

Homework 1

Shun-Lung Chang, Dilip Hiremath

```
load("data/bank.Rdata") # read the data into R
```

1. To start with, you compute the naive model for current salary (SALNOW) as the dependent variable.

(a) Calculate the model and specify the model equation.

```
mod <- lm(SALNOW ~ 1, data = bank)
summary(mod)
```

Call:

```
lm(formula = SALNOW ~ 1, data = bank)
```

Residuals:

Min	1Q	Median	3Q	Max
-7468	-4168	-2218	1007	40232

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	13767.8	313.7	43.88	<2e-16

Residual standard error: 6830 on 473 degrees of freedom

As the summary shows, the model equation is $\widehat{Salary}_i = 13767.8$.

(b) Compute the residual sum of squares for this model, i.e. compute the sum of the squared residuals.

```
deviance(mod)
```

```
[1] 22066639270
```

The sum of the squared residuals is 22066639270.

(c) Compute the residual standard error for this model, i.e. compute the square root of the residual sum of squares divided by $n - 1$, where n is the sample size.

```
sqrt(deviance(mod)/(nrow(bank) - 1))
```

```
[1] 6830.265
```

The residual standard error is 6830.26.

2. As second step, you compute a simple linear regression model for current salary (SALNOW) as the dependent variable using education level (EDLEVEL) as a predictor.

(a) Calculate the model and specify the model equation.

```
mod_2 <- lm(SALNOW ~ EDLEVEL, data = bank)
summary(mod_2)
```

Call:

```
lm(formula = SALNOW ~ EDLEVEL, data = bank)
```

Residuals:

Min	1Q	Median	3Q	Max
-8627	-3284	-1001	2351	31617

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-7332.47	1128.76	-6.496	0.00000000021
EDLEVEL	1563.96	81.82	19.115	< 2e-16

Residual standard error: 5133 on 472 degrees of freedom

Multiple R-squared: 0.4363, Adjusted R-squared: 0.4351

F-statistic: 365.4 on 1 and 472 DF, p-value: < 2.2e-16

The model equation is $\widehat{Salary}_i = -7332.47 + 1563.96 * EDLEVEL_i$.

(b) Compute the residual sum of squares for this model, i.e. compute the sum of the squared residuals.

```
deviance(mod_2)
```

```
[1] 12438124428
```

The residual sum of squares is 12438124428.

(c) Compute the residual standard error for this model, i.e. compute the square root of the residual sum of squares divided by $n - 2$, where n is the sample size.

```
sqrt(deviance(mod_2)/(nrow(bank) - 2))
```

```
[1] 5133.416
```

The residual standard error is 5133.42.

3. In a third model, you add gender (SEX) as an additional predictor to education level.

(a) Calculate the model and specify the model equation.

```
mod_3 <- lm(SALNOW ~ EDLEVEL + SEX, data = bank)
summary(mod_3)
```

Call:

```
lm(formula = SALNOW ~ EDLEVEL + SEX, data = bank)
```

Residuals:

Min	1Q	Median	3Q	Max
-9263.0	-3077.3	-783.3	2054.7	31223.6

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-6369.78	1084.52	-5.873	0.0000000080779
EDLEVEL	1356.67	83.44	16.259	< 2e-16
SEXMale	3369.38	482.81	6.979	0.0000000000102

Residual standard error: 4892 on 471 degrees of freedom

Multiple R-squared: 0.4892, Adjusted R-squared: 0.487

F-statistic: 225.5 on 2 and 471 DF, p-value: < 2.2e-16

The model equation is $\widehat{Salary}_i = -6369.78 + 1356.67 * EDLEVEL_i + 3369.38 * SEX(if Male)_i$.

(b) Compute the residual sum of squares for this model, i.e. compute the sum of the squared residuals.

```
deviance(mod_3)
```

```
[1] 11272531174
```

The residual sum of squares is 11272531174.

(c) Compute the residual standard error for this model, i.e. compute the square root of the residual sum of squares divided by $n - 3$, where n is the sample size.

```
sqrt(deviance(mod_3)/(nrow(bank) - 3))
```

```
[1] 4892.156
```

The residual standard error is 4892.16.

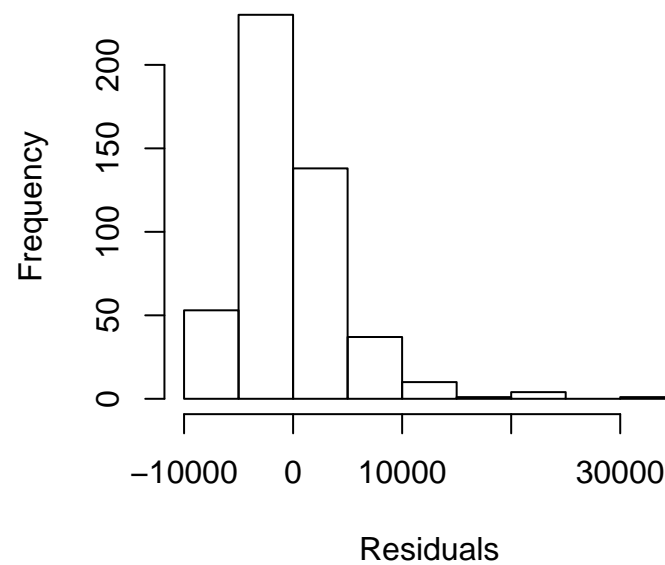
4. You continue with the last model using education level and gender as predictors and investigate the residuals in more detail.

(a) Draw a histogram, a boxplot, a density plot, and a Q-Q-plot to assess normality of the residuals. Give a brief summary report on these plots!

```
res <- resid(mod_3)
```

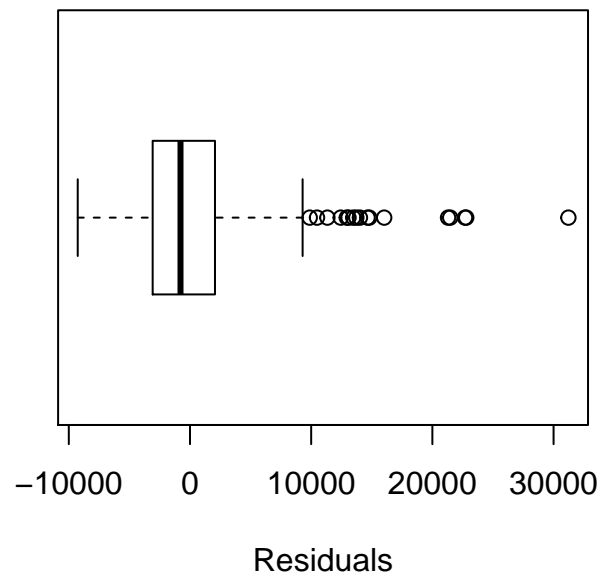
```
hist(res, main = 'Histogram of Residuals', xlab = 'Residuals')
```

Histogram of Residuals

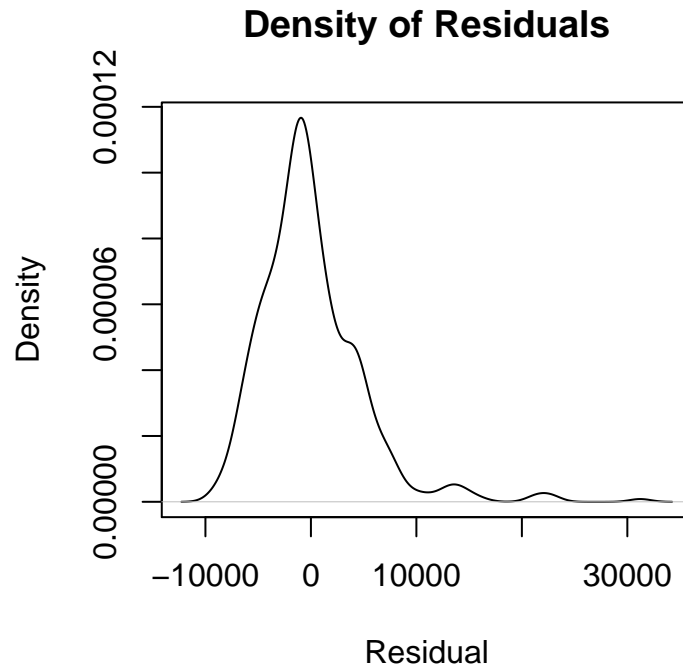


```
boxplot(res, main = 'Boxplot of Residuals', xlab = 'Residuals', horizontal = TRUE)
```

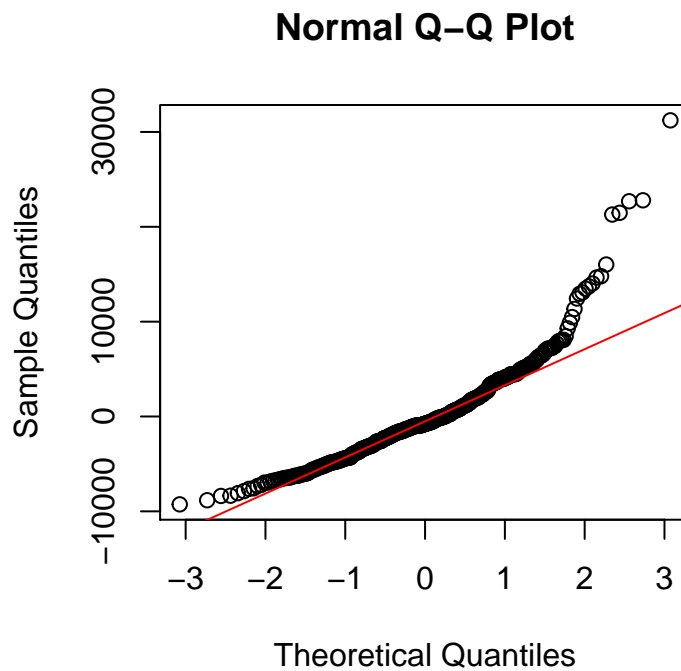
Boxplot of Residuals



```
plot(density(res), main = "Density of Residuals", xlab = "Residual")
```



```
qqnorm(res)
qqline(res, col = "red")
```



The distribution is right-skewed and the points in the Q-Q plot do not all lie on the theoretical line, and hence the residuals do not follow a normal distribution.

(b) Use the Kolmogorov-Smirnov-Test to check whether the residuals follow a normal distribution.

```
ks.test(res, "pnorm")
```

Warning in ks.test(res, "pnorm"): ties should not be present for the Kolmogorov-Smirnov test

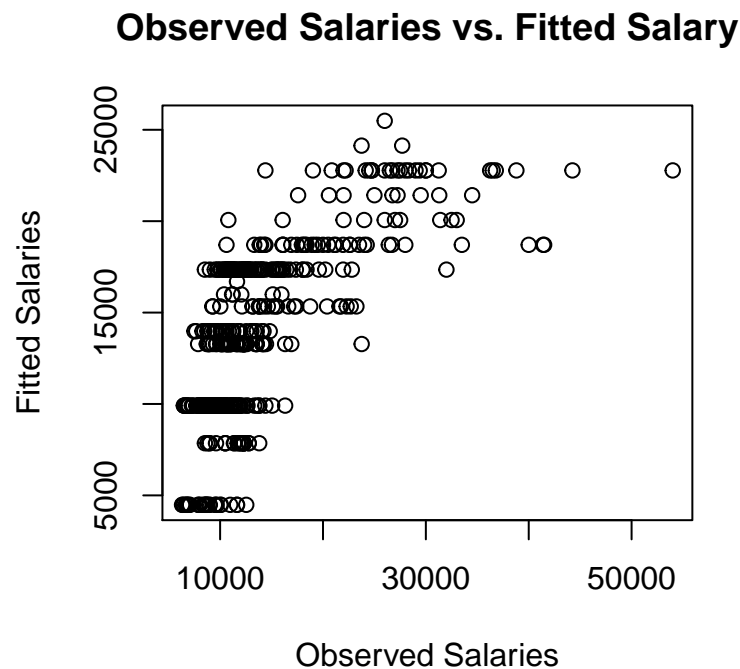
One-sample Kolmogorov-Smirnov test

```
data: res
D = 0.59705, p-value < 2.2e-16
alternative hypothesis: two-sided
```

The p-value is small enough to reject the null hypothesis in Kolmogorov-Smirnov test, thus the residuals are not normally distributed.

5. Plot observed salaries against the ones predicted by the above model (use either the command *fitted* or the stored scores in *modelname\$fitted.values* to obtain the fitted scores). Compute the Pearson correlation coefficient between observed and fitted salaries. How can you check your result using results from the regression table?

```
plot(bank$SALNOW, mod_3$fitted.values,
     main = "Observed Salaries vs. Fitted Salary",
     xlab = "Observed Salaries",
     ylab = "Fitted Salaries")
```



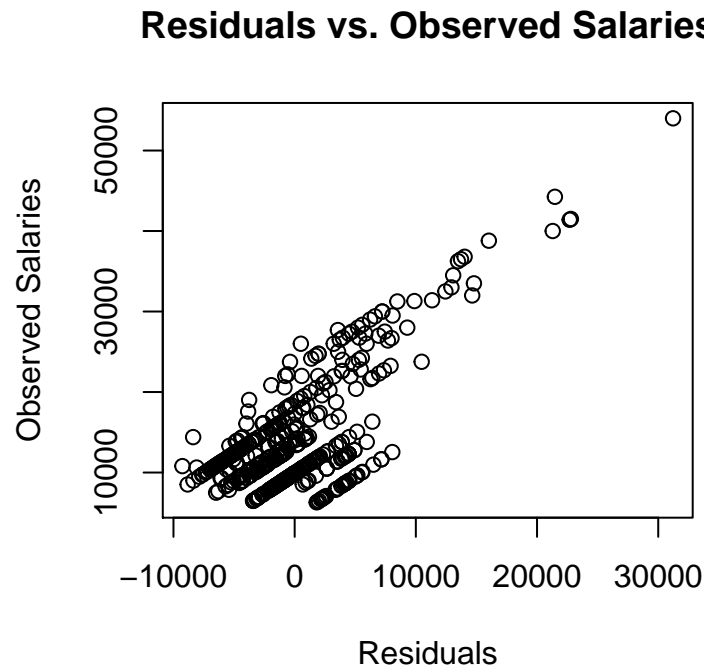
```
cor(bank$SALNOW, mod_3$fitted.values)
```

```
[1] 0.6993994
```

The Pearson correlation coefficient is around 0.699. Besides, the square of the correlation coefficient (0.489) is same as the R-squared reported in Task 3(a).

6. Plot the residuals against the observed salaries. Does the plot look similar to what you had expected? Compute the Pearson correlation coefficient and comment on it!

```
plot(res, bank$SALNOW,  
     main = "Residuals vs. Observed Salaries",  
     xlab = "Residuals",  
     ylab = "Observed Salaries")
```



```
cor(res, bank$SALNOW)
```

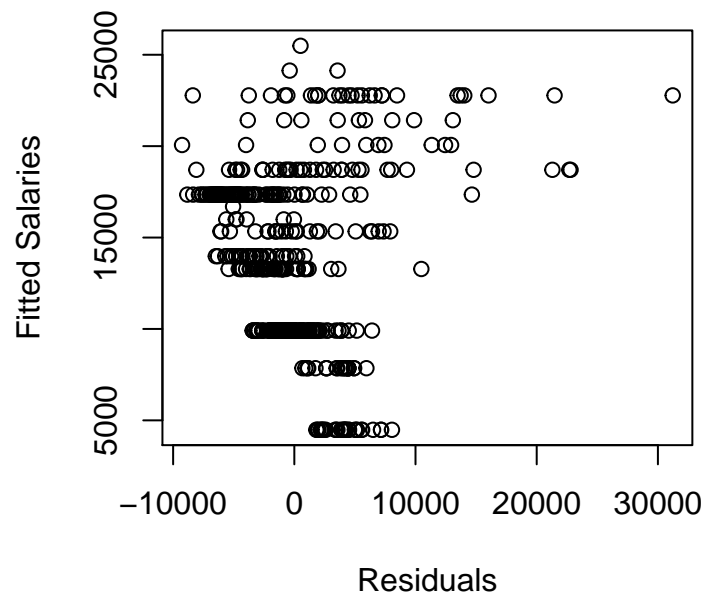
```
[1] 0.714731
```

The Pearson correlation coefficient is 0.71, indicating observed salaries are positively correlated with residuals. The reason lies in the fact that regression models with few predictors tend to predict average values of the response variable. Therefore, the actual salaries deviated greatly from the mean led to larger absolute residuals.

7. Plot the residuals against the fitted salaries. Does the plot look similar to what you had expected? Compute the Pearson correlation coefficient and comment on it!

```
plot(res, mod_3$fitted.values,  
     main = "Residuals vs. Fitted Salaries",  
     xlab = "Residuals",  
     ylab = "Fitted Salaries")
```

Residuals vs. Fitted Salaries



```
cor(res, mod_3$fitted.values)
```

```
[1] 1.491875e-17
```

The infinitesimal coefficient indicates residuals have no linear relationship with fitted values. Thus we can say how fitted values vary do not affect the residuals linearly.

8. In the next analysis step, you want to look at the relationship between the current salary (SALNOW) and all available predictors except ID.

```
mod_4 <- lm(SALNOW ~ . - ID, data = bank)
summary(mod_4)
```

Call:

```
lm(formula = SALNOW ~ . - ID, data = bank)
```

Residuals:

Min	1Q	Median	3Q	Max
-10080.8	-1222.3	-250.1	986.8	18680.1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-2377.95062	1475.56464	-1.612	0.107745
SALBEG	1.41350	0.09043	15.630	< 2e-16
SEXMale	573.73592	326.74078	1.756	0.079765
TIME	58.59362	12.80883	4.574	6.15e-06
AGE	-31.52532	20.29194	-1.554	0.120970
EDLEVEL	181.58295	64.34962	2.822	0.004982
WORK	-66.77509	27.82811	-2.400	0.016812
JOB CATCollegeTrainee	4967.29912	593.55330	8.369	7.05e-16
JOB CATExempt	2792.67798	788.65372	3.541	0.000439

JOB CAT MBATrainee	4022.51936	1366.88277	2.943	0.003417
JOB CAT Office	-190.35313	338.43790	-0.562	0.574086
JOB CAT Security	2517.02042	651.88793	3.861	0.000129
JOB CAT Technical	4142.34777	1615.99866	2.563	0.010684
MINORITY Minority	-391.18092	315.47220	-1.240	0.215613

Residual standard error: 2701 on 460 degrees of freedom
Multiple R-squared: 0.8479, Adjusted R-squared: 0.8436
F-statistic: 197.3 on 13 and 460 DF, p-value: < 2.2e-16

(a) Which variables are significant at the 5% level?

As the table above shows, significant variables are SALBEG, TIME, EDLEVEL, WORK, JOB CAT (except for Office).

(b) How much variability in salaries is explained by this model?

The R-squared shows that 84.79 % of the variability is explained.

(c) Is there evidence for discrimination?

If we consider discrimination with respect to gender, age and ethnic group size, those variables are not statistically significant to conclude discrimination exists.

9. Remove AGE from the previous model.

```
mod_5 <- lm(SALNOW ~ . - ID - AGE, data = bank)
summary(mod_5)
```

Call:

```
lm(formula = SALNOW ~ . - ID - AGE, data = bank)
```

Residuals:

Min	1Q	Median	3Q	Max
-9988.2	-1274.7	-274.2	1002.5	18545.5

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-3402.0567	1322.1600	-2.573	0.01039
SALBEG	1.4078	0.0905	15.556	< 2e-16
SEXMale	733.4405	310.6239	2.361	0.01863
TIME	57.1926	12.7966	4.469	9.89e-06
EDLEVEL	190.2858	64.2035	2.964	0.00320
WORK	-99.8780	17.9271	-5.571	4.30e-08
JOB CAT CollegeTrainee	4996.0276	594.1741	8.408	5.23e-16
JOB CAT Exempt	2816.4936	789.7128	3.566	0.00040
JOB CAT MBATrainee	4040.8572	1368.9259	2.952	0.00332
JOB CAT Office	-7.7030	317.8477	-0.024	0.98068
JOB CAT Security	2616.9634	649.6998	4.028	6.58e-05
JOB CAT Technical	4148.0839	1618.4702	2.563	0.01069

```
MINORITYMinority          -383.6564    315.9183   -1.214   0.22521
```

```
Residual standard error: 2705 on 461 degrees of freedom
Multiple R-squared:  0.8471,    Adjusted R-squared:  0.8431
F-statistic: 212.9 on 12 and 461 DF,  p-value: < 2.2e-16
```

(a) Which variables are now significant at the 5% level?

The significant variables are SALBEG, SEX (if male), TIME, WORK, JOBCAT (except for Office).

(b) How much variability in salaries is explained by this model?

The R-squared suggests that the model explains 84.71% of the variability.

(c) Is there evidence for discrimination?

In this model, we can see that SEX (if male) became a significant variable with a positive coefficient. Therefore, without considering age, women may suffer from discrimination.

10. Compare all models that you have built in this home work assignment using the anova function. Briefly summarize your findings.

From the anova tables below, we can see that the sum of the squared residuals decreases as the number of predictors increases, since the variance of the response variables is explained by the predictors. Also, the larger the variance one predictor explains, the more likely the predictor would be statistically significant.

```
anova(mod)
```

Analysis of Variance Table

```
Response: SALNOW
      Df    Sum Sq Mean Sq F value Pr(>F)
Residuals 473 22066639270 46652514
```

```
anova(mod_2)
```

Analysis of Variance Table

```
Response: SALNOW
      Df    Sum Sq Mean Sq F value    Pr(>F)
EDLEVEL  1  9628514842 9628514842  365.38 < 2.2e-16
Residuals 472 12438124428  26351959
```

```
anova(mod_3)
```

Analysis of Variance Table

```
Response: SALNOW
      Df    Sum Sq Mean Sq F value    Pr(>F)
EDLEVEL  1  9628514842 9628514842 402.308 < 2.2e-16
SEX      1 1165593254 1165593254  48.702 0.00000000001016
Residuals 471 11272531174  23933187
```

```
anova(mod_4)
```

Analysis of Variance Table

Response: SALNOW

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SALBEG	1	17092967800	17092967800	2342.7693	< 2.2e-16
SEX	1	64224764	64224764	8.8027	0.003165
TIME	1	208781551	208781551	28.6157	0.0000001394186974
AGE	1	427757745	427757745	58.6287	0.0000000000001132
EDLEVEL	1	133653116	133653116	18.3186	0.0000227529764503
WORK	1	42296045	42296045	5.7971	0.016445
JOBCAT	6	729555926	121592654	16.6655	< 2.2e-16
MINORITY	1	11218146	11218146	1.5376	0.215613
Residuals	460	3356184177	7296053		

```
anova(mod_5)
```

Analysis of Variance Table

Response: SALNOW

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SALBEG	1	17092967800	17092967800	2335.6072	< 2.2e-16
SEX	1	64224764	64224764	8.7758	0.00321
TIME	1	208781551	208781551	28.5282	0.00000014533853
EDLEVEL	1	330141836	330141836	45.1110	0.00000000005497
WORK	1	258483194	258483194	35.3195	0.00000000552480
JOBCAT	6	727452651	121242108	16.5667	< 2.2e-16
MINORITY	1	10793270	10793270	1.4748	0.22521
Residuals	461	3373794205	7318426		