Introduction to Probability and Statistics for Engineers and Scientists

Sheldon M Ross

Solutions

Anirudh Krishnan

September 13, 2021

Contents

Chapter 1

Regression

"Are you sure you've gotten rid of any multicollinearity in the inputs?"

1.1 Introduction

The problem of determining the relationship between a set of inputs $\{x_i\}$ and the resulting output Y is a frequent problem in engineering. This is further complicated by the lack of prior knowledge about the nature of this dependence.

A simplistic model of response depending on a set of inputs uses a linear combination of these inputs along with some noise e, which is assumed to have zero mean.

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + e \tag{0.1}$$

$$\mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r \tag{0.2}$$

The set of inputs $\{x_i\}$ are called *independent variables* and the response Y which is some function of the inputs is called the *dependent variable*.

The set of coefficients $\{\beta_i\}$ are called the *regression coefficients*, and are to be determined based on an observed data-set. The special case of $r = 1$ is called a *simple* regression, while $r > 1$ is the much more complicated *multiple* regression problem.

$$Y = \alpha + \beta x + e \tag{0.3}$$

The choice of a simple linear regression model is appropriate when the data appears to follow a straight line relationship subject to random error when visualized as a scatter plot.

1.2 Least squares estimators of regression parameters

In a simple regression problem, let the estimators of α, β be A, B. For a given set of inputs $\{x_i\}$ and responses $\{Y_i\}$, the estimated response is

$$\widehat{Y}_i = A + Bx_i \tag{0.4}$$

Using an expression for the squared difference between the observed and estimated responses, it is possible to find an estimator that minimizes the sum of these squares.

$$SS = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - A - Bx_i)^2 \tag{0.5}$$

The usual method of setting partial derivatives to zero in order to find the minimum, yields a system of 2 linear equations in A, B called the *normal equations*

$$\sum Y_i = A\ n + B \sum x_i \tag{0.6}$$

$$\sum x_i Y_i = A \sum x_i + B \sum x_i^2 \tag{0.7}$$

Using the shorthand notation $\overline{Y},\ \overline{x}$ for their sample means, and solving for the least squares estimators gives the estimated regression line as $y = A + Bx$, where

$$A = \overline{Y} - B\overline{x} \qquad\qquad B = \frac{\sum (x_i - \overline{x})Y_i}{\sum x_i^2 - n\overline{x}^2} \tag{0.8}$$

1.3 Distribution of the estimators

In order to determine the distribution of the estimators A, B, an additional assumption about the random errors e is used. For some value σ^2 which is a constant independent of the input,

$$e \sim \mathcal{N}(0,\ \sigma^2) \qquad\qquad Y_i \sim \mathcal{N}(\alpha + \beta x_i,\ \sigma^2) \tag{0.9}$$

From the above expression for the estimator B, it is some linear combination of normal RVs $\{Y_i\}$. Substituting B into A shows that A is also a normal RV,

$$\mathbb{E}[B] = \beta \qquad\qquad \mathrm{Var}(B) = \frac{\sigma^2}{\sum x_i^2 - n\overline{x}^2} \tag{0.10}$$

$$\mathbb{E}[A] = \alpha \qquad\qquad \mathrm{Var}(A) = \frac{\sigma^2}{\sum x_i^2 - n\overline{x}^2} \frac{\sum x_i^2}{n} \tag{0.11}$$

From the above expectation values, A, B are both unbiased estimators of α, β.

Residuals : The difference between each observed value and its estimate is called the residual. To determine the variance of the error, the sum of squares of these residuals can be rearranged into a known RV.

$$r_i = Y_i - \widehat{Y_i} = Y_i - A - Bx_i \tag{0.12}$$

$$SS_R = \sum r_i^2 = \sum (Y_i - A - Bx_i)^2 \tag{0.13}$$

Rearranging the SS_R into a chi-square distribution gives an unbiased estimator of σ^2

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2} \qquad\qquad \mathbb{E}\left[\frac{SS_R}{n-2}\right] = \sigma^2 \tag{0.14}$$

Similar to the sample mean and sample variance from a normal population being independent, the variance of the noise σ^2 is independent of the estimators A, B.

The MLE of α, β also happen to be the least squares estimators A, B.

Shorthand notation for sums of squares : For convenience, some shorthand expressions for the sums of squares are listed here.

$$S_{xY} = \sum (x_i - \overline{x})\ (Y_i - \overline{Y}) = \sum x_i Y_i - n\overline{x}\overline{Y} \tag{0.15}$$

$$S_{xx} = \sum (x_i - \overline{x})^2 = \sum x_i^2 - n\overline{x}^2 \tag{0.16}$$

$$S_{YY} = \sum (Y_i - \overline{Y})^2 = \sum Y_i^2 - n\overline{Y}^2 \tag{0.17}$$

Using the above shorthand, the estimators A, B can be described as

$$B = \frac{S_{xY}}{S_{xx}} \qquad \sim \mathcal{N}\left(\beta,\ \frac{\sigma^2}{S_{xx}}\right) \tag{0.18}$$

$$A = \overline{Y} - B\overline{x} \qquad \sim \mathcal{N}\left(\alpha,\ \frac{\sigma^2}{S_{xx}}\ \frac{\sum x_i^2}{n}\right) \tag{0.19}$$

$$SS_R = \frac{S_{xx}\ S_{YY} - S_{xY}^2}{S_{xx}} \tag{0.20}$$

The above relation for the SS_R can only be established using brute-force computations, so a theoretical justification is not outlined here.

1.4 Statistical Inferences about regression parameters

The problem of constructing hypothesis tests is straightforward given the transformation into known RVs outlined above.

Inferences on β : A hypothesis that the average response is independent of the input requires rearranging the estimator B into a t-RV and then defining the hypothesis test as

$$
\begin{aligned}
&\frac{B-\beta}{\sigma/\sqrt{S_{xx}}} \sim Z \qquad \text{and} \qquad \frac{SS_R}{\sigma^2} \sim \chi^2_{n-2} \\
&\sqrt{\frac{(n-2)S_{xx}}{SS_R}}\,(B-\beta) \sim t_{n-2} \qquad (0.21)\\
&Y = \alpha + \beta x + e
\end{aligned}
$$

Define a hypothesis test of significance level γ as

$$
\begin{aligned}
&H_0 : \beta = 0 \qquad \text{vs.} \qquad H_1 : \beta \neq 0 \\
&\text{reject } H_0 \text{ if} \qquad \sqrt{\frac{(n-2)S_{xx}}{SS_R}}\,|B| > t_{\gamma/2,n-2} \\
&\text{accept } H_0 \qquad \text{otherwise} \qquad (0.22)
\end{aligned}
$$

The corresponding $100(1-\gamma)\%$ confidence interval for β is give by

$$
\beta \in \left[B \pm \sqrt{\frac{SS_R}{(n-2)S_{xx}}}\,t_{\gamma/2,n-2}\right] \qquad (0.23)
$$

Regression to the mean : A linear simple regression with the parameter $\beta \in [0,1]$, displays the following property,

$$
\begin{aligned}
&\mathbb{E}[Y] = \alpha + \beta x \\
&\mathbb{E}[Y] < x \qquad \forall\ x > \frac{\alpha}{1-\beta} \\
&\mathbb{E}[Y] > x \qquad \text{otherwise} \qquad (0.24)
\end{aligned}
$$

This trend is very common in real-world datasets and was first observed in the comparison of height at a given age between successive generations of a family. Even though the variables are positively correlated, $\beta \in [0,1]$ causes the extreme values to regress towards the $y = x$ line.

A test of the hypothesis $\beta \geq 1$ can be rejected in such real-world datasets as confirmation of regression to the mean.

Regression fallacy : The false attribution to some outside influence on the observed phenomenon of regression to the mean, when it might be happening simply because of chance.

Inferences on α : Similar to the previous results for B, the confidence interval for α is given by,

$$\sqrt{\frac{n(n-2)S_{xx}}{SS_R \ \sum x_i^2}} \ (A - \alpha) \sim t_{n-2} \tag{0.25}$$

$$\alpha \in \left[A \pm \sqrt{\frac{SS_R \ \sum x_i^2}{n(n-2)S_{xx}}} \ t_{\gamma/2, n-2} \right] \tag{0.26}$$

Similar to the hypothesis tests on β, the tests on α are constructed as follows.

$$\begin{aligned} &H_0 : \alpha = 0 && \text{vs.} \quad H_1 : \alpha \neq 0 \\ &\text{reject } H_0 \text{ if} && \sqrt{\frac{n(n-2)S_{xx}}{SS_R \ \sum x_i^2}} \ |A| > t_{\gamma/2, n-2} \\ &\text{accept } H_0 && \text{otherwise} \end{aligned} \tag{0.27}$$

Inferences on the mean response : A point estimator for $Y(x_0)$ in a simple linear regression model is clearly $\widehat{Y} = A + Bx_0$. This is additionally an unbiased estimator since $\mathbb{E}[A] = \alpha, \ \mathbb{E}[B] = \beta$.

Since A, B are themselves normal RVs with pre-calculated mean and variance,

$$\mathbb{E}[A + Bx_0] = \alpha + \beta x_0 \tag{0.28}$$

$$\mathrm{Var}(A + Bx_0) = \sigma^2 \left[\frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}}\right] \tag{0.29}$$

$$A + Bx_0 \sim N\left(\alpha + \beta x_0,\ \sigma^2 \left[\frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}}\right]\right) \tag{0.30}$$

Using the fact that the mean response is independent of the random error, which can be rearranged to form a chi-square RV, a t-RV can be formulated and then used to define a confidence interval.

$$\alpha + \beta x_0 \in A + Bx_0 \pm t_{\gamma/2, n-2} \sqrt{\left(\frac{SS_R}{n-2}\right) \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)} \tag{0.31}$$

Note that the lack of direct knowledge about σ^2, forced the use of a t-RV instead of the naive choice of a normal-RV when defining the confidence interval.

Prediction interval of a future response : Unlike the mean response, a point estimate of the response to a new input x_0 is defined as $Y(x_0) = \alpha + \beta x_0 + e$. A point prediction of $Y(x_0)$ is simply $A + Bx_0$, since A, B are the unbiased point estimators of α, β respectively.

Given that the mean, median and mode of a normal RV are all equal, the question of choosing one over the other does not arise here. Using the fact that both the next response Y and the predicted value $A + Bx_0$ are normal RVs, gives

$$Y - A - Bx_0 \sim N\left(0,\ \sigma^2 \left[1 + \frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}}\right]\right) \tag{0.32}$$

Once again rearranging the $Y - A - Bx_0$ and the known chi-square RV SS_R/σ^2 into a t-RV gives the confidence interval for the prediction as

$$Y \in A + Bx_0 \pm t_{\gamma/2, n-2} \sqrt{\left(\frac{SS_R}{n-2}\right) \left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)} \tag{0.33}$$

Note that the results of a linear simple regression cannot be used to make predictions about the response to an input very different from the data used initially to perform the regression.

1.5 Coefficient of Determination and Sample correlation coefficient

The variation in successive values of the response $\{Y_i\}$ can result from the variation in the corresponding inputs $\{x_i\}$ as well as the variance of the random noise e which can cause even the same input to produce different responses.

The coefficient of determination R^2 is defined as the proportion of the variation that is explained by the change in input values.

$$R^2 = \frac{S_{YY} - SS_R}{S_{YY}} \tag{0.34}$$

$R^2 \in [0, 1]$ with values close to 1 implying that the variation in response is almost fully explained by the change in input. A large value of R indicates that the linear regression is a good fit.

Relation to sample correlation coefficient : r has been defined previously as

$$r^2 = \frac{S_{xY}^2}{S_{xx}\ S_{YY}} = \frac{S_{xx}\ S_{YY} - SS_R\ S_{xx}}{S_{xx}\ S_{YY}} = R^2 \tag{0.35}$$

$$|r| = \sqrt{R} \tag{0.36}$$

While $r \in [-1, 1]$ denotes the correspondence between increase in values of one variable and another, its square represents the extent to which a simple linear regression can explain the set of two data points.

1.6 Analysis of residuals:assessing the model

Even thought visual inspection of the scatter plot is a good method to determine how appropriate the choice of linear simple regression model is, any further doubts can be taken care of using the residuals $Y_i - (A + Bx_i)$.

Standardized residuals are obtained by transforming the residuals into standard normal RVs as follows,

$$Z \sim \frac{Y_i - (A + Bx_i)}{\sqrt{SS_R/(n-2)}} \qquad \forall\, i \in \{1, \ldots, n\} \tag{0.37}$$

Any observable pattern in a plot of the residuals that indicates deviation from Z is immediate evidence against choosing a simple linear regression model.

1.7 Transforming to linearity

In cases where the response is not a linear function of the input, it is useful to transform the relation using logarithms, exponentiation or other tools into a linear equation. This extends the power of the linear regression method to systems governed by non-linear relations.

For example, consider an exponential decay relationship between the number of items N and time t,

$$\begin{aligned}
N(t) &= u \ \exp(-vt) \\
\log(N) &= \log(u) - v \ \log(t) \\
\text{substituting } Y &= \log(N) \qquad \alpha = \log(u) \qquad \beta = -v \qquad x = \log(t) \\
Y &= \alpha + \beta x + e
\end{aligned}$$

1.8 Weighted Least Squares

When the assumption that the variance in the response is a constant independent of the input is no longer reasonable, the above procedure has to be modified to incorporate a weighting term in the variance,

$$\mathrm{Var}(Y_i) = \frac{\sigma^2}{w_i} \tag{0.38}$$

The estimators A, B must now minimize the sum of squares with these weights $\{w_i\}$. This is once again solved by setting the partial derivatives to zero and obtaining the *normal equations.*

$$SS_w = \frac{1}{\sigma^2} \sum_{i=1}^{n} w_i \, (Y_i - A - Bx_i)^2 \tag{0.39}$$

$$\sum w_i Y_i = A \sum w_i + B \sum w_i x_i \tag{0.40}$$

$$\sum w_i x_i Y_i = A \sum w_i x_i + B \sum w_i x_i^2 \tag{0.41}$$

The least squares estimator, even if unbiased, need not be the best estimator of the mean of a normal RV.

For a sample from a normal RV, the weighted least squares estimators of α, β happen to also be the MLE. Alternatively, consider the transformation to the error e, which removes its dependence on the input.

$$\begin{aligned} Y &= \alpha + \beta x + e \\ Y\sqrt{w} &= \alpha\sqrt{w} + \beta x\sqrt{w} + e\sqrt{w} \end{aligned} \tag{0.42}$$

The new error term $e\sqrt{w}$ is now independent of the input. It has mean 0 and constant variance. Now, the least squares estimators of α, β would be the ones that minimize

$$\text{minimize} \sum_{i=1}^{n} (Y_i\sqrt{w_i} - \alpha\sqrt{w_i} - \beta x_i\sqrt{w_i})^2$$

$$\text{minimize} \sum_{i=1}^{n} w_i \ (Y_i - \alpha - \beta x_i)^2$$

The weighted least squares method has the added advantage of giving greater emphasis to data points with the lower variance (and thus greater w_i).

For the special case when Y is a Poisson RV, a common transformation is to model $\sqrt{Y}$ as a linear simple regression, with the added approximation that $\text{Var}(\sqrt{Y}) \approx 0.25$ regardless of the parameter λ.

1.9 Polynomial Regression

The extension of the linear regression model to include terms corresponding to higher powers of the single input is straightforward, despite the heavy use of matrix operations.

Polynomial regression follows the same outline as simple regression with the minimization of SS_R leading to a set of *normal equations.*

For convenience, define $\sum_1^n (x_i)^j = K_j$ and for the RHS, $\sum_i^n (x_i)^j \ Y_i = M_j$

$$\begin{bmatrix} K_0 & K_1 & K_2 & \dots & K_r \\ K_1 & K_2 & K_3 & \dots & K_{r+1} \\ K_2 & K_3 & K_4 & \dots & K_{r+2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ K_r & K_{r+1} & K_{r+2} & \dots & K_{2r} \end{bmatrix} \begin{bmatrix} B_0 \\ B_1 \\ B_2 \\ \vdots \\ B_r \end{bmatrix} = \begin{bmatrix} M_0 \\ M_1 \\ M_2 \\ \vdots \\ M_r \end{bmatrix} \tag{0.43}$$

Since the sets $\{K_i\}$, $\{M_i\}$ are both fixed for a given dataset, the above system of linear equations is uniquely solved using matrix inversion. The value of r chosen should be as small as possible in order to avoid overfitting. This choice is usually made through visual inspection of the scatter diagram.

An unnecessarily large choice of degree r makes the regression model significantly worse at predicting the response to inputs far from the set of data points used to calculate it. In real-world problems, this also leads to incorrect inferences about the underlying physical mechanisms.

1.10 Multiple Linear Regression

Most real-world systems of interest are not governed by a single input. This requires generalizing the simple linear regression problem to deal with multiple inputs $\{x_i\}$.

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e \\ e &\sim \mathcal{N}(0, \sigma^2) \\ \mathbb{E}[Y_i] &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \end{aligned} \tag{0.44}$$

Once again, minimizing the SS_R by setting the derivative to zero yields a system of normal equations.

$$\text{minimizing } \sum_{i=1}^{n} (Y - B_0 - B_1 x_{i1} - B_2 x_{i2} - \cdots - B_k x_{ik})^2 \tag{0.45}$$

Defining the matrices for shorthand notation as follows,

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} B_0 \\ B_1 \\ B_2 \\ \vdots \\ B_k \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ 1 & x_{31} & x_{32} & \dots & x_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \tag{0.46}$$

The regression model and thus the normal equations can now be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{0.47}$$

$$\mathbf{X}^\intercal\mathbf{X}\mathbf{B} = \mathbf{X}^\intercal\mathbf{Y} \tag{0.48}$$

Since the existence of the inverse $(\mathbf{X}^\intercal\mathbf{X})^{-1}$ is not an issue in most real world datasets, simple matrix operations can be used to obtain the estimator matrix $\mathbf{B}$.

The set of least squares estimators $\mathbf{B}$ also happen to be unbiased estimators of $\boldsymbol{\beta}$. The variance of these estimators requires further analysis of the matrix $\mathbf{C} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\ \mathbf{X}^\intercal$

$$\mathbf{B} = \mathbf{C}\mathbf{Y}$$

$$B_{i-1} = \sum_{m=1}^{n} C_{im}Y_m \tag{0.49}$$

$$\text{Cov}(B_{i-1}, B_{j-1}) = \sigma^2\ (\mathbf{C}\mathbf{C}^\intercal)_{ij} \tag{0.50}$$

$$\text{Cov}(\mathbf{B}) = \sigma^2\ (\mathbf{X}^\intercal\mathbf{X})^{-1} \tag{0.51}$$

Since $\text{Cov}(B_i, B_i) = \text{Var}(B_i)$, the diagonal elements of the matrix of covariances $\text{Cov}(\mathbf{B})$, give the variances of the least squares estimators (scaled by σ^2).

The value of σ^2 can be estimated by rearranging the sum of squares of residuals SS_R into a chi-squared RV.

$$SS_R = \sum_{i=1}^{n}(Y - B_0 - B_1x_{i1} - B_2x_{i2} - \cdots - B_kx_{ik})^2$$

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-(k+1)} \tag{0.52}$$

$$\mathbb{E}\left[\frac{SS_R}{n-(k+1)}\right] = \sigma^2 \tag{0.53}$$

The above expression is an unbiased estimator of σ^2 and is also independent of the set of estimators $\mathbf{B}$.

Defining the residual as r_i along with the column matrix of the set of residuals $\mathbf{r}$ yields a useful computational formula for SS_R

$$\mathbf{r} = \mathbf{Y} - \mathbf{X}\mathbf{B} \tag{0.54}$$

$$SS_R = \mathbf{r}^\intercal \mathbf{r} \tag{0.55}$$

$$= \mathbf{Y}^\intercal \mathbf{Y} - \mathbf{B}^\intercal \mathbf{X}^\intercal \mathbf{Y} \tag{0.56}$$

Coefficient of multiple determination : The quantity R^2 is defined as follows. It measures the decrease in the sum of squares of residuals SS_r when using a multiple regression model.

$$R^2 = 1 - \frac{SS_R}{\sum (Y_i - \overline{Y})^2} \tag{0.57}$$

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + e \qquad \text{vs} \qquad Y = \beta_0 + e \tag{0.58}$$

Predicting future responses : A point estimate of the mean response is simply

$$\mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

$$\widehat{\mathbb{E}[Y|\mathbf{x}]} = \sum_{i=0}^{k} B_i x_i \tag{0.59}$$

Here the first term is special because $x_o \equiv 1$. In order to find an interval estimate, consider the distribution of the above point estimator, which is a linear combination of normal RVs $\{Y_i\}$.

$$\mathbb{E}\left[\sum_{i=0}^{k} B_i x_i\right] = \sum_{i=0}^{k} \beta_i x_i \tag{0.60}$$

$$\mathrm{Var}\left(\sum_{i=0}^{k} B_i x_i\right) = \mathbf{x}^{\intercal}\ \left(\mathbf{X}^{\intercal}\mathbf{X}\right)^{-1}\ \mathbf{x}\ \sigma^2 \tag{0.61}$$

$$\tag{0.62}$$

Here, $\mathbf{x}$ is the column matrix composed of the set of input variables $\{x_0, \ldots, x_k\}$. Additionally, replacing σ with its estimator $\sqrt{SS_R/(n-k-1)}$, yields a t-RV with $n-k-1$ DOF.

$$\frac{\displaystyle\sum_{i=0}^{k} B_i x_i - \sum_{i=0}^{k} \beta_i x_i}{\sqrt{\mathbf{x}^{\intercal}\ \left(\mathbf{X}^{\intercal}\mathbf{X}\right)^{-1}\ \mathbf{x}\ \left(\dfrac{SS_R}{n-k-1}\right)}} \sim t_{n-k-1} \tag{0.63}$$

The above distribution can be used to construct confidence intervals for $\mathbb{E}[Y|\mathbf{x}]$.

As opposed to predicting the mean value of the response, when the experiment is to be performed only once in order to generate a single data point, it is more relevant to predict the response $Y(\mathbf{x})$ itself.

$$Y(\mathbf{x}) = \sum_{i=0}^{k} \beta_i x_i + e \qquad \text{where} \qquad x_0 \equiv 1 \tag{0.64}$$

The point estimator is clearly just $\sum B_i x_i$ based on the previous n data points. Since $\sum B_i x_i$ are thus independent of $Y(\mathbf{x})$,

$$\text{Var}\left[Y(\mathbf{x}) - \sum_{i=0}^{k} B_i x_i\right] = \sigma^2 + \mathbf{x}^\intercal \ (\mathbf{X}^\intercal \mathbf{X})^{-1} \ \mathbf{x} \ \sigma^2 \tag{0.65}$$

$$\frac{Y(\mathbf{x}) - \sum_{i=0}^{k} B_i x_i}{\sqrt{(1 + \mathbf{x}^\intercal \ (\mathbf{X}^\intercal \mathbf{X})^{-1} \ \mathbf{x}) \ \left(\frac{SS_R}{n-k-1}\right)}} \sim t_{n-k-1} \tag{0.66}$$

Dummy variables for categorical data : A binary categorical variable can be represented as a dummy variable in the regression that takes values of $\{0, 1\}$ depending on the data point belonging to the category or not. This is only necessary when the dataset is too small to be fragmented into multiple datasets each analyzed separately.

For large enough datasets, it is better to avoid the use of dummy variables and simply perform many separate regression analyses on the different data-subsets.

Logistic regression models for binary response data : Consider an experiment which only gives a binary response, with the probability of success $p(x)$ governed by the distribution. Additionally define the odds of success $o(x)$ as the ratio of the win to loss probabilities.

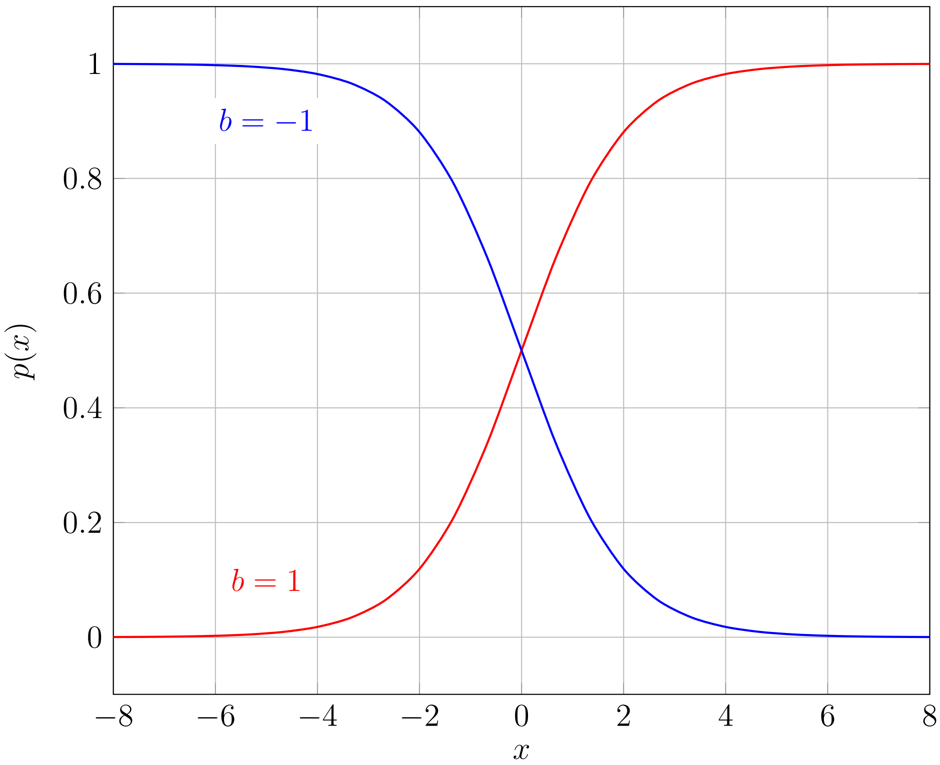
$$p(x) = \frac{\exp(a + bx)}{1 + \exp(a + bx)} \tag{0.67}$$

$$o(x) = \frac{p(x)}{1 - p(x)} = \exp(a + bx) \tag{0.68}$$

$$\log(o(x)) = a + bx \tag{0.69}$$

From the plot below, the asymptotic values of this function at $-\infty, \infty$ are $0, 1$ depending on the sign of b.

logistic function with $a = 1$



In order to find the maximum likelihood estimators, consider the joint PDF of a set of binary responses $\{Y_1, \ldots, Y_k\}$,

$$\log(P\{Y_i = y_i;\ i = 1, \ldots, k\}) = \sum_{i=1}^{k} y_i(a + bx_i) - \sum_{i=1}^{k} \log(1 + \exp(a + bx_i)) \tag{0.70}$$

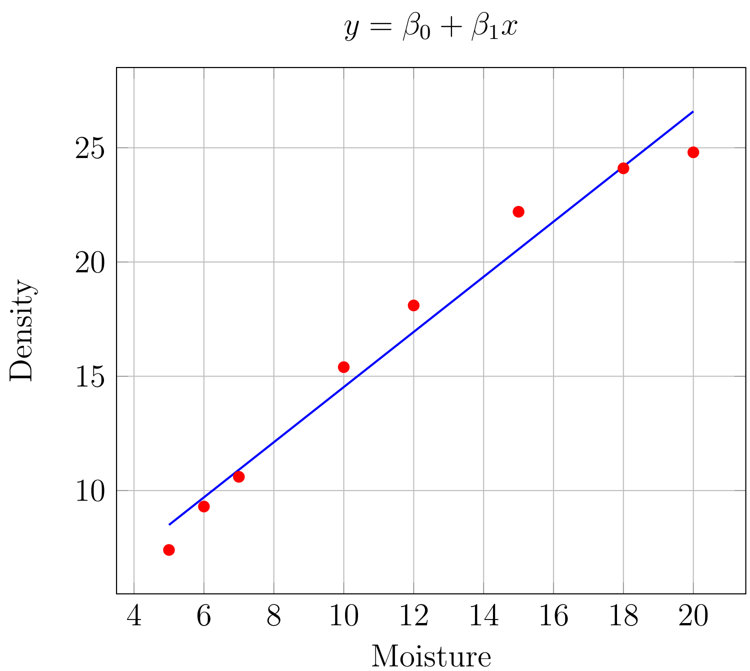
Even though an analytical minimization of the above expression is not possible, there are several iterative computational approaches possible.

Chapter 2

Regression

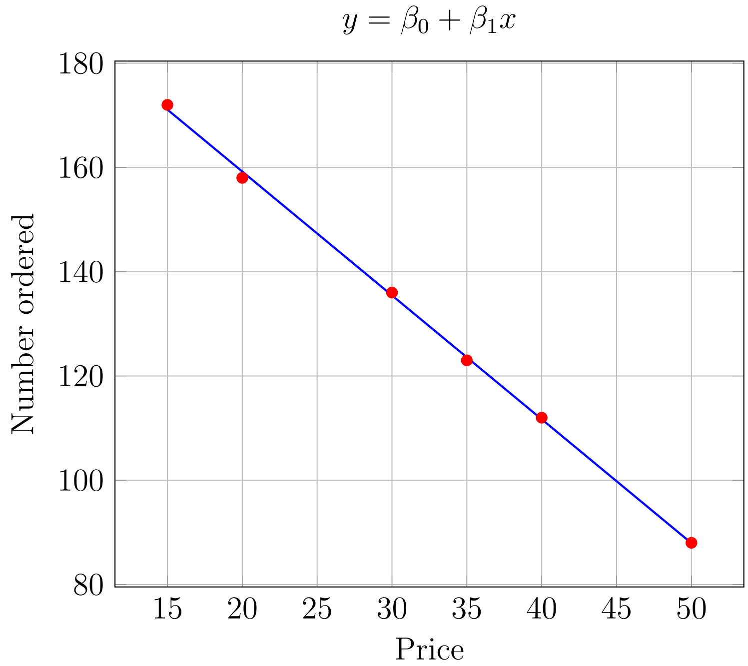
1 Performing the linear regression using `scipy.stats.linregress` outputs.

	`linRegressOutput`
β_1	1.206366
β_0	2.463493
r	0.985651
p	0.000007



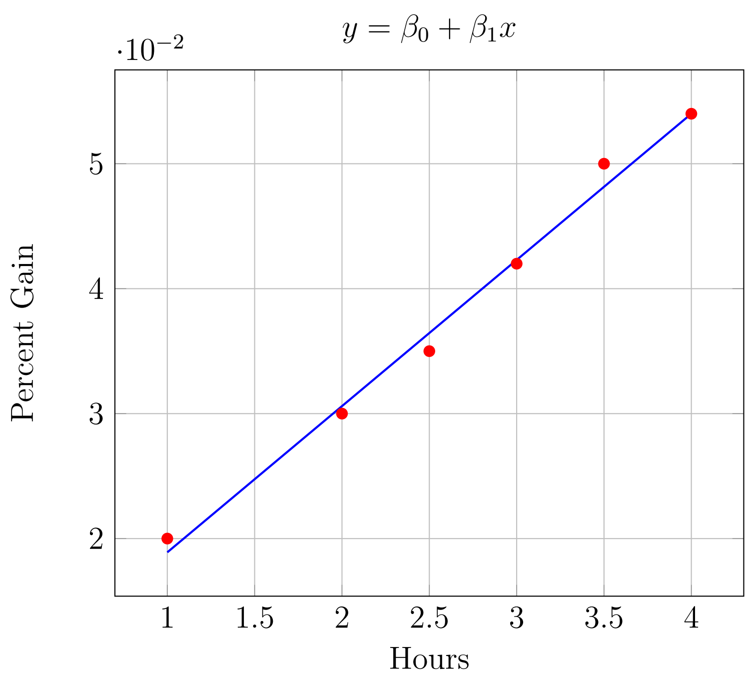
2 Performing the linear regression using `scipy.stats.linregress` outputs.
The value corresponding to $x = 25$ is $y = 147.34$.

linRegressOutput	
β_1	-2.376000e+00
β_0	2.067400e+02
r	-9.996792e-01
p	1.543680e-07



3 Performing the linear regression using `scipy.stats.linregress` outputs.
The value corresponding to $x = 3.2$ is $y = 0.04476$.

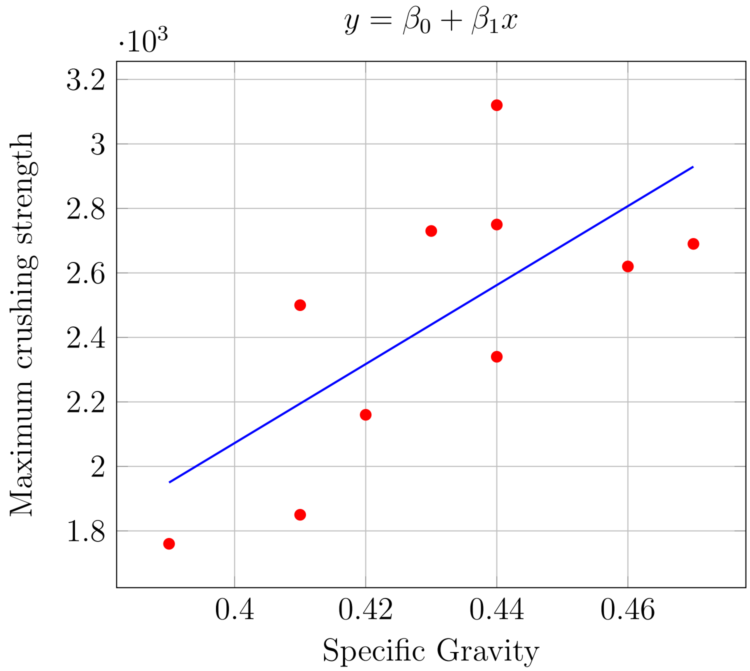
linRegressOutput	
β_1	0.011743
β_0	0.007186
r	0.995607
p	0.000029



4 Performing the linear regression using `scipy.stats.linregress` outputs.

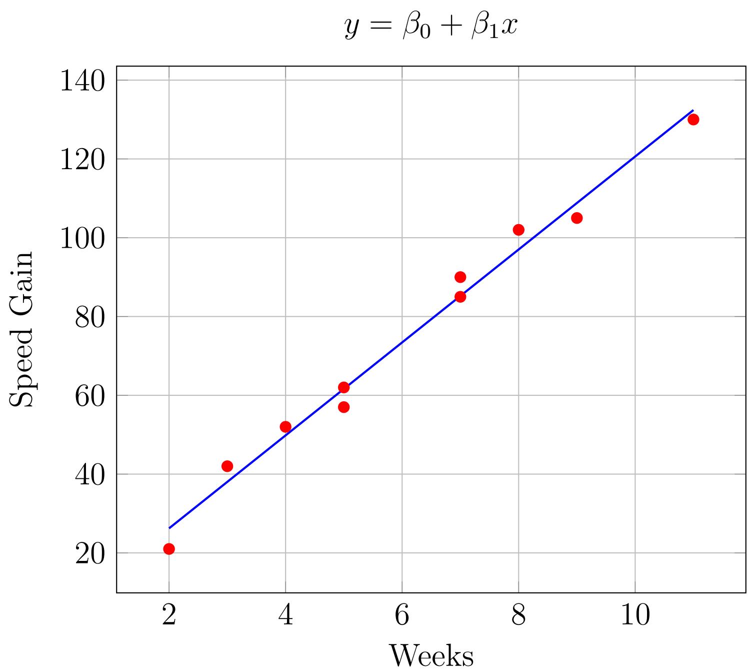
The value corresponding to $x = 0.43$ is $y = 2439.75$.

linRegressOutput	
β_1	12245.746692
β_0	-2825.916824
r	0.695796
p	0.025447



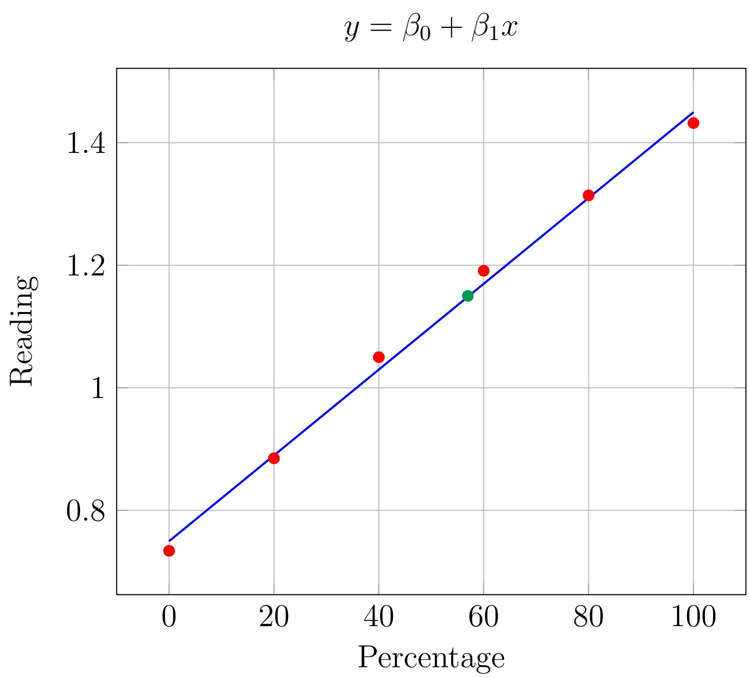
5 Performing the linear regression using `scipy.stats.linregress` outputs.
The least squares estimators are $A = 2.64$ and $B = 11.79$.
The value corresponding to $x = 7$ is $y = 85.22$.

linRegressOutput	
β_1	1.179690e+01
β_0	2.638928e+00
r	9.931055e-01
p	9.803790e-09



6 Performing the linear regression using `scipy.stats.linregress` outputs.
The value corresponding to $y = 1.15$ is $x = 57\%$.

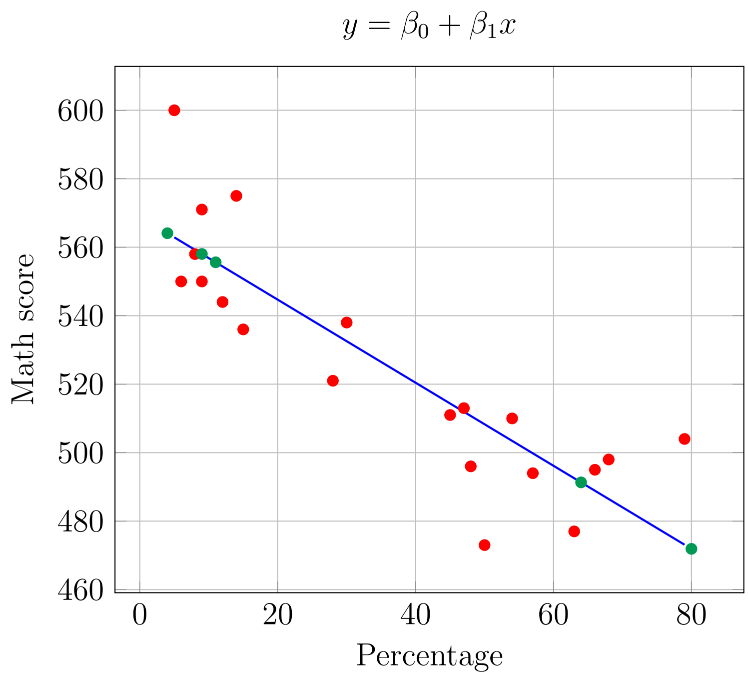
linRegressOutput	
β_1	0.007026
β_0	0.749714
r	0.997906
p	0.000007



7 Performing the linear regression on the first 20 states using `scipy.stats.linregress` out-

puts.
The predicted values for the next 5 states are plotted. mean of the first 20 states' math scores is $\overline{Y} = 525.7$

linRegressOutput	
β_1	-1.212983e+00
β_0	5.689428e+02
r	-8.640923e-01
p	9.078458e-07



8 Verifying the expression for Var(A),

$$\text{Var}(Y_i) = \sigma^2$$

$$\text{Var}(B) = \frac{\text{Var}\left[\sum(x_i - \overline{x})\ (Y_i - \overline{Y})\right]}{S_{xx}^2}$$

$$= \frac{\sum(x_i - \overline{x})^2\ \text{Var}(Y_i - \overline{Y})}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}} \tag{8.1}$$

$$\text{Var}(A) = \overline{x}^2\ \text{Var}(B) + \frac{\text{Var} \sum y_i}{n^2} + 2\ \text{Cov}\left(\frac{\sum y_i}{n}\ ,\ -B\overline{x}\right)$$

$$= \frac{\sigma^2}{n} + \overline{x}^2 \frac{\sigma^2}{S_{xx}} - \frac{2B}{n}\ \text{Cov}\left(\sum y_i\ ,\ B\right)$$

$$= \frac{\sum x_i^2}{n} \frac{\sigma^2}{S_{xx}} \tag{8.2}$$

The covariance term vanishes as shown below.

$$
\begin{aligned}
\mathrm{Cov}\left(\sum y_i \ , \ B\right) &= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}\left(y_i \ , \ \frac{(x_j-\overline{x})\ y_j}{S_{xx}}\right) \\
&= \frac{\sigma^2}{S_{xx}} \ \sum_{i=1}^{n}(x_i-\overline{x}) = 0 \qquad (8.3)
\end{aligned}
$$

9 Using the data from Problem 4,

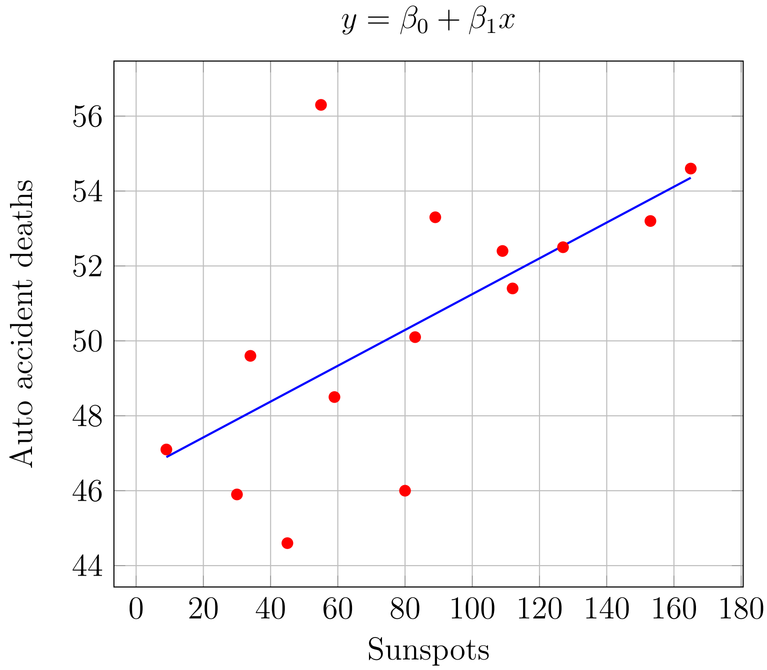
$\sigma^2 = 105660$ is a point estimate along with an interval estimate based on the χ^2 distribution give by $[54508, 309327]$

10 To prove the relation for SS_R using the previously defined shorthand notation S_{xx}, S_{YY}, S_{xY},

$$
\begin{aligned}
SS_R &= \sum(Y_i - A - Bx_i)^2 = \sum(Y_i - \overline{Y} + B\overline{x} - Bx_i)^2 \qquad (10.1) \\
&= \sum(Y_i-\overline{Y})^2 + B^2 \ \sum(\overline{x}-x_i)^2 + 2B \ \sum(Y_i-\overline{Y})(\overline{x}-x_i) \\
&= S_{YY} + \frac{S_{xY}^2}{S_{xx}} - 2 \ \frac{S_{xY}^2}{S_{xx}} \\
&= \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}} \qquad (10.2)
\end{aligned}
$$

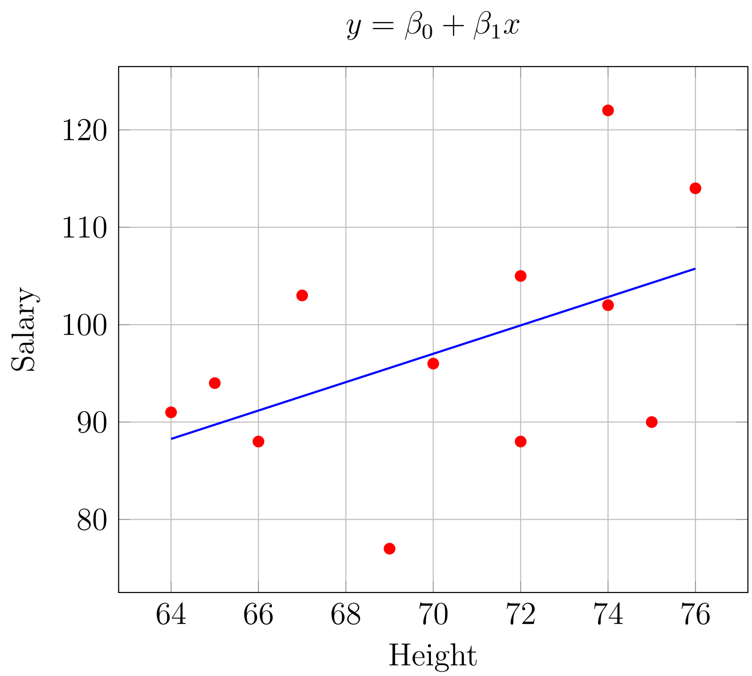
11 Performing the linear regression using `scipy.stats.linregress` outputs.
Testing the hypothesis that $\beta_1 = 0$ gives a p-value of 1.6%. The hypothesis can be rejected at 5% confidence level.

linRegressOutput	
β_1	0.047795
β_0	46.466853
r	0.626258
p	0.016568



12 Performing the linear regression using `scipy.stats.linregress` outputs.
Testing the hypothesis that $\beta_1 = 0$ gives a p-value of 11.1%. The hypothesis cannot be rejected at 5% confidence level. This means that salary is not related to height.

linRegressOutput	
β_1	1.457143
β_0	-4.985714
r	0.483846
p	0.110981



13 Given $0 < \beta < 1$ in a simple linear regression model.

(a)

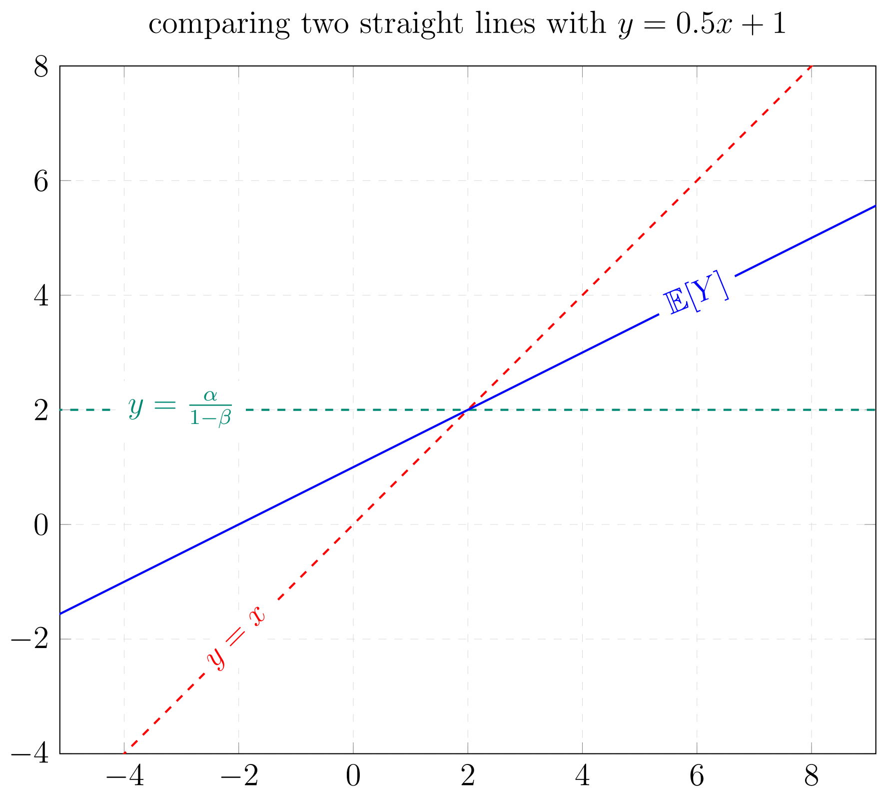
$$x < \frac{\alpha}{(1-\beta)} \rightarrow x < \alpha + \beta x$$

$$\mathbb{E}[Y_i] = \alpha + \beta x \rightarrow \mathbb{E}[Y] > x$$

$$\mathbb{E}[Y] < \frac{\alpha}{1-\beta} \qquad \text{obvious} \tag{13.1}$$

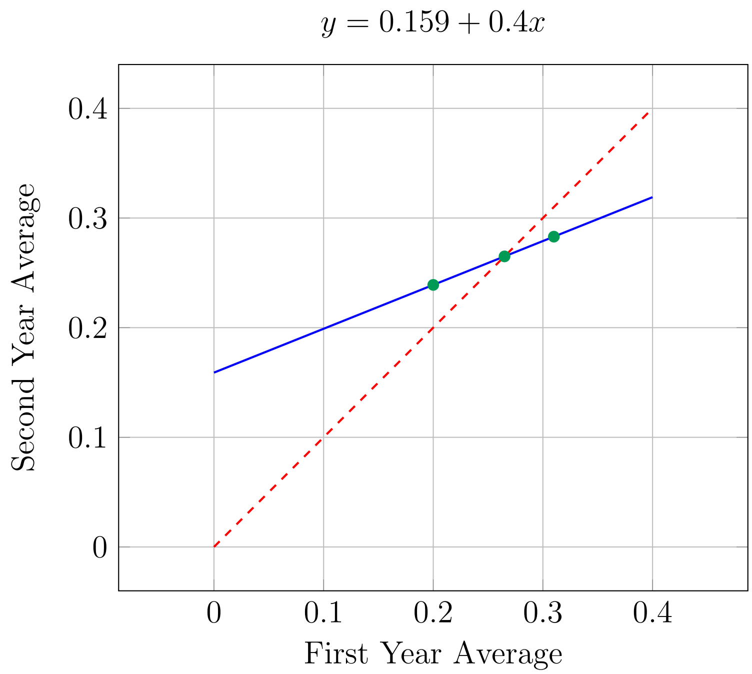
$$\therefore x < \mathbb{E}[Y] < \frac{\alpha}{1-\beta} \tag{13.2}$$

(b) Same as previous part with all inequalities reversed. This proves that $\mathbb{E}[Y]$ always lies in between these two values.



14 Given a linear regression $Y = 0.159 + 0.4X + e$, with $e \sim \mathcal{N}(0, \sigma^2)$. The predictions are outlined in the figure below.

X	Y
0.200	0.239
0.265	0.265
0.310	0.283



15 This is a case of gambler's fallacy. When successive events are independent, it is possible merely through chance for successive outcomes to be good or bad without any causal effect of one outcome on the next. For example, tossing a coin 10 times can result in 10 heads just by chance.

16 In order to rearrange the estimator A of the regression parameter α into a t-RV with $n-2$ DOF,

$$A \sim \mathcal{N}\left(\alpha \ , \ \frac{\sigma^2}{S_{xx}} \frac{\sum x_i^2}{n}\right)$$

$$\frac{(A-\alpha)}{\sigma}\sqrt{\frac{nS_{xx}}{\sum x_i^2}} \sim Z \tag{16.1}$$

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$$

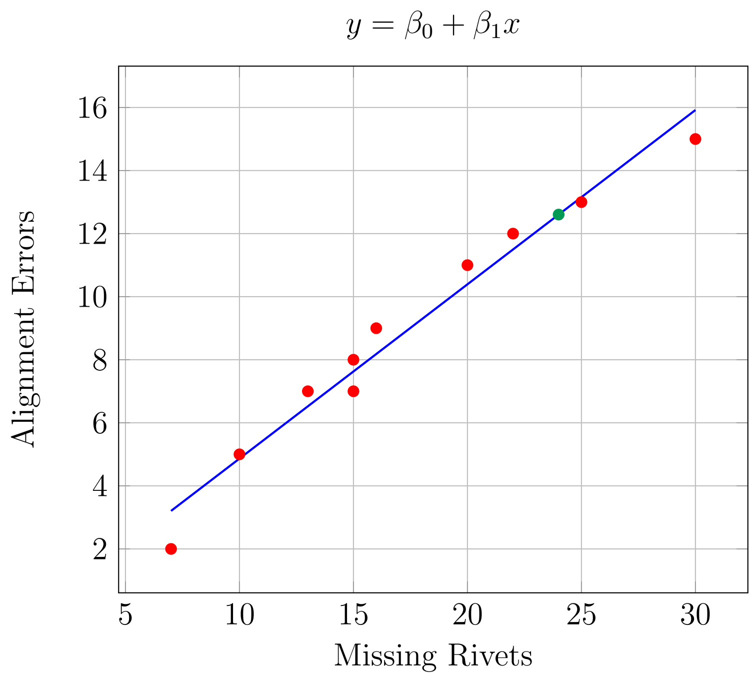
$$\frac{1}{\sigma}\sqrt{\frac{SS_R}{n-2}} \sim \sqrt{\frac{\chi^2_{n-2}}{(n-2)}} \tag{16.2}$$

Rearranging into a t-RV proves the result.

$$
\begin{aligned}
t_{n-2} &\sim \frac{Z}{\sqrt{\chi^2_{n-2}/(n-2)}} \\
t_{n-2} &\sim (A-\alpha)\sqrt{\frac{n(n-2)S_{xx}}{SS_R \sum x_i^2}}
\end{aligned} \tag{16.3}
$$

17 Performing the linear regression using `scipy.stats.linregress` outputs. The value corresponding to $x = 24$ is $y = 12.604$.

linRegressOutput	
β_1	5.528289e-01
β_0	-6.639400e-01
r	9.840458e-01
p	2.780621e-07



Hypothesis testing with $\gamma = 10\%$ for $H_0 : \alpha = 1$ gives a p-value of 3.46%. This means that the hypothesis can be rejected.

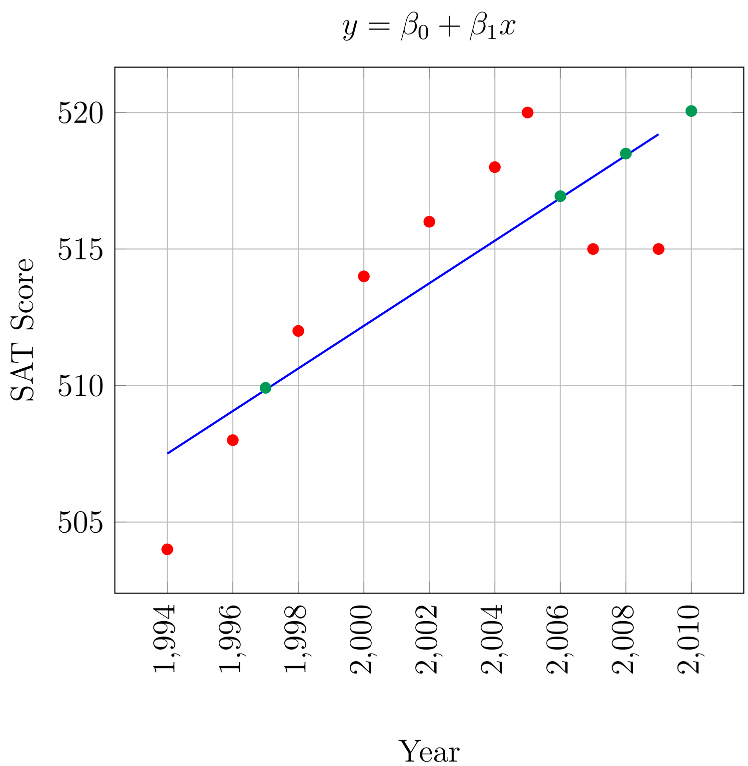
$$
\begin{aligned}
&H_0 : \alpha = 1 \qquad \text{vs.} \qquad H_1 : \alpha \neq 1 \\
&\text{reject } H_0 \text{ if} \qquad \sqrt{\frac{n(n-2)S_{xx}}{SS_R \, \sum x_i^2}}\; |A-\alpha| > t_{\gamma/2,n-2} \\
&\text{accept } H_0 \qquad \text{otherwise}
\end{aligned} \tag{17.1}
$$

In order to find a confidence interval for $x_0 = 24$, using the above t-test,

$$
\begin{aligned}
Y(x_0) &\in A + Bx_0 \pm t_{\gamma/2,n-2}\sqrt{\left(\frac{SS_R}{n-2}\right)\left(\frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right)} \\
Y(x_0) &\in 12.604 \pm 0.6196 = [11.98, 13.22] \qquad (17.2)
\end{aligned}
$$

18 Performing the linear regression using `scipy.stats.linregress` outputs.

	`linRegressOutput`
β_1	0.779935
β_0	-1047.614887
r	0.799131
p	0.009766

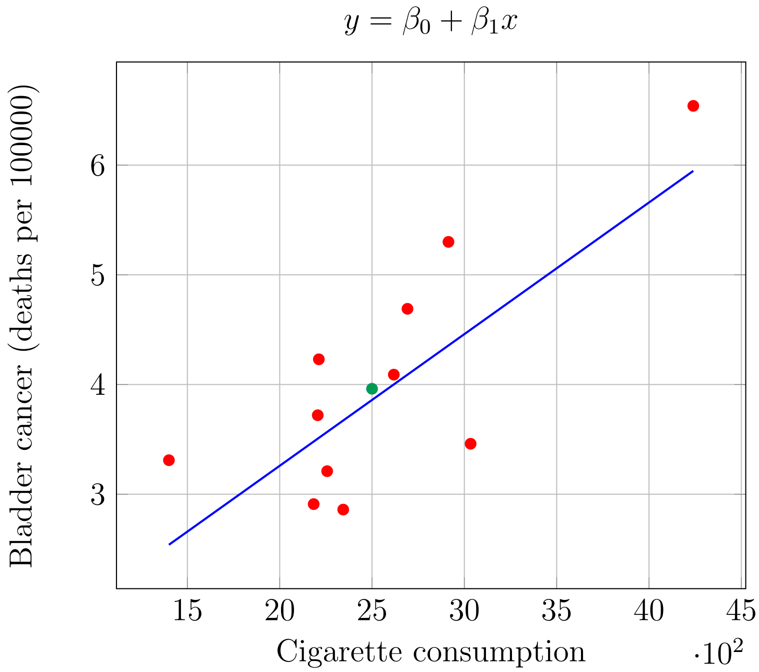


The predicted values are as tabulated as

X	Y
1997	509.90
2006	516.93
2008	518.49
2010	520.05

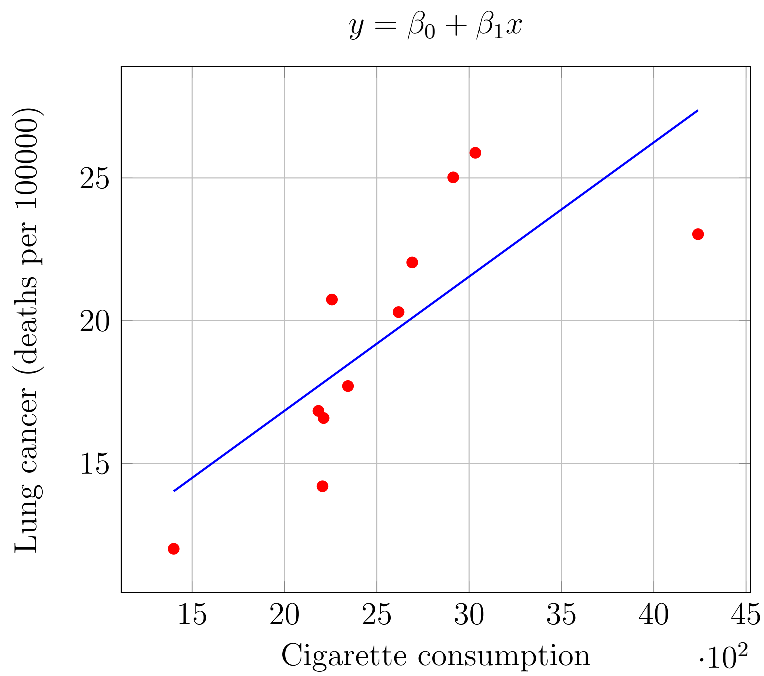
19 Performing the linear regression using `scipy.stats.linregress` outputs for bladder cancer.
Linear relationship is not very strong with p-value of 0.4%
The value corresponding to $x = 2500$ is $y = 3.96$.

`linRegressOutput`	
β_1	0.001241
β_0	0.859260
r	0.785832
p	0.004140



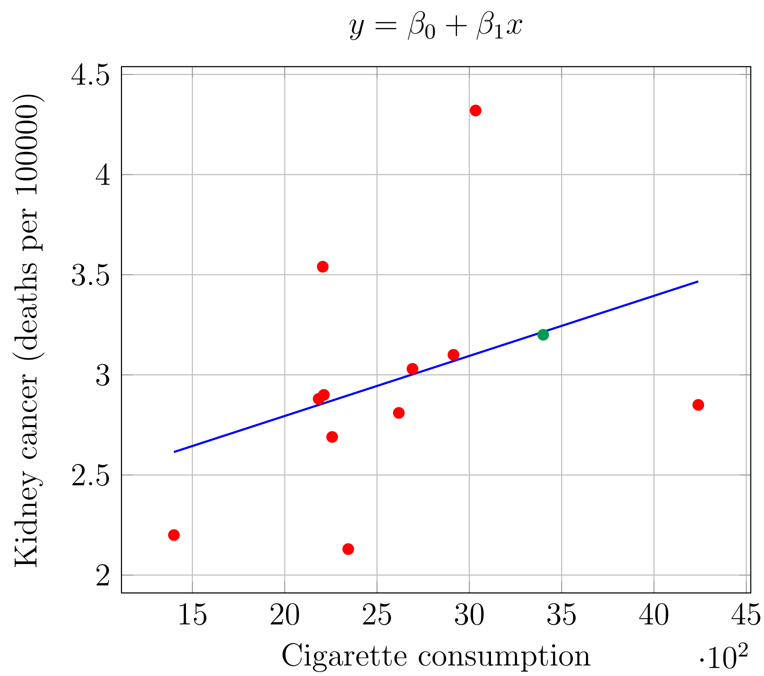
20 Performing the linear regression using `scipy.stats.linregress` outputs for lung cancer.
Linear relationship is very strong with p-value of 0.67%
The value corresponding to $x = 2500$ is $y = 3.96$.

linRegressOutput	
β_1	0.004715
β_0	7.443316
r	0.759411
p	0.006706



21 Performing the linear regression using `scipy.stats.linregress` outputs for kidney cancer.
Linear relationship is nonexistent with p-value of 29%
The value corresponding to $x = 2500$ is $y = 3.96$.

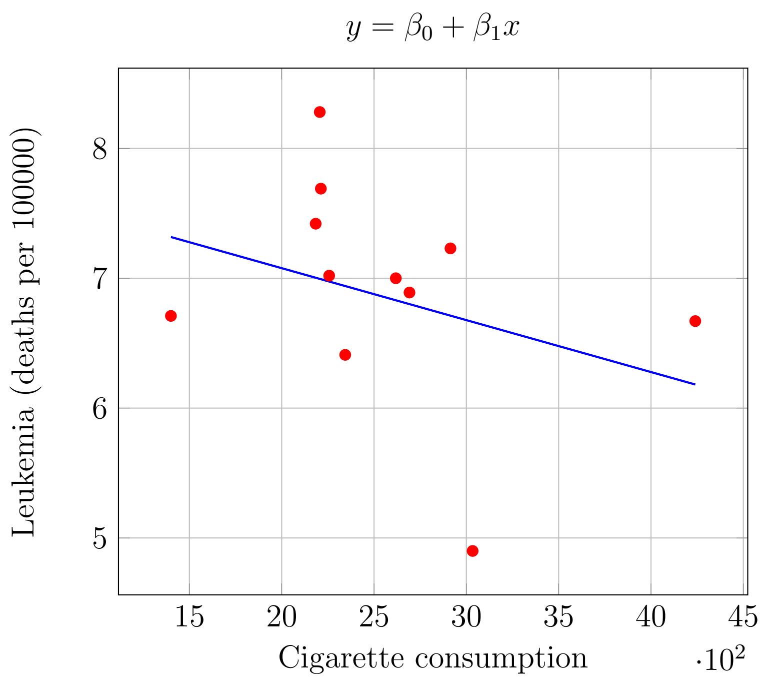
linRegressOutput	
β_1	0.000296
β_0	2.194573
r	0.351066
p	0.289780



The 90% confidence interval for the mean death rate with $x_0 = 3400$ is given by
$Y(x_0) = 3.2 \pm 0.52 = [2.67, 3.72]$

22 Performing the linear regression using `scipy.stats.linregress` outputs for leukemia.
Linear relationship is nonexistent with p-value of 35.31% for the hypothesis $H_0 : \beta = 0$
The value corresponding to $x = 2500$ is $y = 3.96$.
The 90% confidence interval for the mean death rate with $x_0 = 2500$ is given by
$Y(x_0) = 6.95 \pm 0.47 = [6.48, 7.42]$

linRegressOutput	
β_1	-0.000371
β_0	7.877593
r	-0.310281
p	0.353081



23 The variances along with the 95% confidence intervals are as follows,

The predicted values are as tabulated as

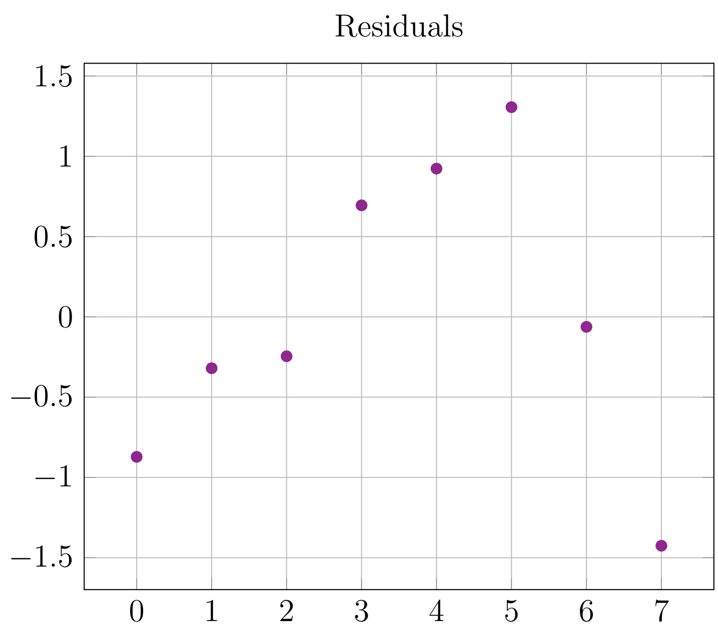
Disease	σ^2	Interval
Bladder Cancer	0.54	[0.2538, 1.7884]
Lung Cancer	9.18	[4.3436, 30.5983]
Kidney Cancer	0.35	[0.1656, 1.1667]
Leukemia	0.73	[0.3445, 2.4270]

The variance for the two sets of data are $\sigma_1^2 = 6.56$ and $\sigma_2^2 = 9.02$. Since the two variances can be arranged into χ^2 RVs, their ratio is an f-RV and this can be used to perform the significance test.

$$\frac{SS_{R1}}{\sigma_1^2}\,\frac{\sigma_2^2}{SS_{R2}} \sim \frac{\chi^2_{n-2}}{\chi^2_{m-2}}$$
$$\frac{SS_{R1}}{\sigma_1^2}\,\frac{\sigma_2^2}{SS_{R2}}\,\frac{m-2}{n-2} \sim f_{n-2,m-2} \tag{23.1}$$

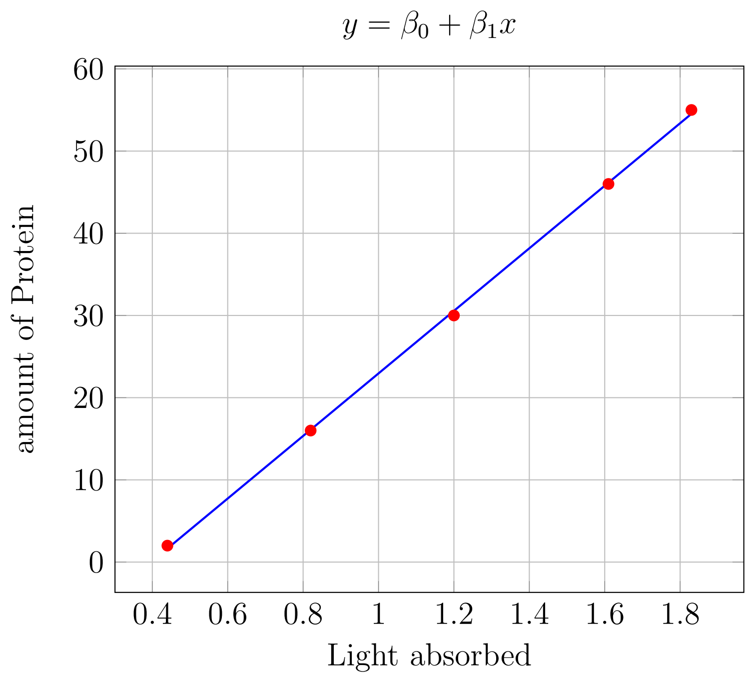
Using the above f-test, the p-value for equality of the two variances is 95.78%. This is sufficient to reject the alternative hypothesis at the 5% confidence level.

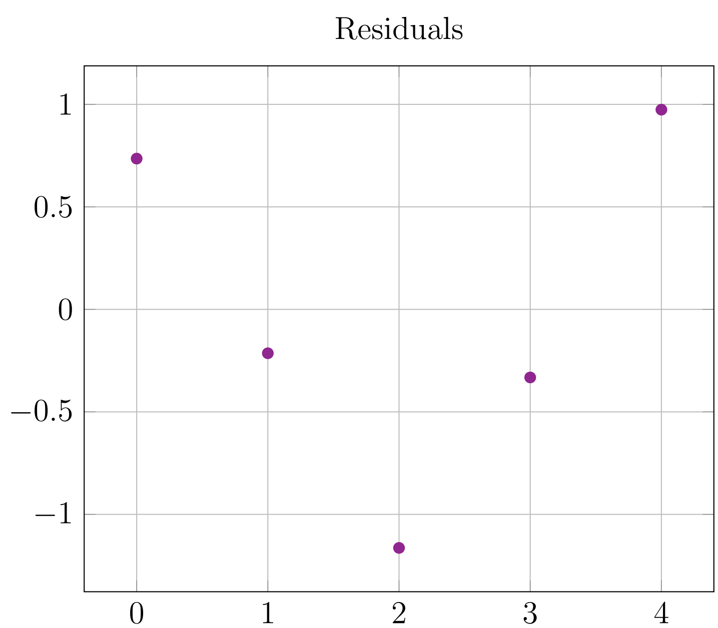
24 Plotting standardized residuals shows that the assumption of linearity is reasonable. There are no discernible trends among the residual data points.



25 $R = r^2 = 0.999625$.
p-value is so small for the linear regression that non-linearity can be strongly rejected, although there is a clear pattern in the residuals plot.
The 90% confidence interval for the amount of protein with $x_0 = 1.5$ is given by
$Y(x_0) = 41.97 \pm 1.2834 = [40.69, 43.26]$

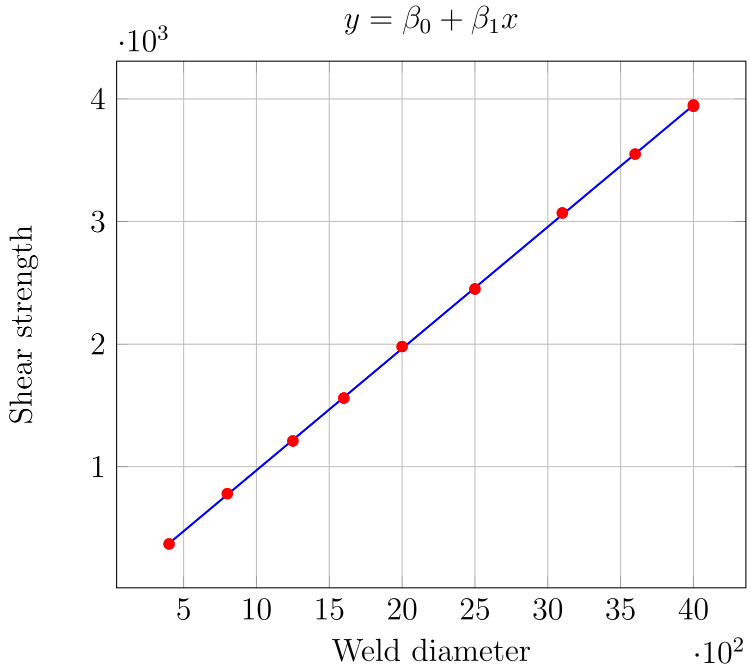
linRegressOutput	
β_1	38.046693
β_0	-15.095097
r	0.999813
p	0.000003



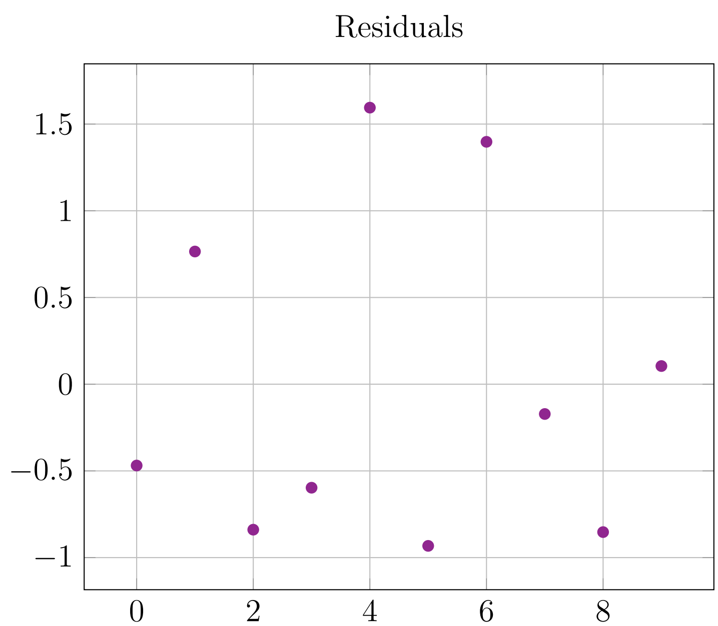


26 The hypothesis $H_0 : \beta = 1$ has a p-value of 2.55% and thus can be rejected at the 5% level of significance.

linRegressOutput	
β_1	9.927803e-01
β_0	-2.221420e+01
r	9.999718e-01
p	2.771209e-18

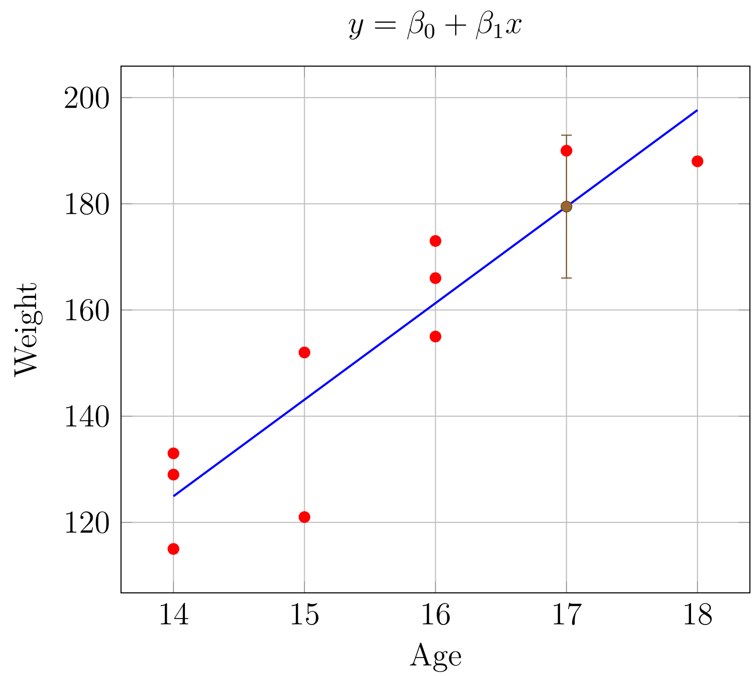


The expected value for $y = 2500$ is $x = 2459.74$
Plotting the residuals shows no discernible pattern, validating the linear model assumption.
Prediction interval for $x = 0.2250$ is $y = 2211.54 \pm 25.26 = [2186.28, 2236.80]$



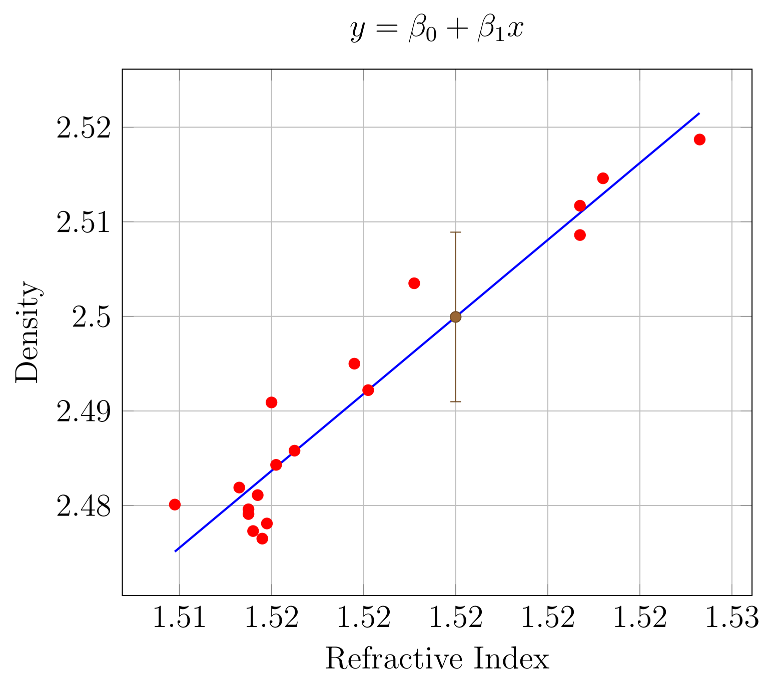
27 Performing the linear regression using `scipy.stats.linregress` outputs
Prediction interval for $x = 17$ is $y = 179.47 \pm 13.45 = [166.02, 192.92]$

linRegressOutput	
β_1	18.181818
β_0	-129.618182
r	0.908705
p	0.000272



28 Performing the linear regression using `scipy.stats.linregress` outputs
Prediction interval for $x = 1.52$ is $y = 2.5 \pm 0.0089 = [2.4909, 2.5089]$

linRegressOutput	
β_1	4.069680e+00
β_0	-3.685973e+00
r	9.607671e-01
p	2.495152e-10



29 (a) Regression through the origin has the model $Y = \beta x + e$ where $e \sim \mathcal{N}(0, \sigma^2)$

$$SS = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - Bx_i)^2$$

$$\text{differentiating, } \frac{\mathrm{d}}{\mathrm{d}B}\ SS = 0$$

$$0 = -2\sum_{i=1}^{n} x_i (Y_i - Bx_i)$$

$$B^* = \frac{\sum x_i Y_i}{\sum x_i^2} \tag{29.1}$$

(b) Given the fact that the individual $Y_i \sim \mathcal{N}(\beta x_i, \sigma^2)$, and B is a linear combination of the set of $\{Y_i\}$,

$$\mathbb{E}[B] = \frac{1}{\sum x_i^2}\ \sum_{i=1}^{n} x_i \mathbb{E}[Y_i]$$

$$= \frac{1}{\sum x_i^2}\ \beta \sum_{i=1}^{n} x_i^2 = \beta \tag{29.2}$$

$$\mathrm{Var}(B) = \frac{\sigma^2 (\sum x_i^2)}{(\sum x_i^2)^2} = \frac{\sigma^2}{\sum x_i^2} \tag{29.3}$$

Thus, B is a normal RV with the above mean and variance.

(c) SS_R is defined as $\sum (Y_i - Bx_i)^2$. Each term is the square of a linear combination of normal RVs and thus is also a squared normal RV. This means that SS_R is a χ^2 RV. The DOF of this χ^2 RV is $(n-1)$ as one of the n DOF is lost in determining β. Same logic as for the general case of nonzero α.

(d) To derive a hypothesis test using the normal RV of B,

$$H_0 : \beta = \beta_0 \qquad \text{and} \qquad H_1 : \beta \neq \beta_0$$

$$\frac{B - \beta}{\sigma / \sqrt{\sum x_i^2}} \sim Z \qquad \text{and} \qquad \frac{SS_R}{\sigma^2} \sim \chi^2_{n-1} \tag{29.4}$$

$$\sqrt{\frac{(n-1)\sum x_i^2}{SS_R}}(B - \beta) \sim t_{n-1}$$

$$\text{reject } H_0 \text{ if} \qquad \sqrt{\frac{(n-1)\sum x_i^2}{SS_R}}\ |B - \beta_0| > t_{\gamma/2, n-1}$$

$$\text{accept } H_0 \qquad \text{otherwise} \tag{29.5}$$

(e) To find a prediction interval for $Y(x_0)$,

$$Y - Y_0 = Y - Bx_0 \sim \mathcal{N}\left(0\ ,\ \sigma^2 + \frac{x_0^2\sigma^2}{\sum x_i^2}\right)$$

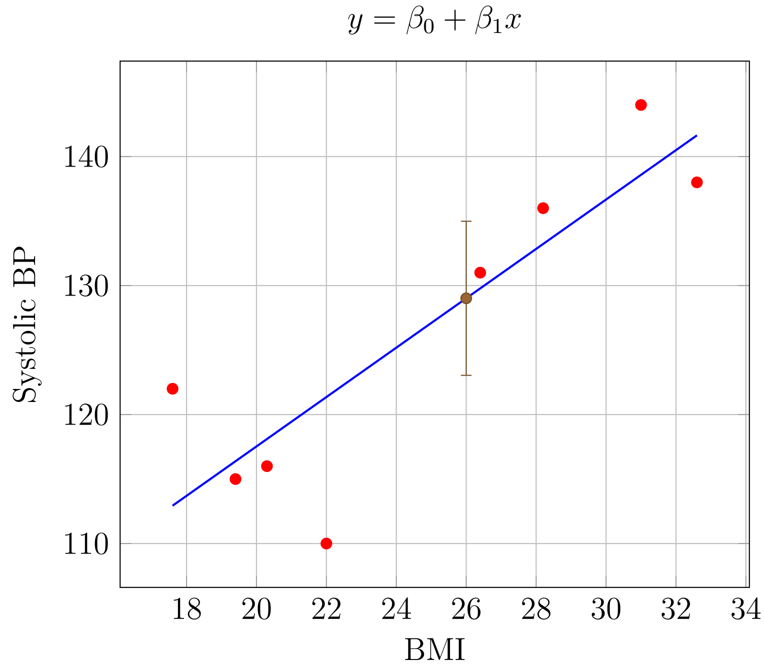
$$t_{n-1} \sim \sqrt{\frac{(n-1)}{SS_R}\ \left(\frac{\sum x_i^2}{\sum x_i^2 + x_0^2}\right)}(Y - Bx_0)$$

$$Y \in Bx_0 \pm t_{\gamma/2, n-1}\sqrt{\frac{SS_R}{(n-1)}\ \left(\frac{\sum x_i^2 + x_0^2}{\sum x_i^2}\right)} \tag{29.6}$$

This gives the $100(1-\gamma)\%$ confidence interval for the prediction Y.

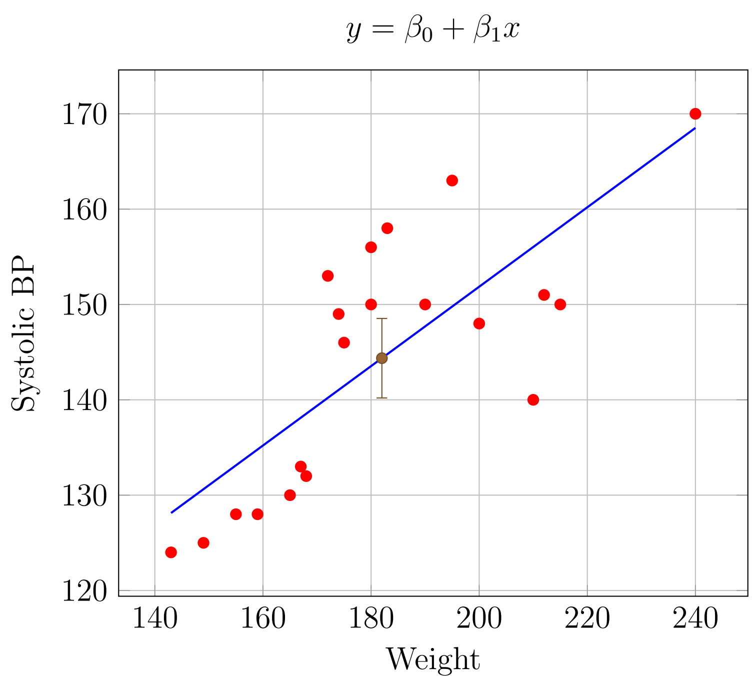
30 Performing the linear regression using `scipy.stats.linregress` outputs
Prediction interval for $x = 1.52$ is $y = 2.5 \pm 0.0089 = [2.4909, 2.5089]$

linRegressOutput	
β_1	1.913991
β_0	79.248355
r	0.866925
p	0.005319

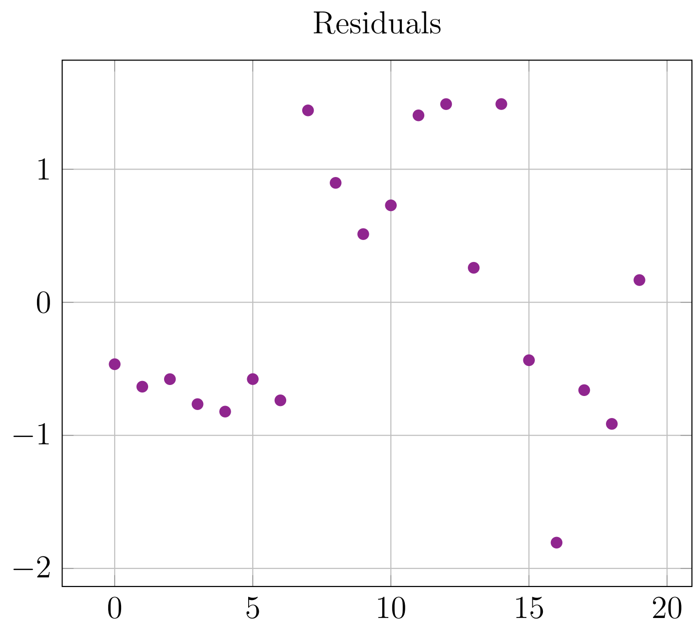


31 Performing the linear regression using `scipy.stats.linregress` outputs
The hypothesis $H_0 : \beta_1 = 0$ has a p-value of 0.01% and can be rejected strongly. The Systolic Bp does depend on the Weight.
Mean response interval at 95% confidence for $x = 182$ is $y = 144.37 \pm 4.17 = [140.2, 148.53]$

linRegressOutput	
β_1	0.416385
β_0	68.584433
r	0.764442
p	0.000087



Plotting the residuals shows no discernible pattern, vindicating the linear model.
The sample correlation coefficient $r = 0.76444$



32 Transforming to linearity,

$$S = \frac{A}{N^m}$$

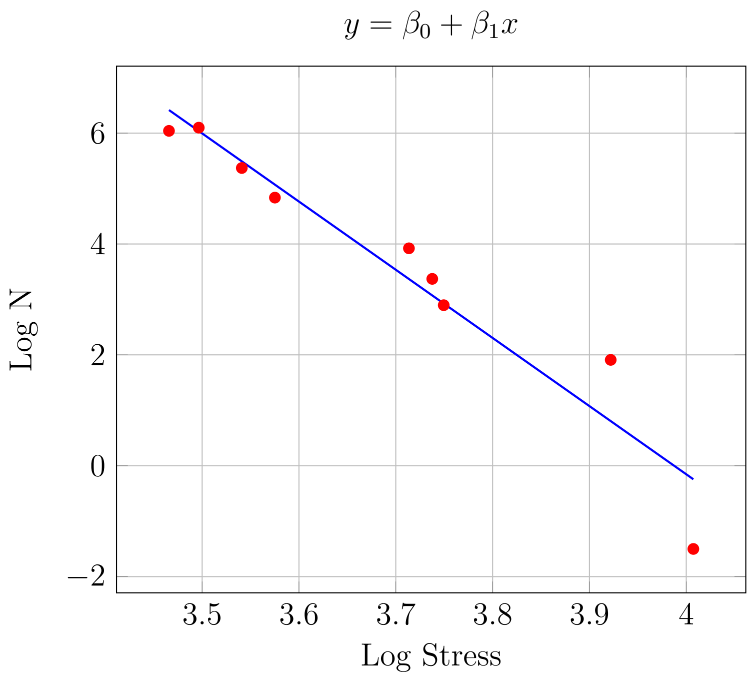
$$\log N = (1/m) \log A - (1/m)\ \log S \tag{32.1}$$

$$m = \frac{-1}{\beta_1} = 0.0813$$

$$A = e^{m\beta_0} = 53.91$$

Performing the linear regression using `scipy.stats.linregress` outputs

linRegressOutput	
β_1	-12.296305
β_0	49.031503
r	-0.962532
p	0.000032

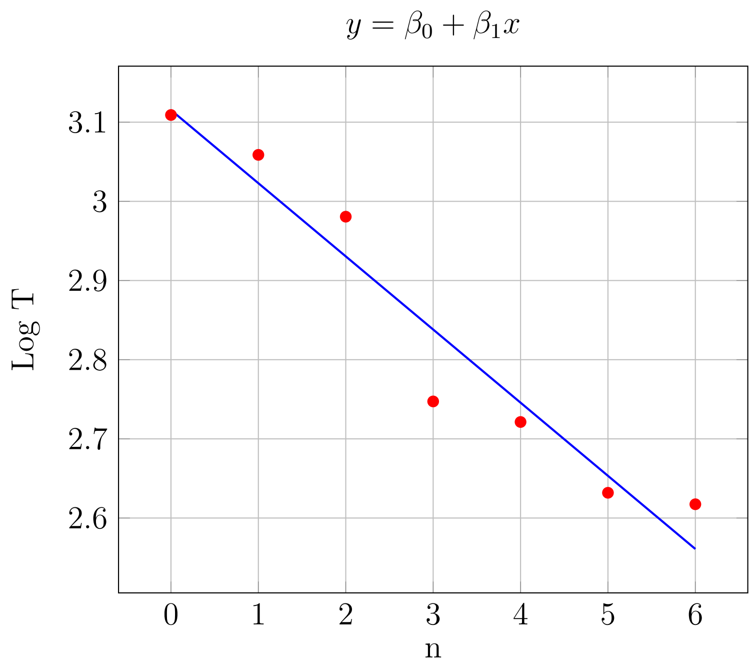


33 Transforming to linearity,

$$
\begin{aligned}
T &= ts^{-n} \\
\log T &= \log t - (n)\ \log s \qquad (33.1) \\
s &= e^{-\beta_1} = 1.0968 \\
t &= e^{\beta_0} = 22.54
\end{aligned}
$$

Performing the linear regression using `scipy.stats.linregress` outputs

linRegressOutput	
β_1	-0.092427
β_0	3.115315
r	-0.967522
p	0.000359

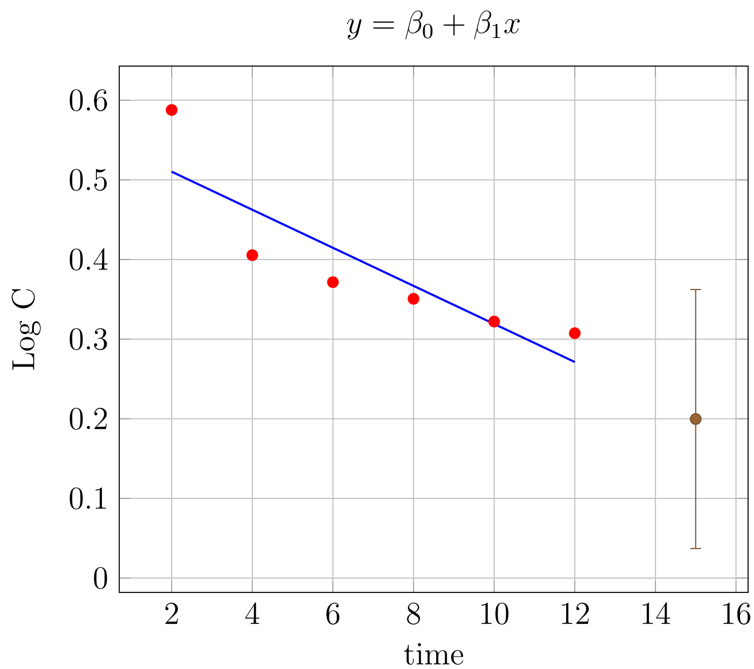


34 Transforming to linearity,

$$
\begin{aligned}
Y &= ae^{-bx} \\
\log Y &= \log a - bx \qquad (34.1) \\
b &= -\beta_1 = 0.0239 \\
a &= e^{\beta_0} = 1.7473
\end{aligned}
$$

The value corresponding to $x = 15$ is $\log Y = 0.1997$ and thus $Y = 1.22$

linRegressOutput	
β_1	-0.023894
β_0	0.558096
r	-0.871119
p	0.023845

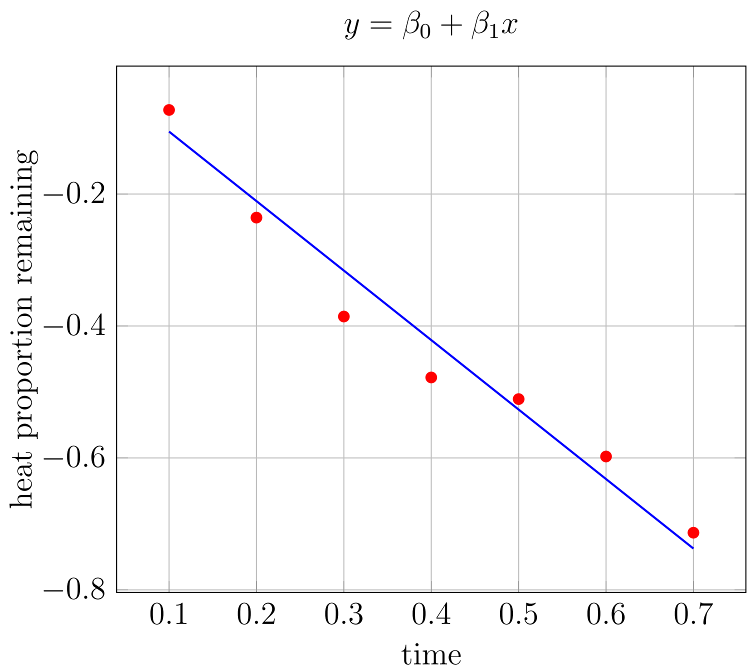


35 Transforming to linearity,

$$P = 1 - e^{-\alpha t}$$
$$\log(1 - P) = -\alpha t \tag{35.1}$$
$$\alpha = -\beta_1 = 1.0534$$

Using the forced pass through origin method from Problem 29,
The value corresponding to $P = 1/2$ is $t = 0.658$

zeroInterceptOutput	
β_1	-1.0534
β_0	0

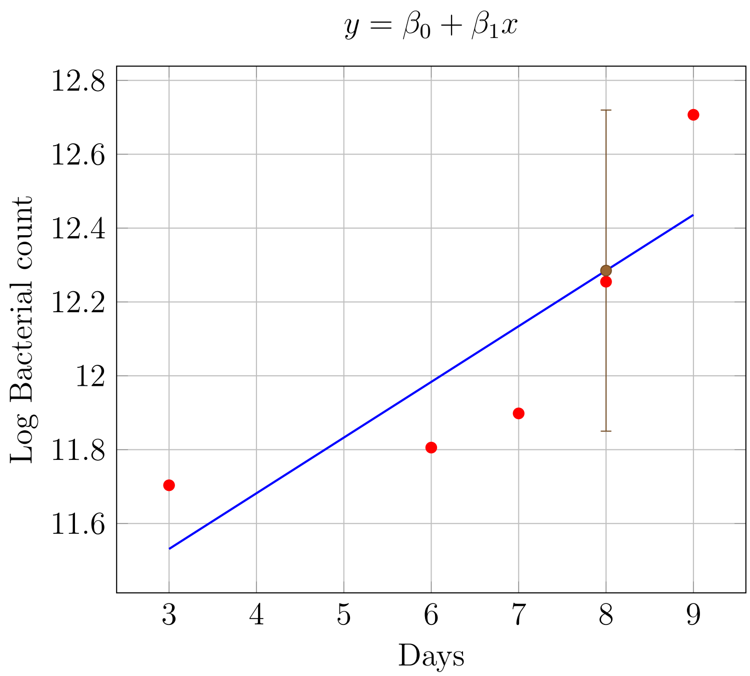


36 Transforming to linearity,

$$
\begin{aligned}
Y &= a\ e^{bx} \\
\log Y &= \log a + bx \\
b &= \beta_1 = 0.151 \\
a &= e^{\beta_0} = 64776
\end{aligned}
\tag{36.1}
$$

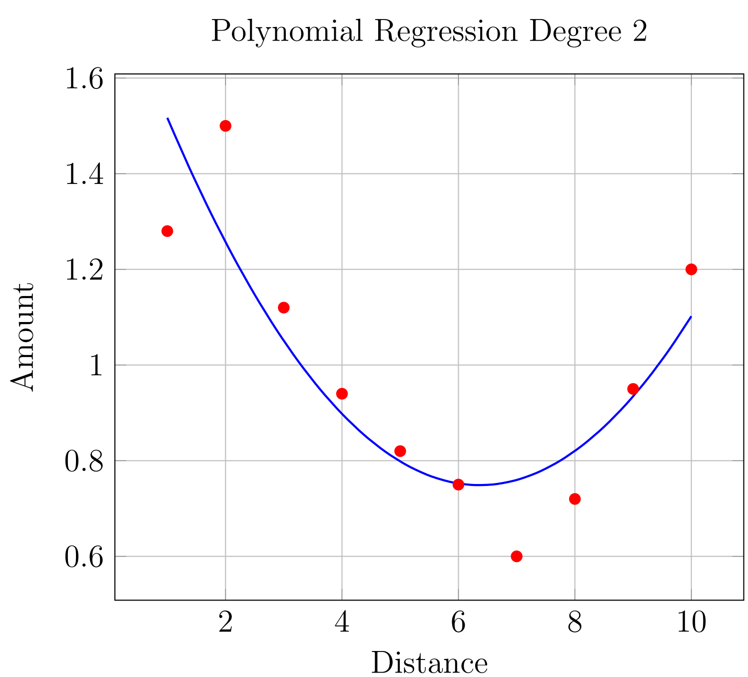
Performing the linear regression using `scipy.stats.linregress` outputs
The value corresponding to $x = 8$ is $y = 216403$

linRegressOutput	
β_1	0.150773
β_0	11.078703
r	0.845975
p	0.070863



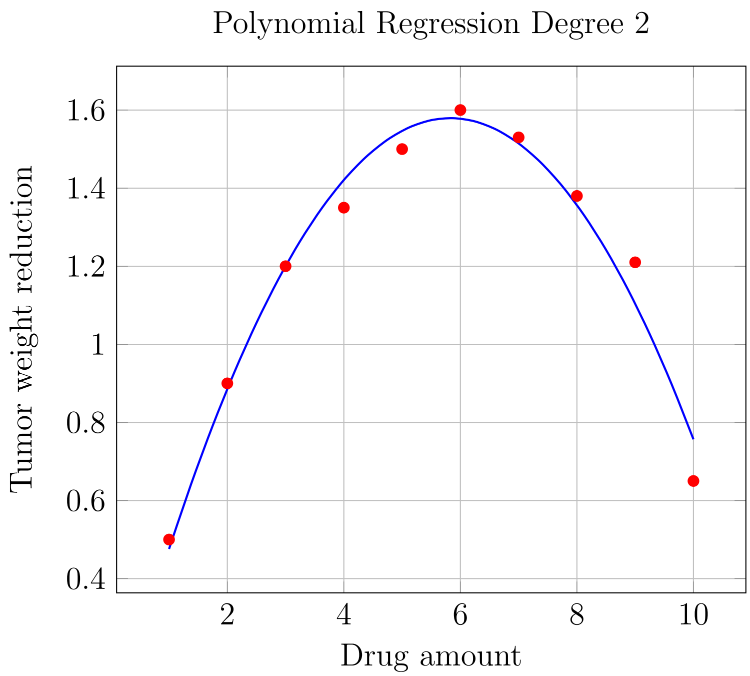
37 Performing the polynomial regression using custom script outputs

PolyRegressOutput	
β_0	1.8300
β_1	-0.3398
β_2	0.0267



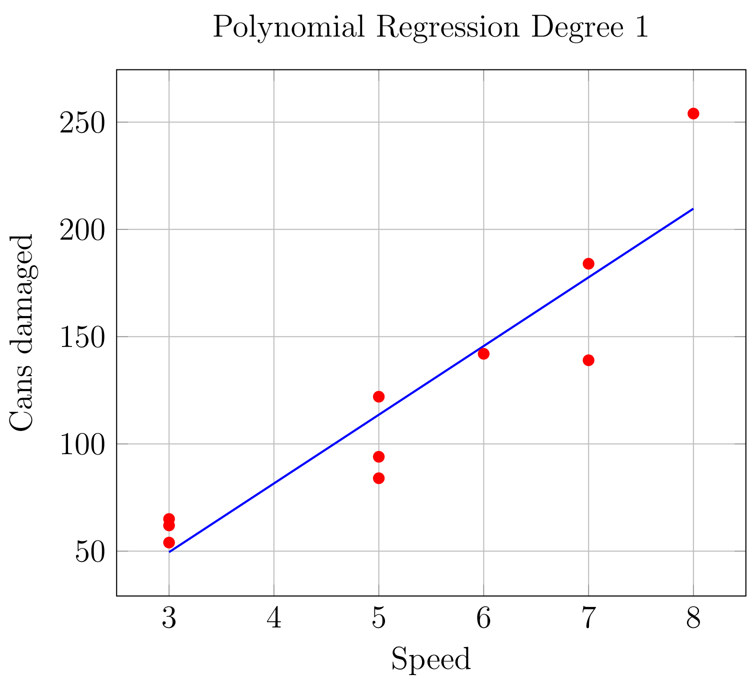
38 Performing the polynomial regression using custom script outputs

PolyRegressOutput	
β_0	-0.0283
β_1	0.5515
β_2	-0.0473

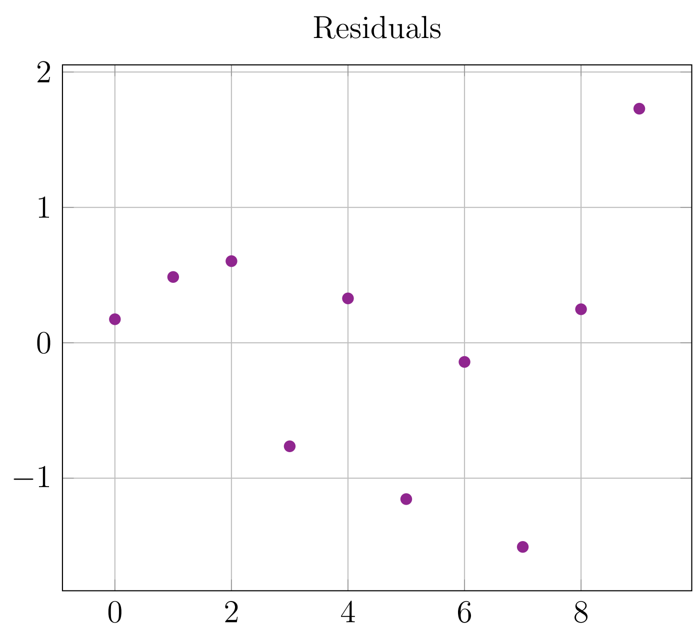


39 Starting with linear regression gives,

PolyRegressOutput	
β_0	-46.5405
β_1	32.0270

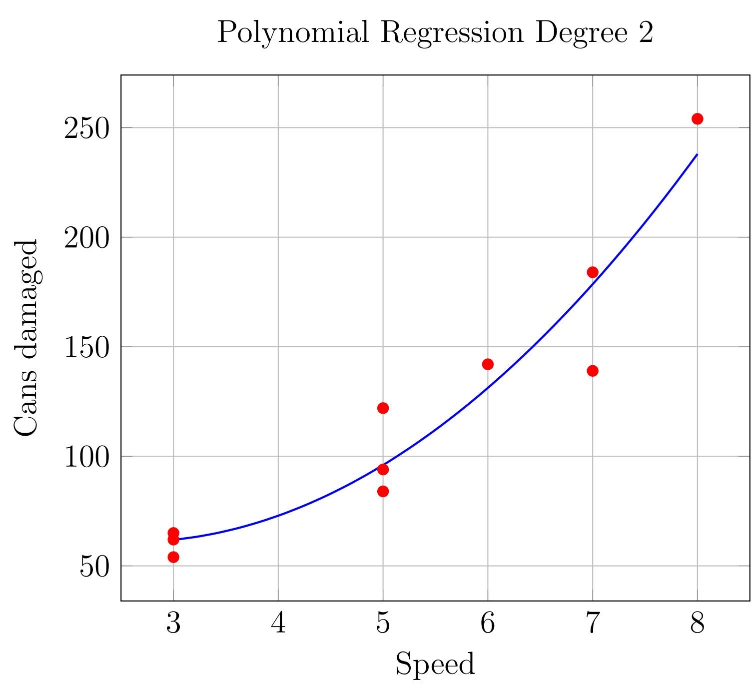


Plotting the standardized residuals



Trying out a degree 2 polynomial regression gives a better looking result.

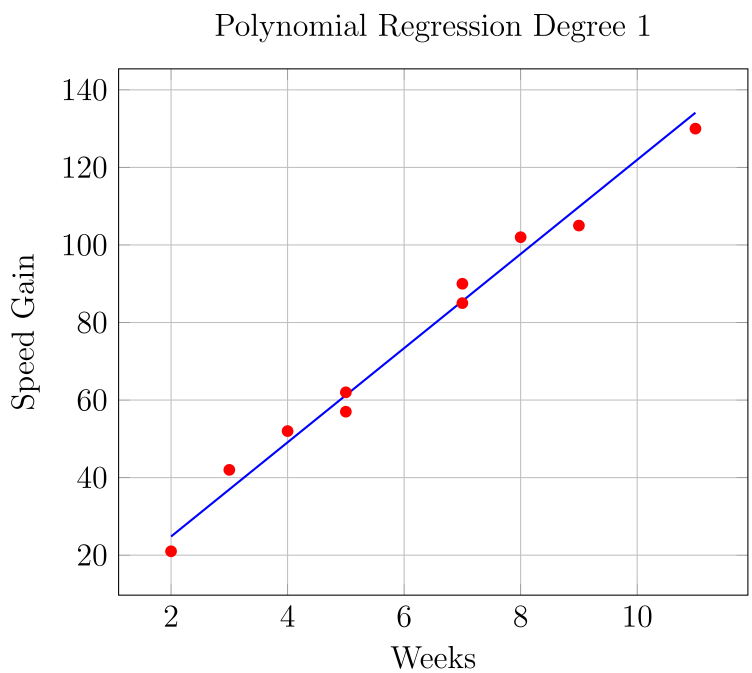
PolyRegressOutput	
β_0	101.4700
β_1	-31.3477
β_2	6.0513



40 Performing linear regression, but with variance in Y proportional to x,

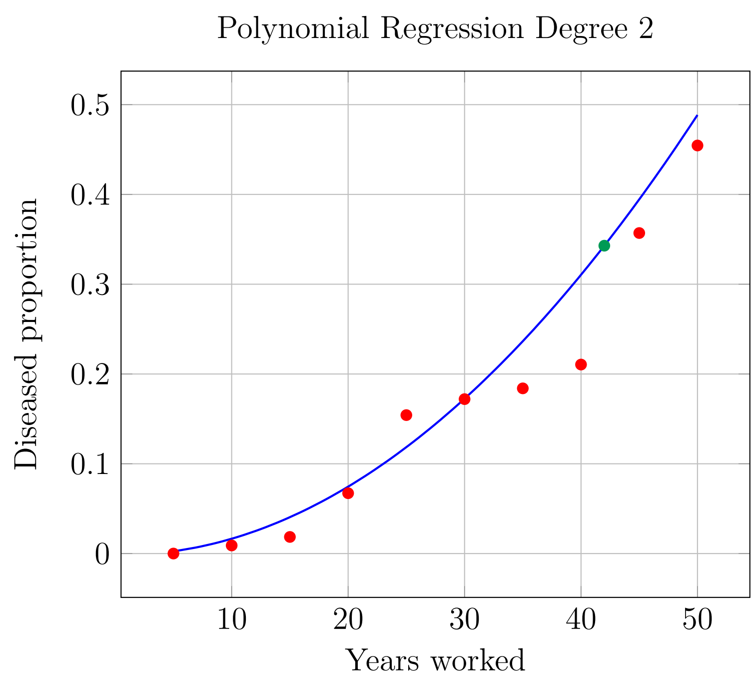
$$
\begin{aligned}
\mathrm{Var}(Y_i) &= \frac{\sigma^2}{w_i} \\
w_i &\propto \frac{1}{x_i} \qquad (40.1)
\end{aligned}
$$

PolyRegressOutput	
β_0	0.5250
β_1	12.1434



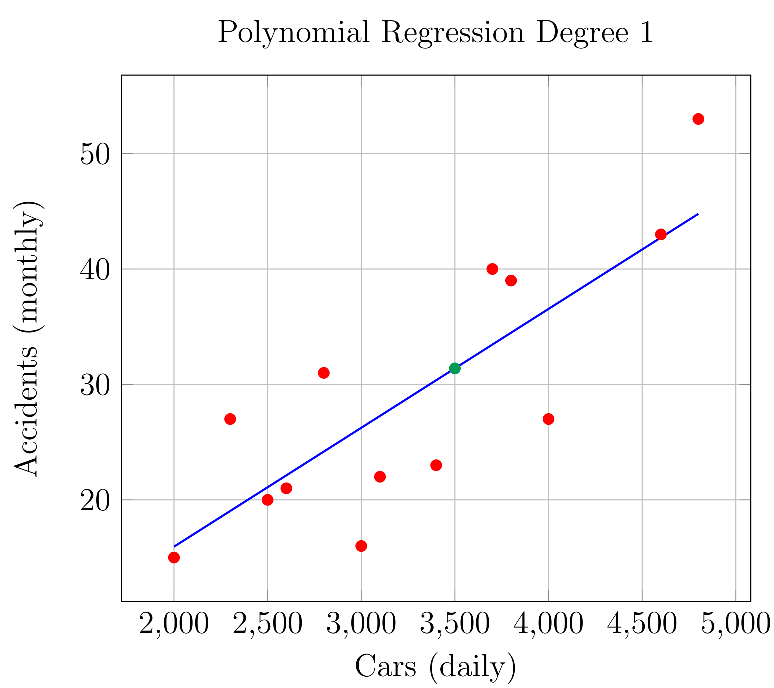
41 Performing the polynomial regression with degree 2 using custom script outputs The value corresponding to $x = 42$ is $y = 0.3429$.

PolyRegressOutput	
β_0	-0.0015
β_1	-0.0002
β_2	0.0002



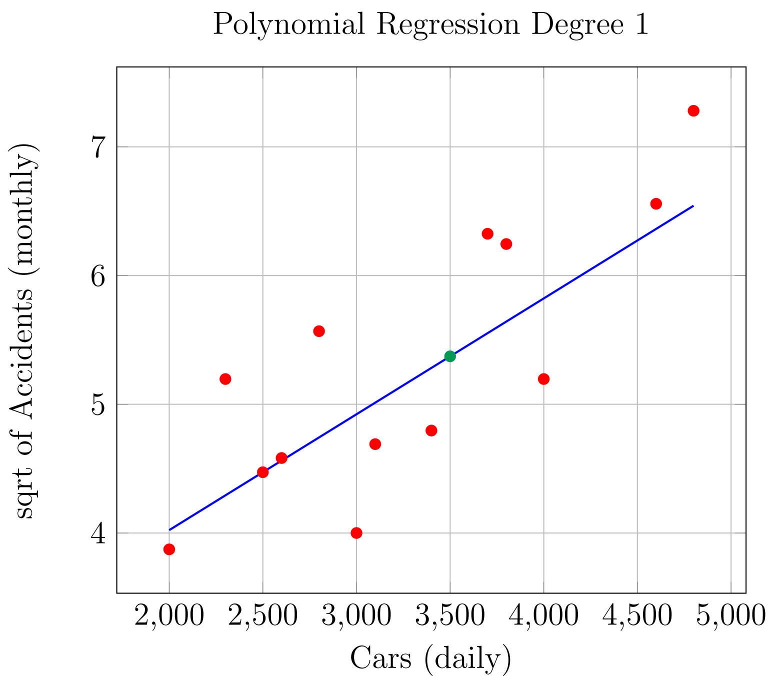
42 Performing linear regression, but with variance in Y proportional to x,
The value corresponding to $x = 3500$ is $y = 31.4$.

PolyRegressOutput	
β_0	-4.6584
β_1	0.0103



Redoing the problem with $\sqrt{Y} = \beta_0 + \beta_1 x + e$, The value corresponding to $x = 3500$ is $\sqrt{Y} = 5.3725$ which corresponds to $Y = 28.86$.

PolyRegressOutput	
β_0	2.2225
β_1	0.0009



43 Using custom multiple linear regression with 2 inputs,

MultiLinRegressOutput	
β_0	150.1414
β_1	0.3621
β_2	-3163.5648

44 Using custom multiple linear regression with 2 inputs,

MultiLinRegressOutput	
β_0	-1.5962
β_1	0.2662
β_2	0.0004

45 Using custom multiple linear regression with inputs 2, 3 and 4 (input 1 is ignored because of collinearity with input 2 as seen in the correlation matrix),

	x1	x2	x3	x4
x1	1.0000	1.0000	0.3725	0.9947
x2	1.0000	1.0000	0.3725	0.9947
x3	0.3725	0.3725	1.0000	0.3986
x4	0.9947	0.9947	0.3986	1.0000

With the collinear variables removed and the model dimensions reduced to $\{x_2, x_3, x_4\}$,

MultiLinRegressOutput	
β_0	18.6060
β_2	9.9179
β_3	14.0753
β_4	-19.1818

46 Using custom multiple linear regression with 2 inputs and log response as output,

MultiLinRegressOutput	
β_0	7.9567
β_1	-1.2047
β_2	-0.0225

The SS_R is 22.08, and using $n = 12, k = 2$, $\sigma^2 = 2.453$.

47 Using custom multiple linear regression with 3 inputs,

MultiLinRegressOutput	
β_0	-2.8278
β_1	5.3707
β_2	9.8157
β_3	0.4482

$SS_R = 201.97$ and the variance in response is $\sigma^2 = 18.361$
$H_0 : \beta_0 = 0$ has uses a t-test and has a p-value of 45.88%
$H_0 : \beta_3 = 0$ has uses a t-test and has a p-value of 36.72%
$H_0 : 8.5 = \beta_0 + \beta_1 + \beta_2 + \beta_3$ uses a t-test and has a p-value of 36.72%

$$\frac{\sum_{i=0}^{k} B_i x_i - \sum_{i=0}^{k} \beta_i x_i}{\sqrt{\mathbf{x}^\intercal \left(\mathbf{X}^\intercal \mathbf{X}\right)^{-1} \mathbf{x} \left(\frac{SS_R}{n-k-1}\right)}} \sim t_{n-k-1} \tag{47.1}$$

48 Performing multiple regression using custom script with 2 inputs

`MultiLinRegressOutput`	
β_0	177.6954
β_1	1.0351
β_2	10.7214

The 90% confidence interval for $\{x_1 = 21, x_2 = 3.6\}$ is
$238.03 \pm 3.961 = [234.069, 241.991]$

49 Using custom multiple linear regression with 3 inputs,
$SS_R = 4614.65$ and $\sigma^2 \approx 769.11$

`MultiLinRegressOutput`	
β_0	1108.7245
β_1	8.6393
β_2	0.2608
β_3	-0.7114

The 95% confidence interval for $\{x_1 = 125, x_2 = 900, x_3 = 160\}$ is
$2309.51 \pm 37.85 = [2271.66, 2347.36]$

50 Same mean and larger variance of the prediction interval means that it is always going to larger than the mean response interval for the same confidence value.

51 Using custom multiple linear regression with 2 inputs,
$SS_R = 619.43$ and $\sigma^2 \approx 68.825$

`MultiLinRegressOutput`	
β_0	5.2387
β_1	5.6968
β_2	9.5501

The 95% confidence interval for the predicted response at $\{x_1 = 10.2, x_2 = 17\}$ is
$225.7 \pm 20.05 = [205.65, 245.75]$

52 Using custom multiple linear regression with 2 inputs,
$SS_R = 0.5793$ and $\sigma^2 \approx 0.0644$
The hypothesis $H_0 : \beta_1 = 0$ has a p-value 0.45% using a t-test. The hypothesis can be safely rejected.

`MultiLinRegressOutput`	
β_0	6.1439
β_1	-0.0376
β_2	0.0850

The 95% confidence interval for the predicted response at $\{x_1 = 85, x_2 = 20\}$ is
$4.645 \pm 0.613 = [4.032, 5.259]$

53 Using custom multiple linear regression with 2 inputs,
The hypothesis $H_0 : \beta_1 = 0$ has a p-value 81% using a t-test. The hypothesis cannot be rejected.

`MultiLinRegressOutput`	
β_0	28.2098
β_1	0.1164
β_2	0.5664

The 95% confidence interval for the mean response at $\{x_1 = 45, x_2 = 180\}$ is
$135.41 \pm 17.18 = [118.22, 152.59]$
The 95% confidence prediction interval for the response at $\{x_1 = 45, x_2 = 180\}$ is
$135.41 \pm 51.28 = [84.12, 186.69]$

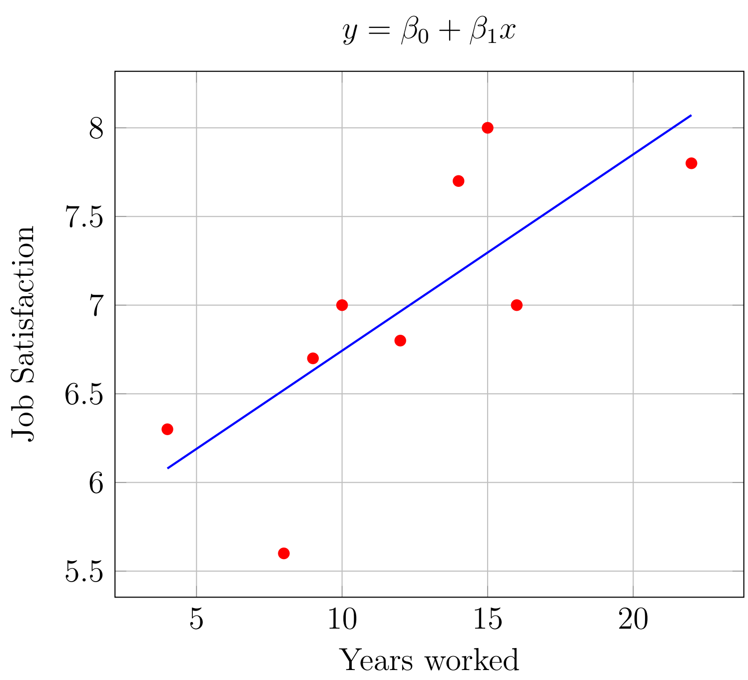
54 Using custom multiple linear regression with 2 inputs,
Since $\beta_2 < 0$, years on the job increasing lead to lesser job satisfaction.

`MultiLinRegressOutput`	
β_0	-1.2050
β_1	0.1619
β_2	-0.1128

The 95% confidence prediction interval for the response at $\{x_1 = 56, x_2 = 5\}$ is $7.3 \pm 1.397 = [5.9, 8.7]$

55 Performing the linear regression using `scipy.stats.linregress` outputs

`linRegressOutput`	
β_1	0.110682
β_0	5.636108
r	0.759205
p	0.017660



Since $\beta_1 > 0$, years on the job increasing lead to greater job satisfaction.

The trend is reversed compared to the relationship between the same input and response in the presence of the other input variable.

The reversal in relationship between x_2 and y upon the introduction of x_1 happens because of significant multicollinearity, as seen in the correlation matrix.

	x1	x2
x1	1.0000	0.7592
x2	0.7592	1.0000

56 To find the value of x corresponding to $p(x) = 0.5$

$$
\begin{aligned}
p(x) &= \frac{\exp(a+bx)}{1+\exp(a+bx)} = 0.5 \\
\frac{y}{1+y} &= \frac{1}{2} \\
y &= 1 = \exp(a+bx) \\
x^* &= -\frac{a}{b}
\end{aligned} \tag{56.1}
$$

57 How to do logistic regression

58 How to do logistic regression