Lecture 26: Model Selection and Regression Diagnostics

- Model selection
- Checking MLR model conditions
- Case study

Introduction

1. Model selection
   1. Model selection criterion depends on goal: significance v.s. prediction
   2. Backward-elimination
   3. Forward-selection
2. Diagnostics for MLR
   1. Checking model conditions using graphs

Recap: Multiple Linear Regression

Recall that the multiple linear regression model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$$

where $\mathbb{E}[\epsilon_i] = 0$, $\mathbb{V}[\epsilon_i] = \sigma^2$, and the ϵ_i's are independent.

The model is useful because:

- it is interpretable—the effect of each explanatory variable is captured by a single coefficient
- theory supports inference and prediction is easy
- simple interactions and transformations are easy (how?)
- dummy variables allow use of categorical information
- computation is fast.

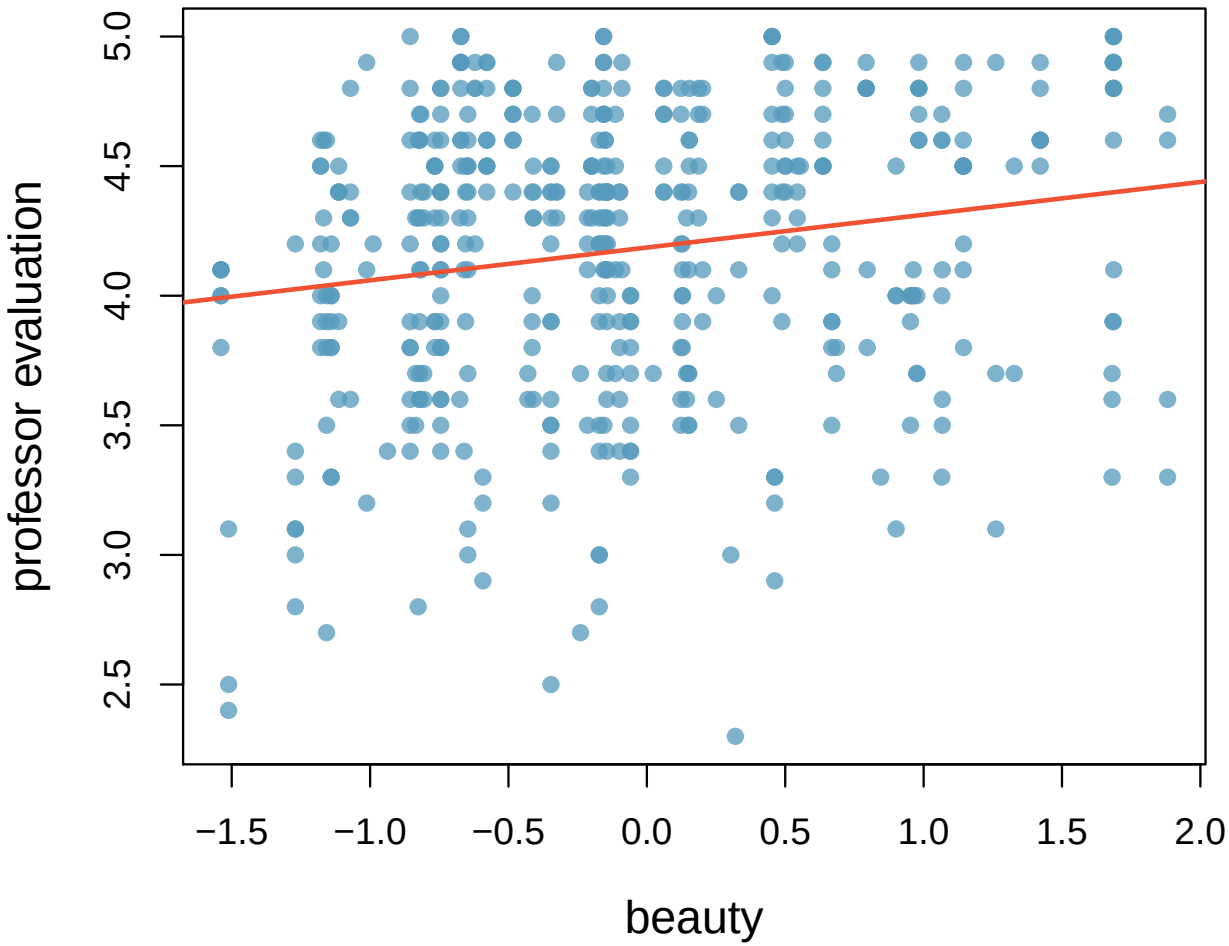
Beauty in the classroom

- Data: Student evaluations of instructors' beauty and teaching quality for 463 courses at the University of Texas. Evaluations conducted at the end of semester.
- The beauty judgements were made later, by six students who had not attended the classes and were not aware of the course evaluations (2 upper level females, 2 upper level males, one lower level female, one lower level male).
- Other potential explanatory variables include Gender, Age, and
  - `formal`: picture wearing tie&jacket/blouse, levels: `yes`, `no`
  - `lower`: lower division course, levels: `yes`, `no`
  - native.non english
  - minority.yes
  - `students`: number of students
  - `tenure`: tenure status, levels: `non-tenure track`, `tenure track`, `tenured`

Hamermesh & Parker. (2004)"Beauty in the Classroom: Professors' Pulchritude and Putative Pedagogical Productivity" Economics Education Review.

Professor rating vs. beauty

Professor evaluation score (higher score means better) vs. beauty score (a score of 0 means average, negative score means below average, and a positive score above average):



Which of the below is correct based on the model output?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.19	0.03	167.24	0.00
beauty	0.13	0.03	4.00	0.00
$R^2 = 0.0336$				

(a) Model predicts 3.36% of professor ratings correctly.

(b) Beauty is not a significant predictor of professor evaluation.

(c) Professors who score 1 point above average in their beauty score are tend to also score 0.13 points higher in their evaluation.

(d) 3.36% of variability in beauty scores can be explained by professor evaluation.

(e) The correlation coefficient could be $\sqrt{0.0336} = 0.18$ or -0.18, we can't tell which is correct.

Which of the below is correct based on the model output?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.19	0.03	167.24	0.00
beauty	0.13	0.03	4.00	0.00
$R^2 = 0.0336$				

(a) Model predicts 3.36% of professor ratings correctly.

(b) Beauty is not a significant predictor of professor evaluation.

(c) *Professors who score 1 point above average in their beauty score are tend to also score 0.13 points higher in their evaluation.*

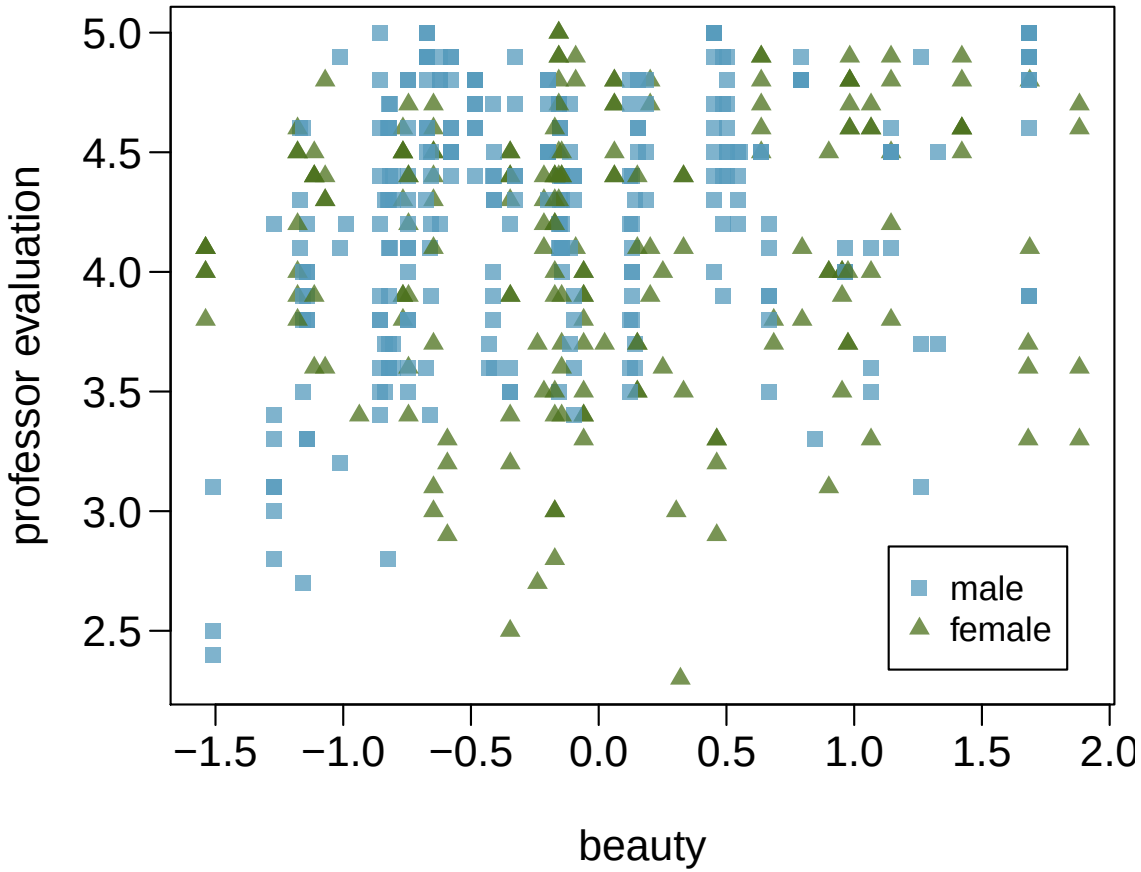
(d) 3.36% of variability in beauty scores can be explained by professor evaluation.

(e) The correlation coefficient could be $\sqrt{0.0336} = 0.18$ or -0.18, we can't tell which is correct.

Exploratory analysis

Any interesting features?

For a given beauty score, are male professors evaluated higher, lower, or about the same as female professors?

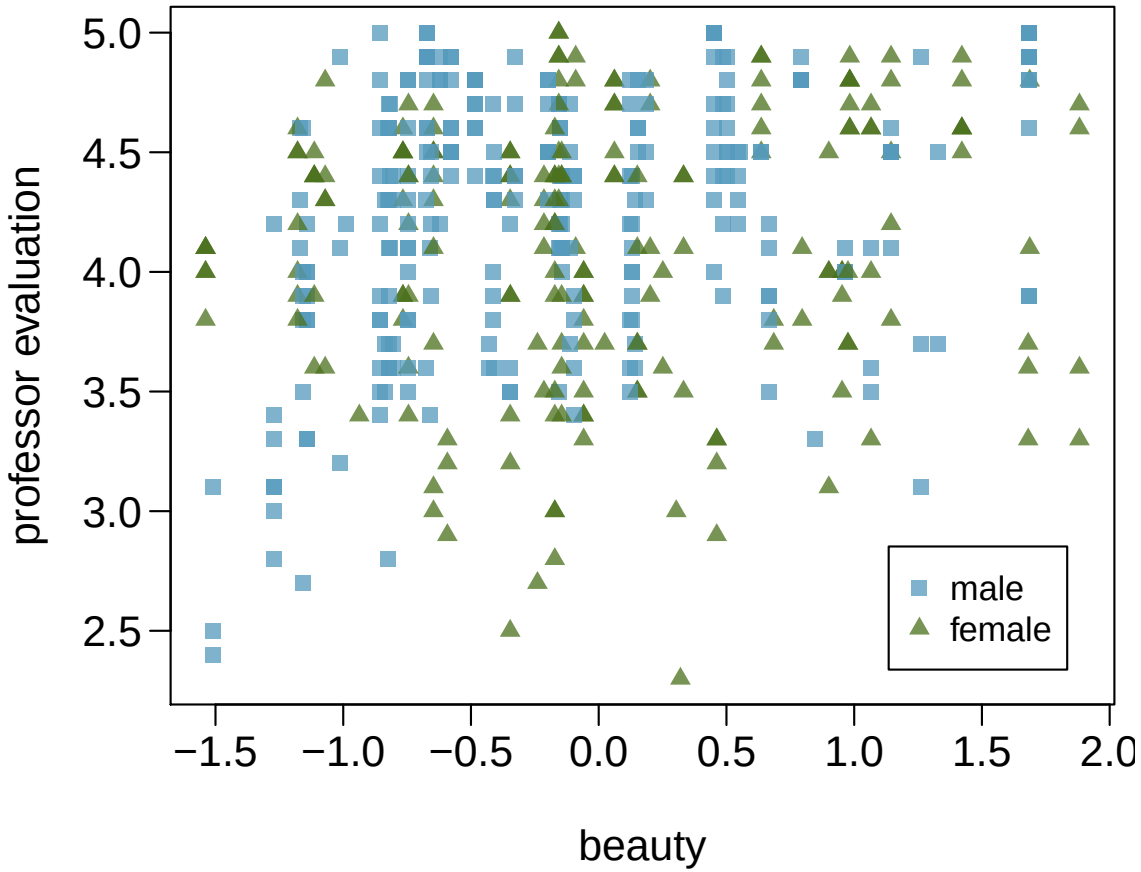


Exploratory analysis

Any interesting features?

Few females with very low beauty scores.

For a given beauty score, are male professors evaluated higher, lower, or about the same as female professors?



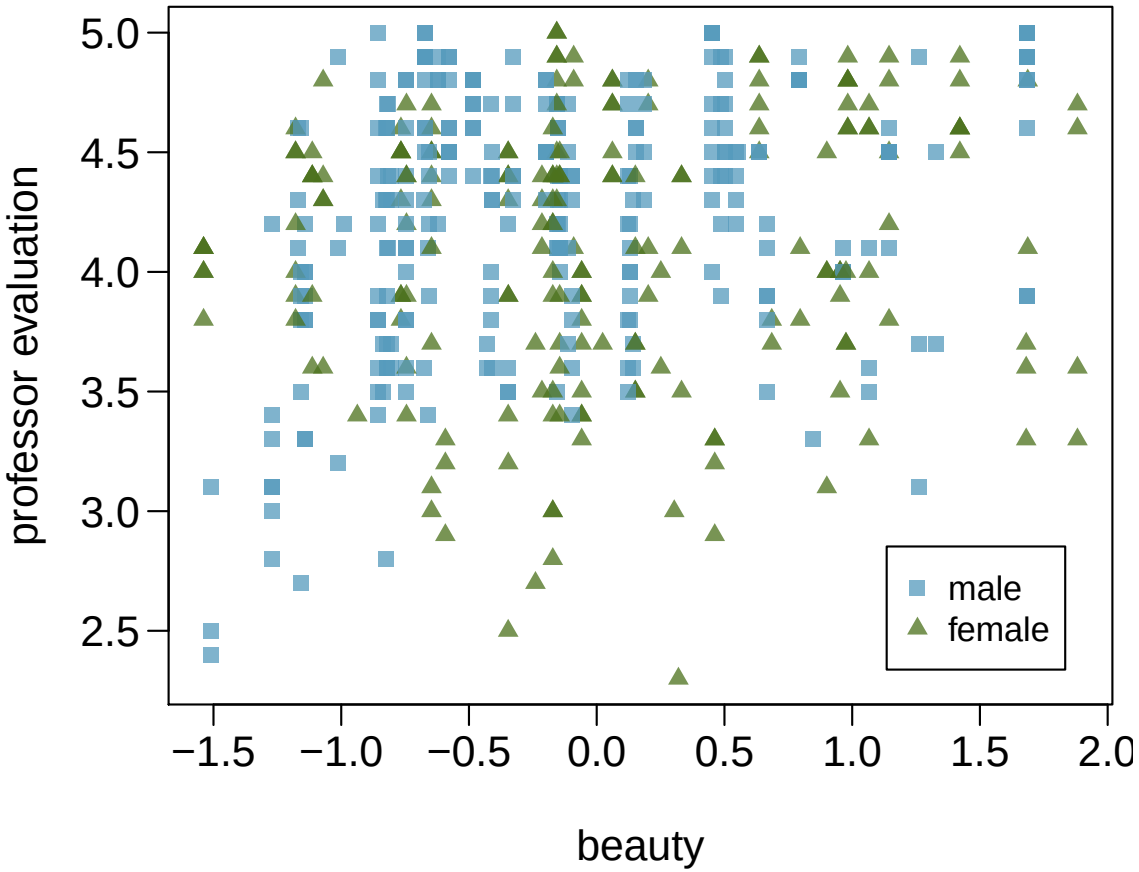
Exploratory analysis

Any interesting features?

Few females with very low beauty scores.

For a given beauty score, are male professors evaluated higher, lower, or about the same as female professors?

Difficult to tell from this plot only.



Professor rating vs. beauty + gender

For a given beauty score, are male professors evaluated higher, lower, or about the same as female professors?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.09	0.04	107.85	0.00
beauty	0.14	0.03	4.44	0.00
gender.male	0.17	0.05	3.38	0.00

$R^2_{adj} = 0.057$

(a) higher

(b) lower

(c) about the same

Professor rating vs. beauty + gender

For a given beauty score, are male professors evaluated higher, lower, or about the same as female professors?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.09	0.04	107.85	0.00
beauty	0.14	0.03	4.44	0.00
gender.male	0.17	0.05	3.38	0.00

$R^2_{adj} = 0.057$

(a) *higher → Beauty held constant, male professors are rated 0.17 points higher on average than female professors.*

(b) lower

(c) about the same

Full model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.6282	0.1720	26.90	0.00
beauty	0.1080	0.0329	3.28	0.00
gender.male	0.2040	0.0528	3.87	0.00
age	-0.0089	0.0032	-2.75	0.01
formal.yes [1]	0.1511	0.0749	2.02	0.04
lower.yes [2]	0.0582	0.0553	1.05	0.29
native.non english	-0.2158	0.1147	-1.88	0.06
minority.yes	-0.0707	0.0763	-0.93	0.35
students [3]	-0.0004	0.0004	-1.03	0.30
tenure.tenure track [4]	-0.1933	0.0847	-2.28	0.02
tenure.tenured	-0.1574	0.0656	-2.40	0.02

[1] `formal`: picture wearing tie&jacket/blouse, levels: `yes`, `no`

[2] `lower`: lower division course, levels: `yes`, `no`

[3] `students`: number of students

[4] `tenure`: tenure status, levels: `non-tenure track`, `tenure track`, `tenured`

Hypotheses

Just as the interpretation of the slope parameters take into account all other variables in the model, the hypotheses for testing for significance of a predictor also takes into account all other variables.

$H_0 : \beta_i = 0$ when other explanatory variables are included in the model.

$H_A : \beta_i \neq 0$ when other explanatory variables are included in the model.

Assessing significance: numerical variables

The p-value for age is 0.01. What does this indicate?

	Estimate	Std. Error	t value	Pr(>|t|)
...				
age	-0.0089	0.0032	-2.75	0.01
...				

(a) Since p-value is positive, higher the professor's age, the higher we would expect them to be rated.

(b) If we keep all other variables in the model, there is strong evidence that professor's age is associated with their rating.

(c) Probability that the true slope parameter for age is 0 is 0.01.

(d) There is about 1% chance that the true slope parameter for age is -0.0089.

Assessing significance: numerical variables

The p-value for age is 0.01. What does this indicate?

	Estimate	Std. Error	t value	Pr(>|t|)
...				
age	-0.0089	0.0032	-2.75	0.01
...				

(a) Since p-value is positive, higher the professor's age, the higher we would expect them to be rated.

(b) *If we keep all other variables in the model, there is strong evidence that professor's age is associated with their rating.*

(c) Probability that the true slope parameter for age is 0 is 0.01.

(d) There is about 1% chance that the true slope parameter for age is -0.0089.

Assessing significance: categorical variables

Tenure is a categorical variable with 3 levels: non tenure track, tenure track, tenured. Based on the model output given, which of the below is false?

	Estimate	Std. Error	t value	Pr(>|t|)
...				
tenure.tenure track	-0.1933	0.0847	-2.28	0.02
tenure.tenured	-0.1574	0.0656	-2.40	0.02

(a) Reference level is non tenure track.

(b) All else being equal, tenure track professors are rated, on average, 0.19 points lower than non-tenure track professors.

(c) All else being equal, tenured professors are rated, on average, 0.16 points lower than non-tenure track professors.

(d) All else being equal, there is a significant difference between the average ratings of tenure track and tenured professors.

Assessing significance: categorical variables

Tenure is a categorical variable with 3 levels: non tenure track, tenure track, tenured. Based on the model output given, which of the below is false?

	Estimate	Std. Error	t value	Pr(>|t|)
...				
tenure.tenure track	-0.1933	0.0847	-2.28	0.02
tenure.tenured	-0.1574	0.0656	-2.40	0.02

(a) Reference level is non tenure track.

(b) All else being equal, tenure track professors are rated, on average, 0.19 points lower than non-tenure track professors.

(c) All else being equal, tenured professors are rated, on average, 0.16 points lower than non-tenure track professors.

(d) *All else being equal, there is a significant difference between the average ratings of tenure track and tenured professors.*

Assessing significance

Which predictors do not seem to meaningfully contribute to the model, i.e. may not be significant predictors of professor's rating score?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.6282	0.1720	26.90	0.00
beauty	0.1080	0.0329	3.28	0.00
gender.male	0.2040	0.0528	3.87	0.00
age	-0.0089	0.0032	-2.75	0.01
formal.yes	0.1511	0.0749	2.02	0.04
lower.yes	0.0582	0.0553	1.05	0.29
native.non english	-0.2158	0.1147	-1.88	0.06
minority.yes	-0.0707	0.0763	-0.93	0.35
students	-0.0004	0.0004	-1.03	0.30
tenure.tenure track	-0.1933	0.0847	-2.28	0.02
tenure.tenured	-0.1574	0.0656	-2.40	0.02

Model selection strategies

Based on what we've learned so far, what are some ways you can think of that can be used to determine which variables to keep in the model and which to leave out?

Backward-elimination

1. R^2_{adj} approach:
   - Start with the full model
   - Drop one variable at a time and record R^2_{adj} of each smaller model
   - Pick the model with the highest increase in R^2_{adj}
   - Repeat until none of the models yield an increase in R^2_{adj}
2. p-value approach:
   - Start with the full model
   - Drop the variable with the highest p-value and refit a smaller model
   - Repeat until all variables left in the model are significant

Backward-elimination: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Full	beauty + gender + age + formal + lower + native + minority + students + tenure	*0.0839*
Step 1	gender + age + formal + lower + native + minority + students + tenure	0.0642
	beauty + age + formal + lower + native + minority + students + tenure	0.0557
	beauty + gender + formal + lower + native + minority + students + tenure	0.0706
	beauty + gender + age + lower + native + minority + students + tenure	0.0777
	beauty + gender + age + formal + native + minority + students + tenure	0.0837
	beauty + gender + age + formal + lower + minority + students + tenure	0.0788
	beauty + gender + age + formal + lower + native + students + tenure	*0.0842*
	beauty + gender + age + formal + lower + native + minority + tenure	0.0838
	beauty + gender + age + formal + lower + native + minority + students	0.0733
Step 2	gender + age + formal + lower + native + students + tenure	0.0647
	beauty + age + formal + lower + native + students + tenure	0.0543
	beauty + gender + formal + lower + native + students + tenure	0.0708
	beauty + gender + age + lower + native + students + tenure	0.0776
	beauty + gender + age + formal + native + students + tenure	*0.0846*
	beauty + gender + age + formal + lower + native + tenure	0.0844
	beauty + gender + age + formal + lower + native + students	0.0725
Step 3	gender + age + formal + native + students + tenure	0.0653
	beauty + age + formal + native + students + tenure	0.0534
	beauty + gender + formal + native + students + tenure	0.0707
	beauty + gender + age + native + students + tenure	0.0786
	beauty + gender + age + formal + students + tenure	0.0756
	beauty + gender + age + formal + native + tenure	*0.0855*
	beauty + gender + age + formal + native + students	0.0713
Step 4	gender + age + formal + native + tenure	0.0667
	beauty + age + formal + native + tenure	0.0553
	beauty + gender + formal + native + tenure	0.0723
	beauty + gender + age + native + tenure	0.0806
	beauty + gender + age + formal + tenure	0.0773
	beauty + gender + age + formal + native	0.0713

`step` function in R

```
Call:
lm(formula = profevaluation ~ beauty + gender + age + formal +
    native + tenure, data = d)

Coefficients:
       (Intercept)              beauty          gendermale
          4.628435            0.105546            0.208079
                 age           formalyes   nativenon english
         -0.008844            0.132422           -0.243003
tenuretenure track       tenuretenured
         -0.206784           -0.175967
```

Best model: beauty + gender + age + formal + native + tenure

Backward-elimination: p-value approach

Step	**Variables included & p-value**									
Full	beauty	gender	age	formal	lower	native	minority	students	tenure	tenure
		male		yes	yes	non english	yes		tenure track	tenured
	0.00	0.00	0.01	0.04	0.29	0.06	*0.35*	0.30	0.02	0.02
Step 1	beauty	gender	age	formal	lower	native		students	tenure	tenure
		male		yes	yes	non english			tenure track	tenured
	0.00	0.00	0.01	0.04	*0.38*	0.03		0.34	0.02	0.01
Step 2	beauty	gender	age	formal		native		students	tenure	tenure
		male		yes		non english			tenure track	tenured
	0.00	0.00	0.01	0.05		0.02		*0.44*	0.01	0.01
Step 3	beauty	gender	age	formal		native			tenure	tenure
		male		yes		non english			tenure track	tenured
	0.00	0.00	0.01	*0.06*		0.02			0.01	0.01
Step 4	beauty	gender	age			native			tenure	tenure
		male				non english			tenure track	tenured
	0.00	0.00	0.01			*0.06*			0.01	0.01
Step 5	beauty	gender	age						tenure	tenure
		male							tenure track	tenured
	0.00	0.00	0.01						0.01	0.01

Best model: beauty + gender + age + tenure

Forward-selection

1. R^2_{adj} approach:
   - Start with regressions of response vs. each explanatory variable
   - Pick the model with the highest R^2_{adj}
   - Add the remaining variables one at a time to the existing model, and once again pick the model with the highest R^2_{adj}
   - Repeat until the addition of any of the remaining variables does not result in a higher R^2_{adj}
2. p-value approach:
   - Start with regressions of response vs. each explanatory variable
   - Pick the variable with the lowest significant p-value
   - Add the remaining variables one at a time to the existing model, and pick the variable with the lowest significant p-value
   - Repeat until any of the remaining variables does not have a significant p-value

In forward-selection the p-value approach isn't any simpler (you still need to fit a bunch of models), so there's almost no incentive to use it.

Selected model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.6284	0.1673	27.66	0.00
beauty	0.1055	0.0328	3.21	0.00
gender.male	0.2081	0.0519	4.01	0.00
age	-0.0088	0.0032	-2.75	0.01
formal.yes	0.1324	0.0714	1.85	0.06
native:non english	-0.2430	0.1080	-2.25	0.02
tenure:tenure track	-0.2068	0.0839	-2.46	0.01
tenure:tenured	-0.1760	0.0641	-2.74	0.01

Modeling conditions

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

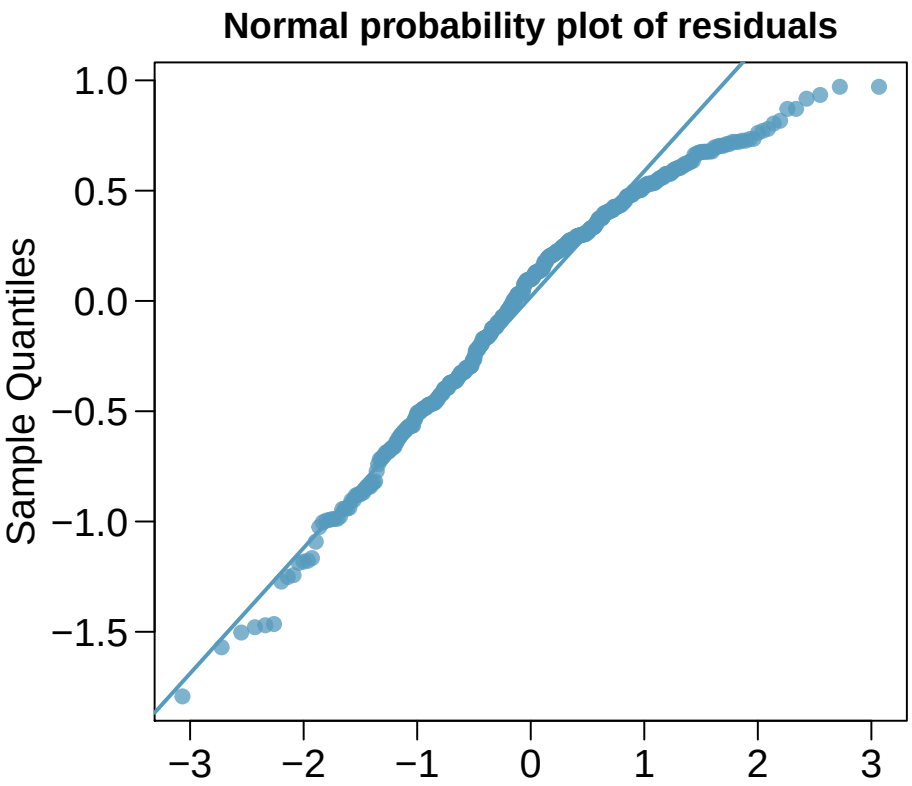
The model depends on the following conditions

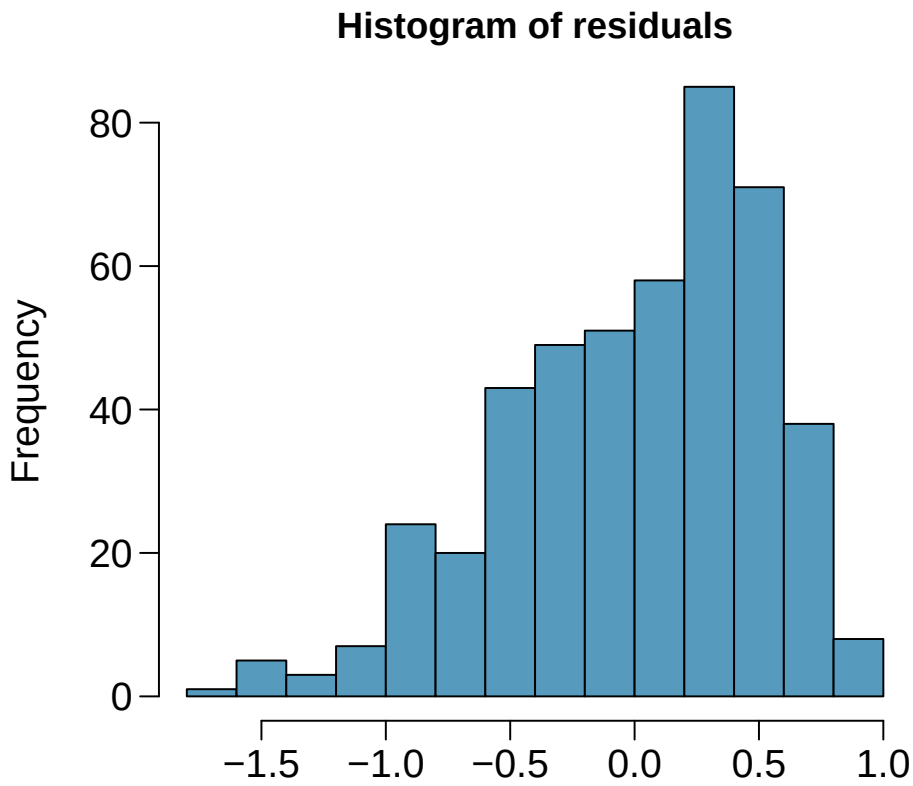
1. Residuals are nearly normal (primary concern relates to residuals that are outliers)
2. Residuals have constant variability
3. Independence of observations (and hence residuals)
4. Each variable is linearly related to the outcome
5. Also important to make sure that your explanatory variables are not collinear

We often use graphical methods to check the validity of these conditions, which we will go through in detail in the following slides.

(1) nearly normal residuals

normal probability plot and/or histogram of residuals:

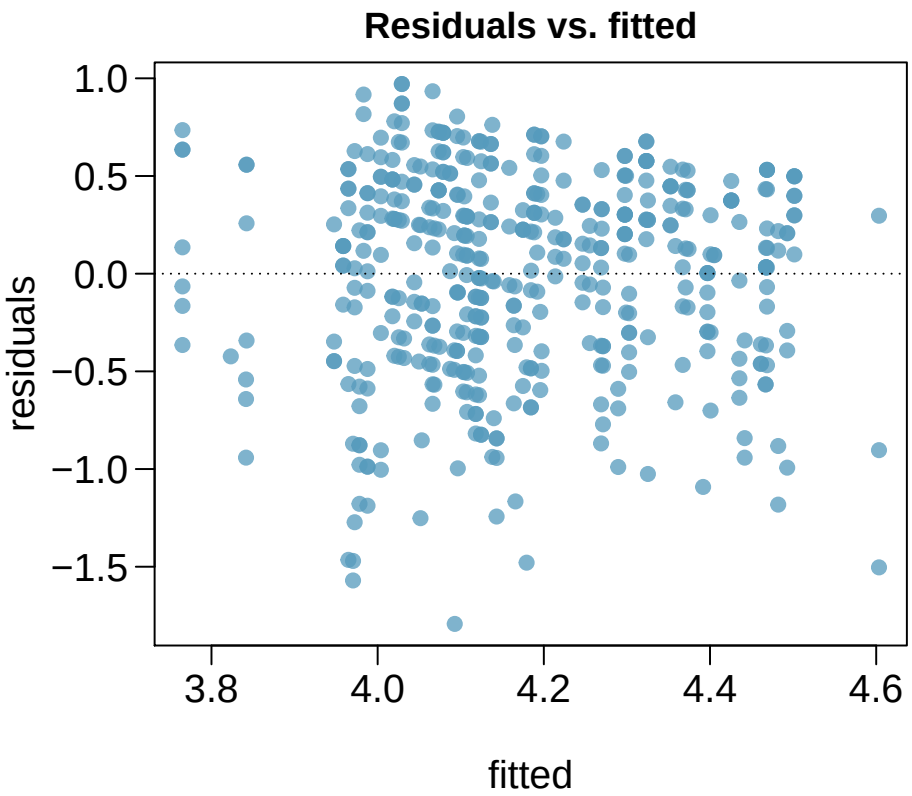


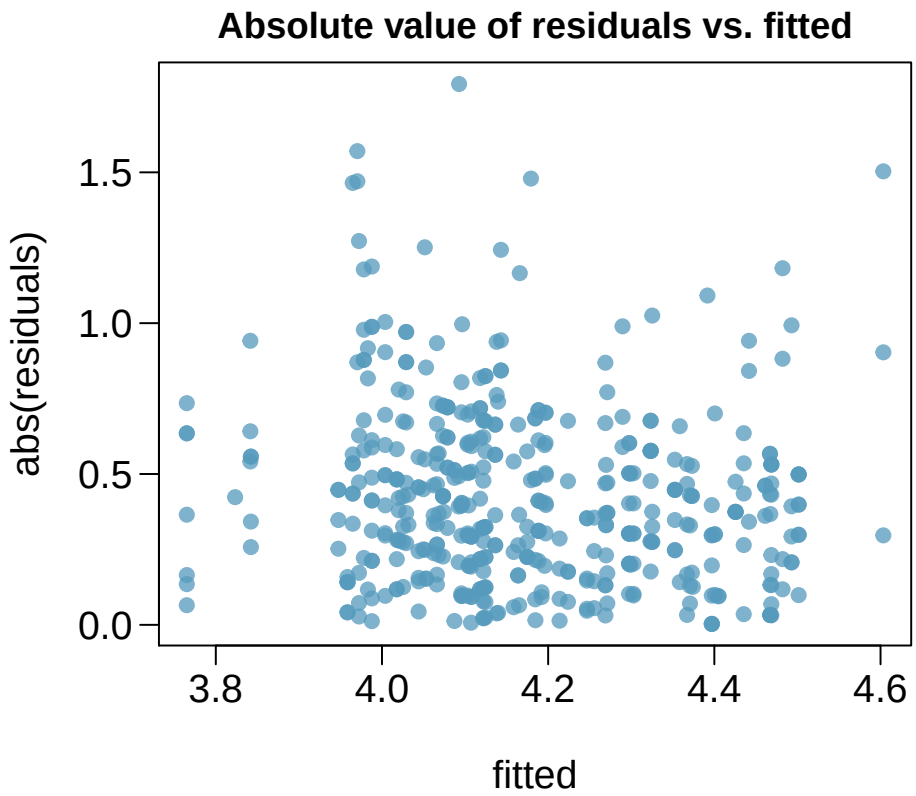


Does this condition appear to be satisfied?

(2) constant variability in residuals

scatterplot of residuals and/or absolute value of residuals vs. fitted (predicted):





Does this condition appear to be satisfied?

Checking constant variance - recap

- When we did simple linear regression (one explanatory variable) we checked the constant variance condition using a plot of *residuals vs. x*.
- With multiple linear regression (2+ explanatory variables) we checked the constant variance condition using a plot of *residuals vs. fitted*.

Why are we using different plots?

Checking constant variance - recap

- When we did simple linear regression (one explanatory variable) we checked the constant variance condition using a plot of *residuals vs. x*.
- With multiple linear regression (2+ explanatory variables) we checked the constant variance condition using a plot of *residuals vs. fitted*.

Why are we using different plots?

In multiple linear regression there are many explanatory variables, so a plot of residuals vs. one of them wouldn't give us the complete picture.

(3) independent residuals

scatterplot of residuals vs. order of data collection:

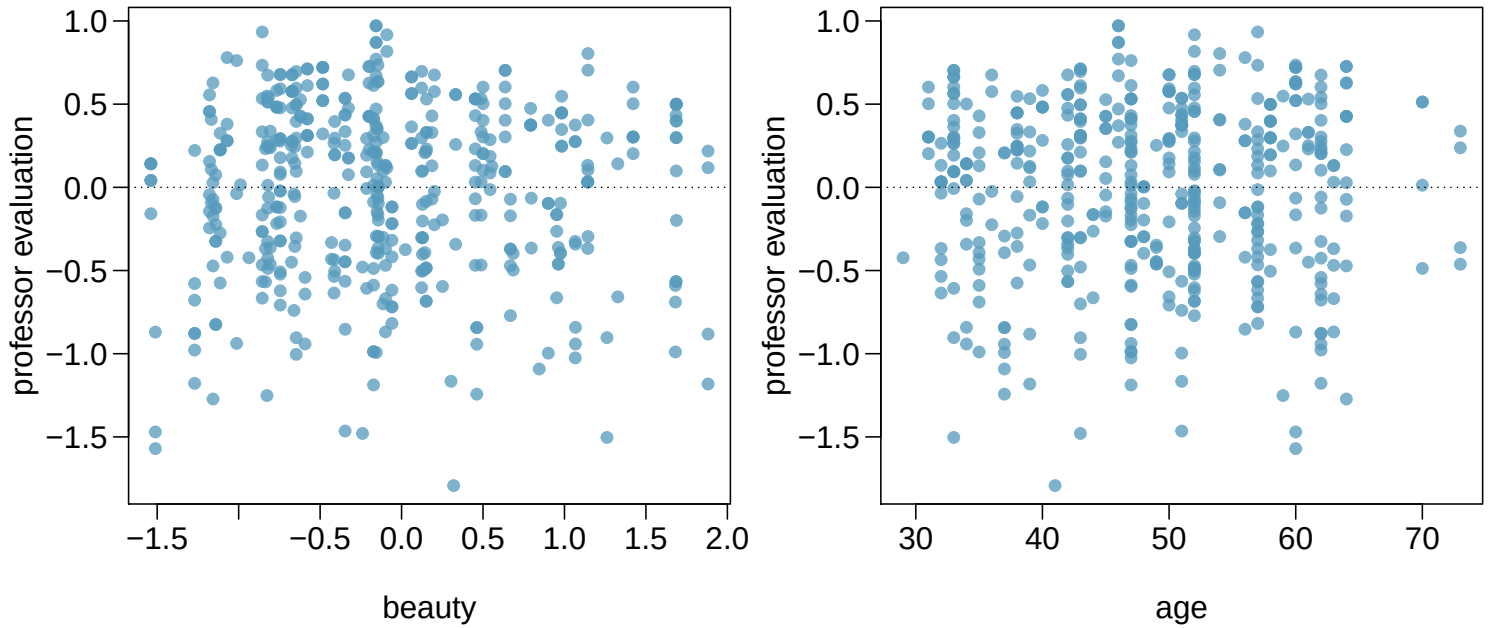
Does this condition appear to be satisfied?

More on the condition of independent residuals

- Checking for independent residuals allows us to indirectly check for independent observations.
- If observations and residuals are independent, we would not expect to see an increasing or decreasing trend in the scatterplot of residuals vs. order of data collection.
- This condition is often violated when we have time series data. Such data require more advanced time series regression techniques for proper analysis.

(4) linear relationships

1. For categorical variable, using boxplot of the residuals against each level to check whether variability fluctuates across levels
2. Using scatterplot of residuals vs. each (numerical) explanatory variable to check if there is some possible structure such as curvature in the residuals.



Does this condition appear to be satisfied?

Practice

Which of the following is the appropriate plot for checking the homoscedasticity condition in MLR

(a) Scatterplot of residuals v.s. $\hat{y}$

(b) Scatterplot of residuals v.s. x

(c) Histogram of residuals

(d) Q-Q plot of residuals

(e) Scatterplot of residuals v.s. order of data collection

Plotting residuals against $\hat{y}$ (predicted, or fitted values of y) allows us to evaluate the whole model as a whole, as opposed to homoscedasticity with regards with regards to just one of the explanatory variables in the model.

Practice

Which of the following is the appropriate plot for checking the homoscedasticity condition in MLR

(a) *Scatterplot of residuals v.s. $\hat{y}$*

(b) Scatterplot of residuals v.s. x

(c) Histogram of residuals

(d) Q-Q plot of residuals

(e) Scatterplot of residuals v.s. order of data collection

Plotting residuals against $\hat{y}$ (predicted, or fitted values of y) allows us to evaluate the whole model as a whole, as opposed to homoscedasticity with regards with regards to just one of the explanatory variables in the model.

Data from ACS

Load data, and subset for those who were employed, if you haven't yet done so.

```
> download.file("http://stat.duke.edu/courses/Spring13/
                sta101.001/labs/acs.RData", destfile = "acs.RData")
> load("acs.RData")
> acs_sub = subset(acs, acs$employment == "employed")
```

1. `income`: Yearly income (wages and salaries) —> *response variable*
2. `employment`: Employment status, not in labor force, unemployed, or employed
3. `hrs_work`: Weekly hours worked
4. `race`: Race, White, Black, Asian, or other
5. `age`: Age
6. `gender`: gender, male or female
7. `citizens`: Whether respondent is a US citizen or not
8. `time_to_work`: Travel time to work
9. `lang`: Language spoken at home, English or other
10. `married`: Whether respondent is married or not
11. `edu`: Education level, hs or lower, college, or grad
12. `disability`: Whether respondent is disabled or not
13. `birth_qrtr`: Quarter in which respondent is born, jan thru mar, apr thru jun, jul thru sep, or oct thru dec

From Lab of STA101, Spring 2013, Prof. Mine Cetinkaya-Rundel

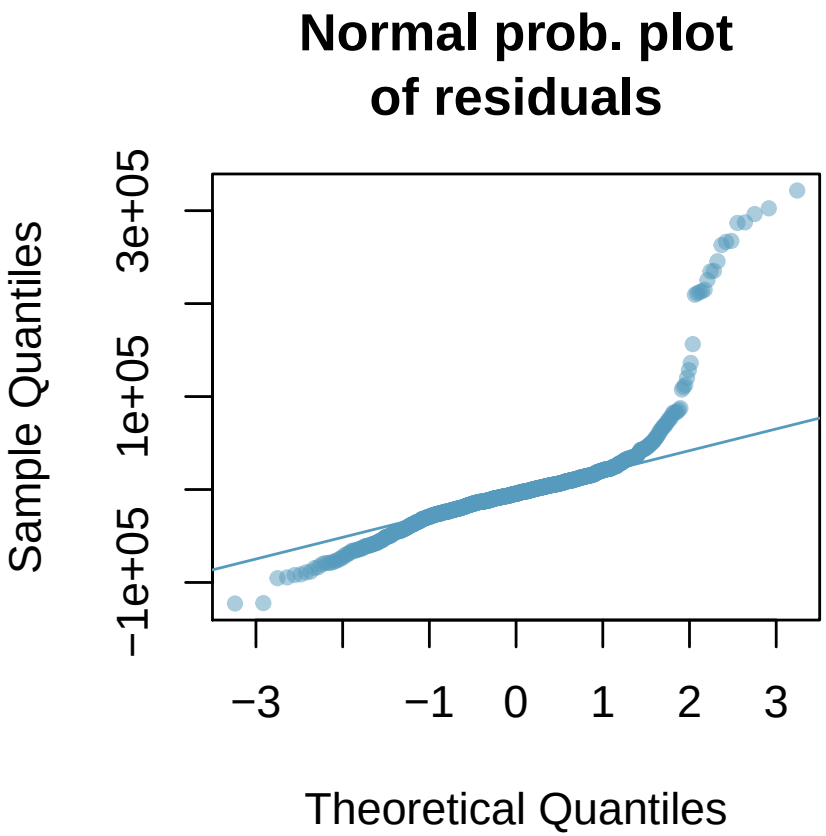
Predicting income

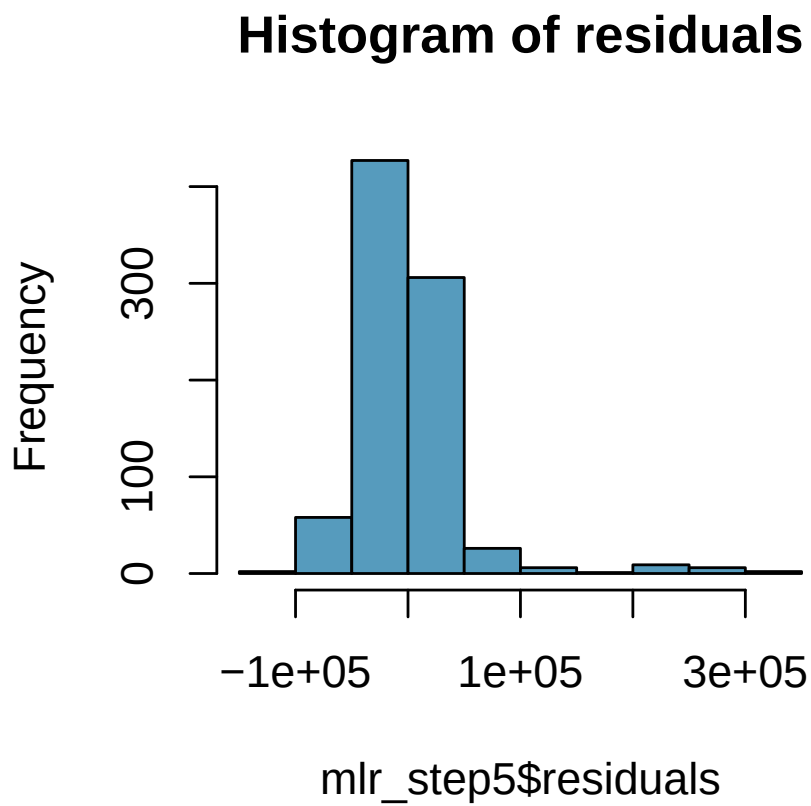
```
> mlr_step5 = lm(income ~ hrs_work + race + age + gender + edu +
                 disability, data = acs_sub)
> summary(mlr_step5)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-21737.27	7719.33	-2.82	0.00
hrs_work	1000.05	135.53	7.38	0.00
race:black	-6015.53	5877.30	-1.02	0.31
race:asian	29595.59	8029.98	3.69	0.00
race:other	-8599.21	6648.63	-1.29	0.20
age	561.57	118.88	4.72	0.00
gender:female	-18120.56	3495.87	-5.18	0.00
edu:college	17273.85	3827.51	4.51	0.00
edu:grad	58551.90	5418.84	10.81	0.00
disability:yes	-15851.98	6209.49	-2.55	0.01

(1) Nearly normal residuals with mean 0

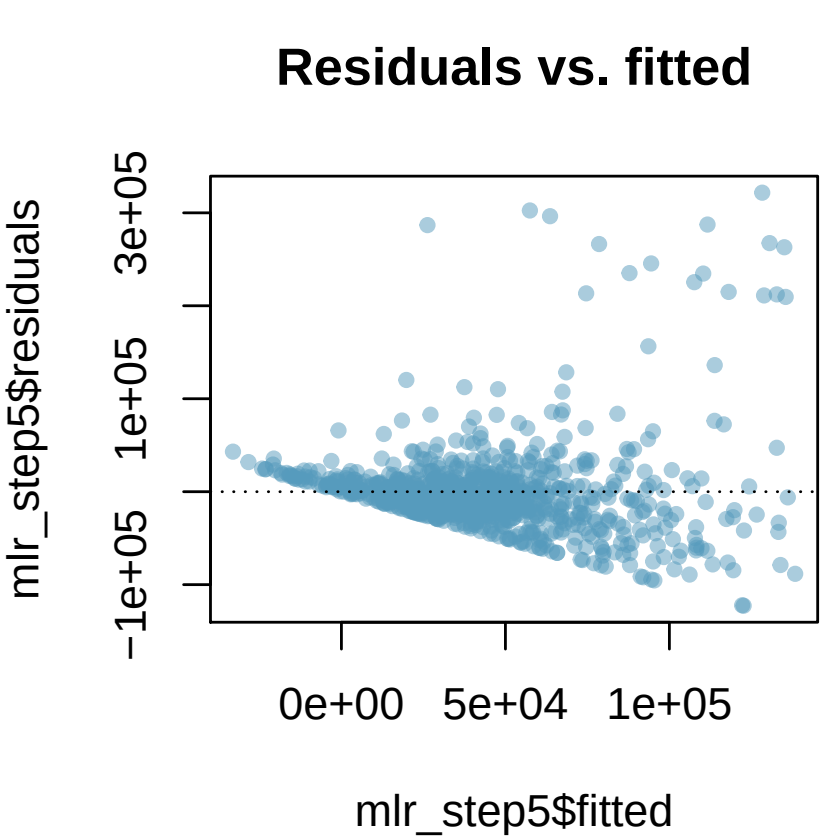
```
> qqnorm(mlr_step5$residuals, main = "Normal prob. plot\nof residuals")
> qqline(mlr_step5$residuals)
> hist(mlr_step5$residuals, main = "Histogram of residuals")
```

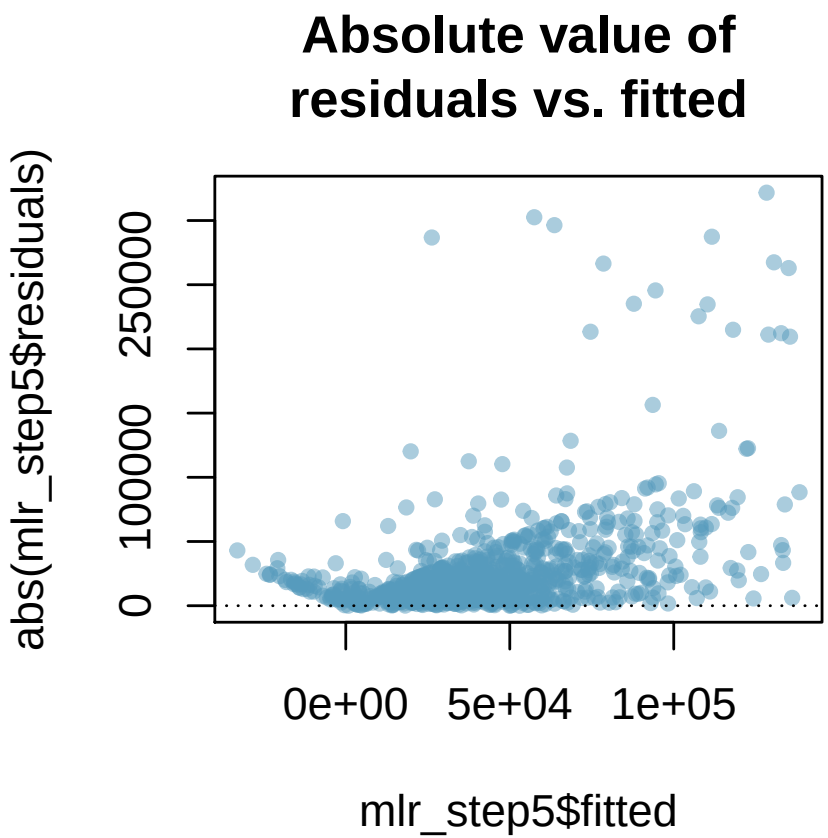




(2) Constant variability of residuals

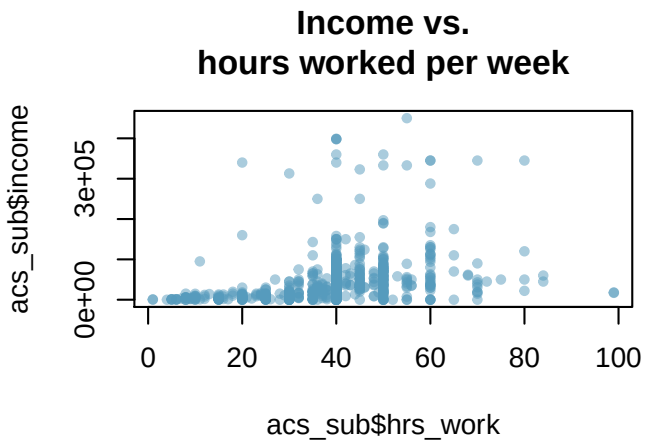
```
> plot(mlr_step5$residuals ~ mlr_step5$fitted,
          main = "Residuals vs. fitted")
> abline(h = 0, lty = 3)
> plot(abs(mlr_step5$residuals) ~ mlr_step5$fitted,
          main = "Absolute value of\nresiduals vs. fitted")
> abline(h = 0, lty = 3)
```

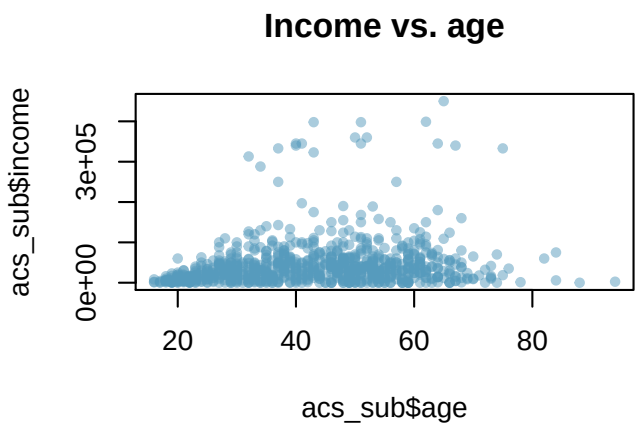


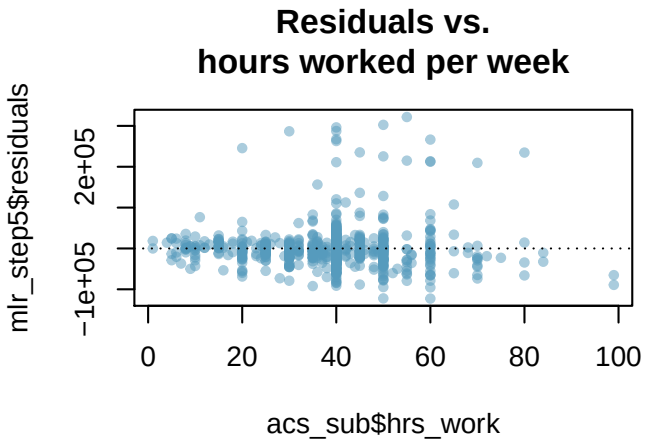


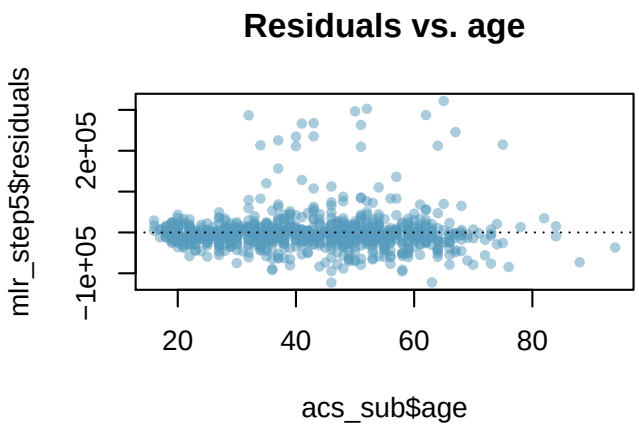
(3) Each (numerical) variable linearly related to outcome

```
> par(mfrow = c(2,2)) # 4 plots in one window, 2 rows, 2 columns
> plot(acs_sub$income ~ acs_sub$hrs_work, main = "Income vs.\nhours worked per week")
> plot(mlr_step5$residuals ~ acs_sub$hrs_work, main = "Residuals vs.\nhours worked per week")
> abline(h = 0, lty = 3)
> plot(acs_sub$income ~ acs_sub$age, main = "Income vs. age")
> plot(mlr_step5$residuals ~ acs_sub$age, main = "Residuals vs. age")
> abline(h = 0, lty = 3)
```





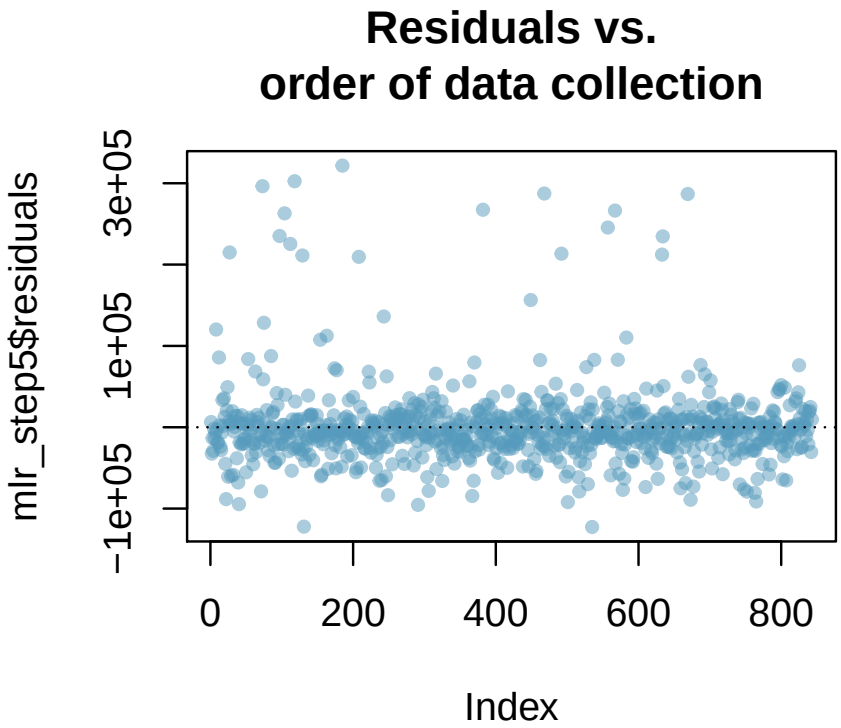




(4) Independence

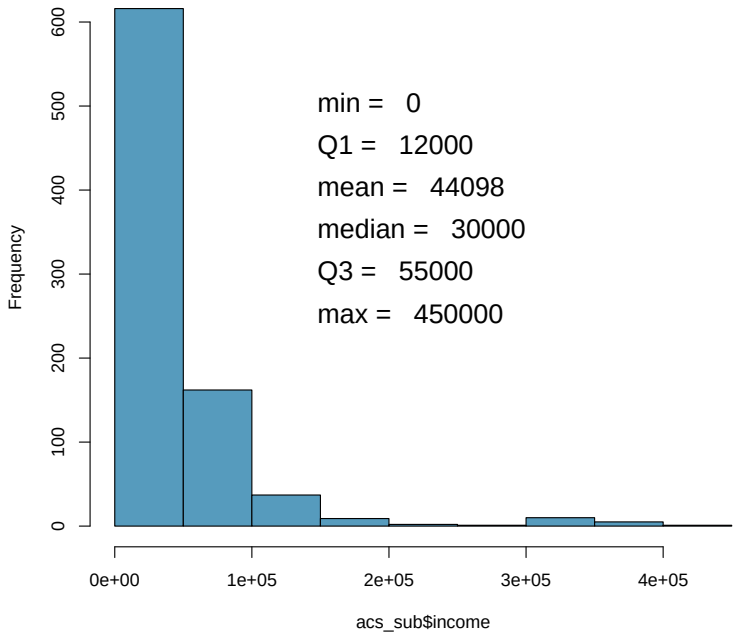
We know that the data are sampled randomly, so there should be no pattern in residuals with respect to the order of data collection.

```
> plot(mlr_step5$residuals,
           main = "Residuals vs.\norder of data collection")
```



Transformations

- We saw that residuals have a right-skewed distribution, and the relationship between hours worked per week and income is non-linear (exponential).
- In these situations a transformation applied to the response variable may be useful.
- In order to decide which transformation to use, we should examine the distribution of the response variable.



- The extremely right skewed distribution suggests that a log transformation may be useful.

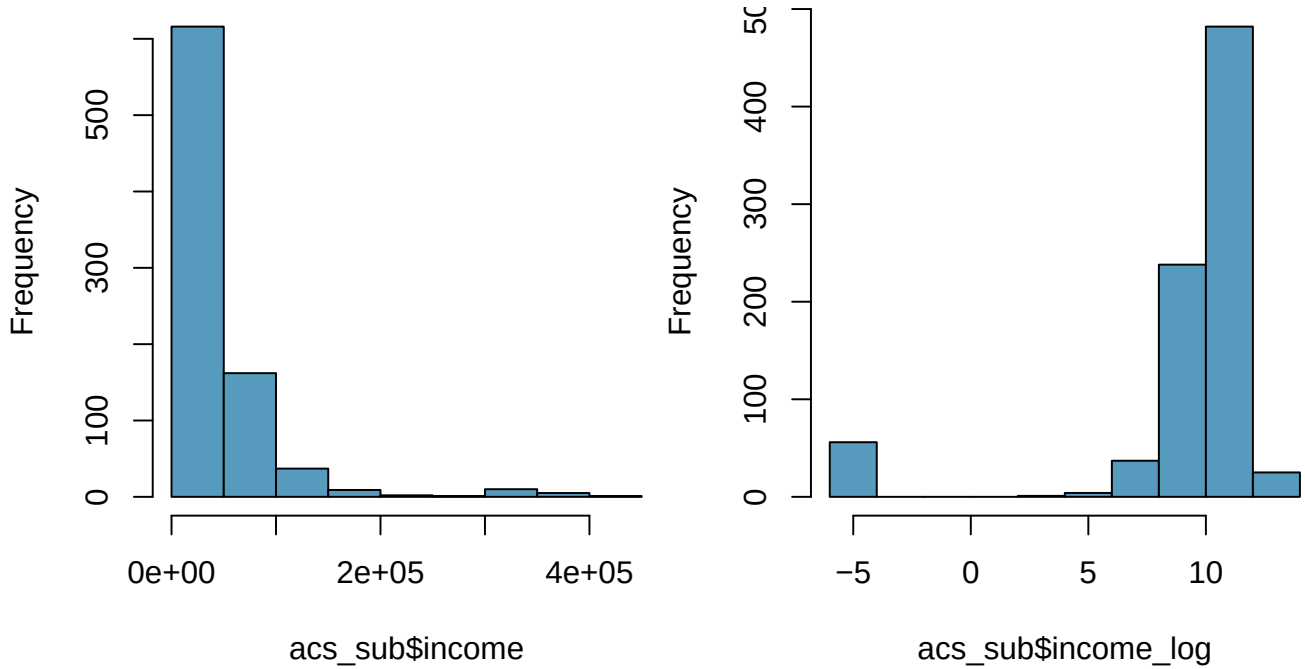
Log of 0

```
> summary(acs_sub$income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0   12000   30000   44100   55000  450000
> log(0)
[1] -Inf
```

- Since there are some individuals who made 0 income (from salaries and wages) last year, we cannot take the log of their income, since log(0) is undefined.
- A commonly used trick is to add a very small number to all values, and then take the log.

Logged distribution

```
> acs_sub$income_log = log(acs_sub$income + 0.01)
> hist(acs_sub$income)
> hist(acs_sub$income_log)
```

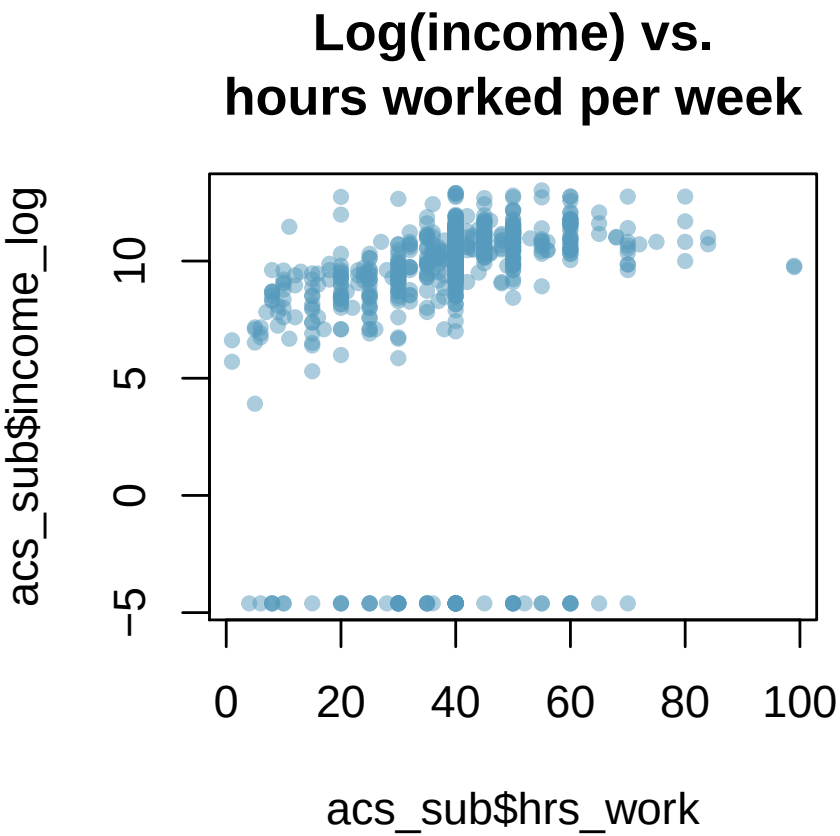


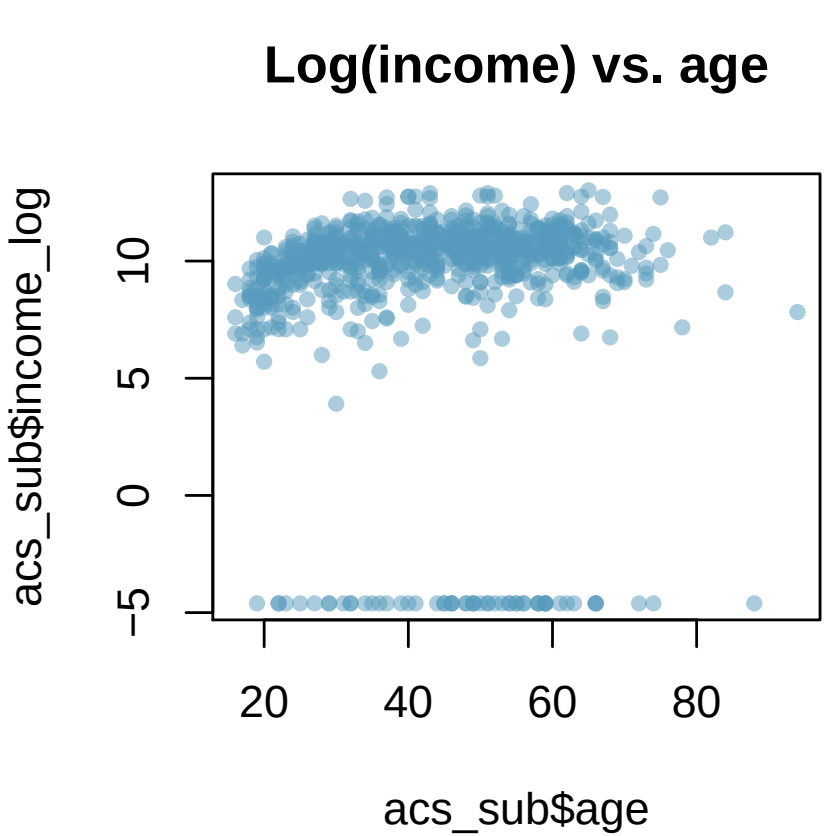
Are there any interesting features in the distribution of the logged income?

The transformation revealed a group of points (people with 0 income) that are unlike the rest of the data.

Logged relationships

```
> plot(acs_sub$income_log ~ acs_sub$hrs_work)
> plot(acs_sub$income_log ~ acs_sub$age)
```





We still might want to do something about those 0 incomes, it doesn't make sense to model them with the rest of the data.

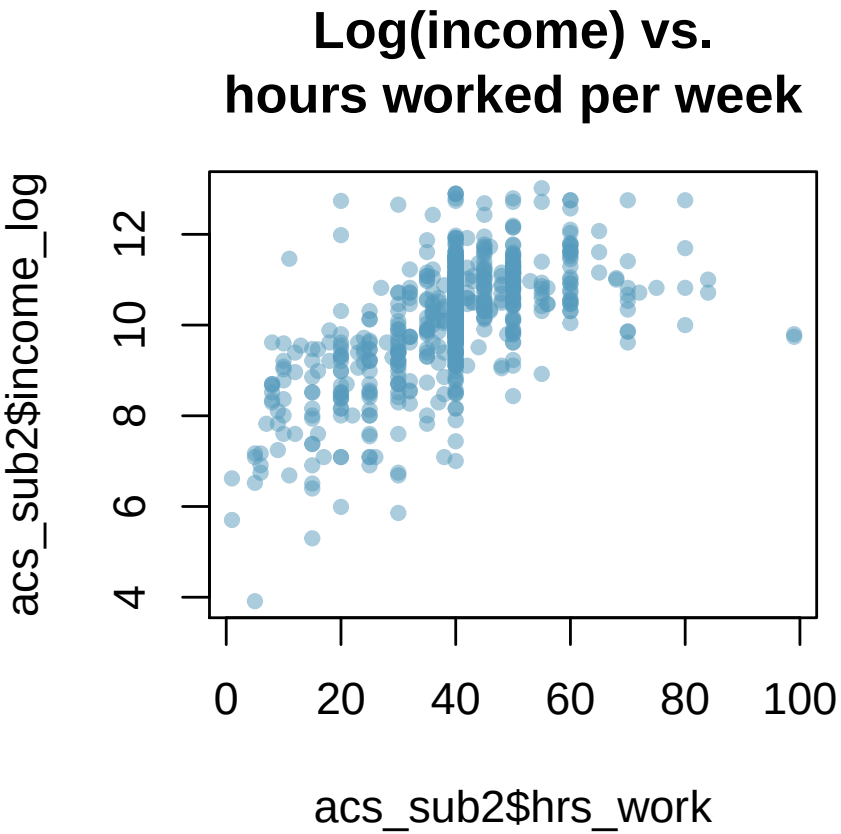
Further subsetting the data

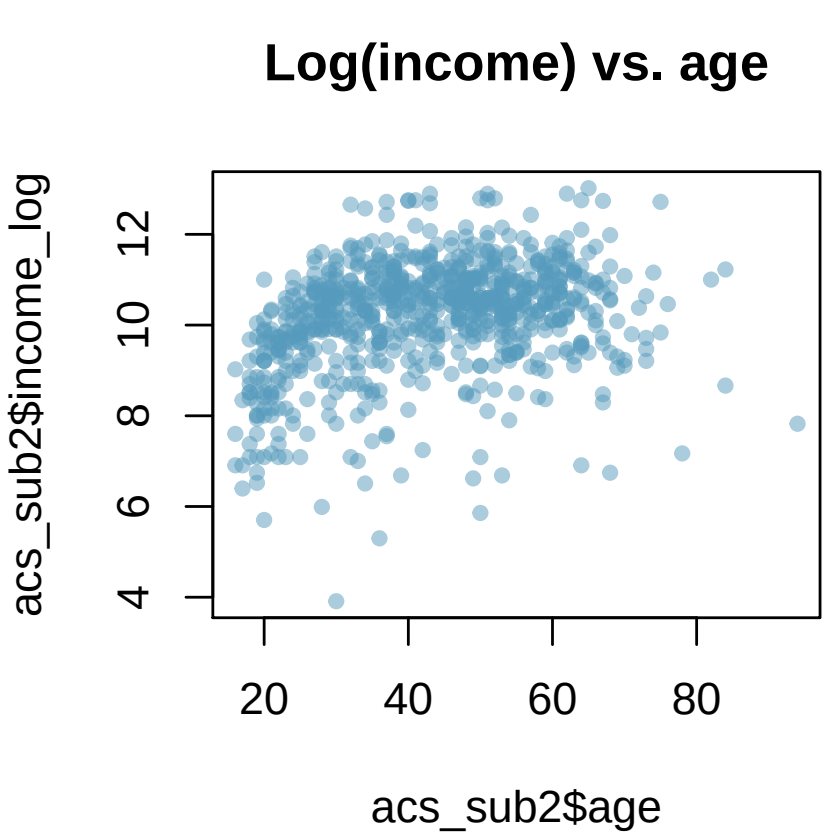
People who work more than 0 hours per week but make 0 income in salaries and wages are different than others whose income is proportional to number of hours they work. So we have reason to omit these people from the analysis (and model their income differently based on other variables).

```
> acs_sub2 = subset(acs_sub, acs_sub$income > 0)
> acs_sub2$income_log = log(acs_sub2$income)
```

Logged relationships - for those with any income

```
> plot(acs_sub2$income_log ~ acs_sub2$hrs_work)
> plot(acs_sub2$income_log ~ acs_sub2$age)
```





Much better...

Final model for log of income

... after a new model selection process (backwards elimination, p-value method):

```
> mlr_log_fin = lm(income_log ~ hrs_work + age + gender + time_to_work +
                   married + edu + disability, data = acs_sub2)
> summary(mlr_log_fin)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.16	0.14	50.46	0.00
hrs_work	0.05	0.00	18.59	0.00
age	0.02	0.00	9.24	0.00
gender:female	-0.26	0.06	-4.00	0.00
time_to_work	0.00	0.00	2.13	0.03
married:yes	0.18	0.07	2.66	0.01
edu:college	0.35	0.07	5.03	0.00
edu:grad	0.86	0.10	8.76	0.00
disability:yes	-0.59	0.12	-5.02	0.00

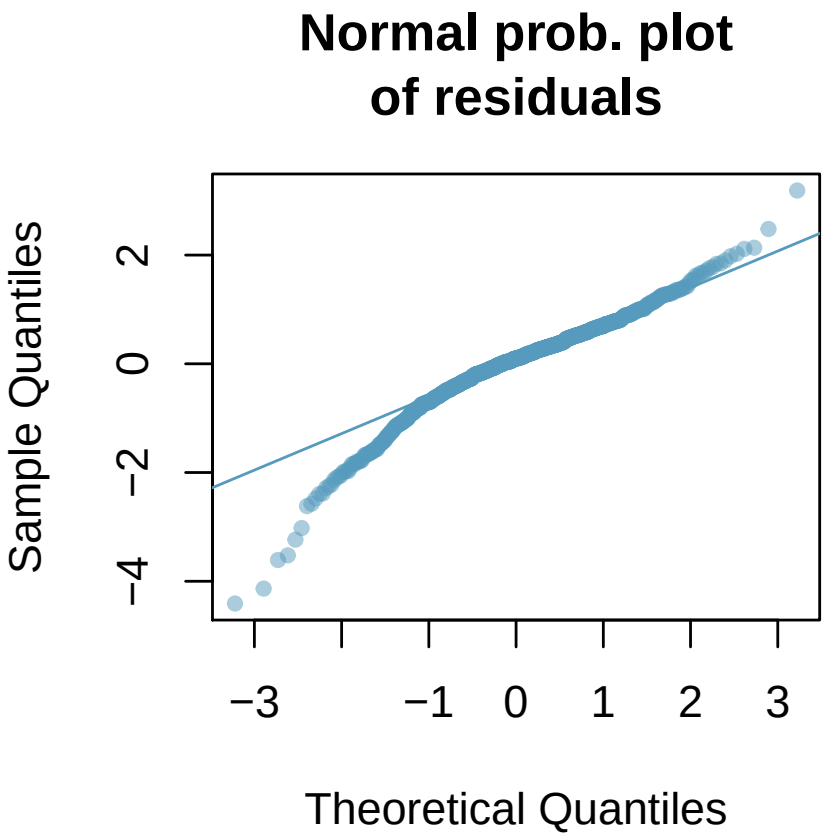
Application exercise:
Diagnostics for model for logged income model

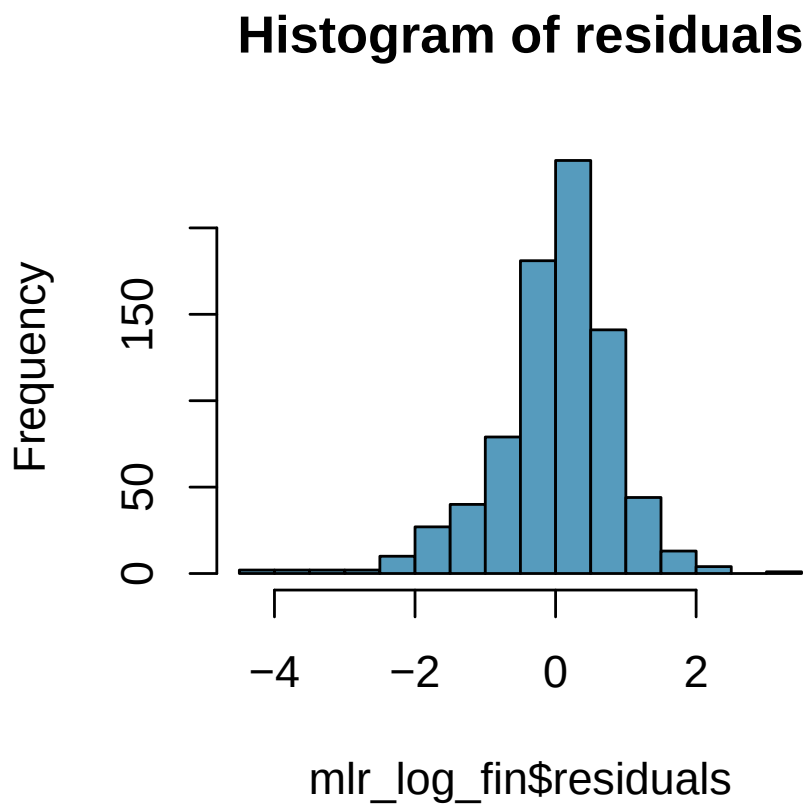
Check the nearly normal residuals with mean 0 and constant variability of residuals conditions for the logged income model using appropriate diagnostics plots.

```
> mlr_log_fin = lm(income_log ~ hrs_work + age + gender + time_to_work +
                   married + edu + disability, data = acs_sub2)
```

(1) Nearly normal residuals with mean 0

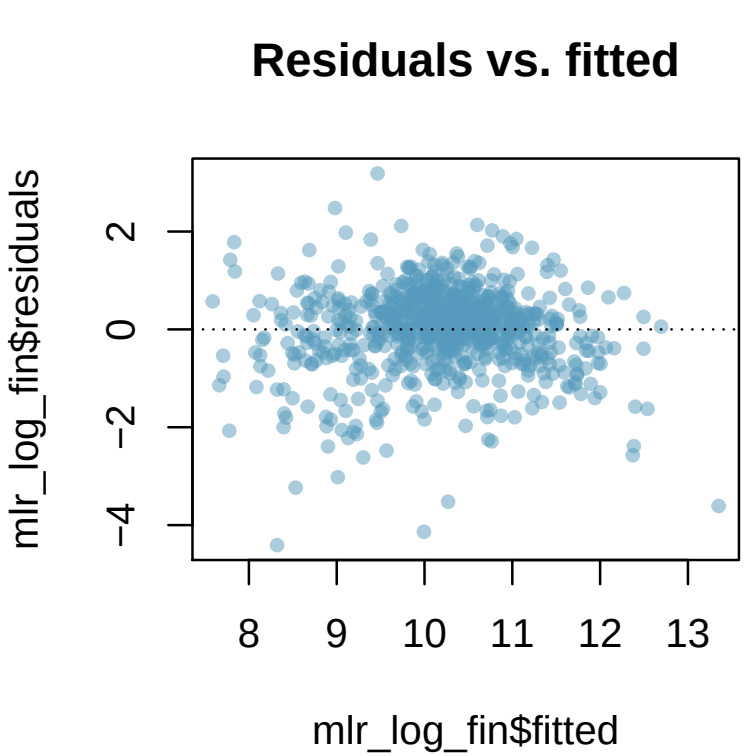
```
> qqnorm(mlr_log_fin$residuals, main = "Normal prob. plot\nof residuals")
> qqline(mlr_log_fin$residuals)
> hist(mlr_log_fin$residuals, main = "Histogram of residuals")
```

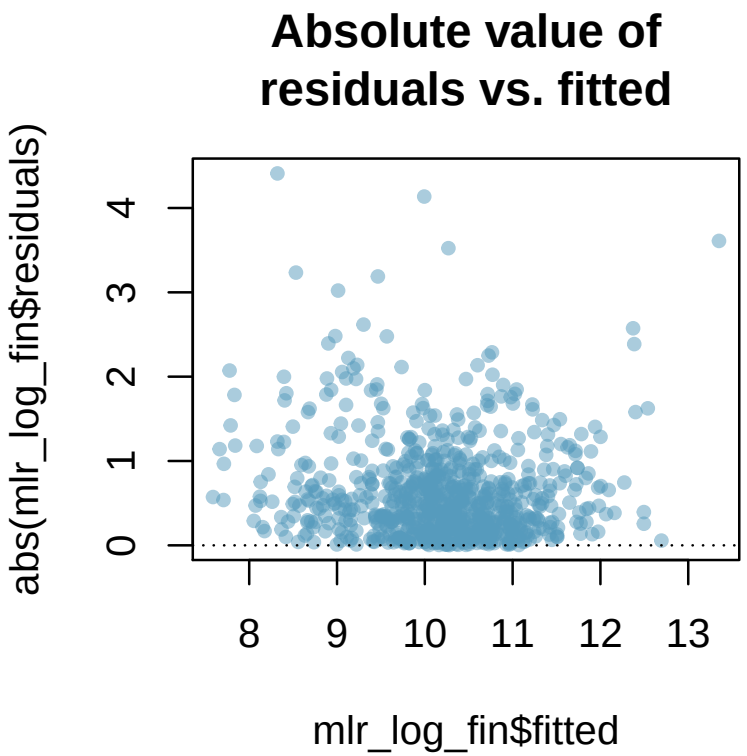




(2) Constant variability of residuals

```
> plot(mlr_log_fin$residuals ~ mlr_log_fin$fitted,
         main = "Residuals vs. fitted")
> abline(h = 0, lty = 3)
> plot(abs(mlr_log_fin$residuals) ~ mlr_log_fin$fitted,
         main = "Absolute value of\nresiduals vs. fitted")
> abline(h = 0, lty = 3)
```





Application exercise:
Interpreting coefficients of logged models (1)

Which is the correct interpretation of the slope of hours worked per week?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.16	0.14	50.46	0.00
hrs_work	0.05	0.00	18.59	0.00
age	0.02	0.00	9.24	0.00
gender:female	-0.26	0.06	-4.00	0.00
time_to_work	0.00	0.00	2.13	0.03
married:yes	0.18	0.07	2.66	0.01
edu:college	0.35	0.07	5.03	0.00
edu:grad	0.86	0.10	8.76	0.00
disability:yes	-0.59	0.12	-5.02	0.00

For each additional hour worked per week,

(a) we would expect income to increase on average by \$0.05.

(b) we would expect income to increase on average by 0.05%.

(c) we would expect income to increase on average by 5.12%.

(d) we would expect income to increase on average by \$18.59

(e) we would expect income to increase on average by a factor of 18.59.

Application exercise:
Interpreting coefficients of logged models (1)

Which is the correct interpretation of the slope of hours worked per week?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.16	0.14	50.46	0.00
hrs_work	0.05	0.00	18.59	0.00
age	0.02	0.00	9.24	0.00
gender:female	-0.26	0.06	-4.00	0.00
time_to_work	0.00	0.00	2.13	0.03
married:yes	0.18	0.07	2.66	0.01
edu:college	0.35	0.07	5.03	0.00
edu:grad	0.86	0.10	8.76	0.00
disability:yes	-0.59	0.12	-5.02	0.00

For each additional hour worked per week,

(a) we would expect income to increase on average by \$0.05.

(b) we would expect income to increase on average by 0.05%.

(c) *we would expect income to increase on average by 5.12%.* *exp(0.05) = 1.0512*

(d) we would expect income to increase on average by \$18.59

(e) we would expect income to increase on average by a factor of 18.59.

Application exercise:
Interpreting coefficients of logged models (2)

Which is the correct interpretation of the slope of gender:female?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.16	0.14	50.46	0.00
hrs_work	0.05	0.00	18.59	0.00
age	0.02	0.00	9.24	0.00
gender:female	-0.26	0.06	-4.00	0.00
time_to_work	0.00	0.00	2.13	0.03
married:yes	0.18	0.07	2.66	0.01
edu:college	0.35	0.07	5.03	0.00
edu:grad	0.86	0.10	8.76	0.00
disability:yes	-0.59	0.12	-5.02	0.00

The model predicts that females, on average, make

(a) \$26,000 less than males.

(b) 23% less than males.

(c) 77% more than males.

(d) \$2,600 less than males.

(e) \$2,600 less more males.

Application exercise:
Interpreting coefficients of logged models (2)

Which is the correct interpretation of the slope of gender:female?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.16	0.14	50.46	0.00
hrs_work	0.05	0.00	18.59	0.00
age	0.02	0.00	9.24	0.00
gender:female	-0.26	0.06	-4.00	0.00
time_to_work	0.00	0.00	2.13	0.03
married:yes	0.18	0.07	2.66	0.01
edu:college	0.35	0.07	5.03	0.00
edu:grad	0.86	0.10	8.76	0.00
disability:yes	-0.59	0.12	-5.02	0.00

The model predicts that females, on average, make

(a) \$26,000 less than males.

(b) *23% less than males.* *exp(-0.26) = 0.77*

(c) 77% more than males.

(d) \$2,600 less than males.

(e) \$2,600 less more males.

Recap

Today we learned about model selection and model diagnostics in multiple linear regression.

Suggested reading:

- OpenIntro3: Sec. 8.2, 8.3