Econ 103: Topics in Single Linear Regression

Manu Navjeevan

UCLA

August 18, 2021

Content Outline

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

Inference: Linear Combinations of Parameters

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

These results were also often presented in the following equivalent manners

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

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

Inference: Linear Combinations of Parameters

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

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

Inference: Linear Combinations of Parameters

However, often we want to preform 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 $\mathbb{E}[Y|X = x]$.

Inference: Linear Combinations of Parameters

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

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

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

Inference: Linear Combinations of Parameters

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

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

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:

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

where $\sigma_\lambda^2 = a^2\sigma_{\beta_0}^2 + b^2\sigma_{\beta_1}^2 + 2ab\sigma_{\beta_{01}}$

Inference: Linear Combinations of Parameters

As a reminder, we can estimate

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

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

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

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

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

Inference: Linear Combinations of Parameters

As $n \to \infty$, $\hat{\sigma}^2_{\beta_0} \to \sigma^2_{\beta_0}$, $\hat{\sigma}^2_{\beta_1} \to \sigma^2_{\beta_1}$, and $\hat{\sigma}_{\beta_{01}} \to \sigma_{\beta_{01}}$ by the Law of Large Numbers. This gives us that $\hat{\sigma}^2_{\lambda} \to \sigma^2_{\lambda}$ as $n \to \infty$ 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.

Inference: Linear Combinations of Parameters

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 α.

- 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 $\hat{\lambda}$.

Inference: Linear Combinations of Parameters

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 $H_1 : \lambda \neq \ell$:
   $$p = \Pr(|Z| \geq |t^*|) = 2\Pr(Z \geq |t^*|).$$
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
   $$|t^*| \geq z_{1-\alpha/2}.$$

Inference: Linear Combinations of Parameters

We can also construct a $100(1-\alpha)\%$ confidence interval for λ in the same way as before: by looking at the values of ℓ 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 α.

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

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

Inference: Linear Combinations of Parameters

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 compating this test statistic to $z_{1-\alpha}$, $-z_{1-\alpha}$, or $z_{1-\alpha/2}$ (depending on what alternate hypothesis we are testing).

Inference: Linear Combinations of Parameters

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

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

Linear Combinations of Parameters: Questions

Questions?

Inference: Linear Combinations of Parameters

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 $\hat{\beta}_0 + 3\hat{\beta}_1$.

Inference: Linear Combinations of Parameters

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

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

Using these estimates we get

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

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

Inference: Linear Combinations of Parameters

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.

We'll run our test in two ways. Next, let's compare our test statistic to $z_{1-\alpha/2}$. Since $\alpha = 0.05$ 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

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 $\mathbb{E}[Y|X=3]$.

From above we have that a 95% confidence interval for λ can be constructed

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

So that we are 95% confident that the true value of $\lambda = \beta_0 + 3\beta_1$ lies in the inteval $[17.88, 20.12]$.

- 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 $\alpha = 0.01$?

Linear Combinations of Parameters: Questions

Questions?

Inference: Heteroskedasticity

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; $(Y_i, X_i) \overset{\text{i.i.d}}{\sim} (Y, X)$
- Homoskedasticity: Assume that $\mathrm{Var}(\epsilon \mid X = x) = \sigma_\epsilon^2$ 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.

Inference: Heteroskedasticity

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

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

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

- 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.

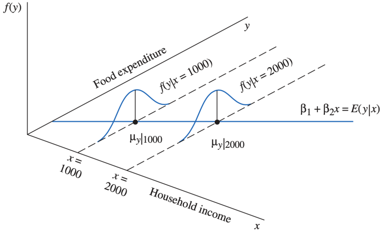
- If homoskedasticity is violated we say that the errors ϵ exhibit heteroskedasticity

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.



Is this a reasonable assumption?

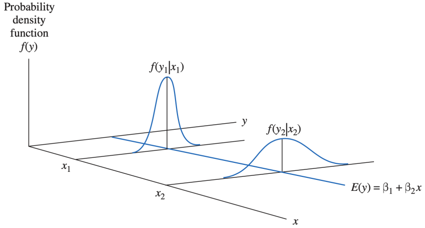
- Low income individuals don't have much choice on how much to spend on food
- High income individuals may have high variance reflecting variance in taste

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. Instead we may expect the spread of food expenditure to depend on income

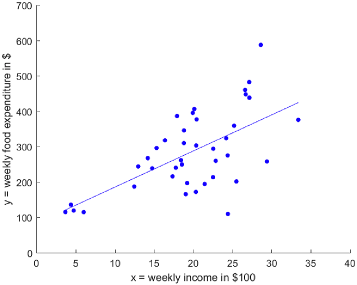

Probability density function $f(y)$
$f(y_1|x_1)$
$f(y_2|x_2)$
y
x_1
x_2
x
$E(y) = \beta_1 + \beta_2 x$

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 againt 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 ϵ), but suggestive

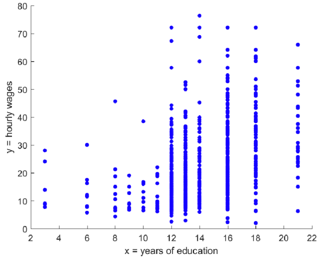
Inference: Heteroskedasticity

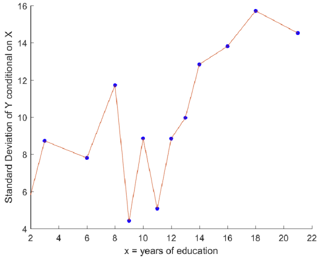
Example 2: Let Y be wages and X be years of education and consider the model

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

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





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

Heteroskedasticity: Questions

Questions?

Inference: Heteroskedasticity

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

- Won't need to adjust our estimator

Recall that when 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,

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

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

Inference: Heteroskedasticity

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

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

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

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

Inference: Heteroskedasticity

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

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

- Already know how to estimate σ_X^2
- To estimate $\mathrm{Var}(\epsilon(X-\mu_X))$ 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 $\overline{\widehat{W}_i} = \frac{1}{n}\sum_{i=1}^n \hat{W}_i = 0$.
  - Can then estimate $\mathrm{Var}(\epsilon(X-\mu_X))$ via

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

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

Inference: Heteroskedasticity

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

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

where $\sigma^2_{\beta_1} = \mathrm{Var}(\epsilon(X - \mu_X))/(\sigma^2_X)^2$.

- This is a different expression for $\sigma^2_{\beta_1}$ 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 $\sigma^2_{\beta_1}$ 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.

Heteroskedasticity: Questions

Questions?

Table of Contents

Advanced Inference Topics

Evaluating our Model

Modeling Choices

Evaluating our Model: Prediction

Let's switch tacks 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:

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

We then spent the next few days conducting inference on the parameters of interest β_0 and β_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.

This is a somewhat different goal than in inference when we are interested in interpreting the parameters β_0 and β_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

Evaluating our Model: Prediction

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.

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

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.

Evaluating our Model: Prediction

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.

Evaluating our Model: Prediction

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

Evaluating our Model: Prediction

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 β_0, β_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

Let's recall the following equalities implied by the first order conditions of our estimating equations

$$\hat{\beta}_0, \hat{\beta}_1 = \arg\min_{b_0, b_1} \frac{1}{n}\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2.$$

From the first order condition for $\hat{\beta}_0$:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i = 0.$$

From the first order condition for $\hat{\beta}_1$:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)X_i = \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i X_i = 0.$$

Evaluating our Model: R^2 and Goodness of Fit

Recall that, by definition of $\hat{\epsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\epsilon}_i = \hat{Y}_i + \hat{\epsilon}_i.$$

and that after solving for $\hat{\beta}_0$ we get

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \implies \bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}.$$

Using these we get that

$$\begin{aligned} Y_i - \bar{Y} &= (\hat{Y}_i - \bar{Y}) + \hat{\epsilon}_i \\ \hat{Y}_i - \bar{Y} &= \hat{\beta}_1 (X_i - \bar{X}) \end{aligned}$$

Evaluating our Model: R^2 and Goodness of Fit

Now let's use these and decompose the sum:

$$\overbrace{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}^{\text{Total variance in } Y} = \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2 + \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})\hat{\epsilon}_i + \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2 + \hat{\beta}_1 \overbrace{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})\hat{\epsilon}_i}^{=0 \text{ by FOCs}} + \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2$$

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}_{\text{Explained by model with } X} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2}_{\text{unexplained by model}}$$

Evaluating our Model: R^2 and Goodness of Fit

Often we multiply both sides of the last equation by n and label the resulting components:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}\hat{\epsilon}_i^2.$$

1. $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$: SST (Total Sum Of Squares)
   - Captures how much total variation there is in Y.
   - Can think of this as the sum of squared errors from just using $\bar{Y}$ to predict Y.
   - Note that this is just the sample variance of Y multiplied by n
2. $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$: SSR (Sum of Squares due to Regression)
   - Captures variation that can be attributed to variation in our predictions
   - This is the sample variance of $\hat{Y}_i$ multiplied by n
3. $\sum_{i=1}^{n}\hat{\epsilon}_i^2$: SSE (Sum of Squared Errors)
   - Variation that is "left over" after using the linear model
   - Note that this is just $\hat{\sigma}_\epsilon^2$ times n

Evaluating our Model: R^2 and Goodness of Fit

Using the decomposition

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n}\hat{\epsilon}_i^2}_{\text{SSE}}.$$

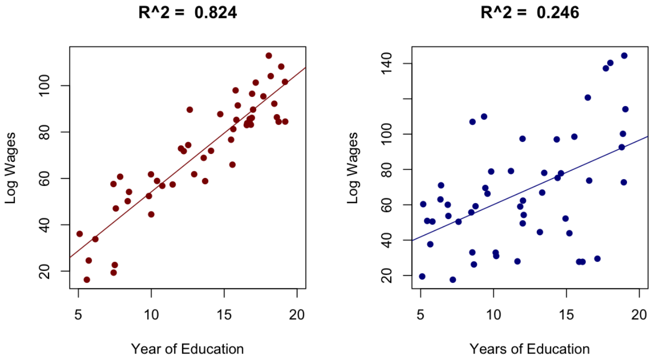
we define the coefficient of determination, R^2, as

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}.$$

- Intuitively R^2 reports what proportion of the total variance in Y can be explained by our linear model with X.
- If $R^2 = 1$ then SSE $= 0$, indicating a perfect fit. If $R^2 = 0$ then SSE $=$ SST, indicating the model does no better than the sample mean.
- R^2 is generally reported when running a regression by almost any statistical software.

Evaluating our Model: R^2 and Goodness of Fit

Let's see how this looks with an example from two different datasets



- In which dataset does the regression line look closer to the data?
- Does homoskedasticity look to be violated here?

Table of Contents

Advanced Inference Topics

Evaluating our Model

Modeling Choices

Modeling Choices: Transforming our Data

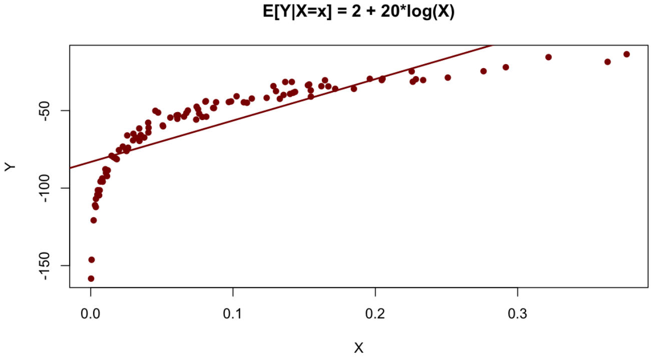
Suppose we fit our linear model and find a low R^2. There are two main reasons this could be happening.

1. Knowing X simply does not give us much information about Y
   - For example, suppose we were trying to predict log wages, Y, using a persons favorite color, X.
2. The true relationship between X and Y is non-linear and we are trying to fit a linear model.

The first problem we can't do much to address, other than trying to collect more right hand side variables. The second problem however we can try and address by transforming our data.

Modeling Choices: Transforming our Data

Let's see an example of this. Suppose we collect our data and it looks like below



The relationship between X and Y is clearly non-linear, but we are trying to fit a line through it. This hurts our model preformance.

Modeling Choices: Transforming our Data

Other reasons that we may think that the relationship between Y and X is non-linear

- Y is bounded and X has a large support.
  - For example, suppose $Y \in \{0, 1\}$ denotes treatment uptake and X denotes income.
  - A linear model of the form $\hat{\beta}_0 + \hat{\beta}_1 \cdot X$ would give $\hat{Y} > 1$ for a sufficiently large value of X.
- We believe that the derivative of Y with respect to X is not constant.
  - Suppose Y is amount spent on groceries and X is income.
  - We expect that as income rises people may start shopping at Whole Foods or spending more on nicer ingredients.
  - But this will probably level off for high levels of income. A linear model imposes that $\frac{d}{dx}\hat{Y}(x) = \frac{d}{dx}(\hat{\beta}_0 + \hat{\beta}_1 x) = \hat{\beta}_1$ for all x.

Modeling Choices: Transforing our Data

So, given that we believe the relationship between Y and X to be nonlinear, what can we do about this?

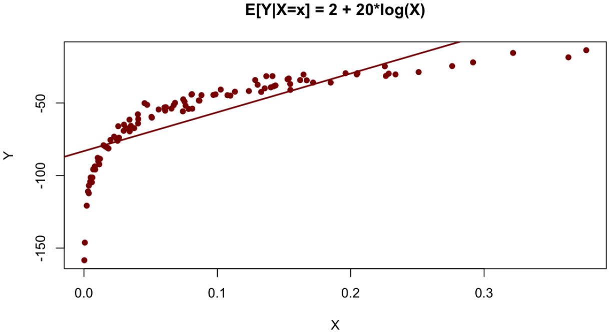
Before we move on to something fancier, let's try just transforming our data and running a linear regression:

$$f(Y) = \beta_0 + \beta_1 g(X) + \epsilon.$$

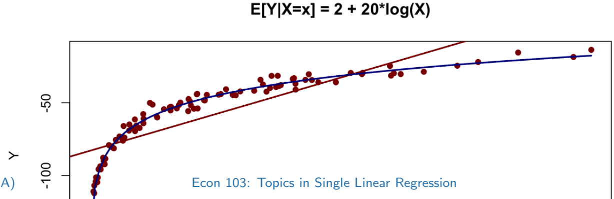
Common functions to be used here are $\ln(z)$, $\sqrt{z}$, and various polynomials; z^2, z^3, etc.

Modeling Choices: Transforming our Data

Let's return to our data from before and see how this would work.



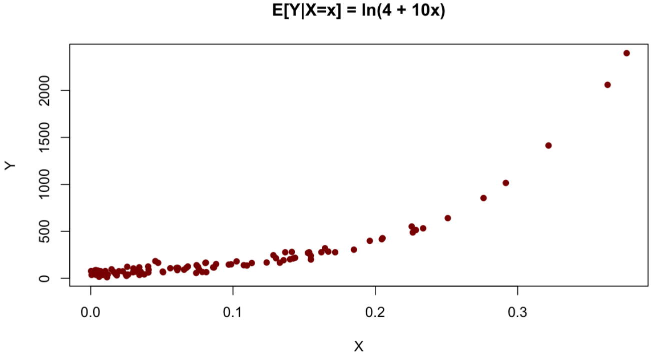
E[Y|X=x] = 2 + 20*log(X)



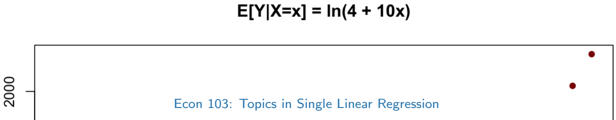
E[Y|X=x] = 2 + 20*log(X)

Modeling Choices: Transforming our Data

Let's see another example. Suppose our data looks like the below.



First, let's try fitting a non-transformed linear regression: $Y = \beta_0 + \beta_1 X + \epsilon$



Modeling Choices: Transforming our Data

In all of the above, notice that from an estimation and inference perspective, nothing much has changed. We can estimate our parameters and conduct inference just as before, but while treating our data as $\{f(Y_i), g(X_i)\}$ as opposed to $\{Y_i, X_i\}$.

Modeling Choices: Transforming our Data

Let's see an example of this. Suppose that we want to estimate the following model:

$$\ln(Y) = \beta_0 + \beta_1 \cdot \ln(X) + \epsilon.$$

Using our data $\{Y_i, X_i\}$ we estimate $\hat{\beta}_0, \hat{\beta}_1$ via

$$\hat{\beta}_0, \hat{\beta}_1 = \arg\min_{b_0, b_1} \frac{1}{n} \sum_{i=1}^{n} (\ln(Y_i) - b_0 - b_1 X_i)^2.$$

we find that, with $n = 49$, $\hat{\beta}_1 = 0.5$, $\sigma^2_{\ln(X)} = \frac{1}{n}\sum_{i=1}^{n}(\ln(X_i) - \overline{\ln(X)})^2 = 4$ and $\hat{\sigma}^2_\epsilon = \frac{1}{n}\sum_{i=1}^{n}(\ln(Y_i) - \hat{\beta}_0 - \hat{\beta}_1 \ln(X_i))^2 = 4$.

Modeling Choices: Transforming our Data

Assuming homoskedasticity, let's use this information to construct a 95% confidence interval for β_1.

We use the same formula from before to calculate $\hat{\sigma}^2_{\beta_1} = \frac{\hat{\sigma}^2_{\epsilon}}{\hat{\sigma}^2_{\ln(X)}} = \frac{4}{4} = 1$.

Then, using $z_{0.975} = 1.96$, a 95% confidence interval is contructed

$$\hat{\beta}_1 \pm 1.96 \frac{\hat{\sigma}_{\beta_1}}{\sqrt{n}} = 0.5 \pm 1.96 \frac{1}{7} = [0.22, 0.78].$$

- Notice that everything is the same as before
- Using this, would we reject $H_0 : \beta_1 \leq 0$ against $H_1 : \beta_1 > 0$ at level $\alpha = 0.5$?
  - Recall that we would reject $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ at level $\alpha = 0.5$.

Transforming our Data: Questions

Questions?

Modeling Choices: Transforming our Data

What changes, however, is our interpretation of our parameters. Before we interpreted:

- β_0: Expected value of Y when $X = 0$.
- β_1: Expected change in Y when X increases by one unit

In the model $f(Y) = \beta_0 + \beta_1 g(X) + \epsilon$, we have the following interpretations of β_0 and β_1.

- β_0: Expected value of $f(Y)$ when $g(X) = 0$
- $\beta_1 g'(x)$: Expected change in $f(Y)$ at $X = x$ when X increases by one unit.

Modeling Choices: Transforming our Data

Example: Suppose X represents the square footage of a house and Y represents is final sales price (in tens of thousands of dollars). We estimate the following model

$$Y = \beta_0 + \beta_1 X^2 + \epsilon$$

and find that $\hat{\beta}_0 = 50$ and $\hat{\beta}_1 = 2$.

How do we interpret these parameter estimates?

- $\hat{\beta}_0 = 50$: The expected sales price of an empty lot $(X^2 = 0 \iff X = 0)$ is \$50,000.
- $\hat{\beta}_1 = 2$: Taking derivatives gives that $\frac{d}{dx}\hat{Y} = 2\hat{\beta}_1 X$. We expect that a one square foot increase in home size at square footage x will be associated with a \$40,000·x increase in sales price.

Modeling Choices: Transforming our Data

A useful approximation that we use here is that a one unit increase in $\ln(Z)$ is about a 100% increase in Z and vice versa, a 1% increase in Z is associated with a $1/100$ unit increase in $\ln(Z)$.

This is useful for interpreting the parameters of various models.

- Suppose we model $Y = \beta_0 + \beta_1 \ln(X) + \epsilon \implies$ a 1% increase in X is associated with a $\beta_1/100$ unit change in Y.
- Suppose we model $\ln(Y) = \beta_0 + \beta_1 X + \epsilon \implies$ a 1 unit increase in X is associated with $100 \cdot \beta_1\%$ increase in Y
- Suppose we model $\ln(Y) = \beta_0 + \beta_1 \ln(X) + \epsilon \implies$ a 1% increase is associated with a $\beta_1\%$ increase in Y.

Transforming our Data: Questions

Questions?

Modeling Choices: Linear Transformations

We may be tempted to try and improve the fit of our model by scaling Y or X up and down. In the below let's see what happens if we try and do so.

First, let's see what happens if we replace Y with $\tilde{Y} = cY$ for some $c \neq 0$. Let's consider esimating the model $\tilde{Y} = \beta_0^\circ + \beta_1^\circ X + \epsilon$

$$\begin{aligned} \hat{\beta}_1^\circ &= \frac{\frac{1}{n}\sum_{i=1}^n (\tilde{Y} - \bar{\tilde{Y}})(X_i - \bar{X})}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})} \\ &= c \underbrace{\frac{\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}}_{\text{estimated slope parameter from } Y \text{ on } X} \end{aligned}$$

So we see that the slope parameter simply get's scaled up or down by c: $\hat{\beta}_1^\circ = c\hat{\beta}_1$.

The intercept parameter also gets scaled by c:

$$\hat{\beta}_0^\circ = \bar{\tilde{Y}} - \hat{\beta}_1^\circ \bar{X} = c\bar{Y} - c\hat{\beta}_1 \bar{X} = c(\bar{Y} - \hat{\beta}_1 \bar{X}) = c\hat{\beta}_0.$$

Modeling Choices: Linear Transformations

Now let's see what happens if we replace X with $\tilde{X} = cX$ for some $c \neq 0$. Let's consider estimating the model $Y = \beta_0^\circ + \beta_1^\circ \tilde{X} + \epsilon$.

$$\begin{aligned}
\hat{\beta}_1^\circ &= \frac{\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})(\tilde{X}_i - \bar{\tilde{X}})}{\frac{1}{n}\sum_{i=1}^n (\tilde{X}_i - \bar{\tilde{X}})^2} \\
&= \frac{c\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{c^2\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \frac{1}{c}\hat{\beta}_1
\end{aligned}$$

Here the slope parameter gets scaled by $\frac{1}{c}$, $\hat{\beta}_1^\circ = \frac{1}{c}\hat{\beta}_1$.

The intercept parameter in this case doesn't get scaled at all:

$$\hat{\beta}_0^\circ = \bar{Y} - \hat{\beta}_1^\circ \bar{\tilde{X}} = \bar{Y} - \frac{1}{c}\hat{\beta}_1 (c\bar{X}) = \beta_0.$$

- Average value of Y when $X = 0$ is the same as the average value of Y when $cX = 0$, they are the same event.

Linear Transformations: Questions

Questions?