Simple Linear Regression.

A. Skripnikov[1]

[1]New College of Florida

IDC 5205

Linear Regression.

Linear regression is a useful tool for predicting/explaining a **quantitative** response based on one or more **predictors**:

- Estimate employee's salary based on experience and education.
- Using house characteristics (size, age, location), evaluate its worth.
- Predict the score differential for a football game based on comparative team statistics and home-field advantage.

Main reference example (*Advertisement.csv***)**:

We're asked to infer the effects that various types of advertisement (TV, radio, newspaper) may have on product sales (**See** *R* **code**).

Simple Linear Regression.

Presume one has

- **Quantitative** response Y
- single predictor variable X

Simple Linear Regression equation:

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

or, in full-blown notation,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \;\; i = 1, \ldots, n$$

Example. Let's just focus on how TV ads affect the sales. Then

- **Quantitative** response is *Sales*,
- predictor variable is *TV*

Simple Linear Regression equation:

$$Sales = \beta_0 + \beta_1 TV + \epsilon,$$

or, in full-blown notation,

Here,

- β_0, β_1 are model **parameters** (their values **unknown**), and
- need to be **estimated** via some values $\hat{\beta}_0, \hat{\beta}_1$.

Estimating β_0 and β_1.

Q: How do we find estimates $\hat{\beta}_0, \hat{\beta}_1$ for **parameters** β_0, β_1?

Task: Find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that

the **estimated** (**"fitted"**) value, $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

is as close as possible to

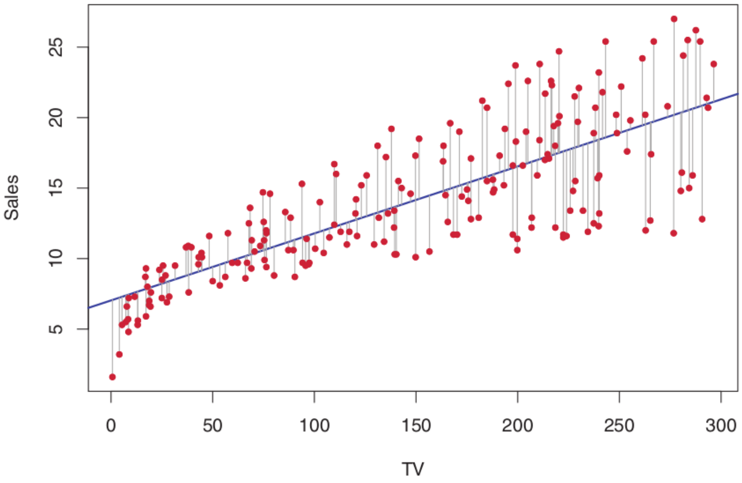
the **observed** (**"true"**) value, Y_i

$$\Downarrow$$

We need to **minimize** magnitude of **residuals**, $e_i = Y_i - \hat{Y}_i, \; i = 1, \ldots, n$

Geometry of (Simple) Linear Regression: Straight Line.

Geometrically, it amounts to finding a **fitted line** $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ that's **closest to the data points** (**vertical lines** are **residuals** $e = \hat{Y} - Y$):



Estimating β_0 and β_1: **Least Squares**.

Note: Can't minimize every single residual e_i individually, but instead use **Method of Least Squares**. Define *Residual Sum of Squares (RSS)*:

$$RSS = \sum_i e_i^2 = \sum_i (Y_i - \hat{Y}_i)^2$$

and formulate the following **optimization task**:

$$\underset{\beta_0, \beta_1}{min} \; RSS =$$

Values $\hat{\beta}_0, \hat{\beta}_1$ solving this criteria are called **least squares** estimates.

Advertising Example: Interpretation of **slope**.

Example. For *Advertising* data (**see** *R* **code**), we got

- $\hat{\beta}_0 = 7.03, \ \hat{\beta}_1 = 0.04754,$
- hence, the **fitted regression equation** is

Task. Noting that the units are $1,000$\$'s for TV & $1,000$ items for *Sales*,

- Interpret the **slope** estimate, $\hat{\beta}_1 = 0.0475$.

Advertising Example: Interpretation of **intercept**.

Example. For *Advertising* data (**see** *R* **code**), we got

- $\hat{\beta}_0 = 7.03, \ \hat{\beta}_1 = 0.04754,$
- hence, the **fitted regression equation** is

$$\widehat{Sales} = 7.03 + 0.04754 \times TV$$

Task. Noting that the units are: $1,000$\$'s for TV & $1,000$ items for *Sales*,

- Interpret the **intercept** estimate, $\hat{\beta}_0 = 7.03$.

Slope and **intercept**: Generic Interpretations.

Say, you have the following general **fitted regression equation**

$$\widehat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \times X$$

Task. Write down a **generic** template for interpretation of

- **Slope** estimate $\hat{\beta}_1$:

- **Intercept** estimate $\hat{\beta}_0$:

Advertising Example: Prediction.

Having fitted the model, one can proceed to **make predictions**:

- What sales are expected for markets that invest 20,000$ in TV ads?

$$\widehat{Sales} = 7.03 + 0.0475 \times 20 = 7.98$$

Interpretation: *On average*, ...

- What sales are expected for markets that invest 100,000$ in TV ads?

See *R* **code**.

Assessing Quality of Fit.

Two main measures to evaluate **the quality of fit** for your linear regression model:

1. Residual Standard Error (RSE):

$$RSE = \sqrt{\frac{1}{n-2} RSS} = \sqrt{\frac{1}{n-2} \sum_i (Y_i - \hat{Y}_i)^2}$$

Interpretation:

Q: Why $n-2$?

A: Explained later, but it does have to do with "degrees of freedom".

R^2 statistic.

Issue with **RSE**? It is measured in units of Y, not standardized.

Alternative:

2. R^2-statistic:

$$R^2 = \frac{TSS - RSS}{TSS} = \frac{\sum_i (Y_i - \bar{\mathbf{Y}})^2 - \sum_i (Y_i - \hat{Y}_i)^2}{\sum_i (Y_i - \bar{\mathbf{Y}})^2}$$

- measures the **proportion of variability in** *response* Y that's explained by the regression model, specifically

- TSS (Total Sum of Squares) $= \sum_i (Y_i - \bar{\mathbf{Y}})^2$ - measures the **initial variability** in *response*
- RSS (Residual Sum of Squares) $= \sum_i (Y_i - \hat{Y}_i)^2$ - measures the amount of **variability** in *response* that is **left unexplained** after performing the **regression**.
- Hence, $TSS - RSS$ measures the amount of **variability** in the *response* that **is explained** (or removed) by performing the **regression**.

R^2 statistic.

Q: What values of R^2 are indicative of a good model? Bad model? Why?

Example. Calculate RSE and R^2 (**see** R **code**) for the $Sales \sim TV$ regression, **interpret**.

R^2 statistic.

While R^2 statistic is more interpretable than RSE, it is still unclear what's a good R^2 value depending on application:

- In certain physics problems, we may know that the data truly comes from a linear model with a small residual error. Then, R^2 is expected ≈ 1, otherwise there's a problem with data generation in the experiment.
- In biology, psychology, marketing & other domains, linear model is at best an extremely rough approximation to the data, and sometimes even an R^2 **value below** 0.1 may be indicative of **an acceptable fit**.

Statistical Inference: Sample ($\hat{\beta}$) to Population (β).

Q: How can **sample estimates** $\hat{\beta}_0, \hat{\beta}_1$ be used to **infer** the **unknown** true parameter values $\hat{\beta}_0, \hat{\beta}_1$?

A: **Statistical inference** techniques, such as

- hypothesis testing,
- confidence intervals.

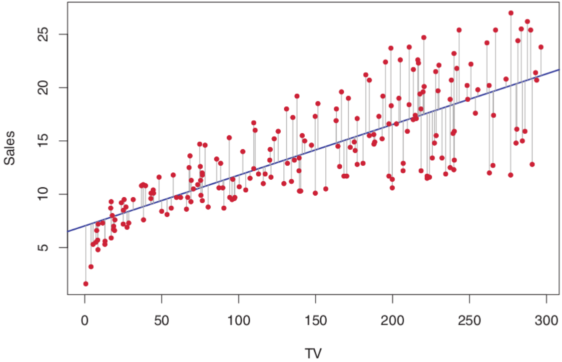
Q: So, what is the difference between β_1 and $\hat{\beta}_1$?

-
-

Statistical Inference: Sample ($\hat{\beta}$) to Population (β).

Our **fitted line**

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \times X$$



is nothing more but a **sample estimate** to the **population line**

$$Y \approx \beta_0 + \beta_1 \times X$$

which we try to **infer about**.

Full Model Equation for Simple Linear Regression is:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \sim_{ind} N(0, \sigma^2), \quad i = 1, \dots, n,$$

Example (cont'd). In our example, we'd have $Y = Sales,\ X = TV$:

Qs:

- (**Once again**) What are β_0, β_1 as opposed to $\hat{\beta}_0, \hat{\beta}_1$? Are the parameters β_0, β_1 **constant** or **random**? Why?

Simple Linear Regression: Full Model Equation.

Qs (cont'd):

- What is the ϵ_i term for? Is it **constant** or **random**? Why?

Simple Linear Regression: Model Assumptions.

Task. Let $Y = Sales, X = TV$. Show that

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

leads to

$$Y_i \sim N(\beta_0 + \beta_1 X_i,\ \sigma^2),$$

hence Y_i is a **random draw** from a **population of response values for <u>all</u> observations with** $X = X_i$, which has **distribution** $N(\beta_0 + \beta_1 X_i, \sigma^2)$.

Note. Y_i's, ϵ_i's are considered **random**. X_i's - **fixed**. **See** R **code**.

Simple Linear Regression: Model Assumptions.

Task (cont'd).

Simple Linear Regression: Model Assumptions.

The fact of

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2), \;\; i = 1, \ldots, n$$

points to **three critical assumptions** of **linear regression**:

1. **Linearity**:

   $$E[Y_i] \; = \; \beta_0 + \beta_1 X_i, \quad AKA \quad E[Y \mid (X = X_i)] \; = \; \beta_0 + \beta_1 X_i$$

   Y, **on average**, represents a linear function of X.

2. **Constant variance**:

   $$V[Y_i] \; = \; \sigma^2, \quad AKA \quad V[Y \mid (X = X_i)] \; = \; \sigma^2$$

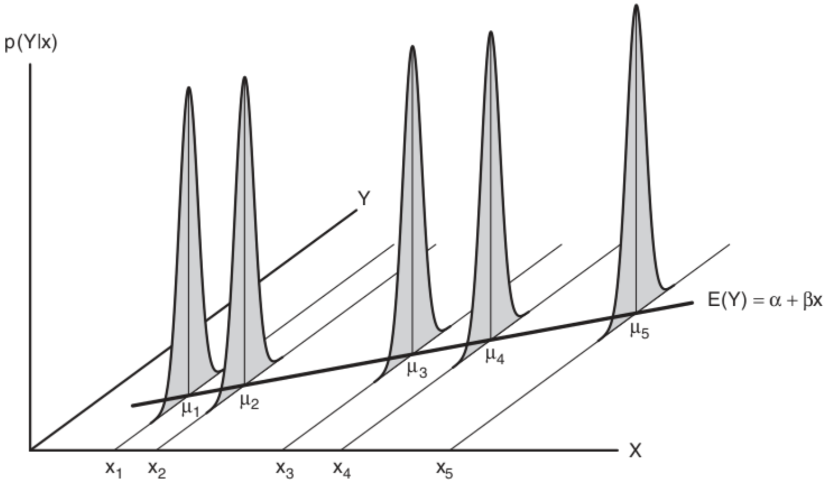
   Y has the same variance across all values of X.

3. **Normality**:

   $$Y_i \; \sim \; N(\beta_0 + \beta_1 X_i, \sigma^2), \quad AKA \quad [Y \mid (X = X_i)] \; \sim \; N(\beta_0 + \beta_1 X_i, \sigma^2)$$

   Y is normally distributed for a fixed value of X (e.g. $X = X_i$).

Simple Linear Regression: Model Assumptions.



Simple Linear Regression: Assumption of Independence.

Another critical assumption is that of:

4. **Independence**:

   $\epsilon_i \sim_{\textbf{ind}} \ldots \quad \Leftrightarrow \quad \epsilon_i$ and ϵ_j are independent for $i \neq j, \;\; i,j = 1, \ldots, n$.

   It also implies that (details left out):

   $$Y_i \text{ and } Y_j \text{ are independent for } i \neq j, \;\; i,j = 1, \ldots, n.$$

This assumption is determined by whether the observations are sampled **independently**, and needs to be justified by procedures of data collection:

- if it's random sample from a large population, then independence is roughly satisfied;
- if it's a time series, or spatial data, then the assumption of independence may be **very wrong**, subsequently affecting legitimacy of your statistical inference (p-values, confidence intervals, etc)

Simple Linear Regression: Model Assumptions.

To recollect all the **model assumptions** of **simple linear regression**.

1. **Linearity**: $E[Y_i] = \beta_0 + \beta_1 X_i$
2. **Constant variance**: $V[Y_i] = \sigma^2$.
3. **Normality**: $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$
4. **Independence**: ϵ_i and ϵ_j ($\Leftrightarrow Y_i$ and Y_j) are independent for $i \neq j$.

NOTE: $Y_i \equiv [Y_i \mid (X = X_i)]$.

The classic model formulation capturing **all** these assumptions is

Simple Linear Regression: Full Modeling Equation

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \sim_{ind} N(0, \sigma^2), \quad i = 1, 2, \ldots, n \tag{1}$$

Why Least Squares? Nice Theoretical Properties.

Q: To estimate regression parameters, why use least squares in particular?

Reason #1: Least squares approach leads to a well-defined, **closed-form**, analytical solution (**see one of HW problems for solution formulas**).

Reason #2: Under the linear regression **model assumptions**, **least squares (LS) estimators $\hat{\beta}_0, \hat{\beta}_1$** have desirable **statistical** properties:

1. Unbiasedness ($E[\hat{\beta}_j] = \beta_j,\ j = 0, 1$).
2. Analytical formulas for sampling variances ($V[\hat{\beta}_j],\ j = 0, 1$)
3. Normality of sampling distribution ($\hat{\beta}_j \sim N,\ j = 0, 1$).

making them **great** for **conducting inference** about population parameters β_0, β_1.

Why Least Squares? Reason #2: Nice Properties.

Example (will be done as a Lab). Presume we know that the true relationship is

$$Y = 2 + 3X + \epsilon, \epsilon \sim N(0, 40^2) \tag{2}$$

with $\beta_0 = 2$, $\beta_1 = 3$.

We proceed to:

1. Generate 200 values of $X = (X_1, X_2, \ldots, X_{200})$. Keep them fixed.
2. Repeat the following process a 1000 times, $j = 1, \ldots, 1000$:
    - **generate a sample** of response values $Y^{(i)} = (Y_1^{(i)}, Y_2^{(i)}, \ldots, Y_{200}^{(i)})$ for those 200 values of X according to equation (2):

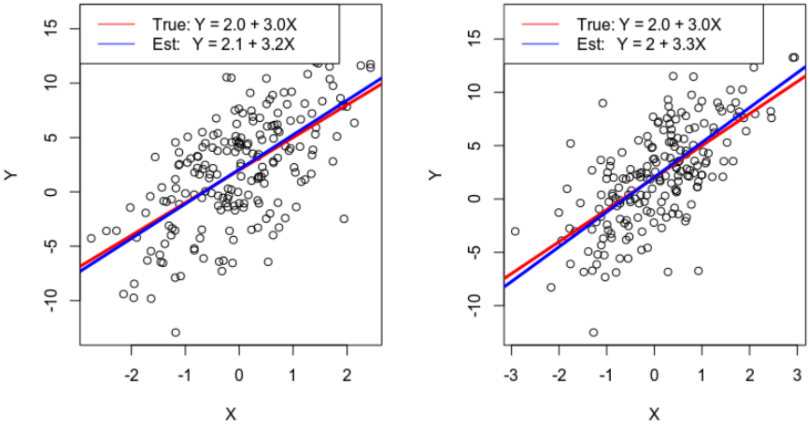
    $$Y^{(i)} = 2 + 3X + \epsilon, \;\epsilon \sim N(0, 40^2)$$

    - calculate the **least squares estimate line** for that j^{th} sample:

    $$\hat{Y}^{(i)} = \hat{\beta}_0^{(i)} + \hat{\beta}_1^{(i)} X$$

Statistical Inference: Population to Sample.

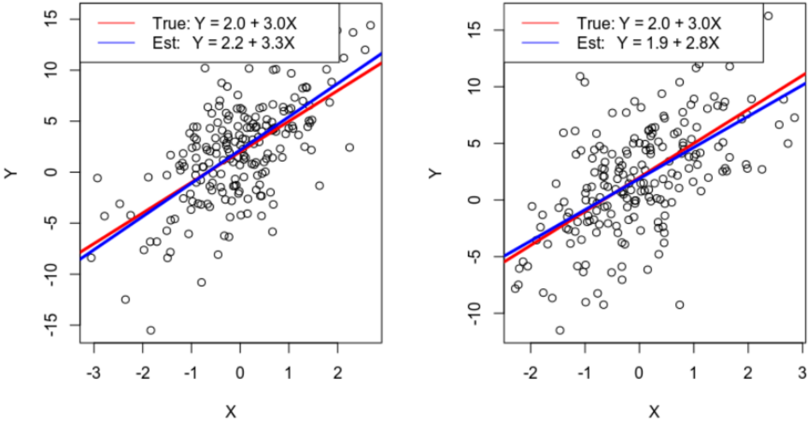
Least squares estimate lines $Y^{(i)} = \hat{\beta}_0^{(i)} + \hat{\beta}_1^{(i)} X$ for each sample won't be exactly the same as the **true population line** $Y = \beta_0 + \beta_1 X$:



but they will be relatively close.

Statistical Inference: Population to Sample.

Least squares estimate lines $Y^{(i)} = \hat{\beta}_0^{(i)} + \hat{\beta}_1^{(i)} X$ for each sample won't be exactly the same as the **true population line** $Y = \beta_0 + \beta_1 X$:



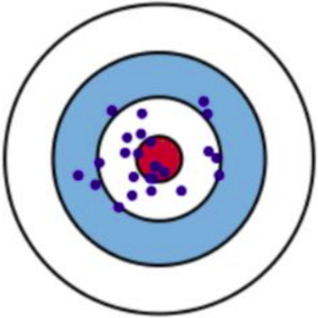
but they will be relatively close.

Unbiasedness of $\hat{\beta}$'s.

After $m = 1000$ **simulations**, we get:

$$\frac{1}{m}\sum_i \hat{\beta}_0^{(i)} \approx \beta_0, \qquad \frac{1}{m}\sum_i \hat{\beta}_1^{(i)} \approx \beta_1$$

which means that $\hat{\beta}_j$ is an **unbiased estimate** of $\beta_j,\ j = 0, 1$.



Practical definition of "**Unbiasedness**"
(for Least Squares Estimates $\hat{\beta}$)

Over many random samples taken from the population, the least squares estimate $\hat{\beta}_j$ will be equal to the population value β_j, **on average**

Theoretical notation: $E[\hat{\beta}_j] = \beta_j,\ j = 0, 1.$

Standard error.

Unbiasedness across many hypothetical samples is great and all, but..

with **real data** we only get to see **one sample**

$\Downarrow$

just **one sample estimate** for each parameter.

Q: How to use that **one sample estimate** (e.g. $\hat{\beta}_1$) in order to infer about the true parameter value (β_1)?

A: We need the **standard error** $SE[\hat{\beta}_1]$ of the estimate, where

$$SE[\hat{\beta}_1] = \{\text{by how much, } \textbf{on average}, \ \hat{\beta}_1 \text{ deviates from } \beta_1\}$$

Task. Check the *summary*() of fitted *sales* $\sim$ *TV* regression in *R*, find and interpret std. errors there.

Origins of $SE(\hat{\beta})$ - FOR CURIOUS.

Q: Where do the $SE(\hat{\beta})$ values come from?

A: They come from **taking a square root** of ($SE(\hat{\beta}) = \sqrt{V[\hat{\beta}]}$)

1. Theoretical formulas for **sampling variance** of **least squares est-s**:

$$V[\hat{\beta}_0] = \sigma^2\left(\frac{1}{n} + \frac{\bar{\mathbf{X}}^2}{\sum_{i=1}^n (X_i - \bar{\mathbf{X}})^2}\right), \quad V[\hat{\beta}_1] = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{\mathbf{X}})^2},$$

Practical definition of **"Sampling Variance"**
(for Least Squares Estimates $\hat{\beta}$)

Variance of $\hat{\beta}$ **estimates** over **many samples** taken from population.

2. Where we substitute unknown population standard deviation σ for

$$\hat{\sigma} = RSE = \sqrt{\frac{1}{n-2}\sum_i (Y_i - \hat{Y}_i)^2}$$

Sampling Distribution of $\hat{\beta}_0, \hat{\beta}_1$.

For **sampling distribution** of **least squares estimates** $\hat{\beta}_j$, we've discussed the

- sampling mean ("unbiasedness"):

$$E[\hat{\beta}_j] = \beta_j, \ j = 0, 1$$

- sampling variance:

$$V[\hat{\beta}_j], \ j = 0, 1$$

Qs:

- What's meant by **sampling distribution** of a *statistic* (e.g. $\bar{\mathbf{x}}$, $\hat{p}$, $\hat{\beta}_1$)?

- To conduct inference on population parameters β_j, what else do we need to know about sampling distributions of $\hat{\beta}_j$?

  A: **Shape**.

Sampling Distribution of $\hat{\beta}_0, \hat{\beta}_1$.

Theorem. For simple linear regression

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \sim_{ind} N(0, \sigma^2), \quad i = 1, 2, \ldots, n$$

the **sampling distributions** of $\hat{\beta}_0, \hat{\beta}_1$ are

$$\hat{\beta}_j \sim N(\beta_j, V[\hat{\beta}_j]),\ j = 0, 1$$

Statistical Inference for $\hat{\beta}_0, \hat{\beta}_1$: Unknown σ.

Sampling distributions of $\hat{\beta}_0, \hat{\beta}_1$ are

$$\hat{\beta}_j \sim N(\beta_j, V[\hat{\beta}_j]),\ j = 0, 1$$

Q: Formulas for $V[\hat{\beta}_j]$ contain σ^2. Why is that an issue when trying to do **inference**? What should we do to address this?

Statistical Inference for $\hat{\beta}_0, \hat{\beta}_1$: Unknown σ.

REMINDER:

- When using sample mean $\bar{\mathbf{X}}$ to infer about μ,

$$\frac{\bar{\mathbf{X}} - \mu}{\sigma/\sqrt{n}} \sim N(0,1),$$

Q: how did we deal with the unknown σ?

- Now, when using $\hat{\beta}_j$ to infer about β_j,

$$\hat{\beta}_j \sim N(\beta_j, V[\hat{\beta}_j]) \quad \Longrightarrow \qquad\qquad\qquad\qquad \sim N(0,1)$$

for the unknown σ we plug in

$$\hat{\sigma} = RSE = \sqrt{\frac{\sum_i (Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\frac{\sum_i (Y_i - [\hat{\beta}_0 + \hat{\beta}_1 X_i])^2}{n-2}}$$

Statistical Inference for $\hat{\beta}_0, \hat{\beta}_1$: Degrees of Freedom.

Q: Why do we use

1. $n-1$ in the denominator of $\hat{\sigma}$ for inference via sample mean $\bar{\mathbf{X}}$,
2. $n-2$ in the denominator of $\hat{\sigma}$ for inference via $\hat{\beta}_j, \ j = 0, 1$?

A: Those are **degrees of freedom**, and each estimated parameter (be it μ, or the β_j's) "takes up" a degree of freedom.

1. For sample mean $\bar{\mathbf{X}}$, when calculating

$$\hat{\sigma} = \sqrt{\frac{\sum_i (X_i - \bar{\mathbf{X}})^2}{n-1}} \equiv \sqrt{\frac{\sum_i (X_i - \hat{\mu})^2}{n-1}},$$

we use the estimate $\bar{\mathbf{X}}$ (or "$\hat{\mu}$") instead of true population mean μ $\Longrightarrow$ 1 degree of freedom lost.

Statistical Inference for $\hat{\beta}_0, \hat{\beta}_1$: Degrees of Freedom.

Q: Why do we use

1. $n - 1$ in the denominator of $\hat{\sigma}$ for inference via sample mean $\bar{\mathbf{X}}$,
2. $n - 2$ in the denominator of $\hat{\sigma}$ for inference via $\hat{\beta}_j, \ j = 0, 1$?

A: Those are **degrees of freedom**, and each estimated parameter (be it μ, or the β_j's) "takes up" a degree of freedom.

2. For least squares estimates $\hat{\beta}_0, \hat{\beta}_1$ in simple linear regression, when calculating

$$\hat{\sigma} = \sqrt{\frac{\sum_i (Y_i - [\hat{\beta}_0 + \hat{\beta}_1 X_i])^2}{n - 2}}$$

....

Sampling Distribution of $\hat{\beta}_0, \hat{\beta}_1$.

Q: In sample mean inference, when plugging in $\hat{\sigma} = s$ for σ, did we have

$$\frac{\bar{\mathbf{X}} - \mu}{s/\sqrt{n}} \sim N(0,1) \quad ???$$

If not, what did we have instead? Why?

Q: Denoting $SE(\hat{\beta}_j)$ as the standard error of $\hat{\beta}_j$, after plugging in $\hat{\sigma} = RSE$ for σ, what should be the distribution of

$$\frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \sim$$

Confidence Intervals.

Now that we've figured out

$$T = \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \sim t_{n-2}, \ j = 0, 1$$

we may proceed to derive the **confidence intervals** for β_j's.

REMINDER:

95% confidence interval for parameter β_j is such interval (c, d) that

$$P(\beta_j \in (c, d)) = 0.95$$

In particular, if we were to

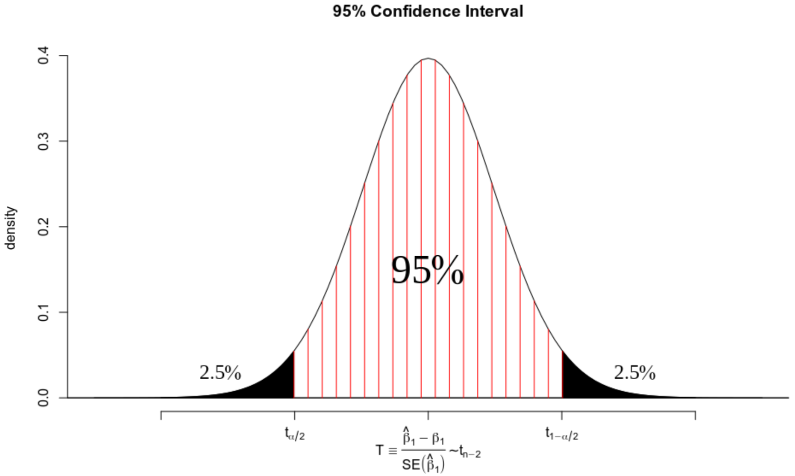
1. (hypothetically) Obtain many samples from the population, and
2. Calculate the confidence interval for each of those samples,

then **parameter** β_j should end up in 95% **of those confidence intervals**. It is also known as 95% **coverage**.

Confidence Intervals.

Formula for $(1-\alpha) \times 100\%$ confidence interval of β_j parameter is

$$(\hat{\beta}_j - t_{1-[\alpha/2]} SE(\hat{\beta}_j), \;\; \hat{\beta}_j + t_{1-[\alpha/2]} SE(\hat{\beta}_j))$$



Statistical Inference: Confidence Intervals.

Example (cont'd). For *Sales* $\sim$ *TV* linear regression (**see** *R* **code**),

- Obtain and **interpret** 95% confidence interval for β_1,

Interpretation: We are 95% confident that ...

Task. Explain what "we are 95% confident" means exactly.

Confidence Intervals: Interpretation.

Example (cont'd). For *Sales* $\sim$ *TV* linear regression (**see** *R* **code**),

- Obtain and **interpret** 90% CI for β_1,

Statistical Inference: Confidence Intervals.

Example (cont'd). For *Sales* $\sim$ *TV* linear regression (**see** *R* **code**),

- Obtain 90%, 95%, 99% (in that order) confidence intervals for β_1.

Q: What happens to confidence interval as confidence level increases? Why does it make sense?

Confidence Intervals: Interpretation.

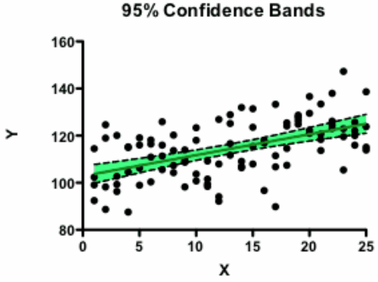
Example (cont'd). For *Sales* $\sim$ *TV* linear regression (**see** *R* **code**),

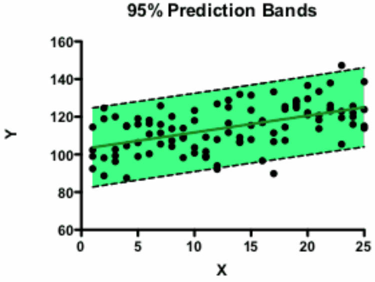
- Obtain and **interpret** 95% CI for β_0.

Prediction and confidence bands.

When providing model **predictions**, one often presents uncertainty bands around them, of which there are two kinds:

- Confidence (narrow) bands, and
- Prediction (wide) bands.





Prediction and confidence bands.

When predicting for an observation with predictor value $X = X_0$:

- **Confidence (narrow)** bands try to capture the **average response** $\bar{Y}$ for all observations with $X = X_0$. E.g., you are 95% sure that the **average** response $\bar{Y}$ for observations with $X = x_0$ would lie within the 95% confidence bands.

  Note: Larger sample size $n \implies$ more narrow the confidence bands.

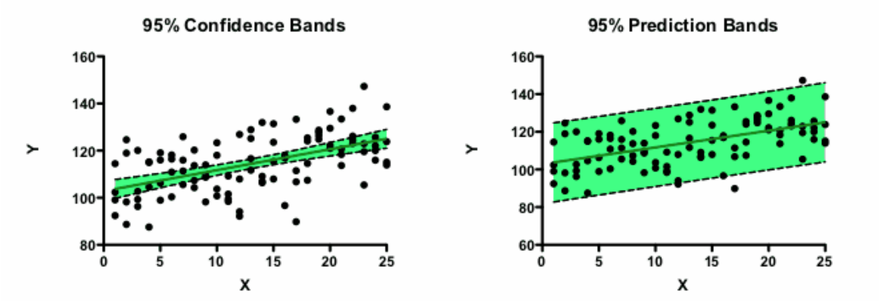
- **Prediction (wide)** bands try to capture **all response values** Y_i (not just their mean $\bar{Y}$) for all observations with $X = X_0$. Hence, 95% prediction bands contain 95% of **all** future responses Y with $X = x_0$.

  Note: Larger sample size $n \implies$ more narrow the prediction bands.

Prediction and confidence bands: illustration.

With many data points, you expect:

- a large fraction of data points to lie **outside** the **confidence** bands, but
- about 95% of the points to lie **within** the **prediction** bands.



See R **code for the** $Sales \sim TV$ **example.**

Prediction and confidence bands: Interpretation.

Task (See R code). In *Advertising* data example, for regression

$$Sales \sim TV$$

we make predictions of markets with $100k$ TV advertisement budget.

Proceed to **interpret** the following results:

- Single prediction of items sold: 11.78.

- 95% **confidence** bands for items sold: $(11.27, 12.30)$

Prediction and confidence bands: Interpretation.

Task (See *R* code). In *Advertising* data example, for regression

$$Sales \sim TV$$

we make predictions of markets with $100k$ TV advertisement budget.

- 95% **prediction bands** for items sold: $(5.34, 18.23)$

Comment on differences between confidence and prediction bands.

summary() Output Breakdown.

Example. *summary*() function output for our *Sales* $\sim$ *TV* linear regression model fitted for *Advertising* data set:

```
> summary(lm.obj)
...
Residuals:
    Min      1Q  Median      3Q     Max
-8.3860 -1.9545 -0.1913  2.0671  7.2124



Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,       Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

summary() Output: Coefficients table,.

Example (cont'd). Focusing on the *Coefficients* table:

```
> summary(lm.obj)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **Estimate**:

- **Std. Error**:

- **t value,** $Pr(>|t|)$: where do these come from? **See following slides**.

Statistical Inference: Hypothesis Testing.

The most common hypothesis test in simple linear regression involves:

H_0: {There is **no** **linear** relationship between X and Y}

vs

H_a: {There **is a** **linear** relationship between X and Y}

Keeping in mind the simple linear regression modeling equation:

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

these hypotheses *mathematically* correspond to

$$H_0\colon\ \beta_1 = 0 \qquad vs \qquad H_a\colon\ \beta_1 \neq 0$$

Why?

Statistical Inference: Hypothesis Testing.

Example (cont'd). For our *Sales* $\sim$ *TV* simple linear regression model:

Task: Formulate the H_0 and H_0 hypotheses

- In plain English.
- Mathematically

Statistical Inference: Hypothesis Testing.

Q: We estimated β_1 with $\hat{\beta}_1 = 0.0475 \neq 0$. Therefore, H_0 should be false and H_a is true, right? Or no? Why?

Hypothesis Testing: Main Steps.

1. State the hypotheses about parameter of interest β_1:

$$H_0\colon \beta_1 = 0, \quad \text{vs} \quad H_a\colon \beta_1 \neq 0,$$

2. Calculate the **observed** test statistic value: $TS = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$
3. If $H_0\colon \beta_1 = 0$ **were** true, test statistic T - depending on a **random sample** drawn from the population - is **expected** to take on values according to t_{n-2} distribution:

$$T = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{n-2}, \quad \textbf{given that } H_0\colon \beta_1 = 0 \textbf{ were true}$$

4. Use t_{n-2} distribution to calculate p-value, which quantifies "how likely it was to witness $T = TS$ (or more extreme) if H_0 were actually true" (**see next slide for illustrations**).
5. For a pre-determined significance level α (usually 0.05, 0.01, or 0.1),

$$\text{if} \begin{cases} p\text{-value} \leq \alpha, & - \text{ reject the } H_0, \text{lean towards } H_a \\ p\text{-value} > \alpha, & - \text{ fail to reject } H_0, \text{ claiming that it is } \textbf{plausible}. \end{cases}$$

Hypothesis Testing: Main Steps.

Illustration (to be filled out during lecture).

Q: What values of $|TS|$ (large? small?) hint at H_0 being false? Why?

Hypothesis Testing: Main Steps.

Example (cont'd). For *Sales* $\sim$ *TV* regression, work through all the steps of the hypothesis test for linear relationship between TV ads and sales.