Formula Sheet for STAT 308: Applied Regression Analysis

Kejin Wu
Department of Mathematics and Statistics, Loyola University Chicago
kwu8@luc.edu

2025-11-29

Abstract

The materials in this formula sheet is collected from the lecture slides of STAT 308. I will keep updating this formula sheet to make it self-contained. If you see any typo, please contact me. Thank you! Partial credit is given to Dr. Stuart at LUC. The course structure is based on his lecture design.

Contents

Simple Linear Regression

Straight Line Model

Mathematically, a straight line is defined as

$$y = \beta_0 + \beta_1 x$$

where

- β_0 is the intercept; the value of y when $x = 0$
- β_1 is the slope; the change in y for a one unit change in x

Simple Linear Regression Motivation

We need to add an additional term to account for the fact that there are differences between a straight line and the actual data. These are defined as the **errors/residuals** of the linear model, and are noted by the greek letter ϵ.

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

Assumptions

- Existence: For any given value of X, Y is a random variable with a certain probability distribution with a finite mean and variance. Define:
  - $\mu_{Y|X}$ - the population mean of Y for a fixed X
  - $\sigma^2_{Y|X}$ - the population variance of Y for a fixed X
- Independence: The observed values of Y are statistically independent of one another given X. Counterexample:
  - X = Daily closing price of the S&P 500
  - Y = Daily closing price of Bitcoin
- Linearity: $\mu_{Y|X}$ is a straight line function of X. In other words we say that

  $$\mu_{Y|X} = \beta_0 + \beta_1 X$$

  where β_0 and β_1 are defined here as the population intercept and slope, respectively.

The next two assumptions discuss the distribution of ϵ.

- Homoscedasticity: The variance of Y is the same for different given values of X. Mathematically, this is equivalent to saying

  $$\sigma^2_{Y|X} = \sigma^2,$$

  or in other words $\sigma^2_{Y|X_i} = \sigma^2_{Y|X_j}$ for different i and j.
- Normality: For any fixed value of x, Y is normally distributed. This fact makes analysis of the data easier.

All of these assumptions put together lead us to the assumption of the distribution of the residuals:

$$\epsilon \sim N(0, \sigma^2).$$

Counter example

Least Square Estimation for Simple Linear Regression

The "least squares" method provides estimates that minimizes the sum of the squared differences between observed y and its estimates from the regression line. In other words, the least squares methods finds $\hat{\beta}_0$ and

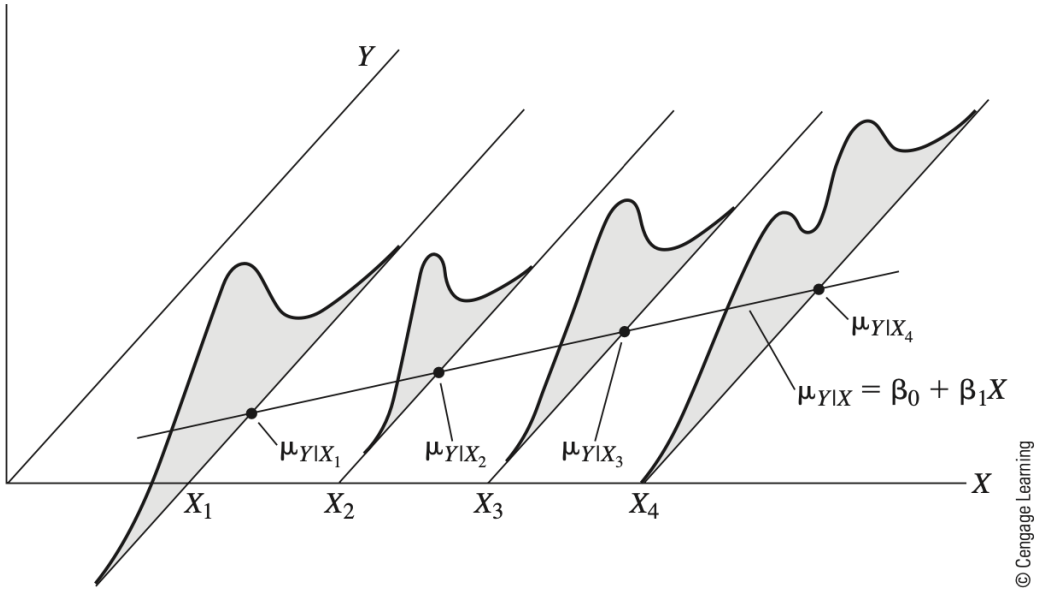


Figure 1: One counter example

$\hat{\beta}_1$ that minimizes the sum of squared (due to) errors

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2,$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Estimators

The least squares method produces the estimates

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

and

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

Equivalent Expression of $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{r s_y}{s_x};$$

where

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right); s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2; s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

We call r the sample correlation between X and Y; s_x^2 and s_y^2 are variance of X and Y, respectively.

Interpretation of the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$

- $\hat{\beta}_1$: When independent variable x increases by 1 unit, the change in the predicted/expected value of dependent variable y is $\hat{\beta}_1$ units.
- $\hat{\beta}_0$: The predicted/expected value of y when $x = 0$.

Estimating the Variance (σ^2)

Recall that, without knowledge of X, our estimate of the variance of Y is

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

The summation can be thought of as the sum of squared distance between an observed y_i and its prediction $\bar{y}$.

The estimate for the variance of the linear regression model, σ^2 is calculated in a similar manner

$$\hat{\sigma}^2 = \frac{1}{n-2}SSE = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2.$$

We also call $\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$ Residual Sum of Squares (RSS).

Property of $\hat{\beta}_1$ and $\hat{\beta}_0$

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are both **normally distributed**.
- $\hat{\beta}_0$ and $\hat{\beta}_1$ are both **unbiased**: $E\left(\hat{\beta}_0\right) = \beta_0$ and $E\left(\hat{\beta}_1\right) = \beta_1$.
- $\text{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$.
- $\text{Var}\left(\hat{\beta}_0\right) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$.

Inference on linear regression model

Recall from previous lecture slides, we said that

$$\frac{\bar{Y} - \mu}{\frac{s_Y}{\sqrt{n}}} \sim t_{df=n-1}$$

where $\bar{Y}$ is the sample mean of normal samples and s_Y is the sample standard deviation of the samples.

We can obtain a similar conclusion regarding the distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ assuming the residuals are normally distributed with a common variance.

$$\frac{\hat{\beta}_0 - \beta_0}{s_{\hat{\beta}_0}} \sim t_{df=n-2}$$

and

$$\frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \sim t_{df=n-2}$$

where $s_{\hat{\beta}_0}^2 = \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_X^2}\right)$ and $s_{\hat{\beta}_1}^2 = \frac{\hat{\sigma}^2}{(n-1)s_X^2}$. They are standard errors of estimator $\hat{\beta}_0$ and $\hat{\beta}_1$ respectively.

(Note: Often times, inference on β_0 is not meaningful to us.).

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

Revisit t distribution

t distribution: $T := \frac{Z}{\sqrt{V/\nu}} \sim t_{df=\nu}$

- $Z \sim N(0,1)$
- $V \sim \chi^2_{df=\nu}$

We also need Z and V are independent.

For $\hat{\beta}_1$

- $\hat{\beta}_1, \bar{Y}$ and $\hat{\sigma}^2$ are mutually independent.
- $\frac{(n-2)\hat{\sigma}^2}{\sigma^2}$ has a chi square distribution with $n-2$ degrees of freedom.
-

Based on these claims, can you show

$$\frac{\hat{\beta}_0 - \beta_0}{s_{\hat{\beta}_0}} \sim t_{df=n-2}$$

Confidence intervals for β_1

Define t^*_α as the quantile from the t-distribution such that $Pr\left(-t^*_\alpha \leq \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \leq t^*_\alpha\right) = 1-\alpha$.

Solving for β_1 in the center, we have $Pr\left(\hat{\beta}_1 - t^*_\alpha \times s_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + t^*_\alpha \times s_{\hat{\beta}_1}\right) = 1-\alpha$.

A $100 \times (1-\alpha)\%$ confidence interval for the population slope, β_1, is

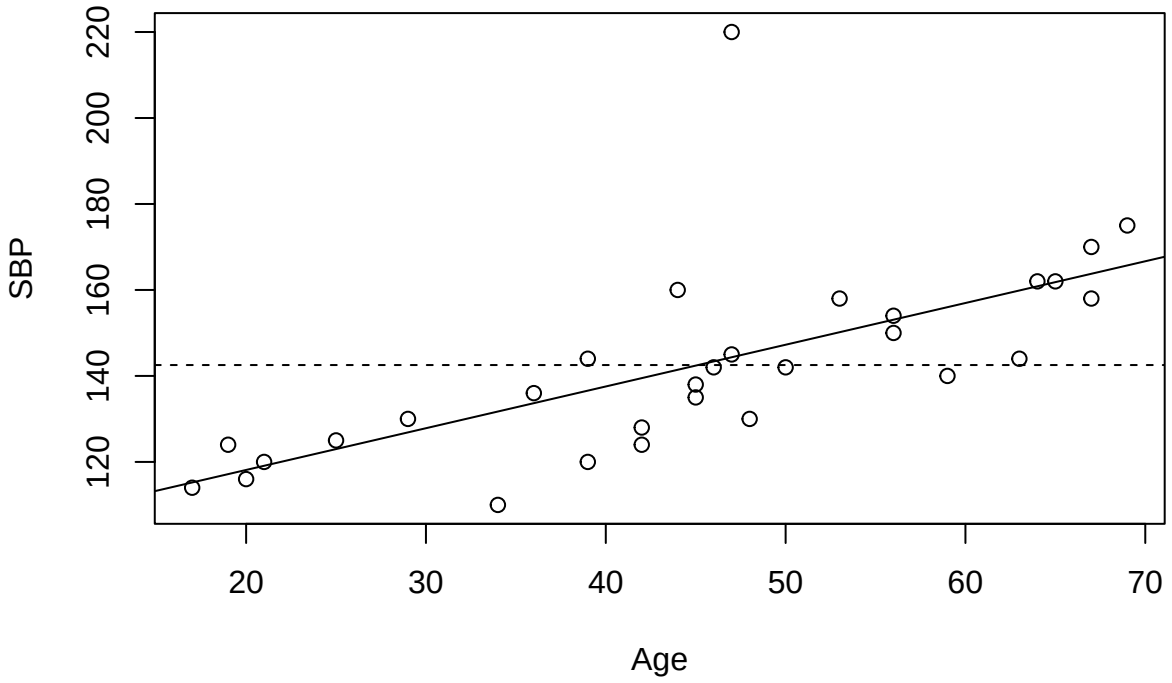
$$\hat{\beta}_1 \pm t_{1-\frac{\alpha}{2}, n-2} \times s_{\hat{\beta}_1}$$

where $t_{1-\frac{\alpha}{2}, n-2}$ is the $1-\frac{\alpha}{2}$ quantile of the t-distribution with $n-2$ degrees of freedom.

Hypothesis testing for β_1

In terms of hypothesis testing, we are considered with testing if a linear model for our response variable with the predictor variable included is statistically significantly better at predicting than the model that does not include the predictor variable.

In other words we are testing to see if the solid line in the below graph ($\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$) is significantly better at predicting Y than the horizontal dashed line ($\hat{y} = \bar{y}$).



Idea We can think of this in terms of a reduced model and a full model:

- Reduced Model: $\hat{Y} = \beta_0$
- Full Model: $\hat{Y} = \beta_0 + \beta_1 X$

The null hypothesis can be thought of as what is missing from the full model to the reduced model.

If $\beta_1 = 0$, the predictor variable goes away.

- $H_0 : \beta_1 = 0$

The alternative hypothesis is the exact opposite of the null hypothesis, and what we are trying to check.

- $H_a : \beta_1 \neq 0$

If H_0 is true, then

$$\frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} \sim t_{df=n-2}$$

Because, under H_0, we are assuming that the true population slope, $\beta_1 = 0$.

Therefore, our test statistic is $t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$.

p-value: $2 \times P(T \geq |t|)$. Note that T represents a t-distributed random variable with $n - 2$ degrees of freedom.

Decision:

- If $p - value < \alpha$, reject H_0. We have statistically significant evidence that X is significant in predicting Y **through a linear relationship**.
- If $p - value \geq \alpha$, do not reject H_0. We do not have statistically significant evidence that X is significant in predicting Y **through a linear relationship**.

Confidence Intervals for $\mu_{Y|X}$ for $X = x_0$

Suppose we are interested in inference for $\mu_{Y|x_0}$, the mean of Y for all members of the population for any given value of $X = x_0$.

We have already shown an estimate of $\mu_{Y|x_0}$,

$$\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

We can also that

$$\frac{\hat{\mu}_{Y|x_0} - \mu_{Y|x_0}}{s_{\hat{\mu}_{Y|x_0}}} \sim t_{df=n-2}$$

where $s^2_{\hat{\mu}_{Y|x_0}} = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_X^2} \right)$ is the estimated variance of $\hat{\mu}_{Y|x_0}$. $s_{\hat{\mu}_{Y|x_0}}$ is the standard error of $\hat{\mu}_{Y|x_0}$.

Therefore, we can calculate a $100 \times (1 - \alpha)\%$ confidence interval for $\mu_{Y|x_0}$ as

$$\hat{\mu}_{Y|x_0} \pm t_{1-\frac{\alpha}{2}, n-2} \times s_{\hat{\mu}_{Y|x_0}}.$$

Prediction Intervals for Y for $X = x_0$

Perhaps instead of calculating an interval for the mean of Y for all individuals where $X = x_0$, we are interesting in an interval for a single individual where $X = x_0$.

Note that the variance of a predict for a single individual is **larger** than the variance of a 'predict' for a group of individuals.

$$Y_{x_0} = \mu_{Y|x_0} + \epsilon = \beta_0 + \beta_1 x_0 + \epsilon.$$

Naturally, we now have to incorporate two sources of variability

- the uncertainty in the estimate of the mean $\mu_{Y|X_0}$, i.e., $\hat{Y}_{x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$.
- the variability in the errors ϵ.

Using these pieces of information, we can say that

$$\frac{\hat{Y}_{x_0} - Y_{x_0}}{s_{\hat{Y}_{x_0} - Y_{x_0}}} \sim t_{df=n-2}.$$

where $s^2_{\hat{Y}_{x_0} - Y_{x_0}} = \hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_X^2} \right)$.

Therefore, we can calculate a $C = 100 \times (1 - \alpha)\%$ **prediction** interval for and individual Y_{x_0} as

$$\hat{Y}_{x_0} \pm t_{1-\frac{\alpha}{2}, n-2} \times s_{\hat{Y}_{x_0} - Y_{x_0}}.$$

Performing Least Squares Regression in R

The function `lm` performs least squares regression in R. Let's run a linear regression for systolic blood pressure by age based on bloodpressure data set.

```
mod_blood <- lm(SBP ~ Age,bloodpressure)
summary(mod)
```

```
##
## Call:
## lm(formula = SBP ~ Age, data = bloodpressure)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -21.724  -6.994  -0.520   2.931  75.654
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  98.7147    10.0005   9.871 1.28e-10 ***
## Age           0.9709     0.2102   4.618 7.87e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.31 on 28 degrees of freedom
## Multiple R-squared:  0.4324, Adjusted R-squared:  0.4121
## F-statistic: 21.33 on 1 and 28 DF,  p-value: 7.867e-05
```

$\hat{SBP} = 98.71 + 0.9709 Age$

R Code to do predictions

Set `interval = "confidence"` and `interval = "prediction"` to make prediction interval for future average value and individual value, respectivaly.

```
newdata <- data.frame(Age = 55)
predict(mod_blood, newdata, interval = "confidence", level = 0.9)
```

```
##        fit     lwr      upr
## 1 152.1126 145.681 158.5442
```

```
predict(mod_blood,newdata,interval = "prediction",level = 0.9)
```

```
##        fit      lwr      upr
## 1 152.1126 121.9656 182.2596
```

Regression to Mean Effects

According to any least squares regression line, what do we expect the dependent variable to be for independent variable as $\bar{x} + s_x$?

$$\begin{aligned}
\hat{y} &= \hat{\beta}_0 + b_1 (\bar{x} + s_x) \\
&= \left(\bar{y} - \hat{\beta}_1 \bar{x}\right) + b_1 \bar{x} + b_1 s_x \\
&= \bar{y} + \frac{r s_y}{s_x} \cdot s_x \\
&= \bar{y} + r s_y.
\end{aligned}$$

We predict y to be r times its standard deviations above the mean.

Because $-1 \leq r \leq 1$, this leads to the result that we never predict y to be more standard deviations above or below the mean than x is, which is called **regression to the mean**.

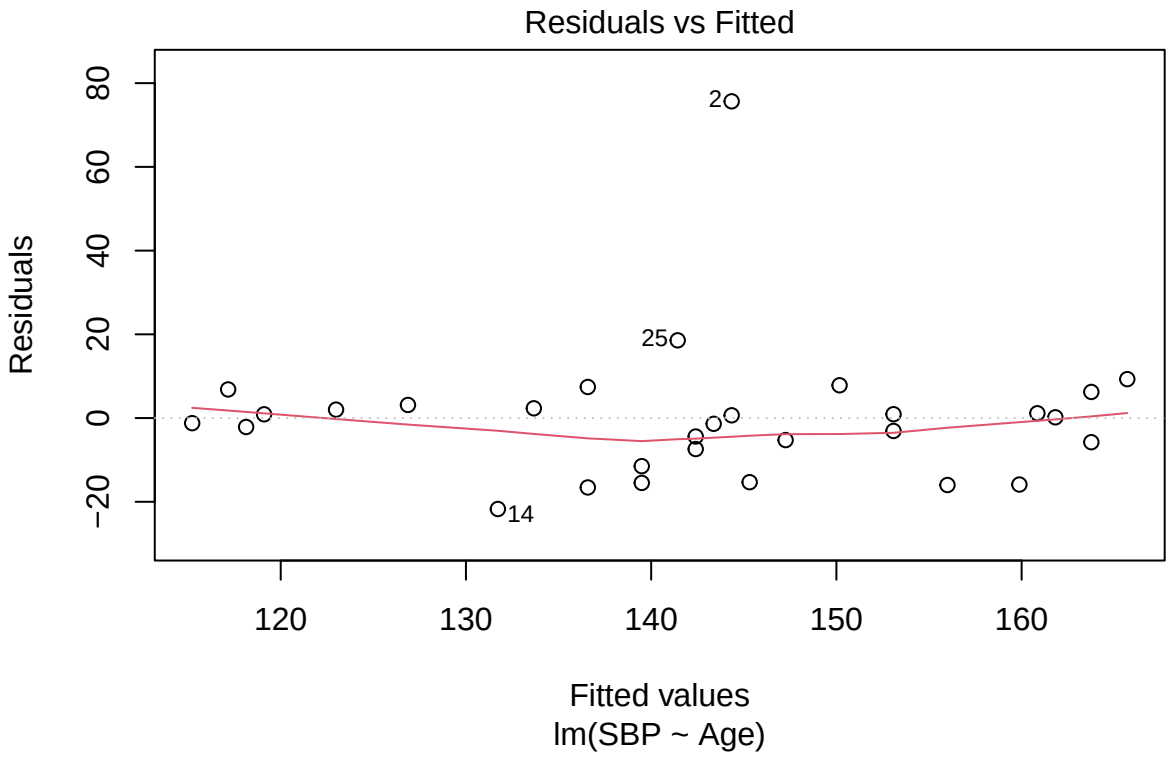
Check Regression Model

- Existence and Independence: Logic based approaches, common sense
- Linearity: Check the scatter plot
- Homoscedasticity (Common Variance): Residual Plot
- Normality: QQ-plot

Residual Plot

Residual Plot: A plot of the residuals from a least squares regression model, plotted against the fitted values $\hat{Y}$.

```
# plot(mod,1) gives a residual plot from a linear model
plot(mod,1)
```



The red curve in this plot is a nonparametric smooth line added to help visualize any patterns or structures in the residuals.

We want to see no pattern at all, or some type of random scatter around a horizontal line at zero. We also want to see the same type of spread of the residuals across different values of $\hat{y}$.

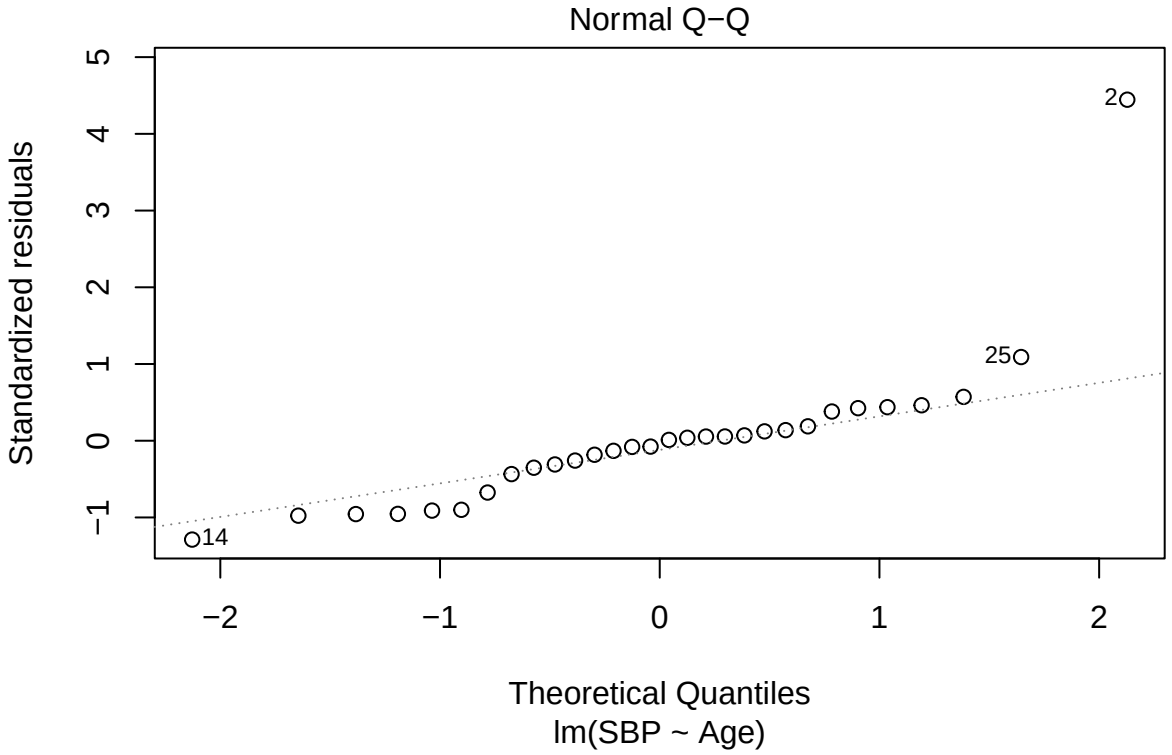
For the residual plot of the blood pressure linear model, the points appear to be evenly spread and randomly scattered around zero. Therefore, the assumption of **homoscedasticity (common variance)** is not violated.

Quantile-Quantile (QQ) Plot

Quantile-Quantile (QQ) Plot: A plot that visualizes the residuals from a model (fitted linear model) against an assumed error distribution (normal distribution for the least squares model).

For our model, we want to check if the residuals are normally distributed.

```
# plot(mod,2) produces a QQ-plot from a linear model
plot(mod,2)
```



We want to see points that fall closely to a line.

Multiple Linear Regression

Multiple Linear Regression: An extension of simple linear regression into where we can estimate a response variable Y based on multiple explanatory variables $X_1, X_2, \ldots, X_{k-1}$. $k-1$ is the number of explanatory variables in our linear structure.

Needed assumptions

- Existence: For any given value of $X_1, \ldots, X_{k-1}$, Y is a random variable with a certain probability distribution with a finite mean and variance.
- Independence: The observed values of Y are statistically independent of one another given $X_1, \ldots, X_{k-1}$
- Linearity: $\mu_{Y|X} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{k-1} X_{k-1}$ or equivalently

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{k-1} X_{k-1} + \epsilon$$

where ϵ is the error of the multiple linear model.

- Homoscedasticity: The variance of Y is the same for different given values of $X_1, \ldots, X_{k-1}$, i.e., $\sigma^2_{Y|X_1,X_2,\ldots,X_{k-1}} = \text{Var}\,(Y \mid X_1, X_2, \ldots, X_{k-1}) \equiv \sigma^2$.
- Normality: For any fixed value of $X_1, \ldots, X_{k-1}$, Y is normally distributed. This fact makes analysis of the data easier.

Note, these assumptions on previous slide are the same for simple linear regression, just extended to multiple linear regression)

All of these assumptions put together lead us to our full mathematical model we will assume:

$$Y \sim N(\beta_0 + \beta_1 X_1 + \cdots + \beta_{k-1} X_{k-1}, \sigma^2).$$

"Least squares" regression estimate

Recall from the simple linear regression notes that, the estimates for the regression parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ can be found by minimizing the sum of squared errors

$$\text{SSE} = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2.$$

For multiple linear regression, this estimation procedure is the same. More specifically, to find estimates for the regression parameters $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{k-1}$, we need to minimize

$$\text{SSE} = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_{k-1} x_{i,k-1}))^2.$$

where $x_{i,j}$ is the i^{th} observation from variable j for $j = 1, \ldots, k-1$.

Confidence Intervals for regression parameters

For $j = 0, 1, \cdots, k-1$, we can say

$$\frac{\hat{\beta}_j - \beta_j}{s_{\hat{\beta}_j}} \sim t_{df=n-k}$$

where $s_{\hat{\beta}_j}$ is the the standard error for $\hat{\beta}_j$.

Therefore, a corresponding $100 \times (1-\alpha)\%$ confidence interval for β_j can be written as

$$\hat{\beta}_j \pm t^*_{1-\frac{\alpha}{2}, n-k} \times s_{\hat{\beta}_j}$$

Use `confint` in R to compute CI for one specific coefficients.

```
allgreen <- read.csv("AllGreen.csv")
mod <- lm(NetSales ~ Advertising + SqFt,allgreen)
# For square footage, let's use the R command confint()
confint(mod,"SqFt",level=0.95)
```

```
##        2.5 %   97.5 %
## SqFt 30.1952 60.98805
```

```
confint(mod,"Advertising",level=0.95)
```

```
##                2.5 %   97.5 %
## Advertising 20.12197 36.52828
```

Confidence Intervals for MLR Predictions

Suppose we have some values of the explanatory variables $\boldsymbol{x}_0$. The mean of the response variable for our linear model is defined as

$$\mu_{Y|\boldsymbol{x}_0} = \boldsymbol{x}_0 \boldsymbol{\beta},$$

an estimate of this mean is

$$\hat{\mu}_{Y|\boldsymbol{x}_0} = \boldsymbol{x}_0 \hat{\boldsymbol{\beta}}.$$

Like in simple linear regression, we can say that

$$\frac{\hat{\mu}_{Y|\boldsymbol{x}_0} - \mu_{Y|\boldsymbol{x}_0}}{\sqrt{\hat{\text{Var}}(\hat{\mu}_{Y|\boldsymbol{x}_0})}} \sim t_{df=n-k},$$

where $\hat{\text{Var}}(\hat{\mu}_{Y|\boldsymbol{x}_0})$ is the estimate of the variance of $\hat{\mu}_{Y|\boldsymbol{x}_0}$

Therefore, we can calculate a $C = 100 \times (1-\alpha)\%$ confidence interval for $\mu_{Y|\boldsymbol{x}_0}$ as

$$\hat{\mu}_{Y|\boldsymbol{x}_0} \pm t_{1-\frac{\alpha}{2},n-k} \times \sqrt{\hat{\text{Var}}(\hat{\mu}_{Y|\boldsymbol{x}_0})}.$$

Similarly, we can estimate a prediction for a particular individual for a given $\boldsymbol{x}_0$ as

$$\hat{Y}_{\boldsymbol{x}_0} = \boldsymbol{x}_0\hat{\boldsymbol{\beta}},$$

$$\frac{\hat{Y}_{\boldsymbol{x}_0} - Y_{\boldsymbol{x}_0}}{\sqrt{\hat{\text{Var}}(\hat{Y}_{\boldsymbol{x}_0} - Y_{\boldsymbol{x}_0})}} \sim t_{df=n-k},$$

where $\hat{\text{Var}}(\hat{Y}_{\boldsymbol{x}_0} - Y_{\boldsymbol{x}_0})$ is the estimate of the variance of $\hat{Y}_{\boldsymbol{x}_0} - Y_{\boldsymbol{x}_0}$ and we can calculate a $100 \times (1-\alpha)\%$ prediction interval for $Y_{\boldsymbol{x}_0}$ as

$$\hat{Y}_{\boldsymbol{x}_0} \pm t_{1-\frac{\alpha}{2},n-k} \times \sqrt{\hat{\text{Var}}(\hat{Y}_{\boldsymbol{x}_0} - Y_{\boldsymbol{x}_0})}.$$

Calculate a prediction, 95% confidence interval, and a 95% prediction interval (PI) for the net sales for stores at 3500 square feet and that spend 10,000 dollars on advertising.

```
# To calculate these confidence and prediction intervals,
# we will use the predict() function
newdata <- data.frame(SqFt = 3.5,
                      Advertising = 10)
predict(mod,newdata)
```

```
##        1
## 348.3281
```

```
predict(mod,newdata,interval="prediction",level = 0.95)
```

```
##        fit      lwr      upr
## 1 348.3281 240.8506 455.8057
```

Multiple Linear Regression in Matrix Form

Everything we need to perform multiple linear regression can be done using linear algebra!

Matrix Transpose

Consider a matrix $\boldsymbol{A}_{2\times 3}$. The transpose of the matrix $\boldsymbol{A}'$ is the same matrix with the rows and columns flipped (i.e. the 1st row of $\boldsymbol{A}$ is the 1st column of $\boldsymbol{A}'$)

Example

$$A = \begin{bmatrix} 2 & 3 & 1 \\ 1 & 1 & 2 \end{bmatrix} \; ; \; \mathbf{A}' = \begin{bmatrix} 2 & 1 \\ 3 & 1 \\ 1 & 2 \end{bmatrix}$$

Matrix Addition

Consider matrices $\boldsymbol{A}_{2\times 3} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$ and $\boldsymbol{B}_{2\times 3} = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix}$, then

$$\boldsymbol{A} + \boldsymbol{B} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & a_{13} + b_{13} \\ a_{21} + b_{21} & a_{22} + b_{22} & a_{23} + b_{23} \end{bmatrix}$$

Matrix Multiplication

Consider matrices $\boldsymbol{A}_{2\times 3} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$ and $\boldsymbol{B}_{3\times 1} = \begin{bmatrix} b_{11} \\ b_{21} \\ b_{31} \end{bmatrix}$, then

$$\boldsymbol{A} \times \boldsymbol{B} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} \end{bmatrix}$$

Diagonal Matrix

A diagonal matrix is a square matrix (i.e. same number of rows and columns) where all of the off-diagonal values is 0 ($a_{ij} = 0$ when $i \neq j$).

Example

$$\begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix}$$

When $a_{11} = a_{22} = a_{33} = 1$, we call the resulting matrix identity matrix, usually denoted by I.

Matrix Inverse

Consider a square matrix $\boldsymbol{A}_{n\times n}$. The inverse of the matrix $\boldsymbol{A}^{-1}$ is the matrix where $\boldsymbol{A} \times \boldsymbol{A}^{-1} = \boldsymbol{A}^{-1} \times \boldsymbol{A} = \boldsymbol{I}$ where $\boldsymbol{I}$ is the identity matrix.

Example

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & 1 \\ 1 & -1 & 2 \\ 1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{A}^{-1} = \begin{bmatrix} 0 & 0 & 1 \\ 2 & -1 & -3 \\ 1 & 0 & -2 \end{bmatrix}$$

Connection to Multiple Linear Regression

Now, consider the following matrices:

$$\boldsymbol{Y}_{n\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \; ; \; \boldsymbol{X}_{n\times k} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,k-1} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,k-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,k-1} \end{bmatrix}$$

$$\boldsymbol{\beta}_{k\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{k-1} \end{bmatrix} \; ; \; \boldsymbol{\epsilon}_{n\times 1} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

For ease of use, I will not include the subscript describing the matrix dimensions in the rest of these lecture notes.

Note that, we previously stated

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{k-1} X_{k-1} + \epsilon.$$

For each observations, we have

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{k-1} X_{ik-1} + \epsilon_i, \text{ for } i = 1, \ldots, n.$$

Using matrix algebra, this is equivalent to saying

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

$$\begin{aligned}
\mathbf{X}\boldsymbol{\beta} + \epsilon &= \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k-1} \\ \vdots & \vdots & & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{k-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \\
&= \begin{bmatrix} \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_{k-1} X_{1k-1} \\ \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_{k-1} X_{2k-1} \\ \vdots \\ \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_{k-1} X_{nk-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \\
&= \begin{bmatrix} \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_{k-1} X_{1k-1} + \epsilon_1 \\ \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_{k-1} X_{2k-1} + \epsilon_2 \\ \vdots \\ \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_{k-1} X_{nk-1} + \epsilon_n \end{bmatrix}
\end{aligned}$$

The sum of squared errors used to find the least squares regression estimates can also be found using matrix notation:

$$SSE = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}).$$

Using this information, the least squares regression estimate for $\boldsymbol{\beta}$ is calculated by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

Inverse of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$

For simple linear regression:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \frac{\sum_{i=1}^n X_i^2}{n\sum_{i=1}^n (X_i - \bar{X})^2} & \frac{-\bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ \frac{-\bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} & \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} \end{bmatrix}$$

Other items from matrix algebra

Vector of Predicted values: $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$

Sum of Squared Errors: SSE $= \boldsymbol{Y}'\boldsymbol{Y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}$

Model Sums of Squares: SSM $= \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} - n\bar{Y}^2$

Variance-covariance matrix of regression parameters is

$$\hat{\text{Var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

- $\hat{\sigma}^2$ is the mean squared error, and the diagonal elements of the above matrix give us the variances (square of the standard errors) for the estimate of the β_j for $j = 0, \ldots, k-1$.

Let's see this in action with the allgreen dataset

```
X <- model.matrix(~ Advertising + SqFt,allgreen)
y <- allgreen$NetSales
betahat <- solve(t(X) %*% X) %*% t(X) %*% y
betahat
```

```
##                   [,1]
## (Intercept) -94.49382
## Advertising  28.32513
## SqFt         45.59163
```

```
# Compare our linear algebra to the R summary output
summary(mod)$coefficients[,"Estimate"]
```

```
## (Intercept) Advertising        SqFt
##   -94.49382    28.32513    45.59163
```

Prediction interval in matrix algebra

For the multiple linear regression,

$$\hat{Y}_{\boldsymbol{x}_0} \pm t_{1-\frac{\alpha}{2},n-k} \cdot s_{Y|\boldsymbol{X}} \cdot \sqrt{1 + \boldsymbol{x}_0^\top \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_0}$$

Calculation by *predict* function directly

```
newdata <- data.frame(SqFt = 3.5,
                      Advertising = 10)
predict(mod,newdata)
```

```
##        1
## 348.3281
```

```
predict(mod,newdata,interval="prediction",level = 0.95)
```

```
##        fit      lwr      upr
## 1 348.3281 240.8506 455.8057
```

Calculation by closed formula.

```
newdata_intercept <- data.frame(inter = 1, Advertising = 10,
                                SqFt = 3.5)
n <- nrow(allgreen)
y_pre <- predict(mod,newdata)
s_yx <- sqrt(sum(mod$residuals^2)/(n-3))
```

```r
sqrt_term <- sqrt(1 +
as.matrix(newdata_intercept)%*% solve(t(X) %*% X) %*% t(as.matrix(newdata_intercept)))
lwr <- y_pre - qt(0.975, n - 3)*s_yx*sqrt_term
upr <- y_pre + qt(0.975, n - 3)*s_yx*sqrt_term
lwr
```

```
##          [,1]
## [1,] 240.8506
```

```r
upr
```

```
##          [,1]
## [1,] 455.8057
```

Regression with Categorical Variables

Indicator Variable

Indicator Variable: A variable that takes only the value 0 or 1 to indicate the absence or presence of some categorical effects.

Baseline: The value of a categorical variable that is omitted to ensure identifiability of the model.

If the categorical variable has m levels, we will only add $m-1$ new indicator variables to our model.

Example

Create a model predicting net personal charges split by whether or not a person is a smoker or not.

```r
mod <- lm(charges ~ smoker,insurance)
summary(mod)
```

```
##
## Call:
## lm(formula = charges ~ smoker, data = insurance)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -19221  -5042   -919   3705  31720
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8434.3      229.0   36.83   <2e-16 ***
## smokeryes    23616.0      506.1   46.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7470 on 1336 degrees of freedom
## Multiple R-squared:  0.6198, Adjusted R-squared:  0.6195
## F-statistic:  2178 on 1 and 1336 DF,  p-value: < 2.2e-16
```

Interpreting model $\hat{charges} = 8434.3 + 23616 * smoker$

$\hat{charges}_{nonsmoker} = 8434.3 + 23616 * 0 = 8434.3$; $\hat{charges}_{smoker} = 8434.3 + 23616 * 1 = 32050.3$

- Interpret the parameter associated with smoker_yes.
  - The interpretation of indicator variables to our linear regression is the expected increase (or decrease) of the response variable for that category **relative** to the baseline category.

- We expect a $\$23,616$ increase in response variable for smokers relative to non-smokers.

- Interpret the intercept of the model.
  - The interpretation of the intercept with categorical variables is the prediction of the response variable for the baseline category.
  - The predicted response variable for a non-smoker is \$8434.30.

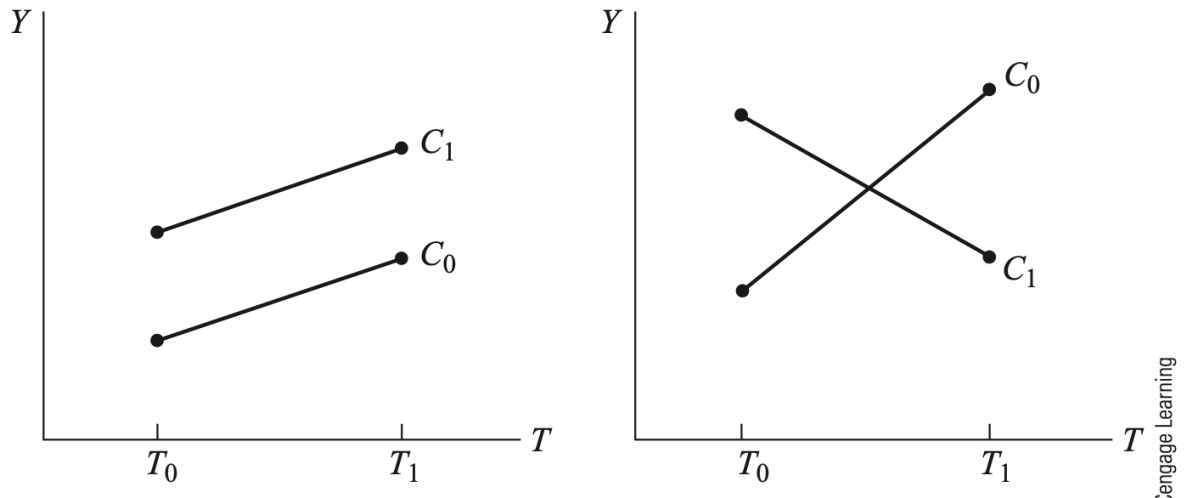
Regression with Interaction Terms

Interaction: An effect in a linear model that quantifies the difference in a linear relationship between an one explanatory variable and the response variable when accounting for the value of another explanatory variable.

Illustration

To illustrate the concept of interaction, let us consider the following example. Suppose that we wish to determine how two independent variables-temperature (T) and catalyst concentration (C)-jointly affect the growth rate (Y) of organisms in a certain biological system. Further, suppose that two particular temperature levels (T_0 and T_1) and two particular levels of catalyst concentration (C_0 and C_1) are to be examined.

We then have two frames to model the relationship between T, C and Y.



Strategy to include interaction terms

- First, one includes only interactions that are reasonable a priori, based on evaluation of the literature and one's expertise. These need not be well-established interactions but rather ones that are worth considering in the model.

  This approach helps to keep regression models parsimonious and to ensure that the model's explanatory variables are readily interpretable.

- Oftentimes, a priori knowledge of possible interactions is unavailable or incomplete, and a second approach is to specify a model with a full set of product terms. For example, if interactions among three variables X_1, X_2, and X_3 are of interest, such a model to consider is

$$\begin{aligned} Y =& \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \beta_6 X_2 X_3 \\ &+ \beta_7 X_1 X_2 X_3 + \epsilon \end{aligned}$$

Here the two-factor products of the form X_iX_j are often referred to as first-order interactions; the three-factor products, like $X_1X_2X_3$, are called second-order interactions; and so on for higher order products. As a general rule, if a higher-order interaction is specified in a model, all lowerorder terms should also be included in the model.

Additional terms such as $X_i^2, X_j^2, X_iX_j^2, X_iX_j^3, X_i^2X_j^2$, and so on can also be included. Nevertheless, there is a limit on the total number of such terms: a model with an intercept (β_0) term cannot contain more than $n-1$ independent variables when n is the total number of observations in the data. Moreover, we may have collinearity issues if we include too many terms.

- The third approach deals with a situation where the association between the dependent variable and a particular factor (or factors) is of primary interest. In this case, we may include only interactions with the primary factor(s). For example, if the purpose of one's study is to describe the relationship between X_1 and Y, we can consider model:

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_1X_2 + \beta_5X_1X_3 + \epsilon.$$

Example

```
# To add an interaction term to our model,
# we use the syntax X1:X2 in addition to X1 and X2
mod_full <- lm(charges ~ bmi + smoker + bmi:smoker,insurance)
# To include both main effects and an interaction,
# use the syntax X1*X2 = X1 + X2 + X1:X2
# The following lm is the same as the one above
mod_full <- lm(charges ~ bmi*smoker,insurance)
# Let's look at a few rows of the model matrix
head(model.matrix(mod_full))
```

```
##   (Intercept)    bmi smokeryes bmi:smokeryes
## 1           1 27.900         1          27.9
## 2           1 33.770         0           0.0
## 3           1 33.000         0           0.0
## 4           1 22.705         0           0.0
## 5           1 28.880         0           0.0
## 6           1 25.740         0           0.0
```

```
coef(mod_full) # Just the estimates of the betas
```

```
##   (Intercept)           bmi     smokeryes bmi:smokeryes
##    5879.42408      83.35056  -19066.00040    1389.75570
```

$\hat{charges} = 5879.42 + 83.35bmi - 19066.00smokeryes + 1389.76bmi \times smokeryes$

$\hat{charges}_{non-smoker} = 5879.42 + 83.35bmi$

$\hat{charges}_{smoker} = (5879.42 - 19066.00) + (83.35 + 1389.76)bmi$

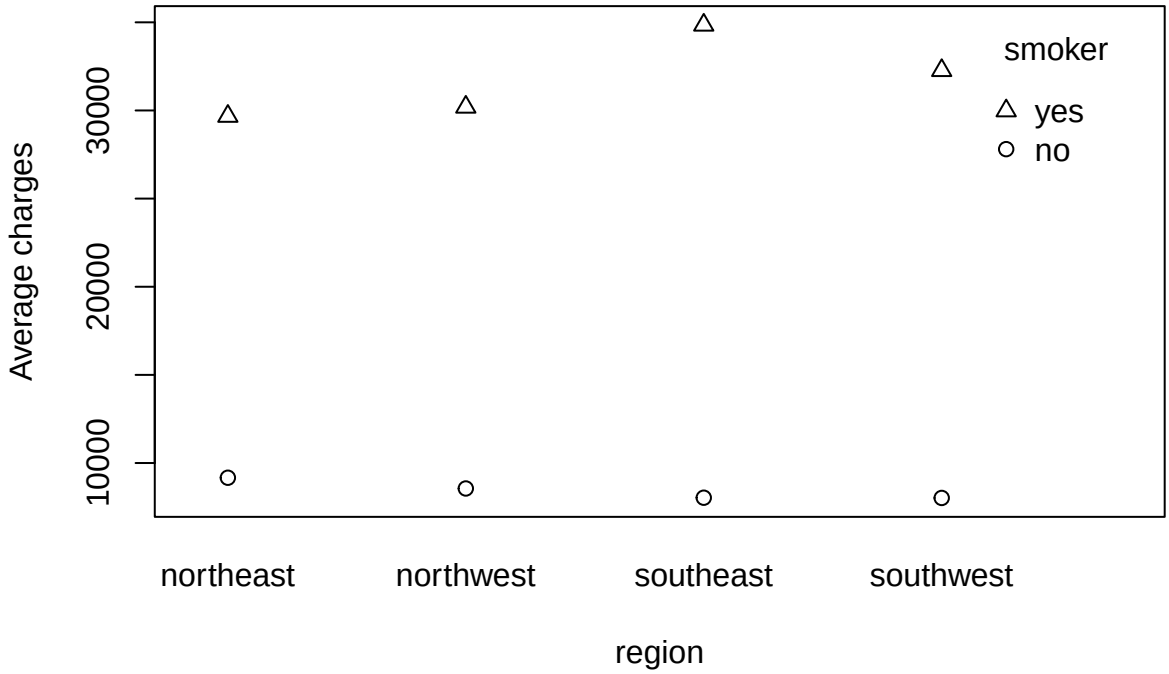
- Interpret the estimate of the parameter associated with the interaction term in the context of the problem.
  - 1389.76: For a 1 kg/m^2 increase in bmi, the predicted health expenditures will increase by \$1389.76 per year more for a smoker than for a non-smoker.
- Interpret the estimate of the parameter associated with bmi in the context of the problem.
  - 83.35: For a 1 kg/m^2 increase in bmi, the predicted health expenditures for non-smokers increases by \$83.35 per year.

- Interpret the estimate of the parameter associated with smokeryes in the context of the problem.
  - -19066.00: The predicted insurance charges for a person with a bmi of $0kg/m^2$ is $\$19,066$ less per year for a person who is a smoker than a non-smoker.

It is counterintuitive. However, we can see it from the graphs.

Visualization to check interaction terms

```
interaction.plot(x.factor = insurance$region,
                 trace.factor = insurance$smoker,
                 response =insurance$charges,
                 type = "p",
                 pch = c(1,2),
                 fun = mean,
                 legend = TRUE,
                 ylab = "Average charges",
                 xlab = "region",
                 trace.label = "smoker")
```



Polynomial Regression

Polynomial Regression: Extension of a linear regression model where we now allow for quadratic, cubic, and higher-order terms in our regression model.

If we have a single explanatory variable X, then we can have

$$Y \sim N\left(\beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_p X^p, \sigma^2\right)$$

or if Y is transformed

$$\tilde{Y} \sim N\left(\beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_p X^p, \sigma^2\right)$$

The method of finding the regression estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ is still minimizing the sum of squared errors:

$$\sum_{i=1}^{n}\left(y_i - (\hat{\beta}_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p)\right)^2$$

or

$$\sum_{i=1}^{n}\left(\tilde{y}_i - (\hat{\beta}_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p)\right)^2$$

Example

Perform a least squares regression analysis for log median home value with a linear and quadratic term for crime rate by answering the following questions:

```
# When you add a polynomial term to a regression model,
# put the term inside of I() so that R knows its a polynomial term
mod_quad <- lm(log(medv) ~ crim + I(crim^2),Boston)
summary(mod_quad)
```

```
##
## Call:
## lm(formula = log(medv) ~ crim + I(crim^2), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.21897 -0.19946 -0.04171  0.17411  1.16083
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.174e+00  1.674e-02 189.629  < 2e-16 ***
## crim        -5.028e-02  3.429e-03 -14.662  < 2e-16 ***
## I(crim^2)    4.847e-04  5.747e-05   8.433 3.57e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3256 on 503 degrees of freedom
## Multiple R-squared:  0.3681, Adjusted R-squared:  0.3656
## F-statistic: 146.5 on 2 and 503 DF,  p-value: < 2.2e-16
```

Remarks

- Transformations should only be conducted **IF AT LEAST ONE OF** the assumption of homoscedasticity OR normality is violated.
- When adding polynomial terms, if you are adding a p-th order polynomial, then add all of the lower polynomials as well (i.e. if you add a quadratic, include a linear OR if you add a cubic, include a linear and quadratic, etc.). **However**, these terms may be removed with variable selection.
- We can look residual plots to see higher order is still needed or we can apply Lack-of-fit tests.

Regression with Transformation

Transformation

Transformation: Inputting the response variable (Y) into a function to obtain a new response ($\tilde{Y}$) that more closely meets the assumptions for a linear regression model.

Common Types of transformations of a response variable:

- Log-Transformation ($\tilde{Y} = \log(Y), Y > 0$): (Note: usually natural logarithm).
  - stabilizes the variance if the variability increases significantly with Y (witnessing a horn shape to the right in the residual plot).
  - normalizes the variable if it is highly right skewed (many points on the upper end of QQ-plot are above the 45-degree line).
  - linearizes the relationship if the relationship between X and Y appears to be exponential (look at this via a scatterplot).
- Square Root-Transformation ($\tilde{Y} = \sqrt{Y}, Y \geq 0$):
  - stabilizes the variance if the variance is proportional to the mean (i.e. $\text{sd}(Y|X_1, \ldots, X_{k-1}) = c\text{E}(Y|X_1, \ldots, X_{k-1})$ where c is a constant). This is particularly appropriate if the dependent variable has the Poisson distribution.
- Squared-Transformation ($\tilde{Y} = Y^2$):
  - stabilizes the variance if the variance decreases with the mean of Y (horn is pointing left in the residual plot).
  - normalizes the dependent variable if the distribution of the residuals for Y is negatively (left) skewed.
  - linearizes the model if the original relationship with some independent variable is curvilinear downward (i.e., if the slope consistently decreases as the independent variable increases).
- Choosing a transformation is subjective, and up to each individuals interpretation of the model assumption checks.

The method of finding the regression estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{k-1}$ is still minimizing the sum of squared errors:

$$\sum_{i=1}^{n} \left(\tilde{y}_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_{k-1} x_{i,k-1}) \right)^2$$

Exercise

Consider the Boston dataset once again. We do logarithm transformation on response variable.

```
# Using the lm function
mod <- lm(log(medv) ~ crim,Boston)
summary(mod)
```

```
##
## Call:
## lm(formula = log(medv) ~ crim, data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.17466 -0.20437 -0.03147  0.17742  1.44893
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.125172   0.016758  186.49   <2e-16 ***
## crim        -0.025089   0.001798  -13.96   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3475 on 504 degrees of freedom
```

```
## Multiple R-squared:  0.2787, Adjusted R-squared:  0.2773
## F-statistic: 194.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
# Use lm for now
```

$\log(\hat{medv}) = 3.125 - 0.02509crim$

Interpret the slope of the least squares regression line in the context of the given problem.

- For a 1% increase in crime rate in a particular suburb, the predicted log-median home value in that suburb decreases by 0.02509 log-thousand dollars.
- We can also write the interpretation in terms of the actual median home value by the following:
- $\hat{medv} = e^{3.125-0.02509crim}$
- For every 1% increase in crime rate, the predicted median home value in Boston suburbs in thousands of dollars will change by a factor of $e^{-0.02509}$
- When doing a log transformation, we are now assuming the relationship between X and Y is exponential, not linear.

Predict the median home value in a Boston suburb with a 0.75 crime rate per capita.

```
library(tidyverse)
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.6.0
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
exp(3.125 - 0.02509*0.75)
```

```
## [1] 22.33562
```

```
# Can still use the predict function, but have to transform to the original units
newdata <- data.frame(crim = 0.75)
predict(mod,newdata) # Incorrect
```

```
##        1
## 3.106355
```

```
predict(mod,newdata) %>% exp() # Correct, need the tidyverse or dplyr to run this
```

```
##        1
## 22.33947
```

Logistic Regression

Logistic Regression: A generalized linear regression technique where the response variable is binary (1 for a "success" or 0 for a "failure")

- Instead of estimating the mean/prediction of Y given $X_1, X_2, \ldots, X_{k-1}$, in logistic regression, we estimate the probability that $Y = 1$ given $X_1, X_2, \ldots, X_{k-1}$.
- However, we cannot set $Pr(Y = 1|X_1, X_2, \ldots, X_{k-1}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_{k-1}$ because the probability is only allowed to be a number between 0 and 1.
- Therefore, in logistic regression, we say that $\text{logit}(Pr(Y = 1)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_{k-1}$.
- In other words, $Pr(Y = 1) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_{k-1}}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_{k-1}}}$.

Note: The logistic regression is one type of generalized linear model.

Example

We take the Titanic dataset as an example to explain the Logistic Regression.

```
Titanic <- read.csv("Titanic.csv"); Titanic <- na.omit(Titanic)
# We are going to remove any rows with missing data
```

- We believe that age, sex, and passenger class are the most important explanatory variables that can be used to predicted whether or not a passenger survived.
- Let's fit a logistic regression model for the Titanic dataset where we want to predict survival rate of passengers by their age, sex, and passenger class.

```
# First, we want passenger class to be a categorical variable
Titanic$Pclass <- as.character(Titanic$Pclass)
# glm runs a generalized linear model, of which logistic regression is an option
# For logistic regression, family = binomial
mod <- glm(Survived ~ Age + Sex + Pclass,
           Titanic,
           family=binomial)
```

```
summary(mod)
```

```
##
## Call:
## glm(formula = Survived ~ Age + Sex + Pclass, family = binomial,
##     data = Titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7303  -0.6780  -0.3953   0.6485   2.4657
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.777013   0.401123   9.416  < 2e-16 ***
## Age         -0.036985   0.007656  -4.831 1.36e-06 ***
## Sexmale     -2.522781   0.207391 -12.164  < 2e-16 ***
## Pclass2     -1.309799   0.278066  -4.710 2.47e-06 ***
## Pclass3     -2.580625   0.281442  -9.169  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 647.28  on 709  degrees of freedom
```

```
## AIC: 657.28
##
## Number of Fisher Scoring iterations: 5
```

- Interpret the parameter estimate associated with Male in the context of the problem.
  - -2.523: The predicted log-odds of surviving the Titanic crash is 2.523 lower for a male than for a female, holding age and passenger class constant.
- Interpret the parameter estimate associated with Age in the context of the problem.
  - -0.03699: For every 1 year increase in age, the predicted log-odds of surviving the Titanic crash is 0.03699 lower, holding sex and passenger class constant.

Hypothesis Testing in Logistic Regression

- Setting up H_0 and H_a.
  - Just like we did in hypothesis testing for linear models, we will write out the reduced and full models using a log-odds regression setup (see previous example).
  - H_0: Values in the full model taken out to get the reduced model.
  - H_a: The exact opposite of H_0.
- We **cannot** use the t-test or the F-test because we do not assume the data are normally distributed.
- We will use (log)-likelihoods to perform hypothesis tests.

Log-likelihood

Log-likelihood: An evaluation of how likely the value of a regression parameter is fitting for the given set of data and the given assumed model.

- Related to the SSE in linear regression analysis.
- The larger the log-likelihood, the better our generalized linear model is at making inference.

In the summary of a glm, the output gives us two deviance values:

- **Null Deviance**: -2*loglikelihood for the glm with no explanatory variables ($df_{Null} = n - 1$). Related to the SST in linear regression.
- **Residual Deviance**: -2*loglikelihood for the glm with the given explanatory variables ($df_{res} = n - k$, k is the number of parameters in the model including intercept). Related to the SSE in linear regression.
- **Null Deviance - Residual Deviance**: A value that is related to the SSM in linear regression (just like SSM = SST - SSE).
- In statistical theory, under H_0, (Null Deviance - Residual Deviance) follows a **Chi-squared** distribution with $(n-1) - (n-k) = k - 1$ degrees of freedom. Difference in the two deviances is related to the SSM in linear regression.
  - Our Test Statistic χ^2 = Null Deviance − Residual Deviance.
  - p-value: $p = Pr(\chi^2_{k-1} > \chi^2)$ (calculating the right tail of the distribution, similar to an F-test).
    - If $p - value < \alpha$, we reject H_0.
    - If $p - value \geq \alpha$, we fail to reject H_0.

Example

Perform a formal hypothesis test to determine if our overall logistic regression model is statistically significant.

- Reduced Model: $\text{logit}(\hat{Survived}) = \beta_0$
- Full Model: $\text{logit}(\hat{Survived}) = \beta_0 + \beta_1 Age + \beta_2 SexMale + \beta_3 Pclass2 + \beta_4 Pclass3$

$$H_0 : \beta_1 = \cdots = \beta_4 = 0; H_a : \text{At least one } \beta \neq 0$$

```
# Test statistic
chi2 <- summary(mod)$null - summary(mod)$deviance
## Null Deviance - Residual Deviance
# p-value, 4 degrees of freedom
p <- pchisq(chi2,df = 4,lower.tail=FALSE)
chi2
```

```
## [1] 317.2328
```

```
p
```

```
## [1] 2.074158e-67
```

$\chi^2 = 317.23, p = 2.0742 \times 10^{-67}$, so we reject H_0 and conclude that a logistic regression model predicting survival probability of the Titanic by sex, age, and passenger class is statistically significant.

Poisson Regression

Poisson regression: a generalized linear regression technique which serves for modeling discrete dependent variables, e.g., $Y = 1, 2, 3, 4, \ldots$.

- For the Poisson distribution, $\Pr(Y = y) = \frac{\mu^y e^{-\mu}}{y!}$, where μ is the mean (rate) of this specific poisson distribution, i.e., $E(Y) = \mu$.
- The rate μ must be positive.
- However, $\beta_0+\beta_1 X_1+\cdots+\beta_{k-1}X_{k-1}$ could be negative. So we can not set $\mu = \beta_0+\beta_1 X_1+\cdots+\beta_{k-1}X_{k-1}$.

Similar with the classical regression for normal distributed Y and logistical regression for 0-1 distributed Y.

We hope to model the mean of Y in the Poisson regression context somehow by predictors $X_1, \ldots, X_{k-1}$.

In poisson regression context, we model the mean of response variable by

$$\ln(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{k-1} X_{k-1}.$$

In other words,

$$\mu = e^{\beta_0+\beta_1 X_1+\cdots+\beta_{k-1}X_{k-1}}.$$

Thus, we can summarize the likelihood function for poisson regression is

$$\begin{aligned} \mathbf{L}(\boldsymbol{\theta} \mid \boldsymbol{X}) &= \prod_{i=1}^{n} Pr\left(Y_i; \boldsymbol{\theta}\right) = \prod_{i=1}^{n} \left\{ \frac{\mu_i^{Y_i} e^{-\mu_i}}{Y_i!} \right\} \\ &= \frac{\left\{ \prod_{i=1}^{n} \mu_i^{Y_i} \right\} e^{-\sum_{i=1}^{n} \mu_i}}{\prod_{i=1}^{n} Y_i!} \end{aligned};$$

where $\mu_i = e^{\beta_0+\beta_1 X_{i,1}+\cdots+\beta_{k-1}X_{i,k-1}}$.

Then, the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ can be obtained by solving k number of equations

$$\frac{\partial}{\partial \beta_j} \ln \mathbf{L}(\boldsymbol{\theta} \mid \boldsymbol{X}) = 0 \quad j = 0, 1, \ldots, k-1;$$

where $\boldsymbol{\theta} = (\beta_0, \ldots, \beta_{k-1})$. See Appendix Maximum Likelihood Estimation for the tutorial for maximum likelihood estimation.

Example

```
library(MASS)
data("ships")

# View structure
str(ships)
```

```
## 'data.frame':    40 obs. of  5 variables:
##  $ type     : Factor w/ 5 levels "A","B","C","D",..: 1 1 1 1 1 1 1 1 2 2 ...
##  $ year     : int  60 60 65 65 70 70 75 75 60 60 ...
##  $ period   : int  60 75 60 75 60 75 60 75 60 75 ...
##  $ service  : int  127 63 1095 1095 1512 3353 0 2244 44882 17176 ...
##  $ incidents: int  0 0 3 4 6 18 0 11 39 29 ...
```

```
head(ships)
```

```
##   type year period service incidents
## 1    A   60     60     127         0
## 2    A   60     75      63         0
## 3    A   65     60    1095         3
## 4    A   65     75    1095         4
## 5    A   70     60    1512         6
## 6    A   70     75    3353        18
```

Data frame giving the number of accidents of ships with other predictors.

```
model_ships <- glm(incidents ~ type + year + period + log(service+1),
data = ships, family = poisson)
# Here, I take log(service+1) is to make all predictors have a similar scale.

summary(model_ships)
```

```
##
## Call:
## glm(formula = incidents ~ type + year + period + log(service +
##     1), family = poisson, data = ships)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2463  -0.8715  -0.3557   0.1153   2.8380
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -8.67747    1.51715  -5.720 1.07e-08 ***
## typeB            -0.33833    0.25953  -1.304  0.19235
```

```
## typeC            -0.73272    0.34110  -2.148  0.03170 *
## typeD            -0.28278    0.29187  -0.969  0.33261
## typeE             0.33785    0.24256   1.393  0.16366
## year              0.03572    0.01379   2.590  0.00960 **
## period            0.02217    0.00811   2.734  0.00625 **
## log(service + 1)  0.89106    0.09813   9.081  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 730.253  on 39  degrees of freedom
## Residual deviance:  58.245  on 32  degrees of freedom
## AIC: 172.11
##
## Number of Fisher Scoring iterations: 5
```

We can also perform a formal hypothesis test to determine if our overall poisson regression model is statistically significant:

```
# Test statistic
chi2 <- summary(model_ships)$null - summary(model_ships)$deviance
## Null Deviance - Residual Deviance
# p-value, 7 degrees of freedom why?
p <- pchisq(chi2,df = 7,lower.tail=FALSE)
chi2
```

```
## [1] 672.0081
```

```
p
```

```
## [1] 7.46111e-141
```

Hypothesis Test on Regression

Fundamental Equation of Regression Analysis

$$\text{SST} = \text{SSM} + \text{SSE}$$

holds for any general regression situation. Figure below illustrates this equation.

R-Squared(R^2)

R-Squared(R^2): the percent of variation in the response variable Y that can be explained through its linear relationship with the explanatory variable X in the data in our sample.

Formally,

$$R^2 = \frac{SSM}{SST} = \frac{SST - SSE}{SST},$$

where $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$, $SSM = \sum_{i=1}^{n}((\hat{\beta}_0 + \hat{\beta}_1 x_i) - \bar{y})^2$, and $SSE = \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$.

- SST: Sum of Squares Total, total variation in the response variable Y
- SSM: Sum of Squares for the Model, total variation in the response variable Y that **IS explained** by its linear relationship with X
- SSE: Sum of Squared Errors, total variation in the response variable Y that **IS NOT explained** by its linear relationship with X

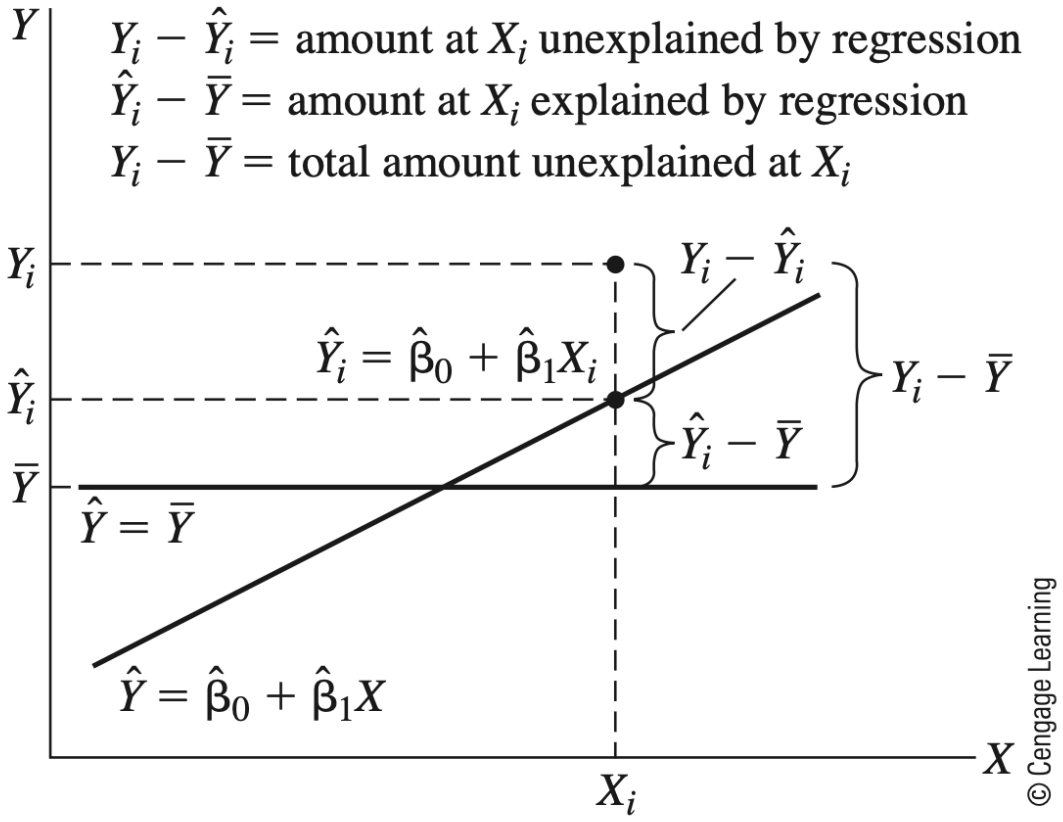


Figure 2: One illustration for simple linear regression

About R-squared

- For simple linear regression, r^2 is the square of the sample correlation, so $0 \leq r^2 \leq 1$.
- The larger the value of r^2, the more variation in Y we can explain through its linear relationship with X, the stronger the linear relationship between them.
  - If $r^2 = 1$, all of the variation in Y can be explained linearly by X (in other words, $SSE = 0$).
  - If $r^2 = 0$, no variation in Y can be explained linearly by X.
- r^2 does NOT measure the magnitude of $\hat{\beta}_1$ (i.e. r^2 can be close to one, but $\hat{\beta}_1$ may still be close to zero, or r^2 can be close to zero, but $\hat{\beta}_1$ may be large.)
- r^2 is NOT a measure of the appropriateness of the linear model.
  - We cannot use r^2 by itself to test for a linear relationship.
  - We can use r^2 along with other pieces of information about the dataset to determine a significantly linear relationship between X and Y.

Adjusted R^2

Adjusted R^2: $R^2_{adj,p} = 1 - \frac{\text{SSE}(p)/(n-p)}{\text{SST}/(n-1)} = 1 - (1 - R^2)\frac{n-1}{n-p}$. Adjusts the regular r^2 for the number of parameters in the model.

ANOVA table for Regression

Mean Squared for Model (MSM): The variance of Y explained by the linear model with X

$$\text{MSM} = \frac{\text{SSM}}{k-1},$$

where $k-1$ is the *model degrees of freedom* for our linear model, and k is the number of parameters (βs) in our linear model.

- You can also think of the *model degrees of freedom* as the number of βs we add to the reduced model to get the full model.

Example

For simple linear regression: $k = 2$ and MSM = SSM.

Mean Squared Error (MSE): The variance of Y not explained by the linear model with X

$$\text{MSE} = \frac{\text{SSE}}{n-k},$$

where $n-k$ is the *error degrees of freedom* for our linear model.

- error degrees of freedom tell us the number of additional parameters we can add to our model without overfitting our model. If we have more parameters than observations (i.e. $k > n$, we cannot properly interpret the estimates of the model parameters)
- n is our sample size and k is the number of parameters in our linear model.

Example

For simple linear regression, $k = 2$, and we said our error degrees of freedom is $n - 2$.

F Statistic

F Statistic: A statistic that measures the ratio of the variances of the response variable Y that is explained/not explained by the model

$$F = \frac{\text{MSM}}{\text{MSE}} = \frac{\text{SSM}/(k-1)}{\text{SSE}/(n-k)} = \frac{\text{Explained Variance}}{\text{Unexplained Variance}}$$

Here, it is equivalent to do F test for overall model significance.

ANOVA Table Format

	df	Sums of Squares	Mean Square	F Value	Pr(>f)
Model	$k-1$	SSM	$\text{MSM} = \frac{\text{SSM}}{k-1}$	$F = \frac{\text{MSM}}{\text{MSE}}$	$P(F_{df1=k-1,df2=n-p} > F)$
Error	$n-k$	SSE	$\text{MSE} = \frac{\text{SSE}}{n-p}$		
Total	$n-1$	SST			

Test for Overall Significance

Example Perform a test for a significant linear regression for All Green's net sales with both advertising dollars and square footage as predictors (explanatory variables, covariates).

As we defined earlier,

$H_0 : \beta_1 = \beta_2 = 0$. H_a : At least one $\beta_j \neq 0$ for $j = 1, 2$.

```
mod <- lm(NetSales ~ Advertising + SqFt,allgreen)
Explained_variance <- sum( (mod$fitted.values - mean(allgreen$NetSales))^2)/2
Unexplained_variance <-(sum(mod$residuals^2)/mod$df.residual)
F_test <- Explained_variance/Unexplained_variance
F_test
```

```
## [1] 174.4105
```

For the overall significance, we can read it from summary directly.

```
mod <- lm(NetSales ~ Advertising + SqFt,allgreen)
# The information you need to perform an
# overall hypothesis (F) test can be found in summary(mod)
summary(mod)$f
```

```
##    value    numdf    dendf
## 174.4105   2.0000  24.0000
```

- F-distribution with 2 and 24 degrees of freedom.

Find the p-value.

```
pf(174.4,2,24,lower.tail = FALSE)
```

```
## [1] 5.067854e-15
```

so we reject H_0 and conclude that a linear model with at least one of advertising spending and square footage is statistically significant in predicted AllGreen net sales for franchises.

Fact about F statistic F-statistic:

$$\frac{\text{Explained Variance}}{\text{Unexplained Variance}} = \frac{\text{SSM}/(k-1)}{\text{SSE}/(n-k)} = \frac{R^2/(k-1)}{\left(1-R^2\right)/(n-k)}$$

Singel Partial F-test

(single) partial F-test: This test assesses whether the addition of any specific independent variable, given others already in the model, significantly contributes to the prediction of Y.

The test, therefore, allows for the deletion of variables that do not help in predicting Y and thus enables one to reduce the set of possible independent variables to an economical set of "important" predictors.

Example Does adding square footage to our linear model that already includes advertising spending significantly improve the prediction of net sales?

- Reduced Model: $\hat{NetSales} = \beta_0 + \beta_1 Advertising$
- Full Model: $\hat{NetSales} = \beta_0 + \beta_1 Advertising + \beta_2 SqFt$

$H_0 : \beta_2 = 0, H_a : \beta_2 \neq 0$

Manually:

```
mod_red <- lm(NetSales ~ Advertising,allgreen)
mod_full <- lm(NetSales ~ Advertising + SqFt,allgreen)
(sum(mod_red$residuals^2) - sum(mod_full$residuals^2) ) /(sum(mod_full$residuals^2) /mod_full$df.residua
```

```
## [1] 37.35142
```

By ANOVA:

```
# We can write two different nested linear models
# and use the ANOVA function to calculate the
# F statistic and p-value
mod_red <- lm(NetSales ~ Advertising,allgreen)
mod_full <- lm(NetSales ~ Advertising + SqFt,allgreen)
anova(mod_red,mod_full)
```

```
## Analysis of Variance Table
##
## Model 1: NetSales ~ Advertising
## Model 2: NetSales ~ Advertising + SqFt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     25 157826
## 2     24  61740  1     96086 37.351 2.591e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The t-test alternative For the test of single partial significance test when number of additional variable is 1, it is equivalent for us to apply the two-sided t-test.

- Reduced Model: $\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_{l-1} X_{l-1}$
- Full Model: $\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_{l-1} X_{l-1} + \beta_l X_l$

We can consider the t-test statistic

$$T = \frac{\hat{\beta}_l}{s_{\hat{\beta}_l}}$$

It can be read from model summary.

Multiple partial F-test

Multiple partial F-test: addresses the simultaneous addition of two or more variables to a model.

- The multiple partial F-test is a straightforward extension of the single partial F test.
- It is often used to test whether a "chunk" (i.e., a group) of variables having some trait in common is important. An example of a chunk is a collection of variables that are all of a certain order, e.g.,$(\text{Advertising})^2$, Advertising $\times$ SqFt, and $(\text{SqFt})^2$ are all of order 2.

In general, for a total of $k-1$ explanatory variables and we assume the first $l-1$ variables are already included in the model ($l < k$):

- Reduced Model: $\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_{l-1} X_{l-1}$
- Full Model: $\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_{l-1} X_{l-1} + \beta_l X_l + \cdots + \beta_{k-1} X_{k-1}$

$H_0 : \beta_l = \beta_{l+1} = \cdots = \beta_{k-1} = 0$, H_a : At least one $\beta_l, \beta_{l+1}, \ldots, \beta_{k-1} \neq 0$

We know that, for any linear model: SST = SSM + SSE

- For the reduced model: $\text{SST} = \text{SSM}_{reduced} + \text{SSE}_{reduced}$
- For the full model: $\text{SST} = \text{SSM}_{full} + \text{SSE}_{full}$

What happens to the sums of squares when we add explanatory variables from the reduced model to get the full model?

- The SSM for the full model will always go up, compared to the reduced model, and the SSE will always go down.
- The difference between $\text{SSE}_{reduced}$ and SSE_{full} is the additional sums of squares explained by the full model that are not explained by the reduced model.

F-statistic

- F-statistic: $\frac{\text{New Explained Variance}}{\text{Still Unexplained Variance}} = \frac{(\text{SSE}_{reduced} - \text{SSE}_{full})/(k-l)}{\text{SSE}_{full}/(n-k)}$
  - $k-l$ are the **additional** degrees of freedom to the linear model (i.e. the number of parameters β_is we are adding to the reduced model to get the full model)
  - $n-k$ are the error degrees of freedom
- $p-value$: Calculated using an F-distribution with $k-l$ and $n-k$ degrees of freedom.

Example Suppose now we want to test whether or not including at least one of square footage and amount of inventory significantly improves the linear model with advertising already included. Perform this hypothesis test.

- Reduced Model: $\hat{NetSales} = \beta_0 + \beta_1 Advertising$
- Full Model: $\hat{NetSales} = \beta_0 + \beta_1 Advertising + \beta_2 SqFt + \beta_3 Inventory$

$H_0 : \beta_2 = \beta_3 = 0, H_a$: At least one $\beta \neq 0$

$F = 25.04$, $p = 1.683 \times 10^{-6}$, so we reject H_0 and conclude that adding at least one of square footage and inventory to our linear model that already includes advertising spending is statistically significant in predicting net sales at All Green franchises.

Test for a linear combination of estimators

Consider the following situations:

1. Suppose that we are interested in if advertising is twice as effective as square footage on the net sales value of store. This conjecture can be addressed by considering the null hypothesis $H_0 : \beta_{\text{advertising}} = 2\beta_{\text{SqFt}}$.

2. We may be interested in estimating how much net sales increases, on average, as a result of 1 unit increase for square footage and advertising. In other words, we would want to estimate the quantity ($\beta_{\text{advertising}} + \beta_{\text{SqFt}}$) and construct a confidence interval for this unknown parameter.
3. We may wish to assess whether or not the joint effect of 1 unit change of square footage and advertising is equal to a hypothesized number q on net sales, where q is a specified mean change. This assessment could be made via a hypothesis test of $H_0 : \beta_{\text{advertising}} + \beta_{\text{SqFt}} = q$.

All of the above situations can be addressed using the same general approach. The quantity of interest is expressed as a linear function L of the β coefficients, with the estimate of L denoted as $\hat{L}$. In the case of hypothesis tests (situations 1 and 3 above), we typically rearrange L as an expression equal to 0.

For example, for situation 1, $L = \beta_{\text{advertising}} - 2\beta_{\text{SqFt}} = 0$ and $\hat{L} = \hat{\beta}_{\text{advertising}} - 2\hat{\beta}_{\text{SqFt}}$.

Hypothesis tests and confidence intervals all utilize the estimated value $\hat{L}$ and its estimated standard error $s_{\hat{L}}$:

- Hypothesis tests use the test statistic $\frac{\hat{L}}{s_{\hat{L}}}$, which follows the t_{n-k} distribution under the null hypothesis that $L = 0$.
- Confidence intervals uses the formula $\hat{L} \pm t_{1-\frac{\alpha}{2},n-k}\left(s_{\hat{L}}\right)$.

A key difference from earlier is that the standard error of $\hat{L}, s_{\hat{L}}$ or, equivalently, its variance, $s^2_{\hat{L}}$ is not readily obtained from standard output.

The formula for $s^2_{\hat{L}} = \widehat{\text{Var}}(\hat{L})$ is

$$s^2_{\hat{L}} = \sum_{i=0}^{k-1} c_i^2 S^2_{\hat{\beta}_i} + 2\sum_{i=0}^{k-2}\sum_{j=i+1}^{k-1} c_i c_j \widehat{\text{cov}}\left(\hat{\beta}_i, \hat{\beta}_j\right)$$

Variance and covariance matrix The variance and covariance matrix corresponding with $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta_{k-1}})$ is $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$; where we can estimate σ^2 by SSE$/(n-k)$, i.e., $s^2_{Y|\boldsymbol{X}}$.

Example Check $H_0 : \beta_{\text{advertising}} = 2\beta_{\text{SqFt}}$

```
n = nrow(allgreen)
X <- model.matrix(~ Advertising + SqFt,allgreen)
mod_full <- lm(NetSales ~ Advertising + SqFt,allgreen)
L_hat = mod_full$coefficients["Advertising"] - 2*mod_full$coefficients["SqFt"]
sse = sum(residuals(mod_full)^2)/(n - 3)
XXinverse = solve(t(X) %*% X)
s_lhat = sqrt( c(0,1,2)^2 %*% c(sse*diag(XXinverse)) - sse*2*1*2*XXinverse[2,3])
L_hat/s_lhat
```

```
##           [,1]
## [1,] -3.475113
```

We can also calculate it by using *linearHypothesis* function from the *car* package.

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
mod_full <- lm(NetSales ~ Advertising + SqFt,allgreen)
linearHypothesis(mod_full, "Advertising = 2*SqFt")
```

```
##
## Linear hypothesis test:
## Advertising - 2 SqFt = 0
##
## Model 1: restricted model
## Model 2: NetSales ~ Advertising + SqFt
##
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1     25 92806
## 2     24 61740  1     31066 12.076 0.001959 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Influential Observations

Influential Observations: Observations that *may* influence the estimation of the least squares regression line.

Leverage(h_i): is a measure of the extremeness of an observation with respect to the independent variables.

We can compute Leverage by `hatvalues` function

```
mod <- lm(SBP ~ Age,bloodpressure)
hatvalues(mod)
```

```
##          1          2          3          4          5          6          7
## 0.03887884 0.03384700 0.03333595 0.03384700 0.09151663 0.03344406 0.10382106
##          8          9         10         11         12         13         14
## 0.03478064 0.10382106 0.05074102 0.08580667 0.05074102 0.06167938 0.05160587
##         15         16         17         18         19         20         21
## 0.03478064 0.03454478 0.03333595 0.15001179 0.12645452 0.13401211 0.04563055
##         22         23         24         25         26         27         28
## 0.03682483 0.03887884 0.11919176 0.03352268 0.04245617 0.08039154 0.07170375
##         29         30
## 0.09308908 0.11730482
```

Jackknife residuals

Generally, large outliers will result in large $\hat{\epsilon}_i$ values. Therefore, the sizes of the $\hat{\epsilon}_i$ (or the standardized or studentized residuals) can be studied in order to detect outliers.

Sometimes, the outlier exerts influence on the fitted regression surface, pulling the fitted regression surface away from the main body of the data and toward itself, thereby reducing the size of the associated residual.

The jackknife residual avoids this problem and, therefore, unmasks outliers.

For the i th observation, the jackknife residual, $r_{(-i)}$, is calculated as

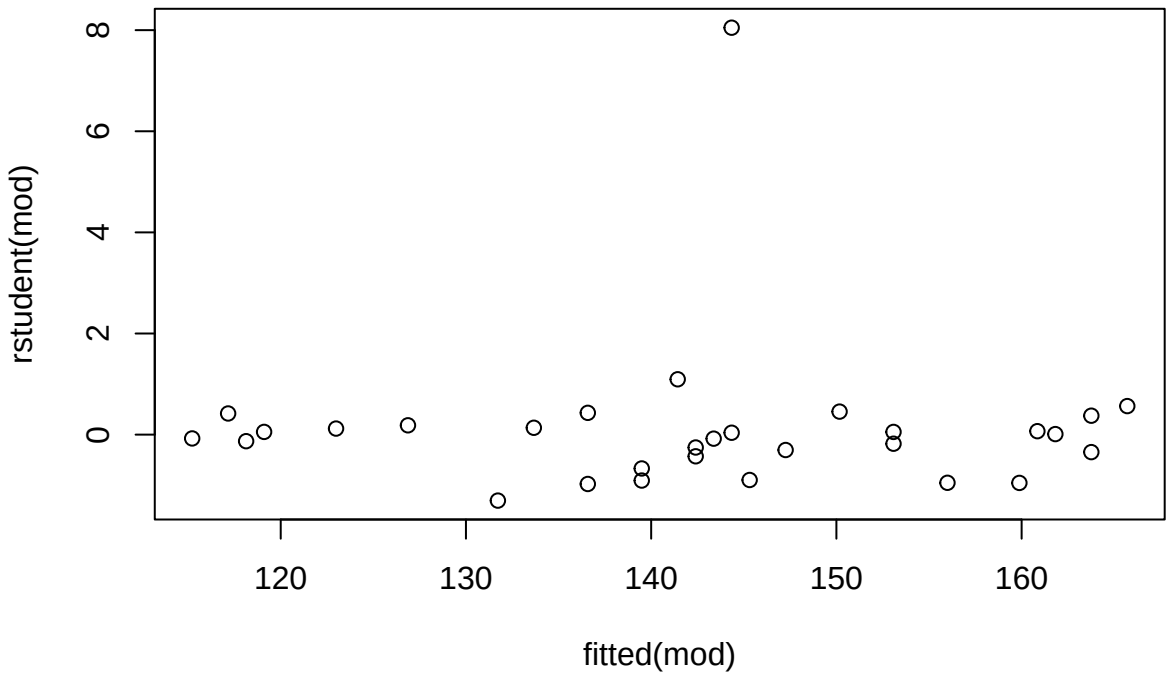
$$r_{(-i)} = \frac{\hat{\epsilon}_i}{s_{(-i)}\sqrt{1-h_i}}$$

The quantity $s^2_{(-i)}$ is the mean square error computed with the i th observation deleted. If the outlier pulls the surface toward itself and away from the main body of the data, s^2, will be larger than $s^2_{(-i)}$ Therefore, for the i th observation, $r_{(-i)}$ will be larger than the studentized residual, r_i, in the case of an outlier.

$$\hat{r}_{(-i)} \sim t_{n-k-1}.$$

Example Plot the studentized residuals for the blood pressure dataset against the predictions from the least squares regression line $\hat{y}_i$ for $i = 1, \ldots, n$

```
# fitted(mod) gives the y-hat values
# rstudent(mod) calculates the studentized residuals
plot(fitted(mod),rstudent(mod))
```



- We have a single point that is way above the others at a studentized residual of 8, the rest are between -2 and 2.

How do we determine statistically which points are influential observations?

We can say that

$$\hat{r}_{(-i)} \sim t_{df=n-k-1}$$

The $n-k-1$ degrees of freedom incorporates the fact that the Jackknife residuals use the leverage values from leave one out regression.

So, the influential points according to studentized residuals are points where $|\hat{r}_i| > t_{1-\frac{\alpha}{2},n-k-1}$ where $t_{1-\frac{\alpha}{2},n-k-1}$ is the $(1-\frac{\alpha}{2})$-th quantile from a t-distribution with $n-k-1$ degrees of freedom.

Collinearity

Collinearity: A phenomenon in regression modelling where two of the independent/explanatory variables in a model are correlated.

Multicollinearity: A phenomenon in regression modelling where more than two of the independent/explanatory variables in a model are correlated.

- If two explanatory variables have a correlation of -1 or 1, they are considered to be *perfectly collinear*.
- If three or more explanatory variables have pairwise correlation of -1 or 1, they are considered to be *perfectly multicollinear*.
- If a regression model has multicollinearity, then we may not be able to easily determine which explanatory variables are most significant in a model.

Variance inflation factor

Variance Inflation Factor (VIF): The VIF quantifies exactly how much the variance of a regression coefficient is inflated due to multicollinearity.

You can say it is comparing the mean squared error of

Reduced Model: $\hat{Y} = \beta_0 + \beta_j X_j$ vs.

Full Model: $\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_j X_j + \cdots + \beta_{p-1} X_{p-1}$

$$VIF(j) = \frac{1}{1 - R_j^2},$$

where R_j^2 is the R-squared of a linear model for X_j with all of the other explanatory variables as predictors, i.e., regressing X_j on $\boldsymbol{X}_{-j}$; $\boldsymbol{X}_{-j}$ represents new $\boldsymbol{X}$ without the j-th column.

Note: R_j^2 is a number between 0 and 1, $VIF(j) > 1$

- If $1 < VIF(j) \leq 5$, then there is no signficant evidence of multicollinearity for variable X_j.
- If $5 < VIF(j) \leq 10$, then there is moderate evidence of multicollinearity for variable X_j, and we should start to consider excluding X_j from our linear model.
- If $VIF(j) > 10$, then there is significant evidence of multicollinearity for variable X_j, and we should definitely consider excluding X_j from our linear model.

```
# Need library(regclass) to run the VIF function
library(regclass)
```

```
## Loading required package: bestglm
## Loading required package: leaps
## Loading required package: VGAM
## Warning: package 'VGAM' was built under R version 4.2.3
## Loading required package: stats4
## Loading required package: splines
##
## Attaching package: 'VGAM'
## The following object is masked from 'package:car':
##
##     logit
## Loading required package: rpart
## Loading required package: randomForest
## randomForest 4.7-1.1
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin

## Important regclass change from 1.3:
## All functions that had a . in the name now have an _
## all.correlations -> all_correlations, cor.demo -> cor_demo, etc.
```

```
mod <- lm(NetSales ~ .,allgreen)
VIF(mod)
```

```
##        SqFt   Inventory Advertising  SizeofDist  NoofStores
##    4.240914   10.122480    7.624391    6.912318    5.818768
```

```
# Let's look at the VIF for the linear model without inventory
mod_less <- lm(NetSales ~ . - Inventory,allgreen)
VIF(mod_less)
```

```
##        SqFt Advertising  SizeofDist  NoofStores
##    3.579850    3.795323    5.861520    5.468943
```

```
# The potential issue of multicollinearity has decreased substantially!
```

- When you find explanatory variable(s) with a high VIF, only eliminate the variable with the *highest* VIF. The other explanatory variables may then exhibit a lower VIF when one is eliminated.

Variable Selection

Maximum Model

Maximum Model: The model with all of the explanatory variables, potentially including polynomial and interaction terms, that we would be okay with including in our regression model.

AIC

Akaike Information Criterion (AIC): $AIC = n\log(\text{SSE}(p)/n) + 2p$; why? Recall that the original definition of AIC is

$$\text{AIC} = 2k - 2\ln(\hat{L}),$$

where $\hat{L}$ is the maximized value of the likelihood function for the model.

Backward Selection

1. Start with a model that contains all possible predictors, i.e., $Y \sim N(\beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1}X_{p-1}, \sigma^2)$.
2. For each of the possible variable to be removed from the model

- Calculate AIC.
- Remove the variable to the model that has the lowest AIC.

3. Repeat (2).

4. If the *AIC* does not decrease any further, do not remove any additional variables from the model, and return the final model.

```
mod_max <- lm(medv ~ .*. + I(crim^2) + I(lstat^2),Boston)
summary(mod_max)
```

```
##
## Call:
## lm(formula = medv ~ . * . + I(crim^2) + I(lstat^2), data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6973 -1.4638 -0.1232  1.3974 18.0726
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.580e+02  6.778e+01  -2.331 0.020238 *
## crim          -1.595e+01  6.552e+00  -2.434 0.015362 *
## zn            -9.695e-02  4.568e-01  -0.212 0.832014
## indus         -2.606e+00  1.693e+00  -1.540 0.124429
## chas           4.228e+01  1.952e+01   2.166 0.030886 *
## nox            3.256e+01  7.512e+01   0.433 0.664883
## rm             2.539e+01  5.680e+00   4.469 1.02e-05 ***
## age            1.256e+00  2.719e-01   4.617 5.21e-06 ***
## dis           -1.191e+00  4.594e+00  -0.259 0.795528
## rad            1.615e+00  2.460e+00   0.657 0.511868
## tax            1.865e-02  1.438e-01   0.130 0.896881
## ptratio        2.988e+00  2.844e+00   1.051 0.293958
## black          9.128e-02  7.464e-02   1.223 0.222048
## lstat          1.154e+00  8.813e-01   1.309 0.191139
## I(crim^2)      1.653e-03  1.282e-03   1.290 0.197901
## I(lstat^2)     9.597e-03  4.955e-03   1.937 0.053450 .
## crim:zn        3.926e-01  1.803e-01   2.178 0.029980 *
## crim:indus    -1.008e-01  4.490e-01  -0.225 0.822479
## crim:chas      2.511e+00  5.723e-01   4.389 1.45e-05 ***
## crim:nox      -9.525e-01  9.355e-01  -1.018 0.309156
## crim:rm        1.469e-01  6.209e-02   2.366 0.018432 *
## crim:age      -3.890e-03  3.876e-03  -1.004 0.316108
## crim:dis      -1.249e-01  1.102e-01  -1.133 0.257777
## crim:rad      -6.996e-01  5.809e-01  -1.204 0.229132
## crim:tax       3.868e-02  4.287e-02   0.902 0.367526
## crim:ptratio   4.169e-01  3.351e-01   1.244 0.214130
## crim:black    -3.390e-04  1.873e-04  -1.809 0.071111 .
## crim:lstat     2.461e-02  6.918e-03   3.558 0.000417 ***
## zn:indus      -6.176e-04  4.646e-03  -0.133 0.894299
## zn:chas       -5.522e-02  6.430e-02  -0.859 0.390991
## zn:nox        -1.995e-02  4.706e-01  -0.042 0.966213
## zn:rm          4.700e-03  2.605e-02   0.180 0.856925
## zn:age        -4.973e-05  8.488e-04  -0.059 0.953310
## zn:dis         1.067e-02  7.526e-03   1.418 0.156850
## zn:rad        -2.057e-03  7.017e-03  -0.293 0.769600
## zn:tax         3.742e-04  1.781e-04   2.101 0.036217 *
## zn:ptratio    -3.931e-03  6.997e-03  -0.562 0.574599
## zn:black       6.068e-05  7.602e-04   0.080 0.936425
## zn:lstat      -9.350e-03  4.693e-03  -1.992 0.047013 *
```

```
## indus:chas    -2.932e-01  3.787e-01  -0.774 0.439239
## indus:nox      3.159e+00  1.444e+00   2.187 0.029300 *
## indus:rm       3.048e-01  1.329e-01   2.293 0.022345 *
## indus:age      9.970e-05  3.660e-03   0.027 0.978282
## indus:dis     -3.758e-02  6.316e-02  -0.595 0.552213
## indus:rad     -2.713e-02  5.016e-02  -0.541 0.588831
## indus:tax      3.143e-04  6.032e-04   0.521 0.602617
## indus:ptratio -5.562e-02  3.785e-02  -1.469 0.142485
## indus:black    8.016e-04  2.033e-03   0.394 0.693601
## indus:lstat   -4.682e-05  1.537e-02  -0.003 0.997571
## chas:nox      -3.176e+01  1.240e+01  -2.561 0.010781 *
## chas:rm       -5.225e+00  1.157e+00  -4.517 8.20e-06 ***
## chas:age       2.639e-02  5.827e-02   0.453 0.650864
## chas:dis       1.110e+00  1.335e+00   0.831 0.406332
## chas:rad      -7.436e-01  5.695e-01  -1.306 0.192354
## chas:tax       4.127e-02  3.642e-02   1.133 0.257693
## chas:ptratio  -5.642e-01  6.908e-01  -0.817 0.414535
## chas:black     2.454e-02  1.562e-02   1.571 0.116984
## chas:lstat    -2.863e-01  1.843e-01  -1.554 0.121008
## nox:rm         4.535e+00  5.490e+00   0.826 0.409214
## nox:age       -7.212e-01  2.332e-01  -3.092 0.002124 **
## nox:dis        5.285e+00  3.721e+00   1.420 0.156308
## nox:rad       -1.908e-01  1.894e+00  -0.101 0.919792
## nox:tax       -2.901e-02  1.310e-01  -0.221 0.824906
## nox:ptratio   -3.732e+00  3.089e+00  -1.208 0.227772
## nox:black     -1.618e-02  3.614e-02  -0.448 0.654649
## nox:lstat      9.074e-01  6.598e-01   1.375 0.169802
## rm:age        -6.275e-02  2.196e-02  -2.858 0.004486 **
## rm:dis         2.315e-01  3.308e-01   0.700 0.484460
## rm:rad        -5.765e-02  1.527e-01  -0.378 0.705945
## rm:tax        -2.112e-02  9.894e-03  -2.135 0.033357 *
## rm:ptratio    -5.249e-01  2.177e-01  -2.412 0.016323 *
## rm:black      -3.188e-03  3.407e-03  -0.936 0.349963
## rm:lstat      -2.238e-01  5.564e-02  -4.022 6.87e-05 ***
## age:dis       -1.617e-02  8.857e-03  -1.825 0.068676 .
## age:rad        1.449e-02  4.201e-03   3.448 0.000623 ***
## age:tax       -3.650e-04  2.183e-04  -1.672 0.095265 .
## age:ptratio   -5.918e-03  6.828e-03  -0.867 0.386570
## age:black     -7.431e-04  2.139e-04  -3.475 0.000566 ***
## age:lstat     -6.509e-03  1.954e-03  -3.331 0.000945 ***
## dis:rad       -7.290e-02  7.147e-02  -1.020 0.308310
## dis:tax       -3.832e-03  2.494e-03  -1.536 0.125204
## dis:ptratio   -4.194e-02  9.954e-02  -0.421 0.673735
## dis:black     -4.430e-03  5.567e-03  -0.796 0.426610
## dis:lstat      1.167e-01  4.933e-02   2.366 0.018437 *
## rad:tax        5.668e-05  1.442e-03   0.039 0.968666
## rad:ptratio   -2.698e-02  8.405e-02  -0.321 0.748347
## rad:black     -8.219e-04  2.516e-03  -0.327 0.744052
## rad:lstat     -2.374e-02  1.812e-02  -1.310 0.190777
## tax:ptratio    7.823e-03  2.496e-03   3.134 0.001849 **
## tax:black      2.694e-05  1.997e-04   0.135 0.892742
## tax:lstat     -1.226e-03  1.206e-03  -1.016 0.310237
## ptratio:black  8.018e-04  3.359e-03   0.239 0.811449
## ptratio:lstat -1.948e-03  2.994e-02  -0.065 0.948159
```

```
## black:lstat   -3.907e-04  4.265e-04  -0.916 0.360207
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.842 on 412 degrees of freedom
## Multiple R-squared:  0.9221, Adjusted R-squared:  0.9045
## F-statistic: 52.44 on 93 and 412 DF,  p-value: < 2.2e-16
```

```
# The step() function in R conducts backwards, forwards,
# or stepwise selection procedures in # determining the "best" regression model

mod_best_backward <- step(mod_max,direction="backward",trace = 0)
# If you do not want to see all of the selection proceedings,
# then input trace = 0 into the function
```

Forward Selection

1. Start with a model that contains no predictors, i.e., $Y \sim N(\beta_0, \sigma^2)$.
2. For each of the possible variables to be added to the model.

- Calculate AIC when adding each variable individually from the model.
- Add the variable from the model that has the smallest AIC.

3. Repeat (2).
4. If the criterion does not change any further, do not add any additional variables to the model, and return the final model.

We can do such selection by setting `step(direction="forward"`

Stepwise Selection

Stepwise selection starts the same as backward selection, except that it incorporates "re-examination". In other words, variables that were previously removed from the model using backward selection are now allowed to be reentered into the model if adding the variable back in minimizes or maximizes the specific criterion.

We can do such selection by setting `step(direction="both").`

Appendix

Maximum Likelihood Estimation

Maximum likelihood method: refers to a very general algorithm for obtaining estimators of population parameters.

To understand this method, we suppose there are n independent data pairs $((Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n))$; where $\boldsymbol{X}_i$ represents $(X_{i,1}, \ldots, X_{i,k-1})$ for $i = 1, \ldots, n$.

Also, we assume that the density or probability mass function for $\boldsymbol{Y}$ given $\boldsymbol{X}$ is

$$\mathbf{f}(\boldsymbol{Y} \mid \boldsymbol{\theta}_0);$$

when $\boldsymbol{\theta}_0$ is the true related parameters and $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ represents all predictors and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ represents all response values.

Likelihood function Likelihood function: is the density function regarded as a function of θ

$$\mathbf{L}(\boldsymbol{\theta} \mid \boldsymbol{Y}) = \mathbf{f}(\boldsymbol{Y} \mid \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta$$

Then, we can define the Maximum Likelihood Estimator (MLE)

The maximum likelihood estimator (MLE):

$$\hat{\boldsymbol{\theta}}(\boldsymbol{Y}) = \arg\max_{\boldsymbol{\theta}} \mathbf{L}(\boldsymbol{\theta} \mid \boldsymbol{Y})$$

Note that if $\hat{\boldsymbol{\theta}}(\boldsymbol{Y})$ is a maximum likelihood estimator for $\boldsymbol{\theta}$, then $g(\hat{\boldsymbol{\theta}}(\boldsymbol{Y}))$ is a maximum likelihood estimator for $g(\boldsymbol{\theta})$. Typically, maximizing the log-likelihood $\ln \mathbf{L}(\boldsymbol{\theta} \mid \boldsymbol{Y})$ will be easier.

- For the classical linear regression, the estimators obtained by least square and maximum likelihood methods are identical.
- For logistic regression, the least squares may result in some predicted probabilities that are negative or greater than 1.0.
- For poisson regression, the least squares may result in negative value of response variable.
- For logistic and poisson regression, the maximum likelihood estimates are more accurate than least squares estimates.