Lecture 18 - More multiple linear regression

Sta102 / BME102

Colin Rundel

April 9, 2014

Poverty vs. region (east, west)

$$\widehat{poverty} = 11.17 + 0.38 \times west$$

- *Explanatory variable*: region
- *Reference level*: east
- *Intercept:* estimated average % poverty in eastern states is 11.17%
  - This is the value we get if we plug in *0* for the explanatory variable
- *Slope:* estimated average % poverty in western states is 0.38% higher than eastern states.
  - Estimated average % poverty in western states is 11.17 + 0.38 = 11.55%.
  - This is the value we get if we plug in *1* for the explanatory variable

Poverty vs. Region (Northeast, Midwest, West, South)

Which region (Northeast, Midwest, West, South) is the reference level?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4midwest	0.03	1.15	0.02	0.98
region4west	1.79	1.13	1.59	0.12
region4south	4.16	1.07	3.87	0.00

Interpretation:

- Predict 9.50% poverty in Northeast
- Predict 9.53% poverty in Midwest
- Predict 11.29% poverty in West
- Predict 13.66% poverty in South

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}
- p-value approach:
  - Pick a critical value α_{crit}
  - Start with the full model
  - Drop the variable with the highest p-value and refit a smaller model
  - Repeat until all variables left have a p-value smaller than α_{crit}
- 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

Backward-selection: p-value approach

Full model:
```
lm(formula = kid_score ~ mom_hs + mom_iq + mom_work + mom_age, data = cognitive)

            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
```

Step 1: `lm(formula = kid_score ~ mom_hs + mom_iq + mom_work, data = cognitive)`
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.17944    6.04319   4.001 7.42e-05 ***
mom_hsyes    5.38225    2.27156   2.369   0.0183 *
mom_iq       0.56278    0.06057   9.291  < 2e-16 ***
mom_workyes  2.56640    2.34871   1.093   0.2751
```

Step 2: `lm(formula = kid_score ~ mom_hs + mom_iq, data = cognitive)`
```
            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 ***
```


adjusted R^2 vs. p-value

- If you're interested in finding out which variables are significant predictors, use p-value approach.
- If you're interested in more reliable predictions, use adjusted R^2 method.
- Most of the time (simple cases) both procedures will arrive at the same (or very similar) models.
- Note that the p-value method depends on the (somewhat arbitrary) α_{crit} cutoff. Using a different significance level you could get a completely different model. It is used commonly since it requires fitting fewer models (in the more commonly used backwards-selection approach).

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}
2. p-value approach:
   - Start with regressions of response vs. each explanatory variable
   - Pick the variable with the smallest p-value
   - Add the remaining variables one at a time to the existing model, and pick the variable with the smallest p-value below α_{crit}
   - Repeat until any of the remaining variables does not have a p-value below α_{crit}

In forward-selection the p-value approach is not any simpler (you still need to fit a bunch of models), so there's little reason to use it.

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**	$\mathbf{R^2_{adj}}$
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: p-value approach

Which variable should be added to the model first?

```
lm(formula = kid_score ~ mom_hs, data = cognitive)

            Estimate Std. Error t value Pr(>|t|)
mom_hsyes     11.771      2.322   5.069 5.96e-07

lm(formula = kid_score ~ mom_iq, data = cognitive)

            Estimate Std. Error t value Pr(>|t|)
mom_iq       0.60997    0.05852   10.42  < 2e-16

lm(formula = kid_score ~ mom_work, data = cognitive)

            Estimate Std. Error t value Pr(>|t|)
mom_workyes    5.832      2.552   2.285   0.0228

lm(formula = kid_score ~ mom_age, data = cognitive)

            Estimate Std. Error t value Pr(>|t|)
mom_age       0.6952     0.3620   1.920   0.0555
```


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

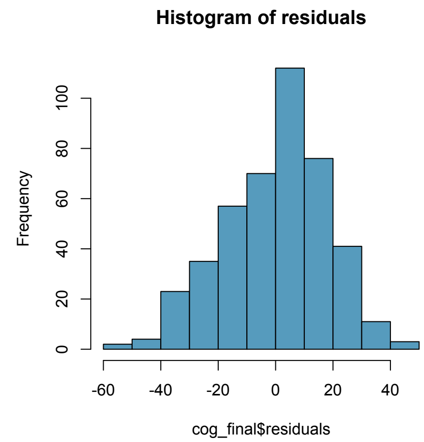
In order to perform inference for multiple regression we require the following conditions:

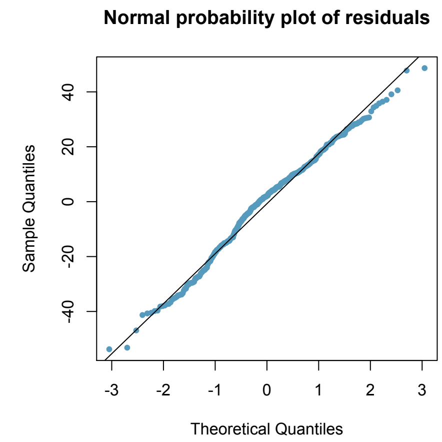
(1) Nearly normal residuals

(2) Constant variability of residuals

(3) Independent residuals

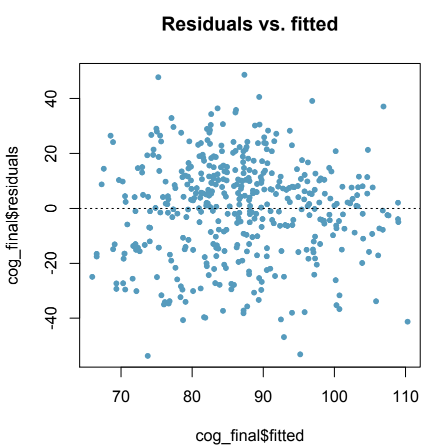
Nearly normal residuals

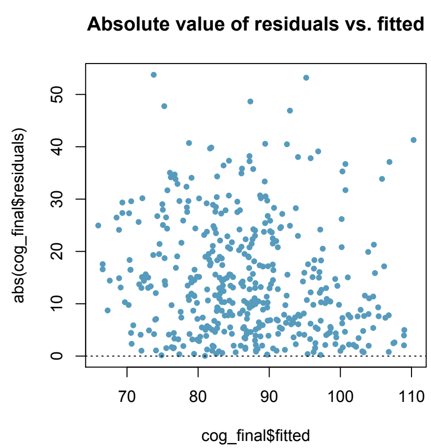




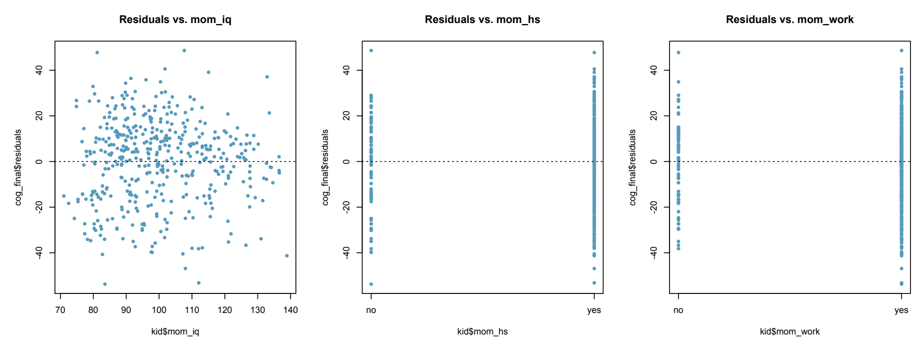
Constant variability of residuals

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



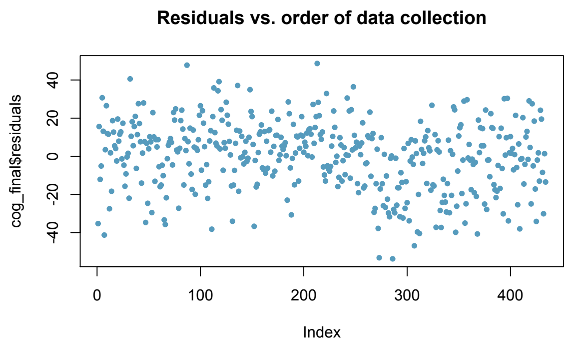


Constant variability of residuals (cont.)



Independent residuals

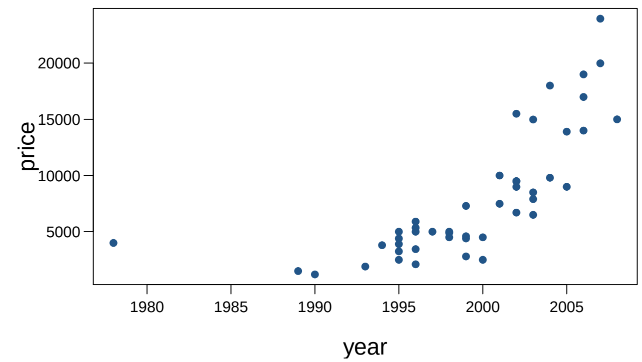
- If we suspect that order of data collection may influence the outcome (mostly in time series data):



- 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.

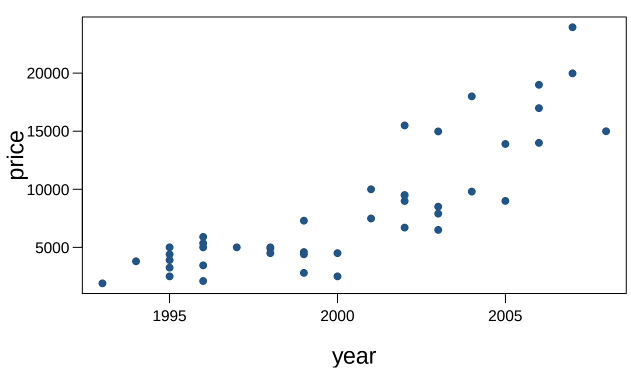


From: *http://faculty.chicagobooth.edu/robert.gramacy/teaching.html*

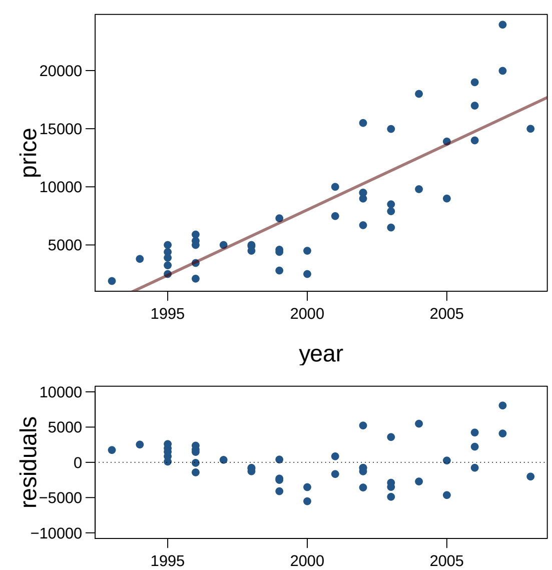
Remove unusual observations

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?



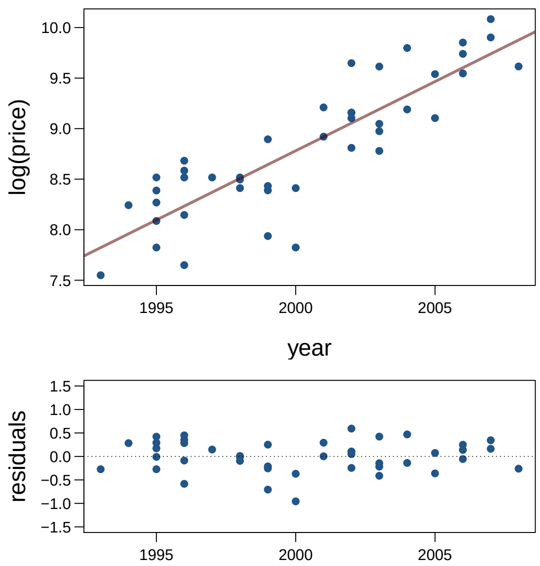
Model:

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

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

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.

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

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

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)