Section 4.6 Simple Linear Regression

Objectives

- Basic philosophy of SLR and the regression assumptions
- Point & interval estimation of the model parameters, and how to make predictions
- Point and interval estimation of future observations from the model
- Regression diagnostics, including R^2 and basic residual analysis

Basic Philosophy

We have two variables X and Y. Here, X is not random (so we will write x), but Y is random. We believe that Y depends in some way on x. Some typical examples of (x, Y) pairs are

- x = study time and Y = score on a test.
- x = height and Y = weight.
- x = father's height and Y = son's height.

We focus our efforts on estimating two parameters, β_0 and β_1 in the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{where } \varepsilon_i \sim N\left(0, \sigma^2\right)$$

- Y_i is the (random) response for the ith case.
- β_0, β_1 are unknown parameters that we want to estimate. β_0 = (unknown) intercept, and β_1 = (unknown) slope.
- X_i is the value of the predictor variable for the ith case.
- ε_i is a (random) error term for the ith case, such that the mean = 0, variance = the "same" for all the cases, and the covariance between the ith and jth case = 0.

Least Squares Estimates

We begin with the likelihood function

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} f\left(y_i; \beta_0, \beta_1, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\left(y_i - \beta_0 - \beta_1 x_i\right)^2}{2\sigma^2}\right] \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{\sum\left(y_i - \beta_0 - \beta_1 x_i\right)^2}{2\sigma^2}\right] \\ -\ln L(\beta_0, \beta_1, \sigma^2) &= \frac{n}{2}\ln\left(2\pi\sigma^2\right) + \sum_{i=1}^{n} \frac{\left(y_i - \beta_0 - \beta_1 x_i\right)^2}{2\sigma^2} \end{aligned}$$

To maximize the log likelihood, let's *minimize* the summand, i.e., $H = \sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2$. That is, let's find β_0 and β_1 that minimize H. Because of the two parameters, we differentiate this wrt

β_0, β_1 and set them equal to zero, we get

$$\begin{aligned}
\frac{\partial}{\partial\beta_0}H &= -2\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)=0 \quad \Rightarrow \quad n\beta_0+\beta_1\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}y_i \\
\frac{\partial}{\partial\beta_1}H &= -2\sum_{i=1}^{n}x_i\left(y_i-\beta_0-\beta_1x_i\right) \\
&= -2\sum_{i=1}^{n}\left(x_iy_i-\beta_0x_i-\beta_1x_i^2\right)=0 \quad \Rightarrow \quad \beta_0\sum_{i=1}^{n}x_i+\beta_1\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}x_iy_i
\end{aligned}$$

Organizing these two equations, we get

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y}-\hat{\beta}_1\bar{x} \\
\hat{\beta}_1 &= \frac{\displaystyle\sum_{i=1}^{n}x_iy_i-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)/n}{\displaystyle\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2/n}=\frac{\displaystyle\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\displaystyle\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}
\end{aligned}$$

Shown below are the second derivatives:

$$\begin{aligned}
\frac{\partial^2}{\partial\beta_0^2}H &= 2n, & \frac{\partial^2}{\partial\beta_0\beta_1}H &= 2\sum_{i=1}^{n}x_i \\
\frac{\partial^2}{\partial\beta_1\beta_0}H &= 2\sum_{i=1}^{n}x_i, & \frac{\partial^2}{\partial\beta_1^2}H &= 2\sum_{i=1}^{n}x_i^2
\end{aligned}$$

And the 2×2 matrix consisting of these second-derivatives is positive definite because the (1,1)th element >0 and its determinant is also >0.

$$\begin{bmatrix} 2n, & 2\displaystyle\sum_{i=1}^{n}x_i \\ 2\displaystyle\sum_{i=1}^{n}x_i, & 2\displaystyle\sum_{i=1}^{n}x_i^2 \end{bmatrix} \quad \Rightarrow \quad \det>0$$

The conclusion? Use $\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_i$ line to ensure the line that fits the (x,y) pattern the best, i.e., the estimated line we have will leave the "smallest" gap between the observed y's and the estimated line. For this reason, they are also called the **least squares** estimates.

Next, let's find the mle of σ^2.

$$\frac{\partial}{\partial \sigma^2}\left\{-\ln L(\beta_0, \beta_1, \sigma^2)\right\} = \frac{n}{2\sigma^2} - \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}{2(\sigma^2)^2} = 0$$

We get

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

One note:

- In statistics, the "gap" between the observed value (y_i) and the expected (or predicted) value ($\hat{y_i}$) is called the "residual". So $\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the sum of squared residuals, and it's commonly called SS_E. For a point estimate of σ^2, we use $\frac{SS_E}{n-2}$, i.e., $\hat{\sigma} = s = \sqrt{\frac{SS_E}{n-2}}$.
- There are many equivalent formulas for $\hat{\beta}_1$ that are more intuitive, or at the least are easier to remember. One of the popular ones is $\hat{\beta}_1 = r \cdot \frac{SD_y}{SD_x}$, where r = correlation coefficient between x and y, SD_y = sd of y and SD_x = sd of x.

Inferences about the Parameters

Let's learn some more notations:

$$b_1 = \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\, y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b_0 = \hat{\beta}_0 = \bar{y} - b_1\bar{x}$$

Here is how to derive the expectation and the variance of the estimates:

$$\begin{aligned}
E(b_1) &= E\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{1}{S_{xx}} E\left\{\sum_{i=1}^{n}(x_i-\bar{x})\,y_i\right\} = \frac{1}{S_{xx}} E\left\{\sum_{i=1}^{n}(x_i y_i - \bar{x} y_i)\right\} \\
&= \frac{1}{S_{xx}}\left[\sum_{i=1}^{n} x_i E(y_i) - nE(\bar{x}\bar{y})\right] \\
&= \frac{1}{S_{xx}}\left[\sum_{i=1}^{n} x_i(\beta_0+\beta_1 x_i) - n\bar{x}(\beta_0+\beta_1\bar{x})\right] \\
&= \frac{1}{S_{xx}}\left[\beta_0\left(\sum_{i=1}^{n} x_i - n\bar{x}\right) + \beta_1\left(\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right)\right] \\
&= \frac{1}{S_{xx}}(0+\beta_1 S_{xx}) = \beta_1 \\
E(b_0) &= E(\bar{y}-b_1\bar{x}) = E(\beta_0+\beta_1\bar{x}) - \bar{x}E(b_1) = \beta_0 \\
Var(b_1) &= Var\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{1}{S_{xx}^2} Var\left\{\sum_{i=1}^{n}(x_i-\bar{x})\,y_i\right\} = \frac{1}{S_{xx}^2}\left\{\sum_{i=1}^{n}(x_i-\bar{x})^2\sigma^2\right\} = \frac{\sigma^2}{S_{xx}} \\
Var(b_0) &= Var(\bar{y}-b_1\bar{x}) = \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)
\end{aligned}$$

Furthermore, it can be shown that

$$b_1 \sim N(\text{mean}=\beta_1,\ \text{sd}=\sigma_{b_1}),\ \text{where } \sigma_{b_1} = \frac{\sigma}{\sqrt{S_{xx}}}$$

$\sigma_{b_1} = \dfrac{\sigma}{\sqrt{S_{xx}}}$ is also called the standard error of b_1 and we can estimate σ from the previous description by $s = \sqrt{\dfrac{SS_E}{n-2}}$. So the SE of b_1 becomes

$$s_{b_1} = \frac{s}{\sqrt{S_{xx}}}$$

See textbook page 204-205. It can be shown that $\dfrac{SS_E}{\sigma^2} = \dfrac{n\hat{\sigma^2}}{\sigma^2} \sim \chi^2_{(n-2)}$. Also, it turns out that b_0, b_1, and s are mutually independent. Therefore, we have the following t-distribution.

$$T_1 = \frac{\frac{b_1-\beta_1}{\sigma_{b_1}}}{\sqrt{\frac{SS_E}{\sigma^2}/(n-2)}} = \frac{\frac{b_1-\beta_1}{\sigma/\sqrt{S_{xx}}}}{\sqrt{\frac{SS_E}{\sigma^2}/(n-2)}} = \frac{b_1-\beta_1}{s/\sqrt{S_{xx}}} = \frac{b_1-\beta_1}{s_{b_1}} \sim t_{df=(n-2)}$$

Therefore, a $100(1-\alpha)\%$ confidence interval for β_1 is given by

$$b_1 \pm t_{\alpha/2}^{df=(n-2)} s_{b_1}$$

It can also be shown in a similar way,

$$b_0 \sim N\left(\text{mean} = \beta_0,\ \text{sd} = \sigma_{b_0}\right),\ \text{where } \sigma_{b_0} = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

$\sigma_{b_0} = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$ is the standard error of b_0 and the SE of b_1 becomes

$$s_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

Therefore, we have another t-distribution.

$$T_0 = \frac{\frac{b_0-\beta_0}{\sigma_{b_0}}}{\sqrt{\frac{SS_E}{\sigma^2}/(n-2)}} = \frac{\frac{b_0-\beta_0}{\sigma\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}}}{\sqrt{\frac{SS_E}{\sigma^2}/(n-2)}} = \frac{b_0-\beta_0}{s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}} = \frac{b_0-\beta_0}{s_{b_0}} \sim t_{df=(n-2)}$$

Therefore, a $100(1-\alpha)\%$ confidence interval for β_1 is given by

$$b_0 \pm t_{\alpha/2}^{df=(n-2)} s_{b_0}$$

We have seen how to estimate the coefficients of a regression line with both point estimates and confidence intervals. We have learned how to estimate a value $\hat{y}$ on the regression line for a given value of x, such as $x = x_0$. But how good is our estimate $\hat{y}$ at $x = x_0$? How much confidence do we have in this estimate? Furthermore, suppose we were going to observe another value of y at $x = x_0$. What can we say?

Intuitively, it should be easier to get bounds on the mean (average) value of y at x_0 (called a ***confidence interval*** for the mean value of y at x_0) than it is to get bounds on a future observation of y (called a ***prediction interval*** for y at x_0). It turns out the confidence intervals are narrower for the mean value, wider for the individual value. Our point estimate of y at x_0 is, of course, $\hat{y}$ at x_0, so for a confidence interval we will need to know the sampling distribution of $\hat{y}$'s. It turns out that $\hat{y}$ at x_0 is distributed as

$$\hat{y} \sim N\left(\text{mean} = E\left(y_{x_0}\right),\ \text{sd} = \sigma_{\hat{y}_{x_0}}\right),\ \text{where } \sigma_{\hat{y}_{x_0}} = \sigma\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x}^2)}{S_{xx}}}$$

$\sigma_{\hat{y}_{x_0}} = \sigma\sqrt{\dfrac{1}{n} + \dfrac{(x_0 - \bar{x}^2)}{S_{xx}}}$ is the standard error of $\hat{y}_{x_0}$ and the estimate is

$$s_{\hat{y}_{x_0}} = s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x}^2)}{S_{xx}}}$$

Therefore, we have the following t-distribution.

$$T_2 = \frac{\frac{\hat{y}_{x_0} - E\left(y_{x_0}\right)}{\sigma_{\hat{y}_{x_0}}}}{\sqrt{\frac{SS_E}{\sigma^2}/(n-2)}} = \frac{\frac{\hat{y}_{x_0} - E\left(y_{x_0}\right)}{\sigma\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x}^2)}{S_{xx}}}}}{\sqrt{\frac{SS_E}{\sigma^2}/(n-2)}} = \frac{\hat{y}_{x_0} - E\left(y_{x_0}\right)}{s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x}^2)}{S_{xx}}}} = \frac{\hat{y}_{x_0} - E\left(y_{x_0}\right)}{s_{\hat{y}_{x_0}}} \sim t_{df=(n-2)}$$

Therefore, a $100(1-\alpha)\%$ confidence interval (C.I.) for $E\left(y\right)$ at x_0 is given by

$$\hat{y}_{x_0} \pm t_{\alpha/2}^{df=(n-2)} s_{\hat{y}_{x_0}}$$

Next, the prediction intervals are slightly different. In order to find confidence bounds for a new observation of y (we will denote it y_{future}) we use the fact that

$$\hat{y}_{future} \sim N\left(\text{mean} = E\left(y_{future}\right),\ \text{sd} = \sigma_{\hat{y}_{future}}\right),\ \text{where } \sigma_{\hat{y}_{future}} = \sigma\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x}^2)}{S_{xx}}}$$

Of course σ is unknown and we estimate it with s. Therefore, a $100(1-\alpha)\%$ prediction interval (P.I.) for a future value of y at x_0 is given by

$$\hat{y}_{x_0} \pm t_{\alpha/2}^{df=(n-2)} s_{\hat{y}_{future}}$$

Take note that the prediction interval is "wider" than the confidence interval, as its SE is greater.

Ex 1. Consider the following sample data and carry out all the inferences involved.

Midterm (X)	Final (Y)
70	87
74	79
80	88
84	98
80	96
67	73
70	83
64	79
74	91
82	94

```
> x <- c(70,74,80,84,80,67,70,64,74,82)
> y <- c(87,79,88,98,96,73,83,79,91,94)
> model1 <- lm(y~x)
> summary(model1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.132     17.422   0.639  0.54072
x              1.016      0.233   4.359  0.00241 **
---
Residual standard error: 4.743 on 8 degrees of freedom
Multiple R-squared:  0.7038,    Adjusted R-squared:  0.6667
F-statistic:    19 on 1 and 8 DF,  p-value: 0.002414

> plot(y~x,pch=16,col=2)
> abline(model1,col=4)
> predict(model1,interval="confidence")
        fit      lwr       upr
1  82.22943 78.00927  86.44960
2  86.29216 82.82276  89.76156
3  92.38625 87.83693  96.93557
4  96.44897 90.28331 102.61464
5  92.38625 87.83693  96.93557
6  79.18239 73.87191  84.49287
7  82.22943 78.00927  86.44960
8  76.13534 69.51806  82.75262
9  86.29216 82.82276  89.76156
10 94.41761 89.10713  99.72809
> predict(model1,interval="prediction")
        fit      lwr       upr
1  82.22943 70.50528  93.95359
2  86.29216 74.81685  97.76747
3  92.38625 80.53963 104.23286
4  96.44897 83.89265 109.00530
5  92.38625 80.53963 104.23286
6  79.18239 67.02315  91.34163
7  82.22943 70.50528  93.95359
8  76.13534 63.35120  88.91949
9  86.29216 74.81685  97.76747
10 94.41761 82.25837 106.57685
> newx <- seq(60,95,0.2)
> ci <- predict(model1,list(x=newx), interval="confidence")
> pi <- predict(model1,list(x=newx), interval="prediction")
> plot(x,y,pch=16,col=2)
> matplot(newx,ci,type="l",lty=c(1,2,2),col=c(1,2,2),add=T)
> matplot(newx,pi,type="l",lty=c(1,3,3),col=c(1,4,4),add=T)
> legend(locator(1),c("regression line","95% ci","95% pi"),cex=0.8,lty=1:3,col=c(1,2,4))
> #The following command creates four diagnostic plots.
> par(mfrow=c(1,4))
```

```
> plot(model1)
```

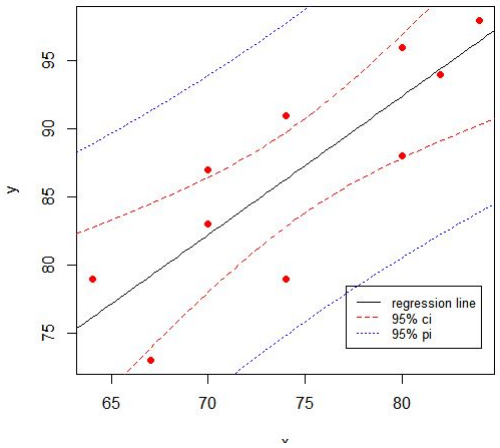

Figure 1: Regression line, 95% CI & 95% PI

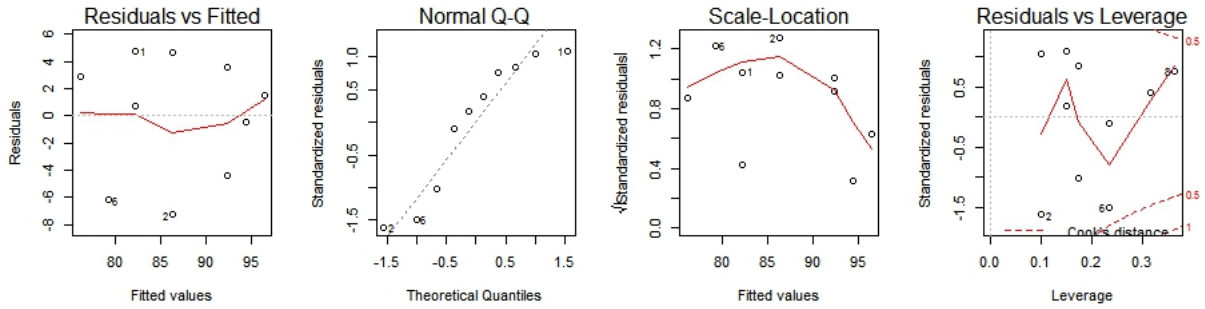

Figure 2: Diagnostic plots of a regression model

Section 4.8 One-Factor ANOVA

One-Factor Samples

Suppose you have collected n_i, where $(i = 1, 2, \ldots, m)$ samples from m groups:

Groups					Means
Y_1:	Y_{11}	Y_{12}	$\cdots$	Y_{1n_1}	$\bar{Y}_{1\cdot}$
Y_2:	Y_{21}	Y_{22}	$\cdots$	Y_{2n_2}	$\bar{Y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_m:	Y_{m1}	Y_{m2}	$\cdots$	Y_{mn_m}	$\bar{Y}_{m\cdot}$
Grand Mean:					$\bar{Y}_{\cdot\cdot}$

The hypotheses we want to test are:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_m \text{ (i.e., all group means are the same.)}$$
$$H_1 : not\, H_0 \text{ (i.e., some group means are significantly different.)}$$

In the end, all will be summarized in the following ANOVA table:

source	SS	df	MS	F-value	p-value
Treatment	SS_{trt}	$m-1$	$MS_{trt} = \frac{SS_{trt}}{m-1}$	MS_{trt}/MS_E	
Error	SS_E	$n-m$	$MS_E = \frac{SS_E}{n-m}$		
Total	SS_{tot}	$n-1$			

Here are all the SS (sum of squares) numbers and how the SS_{tot} is partitioned:

$$\begin{aligned}
SS_{tot} &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{\cdot\cdot}\right)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{i\cdot} + \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)^2 \quad \because \text{cross-product term} = 0 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{\cdot\cdot}\right)^2 + \sum_{i=1}^{m} n_i\left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)^2 \\
&= \qquad SS_E \qquad + \qquad SS_{trt}
\end{aligned}$$

We also have

$$\frac{SS_{trt}}{\sigma^2} \sim \chi^2_{(m-1)}, \qquad \frac{SS_E}{\sigma^2} \sim \chi^2_{(n-m)}$$

$$\Rightarrow \qquad \frac{\frac{SS_{trt}}{\sigma^2}/(m-1)}{\frac{SS_E}{\sigma^2}/(n-m)} = \frac{SS_{trt}/(m-1)}{SS_E/(m-m)} = \frac{MS_{trt}}{MS_E} \sim F_{(m-1),(n-m)}$$

Ex 2. Consider the following sample data and carry out all the inferences involved.

Observations							
Group 1:	92	90	87	105	86	83	102
Group 2:	100	108	98	110	114	97	94
Group 3:	143	149	138	136	139	120	145
Group 4:	147	144	160	149	152	131	134
Group 5:	142	155	119	134	133	146	152

```
> grp <- c(rep(1,7),rep(2,7),rep(3,7),rep(4,7),rep(5,7))
> y <- c(92,90,87,105,86,83,102,100,108,98,110,114,97,94,
+ 143,149,138,136,139,120,145,147,144,160,149,152,131,134,
+ 142,155,119,134,133,146,152)
> data1 <- data.frame(cbind(grp,y))
> head(data1)
> attach(data1)
> grp <- factor(grp)
> boxplot(y~grp,col="pink")
> model1 <- lm(y~grp)
> summary(model1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   92.143      3.670  25.105  < 2e-16 ***
grp2          10.857      5.191   2.092    0.045 *
grp3          46.429      5.191   8.945 5.74e-10 ***
grp4          53.143      5.191  10.238 2.64e-11 ***
grp5          48.000      5.191   9.248 2.74e-10 ***
---
Residual standard error: 9.711 on 30 degrees of freedom
Multiple R-squared:  0.8549,    Adjusted R-squared:  0.8356
F-statistic:  44.2 on 4 and 30 DF,  p-value: 3.664e-12

> anova(model1)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
grp        4 16672.1  4168.0  44.202 3.664e-12 ***
Residuals 30  2828.9    94.3
---
> summary(aov(y~grp))
            Df Sum Sq Mean Sq F value   Pr(>F)
grp          4  16672    4168    44.2 3.66e-12 ***
Residuals   30   2829      94
---
> boxplot(y~grp,col="pink")
> plot(TukeyHSD(aov(y~grp)))
```

```
> par(mfrow=c(1,4))
> plot(model1)
```

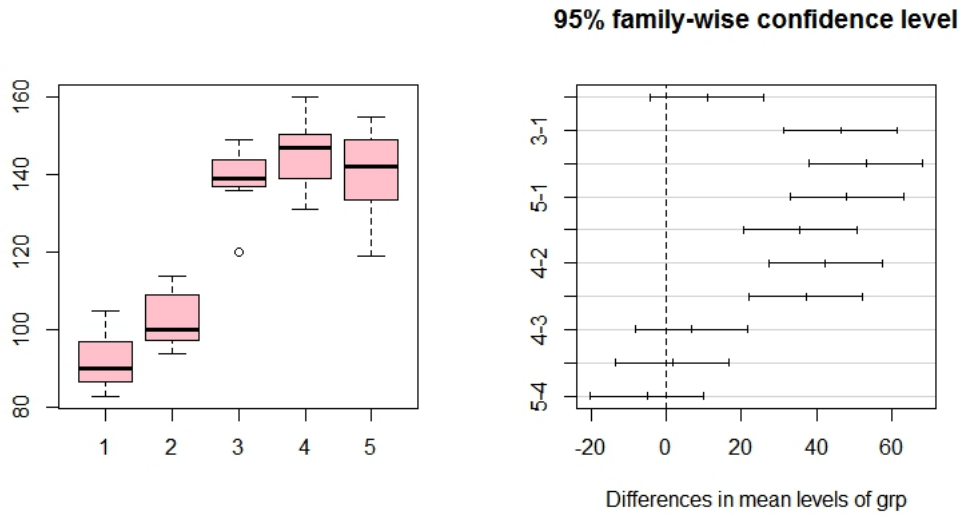

Figure 3: Boxplot & Tukey's pairwise comparison

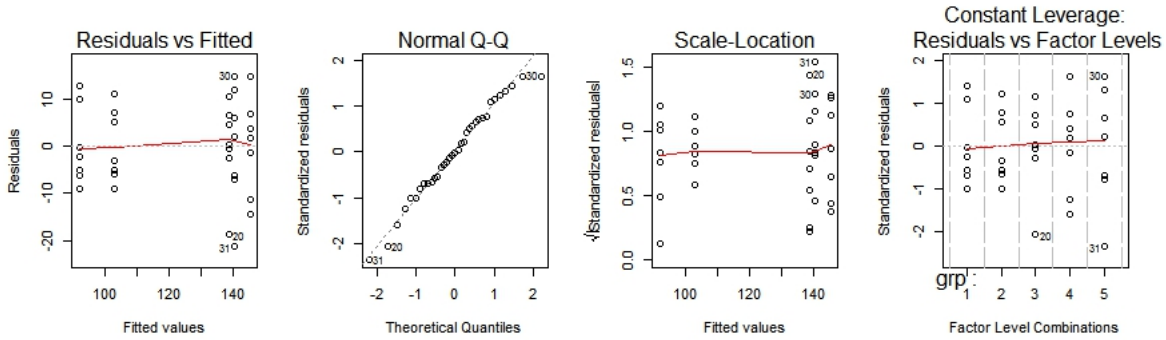

Figure 4: Diagnostic plots of ANOVA

Section 4.10 χ^2 Tests

Review: Facts about χ^2-Distribution

In the χ^2 distribution $X \sim \chi^2_{df}$, where df (= degrees of freedom) is the only parameter that uniquely determines the shape. The (theoretical) population mean is $\mu = df$ and the (theoretical) population standard deviation is $\sigma = \sqrt{2 \cdot df}$.

- If you square a random variable that has the standard normal distribution, it has $\chi^2_{(df=1)}$ distribution. This is often written as $Z^2 \sim \chi^2_{(df=1)}$.
- The random variable with a χ^2 distribution with k degrees of freedom is the sum of k independent, squared standard normal variables, i.e., $\chi^2_{(df=k)} = Z_1^2 + Z_2^2 + \cdots + Z_k^2$, where $Z \sim N(0,1)$.
- The curve is nonsymmetrical and skewed to the right.
- The mean, μ, is always located just to the right of the peak.
- The χ^2 test statistic is always greater than or equal to zero.
- When $df > 90$, the χ^2 curve is approximated by the normal distribution. For example, $X \sim \chi^2_{(df=1000)}$, then, $X \dot{\sim} N(\mu = 1000, \sigma = \sqrt{2000} = 44.7)$.

χ^2 Goodness of Fit Test

We test whether the data "fits" a particular distribution or not. For example, we can test if the color distribution of M&M bags fits what the company claims on their webpage. After flipping a coin many times, we can test if it fits a binomial distribution. We use a χ^2 test statistic to determine if there is a "good" fit or not.

Why χ^2? Demo for a binomial case

Let $Y_1 \sim binomial(n, p_1)$, then $Z = \dfrac{Y_1 - np_1}{\sqrt{np_1(1-p_1)}}$ has an approximate $N(0,1)$ distribution due to the CLT. Consider the following:

$$
\begin{aligned}
Q_1 &= \frac{(Y_1 - np_1)^2}{np_1(1-p_1)} = \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_1 - np_1)^2}{n(1-p_1)} \qquad \text{(Why?)} \\
&= \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_2 - np_2)^2}{np_2} \qquad \left[\because (Y_1 - np_1)^2 = \{n - Y_1 - n(1-p_1)\}^2 = (Y_2 - np_2)^2\right] \\
&= \sum_{i=1}^{2} \frac{(Y_i - np_i)^2}{np_i} \sim \chi^2_{(df=1)}
\end{aligned}
$$

This will be generalized to when there are k many categories. It can be shown:

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i} \sim \chi^2_{(df=k-1)}$$

The null and the alternative hypotheses for the goodness-of-fit test can be written as:

$H_0 : p_i = p_{i0},$ where $i = 1, 2, \ldots, k,$ (i.e., data fits the hypothesized distribution)

$H_1 : p_i \neq p_{i0}$ (i.e., at least in some cases, data does NOT fit the hypothesized distribution)

Ex 1. People were asked to write bunch of random digits. The result:

```
5 8 3 1 9 4 6 7 9 2 6 3 0 8 7 5 1 3 6 2 1 9 5 4 8 0
3 7 1 4 6 0 4 3 8 2 7 3 9 8 5 6 1 8 7 0 3 5 2 5 2
```

If these digits are truly random, the probability of the next digit can be either the same as the preceding one with the probability of 1/10 or "one" away from the preceding one with the probability of 2/10, or neither cases with the probability of 7/10. We want to test whether the data fits this thinking (i.e., random sequence examined by this idea), i.e.,

$H_0 : p_1 = \frac{1}{10}, p_2 = \frac{2}{10}, p_3 = \frac{7}{10}$

H_1 : At least one of the cases is significantly different from the hypothesized proportion.

Here is summary:

	observed freq	expected freq
same digit	0	$51 \times (1/10) = 5.1$
one-away digit	8	$51 \times (2/10) = 10.2$
others	43	$51 \times (7/10) = 35.7$

Test statistic:

$$\chi^2 = \sum_{i=1}^{3} \frac{(observed - expected)^2}{expected} = \frac{(0-5.1)^2}{5.1} + \frac{(8-10.2)^2}{10.2} + \frac{(43-35.7)^2}{35.7} = 7.07 \sim \chi^2_{(df=2)}$$

p-value $= 0.0292$, so we reject H_0 and conclude that the data didn't follow the hypothesized proportion, i.e., the data doesn't seem random. The whole thing can be done in R as shown below.

```
> x <- c(0,8,43)
> chisq.test(x,p <- c(0.1, 0.2, 0.7))

        Pearson's Chi-squared test

data:  x and p <- c(0.1, 0.2, 0.7)
X-squared = 6, df = 4, p-value = 0.1991
```

Ex 2. You flipped a coin 4 times a day and counted total number of H's every day. You did this for 100 days. The result:

Number of H's	observed freq
0	7
1	18
2	40
3	31
4	4

Test if the result agrees with X (total number of H's) being a binomial $(4, 1/2)$.

Answer:

Number of H's	observed freq	expected freq
0	7	6.25
1	18	25.0
2	40	37.5
3	31	25.0
4	4	6.25

Test statistic:

$$\chi^2 = \sum_{i=1}^{5} \frac{(obs - exp)^2}{exp} = \frac{(7-6.25)^2}{6.25} + \frac{(18-25)^2}{25} + \cdots + \frac{(4-6.25)^2}{6.25} = 4.47 \sim \chi^2_{(df=4)}$$

p-value $= 0.3462$, so we do not reject H_0 and conclude that the data supports the hypothesis of binomial $(4, 0.5)$.

Ex 3. You lose one more df by estimating another parameter!
Shown below are X, number of α particles emitted by barium-133 in 1/10 of a second, and counted by a Geiger counter.

```
7 4 3 6 4 4 5 3 5 3 5 5 3 2 5 4 3 3 7 6 6 4 3 11 9
6 7 4 5 4 7 3 2 8 6 7 4 1 9 8 4 8 9 3 9 7 7 9 3 10
```

Test $H_0 : X \sim Poisson$.

Answer: We first have to estimate the Poisson parameter λ by the mean of data, i.e., $\hat{\lambda} = \bar{x} = 5.4$. Then, we calculate the expected probabilities for each case and expected frequencies.

Cases	observed freq	expected freq
{0,1,2,3}	13	10.65
{4}	9	8.00
{5}	6	8.65
{6}	5	7.80
{7}	7	6.00
{8, 9, ...}	10	8.90

Test statistic:

$$\chi^2 = \sum_{i=1}^{6} \frac{(obs - exp)^2}{exp} = \frac{(13 - 10.65)^2}{10.65} + \cdots + \frac{(10 - 8.90)^2}{8.90} = 2.763 \sim \chi^2_{(df=4)}$$

p-value = 0.4018, so we do not reject H_0 and conclude that the data cannot reject the hypothesis that the counts form a Poisson distribution.

χ^2 Test for Homogeneity

The goodness-of-fit test can be used to decide whether a data fits a given distribution, but it will not suffice to decide whether two populations follow the same unknown distribution. A different test, called the "test for homogeneity," can be used to draw a conclusion about whether two populations have the same distribution. Here we're concerned about:

H_0 : The distributions of the two populations are the same.

H_1 : The distributions of the two populations are NOT the same.

Ex 4. Shown below are grade distribution of two groups of students.

observed freq	A	B	C	D	F	total
Group I	8	13	16	10	3	50
Group II	4	9	14	16	7	50

Test H_0 : Grade distribution of the two groups are the same.

Answer: Under H_0 that the probabilities of each grade is equal, the respective estimates of the probabilities are: 12/100=0.12, 22/100=0.22, 30/100=0.3, 26/100=0.26, and 10/100=0.1. Note also, since we have estimated these probabilities, the χ^2 test statistic will have a df $= (5-1)+(5-1)-4 = 4$. Here are the expected frequencies for each case.

expected freq	A	B	C	D	F
Group I	6	11	15	13	5
Group II	6	11	15	13	5

Test statistic:

$$\chi^2 = \sum_{j=1}^{2} \sum_{i=1}^{5} \frac{(obs - exp)^2}{exp} = \frac{(8 - 6)^2}{6} + \cdots + \frac{(7 - 5)^2}{5} = 5.18 \sim \chi^2_{(df=4)}$$

p-value = 0.2693, so we do not reject H_0 and conclude that we cannot say there is a significant difference in grade distribution between the two groups.

```
> data1 <- matrix(c(8,4,13,9,16,14,10,16,3,7), nrow=2, ncol=5)
> chisq.test(as.table(data1))$observed
   A  B  C  D  E
A  8 13 16 10  3
B  4  9 14 16  7
> chisq.test(as.table(data1))$expected
  A  B  C  D E
A 6 11 15 13 5
B 6 11 15 13 5
> chisq.test(as.table(data1))$residual
           A          B          C          D          E
A  0.8164966  0.6030227  0.2581989 -0.8320503 -0.8944272
B -0.8164966 -0.6030227 -0.2581989  0.8320503  0.8944272
> chisq.test(as.table(data1))

        Pearson's Chi-squared test

data:  as.table(data1)
X-squared = 5.1786, df = 4, p-value = 0.2695
```

χ^2 Test for Independence

Test of independence involves using a **contingency table** of observed (data) values. The test statistic for a test of independence is similar to that of a goodness-of-fit test:

$$\chi^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{(obs - exp)^2}{exp} \sim \chi^2_{df=(r-1)(c-1)},$$

where r = number of rows, and c = number of columns.

Ex 5. A random sample of 400 students at the University of Iowa shows the following breakdown of gender and colleges where they study.

observed freq	Business	Engineering	Liberal Arts	Nursing	Pharmacy	total
Male	21	16	145	2	6	190
Females	14	4	175	13	4	210

Test $H_0 : p_{ij} = p_{i\cdot}p_{\cdot j}$ (i.e., the college where a student studies is *independent* of the gender.)

Answer:

```
> data2 <- matrix(c(21,14,16,4,145,175,2,13,6,4), nrow=2, ncol=5)
> chisq.test(as.table(data2))$observed
    A   B   C   D   E
A  21  16 145   2   6
B  14   4 175  13   4
> chisq.test(as.table(data2))$expected
       A    B   C     D    E
```

```
A 16.625  9.5 152 7.125 4.75
B 18.375 10.5 168 7.875 5.25
> chisq.test(as.table(data2))$residual
           A          B          C          D          E
A  1.0729938  2.1088785 -0.5677750 -1.9200009  0.5735393
B -1.0206207 -2.0059435  0.5400617  1.8262852 -0.5455447
> chisq.test(as.table(data2))

        Pearson's Chi-squared test

data:  as.table(data2)
X-squared = 18.9265, df = 4, p-value = 0.0008125
```

We do reject H_0 and conclude that the number of students in colleges is highly dependent on gender, i.e., the two variables (gender and which college) are NOT independent.

Section 4.9 Distribution-Free CI & TI

Basics

Let $Y_1 < Y_2 < Y_3 < Y_4 < Y_5$ be the *order statistics* of a random sample of size $n = 5$ from any continuous distribution. Also, let $m = \pi_{0.5}$ (i.e., the 50th percentile) be the median. For example, we can find the following probability:

$$\begin{aligned} P\left(Y_1 < m < Y_5\right) &= \sum_{k=1}^{4} \binom{5}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{5-k} \\ &= 1 - P(X = 0) - P(X = 5), \text{ where } X \sim binomial(5, 1/2) \\ &= 1 - \left(\frac{1}{2}\right)^5 - \left(\frac{1}{2}\right)^5 \\ &= \frac{15}{16} = 0.94 \end{aligned}$$

Why $P\left(Y_1 < m < Y_5\right)$ is calculated like this? First, for any individual observation, say X, has $P(X < m) = 0.5$, and in order for Y_1 to be less than m and Y_5 to be greater than m, we must have 1, 2, 3, or 4 observations to be less than m. And we say (y_1, y_5) is a 94% (*distribution-free*) confidence interval for m.

In a similar way, when there are n independent trials, we calculate:

$$\begin{aligned} P\left(Y_i < m < Y_j\right) &= \sum_{k=i}^{j-1} \binom{n}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k} \\ &= 1 - \alpha \end{aligned}$$

and (y_i, y_j) is a $100(1-\alpha)\%$ (*distribution-free*) confidence intervals for the median m.

Ex 1. Suppose we have an ordered set of data ($n = 9$) like:

```
15.5    19.0    21.2    21.7    22.8    27.6    29.3    30.1    32.5
```

Let's calculate:

$$P\left(Y_2 < m < Y_8\right) = \sum_{k=2}^{7} \binom{9}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{9-k} = 0.9610$$

and $(y_2, y_8) = (19.0, 30.1)$ is a 96.1% (*distribution-free*) confidence intervals for the median m.

□

It turns out we can argue the same thing for any percentile π_p. In this case, any individual observation X has $P(X < \pi_p) = p$, so when there are n independent trials, we calculate:

$$P\left(Y_i < \pi_p < Y_j\right) = \sum_{k=i}^{j-1} \binom{n}{k} p^k \left(1-p\right)^{n-k} = 1 - \alpha$$

and (y_i, y_j) is a $100(1-\alpha)\%$ (*distribution-free*) confidence intervals for the percentile π_p.

Ex 2. Suppose we have an ordered set of data ($n = 27$) like:

```
161   169   171   174   179   180   183   184   186   187   192   193   196   200
204   205   213   221   222   229   241   243   256   264   291   317   376
```

First, note that $\pi_{0.25}$ (i.e., the first quartile) $= (n+1)p = (27+1)(0.25) = 7$, and we have $\hat{\pi}_{0.25} = y_7 = 183$. Now, let's see how much confidence we can have with (y_4, y_{10}) being a confidence interval for y_7.

$$P\left(Y_4 < \pi_{0.25} < Y_{10}\right) = \sum_{k=4}^{9} \binom{27}{k} \left(0.25\right)^k \left(0.75\right)^{27-k} = 0.8201$$

i.e., $(y_4, y_{10}) = (174, 187)$ is a 82.01% (*distribution-free*) confidence intervals for the 25th percentile $\pi_{0.25}$.

One note: For some of these binomial probability calculations, it's OK to use the normal approximation. For example, in the last problem where we calculate $P(4 \le X \le 9)$, where $X \sim$ binomial $(n = 27, p = 1/4)$, and $X \dot{\sim} N\left(\mu = 27/4 = 6.75,\, \sigma = \sqrt{27 \times (1/4) \times (3/4)} = 2.25\right)$. Finding the same probability by normal approximation, we have:

$$P\left(4 \le X \le 9\right) = P\left(3.5 \le X \le 9.5\right) \approx P\left(\frac{3.5 - 6.75}{2.25} \le Z \le \frac{9.5 - 6.75}{2.25}\right) = 0.8149$$

i.e., the normal approximation works rather well for such a case.

□

Theorem 1. *Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics (based on random samples $x_1, x_2, \ldots, x_n$). Then the pdf of Y_k is*

$$g_k(y) = \frac{n!}{(k-1)!(n-k)!} [F(y)]^{k-1} f(y) [1 - f(y)]^{n-k}, \text{ where } f(\cdot), F(\cdot) = \text{ pdf and cdf of } X.$$

Proof. □

Theorem 2. *Let $U_{(1)} < U_{(2)} < \cdots < U_{(n)}$ be the order statistics, where $U_i \sim uniform(0,1)$. Then $U_{(k)}$ has a beta distribution with two parameters k and $(n-k+1)$.*

Proof. From Theorem 1, we have

$$\begin{aligned} g_k(y) &= \frac{n!}{(k-1)!(n-k)!} (y)^{k-1} (1-y)^{n-k}, \quad 0 < y < 1 \\ &= \text{ pdf of } \beta(k, n-k+1) \end{aligned}$$

□

Theorem 3. *Let $X_1, X_2, \ldots, X_n$ be random variables with cdf $F(\cdot)$, then , where $U_i \sim uniform(0,1)$. Then $F\left\{X_{(k)}\right\}$ has a beta distribution with two parameters k and $(n-k+1)$.*

Proof. First, note that $U_i = F(X_i)$ is iid uniform (0, 1) due to the probability integral transformation. Furthermore, $F(\cdot)$ is a nondecreasing function, i.e., $F(\cdot)$ preserves order. So, $U_{(i)} = F\left\{X_{(i)}\right\}$. That is,

$$\begin{aligned} &\left\{U_{(1)}, U_{(2)}, \ldots, U_{(n)}\right\} \sim \left\{F\left(X_{(1)}\right), F\left(X_{(2)}\right), \ldots, F\left(X_{(n)}\right)\right\} \\ &\therefore \quad F\left(X_{(k)}\right) \sim \beta(k, n-k+1) \end{aligned}$$

□

Application: Let Y_k be the order statistic of X_k, i.e., $Y_k = X_{(k)}$. Consider the following $n+1$

random variables:

$$\begin{aligned}
W_1 &= F(Y_1) \\
W_2 &= F(Y_2) - F(Y_1) \\
W_3 &= F(Y_3) - F(Y_2) \\
&\cdots \\
W_n &= F(Y_n) - F(Y_{n-1}) \\
W_{n+1} &= 1 - F(Y_n)
\end{aligned}$$

- These $W_1, W_2, \ldots, W_{n+1}$ are called the "coverage" of intervals, for example $(Y_i, Y_{i+1}]$.
- Note that sum of k of thse intervals, i.e., $W_1 + \cdots W_k = F(Y_k) \sim \beta(k, n-k+1)$
- $F(Y_j) - F(Y_i)$, $i < j$ is the sum of $k = j - i$ coverages, so that it will have $\beta(j-i, n-j+i+1)$, i.e.,

$$\gamma = P\left\{F(Y_j) - F(Y_i) \geq p\right\} = \int_p^1 \frac{\Gamma(n+1)}{\Gamma(j-i)\Gamma(n-j+i+1)} v^{j-i-1}(1-v)^{n-j+i} dv$$

  and this is called a $100\gamma\%$ "tolerance interval" for $100p\%$ of the distribution.

Ex 3. Let $Y_1 < Y_2 < \cdots < Y_6$ be the *order statistics* of a random sample of size $n = 6$ from any continuous distribution. Also, let $p = 0.8$, then

$$\gamma = P\left\{F(Y_6) - F(Y_1) \geq 0.8\right\} = \int_{0.8}^1 \frac{\Gamma(7)}{\Gamma(5)\Gamma(2)} v^4(1-v) dv = 0.34$$

i.e., (y_1, y_6) is a 34% (*distribution-free*) tolerance interval for 80% of the distribution.

□