Math 244 Lecture Notes

Chapter 25 Day One: Linear Regression Day 1

Overview: Today, we will practice hypothesis tests for two proportions. There are four steps to hypothesis tests:
Linear Regression is used for examining the relationship between two variables–an **explanatory variable**, x, and a **response variable**, y. When done correctly, it allows us to make predictions about what should happen given a limited amount of information. Let's look at a quick review:

Example 1. Find the equation of the line that runs through the points $(1, 10)$ and $(3, 6)$.

Example 2. Plot the data using a scatter plot

Year	Average Cost of Bread $
1930	0.09
1940	0.10
1950	0.12
1960	0.22
1970	0.25
1980	0.50
1990	0.70
2000	1.64
2010	2.79
2015	1.98

Example 3. Describe the above scatter plot

We describe scatter plot by looking at...

- Shape: Does it look like a line? Parabola? Exponential? Etc.
- Direction: Is it mostly increasing or decreasing globally?
- Fit: How much does it look like the indicated shape (e.g. linear)? Is it a good fit? Okay fit? Poor fit?

Examples:

NOTE: We call the input the EXPLANATORY VARIABLE (it is the variable we control or independent variable) and the output the RESPONSE VARIABLE (it is the variable we measure or dependent variable). For example, the Year might be the explanatory variable and the Cost of Bread might be the response variable.

When describing the fit for a linear relationship, we use the terminology of "correlation". It simply tells us how to variables are related (it does not tell us the cause of the relationship though). r is the correlation coefficient. Here are properties of r:

- $-1 \leq r \leq 1$.
- $r > 0$ means the data is increasing. Think of it like slope!
- $r < 0$ means the data is decreasing. Think of it like slope!
- $r \approx 1$ is an almost perfect increasing linear fit.
- $r \approx -1$ is an almost perfect decreasing linear fit.
- $r \approx 0$ indicates that the variables are not correlated [may look like a cloud of points].

Examples:

By hand, we would calculate it using the formula

$$r = \Sigma \frac{Z_x * Z_y}{(n-1)}$$

where Z_x and Z_y are the z-scores for x-values and y-values respectively. This can be time consuming, so we often find it using technology instead...

TI-83/84: Go into [Stat]→[Edit] and type in your data in L1 and L2. Then,
go [Stat]→[Calc]→[LinReg(ax+b)].

TI-89: Go into [Stat/list Editor]→[Edit] and type in your data in list1 and list2. Then, go [F4:Calc]→[Regression

Our r for the Cost of Bread example was...

Since both $r = -1$ and $r = 1$ are considered good fits, we often consider R^2 instead [takes care of the negatives]. By design, $0 \leq R^2 \leq 1$. It's actually a probability!

- R^2 is the percentage of variation accounted for by our model.
- $1 - R^2$ is the percentage of variation left to luck or a lurking variable.
- If $R^2 \geq 0.85$ or 85%, we consider it a good fit.
- If $0.75 < R^2 < 0.85$, we consider it an okay fit.
- If $R^2 \leq 0.75$, we consider it a poor fit.

For our Cost of Bread example, our R^2 was...

Your calculator also gives you R^2 when you calculate r. Go back and verify it!
You may have noticed that your calculator does one better than that–it actually gives us our LINE of best fit! For our Cost of Bread example...

Example 4. We can use this line to make predictions. How much did bread cost in 1960? How much will it cost in 2016? How much did it cost in 1800?

Recall that the equation of a line is $y = mx + b$ where m is the slope and b is part of the y-intercept $(0, b)$. In Stats, we write this as $\hat{y} = b_1(x) + b_0$ where b_1 is our slope and b_0 is our y-intercept. Here's why:

Example 5. How do we find the slope?

$b_1 =$

Example 6. How do we find the y-intercept?

$b_0 =$

Recall that $\hat{y} = b_1(x) + b_0$. We calculate the slope using the formula $b_1 = r\frac{s_y}{s_x}$. We found the slope by plugging in the point $(\bar{x}, \bar{y})$ and solving for b_0.

Example 7. Find the equation of a line of best fit if $r = 0.80$, $\bar{x} = 10$, $s_x = 5$, $\bar{y} = 4$, and $s_y = 3$.

Let's verify our Cost of Bread equation:

More predictions: What will the cost of bread be in 1970?

Predictions are rarely perfect. Residuals are the error in our predictions. We find these by $e = y - \hat{y} =$ data value $-$ line value.

Example 8. Let's find the error in our prediction for the year 1970.

What do the errors typically look like?

Model: