Regression Analysis

Liming Feng

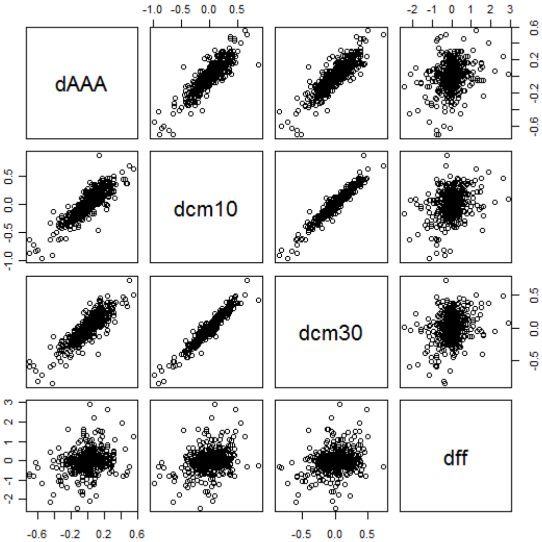
Dept. of Industrial & Enterprise Systems Engineering
University of Illinois at Urbana-Champaign

AAA corporate bond yield

- How is the change in AAA corporate bond yield related to changes in 10-year treasury rate, 30-year treasure rate, overnight federal funds rate, etc.?
- Weekly changes in interest rates from 2/16/1977 to 12/29/1993: dAAA (weekly change in AAA corporate bond yield), dcm10 (weekly change in 10-year treasury rate), dcm30 (weekly change in 30-year treasury rate), dff (weekly change in overnight federal funds rate)

```
> data=read.csv("InterestRates.csv",header=T)
> dAAA=diff(data$aaa)
> dcm10=diff(data$cm10)
> dcm30=diff(data$cm30)
> dff=diff(data$ff)
> pairs(cbind(dAAA,dcm10,dcm30,dff))
```

Scatterplot matrix



- From the scatterplot matrix, it seems that there is some linear relationship between change in AAA corporate bond yield and changes in 10-year and 30-year treasure rates
- Observing changes in 10-year and 30-year treasure rates, how much can we say about the change in AAA corporate bond yield
- Regression analysis studies how a **dependent variable** (dAAA) is related to some **explanatory variables/regressors** (dcm10, dcm30)

1. Linear regression model: assumptions

Ruppert 2011, §12.1; Hayashi 2000, §1.1

Linearity assumption

- Suppose there are K regressors. For $1 \leq i \leq n$, denote

$$Y_i: \text{ the } i\text{th observation of the dependent variable}$$

$$X_{ik}: \text{ the } i\text{th observation of the } k\text{th regressor, } 1 \leq k \leq K$$

- We assume the following **linear regression model**

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_K X_{iK} + \epsilon_i, \ 1 \leq i \leq n$$

  A constant term can be included by letting $X_{i1} \equiv 1$

- **Regression coefficient** β_k: the amount of change in the dependent variable when the kth regressor changes by 1 unit
- ϵ_i: the ith error term; the part of the dependent variable left unexplained by the regressors

Nonlinear transformations

- How restrictive is the linearity assumption?
- Suppose the relationship between the dependent variable Y and explanatory variable W is

$$Y_i = \beta_1 + \beta_2 W_i^2 + \epsilon_i$$

- Define $X_i := W_i^2$, we return to the linear regression model
- Power, logarithm, exponential functions are commonly used transformations

Matrix form

- Matrix form useful for clarity
- In matrix form, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

$$\mathbf{X}_i = \begin{bmatrix} X_{i1} \\ \vdots \\ X_{iK} \end{bmatrix}, \; \mathbf{X} = \begin{bmatrix} \mathbf{X}_1^\top \\ \vdots \\ \mathbf{X}_n^\top \end{bmatrix} = \begin{bmatrix} X_{11} & \cdots & X_{1K} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{nK} \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \; \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}, \; \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

where $\mathbf{X}_i$ is the ith observation of the regressors

Zero mean assumption

- The error term has **zero conditional mean**

$$\mathbb{E}[\epsilon_i|\mathbf{X}] := \mathbb{E}[\epsilon_i|\mathbf{X}_1, \cdots, \mathbf{X}_n] = 0, \ 1 \leq i \leq n$$

- By iterated conditioning, the error term has **zero unconditional mean**

$$\mathbb{E}[\epsilon_i] = \mathbb{E}[\mathbb{E}[\epsilon_i|\mathbf{X}]] = 0$$

- The regressors are **orthogonal** to the error term: $\forall 1 \leq k \leq K$, $1 \leq i, j \leq n$,

$$\mathbb{E}[\epsilon_i X_{jk}] = \mathbb{E}[\mathbb{E}[\epsilon_i X_{jk}|\mathbf{X}]] = \mathbb{E}[X_{jk}\mathbb{E}[\epsilon_i|\mathbf{X}]] = 0$$

- The regressors are **uncorrelated** with the error term:

$$\text{cov}(\epsilon_i, X_{jk}) = \mathbb{E}[\epsilon_i X_{jk}] - \mathbb{E}[\epsilon_i]\mathbb{E}[X_{jk}] = 0$$

No multicollinearity assumption

- We assume no **multicollinearity**: the rank of the $n \times K$ matrix $\mathbf{X}$ is K ($n \geq K$ then must hold)
- $\mathbf{X}$ has K columns, with kth column containing the n observations of the kth regressor
- No **multicollinearity** requires that the K columns be linearly independent: otherwise, some kth column can be expressed as a linear combination of the other columns; kth regressor is **redundant**
- For example, in the following model, the 2nd and 3rd regressors are identical; one is redundant

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i2} + \epsilon_i$$

Homoskedasticity and zero correlation assumption

- We assume conditional **homoskedasticity**

$$\mathrm{var}(\epsilon_i|\mathbf{X}) = \mathbb{E}[\epsilon_i^2|\mathbf{X}] = \sigma^2 > 0,\ 1 \le i \le n$$

and **zero conditional correlation** in the error term:

$$\mathrm{cov}(\epsilon_i, \epsilon_j|\mathbf{X}) = \mathbb{E}[\epsilon_i\epsilon_j|\mathbf{X}] = 0,\ 1 \le i \ne j \le n$$

- The above is equivalent to

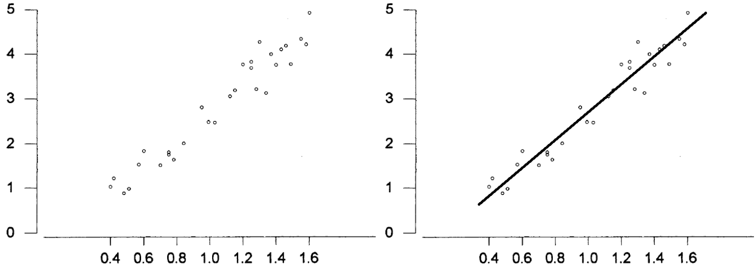
$$\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top|\mathbf{X}] = \sigma^2 I_n$$

where I_n is a $n \times n$ identity matrix

2. Simple regression model: estimation

Ruppert 2011, §12.2; Hayashi 2000, §1.2

- Suppose Y is the dependent variable, X is an explanatory variable. n pairs are observed: (X_i, Y_i) for $1 \leq i \leq n$. A scatter plot shows



- Data suggest that the dependent variable is a linear function of the regressor, plus a random deviation

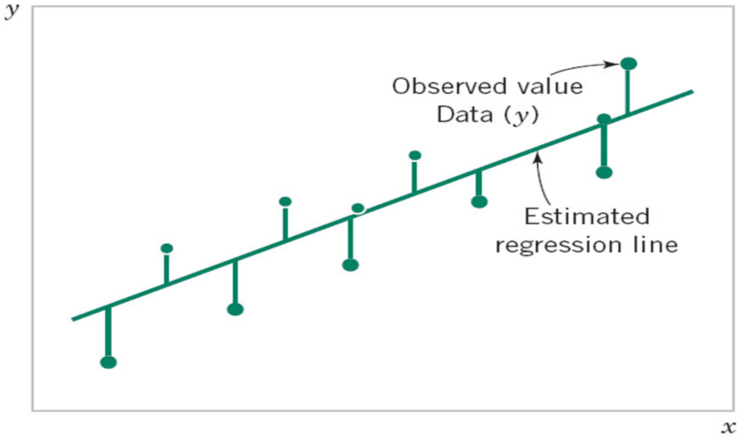
- **Simple regression** model:

$$Y_i = \beta_1 + \beta_2 X_i + \epsilon_i, \; 1 \leq i \leq n$$

- Assumptions for the simple regression model:

  (1) $\mathbb{E}[\epsilon_i | X_1, \cdots, X_n] = 0$;
  (2) X_i's are not constant;
  (3) $\mathbb{E}[\epsilon_i^2 | X_1, \cdots, X_n] = \sigma^2 > 0$;
  (4) $\mathbb{E}[\epsilon_i \epsilon_j | X_1, \cdots, X_n] = 0, 1 \leq i \neq j \leq n$

- Determine the regression line $y = \beta_1 + \beta_2 x$ by minimizing the distance between the observations and the fitted line



Ordinary least squares estimation

- Distance between Y_i and $\beta_1 + \beta_2 X_i$

$$Y_i - \beta_1 - \beta_2 X_i$$

- Ordinary least squares estimation: minimize

$$L(\beta_1, \beta_2) = \sum_{i=1}^{n} (Y_i - \beta_1 - \beta_2 X_i)^2$$

- From calculus,

$$\frac{\partial L}{\partial \beta_2} = 0 \Rightarrow \beta_2 \sum_{i=1}^{n} X_i^2 + \beta_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} X_i Y_i$$

$$\frac{\partial L}{\partial \beta_1} = 0 \Rightarrow \beta_1 + \beta_2 \bar{X}_n = \bar{Y}_n$$

where $\bar{X}_n$, $\bar{Y}_n$ are sample means

- **Ordinary least squares estimators**

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}, \quad \hat{\beta}_1 = \bar{Y}_n - \hat{\beta}_2 \bar{X}_n$$

- Denote

$$S_{XX} = \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

$$S_{XY} = \sum_{i=1}^{n} (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)$$

Then

$$\hat{\beta}_2 = \frac{S_{XY}}{S_{XX}}, \quad \hat{\beta}_1 = \bar{Y}_n - \hat{\beta}_2 \bar{X}_n$$

Fitted values and residuals

- ith **fitted value**

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

  Part of the dependent variable explained by the regressors

- ith OLS **residual**

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$$

  Part of the dependent variable not explained by the regressors

- Residuals can be used to estimate σ^2

Estimating σ^2

- Note that sample mean of $\hat{\epsilon}_i$'s is zero

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i = \bar{Y}_n - \hat{\beta}_1 - \hat{\beta}_2\bar{X}_n = 0$$

- OLS estimator for σ^2

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\epsilon}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$\hat{\sigma}^2$ is an **unbiased** estimator of σ^2; $n-2$ is the **degrees of freedom**

Degrees of freedom

- Estimating the unknown variance σ^2 of a distribution with known mean μ and random sample $X_1, \cdots, X_n$

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

is unbiased; each term in the summation represents independent information; degrees of freedom is n

- To see unbiasedness,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(X_i - \mathbb{E}[X_i])^2] = \sigma^2$$

- Estimating the variance of a distribution with unknown mean and random sample $X_1, \cdots, X_n$

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

is unbiased; only $n-1$ terms in the summation represent independent information; degrees of freedom is $n-1$

- The last term $X_n - \bar{X}_n$ is a linear combination of the previous $n-1$ terms since

$$X_1 - \bar{X}_n + \cdots + X_{n-1} - \bar{X}_n + X_n - \bar{X}_n = 0$$

- For the OLS estimator for σ^2,

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

only $n-2$ terms in the following summation represent independent information; degrees of freedom is $n-2$

$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$

$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)X_i = 0$$

(Recall how OLS estimators were obtained)

3. Simple regression model: analysis

Ruppert 2011, §12.2; Hayashi 2000, §1.3

Unbiasedness

- OLS estimator $\hat{\beta}_2$ is unbiased with $\mathbb{E}[\hat{\beta}_2] = \beta_2$ since

$$
\begin{aligned}
\mathbb{E}[\hat{\beta}_2|\mathbf{X}] &= \mathbb{E}\left[\left.\frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{S_{XX}}\right|\mathbf{X}\right] \\
&= \frac{1}{S_{XX}}\sum_{i=1}^{n}(X_i - \bar{X}_n)\mathbb{E}[(Y_i - \bar{Y}_n)|\mathbf{X}] \\
&= \frac{1}{S_{XX}}\sum_{i=1}^{n}(X_i - \bar{X}_n)(\beta_1 + \beta_2 X_i - (\beta_1 + \beta_2 \bar{X}_n)) \\
&= \beta_2
\end{aligned}
$$

- $\hat{\beta}_1$ is unbiased with $\mathbb{E}[\hat{\beta}_1] = \beta_1$ since

$$
\mathbb{E}[\hat{\beta}_1|\mathbf{X}] = \mathbb{E}[\bar{Y}_n - \hat{\beta}_2\bar{X}_n|\mathbf{X}] = \beta_1 + \beta_2\bar{X}_n - \beta_2\bar{X}_n = \beta_1
$$

Standard error for $\hat{\beta}_2$

- Note that

$$\begin{aligned}\hat{\beta}_2 &= \sum_{i=1}^{n} \frac{X_i - \bar{X}_n}{S_{XX}} Y_i \\ &= \sum_{i=1}^{n} \frac{X_i - \bar{X}_n}{S_{XX}} (\beta_1 + \beta_2 X_i + \epsilon_i)\end{aligned}$$

- Recall the zero conditional correlation assumption:

$$\text{var}(\hat{\beta}_2 | \mathbf{X}) = \frac{1}{S_{XX}^2} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \text{var}(\epsilon_i | \mathbf{X}) = \frac{\sigma^2}{S_{XX}}$$

- Estimated standard error: $se(\hat{\beta}_2 | \mathbf{X}) = \frac{\hat{\sigma}}{\sqrt{S_{XX}}}$

- Recall that $\hat{\beta}_1 = \bar{Y}_n - \hat{\beta}_2 \bar{X}_n$

$$\begin{aligned}
\text{cov}(\bar{Y}_n, \hat{\beta}_2 | \mathbf{X}) &= \sum_{i=1}^{n} \frac{X_i - \bar{X}_n}{S_{XX}} \text{cov}(\bar{Y}_n, Y_i | \mathbf{X}) \\
&= \frac{\sigma^2}{n S_{XX}} \sum_{i=1}^{n} (X_i - \bar{X}_n) \; = \; 0
\end{aligned}$$

- Variance and estimated standard error of $\hat{\beta}_1$

$$\text{var}(\hat{\beta}_1 | \mathbf{X}) = \text{var}(\bar{Y}_n | \mathbf{X}) + \bar{X}_n^2 \text{var}(\hat{\beta}_2 | \mathbf{X}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}_n^2}{S_{XX}} \right)$$

$$se(\hat{\beta}_1 | \mathbf{X}) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{X}_n^2}{S_{XX}}}$$

```
> fit=lm(dAAA~dcm30)
> summary(fit)

Call:
lm(formula = dAAA ~ dcm30)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34351 -0.03247 -0.00156  0.02961  0.40110

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0001208  0.0022570  -0.054    0.957
dcm30        0.6853163  0.0137830  49.722   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06695 on 878 degrees of freedom
Multiple R-squared: 0.7379,     Adjusted R-squared: 0.7376
F-statistic:  2472 on 1 and 878 DF,  p-value: < 2.2e-16
```

Reading R outputs

- Let Y be change in AAA corporate bond yield, X be change in 30-year treasury rate. Fit a simple regression model

$$Y_i = \beta_1 + \beta_2 X_i + \epsilon_i$$

- R outputs: $\hat{\beta}_1 = -0.0001$; $\hat{\beta}_2 = 0.6853$; $\hat{\sigma} = 0.06695$; number of observations $n = 880$; number of regressors $K = 2$ (including a constant regressor); degrees of freedom for estimating σ^2: $n - K = 878$
- How well does the model fit the data

Goodness of fit

- The dependent variable Y varies: does it vary because the (non-constant) regressor X vary?
- If there is no linear relationship between Y and X, $\hat{\beta}_2 \approx 0$: σ^2 will be close to the variance of Y
- If there is a perfect linear relationship between Y and X, variation in Y completely explained by variation in X: $\sigma^2 = 0$
- How much variation in the dependent variable can be explained by variation in the regressors?
- A model is better if most of the variation in the dependent variable can be explained by regressors

Analysis of variance (ANOVA)

- Variation in the dependent variable: "total sum of squares"

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2$$

- Variation explained by the regression model: regression sum of squares

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_n)^2 \quad \left(\text{note that } \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i = \bar{Y}_n\right)$$

- Variation due to the noise: error sum of squares (or residual sum of squares)

$$SSE = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \quad \left(\text{note that } \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i = 0\right)$$

- Decomposition of variation: $SST = SSR + SSE$

$$
\begin{aligned}
SST &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y}_n)^2 \\
&= SSE + SSR + 2\sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}_n)
\end{aligned}
$$

- Recall $\hat{\beta}_1 = \bar{Y}_n - \hat{\beta}_2 \bar{X}_n \Rightarrow \hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i = \bar{Y}_n + \hat{\beta}_2 (X_i - \bar{X}_n)$

$$
\begin{aligned}
\sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}_n) &= \hat{\beta}_2 \sum_{i=1}^{n} (Y_i - \bar{Y}_n - \hat{\beta}_2 (X_i - \bar{X}_n))(X_i - \bar{X}_n) \\
&= \hat{\beta}_2 (S_{XY} - \hat{\beta}_2 S_{XX}) \; = \; 0
\end{aligned}
$$

R squared

- R^2 (coefficient of determination) measures the proportion of variation in Y that is explained by the regression

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- R^2 close to 1 indicates a good fitting
- R = sample correlation in the simple regression model

$$\rho^2 = \frac{(\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n))^2}{(\sum_{i=1}^n (X_i - \bar{X}_n)^2)(\sum_{i=1}^n (Y_i - \bar{Y}_n)^2)} = \frac{S_{XY}^2}{S_{XX} SST}$$

$$= \frac{\hat{\beta}_2^2 S_{XX}}{SST} = \frac{SSR}{SST} = R^2, \left(\text{Recall } \hat{Y}_i = \bar{Y}_n + \hat{\beta}_2 (X_i - \bar{X}_n)\right)$$

- When dAAA is regressed on dff

  sample correlation: 25.0%
  R^2 : 6.254%
  sample standard deviation of Y : 0.1307
  $\hat{\sigma}$: 0.1266

- When dAAA is regressed on dcm30

  sample correlation: 85.9%
  R^2 : 73.79%
  sample standard deviation of Y : 0.1307
  $\hat{\sigma}$: 0.06695

Degrees of freedom

- Degrees of freedom for each sum of squares: the number of independent pieces of information
- For SST, the n terms are $Y_1 - \bar{Y}_n, \cdots, Y_n - \bar{Y}_n$; there are **$n-1$ degrees of freedom** since

$$(Y_1 - \bar{Y}_n) + \cdots + (Y_n - \bar{Y}_n) = 0$$

- For SSR, the n terms are $\{\hat{Y}_i - \bar{Y}_n = \hat{\beta}_2(X_i - \bar{X}_n), 1 \leq i \leq n\}$; there is **1 degree of freedom**
- For SSE, recall that SSE$= (n-2)\hat{\sigma}^2$, there are **$n-2$ degrees of freedom**

4. Simple regression model: inference

Ruppert 2011, §12.2; Hayashi 2000, §1.4

Statistical inference

- Hypothesis testing: e.g., is $\beta_2 = 0$?
- Confidence interval: construct a 95% CI for β_2
- How can you predict Y from the value of X by constructing a prediction interval
- Need to specify the distribution for ϵ_i's
- In this part: we assume that $\boldsymbol{\epsilon}|\mathbf{X} \sim N(0, \sigma^2 I_n)$ (this implies that $\boldsymbol{\epsilon}$ is independent of $\mathbf{X}$ and $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_n)$)

Inference on β_2

- Since $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_n)$

$$\mathbf{Y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n)$$

  Each $Y_i|\mathbf{X} \sim N(\beta_1 + \beta_2 X_i, \sigma^2)$, and $(Y_1, \cdots, Y_n)^\top|\mathbf{X}$ is multivariate normal
- Note that

$$\hat{\beta}_2 = \sum_{i=1}^{n} \frac{X_i - \bar{X}_n}{S_{XX}} Y_i, \ \hat{\beta}_1 = \bar{Y}_n - \hat{\beta}_2 \bar{X}_n$$

- $\hat{\beta}_1, \hat{\beta}_2$ as linear combinations of Y_i's are also normal conditional on $\mathbf{X}$

Distribution of $\hat{\beta}_2$

- Recall that

$$\text{var}(\hat{\beta}_2|\mathbf{X}) = \frac{\sigma^2}{S_{XX}}$$

and $\hat{\beta}_2$ is unbiased estimator of β_2. Therefore,

$$\hat{\beta}_2|\mathbf{X} \sim N(\beta_2, \frac{\sigma^2}{S_{XX}})$$

- By standardization

$$\frac{\hat{\beta}_2 - \beta_2}{\sigma/\sqrt{S_{XX}}} \sim N(0, 1)$$

- Replace σ by its OLS estimator $\hat{\sigma}$, $\frac{\hat{\beta}_2 - \beta_2}{\hat{\sigma}/\sqrt{S_{XX}}} \sim t_{n-2}$

Testing $\beta_2 = 0$

- Is there strong enough evidence that $\beta_2 \neq 0$?
- Reject the null $H_0 : \beta_2 = 0$ in support of $H_1 : \beta_2 \neq 0$ at significance level α if $t_0 = \hat{\beta}_2 / (\hat{\sigma}/\sqrt{S_{XX}})$ satisfies

  $$|t_0| > t_{\alpha/2,n-2}$$

  where $-t_{\alpha/2,n-2}$ is the $\alpha/2$ quantile of t_{n-2}
- Equivalently, reject H_0 if p-value is less than α

  $$\text{p-value} = \mathbb{P}(|T| > |t_0|),\ T \sim t_{n-2}$$

- Equivalently, reject H_0 if 0 is not in the $100(1-\alpha)\%$ confidence interval for β_2: $\hat{\beta}_2 \pm t_{\alpha/2,n-2}\hat{\sigma}/\sqrt{S_{XX}}$
- Testing whether $\beta_2 = \beta_2^*$ for arbitrary β_2^* can be done similarly

Example: dAAA and dcm30

- When dAAA is regressed on dcm30, there is strong evidence that $\beta_2 \neq 0$, with p-value $2e-16$

```
> fit=lm(dAAA~dcm30)
> summary(fit)

Call:
lm(formula = dAAA ~ dcm30)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.34351 -0.03247 -0.00156  0.02961  0.40110 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.0001208  0.0022570  -0.054    0.957    
dcm30        0.6853163  0.0137830  49.722   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06695 on 878 degrees of freedom
Multiple R-squared: 0.7379,     Adjusted R-squared: 0.7376 
F-statistic:  2472 on 1 and 878 DF,  p-value: < 2.2e-16

> confint(fit, 'dcm30', level=0.95)
          2.5 %    97.5 %
dcm30 0.6582648 0.7123677
```

Inference on β_1

- $\hat{\beta}_1|\mathbf{X}$ is normal with mean β_1 and variance

$$\operatorname{var}(\hat{\beta}_1|\mathbf{X}) = \sigma^2\Big(\frac{1}{n} + \frac{\bar{X}_n^2}{S_{XX}}\Big)$$

- By standardization,

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma\sqrt{\frac{1}{n} + \frac{\bar{X}_n^2}{S_{XX}}}} \sim N(0,1), \quad \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{X}_n^2}{S_{XX}}}} \sim t_{n-2}$$

- Testing whether $\beta_1 = 0$ (or any β_1^*) can be done similarly

Prediction

- For new observation $X = X^*$, one gets a prediction

$$\hat{Y}^* = \hat{\beta}_1 + \hat{\beta}_2 X^*$$

- True value is $Y^* = \beta_1 + \beta_2 X^* + \epsilon^*$. Let $\mathbf{X}^* = \{\mathbf{X}, X^*\}$. Then $(Y^* - \hat{Y}^*)|\mathbf{X}^*$ is normal with mean 0 and variance

$$\begin{aligned}
\text{var}(Y^* - \hat{Y}^*|\mathbf{X}^*) &= \text{var}(\epsilon^* - \hat{\beta}_1 - \hat{\beta}_2 X^*|\mathbf{X}^*) \\
&= \sigma^2 + \text{var}(\hat{\beta}_1|\mathbf{X}) + (X^*)^2 \text{var}(\hat{\beta}_2|\mathbf{X}) \\
&\quad +2X^* \text{cov}(\hat{\beta}_1, \hat{\beta}_2|\mathbf{X}) \\
&= \sigma^2 + \text{var}(\hat{\beta}_1|\mathbf{X}) + ((X^*)^2 - 2X^*\bar{X}_n)\text{var}(\hat{\beta}_2|\mathbf{X}) \\
&= \sigma^2 \left(1 + \frac{1}{n} + \frac{(\bar{X}_n - X^*)^2}{S_{XX}}\right)
\end{aligned}$$

For $\text{cov}(\hat{\beta}_1, \hat{\beta}_2|\mathbf{X})$, use $\hat{\beta}_1 = \bar{Y}_n - \hat{\beta}_2 \bar{X}_n$ and $\text{cov}(\bar{Y}_n, \hat{\beta}_2|\mathbf{X}) = 0$

Prediction interval

- When replacing σ by $\hat{\sigma}$, we have

$$\frac{Y^* - \hat{Y}^*}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(\bar{X}_n - X^*)^2}{S_{XX}}}} \sim t_{n-2}$$

- Prediction interval around $\hat{Y}^*$ that contains Y^* with probability $1 - \alpha$

$$Y^* \in \hat{Y}^* \pm t_{\alpha/2, n-2} \cdot \hat{\sigma}_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(\bar{X}_n - X^*)^2}{S_{XX}}}$$

Predicting dAAA using dcm30

- Regress dAAA on dcm30 using the first 879 observations, use the 880th observation for dcm30 to predict dAAA

```
> Y=dAAA[1:879]
> X=dcm30[1:879]
> fit=lm(Y~X)
> newdata=data.frame(X=dcm30[880])
> predict(fit,newdata,interval="predict")
          fit        lwr        upr
1 -0.03441676 -0.1659662 0.09713264
> dAAA[880]
[1] -0.01
```

- Recall that $SST = SSR + SSE$, $SSE = (n-2)\hat{\sigma}^2$, $SSR = \hat{\beta}_2^2 S_{XX}$
- Define

$$F = \frac{SSR}{SSE/(n-2)} = \frac{\hat{\beta}_2^2 S_{XX}}{\hat{\sigma}^2} = \Big(\frac{\hat{\beta}_2}{\hat{\sigma}/\sqrt{S_{XX}}}\Big)^2$$

Recall that

$$\frac{\hat{\beta}_2 - \beta_2}{\hat{\sigma}/\sqrt{S_{XX}}} \sim t_{n-2}$$

- Under $H_0 : \beta_2 = 0$, F has the same distribution as the square of t_{n-2}: $F \sim F_{1,n-2}$; Reject H_0 when p-value $\mathbb{P}(F_{1,n-2} > F)$ is less than α

- ANOVA table

Source of variation	Sum of squares	Degrees of freedom	Mean square	F
Regression	SSR	1	$MSR = SSR$	$F = \frac{MSR}{MSE}$
Error	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$	
Total	SST	$n-1$		

- ANOVA when dAAA is regressed on dcm30

```
> fit=lm(dAAA~dcm30)
> anova(fit)
Analysis of Variance Table

Response: dAAA
           Df  Sum Sq Mean Sq F value    Pr(>F)
dcm30       1 11.0818 11.0818  2472.3 < 2.2e-16 ***
Residuals 878  3.9356  0.0045
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Verify the degrees of freedom
- In simple regression models, the F-test is the same as the t-test for $H_0 : \beta_2 = 0; H_1 : \beta_2 \neq 0$
- $SST = 15.02$, of which SSR is of a significant portion, indicating a reasonab fit

5. Multiple regression model

Hayashi 2000, §1.2, §1.3, §1.4;
Ruppert 2011, §12.3, §12.4, §12.5

Multiple linear regression

- Can dcm10 and dcm30 together better explain dAAA?
- Multiple linear regression

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_K X_{iK} + \epsilon_i, 1 \leq i \leq n$$

  Assume that the first regressor is constant with $X_{i1} \equiv 1$
- In matrix notation, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
- Recall the assumptions: (1) $\mathbb{E}[\epsilon_i|\mathbf{X}] = 0$; (2) no multicollinearity; (3) $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top|\mathbf{X}] = \sigma^2 I_n$

Ordinary least squares estimator

- Minimize

$$
\begin{aligned}
L(\boldsymbol{\beta}) &= \sum_{i=1}^{n}(Y_i - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots - \beta_K X_{iK})^2 \\
&= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \mathbf{Y}^\top\mathbf{Y} - 2\mathbf{Y}^\top\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}
\end{aligned}
$$

  Normal equations

$$
\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^\top\mathbf{Y} + 2\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = 0 \Rightarrow \mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{Y}
$$

- With no multicollinearity, $\mathbf{X}^\top\mathbf{X}$ is invertible
- Ordinary least squares estimator: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$

Fitted values

- Fitted values using matrix notation

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} = \mathbf{P}\mathbf{Y}$$

Recall the derivation for OLS estimator $\hat{\beta}$

$$\sum_{i=1}^{n} X_{i1}(Y_i - \hat{Y}_i) = \sum_{i=1}^{n}(Y_i - \hat{Y}_i) = 0 \Rightarrow \bar{\hat{Y}}_n := \frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i = \bar{Y}_n$$

- $\mathbf{P} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is called the **projection matrix**, $\mathbf{M} = I_n - \mathbf{P} = I_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is called the **annihilator matrix**. They are symmetric and idempotent

$$\mathbf{P}^\top = \mathbf{P}, \quad \mathbf{P}^2 = \mathbf{P}, \quad \mathbf{P}\mathbf{X} = \mathbf{X}$$

$$\mathbf{M}^\top = \mathbf{M}, \quad \mathbf{M}^2 = \mathbf{M}, \quad \mathbf{M}\mathbf{X} = 0$$

OLS estimator for σ^2

- Residuals using matrix notation

$$\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = (I_n - \mathbf{P})\mathbf{Y} = \mathbf{M}\mathbf{Y}$$

- Error sum of squares

$$SSE = \hat{\boldsymbol{\epsilon}}^\top \hat{\boldsymbol{\epsilon}} = \mathbf{Y}^\top \mathbf{M}\mathbf{Y} = \boldsymbol{\epsilon}^\top \mathbf{M}\boldsymbol{\epsilon}$$

  with $n - K$ degrees of freedom. It can be shown that $\mathbb{E}[\boldsymbol{\epsilon}^\top \mathbf{M}\boldsymbol{\epsilon}|\mathbf{X}] = \sigma^2(n - K)$

- The following OLS estimator for σ^2 is unbiased

$$\hat{\sigma}^2 = \frac{SSE}{n - K}$$

$\hat{\beta}$ is unbiased

- Proof of unbiasedness using matrix notation

$$
\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= \mathbb{E}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}|\mathbf{X}] \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbb{E}[\mathbf{X}\beta + \epsilon|\mathbf{X}] \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\beta \\
&= \boldsymbol{\beta}
\end{aligned}
$$

Therefore, $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$

Variance of $\hat{\beta}$

- Computing the variance of $\hat{\beta}$ using the matrix notation

$$
\begin{aligned}
\mathrm{var}(\hat{\beta}|\mathbf{X}) &= \mathrm{var}(\mathbf{AY}|\mathbf{X}), \quad \mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \\
&= \mathbf{A}\mathrm{var}(\mathbf{Y}|\mathbf{X})\mathbf{A}^\top \\
&= \mathbf{A}\sigma^2 I_n \mathbf{A}^\top \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}
\end{aligned}
$$

- $SST = SSR + SSE$

$$
\begin{aligned}
\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2 &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y}_n)^2 \\
&= SSE + SSR + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}_n) \\
&= SSE + SSR + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)\hat{Y}_i \\
&= SSE + SSR + 2(\mathbf{Y} - \hat{\mathbf{Y}})^\top \hat{\mathbf{Y}} \\
&= SSE + SSR + 2\mathbf{Y}^\top \mathbf{P}\mathbf{Y} - 2\mathbf{Y}^\top \mathbf{P}^\top \mathbf{P}\mathbf{Y} \\
&= SSE + SSR
\end{aligned}
$$

- SST has $n-1$ degrees of freedom

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2$$

- SSE has $n-K$ degrees of freedom

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Recall the derivation for OLS estimator $\hat{\beta}$

$$\sum_{i=1}^{n} X_{ik}(Y_i - \hat{Y}_i) = 0, \quad k = 1, \cdots, K$$

- SSR has $K-1$ degrees of freedom

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_n)^2$$

Recall that

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_K X_{iK}$$

$$\bar{Y}_n = \bar{\hat{Y}}_n = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}_{\cdot 2} + \cdots + \hat{\beta}_K \bar{X}_{\cdot K}, \quad \bar{X}_{\cdot k} = \frac{1}{n} \sum_{i=1}^{n} X_{ik}$$

$$\hat{Y}_i - \bar{Y}_n = \sum_{k=2}^{K} \hat{\beta}_k (X_{ik} - \bar{X}_{\cdot k})$$

Goodness of fit

- Goodness of fit measured by R^2

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- SSE will not increase when K increases (recall how least squares estimates are obtained); R^2 will not decrease when K increases
- R^2 favors models with more predictors

- Adjusted R^2 takes the number of parameters into account

$$R^2_{adj} = 1 - \frac{MSE}{MST} = 1 - \frac{(n-K)^{-1} SSE}{(n-1)^{-1} SST}$$

- Decrease in SSE by increasing K is accompanied by increase in $(n-K)^{-1}$
- Introducing an extra predictor is beneficial only if there is sufficient reduction in SSE

Inference

- For statistical inference on $\boldsymbol{\beta}$: assume that $\boldsymbol{\epsilon}|\mathbf{X} \sim N(0, \sigma^2 I_n)$
- Note that $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

$$\mathbf{Y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n)$$

- Recall that $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}$ with mean $\boldsymbol{\beta}$ and covariance matrix $\sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$ (conditional on $\mathbf{X}$)

$$\hat{\boldsymbol{\beta}}|\mathbf{X} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1})$$

Inference on β_k

- For inference on β_k:

$$\frac{\hat{\beta}_k - \beta_k}{\sigma\sqrt{(\mathbf{X}^\top\mathbf{X})^{-1}_{kk}}} \sim N(0,1)$$

where $(\mathbf{X}^\top\mathbf{X})^{-1}_{kk}$ is the kth row and kth column entry of the matrix $(\mathbf{X}^\top\mathbf{X})^{-1}$

- Replacing σ^2 by its OLS estimator $\hat{\sigma}^2 = SSE/(n-K)$

$$\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}\sqrt{(\mathbf{X}^\top\mathbf{X})^{-1}_{kk}}} = \frac{\frac{\hat{\beta}_k - \beta_k}{\sigma\sqrt{(\mathbf{X}^\top\mathbf{X})^{-1}_{kk}}}}{\sqrt{\frac{SSE}{\sigma^2(n-K)}}} \sim t_{n-K}$$

The numerator $\sim N(0,1)$, $SSE/\sigma^2|\mathbf{X} \sim \chi^2_{n-K}$, and they are independent (conditional on $\mathbf{X}$)

Testing β_k

- Testing $H_0 : \beta_k = 0; H_1 : \beta_k \neq 0$ can be done by
  - computing confidence interval for β_k and checking whether it contains 0
  - computing
    $$t_{0k} := \frac{\hat{\beta}_k - 0}{\hat{\sigma}\sqrt{(\mathbf{X}^\top \mathbf{X})^{-1}_{kk}}}$$
    and comparing it with $\pm t_{\alpha/2, n-K}$
  - computing p-value
    $$\mathbb{P}(|t_{n-K}| > |t_{0k}|)$$
    and comparing it with α

Testing $\mathbf{R}\beta = \mathbf{r}$

- More general tests on the coefficients
  - $K = 4$ and $H_0 : \beta_2 = \beta_3, \quad \beta_4 = 0$ (the 2nd and 3rd regressors have the same effect and the 4th regressor is not useful for explaining the dependent variable)

$$\mathbf{R} = \begin{pmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{r} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

  - $H_0 : \beta_2 = \cdots = \beta_K = 0$ (none of the regressors is useful for explaining the dependent variable)

$$\mathbf{R} = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}, \quad \mathbf{r} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}_{K-1}$$

- Distribution of $\mathbf{R}\hat{\beta}$

$$\mathbf{R}\hat{\beta}|\mathbf{X} \sim N(\mathbf{R}\beta, \sigma^2\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top)$$

- For any $d-$dimensional random vector $X \sim N(\mu, \Sigma)$,

$$(X-\mu)^\top\Sigma^{-1}(X-\mu) \sim \chi^2_d$$

- Consequently,

$$(\mathbf{R}\hat{\beta}-\mathbf{R}\beta)^\top(\sigma^2\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top)^{-1}(\mathbf{R}\hat{\beta}-\mathbf{R}\beta) \sim \chi^2_{dim(\mathbf{r})}$$

where $dim(\mathbf{r})$ is the dimension of $\mathbf{r}$

F test

- Consider $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$; $H_1 : \mathbf{R}\boldsymbol{\beta} \neq \mathbf{r}$. Under the null hypothesis H_0, the following has an F distribution

$$\frac{\frac{1}{dim(\mathbf{r})}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})^\top(\sigma^2\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top)^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})}{\frac{SSE}{\sigma^2(n-K)}} \sim F_{dim(\mathbf{r}),n-K}$$

  Numerator $\sim \chi^2_{dim(\mathbf{r})}/dim(\mathbf{r})$; denominator $\sim \chi^2_{n-K}/(n-K)$; they are independent (conditional on $\mathbf{X}$)
- Reject H_0 is the above value is larger than $F_{\alpha,dim(\mathbf{r}),n-K}$

Testing $H_0 : \beta_2 = \cdots = \beta_K = 0$

- In the simple linear regression model ($K = 2$), under $H_0 : \beta_2 = 0$, the test statistic on the previous slide becomes $MSR/MSE \sim F_{1,n-2}$; reject H_0 if this value is above $F_{\alpha,1,n-2}$
- In the multiple linear regression model, under $H_0 : \beta_2 = \cdots = \beta_K = 0$, the test statistic becomes

$$\frac{MSR}{MSE} = \frac{SSR/(K-1)}{SSE/(n-K)} \sim F_{K-1,n-K}$$

  Reject H_0 if the above value is larger than $F_{\alpha,K-1,n-K}$
- If $H_0 : \beta_2 = \cdots = \beta_K = 0$ were rejected, at least one of the regressors is useful in explaining the dependent variable

- ANOVA table

Source of variation	Sum of squares	Degrees of freedom	Mean square	F
Regression	SSR	$K-1$	$MSR = \frac{SSR}{K-1}$	$F = \frac{MSR}{MSE}$
Error	SSE	$n-K$	$MSE = \frac{SSE}{n-K}$	
Total	SST	$n-1$		

dAAA on dcm30

```
> summary(lm(dAAA~dcm30))

Call:
lm(formula = dAAA ~ dcm30)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34351 -0.03247 -0.00156  0.02961  0.40110

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0001208  0.0022570  -0.054    0.957
dcm30        0.6853163  0.0137830  49.722   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06695 on 878 degrees of freedom
Multiple R-squared: 0.7379,     Adjusted R-squared: 0.7376
F-statistic:  2472 on 1 and 878 DF,  p-value: < 2.2e-16
```

dAAA on dcm10, dcm30 and dff

```
> summary(lm(dAAA~dcm30+dcm10+dff))

Call:
lm(formula = dAAA ~ dcm30 + dcm10 + dff)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.33484 -0.03115 -0.00059  0.03062  0.39986 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -9.069e-05  2.179e-03  -0.042    0.967    
dcm30        3.002e-01  5.003e-02   6.000 2.88e-09 ***
dcm10        3.545e-01  4.512e-02   7.858 1.14e-14 ***
dff          4.122e-03  5.282e-03   0.780    0.435    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06463 on 876 degrees of freedom
Multiple R-squared: 0.7563,     Adjusted R-squared: 0.7555 
F-statistic: 906.2 on 3 and 876 DF,  p-value: < 2.2e-16
```

dAAA on dcm10 and dcm30

```
> summary(lm(dAAA~dcm30+dcm10))

Call:
lm(formula = dAAA ~ dcm30 + dcm10)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.33450 -0.03139 -0.00049  0.03032  0.40748 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -9.376e-05  2.178e-03  -0.043    0.966    
dcm30        2.968e-01  4.983e-02   5.956 3.73e-09 ***
dcm10        3.602e-01  4.452e-02   8.092 1.96e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.06462 on 877 degrees of freedom
Multiple R-squared: 0.7561,     Adjusted R-squared: 0.7556 
F-statistic:  1360 on 2 and 877 DF,  p-value: < 2.2e-16
```

Model selection

- Regress dAAA on dcm30, dcm10 and dff: p-value for the F-test on $H_0 : \beta_{dcm30} = \beta_{dcm10} = \beta_{dff} = 0$ is less than 2.2×10^{-16}; H_0 is rejected; some of the regressors are useful in explaining dAAA
- dff may not be useful in explaining dAAA when dcm30 and dcm10 are present: p-value for the t-test on $H_0 : \beta_{dff} = 0$ is 0.435; fail to reject $H_0 : \beta_{dff} = 0$
- Regress dAAA on dcm10 and dcm30: R^2_{adj} chooses this model with a value of 0.7556
- R^2 always favors models with more regressors and is not suitable for model selection

Interpreting outputs

- dAAA regressed on dcm30: $R^2 = 0.7379$
- dAAA regressed on dcm30 and dcm10: $R^2 = 0.7561$
- R^2 not increasing significantly doesn't mean dcm10 is less useful in explaining dAAA
- dAAA regressed on dcm10: $R^2 = 0.7463$
- dcm10 and dcm30 are highly correlated (correlation coefficient 96.4%); adding dcm10 to a model with regressor dcm30 is useful but improvement is limited

- dAAA regressed on dcm30, dcm10 and dff: p-value for $H_0 : \beta_{dff} = 0$ is 0.435; H_0 rejected
- dAAA regressed on dff: p-value for $H_0 : \beta_{dff} = 0$ is 5.2×10^{-14}; H_0 not rejected
- If dff is the only regressor, it provides useful info; when dcm30 and dcm10 are present, it is redundant
- A small p-value for testing $H_0 : \beta_k = 0$ doesn't exclude nonlinear relation between kth regressor and the dependent variable; similarly for a large p-value
- Graphical analysis is useful (e.g., scatter plots)

6. Regression diagnostics

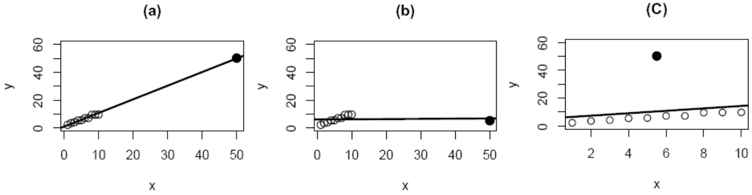
Ruppert 2011, Chapter 13

What can go wrong

- Data entered incorrectly
- Data used are not what one thinks
- Assumptions of linear regression violated
  - Linearity
  - $\mathbb{E}[\boldsymbol{\epsilon}|\mathbf{X}] = \mathbf{0}$
  - No multicollinearity and $\mathbf{X}^\top\mathbf{X}$ is non-singular
  - $E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top|\mathbf{X}] = \sigma^2 I_n$
  - Normality: $\boldsymbol{\epsilon}|\mathbf{X} \sim N(0, \sigma^2 I_n)$

Leverage points and residual outliers

- Problems with data may lead to unusual observations
- Unusual observations may significantly affect the outcome and should be treated carefully
- Consider $Y_i = 1 + X_i + \epsilon_i$, where $0 \leq X_i \leq 10$, $1 \leq i \leq 10$, $X_{11} = 50$



- (a): the 11th observation is a leverage point; not causing problem; (b): Y_{11} recorded mistakenly; a leverage point; significant bias in $\hat{\beta}_2$; (c): X_{11} recorded mistakenly; a residual outlier; bias in $\hat{\beta}_1$

Example

- Leverage points could be identified by computing their leverages or using graphics; residual outliers can be identified from residual plots
- They may distort the fitting and are worth further investigation
- Should be eliminated from the data if they were included by mistake
- **Example**: a simple linear regression model is used to study the relation between the coupon rate of a bond and its price. Data can be found on http://www.stat.tamu.edu/ sheather/book/docs/datasets/bonds.txt

```
> bonds <- read.table("bonds.txt",header=TRUE)
> slg <- lm(bonds$BidPrice~bonds$CouponRate)
> summary(slg)
```

Direct simple linear regression fitting

```
Call:
lm(formula = bonds$BidPrice ~ bonds$CouponRate)

Residuals:
   Min      1Q Median      3Q     Max
-8.249 -2.470 -0.838  2.550 10.515

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       74.7866     2.8267  26.458  < 2e-16 ***
bonds$CouponRate   3.0661     0.3068   9.994 1.64e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.175 on 33 degrees of freedom
Multiple R-squared: 0.7516,     Adjusted R-squared: 0.7441
F-statistic: 99.87 on 1 and 33 DF,  p-value: 1.645e-11
```

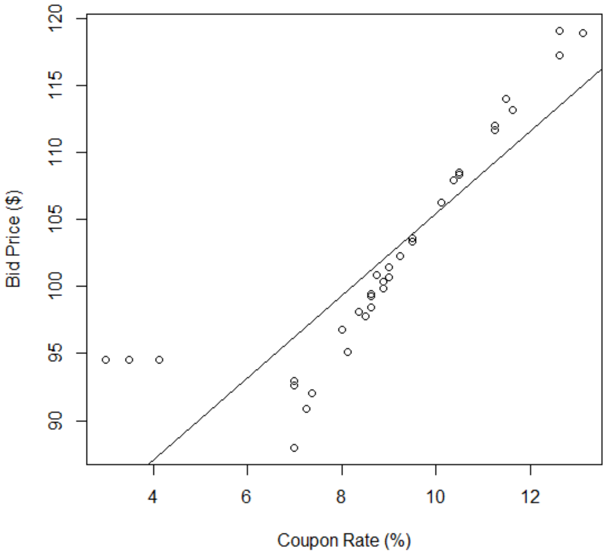
- Coupon rate seems useful in explaining bond price: p-value for testing $H_0 : \beta_2 = 0$ is 1.64e-11; $R^2_{adj} = 0.74$

$$\text{Bond price} = 74.79 + 3.07\text{Coupon rate} + \epsilon$$

- A scatter plot reveals that something might be wrong

```
> plot(bonds$CouponRate,bonds$BidPrice,xlab="Coupon Rate (%)", ylab="Bid Price ($)")
> abline(lsfit(bonds$CouponRate,bonds$BidPrice))
```

- The three data points on the left seem to affect the fitting significantly by dragging the regression line towards themselves


Bid Price ($)
90
95
100
105
110
115
120
4
6
8
10
12
Coupon Rate (%)

- Further investigation reveals that the 4th, 13th and 35th observations correspond to bonds with tax advantages (thus more expensive) and shouldn't have been included for studying regular bonds
- Eliminating the 4th, 13th and 35th observations
- Newly fitted model:
  Bond price $= 57.29 + 4.83$Coupon rate $+ \epsilon$. $R^2_{adj} = 98.47\%$

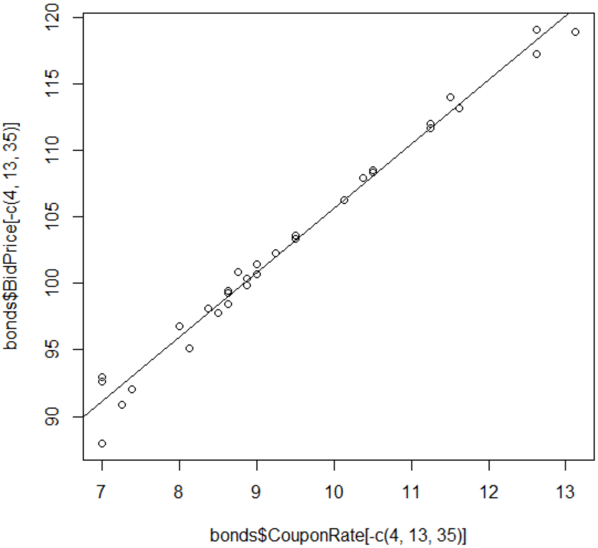
```
> slg1 <- update(slg, subset=(1:35)[-c(4,13,35)])
> plot(bonds$CouponRate[-c(4,13,35)],bonds$BidPrice[-c(4,13,35)])
> abline(slg1)
> summary(slg1)

Call:
lm(formula = bonds$BidPrice ~ bonds$CouponRate, subset = (1:35)[-c(4,
    13, 35)])

Residuals:
    Min      1Q  Median      3Q     Max
-3.1301 -0.3789  0.2240  0.4576  1.8099

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       57.2932     1.0358   55.31   <2e-16 ***
bonds$CouponRate   4.8338     0.1082   44.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.024 on 30 degrees of freedom
Multiple R-squared: 0.9852,     Adjusted R-squared: 0.9847
F-statistic:  1996 on 1 and 30 DF,  p-value: < 2.2e-16
```


bonds$BidPrice[-c(4, 13, 35)]
bonds$CouponRate[-c(4, 13, 35)]
90
95
100
105
110
115
120
7
8
9
10
11
12
13

Residual plots

- **Residual plots** visualize whether the assumptions are violated

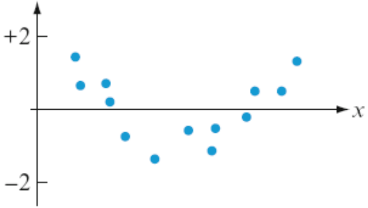
$$\hat{\epsilon}_i = Y_i - \hat{Y}_i$$

- Studentized residuals usually used to eliminate potential impact of leverage points and residual outliers
- Roughly speaking,
  - Residuals should randomly scatter around 0
  - **nonlinear pattern** in residuals indicates nonlinear terms might be needed
  - if residuals exhibit **larger (or smaller) variability for larger x's**, the constant variance assumption might have been violated
  - if **normal plot** of residuals does not exhibit a straight line, normality might have been violated

Checking linearity assumption

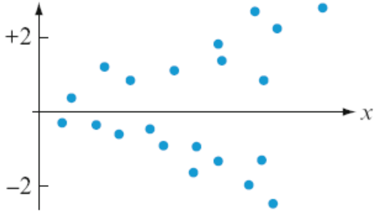
- Scatterplot matrix may identify possible nonlinearity
- Plot residuals against fitted values and regressors
- If the model were true: residuals are randomly scattered around 0
- Nonlinear patterns indicate the necessity for nonlinear terms

- The following residual plot illustrates that a quadratic term X^2 might be needed in the model



- When linearity is violated, transform the data (e.g., consider polynomial regression)

- Plot residuals against fitted values and regressors: if the constant variance assumption were true, residuals have no trend of increasing/decreasing
- The following residual plot illustrates that σ^2 is an increasing function of the regressor



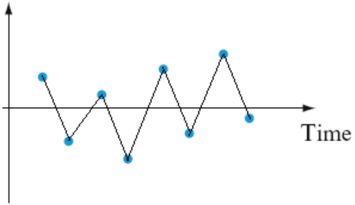
- When constant variance assumption is violated, consider transformation of data or weighted least squares

Checking normality assumption

- Assumption: $\epsilon_i | \mathbf{X} \sim N(0, \sigma^2)$
- Construct a normal probability plot of residuals: if the assumption were true, expect a straight line
- If normality is violated, consider transformation of data
- Or use more realistic distributions and maximum likelihood for parameter estimation

Checking correlation assumption

- Assumption: $\epsilon_i | \mathbf{X}$ **are uncorrelated**
- When data are collected over time, serial correlation is likely
- Compute sample ACF of residuals; plot residuals against time
- The following residual plot illustrates that there is negative serial correlation



- Consider regression with time series data