Group Make-up

The dataset `College` contains a number of variables for 777 different universities and colleges in the US.

Data: `data(College, package="ISLR)`
Source: James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) An Introduction to Statistical Learning with applications in R, www.StatLearning.com, Springer-Verlag, New York
Variables: Description of variables:
- **Private:** A factor with levels No and Yes indicating private or public university
- **Apps:** Number of applications received
- **Accept:** Number of applicants accepted
- **Enroll:** Number of applicants enrolled
- **Top10perc:** New students from top 10 % of high school class
- **Top25perc:** New students from top 25 % of high school class
- **F.Undergrad:** Number of full-time undergraduates
- **P.Undergrad:** Number of part-time undergraduates
- **Outstate:** Out-of-state tuition
- **Room.Board:** Room and board costs
- **Books:** Estimated book costs
- **Personal:** Estimated personal spending
- **PhD:** Percent of faculty with Ph.D.s
- **Terminal:** Percent of faculty with terminal degree
- **S.F.Ratio:** Student/faculty ratio
- **perc.alumni:** Percent of alumni who donate
- **Expend:** Instructional expenditure per student
- **Grad.Rate:** Graduation rate

```
library(car)
data(College, package = "ISLR")
```

(1) First of all, load the data file `College` (the data frame is called `College`) and the libraries you typically use.

Plot a box plot of the graduation rate (variable `Grad.Rate` using the public/private indicator (variable `Private`) as grouping variable.

```
options(width = 60)
par(mfrow = c(1, 1))
boxplot(Grad.Rate ~ Private, data = College, varwidth = TRUE,
    main = "Graduation Rate by college type", ylab = "Graduation Rate",
    xlab = "Private institution")
```

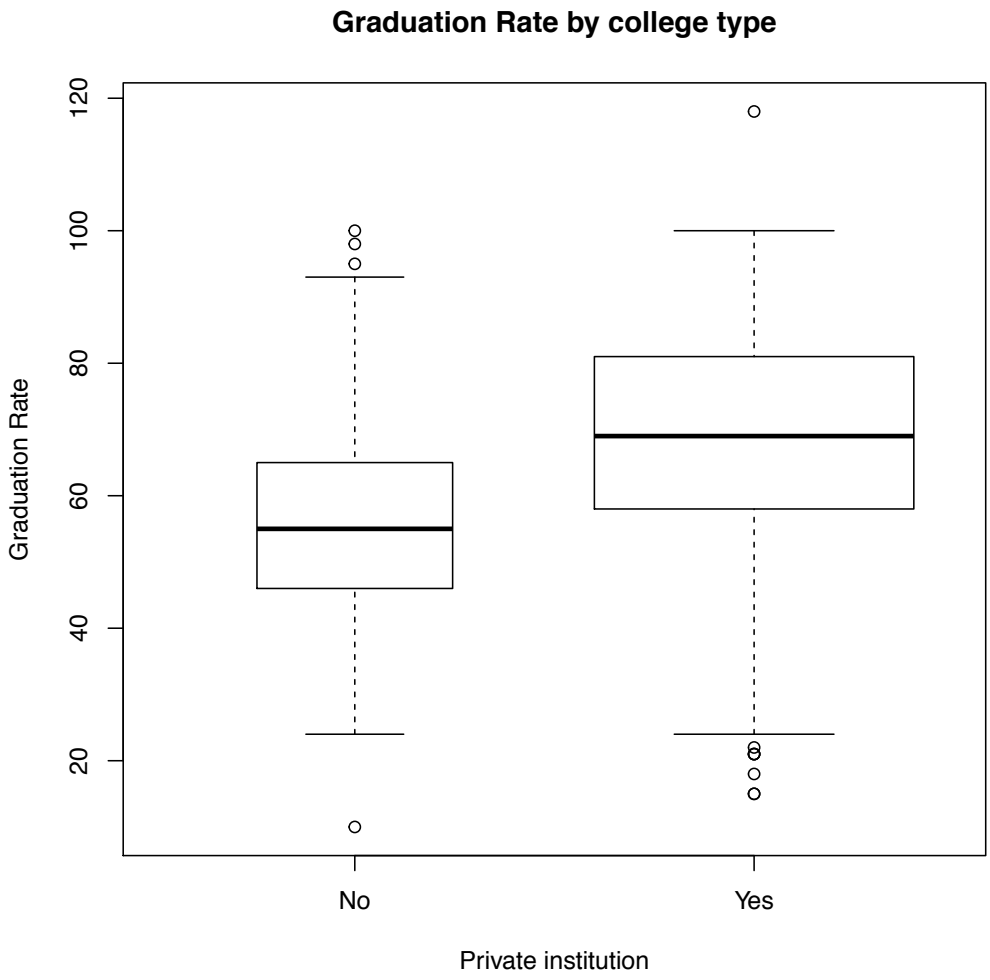

FIGURE 7. Boxplot of graduation rate by type of institution.

```
College.gr.sort <- College[order(College$Grad.Rate, decreasing = TRUE,
    na.last = NA), ]
```

(a) (half a point) Do private institutions have higher median graduation rates than public ones?
Yes

(b) (half a point) Are there any unusual graduation rates? Which ones are unusual?
Yes, the ones above 100%.

(c) (half a point) Which institution has the highest graduation rate?
Cazenovia College

(d) (half a point) Which institution has the lowest graduation rate?

Texas Southern University

(e) (1 point) Looking at the boxplot does homoscedasticity hold for the two groups? Give reasons for your answer!
Overall range differs. Interquartile range a bit larger for private institutions. Upper whiskers for private institutions shorter than the other three whiskers, wich are all pretty similar in length. Both groups have outliers at both ends. Overall, tendency towards heteroscedasticity.

(2) (1 point) Using an appropriate statistical test check whether variances of graduation rates are equal for private and public institutions. Are they? Report the test statistic!

```
grad.rate.var <- tapply(College$Grad.Rate, College$Private, var,
    na.rm = TRUE)
grad.rate.var.test <- bartlett.test(Grad.Rate ~ Private, data = College)
```

No, they are not. There is a highly significant difference in the variances of graduation rates for public and private insitutions as given by the Bartlett's test with a test-statistic of $K^2 = 5.6261$ with 1 degrees of freedom yielding a p-value of $p < 0.0176947$.

(3) Using the appropriate version of the t-test, you want to check whether average graduation rates differ between private and public institutions.

```
grad.rate.t <- t.test(Grad.Rate ~ Private, data = College, var.equal = FALSE)
grad.rate.t
##
##  Welch Two Sample t-test
##
## data:  Grad.Rate by Private
## t = -10.5793, df = 431.974, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -15.36276 -10.54880
## sample estimates:
##  mean in group No mean in group Yes
##          56.04245           68.99823
```

(a) (1 point) Are graduation rates on average the same at public and private institutions? Report the test statistic and the p-value!
No, they are not. There is a highly significant difference in the mean graduation rates for public and private insitutions as given by the Welch t-test with a test-statistic of $t = -10.5793$ with 431.9739 degrees of freedom yielding a p-value of $p < 2.0522429 \times 10^{-23}$.

(b) (half a point) Which institutions have the lower average graduation rates? Private ones or public ones?
public ones ($\bar{x}_{pub} = 56.0424528$ as opposed to $\bar{x}_{priv} = 68.9982301$)

(c) (1 point) Report the 95% confidence interval for the difference in means.
The 95% confidence interval is given by $-15.3627584, -10.5487961$.

(4) Compute a linear model for graduation rate as dependent variable using all variables in the data set as predictors. [The data set also contains the names of the institution, but these will be ignored by default. So, you should not have to worry about them.]

```
grad.rate.lm <- lm(Grad.Rate ~ ., data = College)
anova(grad.rate.lm)
## Analysis of Variance Table
##
## Response: Grad.Rate
##              Df Sum Sq Mean Sq  F value    Pr(>F)
## Private       1  25876 25875.6 159.2894 < 2.2e-16 ***
## Apps          1  24007 24006.9 147.7856 < 2.2e-16 ***
## Accept        1   3160  3159.8  19.4515 1.181e-05 ***
## Enroll        1    247   246.9   1.5198 0.2180317
## Top10perc     1  24486 24486.1 150.7359 < 2.2e-16 ***
## Top25perc     1   2089  2089.2  12.8611 0.0003570 ***
## F.Undergrad   1   2195  2195.2  13.5138 0.0002535 ***
## P.Undergrad   1   2731  2730.8  16.8104 4.576e-05 ***
## Outstate      1  10483 10483.2  64.5345 3.620e-15 ***
## Room.Board    1    897   897.5   5.5247 0.0190047 *
## Books         1    634   633.6   3.9006 0.0486314 *
## Personal      1   1559  1559.1   9.5979 0.0020199 **
## PhD           1    217   216.6   1.3335 0.2485450
## Terminal      1    166   165.6   1.0195 0.3129548
## S.F.Ratio     1    282   281.8   1.7350 0.1881724
## perc.alumni   1   5230  5230.1  32.1966 1.982e-08 ***
## Expend        1   1424  1424.2   8.7671 0.0031628 **
## Residuals   759 123295   162.4
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
summary(grad.rate.lm)
##
## Call:
## lm(formula = Grad.Rate ~ ., data = College)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.897  -7.132  -0.292   7.213  54.056
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.8736716  4.8480858   6.987 6.15e-12 ***
```

```
## PrivateYes   3.3813758  1.6965147   1.993 0.046605 *
## Apps         0.0012984  0.0004418   2.939 0.003390 **
## Accept      -0.0006961  0.0008627  -0.807 0.419995
## Enroll       0.0021593  0.0023081   0.936 0.349814
## Top10perc    0.0548964  0.0717587   0.765 0.444501
## Top25perc    0.1351288  0.0549667   2.458 0.014179 *
## F.Undergrad -0.0004712  0.0004008  -1.176 0.240138
## P.Undergrad -0.0014836  0.0003902  -3.802 0.000155 ***
## Outstate     0.0010174  0.0002334   4.359 1.49e-05 ***
## Room.Board   0.0019143  0.0005908   3.240 0.001246 **
## Books       -0.0022205  0.0029168  -0.761 0.446739
## Personal    -0.0016635  0.0007698  -2.161 0.031000 *
## PhD          0.0872827  0.0568102   1.536 0.124859
## Terminal    -0.0747023  0.0623172  -1.199 0.231002
## S.F.Ratio    0.0758222  0.1593102   0.476 0.634254
## perc.alumni  0.2793343  0.0491750   5.680 1.91e-08 ***
## Expend      -0.0004565  0.0001542  -2.961 0.003163 **
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.75 on 759 degrees of freedom
## Multiple R-squared:  0.4615,Adjusted R-squared:  0.4495
## F-statistic: 38.27 on 17 and 759 DF,  p-value: < 2.2e-16
```

(a) (2 points) Which predictors are statistically significant (at least) at the 0.1% level?
Predictors: `P.Undergrad`, `Outstate` and `perc.alumni` are significant at the 0.1% level.

(b) (1 point) Looking at the regression coefficients, do the signs of the coefficients for these predictors (i.e. the ones significant at least at the 0.1% level) make intuitive sense?
Yes, the coefficient for `P.Undergrad` is negative. Part-time students typically work in parallel to their studies. Due to time constraints they more often tend to not graduate. The other two coefficients are positive, which seems reasonabel for the percent of alumni donating and also for the outstate tuition fees which represent both reputation but also means that the universities care about their students.

(c) (1 point) How good does this model fit? On what do you base your judgment?
As indicated by *adjusted R-squared*, the model fits fairly well with explaining about 44.9480188% variability in the wages.

(d) (1 point) Do the residuals of this model follow a normal distribution? Which tool did you use for checking this?

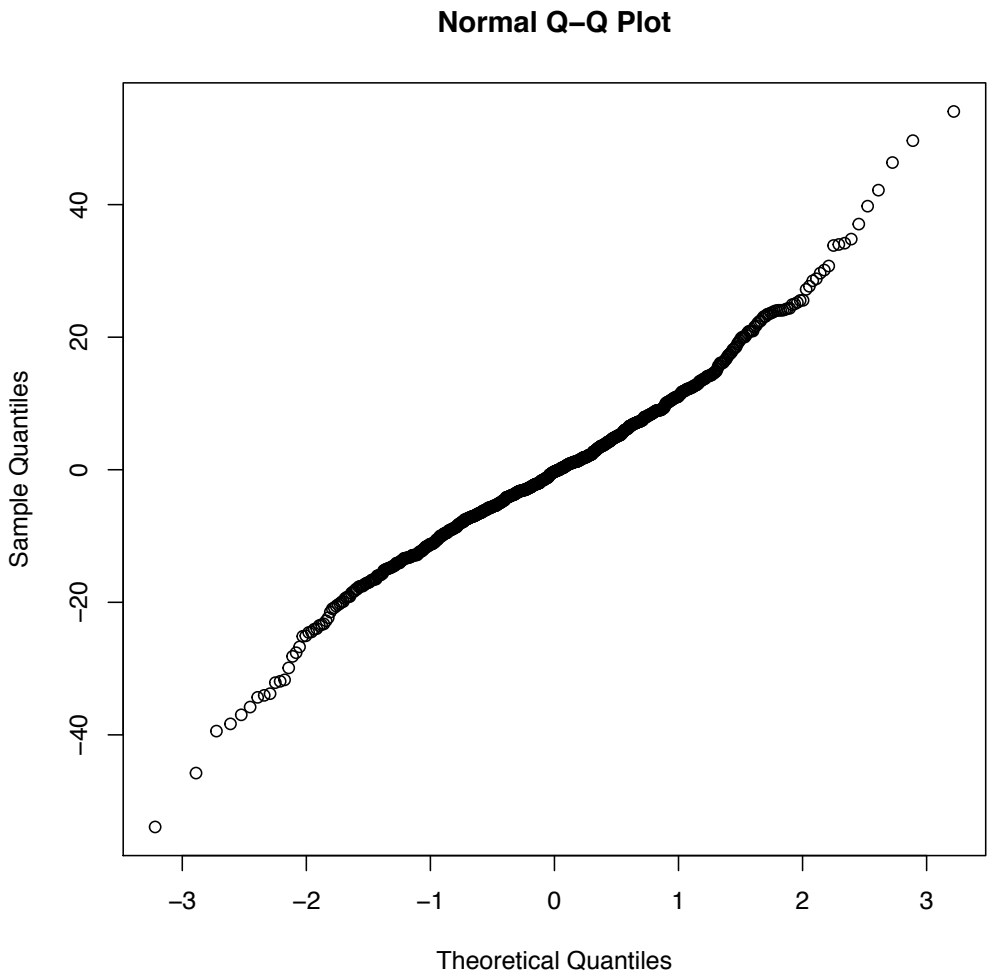

FIGURE 8. Normal Q-Q plots for the model residuals indicating a deviation from a straight line.

```
qqnorm(grad.rate.lm$residuals)
```

```
sd(grad.rate.lm$residuals)
## [1] 12.60497
shapiro.test(grad.rate.lm$residuals)
##
##  Shapiro-Wilk normality test
##
## data:  grad.rate.lm$residuals
## W = 0.9815, p-value = 2.355e-08
```

Using a Q-Q plot, I found some minor deviations from a straight line, so in essence residuals can be considered normal.

(5) Starting with the null model and taking the current model as upper bound, run a stepwise model selection procedure to find the best model according to the AIC criterion.

```
library(MASS)
grad.rate.null <- lm(Grad.Rate ~ 1, data = College)
grad.rate.best <- stepAIC(grad.rate.null, scope = list(upper = grad.rate.lm,
    lower = ~1), direction = "both")
## Start:  AIC=4419.97
## Grad.Rate ~ 1
##
##               Df Sum of Sq    RSS    AIC
## + Outstate     1     74732 154245 4115.0
## + Top10perc    1     56103 172875 4203.6
## + perc.alumni  1     55179 173798 4207.7
## + Top25perc    1     52160 176817 4221.1
## + Room.Board   1     41348 187630 4267.2
## + Expend       1     34889 194089 4293.5
## + Private      1     25876 203102 4328.8
## + S.F.Ratio    1     21540 207437 4345.2
## + PhD          1     21306 207671 4346.1
## + Terminal     1     19194 209783 4353.9
## + Personal     1     16611 212366 4363.5
## + P.Undergrad  1     15124 213853 4368.9
## + Apps         1      4931 224046 4405.1
## + F.Undergrad  1      1421 227556 4417.1
## + Accept       1      1037 227940 4418.4
## <none>                     228977 4420.0
## + Enroll       1       114 228863 4421.6
## + Books        1         0 228977 4422.0
##
## Step:  AIC=4115
## Grad.Rate ~ Outstate
##
##               Df Sum of Sq    RSS    AIC
## + Top25perc    1     11768 142478 4055.3
## + Top10perc    1     10108 144138 4064.3
## + perc.alumni  1      9445 144800 4067.9
## + Apps         1      3202 151044 4100.7
## + P.Undergrad  1      3079 151166 4101.3
## + Personal     1      2439 151807 4104.6
```

```
## + PhD          1      1996 152250 4106.9
## + Accept       1      1542 152704 4109.2
## + Room.Board   1      1048 153197 4111.7
## + Enroll       1      1037 153208 4111.8
## + Terminal     1       875 153370 4112.6
## + F.Undergrad  1       475 153770 4114.6
## <none>                     154245 4115.0
## + Private      1       138 154108 4116.3
## + Books        1       102 154143 4116.5
## + S.F.Ratio    1        35 154211 4116.8
## + Expend       1        15 154230 4116.9
## - Outstate     1     74732 228977 4420.0
##
## Step:  AIC=4055.34
## Grad.Rate ~ Outstate + Top25perc
##
##               Df Sum of Sq    RSS    AIC
## + perc.alumni  1      5998 136480 4023.9
## + P.Undergrad  1      4196 138281 4034.1
## + Personal     1      3377 139101 4038.7
## + Private      1      1517 140960 4049.0
## + Expend       1       958 141520 4052.1
## + Room.Board   1       932 141546 4052.2
## + Books        1       496 141982 4054.6
## <none>                     142478 4055.3
## + F.Undergrad  1       334 142144 4055.5
## + Apps         1       292 142186 4055.7
## + Terminal     1       260 142218 4055.9
## + Top10perc    1       224 142254 4056.1
## + S.F.Ratio    1        88 142390 4056.9
## + Accept       1        52 142425 4057.1
## + Enroll       1        40 142438 4057.1
## + PhD          1        16 142462 4057.2
## - Top25perc    1     11768 154245 4115.0
## - Outstate     1     34339 176817 4221.1
##
## Step:  AIC=4023.92
## Grad.Rate ~ Outstate + Top25perc + perc.alumni
##
##               Df Sum of Sq    RSS    AIC
## + P.Undergrad  1    2580.2 133900 4011.1
## + Personal     1    2096.8 134383 4013.9
## + Room.Board   1    1919.6 134561 4014.9
```

```
## + Apps          1    1347.5 135133 4018.2
## + Expend        1     958.1 135522 4020.4
## + Accept        1     758.8 135721 4021.6
## + Private       1     555.4 135925 4022.8
## + S.F.Ratio     1     370.3 136110 4023.8
## <none>                      136480 4023.9
## + Books         1     213.8 136266 4024.7
## + Terminal      1     178.3 136302 4024.9
## + Top10perc     1      97.2 136383 4025.4
## + Enroll        1      85.9 136394 4025.4
## + F.Undergrad   1       1.2 136479 4025.9
## + PhD           1       0.0 136480 4025.9
## - perc.alumni   1    5997.5 142478 4055.3
## - Top25perc     1    8320.2 144800 4067.9
## - Outstate      1   16770.5 153251 4112.0
##
## Step:  AIC=4011.09
## Grad.Rate ~ Outstate + Top25perc + perc.alumni + P.Undergrad
##
##               Df Sum of Sq    RSS    AIC
## + Apps          1    3864.5 130035 3990.3
## + Accept        1    2892.9 131007 3996.1
## + Room.Board    1    2493.1 131407 3998.5
## + Enroll        1    1470.5 132430 4004.5
## + Personal      1    1213.4 132687 4006.0
## + F.Undergrad   1     953.7 132946 4007.5
## + Expend        1     668.5 133231 4009.2
## + S.F.Ratio     1     586.1 133314 4009.7
## <none>                      133900 4011.1
## + PhD           1     189.5 133711 4012.0
## + Books         1     125.1 133775 4012.4
## + Top10perc     1      64.3 133836 4012.7
## + Private       1      35.8 133864 4012.9
## + Terminal      1       0.0 133900 4013.1
## - P.Undergrad   1    2580.2 136480 4023.9
## - perc.alumni   1    4381.3 138281 4034.1
## - Top25perc     1    9352.2 143252 4061.5
## - Outstate      1   14439.6 148340 4088.7
##
## Step:  AIC=3990.33
## Grad.Rate ~ Outstate + Top25perc + perc.alumni + P.Undergrad +
##     Apps
##
```

```
##               Df Sum of Sq    RSS    AIC
## + Room.Board   1    1862.6 128173 3981.1
## + Personal     1    1533.4 128502 3983.1
## + Expend       1    1519.1 128516 3983.2
## + Private      1    1163.7 128872 3985.4
## + F.Undergrad  1     736.1 129299 3987.9
## + Enroll       1     389.7 129646 3990.0
## <none>                     130035 3990.3
## + S.F.Ratio    1     282.4 129753 3990.6
## + Books        1     210.1 129825 3991.1
## + Terminal     1     134.4 129901 3991.5
## + Accept       1     107.6 129928 3991.7
## + Top10perc    1       4.5 130031 3992.3
## + PhD          1       2.2 130033 3992.3
## - Apps         1    3864.5 133900 4011.1
## - Top25perc    1    4004.1 134040 4011.9
## - P.Undergrad  1    5097.2 135133 4018.2
## - perc.alumni  1    5799.6 135835 4022.2
## - Outstate     1   13840.5 143876 4066.9
##
## Step:  AIC=3981.13
## Grad.Rate ~ Outstate + Top25perc + perc.alumni + P.Undergrad +
##     Apps + Room.Board
##
##               Df Sum of Sq    RSS    AIC
## + Expend       1    1814.5 126358 3972.0
## + Personal     1    1305.6 126867 3975.2
## + Private      1     890.3 127283 3977.7
## + F.Undergrad  1     465.6 127707 3980.3
## + Books        1     372.3 127801 3980.9
## + S.F.Ratio    1     353.0 127820 3981.0
## <none>                     128173 3981.1
## + Terminal     1     280.1 127893 3981.4
## + Enroll       1     165.7 128007 3982.1
## + Accept       1      34.8 128138 3982.9
## + PhD          1       1.6 128171 3983.1
## + Top10perc    1       1.4 128171 3983.1
## - Room.Board   1    1862.6 130035 3990.3
## - Apps         1    3234.1 131407 3998.5
## - Top25perc    1    4077.1 132250 4003.5
## - Outstate     1    4542.1 132715 4006.2
## - P.Undergrad  1    5455.8 133629 4011.5
## - perc.alumni  1    6523.8 134697 4017.7
```

```
##
## Step:  AIC=3972.05
## Grad.Rate ~ Outstate + Top25perc + perc.alumni + P.Undergrad +
##     Apps + Room.Board + Expend
##
##               Df Sum of Sq    RSS    AIC
## + Personal     1    1013.1 125345 3967.8
## + Private      1     838.5 125520 3968.9
## + F.Undergrad  1     639.7 125719 3970.1
## + Accept       1     348.8 126010 3971.9
## <none>                     126358 3972.0
## + Top10perc    1     289.0 126069 3972.3
## + Books        1     275.2 126083 3972.4
## + Enroll       1     271.3 126087 3972.4
## + Terminal     1     189.9 126169 3972.9
## + PhD          1       4.2 126354 3974.0
## + S.F.Ratio    1       3.9 126354 3974.0
## - Expend       1    1814.5 128173 3981.1
## - Room.Board   1    2158.0 128516 3983.2
## - Apps         1    4083.7 130442 3994.8
## - Top25perc    1    5047.0 131405 4000.5
## - P.Undergrad  1    5385.1 131743 4002.5
## - Outstate     1    6221.0 132579 4007.4
## - perc.alumni  1    6952.7 133311 4011.7
##
## Step:  AIC=3967.79
## Grad.Rate ~ Outstate + Top25perc + perc.alumni + P.Undergrad +
##     Apps + Room.Board + Expend + Personal
##
##               Df Sum of Sq    RSS    AIC
## + Private      1     804.4 124541 3964.8
## + F.Undergrad  1     461.7 124884 3966.9
## + Accept       1     325.6 125020 3967.8
## <none>                     125345 3967.8
## + Top10perc    1     317.3 125028 3967.8
## + Terminal     1     188.2 125157 3968.6
## + Enroll       1     177.9 125167 3968.7
## + Books        1     128.7 125217 3969.0
## + PhD          1       5.4 125340 3969.8
## + S.F.Ratio    1       0.0 125345 3969.8
## - Personal     1    1013.1 126358 3972.0
## - Expend       1    1522.0 126867 3975.2
## - Room.Board   1    1911.9 127257 3977.6
```

```
## - Apps          1    4298.5 129644 3992.0
## - P.Undergrad   1    4322.9 129668 3992.1
## - Top25perc     1    5168.7 130514 3997.2
## - Outstate      1    5492.6 130838 3999.1
## - perc.alumni   1    6198.0 131543 4003.3
##
## Step:  AIC=3964.79
## Grad.Rate ~ Outstate + Top25perc + perc.alumni + P.Undergrad +
##     Apps + Room.Board + Expend + Personal + Private
##
##               Df Sum of Sq    RSS    AIC
## <none>                     124541 3964.8
## + Top10perc    1     280.4 124261 3965.0
## + Accept       1     240.4 124301 3965.3
## + F.Undergrad  1     216.4 124325 3965.4
## + Books        1     159.1 124382 3965.8
## + PhD          1     154.1 124387 3965.8
## + Enroll       1      60.7 124480 3966.4
## + S.F.Ratio    1      37.7 124503 3966.6
## + Terminal     1      21.1 124520 3966.7
## - Private      1     804.4 125345 3967.8
## - Personal     1     979.0 125520 3968.9
## - Expend       1    1480.2 126021 3972.0
## - Room.Board   1    1648.1 126189 3973.0
## - P.Undergrad  1    3464.5 128005 3984.1
## - Outstate     1    3498.1 128039 3984.3
## - Apps         1    5102.6 129644 3994.0
## - Top25perc    1    5446.3 129987 3996.0
## - perc.alumni  1    5756.0 130297 3997.9
```

```
summary(grad.rate.best)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ Outstate + Top25perc + perc.alumni +
##     P.Undergrad + Apps + Room.Board + Expend + Personal + Private,
##     data = College)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.345  -7.551  -0.426   7.040  51.789
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32.9174907  2.5635570  12.841  < 2e-16 ***
```

```
## Outstate     0.0010226  0.0002203   4.641 4.07e-06 ***
## Top25perc    0.1763996  0.0304582   5.792 1.02e-08 ***
## perc.alumni  0.2876422  0.0483114   5.954 3.98e-09 ***
## P.Undergrad -0.0016678  0.0003611  -4.619 4.52e-06 ***
## Apps         0.0009022  0.0001609   5.606 2.89e-08 ***
## Room.Board   0.0018262  0.0005732   3.186  0.00150 **
## Expend      -0.0003888  0.0001288  -3.019  0.00262 **
## Personal    -0.0018394  0.0007491  -2.455  0.01429 *
## PrivateYes   3.3935160  1.5246563   2.226  0.02632 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.74 on 767 degrees of freedom
## Multiple R-squared:  0.4561,Adjusted R-squared:  0.4497
## F-statistic: 71.46 on 9 and 767 DF,  p-value: < 2.2e-16
```

(a) (2 points) Which predictors are not included in the optimal model?
Predictors `Top10per, Accept, F.Undergrad, Books, PhD, Enroll, S.F.Ratio, Terminal` are not included.

(b) (half a point) Report the adjusted R-squared of the final model?
0.4497169

(c) (half a point) Report the AIC of the final model?
3964.7895609

(d) (half a point) Using an F-test check whether the final model is significantly different from the model using all predictors (as computed in Question 4)?

```
anova(grad.rate.best, grad.rate.lm, test = "F")
## Analysis of Variance Table
##
## Model 1: Grad.Rate ~ Outstate + Top25perc + perc.alumni + P.Undergrad +
##     Apps + Room.Board + Expend + Personal + Private
## Model 2: Grad.Rate ~ Private + Apps + Accept + Enroll + Top10perc + Top25pe
##     F.Undergrad + P.Undergrad + Outstate + Room.Board + Books +
##     Personal + PhD + Terminal + S.F.Ratio + perc.alumni + Expend
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    767 124541
## 2    759 123295  8      1246 0.9588 0.4672
```

No, it is not.

(e) (half a point) According to this model, how does the instructional expenditure per student influence the graduation rate?
The higher the instructional expenditure per student, the lower the graduation rate.

(f) (half a point) According to this model, how does the number of applications received influence the graduation rate?
The higher the number of applications received, the higher the graduation rate.

(6) (half a point) Amend your model by adding an interaction term between `Room.Board` and `Private`.

```
grad.rate.best.up <- update(grad.rate.best, ~. + Room.Board:Private,
    data = College)
summary(grad.rate.best.up)
##
## Call:
## lm(formula = Grad.Rate ~ Outstate + Top25perc + perc.alumni +
##     P.Undergrad + Apps + Room.Board + Expend + Personal + Private +
##     Room.Board:Private, data = College)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.046  -7.488  -0.353   7.180  51.439
##
## Coefficients:
##                          Estimate Std. Error t value
## (Intercept)           39.2899007  4.3583599   9.015
## Outstate               0.0009672  0.0002221   4.354
## Top25perc              0.1811139  0.0305252   5.933
## perc.alumni            0.2836184  0.0482917   5.873
## P.Undergrad           -0.0016758  0.0003606  -4.648
## Apps                   0.0009512  0.0001630   5.836
## Room.Board             0.0001624  0.0010844   0.150
## Expend                -0.0004059  0.0001289  -3.148
## Personal              -0.0018912  0.0007485  -2.527
## PrivateYes            -4.4490461  4.6004435  -0.967
## Room.Board:PrivateYes  0.0021235  0.0011754   1.807
##                       Pr(>|t|)
## (Intercept)            < 2e-16 ***
## Outstate              1.52e-05 ***
## Top25perc             4.50e-09 ***
## perc.alumni           6.38e-09 ***
## P.Undergrad           3.95e-06 ***
## Apps                  7.87e-09 ***
## Room.Board             0.88097
## Expend                 0.00171 **
## Personal               0.01172 *
## PrivateYes             0.33380
```

```
## Room.Board:PrivateYes  0.07123 .
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.72 on 766 degrees of freedom
## Multiple R-squared:  0.4584,Adjusted R-squared:  0.4513
## F-statistic: 64.83 on 10 and 766 DF,  p-value: < 2.2e-16
```

(a) (half a point) Is the interaction term significant at the 10% level?
Yes.

(b) (1 point) According to this model and all other things equal, by how much will graduation rates differ for two public colleges that differ in room & board by $ 1000?
Depending on type of the institution: for public colleges: About 1.08 percentage points ($1000 \cdot 1.6243 \times 10^{-4}$)

(c) (1 point) According to this model and all other things equal, by how much will graduation rates differ for two private colleges that differ in room & board by $ 1000?
for private colleges: About 3.2 percentage points units ($1000 \cdot 1.6243 \times 10^{-4} + 1000 \cdot 0.0021235$)

(7) You now want to generate a classification model that tells you whether an institution is private or public based on the following predictors: `Room.Board`, `Apps`, `Enroll`, `Outstate`, `perc.alumni`, `Personal`, `Books`. Use the `probit` link here!

```
private.prob <- glm(Private ~ Room.Board + Apps + Enroll + Outstate +
    perc.alumni + Personal + Books, family = binomial(probit),
    data = College)
## Warning:  glm.fit:  fitted probabilities numerically 0 or 1 occurred
summary(private.prob)
##
## Call:
## glm(formula = Private ~ Room.Board + Apps + Enroll + Outstate +
##     perc.alumni + Personal + Books, family = binomial(probit),
##     data = College)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.9889  -0.0468   0.0437   0.2454   3.8557
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.639e+00  4.952e-01  -5.329 9.87e-08 ***
```

```
## Room.Board   2.710e-05  1.131e-04   0.240 0.810587
## Apps        -2.323e-04  6.478e-05  -3.586 0.000335 ***
## Enroll      -3.897e-04  2.413e-04  -1.615 0.106393
## Outstate     3.830e-04  4.368e-05   8.768  < 2e-16 ***
## perc.alumni  3.315e-02  9.268e-03   3.577 0.000347 ***
## Personal    -1.953e-04  1.105e-04  -1.767 0.077155 .
## Books        9.584e-04  6.071e-04   1.579 0.114393
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 910.75  on 776  degrees of freedom
## Residual deviance: 316.52  on 769  degrees of freedom
## AIC: 332.52
##
## Number of Fisher Scoring iterations: 8
Anova(private.prob)
## Warning:  glm.fit:  fitted probabilities numerically 0 or 1 occurred
## Warning:  glm.fit:  fitted probabilities numerically 0 or 1 occurred
## Warning:  glm.fit:  fitted probabilities numerically 0 or 1 occurred
## Warning:  glm.fit:  fitted probabilities numerically 0 or 1 occurred
## Warning:  glm.fit:  fitted probabilities numerically 0 or 1 occurred
## Analysis of Deviance Table (Type II tests)
##
## Response: Private
##             LR Chisq Df Pr(>Chisq)
## Room.Board     0.058  1  0.8091937
## Apps          17.066  1  3.609e-05 ***
## Enroll         3.461  1  0.0628445 .
## Outstate     113.668  1  < 2.2e-16 ***
## perc.alumni   13.668  1  0.0002182 ***
## Personal       2.897  1  0.0887256 .
## Books          2.700  1  0.1003687
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) (1 point) Which error distribution have you chosen to create this model?
binomial distribution

(b) (1 point) According to the Wald tests (Table of coefficients): which predictors are significant (at least at the 5% level)?
`Apps`, `Outstate` and `perc.alumni` are significant at the 5% level.

(c) (1 point) According to the Likelihood-Ratio Test (LR-test as given in the Deviance Table): which predictors are significant (at least at the 5% level)?
`Apps, Outstate` and `perc.alumni` are significant at the 5% level.

(d) (half a point) Report the residual deviance of your model?
316.5238471

(e) (half a point) Report the Null deviance of your model?
910.7486478

(f) (half a point) Using a χ^2-test check whether your model is significantly better than the null model?

```
private.null <- glm(Private ~ 1, family = binomial(probit), data = College)
anova(private.null, private.prob, test = "Chisq")
## Analysis of Deviance Table
##
## Model 1: Private ~ 1
## Model 2: Private ~ Room.Board + Apps + Enroll + Outstate + perc.alumni +
##     Personal + Books
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       776     910.75
## 2       769     316.52  7   594.22 < 2.2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Yes, it is.

(g) (half a point) According to your model, how do the number of applications to a college influence the likelihood of it being a private institution?
The higher the number of applications the less likely the institution is private.

(h) (2 points) Based on your model's fitted probabilites for being a private institution create an indicator for private/public institution using the probability 0.5 as threshold. Create a frequency table of the predicted and the observed private/public indicator. Calculate all misclassification rates.

```
a <- table(College$Private, private.prob$fitted > 0.5)
a
##
##       FALSE TRUE
##   No    179   33
##   Yes    32  533
```

The misclassification rate for 'public' (i.e. 'No') cases is 15.17%.
The misclassification rate for 'private' (i.e. 'Yes') cases is 5.83%.
The overall misclassification rate is 8.37%.