Problem set 10

```r
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.3.6      v purrr   0.3.5
v tibble  3.1.8      v dplyr   1.0.10
v tidyr   1.2.1      v stringr 1.4.1
v readr   2.1.3      v forcats 0.5.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```r
library(Stat2Data)
library(skimr)
library(broom)
library(lmtest)
```

```
Loading required package: zoo

Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

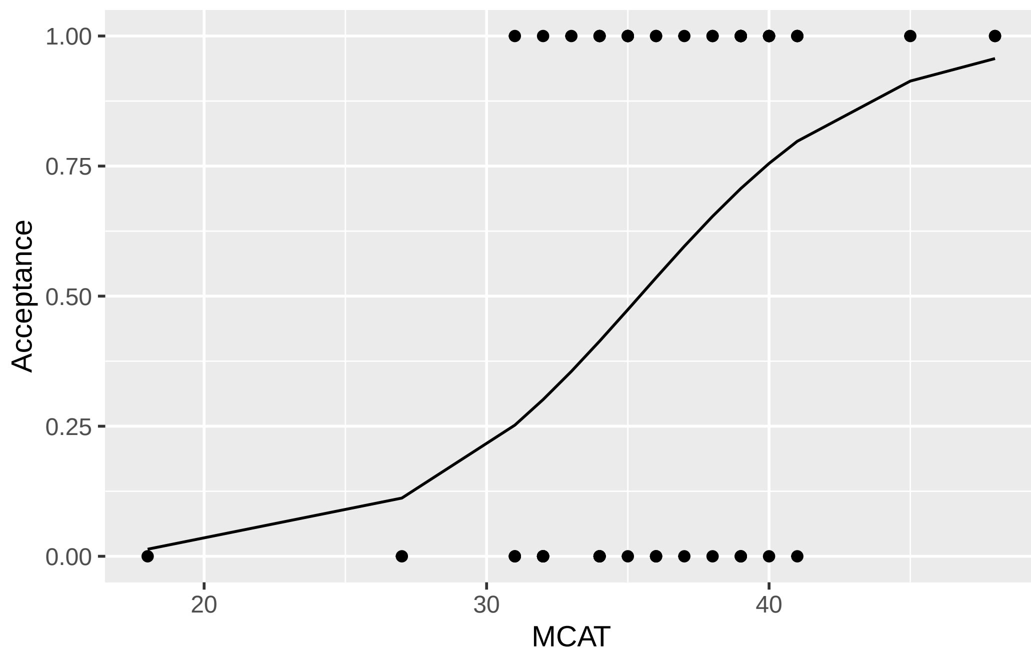
Exercises to hand in: 9.19, 9.40, question made up by Dr. M

9.19 Medical school acceptance

```r
data("MedGPA")
```

a. Model

```r
medgpal<-glm(Acceptance ~ MCAT, data = MedGPA, family = binomial)
medgpal2<-augment(medgpal, data= MedGPA)
medgpal2<- medgpal2 %>%
  mutate(
    odds = exp(.fitted),
    probability = odds/(1+odds))
ggplot(medgpal2, aes(x =MCAT))+ geom_point(aes(y=Acceptance))+geom_line(aes(y=probability)
```



b. Odds ratio

```r
exp(0.24596)
```

```
[1] 1.278848
```

c. Prediction

```r
pred = -8.71245 + 0.24596 *40
exp(pred)
```

```
[1] 3.083144
```

d. 50/50 point

```r
summary(medgpal)
```

```
Call:
glm(formula = Acceptance ~ MCAT, family = binomial, data = MedGPA)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7878  -1.0330   0.4256   0.9225   1.6601

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.71245    3.23645  -2.692  0.00710 **
MCAT         0.24596    0.08938   2.752  0.00592 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 75.791  on 54  degrees of freedom
Residual deviance: 64.697  on 53  degrees of freedom
AIC: 68.697

Number of Fisher Scoring iterations: 4
```

9.40 Levee failures

(No dataset)

a. State the null hypothesis that the P-value 0.046 allows you to test.

$H_0 : \beta_1 = 0$

$H_A : \beta_1 \neq 0$

According to the null hypothesis, we don't have a linear relationship between levee failure and constriction of the flood way over time. But since the p-value is less than 0.05, we can reject the null hypothesis. Therefore, there is a linear relationship between levee failure and constriction of the flood way over time.

b. What happens to the probability of a levee failure as the constriction factor gets larger? Explain

```
0.571 - 0.691 * 1
```

```
[1] -0.12
```

```
0.571 - 0.691 * 2
```

```
[1] -0.811
```

```
0.571 - 0.691 * 3
```

```
[1] -1.502
```

As we can see, as the constriction factor gets larger, the failure decreases.

As the constriction of the flood way increases, the chances of levee failure become lesser since constriction factor has a negative coefficient so the more it gets, the lesser the response variable i.e. failure would get.

c. Find a 95% confidence interval for the slope parameter in the logistic model.

```
-0.691+1.96*0.346
```

```
[1] -0.01284
```

```
-0.691-1.96*0.346
```

```
[1] -1.36916
```

The 95% confidence interval for the slope parameter is between -0.01284 and -1.36916

Problem made up by Dr. M: Empirical logit of levee failures

Okay, in the last problem I said there was no data because I wanted you to work from the table in the book, but the dataset does exist.

```
data("LeveeFailures")
```

In this problem, we are concerned with the same issue as before– trying to predict if a levee will fail or not, based on the constriction factor of the floodway. If you want to read more about the data, use the `?` operator,

```
?LeveeFailures
```

a. Reproduce the logistic regression model from the previous problem, using R.

```
logisiticModel <- glm(Failure ~ ConstrictionFactor, data = LeveeFailures, family = binomia
summary(logisiticModel)
```

```
Call:
glm(formula = Failure ~ ConstrictionFactor, family = binomial,
    data = LeveeFailures)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4095  -1.1730   0.2682   1.1358   1.7781

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.5708     0.3496   1.633   0.1025
ConstrictionFactor  -0.6906     0.3457  -1.998   0.0457 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 113.68  on 81  degrees of freedom
Residual deviance: 108.46  on 80  degrees of freedom
AIC: 112.46

Number of Fisher Scoring iterations: 3
```

```r
exp(coef(logisiticModel))
```

```
      (Intercept) ConstrictionFactor 
        1.7697429          0.5012851 
```

b. Produce an empirical logit plot to check for linearity of the logit with respect to ConstrictionFactor. What conclusion do you reach about the appropriateness of logistic regression?

```r
LeveeFailures <- LeveeFailures %>%
  mutate(ConstrictionFactorGroup = cut(ConstrictionFactor, breaks = 10))
LeveeFailures %>%
  group_by(ConstrictionFactorGroup)
```

```
# A tibble: 82 x 15
# Groups:   ConstrictionFactorGroup [7]
   Failure  Year River~1 Sedim~2 Borro~3 Meander Chann~4 Flood~5 Const~6 LandC~7
     <int> <int>   <dbl>   <int>   <int>   <int>   <dbl>   <dbl>   <dbl>   <int>
 1       1  1890     847       0       0       4   1347    2026.   1           4
 2       1  1890     787       1       0       3   2581.   4123.   1           2
 3       1  1890     776       1       0       3   3379.   7999.   1           4
 4       1  1890     776       1       0       3   3507.   8538.   1           2
 5       1  1890     773       1       0       2   1704.   4174.   1           2
 6       1  1910     830       1       0       1   2823.   5207.   0.700       3
 7       1  1910     785       0       0       1   1707.   3316.   0.824       3
 8       1  1910     785       0       0       1   1741.   3149.   0.824       3
 9       1  1910     785       1       0       1   1713.   3097.   0.824       3
10       1  1910     784       1       0       1   1826.   3244.   0.890       4
# ... with 72 more rows, 5 more variables: VegWidth <dbl>, Sinuosity <dbl>,
#   Dredging <int>, Revetement <int>, ConstrictionFactorGroup <fct>, and
#   abbreviated variable names 1: RiverMile, 2: Sediments, 3: BorrowPit,
#   4: ChannelWidth, 5: FloodwayWidth, 6: ConstrictionFactor, 7: LandCover
```

```r
Levee_binned <- LeveeFailures %>%
  group_by(ConstrictionFactorGroup) %>%
  summarize(binnedFail = mean(Failure), binnedCF = mean(ConstrictionFactor)) %>%
  mutate(logit = log(binnedFail/(1-binnedFail)))
```

```r
logm1 <- augment(logisiticModel, data = LeveeFailures)
logm1 <- logm1 %>%
  mutate(odds = exp(.fitted),
         probability = odds / (1 + odds))
```

```r
ggplot(Levee_binned) +
  geom_point(aes(x = binnedCF, y = logit)) +
  geom_line(data = logm1, aes(x = ConstrictionFactor, y = .fitted))
```