Exam 03

STA 198 Spring 2020 (Jiang)

Monday, July 27, 2020

```
library(class)
library(tidyverse)
library(broom)
```

Q2 T: 2.1, 2.2, 2.5, 2.6, 2.7, 2.9 F: 2.3, 2.4, 2.8, 2.10, 2.11, 2.12

Q3.1 The Kaplan-Meier plot needs to be a non-decreasing step function, but the plot depicted was increasing and was not a step function. As well, all survival estimates need to start at 1 at time 0, and cannot be negative (as it is an estimated probability). However, the plots depicted were both greater than 1 and less than 0. The y-axis should be estimated survival probability, not estimated event probabilty (a correct axis would have been "Estimated probability of remaining recurrence-free"), and finally, the title is not supported by the data in the graphs.

Q3.2 The placebo group has an estimated median survival of 375 days compared to the chemotherapy group of 500 days.

```
wine <- read.csv("wine.csv")

m4.1 <- lm(quality ~ citric_acid + sugar + chlorides + so2 + pH +
             alcohol + as.factor(region), data = wine)

tidy(m4.1)
```

Q4.1

```
## # A tibble: 9 x 5
##   term                estimate std.error statistic     p.value
##   <chr>                  <dbl>     <dbl>     <dbl>       <dbl>
## 1 (Intercept)         1.00       1.35       0.742  0.458
## 2 citric_acid         0.0386     0.464      0.0832 0.934
## 3 sugar               0.0231     0.0123     1.88   0.0607
## 4 chlorides          -0.211      2.79      -0.0757 0.940
## 5 so2                 0.000922   0.00148    0.624  0.533
## 6 pH                  0.495      0.362      1.37   0.172
## 7 alcohol             0.275      0.0532     5.17   0.000000277
## 8 as.factor(region)2  0.159      0.133      1.20   0.231
## 9 as.factor(region)3  0.215      0.124      1.73   0.0842
```

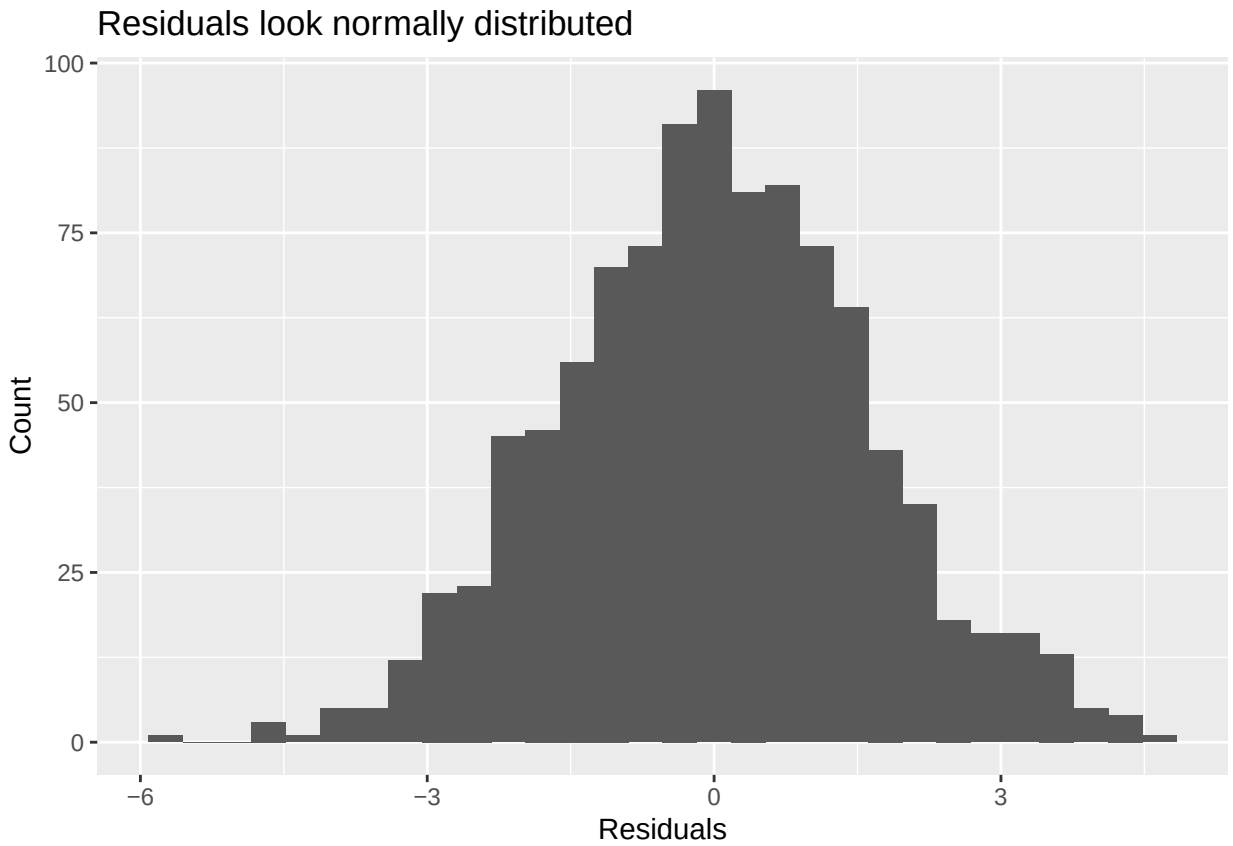
From the model output, we see that alcohol content is the only predictor that has a statistically significant linear association with quality score, conditionally on the other predictors (higher alcohol content being associated with higher scores).

That being said, there are also directional relationships with the other variables that may still be of interest to producers, regardless of statistical significance. In particular, higher citric acid, sugar, sulfur dioxide levels, and pH are also associated with higher score (holding all else constant), as is lower salt content. Finally, region 3 is associated with higher scores as well – producers might consider picking up a plot of land there!

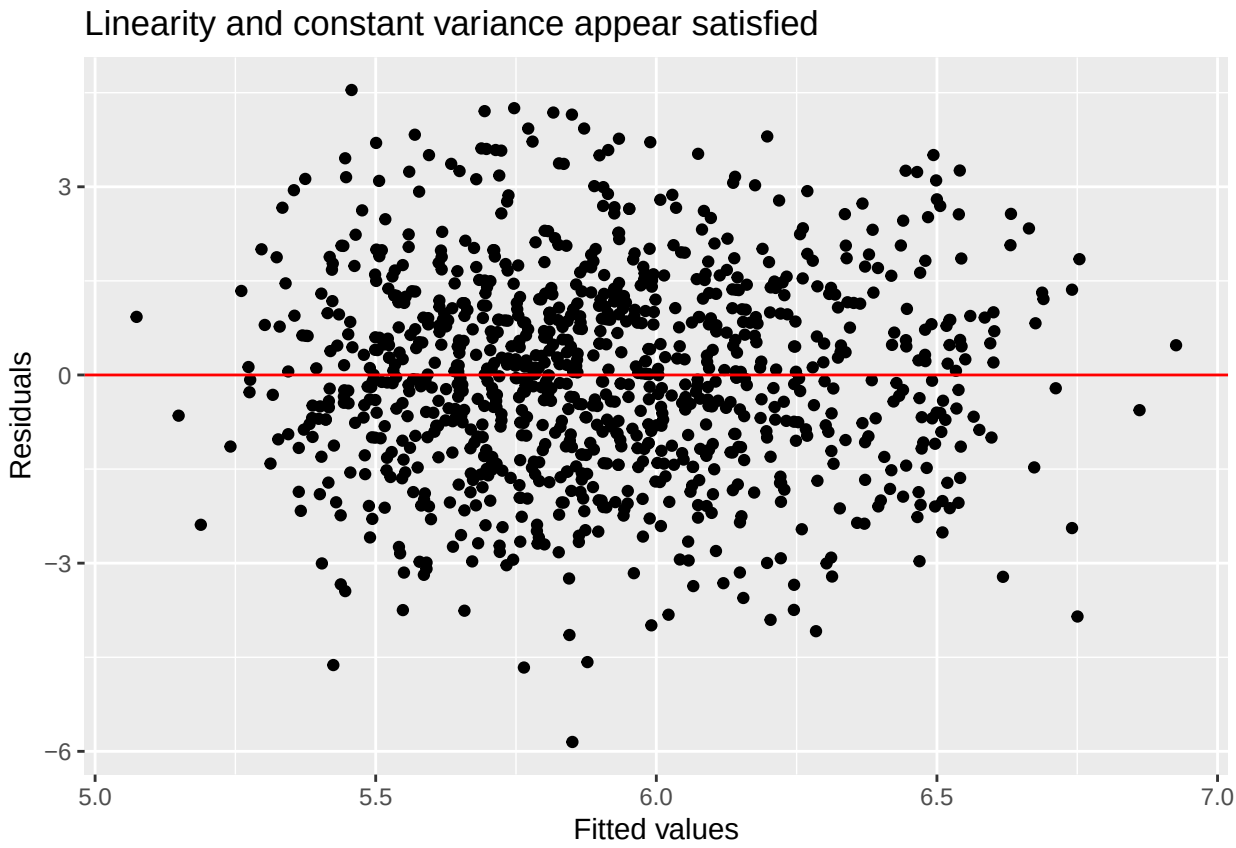
```
m4.1_res <- augment(m4.1)
ggplot(data = m4.1_res, aes(x = .resid)) +
  geom_histogram() +
  labs(x = "Residuals", y = "Count",
       title = "Residuals look normally distributed")
```

Q4.2

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(data = m4.1_res, aes(x = .fitted, y = .resid)) +
  geom_point() +
  labs(x = "Fitted values", y = "Residuals",
       title = "Linearity and constant variance appear satisfied") +
  geom_hline(yintercept = 0, color = "red")
```



The residuals look approximately normally distributed, with constant variance for all fitted values. Linearity also appears satisfied due to the symmetry about the x-axis. As for independence, we must assume that each sample was taken independently from a separate batch of wines and that each observation does not affect any others. This is a reasonable assumption to make, though we do not know exactly how these samples were collected.

Q4.3 A wine grown in region 1 with zero citric acid, sugar, salt, sulfur dioxide, or alcohol content, and with a pH of 0 is expected to have a quality score of 1. For each additional percentage point increase of alcohol by volume, we would expect the quality score to increase by 0.275, holding all other variables constant.

Q4.4 Yes. We are testing the null hypothesis that

$$\beta_{sugar} = 0$$

against the alternative that

$$\beta_{sugar} \neq 0$$

. The test statistic is 1.878, which has a t_{991} distribution under the null hypothesis. This test statistic corresponds to a p-value of 0.0607. At the $\alpha = 0.05$ level, we fail to reject the null hypothesis. There is insufficient evidence to suggest that there is a linear association between sugar content and wine quality score conditionally on the other predictors in our model.

```
est <- tidy(m4.1) %>%
  slice(1) %>%
  pull(estimate)
```

```r
se <- tidy(m4.1) %>%
  slice(1) %>%
  pull(std.error)

round(c(est - qt(0.975, 991)*se, est + qt(0.975, 991)*se), 3)
```

Q4.5

```
## [1] -1.647  3.652
```

This is not a useful confidence interval because we are not interested in inference on the intercept term – such a wine (a region 1 wine with 0 for all of the other predictors) simply does not exist.

```r
transplant <- read.csv("transplant.csv")

m5.1 <- glm(post_dm ~ eti + sex + bmi + meld + pre_dm + cold_isch + don_bmi,
            data = transplant, family = "binomial")

tidy(m5.1) %>%
  filter(term %in% c("etiother", "pre_dm", "cold_isch")) %>%
  mutate(exp_beta = exp(estimate)) %>%
  select(term, exp_beta)
```

Q5.1

```
## # A tibble: 3 x 2
##   term      exp_beta
##   <chr>        <dbl>
## 1 etiother     0.466
## 2 pre_dm       6.35
## 3 cold_isch    1.13
```

Patients who have non-cirrhosis transplant etiologies is expected to have approximately 0.466 times the odds of having PTDM within two years of transplant compared to patients with alcoholic cirrhosis transplant etiologies, holding all other predictors constant.

Patients who have diabetes prior to their transplant is expected to have approximately 6.348 times the odds of having PTDM within two years of transplant compared to patients who do not have diabetes prior to their transplant, holding all other predictors constant.

A patient who has a cold ischemia time one hour greater than another patient is expected to have approximately 1.130 times the odds of having PTDM within two years, holding all other predictors constant.

Q5.2 A predicted probability of 0.1 corresponds to a logit of $\log(0.1/0.9)$, which is approximately -2.197. Note that the mean cold ischemia time in our dataset is 6.86 hours.

```r
transplant %>%
  summarize(mean_ct = mean(cold_isch))
```

```
##    mean_ct
## 1 6.859851
```

Let's now plug in values into our model and solve for when the estimated logit is equal to -2.197. Note that this patient has the baseline values for etiology, sex, and prior diabetes status.

$$
\begin{aligned}
-2.19722 &= -1.86674 - 0.00269 \times 26.5 - 0.01448 \times MELD + 0.12243 \times 6.86 + 0.03663 \times 26.5 \\
&= -0.12746 - 0.01448 \times MELD \\
2.06976 &= 0.01448 \times MELD
\end{aligned}
$$

Solving, we find that this border occurs around 142.94. Thus, MELD scores of more than this value would correspond to predicted probabilities of 10% or less.

Q5.3 Yes. We are testing the null hypothesis $\beta_{cold\_isch} = 0$ against the alternative hypothesis $\beta_{cold\_isch} \neq 0$. Our test statistic is 2.02, which under the null hypothesis has a standard normal distribution. This test statistic corresponds to a p-value of 0.0438, which is significant at the $\alpha = 0.05$ level. Thus, we reject the null hypothesis. There is sufficient evidence to conclude that cold ischemia time is associated with developing PTDM within two years of transplant, adjusting for all other predictors.

Q5.4 We fit a new model that includes an interaction between cold ischemia time and prior diabetes status:

```
m5.4 <- glm(post_dm ~ eti + sex + bmi + meld + pre_dm + cold_isch + don_bmi +
              cold_isch*don_bmi,
            data = transplant, family = "binomial")

tidy(m5.4)
```

```
## # A tibble: 10 x 5
##    term              estimate std.error statistic    p.value
##    <chr>                <dbl>     <dbl>     <dbl>      <dbl>
##  1 (Intercept)       -0.139      2.26     -0.0616 0.951
##  2 eticirr_non       -0.0255     0.335    -0.0761 0.939
##  3 etiother          -0.771      0.458    -1.68   0.0924
##  4 sexM               0.182      0.301     0.605  0.545
##  5 bmi               -0.00396    0.0243   -0.163  0.871
##  6 meld              -0.0139     0.0152   -0.915  0.360
##  7 pre_dm             1.87       0.395     4.74   0.00000215
##  8 cold_isch         -0.127      0.299    -0.424  0.672
##  9 don_bmi           -0.0315     0.0833   -0.378  0.705
## 10 cold_isch:don_bmi  0.00996    0.0118    0.844  0.398
```

We are testing the null hypothesis that the slope parameter corresponding to the interaction term is zero vs. the alternative hypothesis that it is not. Our test statistic is 0.844, which has a standard normal distribution under the null hypothesis. This corresponds to a p-value of 0.398, which is not significant at the $\alpha = 0.05$ level. Thus, we fail to reject the null hypothesis. There is insufficient evidence for us to suggest that the relationship between cold ischemia time and PTDM status depends on a patient's diabetes status prior to transplant.

```
test <- read_csv("new_patients.csv")
```

Q5.5

```
## Parsed with column specification:
## cols(
##   eti = col_character(),
```

```
##   sex = col_character(),
##   bmi = col_double(),
##   meld = col_double(),
##   pre_dm = col_double(),
##   cold_isch = col_double(),
##   don_bmi = col_double(),
##   post_dm = col_double()
## )
```

```
preds_5.1 <- augment(m5.1, newdata = test) %>%
  mutate(p = exp(.fitted)/(1 + exp(.fitted)),
         pred = ifelse(p > 0.5, 1, 0)) %>%
  pull(pred)
mean(preds_5.1 == test %>% pull(post_dm))
```

```
## [1] 0.8
```

The classification accuracy is 80%.

```
library(class)

train_var <- transplant %>%
  select(bmi, meld, cold_isch, don_bmi)
train_status <- transplant %>%
  pull(post_dm)

test_var <- test %>%
  select(bmi, meld, cold_isch, don_bmi)

preds_5.6 <- knn(train_var, test_var, train_status, k = 15)
mean(preds_5.6 == test %>% pull(post_dm))
```

Q5.6

```
## [1] 0.54
```

The classification accuracy is 54%.

Q5.7 The k-NN model only used the four continuous predictors. Importantly, it did not use the best predictor in the model, prior diabetes status. This is likely the reason for the relatively low classification accuracy (54%; barely doing better than random chance!) compared to the logistic regression model (80%).