Back to course

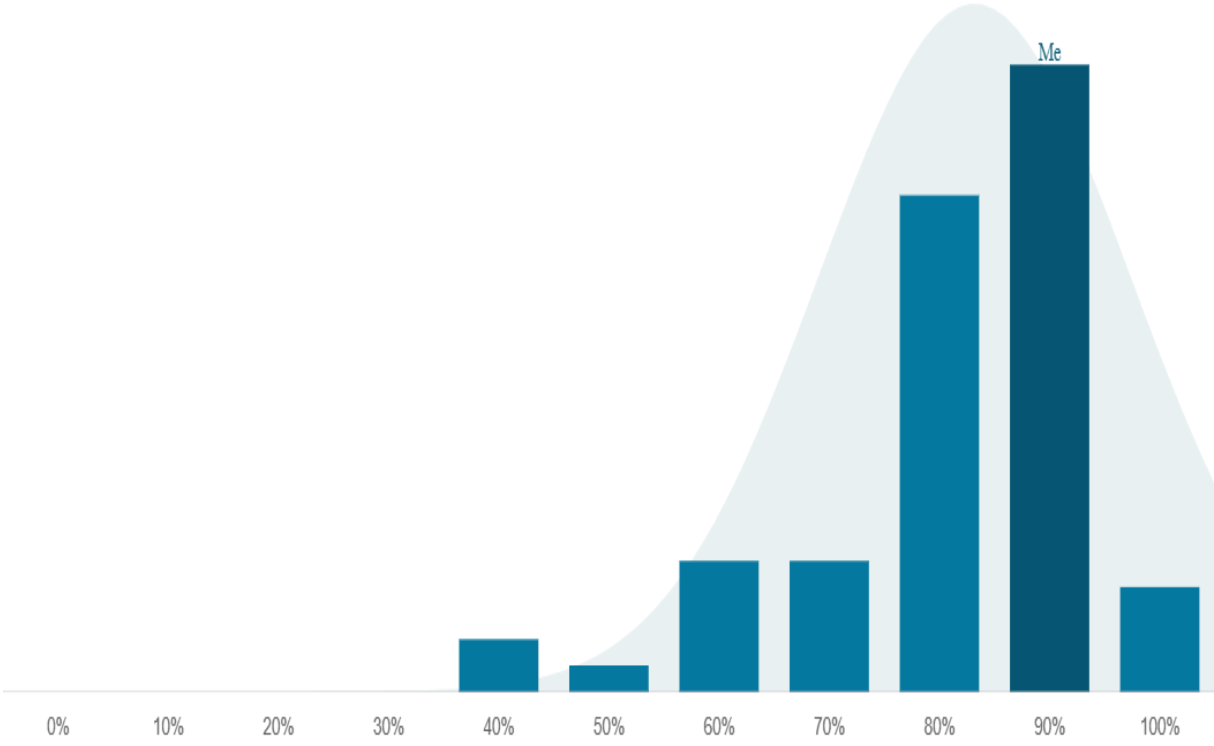
Assignment 5

Class scores distribution Hide

Percentages | Points

Total | Q1 | Q2 | Q3 | Q4

Students: 60 Mean: 80.9 Median: 83.1 Std. Dev: 13.3



My score

92.4% (39.75/43)

Q1

3 / 6

Q1

Q1. a) $\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'Y$

$$E[\hat{\beta}] = E[(X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + \varepsilon)]$$

$$= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + E(\varepsilon))$$

$E(\varepsilon) = 0$

$$= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2)$$

Therefore, the estimator $\hat{\beta}_1$ is biased.

-1 -1

b) If X_1 and X_2 are uncorrelated or orthogonal, $\hat{\beta}_1$ is unbiased because it can expressed as $X_1'X_2 = 0$

uncorrelated is not the right term. should be orthogonal

c) When $\beta_2 = 0$, X_2 would be 0. This means the predictors in X_2 have no effect on Y.

The estimate of σ^2 based on the reduced model is likely to be less than that based on the true model

-2 -2

CamScanner로 스캔하기

Q2

17 / 17

Q2

Q2 a)

Let Y_i be the log selling price, respectively, of the i^{th} diamond

Let $x_{1i} = 1$ if the i^{th} diamond is E colour and be 0 otherwise

Let $x_{2i} = 1$ if the i^{th} diamond is F colour and be 0 otherwise

Let $x_{3i} = 1$ if the i^{th} diamond is G colour and be 0 otherwise

Let $x_{4i} = 1$ if the i^{th} diamond is H colour and be 0 otherwise

Let $x_{5i} = 1$ if the i^{th} diamond is I colour and be 0 otherwise

Let $x_{6i} = 1$ if the i^{th} diamond is VS1 and be 0 otherwise

Let $x_{7i} = 1$ if the i^{th} diamond is VS2 and be 0 otherwise

Let $x_{8i} = 1$ if the i^{th} diamond is VVS1 and be 0 otherwise

Let $x_{9i} = 1$ if the i^{th} diamond is VVS2 and be 0 otherwise

Let x_{10i} be the weight, respectively, of the i^{th} diamond

$$Y_i = \beta_0 + \sum_{j=1}^{10} \beta_j x_{ji} + \beta_{11} x_{10i}^2 + \varepsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ and the ε_i's are independent

b) In general, this model

- Allows for a different intercept for each combination of levels
- Assumes the same effect of weight and weight^2 for each combination of levels
- Assumes that mean log selling price lies on a plane determined by weight and weight^2, with intercept and orientation (in part) determined by colour and clarity.

A5_Q2

2023-11-27

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.4.4     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
df=read.table("diamond-ext.txt",header=TRUE)

df = df %>% mutate(Colour=factor(Colour))

df = df %>% mutate(Clarity=factor(Clarity))

contrasts(df$Colour)
```

```
##   E F G H I
## D 0 0 0 0 0
## E 1 0 0 0 0
## F 0 1 0 0 0
## G 0 0 1 0 0
## H 0 0 0 1 0
## I 0 0 0 0 1
```

```
contrasts(df$Clarity)
```

```
##      VS1 VS2 VVS1 VVS2
## IF     0   0    0    0
## VS1    1   0    0    0
## VS2    0   1    0    0
## VVS1   0   0    1    0
## VVS2   0   0    0    1
```

```
contrasts(df$Colour)=contr.treatment(6)
contrasts(df$Clarity)=contr.treatment(5)

df = df %>% mutate(LogPrice=log(Price))

fit=lm(LogPrice~Colour+Clarity+Carat+I(Carat^2),df)
```

c.

```
summary(fit)
```

```
##
## Call:
## lm(formula = LogPrice ~ Colour + Clarity + Carat + I(Carat^2),
##     data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.164687 -0.037314 -0.000847  0.036397  0.123827
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.90902    0.02752 214.727  < 2e-16 ***
## Colour2     -0.04975    0.01908  -2.608  0.00973 **
## Colour3     -0.13150    0.01795  -7.325 4.21e-12 ***
## Colour4     -0.22777    0.01843 -12.357  < 2e-16 ***
## Colour5     -0.32852    0.01875 -17.520  < 2e-16 ***
## Colour6     -0.43036    0.02026 -21.244  < 2e-16 ***
## Clarity2    -0.24738    0.01234 -20.051  < 2e-16 ***
## Clarity3    -0.33934    0.01391 -24.396  < 2e-16 ***
## Clarity4    -0.10729    0.01323  -8.109 3.28e-14 ***
## Clarity5    -0.18138    0.01224 -14.820  < 2e-16 ***
## Carat        6.34001    0.10732  59.075  < 2e-16 ***
## I(Carat^2)  -2.76293    0.10121 -27.298  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05442 on 225 degrees of freedom
## Multiple R-squared:  0.9943, Adjusted R-squared:  0.994
## F-statistic:  3551 on 11 and 225 DF,  p-value: < 2.2e-16
```

```
head(model.matrix(fit))
```

```
##   (Intercept) Colour2 Colour3 Colour4 Colour5 Colour6 Clarity2 Clarity3
## 1           1       0       0       0       0       0        0        1
## 2           1       1       0       0       0       0        1        0
## 3           1       0       0       1       0       0        0        0
## 4           1       0       0       1       0       0        1        0
## 5           1       0       0       0       0       0        1        0
## 6           1       1       0       0       0       0        1        0
##   Clarity4 Clarity5 Carat I(Carat^2)
## 1        0        0  0.30     0.0900
## 2        0        0  0.30     0.0900
## 3        1        0  0.30     0.0900
## 4        0        0  0.30     0.0900
## 5        0        0  0.31     0.0961
## 6        0        0  0.31     0.0961
```

```
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: LogPrice
##             Df Sum Sq Mean Sq  F value    Pr(>F)
## Colour       5  2.181   0.436   147.29 < 2.2e-16 ***
## Clarity      4 15.384   3.846  1298.56 < 2.2e-16 ***
## Carat        1 95.907  95.907 32382.89 < 2.2e-16 ***
## I(Carat^2)   1  2.207   2.207   745.16 < 2.2e-16 ***
## Residuals  225  0.666   0.003
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit1=lm(LogPrice~+Clarity+Carat+I(Carat^2),df)

anova(fit1,fit)
```

```
## Analysis of Variance Table
##
## Model 1: LogPrice ~ +Clarity + Carat + I(Carat^2)
## Model 2: LogPrice ~ Colour + Clarity + Carat + I(Carat^2)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    230 3.9935
## 2    225 0.6664  5    3.3272 224.68 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

d. i. H0: $\beta1 = \beta2 = \beta3 = \beta4 = \beta5 = 0$ vs Ha: at least one of $\beta1,\beta2,\beta3,\beta4,\beta5$ not 0

ii. From output, F=224.68

iii. If H0 is true, F is a draw from an F5,225 distribution

iv. p-value < 0.001

v. Since p-value<0.05, we reject Ho. We have evidence that the overall effect of colour is significant at the 5% level.

```
pi=predict(fit,data.frame(Carat=0.5,Colour="F",Clarity="VS1"))
exp(pi)
```

```
##        1
## 3009.163
```

e. The predicted selling price of a 0.5-carat diamond of colour F and clarity VS1 is 3009.2

good job 17

Q3

10.75 / 11

Q3

Q3 a)

Let $Y_i = 1$ if the i^{th} patient has hyperglycemia and let $Y_i = 0$ otherwise,

Let $\pi_i = P(Y_i = 1) = E[Y_i]$

Let $x_{1i} = 1$ if the i^{th} patient has periodontitis and 0 otherwise

Let $x_{2i} = 1$ if the i^{th} patient has BMI greater than 23 kg/m^2 and 0 otherwise

Let $x_{3i} = 1$ if the i^{th} patient has a family history of diabetes and 0 otherwise

Let x_{4i} be the age of the i^{th} patient

correct 3

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i}$$

Where $Y_i \sim Binary(\pi_i)$ and the Y_i's are independent

b) Periodontitis is associated with a change of β_1 in the log-odds of hyperglycemia, holding the other predictors constant.

Correct 2

Q3

2023-11-29

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.4.4     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
df=read.table("dental.txt",header=TRUE)

fit=glm(Hyperglycemia~Periodontitis+highBMI+FamilyHistory+age, family=binomial(logit),df)
```

c.

```
summary(fit)
```

```
##
## Call:
## glm(formula = Hyperglycemia ~ Periodontitis + highBMI + FamilyHistory +
##     age, family = binomial(logit), data = df)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -5.09803    0.69116  -7.376 1.63e-13 ***
## Periodontitis   0.64757    0.26611   2.433    0.015 *
## highBMI       -13.55378  535.41118  -0.025    0.980
## FamilyHistory  14.88827  535.41121   0.028    0.978
## age             0.06510    0.01383   4.706 2.52e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 875.54  on 723  degrees of freedom
## Residual deviance: 741.62  on 719  degrees of freedom
## AIC: 751.62
##
## Number of Fisher Scoring iterations: 12
```

Correct 1

```
fit1=glm(Hyperglycemia~Periodontitis, family=binomial(logit),df)

anova(fit1,fit,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Hyperglycemia ~ Periodontitis
## Model 2: Hyperglycemia ~ Periodontitis + highBMI + FamilyHistory + age
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       722     870.45
## 2       719     741.62  3   128.83 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pchisq(128.83,2,lower.tail=FALSE)
```

```
## [1] 1.059061e-28
```

Grade 4.75

d.
i. H0: β2 = β3 = β4 = 0 vs Ha: at least one of β2,
ii. The value of the test statistic is 128.83
iii. p-value < 0.001
iv. Since p-value < 0.001 < 0.05, we reject H0. We
family history of diabetes, and age on the proba
periodontitis

We have evidence at the 5% level that at least one of BMI, family history of diabetes, and age has an effect on the probability of hyperglycemia, after adjusting for periodontitis.

your conclusion is slightly incorrect or incomplete. -0.25

Q4

9 / 9

Q4

Q4

a) Let Y_i be the number of car insurance claims of the i^{th} individual, respectively, $i = 1, \cdots, 400$

Let x_i be the annual cost (in dollars) of the insurance premium of the i^{th} individual, respectively, $i = 1, \cdots, 400$

$Y_i \sim \text{Poisson}(\mu_i)$, the Y_i's are independent

$\log(\mu_i) = \beta_0 + \beta_1 x_{1i}$

Correct 3

b) A 1 dollar increase in premium is associated with a change of β_1 in the log mean number of car insurance claims.

Correct 2

CamScanner로 스캔하기

Q4

2023-12-02

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.4.4     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
df=read.table("insurance.txt",header=TRUE)

fit=glm(claims~premium,poisson(log),df)
```

c.

```
summary(fit)
```

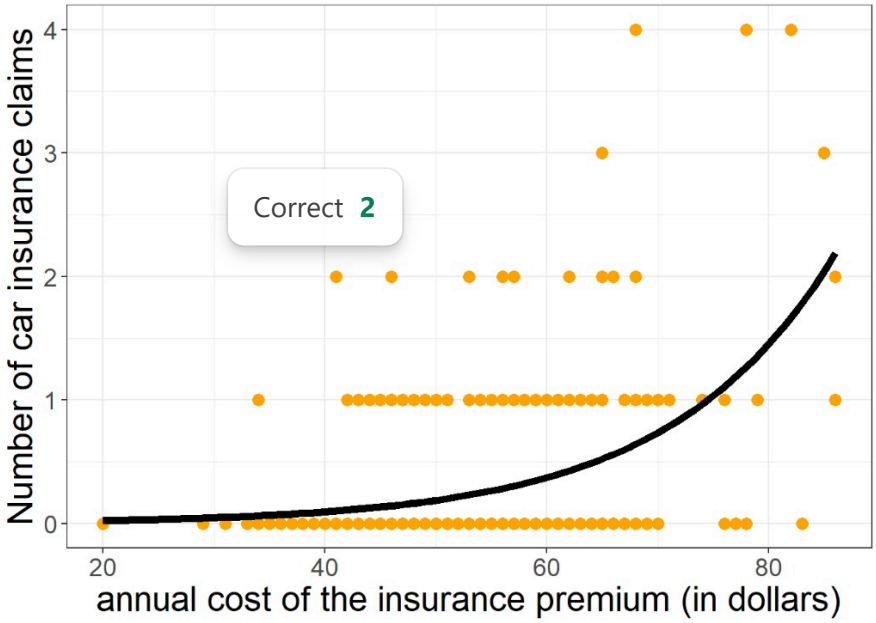
```
##
## Call:
## glm(formula = claims ~ premium, family = poisson(log), data = df)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.07087    0.49139 -10.319   <2e-16 ***
## premium      0.06807    0.00788   8.639   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 372.60  on 399  degrees of freedom
## Residual deviance: 303.41  on 398  degrees of freedom
## AIC: 515.17
##
## Number of Fisher Scoring iterations: 6
```

Correct 1

d.

```
n=100
x=seq(min(df$premium),max(df$premium),length=n)
df1=data.frame(x,y=exp(coef(fit)[1]+coef(fit)[2]*x))

#Plot the data and the fitted regression function.
ggplot()+
  geom_point(data=df,aes(x=premium,y=claims),colour="orange",size=3)+
  geom_line(data=df1,aes(x=x,y=y),linewidth=2)+
  labs(y="Number of car insurance claims",x="annual cost of the insurance premium (in dollars)")
+
  theme_bw()+
  theme(axis.title=element_text(size=20),
        axis.text=element_text(size=14))
```



Correct 2

e.

```
confint(fit,parm="premium",level=0.95)
```

```
## Waiting for profiling to be done...
```

```
##      2.5 %     97.5 %
## 0.05249640 0.08339861
```

Correct 1