Linear Regression

Professor Jie Peng, PhD

Department of Statistics

University of California, Davis

Normal Error Model

Normal Error Model

Simple regression model $+$ Normality assumption:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where the error terms ϵ_is are *independently and identically distributed (i.i.d.)* $N(0, \sigma^2)$ random variables.

MLE

Under the Normal error model:

- LS estimators $\hat{\beta}_0, \hat{\beta}_1$ are the *maximum likelihood estimator (MLE)* of β_0, β_1, respectively.
- The MLE of σ^2 is SSE/n.

MSE=SSE/(n-2)

Sampling Distributions

b0=bar y - b1 bar x
b1= Sxy/Sxx=Σ(xi-bar x)yi/Sxx=yi s' linerar combination

Under the Normal error model: Yi~iid~N

- $\hat{\beta}_0, \hat{\beta}_1$ are normally distributed:

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2\{\hat{\beta}_0\}), \quad \hat{\beta}_1 \sim N(\beta_1, \sigma^2\{\hat{\beta}_1\}).$$

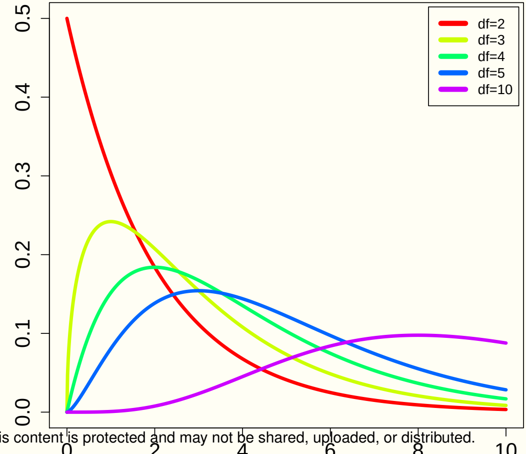
- SSE/σ^2 follows a χ^2 distribution with $n-2$ degrees of freedom, denoted by $\chi^2_{(n-2)}$. MSE=SSE/(n-2)
- SSE is independent with both $\hat{\beta}_0$ and $\hat{\beta}_1$.

MSE/

χ^2 Distributions

Figure: χ^2 distributions: probability density function

[0,+infty)



longer right-tail

right-skewed

confidence interval的目标是
量化estimator与被估计量之间的差距【err】
所以是对err建模，还原成标注xx变量
如果被估计量是常量，err的var就是estimator的var
如果被估计量是随机变量，err的var将由estimator的var、被估计量的var与Cov(estimator, 被估计量)组成

Confidence Intervals of Regression Coefficients

Pivotal Quantity

not a statistic
because involves unknown para.

Sxx=Σ(xi-bar x)^2

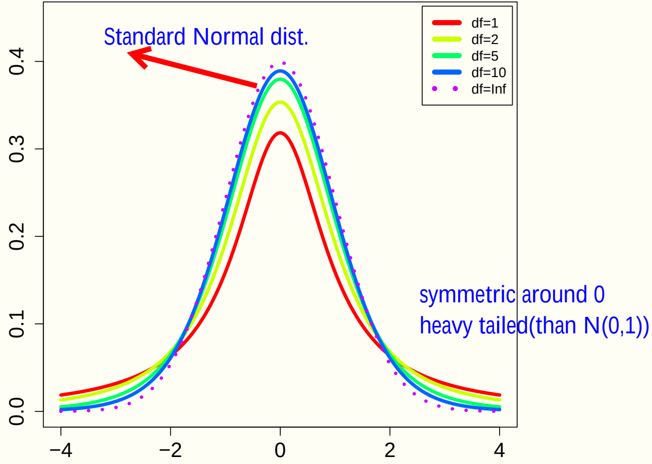
Hat b1 - b1
------------------- =
s(Hat b1 - b1)

$$\frac{\hat{\beta}_1 - \boxed{\beta_1}}{s\{\hat{\beta}_1\}}$$

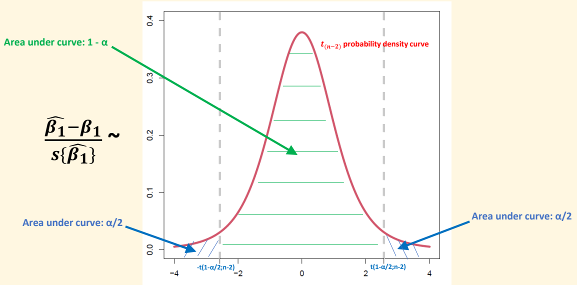
se b1= s{b1}=sqrt{MSE/Sxx}

- The numerator is the difference between the LS estimator $\hat{\beta}_1$ and its mean β_1.
- The denominator is the standard error of $\hat{\beta}_1$.
- This quantity follows a **known distribution**, $t_{(n-2)}$, t-distribution with $n-2$ degrees of freedom.

Figure: t distributions: probability density function*



*t *distribution with* ∞ *degrees of freedom is the standard normal* $N(0, 1)$ *distribution.*



$$P\left(\left|\frac{\hat{\beta}_1-\beta_1}{s\{\hat{\beta}_1\}}\right| \leq t(1-\alpha/2; n-2)\right) = 1-\alpha \Rightarrow$$

$$P\left(\hat{\beta}_1 - t(1-\alpha/2; n-2)s\{\hat{\beta}_1\} \leq \beta_1 \leq \hat{\beta}_1 + t(1-\alpha/2; n-2)s\{\hat{\beta}_1\}\right) = 1-\alpha$$

Confidence Interval

The $(1-\alpha)100\%$-confidence interval of β_1:

$$\hat{\beta}_1 \pm t(1-\alpha/2; n-2)s\{\hat{\beta}_1\},$$

where $t(1-\alpha/2; n-2)$ is the $(1-\alpha/2)100$th percentile of $t_{(n-2)}$.

Confidence Coefficient: Accuracy

- $(1-\alpha)100\%$ is called the *confidence coefficient* or the *confidence level*.
- Commonly used confidence coefficients are 95% ($\alpha = 0.05$), 90% ($\alpha = 0.1$), 99% ($\alpha = 0.01$).
- Confidence coefficient reflects **accuracy of the C.I.**: the larger (i.e., the smaller the α), the more accurate.

Confidence Interval Width: Precision

se b1= s{b1}=sqrt{MSE/Sxx}

- The half-width: $t(1-\alpha/2; n-2)s\{\hat{\beta}_1\}$

MSE/(Sx^2(n-1))

- The width reflects **precision of the C.I.**: the narrower, the more precise

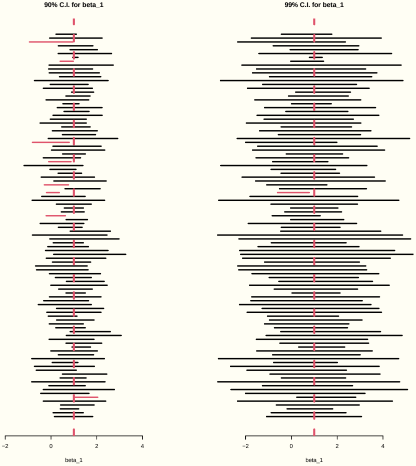
Sx^2=Σ(xi-bar x)^2/(n-1)

MSE=SSE/(n-2)=Σ(yi-hat y)^2/(n-2)

- Factors influencing the precision:
  - The larger the confidence coefficient (more accurate), the wider the C.I. (less precise)
  - The larger the sample size n (more data), the narrower the C.I. (more precise)
  - The larger the SE (more uncertainty), the wider the C.I. (less precise)

Simulation Experiment

Figure: C.I.s of β_1: Left: 90% C.I.; Right: 99% C.I.



Heights

- $n = 928$, $\overline{X} = 68.316$, $\sum_{i=1}^{n}(X_i - \overline{X})^2 = 3038.761$, and

$$\hat{\beta}_0 = 24.54, \;\; \hat{\beta}_1 = 0.637, \;\;\; MSE = 5.031.$$

- $s\{\hat{\beta}_1\} = \sqrt{\frac{5.031}{3038.761}} = 0.0407.$
- 95%-confidence interval of β_1:

$$\begin{aligned} 0.637 \pm t(0.975; 926) \times 0.0407 &= 0.637 \pm 1.963 \times 0.0407 \\ &= [0.557, 0.717]. \end{aligned}$$

- We are 95% confident that the regression slope is between 0.557 and 0.717.

T-test for β_1

- Null hypothesis: $H_0 : \beta_1 = \beta_1^{(0)}$, where $\beta_1^{(0)}$ is a given constant.
- **T-statistic**:

$$T^* = \frac{\hat{\beta}_1 - \beta_1^{(0)}}{s\{\hat{\beta}_1\}}.$$

- **Null distribution**:

$$\text{Under } H_0 : \beta_1 = \beta_1^{(0)}, \quad T^* \sim t_{(n-2)}.$$

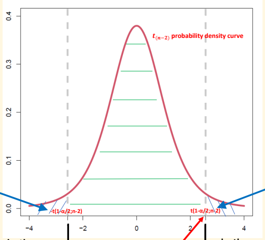
Decision Rules

Ha备择假设是我们想要接受的假设
但由于统计学家的严谨性
我们改为说，我们想要拒绝H0零假设

At significance level α:

- *Two-sided alternative* $H_a : \beta_1 \neq \beta_1^{(0)}$: Reject H_0 if and only if $|T^*| > t(1 - \alpha/\mathbf{2}; n - 2)$; Or equivalently, reject H_0 if and only if pvalue$:= P(|t_{(n-2)}| > |T^*|) < \alpha$.
- *Left-sided alternative* $H_a : \beta_1 < \beta_1^{(0)}$: Reject H_0 if and only if $T^* < t(\alpha; n - 2)$; Or equivalently, reject H_0 if and only if pvalue$:= P(t_{(n-2)} < T^*) < \alpha$.

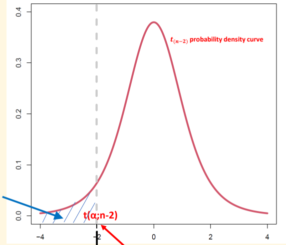
$$under\ H_0: T^* = \frac{\widehat{\beta_1} - \beta_1^{(0)}}{s\{\widehat{\beta_1}\}} \sim$$



two-sided alternative

H_a: $\beta_1 \neq \beta_1^{(0)}$

$$under\ H_0: T^* = \frac{\widehat{\beta_1} - \beta_1^{(0)}}{s\{\widehat{\beta_1}\}} \sim$$



Left-sided alternative

H_a: $\beta_1 < \beta_1^{(0)}$

Heights

Test whether there is a linear association between parent's height and child's height at significance level $\alpha = 0.01$.

- $H_0 : \beta_1 = 0$ *vs.* $H_a : \beta_1 \neq 0$.
- $T^* = \frac{\hat{\beta}_1 - 0}{s\{\hat{\beta}_1\}} = \frac{0.637}{0.0407} = 15.7$.
- **Critical value**: $t(1 - 0.01/2; 928 - 2) = 2.58$. Since the observed $|T^*| = |15.7| > 2.58$, reject the null hypothesis at level 0.01.
- **Pvalue**: $P(|t_{(926)}| > |15.7|) \approx 0$. Since $pvalue < \alpha = 0.01$, reject the null hypothesis at level 0.01.
- Conclusion: There is a significant association between parent's height and child's height at level 0.01.

Mean Response

Estimation of Mean Response

说法不自然，但是内容很自然，需要特别记忆一下

The mean response at $X = X_h$ is $E(Y_h) = \beta_0 + \beta_1 X_h$.

- An unbiased estimator of $E(Y_h)$:

HatYh~N(b0+b1Xh, σ^2{HatYh})

其实这里写HatE(Yh)更好！

Hat E(Yh)= $\widehat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h = \overline{Y} + \hat{\beta}_1 (X_h - \overline{X})$.

- $\sigma^2\{\widehat{Y}_h\} = \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]$.

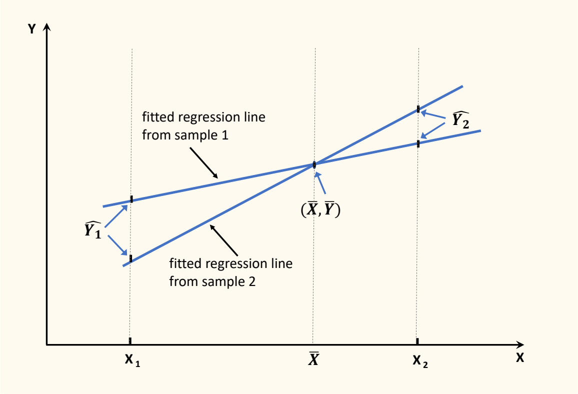
数学上需要记忆 的只有这个结论：
离MeanX 越远的Xh
这个信息越不够，
HatYh越不准/se/var越大

- Standard error of $\widehat{Y}_h$:

与σ2(Hat b1)不同哦
见week1-lectures

$$s\{\widehat{Y}_h\} = \sqrt{MSE \left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]}.$$

- The larger the sample size, or the larger the dispersion of X values, the smaller the SE of $\widehat{Y}_h$.
- The further X_h from $\overline{X}$, the larger the SE of $\widehat{Y}_h$.



Sampling Distribution of $\widehat{Y}_h$

Under the Normal error model:

- $\widehat{Y}_h$ is normally distributed:

EYh is constant

$$\widehat{Y}_h \sim \text{Normal}(E(Y_h), \sigma^2\{\widehat{Y}_h\})$$

这个说法记忆一下

- Pivotal quantity:

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

VarEY=0，Cov(HatYh,EYh)=0

Confidence Intervals of $E(Y_h)$

The $(1-\alpha)100\%$ confidence interval of $E(Y_h)$:

$$\widehat{Y}_h \pm t(1-\alpha/2; n-2)s(\widehat{Y}_h)$$

Heights

What is the average height of children of $70in$ parents?

- $n = 928,\ \overline{X} = 68.316,\quad \sum_{i=1}^{n}(X_i - \overline{X})^2 = 3038.761$ and $\hat{\beta}_0 = 24.54,\ \hat{\beta}_1 = 0.637,\ MSE = 5.031$
- $\widehat{Y}_h = 24.54 + 0.637 \times 70 = 69.2$
- $s\{\widehat{Y}_h\} = \sqrt{5.031 \times \left\{\frac{1}{928} + \frac{(70-68.316)^2}{3038.761}\right\}} = 0.1$
- 95%-confidence interval: $69.2 \pm 1.963 \times 0.1 = [69, 69.40]$
- We are 95% confident that the average height of children of $70in$ parents is between $[69in, 69.40in]$.

Prediction of New Outcome

Predict a **future outcome** at $X = X_h$:

$$Y_{h(new)} = \beta_0 + \beta_1 X_h + \epsilon_h$$

- Predict $Y_{h(new)}$ by the estimated mean response at $X = X_h$:

$$\widehat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h = \overline{Y} + \hat{\beta}_1 (X_h - \overline{X})$$

- ϵ_h is assumed to be uncorrelated with ϵ_is $\rightarrow$ $Y_{h(new)}$ is uncorrelated with the observed Y_is.

Pivotal Quantity

Var(err_h)=Var(Hat E Yh-Yh)

我们的目标是把估计的err构造成一个标准xx随机变量
so that 我们可以转换出一个confidence interval
所以这里要减去Yh_(new)

Under Normal error model:

随机变量

- err_h = $\widehat{Y}_h - Y_{h(new)} \sim \text{Normal}(0, \sigma^2(pred_h))$, where

HatYh只是由过去的epsilon数据推出的，与新的epsilon无关

$$\sigma^2(pred_h) := Var(\widehat{Y}_h - Y_{h(new)}) = \sigma^2(\widehat{Y}_h) + \sigma^2(Y_{h(new)})$$
$$= \sigma^2(\widehat{Y}_h) + \sigma^2 = \sigma^2\left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]$$

- Pivotal quantity: $\frac{\widehat{Y}_h - Y_{h(new)}}{s(pred_h)} \sim t_{(n-2)}$, where

$$s(pred_h) = \sqrt{MSE\left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]}$$

Prediction Intervals

对随机变量的估计

The $(1-\alpha)100\%$ prediction interval of $Y_{h(new)}$:

$$\widehat{Y}_h \pm t(1-\alpha/2; n-2)s(pred_h)$$

Use Hat Yh to est. Yhnew 的随机性大于
比用 Hat Yh to est. E Yh的随机性

Prediction vs. Estimation

- $Y_{h(new)}$ – a “moving target” (random variable) vs. $E(Y_h)$ – a fixed quantity (non-random).
- Two sources of variations in the prediction process: Variability from $\widehat{Y}_h$ and variability from the target $Y_{h(new)} \rightarrow s(pred_h) > s(\widehat{Y}_h)$. 见上一张ppt
- At a given X value, the prediction interval of a new outcome is wider than the confidence interval of the mean response.

Heights

What would be the predicted height of the child of a $70in$ couple?

- $n = 928$, $\overline{X} = 68.316$, $\sum_{i=1}^{n}(X_i - \overline{X})^2 = 3038.761$, and $\hat{\beta}_0 = 24.54$, $\hat{\beta}_1 = 0.637$, $MSE = 5.031$
- Predicted height: $\widehat{Y}_h = 24.54 + 0.637 \times 70 = 69.2$
- Standard error:

$$s\{pred_h\} = \sqrt{5.031 \times \left\{\mathbf{1} + \frac{1}{928} + \frac{(70 - 68.316)^2}{3038.761}\right\}} = 2.25$$

- 95% prediction interval: $69.2 \pm 1.8831 \times 2.25 = [64.75, 73.56]$
- We are 95% confident that the child's height will be between $[64.75in, 73.56in]$.

Extrapolation

Extrapolation occurs when predicting the outcome at an X value that lies outside of the observed data range.

- Every model has a **range of validity**.
- A model may be inappropriate when it is extended outside of the range of the observations upon which it was built.
- Extrapolation is less reliable than interpolation and need to be handled with caution.

Analysis of Variance

Analysis of Variance

- Basic idea: attributing variation in the data to different sources through **decomposition of the total variation**.
- In regression, the variation in the observations comes from:
  - variation in the error term
  - variation in X

Partition of Total Deviation

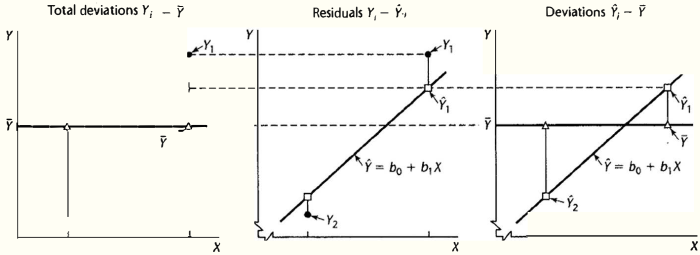
- **Total deviation:** difference between Y_i and the sample mean $\overline{Y}$:

$$Y_i - \overline{Y}, \quad i = 1, \cdots, n.$$

- Total deviation can be decomposed into the sum of two terms:

$$Y_i - \overline{Y} = (Y_i - \widehat{Y}_i) + (\widehat{Y}_i - \overline{Y}), \qquad i = 1, \ldots, n$$

- I.e., the *deviation of the observed value around the fitted regression line (residual)* and the *deviation of the fitted value from the sample mean*.


Total deviations $Y_i - \bar{Y}$
Residuals $Y_i - \hat{Y}_i$
Deviations $\hat{Y}_i - \bar{Y}$
Y
X
$\bar{Y}$
Y_1
Y_2
$\hat{Y}_1$
$\hat{Y}_2$
$\hat{Y} = b_0 + b_1 X$

Decomposition of Total Variation

- Taking sum of squares of the total deviations and noting that the sum of the cross product terms vanishes:

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 + \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2.$$

- Decomposition of total variation:

$$SSTO = SSE + SSR$$

ANOVA: Sums of Squares

Total Sum of Squares (SSTO)

Quantify variation of the observations around the sample mean:

$$SSTO := \sum_{i=1}^{n}(Y_i - \overline{Y})^2, \quad d.f.(SSTO) = n - 1.$$

Σ (Yi-BarY)=0　　1个线性限制

Error Sum of Squares (SSE)

Quantify variation of the observations around the fitted regression line:

$$SSE = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2, \quad d.f.(SSE) = n - 2.$$

Σ ei^2=0, Σ eiXi = 0 2个线性限制

Regression Sum of Squares (SSR)

Quantify variation of the fitted values around the sample mean:

Hat b1只有一个自由度

$$SSR = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 = \hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2, \quad d.f.(SSR) = 1.$$

- $SSR = SSTO - SSE$: reduction of uncertainty in Y by utilizing the predictor X through a linear regression model
- The larger the fitted regression slope or the more the dispersion of X values, the larger SSR

Mean Squares

Sum of Squares divided by its degree of freedom:

$$MS = SS/d.f.(SS).$$

- Mean squared error:

$$MSE = \frac{SSE}{\text{d.f.}(SSE)} = \frac{SSE}{n-2}$$

- Regression mean square:

$$MSR = \frac{SSR}{\text{d.f.}(SSR)} = \frac{SSR}{1}$$

ANOVA: F Tests

Expected Values of SS and MS

Under simple regression model:

- Expected values of SS:

$$E(SSE) = (n-2)\sigma^2, \quad E(SSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

- Expected values of MS:

$$E(MSE) = \sigma^2, \qquad E(MSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

- $E(MSR) \geq E(MSE)$ and " $=$ " holds iff $\beta_1 = 0$.

Sampling Distributions of SS

Under Normal error model:

- $SSE \sim \sigma^2 \chi^2_{(n-2)}$
- SSE and SSR are independent.

F Test

$$SSR = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 = \hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2, \quad d.f.(SSR) = 1.$$

- $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$
- F ratio: $F^* = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)}$ df1 df2
- Null distribution: $F^* \underset{H_0:\beta_1=0}{\sim} F_{1,n-2}$.
- Decision rule at the significance level α:

F test 只能two-sided test
i.e. H1:b1≠0
因为是var，所有都平方了

这里是因为
F*不可能比1小
所以尽管Ha是不等号的
这边只能one sided【df1=1的F分布domain是(0,infty)，
所以这里很神奇的是，尽管SSE限制了SSR的最小值，SSE与SSR是相互独立的】

df1 df2

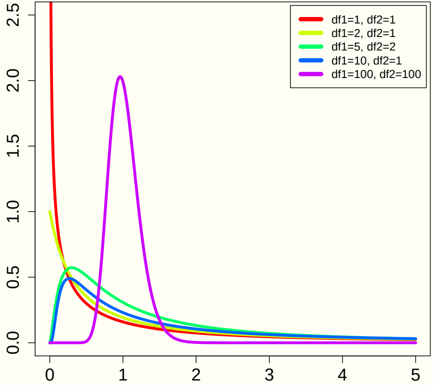
reject H_0 if $F^* > F(1-\alpha; 1, n-2)$,

where $F(1-\alpha; 1, n-2)$ is the $(1-\alpha)100$th percentile of the

$F_{1,n-2}$ distribution.

F Distributions

Figure: F distributions: probability density function



In simple linear regression, the F-test is equivalent to the two-sided t-test for testing $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$.

- $F^* = (T^*)^2$
- $F(1 - \alpha; 1, n - 2) = t^2(1 - \alpha/2; n - 2)$.

ANOVA Table for Simple Regression

Source of Variation	SS	d.f.	MS=SS/d.f.	F^*
Regression	$SSR = \sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2$	1	$MSR = SSR/1$	MSR/MSE
Error	$SSE = \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2$	$n - 2$	$MSE = SSE/(n-2)$	
Total	$SSTO = \sum_{i=1}^n (Y_i - \overline{Y})^2$	$n - 1$		

Heights

Source of Variation	SS	d.f.	MS=SS/d.f.	F^*
Regression	$SSR = 1234$	1	$MSR = 1234$	245
Error	$SSE = 4659$	926	$MSE = 5.03$	
Total	$SSTO = 5893$	927		

- Test whether there is a linear association between parent's height and child's height at significance level $\alpha = 0.01$.
- $F(0.99; 1, 926) = 6.66 < F^* = 245$, so reject $H_0 : \beta_1 = 0$ and conclude that there is a significant linear association between parent's height and child's height.

Coefficient of Determination

Coefficient of Determination R^2

A descriptive measure for **linear association** between X and Y:

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

- Heights: $R^2 = \frac{1234}{5893} = 0.209$. 20% of variation in child's height may be explained by the variation in parent's height.

Properties of R^2

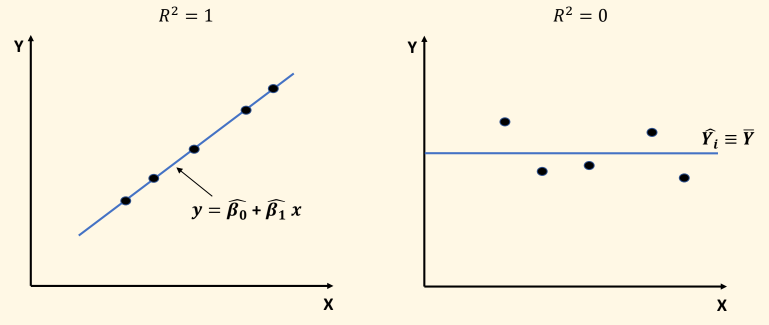
Special: when all Ys are the same, we can not calculate R2(SSTO=0) which is case 1+case 2

- $0 \leq R^2 \leq 1$.

case 1
- If all observations fall on one straight line, then $R^2 = 1$.
  - X accounts for all variation in the observations.

case 2
- If the fitted regression line is horizontal, i.e., $\hat{\beta}_1 = 0$, then $R^2 = 0$.
  - X is of no use in explaining variation in the observations.
  - There is no evidence of linear association between X and Y in the data.

Figure:



$R^2 = 1$
Y
X
$y = \widehat{\beta_0} + \widehat{\beta_1}\, x$
$R^2 = 0$
Y
X
$\widehat{Y}_i \equiv \overline{Y}$

Caution with Interpreting R^2

When the relationship between X and Y is nonlinear, R^2 is not a meaningful measure.

- *"A large R^2 means that the estimated regression line must be a good fit of the data". Not necessarily!*
- *"A near zero R^2 means that X and Y are not related". Not necessarily!*