

SURV686-HW4

Sagnik Chakravarty

I Pledge on my honor that I have not given or
received any unauthorized assistance on this
assignment/examination.


Signature: 
Date: 02/16/2025

Table of contents

Question 1	3
a. Create an empirical logit plot with the response variable (y) subscribed by the predictor variable (x) age. You may want to convert subscribed to a numeric variable first.	3
b. Report the proportion (tabular format is fine) subscribed="yes" for each of the categories for job, marital, education, default, housing, loan, and contact.	5
c. Plot the response variable, proportion subscribed="yes", for each of the following values of campaign: 1, 2, 3, 4, 5, 6+	8
d. Next, we want to evaluate if campaign contacts are effective. Estimate a logistic regression model using the variable campaign as a predictor of subscribed=yes. Are more campaign contacts effective at producing subscriptions to term deposits?	9
e. What is the probability of a person with zero contacts (i.e. campaign=0) subscribing to a term deposit? What is the probability of a person with one contacts (i.e. campaign=1) subscribing to a term deposit? What is the probability of a person with two contacts (i.e. campaign=2) subscribing to a term deposit?	11
f. Estimate a logistic regression model using the variable campaign as a predictor along with the following other variables: job, marital, education, default, housing, loan, contact, age, and campaign. Consider the form in which age should enter the model (i.e. categorical, continuous, transformed) and choose the best option for this model. Are more campaign contacts effective at producing subscriptions to term deposits conditional on the additional predictors?	12
g. For the model estimated in 1f, what is an interpretation of the coefficient for campaign?	17
h. Use the likelihood ratio test discussed in class to evaluate whether the model in 1f is a better fit than the model in 1d.	18

Question 1

Data

```
deposit <- read.csv('deposit-1.csv')
kable(head(deposit, 5), format = 'latex',
        booktabs = TRUE,
        caption = 'Deposit Data Snippet') %>%
kable_styling(latex_options = c("scale_down", "hold_position"))
```

Table 1: Deposit Data Snippet

ID	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	subscribed
26110	56	admin.	married	unknown	no	1933	no	no	telephone	19	nov	44	2	-1	0	unknown	no
40576	31	unknown	married	secondary	no	3	no	no	cellular	20	jul	91	2	-1	0	unknown	no
15320	27	services	married	secondary	no	891	yes	no	cellular	18	jul	240	1	-1	0	unknown	no
43962	57	management	divorced	tertiary	no	3287	no	no	cellular	22	jun	867	1	84	3	success	yes
29842	31	technician	married	secondary	no	119	yes	no	cellular	4	feb	380	1	-1	0	unknown	no

```
print(dim(deposit))
```

```
[1] 31647    18
```

```
deposit <- deposit[complete.cases(deposit), ]
```

a. Create an empirical logit plot with the response variable (y) subscribed by the predictor variable (x) age. You may want to convert subscribed to a numeric variable first.

```
kable(table(deposit$subscribed), format = 'latex')
```

Var1	Freq
no	27932
yes	3715

```
deposit$subscribed_numeric <- ifelse(deposit$subscribed == "yes", 1, 0)

# Calculating empirical logits by age
emp_logits <- deposit %>%
  group_by(age) %>%
  summarize(
```

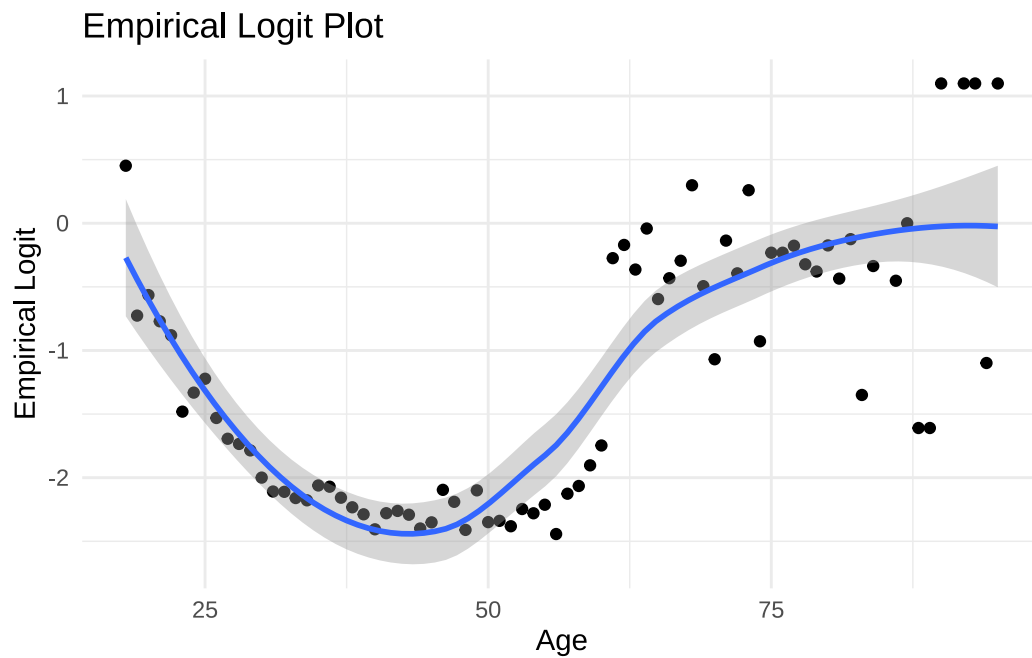
```

n = n(),
successes = sum(subscribed_numeric),
p = successes / n,
emp_logit = log((p + 1/(2*n)) / (1 - p + 1/(2*n)))
)

# Creating the empirical logit plot
ggplot(emp_logits, aes(x = age, y = emp_logit)) +
  geom_point() +
  geom_smooth(method = "loess", se = TRUE) +
  labs(
    title = "Empirical Logit Plot",
    x = "Age",
    y = "Empirical Logit"
  ) +
  theme_minimal()

```

`geom_smooth()` using formula = 'y ~ x'



b. Report the proportion (tabular format is fine) subscribed="yes" for each of the categories for job, marital, education, default, housing, loan, and contact.

```
calculate_proportions <- function(data, var_name) {
  props <- data %>%
    group_by(!!sym(var_name)) %>%
    summarize(
      total = n(),
      subscribed_yes = sum(subscribed == "yes"),
      proportion = round(subscribed_yes / total, 4)
    ) %>%
    arrange(desc(proportion)) %>%
    mutate(variable = var_name)

  return(props)
}

cat_vars <- c("job", "marital", "education", "default", "housing", "loan", "contact")

proportion_tables <- lapply(cat_vars, function(var) {
  calculate_proportions(deposit, var)
})

for(i in seq_along(cat_vars)) {
  cat("\n### Proportions for", cat_vars[i], "\n")
  print(
    kable(
      proportion_tables[[i]],
      col.names = c("Category", "Total", "Subscribed Yes", "Proportion", "Variable"),
      caption = paste("Subscription proportions by", cat_vars[i]),
      digits = 4
    )
  )
  cat("\n")
}
```

Proportions for job

Table: Subscription proportions by job

Category	Total	Subscribed Yes	Proportion	Variable
----------	-------	----------------	------------	----------

:-----	-----:	-----:	-----:	:-----
student	635	182	0.2866	job
retired	1574	362	0.2300	job
unemployed	905	129	0.1425	job
management	6639	923	0.1390	job
unknown	206	26	0.1262	job
self-employed	1123	140	0.1247	job
admin.	3631	452	0.1245	job
technician	5307	594	0.1119	job
housemaid	874	79	0.0904	job
services	2903	254	0.0875	job
entrepreneur	1008	85	0.0843	job
blue-collar	6842	489	0.0715	job

Proportions for marital

Table: Subscription proportions by marital

Category	Total	Subscribed Yes	Proportion	Variable
:-----	-----:	-----:	-----:	:-----
single	8922	1351	0.1514	marital
divorced	3630	445	0.1226	marital
married	19095	1919	0.1005	marital

Proportions for education

Table: Subscription proportions by education

Category	Total	Subscribed Yes	Proportion	Variable
:-----	-----:	-----:	-----:	:-----
tertiary	9301	1415	0.1521	education
unknown	1314	176	0.1339	education
secondary	16224	1697	0.1046	education
primary	4808	427	0.0888	education

Proportions for default

Table: Subscription proportions by default

Category	Total	Subscribed Yes	Proportion	Variable
:-----	-----:	-----:	-----:	:-----
no	31062	3674	0.1183	default
yes	585	41	0.0701	default

Proportions for housing

Table: Subscription proportions by housing

Category	Total	Subscribed Yes	Proportion	Variable
:-----	-----:	-----:	-----:	:-----
no	14063	2365	0.1682	housing
yes	17584	1350	0.0768	housing

Proportions for loan

Table: Subscription proportions by loan

Category	Total	Subscribed Yes	Proportion	Variable
:-----	-----:	-----:	-----:	:-----
no	26516	3384	0.1276	loan
yes	5131	331	0.0645	loan

Proportions for contact

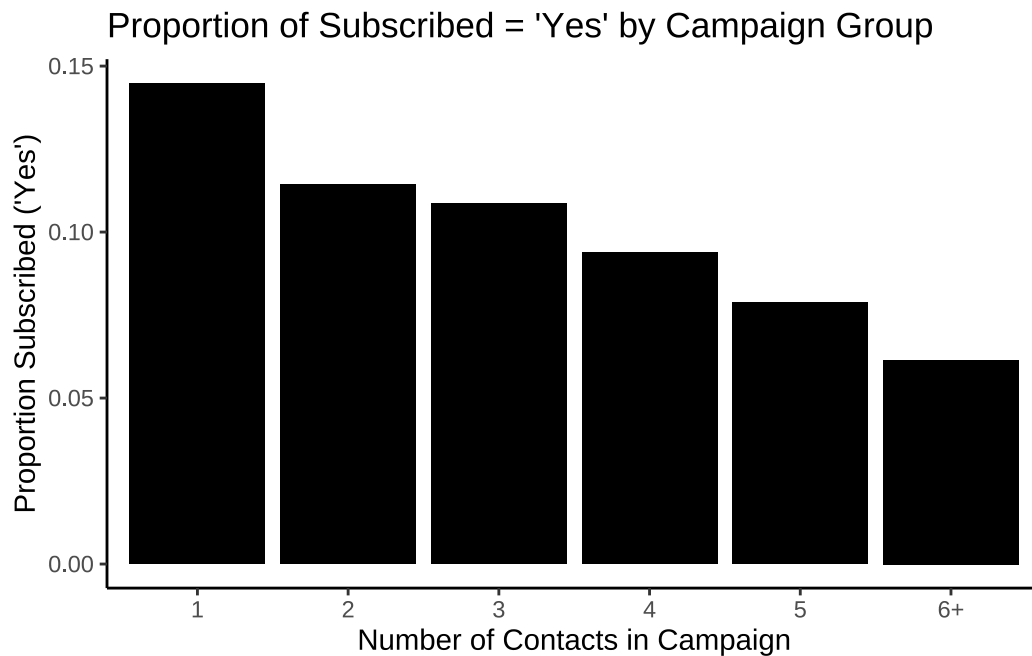
Table: Subscription proportions by contact

Category	Total	Subscribed Yes	Proportion	Variable
:-----	-----:	-----:	-----:	:-----
cellular	20423	3071	0.1504	contact
telephone	2047	268	0.1309	contact
unknown	9177	376	0.0410	contact

c. Plot the response variable, proportion subscribed="yes", for each of the following values of campaign: 1, 2, 3, 4, 5, 6+

```
campaign_summary <- deposit %>%
  mutate(campaign_group = if_else(campaign >= 6, "6+", as.character(campaign))) %>%
  group_by(campaign_group) %>%
  summarise(
    total = n(),
    subscribed_yes = sum(subscribed == "yes"),
    proportion_yes = subscribed_yes / total,
    .groups = "drop"
  )

ggplot(campaign_summary, aes(x = campaign_group, y = proportion_yes)) +
  geom_bar(stat = "identity", fill = "black") +
  labs(title = "Proportion of Subscribed = 'Yes' by Campaign Group",
       x = "Number of Contacts in Campaign",
       y = "Proportion Subscribed ('Yes')") +
  theme_classic()
```



1. The proportion of clients subscribing declines as the number of campaign contacts increases. Specifically:

- The highest subscription rate occurs with just one contact.
 - As the number of contacts rises, the subscription rate steadily decreases, reaching its lowest point for six or more contacts.
2. The diminishing returns from additional contacts suggest that after a certain threshold, making more calls is unlikely to improve conversions and may even have a negative effect.
 3. A more effective approach could be to focus on fewer, well-targeted contacts to maximize subscriptions while keeping costs low.

d. Next, we want to evaluate if campaign contacts are effective. Estimate a logistic regression model using the variable `campaign` as a predictor of `subscribed=yes`. Are more campaign contacts effective at producing subscriptions to term deposits?

```
model2 <- deposit %>% glm(subscribed_numeric~campaign,
                           family = binomial, data = .)
summary(model2)
```

Call:

```
glm(formula = subscribed_numeric ~ campaign, family = binomial,
     data = .)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.718972	0.027482	-62.55	<2e-16 ***
campaign	-0.122504	0.009629	-12.72	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 22893 on 31646 degrees of freedom
 Residual deviance: 22668 on 31645 degrees of freedom
 AIC: 22672

Number of Fisher Scoring iterations: 5

```
print(exp(coef(model2)))
```

(Intercept)	campaign
0.1792504	0.8847028

From your logistic regression model:

- **Intercept:** The estimate for the intercept is -1.718972 , which represents the log-odds of subscribing when `campaign = 0`. Since this is negative, it suggests that without any contacts, the odds of subscription are quite low.
- **Campaign Coefficient:** The estimate for `campaign` is -0.122504 . This is the change in the log-odds of subscribing for each additional contact. A **negative coefficient** means that as the number of campaign contacts increases, the **likelihood of subscribing decreases**.

Statistical Significance

The p-value for the `campaign` variable is $< 2e-16$, which is highly significant. This indicates that `campaign` is a statistically significant predictor of subscription (`subscribed = "yes"`), i.e., the relationship between campaign contacts and subscription is unlikely to be due to random chance.

For the `campaign` coefficient:

- The **odds ratio** is $\exp(-0.122504) = 0.884$ which suggests that for each additional contact, the odds of a person subscribing decrease by about **11.6%** (since the odds ratio is less than 1).

Conclusion

- Based on the logistic regression model, **more campaign contacts are associated with a decrease in the likelihood of subscription**. The negative coefficient and odds ratio below 1 indicate that as the number of contacts increases, the probability of subscribing to the term deposit **decreases**.
- Despite this, since the relationship is statistically significant, it suggests that campaign contacts are an important factor in the model, even though more contacts might not be effective at increasing subscriptions.

e. What is the probability of a person with zero contacts (i.e. $\text{campaign}=0$) subscribing to a term deposit? What is the probability of a person with one contacts (i.e. $\text{campaign}=1$) subscribing to a term deposit? What is the probability of a person with two contacts (i.e. $\text{campaign}=2$) subscribing to a term deposit?

```
# Coefficients from the model
intercept <- -1.718972
campaign_coef <- -0.122504

# Function to calculate probability based on campaign value
calculate_prob <- function(campaign) {
  log_odds <- -(intercept + campaign_coef * campaign)
  prob <- 1 / (1 + exp(log_odds))
  return(prob)
}

# Probabilities for campaign = 0, 1, and 2
prob_0 <- calculate_prob(0)
prob_1 <- calculate_prob(1)
prob_2 <- calculate_prob(2)

cat('P(Campaign = 0):\t', prob_0,
    '\nP(Campaign = 1):\t', prob_1,
    '\nP(Campaign = 2):\t', prob_2)
```

```
P(Campaign = 0):      0.1520036
P(Campaign = 1):      0.1368768
P(Campaign = 2):      0.123037
```

Calculation

$$\begin{aligned} P(\text{Campaign} = 0) &= \frac{1}{1 + e^{-(-1.719 + 0.123 \times \text{campaign})}} = \frac{1}{1 + e^{-(-1.719 + 0.123 \times 0)}} = 0.152 \\ P(\text{Campaign} = 1) &= \frac{1}{1 + e^{-(-1.719 + 0.123 \times \text{campaign})}} = \frac{1}{1 + e^{-(-1.719 + 0.123 \times 1)}} = 0.137 \\ P(\text{Campaign} = 2) &= \frac{1}{1 + e^{-(-1.719 + 0.123 \times \text{campaign})}} = \frac{1}{1 + e^{-(-1.719 + 0.123 \times 2)}} = 0.123 \end{aligned}$$

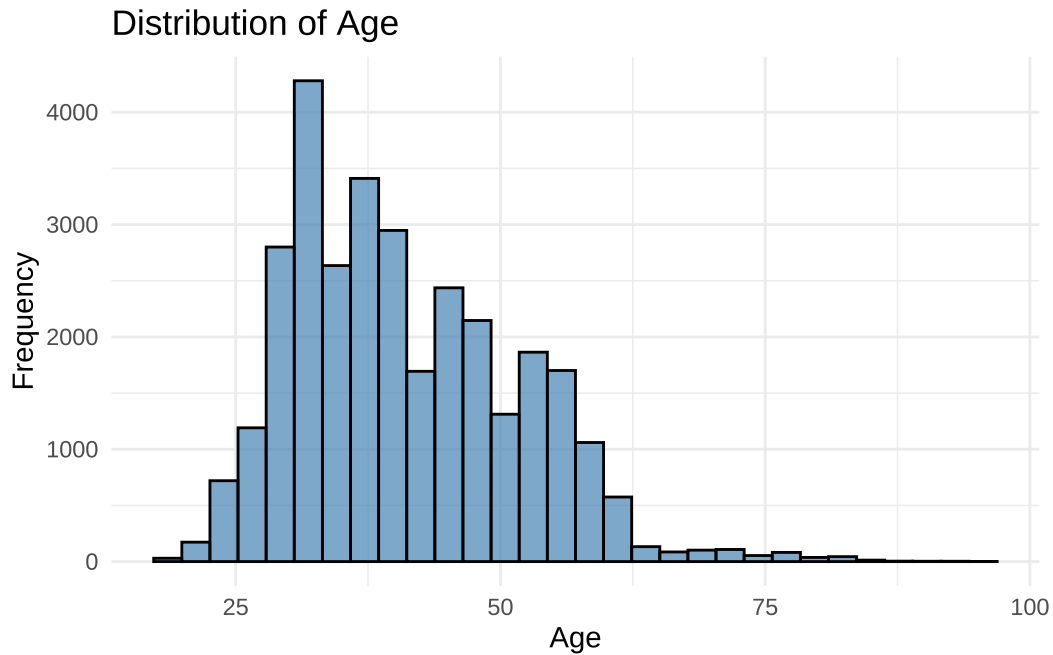
Using the logistic regression model, I estimated the probabilities of a person subscribing to a term deposit based on the number of campaign contacts.

- With zero contacts, the probability of subscribing is about **15.2%**.
- If contacted once, the probability drops to approximately **13.7%**.
- With two contacts, it decreases further to around **12.3%**.

These findings reinforce the earlier observation that increasing the number of contacts tends to lower the likelihood of subscription. This suggests that the **first contact has the greatest impact**, while additional calls may not only be ineffective but could even discourage potential subscribers. This insight highlights the importance of **optimizing the initial contact** rather than relying on multiple follow-ups to drive conversions.

f. Estimate a logistic regression model using the variable campaign as a predictor along with the following other variables: job, marital, education, default, housing, loan, contact, age, and campaign. Consider the form in which age should enter the model (i.e. categorical, continuous, transformed) and choose the best option for this model. Are more campaign contacts effective at producing subscriptions to term deposits conditional on the additional predictors?

```
ggplot(deposit, aes(x = age)) +
  geom_histogram(bins = 30, fill = "steelblue", color = "black", alpha = 0.7) +
  labs(title = "Distribution of Age", x = "Age", y = "Frequency") +
  theme_minimal()
```



```
deposit <- deposit %>%
  mutate(
    job = as.factor(job),
    marital = as.factor(marital),
    education = as.factor(education),
    default = as.factor(default),
    housing = as.factor(housing),
    loan = as.factor(loan),
    contact = as.factor(contact)
  )

model_full <- glm(subscribed_numeric ~ campaign + job + marital + education + default + ho
                  data = deposit,
                  family = binomial)

summary(model_full)
```

Call:

```
glm(formula = subscribed_numeric ~ campaign + job + marital +
    education + default + housing + loan + contact + age, family = binomial,
    data = deposit)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-1.302773	0.144283	-9.029	< 2e-16	***
campaign	-0.122479	0.009929	-12.336	< 2e-16	***
jobblue-collar	-0.308171	0.073844	-4.173	3.00e-05	***
jobentrepreneur	-0.425678	0.129017	-3.299	0.000969	***
jobhousemaid	-0.452775	0.134308	-3.371	0.000749	***
jobmanagement	-0.203276	0.073137	-2.779	0.005446	**
jobretired	0.441899	0.094973	4.653	3.27e-06	***
jobself-employed	-0.226625	0.109192	-2.075	0.037942	*
jobservices	-0.256576	0.084861	-3.023	0.002499	**
jobstudent	0.444423	0.110299	4.029	5.59e-05	***
jobtechnician	-0.241165	0.068935	-3.498	0.000468	***
jobunemployed	-0.108450	0.111729	-0.971	0.331723	
jobunknown	-0.190315	0.225953	-0.842	0.399633	
maritalmarried	-0.173769	0.058317	-2.980	0.002885	**
maritalsingle	0.200246	0.066583	3.007	0.002634	**
educationsecondary	0.098794	0.064555	1.530	0.125921	
educationtertiary	0.353441	0.074361	4.753	2.00e-06	***
educationunknown	0.192037	0.104020	1.846	0.064869	.
defaultyes	-0.315280	0.166802	-1.890	0.058738	.
housingyes	-0.610654	0.039210	-15.574	< 2e-16	***
loanyes	-0.658392	0.061371	-10.728	< 2e-16	***
contacttelephone	-0.243843	0.072658	-3.356	0.000791	***
contactunknown	-1.213220	0.057666	-21.039	< 2e-16	***
age	0.005484	0.002199	2.494	0.012616	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 22893 on 31646 degrees of freedom
 Residual deviance: 20881 on 31623 degrees of freedom
 AIC: 20929

Number of Fisher Scoring iterations: 6

```
library(car)
vif(model_full)
```

GVIF Df GVIF^(1/(2*Df))

campaign	1.013480	1	1.006717
job	3.787707	11	1.062404
marital	1.429676	2	1.093476
education	2.279160	3	1.147174
default	1.008157	1	1.004070
housing	1.124431	1	1.060392
loan	1.020098	1	1.009999
contact	1.103343	2	1.024891
age	2.107999	1	1.451895

We are considering age to be continuous from the histogram

Model Coefficients Interpretation:

- **Campaign (campaign):**
 - The coefficient for **campaign** is **-0.122479** with a very significant p-value ($< 2e-16$). This suggests that for each additional campaign contact, the log-odds of subscribing decrease by 0.1225, holding other factors constant.
 - This indicates a **negative** relationship between the number of contacts and the probability of subscribing.
- **Job:**
 - Some job categories, such as **jobblue-collar**, **jobentrepreneur**, and **jobmanagement**, have **negative coefficients**, indicating that individuals in these jobs are less likely to subscribe compared to the reference category.
 - Conversely, **jobretired**, **jobstudent**, and **jobtechnician** have **positive coefficients**, meaning these job types are more likely to subscribe to a term deposit.
- **Marital:**
 - **maritalmarried** has a negative coefficient, meaning married individuals are less likely to subscribe, while **maritalsingle** has a positive coefficient, indicating single individuals are more likely to subscribe.
- **Education:**
 - **educationtertiary** has a positive coefficient, suggesting individuals with tertiary education are more likely to subscribe compared to those with other education levels.
 - **educationsecondary** does not show a statistically significant effect.

- **Default:**
 - `defaultyes` has a **negative coefficient**, suggesting that individuals with a history of default are less likely to subscribe. However, it is only marginally significant (p-value = 0.0587).
- **Housing and Loan:**
 - Both `housingyes` and `loanyes` have **negative coefficients**, indicating that individuals who already have a housing loan or other loan are less likely to subscribe.
- **Contact:**
 - `contacttelephone` and `contactunknown` show a negative relationship with subscription likelihood, especially `contactunknown` with a large negative coefficient, suggesting that individuals who were contacted via unknown channels have much lower subscription probabilities.
- **Age:**
 - `age` has a positive coefficient (**0.005484**) and is statistically significant (p-value = 0.0126). This means that older individuals are slightly more likely to subscribe, holding other factors constant.

Model Performance:

- **Null deviance:** 22,893 (this is the deviance of the null model, i.e., a model with no predictors).
- **Residual deviance:** 20,881 (after including all predictors).
- The reduction in deviance suggests the model explains the data better than a null model.
- **AIC:** 20,929 (lower AIC values indicate a better model fit).

Multicollinearity Check:

- **GVIF** values are provided for each predictor. A GVIF (Generalized Variance Inflation Factor) greater than 1 suggests possible multicollinearity.
 - For example, `job` has a GVIF of 3.79, which indicates that multicollinearity might be a concern for this variable, as it has multiple categories.

Conclusion:

- **Effectiveness of Campaign:** The coefficient for `campaign` is **negative and significant**, meaning that increasing the number of campaign contacts is associated with a **decrease** in the probability of subscribing to a term deposit when controlling for other variables. This suggests that more contacts might not be effective in generating subscriptions, and there could be diminishing returns after a certain number of contacts.

g. For the model estimated in 1f, what is an interpretation of the coefficient for `campaign`?

Log-Odds Interpretation:

In a logistic regression model, the coefficients represent the change in the **log-odds** of the outcome (in this case, the probability of subscribing to a term deposit) for a one-unit change in the predictor variable, holding all other variables constant.

For the `campaign` variable, the coefficient is negative (-0.122479), which means that for each additional contact (i.e., as the number of campaign contacts increases by 1), the **log-odds of subscribing to a term deposit decrease by 0.1225**, holding all other variables in the model constant.

Odds Ratio Interpretation:

To better understand this coefficient in terms of probabilities, we can compute the **odds ratio**, which is the exponentiation of the coefficient:

$$\text{Odds Ratio} = e^{-0.122479} \approx 0.885$$

This means that for each additional contact in the campaign, the **odds of subscribing to a term deposit decrease by approximately 12%**, holding all other factors constant.

Practical Interpretation:

- If a person receives 1 more contact from the campaign, the odds of them subscribing to a term deposit are approximately **12% lower** compared to a person with one less contact, assuming other variables (like job, marital status, age, etc.) are held constant.
- Since the coefficient for `campaign` is negative, it suggests that **more campaign contacts** might actually **decrease** the likelihood of subscription, possibly because the additional contacts could be perceived as excessive or annoying by some individuals.

Summary:

- **Effect of campaign:** The more campaign contacts a person receives, the less likely they are to subscribe to a term deposit, after adjusting for other factors such as job, marital status, education, and so on.
- **Magnitude of effect:** The odds decrease by about 12% for each additional campaign contact.

h. Use the likelihood ratio test discussed in class to evaluate whether the model in 1f is a better fit than the model in 1d.

```
model_d <- glm(subscribed_numeric ~ campaign, data = deposit, family = binomial)

model_f <- glm(subscribed_numeric ~ campaign + job + marital + education + default + housi
              data = deposit,
              family = binomial)
lrt <- anova(model_d, model_f, test = "LRT")
print(lrt)
```

Analysis of Deviance Table

Model 1: subscribed_numeric ~ campaign

Model 2: subscribed_numeric ~ campaign + job + marital + education + default +
housing + loan + contact + age

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	31645	22668			
2	31623	20881	22	1786.7	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Key Components:

1. Residual Degrees of Freedom (Resid. Df)

- Model 1: 31,645
- Model 2: 31,623
- **Difference (Df) = 22** (indicating that Model 2 has 22 additional parameters)

2. Residual Deviance (Resid. Dev)

- Model 1: 22,668
- Model 2: 20,881
- **Deviance Reduction (Deviance)** = 1,786.7

3. Chi-Square Test ($\text{Pr}(>\text{Chi})$)

- The p-value is $< \mathbf{2.2e-16}$, which is extremely small.
- This suggests that the additional variables in Model 2 **significantly improve the fit** of the model compared to Model 1.

Conclusion:

Since the p-value is **very small** (< 0.001), we **reject the null hypothesis** that the simpler model (Model 1) is as good as the more complex model (Model 2). This means that adding **job, marital, education, default, housing, loan, contact, and age** significantly improves the model's ability to predict whether a person subscribes to a term deposit. Hence **Model 2** is better