HAD5772: Intermediate Statistics

Regression

Alex Hoagland, University of Toronto

February 28, 2024

Introduction

Let's start with **associative questions**:

1. How are one's social determinants of health associated with access and utilization?
2. What are the associations between expanded coverage and takeup of services?
3. What factors are associated with improved hospital quality and lower costs?

Introduction

Let's start with **associative questions**:

1. How are one's social determinants of health associated with access and utilization?
2. What are the associations between expanded coverage and takeup of services?
3. What factors are associated with improved hospital quality and lower costs?

Regression analysis allows us to explore how one random variable X *is associated* with another random variable Y

Introduction

Let's start with **associative questions**:

1. How are one's social determinants of health associated with access and utilization?
2. What are the associations between expanded coverage and takeup of services?
3. What factors are associated with improved hospital quality and lower costs?

Regression analysis allows us to explore how one random variable X *is associated* with another random variable Y

All of our previous work has gone into this:

- Knowledge about random variables (correlation, etc.)
- Estimating relationship (MLE)
- Confidence intervals, and Hypothesis Testing

The Regression: Basics

A **regression equation** is an equation that relates random variables in a **linear** fashion:

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

- Y is the **dependent** variable or the **outcome** variable
- X is the **independent** variable or the **regressor/covariate**
- ε denotes **error**—how randomness is preserved

A **regression equation** is an equation that relates random variables in a **linear** fashion:

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

- Y is the **dependent** variable or the **outcome** variable
- X is the **independent** variable or the **regressor/covariate**
- ε denotes **error**—how randomness is preserved

β_1 tells us the **strength of the relationship** between X and Y

Regression & Causality

We have to be **careful** when interpreting β_1:

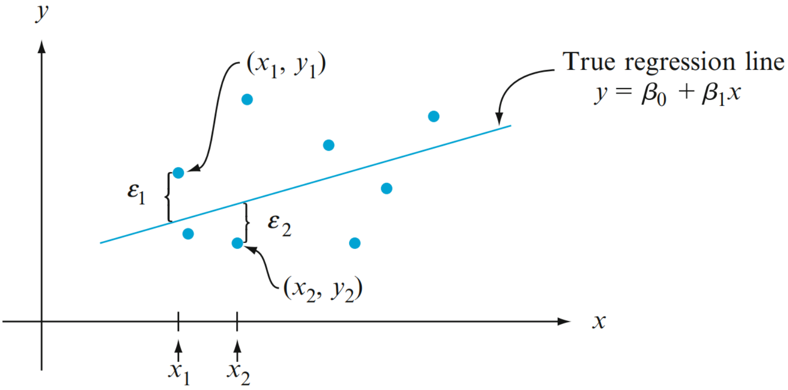
- **Correlation** is *not* **causation**!
- Many potential roadblocks between regression and causality:
  - Lack of control variables (confounders)
  - Reverse causality
  - Selection bias (non-random data)

Remember to **interpret results with caution** throughout

12.1–12.2: SIMPLE LINEAR MODELS

The Regression Equation

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



The Regression Equation

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

- β_1 is interpreted as **the change in Y associated with a 1 unit change in X**
- β_0 is the **expected value** of Y when $X = 0$.
  - **Q:** When would this be meaningful?

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

- β_1 is interpreted as **the change in Y associated with a 1 unit change in X**
- β_0 is the **expected value** of Y when $X = 0$.
  - **Q:** When would this be meaningful?

Theorem: Simple Linear Regression

There exist $\theta = (\beta_0, \beta_1, \sigma^2)$ solving Equation 1.

Using Info about X to get Y

How can we use the regression equation?

1. Easy translation from moments of X to moments of Y

Using Info about X to get Y

How can we use the regression equation?

1. Easy translation from moments of X to moments of Y

Notice:

$$\begin{aligned}
\mathbb{E}[Y|X] &= \mathbb{E}[\beta_0 + \beta_1 X + \varepsilon|X] \\
&= \beta_0 + \beta_1 X + \mathbb{E}[\varepsilon] \\
&= \beta_0 + \beta_1 X \\
\mathbb{V}[Y|X] &= \mathbb{V}[\beta_0] + \mathbb{V}[\beta_1 X|X] + \mathbb{V}[\varepsilon] \\
&= 0 + 0 + \sigma^2 \\
= \sigma^2
\end{aligned}$$

Using Info about X to get Y

How can we use the regression equation?

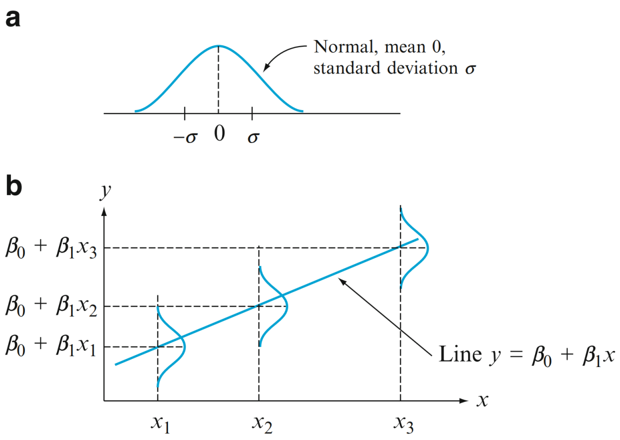
1. Easy translation from moments of X to moments of Y

Notice:

$$\begin{aligned}
\mathbb{E}[Y|X] &= \mathbb{E}[\beta_0 + \beta_1 X + \varepsilon|X] \\
&= \beta_0 + \beta_1 X + \mathbb{E}[\varepsilon] \\
&= \beta_0 + \beta_1 X \\
\mathbb{V}[Y|X] &= \mathbb{V}[\beta_0] + \mathbb{V}[\beta_1 X|X] + \mathbb{V}[\varepsilon] \\
&= 0 + 0 + \sigma^2 \\
= \sigma^2
\end{aligned}$$

The regression line is the line of mean Y's given data

Homogeneous Errors Visualized



Visualizing Regression

- This website can be used to visualize what's going on
- How many lines can I construct? Which one should I pick?

Estimating a Regression

We want a line that correctly reflects that **linear association** between X and Y

- That is, want a line that **best fits** our data
- Want to minimize deviations (ε_i)
- Can't minimize these directly, because they should sum to 0
- Instead, minimize the **squared residuals**

Estimating a Regression

We want a line that correctly reflects that **linear association** between X and Y

- That is, want a line that **best fits** our data
- Want to minimize deviations (ε_i)
- Can't minimize these directly, because they should sum to 0
- Instead, minimize the **squared residuals**

Proposition: Ordinary Least Squares (OLS)

For a data set $\{(x_i, y_i)\}_{i=1}^n$, the **OLS** regression parameters $(\hat{\beta}_0, \hat{\beta}_1)$ are those that minimize $\sum_{i=1}^n u_i^2$

Ordinary Least Squares Regression

We can solve for these directly:

$$(\hat{\beta}_0, \hat{\beta}_1) = \text{argmin}_{\beta_0, \beta_1} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right\}$$

Ordinary Least Squares Regression

We can solve for these directly:

$$(\hat{\beta}_0, \hat{\beta}_1) = \text{argmin}_{\beta_0,\beta_1} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right\}$$

Taking derivatives:

$$\frac{\partial}{\partial \beta_0} = \sum_{i=1}^{n} -2(y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial}{\partial \beta_1} = \sum_{i=1}^{n} -2x_i(y_i - \beta_0 - \beta_1 x_i)$$

Ordinary Least Squares Regression

We can solve for these directly:

$$(\hat{\beta}_0, \hat{\beta}_1) = \text{argmin}_{\beta_0,\beta_1} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right\}$$

Taking derivatives:

$$\frac{\partial}{\partial \beta_0} = \sum_{i=1}^{n} -2(y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial}{\partial \beta_1} = \sum_{i=1}^{n} -2x_i(y_i - \beta_0 - \beta_1 x_i)$$

Setting these equal to 0 and solving yields:

$$\hat{\beta}_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

How much noise is there in my regression?

To form confidence intervals around $\vec{\beta}$, we need an estimate of σ^2:

- Define a regression's **residuals** as the differences between Y and **predicted** Y:

$$\begin{aligned} \hat{u}_i &= y_i - \hat{y}_i \\ &= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \end{aligned}$$

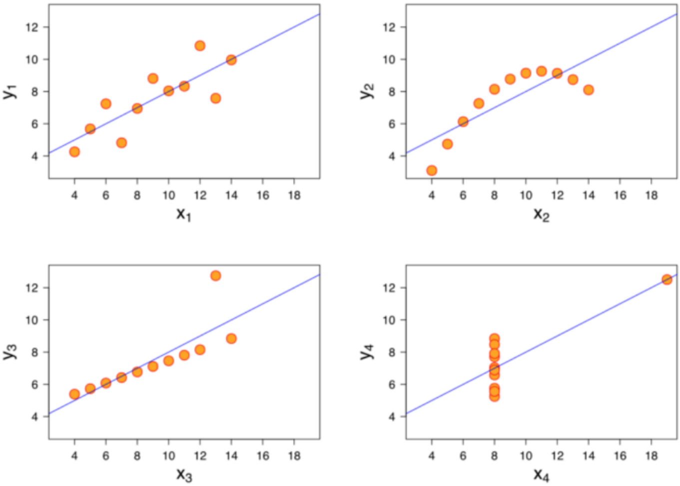
- Then, we can define the corresponding **sum of squared errors (SSE)** as $\sum u_i^2$
  - This should be minimized by our construction
  - Shortcut formula: $\text{SSE} = \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i$
- From this, we can estimate σ^2 by

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n-2}$$

How good is my regression fit?

The OLS prediction is not made equal for all data!

- Remember **Anscombe's Quartet?**



How good is my regression fit?

How can we determine **how well** my regression line matches up with data?

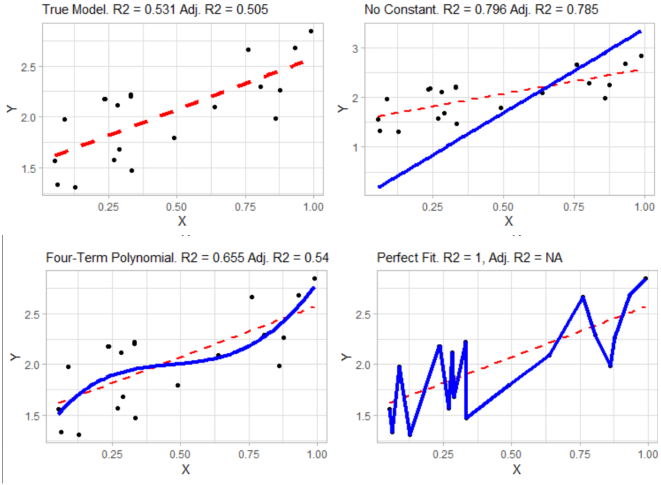
- Want a ratio of **variation in Y** that can be explained by **variation in X**
- If SSE is variation in **error**, then:

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

  where we define $\text{SST} = \sum(y_i - \bar{y})^2$
- The R^2 is a fraction (percentage) of overall movement in one r.v. explained by movement in the other. Which of the 4 regression lines should have the highest R^2?

Is the R^2 a panacea?

The R^2 is a pretty **unreliable** estimate of model fit:



Is the R^2 a panacea?

The R^2 is a pretty **unreliable** estimate of model fit:

- Don't use the R^2 as the only metric to pick a model!
- Good models can have **low** R^2 (especially if problem is noisy)
- Bad models can have **high** R^2 in cases of **overfitting**

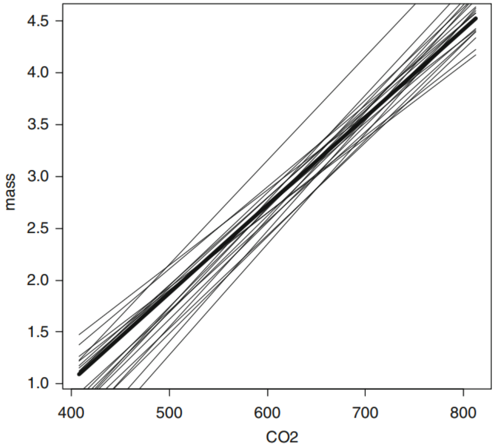
In general, R^2 is very **context dependent**

- Check out this article for a more in-depth discussion of its drawbacks
- And this one for a discussion of alternative measures

12.3: INFERENCE ON β_1

Just like all other parameters so far, $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)$ are **statistics** b/c they are **functions of random data**

- Hence, regression line **varies** based on our sample



Dealing with variation: sampling

Hence, just as for other parameters (e.g., $\mu, \sigma^2, p, \ldots$), we can **infer** things about $(\vec{\beta}, \sigma)$ from our estimates.

- We will focus mainly on $\hat{\beta}_1$.
- **Q:** Why do you think this is the most important estimate?

Dealing with variation: sampling

Proposition: Sampling Distribution of $\hat{\beta}_1$

1. $\mathbb{E}[\hat{\beta}_1] = \beta_1$ **(unbiasedness)**
2. $\mathbb{V}[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}}$, where

$$S_{xx} = \sum (x_i - \overline{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

   - Note: **does this show consistency? Why/why not?**

3. $\hat{\beta}_1 \sim \mathcal{N}$ **immediately**, not asymptotically. In particular,

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{S_{xx}}} \sim \mathcal{N}(0, 1)$$

Dealing with variation: sampling

These properties all come from Chapter 6!

These properties all come from Chapter 6!

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} \\
&= \frac{\sum\left[(x_i - \overline{x})y_i - (x_i - \overline{x})\overline{y}\right]}{S_{xx}} \\
&= \frac{\sum(x_i - \overline{x})y_i - \overline{y}\sum(x_i - \overline{x})}{S_{xx}} \\
&= \frac{\sum(x_i - \overline{x})y_i}{S_{xx}} \\
&= \sum c_i y_i, \text{ where } c_i = \frac{x_i - \overline{x}}{S_{xx}}
\end{aligned}
$$

Dealing with variation: sampling

These properties all come from Chapter 6!

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} \\
&= \frac{\sum\left[(x_i - \overline{x})y_i - (x_i - \overline{x})\overline{y}\right]}{S_{xx}} \\
&= \frac{\sum(x_i - \overline{x})y_i - \overline{y}\sum(x_i - \overline{x})}{S_{xx}} \\
&= \frac{\sum(x_i - \overline{x})y_i}{S_{xx}} \\
&= \sum c_i y_i, \text{ where } c_i = \frac{x_i - \overline{x}}{S_{xx}}
\end{aligned}
$$

Since $\hat{\beta}_1$ is a linear combination of independent, normally distributed random variables, we get the properties above!

Proposition: Asymptotic Distribution of $\hat{\sigma}^2$

1. $\hat{\beta}_1$ and $\hat{\sigma}^2$ are independent, and
2. $(n-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$

Proposition: Asymptotic Distribution of $\hat{\sigma}^2$

1. $\hat{\beta}_1$ and $\hat{\sigma}^2$ are independent, and
2. $(n-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$

Because of this, it follows (with some omitted algebra) that

$$T = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/S_{xx}} \sim t(n-2)$$

We can therefore **build confidence intervals** and **do hypothesis testing** on $\hat{\beta}_1$!

To build a confidence interval, we start with

$$P\left(-t_{\alpha/2}(n-2) < \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/S_{xx}} < t_{\alpha/2}(n-2)\right) = 1 - \alpha$$

Confidence Intervals for β_1

To build a confidence interval, we start with

$$P\left(-t_{\alpha/2}(n-2) < \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/S_{xx}} < t_{\alpha/2}(n-2)\right) = 1 - \alpha$$

Solving this for β_1 yields

$$\beta_1 \in \left[\hat{\beta}_1 \pm t_{\alpha/2}(n-2) \times \frac{\hat{\sigma}}{S_{xx}}\right]$$

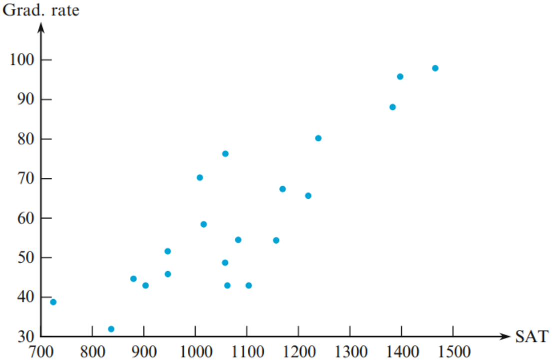
An Example Confidence Interval

We have data on graduation rates and freshman test scores:

	Rank	University	Grad rate	SAT	Private or State
1	2	Princeton	98	1465.00	P
2	13	Brown	96	1395.00	P
3	15	Johns Hopkins	88	1380.00	P
4	69	Pittsburgh	65	1215.00	S
5	77	SUNY-Binghamton	80	1235.00	S
6	94	Kansas	58	1011.10	S
7	102	Dayton	76	1055.54	P
8	107	Illinois Inst Tech	67	1166.65	P
9	125	Arkansas	48	1055.54	S
10	139	Florida Inst Tech	54	1155.00	P
11	147	New Mexico Inst Mining	42	1099.99	S
12	158	Temple	54	1080.00	S
13	172	Montana	45	944.43	S
14	174	New Mexico	42	899.99	S
15	178	South Dakota	51	944.43	S
16	183	Virginia Commonwealth	42	1060.00	S
17	186	Widener	70	1005.00	P
18	187	Alabama A&M	38	722.21	S
19	243	Toledo	44	877.77	S
20	245	Wayne State	31	833.32	S

An Example Confidence Interval

We have data on graduation rates and freshman test scores:



An Example Confidence Interval

We have data on graduation rates and freshman test scores: What is the 95% confidence interval for $\hat{\beta}_1$?

1. Calculate **relevant summary statistics** for the regression

$$\sum x_i = 21{,}600.97 \qquad \sum y_i = 1189 \qquad \sum x_i^2 = 24{,}034{,}220.545$$
$$\sum x_i y_i = 1{,}346{,}524.53 \qquad \sum y_i^2 = 78{,}113$$

An Example Confidence Interval

We have data on graduation rates and freshman test scores: What is the 95% confidence interval for $\hat{\beta}_1$?

1. Calculate **relevant summary statistics** for the regression
2. Obtain **estimates** for $\hat{\beta}_0, \hat{\beta}_1, S_{xx}, \mathrm{SST}, \mathrm{SSE}, R^2$, and $\hat{\sigma}^2$
    - Which of these do we need for the CI, and which are superfluous?
    - In this case, $\hat{\beta}_1 = 0.089$, $S_{xx} = 704125.298$, and $\hat{\sigma}^2 = 105.9$.

An Example Confidence Interval

We have data on graduation rates and freshman test scores:
What is the 95% confidence interval for $\hat{\beta}_1$?

1. Calculate **relevant summary statistics** for the regression
2. Obtain **estimates** for $\hat{\beta}_0, \hat{\beta}_1, S_{xx}, \text{SST}, \text{SSE}, R^2$, and $\hat{\sigma}^2$
   - In this case, $\hat{\beta}_1 = 0.089$, $S_{xx} = 704125.298$, and $\hat{\sigma}^2 = 105.9$.
3. Estimate the **standard error** of $\hat{\beta}_1$

$$\hat{\sigma}_{\hat{\beta}_1} = \frac{\hat{\sigma}}{\sqrt{S_{xx}}} = \frac{10.29}{\sqrt{704125.298}} = 0.0123.$$

An Example Confidence Interval

We have data on graduation rates and freshman test scores: What is the 95% confidence interval for $\hat{\beta}_1$?

1. Calculate **relevant summary statistics** for the regression
2. Obtain **estimates** for $\hat{\beta}_0, \hat{\beta}_1, S_{xx}, \text{SST}, \text{SSE}, R^2$, and $\hat{\sigma}^2$
   - In this case, $\hat{\beta}_1 = 0.089$, $S_{xx} = 704125.298$, and $\hat{\sigma}^2 = 105.9$.
3. Estimate the **standard error** of $\hat{\beta}_1$
   - The standard error of $\hat{\beta}_1$ is about 0.0123.
4. Identify the appropriate **critical value**
   - We look for the value $t_{0.025}(18) = 2.101$.

An Example Confidence Interval

We have data on graduation rates and freshman test scores: What is the 95% confidence interval for $\hat{\beta}_1$?

1. Calculate **relevant summary statistics** for the regression
2. Obtain **estimates** for $\hat{\beta}_0, \hat{\beta}_1, S_{xx}, \text{SST}, \text{SSE}, R^2$, and $\hat{\sigma}^2$
   - In this case, $\hat{\beta}_1 = 0.089$, $S_{xx} = 704125.298$, and $\hat{\sigma}^2 = 105.9$.
3. Estimate the **standard error** of $\hat{\beta}_1$
   - The standard error of $\hat{\beta}_1$ is about 0.0123.
4. Identify the appropriate **critical value**
   - We look for the value $t_{0.025}(18) = 2.101$.
5. Calculate!
   - $\beta_1 \in [0.0885 \pm (2.101)(0.0123)] = (0.063, 0.114)$

What does this mean?

Hypothesis Testing and $\hat{\beta}_1$

We can do **any kind** of hypothesis test about β_1 given our sample

- Most often, we care about **model utility**:

$$\mathcal{H}_0 : \beta_1 = 0$$
$$\mathcal{H}_1 : \beta_1 \neq 0$$

Hypothesis Testing and $\hat{\beta}_1$

We can do **any kind** of hypothesis test about β_1 given our sample

- Most often, we care about **model utility**:

$$\mathcal{H}_0 : \beta_1 = 0$$
$$\mathcal{H}_1 : \beta_1 \neq 0$$

- Under this null hypothesis, $\mathbb{E}[y|x] = \beta_0 + 0X = \beta_0$, so that **$X$ and Y are independent**
- The associated test statistic is (you guessed it!):

$$t = \frac{\hat{\beta}_1 - \mathbf{0}}{\hat{\sigma}/\sqrt{S_{xx}}}$$

An Example Hypothesis Test

In our example from before,

- We estimated a coefficient $\hat{\beta}_1 = 0.089$, and
- a standard error of $\hat{\sigma}/\sqrt{S_{xx}} = 0.0123$.

Hence, the test statistic for the **model utility test** is

$$t = \frac{0.089}{0.01226} = 7.238.$$

An Example Hypothesis Test

In our example from before,

- We estimated a coefficient $\hat{\beta}_1 = 0.089$, and
- a standard error of $\hat{\sigma}/\sqrt{S_{xx}} = 0.0123$.

Hence, the test statistic for the **model utility test** is

$$t = \frac{0.089}{0.01226} = 7.238.$$

- Clearly, this is bigger than a critical value of about 2, so we reject the null hypothesis
- p-value is about 0.000, so we'd reject with almost any confidence level.
- We conclude that X **gives information** about Y.
- **Could we have seen this already? How?**

Predicting Values with a Regression Line

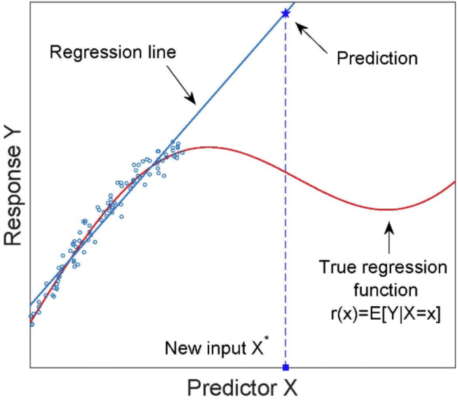
We won't cover prediction much (see 12.4 if you're interested)

- Regression lines can be used to predict $\hat{y}$ given some x^*
- To get a prediction, **just plug x^* into the regression equation!**
- Danger: regression fit is only good **where you have data**

Predicting Values with a Regression Line

We won't cover prediction much (see 12.4 if you're interested)

- Regression lines can be used to predict $\hat{y}$ given some x^*
- To get a prediction, **just plug x^* into the regression equation!**
- Danger: regression fit is only good **where you have data**



Predicting Values with a Regression Line

We won't cover prediction much (see 12.4 if you're interested)

- Regression lines can be used to predict $\hat{y}$ given some x^*
- To get a prediction, **just plug x^* into the regression equation!**
- Danger: regression fit is only good **where you have data**

12.5–12.6: CORRELATION & MODEL SELECTION

Why are we Running Regressions?

- Generally, we use regressions to tell us about the **associations** between X and Y.
- That is, we care about the **correlation** of the two variables

Why are we Running Regressions?

- Generally, we use regressions to tell us about the **associations** between X and Y.
- That is, we care about the **correlation** of the two variables
- Recall that the R^2 measured how much **variation in Y** could be explained by **variation in X**

Why are we Running Regressions?

- Generally, we use regressions to tell us about the **associations** between X and Y.
- That is, we care about the **correlation** of the two variables
- Recall that the R^2 measured how much **variation in Y** could be explained by **variation in X**

In this section, we'll relate R^2 to the **correlation** $\rho_{X,Y}$

From Summary Stats to Sample Correlation

When estimating a regression, we calculated

$$
\begin{aligned}
S_{xy} &= \sum (x_i - \overline{x})(y_i - \overline{y}) \\
&= \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}
\end{aligned}
$$

From Summary Stats to Sample Correlation

When estimating a regression, we calculated

$$S_{xy} = \sum (x_i - \overline{x})(y_i - \overline{y})$$
$$= \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

- This behaves just like a **covariance**
- Whenever x is larger than usual, y will be larger/small than its mean based on the sign of S_{xy}
- Just like covariance, this is **unit sensitive**

From Summary Stats to Sample Correlation

When estimating a regression, we calculated

$$S_{xy} = \sum (x_i - \overline{x})(y_i - \overline{y})$$
$$= \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

- This behaves just like a **covariance**
- Whenever x is larger than usual, y will be larger/small than its mean based on the sign of S_{xy}
- Just like covariance, this is **unit sensitive**

We therefore standardize this measure to the **sample correlation**:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Properties of the Sample Correlation

Once we standardize the sample covariance, we obtain something **very similar** to a correlation coefficient:

1. r is indepdent of the units of x and y, and $r \in [-1, 1]$ **always**
2. $r(x, y) = r(y, x)$, so that the ordering of y and x as dependent/independent variable is **irrelevant**
3. $r = \pm 1 \Leftrightarrow x$ and y have a **perfectly linear** relationship
4. $(r)^2 = R^2$

Properties of the Sample Correlation

Once we standardize the sample covariance, we obtain something **very similar** to a correlation coefficient:

1. r is indepdent of the units of x and y, and $r \in [-1, 1]$ **always**
2. $r(x, y) = r(y, x)$, so that the ordering of y and x as dependent/independent variable is **irrelevant**
3. $r = \pm 1 \Leftrightarrow x$ and y have a **perfectly linear** relationship
4. $(r)^2 = R^2$

Some comments:

- What does property 2 tell us about inferring **causation**?
- What kind of relationship is r measuring? (from property 3)

How are r and ρ related?

The sample correlation r is **sample dependent**—hence, it provides an **estimate** of ρ:

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

How are r and ρ related?

The sample correlation r is **sample dependent**—hence, it provides an **estimate** of ρ:

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

- Hence, with the proper assumptions, we can use the data to **conduct inference** on ρ!
- In this case, we need to assume that $(x_i, y_i) \overset{i.i.d.}{\sim} BVN(\vec{\mu}, \Sigma)$

How are r and ρ related?

The sample correlation r is **sample dependent**—hence, it provides an **estimate** of ρ:

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

- Hence, with the proper assumptions, we can use the data to **conduct inference** on ρ!
- In this case, we need to assume that $(x_i, y_i) \overset{i.i.d.}{\sim} BVN(\vec{\mu}, \Sigma)$
  - Recall that in the BVN,

$$\begin{aligned} \mathbb{E}[Y|x] = \mu_y + (\rho\sigma_y/\sigma_x)(x - \mu_x) &\Rightarrow \beta_0 = \mu_y - \rho\mu_x\sigma_y/\sigma_x \\ &\Rightarrow \beta_1 = \rho\sigma_y/\sigma_x \end{aligned}$$

  - Hence, when $(x_i, y_i) \sim BVN$, linear regression is **the right way** to think about conditional behavior!

Hypothesis Testing and Correlations

Given the assumption of BVN, an appropriate test is:

$$\mathcal{H}_0 : \rho = 0$$
$$\mathcal{H}_1 : \rho \lesseqgtr 0$$
$$T = \frac{r\sqrt{n-2}}{\sqrt{1-R^2}} \sim t(n-2)$$

Hypothesis Testing and Correlations

Given the assumption of BVN, an appropriate test is:

$$\mathcal{H}_0 : \rho = 0$$
$$\mathcal{H}_1 : \rho \lessgtr 0$$
$$T = \frac{r\sqrt{n-2}}{\sqrt{1-R^2}} \sim t(n-2)$$

- Super easy to calculate!
- However, this test statistic only works when testing $\rho = 0$
- For more general ρ_0, need to use **Fisher transformation** (in textbook)
- Remember that even **high correlations** don't imply **causation**—see this site for examples

Model Adequacy

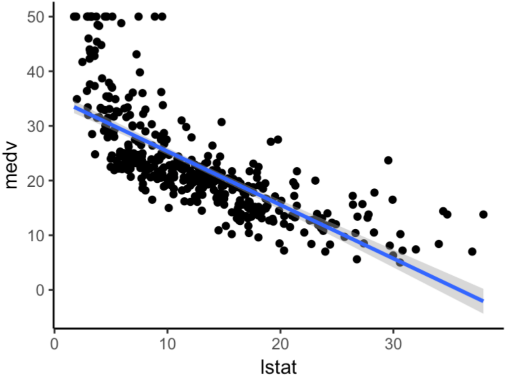
How do I know if my regression line is **useful**?

1. R^2 is easily manipulated
2. Still, can test for significance of $\hat{\beta}_1$ or r as a **first pass**
3. **What else can I do?** $\Rightarrow$ basically all of econometrics. Today:
   a. Check for linear relationship
   b. Analyze residuals
   c. Diagnostic Plots

Check for Linear Relationship

Regression only spots **linear relationships** among data.

- This is why we **always** plot our data before typing "reg"!
- The "twoway" command is helpful for super-imposing a regression line (and CI's) over a scatter plot



12.7–12.8: MULTIPLE REGRESSION

Expanding Our Horizons

What if we want to relate an outcome (y) to **more than 1** covariate ($\vec{x} = x_1, x_2, ...$)?

Example: What determines the price of yogurt?

Expanding Our Horizons

What if we want to relate an outcome (y) to **more than 1** covariate ($\vec{x} = x_1, x_2, ...$)?

Example: What determines the price of yogurt?

- The price of **complements** and **substitutes** (other brands/flabors, granola, etc.)
- **Input prices** (milk, fruit, etc.)
- Consumer **preferences**
- etc.

Expanding Our Horizons

What if we want to relate an outcome (y) to **more than 1** covariate ($\vec{x} = x_1, x_2, ...$)?

Example: What determines the price of yogurt?

- The price of **complements** and **substitutes** (other brands/flabors, granola, etc.)
- **Input prices** (milk, fruit, etc.)
- Consumer **preferences**
- etc.

Which factors are the most important? These types of questions can be answered using **multiple regression**

Multiple Regression

We expand the simple linear model in an additive fashion:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon,$$

where we continue to assume:

- $\mathbb{E}[\varepsilon] = 0$
- $\mathbb{V}[\varepsilon] = \sigma^2$
- $\varepsilon \sim^{\text{i.i.d.}} \mathcal{N}(0, \sigma^2)$

Multiple Regression

We expand the simple linear model in an **additive** fashion:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon,$$

where we continue to assume:

- $\mathbb{E}[\varepsilon] = 0$
- $\mathbb{V}[\varepsilon] = \sigma^2$
- $\varepsilon \sim^{\text{i.i.d.}} \mathcal{N}(0, \sigma^2)$

Once estimated, each $\hat{\beta}_i$ tells us the **average change in y** associated with a **one-unit change in x_i** *holding the other x's constant*

Estimating k Parameters

As we add covariates, our data expands:

$$\text{One } \textbf{observation}: \ (y_i, x_i) \Rightarrow (y_i, x_i^1, ..., x_i^k)$$

$$\textbf{Data:} \ \ (\{y_i\}_{i=1}^n, \{x_i\}_{i=1}^n) \Rightarrow (\vec{y}_{1\times n}, \boldsymbol{X}_{n\times k})$$

Estimating k Parameters

As we add covariates, our data expands:

$$\text{One } \textbf{observation}\text{: } (y_i, x_i) \Rightarrow (y_i, x_i^1, ..., x_i^k)$$
$$\textbf{Data: } (\{y_i\}_{i=1}^n, \{x_i\}_{i=1}^n) \Rightarrow (\vec{y}_{1\times n}, \boldsymbol{X}_{n\times k})$$

Our goal is still to **minimize sum of squared errors.** We can therefore estimate in **two ways**:

- Taking k derivatives and solving a system of equations
- Using **linear algebra**

Estimating k Parameters

As we add covariates, our data expands:

$$\text{One } \textbf{observation}: \; (y_i, x_i) \Rightarrow (y_i, x_i^1, ..., x_i^k)$$
$$\textbf{Data:} \; (\{y_i\}_{i=1}^n, \{x_i\}_{i=1}^n) \Rightarrow (\vec{y}_{1\times n}, \boldsymbol{X}_{n\times k})$$

Our goal is still to **minimize sum of squared errors.** We can therefore estimate in **two ways**:

- Taking k derivatives and solving a system of equations
- Using **linear algebra**
  - We will use this method to introduce the techniques

Some Linear Algebra Tools

To perform this estimation, we want to highlight **a few** linear algebra techniques:

1. **Matrix Multiplication**: Two matrices can be multiplied if their **dimensions** match in a certain way:
   - The # of **columns** of the first = the # of **rows** of the second
   - Easy test: Write the dimensions side by side!
   - Careful: Multiplication does *not* commute ($AB \neq BA$)!

Some Linear Algebra Tools

To perform this estimation, we want to highlight a few linear algebra techniques:

1. **Matrix Multiplication**: Two matrices can be multiplied if their **dimensions** match in a certain way:
2. **Matrix Transposes**: Flips the rows and columns of a matrix!:
   - So any cell x_{ij} becomes x_{ji}
   - Hence, any matrix $\boldsymbol{X}_{n \times k}$ has a transpose $\boldsymbol{X}'_{k \times n}$
   - In general, $(AB)' = B'A'$ and $(A - B)' = A' - B'$

Some Linear Algebra Tools

To perform this estimation, we want to highlight a few linear algebra techniques:

1. **Matrix Multiplication**: Two matrices can be multiplied if their **dimensions** match in a certain way:
2. **Matrix Transposes**: Flips the rows and columns of a matrix!:
3. **Matrix Inverses**: Two matrices who "undo" each other
   - An pair of invertible matrices satisfy $\boldsymbol{X}\boldsymbol{X}^{-1} = I_n$, where

$$I_n = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

   - To find the inverse of a 2×2 matrix A:

$$A^{-1} = \frac{1}{|A|} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

   - Other inverses are harder

Some Linear Algebra Tools

To perform this estimation, we want to highlight a few linear algebra techniques:

1. **Matrix Multiplication**: Two matrices can be multiplied if their **dimensions** match in a certain way:
2. **Matrix Transposes**: Flips the rows and columns of a matrix!:
3. **Matrix Inverses**: Two matrices who "undo" each other
4. **Vector Derivatives**: Generalizing calculus
   - In general, follow same pattern as regular derivatives:

$$\frac{\partial}{\partial \vec{v}}(\boldsymbol{A}\vec{v}) = \boldsymbol{A}$$

$$\frac{\partial}{\partial \vec{v}}(\vec{v}'\boldsymbol{A}\boldsymbol{A}'\vec{v}) = 2\boldsymbol{A}\vec{v}$$

We have 4 objects of interest: $\left\{\vec{y}, \vec{\beta}, \boldsymbol{X}, \vec{\varepsilon}\right\}$:

$$\vec{y}_{n\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \vec{\beta}_{k\times 1} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \vec{\varepsilon}_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\boldsymbol{X}_{n\times k} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$$

We have 4 objects of interest: $\left\{\vec{y}, \vec{\beta}, \boldsymbol{X}, \vec{\varepsilon}\right\}$:
We can therefore write our regression as

$$\vec{y} = \boldsymbol{X}\vec{\beta} + \vec{\varepsilon}$$

Our goal is to minimize the sum of squared errors:

$$\begin{aligned}\min_{\vec{\beta}}\ & (\vec{y} - \boldsymbol{X}\vec{\beta})'(\vec{y} - \boldsymbol{X}\vec{\beta}) \\ & = (\vec{y}' - \vec{\beta}'\boldsymbol{X}')(\vec{y} - \boldsymbol{X}\vec{\beta}) \\ & = \vec{y}'\vec{y} - 2\vec{\beta}'\boldsymbol{X}'\vec{y} + \vec{\beta}'\boldsymbol{X}'\boldsymbol{X}\vec{\beta}\end{aligned}$$

Re-writing the problem this way lets us take **only one derivative** w.r.t $\vec{\beta}$:

$$\min_{\vec{\beta}} \left(\vec{y}'\vec{y} - 2\vec{\beta}'\boldsymbol{X}'\vec{y} + \vec{\beta}'\boldsymbol{X}'\boldsymbol{X}\vec{\beta} \right)$$

$$\Rightarrow -2\boldsymbol{X}'\vec{y} + 2\boldsymbol{X}'\boldsymbol{X}\vec{\beta} \equiv 0$$

$$\boldsymbol{X}'\boldsymbol{X}\vec{\beta} = \boldsymbol{X}'\vec{y}$$

$$\vec{\beta} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\vec{y}$$

Example: Price of Yogurt

Suppose we want to regress **yogurt demand** on two factors: **price of milk** and **quantity of granola sold**. Our data is:

$$y = \begin{bmatrix} 30 \\ 55 \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 2 & 50 \\ 3 & 100 \end{bmatrix}$$

What is the associated $\vec{\beta}$ vector?

Suppose that we estimate a more complicated model and find:

$$\underset{(0.05)}{Q_{\text{Yogurt}} = 3} - \underset{(1.3)}{5P_{\text{dairy}}} + \underset{(1.5)}{2Q_{\text{granola}}} + \underset{(0.8)}{2.5P_{\text{Cottage Cheese}}} - \underset{(1.2)}{0.5P_{\text{gas}}}$$

1. How would we interpret β_3? What does this tell us about yogurt and cottage cheese?
2. Individual coefficients can be evaluated using **simple t-tests**
3. Confidence intervals are also straightforward
4. SSR, SSE, and R^2 calculated the same way

How Useful is Our Model? The F Test

Even if we throw in a bunch of regressors (a **kitchen sink** regression), how can we know our model is **useful**?

- In a simple model, we tested $\mathcal{H}_0 : \beta_1 = 0$.
- For multiple regressors, the test becomes

$$\mathcal{H}_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$\mathcal{H}_a : \text{At least one } \beta_i \neq 0 \text{ for } i \in \{1, 2, ..., k\}$$

How Useful is Our Model? The F Test

Even if we throw in a bunch of regressors (a **kitchen sink** regression), how can we know our model is **useful**?

- In a simple model, we tested $\mathcal{H}_0 : \beta_1 = 0$.
- For multiple regressors, the test becomes

$$\mathcal{H}_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$\mathcal{H}_a : \text{At least one } \beta_i \neq 0 \text{ for } i \in \{1, 2, ..., k\}$$

This is a **joint hypothesis test**, which has the following test statistic and distribution:

$$f = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{\text{SSR}/k}{\text{SSE}/(n-k-1)} \sim F_{\alpha,k,n-k-1}$$

QUESTIONS?