Lecture 25:
Regression
Definition

Suppose $Y$ is some r.v.

What's the "best" guess about $Y$'s value?

By "best" we mean minimizes mean squared error.
Suppose $Y$ is some r.v.

What's the "best" guess about $Y$'s value?

By "best" we mean minimizes mean squared error.

**Theorem:** the minimizer of $\sqrt{\text{E}[(Y-\alpha)^2]}$ is $\alpha = \text{E}Y$.

The minimum mean squared error is $\sqrt{\text{Var}(Y)}$. "
Prediction

Suppose $Y$ is some r.v.

What’s the “best” guess about $Y$’s value?

By “best” we mean minimizes mean squared error.

**Theorem:** the minimizer of $\mathbb{E}[(Y-\alpha)^2]$ is

$$\alpha = \mathbb{E}Y.$$

the minimum mean squared error is

$$\mathbb{E}[(Y-\alpha)^2] = \mathbb{E}[Y^2] - 2\alpha \mathbb{E}[Y] + \alpha^2$$

**Proof:**

$$\mathbb{E}[(Y-\alpha)^2] = \mathbb{E}[Y^2] - 2\alpha \mathbb{E}[Y] + \alpha^2$$

This is a parabola in $\alpha$:

$$a\alpha^2 + b\alpha + c$$

with minimal value at

$$\alpha = \frac{-b}{2a} = \frac{2 \mathbb{E}Y}{2} = \mathbb{E}Y.$$
Prediction

Suppose $X, Y$ are some r.v.
Suppose we learned $X$'s value.
What's the best guess for $Y$'s value?
Multiple Random Variables

Joint Distribution: If $X$ and $Y$ are two r.v.s over the same probability space then their joint distribution is defined as

$$\{(a, b, Pr[X=a, Y=b]) : a \in \text{range}(X), b \in \text{range}(Y)\}$$

Marginal Distributions

Marginal for $X$: $Pr[X=a] = \sum_{b \in \text{range}(Y)} Pr[X=a, Y=b]$  
Marginal for $Y$: $Pr[Y=b] = \sum_{a \in \text{range}(X)} Pr[X=a, Y=b]$
Multiple Random Variables

Joint Distribution:
\[
\{(a, b, \Pr[X=a, Y=b]) : a \in \text{range}(x), b \in \text{range}(y)\}
\]

Marginal for X: \(\Pr[X=a] = \sum_{b \in \text{range}(y)} \Pr[X=a, Y=b]\)

Marginal for Y: \(\Pr[Y=b] = \sum_{a \in \text{range}(x)} \Pr[X=a, Y=b]\)

Example:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>0.3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0.1</td>
<td>0.2</td>
<td>0.1</td>
</tr>
</tbody>
</table>

\(\Pr[X=1]\) = \\
\(\Pr[Y=3]\) = \\
\(\Pr[X=1 | Y=3]\) =
**Conditional Expectation**

**Defn:** Let $X$ and $Y$ be r.v.s over $\Omega$. The conditional expectation of $Y$ given $X$ is defined as

$$E[Y|X] = g(x)$$

where

$$g(x) = E[Y|X=x] = \sum_y y \cdot P_r[Y=y | X=x].$$
**Conditional Expectation**

**Defn:** Let $X$ and $Y$ be r.v.s over $\mathbb{R}$. The conditional expectation of $Y$ given $X$ is defined as

$$E[Y | X] = g(x)$$

where

$$g(x) = E[Y | X = x] = \sum_y y \cdot P_r(Y = y | X = x).$$

**Note:** $E[Y | X]$ is a r.v. that is a func. of $X$. For $x$: $E[Y | X = x]$ is a number.
Properties of Conditional Expectation

\[ E[Y \mid X = x] = \sum_y y \cdot \Pr[Y = y \mid X = x] \]

1. If \( X, Y \) indep. \( \Rightarrow \) \( E[Y \mid X] = E[Y] \).

2. \( E[ay + b \mid X] = a E[Y \mid X] + b \)

3. \( \forall \text{func } h(.) \) \( E[h(X)Y \mid X] = h(X) \cdot E[Y \mid X] \).

4. \( E[E[Y \mid X]] = E[Y] \)

5. \( \forall \text{func } h(.) \) \( E[h(X)E[Y \mid X]] = E[h(X)Y] \).
Properties of Conditional Expectation

\[ E[Y|X=x] = \sum_y y \cdot Pr[Y=y|X=x] \]

1. If \( X, Y \) indep. \( \Rightarrow \) \( E[Y|X] = E[Y] \).
2. \( E[aY+b|X] = aE[Y|X]+b \)
3. \( \forall \text{ func } h(\cdot) \quad E[h(X)Y|X] = h(X) \cdot E[Y|X] \)
4. \( E\left[ E[Y|X] \right] = E[Y] \)
5. \( \forall \text{ func } h(\cdot) \quad E\left[h(X)E[Y|X]\right] = E[h(X)Y] \)

**Proof:**

1. true, by definition.
2. true, by linearity of expectation.
3. For any \( x, \) \( E[h(x) \cdot Y|X=x] = h(x) \cdot E[Y|X=x] \).
4. \( E\left[ E[Y|X] \right] = \sum_x \sum_y y \cdot Pr[X=x] \cdot Pr[Y=y|X=x] \)
   \[ = \sum_y y \cdot \sum_x Pr[X=x] \cdot Pr[Y=y|X=x] \]
   \[ = \sum_y y \cdot \sum_x Pr[X=x] \cdot Pr[Y=y] \]
   \[ = E[Y] \]
   \[ Pr[Y=y] \]
5. For any $h(\cdot)$

$$\mathbb{E} \left[ h(x) \mathbb{E}[y | x] \right] = \mathbb{E} \left[ h(x) y \right]$$
5. For any $h(.)$ 
\[ \mathbb{E}_x \left[ h(x) \mathbb{E}_y[Y|x] \right] = \mathbb{E}_{x,y} [h(x) Y] \]

**Proof:** 
\[ \mathbb{E}_x \left[ h(x) \mathbb{E}_y[Y|x] \right] = \mathbb{E}_x \left[ \sum_y \mathbb{P}(Y=y|x) \cdot h(x) \cdot y \right] \]
\[ = \sum_x \mathbb{P}(X=x) \cdot h(x) \cdot \sum_y \mathbb{P}(Y=y|x) \cdot y \]
\[ = \sum_x \mathbb{P}(X=x, Y=y) \cdot h(x) \cdot y \]
\[ = \mathbb{E}_{x,y} [h(x) Y] \]
Corollary. For all \( h(.) \) \( \mathbb{E}[(y - \omega h(x)) \cdot h(x)] = 0 \)
Corollary: For all \( h(.) \), \( \mathbb{E}[(Y - \mathbb{E}[Y|x]) \cdot h(x)] = 0 \)

Proof:

\[
\mathbb{E}[(Y - \mathbb{E}[Y|x]) h(x)] = \\
= \mathbb{E}[Y \cdot h(x)] - \mathbb{E}[\mathbb{E}[Y|x] h(x)] \\
= \mathbb{E}[Y \cdot h(x)] - \mathbb{E}[Y \cdot h(x)].
\]
Theorem: Let $X,Y$ be two r.v.s over $\mathcal{X}$. The best predictor of $Y$ from $X$ (minimizes mean squared error) is 

$$g(x) = \mathbb{E}[Y|X].$$
Theorem: Let $X,Y$ be two r.v.s over $\mathcal{S}$.

The best predictor of $Y$ from $X$ (minimizes mean squared error) is
\[ g(x) = \mathbb{E}[Y \mid X]. \]

Proof: Let $h(X)$ be any function of $X$.

\[
\mathbb{E}[(Y-h(X))^2] = \mathbb{E}[(Y-g(X) + g(X)-h(X))^2]
\]

\[
= \mathbb{E}[(Y-g(X))^2] + 2 \mathbb{E}[(Y-g(X)) \cdot (g(X)-h(X))] + \mathbb{E}[(g(X)-h(X))^2]
\]

\[ \geq \mathbb{E}[(Y-g(X))^2] \]
Theorem: Let $X, Y$ be two r.v.s over $\mathcal{S}$. The best predictor of $Y$ from $X$ (minimizes mean squared error) is $g(x) = \mathbb{E}[Y|X]$.

Proof: Let $h(X)$ be any function of $X$.

$$\mathbb{E}[(Y-h(X))^2] = \mathbb{E}[(Y-g(x) + g(x)-h(x))^2]$$

$$= \mathbb{E}[(Y-g(x))^2] + 2 \mathbb{E}[(Y-g(x))(g(x)-h(x))] + \mathbb{E}[(g(x)-h(x))^2]$$

$$\geq \mathbb{E}[(Y-g(x))^2]$$
Theorem: Let \( X, Y \) be two r.v.s over \( \mathbb{R} \).

The best predictor of \( Y \) from \( X \) (minimizes mean squared error) is

\[
g(x) = \mathbb{E}[Y|X].
\]

Proof: Let \( h(X) \) be any function of \( X \).

\[
\mathbb{E} [(Y-h(X))^2] = \mathbb{E} [(Y-g(x)+g(x)-h(x))^2]
\]

\[
= \mathbb{E} [(Y-g(x))^2] + 2 \mathbb{E} [(Y-g(x))(g(x)-h(x))] + \mathbb{E} [(g(x)-h(x))^2]
\]

\[
\geq \mathbb{E} [(Y-g(x))^2]
\]

Alternative Proof: For any \( x \in \text{range } (X) \), given \( X=x \)

the best predictor of \( Y \) is \( \mathbb{E}[Y|X=x] \) from earlier.
Linear Regression

So far, we've seen:

• If we want to guess $Y$ without knowing anything else, the best guess is $\mathbb{E}Y$.

• If we make some observation $X$ related to $Y$, the best guess is $g(x) = \mathbb{E}[Y|X]$. 
Linear Regression

So far, we've seen:

- If we want to guess \( Y \) without knowing anything else, the best guess is \( \mathbb{E}[Y] \).
- If we make some observation \( X \) related to \( Y \), the best guess is \( g(x) = \mathbb{E}[Y | X] \).

The latter is optimal but can be complicated.

What if we want a simpler function of \( X \) explaining \( Y \). For example: a linear function.
Motivation: Statistics

In real-life applications, we don't necessarily know the joint dist. of $x, y$.

We can get estimates for $E[X], E[Y], \text{etc.}$ from observations.
Motivation: Statistics

In real-life applications, we don't necessarily know the joint dist. of $x, y$.
We can get estimates for $E[X], E[Y]$, etc. from observations.

To estimate $E[Y | X=x]$ we need samples such that $X=x$, but typically we'll have few or no such examples.
Motivation: Statistics

In real-life applications, we don't necessarily know the joint dist. of $x, y$.

We can get estimates for $EX$, $EY$, etc. from observations.

To estimate $\pi_1, \pi_2$, ...

Samples sum --, but typically we'll have few or no such examples.

For a simpler model, $\alpha x + \beta$, we can use all the samples to get a good estimate of the two parameters: $\alpha, \beta$. 
Theorem: The best linear predictor of $Y$ as a function of $X$ is 
\[ \hat{Y} = E[Y] + \frac{\text{cov}(X,Y)}{\text{var}(X)} \cdot (X - E[X]). \]
Theorem: The best linear predictor of $Y$ as a function of $X$ is

$$\hat{Y} = \mathbb{E}[Y] + \frac{\text{cov}(X,Y)}{\text{var}(X)} \cdot (X - \mathbb{E}X).$$

Proof: We'll first consider predicting $\hat{Y} = Y - \mathbb{E}Y$
from $\hat{X} = X - \mathbb{E}X$. 
**Theorem:** The best linear predictor of \( Y \) as a function of \( X \) is
\[
E[Y] + \frac{\text{cov}(x,y)}{\text{var}(x)} \cdot (X - E[X]).
\]

**Proof:** We'll first consider predicting \( \hat{Y} = Y - EY \)
from \( \hat{X} = X - E[X] \).

\[
\min_{\alpha, \beta} E[(\hat{Y} - (\alpha \hat{X} + \beta))^2] = \\
= \min_{\alpha, \beta} E[\hat{Y}^2] - 2E[\hat{Y}(\alpha \hat{X} + \beta)] + E[(\alpha \hat{X} + \beta)^2].
\]
**Theorem:** The best linear predictor of $Y$ as a function of $X$ is

$$\hat{Y} = \mathbb{E}[Y] + \frac{\text{cov}(X,Y)}{\text{var}(X)} \cdot (X - \mathbb{E}X).$$

**Proof:** We'll first consider predicting $\hat{Y} = Y - \mathbb{E}Y$ from $\hat{X} = X - \mathbb{E}X$.

$$\min_{\alpha, \beta} \mathbb{E}[(\hat{Y} - (\alpha \hat{X} + \beta))^2] =$$

$$\min_{\alpha, \beta} \mathbb{E}[\hat{Y}^2] - 2\mathbb{E}[\hat{Y}(\alpha \hat{X} + \beta)] + \mathbb{E}[(\alpha \hat{X} + \beta)^2]$$

$$= \min_{\alpha, \beta} \mathbb{E}[\hat{Y}^2] - 2\alpha \mathbb{E}[\hat{X}\hat{Y}] + \alpha^2 \mathbb{E}[\hat{X}^2] + \beta^2$$
Theorem: The best linear predictor of $Y$ as a function of $X$ is
\[ E[Y] + \frac{\text{cov}(X,Y)}{\text{var}(X)} \cdot (X - E[X]). \]

Proof: We'll first consider predicting $\hat{Y} = Y - E[Y]
from $ \(X = x - E[X]. \)

\[
\min_{\alpha, \beta} E \left[ (\hat{y} - (\alpha x + \beta))^2 \right] = \\
= \min_{\alpha, \beta} E[\hat{y}^2] - 2E[\hat{y}(\alpha x + \beta)] + E[(\alpha x + \beta)^2] \\
= \min_{\alpha, \beta} E[\hat{y}^2] - 2\alpha E[\hat{y}x] + \beta^2 \\
= \min_{\alpha, \beta} E[\hat{y}^2] - 2\alpha E[\hat{y}x] + \alpha^2 E[x^2] + \beta^2 \\

\text{Solution:} \quad \beta = 0, \quad \alpha = \frac{E[\hat{y}x]}{E[x^2]} = \frac{\text{cov}(x,y)}{\text{var}(x)}
Theorem: The best linear predictor of $Y$ as a function of $X$ is
\[ \mathbb{E}[Y] + \frac{\text{cov}(X,Y)}{\text{var}(X)} \cdot (X - \mathbb{E}X). \]

Proof: We'll first consider predicting $\hat{Y} = Y - \mathbb{E}Y$
from $\mathbf{x} = X - \mathbb{E}X$.

\[
\begin{align*}
\min_{\alpha, \beta} & \quad \mathbb{E}[(Y - (\alpha X + \beta))^2] \\
= & \quad \min_{\alpha, \beta} \mathbb{E}[Y^2] - 2\alpha \mathbb{E}[Y X] + \alpha^2 \mathbb{E}[X^2] + \beta^2 \\
= & \quad \min_{\alpha, \beta} \mathbb{E}[Y^2] - 2\alpha \mathbb{E}[X^2] + \alpha^2 \mathbb{E}[X^2] + \beta^2 \\
\text{solution:} & \quad \beta = 0, \quad \alpha = \frac{\mathbb{E}[X,Y]}{\mathbb{E}[X^2]} = \frac{\text{cov}(X,Y)}{\text{var}(X)} \\
\Rightarrow & \quad \text{Best linear predictor for } Y - \mathbb{E}Y \text{ is } \alpha \cdot (X - \mathbb{E}X) \\
\Rightarrow & \quad \text{" " " " " " } Y \text{ is } \mathbb{E}[Y] + \alpha \cdot (X - \mathbb{E}X) \]
Theorem: The best linear predictor of $Y$ as a function of $X$ is $l(x) = \mathbb{E}[Y] + \frac{\text{cov}(X,Y)}{\text{var}(X)} \cdot (x - \mathbb{E}X)$.

Corollary: The minimum $\mathbb{E}[(Y - l(x))^2] = \text{var}(Y) \cdot (1 - \text{corr}(X,Y))$. 
**Theorem:** The best linear predictor of $Y$ as a function of $X$ is

$$
EY + (X - EX) \cdot \frac{Cov(X, Y)}{Var(X)}.
$$

**Corollary:** The minimum $E[(Y - l(x))^2] = Var(Y) \cdot (1 - Corr(X,Y)^2)$.

**Proof:** By the theorem $l(x) = EY + (X - EX) \frac{Cov(X, Y)}{Var(X)}$

$$
E[(Y - l(x))^2] = E[(Y - EY - (X - EX) \frac{Cov(X, Y)}{Var(X)})^2]
$$

$$
= E[(Y - EY)^2] - 2 \frac{Cov(X, Y)}{Var(X)} E[(Y - EY)(X - EX)]
$$

$$
+ \left( \frac{Cov(X, Y)}{Var(X)} \right)^2 \cdot E[(X - EX)^2]
$$

$$
= Var(Y) - \frac{Cov(X, Y)^2}{Var(X)}
$$

$$
= Var(Y) - Var(Y) \cdot \frac{Cov(X, Y)^2}{Var(X) \cdot Var(Y)} = Var(Y) \cdot (1 - Corr(X,Y)^2).
$$
\[ l(x) = \mathbb{E}Y + \frac{\text{Cov}(x, y)}{\text{Var}(x)} (x - \mathbb{E}x) \]

The line goes through \((\mathbb{E}x, \mathbb{E}y)\).

The slope of the line is \(\frac{\text{Cov}(x, y)}{\text{Var}(x)}\).
Theorem: The best linear predictor of $Y$ as a function of $X$ is
$$ E(Y) + (x - E(x)) \cdot \frac{\text{cov}(x, y)}{\text{Var}(x)}. $$

Corollary: The minimum $E[(Y - f(x))^2] = \text{Var}(Y) \cdot (1 - \text{corr}(x,y)^2).$

This is what we mean by

"$X$ explains 80\% of the variance of $Y$"

$$ \text{corr}(x, y) = 0.8. $$
Minimizing Given Data

Suppose you get samples $(x_1, y_1), \ldots, (x_n, y_n)$ from the joint distribution, and you want to minimize

$$\sum_{i=1}^{n} (y_i - \alpha x_i + \beta)^2$$
Minimizing Given Data

Suppose you get samples \((x_1, y_1), \ldots, (x_n, y_n)\) from the joint distribution, and you want to minimize

\[
\sum_{i=1}^{n} (y_i - \alpha x_i + \beta)^2
\]

You'll get:

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i
\]

\[
\alpha = \frac{\frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x} \bar{y}}{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}
\]

\[
\beta = \bar{y} - \alpha \bar{x}
\]

(estimate to \(\text{Cov}(x, y)\))

(estimate to \(\text{Cov}(x, y) \div \text{Var}(x)\))