Advanced Topics in Regression

W. Evan Johnson, Ph.D.
Professor, Division of Infectious Disease
Director, Center for Data Science
Co-Direcor, Center for Biomedical Informatics and Health AI
Rutgers University – New Jersey Medical School

2025-07-30

Introduction to Regression

In data science applications, it is very common to be interested in the relationship between two or more variables. For example, we we might want to use a data-driven approach that examines the relationship between baseball player statistics and success to guide the building of a baseball team with a limited budget. Before delving into this more complex example, we introduce necessary concepts needed to understand regression.

Introduction to Regression

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome' or 'response' variable, or a 'label' in machine learning parlance) and one or more independent variables (often called 'predictors', 'covariates', 'explanatory variables' or 'features').

Introduction to Regression

The most common form of regression analysis is **linear regression**, in which one finds the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion.

Introduction to Regression

In your math classes, you used the following model for a line, where m represents the slope and b represents the y-intercept:

$$y = mx + b.$$

In statistics, we use a slightly different formulation and notation:

$$y_i = \beta_0 + x_i\beta_1,$$

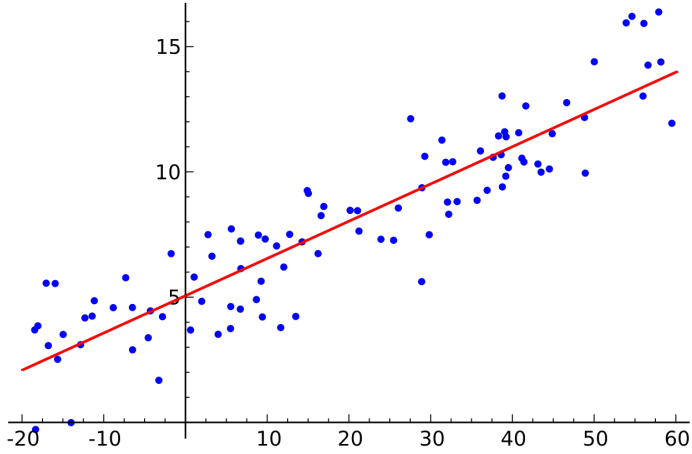
which describes a line with slope β_1 and y-intercept β_0, and the y_i and x_i represent the observed points from your dataset.

Introduction to Regression

One other important difference between your math class and statistics is that in math you are usually fitting lines that fit your points *exactly*, whereas in statistics your points usually *do not* fall on directly on your line.

In statistics, we are attempting to find the **best fit** line that uses your data to estimate the unobserved relationship (slope and intercept) between the independent and dependent variables. As defined above, this relationship between parameters β_0 and β_1 and the data points is called a **linear regression model**.

Introduction to Regression



Introduction to Regression

Formal Definition: suppose we observe n data pairs and call them

$$(x_i, y_i), i = 1, ..., n.$$

We can describe the underlying relationship between y_i and x_i involving this error term ϵ_i by

$$y_i = \beta_0 + x_i\beta_1 + \epsilon_i.$$

We call the deviations of the data from the line the **errors**. It is also important to note that the true underlying parameters β_0 and β_1 are not observed, but must be *estimated* using the data.

Introduction to Regression

The goal is to find estimated values $\hat{\beta}_0$ and $\hat{\beta}_1$ for the parameters β_0 and β_1 which would provide the "best fit" in some sense for the data points.

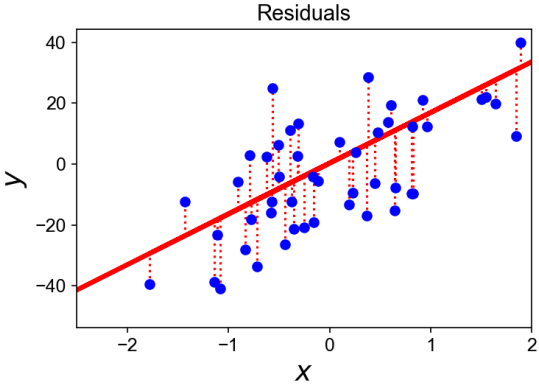
Here, the "best fit" will be understood as in the **least-squares** approach: a line that minimizes the sum of squared residuals

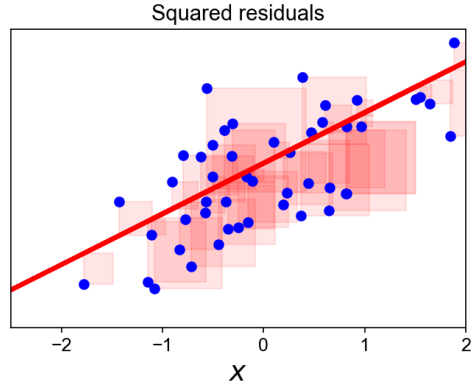
$$\min_{\beta_0,\beta_1} Q(\beta_0,\beta_1) = \inf_{\beta_0,\beta_1} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - x_i\beta_1)^2 \right\} = \inf_{\beta_0,\beta_1} \sum_{i=1}^{n} \epsilon_i^2.$$

Introduction to Regression

The least squares estimator (LSE) is the most commonly used method due to its favorable mathematical properties (not shown here). For example, it can be mathematically shown that the LSE is the *minimum variance unbiased linear estimator*–meaning that no unbiased estimator can have a smaller (sampling distrbution) variance for $\hat{\beta}_0$ and $\hat{\beta}_1$.

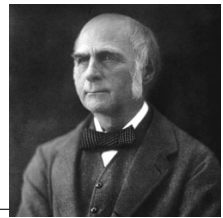
Introduction to regression





Case study: is height hereditary?

History of regression: Francis Galton[1] studied the variation and heredity of human traits, including height data from families to try to understand heredity. While doing this, he developed the concepts of correlation and regression, as well as a connection to pairs of data that follow a normal distribution.



[1] https://en.wikipedia.org/wiki/Francis_Galton

Case study: is height hereditary?

A very specific question Galton tried to answer was: how well can we predict a child's height based on the parents' height? The technique he developed to answer this question was called **regression**!

Historical note: Galton made important contributions to statistics and genetics, but he was also a proponent of eugenics, a scientifically flawed philosophical movement with horrific historical consequences: https://pged.org/history-eugenics-and-genetics/.

Case study: is height hereditary?

Galton's height data is in the `HistData` package. We will create a dataset with the heights of fathers and a randomly selected son:

```
library(tidyverse)
library(HistData)
data("GaltonFamilies")

set.seed(1983)
galton_heights <- GaltonFamilies %>%
  filter(gender == "male") %>%
  group_by(family) %>%
  sample_n(1) %>%
  ungroup() %>%
  select(father, childHeight) %>%
  rename(son = childHeight)
```

Case study: is height hereditary?

Suppose we were asked to summarize the father and son data. Since both distributions are well approximated by the normal distribution, we could use the two averages and two standard deviations as summaries:

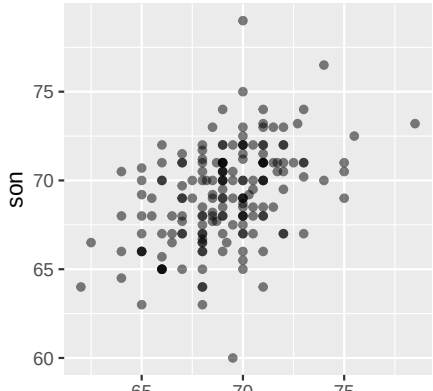
```
galton_heights %>%
  summarize(mean(father), sd(father), mean(son), sd(son))
```

```
## # A tibble: 1 x 4
##   `mean(father)` `sd(father)` `mean(son)` `sd(son)`
##            <dbl>        <dbl>       <dbl>     <dbl>
## 1           69.1         2.55        69.2      2.71
```

However, this summary fails to describe an important characteristic of the data: the trend that the taller the father, the taller the son.

Case study: is height hereditary?

```
galton_heights %>% ggplot(aes(father, son)) +
  geom_point(alpha = 0.5)
```



Case study: is height hereditary?

The correlation between father and son's heights is:

```
galton_heights %>%
    summarize(r = cor(father, son)) %>%
    pull(r)
```

```
## [1] 0.4334102
```

Case study: is height hereditary?

In R, we can obtain the least squares estimates using the `lm` function. To fit the model:

$$y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$$

with y_i the son's height and x_i the father's height, we can use this code to obtain the least squares estimates:

```
fit <- lm(son ~ father, data = galton_heights)
fit$coef
```

```
## (Intercept)      father
##   37.287605    0.461392
```

Case study: is height hereditary?

The object `fit` includes more information about the fit. We can use the function `summary` to extract more of this information:

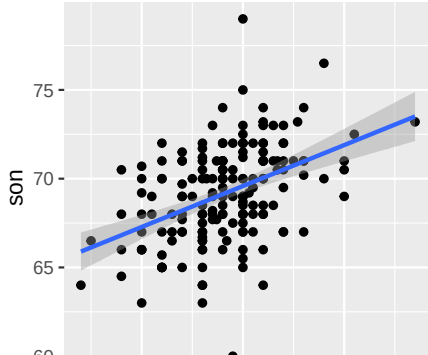
```
summary(fit)
```

```
##
## Call:
## lm(formula = son ~ father, data = galton_heights)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3543 -1.5657 -0.0078  1.7263  9.4150
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.28761    4.98618   7.478 3.37e-12 ***
## father       0.46139    0.07211   6.398 1.36e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Case study: is height hereditary?

We can use ggplot2 layers to plot $\hat{Y}$ with its confidence intervals:

```
galton_heights %>% ggplot(aes(father, son)) +
  geom_point() + geom_smooth(method = "lm")
```



Case study: is height hereditary?

The R function `predict` takes an `lm` object as input and returns the prediction. If requested, the standard errors and other information from which we can construct confidence intervals is provided:

```
fit <- galton_heights %>% lm(son ~ father, data = .)
y_hat <- fit %>% predict(se.fit = TRUE)
names(y_hat)
```

```
## [1] "fit"            "se.fit"         "df"             "residual.scale"
```

Introduction to Regression

Now, we can extend our regression to include multiple (p) predictors:

$$y_i = x_{i1}\beta_1 + \ldots + x_{ip}\beta_p + \epsilon_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i.$$

Note that this model can still include a y-intercept if one of your predictors is a constant for all x, y data pairs, e.g., $x_{ij} = 1$ for all i for a particular j. In this case, $x_{ij}\beta_j = (1)\beta_j = \beta_0$.

Introduction to Regression

Fitting the entire dataset using **multiple regression**:

```
fit2 <- GaltonFamilies %>%
  lm(childHeight ~ mother + father + gender + children, data = .)
fit2 %>% summary()
```

```
##
## Call:
## lm(formula = childHeight ~ mother + father + gender + children,
##     data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4759 -1.4743  0.0906  1.4789  9.1734
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.43103    2.77407   6.284 5.08e-10 ***
## mother       0.31619    0.03098  10.207  < 2e-16 ***
## father       0.38521    0.02898  13.292  < 2e-16 ***
## gendermale   5.19852    0.14197  36.617  < 2e-16 ***
## children    -0.04573    0.02631  -1.738   0.0825 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.162 on 929 degrees of freedom
```

Introduction to Regression

Interpreting this model:

- The intercept, $\beta_0 = 17.4$ inches (p-value=`5.08e-10`).
- Every inch increase in mother's height leads to a $\beta_{mother} = 0.32$ (p-value<`2e-16`) in `childHeight`.
- Every inch increase in father's height leads to a $\beta_{father} = 0.39$ (p-value<`2e-16`) in `childHeight`.
- `Gender=female` was selected as the reference level, and `gender=male` results in an increase in $\beta_{male} = 5.2$ inches (p-value<`2e-16`). *R always chooses the first level of a factor (alphabetically) as the reference.*
- The number of children $\beta_{children}$ is not a significant predictor of child height (p-value=0.0825).
- Multiple $R^2 = 0.637$ and adjusted $R^2 = 0.635$, thus `father`, `mother`, and `gender` are moderately strong predictors of `childHeight`.

Introduction to Regression

In addition, we can use matrices to represent our model. Assume $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ with a matrix of predictors $\mathbf{X}$ and coefficient vector β ($n \times 1$ vector). Then we can define our regression equation as:

$$\mathbf{y} = X\beta + \epsilon,$$

where $\epsilon \sim N(\mathbf{0}, \sigma^2 I_n)$ and I_n is an identity matrix with dimension n.

Introduction to Regression

Minimizing the least squared error is can be represented by

$$\hat{\beta} = \inf_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\} = \inf_{\beta} \{(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)\}$$

Diversion: Maximum Likelihood for a Normal mean

Assume $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, where x_is are independent from a Normal(μ, σ^2) distribution, where $\mathbf{x}$ is observed and σ^2 is known. Then

$$L(\mu|\mathbf{x}) = \left\{\frac{1}{2\pi\sigma^2}\right\}^{N/2} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2\right\}$$

Diversion: Maximum Likelihood for a Normal mean

Assume $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, where x_is are independent from a Normal(μ, σ^2) distribution, where $\mathbf{x}$ is observed and σ^2 is known. Then

$$L(\mu|\mathbf{x}) = \left\{\frac{1}{2\pi\sigma^2}\right\}^{N/2} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2\right\}$$

Notice we can maximize $L(\mu|\mathbf{x})$ by minimizing $\sum_{i=1}^{N}(x_i - \mu)^2$ for μ.

Diversion: Maximum Likelihood for a Normal mean

Note the following:

$$\sum_{i=1}^{N}(x_i - \mu)^2 = \sum_{i=1}^{N}(x_i^2 - 2\mu x_i + \mu^2)$$

Diversion: Maximum Likelihood for a Normal mean

Note the following:

$$\begin{aligned}\sum_{i=1}^{N}(x_i-\mu)^2 &= \sum_{i=1}^{N}(x_i^2-2\mu x_i+\mu^2)\\ &= \sum_{i=1}^{N}x_i^2-2\mu\sum_{i=1}^{N}x_i+N\mu^2\end{aligned}$$

Diversion: Maximum Likelihood for a Normal mean

Note the following:

$$\begin{aligned}\sum_{i=1}^{N}(x_i-\mu)^2 &= \sum_{i=1}^{N}(x_i^2-2\mu x_i+\mu^2)\\ &= \sum_{i=1}^{N}x_i^2-2\mu\sum_{i=1}^{N}x_i+N\mu^2\end{aligned}$$

Diversion: Maximum Likelihood for a Normal mean

therefore:

$$\frac{\partial}{\partial \mu} = -2 \sum x_i + 2N\hat{\mu} \stackrel{set}{=} 0$$

Diversion: Maximum Likelihood for a Normal mean

therefore:

$$\frac{\partial}{\partial \mu} = -2 \sum x_i + 2N\hat{\mu} \stackrel{set}{=} 0$$

$$\Rightarrow N\hat{\mu} = \sum x_i$$

Diversion: Maximum Likelihood for a Normal mean

therefore:

$$\begin{aligned}\frac{\partial}{\partial \mu} &= -2\sum x_i + 2N\hat{\mu} \overset{set}{=} 0 \\ &\Rightarrow N\hat{\mu} = \sum x_i \\ &\Rightarrow \hat{\mu} = \frac{\sum x_i}{N}\end{aligned}$$

Diversion: Maximum Likelihood for Regression

Assume $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ with a matrix of predictors $\mathbf{X}$ and coefficient vector β ($n \times 1$ vector). Then we can define our regression equation as:

$$\mathbf{y} = X\beta + \epsilon,$$

where $\epsilon \sim N(\mathbf{0}, \sigma^2 I_N)$ and I_N is an identity matrix with dimension N.

Diversion: Maximum Likelihood for Regression

Now, extending the Normal mean MLE to regression, we note that

$$\sum_{i=1}^{N}(y_i - x_1\beta_1 - \ldots - x_N\beta_N)^2 = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta),$$

Diversion: Maximum Likelihood for Regression

Now, extending the Normal mean MLE to regression, we note that

$$\sum_{i=1}^{N}(y_i - x_1\beta_1 - \ldots - x_N\beta_N)^2 = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta),$$

so

$$L(\beta|\mathbf{X}, \mathbf{y}) = \left\{\frac{1}{2\pi\sigma^2}\right\}^{N/2} exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)\right\}.$$

Thus maximizing the Likelihood is equivalent to minimizing $(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$, or in other words, minimizing the sum of the squared error!

Diversion: Maximum Likelihood for Regression

Note

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) = \mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta,$$

Diversion: Maximum Likelihood for Regression

Note

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) = \mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta,$$

and therefore

$$\frac{\partial}{\partial \beta} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\beta} \stackrel{set}{=} 0$$

Diversion: Maximum Likelihood for Regression

Note

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) = \mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta,$$

and therefore

$$\begin{aligned} \frac{\partial}{\partial \beta} &= -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\beta} \stackrel{set}{=} 0 \\ &\Rightarrow \mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y} \end{aligned}$$

Diversion: Maximum Likelihood for Regression

Note

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) = \mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta,$$

and therefore

$$\begin{aligned}\frac{\partial}{\partial\beta} &= -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\beta} \stackrel{set}{=} 0 \\ &\Rightarrow \mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y} \\ &\Rightarrow (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\end{aligned}$$

Diversion: Maximum Likelihood for Regression

Note

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) = \mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta,$$

and therefore

$$\begin{aligned}
\frac{\partial}{\partial \beta} &= -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\beta} \stackrel{set}{=} 0 \\
&\Rightarrow \mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y} \\
&\Rightarrow (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&\Rightarrow \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}
\end{aligned}$$

Diversion: Maximum Likelihood for Regression

Now note:

$$E[\hat{\beta}] = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}]$$

Diversion: Maximum Likelihood for Regression

Now note:

$$
\begin{aligned}
E[\hat{\beta}] &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{y}]
\end{aligned}
$$

Diversion: Maximum Likelihood for Regression

Now note:

$$
\begin{aligned}
E[\hat{\beta}] &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{y}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta
\end{aligned}
$$

Diversion: Maximum Likelihood for Regression

Now note:

$$
\begin{aligned}
E[\hat{\beta}] &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{y}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta \\
&= \beta
\end{aligned}
$$

Diversion: Maximum Likelihood for Regression

And, remembering that

$$Var[\mathbf{Ay}] = \mathbf{A}\{Var[\mathbf{y}]\}\mathbf{A}',$$

so therefore,

$$Var[\hat{\beta}] = Var[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}]$$

Diversion: Maximum Likelihood for Regression

And, remembering that

$$Var[\mathbf{Ay}] = \mathbf{A}\{Var[\mathbf{y}]\}\mathbf{A}',$$

so therefore,

$$\begin{aligned} Var[\hat{\beta}] &= Var[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\{Var[\mathbf{y}]\}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

Diversion: Maximum Likelihood for Regression

And, remembering that

$$Var[\mathbf{Ay}] = \mathbf{A}\{Var[\mathbf{y}]\}\mathbf{A}',$$

so therefore,

$$\begin{aligned} Var[\hat{\beta}] &= Var[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\{Var[\mathbf{y}]\}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\{\sigma^2 I_N\}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

Diversion: Maximum Likelihood for Regression

And, remembering that

$$Var[\mathbf{Ay}] = \mathbf{A}\{Var[\mathbf{y}]\}\mathbf{A}',$$

so therefore,

$$\begin{aligned} Var[\hat{\beta}] &= Var[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\{Var[\mathbf{y}]\}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\{\sigma^2 I_N\}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

Diversion: Maximum Likelihood for Regression

And, remembering that

$$Var[\mathbf{Ay}] = \mathbf{A}\{Var[\mathbf{y}]\}\mathbf{A}',$$

so therefore,

$$\begin{aligned} Var[\hat{\beta}] &= Var[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\{Var[\mathbf{y}]\}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\{\sigma^2 I_N\}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

Diversion: Maximum Likelihood for Regression

So we can conduct a hypothesis test β, where $H_0 : \beta_j = 0$ versus $H_a : \beta_j \neq 0$ using the statistic:

$$t_{\beta_j} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_{jj}}},$$

where v_{jj} is the jth diagonal element of $\mathbf{V} = (\mathbf{X}'\mathbf{X})^{-1}$. Under H_0, t_{β_j} will follow a t distribution with $N - p - 1$ degrees of freedom.

Case study: is height hereditary?

Thus using our estimator for β:

$$\widehat{\beta} = (X'X)^{-1}X'\mathbf{y}$$

```
X <- cbind(1,galton_heights$father)
y <- galton_heights$son
beta_hat <- solve(t(X)%*%X)%*%t(X)%*%y
beta_hat
```

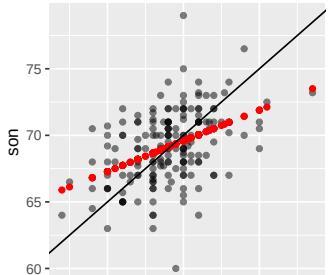
```
##             [,1]
## [1,] 37.287605
## [2,]  0.461392
```

Case study: is height hereditary?

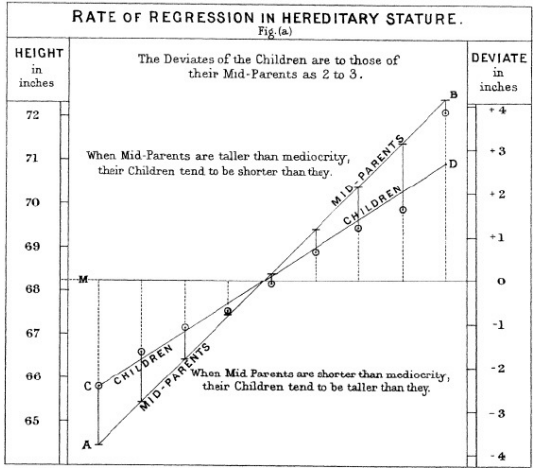
Predicted values can be obtained by: $\hat{y} = X\hat{\beta}$, In R:

```
y_hat <- X %*% beta_hat

galton_heights %>% ggplot(aes(father, son)) +
  geom_point(alpha = 0.5) +
  geom_point(aes(y = y_hat), col = "red") +
  geom_abline(slope = 1, intercept = 0)
```



Case study: is height hereditary?



Case study: is height hereditary?

For hypothesis testing we can optain a standard error:

$$SE(\hat{\beta}_i) = \hat{\sigma}\sqrt{v_{ii}},$$

where v_{ii} is the ith diagonal element of $(X'X)^{-1}$, and

$$\hat{\sigma}^2 = \frac{1}{N-p-1}(\mathbf{y} - X\widehat{\beta})'(\mathbf{y} - X\widehat{\beta})$$

```
N <- length(y)
p <- length(beta_hat)
sigma2_hat <- 1 / (N - p - 1) *
  t(y - X %*% beta_hat) %*% (y - X %*% beta_hat)
V <- solve(t(X) %*% X)
Z2 <- beta_hat[2] / sqrt(sigma2_hat * V[2, 2])
Z2
```

The Regression Fallacy

Wikipedia defines the *sophomore slump* as:

> *A sophomore slump or sophomore jinx or sophomore jitters refers to an instance in which a second, or sophomore, effort fails to live up to the standards of the first effort. It is commonly used to refer to the apathy of students (second year of high school, college or university), the performance of athletes (second season of play), singers/bands (second album), television shows (second seasons) and movies (sequels/prequels).*

The Regression Fallacy

In Major League Baseball, the rookie of the year (ROY) award is given to the first-year player who is judged to have performed the best. The *sophmore slump* phrase is used to describe the observation that ROY award winners don't do as well during their second year. For example, this Fox Sports article[2] asks "Will MLB's tremendous rookie class of 2015 suffer a sophomore slump?".

[2] http://www.foxsports.com/mlb/story/kris-bryant-carlos-correa-rookies-of-year-award-matt-duffy-francisco-lindor-kang-sano-120715

The Regression Fallacy

Does the data confirm the existence of a sophomore slump? Let's take a look. Examining the data for batting average, we see that this observation holds true for the top performing ROYs:

nameFirst	nameLast	rookie_year	rookie	sophomore
Willie	McCovey	1959	0.3541667	0.2384615
Ichiro	Suzuki	2001	0.3497110	0.3214838
Al	Bumbry	1973	0.3370787	0.2333333
Fred	Lynn	1975	0.3314394	0.3136095
Albert	Pujols	2001	0.3288136	0.3135593

The Regression Fallacy

In fact, the proportion of players that have a lower batting average their sophomore year is 0.7090909.

o is it “jitters” or “jinx”? To answer this question, let’s turn our attention to all players that played the 2013 and 2014 seasons and batted more than 130 times (minimum to win Rookie of the Year).

The Regression Fallacy

The same pattern arises when we look at the top performers: batting averages go down for most of the top performers.

nameFirst	nameLast	2013	2014
Miguel	Cabrera	0.3477477	0.3126023
Hanley	Ramirez	0.3453947	0.2828508
Michael	Cuddyer	0.3312883	0.3315789
Scooter	Gennett	0.3239437	0.2886364
Joe	Mauer	0.3235955	0.2769231

But these are not rookies!

The Regression Fallacy

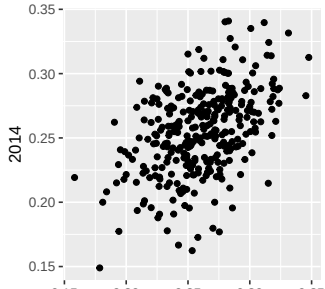
Also, look at what happens to the worst performers of 2013:

nameFirst	nameLast	2013	2014
Danny	Espinosa	0.1582278	0.2192192
Dan	Uggla	0.1785714	0.1489362
Jeff	Mathis	0.1810345	0.2000000
B. J.	Upton	0.1841432	0.2080925
Adam	Rosales	0.1904762	0.2621951

Their batting averages mostly go up!

The Regression Fallacy

Is this some sort of reverse sophomore slump? It is not. There is no such thing as the sophomore slump. This is all explained with a simple statistical fact: the correlation for performance in two separate years is high, but not perfect:



The Regression Fallacy

The correlation is 0.460254 and the data look very much like a bivariate normal distribution, which means we predict a 2014 batting average Y for any given player that had a 2013 batting average X with:

$$\frac{Y - .255}{.032} = 0.46 \left(\frac{X - .261}{.023} \right)$$

The Regression Fallacy

Because the correlation is not perfect, regression tells us that, on average, expect high performers from 2013 to do a bit worse in 2014. It's not a jinx; it's just due to chance. The ROY are selected from the top values of X so it is expected that Y will **regress to the mean**.

Session Info

```
sessionInfo()
```

```
## R version 4.5.1 (2025-06-13)
## Platform: aarch64-apple-darwin20
## Running under: macOS Sequoia 15.5
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.5-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.5-arm64/Resources/lib/libRlapack.dylib;  LAPACK version 3.12.1
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/New_York
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] Lahman_12.0-0   HistData_0.9-3  lubridate_1.9.4 forcats_1.0.0
##  [5] stringr_1.5.1   dplyr_1.1.4     purrr_1.1.0     readr_2.1.5
##  [9] tidyr_1.3.1     tibble_3.3.0    ggplot2_3.5.2   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] generics_0.1.4     xml2_1.3.8         stringi_1.8.7      lattice_0.22-7
##  [5] hms_1.1.3          digest_0.6.37      magrittr_2.0.3     evaluate_1.0.4
```