Statistical Foundations of Data Science

Hypothesis testing

University of the Witwatersrand

2025

Review Question

- Three players roll dice. What's the probability that they all get different numbers?

Lesson Plan

- Review Question

Lesson Plan

- Review Question
- Hypothesis Tests

Lesson Plan

- Review Question
- Hypothesis Tests
- Z - Test

Lesson Plan

- Review Question
- Hypothesis Tests
- Z - Test
- T - Test

Review Question

- Three players roll dice. What's the probability that they all get different numbers?

Review Question

- Three players roll dice. What's the probability that they all get different numbers?
- $\frac{6}{6}\frac{5}{6}\frac{4}{6} = \frac{5}{9}$

The idea of a hypothesis test

- We have some natural idea of how the world is.

The idea of a hypothesis test

- We have some natural idea of how the world is.
- For example maybe we have a potion that we think cures cancer. The natural idea is that it doesn't do anything.

The idea of a hypothesis test

- We have some natural idea of how the world is.
- For example maybe we have a potion that we think cures cancer. The natural idea is that it doesn't do anything.
- This natural idea is called "the null hypothesis".

The idea of a hypothesis test

- We have some natural idea of how the world is.
- For example maybe we have a potion that we think cures cancer. The natural idea is that it doesn't do anything.
- This natural idea is called "the null hypothesis".
- We collect some data, and see if it's consistent with the natural idea (null hypothesis).

The idea of a hypothesis test

- We have some natural idea of how the world is.
- For example maybe we have a potion that we think cures cancer. The natural idea is that it doesn't do anything.
- This natural idea is called "the null hypothesis".
- We collect some data, and see if it's consistent with the natural idea (null hypothesis).
- If it isn't we reject the null hypothesis. The potion becomes a medicine.

The idea of a hypothesis test

- We have some natural idea of how the world is.
- For example maybe we have a potion that we think cures cancer. The natural idea is that it doesn't do anything.
- This natural idea is called "the null hypothesis".
- We collect some data, and see if it's consistent with the natural idea (null hypothesis).
- If it isn't we reject the null hypothesis. The potion becomes a medicine.
- If the data does fit the hypothesis we'll say we "fail to reject". Don't trust the random guy claiming to have a magical cure!

Hypothesis Testing - Example

- We have a coin. We suspect it's a fair coin. Most coins are (at least most coins are almost).

Hypothesis Testing - Example

- We have a coin. We suspect it's a fair coin. Most coins are (at least most coins are almost).
- We flip it 10 times to start with and get 6 heads. Is this consistent with it being fair?

Hypothesis Testing - Example

- We have a coin. We suspect it's a fair coin. Most coins are (at least most coins are almost).
- We flip it 10 times to start with and get 6 heads. Is this consistent with it being fair?
- Intuitively it really is. So we say that we "fail to reject" the null hypothesis that it's fair.

Hypothesis Testing - Example

- We have a coin. We suspect it's a fair coin. Most coins are (at least most coins are almost).
- We flip it 10 times to start with and get 6 heads. Is this consistent with it being fair?
- Intuitively it really is. So we say that we "fail to reject" the null hypothesis that it's fair.
- Why don't we say that we accept the null hypothesis???

Hypothesis Testing - Example

- We have a coin. We suspect it's a fair coin. Most coins are (at least most coins are almost).
- We flip it 10 times to start with and get 6 heads. Is this consistent with it being fair?
- Intuitively it really is. So we say that we "fail to reject" the null hypothesis that it's fair.
- Why don't we say that we accept the null hypothesis???
- Because as we continue to flip it we might reach 1000 flips and 600 heads.

Hypothesis Testing - Example

- We have a coin. We suspect it's a fair coin. Most coins are (at least most coins are almost).
- We flip it 10 times to start with and get 6 heads. Is this consistent with it being fair?
- Intuitively it really is. So we say that we "fail to reject" the null hypothesis that it's fair.
- Why don't we say that we accept the null hypothesis???
- Because as we continue to flip it we might reach 1000 flips and 600 heads.
- So at the point where we've got 6 out of 10 heads we reserve judgement.

Hypothesis Testing - Example

- We have a coin. We suspect it's a fair coin. Most coins are (at least most coins are almost).
- We flip it 10 times to start with and get 6 heads. Is this consistent with it being fair?
- Intuitively it really is. So we say that we "fail to reject" the null hypothesis that it's fair.
- Why don't we say that we accept the null hypothesis???
- Because as we continue to flip it we might reach 1000 flips and 600 heads.
- So at the point where we've got 6 out of 10 heads we reserve judgement.
- At the point where we have 9 out of 10 we do not.

Hypothesis Testing - Formalisation

- OK that's well and good but when do we actually decide to reject?

Hypothesis Testing - Formalisation

- OK that's well and good but when do we actually decide to reject?
- Well we use a tool called a p-value. This means the probability of seeing a result this extreme or more when the null hypothesis is true.

Hypothesis Testing - Formalisation

- OK that's well and good but when do we actually decide to reject?
- Well we use a tool called a p-value. This means the probability of seeing a result this extreme or more when the null hypothesis is true.
- Does random chance explain the potion curing 67 out of 100 patients when we usually see 58 recover?

Hypothesis Testing - Formalisation

- OK that's well and good but when do we actually decide to reject?
- Well we use a tool called a p-value. This means the probability of seeing a result this extreme or more when the null hypothesis is true.
- Does random chance explain the potion curing 67 out of 100 patients when we usually see 58 recover?
- Compare this to the Bayesian approach where we'd have a distribution on the coins probability of coming up heads and adjust.

Hypothesis Testing - Example

- Usually 58 percent of patients get healthy without intervention. Of the 100 who took the intervention 67 recovered.

Hypothesis Testing - Example

- Usually 58 percent of patients get healthy without intervention. Of the 100 who took the intervention 67 recovered.
- Null Hypothesis $H_0 : p = 0.58$.

Hypothesis Testing - Example

- Usually 58 percent of patients get healthy without intervention. Of the 100 who took the intervention 67 recovered.
- Null Hypothesis $H_0 : p = 0.58$.
- Altenrate Hypothesis $H_1 : p > 0.58$.

Hypothesis Testing - Example

- Usually 58 percent of patients get healthy without intervention. Of the 100 who took the intervention 67 recovered.
- Null Hypothesis $H_0 : p = 0.58$.
- Altenrate Hypothesis $H_1 : p > 0.58$.
- Under this hypothesis we get that the number of recoveries is distributed binomial with $n = 100$ and $p = 58$. So approximately normal (CLT) with mean 58 and variance $100 \times 0.58 \times 0.42 = 24.36$ (so standard deviation 4.9355850717)

Hypothesis Testing - Example

- Usually 58 percent of patients get healthy without intervention. Of the 100 who took the intervention 67 recovered.
- Null Hypothesis $H_0 : p = 0.58$.
- Altenrate Hypothesis $H_1 : p > 0.58$.
- Under this hypothesis we get that the number of recoveries is distributed binomial with $n = 100$ and $p = 58$. So approximately normal (CLT) with mean 58 and variance $100 \times 0.58 \times 0.42 = 24.36$ (so standard deviation 4.9355850717)
- This is a Z-score of $\frac{9}{4.9355850717} \approx 1.82$

Hypothesis Testing - Example

- Usually 58 percent of patients get healthy without intervention. Of the 100 who took the intervention 67 recovered.
- Null Hypothesis $H_0 : p = 0.58$.
- Altenrate Hypothesis $H_1 : p > 0.58$.
- Under this hypothesis we get that the number of recoveries is distributed binomial with $n = 100$ and $p = 58$. So approximately normal (CLT) with mean 58 and variance $100 \times 0.58 \times 0.42 = 24.36$ (so standard deviation 4.9355850717)
- This is a Z-score of $\frac{9}{4.9355850717} \approx 1.82$
- What's the probability of getting a more extreme than that? use a table and it's 0.034383.

Hypothesis Testing - Example repeat

- Usually 58 percent of patients get healthy without intervention. Of the 100 who took the intervention 67 recovered.

Hypothesis Testing - Example repeat

- Usually 58 percent of patients get healthy without intervention. Of the 100 who took the intervention 67 recovered.
- Null Hypothesis $H_0 : p = 0.58$.

Hypothesis Testing - Example repeat

- Usually 58 percent of patients get healthy without intervention. Of the 100 who took the intervention 67 recovered.
- Null Hypothesis $H_0 : p = 0.58$.
- Altenrate Hypothesis $H_1 : p \neq 0.58$.

Hypothesis Testing - Example repeat

- Usually 58 percent of patients get healthy without intervention. Of the 100 who took the intervention 67 recovered.
- Null Hypothesis $H_0 : p = 0.58$.
- Altenrate Hypothesis $H_1 : p \neq 0.58$.
- Under this hypothesis we get that the number of recoveries is distributed binomial with $n = 100$ and $p = 58$. So approximately normal (CLT) with mean 58 and variance $100 \times 0.58 \times 0.42 = 24.36$ (so standard deviation 4.9355850717)

Hypothesis Testing - Example repeat

- Usually 58 percent of patients get healthy without intervention. Of the 100 who took the intervention 67 recovered.
- Null Hypothesis $H_0 : p = 0.58$.
- Altenrate Hypothesis $H_1 : p \neq 0.58$.
- Under this hypothesis we get that the number of recoveries is distributed binomial with $n = 100$ and $p = 58$. So approximately normal (CLT) with mean 58 and variance $100 \times 0.58 \times 0.42 = 24.36$ (so standard deviation 4.9355850717)
- This is a Z-score of $\frac{9}{4.9355850717} \approx 1.82$

Hypothesis Testing - Example repeat

- Usually 58 percent of patients get healthy without intervention. Of the 100 who took the intervention 67 recovered.
- Null Hypothesis $H_0 : p = 0.58$.
- Altenrate Hypothesis $H_1 : p \neq 0.58$.
- Under this hypothesis we get that the number of recoveries is distributed binomial with $n = 100$ and $p = 58$. So approximately normal (CLT) with mean 58 and variance $100 \times 0.58 \times 0.42 = 24.36$ (so standard deviation 4.9355850717)
- This is a Z-score of $\frac{9}{4.9355850717} \approx 1.82$
- What's the probability of getting a more extreme than that? use a table and it's 0.068766.

Hypothesis Testing - When to actually reject???

- That's an engineering decision. By convention the cut-off is usually at five percent.

Hypothesis Testing - When to actually reject???

- That's an engineering decision. By convention the cut-off is usually at five percent.
- It's also convention to made it very clear that this is by convention and to complain about people not understanding that it's arbitrary (even though everyone seems to).

Hypothesis Testing - When to actually reject???

- That's an engineering decision. By convention the cut-off is usually at five percent.
- It's also convention to made it very clear that this is by convention and to complain about people not understanding that it's arbitrary (even though everyone seems to).
- That's probably sample bias because I only take abotu this stuff to data-scientists and statisticians. Maybe in other fields this is taken as some kind of weird gospel.

Hypothesis Testing - When to actually reject???

- That's an engineering decision. By convention the cut-off is usually at five percent.

Hypothesis Testing - When to actually reject???

- That's an engineering decision. By convention the cut-off is usually at five percent.
- It's also convention to made it very clear that this is by convention and to complain about people not understanding that it's arbitrary (even though everyone seems to).

Hypothesis Testing - When to actually reject???

- That's an engineering decision. By convention the cut-off is usually at five percent.
- It's also convention to made it very clear that this is by convention and to complain about people not understanding that it's arbitrary (even though everyone seems to).
- That's probably sample bias because I only take abotu this stuff to data-scientists and statisticians. Maybe in other fields this is taken as some kind of weird gospel.

Hypothesis Testing - Known variance

- First test. Assume that we have data which we know to be normal with known variance.

Hypothesis Testing - Known variance

- First test. Assume that we have data which we know to be normal with known variance.
- This is not a realistic test as if you think you know the variance, you probably know the mean.

Hypothesis Testing - Known variance

- First test. Assume that we have data which we know to be normal with known variance.
- This is not a realistic test as if you think you know the variance, you probably know the mean.
- But it's illustrative.

Hypothesis Testing - Known variance

- First test. Assume that we have data which we know to be normal with known variance.
- This is not a realistic test as if you think you know the variance, you probably know the mean.
- But it's illustrative.
- $H_0 : \mu = \mu_0$

Hypothesis Testing - Known variance

- First test. Assume that we have data which we know to be normal with known variance.
- This is not a realistic test as if you think you know the variance, you probably know the mean.
- But it's illustrative.
- $H_0 : \mu = \mu_0$
- $H_1 : \mu > \mu_0$

Hypothesis Testing - Known variance

- First test. Assume that we have data which we know to be normal with known variance.
- This is not a realistic test as if you think you know the variance, you probably know the mean.
- But it's illustrative.
- $H_0 : \mu = \mu_0$
- $H_1 : \mu > \mu_0$
- Compute $\overline{X}$

Hypothesis Testing - Known variance

- First test. Assume that we have data which we know to be normal with known variance.
- This is not a realistic test as if you think you know the variance, you probably know the mean.
- But it's illustrative.
- $H_0 : \mu = \mu_0$
- $H_1 : \mu > \mu_0$
- Compute $\overline{X}$
- Under H_0 we have $\overline{X}\tilde{N}(\mu, \frac{\sigma^2}{n})$

Hypothesis Testing - Known variance

- First test. Assume that we have data which we know to be normal with known variance.
- This is not a realistic test as if you think you know the variance, you probably know the mean.
- But it's illustrative.
- $H_0 : \mu = \mu_0$
- $H_1 : \mu > \mu_0$
- Compute $\overline{X}$
- Under H_0 we have $\overline{X}\tilde{N}(\mu, \frac{\sigma^2}{n})$
- Compute Z-score and p-value

Example

- We have a company that sells boxes of cereal and claims that on average they contain $500g$ of cereal.

Example

- We have a company that sells boxes of cereal and claims that on average they contain $500g$ of cereal.
- We think that they might be lying.

Example

- We have a company that sells boxes of cereal and claims that on average they contain $500g$ of cereal.
- We think that they might be lying.
- We (somehow) know that the variance is 100 and that the distribution is normal

Example

- We have a company that sells boxes of cereal and claims that on average they contain $500g$ of cereal.
- We think that they might be lying.
- We (somehow) know that the variance is 100 and that the distribution is normal
- $H_0 : \mu = 500$

Example

- We have a company that sells boxes of cereal and claims that on average they contain $500g$ of cereal.
- We think that they might be lying.
- We (somehow) know that the variance is 100 and that the distribution is normal
- $H_0 : \mu = 500$
- $H_1 : \mu < 500$

Example

- We have a company that sells boxes of cereal and claims that on average they contain $500g$ of cereal.
- We think that they might be lying.
- We (somehow) know that the variance is 100 and that the distribution is normal
- $H_0 : \mu = 500$
- $H_1 : \mu < 500$
- $\overline{X} = 490$ and $n = 25$.

Example

- We have a company that sells boxes of cereal and claims that on average they contain $500g$ of cereal.
- We think that they might be lying.
- We (somehow) know that the variance is 100 and that the distribution is normal
- $H_0 : \mu = 500$
- $H_1 : \mu < 500$
- $\overline{X} = 490$ and $n = 25$.
- $Z = \frac{490-500}{\sqrt{\frac{100}{25}}} = -5$

Example

- We have a company that sells boxes of cereal and claims that on average they contain 500g of cereal.
- We think that they might be lying.
- We (somehow) know that the variance is 100 and that the distribution is normal
- $H_0 : \mu = 500$
- $H_1 : \mu < 500$
- $\overline{X} = 490$ and $n = 25$.
- $Z = \frac{490-500}{\sqrt{\frac{100}{25}}} = -5$
- Yeah they lied!

One Sample t-test

- More often we'll know the a population mean μ but not a variance σ^2.

One Sample t-test

- More often we'll know the a population mean μ but not a variance σ^2.
- We'll compute $T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$

One Sample t-test

- More often we'll know the a population mean μ but not a variance σ^2.
- We'll compute $T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$
- Distributed T with $n - 1$ degrees of freedom

One Sample t-test

- More often we'll know the a population mean μ but not a variance σ^2.
- We'll compute $T = \frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}$
- Distributed T with $n-1$ degrees of freedom
- s is the sample standard deviation.

One Sample t-test

- More often we'll know the a population mean μ but not a variance σ^2.
- We'll compute $T = \frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}$
- Distributed T with $n-1$ degrees of freedom
- s is the sample standard deviation.
- $s^2 = \frac{(x_i-\overline{X})^2}{n-1}$

Example

- Four sumo wrestlers weigh 160kg, 165kg, 170kg and 180kg

Example

- Four sumo wrestlers weigh 160kg, 165kg, 170kg and 180kg
- The population of sumo wrestlers has average weight is 175kg. Is this group below average?

Example

- Four sumo wrestlers weigh 160kg, 165kg, 170kg and 180kg
- The population of sumo wrestlers has average weight is 175kg. Is this group below average?
- $H_0 : \mu_0 = \mu$

Example

- Four sumo wrestlers weigh 160kg, 165kg, 170kg and 180kg
- The population of sumo wrestlers has average weight is 175kg. Is this group below average?
- $H_0 : \mu_0 = \mu$
- $H_1 : \mu_0 < \mu$

Example

- Four sumo wrestlers weigh 160kg, 165kg, 170kg and 180kg
- The population of sumo wrestlers has average weight is 175kg. Is this group below average?
- $H_0 : \mu_0 = \mu$
- $H_1 : \mu_0 < \mu$
- $\overline{X} = 170$. $s^2 = \frac{100+25+100}{3} = 75$

Example

- Four sumo wrestlers weigh 160kg, 165kg, 170kg and 180kg
- The population of sumo wrestlers has average weight is 175kg. Is this group below average?
- $H_0 : \mu_0 = \mu$
- $H_1 : \mu_0 < \mu$
- $\overline{X} = 170.\ s^2 = \frac{100+25+100}{3} = 75$
- $T = \frac{170-175}{\frac{\sqrt{75}}{\sqrt{3}}} = -1$

Example

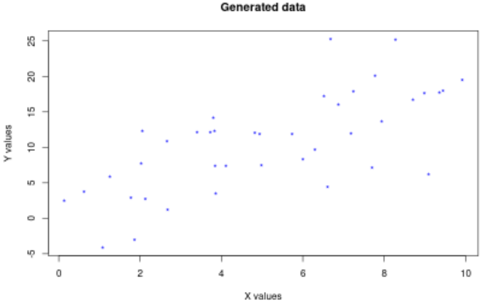
- Four sumo wrestlers weigh 160kg, 165kg, 170kg and 180kg
- The population of sumo wrestlers has average weight is 175kg. Is this group below average?
- $H_0 : \mu_0 = \mu$
- $H_1 : \mu_0 < \mu$
- $\overline{X} = 170$. $s^2 = \frac{100+25+100}{3} = 75$
- $T = \frac{170-175}{\frac{\sqrt{75}}{\sqrt{3}}} = -1$
- p-value is around about 0.19 We don't have evidence to comclude our guys are too small

Given some data

- Given some data we'd like to predict Y from X

Given some data

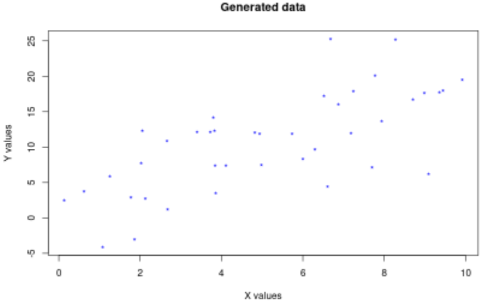
- Given some data we'd like to predict Y from X



-

Given some data

- Given some data we'd like to predict Y from X



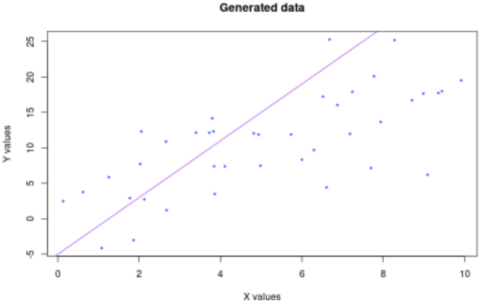
-
- Simplest idea is to fit a line

Given some data

- So we're going to fit a two parameter model $Y = \beta_0 + \beta_1 X$

Given some data

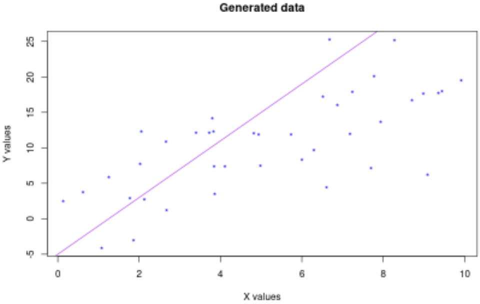
- So we're going to fit a two parameter model $Y = \beta_0 + \beta_1 X$



-

Given some data

- So we're going to fit a two parameter model $Y = \beta_0 + \beta_1 X$



-

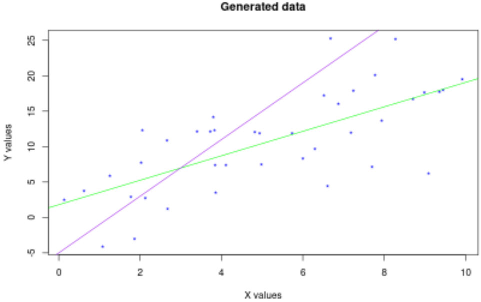
- Idea is we'll predict Y from X. As no line is exact we'll really use $Y = \beta_0 + \beta_1 X + \epsilon$. Where the ϵ are distributed normal, independent with mean zero and common variance.

Model fitting

- β_0 and β_1 can we fit with something like gradient descent. Some values are better than others

Model fitting

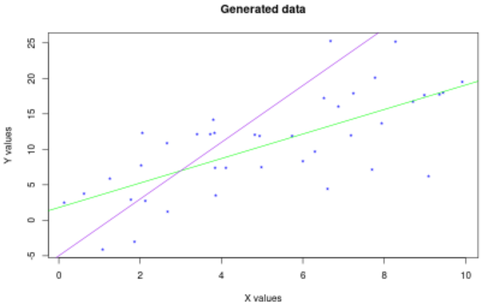
- β_0 and β_1 can we fit with something like gradient descent. Some values are better than others



-

Model fitting

- β_0 and β_1 can we fit with something like gradient descent. Some values are better than others



-

- In the case of single variable linear regression we have mathematical tools to get exact formulas. Which provide interruptibility and intuition for what's going on.

Model fitting

- β_0 and β_1 are still best guesses. I generated them with $\beta_0 = 0$ and $\beta_1 = 2$. The values that the computer worked out given the data were $\hat{\beta}_0 = 1.801564$ and $\hat{\beta}_1 = 1.727012$

Model fitting

- β_0 and β_1 are still best guesses. I generated them with $\beta_0 = 0$ and $\beta_1 = 2$. The values that the computer worked out given the data were $\hat{\beta}_0 = 1.801564$ and $\hat{\beta}_1 = 1.727012$
- Given a data set (X_i, Y_i) we want to fit β_0 and β_1 to minimize $\sum_i (Y_i - \hat{Y}_i)^2$

Model fitting

- β_0 and β_1 are still best guesses. I generated them with $\beta_0 = 0$ and $\beta_1 = 2$. The values that the computer worked out given the data were $\hat{\beta}_0 = 1.801564$ and $\hat{\beta}_1 = 1.727012$
- Given a data set (X_i, Y_i) we want to fit β_0 and β_1 to minimize $\sum_i (Y_i - \hat{Y}_i)^2$
- That is minimize $L = \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$

Model fitting

- β_0 and β_1 are still best guesses. I generated them with $\beta_0 = 0$ and $\beta_1 = 2$. The values that the computer worked out given the data were $\hat{\beta}_0 = 1.801564$ and $\hat{\beta}_1 = 1.727012$
- Given a data set (X_i, Y_i) we want to fit β_0 and β_1 to minimize $\sum_i (Y_i - \hat{Y}_i)^2$
- That is minimize $L = \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$
- $\frac{\partial L}{\partial \beta_0} = 0$

Model fitting

- β_0 and β_1 are still best guesses. I generated them with $\beta_0 = 0$ and $\beta_1 = 2$. The values that the computer worked out given the data were $\hat{\beta}_0 = 1.801564$ and $\hat{\beta}_1 = 1.727012$
- Given a data set (X_i, Y_i) we want to fit β_0 and β_1 to minimize $\sum_i (Y_i - \hat{Y}_i)^2$
- That is minimize $L = \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$
- $\frac{\partial L}{\partial \beta_0} = 0$
- $\frac{\partial L}{\partial \beta_1} = 0$

Model fitting

- $\frac{\partial L}{\partial \beta_0} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-1) = 0$

Model fitting

- $\frac{\partial L}{\partial \beta_0} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-1) = 0$
- Solves to $\overline{Y} = \beta_0 + \beta_1 \overline{X}$

Model fitting

- $\frac{\partial L}{\partial \beta_0} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-1) = 0$
- Solves to $\overline{Y} = \beta_0 + \beta_1 \overline{X}$
- $\beta_0 = \overline{Y} - \beta_1 \overline{X}$

Model fitting

- $\frac{\partial L}{\partial \beta_0} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-1) = 0$
- Solves to $\overline{Y} = \beta_0 + \beta_1 \overline{X}$
- $\beta_0 = \overline{Y} - \beta_1 \overline{X}$
- $\frac{\partial L}{\partial \beta_1} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0$

Model fitting

- $\frac{\partial L}{\partial \beta_0} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-1) = 0$
- Solves to $\overline{Y} = \beta_0 + \beta_1 \overline{X}$
- $\beta_0 = \overline{Y} - \beta_1 \overline{X}$
- $\frac{\partial L}{\partial \beta_1} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0$
- Solves to $\overline{XY} = \beta_0 \overline{X} + \beta_1 \overline{X^2}$

Model fitting

- $\frac{\partial L}{\partial \beta_0} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-1) = 0$
- Solves to $\overline{Y} = \beta_0 + \beta_1 \overline{X}$
- $\beta_0 = \overline{Y} - \beta_1 \overline{X}$
- $\frac{\partial L}{\partial \beta_1} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0$
- Solves to $\overline{XY} = \beta_0 \overline{X} + \beta_1 \overline{X^2}$
- $\overline{XY} = (\overline{Y} - \beta_1 \overline{X})\overline{X} + \beta_1 \overline{X^2}$

Model fitting

- $\frac{\partial L}{\partial \beta_0} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-1) = 0$
- Solves to $\overline{Y} = \beta_0 + \beta_1 \overline{X}$
- $\beta_0 = \overline{Y} - \beta_1 \overline{X}$
- $\frac{\partial L}{\partial \beta_1} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0$
- Solves to $\overline{XY} = \beta_0 \overline{X} + \beta_1 \overline{X^2}$
- $\overline{XY} = (\overline{Y} - \beta_1 \overline{X})\overline{X} + \beta_1 \overline{X^2}$
- $\beta_1 = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \overline{X}^2}$

Model fitting

- $\frac{\partial L}{\partial \beta_0} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-1) = 0$
- Solves to $\overline{Y} = \beta_0 + \beta_1 \overline{X}$
- $\beta_0 = \overline{Y} - \beta_1 \overline{X}$
- $\frac{\partial L}{\partial \beta_1} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0$
- Solves to $\overline{XY} = \beta_0 \overline{X} + \beta_1 \overline{X^2}$
- $\overline{XY} = (\overline{Y} - \beta_1 \overline{X})\overline{X} + \beta_1 \overline{X^2}$
- $\beta_1 = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \overline{X}^2}$
- Plugging in for $\beta_0 = \overline{Y} - \frac{cov(X,Y)}{V(X)}\overline{X}$

Model fitting

- $\frac{\partial L}{\partial \beta_0} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-1) = 0$
- Solves to $\overline{Y} = \beta_0 + \beta_1 \overline{X}$
- $\beta_0 = \overline{Y} - \beta_1 \overline{X}$
- $\frac{\partial L}{\partial \beta_1} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0$
- Solves to $\overline{XY} = \beta_0 \overline{X} + \beta_1 \overline{X^2}$
- $\overline{XY} = (\overline{Y} - \beta_1 \overline{X})\overline{X} + \beta_1 \overline{X^2}$
- $\beta_1 = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \overline{X}^2}$
- Plugging in for $\beta_0 = \overline{Y} - \frac{cov(X,Y)}{V(X)} \overline{X}$
- In some sense β_1 is is the fraction of the variance of X explained by Y

Model fitting - distributions

- $\frac{\partial L}{\partial \beta_0} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-1) = 0$

Model fitting - distributions

- $\frac{\partial L}{\partial \beta_0} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-1) = 0$
- Solves to $\overline{Y} = \beta_0 + \beta_1 \overline{X}$

Model fitting - distributions

- $\frac{\partial L}{\partial \beta_0} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-1) = 0$
- Solves to $\overline{Y} = \beta_0 + \beta_1 \overline{X}$
- $\beta_0 = \overline{Y} - \beta_1 \overline{X}$

Model fitting - distributions

- $\frac{\partial L}{\partial \beta_0} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-1) = 0$
- Solves to $\overline{Y} = \beta_0 + \beta_1 \overline{X}$
- $\beta_0 = \overline{Y} - \beta_1 \overline{X}$
- $\frac{\partial L}{\partial \beta_1} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0$

Model fitting - distributions

- $\frac{\partial L}{\partial \beta_0} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-1) = 0$
- Solves to $\overline{Y} = \beta_0 + \beta_1 \overline{X}$
- $\beta_0 = \overline{Y} - \beta_1 \overline{X}$
- $\frac{\partial L}{\partial \beta_1} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0$
- Solves to $\overline{XY} = \beta_0 \overline{X} + \beta_1 \overline{X^2}$

Model fitting - distributions

- $\frac{\partial L}{\partial \beta_0} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-1) = 0$
- Solves to $\overline{Y} = \beta_0 + \beta_1 \overline{X}$
- $\beta_0 = \overline{Y} - \beta_1 \overline{X}$
- $\frac{\partial L}{\partial \beta_1} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0$
- Solves to $\overline{XY} = \beta_0 \overline{X} + \beta_1 \overline{X^2}$
- $\overline{XY} = (\overline{Y} - \beta_1 \overline{X})\overline{X} + \beta_1 \overline{X^2}$

Model fitting - distributions

- $\frac{\partial L}{\partial \beta_0} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-1) = 0$
- Solves to $\overline{Y} = \beta_0 + \beta_1 \overline{X}$
- $\beta_0 = \overline{Y} - \beta_1 \overline{X}$
- $\frac{\partial L}{\partial \beta_1} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0$
- Solves to $\overline{XY} = \beta_0 \overline{X} + \beta_1 \overline{X^2}$
- $\overline{XY} = (\overline{Y} - \beta_1 \overline{X})\overline{X} + \beta_1 \overline{X^2}$
- $\beta_1 = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \overline{X}^2}$

Model fitting - distributions

- $\frac{\partial L}{\partial \beta_0} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-1) = 0$
- Solves to $\overline{Y} = \beta_0 + \beta_1 \overline{X}$
- $\beta_0 = \overline{Y} - \beta_1 \overline{X}$
- $\frac{\partial L}{\partial \beta_1} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0$
- Solves to $\overline{XY} = \beta_0 \overline{X} + \beta_1 \overline{X^2}$
- $\overline{XY} = (\overline{Y} - \beta_1 \overline{X})\overline{X} + \beta_1 \overline{X^2}$
- $\beta_1 = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \overline{X}^2}$
- Plugging in for $\beta_0 = \overline{Y} - \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \overline{X}^2} \overline{X}$

Model fitting - distributions

- $\frac{\partial L}{\partial \beta_0} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-1) = 0$
- Solves to $\overline{Y} = \beta_0 + \beta_1 \overline{X}$
- $\beta_0 = \overline{Y} - \beta_1 \overline{X}$
- $\frac{\partial L}{\partial \beta_1} = \sum_i 2(Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0$
- Solves to $\overline{XY} = \beta_0 \overline{X} + \beta_1 \overline{X^2}$
- $\overline{XY} = (\overline{Y} - \beta_1 \overline{X})\overline{X} + \beta_1 \overline{X^2}$
- $\beta_1 = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \overline{X}^2}$
- Plugging in for $\beta_0 = \overline{Y} - \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \overline{X}^2} \overline{X}$
- In some sense β_1 is is the fraction of the variance of X explained by Y

Model fitting - distributions

- If we do this computation with distributions instead of data we have the same calculation and get:

Model fitting - distributions

- If we do this computation with distributions instead of data we have the same calculation and get:
- $\beta_1 = \frac{\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[X^2]-\mathbb{E}[X]^2} = \frac{cov(X,Y)}{Var(X)}$

Model fitting - distributions

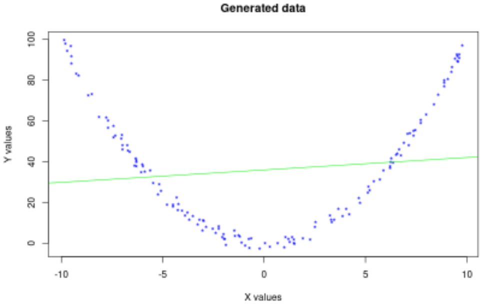
- If we do this computation with distributions instead of data we have the same calculation and get:
- $\beta_1 = \frac{\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[X^2]-\mathbb{E}[X]^2} = \frac{cov(X,Y)}{Var(X)}$
- $\beta_0 = \mathbb{E}[Y] - \frac{cov(X,Y)}{Var(X)}\mathbb{E}[X]$

Cavaets

- Sometimes straight lines are not the best predictors.

Cavaets

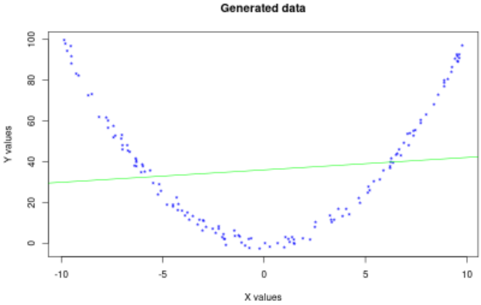
- Sometimes straight lines are not the best predictors.



-

Cavaets

- Sometimes straight lines are not the best predictors.



-

- Quadratic signal. Slope insignificant! But you can still get a good estimate from x just not with a linear model.

Inference

- OK we can estimate βs from the data. We know it's not perfect, but with enough data it's pretty good.

Inference

- OK we can estimate βs from the data. We know it's not perfect, but with enough data it's pretty good.
- One thing we might want to do is to test some hypothesis' about our coefficents.

Inference

- OK we can estimate βs from the data. We know it's not perfect, but with enough data it's pretty good.
- One thing we might want to do is to test some hypothesis' about our coefficents.
- Particularly popular is testing weather $\beta_i = 0$. We use the usual hypothesis testing notions.

Inference

- OK we can estimate βs from the data. We know it's not perfect, but with enough data it's pretty good.
- One thing we might want to do is to test some hypothesis' about our coefficents.
- Particularly popular is testing weather $\beta_i = 0$. We use the usual hypothesis testing notions.
- Null hypothesis $\beta_1 = 0$.

Inference

- OK we can estimate βs from the data. We know it's not perfect, but with enough data it's pretty good.
- One thing we might want to do is to test some hypothesis' about our coefficents.
- Particularly popular is testing weather $\beta_i = 0$. We use the usual hypothesis testing notions.
- Null hypothesis $\beta_1 = 0$.
- Alternate hypothesis $\beta_1 \neq 0$.

Inference

- OK we can estimate βs from the data. We know it's not perfect, but with enough data it's pretty good.
- One thing we might want to do is to test some hypothesis' about our coefficents.
- Particularly popular is testing weather $\beta_i = 0$. We use the usual hypothesis testing notions.
- Null hypothesis $\beta_1 = 0$.
- Alternate hypothesis $\beta_1 \neq 0$.
- To do this we usually assume that the ϵs are i.i.d. normals with mean 0 and constant variance.

Inference

- OK we can estimate βs from the data. We know it's not perfect, but with enough data it's pretty good.
- One thing we might want to do is to test some hypothesis' about our coefficents.
- Particularly popular is testing weather $\beta_i = 0$. We use the usual hypothesis testing notions.
- Null hypothesis $\beta_1 = 0$.
- Alternate hypothesis $\beta_1 \neq 0$.
- To do this we usually assume that the ϵs are i.i.d. normals with mean 0 and constant variance.
- We're not going to go into the mathematics of this one, but you should understand that some of it is hiding in the background.

Inference

- OK we can estimate βs from the data. We know it's not perfect, but with enough data it's pretty good.
- One thing we might want to do is to test some hypothesis' about our coefficents.
- Particularly popular is testing weather $\beta_i = 0$. We use the usual hypothesis testing notions.
- Null hypothesis $\beta_1 = 0$.
- Alternate hypothesis $\beta_1 \neq 0$.
- To do this we usually assume that the ϵs are i.i.d. normals with mean 0 and constant variance.
- We're not going to go into the mathematics of this one, but you should understand that some of it is hiding in the background.
- Important bit here is that we care if a variable is doing significantly better than random.