ST3131 Regression Analysis
AY23/24 S1

by Isaac Lai

Statistics fundamentals

Let $Y_1, \ldots, Y_n$ be a random sample from normal distribution with mean μ and variance σ^2. Then

- $\bar{Y}$ is normally distributed with mean μ and variance $\frac{\sigma^2}{n}$
- $Z_i = \frac{Y_i - \mu}{\sigma}$ are independent standard normal random variables and $\sum_{i=1}^n Z_i^2$ has a χ^2 distribution with n degrees of freedom
- $\frac{(n-1)S^2}{\sigma} = \frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \bar{Y})^2$ has a χ^2 distribution with $n-1$ degrees of freedom. Also, $\bar{Y}$ and S^2 are independent random variables. Here $S^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$ is the sample variance

Let Z be a standard normal random variable and W be a χ^2-distributed variable with ν df. Then if Z and W are independent, $T = \frac{Z}{\sqrt{W/\nu}}$ has a t distribution with ν df.

Let W_1, W_2 be independent χ^2-distributed random variables with nu_1 and nu_2 df respectively. Then $F = \frac{W_1/\nu_1}{W_2/\nu_2}$ has an F distribution with ν_1 numerator df and ν_2 denominator df.

Simple Linear Regression

The **simple linear regression model** for response variable y and regressor variable x based on observations $(x_1, y_1), \ldots, (x_n, y_n)$ is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

where ϵ_i is a random variable such that $E(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$ and ϵ_i's are independent.

Least squares

Function

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

Normal equations

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

Estimators $\hat{\beta}_0, \hat{\beta}_1$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$
$$S_{xx} = \sum_{i=1}^n x_i^2 - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2 = \sum_{i=1}^n (x_i - \bar{x})^2$$
$$S_{xy} = \sum_{i=1}^n y_i x_u - \frac{1}{n}\left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n x_i\right) = \sum_{i=1}^n y_i(x_i - \bar{x})$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

i-th residual $e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ **Properties of least-squares estimators**

- $\hat{\beta}_0, \hat{\beta}_1$ are **linear combinations** of y_i

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

- $\hat{\beta}_0, \hat{\beta}_1$ are **unbiased estimators** of β_0, β_1.

$$\begin{aligned} E(\hat{\beta}_1) &= E\left(\sum_{i=1}^n c_i y_i\right) = \sum_{i=1}^n c_i E(y_i) \\ &= \sum_{i=1}^n c_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^n c_i + \beta_1 \sum_{i=1}^n c_i x_i \\ &= \beta_1 \text{ since } \sum_{i=1}^n c_i = 0 \text{ and } \sum_{i=1}^n c_i x_i = 1 \end{aligned}$$

$$E(\hat{\beta}_0) = \beta_0$$

$$\begin{aligned} \mathrm{Var}(\hat{\beta}_1) &= \mathrm{Var}\left(\sum_{i=1}^n c_i y_i\right) = \sum_{i=1}^n c_i^2 \mathrm{Var}(y_i) = \sigma^2 \sum_{i=1}^n c_i^2 \\ &= \frac{\sigma^2}{S_{xx}^2}\sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \frac{\sigma^2}{S_{xx}} \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(\hat{\beta}_0) &= \mathrm{Var}(\bar{y} - \hat{\beta}_1 \bar{x}) \\ &= \mathrm{Var}(\bar{y}) + \bar{x}^2 \mathrm{Var}(\hat{\beta}_1) - 2\bar{x}\mathrm{Cov}(\bar{y}, \hat{\beta}_1) \\ &= \mathrm{Var}(\bar{y}) + \bar{x}^2 \mathrm{Var}(\hat{\beta}_1) \\ &= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) \end{aligned}$$

Gauss-Markov Theorem: $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased and have minimum variance when compared with all other unbiased estimators that are linear combinations of the y_i. Thus least-squares estimators are the **best linear unbiased estimators**.

Other useful properties

- $\sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n e_i = 0$
- $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$
- The least-squares regression line always passes through the centroid $(\bar{x}, \bar{y})$
- $\sum_{i=1}^n x_i e_i = 0$
- $\sum_{i=1}^n \hat{y}_i e_i = 0$

Estimation of σ^2

- Corrected sum of squares or total variation in y

$$SS_T \equiv \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2.$$

- Residual sum of squares

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

- Sum of squares due to regression

$$SS_R = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1 S_{xy}$$
$$SS_{Res} = SS_T - \hat{\beta}_1 S_{xy}$$
$$SS_T = SS_{Res} + SS_R$$

- $E(SS_{Res}) = (n-2)\sigma^2$

Unbiased estimator of σ^2

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = MS_{Res}.$$

MS_{Res} is the **residual mean square**. The square root of $\hat{\sigma}^2$ is the **standard error of regression**. $\hat{\sigma}^2$ is a **model-dependent** estimate of σ^2.

Hypothesis testing and confidence interval of the slope and intercept

- Assume the ϵ_i's in a simple linear regression model are normally distributed. Suppose we want to test $H_0 : \beta_1 = \beta_{10}$ versus $H_1 : \beta_1 \neq \beta_{10}$. If H_0 is true, $T = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}}$
- Reject H_0 if

$$\frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}} < -t_{\alpha/2,n-2} \text{ or } \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}} > t_{\alpha/2,n-2}$$

  where α is the level of significance.
- A $100(1-\alpha)\%$ confidence interval for β_1 is given as

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{MS_{Res}/S_{xx}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{MS_{Res}/S_{xx}}$$

- R commands: `summary(lm(y~x)`, `confint(lm(y~x), level=0.95)`
- Suppose we want to test $H_0 : \beta_0 = \beta_{00}$ versus $H_1 : \beta_0 \neq \beta_{00}$. If H_0 is true, $T = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{Res}(1/n + \bar{x}^2/S_{xx})}}$ follows the t distribution with $n-2$ df.
- Reject h_0 if

$$\frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{Res}(1/n + \bar{x}^2/S_{xx})}} < -t_{\alpha/2,n-2} \text{ or}$$
$$\frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{Res}(1/n + \bar{x}^2/S_{xx})}} > t_{\alpha/2,n-2}$$

- A $100(1-\alpha)\%$ confidence interval for β_0 is given as

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{MS_{Res}(1/n + \bar{x}^2/S_{xx}} \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{MS_{Res}(1/n + \bar{x}^2/S_{xx})}$$

Analysis of Variance

- To test $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 \neq 0$, use $F = \frac{SS_R/1}{SS_{Res}/(n-2)} = \frac{MS_R}{MS_{Res}}$. F follows the F distribution with df 1 and $n-2$ when H_0 is true. For a given level of significance α, reject H_0 if $F > F_{\alpha,1,n-2}$.
- R commands: `summary.aov(lm(y~x))`, `anova(lm(y~x))`
- Equivalence between t and F tests

$$T = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{MS_{Res}/S_{xx}}}$$

$$T^2 = \frac{\hat{\beta}_1^2 S_{xx}}{MS_{Res}} = \frac{\hat{\beta}_1 S_{xy}}{MS_{Res}} = \frac{MS_R}{MS_{Res}} = F$$

ANOVA table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Regression	SS_R	1	MS_R	MS_R/MS_{Res}
Residual	SS_{Res}	$n-2$	MS_{Res}	
Total	SS_T	$n-1$		

Hypothesis testing and confidence interval of σ^2

- Assume the ϵ_i's in a simple linear regression model are normally distributed. Suppose we want to test $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 \neq \sigma_0^2$. If H_0 is true, $X = \frac{1}{\sigma_0^2}(n-2)MS_{Res}$ follows the χ^2 distribution with $n-2$ df.
- Reject H_0 if

$$\frac{(n-2)MS_{Res}}{\sigma_0^2} < \chi^2_{1-\alpha/2,n-2} \text{ or}$$

$$\frac{(n-2)MS_{Res}}{\sigma_0^2} > \chi^2_{\alpha/2,n-2}$$

  where α is the level of significance.
- A $100(1-\alpha)\%$ confidence interval for σ^2 is given as

$$\frac{(n-2)MS_{Res}}{\chi^2_{\alpha/2,n-2}} \leq \sigma^2 \leq \frac{(n-2)MS_{Res}}{\chi^2_{1-\alpha/2,n-2}}$$

Confidence interval of mean response $E(y)$ at x_0

$$\begin{aligned}
E(y|x_0) &= \mu_{y|x_0} = \beta_0 + \beta_1 x_0 \\
\widehat{E(y|x_0)} &= \hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0 \\
\text{Var}(\hat{\mu}_{y|x_0}) &= \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) \\
&= \text{Var}(\bar{y} + \hat{\beta}_1(x_0 - \bar{x})) \text{ since } \text{Cov}(\bar{y}, \hat{\beta}_1) = 0 \\
&= \frac{\sigma^2}{n} + \frac{\sigma^2(x_0-\bar{x})^2}{S_{xx}} \\
&= \sigma^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)
\end{aligned}$$

$\hat{\mu}_{y|x_0}$ is a normally distributed random variable because it is a linear combination of the obesrvations y_i.

$\frac{\hat{\mu}_{y|x_0} - E(y|x_0)}{\sqrt{MS_{Res}(1/n + (x_0-\bar{x})^2/S_{xx})}}$ is t with $n-2$ df

A $100(1-\alpha)$ percent CI on the mean response at $x = x_0$ is given by

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

$$\leq \hat{\mu}_{y|x_0} + t_{\alpha/2,n-2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

R code:

```
x0 <- mean(x)
data <- data.frame(x=x0)
predict.lm(model,data,interval="confidence",level=0.95)
```

Prediction interval for future observation y_0

$\psi = y_0 - \hat{y}_0$ is normally distributed with mean zero and variance

$$\text{Var}(\psi) = \text{Var}(y - 0\hat{y}_0) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right).$$

Note also that y_0 and $\hat{y}_0$ are independent. A $100(1-\alpha)$ percent prediction interval on a future observation y_0 is given by

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

$$\leq \hat{y}_0 + T_{\alpha/2,n-2}\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

Construction using R is the same as for confidence intervals, but with the interval parameter sent to be `"prediction"`

Let $\bar{y}_0$ be the **mean** of m future observations at $x = x_0$. A point estimator of $\bar{y}_0$ is $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$. A $100(1-\alpha)\%$ prediction interval is

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

$$\leq \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{MS_{Res}\left(\frac{1}{m} + \frac{1}{n} + \frac{(x-0\bar{x})^2}{S_{xx}}\right)}$$

Coefficient of determination R^2

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$$

$$SS_T = SS_R + SS_{Res}$$

- SS_T is a measure of variability in y without considering the effect of x.
- SS_{Res} is a measure of variability in y remaining after x has been considered.
- R^2 is often called the proportion of variation explained by the regressor x.
- $0 \leq R^2 \leq 1$, $0 \leq SS_{Res} \leq SS_T$

Considerations in using regression

- Regression models are intended as interpolation equations over the range of regressor variable(s).
- Disposition of x values plays an important role in least-squares fit.
- **Outliers** can seriously disturb least-squares fit.
- Strong relationship between two variables does not imply causal relationship between variables.
- In some applications of regression, the value of x required to predict y is unknown.

Regression through the origin

- No-intercept model: $y = \beta_1 x + \epsilon$
- Least-squares function: $S(\beta_1) = \sum_{i=1}^n (y_i - \beta_1 x_i)^2$
- Normal equation $\hat{\beta}_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i x_i$
- **Least-squares estimator of the slope**

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2}.$$

- Note that $\hat{\beta}_1$ is unbiased for β_1
- Estimator of σ^2

$$\begin{aligned}
\hat{\sigma}^2 = MS_{Res} &= \frac{1}{n-1}\sum_{i=1}^n (y_i - \hat{y}_i)^2 \\
&= \frac{1}{n-1}\left(\sum_{i=1}^n y_i^2 - \hat{\beta}_1 \sum_{i=1}^n y_i x_i\right)
\end{aligned}$$

- $100(1-\alpha)$ percent CI on β_1

$$\hat{\beta}_1 - t_{\alpha/2,n-1}\sqrt{\frac{MS_{Res}}{\sum_{i=1}^n x_i^2}} \leq \beta_1$$

$$\leq \hat{\beta}_1 + t_{\alpha/2,n-1}\sqrt{\frac{MS_{Res}}{\sum_{i=1}^n x_i^2}}$$

- $100(1-\alpha)$ percent CI on $E(y|x_0)$, the mean response at $x = x_0$

$$\hat{\mu}_{y|x_0} - t_{\alpha/2,n-1}\sqrt{\frac{x_0^2 MS_{Res}}{\sum_{i=1}^n x_i^2}} \leq E(y|x_0)$$

$$\leq \hat{\mu}_{y|x_0} + t_{\alpha/2,n-1}\sqrt{\frac{x_0^2 MS_{Res}}{\sum_{i=1}^n x_i^2}}$$

- $100(1-\alpha)$ percent prediction interval on a future observation at $x = x_0$

$$\hat{y}_0 - t_{\alpha/2,n-1}\sqrt{MS_{Res}\left(1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}\right)} \leq y_0$$

$$\leq \hat{y}_0 + t_{\alpha/2,n-1}\sqrt{MS_{Res}\left(1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}\right)}$$

Maximum likelihood estimation

- Assume that the ϵ_i's of a simple linear regression model are normally distributed. The responses y_i's are independently and normally distributed with mean $E(y_i) = \beta_0 + \beta_1 x_i$ and variance $\text{Var}(y_i) = \sigma^2$. The probability density of y_i is

$$f(y_i) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right).$$

- Likelihood function

$$L(y_i, x_i, \beta_0, \beta_i, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right]$$

- To find maximum-likelihood estimators $\tilde{\beta}_0, \tilde{\beta}_1$, and $\tilde{\sigma}^2$, maximise $\ln L$:

$$\begin{aligned}
\ln L(y_i, x_i, \beta_0, \beta_1, \sigma^2) &= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 \\
&\quad - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \\
\tilde{\beta}_0 &= \bar{y} - \bar{\beta}_1 \bar{x} \\
\tilde{\beta} &= \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
\tilde{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2 \\
\tilde{\sigma}^2 i &= \frac{n-1}{n}\tilde{\sigma}^2
\end{aligned}$$

  The maximum likelihood estimates are the same as the least-squares estimates except for σ^2.

Case where the regressor x is random

Suppose x and y are jointly distributed random variables but the form of the joint distribution is not known. All previous regression results hold if the following conditions are satisfied:

1. The conditional distribution of y given x is normal with conditional mean $\beta_0 + \beta_1 x$ and conditional variance σ^2
2. The x's are independent random variables whose probability distribution does not involve β_0, β_1, and σ^2.

The maximum-likelihood estimators are identical to those produced by least-squares. The sample correlation coefficient r^2 is also equal to the coefficient of determination R^2:

$$\begin{aligned}
r &= \frac{S_{xy}}{S_{xx}SS_T}^{1/2} \\
\hat{\beta}_1 &= \left(\frac{SS_T}{S_{xx}}\right)^{1/2} r \\
r^2 &= \hat{\beta}_1 \frac{S_{xx}}{SS_T} = \frac{\hat{\beta}_1 S_{xy}}{SS_T} = \frac{SS_R}{SS_T} = R^2
\end{aligned}$$

Hypothesis testing of population correlation coefficient

- $H : \rho = 0$, $H_1 : \rho \neq 0$. $t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ follows the t distribution with $n-2$ degrees of freedom if H_0 is true. Reject H_0 if $|t_0| > t_{\alpha/2, n-2}$
- $H_0 : \rho = \rho_0, H_1 : \rho \neq \rho_0$. For samples with $n \geq 25$, $Z = \text{arctanh } r = \frac{1}{2}\ln\frac{1+r}{1-r}$ is approximately normally distributed with $\mu_Z = \text{arctanh } \rho = \frac{1}{2}\ln\frac{1+\rho}{1-\rho}$ and $\sigma_Z^2 = (n-3)^{-1}$. Test statistic:
$$Z_0 = (\text{arctanh } r - \text{arctanh } \rho_0)(n-3)^{1/2}.$$
  Reject H_0 if $|Z_0| > Z_{\alpha/2}$.
- $100(1-\alpha)$ percent confidence interval for ρ:

$$\begin{aligned}
&\tanh\left(\text{arctanh } r - \frac{Z_{\alpha/2}}{\sqrt{n-3}}\right) \leq \rho \\
&\leq \tanh\left(\text{arctanh } r + \frac{Z_{\alpha/2}}{\sqrt{n-3}}\right)
\end{aligned}$$

  where $\tanh u = (e^u - e^{-u})/(e^u + e^{-u})$.

Extractor functions for `lm()`

`summary()`	returns summary information
`plot()`	makes diagnostic plots
`coef()`	returns the coefficients
`residuals()`	returns the residuals (also `resid()`
`fitted()`	returns fitted values $\hat{y}_i$
`deviance()`	returns RSS
`predict()`	performs predictions
`anova()`	finds various sums of squares

Multiple linear regression

Multiple linear regression model for response variable y and regressor variables $x_1, \ldots, x_k$:

$$\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i \\
&= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i
\end{aligned}$$

where $E(\epsilon_i) = 0$, $\text{Var}(\epsilon_i) = \sigma^2$, and ϵ_i's are independent. Number of regressor variables is k and number of regression coefficients is $p = k + 1$. β_j are partial regression coefficients which represent the expected change in the response per unit change in x_j when all remaining regressor variables are held constant.

Least-squares

Function

$$S = \sum_{i=1}^{n}\epsilon_i^2 = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2.$$

Matrix approach

Obtain least-squares estimates $\hat{\beta}_0, \ldots, \hat{\beta}_k$ by minimising error sum of squares with respect to $\beta_0, \ldots, \beta_k$. We get $y = X\beta + \epsilon$ where

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \; X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \; \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Least squares function can thus be written as

$$S = \sum_{i=1}^{n}\epsilon_i^2 = \epsilon'\epsilon = y'y - 2y'X\beta + \beta'X'X\beta$$

Normal equation: $X'X\hat{\beta} = X'y$

Least-squares estimate of $\hat{\beta}$: $\hat{\beta} = (X'X)^{-1}X'y$

Predicted response $\hat{y}$ and residual e

$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y = Hy$ where H is called the hat matrix.
$e = y - \hat{y} = (I_n - H)y$

Properties of least-squares estimators

- Let A be a $k \times k$ matrix of constants and y be a $k \times 1$ random vector with mean $E(y) = \mu$ and non-singular variance-covariance matrix $\text{Var}(y) = V$. Then $E(Ay) = A\mu$ and $\text{Var}(Ay) = AVA'$.
- $\hat{\beta}$ is an unbiased estimator of β i.e. $E(\hat{\beta}) = \beta$
- Variance-covariance matrix of $\hat{\beta}$ $\text{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}$

$$\begin{aligned}
\text{Var}(\hat{\beta}) &= \sigma^2(X'X)^{-1} = \sigma^2 C \\
&= \sigma^2 \begin{pmatrix} C_{00} & C_{01} & \ldots & C_{0k} \\ C_{10} & C_{11} & \ldots & C_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ C_{k0} & C_{k1} & \ldots & C_{kk} \end{pmatrix}
\end{aligned}$$

  Note that $\text{Var}(\hat{\beta}_j) = \sigma^2 C_{jj}$, $\text{Cov}(\hat{\beta}_i, \hat{\beta}_k) = \sigma^2 C_{ij}$

Residual sum of squares and estimation of σ^2

$$\begin{aligned}
SS_{Res} &= \sum_{i=1}^{n} e_i^2 = e'e = y'y - \hat{\beta}'X'y \\
\hat{\sigma}^2 &= MS_{Res} = \frac{SS_{Res}}{n-p} \\
E(MS_{Res}) &= \sigma^2
\end{aligned}$$

Maximum-Likelihood Estimation

ϵ is normally distributed with mean 0 and variance $\sigma^2 I$. The likelihood function is the joint density of $\epsilon_1, \ldots, \epsilon_n$

$$L(\epsilon, \beta, \sigma^2) = \prod_{i=1}^{n} f(\epsilon_i) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\epsilon'\epsilon\right)$$

$$L(y, X, \beta, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right)$$

$$\ln L(y, X, \beta, \sigma^2) = -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)$$

Maximising $\ln L(y, X, \beta, \sigma^2)$ is the same as minimising the least-squares function $(y - X\beta)'(y - X\beta)$ for a fixed value of σ. Thus the MLE and least-squares estimator are both $\hat{\beta} = (X'X)^{-1}X'y$. MLE for σ^2 is $\tilde{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})'(y - X\hat{\beta})$

Decomposition of variance

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

$$SS_{Res} \equiv y'y - \hat{\beta}'X'y$$

$$SS_R \equiv \hat{\beta}'X'y - n\bar{y}^2$$

$$SS_T = SS_{Res} + SS_R$$

Test of overall fit of model - analysis of variance

Assume $\epsilon \sim N(0, \sigma^2 I)$. Test $H_0 : \beta_1 = \cdots = \beta_k = 0$ against H_1 : At least one β_j is not equal to zero. Similar to the case for simple linear regression, we test $F = \frac{SS_R/k}{SS_{Res}/(n-p)} = \frac{MS_R}{MS_{Res}}$ since F follows the F distribution with degrees of freedom k and $n - p$ when H_0 is true. For a given level of significance α, reject H_0 if $F > F_{\alpha,k,n-k-1}$.

ANOVA table:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Regression	SS_R	k	MS_R	MS_R/MS_{Res}
Residual	SS_{Res}	$n-(k+1)$	MS_{Res}	
Total	SS_T	$n-1$		

R^2 and adjusted R^2 for assessing overall adequacy of model

- $R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$, adjusted $R^2 = 1 - \frac{SS_{Res}/(n-p)}{SS_T/(n-1)} = 1 - \frac{MS_Res}{SS_T/(n-1)}$
- R^2 is useful in assessing the contribution of an additional variable.

Testing and confidence intervals on individual regression coefficients

- $T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}$ follows the t distribution with $n - p$ degrees of freedom.
- To test $H_0 : \beta_j = c$ versus $H_1 : \beta_j \neq c$, reject H_0 if

$$\frac{\hat{\beta}_j - c}{\sqrt{\hat{\sigma}^2 C_{jj}}} < -t_{\alpha/2,n-p} \text{ or } \frac{\hat{\beta}_j - c}{\sqrt{\hat{\sigma}^2 C_{jj}}} > t_{\alpha/2,n-p}$$

- A $100(1-\alpha)\%$ confidence interval for β_j is given as

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

- R code: `summary(lm(y~x1+x2+x3))`, `confint(lm(y~x1+x2+x3), level=0.95)`

$SS_R(\beta)$ and $SS_R(\beta_2|\beta_1)$

Let β_1 contain the first $p - r$ regression coefficients and β_2 contain the last r coefficients. We have

$$y = X\beta + \epsilon = X_1\beta_1 + X_2\beta_2 + \epsilon$$

$$SS_R(\beta) = \hat{\beta}'X'y,\ SS_R(\beta_1) = \hat{\beta}_1'X_1'y,\ SS_R(\beta_2) = \hat{\beta}_2'X_2'y$$

$$SS_R(\beta_2|\beta_1) \equiv SS_R(\beta_1, \beta_2) - SS_R(\beta_1)$$

$SS_R(\beta)$ denotes the regression sum of squares due to β. $SS_R(\beta_2|\beta_1)$ denotes the regression sum of squares due to β_2 given that β_1 is already in the model.

Fitting $y = \beta_0 + \epsilon$ and testing $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$

- Note that $SS_R(\beta_0) = n\bar{y}^2$ ($SS_R(\beta_0)$ is the regression sum of squares due to β_0)
- Since $E(y) = \beta_0$, testing $H_0 : \beta_0 = 0$ is the same as testing whether the sample was taken from a normal population with mean β_0. Use a t-test of population mean assuming population variance is unknown:

$$t = \frac{\bar{y} - 0}{s_y/\sqrt{n}} \text{ where } s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

  Reject H_0 if $|t| > t_{\alpha/2,n-1}$.
- According to decomposition of variance, test $F = \frac{SS_R(\beta_0)/1}{SS_{Res}/(n-1)}$. Reject H_0 if $F > F_{\alpha,1,n-1}$. It can be shown that $F = t^2$.

SS_R and $SS_R(\beta)$

We have $SS_R = SS_R(\beta_1, \ldots, \beta_k|\beta_0)$, so SS_R is the regression sum of squares due to $\beta_1, \ldots, \beta_k$ given that β_0 is already in the model. Thus we use SS_R to test $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$.

Extra sum of squares

- Consider the multiple linear regression model with k regressor variables $y = X_1\beta_1 + X_2\beta_2 + \epsilon$. Let this be the full model (FM), and let the reduced model (RM) be $y = X_1\beta_1 + \epsilon$.
- To test the last r regression coefficients $H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 \neq 0$ use

$$F = \frac{SS_R(\beta_2|\beta_1)/r}{SS_{Res}(FM)/(n-p)}.$$

  Reject H_0 if $F > F_{\alpha,r,n-p}$.
- $SS_R(\beta_2|\beta_1)$ is known as the **extra** sum of squares due to β_2 because it measures the increase in regression sum of squares that results from adding β_2 to a model that already contains β_1.
- $SS_R(\beta_2|\beta_1) = SS_{Res}(RM) - SS_{Res}(FM)$
- The df of $SS_R(\beta_2|\beta_1)$ can be found as

$$\begin{aligned} df(SS_R(\beta_2|\beta_1)) &= df(SS_{Res}(RM)) - df(SS_{Res}(FM)) \\ &= [n - (p - r)] - [n - p] \\ &= r \end{aligned}$$

Testing general hypothesis about β

- Suppose we have a full model with $\epsilon \sim N(0, \sigma^2 I)$. We want to test $H_0 : T\beta = 0$ where T is an $r \times p$ matrix such that all r equations in $T\beta = 0$ are independent.
- By applying $T\beta = 0$ to the full model, we obtain a reduced model $y_{n\times 1} = Z_{n\times(p-r)}\Gamma_{(p-r)\times 1} + \epsilon_{n\times 1}$ where $\Gamma = (\gamma_0, \ldots, \gamma_{p-r-1})$.
- To test $H_0 : T\beta = 0$ versus $H_1 : T\beta \neq 0$, consider the FM $y = X\beta + \epsilon$ and RM $y = Z\Gamma + \epsilon$.
- $SS_R = SS_{Res}(RM) - SS_{Res}(FM)$, df for $SS_R = r$
- Test using $F = \frac{SS_R/r}{SS_{Res}(FM)/(n-p)}$. Reject H_0 if $F > F_{\alpha,r,n-p}$

Tests and confidence intervals on individual regression coefficients

- $T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}$ follows the t distribution with $n - p$ df.
- To test $H_0 : \beta_k = c$ vs $H_1 : \beta_k \neq c$, reject H_0 if

$$\frac{\hat{\beta}_j - c}{\sqrt{\hat{\sigma}^2 C_{jj}}} < -t_{\alpha/2,n-p} \text{ or } \frac{\hat{\beta}_j - c}{\sqrt{\hat{\sigma}^2 C_{jj}}} > t_{\alpha/2,n-p}.$$

- A $100(1-\alpha)\%$ confidence interval for β_j is given as

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

Estimation of mean response

- Suppose we want to predict a future response y_0 at the point $x_0' = [1\ x_{01} \ldots\ x_{0k}]$. The predicted response is $\hat{y}_0 = x_0'\hat{\beta}$.
- The mean of $y_0 - \hat{y}_0$ is zero since $E(y_0) = E(\hat{y}_0) = x_0'\beta$.
-

$$\begin{aligned} \text{Var}(y_0 - \hat{y}_0) &= \text{Var}(y_0) + \text{Var}(\hat{y}_0) - 2\text{Cov}(y_0, \hat{y}_0) \\ &= \text{Var}(y_0) + \text{Var}(\hat{y}_0) \\ &= \sigma^2(1 + x_0'(X'X)^{-1}x_0) \end{aligned}$$

  Also, $\widehat{\text{Var}(y_0 - \hat{y}_0)} = \hat{\sigma}^2(1 + x_0'(X'X)^{-1}x_0)$
- $\frac{y_0 - \hat{y}_0}{\sqrt{\widehat{\text{Var}(y_0 - \hat{y}_0)}}}$ follows the t distribution with $n - p$ degrees of freedom.
- $100(1-\alpha)\%$ prediction interval for y_0:

$$\hat{y}_0 - t_{\alpha/2,n-p}\sqrt{\widehat{\text{Var}(y_0 - \hat{y}_0)}} \leq y_0 \leq \hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\widehat{\text{Var}(y_0 - \hat{y}_0)}}$$

Simultaneous confidence intervals on regression coefficients

A simultaneous confidence interval for β is a joint region that applies simultaneously to the entire set of regression coefficients.

Ellipsoidal confidence region: $\frac{(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta)}{p\ MS_{Res}} \sim F_{p,n-p}$

$$P\left(\frac{(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta)}{p\ MS_{Res}} \leq F_{p,n-p}\right) = 1-\alpha.$$

A $100(1-\alpha)\%$ joint confidence region for all parameters in β is $\frac{(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta)}{p\ MS_{Res}} \leq F_{\alpha,p,n-p}$. The joint confidence region is an ellipsoidal region.

Bonferroni intervals

- This is based on the Bonferroni inequality:

$$P\left(\bigcup_{i=1}^{m} A_i^c\right) \leq \sum_{i=1}^{m} P(A_i^c)$$

which gives us

$$P\left(\bigcap_{i=1}^{m} A_i\right) \geq 1 - \sum_{i=1}^{m} P(A_i^c).$$

Suppose $P(A_i) = 1-\alpha/m$. Then $P(A_i^c) = \alpha/m$ and $P\left(\bigcap_{i=1}^{m} A_i\right) \geq 1-\alpha$. In order to construct $(1-\alpha)100\%$ simultaneous confidence intervals for m parameters, construct $(1-\alpha/m)100\%$ confidence intervals for each of the m parameters. Thus the joint coverage probability of the m intervals will be at least $1-\alpha$.

- The Bonferroni confidence region is rectangular.
- To consruct a $(1-\alpha)100\%$ joint confidence region for $\beta_0, \ldots, \beta_k$, construct $(1-\alpha/p)100\%$ confidence intervals for $\beta_0, \ldots, \beta_k$

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

The p intervals form the Bonferroni confidence region that contains $(\beta_0, \ldots, \beta_k)'$ with probability of at least $1-\alpha$. This means the intervals are wider than they are supposed to be, and thus the Bonferroni confidence region is conservative.

Hidden extrapolation in multiple regression

- The **regressor variable hull (RVH)** is the smallest convex set containing all the original n data points. If a point lies inside or on the boundary of the RVH, prediction involves interpolation. If it lies outside the RVH, the prediction is based on extrapolation.
- The diagonal elements of the hat matrix $H_{n\times n} = X_{n\times p}(X'X)^{-1}_{p\times p}X'_{p\times n}$ are useful in detecting hidden extrapolation. h_{ii} depends on the Eucildean distance of the point x_i from the centroid and on the density of the points in the RVH. In general, the point with the largest value of h_{ii} (h_{max}) will lie on the boundary of the RVH in a region of the x space where the density of observations is low.
- If x satisfies $x'(X'X)^{-1}x \leq h_{max}$, it is in the ellipsoid enclosing the RVH, and possibly in the RVH.

Standardising regression coefficients

Unit normal scaling:

$$y_i^* = \frac{y_i - \bar{y}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}},\ z_{ik} = \frac{x_{ik}-\bar{x}_k}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}}$$

Least-squares estimate of b_0 is zero if all variables are unit normal scaled. The least-squares regression coefficients are $\hat{b} = (Z'Z)^{-1}Z'y^*$.
Unit length scaling:

$$y_i^* = \frac{y_i - \bar{y}}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}},\ w_{ik} = \frac{x_{ik}-\bar{x}_k}{\sqrt{\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}}$$

Least-squares estimate of b_0 is zero if all variables are unit length scaled. The least-squares regression coefficients are $\hat{b} = (W'W)^{-1}W'y$. For unit length scaled variables, the off-diagonal elements r_{ij} of $W'W$ are the correlation coefficients of x_i and x_j.

Multicollinearity

- The main diagonal elements of $C = (X'X)^{-1}$ are called the **variance inflation factors (VIFs)**. $VIF_j = C_{jj}$.
- It can be shown that $VIF_j = \frac{1}{1-R_j^2}$ where R_j^2 is the coefficient of determination obtained when x_j is regressed on the remaining $k-1$ regressor variables.
- If x_j is nearly dependent on some subset of the remaining regressor variables, R_j^2 will be near one and VIF_j will be much greater than one. If x_j is orthogonal to the remaining regressor variables, R_j^2 will be near zero and VIF_j will be near one.
- VIF_j measures how much the variance of $\hat{\beta}_j$ is affected by the relationship of x_j with the other regressor variables.
- VIF_j can be used to detect multicollinearity. In general, $V_F \geq 2.5$ provides some evidence of multicollinearity.
- R code: `vif(fitted.model)` returns VIFs for unit length scaled regressor variables.

Regression coefficients having the wrong sign

- Range of regressors is too small
- Important regressors have not been included in the model
- Multicollinearity is present

X matrix with orthogonal columns

- Consider the model $y = X\beta + \epsilon = X_1\beta_1 + X_2\beta_2 + \epsilon$. If the columns of X_1 are orthogonal to those in X_2, the normal equations become $X_1'X_1\hat{\beta}_1 = X_1'y$ and $X_2'X_2\hat{\beta}_2 = X_2'y$.
- We can obtain $SS_R(\beta) = SS_R(\beta_1) + SS_R(\beta_2)$, which gives us $SS_R(\beta_1|\beta_2) = SS_R(\beta) - SS_R(\beta_2) = SS_R(\beta_1)$ and $SS_R(\beta_2|\beta_1) = SS_R(\beta) - SS_R(\beta_1) = SS_R(\beta_2)$.

Model Adequacy Checking

Major assumptions so far:

- Relationship between response y amd regressors is (approximately) linear
- Error term ϵ has zero mean and constant variance σ^2
- Errors are uncorrelated and normally distributed

Residual Analysis

- Residual $e_i = y_i - \hat{y}_i$. They have zero mean and approximate average variance estimated by

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

However the residuals are not independent as the n residuals have only $n-p$ df.

Methods for scaling residuals:

- Standardised residuals: $\frac{e_i}{\sqrt{MS_{Res}}}$
- Studentised residuals: note that $E(e) = 0$ and $\text{Var}(e) = (I-H)\sigma^2$ where H is the hat matrix. Since H is both symmetric and idempotent, so is $I-H$. This gives us the studentised residual

$$r_i = \frac{e_i}{\sqrt{MS_{Res}(1-h_{ii})}}.$$

Note that the value of h_{ii} is a measure of the Euclidean distance of the ith observation from the centroid of the data. Hence an outlier will result in a large studentised residual.

- PRESS residuals: $e_{(i)} = y_i - \hat{y}_{(i)}$ where $\hat{y}_{(i)}$ is the fitted value of the ith response based on all observations except the ith one. It can be shown that $e_{(i)} = \frac{e_i}{1-h_{ii}}$. We have $\text{Var}(e_{(i)}) = \frac{\sigma^2}{1-h_{ii}}$. Standardised PRESS residual $= \frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}}$
- R-Student residual: $t_i = \frac{e_i}{\sqrt{S^2_{(i)}(1-h_{ii})}}$

Plots of residuals in time sequence

One key assumption in linear regression is that responses are independent. In practice, they are often correlated. The correlation between model errors at different time periods is called autocorrelation. This can be detected by plotting the residual against time.

PRESS statistic

$$PRESS = \sum_{i=1}^{n} e_{(i)}^2 = \sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2$$

The PRESS statistic is used as a measure of model quality, and a small value is desirable.

Normal probability plot

Suppose $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Then $\frac{X_1-\mu}{\sigma}, \ldots, \frac{X_n-\mu}{\sigma} \sim N(0,1)$
Let cdf of $\frac{X_i-\mu}{\sigma}$ be $F(x)$ and order in increasing order $F\left(\frac{X_{(1)}-\mu}{\sigma}\right), \ldots, F\left(\frac{X_{(n)}-\mu}{\sigma}\right)$. Then
$F\left(\frac{X_{(i)}-\mu}{\sigma}\right) \sim Beta(i, n+1-i)$, $E\left[F\left(\frac{X_{(i)}-\mu}{\sigma}\right)\right] = \frac{i}{n+1}$,
$F\left(\frac{X_{(i)}-\mu}{\sigma}\right) \approx \frac{i}{n+1}$, $X_{(i)} \approx \sigma F^{-1}\left(\frac{i}{n+1}\right) + \mu$.
Q-Q plot: $X_{(i)}$ vs $F^{-1}\left(\frac{i}{n+1}\right)$, P-P plot: $F\left(\frac{X_{(i)}-\mu}{\sigma}\right)$ vs $\frac{i}{n+1}$. If the data set has a normal distribution, plots should be linear

LOF test

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

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

H_0: no LOF, H_1: LOF. $F_0 = \frac{SS_{LOF}/(m-2)}{SS_{PE}/(n-m)} = \frac{MS_{LOF}}{MS_{PE}}$

$$E(MS_{LOF}) = \sigma^2 + \frac{\sum_{i=1}^{m} n_i[E(y_i)-\beta_0-\beta_1 x_i]^2}{m-2}$$

If $F_0 > F_{\alpha,m-2,n-m}$ conclude regression function not linear, otherwise no strong evidence of LOF and MS_{PE} and MS_{LOF} are often combined to estimate σ^2

Transformations and weighting to correct model inadequacies

Assumptions in doing regression:

- Model errors have mean zero, constant variance, and are uncorrelated
- Model errors have normal distribution and are independent
- Form of the model is correct

Transformations help us build models when these are violated.

Variance stabilising

Relationship of σ^2 to $E(y)$	Transformation
$\sigma^2 \propto$ constant	$y' = y$ (no transformation)
$\sigma^2 \propto E(y)$	$y' = \sqrt{y}$ (square root; Poisson)
$\sigma^2 \propto E(y)[1-E(y)]$	$y' = \sin^{-1}(\sqrt{y})$ (arcsin; binomial proportions $0 \le y_i \le 1$)
$\sigma^2 \propto [E(y)]^2$	$y' = \ln(y)$(log)
$\sigma^2 \propto [E(y)]^3$	$y' = y^{-1/2}$ (reciprocal square root)
$\sigma^2 \propto [E(y)]^4$	$y' = y^{-1}$ (reciprocal)

Linearising

Function	Transformation	Linear form
$y = \beta_0 x^{\beta_1}$	$y' = \log y$, $x' = \log x$	$y' = \log\beta_0 + \beta_1 x'$
$y = \beta_0 e^{\beta_1 x}$	$y' = \ln y$	$y' = \ln\beta_0 + \beta_1 x$
$y = \beta_0 + \beta_1 \log x$	$x' = \log x$	$y' = \beta_0 + \beta_1 x'$
$y = \frac{x}{\beta_0 x - \beta_1}$	$y' = \frac{1}{y}$,$x' = \frac{1}{x}$	$y' = \beta_0 - \beta_1 x'$

Transformations on y (Box-Cox method)

- For the model $y = X\beta + \epsilon$, use the power transformation

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

- Parameters of the model and λ can be estimated simultaneously using maximum likelihood. Max-likelihood estimate of λ is the value when SS_{Res} is minimum, and can be obtained by fitting a model to $y^{(\lambda)}$ for various values of λ, plotting SS_{Res} vs λ, and then reading the value of λ that minimises SS_{Res} from the graph.
- λ cannot be selected by directly comparing SS_{Res} from the regression of y^λ because SS_{Res} for each λ is measured on a different scale. Instead, use

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

where $\tilde{y} = e^{[\frac{1}{n}\sum_{i=1}^{n}\log(y_i)]}$

Generalised least squares

- $y = X\beta + \epsilon$ where $E(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2 V$ (vs $\sigma^2 I$) for OLS
- V nonsingular $\implies \exists n \times n$ nonsingular K such that $K'K = KK = V$. Thus $K = K'$ and $V^{-1} = K^{-1}K^{-1}$
- Gen least sq model can be converted into OLS using the transformation $z = K^{-1}y$, $B = K^{-1}X$, $g = K^{-1}\epsilon$ to get $Z = B\beta + g$
- Least squares function $S(\beta) = g'g = (y - X\beta)'V^{-1}(y - X\beta)$
- Least squares normal equation $X'V^{-1}X\hat{\beta} = X'V^{-1}y$
- $\hat{\beta}$ is an unbiased estimator of β – $\text{Var}(\hat{\beta}) = \sigma^2 X'V^{-1}X$

ANOVA for generalised least squares

Source	Sum of squares		df	Mean square	F_0
Regression	SS_R $\hat{\beta}'B'z$	=	p	SS_R/p	MS_R/MS_{Res}
Error	SS_{Res} $z'z - \hat{\beta}'B'z$	=	$n-p$	$SS_{Res}/(n-p)$	
Total	$z'z$ $y'V^{-1}y$	=	n		

Weighted least squares

- ϵ uncorrelated but have unequal variances so that covariance matrix of ϵ is

$$\sigma^2 V = \sigma^2 \begin{pmatrix} 1/w_1 & & & 0 \\ & 1/w_2 & & \\ & & \ddots & \\ 0 & & & 1/w_n \end{pmatrix}$$

- $W = V^{-1}$, $(X'WX)\hat{\beta} = X'Wy$. $\hat{\beta} = (X'WX)^{-1}X'Wy$ is the weighted least-squares estimator
- $z = K^{-1}y$, $B = K^{-1}X$

$$B = \begin{pmatrix} 1\sqrt{w_1} & \cdots & x_{1k}\sqrt{w_1} \\ \vdots & \ddots & \vdots \\ 1\sqrt{w_n} & \cdots & x_{nk}\sqrt{w_n} \end{pmatrix}, z = (y_1\sqrt{w_1}, \ldots, y_n\sqrt{w_n})'$$

- Apply OLS to transformed data to get $\hat{\beta} = (B'B)^{-1}B'z = (X'WX)^{-1}X'Wy$ which is the weighted least-squares estimate of β

Diagnostics for leverage and influence

- Diagonal elements h_{ii} of the hat matrix $H = X(X'X)^{-1}X'$ are a measure of the distsance of the ith observation (ith row) from the centroid of the x space
- Note that $h_{ii} = x_i'(X'X)^{-1}x_i$, $\sum_{i=1}^{n} h_{ii} = \text{rank}(H) = p$, $\bar{h} = p/n$
- If h_{ii} exceeds twice the average $2p/n$, it is a **leverage point**
- An **influential point** is a leverage point that affects the regression coefficients significantly

Cook's distance

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'X'X(\hat{\beta}_{(i)} - \hat{\beta})}{pMS_{Res}}$$

- Measure of squared distance between least-squares estimate based on all n points $\hat{\beta}$ and estimate obtained by deleting ith point $\hat{\beta}_{(i)}$
- Large value of D_i means considerable influence on $\hat{\beta}$
- If $D_i = F_{0.5,p,n-p} \approx 1$, $\hat{\beta}_{(i)}$ is on the boundary of an approximate 50% confidence region for β
- Any $\hat{\beta}_{(i)}$ beyond the 50% confidence region is influential

Alternative formulae for Cook's distance

$$D_i = \frac{r_i^2}{p}\frac{\text{Var}(\hat{y}_i)}{\text{Var}(e_i)} = \frac{r_i^2}{p}\frac{h_{ii}}{1-h_{ii}}$$

D_i is the product of the square of the ith studentised residual and $h_{ii}/(1-h_{ii})$ apart from the constant p. This ratio is the distance from x_i to the centroid of the remaining data

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'X'X(\hat{\beta}_{(i)} - \hat{\beta})}{pMS_{Res}} = \frac{(\hat{y}_{(i)} - \hat{y})'(\hat{y}_{(i)} - \hat{y})}{pMS_{Res}}$$

Squared Euclidean distance that vector of fitted values moves when ith observation is deleted

DFBETAS

$$DFBETAS_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{S_{(i)}^2 C_{jj}}}$$

where C_{jj} is the jth diagonal element of $(X'X)^{-1}$, $S_{(i)}^2$ is the estimate of σ^2 based on the dataset with the ith observation removed, $\hat{\beta}_{j(i)}$ is the jth regression coefficient computed without use of the ith observation.

- Large $DFBETAS_{j,i}$ means observation i has considerable influence over the jth regression coefficent
- Belsey, Kuh, Welsh: $|DFBETAS_{j,i}| > 2/\sqrt{n} \implies$ influential

DFFITS

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{S_{(i)}^2 h_{ii}}} = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} t_i$$

where $\hat{y}_{(i)}$ is the fitted value of y_i obtained without use of the ith observation and t_i is R-student

- $DFFITS_i$ is the number of standard deviations that the fitted value $\hat{y}_i$ changes if observation i is removed
- Belsey, Kuh, Welsch: $|DFFITS_i| > 2\sqrt{p/n} \implies$ influential

Measure of model performance

$$\text{Generalised variance } = \left|\text{Var}(\hat{\beta})\right| = \det(\sigma^2(X'X)^{-1})$$

A small generalised variance is desirable.

$$COVRATIO_i = \frac{\left|(X'_{(i)}X_{(i)})^{-1}S^2_{(i)}\right|}{|(X'X)^{-1}MS_{Res}|} = \frac{(S^2_{(i)})^p}{MS^p_{Res}}\left(\frac{1}{1-h_{ii}}\right)$$

where $S^2_{(i)} = MS_{Res}$ calculated without the ith observation.

- $COVRATIO_i > 1 \implies i$th observation improves precision of estimation
- $COVRATIO_i < 1 \implies$ inclusion of ith point degrades precision
- A high leverage point will make $COVRATIO_i$ large
- Belsey, Kuh, Welsch: $COVRATIO_i > 1 + 3p/n$ or $COVRATIO_i < 1 - 3p/n \implies i$th point should be considered influential. Lower bound only appropriate when $n > 3p$

Polynomial regression

- kth order polynomial model in 1 variable: $y = \beta_0 + \beta_1 x + \cdots + \beta_k x^k + \epsilon$
- second order polynomial in 2 variables: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11}x_1^2 + \beta_{22}x_2^2 + \beta_{12}x_1x_2 + \epsilon$

Order

- As order increases, R^2 increases, MS_{Res} decreases
- Poly model of order $n-1$ can be fitted through n points such that $R^2 = 1$, $MS_{Res} = 0$. But this is not useful
- Use of $k > 2$ polynomials should be avoided

Model building

- Forward selection: start with lowest order, increase until fitted highest order term is non significant
- Backward elimination: start with highest order, delete terms starting with highest order one at a time till highest order remaining term is significant
- Two approaches may not yield same model, should try first or second order poly

Extrapolation: Can be hazardous

Ill-conditioning I

- As order increases, $X'X$ becomes ill-conditioned
- One way to reduce ill-conditioning of $X'X$ is to use $x - \bar{x}$ as the regressor variable instead of x

Ill-conditioning II: if range of x is small, x and x^2 can be highly correlated and cause ill-conditioned $X'X$

Hierarchy

- Hierarhical model: for order k, must include all terms with exponents $1, \ldots, k$
- Hierarchical models are invariant under lin transformation of regressor variable

Splines

- Piecewise polynomial fitting: split range of x into different segments and fit a spline in each one
- Joint points of the segments are **knots**
- Cubic spline: polynomial in each segment has order 3. h knots $t_1 < \cdots < t_h$ with cont first and second derivatives

$$E(y) = S(x) = \sum_{j=0}^{3}\beta_{0_j}x^j + \sum_{i=1}^{h}\beta_i(x-t_i)^3_+$$

  where $(x-t_i)_+ = x - t_i$ if > 0 and 0 otherwise
- To test $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$, H_1 : at least one β is nonzero use anova
- To compare simple cubic and cubic spline test $H_0 = \beta_4 = \beta_5 = 0$, H_1 : at least one β is nonzero
- Cubic spline model with no continuitiy restrictions

$$E(y) = S(x) = \sum_{j=0}^{3}\beta_{0_j}x^j + \sum_{i=1}^{h}\sum_{j=0}^{3}\beta_{ij}(x-t_i)^j_+$$

  where $(x-t)^0_+ = 1$ if $x > t$ and 0 otherwise. If a term $\beta_{i0}(x-t_i)^0_+$ is in the model, it forces a discontinuity at t_i

Nonparametric regression

- Both parametric and nonparametric models are linear combinations of the data, but nonparametric models set the weights differently
- Kernel regression for bandwidth b

$$\tilde{y} = \sum_{j=1}^{n} w_j y_j$$

$$w_j = K\left(\frac{x-x_h}{b}\right)/\sum_{i=1}^{n}K\left(\frac{x-x_j}{b}\right),\ \sum_{j=1}^{n}w_j = 1$$

$$K(t) = 1 \text{ if } |t| \le 0.5,\ 0 \text{ if } |t| > 0.5$$

$$K\left(\frac{x-x_k}{b}\right) = 1 \iff x - 0.5b \le x_k \le x + 0.5b$$

- Kernel functions:

Box	$K(t) = 1$ if $\lvert t\rvert \le 0.5$ else 0
Triangle	$K(t) = 1 - \lvert t\rvert /c$ if $\lvert t\rvert \le c$ else 0
Normal	$K(t) = \frac{1}{\sqrt{2\pi}k_6}\exp\{\frac{-t^2}{2k_6^2}\}$

  Requirements: $K(t) \ge 0\ \forall t$, $\int_{-\infty}^{\infty}K(t)dt = 1$, $K(-t) = K(t)$ (symmetry). But properties of kernel smoother depend more on choice of bandwidth than kernel function
- Locally weighted regressoin (loess) uses data from the neighbourhood around a point x_0 (span) which is the fraction of total points used to form neighbourhoods
- Let $\Delta(x_0)$ be the dist between x_0 and the furthest point in x_0's neighbourhood. Tri-cube weight function is $W\left[\frac{[x_0-x_j]}{\Delta(x_0)}\right]$ where $W(t) = (1-t^3)^3$ for $0 \le t < 1$ and 0 elsewhere
- Weighted least squares: covariance matrix of ϵ is $\sigma^2\text{diag}[\frac{1}{w_1}, \ldots, \frac{1}{w_n}]$
- Estimating σ^2 based on loess: $\tilde{y} = Sy$, need to fit n weighted least sq models to get S

$$SS_{Res} = y'[I - S' - S + S'S]y$$

$$E(SS_{Res}) \approx \sigma^2[n - 2\text{trace}(S) + \text{trace}(S'S)]$$

$$R^2 = \frac{SS_T - SS_{Res}}{SS_T}$$

- Ordinary least sq: $\hat{y} = Hy$. Loess estimate is asymptotically unbiased for $X\beta$ i.e. for large sample size $S \approx H$

Second order model

- Second-order polynomial in 2 variables:

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_{11}x_1^2 + \beta_{22}x_2^2 + \beta_{12}x_1x_2 + \epsilon$$

- $E(y) = y - \epsilon$ is the response surface
- **ADD MORE STUFF FROM TUTORIAL**

Indicator variables

- Indicator variable: categorical variable coded for fitting a multiple linear regression model
- E.g. variable with value 1 if variable is of type A, 0 if variable is of type B

Interaction

- Interaction term: if two variables x_1, x_2, model has βx_1x_2
- Comparing regression models: parallel ($H_0 : \beta_{12} = \cdots = \beta_{1M} = 0$), concurrent ($H_0 : \beta_2 = \cdots = \beta_M = 0$), coincident ($H_0 : \beta_2 = \cdots = \beta_M = 0, \beta_{12} = \cdots = \beta_{1M} = 0$)
- Allocated codes impose a particular metric with different distances between each qualitative factor. No guarantee this spacing is correct

Anova

- Factorial design: $y_{ij} = \mu + \tau_i + \epsilon_{ij}$ for $i = 1, \ldots, a$, $j = 1, \ldots, n$, $\sum_{i=1}^{a}\tau_i = 0$ where μ is the overall mean, τ_i is the effect due to level i of the factor, ϵ_{ij} is random error. Number of levels is a, n experiments conducted

Anova table

Source of variation	Sum of Squares	Df	Mean sq	F_0
Between treatments	$SS_{Treatments} = n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2$	$k-1$	$MS_{Treatments}$	$\frac{MS_{Treatments}}{MS_{Res}}$
Error (within treatments)	$SS_E = SS_T - SS_{Treatments}$	$N-a$	MS_E	
Total	$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2$	$N-1$		

$$y_{i.} = \sum_{j=1}^{n}y_{ij},\ \bar{y}_{i.} = y_{i.}/n,\ y_{..} = \sum_{i=1}^{a}\sum_{j=1}^{n}y_{ij},\ \bar{y}_{..} = y_{..}/N$$

- One-way factorial design can be treated as a regression problem where all regressors are indicator variables. For one-way factorial design:

$$H_0 : \tau_1 = \tau_2 = \tau_3 = 0, H_1 : \tau_i \neq 0 \text{ for at least one } i$$

for regression model:

$$H_0 : \beta_1 = \beta_2 = 0, H_1 \beta_i \neq 0 \text{ for at least one } i$$

$$\begin{aligned}\beta_0 =& \mu_3, \beta_1 = \mu_1 - \mu_3, \beta_2 = \mu_2 - \mu_3 \\ \beta_1 = \beta_2 = 0 &\implies \mu_1 = \mu_2 \mu_3 \implies \tau_1 = \tau_2 = \tau_3 \\ &\implies \tau_1 = \tau_2 = \tau_3 = 0\end{aligned}$$

SO testing $\beta_1 = \beta_2 = 0$ is the same as testing $\tau_1 = \tau_2 = \tau_3 = 0$

Multicollinearity

- Linear dependence among regressor variables (correlation 1)
- High dependency amongst regressor variables means $X'X$ will be near singular so β's will be estimated inaccuractely
- If X_1, X_2 are columns of X and $X_1^T X_2 = 0$, the two variables are orthogonal and there is no linear relationship between them
- Squared distance from $\hat{\beta}$ to true β $L_1^2 = (\hat{\beta} - \beta)'(\hat{\beta} - \beta)$.

$$E(L_1^2) = \sum_{j=1}^{p} \text{Var}(\hat{\beta}_j) = \sigma^2 \text{Tr}[(X'X)^{-1}] = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}$$

So small eigenvalues result in poorly estimated β

- Small eigenvalue $\implies$ high dependency among columns of X
- Centering of regressor variables can help reduce multicollinearity
- Sources: data collection method (only a subspace of the region is sampled), constraints on model or population, model specification, overdefined model (more regressor variables than observations)

Diagnostics

- Examination of correlation matrix. x_i, x_j nearly lin dep $\implies |r_{ij}|$ near unity, but only helpful for detecting between pairs of regressors
- VIFs. See above.
- Eigensystem of $X'X$. $E(L_1^2)$ will be large if at least on eigenvalue is small. Condition number of $X'X$ $\kappa = \frac{\lambda_{max}}{\lambda_{min}}$
  - $\kappa < 100$ – no serious problem
  - $100 \leq \kappa \leq 1000$ moderate to strong
  - $\kappa > 1000$ severe

  Condition indices of $X'X$ $\kappa_j = \frac{\lambda_{max}}{\lambda_j}$. $\kappa_j > 1000$ indicates near linear dependencies in $X'X$

Remedies

- Collection additional data
- Model respecification
  - Regressor variable elimination. If x_1, x_2 highly correlated, drop one of them. Not satisfactory if one variable has significant explanatory power
  - New regressor variable as a function of linearly dependent variables
- Ridge regression: $(X'X + kI)\hat{\beta}_R = X'y$, $k \geq 0$
  - Ridge of $X'X$ – its diagonal elements
  - Ridge estimator is lin transf of least sq estimator

$$\hat{\beta}_R = (X'X + kI)^{-1}X'y = (X'X + kI)^{-1}X'X\hat{\beta} = Z_k\hat{\beta}$$

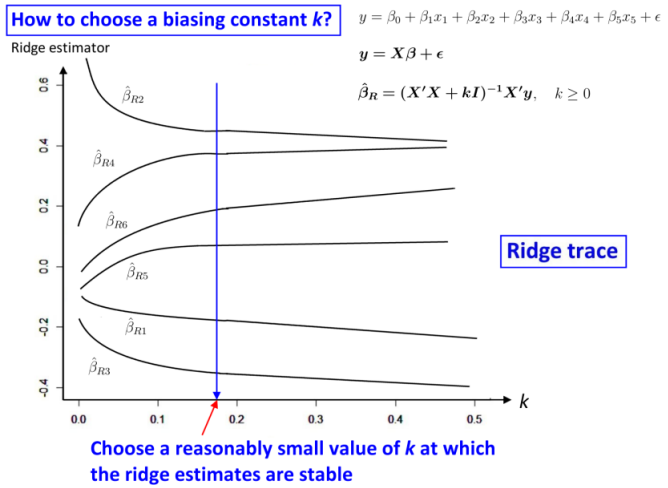
  - $\hat{\beta}_R$ is a biased estimator of β since $E(\hat{\beta}_R) = Z_k\beta \neq \beta$ unless $k = 0$
  - $\text{Var}(\hat{\beta}_R) = \sigma^2(X'X + kI)^{-1}X'X(X'X + kI)^{-1}$

$$\begin{aligned}MSE(\hat{\beta}) &= E[(\hat{\beta} - \beta)^2] = \text{Var}(\hat{\beta}) + (E(\hat{\beta}) - \beta)^2 \\ MSE(\hat{\beta}_R) &= E[(\hat{\beta}_R - \beta)'(\hat{\beta}_R - \beta)] \\ &= \text{tr}[\text{Var}(\hat{\beta}_{R,j}) + (E(\hat{\beta}_R) - \beta)'(E(\hat{\beta}_R - \beta) \\ &= \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2} + k^2\beta'(X'X + kI)^{-2}\beta \\ SS_{Res} &= (y - X\hat{\beta})'(y - X\hat{\beta}) + (\hat{\beta}_R - \hat{\beta})'X'X(\hat{\beta}_R - \hat{\beta})\end{aligned}$$

Ridge regression estimates can be computed using OLS:
$y_A = (y, 0_p)^T$, $X_A = (X, \sqrt{k}I_p)^T$,
$\hat{\beta}_R = (X_A'X_A)^{-1}X_A'y_A = (X'X + kI_p)^{-1}X'y$

Choosing k

- Make a ridge trace i.e. plot elements of ridge estimate $\hat{\beta}_R$ vs k for $0 < k < 1$
- As k increases, ridge estimates will vary but stabilise for larger values of k
- Choose reasonably small value of k at which ridge estimates are stable



Principal component regression

- Arrange eigenvalues in decreasing order, set the last n that are near zero to be zero. Perform least sq on resulting multiple linear regression model
- Ridge regression does not remove dependency among regressor variables, it just makes $X'X$ less singular
- Principal component regression does remove dependency because it removes the corresponding eigenvectors
- Scaling of regressor variables is not required for the two methods to work

Variable selection and model building

- K regressor variables, r deleted, $p = K + 1 - r$ retained. Full model written as $y = X_p\beta_p + X_r\beta_r + \epsilon$. Subset model given by $y = X_p\beta_p + \epsilon$
- $\hat{\beta}_p$ is a biased estimator of β_p unless $\beta_r = 0$ or $X_p'X_r = 0$
- $\text{Var}(\hat{\beta}_p) = \sigma^2(X_p'X_p)^{-1}$, $\text{Var}(\hat{\beta}^*) = \sigma^2(X'X)^{-1}$
- $\hat{y}$ is a biased estimate of $x_p\beta_p$ unless $x_p'A\beta_r = 0$ which is only true if $X_p'X_r\beta_r = 0$
- $\text{Var}(\hat{y}^*) \geq MSE(\hat{y})$. $\hat{y}$ has smaller variance than that of the full model
- Choosing subsets: add regressor variables until it provides only a small increase in R_p^2, or choose p that minimises MS_{Res}

Stepwise regression methods

- **Forward selection**: Start with no regressor variables. Choose small α_{IN}
- Fit models with one variable. Calculate p-value using $F = SS_R(x_j)/MS_{Res}(x_j)$ for each model, and only consider models with p-value $< \alpha_{IN}$. Add the one with the smallest p-value.
- Repeat with two variables using $F = SS_R(x_j \mid x_1)/MS_{Res}(x_1, x_j)$
- **Backward selection**: Same as forward, but start with K regressor variables and eliminate the one with highest p-value
- **Stepwise regression**: Start with forward selection, then use backward selection to check whether previous variables can be removed