# Week 11

# Logistic Regression and Non-parametric Regression

# F tests

- Two types of F test:
  - Overall F test – test for the usefulness of the model
  - Partial F test – test for linear restrictions
- $S_y^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} = \frac{S_{yy}}{n-1} = \frac{TSS}{n-1}$

  $S_{yy} = TSS = (n - 1)\, S_y^2$ (total variation in Y)
- TSS = RegSS + RSS

| y | X1 | X2 | X3 | X4 |
|---|---|---|---|---|
| 69 | 6 | 1 | 2 | 1 |
| 118.5 | 10 | 1 | 2 | 2 |
| 116.5 | 10 | 1 | 3 | 2 |
| 125 | 11 | 1 | 3 | 2 |
| 129.9 | 13 | 1 | 3 | 1.7 |
| 135 | 13 | 2 | 3 | 2.5 |
| 139.9 | 13 | 1 | 3 | 2 |
| 147.9 | 17 | 2 | 3 | 2.5 |
| 160 | 19 | 2 | 3 | 2 |
| 169.9 | 18 | 1 | 3 | 2 |
| 134.9 | 13 | 1 | 4 | 2 |
| 155 | 18 | 1 | 4 | 2 |
| 169.9 | 17 | 2 | 4 | 3 |
| 194.5 | 20 | 2 | 4 | 3 |
| 209.9 | 21 | 2 | 4 | 3 |

```
> reg1 <- lm(y ~ X1+X2+X3+X4, data=hprice)
> summary(reg1)

Call:
lm(formula = y ~ X1 + X2 + X3 + X4, data = hprice)

Residuals:
    Min      1Q  Median      3Q     Max
-12.700  -1.616   0.984   2.510  11.759

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.7633     9.2074   2.038  0.06889 .
X1            6.2698     0.7252   8.645 5.93e-06 ***
X2          -16.2033     6.2121  -2.608  0.02611 *
X3           -2.6730     4.4939  -0.595  0.56519
X4           30.2705     6.8487   4.420  0.00129 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.849 on 10 degrees of freedom
Multiple R-squared:  0.9714,    Adjusted R-squared:  0.9599
F-statistic:  84.8 on 4 and 10 DF,  p-value: 1.128e-07
```

Multiple correlation

$\text{Corr}(Y, \hat{Y}) = \text{Corr}[(Y, (X_1, X_2, X_3, X_4)] = \sqrt{0.9714} = 0.9856$

```
> round(cor(hprice$y, reg1$fitted.values)^2, 4)
[1] 0.9714
```

# Multiple $R^2$ versus Multiple R

- $R^2 = \frac{\text{RegSS}}{\text{TSS}}$, $0 \le R^2 \le 1$
- Another way of viewing $R^2$ is to note that it is the squared correlation of Y (observed Y's) and $\hat{Y}$ (predicted Y's), which are of course a linear combination of the X's.
- $R^2 = \frac{(S_{Y\hat{Y}})^2}{S_{YY}S_{\hat{Y}\hat{Y}}}$ where

  $S_{Y\hat{Y}} = \sum Y\hat{Y} - \frac{(\sum Y)(\sum \hat{Y})}{n}$, $S_{YY} = \sum Y^2 - \frac{(\sum Y)^2}{n}$, and $S_{\hat{Y}\hat{Y}} = \sum \hat{Y}^2 - \frac{(\sum \hat{Y})^2}{n}$

- The positive square root of $R^2$ gives R. Unlike r, which can take positive as well as negative values, R may vary from 0 to 1. The closer the value of R to 1, the greater the linear relationship between the independent variables and the dependent variable.
- R = 1 indicates that the predictions are exactly correct.
- R = 0 indicates that no linear combination of the independent variables is a better predictor than is the fixed mean of the dependent variable.

# Multiple $R^2$ versus Multiple R

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.08411    0.12136  132.53 1.17e-14 ***
x           -1.76387    0.03155  -55.91 1.16e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02091 on 8 degrees of freedom
Multiple R-squared:  0.9974,    Adjusted R-squared:  0.9971
F-statistic:  3126 on 1 and 8 DF,  p-value: 1.162e-11
```

$\sqrt{\text{Multiple } R^2}$ = Corr(Y, $\hat{Y}$) = $\sqrt{0.9974}$ = 0.9987

$0 \le \text{Corr}(Y, \hat{Y}) \le 1$

For simple linear regression,

$\hat{\beta}_1 = r\frac{S_y}{S_x}$

Since $\frac{S_y}{S_x} > 0$, r and $\hat{\beta}_1$ must have the same sign. r > 0, then $\hat{\beta}_1 > 0$; r < 0, then $\hat{\beta}_1 < 0$.

$\hat{\beta}_1 = -1.76387$

Corr(X, Y) = -0.9987.

# ANOVA table

```
> reg1 <- lm(y ~ X1+X2+X3+X4, data=hprice)
> summary(reg1)

Call:
lm(formula = y ~ X1 + X2 + X3 + X4, data = hprice)

Residuals:
    Min      1Q  Median      3Q     Max
-12.700  -1.616   0.984   2.510  11.759

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.7633     9.2074   2.038  0.06889 .
X1            6.2698     0.7252   8.645 5.93e-06 ***
X2          -16.2033     6.2121  -2.608  0.02611 *
X3           -2.6730     4.4939  -0.595  0.56519
X4           30.2705     6.8487   4.420  0.00129 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.849 on 10 degrees of freedom
Multiple R-squared:  0.9714,    Adjusted R-squared:  0.9599
F-statistic:  84.8 on 4 and 10 DF,  p-value: 1.128e-07
```

```
> anova(reg1)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq  F value   Pr(>F)
X1         1 14829.3 14829.3 316.1025 6.76e-09 ***
X2         1     0.9     0.9   0.0184 0.894652
X3         1   166.4   166.4   3.5472 0.089023 .
X4         1   916.5   916.5  19.5356 0.001294 **
Residuals 10   469.1    46.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

RegSS = 14829.3 + 0.9 +166.4 + 916.5 = 15913.1

TSS = RegSS + RSS = 15913.1 + 469.1 = 16382.2

| Source | SS | df | MS | F |
|---|---|---|---|---|
| Regression | 15913.1 | 4 | 3978.275 | 84.8 |
| Residual | 469.1 | 10 | 46.91 | |
| Total | 16382.2 | 14 | | |

# Partial F test vs Individual t test

```
> fm <- lm(y ~ X1+X2+X3+X4, data=hprice)
> summary(fm)

Call:
lm(formula = y ~ X1 + X2 + X3 + X4, data = hprice)

Residuals:
    Min      1Q  Median      3Q     Max
-12.700  -1.616   0.984   2.510  11.759

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.7633     9.2074   2.038  0.06889 .
X1            6.2698     0.7252   8.645 5.93e-06 ***
X2          -16.2033     6.2121  -2.608  0.02611 *
X3           -2.6730     4.4939  -0.595  0.56519
X4           30.2705     6.8487   4.420  0.00129 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.849 on 10 degrees of freedom
Multiple R-squared:  0.9714,    Adjusted R-squared:  0.9599
F-statistic:  84.8 on 4 and 10 DF,  p-value: 1.128e-07
```

- Test $H_0$: $\beta_1 = 0$ versus $H_1$: $\beta_1 \neq 0$
- $F_{stat} = \frac{(RSS_r - RSS_f)/q}{RSS_f/(n-p)} = 74.741$

```
> rm <- lm(y ~ X2+X3+X4, data=hprice)
> summary(rm)

Call:
lm(formula = y ~ X2 + X3 + X4, data = hprice)

Residuals:
    Min      1Q  Median      3Q     Max
-20.306 -14.133   2.254   6.364  36.028

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.709     25.261   0.266    0.795
X2             3.129     16.086   0.195    0.849
X3            17.626     10.635   1.657    0.126
X4            35.578     18.932   1.879    0.087 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.01 on 11 degrees of freedom
Multiple R-squared:  0.7573,    Adjusted R-squared:  0.6911
F-statistic: 11.44 on 3 and 11 DF,  p-value: 0.001043

> fm <- lm(y ~ x1+x2+x3+x4, data=dat)
> rm <- lm(y ~ x2+x3+x4,data=dat)
> fobs <- ((deviance(rm)-deviance(fm))/1)/(deviance(fm)/10)
> c(deviance(rm), deviance(fm), fobs)
[1] 3975.44472  469.12948   74.74089
> pf(fobs,1,10,lower.tail=F)
[1] 5.931552e-06
> anova(rm,fm)
Analysis of Variance Table

Model 1: y ~ x2 + x3 + x4
Model 2: y ~ x1 + x2 + x3 + x4
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     11 3975.4
2     10  469.1  1    3506.3 74.741 5.932e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

# Linear probability model

- Up until now, we have assumed that the dependent variable is continuous (e.g. quantities, prices, etc.).
- However, many choices cannot be measured by continuous variables but rather are of either/or nature
- Examples:  to go to uni or not; to purchase a house or rent; to approve loan application or not; to vote for the labour party or not, etc.
- We want to explain why such choices are made, what factors enter into the decision process, and how much each factor affects the outcome. Sometimes we want to predict such choices.
- Such choices lead to models in which the dependent variable Y is binary in nature (i.e. is equal to either zero or one).
- In a model where Y is continuous, our objective is to estimate its expected, or mean, value given the values of the regressors; i.e., we want $E(Y \mid X_1, X_2, \ldots, X_k)$, where X's can be qualitative or quantitative.
- In models where Y is binary, our objective is to estimate the **probability** of something happening; i.e., $P(Y = 1 \mid X_1, X_2, \ldots, X_k)$.  Hence, the binary response regression models are often known as probability models.

# Binary Models

- We first consider the binary response regression model. There are 3 approaches to developing a probability model for a binary response variable:
  - The **linear probability model** (LPM)
  - The **logit model**
  - The **Probit model**

# The Linear Probability Model (LPM)

- Consider $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ (1)
  - X = family income
  - $Y_i$ = 1 if the family owns a house, 0 if it does not
  - ε is a random error with E(ε|X) = 0
- Because of the binary nature of the dependent variable model (1) is called Linear Probability Model (LPM)
- Let $P_i$ = probability that $Y_i$ = 1 and $(1 - P_i)$ = probability that $Y_i$ = 0. The distribution of Y is as follows:

| $Y_i$ | Probability |
|---|---|
| 0 | 1- $P_i$ |
| 1 | $P_i$ |
| Total | 1 |

- Hence, $Y_i$ follows the Bernoulli probability distribution with $E(Y_i) = P_i = P(Y_i = 1)$

# The Linear Probability Model (LPM)

- Hence, in the LPM (1), $\mathrm{E}(\mathrm{Y_i}|\mathrm{X_i}) = \beta_0 + \beta_1 \mathrm{X_i} = \mathrm{P_i}$

- Therefore, in LPM the conditional expectation of $Y_i$ given $X_i$, $\mathrm{E}(\mathrm{Y_i}|\mathrm{X_i})$ can be interpreted as the conditional probability that the event will occur given $X_i$; i.e. $\mathrm{P}(\mathrm{Y_i} = 1|\mathrm{X_i})$.
  - In the example, $\mathrm{E}(\mathrm{Y_i}|\mathrm{X_i})$ gives the probability of a family owning a house conditional that income is the given amount $X_i$.
- Does the LPM model satisfy the classical linear normal regression normal (CLNRM) assumptions to be reliably estimable by OLS? It turns out that it does not!
- Extending OLS to binary dependent variable regression models poses several problems.

# Linear Probability Model

**Non-normality of the random errors $\varepsilon_i$**

- The assumption of normality for $\varepsilon_i$ is not tenable for the LPMs because, like $Y_i$, the random errors also take only 2 values.
- If we rewrite our model as: $\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$, the probability distribution of $\varepsilon_i$ is

| | $\varepsilon_i$ | **Probability** |
|---|---|---|
| $Y_i = 1$ | $1 - \beta_0 - \beta_1 X_i$ | $P_i$ |
| $Y_i = 0$ | $- \beta_0 - \beta_1 X_i$ | $(1 - P_i)$ |

- Thus $\varepsilon_i$ cannot be assumed to be normally distributed. Rather, they follow Bernoulli distribution.
- Is the nonfulfillment of the normality assumption so critical?
  - We know that the OLS point estimates still remain unbiased.
  - As the sample size increases indefinitely, OLS estimators tend to be normally distributed
  - Thus in large samples the statistical inference of the LPM will follow the usual OLS procedure under the normality assumption

# Linear Probability Model

**Heteroscedastic Variances of the Disturbances**

- It can no longer be maintained that in the LPM the random errors are homoscedastic.
- The variance of the error term $\varepsilon_i$ (which follows Bernoulli distribution) is: $\text{Var}(\varepsilon_i) = P_i(1 - P_i)$
- Since $P_i = E(Y_i|X_i) = \beta_0 + \beta_1 X_i$, the variance of $\varepsilon_i$ ultimately depends on the values of X and hence is not homoscedastic:
- $\text{Var}(\varepsilon_i) = P_i(1 - P_i) = (\beta_0 + \beta_1 X_i)(1 - \beta_0 - \beta_1 X_i)$
- Hence, OLS estimators will be unbiased but inefficient

# Linear Probability Model

- **Nonfulfillment of $\mathbf{0 \leq E(Y_i|X_i) \leq 1}$**
- Since $E(Y_i|X_i)$ in the linear probability model measures the conditional probability of the event Y occurring given X, it must lie between 0 and 1, inclusive (as any probability).
- However, there is no guarantee that $\widehat{Y}_i$, the estimators of $\mathbf{E(Y_i|X_i)}$, will necessarily fulfil this restriction, and this is the real problem with the OLS estimation of the LPM.
- This happens because OLS does not take into account the restriction that $\mathbf{0 \leq E(Y_i|X_i) \leq 1.}$

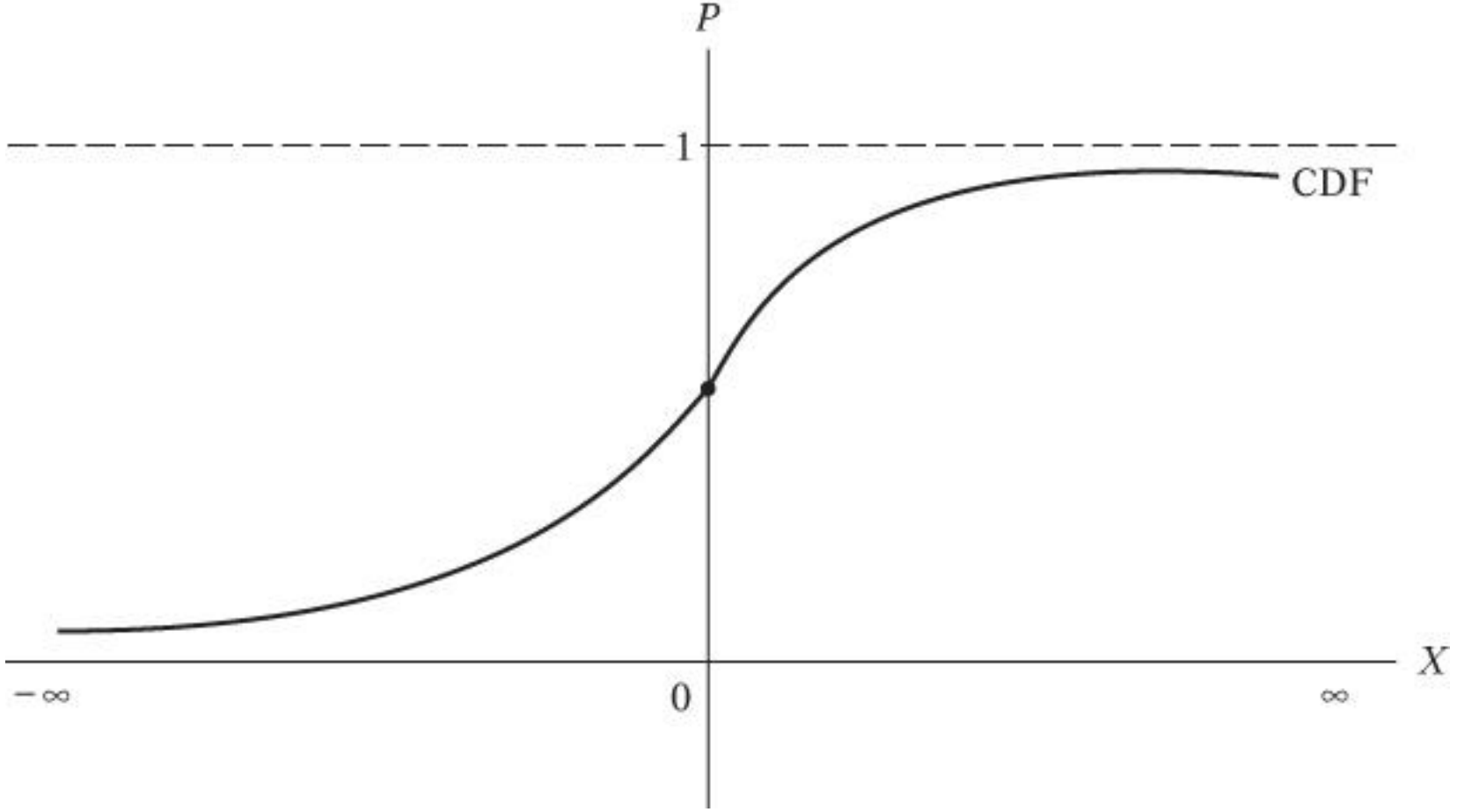
# Alternatives to LPM

- LPM has several problems:
  - Non-normality of $\varepsilon_i$ (not crucial if sample is large)
  - Heteroscedasticity of $\varepsilon_i$ (can be mitigated by GLS)
  - Possibility of $\widehat{Y}_i$ lying outside the 0-1 range (can adjust fitted values)
- The fundamental problem with the LPM is that it is not very attractive because it assumes that $P_i = E(Y = 1|X)$ increases linearly with X, i.e. the marginal effect of X remains constant throughout. That is, a given increase in X leads to the same change in $P_i$, for large and low values of X. Unrealistic!

# Alternatives to LPM

- We need a probability model that has 2 features:
  - As $X_i$ increases, $P_i = E(Y = 1|X)$ increases but never steps outside the 0-1 interval.
  - The relationship between $P_i$ and $X_i$ is nonlinear.
- Geometrically, we want something like the following diagram (next graph):
  - The probability lies between 0 and 1 and varies nonlinearly with X
  - Resembles the cumulative distribution function (CDF) of a random variable
- But – which CDF?  Logistic (logit model) or normal (probit model).

# A cumulative distribution function



# Odds of an event

Odds are used to express the likelihood of an event and are expressed as the **ratio** of number of successes (s) to the number of failures (f) or vice versa.

Odds in favour:

$O_s(\text{event}) = \frac{\text{number of successes}}{\text{number of failures}}$ or number of successes **:** number of failures

Odds against:

$O_f(\text{event}) = \frac{\text{number of failures}}{\text{number of successes}}$ or number of failures **:** number of successes

Example

A jewellery box contains 5 white pearl, 2 gold rings, and 6 silver rings. What are the odds of drawing a white pearl from the jewellery box?

$O_s(\text{white pearl}) = \frac{5}{8}$ or 5 : 8

So, there are 5 chances that the event will happen to every 8 chances that it will not happen.

# Odds ratio

The odds ratio (OR) of an event compares the odds, conditional on a second event occurring, to the conditional odds when that second event does not occur; i.e.,

$$\text{OR(event 1 occurs)} = \frac{\frac{P(\text{event 1 occur} \mid \text{event 2 occurs})}{1 - P(\text{event 1 occur} \mid \text{event 2 occurs})}}{\frac{P(\text{event 1 occur} \mid \text{event 2 does not occur})}{1 - P(\text{event 1 occur} \mid \text{event 2 does not occur})}} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{ad}{bc}$$

| | Event 2 | | |
|---|---|---|---|
| Event 1 | Occur | Do not occur | Total |
| Occur | a | b | a + b |
| Do not occur | c | d | c + d |
| | a + c | b + d | n |

# Odds ratio

The following table shows the number of students at a high school who became ill after eating lunch bought at the school cafeteria.

$$\text{OR} = \frac{\frac{\text{P(ill} \mid \text{ate sandwich)}}{1 - \text{P(ill} \mid \text{ate sandwich)}}}{\frac{\text{P(ill} \mid \text{did not eat sandwich)}}{1 - \text{P(ill} \mid \text{did not eat sandwich)}}} = \frac{\frac{109}{116}}{\frac{4}{34}} = 7.987$$

| | Ate Sandwich | | |
|---|---|---|---|
| Ill | Yes | No | Total |
| Yes | 109 | 4 | 113 |
| No | 116 | 34 | 150 |
| | 225 | 38 | 263 |

The odds of becoming ill for students who ate sandwiches are 7.987 higher than the odds for students who did not eat the sandwiches.

# Probability versus odds

Probability and odds are not the same thing. They contain the same information, but they express it differently.

Probability and odds can be interconverted using the following formulas:

$$\text{Odds} = \frac{\text{probability}}{1 - \text{probability}}$$

So, the odds are the ratio of 2 complementary probabilities.

$$\text{Probability} = \frac{\text{odds}}{1 + \text{odds}}$$

While probabilities are bound to [0, 1], odds are bound to [0, ∞).

- A probability of 0 is the same as odds of 0.
- A probability of 0.5 is the same as odds of 1.
- Probabilities between 0 and 0.5 equal odds less than 1.
- As the probability goes up from 0.5 to 1, the odds increase from 1 to approach infinity.
- High odds correspond to high probabilities, low odds to low probabilities.

# Binary Dependent Variables

- Recall the linear probability model, which can be written as

  $P(Y = 1|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$

- A drawback to the linear probability model is that predicted values are not constrained to lie between 0 and 1
- An alternative is to use model the probability as a function F:
  - $P(Y = 1|X) = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)$
  - F(.) is known as a transformation. Let $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$. It is a (nonlinear) function that has the following 2 good properties:
    i. $0 < F(z) < 1$ for all values of z
    ii. F(z) is increasing in z.
  - In the above model, the predicted probability will never be “absurd”!
  - How to choose F(z)?

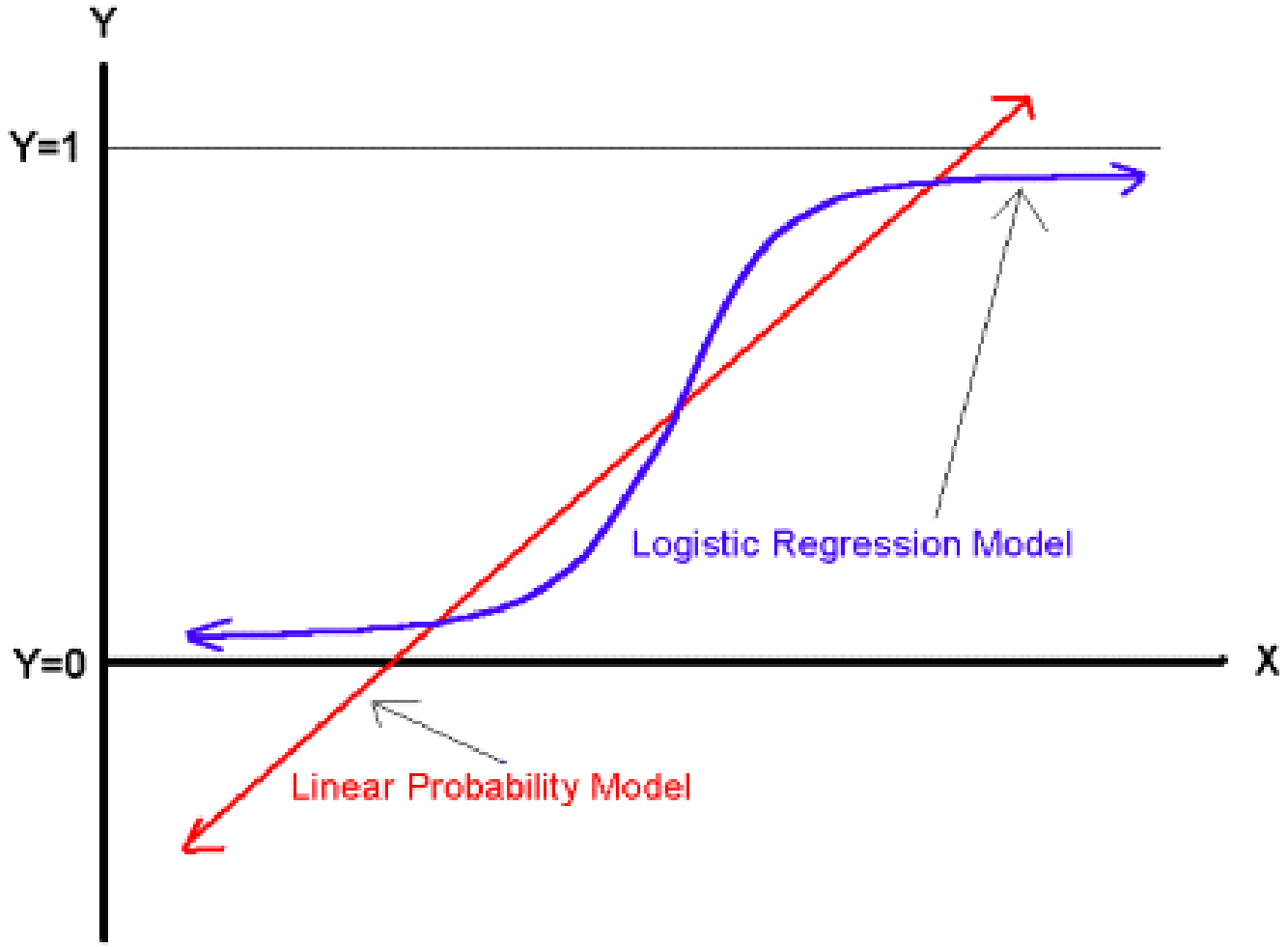
# The Logit Model / logistic regression model

- One choice for F(z) is the logistic function, which is the cdf for a standard logistic random variable.
- $F(z) = \frac{\exp(z)}{1 + \exp(z)} = \Lambda(z)$
- Hence, $P(Y = 1 \mid X) = p = \Lambda(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)$

$$= \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}$$

- In logistic regression, a logistic transformation of the odds (referred to as logit) served as the dependent variable; i.e.,
- $\ln(\text{odds}) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$
- The left hand side (L.H.S.) of the equation is called the “logit function” or “log odds”, which is the $\log_e$ of the odds.
- $p = \frac{\text{odds}}{1 + \text{odds}}$ or $\text{odds} = \frac{p}{1-p}$

# Comparing the LP and Logit Models



# Example 1

A study was designed to compare two energy drink commercials. Each participant was shown the commercials, A and B, in random order and asked to select the better one. There were 100 women and 140 men who participated in the study. Commercial A was selected by 45 women and by 80 men.

| | Commercials | | |
|---|---|---|---|
| Gender | A | B | |
| Women | 45 | 55 | 100 |
| Men | 80 | 60 | 140 |
| | 125 | 115 | 240 |

- Find the log odds of selecting Commercial A for the men.
- Find the log odds of selecting Commercial A for the women.
- Find ratio of the odds that a man prefers A to the odds that a woman prefers A.
- Let X = 1 if men; 0 otherwise. If the response variable is the proportion of Commercial A, find the estimated logistic regression equation.

# Example 1

| | Commercials | | |
|---|---|---|---|
| Gender | A | B | |
| Women | 45 | 55 | 100 |
| Men | 80 | 60 | 140 |
| | 125 | 115 | 240 |

- Find the log odds of selecting Commercial A for the men.

For men, odds $= \frac{80}{60} = \frac{4}{3}$

$\log(\text{odds})_m = 0.2877$

- Find the log odds of selecting Commercial A for the women.

For women, odds $= \frac{45}{55} = \frac{9}{11}$

$\log(\text{odds})_w = -0.2007$

# Example 1

| | Commercials | | |
|---|---|---|---|
| Gender | A | B | |
| Women | 45 | 55 | 100 |
| Men | 80 | 60 | 140 |
| | 125 | 115 | 240 |

- Find ratio of the odds that a man prefers A to the odds that a woman prefers A.

$$OR = \frac{\frac{4}{3}}{\frac{9}{11}} = \frac{44}{27} = 1.63$$

The odds for men are 1.63 times the odds for women.

ln(OR) = ln(1.63) = 0.4884

# Example 1

- Let X = 1 if men; 0 otherwise. If the response variable is the proportion of Commercial A, find the estimated logistic regression equation.

```
> y = c(1,1,0,0)
> x = c(0,1,0,1)
> Freq = c(45,80,55,60)
> comm <- data.frame(y,x,Freq)
> fit <- glm(y ~ x, weights = Freq, data = comm, family = binomial)
> summary(fit)

Call:
glm(formula = y ~ x, family = binomial, data = comm, weights = Freq)

Deviance Residuals:
      1       2       3       4
  8.477   9.463  -8.109 -10.083

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.2007     0.2010  -0.998   0.3181
x             0.4884     0.2638   1.851   0.0641 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 332.29  on 3  degrees of freedom
Residual deviance: 328.84  on 2  degrees of freedom
AIC: 332.84

Number of Fisher Scoring iterations: 4
```

$$\ln(\widehat{\text{odds}}) = -0.2007 + 0.4884X$$

# Example 2

- Military pilots sometimes black out when their brains are deprived of oxygen due to G-Forces during violent manoeuvres. The following data was obtained by producing similar symptoms to volunteers by exposing their lower body to negative air pressure.
- The response variable has only 2 possible outcomes (yes/no) and the explanatory variable (age) is numerical/continuous.

| Subject | JW | JM | DT | LK | JK | MK | FP | DG |
|---|---|---|---|---|---|---|---|---|
| Signs (Y) | No | Yes | No | Yes | Yes | No | Yes | Yes |
| Age (X) | 39 | 42 | 20 | 37 | 20 | 21 | 41 | 51 |

# Example 2

```
> y <- c(0,1,0,1,1,0,1,1)
> x <- c(39,42,20,37,20,21,41,51)
>
> reg1 <- glm(y ~ x, family="binomial"); summary(reg1)

Call:
glm(formula = y ~ x, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7152  -0.8686   0.5134   0.6911   1.5336

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.9305     2.6535  -1.104    0.269
x             0.1062     0.0806   1.317    0.188

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10.5850  on 7  degrees of freedom
Residual deviance:  8.4502  on 6  degrees of freedom
AIC: 12.45

Number of Fisher Scoring iterations: 4
```

The model for showing signs of blackout:

$\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$ = -2.9305 + 0.1062X

$\hat{\beta}_1$ = 0.1062

An increase of 1 year unit in Age will increase the odds of showing signs of blackout by a factor of $e^{0.1062}$ = 1.112

Estimate the probability of a test subject showing signs of blackout if they were 30 years old.

$\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$ = -2.9305 + 0.1062(30) = 0.2555

Odds = $\frac{\hat{p}}{1-\hat{p}} = e^{0.2555}$

$\hat{p} = \frac{\widehat{odds}}{1+\widehat{odds}} = \frac{e^{0.2555}}{1+e^{0.2555}} = 0.5635$

The probability of showing signs of blackout, given the pilot is 30 years old is approximately equal to 0.5635.

# Example 2 – Inference for Logistic Model

```
> y <- c(0,1,0,1,1,0,1,1)
> x <- c(39,42,20,37,20,21,41,51)
>
> reg1 <- glm(y ~ x, family="binomial"); summary(reg1)

Call:
glm(formula = y ~ x, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7152  -0.8686   0.5134   0.6911   1.5336

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.9305     2.6535  -1.104    0.269
x             0.1062     0.0806   1.317    0.188

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10.5850  on 7  degrees of freedom
Residual deviance:  8.4502  on 6  degrees of freedom
AIC: 12.45

Number of Fisher Scoring iterations: 4
```

z value is given by $z^* = \dfrac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$.

For large sample size, $n$, we assume $z^* \sim N(0,1)$ and so for $H_0 : \beta_j = 0$, the $p$-value is given by $P(|Z| > |z^*|)$.

Why don't we use a $t$-test? We ought to but this is not straight forward for logistic regression so care should be taken for small $n$.

# Example 2 – Inference for Logistic Model

```
> y <- c(0,1,0,1,1,0,1,1)
> x <- c(39,42,20,37,20,21,41,51)
>
> reg1 <- glm(y ~ x, family="binomial"); summary(reg1)

Call:
glm(formula = y ~ x, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7152  -0.8686   0.5134   0.6911   1.5336

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.9305     2.6535  -1.104    0.269
x             0.1062     0.0806   1.317    0.188

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10.5850  on 7  degrees of freedom
Residual deviance:  8.4502  on 6  degrees of freedom
AIC: 12.45

Number of Fisher Scoring iterations: 4
```

We want to test if there is a relationship between age and signs of blackout.

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \varepsilon$$

$H_0$: $\beta_1 = 0$ against $H_1$: $\beta_1 \neq 0$

Decision rule based on p-value:

Reject $H_0$ if p-value < α

Decision:

Retain $H_0$ because p-value = 0.188 ⋪ α = 0.05

Conclusion:

There is not sufficient evidence to show a relationship between age and signs of blackouts.

# Example 3

Before a bank approves your application for a loan (personal, car, home, etc.), a credit card, etc., it will use your information provided to assess whether you are trustworthy. Suppose that the bank finds that the age, $X_1$, income, $X_2$ (in \$'000), time at current job, $X_3$ (in years) and time at current address, $X_4$ (in years) can be used to determine the probability that a loan is repaid (Y = 1) (Y = 0 for default). Based on a random sample of observations, a logistic regression analysis gives

$\hat{\beta}_0 = 0.1524$, $\hat{\beta}_1 = 0.0281$, $\hat{\beta}_2 = 0.0223$, $\hat{\beta}_3 = 0.0152$, and $\hat{\beta}_4 = 0.0114$.

For a 48-year-old applicant with \$78,000 annual income, at current job for 3 years and at the current address for 12 years, is his/her application likely to be successful?

$\ln(\widehat{\text{Odds}}) = 0.1524 + 0.0281(48) + 0.0223(78) + 0.0152(3) + 0.0114(12) = 3.423$

$\widehat{\text{Odds}} = e^{3.423}$

$\hat{p} = \frac{\widehat{\text{Odds}}}{1+\widehat{\text{Odds}}} = \frac{e^{3.423}}{1+e^{3.423}} = 0.9684$

The probability of repaying the loan for this applicant is predicted to be close to 100%. This application should be successful.

# Note:

In linear regression, the coefficients describe the relationship between each of the independent variables and dependent variable.

- A positive coefficient of an independent variable means that when that variable increases by 1 unit, on average, the dependent variable will increase by the amount of the coefficient (holding all the other variables constant).
- If the coefficient is negative, an increase of 1 unit in the independent variable will result in a decrease by the amount equal to the coefficient in the dependent variable (holding all the other variables constant).

Interpreting the coefficients of the logistic regression is somewhat more complex. Rather than interpreting the magnitude of the coefficient, we will instead describe the effect of the sign.

- If the coefficient is positive, an increase in that independent variable will result in an increase in the probability of the event.