Inferences in Regression and Correlation Analysis

2.1 Inference Concerning β_1

Definition. *Inference concerning* β_1 *Test concerning β_1 is of the form*

$$\begin{cases} \mathcal{H}_0 : \beta_1 = 0 \\ \mathcal{H}_\alpha : \beta_1 \neq 0 \end{cases}$$

$\beta_1 = 0$ implies that there is no linear association between Y and X. On assumption of gaussian noise, there is also no relation of any type between Y and X, since probability of Y are identical for all levels of X

Definition. *Linear estimators* *Least square estimator $\hat{\beta}_1$ is a linear estimator*

$$\hat{\beta}_1 = \sum_i k_i y_i \qquad k_i = \frac{x_i - \overline{x}}{\sum_j (x_j - \overline{x})^2}$$

with properties

$$\sum k_i = 0 \quad \sum k_i x_i = 1 \quad \sum k_i^2 = \frac{1}{S_{XX}}$$

Proof. Note

$$\sum (x_i - \overline{x})(y_i - \overline{y}) = \sum (x_i - \overline{x})y_i - \sum (x_i - \overline{x})\overline{y} = \sum (x_i - \overline{x})y_i \quad \text{by } \sum (x_i - \overline{x}) = 0$$

the result follows. Proof of properties are simple, i.e.

$$\sum k_i = \sum_i \left(\frac{x_i - \overline{x}}{\sum_j (x_j - \overline{x})^2} \right) = \frac{1}{\sum_j (x_j - \overline{x})^2} \sum_i (x_i - \overline{x}) = 0$$

For second property

$$\sum k_i x_i = \frac{1}{\sum_j (x_j - \overline{x})^2} \sum_i (x_i - \overline{x})x_i = \frac{1}{S_{XX}} \sum_i x_i^2 - n\overline{x}^2 = \frac{1}{S_{XX}} S_{XX} = 1$$

$\square$

Definition. *Sampling distribution of* $\hat{\beta}_1$ *The sampling distribution of $\hat{\beta}_1$ refers to different values of the estimator obtained with repeated sampling when the levels of the predictor variable X are held constant from sample to sample. Given $\hat{\beta}_1$*

$$\hat{\beta}_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$$

*The sampling distribution of $\hat{\beta}_1$ is **normal** with mean and variance,*

$$\mathbb{E}(\hat{\beta}_1 | X) = \beta_1 \quad Var(\hat{\beta}_1 | X) = \frac{\sigma^2}{\sum (x_i - \overline{x})^2} = \frac{\sigma^2}{S_{XX}}$$

$$\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \frac{\sigma^2}{S_{XX}})$$

Proof.

1. **Mean and variance**

$$\mathbb{E}(\hat{\beta}_1) = \mathbb{E}\{\sum k_i y_i\} \stackrel{ind}{=} \sum (k_i \mathbb{E}\{y_i\}) = \sum k_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum k_i + \beta_1 \sum k_i x_i = \beta_1$$

$$Var(\hat{\beta}_1) = Var\left(\sum k_i y_i\right) = \sum k_i^2 Var(y_i) = \sum k_i^2 \sigma^2 = \sigma^2 \sum k_i^2 = \frac{\sigma^2}{S_{XX}}$$

   with last step of both derivation given by properties of $\hat{\beta}_1$ as a linear estimator.

2. **Normality** of sampling distribution given by the fact that $\hat{\beta}_1$ is a linear combination of y_is. Since $y_i \stackrel{i.i.d}{\sim} \mathcal{N}(\beta_0 + \beta 1 x_i, \sigma^2)$, the estimator is normaly distributed because it is a linear combination of independent normal random variables

3. **Estimated variance** We can estimate the variance of $\hat{\beta}_1$ by substituting σ^2 (unknown) with its unbiased estimator MSE s^2

$$s^2(\hat{\beta}_1) = \frac{MSE}{S_{XX}} \quad \text{where} \quad MSE = \frac{\sum e_i^2}{n-2}$$

   Note s^2 carries a denominator of S_{XX}, this is from the variance of $\hat{\beta}_1$, we are simply substituting the unknown σ^2 with MSE

   $\square$

Definition. ***standardization of*** $\hat{\beta}_1$
Standardization of sampling distribution of $\hat{\beta}_1$ gives

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma(\hat{\beta}_1)} = \mathcal{N}(0,1) \quad wher \quad \sigma(\hat{\beta}_1) = \frac{\sigma}{\sqrt{S_{XX}}}$$

*Usually, have to estimate standard error $\sigma(\hat{\beta}_1)$ with $s(\hat{\beta}_1)$. Standardization where the denominator is an estimated standard error is called **studentized statistic**, given by*

$$T = \frac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)} \sim t_{n-2} \quad where \quad s^2(\hat{\beta}_1) = \frac{MSE}{S_{XX}}$$

Proof. Assume proposition

$$\frac{\sum(\hat{e}_i)^2}{\sigma^2} \sim \chi^2_{n-2}$$

Then we have

$$\frac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{\sigma(\hat{\beta}_1)} \bigg/ \frac{s(\hat{\beta}_1)}{\sigma(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{\sigma(\hat{\beta}_1)} \bigg/ \sqrt{\frac{\sum e_i^2 S_{XX}}{(n-2)S_{XX}\sigma^2}} \sim \frac{Z}{\sqrt{\chi^2_{n-2}\big/(n-2)}} = t_{n-2}$$

   $\square$

Definition. *Confidence Interval and test for* β_1
We use the previously derived distribution as a pivot

$$\Pr\left(t_{\alpha/2,n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)} \leq t_{1-\alpha/2,n-2}\right) = 1 - \alpha$$

$$\Pr\left(\hat{\beta}_1 - s(\hat{\beta}_1)t_{1-\alpha/2,n-2} \leq \beta_1 \leq \hat{\beta}_1 + s(\hat{\beta}_1)t_{1-\alpha/2,n-2}\right) = 1 - \alpha$$

Hence the confidence interval is given by

$$\hat{\beta}_1 \pm s(\hat{\beta}_1)t_{1-\alpha/2,n-2} \quad where \quad s^2(\hat{\beta}_1) = \frac{MSE}{S_{XX}}$$

For 2-sided tests

$$\begin{cases} \mathcal{H}_0 : \beta_1 = b \\ \mathcal{H}_\alpha : \beta_1 \neq b \end{cases}$$

We compute test statistics

$$t^* = \frac{\hat{\beta}_1 - b}{s(\hat{\beta}_1)} \quad and \ reject \ \mathcal{H}_0 \ if \ |t^*| > t_{1-\alpha/2,n-2}$$

2.2 Inference Concerning of $\hat{\beta}_0$

Definition. *Sampling Distribution of* $\hat{\beta}_0$
Given point estimator

$$\hat{\beta}_0 = \overline{y} - \beta_1\overline{x}$$

$\hat{\beta}_0$ *refers to different values of* β_0 *that would be obtained with repeated sampling when levels of predictor variable* x *are held constant from sample to sample. The* **sampling distribution of** $\hat{\beta}_0$ *is* **normal** *with mean and variance*

$$\mathbb{E}(\hat{\beta}_0) = \beta_0 \qquad Var(\hat{\beta}_0) = \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{\sum(x_i - \overline{x})^2}\right] = \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}}\right]$$

Proof.

$$\mathbb{E}(\hat{\beta}_0) = \mathbb{E}(\overline{y}) - \mathbb{E}(\hat{\beta}_1\overline{x}) = \frac{1}{n}\sum \mathbb{E}(y_i) - \beta_1\overline{x} = \beta_0 + \beta_1\frac{1}{n}\sum x_i - \beta_1\overline{x} = \beta_0$$

$$
\begin{aligned}
Var(\hat{\beta}_0) &= Var(\overline{y} - \hat{\beta}_1 \overline{x}) \\
&= Var(\overline{y}) + \overline{x}^2 Var(\hat{\beta}_1) - 2\overline{x}Cov(\overline{y}, \hat{\beta}_1) \\
&= Var(\frac{1}{n}\sum y_i) + \frac{\sigma^2}{S_{XX}} - 2\overline{x}Cov(\frac{1}{n}\sum y_i, \sum_i k_i y_i) \\
&= \frac{\sigma^2}{n^2} + \frac{\sigma^2}{S_{XX}} - 2\overline{x}\frac{1}{n}\sum k_i Cov(y_i, y_i) \\
&= \frac{\sigma^2}{n^2} + \frac{\sigma^2}{S_{XX}} - 2\overline{x}\frac{\sigma^2}{n}\sum k_i \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}} \right)
\end{aligned}
$$

The normality follows because $\hat{\beta}_0$ is also a linear estimator of y_is. We can estimate $Var(\hat{\beta}_0)$ by replacing σ^2 with MSE, as before

$$
s^2(\hat{\beta}_0) = MSE \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}} \right]
$$

$\square$

Definition. ***Standardization of*** $\hat{\beta}_0$

$$
Z = \frac{\hat{\beta}_0 - \beta_0}{\sigma(\hat{\beta}_0)} \sim t_{n-2} \quad where \quad \sigma^2(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}} \right]
$$

$$
T = \frac{\hat{\beta}_0 - \beta_0}{s(\hat{\beta}_0)} \sim t_{n-2} \quad where \quad s^2(\hat{\beta}_0) = MSE \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}} \right]
$$

Definition. ***Confidence Interval for*** $\hat{\beta}_0$

$$
\hat{\beta}_0 \pm s(\hat{\beta}_0) t_{1-\alpha/2, n-2} \quad where \quad s(\hat{\beta}_0) = \sqrt{\frac{\sum e_i^2}{n-2}} \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}}}
$$

Definition. ***Consideration about inference***

1. ***Spacing of*** x ***levels*** Looking at variance of $\hat{\beta}_0$ and $\hat{\beta}_1$, the larger the spread in x levels, the larger S_{XX} and smaller the variance.

2.4 Interval Estimation of $\mathbb{E}(Y|X = x^*)$, the population regression line

Definition. ***Mean estimation*** Often times want to estimate mean for one or more probability distribution of Y. (i.e. mean Y for low and high X levels). Let x^* be level of X for which we want to estimate the mean response $\hat{y}^*$. The mean response is given by

$$
\mathbb{E}(Y|X = x^*) = E(y^*) = \beta_0 + \beta_1 x^*
$$

The idea is that the expectation is a random variable because of the estimated correlation coefficients. We would want to do inference on the mean response. We have a point estimator of the the mean response

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

which is simply the estimated response from estimated correlation coefficients and regression function. Note $X = x^*$ is a known constant

Definition. *Sampling Distribution of mean response estimator $\hat{y}^*$*
*The sampling distribution of $\hat{y}^*$ is **normal** with mean and variance*

$$\mathbb{E}(\hat{y}^*) = \mathbb{E}(\hat{y}|X = x^*) = \beta_0 + \beta_1 x^* \quad (= \mathbb{E}(y^*) \text{ so unbiased})$$

$$Var(\hat{y}^*) = Var(\hat{y}|X = x^*) = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum(x_i - \overline{x})^2} \right] = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}} \right]$$
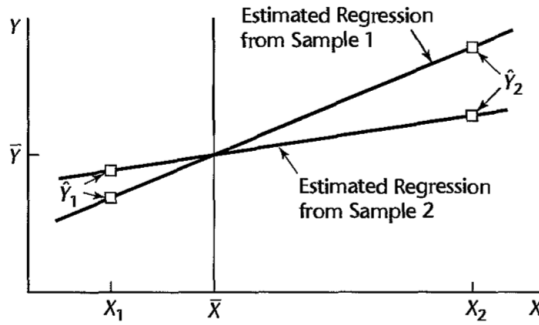
$$\hat{y}^* = (\hat{y}|X = x^*) \sim \mathcal{N}(\beta_0 + \beta_1 x^*, \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}} \right])$$

Note when $x^* = 0$, $Var(\hat{y}^*)$ reduces to variance of $\hat{\beta}_0$.

Proof. 3 parts

1. **Normality** of $\hat{y}^*$ follows from the fact that it is composed of $\hat{\beta}_0$ and $\hat{\beta}_1$, both of which are linear estimators of y_i

2. **Mean**

$$\mathbb{E}(\hat{y}^*) = \mathbb{E}\left(\hat{\beta}_0 + \hat{\beta}_1 x^* \right) = \mathbb{E}(\hat{\beta}_0) + x^* \mathbb{E}(\hat{\beta}_1) = \beta_0 + \beta_1 x^* = \mathbb{E}(y^*)$$

3. **Variance**

$$Var(\hat{y}|X = x^*) = Var(\hat{\beta}_0 + \hat{\beta}_1 x|X = x^*)$$
$$= Var(\hat{\beta}_0|X = x^*) + (x^*)^2 Var(\hat{\beta}_1|X = x^*) + 2x^* Cov(\hat{\beta}_0, \hat{\beta}_1|X = x^*)$$

with

$$Cov(\hat{\beta}_0, \hat{\beta}_1 | X = x^*) = Cov(\overline{y} - \hat{\beta}_1 \overline{x} | X = x^*)$$
$$= Cov(\overline{y}, \hat{\beta}_1 | X = x^*) - \overline{x} Cov(\hat{\beta}_1, \hat{\beta}_1)$$
$$= 0 - \overline{x} Var(\hat{\beta}_1)$$
$$= \frac{-\overline{x}\sigma^2}{S_{XX}}$$

So

$$Var(\hat{y} | X = x^*) = \sigma^2 \left(\frac{1}{2} + \frac{\overline{x}^2}{S_{XX}} \right) + (x^*)^2 \frac{\sigma^2}{S_{XX}} - \frac{2x^* \overline{x} \sigma^2}{S_{XX}} = \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}} \right)$$

Idea is variability of $\hat{y}^*$ is affected by how far x^* is from $\overline{x}$, via

$$(x^* - \overline{x})^2 = S_{XX}$$

The further x^* is from $\overline{x}$, the greater the variability. Note in plot, x_1 near $\overline{x}$, the fitted value $\hat{y}_1$ for two sample regression line (from 2 experiments) are close to each other; the fitted values $\hat{y}_2$ differ substantially due to the fact that x_2 is far from $\overline{x}$. In summary,

variation in $\hat{y}^*$ value from sample to sample will be greater when x^* is far from mean than when x^* is near mean

We can substitute MSE for σ^2 to obtain $s^2(\hat{y}^*)$. The estimated variance is given by

$$s^2(\hat{y}^*) = MSE \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}} \right]$$

$\square$

Definition. *Standardization of mean response estimator $\hat{y}^*$*

$$Z = \frac{\hat{y}^* - (\beta_0 + \beta_1 x^*)}{\sigma(\hat{y}^*)} \sim \mathcal{N}(0, 1)$$

note $\mathbb{E}(y^*) = \beta_0 + \beta_1 x^*$

$$T = \frac{\hat{y}^* - (\beta_0 + \beta_1 x^*)}{s(\hat{y}^*)} \sim t_{n-2}$$

Definition. *Confidence Interval for $\hat{y}^*$*
The $100(1 - \alpha)\%$ confidence interval for $\mathbb{E}(Y | X = x^*) = \beta_0 + \beta_1 x^*$ is given by

$$\hat{y}^* \pm s(\hat{y}^*) t_{1-\alpha/2, n-2} = (\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{1-\alpha/2, n-2} \sqrt{\frac{\sum e_i^2}{n-2}} \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}}}$$

Definition. *Obervations*

1. *variance of $\hat{y}^*$ is smallest when $x^* = \overline{x}$. So in an experiment to estimate mean response at a particular level x^* of predictor variable, the **precision of the estimate is greatest** if (everything else remain equal) the observations on X are spaced so that $\overline{x} = x^*$*

2. *confidence interval for $\hat{y}^*$ not sensitive to moderate departures from assumption of error being normally distributed. The robustness in estimating mean response is related to robustness of Confidence interval for β_0 and β_1*

2.5 Prediction of New Observations

Definition. *Prediction of New Observations*

1. ***Motivation*** *Idea is that we have a model set up given a set of data, and we would want to extrapolate to new data points. The new observation Y is viewed as the result of a new trial, independent of trials on which the regression analysis is based. Let level of X be x^* and the new observation y^*_{new} (which is unknown, and which we want to characterize), assuming that the underlying regression model is still appliacble for basic sample data*

2. ***Estimate of mean response*** $\mathbb{E}(y^*)$ ***vs. Prediction of new response*** y^*_{new}
*In the former case, we estimate **mean of distribution of** Y. In latter case, we predict an **individual outcome** drown from the distribution of Y. Idea is we have to take into account of the fact that the majority of individual outcomes deviate from the mean response*

Definition. *Prediction Interval for y^*_{new} when parameter is known*
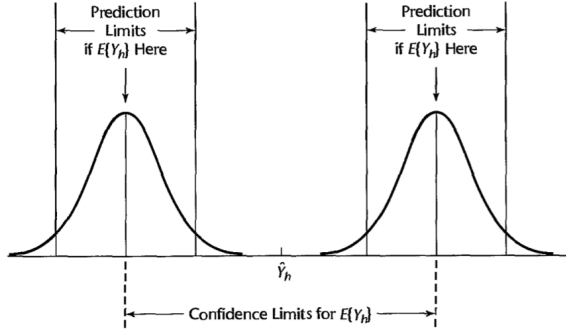*Assume all parameters are known, we have y^* follow a normal distribution*

$$y^*_{new} \sim \mathcal{N}(\beta_0 + \beta_1 x^*, \sigma^2)$$

So we have confidence interval,

$$(\beta_0 + \beta_1 x^*) \pm \sigma z_{1-\alpha/2}$$

Definition. *Prediction Interval for y^*_{new} when parameter is unknown*
When parameter is unknown, we must estimate regression parameters. We might want to estimate mean distribution of Y with $\hat{y}^*$ and variance of distribution of Y with MSE.

However, we cannot substitute these estimate into the previous distribution because $\mathbb{E}(y^*)$ is a random variable. Since we do not know the mean $\mathbb{E}(y^*)$, and only estiamte it by a confidence interval (shown previously), we cannot be certain of the distribution of Y. It could be anywhere along within its confidence intervals (for $\mathbb{E}\{Y_h\}$). Hence **prediction limit** for y^*_{new} must take into account two elements

1. variation in possible location of distribution of Y (i.e. sampling distributino of $\hat{y}^*$)

2. variation within the probability distribution of Y (namely σ^2, same as that of error terms')

We can prove that

$$\mathbb{E}(y^*_{new} - \hat{y}^*) = \mathbb{E}(y_{new} - \hat{y}|X = x^*) = \mathbb{E}(\hat{\beta}_0 + \hat{\beta}_1 x^*) - \mathbb{E}(\hat{y}^*) = 0$$

$$Var(y^*_{new} - \hat{y}^*) = Var(\hat{\beta}_0 + \hat{\beta}_1 x^*) + Var(\hat{y}^*) = \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}} \right] = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}} \right]$$

Note $Cov(y^*_{new}, \hat{y}^*) = 0$ by independence

$$y^*_{new} - \hat{y}^* \sim \mathcal{N}\left(0, \sigma^2 \left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}} \right]\right)$$

An unbiased estimtor of the variance is given by

$$s^2(y^*_{new} - \hat{y}^*) = MSE \left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}} \right]$$

Standardization of y^*_{new} gives

$$T = \frac{y^*_{new} - \hat{y}^*}{s^2(y^*_{new} - \hat{y}^*)} \sim t_{n-2}$$

The $(100 - \alpha)\%$ **Prediction limit** for y^*_{new} at $X = x^*$ is thus given by,

$$(\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{1-\alpha/2, n-2} MSE \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}}}$$

8

Definition. *Comments*

1. *prediction limit is subject to departure from normality of error term distributions*

2. *A confidence interval represents an inference on a paramter and is an interval that is intended to cover the value of **parameter**. A prediction interval, is a statement about the value to be taken by a **random variable**, the new observation y^*_{new}*

Confidence-band for Regression line

Definition. *Confidence-band* *represents uncertainty in the estimate of regression line (i.e. $\mathbb{E}(Y) = \beta_0 + \beta_1 X$). Used to determine appropriateness of a fitted regression function. **The Working-Hotelling** $1 - \alpha$ confidence band for regression line has boundary values at any level x^**

$$\hat{y}^* \pm W s\{\hat{y}^*\} \quad where \quad W^2 = 2F_{1-\alpha, n-2}$$

Analysis of Variance Approach to Regression Analysis

Definition. *Partitioning Total sum of Squares* *Idea is to partititon sums of squares and degree of freedom associated with response variable Y*
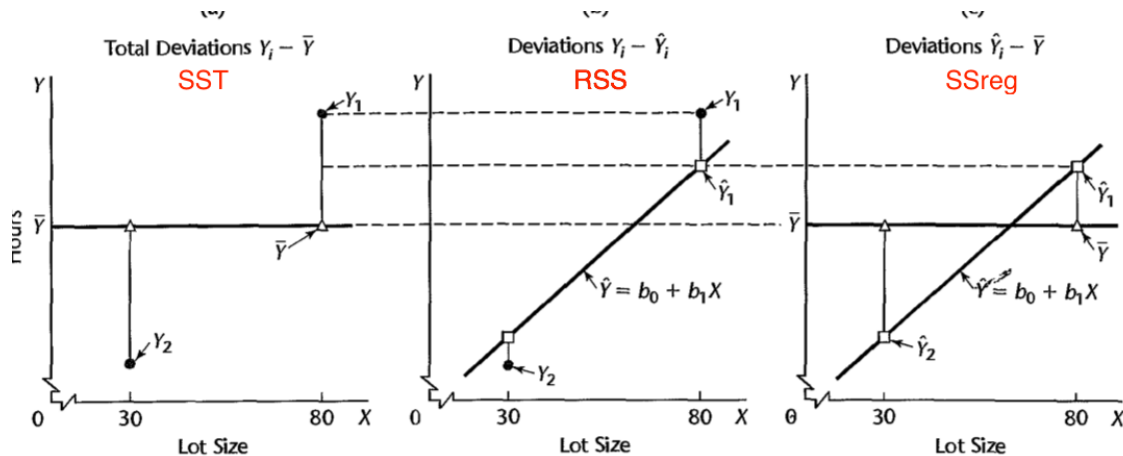
1. ***Total sum of squares** (SST)*

$$SST = \sum (y_i - \overline{y})^2$$

   is a measure of uncertainty of Y, without taking X into account

2. ***Residual sum of squares** (RSS, SSE)*

$$RSS = \sum (y_i - \hat{y}_i)^2$$

9

is a measure of variation of Y, with X taken into account. The greater variation of Y around fitted regression line, the larger the RSS.

3. **Regression sum of squares** *(SSR, SSreg)*

$$SSreg = \sum(\hat{y}_i - \overline{y})^2$$

Alternatively we can expand and get to an equivalent expression

$$SSreg = \sum((\hat{\beta}_0 - \hat{\beta}_1 x_i) - (\hat{\beta}_0 + \hat{\beta}_1 \overline{x}))^2 = \hat{\beta}_1^2 \sum(x_i - \overline{x})^2$$

SST can be broken down into 2 components, i.e. deviation of fitted value $\hat{Y}_i$ around mean $\overline{Y}$ and deviation of observed Y_i around fitted regression line

$$\underbrace{Y_i - \overline{Y}}_{\substack{\text{Total} \\ \text{deviation}}} = \underbrace{\hat{Y}_i - \overline{Y}}_{\substack{\text{Deviation} \\ \text{of fitted} \\ \text{regression} \\ \text{value} \\ \text{around mean}}} + \underbrace{Y_i - \hat{Y}_i}_{\substack{\text{Deviation} \\ \text{around} \\ \text{fitted} \\ \text{regression} \\ \text{line}}}$$

The sum of these squared deviations have the same relationship!

$$\sum(Y_i - \hat{Y})^2 = \sum(\hat{Y}_i - \overline{Y})^2 + \sum(Y_i - \hat{Y}_i)^2$$

$$SST = RSS + SSreg$$

total variability = unexplained variability + variability explained by model

Proof.

$$
\begin{aligned}
\sum(y_i - \overline{y})^2 &= \sum[(\hat{y}_i - \overline{y}) + (y_i - \hat{y}_i)]^2 \\
&= \sum(\hat{y}_i - \overline{y})^2 + \sum(y_i - \hat{y}_i)^2 + 2\sum(\hat{y}_i - \overline{y})(y_i - \hat{y}_i) \\
&= SSreg + RSS
\end{aligned}
$$

where

$$
\begin{aligned}
2\sum(\hat{y}_i - \overline{y})(y_i - \hat{y}_i) &= 2\sum \hat{y}_i(y_i - \hat{y}_i) - 2\overline{y}\sum(y_i - \hat{y}_i) \\
&= 2\sum \hat{y}_i e_i - 2\overline{y}\sum e_i \\
&= 0
\end{aligned}
$$

$\square$

Definition. *Break down of degress of freedom*

1. *SST has $n-1$ degrees of freedom (1 df is lost because sample mean $\overline{y}$ is used to estimate population mean)*

2. *RSS has $n-2$ degrees of freedom (2 df lost because β_0 and β_1 are estimated)*

3. *SSreg has 1 degree of freedom.*

$$n - 1 = 1 + (n - 2)$$

Definition. *Mean Squares* *A sum of squares divided by its associated degree of freedom is mean square*

$$MSR = \frac{SSreg}{1} = SSreg \qquad MSE = \frac{RSS}{n - 2}$$

MSE of an estimator measures the average of the squares of the errors or deviations, i.e. the difference between the estimator and what is estimated. MSE is a measure of quality of estimator

Definition. *Analysis of Variance Table* *displays breakdowns of total sum of squares and associated degrees of freedom*

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

Definition. *Expected Mean square* *Expected value of a mean square is the mean of its sampling distribution and tells us what is being estimated by the mean square.*

$$\mathbb{E}(MSE) = \sigma^2$$
$$\mathbb{E}(MSR) = \sigma^2 + \beta_1^2 \sum (x_i - \overline{x})^2$$

Proof. Given

$$\frac{RSS}{\sigma^2} \sim \chi_{n-2}^2$$

We have

$$\mathbb{E}(\frac{RSS}{\sigma^2}) = n - 2 \quad \rightarrow \quad \mathbb{E}(\frac{RSS}{n - 2}) = \mathbb{E}(MSE) = \sigma^2$$

11

To prove $\mathbb{E}(MSR)$, we first note form of $SSreg$,

$$SSreg = \hat{\beta}_1^2 \sum (x_i - \overline{x})^2$$

to find its expected value, we first find

$$\mathbb{E}(\hat{\beta}_1^2) = Var\hat{\beta}_1 + (\mathbb{E}(\hat{\beta}_1))^2 = \frac{\sigma^2}{S_{XX}} + \beta_1^2$$

Then we have

$$\mathbb{E}(SSreg) = \mathbb{E}(\hat{\beta}_1^2) \sum (x_i - \overline{x})^2 = (\frac{\sigma^2}{S_{XX}} + \beta_1^2)S_{XX} = \sigma^2 + \beta_1^2 S_{XX}$$

Then we find mean squared regression

$$\mathbb{E}(MSR) = \mathbb{E}(\frac{SSreg}{1}) = \sigma^2 + \beta_1^2 S_{XX}$$

$\square$

Definition. *F **test of** $\beta_1 = 0$ **versus** $\beta_1 \neq 0$*
To test for if there is a linear relationship between X and Y, We can test for this with t test

$$T = \frac{\hat{\beta}_1 - 0}{\sigma/\sqrt{S_{XX}}} \sim t_{n-2}$$

Alternatively, to generalize the test to multiple regression, we can use F test where test statistics is given by
$$F^* = \frac{MSR}{MSE} = \frac{SSreg}{RSS/(n-2)}$$

Intuitively, larger values of F^* (larger MSR) supports H_α and values of F^* near 1 supports H_0. It can be shown that sampling distribution for F^* is given by

$$F^* \sim F_{1,n-2}$$

The test is then given by
$$\textbf{Reject if } F^* > F_{1-\alpha;1,n-2}$$

Definition. *Equivalence of F test and t test*
Given α confidence level, F test of $\beta_1 = 0$ versus $\beta_1 \neq 0$ is equivalent to two-sided t test

Proof.
$$F^* = \frac{SSreg/1}{RSS/(n-2)} = \frac{\hat{\beta}_1 \sum (x_i - \overline{x})^2}{MSE} = \frac{\hat{\beta}_1 S_{XX}}{MSE}$$

Note
$$s^2(\hat{\beta}_1) = \frac{MSE}{S_{XX}}$$

we have
$$F^* = \left(\frac{\hat{\beta}_1^2 - 0}{s(\hat{\beta}_1)}\right)^2 = (t^*)^2$$

where t is the t statistics $\qquad\square$

2.9 Descriptive Measures of Linear Association between X and Y

Definition. *Coefficient of Determination*

1. *SST measures variation in Y without taken into account of X*

2. *RSS is a measure of variation in Y when taken into account of X*

3. *A natural measure of effect of X in reducing variation in Y is in the difference between SST and RSS as a proportion of total variation*

*The **coefficient of determination** R^2 is given by*
$$R^2 = \frac{SST - RSS}{SST} = \frac{SSreg}{SST}$$

with
$$0 \leq R^2 \leq 1$$

R^2 can be interpreted as the proportionate reduction of total variation associated with use of X predictor variable

Definition. *Coefficient of Correlation*
A measure of linear association between Y and X when both Y and X are random is coefficient of correlation.
$$r = \pm\sqrt{R^2}$$

Dummy variable regression