Recap: Joint Distribution: $(\hat{Y}, \hat{\epsilon})$

Theorem

Under the Gaussian linear model,

$$\begin{pmatrix} \hat{Y} \\ \hat{\epsilon} \end{pmatrix} \sim \mathrm{N}\left\{ \begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} H & 0 \\ 0 & I_n - H \end{pmatrix} \right\}$$

so $\hat{Y} \perp\!\!\!\perp \hat{\epsilon}$

- Orthogonal: a linear algebra fact without assumptions. $< x, y >= 0$
- Independent: a statistical property under the Gaussian linear model. $f(x, y) = f(x)f(y)$
- Uncorrelated: $\mathrm{cov}(x, y) = 0$

Statistical Inference for β_j

- $\hat{\beta} \sim \mathrm{N}(\beta, \sigma^2(X^TX)^{-1})$
- We can also denote it as $\hat{\beta}|X \sim \mathrm{N}(\beta, \sigma^2(X^TX)^{-1})$
- Therefore, for $j = 0, 1, \ldots, p$:

$$\hat{\beta}_j|X \sim \mathrm{N}(\beta_j, \mathrm{Var}(\hat{\beta}_j|X))$$

  $\mathrm{Var}(\hat{\beta}_j|X)$: j+1 th diagonal entry of $\sigma^2(X^TX)^{-1}$

- $\dfrac{\hat{\beta}_j - \beta_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j|X)}} \sim \mathrm{N}(0, 1)$
- Issue: σ^2 is unknown
- $\dfrac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\mathrm{Var}}(\hat{\beta}_j|X)}} = \dfrac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} \sim t_{n-p-1}$

Hypothesis test with $\alpha = 0.05$

```r
summary(lmtrees)$coef
```

```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)    -6.631617 0.79978973 -8.291701 5.057138e-09
## log(Diameter)   1.982650 0.07501061 26.431592 2.422550e-21
## log(Height)     1.117123 0.20443706  5.464388 7.805278e-06
```

Suppose now we want to test

- $H_0 : \beta_1 = 2$, $H_1 : \beta_1 \neq 2$

$$t_1 = \frac{\hat{\beta}_1 - 2}{s.e.(\hat{\beta}_1)} = \frac{1.983 - 2}{0.075} \approx -0.227 \text{ with } df = 31 - 2 - 1 = 28$$

```r
# Compute p value
2*pt(0.227,df=28, lower.tail=F)
```

```
## [1] 0.822073
```

Conclusion: Since $0.82 > 0.05$, we can not reject H_0

- Exercise: $H_0 : \beta_2 = 1$, $H_1 : \beta_2 \neq 1$

Variance of estimation and prediction

For SLR, $(X^TX)^{-1}$ equals

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} & -\frac{\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ -\frac{\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} & \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{bmatrix}$$

We get $\mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right), \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$

$\mathrm{cov}(\hat{\beta}_1, \hat{\beta}_0) = -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Variance of estimation and prediction

- $\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$
- $\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0 + \epsilon) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) + \sigma^2$
- As n gets larger,
  - $\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0)$ would go down to zero
  - $\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0 + \epsilon)$ just goes down to σ^2

Confidence Intervals VS Prediction Intervals

- $100(1-\alpha)\%$ confidence interval for $\beta_0 + \beta_1 x_0$ is:

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{n-2,\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

- $100(1-\alpha)\%$ prediction interval for $Y = \beta_0 + \beta_1 x_0 + \epsilon$ is:

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{n-2,\alpha/2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

Just like SLR, two problems have identical estimated/predicted values:

$$\hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_p x_{0p}$$

but their standard errors are different

$$s.e.(E(\hat{Y}|X_0)) = \hat{\sigma}\sqrt{x_0^T (X^T X)^{-1} x_0}$$

$$s.e.(\hat{Y}|X_0) = \hat{\sigma}\sqrt{1 + x_0^T (X^T X)^{-1} x_0}$$

where $x_0^T = (1, x_{01}, \ldots, x_{0p})^T$

Confidence Intervals and Prediction Intervals

- $100(1-\alpha)\%$ confidence interval for $E(Y|X_0) = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p}$ is:

$$\hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_p x_{0p} \pm t_{n-p-1,\alpha/2} s.e.(E(\hat{Y}|X_0))$$

- $100(1-\alpha)\%$ prediction interval for $Y = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p} + \epsilon$ is:

$$\hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_p x_{0p} \pm t_{n-p-1,\alpha/2} s.e.(\hat{Y}|X_0)$$

Sum of Squares

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SST} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{SSR} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{SSE}$$

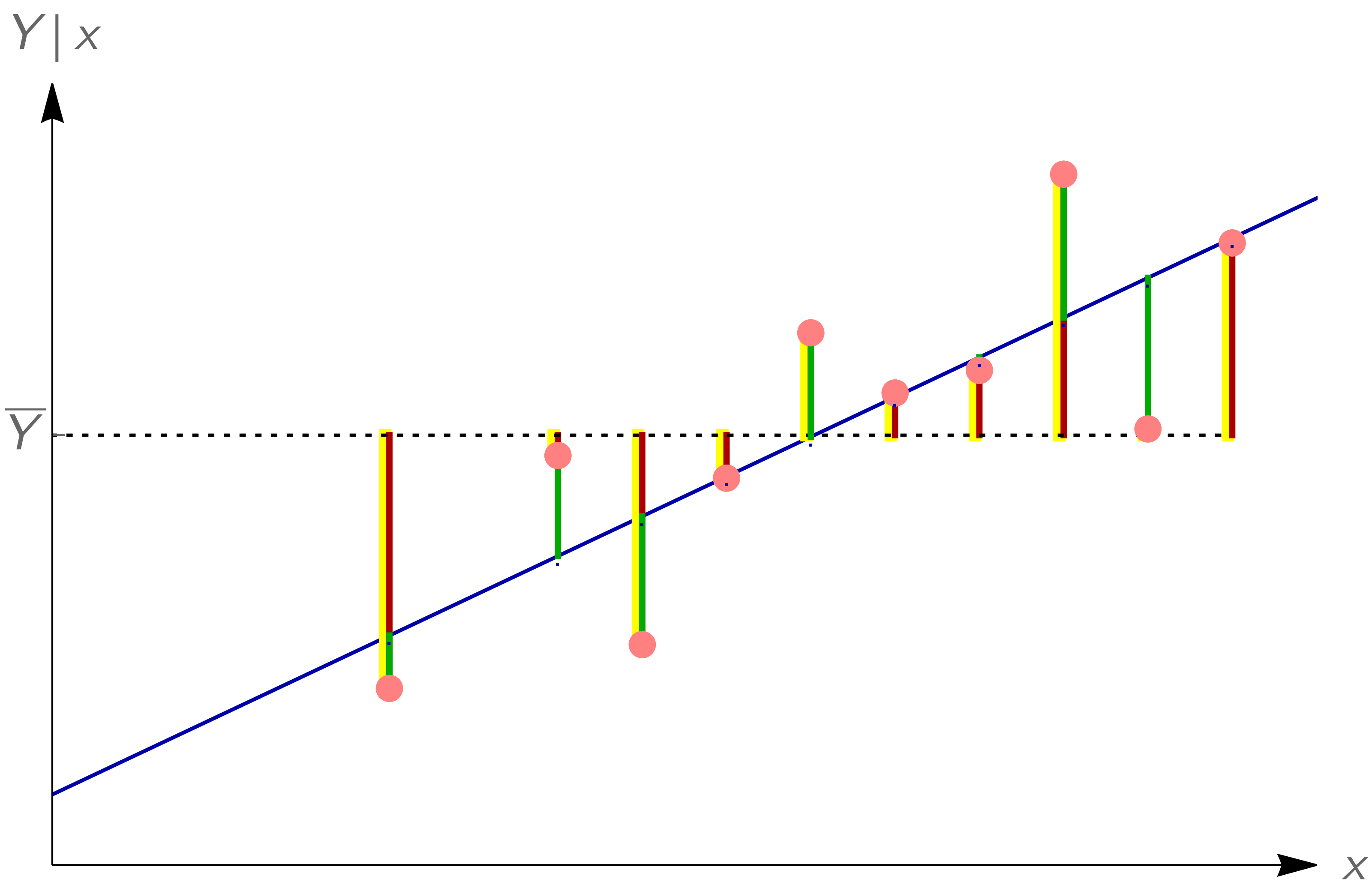
(Proof on board)

- SST = total sum of squares
    - total variability of Y
    - depends on the response Y only, not on the form of the model
- SSR = regression sum of squares
    - variability of Y explained by $X_1, \dots, X_p$
- SSE = error (residual) sum of squares
    - variability of Y not explained by $X_1, \dots, X_p$

Fundamental Sum-of-Squares Error Decomposition

$$\sum(\text{yellow lengths})^2 = \sum(\text{green lengths})^2 + \sum(\text{red lengths})^2$$

SST SSE SSR



Multiple R^2 and Adjusted R^2

- Multiple R^2 also called the coefficient of determination, is defined as

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

- Interpretation: proportion of variability in Y explained by $X_1, \ldots, X_p$
- $0 \leq R^2 \leq 1$
- For SLR, $R^2 = r_{xy}^2$ is the square of the correlation between X and Y. So multiple R^2 is a generalization of the correlation
- For MLR, R^2 is the square of the correlation between Y and $\hat{Y}$
- When more terms are added into a model, R^2 may increase or stay the same but never decrease.
  - Is large R^2 always preferable?

Adjusted R^2

Since R^2 always increases as we add terms to the model, some people prefer to use an adjusted R^2 defined as

$$R^2_{adj} = 1 - \frac{SSE/df_{SSE}}{SST/dt_{SST}} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$
$$= 1 - \frac{n-1}{n-p-1}(1-R^2)$$

- $-\frac{p}{n-p-1} \leq R^2_{adj} \leq R^2 \leq 1$
- R^2_{adj} can be negative
- R^2_{adj} does not always increase as more variables are added. In fact, if unnecessary terms are added, R^2_{adj} may decrease.

General Framework for Testing Nested Models

H_0: Reduced model is true v.s. H_1: Full model is true

- Simplicity or Accuracy?
  - The full model fits the data better (with a smaller SSE) but it is more complex.
  - The reduced model doesn't fit as well but it is simpler.
  - If $\text{SSE}_{reduced} \approx \text{SSE}_{full}$: one can sacrifice a bit of accuracy in exchange for simplicity.
  - If $\text{SSE}_{reduced} \gg \text{SSE}_{full}$:it would sacrifice too much in accuracy in exchange for simplicity. The full model is preferred.
- How to quantify?
  - F statistic!

The F-Statistic

$$F = \frac{(\text{SSE}_{reduced} - \text{SSE}_{full})/(df_{reduced} - df_{full})}{\text{MSE}_{full}}$$

- Under H_0, F statistic has an F-distribution with $(df_{reduced} - df_{full}, df_{full})$ degrees of freedom
- $F \geq 0$ since $\text{SSE}_{reduced} \geq \text{SSE}_{full}$
- The smaller the F-statistic, the more the reduced model is favored

Example 1: Testing All Coefficients Equal to Zero

$H_0 : \beta_1 = \beta_2 \cdots = \beta_p = 0$ v.s. H_a: not all $\beta_1, \ldots, \beta_p = 0$

- The is a test to evaluate the **overall significance** of a model.
  - Full: $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$
  - Reduced: $y_i = \beta_0 + \epsilon_i$ (All predictors are unnecessary)
- The OLS estimate for β_0 in the reduced model is $\hat{\beta}_0 = \bar{y}$, so $SSE_{reduced} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SST_{full}$
- $F = \dfrac{(SSE_{reduced} - SSE_{full})/(df_{reduced} - df_{full})}{MSE_{full}} = \dfrac{(SST_{full} - SSE_{full})/(n - 1 - (n - p - 1))}{MSE_{full}} = \dfrac{SSR_{full}/p}{MSE_{full}} = \dfrac{MSR_{full}}{MSE_{full}}$
- Moreover, $F \sim F_{p,n-p-1}$ under $H_0 : \beta_1 = \beta_2 \cdots = \beta_p = 0$.

Example 1: Testing All Coefficients Equal to Zero

In R, the F statistic and p-value are displayed in the last line of the output of the `summary()` command.

```
data(trees)
trees$Diameter = trees$Girth
lmtrees = lm(log(Volume) ~ log(Diameter) + log(Height), data=trees)
summary(lmtrees)
```

```
##
## Call:
## lm(formula = log(Volume) ~ log(Diameter) + log(Height), data = trees)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.168561 -0.048488  0.002431  0.063637  0.129223
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.63162    0.79979  -8.292 5.06e-09 ***
## log(Diameter)  1.98265    0.07501  26.432  < 2e-16 ***
## log(Height)    1.11712    0.20444   5.464 7.81e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08139 on 28 degrees of freedom
## Multiple R-squared:  0.9777, Adjusted R-squared:  0.9761
## F-statistic: 613.2 on 2 and 28 DF,  p-value: < 2.2e-16
```

Example 2: General Test on Full and Reduced Models

anova(model1,model2)

```
lmfull = lm(log(Volume) ~ log(Diameter) + log(Height),
            data=trees)
lmreduced = lm(log(Volume) ~ 1, data=trees)
anova(lmreduced, lmfull)
```

```
## Analysis of Variance Table
##
## Model 1: log(Volume) ~ 1
## Model 2: log(Volume) ~ log(Diameter) + log(Height)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 8.3087
## 2     28 0.1855  2    8.1232 613.19 < 2.2e-16 ***
## ---
```

$RSS(\beta) = (y - X\beta)^T (y - X\beta)$ $\quad \nabla_\beta RSS(\beta) = 0$

$\Rightarrow$ If $(X^TX)^{-1}$ exists, $\boxed{\hat\beta = (X^TX)^{-1}X^Ty}$

$\nabla_B RSS(B) = \nabla_B \operatorname{Tr}[(y - XB)^T(y - XB)]$

$= \nabla_B \operatorname{Tr}[y^Ty - y^T(XB) - (XB)^Ty + (XB)^TXB]$

$= \nabla_B [\operatorname{Tr}(y^Ty) - 2\operatorname{Tr}(y^T(XB)) + \operatorname{Tr}(B^TX^TXB)]$

$= 0 - 2(y^TX)^T + 2(X^TX)B = 0$

$\Rightarrow \boxed{\hat B = (X^TX)^{-1}X^Ty}$

$\boxed{\hat Y = X\hat\beta + \hat e \quad e \sim N(0, \hat\sigma^2)}$

Simple Linear Regression:

$RSS(\beta) = \Sigma(y_i - \hat\beta_0 - \hat\beta_1 \cdot x_i)\begin{pmatrix}1\\x_i\end{pmatrix} = 0$ $\quad \nabla_{\beta_0,\beta_1} RSS(\beta) = 0$

$\Rightarrow \boxed{\hat\beta_0 = \bar y - \hat\beta_1 \bar x}$ $\quad \boxed{\hat\beta_1 = (\Sigma(x_i - \bar x)(y_i - \bar y)) / (\Sigma(x_i - \bar x)^2)}$

unbiased $\Leftarrow E[\hat\sigma^2] = \sigma^2$ $\quad = RSS(\hat\beta)$

$\sigma^2 = (\Sigma(e_i - \bar e)^2)/(n-1) \Rightarrow \boxed{\hat\sigma^2 = (\Sigma \hat e_i^2)/[n-(p+1)]}$

$\bar{\hat e} = \frac{1}{n}\Sigma \hat e_i = 0$

$\hat\beta = (X^TX)^{-1}X^Ty = (X^TX)^{-1}X^T(X\beta + e) = \beta + (X^TX)^{-1}X^Te$

$\Rightarrow E[\hat\beta] = \beta + (X^TX)^{-1}X^T \cdot E(e|X) = \beta \Rightarrow$ unbiased

$\Rightarrow Var[\hat\beta] = (X^TX)^{-1}X^T \cdot Var(e|X) \cdot [(X^TX)^{-1}X^T]^T = \sigma^2 \cdot (X^TX)^{-1}$

$\Rightarrow Cov(\hat\beta_k, \hat\beta_j) = (k+1, j+1)^{th}$ entry of matrix $\sigma^2(X^TX)^{-1}$

$Cov(\hat\beta_i, \hat\beta_i) = Var(\hat\beta_i) = i^{th}$ diag entry of $\sigma^2(X^TX)^{-1}$

$\because \sigma^2$ is unknown $\therefore \widehat{Var}(\hat\beta|X) = \hat\sigma^2(X^TX)^{-1}$

If assume $e_{n\times1}|X \sim N(0, \sigma^2 I_n)$, then $\hat\beta|X \sim N(\beta, \sigma^2(X^TX)^{-1})$

$\Rightarrow \hat\beta_i|X \sim N(\beta_i, Var(\hat\beta_i|X))$

$\Rightarrow (\hat\beta_i - \beta_i)/\sqrt{Var(\hat\beta_i|X)} \sim N(0,1)$ $\quad$ preknown/assumed, usually 0

$(\hat\beta_i - \beta_i)/s.e.(\hat\beta_i) \sim t$-distribution with $d.f. = n-(p+1)$

$= \sqrt{\widehat{Var}(\hat\beta_j|X)}$ $\quad \frac{\hat\beta_j - \beta_j^*}{s.e.(\hat\beta_j)} \in [-t_{\frac{\alpha}{2}}, t_{\frac{\alpha}{2}}]$

Summary():

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	$\hat\beta_0$	$\div$ $s.e.(\hat\beta_0)$	= T-Test statistics for $\begin{cases}H_0: \beta_i = 0\\H_1: \beta_i \neq 0\end{cases}$	p-value
predictor	$\hat\beta_1$	$=\sqrt{\widehat{Var}(\hat\beta_0|X)}$		

Residual standard error: $\sqrt{\hat\sigma^2} = \sqrt{RSS/[n-(p+1)]}$ on $n-(p+1)$ d.f.

F-stats: f-stats on $p-1$ & $n-(p+1)$ DF, p-value: $\begin{matrix}H_0: E[Y|X] = \beta_0\\H_1: E = \beta_0 + \beta^TX\end{matrix}$

$TSS = \Sigma(y_i - \bar y)^2$, Sample Variance $= TSS/(n-1) = \sigma^2$

$R^2 = 1 - \frac{RSS}{TSS} = \frac{SSreg}{TSS}$ $\quad SSreg = \Sigma(\hat y_i - \bar y)^2$

$Adj\text{-}R^2 = 1 - \frac{RSS/[n-(p+1)]}{TSS/(n-1)} = 1 - \left(\frac{n-1}{n-p-1}\right)\cdot(1 - R^2)$ $\quad \leftarrow \hat\sigma^2$ $\quad \leftarrow$ sample variance of y

Proportion of variability of y's that can be explained by x's

F-Test: H_0: model 1 (submodel) v.s. H_1: model 2

$F\text{-stats} = \frac{(RSS_{H_0} - RSS_{H_1})/(df_{H_0} - df_{H_1})}{RSS_{H_1}/df_{H_1}}$

If H_0 is true, F-stats $\overset{H_0}{\sim}$ F-dist with $(df_{H_0} - df_{H_1}, df_{H_1})$

anova(model 1, model 2): $\quad$ (model 1 = H_0, model 2 = H_1)

	Res. DF	RSS	DF	Sum of Sq	F	Pr(>F)
1	df_{H_0}	RSS_{H_0}			$\frac{(RSS_{H_0} - RSS_{H_1})/(df_{H_0} - df_{H_1})}{RSS_{H_1}/df_{H_1}}$	
2	df_{H_1}	RSS_{H_1}	$df_{H_0} - df_{H_1}$	$RSS_{H_0} - RSS_{H_1}$	= F-stats	P-value

$\sim F\text{-dist}(df_{H_0} - df_{H_1}, df_{H_1})$

anova(model) Type I Anova

	DF	Sum Sq	Mean Sq	F value $= \frac{\text{Mean Sq}}{\hat\sigma^2}$
X1	$df_{H_0} - df_{H_1} = 1$	$RSS_{H_0} - RSS_{H_1}$	$\frac{RSS_{H_0} - RSS_{H_1}}{df_{H_0} - df_{H_1}}$	$\overset{H_0}{\sim} F\text{-dist}(1, n-p-1)$
X2	1			
X3	1			
Residuals	$n-(p+1)$	RSS_{model}	$\hat\sigma^2 = \frac{RSS_{model}}{n-(p+1)}$	

$H_0: E(Y) = \beta_0$
$H_1: E(Y) = \beta_0 + \beta_1X_1$
$H_0: E(Y) = \beta_0 + \beta_1X_1$
$H_1: E(Y) = \beta_0 + \beta_1X_1 + \beta_2X_2$
$H_0: E(Y) = \beta_0 + \beta_1X_1 + \beta_2X_2$
$H_1: E(Y) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3$

Prediction: suppose new observation (x^*, y^*).

$y^* = (x^*)^T\beta + e^*$ $\quad \hat y^* = (x^*)^T\hat\beta$

$\Rightarrow s.e.(\hat y^*) = \sqrt{Var((x^*)^T\hat\beta) + Var(e^*)} = \sqrt{\hat\sigma^2(x^*)^T(X^TX)^{-1}(x^*) + \hat\sigma^2}$

$\Rightarrow$ Prediction CI with $(1-\alpha)\%$ level: $\hat y^* \pm t_{\alpha/2, n-(p+1)} \cdot s.e.(\hat y^*)$

For fitted value. $s.e.(\text{fitted}) = \sqrt{Var((x^*)^T\hat\beta)}$

Model Diagnosis:

Assumption: ① $E[e|X] = 0$ ② $Var[e|X] = \sigma^2 I_n$

③ For hypothesis testing, $e_{n\times1} \sim N(0, \sigma^2 I_n)$

$\Rightarrow E(e|X) = 0 \Rightarrow E(e_i|X) = 0 \Rightarrow E(e_ix_i|X) = 0 \Rightarrow E(e_ix_i) = 0$

$\Rightarrow cov(e_i, x_i) = E(e_ix_i) - E(e_i)\cdot E(x_i) = 0$

Property of Residuals:

① $\frac{1}{n}\Sigma\hat e_i = 0 \Rightarrow \Sigma\hat e_i \cdot x_i = 0 \Rightarrow X^T\hat e = 0$

② sample $cov(x, \hat e) = \frac{1}{n-1}\cdot\Sigma(x_i - \bar x)(\hat e_i - \bar{\hat e}) = 0$

③ sample $cov(\hat y, \hat e) = \cdots = \frac{1}{n-1}(\Sigma\hat y_i\cdot\hat e_i - \Sigma\bar y\cdot\hat e_i) = \frac{1}{n-1}(\Sigma\hat e_ix_i^T)\hat\beta = 0$

④ $\hat e = y - \hat y = (I - X(X^TX)^{-1}X^T)y = (I - H)y$ $\quad HX = X \quad HH = H$

$= (I - X(X^TX)^{-1}X^T)(X\beta + e) = (I - H)e$ $\quad H^T = H \quad \hat y \perp \hat e$

⑤ $\boxed{\begin{matrix}E(\hat e|X) = E[(I_n - H)e|X] = (I_n - H)\cdot E(e|X) = 0 \quad Var(\hat e_i|X) = \sigma^2(1-h_{ii})\\ Var(\hat e|X) = (I_n - H)\,Var[e|X]\cdot(I_n - H)^T = (I_n - H)\cdot\sigma^2 I_n\cdot(I_n - H) = \sigma^2(I_n - H)\end{matrix}}$

⑥ $h_{ij} = x_i^T(X^TX)^{-1}x_j$, h_{ii} is the leverage of i^{th} observation.

Property of h_{ii}: $h_{ii} = (1 \ \ x_i)(X^TX)^{-1}\begin{pmatrix}1\\x_i\end{pmatrix}$ (Simple Linear R)

① $(X^TX)^{-1} = \frac{1}{n\Sigma x_i^2 - (\Sigma x_i)^2}\cdot\begin{pmatrix}\Sigma x_i^2 & -\Sigma x_i\\ -\Sigma x_i & n\end{pmatrix} \Rightarrow h_{ii} = \frac{1}{n} + \frac{(x_i - \bar x)^2}{\Sigma(x_j - \bar x)^2} \in [\frac{1}{n}, 1]$

How large is leverage $\Leftrightarrow$ How far x_i is from the center.

② $\widehat{Var}\begin{pmatrix}\hat\beta_0\\\hat\beta_1\end{pmatrix} = \hat\sigma^2(X^TX)^{-1} \Rightarrow \widehat{Var}(\hat\beta_0) = \hat\sigma^2\cdot\frac{\Sigma x_i^2}{n\Sigma x_i^2 - (\Sigma x_i)^2}$, $\widehat{Var}(\hat\beta_1) = \hat\sigma^2\cdot\frac{1}{\Sigma x_i^2 - n\cdot\bar x^2}$

③ $Var(\hat e_i|X) = \sigma^2(1 - h_{ii})$: $h_{ii} \to 1 \Leftrightarrow Var(\hat e_i|X) \to 0 \Rightarrow$ regression line pass the point

Z-Test: $H_0: \beta_0 = 0$ $\quad H_A: \beta_0 \neq 0$, If H_0 is true, $Z_{n-1} \overset{H_0}{\sim} N(0,1)$

σ^2 is known: $Z_{n-1} = \frac{\bar y}{\sigma/\sqrt n}$, $\bar y = \hat\beta_0$ $\quad (y_i = \beta_0 + e_i)$

T-Test: $T_{n-1} = \frac{\bar y}{\hat\sigma/\sqrt n}$, $\bar y = \hat\beta_0$, $\hat\sigma = \sqrt{\frac{\Sigma(y_i - \bar y)^2}{n-1}}$, $se(\hat\beta_0) = \hat\sigma/\sqrt n$

If H_0 is true, $T_{n-1} \overset{H_0}{\sim}$ t-dist with $df = n-1$

T-Test: $H_0: \beta_j = \beta_j^*$. $H_A: \beta_j \neq \beta_j^* \Rightarrow T_{n-1} = \frac{\hat\beta_j - \beta_j^*}{se(\hat\beta_j)}$

If H_0 is true, $e_{n\times1} \overset{iid}{\sim} N(0, \sigma^2 I_n)$

$\Rightarrow T_{n-1} \overset{H_0}{\sim}$ t-dist with $df = n-(p+1)$

Accept H_0 if $\beta_j^* \in \hat\beta_j \pm t_{\frac{\alpha}{2}}\cdot se(\hat\beta_j)$

$\nabla_B \operatorname{Tr}(AB) = A^T$ (mxd, dxm) $\quad \nabla\operatorname{Tr}(B^TCB) = 2CB$ (mxd, dxd, dxm; symmetric)

$\bar y = \hat\beta_0 + \hat\beta_1\bar x$ $\quad \hat y = y + \hat e$ $\quad \frac{1}{n}\Sigma\hat y_i = \bar y$

β_0 is the expected value of the response when $x = \bar 0$

β_j gives the change in the expected value when x_j increases by 1 units and others hold

Anova(model) Type II Anova

	Sum Sq	DF	F value
X1	$RSS_{H_0} - RSS_{H_1}$	1	$\frac{RSS_{H_0} - RSS_{H_1}}{RSS_{H_1}/df_{H_1}} = \hat\sigma^2$ $\overset{H_0}{\sim} F\text{-dist}(1, n-p-1)$
X2		1	
X3		1	
Residuals	RSS_{model}	$n-(p+1)$	

$H_0: E(Y) = \beta_0 + \beta_2X_2 + \beta_3X_3$
$H_1: E(Y) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3$
$H_0: E(Y) = \beta_0 + \beta_1X_1 + \beta_3X_3$
$H_1: E(Y) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3$
$H_0: E(Y) = \beta_0 + \beta_1X_1 + \beta_2X_2$
$H_1: E(Y) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3$