# ST3131 Regression Analysis

Lingjie, April 27, 2021

## ATQ

### Comment on relationship using Scatter Plot

1. There is/isn't a
2. positive/ negative
3. linear/ non-linear relationship between $x$ and $y$
4. with a constant/ non constant Variance

### Fit a linear model

1. Write down fitted model
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 I(X_2 = 1)$$
2. Note: if cat var, write $I(X = \text{Type A})$

### Test model

1. Single regressor: t-test, Multiple regressors: F-test
2. State $H_0, H_1$
3. Find t-stat/ F-stat, p-value
4. Conclude reject/ do not reject $H_0$

### Comment on model Fit

1. $R^2$ is large/ small
2. F-test is significant/ not significant

### Checking on assumptions

1. Residual plot shows residuals randomly scattered/ some trend about 0, a linear model is sufficient/ insufficient
2. Residuals plot shows constant/ non constant variance
3. Standardised residuals shows there are/ no outliers
4. Residual Normal probability plot shows residuals violate/ does not violate normality assumption
5. VIF is small suggesting there is/ no multicollinearity

## Hypothesis testing

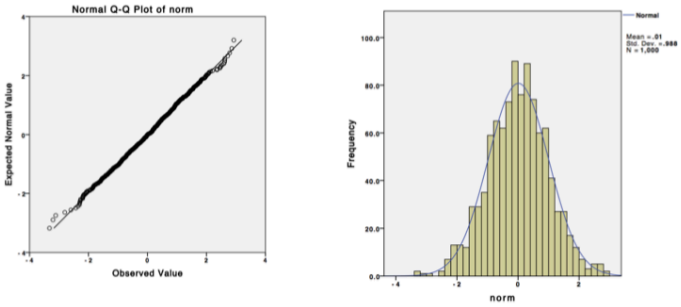
Statistical Inference: using probability to quantify how plausible a value for a parameter is

1. Assumptions
   - Randomised data
   - Required sample size
   - Shape of distribution
   - Note: transformed regressor need to be transformed back
2. Hypothesis
   - $H_0$ := null hypothesis, represents no effect
   - $H_1$ := alternative hypothesis, effect of some type
   - $H_1 : p > 0.5, H_1 : p < 0.5$ := one-sided alternative hypothesis
   - $H_1 : p \neq 0.5$ := two-sided alternative hypothesis
   - Note: $H_0, H_1$ cannot have any overlap
3. Test Statistics
   - Describes how far that point estimate falls from the parameter value given in $H_0$
   - Required:
     1. value in the sample,
     2. parameter value in $H_0$
4. $p$-value
   - Probability in sampling distribution that test statistic assumes a value more extreme (tail probability) than observed
   - Small $p$-value, either reject $H_0$ or sample is not representative of the population
   - Note: for $F$-test, it is only the right tail
5. Conclusion
   Reports the $p$-value, and summarises how much evidence there is against $H_0$

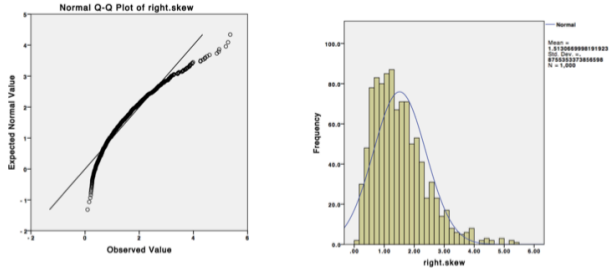
## Checking Normality using QQ-plots

QQ-plots: standardised sample quantiles vs theoretical quantiles of a $N(0,1)$ distribution. When compared to the straight line:

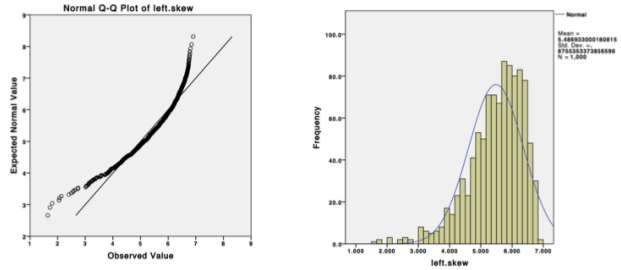
```
qqnorm(X)   # qq plot
qqline(X)   # line
```



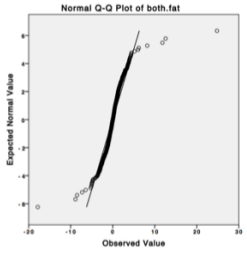

- Left tail shorter, Right tail longer: Right-skewed

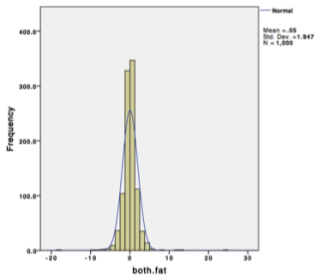



- Left tail longer, Right tail shorter: Left-skewed

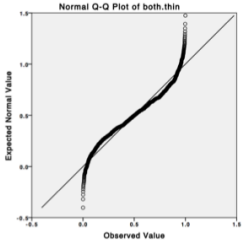


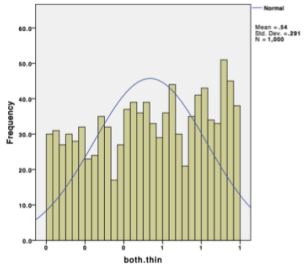

- Both tails longer




- Both tail shorter




- Right tail is below: longer than Normal
- Right tail is above: shorter than Normal
- Left tail is below: shorter than Normal
- Left tail is above: longer than Normal

Alternative method: Anderson-Darling test
$H_0$ : data is normal vs $H_1$ : data is not normal

## Hypothesis Testing for Means

$\mu :=$ population mean

### One-Sample t-Test for Means

1. Assumption
   - variable measured is quantitative
   - data obtained through randomization
   - population distribution is approximately Normal, crucial when $n$ is small
2. Hypothesis
   $H_0 : \mu = \mu_0$, $H1 : \mu \neq \mu_0$
3. Test Statistics
   point estimate: $\bar{X}$
   $T = \frac{\bar{X}-\mu_0}{s/\sqrt{(n)}} \sim t_{n-1}$
4. $p$-Value

```r
t.test(data, mu=0) # default: 2-sided
```

5. Conclusion
   Report and interpret $p$-value in the context of the experiment

### Confidence Interval

95% Confidence Interval for the Mean
$\bar{X} \pm t_{n-1,0.975} \times s/\sqrt{n}$
$\bar{X} \pm t_{n-1,0.975} \times SE(\bar{X})$

## Comparing Means

### Tests for Comparing Two Group Means

Not the focus of this module, but we can use
$H_0 : \mu_1 = \mu_2 \Rightarrow \mu_1 - \mu_2 = 0$ and apply $t$-test

### ANOVA

Let $I :=$ num of groups and $J :=$ num of obs in each group

1. Assumption
   errors are normally distributed
2. Hypothesis
   $H_0 : \beta_1 = \beta_2 = \cdots = \beta_I = 0$
3. Test Statistics
   $F = \frac{SS_B/(I-1)}{SS_W[I(J-1)]} \sim F_{I-1,I(J-1)}$
4. $p$-value
   reject $H_0$ if $F > F_{I-1,I(J-1)}(\alpha)$
   Note: F-stat test only on the right tail, even if test is 2-sided
5. Conclusion

# Linear Regression

```r
model <- lm(y~x+I(x^2)+log(x), data=data)
summary(model) # regression results
anova(model) # ANOVA table
predict(model, data.frame(x1=x),
  interval="confidence", level=0.95)
predict(model, data.frame(x1=x),
  interval="prediction", level=0.95)
qt(alpha/2, df) # quantile
pt(t0, df)  # probability
```

| | |
|---|---|
| Population model | : $\mathbf{y_{n\times 1}} = \mathbf{X_{n\times p}}\beta_{\mathbf{p}\times\mathbf{1}} + \epsilon_{\mathbf{n}\times\mathbf{1}}$ |
| Fitted model | : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$ |
| Hat matrix ($\mathbf{H}$) | : $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{y}$ |
| | : $\mathbf{Hy}$ |
| Sample regression model | : $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ |

Note:
linearity refers to parameters (coefficients)
fitted model has no error term

Main Objective of Regression Analysis

- model fitting
  estimate unknown parameters (OLS, MLE)
- model adequacy checking
  validate appropriateness of the model

### Notations

| | |
|---|---|
| $\epsilon$ | := random (uncorrelated) error term |
| | := $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$ |
| $y_i$ | := $i$th dependent (response/target/output) variable |
| $x_{ij}$ | := independent (predictor/input/covariate/regressor) |
| | := $i$th observation of regressor $x_j$ |
| $k$ | := num of regressors, excl $\beta_0$ |
| $p$ | := $k+1$, incl $\beta_0$ |
| $\beta_0$ | := intercept, always assumed exist |
| $\beta_i$ | := slope |
| $E(y\|x)$ | := $\mu_{y\|x} = E(\beta_0 + \beta_1 x + \epsilon\|x) = \beta_0 + \beta_1 x$ |
| $Var(y\|x)$ | := $\sigma^2_{y\|x} = Var(\beta_0 + \beta_1 x + \epsilon\|x) = \sigma^2$ |

## Estimation of the Model Parameters

### Simple regression

Objective:

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

Solution:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i - \frac{(\sum_{i=1}^n y_i)(\sum_{i=1}^n x_i)}{n}}{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}} = \frac{Cov(X,Y)}{Var(X)} = \frac{S_{xy}}{S_{xx}}$$

$$Var(y) = \frac{SS_T}{n-1}$$

$$Var(X) = \frac{S_{xx}}{n-1}$$

Notation:

$$S_{xy} = \sum_{i=1}^n y_i x_i - \frac{(\sum_{i=1}^n y_i)(\sum_{i=1}^n x_i)}{n} = \sum_{i=1}^n y_i(x_i - \bar{x})$$

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

## Multiple regression

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}_{n\times p}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}_{p\times 1} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}_{p\times 1}$$

Objective:

$$\min S(\beta) = \sum_{i=1}^n \epsilon^2 = \epsilon^T \epsilon = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$
$$= \mathbf{y}^T\mathbf{y} - 2\beta^T\mathbf{X}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta$$

Note: $(\beta^T\mathbf{X}^T\mathbf{y})_{1\times 1}$ is a scalar

Solution:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$(\mathbf{X}^T\mathbf{X})^{-1}$ exist $\Leftrightarrow$ regressors are linearly independent

## Properties of the Least-Squares Estimation

Key: unbiased mean + small variance
$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^n \frac{(x_i - \bar{x})}{S_{xx}} y_i = \sum_{i=1}^n c_i y_i$
$\because \sum_{i=1}^n c_i = 0, \sum_{i=1}^n c_i x_i = 1, \sum_{i=1}^n c_i^2 = \frac{1}{S_{xx}}$

$$E(\hat{\beta}_1) = \beta_1, \quad E(\hat{\beta}_0) = \beta_0$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}, \quad Var(\hat{\beta}_0) = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})$$

## Mean and Variance of $\hat{\beta}$

$\because E(\epsilon) = 0, (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{X} = \mathbf{I}$
$E(\hat{\beta}) = E[(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{y}] = E[(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}(\mathbf{X}\beta + \epsilon)] =$
$E[(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{X}\beta + (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\epsilon] = \beta$
$Cov(\hat{\beta}) = E\left\{[\hat{\beta} - \mathbf{E}(\hat{\beta})][\hat{\beta} - \mathbf{E}(\hat{\beta})]^T\right\}_{p\times p}$ symmetric matrix
the $j^{th}$ diagonal element is $Var(\hat{\beta}_i)$ and $(ij)$th off-diagonal element is $Cov(\hat{\beta}_i, \hat{\beta}_j)$ Hence
$Cov(\hat{\beta}) = Var(\hat{\beta}) = Var\left[(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{y}\right] =$
$(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}Var(\mathbf{y})\left[(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^T\right]^T =$
$\sigma^2(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-\mathbf{1}} = \sigma^2(\mathbf{X}^\mathbf{T}\mathbf{X})^{-\mathbf{1}}$
Denote $\mathbf{C} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-\mathbf{1}}$, then $Var(\hat{\beta}_j) = \sigma^2 C_{jj}$ and $Cov(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 C_{ij}$

$$E(\hat{\beta}) = \beta, \quad Cov(\hat{\beta}) = \sigma^2(\mathbf{X}^\mathbf{T}\mathbf{X})^{-\mathbf{1}}$$
$$Var(\hat{\beta}_j) = \sigma^2 C_{jj}, \quad Cov(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 C_{ij}$$

## Estimation of $\sigma^2$

Denote $SS_T = \sum_{i=1}^n (y_i - \bar{y})^2$ (not model dependent)

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

$\because E(SS_{Res}) = (n-p)\sigma^2, p :=$ num of regressors (includ $\beta_0$)
$\therefore \hat{\sigma}^2 = SS_{Res}/(n-p)$, unbiased estimator of $\sigma^2$
Denote $MS_{Res} = SS_{Res}/(n-p)$, residual mean square
$\hat{\sigma} :=$ residual standard error or standard error of regression

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

## Estimation of $\sigma^2$, multiple regression

$SS_{Res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 = \mathbf{e}^\mathbf{T}\mathbf{e}$
Sub $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\beta}$
$SS_{Res} = (\mathbf{y} - \mathbf{X}\hat{\beta})^\mathbf{T}(\mathbf{y} - \mathbf{X}\hat{\beta})$
$= \mathbf{y}^\mathbf{T}\mathbf{y} - \hat{\beta}^\mathbf{T}\mathbf{X}^\mathbf{T}\mathbf{y} - \mathbf{y}^\mathbf{T}\mathbf{X}\hat{\beta} + \hat{\beta}^\mathbf{T}\mathbf{X}^\mathbf{T}\mathbf{X}\hat{\beta}$
$= \mathbf{y}^\mathbf{T}\mathbf{y} - \mathbf{2}\hat{\beta}^\mathbf{T}\mathbf{X}^\mathbf{T}\mathbf{y} + \hat{\beta}^\mathbf{T}\mathbf{X}^\mathbf{T}\mathbf{X}\hat{\beta}$
Since $\mathbf{X}^\mathbf{T}\mathbf{X}\hat{\beta} = \mathbf{X}^\mathbf{T}\mathbf{y} \Rightarrow SS_{Res} = \mathbf{y}^\mathbf{T}\mathbf{y} - \hat{\beta}^\mathbf{T}\mathbf{X}^\mathbf{T}\mathbf{y}$
$MS_{Res} = \frac{SS_{Res}}{n-p} \Rightarrow E(MS_{Res}) = \sigma^2 \Rightarrow \hat{\sigma}^2 = MS_{Res}$ is unbiased estimator of $\sigma^2$

## Hypothesis Testing

$$y_i \sim N(\beta_0 + \sum_{i=1}^k \beta_j x_{ij}, \sigma^2)$$
$$\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}^\mathbf{T}\mathbf{X})^{-\mathbf{1}}), \hat{\beta}_j \sim N(\beta_j, \sigma^2 C_{jj})$$

Assumption:

- $x, y$ shares linear relationship
- uncorrelated errors with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$
- $\epsilon_i \sim N(0, \sigma^2)$
  Note: for each sub-population, they have the same variance

### $t$ test: Individual Regression Coefficient

$H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{MS_{Res}/S_{xx}}} = \frac{\beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim t_{n-p}$$

$C_{jj}$ is diagonal element $(\mathbf{X}^\mathbf{T}\mathbf{X})^{-\mathbf{1}} \Leftrightarrow \hat{\beta}_j$
$H_0 : \beta_0 = 0, H_1 : \beta_0 \neq 0$

$$t_0 = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)} = \frac{\hat{\beta}_0}{\sqrt{MS_{Res}(1/n + \bar{x}^2/S_{xx})}} \sim t_{n-p}$$

rejected $H_0$ if $|t_0| > t_{n-k-1(\alpha/2)}$
Note: the test is given other regressors in the model

### $F, t$ test: Significance of Regression

$H_0 : \beta_1 = \beta_2 = \cdots \beta_j = 0, H_1 :$ at least 1 $\beta_i \neq 0$

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

Reject $H_0$ if $F_0 > F_{k,n-k-1}(\alpha)$

ANOVA table

| Source of Variation | Sum of Squares | DF | MS | $F_\alpha$ |
|---|---|---|---|---|
| Regression | $SS_R$ | k | $MS_R$ | $MS_R/MS_{Res}$ |
| Residual | $SS_{Res}$ | n-p | $MS_{Res}$ | |
| Total | $SS_T$ | n-1 | | |

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

### Extra-sum-of-squares Method

Useful to select a subset of regressors
Extra-sum-of-square due to
$\beta_2 : SS_R(\beta_2|\beta_1) = SS_R(\beta) - SS_R(\beta_1)$ with $df = r$
$H_0 : \beta_2 = 0, H_1 : \beta_2 \neq 0$

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

Note: in R the order we put regressors matters
$t^2 = F_0 \sim F_{1,n-p}$

### ESS: Orthogonal Col in X

consider $\mathbf{y} = \mathbf{X}\beta + \epsilon = \mathbf{X_1}\beta_1 + \mathbf{X_2}\beta_2 + \epsilon$, if $\mathbf{X_1} \perp \mathbf{X_2}$, Sum of square and model is the same regardless of order to add in regressors ⇒ no need to refit (rare case)

## Confidence Intervals

### CI on the Regression Coefficients

$$\beta_i \in \hat{\beta}_i \pm t_{n-p}(\alpha/2) \times se(\hat{\beta}_i)$$
$$\sigma^2 \in \left( \frac{(n-p)MS_{Res}}{\chi^2_{n-p}(\alpha/2)}, \frac{(n-p)MS_{Res}}{\chi^2_{n-p}(1-\alpha/2)} \right)$$

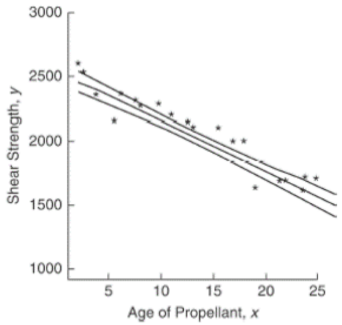
### CI Estimation of the Mean Response

Mean Response: $E(y|x_0)$
$E(\hat{y}|x_0) = \hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$
$Var(\hat{\mu}_{y|x_0}) = Var(\hat{\beta}_0 + \hat{\beta}_1 x_0) = Var(\bar{y} + \hat{\beta}_1(x_0 - \bar{x}))$
$= \sigma^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) = \sigma^2 \mathbf{x_0^T}(\mathbf{X^T X})^{-1}\mathbf{x_0}$

$$E(y|x_0) \in \hat{\mu}_{y|x_0} \pm t_{n-p}(\alpha/2) \sqrt{MS_{Res}\left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}$$
$$E(y|x_0) \in \hat{y}_0 \pm t_{n-p}(\alpha/2) \sqrt{\hat{\sigma}^2 \mathbf{x_0^T}(\mathbf{X^T X})^{-1}\mathbf{x_0}}$$

Note

- CI for $E(y|x_0)$ is a function of $x_0$
- interval width minimised at $x_0 = \bar{x}$ and widens as $|x_0 - \bar{x}|$ increases
- standard error for a CI on mean responses takes into account the uncertainty due to sampling ⇒ confidence band is not parallel




### Simultaneous CI on Regression Coefficients

Bonferroni Method, useful when trying to find the joint CI of $\beta_i, \beta_j, i \neq j$

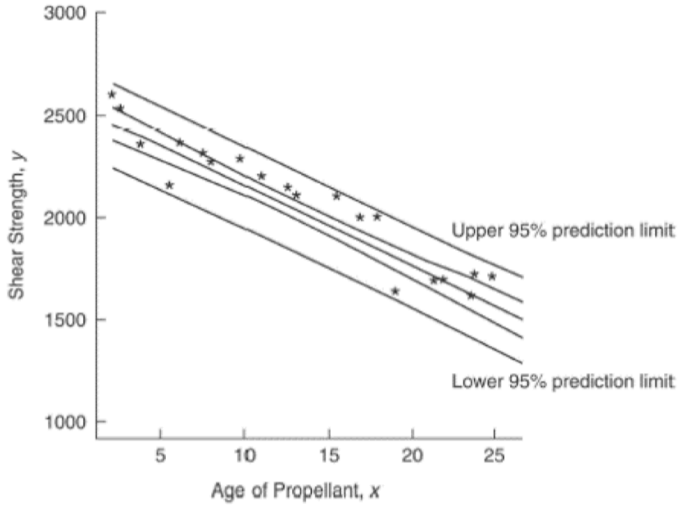
$$\hat{\beta}_j \pm t_{n-p}(\alpha/2p) \times se(\hat{\beta}_j), j \in [0, k]$$

### Prediction Interval

Note: check if $X$ is in the range of dataset
Prediction: $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$, let $\psi = y_0 - \hat{y}_0$
Note: Prediction interval is for the future observation $y_0$ can be obtained, and future observation $y_0$ is independent of our prediction $\hat{y}_0$
$Var(\psi) = Var(y_0) - 2Cov(y_0, \hat{y}_0) + Var(\hat{y}_0)$
$= \sigma^2 + \sigma^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$

$$y_0 \in \hat{y}_0 \pm t_{n-p}(\alpha/2) \sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$
$$y_0 \in \hat{y}_0 \pm t_{n-p}(\alpha/2) \sqrt{\hat{\sigma}^2(1 + \mathbf{x_0^T}(\mathbf{X^T X})^{-1}\mathbf{x_0})}$$

Note: standard error for a prediction interval of a future observation takes into account the uncertainty due to sampling and variability of the individual around the predicted mean.




## Coefficient of Determination $R^2$

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$$
$$R^2_{Adj} = 1 - \frac{SS_{Res}/(n-p)}{SS_T/(n-1)}$$
$$= 1 - \frac{n-1}{n-p}(1 - R^2)$$
$$cor(y, \hat{y}) = |cor(y, x)| = \sqrt{R^2}$$

proportion of variation explained by the regressors

## Interpretation of Regression Coefficients

### First Interpretation

changes in mean response $\frac{\partial}{\partial x_i} E(y|X)$

### Second Interpretation

contribution of $x_j$ to $y$ after both $y, x_j$ have been linearly adjusted for all other regressors
$e_{y,x_1,x_2,\cdots,x_{k-1}} = \beta_k e_{x_k,x_1,x_2,\cdots,x_{k-1}}$

## Hidden Extrapolation in Multiple Regression

### Regressor Variables Hull (RVH)

Using the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T}$
and set $\max h_{ii} := h_{max}$
To be within the ellipsoid, $\mathbf{x^T}(\mathbf{X^T X})^{-1}\mathbf{x} \leq h_{max}$
Therefore, the new estimation at point
$\mathbf{x_0^T} = [1, x_{01}, x_{02}, \cdots, x_{0k}]$ will have $h_{00} = \mathbf{x_0^T}(\mathbf{X^T X})^{-1}\mathbf{x_0}$
if $h_{00} > h_{max}$, then the point is a point of extrapolation

## Indicator Variables

Categorical variables, we assign dummy variables, $\alpha - 1$

$$x = \begin{cases} 0 & \text{if obs from type A} \\ 1 & \text{if obs from type B} \end{cases} = I(\text{obs from type B})$$

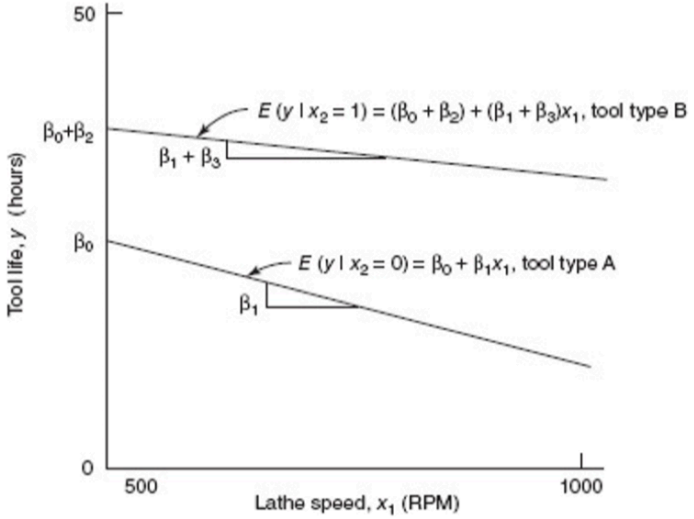
Same slope, different intercept for different type
To test for significance of 3 types:

$$H_0 : \beta_2 = \beta_3 = 0$$
$$H_1 : \beta_2 \neq 0 \text{ and/or } \beta_3 \neq 0$$

```r
# recode to cat var
data$x11 <- as.factor(data$x11)
```

## Interaction Term

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$
$\frac{\delta}{\delta x_1} E(y|x_1, x_2) = \beta_1 + \beta_3 x_2$




To test if diff types are identical:
$H_0 : \beta_2 = \beta_3 = 0$ vs $H_1 : \beta_2$ and/or $\beta_3 \neq 0$
Note: we should not include $x_1 x_2$ if there is no $x_2, x_1$

## Standardised Regression Coefficients

### Unit normal Scaling

$$x_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, i \in [1, n], j \in [1, k]$$

$\bar{x}_j = (\sum_{i=1}^n x_{ij})/n, s_j^2 = (\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2)/(n-1)$
Note: model using scaled response and scaled regressors does not have intercept

### Unit Length Scaling

$$w_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{S_{jj}}}, i \in [1, n], j \in [1, k]$$

$S_{jj} = \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 = (n-1)s_j^2$
Note: each new regressor $w_j$ has mean $\bar{w}_j = 0$ and length $\sqrt{\sum_{i=1}^n (w_{ij} - \bar{w}_j)^2} = 1$
The resultant $\mathbf{W^T W}$ matrix is a correlation matrix where $r_{lj}$ is simple correlation between $x_l, x_j$, $l, j = 1, \cdots, k, l \neq j$

## Other topics

### Regression Through the Origin

$y = \beta_1 x + \epsilon$, only appropriate when origin $(0, 0)$ has meaning
$\min S(\beta_1) = \sum_{i=1}^n (y_i - \beta_1 x_i)^2$
$\hat{\beta}_1 = \sum_{i=1}^n y_i x_i / \sum_{i=1}^n x_i^2$
$\hat{\sigma}^2 = MS_{Res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 / n - 1$
$= \left(\sum_{i=1}^n y_i^2 - \hat{\beta}_1 \sum_{i=1}^n y_i x_i\right) / (n-1)$

### Estimation by Maximum Likelihood

Only when the response (or error) distribution is known
i.e. $\epsilon \sim N(0, \sigma^2) \Rightarrow y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$
$\max L(y_i, x_i, \beta_0, \beta_1, \sigma^2) =$
$\prod_{i=1}^n (2\pi\sigma^2)^{-1/2} exp\left[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right]$
$= (2\pi\sigma^2)^{-n/2} exp\left[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right]$
$\Leftrightarrow \max \log L =$
$-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$
$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
$\hat{\beta}_1 = \sum_{i=1}^n y_i (x_i - \bar{x}) / \sum_{i=1}^n (x_i - \bar{x})^2$
$\hat{\sigma}^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 / n = [(n-2)/n]\hat{\sigma}^2_{OLS}$, biased

# Model Adequacy Checking

Model assumption:

- Linearity assumption $\mathbf{Y} = \beta\mathbf{X}$
- Normality assumption $\epsilon_i \sim N$
- $E(\epsilon) = 0$
- Homogeneity assumption $Var(\epsilon) = \sigma^2$
- Independent error assumption $Cov(\epsilon_i, \epsilon_j) = 0$

## Graphical Methods before model fitting

- One-dimensional graphs

  Histogram, Stem-and-leaf-display, Dot plot, Box plot (for outlier detection)

  Check for: 1) distribution (symmetric or skewed) to determine if transformation is required 2) outliers

- Two-dimensional graphs

  pairwise plots (e.g. scatter plots)

  Check for: 1) Multicollinearity 2) Linear relationship $y \sim x$

- Rotating plots
- Dynamic graphs

## Residual Analysis

Diagnostic methods primarily based on studying model residuals $(e_i = y_i - \hat{y}_i)$

$$E(e_i) = 0$$
$$Var(\bar{e}) = \frac{\sum_{i=1}^n (e_i - \bar{e})^2}{n-p} = \frac{\sum_{i=1}^n e_i^2}{n-p} = \frac{SS_{Res}}{n-p} = MS_{Res}$$

The residuals are not independent. However, when $p << n$ the nonindependence has little effect

### Hat Matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$
$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$
$$\hat{y}_i = h_{i1}y_1 + \cdots + h_{in}y_n$$
$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

$\hat{y}_i$ := weighted sum of all given observation
$h_{ii}$ := leverage value for $i$th observation
:= weight given to $y_i$ in determining the $i$th fitted values $\hat{y}_i$
$\mathbf{H}$ := is symmetric and idempotent $(\mathbf{HH} = \mathbf{H})$
$\mathbf{I} - \mathbf{H}$ := is symmetric and idempotent

### Variance of Residuals

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})(\mathbf{X}\beta + \epsilon)$$
$$= (\mathbf{I} - \mathbf{H})\epsilon$$
$$Var(\epsilon) = \sigma^2\mathbf{I}$$
$$Var(\mathbf{e}) = (\mathbf{I} - \mathbf{H})Var(\epsilon)(\mathbf{I} - \mathbf{H})' = \sigma^2(\mathbf{I} - \mathbf{H})$$
$$Var(e_i) = \sigma^2(1 - h_{ii})$$
$$Cov(e_i, e_j) = -\sigma^2 h_{ij}$$

## Standardized Residual

$$\frac{e_i}{\sigma\sqrt{(1-h_{ii})}}$$

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

Estimate $\sigma^2$: $MS_{Res}$ or $\hat{\sigma}^2_{(i)}$, both unbiased estimator of $\sigma^2$
$SS_{Res(i)}$ := sum of squared residual when fit model with $(n-1)$ obs without $i$th obs

## Studentized Residuals

Internally studentized residuals

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

```
rstandard(model)
```

Externally studentized residuals

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

$$r_i^* = r_i\sqrt{\frac{n-p-1}{n-p-r_i^2}}$$

Since both internal and external studentized residuals are monotone transformation of another, either is fine

## Residual plots

Normal Probability Plot (QQ plot)

- Plot ordered S.R (x-axis) vs cumulative probability or normal scores (y-axis)

  Expect: normally distributed residuals, nearly straight line with an intercept of 0 and slope 1

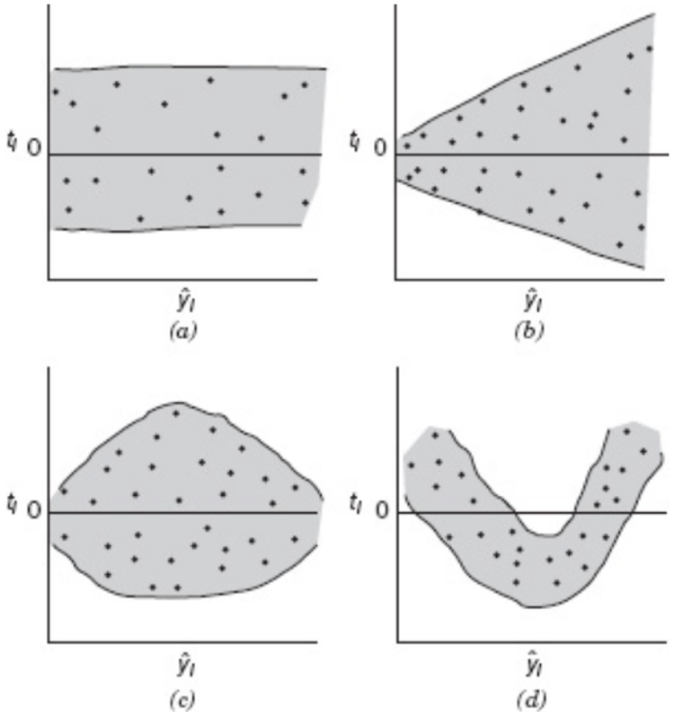
  Note: common defect is due to large residuals arise from outliers

Scatter plot of S.R. vs fitted values

- Plot residuals vs fitted values

  Expect: random scatter plot

  Note: observe for potential patterns, such as non-constant variance, nonlinearity




a ideal

b non constant variance (heterogeneity)

c variance follow binomial distribution

d non linear relationship

Scatter plots of S.R. vs each predictor

- Plot S.R. with each regressors

  Except: randomly scattered plot

  Note: transformation might be needed for the regressors

Plot of Residuals in Time Sequence

- Plot residuals against time order

  Expect: randomly scatter plot

  Note: if not random, variance might be changing with time then linear or quadratic terms in time should be added to the model

- Autocorrection (errors are related to each other) is serious violation of basic regression assumptions

## Comments on Residual plots

- Strong indication of linear relationship between variables
- Normal probability plot indicates a deviation from normality assumption, it also shows potential outliers
- plot of residuals vs fitted values has large residuals, suggesting outliers
- Scatter plot between regressors show linear relationship between them, a deeper investigation about data is needed. There might be ommitted variable bias (factors that linked to both regressors such as location), extra variables should be added to fix this

## Detection and Treatment of Outliers

Outlier:

- extreme observation, one that is considerably different from the majority of data
- Residuals >> 3 sd from mean indicate potential $y$ space outliers
- 1) Examine Residual plots against fitted values
- 2) QQ-plot
- "bad" values: result of unusual but explainable events, can be deleted and corrected

  e.g. faulty measurement or analysis, incorrect recording data, failure of measuring instrument

- "normal" values: unusual but perfectly pausible observation

  should be kept since deleting will provide false sense of precision in estimation/prediction

Observe (compare with and without outliers):

- Changes in goodness of fit ($R^2$, $MS_{Res}$)
- Effect on model (magnitude and sign of $\beta$)

## Lack of Fit of the Regression Model

Formal test for Lack of fit (only for simple model)

$$y_{ij} - \hat{y}_i = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \hat{y}_i)$$

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

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

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

Assumption: normality, independence and constant variance. Test only the first order relationship

$$H_0 : \beta_1 \neq 0$$

$m$ := num of levels ($x$ into $m$ groups)
$n$ := $\sum_{i=1}^{m} n_i$ total observation
$ij$ residual := $y_{ij} - \hat{y}_i$
$SS_{PE}$ := sum of squaress due to pure error
$SS_{LOF}$ := sum of squares due to lack of fit

## Leverage and Influence

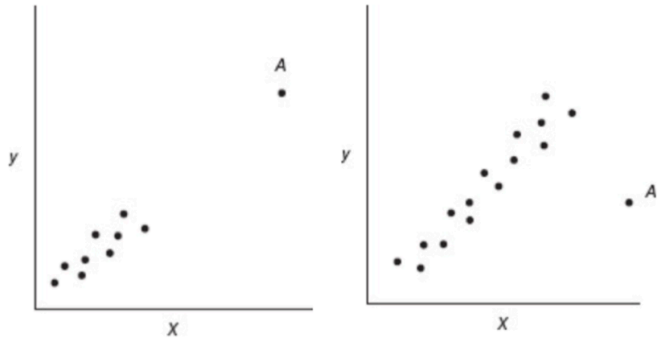



Figure: Examples of a leverage point (left) and an influential point (right).

### Diagnostics for Leverage

$$h_{ii} = x_i(\mathbf{X}'\mathbf{X})^{-1}x_i'$$

- standardised distance of $i$th observation from the center of the $x$ space
- the amount of leverage exerted by $j$th observation $y_j$ on the fitted value $\hat{y}_i$.
- Large $h_{ii}$ indicate potential influential
- Not all leverage points are influential points. Should consider $h_{ii}$ with standardized residuals. Large for both are likely to be influential
- since $\sum h_{ii} = rank(\mathbf{H}) = rank(\mathbf{X}) = p$, average size of $h_{ii}$ is $\bar{h} = p/n$
- $h_{ii} > 2p/n$ is consider leverage point (for small sample where $2p/n > 1$ then this cutoff doesn't apply)

```
# include beta0
x <- cbind(c(rep(1, n), x1, x2)
# hat matrix
hat <- x%*%solve(t(x)%*%x)%*%t(x)
which(diag(hat)>(2p/n))
```

### Cook's Distance

Measure of influence
$\beta_{(i)}$ := beta without the $i$th data
Usual choice of $\mathbf{M} = \mathbf{X}'\mathbf{X}, c = pMS_{res}$

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

Points with large $D_i$ have considerable influence on the least square estimate $\hat{\beta}$

$$D_i = \frac{r_i^2}{p}\frac{h_{ii}}{1 - h_{ii}} \sim F_{p,n-p}(\alpha)$$

Deleting point $i$ would move the estimate of $\beta$ to approximately the edge of a $F_{p,n-p}^{-1}(D_i)$ confidence region.

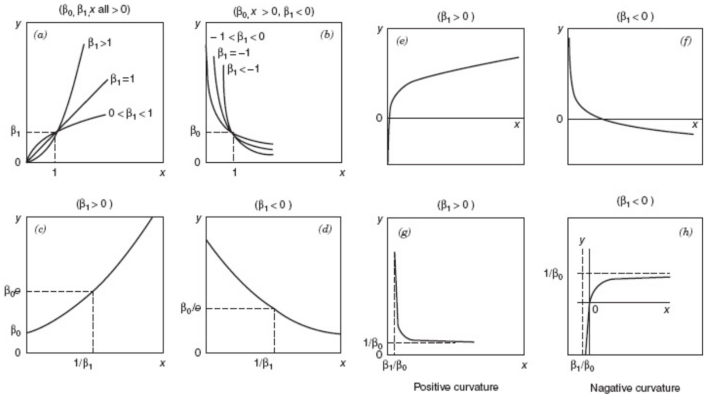
```
cook.distance(model)
```

### DFFITS, DFBETAS

- Statistic that indicate how much $\hat{\beta}_j$ changes in standard deviation if $i$ was deleted
- Number of standard deviation that the fitted value $\hat{y}_i$ changes if observation $i$ was removed

# To Correct Model Inadequacies

## Transformation to Linearize the model




| Figure | Linearizable Function | Transformation | Linear Form |
|---|---|---|---|
| 5.4a, b | $y = \beta_0 x^{\beta_1}$ | $y' = \log y, x' = \log x$ | $y' = \log \beta_0 + \beta_1 x'$ |
| 5.4c, d | $y = \beta_0 e^{\beta_1 x}$ | $y' = \ln y$ | $y' = \ln \beta_0 + \beta_1 x$ |
| 5.4e, f | $y = \beta_0 + \beta_1 \log x$ | $x' = \log x$ | $y' = \beta_0 + \beta_1 x'$ |
| 5.4g, h | $y = \frac{x}{\beta_0 x - \beta_1}$ | $y' = \frac{1}{y}, x' = \frac{1}{x}$ | $y' = \beta_0 - \beta_1 x'$ |

```
model <- lm(y~I(x^2)+I(1/X))
```

## Analytical Methods for Selecting a Transformation

### Transformation on $y$: Box-Cox Method

Power transformation $y^\lambda$ using maximum likelihood
Estimate $\beta, \lambda$

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

$$\dot{y} = exp[\frac{1}{n}\sum_{i=1}^{n} log(y_i)]$$

$$\mathbf{y}^{(\lambda)} = \mathbf{X}\beta + \epsilon$$

Transform $y = y^{(\lambda)}$ if $\lambda \neq 0$ else $log(y)$

```
library(MASS)
boxcox(model, lambda=seq(-2, 2, 0.5),
       optimize=TRUE, plotit=TRUE)
```

### Transformation on $X$: Box and Tidwell

Estimate $\beta_0, \beta_1, \alpha$

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

To estimate $\alpha$

1. Initial a model by least square method
   $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
2. fit a new model adding $w = xlog(x)$
   $\hat{y} = \hat{\beta}_0' + \hat{\beta}_1' x + \hat{\gamma} w$
3. then take
   $\alpha_1 = (\hat{\gamma}/\hat{\beta}_1) + 1$

Repeat procedure (1-3) with $x' = x^{\alpha_1}$
Procedure converge rapidly, often first stage result $\alpha$ is a satisfactory estimate of $\alpha$

# Generalized and Weighted Least Squares

## Weighted Least Squares

Linear regression models with nonconstant variance.
Deviation between $y_i, \hat{y}_i$ is multiplied by a weight
$w_i \propto 1/Var(y_i)$
WLS estimators are unbiased
$MS_{(w)RES}$ is unbiased estimator of $\sigma^2$

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_i)^2$$
$$w_i = \frac{1}{\sigma_i^2}$$
$$w_i = w_j \Leftrightarrow \sigma_i = \sigma_j$$
$$\hat{\beta}_0 = \bar{y}_w - \hat{\beta}_1 \bar{x}_w$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} w_i (x_i - \bar{y}_w)(y_i - \bar{x}_w)}{\sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2}$$
$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \quad \bar{y}_w = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

Benefits

- In transformed model, coefficient estimates are hard to interpret. Interpretation for WLS remain the same
- WLS can remove an observation by setting weight = 0
- WLS also can weight down outlier and influential point

## Generalized Least Squares

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$
$$E(\epsilon) = 0 \quad Var(\epsilon) = \sigma^2 \mathbf{V}$$

- OLS: $\mathbf{V} = \mathbf{I}$
- WLS: $\mathbf{V}$ is diagonal, $\mathbf{y}$ are uncorrelated but have unequal variances
- if off-diagonal $\mathbf{V}$ are nonzero, then observations are correlated

Least squares normal equation

$$(\mathbf{X'V^{-1}X})\hat{\beta} = \mathbf{X'V^{-1}y}$$
$$\hat{\beta} = (\mathbf{X'V^{-1}X})^{-1}\mathbf{X'V^{-1}y}$$

$\hat{\beta}$ is the generalized least squares estimate of $\beta$

# Multicollinearity

Multicollinearity occur when regressors are not orthogonal to each other
Let $X_j$ be the $j$th column of the matrix $X$, if

$$\sum_{j=1}^{k} t_j \mathbf{X_j} = 0$$

where $t_j$ are not all zero is approximately true then $\mathbf{X'X}^{-1}$ does not exist

## Sources of Multicollinearity

- Data collection method

  occur when only subsample of the entire sample space has been selected

  if positive correlation is strong enough, multicollinearity problem will occur

- Constraints on the model or in the population

  physical constraint regardless of collection method

- Model specification

  adding polynomial term of regressors

- Overdefined model

  fit more regressors than observations

  solve by: 1. using lesser regressors 2. do preliminary studies using subset of regressors 3. use principal components

## Effects of Multicollinearity

### Poor Coefficient Estimate

$\mathbf{C} := (\mathbf{X'X})^{-1}$ with only $k$ regressors
$L_1^2 :=$ squared distance between $\hat{\beta}$ and $\beta$

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

$\lambda_j :=$ $j$th eigenvalues of $\mathbf{X'X}$
Trace of matrix is sum of the eigenvalues of matrix
When multicollinearity presents, at least one of $\lambda_j$ is large.
Therefore, distance from $\hat{\beta}$ to $\beta$ is large.

### Large Variances and Covariances

if $\mathbf{X}$ is unit length scaled matrix

$$Cor(\mathbf{X}) = \mathbf{X'X}$$
$$[Cor(\mathbf{X})]^{-1} = \mathbf{C} = \frac{1}{1 - R_j^2}$$
$$Var(\hat{\beta}_j) = \sigma^2 C_{jj}$$

$R_j^2 :=$ multiple coefficient of determination from the regression of $x_j$ on the remaining $k-1$ regressor variables
If multicollinearity presents, one of $R_j^2$ will be large and $Var(\beta_j)$ will be large

## Multicollinearity Diagnostics

### Examination of the Correlation Matrix

Unit scaled $\mathbf{X}$ without intercept

$$r_{ij} \text{ of } \mathbf{X'X} = cor(\mathbf{X})$$

if $|r_{ij}|$ is close to 1, there might be multicollinearity.

### Variance Inflation Factors

Since $Var(\beta_j) = \sigma^2 C_{jj}$. $C_j j$ increase with VIF.

VIF is the factor by which variance increase due to near-linear dependence among the regressors

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

If $VIF_j > 5$, associated regression coefficient are poorly estimated due to multicollinearity

```
diag(solve(t(X)%*%X))
diag(solve(cor(X)))
```

### Eigensystem Analysis of X′X

eigenvalues of $\mathbf{A}_{k\times k}$ are the $k$ roots of the equation $|\mathbf{A}-\lambda\mathbf{I}|=0$

```
eigen(X)$values
```

small eigenvalues are indications of multicollinearity

$$\text{Kappa } k = \frac{\lambda_{max}}{\lambda_{min}}$$

$$k_j = \frac{\lambda_{max}}{\lambda_j}$$

- $K < 100$: no serious problem
- $100 < l < 1000$: moderate to strong multicollinearity
- $k > 1000$: strong multicollinearity

Eigensystem analysis can be used to identify the nature of the near-linear dependence in data (self-study)

## Methods for Dealing with Multicollinearity

- Collect more data
- Respecify the model
- Ridge Regression or Principal Component Regression

### Ridge Regression

$\mathbf{X}'\mathbf{X}$ in correlation form

$$\begin{aligned}(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I})\hat{\beta}_\mathbf{R} &= \mathbf{X}'\mathbf{y}\\ \hat{\beta}_\mathbf{R} &= (\mathbf{X}'\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}\\ &= (\mathbf{X}'\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}\hat{\beta}\\ &= \mathbf{Z}_\lambda\hat{\beta}\end{aligned}$$

when $\lambda = 0$, ridge is least squares estimate

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

$\lambda$ increase: Var decrease bias increase
Hoerl and Kennard proved there exist a non-zero $\lambda$ s.t. MSE of $\hat{\beta}_R < Var(\hat{\beta})$ from OLS. provided $\hat{\beta}'\beta$ is bounded

$$\begin{aligned}SS_{Res} &= (\mathbf{y}-\mathbf{X}\hat{\beta}_\mathbf{R})'(\mathbf{y}-\mathbf{X}\hat{\beta}_\mathbf{R})\\ &= (\mathbf{y}-\mathbf{X}\hat{\beta})(\mathbf{y}-\mathbf{X}\hat{\beta}) + (\hat{\beta}_\mathbf{R}-\hat{\beta})'\mathbf{X}'\mathbf{X}(\hat{\beta}_\mathbf{R}-\hat{\beta})\end{aligned}$$

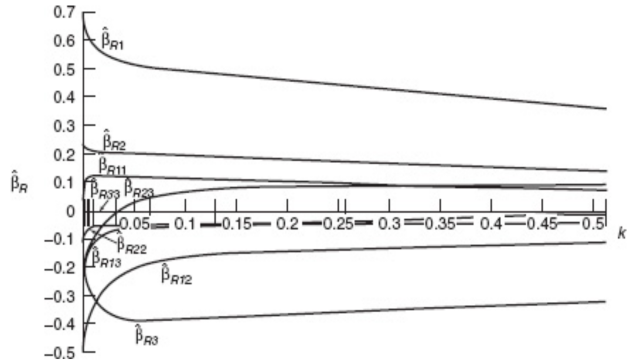
Term of LHS is $SS_{Res}$ for OLS $\hat{\beta}$
Ridge estimates will not necessarily provide the best "fit" to the data. When $\lambda$ increase, $SS_{Res}$ of $\hat{\beta}_R$ increase and $R^2$decrease

```
library(MASS)
lm.ridge(y~x, data=data, lambda=)
library(lmridge)
lmridge(y~x, data=data, K=)
summary(model) # only lmridge
```

### Ridge Trace

Objective: finding smallest possible $\lambda$ s.t. line is stable

**Figure 9.5** Ridge trace for acetylene data using nine regressors.




```
library(MASS)
# only mass
plot(lm.ridge(y~x,data=data,
        lambda=seq(0, 0.5, 0.01)))
select(model)
```

# Variable Selection

Basic Strategy for Variable Selection

1. Fit full model
2. Perform analysis (full residual analysis and investigate collinearity)
3. Determine transformation on response and some regressors
4. Use $t$-test on individual regressors
5. Perform analysis on edited model (esp residual analysis)

## Consequence of Model Misspecfication

1. $\hat{\beta}_q$ is biased for subset model
2. $Var(\hat{\beta}_q^*) \geq Var(\hat{\beta}_q)$
3. $MSE(\hat{\beta}_q^*) \geq MSE(\hat{\beta}_q)$
4. $\hat{\sigma}$ is biased for subset model
5. $MSE(\hat{y}^*) = Var(\hat{y}^*) \geq MSE(\hat{y})$

## Criteria for Evaluating Subset Regression models

- $R^2$

$$R_q^2 = \frac{SS_R(q)}{SS_T} = 1 - \frac{SS_{Res(q)}}{SS_T}$$

- Adjusted $R^2$

$$R_{adj,p}^2 = 1 - \frac{n-1}{n-p}(1-R_p^2)$$

- $MS_{Res}$

$$MS_{Res} = \frac{SS_{Res}(p)}{n-p}$$

- Akaike Information Criterion

$$AIC = nlog(\frac{SS_{Res}}{n}) + 2p$$

- Bayesian Information Criterion (OLS) by Schwartz and Sawa

$$BIC = nlog(\frac{SS_{Res}}{n}) + plog(n)$$

AIC, BIC must be comparing the same response of the same data size

## Computational Techniques for Variable Selection

- Evaluating All Possible models
- Stepwise Regression Methods
  - Forward Selection
  - Backward Elimination

### Forward Selection

1. Derived the fitted values and residuals from $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x - 1$ (model1)
2. Fit the regression model: $\hat{x}_j = \hat{\alpha}_{0j} + \hat{\alpha}_{1j} x_1, j \in [2, k]$
3. Derive the simple correlation between the residuals of Model 1 and the residuals from $k-1$ models above
4. The $x_j$ that give the largest correlation will be the next regressor to enter the model

Add to model if

$$F = \frac{SS_R(x_2|x_1)}{MS_{Res}(x_1, x_2)} > F_{in}$$

```
library(leaps)
model <- lm(y~1) # only intercept
step(model, direction="forward",
            scope=y~x1+x2+x3)
```

### Backward Elimination

1. Start model with $k$ regressor, compute partial $F$ statistic for each regressors as if it was the last varaible to enter the model
2. regressor with the smallest F stat is examined first and removed if $F < F_{out}$
3. Fit new model with rest of $k-1$ regressors and calculate partial F stat. Regressor with smallest partial F stat is remove if $F < F_{out}$

```
library(leaps)
model <- lm(y~x1+x2+x3)
step(model, direction="backward")
```

### Stepwise Regression Methods

1. At each step, all the regressors entered model previously are reassesssed via their partial F stat
2. A regressor added at earlier step may now be redundant because of the relationship between it and regressors now in the model
3. If the partial F stat for a variables is less than $F_{out}$ then variable is dropped from the model
4. Stepwise regression method requires two cutoff values. One for entering variables and one for removing them. It's often $F_{in} > F_{out}$

```
library(leaps)
model <- lm(y~x1+x2+x3)
step(model, direction="both")
```

## Strategy for Regression Model Building

Always choose the model passing model adequcy instead of just high $r^2$