

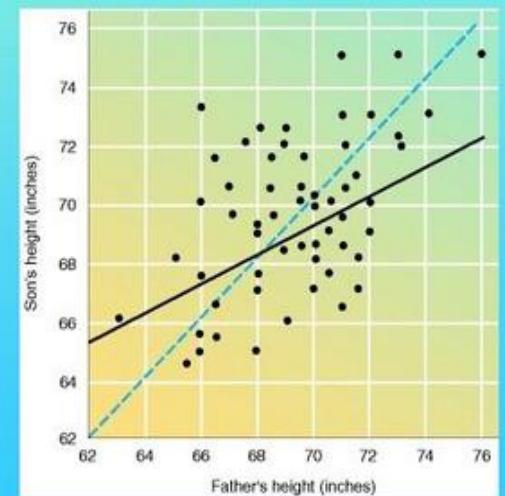
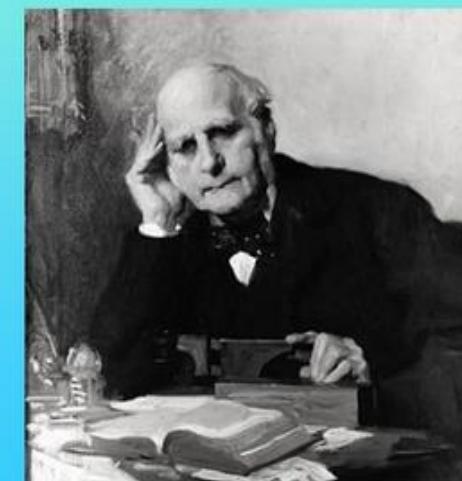
Simple Linear Regression (SLR)

Simple Linear Regression

The purpose of regression?

- Describe functional relationships between variables
- Control
- Prediction of outcomes

Developed by Sir Francis Galton (1822-1911) in his article “Regression towards mediocrity in hereditary structure”



Simple Linear Regression

The basic concepts of regression

- Describe statistical relationships between variables
- The statistical relation has **two essential ingredients**
 - A tendency of the response variable Y to vary with the predictor variable X
 - There is a probability distribution of Y for each level of X .
 - A scattering of points around the curve of statistical relationship.
 - The means of these probability distributions vary in some systematic fashion with X

Example (diamonds.csv)

Variables:

Response Variable: Price in Singapore dollars (Y)

Explanatory Variable: Weight of diamond in carats (X)

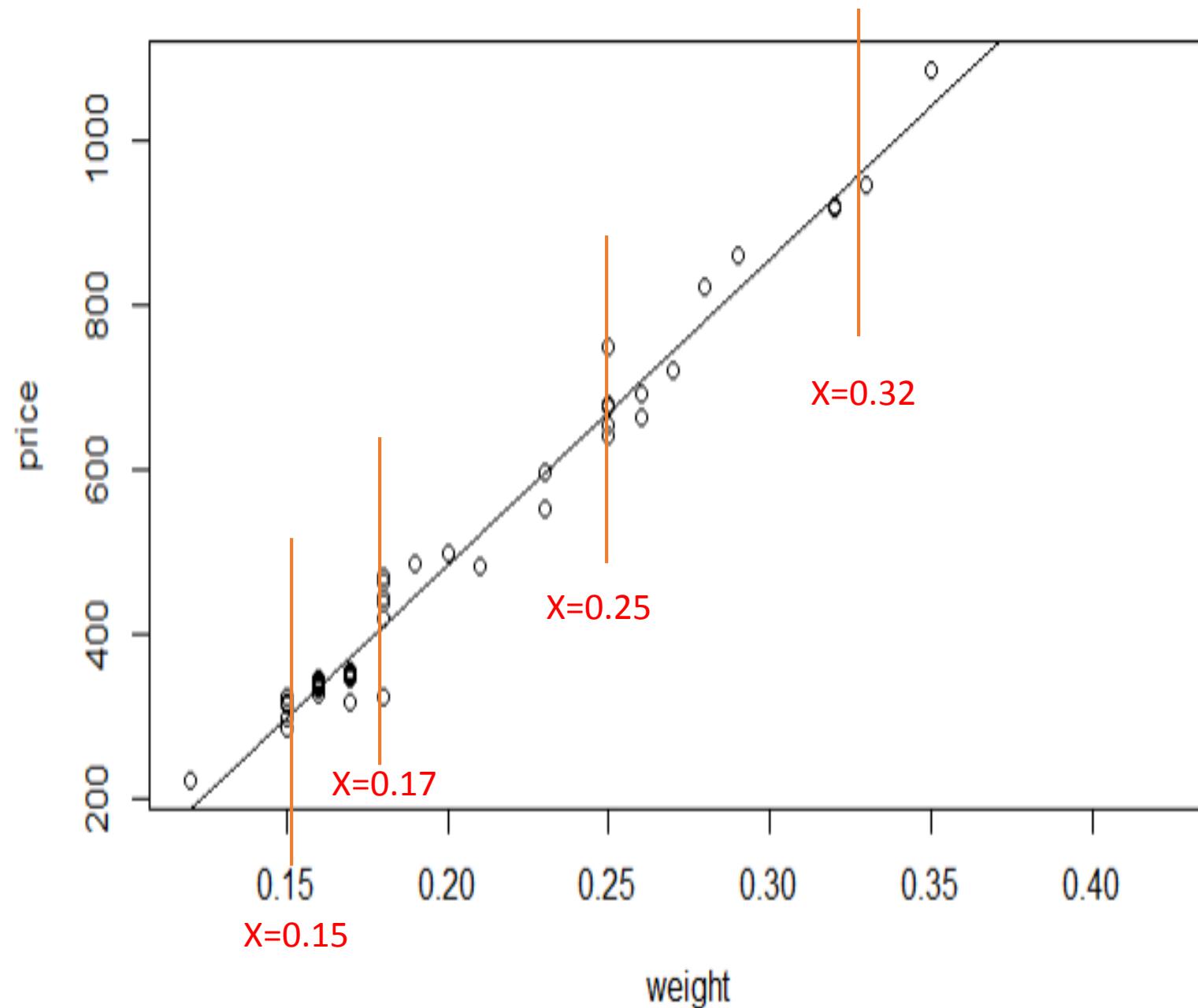
Goal:

Predict the price of a sale for a 0.43 carat diamond ring

What are the two ingredients in understanding statistical relationship between price and weight ?

Scatter plot

Mean price = intercept + slope (weight)



- The means of the price distributions increase linearly with the weight
- For any given weight, the distribution of price varies, and we can see later that the distribution is Normal (the bell-shape distribution).

Notation for Simple Linear Regression (SLR)

- Observe a pair of variables (explanatory and response) on each of $i = 1, 2, \dots, n$ samples
- Each pair often called a **case** or a **data point** (X_i, Y_i)
- Y_i is the value of the response for the i -th case
- X_i is the value of the explanatory variable for the i -th case

Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad \text{for } i = 1, 2, \dots, n$$

Simple Linear Regression Model Parameters

- β_0 is the intercept.
- β_1 is the slope.
- ϵ_i are independent, normally distributed random errors with mean 0 and variance σ^2 ,

$$\epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

Features of Simple Linear Regression Model

- Individual observations: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- Since ε_i are random, Y_i are also random and

$$\begin{aligned}E(Y_i) &= \beta_0 + \beta_1 X_i + E(\varepsilon_i) = \beta_0 + \beta_1 X_i \\ \text{Var}(Y_i) &= 0 + 0 + \text{Var}(\varepsilon_i) = \sigma^2.\end{aligned}$$

Since ε_i is Normally distributed, $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$

Fitted Regression Equation and Residuals

The parameters β_0 , β_1 , and σ^2 are unknown and must be estimated from the data.

The “hat” symbol
is “point estimation”



$$\hat{Y} = b_0 + b_1 X$$

- b_0 estimates β_0 (intercept) $\hat{\beta}_0 = b_0 \text{ or } \hat{\beta}_1 = b_1$
- b_1 estimates β_1 (slope)
- $\hat{Y}_i = b_0 + b_1 X_i$ gives the estimated mean of Y when the predictor is X_i .
- The *residual* for the i -th case is $e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i)$
- $s^2 = \text{Var}(e_i)$ estimates the error variance σ^2 $\widehat{\sigma^2} = s^2$

The residual e_i (in one sample) is NOT the same as the error ε_i (in population) !

$$\hat{\varepsilon} = e$$

Estimating the parameters with Least Squares (LS) Solution

- We want to find the “best” estimates, b_0 and b_1 .
- Minimize the sum of the squared residuals, $\sum_{i=1}^n e_i^2$, i.e., find

$$\arg \min_{(b_0, b_1)} = \sum_{i=1}^n [Y_i - (b_0 + b_1 X_i)]^2$$

- How? Calculus!
 1. Take derivatives with respect to b_0 and with respect to b_1 .
 2. Set equations equal to zero and solve for both b_0 and b_1 .

Estimating the parameters with Least Squares (LS) Solution

- The best estimates of β_1 and β_0 given the data (X, Y) are:

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{SS_{XY}}{SS_X}$$

SS is “sum of squares”

- $b_0 = \bar{Y} - b_1 \bar{X}$
- This estimate is the “best” because it
 - is *unbiased* (its expected value is equal to the true value)
 - has *minimum variance*

Estimate the parameters with Maximum Likelihood Estimation (MLE)

Our model says that $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$.

Given X_i , the probability of data point i is,

$$f_i = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2}$$

β_0 and β_1 are unknown, but the *likelihood* of the proposed values (β_0^*, β_1^*) given the data is,

$$L(\beta_0^*, \beta_1^* | X, Y) = f_1 \times f_2 \times \dots \times f_n = \prod_{i=1}^n f_i$$

L is maximized when $\beta_0^* = b_0$ and $\beta_1^* = b_1$. Thus, the LS estimates, b_0 and b_1 , are also the estimated parameter values that are most (probabilistically) consistent with the data!

Estimation of stochastic variance, σ^2

We estimate σ^2 as the sum of the squared residuals, SSE , divided by the degrees of freedom:

$$s^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n - 2} = \frac{\sum e_i^2}{n - 2} = \frac{SSE}{n - 2} = \frac{SSE}{DFE} = MSE$$

SSE stands for “sum of squares error”

DFE stands for “degree of freedom of error”

MSE stands for “mean squared error”

$$E\{MSE\} = \sigma^2 \rightarrow \text{MSE is an unbiased estimator of } \sigma^2$$

$$s = \sqrt{MSE} \rightarrow \text{This is the residual standard error, which estimates the residual standard deviation } (\sigma)$$

Estimation of stochastic variance, σ^2

We estimate σ^2 as the sum of the squared residuals, SSE , divided by the degrees of freedom:

$$s^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n - 2} = \frac{\sum e_i^2}{n - 2} = \frac{SSE}{n - 2} = \frac{SSE}{DFE} = MSE$$

SSE stands for “sum of squares error”

DFE stands for “degree of freedom of error”

MSE stands for “mean squared error”

$$E\{MSE\} = \sigma^2 \rightarrow \text{MSE is an unbiased estimator of } \sigma^2$$

$$s = \sqrt{MSE}$$

MSE measures variability around the fitted regression line,
A _____ (A. smaller/B larger) MSE is preferred and often used as
a criterion for model selection

A comment on the notation

We will also estimate variances for other quantities.

These will also be denoted S^2 , but will have a subscript to identify them, e.g. $S^2_{\{b_1\}}$.

Without any subscript, S^2 refers to the estimated variance of the residuals.

And S refers to the standard error of the residuals.

Identifying statistics and estimates in the R output

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-259.63	17.32	-14.99	<2e-16 ***
weight	3721.02	81.79	45.50	<2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.84 on 46 degrees of freedom
(1 observation deleted due to missingness)

Multiple R-squared: 0.9783, Adjusted R-squared: 0.9778

F-statistic: 2070 on 1 and 46 DF, p-value: < 2.2e-16

```
diamond.mod<-lm(price~weight, diamond)
summary(diamond.mod)
anova(diamond.mod)
```

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

Analysis of Variance Table

Response: price

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
weight	1	2098596	2098596	2070	< 2.2e-16 ***
Residuals	46	46636	1014		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

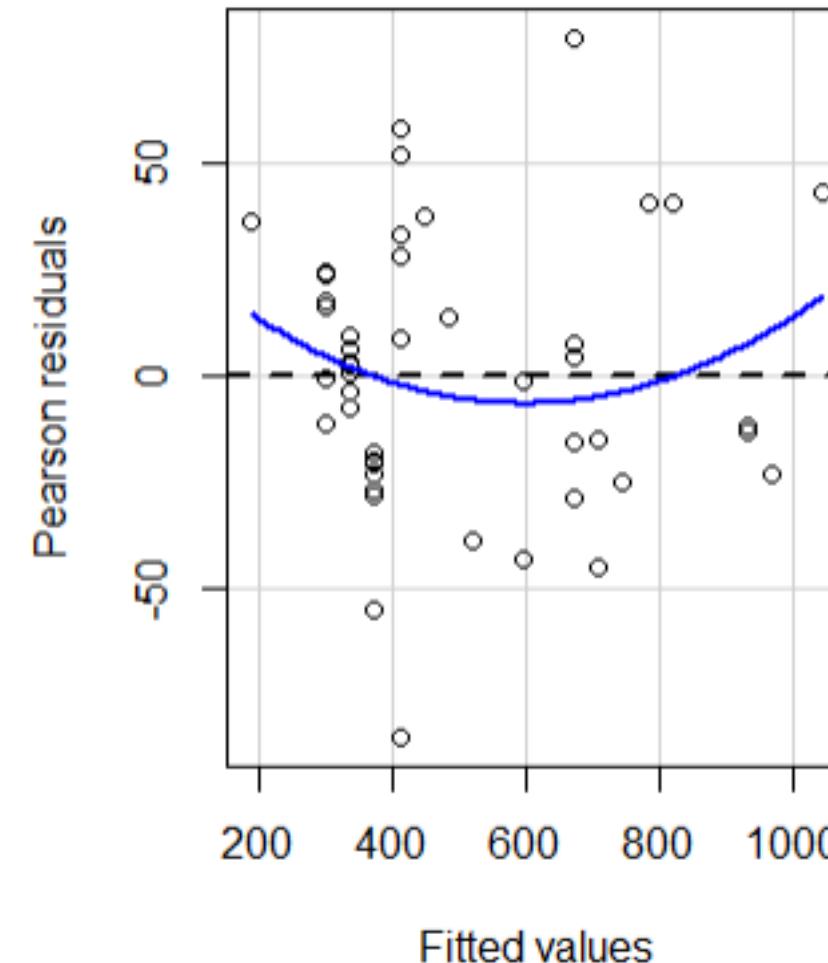
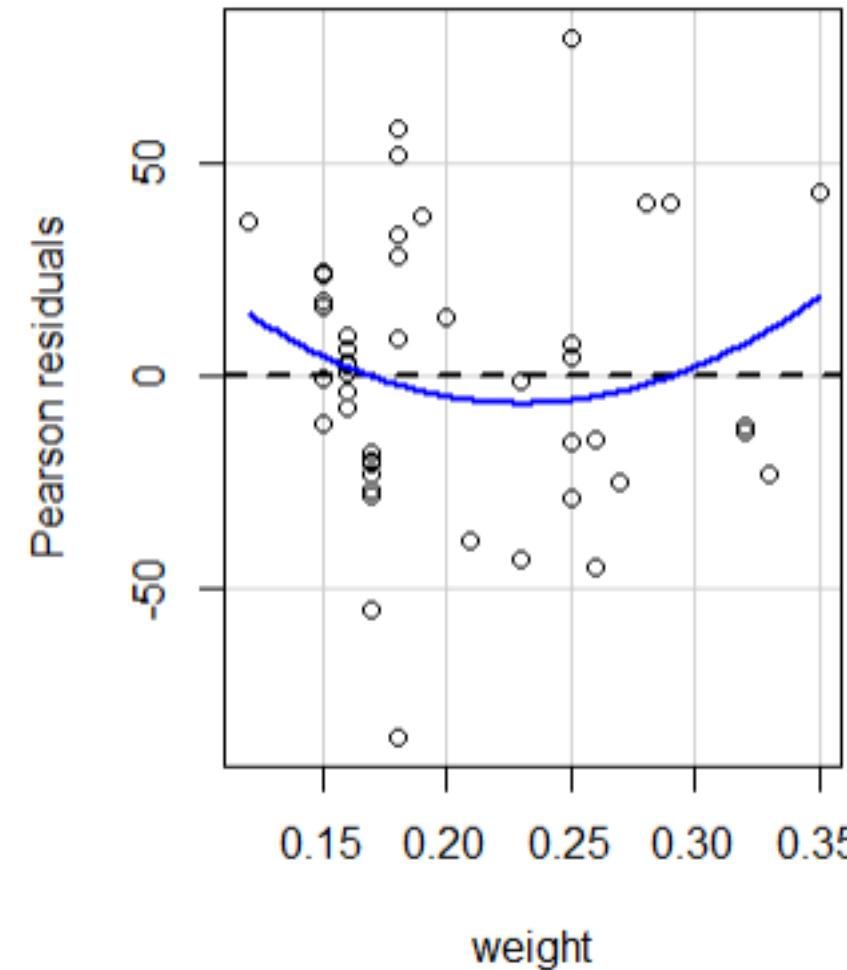
$$MSE =$$

$$s = \sqrt{MSE} =$$

$$DF = n - 2 = \\ \text{after remove 1 observation}$$

Residual plots

`residualPlots(diamond.mod)`



Residuals show a random pattern.

Properties of the LS Line

- The least-squares line always passes through the point (\bar{X}, \bar{Y}) .
- The residuals always sum to zero:

$$\begin{aligned}
 \sum e_i &= \sum [Y_i - (b_0 + b_1 X_i)] \\
 &= \sum Y_i - b_0 - b_1 \sum X_i \\
 &= n \bar{Y} - nb_0 - nb_1 \bar{X} \\
 &= n[(\bar{Y} - b_1 \bar{X}) - b_0] \\
 &= 0
 \end{aligned}$$

- $\sum Y_i = \sum \hat{Y}_i$
- $\sum X_i e_i = 0$
- $\sum \hat{Y}_i e_i = 0$

Statistical inference for the slope and intercept in SLR

Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \text{ for } i = 1, \dots, n$$

Simple Linear Regression Model Parameters

- β_0 is the intercept.
- β_1 is the slope.
- ε_i are independent, normally distributed random errors with mean 0 and variance σ^2 ,

$$\varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

The point estimate of β_1 is b_1

- Recall that,

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

which we can rewrite as,

$$= \sum c_i (Y_i - \bar{Y}) = \sum c_i Y_i - \bar{Y} \sum c_i = \sum c_i Y_i$$

$$\text{where } c_i = \frac{(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$$

It can be proved that, $E(b_1) = \beta_1$ and $\sigma^2(b_1) = \sigma^2 \frac{1}{\sum (X_i - \bar{X})^2}$, therefore

By replacing the parameter σ^2 with MSE , the unbiased estimator of $\sigma^2\{b_1\}$,

$$s^2\{b_1\} = \frac{MSE}{\sum (X_i - \bar{X})^2}$$

$$s\{b_1\} = \sqrt{\frac{MSE}{\sum (X_i - \bar{X})^2}}$$

The Sampling Distribution of b_1 is Normal ($\beta_1, \sigma^2(b_1)$)

$$\frac{b_1 - \beta_1}{s\{b_1\}} \sim t(n - 2)$$

Confidence Interval for β_1

Since $t^* = \frac{b_1 - \beta_1}{s\{b_1\}} \sim t(n - 2)$

$$P \left\{ t \left(\frac{\alpha}{2}; n - 2 \right) \leq \frac{b_1 - \beta_1}{s\{b_1\}} \leq t \left(1 - \frac{\alpha}{2}; n - 2 \right) \right\} = 1 - \alpha$$

Where $t(\frac{\alpha}{2}; n - 2)$ denotes the $(\frac{\alpha}{2})$ 100 percentile of the t distribution with $n - 2$ degrees of freedom.

Because of the symmetry of the t distribution around its mean 0, it follows that:

$$t \left(\frac{\alpha}{2}; n - 2 \right) = -t \left(1 - \frac{\alpha}{2}; n - 2 \right)$$

Hence the $1 - \alpha$ confidence interval for β_1 are:

$$\mathbf{b}_1 \pm t \left(1 - \frac{\alpha}{2}; n - 2 \right) s\{\mathbf{b}_1\}$$

Point estimate \pm Margin error, where Margin error (denoted by ME) = $t^* \text{ standard error}$

Significance Tests for β_1

$$H_0: \beta_1 = \beta_1^* \quad H_a: \beta_1 \neq \beta_1^*$$

The test statistic $t^* = (b_1 - \beta_1^*) / s\{b_1\} \sim t(n - 2)$

For two sided test

Reject H_0 if $|t^*| \geq t_c$, $t_c = t_{n-2}(1 - \alpha/2)$

Or, reject H_0 if $p-value \leq \alpha$

For one sided test

Reject H_0 if $t^* \geq t_c$, $t_c = t_{n-2}(1 - \alpha)$

Or, reject H_0 if $p-value \leq \alpha$

Inference for the intercept, β_0

$$b_0 = \bar{Y} - b_1 \bar{X}$$

It can be proved that, $E(b_0) = \beta_0$ and $\sigma^2\{b_0\} = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right]$

$$s^2\{b_0\} = MSE\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right]$$

$$s\{b_0\} = \sqrt{MSE\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right]}$$

Analogous to theorem for b_1 , $t^* = (b_0 - \beta_0)/s\{b_0\} \sim t(n - 2)$

Confidence Interval for β_0

$$b_0 \pm t\left(1 - \frac{\alpha}{2}; n - 2\right) s\{b_0\}$$

Significance Tests for β_0

$$Ho: \beta_0 = \beta_0^* \quad Ha: \beta_0 \neq \beta_0^*$$

The test statistic $t^* = (b_0 - \beta_0^*) / s\{b_0\} \sim t(n - 2)$

Comments on the inference assumptions

- Both b_1 and b_0 follow *Normal distribution* because they are based on ε which is normally distributed.
- As long as the ε s are **close to normal**, the t-method for the inferences based on b_1 and b_0 is approximately correct, even with small sample sizes.

Comments on the inference assumptions

- Often, the value of the intercept is not of direct interest, so there is no need to calculate CIs or hypothesis tests β_0 . Because it is just a single value of Y when X=0 and will be of no much value to predict other Y values.
- Reduce the standard error for estimating the linear impact, β_1 , by increasing the dilation in X, i.e., bigger $SSX = \sum(X_i - \bar{X})^2$, since $s\{b_1\} = \frac{s}{\sqrt{SSX}}$

One way to do confidence interval for β_1

```
alpha=0.05
n=48
qt(1-0.5*alpha,n-2)
confint(lm(price~weight, diamond), "weight", level=0.95)
```

$$\alpha = 0.05$$

$$n = 48$$

$$t\left(1 - \frac{\alpha}{2}, n - 2\right) = t(0.975, 46) = 2.013$$

$$b_1 \pm t\left(1 - \frac{\alpha}{2}; n - 2\right) s\{b_1\}$$

	2.5 %	97.5 %
weight	3556.398	3885.651

Conclusion: we are 95% confident that,
the average price will
increase by at least 3556 and at most 3889
when the weight increase by 1 carat .

The other way to do confidence interval for β_1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-259.63	17.32	-14.99	<2e-16 ***
weight	3721.02	81.79	45.50	<2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.84 on 46 degrees of freedom

(1 observation deleted due to missingness)

Multiple R-squared: 0.9783, Adjusted R-squared: 0.9778

F-statistic: 2070 on 1 and 46 DF, p-value: < 2.2e-16

$$b_1 \pm t\left(1 - \frac{\alpha}{2}; n - 2\right) s\{b_1\}$$

$$= 3721 \pm 2.013 (81.79) = 3556.4, 3885.65$$

Do hypothesis test for β_1

$H_0: \beta_1 = 3500$ vs $H_a: \beta_1 \neq 3500$

$$\text{The test statistic: } t_s = \frac{b_1 - 3500}{S\{b_1\}} = \frac{3721 - 3500}{81.79} = 2.702$$

The reject region: reject H_0 , if $|t_s| > t\left(1 - \frac{\alpha}{2}, n - 2\right) = t(0.975, 46) = 2.103$

The p value = $2Pr(T > 2.702) = 0.00962$

```
2*(1-pt(2.702,46))
```

```
[1] 0.00962015
```

Conclusion: at a significant level of 5%,
when the weight increases by 1 caret,
the incensement in the average price
is not statistically different
from 3500 dollars.

One sided hypothesis test for β_1

$H_0: \beta_1 = 3500$ vs $H_a: \beta_1 > 3500$

The test statistic: $t_s = \frac{b_1 - 3500}{S\{b_1\}} = \frac{3721 - 3500}{81.79} = 2.702$

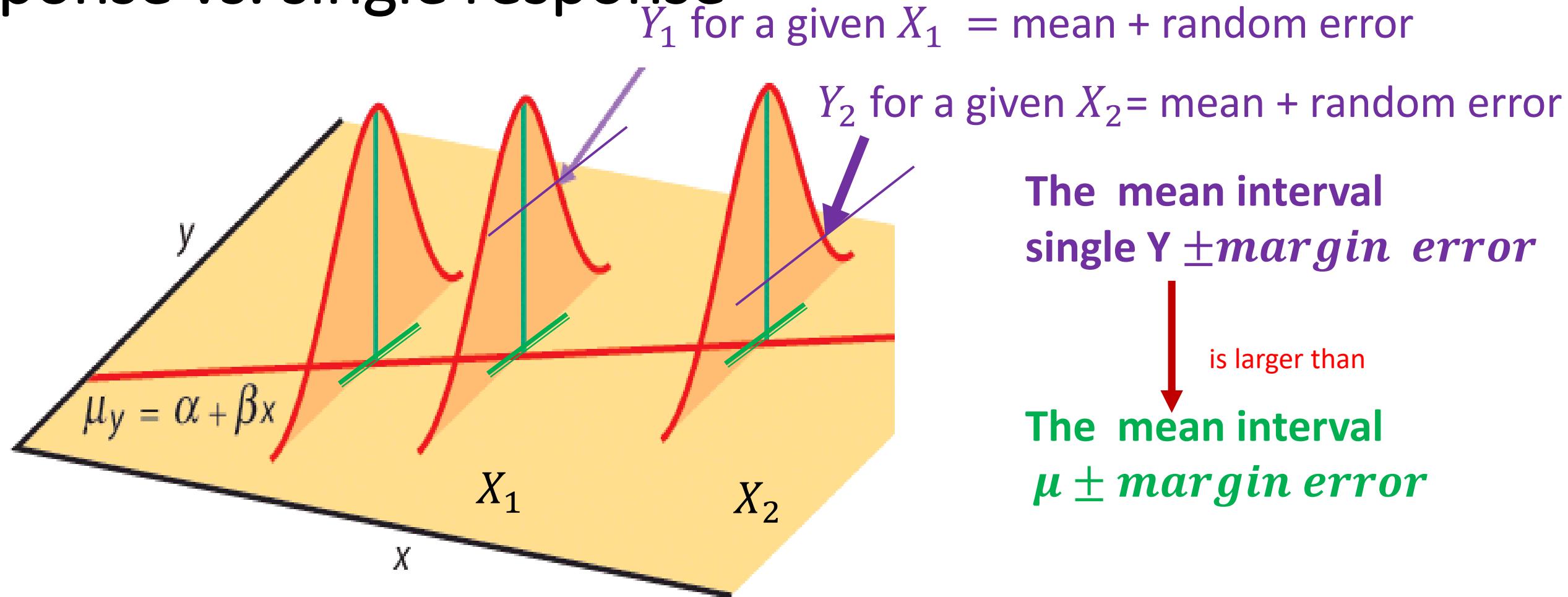
The reject region: reject H_0 , if $|t_s| > t(1 - \alpha, n - 2) = t(0.95, 46) = 1.679$

The p value = $Pr(T > 2.702) = 0.0048$

**Conclusion: at a significant level of 5%,
when the weight increases by 1 caret,
the incensement in the average price
is not statistically greater than
3500 dollars.**

**Interval Estimation of mean response $E\{Y_h\}$, or \hat{Y}_h
and single response $\hat{Y}_h(\text{new})$ when $X = X_h$**

Mean response vs. single response



- ❖ Predict the mean response of Y on X

$$E\{Y_h\} \text{ or } \hat{\mu}_h$$

- ❖ Predict the single response of Y on X

$$\hat{Y}_h$$

} Predict in the **same manner**,
Same value;
But **different precision**.

Recall that

The best point estimates of β_1 and β_0 given the data (X, Y) are:

$$b_1 = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2} = \frac{SS_{XY}}{SS_X} = \sum c_i Y_i \quad E(b_1) = \beta_1 \text{ and } Var(b_1) = \sigma^2 \frac{1}{\sum (X_i - \bar{X})^2}$$

$$b_0 = \bar{Y} - b_1 \bar{X} = \frac{\sum Y_i}{n} - \sum c_i \bar{X} Y_i = \sum d_i Y_i \quad E(b_0) = \beta_0 \text{ and } Var(b_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right]$$

Hence, $\hat{Y}_h = b_0 + b_1 X_h$ is a linear combination of the observations Y_i

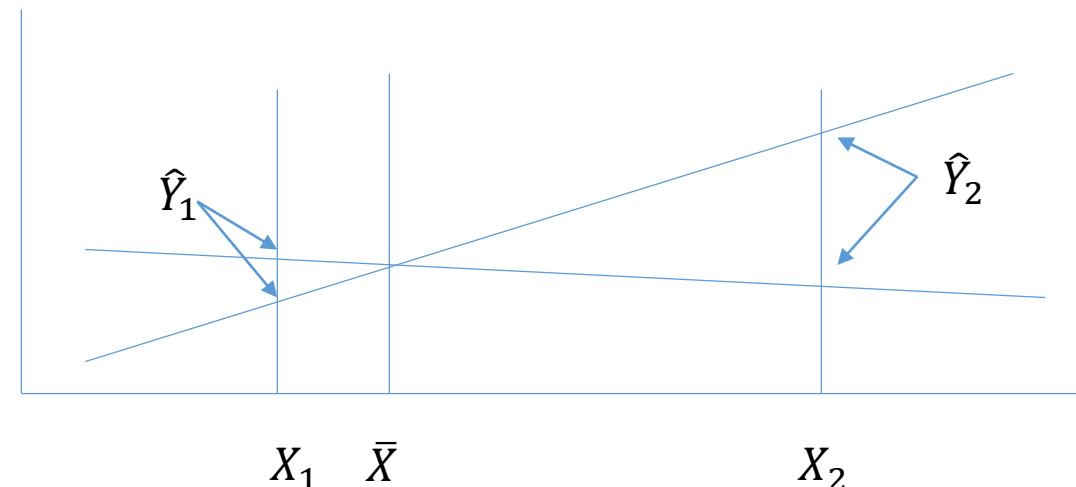
Question1: Does \hat{Y}_h follow normal distribution?

Question2: are b_0 and b_1 independent?

Prediction of the mean response

$$\hat{Y}_h = b_0 + b_1 X_h$$

- For normal error (ε) regression model, $\hat{Y}_h \sim \text{Normal}$, with mean and variance:
- $E\{\hat{Y}_h\} = E\{Y_h\} = \mu_h$
- $\sigma^2\{\hat{Y}_h\} = \sigma^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$
- \hat{Y}_h is normal because $b_0 + b_1 X_h$ is a linear combination of independent, normal Y_i 's.
- Its variance is affected by how far X_h is from \bar{X} , through the term $(X_h - \bar{X})^2$.
- Estimation is more precise near \bar{X}



Prediction of the mean response

$$\hat{Y}_h = b_0 + b_1 X_h$$

- For normal error (ε) regression model, $\hat{Y}_h \sim Normal$, with mean and variance:

$$E\{\hat{Y}_h\} = E\{Y_h\} = \mu_h$$

$$\sigma^2\{\hat{Y}_h\} = \sigma^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

- When replace σ^2 with MSE $s^2\{\hat{Y}_h\} = MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] = s^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$

Therefore, it follows that

$$\frac{\hat{Y}_h - E\{Y_h\}}{s\{\hat{Y}\}} \sim t(n - 2)$$

Prediction confidence interval of mean response, $E\{\hat{Y}_h\}$

$$\frac{\hat{Y}_h - E\{\hat{Y}_h\}}{s\{\hat{Y}_h\}} \sim t(n-2)$$

The confidence interval of $E\{\hat{Y}_h\}$

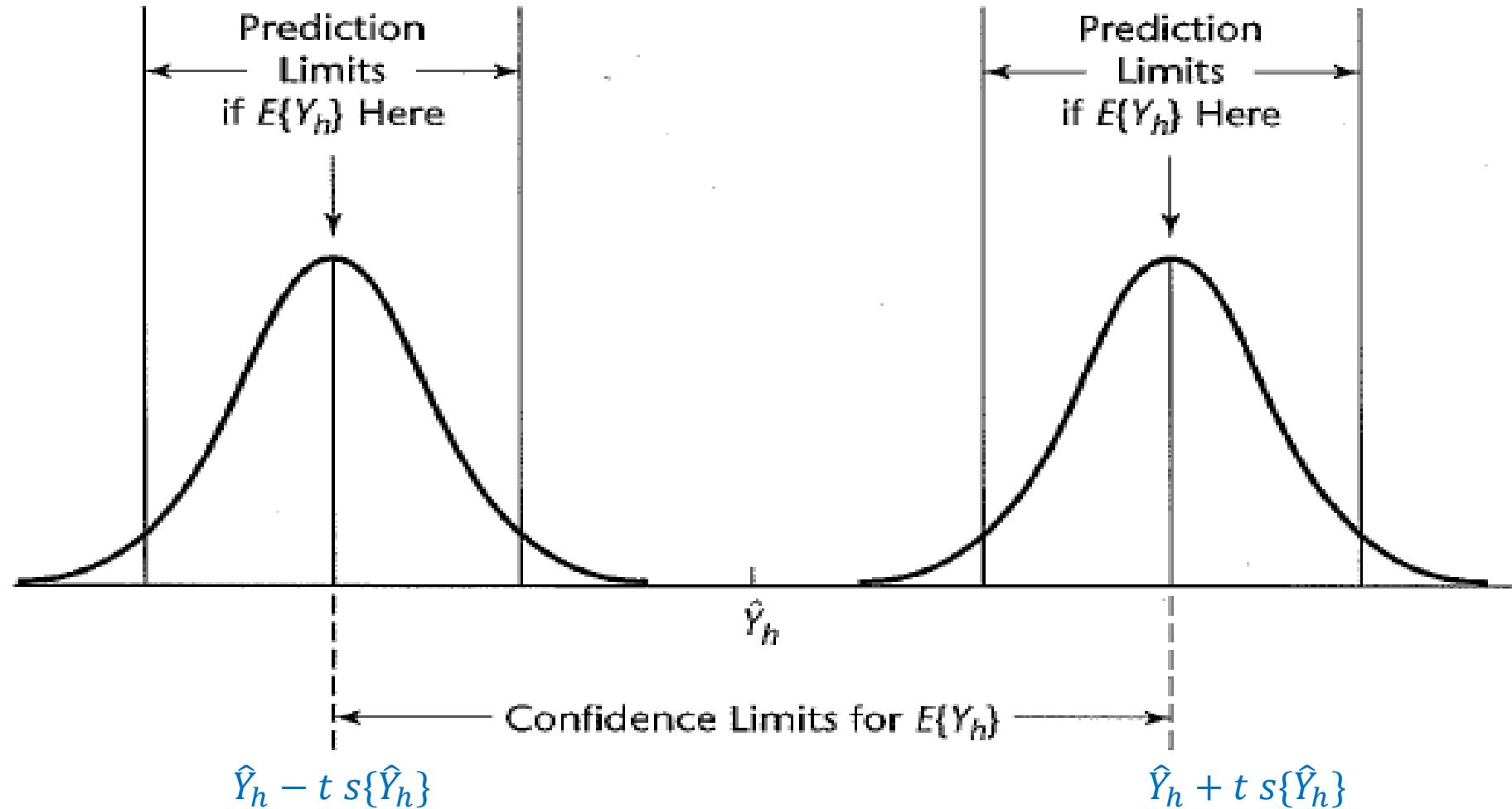
$$\hat{Y}_h \pm t\left(1 - \frac{\alpha}{2}; n-2\right) s\{\hat{Y}_h\}$$

$$\text{where } s_{\{\hat{Y}_h\}}^2 = s^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right]$$

$s\{\hat{Y}_h\}$ is the “standard error of the mean response value at $X = X_h$ ”

s is the “standard error of the residuals”

Prediction of single response $\hat{Y}_{h(new)}$



$$\sigma^2\{Y_{h(new)}\} = \sigma^2\{\hat{Y}_h\} + \sigma^2$$

The variance of prediction = variance in possible location of the distribution + variance within the distribution

We estimate the variance of the single prediction as

$$s_{\{pred\}}^2 = s_{\{\hat{Y}_h\}}^2 + s^2 = s^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} + 1 \right]$$

For normal error regression model

$$\frac{Y_{h(new)} - \hat{Y}_h}{s_{\{pred\}}} \sim t(n-2)$$

$s_{\{pred\}}$ is the “standard error for predicting one new response value at X_h .”

Prediction interval of single response $\hat{Y}_{h(new)}$

$$\hat{Y}_h \pm t\left(1 - \frac{\alpha}{2}; n - 2\right) s_{\{pred\}}$$

$$s_{\{pred\}}^2 = s_{\{\hat{Y}_h\}}^2 + s^2 = s^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} + 1 \right]$$

- More sensitive to departure of normal in error terms distribution.
- Predictions are more precise near \bar{X} because σ^2 decreases with $|X_h - \bar{X}|$.

Prediction interval of mean of m new response $\bar{Y}_{h\{new\}}$ not \hat{Y}_h , or $\hat{Y}_h\{new\}$

$$\hat{Y}_h \pm t \left(1 - \frac{\alpha}{2}; n - 2 \right) s\{predmean\}$$

Where: $s^2\{predmean\} = \frac{MSE}{m} + s^2\{\hat{Y}_h\}$
 $= MSE \left[\frac{1}{m} + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right]$

- Predict the mean of m new observations on Y for a given level of the predictor variables.
- The variance $s^2\{predmean\}$ has two components: variance between the distribution and variance within a distribution.

$s\{predmean\}$ is the “standard error for predicting the mean of m new response value.”

The Diamond example, if $X_h = 0.43$, compute

1. The confidence interval for the mean predicted value $E(\hat{Y}_h)$

$$\hat{Y}_h \pm t_c s_{\{\hat{Y}_h\}}$$

Where $s_{\{\hat{Y}_h\}}^2 = s^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right]$

2. The confidence interval for the single predicted value (\hat{Y}_h)

$$\hat{Y}_h \pm t_c s_{\{pred\}}$$

Where $s_{\{pred\}}^2 = s_{\{\hat{Y}_h\}}^2 + s^2 = s^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} + 1 \right]$

3. The confidence interval for the mean price $\bar{Y}_{h(new)}$ of three diamonds with the same weight (0.43)

$$\hat{Y}_h \pm t_c s_{\{predmean\}}$$

Where: $s^2_{\{predmean\}} = \frac{MSE}{m} + s^2_{\{\hat{Y}_h\}} = MSE \left[\frac{1}{m} + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right]$

Recall that in the Diamond example

The lm output

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-259.63	17.32	-14.99	<2e-16 ***
weight	3721.02	81.79	45.50	<2e-16 ***
<hr/>				
Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’
	0.1 ‘ ’	1		

Residual standard error: 31.84 on 46 degrees of freedom
 (1 observation deleted due to missingness)

Multiple R-squared: 0.9783, Adjusted R-squared: 0.9778
 F-statistic: 2070 on 1 and 46 DF, p-value: < 2.2e-16

$$MSE = s^2 = 31.84^2 = 1013.8$$

$$\bar{X} = 0.204, s_X = 0.0568, n=48$$

Recall that in the diamond example

Where $\alpha = 0.05, n = 48, df = 46$ round down to 40

Degrees of freedom	Confidence level C											
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
One-sided P	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
Two-sided P	.50	.40	.30	.20	.10	.05	.04	.02	.01	.005	.002	.001

Or use R

$$t\left(1 - \frac{\alpha}{2}, n - 2\right) = t(0.975, 46)$$

= 2.021 (estimation using the t table)

```
qt(1 - 0.5 * alpha, n - 2)
```

= 2.013 (exact value using R)

[1] 2.012896

The Diamond example, if $X_h = 0.43$, compute

1. The confidence interval for the mean predicted value $E(\hat{Y}_h)$

$$\hat{Y}_h \pm t_c s_{\{\hat{Y}_h\}} \quad \text{Where } s_{\{\hat{Y}_h\}}^2 = s^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right]$$

$$SS_X = s_X^2(n - 1) =$$

$$t\left(1 - \frac{\alpha}{2}, n - 2\right) = t(0.975, 46) = 2.021 \text{ (estimation from T-table) or } 2.013 \text{ (exact value from R)}$$

$$s_{\{\hat{Y}_h\}}^2 = s^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right] = s_{\{\hat{Y}_h\}} =$$

$$\hat{Y}_h \pm t_c s_{\{\hat{Y}_h\}} = 1340.415 \pm 2.013(19.03) = 1302.1, 1378.73.$$

```
ci.reg(diamond.mod, new, type='m', alpha=0.05)
```

The Diamond example, if $X_h = 0.43$, compute

1. The confidence interval for the mean predicted value $E(\hat{Y}_h)$

$$\hat{Y}_h \pm t_c s_{\{\hat{Y}_h\}} \quad \text{Where } s_{\{\hat{Y}_h\}}^2 = s^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right]$$

$$SS_X = s_X^2(n-1) = 0.0568^2(48-1) = 0.152$$

$$t\left(1 - \frac{\alpha}{2}, n-2\right) = t(0.975, 46) = 2.021 \text{ (estimation from T-table) or } 2.013 \text{ (exact value from R)}$$

$$s_{\{\hat{Y}_h\}}^2 = s^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right] = 31.84^2 \left[\frac{1}{48} + \frac{(0.43 - 0.204)^2}{0.152} \right] = 362.14 \quad s_{\{\hat{Y}_h\}} = \sqrt{362.14} = 19.03$$

$$\hat{Y}_h \pm t_c s_{\{\hat{Y}_h\}} = 1340.415 \pm 2.013(19.03) = 1302.1, 1378.73.$$

```
ci.reg(diamond.mod, new, type='m', alpha=0.05)
```

The Diamond example, if $X_h = 0.43$, compute

2. The confidence interval for the single predicted value (\hat{Y}_h)

$$\hat{Y}_h \pm t_c s_{\{pred\}} \quad \text{Where } s_{\{pred\}}^2 = s_{\{\hat{Y}_h\}}^2 + s^2 = s^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} + 1 \right]$$

$$s_{\{pred\}}^2 = s_{\{\hat{Y}_h\}}^2 + s^2 =$$

$$\hat{Y}_h \pm t_c s_{\{pred\}} = 1340.415 \pm 2.013(37.093) = 1265.75, 1415.08$$

```
ci.reg(diamond.mod, new, type='n', alpha=0.05)
```

The Diamond example, if $X_h = 0.43$, compute

2. The confidence interval for the single predicted value (\hat{Y}_h)

$$\hat{Y}_h \pm t_c s_{\{pred\}} \quad \text{Where } s_{\{pred\}}^2 = s_{\{\hat{Y}_h\}}^2 + s^2 = s^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} + 1 \right]$$

$$s_{\{pred\}}^2 = s_{\{\hat{Y}_h\}}^2 + s^2 = 362.14 + 1013.78 = 1375.92$$

$$\hat{Y}_h \pm t_c s_{\{pred\}} = 1340.415 \pm 2.013(37.093) = 1265.75, 1415.08$$

```
ci.reg(diamond.mod, new, type='n', alpha=0.05)
```

The Diamond example, if $X_h = 0.43$, compute

3. The confidence interval for the mean price $\bar{Y}_{h(new)}$ of three diamonds with the same weight (0.43)

$$\hat{Y}_h \pm t_c s\{predmean\} \quad \text{where: } s^2\{predmean\} = \frac{MSE}{m} + s^2\{\hat{Y}_h\} = MSE \left[\frac{1}{m} + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right]$$

$$s^2\{predmean\} = \frac{MSE}{m} + s^2\{\hat{Y}_h\} =$$

$$\hat{Y}_h \pm t_c s\{predmean\} = 1340.415 \pm 2.013(26.46) = (1287.151, 1393.679)$$

```
ci.reg(diamond.mod, new, type='nm', m=3, alpha=0.05)
```

The Diamond example, if $X_h = 0.43$, compute

3. The confidence interval for the mean price $\bar{Y}_{h(new)}$ of three diamonds with the same weight (0.43)

$$\hat{Y}_h \pm t_c s\{predmean\} \quad \text{Where: } s^2\{predmean\} = \frac{MSE}{m} + s^2\{\hat{Y}_h\} = MSE \left[\frac{1}{m} + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right]$$

$$s^2\{predmean\} = \frac{MSE}{m} + s^2\{\hat{Y}_h\} = \frac{31.84^2}{3} + 362.14 = 700.07$$

$$\hat{Y}_h \pm t_c s\{predmean\} = 1340.415 \pm 2.013(26.46) = (1287.151, 1393.679)$$

```
ci.reg(diamond.mod, new, type='nm', m=3, alpha=0.05)
```

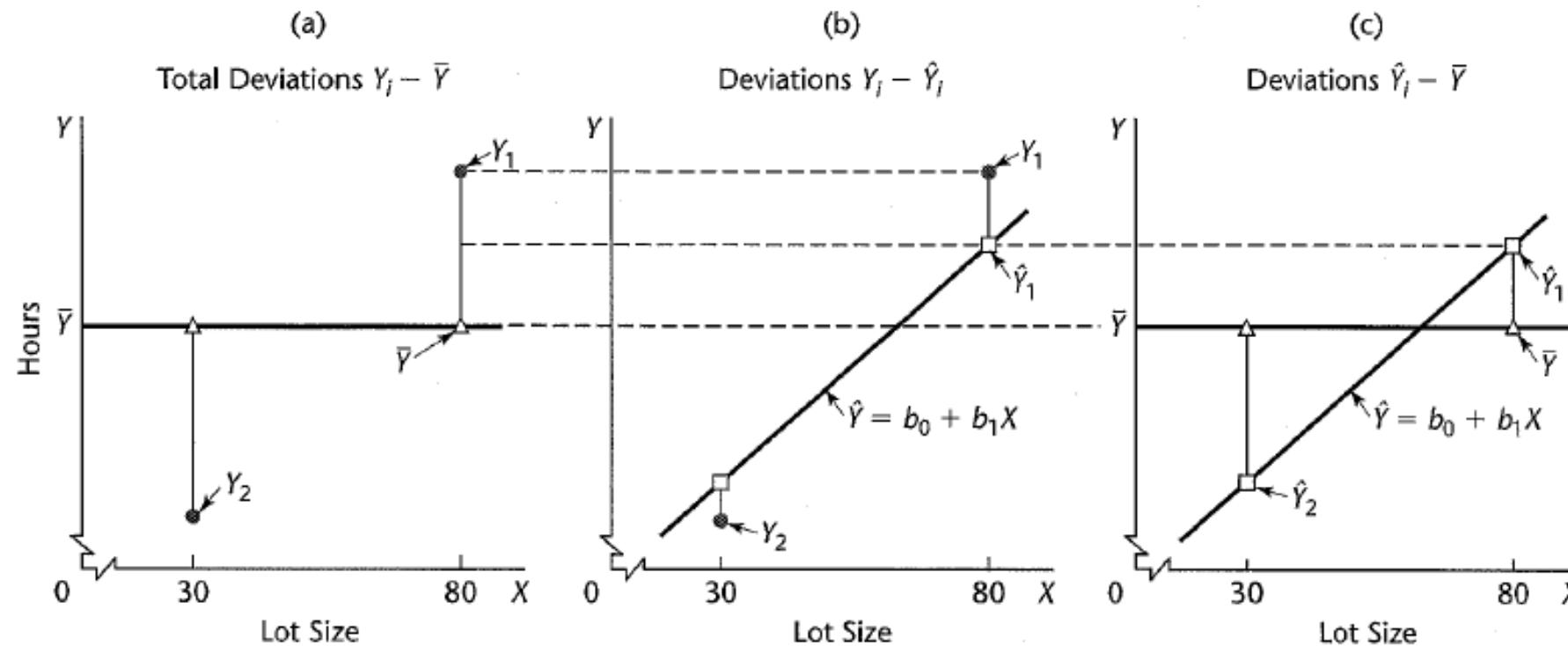
The ANOVA F test and the General Linear Test (GLT)

The Analysis of Variance Test (ANOVA)

- The ANOVA test is a hypothesis test to study different variances from different resource in the data
 - The most common type of ANOVA test is the **Global F test**, also known as **the significance test of the model**.
 - The test statistic follows a F distribution; therefore, it is a F test which **sometimes** can be replaced by a T-test.
- The General Linear Test (GLT) is a test to study different variances in different models defined in H_0 (Reduced model) and H_a (Full model), respectively.
 - It uses a F test to analyze the variances, hence it is **essentially an ANOVA test**.
 - GLT test is usually used in model improvement.
 - It is different from the Generalized Linear Model (GLM).

Partitioning variance in the total sum of squares

$$Y_i - \bar{Y}$$



$$\Sigma(Y_i - \bar{Y})^2 = \Sigma(Y_i - \hat{Y}_i)^2 + \Sigma(\hat{Y}_i - \bar{Y})^2$$

$$SSTO = SSE + SSR \quad \text{Also known as SSM (model)}$$

"Total sum of squares"

"error sum of squares"

"regression sum of squares"

Partitioning Degree of freedom

$$\Sigma(Y_i - \bar{Y})^2 = \Sigma(Y_i - \hat{Y}_i)^2 + \Sigma(\hat{Y}_i - \bar{Y})^2$$

$$SSTO = SSE + SSR$$

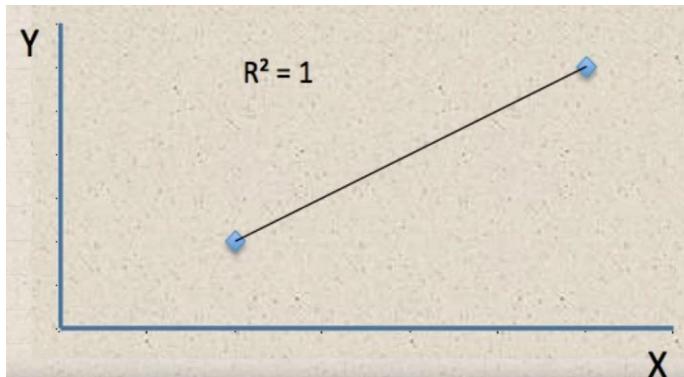
Degree of freedom	$n - 1$	=	$n - 2$	+	1
-------------------	---------	---	---------	---	---

Degree of freedom of error (an intuitive flavor)

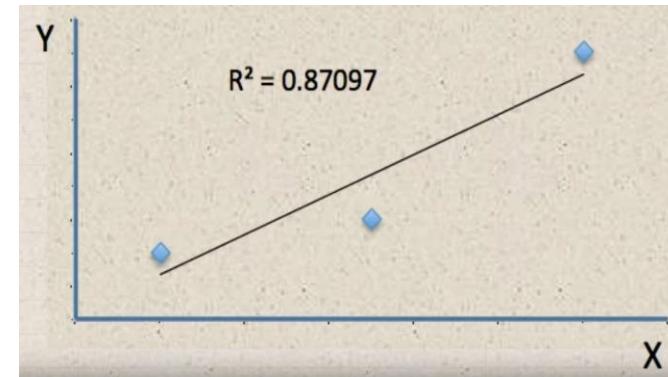
Q: what is the minimum requirement on data points to estimate this regression?

$$Y = \beta_0 + \beta_1 X + \epsilon, \text{ and } \epsilon = Y - \beta_0 - \beta_1 X$$

$$n = 2, dfE = 0$$



$$n = 3, dfE = 1$$



$$\text{So, } dfE = n - 2 = n - p$$

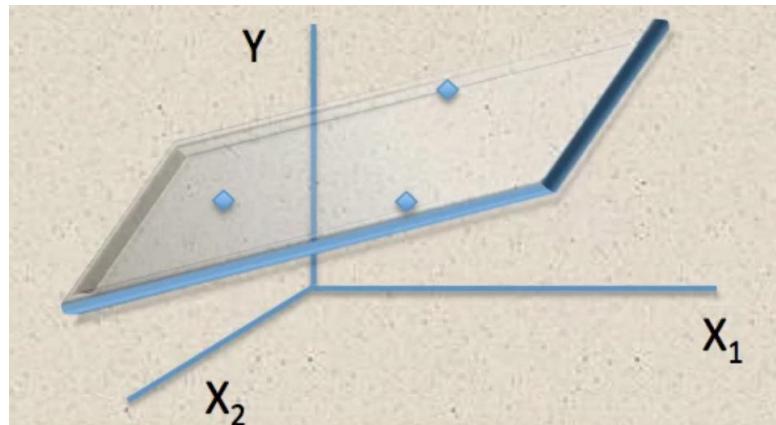
Where p is the number of parameters ($p = 2$ in this case)

Degree of freedom of error (an intuitive flavor)

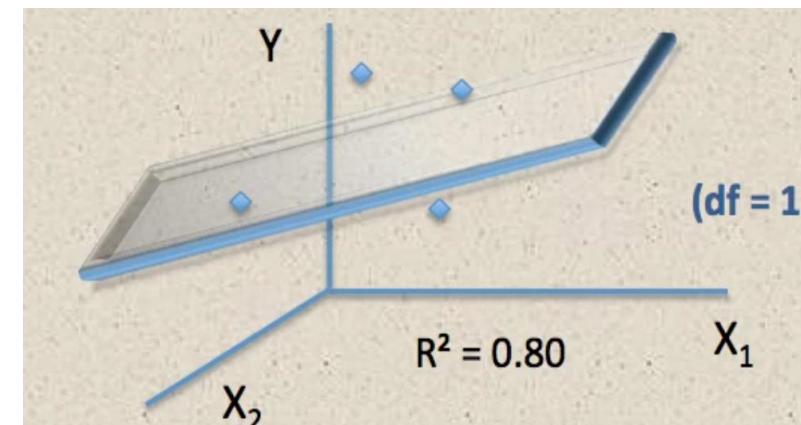
Q: what is the minimum requirement on data points to estimate this regression? What is the degree of freedom left?

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon, \text{ and } \epsilon = Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2$$

$$n = 3, dfE = 0$$



$$n = 4, dfE = 1$$



$$\text{So, } dfE = n - 3$$

$$= n - p$$

Where p is the number of parameters ($p = 3$ in this case)

The F test of $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$

This F test is also known as the Significant test of a SLR model, or the significant linear impact of the independent variable.

Source of Variation	SS	df	MS	$E\{MS\}$
Regression	$SSR = \sum (\hat{Y}_i - \bar{Y})^2$ $= b_1^2 \sum (X_i - \bar{X})^2$	1	$MSR = \frac{SSR}{1}$	$\sigma^2 + \beta_1^2 (X_i - \bar{X})^2$
Error	$SSE = \sum (Y_i - \hat{Y}_i)^2$	$n - 2$	$MSE = \frac{SSE}{n-2}$	σ^2
Total	$SSTO = \sum (Y_i - \bar{Y})^2$	$n - 1$		

The test statistic is denoted by F^* or $F_s = \frac{MSR}{MSE} \sim F(1, n - 2)$

Reject H_0 if $F^* > F(1 - \alpha; 1, n - 2)$

Example 1 Complete the hypothesis test $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$
 On a (partial) given ANVOA table.

Source of Variation	SS	df	MS	F
Regression	252378	1	252378/1=252378	232378/2384=105.88
Error	54825	23	54825/23=2384	
Total	307203	24		

$$F_s = \frac{MSR}{MSE} = 105.88 \sim F(1, 23)$$

Reject H_0 if

$$F_s > F(0.95; 1, 23) = 4.28$$

qf(0.95, 1, 23)

Conclude that X has a significant linear impact on Y, or the SLR model is statistically significant.

Example 2 Complete the hypothesis test $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$
On a model summary output.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-259.63	17.32	-14.99	<2e-16 ***
weight	3721.02	81.79	45.50	<2e-16 ***
<hr/>				
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

Residual standard error: 31.84 on 46 degrees of freedom
(1 observation deleted due to missingness)

Multiple R-squared: 0.9783, Adjusted R-squared: 0.9778

F-statistic: 2070 on 1 and 46 DF, p-value: < 2.2e-16

$$F_s = 2070 \sim F(1, 46)$$

Conclude that X has a significant linear impact on Y, or the SLR model is statistically significant.

Equivalence of a two-sided F test (ANOVA) and t test (SLR)

$$H_0: \beta_1 = 0 \quad H_a: \beta_1 \neq 0$$

The T test statistic $t_s = \frac{b_1}{s\{b_1\}} \sim t(n - 2)$

The F test statistic $F_s = \frac{MSR}{MSE} \sim F(1, n - 2)$

$$F_s = \frac{MSR}{MSE} = \frac{b_1^2 \Sigma(X_i - \bar{X})^2}{MSE} = \frac{b_1^2}{s^2\{b_1\}} = t_s^2$$



Since $s^2\{b_1\} = MSE / \Sigma(X_i - \bar{X})^2$

The T test and F tests are equivalent in SLR $F_s = (t_s)^2$ for two sided test,
the critical values:

$$t\left(1 - \frac{\alpha}{2}, n - 2\right)^2 = F(1 - \alpha; 1, n - 2)$$

For example, at $\alpha = 0.05, dfe = 23$: $t(0.975; 23)^2 = (2.069)^2 = 4.28 = F(0.95, 1, 23)$

Example 3 The equivalence of the F and the T test in the SLR.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)							
(Intercept)	-259.63	17.32	-14.99	<2e-16	***						
weight	3721.02	81.79	45.50	<2e-16	***						

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'. '	0.1	' '	1

$$ts = \frac{3721.02}{81.79} = 45.5 \sim t(46)$$

Residual standard error: 31.84 on 46 degrees of freedom

(1 observation deleted due to missingness)

Multiple R-squared: 0.9783, Adjusted R-squared: 0.9778

F-statistic: 2070 on 1 and 46 DF, p-value: < 2.2e-16

$$Fs = 2070 \sim F(1, 46)$$

The T-test and the F-test are the same because $45.5^2 = 2070$,
 And they both have the same p-value.

F Test (ANOVA) and T test are **not always equivalent**

1. They are equivalent in simple linear regression (SLR) problem and will not be so for Multiple regression.
2. They are equivalent when $H_0 : \beta_1 = 0$.
 - $H_0 : \beta_1 = \beta_1^* (\neq 0)$ can be tested with a *t*-test.
 - In $H_0 : \beta_1 = \beta_1^* (\neq 0)$, the test statistic F^* has a *non-central F* distribution and require extra steps and not covered in the course.
3. In SLR, the T test is more flexible and more commonly used than the F test. We will continue to compare them in MLR.

The General Linear Test (GLT) approach

Ho: $\beta_1 = 0$ versus **Ha:** $\beta_1 \neq 0$

Full model:

$$Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i$$

Under Ha

$$SSE(F) = \sum(Y_i - \hat{Y}_i)^2 = SSE, \quad df_F = n - 2$$

Reduced model:

$$Y_i = \beta_0 + \epsilon_i$$

Under Ho

$$SSE(R) = \sum(Y_i - \bar{Y}_i)^2 = SSTO, \quad df_R = n - 1$$

“Significant reduction in SSE?”

$$F^* = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)/df_F}{MSE}} = \frac{MSR}{MSE} \sim F(1, n - 2)$$

The test statistic of the general linear test in simple linear regression is identical to the ANOVA test statistic.

Example 4 The global F test in example 1 can convert to a GLT test

Source of Variation	SS	df	MS	F
Regression	252378	1	252378	105.88
Error	54825	23	2384	
Total	307203	24		

Full model: $Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i$

Under H_a

$$SSE(F) = \sum(Y_i - \hat{Y}_i)^2 = SSE = 54825, \quad df_F = n - 2 = 25 - 2 = 23$$

Reduced model: $Y_i = \beta_0 + \epsilon_i$

Under H₀

$$SSE(R) = \sum(Y_i - \bar{Y}_i)^2 = SSTO = 307203, \quad df_R = n - 1 = 25 - 1 = 24$$

$$F_s = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)/df_F}{23}} = \frac{\frac{307203 - 54825}{24 - 23}}{\frac{54825}{23}} = \frac{\frac{252378}{1}}{2384} = 105.88, \text{ which is same as the test statistic in the Global F test, } F_s = \frac{MSR}{MSE}$$

General Linear Test can be extended to multiple parameters (β_1, β_2, \dots)

Given the number of additional parameters in the full (more complex) model compared to the reduced model, does the full model yield a larger reduction in SSE than we would expect to get by adding a similar number of unrelated (i.e., useless) predictor variables?

$$F^* = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)/df_F}{MSE}} = \frac{MSR}{MSE} \sim F(df_R - df_F, df_F)$$

The GLT is a very general tool.

We will see it again in Multiple Linear Regression.

Summary

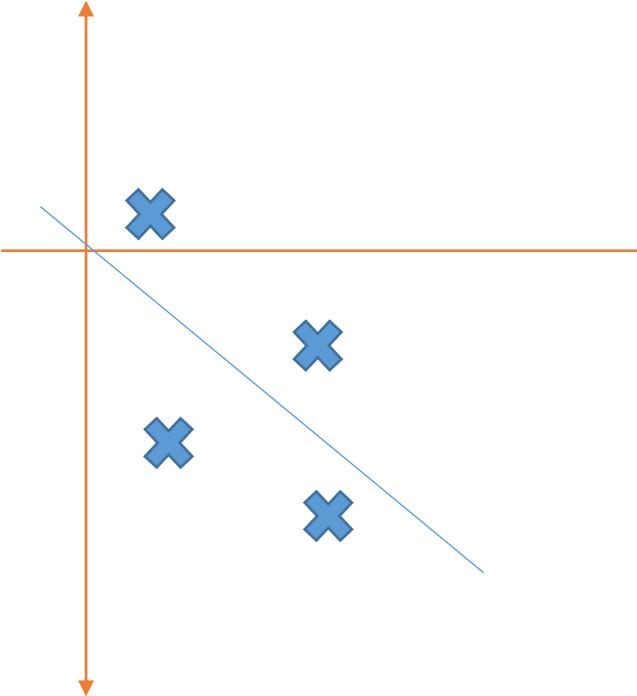
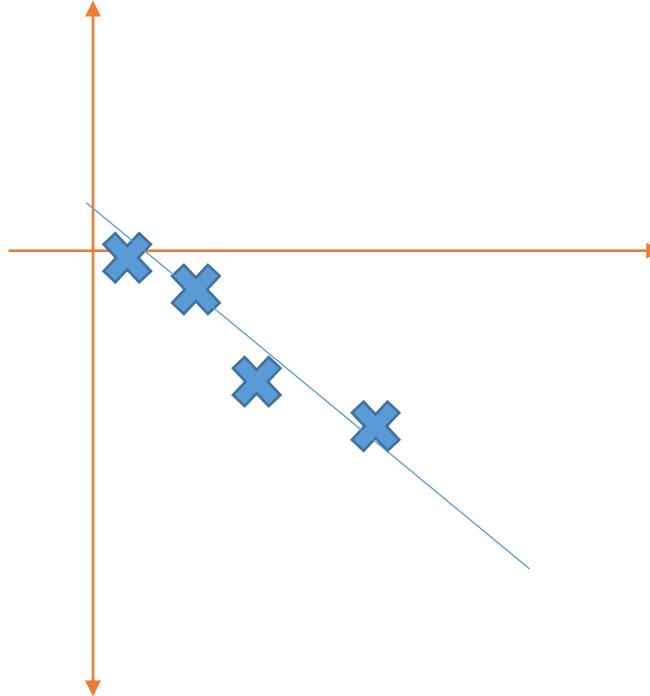
- The basic idea of ANOVA is to study the source and proportion of variance in data
- F test (ANOVA) and T test (Simple Linear model) are not always equivalent
- GLT can be used to compare two models that containing different X variables, and decide whether (dropping) some of the X variable affect the effectiveness of the linear model to explain the variance in Y.

Linear Association, Pearson Correlation and R^2

The *linear impact* and *the linear association measures different perspectives of the linear relationship between two continuous variables.*

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \quad \text{vs.}$$

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$



The *correlation coefficient* ρ (estimated by r) varies from -1 to 1 :

$$r = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2 \Sigma(Y_i - \bar{Y})^2}}$$

$$b_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2} \quad \rightarrow \quad = b_1 \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{\Sigma(Y_i - \bar{Y})^2}} = b_1 \frac{s\{X\}}{s\{Y\}}$$

The correlation coefficient measures the *direction* and *strength* of the “*mutual*” *linear* relationship between two *continuous* variables

The Coefficient of Determination, R^2

$$R^2 = \frac{SSR}{SST} \quad \text{measures the } \underline{\text{proportion of variation in } Y \text{ explained by the model}}.$$

In SLR, $R^2 = r^2$

$$r = b_1 \sqrt{\frac{\sum(X_i - \bar{X})^2}{\sum(Y_i - \bar{Y})^2}}$$

$$r^2 = b_1^2 \frac{\sum(X_i - \bar{X})^2}{\sum(Y_i - \bar{Y})^2} = \frac{SSR}{SST}$$

In MLR, there is only one R^2 , but could be multiple r^2 for any pair of continuous variables, either between X and Y or X_i and X_j .

The Toluca Company example

The Toluca company manufactures refrigeration equipment as well as many replacement parts. In the past, one of the Replacement parts has been produced periodically in lots of varying sizes. When a cost improvement program was undertaken, company officials wished to determine the optimum **lot size (X)** for producing this part. The production of this part involves setting up the production process and machining and assembly operations. One key input for the model to ascertain the optimum lot size was the relationship between lot size and **labor hours (Y)** required to produce the lot.

To determine this relationship, data on lot size and work hours for **25 (n)** recent production runs were utilized.

	x	y
1	80	399
2	30	121
3	50	221
4	90	376
5	70	361
6	60	224
7	120	546
8	80	352
9	100	353
10	50	157

R^2 for Toluca Example

Source of Variation	SS	df	MS	F
Regression	252378	1	252378	105.88
Error	54825	23	2384	
Total	307203	24		

```
toluca.mod<-lm(hour~size, toluca)
summary(toluca.mod)
anova(toluca.mod)
```

$$R^2 = \frac{SSR}{SST} = \frac{252378}{307203} = 0.8215$$

From SLR output:

Residual standard error: 48.82 on 23 degrees of freedom

Multiple R-squared: 0.8215, Adjusted R-squared: 0.8138

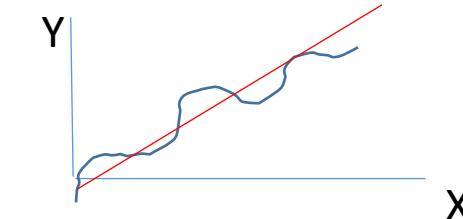
F-statistic: 105.9 on 1 and 23 DF, p-value: 4.449e-10

Thus, *the variation in the workload (Y) is explained by 82 percent by X through a linear model.*

R^2 can be used as a criterion to access a linear regression model,

if and only if the relationship between X and Y is linear;

then we can safely conclude that the model is good due to a large R^2 .



The adjusted R²

Source of Variation	SS	df	MS	F
Regression	252378	1	252378	105.88
Error	54825	23	2384	
Total	307203	24		

From SLR output:

Residual standard error: 48.82 on 23 degrees of freedom

Multiple R-squared: 0.8215, Adjusted R-squared: 0.8138

F-statistic: 105.9 on 1 and 23 DF, p-value: 4.449e-10

$$\text{Adj } R^2 = 1 - \frac{\text{SSE/DfE}}{\text{SST/DfT}} = 1 - \frac{(1 - R^2)(n - 1)}{n - p} = 1 - \frac{(1 - 0.8215^2)(25 - 1)}{25 - 2} = 0.8138$$

Since R^2 usually can be made larger by including a larger number of predictor variables, the adjusted coefficient of Determination (*the Adj R² or R_a^2*) is used to adjust for the number of X values in the model.

Compute r in R

This function is also useful to check the linear association among the independent variables X.

```
cor(toluca)
```

```
cor(toluca$hour,toluca$size)
```

	size	hour
size	1.000000	0.9063848
hour	0.9063848	1.0000000

```
cor(toluca)^2
```

	size	hour
size	1.000000	0.8215335
hour	0.8215335	1.0000000

Inference on *correlation coefficients*

Research question: Points deviate far from the line?

The following can be used to test if
X and Y have no linear association

$$H_0: \rho = 0 \\ H_a: \rho \neq 0$$

is equivalent to

$$r = b_1 \frac{s\{X\}}{s\{Y\}}$$

$$t_s = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

is equivalent to

Research question: X has low impact on Y?

The following can be used to test if
X has no impact on Y

$$H_0: \beta_1 = 0 \\ H_a: \beta_1 \neq 0$$

$$t_s = \frac{b_1}{s\{b_1\}}$$

Inference on *correlation coefficients*

$$H_0: \rho = 0 \\ H_a: \rho \neq 0$$

is equivalent to

$$r = b_1 \frac{s\{X\}}{s\{Y\}}$$

$$H_0: \beta_1 = 0 \\ H_a: \beta_1 \neq 0$$

$$t_s = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

is equivalent to

$$t_s = \frac{b_1}{s\{b_1\}}$$

Pearson's product-moment correlation

```
data: toluca$hour and toluca$size
t = 10.29, df = 23, p-value = 4.449e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.7965202 0.9583070
sample estimates:
cor
0.9063848
```

```
cor.test(toluca$hour, toluca$size, conf.level=0.95)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	62.366	26.177	2.382	0.0259 *
size	3.570	0.347	10.290	4.45e-10 ***

```
toluca.mod<-lm(hour~size, toluca)
summary(toluca.mod)
```

Inference on *correlation coefficients*

Research question: Points deviate far from the line?

The following can be used to test if
X and Y have a certain amount of linear association

$$\begin{aligned} H_0: \rho = \rho^* \\ H_a: \rho \neq \rho^* \end{aligned}$$

is NOT equivalent to

Research question: X has low impact on Y?

The following can be used to test if
X has a certain impact on Y

$$\begin{aligned} H_0: \beta_1 = \beta^* \\ H_a: \beta_1 \neq \beta^* \end{aligned}$$

$$t_s = \frac{b_1 - \beta^*}{s\{b_1\}}$$

Compute the confidence interval on *correlation coefficients*

- $z' = \frac{1}{2} \log_e(\frac{1+r}{1-r})$ is the **Fisher z transformation**.
- $E\{z'\} = \frac{1}{2} \log_e(\frac{1+\rho}{1-\rho})$
- $\sigma^2\{z'\} = \frac{1}{n-3}$
- $(z' - E\{z'\})/\sigma\{z'\}$ is approximately Normal (0,1)

$$z' \pm z(1 - \alpha/2)\sigma\{z'\}$$

- Remember, we still need to transform back to ρ !

$$r = (e^{2z'} - 1)/(e^{2z'} + 1)$$

Example: what is a 90% CI for correlation coefficients in the Toluca example

- $z' = \frac{1}{2} \log_e(\frac{1+r}{1-r})$ is the **Fisher z transformation**.
- $E\{z'\} = \frac{1}{2} \log_e(\frac{1+\rho}{1-\rho})$
- $\sigma^2\{z'\} = \frac{1}{n-3}$
- $(z' - E\{z'\})/\sigma\{z'\}$ is approximately Normal (0,1)

$$z' \pm z(1 - \alpha/2)\sigma\{z'\}$$

$$r = (e^{2z'} - 1)/(e^{2z'} + 1)$$

The CI for ρ is (_____, _____)

$$z' = \frac{1}{2} \log_e(\frac{1+r}{1-r}) = \frac{1}{2} \log_e\left(\frac{1+0.9064}{1-0.9064}\right) = 1.505$$

$$\sigma^2\{z'\} = \frac{1}{n-3} = \frac{1}{25-3} = 0.045$$

$$\sigma\{z'\} = \sqrt{0.045} = 0.213$$

$$\begin{aligned} z' &\pm z(1 - \alpha/2)\sigma\{z'\} = 1.505 \pm z(0.95)(0.213) \\ &= 1.505 \pm 1.645(0.213) \end{aligned}$$

$$=(1.154, 1.856)$$

$$r = (e^{2z'} - 1)/(e^{2z'} + 1) = (e^{2*1.154} - 1)/(e^{2*1.154} + 1) = 0.819$$

The CI for ρ is (0.819, 0.952)

Example: what is a 90% CI for correlation coefficients in the Toluca example

```
r=cor(toluca$hour,toluca$size);  
cor.test(toluca$hour,toluca$size, conf.level=0.90)
```

```
Pearson's product-moment correlation  
  
data: toluca$hour and toluca$size  
t = 10.29, df = 23, p-value = 4.449e-10  
alternative hypothesis: true correlation is not equal to 0  
90 percent confidence interval:  
 0.8197982 0.9524538  
sample estimates:  
      cor  
0.9063848
```

- Conclusion: we are 90% confident that the correlation between hour and size is at least 0.82 and at most 0.95.
- The CI doesn't contain 0 so you may also conclude that the linear association is significant.

Notes on R^2

- The coefficient (of) determination
- R^2 is often expressed as a percentage instead of a proportion. It measures the linear determination of variation of Y by a linear model (not by X)
- In MLR there will be a different r between Y and each predictor variable X, but only one R^2 for the whole model.
- In MLR, we often use *adjusted R²* which has been adjusted to account for the number of variables in the model
- Low or high R^2 does not imply no or high functional relationship without checking linearity first.

The Lack of Fit Test

Review: use the General Linear Test (GLT) approach to test the linear impact

Ho: $\beta_1 = 0$ versus Ha: $\beta_1 \neq 0$

Full model:

$$Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i$$

Under Ha

$$SSE(F) = \sum(Y_i - \hat{Y}_i)^2 = SSE, \quad df_F = n - 2$$

Reduced model:

$$Y_i = \beta_0 + \epsilon_i = \bar{Y}_{grand\ mean} + \varepsilon_i$$

Under Ho

$$SSE(R) = \sum(Y_i - \bar{Y}_{grand\ mean})^2 = SSTO, \quad df_R = n - 1$$

"Significant reduction in SSE?"

$$F^* = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)/df_F}{MSE}} = \frac{MSR}{MSE} \sim F(1, n - 2)$$

In SLR, the global test (the significance of a model test), the ANVOA F test or the T test for the linear impact are equivalent.

The F test for Lack of Fit

- ▶ Formal test for determining whether a specific type of regression function adequately fits the data.
- ▶ Assumptions (usual):
 - observations $Y|X$ are
 1. i.i.d.
 2. normally distributed
 3. same variance σ^2
 - ▶ Requires: repeat observations at one or more X levels (called replicates)

The Bank example

- ▶ **11** similar branches of a bank offered gifts for setting up money market accounts
- ▶ Minimum initial deposits were specific to qualify for the gift
- ▶ Value of gift was proportional to the specified minimum deposit
- ▶ Interested in: relationship between specified minimum deposit and number of new accounts opened

Notation

Minimum deposit	Number of new accounts
75	28
75	42
100	112
100	136
125	160
125	150
150	152
175	156
175	124
200	124
200	104

- Y_{11} denotes the first measurement (28) made at the first X level (75).
- Y_{12} denotes the second measurement (42) made at the first X level (75).
- \bar{Y}_1 denotes the average $(\frac{28+42}{2} = 35)$ of all y values at the first X level (75).
- \hat{Y}_{11} denotes the predicted response ($b_0 + b_1X = 87.5$) for the first measurement at the first X level (75).
- \hat{Y}_{12} denotes the predicted response ($b_0 + b_1X = 87.5$) for the second measurement at the first X level (75).
- \hat{Y}_{ij} denotes the predicted response for the jth measurement at the ith X level.
 $\hat{Y}_{ij} = b_0 + b_1X_i = \hat{Y}_i$ is the same for all j at the same X_i value.
- \bar{Y}_i denotes the average of all y values at the ith X level, or the group mean.
- \bar{Y} denotes the average of all y values at all X levels, or the grand mean.
- C denotes the number of distinct X levels.
 $c = 6, X_1 = 75, X_2 = 100, X_3 = 125, X_4 = 150, X_5 = 175, X_6 = 200$
- Most X_i has two replicates except X_4

$$X_4 = 150, Y_4 = 152 = \bar{Y}_4 = 152, \hat{Y}_4 = 51 + 0.5(150) = 126$$

$$X_3 = 125, Y_{31} = 160, Y_{32} = 150, \bar{Y}_3 = 155, \hat{Y}_3 = 51 + 0.5(125) = 114$$

The F test of ANOVA for $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$

Q: Does X have significant linear impact on Y?

Source of Variation	SS	df	MS	F	Conclusion
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$	$MSR / MSE \sim F(1, n-2)$	Reject H_0 means X has significant Linear impact on Y
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n - 2$	$MSE = \frac{SSE}{n-2}$		
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n - 1$			

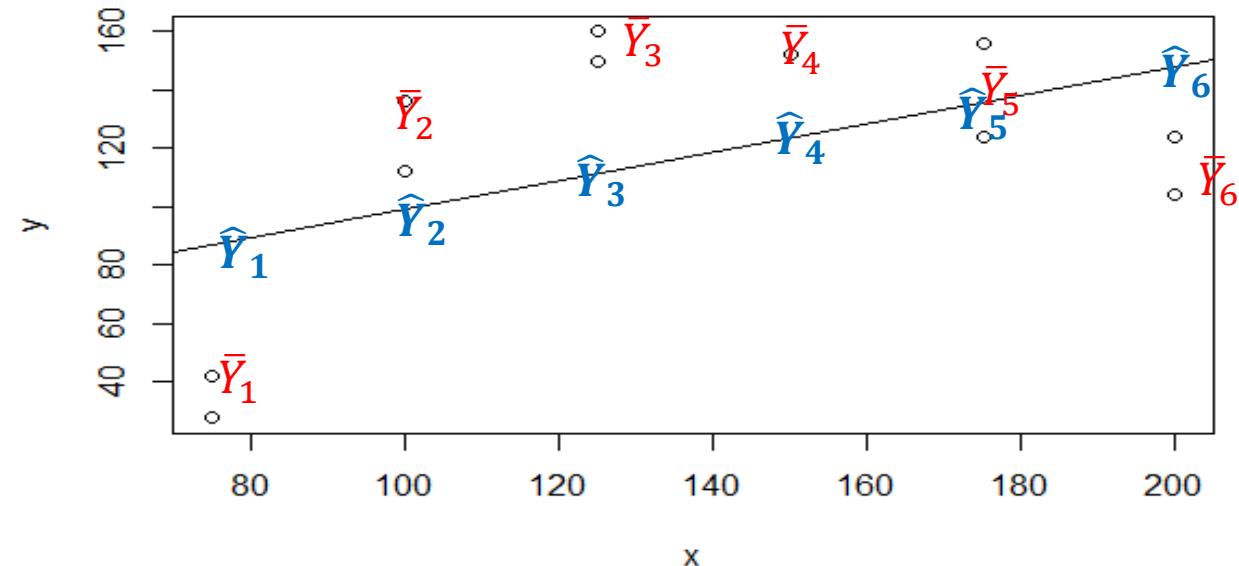
The bank example

Response: y

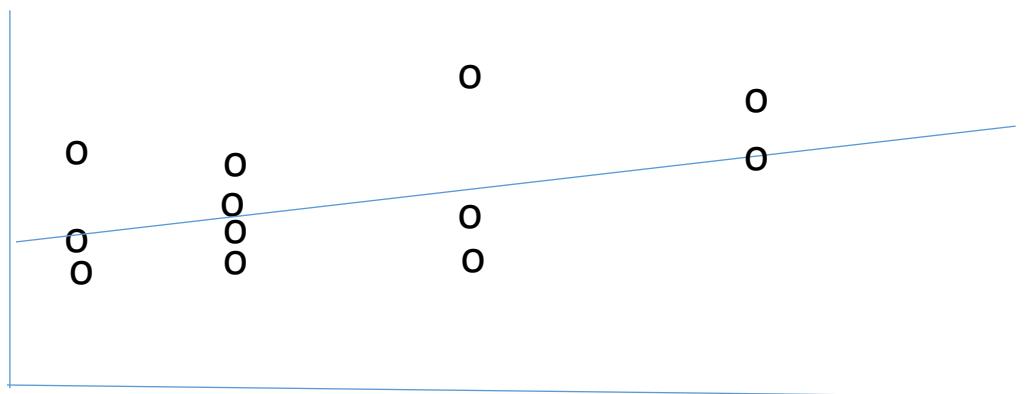
	df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	5141.3	5141.3	3.1389	0.1102
Residuals	9	14741.6	1638.0		

There is no evidence to reject $\beta_1 = 0$, X seems to have no significant linear impact on Y.

The lack-of-fit property



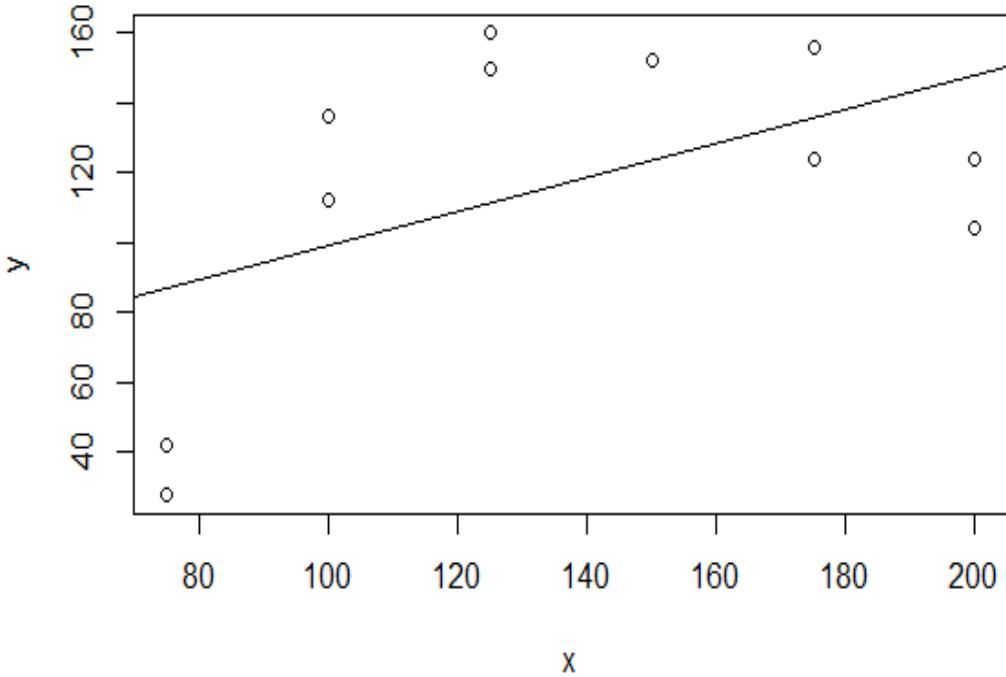
- The linear line is rather flat. But there seems to be more issue found in the scatter plot
- The predictor value $\hat{Y}_i = b_0 + b_1 X_i$ is systematically off from the actual sample mean \bar{Y}_i . Such model has a poor fit on the data, or **lack of fit**.
- This linear model shows X has little impact on Y, and has a lack of fit.



- This model demonstrates X has little impact on Y, but **doesn't have a lack of fit issue**.

The lack of fit test $H_0: E\{Y\} (= \mu) = \beta_0 + \beta_1 X$, $H_a: E\{Y\} (= \mu) \neq \beta_0 + \beta_1 X$

Q: Does the linear model fit the data, or is the predicted mean response value the same as the actual mean response value?



Reduced model (H_0) : $\hat{Y}_{ij} = \beta_0 + \beta_1 X_i$

$$SSE(\text{Reduced}) = \sum \sum (Y_{ij} - \hat{Y}_i)^2 = SSE, \quad dfE_{\text{Reduced}} = n - 2$$

Full model (H_a): $\hat{Y}_{ij} = \mu_i + \varepsilon_{ij}$

Specifically,

$\hat{Y}_{1j} = \bar{Y}_1$, the residual $= Y_{1j} - \bar{Y}_1$ for $j = 1$ or 2

$\hat{Y}_{2j} = \bar{Y}_2$, the residual $= Y_{2j} - \bar{Y}_2$ for $j = 1$ or 2

$\hat{Y}_{3j} = \bar{Y}_3$, the residual $= Y_{3j} - \bar{Y}_3$ for $j = 1$ or 2

$\hat{Y}_{4j} = \bar{Y}_4$, the residual $= Y_{4j} - \bar{Y}_4 = 0$ for no replicate

$\hat{Y}_{5j} = \bar{Y}_5$, the residual $= Y_{5j} - \bar{Y}_5$ for $j = 1$ or 2

$\hat{Y}_{6j} = \bar{Y}_6$, the residual $= Y_{6j} - \bar{Y}_6$ for $j = 1$ or 2

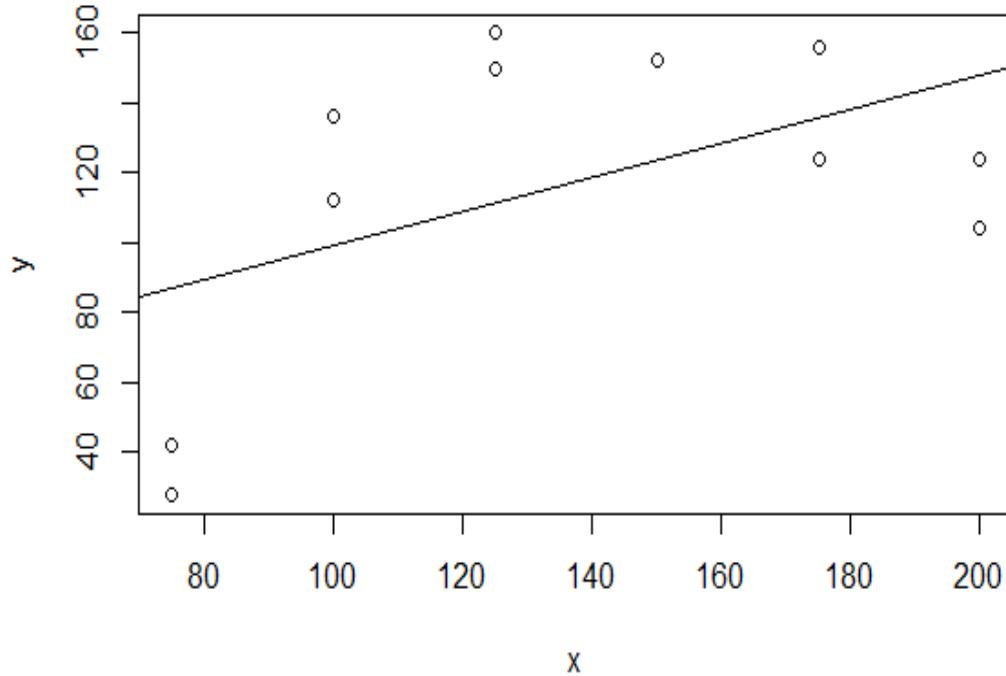
$SSE(\text{Full}) = \text{Total residuals summing up } i \text{ and } j$

$$= \sum \sum (Y_{ij} - \bar{Y}_i)^2,$$

$$dfE_{\text{full}} = n - 1 + \dots + (n - 1) = n - 6 = n - c$$

The lack of fit test $H_0: E\{Y\} (= \mu) = \beta_0 + \beta_1 X$, $H_a: E\{Y\} (= \mu) \neq \beta_0 + \beta_1 X$

Q: Does the linear model fit the data, or is the predicted mean response value the same as the actual mean response value?



Reduced model (H_0) : $\hat{Y}_{ij} = \beta_0 + \beta_1 X_i$

$$SSE(\text{Reduced}) = \sum \sum (Y_{ij} - \hat{Y}_i)^2 = SSE, \quad df_R = n - 2$$

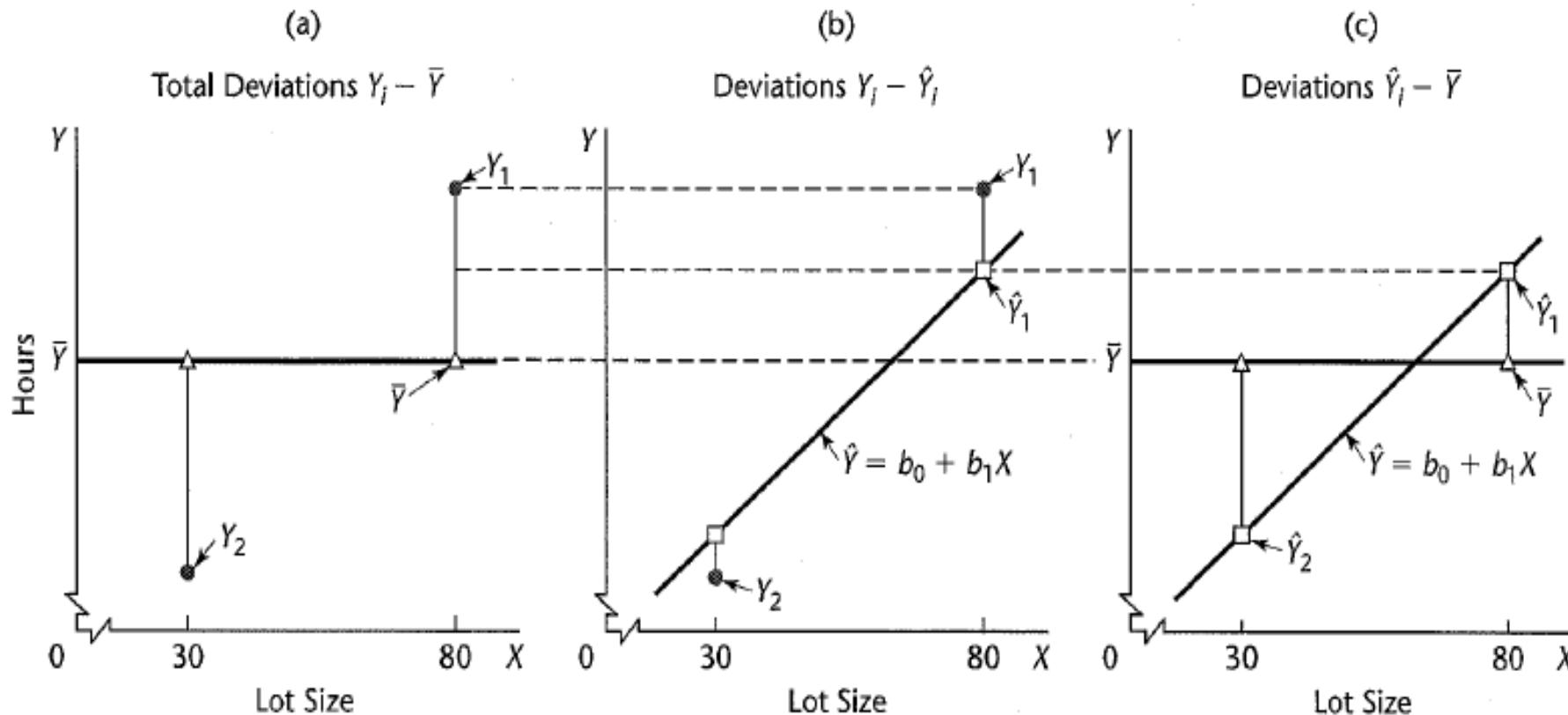
Full model (H_a): $Y_{ij} = \mu_i + \varepsilon_{ij}$

$$SSE(\text{Full}) = \sum \sum (Y_{ij} - \bar{Y}_i)^2, \quad df_F = n - c$$

$$F = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{SSE(F)/df_F} = \frac{\frac{SSE(R) - SSE(F)}{n - 2 - (n - c)}}{\frac{SSE(F)}{n - c}}$$

$$\sim F(c - 2, n - c)$$

Partition the variances



$$\Sigma(Y_i - \bar{Y})^2 = \Sigma(Y_i - \hat{Y}_i)^2 + \Sigma(\hat{Y}_i - \bar{Y})^2$$

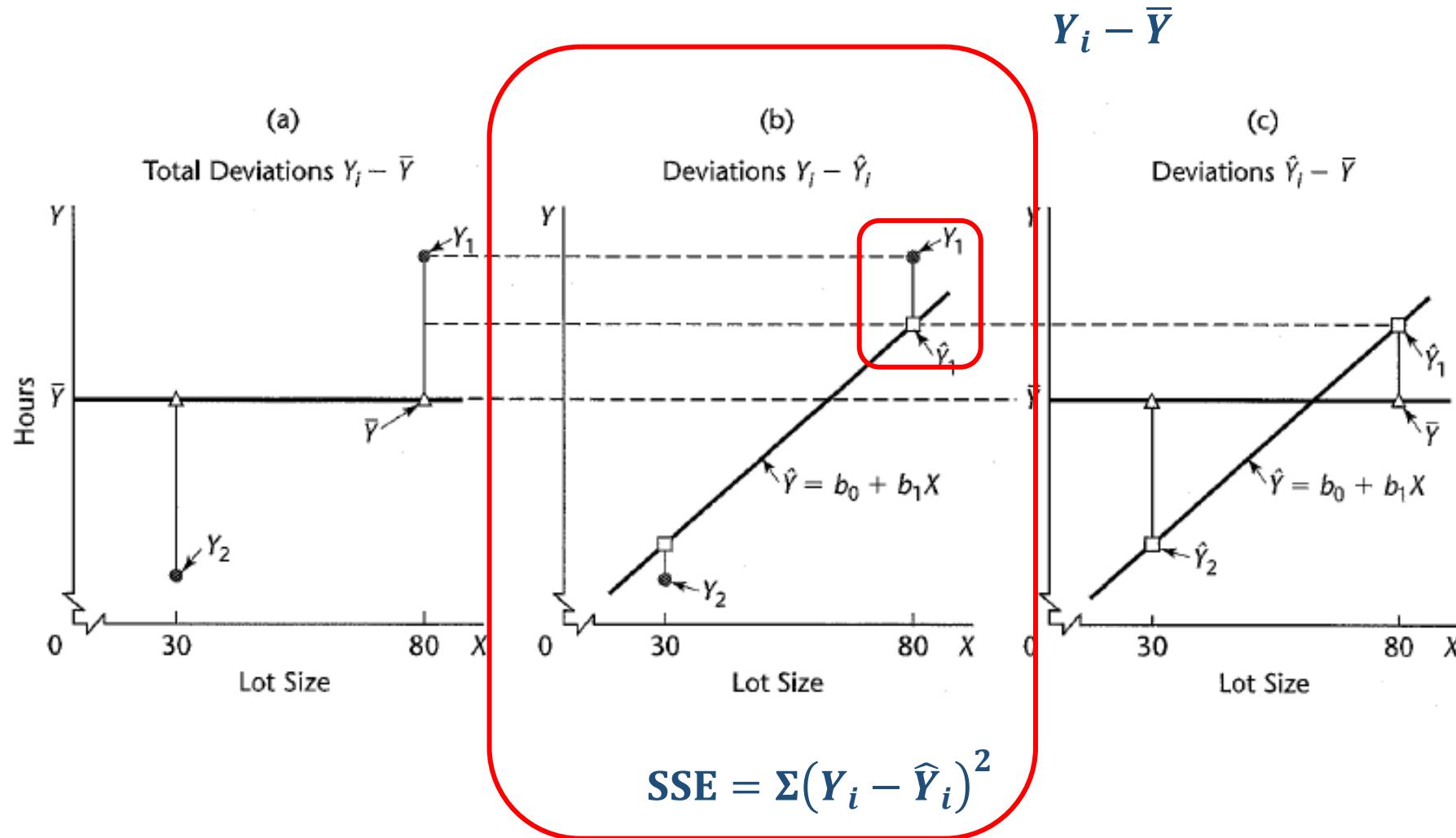
$$SSTO = SSE + SSR$$

"Total sum of squares"

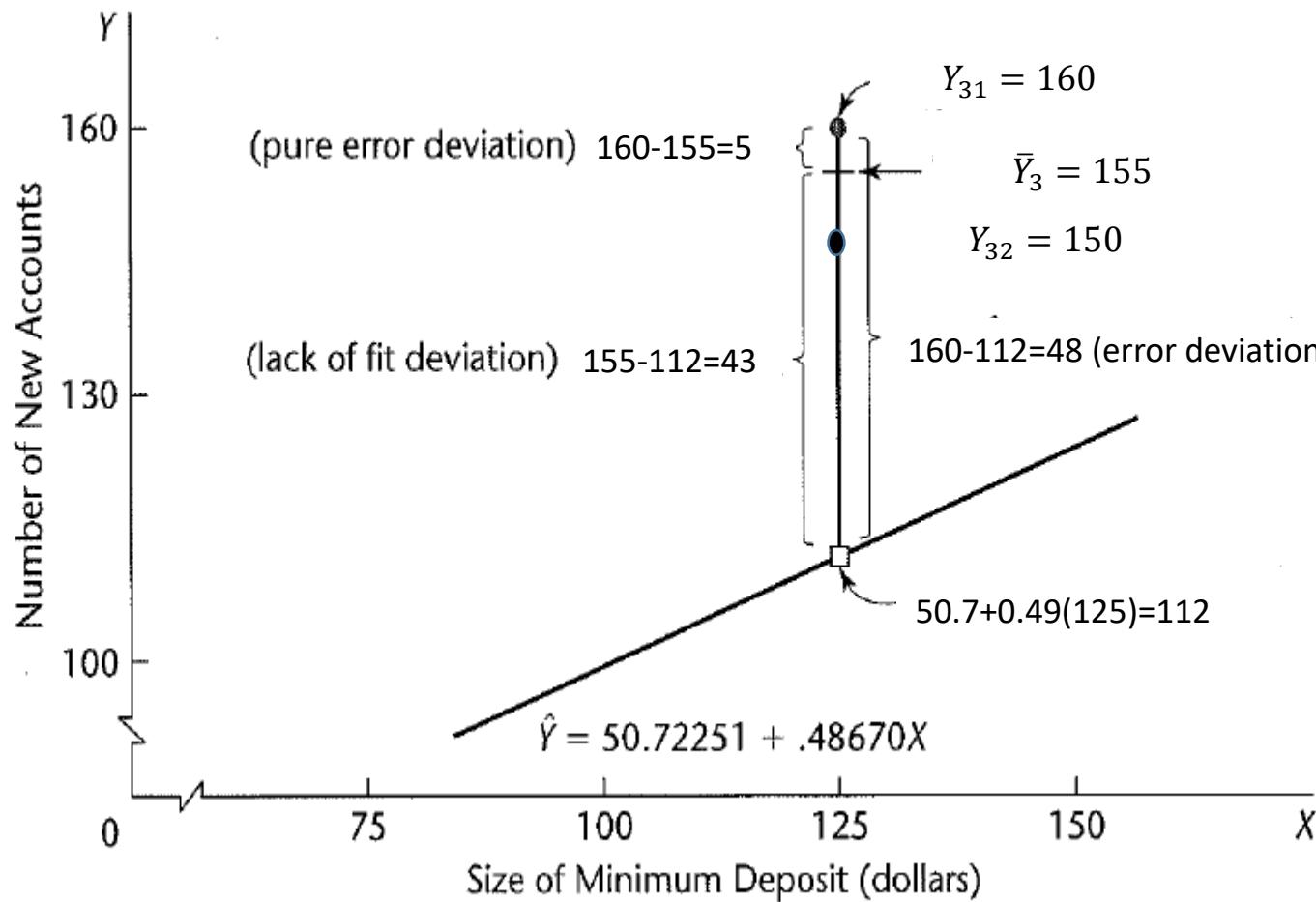
"error sum of squares"

"regression sum of squares"

Partition the residual errors for lack of fit



Partition the residual errors for lack of fit, SSPE and SSLF



$(Y_{ij} - \hat{Y}_i)$ measures the total error deviation in one observation.

$(Y_{ij} - \bar{Y}_i)$ measure the pure error deviation, which is the randomness result from the data, not from the choice of model.

$(\bar{Y}_i - \hat{Y}_{ij})$ measure the lack of fit deviation, which is the error result from the choice of model and could be improved with a better model.

Do this for every data point, and sum, we have

$$\sum \sum (Y_{ij} - \hat{Y}_{ij})^2 = \sum \sum (Y_{ij} - \bar{Y}_j)^2 + \sum \sum (\bar{Y}_j - \hat{Y}_{ij})^2$$

$$SSE = SSPE + SSLF$$

Partition the previous ANOVA table on the SSE term further into SSLF and SSPE

Source of Variation	SS	df	MS	F	Conclusion
Regression	$SSR = \Sigma \sum (\hat{Y}_{ij} - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$	$MSR / MSE \sim F(1, n-2)$	Reject H_0 means X has significant Linear impact on Y
Error	$SSE = \Sigma \sum (Y_{ij} - \hat{Y}_{ij})^2$	$n - 2$	$MSE = \frac{SSE}{n-2}$		
Lack of fit (in Error)	$SSLF = \Sigma \sum (\bar{Y}_i - \hat{Y}_{ij})^2$	$c - 2$	$MSLF = \frac{SSLF}{c-2}$	$MSLF / MSPE \sim F(c-2, n-c)$	Reject H_0 means the current model does not fit the data
Pure error (in Error)	$SSPE = \Sigma \sum (Y_{ij} - \bar{Y}_i)^2$	$n - c$	$MSPE = \frac{SSPE}{n-c}$		
Total	$SSTO = \Sigma \sum (Y_{ij} - \bar{Y})^2$	$n - 1$			

Example 1, the R output on the linear impact, or the model significance test

Source of Variation	SS	df	MS	F	Conclusion
Regression	5141	1	5141	?	?
Error	14742	11-2=9	1638		
Lack of fit(in Error)	13594	6-2=4	3398.5		
Pure error(in Error)	1148	11-6=5	229.6		
Total	19883	10			

```
bankR.mod<-lm(y~x, bank)
anova(bankR.mod)
```

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	5141.3	5141.3	3.1389	0.1102
Residuals	9	14741.6	1638.0		

Example 2, the R output on the lack of fit test

Source of Variation	ss	df	MS	F	Conclusion
Regression	5141	1	5141	3.14 (p=0.11)	X does not have significant linear impact on Y
Error	14742	n-2=11-2=9	1638		
Lack of fit(in Error)	13594	c-2=6-2=4	3398.5	?	?
Pure error(in Error)	1148	N-c=11-6=5	229.6		
Total	19883	10			

Build the reduced model under $H_0: \hat{Y} = \beta_0 + \beta_1 X$

```
bankR.mod<-lm(y~x, bank)
anova(bankR.mod)
```

Build the full model under $H_a: \hat{Y} = \mu$

```
bankF.mod<-lm(y~as.factor(x), bank)
anova(bankR.mod, bankF.mod)
```

```
Response: y
          Df  Sum Sq Mean Sq F value Pr(>F)
x           1  5141.3  5141.3  3.1389 0.1102
Residuals   9 14741.6  1638.0
```

```
Model 1: y ~ x
Model 2: y ~ as.factor(x)
  Res.Df  RSS  Df Sum of Sq    F    Pr(>F)
1      9 14742
2      5 1148  4     13594 14.801 0.005594 ***

```

$$Fs = MSLF/MSPE = \frac{13594}{4} \div \frac{1148}{5} = 14.801, \text{ this model has a lack of fit issue.}$$

The lack of fit test is not valid without replicates. But we can manually create replicates by grouping.

- $SSPE = \sum \sum (Y_{ij} - \bar{Y}_{ij})^2 = 0$

size	hour
20	113
30	121
40	160
50	221
60	224
70	361
80	399
90	376
100	353
110	435
120	546

- Solution: grouping

```

g<-c(30,30,30,60,60,60,90,90,90,115)
tolucanr$g<-g
tolucanrgR.mod<-lm(y~g, data=tolucanr)
tolucanrgF.mod<-lm(y~factor(g), data=tolucanr)
summary(tolucanrgR.mod)
anova(tolucanrgR.mod)
anova(tolucanrgR.mod, tolucanrgF.mod)

```

Model 1: $y \sim x$

Model 2: $y \sim \text{factor}(x)$

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	16602	9			
2	16602	0	0	9	

Model 1: $y \sim g$

Model 2: $y \sim \text{factor}(g)$

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	21775	9			
2	21276	7	498.74	2	0.082 0.9221

Diagnostics in SLR

“Things to check behind those significant T-test and F test”

Diagnostics – Methods to check whether our model is reasonable for our data and representative of the system that we are studying.

Some diagnostics check the *assumptions* of our model. Other diagnostics check the *influence* of different data points.

Remedies – Analytic strategies used to fix problems identified by the diagnostics.

Why do we need to check the model?

All models are wrong. Some models are useful.

— George Box

The goal of building a model is to:

- *learn something* about the real world
- *predict outcomes* in the real world

To use a model successfully, we need to know its limitations:

- Does it adequately describe the functional relationship of interest?
- Is there reason to worry that inferences about the parameters might be flawed?
- Is the error distribution appropriate?

What do we need to check?

- Is the *functional form* of the model appropriate?
- Do any of the data points have a *disproportionate influence* on the parameter estimates?
- Are there *outliers*?
- Are the residuals in the data consistent with our model for random error:
 - Errors are *independent*,
 - Errors all have the *same variance*, σ^2
 - Errors are *normally distributed*.

How do we check?

- Diagnostic plots and tests on **residuals (major content in this topic)**
 - Plot of residual vs predictor variable.
 - Plot of residual against fitted value
 - Plot of residual against time or other sequence
 - Plot of residual against omitted predictor variable.
 - Normal probability plot of residual
 - Brown-Forsythe test for non-constant variance (**heteroscedasticity**)
 - **Shapiro test for non-Normality**
- Diagnostic plots and tests **on dependent and independent variables**

Diagnostics based on residuals

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

The random errors ε_i are *independent*, *normal*, and should have *constant variance*.

$$\varepsilon \sim Normal(0, \sigma)$$

The residuals are computed from sample to simulate ε_i ,

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

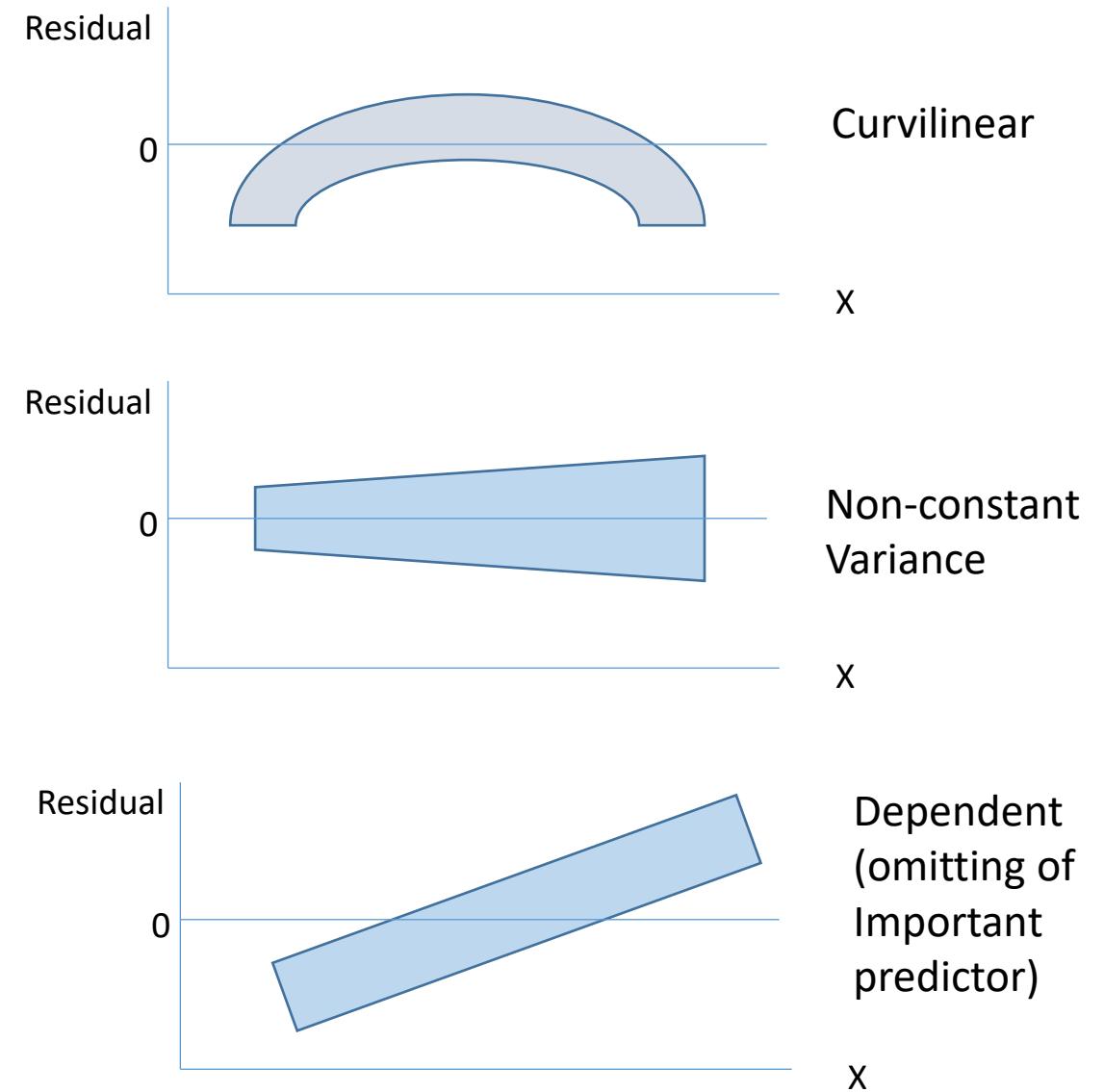
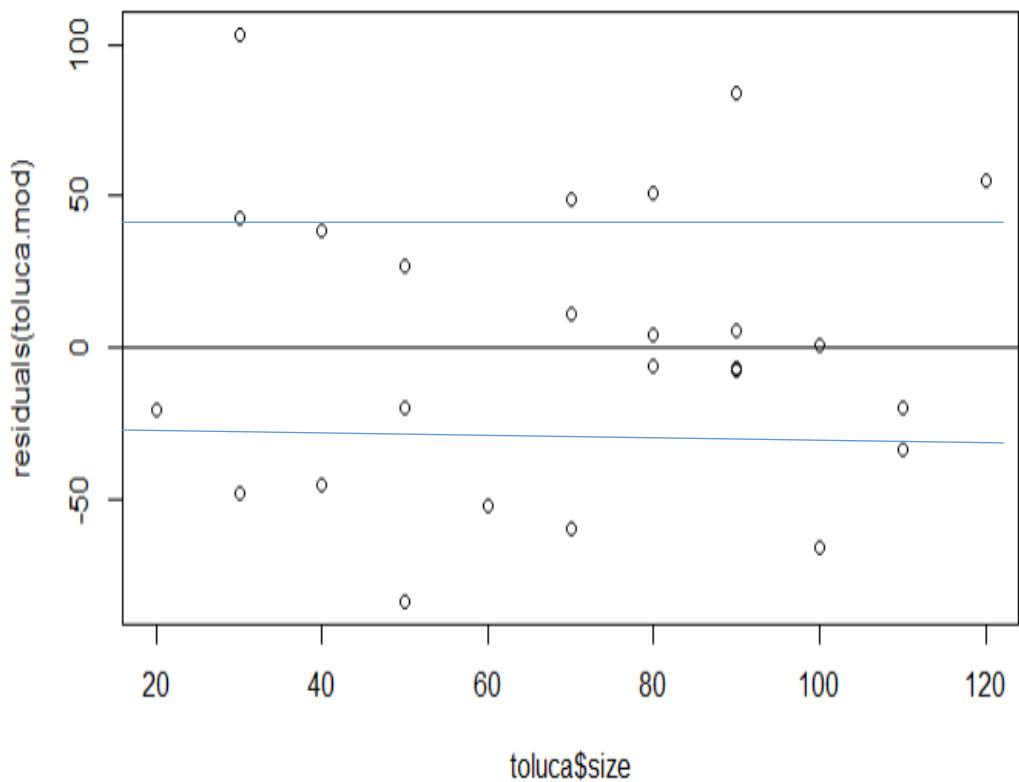
If the model fits, e_i should reflects the properties assumed for ε_i .

Questions Addressed by Diagnostics from **Residuals plot**

- Is the functional form appropriate (in SLR, that means linear)? → **is there no curvature pattern?**
- Does the variance depend on X ? → **is there increase or decrease in average magnitude with the fitted values?**
- Are the errors normal? → **The normal plot of residual is straight?**
- Are there outliers? → **is there relatively large residual?**
- Are the errors appearing independent? → **Any patterns in the residual plot?**

Note that $e_i = Y_i - \hat{Y}_i$ are not independent since each \hat{Y}_i is computed with the same b_0 and b_1 . However, when the sample size (n) is much larger than the number of parameters (p), We can ignore the minor dependence.

Prototype of good or problematic residual plot

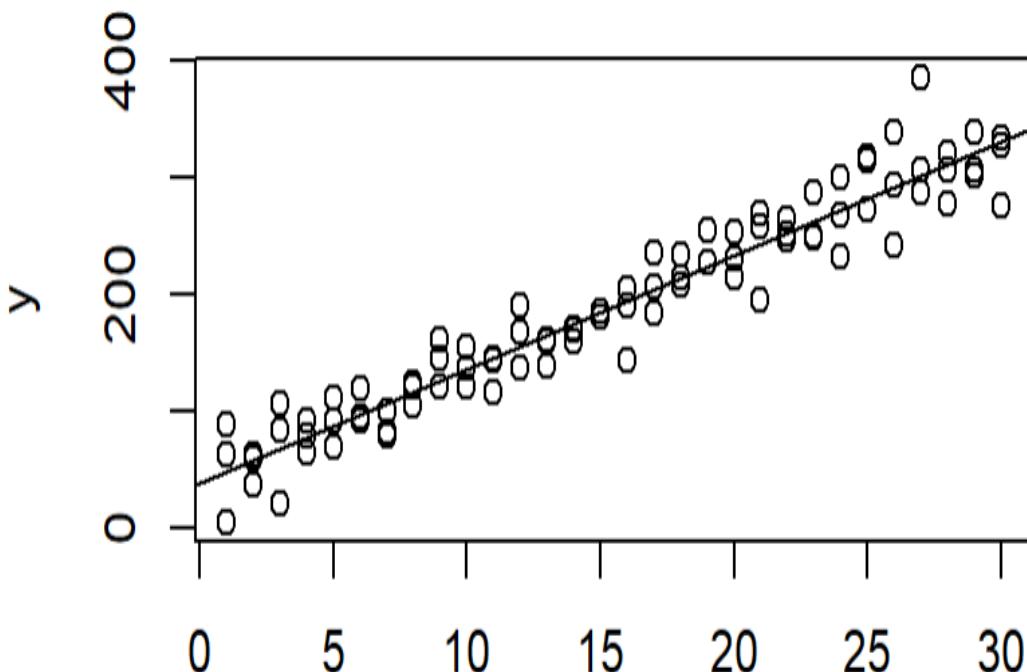


Some example of violations and diagnostics from residual plot

- linear relationship
- constant variance
- normal errors
- Independence
- outliers

A case with perfect linear relationship

$$Y = 10X + 30 + N(0,25)$$



```
x<-rep(seq(1:30),3)
y<-10*x+30+rnorm(90, 0, 25)
SLRdata<-data.frame(x,y)
SLRdataRM<-lm(y~x, SLRdata)
plot(x,y)
abline(SLRdataRM)
summary(SLRdataRM)
anova(SLRdataRM)
```

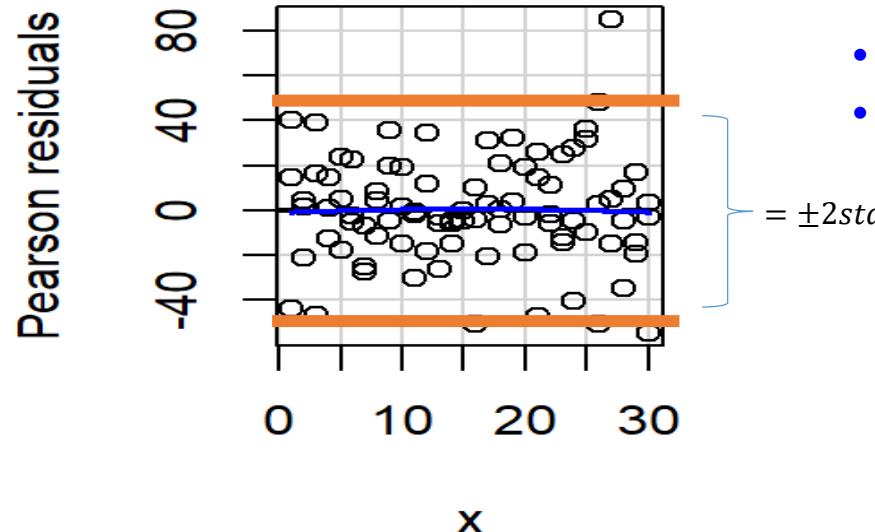
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	38.0032	5.1771	7.341	1.01e-10 ***
x	9.7570	0.2916	33.458	< 2e-16 ***
<hr/>				
Signif. codes:	0	***	0.001	**
			0.01	*
			0.05	.
			0.1	'
			1	'

Residual standard error: 23.95 on 88 degrees of freedom
 Multiple R-squared: 0.9271, Adjusted R-squared: 0.9263
 F-statistic: 1119 on 1 and 88 DF, p-value: < 2.2e-16

A case with perfect linear relationship

$$Y = 10X + 30 + N(0,25)$$



- The residual plot shows the residuals scatter evenly around a flat line of 0,
- The variance is constant across X, no obvious pattern.

Analysis of Variance Table

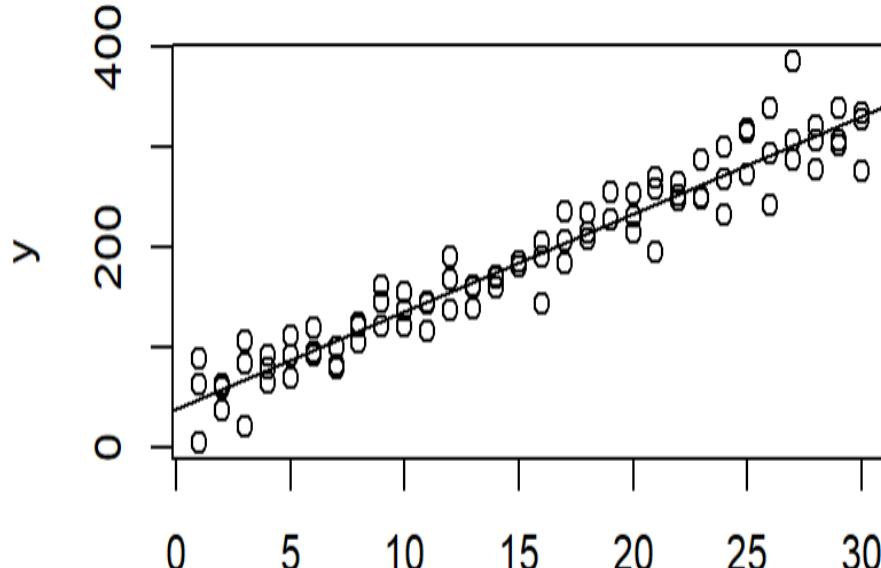
Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	641876	641876	1119.4	< 2.2e-16 ***
Residuals	88	50458	573		

- $\sqrt{MSE} = s\{\text{residual}\} = \sqrt{573} = 24$ which is a good estimate of the actual $s\{\text{random error}\}=25$. The model is **efficient**.

A case with perfect linear relationship

$$Y = 10X + 30 + N(0,25)$$



- From the scatter plot and the residual plot, we can observe that the mean response prediction and the actual mean For each X level are close. Hence no lack-of-fit is expected.

Analysis of variance Table

Model 1: $y \sim x$

Model 2: $y \sim \text{as.factor}(x)$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	88	50458				
2	60	41055	28	9403.5	0.4908	0.9793

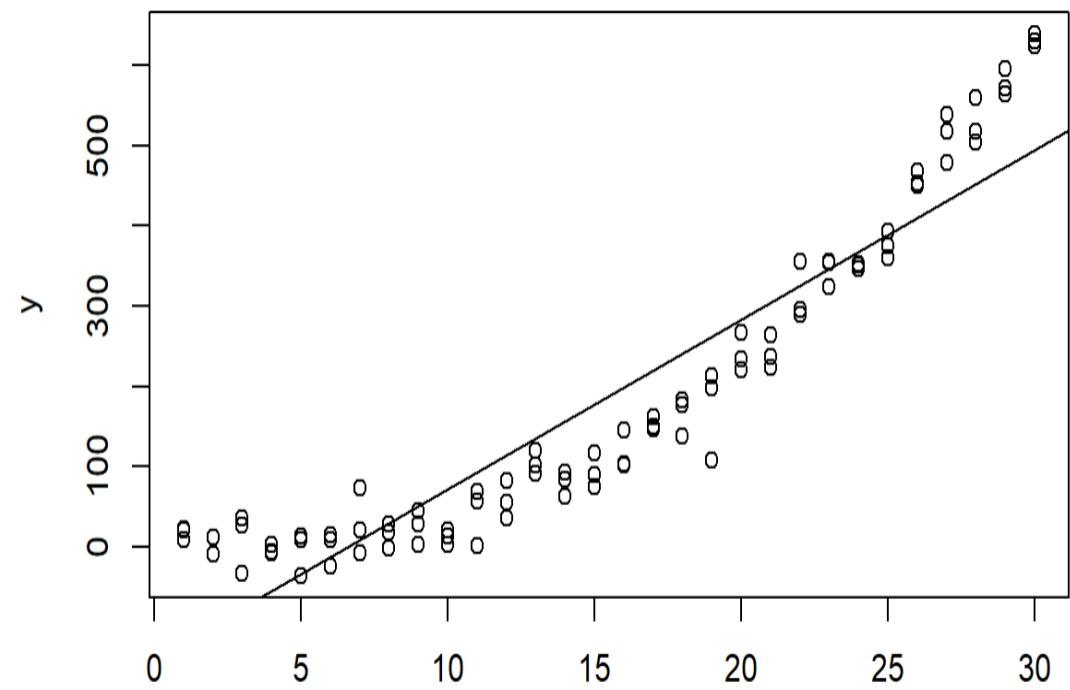
$$\bullet \quad F_S = \frac{MSLF}{MSPE} = \frac{9403/28}{41055/60} = \frac{336}{684} = 0.491$$

$$\bullet \quad \sqrt{MSPE} = s\{\text{pure error}\} = \sqrt{684} = 26.15 \text{ which is also a good estimate of the actual } s\{\text{random error}\} = 25$$

Make a dataset that we know is quadratic, not linear.

$$Y = 30 - 10X + X^2 + N(0, 25)$$

```
x<-rep(seq(1:30),3)
y<-x^2-10*x+30+rnorm(90, 0, 25)
n1<-data.frame(x,y)
nonlinearRM<-lm(y~x, n1)
plot(x,y)
abline(nonlinearRM)
summary(nonlinearRM)
anova(nonlinearRM)
```



Coefficients:

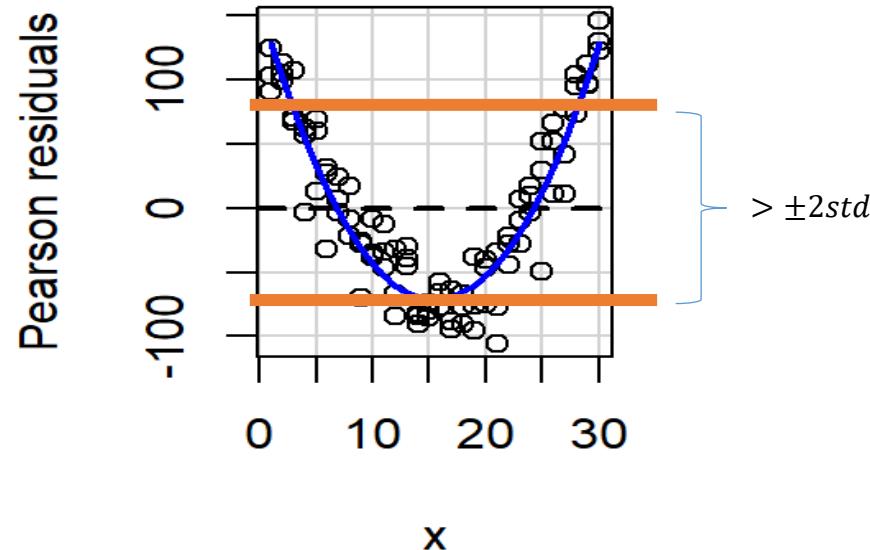
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-138.6100	15.4200	-8.989	4.32e-14 ***
x	21.0999	0.8686	24.292	< 2e-16 ***

Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’
	0.1 ‘ ’	1		

Residual standard error: 71.32 on 88 degrees of freedom
 Multiple R-squared: 0.8702, Adjusted R-squared: 0.8688
 F-statistic: 590.1 on 1 and 88 DF, p-value: < 2.2e-16

- The summary shows that the linear impact is significant, the R-square is 88.26%, and the model is significant.
- Is the model good?

$$Y = 30 - 10X + X^2 + N(0, 25)$$

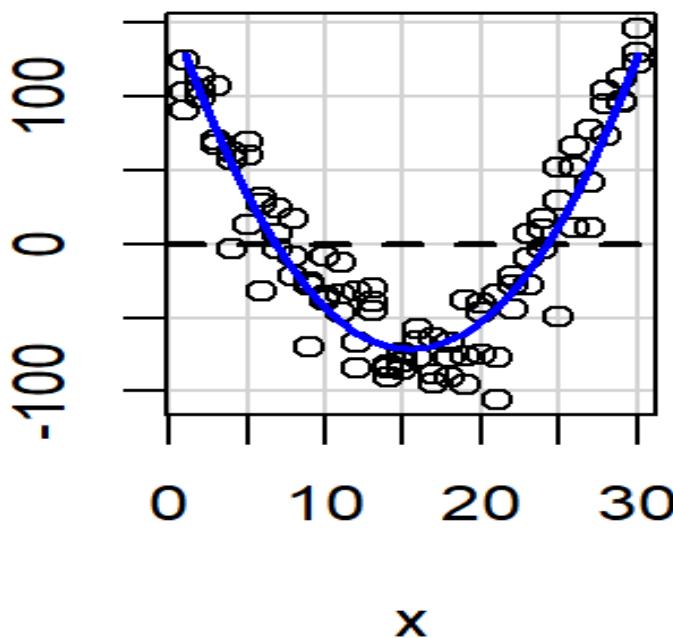
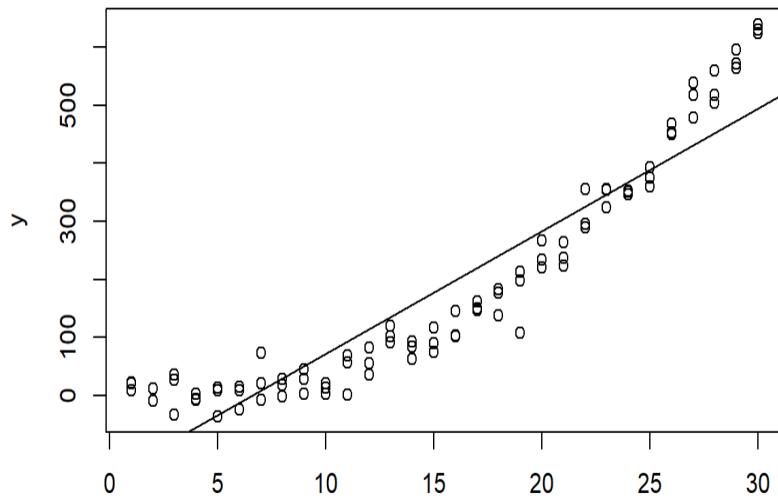


- The residual plot shows that the residuals do not scatter evenly around the flat line of 0.
- The variances are not constant across X .

Response: y						
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
x	1	3001798	3001798	590.11	< 2.2e-16	***
Residuals	88	447641	5087			

- $\sqrt{MSE} = s\{\text{residual}\} = \sqrt{5087} = 71$, while the actual $S\{\text{random error}\}=25$.
- The actual deviation is overestimated. The model is **inefficient**.

$$Y = 30 - 10X + X^2 + N(0, 25)$$



- From the scatter plot and the residual plot, we can observe that the mean response prediction and the actual mean For each X level are always off. Hence lack-of-fit is expected.

The lack of fit test

Model 1: $y \sim x$

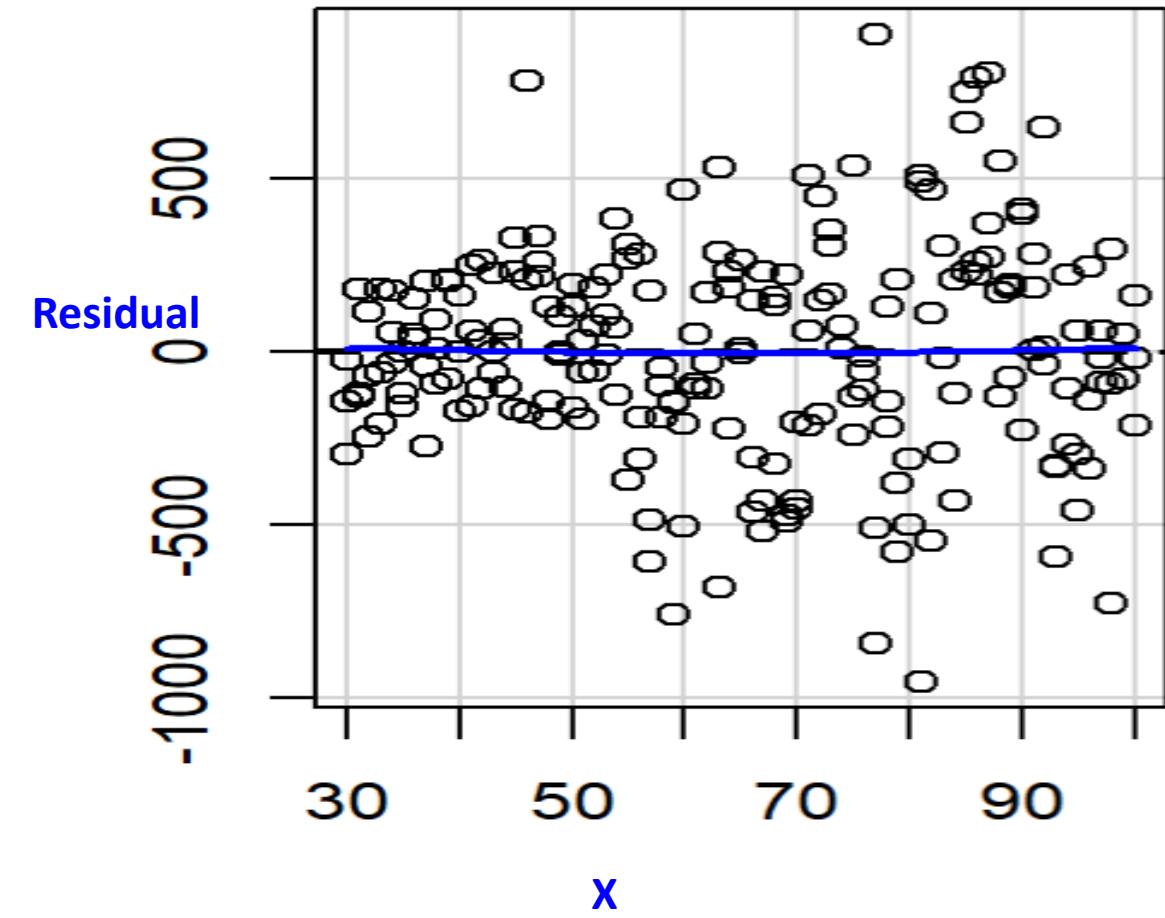
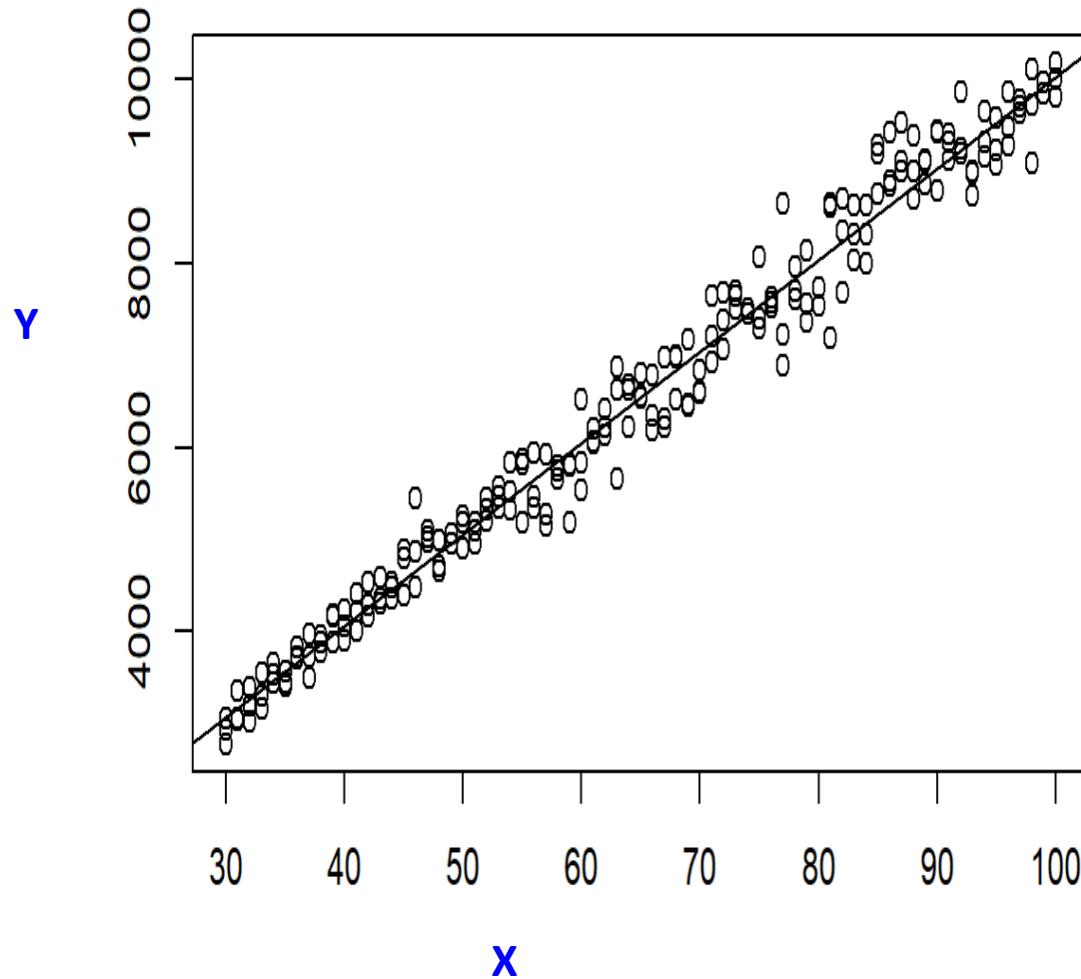
Model 2: $y \sim \text{as.factor}(x)$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	88	447641				
2	60	35301	28	412340	25.03	< 2.2e-16 ***

- $F_S = \frac{MSLF}{MSPE} = \frac{412340/28}{35301/60} = \frac{14726}{588} = 25.03$
- $\sqrt{MSPE} = s\{\text{pure error}\} = \sqrt{588} = 24.25$, close to 25.
- This is because the Pure error is the variation among Y in each X value, $\sigma\{Y\} = \sigma\{30 - 10X + X^2 + \varepsilon\} = \sigma\{\varepsilon\} = 25$

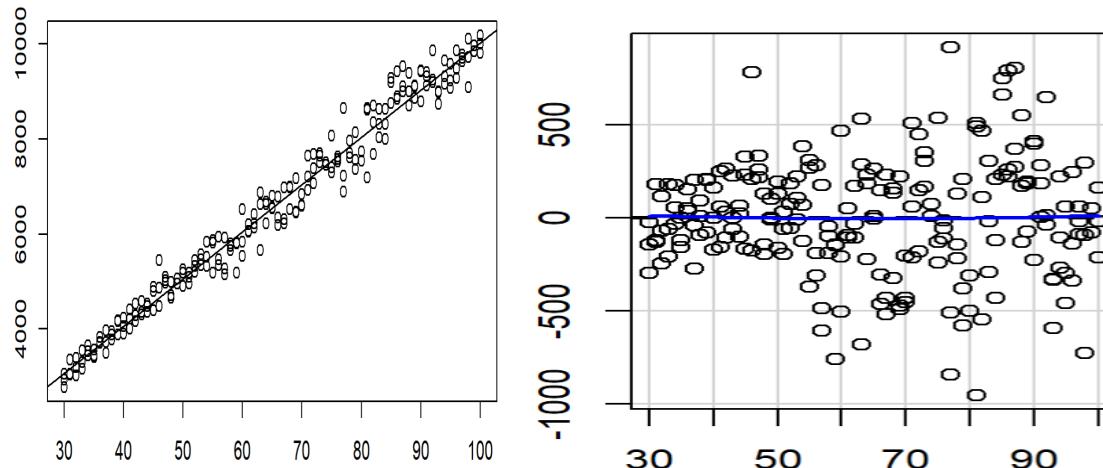
A case with non-constant variance (heteroscedasticity)

$Y = 30 + 100X + N(0, 5x)$, X ranges from 30 to 100



Fit a SLR on the heteroscedasticity data

$$Y = 30 + 100X + N(0, 5x), \text{ X ranges from 30 to 100}$$



Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	77.614	82.189	0.944	0.346
x	98.787	1.206	81.918	<2e-16 ***

Residual standard error: 360.7 on 211 degrees of freedom
 Multiple R-squared: 0.9695, Adjusted R-squared: 0.9694
 F-statistic: 6711 on 1 and 211 DF, p-value: < 2.2e-16

Response: y

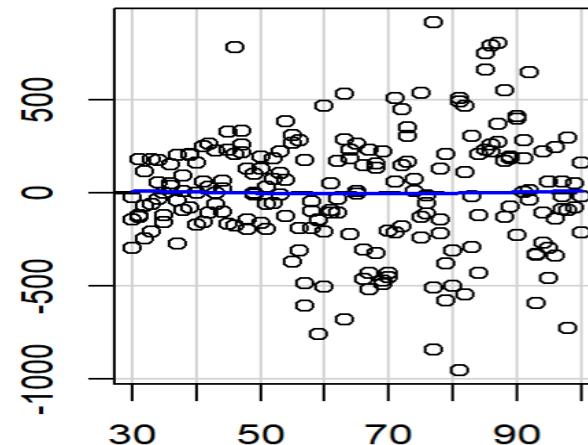
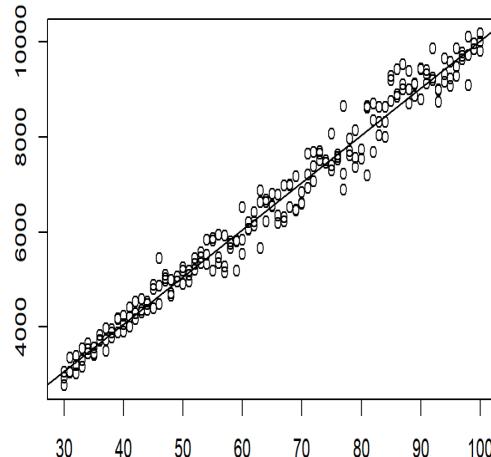
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	873033514	873033514	6710.5	< 2.2e-16 ***
Residuals	211	27450926	130099		

- The actual random error deviates from 150 to 500, or $650/2=325$ on average.

- The linear impact (slope) is estimated well: 98.8 ± 1.2
- The intercept is not estimated well: 77.6 ± 82.2 , with a P-value of 0.346.
- The residual standard error is 360.7, this is a good estimate for the **average** random error standard deviation. But not a good estimate for the actual random error standard deviation which has a changing value.

The lack of fit test on the heteroscedasticity data

$$Y = 30 + 100X + N(0, 5x), \text{ X ranges from 30 to 100}$$



Model 1: $y \sim x$

Model 2: $y \sim \text{as.factor}(x)$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	211	27450926				
2	142	19030337	69	8420590	0.9106	0.6641

- The actual random error deviates from 150 to 500, or $650/2=325$ on average.
- $\sqrt{MSPE} = s\{\text{pure error}\} = \sqrt{19030337/142} = 366.08$
- Similar as the SLR model, this is a good estimate for the **average** random error standard deviation.

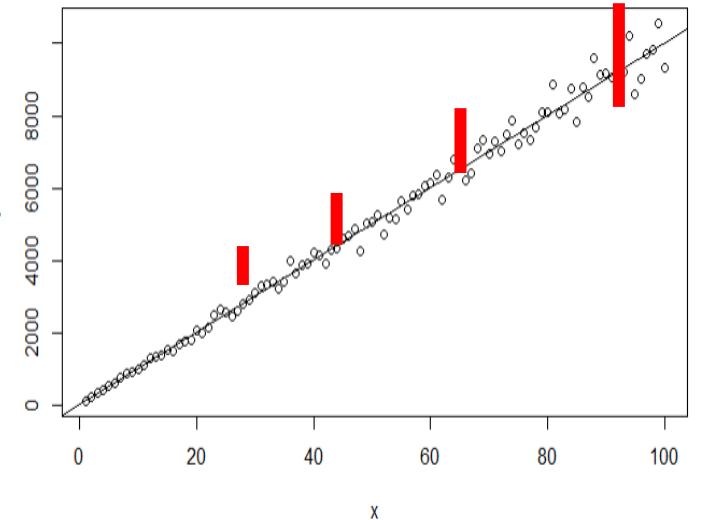
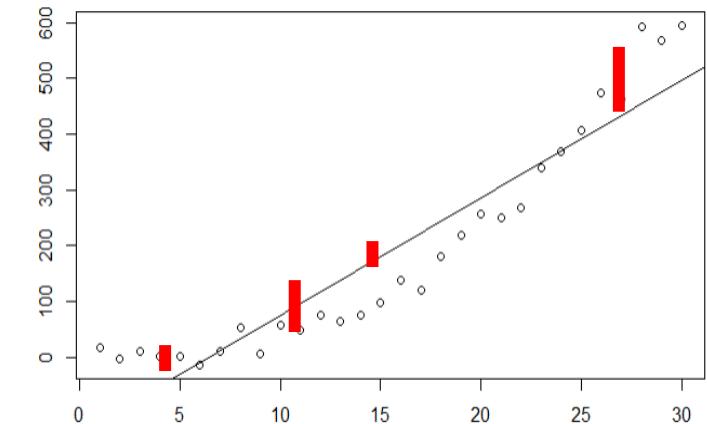
Compare the consequence with the non-linear and the non-constant variance

Systematic deviations from the functional form of the model and non-constant variance are both examples of *model misspecification*, and are both magnified in residual plots, they both cause prediction problems. Specifically,

- Systematic deviation from the linear form
 1. Different systematic biases at different values of X (i.e., $b_0 + b_1X$ is no good).
 2. A higher overall estimate of error variance (the model is inefficient).

- Non-constant variance
 1. Does **not** cause bias in the point estimates. (i.e., $b_0 + b_1X$ could still be good) ➤
 - 2. But it does** invalidate estimates for the standard errors of the parameters.

For example, $s\{b_0\}$ is large.



Statistical test for heteroscedasticity

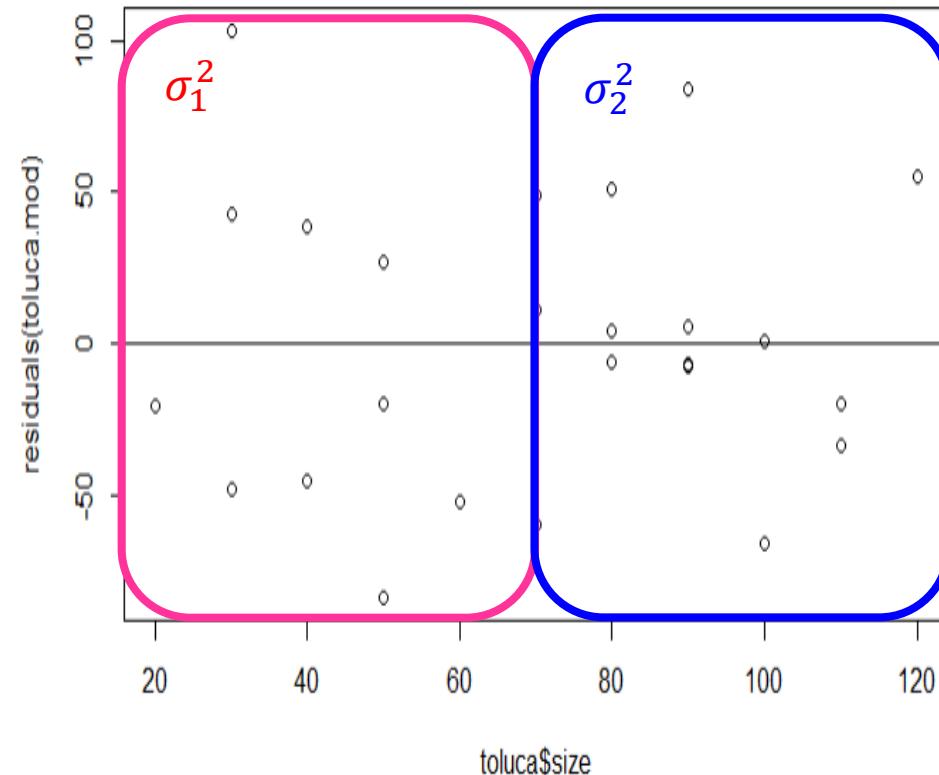
H₀: residuals have constant variance H_a: residuals have non – constant variances

1. Brown Forsythe (BF) test
 - Does not depend on normality of the error terms.
 - BF test is usually used for case with categorical predictors (X).
 - We need to adjust the continuous prediction variable X in the LR **into two or n groups** or categories.
 - Convenient for SLR, but not for MLR.
2. Breusch-Pagan (BP) test
 - Assumes that the error terms are independent and normal and that the variance of the error terms is related to X.
 - No need to group X so more convenient for MLR.

Brown-Forsythe test (2 groups)

```
library(ALSM)
g<-rep(1,25)
g[toluca$size<=70]=0      #form two groups
bftest(lm(hour~size, toluca),g)
```

	t.value	P.value	alpha	df
[1,]	1.316482	0.2009812	0.05	23



Brown-Forsythe test (n groups)

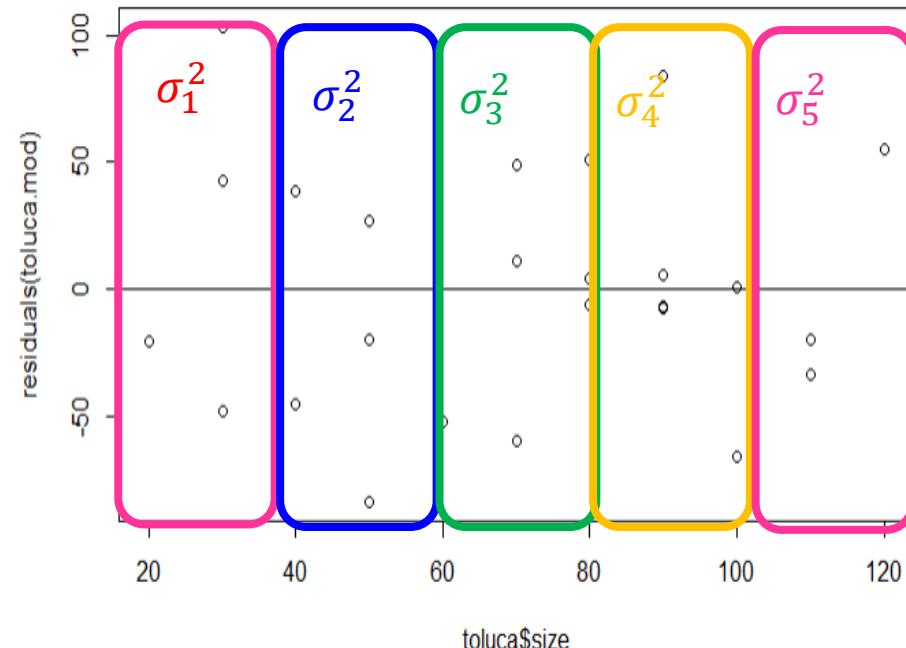
```
library(onewaytests)
toluca$group<-cut(toluca$size, 5) #form five groups
toluca$residual<-toluca.mod$residuals
bf.test(residual~group, toluca )
```

Brown-Forsythe Test

data : residual and group

statistic : 0.5567856
 num df : 4
 denom df : 16.50945
 p.value : 0.6970538

Result : difference is not statistically significant.



Breusch-Pagan Test (BP test)

```
library(lmtest)
bptest(lm(hour~size, toluca))
```

studentized Breusch-Pagan test

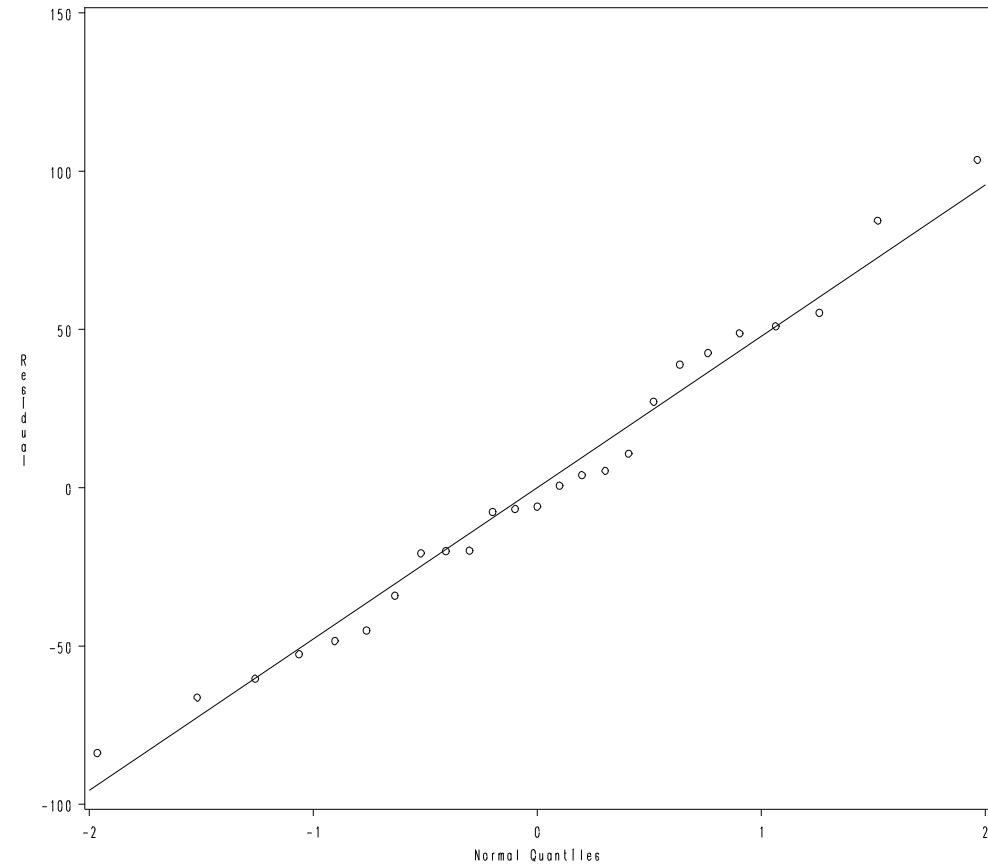
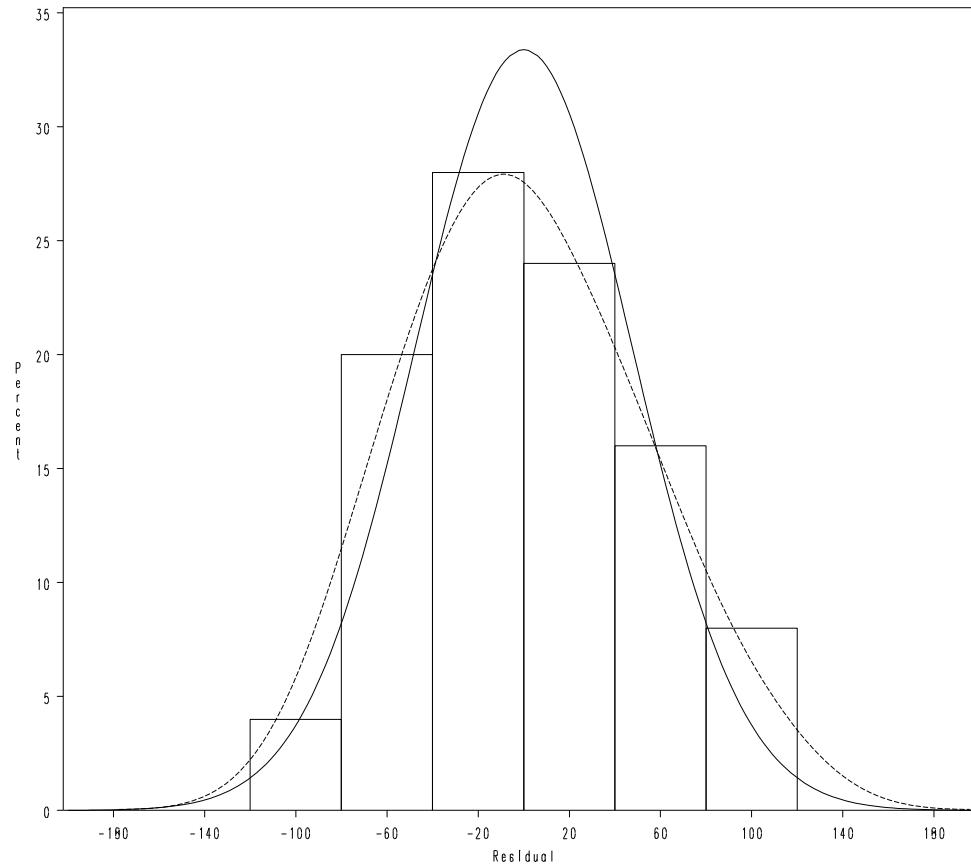
```
data: lm(hour ~ size, toluca)
BP = 1.1326, df = 1, p-value = 0.2872
```

Normal assumption and diagnostics (check residuals)

"**Are the errors normal?**" is not actually a very helpful question. More useful questions are:

- How far is the distribution of the errors from normal?
- In what way is it non-normal? Is it heavy-tailed? light-tailed? skewed? discrete?
- How will the non-normality affect our inferences? Is it bad enough to invalidate our confidence intervals and hypothesis tests?

Normal quantile plot appears to be a straight line

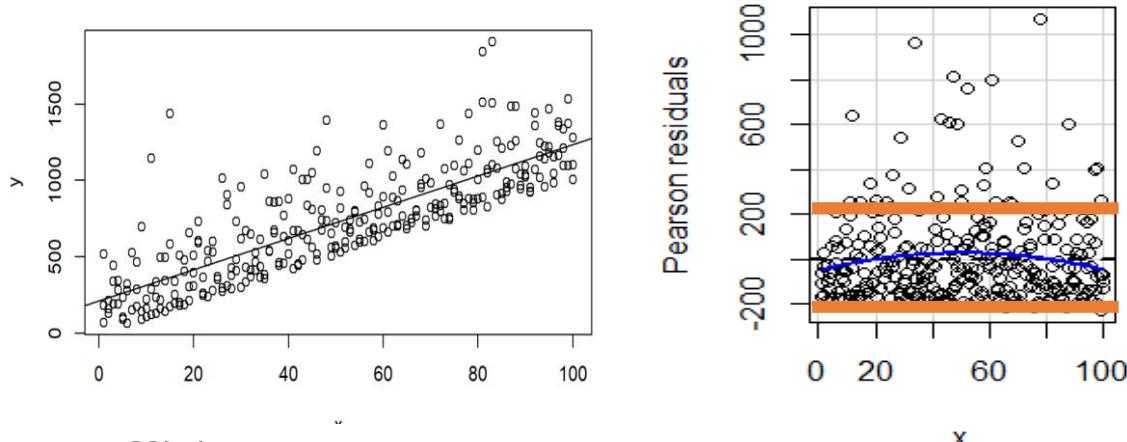


The Toluca example looks pretty good.

A case with non-normal errors

$$Y = \beta_0 + \beta_1 X + \varepsilon, \varepsilon \sim \exp\left(\frac{1}{200}\right), \mu\{\varepsilon\} = 200, \sigma\{\varepsilon\} = 200$$

X ranges from 1 to 100, replicate=3



Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	212.4916	24.5219	8.665	2.92e-16 ***
x	10.2293	0.4216	24.265	< 2e-16 ***

Residual standard error: 221.8 on 298 degrees of freedom
 Multiple R-squared: 0.644. Adjusted R-squared: 0.6428
 F-statistic: 539.2 on 1 and 298 DF, p-value: < 2.2e-16

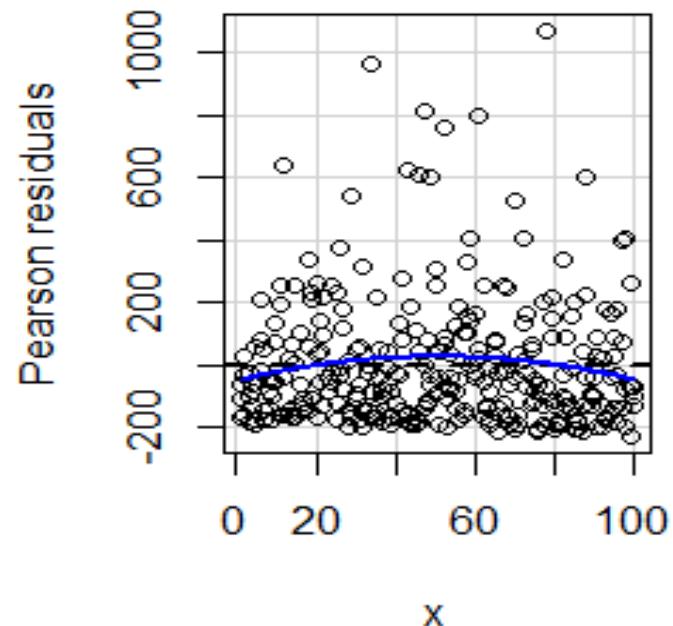
Model 1: $y \sim x$

Model 2: $y \sim \text{as.factor}(x)$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	298	14659698				
2	200	9815558	98	4844140	1.0072	0.4757

- The problem with this case is the outliers.

Fit a SLR to a non-normal data



Model 1: $y \sim x$

Model 2: $y \sim \text{as.factor}(x)$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	298	14659698				
2	200	9815558	98	4844140	1.0072	0.4757

Shapiro test for normality

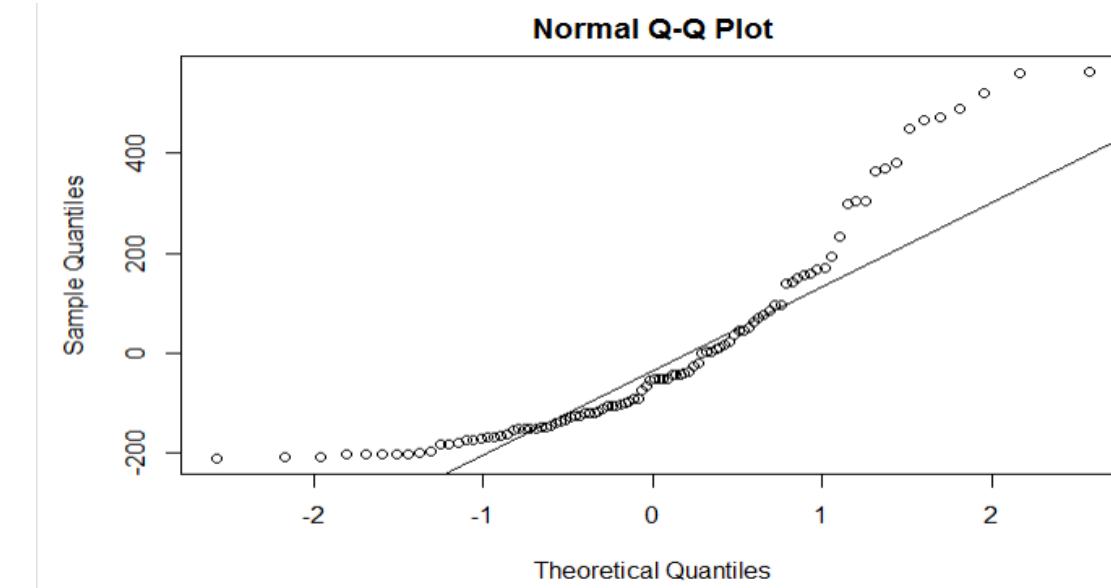
H₀: Data follows normal distribution.

H_a: Data violates from normal distribution.

```
exporesid<-residuals(lm(y~x, data))
shapiro.test(exporesid)
qqnorm(exporesid)
qqline(exporesid)
```

Shapiro-Wilk normality test

```
data: exporesid
W = 0.79732, p-value < 2.2e-16
```



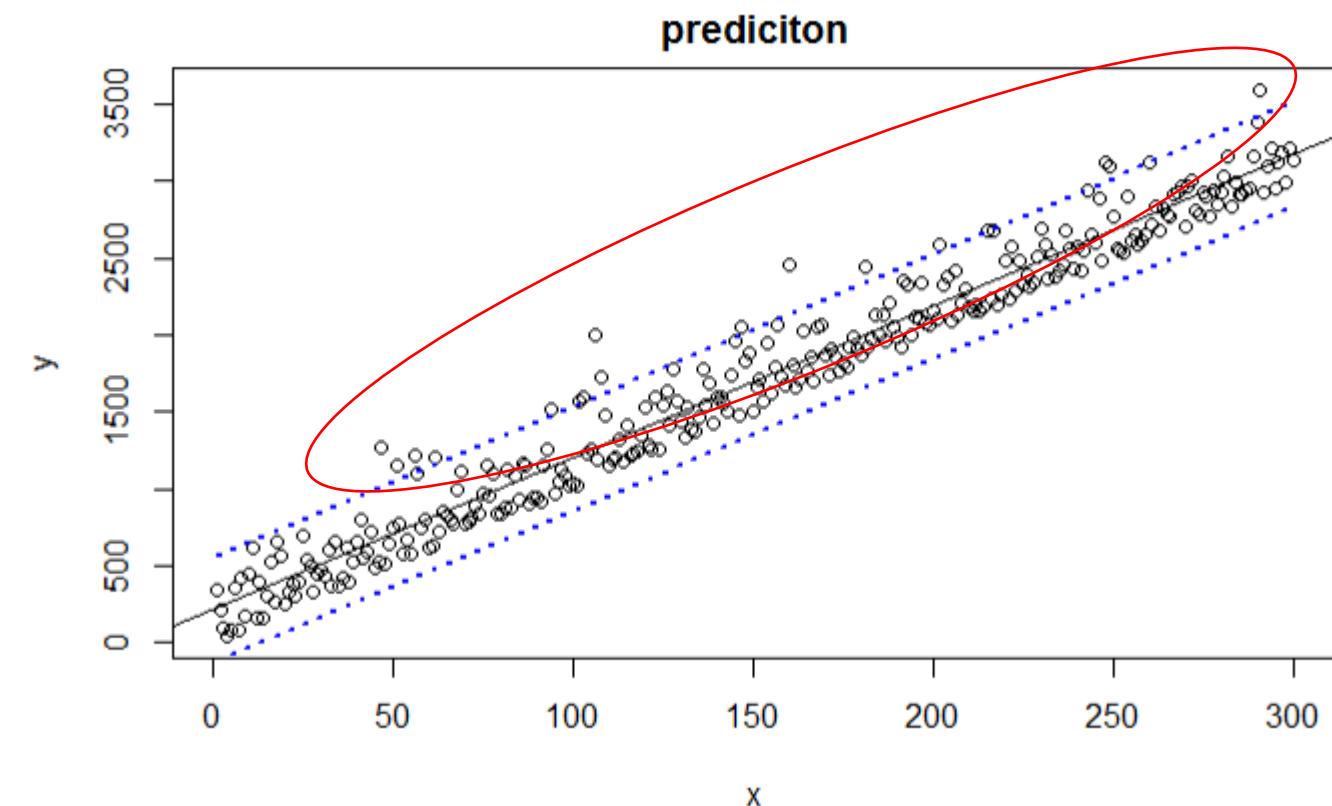
Simulating consequence with non-normal violation (another dataset)

```
library(ALSM)
x<-seq(1:300)
y<-10*x+rexp(300, rate=1/200)
expo<-data.frame(x,y)
expo.mod<-lm(y~x, expo)
cin<-ci.reg(expo.mod, expo$x, type='n',alpha=0.05)
plot(y~x, expo, main="prediciton")
abline(expo.mod)
lines(expo$x, cin$Lower.Band,col="blue", lwd=2, lty=3)
lines(expo$x, cin$Upper.Band, col="blue", lwd=2, lty=3)
```

*We create confidence band for
predicting a single response variable,
 $\hat{Y}_h\{new\}$*

$$\hat{Y}_h\{new\} \pm ts^2\{pred\}$$

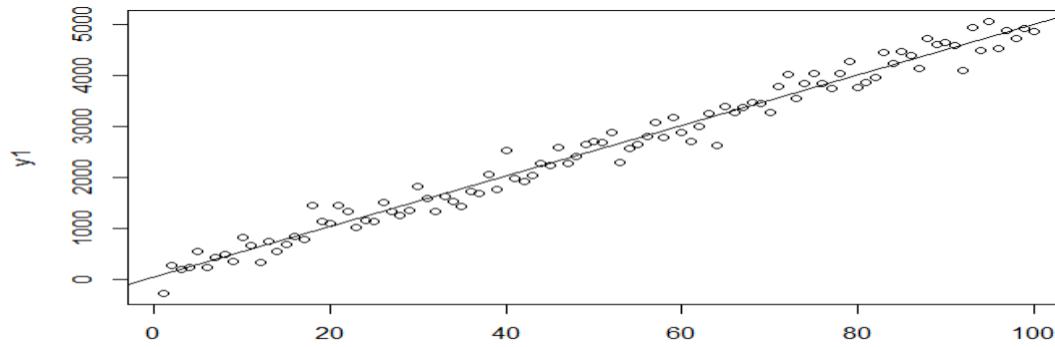
$$\text{Where } s^2_{\{pred\}} = s^2_{\{\hat{Y}_h\}} + s^2 = s^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} + 1 \right]$$



Diagnostic procedure on outliers (the informal procedure)

- Outliers are extreme observations.
- Residual outliers can be identified from ***residual plot, boxplot, stem-and-leaf plot*** etc.
- Under the least square (LS) method, a fitted line may be pulled disproportionately toward an outlier.
- Outlier may convey significant information because of an interaction with another predictor variable.
- Outlier that stand *near and far to \bar{X} has different impact*

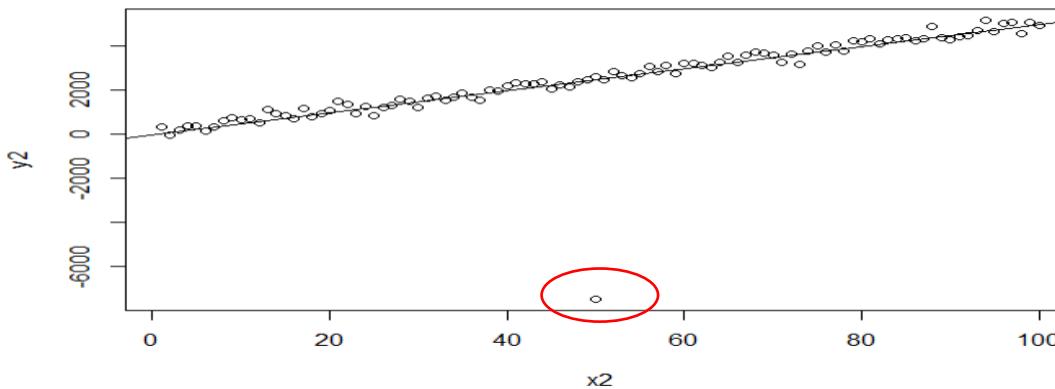
Impact of outliers: **true model** $Y = 30 + 50X + N(0, 200)$



Without outlier

$$\hat{Y} = 42.03 + 49.8X$$

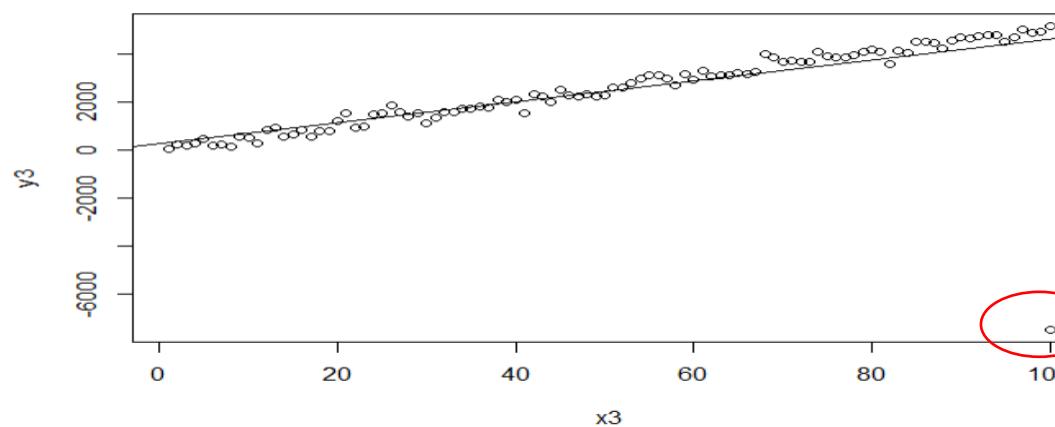
$$R^2 = 0.99, s = 191.8 \quad (\sigma = 200)$$



With outlier near \bar{X}

$$\hat{Y} = -59.98 + 49.8X$$

$$R^2 = 0.67, s = 1018$$



With outlier near X_{max}

$$\hat{Y} = 235.1 + 43.5X$$

$$R^2 = 0.51, s = 1251$$

Outlier nears the edge has more impact.

Different kinds of outliers and impact

Outliers near the mean of \bar{X} can **influence** the intercept but lack the **leverage** to strongly affect the slope. However, they still inflate the standard errors for both parameters.

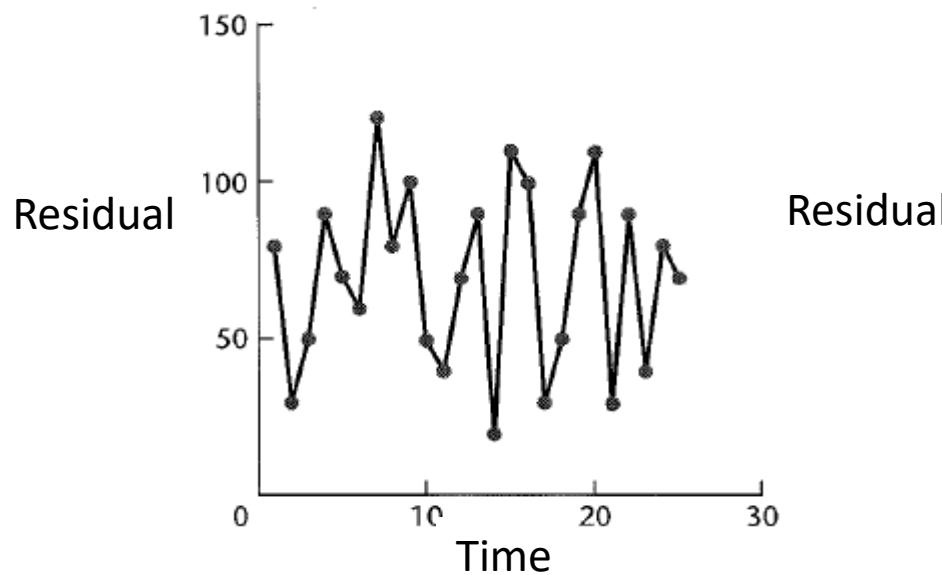
Outliers that are located further from \bar{X} has **greater leverage**, and thus a greater effect on the estimated slope for the same “extremeness”. They increase MSE and reduce the precision of estimates.

Diagnostic procedure on dependent Errors

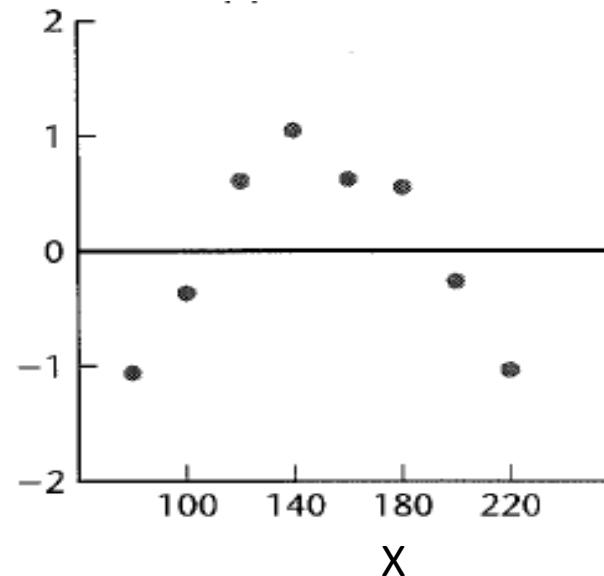
- Ideally, any potential source of dependence is handled at the experimental design stage, so that it is either eliminated by randomization or explicitly included in the data.
- Always watch out for trends or cyclical patterns in the residuals, and plot the residuals **against time, collection order, spatial coordinates, etc.** if you think these might affect the data.
- More subtle dependences can be difficult to detect, especially if the information needed to detect them has not been included with the dataset.

Example

Residual plot A



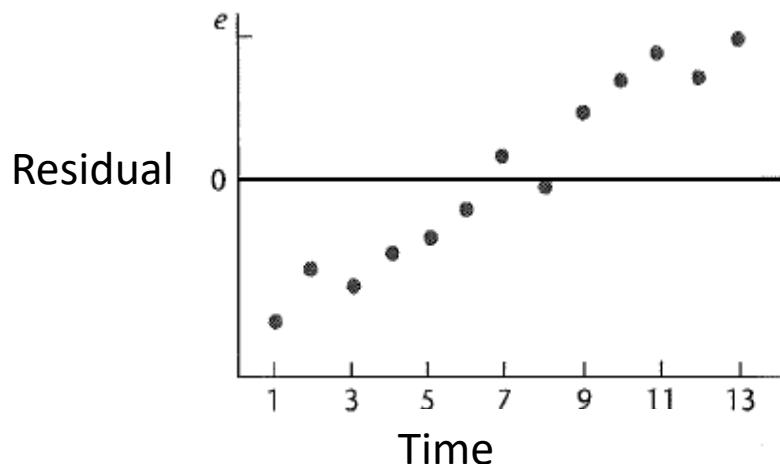
Residual plot B



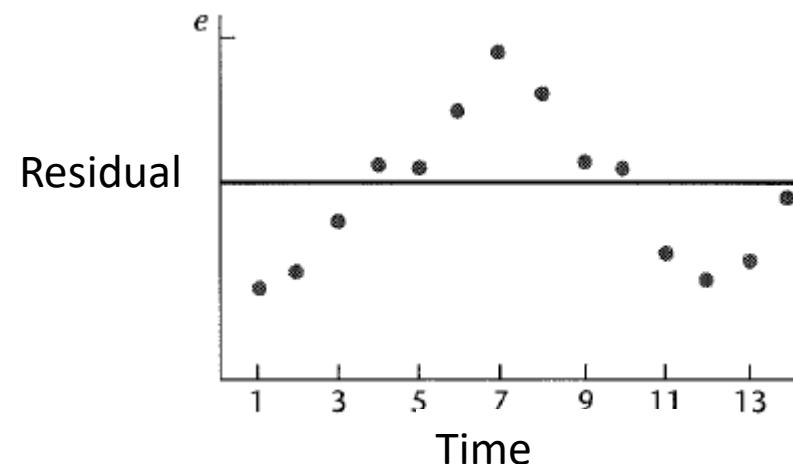
Which of the residual plots provide(s) strong evidence of dependent errors?

--C and D

Residual plot C



Residual plot D



Diagnostic procedure on distribution of predictors

Linear models do not make any assumptions about X ,
but the distribution of X in the data can affect

- the *scope* of the model
- the *accuracy* of inferences for the values of parameters
- the *efficiency* (and accuracy) of inferences for \hat{Y}_h and $Y_{h(new)}$

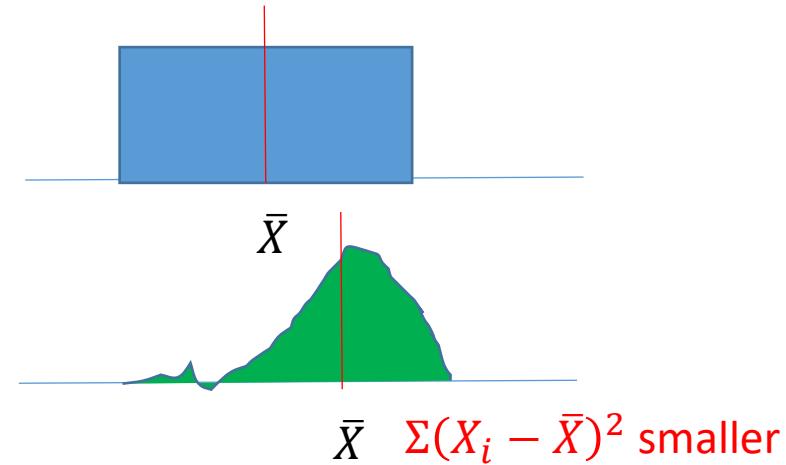
Why does the distribution of X matter?

Even though we make no assumptions about the distribution of X , our estimators depend on its sample mean and variance (technically, on $SS_X = \sum(X_i - \bar{X})^2$).

If the range of X in the data is held constant, then relative to a dataset where X is uniformly distributed, a more skewed distribution (or outliers) will:

- pull \bar{X} toward the *body* of the distribution
- cause SS_X to be smaller

As a result . . .



Datasets in which X is skewed will generally yield *less precise estimates for the slope* parameter(s) compared with datasets in which X is more uniformly distributed:

$$s^2\{b_1\} = \frac{MSE}{\sum(X_i - \bar{X})^2}$$

If the estimate for β_1 is less precise, then estimates for \hat{Y}_h and $Y_{h(new)}$ will also suffer.

$$s^2_{\{\hat{Y}_h\}} = s^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right] \text{ and } s^2_{\{pred\}} = s^2_{\{\hat{Y}_h\}} + s^2 = s^2$$

In addition, outliers or skew increase the risk that a small number of highly influential data points can dominate the fit and degrade accuracy.

Summary

- We discuss model departures one at a time, although in actuality, **several types of departures may occur together**. For instance, a linear regression function may be a poor fit and the variance of the error terms may not be constant.
- Graphic analysis of residual analysis is one informal method of analysis, but in many cases, it suffices for examining the aptness of a model.
- The basic approach to residual analysis explained here applies not only simple linear regression but also to more complex regression and other types of statistical models.

Summary

- Several types of departures from the simple linear regression model have been identified by diagnostic tests of the residuals. **Model misspecification due to either nonlinearity or the omission of important predictor variables tends to be serious**, leading to biased estimates of the regression parameters and error variances.
- **Non-constancy of errors variance tends to be less serious**, leading to less efficient estimates and invalid error variance estimates.
- **The presence of outliers can be serious** for smaller data sets when their influence is large.
- Finally, the **dependence of error terms** results in estimators that are unbiased but whose variances are seriously biased. These problems will be discussed later in depth.

Remedial Procedure in SLR: transformation

Overview of remedial measures

If the simple linear regression model is not appropriate for a data set

- Abandon regression model and develop a more appropriate model
- Employ some transformation on the data so that regression model is appropriate
for the transformed data

- Nonlinearity of regression function → Transformations
- Non-constancy of error variance → Transformations and Weighted least squares
- Non-normality of error terms → Transformations
- Outliers → Transformations or Robust regression
- Non-independence of Error terms → Autocorrelation, time series analysis

When the error terms approximately have a Normal distribution with constant variance

- Transformation on X should be attempted (at first).

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon \rightarrow Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \text{ where } X_1 = X, X_2 = X^2$$

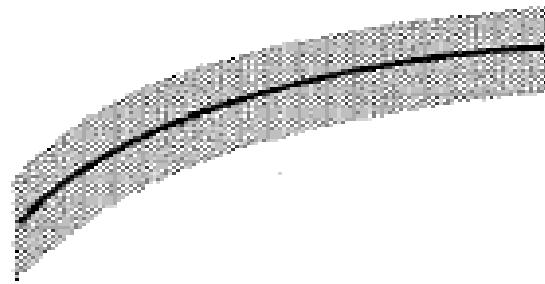
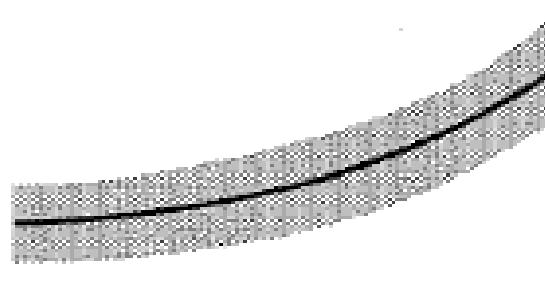
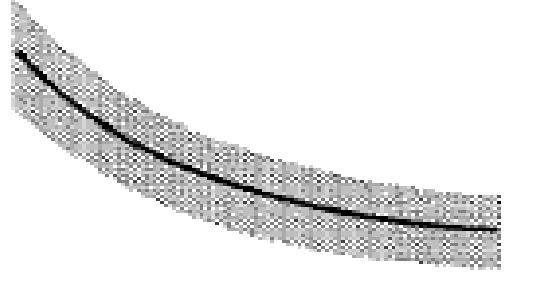
$$Y = \beta_0 + \beta_1 \log(X) + \varepsilon \rightarrow Y = \beta_0 + \beta_1 X_1 + \varepsilon, \text{ where } X_1 = \log(X)$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon \rightarrow Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon, \text{ where } X_3 = X_1 X_2$$

- The reason why transformation on Y may not be desirable is that transformation of Y may materially change the shape of distribution of the error terms from the Normal distribution and lead to differing error term variances.

$$Y = \beta_0 + \beta_1 X + \varepsilon \rightarrow \sqrt{Y} = \beta_0 + \beta_1 X_1 + \text{new } \varepsilon$$

Some common transformation form on X

	Prototype Regression Pattern	Transformations of X	Comment
(a)		$X' = \log_{10} X$ $X' = \sqrt{X}$	<ul style="list-style-type: none"> If some of the X data are near 0 and reciprocal transformation is desired. Shift the origin by $X' = \frac{1}{X + k}$
(b)		$X' = X^2$ $X' = \exp(X)$	Where $k \neq 0$
(c)		$X' = 1/X$ $X' = \exp(-X)$	

Unequal error variance and nonnormality of the error terms frequently appear together, and we need a transformation on Y

	Prototype Regression Pattern	Transformations on Y	Comment
(a)		$Y' = Y^\lambda$ (Box-Cox Transformation) For example, $\lambda = 2 \quad Y' = Y^2$ $\lambda = 0.5 \quad Y' = \sqrt{Y}$ $\lambda = 0 \quad Y' = \ln Y$	<ul style="list-style-type: none"> Consider use constant values to validate the transformation function
(b)		$\lambda = -0.5 \quad Y' = \frac{1}{\sqrt{Y}}$ $\lambda = -1 \quad Y' = \frac{1}{Y}$	$Y' = \log_{10}(Y + k)$ <i>k is selected such that $Y + k > 0$ for all Y.</i> <ul style="list-style-type: none"> Can be combined with transformation on X
(c)			

Box-Cox Procedure

Transformations on Y sometimes help with variance issue: non-normality and non-constant.

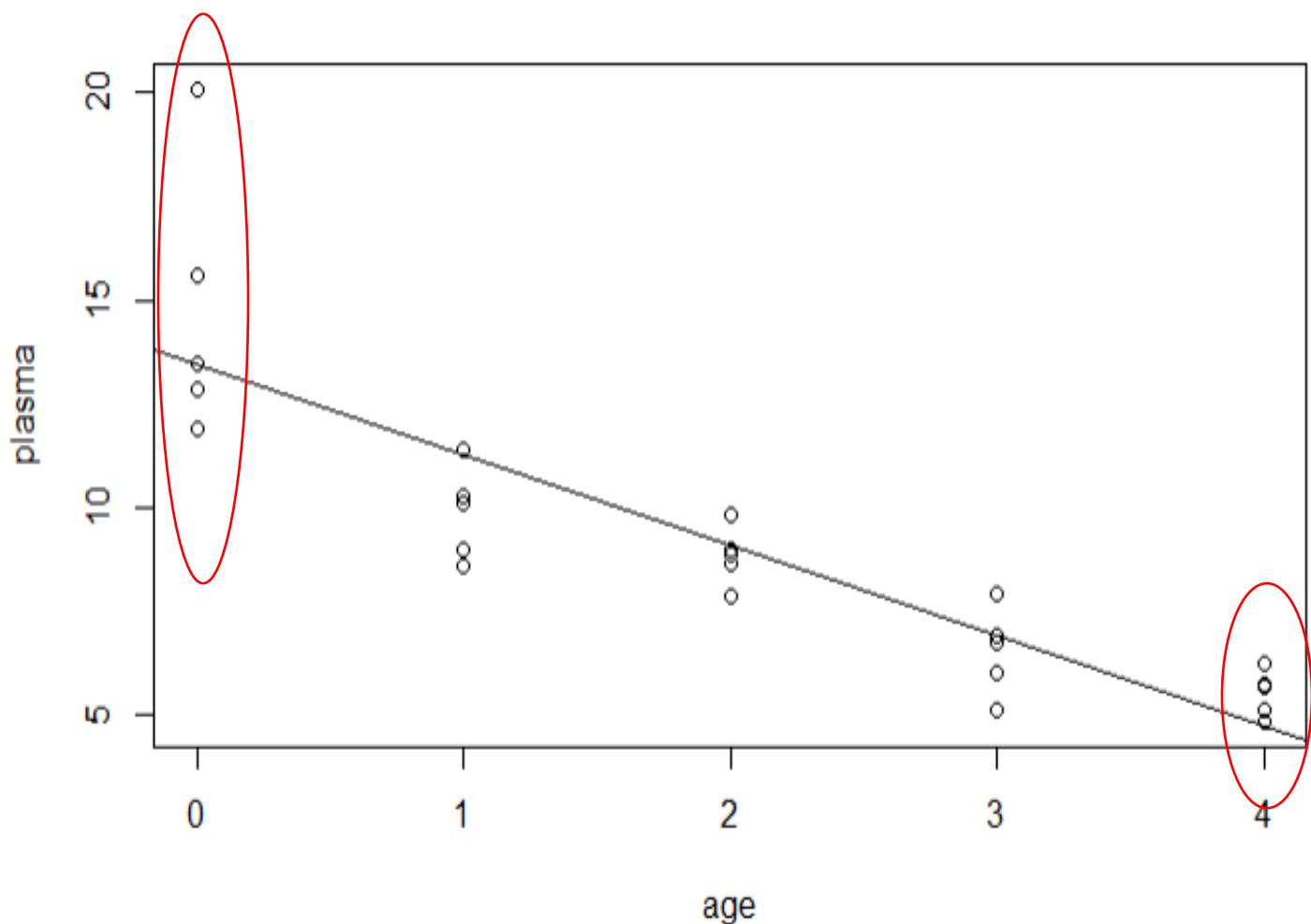
- Box-Cox considers a family of so-called “power transformations”,

$$Y' = Y^\lambda$$

- “Works by using the method of **maximum likelihood** or **minimum SSE** to find the value of λ that produces the best (transformed) regression $Y^\lambda = \beta_0 + \beta X + \varepsilon$
- Need to check assumptions for the transformed regression model.

The Plasma example

Age (X) and plasma level of a poly amine (Y) for a portion of the 25 healthy children are studied.
 Scatter plot shows there is greater variability for younger children than for older ones



Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	13.4752	0.6379	21.126	< 2e-16 ***
age	-2.1820	0.2604	-8.379	1.92e-08 ***

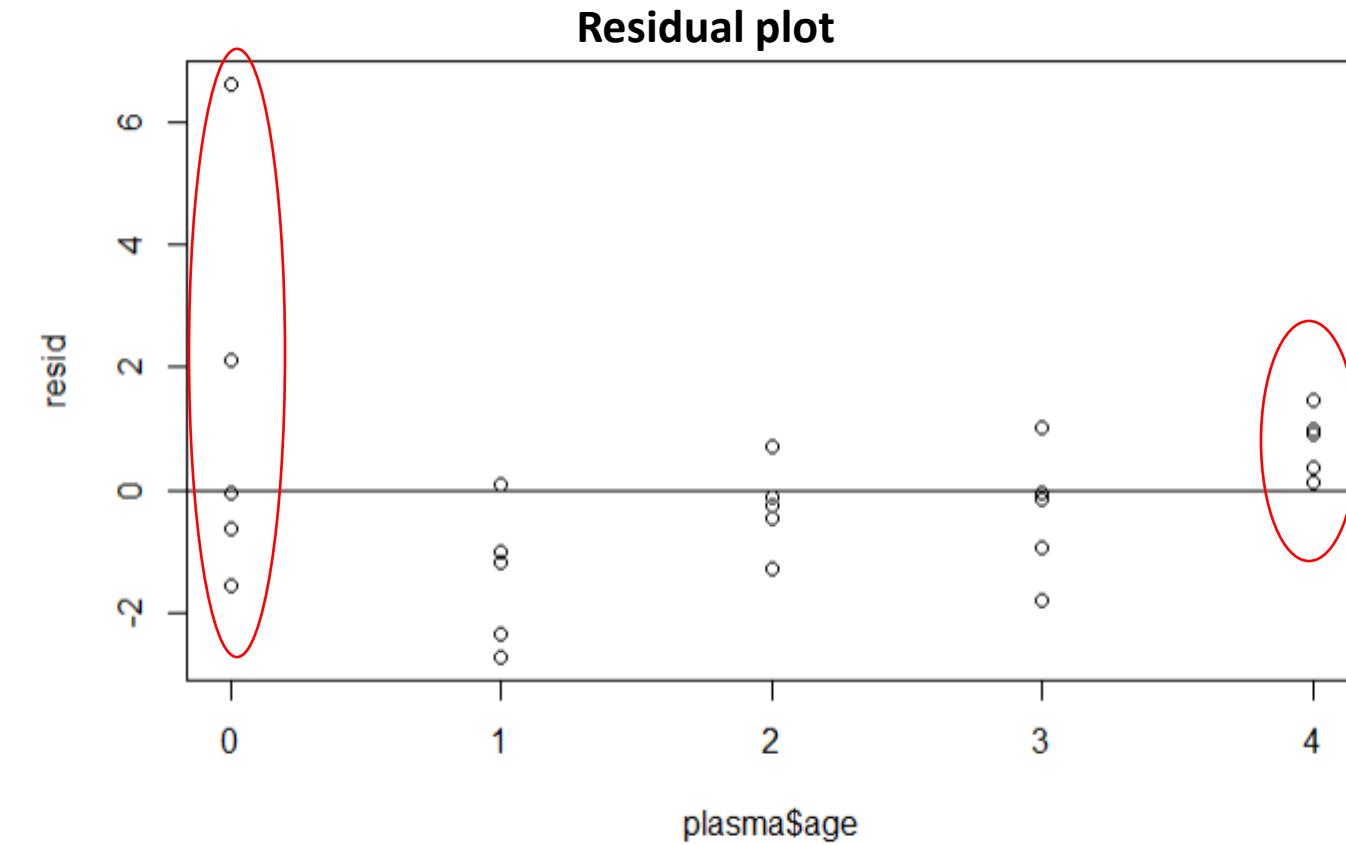
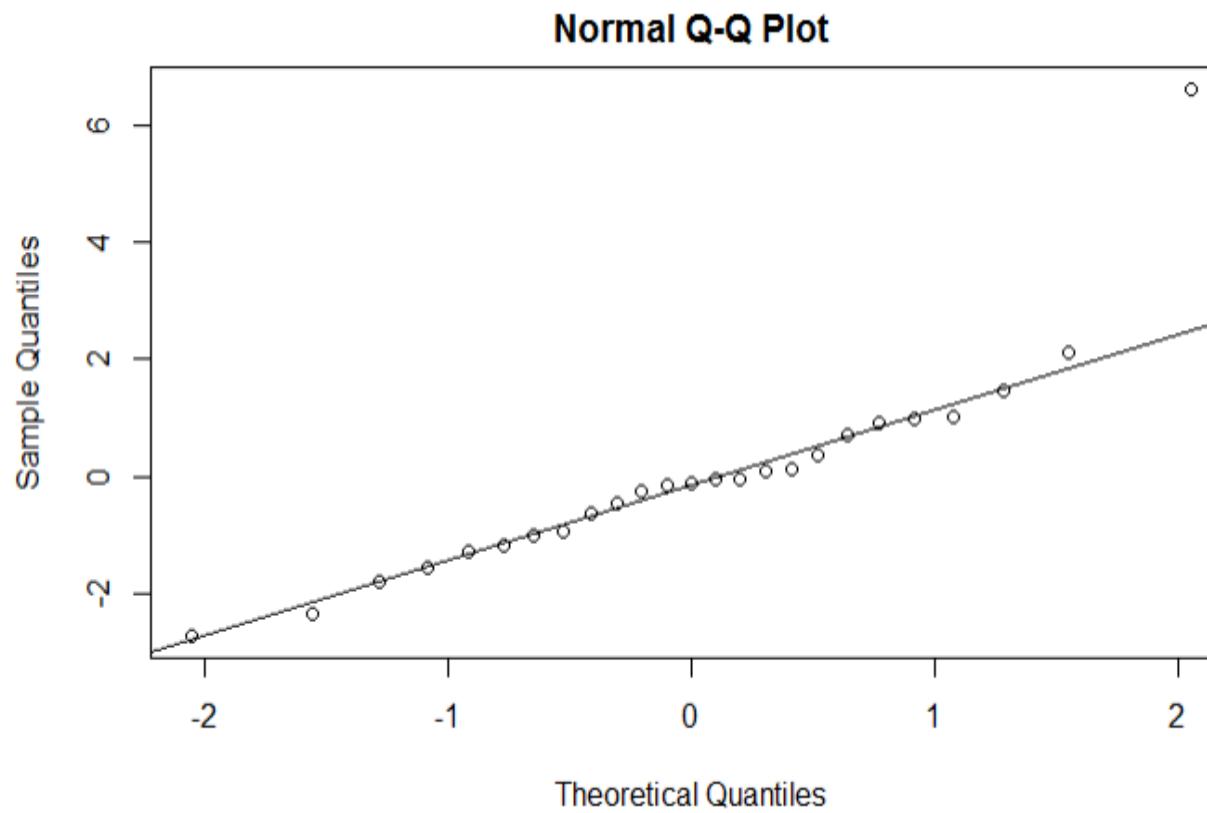
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.841 on 23 degrees of freedom

Multiple R-squared: 0.7532, Adjusted R-squared: 0.7425

F-statistic: 70.21 on 1 and 23 DF, p-value: 1.92e-08

Check Normality and constancy on the residuals



Shapiro-Wilk normality test

```
data: residuals(plasma.mod)
W = 0.83903, p-value = 0.001098
```

```
shapiro.test(residuals(plasma.mod))
qqnorm(residuals(plasma.mod))
qqline(residuals(plasma.mod))
```

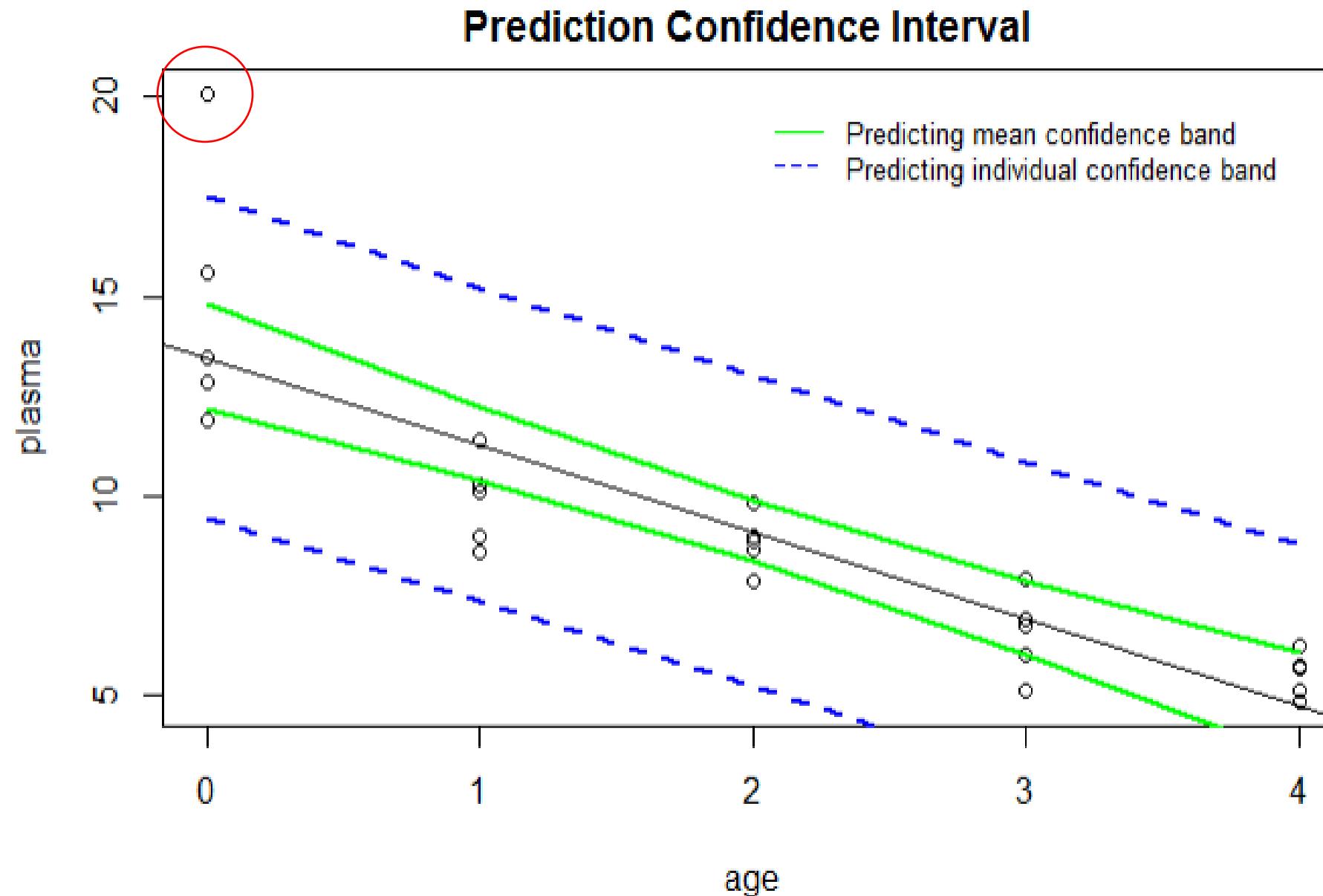
Brown-Forsythe Test

```
data : residual and agef
statistic : 2.059299
num df   : 4
denom df : 6.526859
p.value  : 0.1965498
```

Result : difference is not statistically significant.

```
bf.test(residual~agef, plasma)
```

Confident interval band



$$\hat{Y}_h \pm t_c s_{\{\hat{Y}_h\}}$$

$$\text{Where } s_{\{\hat{Y}_h\}}^2 = s^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right]$$

$$\hat{Y}_h \pm t_c s_{\{pred\}}$$

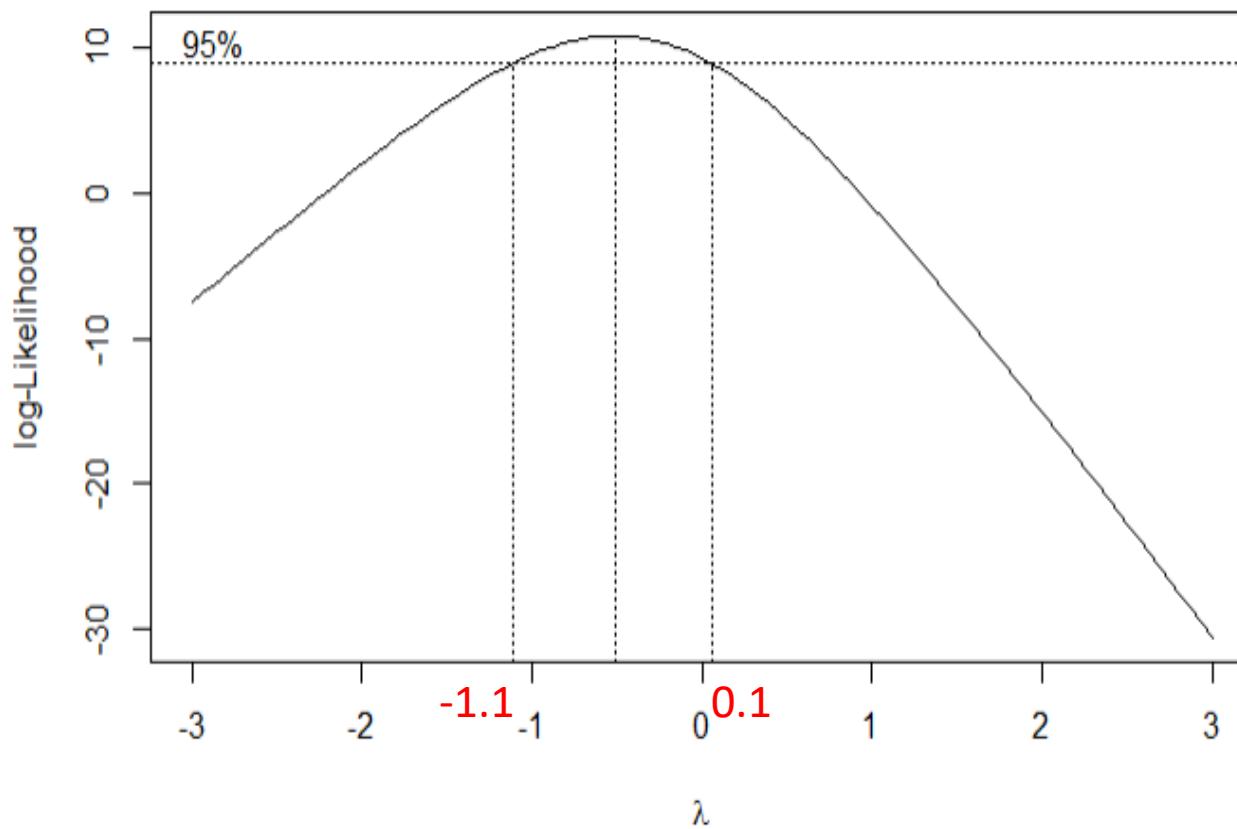
$$\text{Where } s_{\{pred\}}^2 = s^2 + s_{\{\hat{Y}_h\}}^2$$

After a careful exam on the experiment procedure, no mistake has been found, hence we should keep this observation.

Box-cox procedure

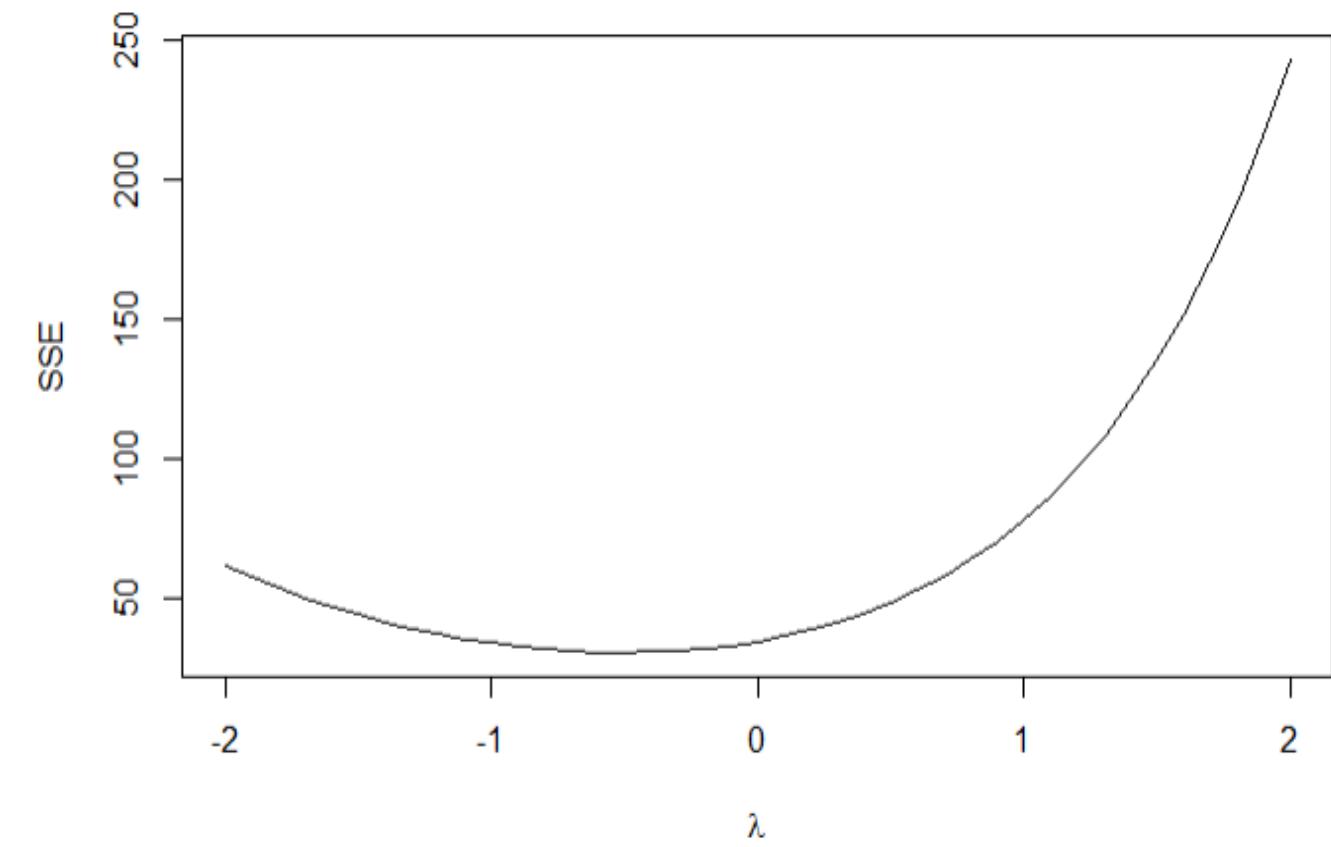
$$\lambda = -0.5 \quad Y' = \frac{1}{\sqrt{Y}}$$

The best $\lambda = -0.515$ (biggest log – likelihood)



```
library(MASS)
bcmle<-boxcox(lm(plasma~age,data=orig),lambda=seq(-3,3, by=0.1))
lambda<-bcmle$x[which.max(bcmle$y)]
lambda
```

The best $\lambda = -0.5$ (smallest SSE)

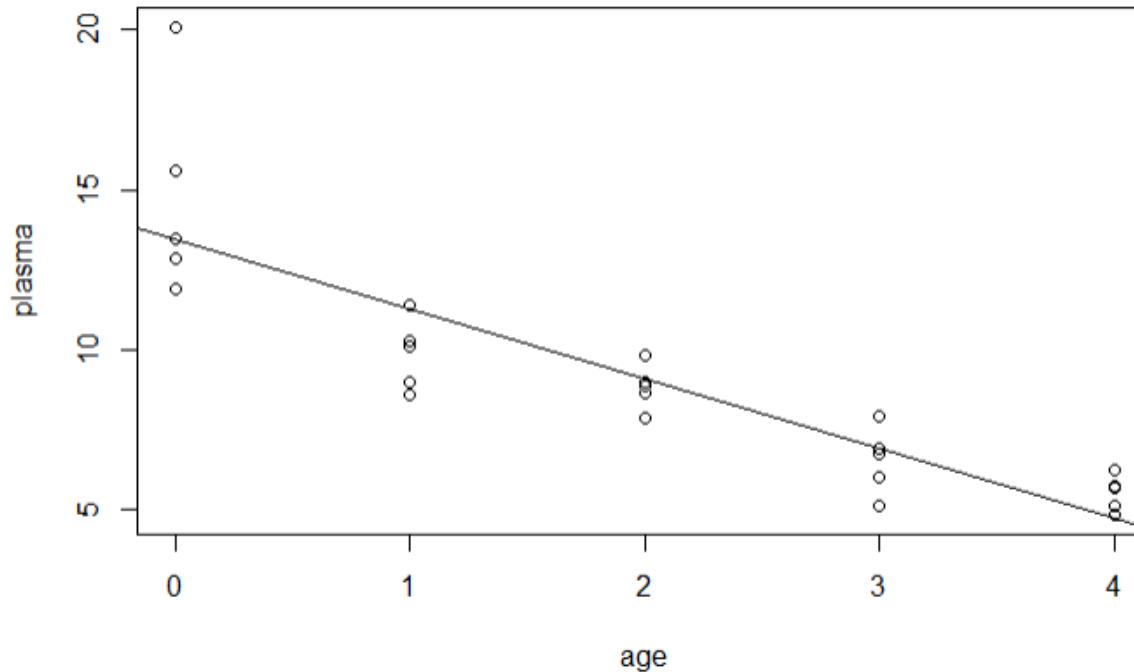


```
library(ALSM)
bcsse<-boxcox.sse(plasma$age,plasma$plasma,l=seq(-2,2,0.1))
lambda<-bcsse$lambda[which.min(bcsse$SSE)]
lambda
```

$Y^\lambda = \beta_0 + \beta_1 X$, where λ ranges from – 3 to 3, increases by 0.1)

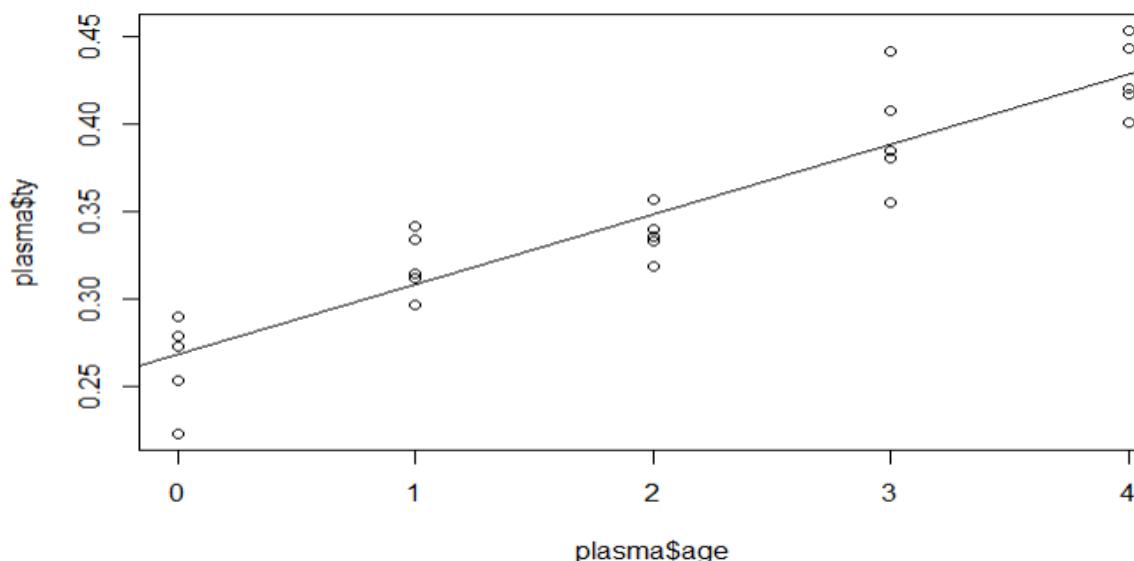
$$\text{Transform } Y' = \frac{1}{\sqrt{Y}}$$

Because `s(residual)` has the same unit as the response variable, but transformation alters that.



```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.4752    0.6379 21.126 < 2e-16 ***
age         -2.1820    0.2604 -8.379 1.92e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.841 on 23 degrees of freedom
Multiple R-squared:  0.7532, Adjusted R-squared:  0.7425
F-statistic: 70.21 on 1 and 23 DF, p-value: 1.92e-08
```

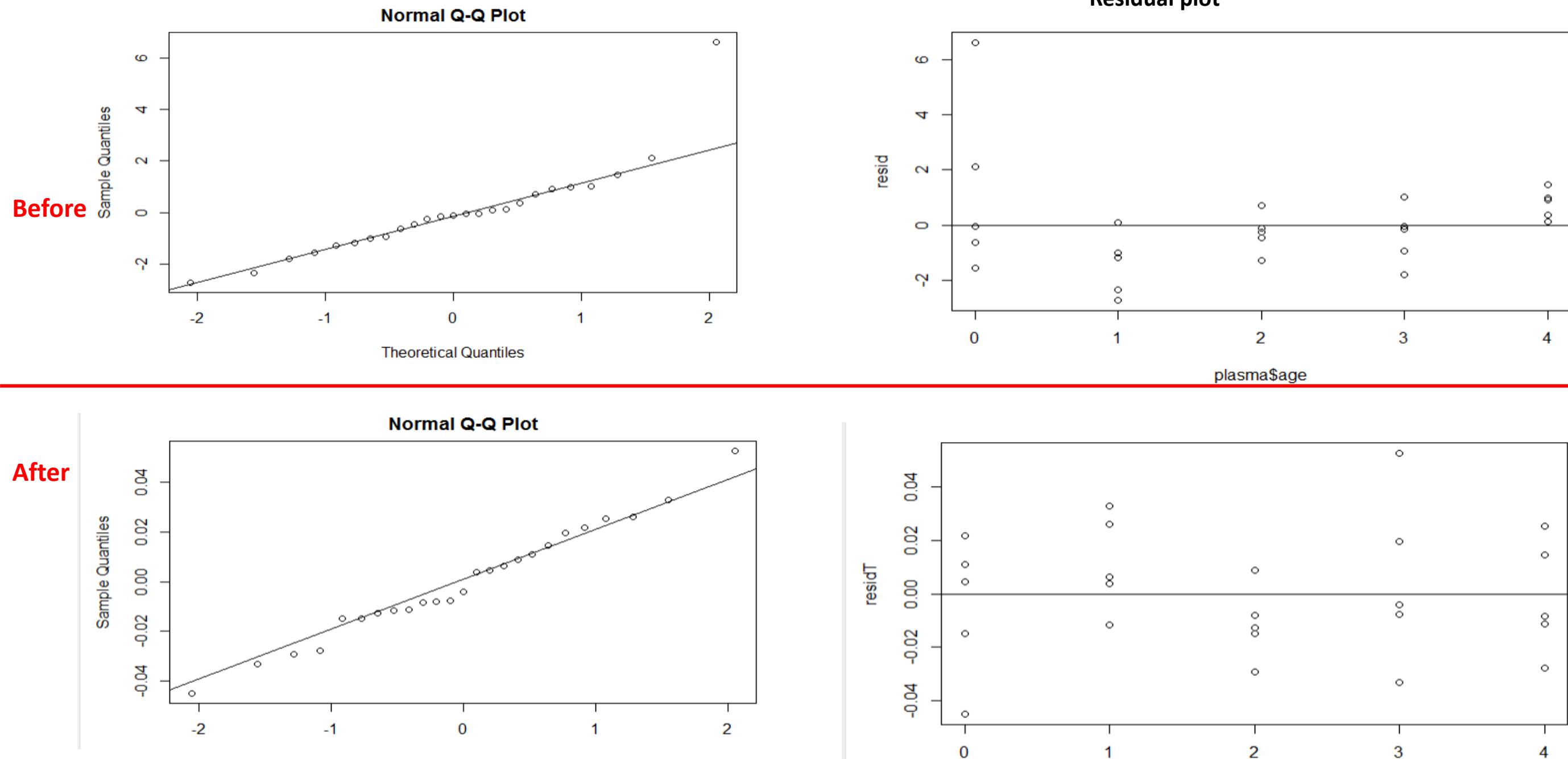


```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.268026   0.008033 33.36 < 2e-16 ***
age         0.040062   0.003280 12.22 1.55e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02319 on 23 degrees of freedom
Multiple R-squared:  0.8665, Adjusted R-squared:  0.8606
F-statistic: 149.2 on 1 and 23 DF, p-value: 1.548e-11
```

Why the `s(residual)` is not comparable?

Re-Check Normality and constancy on the residuals



Back-transformations

Transformations can improve model performance, but make interpretation hard.

Back transformation lets us make inferences (and **graphs!**) on the original scale.

Very helpful for communicating results to the public.

Interpreting the confidence interval for the mean and single prediction

- In general, let $Y' = f(Y)$ and let f' be the **back-transformation function**.

For example,

$Y' = f(Y) = Y^2$, the back-transformation function f' does

$f'(Y') = Y$, so $f'(Y') = \sqrt{Y^2} = Y$

- Then, back transform the mean and single response confidence interval (a, b) as following

$(f'(a), f'(b))$, For example, (\sqrt{a}, \sqrt{b})

- Back transforming the coefficients or the standard error is not accurate.

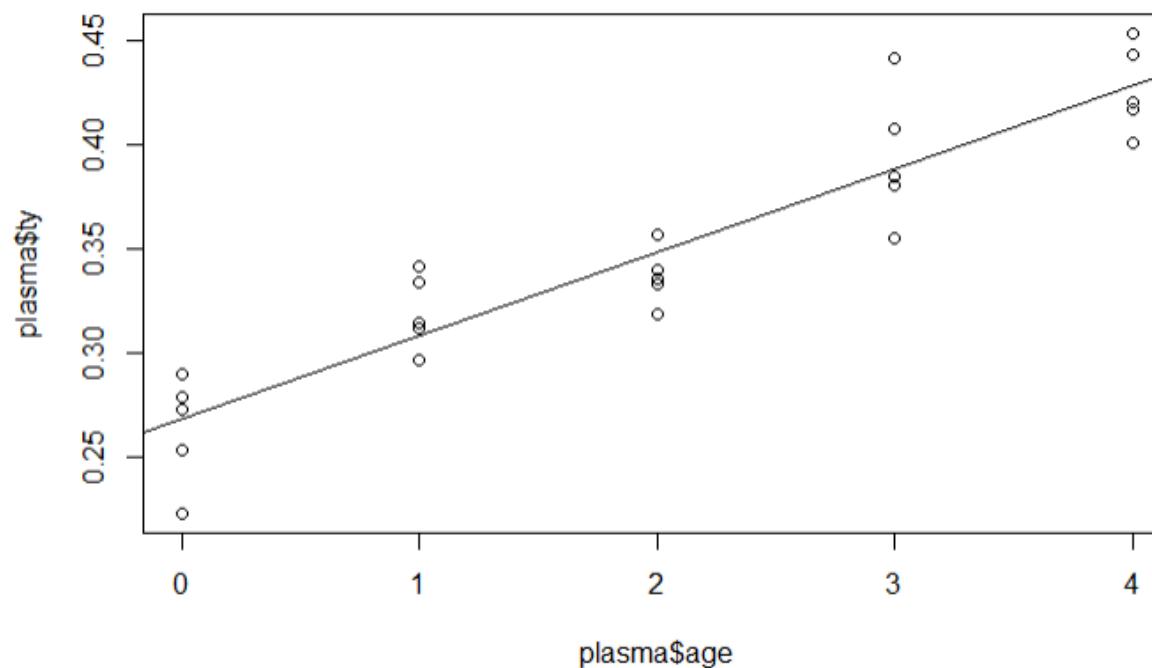
For example, do not back transform the point estimate with $\hat{Y} = f'(b_0) + f'(b_1)X = \sqrt{b_0} + \sqrt{b_1}X$,
instead, do $\sqrt{b_0 + b_1 X}$

- If only X is transformed to X' , then no need to back transform Y 's estimation because Y hasn't been transformed.

For example, $\hat{Y} = b_0 + b_1 X'$

Back-transform $Y' = \frac{1}{\sqrt{Y}}$

1. The back transform function $f' = \frac{1}{Y'^2} = (Y')^{-2}$
2. The predicted value should be $\hat{Y} = (\hat{Y}')^{-2}$
3. The confidence interval for the prediction, either for the mean or single response, should also be back transformed with $(value)^{-2}$.

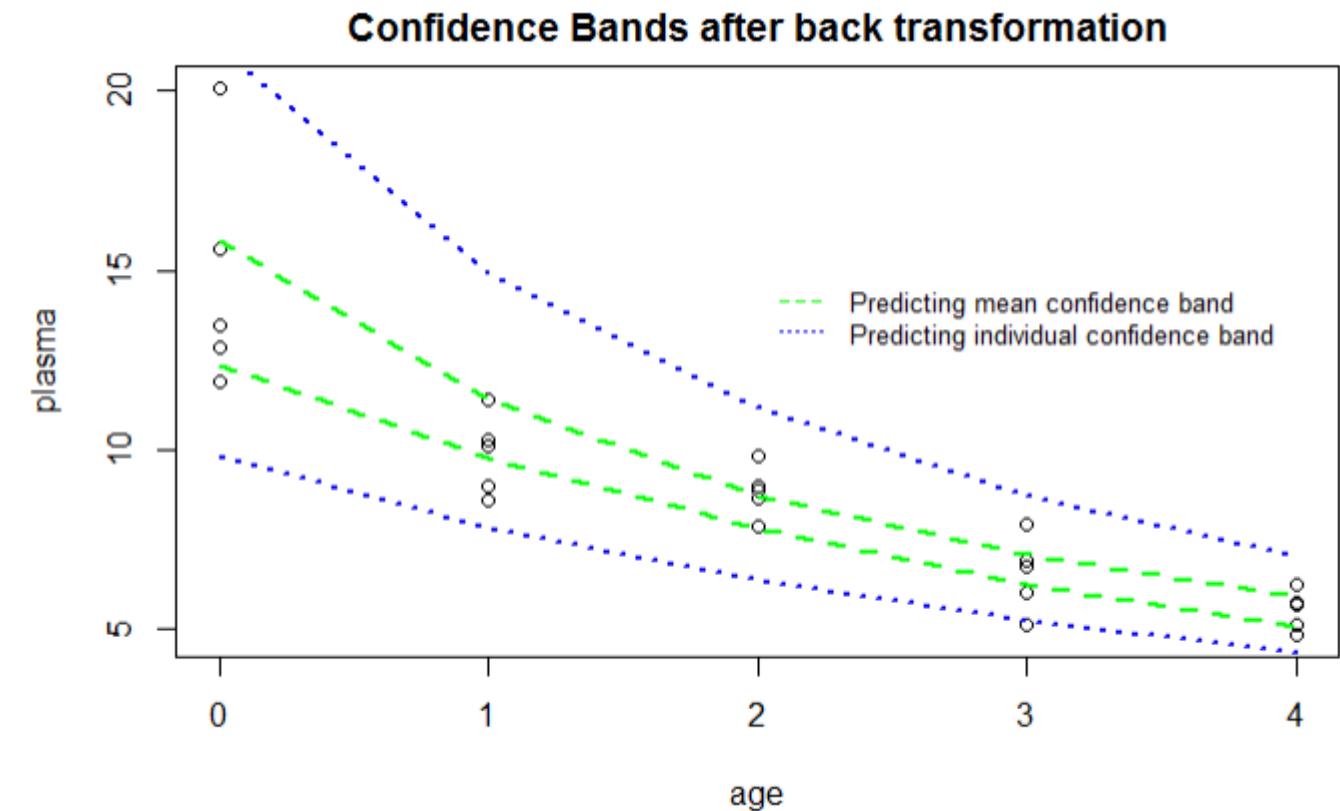
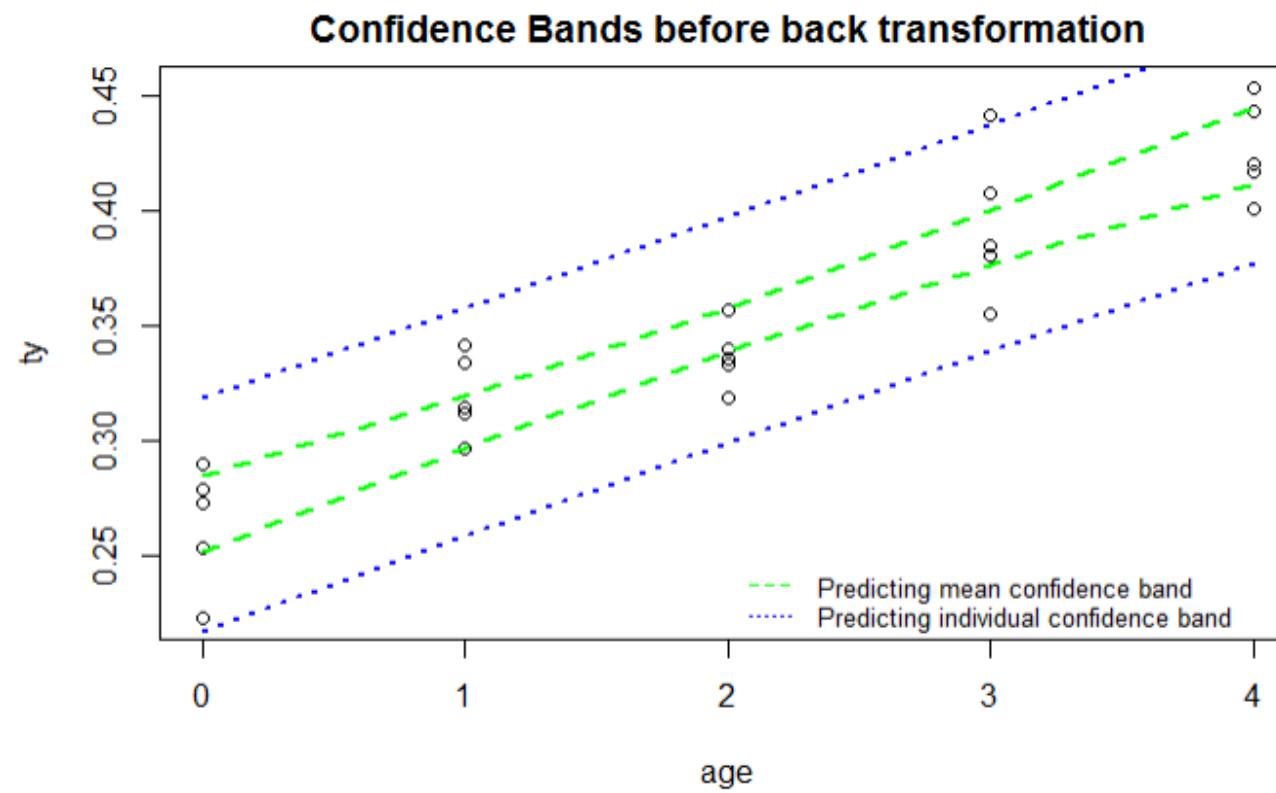


```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.268026  0.008033 33.36 < 2e-16 ***
age         0.040062  0.003280 12.22 1.55e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02319 on 23 degrees of freedom
Multiple R-squared:  0.8665,    Adjusted R-squared:  0.8606
F-statistic: 149.2 on 1 and 23 DF,  p-value: 1.548e-11
```

$$\frac{1}{\sqrt{Y}} = 0.268 + 0.04(X)$$

Back-transform $Y' = \frac{1}{\sqrt{Y}}$, then $Y = (Y')^{-2}$



```
plot(ty ~ age, plasma, main="Confidence Bands before back transformation")
lines(cim$age, cim$Lower.Band,col="green", lwd=2, lty=2)
lines(cim$age, cim$Upper.Band, col="green", lwd=2, lty=2)
lines(cin$age, cin$Lower.Band,col="blue", lwd=2, lty=3)
lines(cin$age, cin$Upper.Band, col="blue", lwd=2, lty=3)
```

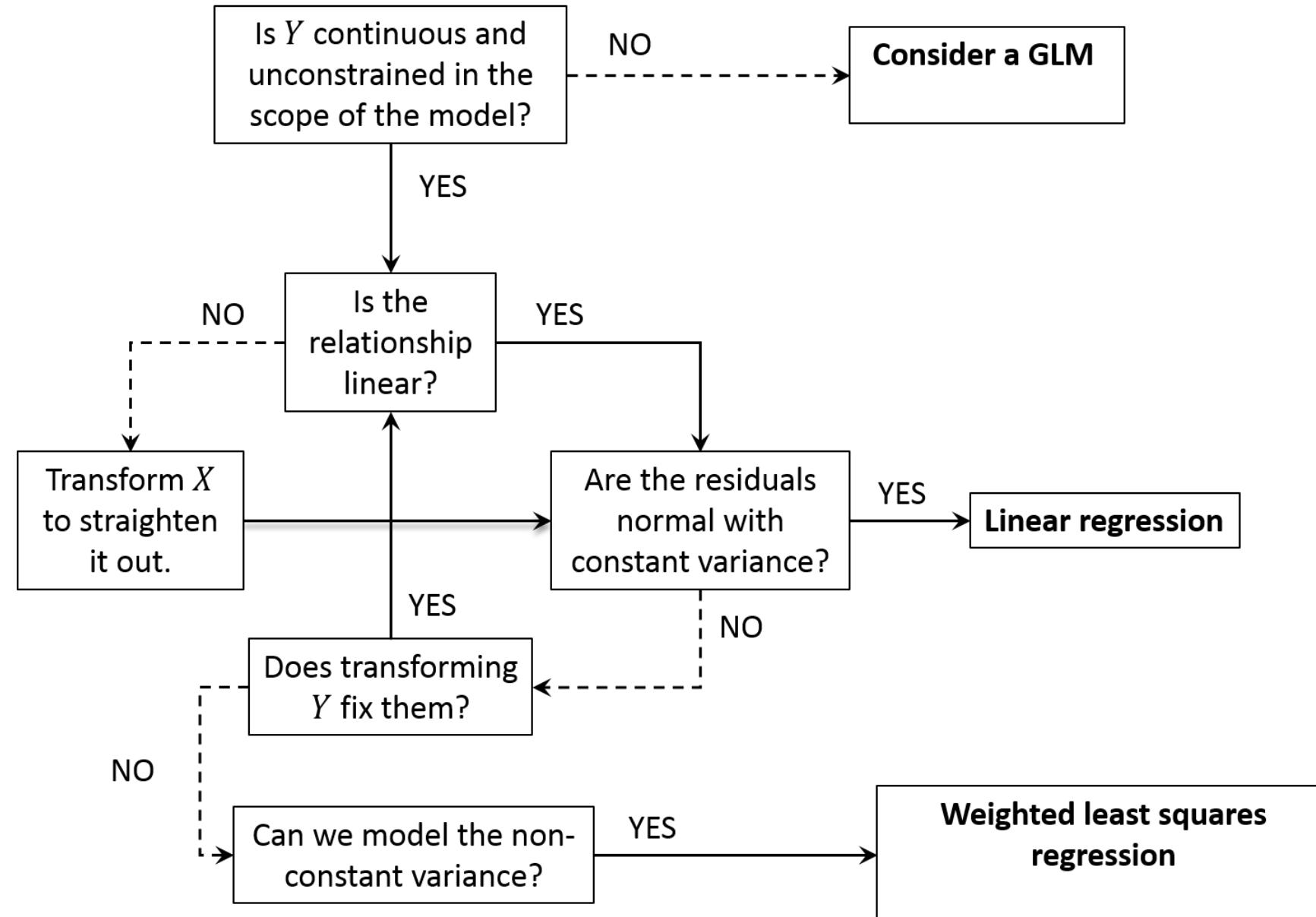
```
plot(plasma ~ age, plasma, main="Confidence Bands after back transformation")
lines(cim$age, (cim$Lower.Band)^(-2),col="green", lwd=2, lty=2)
lines(cim$age, (cim$Upper.Band)^(-2), col="green", lwd=2, lty=2)
lines(cin$age, (cin$Lower.Band)^(-2),col="blue", lwd=2, lty=3)
lines(cin$age, (cin$Upper.Band)^(-2), col="blue", lwd=2, lty=3)
```

Summary of remedial measures

- For nonlinear functional relationships with well behaved residuals
 - Try transforming X
 - May require a polynomial or piecewise fit (we will cover these later)
- For non-constant or non-normal variance, possibly with a nonlinear functional form
 - Try transforming Y
 - The Box-Cox procedure may be helpful
 - If the transformation on Y doesn't fix the non constant variance problem, weighted least squares can be used (we will cover this later).

- Transformations of X and Y can be used together.
- Any time you consider a transformation
 - Remember to recheck all the diagnostics.
 - Consider whether you gain enough to justify losing interpretability.
 - Reciprocal transformations make interpretation especially hard.
 - Consider back-transforming the results of the final model for presentation.
- For very non-normal errors, especially those arising from discrete responses, generalized linear models are often a better option, but linear regression may be “good enough.”

Transformation – our primary tool to improve model fit



Always repeat diagnostic process after transformation

Confidence Band and Simultaneous Confidence Inference in SLR

The Simultaneous(as known as the Family or Joint) Confidence Interval Problem

1. Simultaneously (joint) estimation of p parameters, in SLR, $p = 2$ with β_0 and β_1 , with all X values.
 - Provide confidence that the conclusions for both β_0 and β_1 are correct.
2. Simultaneously estimation of g mean response \hat{Y}_h with g (usually different) X_h values.
 - Only one mean response value is made at one X_h value, $s^2\{\hat{Y}_h\} = s^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right]$
 - The g different mean responses made at different X_h values might not all be correct at the same time, even though all estimates are based on the same fitted regression line, depending on p parameters, e.g., b_0 and b_1 , $p = 2$ in SLR.
3. Simultaneously estimation of g single response $\hat{Y}_h\{new\}$ with g (usually different) X_h values
 - Only one single response value is made at one X_h value, $s^2\{pred\} = MSE \left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right] = s^2 + s^2\{\hat{Y}_h\}$

The Method

1. Simultaneously (joint) estimation of p parameters, in SLR, $p = 2$ with β_1 and β_2 , with **all** X values.
 - a) Bonferroni
 - b) Working-hoteling
2. Simultaneously estimation of mean response \hat{Y}_h with g different X_h values.
 - a) Bonferroni
 - b) Working-hoteling
3. Simultaneously estimation of single response $\hat{Y}_h\{new\}$ with g different X_h values
 - a) Bonferroni
 - b) Schefft

The Method

1. Simultaneously (joint) estimation of p parameters, in SLR, $p = 2$ with β_1 and β_2 , with **all** X values.
 - a) Bonferroni
 - b) Working-hoteling

The Individual Confidence Interval

- Consider to estimate β_0 and β_1 respectively
- The two individual intervals are

$$b_0 \pm t(1 - \alpha/2; n - 2)s\{b_0\} = I_0$$

$$b_1 \pm t(1 - \alpha/2; n - 2)s\{b_1\} = I_1$$

The event I_0^+ means the I_0 sucessfully contains the true intercept, β_0

The event I_1^+ means the I_1 sucessfully contains the true slope, β_1

$\Pr(I_0^+) = \Pr(I_1^+) = 1 - \alpha_i$, where α_i is used for individual interval

The event I_0^- and I_1^- define the complement events, and

$\Pr(I_0^-) = \Pr(I_1^-) = \alpha_i$

For example,

$$\Pr(I_0^+) = \Pr(I_1^+) = 0.95$$

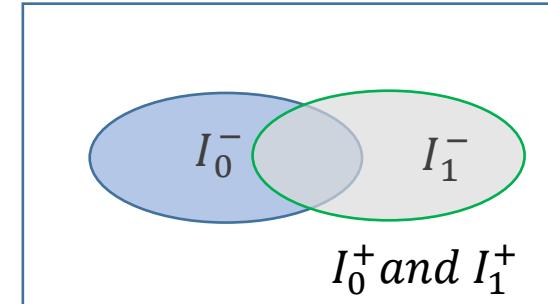
$$\Pr(I_0^-) = \Pr(I_1^-) = 0.05$$

Bonferroni Inequality

When $\Pr(I_0^+) = \Pr(I_1^+) = 0.95$, and $\Pr(I_0^-) = \Pr(I_1^-) = 0.05$

Q: what of the following is true about the probability that both individual intervals are correct, i.e., $= P(I_0^+ \text{ and } I_1^+)$.

- A) $0.95(0.95)$
- B) $1 - 2(0.05)$
- C) $1 - 2(0.05) + (0.05)(0.05)$
- D) $\geq 1 - 2(0.05)$



The answer is D). Note that A) or C) is true only when the individual estimates are independent.

$$P(I_0^+ \text{ and } I_1^+) = 1 - [P(I_0^-) + P(I_1^-) - P(I_0^- \text{ and } I_1^-)] \geq 1 - (P(I_0^-) + P(I_1^-)) = 1 - 2\alpha_i$$

$$\begin{aligned} P(I_0^+ \text{ and } I_1^+) &\geq 1 - 2\alpha_i \\ &= 1 - 2(0.05) \\ &= 1 - 0.1 = 0.9 \end{aligned}$$

- The α value for the joint confidence interval is at most 0.1, and the joint confidence level is at least 0.9
- This feature is called the Bonferroni Inequality.
- Bonferroni Inequality extends to k groups

$$P(I_1^+ \text{ and } I_2^+ \dots \text{ and } I_k^+) \geq 1 - k(\alpha_i)$$

The Bonferroni Joint Confidence Interval

$$P(I_0^+ \text{ and } I_1^+) \geq 1 - 2\alpha_i = 1 - \alpha, \quad \text{where } \alpha_i = \frac{\alpha}{2}$$

Thus, if we want the joint confidence interval $P(I_0^+ \text{ and } I_1^+)$ to be at least $1 - \alpha$, set the individual alpha value $\alpha_i = \frac{\alpha}{2}$, note that the **individual confidence level** is now at least $1 - \alpha_i = 1 - \frac{\alpha}{2}$

For MLR with p parameters, if we want $P(I_0^+ \text{ and } I_1^+ \dots \text{ and } I_{p-1}^+)$ to be at least $1 - \alpha$, set the individual alpha value $\alpha_i = \frac{\alpha}{p}$ such that the **joint confidence level**

$$P(I_0^+ \text{ and } I_1^+ \dots \text{ and } I_{p-1}^+) \geq 1 - p(\alpha_i) = 1 - \alpha .$$

This estimation is a **very conservative** because it overestimates the actual confidence level.

The Bonferroni Critical Value

The $1 - \alpha$ joint confidence interval is done by estimating β_0 and β_1 ($p = 2$) each with the individual confidence level of at least $1 - \frac{\alpha}{2}$, or the alpha value, $\alpha_i = \frac{\alpha}{2}$

We can now adjust the individual confidence interval to the Bonferroni Joint Confidence Interval:

$$b_0 \pm t\left(1 - \frac{\alpha_i}{2}, dfE\right) s\{b_0\} = b_0 \pm t\left(1 - \frac{\frac{\alpha}{2}}{2}, dfE\right) s\{b_0\} = b_0 \pm t(1 - \alpha/4, dfE) s\{b_0\}$$

$$b_1 \pm t\left(1 - \frac{\alpha_i}{2}, dfE\right) s\{b_1\} = b_1 \pm t\left(1 - \frac{\frac{\alpha}{2}}{2}, dfE\right) s\{b_1\} = b_1 \pm t(1 - \alpha/4, dfE) s\{b_1\}$$

In general, the Bonferroni Joint Confidence Interval for p parameters consists of p intervals,

$$b_k \pm t\left(1 - \frac{\alpha_i}{2}, dfE\right) s\{b_k\} = b_k \pm t\left(1 - \frac{\frac{\alpha}{p}}{2}, dfE\right) s\{b_k\} = b_k \pm t(1 - \alpha/2p, dfE) s\{b_k\}, \text{ where } k \text{ ranges from 0 to } p - 1.$$

The Bonferroni critical value is defined as

$$B = t(1 - \alpha/2p, dfE), \text{ or } t(1 - \frac{\alpha}{4}; n - 2) \text{ in SLR.}$$

Example: in the Toluca company case, compute the 90% family confidence interval for (β_0, β_1)

$$B = t\left(1 - \frac{\alpha}{4}; n - 2\right) = t(0.975, 23) = 2.069$$

$$b_0 \pm Bs\{b_0\} = 62.366 \pm 2.069 * 26.177 = (8.2, 116.5)$$

$$b_1 \pm Bs\{b_1\} = 3.570 \pm 2.069 * 0.347 = (2.85, 4.29)$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	62.366	26.177	2.382	0.0259 *
size	3.570	0.347	10.290	4.45e-10 ***

Residual standard error: 48.82 on 23 degrees of freedom
 Multiple R-squared: 0.8215, Adjusted R-squared: 0.8138
 F-statistic: 105.9 on 1 and 23 DF, p-value: 4.449e-10

- This is essentially setting the two individual confidence level as 95% for both β_0 and β_1 , to ensure the family have at least a 90% confidence level.

Some Example on the Bonferroni Critical Value Notation

$$\alpha = 0.1, \quad g = 2$$

$$B = t\left(1 - \frac{\alpha}{2g}; n - p\right) = t(1 - 0.025; n - p) = t(0.975, n - p)$$

$$\alpha = 0.05, \quad g = 2$$

$$B = t\left(1 - \frac{\alpha}{2g}; n - p\right) = t(1 - 0.0125; n - p) = t(0.9875, n - p)$$

$$\alpha = 0.15, \quad g = 3$$

$$B = t\left(1 - \frac{\alpha}{2g}; n - p\right) = t(1 - 0.025; n - p) = t(0.975, n - p)$$

$$\alpha = 0.1, \quad g = 3$$

$$B = t\left(1 - \frac{\alpha}{2g}; n - p\right) = t(1 - 0.0167; n - p) = t(0.983, n - p)$$

Working-hotelng Joint confidence interval (confidence band)

$$F = (\mathbf{b} - \boldsymbol{\beta}^*)'(\mathbf{X}'\mathbf{X})(\mathbf{b} - \boldsymbol{\beta}^*)/p\text{MSE},$$

$$F \sim F(p, n - p).$$

$F \leq F_\alpha$, Where $F_\alpha = F(1 - \alpha; p, n - p)$ is the $(1 - \alpha)100$ th percentile of the F – distribution $(p, n - p)$

$$(b_j - \beta_j)^2 \leq pF_\alpha \text{MSE} = W^2 \text{MSE} \quad \text{Where } W^2 = pF_\alpha(p, n - p)$$

which gives $b_j - Ws(b_j) \leq \beta_j \leq b_j + Ws(b_j)$ where $j = 1, 2, \dots, p$

In SLR, $p = 2$ $W^2 = pF_\alpha(p, n - p) = 2F_\alpha(2, n - p)$

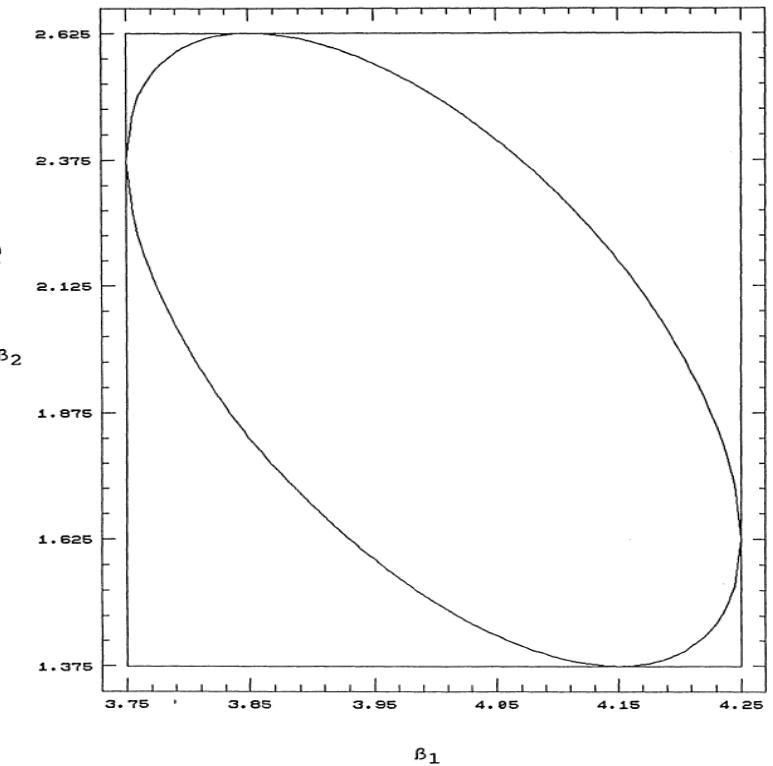


Figure 1. Exact Versus Conservative Confidence Regions. The plot compares a typical exact confidence ellipse with a conservative confidence rectangle for the case $p = 2$; that is, the case $Y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$. The projection of the ellipse onto the β_1 axis is the interval 3.75–4.25, whereas the projection of the ellipse onto the β_2 axis is 1.375–2.625.

Reference: David M. Nickerson, Construction of a conservative confidence region from projections of an exact confidence region in Multiple Linear Regression, The American Statistician, Vol. 48, No.2 (May, 1994)

Working hoteling confidence band on estimating (β_0, β_1)

- Working hoteling method is based on a F distribution: $W^2 = pF_\alpha(p, n - p)$

$$\beta_j \pm Ws\{\beta_j\} \quad \text{In SLR,} \quad j = 0 \text{ or } 1, \text{ and } W^2 = 2 F(1 - \alpha; 2, n - 2),$$

- In the Toluca case, find the joint 90% confidence interval for the parameters (β_0, β_1) .

$$W = \sqrt{2 F(1 - \alpha; 2, n - 2)} = 2.258$$

`sqrt(2*qf(0.9, 2, 23))`

$$b_0 \pm Ws\{b_0\} = 62.37 \pm W(26.18) = 62.3 \pm 59.11$$

$$b_1 \pm Ws\{b_1\} = 3.57 \pm W(0.347) = 3.57 \pm 0.78$$

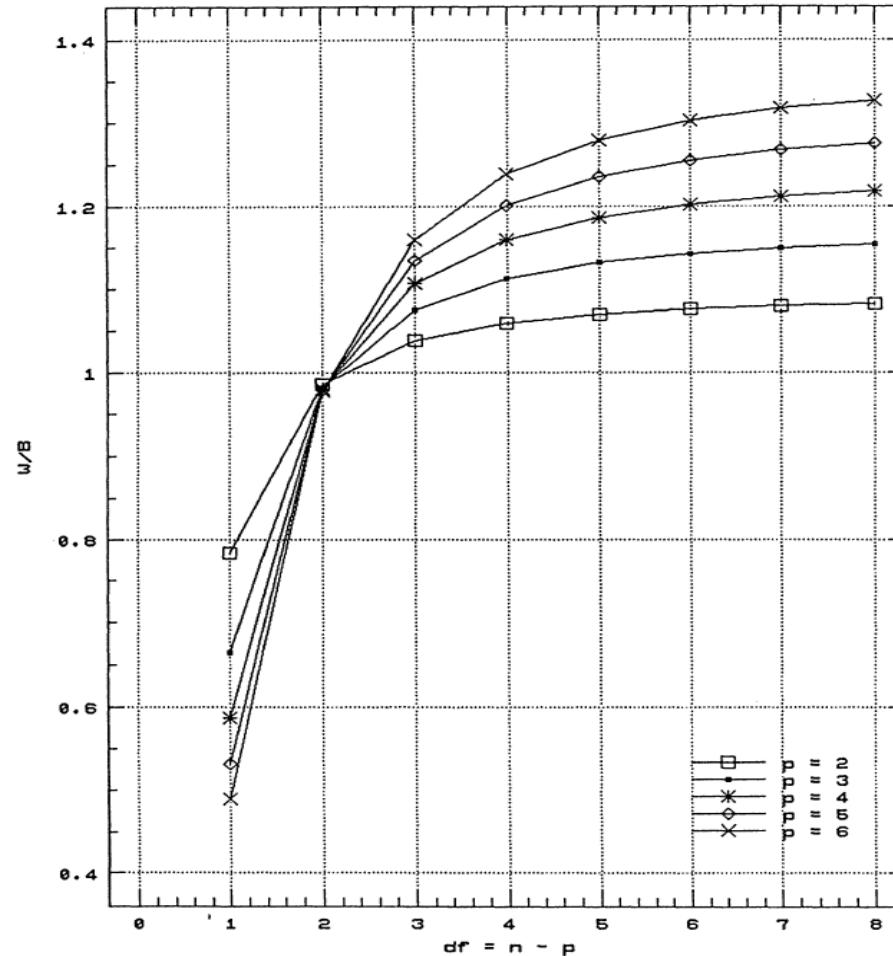
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	62.366	26.177	2.382	0.0259 *
size	3.570	0.347	10.290	4.45e-10 ***

Residual standard error: 48.82 on 23 degrees of freedom
 Multiple R-squared: 0.8215, Adjusted R-squared: 0.8138
 F-statistic: 105.9 on 1 and 23 DF, p-value: 4.449e-10

In this case of $n = 25, p = 2$, the Bonferroni procedure is better because the critical value is smaller.

Working hoteling confidence band on estimating (β_0, β_1)



- For $n - p \geq 2$, the Bonferroni procedure is better because $\frac{W}{B} > 1$
- This conclusion holds true when we try to estimate the **coefficients**, or the **linear impacts** (β) simultaneously in a linear model.

Figure 2. Working-Hotelling-Scheffè/Bonferroni Versus Degrees of Freedom and Dimension. This compares the ratio of the Working-Hotelling-Scheffè multiplier, W , with the Bonferroni multiplier, B , across values of $df = (n - p)$ for $p = 2, 3, 4, 5, 6$ at $\alpha = 0.10$.

The method

1. Simultaneously (joint) estimation of p parameters, in SLR, $p = 2$ with β_1 and β_2 , with **all** X values.
 - a) Bonferroni
 - b) Working-hoteling
2. Simultaneously estimation of mean response \hat{Y}_h with g different X_h values.
 - a) Bonferroni
 - b) Working-hoteling
3. Simultaneously estimation of single response $\hat{Y}_h\{\text{new}\}$ with g different X_h values
 - a) Bonferroni
 - b) Scheffe

Bonferroni Joint (or Family) Confidence Interval to predict the mean response, \hat{Y}_h $(\hat{Y}_h \text{ given } X_1, \hat{Y}_h \text{ given } X_2, \dots, \hat{Y}_h \text{ given } X_g)$

The Bonferroni procedure is very general. To make joint confidence interval for multiple (g) simultaneous prediction

For each mean response \hat{Y}_h for a given X_h

$$\hat{Y}_h \pm t \left(1 - \frac{\alpha}{2}; n - 2 \right) s\{\hat{Y}_h\} \quad s^2\{\hat{Y}_h\} = s^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right]$$

Then for the $1 - \alpha$ joint CI for g predictions, change the confidence level of each individual CI to be $1 - \alpha/g$

$$\hat{Y}_h \pm t \left(1 - \frac{\alpha}{2g}; n - 2 \right) s\{\hat{Y}_h\}$$

Note: if a sufficiently large number of simultaneous predictions are made, the width of the individual confidence Intervals may become so wide that they are no longer useful.

Toluca Company Example: Bonferroni Joint (or Family) Confidence Interval on Mean Response \hat{Y}_h

What is the simultaneous estimates for the mean number of work hours for $X_h = 30, 65 \text{ and } 100$ (i.e., $g = 3$) with family confidence level **0.9** ($\alpha = 0.1$). Suppose $\bar{X} = 70$, $SSX = 19800$, $s = 48.82$, $n = 25$

$$\hat{Y}_h \pm t\left(1 - \frac{\alpha}{2g}; n - 2\right) s\{\hat{Y}_h\} \quad X_h = 30, 65 \text{ and } 100 \quad (\hat{Y}_h = 169.5, 294.4, \text{ and } 419.4 \text{ respectively})$$

- Set the **confidence level** for each of the g individual estimate to be $1 - \frac{\alpha}{g} = 1 - \frac{0.1}{3} = 96.7\%$
- Then the **confidence level** for the family estimate is at least $1 - g * \frac{\alpha}{g} = 1 - 0.1 = 90\%$

Toluca Company Example: Bonferroni Joint (or Family) Confidence Interval on Mean Response \hat{Y}_h

What is the simultaneous estimates for the mean number of work hours for $X_h = 30, 65$ and 100 (i.e., $g = 3$) with family confidence level **0.9** ($\alpha = 0.1$). Suppose $\bar{X} = 70$, $SSX = 19800$, $s = 48.82$, $n = 25$

$$\hat{Y}_h \pm t\left(1 - \frac{\alpha}{2g}; n - 2\right)s\{\hat{Y}_h\} \quad X_h = 30, 65 \text{ and } 100 \quad (\hat{Y}_h = 169.5, 294.4, \text{ and } 419.4 \text{ respectively})$$

$$1. \quad X_h = 30 \quad s^2\{\hat{Y}_h\} = s^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] = 48.82^2\left[\frac{1}{25} + \frac{(30 - 70)^2}{s^2(n - 1)}\right] = 48.82^2\left[\frac{1}{25} + \frac{(30 - 70)^2}{19800}\right] = 288.169$$

$$\hat{Y}_h \pm t\left(1 - \frac{\alpha}{2g}; n - 2\right)s\{\hat{Y}_h\} = 169.5 \pm t\left(1 - \frac{0.1}{6}; 23\right)\sqrt{288.169} = 169.5 \pm 2.263(16.97) = (131.1, 207.9)$$

$$2. \quad X_h = 65 \quad \hat{Y}_h \pm t\left(1 - \frac{\alpha}{2g}; n - 2\right)s\{\hat{Y}_h\} = 294.4 \pm 2.263(9.92) = (272, 316.8)$$

$$3. \quad X_h = 100 \quad \hat{Y}_h \pm t\left(1 - \frac{\alpha}{2g}; n - 2\right)s\{\hat{Y}_h\} = 419.4 \pm 2.263(14.27) = (387.1, 451.7)$$

```
d<-data.frame(c(30,65,100))
ci.reg(tol.mod, d, type='b',alpha=0.1)
```

	size <dbl>	Fit <dbl>	Lower.Band <dbl>	Upper.Band <dbl>
1	30	169.4719	131.0570	207.8868
2	65	294.4290	271.9783	316.8797
3	100	419.3861	387.0774	451.6947

Working-Hoteling Joint Confidence Interval to predict the mean response, \hat{Y}_h

$(\hat{Y}_h \text{ given } X_1, \hat{Y}_h \text{ given } X_2, \dots, \hat{Y}_h \text{ given } X_g)$

- The WH procedure estimate β_0 and β_1 with a F distribution using the whole scale of X.
- The estimation of true mean $\mu_h = \beta_0 + \beta_1 X_h$ depends on the estimation of β_0 and β_1 and a given constant X_h
- The WH procedure for estimating mean for whatever X_h level is through a conservative estimation of β_0 and β_1
- For SLR, $\hat{Y}_h \pm W s\{\hat{Y}_h\}$ $W^2 = 2F(1 - \alpha; 2, n - 2)$
- For MLR, $\hat{Y}_h \pm W s\{\hat{Y}_h\}$ $W^2 = pF(1 - \alpha; p, n - p)$, where p is the number of parameters, not the number of predictions to make, i.e., g .
- Note that for MLR, the $s^2\{\hat{Y}_h\} = s^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right]$ formula should be modified.

Toluca Company Example: Working-Hoteling Confidence band on Mean Response \hat{Y}_h

What is the simultaneous estimates for the mean number of work hours for $X_h = 30, 65$ and 100 (i.e., $g = 3$) with family confidence level **0.9** ($\alpha = 0.1$). Suppose $\bar{X} = 70$, $SSX = 19800$, $s = 48.82$, $n = 25$

$X_h = 30, 65$ and 100 ,

$\hat{Y}_h = 169.5, 294.4$, and 419.4 respectively,

$s^2\{\hat{Y}_h\} = s^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right] = 288.169, 98.41$ and 203.63 respectively,

$Wc = 2.258$ (for any given X_h)

$$1. X_h = 30 \quad \hat{Y}_h \pm W s\{\hat{Y}_h\} = (131.15, 207.79)$$

$$2. X_h = 65 \quad \hat{Y}_h \pm W s\{\hat{Y}_h\} = (272.04, 316.82)$$

$$3. X_h = 100 \quad \hat{Y}_h \pm W s\{\hat{Y}_h\} = (387.16, 465.61)$$

```
sqrt(2*qf(0.9,2,23))
x<-data.frame(size=c(30,65,100))
ci.reg(toluca.mod, x, type='w',alpha=0.1) # Working hotelling
```

Lower.Band <dbl>	Upper.Band <dbl>
131.1542	207.7897
272.0351	316.8229
387.1591	451.6130

Bonferroni simultaneous CI vs Working hoteling confidence band on estimating mean \hat{Y}_h

Working hoteling procedure: $\hat{Y}_h \pm W s\{\hat{Y}_h\}$ where $W^2 = p F(1 - \alpha; p, n - p)$

Bonferroni procedure: $\hat{Y}_h \pm t\left(1 - \frac{\alpha}{2g}; n - 2\right) s\{\hat{Y}_h\}$

Compare the critical value with 90% confidence level, $dfE = 23$

$$W = 2.258 \text{ (for all } g\text{)}$$

$$B=t\left(1 - \frac{\alpha}{2g}; n - 2\right) = 1.714 \text{ (} g = 1 \text{)}$$

- WH is better than Bonferroni when $g \geq 3$

$$\begin{aligned} W &= \sqrt{pF(1 - \alpha, p, n - p)} \\ &= \sqrt{2F(0.9, 2, 23)} = 2.258 \end{aligned}$$

$$B=t(0.975, 23) = 2.069 \text{ (} g = 2 \text{)}$$

$$B=t(0.9833, 23) = 2.263 \text{ (} g = 3 \text{)}$$

$$B=t(0.99, 23) = 2.5 \text{ (} g = 5 \text{)}$$

$$B=t(0.995, 23) = 2.807 \text{ (} g = 10 \text{)}$$

Bonferroni simultaneous CI vs Working hoteling confidence band on estimating mean \hat{Y}_h

Comments:

- Both the WH and Bonferroni procedures provide wider bounds to the actual family confidence level.
- For larger families (g), the WH confidence band will be narrower since W stays the same for all g , while B gets larger.
- The levels for the predictor variable (X_h) are sometimes not known in advance. In such case, it is better to use the WH procedure because the family confidence interval encompasses all possible levels of X .
- The estimation is the better when X_h is closer to the mean, since $s_{\{\hat{Y}_h\}}^2 = s^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right]$

Bonferroni simultaneous CI vs Working hoteling confidence band on estimating mean \hat{Y}_h

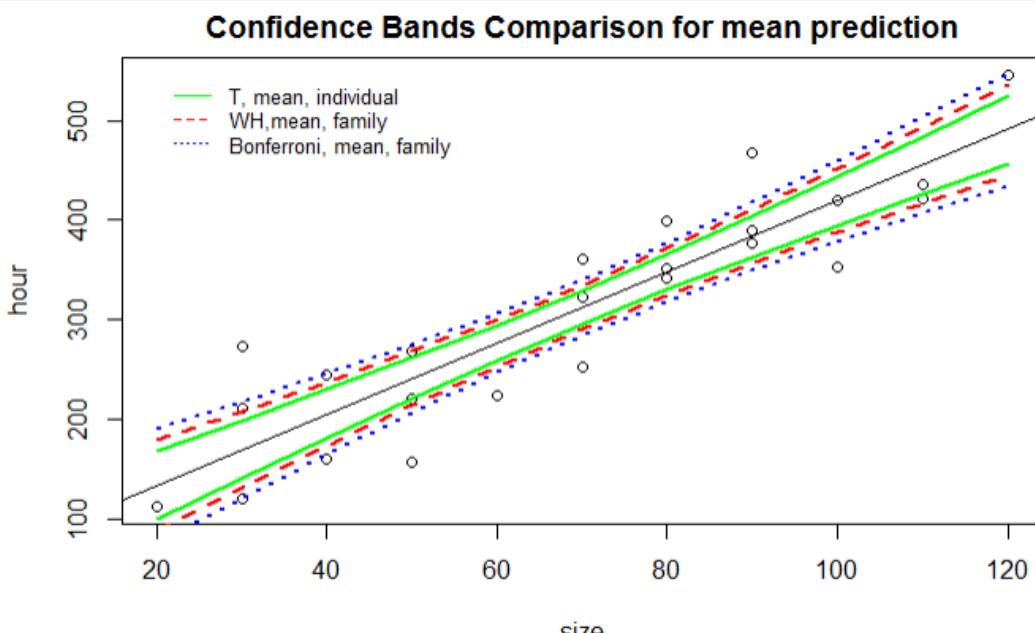
```

library(ALSM)
tol.mod<-lm(hour~size, toluca)
ox<-data.frame(size=unique(toluca$size))
x<-sort(ox$size)
x
[1] 20 30 40 50 60 70 80 90 100 110 120

plot(hour ~ size, toluca, main="Confidence Bands Comparison for mean prediction")
abline(lm(tol.mod))
lines(x, cim$Lower.Band, col="green", lwd=2, lty=1)
lines(x, cim$Upper.Band, col="green", lwd=2, lty=1)
lines(x, ciw$Lower.Band, col="red", lwd=2, lty=2)
lines(x, ciw$Upper.Band, col="red", lwd=2, lty=2)
lines(x, cib$Lower.Band, col="blue", lwd=2, lty=3)
lines(x, cib$Upper.Band, col="blue", lwd=2, lty=3)

legend(x=20, y=550, legend=c("T, mean, individual", "WH,mean, family","Bonferroni, mean, family"), lty=c(1,2,3), col=c("green","red","blue"),cex=0.8,
bty="n")

```



$$t \left(1 - \frac{\alpha}{2g}; n - 2 \right) = 1.714 \text{ } (g = 1)$$

$$W = 2.258 \text{ (for all } X_h \text{)}$$

$$t \left(1 - \frac{\alpha}{2g}; n - 2 \right) = 2.807 \text{ } (g = 10)$$

The method

1. Simultaneously (joint) estimation of p parameters, in SLR, $p = 2$ with β_1 and β_2 , with **all** X values.
 - a) Bonferroni
 - b) Working-hoteling
2. Simultaneously estimation of mean response \hat{Y}_h with g different X_h values.
 - a) Bonferroni
 - b) Working-hoteling
3. Simultaneously estimation of single response $\hat{Y}_h\{\text{new}\}$ with g different X_h values
 - a) Bonferroni
 - b) Schefft

Bonferroni Joint (or Family) Confidence Interval to predict the g single response, $\hat{Y}_h\{\text{new}\}$ ($\hat{Y}_h\{\text{new}\}$ given $X_1, \hat{Y}_h\{\text{new}\}$ given $X_2, \dots, \hat{Y}_h\{\text{new}\}$ given X_g)

To make joint confidence interval for multiple (g) simultaneous prediction for

For each \hat{Y}_h

$$\hat{Y}_h \pm t\left(1 - \frac{\alpha}{2}; n - 2\right) s\{\text{pred}\} \quad s^2\{\text{pred}\} = MSE[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}] = s^2 + s^2\{\hat{Y}_h\} \text{ in SLR}$$

Then for g prediction with the same X_h

$$\hat{Y}_h \pm t\left(1 - \frac{\alpha}{2g}; n - 2\right) s\{\text{pred}\}$$

Note: if a sufficiently large number of simultaneous predictions are made, the width of the individual confidence Intervals may become so wide that they are no longer useful.

Toluca Company Example: Bonferroni Joint (or Family) Confidence Interval on a single Response

What is the simultaneous estimates for the single prediction of number of work hours for $X_h = 30, 65$ and 100 with family confidence level **0.9** ($\alpha = 0.1$). Suppose $\bar{X} = 70$, $SSX = 19800$, $s = 48.82$, $n = 25$

$$\hat{Y}_h \pm t\left(1 - \frac{\alpha}{2g}; n - 2\right)s\{pred\} \quad X_h = 30, 65 \text{ and } 100 \quad (\hat{Y}_h = 169.5, 294.4, \text{ and } 419.4 \text{ respectively})$$

$$1. \quad X_h = 30 \quad s^2\{pred\} = s^2 + s^2\{\hat{Y}\} = 48.82^2 + 288.169 = 2671.56$$

$$\hat{Y}_h \pm t\left(1 - \frac{\alpha}{2g}; n - 2\right)s\{pred\} = 169.5 \pm t\left(1 - \frac{0.1}{6}; 23\right)\sqrt{2671.56} = 169.5 \pm 2.263(51.69) = (52.5, 286.5)$$

$$2. \quad X_h = 65 \quad \hat{Y}_h \pm t\left(1 - \frac{\alpha}{2g}; n - 2\right)s\{pred\} = 294.4 \pm 2.263(49.82) = (181.6, 407.2)$$

$$3. \quad X_h = 100 \quad \hat{Y}_h \pm t\left(1 - \frac{\alpha}{2g}; n - 2\right)s\{pred\} = 419.4 \pm 2.263(50.86) = (304.2, 534.5)$$

```
d<-data.frame(c(30,65,100))
ci.reg(tol.mod, d, type='gn',alpha=0.1)
```

size <dbl>	Fit <dbl>	Lower.Band <dbl>	Upper.Band <dbl>
30	169.4719	52.46349	286.4804
65	294.4290	181.64910	407.2089
100	419.3861	304.23781	534.5343

Scheffe procedure of Joint (or Family or simultaneous) Confidence Interval to predict the g single responses

$$\hat{Y}_h \pm S s\{pred\}$$

$$S^2 = gF(1 - \alpha; g, n - 2)$$

$$s_{\{pred\}}^2 = s_{\{\hat{Y}_h\}}^2 + s^2 = s^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} + 1 \right] \text{ in SLR.}$$

- For SLR, $\hat{Y}_h \pm S s\{pred\}$ $S^2 = gF(1 - \alpha; g, n - 2)$
- For MLR, $\hat{Y}_h \pm S s\{pred\}$ $S^2 = gF(1 - \alpha; g, n - p)$, where p is the number of parameters, and g is the number of predictions to make.
- In the WH procedure for estimating mean where all mean estimations are done through a conservative estimation of the p parameters, e.g. $E(\hat{Y}_h) = \beta_0 + \beta_1 X_h$, we only need to concern a combined distribution of p variables, e.g., b_0 and b_1
- On the other hand, in the Scheff procedure, the g single responses each has its own distribution. We need to concern the combined distribution of g variables, $Y_1, Y_2, Y_3, \dots, Y_g$.

Toluca Company Example: Scheffé Joint (or Family) Confidence Interval on g single Response \hat{Y}_h

What is the simultaneous estimates for the single number of work hours for $X_h = 30, 65$ and 100 (i.e., $g = 3$) with family confidence level **0.9** ($\alpha = 0.1$). Suppose $\bar{X} = 70$, $SSX = 19800$, $s = 48.82$

$\hat{Y}_h \pm S s\{pred\}$ $X_h = 30, 65$ and 100 ($\hat{Y}_h = 169.5, 294.4$, and 419.4 respectively)

1. $X_h = 30$

$$s^2\{pred\} = s^2 + s^2\{\hat{Y}_h\} = 48.82^2 + 288.169 = 2671.56 \quad S^2 = 3F(1 - 0.1; 3, n - 2) = 3(3.028) = 9.084 \Rightarrow S = 3.01$$

$$\hat{Y}_h \pm S s\{pred\} = 169.5 \pm 3.01(51.69) = (13.68, \quad 325.26)$$

2. $X_h = 65$

$$\hat{Y}_h \pm S s\{pred\} = 294.4 \pm 3.01(49.82) = (144.27, \quad 444.59)$$

3. $X_h = 100$

$$\hat{Y}_h \pm S s\{pred\} = 419.4 \pm 3.01(50.86) = (266.08, \quad 572.7)$$

```
d<-data.frame(c(30,65,100))
ci.reg(tol.mod, d, type='s',alpha=0.1)
```

- Unfortunately, the ci.reg function in R is wrong when computing this method.
- You can compute by hand or use the self-defined function on the next page.

	size <dbl>	Fit <dbl>
1	30	169.4719
2	65	294.4290
3	100	419.3861

	Lower.Band <dbl>	Upper.Band <dbl>
	-2502.216	2841.160
	-2187.645	2776.503
	-2168.029	3006.801

Family confidence interval, Bonferroni and Scheffe procedure

28

```
ci.sim <- function(model, newdata, type = c("B", "S"), alpha = 0.05)
{
  g <- nrow(newdata)
  CI <- predict(model, newdata, se.fit = TRUE)
  M <- ifelse(match.arg(type) == "B",
               qt(1 - alpha / (2*g), model$df),           # B "Bonferroni"
               sqrt(g * qf(1 - alpha, g, model$df)))      # S "scheffe"

  spred <- sqrt( CI$residual.scale^2 + (CI$se.fit)^2 )
  x <- data.frame(
    "x"      = newdata,
    "credV"  = M,
    "fit"    = CI$fit,
    "lower"  = CI$fit - M * spred,
    "upper"  = CI$fit + M * spred)

  return(x)
}

toluca<-read.table("U:/data/Toluca.txt", header=FALSE)
colnames(toluca)<-c("size","hour")

toluca.mod<-lm(hour~size, data=toluca)

new <- data.frame(size= c(30, 65, 100))
ci.sim(toluca.mod, new, type = "B")
ci.sim(toluca.mod, new, type = "S")
```

SLR in the Matrix Form

If you are not familiar with matrix algebra,
please read KNNL Sections 5.1-5.7.

Matrix form

- Matrices and vectors will be written in **bold face** type.
- Subscripts will indicate the *dimension* of the matrix. For example,

$$A_{3 \times 2} = \begin{bmatrix} 1 & 0.2 \\ 1 & 3.4 \\ 1 & 2.1 \end{bmatrix}$$

is a 3-row by 2-column matrix.

- Dimension subscripts will only be used when needed for clarity.

Warm up exercise

$$\mathbf{A}_{2 \times 2} = \begin{bmatrix} 2 & 5 \\ 4 & 1 \end{bmatrix} \quad \mathbf{B}_{2 \times 2} = \begin{bmatrix} 4 & 6 \\ 5 & 8 \end{bmatrix}$$

A) $\mathbf{AB} = \begin{bmatrix} 8 & 30 \\ 2 & 28 \end{bmatrix}$

B) $\mathbf{AB} = \begin{bmatrix} 38 \\ 28 \end{bmatrix}$

C) $\mathbf{AB} = \begin{bmatrix} 38 \\ 28 \end{bmatrix}$

D) $\mathbf{AB} = \begin{bmatrix} 33 & 52 \\ 21 & 32 \end{bmatrix}$

Warm up exercise

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_i \end{bmatrix}$$

$$\mathbf{Y}'\mathbf{Y} = \Sigma y_i^2$$

$$\mathbf{Y}' \begin{bmatrix} 1..1 \\ 1..1 \\ 1..1 \end{bmatrix} \mathbf{Y} = (\Sigma y_i)^2$$

SST = $\Sigma y_i^2 - (\Sigma y_i)^2 = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}' \begin{bmatrix} 1..1 \\ 1..1 \\ 1..1 \end{bmatrix} \mathbf{Y}$

The SLR Model in Scalar Form

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \text{where} \quad \varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

Consider now writing an equation for each observation:

$$Y_1 = \beta_0 + \beta_1 X_1 + \varepsilon_1$$

$$Y_2 = \beta_0 + \beta_1 X_2 + \varepsilon_2$$

$$\vdots \quad \vdots \quad \vdots$$

$$Y_n = \beta_0 + \beta_1 X_n + \varepsilon_n$$

The SLR Model in Matrix Form

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

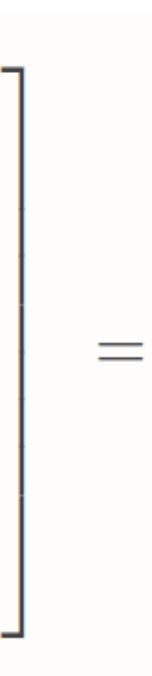
In matrix notation, the simple linear regression model is written as,

$$Y_{n \times 1} = X_{n \times 2} \beta_{2 \times 1} + \varepsilon_{n \times 1}$$

Or more simply as,

- Y is the *response vector*
- X is called the *design matrix*
- β is the *parameter vector*
- ε is the *error vector*

$$Y = X \beta + \varepsilon$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$


Example (Flavor deterioration). The results shown below were obtain in a small-scale experiment to study the relation between F of storage temperature (X) and number of weeks before flavor deterioration of a food product begins occur (Y).

Observation	1	2	3	4	5
Temp (X)	8	4	0	-4	-8
Week (Y)	7.8	9	10.2	11	11.7

Write the matrix notation for X and Y .

$$Y = X \beta + \varepsilon$$

```
week<-c(7.8,9,10.2,11,11.7)
```

```
[1] 7.8 9.0 10.2 11.0 11.7
```

```
week
```

```
y<-as.matrix(week)
colnames(y)<-c("week")
y
```

```
week
[1,] 7.8
[2,] 9.0
[3,] 10.2
[4,] 11.0
[5,] 11.7
```

```
temp<-c(8,4,0,-4,-8)
Intercept<-rep(1,5)
x<-cbind(Intercept,temp)
x
```

```
Intercept temp
[1,] 1 8
[2,] 1 4
[3,] 1 0
[4,] 1 -4
[5,] 1 -8
```

Example (Flavor deterioration). The results shown below were obtain in a small-scale experiment to study the relation between F of storage temperature (X) and number of weeks before flavor deterioration of a food product begins occur (Y).

Observation	1	2	3	4	5
Temp (X)	8	4	0	-4	-8
Week (Y)	7.8	9	10.2	11	11.7

Use the matrix notation to compute SST.

$$SST = \sum(Y - \bar{Y})^2 = \sum Y^2 - \frac{(\sum Y)^2}{n} = \mathbf{y}^t \mathbf{y} - \frac{1}{n} \mathbf{y}^t \mathbf{J} \mathbf{y} = 9.752$$

Where \mathbf{J} is the n by n matrix of 1s

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

$$\mathbf{y} = \begin{pmatrix} \text{week} \\ 7.8 \\ 9.0 \\ 10.2 \\ 11.0 \\ 11.7 \end{pmatrix}$$

```
J<-matrix(1,nrow=5,ncol=5)
SST<-t(y)%*%y-(1/5)*t(y)%*%J%*%y
```

Introducing the Variance-Covariance Matrix for a random multivariable, \mathbf{U}

- $\mathbf{U} = \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ \dots \\ U_n \end{bmatrix}$ is a multivariate. The mean, $E\{\mathbf{U}\} = \begin{bmatrix} E(U_1) \\ E(U_2) \\ E(U_3) \\ \dots \\ E(U_n) \end{bmatrix}$

- The variance,

$$Var(U_i) = \sigma^2(U_i) = E[(U_i - E(U_i))(U_i - E(U_i))'] = E(U_i^2) - [E(U_i)]^2$$

- The covariance of two random multivariate, \mathbf{U} and \mathbf{V} , describes the linear relationship between U and V

$$\text{Cov}(U_i, U_j) = \sigma(U_i, U_j) = E[(U_i - E(U_i))(U_j - E(U_j))] = E(U_i U_j) - E(U_i)E(U_j)$$

- The variance-Covariance Matrix is a n-by-n matrix consisting both variance and covariance in the multivariate.

$$\Sigma\{\mathbf{U}\}_{n \times n} = \begin{bmatrix} Var(U_i), Cov(U_i, U_j) \\ Cov(U_j, U_i), Var(U_j) \end{bmatrix}$$

Features on Normal Variable

- Usually, independent variables are always uncorrelated, but uncorrelated variables are not necessarily independent. That is, correlation only measures dependence in the linear dimension.
- Except when the variables follow Normal distribution, in which case correlation and dependence are the same.
- If variables are (completely) uncorrelated, their covariance is 0.
- The variance-covariance matrix of uncorrelated variables will therefore be a diagonal matrix , since all the covariances are 0.

$$\Sigma\{\mathbf{U}\}_{n \times n} = \begin{bmatrix} Var(U_i), 0 \\ 0, Var(U_j) \end{bmatrix}$$

More General Notations

Let $\mathbf{U} \sim N(\mathbf{E}(\mathbf{U}), \Sigma(\mathbf{U}))$ be a multivariate normal vector,
and let $\mathbf{V} = \mathbf{c} + \mathbf{D} \mathbf{U}$ be a linear transformation of \mathbf{U}
where \mathbf{c} is a vector of constants and \mathbf{D} is a matrix of constants.

Then $\mathbf{V} \sim N(\mathbf{c} + \mathbf{D} \boldsymbol{\mu}, \mathbf{D} \Sigma \mathbf{D}^t)$.

The Variance-Covariance Matrix for the Random Error, $\Sigma\{\varepsilon\}$

$$\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix} \sim \text{Normal}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma^2 & \dots & 0 \\ \vdots & & & \ddots & \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix}\right)$$

$$\Sigma\{\varepsilon\}_{n \times n} = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma^2 & \dots & 0 \\ \vdots & & & \ddots & \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 I_{n \times n}$$

This is true when the random errors are

- independent,
- have a mean of 0, and
- a constant variance

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} \sim \text{Normal}\left(\begin{bmatrix} X_1\beta \\ X_2\beta \\ X_3\beta \\ \vdots \\ X_n\beta \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma^2 & \dots & 0 \\ \vdots & & & \ddots & \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix}\right)$$

$$\Sigma\{Y\}_{n \times n} = \Sigma\{\varepsilon\}_{n \times n}$$

- $Y = X\beta + \varepsilon = X\beta + I\varepsilon$, and
- No assumption violation

Least Squares Parameter Estimation

We want to minimize the sum of residuals.

To find the solution, set the derivative with respect to the vector β equal to a zero vector and solve:

$$\begin{aligned}\frac{d}{d\beta} (\varepsilon^t \varepsilon) &= \frac{d}{d\beta} ((Y - X\beta)^t (Y - X\beta)) \\ &= -2X^t(Y - X\beta)\end{aligned}$$

(Partially) Solving for β yields the so-called *Normal Equations* :

$$\begin{aligned}-2X^t(Y - X\beta) &= \mathbf{0} \\ X^t Y &= X^t X \beta\end{aligned}$$

$$\widehat{\beta} = (X^t X)^{-1} (X^t Y) = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} (\bar{Y}) - \frac{(\bar{X})SS_{XY}}{SS_X} \\ \frac{SS_{XY}}{SS_X} \end{bmatrix}$$

The details

$$\beta = (X^t X)^{-1} (X^t Y) = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

$$X^t X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_2 & \dots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} = \begin{bmatrix} n & \Sigma X_i \\ \Sigma X_i & \Sigma X_i^2 \end{bmatrix}$$

$$(X^t X)^{-1} = \frac{1}{n\Sigma X_i^2 - (\Sigma X_i)^2} \begin{bmatrix} \Sigma X_i^2 & -\Sigma X_i \\ -\Sigma X_i & n \end{bmatrix} = \frac{1}{nSS_X} \begin{bmatrix} \Sigma X_i^2 & -\Sigma X_i \\ -\Sigma X_i & n \end{bmatrix}$$

$$X^t Y = \begin{bmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_2 & \dots & X_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \Sigma Y_i \\ \Sigma X_i Y_i \end{bmatrix}$$

Plug these into the equation for b:

$$\begin{aligned}
 \mathbf{b} &= (\mathbf{X}^t \mathbf{X})^{-1} (\mathbf{X}^t \mathbf{Y}) \\
 &= \frac{1}{nSS_X} \begin{bmatrix} \Sigma X_i^2 & -\Sigma X_i \\ -\Sigma X_i & n \end{bmatrix} \begin{bmatrix} \Sigma Y_i \\ \Sigma X_i Y_i \end{bmatrix} \\
 &= \frac{1}{nSS_X} \begin{bmatrix} (\Sigma X_i^2)(\Sigma Y_i) - (\Sigma X_i)(\Sigma X_i Y_i) \\ -(\Sigma X_i)(\Sigma Y_i) + n(\Sigma X_i Y_i) \end{bmatrix} \\
 &= \frac{1}{SS_X} \begin{bmatrix} (\Sigma X_i^2)(\bar{Y}) - (\bar{X})(\Sigma X_i Y_i) \\ -n(\bar{X})(\bar{Y}) + (\Sigma X_i Y_i) \end{bmatrix} \\
 &= \frac{1}{SS_X} \begin{bmatrix} (\Sigma X_i^2)(\bar{Y}) - \bar{Y}(n \bar{X}^2) + \bar{X}(n \bar{X} \bar{Y}) - (\bar{X})(\Sigma X_i Y_i) \\ SS_{XY} \end{bmatrix} \\
 &= \frac{1}{SS_X} \begin{bmatrix} SS_X(\bar{Y}) - (\bar{X})SP_{XY} \\ SS_{XY} \end{bmatrix} = \begin{bmatrix} (\bar{Y}) - \frac{(\bar{X})SS_{XY}}{SS_X} \\ \frac{SS_{XY}}{SS_X} \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \text{ Finally!}
 \end{aligned}$$

Example (Flavor deterioration). The results shown below were obtain in a small-scale experiment to study the relation between F of storage temperature (X) and number of weeks before flavor deterioration of a food product begins occur (Y).

i	1	2	3	4	5
Xi	8	4	0	-4	-8
Yi	7.8	9	10.2	11	11.7

Compute estimators in matrix form

$$Y_{5 \times 1} = \begin{bmatrix} \text{week} \\ 7.8 \\ 9.0 \\ 10.2 \\ 11.0 \\ 11.7 \end{bmatrix}$$

```
xtx<-t(x)%*%x
xtxinv<-solve(xtx)
xtxinv
```

Intercept	temp
0.2	0.00000
0.0	0.00625

$$X_{5 \times 2} = \begin{bmatrix} \text{Intercept} & \text{temp} \\ 1 & 8 \\ 1 & 4 \\ 1 & 0 \\ 1 & -4 \\ 1 & -8 \end{bmatrix}$$

```
xty<-t(x)%*%y
xty
```

week
49.7
-39.2

$$\hat{\beta} = (X'X)^{-1}(X'Y) = \begin{bmatrix} 9.94 \\ -0.245 \end{bmatrix}$$

The Hat Matrix

$$\begin{aligned}\hat{Y} &= X b \\ &= X (X^t X)^{-1} X^t Y \\ &= H Y\end{aligned}$$

where $H = X (X^t X)^{-1} X^t$ is called the *hat matrix* because it turns Y 's into \hat{Y} 's.

The hat matrix will give us many useful diagnostic tools in MLR.

In the flavor example

$$H = \begin{matrix} 5 \times 5 & \begin{array}{ccccc} 0.6 & 0.4 & 0.2 & 0.0 & -0.2 \\ 0.4 & 0.3 & 0.2 & 0.1 & 0.0 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.0 & 0.1 & 0.2 & 0.3 & 0.4 \\ -0.2 & 0.0 & 0.2 & 0.4 & 0.6 \end{array} \end{matrix}$$

The matrix H is symmetric and has the special property called **idempotent**:

$$HH = H$$

Predicted value, residual, SSE, and MSE

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \begin{bmatrix} b_0 + b_1 X_1 \\ b_0 + b_1 X_2 \\ \vdots \\ b_0 + b_1 X_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \mathbf{X}\mathbf{b}$$

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

$$= \begin{bmatrix} -0.18 \\ 0.04 \\ 0.26 \\ 0.08 \\ -0.20 \end{bmatrix}$$

$$\begin{aligned} SSE &= \mathbf{e}' \mathbf{e} = \begin{bmatrix} -0.18 & 0.04 & 0.26 & 0.08 & -0.20 \end{bmatrix} \begin{bmatrix} -0.18 \\ 0.04 \\ 0.26 \\ 0.08 \\ -0.20 \end{bmatrix} \\ &= 0.148 \end{aligned}$$

$$MSE = SSE/(n - p) = 0.0493$$

The Matrix Form of the coefficients

$$\mathbf{b} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y}$$

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

Since \mathbf{b} is a linear combination of \mathbf{Y} , \mathbf{b} follows Normal distribution.

$$\begin{aligned} E(\mathbf{b}) &= (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t E(\mathbf{Y}) \\ &= (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{X} \boldsymbol{\beta} = \boldsymbol{\beta} \end{aligned}$$

Therefore, \mathbf{b} is an unbiased estimator of $\boldsymbol{\beta}$.

The variance-covariance matrix for the coefficient

$$\begin{aligned}
 \Sigma_{\{b\}} &= [(X'X)^{-1}X'] \sigma^2 I [(X'X)^{-1}X']' \\
 &= \sigma^2 [(X'X)^{-1}X'] I [(X'X)^{-1}X']' \\
 &= \sigma^2 (X'X)^{-1} (X'X) [(X'X)^{-1}]' \\
 &= \sigma^2 (X'X)^{-1}
 \end{aligned}$$

- In the last step, $X'X$ and its inverse are both symmetric, therefore
 $(X'X)^{-1t} = (X'X)^{-1}$
- $\sigma^2(X'X)^{-1}$ generally is **NOT** a diagonal matrix, because the estimates b_0 and b_1 are generally **not** independent of each other.
- $\Sigma\{b\}$ has a dimension of p by p, where p is the number of parameters.

Example (Flavor deterioration).

Compute the covariance matrix for $\hat{\beta}$

$$Y = \begin{matrix} \text{week} \\ 7.8 \\ 9.0 \\ 10.2 \\ 11.0 \\ 11.7 \end{matrix}$$

$$\hat{\beta} = (X^t X)^{-1} (X^t Y) = \begin{matrix} \text{Intercept} \\ \text{temp} \end{matrix} \quad \begin{matrix} \text{week} \\ 9.940 \\ -0.245 \end{matrix}$$

$$\Sigma\{\hat{\beta}\} = \sigma^2 (X^t X)^{-1} = MSE(X^t X)^{-1}$$

$$= 0.0493 \begin{pmatrix} 0.2 & 0.00000 \\ 0.0 & 0.00625 \end{pmatrix} = \begin{pmatrix} 0.009866667 & 0.000000000 \\ 0.000000000 & 0.0003083333 \end{pmatrix}$$

$$X = \begin{matrix} \text{Intercept} & \text{temp} \\ 1 & 8 \\ 1 & 4 \\ 1 & 0 \\ 1 & -4 \\ 1 & -8 \end{matrix}$$

- The diagonal elements in $\Sigma\{\hat{\beta}\}$ are $s^2\{b_0\}$ and $s^2\{b_1\}$
- The non-diagonal elements are generally not 0 due to dependency.

The Matrix Form of the Mean Response at One X

To estimate the mean response at X_h , define a matrix form

$$\mathbf{X}_h = \begin{bmatrix} 1 \\ X_h \end{bmatrix}$$

$$\hat{Y}_h = \mathbf{X}'_h \mathbf{b} = [1 \quad X_h] \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = [b_0 + b_1 X_h]$$

$$\Sigma\{\hat{Y}_h\} = X'_h \Sigma\{\mathbf{b}\} X_h$$

Example (Flavor deterioration): find the point estimator and standard error of \hat{Y}_h when $X_h = -6$

$$\hat{Y}_h = \mathbf{X}'_h \mathbf{b} = [1, -6] \begin{bmatrix} 9.94 \\ -0.245 \end{bmatrix} = 11.41$$

$$\begin{aligned} \Sigma\{\hat{Y}_h\} &= X'_h \Sigma\{\mathbf{b}\} X_h = [1 \quad -6] \begin{bmatrix} 0.00986 & 0 \\ 0 & 0.00031 \end{bmatrix} \begin{bmatrix} 1 \\ -6 \end{bmatrix} \\ &= 0.021 \\ &= s^2\{\hat{Y}_h\} \end{aligned}$$

- The variance covariance matrix of the mean response has a dimension of 1 by 1, and reduces to $s^2\{\hat{Y}_h\}$
- Note that the X_h matrix differs from the design matrix!

The Matrix Form of the Mean Response at Multiple Xs

To estimate the mean response at m multiple points, define a 2 by m matrix form for X_h

Example (Flavor deterioration): find the point estimator and standard error of \hat{Y}_h when $X = -6$ and -5

$$X_h = \begin{bmatrix} 1 & 1 \\ -6 & -5 \end{bmatrix} \quad \hat{Y}_h = X'_h \mathbf{b} = \begin{bmatrix} 1 & -6 \\ 1 & -5 \end{bmatrix} \begin{bmatrix} 9.94 \\ -0.245 \end{bmatrix} = \begin{bmatrix} 11.41 \\ 11.17 \end{bmatrix}$$

$$\Sigma\{\hat{Y}_h\} = X'_h \Sigma\{\mathbf{b}\} X_h = \begin{bmatrix} 1 & -6 \\ 1 & -5 \end{bmatrix} \begin{bmatrix} 0.00986 & 0 \\ 0 & 0.00031 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -6 & -5 \end{bmatrix} = \begin{bmatrix} 0.021 & * \\ * & 0.018 \end{bmatrix}$$

- The diagonal values, $s^2\{\hat{Y}_h\} = 0.021$ and 0.018 for the $X=-6$ and -5 , respectively.
 - Use the $s^2\{\hat{Y}_h\}$ to compute the confidence interval for the mean response: $\hat{Y}_h \pm ts\{\hat{Y}_h\}$
 - Use the $s^2\{\hat{Y}_h\}$ to compute the variance (and the CI) for the single response: $s^2\{\text{pred}\} = s^2 + s^2\{\hat{Y}_h\}$
- The off-diagonal values, or the “*” part is not applicable under the assumption of independence.
- $\Sigma\{\hat{Y}_h\}$ has a dimension of m by m , where m is the number of X levels to predict.

Extending SLR to MLR through the Matrix Form

Overview of Multiple Linear Regression (MLR)

Data for Multiple Regression

- Y_i is the response variable (as usual)
- $X_{i,1}, X_{i,2}, \dots, X_{i,p-1}$ are the $p - 1$ explanatory variables for cases $i = 1$ to n .
- Example – In Homework #1 you modeled GPA as a function of entrance exam score. We could also consider an aptitude test and high school GPA as potential predictors. With the entrance exam score, this would be 3 variables, so $p = 4$.
- *Potential problem to remember!!!* These predictor variables are probably correlated with each other.

The Multiple Regression Model in Scalar Form (MLR: multiple linear regression)

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \dots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \text{ for } i = 1, 2, \dots, n$$

where

- Y_i is the value of the response variable for the i th case.
- $\varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$ (exactly as before!)
- β_0 is the intercept (think multidimensionally and look at the equation).
- $\beta_1, \beta_2, \dots, \beta_{p-1}$ are the regression coefficients for the explanatory variables.
- $X_{i,k}$ is the value of the k th explanatory variable for the i th case.

Special Cases

- *Polynomial model*

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \dots + \beta_{p-1} X_i^{p-1} + \varepsilon_i$$

- *Interactions* between explanatory variables are expressed as a product of the X 's:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,1} X_{i,2} + \varepsilon_i$$

- *ANOVA Models* with discrete predictors can be encoded by defining the X 's as *indicator* or *dummy variables* where $X_{i,k} = 1$ if case i belongs to the k -th group, and $X = 0$ otherwise.

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,1} X_{i,2} + \varepsilon_i$$

- *Linear model vs Nonlinear model (not covered in the course)*

e.g. $Y_i = \beta_0 \exp(\beta_1 X_i) + \varepsilon_i$

Multiple Regression Model in Matrix Form

$$\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$$

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n \times n})$$

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

Design Matrix \mathbf{X} :

$$\mathbf{X} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,p-1} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \ddots & & \vdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,p-1} \end{bmatrix}$$

Matrix Forms for the MLR are Identical to the SLR

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \dots \\ Y_n \end{bmatrix} \sim \text{Normal}\left(\begin{bmatrix} X_1\beta \\ X_2\beta \\ X_3\beta \\ \dots \\ X_n\beta \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma^2 & \dots & 0 \\ \dots & & & \dots & \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix}\right)$$

$$\Sigma\{Y\}_{n \times n} = \Sigma\{\varepsilon\}_{n \times n} = \sigma^2 I$$

- $Y = X\beta + \varepsilon = X\beta + I\varepsilon$, and
- No assumption violation

$$e = Y - \hat{Y} = (I - H)Y$$

$$b = (X^t X)^{-1} X^t Y$$

$$\Sigma\{b\}_{p \times p} = \sigma^2 (X' X)^{-1} = \text{MSE} (X' X)^{-1}$$

$$\hat{Y}_h = X'_h b$$

$$\Sigma\{\hat{Y}_h\} = X'_h \Sigma\{b\} X_h$$

Matrix Forms for the Residuals

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

$$\begin{aligned}\Sigma\{\mathbf{e}\}_{n \times n} &= (\mathbf{I} - \mathbf{H})\Sigma\{\mathbf{Y}\}(\mathbf{I} - \mathbf{H})' \\ &= (\mathbf{I} - \mathbf{H})\sigma^2 \mathbf{I}(\mathbf{I} - \mathbf{H})' = \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})' \\ &= \sigma^2(\mathbf{I} - \mathbf{H}) = \text{MSE}(\mathbf{I} - \mathbf{H})\end{aligned}$$

$\sigma^2(e_i) = \text{MSE}(1 - h_{ii})$, where h_{ii} is the i-th diagonal element of \mathbf{H} .

$h_{ii} = \mathbf{X}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i$ and $\mathbf{X}'_i = (1 \ X_{i1}, \dots, X_{ip-1})$ is taken from the i-th data point.

The covariance $\sigma(e_i, e_j) = \text{MSE}(-h_{ii})$ is usually not 0, but we can ignore this with a reasonably large n.

In the flavor example,

$$\text{MSE}(\mathbf{I} - \mathbf{H}) = 0.0493$$

0.4	-0.4	-0.2	0.0	0.2
-0.4	0.7	-0.2	-0.1	0.0
-0.2	-0.2	0.8	-0.2	-0.2
0.0	-0.1	-0.2	0.7	-0.4
0.2	0.0	-0.2	-0.4	0.4

The ANOVA Table is Identical to the SLR

Source	df	SS	MSE	F
Model	$df_M = p - 1$	SSM	MSM	$\frac{MSM}{MSE}$
Error	$df_E = n - p$	SSE	MSE	
Total	$df_T = n - 1$	SST		

The Global F test or the Significant Test

$H_0 : \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$ (β_0 is not in this list!)

$H_A : \beta_k \neq 0$, for at least one $k = 1, \dots, p - 1$

$$F^* = M\text{SM}/M\text{SE}$$

Under H_0 , $F^* \sim F_{p-1, n-p}$.

We test it the usual way (reject H_0 if $p \leq \alpha$).

Interpreting the p -value of the Global F -test

If the p -value for the F -test . . .

- is $> \alpha$, we lack evidence to conclude that *any* of our explanatory variables can help to predict or explain the response variable using a linear regression model.
- is $\leq \alpha$, *one or more* of the explanatory variables in our model *is* potentially useful for predicting the response in a linear model (but F does not say which ones).

Coefficient of Multiple Determination, R^2

As in SLR, R^2 is the proportion of variation in the response explained by the model. R^2 and F s are related.

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

Note that “explained by the model” means “explained by these predictors using this specific regression equation,” *NOT* just “explained by these predictor variables.”

Finally, a large value of R^2 does not necessarily imply that the fitted model is a useful one.

Adjusted Coefficient of Determination, R^2_{adj}

It is sometimes suggested that a modified measure be used that adjusts for the number of X variables in the model.

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

- $1/df_E = 1/(n - p)$ increases as a hyperbolic function of p .
- Increasing p by 1 *always* makes SSE smaller, but not always by the same amount.
- If the decline in SSE is large enough to cancel the increase in $1/df_E$, then MSE will get smaller, so R^2 will get bigger.
- If the fit does not improve sufficiently to overcome $1/df_E$, then R^2_{Adj} will remain the same, or might even get smaller!

The T test on the Individual Regression Coefficients (Parameters)

As usual, the CI for β_k is,

$$b_k \pm t_c s_{\{b_k\}}, \text{ where } t_c = t_{n-p}(1 - \alpha/2)$$

We know that $\mathbf{b} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^t \mathbf{X})^{-1})$

We estimate the variance-covariance matrix for the parameter vector as,

$$\begin{aligned}\Sigma\{b\}_{p \times p} &= MSE (\mathbf{X}^t \mathbf{X})^{-1} \\ &= \left(\frac{1}{n-p} \right) Y'(I - H)Y (\mathbf{X}^t \mathbf{X})^{-1}\end{aligned}$$

For an individual coefficient β_k , where $k = (0, \dots, p-1)$,

$s_{\{b_k\}}^2$ is the $(k+1)$ -th diagonal element of the variance-covariance matrix $\Sigma\{b\}_{p \times p}$

The T test on the Individual Regression Coefficients (Parameters)

- The hypothesis test is defined as $H_0: \beta_k = \beta_k^*$. By default, $\beta_k^* = 0$ and two-sided.
- This tests the significance of this β_k , given all other β_s in the model.

For example: $H_0: \beta_3 = 0 | \beta_1, \beta_2, \beta_4$ in the model $H_a: \beta_3 \neq 0 | \beta_1, \beta_2, \beta_4$ in the model

- The test statistic is $t_s = \frac{b_k}{s\{b_k\}} \sim t(n - p)$

$$ts^2 \neq Fs = \frac{MSR}{MSE}$$
- The result of the Significant test of beta could be misleading when the impact of X_k overlaps with other predictors.

The Dwaine Studios example: The Dwaine operates studios that specialize in portraits of children. The company is considering whether sales (Y) in a community can be predicted from the number of persons aged 16 or younger in the community (X1) and the per capita disposable personal income in the community (X2). Data is Dwaine.csv, n=21, p=3

X1	X2	Y
68.5	16.7	174.4
45.2	16.8	164.4
91.3	18.2	244.2
47.8	16.3	154.6
46.9	17.3	181.6
66.1	18.2	207.5
49.5	15.9	152.8
52	17.2	163.2
48.9	16.6	145.4
38.4	16	137.2
87.9	18.3	241.9
72.8	17.1	191.1
88.4	17.4	232
42.9	15.8	145.3
52.5	17.8	161.1
85.7	18.4	209.7
41.3	16.5	146.4
51.7	16.3	144
89.6	18.1	232.6
82.7	19.1	224.1
52.3	16	166.5

The Dwaine Studios example: The Dwaine operates studios that specialize in portraits of children. The company is considering whether sales (Y) in a community can be predicted from the number of persons aged 16 or younger in the community (X1) and the per capita disposable personal income in the community (X2). Data is Dwaine.csv, n=21, p=3

1). Predict the mean Y when X1=65.4 and X2=17.6

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 \\ = 191$$

2). Find SSE, df_E , SSR, df_R , MSE, MSR, R^2 , R_{adj}^2

$$SSE = 2180.9 \quad df_E = 18 \quad MSE = 121.2$$

$$SSR = 23371.8 + 643.5 = 24015.3, \quad df_R = p-1 = 2$$

$$MSR = 24015.3/2 = 12007.6$$

$$SST = SSR + SSE = 26195.3$$

$$R^2 = SSR/SS = 0.9167$$

$$R_{adj}^2 = 1 - \left(\frac{n-1}{n-p} \right) SSE/SST = 0.9075$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-68.8571	60.0170	-1.147	0.2663
x1	1.4546	0.2118	6.868	2e-06 ***
x2	9.3655	4.0640	2.305	0.0333 *

Residual standard error: 11.01 on 18 degrees of freedom
 Multiple R-squared: 0.9167, Adjusted R-squared: 0.9075
 F-statistic: 99.1 on 2 and 18 DF, p-value: 1.921e-10

Response: y

	df	sum Sq	Mean Sq	F value	Pr(>F)
x1	1	23371.8	23371.8	192.8962	4.64e-11 ***
x2	1	643.5	643.5	5.3108	0.03332 *
Residuals	18	2180.9	121.2		

```
dwa.mod<-lm(y~x1+x2, dwaine)
summary(dwa.mod)
anova(dwa.mod)
```

The Dwaine Studios example: The Dwaine operates studios that specialize in portraits of children. The company is considering whether sales (Y) in a community can be predicted from the number of persons aged 16 or younger in the community (X1) and the per capita disposable personal income in the community (X2). Data is Dwaine.csv, n=21, p=3

3). Find the estimated variance-covariance matrix for the parameter $\Sigma^2\{b\} = MSE(X'X)^{-1}$

$$s^2\{b_1\} = 0.04485$$

$$s^2\{b_2\} = 16.5158$$

	(Intercept)	x1	x2
(Intercept)	3602.03467	8.74593958	-241.4229923
x1	8.74594	0.04485151	-0.6724426
x2	-241.42299	-0.67244260	16.5157558

$$\text{Cov}(b_1, b_2) = -0.67$$

`vcov(dwa.mod)`

4). Test whether sales are related to the target population and per capita disposable income.

$$H_0: \beta_1 = 0 \text{ and } \beta_2 = 0,$$

$$H_a: \text{not both } \beta_1 \text{ and } \beta_2 \text{ equal zero}$$

$$F^* = MSR/MSE = 99.1$$

$$\text{For } \alpha = 0.05, \text{ we require } F(0.95; 2, 18) = 3.55.$$

The sales are related to (at least one or both) target population and per capita disposable income

Coefficients:

	Estimate	std. Error	t value	Pr(> t)
(Intercept)	-68.8571	60.0170	-1.147	0.2663
x1	1.4546	0.2118	6.868	2e-06 ***
x2	9.3655	4.0640	2.305	0.0333 *

Residual standard error: 11.01 on 18 degrees of freedom

Multiple R-squared: 0.9167, Adjusted R-squared: 0.9075

F-statistic: 99.1 on 2 and 18 DF, p-value: 1.921e-10

Response: y

	df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	23371.8	23371.8	192.8962	4.64e-11 ***
x2	1	643.5	643.5	5.3108	0.03332 *
Residuals	18	2180.9	121.2		

The Dwaine Studios example: The Dwaine operates studios that specialize in portraits of children. The company is considering whether sales (Y) in a community can be predicted from the number of persons aged 16 or younger in the community (X1) and the per capita disposable personal income in the community (X2). Data is Dwaine.csv, n=21, p=3

5) Find the 95% *individual* CI for each of the parameter.

$$\text{95\% CI for } \beta_1: b_1 \pm t s\{b1\} = 1.455 \pm 2.101(0.2118) \\ = (1.01, 1.9)$$

$$\text{95\% CI for } \beta_2: b_2 \pm t s\{b2\} = 9.366 \pm 2.101(4.064) \\ = (0.83, 17.9)$$

	(Intercept)	x1	x2
(Intercept)	3602.03467	8.74593958	-241.4229923
x1	8.74594	0.04485151	-0.6724426
x2	-241.42299	-0.67244260	16.5157558
	2.5 %	97.5 %	
(Intercept)	-194.9480130	57.233867	
x1	1.0096226	1.899497	
x2	0.8274411	17.903560	
<code>confint(dwa.mod, level=0.95)</code>			

6) Find the 95% simultaneous Bonferroni CI for the parameter (β_1 and β_2), g=2

$$B = t \left(1 - \frac{\alpha}{2g}, dfE \right) = t(1 - 0.9875, 18) = 2.445$$

$$\text{95\% simultaneous Bonferroni CI for } \beta_1 : \\ b_1 \pm t(1 - \alpha/2g, df)s\{b1\} \\ = 1.45 \pm 2.445(0.2118) = (0.932, 1.968)$$

$$\text{95\% simultaneous Bonferroni CI for } \beta_2 : \\ b_2 \pm t(1 - \alpha/2g, df)s\{b2\} \\ = 9.37 \pm 2.445(4.064) = (-0.566, 19.306)$$

Prediction in Multiple Linear Regression (MLR)

Prediction of the response variable

1. The $1 - \alpha$ prediction limits for **mean response** $E\{Y_h\}$ corresponding to X_h are

$$\hat{Y}_h \pm t(1 - \alpha/2; n - p)s\{\hat{Y}_h\} \quad s^2\{\hat{Y}_h\} = X'_h \Sigma\{b\} X_h \quad \text{Where } \Sigma\{b\} = MSE(X'X)^{-1}$$

2. The $1 - \alpha$ prediction limits for **single response** $Y_{h(\text{new})}$ corresponding to X_h are

$$\hat{Y}_h \pm t(1 - \alpha/2; n - p)s\{\text{pred}\} \quad s^2\{\text{pred}\} = MSE + s^2\{\hat{Y}_h\} = MSE(1 + \mathbf{X}'_h (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_h)$$

3. The $1 - \alpha$ prediction limits for **means of m new responses at X_h** are

$$\hat{Y}_h \pm t(1 - \alpha/2; n - p)s\{\text{predmean}\} \quad s^2\{\text{predmean}\} = \frac{MSE}{m} + s^2\{\hat{Y}_h\} = MSE \left(\frac{1}{m} + \mathbf{X}'_h (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_h \right)$$

4. The **simultaneous** $1 - \alpha$ prediction limits for **g mean observations at X_h** (Bonferroni procedure) are

$$\hat{Y}_h \pm Bs\{\text{pred}\} \quad B = t(1 - \alpha/2g; n - p)$$

The Dwaine Studios example: The Dwaine operates studios that specialize in portraits of children. The company is considering whether sales (Y) in a community can be predicted from the number of persons aged 16 or younger in the community (X_1) and the per capita disposable personal income in the community (X_2). Data is Dwaine.csv $n=21$

1). Estimate the 95% CI for the mean response when $X_1=65.4$ and $X_2=17.6$

The $1 - \alpha$ confidence limits for $E\{\hat{Y}_h\}$ are:

$$\hat{Y}_h \pm t(1 - \alpha/2; n - p)s\{\hat{Y}_h\}$$

$$dfE = 18 \quad t(0.975, dfE) = 2.101$$

$$s^2\{\hat{Y}_h\} = X_h' \Sigma \{b\} X_h = 7.656$$

$$\begin{aligned} \text{The CI, } \hat{Y}_h &\pm ts\{\hat{Y}_h\} = \\ &= (185.3, 196.9) \end{aligned}$$

We are 95% confident that the average sale will be between 185 and 197 when the population is 65.4 unit and personal income is 17.6 unit.

```
ci.reg(dwa.mod, new, type='m', alpha=0.05)
```

The Dwaine Studios example: The Dwaine operates studios that specialize in portraits of children. The company is considering whether sales (Y) in a community can be predicted from the number of persons aged 16 or younger in the community (X_1) and the per capita disposable personal income in the community (X_2). Data is Dwaine.csv $n=21$

2). Estimate the 95% CI for single response $\text{at } X_h = (65.4, 17.6)$ are

$$\begin{aligned}s^2\{pred\} &= \text{MSE} + X'_h \Sigma\{b\} X_h = \\ &= 128.82\end{aligned}$$

$$\begin{aligned}\text{The CI, } \hat{Y}_h \pm ts\{\hat{Y}_h\} &= \\ &= (167.3, 214.9)\end{aligned}$$

```
ci.reg(dwa.mod, new, type='n', alpha=0.05)
```

The Dwaine Studios example: The Dwaine operates studios that specialize in portraits of children. The company is considering whether sales (Y) in a community can be predicted from the number of persons aged 16 or younger in the community (X_1) and the per capita disposable personal income in the community (X_2). Data is Dwaine.csv $n=21$

3). Estimate the 95% CI for the mean of m (e.g, 2) new observations at the same $X_h = (65.4, 17.6)$ are

$$s^2\{\text{predmean}\} = \frac{MSE}{m} + s^2\{\hat{Y}_h\} = MSE \left(\frac{1}{m} + \mathbf{X}'_h (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}_h \right)$$

$$= 68.24$$

$$\text{The CI, } \hat{Y}_h \pm ts\{\text{predmean}\} =$$

$$= (173.75, 208.46)$$

```
ci.reg(dwa.mod, new, type='nm', m=2, alpha=0.05)
```

The Dwaine Studios example: The Dwaine operates studios that specialize in portraits of children. The company is considering whether sales (Y) in a community can be predicted from the number of persons aged 16 or younger in the community (X_1) and the per capita disposable personal income in the community (X_2). Data is Dwaine.csv $n=21$

A supplement question to 4). Estimate the 95% **simultaneous** CI for predicting mean responses when $(X_1, X_2)=(65.4, 17.6)$ and $(66, 20)$

$$s^2\{\hat{Y}_h\} = X_h' \Sigma \{b\} X_h \quad \text{Where } B = t\left(1 - \frac{0.05}{2g}; 18\right) = 2.445$$

$$X_h' \Sigma \{b\} X_h = \begin{pmatrix} 1 & 65.4 & 17.6 \\ 1 & 66.0 & 20.0 \end{pmatrix} \begin{pmatrix} 3602.03467 & 8.74593958 & -241.4229923 \\ 8.74594 & 0.04485151 & -0.6724426 \\ -241.42299 & -0.67244260 & 16.5157558 \end{pmatrix} \begin{pmatrix} 1.0 & 1 \\ 65.4 & 66 \\ 17.6 & 20 \end{pmatrix} = \begin{pmatrix} 7.65517 & 20.22547 \\ 20.22547 & 126.00603 \end{pmatrix}$$

$$s^2\{\hat{Y}_h\} = X_h' \Sigma \{b\} X_h = 7.655 \quad \text{when } (X_1, X_2)=(65.4, 17.6)$$

$$s^2\{\hat{Y}_h\} = X_h' \Sigma \{b\} X_h = 126.006 \quad \text{when } (X_1, X_2)=(66, 20)$$

$$\text{The simultaneous CI, } \hat{Y}_h \pm ts\{\hat{Y}_h\} = 190.0 \pm 2.445 * \sqrt{7.655} = (184.13 \ 197.66)$$

$$\hat{Y}_h \pm ts\{\hat{Y}_h\} = 214.45 \pm 2.445 * \sqrt{126.006} = (187, \ 241.9)$$

```
ci.reg(dwa.mod, new, type='gn', alpha=0.05)
```

The Dwaine Studios example: The Dwaine operates studios that specialize in portraits of children. The company is considering whether sales (Y) in a community can be predicted from the number of persons aged 16 or younger in the community (X_1) and the per capita disposable personal income in the community (X_2). Data is Dwaine.csv $n=21$

4). Estimate the 95% **simultaneous** CI for predicting single responses when $(X_1, X_2)=(65.4, 17.6)$ and $(66, 20)$

$$\hat{Y}_h \pm Bs\{\text{pred}\} \quad \text{Where } B = t\left(1 - \frac{0.05}{2g}; 18\right) = 2.445$$

$$s^2\{\text{pred}\} = \text{MSE} + X_h' \Sigma \{b\} X_h = 121.1626 + 7.656 = 128.82 \quad \text{when } (X_1, X_2)=(65.4, 17.6)$$

$$s^2\{\text{pred}\} = \text{MSE} + X_h' \Sigma \{b\} X_h = 247.16 \quad \text{when } (X_1, X_2)=(66, 20)$$

$$X_h' \Sigma \{b\} X_h = \begin{pmatrix} & & \end{pmatrix} \begin{pmatrix} 3602.03467 & 8.74593958 & -241.4229923 \\ 8.74594 & 0.04485151 & -0.6724426 \\ -241.42299 & -0.67244260 & 16.5157558 \end{pmatrix} \begin{pmatrix} 1.0 & 1 \\ 65.4 & 66 \\ 17.6 & 20 \end{pmatrix} = \begin{pmatrix} 7.65517 & 20.22547 \\ 20.22547 & 126.00603 \end{pmatrix}$$

$$\text{The simultaneous CI, } \hat{Y}_h \pm ts\{\hat{Y}_h\} = = (163.4, 218.9)$$

$$\hat{Y}_h \pm ts\{\hat{Y}_h\} = = (176, 252.9)$$

```
ci.reg(dwa.mod, new, type='gn', alpha=0.05)
```

Simultaneous confidence intervals for g mean response, at different X_h levels

1. Use the Working-Hotelling method

$$\hat{Y}_h \pm Ws\{\hat{Y}_h\} \quad \text{Where } W^2 = pF(1 - \alpha; p, n - p)$$

2. Use the Bonferroni method

$$\hat{Y}_h \pm Bs\{\hat{Y}_h\}$$

where:

$$B = t(1 - \alpha/2g; n - p)$$

Extra Sums of Squares and Marginal Effect

Extra Sum of Squares

- Measures the marginal reduction in the error sum of square when one or several predictor variables are added to the regression model, given that other predictor variables are already in the model.
- Can view as measuring the **marginal effect** in the regression sum of squares when one or several predictor variables are added to the regression model.
- The body fat example: a study of the relation of amount of **body fat (Y)** to several possible predictor variables, based on a sample of n=20 healthy females 25-34 years old. The possible predictors are
 - X1:** The triceps skinfold thickness;
 - X2:** The thigh circumference;
 - X3:** mid-arm circumference.
- We now construct 4 models, Y is regressed
 - (model1) on X1 alone; $lm(y \sim x1)$
 - (model2) on X2 alone; $lm(y \sim x2)$
 - (model3) on X1 and X2 only; and $lm(y \sim x1 + x2)$
 - (model4) on X1, X2 and X3. $lm(y \sim x1 + x2 + x3)$
- There could be 2^3 ways to construct MLR (including the null set model).

The scatter plot

X1: The triceps skinfold thickness;
X2: The thigh circumference;
X3: mid-arm circumference;
Y: body fat

```
plot(bodyfat)
```



Extra sum of squares

4

Model 1, $Y \sim X_1$

	Df	SS	MS
x1	1	352	352
Residuals	18	143	7.9
Total	19	495	

R^2

$$R^2 = \frac{352}{495} = 71\%$$

Reductions in error variance
Extra sum of square of Error

Increase in regression variance
Extra sum of square of Regression

SSE(X1)=143

SSR(X1)=352

Model 3, $Y \sim X_1 + X_2$

	Df	SS	MS
x1	1	352	352
x2	1	33	33
Residuals	17	110	6.5
Total	19	495	

$$R^2 = \frac{385}{495} = 78\%$$

SSE(X1) = 143
— SSE(X1, X2) = 110

SSE (X2 | X1) = 33

SSR(X1, X2) = 352 + 33 = 385
— SSM(X1) = 352

SSR (X2 | X1) = 33

Model 4, $Y \sim X_1 + X_2 + X_3$

	Df	SS	MS
x1	1	352	352
x2	1	33	33
x3	1	12	12
Residuals	16	98	6.1
Total	19	495	

$$R^2 = \frac{397}{495} = 80\%$$

SSE(X1, X2) = 110
— SSE(X1, X2, X3) = 98

SSE (X3 | X1, X2) = 12

SSR(X1, X2, X3) = 385 + 12 = 397
— SSR(X1, X2) = 385

SSR (X3 | X1, X2) = 12

Note that the extra sum of squares, $SSR(A|B)$ notation is equivalent to $SSE(A|B)$
But the $SSR(A)$ notation is not equivalent to $SSE(A)$.

ANOVA table containing decomposition of SSR (Type I ANOVA, entering order X1, X2, X3)

5

	Df	SS	MS
x1	1	352	352
x2	1	33	33
x3	1	12	12
Residuals	16	98	6.1
Total	19	495	

Source of variation	Df	SS	MS
Regression	3	SSR(X1, X2, X3)=397	MSR(X1, X2, X3)=397/3=132.33
X1	1	SSR(X1) = 352	MSR(X1)=352
X2 X1	1	SSR(X2 X1)=SSE(X2 X1)=33	MSR(X2 X1)=MSE(X2 X1)=33
X3 X1, X2	1	SSR(X3 X1, X2)=SSE(X3 X1, X2)=12	MSR(X3 X1, X2)=MSE(X3 X1, X2)=12
Residuals	16	SSE(X1, X2, X3)=98	MSE(X1, X2, X3)=98/16=6.13
Total	19	SSTO=495	

Note that the extra sum of squares, $SSR(A|B)$ notation is equivalent to $SSE(A|B)$
 But the $SSR(A)$ notation is not equivalent to $SSE(A)$.

ESS Terms Decomposed in Different Ways

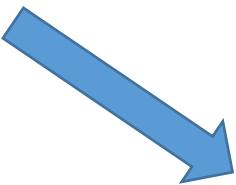
6

	Df	SS	MS
x1	1	352	352
x2	1	33	33
x3	1	12	12
Residuals	16	98	6.1
Total	19	495	



	Df	SS	MS
X1	1	352	352
X2, X3 X1	2	45	22.5
Residuals	16	98	6.1
Total	19	495	

$$\begin{aligned} \text{SSR}(X2, X3 | X1) &= \text{SSR}(X2 | X1) + \text{SSR}(X3 | X1, X2) = 33 + 12 = 45 \\ \text{OR} &= \text{SSR}(X1, X2, X3) - \text{SSR}(X1) = 397 - 352 = 45 \\ \text{OR} &= \text{SSE}(X1) - \text{SSE}(X1, X2, X3) = 143 - 98 = 45 \\ \text{MSR}(X2, X3 | X1) &= \text{SSR}(X2, X3 | X1) / 2 = 22.5 \end{aligned}$$



	Df	SS	MS
X1, X2	2	385	192.5
X3 X1, X2	1	12	12
Residuals	16	98	6.1
Total	19	495	

$$\begin{aligned} \text{SSR}(X1, X2) &= \text{SSR}(X1) + \text{SSR}(X2 | X1) = 352 + 33 = 385 \\ \text{OR} &= \text{SSR}(X1, X2, X3) - \text{SSR}(X3 | X1, X2) = 397 - 12 = 385 \\ \text{OR} &= \text{SST} - \text{SSE}(X1, X2) = \text{SST} - (\text{SSE}(X1, X2, X3) + \text{SSE}(X3 | X1, X2)) = 497 - (98 + 12) = 385 \end{aligned}$$

$$\text{MSR}(X1, X2) = \text{SSR}(X1, X2) / 2 = 192.5$$

Comments:

- Note that the extra sum of squares can be denoted as either $SSR(A|B)$ or $SSE(A|B)$ and should not be confused with the usual sum of squares, including $SSR(A)$, $SSE(A)$, $SSR(B)$, $SSE(B)$, $SSR(A, B)$ and $SSE(A, B)$
- The extra sum of squares can be decomposed in multiple ways in multiple steps.
 - ❖ $SSR(B, C|A)=SSR(B|A)+SSR(C|A, B)$
- The order in which the variables are presented matters in the extra sum of square terms.
 - ❖ $SSR(A|B)$ is not usually the same as $SSR(B|A)$.
 - ❖ However, $SSR(A,B) = SSR(B,A)$, and $SSE(A, B)=SSE(B, A)$
- The total sum of squares, $SST = \sum(Y_i - \bar{Y})^2$ always remains the same.
 - ❖ $SST=SSR(A)+SSE(A)=SSR(B)+SSE(B)=SSR(A,B)+SSE(A, B)=SSR(B, A)+SSE(B, A)$
 - ❖ $SST=SSR(A)+SSR(B|A)+SSE(A, B)$, or $SST=SSR(B)+SSR(A|B)+SSE(B, A)$.

A GLT Test for all $\beta_k = 0$ (e.g., $\beta_1 = 1, \beta_2 = 0, \beta_3 = 0$)

8

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$ (Reduced model) $H_a: \text{not all } \beta_k \text{ equal 0}$ (Full model)

$$Y_i = \beta_0 + \epsilon_i \text{ (The null model)}$$

$$dfE(\text{Reduced}) = n - p = n - 1$$

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \epsilon_i$$

$$dfE(\text{Full}) = n - p = n - 4$$

$$F_s = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)/df_F}{MSR}} = \frac{MSR}{MSE} \sim F(df_R - df_F, df_F)$$

$$F_s = \frac{\frac{SSR(X1, X2, X3)}{p - 1}}{\frac{SSE(X1, X2, X3)}{n - 4}}$$

$$\begin{aligned} F_s &= \frac{352.27 + 33.17 + 11.55}{\frac{3}{\frac{98.41}{16}}} \\ &= 21.52 \end{aligned}$$

- The critical value $F(0.95, 3, 16) = 3.24$, we reject the H_0 , not all betas are zero.
- This is the result shown in the R output.
- Since it is for all predictors, it is also known as the **global test**.

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	352.27	352.27	57.2768	1.131e-06 ***
x2	1	33.17	33.17	5.3931	0.03373 *
x3	1	11.55	11.55	1.8773	0.18956
Residuals	16	98.40	6.15		

Residual standard error: 2.48 on 16 degrees of freedom
 Multiple R-squared: 0.8014, Adjusted R-squared: 0.7641
 F-statistic: 21.52 on 3 and 16 DF, p-value: 7.343e-06

A GLT test for whether a single $\beta_k = 0$ (e.g., $\beta_3 = 0$), given all other predictors have been considered

9

$H_0: \beta_3 = 0$ (Reduced model)

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \varepsilon_i$$

$$dfE(\text{Reduced}) = n - p = n - 3$$

$H_a: \beta_3 \neq 0$ (Full model)

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \varepsilon_i$$

$$dfE(\text{Full}) = n - p = n - 4$$

$$F_s = \frac{\frac{SSE(R) - SSE(F)}{dfE_R - dfE_F}}{\frac{SSE(F)/df_F}{MSR}} = \frac{MSR}{MSE} \sim F(df_R - df_F, df_F)$$

$$F_s = \frac{\frac{SSR(X3|X1, X2)}{n - 3 - n + 4}}{\frac{SSE(X1, X2, X3)}{n - 4}}$$

$$F_s = \frac{\frac{11.55}{1}}{\frac{98.41}{16}}$$

$$= 1.88$$

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	352.27	352.27	57.2768	1.131e-06 ***
x2	1	33.17	33.17	5.3931	0.03373 *
x3	1	11.55	11.55	1.8773	0.18956
Residuals	16	98.40	6.15		

- Based on the critical value of $F(0.95, 1, 16) = 4.49$, and a p-value of 0.19. We can conclude that X3 can be removed from the MLR that already includes X1 and X2.
- When determining the significance of a predictor in an MLR, it is assumed that all other predictors have already been considered, this this predictor being the last to be evaluated.
- This process is also a test of **the predictor's marginal effect**.
- By default, in an MLR, the significance of a predictor is evaluated based on its marginal effect.**

A GLT test for whether a single $\beta_k = 0$ (e.g., $\beta_2 = 0$)

10

$H_0: \beta_2 = 0$ (Reduced model)

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_3 X_{i,2} + \varepsilon_i$$

$$dfE(\text{Reduced}) = n - p = n - 3$$

$$F_s = \frac{\frac{SSE(R) - SSE(F)}{dfE_R - dfE_F}}{\frac{SSE(F)/df_F}{MSR}} = \frac{MSR}{MSE} \sim F(df_R - df_F, df_F)$$

$$F_s = \frac{\frac{SSR(X2|X1, X3)}{n - 3 - n + 4}}{\frac{SSE(X1, X2, X3)}{n - 4}}$$

$$F_s = \frac{\frac{7.53}{1}}{\frac{98.41}{16}}$$

$$= 1.22$$

$H_a: \beta_2 \neq 0$ (Full model)

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_3 X_{i,2} + \beta_2 X_{i,3} + \varepsilon_i$$

$$dfE(\text{Full}) = n - p = n - 4$$

Analysis of variance Table

Response: y		Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1		1	352.27	352.27	57.2768	1.131e-06 ***
x2		1	33.17	33.17	5.3931	0.03373 *
x3		1	11.55	11.55	1.8773	0.18956
Residuals		16	98.40	6.15		

Response: y		Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1		1	352.27	352.27	57.2768	1.131e-06 ***
x3		1	37.19	37.19	6.0461	0.02571 *
x2		1	7.53	7.53	1.2242	0.28489
Residuals		16	98.40	6.15		

- Need to refit the model as $Y \sim X_1 + X_3 + X_2$, or $Y \sim X_3 + X_1 + X_2$.
- It is important to note that $SSR(X_2|X_1, X_3) \neq SSR(X_2|X_1)$, indicating that the effect of X_2 is not independent of the other variable in the model.
- With a p-value of 0.28, we can conclude that X_2 can be removed from the MLR that already includes X_1 and X_3 .

A GLT test for whether a subset of $\beta_k = 0$ (e.g., $\beta_2 = \beta_3 = 0$)

11

$H_0: \beta_2 = \beta_3 = 0$ (Reduced model)

$H_a: \text{not both } \beta_2 \text{ and } \beta_3 \text{ equal 0}$ (Full model)

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \epsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \epsilon_i$$

$$dfE_R = n - 2$$

$$dfE_F = n - 4$$

$$F_s = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)/df_F}{MSE}} = \frac{MSR}{MSE} \sim F(df_R - df_F, df_F)$$

$$F_s = \frac{\frac{SSR(X2, X3|X1)}{2}}{\frac{SSE(X1, X2, X3)}{n - 4}}$$

$$F_s = \frac{\frac{33.17 + 11.55}{2}}{\frac{98.41}{16}}$$

$$= 3.635$$

Analysis of variance Table

		Response: y				
	DF	Sum Sq	Mean Sq	F value	Pr(>F)	
x1	1	352.27	352.27	57.2768	1.131e-06	***
x2	1	33.17	33.17	5.3931	0.03373	*
x3	1	11.55	11.55	1.8773	0.18956	
Residuals	16	98.40	6.15			

- When evaluating the significance of a subset of predictors in an MLR, it is assumed that all other predictors have already been considered, with this subset being the last one to be evaluated.
- The critical value of $F(0.95, 2, 16) = 3.63$, indicating that further analysis maybe required before deciding whether X_2 and X_3 should be dropped from the regression model that already includes X_1 .
- In R this can also be done with function ***anova(reduced model, full model)***

Model 1: $y \sim x_1$		Model 2: $y \sim x_1 + x_2 + x_3$		Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	18	143.120	2						
2	16	98.405	2	44.715	3.6352	0.04995	*		

Understanding the F tests and P-values in the ANOVA Table for a predictor in an SLR

12

Analysis of variance Table

Response: y					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	352.27	352.27	44.305	3.024e-06 ***
Residuals	18	143.12	7.95		

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

$$Y_i = \beta_0 + \varepsilon_i \text{ The null model}$$

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \varepsilon_i$$

$$F_s = \frac{MSR}{MSE(\text{Full model})} = \frac{352.27}{7.95} = 44.305 \sim F(1, 18)$$

- The F test is used to evaluate the overall significance of the regression model. It considers the joint effect of all predictors in the model. In SLR, F test for the predictor is testing the significance of the predictor.
- The F test is equivalent to the t test corresponding to the predictor

Analysis of Variance Table

Response: y					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	352.27	352.27	57.2768	1.131e-06 ***
x2	1	33.17	33.17	5.3931	0.03373 *
x3	1	11.55	11.55	1.8773	0.18956
Residuals	16	98.40	6.15		

1. The first predictor, e.g. X1, in the model $X_1 + X_2 + X_3$

$$F_s = \frac{SSR(X_1)/1}{MSE(\text{full model})} = \frac{352.27}{6.15} = 57.28 \sim F(1, 16)$$

- A significant F-value for the first predictor indicates that the inclusion X_1 contributes significantly to explaining the variance in Y , after accounting for the other predictors in the model, in this specific order, X_1 , X_2 and X_3 .
- Note that is not a test for the marginal effect of X_1 , which is testable in the model where X_1 is last predictor in the model, e.g. $Y \sim X_2 + X_3 + X_1$:

Analysis of Variance Table

Response: y					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x2	1	381.97	381.97	62.1052	6.735e-07 ***
x3	1	2.31	2.31	0.3762	0.5483
x1	1	12.70	12.70	2.0657	0.1699
Residuals	16	98.40	6.15		

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

$$Y_i = \beta_0 + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \varepsilon_i$$

$$F_s = \frac{SSR(X_1|X_2, X_3)/1}{MSE(\text{full model})} = \frac{12.7}{6.15} = 2.065 \sim F(1, 16)$$

2. The middle predictor, e.g. X_2 , in the model $X_1 + X_2 + X_3$

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	352.27	352.27	57.2768	1.131e-06 ***
x2	1	33.17	33.17	5.3931	0.03373 *
x3	1	11.55	11.55	1.8773	0.18956
Residuals	16	98.40	6.15		

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	352.27	352.27	57.2768	1.131e-06 ***
x3	1	37.19	37.19	6.0461	0.02571 *
x2	1	7.53	7.53	1.2242	0.28489
Residuals	16	98.40	6.15		

$$F_s = \frac{SSR(X_2|X_1)/1}{MSE(\text{full model})} = \frac{33.17}{6.15} = 5.3921 \sim F(1, 16)$$

- A significant F-value for the middle predictor indicates that the inclusion X_2 contributes significantly to explaining the variance in Y , after accounting for the other predictors in the model, in this specific order, X_1 , X_2 and X_3 .
- Note that is not a test for the marginal effect of X_2 , which is testable in the model where X_1 is last predictor in the model, e.g., $Y \sim X_1 + X_3 + X_2$:

$$H_0: \beta_2 = 0$$

$$H_a: \beta_2 \neq 0$$

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_3 X_{i,3} + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \varepsilon$$

$$F_s = \frac{SSR(X_2|X_1, X_3)/1}{MSE(\text{full model})} = \frac{7.53}{6.15} = 1.2242 \sim F(1, 16)$$

3. last predictor, e.g. X_3 , in the model $X_1 + X_2 + X_3$

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	352.27	352.27	57.2768	1.131e-06 ***
x2	1	33.17	33.17	5.3931	0.03373 *
x3	1	11.55	11.55	1.8773	0.18956
Residuals	16	98.40	6.15		

$$F_s = \frac{SSR(X_3|X_1, X_2)/1}{MSE(\text{full model})} = \frac{11.55}{6.15} = 1.8773 \sim F(1, 16)$$

- Note that is the test for the marginal effect of X_3 in a full model that Consists of X_1 , X_2 and X_3 .

$$H_0: \beta_3 = 0$$

$$H_a: \beta_3 \neq 0$$

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \varepsilon$$

$$F_s = \frac{SSR(X_3|X_1, X_2)/1}{MSE(\text{full model})} = \frac{11.55}{6.15} = 1.8773 \sim F(1, 16)$$

Understanding the T tests and P-values in an MLR Model Summary Table

16

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	117.085	99.782	1.173	0.258
x1	4.334	3.016	1.437	0.170
x2	-2.857	2.582	-1.106	0.285
x3	-2.186	1.595	-1.370	0.190

Residual standard error: 2.48 on 16 degrees of freedom

Multiple R-squared: 0.8014, Adjusted R-squared: 0.7641

F-statistic: 21.52 on 3 and 16 DF, p-value: 7.343e-06

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	117.085	99.782	1.173	0.258
x1	4.334	3.016	1.437	0.170
x3	-2.186	1.595	-1.370	0.190
x2	-2.857	2.582	-1.106	0.285

- Order doesn't matter for the T test of a predictor because it is for the marginal effect of a single predictor.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.4961	3.3192	-0.451	0.658
x1	0.8572	0.1288	6.656	3.02e-06 ***

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-19.1742	8.3606	-2.293	0.0348 *
x1	0.2224	0.3034	0.733	0.4737
x2	0.6594	0.2912	2.265	0.0369 *

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.7916	4.4883	1.513	0.1486
x1	1.0006	0.1282	7.803	5.12e-07 ***
x3	-0.4314	0.1766	-2.443	0.0258 *

Important Applications of ESS terms and the GLT test!

17

$$Ha: Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i \quad (\text{Full model})$$

Case 1. $H_o: \beta_1 = \beta_2 = \beta_{new}$ *Not zero*
 $Ha: \beta_1 \neq \beta_2$

Case 2. $H_o: \beta_1 = 3, \beta_2 = 5$
 $Ha: \text{not both equalities in } H_o \text{ hold}$

$$Y_i = \beta_0 + \beta_{new}(X_{i1} + X_{i2}) + \beta_3 X_{i3} + \epsilon_i \quad (\text{Reduced model})$$

$$Y_i = \beta_0 + 3X_{i1} + 5X_{i2} + \beta_3 X_{i3} + \epsilon_i \quad , \text{ or}$$
$$(Y_i - 3X_{i1} - 5X_{i2}) = \beta_0 + \beta_3 X_{i3} + \epsilon_i \quad (\text{Reduced model})$$

- The matrix form (Y and/or the design matrix) needs to be modified.
- In case 1, where the full and reduced models have the same response variable (Y), the **anova(reduced model, full model)** can be used in this scenario to evaluate the hypothesis.
- In case 2, where the response variable changes, direct comparison of the two models is not possible. However, the GLT test can still be performed By calculating the test statistics.

$$F_s = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)/df_F}{MSE}} = \frac{MSR}{MSE} \sim F(df_R - df_F, df_F)$$

A Simulated Case 1. $H_0: \beta_1 = \beta_2 = \beta_{new}$
 $H_a: \beta_1 \neq \beta_2$

```
```{r}
n = 30
set.seed(123)
x1 = runif(n)
x2 = runif(n, max=5)
x3 = runif(n, max=10)
```
```

```
```{r}
set.seed(123)
b0 = 1
b1 = 2
b2 = 2
b3 = 5
Y = b0 + b1*x1 + b2*x2 + b3*x3 + rnorm(n)
m1 = lm(Y~x1+x2+x3)
m1_reduced = lm(Y~I(x1+x2)+x3)
anova(m1_reduced, m1)
```
```

Analysis of Variance Table

| | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|--------|--------|----|-----------|--------|--------|
| 1 | 27 | 27.081 | | | | |
| 2 | 26 | 26.878 | 1 | 0.20344 | 0.1968 | 0.661 |

A Simulated Case 2. $H_0: \beta_1 = 3, \beta_2 = 5$ $H_a: \text{not both equalities in } H_0 \text{ hold}$

19

```
```{r}
set.seed(123)
b0 = 1
b1 = 3
b2 = 5
b3 = 8
Y = b0 + b1*x1 + b2*x2 + b3*x3 + rnorm(n)

m2 = lm(Y~x1+x2+x3)
Y_new = Y-3*x1-5*x2
m2_reduced = lm(Y_new~x3)
```
```

In this case you may not directly use anova function because in R it requires the response to be the same. Therefore, we need to compute the F statistics.

```
```{r}
MSR = (sum(m2_reduced$residuals^2)-sum(m2$residuals^2))/(m2_reduced$df.residual-m2$df.residual)
MSE = sum(m2$residuals^2)/m2$df.residual
Fs = MSR/MSE
Fs
```
[1] 0.3404
```

The .95 quantile for F distribution in this case:

```
```{r}
qf(0.95, m2_reduced$df.residual-m2$df.residual, m2$df.residual)
```
[1] 3.369016
```

The F statistics is smaller than the threshold, so we do not reject the null hypothesis.

Or we can use the p-value as well.

```
```{r}
pf(Fs, m2_reduced$df.residual-m2$df.residual, m2$df.residual)
```
[1] 0.2853909
```

The p-value is large.

Type I and Type II Sum of Squares and Partial R^2

Type I vs. Type II Sum of Square Terms

| Variable | Type I SS | Type II SS |
|----------|-----------------------|-----------------------|
| X_1 | $SSR(X_1)$ | $SSR(X_1 X_2, X_3)$ |
| X_2 | $SSR(X_2 X_1)$ | $SSR(X_2 X_1, X_3)$ |
| X_3 | $SSR(X_3 X_1, X_2)$ | $SSR(X_3 X_1, X_2)$ |

```
anova(model4) #type I
```

Analysis of Variance Table

Response: y

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------|----|--------|---------|---------|---------------|
| x1 | 1 | 352.27 | 352.27 | 57.2768 | 1.131e-06 *** |
| x2 | 1 | 33.17 | 33.17 | 5.3931 | 0.03373 * |
| x3 | 1 | 11.55 | 11.55 | 1.8773 | 0.18956 |
| Residuals | 16 | 98.40 | 6.15 | | |

```
library(car)
Anova(model4, type="II") #type II
```

Anova Table (Type II tests)

Response: y

| | Sum Sq | Df | F value | Pr(>F) |
|-----------|--------|----|---------|--------|
| x1 | 12.705 | 1 | 2.0657 | 0.1699 |
| x2 | 7.529 | 1 | 1.2242 | 0.2849 |
| x3 | 11.546 | 1 | 1.8773 | 0.1896 |
| Residuals | 98.405 | 16 | | |

The F tests in the Type II ANOVA table are equivalent to The T tests in the Model summary.

The effect of order of predictors entering the model

| Variable | Type I SS | Type II SS |
|----------|-----------------------|-----------------------|
| X_3 | $SSR(X_3)$ | $SSR(X_3 X_1, X_2)$ |
| X_2 | $SSR(X_2 X_3)$ | $SSR(X_2 X_1, X_3)$ |
| X_1 | $SSR(X_1 X_2, X_3)$ | $SSR(X_1 X_2, X_3)$ |

Analysis of Variance Table

```
Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x3         1 10.05   10.05  1.6343  0.2193
x2         1 374.23  374.23 60.8471 7.684e-07 ***
x1         1 12.70   12.70  2.0657  0.1699
Residuals 16 98.40    6.15

```

Anova Table (Type II tests)

```
Response: y
          Sum Sq Df F value    Pr(>F)
x3        11.546  1 1.8773 0.1896
x2         7.529  1 1.2242 0.2849
x1        12.705  1 2.0657 0.1699
Residuals 98.405 16

```

Sum up to $SSR = SSTO - SSE$

Analysis of Variance Table

```
Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 352.27  352.27 57.2768 1.131e-06 ***
x2         1 33.17   33.17  5.3931  0.03373 *
x3         1 11.55   11.55  1.8773  0.18956
Residuals 16 98.40    6.15

```

Anova Table (Type II tests)

```
Response: y
          Sum Sq Df F value    Pr(>F)
x1        12.705  1 2.0657 0.1699
x2         7.529  1 1.2242 0.2849
x3        11.546  1 1.8773 0.1896
Residuals 98.405 16

```

Type I SS would change by the order

Type II SS would not change.

After

Before

Comments

- Type I SS always sum to SSR for the model with all predictors.
- Type I SS can give different values depending on the order in which variables are specified (e.g., switching X_1 and X_2).
- Type I SS are generally less useful than Type II SS unless you are specifically interested in partitioning variation among an ordered set of predictors.
- Type II SS can be considered a special case of Type I SS.
- When there is no assumption violation and the Type I and II ANOVA tables are similar, the order doesn't matter in the marginal effect of the predictors given others. We can conclude that the predictors are independent.

Coefficients of partial determination

The relative marginal reduction in the variation in Y associated with some predictor when others are already in the model is

$$R^2_{Y2|1} = 0.232$$

$$R^2_{Y3|12} = 0.105$$

$$R^2_{Y1|2} = 0.031$$

When X2 is added to the model containing X1, the error sum of squares is reduced by 23.2%.

The error sum of squares containing both X1 and X2 is reduced by 10.5% when X3 is added.

Adding X1 to the model containing X2, the error sum of squares is reduced only by 3.1%.

Coefficients of partial determination (example 1)

The relative marginal reduction in the variation in Y associated with some predictor when others are already in the model is

$$R^2_{Y2|1} = 0.232$$

When X2 is added to the model containing X1, the **existing** error sum of squares is reduced by 23.2%.. .

Model 1, Y~X1

| | Df | SS | MS |
|-----------|----|-----|-----|
| x1 | 1 | 352 | 352 |
| Residuals | 18 | 143 | 7.9 |
| Total | 19 | 495 | |

$$R^2$$

$$R^2 = \frac{352}{495} = 71\%$$

$$SSE(X_1) = 143 \quad SSE(X_2|X_1) = 33$$

Model 3, Y~X1+X2

| | Df | SS | MS |
|-----------|----|-----|-----|
| x1 | 1 | 352 | 352 |
| x2 | 1 | 33 | 33 |
| Residuals | 17 | 110 | 6.5 |
| Total | 19 | 495 | |

$$R^2 = \frac{385}{495} = 78\%$$

$$R^2_{Y2|1} = \frac{SSE(X_2|X_1)}{SSE(X_1)} = \frac{33}{143} = 0.232$$

Model 4, Y~X1+X2+X3

| | Df | SS | MS |
|-----------|----|-----|-----|
| x1 | 1 | 352 | 352 |
| x2 | 1 | 33 | 33 |
| x3 | 1 | 12 | 12 |
| Residuals | 16 | 98 | 6.1 |
| Total | 19 | 495 | |

$$R^2 = \frac{397}{495} = 80\%$$

Coefficients of partial determination (example 2)

The relative marginal reduction in the variation in Y associated with some predictor when others are already in the model is

$$R^2_{Y3|12} = 0.105$$

When X3 is added to the model containing X1 X2, the **existing** error sum of squares is reduced by 10.5%.. .

Model 1, $Y \sim X_1$

| | Df | SS | MS |
|-----------|----|-----|-----|
| x1 | 1 | 352 | 352 |
| Residuals | 18 | 143 | 7.9 |
| Total | 19 | 495 | |

$$R^2$$

$$R^2 = \frac{352}{495} = 71\%$$

Model 3, $Y \sim X_1 + X_2$

| | Df | SS | MS |
|-----------|----|-----|-----|
| x1 | 1 | 352 | 352 |
| x2 | 1 | 33 | 33 |
| Residuals | 17 | 110 | 6.5 |
| Total | 19 | 495 | |

$$R^2 = \frac{385}{495} = 78\%$$

$$SSE(X_1, X_2) = 110 \quad SSE(X_3 | X_1 X_2) = 12$$

Model 4, $Y \sim X_1 + X_2 + X_3$

| | Df | SS | MS |
|-----------|----|-----|-----|
| x1 | 1 | 352 | 352 |
| x2 | 1 | 33 | 33 |
| x3 | 1 | 12 | 12 |
| Residuals | 16 | 98 | 6.1 |
| Total | 19 | 495 | |

$$R^2 = \frac{397}{495} = 80\%$$

$$R^2_{Y3|12} = \frac{SSE(X_3 | X_1 X_2)}{SSE(X_1, X_2)} = \frac{12}{110} = 0.105$$

Coefficients of partial determination

A coefficient of partial determination measures the marginal contribution of one X variable when all others Are already included in the model

For example,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

The relative marginal reduction in the variation in Y associated with X1 when X2 is already in the model is

$$R^2_{Y1|2} = \frac{SSE(X2) - SSE(X1, X2)}{SSE(X2)} = \frac{SSR(X1|X2)}{SSE(X2)}$$

$$R^2_{Y2|1} = \frac{SSE(X1) - SSE(X1, X2)}{SSE(X1)} = \frac{SSR(X2|X1)}{SSE(X1)}$$

Q: Which of following represents the relative marginal reduction in the variation in Y associated with X3 when X1 and X2 are already in the model

A) $R^2_{Y12|3} = \frac{SSR(X1 X2|X3)}{SSE(X3)}$

B) $R^2_{Y12|3} = \frac{SSR(X3|X1 X2)}{SSE(X1 X2)}$

C) $R^2_{Y3|12} = \frac{SSR(X1 X2|X3)}{SSE(X3)}$

D) $R^2_{Y3|12} = \frac{SSR(X3|X1 X2)}{SSE(X1 X2)}$

Type I and Type II Partial coefficient determination R^2

Partial determination can be calculated from the Type I and Type II SS:

- Type I Squared Partial Correlation uses Type I SS: $R^2 = \frac{SS1}{SS1 + SSE}$
- Type II Squared Partial Correlation uses Type II SS: $R^2 = \frac{SS2}{SS2 + SSE}$

Where SS1 and SS2 are the Type I and Type II sums of squares for a particular predictor variable, and SSE is the sum of squared error for the full model.

The partial correlation $r = \sqrt{R^2}$

Type I Coefficients of Partial Determination

- The order matters! Suppose the variables enter the model in the order of X3, X2, X1

$$R_{Y|3}^2 = \frac{SSR(X3)}{SST} = \frac{10.05}{495} = 0.02$$

$$\begin{aligned} R_{Y|2|3}^2 &= \frac{SSR(X2|X3)}{SSE(X3)} \\ &= \frac{SSR(X2|X3)}{SSR(X2|X3) + SSE(X3|X2)} = \frac{374.23}{374.23 + 12.7 + 98.4} = 0.77 \end{aligned}$$

$$\begin{aligned} R_{Y|1|3|2}^2 &= \frac{SSR(X1|X3|X2)}{SSE(X3|X2)} \\ &= \frac{SSR(X1|X3|X2)}{SSR(X1|X3|X2) + SSE(X3|X2|X1)} = \frac{12.7}{12.7 + 98.4} = 0.114 \end{aligned}$$

| Response: y | | | | | | |
|-------------|----|--------|---------|---------|-----------|-----|
| | Df | Sum Sq | Mean Sq | F value | Pr (>F) | |
| x3 | 1 | 10.05 | 10.05 | 1.6343 | 0.2193 | |
| x2 | 1 | 374.23 | 374.23 | 60.8471 | 7.684e-07 | *** |
| x1 | 1 | 12.70 | 12.70 | 2.0657 | 0.1699 | |
| Residuals | 16 | 98.40 | 6.15 | | | |

- Type II coefficients of partial determination can be denoted and computed the same way as type I.
 - $R_{Y|2|3}^2$ is also the type II coefficient of partial determination of X2 in the MLR with just X2 and X3 predictors.
 - $R_{Y|1|3,2}^2$ is also the type II coefficient of partial determination of X1 in the MLR with just X1, X2 and X3 predictors.

Type II Coefficients of Partial Determination

- Measures the marginal contribution of one X variable **when all other variables** are already included in the model.

$$R_{Y|1|2,3}^2 = \frac{SSR(X_1|X_2, X_3)}{SSR(X_1|X_2, X_3) + SSE}$$

$$R_{Y|2|1,3}^2 = \frac{SSR(X_2|X_1, X_3)}{SSR(X_2|X_1, X_3) + SSE}$$

$$R_{Y|3|1,2}^2 = \frac{SSR(X_3|X_1, X_2)}{SSR(X_3|X_1, X_2) + SSE}$$

- Type II coefficient** of partial determination of a predictor (X_i) is **its Type I coefficient** when it entering the model last.
 - ❖ The order of other predictors entering the model doesn't matter.

Compute the Type II Coefficients of Partial Determination of X1 and X3 from the type I ANOVA table

$$R^2_{Y|1|2,3} = \frac{SSR(X1|X2,X3)}{SSE(X2,X3)} = 0.114$$

| Response: y | | | | | | |
|-------------|----|--------|---------|---------|-----------|-----|
| | Df | Sum Sq | Mean Sq | F value | Pr(>F) | |
| x3 | 1 | 10.05 | 10.05 | 1.6343 | 0.2193 | |
| x2 | 1 | 374.23 | 374.23 | 60.8471 | 7.684e-07 | *** |
| x1 | 1 | 12.70 | 12.70 | 2.0657 | 0.1699 | |
| Residuals | 16 | 98.40 | 6.15 | | | |

$$R^2_{Y|3|1,2} = \frac{SSR(X3|X1 X2)}{SSE(X1,X2)} = 0.105$$

| Response: y | | | | | | |
|-------------|----|--------|---------|---------|-----------|-----|
| | Df | Sum Sq | Mean Sq | F value | Pr(>F) | |
| x1 | 1 | 352.27 | 352.27 | 57.2768 | 1.131e-06 | *** |
| x2 | 1 | 33.17 | 33.17 | 5.3931 | 0.03373 | * |
| x3 | 1 | 11.55 | 11.55 | 1.8773 | 0.18956 | |
| Residuals | 16 | 98.40 | 6.15 | | | |

Partial Coefficient of Determination R^2 , connection between type I and type II

| Variable | Type I
(order 1,2,3) | Type I
(order 3,2,1) | Type I
(order 1,3,2) | Type II |
|----------|-------------------------|-------------------------|-------------------------|---------|
| X1 | 0.711 | 0.114 | 0.711 | 0.114 |
| X2 | 0.231 | 0.771 | 0.071 | 0.071 |
| X3 | 0.105 | 0.02 | 0.26 | 0.105 |

Type II R^2 is the same as type I for a predictor when it is the last one entering the model.

Partial Correlation Coefficient r , (in population, ρ) and Coefficients of Determination (R^2)

- Correlation coefficient measures the linear association between two (continuous) variables.
- *Partial correlation* measures the strength and direction of a linear association between two continuous variables while controlling one or more other continuous variables.

$$r_{Y|3} = \pm \sqrt{R_{Y|3}^2} \quad r_{Y|2|3} = \pm \sqrt{R_{Y|2|3}^2} \quad r_{Y|1|2,3} = \pm \sqrt{R_{Y|1|2,3}^2}$$

- In MLR, $R^2 = SSR/SST$ is the proportion of variation explained by the linear model, while the *coefficient of partial determination* for X_k measures the marginal increase in SSR that results from including X_k in the model.

Effects of Multicollinearity

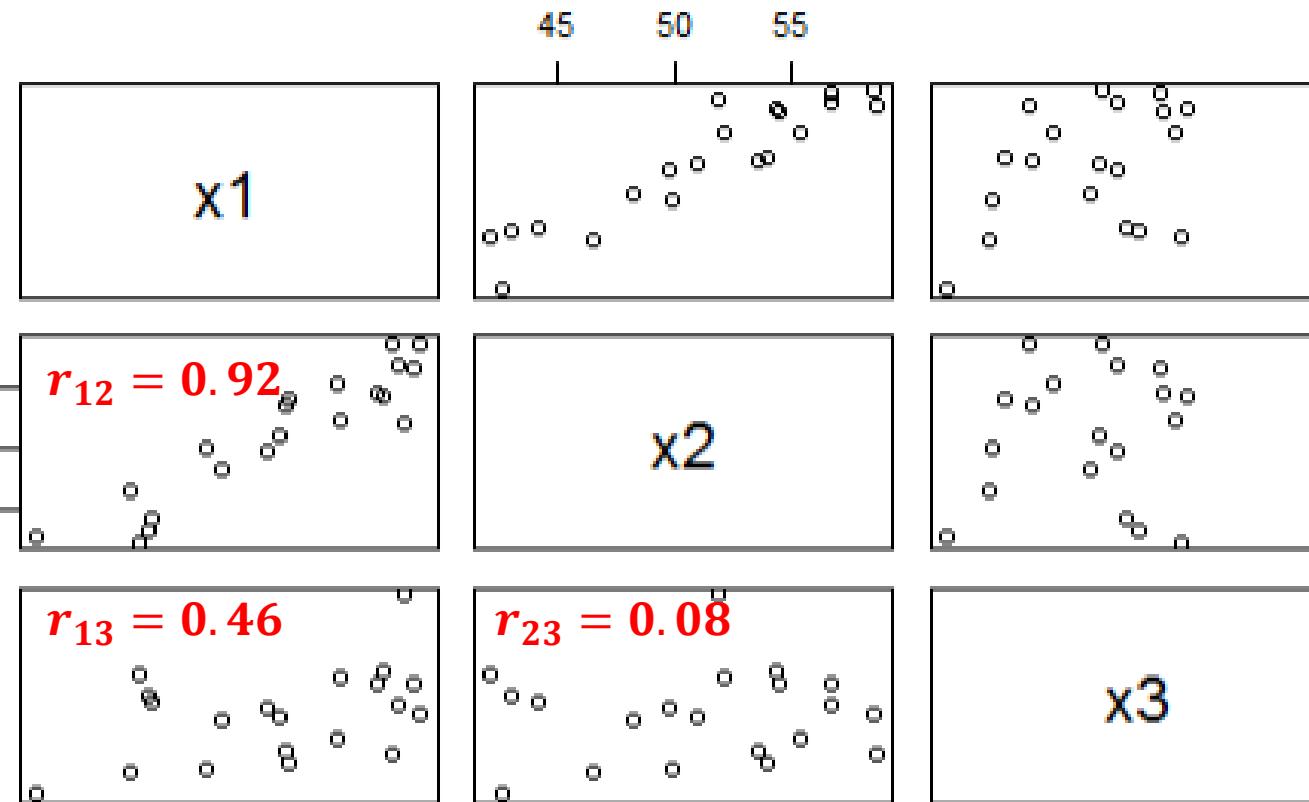
The body fat example

- The body fat example: a study of the relation of amount of **body fat (Y)** to several possible predictor variables, based on a sample of n=20 healthy females 25-34 years old. The possible predictors are

X1: The triceps skinfold thickness;

X2: The thigh circumference;

X3: midarm circumference.



X1 and X2 are highly correlated;
X3 is not so related to X1 and X2 individually

Effects of multicollinearity on regression coefficients, b_k

- The regression coefficient of one variable (eg. For X1) varies markedly depending on other variables in the model.

- If predictors are correlated, the regression coefficient cannot capture the true effect of the individual predictor variable, and instead, only represents a marginal or partial effect.

As a result, the coefficients in the MLR model do not accurately reflect the linear impact of the variable on Y.

- If intercorrelated predictor variables are left out of the model, they can still affect the coefficients of the remaining variables in the model.

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|--------------|
| (Intercept) | -1.4961 | 3.3192 | -0.451 | 0.658 |
| x1 | 0.8572 | 0.1288 | 6.656 | 3.02e-06 *** |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|--------------|
| (Intercept) | -23.6345 | 5.6574 | -4.178 | 0.000566 *** |
| x2 | 0.8565 | 0.1100 | 7.786 | 3.6e-07 *** |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|----------|
| (Intercept) | -19.1742 | 8.3606 | -2.293 | 0.0348 * |
| x1 | 0.2224 | 0.3034 | 0.733 | 0.4737 |
| x2 | 0.6594 | 0.2912 | 2.265 | 0.0369 * |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|----------|
| (Intercept) | 117.085 | 99.782 | 1.173 | 0.258 |
| x1 | 4.334 | 3.016 | 1.437 | 0.170 |
| x2 | -2.857 | 2.582 | -1.106 | 0.285 |
| x3 | -2.186 | 1.595 | -1.370 | 0.190 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|--------------|
| (Intercept) | 6.7916 | 4.4883 | 1.513 | 0.1486 |
| x1 | 1.0006 | 0.1282 | 7.803 | 5.12e-07 *** |
| x3 | -0.4314 | 0.1766 | -2.443 | 0.0258 * |

Effects of multicollinearity on the standard error of the coefficients, $s\{b_k\}$

When only X1 in the model $s\{b_1\} = 0.1288$

| Coefficients: | | | | | |
|---------------|----------|------------|---------|--------------|--|
| | Estimate | Std. Error | t value | Pr(> t) | |
| (Intercept) | -1.4961 | 3.3192 | -0.451 | 0.658 | |
| x1 | 0.8572 | 0.1288 | 6.656 | 3.02e-06 *** | |

When only X2 in the model $s\{b_2\} = 0.11$

| Coefficients: | | | | | |
|---------------|----------|------------|---------|--------------|--|
| | Estimate | Std. Error | t value | Pr(> t) | |
| (Intercept) | -23.6345 | 5.6574 | -4.178 | 0.000566 *** | |
| x2 | 0.8565 | 0.1100 | 7.786 | 3.6e-07 *** | |

When only X1, x2 in the model $s\{b_1\} = 0.3034$

$s\{b_2\} = 0.2912$

| Coefficients: | | | | | |
|---------------|----------|------------|---------|----------|--|
| | Estimate | Std. Error | t value | Pr(> t) | |
| (Intercept) | -19.1742 | 8.3606 | -2.293 | 0.0348 * | |
| x1 | 0.2224 | 0.3034 | 0.733 | 0.4737 | |
| x2 | 0.6594 | 0.2912 | 2.265 | 0.0369 * | |

When X1, x2, and x3 in the model $s\{b_1\} = 3.016$

$s\{b_2\} = 2.582$

| Coefficients: | | | | | |
|---------------|----------|------------|---------|----------|--|
| | Estimate | Std. Error | t value | Pr(> t) | |
| (Intercept) | 117.085 | 99.782 | 1.173 | 0.258 | |
| x1 | 4.334 | 3.016 | 1.437 | 0.170 | |
| x2 | -2.857 | 2.582 | -1.106 | 0.285 | |
| x3 | -2.186 | 1.595 | -1.370 | 0.190 | |

- High degree of multicollinearity among the predictor variables is responsible for the inflated variability of the estimated regression coefficients, resulting in an increased standard error of the estimates.

Effects of multicollinearity on sums of squares

- In the presence of correlated predictors, the impact of a single predictor variable on reducing the error sum of squares can differ depending on the other variables included in the model. Therefore, when evaluating the reduction in total variance attributed to a particular predictor, it must be considered in the context of the other correlated predictors.

$$\text{SSR}(X_1) = 352.27$$

$$\text{SSR}(X_1 | X_2) = 3.47$$

$$\text{SSR}(X_1) > \text{SSR}(X_1 | X_2)$$

$$R^2_1 = 0.72 > R^2_{1|2} = \frac{3.47}{3.47 + 109.95} = 0.03$$

- When $\text{SSR}(X_1) < \text{SSR}(X_1 | X_2)$, X_2 is called **suppressor** variable.
- In other words, a suppressor variable enhances the relationship between the Y variable and another predictor variable by removing the influence of extraneous or confounding variables, which allows for a more accurate prediction of the outcome variable. The evidence of suppressor variable can be verified via the corresponding ESS terms or partial correlation coefficients values.

Suppressor Variable Example:

- The dependent variable, Y (Chance to cancel a streaming service subscription)
- The Predictors variables X1(age), X2(subscription plan cost), X3(frequency of watching)
- The confounding variable Xc (the availability of alternative service in the location)
- The suppressor variable Xs (location)

```
anova(lm(y~x1, bodyfat))
```

Analysis of Variance Table

Response: y

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------|----|--------|---------|---------|---------------|
| x1 | 1 | 352.27 | 352.27 | 44.305 | 3.024e-06 *** |
| Residuals | 18 | 143.12 | 7.95 | | |

```
anova(lm(y~x2+x1, bodyfat))
```

Analysis of Variance Table

Response: y

| | Df | Sum Sq | Mean Sq | F value | Pr (>F) |
|-----------|----|--------|---------|---------|---------------|
| x2 | 1 | 381.97 | 381.97 | 59.057 | 6.281e-07 *** |
| x1 | 1 | 3.47 | 3.47 | 0.537 | 0.4737 |
| Residuals | 17 | 109.95 | 6.47 | | |

```
anova(lm(y~x2, bodyfat))
```

Analysis of Variance Table

Response: y

| | Df | Sum Sq | Mean Sq | F value | Pr (>F) |
|-----------|----|--------|---------|---------|-------------|
| x2 | 1 | 381.97 | 381.97 | 60.617 | 3.6e-07 *** |
| Residuals | 18 | 113.42 | 6.30 | | |

Effects of multicollinearity on mean response estimate

When only X1 in the model MSE= 7.95

When $X_1 = 25$, $\hat{Y}_h = -1.4961 + 0.8572(25) = 19.93$

$$s\{\hat{Y}_h\} = \sqrt{\mathbf{X}'_h \boldsymbol{\Sigma}\{\mathbf{b}\} \mathbf{X}_h} = \sqrt{s^2 \left(\frac{1}{n} + \frac{(25 - \bar{X})^2}{SSX} \right)} = 0.63$$

When only X1, x2 in the model MSE=6.47

When $X_1 = 25$, $X_2 = 50$, $\hat{Y}_h = -19.1742 + 0.2224(25) + 0.6594(50) = 19.39$

$$s\{\hat{Y}_h\} = \sqrt{\mathbf{X}'_h \boldsymbol{\Sigma}\{\mathbf{b}\} \mathbf{X}_h} = 0.624$$

When X1, x2, and x3 in the model MSE = 6.15

When $X_1 = 25$, $X_2 = 50$, $X_3 = 29$, $\hat{Y}_h = 19.19$ $s\{\hat{Y}_h\} = 0.619$

Analysis of variance Table

Response: y

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------|----|--------|---------|---------|---------------|
| x1 | 1 | 352.27 | 352.27 | 44.305 | 3.024e-06 *** |
| Residuals | 18 | 143.12 | 7.95 | | |

Analysis of Variance Table

Response: y

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------|----|--------|---------|---------|---------------|
| x2 | 1 | 381.97 | 381.97 | 59.057 | 6.281e-07 *** |
| x1 | 1 | 3.47 | 3.47 | 0.537 | 0.4737 |
| Residuals | 17 | 109.95 | 6.47 | | |

Analysis of Variance Table

Response: y

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------|----|--------|---------|---------|---------------|
| x1 | 1 | 352.27 | 352.27 | 57.2768 | 1.131e-06 *** |
| x2 | 1 | 33.17 | 33.17 | 5.3931 | 0.03373 * |
| x3 | 1 | 11.55 | 11.55 | 1.8773 | 0.18956 |
| Residuals | 16 | 98.40 | 6.15 | | |

When more variables are added to the model, the high degree of multicollinearity

- does not prevent the SSE from being steadily reduced.
- could prevent the MSE from being steadily reduced.
- will increase the standard error of the mean response estimate.

Need for more powerful diagnostics for multicollinearity

- Multicollinearity in predictor variables can significantly impact the interpretation and utilization of a regression model.
- The correlation coefficient and partial correlation coefficient are common diagnostic tools used to identify multicollinearity and can be useful in detecting the issue.
- However, it is possible for serious multicollinearity to exist without being detected by these methods. Later, we will discuss some remedial measures that can be taken to address this issue.

Polynomial Regression

Polynomial Regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \beta_4 X_{i2}^2 + \beta_5 X_{i1} X_{i2} + \epsilon_i$$

- With enough data, can approximate arbitrary nonlinear response functions
- Most useful when the mean response is non-linear but error variance is constant
- In many cases, transformation of Y may make more sense
 - Regression on transformed variables may use fewer degrees of freedom
 - Polynomial regression will **not** correct non-constant variance

- Polynomials and transformations can be used together.
- Polynomials of several predictors can be combined (*response surface methodology*)
- this may or may not involve interactions between variables

Cautions

- Polynomials generally create a multicollinearity problem, which can often be corrected by centering the data ($x_i = X_i - \bar{X}$) , or standardization ($x_i = (X_i - \bar{X})/s$) . In R, use the `scale()` function
- Extrapolation beyond the scope of the data is a **very bad idea**

Example: Battery life

A researcher studied the effect of the charge rate and temperature on the life of a new type of power cell in a preliminary small-scale experiment.

X1:Charge rate (3 levels)

X2:Temperature (3 levels)

Y: Number of cycles

- The levels of charge and temperature are planned.
- We want to know:
 1. whether a linear or quadratic function is appropriate
 2. if there is a significant interaction between charge rate and temperature

Starting with the second-order polynomial regression model with interaction, the researcher aimed to better understand the response function (e.g., the MLR function) within the range of the factor being studied. Despite uncertainty about its nature, this was seen as a necessary step in the research process.

$$Y_i = old\beta_0 + old\beta_1 X_{i1} + old\beta_2 X_{i2} + old\beta_3 X_{i1}^2 + old\beta_4 X_{i2}^2 + old\beta_5 X_{i1}X_{i2} + \epsilon_i$$

Correlation between X_1 and X_1^2 is 0.991, between X_2 and X_2^2 is 0.986

To reduce multicollinearity, the researcher decided to center the variables. For the sake of demonstration, they also scaled the variables using convenient units.

$$x_{i1} = \frac{X_{i1}-\bar{X}}{0.4} = \frac{X_{i1}-1}{0.4}, \quad x_{i2} = \frac{X_{i2}-\bar{X}}{10} = \frac{X_{i2}-20}{10}, \quad \text{hence}$$

- The actual $sd[x1]=0.31$, and $sd[x2]=7.75$
- If use $scale(x)$, there will be slightly different
- Choose 0.4 and 10 for convenient.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \beta_4 x_{i2}^2 + \beta_5 x_{i1}x_{i2} + \epsilon_i$$

Correlation between x_1 and x_1^2 is now <0.001, between x_2 and x_2^2 is now <0.001

Notes:

1. $old\beta \neq \beta$
2. For simplicity in coding, sometimes we exchange notations:

$$x_1^2 = x_{11}, \quad x_2^2 = x_{22}, \quad x_1x_2 = x_{12}, \quad \beta_3 = \beta_{11}, \beta_4 = \beta_{22}, \beta_5 = \beta_{12},$$

Regression Result Based on the Scaled X variables

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|-------------|----------|------------|---------|----------|-----|
| (Intercept) | 162.84 | 16.61 | 9.805 | 0.000188 | *** |
| x1 | -55.83 | 13.22 | -4.224 | 0.008292 | ** |
| x2 | 75.50 | 13.22 | 5.712 | 0.002297 | ** |
| x11 | 27.39 | 20.34 | 1.347 | 0.235856 | |
| x22 | -10.61 | 20.34 | -0.521 | 0.624352 | |
| x12 | 11.50 | 16.19 | 0.710 | 0.509184 | |

```
summary(lm(cycle~rate+temp+rate2+temp2+rate*temp, data=cell))
```

Residual standard error: 32.37 on 5 degrees of freedom

Multiple R-squared: 0.9135, Adjusted R-squared: 0.8271

F-statistic: 10.57 on 5 and 5 DF, p-value: 0.01086

Type I SS

Analysis of Variance Table

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) | |
|-----------|----|--------|---------|---------|----------|----|
| x1 | 1 | 18704 | 18704 | 17.8460 | 0.008292 | ** |
| x2 | 1 | 34202 | 34202 | 32.6323 | 0.002297 | ** |
| x11 | 1 | 1646 | 1646 | 1.5704 | 0.265552 | |
| x22 | 1 | 285 | 285 | 0.2719 | 0.624352 | |
| x12 | 1 | 529 | 529 | 0.5047 | 0.509184 | |
| Residuals | 5 | 5240 | 1048 | | | |

Type II SS

Anova Table (Type II tests)

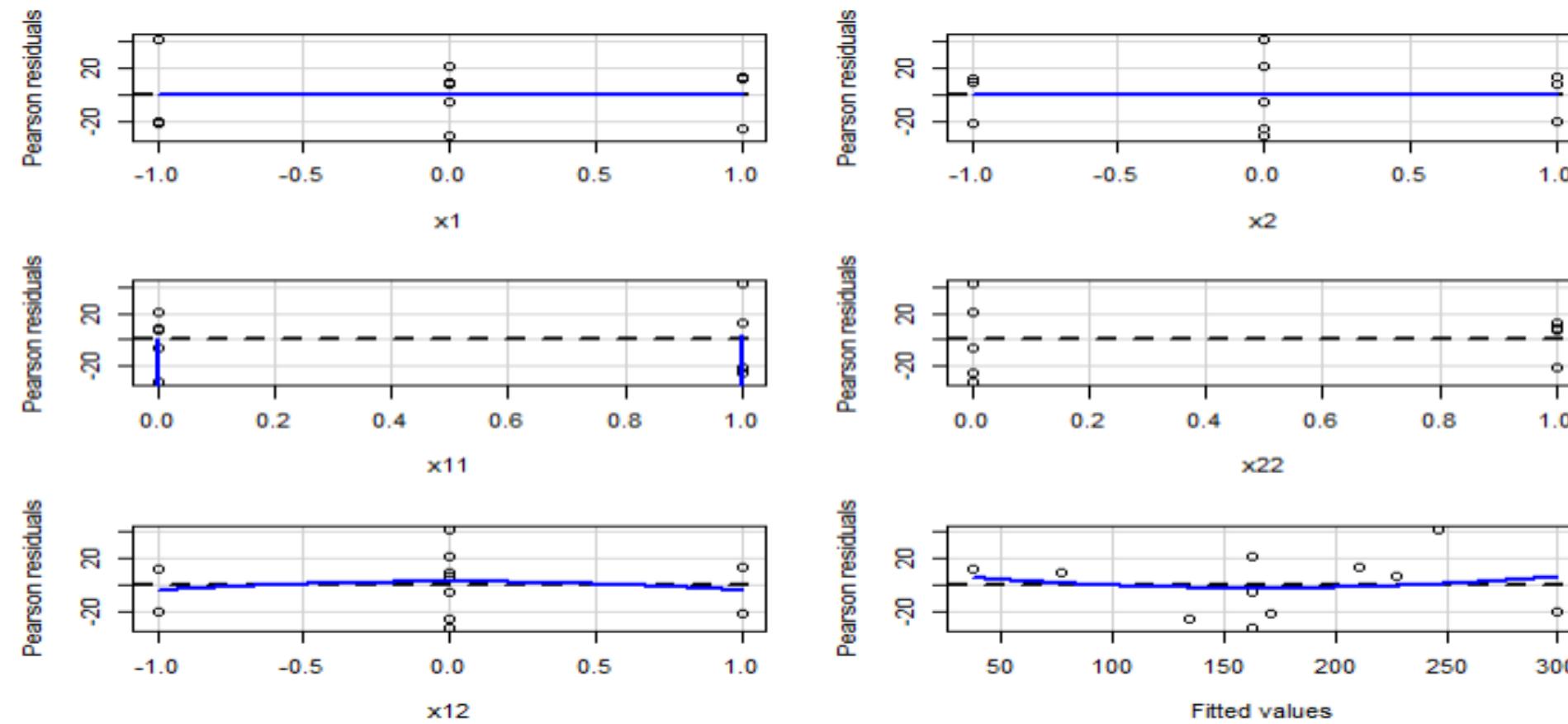
| | sum Sq | Df | F value | Pr (>F) | |
|-----------|--------|----|---------|----------|----|
| x1 | 18704 | 1 | 17.8460 | 0.008292 | ** |
| x2 | 34202 | 1 | 32.6323 | 0.002297 | ** |
| x11 | 1901 | 1 | 1.8140 | 0.235856 | |
| x22 | 285 | 1 | 0.2719 | 0.624352 | |
| x12 | 529 | 1 | 0.5047 | 0.509184 | |
| Residuals | 5240 | 5 | | | |

The Type I and II ANOVA tables may be similar due to a potential lack of multicollinearity issues among the variables.

Fitting of Model

- $\hat{Y} = 162.84 - 55.83x_1 + 75.50x_2 + 27.39x_1^2 - 10.61x_2^2 + 11.5x_1x_2$

Residual Plots



Lack of Fit test

- $H_0: Y_{ij} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon_{ij}$ (Reduced model)

$H_a: Y_{ij} = \mu_j + \epsilon_{ij}$ (Full model)

```
reducedModel1<-lm(Y~x1+x2+x11+x22+x12, data=celln)
fullModel1<-lm(Y~factor(x1)*factor(x2)*factor(x11)*factor(x22)*factor(x12), data=celln)
anova(reducedModel1, fullModel1)
```

Analysis of Variance Table

Model 1: $Y \sim x1 + x2 + x11 + x22 + x12$

Model 2: $Y \sim \text{factor}(x1) * \text{factor}(x2) * \text{factor}(x11) * \text{factor}(x22) * \text{factor}(x12)$

| | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|--------|--------|----|-----------|--------|--------|
| 1 | 5 | 5240.4 | | | | |
| 2 | 2 | 1404.7 | 3 | 3835.8 | 1.8205 | 0.3738 |

$$F_S = \frac{\text{SSLF}}{c-p} / \frac{\text{SSPE}}{n-c} = \frac{3835.77}{3} / \frac{1404.7}{2} = 1.82 \sim F(3,2)$$

Do not reject the H_0 , conclude that there isn't a lack of fit issue.

The GLT (Partial F test): Now consider whether a first-order model would be sufficient

- $H_0: \beta_3 = \beta_4 = \beta_5 = 0$; (Reduced model)
- $H_a: \text{not all } \beta\text{s in } H_0 \text{ equal zero}$ (Full model)

Analysis of Variance Table

| | Response: Y | df | Sum Sq | Mean Sq | F value | Pr (>F) |
|-----------|-------------|----|--------|---------|---------|--------------|
| x1 | | 1 | 18704 | 18704 | 17.8460 | 0.008292 *** |
| x2 | | 1 | 34202 | 34202 | 32.6323 | 0.002297 *** |
| x11 | | 1 | 1646 | 1646 | 1.5704 | 0.265552 |
| x22 | | 1 | 285 | 285 | 0.2719 | 0.624352 |
| x12 | | 1 | 529 | 529 | 0.5047 | 0.509184 |
| Residuals | | 5 | 5240 | 1048 | | |

$$F_S = \frac{\frac{SSR(x_1^2, x_2^2, x_1x_2 | x_1, x_2)}{3}}{\frac{SSE(x_1, x_2, x_1^2, x_2^2, x_1x_2)}{dfE}} = \frac{\frac{1646+285+519}{3}}{1048} = 0.78$$

At $\alpha = 0.05$, the critical value $F(0.95; 3,5) = 5.41$, since $F_S < 5.44$, we do not reject the H_0 .

We conclude that the first-order model is adequate for the range of the charge rates and temperatures considered.

Fit the first-order Model

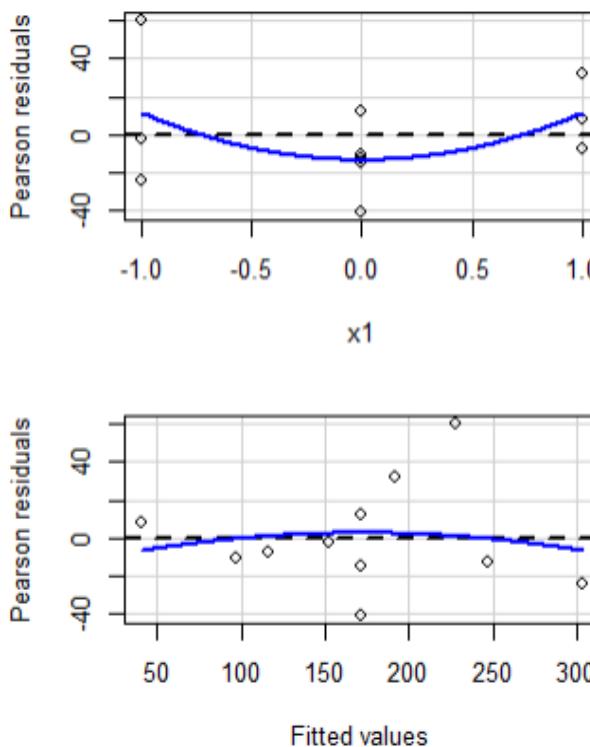
$$Y_{ij} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_{ij}$$

$$\hat{Y} = 172 - 55.83x_1 + 75.5x_2$$

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|----------------|----------|------------|----------|--------------|
| (Intercept) | 172.000 | 9.354 | 18.387 | 7.88e-08 *** |
| x1 | -55.833 | 12.666 | -4.408 | 0.002262 ** |
| x2 | 75.500 | 12.666 | 5.961 | 0.000338 *** |
| <hr/> | | | | |
| signif. codes: | 0 '***' | 0.001 '**' | 0.01 '*' | 0.05 '.' |
| | 0.1 ' ' | 1 | | |

Residual standard error: 31.02 on 8 degrees of freedom
 Multiple R-squared: 0.8729, Adjusted R-squared: 0.8412
 F-statistic: 27.48 on 2 and 8 DF, p-value: 0.0002606



The Lack of Fit test

Model 1: $Y \sim x_1 + x_2$

Model 2: $Y \sim \text{factor}(x_1) * \text{factor}(x_2)$

| Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|--------|-----|--------|-----------|--------|--------------|
| 1 | 8 | 7700.3 | | | |
| 2 | 2 | 1404.7 | 6 | 6295.7 | 1.494 0.4535 |

$$F_S = \frac{SSLF}{c-p} / \frac{SSPE}{n-c} = \frac{6295.7}{6} / \frac{1404.7}{2} = 1.49$$

Do not reject the H_0 , conclude that there isn't a lack of fit.

Simultaneously estimation the regression coefficients (the linear impacts, the betas, etc.). Use the Bonferroni method with 90% confidence level.

$$B = t \left(1 - \frac{\alpha}{2g}, df \right) = t(0.975; 8) = 2.306$$

Where $g = 2$, and $df = n - p = 11 - 3 = 8$

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|--------------|
| (Intercept) | 172.000 | 9.354 | 18.387 | 7.88e-08 *** |
| x1 | -55.833 | 12.666 | -4.408 | 0.002262 ** |
| x2 | 75.500 | 12.666 | 5.961 | 0.000338 *** |

The confidence interval, $b_k \pm B_s\{b_k\}$, represents the estimate of the impact of the transformed X on Y. However, our objective is to determine the impact of the original X on Y.

The original variable X1 and X2 are transformed to x_1 and x_2 through the following transformation function.

$$x_{i1} = \frac{X_{i1} - \bar{X}}{0.4} = \frac{X_{i1} - 1}{0.4}, \quad x_{i2} = \frac{X_{i2} - \bar{X}}{10} = \frac{X_{i2} - 20}{10}$$

We need to back-transform the β to the $old\beta$.

The back transformation process on the betas

$$Y_{ij} = old\beta_0 + old\beta_1 X_{i1} + old\beta_2 X_{i2} + \epsilon_{ij}$$

$$Y_{ij} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_{ij} \quad \text{where } x_{i1} = \frac{X_{i1} - \bar{X}}{0.4} = \frac{X_{i1} - 1}{0.4}, \quad x_{i2} = \frac{X_{i2} - \bar{X}}{10} = \frac{X_{i2} - 20}{10}$$

$$= \beta_0 + \beta_1 \frac{X_{i1}-1}{0.4} + \beta_2 \frac{X_{i2}-20}{10} + \epsilon_{ij} = (\beta_0 - \frac{\beta_1}{0.4} - 2\beta_2) + \frac{\beta_1}{0.4} X_{i1} + \frac{\beta_2}{10} X_{i2}$$

Hence

$$old\beta_1 = \frac{\beta_1}{0.4}, \quad \sigma(old\beta_1) = \frac{\sigma(\beta_1)}{0.4}$$

$$old\beta_2 = \frac{\beta_2}{10}, \quad \sigma(old\beta_2) = \frac{\sigma(\beta_2)}{10}$$



$$oldb_1 = \frac{b_1}{0.4}, \quad s\{oldb_1\} = \frac{s\{b_1\}}{0.4}$$

$$oldb_2 = \frac{b_2}{10}, \quad s\{oldb_2\} = \frac{s\{b_2\}}{10}$$

Estimate the Regression coefficients:

$$oldb_1 = \frac{b_1}{0.4}, \quad s\{oldb_1\} = \frac{s\{b_1\}}{0.4}$$

$$oldb_2 = \frac{b_2}{10}, \quad s\{oldb_2\} = \frac{s\{b_2\}}{10}$$

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|-------------|----------|------------|---------|----------|-----|
| (Intercept) | 172.000 | 9.354 | 18.387 | 7.88e-08 | *** |
| x1 | -55.833 | 12.666 | -4.408 | 0.002262 | ** |
| x2 | 75.500 | 12.666 | 5.961 | 0.000338 | *** |

The Bonferroni Confidence interval for $old\beta$:

$$B = t\left(1 - \frac{\alpha}{2g}, df\right) = t(0.975; 8) = 2.306$$

$$\frac{b_1}{0.4} \pm \frac{Bs\{b_1\}}{0.4} = -\frac{55.833}{0.4} \pm 2.306\left(\frac{12.666}{0.4}\right) = (-212.6, -66.5)$$

$$\frac{b_2}{10} \pm \frac{Bs\{b_2\}}{10} = \frac{75.5}{10} \pm 2.306\left(\frac{12.666}{10}\right) = (4.6, 10.5)$$

Based on the analysis, we can conclude that an increase in charge rate by 1-unit results in a decrease in battery life ranging from at least 66.5 cycles to at most 212.6 cycles. Additionally, an increase in temperature by 1-unit results in an increase in battery life ranging from at least 4.6 cycles to at most 10.5 cycles.

Summary on the battery case

- Linear terms are significant
- Quadratic terms are not significant
- Interaction is not significant
- General linear test shows that quadratic and interaction terms can be omitted
- Type I and Type II SS are almost identical

Caution: the coefficient of estimate of high order item is 0 doesn't necessarily establish that a linear response function is appropriate. Examination of residuals would disclose this lack of fit and should always accompany formal testing of polynomial regression coefficients.

MLR with Qualitative Predictors

Dummy Variable and Baseline Category

An economist conducting a study on the insurance industry aimed to establish a relationship between the adoption speed of a specific insurance innovation (Y) and the size of the insurance firm (X_1) as well as the type of firm (X_2), which could be either a stock company or a mutual company. Notably, X_2 is a qualitative variable.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

X_1 = size of firm

X_2 = mutual company or stock company

Comment:

- Indicator variables with c classes will be represented by $c-1$ indicator variables, each taking on binary values of either 0 or 1. **The 0 status is generally considered the “baseline” by default. In R, it is based on the alphabetical order by default.**
- Indicator variables are frequently called **dummy variables**, or **binary variables**.

$X_2 = 0$ (*The mutual company*). This is the baseline.

$X_2 = 1$ (*The stock company*)

- In situations where there are three kinds of companies, namely mutual, stock, and other (i.e., $c=3$), it is necessary to define two indicator variables (i.e., $c-1$) to represent them.

$$X_2 = \begin{cases} 1 & \text{if mutual company} \\ 0 & \text{Elsewise} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if stock company} \\ 0 & \text{Elsewise} \end{cases}$$

$X_2 = 0, X_3 = 0$ (*The "elsewise"*) This is the baseline.

$X_2 = 1, X_3 = 0$ (*The mutual company*)

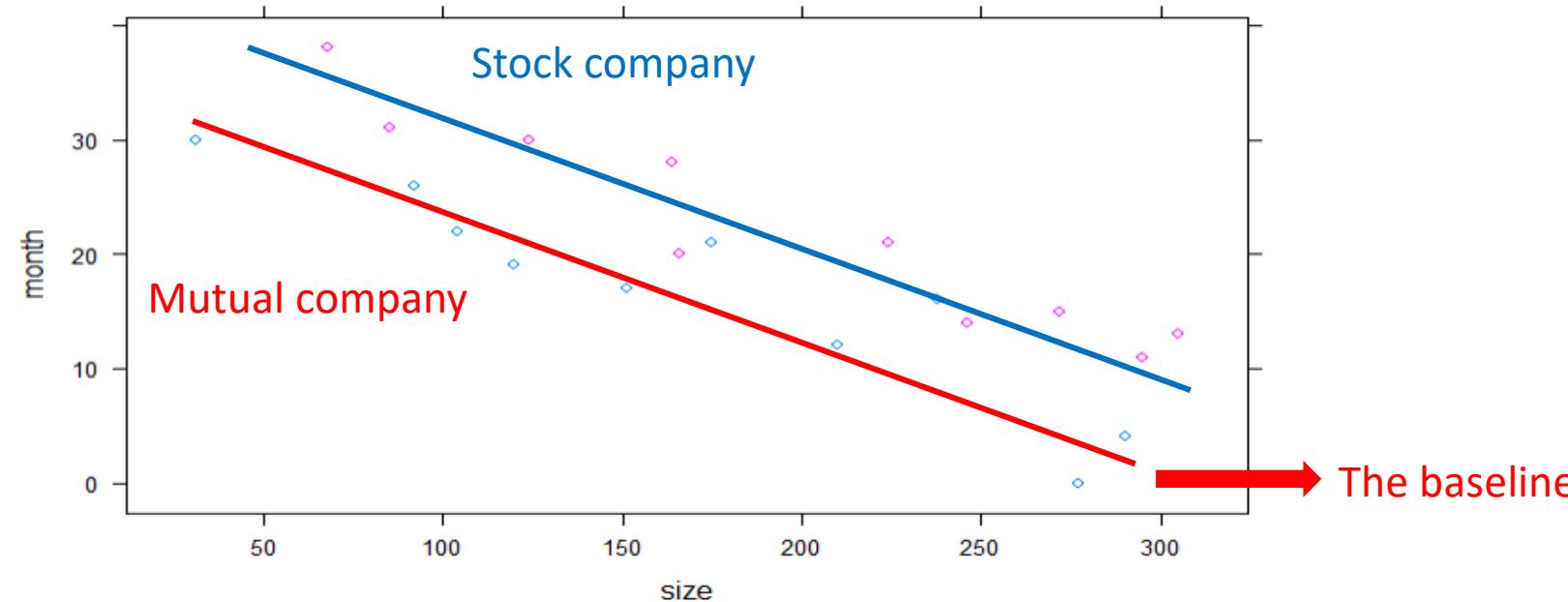
$X_2 = 0, X_3 = 1$ (*The Stock company*)

The Scatter Plot of the MLR with Qualitative Predictors

$X_1 = \text{size of firm}$

$$X_2 = \begin{cases} 0 & \text{if mutual company} \\ 1 & \text{if stock company} \end{cases}$$

| Month (Y) | Size (X1) | Type (X2) |
|-----------|-----------|-----------|
| 17 | 151 | 0 |
| 26 | 92 | 0 |
| 28 | 164 | 1 |
| 11 | 295 | 1 |



The response function: $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

When $X_2 = 0$ (mutual company), the model becomes **the baseline function**

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2(0) + \varepsilon \\ &= \beta_0 + \beta_1 X_1 + \varepsilon \quad (1) \end{aligned}$$

When $X_2 = 1$ (stock company)

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2(1) + \varepsilon \\ &= \beta_0 + \beta_1 X_1 + \beta_2 + \varepsilon \\ &= (\beta_0 + \beta_2) + \beta_1 X_1 + \varepsilon \quad (2) \end{aligned}$$

- The response function for the baseline (1) and the next category (2) exhibit an equivalent slope denoted by β_1 . This implies that the adoption time(Y) changes uniformly with a change in the company size (X1).
- The difference in intercepts, β_2 , reveals the duration difference in adopting a new technology between a stock company ($x_2=1$) and a mutual company ($x_2=0$), considering any given firm size (X1). If $\beta_2 < 0$, it indicates a shorter adoption time for stock companies than mutual companies.
- The effect of company size (x1) on Y is similar for both mutual and stock companies (x2). This characteristic is commonly referred to as the absence of an **interaction effect** between x1 and x2 on Y.
- In the absence of an interaction effect, the distinction in the mean adoption time (Y) between the two types of companies for any specific X1 value is denoted as the **main effect**, β_2 .

Estimate the Coefficients for the MLR

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|-----------|------------|---------|--------------|
| (Intercept) | 33.874069 | 1.813858 | 18.675 | 9.15e-13 *** |
| size | -0.101742 | 0.008891 | -11.443 | 2.07e-09 *** |
| type | 8.055469 | 1.459106 | 5.521 | 3.74e-05 *** |

The 95% CI for β_1 is, $b_1 \pm ts\{b_1\} = -0.102 \pm 2.11(0.00889) = -0.102 \pm 0.0188 = (-0.12, -0.08)$

The 95% CI for β_2 is, $b_2 \pm ts\{b_2\} = 8.06 \pm 2.11(1.46) = 8.06 \pm 3.08 = (5, 11)$

With 95% confidence level, we conclude that

- For both types of companies, the average adoption time decreases by at least 0.08 and at most 0.12 when the company size increases by 1 unit.
- Additionally, at any given level of company size, we observe that stock companies tend to adopt the innovation at least 5 months and at most 11 months later than mutual companies.

Adding the Interaction Term, X_1X_2

`lm(Y~X1 + X2 + X1*X2)`

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|------------|------------|---------|--------------|
| (Intercept) | 33.8383695 | 2.4406498 | 13.864 | 2.47e-10 *** |
| size | -0.1015306 | 0.0130525 | -7.779 | 7.97e-07 *** |
| type | 8.1312501 | 3.6540517 | 2.225 | 0.0408 * |
| size:type | -0.0004171 | 0.0183312 | -0.023 | 0.9821 |

Analysis of Variance Table

| Response: month | | df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------------|----|---------|---------|----------|---------------|--------|
| size | 1 | 1188.17 | 1188.17 | 107.7819 | 1.627e-08 *** | |
| type | 1 | 316.25 | 316.25 | 28.6875 | 6.430e-05 *** | |
| size:type | 1 | 0.01 | 0.01 | 0.0005 | 0.9821 | |
| Residuals | 16 | 176.38 | 11.02 | | | |

$H_0: \beta_3 = 0, H_a: \beta_3 \neq 0$

$$t_s = \frac{b_3}{s\{b_3\}} = -\frac{0.0004171}{0.01833} = -0.02$$

- Do not reject H_0 (p-value = 0.9821)
- The interaction is insignificant
- Can also do a GLT test.

Understand the Coefficients in the MLR with a categorical variable

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

To comprehend the coefficients, we break down the equation:

$$E(Y) = \beta_0 + \beta_1 X_1 \quad (\text{Mutual})$$

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_3 X_1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 \quad (\text{Stock})$$

- Firstly, $\beta_0 + \beta_1 X_1$ describes the linear model for the baseline category (i.e., the mutual company), where the linear impact of X_1 on Y is β_1 for this baseline.
- Secondly, β_2 describes the main category effect difference between the other category (i.e., the stock company) and the baseline. This main effect difference is associated with the category (X_2), not with the other predictor (i.e., X_1).
- Lastly, β_3 describes the interaction effect between X_1 and X_2 , which is associated with X_1 . The linear impact of X_1 on Y is $\beta_1 + \beta_3$ in this category.
- To define the linear model for the other category (i.e., the stock company), we can sum up the above three points and write:

$$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1$$

Understand the Coefficients in the MLR with a categorical variable

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|------------|------------|---------|--------------|
| (Intercept) | 33.8383695 | 2.4406498 | 13.864 | 2.47e-10 *** |
| size | -0.1015306 | 0.0130525 | -7.779 | 7.97e-07 *** |
| type | 8.1312501 | 3.6540517 | 2.225 | 0.0408 * |
| size:type | -0.0004171 | 0.0183312 | -0.023 | 0.9821 |

$$\hat{Y} = b_0 + b_1 X_1 = 33.8 - 0.1X_1 \quad \text{For mutual firms } (X_2 = 0)$$

$$\hat{Y} = b_0 + b_2 + (b_1 + b_3)X_1 = (33.8 + 8.1) - (0.1 + 0.0004)X_1 \quad \text{For stock firms } (X_2 = 1)$$

- When X_1 increases by 1 unit, the mutual firm experiences a decrease of 0.1 in Y , while the stock firm experiences a reduction of 0.0004 more than the mutual firm.
- For a given value of X_1 , the average Y response for the mutual firm is $33.8 - 0.1X_1$, while the stock firm's mean Y response is $(33.8+8.1)-(0.1+0.0004) X_1$ for the stock firm. The difference is $8.1-0.0004 X_1 = b_2 + b_3 X_1$
- β_3 , or the interaction effect is insignificant.

Construct the regression model with qualitative predictors **with three (or more) categories**

Example (insurance): in a study of insurance industry, an economist wished to relate the speed with which a particular insurance innovation is adopted (Y) to the size of the insurance firm (X₁) and the type of firm (type 1, 2 and 3)

Comment:

- Indicator variables with c classes will be represented by c-1 indicator variables, each taking on the values 0 and 1.
- Two dummy variables, X₂ and X₃ are required to describe the categorical variable. X₂=1 only for type 2, and X₃=1 only for type 2. The baseline is type 1 (X₂=0, X₃=0).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \varepsilon$$

The first category is treated as a base line for other categories to compare to

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \varepsilon$$

$$E(Y) = \beta_0 + \beta_1 X_1 \quad \text{For type 1}$$

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_{12} X_1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_{12}) X_1 \quad \text{For type 2}$$

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_3 + \beta_{13} X_1 = (\beta_0 + \beta_3) + (\beta_1 + \beta_{13}) X_1 \quad \text{For type 3}$$

1. $\beta_0 + \beta_1 X_1$ describe the linear model for the baseline category (type1). The linear impact of X_1 on Y is β_1 in the baseline.
2. β_2 describes the main category effect difference between the second category (type 2) and the baseline.
3. β_{12} describes the interaction effect and represents the linear impact difference between type 2 and the baseline.

The response function for type 2 finally sums up to: $Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_{12}) X_1$

Similarly, the linear model for the third category is

$$Y = (\beta_0 + \beta_3) + (\beta_1 + \beta_{13}) X_1$$

Set up hypotheses to compare the two company

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}X_{i2} + \epsilon_i, \text{ where } \beta_3 \neq 0$$

$$\left\{ \begin{array}{l} E(Y) = \beta_0 + \beta_1 X_1 \quad \text{For mutual firm, } X_2 = 0 \\ E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_3 X_1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 \quad \text{For stock firm, } X_2 = 1 \end{array} \right.$$

Question 1 “the mutual firm and the stock firm have the same average adopt time for any firm size.”

Can be tested by

$$H_0: \beta_2 = \beta_3 = 0$$

$$\text{Reduced model } Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$$

$$H_a: \text{Not } H_0$$

$$\text{Full model } Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}X_{i2} + \epsilon_i$$

Set up hypotheses to compare the two company

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}X_{i2} + \epsilon_i, \text{ where } \beta_3 \neq 0$$

$$\left\{ \begin{array}{l} E(Y) = \beta_0 + \beta_1 X_1 \quad \text{For mutual firm, } X_2 = 0 \\ E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_3 X_1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 \quad \text{For stock firm, } X_2 = 1 \end{array} \right.$$

Question 2 “the firm size (X_1) has no impact on the adopt time in mutual firm and stock firm.”

Can be tested by

$$H_0: \beta_1 = \beta_3 = 0$$

$$\text{Reduced model } Y_i = \beta_0 + \beta_2 X_{i2} + \epsilon_i$$

$$H_a: \text{Not } H_0$$

$$\text{Full model } Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}X_{i2} + \epsilon_i$$

Set up hypotheses to compare the two company

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}X_{i2} + \epsilon_i, \text{ where } \beta_3 \neq 0$$

$$\left\{ \begin{array}{l} E(Y) = \beta_0 + \beta_1 X_1 \quad \text{For mutual firm, } X_2 = 0 \\ E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_3 X_1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 \quad \text{For stock firm, } X_2 = 1 \end{array} \right.$$

Question 3 “the firm size (X1) has the same impact on the adopt time in mutual firm and stock firm.”

Can be tested by

$$H_0: \beta_3 = 0$$

$$\text{Reduced model } Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

$$H_a: \text{Not } H_0$$

$$\text{Full model } Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}X_{i2} + \epsilon_i$$

Set up hypotheses to compare the two company

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i, \text{ where } \beta_3 \neq 0$$

$$\left\{ \begin{array}{l} E(Y) = \beta_0 + \beta_1 X_1 \quad \text{For mutual firm, } X_2 = 0 \\ E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_3 X_1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 \quad \text{For stock firm, } X_2 = 1 \end{array} \right.$$

Question 4 “Given the firm size (X_1) has the same impact on the two companies (i.e., $\beta_3 = 0$) , the average adoption time for the stock firm, at any given firm size, is also the same as the mutual firm.”

Can be tested by

$$H_0: \beta_2 = 0$$

$$\text{Reduced model } Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$$

$$H_a: \text{Not } H_0$$

$$\text{Full model } Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

Some considerations in using indicator variables

Many different coding of indicator variables are possible. For example, consider a variable X be the “frequency of product use”

| Code1 | X |
|-----------------|---|
| Frequent user | 3 |
| Occasional user | 2 |
| Nonuser | 1 |

Or

| Code2 | X |
|-----------------|---|
| Frequent user | 6 |
| Occasional user | 3 |
| Nonuser | 1 |

Or

| Code3 | X1 | X2 |
|-----------------|----|----|
| Frequent user | 1 | 0 |
| Occasional user | 0 | 1 |
| Nonuser | 0 | 0 |

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

| Mean | $E\{Y\}=\beta_0 + \beta_1 X_i$ |
|-----------------|--------------------------------|
| Frequent user | $\beta_0 + 3\beta_1$ (1) |
| Occasional user | $\beta_0 + 2\beta_1$ (2) |
| Nonuser | $\beta_0 + 1\beta_1$ (3) |

| Mean | $E\{Y\}=\beta_0 + \beta_1 X_i$ |
|-----------------|--------------------------------|
| Frequent user | $\beta_0 + 6\beta_1$ (4) |
| Occasional user | $\beta_0 + 3\beta_1$ (5) |
| Nonuser | $\beta_0 + 1\beta_1$ (6) |

| Mean | $E\{Y\}=\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$ |
|-----------------|--|
| Frequent user | $\beta_0 + \beta_1$ (7) |
| Occasional user | $\beta_0 + \beta_2$ (8) |
| Nonuser | β_0 (9) |

Note the key implication:

| | Allocation code 1 | Allocation code 2 | Allocation code 3 |
|---|-------------------|--------------------|-----------------------------|
| $E(Y frequent user) - E(Y occasional user)$ | $(1)-(2)=\beta_1$ | $(4)-(5)=3\beta_1$ | $(7)-(8)=\beta_1 - \beta_2$ |
| $E(Y occasional user) - E(Y non user)$ | $(2)-(3)=\beta_1$ | $(5)-(6)=2\beta_1$ | $(8)-(9)=\beta_2$ |

Only code 3 makes no assumption about the spacing of the categories.

Model Building Process, Model Selection Guideline and Criteria

The Model Building Process

1. Planning and Data Collection:

- Identify research questions and objectives
- Plan data collection (decide sample unit, variable, sample size)
- Collect data
- Clean data (check for errors and organize database)

2. Model Exploration:

- Use graphical screening and bivariate modeling to explore data
- Identify relationships and potential outliers
- Recognize possible interactions, especially for qualitative variables
- Discuss possible sources of multicollinearity or other issues
- Use various methods including histograms, scatterplots, contingency tables, boxplots, pairwise correlations, SLR results, residual plots, and diagnostics for individual predictors

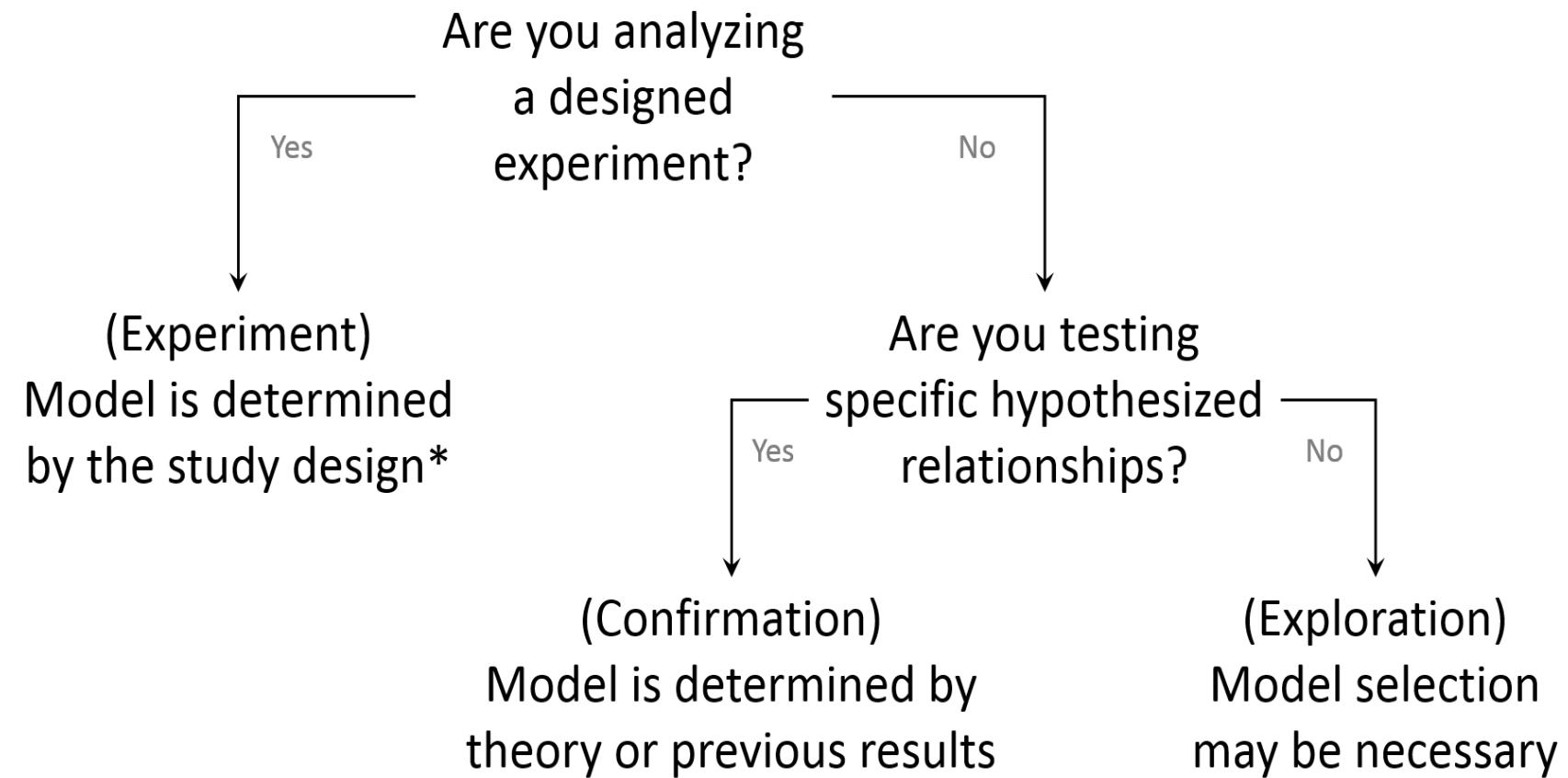
3. Model Selection:

- Fit various regression models
- Compare results to identify best models that align with study objectives
- **Reduce explanatory variables depending on the nature of the study**

4. Model Validation:

- Compare model predictions against theoretical expectations
- Check model's predictive ability with cross-validation

Model Selection Depends on the Nature of Study



* Model selection on *covariates* may be helpful.

The Nature of Study

I. Controlled experiment

- This study type involves controlling the levels of explanatory variables and assigning a treatment to each experimental unit to observe its response. In controlled experiments, the explanatory variables are often called factors or control variables. In controlled experiments, the explanatory variables are often called **factors** or **control variables**.
- For instance, an experiment that examines the impact of graphic presentation size (X_1) and analysis time (X_2) on accuracy (Y). A treatment consists of a specific combination of size and time.
- $Y \sim X_1 + X_2$

II. Controlled Experiments with covariates

- In this study, **uncontrolled variables or covariates** are included to reduce error variance.
- For example, in the previous experiment, gender (X_3) and years of experience (X_4) are measured as uncontrolled variables from each unit.
- $Y \sim X_1 + X_2 + X_3 + X_4$

The Nature of Study

III. Confirmatory observational study

- This type of study is intended to test hypotheses based on observational data, not experimentation.
- The explanatory variables are called **primary variables**, and the variables included to reflect existing knowledge are called **control variables** (or known risk factors in epidemiology).
- In this study, the control variables are not controlled, but they reflect the known influence.
- For instance, in an observational study of the effect of vitamin E supplements (X1) on a certain type of cancer (Y), known risk factors such as age (X2), gender (X3), and race (X4) would be included as control variables, while the amount of vitamin E supplements taken daily would be the primary explanatory variable.
- $Y \sim X_1 + X_2 + X_3 + X_4$

The Nature of Study

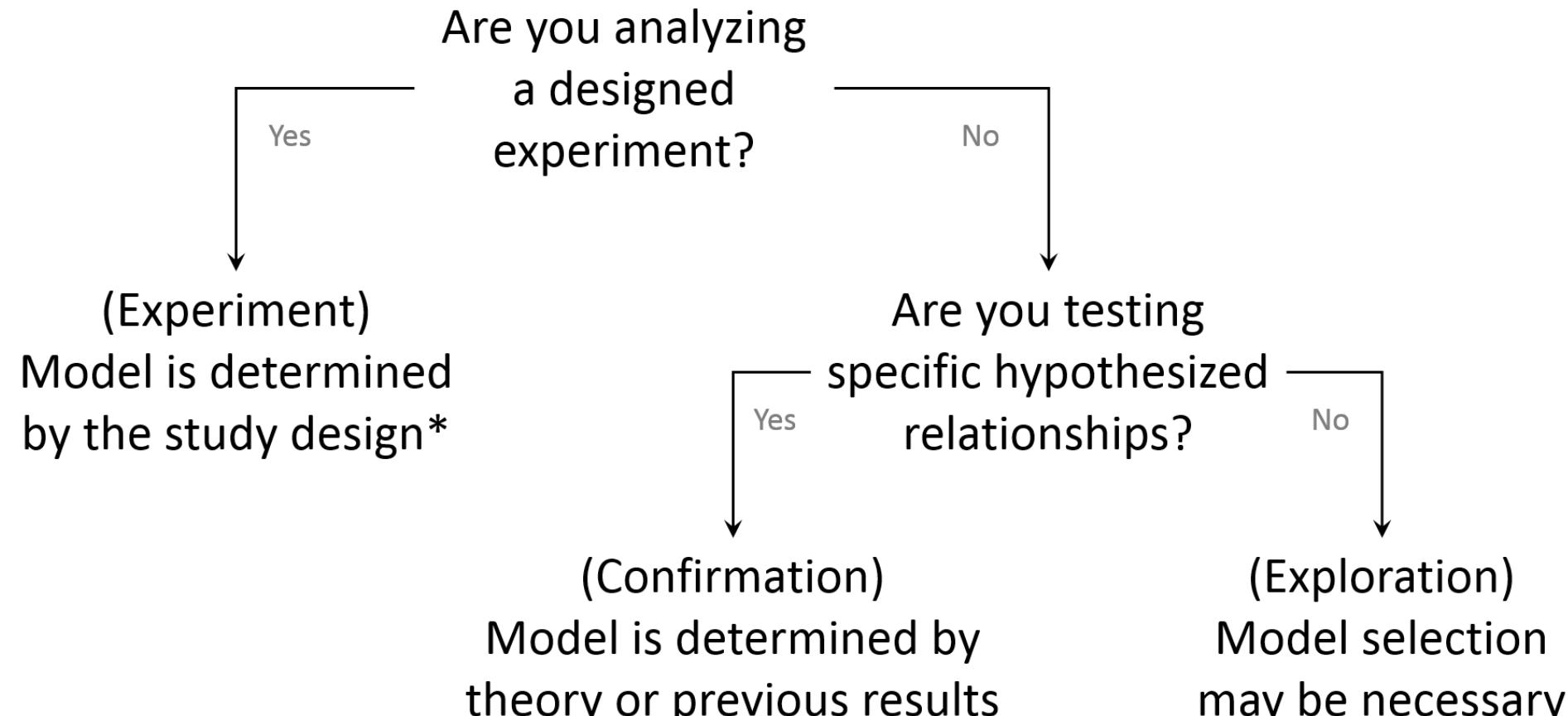
IV. Exploratory observational study

- This study is often used in social, behavioral, health science, management, and other fields when conducting controlled experiments is not possible, or when adequate knowledge for confirmatory observational studies is lacking.
- In this type of study, explanatory variables that are not directly measurable could be involved in any available theoretical model. Variables that could be conceivably related to the response variable are studied.
- The number of cases collected for an exploratory observational regression study depends on the size of the pool of variables. A general rule of thumb suggests that there should **be at least 6 to 10 cases for every variable in the pool.**
- $Y \sim X_1 + X_2 + X_3 + X_1^2 + X_1X_2 + X_1X_3$

Model Selection Guideline on Reduction of Predictors

- In a controlled experiment, the reduction of explanatory variables is usually not an essential issue.
- In controlled experiments with covariates, some reduction of the covariates may take place.
- In a confirmatory observational study, no reduction of primary explanatory variables should generally take place. Even the controlled variables should be retained for comparison with earlier studies.
- In an exploratory observational study, many variables are frequently highly inter-correlated. The main goal is to determine the functional form, interactions, and reduce the variables accordingly.

When is Model Selection Needed?



* Model selection on *covariates* may be helpful.

Methods of Model Selection

1. Selection by design (experiments)

- One or a few specific models based on the design of the experiment

2. Interest/previous knowledge/expert opinion (confirmation, covariates)

Selection informed by study objectives or previous experience

3. Best subsets algorithms

- identify the “best” model with a subset of $p - 1$ predictors, according to some criterion

4. Stepwise algorithms

- construct the model by adding or removing variables one at a time and monitoring changes in a criterion

Some comments on model selection

Many criteria have been proposed to help identify the “best” subset of predictor variables

- Each has benefits and drawbacks
- In some cases, they may lead to different conclusions

In general, you should think of model selection criteria as tools that provide insights about your regression problem, not as magical oracles. Model building is about choices, determined by *you*.

Case Study: Surgical Unit Example

A hospital surgical unit was interested in predicting survival in patient undergoing a particular type of liver operation. A random number of 108 patients was available for analysis, but we only study ($n=$)54. For each patient record, the following information was extracted (data: surgery.csv):

Potential predictors include,

- Blood clotting score (X_1 , `blood`)
- A prognostic index (X_2 , `prog`)
- Enzyme function test (X_3 , `enz`)
- Liver function test (X_4 , `liver`)

The response variable is survival time in days (Y , `surv`)

We skip the model exploration process in this topic.

Check out the MLR diagnostic procedure case for the process of transforming the regression function

Current model $\ln(Y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_4 liver + \epsilon$

Should we delete some predictors?

We now proceed with the model selection process.

Criteria for Model Selection

- R^2
- *Adjusted R² (MSE)*
- *Mallows' C_p*
- *AIC*
- *SBC*
- *PRESS*

Model selection: R_p^2 or SSE_p criterion

We will assume that the number of observations (n) exceeds the maximum number of potential parameters (P): $n > P$

R_p^2 : The multiple determination for p parameters or $p - 1$ predictors

$$R_p^2 = 1 - \frac{SSE_p}{SSTO}$$

Model selection: $R_{a,p}^2$ or MSE_p criterion

Analysis of Variance Table

| | | Response: lny | | | | |
|-----------|----|---------------|---------|----------|-----------|-----|
| | Df | Sum Sq | Mean Sq | F value | Pr(>F) | |
| blood | 1 | 0.7763 | 0.7763 | 12.3337 | 0.0009661 | *** |
| prog | 1 | 2.5888 | 2.5888 | 41.1325 | 5.377e-08 | *** |
| enz | 1 | 6.3341 | 6.3341 | 100.6408 | 1.810e-13 | *** |
| liver | 1 | 0.0246 | 0.0246 | 0.3905 | 0.5349320 | |
| Residuals | 49 | 3.0840 | 0.0629 | | | |

SSTO= 12.8078

The R_p^2 criterion is not intended to identify the subsets since it never decreases.

$$R_{a,p}^2 = 1 - \left(\frac{n-1}{n-p} \right) \frac{SSE_p}{SSTO} = 1 - \frac{MSE_p}{SSTO/(n-1)}$$

Example:

$$\ln(Y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_4 liver + \epsilon$$

$$R_{a,p}^2 = 1 - \left(\frac{n-1}{n-p} \right) \frac{SSE_p}{SSTO}$$

$$= 1 - \left(\frac{54-1}{54-5} \right) \frac{3.084}{12.8078} = 0.7396$$

$$\ln(Y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \epsilon$$

$$R_{a,p}^2 = 1 - \left(\frac{n-1}{n-p} \right) \frac{SSE_p}{SSTO}$$

$$= 1 - \left(\frac{54-1}{54-4} \right) \frac{3.084 + 0.0246}{12.8078} = 0.743$$

Model selection: **Mallows' C_p** criterion

The squared error of the i th fitted value:

$$(\hat{Y}_i - \mu_i)^2$$

The mean(expected) squared error of the i th fitted value :

$$\mathbb{E}(\hat{Y}_i - \mu_i)^2 = (E\{\hat{Y}_i\} - \mu_i)^2 + \sigma^2\{\hat{Y}_i\}$$

The total mean squared error:

$$\Sigma[\mathbb{E}(\hat{Y}_i - \mu_i)^2] = \Sigma(E\{\hat{Y}_i\} - \mu_i)^2 + \Sigma\sigma^2\{\hat{Y}_i\}$$

The total mean squared error divided by the error variance (σ^2):

$$\Gamma_p = \frac{1}{\sigma^2} [\Sigma(E\{\hat{Y}_i\} - \mu_i)^2 + \Sigma\sigma^2\{\hat{Y}_i\}]$$

Which can then be estimated by C_p

$$C_p = \frac{SSE_p}{MSE_{full}} - (n - 2p)$$

Comments:

- When there is no bias in the model with $p-1$ predictors and $E\{\hat{Y}_i\} = \mu$ $C_p \approx P$
- Model is better when C_p is : 1) small and 2) near p
 - It may sometimes occur that the regression model based on a subset of X variables with a small C_p but large bias.
 - One may prefer a model based on a somewhat more X with a slightly larger C_p but smaller bias.

Model selection: Mallows' C_p criterion (should be small and near p)

$$C_p = \frac{SSE_p}{MSE_{full}} - (n - 2p)$$

For example,

$$\ln(Y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_4 liver + \epsilon$$

| Response: lny | | | | | | |
|---------------|----|--------|---------|----------|-----------|-----|
| | Df | Sum Sq | Mean Sq | F value | Pr(>F) | |
| blood | 1 | 0.7763 | 0.7763 | 12.3337 | 0.0009661 | *** |
| prog | 1 | 2.5888 | 2.5888 | 41.1325 | 5.377e-08 | *** |
| enz | 1 | 6.3341 | 6.3341 | 100.6408 | 1.810e-13 | *** |
| liver | 1 | 0.0246 | 0.0246 | 0.3905 | 0.5349320 | |
| Residuals | 49 | 3.0840 | 0.0629 | | | |

$$\begin{aligned} C_p &= \frac{SSE_p}{MSE_{full}} - (n - 2p) \\ &= \frac{SSE_{full}}{MSE_{full}} - (n - 2 * 5) \\ &= n - p - (n - 2p) = p = 5 \end{aligned}$$

C_p of the full model is exactly p.

$$\ln(Y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \epsilon$$

| Response: lny | | | | | | |
|---------------|----|--------|---------|---------|-----------|-----|
| | Df | Sum Sq | Mean Sq | F value | Pr(>F) | |
| blood | 1 | 0.7763 | 0.7763 | 12.486 | 0.0008931 | *** |
| prog | 1 | 2.5888 | 2.5888 | 41.640 | 4.307e-08 | *** |
| enz | 1 | 6.3341 | 6.3341 | 101.883 | 1.174e-13 | *** |
| Residuals | 50 | 3.1085 | 0.0622 | | | |

$$\begin{aligned} C_p &= \frac{SSE_p}{MSE_{full}} - (n - 2p) \\ &= \frac{3.1085}{0.0629} - (54 - 2 * 4) = 3.4 \end{aligned}$$

C_p is close to p=4: indicating little or no bias in this model.

$$\ln(Y) = \beta_0 + \beta_2 prog + \beta_4 liver + \epsilon$$

| Response: lny | | | | | | |
|---------------|----|--------|---------|---------|-----------|-----|
| | Df | Sum Sq | Mean Sq | F value | Pr(>F) | |
| prog | 1 | 2.8285 | 2.8285 | 21.784 | 2.247e-05 | *** |
| liver | 1 | 3.3572 | 3.3572 | 25.855 | 5.321e-06 | *** |
| Residuals | 51 | 6.6220 | 0.1298 | | | |

$$\begin{aligned} C_p &= \frac{SSE_p}{MSE_{full}} - (n - 2p) \\ &= \frac{6.622}{0.0629} - (54 - 2 * 3) = 57.28 \end{aligned}$$

C_p is larger than in the second model.
Plus, it is biased because C_p is much larger than $p(=3)$ in this case.

Model selection: AIC_p and SBC_p criteria

Akaike's information criterion

$$AIC_p = n * \ln(SSE_p) - n * \ln(n) + 2p$$

Schwarz's Bayesian criterion

$$SBC_p = n * \ln(SSE_p) - n * \ln(n) + [\ln(n)] * p$$

Aka Bayesian information criterion (BIC)

Comments:

- Both methods based on the Maximum Likelihood method.
 - The model does a good job explaining the ***current*** data. But there is chance of overfitting for the future data.
 - Can be used to compare candidate models with different error distributions which ***may not be Normal***.
 - ***Do not*** assume any form of nesting, i.e., the p predictors are a subset of the full model. But all models need to be trained on the same data.
- The better the model, the smaller AIC_p or SBC_p is.
- AIC_p and SBC_p differ in the way they penalize for model complexity.
 - The AIC_p penalizes for the number of parameters in the model, while the SBC_p penalizes for both the number of parameters and the number of observations in the model.
 - In general, AIC_p is more suitable for small datasets, while SBC_p is more suitable for large datasets.
 - If $n \geq 8$, the penalty for SBC_p is larger than that for AIC_p ; hence the SBC_p tends to favor simpler models
- AIC_p and C_p will tend to pick the same model.
- If the true model is a candidate,
 - AIC_p and C_p will tend to pick more complex models than the truth
 - SBC_p will tend to pick the true model more often

Model selection: AIC_p and SBC_p criteria

Akaike's information criterion

$$AIC_p = n * \ln(SSE_p) - n * \ln(n) + 2p$$

Schwarz's Bayesian criterion

$$SBC_p = n * \ln(SSE_p) - n * \ln(n) + [\ln(n)] * p$$

Example

$$\ln(Y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \epsilon$$

$$AIC_4 = 54 * \ln(3.1085) - 54 * \ln(54) + 2(4) = -146.162$$

| Response: lny | | | | | | |
|---------------|----|--------|---------|---------|-----------|-----|
| | Df | Sum Sq | Mean Sq | F value | Pr(>F) | |
| blood | 1 | 0.7763 | 0.7763 | 12.486 | 0.0008931 | *** |
| prog | 1 | 2.5888 | 2.5888 | 41.640 | 4.307e-08 | *** |
| enz | 1 | 6.3341 | 6.3341 | 101.883 | 1.174e-13 | *** |
| Residuals | 50 | 3.1085 | 0.0622 | | | |

$$SBC_4 = 54 * \ln(3.1085) - 54 * \ln(54) + \ln(54)*4 = -138.206$$

$$\ln(Y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_4 liver + \epsilon$$

$$AIC_5 = 54 * \ln(3.084) - 54 * \ln(54) + 2(5) = -144.59$$

| Response: lny | | | | | | |
|---------------|----|--------|---------|----------|-----------|-----|
| | Df | Sum Sq | Mean Sq | F value | Pr(>F) | |
| blood | 1 | 0.7763 | 0.7763 | 12.3337 | 0.0009661 | *** |
| prog | 1 | 2.5888 | 2.5888 | 41.1325 | 5.377e-08 | *** |
| enz | 1 | 6.3341 | 6.3341 | 100.6408 | 1.810e-13 | *** |
| liver | 1 | 0.0246 | 0.0246 | 0.3905 | 0.5349320 | |
| Residuals | 49 | 3.0840 | 0.0629 | | | |

$$SBC_5 = 54 * \ln(3.084) - 54 * \ln(54) + \ln(54)*5 = -134.645$$

Model selection: PRESS_p criterion

- The Prediction Sum of Squares (PRESS) criterion measures the effectiveness of using the fitted values from a subset model to predict the observed response.
- It differs from the Sum of Squares Error (SSE) in that each fitted value is obtained by **excluding the i th observation from the dataset**, and the model is estimated using the remaining **$n-1$ observations**, this predicted value is denoted by $\hat{Y}_{i(i)}$.
- PRESS is also referred to as "leave-one-out-cross-validation."

$$\text{PRESS}_p = \sum (Y_i - \hat{Y}_{i(i)})^2 \quad \text{SSE}_p = \sum (Y_i - \hat{Y}_i)^2$$

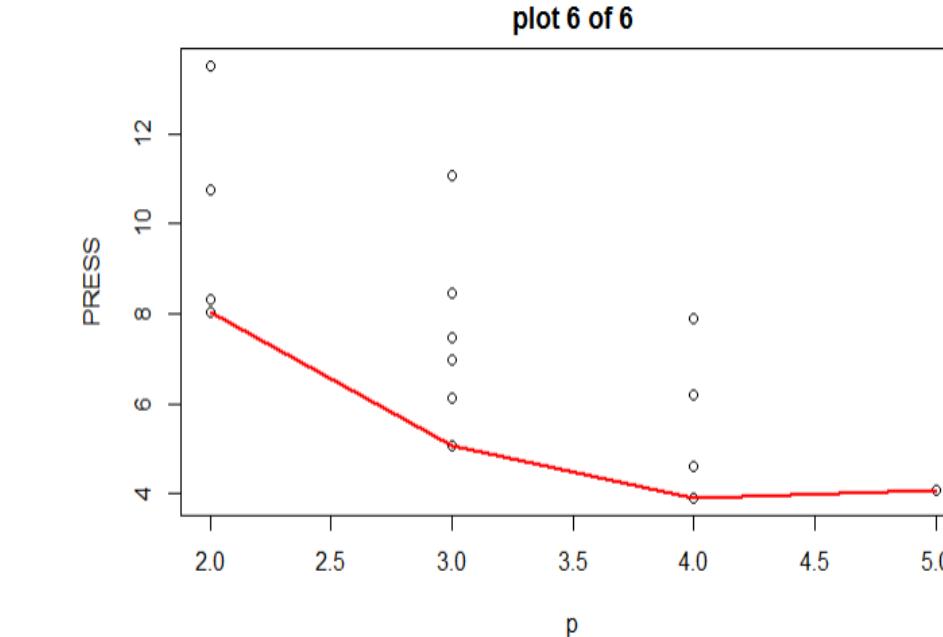
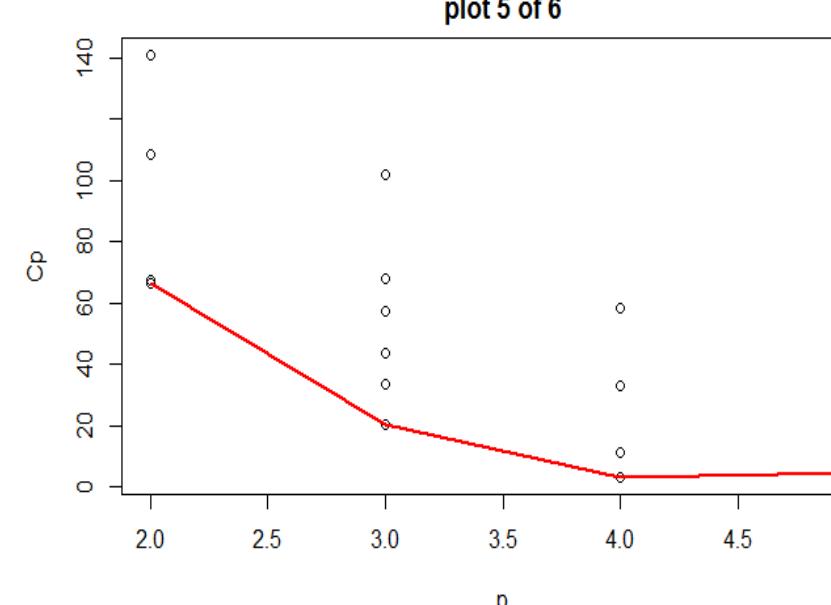
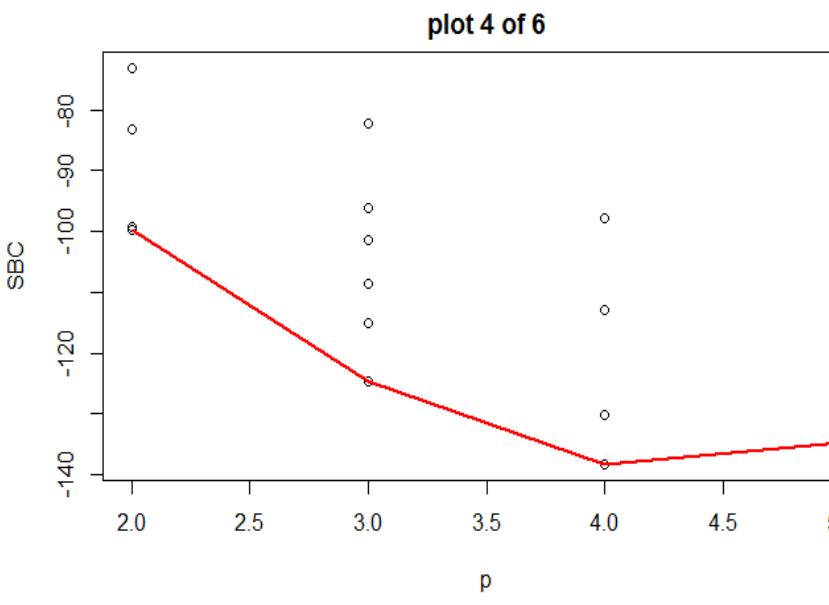
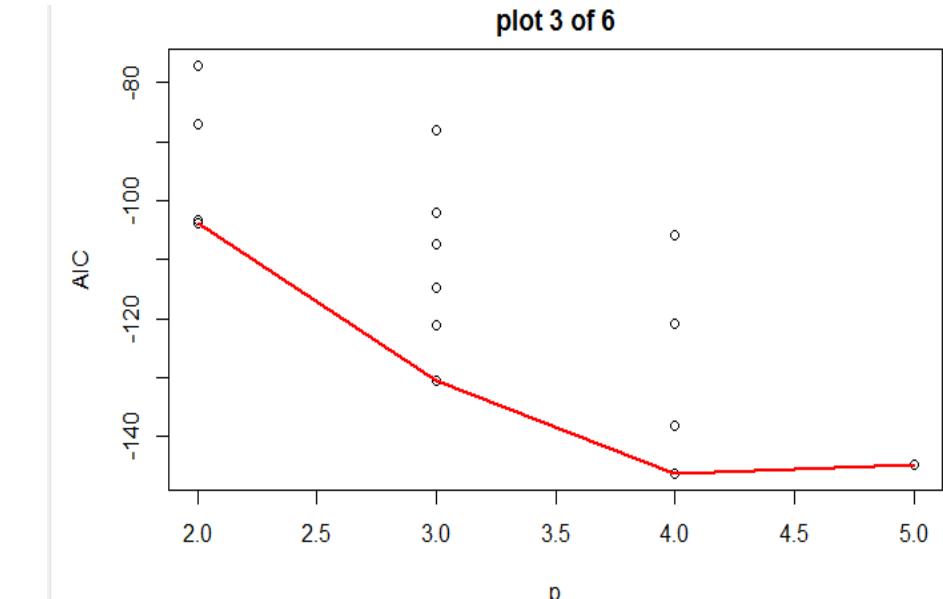
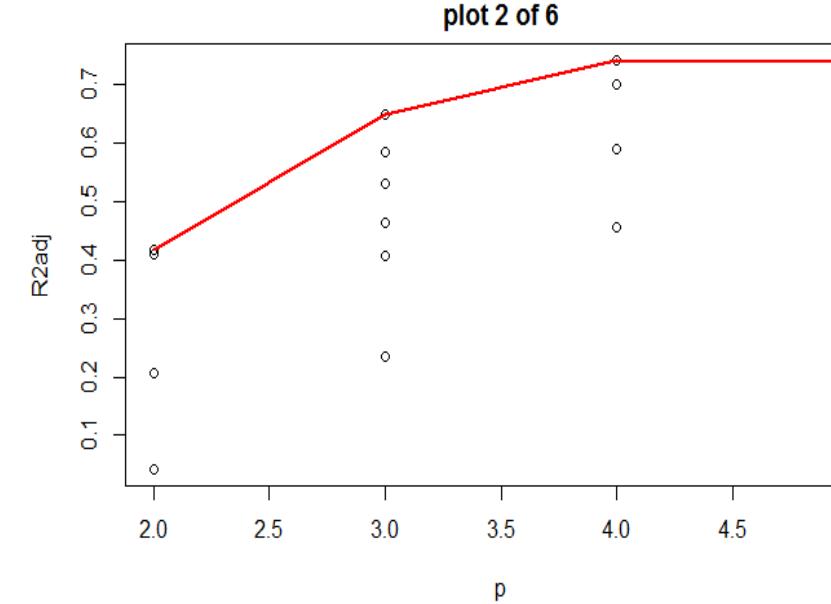
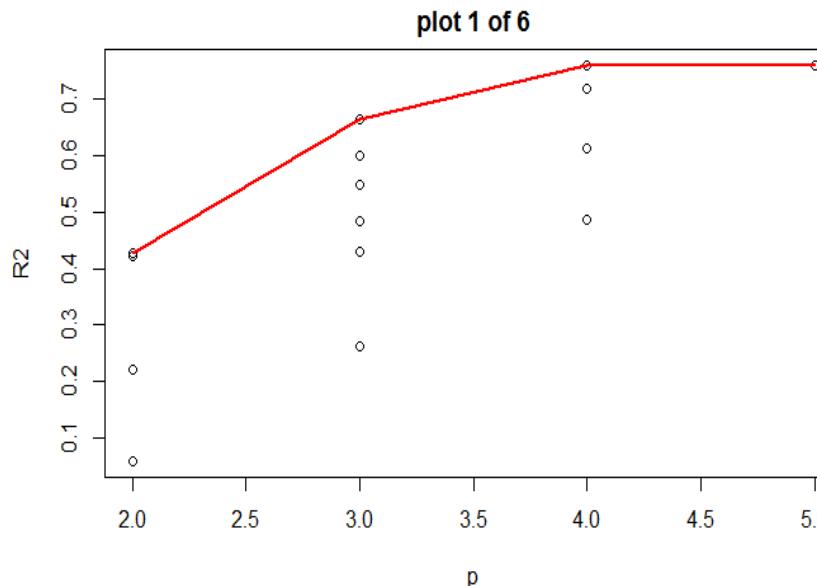
- Models with small PRESS_p values are considered good.
- It is not necessary to refit the model n times, PRESS can be calculated using the information in the Hat matrix

$$\text{PRESS}_p = \sum \left(\frac{Y_i - \hat{Y}_i}{1 - H_{ii}} \right)^2, \text{ where } H_{ii} \text{ is the } i\text{th diagonal element of the Hat matrix.}$$

- When the purpose of multiple linear regression (MLR) is to make predictions, it is recommended to use the PRESS criterion for model selection, since it is a **measure of the predictive accuracy of the model**, which is what **matters most**.

Plot of variables selection criteria-Surgical Unit Example

```
library(ALSM)
plotmodel.s(sur[,1:4], sur$lyn)
```



Plot of variables selection

- Plots of variable selection criteria show the criteria for each possible subset of variables.
- There are six criteria used in the plots.
- The subset with the optimal criterion can be chosen based on the plots.
- Note that the plots do not tell you exactly which variables are selected, only the number of variables in the best subset.
- For example, subsets x_1, x_2, x_3 and x_1, x_2, x_4 both have the same number of variables, but they are different subsets. More on this will be introduced in the next topic.

Model Selection Algorithm and Guideline

Model selection algorithm

- “Best” subsets algorithms
 - Provide the best subsets according to the specified criterion and identify several good subsets for each possible number of X variables to give the additional information.
 - Use when the potential X variables is relatively small, <30.
- Stepwise regression methods
 - Develops the best subsets sequentially.
 - Contain both forward selection and backward elimination.
 - Only a single regression model is identified by the stepwise regression method. The last model Could be suboptimal.
- When the pool of X variables is very large, we should use the subset identified by the stepwise search procedures as a starting point. One may treat the selected number of X variables in the regression model as the right subset size and then use the “best” subsets procedures.

Case Study: Surgical Unit Example (8 X variables, $p = 9$)

A hospital surgical unit was interested in predicting survival in patient undergoing a particular type of liver operation. A random number of 108 patients was available for analysis, but we only study ($n=$)54. For each patient record, the following information was extracted (data: surgery.csv):

Potential predictors include,

- Blood clotting score (X_1 , [blood](#))
- A prognostic index (X_2 , [prog](#))
- Enzyme function test (X_3 , [enz](#))
- Liver function test (X_4 , [liver](#))

- Age (X_5 , [age](#))
- Gender (X_6 , [gender](#) 0 = *male*, 1 = *female*)
- History of alcohol use (3 levels 2 indicator variables X_7 and X_8)

| Alcohol Use | X7 | X8 |
|-------------|----|----|
| None | 0 | 0 |
| Moderate | 1 | 0 |
| Severe | 0 | 1 |

The response variable is survival time in days (Y , [surv](#))

“Best” subsets algorithms

Y: (ln transformed) survival in days (sur)

X1: blood clotting score (blood)
X2: prognostic index (prog)
X3: enzyme function (enz)
X4: liver function test (liver)

X5: age (age)
X6: gender (gender 0=male, 1=female)
X7: history of alcohol use (X7, 1=moderate, 0=otherwise)
X8: history of alcohol use (X8, 1= severe, 0=otherwise)

```
library(ALSM)
bs<-BestSub(sur[,1:8], sur$lny, num=1)
```

| | p | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | SSEp | r2 | r2.adj | Cp | AICp | SBCp | PRESSp |
|---|---|---|---|---|---|---|---|---|---|----------|-----------|-----------|------------|-----------|------------|----------|
| 1 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 7.331575 | 0.4275662 | 0.4165579 | 117.409441 | -103.8269 | -99.84889 | 8.326716 |
| 2 | 3 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 4.312491 | 0.6632899 | 0.6500855 | 50.471575 | -130.4833 | -124.51634 | 5.065339 |
| 3 | 4 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 2.842883 | 0.7780337 | 0.7647157 | 18.914496 | -150.9849 | -143.02899 | 3.469403 |
| 4 | 5 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 2.178799 | 0.8298840 | 0.8159970 | 5.750774 | -163.3514 | -153.40643 | 2.737771 |
| 5 | 6 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 2.082008 | 0.8374413 | 0.8205081 | 5.540639 | -163.8052 | -151.87127 | 2.782713 |
| 6 | 7 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 2.005225 | 0.8434363 | 0.8234494 | 5.787389 | -163.8343 | -149.91140 | 2.772325 |
| 7 | 8 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1.972032 | 0.8460279 | 0.8225974 | 7.029455 | -162.7356 | -146.82378 | 2.808705 |
| 8 | 9 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1.970742 | 0.8461286 | 0.8187737 | 9.000000 | -160.7710 | -142.87013 | 2.931232 |

Q: Among these 8 subsets, which is identified as the best under each criterion.

SSEp 8 R2 8 R2.adj 6 Cp 5, AICp 6, SBC 4, PRESSp 4

“Best” subsets algorithms

Y: (ln transformed) survival in days (sur)

X1: blood clotting score (blood)
X2: prognostic index (prog)
X3: enzyme function (enz)
X4: liver function test (liver)

X5: age (age)
X6: gender (gender 0=male, 1=female)
X7: history of alcohol use (X7, 1=moderate, 0=otherwise)
X8: history of alcohol use (X8, 1= severe, 0=otherwise)

| | p | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | SSEp | r2 | r2.adj | Cp | AICp | SBCp | PRESSp |
|---|---|---|---|---|---|---|---|---|---|----------|-----------|-----------|------------|------------|------------|-----------|
| 1 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 7.331575 | 0.4275662 | 0.4165579 | 117.409441 | -103.82686 | -99.84889 | 8.326716 |
| 1 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 7.408731 | 0.4215420 | 0.4104178 | 119.171240 | -103.26154 | -99.28357 | 8.024956 |
| 1 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 9.979182 | 0.2208467 | 0.2058629 | 177.865004 | -87.17808 | -83.20011 | 10.743872 |
| 2 | 3 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 4.312491 | 0.6632899 | 0.6500855 | 50.471575 | -130.48329 | -124.51634 | 5.065339 |
| 2 | 3 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 5.129702 | 0.5994837 | 0.5837772 | 69.131808 | -121.11257 | -115.14561 | 6.120508 |
| 2 | 3 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 5.780964 | 0.5486346 | 0.5309340 | 84.002738 | -114.65834 | -108.69138 | 6.987582 |
| 3 | 4 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 2.842883 | 0.7780337 | 0.7647157 | 18.914496 | -150.98493 | -143.02899 | 3.469403 |
| 3 | 4 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 3.108539 | 0.7572918 | 0.7427294 | 24.980500 | -146.16088 | -138.20494 | 3.914240 |
| 3 | 4 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 3.614131 | 0.7178164 | 0.7008853 | 36.525190 | -138.02317 | -130.06723 | 4.596928 |
| 4 | 5 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 2.178799 | 0.8298840 | 0.8159970 | 5.750774 | -163.35135 | -153.40643 | 2.737771 |
| 4 | 5 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 2.376584 | 0.8144413 | 0.7992937 | 10.267014 | -158.65926 | -148.71434 | 3.021034 |
| 4 | 5 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 2.705477 | 0.7887621 | 0.7715182 | 17.776952 | -151.66012 | -141.71520 | 3.505131 |
| 5 | 6 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 2.082008 | 0.8374413 | 0.8205081 | 5.540639 | -163.80517 | -151.87127 | 2.782713 |
| 5 | 6 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 2.102923 | 0.8358083 | 0.8187050 | 6.018212 | -163.26542 | -151.33152 | 2.738932 |
| 5 | 6 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 2.137098 | 0.8331399 | 0.8157587 | 6.798576 | -162.39490 | -150.46099 | 2.829352 |
| 6 | 7 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 2.005225 | 0.8434363 | 0.8234494 | 5.787389 | -163.83429 | -149.91140 | 2.772325 |
| 6 | 7 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 2.059621 | 0.8391892 | 0.8186601 | 7.029456 | -162.38896 | -148.46607 | 2.839169 |
| 6 | 7 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 2.065608 | 0.8387217 | 0.8181330 | 7.166172 | -162.23220 | -148.30932 | 2.874944 |
| 7 | 8 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1.972032 | 0.8460279 | 0.8225974 | 7.029455 | -162.73565 | -146.82378 | 2.808705 |
| 7 | 8 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 2.002941 | 0.8436146 | 0.8198168 | 7.735230 | -161.89584 | -145.98397 | 2.882665 |
| 7 | 8 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 2.044284 | 0.8403866 | 0.8160976 | 8.679263 | -160.79256 | -144.88069 | 2.943495 |
| 8 | 9 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1.970742 | 0.8461286 | 0.8187737 | 9.000000 | -160.77098 | -142.87013 | 2.931232 |

Stepwise regression model (both directions)

- The stepwise regression first fits a SLR for each of the $p - 1$ X variables.

$$t_k^* = b_k / s\{b_k\}$$

- The X variable with the largest t^* value is the candidate for the first addition. If the t^* is large, or the P-value is less than some predetermined α , the X variable is added.
- The second variable is added to the model with the first variable. The T test or Partial F test is obtained to determine its significance.

$$F_{new}^* = \frac{MSR(X_{new}|X_{old})}{MSE(X_{old}, X_{new})} \text{ or } t_{new}^* = \sqrt{F_{new}^*}$$

- After two variables are added, the algorithm examines whether any of the variables already in the model should now be dropped.
- Continue till no further variables can either be added or deleted, then the process terminates. AIC is computed for each model in each step
- Since variable can be added and/or removed in each step, this is also the “both” stepwise.
 - If variable can only be added in each step, the method is called “forward” stepwise.
 - If variable can only be removed in each step, the method is called “backward” stepwise.

Stepwise regression model

```
step(lm(lny~blood+prog+enz+liver+age+gender+x7+x8, data=sur), method="both", trace=1)
```

Y: (ln transformed) survival in days (sur)

X1: blood clotting score (blood)

X2: prognostic index (prog)

X3: enzyme function (enz)

X4: liver function test (liver)

X5: age (age)

X6: gender (gender 0=male, 1=female)

X7: history of alcohol use (X7, 1=moderate, 0=otherwise)

X8: history of alcohol use (X8, 1= severe, 0=otherwise)

$\ln(y) \sim \text{blood} + \text{prog} + \text{enz} + \text{age} + \text{gender} + X8$

Start: AIC=-160.77

$\ln y \sim \text{blood} + \text{prog} + \text{enz} + \text{liver} + \text{age} + \text{gender} + x7 + x8$

| | DF | Sum of Sq | RSS | AIC |
|----------|----|-----------|--------|---------|
| - liver | 1 | 0.00129 | 1.9720 | -162.74 |
| - x7 | 1 | 0.03220 | 2.0029 | -161.90 |
| - age | 1 | 0.07354 | 2.0443 | -160.79 |
| <none> | | | 1.9707 | -160.77 |
| - gender | 1 | 0.08415 | 2.0549 | -160.51 |
| - blood | 1 | 0.31809 | 2.2888 | -154.69 |
| - x8 | 1 | 0.84573 | 2.8165 | -143.49 |
| - prog | 1 | 2.09045 | 4.0612 | -123.72 |
| - enz | 1 | 2.99085 | 4.9616 | -112.91 |

Step: AIC=-162.74

$\ln y \sim \text{blood} + \text{prog} + \text{enz} + \text{age} + \text{gender} + x7 + x8$

| | DF | Sum of Sq | RSS | AIC |
|----------|----|-----------|--------|----------|
| - x7 | 1 | 0.0332 | 2.0052 | -163.834 |
| <none> | | | 1.9720 | -162.736 |
| - age | 1 | 0.0876 | 2.0596 | -162.389 |
| - gender | 1 | 0.0971 | 2.0691 | -162.141 |
| - blood | 1 | 0.6267 | 2.5988 | -149.833 |
| - x8 | 1 | 0.8446 | 2.8166 | -145.486 |
| - prog | 1 | 2.6731 | 4.6451 | -118.471 |
| - enz | 1 | 5.0986 | 7.0706 | -95.784 |

Step: AIC=-163.83

$\ln y \sim \text{blood} + \text{prog} + \text{enz} + \text{age} + \text{gender} + x8$

| | DF | Sum of Sq | RSS | AIC |
|----------|----|-----------|----------|----------|
| <none> | | 2.0052 | -163.834 | |
| - age | 1 | 0.0768 | 2.0820 | -163.805 |
| - gender | 1 | 0.0977 | 2.1029 | -163.265 |
| - blood | 1 | 0.6282 | 2.6335 | -151.117 |
| - x8 | 1 | 0.9002 | 2.9055 | -145.809 |
| - prog | 1 | 2.7626 | 4.7678 | -119.064 |
| - enz | 1 | 5.0801 | 7.0853 | -97.672 |

Model Selection Guideline Continued

- It is important to consider the fundamental nature of the explanatory variables.
 - For example, all indicator variables that define a qualitative predictor should be retained in the model.
 - In situations where second order terms X_k^2 or interaction terms $X_k X_m$ are necessary, it is generally recommended to also include the first-order terms, such as X_k and X_m .
- It is crucial to carefully consider the relevance and significance of each variable and to avoid overfitting the model to the data.

For example, after the stepwise algorithm

The algorithm suggests

$$\ln(y) \sim \text{blood} + \text{prog} + \text{enz} + \text{age} + \text{gender} + X_8$$

We should suggest

$$\ln(y) \sim \text{blood} + \text{prog} + \text{enz} + \text{age} + \text{gender} + X_7 + X_8$$

| p | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | SSEp | r2 | r2.adj | Cp | AICp | SBCp | PRESSp |
|---|---|---|---|---|---|---|---|---|----------|-----------|-----------|------------|-----------|------------|----------|
| 1 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 7.331575 | 0.4275662 | 0.4165579 | 117.409441 | -103.8269 | -99.84889 | 8.326716 |
| 2 | 3 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 4.312491 | 0.6632899 | 0.6500855 | 50.471575 | -130.4833 | -124.51634 | 5.065339 |
| 3 | 4 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 2.842883 | 0.7780337 | 0.7647157 | 18.914496 | -150.9849 | -143.02899 | 3.469403 |
| 4 | 5 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2.178799 | 0.8298840 | 0.8159970 | 5.750774 | -163.3514 | -153.40643 | 2.737771 |
| 5 | 6 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 2.082008 | 0.8374413 | 0.8205081 | 5.540639 | -163.8052 | -151.87127 | 2.782713 |
| 6 | 7 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 2.005225 | 0.8434363 | 0.8234494 | 5.787389 | -163.8343 | -149.91140 | 2.772325 |
| 7 | 8 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1.972032 | 0.8460279 | 0.8225974 | 7.029455 | -162.7356 | -146.82378 | 2.808705 |
| 8 | 9 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1.970742 | 0.8461286 | 0.8187737 | 9.000000 | -160.7710 | -142.87013 | 2.931232 |

```
library(ALSM)
reducedm<-lm(lny~blood+prog+enz+age+gender+x7+x8, data=sur)
fullm<-lm(lny~blood+prog+enz+liver+age+gender+x7+x8, data=sur)
cpc(reducedm, fullm)
AICp(reducedm)
SBCp(reducedm)
pressc(reducedm)
```

```
[1] 7.029455
[1] -162.7356
[1] -146.8238
[1] 2.808705
```

Model selected by different approaches in the surgical unit example

Y: (ln transformed) survival in days (sur)

X1: blood clotting score (blood)
 X2: prognostic index (prog)
 X3: enzyme function (enz)
 X4: liver function test (liver)

X5: age (age)
 X6: gender (gender 0=male, 1=female)
 X7: history of alcohol use (X7, 1=moderate, 0=otherwise)
 X8: history of alcohol use (X8, 1= severe, 0=otherwise)

| | p | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 55Ep | r2 | r2.adj | Cp | AICp | SBCp | PRESSp | |
|---------------------------|---|---|---|---|---|---|---|---|---|----------|-----------|-----------|------------|-----------|------------|------------|----------|
| 1 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 7.331575 | 0.4275662 | 0.4165579 | 117.409441 | -103.8269 | -99.84889 | 8.326716 | |
| 2 | 3 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 4.312491 | 0.6632899 | 0.6500855 | 50.471575 | -130.4833 | -124.51634 | 5.065339 | |
| 3 | 4 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 2.842883 | 0.7780337 | 0.7647157 | 18.914496 | -150.9849 | -143.02899 | 3.469403 | |
| Model 1 | 4 | 5 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2.178799 | 0.8298840 | 0.8159970 | 5.750774 | -163.3514 | -153.40648 | 2.737771 | |
| Model 2 | 5 | 6 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 2.082008 | 0.8374413 | 0.8205081 | 5.540639 | -163.8052 | -151.87127 | 2.782713 | |
| Model 3 | 6 | 7 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 2.005225 | 0.8434363 | 0.8234494 | 5.787389 | -163.8343 | -149.91140 | 2.772325 |
| Model 4 (stepwise) | 7 | 8 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1.972032 | 0.8460279 | 0.8225974 | 7.029455 | -162.7356 | -146.82378 | 2.808705 | |
| | 8 | 9 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1.970742 | 0.8461286 | 0.8187737 | 9.000000 | -160.7710 | -142.87013 | 2.931232 | |

$$Model1: \ln(y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_8 X8 + \beta_7 X7$$

X7 can also be added as a option because
 X7 and X8 both belong to one indicator.

$$Model2: \ln(y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_6 gender + \beta_8 X8 + \beta_7 X7$$

$$Model3: \ln(y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_5 age + \beta_6 gender + \beta_8 X8 + \beta_7 X7$$

$$Model4: \ln(y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_5 age + \beta_6 gender + \beta_7 X7 + \beta_8 X8$$

Model Validation: K-fold Cross-validation

- A method for cross-validation that aims to reduce similarity among training datasets is k-fold cross validation. This method involves the following steps:
 1. Randomly split the dataset into k parts, or folds, where k is typically 5 or 10.
 2. Fit the model using $k-1$ folds, and then calculate the predictive mean squared error (MSE) for the remaining testing set.
 3. Repeat steps 1 and 2 k times, using each fold as the testing set exactly once.
 4. Take the average MSE over the k folds to obtain an estimate of the generalization error of the model.
- K-fold cross-validation is a widely used technique for assessing the performance of a model and selecting optimal hyperparameters. It can help to reduce overfitting by providing a more realistic estimate of the model's performance on unseen data. The choice of the number of folds (k) should depend on the size and complexity of the dataset and the computational resources available.

K-fold Cross-validation result

```
library(MASS)
library(leaps)
library(caret)

set.seed(123) #set seed for reproducibility

train.control<-trainControl(method="cv", number=10) #10 fold cross validation

step.model1<-train(lny~blood+prog+enz+x7+x8, data=sur, method="leapBackward",
                     tuneGrid=data.frame(nvmax=5),
                     trControl=train.control)
step.model1$results

step.model2<-train(lny~blood+prog+enz+gender+x7+x8, data=sur, method="leapBackward",
                     tuneGrid=data.frame(nvmax=6),
                     trControl=train.control)
step.model2$results

step.model3<-train(lny~blood+prog+enz+age+gender+x7+x8, data=sur, method="leapBackward",
                     tuneGrid=data.frame(nvmax=7),
                     trControl=train.control)
step.model3$results
```

Model1: $\ln(y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_8 X8 + \beta_7 X7$

Root MSE (RMSE)=0.218

Model2: $\ln(y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_6 gender + \beta_8 X8 + \beta_7 X7$

Root MSE (RMSE)=0.225

Model3: $\ln(y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_5 age + \beta_6 gender + \beta_8 X8$

Root MSE (RMSE)=0.222

Advanced Diagnostic Measurement in MLR

Why do we perform diagnostics?

Looking for,

- Outliers
- Evidence of a non-normal error distribution
- Evidence of non-independence in the errors
- Evidence of a disproportionate influence by one or more individual data points
- Evidence of multicollinearity

Previously we have used:

- Plots of residuals against predicted values for multiple symptoms
- Plots of residuals against independent variables for functional relationship
- Plots of residuals against time, collection order for independency checking
- Normal quantile plots of residuals for Normality
- Histograms of residuals for Normality, outliers
- Plots of independent variables against each other for multicollinearity

However, these plots have a limitation in that they do not reveal the marginal effect of a predictor variable when other variables are present.

To overcome this limitation, we require more advanced models and specialized tools.

- Added-variable plots to observe the marginal effect of a predictor variable
- Studentized deleted residuals, hat matrix diagonals to observe the outliers
- Cook's D, DFFITS, and DFBETAS to detect influential points
- Variance inflation factor (VIF) and tolerance to diagnose multicollinearity

These tools can provide more insights into the relationships between variables and improve the accuracy of our model.

Added-variable plot

Added-variable plots, also called *partial regression plots* or *adjusted variable plots*, are refined residual plots that provide graphic information about the marginal importance of a predictor variable X_k , given the other predictors are already in.

$$\text{Regress } Y \text{ on } X_2 \quad \hat{Y}_i(X_2) = b_0 + b_2 X_{i2}$$

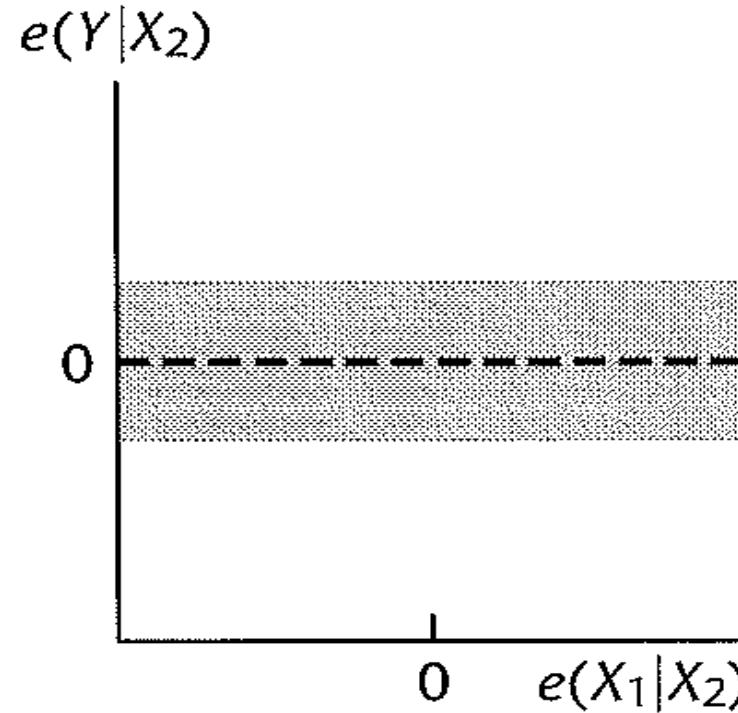
$$\text{Regress } X_1 \text{ on } X_2 \quad \hat{X}_{i1}(X_2) = b_0 + b_2 X_{i2}$$

$$e_i(\hat{Y}_i|X_2) = Y_i - \hat{Y}_i(X_2)$$

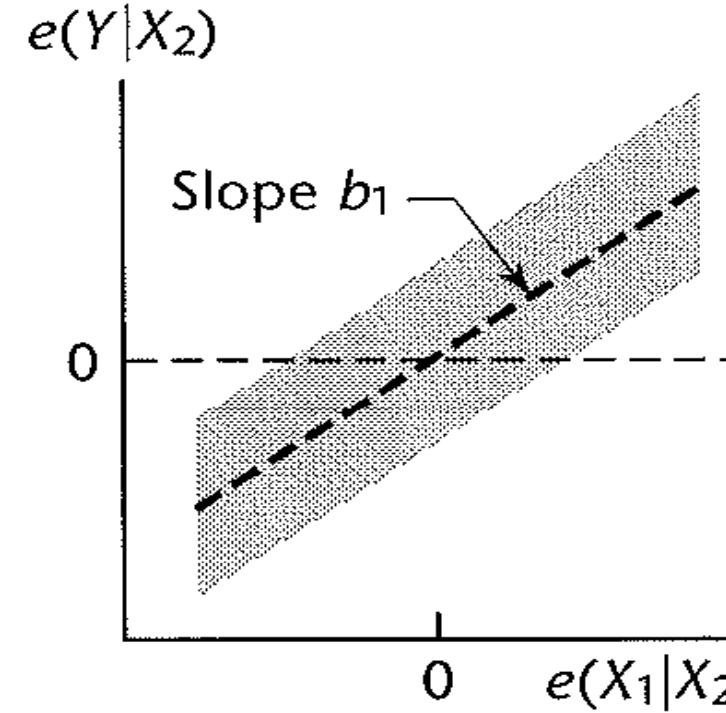
$$e_i(X_1|X_2) = X_{i1} - \hat{X}_{i1}(X_2)$$

- The added-variable plot for X_1 is a plot of the $e_i(\hat{Y}_i|X_2)$ against $e_i(X_1|X_2)$ which shows the relationship between the residual error of the response variable Y and the residual error of the explanatory variable X_1 , while holding all other explanatory variables (X_2, X_3, \dots) constant.
- The plot can help identify any patterns or trends in the relationship between the **marginal effect** of X_1 and Y , which may not be apparent from simple scatter plots or other graphical tools.
- It can also help evaluate the validity of the linear regression model assumptions, such as linearity, independence, and homoscedasticity.

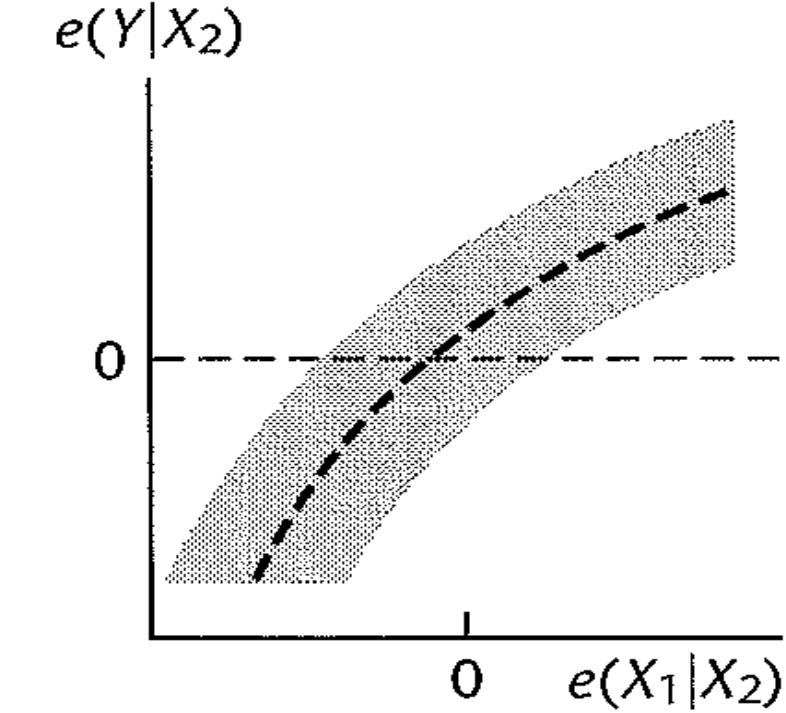
Added-variable plot prototypes



(a)



(b)



(c)

Plot (a) indicates that X_1 contain no additional information useful for predicting Y beyond that contained in X_2

Plot (b) indicates that a linear term in X_1 may be a helpful addition to the regression model already containing X_2 .

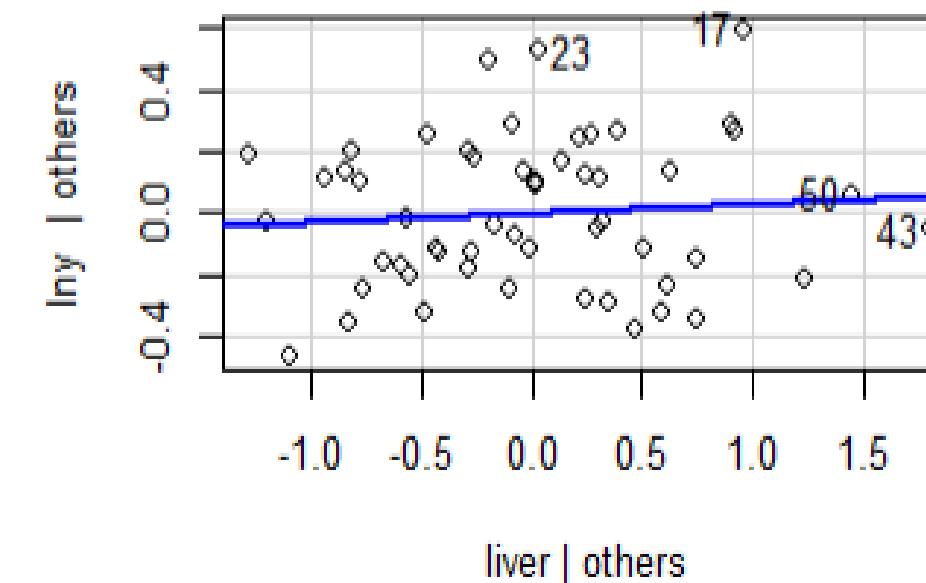
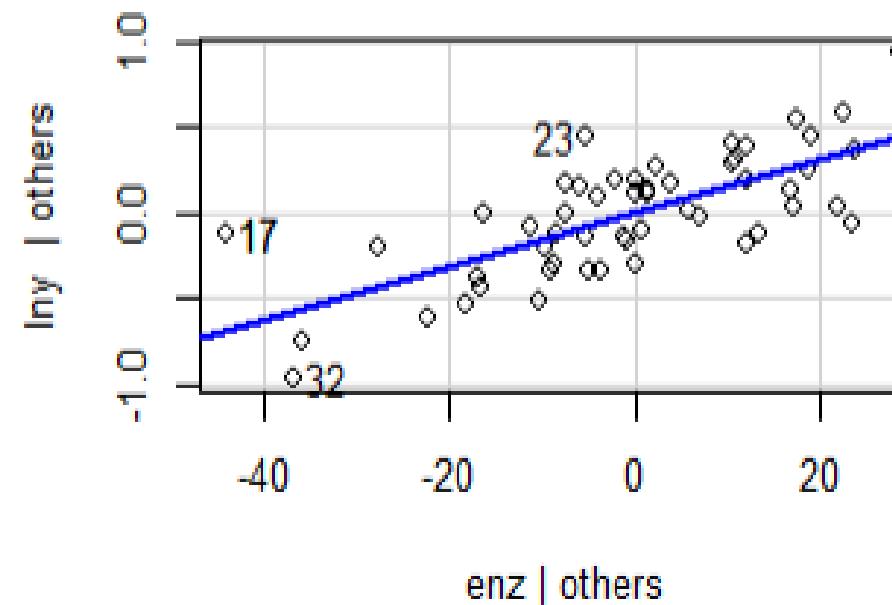
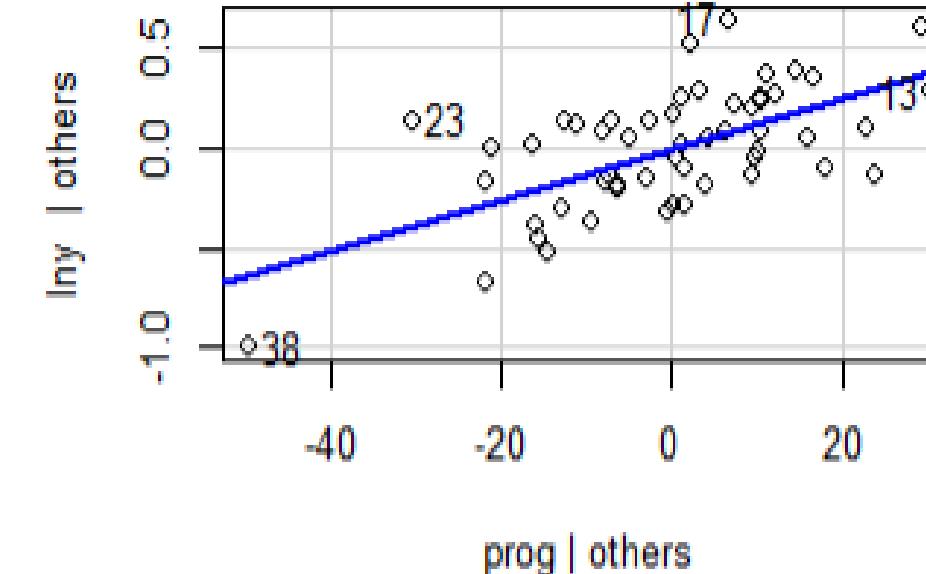
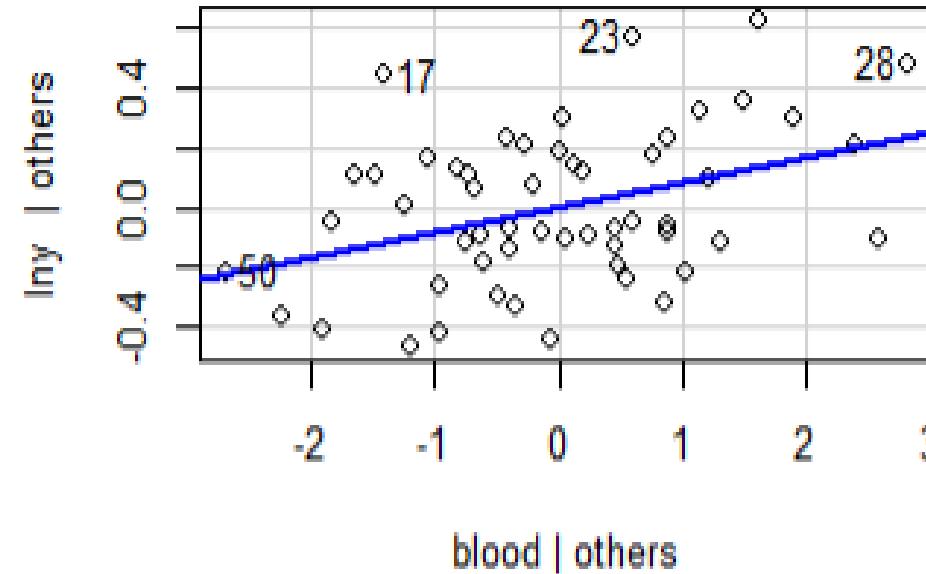
Plot (c) indicates that a nonlinear term in X_1 may be a helpful addition to the regression model already containing X_2 .

Added-variable plots (surgical example)

library(car)

```
avPlots(lm(y~blood+prog+enz+liver , data=sur))
```

Added-Variable Plots



- The added-on effect can be seen on Prog, Enz, Blood
- The first order seems sufficient
- Liver doesn't show added-on effect

Comments

- Added-variable plots need to be used with caution. They may not show the proper form of the marginal effect if the functional relations for some of the predictor variables are miss specified.

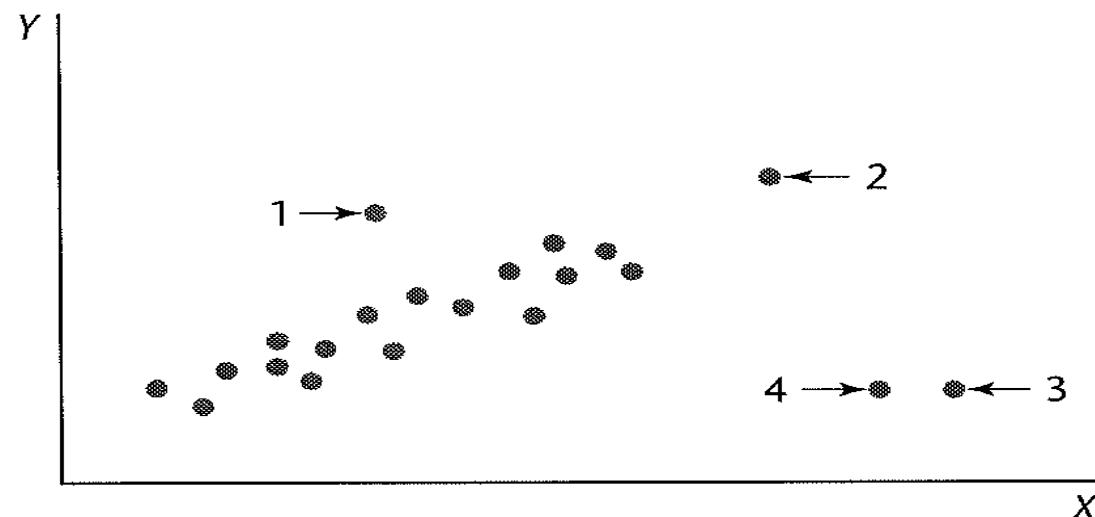
E.g., if X_1 and X_2 are related in a curvilinear fashion to Y but the regression model is linear term only $e_i(X_1|X_2) \neq X_{i1} - \hat{X}_{i1}(X_2)$

- High multicollinearity among the predictor variables may cause the added-variable plots to show an improper functional relation for the marginal effect of a predictor.

$$e_i(\hat{Y}_i|X_2) \neq Y_i - \hat{Y}_i(X_2)$$

Identifying outliers in MLR

- In SLR, we can identify outliers by means of boxplots, scatter plots, residual plots etc.
- In MLR, it is difficult to identify outliers by simple graphic means because it might not be extreme in a multiple regression model anymore.
- We now discuss the use of some refined measures for identifying outliers
 - Residuals and semi-studentized residuals (*outlying Y observation*)
 - Studentized deleted residuals (*outlying Y observation*)
 - Hat Matrix (*outlying X observation*)
 - Identifying influential cases (DFFITS, Cook's distance and DFBETAS measures)



A simple review on the Hat matrix and residuals

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

$$\widehat{\mathbf{Y}} = \mathbf{HY}$$

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

$$\sigma^2\{\mathbf{e}\} = \sigma^2 (\mathbf{I} - \mathbf{H})$$

- The hat matrix (\mathbf{H}) is a matrix that is used to compute the predicted values ($\widehat{\mathbf{y}}$) for the dependent variable based on the observed values of the independent variables.
- It can be interpreted as a matrix that projects the observed values of the dependent variable (\mathbf{y}) onto the predicted values ($\widehat{\mathbf{y}}$) based on the observed values of the independent variables (\mathbf{x})

The variance of residual e_i is:

$$\sigma^2\{e_i\} = \sigma^2 (1 - h_{ii}) \xrightarrow{\text{Estimated by}} s^2\{e_i\} = MSE (1 - h_{ii})$$

The covariance between e_i and e_j is:

$$\sigma^2\{e_i, e_j\} = \sigma^2 (0 - h_{ij}) = -h_{ij}\sigma^2 \xrightarrow{} s^2\{e_i, e_j\} = -h_{ij}(MSE)$$

Comment:

$$h_{ii} = \mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i \quad \text{Where } \mathbf{X}_i = \begin{pmatrix} 1 \\ X_{i,1} \\ X_{i,2} \\ \vdots \\ X_{i,p-1} \end{pmatrix}$$

i.e., $p = 3$

```
lm.influence(lm(y~x1+x2, bodyfat))$hat
```

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| 0.20101253 | 0.05889478 | 0.37193301 | 0.11094009 | 0.24801034 | 0.12861620 | 0.15551745 | 0.09628780 | 0.11463564 |
| 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| 0.11024435 | 0.12033655 | 0.10926629 | 0.17838181 | 0.14800684 | 0.33321201 | 0.09527739 | 0.10559466 | 0.19679280 |
| 19 | 20 | | | | | | | |
| 0.06695419 | 0.05008526 | | | | | | | |

Studentized residuals (r_i)

Studentized residuals: $r_i = \frac{e_i}{s\{e_i\}}$, where $s\{e_i\} = \sqrt{MSE(1 - h_{ii})}$

The body fat example

To compute r_3 for case 3

$$e_3 = -3.176 \quad h_{33} = 0.372$$

$$MSE = 6.4677$$

$$r_i = \frac{e_i}{s\{e_i\}} = -\frac{3.176}{\sqrt{MSE(1 - h_{ii})}} = -1.576$$

- This means that the residual for case 3 is 1.576 standard deviations smaller than what would be expected based on the overall distribution of residuals.
- Studentized residuals mainly measures the outliers in the Y scale.
- When the Normal assumption is met and a significant value of 0.05 is used, a case with $|r_i| > 2$ could be considered an outlier for a reasonably large sample.

Studentized deleted residual (t_i)

$$d_i = Y_i - \hat{Y}_{i(i)} = \frac{e_i}{1 - h_{ii}}$$

$$s^2\{d_i\} = \frac{MSE_{(i)}}{1 - h_{ii}}$$

$$t_i = \frac{d_i}{s\{d_i\}} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}} = e_i \left[\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2} \right]^{1/2}$$

- $\hat{Y}_{i(i)}$ and $MSE_{(i)}$ are computed from data set after the i th case is deleted in fitting the LM.
- The larger is the value h_{ii} , the larger will be the deleted residual as compared to the ordinary residual of e_i
- The studentized deleted residual mainly measure the outliers in the Y scale. It is also considered a balanced measurement for other outlying situation such as the impact on the estimation of regression coefficients.
- Like the studentized residual (r_i), a value of 2 is often used as a threshold to determine outliers with the studentized deleted residual (t_i) when the sample size is large. A more precise method is the Bonferroni procedure, especially for small sample size or when the Normality assumption is violated.

The Bonferroni Procedure to determine Y Outliers with the Studentized deleted residual ($\alpha = 0.1, n = 20, p = 3$)

H_0 : Case i is not outlying in Y-scale

$$t_i = \frac{d_i}{s\{d_i\}} \sim t(n - p - 1)$$

H_a : Case i is outlying in Y-scale

With Bonferroni procedure: $g = n$
 Bonferroni critical value = $t(1 - \frac{\alpha}{2n}; n - 1 - p)$
 $= t(0.9975; 16) = 3.252$

The body fat example

$$Y = -17.174 + 0.2224X_1 + 0.6594X_2 + \epsilon$$

To test if case 1 ($X_{11} = 19.5, X_{12} = 43.1$) is an outlier: $e_1 = -1.683$ $h_{11} = 0.201$ $SSE = 195.9508$

$$t_1 = e_i \left[\frac{n - p - 1}{SSE(1 - h_{ii}) - e_i^2} \right]^{1/2} = -1.683 \left[\frac{20 - 3 - 1}{195.9508(1 - 0.201) - (-1.683)^2} \right]^{1/2} = -0.73$$

Since $|t_1| = 0.73 < 3.252$, we conclude that case1 is not an outlier.

```
rstudent(lm(y~x1+x2, bodyfat))
```

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---------------|--------------|---------------|---------------|---------------|---------------|--------------|
| -0.7299854027 | 1.5342541325 | -1.6543295725 | -1.3484842072 | -0.0001269809 | -0.1475490938 | 0.2981276214 |
| 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 1.7600924916 | 1.1176487404 | -1.0337284208 | 0.1366610657 | 0.9231785040 | -1.8259027246 | 1.5247630510 |
| 15 | 16 | 17 | 18 | 19 | 20 | |
| 0.2671500921 | 0.2581323416 | -0.3445090997 | -0.3344080836 | -1.1761712768 | 0.4093564171 | |

Comment

In addition to outliers, large Studentized deleted residuals can be caused by a non-normal error distribution or non-constant variance.

If the data for a potential outlier does not have any obvious problems, consider transforming Y .

Identifying X Outliers with Hat Matrix leverage values

- The hat matrix projects Y to \hat{Y} based on the value of X: $\hat{Y} = HY$
- The i^{th} diagonal element, h_{ii} is called the leverage of the i^{th} x value has on the predicted values of Y.
- The larger is h_{ii} , the smaller is the variance of the residual e_i : $\sigma^2\{e_i\} = \sigma^2 (1 - h_{ii})$
- Observations with extreme values on one or more X values tend to have large h_{ii}

$$h_{ii} = X_i'(X'X)^{-1}X_i$$
- h_{ii} is considered *large if* it is more than twice as large as the mean leverage value $\bar{h} = p/n$,

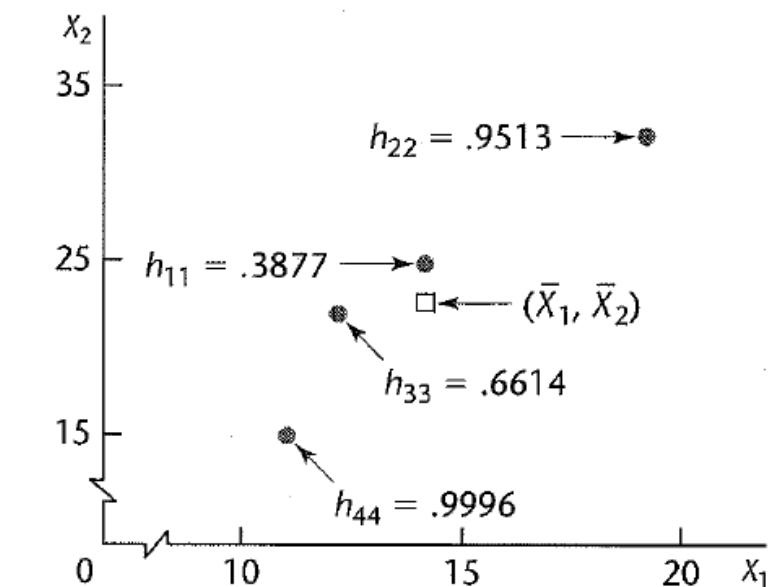
$$h_{ii} > 2p/n$$

In the body fat example, $\bar{h} = \frac{\sum h_{ii}}{n} = \frac{p}{n} = \frac{3}{20} = 0.15$

Hence, any case with a $h_{ii} > 2(0.15) = 0.3$ is considered outlying in term of their X values.

```
lm.influence(lm(y~x1+x2, bodyfat))$hat
```

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| 0.20101253 | 0.05889478 | 0.37193301 | 0.11094009 | 0.24801034 | 0.12861620 | 0.15551745 | 0.09628780 | 0.11463564 |
| 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| 0.11024435 | 0.12033655 | 0.10926629 | 0.17838181 | 0.14800684 | 0.33321201 | 0.09527739 | 0.10559466 | 0.19679280 |
| 19 | 20 | | | | | | | |
| 0.06695419 | 0.05008526 | | | | | | | |



Comments

- High leverage does not necessarily mean that an observation is an outlier in the y-scale (the space of the dependent variable).
- If the dataset has small n or large p, a lower cutoff value may be appropriate; if the dataset has a large n or small p, a higher cutoff value may be appropriate. (0.2-0.5)
- For observations with very high leverage, examine the pattern of leverage values across the independent variables to determine which independent variable(s) may be driving the high leverage values.
 - If multiple observations have high leverage values that are spread out across multiple independent variables, it may be an indication of collinearity in the dataset.

Identifying influential cases—DFFITS, Cook's Distance, and DFBETAS Measures

A case is **influential** if its exclusion causes major changes in the fitted regression function: either on the coefficients or the fitted values.

We take up three measures of influence that are widely used in practice, each based on the omission of a single case to measure its influence.

Influence on Single Fitted Value--DFFITS

$$(DFFITS)_i = \frac{(\hat{Y}_i - \hat{Y}_{i(i)})}{\sqrt{MSE_{(i)} h_{ii}}}$$

It represents the estimated standard deviation of the fitted value increases or decrease with the inclusion (\hat{Y}_i) or exclusion ($\hat{Y}_{i(i)}$) of the i th case. i.e., the influence on prediction

$$(DFFITS)_i = e_i \left[\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2} \right]^{\frac{1}{2}} \left(\frac{h_{ii}}{1-h_{ii}} \right)^{\frac{1}{2}} = t_i \left(\frac{h_{ii}}{1-h_{ii}} \right)^{\frac{1}{2}}$$

A guideline for identifying influential cases is that

$$|(DFFITS)_i| > 1 \text{ for small to medium data set, and } > 2\sqrt{\frac{p}{n}} \text{ for large data sets.}$$

In the body fat example, $t_3 = -1.656, h_{33} = 0.372$

$$(DFFITS)_i = t_i \left(\frac{h_{ii}}{1-h_{ii}} \right)^{\frac{1}{2}} = -1.656 \left(\frac{0.372}{1-0.372} \right)^{\frac{1}{2}} = -1.27$$

```
dffits(lm(y~x1+x2, bodyfat))
```

Since 1.27 is somewhat large than 1, but since it is not too far greater than 1, the case may not be influential enough to require remedial action

Influence on All Fitted Value—Cook's distance

$$D_i = \frac{\sum (\hat{Y}_i - \hat{Y}_{i(i)})^2}{pMSE} = \frac{e_i^2}{pMSE} \left(\frac{h_{ii}}{(1 - h_{ii})^2} \right)$$

- An aggregate measure represents the influence of the i th case on all n fitted values
- For interpreting Cook's distance measure, relate D_i to the $F(p, n - p)$ distribution, and ascertain the corresponding percentile value.
 - If the percentile < 20%, the i th case has no influence on the fitted values
 - If the percentile is between 20% and 50%, the i th case has minor influence on the fitted values
 - If the percentile > 50%, the i th case has major influence on the fitted values
- Large influence depending on e_i and h_{ii}

In the body fat example, $e_3 = -3.176, h_{33} = 0.372$

$$D_3 = \frac{e_3^2}{pMSE} \left(\frac{h_{33}}{(1 - h_{33})^2} \right) = \frac{(-3.176)^2}{3(6.47)} \left(\frac{0.372}{(1 - 0.372)^2} \right) = 0.49$$

| | | |
|------------------------------|-----------|---|
| <code>pf(0.49, 3, 17)</code> | 0.3061611 | • Case 3 with a $D=0.49$, is the 30.6% percentile of the distribution, hence has a minor influence on the fitted values. |
| <code>qf(0.2, 3, 17)</code> | 0.3352959 | • In general, case with $D_i < 0.33$ has no influence on the fitted values and, |
| <code>qf(0.5, 3, 17)</code> | 0.8212088 | a case with $D_i > 0.82$ has major influence on the fitted values. |

```
cooks.distance(lm(y~x1+x2, bodyfat))
```

Influence on the Regression Coefficients—DFBETAS

“S” means “standardized”.

$$(DFBETAS)_{k(i)} = \frac{(b_k - b_{k(i)})}{\sqrt{MSE_{(i)} c_{kk}}} , \quad k = 0, 1, \dots, p-1$$

- A measure of the influence of the i th case on all each regression coefficients.
- c_{kk} is the k th diagonal element of $(X'X)^{-1}$.
- $\sigma^2\{b\} = \sigma^2(X'X)^{-1}$, hence $\sigma^2\{b_k\} = \sigma^2 c_{kk}$ (estimated by $MSE_{(i)} c_{kk}$)
- Large value of $(DFBETAS)_{k(i)}$ is indicative of a large impact of the i th case on the k th regression coefficient. A large value means

$(DFBETAS)_{k(i)} > 1$ for small to medium data set and $> 2/\sqrt{n}$ for large data set.

```
dfbetas(lm(y~x1+x2, bodyfat))
```

| | (Intercept) | x1 | x2 |
|----|---------------|---|---------------|
| 1 | -3.051821e-01 | -1.314856e-01 | 2.320319e-01 |
| 2 | 1.725732e-01 | 1.150251e-01 | -1.426129e-01 |
| 3 | -8.471013e-01 | -1.182525e+00 | 1.066903e+00 |
| 4 | -1.016120e-01 | -2.935195e-01 | 1.960719e-01 |
| 5 | -6.372122e-05 | -3.052747e-05 | 5.023715e-05 |
| 6 | 3.967715e-02 | 4.008114e-02 | -4.426759e-02 |
| 7 | -7.752748e-02 | -1.561293e-02 | 5.431634e-02 |
| 8 | 2.614312e-01 | 3.911262e-01 | -3.324533e-01 |
| 9 | -1.513521e-01 | -2.946556e-01 | 2.469091e-01 |
| 10 | 2.377492e-01 | 2.446010e-01 | -2.688086e-01 |
| 11 | -9.020885e-03 | 1.705640e-02 | -2.484518e-03 |
| 12 | -1.304933e-01 | 2.245800e-02 | 6.999608e-02 |
| 13 | 1.194147e-01 | 5.924202e-01 | -3.894913e-01 |
| 14 | 4.517437e-01 | 1.131722e-01 | -2.977042e-01 |
| 15 | -3.004276e-03 | -1.247567e-01 | 6.876929e-02 |
| 16 | 9.308463e-03 | 4.311347e-02 | -2.512499e-02 |
| 17 | 7.951208e-02 | 5.504357e-02 | -7.609008e-02 |
| 18 | 1.320522e-01 | 7.532874e-02 | -1.161003e-01 |
| 19 | -1.296032e-01 | -4.072030e-03 | 6.442931e-02 |
| 20 | 1.019045e-02 | 2.290797e-03 | -3.314146e-03 |

Diagnostic with the influencePlot() Output

```
library(car)
influencePlot(lm(y~x1+x2, bodyfat))
```

| | StudRes
<dbl> | Hat
<dbl> | CookD
<dbl> |
|----|------------------|--------------|----------------|
| 3 | -1.6543296 | 0.3719330 | 0.49015668 |
| 8 | 1.7600925 | 0.0962878 | 0.09793853 |
| 13 | -1.8259027 | 0.1783818 | 0.21215024 |
| 15 | 0.2671501 | 0.3332120 | 0.01257530 |

List all possible cases that are identified as outlying regarding its Y value.

None, since the studentized deleted residuals do not exceed the Bonferroni critical value.

$$\text{With the Bonferroni procedure, a case is possibly an outlier if } |t_i| > t\left(1 - \frac{\alpha}{2n}; n - 1 - p\right) = t(1 - 0.05/2(20); 20 - 1 - 3) = t(0.99875; 16) = 3.58$$

List all possible cases that are identified as outlying with regard to its X value.

Case 3 and 15

A case is considered an outlier if $h_{ii} > 2\bar{h} = 2p/n=2(3)/20=0.3$

Diagnostic with the influencePlot() Output

```
library(car)
influencePlot(lm(y~x1+x2, bodyfat))
```

| | StudRes
<dbl> | Hat
<dbl> | CookD
<dbl> |
|----|------------------|--------------|----------------|
| 3 | -1.6543296 | 0.3719330 | 0.49015668 |
| 8 | 1.7600925 | 0.0962878 | 0.09793853 |
| 13 | -1.8259027 | 0.1783818 | 0.21215024 |
| 15 | 0.2671501 | 0.3332120 | 0.01257530 |

A case has major influence if $D > F(0.5; p, n - p) = F(0.5; 3, 20 - 3) = 0.821$;

Moderate influence if D is less than 0.821 but greater than $F(0.2; p, n - p) = F(0.2; 3, 20 - 3) = 0.335$

Are there any potential influential points? Case 3 has moderate influence.

Diagnostic with the influencePlot() Output

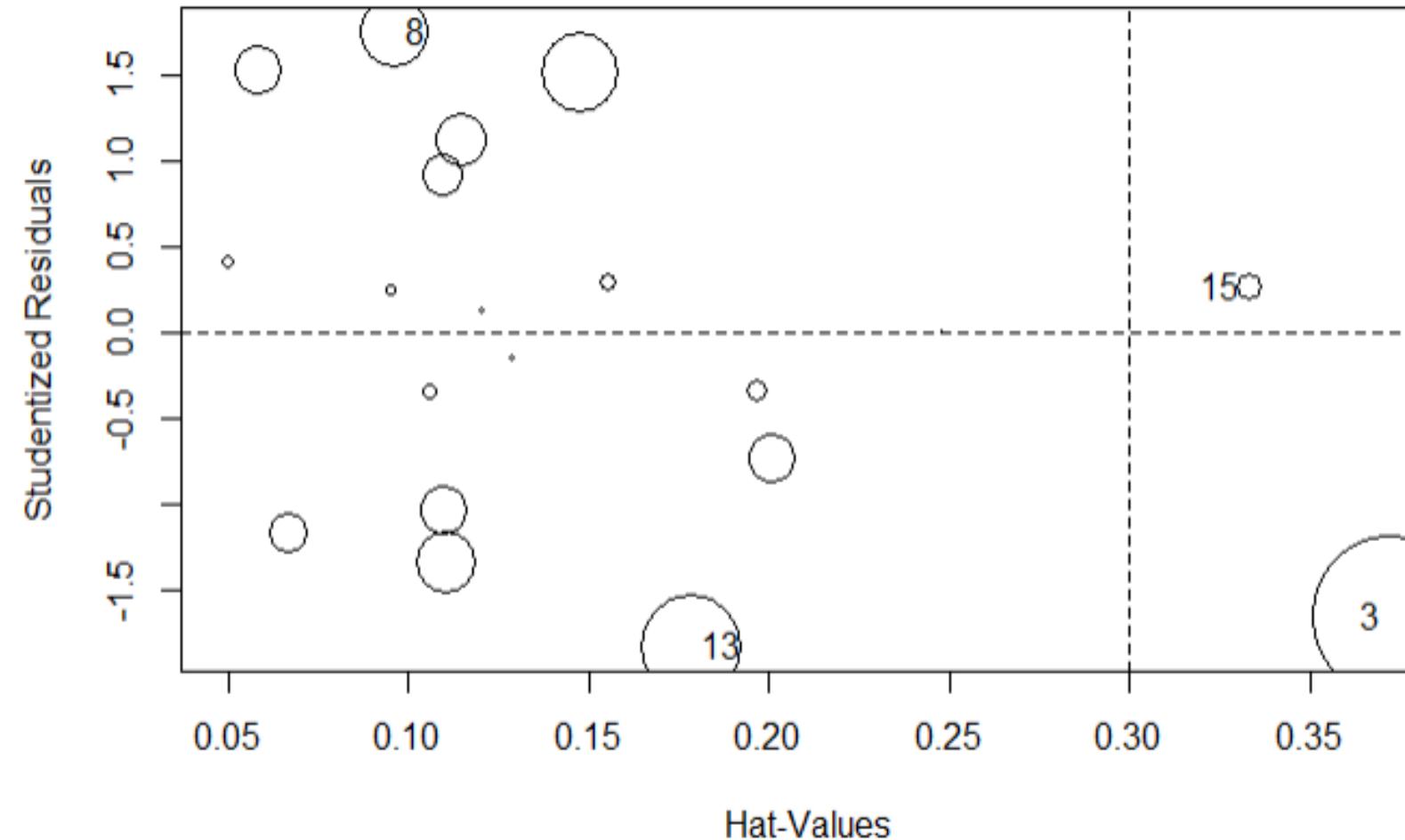
| | (Intercept) | x1 | x2 |
|----|---------------|---------------|---------------|
| 1 | -3.051821e-01 | -1.314856e-01 | 2.320319e-01 |
| 2 | 1.725732e-01 | 1.150251e-01 | -1.426129e-01 |
| 3 | -8.471013e-01 | -1.182525e+00 | 1.066903e+00 |
| 4 | -1.016120e-01 | -2.935195e-01 | 1.960719e-01 |
| 5 | -6.372122e-05 | -3.052747e-05 | 5.023715e-05 |
| 6 | 3.967715e-02 | 4.008114e-02 | -4.426759e-02 |
| 7 | -7.752748e-02 | -1.561293e-02 | 5.431634e-02 |
| 8 | 2.614312e-01 | 3.911262e-01 | -3.324533e-01 |
| 9 | -1.513521e-01 | -2.946556e-01 | 2.469091e-01 |
| 10 | 2.377492e-01 | 2.446010e-01 | -2.688086e-01 |
| 11 | -9.020885e-03 | 1.705640e-02 | -2.484518e-03 |
| 12 | -1.304933e-01 | 2.245800e-02 | 6.999608e-02 |
| 13 | 1.194147e-01 | 5.924202e-01 | -3.894913e-01 |
| 14 | 4.517437e-01 | 1.131722e-01 | -2.977042e-01 |
| 15 | -3.004276e-03 | -1.247567e-01 | 6.876929e-02 |
| 16 | 9.308463e-03 | 4.311347e-02 | -2.512499e-02 |
| 17 | 7.951208e-02 | 5.504357e-02 | -7.609008e-02 |
| 18 | 1.320522e-01 | 7.532874e-02 | -1.161003e-01 |
| 19 | -1.296032e-01 | -4.072030e-03 | 6.442931e-02 |
| 20 | 1.019045e-02 | 2.290797e-03 | -3.314146e-03 |

A case i is considered has a large impact on the regression coefficient if the $|DEBETAS| > 1$, or $> \frac{2}{\sqrt{n}}$ ($=0.45$) for large data set.

Any case has significant impact? Case 3 since $|DEBETAS|$ for beta1 and beta2 are bigger than 1 but not by much.

Diagnostic with the influencePlot() Output

```
library(car)
influencePlot(lm(y~x1+x2, bodyfat))
```



- The areas of the circles are proportional to DFFIT and Cooks distance.
- Vertical reference are drawn at twice and three times the average hat value
- Horizontal reference lines at -2, 0 and 2 on the Studentized-residual scale (for reference).

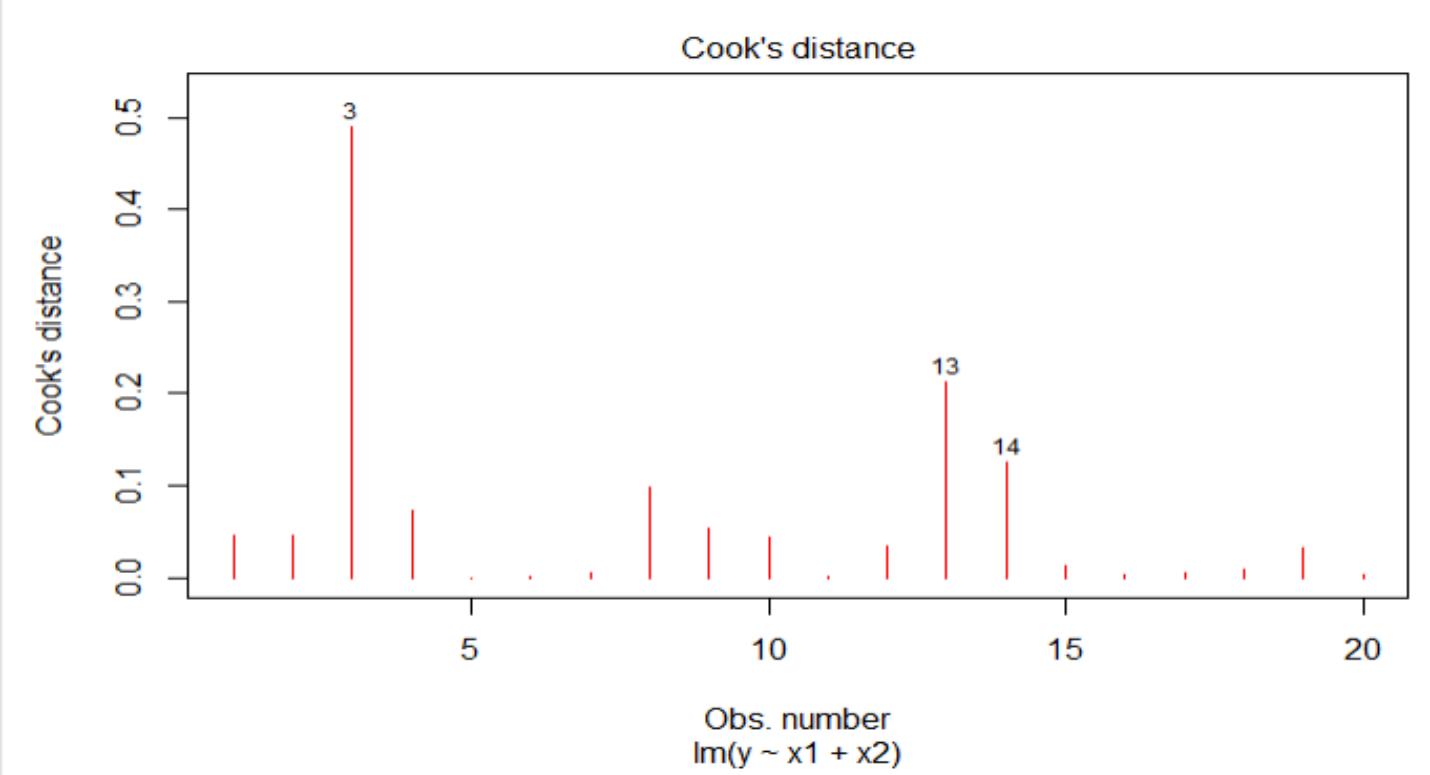
Outlying Y observations deviate in the Y axis (Studentized residuals)

Outlying X observations deviate in the X axis (Hat values)

The areas of the circles imply potential influential point, the bigger it is, the more likely that the point is influential.

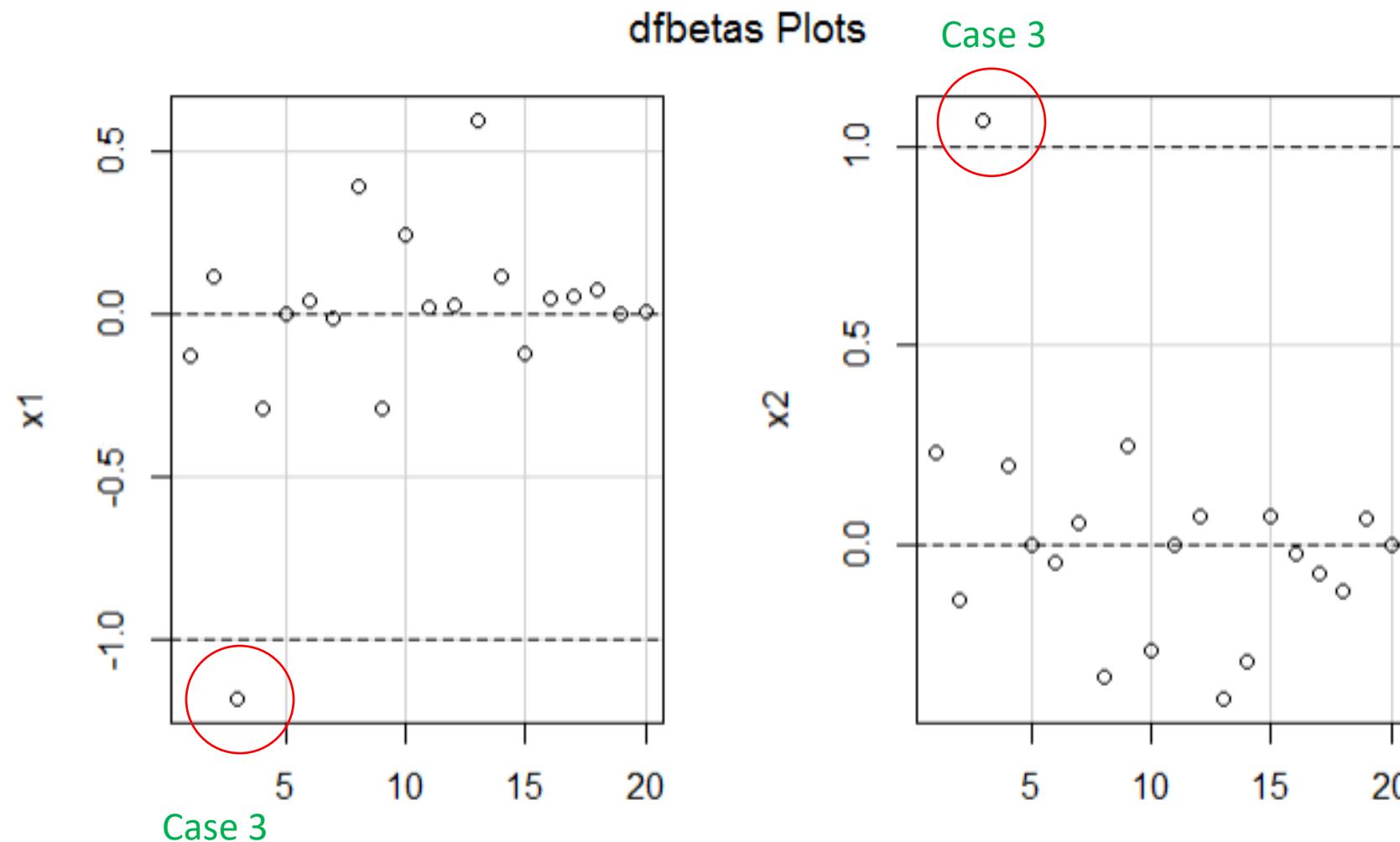
Diagnostic with Cook's distance Plot

```
plot(lm(y~x1+x2, bodyfat), pch=18, col="red", which=c(4))
```



Diagnostic with dfbetasPlots() Plot

```
dfbetasPlots(lm(y~x1+x2, bodyfat))
```



- Horizontal reference lines at 0, and +/- 1

```
dfbetas(lm(y~x1+x2, bodyfat))
```

| | (Intercept) | x_1 | x_2 |
|----|---------------|---------------|---------------|
| 1 | -3.051821e-01 | -1.314856e-01 | 2.320319e-01 |
| 2 | 1.725732e-01 | 1.150251e-01 | -1.426129e-01 |
| 3 | -8.471013e-01 | -1.182525e+00 | 1.066903e+00 |
| 4 | -1.016120e-01 | -2.935195e-01 | 1.960719e-01 |
| 5 | -6.372122e-05 | -3.052747e-05 | 5.023715e-05 |
| 6 | 3.967715e-02 | 4.008114e-02 | -4.426759e-02 |
| 7 | -7.752748e-02 | -1.561293e-02 | 5.431634e-02 |
| 8 | 2.614312e-01 | 3.911262e-01 | -3.324533e-01 |
| 9 | -1.513521e-01 | -2.946556e-01 | 2.469091e-01 |
| 10 | 2.377492e-01 | 2.446010e-01 | -2.688086e-01 |
| 11 | -9.020885e-03 | 1.705640e-02 | -2.484518e-03 |
| 12 | -1.304933e-01 | 2.245800e-02 | 6.999608e-02 |
| 13 | 1.194147e-01 | 5.924202e-01 | -3.894913e-01 |
| 14 | 4.517437e-01 | 1.131722e-01 | -2.977042e-01 |
| 15 | -3.004276e-03 | -1.247567e-01 | 6.876929e-02 |
| 16 | 9.308463e-03 | 4.311347e-02 | -2.512499e-02 |
| 17 | 7.951208e-02 | 5.504357e-02 | -7.609008e-02 |
| 18 | 1.320522e-01 | 7.532874e-02 | -1.161003e-01 |
| 19 | -1.296032e-01 | -4.072030e-03 | 6.442931e-02 |
| 20 | 1.019045e-02 | 2.290797e-03 | -3.314146e-03 |

Measures of Multicollinearity

We can already diagnose multicollinearity by observing:

- A significant global F -test alongside non-significant t -tests for all individual β s
- Parameter estimates that change greatly when predictors are added to the model or removed
- Parameter estimates that “don’t make sense” and the standard errors become large.
- Large differences between Type I and Type II extra sums of squares

Variance Inflation Factor (Tolerance)

The VIF measures the extent to which the variance of the estimated regression coefficient of a predictor variable is increased due to multicollinearity.

$$\Sigma\{b\} = \sigma^2(X'X)^{-1}, \text{ and } Var(b_i) = \sigma^2(X'X)^{-1}[i,i] \text{ in MLR, and reduced to } \sigma^2/SS_X \text{ in SLR.}$$

Furthermore, $Var(b_i) = \frac{\sigma^2}{SS_{X_i}(1-R_i^2)} = \frac{\sigma^2}{SS_X} \times \frac{1}{1-R_i^2}$, where R_i^2 is the correlation determination of X_i and other predictors.

Example, regress X_1 on X_2 and X_3 , then $R_{1|23}^2$ is the coefficient determination for $X_1 \sim X_2 + X_3$

- If *none* multicollinearity between X_1 and (X_2, X_3) , $R_{1|23}^2 = 0$, then $Var(b_1) = \frac{\sigma^2}{SS_{X_1}(1-R_1^2)} = \frac{\sigma^2}{SS_X}$.
- If *all* multicollinearity between X_1 and (X_2, X_3) , $R_{1|23}^2 = 1$, then $Var(b_1) = \frac{\sigma^2}{SS_{X_1}(1-R_1^2)}$ becomes indeterminate.

Define the Variance Inflation Factor, VIF_i as $\frac{1}{1-R_i^2} = \frac{1}{Tolerance}$

As a rule of thumb, a $VIF \geq 10$, or $R_k^2 > 0.9$ indicate excessive multilinearity.

The Body-fat example

$$VIF_{1|23} = 708.84$$

$$VIF_{2|13} = 564.34$$

$$VIF_{3|12} = 104.61$$

$$VIF_{3|1} = 1.265$$

$$VIF_{3|2} = 1.007$$

```
library(fmsb)
vif(lm(x1~x2+x3, data=bodyfat))
vif(lm(x2~x1+x3, data=bodyfat))
vif(lm(x3~x1+x2, data=bodyfat))

vif(lm(x3~x1, data=bodyfat))
vif(lm(x3~x2, data=bodyfat))
```

- The result proves that multicollinearity issue exists between X_3 and (X_1 and X_2) combined.
- It is a common practice to compute the VIF of a predictor with all other predictors considered in the full model.

```
library(car)
vif<-vif(lm(y~x1+x2+x3, data=bodyfat))
vif
```

| | x1 | x2 | x3 |
|--|----------|----------|----------|
| | 708.8429 | 564.3434 | 104.6060 |

Advanced Remedial Measures

Previously, we have addressed violations of the model assumptions by

- Transforming Y (e.g., Box-Cox, natural log)
- Transforming one or more X variables
- Adding terms to the model (polynomials, additional predictors, interactions)
- Variable selection (to minimize multicollinearity)

Sometimes these tools fail...

Advanced remedial measures

- Weighted least squares (WLS)
- Ridge regression
- Robust regression
- Nonparametric methods: Bootstrapping

Unequal Error variances Remedial Measures--Weighted Least Squares (WLS)

WLS

Our model assumes that the errors are *iid* with constant variance, σ^2

$$\varepsilon \sim N(0, \sigma^2 I)$$

But what if each subpopulation (i.e., each unique combination of X values) has its own, potentially unique error variance instead?

$$\varepsilon \sim N(0, \sigma^2 D)$$

Where the diagonal matrix D reflects that the variance could be non-consistent.

$$\sigma^2(\varepsilon) = \sigma^2 D = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

Define a diagonal weight matrix W , such that $w_i = 1/\sigma_i^2$

$$W = \begin{bmatrix} 1/\sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma_n^2 \end{bmatrix} \quad \sigma^2(\varepsilon) = \sigma^2 D = W^{-1} \quad W^{1/2} = \begin{bmatrix} 1/\sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma_n \end{bmatrix}$$

The weighted matrix W can be used to create a (weighted) data with constant variance

Multiple $W^{1/2}$ to $Y = X\beta + \varepsilon$, we obtain $W^{1/2}Y = W^{1/2}X\beta + W^{1/2}\varepsilon$

This becomes $Y_w = X_w\beta + \varepsilon_w$ where

$$Y_w = W^{1/2}Y$$

$$X_w = W^{1/2}X$$

$$\varepsilon_w = W^{1/2}\varepsilon$$

$$E(\varepsilon_w) = E(W^{1/2}\varepsilon) = W^{1/2}E(\varepsilon) = \mathbf{0}$$

$$\sigma^2(\varepsilon_w) = \sigma^2(W^{1/2}\varepsilon) = W^{1/2}\sigma^2(\varepsilon)W^{1/2} = W^{1/2}W^{-1}W^{1/2} = I$$

$$b_w = (X'_w X_w)^{-1} X'_w Y_w = (X' W X)^{-1} X' W Y$$

$$s^2\{b_w\} = \text{MSE}_w (X' W X)^{-1} = \frac{\sum w_i(Y - \hat{Y})^2}{n-p} (X' W X)^{-1}$$

If all weights are equal, w_i is identically equal to a constant, and WLS reduces to OLS.

Advantage

Valid inference in presence of non-constant variance (heteroscedasticity).

Disadvantage

$$\sigma^2(\varepsilon_w) = \sigma^2(W^{1/2}\varepsilon) = W^{1/2}\sigma^2(\varepsilon)W^{1/2} = W^{1/2}W^{-1}W^{1/2} = I$$

MSE_w is close to 1 in a good WLS model. Therefore, MSE_w could be used in model diagnosis but it has no clear contextual interpretation and cannot be used to compare models.

Next, we need to estimate the variance matrix D , or W^{-1}

Method 1: use replicated observations at each X_i to estimate each σ_i^2 , which may require new data.

Method 2: regress the residual $|e|$ with a MLR function on X variables.

$$|e| = U_0 + U_1X_1 + \dots + U_{p-1}X_{p-1}$$

since $D = \frac{1}{W} = \sigma^2(\varepsilon) = E(\varepsilon^2) - [E(\varepsilon)]^2 = E(\varepsilon^2)$,

and $E(\varepsilon^2)$ is estimated by $|e|^2 = (\hat{U}_0 + \hat{U}_1X_1 + \dots + \hat{U}_{p-1}X_{p-1})^2$

Then, W is estimated by $1/|e|^2$

The process could take several interactions with a weights added in the MLR model until the estimates become stable. This process is also known as the Interactively Reweighted Least Squares (IRLS)

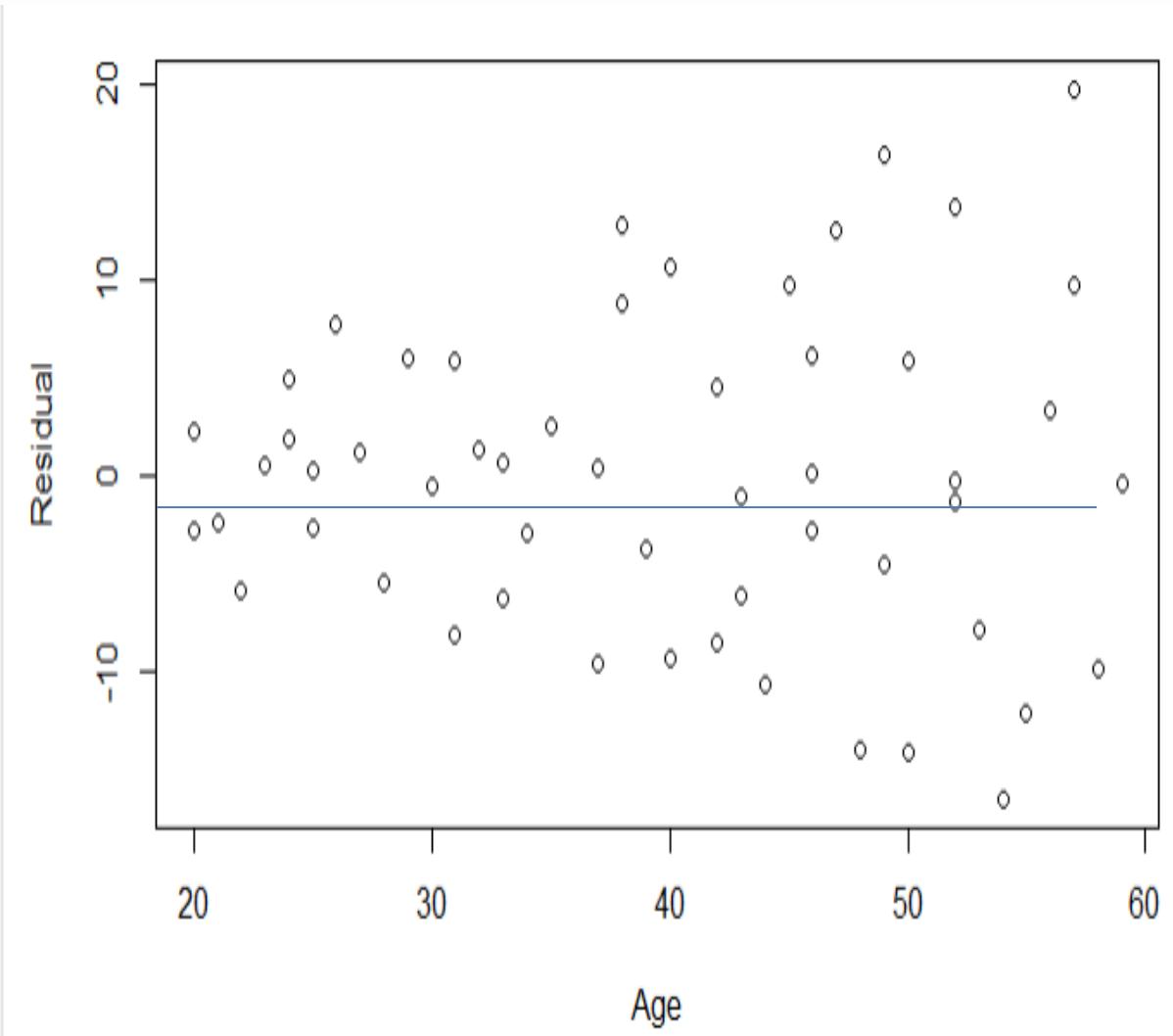
Although this method does not require new data, it does assume that residuals can be predicted with a MLR function on the predictors.

Example: Modeling blood pressure as a function of age

A health researcher who is interested in studying the relationship between diastolic blood pressure and age among healthy adult women 2- to 60 years old, collected data on 54 subjects.

- Y is diastolic blood pressure
- X is age in years
- $n = 54$ healthy adult women aged 20 to 60 years old

Diagnostic plots detecting unequal error variance



The algorithm run down

```
pres.mod<-lm(bp~age, pres)
wts1<-1/fitted(lm(abs(residuals(pres.mod))~age, pres))^2
pres.mod2<-lm(bp~age, weight=wts1, data=pres)
```

1. Fit $Y \sim X\beta + \varepsilon$ by unweighted LS
2. Save the residuals e_i
3. Fit the model $|e_i| \sim U_i'X + \phi_i$
4. Use the fitted values from Step 3 to calculate weights

$$W_i = \frac{1}{\widehat{\sigma}_i^2}$$

5. Use the estimated weights to fit $Y = X\beta + \varepsilon$ by WLS
6. (If necessary) Repeat Steps 2–5 until the values of β stabilize (typically, 1–3 iterations).

Unweighted linear model (OLS)

vs. Weighted linear model (WLS)

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|-------------|----------|------------|---------|----------|-----|
| (Intercept) | 56.15693 | 3.99367 | 14.061 | < 2e-16 | *** |
| age | 0.58003 | 0.09695 | 5.983 | 2.05e-07 | *** |

Residual standard error: 8.146 on 52 degrees of freedom

Multiple R-squared: 0.4077, Adjusted R-squared: 0.3963

F-statistic: 35.79 on 1 and 52 DF, p-value: 2.05e-07

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|-------------|----------|------------|---------|----------|-----|
| (Intercept) | 55.56577 | 2.52092 | 22.042 | < 2e-16 | *** |
| age | 0.59634 | 0.07924 | 7.526 | 7.19e-10 | *** |

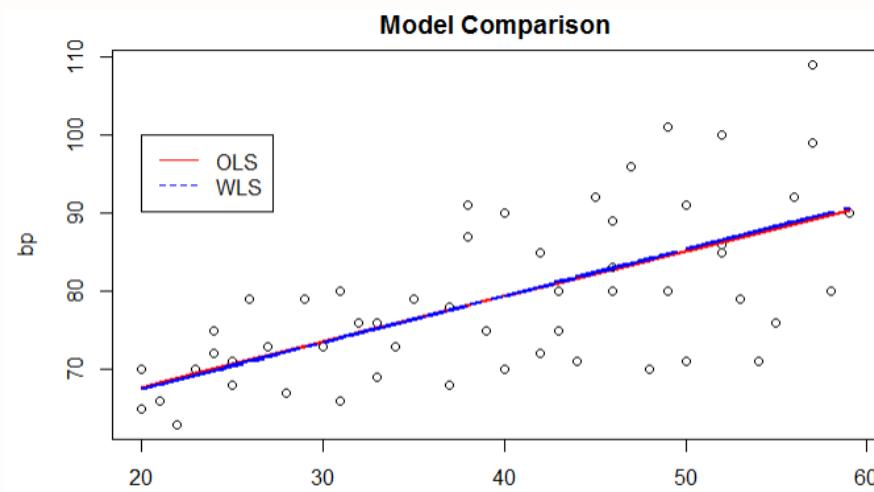
Residual standard error: 1.213 on 52 degrees of freedom

Multiple R-squared: 0.5214, Adjusted R-squared: 0.5122

F-statistic: 56.64 on 1 and 52 DF, p-value: 7.187e-10

- Comparing to the OLS, in the WLS,
 - the standard error of the coefficient is smaller, and
 - the Multiple R-square and the F-statistic are larger probably because the heteroscedasticity in the errors is accounted for by the chosen weighting scheme.
- We cannot compare the residual standard error (1.213 in WLS vs 8.146 in OLS) because the residuals have been altered and not comparable.

OLS and WLS Comparison via the Standardized Residual Plot



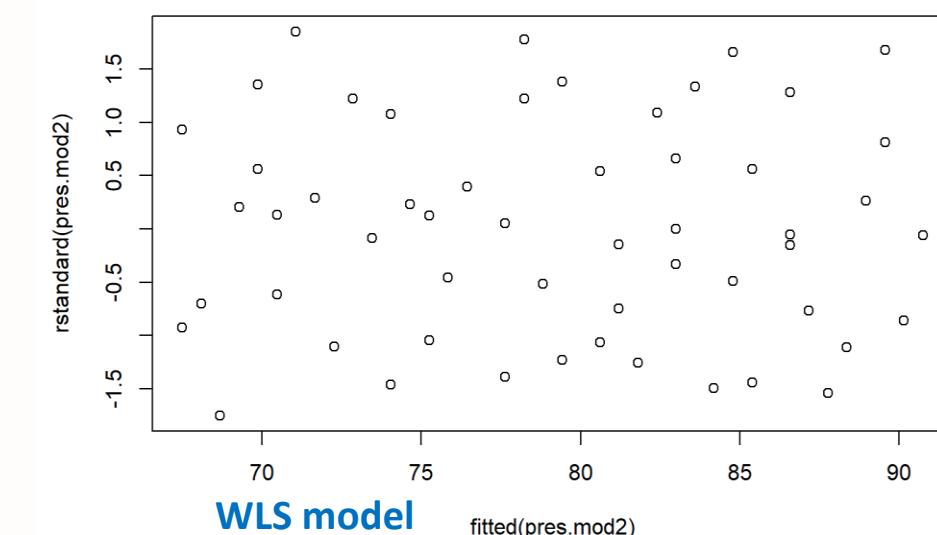
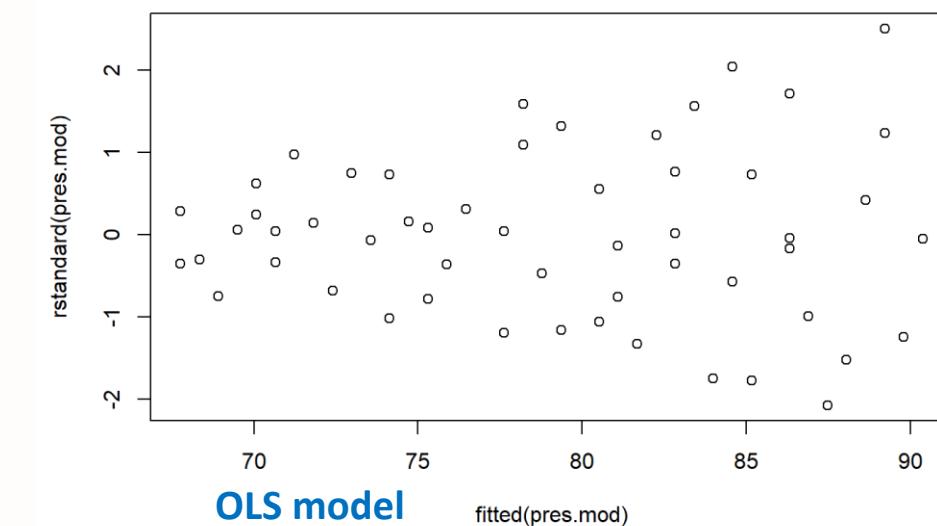
- In the scatter plot, the OLS and WLS do not show much difference.

```
plot(fitted(pres.mod), rstandard(pres.mod))
plot(fitted(pres.mod2), rstandard(pres.mod2))
```

- Rstandard residuals are standardized residuals, which means that they are scaled by the estimated standard deviation of the residuals.

This makes them more appropriate for assessing homogeneity of variance, as they consider the potential differences in variances at different levels of the predictor variables.

- From the plot of the (standardized residual, fitted value), we can see that the Rstandard residuals are constant across the range of the fitted values in the WLS.
- In WLS, the weights used to estimate the regression coefficients also affect the estimated standard deviation of the error term, which can lead to a change in the magnitude of the rstandard residual.



OLS and WLS Comparison via the Studentized Breusch-Pagan test

```
library(lmtest)
bptest(pres.mod)
bptest(pres.mod2)
```

studentized Breusch-Pagan test

```
data: pres.mod
BP = 12.541, df = 1, p-value = 0.0003981
```

studentized Breusch-Pagan test

```
data: pres.mod2
BP = 0.43608, df = 1, p-value = 0.509
```

- The null hypothesis of the BP test is homoscedasticity, so, a significant p-value in the OLS model would suggest that the data exhibit heteroscedasticity. But WLS model doesn't have the issue.

Confidence inference for coefficient in the weighted linear model (WLS)

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|-------------|----------|------------|---------|----------|-----|
| (Intercept) | 55.56577 | 2.52092 | 22.042 | < 2e-16 | *** |
| age | 0.59634 | 0.07924 | 7.526 | 7.19e-10 | *** |

Residual standard error: 1.213 on 52 degrees of freedom

Multiple R-squared: 0.5214, Adjusted R-squared: 0.5122

F-statistic: 56.64 on 1 and 52 DF, p-value: 7.187e-10

$$b_{w1} \pm t(0.975; 52)SE\{b_{w1}\} = 0.59634 \pm 2.007(0.07924) = (0.437, 0.755)$$

- (IMPORTANT) The T and F method here still based on the assumption that the random error follows Normal with constant variance! We could consider **Bootstrapping** for a more precise evaluation.

Multicollinearity Remedial Measures--Ridge regression

Multicollinearity and Ridge Regression

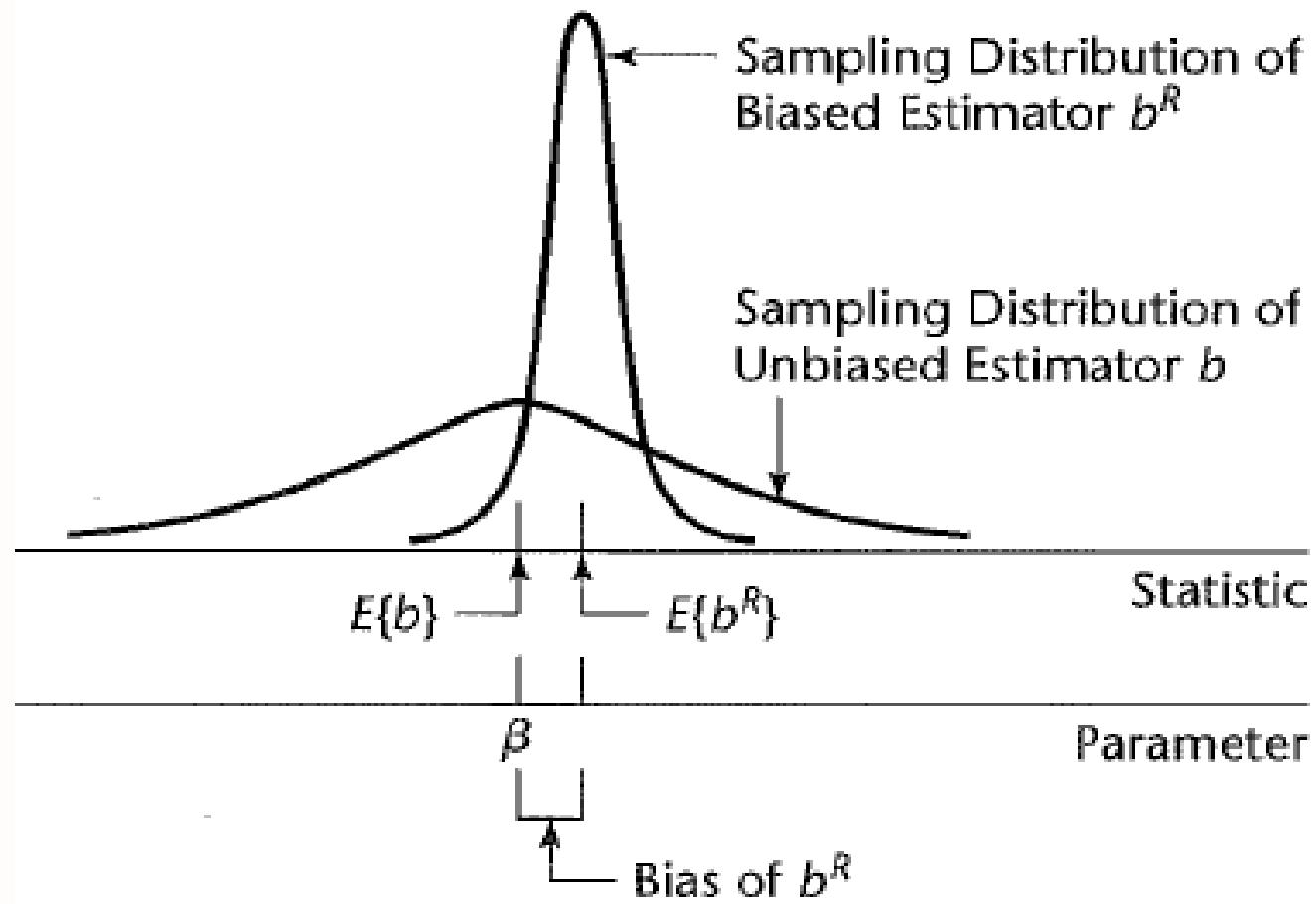
Previous approaches for dealing with serious multicollinearity include,

- Keeping the collinear predictors and restricting prediction to similarly collinear cases
- Centering of predictors in polynomial regression
- Model selection (drop some of the predictors)

Some additional possibilities include:

- Add new data points that break the pattern of collinearity (this can be difficult)
- Use of supplementary data that come from other contexts
- Principle components analysis (PCA) to create one or more composite variables that combine the collinear predictors
- Biased (a.k.a. “shrinkage”) estimation methods such as *ridge regression*

Ridge Regression



$$E\{b^R\} \neq \beta$$
$$E\{b\} = \beta$$

But $s\{b^R\} < s\{b\}$

Then on average, estimates based on the biased estimator, b^R , Will be closer to the true parameter β , than those based on The unbiased estimator, b .

Two equivalent formulations of ridge regression

Ridge regression shrinks estimators by adding a size penalty: $\lambda \sum_{j=1}^p \beta_j^2$

Penalized Residual Sum of Squares:

$$b^R = \arg \beta \min \left\{ \sum_{i=1}^n \left(Y_i - \left(\beta_0 + \sum_{j=1}^p X_{ij} \beta_j \right) \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}$$

Or in matrix form: $b^R = (X'X + \lambda I)^{-1} X'Y$

- λ controls the amount of bias (shrinkage) of the parameter estimates.
- Large $\lambda \rightarrow$ greater shrinkage (toward zero), the less variable of the coefficients.
- A commonly used method to determine λ is the *ridge trace, which simultaneously trace the b^R with different λ .*
- The value of VIF also tend to reduce as λ (*also denoted by k or c*) is increased.
- You will choose the smallest value when the regression confidents become stable in the ridge trace and the VIFs become sufficiently small.

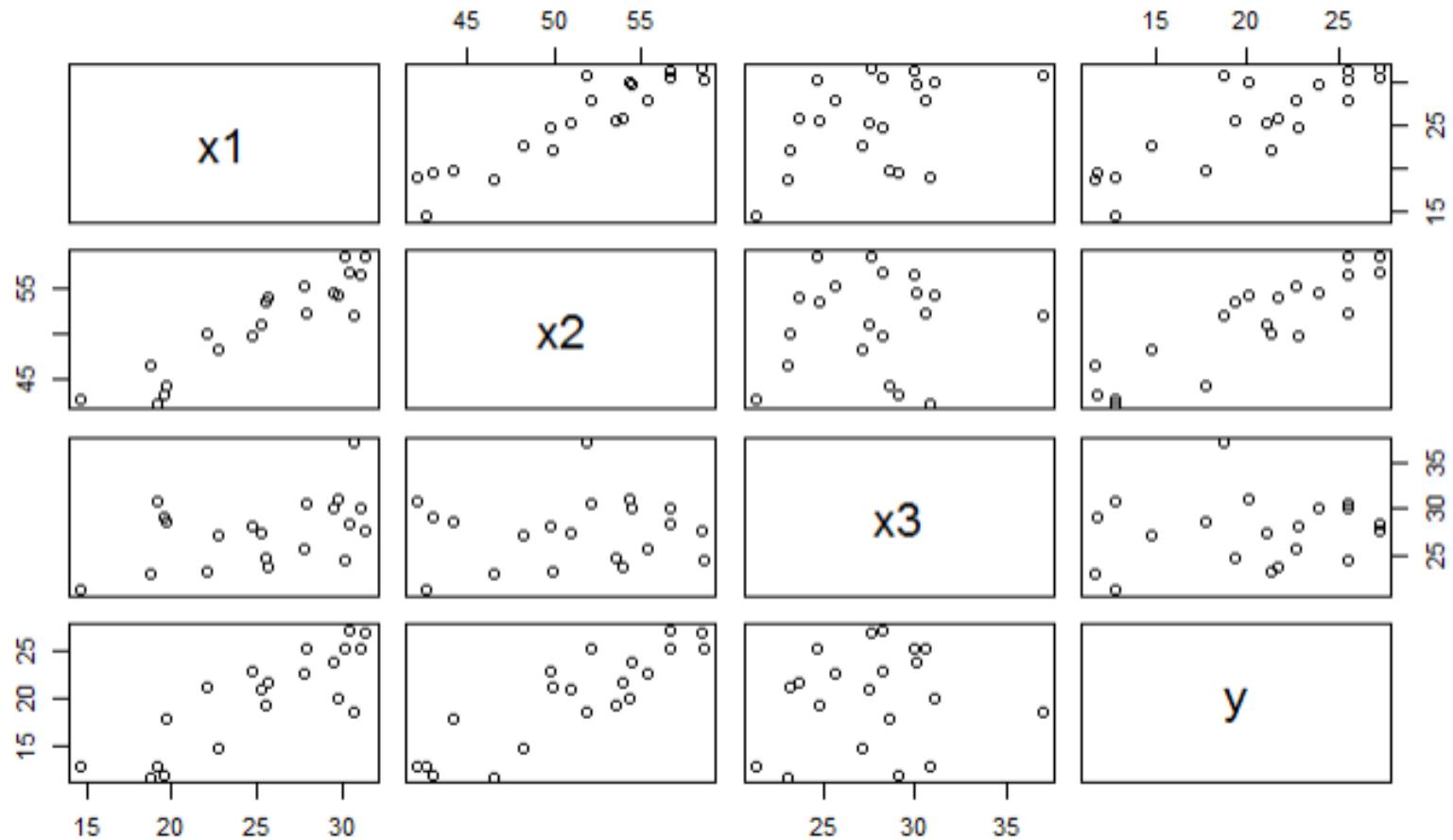
Choosing a value of λ

This is a judgment call. Select the smallest value of λ for which:

- Variance inflation factors are close enough to 1.
- Estimated coefficients are stable (trace lines approximately horizontal).
- Either R^2 or $\hat{\sigma}$ ($RM\ SE$) are changing slowly.
- From cross validation

The body fat example

- 20 healthy female subjects ages 25-34
- Y is fraction body fat
- X_1 is triceps skin fold thickness
- X_2 is thigh circumference
- X_3 is midarm circumference



$$VIF_1 = 708.84 \quad VIF_2 = 564.34 \quad VIF_3 = 104.61$$

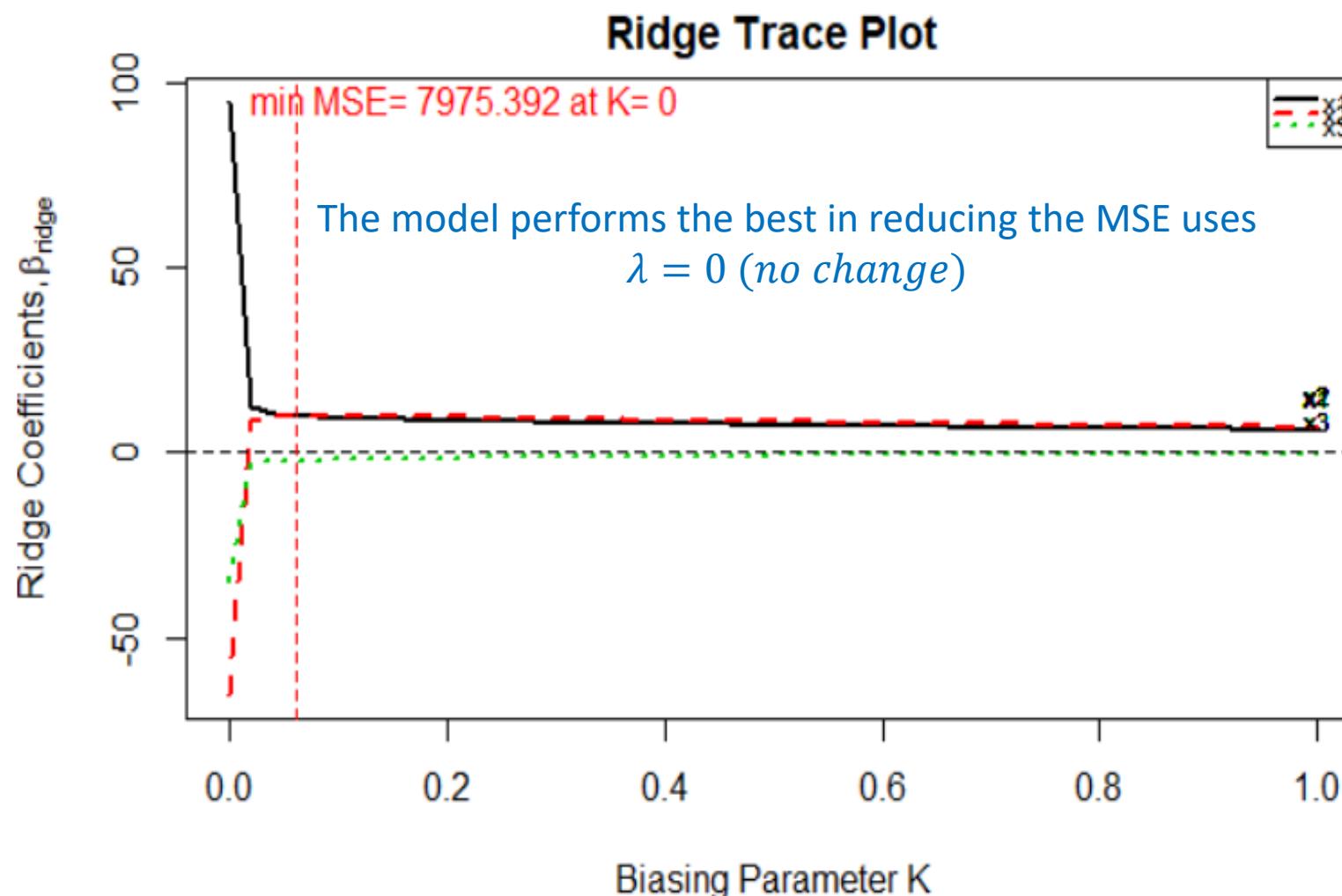
The selection of λ is subjective

```
library(MASS)
mod1<-lm.ridge(y~x1+x2+x3, data=bodyfat, lambda=seq(0, 1, 0.02))
plot(mod1)
select(mod1)
```



The selection of λ is subjective

```
library(lmridge)
mod2<-lmridge(y~x1+x2+x3, data=as.data.frame(bodyfat), K=seq(0,1, 0.02))
plot(mod2)
vif(mod2)
```



VIF
The model performs the best in reducing the multicollinearity uses $\lambda = 0.02$

| | x1 | x2 | x3 |
|--------|-----------|-----------|-----------|
| k=0 | 708.84291 | 564.34339 | 104.60601 |
| k=0.02 | 1.10255 | 1.08054 | 1.01051 |
| k=0.04 | 0.45279 | 0.55529 | 0.88140 |
| k=0.06 | 0.32437 | 0.44543 | 0.83060 |
| k=0.08 | 0.27615 | 0.39984 | 0.79339 |
| k=0.1 | 0.25155 | 0.37347 | 0.76137 |
| k=0.12 | 0.23639 | 0.35499 | 0.73232 |
| k=0.14 | 0.22577 | 0.34047 | 0.70540 |
| k=0.16 | 0.21762 | 0.32825 | 0.68022 |
| k=0.18 | 0.21096 | 0.31751 | 0.65652 |
| k=0.2 | 0.20525 | 0.30782 | 0.63415 |

VIF < 1 because VIF is closely related to $1/(1+\lambda^2)$ in the algorithm

Model summary for different λ

```
summary(lmridge(y~x1+x2+x3, data=as.data.frame(bodyfat), K=seq(0,1, 0.02)))
```

Coefficients: for Ridge parameter K= 0

| | Estimate | Estimate (Sc) | StdErr (Sc) | t-value (Sc) | Pr(> t) |
|-----------|----------|---------------|-------------|--------------|----------|
| Intercept | 117.0847 | 1914.1817 | 3412.1592 | 0.5610 | 0.5826 |
| x1 | 4.3341 | 94.8988 | 64.0559 | 1.4815 | 0.1579 |
| x2 | -2.8569 | -65.1851 | 57.1552 | -1.1405 | 0.2709 |
| x3 | -2.1861 | -34.7530 | 24.6072 | -1.4123 | 0.1770 |

Ridge Summary

| R2 | adj-R2 | DF ridge | F | AIC | BIC |
|---------|---------|----------|----------|----------|-----------|
| 0.80140 | 0.77800 | 3.00001 | 22.86042 | 37.86718 | 100.76904 |

Ridge minimum MSE= 7975.392 at K= 0

Coefficients: for Ridge parameter K= 1

| | Estimate | Estimate (Sc) | StdErr (Sc) | t-value (Sc) | Pr(> t) |
|-----------|----------|---------------|-------------|--------------|------------|
| Intercept | -2.2485 | -486.8614 | 71.6147 | -6.7983 | <2e-16 *** |
| x1 | 0.2844 | 6.2268 | 0.9584 | 6.4969 | <2e-16 *** |
| x2 | 0.3025 | 6.9013 | 1.0795 | 6.3930 | <2e-16 *** |
| x3 | -0.0083 | -0.1322 | 1.3966 | -0.0947 | 0.9256 |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Ridge Summary

| R2 | adj-R2 | DF ridge | F | AIC | BIC |
|---------|---------|----------|----------|----------|-----------|
| 0.33290 | 0.25440 | 1.15722 | 15.42241 | 43.60628 | 104.67321 |

Ridge minimum MSE= 7975.392 at K= 0

P-value for F-test (1.15722 , 18.37261) = 0.0006391673

Coefficients: for Ridge parameter K= 0.02

| | Estimate | Estimate (Sc) | StdErr (Sc) | t-value (Sc) | Pr(> t) |
|-----------|----------|---------------|-------------|--------------|------------|
| Intercept | -7.4034 | -633.1991 | 161.1205 | -3.9300 | 0.0011 ** |
| x1 | 0.5554 | 12.1599 | 2.5781 | 4.7167 | 0.0002 *** |
| x2 | 0.3681 | 8.4000 | 2.5522 | 3.2913 | 0.0043 ** |
| x3 | -0.1916 | -3.0464 | 2.4681 | -1.2343 | 0.2339 |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Ridge Summary

| R2 | adj-R2 | DF ridge | F | AIC | BIC |
|---------|---------|----------|----------|----------|----------|
| 0.76340 | 0.73560 | 2.00448 | 21.95136 | 37.75478 | 99.66535 |

Ridge minimum MSE= 7975.392 at K= 0

P-value for F-test (2.00448 , 17.93165) = 1.500203e-05

Statistical Inference from the Ridge Regression at $\lambda = 0.02$

Coefficients: for Ridge parameter K= 0.02

| | Estimate | Estimate (Sc) | StdErr (Sc) | t-value (Sc) | Pr(> t) | |
|-----------|----------|---------------|-------------|--------------|----------|-----|
| Intercept | -7.4034 | -633.1991 | 161.1205 | -3.9300 | 0.0011 | ** |
| x1 | 0.5554 | 12.1599 | 2.5781 | 4.7167 | 0.0002 | *** |
| x2 | 0.3681 | 8.4000 | 2.5522 | 3.2913 | 0.0043 | ** |
| x3 | -0.1916 | -3.0464 | 2.4681 | -1.2343 | 0.2339 | |
| --- | | | | | | |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Ridge Summary

| R2 | adj-R2 | DF ridge | F | AIC | BIC |
|---------|---------|----------|----------|----------|----------|
| 0.76340 | 0.73560 | 2.00448 | 21.95136 | 37.75478 | 99.66535 |

Ridge minimum MSE= 7975.392 at K= 0

P-value for F-test (2.00448 , 17.93165) = 1.500203e-05

- The Estimate (Sc) column shows that the coefficient estimate is scaled to reduce the impact of the predictor's unit (i.e., Km vs m)
- The Pr(> |t|) gives a pvalue for the significant marginal Effect test for X_i .

➤ E.g., X_3 has little marginal effect for a model with X_1 and X_2

- The Confidence interval

$$\text{Estimate (SC)} \pm t\left(1 - \frac{\alpha}{2}, n - df.\text{ridge}\right) \text{StdErr(SC)}$$

Could be used to estimate the linear impact of the predictor

➤ E.g., a 95% CI for X3's impact is estimated by

$$-3.0464 \pm t(0.975, 20 - 2.00448)2.4681 = (-8.25, 2.16)$$

- (IMPORTANT) The T and F method here still based on the assumption that the random error follows Normal with constant variance!
We could consider **Bootstrapping** for a more precise evaluation.

Remedial measures for influential cases—Robust regression

Robust regression

Tools that have been used to detect outliers and influential points.

- Hat matrix, studentized deleted residuals
- DFFITS, Cook's distance, and DEBETAS measures.
- LS method is particularly susceptible to outliers and influential cases.

Outlying and influential case may lead to the finding of model inadequacies.

- Missing interaction, missing important predictors or choice of an incorrect functional form

In OLS, using least square errors is not robust

- Outliers are heavily weighted

An alternative to discarding outlying cases that is less severe is to dampen the influence of these cases.

Iteratively reweighted least squares (IRLS) robust regression

1. Choose a ***weight function*** for weighting the case
2. Obtain the ***starting weights*** for all cases.
3. Use the starting weights in weighted least squares and ***obtain the residuals*** from the fitted Regression function.
4. Use the residuals in step 3 to obtain revised weights.
5. Continue the iterations until convergence, which can be judged by whether
 - The weights change relatively little, or
 - The residuals change relatively little, or
 - The estimated regression coefficients change relatively little, or
 - The fitted values change relatively little

$$W = \begin{bmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_n \end{bmatrix}$$

Where the w_i is computed differently for outliers or non-outliers.

Weight function (Huber estimator and Bisquare estimator)

Many weight functions have been proposed for dampening the influence of outlying cases.
Two widely used weight functions are the Huber and Tukey's Bisquare weight functions

$$\text{Huber: } w = \begin{cases} 1 & |u| \leq 1.345 \\ \frac{1.345}{|u|} & |u| > 1.345 \end{cases}$$

$$\text{Bisquare: } w = \begin{cases} \left(1 - \left(\frac{u}{4.685}\right)^2\right)^2 & |u| \leq 4.685 \\ 0 & |u| > 4.685 \end{cases}$$

w denotes the weight the u denotes the **scale residual**:

$u_i = e_i / MAD$, where MAD , the median absolute deviation estimator is

$$MAD = \frac{1}{0.6745} \text{median}\{|e_i - \text{median}\{e_i\}|\}$$

Comments on the scale residual:

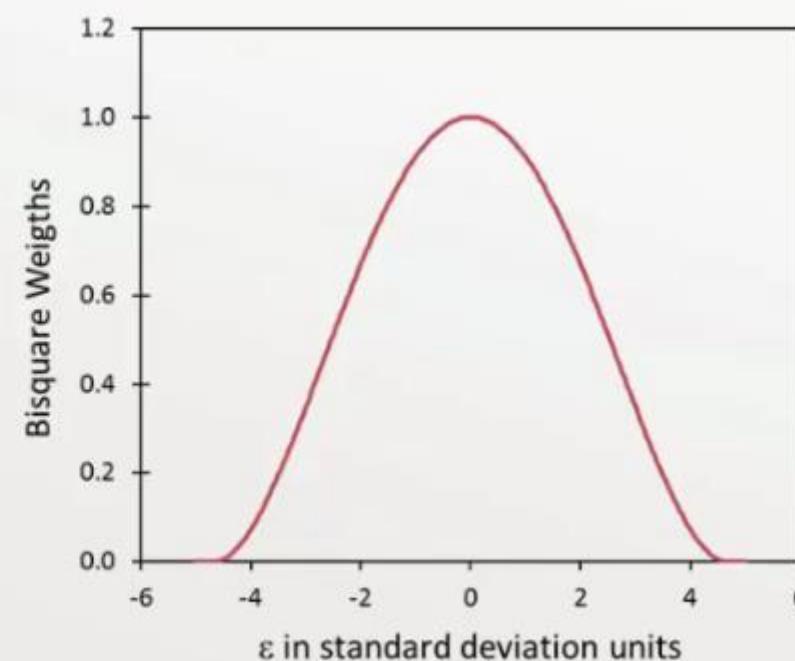
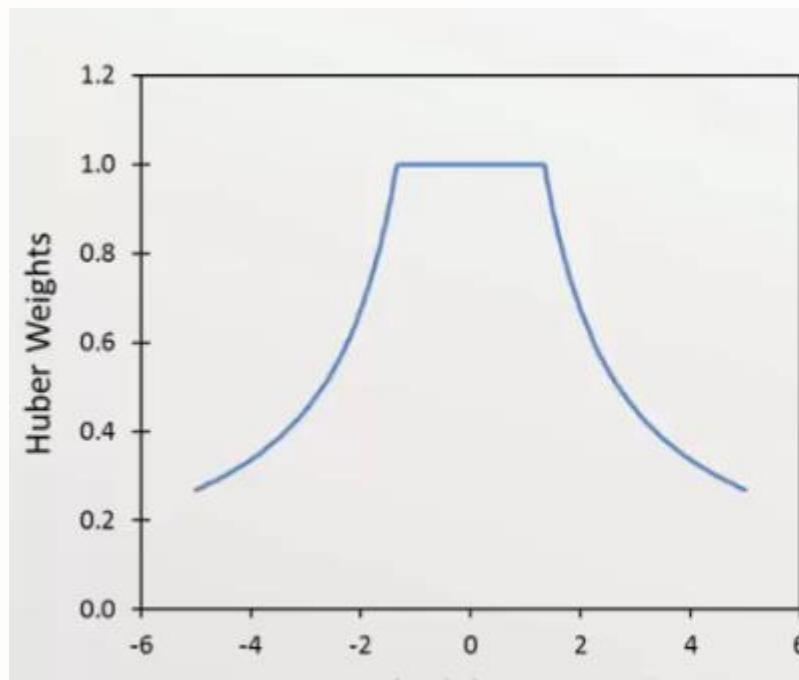
- We want to use some measurement that is more resistant to outliers, i.e., the median.
- The constant 0.6745 provides an unbiased estimate of σ for independent observations from a Normal Distribution

Weight function (Huber estimator and Bisquare estimator)

Many weight functions have been proposed for dampening the influence of outlying cases. Two widely used weight functions are the Huber and Tukey's Bisquare weight functions

$$\text{Huber: } w = \begin{cases} 1 & |u| \leq 1.345 \\ \frac{1.345}{|u|} & |u| > 1.345 \end{cases}$$

$$\text{Bisquare: } w = \begin{cases} \left(1 - \left(\frac{u}{4.685}\right)^2\right)^2 & |u| \leq 4.685 \\ 0 & |u| > 4.685 \end{cases}$$



With Bisquare we can throw in very extreme values

WLS and Robust Regression

- Both methods use weight function to adjust the influence of the observation on the Estimations. Both can handle unequal variance of in the error terms.
- WLS applies re-weighting on each observation in the sample, assuming the errors have a known variance structure, $|e_i| \sim U_i'X + \phi_i$. If the primary issue is heteroscedasticity, WLS should be considered.
- On the other hand, robust regression uses the weight function to trim the influence of outliers or influential observations based on their residuals. but not for other observations in the sample. If the main issue is the presence of outliers or influential observations, robust regression should be considered.
- In some cases, both robust regression and WLS may be used together.

Case study (Math proficiency)

The educational testing service study *America's smallest school: the family* investigated the relation of educational achievement of students to their home environment. Data on average mathematics proficiency (Mathprof, Y) and five home environment variables were obtained. The sample size **n=40**

Parents (X1): percentage of eighth-grade students with both parents living at home

Homelib (X2): percentage of eighth-grade students with three or more types of reading materials at home

Reading (X3): percentage of eighth-grade students who read more than 10 pages a day

Tvwatch (X4): percentage of eighth-grade students who watch TV for six hours or more a day

Absences (X5): percentage of eighth-grade students absent three days or more last month

| state | math proficiency | parents | home library | reading | TV watch | absence |
|-------------|------------------|---------|--------------|---------|----------|---------|
| Alabama | 252 | 75 | 78 | 34 | 18 | 18 |
| Arizona | 259 | 75 | 73 | 41 | 12 | 26 |
| Arkansas | 256 | 77 | 77 | 28 | 20 | 23 |
| California | 256 | 78 | 68 | 42 | 11 | 28 |
| Colorado | 267 | 78 | 85 | 38 | 9 | 25 |
| Connecticut | 270 | 79 | 86 | 43 | 12 | 22 |

The initial model

Parents (X1): percentage of eighth-grade students with both parents living at home

Homelib (X2): percentage of eighth-grade students with three or more types of reading materials at home

Reading (X3): percentage of eighth-grade students who read more than 10 pages a day

Tvwatch (X4): percentage of eighth-grade students who watch TV for six hours or more a day

Absences (X5): percentage of eighth-grade students absent three days or more last month

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|--------------|
| (Intercept) | 155.0304 | 36.2383 | 4.278 | 0.000145 *** |
| x1 | 0.3911 | 0.2571 | 1.521 | 0.137399 |
| x2 | 0.8639 | 0.1797 | 4.807 | 3.05e-05 *** |
| x3 | 0.3616 | 0.2690 | 1.345 | 0.187679 |
| x4 | -0.8467 | 0.3525 | -2.402 | 0.021927 * |
| x5 | 0.1923 | 0.2636 | 0.729 | 0.470718 |

Residual standard error: 5.268 on 34 degrees of freedom

Multiple R-squared: 0.861, Adjusted R-squared: 0.8406

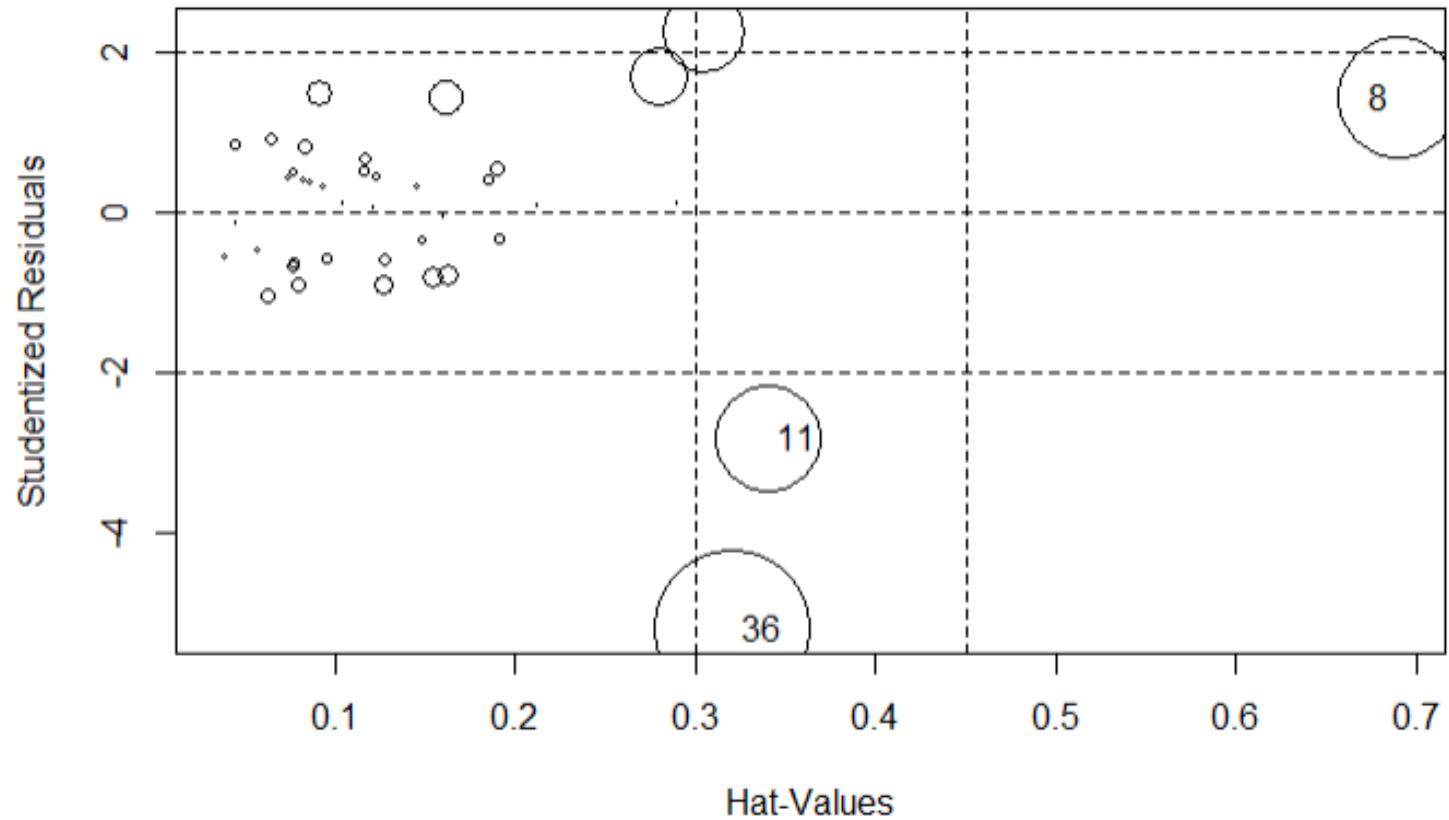
F-statistic: 42.13 on 5 and 34 DF, p-value: 1.276e-13

Analysis of variance Table

Response: y

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------|----|--------|---------|----------|---------------|
| x1 | 1 | 3732.4 | 3732.4 | 134.4896 | 2.303e-13 *** |
| x2 | 1 | 1647.0 | 1647.0 | 59.3468 | 5.863e-09 *** |
| x3 | 1 | 290.5 | 290.5 | 10.4693 | 0.002705 ** |
| x4 | 1 | 161.6 | 161.6 | 5.8245 | 0.021341 * |
| x5 | 1 | 14.8 | 14.8 | 0.5321 | 0.470718 |
| Residuals | 34 | 943.6 | 27.8 | | |

Diagnostics for the outlying and influential case



| State | studRes | Hat | CookD |
|------------------|---------|------|-------|
| 8D.C. | 1.41 | 0.69 | 0.72 |
| 11Guam | -2.83 | 0.34 | 0.57 |
| 36Virgin_islands | -5.21 | 0.32 | 1.21 |

A case i is considered influence point if $D_i >$

Major influence if $> F(0.5; 6, 34) = 0.91$

Moderate influence if less than 0.91 but greater than
 $F(0.2; 6, 34) = 0.51$

Any influence case?

D.C. Guam and Virgin-Islands

Best model selection

Parents (X1): percentage of eighth-grade students with both parents living at home

Homelib (X2): percentage of eighth-grade students with three or more types of reading materials at home

Reading (X3): percentage of eighth-grade students who read more than 10 pages a day

Tvwatch (X4): percentage of eighth-grade students who watch TV for six hours or more a day

Absences (X5): percentage of eighth-grade students absent three days or more last month

| p | 1 | 2 | 3 | 4 | 5 | SSEp | r2 | r2.adj | Cp | AICp | SBCp | PRESSp |
|---|---|---|---|---|---|-----------|-----------|-----------|-----------|----------|----------|----------|
| 1 | 2 | 0 | 0 | 1 | 0 | 1609.4257 | 0.7629677 | 0.7567300 | 21.992880 | 151.7901 | 155.1679 | 1883.644 |
| 2 | 3 | 0 | 1 | 0 | 1 | 1071.3398 | 0.8422157 | 0.8336868 | 4.603884 | 137.5114 | 142.5781 | 1392.568 |
| 3 | 4 | 0 | 1 | 1 | 1 | 1008.8965 | 0.8514122 | 0.8390299 | 4.353845 | 137.1093 | 143.8648 | 1412.810 |
| 4 | 5 | 1 | 1 | 1 | 1 | 958.3394 | 0.8588581 | 0.8427276 | 4.532109 | 137.0529 | 145.4973 | 1629.298 |
| 5 | 6 | 1 | 1 | 1 | 1 | 943.5723 | 0.8610330 | 0.8405966 | 6.000000 | 138.4317 | 148.5650 | 1832.519 |

We consider the model: $\hat{Y} = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$

Robust regression

Parents (X1): percentage of eighth-grade students with both parents living at home

Homelib (X2): percentage of eighth-grade students with three or more types of reading materials at home

Reading (X3): percentage of eighth-grade students who read more than 10 pages a day

Tvwatch (X4): percentage of eighth-grade students who watch TV for six hours or more a day

Absences (X5): percentage of eighth-grade students absent three days or more last month

We consider the model: $\hat{Y} = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$

The Robust Model Summary

```
library(MASS)
r<-rlm(y~x2+x3+x4, data=mathpro, psi=psi.bisquare)
```

Coefficients:

| | Value | Std. Error | t value |
|-------------|----------|------------|---------|
| (Intercept) | 207.6806 | 17.6965 | 11.7357 |
| x2 | 0.7972 | 0.1399 | 5.6982 |
| x3 | 0.1609 | 0.2209 | 0.7282 |
| x4 | -1.1692 | 0.2231 | -5.2412 |

Residual standard error: 4.342 on 36 degrees of freedom

OLS Model Summary

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|----------------|----------|------------|----------|--------------|
| (Intercept) | 199.6107 | 21.5289 | 9.272 | 4.50e-11 *** |
| x2 | 0.7804 | 0.1702 | 4.585 | 5.29e-05 *** |
| x3 | 0.4012 | 0.2688 | 1.493 | 0.14423 |
| x4 | -1.1565 | 0.2714 | -4.261 | 0.00014 *** |
| --- | | | | |
| Signif. codes: | 0 ‘***’ | 0.001 ‘**’ | 0.01 ‘*’ | 0.05 ‘.’ |
| | 0.1 ‘ ’ | 1 | | |

Residual standard error: 5.294 on 36 degrees of freedom

Multiple R-squared: 0.8514, Adjusted R-squared: 0.839

F-statistic: 68.76 on 3 and 36 DF, p-value: 5.646e-15

- The residual standard error for the OLS is 5.294, the robust model is better with a smaller s of 4.342.
- Access the robust model based on the residuals in a similar way as the OLS.

Remedial Measures for evaluating precision in Nonstandard situations:
Bootstrapping

Bootstrap method introduction

Conceptually simple but extremely powerful, nonparametric method for estimating precision when the standard approaches are unavailable.

Bootstrap methods allow approximate estimation of

- Confidence and prediction intervals in weighted regression, robust regression, or ridge regression
- Correct intervals when the errors are strongly non-normal

Robust regression

Parents (X1): percentage of eighth-grade students with both parents living at home

Homelib (X2): percentage of eighth-grade students with three or more types of reading materials at home

Reading (X3): percentage of eighth-grade students who read more than 10 pages a day

Tvwatch (X4): percentage of eighth-grade students who watch TV for six hours or more a day

Absences (X5): percentage of eighth-grade students absent three days or more last month

We consider the model: $\hat{Y} = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$

```
library(MASS)
r<-rlm(y~x2+x3+x4, data=mathpro, psi=psi.bisquare)
```

Coefficients:

| | value | std. Error | t value |
|-------------|----------|------------|---------|
| (Intercept) | 207.6806 | 17.6965 | 11.7357 |
| x2 | 0.7972 | 0.1399 | 5.6982 |
| x3 | 0.1609 | 0.2209 | 0.7282 |
| x4 | -1.1692 | 0.2231 | -5.2412 |

Residual standard error: 4.342 on 36 degrees of freedom

The confidence interval for the coefficients after the Robust regression

Coefficients:

| | value | std. Error | t value |
|-------------|----------|------------|---------|
| (Intercept) | 207.6806 | 17.6965 | 11.7357 |
| x2 | 0.7972 | 0.1399 | 5.6982 |
| x3 | 0.1609 | 0.2209 | 0.7282 |
| x4 | -1.1692 | 0.2231 | -5.2412 |

Residual standard error: 4.342 on 36 degrees of freedom

CI for the linear impact for X2, e.g., β_1 :

$$\begin{aligned} b_1 \pm t(0.975, 36)S(b_1) &= 0.7972 \pm 2.028(0.1399) \\ &= 0.7972 \pm 0.2837 = (0.5135, 1.0809) \end{aligned}$$

- (IMPORTANT) The T method here is still based on the assumption

We now evaluate the precision of the estimate $b_1=0.7972$ by the bootstrap method.

Basic bootstrap algorithm to evaluate the precision of the estimated coefficients

1. Randomly resample the available data ***with replacement*** to generate a new bootstrap sample with n equal to the original sample
2. Run regression on the bootstrap sample and save \hat{b} or \hat{Y}
3. Repeat Steps 1-2 B times to obtain an empirical sampling distribution for the parameters or fitted values (the bootstrap samples) $b_1^* \ b_1^* \dots \ b_1^*$
4. The standard deviation of the bootstrapped samples estimates standard error $s^*\{b_1^*\}$, and the quantiles of the bootstrapped values give approximate confidence intervals

For example, the 90% confidence interval is given by

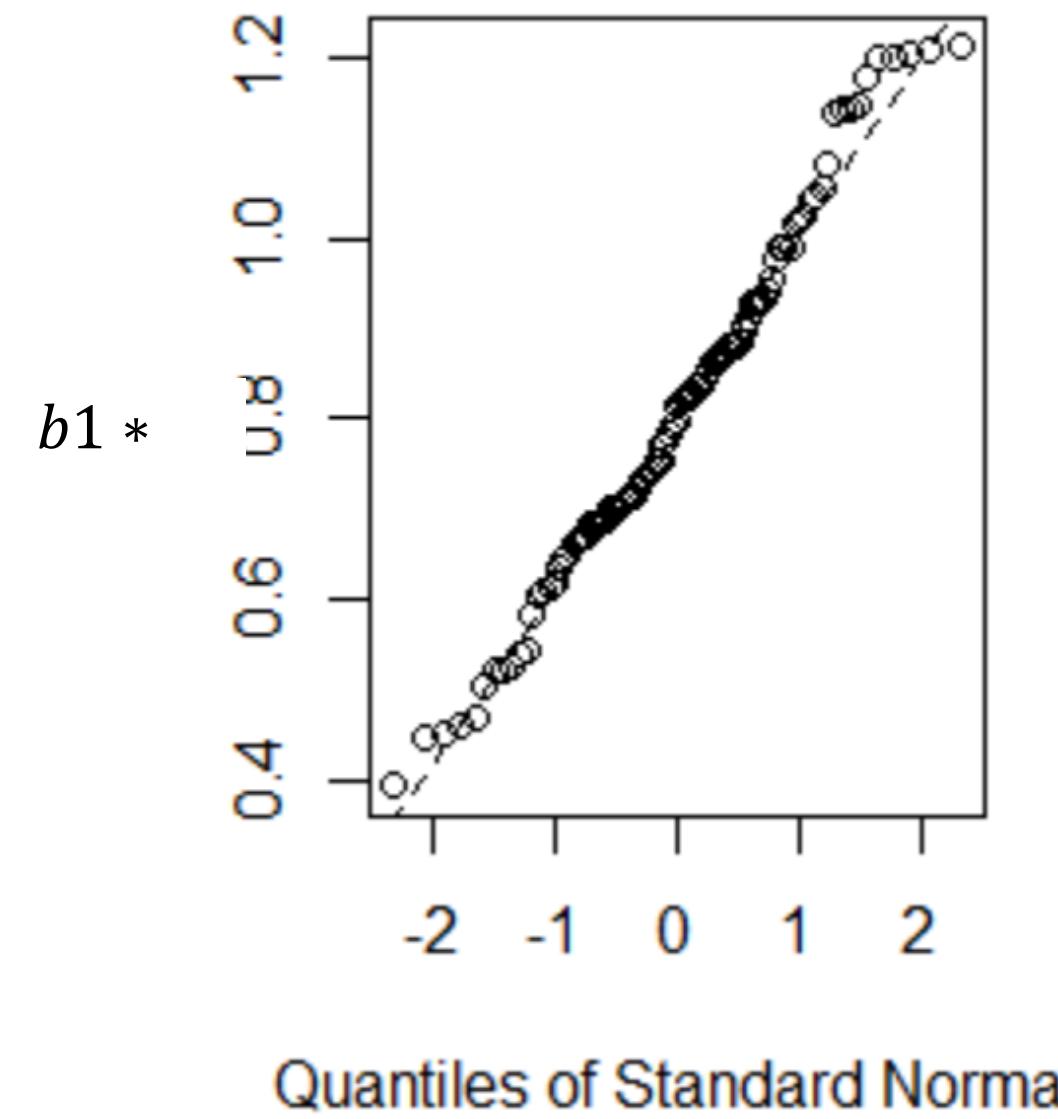
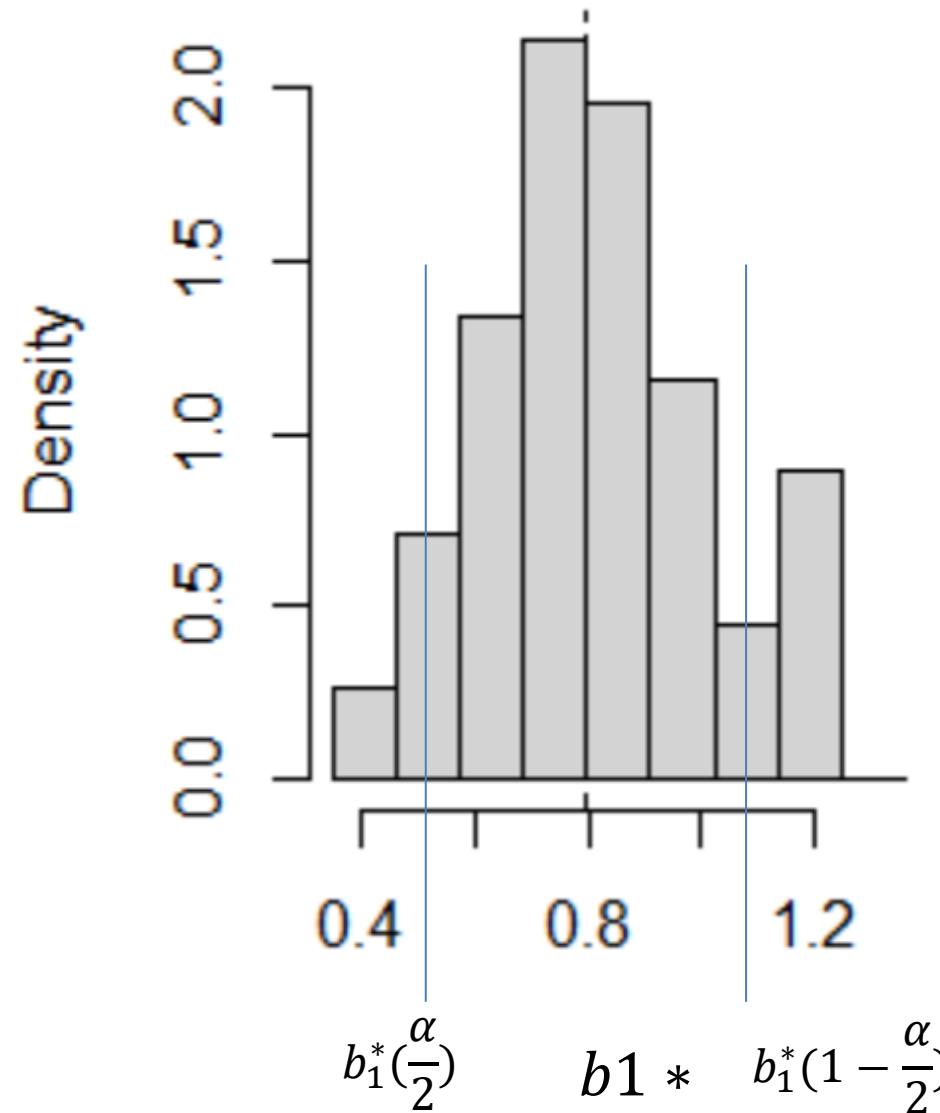
(the 5th percentile, the 95% percentile) denoted by $(b_1^*(0.05), b_1^*(0.95))$

In comparison, the 90% confidence interval, under Normal distribution, is given by a symmetric interval:
 $(b_1 - tSE(b_1), b_1 + tSE(b_1))$

The bootstrap resampling distribution of b_1

```
plot(mathpro.boot, index=2)
```

*Histogram of b_1^**



```

library(boot)
boot.huber <- function(data, indices, maxit=100){
  data <- data[indices,] # select obs. in bootstrap sample
  mod <- rlm(y ~ x2+x3+x4, data=data, maxit=maxit)
  coefficients(mod) # return coefficient vector
}

mathpro.boot<-boot(data=mathpro,statistic=boot.huber, R=100,maxit=100)

```

Bootstrap Statistics :

| | original | bias | std. error |
|-----|-------------|-------------|------------|
| t1* | 207.6806290 | -3.10601076 | 22.6057370 |
| t2* | 0.7971940 | 0.01232840 | 0.1951210 |
| t3* | 0.1608632 | 0.04356760 | 0.2686475 |
| t4* | -1.1692169 | 0.02979697 | 0.2174511 |

- “original” is the value of the estimates computed from Robust model.
- “bias” is the difference between the average of the bootstrap samples and the original.
- “std. error” is the standard deviation of the bootstrap samples.
- Check out the course website for R markdown file for examples on how to apply Bootstrapping on WLS and Ridge model.

The bootstrap confidence interval for β_1

```
boot.ci(mathpro.boot, index=2, type="perc")
```

Intervals :

Level Percentile

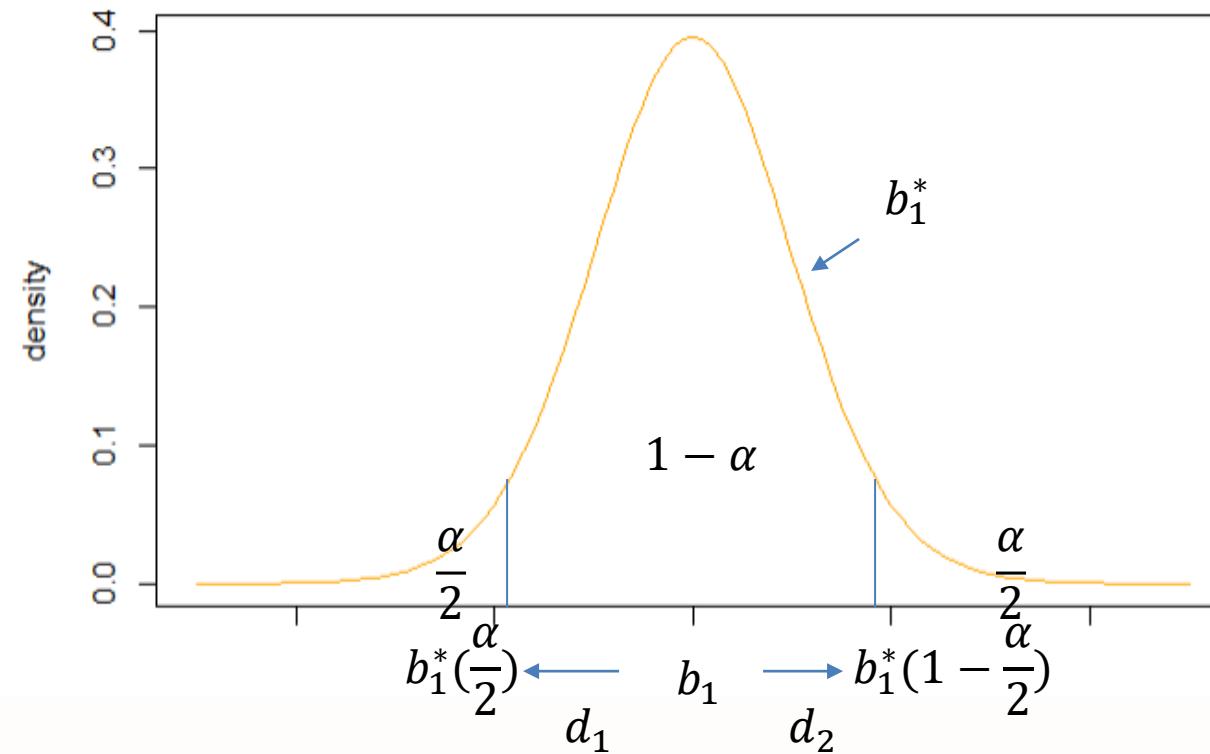
95% (0.4493, 1.2068) ← Computing by hand is not required.

Calculations and Intervals on Original Scale

Some percentile intervals may be unstable

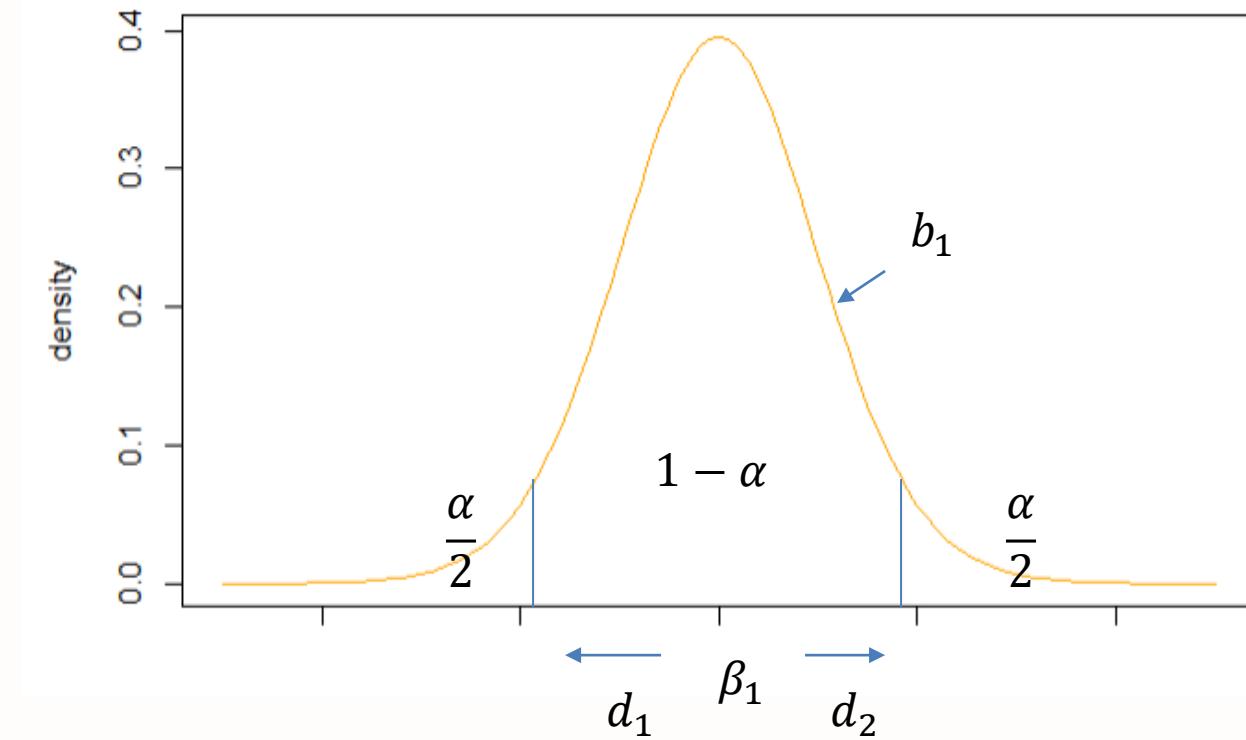
Comparing to the Robust CI for β_1 : $b_1 \pm t(0.975, 36)S(b_1) = 0.7972 \pm 2.028(0.1399)$
 $= 0.7972 \pm 0.2837 = (0.5135, 1.0809)$

(Optional) Use the reflection method to estimate the empirical confidence interval for β_1



$$d_1 = b_1 - b_1^*\left(\frac{\alpha}{2}\right)$$

$$d_2 = b_1^*\left(1 - \frac{\alpha}{2}\right) - b_1$$



$$\beta_1 - d_1 \leq \beta_1 \leq \beta_1 + d_2$$

Hence the $1 - \alpha$ confidence interval for β_1 is

$$b_1 - d_2 \leq \beta_1 \leq b_1 + d_1$$

One-Way ANOVA

Factor effect model

One-way Analysis of Variance (ANOVA)

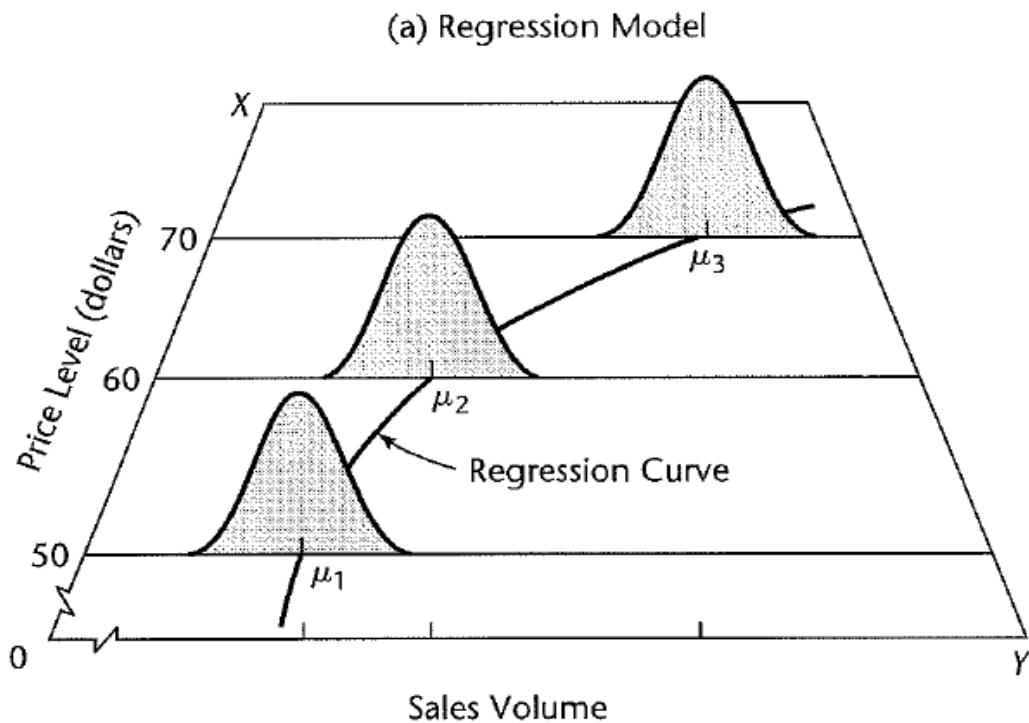
- Y is a continuous variable (just like regular regression)
- X is a categorical variable with $r \geq 2$ distinct values
- In ANOVA terminology, X is a *factor* with *r levels*
- Typically, the levels represent different groups, subpopulations, or treatments
- Because X is no longer continuous, our model no longer expresses \hat{Y} as a smooth function of X . We are now interested in *differences* among the mean responses for the various factor levels.

Relation between Regression and ANOVA

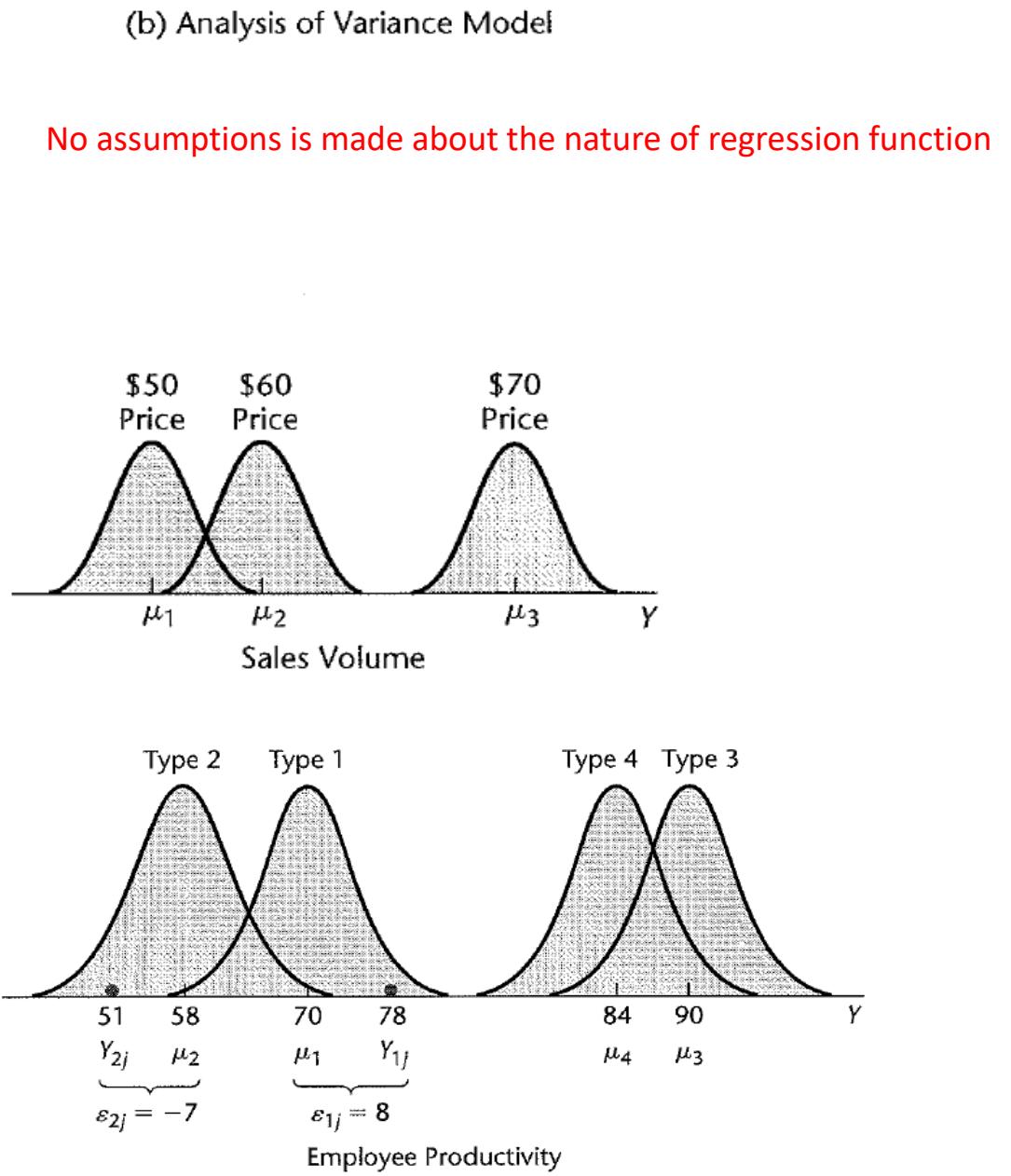
In regression, we aimed to estimate the parameters of a deterministic equation that expressed the conditional expectation of Y as a function of X .

In ANOVA, our common objectives are slightly different:

1. Determine whether any differences in $E(Y)$ exist among the factor levels
2. Determine which specific factor levels differ from each other
3. Estimate the differences in $E(Y)$ among various levels
(or equivalently, estimate the population means for Y within different levels)



$$\varepsilon = Y - \hat{Y} = Y - X\beta$$



The Cell means model

- Until now, we have used the index i to represent individual cases in the data.
- For ANOVA, use i to represent a factor level, $i = 1, \dots, r$
- Individual cases within each level are represented by the index $j, j = 1, 2, \dots, n_i$
- Y_{ij} is the value of the response for the j -th individual in factor level i .
- We will also (eventually) transition away from representing parameters as β to representing them as μ or T

The Cell means model

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad \varepsilon_{ij} \sim \text{Normal}(0, \sigma^2)$$

$$E(Y_{ij}) = \mu_i \quad \sigma(Y_{ij}) = \sigma^2 \quad Y_{ij} \sim \text{Normal}(\mu_i, \sigma^2)$$

$\bar{Y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{i,j}$ is the sample mean for observations from level i .

$\bar{Y}_{..} = \frac{1}{n_T} \sum_{i=1}^r \sum_{j=1}^{n_i} Y_{i,j}$ is the mean over all of the observations.

$n_T = \sum_{i=1}^r n_i$ is the total sample size.

The Cell means model is an essentially a linear model, $Y = X\beta + \varepsilon$

For example, if $r = 3, n_1 = n_2 = n_3 = 2, n = 6$

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}$$

$$\mathbf{E}\{\mathbf{Y}\} = \begin{bmatrix} E\{Y_{11}\} \\ E\{Y_{12}\} \\ E\{Y_{21}\} \\ E\{Y_{22}\} \\ E\{Y_{31}\} \\ E\{Y_{32}\} \end{bmatrix} = \mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_1 \\ \mu_2 \\ \mu_2 \\ \mu_3 \\ \mu_3 \end{bmatrix}$$

$$\sigma^2\{\mathbf{Y}\} = \sigma^2\{\boldsymbol{\varepsilon}\} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}$$

The Cell means model is a linear model, $Y = X\mu + \varepsilon$, with *no intercepts*

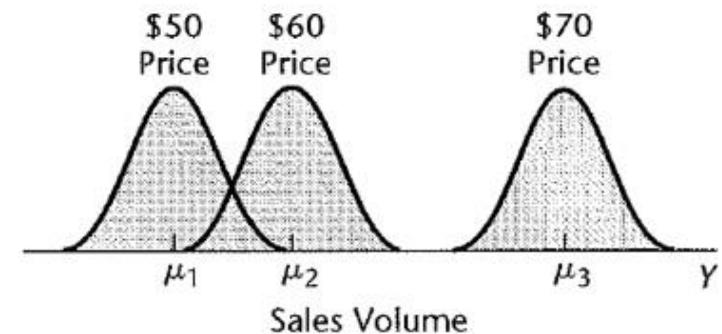
For example, if $r = 3, n_1 = n_2 = n_3 = 2, n = 6$

$$\begin{bmatrix} Y_{1,1} \\ Y_{1,2} \\ Y_{2,1} \\ Y_{2,2} \\ \vdots \\ Y_{r,n_r} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \cdot \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_r \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,1} \\ \varepsilon_{1,2} \\ \varepsilon_{2,1} \\ \varepsilon_{2,2} \\ \vdots \\ \varepsilon_{r,n_r} \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\mu} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I})$$

The model assumes that,

- The errors (and therefore the observations) are independent
- The errors are normally distributed (but the CLT still applies)
- The errors have constant variance
- Subpopulations associated with different levels of the factor *might* have different mean responses



Estimation for the cell means model

Because X is discrete, estimates for the ANOVA model can be computed without linear algebra by using the standard equations for the sample mean and sample variance.

For example, minimize $Q_i = \sum(Y_{ij} - \mu_i)^2$ with respect to μ_i

For each level i , the true *within-group mean*, μ_i , is estimated as,

$$\hat{\mu}_i = \bar{Y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{i,j}$$

and the *within-group sample variance* is

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{i,j} - \bar{Y}_{i\cdot})^2$$

The within-group sample variances are treated as “data” about the value of the true error variance, σ^2 , which is estimated by taking a weighted average (“pooling” the variances):

$$\begin{aligned}
 s^2 &= \frac{\sum_{i=1}^r (n_i - 1)s_i^2}{\sum_{i=1}^r (n_i - 1)} \\
 &= \frac{1}{n_T - r} \sum_{i=1}^r (n_i - 1)s_i^2 \\
 &= \frac{1}{n_t - r} \sum_{i=1}^r \sum_{j=1}^{n_i} (Y_{i,j} - \bar{Y}_{i\cdot})^2 \\
 &= MSE
 \end{aligned}$$

Note: If $n_i = n$ for all i , this equation reduces to a simple mean, $s^2 = \frac{1}{r} \sum s_i^2$

This is also known as the “balanced design”. If $n_i \neq n$, s^2 will be weighted by group size.

ANOVA table

| Source of Variation | SS | df | MS | $E\{MS\}$ |
|--------------------------|--|---------|-------------------------|--|
| ANOVA model fixed effect | $SSR = \sum n_i (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2$ | $r - 1$ | $MSR = \frac{SSR}{r-1}$ | $\sigma^2 + \frac{n \sum (\mu_i - \mu_{\cdot\cdot})^2}{r - 1}$ |
| Error | $SSE = \sum (Y_{ij} - \bar{Y}_{i\cdot})^2$ | $n - r$ | $MSE = \frac{SSE}{n-r}$ | σ^2 |
| Total | $SSTO = \sum (Y_{ij} - \bar{Y}_{..})^2$ | $n - 1$ | | |

Under $H_0 : (\mu_1 = \mu_2 = \dots = \mu_r)$,

$$F^* = \frac{MSM}{MSE} \sim F_{r-1, n_t - r}$$

If $p = P(F_{r-1, n_t - r} \geq F^*) \leq \alpha \rightarrow \text{reject } H_0$

If we reject H_0 , we conclude that *at least one* of the factor levels has a group mean that is different from the others.

Factor Effects Model

Factor effects simply reparameterize the cell means model so that the parameters now represent differences (i.e., “effects”) relative to a selected baseline reference.

Advantages:

- easier to interpret null hypotheses
- an effect of 0 for a particular level indicates that the level is not different from the **reference**
- positive and negative effects have similarly natural interpretations

Disadvantage:
somewhat more convoluted notation. Choice of reference matters.

Factor Effects Model

$$\mu_i = \mu_{\cdot} + (\mu_i - \mu_{\cdot})$$

Let $\tau_i = \mu_i - \mu_{\cdot}$.

$$\mu_i = \mu_{\cdot} + \tau_i$$

$$\text{Then } Y_{ij} = \mu_{\cdot} + \tau_i + \varepsilon_{ij}$$

- μ_{\cdot} is the (unknown) population mean for the **baseline reference**, common to all observations
- τ_i is the *i th factor level effect* or the *i th treatment effect*.
- ε_{ij} are independent $N(0, \sigma^2)$ $i = 1, \dots, r$ and $j = 1, \dots, n_i$
- Factor effects model and cell means model are equivalent for modeling data.

Factor effects model $Y_{ij} = \mu_{\cdot} + \tau_i + \varepsilon_{ij}$

Cell means model $Y_{ij} = \mu_i + \varepsilon_{ij}$

- Factor effect model uses intercept (β_0 or μ_{\cdot}) to represent the baseline level. Other levels are compared to the baseline ($\beta_i = \tau_i = \mu_i - \mu_{\cdot}$) $i = 1, \dots, r$
- Cell mean model doesn't use intercept. All levels are estimated with β_i , $i = 1, \dots, r$

Some basic choices of the reference and the response function

Example: suppose r=3

Unweighted mean

$$\mu_{\cdot} = \frac{(\sum_{i=1}^r \mu_i)}{r}$$

Subject to restriction that $(\sum_{i=1}^r \tau_i = 0)$

- The parameter vector is $(\mu_{\cdot}, \tau_1, \tau_2)$
- For level 1: $E(Y) = \mu_1 = \mu_{\cdot} + \tau_1$
- For level 2: $E(Y) = \mu_2 = \mu_{\cdot} + \tau_2$
- For level 3: $E(Y) = \mu_3 = \mu_{\cdot} + \tau_3 = \mu_{\cdot} - \tau_1 - \tau_2$

The first factor mean

$$\mu_{\cdot} = \mu_1$$

- The parameter vector is $(\mu_{\cdot}, \tau_2, \tau_3)$
- For level 1: $E(Y) = \mu_1 = \mu_{\cdot} + \tau_1 = \mu_1 + 0$
- For level 2: $E(Y) = \mu_2 = \mu_{\cdot} + \tau_2 = \mu_1 + (\mu_2 - \mu_1)$
- For level 3: $E(Y) = \mu_3 = \mu_{\cdot} + \tau_3 = \mu_1 + (\mu_3 - \mu_1)$

The second factor mean

$$\mu_{\cdot} = \mu_2$$

- The parameter vector is $(\mu_{\cdot}, \tau_1, \tau_3)$
- For level 1: $E(Y) = \mu_1 = \mu_{\cdot} + \tau_1 = \mu_2 + (\mu_1 - \mu_2)$
- For level 2: $E(Y) = \mu_2 = \mu_{\cdot} + \tau_2 = \mu_2 + 0$
- For level 3: $E(Y) = \mu_3 = \mu_{\cdot} + \tau_3 = \mu_2 + (\mu_3 - \mu_2)$

Factor Effects Model with Unweighted Mean

$$\mu_{\cdot} = \frac{(\sum_{i=1}^r \mu_i)}{r} \quad \text{Subject to restriction that } (\sum_{i=1}^r \tau_i = 0)$$

We shall use only the parameters $\mu_{\cdot}, \tau_1, \dots, \tau_{r-1}$ for the linear model, since $\tau_r = -\tau_1 - \tau_2 - \dots - \tau_{r-1}$

Consider a single factor study with $r = 3$ factor levels when $n_1 = n_2 = n_3 = 2$.

The matrix form $Y = X\beta + \varepsilon$ can be specified as

$$Y = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{pmatrix}$$

$$\beta = \begin{pmatrix} \mu_{\cdot} \\ \tau_1 \\ \tau_2 \end{pmatrix}$$

$$\varepsilon = \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{pmatrix}$$

$$E\{Y\} = X\beta = \begin{pmatrix} \mu_{\cdot} + \tau_1 \\ \mu_{\cdot} + \tau_1 \\ \mu_{\cdot} + \tau_2 \\ \mu_{\cdot} + \tau_2 \\ \mu_{\cdot} - \tau_1 - \tau_2 \\ \mu_{\cdot} - \tau_1 - \tau_2 \end{pmatrix}$$

The intercept is back for the reference mean

Factor Effects Model with Unweighted Mean

$H_0: \tau_1 = \tau_2 = \dots = \tau_{r-1} = 0$

$H_0: \text{not all } \tau_i = 0$

| | x_1 | x_2 | x_3 | |
|----|-------------|---------|---------|---------|
| | (Intercept) | design1 | design2 | design3 |
| 1 | 1 | 1 | 0 | 0 |
| 2 | 1 | 1 | 0 | 0 |
| 3 | 1 | 1 | 0 | 0 |
| 4 | 1 | 1 | 0 | 0 |
| 5 | 1 | 1 | 0 | 0 |
| 6 | 1 | 0 | 1 | 0 |
| 7 | 1 | 0 | 1 | 0 |
| 8 | 1 | 0 | 1 | 0 |
| 9 | 1 | 0 | 1 | 0 |
| 10 | 1 | 0 | 1 | 0 |
| 11 | 1 | 0 | 0 | 1 |
| 12 | 1 | 0 | 0 | 1 |
| 13 | 1 | 0 | 0 | 1 |
| 14 | 1 | 0 | 0 | 1 |
| 15 | 1 | -1 | -1 | -1 |
| 16 | 1 | -1 | -1 | -1 |
| 17 | 1 | -1 | -1 | -1 |
| 18 | 1 | -1 | -1 | -1 |
| 19 | 1 | -1 | -1 | -1 |

Analysis of Variance Table

Response: y

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------|----|--------|---------|---------|---------------|
| Design_uw | 3 | 588.22 | 196.074 | 18.591 | 2.585e-05 *** |
| Residuals | 15 | 158.20 | 10.547 | | |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|---------------------------|----------|------------|---------|--------------|
| $\mu.$ (Intercept) | 18.6750 | 0.7485 | 24.949 | 1.25e-13 *** |
| τ_1 Design_uwdesign1 | -4.0750 | 1.2708 | -3.207 | 0.005884 ** |
| τ_2 Design_uwdesign2 | -5.2750 | 1.2708 | -4.151 | 0.000854 *** |
| τ_3 Design_uwdesign3 | 0.8250 | 1.3706 | 0.602 | 0.556221 |
| --- | | | | |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.248 on 15 degrees of freedom
Multiple R-squared: 0.7881, Adjusted R-squared: 0.7457
F-statistic: 18.59 on 3 and 15 DF, p-value: 2.585e-05

$$E\{Y_1\} = \underline{\mu + \tau_1} = 18.675 - 4.075 = 14.6 \quad E\{Y_2\} = \underline{\mu + \tau_2} = 18.675 - 5.275 = 13.4$$

$$E\{Y_3\} = \underline{\mu + \tau_3} = 18.675 + 0.825 = 19.5 \quad E\{Y_4\} = \underline{\mu - \tau_1 - \tau_2 - \tau_3} = 18.675 + 4.075 + 5.275 - 0.825 = 27.2$$

Note: the t value and the p value are for testing the significance of the corresponding coefficients of the same row.

Factor Effects Model with the first group 1 as reference mean (default)

$$\mu = \mu_1 \quad \tau_1 = 0$$

| | (Intercept) | design2 | design3 | design4 |
|----|-------------|---------|---------|---------|
| 1 | 1 | 0 | 0 | 0 |
| 2 | 1 | 0 | 0 | 0 |
| 3 | 1 | 0 | 0 | 0 |
| 4 | 1 | 0 | 0 | 0 |
| 5 | 1 | 0 | 0 | 0 |
| 6 | 1 | 1 | 0 | 0 |
| 7 | 1 | 1 | 0 | 0 |
| 8 | 1 | 1 | 0 | 0 |
| 9 | 1 | 1 | 0 | 0 |
| 10 | 1 | 1 | 0 | 0 |
| 11 | 1 | 0 | 1 | 0 |
| 12 | 1 | 0 | 1 | 0 |
| 13 | 1 | 0 | 1 | 0 |
| 14 | 1 | 0 | 1 | 0 |
| 15 | 1 | 0 | 0 | 1 |
| 16 | 1 | 0 | 0 | 1 |
| 17 | 1 | 0 | 0 | 1 |
| 18 | 1 | 0 | 0 | 1 |
| 19 | 1 | 0 | 0 | 1 |
| | ... | | | |

Analysis of Variance Table

Response: y

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------|----|--------|---------|---------|---------------|
| Design_uw | 3 | 588.22 | 196.074 | 18.591 | 2.585e-05 *** |
| Residuals | 15 | 158.20 | 10.547 | | |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | | | | | | | | |
|------------------------|----------|------------|---------|--------------|------|-----|------|---|-----|---|---|---|
| μ . (Intercept) | 14.600 | 1.452 | 10.053 | 4.66e-08 *** | | | | | | | | |
| τ_2 Designdesign2 | -1.200 | 2.054 | -0.584 | 0.5677 | | | | | | | | |
| τ_3 Designdesign3 | 4.900 | 2.179 | 2.249 | 0.0399 * | | | | | | | | |
| τ_4 Designdesign4 | 12.600 | 2.054 | 6.135 | 1.91e-05 *** | | | | | | | | |
| --- | | | | | | | | | | | | |
| Signif. codes: | 0 | '***' | 0.001 | '**' | 0.01 | '*' | 0.05 | . | 0.1 | ' | ' | 1 |

Residual standard error: 3.248 on 15 degrees of freedom
Multiple R-squared: 0.7881, Adjusted R-squared: 0.7457
F-statistic: 18.59 on 3 and 15 DF, p-value: 2.585e-05

$$E\{Y_1\} = \underline{\mu + \tau_1} = 14.6 + 0 = 14.6 \quad E\{Y_2\} = \underline{\mu + \tau_2} = 14.6 - 1.2 = 13.4$$

$$E\{Y_3\} = \underline{\mu + \tau_3} = 14.6 + 4.9 = 19.5 \quad E\{Y_4\} = \underline{\mu + \tau_4} = 14.6 + 12.6 = 27.2$$

Note: the t value and the p value are for testing the significance of the corresponding coefficients of the same row.

Factor Effects Model with the first group 2 as reference mean (Relevel)

$$\mu_0 = \mu_2 \quad \tau_2 = 0$$

| | (Intercept) | design1 | design3 | design4 |
|----|-------------|---------|---------|---------|
| 1 | 1 | 1 | 0 | 0 |
| 2 | 1 | 1 | 0 | 0 |
| 3 | 1 | 1 | 0 | 0 |
| 4 | 1 | 1 | 0 | 0 |
| 5 | 1 | 1 | 0 | 0 |
| 6 | 1 | 0 | 0 | 0 |
| 7 | 1 | 0 | 0 | 0 |
| 8 | 1 | 0 | 0 | 0 |
| 9 | 1 | 0 | 0 | 0 |
| 10 | 1 | 0 | 0 | 0 |
| 11 | 1 | 0 | 1 | 0 |
| 12 | 1 | 0 | 1 | 0 |
| 13 | 1 | 0 | 1 | 0 |
| 14 | 1 | 0 | 1 | 0 |
| 15 | 1 | 0 | 0 | 1 |
| 16 | 1 | 0 | 0 | 1 |
| 17 | 1 | 0 | 0 | 1 |
| 18 | 1 | 0 | 0 | 1 |
| 19 | 1 | 0 | 0 | 1 |

Analysis of Variance Table

Response: y

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------|----|--------|---------|---------|---------------|
| Design_uw | 3 | 588.22 | 196.074 | 18.591 | 2.585e-05 *** |
| Residuals | 15 | 158.20 | 10.547 | | |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------------------|----------|------------|---------|--------------|
| μ_0 (Intercept) | 13.400 | 1.452 | 9.226 | 1.43e-07 *** |
| τ_1 Design2design1 | 1.200 | 2.054 | 0.584 | 0.5677 |
| τ_3 Design2design3 | 6.100 | 2.179 | 2.800 | 0.0135 * |
| τ_4 Design2design4 | 13.800 | 2.054 | 6.719 | 6.88e-06 *** |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.248 on 15 degrees of freedom
Multiple R-squared: 0.7881, Adjusted R-squared: 0.7457
F-statistic: 18.59 on 3 and 15 DF, p-value: 2.585e-05

$$E\{Y_{1.}\} = \underline{\mu_0 + \tau_1} = 13.4 + 1.2 = 14.6$$

$$E\{Y_{2.}\} = \underline{\mu_0 + \tau_2} = 13.4$$

$$E\{Y_{3.}\} = \underline{\mu_0 + \tau_3} = 13.4 + 6.1 = 19.5$$

$$E\{Y_{4.}\} = \underline{\mu_0 + \tau_4} = 13.4 + 13.8 = 27.2$$

Note: the t value and the p value are for testing the significance of the corresponding coefficients of the same row.

Estimation and hypotheses on the following effects

- A single factor level mean μ_i
- A difference between two factor level means
- A contrast among factor level means
- A linear combination of factor level means.
- Multiple and simultaneous comparison

A single factor level and difference between two factor levels

$$Ho: \mu_i = c \quad Ha: \mu_i \neq c$$

$$ts = \frac{\bar{Y}_i - c}{s\{\bar{Y}_i\}} \quad s^2\{\bar{Y}_i\} = \frac{MSE}{n_i} \quad \leftarrow \sigma^2(\bar{Y}) = \frac{\sigma^2}{n}$$

$$CI: \bar{Y}_i \pm t\left(1 - \frac{\alpha}{2}; n_T - r\right) s\{\bar{Y}_i\}$$

$$Ho: \mu_2 = 0 \quad Ha: \mu_2 \neq 0$$

$$\bar{Y}_2 = 13.4 \quad s^2\{\bar{Y}_2\} = \frac{10.55}{5} = 2.11, \text{ so } s\{\bar{Y}_2\} = 1.453$$

$$ts = \frac{\bar{Y}_i}{s\{\bar{Y}_i\}} = 9.22 \quad \text{Reject if } t_s > t(0.975; 15) = 2.131$$

$$CI: \bar{Y}_i \pm t\left(1 - \frac{\alpha}{2}; n_T - r\right) s\{\bar{Y}_i\} = 13.4 \pm 2.131(1.453) \\ = 13.4 \pm 3.096 = 10.3, 16.6$$

$$Ho: \mu_i - \mu_j = 0 \quad Ha: \mu_i - \mu_j \neq 0$$

$$ts = \frac{\bar{Y}_i - \bar{Y}_j}{s\{\bar{Y}_i - \bar{Y}_j\}} \quad s^2\{\bar{Y}_i - \bar{Y}_j\} = MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right) \quad \leftarrow \sigma^2(\bar{Y}_1 \pm \bar{Y}_2) \\ = \sigma^2(\bar{Y}_1) + \sigma^2(\bar{Y}_2)$$

$$CI: \bar{Y}_i - \bar{Y}_j \pm t\left(1 - \frac{\alpha}{2}; n_T - r\right) s\{\bar{Y}_i - \bar{Y}_j\}$$

$$Ho: \mu_2 - \mu_1 = 0 \quad Ha: \mu_2 - \mu_1 \neq 0$$

$$\bar{Y}_2 - \bar{Y}_1 = 13.4 - 14.6 = -1.2$$

$$s^2\{\bar{Y}_i - \bar{Y}_j\} = MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right) = 4.22$$

$$ts = \frac{-1.2}{2.054} = -0.584 \quad \text{Reject if } |t_s| > t(0.975; 15) = 2.131$$

$$CI: \bar{Y}_i - \bar{Y}_j \pm t\left(1 - \frac{\alpha}{2}; n_T - r\right) s\{\bar{Y}_i - \bar{Y}_j\} = -1.2 \pm 2.131(2.054) \\ = -1.2 \pm 4.377 = -5.58, 3.18$$

Contrast of factor level means (not simultaneous comparison)

A **contrast** is a comparison involving two or more factor level means. A contrast will be denoted by L , and is defined as

$$L = \sum_{i=1}^r c_i \mu_i \quad \text{Where } \sum_{i=1}^r c_i = 0$$

For example:

$$1. L = \mu_1 - \mu_2 \quad c_1 = 1, c_2 = -1, c_3 = 0, c_4 = 0$$

$$2. L = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2} \quad c_1 = \frac{1}{2}, c_2 = \frac{1}{2}, c_3 = -\frac{1}{2}, c_4 = -\frac{1}{2}$$

$$3. L = \mu_1 - \frac{\mu_1 + \mu_2 + \mu_3 + \mu_4}{4} \quad c_1 = \frac{3}{4}, c_2 = -\frac{1}{4}, c_3 = -\frac{1}{4}, c_4 = -\frac{1}{4}$$

$$\hat{L} = \sum_{i=1}^r c_i \bar{Y}_i$$

$$s^2\{L\} = \text{MSE} \sum_{i=1}^r c_i^2 / n_i$$

$$\frac{\hat{L} - L}{s\{L\}} \sim t(n_T - r) \text{ for ANOVA}$$

For example: Ho: $\frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2} = 0$ and Ha: $\frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2} \neq 0$

$$\hat{L} = \sum_{i=1}^r c_i \bar{Y}_i = -9.35 \quad s^2\{L\} = \text{MSE} \sum_{i=1}^r c_i^2 / n_i = 2.242 \quad t_s = \frac{\hat{L} - L}{s\{L\}} = -6.23 \sim t(15) \quad \text{The CI for L: } (-12.54, -6.16)$$

```
oneway(cereal$y, cereal$design, mc=matrix(c(0.5, 0.5, -0.5, -0.5), 1, 4))$contrast.NOT.simultaneous
```

```
$Contrast.NOT.simultaneous
```

| L | lower | upper | t | p-value |
|-----------|------------|-----------|-----------|----------|
| -9.350000 | -12.540892 | -6.159108 | -6.245605 | 0.000016 |

Bonferroni multiple comparison

We want compare g linear combination Ls'. $L = \sum_{i=1}^r c_i \mu_i$ where $\sum_{i=1}^r c_i = 0$

$$\hat{L} \pm Bs\{\hat{L}\}, \text{where } B = t\left(1 - \frac{\alpha}{2g}; n_T - r\right)$$

For example:

$$1. L_1 = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2}$$
$$c_1 = \frac{1}{2}, c_2 = \frac{1}{2}, c_3 = -\frac{1}{2}, c_4 = -\frac{1}{2}$$

$$2. L_2 = \frac{\mu_1 + \mu_3}{2} - \frac{\mu_2 + \mu_4}{2}$$
$$c_1 = \frac{1}{2}, c_2 = -\frac{1}{2}, c_3 = \frac{1}{2}, c_4 = -\frac{1}{2}$$

$$\widehat{L}_1 = \sum_{i=1}^r c_i \bar{Y}_i = -9.35$$

$$s^2\{L_1\} = 2.242$$

$$B = t\left(1 - \frac{\alpha}{2g}; n_T - r\right) = \textcolor{red}{2.84}$$

$$\widehat{L}_2 = \sum_{i=1}^r c_i \bar{Y}_i = -3.25$$

$$s^2\{L_2\} = 2.242$$

The simultaneous CI for L_1 : $(\textcolor{red}{-13.6}, -5.1)$ L_2 : $(\textcolor{red}{-7.5}, \textcolor{blue}{1})$

```
mc2<-matrix(c(0.5,0.5,-0.5,-0.5, 0.5, -0.5, 0.5, -0.5),2,4, byrow=TRUE)
oneway(cereal$y, cereal$design, mc=mc2)
```

The procedure of diagnostic and remedial measures in ANOVA is like regular regression model

- Non-constancy of error variance
- Non-independence of error terms
- Outliers
- Omission of important predictors
- Non-normality of error terms

The `oneway()` function in ALSM package serves multiple purpose for single factor ANOVA

- Fitting of ANOVA model
- ANOVA table
- Test and confidence interval for single factor level mean
- Inferences for difference between two factor level means
- Contrast of factor level means
- ANOVA diagnostic
- Nonparametric Rank F test
- Plots for exploration and residuals

Usage `oneway(y, group, alpha = 0.05, c.vallue = 0, mc = NULL)`

Arguments `y: vector`

`group: vector, factor`

`alpha: 0.05 by default`

`c.value: single factor test: Ho: $\mu_i = c, 0$ by defult`

`mc: matric contrast`

Example

1. Find the test statistic and p value for a hypothesis test $H_0: \mu_1 = \mu_3$, $H_a: \mu_1 \neq \mu_3$

2. Find the test statistic and p value for a hypothesis test $H_0: \mu_2 = \mu_3$, $H_a: \mu_2 \neq \mu_3$

3. Find the test statistic and p value for a hypothesis test $H_0: L = 0$, $H_a: L \neq 0$ where $L = \mu_1 - \frac{\mu_1 + \mu_2 + \mu_3 + \mu_4}{4}$

4. Find the simultaneous confidence interval for $(\mu_1 - \mu_3)$, $(\mu_2 - \mu_3)$

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|----------------|--|------------|---------|--------------|
| (Intercept) | 14.600 | 1.452 | 10.053 | 4.66e-08 *** |
| Designdesign2 | -1.200 | 2.054 | -0.584 | 0.5677 |
| Designdesign3 | 4.900 | 2.179 | 2.249 | 0.0399 * |
| Designdesign4 | 12.600 | 2.054 | 6.135 | 1.91e-05 *** |
| --- | | | | |
| Signif. codes: | 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 | | | |

Residual standard error: 3.248 on 15 degrees of freedom

Multiple R-squared: 0.7881, Adjusted R-squared: 0.7457

F-statistic: 18.59 on 3 and 15 DF, p-value: 2.585e-05

Two-way ANOVA

Factor Level Means Study

Main effects and Interaction effects

- Cells defined by combinations of two or more **discrete factors**
- Allows effects to be decomposed into *main effects* and *interactions*
- Model assumptions remain unchanged

Cell means notation for two-way ANOVA

For $i = 1, \dots, a$ levels in Factor A and $j = 1, \dots, b$ levels in Factor B ,
there are $k = 1, \dots, n_{i,j}$ individual observations in cell (i, j) .

Cell means model: $Y_{i,j,k} = \mu_{i,j} + \varepsilon_{i,j,k}$

- $\mu_{i,j}$ is the expected value (true mean) of cell (i, j) , estimated by $\bar{Y}_{i,j}$
- $\varepsilon \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$

There are $ab + 1$ parameters in this model (including σ^2)

Main effects and Interaction effects

A *main effect* describes the difference between a baseline reference (μ) and the marginal mean for a factor level ($\mu_{i\cdot}$ or $\mu_{\cdot j}$).

The *marginal mean* is the average value of the response across all data points that belong to a particular level of a factor.

An *interaction effect* gives the difference between the mean for a particular cell ($\mu_{i,j}$) and the sum of the baseline and main effects for belonging to level i of factor A and level j of factor B .

Cell Means and Marginal Means in Two-way ANOVA

$$\begin{array}{c} \text{Factor B} \\ \begin{array}{cccc} 1 & 2 & 3 \\ \hline \end{array} \\ \text{Factor A} \quad \begin{array}{ccccc} 1 & \mu_{1,1} & \mu_{1,2} & \mu_{1,3} & \mu_{1\cdot} \\ 2 & \mu_{2,1} & \mu_{2,2} & \mu_{2,3} & \mu_{2\cdot} \\ 3 & \mu_{3,1} & \mu_{3,2} & \mu_{3,3} & \mu_{3\cdot} \\ \hline \mu_{\cdot 1} & \mu_{\cdot 2} & \mu_{\cdot 3} & \mu_{..} & \end{array} \end{array} \quad \left. \begin{array}{l} \mu_{i\cdot} = \frac{1}{n_A} \sum_{j=1}^3 (n_{i,j} \mu_{i,j}) \\ \mu_{..} = \frac{1}{n_T} \sum_{i=1}^3 \sum_{j=1}^3 (n_{i,j} \mu_{i,j}) \end{array} \right\}$$

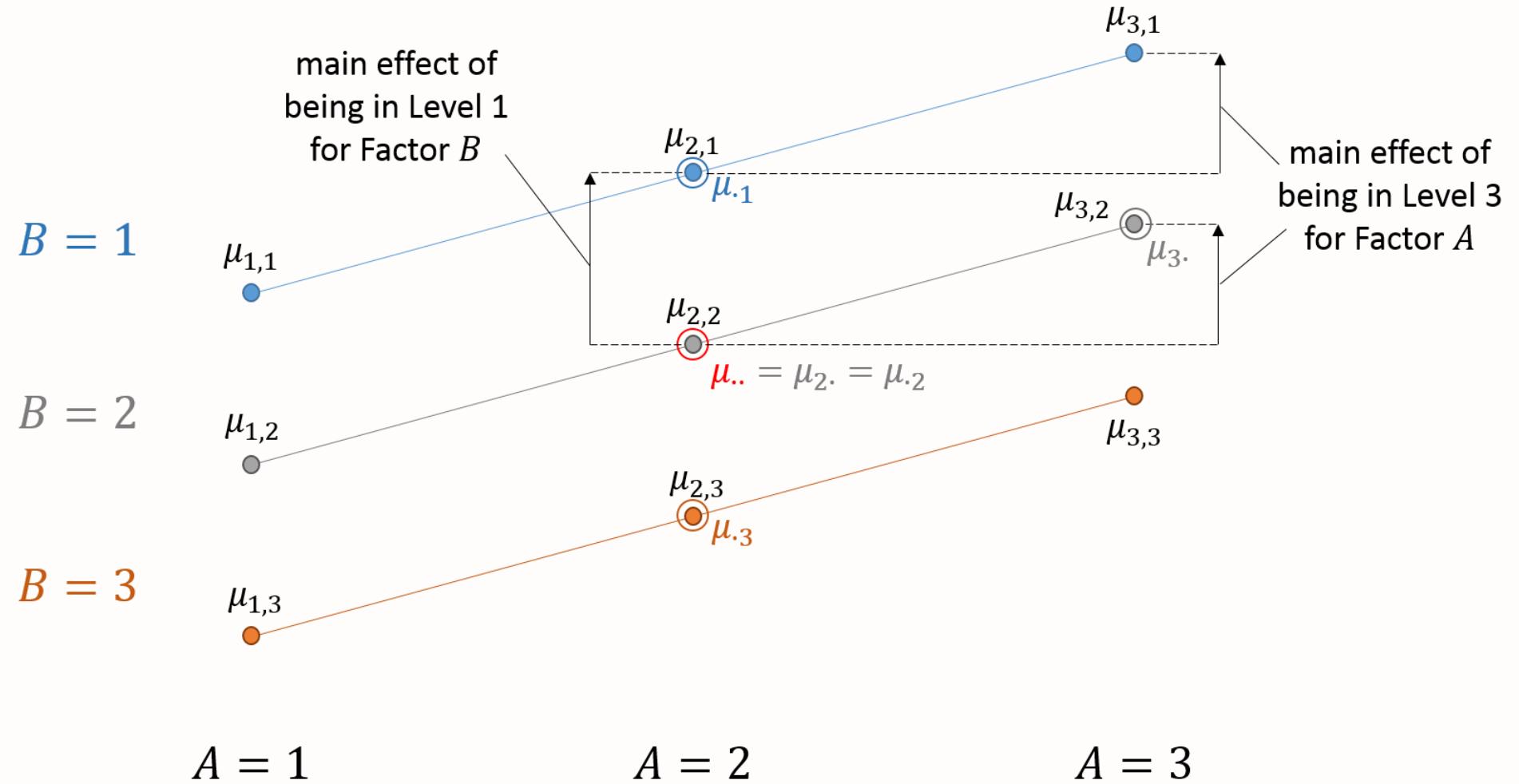
Factor effects notation for two-way ANOVA

Factor effects model: $Y_{i,j,k} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{i,j} + \varepsilon_{i,j,k}$

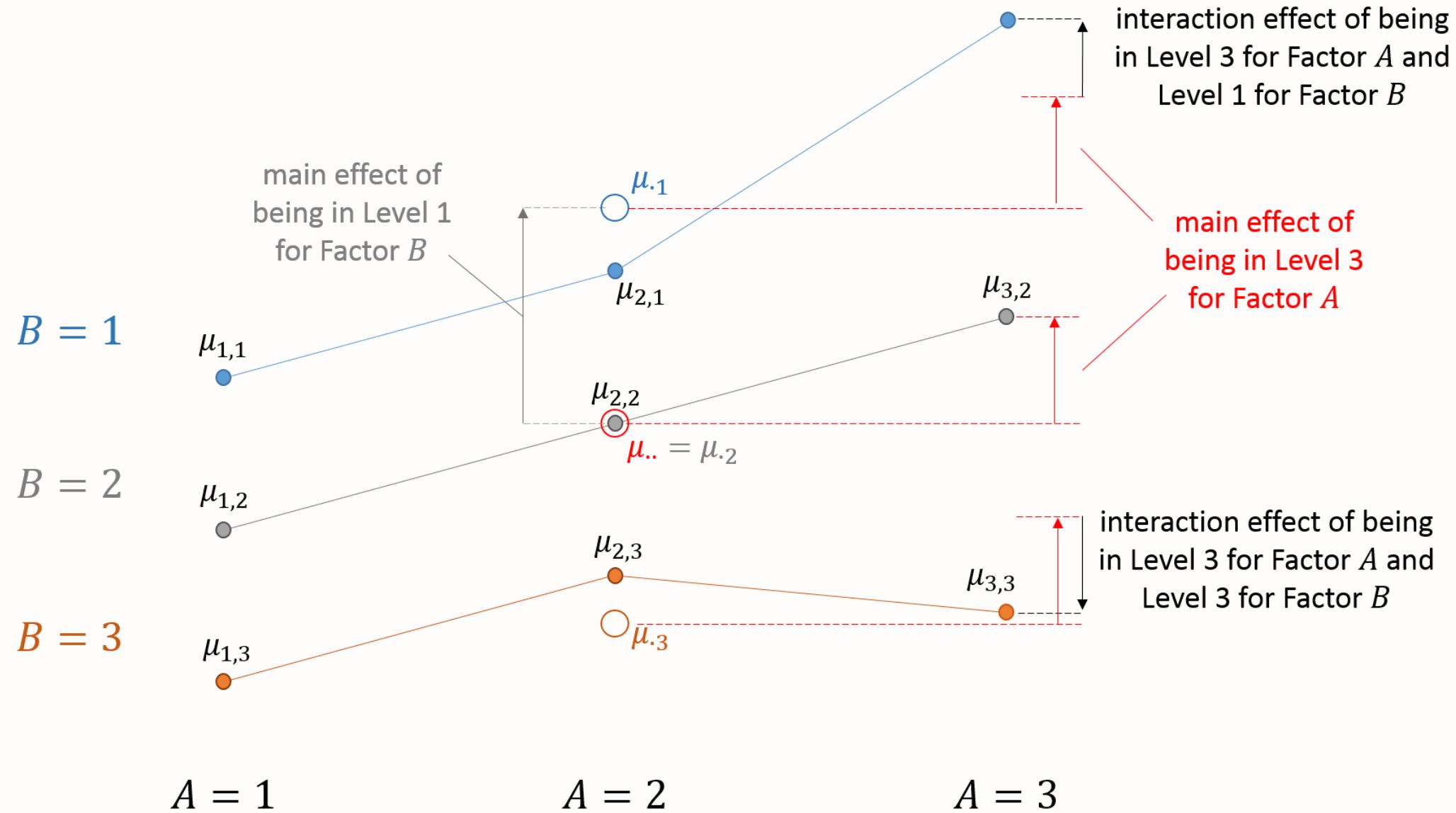
where,

- μ is grand mean, estimated by $\bar{Y}_{...}$
- α_i is the main effect of belonging to level i of factor A , estimated by $\bar{Y}_{i..} - \bar{Y}_{...}$
- β_j is the main effect of belonging to level j of factor B , estimated by $\bar{Y}_{.j} - \bar{Y}_{...}$
- $(\alpha\beta)_{i,j}$ is the interaction effect of belonging to both i and j , estimated by $\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j} + \bar{Y}_{...}$

Note that “ $(\alpha\beta)_{i,j}$ ” is ONE parameter, NOT a product!



Two-way ANOVA with no interactions: $\mu_{i,j} = \mu + \alpha_i + \beta_j$



Two-way ANOVA with interactions: $\mu_{i,j} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{i,j}$

Development of two-way ANOVA Model

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \text{ where } i = 1 \text{ to } a, j = 1 \text{ to } b$$

For two variables with a and b levels respectively, we need to define $(a - 1)$ and $(b - 1)$ dummy variables for main effect, and $(a - 1)(b - 1)$ dummy variables for interaction.

For example, $a = 3, b = 2$, the design matrix X is

| | X1 | X2 | | X3 |
|---------------------|----|----|---------------------|----|
| level 1 in factor A | 1 | 0 | level 1 in factor B | 1 |
| level 2 in factor A | 0 | 1 | level 2 in factor B | -1 |
| level 3 in factor A | -1 | -1 | | |

Main effect

Note: the reference baseline ($\mu_{..}$): unweighted mean

| | X1X3 | X2X3 |
|---|------|------|
| level 1 in factor A and level 1 in factor B | 1 | 0 |
| level 1 in factor A and level 2 in factor B | -1 | 0 |
| level 2 in factor A and level 1 in factor B | 0 | 1 |
| level 2 in factor A and level 2 in factor B | 0 | -1 |
| level 3 in factor A and level 1 in factor B | -1 | -1 |
| level 3 in factor A and level 2 in factor B | 1 | 1 |

Interaction effect

Constraints in two-way ANOVA

Equivalent of Constraint C ($\sum_{i=1}^r \tau_i = 0$) in one-way ANOVA:

$$\sum_{i=1}^a \alpha_i = 0 \quad \sum_{j=1}^b \beta_j = 0 \quad \sum_{i=1}^a (\alpha\beta)_{i,j} = 0 \quad \forall j \quad \sum_{j=1}^b (\alpha\beta)_{i,j} = 0 \quad \forall i$$

- μ is the *grand mean* of the population ($\mu_{..}$, estimated by $\bar{Y}_{..}$)
- $\mu + \alpha_i$ is the *marginal mean* for level i of Factor A ($\mu_{i..}$, estimated by $\bar{Y}_{i..}$)
- $\mu + \beta_j$ is the marginal mean for level j of Factor B ($\mu_{..j}$, estimated by $\bar{Y}_{..j}$)
- $\mu + \alpha_i + \beta_j + (\alpha\beta)_{i,j}$ is the *cell mean* ($\mu_{i,j}$, estimated by $\bar{Y}_{i,j}$)
(in a purely additive model, the cell mean would be $\mu + \alpha_i + \beta_j$)

Constraint $(\alpha\beta)_{a,b} = 0$ appears twice, so this is
a total of $1 + 1 + a + b - 1 = 1 + a + b$ constraints.

Development of the Regression Model

Example: Bread sales In this example, we use data from a designed experiment to determine how the height and width of a display shelf affects bread sales at a bakery. Twelve supermarkets, similar in sales volume and clientele were studied (bakery.txt).

$$a = 3, b = 2, n = 2, n_T = 12$$

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \text{ where } i = 1 \text{ to } 3, j = 1 \text{ to } 2$$

| Factor A (height) | Factor B (width) | | row total | height average |
|-------------------|------------------|----------|-----------|----------------|
| | B1 (regular) | B2(wide) | | |
| A1 (bottom) | 47 | 46 | | |
| | 43 | 40 | | |
| Total | 90 | 86 | 176 | |
| average | 45 | 43 | | 44 |
| A2 (middle) | 62 | 67 | | |
| | 68 | 71 | | |
| Total | 130 | 138 | 268 | |
| average | 65 | 69 | | 67 |
| A3 (top) | 41 | 42 | | |
| | 39 | 46 | | |
| Total | 80 | 88 | 168 | |
| average | 40 | 44 | | 42 |
| Column total | 300 | 312 | 612 | |
| width average | 50 | 52 | | 51 |

| y | a (weight) | b(height) | Int. | x1 | x2 | x3 | x1x3 | x2x3 |
|----|------------|-----------|------|----|----|----|------|------|
| 47 | 1 | | 1 | 1 | 1 | 0 | 1 | 1 |
| 43 | 1 | | 1 | 1 | 1 | 0 | 1 | 1 |
| 46 | 1 | | 2 | 1 | 1 | 0 | -1 | 0 |
| 40 | 1 | | 2 | 1 | 1 | 0 | -1 | 0 |
| 62 | 2 | | 1 | 1 | 0 | 1 | 1 | 0 |
| 68 | 2 | | 1 | 1 | 0 | 1 | 1 | 0 |
| 67 | 2 | | 2 | 1 | 0 | 1 | -1 | 0 |
| 71 | 2 | | 2 | 1 | 0 | 1 | -1 | 0 |
| 41 | 3 | | 1 | 1 | -1 | -1 | 1 | -1 |
| 39 | 3 | | 1 | 1 | -1 | -1 | 1 | -1 |
| 42 | 3 | | 2 | 1 | -1 | -1 | -1 | 1 |
| 46 | 3 | | 2 | 1 | -1 | -1 | -1 | 1 |

Design matrix

Development of the Regression Model

Example: Bread sales In this example, we use data from a designed experiment to determine how the height and width of a display shelf affects bread sales at a bakery. Twelve supermarkets, similar in sales volume and clientele were studied (bakery.txt).

| | Factor B (width) | | row total | height average |
|-------------------|------------------|----------|-----------|----------------|
| Factor A (height) | B1 (regular) | B2(wide) | | |
| A1 (bottom) | 47 | 46 | | |
| | 43 | 40 | | |
| Total | 90 | 86 | 176 | |
| average | 45 | 43 | | 44 |
| A2 (middle) | 62 | 67 | | |
| | 68 | 71 | | |
| Total | 130 | 138 | 268 | |
| average | 65 | 69 | | 67 |
| A3 (top) | 41 | 42 | | |
| | 39 | 46 | | |
| Total | 80 | 88 | 168 | |
| average | 40 | 44 | | 42 |
| Column total | 300 | 312 | 612 | |
| width average | 50 | 52 | | 51 |

```
summary(lm(y~height*width, bakery))
```

Coefficients:

| | Estimate | std. Error | t value | Pr(> t) |
|----------------|----------|------------|---------|--------------|
| (Intercept) | 51.000 | 0.928 | 54.959 | 2.44e-09 *** |
| height1 | -7.000 | 1.312 | -5.334 | 0.00177 ** |
| height2 | 16.000 | 1.312 | 12.192 | 1.85e-05 *** |
| width1 | -1.000 | 0.928 | -1.078 | 0.32261 |
| height1:width1 | 2.000 | 1.312 | 1.524 | 0.17835 |
| height2:width1 | -1.000 | 1.312 | -0.762 | 0.47494 |

$$\mu_{..} = 51$$

$$\alpha_1 = \mu_{1.} - \mu_{..} = -7$$

$$\alpha_2 = \mu_{2.} - \mu_{..} = 16$$

$$\alpha_3 = -\alpha_1 - \alpha_2 =$$

$$(\alpha\beta)_{11} = \mu_{11} - \mu_{1.} - \mu_{.1} + \mu_{..} = 2$$

$$\beta_1 = \mu_{.1} - \mu_{..} = -1$$

$$\beta_2 = -\beta_1 =$$

$$(\alpha\beta)_{21} = \mu_{11} - \mu_{1.} - \mu_{.1} + \mu_{..} = -1$$

$$\hat{Y}_{11} = \mu_{..} + \alpha_1 + \beta_1 + (\alpha\beta)_{11} = 55 - 7 - 1 + 2 = 44$$

$$(\alpha\beta)_{31} = -(\alpha\beta)_{11} - (\alpha\beta)_{21} =$$

$$(\alpha\beta)_{12} = -(\alpha\beta)_{11} =$$

$$(\alpha\beta)_{22} = -(\alpha\beta)_{21} =$$

$$(\alpha\beta)_{32} = -(\alpha\beta)_{31} =$$

Building the analysis of variance table in two-way ANOVA

In regression and one-way ANOVA, we broke the total sum of squares down into the model sum of squares (SSM) and error sum of squares (SSE).

In two-way ANOVA, SSM is further broken down into the main and interaction effects.

(This is really just an application of the extra sum of squares)

Rules for degrees of freedom

Degrees of freedom in the two-way ANOVA analysis are allocated as follows:

- Main effects for each factor take $r - 1$ df, where r is the number of levels in the factor
i.e., $df_A = (a - 1)$ and $df_B = (b - 1)$
- Interactions take df's equal to the product of the main effect df's:
 $(a - 1)(b - 1)$ for the interaction between factors A and B .
- Total sum of squares: $df_T = n_T - 1$ (as usual)
- Model df's are given by the sum of the df's for all main and interactions in the model:
$$df_M = a + b - 2 + (a - 1)(b - 1)$$
- Error: $df_E = df_T - df_M$ (as usual)

F -tests

Two-way ANOVA adds several secondary F -tests to the standard global F -test.

- In *fixed effects* models, all of the F -tests use MSE in the denominator.
- The numerators for the secondary F -tests may use either the Type I (sequential) or Type II (last-variable-added) extra sums of squares.

ANOVA Table

| Source | df | SS | MS | F |
|----------------|------------------|--------|----------------------------------|--------------------|
| Factor A | $a - 1$ | SSA | $MSA = \frac{SSA}{a-1}$ | $\frac{MSA}{MSE}$ |
| Factor B | $b - 1$ | SSB | $MSB = \frac{SSB}{b-1}$ | $\frac{MSB}{MSE}$ |
| Interaction AB | $(a - 1)(b - 1)$ | $SSAB$ | $MSAB = \frac{SSAB}{(a-1)(b-1)}$ | $\frac{MSAB}{MSE}$ |
| Error | $ab(n - 1)$ | SSE | $MSE = \frac{SSE}{ab(n-1)}$ | |
| Total | $n_T - 1$ | SST | $MSA = \frac{SST}{n_T-1}$ | |

Expected mean squares

With the zero-sum constraints and a balanced design (so $n_{i,j} = n \forall i, j$):

$$\mathbb{E}(MSE) = \sigma^2$$

$$\mathbb{E}(MSA) = \sigma^2 + \frac{nb}{(a-1)} \sum_i \alpha_i^2$$

$$\mathbb{E}(MSB) = \sigma^2 + \frac{na}{(b-1)} \sum_j \beta_j^2$$

$$\mathbb{E}(MSAB) = \sigma^2 + \frac{n}{(a-1)(b-1)} \sum_{i,j} (\alpha\beta)_{i,j}^2$$

Analytical strategy

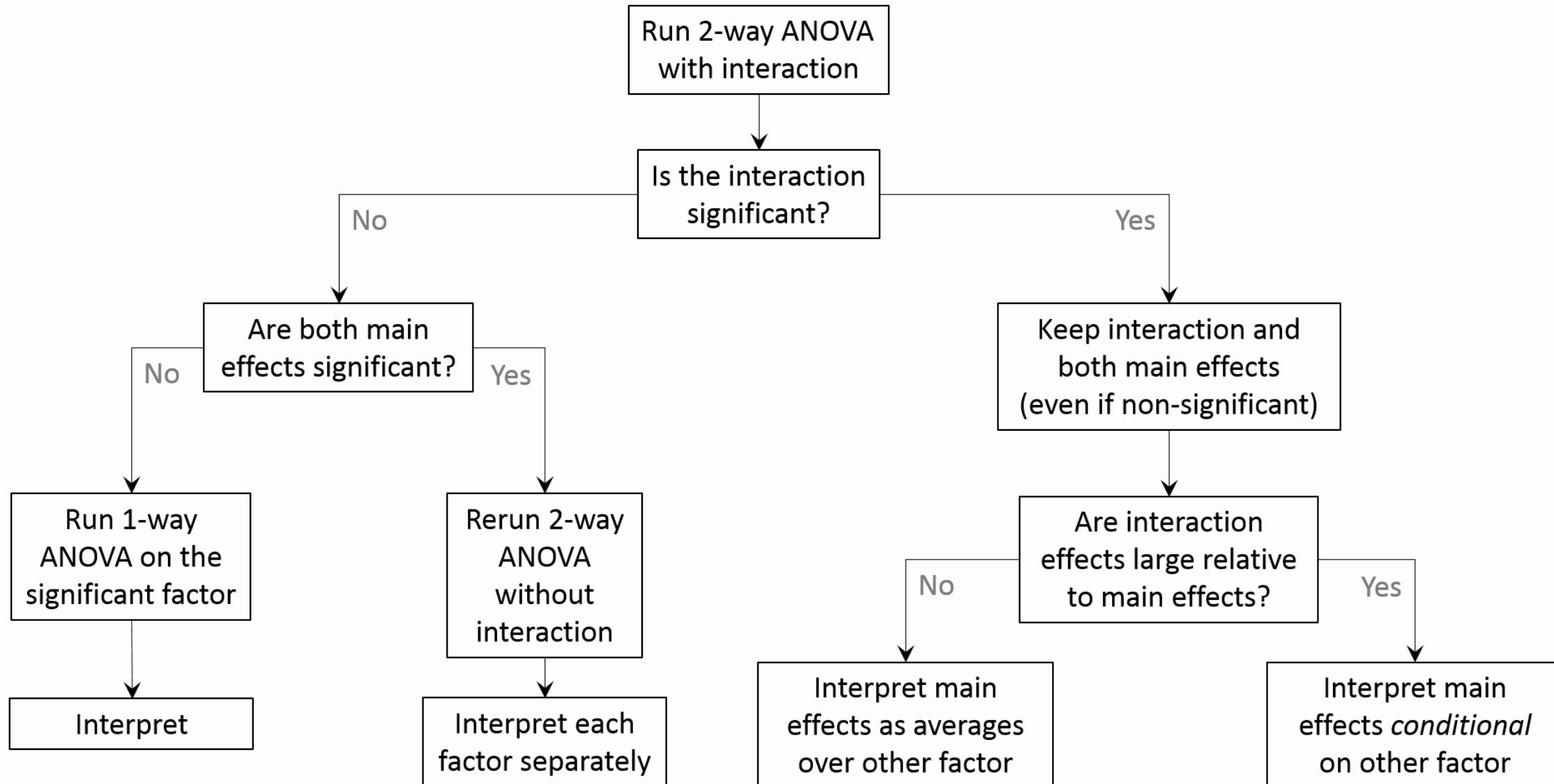
If the model contains interaction terms that are significantly different from zero, then the relationship between each factor and the response is not consistent.

It depends on the level of the other factor.

Always check for an interaction first.

- If the interaction is *not* significant, you can remove it.
- If the interaction term *is* significant, *leave the main effects in the model*, even if they are not significant.

Analytic strategy for two-way ANOVA



Compare main factor effects

Estimation of factor level mean

$$\begin{aligned}\hat{\mu}_{i..} &= \bar{Y}_{i..} & \sigma^2\{\bar{Y}_{i..}\} &= \frac{\sigma^2}{bn} & s^2\{\bar{Y}_{i..}\} &= \frac{MSE}{bn} \\ \hat{\mu}_{.j} &= \bar{Y}_{.j} & \sigma^2\{\bar{Y}_{.j}\} &= \frac{\sigma^2}{an} & s^2\{\bar{Y}_{.j}\} &= \frac{MSE}{an}\end{aligned}$$

Confidence interval for $\mu_{i..}$ and $\mu_{.j}$.

$$\bar{Y}_{i..} \pm t[1 - \alpha/2; (n - 1)ab]s\{\bar{Y}_{i..}\}$$

$$\bar{Y}_{.j} \pm t[1 - \alpha/2; (n - 1)ab]s\{\bar{Y}_{.j}\}$$

Estimation of contrast (or just general linear combination without the constriction) of factor level means

For factor A means μ_i .

$$L = \sum c_i \mu_i. \quad \text{where } \sum c_i = 0 \quad \hat{L} = \sum c_i \bar{Y}_{i..}$$

$$\sigma^2\{\hat{L}\} = \sum c_i^2 \sigma^2\{\bar{Y}_{i..}\} = \frac{\sigma^2}{bn} \sum c_i^2 \quad s^2\{\hat{L}\} = \frac{MSE}{bn} \sum c_i^2$$

For factor B means μ_j

$$L = \sum c_j \mu_{.j} \quad \text{where } \sum c_j = 0 \quad \hat{L} = \sum c_j \bar{Y}_{.j}$$

$$s^2\{\hat{L}\} = \frac{MSE}{an} \sum c_j^2$$

Finally, the appropriate $1 - \alpha$ confidence limits for L are:

$$\hat{L} \pm t[1 - \alpha/2; (n - 1)ab]s\{\hat{L}\}$$

The test statistic is $\frac{L}{s\{\hat{L}\}} \sim t((n - 1)ab)$

Compare main factor effects

Bonferroni procedure comparison of factor level means

Compare g groups in factor A,
each being D

$$D = \mu_{i\cdot} - \mu_{i'\cdot}$$

$$\hat{D} = \bar{Y}_{i\cdot\cdot} - \bar{Y}_{i'\cdot\cdot}$$

$$s^2\{\hat{D}\} = \frac{2MSE}{bn}$$

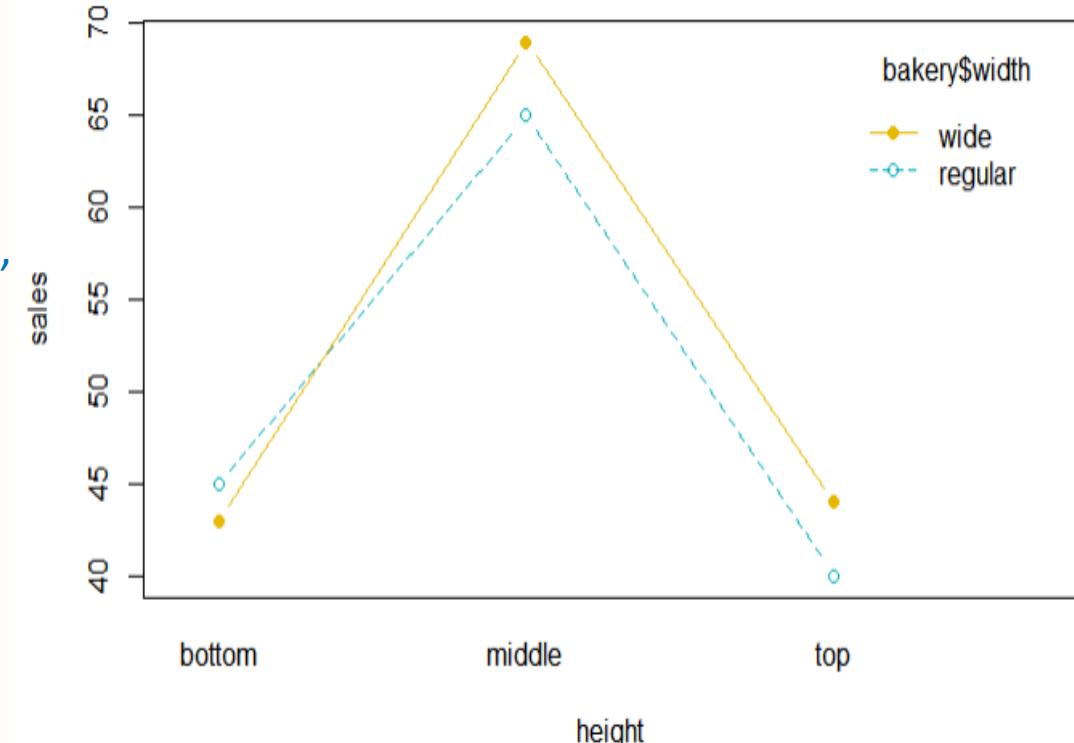
$$B = t[1 - \alpha/2g; (n - 1)ab]$$

The $1 - \alpha$ confidence interval for D are

$$\hat{D} \pm Bs\{\hat{D}\}$$

The test statistic are

$$t^* = \frac{\hat{D}}{s\{\hat{D}\}}; \quad \text{If } |t^*| > t[1 - \alpha/2g; (n - 1)ab], \text{ conclude } H_a$$



For example (main effect comparison)

Which of the following is the correct equation to compare average sale between regular and wide width.

- A) $\mu_{1\cdot} - \mu_{2\cdot}$
- B) $\mu_{\cdot 1} - \mu_{\cdot 2}$
- C) $\mu_{11} - \mu_{21}$
- D) $\mu_{11} - \mu_{12}$

There are also other procedures (Turkey, LSD, Sheaffe etc.) comparison of factor level means. Check out the text book for more details.

Compare interaction effects

Simultaneously compare multiple cell means

$$D = \mu_{ij} - \mu_{i'j'} = \hat{Y}_{ij\cdot} - \hat{Y}_{i'j'\cdot} \quad s^2\{\hat{D}\} = 2\left(\frac{MSE}{n}\right)$$

For example, $D_1 = \mu_{11} - \mu_{12}$

$D_2 = \mu_{22} - \mu_{21}$

$D_3 = \mu_{31} - \mu_{32}$

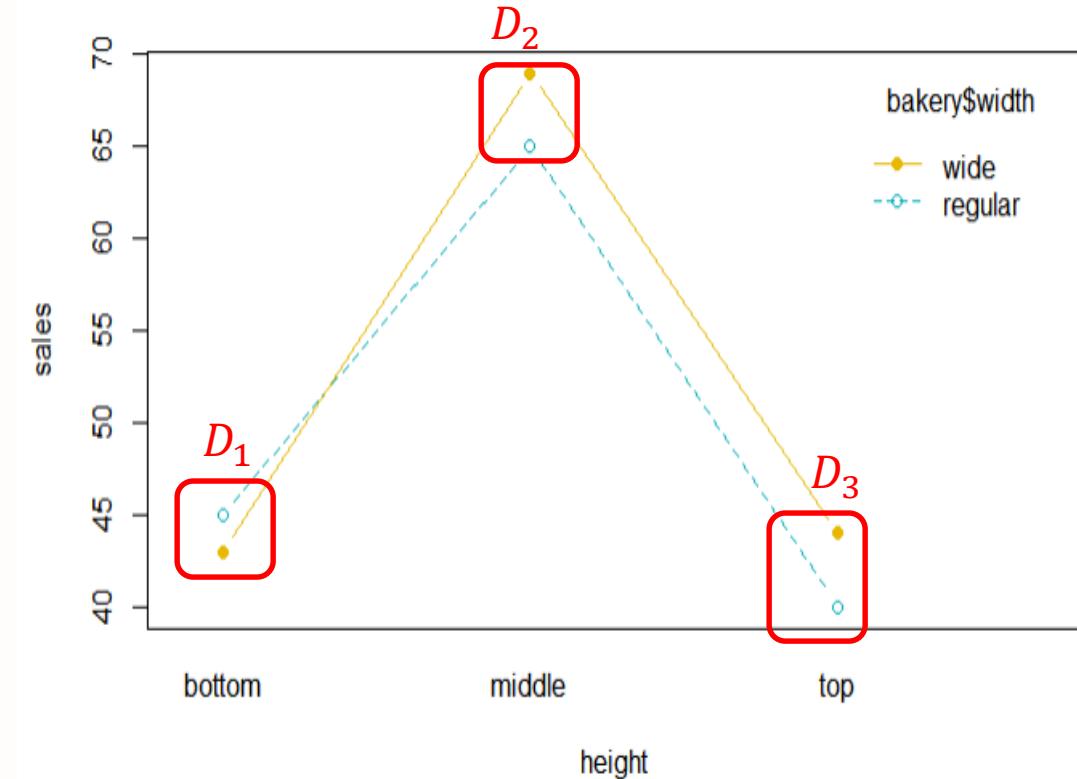
$$B = t[1 - \alpha/2g; (n - 1)ab]$$

The $1 - \alpha$ confidence interval for D are

$$\hat{D} \pm Bs\{\hat{D}\}$$

The test statistic are

$$t^* = \frac{\hat{D}}{s\{\hat{D}\}}; \quad \text{If } |t^*| > t[1 - \alpha/2g; (n - 1)ab], \text{ conclude } H_a$$



Comparison multiple cell means is necessary especially when the interaction effect is significant.

Example: Bread sales

In this example, we use data from a designed experiment to determine how the height and width of a display shelf affects bread sales at a bakery. Twelve supermarkets, similar in sales volume and clientele were studied (bakery.txt).

| | | Factor B (width) | | row total | height average |
|-------------------|-----|------------------|----------|-----------|----------------|
| Factor A (height) | | B1 (regular) | B2(wide) | | |
| A1 (bottom) | 47 | 46 | | | |
| | 43 | 40 | | | |
| | 90 | 86 | 176 | | |
| average | 45 | 43 | | 44 | |
| A2 (middle) | 62 | 67 | | | |
| | 68 | 71 | | | |
| | 130 | 138 | 268 | | |
| average | 65 | 69 | | 67 | |
| A3 (top) | 41 | 42 | | | |
| | 39 | 46 | | | |
| | 80 | 88 | 168 | | |
| average | 40 | 44 | | 42 | |
| Column total | 300 | 312 | 612 | | |
| width average | 50 | 52 | | 51 | |

```
anova(lm(y~height*width, bakery))
```

Analysis of Variance Table

Response: y

| | df | sum Sq | Mean Sq | F value | Pr (>F) |
|--------------|----|--------|---------|---------|---------------|
| height | 2 | 1544 | 772.00 | 74.7097 | 5.754e-05 *** |
| width | 1 | 12 | 12.00 | 1.1613 | 0.3226 |
| height:width | 2 | 24 | 12.00 | 1.1613 | 0.3747 |
| Residuals | 6 | 62 | 10.33 | | |

$H_0: \text{all } (\alpha\beta_{ij}) = 0, \quad H_a: \text{not all } (\alpha\beta_{ij}) = 0$

Not significant, p-value=0.3747

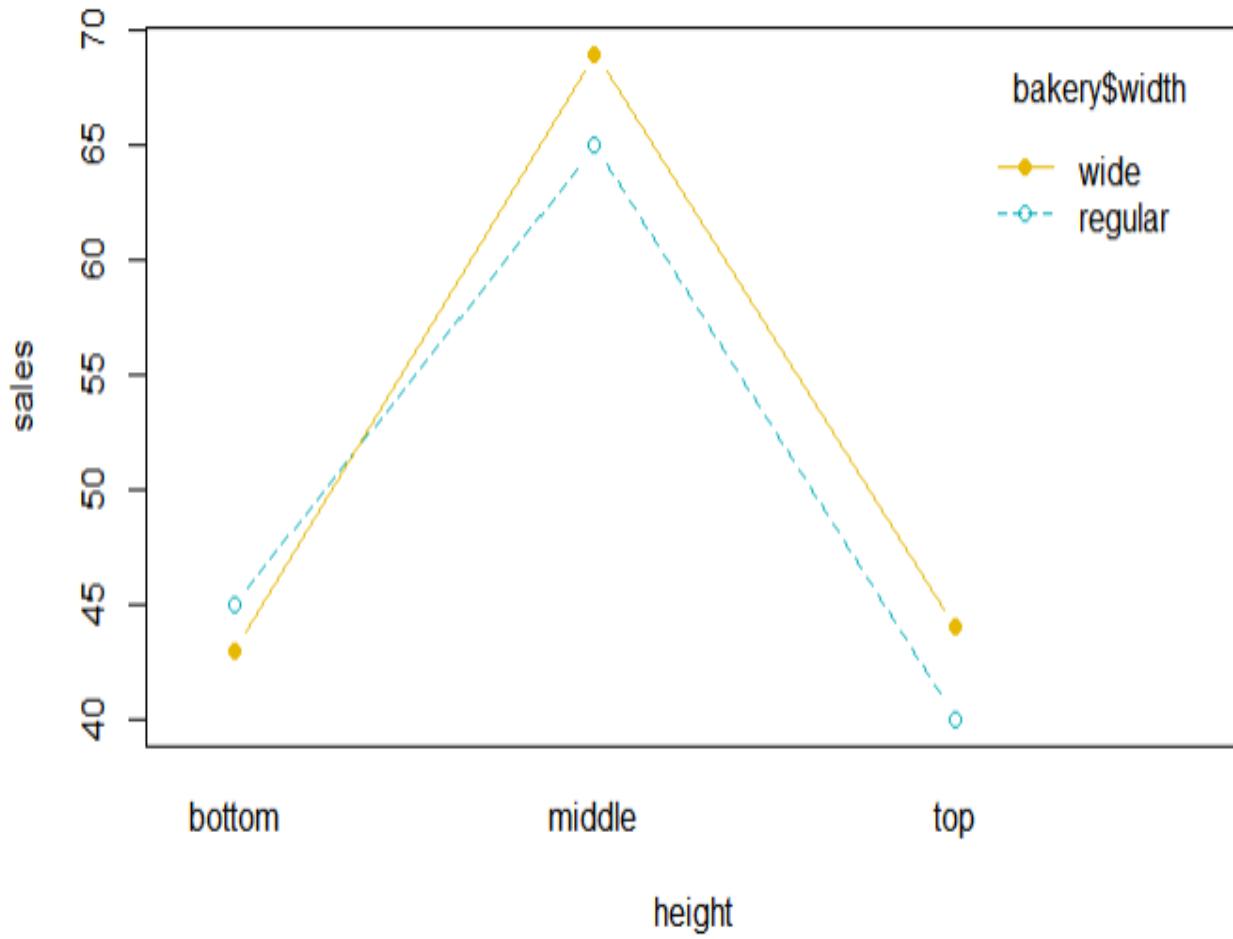
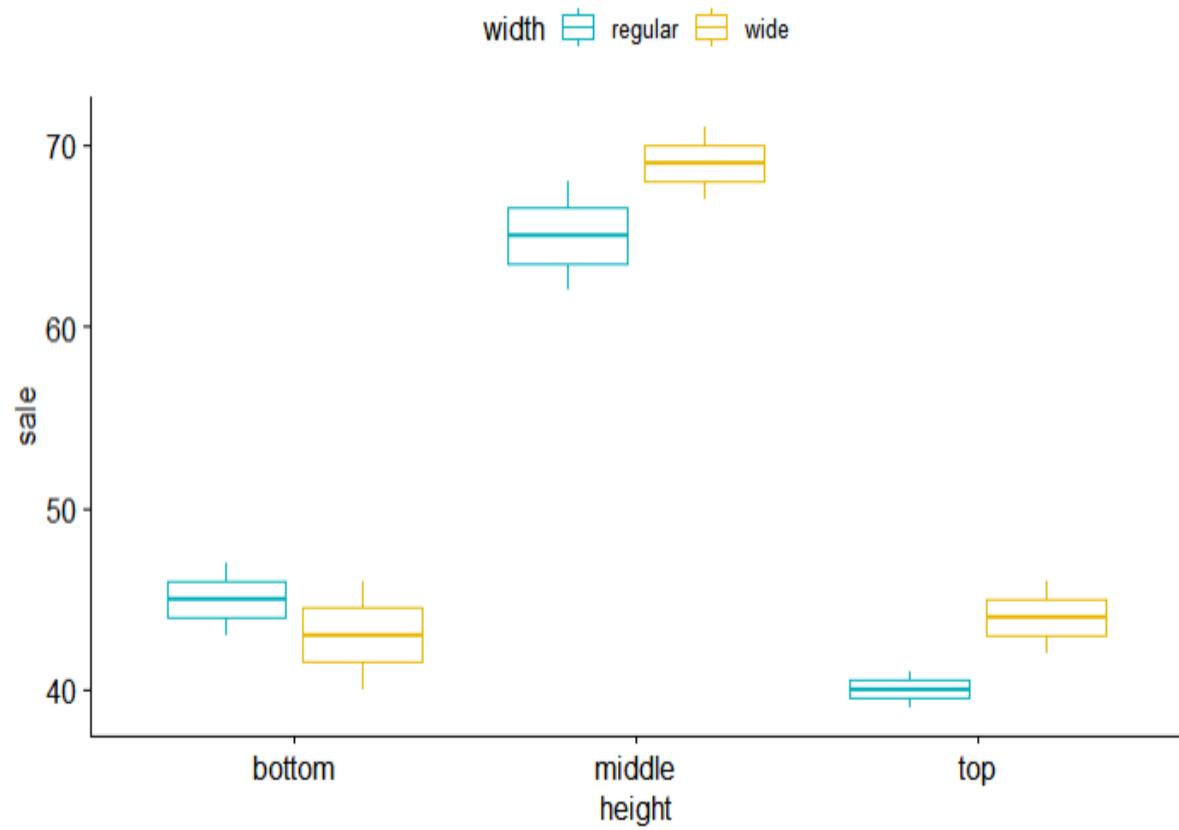
$H_0: \text{all } \alpha_i = 0, \quad H_a: \text{not all } (\alpha_i) = 0$

Significant, p-value<0.0001

$H_0: \text{all } \beta_i = 0, \quad H_a: \text{not all } (\beta_i) = 0$

Not significant, p-value=0.3226

Example: Bread sales



1. Check the interaction term: not significant ($F_{2,6} = 1.16, p = 0.3747$)
2. Since the interaction is not significant, we can interpret the main effects independently of each other.
3. Main effect of height is significant ($F_{2,6} = 74.71, p < 0.0001$)
4. Main effect of width is not significant ($F_{1,6} = 1.16, p = 0.3226$)

Example: Bread sales

The height of the display affects sales, and has a similar effect at both widths.

Width has no effect on sales.

Further analyses are needed to tell which levels of height differ from the others:

1. Rerun the analysis as a one-way ANOVA on height
2. Compare the individual pairwise differences between levels
3. Look at a plot or at the cell means to see which height(s) maximize sales.

Example: Bread sales

Since the interaction effect is not significant. We can do comparison based on the marginal means (main effect)

1. Compare the average sale between bottom and middle height

$$D1 = \mu_1 - \mu_2.$$

2. Compare the average sale between regular and wide width

$$D2 = \mu_{.1} - \mu_{.2}$$

3. Compare the average sale between the average of middle and top regular (21 and 31) and middle and top wide (22 and 32) width

$$D3 = \frac{(\mu_{21} + \mu_{31})}{2} - \frac{(\mu_{22} + \mu_{32})}{2}$$

4. Is the average sale the highest in the middle height? Consider a simultaneous comparison with Bonferroni procedure at 95% level.

| Factor A
(display height, i) | Factor B
(display width, j) | Mean | n |
|---------------------------------|--------------------------------|------|---|
| i=1 bottom | j=1 (regular) | 45 | 2 |
| i=2 middle | j=1 (regular) | 65 | 2 |
| i=3 top | j=1 (regular) | 40 | 2 |
| i=1 bottom | j=2 (wide) | 43 | 2 |
| i=2 middle | j=2 (wide) | 69 | 2 |
| i=3 top | j=2 (wide) | 44 | 2 |
| MSE=10.3 a=3, b=2 | | | |

Example: Bread sales

Since the interaction effect is not significant. We can do comparison based on the marginal means (main effect)

1. Compare the average sale between bottom and middle height

$$D_1 = \mu_{1\cdot} - \mu_{2\cdot}$$

$$\hat{D} = \hat{Y}_{1\cdot} - \hat{Y}_{2\cdot} = (\hat{Y}_{11} + \hat{Y}_{12})/2 - (\hat{Y}_{21} + \hat{Y}_{22})/2 = -23$$

$$s^2\{\hat{D}\} = \frac{2MSE}{bn} = \frac{2(10.3)}{2(2)} = 5.15, \quad \text{so } s\{\hat{D}\} = 2.27$$

$$t_s = \frac{\hat{D}}{s\{\hat{D}\}} = 10.1$$

| Factor A
(display height, i) | Factor B
(display width, j) | Mean | n |
|---------------------------------|--------------------------------|------|---|
| i=1 bottom | j=1 (regular) | 45 | 2 |
| i=2 middle | j=1 (regular) | 65 | 2 |
| i=3 top | j=1 (regular) | 40 | 2 |
| i=1 bottom | j=2 (wide) | 43 | 2 |
| i=2 middle | j=2 (wide) | 69 | 2 |
| i=3 top | j=2 (wide) | 44 | 2 |
| MSE=10.3 a=3, b=2 | | | |

```
library(gmodels)
bm<-lm(y~height*width+0, data=bakery)
summary(bm)
anova(bm)
```

coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|----------------|----------|------------|---------|--------------|
| height1:width1 | 45.000 | 2.273 | 19.80 | 1.08e-06 *** |
| height2:width1 | 65.000 | 2.273 | 28.60 | 1.21e-07 *** |
| height3:width1 | 40.000 | 2.273 | 17.60 | 2.16e-06 *** |
| height1:width2 | 43.000 | 2.273 | 18.92 | 1.41e-06 *** |
| height2:width2 | 69.000 | 2.273 | 30.36 | 8.48e-08 *** |
| height3:width2 | 44.000 | 2.273 | 19.36 | 1.23e-06 *** |

Response: y

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|--------------|----|--------|---------|---------|---------------|
| height:width | 6 | 32792 | 5465.3 | 528.9 | 6.702e-08 *** |
| Residuals | 6 | 62 | 10.3 | | |

To understand the cm setting

$$\begin{aligned}
 D_1 &= \mu_{1\cdot} - \mu_{2\cdot} = (\mu_{11} + \mu_{12})/2 - (\mu_{21} + \mu_{22})/2 \\
 &= \frac{\mu_{11}}{2} + \frac{\mu_{12}}{2} - \frac{\mu_{21}}{2} - \frac{\mu_{22}}{2} \\
 &= \frac{1}{2}(\mu_{11}) - \frac{1}{2}(\mu_{21}) + 0(\mu_{31}) + \frac{1}{2}(\mu_{12}) - \frac{1}{2}(\mu_{22}) + 0(\mu_{22})
 \end{aligned}$$

```
cm<-c(1/2, -1/2, 0, 1/2, -1/2, 0)
estimable(bm, cm)
```

| Estimate
<dbl> | Std. Error
<dbl> | t value
<dbl> | DF
<dbl> | Pr(> t)
<dbl> |
|-------------------|---------------------|------------------|-------------|-------------------|
| -23 | 2.27303 | -10.11865 | 6 | 5.415009e-05 |

Example: Bread sales

Since the interaction effect is not significant. We can do comparison based on the marginal means (main effect)

2. Compare the average sale between regular and wide width

$$D_2 = \mu_{.1} - \mu_{.2}$$

$$\hat{D} = \hat{Y}_{.1} - \hat{Y}_{.2} = (\hat{Y}_{11} + \hat{Y}_{21} + \hat{Y}_{31})/3 - (\hat{Y}_{12} + \hat{Y}_{22} + \hat{Y}_{32})/3 = -2$$

$$s^2\{\hat{D}\} = \frac{2MSE}{an} = \frac{2(10.3)}{3(2)} = 3.43, \quad \text{so } s\{\hat{D}\} = 1.86$$

$$t_s = \frac{\hat{D}}{s\{\hat{D}\}} = \frac{-2}{1.86} = 1.08$$

| Factor A
(display height, i) | Factor B
(display width, j) | Mean | n |
|---------------------------------|--------------------------------|------|---|
| i=1 bottom | j=1 (regular) | 45 | 2 |
| i=2 middle | j=1 (regular) | 65 | 2 |
| i=3 top | j=1 (regular) | 40 | 2 |
| i=1 bottom | j=2 (wide) | 43 | 2 |
| i=2 middle | j=2 (wide) | 69 | 2 |
| i=3 top | j=2 (wide) | 44 | 2 |
| MSE=10.3 a=3, b=2 | | | |

```
library(gmodels)
bm<-lm(y~height*width+0, data=bakery)
summary(bm)
anova(bm)
```

coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|----------------|----------|------------|---------|--------------|
| height1:width1 | 45.000 | 2.273 | 19.80 | 1.08e-06 *** |
| height2:width1 | 65.000 | 2.273 | 28.60 | 1.21e-07 *** |
| height3:width1 | 40.000 | 2.273 | 17.60 | 2.16e-06 *** |
| height1:width2 | 43.000 | 2.273 | 18.92 | 1.41e-06 *** |
| height2:width2 | 69.000 | 2.273 | 30.36 | 8.48e-08 *** |
| height3:width2 | 44.000 | 2.273 | 19.36 | 1.23e-06 *** |

Response: y

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|--------------|----|--------|---------|---------|---------------|
| height:width | 6 | 32792 | 5465.3 | 528.9 | 6.702e-08 *** |
| Residuals | 6 | 62 | 10.3 | | |

```
cm<-c(1/3, 1/3, 1/3, -1/3, -1/3, -1/3)
estimable(bm, cm)
```

| Estimate
(β_6) | Std. Error
(σ_{β_6}) | t value
(t_{β_6}) | DF
($n-2$) | Pr(> t)
($P(t > t_{\beta_6})$) |
|------------------------------------|---|---------------------------------------|--------------------------|---|
| -2 | 1.855921 | -1.077632 | 6 | 0.3226055 |

Example: Bread sales

3. Compare the average sale between the average of middle and top regular (21 and 31) and middle and top wide (22 and 32) width

$$D_3 = \frac{(\mu_{21} + \mu_{31})}{2} - \frac{(\mu_{22} + \mu_{32})}{2}$$

$$\hat{D} = \frac{(\hat{Y}_{21} + \hat{Y}_{31})}{2} - \frac{(\hat{Y}_{22} + \hat{Y}_{32})}{2} = -4$$

$$s^2\{\hat{D}\} = \frac{MSE}{n} \sum c_i^2 = \frac{(10.3)}{2} 1 = 5.15, \quad \text{so } s\{\hat{D}\} = 2.27$$

$$t_s = \frac{\hat{D}}{s\{\hat{D}\}} = -1.76$$

| Factor A
(display height, i) | Factor B
(display width, j) | Mean | n |
|---------------------------------|--------------------------------|------|---|
| i=1 bottom | j=1 (regular) | 45 | 2 |
| i=2 middle | j=1 (regular) | 65 | 2 |
| i=3 top | j=1 (regular) | 40 | 2 |
| i=1 bottom | j=2 (wide) | 43 | 2 |
| i=2 middle | j=2 (wide) | 69 | 2 |
| i=3 top | j=2 (wide) | 44 | 2 |
| MSE=10.3 a=3, b=2 | | | |

```
library(gmodels)
bm<-lm(y~height*width+0, data=bakery)
summary(bm)
anova(bm)
```

coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|----------------|----------|------------|---------|--------------|
| height1:width1 | 45.000 | 2.273 | 19.80 | 1.08e-06 *** |
| height2:width1 | 65.000 | 2.273 | 28.60 | 1.21e-07 *** |
| height3:width1 | 40.000 | 2.273 | 17.60 | 2.16e-06 *** |
| height1:width2 | 43.000 | 2.273 | 18.92 | 1.41e-06 *** |
| height2:width2 | 69.000 | 2.273 | 30.36 | 8.48e-08 *** |
| height3:width2 | 44.000 | 2.273 | 19.36 | 1.23e-06 *** |

Response: y

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|--------------|----|--------|---------|---------|---------------|
| height:width | 6 | 32792 | 5465.3 | 528.9 | 6.702e-08 *** |
| Residuals | 6 | 62 | 10.3 | | |

```
cm<-c(0, 1/2, 1/2, 0, -1/2, -1/2)
estimable(bm, cm)
```

| Estimate
<dbl> | Std. Error
<dbl> | t value
<dbl> | DF
<dbl> | Pr(> t)
<dbl> |
|-------------------|---------------------|------------------|-------------|-------------------|
| -4 | 2.27303 | -1.759765 | 6 | 0.1289371 |

Example: Bread sales

4. Is the average sale the highest in the middle height?

Consider a simultaneous confidence interval with Bonferroni procedure at 0.95 level.

| Factor A
(display height, i) | Factor B
(display width, j) | Mean | n |
|---------------------------------|--------------------------------|------|---|
| i=1 bottom | j=1 (regular) | 45 | 2 |
| i=2 middle | j=1 (regular) | 65 | 2 |
| i=3 top | j=1 (regular) | 40 | 2 |
| i=1 bottom | j=2 (wide) | 43 | 2 |
| i=2 middle | j=2 (wide) | 69 | 2 |
| i=3 top | j=2 (wide) | 44 | 2 |
| MSE=10.3 a=3, b=2 | | | |

$$\hat{D}_1 = \hat{Y}_{1.} - \hat{Y}_{2.} = (\hat{Y}_{11} + \hat{Y}_{12})/2 - (\hat{Y}_{21} + \hat{Y}_{22})/2 = -23$$

$$s^2\{\hat{D}_1\} = \frac{2MSE}{bn} = \frac{2(10.3)}{2(2)} = 5.15, \quad \text{so } s\{\hat{D}_1\} = 2.27$$

$$\hat{D}_2 = \hat{Y}_{3.} - \hat{Y}_{2.} = (\hat{Y}_{31} + \hat{Y}_{32})/2 - (\hat{Y}_{21} + \hat{Y}_{22})/2 = -25$$

$$s^2\{\hat{D}_2\} = \frac{2MSE}{bn} = \frac{2(10.3)}{2(2)} = 5.15, \quad \text{so } s\{\hat{D}_2\} = 2.27$$

The $1 - \alpha$ confidence interval for D are

$$\hat{D} \pm Bs\{\hat{D}\} \quad \text{Where } B = t\left(1 - \frac{\alpha}{2g}; (n - 1)ab\right) = t(0.9875; 6) = 2.97$$

$$\hat{D}_1 \pm Bs\{\hat{D}_1\} = -23 \pm 2.97 * 2.27 = -23 \pm 6.74 = (-29.74, -16.26)$$

$$\hat{D}_2 \pm Bs\{\hat{D}_2\} = -25 \pm 2.97 * 2.27 = -25 \pm 6.74 = (-31.74, -18.26)$$

Example: Teaching method (significant interaction)

A junior college system studies the effects of teaching method (factor A) and student's quantitative ability (factor B) on learning of College mathematics.

- Factor A (teaching methods) Abstract and Standard, $a=2$
- Factor B (quantitative ability) Excellent, Good, and Moderate, $b=3$
- $n=42$ students were selected and randomly placed into classes, with each class containing equal numbers of students of each quantitative ability level.
- Y is the amount of learning of college mathematics, measured by a standard mathematics achievement test.

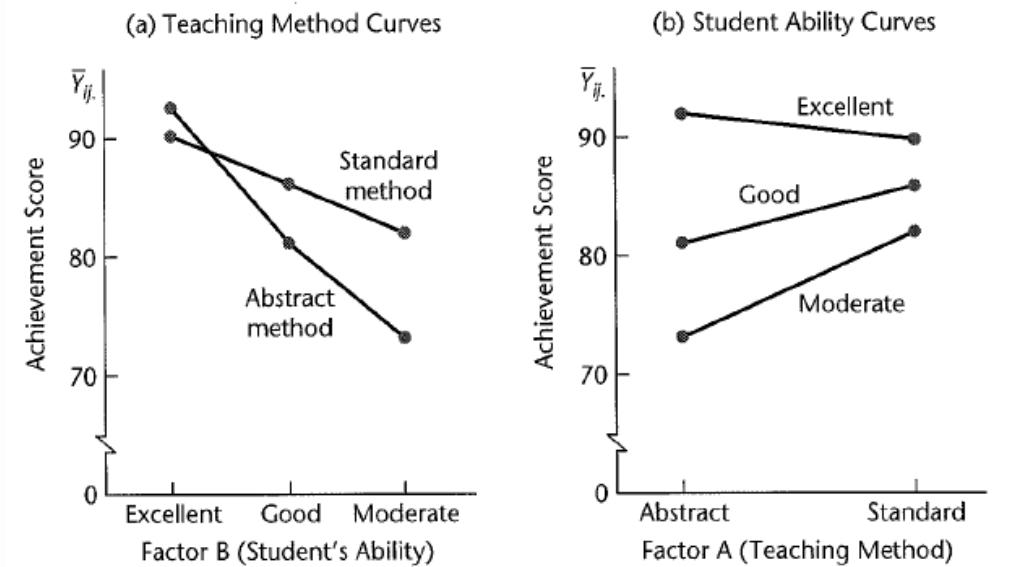
| (a) Mean Learning Scores ($n = 21$) | | | |
|---------------------------------------|-----------------------------------|------------------------|------------------------|
| Teaching Method
<i>i</i> | Quantitative Ability (<i>j</i>) | | |
| | Excellent | Good | Moderate |
| Abstract | 92 ($\bar{Y}_{11.}$) | 81 ($\bar{Y}_{12.}$) | 73 ($\bar{Y}_{13.}$) |
| Standard | 90 ($\bar{Y}_{21.}$) | 86 ($\bar{Y}_{22.}$) | 82 ($\bar{Y}_{23.}$) |

| (b) ANOVA Table | | | |
|---------------------------------|-------|-----|---------|
| Source of Variation | SS | df | MS |
| Factor A (teaching methods) | 504 | 1 | 504 |
| Factor B (quantitative ability) | 3,843 | 2 | 1,921.5 |
| AB interactions | 651 | 2 | 325.5 |
| Error | 3,360 | 120 | 28 |
| Total | 8,358 | 125 | |

Example: Teaching method (significant interaction)

| (a) Mean Learning Scores ($n = 21$) | | | |
|---------------------------------------|------------------------------|------------------------|------------------------|
| Teaching Method
i | Quantitative Ability (j) | | |
| | Excellent | Good | Moderate |
| Abstract | 92 ($\bar{Y}_{11.}$) | 81 ($\bar{Y}_{12.}$) | 73 ($\bar{Y}_{13.}$) |
| Standard | 90 ($\bar{Y}_{21.}$) | 86 ($\bar{Y}_{22.}$) | 82 ($\bar{Y}_{23.}$) |

| (b) ANOVA Table | | | |
|---------------------------------|-------|-----|---------|
| Source of Variation | SS | df | MS |
| Factor A (teaching methods) | 504 | 1 | 504 |
| Factor B (quantitative ability) | 3,843 | 2 | 1,921.5 |
| AB interactions | 651 | 2 | 325.5 |
| Error | 3,360 | 120 | 28 |
| Total | 8,358 | 125 | |



Investigate the nature of the interaction effects: estimating separately for students with excellent, good, and moderate quantitative abilities, how large is the difference in mean learning for the two teaching methods.

$$D_1 = \mu_{11} - \mu_{21}$$

$$\hat{D}_1 = 92 - 90 = 2$$

$$s^2\{\hat{D}_1\} = s^2\{\hat{D}_2\} = s^2\{\hat{D}_3\} = \frac{2(28)}{21} = 2.667$$

$$D_2 = \mu_{12} - \mu_{22}$$

$$\hat{D}_2 = 81 - 86 = -5$$

$$s\{\hat{D}_2\} = s\{\hat{D}_1\} = s\{\hat{D}_3\} = 1.633$$

$$D_3 = \mu_{13} - \mu_{23}$$

$$\hat{D}_3 = 73 - 82 = -9$$

Consider Bonferroni procedure, $B = t\left(1 - \frac{\alpha}{2g}, ab(n-1)\right) = 2.428$

and the 95 percent confidence intervals for the family of comparisons are:

$$-1.96 \leq \mu_{11} - \mu_{21} \leq 5.96$$

$$-8.96 \leq \mu_{12} - \mu_{22} \leq -1.04$$

$$-12.96 \leq \mu_{13} - \mu_{23} \leq -5.04$$