Chapter 14

Linear Regression Analysis

14.1 Introduction

A problem arising in engineering, economics, medicine, and other areas, is that of investigating the relationship between two or more variables. In such settings, the goal is to model a random variable Y (often continuous) as a function of one or more independent variables, say, $x_1, x_2, ..., x_k$. Mathematically, we can express this model as

$$Y = g(x_1, x_2, ..., x_k) + \epsilon,$$

where $g : \mathbb{R}^k \to \mathbb{R}$ is a function (whose form may or may not be specified). This is called a **regression model**.

The presence of the (random) error ϵ conveys the fact that the relationship between the dependent variable Y and the independent variables $x_1, x_2, ..., x_k$ through g is not deterministic. Instead, the term ϵ "absorbs" all variation in Y that is not explained by $g(x_1, x_2, ..., x_k)$.

Terminology: In this course, we will consider models of the form

$$Y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}_{g(x_1, x_2, ..., x_k)} + \epsilon,$$

that is, g is a linear function of $\beta_0, \beta_1, \beta_2, ..., \beta_k$. We call this a **linear regression model**.

- The **response variable** Y is random (but we do get to observe its value).
- The independent variables $x_1, x_2, ..., x_k$ are fixed (and observed).
- The **regression parameters** $\beta_0, \beta_1, \beta_2, ..., \beta_k$ are unknown. These are to be estimated on the basis of the observed data.
- The **error term** ϵ is random (and not observed).

Terminology: More precisely, we call a regression model a **linear regression model** if the regression parameters enter the g function in a linear fashion. For example, each of the models is

a linear regression model:

$$Y = \underbrace{\beta_0 + \beta_1 x}_{g(x)} + \epsilon$$
$$Y = \underbrace{\beta_0 + \beta_1 x + \beta_{11} x^2}_{g(x)} + \epsilon$$
$$Y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2}_{g(x)} + \epsilon$$

The term "linear" does not refer to the shape of the regression function g. It refers to how the regression parameters $\beta_0, \beta_1, \beta_2, ..., \beta_k$ enter the g function.

Important: Regression models (linear or otherwise) are models for a population of individuals. From a statistical inference standpoint, our goal is the same as in previous chapters. We will use sample information to estimate the population parameters in the model. We say that we are "estimating" or "fitting the model" with the observed data

14.2 Simple linear regression model

A simple linear regression model includes only one independent variable x and is of the form

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

The population regression function $g(x) = \beta_0 + \beta_1 x$ is a straight line with intercept β_0 and slope β_1. These parameters describe the population of individuals for which this model is assumed.

Note: If $E(\epsilon) = 0$, then

$$E(Y) = E(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x + E(\epsilon) = \beta_0 + \beta_1 x.$$

Therefore, we have the following interpretations for the population regression parameters β_0 and β_1:

- β_0 quantifies the population mean of Y when $x = 0$.
- β_1 quantifies the population-level change in $E(Y)$ brought about by a one-unit change in x.

Example 14.1. Recall data frome Example 13.10. As part of a waste removal project, a new compression machine for processing sewage sludge is being studied. Engineers are interested in the following variables:

$$Y = \text{moisture control of compressed pellets (measured as a percent)}$$
$$x = \text{machine filtration rate (kg-DS/m/hr)}.$$

Engineers collect observations of (x, Y) from a random sample of $n = 20$ sewage specimens; the data are given below.

Specimen	x	y	Specimen	x	y
1	125.3	77.9	11	159.5	79.9
2	98.2	76.8	12	145.8	79.0
3	201.4	81.5	13	75.1	76.7
4	147.3	79.8	14	151.4	78.2
5	145.9	78.2	15	144.2	79.5
6	124.7	78.3	16	125.0	78.1
7	112.2	77.5	17	198.8	81.5
8	120.2	77.0	18	132.5	77.0
9	161.2	80.1	19	159.6	79.0
10	178.9	80.2	20	110.7	78.6

Table 14.2.1: Sewage data. Moisture (Y , measured as a percentage) and machine filtration rate (x, measured in kg-DS/m/hr). There are n = 20 observations.

```
> filt.rate<-c(125.3,98.2,201.4,147.3,145.9,124.7,112.2,120.2,161.2,178.9,
+              159.5,145.8,75.1,151.4,144.2,125,198.8,132.5,159.6,110.7)
> moisture<-c(77.9,76.8,81.5,79.8,78.2,78.3,77.5,77,80.1,80.2,
+             79.9,79,76.7,78.2,79.5,78.1,81.5,77,79,78.6)
> ggdat<-data.frame(filt.rate=filt.rate,moisture=moisture)
> ggplot(ggdat, aes(x=filt.rate, y=moisture)) +
+   geom_point(shape=1)+
+   geom_smooth(alpha=0.25,color="black",method="lm")+
+   theme_bw()+
+   xlab("Filtration rate (km/m/hr)")+
+   ylab("Pellet Moisture (%)")
```

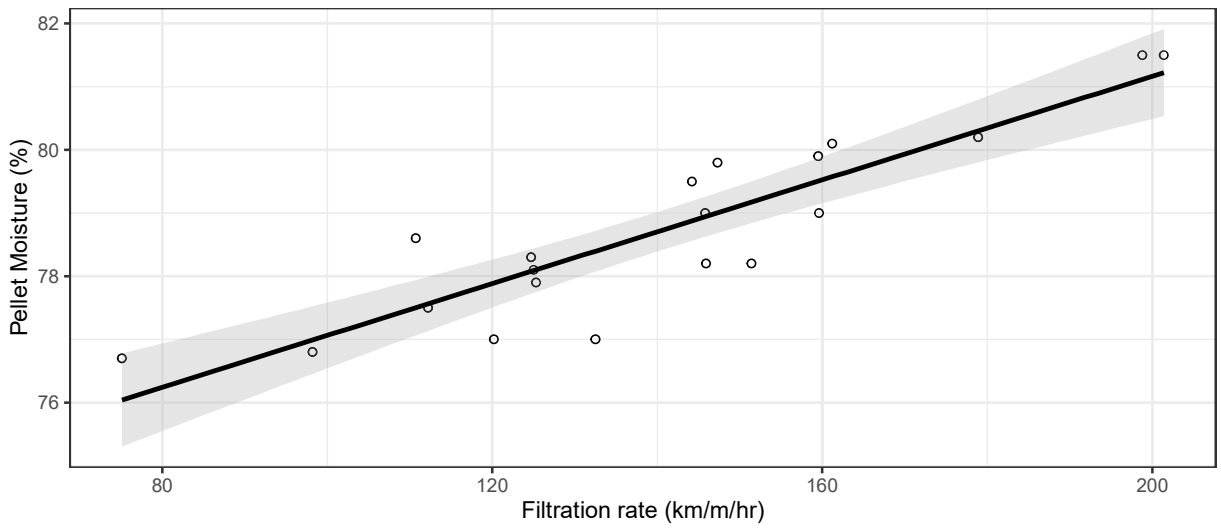

Figure 14.2.1: Scatterplot of pellet moisture Y (measured as a percentage) as a function of machine filtration rate x (measured in kg-DS/m/hr). The estimated linear regression model is superimposed.

Figure 14.2.1 displays the sample data in a scatterplot. This sample information suggests the variables Y and x are linearly related, although there is a large amount of variation that is unexplained.

This unexplained variability could arise from other independent variables (e.g., applied temperature, pressure, sludge mass, etc.) that also influence the moisture percentage Y but are not present in the model. It could also arise from measurement error or just random variation in the sludge compression process.

Inference: What does the sample information suggest about the population? Do we have evidence that Y and x are linearly related in the population?

14.3 Least Squares Estimation

When we say, "fit a regression model," we mean that we are estimating the population regression parameters in the model with the observed sample information (data). Suppose we have a random sample of observations (x_i, Y_i), $i = 1, 2, ..., n$, and postulate the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

for $i = 1, 2, ..., n$. Our first goal is to estimate β_0 and β_1. Formal assumptions for the error terms ϵ_i will be given later.

Terminology: The most common method of estimating the population parameters β_0 and β_1 is least squares. The method of least squares says to choose the values of β_0 and β_1 that minimize

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

Denote the least squares estimators by $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively, that is, the values of β_0 and β_1 that minimize $Q(\beta_0, \beta_1)$. A two-variable calculus minimization argument can be used to find expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$. Taking partial derivatives of $Q(\beta_0, \beta_1)$, we obtain

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_i)] \stackrel{\text{set}}{=} 0$$
$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_i) x_i] \stackrel{\text{set}}{=} 0$$

Solving for β_0 and β_1 gives the least squares estimators

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{SS_{xy}}{SS_{xx}}.$$

The estimated model is written as follows:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Remark: These estimators are equivalent to the maximum likelihood estimators. In OLS regression analysis, we assume that the errors are approximately Gaussian; e.g.,

$$\epsilon_i = (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) \sim \text{Gaussian}\left(\mu_\epsilon = 0, \sigma_\epsilon = \sigma\right),$$

where σ is estimated using the observations.

Under this assumption we take $(\hat{\beta}_0, \hat{\beta}_1)$ to be

$$\arg \max_{(\hat{\beta}_0,\hat{\beta}_1)\in\mathbb{R}} \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\sum_{i=1}^{n} \frac{(Y_i-(\hat{\beta}_0+\hat{\beta}_1 x_i))^2}{2\sigma^2}},$$

the maximum likelihood estimates under the normality assumption on ϵ_i.

We use `R` to calculate the equation of the least squares regression line for the sewage data in Example 14.1. Here is the output:

```
> sewage.mod<-lm(moisture~filt.rate)
> summary(sewage.mod)

Call:
lm(formula = moisture ~ filt.rate)

Residuals:
Min       1Q   Median       3Q      Max
-1.39552 -0.27694  0.03548  0.42913  1.09901

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 72.958547   0.697528 104.596  < 2e-16 ***
filt.rate    0.041034   0.004837   8.484 1.05e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6653 on 18 degrees of freedom
Multiple R-squared:  0.7999,        Adjusted R-squared:  0.7888
F-statistic: 71.97 on 1 and 18 DF,  p-value: 1.052e-07
```

The least squares estimates (to 3 dp) for the sewage data are

$$\hat{\beta}_0 = 72.959$$
$$\hat{\beta}_1 = 0.041.$$

The estimated model (plotted in Figure 14.2.1) is

$$\hat{Y} = 72.959 + 0.041x$$

or, in other words,

$$\widehat{\text{Moisture}} = 72.959 + 0.041\text{Filtration rate}.$$

Note: The estimated model is also called the prediction equation. This is because we can now predict the value of Y (moisture percentage) for a given value of x (filtration rate). For example, when the filtration rate is $x = 150$ kg-DS/m/hr, we would predict the moisture percentage to be

$$\hat{Y}(150) = 72.959 + 0.041(150) \approx 79.11.$$

Of course, this prediction comes directly from the sample of observations used to fit the regression model. Therefore, we will eventually want to quantify the uncertainty in this prediction; e.g., how variable is this prediction?

Remark: While this is a powerful result – we can perhaps provide insight about future results – the model specification itself is often of more interest to us as it describes the association between response and explanatory variables.

14.4 Model Assumptions and Sampling Distributions

We investigate the properties of the least squares estimators b0 and b1 as estimators of the population-level regression parameters β_0 and β_1 in the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, for $i = 1, 2, ..., n$. To do this, we need statistical assumptions on the errors ϵ_i . **Assumptions:** We will assume throughout that the random variables ϵ_i are independent and identically Gaussian.

$$\begin{aligned} \epsilon_i &\sim \text{Gaussian}\,(\mu_\epsilon = 0, \sigma_\epsilon = \sigma) \\ E(\epsilon_i) &= 0, \ \text{for} i = 1, 2, ..., n \\ var(\epsilon_i) &= \sigma^2. \end{aligned}$$

Under these assumptions, we can derive the following results for the simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon$$

Result 1:

$$Y \sim \text{Gaussian}(\mu_y = \beta_0 + \beta_1 x . \sigma_y = \sigma^2].$$

In other words, the response variable Y is Gaussian distributed with mean $\beta_0 + \beta_1 x$ and variance $sigma^2$. Note that the population mean of Y depends on x. The population variance of Y does not depend on x – this makes sense because we assumed constant variance for the ϵ_i.

Result 2: The least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1, respectively, that is,

$$\begin{aligned} E(\hat{\beta}_0) &= \beta_0 \\ E(\hat{\beta}_1) &= \beta_1. \end{aligned}$$

Result 3: The least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ have Gaussian sampling distributions; specifically,

$$\begin{aligned} \hat{\beta}_0 &\sim \text{Gaussian}(\mu_{\hat{\beta}_0} = \beta_0, \sigma_{\beta_0} = c_0\sigma^2) \\ \hat{\beta}_1 &\sim \text{Gaussian}(\mu_{\hat{\beta}_1} = \beta_1, \sigma_{\beta_1} = c_1\sigma^2) \end{aligned}$$

where

$$\begin{aligned} c_0 &= \frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}} \\ c_1 &= \frac{1}{S_{xx}}. \end{aligned}$$

These distributions are needed to write confidence intervals and perform hypothesis tests for β_0 and β_1 (i.e., to perform statistical inference for the population).

14.5 Estimating the error variance

In the simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon$$

where $\epsilon \sim$ Gaussian$(\mu_{\epsilon_i} = 0, \sigma^2_{\epsilon_i} = \sigma^2)$, we now turn our attention to estimating σ^2, the error variance.

As we did in estimating β_0 and β_1 (the population level regression parameters), we will use the observed data (x_i, Y_i), $i = 1, 2, ..., n$, to estimate the error variance σ^2. The error variance is also a population level parameter and quantifies how variable the population is for a given model.

Define the i_th **fitted value** by

$$\hat{Y}_i = \beta_0 + \beta_1 x_i,$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares estimators. Each observation has its own fitted value.

Define the i_th **residual** by

$$e_i = Y_i - \hat{Y}_i.$$

Each observation has its own residual.

Sewage data: Below, we calculated the fitted values and residuals for each observation.

```
> tab<-cbind(filt.rate,moisture,y.hat=sewage.mod$fitted.values,e=sewage.mod$residuals)
```

	x	y	$\hat{Y}$	e		x	y	$\hat{Y}$	e
1	125.3	77.9	78.100	-0.200	11	159.5	79.9	79.503	0.397
2	98.2	76.8	76.988	-0.188	12	145.8	79.0	78.941	0.059
3	201.4	81.5	81.223	0.277	13	75.1	76.7	76.040	0.660
4	147.3	79.8	79.003	0.797	14	151.4	78.2	79.171	-0.971
5	145.9	78.2	78.945	-0.745	15	144.2	79.5	78.876	0.624
6	124.7	78.3	78.075	0.225	16	125.0	78.1	78.088	0.012
7	112.2	77.5	77.563	-0.062	17	198.8	81.5	81.116	0.384
8	120.2	77.0	77.891	-0.891	18	132.5	77.0	78.396	-1.396
9	161.2	80.1	79.573	0.527	19	159.6	79.0	79.508	-0.508
10	178.9	80.2	80.299	-0.099	20	110.7	78.6	77.501	1.099

Table 14.5.2: Sewage data. Fitted values and residuals from the least squares fit.

Note that

- If an observation's Y value is above the least squares regression line, then $Y_i > \hat{Y}_i$ and its residual e_i is positive.
- If an observation's Y value is below the least squares regression line, then $Y_i < \hat{Y}_i$ and its residual e_i is negative.
- If an observation's Y value is on the least squares regression line, then $Y_i = \hat{Y}_i$ and its residual e_i is zero.

In our simple linear regression model,

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (Y_i - \hat{Y}_i) = 0.$$

That is, the residuals sum to zero. For the sewage data in Example 14.2.1,

```
> sum(sewage.mod$residuals)
[1] -9.020562e-17
```

We can also plot the observed residuals to check the Gaussian assumption on the errors. In Figure 14.5.2, we see no severe departures from normality and so we don't have any concerns about the normality of error terms.

```
> g1<-ggplot(data=ggdat,aes(x=residuals))+
+   geom_histogram(aes(y=..density..),
+                  fill="lightblue",color="black",bins=8)+
+   geom_hline(yintercept=0)+
+   geom_density(fill="red",alpha=0.2)+
+   theme_bw()+
+   xlab("Residuals")+
+   ylab("Density")
> library("qqplotr")
> g2<-ggplot(data=ggdat,aes(sample=residuals))+
+   stat_qq_band(alpha=0.25) +
+   stat_qq_line() +
+   stat_qq_point() +
+   theme_bw()+
+   xlab("Gaussian Quantiles")+
+   ylab("Sample Quantiles")
> grid.arrange(g1,g2,ncol=2)
```

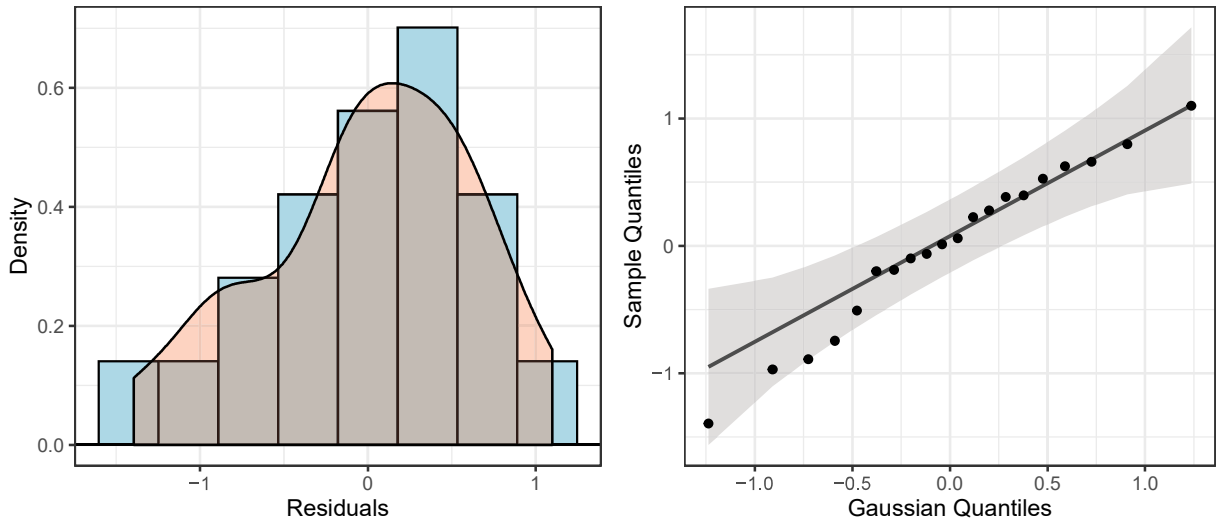

Figure 14.5.2: Sewage data: **Left:** A histogram of the residuals with a superimposed density plot. **Right:** A qq plot of the residuals.

To check for constant variance among residuals, we plot the observed residuals against the fitted values ($\hat{Y}$). In Figure 14.5.3, we see no pattern and a roughly even band around $e = 0$.

```
> ggdat<-data.frame(residuals=sewage.mod$residuals,
+                   fitted=sewage.mod$fitted.values)
```

```
> ggplot(data=ggdat,aes(x=fitted,y=residuals))+
+   geom_point(shape=1)+
+   geom_hline(yintercept=0,linetype="dashed")+
+   theme_bw()+
+   xlab(bquote(hat(Y)))+
+   ylab("Residuals")
```

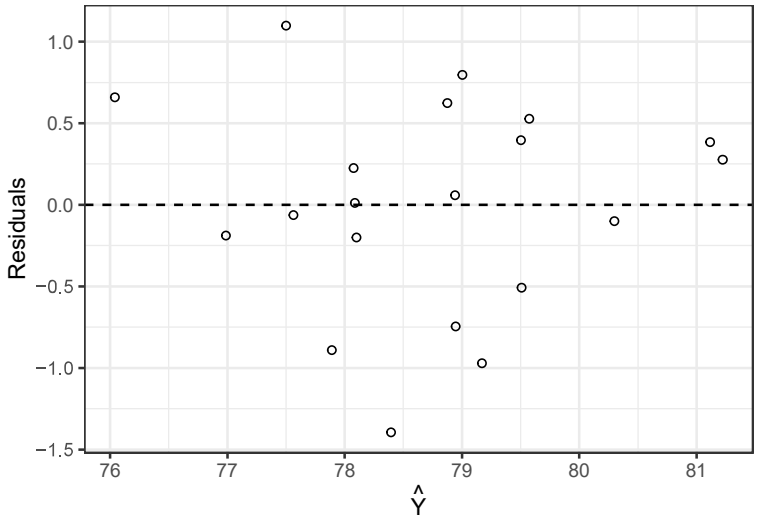

Figure 14.5.3: Sewage data: The observed residuals against the fitted values ($\hat{Y}$).

Define the **residual sum of squares** by

$$\begin{aligned} SS_{res} &= \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2. \end{aligned}$$

In the simple linear regression model, the **residual mean squares**

$$MS_{res} = \frac{SS_{res}}{n-2}$$

is an unbiased estimator of σ^2, that is,

$$E(MS_{res}) = \sigma^2.$$

The quantity

$$\hat{sigma} = \sqrt{MS_{res}} = \sqrt{\frac{SS_{res}}{n-2}}$$

estimates σ and is called the **residual standard error**.

To illustrate for the sewage data in Example 14.2.1, we can calculate the MS_{res} in `R` as follows.

```
> (MS.res<-sum(sewage.mod$residuals^2)/(length(moisture)-2))
[1] 0.4426659
> (sigma.hat<-sqrt(MS.res))
[1] 0.6653314
```

Remark: This is given as the "Residual standard error" when we asked for the summary of the model via `summary(sewage.mod)`.

We can also plot the observed errors to assess the normality assumptions on e_i.

14.6 Statistical Inference for β_0 and β_1

In the simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

we now discuss the formal question:

> "What does the sample information from an estimated regression model suggest about the population?"

In other words, we pursue statistical inference for the population level regression parameters β_0 and β_1. In practice, inference for the slope parameter β_1 is of primary interest because of its connection to the independent variable x in the model. For example, if $\beta_1 = 0$, then Y and x are not linearly related in the population. Statistical inference for β_0 is less meaningful, unless one is explicitly interested in the mean of Y when $x = 0$. We will generally not pursue this.

Confidence interval: Under our regression model assumptions, the following sampling distribution arises:

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{res}}{SS_{xx}}}} \sim t(n-2).$$

This result can be used to derive a $100(1-\alpha)$ percent confidence interval for β_1, which is given by

$$\hat{\beta}_1 \pm t_{n-2,1-\alpha/2}\sqrt{\frac{MS_{res}}{SS_{xx}}}.$$

Note that the form of the interval is similar to our other t intervals.

$$\underbrace{\text{point estimate}}_{\hat{\beta}_1} \pm \underbrace{\text{quantile}}_{t_{n-2,1-\alpha/2}} \underbrace{\text{standard error}}_{\sqrt{\frac{MS_{res}}{SS_{xx}}}}.$$

We interpret the interval in the same way:

> "We are $100(1-\alpha)$ percent confident that the population parameter β_1 is in this interval."

Of particular interest is the value $\beta_1 = 0$:

- If the confidence interval for β_1 contains "0," this suggests that Y and x are not linearly related in the population.
- If the confidence interval for β_1 does not contain "0," this suggests that Y and x are linearly related in the population.

Recall Example 14.2.1. We can use the `confint()` function in `R` to calculate a 95 percent confidence interval for β_1:

```
> confint(sewage.mod,level=0.95)
                  2.5 %      97.5 %
(Intercept) 71.49309400 74.42399995
filt.rate    0.03087207  0.05119547
```

We are 95 percent confident that the population parameter β_1 is between 0.0309 and 0.0511. This means for every one unit increase in the machine filtration rate x, we are 95 percent confident that the population mean absorption $E(Y)$ will increase between 0.0309 and 0.0511 percent.

Note that this interval does not contain "0" and includes only positive values. There is strong evidence that the absorption rate Y is positively linearly related to machine filtration rate x in the population. The confidence interval gives information about how strong this relationship is.

Under our regression model assumptions, if we wanted to formally test

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0,$$

we would use

$$t = \frac{\hat{\beta}_1}{\sqrt{\frac{MS_{res}}{SS_{xx}}}}$$

as a test statistic and reject H_0 if the corresponding p-value was small.

To illustrate for the sewage data in Example 14.2.1, we can ask `R` for these values directly.

```
> summary(sewage.mod)

Call:
lm(formula = moisture ~ filt.rate)

Residuals:
Min       1Q   Median       3Q      Max
-1.39552 -0.27694  0.03548  0.42913  1.09901

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 72.958547   0.697528 104.596  < 2e-16 ***
filt.rate    0.041034   0.004837   8.484 1.05e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.6653 on 18 degrees of freedom
Multiple R-squared:  0.7999,        Adjusted R-squared:  0.7888
F-statistic: 71.97 on 1 and 18 DF,  p-value: 1.052e-07
```

This tells us that

$$t = \frac{\hat{\beta}_1}{\sqrt{MS_{res}}SS_{xx}} = \frac{0.041034}{0.004837} = 8.484.$$

This is not a frequently expected outcome from the $t(18)$ distribution, the sampling distribution of

$$t = \frac{\hat{\beta}_1}{\sqrt{MS_{res}}SS_{xx}},$$

when the null hypothesis is true. This is reflected in a very small p-value,

$$p - \text{value} = 0.000000105.$$

This is strong evidence against H_0. There is strong evidence that the absorption percentage Y is positively linearly related to machine filtration rate x in the population.

```
> b1<-sewage.mod$coefficients[2]
> se.b1<-summary(sewage.mod)$coefficients[2,2]
> t.obs<-b1/se.b1
> n<-length(moisture)
> alpha<-0.05
> ggdat<-data.frame(t=seq(-4.5,4.5,0.01),
+                   f=dt(x=seq(-4.5,4.5,0.01),df=n-2))
> ggdat.highlight<-data.frame(x=c(-t.obs,t.obs),y=c(0,0))
> axis.labels<-round(c(-abs(b1),qt(0.05,df=n-2)*se.b1,
+                                0,
+                                qt(0.95,df=n-2)*se.b1,
+                      abs(b1)),3)
textcolorblue> ggplot(data=ggdat,aes(x=t,y=f))+
+   geom_line()+
+   geom_ribbon(data=subset(ggdat,t>=qt(1-alpha/2,df=n-2)),aes(ymax=f),ymin=0,
+               fill="grey",color=NA)+
+   geom_ribbon(dat=subset(ggdat,t<=qt(alpha/2,df=n-2)),aes(ymax=f),ymin=0,
+               fill="grey",color=NA)+
+   geom_ribbon(data=subset(ggdat,t>=abs(t.obs)),aes(ymax=f),ymin=0,
+               fill="red",color=NA,alpha=0.25)+
+   geom_ribbon(data=subset(ggdat,t<=-abs(t.obs)),aes(ymax=f),ymin=0,
+               fill="red",color=NA,alpha=0.25)+
+   geom_point(data=ggdat.highlight,aes(x=x,y=y),color="red")+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("t")+
+   ylab("Density")+
+   ggtitle("T Test for the Filtration Rate Coefficient",
+           subtitle=bquote(H[0]*":"~beta[1]==0*", versus "*H[a]*":"~beta[1]!=0))+
+   annotate("text", x=4, y=0.05,
+            label= deparse(bquote(alpha==0.05)),parse=TRUE,size=3.5)+
```

```
+   annotate("text", x=8, y=0.05, label="Observation\n P-value<0.0001",size=3.5)+
+   annotate("text", x=8, y=0.05, label="Mirrored\n Observation",size=3.5)+
+   scale_x_continuous(sec.axis = sec_axis(~.,
+        breaks=c(-abs(t.obs),qt(alpha,n-1),0,qt(1-alpha,n-1),abs(t.obs)),
+        labels = axis.labels,name="Coefficient Estimate"))
```

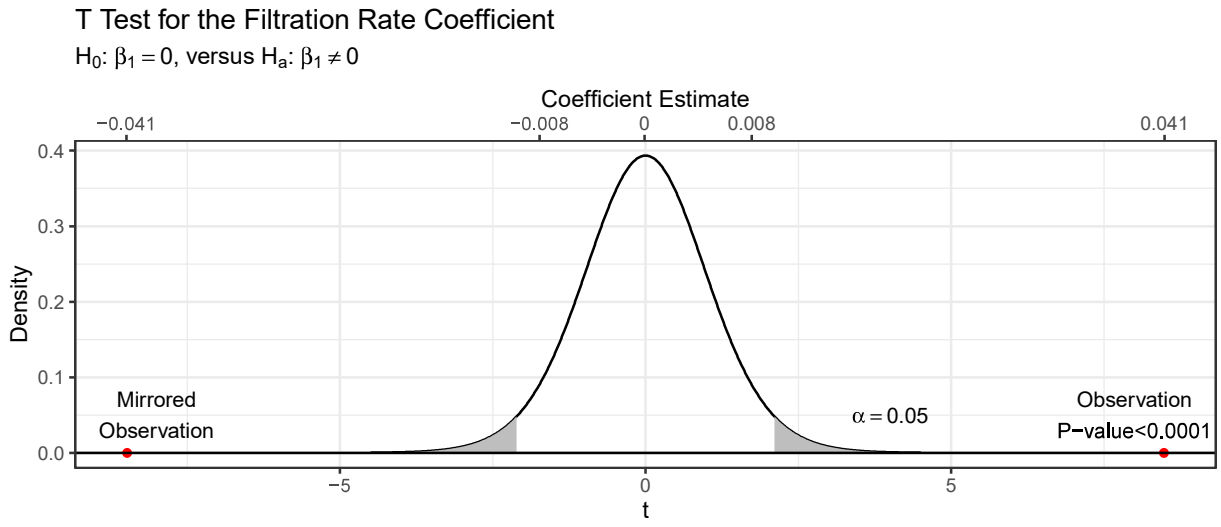

Figure 14.6.4: Sewage data: $t(18)$ PDF. This is the sampling distribution of t when $H_0 : \beta_1 = 0$ is true. A red point at $t = 8.484$ has been added.

14.7 Confidence and Prediction Intervals for a given $x = x_0$

Consider the simple linear regression model

$$Y = \beta_0 + \beta 1_x + \epsilon$$

We are often interested in learning about the response Y at a certain setting of the independent variable, say $x = x_0$. For the sewage data, for example, suppose we are interested in the moisture percentage Y when the filtration rate is $x = 150$ kg-DS/m/hr. Two potential goals arise:

- We might be interested in estimating the population mean of Y when $x = x_0$. This mean response is denoted by $E(Y|x_0)$. **This is the mean of the following probability distribution**:
$$Y|X = x_0 \sim \text{Gaussian}(\mu_{y|x} = \beta 0 + \beta 1 x_0, \sigma^2_{y|x} = \sigma^2)$$
- We might be interested in predicting a new response Y when $x = x_0$. This predicted response is denoted by $Y^*(x_0)$. **This is a new observation from the following probability distribution**:
$$Y|X = x_0 \sim \text{Gaussian}(\mu_{y|x} = \beta 0 + \beta 1 x_0, \sigma^2_{y|x} = \sigma^2).$$

In the first problem, we are estimating the mean of a distribution. In the second problem, we are predicting the value of a new response from this distribution. The second problem is more difficult than the first.

We would like to create $100(1-\alpha)$ percent intervals for the population mean $E(Y|x_0)$ and for the new response $Y^*(x_0)$. The former is called a **confidence interval**, the latter is called a **prediction interval**.

Point Estimator/Predictor: To construct either interval, we start with the same quantity:

$$\hat{Y}(x_0) = b0 + b1x_0,$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares estimates from the fit of the model.

- In the confidence interval for $E(Y|x_0)$, we call $\hat{Y}(x_0)$ a point estimator.
- In the prediction interval for $Y^*(x_0)$, we call $\hat{Y}(x_0)$ a point predictor.

The primary difference in the intervals arises in assessing the variability of $Y^*(x_0)$.

Confidence interval: A $100(1-\alpha)$ percent confidence interval for the population mean $E(Y|x_0)$ is given by

$$\hat{Y}(x_0) \pm t_{n-2,1-\alpha/2}\sqrt{MS_{res}\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{SS_{xx}}\right]}.$$

Prediction interval: A $100(1-\alpha)$ percent prediction interval for the new response $Y^*(x_0)$ is given by

$$\hat{Y}(x_0) \pm t_{n-2,1-\alpha/2}\sqrt{MS_{res}\left[1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{SS_{xx}}\right]}.$$

The two intervals have the same form and are nearly identical. The extra "1" in the prediction interval's standard error arises from the additional uncertainty associated with predicting a new response from the Gaussian($\mu_{y|x} = \beta 0 + \beta 1 x_0, \sigma^2_{y|x} = \sigma^2$) distribution. Therefore, at the same value of x_0, a $100(1-\alpha)$ percent prediction interval for $Y^*(x_0)$ will necessarily be wider than the corresponding $100(1-\alpha)$ percent confidence interval for $E(Y|x_0)$.

The length of both intervals depends on the value of x_0. The standard error in either interval will be smallest when $x_0 = \bar{x}$ and will get larger the farther x_0 is from x in either direction. This implies that the precision with which we estimate $E(Y|x_0)$ or predict $Y^*(x_0)$ decreases the farther we get away from $\bar{x}$. This makes intuitive sense, namely, we would expect to have the most "confidence" in our fitted model near the "center" of the observed data.

It is sometimes desired to estimate $E(Y|x_0)$ or predict $Y^*(x_0)$ for values of x_0 outside the range of x values used in the study. This is called extrapolation and can be very dangerous.

In order for our inferences to be valid, we must believe that the model holds for x values outside the range where we have observed data. In some situations, this may be reasonable. In others, we may have no theoretical basis for making such a claim without data to support it.

In Example 14.1, suppose that we are interested in estimating $E(Y|x_0)$ and predicting a new $Y^*(x_0)$ when the filtration rate is $x_0 = 150$ kgDS/m/hr.

- $E(Y|x_0)$ denotes the population mean moisture percentage when the machine filtration rate is $x_0 = 150$ kg-DS/m/hr.

- $Y^*(x_0)$ denotes the moisture percentage Y for an individual sludge specimen when the filtration rate is $x_0 = 150$ kg-DS/m/hr.

We can ask R to calculate the confidence and prediction intervals directly, as seen below.

```
> newData<-data.frame(filt.rate=150)
> predict(sewage.mod,newdata=newData,level=0.95,interval="confidence")
       fit      lwr      upr
1 79.11361 78.78765 79.43958
> predict(sewage.mod,newdata=newData,level=0.95,interval="prediction")
       fit     lwr      upr
1 79.11361 77.6783 80.54893
```

Note that the point estimate (point prediction) is easily calculated:

$$\hat{Y}(x_0 = 150) = 72.959 + 0.041(150) \approx 79.11361.$$

A 95 percent confidence interval for $E(Y|x_0 = 150)$ is $(78.79, 79.44)$. When the filtration rate is $x_0 = 150$ kg-DS/m/hr, we are 95 percent confident that the population mean moisture percentage is between 78.79 and 79.44 percent.

A 95 percent prediction interval for $Y^*(x_0 = 150)$ is $(77.68, 80.55)$. When the filtration rate is $x_0 = 150$ kg-DS/m/hr, we are 95 percent confident that the moisture percentage for a single specimen will be between 77.68 and 80.55 percent.

Figure **??** shows 95 percent confidence bands for $E(Y|x_0)$ and 95 percent prediction bands for $Y^*(x_0 = 150)$. These are not simultaneous bands (i.e., these are not bands for the entire population regression function). We create these plots by calculating confidence and prediction intervals over a grid. This can be done in R as follows.

```
> ggdat<-data.frame(filt.rate=filt.rate,
+                   moisture=moisture)
> newdata<-data.frame(filt.rate=seq(75,202,0.01))
> pred<-predict(sewage.mod,newdata=newdata,level=0.95,interval="prediction")
> conf<-predict(sewage.mod,newdata=newdata,level=0.95,interval="confidence")
> ggdat.intervals<-data.frame(x=newdata,
+                             pred.lower=pred[,2],
+                             pred.upper=pred[,3],
+                             conf.lower=conf[,2],
+                             conf.upper=conf[,3])
> g1<-ggplot(data=ggdat.intervals, aes(x=x))+
+   geom_ribbon(aes(ymin=conf.lower,ymax=conf.upper),fill="grey",alpha=0.5)+
+   geom_point(data=ggdat,aes(x=filt.rate,y=moisture),shape=1)+
+   geom_smooth(data=ggdat,aes(x=filt.rate,y=moisture),color="black",method="lm",se=FALSE)+
+   theme_bw()+
+   ggtitle("95% Confidence Interval",
+           subtitle=bquote("For inference about"~E(Y|X==x[0])))+
+   xlab("Filtration rate (km/m/hr)")+
+   ylab("Pellet Moisture (%)")
> g2<-ggplot(data=ggdat.intervals, aes(x=x))+
```

```
+   geom_ribbon(aes(ymin=pred.lower,ymax=pred.upper),fill="grey",alpha=0.5)+
+   geom_point(data=ggdat,aes(x=filt.rate,y=moisture),shape=1)+
+   geom_smooth(data=ggdat,aes(x=filt.rate,y=moisture),color="black",method="lm",se=FALSE)+
+   theme_bw()+
+   ggtitle("95% Prediction Interval",
+           subtitle=bquote("For inference about"~Y^"*"*(X==x[0])))+
+   xlab("Filtration rate (km/m/hr)")+
+   ylab("Pellet Moisture (%)")
> grid.arrange(g1,g2)
```

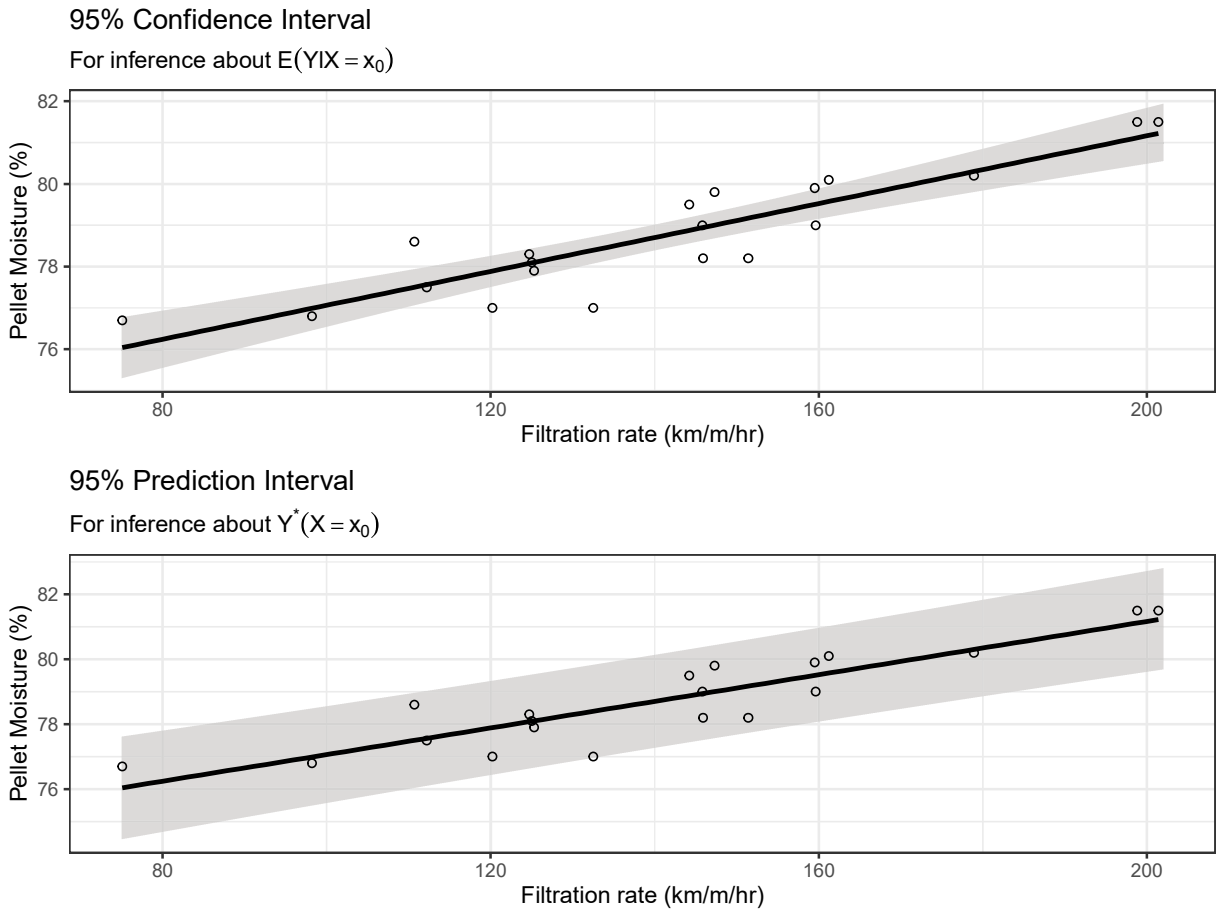

Figure 14.7.5: Scatterplot of pellet moisture Y (measured as a percentage) as a function of machine filtration rate x (measured in kg-DS/m/hr). The estimated linear regression model is superimposed, **Top:** a 95% confidence interval is shaded in grey; **Bottom:** a 95% prediction interval is shaded in grey.

Remark: The confidence interval for the linear regression model is shaded by default when we use supply the `method="lm"` argument in `geom_smooth()`. In the code above, we turned this off by specifying `se=FALSE` to show how we can explicitly plot the confidence band.