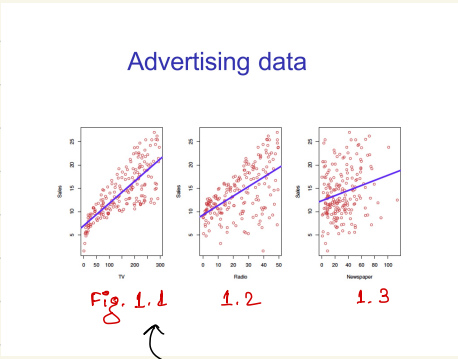
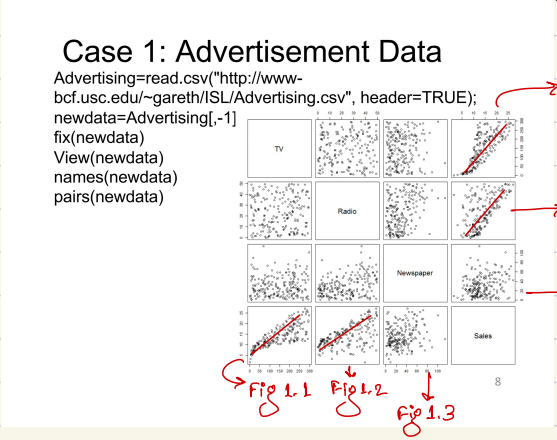
Linear Regression

- Supervised Learning Method
- Predicts response variable from one or more predictors.





* Simple Linear Regression ($p = 1$) : Single predictor

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

ϵ → error term

X → predictor

β_0, β_1 → Parameters

∴ eqⁿ of trained model :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

{ ^ cap denotes estimation }

✡ Residual Sum of Squares :

Residual = Diff. betⁿ actual and pred.

$$e_i = y_i - \hat{y}_i$$

∴ $RSS = e_1^2 + e_2^2 \dots e_n^2$

$$RSS = \frac{1}{n} \Sigma (y_i - \hat{y}_i)^2$$

Estimation of the parameters by least squares

The least squares approach chooses $\beta_0$and $\beta_1$to minimize the RSS. The minimizing values can be shown to be

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = S_{XY}/S_X^2$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x},$$

where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

Sample mean ≠ Population mean

These coef. decide the least square line

Normality of ε

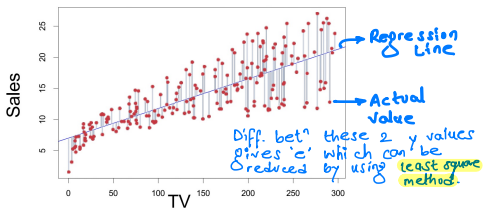
- Note that in the following, the statistical results including confidence intervals, hypothesis testing assume that $\varepsilon$is normally distributed with mean zero and standard deviation σ.

13

Relⁿ betⁿ sample mean and population mean :

✡ $\{ var(\hat{\mu}) = \frac{\sigma^2}{n} = SE(\hat{\mu})^2 \}$

Example: advertising data



The least squares fit for the regression of sales onto TV. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

15

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, = S_{XY}/S_X^2$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

Assessing the Accuracy of the Coefficient Estimates

- The standard error of an estimator reflects how it varies under repeated sampling. We have

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

where $\sigma^2 = \mathrm{Var}(\varepsilon)$

Lower the SEs, better the least square line

16

✡ Confidence interval

95% conf. int. ⇒ 95% prob. of finding actual value in the interval.

∴ Simple LR ($p = 1$)

95% conf. int. ⇒ $\hat{\beta}_1 \pm 2\,\mathrm{SE}(\hat{\beta}_1)$

Confidence intervals — continued

That is, there is approximately a 95% chance that the interval

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$

will contain the true value of β_1 (under a scenario where we got repeated samples like the present sample)

18

- In general, for $p > 1$:

$$\left\{ CI = \left[\hat{\beta}_j - t_{(n-p-1,\alpha/2)} \cdot \hat{\beta}_j + t_{n-p-1,\alpha/2} \right] \right\}$$

Where, $\hat{\beta}_j$ = parameter of j^{th} predictor
n = Sample size
p = no. of parameters
$\therefore$ $n-p-1$ = Degrees of freedom
α = 100 − confidence level
↳ eg. 95% CL $\Rightarrow \alpha = 0.05$

Student's t distribution

Student's t-distribution (or simply the t-distribution) is any member of a family of continuous probability distributions that arises when estimating the mean of a normally distributed population in situations where the sample size is small and the population standard deviation is unknown.

20

Student's t distribution

It was developed by William Sealy Gosset under the pseudonym Student.

The family is parameterized by a parameter ν, which is called the degrees of freedom.

$\nu = n-p-1$

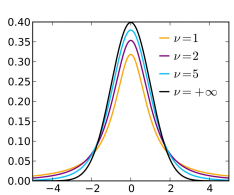
The distribution is bell-shaped and has a zero mean, but its tails are heavier than the standard normal distribution.

$\hat{\mu} = 0$

21

Student's t distribution

t



22

Student's t distribution

When $\nu \to \infty$, $t_\nu \to Z$, where Z is a standard normal distribution, i.e., a normal distribution with mean zero and standard deviation 1.

The black curve on the LEFT.
- The peak (mean) is at '0'

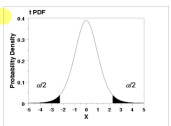
23

Student's t distribution-cut off points

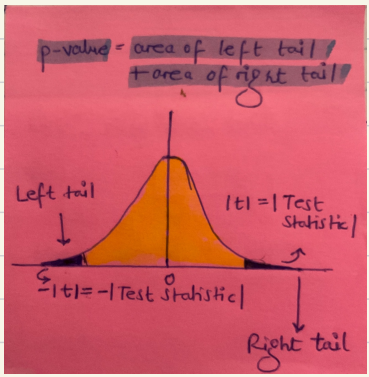
By $t_{n-2,\alpha/2}$, we mean: Imp.

$Pr(t_{n-2} > t_{n-2,\alpha/2}) = \alpha/2$

In other words, the area under the pdf of the t distribution with n-1 degrees of freedom is $\alpha/2$ to the right of $t_{n-2,\alpha/2}$.



24



✡ Hypothesis testing :

a) NULL hypothesis : (H_0)

- It means there is no rel^n bet^n the predictor and response variable.
  { i.e. predictor is insignificant }
- $H_0 \Rightarrow \beta_1 = 0$

NOTE: REJECTING NULL hypothesis means $\beta_1 \neq 0$, means predictor is significant.

b) Alternative hypo. : (H_A)

- $\beta_1 \neq 0$

Hypothesis testing — continued

- To test the null hypothesis, we compute a *t-statistic*, given by

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

- This will have a t-distribution with $n-2$ degrees of freedom, assuming $\beta_1 = 0$.

$n-2 = n-p-1$ (Simple LR)

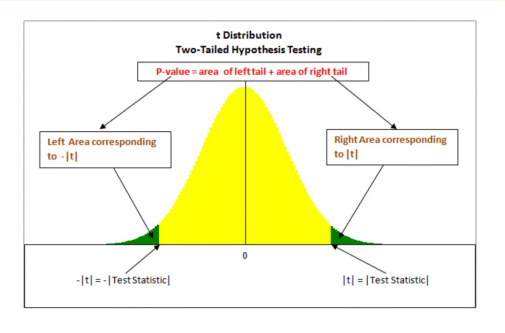
Hypothesis testing — continued

- If the null hypothesis is true, the probability of observing $t > t_{n-2,\alpha/2}$ or $t < t_{n-2,\alpha/2}$ would be α. α is the probability of rejecting a true null hypothesis, i.e. a *Type-I error*, and should be set ahead of time (metaphorically, by your boss). Why? Usually, α is selected to be 5%.

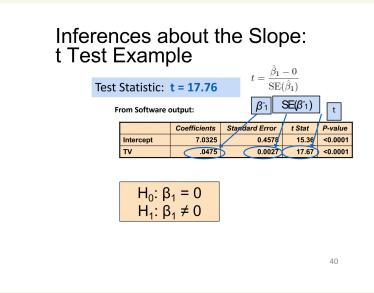
- t-statistic measures no. of std. dev. that β_1 is away from 0

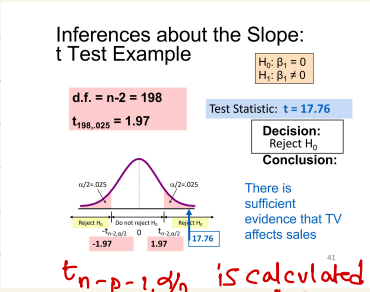
$$\left\{ t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \right\}$$

- t-value ~ p-value
  - Small p-value ⇒ asso. bet^n response & predictor.
    ⇒ Reject NULL hypo.

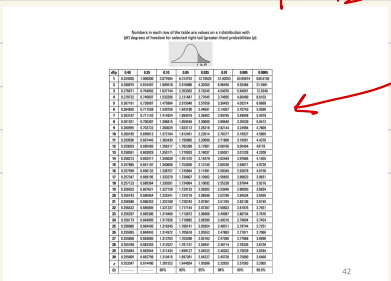


Example:





$t_{n-p-1,\alpha/2}$ is calculated using



Assessing the Overall Accuracy of the Model

- We compute the *Residual Standard Error*

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

where the *residual sum-of-squares* is $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

n-2 considering simple LR

$$RSE = \sqrt{\frac{1}{(n-p-1)}\,RSS}$$

$$= \sqrt{\frac{1}{(n-p-1)}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

RSS = Residual Sum of Squares

★ **Multiple Regression :** $(p > 1)$

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_p X_p$$

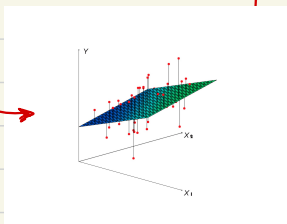
(p predictors)

RSS for multiple regression :

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \dots \hat{\beta}_p x_p)^2$$

Note: In case of Multiple Regression, we have 'regression plane' instead of 'regression line'



✡ Confidence Interval:

Already discussed earlier.

An interval that will contain the true unknown value of the parameter β_i in 1-α percent of times is

$$\left[\hat{\beta}_i - t_{n-p-1,\alpha/2} \cdot \hat{\mathrm{SE}}(\hat{\beta}_i), \hat{\beta}_i + t_{n-p-1,\alpha/2} \cdot \hat{\mathrm{SE}}(\beta_i)\right]$$

✡ Hypothesis Testing:

Same as for $p = 1$

Hypothesis testing — continued

- Using statistical software, it is easy to compute the probability of observing any value equal to |t| or larger. We call this probability the *p-value*.

$$t = \frac{\hat{\beta}_i - \beta_i}{\mathrm{SE}(\hat{\beta}_i)}$$ (β_i: Usually zero)

→

Hypothesis testing — continued

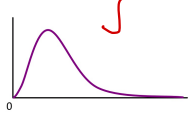
- If the p-value is very small, it means that the probability of seeing a t statistic extremer than what was observed (assuming that $\beta_i = \beta$) is very small. So we reject the null.

Q. Is at least one predictor useful?

For the first question, we can use the F-statistic

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)} \sim F_{p,n-p-1}$$

Quantity	Value
Residual Standard Error	1.69
R^2	0.897
F-statistic	570



Ans: Using F-statistic, if $F > F_{p,n-p-1,\alpha}$,

→ reject NULL hypo. means at least one predictor is significant.

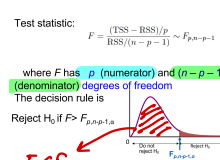
F-Test for Overall Significance

Test statistic: $F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)} \sim F_{p,n-p-1}$

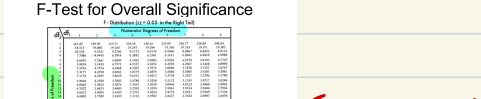
where F has p (numerator) and $(n-p-1)$ (denominator) degrees of freedom

The decision rule is

Reject H_0 if $F > F_{p,n-p-1,\alpha}$



$F < F_{p,n-p-1}$



$F_{p,n-p-1}$ is calculated using this table

F-Test for Overall Significance

F-Distribution (α = 0.05, in the Right Tail)

Numerator Degrees of Freedom

Denominator Degrees of Freedom

✡ Example:

F-Test for Overall Significance

p = 3

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$
H_1: Not all three of β_1, β_2, β_3 are zero

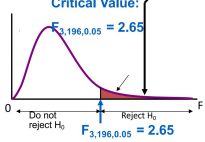
Test Statistic: F=570

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)} \sim F_{p,n-p-1}$$

$df_1 = 3$ $df_2 = 200\text{-}3\text{-}1$

Critical Value:

$F_{3,196,0.05} = 2.65$



Decision:

Since F test statistic is in the rejection region (p-value < .05), reject H_0

Conclusion:

There is evidence that at least one independent variable affects Y

80

Note:

Deciding on the important variables

- The most direct approach is called *all subsets* or *best subsets* regression: we compute the least squares fit for all possible subsets and then choose between them based on some criterion that balances training error with model size.

81

Deciding on the important variables

- However we often can't examine all possible models, since they are 2^p of them; for example when $p = 40$ there are over a trillion models!
- Instead we need an automated approach that searches through a subset of them. We discuss two commonly use approaches next.

82

Forward selection

- Add to that model the variable that results in the lowest RSS amongst all two-variable models.
- Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold.

AND

Backward selection

- Start with all variables in the model.
- Remove the variable with the largest p-value — that is, the variable that is the least statistically significant.
- The new ($p - 1$)-variable model is fit, and the variable with the largest p-value is removed.
- Continue until a stopping rule is reached. For instance, we may stop when all remaining variables have a significant p-value defined by some significance threshold.

Linear Regression for qualitative data :

Qualitative Data = Categorical Data

Qualitative Predictors — cont'd

Example: investigate differences in credit card balance between males and females, ignoring the other variables. We create a new variable

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

Resulting model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

Intrepretation?

Credit card data — continued

Results for gender model:

	Coefficient	Std. Error	t-statistic	p-value
Intercept	509.80	33.13	15.389	< 0.0001
gender[Female]	19.73	46.05	0.429	0.6690

Qualitative predictors with more than two levels

- With more than two levels, we create additional dummy variables. For example, for the ethnicity variable we create two dummy variables. The first could be

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian,} \end{cases}$$

and the second could be

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian.} \end{cases}$$

Qualitative predictors with more than two levels

- Then both of these variables can be used in the regression equation, in order to obtain the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is AA.} \end{cases}$$

For ex., here there are 2 dummy variable (3-1) as there are 3 classes:

① Asian
② Caucasian
③ Caucasian

Qualitative predictors with more than two levels

One less than total no. of classes

- There will always be one fewer dummy variable than the number of levels. The level with no dummy variable — African American in this example — is known as the *baseline*.

Interactions and non-linearity : (Multinomial LR)

Extensions of the Linear Model

Removing the additive assumption: *interactions* and *nonlinearity*

Interactions:

- In our previous analysis of the Advertising data, we assumed that the effect on sales of increasing one advertising medium is independent of the amount spent on the other media.

It may be possible, that 2 or more more types of media collaboratively affect the sales.

Outcomes of interaction :

Modelling interactions — Advertising data

Model takes the form

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times (\text{radio} \times \text{TV}) + \varepsilon$$
$$= \beta_0 + (\beta_1 + \beta_3 \times \text{radio}) \times \text{TV} + \beta_2 \times \text{radio} + \varepsilon$$

Results:

	Coefficient	Std. Error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

p-value of TV x Radio shows that is is significant.

Extensions of the Linear Model

- For example, the linear model

$$\widehat{\text{sales}} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper}$$

states that the average effect on sales of a one-unit increase in TV is always β_1, regardless of the amount spent on radio.

- But suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for TV should increase as radio increases.
- In this situation, given a fixed budget of \$100, 000, spending half on radio and half on TV may increase sales more than allocating the entire amount to either TV or to radio.
- In marketing, this is known as a *synergy* effect, and in statistics it is referred to as an *interaction* effect.

Hierarchical Principle:

Hierarchy

- Sometimes it is the case that an interaction term has a very small p-value, but the associated main effects (in this case, TV and radio) do not.
- The *hierarchical principle*:

  If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.

Imp.

Interpretation

- The R^2 for the interaction model is 96.8%, compared to only 89.7% for the model that predicts sales using TV and radio without an interaction term.
- This means that (96.8 − 89.7)/(100 − 89.7) = 69% of the variability in sales that remains after fitting the additive model has been explained by the interaction term.

Interaction between Quantitative and Qualitative Variables

Consider the Credit data set, and suppose that we wish to predict balance using income (quantitative) and student (qualitative).

Without an interaction term, the model takes the form

$$\begin{aligned} \texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases} \\ &= \beta_1 \times \texttt{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student.} \end{cases} \end{aligned}$$

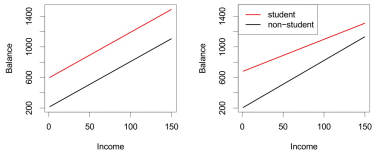
Interaction between Quantitative and Qualitative Variables

With interactions, it takes the form

$$\begin{aligned} \texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases} \\ &= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \texttt{income}_i & \end{cases} \end{aligned}$$

Qualitati

- Challe with c
- A cat levels of the chan on m varia some occu

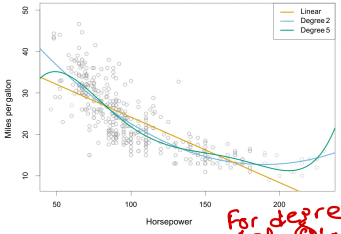


Credit data; Left: no interaction between income and student. Right: with an interaction term between income and student.

Considering if (Right panel) improve the of the model

Non-linear effects of predictors

polynomial regression on Auto data



The figure suggests that

$\text{mpg} = \beta_0 + \beta_1 \times \text{horsepower} + \beta_2 \times \text{horsepower}^2 + \varepsilon$

may provide a better fit.

	Coefficient	Std. Error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	-0.4662	0.0311	-15.0	< 0.0001
horsepower^2	0.0012	0.0001	10.1	< 0.0001

For degree = 5 model fits more datapoints (more flexible)

Considering interaction of 'horsepower with itself can also improve the model.

Challenges faced with categorical data:

1) Categorical data with more than 2 levels.

- A categorical variable has too many levels. This pulls down performance level of the model. For example, a cat. variable “zip code” would have numerous levels.
- A categorical variable has levels which rarely occur. Many of these levels have minimal chance of making a real impact on model fit. For example, a variable ‘disease’ might have some levels which would rarely occur.

Soln: Convert to label encoded using encoding technique.

Methods to deal with Qualitative/ Categorical Variables

Label Encoder:

```
In [53]: train.head(5)
```

Out[53]:

	sex	pclass
0	male	3
1	female	1
2	female	3
3	female	1
4	male	3

```
In [54]: from sklearn.preprocessing import LabelEncoder

number = LabelEncoder()
train['sex'] = number.fit_transform(train['sex'].astype('str'))
test['sex'] = number.fit_transform(test['sex'].astype('str'))

train.head(5)
```

Out[54]:

	sex	pclass
0	1	3
1	0	1
2	0	3
3	0	1
4	1	3