Statistical Inference

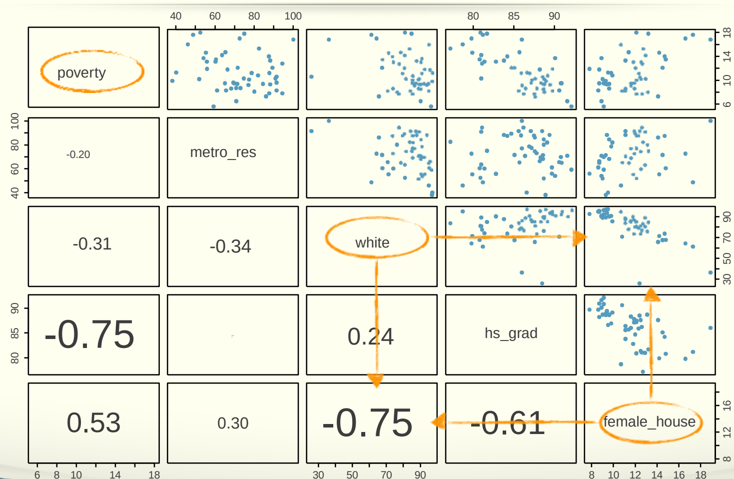
Introduction to Linear Regression

Behnam Bahrak
Spring 2020

Collinearity

- Two predictor variables are said to be collinear when they are correlated with each other.
- Remember: Predictors are also called independent variables, so they should be independent of each other.
- Inclusion of collinear predictors (also called multicollinearity) complicates model estimation.

Pairwise Scatter Plots



Parsimony

- Avoid adding predictors associated with each other because often times the addition of such variable brings nothing new to the table
- Prefer the simplest best model, i.e. the parsimonious model
  - Occam's razor: Among competing hypotheses, the one with the fewest assumptions should be selected
- Addition of collinear variables can result in biased estimates of the regression parameters
- While it's impossible to avoid collinearity from arising in observational data, experiments are usually designed to control for correlated predictors

Modeling cognitive test scores of children

- Data: Cognitive test scores of three- and four-year-old children and characteristics of their mothers (from 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
…	…	…	…	…	…
6	98	no	107.90	no	18
…	…	…	…	…	…
434	70	yes	91.25	yes	25

Fit a Model using R

```
R
# full model
> cog_full = lm(kid_score ~ mom_hs + mom_iq + mom_work + mom_age, data = cognitive)
> summary(cog_full)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.59241    9.21906   2.125   0.0341 *
mom_hs:yes   5.09482    2.31450   2.201   0.0282 *
mom_iq       0.56147    0.06064   9.259   <2e-16 ***
mom_work:yes 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
```

Inference for the model as a whole

H_0: $\beta_1 = \beta_2 = \ \ldots = \beta_k = 0$

H_A: at least one β_i is different than 0

```
F-statistic: 29.74 on 4 and 429 DF, p-value: < 2.2e-16
```

- Since p-value < 0.05, the model as a whole is significant.
- The F test yielding a significant result doesn't mean the model fits the data well, it just means at least one of the β_is is non-zero.
- The F test not yielding a significant result doesn't mean individual variables included in the model are not good predictors of y, it just means that the combination of these variables doesn't yield a good model.

Hypothesis testing for slopes

- Is whether or not the mother went to high school a significant predictor of the cognitive test scores of children, given all other variables in the model?

H_0: $\beta_1 = 0$, when all other variables are included in the model
H_A: $\beta_1 \neq 0$, when all other variables are included in the model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59241	9.21906	2.125	0.0341
mom_hs:yes	5.09482	2.31450	2.201	0.0282
mom_iq	0.56147	0.06064	9.259	<2e-16
mom_work:yes	2.53718	2.35067	1.079	0.2810
mom_age	0.21802	0.33074	0.659	0.5101

- Whether or not mom went to high school is a significant predictor of the cognitive test scores of children, given all other variables in the model.

Testing for the slope - mechanics

➢ Use a t-statistic in inference for regression

$$T = \frac{\text{point estimate} - \text{null value}}{SE}$$

b_1 SE_{b_1}

***t*-statistic for the slope:**	$T = \dfrac{b_1 - 0}{SE_{b_1}}$	$df = n - k - 1$

Degrees of Freedom

➢ Multiple predictors:

$$df = n - k - 1$$

➢ Single predictors:

$$df = n - 1 - 1 = n - 2$$

➢ Lose 1 df for each parameter estimated, and one for the intercept.

Example

➢ Verify the T score and the p-value for the slope of `mom_hs`.

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   19.59241    9.21906   2.125   0.0341
mom_hs:yes     5.09482    2.31450   2.201   0.0282
mom_iq         0.56147    0.06064   9.259   <2e-16
mom_work:yes   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
```

$$T = \frac{5.095 - 0}{2.315} = 2.201$$

$$df = n - k - 1 = 434 - 4 - 1 = 429$$

R

```
> pt(2.201,df = 429, lower.tail = FALSE) * 2
[1] 0.0282
```


Confidence Intervals for Slopes

point estimate ± margin of error

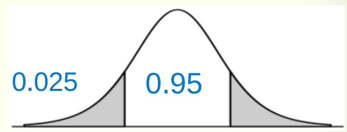
$$b_1 \pm t^\star_{df} SE_{b_1}$$

Example

- Calculate the 95% confidence interval for the slope of `mom_work`.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59241	9.21906	2.125	0.0341
mom_hs:yes	5.09482	2.31450	2.201	0.0282
mom_iq	0.56147	0.06064	9.259	<2e-16
mom_work:yes	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



R

```
> qt(0.025, df = 429)
[1] -1.97
```

$df = 434 - 4 - 1 = 429$

$t^*_{429} = 1.97$

$2.54 \pm 1.97 \times 2.35 \approx (-2.09\,, 7.17)$

Example

- Interpret the 95% confidence interval for the slope of `mom_work`.

CI: (-2.09, 7.17)

- We are 95% confident that, all else being equal, the model predicts that children whose moms worked during the first three years of their lives score 2.09 points lower to 7.17 points higher than those whose moms did not work.

Stepwise Model Selection

- Backwards elimination: start with a full model (containing all predictors), drop one predictor at a time until the parsimonious model is reached.
- Forward selection: start with an empty model and add one predictor at a time until the parsimonious model is reached.
- Criteria:
  - p-value, adjusted R^2
  - AIC, BIC, DIC, Bayes factor, Mallow's C_p (beyond the scope of this course)

Backwards Elimination - Adjusted R^2

- Start with the full model
- Drop one variable at a time and record adjusted R^2 of each smaller model
- Pick the model with the highest increase in adjusted R^2
- Repeat until none of the models yield an increase in adjusted R^2

Backwards Elimination - Adjusted R^2

step	variables included	removed	adjusted R
FULL	kid_score ~ mom_hs + mom_iq + mom_work + mom_age		0.2098
STEP 1	kid_score ~ mom_iq + mom_work + mom_age	[-mom_hs]	0.2027
	kid_score ~ mom_hs + mom_work + mom_age	[-mom_iq]	0.0541
	kid_score ~ mom_hs + mom_iq + mom_age	[-mom_work]	0.2095
	kid_score ~ mom_hs + mom_iq + mom_work	[-mom_age]	0.2109
STEP 2	kid_score ~ mom_iq + mom_work	[-mom_hs]	0.2024
	kid_score ~ mom_hs + mom_work	[-mom_iq]	0.0546
	kid_score ~ mom_hs + mom_iq	[-mom_work]	0.2105

Backwards Elimination - p-value

- 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

Backwards Elimination - p-value

FULL	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.5924	9.2191	2.13	0.0341
mom_hs:yes	5.0948	2.3145	2.20	0.0282
mom_iq	0.5615	0.0606	9.26	0.0000
mom_work:yes	2.5372	2.3507	1.08	0.2810
~~mom_age~~	~~0.2180~~	~~0.3307~~	~~0.66~~	~~0.5101~~

STEP 1	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	24.1794	6.0432	4.00	0.0001
mom_hs:yes	5.3823	2.2716	2.37	0.0183
mom_iq	0.5628	0.0606	9.29	0.0000
~~mom_work:yes~~	~~2.5664~~	~~2.3487~~	~~1.09~~	~~0.2751~~

STEP 2	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	25.7315	5.8752	4.38	0.0000
mom_hsyes	5.9501	2.2118	2.69	0.0074
mom_iq	0.5639	0.0606	9.31	0.0000

Example

- The following model uses data from the American Community Survey to predict income from hours worked per week, race, and gender. Which variable (if any) should be dropped from the model first when doing backwards elimination using the p-value approach?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2782.5726	6676.5534	0.42	0.6770
hrs_work	1247.2128	146.2013	8.53	0.0000
race:black	-9565.3090	6393.2168	-1.50	0.1350
race:asian	35816.6156	8690.3484	4.12	0.0000
race:other	-11112.8617	7213.3220	-1.54	0.1238
gender:female	-16430.0916	3803.4700	-4.32	0.0000

don't drop any variables

Adjusted R^2 vs. p-value

- p-value: statistically significant predictors
- Adjusted R^2: more reliable predictions
- p-value method depends on the (somewhat arbitrary) 5% significance level cutoff
- Different significance level → different model
- P-value is used commonly since it requires fitting fewer models (in the more commonly used backwards-selection approach)

Forward Selection - Adjusted R^2

- Start with single predictor regressions of response vs. each explanatory variable
- Pick the model with the highest adjusted R^2
- Add the remaining variables one at a time to the existing model, and pick the model with the highest adjusted R^2
- Repeat until the addition of any of the remaining variables does not result in a higher adjusted R^2

Forward Selection - Adjusted R^2

step	variables included	adjusted R
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_hs + mom_iq + mom_work + mom_age	0.2098

Forward Selection - p-value

- Start with single predictor 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 do not have a significant p-value

Expert Opinion

- Variables can be included in (or eliminated from) the model based on expert opinion
- If you are studying a certain variable, you might choose to leave it in the model regardless of whether it's significant or yield a higher adjusted R^2

Final Model

R

```
> cog_final = lm(kid_score ~ mom_hs + mom_iq + mom_work, data = cognitive)
> summary(cog_final)

Coefficients:
            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

Residual standard error: 18.13 on 430 degrees of freedom
Multiple R-squared:  0.2163, Adjusted R-squared:  0.2109
F-statistic: 39.57 on 3 and 430 DF,  p-value: < 2.2e-16
```

Diagnostics for MLR

- Linear relationships between x and y
- Nearly normal residuals
- Constant variability of residuals
- Independence of residuals

Condition 1

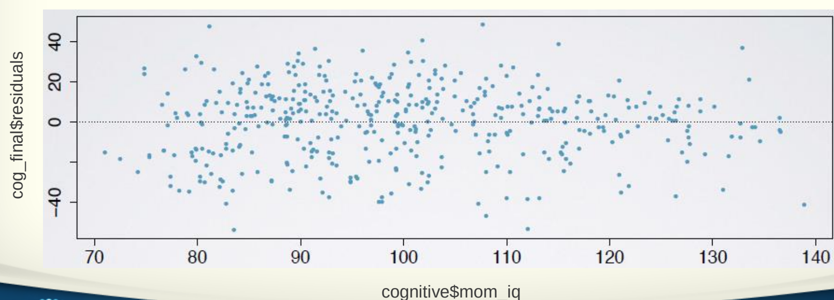
(1) Linear relationships between (numerical) x and y

- Each (numerical) explanatory variable linearly related to the response variable
- Check using residuals plots (e vs. x)
  - Looking for a random scatter around 0
- Instead of scatterplot of y vs. x: allows for considering the other variables that are also in the model, and not just the bivariate relationship between a given x and y

Example

R
```
> cog_final = lm(kid_score ~ mom_hs + mom_iq + mom_work, data = cognitive)
> plot(cog_final$residuals ~ cognitive$mom_iq)
```

Residuals vs. mom_iq



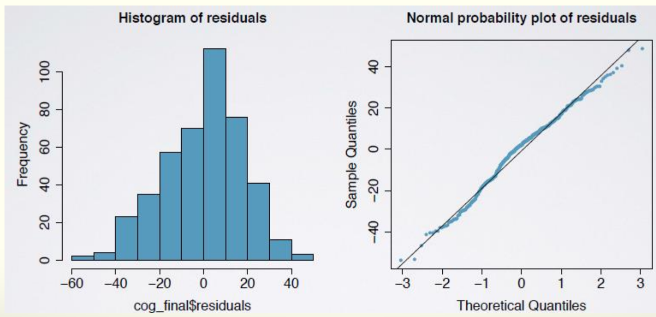
Condition 2

(2) Nearly normal residuals with mean 0

- Some residuals will be positive and some negative
- On a residuals plot we look for random scatter of residuals around 0
- This translates to a nearly normal distribution of residuals centered at 0
- Check using histogram or normal probability plot

Example

R
```
> hist(cog_final$residuals)
> qqnorm(cog_final$residuals)
> qqline(cog_final$residuals)
```



Condition 3

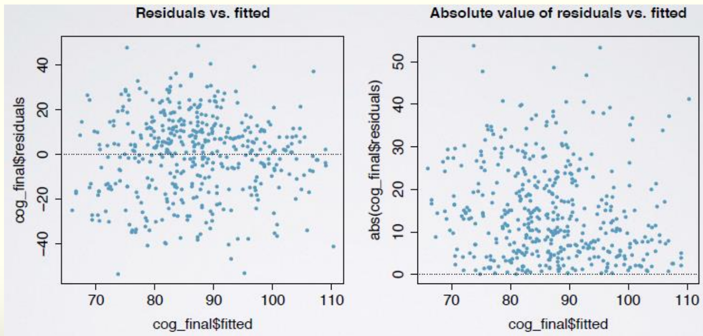
(3) Constant variability of residuals

- Residuals should be equally variable for low and high values of the predicted response variable
- Check using residuals plots of residuals vs. predicted (e vs. $\hat{y}$)
- Residuals vs. predicted instead of residuals vs. x because it allows for considering the entire model (with all explanatory variables) at once.
- Residuals randomly scattered in a band with a constant width around 0 (no fan shape)
- Also worthwhile to view absolute value of residuals vs. predicted to identify unusual observations easily

Example

R

```
> plot(cog_final$residuals ~ cog_final$fitted)
> plot(abs(cog_final$residuals) ~ cog_final$fitted)
```



Condition 4

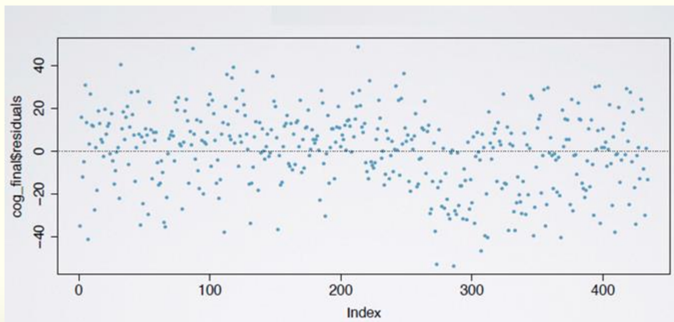
(4) independent residuals

- Independent residuals → independent observations
- If time series structure is suspected check using residuals vs. order of data collection
- If not, think about how the data are sampled

Example

R

```
> plot(cog_final$residuals)
```



Ridge Regression

- Recall that the least squares fitting procedure estimates $\beta_0, \beta_1, \ldots, \beta_p$ using the values that minimize:

$$RSS = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_i x_{ij}\right)^2$$

- In contrast the ridge regression coefficient estimates β_i that minimize:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_i x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

where λ is a tuning parameter.

The Lasso

- LASSO: Least Absolute Shrinkage and Selection Operator
- The Lasso is a relatively recent alternative to ridge regression that minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_i x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

- The Lasso performs variable selection much better than ridge regression.