Econ 103: Multiple Linear Regression I

Manu Navjeevan

UCLA

September 1, 2021

Content Outline

The Model:

- Adding more covariates
- Assumptions needed for inference

The Estimator:

- Relation to Single Linear Regression Estimator
- Asymptotic Dsitribution

Inference:

- Hypothesis Tests and Linear Combinations
- Confidence Inervals

Modeling Choices:

- Polynomial Equations, transformations, and interactions
- R^2 and goodness of fit

Table of Contents

The Model: Introduction

So far we have used the model $Y = \beta_0 + \beta_1 X + \epsilon$ defined by the line of best fit parameters

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

to learn about the relationship between a single random variable X and Y and to use X to predict Y.

Examples:

- Using education to predict income or interpreting the coeffecient $\hat{\beta}_1$ to learn about the relationship between the two.
- Learning about the relationship between smoking and heart disease.

The Model: Introduction

However, what happens if we have access to multiple explanatory variables $X_1, \ldots, X_p$?

Examples:

- Suppose we wanted to impact the joint effect of education and experience on income?
- Learn about the relationship between smoking, genetic risk, and heart disease

The Model: Introduction

As before, we may be interested in the parameters of a "line of best fit" between Y and our explantory variables $X_1, \ldots, X_p$:

$$\beta_0, \beta_1, \ldots, \beta_p = \arg \min_{b_0, \ldots, b_p} \mathbb{E}\left[(Y - b_0 - b_1 X_1 - b_2 X_2 - \cdots - b_p X_p)^2\right].$$

Again defining $\epsilon = Y - \beta_0 - \beta_1 X_1 - \cdots - \beta_p X_p$ these parameters generate the linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

where, by the first order conditions for β, $\mathbb{E}[\epsilon] = \mathbb{E}[\epsilon X_j] = 0$ for all $j = 0, 1, \ldots, p$.

The Model: Introduction

Example 1: Let Y be log wages, EDU be years of college education, and EXP be years of experience. Prior to this we have estimated the equation

$$Y = \beta_0 + \beta_1 EDU + \epsilon. \tag{1}$$

Now, we will consider estimation and inference on the model

$$Y = \beta_0 + \beta_1 EDU + \beta_2 EXP + \epsilon. \tag{2}$$

Note that β_0, β_1 in model (1) will differ from β_0, β_1 in model (2).

- In (1) β_0 corresponds to the average log wage for someone with no college education
- In (2) β_0 will correspond to the average log wage for someone with no college education and no experience
- In (1) β_1 corresponds to the expected change in log wage for an additional year of college education
- In (2) β_1 corresponds to the expected change in log wage for an additional year of college education after controlling for years of experience

The Model: Introduction

Example 2: Let Y be the (log) final sales price of a home, $SQFT$ be the square footage of the house, and $DAYS$ be the number of days the house has been on the market. Before we estimated and interpreted the linear model:

$$Y = \beta_0 + \beta_1 SQFT + \epsilon. \tag{3}$$

Now, we will consider estimation and inference on the model

$$Y = \beta_0 + \beta_1 SQFT + \beta_2 DAYS + \epsilon \tag{4}$$

- In (3) β_0 is interpreted as the average log sales price for a home with zero square feet (an empty lot) regardless of how long it's been on the market.
- In (4) β_0 is interpreted as the average log sales price for a home with zero square feet that has just entered the market
- In (3) β_1 is interpeted as the average change in log sales price for a one unit increase in square footage
- In (4) β_1 is interpreted as the average change in log sales price for a one unit increase in square footage, holding the number of days on the market constant

The Model: Introduction

Example 3: Finally, let's return to an example from Week 1. Let Y be a measure of anxiety levels, ENG be the number of energy drinks consumed per day, and CLS be the number of courses being taken. Before we may have estimated the model:

$$Y = \beta_0 + \beta_1 ENG + \epsilon \tag{5}$$

Now, we may consider the model

$$Y = \beta_0 + \beta_1 ENG + \beta_2 CLS + \epsilon \tag{6}$$

- In (5) we can interpret β_0 as the average anxiety level for someone who drinks no energy drinks
- In (6) we can interpret β_0 as the average anxiety level for someone who drinks no energy drinks and takes no classes
- In (5) we can interpret β_1 as the expected change in anxiety levels for someone who drinks one more energy drink per day
- In (6) we can interpret β_1 as the expected change in anxiety levels for an additional energy drink holding the number of courses being taken constant.

Question: How may we expect the signs/magnitutes of the parameters to change when going from model (5) to model (6)?

The Model: Questions

Questions?

Table of Contents

Estimation: Introduction

Before, in single linear regression when we were interested in the population line of best fit parameters

$$\beta_0, \beta_1 = \arg\min_{b_0, b_1} \mathbb{E}\left[(Y - b_0 - b_1 X)^2\right],$$

we estimated them by finding the line of best fit through our sample $\{Y_i, X_i\}_{i=1}^n$:

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

$\rightarrow$ Have to estimate these parameters using the sample because we don't know the population distribution of (Y, X)

Estimation: Introduction

Now, we are interested in the population line of best fit parameters:

$$\beta_0, \beta_1, \ldots, \beta_p = \arg \min_{b_0, b_1, \ldots, b_p} \mathbb{E}\left[(Y - b_0 - b_1 X_1 - \cdots - b_p X_p)^2\right].$$

Question: How should we estimate these using our sample $\{Y_i, X_{1,i}, \ldots, X_{p,i}\}_{i=1}^n$?

Estimate $\beta_0, \ldots, \beta_p$ by finding the line of best fit through our sample:

$$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p = \arg \min_{b_0, b_1, \ldots, b_p} \frac{1}{n} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1,i} - \cdots - b_p X_{p,i})^2.$$

Estimation: Introduction

Taking first order conditions for $\hat{\beta}_0, \ldots, \hat{\beta}_p$ above gives us

$$\frac{\partial}{\partial b_0} : \frac{1}{n} \sum_{i=1}^{n} \overbrace{(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1,i} - \cdots - \hat{\beta}_p X_{p,i})}^{\hat{\epsilon}_i} = 0$$

$$\frac{\partial}{\partial b_1} : \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1,i} - \cdots - \hat{\beta}_p X_{p,i}) X_{1,i} = 0$$

$$\vdots$$

$$\frac{\partial}{\partial b_p} : \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1,i} - \cdots - \hat{\beta}_p X_{p,i}) X_{p,i} = 0$$

This gives us $p + 1$ linear equations to solve for our $p + 1$ parameters. Computers can solve these very quickly, but the explicit formulas for $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ become very cumbersome if we don't use linear algebra notation.

Estimation: Aside

Quickly, it is useful to note the following implication from the first order conditions for $\hat{\beta}_0, \ldots, \hat{\beta}_p$. Define $\hat{\epsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1,i} - \cdots - \hat{\beta}_p X_{p,i}$. Then

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i = 0$$

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i X_{1,i} = 0$$

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i X_{2,i} = 0$$

$$\vdots$$

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i X_{p,i} = 0$$

Estimation: Asymptotic Distribution

Just as in single linear regression, however, the solutions for $\hat{\beta}_0, \ldots, \hat{\beta}_p$ depend on the data. That is $\hat{\beta}_0, \ldots, \hat{\beta}_p$ are functions of our sample $\{Y_i, X_{1,i}, \ldots, X_{p,i}\}_{i=1}^n$.

- If we collected a different sample $\{Y_i, X_{1,i}, \ldots, X_{p,i}\}_{i=1}^n$, we would get different values for our estimators $\hat{\beta}_0, \ldots, \hat{\beta}_p$.

For hypothesis testing we would still like to know the (approximate) distribution of our estimates $\hat{\beta}_0, \ldots, \hat{\beta}_p$. This will be useful later on as we'd like to calculate objects such as

$$\Pr(|\hat{\beta}_1| > 5 | \beta_1 = -2).$$

Estimation: Asymptotic Distribution

In order for the estimates $\hat{\beta}_0, \ldots, \hat{\beta}_p$ to have a stable asymptotic distribution and to converge to the true parameters $\beta_0, \ldots, \beta_p$, we need to make some (light) assumptions about the underlying distribution of $(Y, X_1, \ldots, X_p)$ from which our sample is drawn.

Estimation: Asymptotic Distribution

Assumptions needed for valid inference:

- Random Sampling: The data $\{Y_i, X_{1,i}, \ldots, X_{p,i}\}$ is independently and identically sampled from the population distribution $(Y, X_1, \ldots, X_p)$
  - Needed to make sure that we are making inferences on the correct population
  - Question: When would this be violated?
- Rank Condition: The right hand side variables $X_1, \ldots, X_p$ are not linearly dependent, i.e we cannot write

  $$a_1 X_1 + a_2 X_2 + \cdots + a_p X_p = 0$$

  for some constants $a_1, \ldots, a_p$ with at least one $a_k \neq 0$.
  - If this is violated then we can write one random variable as a linear combination of the other ones.
  - To see why this is problematic, suppose that we could write $X_1 = 2X_2$. Then these two linear models are equivalent

    $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \iff Y = \beta_0 + (2\beta_1 + \beta_2) X_2 + \epsilon.$$

    The "line of best fit" solution is then not unique. We can achieve the same fit by setting the coeffecient on X_1 to be β_1 and the coeffecient on X_2 to be β_2 or by setting the coeffecient on X_1 to be zero and the coeffecient on X_2 to be $2\beta_1 + \beta_2$.

- Homoskedasticity: $\mathrm{Var}(\epsilon | X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p) = \sigma^2_\epsilon$ for all

Estimation: Asymptotic Distribution

Under the assumptions Random Sampling and Rank Condition we get the following result for any $\hat{\beta}_k$, $k = 0, 1, \ldots, p$.

Approximately, for large n:

$$\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}_{\beta_k}/\sqrt{n}} \sim N(0,1) \iff \hat{\beta}_k \sim N\Big(\beta_k, \underbrace{\sigma^2_{\beta_k}/n}_{=\text{Var}(\hat{\beta}_k)}\Big).$$

- The assumption Homoskedasticity simply changes the form of $\sigma^2_{\beta_k}$ and thus it's estimator $\hat{\sigma}^2_{\beta_k}$.
- Unlike in single linear regression we will not go over a general form for $\hat{\sigma}_{\beta_k}$
  - Typically, all that you need to know is that $\hat{\sigma}_{\beta_k}$ (or the standard error, $\hat{\sigma}_{\beta_k}/\sqrt{n}$ or the variance $\hat{\sigma}^2_{\beta_k}/n$) will either be given to us directly or found in R output.
- In addition, we will be able to estimate the asymptotic covariance between any two estimates $\hat{\beta}_j, \hat{\beta}_k$ for $j, k = 0, 1, \ldots, p$.

Estimation: Asymptotic Variance

To consolidate notation, the variances and covariances are often presented as a Variance-Covariance matrix. For example, when $p = 2$ the variance covariance matrix looks like

$$\text{Cov}(\hat{\beta}) = \begin{pmatrix} \text{Var}(\hat{\beta}_0) & \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \text{Cov}(\hat{\beta}_0, \hat{\beta}_2) \\ \text{Cov}(\hat{\beta}_1, \hat{\beta}_0) & \text{Var}(\hat{\beta}_1) & \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) \\ \text{Cov}(\hat{\beta}_2, \hat{\beta}_0) & \text{Cov}(\hat{\beta}_2, \hat{\beta}_1) & \text{Var}(\hat{\beta}_2) \end{pmatrix}.$$

- Note that $\text{Cov}(X, Y) = \text{Cov}(Y, X)$ so this is a symmetric matrix
- Also note that $\text{Var}(X) = \text{Cov}(X, X)$ which is why we sometimes just call this the Covariance matrix.
- In general the Variance-Covariance matrix will be a $(p + 1) \times (p + 1)$ matrix (one dimension for each of the slope coefficients and the intercept).

Estimation: Asymptotic Variance

Question: What influences the asymptotic variance?

In order to get some intuition for this, we will go over a particular example when $p = 2$. That is when we want to estimate the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$

In this case, we will be able to get some simple closed form expressions for $\sigma^2_{\beta_1}$ and $\sigma^2_{\beta_2}$.

- Will provide some insight into what drives the asymptotic variance

Estimation: Asymptotic Variance

Before doing so, let's review the correlation coeffecient. Recall that for two random variables X_1 and X_2 the correlation coeffecient ρ_{12} is defined

$$\rho_{12} = \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{Var}(X_1)}\sqrt{\text{Var}(X_2)}}.$$

The correlation coeffecient is a measure of the linear dependence between X_1 and X_2

- If $\rho_{12} = 1$ then X_1 and X_2 are perfectly linearly dependent, that is $X_1 = cX_2$ for some constant $c \neq 0$
- If $\rho_{12} = 0$ then X_1 and X_2 are have no linear dependence, that is $\text{Cov}(X_1, X_2) = 0$.

Estimation: Asymptotic Variance

With this in mind, the asymptotic variance (under homoskedasticity) $\hat{\sigma}^2_{\beta_1}$ for β_1 in the linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon,$$

is given

$$\sigma^2_{\beta_1} = \frac{\sigma^2_\epsilon}{(1-\rho^2_{12})\sigma^2_{X_1}} \iff \sqrt{n}(\hat{\beta}_1 - \beta_1) \sim N(0, \sigma^2_{\beta_1}),$$

where ρ_{12} is the correlation coeffecient between X_1 and X_2.

Notice:

- As before $\text{Var}(\hat{\beta}_1) = \sigma^2_{\beta_1}/n$ is decreasing with n, $\hat{\beta}_1 \to \beta_1$ as $n \to \infty$.
- As before $\sigma^2_{\beta_1}$ is increase with σ^2_ϵ and decreasing with $\sigma^2_{X_1}$
  - σ^2_ϵ : If points are closer to the line it is easier to make out the line
  - $\sigma^2_{X_1}$: If points are more spread out, it is easier to make out the line
- However, now we see that the variance $\sigma^2_{\beta_1}$ is increasing also as $\rho_{12} \uparrow 1$.
  - Intuition: If X_1 and X_2 are highly correlated, it is difficult to parse out the relationship of X_1 on Y holding X_2 constant.

Estimation: Asymptotic Variance

To estimate $\sigma^2_{\beta_1}$ we can estimate each of it's components.

$$\hat{\sigma}^2_{\beta_1} = \frac{\hat{\sigma}^2_\epsilon}{(1-\hat{\rho}^2_{12})\hat{\sigma}^2_{X_1}}.$$

- For $\hat{\sigma}^2_\epsilon$ generate the estimated residuals $\hat{\epsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1,i} - \hat{\beta}_2 X_{2,i}$ and calculate the sample variance of the estimated residuals:

$$\hat{\sigma}^2_\epsilon = \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}^2_i.$$

  - Recall that by the first order conditions for $\hat{\beta}_0$, $\frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i = 0$ so that $\bar{\hat{\epsilon}} = 0$

- For $\hat{\sigma}^2_{X_1}$ calculate the sample variance of X_1

$$\hat{\sigma}^2_{X_1} = \frac{1}{n}\sum_{i=1}^n (X_{1,i} - \bar{X}_1)^2.$$

Estimation: Asymptotic Variance

To estimate $\sigma^2_{\beta_1}$ we can estimate each of it's components.

$$\hat{\sigma}^2_{\beta_1} = \frac{\hat{\sigma}^2_\epsilon}{(1-\hat{\rho}^2_{12})\hat{\sigma}^2_{X_1}}.$$

- To estimate $\hat{\rho}^2_{12}$ recall that

$$\rho_{12} = \frac{\text{Cov}(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}} \implies \hat{\rho}_{12} = \frac{\widehat{\text{Cov}}(X_1, X_2)}{\hat{\sigma}_{X_1}\hat{\sigma}_{X_2}}.$$

  - We have already covered how to estiamte $\hat{\sigma}^2_{X_1}$. Estimating $\hat{\sigma}^2_{X_2}$ follows the same formula

$$\hat{\sigma}^2_{X_2} = \frac{1}{n}\sum_{i=1}^{n}(X_{2,i} - \bar{X}_2)^2.$$

    Then, take square roots $\hat{\sigma}_{X_1} = \sqrt{\hat{\sigma}^2_{X_1}}$ and $\hat{\sigma}_{X_2} = \sqrt{\hat{\sigma}^2_{X_2}}$.

  - To estimate the covariance note $\text{Cov}(X_1, X_2) = \mathbb{E}[(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})]$ so

$$\widehat{\text{Cov}}(X_1, X_2) = \frac{1}{n}\sum_{i=1}^{n}(X_{1,i} - \bar{X}_1)(X_{2,i} - \bar{X}_2).$$

Estimation: Asymptotic Variance

Let's see an example of this. Suppose we are interested in the joint effect of smoking heavily and drinking heavily on liver failure.

That is let $Y \in \{0, 1\}$ denote liver failure, $X_1 \in \{0, 1\}$ denote being a heavy smoker, and $X_2 \in \{0, 1\}$ denote being a heavy drinker and suppose we want to estimate the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$

Estimation: Asymptotic Variance

After collecting a sample of size $n = 64$ we estimate $\hat{\sigma}^2_\epsilon = 0.25$, $\hat{\sigma}^2_{X_1} = 0.1$, and $\hat{\rho}_{12} = 0.5$, where

$$\hat{\rho}_{12} = \frac{\widehat{\text{Cov}}(X_1, X_2)}{\hat{\sigma}_{X_1}\hat{\sigma}_{X_2}}.$$

Question: What is the standard error of $\hat{\beta}_1$?

Answer: Recall that the standard error is given $\hat{\sigma}_{\beta_1}/\sqrt{n}$. Using the above we get that

$$\hat{\sigma}^2_{\beta_1} = \frac{\hat{\sigma}^2_\epsilon}{(1-\hat{\rho}^2_{12})\hat{\sigma}^2_{X_1}} = \frac{0.25}{(1-0.25)0.1} = \frac{10}{3}.$$

The standard error is then $\hat{\sigma}_{\beta_1}/\sqrt{n} = \sqrt{10/3}/\sqrt{64} \approx 0.228$

Estimation: Asymptotic Variance

Now suppose that after collecting a sample of size $n = 100$ we estimate $\hat{\sigma}^2_\epsilon = 0.25$, $\hat{\sigma}^2_{X_1} = 0.1$. This time however, we estimate $\hat{\rho}_{12} = 0.75$, where

$$\hat{\rho}_{12} = \frac{\widehat{\text{Cov}}(X_1, X_2)}{\hat{\sigma}_{X_1}\hat{\sigma}_{X_2}}.$$

Question: In this case, what is the standard error of $\hat{\beta}_1$?

Answer: Using the formula above

$$\hat{\sigma}^2_{\beta_1} = \frac{\hat{\sigma}^2_\epsilon}{(1-\hat{\rho}^2_{12})\hat{\sigma}^2_{X_1}} = \frac{0.25}{(1-0.5625)0.1} = \approx 5.714.$$

The standard error is then $\hat{\sigma}_{\beta_1}/\sqrt{n} \approx \sqrt{5.714}/\sqrt{100} = 0.239$.

Notice that the standard error is larger now than it was when $n = 64$, despite the fact that our sample size has grown by about 50%!

Table of Contents

Inference: Introduction

Testing single hypothesis about the coeffecients of our regression or linear combinations of coeffecients follows the same procedure.

If we recall, this procedure consists of constructing a test statistic of the form

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

and then either computing a p-value or comparing the test statistic directly to a quantile of the standard normal distribution.

Inference: Single Coefficient Testing

Let's first see how this looks like with a hypothesis test for single coeffecient. Returning to the example from before let $Y \in \{0, 1\}$ be an indicator the existence of liver disease, $X_1 \in \{0, 1\}$ indicate whether or not someone is a heavy smoker, and $X_2 \in \{0, 1\}$ indicate whether someone is a heavy drinker.

We want to know if being a heavy smoker is a significant predictor of having liver disease after controlling for whether someone is a heavy drinker. To do so we estimate the following model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$

We want to test the following hypothesis at level $\alpha = 0.05$:

$$H_0 : \beta_1 \leq 0 \quad \text{vs.} \quad H_1 : \beta_1 > 0.$$

- Recall that this essentially amounts to computing the probability that we would observe our estimated value of $\hat{\beta}_1$ (or something even further from the null/more positive) if the true value $\beta_1 = 0$.
- Use the result that approximately for large n

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\beta_1}/\sqrt{n}} \sim N(0, 1).$$

Inference: Single Coefficient Testing

Let $Z \sim N(0,1)$. After collecting a sample of size $n = 100$ we estimate that $\hat{\beta}_1 = 0.03$ and that $\hat{\sigma}^2_{\beta_1} = 4$. Given this we want to compute the p-value:

$$\begin{aligned}
\Pr(\hat{\beta}_1 > 0.03 | \beta_1 = 0) &= \Pr(\hat{\beta}_1 - \beta_1 > 0.03 - \overbrace{\beta_1}^{=0}) \\
&= \Pr\left(\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\beta_1}/\sqrt{n}} > \underbrace{\frac{0.03 - 0}{\sqrt{4}/\sqrt{100}}}_{=t^*} \right) \\
&= \Pr(Z > 0.2) \text{ or, equivalently } \Pr(Z > t^*)
\end{aligned}$$

and reject the null hypothesis if this probability is less than $\alpha = 0.05$.

Computing this probability using R we find that $p = \Pr(Z > 0.2) \approx 0.42$, since this is larger than $\alpha = 0.05$ we fail to reject the null hypothesis.

Inference: Single Coefficient Testing

As in single linear regression, this is equivalent to checking whether the test statistic t^* is larger than $z_{1-\alpha}$ where $z_{1-\alpha}$ is such that $\Pr(Z \leq z_{1-\alpha}) = 1 - \alpha$ so that $\Pr(Z > z_{1-\alpha}) = \alpha$.

Understanding Check: Why is checking if $t^* > z_{1-\alpha}$ equivalent to checking if $p = \Pr(Z > t^*) < \alpha$? Since $\alpha = 0.05$ using the command **qnorm(0.95)** in R we find that $z_{0.95} = 1.64$. Computing the test statistic as before we find that

$$t^* = \frac{\hat{\beta}_1 - \text{null hypothesis}}{\hat{\sigma}_{\beta_1}/\sqrt{n}} = \frac{\hat{\beta}_1 - 0}{\sqrt{4}/\sqrt{100}} = 0.2.$$

Since $t^* = 0.2 \leq 1.64 = z_{0.95}$ we fail to reject the null hypothesis that $\beta_1 \leq 0$. We cannot reject the claim that heavy smoking has no association with having liver disease after controlling for heavy drinking.

Inference: Single Hypothesis Testing

We can see that this procedure is exactly the one we considered when we were doing hypothesis tests in single linear regression. In fact we will now go over a general framework for testing a single null hypothesis of a parameter θ whenever we have a result of the form (approximately for large n):

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

where $\hat{\theta}$ is an estimator of θ and $\hat{\sigma}_\theta / \sqrt{n}$ is the standard error of $\hat{\theta}$.

Examples:

- By the central limit theorem

$$\frac{\bar{X} - \mu_X}{\hat{\sigma}_X^2 / \sqrt{n}} \sim N(0, 1)$$

- For a specific parameter β_k in our multiple linear regression we have that

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

- For a linear combination of regression coefficients $\lambda = a\beta_j + b\beta_k$ and $\hat{\lambda} = a\hat{\beta}_j + b\hat{\beta}_k$ we have that

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

Inference: Single Hypothesis Testing

As before we have two standard procedure for testing the null hypotheses: $H_0 : \theta \leq b$, $H_0 : \theta \geq b$, or $H_0 : \theta = b$. The first one involves computing p-values:

1. Compute the test statistic
$$t^* = \frac{\hat{\theta} - b}{\hat{\sigma}_\theta / \sqrt{n}}.$$

2. Compute the p-value, the probability that we would obtain our observed value of $\hat{\theta}$, or something even further from the null hypothesis, if the null hypothesis was correct
   - If $H_0 : \theta = b$ and $H_1 : \theta \neq b$ compute
   $$p = \Pr(|Z| > |t^*|) = 2\Pr(Z > |t^*|).$$
   - If $H_0 : \theta \leq b$ and $H_1 : \theta > b$ compute
   $$p = \Pr(Z > t^*).$$
   - If $H_0 : \theta \geq b$ and $H_1 : \theta < b$ compute
   $$p = \Pr(Z < t^*).$$

3. Reject the null hypothesis in favor of the alternative hypothesis if $p < \alpha$. Otherwise fail to reject the null hypothesis.

Inference: Single Coeffecient Testing

As before this gives us two standard procedure for testing the null hypotheses: $H_0 : \theta \leq b$, $H_0 : \theta \geq b$, or $H_0 : \theta = b$. The second one involves comparing the test statistic to quantiles of the standard normal distribution:

1. Compute the test statistic or "t-statistic"

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

2. For a given level α compute $z_{1-\alpha}$ for a one sided alternative or $z_{1-\alpha/2}$ for a 2 sided alternative, where $z_{1-\alpha}$ and $z_{1-\alpha/2}$ are such that

$$\Pr(Z > z_{1-\alpha}) = \alpha \quad \text{and} \quad \Pr(Z > z_{1-\alpha/2}) = \frac{\alpha}{2}.$$

   These are called the $1 - \alpha$ and $1 - \alpha/2$ quantiles of the standard normal distribution, respectively.

   - $z_{0.9} \approx 1.28$
   - $z_{0.95} \approx 1.64$
   - $z_{0.975} \approx 1.96$
   - $z_{0.99} \approx 2.32$
   - $z_{0.995} \approx 2.57$

3. Compare the test statistic t^* to the quantile $z_{1-\alpha}$ or $z_{1-\alpha/2}$.

Single Hypothesis Testing: Questions

Questions?

Inference: Single Hypothesis Testing

Let's see how this works when testing a linear combination of parameters. Returning to our example let $Y \in \{0, 1\}$ denote having liver disease, $X_1 \in \{0, 1\}$ denote being a heavy smoker, and $X_2 \in \{0, 1\}$ denote being a heavy drinker. We estimate the following linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$

Inference: Single Hypothesis Testing

We want to test at level $\alpha = 0.1$ the null hypothesis that being a heavy smoker and a heavy drinker makes you at most 10% more likely to develop heart disease compared to someone who is neither a heavy smoker nor a heavy drinker. Using our linear model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ we approximate

$$\begin{aligned} &\mathbb{E}[Y|X_1 = 1, X_2 = 1] = \beta_0 + \beta_1 + \beta_2 \\ -\ &\mathbb{E}[Y|X_1 = 0, X_2 = 0] = \beta_0 \\ \hline &= \beta_1 + \beta_2 \end{aligned}$$

So, we can state our hypotheses as

$$H_0 : \lambda = \beta_1 + \beta_2 \leq 0.1 \quad \text{vs.} \quad H_1 : \lambda = \beta_1 + \beta_2 > 0.1.$$

We will aim to test this hypothesis by using the procedure outline above. We will construct $\hat{\lambda} = \hat{\beta}_1 + \hat{\beta}_2$. Because $\hat{\beta}_1$ and $\hat{\beta}_2$ are both (jointly) approximately normal for large n we get the following result that

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

where we can calculate $\hat{\sigma}_\lambda / \sqrt{n} = \sqrt{\text{Var}(\hat{\lambda})}$. We can then use this to do hypothesis testing following the steps above.

Aside: I know we have a lot of forms for the variance floating around. To reiterate, basically we use something like σ_θ^2 to say that

$$\sqrt{n}(\hat{\theta} - \theta) \sim N(0, \sigma_\theta^2)$$

then after rearranging this we get that $\text{Var}(\hat{\theta}) = \sigma_\theta^2 / n$ so that $\sigma_\theta / \sqrt{n} = \sqrt{\text{Var}(\hat{\theta})}$.

Inference: Single Hypothesis Testing

After collecting a sample of size $n = 100$ of $\{Y_i, X_{1,i}, X_{2,i}\}_{i=1}^{100}$ we find that

$$\hat{\beta}_0 = 0.05 \;\; \hat{\beta}_1 = 0.02 \;\; \hat{\beta}_2 = 0.15 \implies \hat{\lambda} = \hat{\beta}_1 + \hat{\beta}_2 = 0.17.$$

Question: How do we interpret $\hat{\lambda}$? We also estimate the following covariance matrix:

$$\text{Cov}(\hat{\beta}) = \begin{array}{c} \\ \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{array} \begin{array}{c} \begin{array}{ccc} \hat{\beta}_0 & \hat{\beta}_1 & \hat{\beta}_2 \end{array} \\ \begin{pmatrix} 0.05 & 0.25 & 0.16 \\ 0.25 & 0.08 & 0.1 \\ 0.16 & 0.1 & 0.36 \end{pmatrix} \end{array}$$

Recall that $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X, Y)$. Using this we calculate

$$\begin{aligned} \text{Var}(\hat{\lambda}) &= \text{Var}(\hat{\beta}_1) + \text{Var}(\hat{\beta}_2) + 2\,\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) \\ &= 0.08 + 0.36 + 2 \cdot 0.1 = \underline{0.64} \end{aligned}$$

The standard error is then $\hat{\sigma}_\lambda / \sqrt{n} = \sqrt{\text{Var}(\hat{\lambda})} = 0.8$.

Inference: Single Hypothesis Testing

So we have that $\hat{\lambda} = 0.17$ and $\hat{\sigma}_\lambda/\sqrt{n} = 0.8$. We would like to use this to test at level $\alpha = 0.1$ the hypotheses

$$H_0 : \lambda \leq 0.1 \quad \text{vs.} \quad H_1 : \lambda > 0.1.$$

We will do so two ways, first using the p-value and second by directly comparing to a critical value $z_{1-\alpha} = z_{0.9}$.

First, construct the test statistic

$$t^* = \frac{\hat{\lambda} - 0.1}{\hat{\sigma}_\lambda/\sqrt{n}} = \frac{0.17 - 0.1}{0.8} = 0.0875.$$

Now, compute the p-value. Since this is a one sided test with $H_1 : \lambda > 0.1$ the p-value is computed

$$\Pr(Z > t^*) = 1 - \Pr(Z \leq 0.0875) = \textbf{1- pnorm(0.0875)} = 0.465$$

Since this p-value is larger than $\alpha = 0.1$ we fail to reject the null hypothesis.

Inference: Single Hypothesis Testing

Let's try running this test directly by comparing the critical value t^* to $z_{1-\alpha}$ where $z_{1-\alpha}$ is such that

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

Because $H_1 : \lambda > 0.1$ and $\alpha = 0.1$ we reject $H_0 : \lambda \leq 0.1$ if $t^* > z_{0.9}$.

To compute $z_{1-\alpha}$ we run **qnorm(0.9)** which outputs 1.28. Since our test statistic $t^* = 0.0875$ is less than $z_{1-\alpha}$ we again fail to reject this null hypothesis.

Inference: Confidence Intervals

As before, we may also want to compute a $100(1-\alpha)\%$ confidence interval for θ, where again we are in the setting where (approximately, for large n:)

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

As before, we want to include potential values in this interval that we are "reasonably confident" that θ could be.

- That is values, b, for which we would not reject $H_0 : \theta = b$ against an alternative $H_1 : \theta \neq b$

Inference: Confidence Intervals

Recall from the discussion above that we fail to reject $H_0 : \theta = b$ in favor of the alternative hypothesis $H_1 : \theta \neq b$ if the absolute value of the test statistic is less than $z_{1-\alpha/2}$

$$\left| \frac{\hat{\theta} - b}{\hat{\sigma}_\theta / \sqrt{n}} \right| \leq z_{1-\alpha/2} \implies b \in \left[\hat{\theta} - z_{1-\alpha/2} \hat{\sigma}_\theta / \sqrt{n}, \hat{\theta} + z_{1-\alpha/2} \hat{\sigma}_\theta / \sqrt{n} \right].$$

Inference: Confidence Intervals

So, our $100(1-\alpha)\%$ confidence interval for θ is given

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

We interpret this as being $100(1-\alpha)\%$ confident that the true value of θ lies in the interval

$$\left[\hat{\theta} - z_{1-\alpha/2}\hat{\sigma}_\theta/\sqrt{n}, \hat{\theta} + z_{1-\alpha/2}\hat{\sigma}_\theta/\sqrt{n}\right].$$

Inference: Confidence Intervals

Let's see an example of this. Suppose we are trying to examine the relationship between log home sales price, square footage, and days on the market. To do so we estimate the following regression equation

$$SALES = \beta_0 + \beta_1 SQFT + \beta_2 DAYS + \epsilon.$$

Inference: Confidence Intervals

After collecting a sample of size $n = 100$ we obtain the following parameter estimates

$$\hat{\beta}_0 = 2, \quad \hat{\beta}_1 = 5, \quad \hat{\beta}_2 = 8.$$

and estimated covariance matrix

$$\text{Cov}(\hat{\beta}) = \begin{array}{c} \\ \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{array} \begin{array}{c} \begin{array}{ccc} \hat{\beta}_0 & \hat{\beta}_1 & \hat{\beta}_2 \end{array} \\ \begin{pmatrix} 2.5 & 1.6 & 0.5 \\ 1.6 & 1.96 & 0.1 \\ 0.5 & 0.1 & 1.9 \end{pmatrix} \end{array}$$

Question: How would we compute a 99% confidence interval for $\hat{\beta}_1$?

Inference: Confidence Intervals

To create a 99% confidence interval we first need to compute $z_{1-\alpha/2} = z_{0.995}$. To do so, we use $z_{0.995} =$ **qnorm(0.995)** $= 2.57$.

Next, we need to compute the standard error $\hat{\sigma}_{\beta_1}/\sqrt{n} = \sqrt{\hat{\sigma}^2_{\beta_1}/n} = \sqrt{\text{Var}(\hat{\beta}_1)}$.

From the covariance matrix we know that $\text{Var}(\hat{\beta}_1) = 1.96$ so that the standard error is given $\sqrt{1.96} = 1.4$.

Finally, we put these together to construct our confidence interval

$$\hat{\beta}_1 \pm z_{1-\alpha/2}\hat{\sigma}_{\beta_1}/\sqrt{n} = 5 \pm 2.57 \cdot 1.4 = [1.402, 8.598].$$

Question: Would we reject the null hypothesis $H_0 : \beta_1 = 0$ in favor of the alternative $H_1 : \beta_1 \neq 0$ at level $\alpha = 0.01$?

Confidence Intervals: Questions

Questions?

Table of Contents

Modeling Choices: Functional Forms

So far we have mainly covered regression specifications of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon.$$

As we covered briefly in single linear regression, this functional form is restrictive for a few reasons.

1. Marginal effect of any variable X_k on Y is modeled to be constant:

$$\frac{\partial}{\partial X_k} \widehat{Y} = \beta_k.$$

   - In reality, this assumption may be violated: marginal effect of education on income may be diminishing after the college.

2. Marginal effect of any variable X_k does not depend on any other variable X_j

$$\frac{\partial}{\partial X_k \partial X_j} = 0.$$

   - In reality, this assumption may also be violated: marginal effect of experience may depend on education levels

3. And in general, the relationship between Y and $(X_1, \ldots, X_k)$ may not be linear
   - If $Y \in \{0, 1\}$ and $\text{supp}(X_1, \ldots, X_k)$ is quite large, a linear model may predict $\widehat{Y} > 1$ for some values of X.

Modeling Choices: Functional Forms

In the single linear regression case we tried to account for these restrictions by taking nonlinear transformations of our variables, i.e allowing for specifications like

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

Question: Why limit ourselves to just including a single transformation?

Now that we know how to do estimation and inference on linear regression models with multiple right hand side variables, we will study two types of more advanced modeling techniques

- Polynomial modeling, where we include higher order polynomial terms: $X_1, X_1^2, X_1^3, \ldots$.

$$Y = \beta_0 + \beta_1 X_1 + + \beta_2 X_1^2 + \epsilon.$$

  - Useful for modeling diminishing/increasing marginal returns

- Interaction modeling, where we include terms like $X_1 \cdot X_2$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \cdot X_2 + \epsilon.$$

  - Useful for modeling non-zero cross derivatives

Of course, we can also always combine these two techniques!

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_1 \cdot X_2 + \epsilon.$$

Modeling Choices: Polynomial Modeling

Let's first look at an example where we might want to use polynomial terms. Suppose we want to estiamte the relationship between S, the square footage of a house, D, the number of days the house has been on the market, and its final log sales price, P.

We suspect that there are diminishing returns to square footage so we estimate the following model:

$$P = \beta_0 + \beta_1 D + \beta_2 S + \beta_3 S^2 + \epsilon.$$

Questions:

- How does this model allow for diminishing marginal returns to square footage?
- What values of β_3 would indicate diminishing marginal returns?
- How could we formally test for diminishing marginal returns?

Modeling Choices: Polynomial Modeling

Suppose after collecting a sample of size $n = 81$ and fitting this model we estimate $\hat{\beta}_3 = -0.1$ and the following covariance matrix:

$$\text{Cov}(\hat{\beta}) = \begin{array}{c} \\ \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{array} \begin{array}{c} \begin{array}{cccc} \hat{\beta}_0 & \hat{\beta}_1 & \hat{\beta}_2 & \hat{\beta}_3 \end{array} \\ \begin{pmatrix} 0.05 & 0.25 & 0.16 & 0.26 \\ 0.25 & 0.8 & 0.1 & 0.3 \\ 0.16 & 0.1 & 0.36 & 0.3 \\ 0.26 & 0.3 & 0.3 & 0.16 \end{pmatrix} \end{array}$$

Recall that the standard error is given $\hat{\sigma}_{\beta_1}/\sqrt{n} = \sqrt{\text{Var}(\hat{\beta}_k)}$! Let's use this information to test the null hypothesis $H_0 : \beta_3 \geq 0$ against the alternative hypothesis $H_1 : \beta_3 < 0$ at level $\alpha = 0.05$. Our test statistic is given

$$t^* = \frac{\hat{\beta}_3 - 0}{\sqrt{\text{Var}(\hat{\beta}_3)}} = \frac{-0.1}{\sqrt{0.16}} = -2.5.$$

Since $-2.5 < -z_{1-\alpha/2} = -z_{0.975} = -1.96$ we are further from the null than would be reasonably expected if $\hat{\beta}_3 \geq 0$, so we reject in favor of $H_1 : \beta_3 < 0$.

Modeling Choices: Interaction Modeling

Let's see an example of interaction modeling. Suppose we are interested in the effect of EDU, years of education, and EXP, years of experience, on INC, log income.

We suspect that the returns to experience differ based on people's level of education, so we include and interaction term and estimate the following model:

$$INC = \beta_0 + \beta_1 EXP + \beta_2 EDU + \beta_3 EDU \cdot EXP + \epsilon.$$

Questions:

- Given estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$, and $\hat{\beta}_3$, what is the estimated marginal return to experience for a given level of education?
- How does this model allow for returns to experience to differ based on one's education level?
- What does the sign of $\hat{\beta}_3$ tell us about the estimated relationship between returns to experience and level of education?

Modeling Choices: Interaction Modeling

Suppose after collecting a sample of size $n = 49$ and fitting the model we estimate $\hat{\beta}_3 = 0.07$ and the following covariance matrix.

$$\text{Cov}(\hat{\beta}) = \begin{array}{c} \\ \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{array} \begin{array}{c} \begin{array}{cccc} \hat{\beta}_0 & \hat{\beta}_1 & \hat{\beta}_2 & \hat{\beta}_3 \end{array} \\ \begin{pmatrix} 0.05 & 0.25 & 0.16 & 0.26 \\ 0.25 & 0.8 & 0.1 & 0.3 \\ 0.16 & 0.1 & 0.36 & 0.3 \\ 0.26 & 0.3 & 0.3 & 0.25 \end{pmatrix} \end{array}$$

We want to test whether returns to experience differs based on education level, that is test $H_0 : \beta_3 = 0$ against $H_1 : \beta_3 \neq 0$ at level $\alpha = 0.1$. To do so, let's construct our test statistic

$$t^* = \frac{\hat{\beta}_3 - 0}{\sqrt{\text{Var}(\hat{\beta}_3)}} = \frac{0.07}{0.5} = 9,14.$$

Modeling Choices: Interaction Modeling

Now, given $t^* = 0.14$ let's construct a p-value. For a two sided test, the p-value is given

$$\begin{aligned}\Pr(|Z| \geq |t^*|) = 2\Pr(Z \geq 0.14) &= 2\left(1 - \Pr(Z \leq .14)\right) \\ &= 2\left(1 - \textbf{pnorm}(0.14)\right) \\ &= 0.88866\end{aligned}$$

Since this p-value is larger than $\alpha = 0.1$ we fail to reject this null hypothesis and conclude that there is no evidence that returns to experience differ based on education levels.

Polynomial and Interaction Modeling: Questions

Questions?

Modeling Choices: Managing Model Complexity

As we have seen in the last couple examples, adding polynomial and interaction terms can allow us to build more flexible models that can better approximate the "true" relationship between our explanatory and response variables.

Question: Why stop at one interaction term and a polynomial? Why not keep adding terms?

- If we have enough data, this can be a good idea. Adding more terms can lead to a more accurate model. However:
- Adding more terms can reduce the interpetability of our model
  - How do we interpret a coeffecient on $X_1^2 \cdot X_2^3$?
- Adding more terms can lead to overfitting.

Modeling Choices: Managing Model Complexity

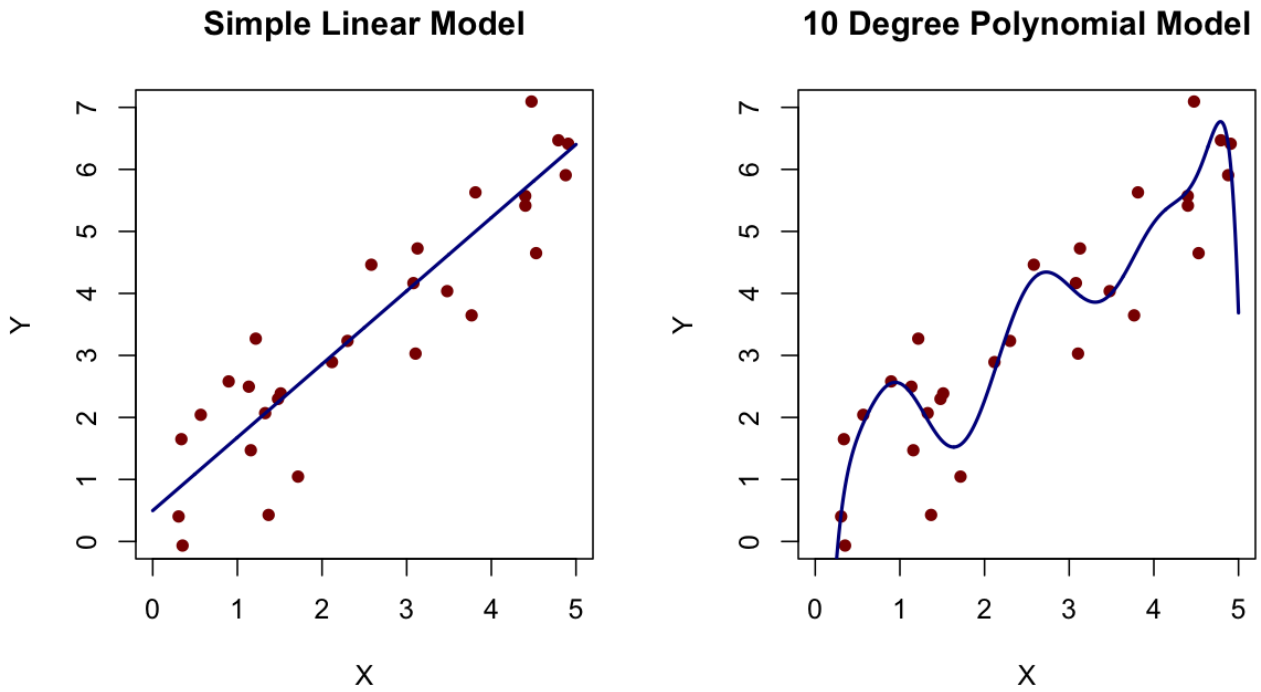
Essentially, when we add more terms to our model, we are balancing two effects

1. Adding more terms can allow the model to better approximate the true relationship between the outcome and the explanatory variable (Good).
    - We can allow for more general marginal effects, etc.
    - Often this is referred to as reducing the "bias" or reducing the "approximation error"
2. On the other hand, adding more terms means that we have more parameters that we need to estimate using the same amount of data. This means that our coefficient estimates are, on average, further from their true values (Bad).
    - This is often referred to as increasing the "variance" of our model, balancing these two effects is referred to as managing the "bias-variance" tradeoff.

Modeling Choices: Managing Model Complexity

In general our estimates are going to provide a better fit to our data than to the true population, this problem is exacerbated when we start to add more terms. Let's see an example of this in practice.

On the left hand side we see a simple linear model, whereas on the right hand side we see a 10 degree polynomial model.



The polynomial model fits the initial data much better than the linear one.

Modeling Choices: Managing Model Complexity

How do we determine whether we are overfitting? A first guess might be to take the approach we took to model evaluation in Single Linear Regression and stop adding terms when our R^2 falls. Just like in Single Linear Regression we can define

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

where as before $\text{SSR} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$, the sum of squares due to regression, $\text{SST} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$, the total sum of squares, and $\text{SSE} = \sum_{i=1}^{n}\hat{\epsilon}_i^2$ is the sum of squared errors.

Modeling Choices: Managing Model Complexity

Problem: The R^2 will fall as long as we keep adding terms.

Why: To see this let's use the representation $R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$. Across all models, $\text{SST} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ will remain the same. However, the SSE will keep falling as we add more terms.

Modeling Choices: Managing Model Complexity

Why does the SSE fall as we add more terms? Note that we essentially choose our estimated parameters to minimize the SSE

$$\hat{\beta}_0, \hat{\beta}_1 = \arg\min_{b_0, b_1} \frac{1}{n} \sum_{i=1}^{n} (\underbrace{Y_i - b_0 - b_1 X_{1,i}}_{\hat{\epsilon}_i})^2$$

$$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2 = \arg\min_{b_0, b_1, b_2} \sum_{i=1}^{n} (\underbrace{Y_i - b_0 - b_1 X_{1,i} - b_2 X_{2,i}}_{\hat{\epsilon}_i})^2$$

When using a third term, we could always recover the SSE from using two terms by keeping $\hat{\beta}_0$ and $\hat{\beta}_1$ the same and setting $\hat{\beta}_2 = 0$. The SSE mechanically must (weakly) fall when adding an additional term.

Since

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

this means that R^2 also must increase as we add more terms.

Modling Choices: Managing Model Complexity

For this reason, when evaluating multiple linear regression model we often use the adjusted R-squared. The adjusted R^2 for a model with p terms and an intercept is given:

$$\text{adj. } R^2 = 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)}.$$

Comments:

- The adjusted R^2 penalizes model complexity, it falls as p increases.
- A first order approach to determining what model to use would be to add a potential (carefully thought through) term only if the adjusted R^2 increases after adding the term.
    - We will go over more formal approaches in the next lecture.
- Adjusted R^2 will be reported by most statistical software when running a multiple linear regression model.

Managing Model Complexity: Questions

Questions?