Problems with Linear Regression

Today:

- Define six major problems analysts using regression run into;
- Understand what their effects will be on your regression.

So far, we've used linear regression to:

1. Model a linear relationship between dependent variable y and independent variable x:

$$y = \beta_0 + \beta_1 x + \varepsilon;$$

2. Model a linear relationship between dependent variable y and many independent variables $x_1, x_2, x_3, \ldots$:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \varepsilon;$$

3. Model a NON-linear relationship between dependent variable y and many independent variables x:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \ldots + \varepsilon;$$

4. Model relationships between independent variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon;$$

Reminder: Assumptions of Linear Regression

1. Dependent variable is a **linear** function of independent variables plus noise;

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + \varepsilon$$

2. Independent variables are not related to each other – **no multicollinearity**;

3. Independent variables have **no measurement error**;

4. Noise term is a random variable following the **normal distribution**;

Reminder: Assumptions of Linear Regression

1. Dependent variable is a **linear** function of independent variables plus noise;

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + \varepsilon$$

   - Violated by: selecting on the dep var, model specification, endogeneity;

2. Independent variables are not related to each other – **no multicollinearity**;

3. Independent variables have **no measurement error**;

4. Noise term is a random variable following the **normal distribution**;

Reminder: Assumptions of Linear Regression

1. Dependent variable is a **linear** function of independent variables plus noise;

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + \varepsilon$$

    - Violated by: selecting on the dep var, model specification, endogeneity;

2. Independent variables are not related to each other – **no multicollinearity**;
    - Violated by: Multicollinearity between indp vars;

3. Independent variables have **no measurement error**;

4. Noise term is a random variable following the **normal distribution**;

Reminder: Assumptions of Linear Regression

1. Dependent variable is a **linear** function of independent variables plus noise;

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + \varepsilon$$

   - Violated by: selecting on the dep var, model specification, endogeneity;

2. Independent variables are not related to each other – **no multicollinearity**;
   - Violated by: Multicollinearity between indp vars;

3. Independent variables have **no measurement error**;
   - Violated by: Measurement error of indp vars;

4. Noise term is a random variable following the **normal distribution**;

Reminder: Assumptions of Linear Regression

1. Dependent variable is a **linear** function of independent variables plus noise;

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + \varepsilon$$

   - Violated by: selecting on the dep var, model specification, endogeneity;

2. Independent variables are not related to each other – **no multicollinearity**;
   - Violated by: Multicollinearity between indp vars;
3. Independent variables have **no measurement error**;
   - Violated by: Measurement error of indp vars;
4. Noise term is a random variable following the **normal distribution**;
   - Violated by: Heteroscedasticity.

Problem 0: Selecting on the dependent variable

Suppose you are an analyst trying to understand the effect of a certain chemical on mortality. You collect a sample of deceased persons and you observe that every single one of them was exposed to the chemical in large quantities. What do you conclude?

Problem 0: Selecting on the dependent variable

- **Selecting on the dependent variable** = collecting only one value of y when making the data;
- **Effect**: well, you can't run linear regression, and any inferences you try to draw will be unrelated to the data;
- **Fix**: collect more data.

Problem 1: Model Specification

- **Model specification** $=$ which independent variables you choose to include – leaving out an independent variable out that should be there is called **omitted variable bias**;

- **Effect**: the independent variable effects (the β's) that linear regression estimates will be wrong;

- **Fix**: theorizing about why variables are/not included, advanced techniques.

Problem 1: Model Specification

- True relationship between y and x's is:

$$y = 1 + 2x_1 - x_2 + \varepsilon;$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.9623	0.0317	30.40	0.0000
x1	2.0401	0.1030	19.80	0.0000
x2	-1.0055	0.0322	-31.19	0.0000

Problem 1: Model Specification

- True relationship between y and x's is:

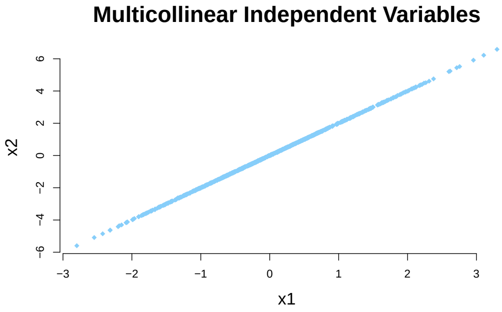
$$y = 1 + 2x_1 - x_2 + \varepsilon;$$

- Suppose we omit x_2:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.9698	0.0445	21.81	0.0000
x1	-0.9945	0.0475	-20.95	0.0000

Problem 2: Multicollinearity

- **Multicollinearity** = two (or more) independent variables correlated with each other;
- **Effect**: the independent variable effects (the β's) that linear regression estimates will be wrong;
- **Fix**: drop one of the independent variables, advanced techniques.



Problem 2: Multicollinearity

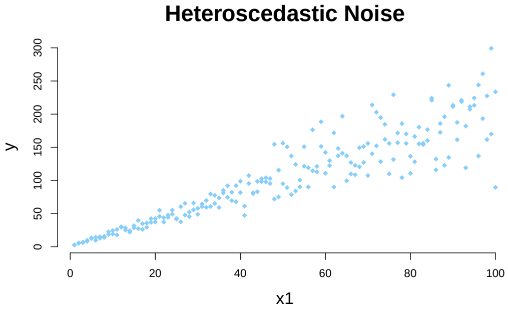
- True relationship between y and x's is:

$$y = 1 + 2x_1 - x_2 + \varepsilon;$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.0314	0.0314	32.88	0.0000
x1	-4.2018	6.2780	-0.67	0.5035
x2	2.0998	3.1390	0.67	0.5037

Problem 3: Heterscedasticity

- **Heterscedasticity** = the standard deviation of the noise is not constant;
- **Effect**: the *p*-values will be too large leading you to fail to reject the null when you really should;
- **Fix**: transform variables (e.g. log y), advanced techniques.



Problem 3: Heterscedasticity

- True relationship between y and x's is:

$$y = 1 + 2x_1 - x_2 + \varepsilon;$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6148	3.8241	0.16	0.8724
x1	2.0667	0.0658	31.43	0.0000
x2	-0.3911	1.8852	-0.21	0.8359

Problem 4: Measurement error

- **Measurement error** = we make mistakes measuring the independent variables when collecting the data;
- **Effect**: the independent variable effect (the β) on the badly measured variable that linear regression estimates will be too small;
- **Fix**: advanced techniques.

Problem 4: Measurement error

- True relationship between y and x's is:

$$y = 1 + 2x_1 - x_2 + \varepsilon;$$

- When we collect the data we make mistakes in measuring so that we collect $x_2 + e$;

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.4749	0.0361	40.89	0.0000
x1	2.0239	0.0327	61.87	0.0000
x2	-0.9011	0.0306	-29.43	0.0000

Problem 4: Measurement error

- True relationship between y and x's is:

$$y = 1 + 2x_1 - x_2 + \varepsilon;$$

- When we collect the data we make mistakes in measuring so that we collect $x_2 + e$;

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.7834	0.0469	38.05	0.0000
x1	1.9934	0.0369	54.09	0.0000
x2	-0.7493	0.0312	-23.98	0.0000

Problem 4: Measurement error

- True relationship between y and x's is:

$$y = 1 + 2x_1 - x_2 + \varepsilon;$$

- When we collect the data we make mistakes in measuring so that we collect $x_2 + e$;

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.9629	0.0581	33.81	0.0000
x1	1.9710	0.0385	51.22	0.0000
x2	-0.6529	0.0293	-22.28	0.0000

Problem 4: Measurement error

- True relationship between y and x's is:

$$y = 1 + 2x_1 - x_2 + \varepsilon;$$

- When we collect the data we make mistakes in measuring so that we collect $x_2 + e$;

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.7451	0.0745	23.43	0.0000
x1	2.0534	0.0397	51.72	0.0000
x2	-0.3123	0.0245	-12.72	0.0000

Problem 5: Endogeneity

- **Endogeneity** = the independent variables cause the dependent variable AND the dependent variable also causes one of the independent variables;
- **Effect**: the independent variable effects (the βs) that linear regression estimates will be wrong;
- **Fix**: theorizing about variable relationships, restructuring the data, removing independent variables thought to be caused by the dependent variable from the regression, advanced techniques.

Problem 5: Endogeneity

- True relationship between y and x's is:

$$\begin{aligned} y &= 1 + 2x_1 - x_2 + \varepsilon; \\ x_2 &= 3y + e; \end{aligned}$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0244	0.0101	2.42	0.0155
x1	0.0647	0.0107	6.07	0.0000
x2	0.3017	0.0032	95.35	0.0000

Why should we care?

Linear regression – so powerful...and so easy to mess up!