Unit 2: Review of Probability

Statistics 537: Statistics for Research
Ramón V. León
Poisson Distribution:

\[ f(x) = P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad \text{for } x = 0, 1, 2, \ldots \]

\[ E(X) = \lambda, \quad Var(X) = \lambda \]
Poisson Distribution:

\[ f(x) = P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad \text{for } x = 0, 1, 2, \ldots \]

\[ E(X) = \lambda, \quad Var(X) = \lambda \]

**Example:** On the average five Prussian soldiers die from horse kicks in a year. What is the probability that exactly four soldiers are killed this way in a given year?

\[ P(X = 4) = \frac{e^{-5} (5)^4}{4!} = .175 \]
Geometric Distribution

Probability of waiting time to an event in independent trials

\[ P(X = x) = (1 - p)^{x-1} p, \quad x = 1, 2, ... \]

\[ E(X) = \frac{1}{p} \quad \text{and} \quad Var(X) = \frac{1-p}{p^2} \]
Geometric Distribution

Probability of waiting time to an event in independent trials

\[ P(X = x) = (1 - p)^{x-1} p, \quad x = 1, 2, \ldots \]

\[ E(X) = \frac{1}{p} \quad \text{and} \quad Var(X) = \frac{1-p}{p^2} \]

Example: Suppose the probability of winning the jackpot in a slot machine is .01. What is the expected number of tries to win the jackpot?

What is the probability that you hit the jackpot for the first time on your fifth try?

\[ E(X) = \frac{1}{.01} = 100, \quad P(X = 5) = (.99)^4 (.01) = .0096 \]
Uniform Distribution

Distribution when all values in an interval are equally likely

Suppose that you select a real number at random in the interval [1,5]. What is the probability that it turns out to be between 2 and 4?

\[ P(2 \leq X \leq 4) = \frac{4 - 2}{5 - 1} = 0.5 \]

Proportion of the lengths of the intervals [2,4] to the length of the interval [1,5]
Exponential Distribution

Distribution of waiting time when arrivals occur at random

\[ f(x) = \lambda e^{-\lambda x}, \quad F(x) = \int_0^x \lambda e^{-\lambda t} \, dt = 1 - e^{-\lambda x} \quad \text{for } x \geq 0 \]

\[ E(X) = \frac{1}{\lambda} \quad \text{and} \quad Var(X) = \frac{1}{\lambda^2} \]
Exponential Distribution

Distribution of waiting time when arrivals occur at random

\[ f(x) = \lambda e^{-\lambda x}, \quad F(x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x} \quad \text{for } x \geq 0 \]

\[ E(X) = \frac{1}{\lambda} \quad \text{and} \quad Var(X) = \frac{1}{\lambda^2} \]

**Example 2.35**

Suppose that airport shuttle buses arrive at a terminal at the rate of one every 10 minutes with exponentially distributed interarrival times. If a person arrives at the bus stop and sees a bus leaving, what is the probability that he must wait for more than 10 minutes for the next bus? What if the person does not see a bus leaving?

Let \( X \) be the time between arrivals of buses; \( X \sim \text{Exp}(\lambda = 1/10) \). The first probability is

\[ P(X > 10) = e^{-10/10} = 0.368. \]

One might think that the second probability should be smaller, since the person arrived after the previous bus had left and so there is a smaller chance that he will have to wait for more than 10 minutes. However, because of the memoryless property of the exponential distribution this probability is also 0.368.

\[ P(X > x) = 1 - P(X \leq x) = 1 - F(x) = e^{-\lambda x} \]
Memoryless Property of the Exponential Distribution

\[ P(X > s + t \mid X > s) = P(X > t) \]
\[ P(X > 5 + 3 \mid X > 5) = P(X > 3) \]

The probability of having to wait \( t \) additional minutes after having waited \( s \) minutes is the same as the probability of having to wait \( t \) minutes to begin with.
Normal Distribution
A continuous r.v. $X$ has a normal distribution with parameter $\mu$ and $\sigma^2$ if its p.d.f. is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ for } -\infty < x < \infty$$

$$E(X) = \mu \quad \text{and} \quad Var(X) = \sigma^2$$

Notation:

$$X \sim N(\mu, \sigma^2)$$
Normal Distribution

Mean: 0, Sd: 1

- Above: 1
- Below: 1
- Between: -1 and 1
- Outside: -1 or 1

Shaded area: 0.682689

Normal Distribution

Mean: 5, Sd: 2

- Above: 1
- Below: 1
- Between: 3 and 7
- Outside: -1 or 1

Shaded area: 0.682689
Empirical Rule
Mean of i.i.d. Normal Random Variable

Let $X_1, X_2, ..., X_n$ be independent, indentically distributed $N(\mu, \sigma^2)$. We say that $X_1, X_2, ..., X_n$ is a random sample from a $N(\mu, \sigma^2)$ population.

Then for $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ we have $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

Hint: Use this result to do homework problem 2.83
Percentiles of the Normal Distribution

Suppose that the scores on a standardized test are normally distributed with mean 500 and standard deviation 100. What is the 75\(^{th}\) percentile score of this test?

For \( x = 75^{th} \) percentile means that \( P(X \leq x) = .75 \). So
Percentiles of the Normal Distribution

Suppose that the scores on a standardized test are normally distributed with mean 500 and standard deviation 100. What is the 75th percentile score of this test?

For $x = 75^{th}$ percentile means that $P(X \leq x) = .75$. So
Percentiles of the Normal Distribution

Suppose that the scores on a standardized test are normally distributed with mean 500 and standard deviation 100. What is the 75th percentile score of this test?

For $x = 75^{th}$ percentile means that $P(X \leq x) = .75$. So
Percentiles of the Normal Distribution

Suppose that the scores on a standardized test are normally distributed with mean 500 and standard deviation 100. What is the 75th percentile score of this test?

For $x = 75^{th}$ percentile means that $P(X \leq x) = 0.75$. So

The score is 567
Introductory Remarks

• Regression analysis is a method for studying the relationship between two or more **numerical** variables
• In regression analysis one of the variables is regarded as a **response**, **outcome** or **dependent variable**
• The other variables are regarded as **predictor**, **explanatory** or **independent** variables
• An empirical model approximately captures the main features of the relationship between the response variable and the key predictor variables
• Sometimes it is not clear which of two variable should be the response. In this case **correlation analysis** is used.
• In this unit we consider only two variables
A Probabilistic Model for Simple Linear Regression

Note that the $Y_i$ are independent $N(\mu_i = \beta_0 + \beta_1 x_i, \sigma^2)$ r.v.'s

![Figure 10.1 Simple Linear Regression Model]

Constant variability about the regression line

Notice that the predictor variable is regarded as nonrandom because it is assumed to be set by the investigator
A Probabilistic Model for Simple Linear Regression

Let $x_1, x_2, ..., x_n$ be specific settings of the predictor variable.
Let $y_1, y_2, ..., y_n$ be the corresponding values of the response variable.
Assume that $y_i$ is the observed value of a random variable (r.v.) $Y_i$, which depends on $x_i$ according to the following model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, 2, ..., n).$$

Here $\varepsilon_i$ is the random error with $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$.

Thus $E(Y_i) = \mu_i = \beta_0 + \beta_1 x_i$ (true regression line).

We assume that $\varepsilon_i$ are independent, identically distributed (i.i.d.) r.v.'s.
We also usually assume that the $\varepsilon_i$ are normally distributed.
Reflections on the Method of Tire Data Collection: Experimental Design

<table>
<thead>
<tr>
<th>Mileage Depth</th>
<th>Depth in mils</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>394.33</td>
</tr>
<tr>
<td>4</td>
<td>329.5</td>
</tr>
<tr>
<td>8</td>
<td>291</td>
</tr>
<tr>
<td>12</td>
<td>255.17</td>
</tr>
<tr>
<td>16</td>
<td>229.33</td>
</tr>
<tr>
<td>20</td>
<td>204.83</td>
</tr>
<tr>
<td>24</td>
<td>179</td>
</tr>
<tr>
<td>28</td>
<td>163.83</td>
</tr>
<tr>
<td>32</td>
<td>150.33</td>
</tr>
</tbody>
</table>

- Was one tire or nine used in the experiment? (Read the book!)
- What effect would this answer have
  - on the experimental error?
  - on our ability to generalize to all tires of this brand?
- Was the data collected on one or several cars? Does it matter?
- Was the data collected in random order or in the order given? Does it matter?
- Is there a possible confounding problem?

• Mileage in 1000 miles
• Depth in mils
Example 10.1: Tread Wear vs. Mileage

Using the **Fit Y by X** platform in the Analyze menu.
Scatter Plot: Tread Wear vs. Mileage
Least Squares Line Fit

\[ y = \hat{\beta}_0 + \hat{\beta}_1 x \]
Illustration: Least Squares Line Fit

The least squares line fit is the line that minimizes the sum of the squares of the lengths of the vertical segments.

Mathematics of Least Squares

Find the line, i.e., values of $\beta_0$ and $\beta_1$ that minimizes the sum of the squared deviations:

$$Q = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

How?

Solve for values of $\beta_0$ and $\beta_1$ for which

$$\frac{\partial Q}{\partial \beta_0} = 0 \quad \text{and} \quad \frac{\partial Q}{\partial \beta_1} = 0$$
How Least Squares Can Be Misleading
Predict Mean Grove Depth for a Given Mileage

For $x = 25$ (25,000 miles) the predicted $y$:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 (25) = 360.64 - 7.281 \times 25 = 178.62\text{ mils.}$$

Add a row to the table and leave the $y$ value blank.
Goodness of Fit of the LS Line

Fitted values of $y_i$: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \ i = 1, 2, ..., n.$

Residuals: $e_i = y_i - \hat{y}_i = y_i - \left( \hat{\beta}_0 + \hat{\beta}_1 x_i \right) = \text{segment length}$
Sums of Squares

Error Sum of Squares (SSE) = \( \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \)

Total Sum of Squares (SST) = \( \sum_{i=1}^{n} (y_i - \bar{y})^2 \)

Regression Sum of Squares (SSR) = \( \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \)

\[ \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]

\( SST = SSR + SSE \)

These definitions apply even if the model is more complex than linear in x, for example, if it is quadratic in x.

SST is interpreted as the total variation in y.

JMP calls SSR the Model Sums of Square.
Illustration of Sums of Squares
Coefficient of Determination

\[ SST = SSR + SSE \]

\[ r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \]

proportion of the variation in \( y \) that is accounted for by regression on \( x \)
Estimation of $\sigma^2$

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} \left( y_i - \left[ \hat{\beta}_0 + \hat{\beta}_1 x_i \right] \right)^2}{n-2}$$

This estimate has $n-2$ degrees of freedom because two unknown parameters are estimated.
Statistical Inference for $\beta_0$ and $\beta_1$

$$\frac{\hat{\beta}_0 - \beta_0}{SD(\hat{\beta}_0)} \sim N(0,1)$$ and $$\frac{\hat{\beta}_1 - \beta_1}{SD(\hat{\beta}_1)} \sim N(0,1)$$

$$\frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} \sim t_{n-2}$$ and $$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

100(1 - $\alpha$)% CI:

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2} SE(\hat{\beta}_1) = -7.28 \pm t_{7,0.025} \times 0.614 = -7.28 \pm 2.365 \times 0.614 = [-8.733, -5.829]$$

$$\hat{\beta}_0 \pm t_{n-2,\alpha/2} SE(\hat{\beta}_0) = 360.63 \pm t_{7,0.025} \times 11.69$$

1. Right click on the Parameter Estimate area
2. Select
Test of Hypothesis for $\beta_0$ and $\beta_1$

\[ H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0 \]

Reject $H_0$ at $\alpha$ level if

\[ |t| = \left| \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \right| = \left| \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \right| > t_{n-2,\alpha/2} \]

Tire example:

\[ t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{-7.28}{.614} = -11.86, \quad t_{7,.025} = 2.365 \]

Strong evidence that mileage affects tread depth. (Two-sided p-value)
Analysis of Variance

\[ H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_0 : \beta_1 \neq 0 \]

\[ F = \frac{MSR}{MSE} \]

The Mean Square (MS) is the Sums of Square divided by its degrees of freedom, e.g.

\[ \text{MSE} = \frac{\text{SSE}}{\text{df}} = \frac{2531.529}{7} = 361.6 \]

\[ \frac{50887.2}{361.6} = 140.71 \]

\[ F = 140.70 = (-11.8621)^2 = t^2 \]

\( F \) and \( t \) relationship holds only when the \( F \) numerator has one degree of freedom
Prediction of a Future $y$ or Its Mean

For a fixed value of $x^*$, are we trying to predict
- the average value of $y$?
- the value of a future observation of $y$?

Example:

- Do I want to predict the average selling price of all 4,000 square feet houses in my neighborhood.
- Or do I want to predict the particular future selling price of my 4,000 square feet house?

Which prediction is subject to the most error?
Prediction of a Future $y$ or Its Mean: Prediction Interval or Confidence Interval

For a fixed value of $x^*$, are we trying to predict
- the average value of $y$?
- the value of a future observation of $y$?
JMP: Prediction of the Mean of \( y \)
JMP: Prediction of a Future Value of $y$
Formulas for Confidence and Prediction Intervals

A 100\((1 - \alpha)\)% CI for \(\mu^*\) is given by

\[
\hat{\mu}^* - t_{n-2,\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} \leq \mu^* \leq \hat{\mu}^* + t_{n-2,\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}
\]  

(10.18)

where \(\hat{\mu}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*\) and \(s = \sqrt{\text{MSE}}\) is the estimate of \(\sigma\).

A 100\((1 - \alpha)\)% PI for \(Y^*\) is given by

\[
\hat{Y}^* - t_{n-2,\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} \leq Y^* \leq \hat{Y}^* + t_{n-2,\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}
\]  

(10.19)

where \(\hat{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*\).

Prediction Interval of Chapter 7:

100\((1 - \alpha)\)% PI for a future observation \(X \sim N(\mu, \sigma^2)\) is given by

\[
\bar{x} - t_{n-1,\alpha/2} s \sqrt{1 + \frac{1}{n}} \leq X \leq \bar{x} + t_{n-1,\alpha/2} s \sqrt{1 + \frac{1}{n}}
\]
Confidence and Prediction Intervals with JMP

Using the **Fit Model** Platform not the **Fit Y by X** platform
## CI for Mean for μ*

![Image of a computer interface with columns and formulas]

<table>
<thead>
<tr>
<th>Mileage (in 1000 Miles)</th>
<th>Grove Depth (in mils)</th>
<th>Predicted Grove Depth (in mils)</th>
<th>Lower 95% Mean Grove Depth (in mils)</th>
<th>Upper 95% Mean Grove Depth (in mils)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>394.33</td>
<td>332.997636</td>
<td>388.275698</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>329.5</td>
<td>308.535899</td>
<td>354.492434</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>291</td>
<td>283.431448</td>
<td>321.351886</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>255.17</td>
<td>257.194859</td>
<td>289.343474</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>229.33</td>
<td>229.157298</td>
<td>259.136036</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>204.83</td>
<td>198.949859</td>
<td>231.098474</td>
</tr>
<tr>
<td>7</td>
<td>24</td>
<td>179</td>
<td>166.941448</td>
<td>204.861886</td>
</tr>
<tr>
<td>8</td>
<td>28</td>
<td>163.83</td>
<td>133.800899</td>
<td>179.757434</td>
</tr>
<tr>
<td>9</td>
<td>32</td>
<td>150.33</td>
<td>100.017636</td>
<td>155.295698</td>
</tr>
<tr>
<td>10</td>
<td>25</td>
<td>*</td>
<td>178.621042</td>
<td>198.503156</td>
</tr>
</tbody>
</table>

---

6/18/2012 Unit 10 - Stat 571 - Ramón V. León 44
**Prediction Interval for \( Y^* \)**

<table>
<thead>
<tr>
<th>Mileage (in 1000 Miles)</th>
<th>Grove Depth (in mils)</th>
<th>Predicted Grove Depth (in mils)</th>
<th>Lower 95% Indiv Grove Depth (in mils)</th>
<th>Upper 95% Indiv Grove Depth (in mils)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>394.33</td>
<td>307.853651</td>
<td>413.419682</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>329.5</td>
<td>281.01534</td>
<td>382.012993</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>291</td>
<td>253.589817</td>
<td>351.193517</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>255.17</td>
<td>225.514439</td>
<td>321.023895</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>229.33</td>
<td>196.74612</td>
<td>291.547214</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>204.83</td>
<td>167.269439</td>
<td>262.778895</td>
</tr>
<tr>
<td>7</td>
<td>24</td>
<td>179</td>
<td>137.099817</td>
<td>234.703517</td>
</tr>
<tr>
<td>8</td>
<td>28</td>
<td>163.83</td>
<td>106.28034</td>
<td>207.277993</td>
</tr>
<tr>
<td>9</td>
<td>32</td>
<td>150.33</td>
<td>74.8736513</td>
<td>180.439682</td>
</tr>
<tr>
<td>10</td>
<td>25</td>
<td>*178.621042</td>
<td>*129.453683</td>
<td>*227.7884</td>
</tr>
</tbody>
</table>
Prediction for the Mean of Y or a Future Observation of Y

• Point estimate prediction is the same in both cases: 178.62

• But the error bands are different
  – Narrower for the mean of Y: [158.73, 198.51]
  – Wider for a future value of Y: [129.44, 227.80]
Calibration (Inverse Regression)

We have looked at the problem of estimating $\mu^*$ for given $x^*$. The opposite problem is to estimate $x^*$ for given $\mu^*$. This is called the calibration or inverse regression problem.

The formula for an estimate of $x^*$ is obtained by solving for $x^*$ from $\mu^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$, which yields

$$\hat{x}^* = \frac{\mu^* - \hat{\beta}_0}{\hat{\beta}_1}.$$

**Example 10.9 (Tire Tread Wear vs. Mileage: Inverse Regression)**

Estimate the mean life of a tire at wearout (62.5 mils remaining) using the results from Example 10.2.

We want to estimate $x^*$ when $\mu^* = 62.5$. Substituting $\hat{\beta}_0 = 360.64$ and $\hat{\beta}_1 = -7.281$ in the above formula we obtain

$$\hat{x}^* = \frac{62.5 - 360.64}{-7.281} \times 1000 = 40,947.67 \text{ miles}.$$

Caution should be exercised in using this projection, since we do not have any data past 32,000 miles.
Inverse Regression in JMP 5: Confidence Interval

In Fit Model Platform

Confidence Interval with respect to an expected response

Unchecked
Inverse Regression: Two Types of Confidence Intervals

Calibration is usually used in measurement. Given that you got a certain measurement what is the “true” value of the measured quantity? Which inverse prediction interval makes more sense?
Measurement Application

Suppose you get a measurement of 6.50 on an object. What is a calibration interval for its true value?

Data was supposedly obtained by measuring standards of known “true” measurement with a calibrated instrument.

How the data was really generated

Confidence Interval with respect to an individual response
Residuals

The residuals are the differences between the observed value of $y$ and the predicted value of $y$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 (32) = 360.64 - 7.281 \times 32 = 127.65 \text{ mils}$$
Regression Diagnostics

The residuals $e_i = y_i - \hat{y}_i$ can be viewed as the "left over" after the model is fitted. If the model is correct, then the $e_i$ can be viewed as the "estimates" of the random errors $\epsilon_i$'s. As such, the residuals are vital for checking the model assumptions.

Residual plots

If the model is correct the residuals should have no structure, that is, they should look random on their plots.
Mathematics of Residuals

If the assumed regression model is correct, then the $e_i$'s are normally distributed with $E(e_i) = 0$ and

$$Var(e_i) = \sigma^2 \left[ 1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}} \right] \approx \sigma^2 \text{ if } n \text{ is large.}$$

The $e_i$'s are not independent (even though the $\varepsilon_i$'s are independent) because they satisfy the following constraints:

$$\sum_{i=1}^{n} e_i = 0, \quad \sum_{i=1}^{n} x_i e_i = 0.$$ 

However, the dependence introduced by these constraints is negligible in $n$ is modestly large, say greater than 20 or so.

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$
Basic Principle Underlying Residual Plots

If the assumed model is correct then the residuals should be randomly scattered around 0 and should show no obvious systematic pattern.
Residuals in JMP
Checking for Linearity

Should

\[ E(Y) = \beta_0 + \beta_1 x + \beta_2 x^2 \]

be fitted rather than

\[ E(Y) = \beta_0 + \beta_1 x ? \]
Predicted by Residual Plot

What are the advantages of this plot?
Tread Wear Quadratic Model
Residuals for Quadratic Model

Residuals have a random pattern

The Fit Y by X platform was used to get these plots

The Fit Y by X platform was used to get these plots

Left plot can also be obtained using the Plot Residuals option

<table>
<thead>
<tr>
<th>Mileage (in 1000 Miles)</th>
<th>Grove Depth (in mls)</th>
<th>Predicted Grove Depth (in mls)</th>
<th>Residuals Grove Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>394.33</td>
<td>386.264848</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>329.5</td>
<td>337.921212</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>291</td>
<td>295.069329</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>255.17</td>
<td>257.709199</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>229.33</td>
<td>225.840823</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>204.83</td>
<td>199.464199</td>
</tr>
<tr>
<td>7</td>
<td>24</td>
<td>179</td>
<td>178.579329</td>
</tr>
<tr>
<td>8</td>
<td>28</td>
<td>163.83</td>
<td>163.186212</td>
</tr>
<tr>
<td>9</td>
<td>32</td>
<td>150.33</td>
<td>153.284848</td>
</tr>
</tbody>
</table>
Tread Wear Quadratic Model

\[ \text{Grove Depth (in mils)} = 342.33082 - 7.280625 \text{ Mileage (in 1000 Miles)} + 0.1716173 (\text{Mileage (in 1000 Miles-16)})^2 \]

**Summary of Fit**
- \( R^2 = 0.996083 \)
- \( R^2 \text{ Adj} = 0.994777 \)
- Root Mean Square Error: 5.055636
- Mean of Response: 244.1467
- Observations (or Sum Wgts): 9

**Analysis of Variance**
- Source: Model, DF: 2, Sum of Squares: 53209.471, Mean Square: 26604.7, F Ratio: 762.8263
- Source: Error, DF: 6, Sum of Squares: 209.259, Mean Square: 34.9, F Ratio: 762.8263
- C. Total, DF: 8, Sum of Squares: 53418.730, Mean Square: 1662.704
- Prob > F: <.0001

**Parameter Estimates**
- **Intercept**: 342.33082, Std Error: 4.267115, t Ratio: 80.23, Prob>|t|: <.0001
- **Mileage (in 1000 Miles)**: -7.280625, Std Error: 0.190604, t Ratio: -38.20, Prob>|t|: <.0001
- **(Mileage (in 1000 Miles-16))^2**: 0.1716173, Std Error: 0.021032, t Ratio: 8.16, Prob>|t|: 0.0002

\( R^2 \) was .95261 for model linear in x
Checking for Constant Variance: Plot of Predicted vs. Residual

If assumption is incorrect, often $Var(Y)$ is some function of $E(Y) = \mu$

Figure 10.7 Plots of Residuals $e_i$ vs. $\hat{y}_i$ Corresponding to Different Functional Relationships Between $Var(Y)$ and $E(Y)$: (a) Constant variance; (b) Variance proportional to $\mu^2$; (c) Variance proportional to $\mu$
Other Model Diagnostics Based on Residuals

• **Checking for normality of errors:** Do a normal plot of the residuals

• **Warning:** Don’t plot the response $y$ on a normal plot
  - This plot has no meaning when one has regressors
  - Don’t transform the data on the basis of this plot
  - Many students make this mistake in their project. Don’t be one of them.
Other Model Diagnostics Based on Residuals

- **Checking for independence of errors:** Plot the residual in time order. The order of data collection should be always recorded. Use Durbin-Watson statistic to test for autocorrelation.
Other Model Diagnostics Based on Residuals

• **Checking for outliers:** See if any standardized (Studentized) residual exceeds 2 (standard deviations) in absolute value.

\[
e_i^* = \frac{e_i}{SE(e_i)} = \frac{e_i}{s \sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}}} \approx \frac{e_i}{s}, \quad i = 1, 2, \ldots, n.
\]

Also do a box plot of residuals to check for outliers.
Standardized (Studentized) Residuals in JMP 4

Using the Fit Model JMP platform with linear model

Should we omit the first observation and refit the model?

Possible outlier

The model fitted here is the one linear in x, not the quadratic model
Checking for Influential Observations

An observation can be influential because it has an extreme \( x \)-value, an extreme \( y \)-value, or both

\[
\text{Figure 10.8} \quad \text{Illustration of Influential Observations}
\]

Note that the influential observation above are not outliers since their residuals are quite small (even zero)
Checking for Influential Observations

\[ \hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j \] where the \( h_{ij} \) are some functions of the \( x \)'s

\[ H = \begin{bmatrix} h_{ij} \end{bmatrix} \] is called the hat matrix

We can think of the \( h_{ij} \) as the leverage exerted by \( y_j \) on the fitted value \( \hat{y}_i \).

If \( \hat{y}_i \) is largely determined by \( y_i \) with very small contribution from the other \( y_j \)'s, then we say that the \( i \)th observations is influential (also called high leverage).
Checking for Influential Observations

Since $h_{ii}$ determines how much $\hat{y}_i$ depends on $y_i$ it is used as a measure of the leverage of the $i$-th observation.

Now $\sum_{i=1}^{n} h_{ii} = k + 1$ or the average of $h_{ii}$ is $(k + 1) / n$, where $k$ is the number of predictor variables.

**Rule of thumb:**

Regard any $h_{ii} > 2(k + 1) / n$ as high leverage. In simple regression $k = 1$ and so $h_{ii} > 4 / n$ is regarded as high leverage.

$$e_i^* = \frac{e_i}{SE(e_i)} = \frac{e_i}{s\sqrt{1 - h_{ii}}}$$

Standardized (Studentized) residuals will be large relative to the residual for high leverage observations.
Graph of Anscombe Data

Observation in row 8
Checking for Influential Observations

Fit Model Platform:

Anscombe Data Illustration:

\[ h_{ii} > \frac{4}{n} = \frac{4}{11} = 0.36 \]
Thus observation 8 is influential

Notice: Observation No. 8 is not an outlier since \[ e_8 = y_8 - \hat{y}_8 = 0 \]
How to Deal with Outliers and Influential Observations

- They can give a misleading picture of the relationship between $y$ and $x$
- Are they erroneous observations? If yes, they should be discarded.
- If the observation are valid then they should be included in the analysis
- Least Squares is very sensitive to these observations. Do analysis with and without outlier or influential observation.
- Or use robust method: Minimize the sum of the absolute deviations

$$\sum_{i=1}^{n} |y_i - (\beta_0 + \beta_1 x_i)|$$

These estimates are analogous to the sample median. LS estimates are analogous to the sample mean.
Data Transformations

If the functional relationship between $x$ and $y$ is known it may be possible to find a linearizing transformation analytically:

For $y = \alpha x^\beta$ take log on both sides: $\log y = \log \alpha + \beta \log x$

For $y = \alpha e^{\beta x}$ take log on both sides: $\log y = \log \alpha + \beta x$

If we are fitting an empirical model, then a linearizing transformation can be found by trial and error using the scatter plot as a guide.
Scatter Plots and Linearizing Transformations

Figure 10.10 Typical Scatter Plot Shapes and Corresponding Linearizing Transformations
Tire Tread Wear vs. Mileage: Exponential Model

\[ y = \alpha e^{-\beta x} \]

\[ \Rightarrow \log y = \log \alpha - \beta x \]
Tire Tread Wear vs. Mileage: Exponential Model

\[
\Rightarrow WEAR = e^{5.93} \times e^{-0.0298 MILES} = 376.15 e^{-0.0298 MILES}
\]
Tire Tread Wear vs. Mileage: Exponential Model

For linear model: $R^2 = .953$

Residual plot is still curved mainly because observation 1 is an outlier
Weights and Gas Mileages of 38 1978-79 Model Year Cars

<table>
<thead>
<tr>
<th>Year Cars</th>
<th>WT (1000 lb.)</th>
<th>MPG</th>
<th>WT (1000 lb.)</th>
<th>MPG</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>2.795</td>
<td>21.6</td>
<td>2.795</td>
<td>26.8</td>
</tr>
<tr>
<td>12</td>
<td>3.41</td>
<td>16.2</td>
<td>3.41</td>
<td>33.5</td>
</tr>
<tr>
<td>13</td>
<td>3.38</td>
<td>20.6</td>
<td>3.38</td>
<td>34.2</td>
</tr>
<tr>
<td>14</td>
<td>3.07</td>
<td>20.8</td>
<td>3.07</td>
<td>31.8</td>
</tr>
<tr>
<td>15</td>
<td>6.05</td>
<td>19.2</td>
<td>6.05</td>
<td>21.9</td>
</tr>
<tr>
<td>16</td>
<td>3.50</td>
<td>18.5</td>
<td>3.50</td>
<td>34.1</td>
</tr>
<tr>
<td>17</td>
<td>2.155</td>
<td>30.0</td>
<td>2.155</td>
<td>35.1</td>
</tr>
<tr>
<td>18</td>
<td>2.560</td>
<td>27.5</td>
<td>2.560</td>
<td>27.4</td>
</tr>
<tr>
<td>19</td>
<td>2.300</td>
<td>27.2</td>
<td>2.300</td>
<td>31.5</td>
</tr>
<tr>
<td>20</td>
<td>2.230</td>
<td>30.9</td>
<td>2.230</td>
<td>29.5</td>
</tr>
<tr>
<td>21</td>
<td>2.830</td>
<td>20.3</td>
<td>2.830</td>
<td>28.4</td>
</tr>
<tr>
<td>22</td>
<td>3.140</td>
<td>17.0</td>
<td>3.140</td>
<td>28.8</td>
</tr>
<tr>
<td>23</td>
<td>2.795</td>
<td>21.6</td>
<td>2.795</td>
<td>26.8</td>
</tr>
<tr>
<td>24</td>
<td>3.410</td>
<td>16.2</td>
<td>3.410</td>
<td>33.5</td>
</tr>
<tr>
<td>25</td>
<td>3.380</td>
<td>20.6</td>
<td>3.380</td>
<td>34.2</td>
</tr>
<tr>
<td>26</td>
<td>3.070</td>
<td>20.8</td>
<td>3.070</td>
<td>31.8</td>
</tr>
<tr>
<td>27</td>
<td>3.620</td>
<td>18.6</td>
<td>3.620</td>
<td>37.3</td>
</tr>
<tr>
<td>28</td>
<td>3.410</td>
<td>18.1</td>
<td>3.410</td>
<td>30.5</td>
</tr>
<tr>
<td>29</td>
<td>3.840</td>
<td>17.0</td>
<td>3.840</td>
<td>22.0</td>
</tr>
<tr>
<td>30</td>
<td>3.725</td>
<td>17.6</td>
<td>3.725</td>
<td>21.5</td>
</tr>
<tr>
<td>31</td>
<td>3.955</td>
<td>16.5</td>
<td>3.955</td>
<td>31.9</td>
</tr>
</tbody>
</table>
Weights and Gas Mileages of 38 1978-79 Model Year Cars

Use transformation $y \rightarrow \frac{100}{y}$. The new variable has units of gallons per 100 miles.

This transformation has the advantage of being physically interpretable.

Using *Fit Y by X* platform

A spline fits the data locally. Our prior curve fits have been global.
Weights and Gas Mileages of 38 1978-79 Model Year Cars
Variance Stabilizing Transformation

Suppose $\sigma_y \propto \mu^\alpha$

We want to find a transformation of $y$ that yields a constant variance. Suppose the transformation is a power of the original data, say

$$y^* = y^\lambda$$

Table 3–8  Variance-Stabilizing Transformations

<table>
<thead>
<tr>
<th>Relationship Between $\sigma_y$ and $\mu$</th>
<th>$\alpha$</th>
<th>$\lambda = 1 - \alpha$</th>
<th>Transformation</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_y \propto \text{constant}$</td>
<td>0</td>
<td>1</td>
<td>No transformation</td>
<td></td>
</tr>
<tr>
<td>$\sigma_y \propto \mu^{1/2}$</td>
<td>1/2</td>
<td>1/2</td>
<td>Square root</td>
<td>Poisson (count) data</td>
</tr>
<tr>
<td>$\sigma_y \propto \mu$</td>
<td>1</td>
<td>0</td>
<td>Log</td>
<td></td>
</tr>
<tr>
<td>$\sigma_y \propto \mu^{3/2}$</td>
<td>3/2</td>
<td>-1/2</td>
<td>Reciprocal square root</td>
<td></td>
</tr>
<tr>
<td>$\sigma_y \propto \mu^2$</td>
<td>2</td>
<td>-1</td>
<td>Reciprocal</td>
<td></td>
</tr>
</tbody>
</table>

For example: Take the square root of Poisson count data since for data so distributed the mean and the variance are the same.
Box-Cox Transformation:
Empirical Determination of Best Transformation

Fit Model Platform:

$$y^{(\lambda)} = \begin{cases} \left( y^\lambda - 1 \right) / \lambda \hat{y}^{\lambda-1} & \lambda \neq 0 \\ \hat{y} \ln y & \lambda = 0 \end{cases}$$

where $$\hat{y} = \sqrt[n]{\prod_{i=1}^{n} y_i}$$ = geometric mean
Transformation in Tire Data

A better fit than the polynomial and with one less parameter (3 versus 4)
Weights and Gas Mileages of 38 1978-79 Model Year Cars: Box-Cox Transformation

Reciprocal transformation among suggested transformations
Correlation Analysis

• When it is not clear which is the predictor variable and which is the response variable
• When both variables are random
• Try the bivariate normal distribution as a probability model for the joint distribution of two r.v.’s

Parameters: $\mu_X, \mu_Y, \sigma_X, \sigma_Y$ and

$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

The correlation $\rho$ is a measure of association between the two random variables $X$ and $Y$. Zero correlation corresponds to no association. Correlation of 1 or -1 represents perfect association.
Bivariate Normal Density Function

Figure 10.15  Bell Shaped Surface Defined by the Bivariate Normal Density Function
Statistical Inference on the Correlation Coefficient

\[ R = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}} \]

\[ H_0 : \rho = 0 \, \text{vs} \, H_1 : \rho \neq 0 \]

Test statistic \( T = \frac{R \sqrt{n-2}}{\sqrt{1 - R^2}} \sim t_{n-2} \)

Equivalent to testing \( H_0 : \beta_1 = 0 \) vs. \( H_1 : \beta_1 \neq 0 \)

Approximate test statistic available to

- Test nonzero null hypotheses
- Obtain confidence intervals for \( \rho \)

See textbook
Exercise 10.28

The following are the heights and weights of 30 eleven year old girls.

<table>
<thead>
<tr>
<th>Height (cm)</th>
<th>Weight (kg)</th>
<th>Height (cm)</th>
<th>Weight (kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>135</td>
<td>26</td>
<td>133</td>
<td>31</td>
</tr>
<tr>
<td>146</td>
<td>33</td>
<td>149</td>
<td>34</td>
</tr>
<tr>
<td>153</td>
<td>55</td>
<td>141</td>
<td>32</td>
</tr>
<tr>
<td>154</td>
<td>50</td>
<td>164</td>
<td>47</td>
</tr>
<tr>
<td>139</td>
<td>32</td>
<td>146</td>
<td>37</td>
</tr>
<tr>
<td>131</td>
<td>25</td>
<td>149</td>
<td>46</td>
</tr>
<tr>
<td>149</td>
<td>44</td>
<td>147</td>
<td>36</td>
</tr>
<tr>
<td>137</td>
<td>31</td>
<td>152</td>
<td>47</td>
</tr>
<tr>
<td>143</td>
<td>36</td>
<td>140</td>
<td>33</td>
</tr>
<tr>
<td>146</td>
<td>35</td>
<td>143</td>
<td>42</td>
</tr>
<tr>
<td>141</td>
<td>28</td>
<td>148</td>
<td>32</td>
</tr>
<tr>
<td>136</td>
<td>28</td>
<td>149</td>
<td>32</td>
</tr>
<tr>
<td>154</td>
<td>36</td>
<td>141</td>
<td>29</td>
</tr>
<tr>
<td>151</td>
<td>48</td>
<td>137</td>
<td>34</td>
</tr>
<tr>
<td>155</td>
<td>36</td>
<td>135</td>
<td>30</td>
</tr>
</tbody>
</table>

(a) Plot weights vs. heights.

(b) Calculate the correlation coefficient. Test if it is significantly greater than 0.7.
Exercise 10.28
Exercise 10.28

By default, a 95% bivariate normal density ellipse is imposed on each scatterplot. If the variables are bivariate normally distributed, this ellipse encloses approximately 95% of the points. The correlation of the variables is seen by the collapsing of the ellipse along the diagonal axis. If the ellipse is fairly round and is not diagonally oriented, the variables are uncorrelated.