Methods of Data Analysis I

Derek Li

Contents

1 Review

1.1 Expectation

- $\mathbb{E}[a] = a, a \in \mathbb{R}$.
- $\mathbb{E}[aY] = a\mathbb{E}[Y]$.
- $\mathbb{E}[X \pm Y] = \mathbb{E}[x] \pm \mathbb{E}[Y]$.
- $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ if X and Y are independent.
- Tower rule: $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$.

1.2 Variance and Covariance

- $\text{Var}[a] = 0, a \in \mathbb{R}$.
- $\text{Var}[aY] = a^2\text{Var}[Y]$.
- $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.
- $\text{Cov}(Y, Y) = \text{Var}[Y]$.
- $\text{Var}[Y] = \text{Var}[\mathbb{E}[Y|X]] + \mathbb{E}[\text{Var}[Y|X]]$.
- $\text{Var}[X \pm Y] = \text{Var}[X] + \text{Var}[Y] \pm 2\text{Cov}(X, Y)$.
- $\text{Cov}(X, Y) = 0$ if X and Y are independent.
- $\text{Cov}(aX + bY, cU + dW) = ac\text{Cov}(X, U) + ad\text{Cov}(X, W) + bc\text{Cov}(Y, U) + bd\text{Cov}(Y, W)$.

1.3 Correlation

If X and Y are random variables, a symmetric measure of the direction and strength of the linear dependence between them is their correlation

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}[X]\text{Var}[Y]}}.$$

1.4 Distributions

- If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$.
- Let $U = Z^2$, then $U \sim \chi^2_{(1)}$.
- If Z and $X \sim \chi^2_{(m)}$ are independent, then $\frac{Z}{\sqrt{X/m}} \sim t_{(m)}$.
- If $X \sim \chi^2_{(m)}, Y \sim \chi^2_{(n)}$ are independent, then $\frac{X/m}{Y/n} \sim F_{(m,n)}$.
- $t_{(m)} \xrightarrow{D} Z$, as $m \to \infty$.

1.4.1 Bivariate Normal Distribution

X and Y are jointly normally distributed is their joint density function is

$$f(x,y) = \frac{e^{-\frac{Q}{2}}}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}},$$

where

$$Q = \frac{1}{1-\rho^2}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right].$$

Two marginal distributions are

$$X \sim \mathcal{N}(\mu_x, \sigma_x^2) \text{ and } Y \sim \mathcal{N}(\mu_y, \sigma_y^2).$$

The conditional distribution of Y given $X = x$ is

$$Y|X = x \sim \mathcal{N}\left(\mu_y + \rho\sigma_y\left(\frac{x-\mu_x}{\sigma_x}\right), (1-\rho^2)\sigma_y^2\right).$$

Theorem 1.1. If X and Y are jointly normally distributed, then a zero covariance between X and Y implies that they are statistically independent.

2 Sample Linear Regression

2.1 Statistical Model

$$Y = \beta_0 + \beta_1 X + e,$$

where Y is dependent or response variable, X is independent or explanatory variable, β_0 is intercept parameter, β_1 is slope parameter, and e is random error or noise (variation in measures that we cannot account for).

Given a specific value of $X = x$, we want to find the expected value of Y

$$\mathbb{E}[Y|X = x].$$

2.2 Estimating β_0, β_1

Given n pairs bivariate data $(x_1, y_1), \cdots, (x_n, y_n)$, we want to use $\widehat{\beta}_0$ and $\widehat{\beta}$ to estimate β_0 and β_1.

Consider the residual sum of squares

$$RSS = \sum_{i=1}^{n} \widehat{e}_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} \left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i \right) \right]^2,$$

we can use least squares method that minimizes the criterion RSS to find the estimators.

2.2.1 Least Squares Method

Least squares method makes no statistical assumptions. We have

$$\frac{\partial RSS}{\partial \widehat{\beta}_0} = -2 \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \right) \text{ and } \frac{\partial RSS}{\partial \widehat{\beta}_1} = -2 \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \right) x_i.$$

Let $\frac{\partial RSS}{\partial \widehat{\beta}_0}$ and $\frac{\partial RSS}{\partial \widehat{\beta}_1}$ be 0, we get the normal equations

$$\sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \right) = 0 \text{ and } \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \right) x_i = 0.$$

Therefore, we have

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \widehat{\beta}_0 - \sum_{i=1}^{n} \widehat{\beta}_1 x_i = n\overline{y} - n\widehat{\beta}_0 - n\widehat{\beta}_1 \overline{x} = 0 \Rightarrow \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}.$$

Besides,

$$\begin{aligned} \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \widehat{\beta}_0 x_i - \sum_{i=1}^{n} \widehat{\beta}_1 x_i^2 &= \sum_{i=1}^{n} x_i y_i - \left(\overline{y} - \widehat{\beta}_1 \overline{x} \right) \sum_{i=1}^{n} x_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\ &= \sum_{i=1}^{n} x_i y_i - n\overline{xy} + n\widehat{\beta}_1 \overline{x}^2 - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0, \end{aligned}$$

i.e.,

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} := \frac{SXY}{SXX}.$$

2.2.2 Interpretation

$\widehat{\beta}_0$: The expected value of y when $x = 0$. No practical interpretation unless 0 is within the range of the predictor values.

$\widehat{\beta}_1$: When x changes by 1 unit, the corresponding average change in y is the slope.

2.2.3 Estimation in R

```
model=lm(y~x)
summary(model)
```

2.3 Properties of Fitted Regression Line

Property 2.1.

$$\sum_{i=1}^{n} \widehat{e}_i = 0.$$

Proof. By definition,

$$\begin{aligned}\sum_{i=1}^{n} \widehat{e}_i &= \sum_{i=1}^{n}(y_i - \widehat{y}_i) = \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right) = \sum_{i=1}^{n} \left(y_i - \overline{y} + \widehat{\beta}_1 \overline{x} - \widehat{\beta}_1 x_i\right) \\ &= n\overline{y} - n\overline{y} + n\widehat{\beta}_1\overline{x} - n\widehat{\beta}_1\overline{x} = 0.\end{aligned}$$

□

Property 2.2. The sum of squares of residuals is not 0 unless the fit to the data is perfect.

Property 2.3.

$$\sum_{i=1}^{n} \widehat{e}_i x_i = 0.$$

Proof. By definition,

$$\begin{aligned}\sum_{i=1}^{n} \widehat{e}_i x_i &= \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right) x_i = \sum_{i=1}^{n} x_i y_i - \overline{y}\sum_{i=1}^{n} x_i + \widehat{\beta}_1 \overline{x} \sum_{i=1}^{n} x_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\ &= \sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y} - \widehat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) = 0.\end{aligned}$$

□

Property 2.4.

$$\sum_{i=1}^{n} \widehat{e}_i \widehat{y}_i = 0.$$

Proof. By definition,

$$\sum_{i=1}^{n} \widehat{e}_i \widehat{y}_i = \sum_{i=1}^{n} \widehat{e}_i (\widehat{\beta}_0 + \widehat{\beta}_1 x_i) = \widehat{\beta}_0 \sum_{i=1}^{n} \widehat{e}_i + \widehat{\beta}_1 \sum_{i=1}^{n} \widehat{e}_i x_i = 0 + 0 = 0.$$

$\square$

Property 2.5.

$$\sum_{i=1}^{n} \widehat{y}_i = \sum_{i=1}^{n} y_i.$$

Proof. We have

$$\sum_{i=1}^{n} \widehat{e}_i = 0 = \sum_{i=1}^{n} (y_i - \widehat{y}_i) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \widehat{y}_i \Rightarrow \sum_{i=1}^{n} \widehat{y}_i = \sum_{i=1}^{n} y_i.$$

$\square$

2.4 Assumptions

The Gauss-Markov conditions are:

1. $\mathbb{E}[e_i] = 0$.

2. $\mathrm{Var}[e_i] = \sigma^2$, i.e., homoscedastic.

3. The errors are uncorrelated or $\mathrm{Cov}(e_i, e_j) = \rho(e_i, e_j) = 0$.

Theorem 2.1 (Gauss-Markov Theorem)**.** Under the conditions or the simple linear regression model, the least-squares parameter estimators are best linear unbiased estimators.

We assume that Y is relate to x by the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + e_i, i = 1, \cdots, n.$$

Under the conditions we have

$$\mathbb{E}[Y|X = x_i] = \beta_0 + \beta_1 x_i$$

and

$$\mathrm{Var}[Y|X = x_i] = \mathrm{Var}[\beta_0 + \beta_1 x_i + e_i|X = x_i] = \mathrm{Var}[e_i] = \sigma^2.$$

2.5 Estimating the Variance of the Random Error Term

The variance σ^2 is another parameter of the SLR model and we want to estimate σ^2 to measure the variability of our estimates of Y, and carry out inference on the model.

An unbiased estimate of σ^2 is

$$S^2 = \frac{\sum_{i=1}^{n} \widehat{e}_i^2}{n-2} = \frac{RSS}{n-2}.$$

2.6 Properties of Least Squares Estimators

Since $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$,

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n}(x_i - \overline{x})y_i - \overline{y}\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n}(x_i - \overline{x})y_i.$$

Let $c_i = \frac{x_i - \overline{x}}{SXX}$, we can rewrite $\widehat{\beta}_1$ as

$$\widehat{\beta}_1 = \sum_{i=1}^{n} c_i y_i,$$

which is a linear combination of y_i.

We have

$$\begin{aligned}
\mathbb{E}\left[\widehat{\beta}_1|X\right] &= \mathbb{E}\left[\sum_{i=1}^{n} c_i y_i | X = x_i\right] = \sum_{i=1}^{n} c_i \mathbb{E}[y_i|X = x_i] \\
&= \sum_{i=1}^{n} c_i \mathbb{E}[\beta_0 + \beta_1 x_i] = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i \\
&= \frac{\beta_0}{SXX}\sum_{i=1}^{n}(x_i - \overline{x}) + \beta_1 \sum_{i=1}^{n}\frac{(x_i - \overline{x})x_i}{SXX} \\
&= \beta_1 \frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{SXX} = \beta_1.
\end{aligned}$$

Therefore, $\widehat{\beta}_1$ is unbiased for β_1. Besides,

$$\begin{aligned}
\mathrm{Var}\left[\widehat{\beta}_1|X\right] &= \mathrm{Var}\left[\sum_{i=1}^{n} c_i y_i | X\right] = \sum_{i=1}^{n} c_i^2 \mathrm{Var}[y_i|X = x_i] \\
&= \sigma^2 \sum_{i=1}^{n} c_i^2 = \sigma^2 \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{SXX^2} = \frac{\sigma^2}{SXX}.
\end{aligned}$$

We have

$$\begin{aligned}
\mathbb{E}\left[\widehat{\beta}_0|X\right] &= \mathbb{E}\left[\overline{y} - \widehat{\beta}_1\overline{x}|X = x_i\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} y_i - \widehat{\beta}_1\overline{x}|X = x_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\beta_0 + \beta_1 x_i + e_i|X = x_i] - \overline{x}\mathbb{E}\left[\widehat{\beta}_1|X = x_i\right] \\
&= \frac{1}{n}n\beta_0 + \frac{1}{n}n\beta_1\overline{x} - \overline{x}\beta_1 = \beta_0.
\end{aligned}$$

Therefore, $\widehat{\beta}_0$ is unbiased for β_0. Besides,

$$\begin{aligned}
\mathrm{Var}\left[\widehat{\beta}_0|X\right] &= \mathrm{Var}\left[\overline{y} - \widehat{\beta}_1\overline{x}|X = x_i\right] \\
&= \mathrm{Var}\left[\overline{y}|X = x_i\right] + \mathrm{Var}\left[\widehat{\beta}_1\overline{x}|X = x_i\right] - 2\mathrm{Cov}\left(\overline{y}, \widehat{\beta}_1\overline{x}|X = x_i\right) \\
&= \frac{\sigma^2}{n} + \frac{\overline{x}^2\sigma^2}{SXX} - 0 = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{SXX}\right).
\end{aligned}$$

Note that $\text{Cov}\left(\overline{y}, \widehat{\beta}_1 \overline{x} | X = x_u\right) = \frac{\overline{x}\sigma^2}{n} \sum_{i=1}^{n} c_i = 0$.

2.7 Normal Error Regression Model

Given distributional assumption:

$$e_i \sim \mathcal{N}(0, \sigma^2),$$

we know:

(1) the errors are independent since $\rho = 0$;

(2) since $y_i = \beta_0 + \beta_1 x_i + e_i$, then $Y_i | X \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$;

(3) the least squares estimates of β_0, β_1 are equivalent to their maximum likelihood estimators.

(4) since $\widehat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$ is a linear combination of the y_i's, $\widehat{\beta}_1 | X \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{SXX}\right)$; since $\overline{y}$ is normally distributed, $\widehat{\beta}_0 | X \sim \mathcal{N}\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{SXX}\right)\right)$.

Property 2.6. Under the normal error SLR model, where

$$e_i \sim \mathcal{N}(0, \sigma^2) \text{ and } S^2 = \frac{1}{n-2} \sum_{i=1}^{n} \widehat{e}_i^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2,$$

we have

$$\frac{(n-2)S^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{Y_i - \widehat{Y}_i}{\sigma^2}\right)^2 \sim \chi^2_{(n-2)}.$$

Property 2.7. Under the normal error SLR model,

$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{S^2}{SXX}}} \sim t_{(n-2)}.$$

Proof. We have $\widehat{\beta}_1 | X = x_i \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{SXX}\right)$, and thus

$$\frac{\widehat{\beta}_1 - \beta_1}{\sigma / \sqrt{SXX}} \sim \mathcal{N}(0, 1).$$

Wherefore

$$\frac{\frac{\widehat{\beta}_1 - \beta_1}{\sigma/\sqrt{SXX}}}{\sqrt{(n-2)S^2/\sigma^2/(n-2)}} = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{S^2}{SXX}}} \sim t_{(n-2)}.$$

□

2.8 Inference for the Parameter

2.8.1 Significance Test

- Step 1: $H_0 : \beta_1 = \beta_1^0$ against $H_a : \beta_1 \neq \beta_1^0$.
- Step 2: Test statistic $t = \frac{\widehat{\beta}_1 - \beta_1^0}{\sqrt{S^2/SXX}}$, and under $H_0, t \sim t_{(n-2)}$.
- Step 3: $p-$value $= 2P(t_{(n-2)} \geqslant |t|)$.
- Step 4: The smaller the p-value, the greater the evidence against H_0 and the larger p-value indicate that the data is consistent with H_0.

p-value	Evidence against H_0
< 0.001	Very strong
(0.001, 0.01)	Strong
(0.01, 0.05)	Moderate
(0.05, 0.1)	Weak
> 0.1	None

Note that the test statistic for $\widehat{\beta}_0$ is $t = \frac{\widehat{\beta}_0 - \beta_0^0}{\sqrt{S^2\left(\frac{1}{n} + \frac{\overline{x}^2}{SXX}\right)}}$

2.8.2 Confidence Interval

The CI is

$$\text{Estimate} \pm 100\left(1 - \frac{\alpha}{2}\right) \text{th quantile} \times \text{Standard Error (Estimate)},$$

where α is the critical value.

For β_1, the CI is

$$\left[\widehat{\beta}_1 \pm t_{\frac{\alpha}{2}(n-2)}\sqrt{\frac{S^2}{SXX}}\right].$$

For β_0, the CI is

$$\left[\widehat{\beta}_0 \pm t_{\frac{\alpha}{2}(n-2)}\sqrt{S^2\left(\frac{1}{n} + \frac{\overline{x}^2}{SXX}\right)}\right].$$

Note that a $100(1-\alpha)\%$ CI for θ consists of all those values of θ_0 for which $H_0 : \theta = \theta_0$ will not be rejected at level α. In other words, we do not reject H_0 is θ_0 lies within the CI, and we reject H_0 is the CI does not include θ_0.

2.9 The Pooled Two-Sample t-Procedure

We want to test $H_0 : \mu_x = \mu_y$, where

$$X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_x, \sigma_x^2) \text{ and } Y_1, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_y, \sigma_y^2).$$

Suppose two samples are independent and $\sigma_x^2 = \sigma_y^2 = \sigma^2$, then we have

$$t = \frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{s_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \sim t_{(n_x+n_y-2)},$$

where $s_p^2 = \frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_x+n_y-2}$.

2.10 Regression Analysis of Variance

Notice that $y_i - \overline{y} = (y_i - \widehat{y}_i) + (\widehat{y}_i - \overline{y})$. We have

$$\begin{aligned} TSS &= \sum_i^n (y_i - \overline{y})^2, \\ RSS &= \sum_i^n (y_i - \widehat{y}_i)^2 = \sum_i^n \widehat{e}_i^2, \\ RegSS &= \sum_i^n (\widehat{y}_i - \overline{y})^2. \end{aligned}$$

RSS, residual SS, is the least square criterion, representing the unexplained variation in y's. $RegSS$, regression SS, is the amount of variation in y's explained by regression line.

Property 2.8. *$RegSS = \widehat{\beta}_1^2 SXX$.*

Proof. We have

$$\begin{aligned} RegSS &= \sum_i^n (\widehat{y}_i - \overline{y})^2 = \sum_i^n (\widehat{\beta}_0 + \widehat{\beta}_1 x_i - \overline{y})^2 \\ &= \sum_i^n (\overline{y} - \widehat{\beta}_1 \overline{x} + \widehat{\beta}_1 x_i - \overline{y})^2 = \widehat{\beta}_1^2 \sum_i^n (x_i - \overline{x})^2 = \widehat{\beta}_1^2 SXX. \end{aligned}$$

□

Property 2.9. *$TSS = RSS + RegSS$.*

Proof. We have

$$\begin{aligned} \sum_i^n (y_i - \overline{y})^2 &= \sum_i^n ((y_i - \widehat{y}_i) + (\widehat{y}_i - \overline{y}))^2 \\ &= \sum_i^n (y_i - \widehat{y})^2 + \sum_i^n (\widehat{y}_i - \overline{y})^2 + 2\sum_i^n (y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}) \\ &= RSS + RegSS + 2\sum_i^n \widehat{e}_i(\widehat{y}_i - \overline{y}) \\ &= RSS + RegSS + 2\sum_i^n \widehat{e}_i \widehat{y}_i - 2\overline{y}\sum_i^n \widehat{e}_i \\ &= RSS + RegSS. \end{aligned}$$

□

2.10.1 Regression ANOVA Table

Source	SS	df	Mean SS
Regression Line	$RegSS = \widehat{\beta}_1^2 SXX$	1	$\widehat{\beta}_1^2 SXX$
Error	$RSS = \sum_{i=1}^{n} \widehat{e}_i^2$	$n-2$	$\frac{\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}{n-2} = S^2$
Total	$TSS = \sum_{i}^{n}(y_i - \overline{y})^2$		

Property 2.10. Let

$$F = \frac{MRegSS}{MRSS} = \frac{RegSS/1}{RSS/(n-2)}.$$

If $\beta_1 = 0$, then

$$F \sim F_{(1,n-2)}.$$

Proof. If $\beta_1 = 0$, then $\widehat{\beta}_1 \sim \mathcal{N}\left(0, \frac{\sigma^2}{SXX}\right)$, i.e.,

$$\frac{\widehat{\beta}_1}{\sqrt{\sigma^2/SXX}} \sim \mathcal{N}(0,1) \Rightarrow \frac{\widehat{\beta}_1^2}{\sigma^2/SXX} \sim \chi^2_{(1)}.$$

Besides, $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{(n-2)}$, and we have

$$\frac{\frac{\widehat{\beta}_1^2}{\sigma^2/SXX}}{\frac{(n-2)S^2}{\sigma^2}/(n-2)} = \frac{\widehat{\beta}_1^2 SXX}{S^2} = F \sim F_{(1,n-2)}.$$

□

Note that F is another test of $H_0 : \beta_1 = 0$, and in R, we have:

```
anova(model)
```

2.10.2 Coefficient of Determination

Let

$$R^2 = \frac{RegSS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}.$$

Here are some comments about R^2:

- $R^2 \in [0,1]$.
- R^2 gives percentage of variation in y's explained by regression line.
- R^2 is not resistant to outliers.
- A high R^2 does not indicate that the estimated regression line is a good fit since:
  - ∗ we do not have absolute rules about how large it should be;
  - ∗ R^2 can get very high by overfitting.

- It is not meaningful for models without intercept.
- To compare 2 models, R^2 is only useful:
  * same observations, y's in original units (not transformed);
  * one set of predictor variables is a subset of the other.

2.10.3 Sample Correlation Coefficient

The estimate of the population correlation is Pearson's Product-Moment Correlation Coefficient

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{SXY}{\sqrt{SXX \cdot SYY}},$$

which is the MLE of ρ. r is distribution free and is always a number between -1 and 1.

Theorem 2.2. $R^2 = r^2$.

Proof. We have

$$R^2 = \frac{RegSS}{TSS} = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{\widehat{\beta}_1^2 SXX}{SYY} = \frac{\frac{SXY^2}{SXX^2} \cdot SXX}{SYY} = \frac{SXY^2}{SXX \cdot SYY} = r^2.$$

□

Property 2.11. If $\rho = 0$,

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{\widehat{\beta}_1}{\sqrt{S^2/SXX}} \sim t_{(n-2)},$$

where $\widehat{\beta}_1$ is the slope estimate for the normal error SLR model.

Proof. Since $r^2 = R^2$, then

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{\frac{\widehat{\beta}_1\sqrt{SXX}}{\sqrt{SXY}}\sqrt{n-2}}{\sqrt{(n-2)S^2/SXY}} = \frac{\widehat{\beta}_1}{\sqrt{S^2/SXX}}.$$

If $\rho = 0$, then $\beta_1 = 0$, i.e.,

$$\frac{\widehat{\beta}_1}{\sqrt{S^2/SXX}} \sim t_{(n-2)}.$$

□

2.11 Confidence Interval for the Population Regression Line

We want to find a CI for the unknown population regression line at a given value of X, denoted by x^*, i.e.,

$$\mathbb{E}[Y|X = x^*] = \beta_0 + \beta_1 x^*.$$

The point estimate for $\mathbb{E}[Y|X = x^*]$ is

$$\widehat{y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x^*.$$

We have

$$\mathbb{E}\left[\widehat{y}^*\right] = \mathbb{E}\left[\widehat{y}|X = x^*\right] = \beta_0 + \beta_1 x^*,$$

i.e., $\widehat{y}^*$ is unbiased for $\mathbb{E}[Y|X = x^*]$.

Recall that $\text{Var}\left[\widehat{\beta}_0|X\right] = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{SXX}\right), \text{Var}\left[\widehat{\beta}_1|X\right] = \frac{\sigma^2}{SXX}$, then

$$\text{Cov}\left[\widehat{\beta}_0, \widehat{\beta}_1|X\right] = \text{Cov}\left[\overline{y} - \widehat{\beta}_1\overline{x}, \widehat{\beta}_1|X\right] = -\overline{x}\text{Var}\left[\widehat{\beta}_1|X\right] = -\frac{\overline{x}\sigma^2}{SXX}.$$

Wherefore

$$\begin{aligned}
\text{Var}\left[\widehat{y}^*\right] &= \text{Var}\left[\widehat{y}|X = x^*\right] = \text{Var}\left[\widehat{\beta}_0 + \widehat{\beta}_1 x|X = x^*\right] \\
&= \text{Var}\left[\widehat{\beta}_0|X = x^*\right] + (x^*)^2\text{Var}\left[\widehat{\beta}_1|X = x^*\right] + 2x^*\text{Cov}\left[\widehat{\beta}_0, \widehat{\beta}_1|X = x^*\right] \\
&= \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{SXX}\right) + (x^*)^2\frac{\sigma^2}{SXX} - \frac{2x^*\overline{x}\sigma^2}{SXX} = \sigma^2\left(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{SXX}\right).
\end{aligned}$$

Hence, as $n \uparrow, \text{Var}\left[\widehat{y}^*\right] \downarrow$; as x^* closer to $\overline{x}, \text{Var}\left[\widehat{y}^*\right] \downarrow$.

Using $S^2 = MRSS$, we get the standard error of the estimate of $\mathbb{E}[Y|X = x^*]$,

$$\sqrt{S^2\left(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{SXX}\right)}.$$

Hence, a $100(1-\alpha)\%$ CI for $\mathbb{E}[Y|X = x^*]$, the mean response for all the elements in the population with $X = x^*$ is

$$\left[\widehat{y}^* \pm t_{\frac{\alpha}{2}(n-2)}\sqrt{S^2\left(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{SXX}\right)}\right].$$

Notice that it is only valid for x^* in the range of the original data values of X but not for extrapolation.

2.12 Prediction Interval for Actual Value of Y

A confidence interval is always reported for a parameter while a prediction interval is reported for the value of a random variable. We want to find a PI for the actual value of Y at $X = x^*$, i.e., $Y^* = Y|X = x^*$.

The point estimate for Y^* is

$$\widehat{y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x^*.$$

The error in our prediction is

$$\varepsilon^* = Y^* - \widehat{y}^*.$$

The predicted value $\widehat{y}^*$ has two sources of variability:

- Since the regression line is estimated at $\widehat{\beta}_0 + \widehat{\beta}_1 X$;
- due to ε^*, some points do not fall exactly on the line.

We have

$$\begin{aligned}
\mathrm{Var}\left[Y^* - \widehat{y}^*\right] &= \mathrm{Var}\left[Y - \widehat{y}|X = x^*\right] \\
&= \mathrm{Var}[Y|X = x^*] + \mathrm{Var}\left[\widehat{y}|X = x^*\right] - 2\mathrm{Cov}(Y, \widehat{y}|X = x^*) \\
&= \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{SXX}\right) - 0 = \sigma^2\left(1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{SXX}\right).
\end{aligned}$$

Notice that $\mathrm{Cov}(Y, \widehat{y}|X = x^*) = 0$ since Y^* is a new observation.

Hence, a $100(1-\alpha)\%$ PI for $Y|X = x^*$ is

$$\left[\widehat{y}^* \pm t_{\frac{\alpha}{2}(n-2)}\sqrt{S^2\left(1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{SXX}\right)}\right].$$

PIs for Y^* have the same center but are wider than CIs for $\mathbb{E}[Y|X = x^*]$.