Lecture 20 - Model Selection

Sta102 / BME102

Colin Rundel

November 17, 2014

Modeling kid's test scores (revisited)

Predicting cognitive test scores of three- and four-year-old children using characteristics of their mothers. Data are from a survey of adult American women and their children - a subsample from the National Longitudinal Survey of Youth.

	kid_score	mom_hs	mom_iq	mom_work	mom_age
1	65	yes	121.12	yes	27
⋮	⋮	⋮	⋮	⋮	⋮
5	115	yes	92.75	yes	27
6	98	no	107.90	no	18
⋮	⋮	⋮	⋮	⋮	⋮
434	70	yes	91.25	yes	25

Gelman, Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. (2007) Cambridge University Press.

Model output

```
cog_full = lm(kid_score ~ mom_hs + mom_iq + mom_work + mom_age,
              data = cognitive)
summary(cog_full)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.59241    9.21906   2.125   0.0341
## mom_hsyes    5.09482    2.31450   2.201   0.0282
## mom_iq       0.56147    0.06064   9.259   <2e-16
## mom_workyes  2.53718    2.35067   1.079   0.2810
## mom_age      0.21802    0.33074   0.659   0.5101
##
## Residual standard error: 18.14 on 429 degrees of freedom
## Multiple R-squared: 0.2171, Adjusted R-squared: 0.2098
## F-statistic: 29.74 on 4 and 429 DF,  p-value: < 2.2e-16
```


Backward-elimination

- Adjusted R^2 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 largest increase in R^2_{adj}
  - Repeat until none of the reduced models yield an increase in R^2_{adj}
- When removing a categorical variable all levels should be included or removed (may not be clear what to do with the p-value approach)

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

Step	Variables included	R^2_{adj}
Full	kid_score ~ mom_hs + mom_iq + mom_work + mom_age	*0.2098*
Step 1	kid_score ~ mom_iq + mom_work + mom_age	0.2027
	kid_score ~ mom_hs + mom_work + mom_age	0.0541
	kid_score ~ mom_hs + mom_iq + mom_age	0.2095
	kid_score ~ mom_hs + mom_iq + mom_work	*0.2109*
Step 2	kid_score ~ mom_iq + mom_work	0.2024
	kid_score ~ mom_hs + mom_work	0.0546
	kid_score ~ mom_hs + mom_iq	*0.2105*
Step 3*	kid_score ~ mom_hs	0.2024
	kid_score ~ mom_iq	0.0546

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

Step	Variables included	R^2_{adj}
Full	kid_score ~ mom_hs + mom_iq + mom_work + mom_age	0.2098
Step 1	kid_score ~ mom_iq + mom_work + mom_age	0.2027
	kid_score ~ mom_hs + mom_work + mom_age	0.0541
	kid_score ~ mom_hs + mom_iq + mom_age	0.2095
	kid_score ~ mom_hs + mom_iq + mom_work	0.2109
Step 2	kid_score ~ mom_iq + mom_work	0.2024
	kid_score ~ mom_hs + mom_work	0.0546
	kid_score ~ mom_hs + mom_iq	0.2105
Step 3*	kid_score ~ mom_hs	0.2024
	kid_score ~ mom_iq	0.0546

Forward-selection

1. Adjusted R^2 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 remanning variables does not result in a higher R^2_{adj}

Forward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Step 1	kid_score ~ mom_hs	0.0539
	kid_score ~ mom_work	0.0097
	kid_score ~ mom_age	0.0062
	kid_score ~ mom_iq	*0.1991*
Step 2	kid_score ~ mom_iq + mom_work	0.2024
	kid_score ~ mom_iq + mom_age	0.1999
	kid_score ~ mom_iq + mom_hs	*0.2105*
Step 3	kid_score ~ mom_iq + mom_hs + mom_age	0.2095
	kid_score ~ mom_iq + mom_hs + mom_work	*0.2109*
Step 4*	kid_score ~ mom_iq + mom_hs + mom_age + mom_work	0.2098

Forward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Step 1	kid_score ~ mom_hs	0.0539
	kid_score ~ mom_work	0.0097
	kid_score ~ mom_age	0.0062
	kid_score ~ mom_iq	0.1991
Step 2	kid_score ~ mom_iq + mom_work	0.2024
	kid_score ~ mom_iq + mom_age	0.1999
	kid_score ~ mom_iq + mom_hs	0.2105
Step 3	kid_score ~ mom_iq + mom_hs + mom_age	0.2095
	kid_score ~ mom_iq + mom_hs + mom_work	0.2109
Step 4*	kid_score ~ mom_iq + mom_hs + mom_age + mom_work	0.2098

Expert opinion as criterion for model selection

In addition to the quantitative approaches we discussed, variables can be included in (or eliminated from) the model based on expert opinion.

Final model choice

```
cog_final = lm(kid_score ~ mom_hs + mom_iq, data = kid)
summary(cog_final)

## Call:
## lm(formula = kid_score ~ mom_hs + mom_iq, data = kid)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.73154    5.87521   4.380 1.49e-05 ***
## mom_hsyes    5.95012    2.21181   2.690  0.00742 **
## mom_iq       0.56391    0.06057   9.309  < 2e-16 ***
##
## Residual standard error: 18.14 on 431 degrees of freedom
## Multiple R-squared: 0.2141, Adjusted R-squared: 0.2105
## F-statistic: 58.72 on 2 and 431 DF,  p-value: < 2.2e-16
```


Conditions for MLR

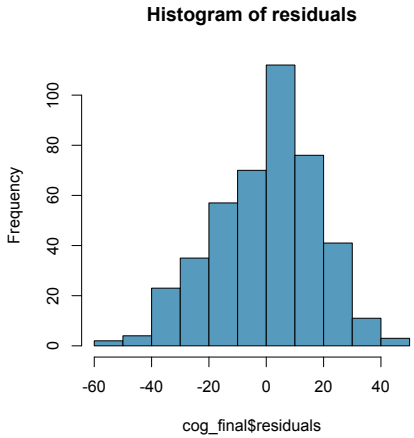
In order to perform inference for multiple regression we require the following conditions:

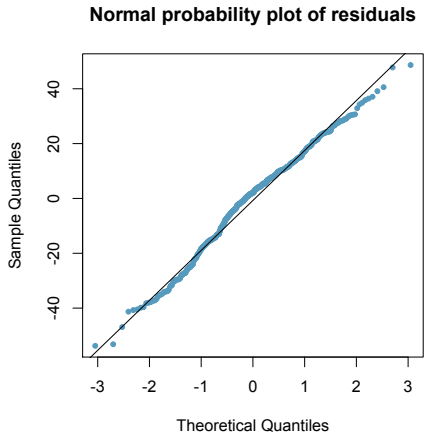
(1) Nearly normal residuals

(2) Constant variability of residuals

(3) Independent residuals

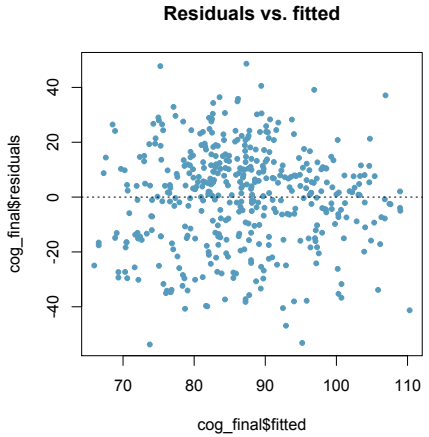
Nearly normal residuals

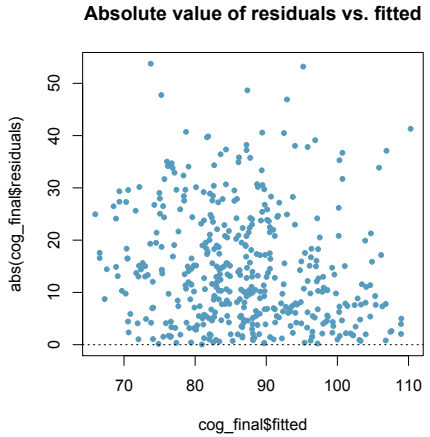




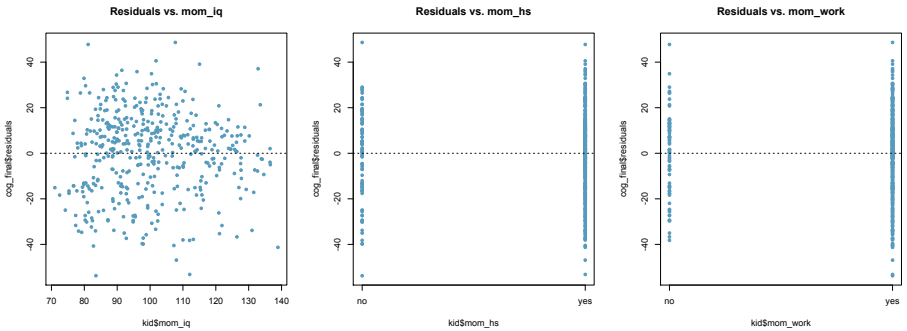
Constant variability of residuals

Why do we use the fitted (predicted) values in MLR?



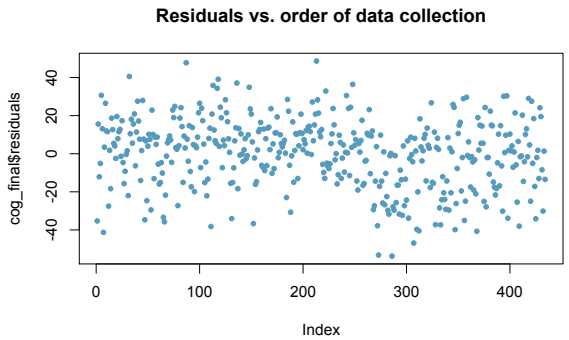


Constant variability of residuals (cont.)



Independent residuals

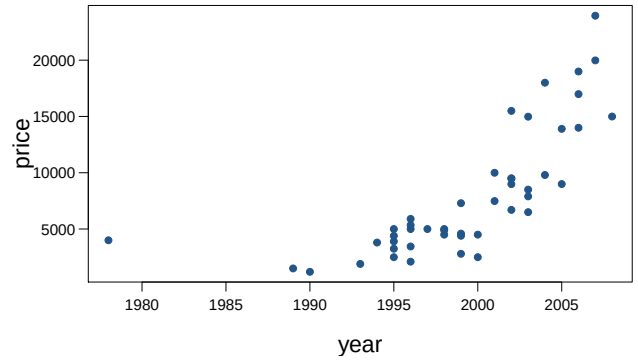
- If we suspect that order of data collection may influence the outcome (mostly in time series data):



- If not, think about how data are sampled.

Truck prices

The scatterplot below shows the relationship between year and price of a random sample of 43 pickup trucks. Describe the relationship between these two variables.

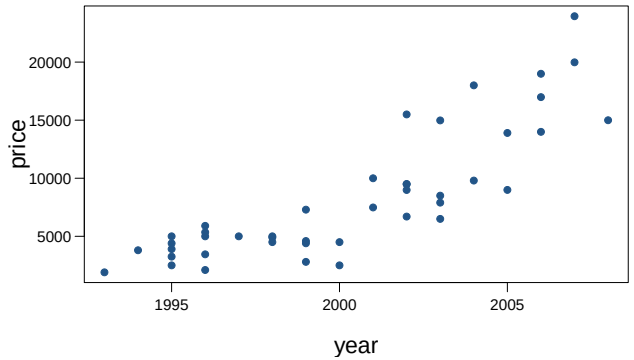


From: *http://faculty.chicagobooth.edu/robert.gramacy/teaching.html*

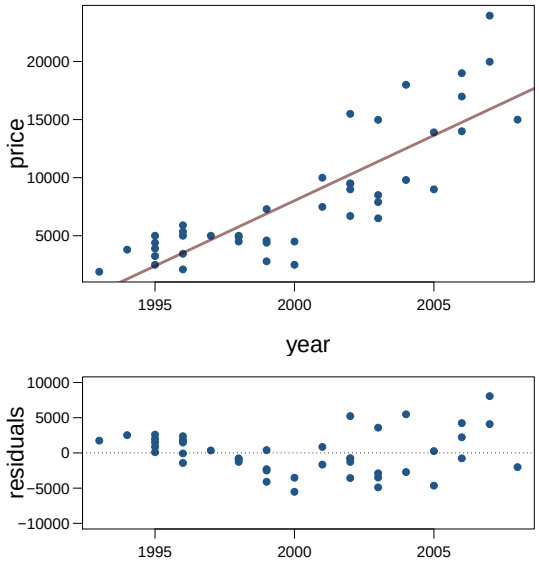
Remove unusual observations

Let's remove trucks older than 20 years, and only focus on trucks made in 1992 or later.

Now what can you say about the relationship?



Truck prices - linear model?



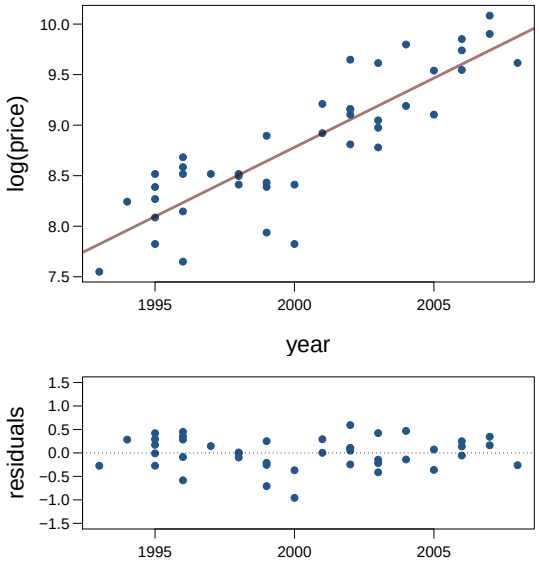
Model:

$$\widehat{price} = b_0 + b_1 \ year$$

The linear model doesn't appear to be a good fit since the residuals have non-constant variance.

In particular residuals for newer cars (to the right) have a larger variance than the residuals for older cars (to the left).

Truck prices - log transform of the response variable



Model:

$$\widehat{log(price)} = b_0 + b_1 \ year$$

We have applied a log transformation to the response variable. The relationship now seems linear, and the residuals have (more) constant variance.

Interpreting models with log transformation

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-265.07	25.04	-10.59	0.00
pu$year	0.14	0.01	10.94	0.00

$$\textit{Model: } \widehat{log(price)} = -265.07 + 0.14\ year$$

- For each additional year the car is newer (for each year decrease in car's age) we would expect the log price of the car to increase on average by 0.14 log dollars.
- which is not very useful ...

Working with logs

- Subtraction and logs:

$$log(a) - log(b) = log(\frac{a}{b})$$

- Natural logarithm:

$$e^{log(x)} = x$$

- We can use these identities to "undo" the log transformation

Interpreting models with log transformation (cont.)

The slope coefficient for the log transformed model is 0.14, meaning the log price difference between cars that are one year apart is predicted to be 0.14 log dollars.

$$\log(\text{price 1}) = -265.07 + 0.14\ y$$
$$\log(\text{price 2}) = -265.07 + 0.14\ (y+1)$$

$$\begin{aligned}
\log(\text{price 2}) - \log(\text{price 1}) &= 0.14 \\
log\left(\frac{\text{price 2}}{\text{price 1}}\right) &= 0.14 \\
e^{log\left(\frac{\text{price 2}}{\text{price 1}}\right)} &= e^{0.14} \\
\frac{\text{price 2}}{\text{price 1}} &= 1.15
\end{aligned}$$

For each additional year the car is newer (for each year decrease in car's age) we would expect the price of the car to increase on average *by a factor of 1.15*.

Recap: dealing with non-constant variance

- Non-constant variance is one of the most common model violations, however it is usually fixable by transforming the response (y) variable
- The most common variance stabilizing transform is the log transformation: $log(y)$, especially useful when the response variable is (extremely) right skewed.
- When using a log transformation on the response variable the interpretation of the slope changes:
  - For each unit increase in x, y is expected on average to decrease/increase by a factor of e^{b_1}.
- Another useful transformation is the square root: $\sqrt{y}$, especially useful when the response variable is counts.
- These transformations may also be useful when the relationship is non-linear, but in those cases a polynomial regression may also be needed (this is beyond the scope of this course, but you're welcomed to try it for your project, and I'd be happy to provide further guidance)

From last lab

Just like in lab we load data, and subset for those who were employed.

```
download.file("http://stat.duke.edu/~cr173/Sta102_Fa14/Lab/acs.RData",
              destfile = "acs.RData", method="wget")
load("acs.RData")
acs_sub = subset(acs, acs$employment == "employed")
```


Predicting income

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


##
## Call:
## lm(formula = income ~ hrs_work + race + age + gender + edu +
##     disability, data = acs_sub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -122650  -20503   -4597   10945  321681
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -21737.3     7719.3  -2.816 0.004978 **
## hrs_work        1000.1      135.5   7.379 3.86e-13 ***
## raceblack      -6015.5     5877.3  -1.024 0.306359
## raceasian      29595.6     8030.0   3.686 0.000243 ***
## raceother      -8599.2     6648.6  -1.293 0.196238
## age              561.6      118.9   4.724 2.71e-06 ***
## genderfemale  -18120.6     3495.9  -5.183 2.74e-07 ***
## educollege     17273.8     3827.5   4.513 7.31e-06 ***
## edugrad        58551.9     5418.8  10.805  < 2e-16 ***
## disabilityyes -15852.0     6209.5  -2.553 0.010861 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48270 on 833 degrees of freedom
## Multiple R-squared:  0.2935, Adjusted R-squared:  0.2859
## F-statistic: 38.46 on 9 and 833 DF,  p-value: < 2.2e-16
```


Categorical variables with multiple levels

In model selection based on R^2_{adj}:

> Leave all levels in or drop the entire variable (even if one level is significant).

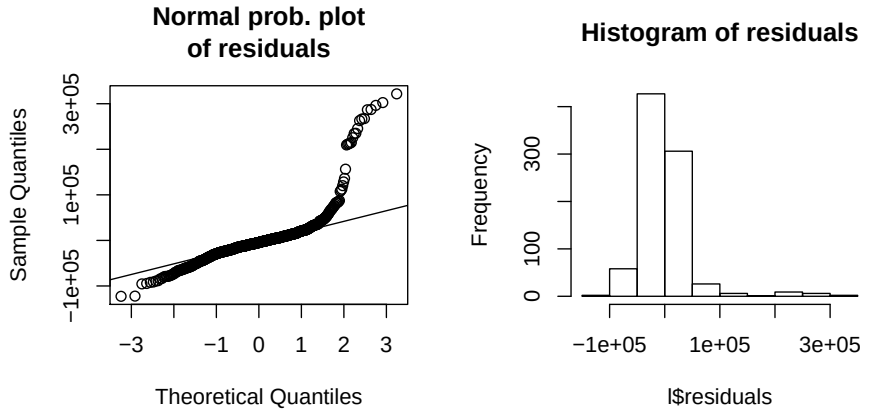
For example, the race variable in our model:

	Estimate	Std. Error	t value	Pr(>|t|)
...				
raceblack	-6015.53	5877.30	-1.02	0.31
raceasian	29595.59	8029.98	3.69	0.00
raceother	-8599.21	6648.63	-1.29	0.20
...				

How do we interpret the slopes associated with the race variable?

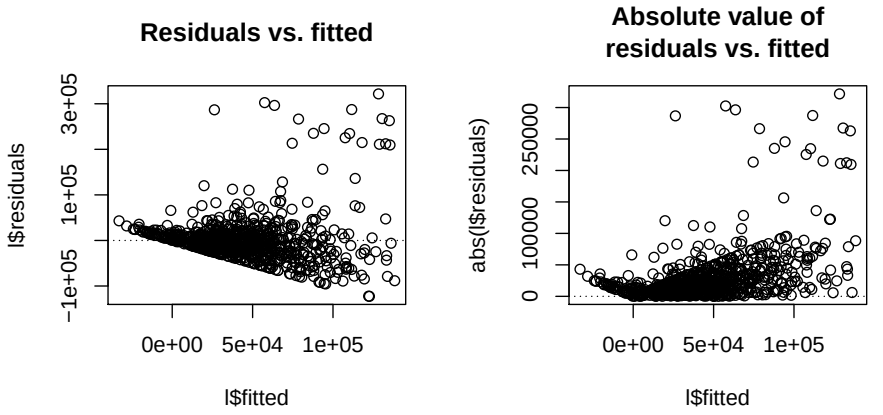
(1) Nearly normal residuals

```
par(mfrow=c(1,2))
qqnorm(l$residuals, main = "Normal prob. plot\nof residuals")
qqline(l$residuals)
hist(l$residuals, main = "Histogram of residuals")
```



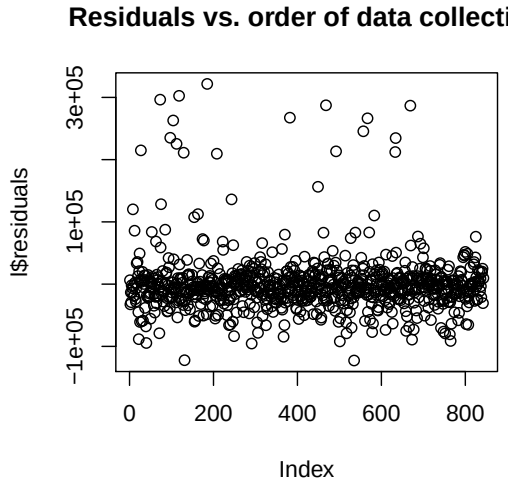
(2) Constant variability of residuals

```
par(mfrow=c(1,2))
plot(l$residuals ~ l$fitted, main = "Residuals vs. fitted")
abline(h = 0, lty = 3)
plot(abs(l$residuals) ~ l$fitted, main = "Absolute value of\nresiduals vs. fitted")
abline(h = 0, lty = 3)
```



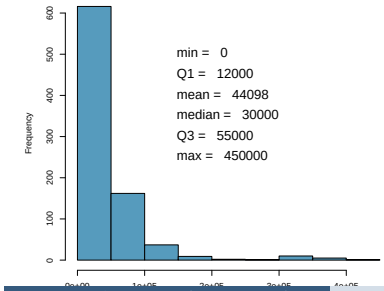
(3) Independence

```
plot(l$residuals, main = "Residuals vs. order 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 distribution is right skewed → suggests that a log transformation may be useful.

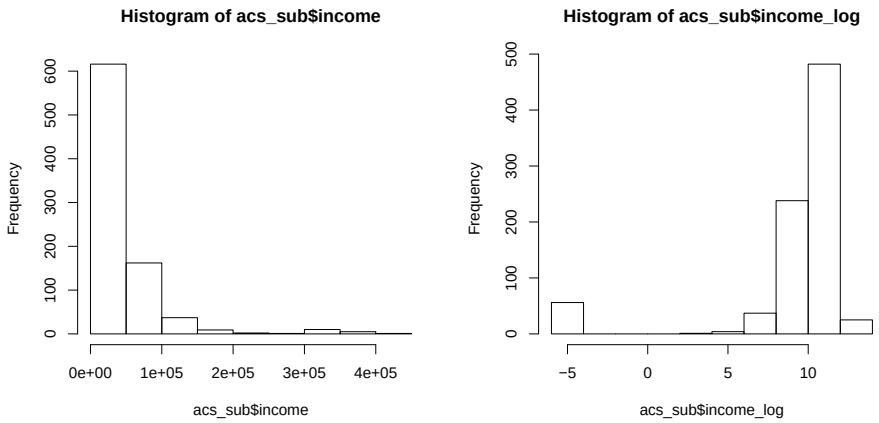
Log of 0

```
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 had 0 income (from salaries and wages) last year, we cannot take the log of their income, since $log(0) = -\infty$.
- A commonly used trick is to add a very small number to all values before taking the log.

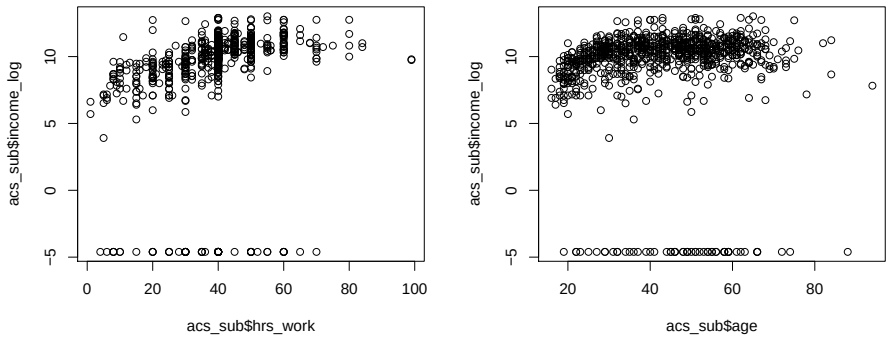
Logged income distribution

```
acs_sub$income_log = log(acs_sub$income + 0.01)
par(mfrow=c(1,2))
hist(acs_sub$income)
hist(acs_sub$income_log)
```



Logged income relationships

```
par(mfrow=c(1,2),mar=c(5, 4, 1, 2) + 0.1)
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.

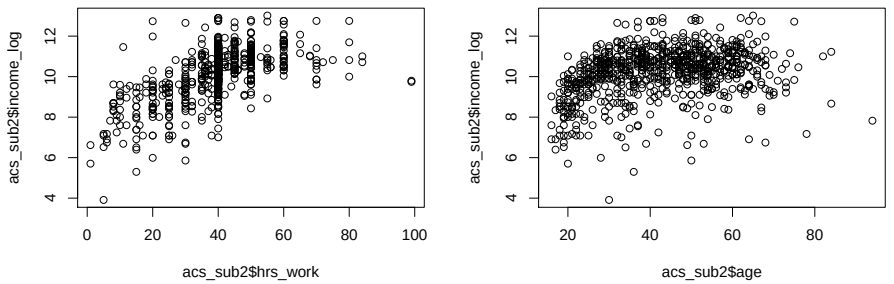
Further subsetting the data

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

```
par(mfrow=c(1,2))
plot(acs_sub2$income_log ~ acs_sub2$hrs_work)
plot(acs_sub2$income_log ~ acs_sub2$age)
```



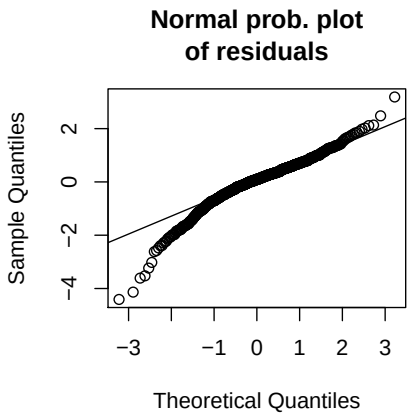
Predicting log of income

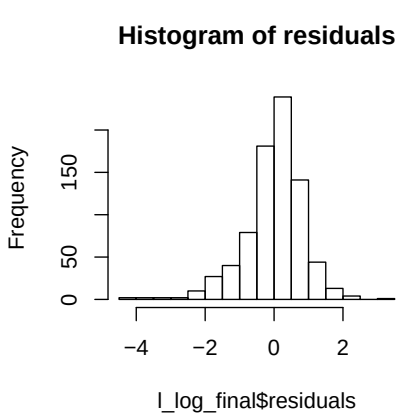
```
l_log = lm(income_log ~ hrs_work + race + age + gender + edu + disability, data = acs_sub2);summary(l_log)


##
## Call:
## lm(formula = income_log ~ hrs_work + race + age + gender + edu +
##     disability, data = acs_sub2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5725 -0.3936  0.0880  0.4993  3.1652
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.313684   0.140742  51.965  < 2e-16 ***
## hrs_work       0.048793   0.002526  19.317  < 2e-16 ***
## raceblack     -0.147579   0.104363  -1.414    0.158
## raceasian      0.136877   0.141616   0.967    0.334
## raceother     -0.192193   0.121193  -1.586    0.113
## age            0.022229   0.002175  10.222  < 2e-16 ***
## genderfemale  -0.276076   0.063702  -4.334 1.66e-05 ***
## educollege     0.399230   0.069932   5.709 1.62e-08 ***
## edugrad        0.833686   0.098711   8.446  < 2e-16 ***
## disabilityyes -0.624479   0.115492  -5.407 8.53e-08 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.849 on 777 degrees of freedom
## Multiple R-squared:  0.5202, Adjusted R-squared:  0.5146
## F-statistic: 93.59 on 9 and 777 DF,  p-value: < 2.2e-16
```


Final model for log of income

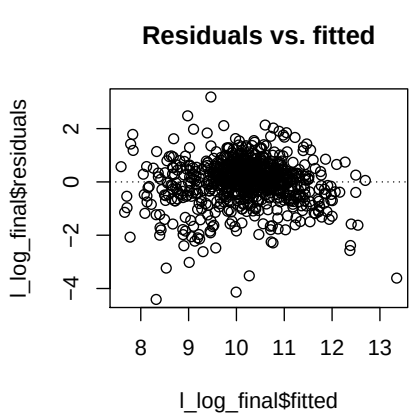
```
l_log_final = lm(income_log ~ hrs_work + age + gender + edu + disability, data = acs_sub2);summary(l_log_final)


##
## Call:
## lm(formula = income_log ~ hrs_work + age + gender + edu + disability,
##     data = acs_sub2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4098 -0.3936  0.0987  0.5124  3.1883
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.281258   0.139082  52.352  < 2e-16 ***
## hrs_work       0.049017   0.002527  19.395  < 2e-16 ***
## age            0.022309   0.002166  10.299  < 2e-16 ***
## genderfemale  -0.287365   0.063569  -4.521 7.13e-06 ***
## educollege     0.413555   0.069741   5.930 4.55e-09 ***
## edugrad        0.844909   0.098323   8.593  < 2e-16 ***
## disabilityyes -0.632040   0.115558  -5.469 6.08e-08 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8503 on 780 degrees of freedom
## Multiple R-squared:  0.5168, Adjusted R-squared:  0.5131
## F-statistic:   139 on 6 and 780 DF,  p-value: < 2.2e-16
```

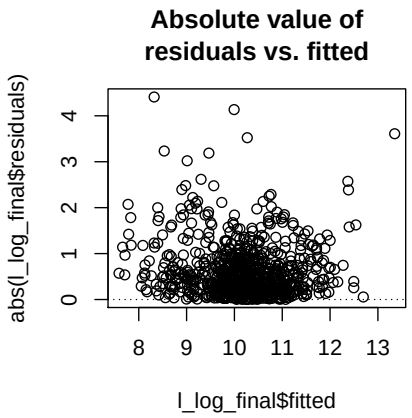

(1) Nearly normal residuals



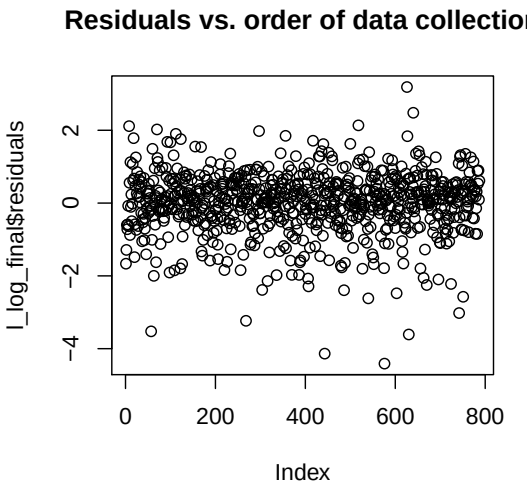


(2) Constant variability of residuals





(3) Independence



Interpretation

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.28	0.14	52.35	0.00
hrs_work	0.05	0.00	19.39	0.00
age	0.02	0.00	10.30	0.00
genderfemale	-0.29	0.06	-4.52	0.00
educollege	0.41	0.07	5.93	0.00
edugrad	0.84	0.10	8.59	0.00
disabilityyes	-0.63	0.12	-5.47	0.00

Interpretation (cont.)

Which of the following is the correct interpretation of the slope of edu:college?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.28	0.14	52.35	0.00
hrs_work	0.05	0.00	19.39	0.00
age	0.02	0.00	10.30	0.00
gender:female	-0.29	0.06	-4.52	0.00
edu:college	0.41	0.07	5.93	0.00
edu:grad	0.84	0.10	8.59	0.00
disability:yes	-0.63	0.12	-5.47	0.00