Last week we learned

The SLR model

Least squares estimation

Properties of the slope and the intercept

Estimating variance of the random error term

Testing hypotheses for the slope and intercept

The SLR Model

- The (normal) simple linear regression model is

$$Y_i = \mathbb{E}(Y_i|X = x_i) + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n$$

  where Y_i are the response variables, X is an predictor variable, and ε_i are random errors, usually assumed to be independent and normally distributed

$$\varepsilon_i \overset{ind}{\sim} N(0, \sigma^2) \quad \Longrightarrow \quad Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$$

- The LSE of β_0 and β_1 minimising the $SS_E = \sum_{i=1}^n e_i^2$, where $e_i = y_i - \hat{y}_i$, are

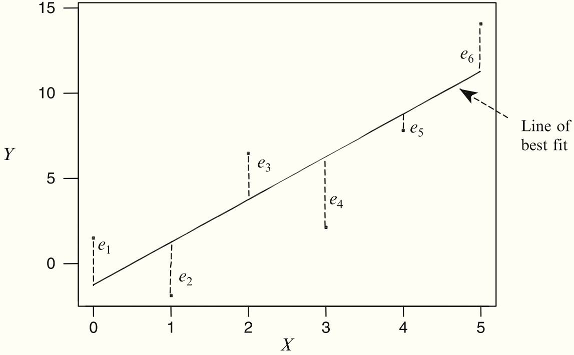
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- $\hat{\beta}_1$ and $\hat{\beta}_0$ are unbiased estimators of β_1 and β_0 distributed normally by

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{s_{xx}}\right) \qquad \hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)\sigma^2\right)$$

The Fitted Model

- The fitted regression model is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ where $\hat{y}_i$ are fitted values.



- The fitted line is also called the mean response line. Points on the line are called mean responses.

Hypothesis Testing

- A test statistic to test the null hypothesis $H_0 : \beta_i = \beta_i^*$ vs $H_1 : \beta_i \neq \beta_i^*$ is

$$T = \frac{\hat{\beta}_i - \beta_i^*}{\mathrm{se}(\hat{\beta}_i)} \sim t_{n-2}$$

  We reject H_0 if $t_{cal} = |t| > t_{\alpha/2,n-2} = t_{crit}$ where t is the sample value of T.

- The estimated standard errors of $\hat{\beta}_1$ and $\hat{\beta}_0$ are

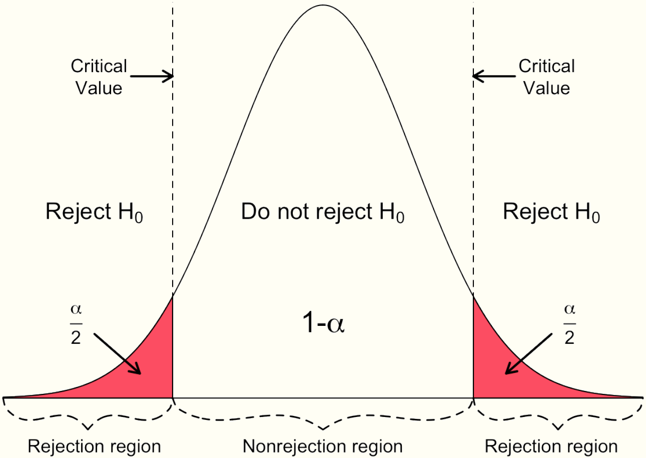
$$\mathrm{se}(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2/s_{xx}} \qquad \mathrm{se}(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\big(1/n + \bar{x}^2/s_{xx}\big)}$$

  where

$$\hat{\sigma}^2 \equiv MS_E = \frac{SS_E}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  is an estimate of σ^2.

Decision Making



Lecture 3

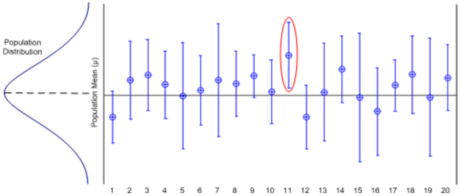
Further inference and significance of regression

Aim: to better understand the SLR model

1. Confidence intervals on the slope and intercept
2. Estimating mean response
3. Prediction of a new observations
4. Testing significance of regression
5. Coefficient of determination R^2

Confidence intervals on the slope and intercept

- In addition to point estimates of β_0 and β_1, we may also estimate confidence intervals (CI's) on these parameters.
- The CI gives the probability that the interval produced by the method employed includes the true value of the parameter.



Loosely speaking, a 95% CI means that there is 95% probability that the true value of the parameter in question is within this CI.

- The width of the CI's of β_0 and β_1 is a measure of the overall quality of the regression line.

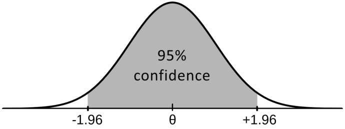
Confidence intervals on the slope and intercept

- To find a $100(1-\alpha)\%$ CI on an unknown parameter θ means to find boundaries a and b such that
$$P(a \leq \theta \leq b) = 1 - \alpha$$

- The boundaries will depend on the data and so using the parameter estimates is a natural way of finding the CI.

- For example, when $a = -1.96$, $b = 1.96$ and θ is an unknown parameter sampled from the standard normal distribution, we have that
$$P(-1.96 \leq \theta \leq 1.96) = 0.95$$

  In other words, 95% of observations of a normal population are within 1.96 standard deviations of the mean.



Confidence intervals on the slope and intercept

- Let us apply $P(a \leq \theta \leq b) = 1 - \alpha$ to β_1:

$$\frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)} \sim t_{n-2} \quad \Longrightarrow \quad P\left(-t_{\alpha/2,\, n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)} \leq t_{\alpha/2,\, n-2}\right) = 1 - \alpha$$

- Solving for β_1 gives

$$P\left(\hat{\beta}_1 - t_{\alpha/2,\, n-2} \cdot \text{se}(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,\, n-2} \cdot \text{se}(\hat{\beta}_1)\right) = 1 - \alpha$$

  This is a probability statement about random variables $\hat{\beta}_1$ and $\hat{\sigma}^2$.

- Upon replacing random variables by their values from the observed data we find a $100(1-\alpha)\%$ confidence interval for β_1 to be

$$\boxed{\text{CI}(\beta_1) = \left[\hat{\beta}_1 - t_{\alpha/2,\, n-2} \cdot \text{se}(\hat{\beta}_1),\ \hat{\beta}_1 + t_{\alpha/2,\, n-2} \cdot \text{se}(\hat{\beta}_1)\right]}$$

  In other words, there is a $100(1-\alpha)\%$ probability that the unknown true value of β_1 is within this CI.

Confidence intervals on the slope and intercept

- We have shown that a $100(1-\alpha)\%$ CI on β_1 is

$$\boxed{\mathrm{CI}(\beta_1) = \left[\hat{\beta}_1 - t_{\alpha/2,\,n-2}\cdot \mathrm{se}(\hat{\beta}_1),\ \hat{\beta}_1 + t_{\alpha/2,\,n-2}\cdot \mathrm{se}(\hat{\beta}_1)\right]}$$

- By the same arguments, a $100(1-\alpha)\%$ CI on β_0 is

$$\boxed{\mathrm{CI}(\beta_0) = \left[\hat{\beta}_0 - t_{\alpha/2,\,n-2}\cdot \mathrm{se}(\hat{\beta}_0),\ \hat{\beta}_0 + t_{\alpha/2,\,n-2}\cdot \mathrm{se}(\hat{\beta}_0)\right]}$$

- **Example.** Consider the manufacturer production data once again. We want to find a 95% CI on β_1. We already know that

$$\hat{\beta}_1 = 0.259, \qquad t_{\alpha/2,\,n-2} = t_{0.025,\,18} = 2.1, \qquad \mathrm{se}(\hat{\beta}_1) = 0.037$$

Therefore

$$\mathrm{CI}(\beta_1) = [0.259 \pm 2.1 \times 0.037] = [0.181,\ 0.337]$$

In other words, there is a 95% probability that $0.181 \le \beta_1 \le 0.337$.

Lecture 3

Further inference and significance of regression

1. Confidence intervals on the slope and intercept
2. Estimating mean response
3. Prediction of a new observations
4. Testing significance of regression
5. Coefficient of determination R^2

- A major use of a regression model is to estimate the mean response μ_0 of Y for a given value of the predictor variable $X = x_0$

$$\mu_0 = \mathbb{E}(Y|X = x_0) = \beta_0 + \beta_1 x_0.$$

- An estimate of this unknown quantity is the value of the estimated regression equation at $X = x_0$,

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

Claim. $\hat{\mu}_0$ is an unbiased estimator of μ_0, that is

$$\hat{\mu}_0 \sim N\left(\mu_0, \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right)\sigma^2\right)$$

Proof. We need to compute $\mathbb{E}(\hat{\mu}_0)$ and $\mathrm{Var}(\hat{\mu}_0)$:

$$\mathbb{E}(\hat{\mu}_0) = \mathbb{E}(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \beta_0 + \beta_1 x_0 = \mu_0$$

To compute $\mathrm{Var}(\hat{\mu}_0)$ we recall that

$$\hat{\beta}_0 = \sum_{i=1}^{n} \left(\frac{1}{n} - c_i \bar{x} \right) Y_i, \qquad \hat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i, \qquad c_i = \frac{x_i - \bar{x}}{s_{xx}}$$

Thus

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_0) &= \mathrm{Var}\left(\sum_{j=1}^{n} \left(\frac{1}{n} - c_j \bar{x} \right) Y_j + \sum_{j=1}^{n} c_j Y_j x_0 \right) \\
&= \mathrm{Var}\left(\sum_{j=1}^{n} \left(\frac{1}{n} + c_j (x_0 - \bar{x}) \right) Y_j \right) \\
&= \sum_{j=1}^{n} \left(\frac{1}{n} + c_j (x_0 - \bar{x}) \right)^2 \mathrm{Var}(Y_j) \\
&= \sum_{j=1}^{n} \left(\frac{1}{n^2} + \frac{2}{n} \cdot \frac{(x_j - \bar{x})(x_0 - \bar{x})}{s_{xx}} + \frac{(x_j - \bar{x})^2}{s_{xx}^2} (x_0 - \bar{x})^2 \right) \sigma^2 \\
&= \left(\frac{1}{n} + 0 + \frac{s_{xx}}{s_{xx}^2} (x_0 - \bar{x})^2 \right) \sigma^2 = \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right) \sigma^2
\end{aligned}$$

Estimating mean response

- We showed that
$$\hat{\mu}_0 \sim N\left(\mu_0, \left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}\right)\sigma^2\right)$$
- We can use this result to test the null hypothesis
$$H_0 : \mu_0 = \mu_0^* \quad \text{vs.} \quad H_1 : \mu_0 \neq \mu_0^*$$
where μ_0^* is some particular constant. The test statistic is
$$T = \frac{\hat{\mu}_0 - \mu_0^*}{\text{se}(\hat{\mu}_0)} \sim t_{n-2}$$
where $se(\hat{\mu}_0) = \sqrt{\left(1/n + (x_0-\bar{x})^2/s_{xx}\right)\hat{\sigma}^2}$
- A $100(1-\alpha)\%$ CI on μ_0 is
$$\boxed{\text{CI}(\mu_0) = \left[\hat{\mu}_0 - t_{\alpha/2,\, n-2} \cdot \text{se}(\hat{\mu}_0),\ \hat{\mu}_0 + t_{\alpha/2,\, n-2} \cdot \text{se}(\hat{\mu}_0)\right]}$$
- Care is needed when estimating the mean response at x_0. It should only be done if x_0 is within the data range. Extrapolation beyond the range of the given x-values is not reliable, as there is no evidence that a linear relationship is appropriate there.

Lecture 3

Further inference and significance of regression

1. Confidence intervals on the slope and intercept
2. Estimating mean response
3. Prediction of a new observations
4. Testing significance of regression
5. Coefficient of determination R^2

Prediction of a new observations

- An important application of the regression model is prediction of a new observation of Y, corresponding to a specified level of the predictor variable X.

- If x_0 is the value of X of interest, then

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

  is the point estimate of the future observation $y_0 = \mathbb{E}(Y|X = x_0)$.

- We want to obtain an interval estimate of this future observation y_0.

- The CI on the mean response μ_0 at $X = x_0$ is inappropriate for this problem because it is an interval estimate on the mean response, not a probability statement about future observations from that distribution.

Prediction of a new observations

- We can write a future observation, viewed as a random variable, as

$$\hat{Y}_0 = \hat{\mu}_0 + \varepsilon_0$$

where $\hat{\mu}_0$ is an estimator of μ_0 at $X = x_0$, and ε_0 is a random error.

- We know that

$$\hat{\mu}_0 \sim N\left(\mu_0, \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right)\sigma^2\right) \qquad \varepsilon_0 \sim N(0, \sigma^2)$$

are independent random variables, implying

$$\hat{Y}_0 = \hat{\mu}_0 + \varepsilon_0 \sim N\left(\mu_0, \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} + 1\right)\sigma^2\right)$$

- Replacing σ^2 by its estimate $\hat{\sigma}^2$ gives

$$\boxed{\frac{\hat{Y}_0 - \mu_0}{\mathrm{se}(\hat{y}_0)} \sim t_{n-2}, \qquad \mathrm{se}(\hat{y}_0) = \sqrt{\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right)\hat{\sigma}^2}}$$

Prediction of a new observations

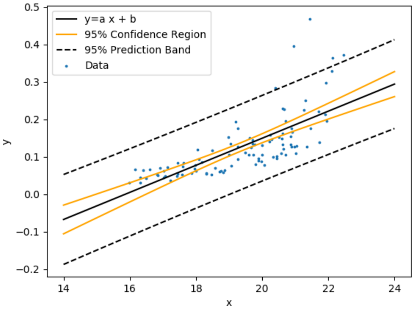
- A $100(1-\alpha)\%$ PI on a new observation y_0 is

$$\boxed{\mathrm{PI}(y_0) = \left[\hat{y}_0 - t_{\alpha/2,\,n-2}\cdot \mathrm{se}(\hat{y}_0),\ \hat{y}_0 + t_{\alpha/2,\,n-2}\cdot \mathrm{se}(\hat{y}_0)\right]}$$

where

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0, \qquad \mathrm{se}(\hat{y}_0) = \sqrt{\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}\right)\hat{\sigma}^2}$$

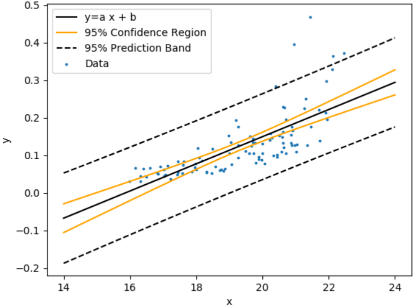
- This interval is usually much wider than the CI for the mean response $\hat{\mu}_0$. This is because of the random error ε_0 reflecting the random source of variability in the data.



Prediction of a new observations

- You should only make predictions for values of x_0 within the range of the data.

- Prediction interval relies strongly on the assumption that the residual errors are normally distributed with a constant variance. So, you should only use such intervals if you believe that the assumption is approximately met for the data at hand.



Lecture 3

Further inference and significance of regression

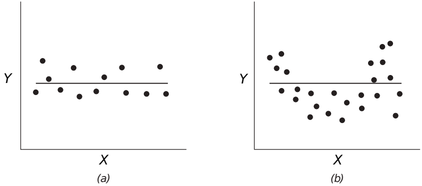
1. Confidence intervals on the slope and intercept
2. Estimating mean response
3. Prediction of a new observations
4. Testing significance of regression
5. Coefficient of determination R^2

- A very important special case of the null hypothesis for β_1 is

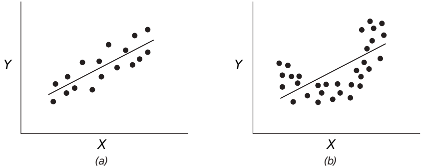
$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$$

This null hypothesis tests the **significance of regression**. Failing to reject H_0 implies that there is no linear relationship between X and Y.

- Situations when H_0 is not rejected, i.e. $\beta_1 = 0$



- Situations when H_0 is rejected, i.e. $\beta_1 \neq 0$



Analysis of variance

- We typically use the **analysis of variance** (ANOVA) approach to test the significance of regression.

- The analysis of variance is based on a partitioning of total variability in the response variable, SS_T, into variability explained by the regression model, SS_R, and the residual, or error, variability, SS_E, that is

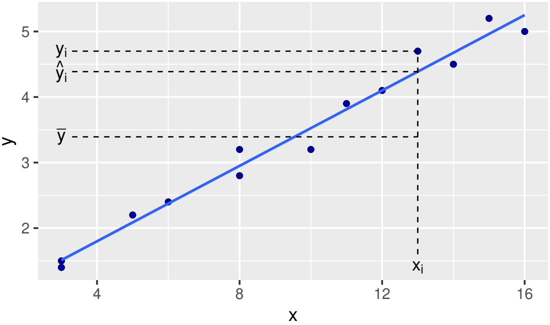
$$SS_T = SS_R + SS_E$$

where

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$



Claim. The Analysis of Variance Identity, $SS_T = SS_R + SS_E$, holds true.

Proof.

$$
\begin{aligned}
SS_T &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}\Big((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})\Big)^2 \\
&= \sum_{i=1}^{n}\Big((y_i - \hat{y}_i)^2 + 2\,(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2\Big) \\
&= SS_E + SS_R + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})
\end{aligned}
$$

where

$$
\begin{aligned}
\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n}(y_i - \hat{y}_i)\hat{y}_i - \bar{y}\sum_{i=1}^{n}(y_i - \hat{y}_i) \\
&= \sum_{i=1}^{n} e_i \hat{y}_i - \bar{y}\sum_{i=1}^{n} e_i \\
&= \sum_{i=1}^{n} e_i(\hat{\beta}_0 + \hat{\beta}_1 x_i) = \hat{\beta}_0 \sum_{i=1}^{n} e_i + \hat{\beta}_1 \sum_{i=1}^{n} e_i x_i = 0
\end{aligned}
$$

since $\sum_{i=1}^{n} e_i = 0$ and $\sum_{i=1}^{n} e_i x_i = 0$.

- The Analysis of Variance identity is used to draw the Analysis of Variance table

Source of variation	d.o.f.	SS	MS	F
Regression	$\nu_R = 1$	SS_R	$MS_R = \frac{SS_R}{\nu_R}$	$F = \frac{MS_R}{MS_E}$
Residual (Error)	$\nu_E = n-2$	SS_E	$MS_E = \frac{SS_E}{\nu_E}$	
Total	$\nu_T = n-1$	SS_T		

- This table shows the sources of variation, the sums of squares, and the statistic, based on the sums of squares, for testing significance of regression slope

$$F = \frac{MS_R}{MS_E} = \frac{SS_R/\nu_R}{SS_E/\nu_E}$$

It measures variation explained by the model relative to variation due to residuals, and can be used to test the hypothesis

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$$

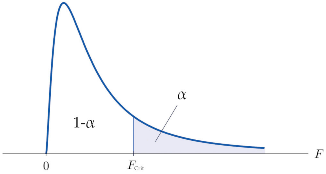
- It can be shown that, assuming the null hypothesis $H_0 : \beta_1 = 0$ is true, then

$$\frac{SS_R}{\sigma^2} \sim \chi^2_1, \qquad \frac{SS_E}{\sigma^2} \sim \chi^2_{n-2}$$

and SS_R and SS_E are independent, and

$$F = \frac{MS_R}{MS_E} \sim F_{1,n-2}$$

- The test procedure computes the value F_{cal} of F for a given data set, and compares with $F_{\alpha,1,n-2}$, the percentile of the $F_{1,\,n-2}$ distribution corresponding to a cumulative probability of $(1-\alpha)$.



- We reject H_0 if $F_{cal} > F_{\alpha,1,n-2}$. Rejecting H_0 means that the slope $\beta_1 \neq 0$ and the full model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ is better than the constant model $Y_i = \beta_0 + \varepsilon_i$.

Example

ANOVA table for the manufacturer production time data ($n = 20$):

Source of variation	d.o.f.	SS	MS	F
Regression	1	12868.37	12868.37	48.72
Residual (Error)	18	4754.58	264.14	
Total	19	17622.95		

Assuming $\alpha = 5\%$, we have $F_{crit} = F_{\alpha,1,18} = 4.41$. Since $F_{cal} = 48.72 > 4.41$, we conclude that regression is significant, i.e. we reject $H_0 : \beta_1 = 0$.

Lecture 3

Further inference and significance of regression

1. Confidence intervals on the slope and intercept
2. Estimating mean response
3. Prediction of a new observations
4. Testing significance of regression
5. Coefficient of determination R^2

Coefficient of determination R^2

- The coefficient of determination, denoted by R^2, is the percentage of total variation in y_i explained by the fitted model, that is

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SS_R}{SS_T} = \frac{SS_T - SS_E}{SS_T} = \left(1 - \frac{SS_E}{SS_T}\right)$$

- For the SLR model R^2 is a square of the Pearson correlation coefficient

$$r^2 = \frac{s_{xy}^2}{s_{xx}s_{yy}}$$

- Note that:
  - $R^2 \in [0,1]$ (or $[0,100]\%$)
  - $R^2 = 0$ (0%) indicates that none of the variability in the data (y) is explained by the regression model.
  - $R^2 = 1$ (or 100%) indicates that $SS_E = 0$ and all observations fall on the fitted line exactly.
- R^2 is a measure of the linear association between Y and X. A small R^2 does not always imply a poor relationship between Y and X, which may, for example, be quadratic.

Summary

- The LSE of the mean response μ_0 at $X = x_0$ is

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \sim N\left(\mu_0, \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right)\sigma^2\right)$$

- A $100(1-\alpha)\%$ CI on μ_0 is

$$\mathrm{CI}(\mu_0) = \left[\hat{\mu}_0 - t_{\alpha/2,\, n-2} \cdot \mathrm{se}(\hat{\mu}_0),\ \hat{\mu}_0 + t_{\alpha/2,\, n-2} \cdot \mathrm{se}(\hat{\mu}_0)\right]$$

- The LSE of a new observation y_0 at $X = x_0$ is

$$\hat{Y}_0 = \hat{\mu}_0 + \varepsilon_0 \sim N\left(\mu_0, \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} + 1\right)\sigma^2\right)$$

- A $100(1-\alpha)\%$ PI on y_0 is

$$\mathrm{PI}(y_0) = \left[\hat{y}_0 - t_{\alpha/2,\, n-2} \cdot \mathrm{se}(\hat{y}_0),\ \hat{y}_0 + t_{\alpha/2,\, n-2} \cdot \mathrm{se}(\hat{y}_0)\right]$$

Summary

- Analysis of Variance Identity

$$SS_T = SS_R + SS_E$$

where

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

– SS_T measures the total variation in y around its mean $\bar{y}$.

– SS_R measures the total variation in $\hat{y}$ around the mean $\bar{y}$.

– SS_E measures how closely model fits the data.

- Statistic for testing significance of regression, $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$,

$$F = \frac{MS_R}{MS_E} = \frac{SS_R / \nu_r}{SS_E / \nu_E} \sim F_{1,n-2}$$

- Coefficient of determination

$$R^2 = 1 - \frac{SS_E}{SS_T} \in [0, 1]$$

Next week

Diagnostics and transformations