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.)* **Normal**$(0, \sigma^2)$ random variables.

$$\Longrightarrow Y_i \sim_{independent} N(\beta_0 + \beta_1 X_i, \sigma^2).$$

Optional Reading: 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.

Sampling Distributions

Under the Normal error model:

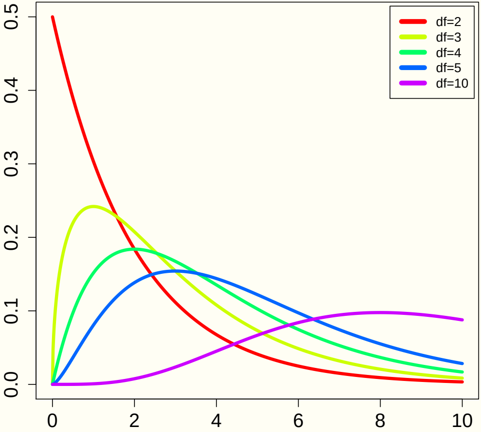
- $\hat{\beta}_0, \hat{\beta}_1$ are normally distributed (as they are linear combinations of independent Normal random variables, i.e., Y_is):

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

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

χ^2 Distributions

Figure: χ^2 distributions: probability density function



Confidence Intervals of Regression Coefficients

Confidence Intervals for β_1

Under the Normal error model, a $(1-\alpha)100\%$-confidence interval for β_1 is:

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

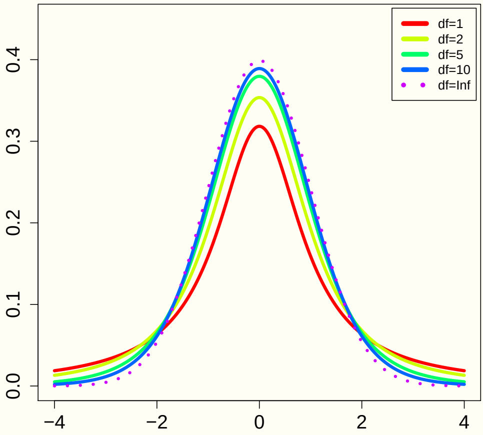
where $t(1-\alpha/2; n-2)$ is the $(1-\alpha/2)100$th percentile of $t_{(n-2)}$.

- The coverage probability is $1-\alpha$:

$$P(\beta_1 \in \hat{\beta}_1 \pm t(1-\alpha/2; n-2) \cdot s\{\hat{\beta}_1\}) = 1-\alpha$$

How to construct confidence intervals for β_0?

Figure: t distributions: probability density function[1]

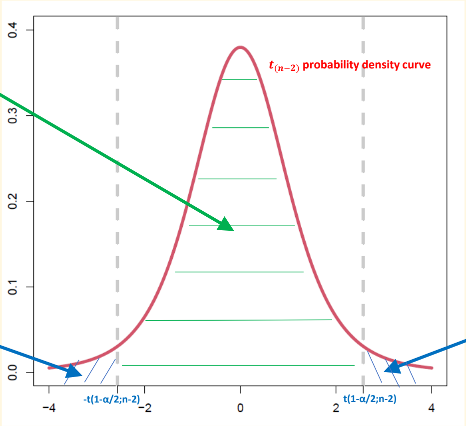


[1] t-distribution with df=∞ is the standard normal $N(0, 1)$ distribution.

$$\frac{\widehat{\beta_1}-\beta_1}{s\{\widehat{\beta_1}\}} \sim$$

Area under curve: 1 - α

Area under curve: α/2



Area under curve: α/2

Optional Reading: Deriving Confidence Intervals via Pivotal Quantity

Pivotal quantities are intermediate objects used in derivations of confidence intervals:

- They involve **both** observed data and unknown parameters, so they are **not** statistics themselves.
- They have known distributions.
- More in STA200B.

Look at the following quantity:

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

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

Remark: followed from the fact that if $Z \sim N(0, 1)$, $S^2 \sim \chi^2_{(k)}$ and Z, S^2 are independent, then $\frac{Z}{\sqrt{S^2/k}} \sim t_{(k)}$.

Confidence intervals can be derived from "inverting the region under the curve" :

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

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

Confidence Coefficient: Accuracy

- $(1-\alpha)100\%$ is called the *confidence coefficient* or the *confidence level/coverage*.
- 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

- The half-width: $t(1 - \alpha/2; n - 2)s\{\hat{\beta}_1\}$
- The width reflects **precision of the C.I.**: the narrower, the more precise
- 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

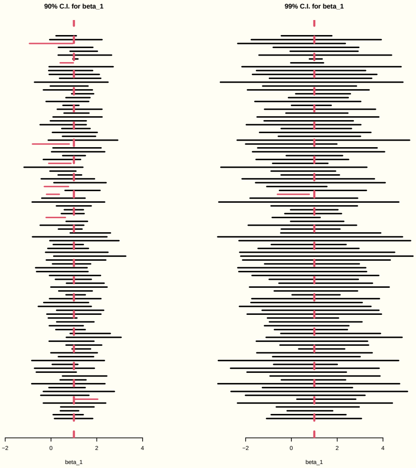


Figure: C.I.s of β_1: Left: 90% C.I.; Right: 99% C.I.

Reading: 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-tests for β_1

- Null hypothesis: $H_0 : \beta_1 = \beta_1^{(0)}$, where $\beta_1^{(0)}$ is a given constant (e.g., 0).
- **T-statistic**: derived from standardization of $\hat{\beta}_1$ under H_0:

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

- **Null distribution** of T^*:

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

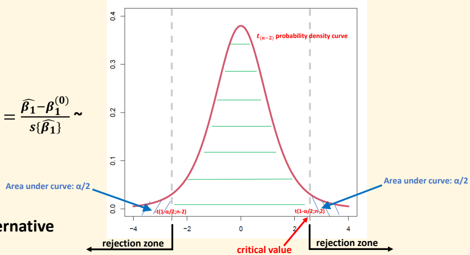
Decision Rules

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$:= P(t_{(n-2)} < T^*) < \alpha$.

What about the right-sided alternative? Why are the critical value approach and the pvalue approach equivalent? How to conduct hypothesis testing with regard to β_0?

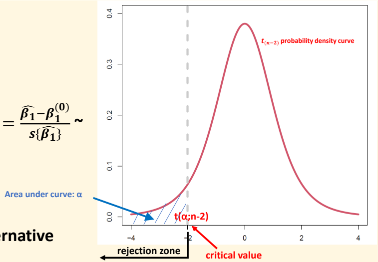
$$under\ H_0: T^* = \frac{\widehat{\beta_1} - \beta_1^{(0)}}{s\{\widehat{\beta_1}\}} \sim$$



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

Reading: 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

Consider an (arbitrary) X value, denoted by X_h. Assume that the model holds at X_h, i.e., $Y_h = \beta_0 + \beta_1 X_h + \epsilon_h$. Then the mean response at $X = X_h$ is $E(Y_h) = \beta_0 + \beta_1 X_h$. The goal is to estimate $E(Y_h)$.

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

- An unbiased estimator of $E(Y_h) = \beta_0 + \beta_1 X_h$ is:

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

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

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

- Under Normal error model $\widehat{Y}_h \sim N(E(Y_h), \sigma^2\{\widehat{Y}_h\})$.

Confidence Intervals for $E(Y_h)$

Under the Normal error model, a $(1-\alpha)100\%$ confidence interval for $E(Y_h)$:

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

- The coverage probability is $1-\alpha$:

$$P(E(Y_h) \in \widehat{Y}_h \pm t(1-\alpha/2; n-2) \cdot s(\widehat{Y}_h)) = 1-\alpha$$

Reading: Heights

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

- $n = 928,\ \overline{X} = 68.316,\ \ \sum_{i=1}^{n}(X_i - \overline{X})^2 = 3038.761$ and $\hat{\beta}_0 = 24.54,\ \hat{\beta}_1 = 0.637,\ MSE = 5.031$
- $\widehat{Y}_h = 24.54 + 0.637 \times 70 = 69.2$
- $s\{\widehat{Y}_h\} = \sqrt{5.031 \times \left\{\frac{1}{928} + \frac{(70-68.316)^2}{3038.761}\right\}} = 0.1$
- 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

Prediction of New Outcome

Suppose instead of estimating $E(Y_h)$, we would like to predict the outcome Y_h at $X = X_h$. We still assume the model holds at X_h, i.e., $Y_h = \beta_0 + \beta_1 X_h + \epsilon_h$.

- What would be the height of a (future) child of a specific $70in$ couple?

- We can predict Y_h by the estimated mean response at $X = X_h$:

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

- Note that $E(\widehat{Y}_h) = \beta_0 + \beta_1 X_h = E(Y_h)$, so $\widehat{Y}_h$ is an *unbiased predictor* of Y_h.

Prediction Intervals

How reliable is this prediction?

- We can use *prediction intervals* to answer this question.
- In order to derive prediction intervals, we need to further **assume that the error ϵ_h pertained to this specific case is uncorrelated with errors ϵ_is associated with the observations Y_is in the current data set**.
- This is a reasonable assumption if we are predicting a future outcome.

Under the Normal error model:

- $\widehat{Y}_h - Y_h \sim \text{Normal}(0, \sigma^2(pred_h))$, where

$$\begin{aligned}\sigma^2(pred_h) &:= Var(\widehat{Y}_h - Y_h) = \sigma^2(\widehat{Y}_h) + \sigma^2(Y_h)\\ &= \sigma^2(\widehat{Y}_h) + \sigma^2 = \sigma^2\left[\mathbf{1} + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]\end{aligned}$$

- Standard error of $\widehat{Y}_h - Y_h$ is then

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

Prediction Intervals

Under the Normal error model, a $(1-\alpha)100\%$ prediction interval for Y_h is:

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

- The coverage probability is $1-\alpha$:

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

Optional Reading: Deriving Prediction Intervals

The derivation of prediction intervals follows the same type of calculation for the derivation of C.Is through a pivotal quantity, namely, $\frac{\widehat{Y}_h - Y_h}{s(pred_h)} \sim t_{(n-2)}$.

Prediction vs. Estimation

- Y_h is a "moving target" (random variable) vs. $E(Y_h)$ is a fixed (non-random) quantity.
- There are two sources of variations in the prediction process: (i) variability from the predictor $\widehat{Y}_h$ due to sampling variability of the data; (ii) variability from the target Y_h.
- In contrast, there is only one source of variation in the estimation process, i.,e., variability from the estimator $\widehat{Y}_h$.
- At any given X value, the prediction process has intrinsically larger variability than the estimation process $\Longrightarrow$ prediction intervals are wider than the corresponding confidence intervals.

Reading: 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: $69.2 \pm 1.963 \times 2.25 = [64.78, 73.62]$
- We are 95% confident that the child's height will be between $[64.78in, 73.62in]$.

Analysis of Variance

Analysis of Variance

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

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:

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

  - *deviation of the observed value around the fitted regression line (residual)*;
  - *deviation of the fitted value from the sample mean*;

Decomposition of Total Variation

- Taking sum of squares of the total deviations and noting that the sum of the cross product terms equal to zero:

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

- 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:

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

Error Sum of Squares (SSE)

Quantify variation of the observations around the fitted regression line:

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

Regression Sum of Squares (SSR)

Quantify variation of the fitted values around the sample mean:

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

- $SSR = SSTO - SSE$: 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 the X values, the larger is SSR

Mean Squares

Sum of Squares divided by its degrees 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:

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

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

Sampling Distributions of SS

Under the Normal error model:

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

F Test for Linear Association between X and Y

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

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

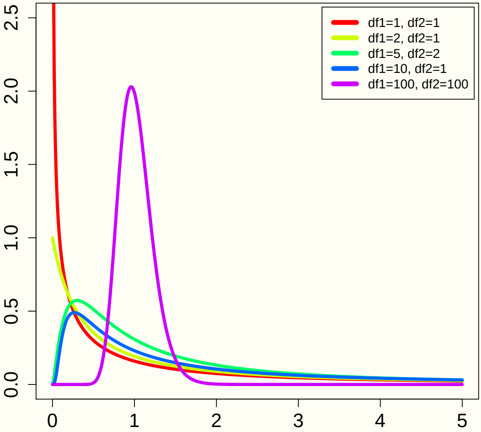
    where $F(1-\alpha; 1, n-2)$ is the $(1-\alpha)100$th percentile of the $F_{1,n-2}$ distribution.
  - P-value approach: reject H_0 if p-value$< \alpha$ where p-value$= P(F_{1,n-2} > F^*)$.

Optional Reading: Definition of F Distributions

If $Z_1 \sim \chi^2_{(df_1)}$, $Z_2 \sim \chi^2_{(df_2)}$ and Z_1, Z_2 are independent, then $\frac{Z_1/df_1}{Z_2/df_2} \sim F_{df_1,df_2}$.

F Distributions

Figure: F distributions: probability density function



Relationship between F Tests and T Tests

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

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

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	$SSTO = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$	$n-1$		

Reading: 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

R^2 is a descriptive measure for **linear association** between X and Y:

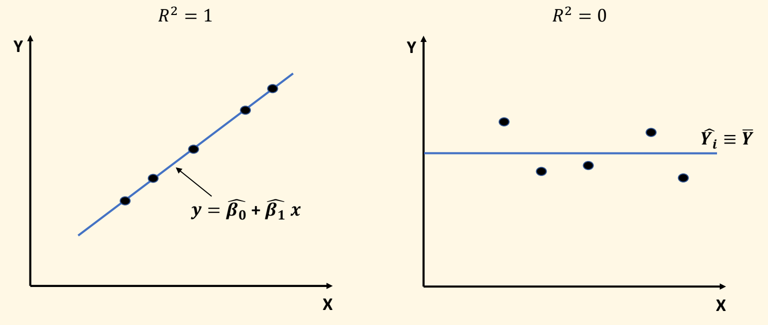
$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}.$$

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

Properties of R^2

- $0 \leq R^2 \leq 1$.
  - In simple linear regression, $R^2 = r_{Xy}^2$.
- If all observations fall on one straight line, then $R^2 = 1$.
  - X accounts for all variation in the observations.
- 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:

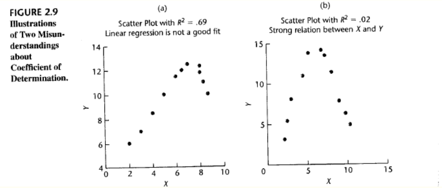

$R^2 = 1$
Y
X
$y = \widehat{\beta_0} + \widehat{\beta_1}\, x$
$R^2 = 0$
Y
X
$\hat{Y}_i \equiv \bar{Y}$

Caution when Interpreting R^2

Is any of the following statements true?

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

Figure:



Ref.: Applied Linear Statistical Models by Kutner, Nachtsheim, Neter and Li

If the relationship between X and Y is indeed linear, is any of the following statements true?

- *"A large R^2 means that there must be a (statistically) significant linear association between X and Y."*
- *"A near zero R^2 means that there is no (statistically) significant linear association between X and Y."*