Regression Analysis Notes

Ziyuan Huang

2025-10-14

Contents

Chapter 1

Prerequisites

1.1 Regression Model

Given *response* $Y \in \mathbb{R}$ and *predictors* $\underline{X} = (X_1, \dots, X_p) \in \mathbb{R}^p$, the goal is to model the relationship between Y and $\underline{X}$ (called to *regress* Y *onto* $\underline{X}$) by

$$Y = f(\underline{X}) + \varepsilon$$

where $f(\cdot)$ is unknown and ε is the noise.

The **input** is a dataset of size n: $\left\{(x_{ij})_{j\in[p]}, y_i\right\}_{i\in[n]}$.

Linear Regression Model

Assume $f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.

- Reduce estimation of functions to estimation of parameters.
- A linear model is linear in *parameters* not linear in predictors!
  - Predictors can be transformed features such as $\ln(X_1)$ or $X_1 X_2$
  - Generalized linear model:

    $$g(\mathbb{E}[Y|X]) = \beta_0 + \sum_{i\in[p]} \beta_i X_i$$

    for some *known* function g.

1.2 Continuous Random Variables

Definition 1.1 (t-distribution)**.** The *t-distribution* with *degrees of freedom* n is determined by the following pdf

$$t_n \sim f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi} \cdot \Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

where $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$.

- t-distribution is symmetric around zero, bell-shaped, but has heavier tails than normal
- $t_n \to \mathcal{N}(0,1)$ as $n \to \infty$

Definition 1.2 (Beta distribution)**.** *Beta distribution* with parameters $a > 0$ and $b > 0$ has pdf

$$\text{Beta}(a,b) \sim f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1.$$

The inverse coefficient is also called the beta-function $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

- Beta distribution is a generalization of uniform distribution to distributions over $[0,1]$
- $\text{Beta}(1,1) = \text{Unif}(0,1)$
- *U-shaped* when $a < 1$ and $b < 1$; and is *bell-shaped* when $a > 1$ and $b > 1$
- Skewed towards 0 when $a < b$; and is skewed towards 1 when $a > b$.

Theorem 1.1 (Beta Conjugate)**.** *Beta distribution is the conjugate distribution for binomial distribution.*

- *If prior is* $p \sim$ *Unif*$(0,1)$*, let* $X \sim$ *Bin*(n,p)*, then*

  $$\mathbb{P}(X = k) = \int \binom{n}{k} x^k (1-x)^{n-k} dx = \frac{1}{n+1}$$

  meaning the expected number of successes are uniformly distributed.
- *Let* $p \sim$ *Beta*(a,b) *be the prior and the number of successes be* k*. Then, the posterior is Beta*$(a+k, b+n-k)$*. Thus,* $a-1$ *represents the number of prior successes and* $b-1$ *represents the number of prior failures.*

Theorem 1.2 (Beta Order Statistic)**.** *Let* X *be the distribution of the* k*-th smallest element of* n *uniform and independent tosses Then,* $X \sim$ *Beta*$(k, n-k)$.

Proof. Observe $\mathbb{P}(X \leq t) = \mathbb{P}(\text{at least k tosses land in [0,t]}) = \sum_{i \geq k} \binom{n}{i} t^i (1-t)^{n-i}$. Thus,

$$\begin{aligned} f_X(t) &= \sum_{i \geq k} \binom{n}{i} \left(i t^{i-1}(1-t)^{n-i} - (n-i) t^i (1-t)^{n-i-1} \right) \\ &= \binom{n}{k} k t^{k-1}(1-t)^{n-k} + \sum_{i=k}^{n-1} \left[-(n-i)\binom{n}{i} + (i+1)\binom{n}{i+1} \right] t^i (1-t)^{n-i-1} \\ &= \frac{n!}{(k-1)!(n-k)!} t^{k-1}(1-t)^{n-k} \sim \text{Beta}(k, n-k+1) \end{aligned}$$

Notice that for integer n, $\Gamma(n) = (n-1)!$. □

Definition 1.3 (Gamma distribution)**.** *Gamma distribution* with *shape* parameter α and *scale* parameter θ has pdf

$$\text{Gamma}(\alpha, \theta) \sim f(x; \alpha, \theta) = \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} e^{-\frac{x}{\theta}}, \quad 0 \leq x < \infty.$$

Lemma 1.1 (Infinite divisibility)**.** *Let $X_i \sim Gamma(\alpha_i, \theta)$ for $i \in [n]$ be* independent*, then*

$$\sum_{i=1}^{n} X_i \sim Gamma\left(\sum_{i \in [n]} \alpha_i, \theta \right).$$

This implies that Gamma distribution is additive in the shape parameter and that it can be decomposed as the sum of arbitrary independent random variables — a property called ***infinite divisibility****.*

Proof. It is sufficient to show the additivity property for X_1 and X_2.

$$\begin{aligned} f_{X_1+X_2}(t) &= \int_0^\infty f_{X_1}(x) f_{X_2}(t-x) dx \\ &= \int_0^t \frac{x^{\alpha_1-1} e^{-x/\theta}}{\Gamma(\alpha_1)\theta^{\alpha_1}} \frac{(t-x)^{\alpha_2-1} e^{-(t-x)/\theta}}{\Gamma(\alpha_1)\theta^{\alpha_2}} dx \\ &= e^{-t/\theta} \int_0^t \frac{x^{\alpha_1-1}(t-x)^{\alpha_2-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\theta^{\alpha_1+\alpha_2}} dx \\ &\overset{x:=tz}{=} \frac{t^{\alpha_1+\alpha_2-1} e^{-t/\theta}}{\Gamma(\alpha_1+\alpha_2)\theta^{\alpha_1+\alpha_2}} \int_0^1 \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} z^{\alpha_1-1}(1-z)^{\alpha_2-1} dz \\ &= \frac{t^{\alpha_1+\alpha_2-1} e^{-t/\theta}}{\Gamma(\alpha_1+\alpha_2)\theta^{\alpha_1+\alpha_2}} \sim \text{Gamma}(\alpha_1+\alpha_2, \theta) \end{aligned}$$

where the last equality used the fact that $\text{Beta}(\alpha_1, \alpha_2) \sim \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} z^{\alpha_1-1}(1-z)^{\alpha_2-1}$. □

Remarks

- $\text{Gamma}(1,\theta) = \text{Exp}(1/\theta)$ and $\text{Gamma}(n,\theta) = \sum_{i=1}^{n} \text{Exp}(1/\theta)$: $\text{Gamma}(n,\theta)$ models the waiting time for *the next* n independent events to occur, compared to the exponential distribution that only models the very *next* event.
- Scaling property: $X \sim \text{Gamma}(\alpha,\theta) \implies cX \sim \text{Gamma}(\alpha, c\theta)$

Theorem 1.3 (Beta and Gamma)**.** *Let $X \sim Gamma(\alpha,\theta)$ and $Y \sim Gamma(\beta,\theta)$ be independent. Then, $\frac{X}{X+Y} \sim Beta(\alpha,\beta)$ and $\frac{X}{X+Y}$ is independent of $X+Y$.*

Proof. We apply the transformation rule: $f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|$. In this case, $x = (X,Y)$ and $y = (\frac{X}{X+Y}, X+Y)$. The mapping $y \mapsto x$ is $x_1 = y_1 y_2$ and $x_2 = (1-y_1)y_2$.

$$\left|\frac{dx}{dy}\right| = \left|\begin{pmatrix} y_2 & y_1 \\ -y_2 & 1-y_1 \end{pmatrix}\right| = \left|\begin{pmatrix} X+Y & \frac{X}{X+Y} \\ -X-Y & \frac{Y}{X+Y} \end{pmatrix}\right| = Y + X.$$

Thus, the joint distribution of $(\frac{X}{X+Y}, X+Y)$ equals

$$\begin{aligned} f_{\frac{X}{X+Y},X+Y}(w,t) &= f_X(wt) f_Y((1-w)t)t \\ &\propto t(wt)^{\alpha-1} e^{-wt/\theta} ((1-w)t)^{\beta-1} e^{-(1-w)t/\theta} \\ &= \underbrace{w^{\alpha-1}(1-w)^{\beta-1}}_{\sim \text{Beta}(\alpha,\beta)} \cdot \underbrace{t^{\alpha+\beta-1} e^{-t/\theta}}_{\sim \text{Gamma}(\alpha+\beta,\theta)} \end{aligned}$$

The separability of the two pdfs shows the independence. □

Definition 1.4 (Chi-squared distribution)**.** A *chi-square distribution* with *degrees of freedom* n is the distribution of a sum of n independent standard normal random variables, denoted as χ^2_n.

- χ^2_n distribution is a special case of the gamma distribution $\chi^2_n = \text{Gamma}\left(\frac{n}{2}, 2\right)$

Theorem 1.4 (Normal over Chi is t-distribution)**.** *Let $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi^2_n$. Then, $\frac{Z}{\sqrt{W/n}} \sim t_n$.*

Proof. We first calculate the distribution of $\sqrt{W/n}$, using the scaling property of the gamma distribution:

$$\mathbb{P}(\sqrt{W/n} \le t) = \mathbb{P}(W/n \le t^2) = \mathbb{P}\left(\text{Gamma}\left(\frac{n}{2}, \frac{2}{n}\right) \le t^2\right)$$

Using the chain rule, we have

$$f_{\sqrt{W/n}}(x) = 2t\left(\frac{n}{2}\right)^{\frac{n}{2}} \cdot \frac{x^{n-2}e^{-\frac{nx^2}{2}}}{\Gamma(\frac{n}{2})} = \frac{n^{n/2}}{2^{n/2-1}}\frac{x^{n-1}e^{-\frac{nx^2}{2}}}{\Gamma(\frac{n}{2})}$$

Define $T := Z/\sqrt{W/n}$. Then, use the ratio distribution formula for independent random variables:

$$\begin{aligned}
f_T(t) &= \int_{-\infty}^{\infty} |x| f_Z(tx) f_{\sqrt{W/n}}(x)dx = \frac{1}{\sqrt{2\pi}}\frac{n^{n/2}}{2^{n/2-1}\Gamma(\frac{n}{2})}\int_0^{\infty} |x|x^{n-1}\exp\left(-\frac{t^2x^2}{2} - \frac{n}{2}x^2\right)dx \\
&= \frac{1}{\sqrt{2\pi}}\frac{n^{n/2}}{2^{n/2-1}\Gamma(\frac{n}{2})}\int_0^{\infty} x^n \exp\left(-\frac{t^2+n}{2}x^2\right)dx \\
&\overset{z:=x^2}{=} \frac{1}{\sqrt{2\pi}}\frac{n^{n/2}}{2^{n/2-1}\Gamma(\frac{n}{2})}\frac{1}{2}\int_0^{\infty} z^{\frac{n-1}{2}}\exp\left(-\frac{t^2+n}{2}z\right)dz \\
&= \frac{1}{\sqrt{2\pi}}\frac{n^{n/2}}{2^{n/2}\Gamma(\frac{n}{2})}\Gamma\left(\frac{n+1}{2}\right)\left(\frac{t^2+n}{2}\right)^{-\frac{n+1}{2}}\int_0^{\infty} f_{\text{Gamma}(\frac{n+1}{2},\frac{2}{t^2+n})}(z)dz \\
&= \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\Gamma(\frac{n}{2})}(1+t^2/n)^{-\frac{n+1}{2}}
\end{aligned}$$

This is exactly the pdf of t_n. □

Definition 1.5 (F-distribution)**.** Let $U \sim \chi^2_m$, $v \sim \chi^2_n$, and U, V be independent. Then, $\frac{U/m}{V/n} \sim F_{m,n}$, an F-distribution with m and n degrees of freedom.

Chapter 2

Estimation

Estimator is a statistic (a function of the observed finite samples/dataset) that is used to estimate an unknown population parameter. **Criterion**: *Least Squares* (LS) or *Sum of Squared Errors* (SSE)

$$\min_{\beta \in \mathbb{R}^{p+1}} \sum_{i \in [n]} \varepsilon_i^2 = \sum_{i \in [n]} \left(y_i - \beta_0 - \sum_{j \in [p]} x_{ij} \beta_j \right)^2 .$$

2.1 Simple Linear Regression

This is the special case when $p = 1$ (i.e., one predictor). The optimization problem is convex and can be solved with the first order condition:

$$\hat{\beta}_1 = \frac{\sum_{i \in [n]} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i \in [n]} (x_i - \bar{x})^2} = \frac{s_{xy}^2}{s_x^2} = r_{xy} \times \frac{s_y}{s_x}$$

where s_{xy}^2, s_x^2, s_y^2, and r_{xy} are sample covariance of X and Y, sample variance of X, sample variance of Y, and sample correlation of X and Y, respectively; and

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

Plugin the above solution into the regression line $y = \hat{\beta}_1 x + \hat{\beta}_0$, we obtain the concise expression

$$\frac{y - \bar{y}}{s_y} = r \frac{x - \bar{x}}{s_x}$$

Remarks

- Let x and y be standardized by sample mean and standard deviation. Then, the regression lines are $y = rx$ and $x = ry$. Notice that the lines are different unless $r = \pm 1$, representing perfect correlation.

2.2 Multiple Linear Regression

This is the general case with $p \geq 1$. Denote $x_i = (1, x_{i1}, \dots, x_{ip})^\top \in \mathbb{R}^{p+1}$ and the data matrix $X := (x_1, \dots, x_n)^\top \in \mathbb{R}^{(p+1)\times n}$. The estimator is given by the MSE solution

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

Remarks

- $H := X(X^\top X)^{-1}X$ is the *projection matrix* onto $\mathrm{Col}(X)$. In other words, linear regression is equivalent to projecting the response vector on to $\mathrm{Col}(X)$ and $\hat{\beta}$ is the projection coefficient.
  - H is also called the *hat matrix* as it converts the response vector y to the "hatted" (i.e., predicted) response vector $\hat{y}$
  - $R := I - H$ is the *residual matrix* as it produces the residual vector $\hat{e} := y - \hat{y}$
- $\hat{\beta}$ is *unbiased*: $\mathbb{E}[\hat{\beta}] = (X^\top X)^{-1}X^\top(X\vec{\beta} + \varepsilon) = \beta$
- When noise are *homoscedasticity* with variance σ^2, then the variance of the estimator equals

$$\mathrm{Var}(\hat{\beta}) = (X^\top X)^{-1}X^\top \mathrm{Var}(\varepsilon) X (X^\top X)^{-1} = \sigma^2 (X^\top X)^{-1}$$

- An unbiased estimator of the noise variance is

$$\hat{\sigma}^2 = \frac{\sum_{i\in[n]}(y_i - \hat{y})^2}{n - (p+1)}.$$

2.3 Error Analysis

Error analysis is a study of the residuals $e_i := y_i - \hat{y}_i$ for all $i \in [n]$.

Lemma 2.1. $\sum_{i\in[n]} e_i = 0$ *or equivalently* $\frac{1}{n}\sum_{i\in[n]} \hat{y}_i = \bar{y}$.

Proof. By the projection theorem, $e \perp \mathrm{Col}(X)$. Then, notice $1 \in \mathrm{Col}(X)$. □

Define *residual sum of squares* (i.e., the LS loss at the optimal estimator) as

$$RSS := \sum_{i\in[n]} e_i^2 = \sum_{i\in[n]} (y_i - \hat{y}_i)^2$$

and the *total sum of squares*

$$TSS := \sum_{i\in[n]} (y_i - \bar{y}_i)^2.$$

We measure the goodness-of-fit using the *coefficient of determination* or R^2

$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2 measures *the proportion of variability of the response variable that can be explained by the predictors.*

Remarks

- RSS alone as the error metric has the problem that it *does not have reasonable units.*
- Variability of the predictors or *regression sum of squares*: $\sum_{i\in[n]}(\hat{y}_i - \bar{y})^2 = TSS - RSS$
- For simple linear regression: $R^2 = r^2$
- Small R^2 does not mean x, y are not linearly related; it just means the variance of noise is high compared to the trend (i.e., low snr)
- Large R^2 does not mean the model is correct (e.g., quadratic model with zero noise)
- R^2 always increase when adding more *predictors* to the model because setting β for the new predictor to zero recovers the old RSS, which equals to the LS under the optimal estimator.

  Adjusted R^2 is used to mitigate the issue of number of predictors:

$$\text{Adjusted } R^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)} = 1 - \frac{\hat{\sigma}^2}{s_y^2}$$

2.4 The Gauss-Markov Theorem

Definition 2.1. A *linear estimator* of the quantity ψ is a linear combination of the *responses*: $\hat{\psi} = c_1 y_1 + \cdots + c_n y_n$.

The linear regression produces the *best linear unbiased estimator* (BLUE) in the following sense.

Theorem 2.1. *Suppose $y = X\beta + \varepsilon$, X is full rank, $\mathbb{E}[\varepsilon] = 0$, and $Var(\varepsilon) = \sigma^2 I$. Consider the quantity $\psi = c^\top \beta$. Then, among all unbiased and linear estimator of ψ, $\hat{\psi} := c^\top \hat{\beta}$ has the minimum variance and is unique.*

Proof. Let $\tilde{\psi} = b^\top y$ be an arbitrary unbiased linear estimator of ψ. Then, we have

$$\text{Var}(\tilde{\psi}) = b^\top \text{Var}(y) b = \sigma^2 \|b\|^2$$

Since $\tilde{\psi}$ is unbiased, $\mathbb{E}[\tilde{\psi}] = b^\top X\beta = c^\top\beta$ for every β. Thus, we must have $b^\top X = c^\top$. The minimum variance of $\tilde{\psi}$ can be obtained from the following optimization problem

$$\min_{b\in\mathbb{R}^n} \|b\|^2 \quad \text{s.t.} \quad b^\top X = c^\top.$$

The lagrangian of this problem is $L(\beta,\lambda) = \|b\|^2 + \lambda^\top(X^\top b - c)$. The first order condition implies $b^* = -\frac{1}{2}X\lambda$ and the dual function is

$$D(\lambda) = -\frac{1}{4}\lambda^\top(X^\top X)\lambda - \lambda^\top c.$$

It is easy to check that the strong duality holds and the dual maximizer satisfies $\lambda^* = -2(X^\top X)^{-1}c$. Thus, we must have $b^* = X(X^\top X)^{-1}c$. So, this minimum variance unbiased linear estimator is

$$\hat{\psi} = (b^*)^\top y = c^\top(X^\top X)^{-1}X^\top y = c^\top\hat{\beta}.$$

The uniqueness is obvious as the optimization is strongly convex. □

Remarks

- In the linear regression model, $f(X)$ is best estimated by $X\hat{\beta}$ by taking $c = (1, X_1, \ldots, X_p)^\top$
- Unbiased cannot be dropped: ridge regression has lower variance than linear regression

2.5 Introducing Predictors

2.5.1 General Theory

Let $X \in \mathbb{R}^{n\times(p+1)}$ be existing data matrix and $Z \in \mathbb{R}^{n\times q}$ be the data matrix consisting additional predictors. The linear regression model can be written as

$$\mathbb{E}[Y|X,Z] = X\delta + Z\gamma$$

Let the optimal (LS-minimal) coefficients be $\hat{\delta}$ and $\hat{\gamma}$. Our goal is to compare $\hat{\delta}$ with $\hat{\beta}$ — the regression coefficients with the old predictors only.

Theorem 2.2 (Coefficients of Additional Predictors)**.** *Assume columns of X and Z are independent so that a unique solution exists. Then,*

- $\hat{\gamma} = (Z^\top RZ)^{-1}Z^\top Ry$ *— projection coefficient of the response vector onto the X-residualized column space of Z (i.e., $\mathcal{C}(X)^\perp \cap \mathcal{C}(Z)$).*

- $\hat{\delta} = (X^\top X)^{-1} X^\top (y - Z\hat{\gamma}) = \hat{\beta} - (X^\top X)^{-1} X^\top Z\hat{\gamma}$ — *The coefficients of the new predictors are determined by the* ***innovations*** *in the new predictors* ***independently****, while the coefficients of the existing predictors should bear the linearly dependent component brought by the new predictors.*

Proof. Rewrite the model as $\mathbb{E}[Y|X,Z] = X\delta + HZ\gamma + (I-H)Z\gamma = (X + HZ)\tilde{\delta} + (I-H)Z\gamma$. Then, it follows

$$\begin{aligned}
\begin{pmatrix} \tilde{\delta} \\ \hat{\gamma} \end{pmatrix} &= \left(\begin{pmatrix} (X+HZ)^\top \\ Z^\top (I-H) \end{pmatrix} \begin{pmatrix} X+HZ & (I-H)Z \end{pmatrix} \right)^{-1} \begin{pmatrix} (X+HZ)^\top \\ Z^\top (I-H) \end{pmatrix} y \\
&= \begin{pmatrix} (X+HZ)^\top (X+HZ) & 0 \\ 0 & Z^\top (I-H) Z \end{pmatrix}^{-1} \begin{pmatrix} (X+HZ)^\top \\ Z^\top (I-H) \end{pmatrix} y \\
\implies \hat{\gamma} &= (Z^\top R Z)^{-1} Z^\top R y.
\end{aligned}$$

The second bullet point is obtained by plugging in $\hat{\gamma}$ in which case the problem becomes regressing $y - Z\hat{\gamma}$ onto X. □

- Due to symmetry, $\hat{\delta}$ can also be viewed as an independent regression coefficients on the innovations in X against Z

Lemma 2.2 (Subspace relationships)**.** *Let H_X, H_Z, and H_{XZ} be the hat matrix when the predictors are X, Z, and both X and Z, respectively. Similarly, we use R_X, R_Z, and R_{XZ} to denote the corresponding orthogonal matrices.*

- *H_{XZ} is the projection onto $\mathcal{C}(X) \oplus \mathcal{C}(Z)$, while R_{XZ} is the projection onto $\mathcal{C}(X) \cap \mathcal{C}(Z)$*
- *$R_{XZ} = R_X R_Z$ and $H_{XZ} = 1 - R_{XZ}$*

Theorem 2.3 (Residuals of additional predictors)**.**

- $R_{XZ} y = R_X (y - Z\hat{\gamma}) = R_Z (y - X\hat{\delta})$
- $RSS = (y - Z\hat{\gamma})^\top R_X (y - Z\hat{\gamma}) = y^\top R_X y - \hat{\gamma}^\top Z^\top R_X y$

Proof. The second bullet point follows from $R_X y - R_X Z\hat{\gamma} \perp R_X y$ as projecting y onto the X-residualized Z is the same as projecting the X-residualized y onto X-residualized Z. □

2.6 R Example

```
library(faraway)
data(gala)
temp = lm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent, data=gala)
summary(temp)
```

```
##
## Call:
## lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
##     data = gala)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -111.679  -34.898   -7.862   33.460  182.584
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.068221  19.154198   0.369 0.715351
## Area        -0.023938   0.022422  -1.068 0.296318
## Elevation    0.319465   0.053663   5.953 3.82e-06 ***
## Nearest      0.009144   1.054136   0.009 0.993151
## Scruz       -0.240524   0.215402  -1.117 0.275208
## Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.98 on 24 degrees of freedom
## Multiple R-squared:  0.7658, Adjusted R-squared:  0.7171
## F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

- *Estimate*: OLS estimation of β
- *Std. Error*: estimated standard deviation of β_i's (i.e., $\sqrt{\hat{\sigma}^2(X^\top X)^{-1}_{ii}}$)
- *Residual standard error*: $\hat{\sigma}$ where *degrees of freedom* is $n-p-1$

Chapter 3

Inference

A methodology to draw conclusions about the unknown population parameters $\beta_0, \dots, \beta_p$ from the estimators $\widehat{\beta}_0, \dots, \widehat{\beta}_p$.

3.1 Hypothesis Tests

Procedure

Consider a null hypothesis H_0 and an alternative hypothesis H_A. Define *p-value* as:

$$\mathbb{P}(\text{observed or more extreme departure from } H_0 \text{ in favor of } H_A | H_0 \text{ is true}).$$

Given a predetermined significance level $\alpha \in (0, 1)$, we reject the null hypothesis if p-value $< \alpha$.

Assumption

To compute the p-value, we need to make assumptions about the model distribution:

$$\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I).$$

Then, $\widehat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2 (X^\top X)^{-1})$ — *sample distribution* of the statistic $\widehat{\beta}$.

3.1.1 T-statistic and F-statistic

Based on the normality assumption, we can now derive the distribution of the t-statistic (Theorem 3.1) and the F-statistic (Theorem 3.2) as follows.

Definition 3.1 (Standard error)**.** The *standard error* (SE) of a statistic (usually an estimator of a parameter, like the average or mean) is the standard deviation of its sampling distribution.

The standard error of $\hat{\beta}_i$ is

$$\mathrm{se}(\hat{\beta}_i) = \sqrt{\sigma^2 (X^\top X)^{-1}_{ii}}$$

which can be estimated by (biased)

$$\widehat{\mathrm{se}}(\hat{\beta}_i) = \sqrt{\hat{\sigma}^2 (X^\top X)^{-1}_{ii}}.$$

Theorem 3.1 (t-statistic)**.** $T := \frac{\hat{\beta}_j - \beta_j}{\widehat{se}(\hat{\beta}_j)} \sim t_{n-(p+1)}$. *This is in contrast to that* $\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim \mathcal{N}(0,1)$

Proof. It is sufficient to show $\frac{\widehat{\mathrm{se}}(\hat{\beta}_j)}{\mathrm{se}(\hat{\beta}_j)} = \sqrt{\frac{\hat{\sigma}^2}{\sigma^2}} \sim \sqrt{\frac{\chi^2_{n-p-1}}{n-p-1}}$. Define the projection matrix $\hat{H} := I - X(X^\top X)^{-1} X^\top$. So, we have

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n-p-1} y^\top H y \\
&= \frac{1}{n-p-1} \left(\varepsilon^\top H \varepsilon + \underbrace{2\varepsilon^\top H X\beta + \beta^\top X^\top H X \beta}_{=0 \text{ by projection theorem}} \right) \\
&= \frac{1}{n-p-1} \varepsilon^\top \left(\sum_{i=1}^{n-p-k} u_i u_i^\top \right) \varepsilon \\
&= \frac{1}{n-p-1} \sum_{i=1}^{n-p-1} (u_i^\top \varepsilon)^2
\end{aligned}$$

where u_i's are the eigenvectors of H associated with the $n-p-1$ non-zero eigenvalues. Since $\|u_i\|_2 = 1$, we have $u_i^\top \varepsilon \sim \mathcal{N}(0, u_i^\top (\sigma^2 I) u_i) = \mathcal{N}(0, \sigma^2 I)$. Therefore, we have $(n-p-1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p-1}$. □

Theorem 3.2 (F-statistic)**.** *Suppose H_0 claims a subset (≥ 1) of predictors as insignificant (e.g., $H_0 : p_1 = p_2 = 0$) and H_A is the alternative. Fit a model under each of H_0 and $H_0 \cup H_A$ and compute RSS_{H_0} and RSS_{H_A} respectively. Then,*

$$F := \frac{(RSS_{H_0} - RSS_{H_0 \cup H_A})/(df_{H_0} - df_{H_0 \cup H_A})}{RSS_{H_0 \cup H_A}/df_{H_0 \cup H_A}}$$

follows F-distribution with degrees of freedom $df_{H_0} - df_{H_0 \cup H_A}$ and $df_{H_0 \cup H_A}$.

Proof. Following the proof of Theorem 3.1, we know that $\frac{\text{RSS}_{H_0}}{\text{df}_{H_0}} = \hat{\sigma}^2_{H_0} \sim \chi^2_{\text{df}_{H_0}}$. Then, the result follows readily from the additivity of chi-squared distribution and the definition of F-distribution. □

- Here, H_0 and H_A should be treated as sets of predictors that are not excluded (due to insignificance). Notice in the F-statistic, the actual alternative hypothesis is $H_0 \cup H_A$, which must be a **superset** of the null hypothesis (or using more predictors). For example, F-test can test $H_0 : \beta_1 = \beta_2 = 0$ and $H_A : \beta_1 = 0$, but cannot test $H_0 : \beta_1 = \beta_2 = 0$ and $H_A : \beta_1 = \beta_3 = 0$.

3.1.2 Test whether a coefficient β_i is significant

$H_0 : \beta_i = 0$ and $H_A : \beta_i \neq 0$ (or more generally $H_0 : \beta_i = c$ for some constant c)

- (two-sided) t-test: p-value $= \mathbb{P}(|\hat{\beta}_i| \geq \hat{\beta}_i(\mathcal{D})|H_0) = 1 - \text{CDF}_{t_{n-p-1}}\left(\frac{\hat{\beta}_i(\mathcal{D})}{\widehat{\text{se}}(\hat{\beta}_i)}\right)$[1]
- F-test: p-value $= 1 - \text{CDF}_{F_{1,n-p-1}}(\hat{F}(\mathcal{D}))$ where $\hat{F}(\mathcal{D})$ is the F-statistic in Theorem 3.2

```
data(savings)
h0 <- lm(sr ~ pop15 + dpi + ddpi, savings)  # Model under H0
h0a <- lm(sr ~ pop15 + pop75 + dpi + ddpi, savings)  # Model under H0 + HA
anova(h0, h0a)
```

```
## Analysis of Variance Table
##
## Model 1: sr ~ pop15 + dpi + ddpi
## Model 2: sr ~ pop15 + pop75 + dpi + ddpi
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     46 685.95
## 2     45 650.71  1    35.236 2.4367 0.1255
```

- For single predictor testing, the F-test and two-sided t-test are equivalent because $T^2 = F$ (statistics equal) and $t_n = F_{1,n}$ (distributions equal)

  By Theorem 2.3,

$$\text{RSS}_{H_0} - \text{RSS}_{H_0 \cup H_A} = \hat{\beta}_i x_i^\top \tilde{R} y = \frac{(x_i^\top \tilde{R} y)^2}{x_i^\top \tilde{R} x_i} = \hat{\beta}_i^2 x_i^\top \tilde{R} x_i$$

[1] The `t value` and `P(>|t|)` fields are respectively the t-statistic and the p-value of two-sided t-test on whether that single parameter is significant (see Section 2.6).

where $\tilde{R}$ is the residual matrix of the $n-1$ predictors without x_i. The last piece is completed by $x_i^\top \tilde{R} x_i = (X^\top X)^{-1}_{ii}$ using block matrix inversion:

$$\begin{aligned}(X^\top X)^{-1} &= \begin{pmatrix} \tilde{X}^\top \tilde{X} & \tilde{X}^\top x_i \\ x_i^\top \tilde{X} & x_i^\top x_i \end{pmatrix}^{-1} \\ &= \begin{pmatrix} * & * \\ * & \left(x_i^\top x_i - x_i^\top \tilde{X}(\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top x_i\right)^{-1} \end{pmatrix} \\ &= \begin{pmatrix} * & * \\ * & \left(x_i \tilde{R} x_i\right)^{-1} \end{pmatrix}.\end{aligned}$$

3.1.3 Test whether a pair/subset can be excluded from the model

$H_0 : \beta_1 = \beta_2 = 0$ and $H_A : \neg H_0$

- Standard application of F-test
- Test whether any of the predictors are useful: $H_0 : \beta_1 = \cdots = \beta_p = 0$ (not including the intercept).[2]

```
data(savings)
h0 <- lm(sr ~ pop15 + ddpi, savings)  # Model under H0
h0a <- lm(sr ~ pop15 + pop75 + dpi + ddpi, savings)  # Model under H0 + HA
anova(h0, h0a)
```

```
## Analysis of Variance Table
##
## Model 1: sr ~ pop15 + ddpi
## Model 2: sr ~ pop15 + pop75 + dpi + ddpi
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     47 700.55
## 2     45 650.71  2    49.839 1.7233   0.19
```

3.1.4 Test whether the pair of predictors has the same effect

$H_0 : \beta_1 = \beta_2$ and $H_A : \beta_1 \neq \beta_2$

Use F-statistic. Under H_0, we have $\mathbb{E}[Y|X] = \beta_0 + \beta_1(X_1 + X_2) + \beta_3 X_3 + \cdots + \beta_p X_p$. H_A is the original regression line.

[2] This result is the last row of R `summary(lm(...))` output (see Section 2.6)

```
h0 <- lm(sr ~ I(pop15 + pop75) + dpi + ddpi, savings)
h0a <- lm(sr ~ pop15 + pop75 + dpi + ddpi, savings)
anova(h0, h0a)
```

```
## Analysis of Variance Table
##
## Model 1: sr ~ I(pop15 + pop75) + dpi + ddpi
## Model 2: sr ~ pop15 + pop75 + dpi + ddpi
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     46 673.63
## 2     45 650.71  1    22.915 1.5847 0.2146
```

3.2 Confidence Interval (CI)

Given the normality assumption of the noise, the CI can be constructed directly from the distributions of $\hat{\beta}_i$'s. Under significance level α or confidence $1-\alpha$:

- Two-sided CI: $\hat{\beta}_i \pm t_{n-p-1}^{(\alpha/2)} \widehat{\text{se}}(\hat{\beta}_i)$
- One-side CI: $(-\infty, \hat{\beta}_i + t_{n-p-1}^{(\alpha)} \widehat{\text{se}}(\hat{\beta}_i)]$
- Simultaneous CI: notice

$$\begin{aligned}
&\frac{(\hat{\beta}-\beta)^\top \left[\widehat{\text{se}}(\hat{\beta})\right]^{-1} (\hat{\beta}-\beta)}{p+1} \\
=&\frac{(\hat{\beta}-\beta)^\top (X^\top X)(\hat{\beta}-\beta)}{(p+1)\hat{\sigma}^2} \\
=&\frac{(\hat{\beta}-\beta)^\top \left[\text{se}(\hat{\beta})\right]^{-1} (\hat{\beta}-\beta)/(p+1)}{\hat{\sigma}^2/\sigma^2} \sim F_{p+1,n-p-1}
\end{aligned}$$

  To see the last term follows $F_{p+1,n-p-1}$, suppose $z \sim \mathcal{N}_{p+1}(0,\Sigma)$, then we have

$$z^\top \Sigma^{-1} z = e^\top \Sigma^{1/2} \Sigma^{-1} \Sigma^{1/2} e = e^\top e \sim \chi^2_{p+1}.$$

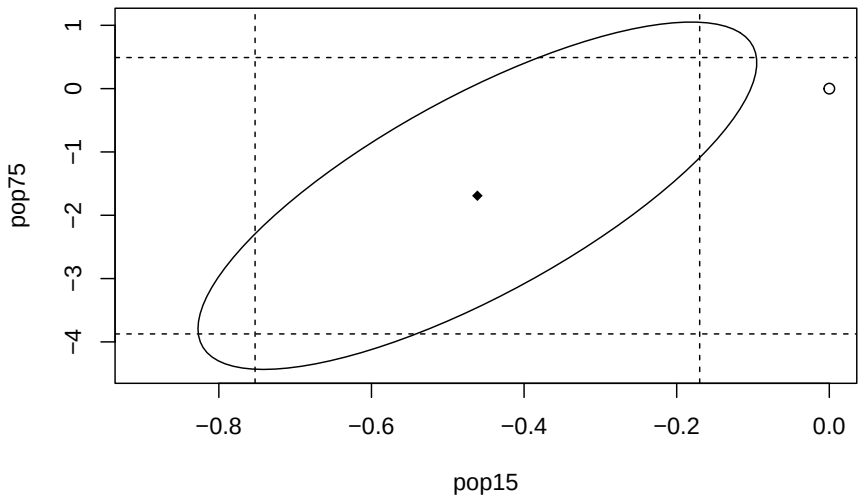
  We bound the statistic from above since F-distribution is one-sided over $[0,\infty)$ and we want to measure how close β is to $\hat{\beta}$. The result is $(\hat{\beta}-\beta)^\top X^\top X(\hat{\beta}-\beta) \le (p+1)\hat{\sigma}^2 F_{p+1,n-p-1}^{(\alpha)}$.

```
library(ellipse)
```

```
##
## Attaching package: 'ellipse'
```

```
## The following object is masked from 'package:graphics':
##
##     pairs
```

```
data(savings)
result <- lm(sr ~ pop15 + pop75 + dpi + ddpi, savings)  # estimate linear regression c
conf <- confint(result)  # compute confidence intervals for every predictor
plot(ellipse(result, c('pop15', 'pop75')), type="l", xlim=c(-0.9,0))  # plot the confi
points(result$coef['pop15'], result$coef['pop75'], pch=18)  # add the estimate to the
points(0, 0, pch=1)  # add the origin to the plot
abline(v=conf['pop15',], lty=2)  # add the confidence interval for pop15
abline(h=conf['pop75',], lty=2)  # add the confidence interval for pop17
```



- Notice that simultaneous CI may contain points that are excluded by individual CI and vice versa.

Chapter 4

Prediction

Two types of predictions:

- Prediction of a *future observation*: y_0
- Prediction of the *future mean response*: $\mathbb{E}[Y|X = x_0]$

4.1 Prediction Interval

We aim to use the predicted value to quantify the high-probability range of the actual response. Let (x_0, y_0) be a future data pair and $\hat{y}_0 := x_0^\top \hat{\beta}$ be its predicted value.

4.1.1 Future observation

We already know that

- $y_0 \sim \mathcal{N}_1(x_0^\top \hat{\beta}, \sigma^2)$
- $\hat{y}_0 \sim \mathcal{N}_1(x_0^\top \hat{\beta}, \sigma^2 x_0^\top (X^\top X)^{-1} x_0)$
- y_0 and $\hat{y}_0$ are independent.

This implies $\hat{y}_0 - y_0 \sim \mathcal{N}_1(0, \sigma^2(1 + x_0^\top (X^\top X)^{-1} x_0))$ and therefore,

$$\frac{\hat{y}_0 - y_0}{\hat{\sigma}\sqrt{1 + x_0^\top (X^\top X)^{-1} x_0}} = \frac{(\hat{y}_0 - y_0)/(\sigma\sqrt{1 + x_0^\top (X^\top X)^{-1} x_0})}{\hat{\sigma}/\sigma} \sim t_{n-p-1}$$

The prediction interval can then be constructed by

$$x_0^\top \hat{\beta} \pm t_{n-p-1}^{(\alpha/2)} \hat{\sigma}\sqrt{1 + x_0^\top (X^\top X)^{-1} x_0}$$

4.1.2 Future mean response

This can be derived directly from the distribution of $\hat{y}_0$:

$$x_0^\top \hat{\beta} \pm t_{n-p-1}^{(\alpha/2)} \hat{\sigma} \sqrt{x_0^\top (X^\top X)^{-1} x_0}$$

- This is exactly the confidence interval for the (BLUE) estimator $\hat{\psi} = x_0^\top \hat{\beta}$ of the quantity $\psi = x_0^\top \beta$.

4.2 R Example

```
data(savings)
result <- lm(sr ~ pop15 + pop75 + dpi + ddpi, savings)
x0 <- data.frame(pop15=35, pop75=2, dpi=1000, ddpi=4)
predict(result, x0, interval="confidence")  # prediction interval for E[Y|X=x0]
```

```
##        fit      lwr      upr
## 1 10.34321 9.093452 11.59297
```

```
predict(result, x0, interval="prediction")  # prediction interval for y0
```

```
##        fit      lwr      upr
## 1 10.34321 2.582946 18.10347
```

Chapter 5

Diagnostics

5.1 Checking Error Assumptions

5.1.1 Constant Noise Variance

Residual plot plots residuals $\hat{\varepsilon}_i$ against predictions $\hat{y}_i$. It can be used to check

- *Linear mean*: linear model (i.e., $\mathbb{E}[Y|X] = X\beta$) implies $\widehat{\mathrm{Cov}}(\varepsilon, \hat{y}) = \widehat{\mathrm{Cov}}(\hat{\varepsilon}, \hat{y}) = 0$. So, the plot should display no pattern but appear as a evenly spread horizontal band of points with mean zero.
  - For population covariance, notice

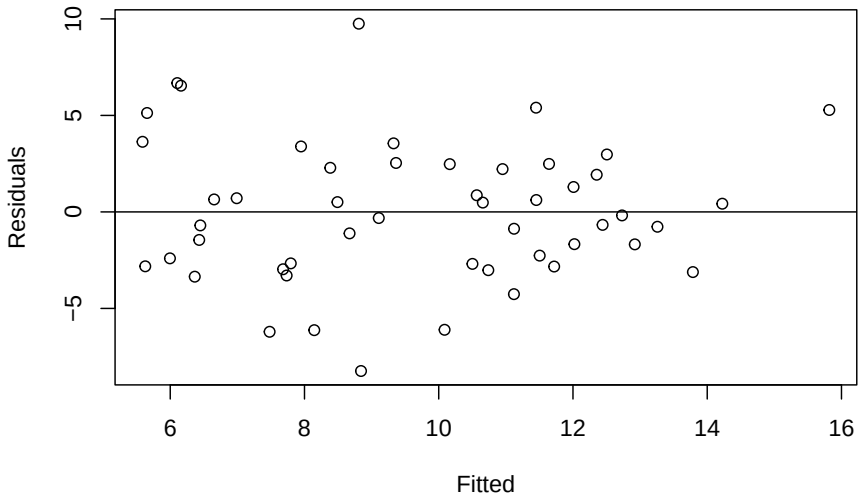
    $$\mathrm{Cov}(\hat{\varepsilon}, \hat{y}) = \mathrm{Cov}((I-H)y, Hy) = (I-H)\mathrm{Var}(y)H^\top = \sigma^2(I-H)H^\top = 0$$

    where H is the hat matrix $X(X^\top X)^{-1}X^\top$.
    This derivation used the homoscedasticity and independence of the noise distribution.
  - For sample covariance, recall $\sum_{i\in[n]} \hat{\varepsilon}_i = 0$ as $\hat{\varepsilon} \in \mathcal{C}(X)^\perp$ and $\mathbb{1} \in \mathcal{C}(X)$. Then, we have

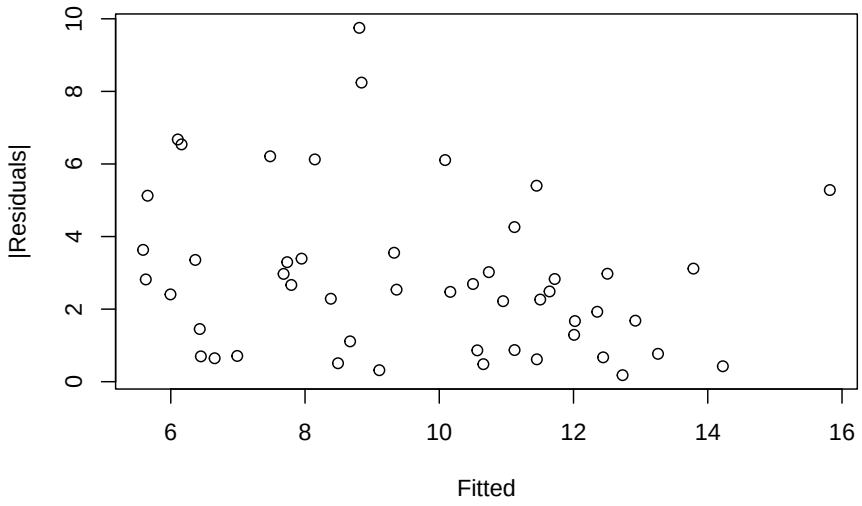
    $$\sum_{i\in[n]} (\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})(\hat{y}_i - \bar{\hat{y}}) = \sum_{i\in[n]} \hat{\varepsilon}_i(\hat{y}_i - \bar{\hat{y}}) = \sum_{i\in[n]} \hat{\varepsilon}_i\hat{y}_i = \hat{\varepsilon}^\top\hat{y} = y^\top(I-H)Hy = 0.$$

    This derivation is independent of the noise distribution as all elements are determined by the algorithm.
- *Homoscedasticity* and *heteroscedasticity*: whether the dispersion of $\hat{\varepsilon}_i$ is constant (resp. increases) in $\hat{y}_i$ for homoscedasticity (resp. heteroscedasticity).

```
library(faraway)
data(savings)
result <- lm(sr ~ pop15 + pop75 + dpi + ddpi, savings)
plot(result$fitted, result$residual, xlab="Fitted", ylab="Residuals")  # ehat v.s. yha
abline(h=0)
```




```
plot(result$fitted, abs(result$residual), xlab="Fitted", ylab="|Residuals|")  # |ehat|
```




```
summary(lm(abs(result$residual) ~ result$fitted))
```

```
##
## Call:
## lm(formula = abs(result$residual) ~ result$fitted)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8395 -1.6078 -0.3493  0.6625  6.7036
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.8398     1.1865   4.079  0.00017 ***
## result$fitted  -0.2035     0.1185  -1.717  0.09250 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.163 on 48 degrees of freedom
## Multiple R-squared:  0.05784,    Adjusted R-squared:  0.03821
## F-statistic: 2.947 on 1 and 48 DF,  p-value: 0.0925
```

- The last method regress $|\hat{\varepsilon}_i|$[1] onto $\hat{y}_i$ and test $H_0 : \beta_1 = 0$ v.s. $H_A : \beta_1 \neq 0$.

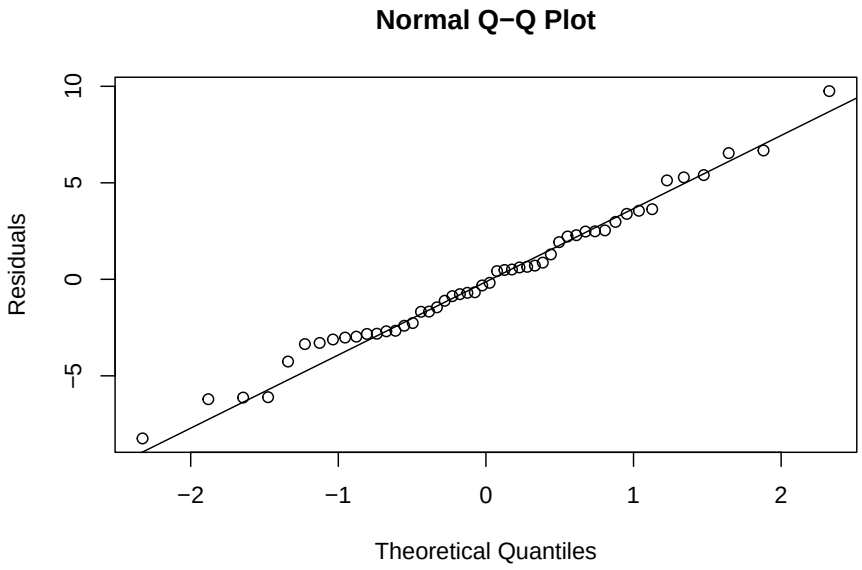
[1] Must have the absolute value as we are testing for *variance*.

5.1.2 Checking Noise Normality

5.1.2.1 QQ-plot

- Sort the residuals $\hat{\varepsilon}_{[1]} \leq \hat{\varepsilon}_{[2]} \leq \cdots \leq \hat{\varepsilon}_{[n]}$ — *order statistic*
- Compute the n percentiles $u_i = F^{-1}\left(\frac{i}{n+1}\right)$ where F is the CDF of the testing distribution
- Plot $\hat{\varepsilon}_{[i]}$ against u_i
- The data follows the distribution if the scatter points lie on a straight line

```r
## Example QQ plot
qqnorm(result$residual, ylab="Residuals")
qqline(result$residual)
```



Remarks

- Should be tested after ensuring linear mean and constant noise variance
- Possibilities of non-normality (both 1 and 2 are heavy-tailed distribution):

  1. Skewed distribution: log-normal
  2. Long-tailed distribution: Cauchy
  3. Short-tailed distribution: uniform with finite support

5.1.3 Shapiro-Wilk test for normality

H_0: $x_1, \dots, x_n$ are sampled from a normally distributed population

```
shapiro.test(result$residual)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  result$residual
## W = 0.98698, p-value = 0.8524
```

Shapiro-Wilk test is not very helpful compared to QQ plot because

- When n is small, the test has little power
- When n is large, we can use the asymptotic distribution (e.g., central limit theorem) to do inference (i.e., hypothesis tests or CI).[2]

5.2 Finding Unusual Points

- *Outliers*: large difference between the response y_i and the mean $x_i^\top \beta$
- *High-leverage points*: large difference between the predictor vector x_i for the ith case and the center of the X-data

5.2.1 Leverage

The *leverage* of a point i is $h_i := H_{ii}$ where H is the hat matrix $H := X(X^\top X)^{-1}X^\top$.

1. h_i depends on on X
2. $\mathrm{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_i)$
3. $\sum_{i\in[n]} h_i = \mathrm{Tr}(H) = p + 1$
4. $\frac{1}{n} \le h_i \le 1$ for all $i \in [n]$

Remarks

- Average leverage is $\frac{p+1}{n}$, so generally leverage larger than $\frac{2(p+1)}{n}$ can be considered high
- Let $\tilde{x}$ be the "reduced" data by removing the first constant value. Then,

$$h_i = \frac{1}{n} + (\tilde{x}_i - \bar{\tilde{x}})^\top (\tilde{X}_C^\top \tilde{X}_C)^{-1} (\tilde{x}_i - \bar{\tilde{x}})$$

  where $\tilde{X}_C = \begin{bmatrix} (\tilde{x}_1 - \bar{x}) & \dots & (\tilde{x}_n - \bar{x}) \end{bmatrix}^\top$

[2] Notice that we only requires normality in the inference step.

5.3 Checking Model Structure

Tests that imply the underlying structure of the model as well as suggestions on how to improve the structure of the model.

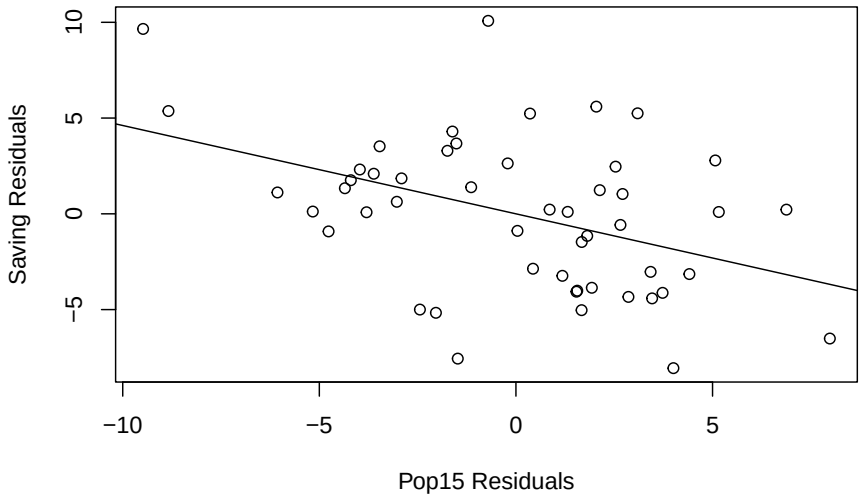
5.3.1 Exploratory analysis

- Plot y v.s. each individual x_i to investigate the relationship and/or linearity between each individual predictor.
- Usually done before fitting a model
- *Drawback*: other predictors may affect the relationship between y and x_i

5.3.2 Partial regression plot

- Isolate the effect of x_i on y
- Regress y on all x except x_i, get residuals $\hat{\delta}$ — take out effect of other X from y
- Regress x_i on all x except x_i, get residuals $\hat{\gamma}$ — take out effect of other X from x_i
- Plot $\hat{\delta}$ v.s. $\hat{\gamma}$
  - The slope is $\hat{\beta}_j$
  - Can be used for linearity, outliers, and influential point tests

```
data(savings)
## Partial regression plot
result <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data=savings)
delta <- residuals(lm(sr ~ pop75 + dpi + ddpi, data=savings))
gamma <- residuals(lm(pop15 ~ pop75 + dpi + ddpi, data=savings))
plot(gamma,delta, xlab="Pop15 Residuals", ylab="Saving Residuals")
temp <- lm(delta ~ gamma)
abline(reg=temp)
```




```
## The slope of the partial regression plot
## Notice this is the same as beta(pop15) in the summary after this
coef(temp)
```

```
##   (Intercept)         gamma
## -1.545720e-16 -4.611931e-01
```

```
coef(result)
```

```
##   (Intercept)         pop15         pop75           dpi          ddpi
## 28.5660865407 -0.4611931471 -1.6914976767 -0.0003369019  0.4096949279
```