, the average size of a hat diagonal is $\bar{h} = p/n$. Hence we traditionally assume that any observation for which the hat diagonal exceeds $2p/n$ is remote enough from the rest of the data to be considered a leverage point (only applies to large sample size where $2p/n < 1$). Not all leverage points are influential points, but observations with large hat diagonals and large residuals are likely to be influential.

Cook's Distance: squared distance between least squares estimate based on all n data points $\hat{\beta}$ and the estimate by deleting the ith point, $\hat{\beta}_{(i)}$:

$$D_i = (\mathbf{M}, c) = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'\mathbf{M}(\hat{\beta}_{(i)} - \hat{\beta})}{c}, \quad i = 1, \cdots, n$$

where usually $\mathbf{M} = \mathbf{X}'\mathbf{X}$ and $c = pMS_{Res}$. Magnitude of D_i is usually assessed by comparing it to $F_{p,n-p}(\alpha)$. If $D_i = F_{p,n-p}(0.5)$, deleting point i would move $\beta_{(i)}$ to the boundary of an approximate 50% confidence region for β based on the complete dataset. Since $F_{p,n-p}(0.5) \simeq 1$, consider points with $D_i > 1$ to be influential. D_i may be rewritten as

$$D_i = \frac{r_i^2}{p}\frac{h_{ii}}{1 - h_{ii}}$$

4 Correcting Model Inadequacies

4.1 Variance-Stabilizing Transformations

Useful guidelines:

Relationship of σ^2 to $\mathbf{E}(y)$	Transformation
$\sigma^2 \propto$ constant	$y' = y$ (no transformation)
$\sigma^2 \propto \mathbf{E}(y)$	$y' = \sqrt{y}$ (Poisson data)
$\sigma^2 \propto \mathbf{E}(y)[1 - \mathbf{E}(y)]$	$y' = \sin^{-1}(\sqrt{y})$
$\sigma^2 \propto [\mathbf{E}(y)]^2$	$y' = \ln(y)$
$\sigma^2 \propto [\mathbf{E}(y)]^3$	$y' = y^{-1/2}$
$\sigma^2 \propto [\mathbf{E}(y)]^4$	$y' = y^{-1}$

4.2 Transformations to Linearize the Model

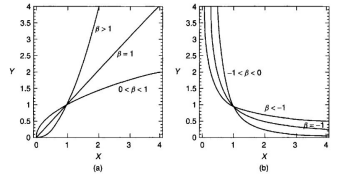

Figure 6.1 Graphs of the linearizable function $Y = \alpha X^{\beta}$.

Figure 6.2 Graphs of the linearizable function $Y = \alpha e^{\beta X}$.

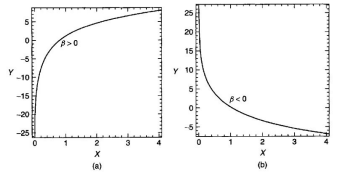

Figure 6.3 Graphs of the linearizable function $Y = \alpha + \beta \log X$.

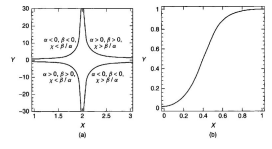

Figure 6.4 Graphs of the linearizable functions: (a) $Y = X/(\alpha X - \beta)$ and (b) $Y = (e^{\alpha+\beta X})/(1 + e^{\alpha+\beta X})$.

Figure	Function	Transformation	Linear Form
6.1	$Y = \alpha X^{\beta}$	$Y' = \log Y, X' = \log X$	$Y' = \log\alpha + \beta X$
6.2	$Y = \alpha e^{\beta X}$	$Y' = \ln Y$	$Y' = \ln\alpha + \beta X$
6.3	$Y = \alpha + \beta \log X$	$X' = \log X$	$Y = \alpha + \beta X'$
6.4(a)	$Y = \frac{X}{\alpha X + \beta}$	$Y' = \frac{1}{Y}, X' = \frac{1}{x}$	$Y' = \alpha - \beta X'$
6.4(b)	$Y = \frac{\exp(\alpha+\beta X)}{1+\exp(\alpha+\beta X)}$	$Y' = \ln\frac{Y}{1-Y}$	$Y' = \alpha + \beta X$

4.3 Analytical Methods

Box-Cox power transformation y^{λ} to correct nonnormality and/or nonconstant variance:

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \log y, & \lambda = 0 \end{cases}$$

to fit the model $y^{(\lambda)} = \mathbf{X}\beta + \epsilon$. Simpler λ for interpretability.

Box-Tidwell regressor transformation (for simple model):

$$\xi = \begin{cases} x^{\alpha}, & \alpha \neq 0 \\ \log x, & \alpha = 0 \end{cases}$$

1. Initially fit a model by least squares, $\hat{y} = \hat{\beta_0} + \hat{\beta_1}x$
2. Fit a new model, adding $w = x\log x$: $\hat{y} = \hat{\beta_0}' + \hat{\beta_1}'x + \hat{\gamma}w$
3. Take $\alpha_1 = \frac{\hat{\gamma}}{\hat{\beta_1}} + 1$

This procedure can be repeated using a new regressor $x' = x^{\alpha_1}$ (fit $y \sim x^{\alpha_1}$), and converges quite rapidly.

4.4 Weighted Least Squares

Linear regression models with nonconstant variance can also be fitted by WLS. The deviation between observed y_i and expected $\hat{y_i}$ is multiplied by weight w_i chosen inversely proportional to the variance of y_i. In the simple model, we minimize weighted sum of squares $S(\beta_0, \beta_1) = \sum_{i=1}^n w_i(y_i - \beta_0 - \beta_1 x_i)^2$.

$$\hat{\beta_0}\sum_{i=1}^n w_i + \hat{\beta_1}\sum_{i=1}^n w_i x_i = \sum_{i=1}^n w_i y_i$$
$$\hat{\beta_0}\sum_{i=1}^n w_i x_i + \hat{\beta_1}\sum_{i=1}^n w_i x_i^2 = \sum_{i=1}^n w_i y_i x_i$$

Solving the above will give WLS estimates of β_0 and β_1. Idea: if each error has variance σ_i^2, choose the weight $w_i = 1/\sigma_i^2$ so that variances will be (approximately) equal; points with low variance will be given larger weights and vice versa. The WLS estimates are:

$$\hat{\beta_0} = \bar{y}_w - \hat{\beta_1}\bar{x}_w, \qquad \hat{\beta_1} = \frac{\sum_{i=1}^n w_i(x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sum_{i=1}^n w_i(x_i - \bar{x}_w)^2}$$

where $\bar{x}_w$ and $\bar{y}_w$ are weighted means: $\bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}, \bar{y}_w = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i}$.
WLS estimators are still unbiased, and weighted mean squared residuals $MS_{(w)Res}$ is also an unbiased estimator of σ^2.
One way to estimate the weights is to use the multiple repeated (or nearly repeated) values of the regressor. For each cluster of x values, obtain sample mean $\bar{x}$ and sample variance s_y^2. Consider s_y^2 as response and $\bar{x}$ as regressor, find a least squares fit, then substituting each x_i value into the LS equation will give an estimate of the variance of the corresponding y_i. The weight w_i is the inverse of this estimated variance.

4.5 Generalized Least Squares

Consider $\mathbf{y} = \mathbf{X}\beta + \epsilon$, where $\mathbf{E}(\epsilon) = 0, \mathrm{Var}(\epsilon) = \sigma^2\mathbf{V}$. Assumptions made for errors correspond to $\mathbf{V} = \mathbf{I}$. The least squares normal equation is $(\mathbf{X'V^{-1}X})\hat{\beta} = \mathbf{X'V^{-1}y}$ with solution $\hat{\beta} = (\mathbf{X'V^{-1}X})^{-1}\mathbf{X'V^{-1}y}$, the GLS estimator of β.
WLS for GLS: when errors ϵ are uncorrelated but have equal variances, the covariance matrix of ϵ is of the form

$$\sigma^2\mathbf{V} = \sigma^2\begin{bmatrix} \frac{1}{w_1} & & & \\ & \frac{1}{w_2} & & \\ & & \ddots & \\ & & & \frac{1}{w_n} \end{bmatrix}$$

Let $\mathbf{W} = \mathbf{V}^{-1}$, then $\mathbf{W}$ is a diagonal matrix with diagonal elements or weights $w_1, w_2, \cdots, w_n$. The WLS equations are $(\mathbf{X'WX})\hat{\beta} = \mathbf{X'Wy}$ with WLS estimator $\hat{\beta} = (\mathbf{X'WX})^{-1}\mathbf{X'Wy}$.

5 Multicollinearity

Sources of MC: the data collection method employed, constraints in the model or population, model specification, or an over-defined model.

5.1 Effects of Multicollinearity

Define matrix $\mathbf{C} = (\mathbf{X'X})^{-1}$. $C_{jj} = \frac{1}{1-R_j^2}$ where R_j^2 is the multiple coefficient of determination from regression of x_j on the remaining $k-1$ regressors. If there is strong MC between x_j and any subset of the other $k-1$ regressors, R_j^2 will be large, thus C_{jj} will be large.

Since $\mathrm{Var}(\hat{\beta_j}) = \sigma^2 C_{jj} = \sigma^2(1 - R_j^2)^{-1}$ for $j = 1, \cdots, k$, strong MC implies that the $\mathrm{Var}(\hat{\beta_j})$ is very large. Generally, $\mathrm{Cov}(\hat{\beta_i}, \hat{\beta_j})$ will also be large if x_i and x_j have MC relationship.

MC also tends to produce $\hat{\beta_j}$ that are too large in absolute value. Squared distance between LS estimate and true paramter is denoted $L_1^2 = (\hat{\beta} - \beta)'(\hat{\beta} - \beta)$, then $\mathbf{E}(L_1^2) = \mathbf{E}[(\hat{\beta} - \beta)'(\hat{\beta} - \beta)] = \sum_{j=1}^k \mathbf{E}(\hat{\beta_j} - \beta_j)^2 = \sum_{j=1}^k \mathrm{Var}(\hat{\beta_j}) = \sigma^2\mathrm{Tr}(\mathbf{X'X})$ where the trace of a matrix is the sum of the main diagonal elements (sum of the eigenvalues). When MC is present, some of the eigenvalues of $\mathbf{X'X}$ will be small. Let λ_j denote the jth eigenvalue of $\mathbf{X'X}$, then $\mathbf{E}(L_1^2) = \sigma^2\sum_{j=1}^k \frac{1}{\lambda_j}$, so L_1^2 may be large.

5.2 Detecting Multicollinearity

Check off-diagonal elements r_{ij} in $\mathbf{X'X}$. If $|r_{ij}| \approx 1$, **may** indicate MC. If regressors are scaled, $\mathrm{Cor}(\mathbf{X}) = \mathbf{X'X}$.

$$VIF_j = C_{jj} = \frac{1}{1 - R_j^2}$$

One or more large values of VIF (> 10 here) indicate multicollinearity. The width of the CI of β_j is $L_j = 2(C_{jj}\sigma^2)^{1/2} \times t_{n-k-1}$, and the width of the corresponding interval based on orthogonal reference design with the same sample size and root-mean-square (rms) values (rms $= \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2/n$, which is a measure of the spread of regressor x_j) is $L^* = 2\sigma t_{n-k-1}$. The ratio of these two widths for x_j is $L_j/L^* = \sqrt{C_{jj}}$. Thus, the square root of the jth VIF indicate how much larger the CI for the jth regression coefficient is because of MC.

The eigenvalues of $k \times k$ matrix $\mathbf{A}$ are all the k roots of the equation $|\mathbf{A} - \lambda\mathbf{I}| = 0$. The eigenvalues of $\mathbf{X'X}$: $\lambda_1, \cdots, \lambda_k$ can be used to measure the extent of MC in the data. Small eigenvalues indicate MC.
Define condition number of $\mathbf{X'X}$ as $\kappa = \frac{\lambda_{max}}{\lambda_{min}}$, which measures the spread in the eigenvalues. $\kappa < 100$: no serious problem; $100 < \kappa < 1000$: moderate to strong MC; $\kappa > 1000$: strong MC. Define condition indices $\kappa_j = \frac{\lambda_{max}}{\lambda_j}$, for $j = 1, \cdots, k$. The number of large condition indices (> 1000) indicate the number of near-linear dependencies in $\mathbf{X'X}$.

5.3 Dealing with Multicollinearity

Collect more data or respecify the model, or use ridge regression (find an estimate that is biased but has smaller variance than unbiased estimator). We want to find a biased estimator $\hat{\beta}^*$ such that $MSE(\hat{\beta}^*) < MSE(\hat{\beta})$. The ridge estimator $\hat{\beta}_R$ is defined as the solution to $(\mathbf{X'X} + k\mathbf{I})\hat{\beta}_R = \mathbf{X'y}$. $\hat{\beta}_R = (\mathbf{X'X} + k\mathbf{I})^{-1}\mathbf{X'y} = (\mathbf{X'X} + k\mathbf{I})^{-1}\mathbf{X'X}\hat{\beta} = \mathbf{Z}_k\hat{\beta}$ where k is to be determined.

$$\begin{aligned} MSE(\hat{\beta}_R) &= \mathrm{Var}(\hat{\beta}_R) + (\text{bias in } \hat{\beta}_R)^2 \\ &= \sigma^2\sum_{j=1}^k \frac{\lambda_j}{(\lambda_j + k)^2} + k^2\beta'(\mathbf{X'X} + k\mathbf{I})^{-2}\beta \end{aligned}$$

where λ_j are the eigenvalues of $\mathbf{X'X}$. As k increases, variance decreases and bias increases.

$$\begin{aligned} SS_{Res} &= (y - \mathbf{X}\hat{\beta}_R)'(y - \mathbf{X}\hat{\beta}_R) \\ &= \underbrace{(y - \mathbf{X}\hat{\beta})'(y - \mathbf{X}\hat{\beta})}_{SS_{Res}\text{ from OLS}} + (\hat{\beta}_R - \hat{\beta})'\mathbf{X'X}(\hat{\beta}_R - \hat{\beta}) \end{aligned}$$

When k increases, SS_{Res} of $\hat{\beta}_R$ increases, R^2 decreases. k can be chosen by inspection of the ridge trace; select a small k at which ridge estimates $\hat{\beta}_R$ are stable.

6 Variable Selection

Deleting variables improves the precision of (1) the parameter estimates of retained variables, (2) the variance of predicted response, but it can introduce bias in them unless the deleted variables are "insignificant". Retaining insignificant variables can increase the variances of estimates and predicted response.

6.1 Criteria for Evaluating Subset Models

A regression model with p regressors has $MS_{Res}(p) = \frac{SS_{Res}(p)}{n-p}$. In general, $MS_{Res}(p)$ increases as p increases. The increase in $MS_{Res}(p)$ occurs when the reduction in $SS_{Res}(p)$ from adding a regressor to the model is not sufficient to compensate for the loss of one degree of freedom. We want a model with minimum $MS_{Res}(p)$ as it equivalently maximises $R_{Adj,p}$.

AIC is based on maximizing the expected entropy of the model. Let L be the likelihood function for a specific model, then $AIC = -2\log(L) + 2p$ where $p = k + 1$. In the case of OLS, $AIC = n\log(\frac{SS_{Res}}{n}) + 2p$. BIC is the Bayesian extension of AIC. $BIC_{Sch} = -2\log(L) + p\log(n)$. For OLS, $BIC_{Sch} = n\log(\frac{SS_{Res}}{n}) + p\log(n)$. Among models, the one with smaller AIC/BIC preferred.

6.2 Stepwise Regression

Forward selection: assume no regressor in the model, variables added one at a time. First regressor to be added is the one with the highest correlation with response. If the F statistic corresponding to the model containing this variable is significant (larger than some pre-selected value F_{in}), then that regressor is entered. The second regressor chosen for entry is the one that now has the largest correlation with y after adjusting for the effect of the first regressor entered on y. This correlation is referred to as partial correlation.

1. Derive fitted values and residuals from Model 1: $\hat{y} = \hat{\beta_0} + \hat{\beta_1}x_1$
2. Fit regression models: $\hat{x}_j = \hat{\alpha}_{0j} + \hat{\alpha}_{1j}x_1$, for $j = 1, \cdots, k$
3. Derive simple correlation between residuals of Model 1 and residuals from $k-1$ models above.
4. The x_j that gives the highest correlation will be the next regressor to enter the model

Suppose at step 2, x_2 has the highest partial correlation with y. This implies that the largest partial F statistic is $F = \frac{SS_R(x_2|x_1)}{MS_{Res}(x_1|x_2)}$. If this F value exceeds F_{in}, x_2 is added in. Procedure terminates when partial F test at a particular step does not exceed F_{in} or when last regressor is added.
Backward elimination: all variables are in the model originally, examined one at at time, removed if insignificant. Starting model has k regressors, partial F statistic are computed for each regressor as if it was the last variable to enter. Regressor with smallest F statistic is examined first and will be removed if this value is less than F_{out}. Fit a new model with rest of $(k-1)$ regressors and calculate partial F statistics again. The process continues until are regressors are examined.
Stepwise: a modification of forward selection. At each step, all regressors are reassessed via their F (or t) statistic. If the partial F (or t) statistic for a variable is less than F_{out} (or t_{out}) then the variable is dropped. Stepwise regression requires two cutoff values, one for entering and one for removing variables. Usually $F_{in} > F_{out}$.