

# Econ 103: Topics in Single Linear Regression

Manu Navjeevan

UCLA

July 26, 2021

## Advanced Inference Topics

- Inference on Linear Combinations of Parameters
- Heteroskedasticity

## Evaluating our Model

- Prediction
- $R^2$  and goodness of fit

## Modeling Choices

- How do results change if we apply linear transformations?
- Useful non-linear transformations of  $X$  and  $Y$

# Table of Contents

---

## Advanced Inference Topics

## Evaluating our Model

## Modeling Choices

Recall that, approximately for large  $n$

$$\sqrt{n}(\hat{\beta}_0 - \beta_0) \sim N\left(0, \mathbb{E}[X^2]\sigma_\epsilon^2/\sigma_X^2\right), \quad \sqrt{n}(\hat{\beta}_1 - \beta_1) \sim N\left(0, \sigma_\epsilon^2/\sigma_X^2\right)$$

and  $\sigma_{\beta_{01}} = \text{Cov}(\sqrt{n}\{\hat{\beta}_0 - \beta_0\}, \sqrt{n}\{\hat{\beta}_1 - \beta_1\}) = -\mathbb{E}[X] \frac{\sigma_\epsilon^2}{\sigma_X^2}.$

Recall that, approximately for large  $n$

$$\sqrt{n}(\hat{\beta}_0 - \beta_0) \sim N\left(0, \mathbb{E}[X^2]\sigma_\epsilon^2/\sigma_X^2\right), \quad \sqrt{n}(\hat{\beta}_1 - \beta_1) \sim N\left(0, \sigma_\epsilon^2/\sigma_X^2\right)$$

and  $\sigma_{\beta_{01}} = \text{Cov}(\sqrt{n}\{\hat{\beta}_0 - \beta_0\}, \sqrt{n}\{\hat{\beta}_1 - \beta_1\}) = -\mathbb{E}[X] \frac{\sigma_\epsilon^2}{\sigma_X^2}.$

These results were also often presented in the following equivalent manners

$$\frac{\hat{\beta}_0 - \beta_0}{\sigma_{\beta_0}/\sqrt{n}} \sim N(0, 1) \quad \text{and} \quad \hat{\beta}_0 \sim N(\beta_0, \sigma_{\beta_0}^2/n)$$
$$\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\beta_1}/\sqrt{n}} \sim N(0, 1) \quad \text{and} \quad \hat{\beta}_1 \sim N(\beta_1, \sigma_{\beta_1}^2/n)$$

where  $\sigma_{\beta_0}^2 = \mathbb{E}[X^2]\sigma_\epsilon^2/\sigma_X^2$  and  $\sigma_{\beta_1}^2 = \sigma_\epsilon^2/\sigma_X^2$ .

Recall that, approximately for large  $n$

$$\sqrt{n}(\hat{\beta}_0 - \beta_0) \sim N\left(0, \mathbb{E}[X^2]\sigma_\epsilon^2/\sigma_X^2\right), \quad \sqrt{n}(\hat{\beta}_1 - \beta_1) \sim N\left(0, \sigma_\epsilon^2/\sigma_X^2\right)$$

and  $\sigma_{\beta_{01}} = \text{Cov}(\sqrt{n}\{\hat{\beta}_0 - \beta_0\}, \sqrt{n}\{\hat{\beta}_1 - \beta_1\}) = -\mathbb{E}[X] \frac{\sigma_\epsilon^2}{\sigma_X^2}$ .

These results were also often presented in the following equivalent manners

$$\frac{\hat{\beta}_0 - \beta_0}{\sigma_{\beta_0}/\sqrt{n}} \sim N(0, 1) \quad \text{and} \quad \hat{\beta}_0 \sim N(\beta_0, \sigma_{\beta_0}^2/n)$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\beta_1}/\sqrt{n}} \sim N(0, 1) \quad \text{and} \quad \hat{\beta}_1 \sim N(\beta_1, \sigma_{\beta_1}^2/n)$$

where  $\sigma_{\beta_0}^2 = \mathbb{E}[X^2]\sigma_\epsilon^2/\sigma_X^2$  and  $\sigma_{\beta_1}^2 = \sigma_\epsilon^2/\sigma_X^2$ .

- Also went over how to estimate these variances

In the last lecture, we used these distributional results to compute objects like

$$\Pr \left( |\hat{\beta}_1| > 5 \mid \beta_1 = 0 \right).$$

which in turn were useful for hypothesis testing

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0.$$

However, often we want to perform inference not just on one parameter, but on a linear combination of parameters, i.e we want to test

$$H_0 : \beta_0 + 5\beta_1 = 0 \text{ vs. } H_1 : \beta_0 + 5\beta_1 \neq 0.$$



However, often we want to perform inference not just on one parameter, but on a linear combination of parameters, i.e we want to test

$$H_0 : \beta_0 + 5\beta_1 = 0 \text{ vs. } H_1 : \beta_0 + 5\beta_1 \neq 0.$$

This is useful, for example, if we are trying to test something like

$$H_0 : \mathbb{E}[Y|X = 5] = 0 \text{ vs. } H_1 : \mathbb{E}[Y|X = 5] \neq 0$$

and we view the linear regression model  $Y = \beta_0 + \beta_1 X + \epsilon$  as a way of approximating the conditional mean function  $\mathbb{E}[Y|X = x]$ .

In order to test such a hypothesis we want to know the distribution of a linear combination of our model parameters. That is, for  $\lambda = a\beta_0 + b\beta_1$  we would like to know the approximate distribution of

$$\hat{\lambda} = a\hat{\beta}_0 + b\hat{\beta}_1$$

so that we can calculate objects like  $\Pr(|\hat{\lambda}| > 0.5 \mid \lambda = 0)$ .

In order to test such a hypothesis we want to know the distribution of a linear combination of our model parameters. That is, for  $\lambda = a\beta_0 + b\beta_1$  we would like to know the approximate distribution of

$$\hat{\lambda} = a\hat{\beta}_0 + b\hat{\beta}_1$$

so that we can calculate objects like  $\Pr(|\hat{\lambda}| > 0.5 \mid \lambda = 0)$ .

Note that

$$\begin{aligned}\sqrt{n}(\hat{\lambda} - \lambda) &= \sqrt{n}(a\hat{\beta}_0 + b\hat{\beta}_1 - a\beta_0 - b\beta_1) \\ &= a\sqrt{n}(\hat{\beta}_0 - \beta_0) + b\sqrt{n}(\hat{\beta}_1 - \beta_1).\end{aligned}$$

and that we know the (joint) distribution of  $\sqrt{n}(\hat{\beta}_0 - \beta_0)$  and  $\sqrt{n}(\hat{\beta}_1 - \beta_1)$ .

Recall from our Econ 41 Review that the sum of two jointly normal random variables is also normally distributed and that if  $X$  and  $Y$  are random variables then

$$\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{Cov}(X, Y).$$

Recall from our Econ 41 Review that the sum of two jointly normal random variables is also normally distributed and that if  $X$  and  $Y$  are random variables then

$$\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{Cov}(X, Y).$$

Using this result along with  $X = \sqrt{n}(\hat{\beta}_0 - \beta_0)$  and  $Y = \sqrt{n}(\hat{\beta}_1 - \beta_1)$  gives us that, for large  $n$ :

$$\sqrt{n}(\hat{\lambda} - \lambda) \sim N(0, \sigma_\lambda^2) \implies \frac{\hat{\lambda} - \lambda}{\sigma_\lambda/\sqrt{n}} \sim N(0, 1),$$

where  $\sigma_\lambda^2 = a^2 \sigma_{\beta_0}^2 + b^2 \sigma_{\beta_1}^2 + 2ab \sigma_{\beta_0 \beta_1}$

As a reminder, we can estimate

$$\sigma_{\beta_0}^2 = \mathbb{E}[X^2] \frac{\sigma_\epsilon^2}{\sigma_X^2} \iff \hat{\sigma}_{\beta_0}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 \cdot \frac{\hat{\sigma}_\epsilon^2}{\hat{\sigma}_X^2}$$

$$\sigma_{\beta_1}^2 = \frac{\sigma_\epsilon^2}{\sigma_X^2} \iff \hat{\sigma}_{\beta_1}^2 = \frac{\hat{\sigma}_\epsilon^2}{\hat{\sigma}_X^2}$$

$$\sigma_{\beta_{01}} = -\mathbb{E}[X] \frac{\sigma_\epsilon^2}{\sigma_X^2} \iff \hat{\sigma}_{\beta_{01}} = \bar{X} \frac{\hat{\sigma}_\epsilon^2}{\hat{\sigma}_X^2}$$

As a reminder, we can estimate

$$\sigma_{\beta_0}^2 = \mathbb{E}[X^2] \frac{\sigma_\epsilon^2}{\sigma_X^2} \iff \hat{\sigma}_{\beta_0}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 \cdot \frac{\hat{\sigma}_\epsilon^2}{\hat{\sigma}_X^2}$$

$$\sigma_{\beta_1}^2 = \frac{\sigma_\epsilon^2}{\sigma_X^2} \iff \hat{\sigma}_{\beta_1}^2 = \frac{\hat{\sigma}_\epsilon^2}{\hat{\sigma}_X^2}$$

$$\sigma_{\beta_{01}} = -\mathbb{E}[X] \frac{\sigma_\epsilon^2}{\sigma_X^2} \iff \hat{\sigma}_{\beta_{01}} = \bar{X} \frac{\hat{\sigma}_\epsilon^2}{\hat{\sigma}_X^2}$$

So, we can use these to estimate  $\sigma_\lambda^2 = a^2\sigma_{\beta_0}^2 + b^2\sigma_{\beta_1}^2 + 2ab\sigma_{\beta_{01}}$  with

$$\hat{\sigma}_\lambda^2 = a^2\hat{\sigma}_{\beta_0}^2 + b^2\hat{\sigma}_{\beta_1}^2 + 2ab\hat{\sigma}_{\beta_{01}}.$$

As  $n \rightarrow \infty$ ,  $\hat{\sigma}_{\beta_0}^2 \rightarrow \sigma_{\beta_0}^2$ ,  $\hat{\sigma}_{\beta_1}^2 \rightarrow \sigma_{\beta_1}^2$ , and  $\hat{\sigma}_{\beta_{01}} \rightarrow \sigma_{\beta_{01}}$  by the **Law of Large Numbers**. This gives us that  $\hat{\sigma}_{\lambda}^2 \rightarrow \sigma_{\lambda}^2$  as  $n \rightarrow \infty$  so that we can say (approximately for large  $n$ ):

$$\frac{\hat{\lambda} - \lambda}{\hat{\sigma}_{\lambda}/\sqrt{n}} \sim N(0, 1).$$



As  $n \rightarrow \infty$ ,  $\hat{\sigma}_{\beta_0}^2 \rightarrow \sigma_{\beta_0}^2$ ,  $\hat{\sigma}_{\beta_1}^2 \rightarrow \sigma_{\beta_1}^2$ , and  $\hat{\sigma}_{\beta_{01}} \rightarrow \sigma_{\beta_{01}}$  by the **Law of Large Numbers**. This gives us that  $\hat{\sigma}_{\lambda}^2 \rightarrow \sigma_{\lambda}^2$  as  $n \rightarrow \infty$  so that we can say (approximately for large  $n$ ):

$$\frac{\hat{\lambda} - \lambda}{\hat{\sigma}_{\lambda}/\sqrt{n}} \sim N(0, 1).$$

As when considering just  $\hat{\beta}_0$  or  $\hat{\beta}_1$ , this distributional result will be useful for hypothesis testing and creating confidence intervals.

Using the distributional result:

$$\frac{\hat{\lambda} - \lambda}{\hat{\sigma}_{\lambda}/\sqrt{n}} \sim N(0, 1),$$

we can test a null hypothesis of the form  $H_0 : \lambda \leq \ell$ ,  $H_0 : \lambda \geq \ell$ , or  $H_0 : \lambda = \ell$  by first constructing our test statistic

$$t^* = \frac{\hat{\lambda} - \ell}{\hat{\sigma}_{\lambda}/\sqrt{n}}.$$

Using the distributional result:

$$\frac{\hat{\lambda} - \lambda}{\hat{\sigma}_{\lambda}/\sqrt{n}} \sim N(0, 1),$$

we can test a null hypothesis of the form  $H_0 : \lambda \leq \ell$ ,  $H_0 : \lambda \geq \ell$ , or  $H_0 : \lambda = \ell$  by first constructing our test statistic

$$t^* = \frac{\hat{\lambda} - \ell}{\hat{\sigma}_{\lambda}/\sqrt{n}}.$$

As before, we want to reject our null hypothesis if the probability of obtaining our test statistic (or something even further from the null hypothesis) under the null hypothesis is less than or equal to some pre-specified value  $\lambda$ .

Using the distributional result:

$$\frac{\hat{\lambda} - \lambda}{\hat{\sigma}_{\lambda}/\sqrt{n}} \sim N(0, 1),$$

we can test a null hypothesis of the form  $H_0 : \lambda \leq \ell$ ,  $H_0 : \lambda \geq \ell$ , or  $H_0 : \lambda = \ell$  by first constructing our test statistic

$$t^* = \frac{\hat{\lambda} - \ell}{\hat{\sigma}_{\lambda}/\sqrt{n}}.$$

As before, we want to reject our null hypothesis if the probability of obtaining our test statistic (or something even further from the null hypothesis) under the null hypothesis is less than or equal to some pre-specified value  $\lambda$ .

- Recall that by the distributional result, under the null  $t^* \sim N(0, 1)$
- The quantity  $\hat{\sigma}_{\lambda}/\sqrt{n}$  is called the **standard error** of  $\hat{\lambda}$ .

Now that we have constructed our test statistic  $t^*$  we can conduct our test in two (equivalent) ways, as before

1. Construct a p-value and reject if  $p < \alpha$ :

- If  $H_0 : \lambda \leq \ell$  and  $H_1 : \lambda > \ell$ :

$$p = \Pr(Z \geq t^*).$$

- If  $H_0 : \lambda \geq \ell$  and  $H_1 : \lambda < \ell$ :

$$p = \Pr(Z \leq t^*).$$

- If  $H_0 : \lambda = \ell$  and  $H_1 : \lambda \neq \ell$ :

$$p = \Pr(|Z| \geq |t^*|) = 2 \Pr(Z \geq |t^*|).$$

Now that we have constructed our test statistic  $t^*$  we can conduct our test in two (equivalent) ways, as before

2. Compare the t statistic to the  $1 - \alpha$  or  $1 - \alpha/2$  quantile of the standard normal distribution:  $z_{1-\alpha}$  or  $z_{1-\alpha/2}$ .

- If  $H_0 : \lambda \leq \ell$  and  $H_1 : \lambda > \ell$  reject if

$$t^* \geq z_{1-\alpha}.$$

- If  $H_0 : \lambda \geq \ell$  and  $H_1 : \lambda < \ell$  reject if

$$t^* \leq -z_{1-\alpha}.$$

- If  $H_0 : \lambda = \ell$  and  $H_1 : \lambda \neq \ell$  reject if

$$|t^*| \geq z_{1-\alpha/2}.$$

As a reminder  $z_{1-\alpha}$  and  $z_{1-\alpha/2}$  are such that

$$\Pr(Z \leq z_{1-\alpha}) = 1 - \alpha \iff \Pr(Z > z_{1-\alpha}) = \alpha$$

$$\Pr(Z \leq z_{1-\alpha/2}) = 1 - \alpha/2 \iff \Pr(|Z| > z_{1-\alpha/2}) = \alpha$$

We can also construct a  $100(1 - \alpha)\%$  confidence interval for  $\lambda$  in the same way as before: by looking at the values of  $\ell$  for which we would **fail to reject** the null hypothesis  $H_0 : \lambda = \ell$  against a two-sided alternative  $H_1 : \lambda \neq \ell$  at level  $\alpha$ .

We can also construct a  $100(1 - \alpha)\%$  confidence interval for  $\lambda$  in the same way as before: by looking at the values of  $\ell$  for which we would **fail to reject** the null hypothesis  $H_0 : \lambda = \ell$  against a two-sided alternative  $H_1 : \lambda \neq \ell$  at level  $\alpha$ .

This gives us a symmetric formula as before, a  $100(1 - \alpha)\%$  confidence interval for  $\lambda$  is given

$$\hat{\lambda} \pm z_{1-\alpha/2} \frac{\hat{\sigma}_{\lambda}}{\sqrt{n}}.$$



As an aside, we can start to see a pattern here. Essentially anytime we have a distributional result like

$$\frac{\text{Estimator} - \text{True Value}}{\text{Standard Error of Estimator}} \sim N(0, 1).$$

we can test a null hypothesis by constructing our test statistic

$$t^* = \frac{\text{Estimator} - \text{Null Hypothesis Value}}{\text{Standard Error of Estimate}}.$$

and then computing a  $p$ -value or directly comparing this test statistic to  $z_{1-\alpha}$ ,  $-z_{1-\alpha}$ , or  $z_{1-\alpha/2}$  (depending on what alternate hypothesis we are testing).

We can also use this distributional result to generate  $100(1 - \alpha)\%$  confidence intervals for the true value via

$$\text{Estimator} \pm z_{1-\alpha/2} \cdot \text{Standard Error of Estimator}.$$

Questions?

**Example:** Suppose we are arguing with our professional colleague Kyle Kuzma about the relationship between number of mental health days taken in a month ( $X$ ) and the average number of points per game scored in the NBA ( $Y$ ). Kuzma claims that  $\mathbb{E}[Y|X = 3] = 20$ , we want to test this claim at level  $\alpha = 0.05$ .

**Example:** Suppose we are arguing with our professional colleague Kyle Kuzma about the relationship between number of mental health days taken in a month ( $X$ ) and the average number of points per game scored in the NBA ( $Y$ ). Kuzma claims that  $\mathbb{E}[Y|X = 3] = 20$ , we want to test this claim at level  $\alpha = 0.05$ .

First we collect a random sample of 49 NBA players and ask them how many mental health days they took this month and their average points per game,  $\{Y_i, X_i\}_{i=1}^{49}$ . Then, since we believe the relationship between  $Y$  and  $X$  to be linear, we estimate the linear model

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon.$$

**Example:** Suppose we are arguing with our professional colleague Kyle Kuzma about the relationship between number of mental health days taken in a month ( $X$ ) and the average number of points per game scored in the NBA ( $Y$ ). Kuzma claims that  $\mathbb{E}[Y|X = 3] = 20$ , we want to test this claim at level  $\alpha = 0.05$ .

First we collect a random sample of 49 NBA players and ask them how many mental health days they took this month and their average points per game,  $\{Y_i, X_i\}_{i=1}^{49}$ . Then, since we believe the relationship between  $Y$  and  $X$  to be linear, we estimate the linear model

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon.$$

We can then estimate  $\mathbb{E}[Y|X = 3]$  by  $\hat{\beta}_0 + 3\hat{\beta}_1$ .

We can test Kuzma's claim that  $\mathbb{E}[Y|X = 3] = 20$  by running the following hypothesis test

$$H_0 : \beta_0 + 3\beta_1 = 20 \quad \text{vs.} \quad H_1 : \beta_0 + 3\beta_1 \neq 20.$$

We can test Kuzma's claim that  $\mathbb{E}[Y|X = 3] = 20$  by running the following hypothesis test

$$H_0 : \beta_0 + 3\beta_1 = 20 \text{ vs. } H_1 : \beta_0 + 3\beta_1 \neq 20.$$

To test this claim we use our data (with  $n = 49$ ) to estimate

$$\begin{aligned}\hat{\beta}_0 &= 10, & \hat{\beta}_1 &= 3 \\ \hat{\sigma}_{\beta_0}^2 &= \hat{\sigma}_{\beta_1}^2 = \hat{\sigma}_{\beta_{01}}^2 = 1\end{aligned}$$



We can test Kuzma's claim that  $\mathbb{E}[Y|X = 3] = 20$  by running the following hypothesis test

$$H_0 : \beta_0 + 3\beta_1 = 20 \quad \text{vs.} \quad H_1 : \beta_0 + 3\beta_1 \neq 20.$$

To test this claim we use our data (with  $n = 49$ ) to estimate

$$\begin{aligned}\hat{\beta}_0 &= 10, & \hat{\beta}_1 &= 3 \\ \hat{\sigma}_{\beta_0}^2 &= \hat{\sigma}_{\beta_1}^2 = \hat{\sigma}_{\beta_{01}} = 1\end{aligned}$$

Using these estimates we get

$$\begin{aligned}\hat{\lambda} &= \hat{\beta}_0 + 3\hat{\beta}_1 = 19 \\ \hat{\sigma}_{\lambda}^2 &= \hat{\sigma}_{\beta_0}^2 + 9\hat{\sigma}_{\beta_1}^2 + 6\hat{\sigma}_{\beta_{01}} = 16\end{aligned}$$

We can test Kuzma's claim that  $\mathbb{E}[Y|X = 3] = 20$  by running the following hypothesis test

$$H_0 : \beta_0 + 3\beta_1 = 20 \quad \text{vs.} \quad H_1 : \beta_0 + 3\beta_1 \neq 20.$$

To test this claim we use our data (with  $n = 49$ ) to estimate

$$\begin{aligned}\hat{\beta}_0 &= 10, & \hat{\beta}_1 &= 3 \\ \hat{\sigma}_{\beta_0}^2 &= \hat{\sigma}_{\beta_1}^2 = \hat{\sigma}_{\beta_{01}}^2 = 1\end{aligned}$$

Using these estimates we get

$$\begin{aligned}\hat{\lambda} &= \hat{\beta}_0 + 3\hat{\beta}_1 = 19 \\ \hat{\sigma}_{\lambda}^2 &= \hat{\sigma}_{\beta_0}^2 + 9\hat{\sigma}_{\beta_1}^2 + 6\hat{\sigma}_{\beta_{01}}^2 = 16\end{aligned}$$

- Notice how much larger  $\hat{\sigma}_{\lambda}^2$  is than  $\hat{\sigma}_{\beta_0}^2$  or  $\hat{\sigma}_{\beta_1}^2$ .

Using  $\hat{\lambda} = 19$ ,  $\hat{\sigma}_{\lambda}^2 = 16$ , and  $n = 49$  we can construct our test statistic for  $H_0 : \lambda = 20$  vs  $H_1 : \lambda \neq 20$

$$t^* = \frac{\hat{\lambda} - 20}{\hat{\sigma}_{\lambda}/\sqrt{n}} = \frac{19 - 20}{\sqrt{16}/\sqrt{49}} = -\frac{1}{4/7} = -\frac{7}{4} = -1.75.$$

Using  $\hat{\lambda} = 19$ ,  $\hat{\sigma}_{\lambda}^2 = 16$ , and  $n = 49$  we can construct our test statistic for  $H_0 : \lambda = 20$  vs  $H_1 : \lambda \neq 20$

$$t^* = \frac{\hat{\lambda} - 20}{\hat{\sigma}_{\lambda}/\sqrt{n}} = \frac{19 - 20}{\sqrt{16}/\sqrt{49}} = -\frac{1}{4/7} = -\frac{7}{4} = -1.75.$$

We'll run our test in two ways. First, let's compute our p-value

$$p = \Pr(|Z| \geq |-1.75|) = 2 \Pr(Z \geq 1.75) = 2(1 - \Pr(Z \leq 1.75)) = 2 \cdot 0.04 = 0.08.$$

Since  $0.08 > 0.05$  we **fail to reject** Kuzma's claim.

Using  $\hat{\lambda} = 19$ ,  $\hat{\sigma}_{\lambda}^2 = 16$ , and  $n = 49$  we can construct our test statistic for  $H_0 : \lambda = 20$  vs  $H_1 : \lambda \neq 20$

$$t^* = \frac{\hat{\lambda} - 20}{\hat{\sigma}_{\lambda}/\sqrt{n}} = \frac{19 - 20}{\sqrt{16}/\sqrt{49}} = -\frac{1}{4/7} = -\frac{7}{4} = -1.75.$$

We'll run our test in two ways. Next, let's compare our test statistic to  $z_{1-\alpha/2}$ . Since  $\alpha = 2$  we get that  $z_{1-\alpha/2} = z_{0.975} = 1.96$ . Because

$$|t^*| = 1.75 < 1.96 = z_{0.975}$$

we again **fail to reject** Kuzma's claim

Let's use these same estimates,  $\hat{\lambda} = 19$  and  $\hat{\sigma}_{\lambda}^2 = 16$ , to construct a 95% confidence interval for the true parameter  $\lambda = \beta_0 + 3\beta_1$ .

## Inference: Linear Combinations of Parameters

---

Let's use these same estimates,  $\hat{\lambda} = 19$  and  $\hat{\sigma}_{\lambda}^2 = 16$ , to construct a 95% confidence interval for the true parameter  $\lambda = \beta_0 + 3\beta_1$ .

- Since we have assumed that the true relationship between  $Y$  (points per game) and  $X$  (number of mental health days taken per month) is linear then  $\lambda = \mathbb{E}[Y|X = 3]$ .
  - By linear we mean that  $\mathbb{E}[Y|X = x] = \beta_0 + \beta_1 \cdot x$
  - Otherwise we can view  $\lambda = \beta_0 + 3\beta_1$  as an approximation of  $\mathbb{E}[Y|X = 3]$ .

Let's use these same estimates,  $\hat{\lambda} = 19$  and  $\hat{\sigma}_{\lambda}^2 = 16$ , to construct a 95% confidence interval for the true parameter  $\lambda = \beta_0 + 3\beta_1$ .

From above we have that a 95% confidence interval for  $\lambda$  can be constructed

$$\hat{\lambda} \pm z_{0.975} \frac{\hat{\sigma}_{\lambda}}{\sqrt{n}} = 19 \pm 1.96 \frac{7}{4}.$$

So that we are 95% confident that the true value of  $\lambda = \beta_0 + 3\beta_1$  lies in the interval  $[15.57, 22.43]$ .

- How could we use this interval to test the hypothesis  $H_0 : \lambda = 20$  vs  $H_1 : \lambda \neq 20$  at level  $\alpha = 0.05$ ?



Let's use these same estimates,  $\hat{\lambda} = 19$  and  $\hat{\sigma}_{\lambda}^2 = 16$ , to construct a 95% confidence interval for the true parameter  $\lambda = \beta_0 + 3\beta_1$ .

From above we have that a 95% confidence interval for  $\lambda$  can be constructed

$$\hat{\lambda} \pm z_{0.975} \frac{\hat{\sigma}_{\lambda}}{\sqrt{n}} = 19 \pm 1.96 \frac{7}{4}.$$

So that we are 95% confident that the true value of  $\lambda = \beta_0 + 3\beta_1$  lies in the interval  $[15.57, 22.43]$ .

- How could we use this interval to test the hypothesis  $H_0 : \lambda = 20$  vs  $H_1 : \lambda \neq 20$  at level  $\alpha = 0.05$ ?
- What about testing this hypothesis at level  $\alpha = 0.01$ ?

Recall that in order to conduct inference and derive an asymptotic distribution for  $\hat{\beta}_0$  and  $\hat{\beta}_1$  we made the following assumptions about the underlying distribution of  $(Y, X)$ :

- **Random Sampling:** Assume that  $\{Y_i, X_i\}$  are independently and identically distributed;  $(Y_i, X_i) \stackrel{\text{i.i.d}}{\sim} (Y, X)$
- **Homoskedasticity:** Assume that  $\text{Var}(\epsilon \mid X = x) = \sigma_\epsilon^2$  for all possible values of  $x$ .
- **Rank Condition:** There must be at least two distinct values of  $X$  that appear in the population.

Recall that in order to conduct inference and derive an asymptotic distribution for  $\hat{\beta}_0$  and  $\hat{\beta}_1$  we made the following assumptions about the underlying distribution of  $(Y, X)$ :

- **Random Sampling:** Assume that  $\{Y_i, X_i\}$  are independently and identically distributed;  $(Y_i, X_i) \stackrel{\text{i.i.d}}{\sim} (Y, X)$
- **Homoskedasticity:** Assume that  $\text{Var}(\epsilon \mid X = x) = \sigma_\epsilon^2$  for all possible values of  $x$ .
- **Rank Condition:** There must be at least two distinct values of  $X$  that appear in the population.

In the last set of slides we mentioned that Homoskedasticity was a strong assumption that we will want to relax. Let's look more into why this is the case and how to relax the assumption.

Since our linear model relates  $Y$  and  $X$  via the following equation

$$Y = \beta_0 + \beta_1 X + \epsilon$$

the requirement that  $\text{Var}(\epsilon|X = x) = \sigma_\epsilon^2$  for all  $x \in \text{supp}(X)$  is implicitly requiring that  $\text{Var}(Y|X = x) = \text{Var}(\epsilon|X = x) = \sigma_\epsilon^2$  for all  $x \in \text{supp}(X)$ .

Since our linear model relates  $Y$  and  $X$  via the following equation

$$Y = \beta_0 + \beta_1 X + \epsilon$$

the requirement that  $\text{Var}(\epsilon|X = x) = \sigma_\epsilon^2$  for all  $x \in \text{supp}(X)$  is implicitly requiring that  $\text{Var}(Y|X = x) = \text{Var}(\epsilon|X = x) = \sigma_\epsilon^2$  for all  $x \in \text{supp}(X)$ .

- Intuitively this means that  $X$  has no information about the spread of  $Y$ .
- The variance of  $Y$  is the same for all values of  $X$ .

Since our linear model relates  $Y$  and  $X$  via the following equation

$$Y = \beta_0 + \beta_1 X + \epsilon$$

the requirement that  $\text{Var}(\epsilon|X = x) = \sigma_\epsilon^2$  for all  $x \in \text{supp}(X)$  is implicitly requiring that  $\text{Var}(Y|X = x) = \text{Var}(\epsilon|X = x) = \sigma_\epsilon^2$  for all  $x \in \text{supp}(X)$ .

- Intuitively this means that  $X$  has no information about the spread of  $Y$ .
- The variance of  $Y$  is the same for all values of  $X$ .

To see why homoskedasticity is a restrictive assumption, let's turn to some examples.

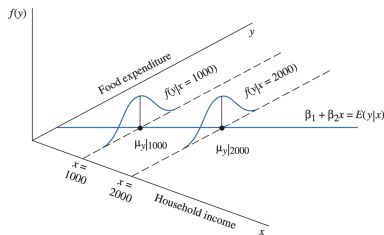
- If homoskedasticity is violated we say that the errors  $\epsilon$  exhibit **heteroskedasticity**

## Inference: Heteroskedasticity

**Example 1:** Let  $Y$  be food expenditure and  $X$  be household income and suppose we want to estimate the linear model.

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

Homoskedasticity requires that the variance of food expenditures is the same for all levels of income.



**Example 1:** Let  $Y$  be food expenditure and  $X$  be household income and suppose we want to estimate the linear model.

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

Homoskedasticity requires that the variance of food expenditures is the same for all levels of income.

Is this a reasonable assumption?



**Example 1:** Let  $Y$  be food expenditure and  $X$  be household income and suppose we want to estimate the linear model.

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

Homoskedasticity requires that the variance of food expenditures is the same for all levels of income.

Is this a reasonable assumption?

- Low income individuals don't have much choice on how much to spend on food

**Example 1:** Let  $Y$  be food expenditure and  $X$  be household income and suppose we want to estimate the linear model.

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

Homoskedasticity requires that the variance of food expenditures is the same for all levels of income.

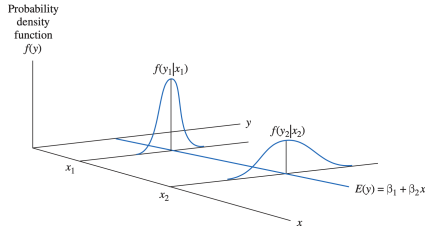
Is this a reasonable assumption?

- Low income individuals don't have much choice on how much to spend on food
- High income individuals may have high variance reflecting variance in taste

**Example 1:** Let  $Y$  be food expenditure and  $X$  be household income and suppose we want to estimate the linear model.

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

Homoskedasticity requires that the variance of food expenditures is the same for all levels of income. Instead we may expect the spread of food expenditure to depend on income



## Inference: Heteroskedasticity

---

**Example 1:** Let  $Y$  be food expenditure and  $X$  be household income and suppose we want to estimate the linear model.

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

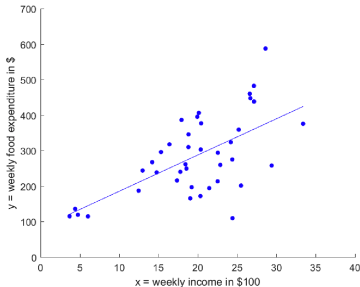
After estimating  $\hat{\beta}_0$  and  $\hat{\beta}_1$  we can check for heteroskedasticity by plotting the estimated residuals against  $X$ .

## Inference: Heteroskedasticity

**Example 1:** Let  $Y$  be food expenditure and  $X$  be household income and suppose we want to estimate the linear model.

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

After estimating  $\hat{\beta}_0$  and  $\hat{\beta}_1$  we can check for heteroskedasticity by plotting the estimated residuals against  $X$ . In this case that looks like

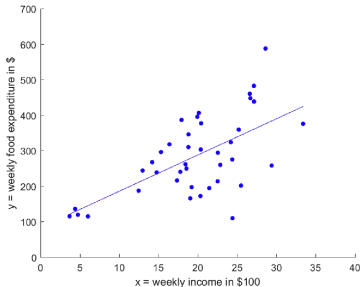


## Inference: Heteroskedasticity

**Example 1:** Let  $Y$  be food expenditure and  $X$  be household income and suppose we want to estimate the linear model.

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

After estimating  $\hat{\beta}_0$  and  $\hat{\beta}_1$  we can check for heteroskedasticity by plotting the estimated residuals against  $X$ . In this case that looks like



- Residuals look more spread out for higher income levels, suggests homoskedasticity is violated
- Not a formal test (using  $\hat{\epsilon}$  instead of  $\epsilon$ ), but suggestive

**Example 2:** Let  $Y$  be wages and  $X$  be years of education and consider the model

$$Y = \beta_1 + \beta_2 X + \epsilon.$$

In this context homoskedasticity requires that the variance in wages is the same for all levels of education.

**Example 2:** Let  $Y$  be wages and  $X$  be years of education and consider the model

$$Y = \beta_1 + \beta_2 X + \epsilon.$$

In this context homoskedasticity requires that the variance in wages is the same for all levels of education. Is this a reasonable assumption?

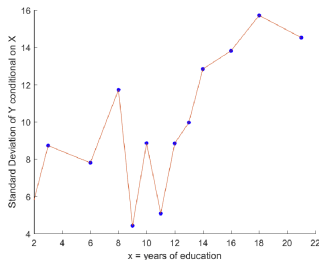
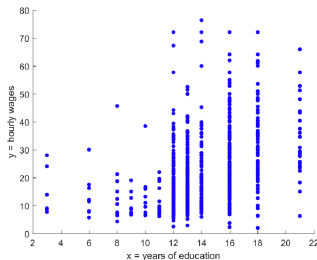
- High school graduates may not have access to as many career paths
- College graduates can range from English PhDs to engineers at Google



**Example 2:** Let  $Y$  be wages and  $X$  be years of education and consider the model

$$Y = \beta_1 + \beta_2 X + \epsilon.$$

In this context homoskedasticity requires that the variance in wages is the same for all levels of education.



- Looking at the data we see that variance in wages increases after high school

Questions?

So, suppose we look at our data and suspect homoskedasticity is violated. What do we do now?

- Won't need to adjust our estimator

So, suppose we look at our data and suspect homoskedasticity is violated. What do we do now?

Recall that when we were deriving the asymptotic distribution of our estimator  $\hat{\beta}_1$  (the approximate distribution for  $n$  large) we applied **law of large numbers** and found that approximately for large  $n$ :

$$\sqrt{n} \left( \hat{\beta}_1 - \beta_1 \right) \approx \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i (X_i - \mu_X)}{\sigma_X^2}.$$

So, suppose we look at our data and suspect homoskedasticity is violated. What do we do now?

Recall that when we were deriving the asymptotic distribution of our estimator  $\hat{\beta}_1$  (the approximate distribution for  $n$  large) we applied **law of large numbers** and found that approximately for large  $n$ :

$$\sqrt{n} \left( \hat{\beta}_1 - \beta_1 \right) \approx \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i (X_i - \mu_X)}{\sigma_X^2}.$$

We then applied the **central limit theorem** to  $\frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i (X_i - \mu_X)$ : approximately for large  $n$ ,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i (X_i - \mu_X) \sim N \left( 0, \text{Var}(\epsilon(X - \mu_X)) \right).$$

So, suppose we look at our data and suspect homoskedasticity is violated. What do we do now?

Recall that when we were deriving the asymptotic distribution of our estimator  $\hat{\beta}_1$  (the approximate distribution for  $n$  large) we applied **law of large numbers** and found that approximately for large  $n$ :

$$\sqrt{n} \left( \hat{\beta}_1 - \beta_1 \right) \approx \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i (X_i - \mu_X)}{\sigma_X^2}.$$

We then applied the **central limit theorem** to  $\frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i (X_i - \mu_X)$ : approximately for large  $n$ ,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i (X_i - \mu_X) \sim N \left( 0, \text{Var}(\epsilon(X - \mu_X)) \right).$$

- The only time we used homoskedasticity was to decompose  $\text{Var}(\epsilon(X - \mu_X)) = \sigma_\epsilon^2 \sigma_X^2$ . This simplified estimation but is not necessary.

Without heteroskedasticity we can use the same logic as before without decomposing  $\text{Var}(\epsilon(X - \mu_X)) = \sigma_\epsilon^2 \sigma_X^2$  to get

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \sim N\left(0, \frac{\text{Var}(\epsilon(X - \mu_X))}{(\sigma_X^2)^2}\right) \implies \hat{\beta}_1 \sim N\left(\beta_1, \frac{\text{Var}(\epsilon(X - \mu_X))}{n(\sigma_X^2)^2}\right).$$

Without heteroskedasticity we can use the same logic as before without decomposing  $\text{Var}(\epsilon(X - \mu_X)) = \sigma_\epsilon^2 \sigma_X^2$  to get

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \sim N\left(0, \frac{\text{Var}(\epsilon(X - \mu_X))}{(\sigma_X^2)^2}\right) \implies \hat{\beta}_1 \sim N\left(\beta_1, \frac{\text{Var}(\epsilon(X - \mu_X))}{n(\sigma_X^2)^2}\right).$$



Without heteroskedasticity we can use the same logic as before without decomposing  $\text{Var}(\epsilon(X - \mu_X)) = \sigma_\epsilon^2 \sigma_X^2$  to get

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \sim N\left(0, \frac{\text{Var}(\epsilon(X - \mu_X))}{(\sigma_X^2)^2}\right) \implies \hat{\beta}_1 \sim N\left(\beta_1, \frac{\text{Var}(\epsilon(X - \mu_X))}{n(\sigma_X^2)^2}\right).$$

- Note that the variance still goes to 0 as  $n \rightarrow \infty$  so that  $\hat{\beta}_1 \rightarrow \beta_1$  as  $n \rightarrow \infty$ .

Without heteroskedasticity we can use the same logic as before without decomposing  $\text{Var}(\epsilon(X - \mu_X)) = \sigma_\epsilon^2 \sigma_X^2$  to get

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \sim N\left(0, \frac{\text{Var}(\epsilon(X - \mu_X))}{(\sigma_X^2)^2}\right) \implies \hat{\beta}_1 \sim N\left(\beta_1, \frac{\text{Var}(\epsilon(X - \mu_X))}{n(\sigma_X^2)^2}\right).$$

- Note that the variance still goes to 0 as  $n \rightarrow \infty$  so that  $\hat{\beta}_1 \rightarrow \beta_1$  as  $n \rightarrow \infty$ .
- Asymptotic variance is different now however, and will require a different estimator.

Without heteroskedasticity we can use the same logic as before without decomposing  $\text{Var}(\epsilon(X - \mu_X)) = \sigma_\epsilon^2 \sigma_X^2$  to get

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \sim N\left(0, \frac{\text{Var}(\epsilon(X - \mu_X))}{(\sigma_X^2)^2}\right) \implies \hat{\beta}_1 \sim N\left(\beta_1, \frac{\text{Var}(\epsilon(X - \mu_X))}{n(\sigma_X^2)^2}\right).$$

- Note that the variance still goes to 0 as  $n \rightarrow \infty$  so that  $\hat{\beta}_1 \rightarrow \beta_1$  as  $n \rightarrow \infty$ .
- Asymptotic variance is different now however, and will require a different estimator.
- Using the wrong variance renders our inference useless as we will not be accurately computing objects like

$$\Pr(|\hat{\beta}_1| > 5|\beta_1 = 0).$$

As mentioned above, all that needs to be done to relax homoskedasticity is find a way to estimate the asymptotic variance of  $\hat{\beta}_1$ :

$$\frac{\text{Var}(\epsilon(X - \mu_X))}{(\sigma_X^2)^2}.$$

As mentioned above, all that needs to be done to relax homoskedasticity is find a way to estimate the asymptotic variance of  $\hat{\beta}_1$ :

$$\frac{\text{Var}(\epsilon(X - \mu_X))}{(\sigma_X^2)^2}.$$

- Already know how to estimate  $\sigma_X^2$

As mentioned above, all that needs to be done to relax homoskedasticity is find a way to estimate the asymptotic variance of  $\hat{\beta}_1$ :

$$\frac{\text{Var}(\epsilon(X - \mu_X))}{(\sigma_X^2)^2}.$$

- Already know how to estimate  $\sigma_X^2$
- To estimate  $\text{Var}(\epsilon(X - \mu_X))$  let  $\hat{W}_i = \hat{\epsilon}_i(X_i - \bar{X})$ .

As mentioned above, all that needs to be done to relax homoskedasticity is find a way to estimate the asymptotic variance of  $\hat{\beta}_1$ :

$$\frac{\text{Var}(\epsilon(X - \mu_X))}{(\sigma_X^2)^2}.$$

- Already know how to estimate  $\sigma_X^2$
- To estimate  $\text{Var}(\epsilon(X - \mu_X))$  let  $\hat{W}_i = \hat{\epsilon}_i(X_i - \bar{X})$ .
  - Note that  $\frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_i X_i = \bar{X} \frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_i = 0$  by the first order conditions for  $\hat{\beta}_1$  and  $\hat{\beta}_0$ , respectively. So  $\widehat{\overline{W}}_i = \frac{1}{n} \sum_{i=1}^n \hat{W}_i = 0$ .

As mentioned above, all that needs to be done to relax homoskedasticity is find a way to estimate the asymptotic variance of  $\hat{\beta}_1$ :

$$\frac{\text{Var}(\epsilon(X - \mu_X))}{(\sigma_X^2)^2}.$$

- Already know how to estimate  $\sigma_X^2$
- To estimate  $\text{Var}(\epsilon(X - \mu_X))$  let  $\hat{W}_i = \hat{\epsilon}_i(X_i - \bar{X})$ .
  - Note that  $\frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_i X_i = \bar{X} \frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_i = 0$  by the first order conditions for  $\hat{\beta}_1$  and  $\hat{\beta}_0$ , respectively. So  $\widehat{\bar{W}_i} = \frac{1}{n} \sum_{i=1}^n \hat{W}_i = 0$ .
  - Can then estimate  $\text{Var}(\epsilon(X - \mu_X))$  via

$$\frac{1}{n} \sum_{i=1}^n \hat{W}_i^2.$$



As mentioned above, all that needs to be done to relax homoskedasticity is find a way to estimate the asymptotic variance of  $\hat{\beta}_1$ :

$$\frac{\text{Var}(\epsilon(X - \mu_X))}{(\sigma_X^2)^2}.$$

- Already know how to estimate  $\sigma_X^2$
- To estimate  $\text{Var}(\epsilon(X - \mu_X))$  let  $\hat{W}_i = \hat{\epsilon}_i(X_i - \bar{X})$ .
  - Note that  $\frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_i X_i = \bar{X} \frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_i = 0$  by the first order conditions for  $\hat{\beta}_1$  and  $\hat{\beta}_0$ , respectively. So  $\widehat{\bar{W}}_i = \frac{1}{n} \sum_{i=1}^n \hat{W}_i = 0$ .
  - Can then estimate  $\text{Var}(\epsilon(X - \mu_X))$  via

$$\frac{1}{n} \sum_{i=1}^n \hat{W}_i^2.$$

- Since  $\hat{\epsilon}_i \rightarrow \epsilon_i$  and  $\bar{X} \rightarrow \mu_X$ ,  $\hat{W}_i \rightarrow \epsilon_i(X - \mu_X)$  and so we have a consistent estimator for  $\text{Var}(\epsilon(X - \mu_X))$  by **law of large numbers**.

To summarize under heteroskedasticity we have that (approximately for large  $n$ )

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \sim N(0, \sigma_{\beta_1}^2) \implies \hat{\beta}_1 \sim N(\beta_1, \sigma_{\beta_1}^2/n),$$

where  $\sigma_{\beta_1}^2 = \text{Var}(\epsilon(X - \mu_X))/(\sigma_X^2)^2$ .

To summarize under heteroskedasticity we have that (approximately for large  $n$ )

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \sim N(0, \sigma_{\beta_1}^2) \implies \hat{\beta}_1 \sim N(\beta_1, \sigma_{\beta_1}^2/n),$$

where  $\sigma_{\beta_1}^2 = \text{Var}(\epsilon(X - \mu_X))/(\sigma_X^2)^2$ .

- This is a different expression for  $\sigma_{\beta_1}^2$  than under homoskedasticity and involves a somewhat more complicated estimation procedure.

To summarize under heteroskedasticity we have that (approximately for large  $n$ )

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \sim N(0, \sigma_{\beta_1}^2) \implies \hat{\beta}_1 \sim N(\beta_1, \sigma_{\beta_1}^2/n),$$

where  $\sigma_{\beta_1}^2 = \text{Var}(\epsilon(X - \mu_X))/(\sigma_X^2)^2$ .

- This is a different expression for  $\sigma_{\beta_1}^2$  than under homoskedasticity and involves a somewhat more complicated estimation procedure.
- From now on we will generally assume heteroskedasticity. It is easy to let the computer handle estimation of the variance  $\sigma_{\beta_1}^2$  and the standard error  $\sigma_{\beta_1}/\sqrt{n}$ .

To summarize under heteroskedasticity we have that (approximately for large  $n$ )

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \sim N(0, \sigma_{\beta_1}^2) \implies \hat{\beta}_1 \sim N(\beta_1, \sigma_{\beta_1}^2/n),$$

where  $\sigma_{\beta_1}^2 = \text{Var}(\epsilon(X - \mu_X))/(\sigma_X^2)^2$ .

- This is a different expression for  $\sigma_{\beta_1}^2$  than under homoskedasticity and involves a somewhat more complicated estimation procedure.
- From now on we will generally assume heteroskedasticity. It is easy to let the computer handle estimation of the variance  $\sigma_{\beta_1}^2$  and the standard error  $\sigma_{\beta_1}/\sqrt{n}$ .
- Formulas for variance of  $\hat{\beta}_0$  and the covariance between  $\hat{\beta}_1$  and  $\hat{\beta}_0$  will also be different. Again computer can handle these easily.

To summarize under heteroskedasticity we have that (approximately for large  $n$ )

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \sim N(0, \sigma_{\beta_1}^2) \implies \hat{\beta}_1 \sim N(\beta_1, \sigma_{\beta_1}^2/n),$$

where  $\sigma_{\beta_1}^2 = \text{Var}(\epsilon(X - \mu_X))/(\sigma_X^2)^2$ .

- This is a different expression for  $\sigma_{\beta_1}^2$  than under homoskedasticity and involves a somewhat more complicated estimation procedure.
- From now on we will generally assume heteroskedasticity. It is easy to let the computer handle estimation of the variance  $\sigma_{\beta_1}^2$  and the standard error  $\sigma_{\beta_1}/\sqrt{n}$ .
- Formulas for variance of  $\hat{\beta}_0$  and the covariance between  $\hat{\beta}_1$  and  $\hat{\beta}_0$  will also be different. Again computer can handle these easily.
- Once the heteroskedastic-consistent variances/standard errors/covariances are computed inference and confidence intervals are computed the same as before.

Questions?

# Table of Contents

---

Advanced Inference Topics

Evaluating our Model

Modeling Choices



Let's switch tasks a bit and turn from inference to evaluating our model. Recall that we chose to model the relationship between  $Y$  and  $X$  by estimating the line of best fit between  $Y$  and  $X$ :

$$\beta_0, \beta_1 = \arg \min_{\tilde{\beta}_0, \tilde{\beta}_1} \mathbb{E} \left[ (Y - \tilde{\beta}_0 - \tilde{\beta}_1 X)^2 \right].$$

We then spent the next few days conducting inference on the parameters of interest  $\beta_0$  and  $\beta_1$ .

Let's switch tasks a bit and turn from inference to evaluating our model. Recall that we chose to model the relationship between  $Y$  and  $X$  by estimating the line of best fit between  $Y$  and  $X$ :

$$\beta_0, \beta_1 = \arg \min_{\tilde{\beta}_0, \tilde{\beta}_1} \mathbb{E} \left[ (Y - \tilde{\beta}_0 - \tilde{\beta}_1 X)^2 \right].$$

We then spent the next few days conducting inference on the parameters of interest  $\beta_0$  and  $\beta_1$ .

Now, however, let's consider a different question. Is this even a good model?

## Evaluating our Model: Prediction

---

When we are answering this question we are essentially interested in how our model does as a tool to **predict**  $Y$  using  $X$ .

When we are answering this question we are essentially interested in how our model does as a tool to **predict**  $Y$  using  $X$ .

This is a somewhat different goal than in **inference** when we are interested in **interpreting** the parameters  $\beta_0$  and  $\beta_1$  to learn about the underlying relationship between  $Y$  and  $X$ .

- Inference tends to be useful when thinking about policies to implement, want to know about average effects

Prediction can also be very useful though. Let's consider some examples.

Prediction can also be very useful though. Let's consider some examples.

**Example:** How can Amazon do two day delivery?

- In order to deliver an item within two days, an item has to be in stock in a warehouse nearby
- If too many people buy the item at once in a certain area the warehouse will run out of stock
- Amazon has to be able to accurately **forecast/predict** demand in certain areas.

Prediction can also be very useful though. Let's consider some examples.

**Example:** How much power should the energy grid generate?

- Energy grid must have enough supply to meet demand
- But it is costly and takes time to adjust production levels
- Energy suppliers must have good forecasts of demand

Prediction can also be very useful though. Let's consider some examples.

**Example:** How much should a pension fund keep in liquid funds?

- Fund must be able to pay all it's obligations
- Some portion of funds are invested and returns are random. People also retire/die at random times.
- Pension fund must forecast both obligations and returns.



Prediction can also be very useful though. Let's consider some examples.

**Example:** How much should a pension fund keep in liquid funds?

- Fund must be able to pay all its obligations
- Some portion of funds are invested and returns are random. People also retire/die at random times.
- Pension fund must forecast both obligations and returns.

Note that in these examples we are not per-se interested in interpreting the parameters of our linear regression model. We are just interested in using our linear regression model to predict  $Y$  using  $X$ .

Suppose we had no information on  $X$ . What is the best we can do in terms of predicting  $Y$ ?

- If we just have information on income and we are given a random name, what should we predict their income to be?

Suppose we had no information on  $X$ . What is the best we can do in terms of predicting  $Y$ ?

- If we just have information on income and we are given a random name, what should we predict their income to be?

Since we have no additional information, we will make the same prediction for all new observations. Want to choose a value  $a^*$  that minimizes

$$a^* = \arg \min_a \mathbb{E}[(Y - a)^2].$$

That is  $a^*$  is “closest” to  $Y$  on average.

Taking first order conditions, we see that  $a^*$  solves

$$-2\mathbb{E}[(Y - a^*)] = 0 \implies a^* = \mathbb{E}[Y].$$

Taking first order conditions, we see that  $a^*$  solves

$$-2\mathbb{E}[(Y - a^*)] = 0 \implies a^* = \mathbb{E}[Y].$$

The best predictor of  $Y$  with no additional information is just  $\mathbb{E}[Y]$ !

- This is intuitive enough

Now that we have information on  $X$  we have tried to use this information to predict  $Y$  by estimating a line of best fit (linear model) between  $Y$  and  $X$ :

$$\beta_0, \beta_1 = \arg \min_{\tilde{\beta}_0, \tilde{\beta}_1} \mathbb{E} \left[ (Y - \tilde{\beta}_0 - \tilde{\beta}_1 X)^2 \right].$$

Now that we have information on  $X$  we have tried to use this information to predict  $Y$  by estimating a line of best fit (linear model) between  $Y$  and  $X$ :

$$\beta_0, \beta_1 = \arg \min_{\tilde{\beta}_0, \tilde{\beta}_1} \mathbb{E} \left[ (Y - \tilde{\beta}_0 - \tilde{\beta}_1 X)^2 \right].$$

**Question:** How much better is this linear model at predicting  $Y$  than just using  $\mathbb{E}[Y]$ ?

Now that we have information on  $X$  we have tried to use this information to predict  $Y$  by estimating a line of best fit (linear model) between  $Y$  and  $X$ :

$$\beta_0, \beta_1 = \arg \min_{\tilde{\beta}_0, \tilde{\beta}_1} \mathbb{E} \left[ (Y - \tilde{\beta}_0 - \tilde{\beta}_1 X)^2 \right].$$

**Question:** How much better is this linear model at predicting  $Y$  than just using  $\mathbb{E}[Y]$ ?

- Obviously cannot evaluate this directly since we don't know  $\beta_0, \beta_1$  and  $\mathbb{E}[Y]$



Now that we have information on  $X$  we have tried to use this information to predict  $Y$  by estimating a line of best fit (linear model) between  $Y$  and  $X$ :

$$\beta_0, \beta_1 = \arg \min_{\tilde{\beta}_0, \tilde{\beta}_1} \mathbb{E} \left[ (Y - \tilde{\beta}_0 - \tilde{\beta}_1 X)^2 \right].$$

**Question:** How much better is this linear model at predicting  $Y$  than just using  $\mathbb{E}[Y]$ ?

- Obviously cannot evaluate this directly since we don't know  $\beta_0, \beta_1$  and  $\mathbb{E}[Y]$
- Instead we will see how much closer  $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_1 X_i$  is to  $Y_i$  on average than  $\bar{Y}$ .

## Evaluating our Model: $R^2$ and Goodness of Fit

---

# Table of Contents

---

Advanced Inference Topics

Evaluating our Model

Modeling Choices