7.3 Regression with Both Categorical and Numerical Predictors

7.3.1 Different Lines for Each Group

7.3.2 A Simpler Model: Parallel Lines

7.3.3 Analysis of Covariance

8 Correlation
7.3 Regression with Both Categorical and Numerical Predictors

It is often the case that a response may depend on both categorical and numerical predictors. This is again easily handled by appropriate use of dummy variables.

**Example.** The file `turkeys.txt` (from Draper and Smith, 1998, *Applied Regression Analysis*, 3rd ed.) contains data on the weight, age, and state of origin of thirteen Thanksgiving turkeys (four from Georgia, four from Virginia, and five from Wisconsin). Treating weight as the response we have a numerical predictor (age) and a categorical predictor (origin).

We might want to model weight as function of age, but what if the same straight line model is not appropriate for turkeys from all three states?

What if we want to compare the three states to see which produces the heaviest turkeys (or better yet, imagine that instead of coming from three different states, the turkeys were fed three different diets)? How can we account for the fact that older turkeys may weigh more in making this comparison?

### 7.3.1 Different Lines for Each Group

In the situation of the last example, we have a numerical predictor and a categorical predictor with three levels (Georgia, Virginia, and Wisconsin). One possibility is that the dependence of weight on age is best described by a separate simple linear regression model for each state, each having its own slope and intercept.

Let $Y$ represent weight, $x$ age, and define the following dummy variables for origin, where just as before we take one level of the factor (Georgia) as a baseline.

\[
d_1 = \begin{cases} 
1, & \text{if turkey is from Virginia} \\
0, & \text{otherwise}
\end{cases}
\]

and

\[
d_2 = \begin{cases} 
1, & \text{if turkey is from Wisconsin} \\
0, & \text{otherwise}
\end{cases}
\]

Then a model allowing different slopes and intercepts for each origin is

\[
Y = \beta_0 + \beta_1 x + \beta_2 d_1 + \beta_3 d_2 + \beta_4 d_1 x + \beta_5 d_2 x + \epsilon
\]

Note that
Fitting this model to the turkey data yields the following estimates:

\[
\hat{\beta}_0 = -0.979 \quad \hat{\beta}_1 = 0.506 \quad \hat{\beta}_2 = 0.679 \quad \hat{\beta}_3 = 3.450 \quad \hat{\beta}_4 = -0.036 \quad \hat{\beta}_5 = -0.061
\]

Thus the three estimated regression lines

<table>
<thead>
<tr>
<th>Origin</th>
<th>( d_1 )</th>
<th>( d_2 )</th>
<th>( E(Y) )</th>
<th>Intercept</th>
<th>Slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>( \hat{\beta}_0 + \hat{\beta}_1 x )</td>
<td>( \hat{\beta}_0 )</td>
<td>( \hat{\beta}_1 )</td>
</tr>
<tr>
<td>V</td>
<td>1</td>
<td>0</td>
<td>( (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_4) x )</td>
<td>( \hat{\beta}_0 + \hat{\beta}_2 )</td>
<td>( \hat{\beta}_1 + \hat{\beta}_4 )</td>
</tr>
<tr>
<td>W</td>
<td>0</td>
<td>1</td>
<td>( (\hat{\beta}_0 + \hat{\beta}_3) + (\hat{\beta}_1 + \hat{\beta}_5) x )</td>
<td>( \hat{\beta}_0 + \hat{\beta}_3 )</td>
<td>( \hat{\beta}_1 + \hat{\beta}_5 )</td>
</tr>
</tbody>
</table>

See Figure 1.

The residual sum of squares for the fitted model is \( \text{SSE} = 0.7062 \).

**Question:** what is the degrees of freedom associated with this residual sum of squares?

A comparison of the three states on the weights of their turkeys is complicated if the slopes are truly different for the three states. (Again, this may be more meaningful if we imagine that we are comparing different turkey feeds rather than states.) For example, the fitted regression lines above for Georgia and Virginia cross at \( x = 18.84 \). Thus we estimate that

- Virginia turkeys less than 18.84 months old are heavier on the average than Georgia turkeys of the same age, whereas

- Virginia turkeys over 18.84 months old are lighter on the average than Georgia turkeys of the same age.

However, Figure 1 suggests that it may be plausible to assume that the three regression lines actually have the same slope.

Note also that it would not be wise to ignore the effect of age on weight in comparing the three states, since it seems clear from Figure 1 that age explains a great deal of the variation in weights. If we ignore age, then the additional variation in weight may mean a loss in precision for our comparison of the three states (see also Figure 2).
Figure 1: Regression lines for turkey weights, with a different slope and intercept for each state of origin.
Figure 2: Turkey weights grouped by state of origin, ignoring age.
In fact, a simple one-way analysis of variance comparing the weights of turkeys from the three states, ignoring age, yields a $P$-value of 0.41, i.e., no evidence of a difference between the three states in average turkey weights.

Question: what are the numerator and denominator degrees of freedom of the $F$ statistic for this test?

Note by the way the the one-way analysis of variance model can be written

$$Y = \beta_0 + \beta_2 d_1 + \beta_3 d_2 + \epsilon$$

i.e., it is a special case of the previous model with $\beta_1 = \beta_4 = \beta_5 = 0$. There is strong evidence that this is not an appropriate model, as the $P$-value for the $F$ test comparing this to the full model is practically zero ($P$-value $= 3.3 \times 10^{-6}$).

Question: The residual sum of squares for this model is $SSE = 33.035$. What is the value of the $F$-statistic for the test just carried out and what are its numerator and denominator degrees of freedom?

### 7.3.2 A Simpler Model: Parallel Lines

A much simpler model results if we can assume that the regression lines for the three states have the same slope, i.e., $\beta_4 = \beta_5 = 0$. This assumption corresponds to the reduced model

$$Y = \beta_0 + \beta_1 x + \beta_2 d_1 + \beta_3 d_2 + \epsilon$$

As usual, we can test whether this simpler model is plausible via an $F$ test of

$$H_0 : \beta_4 = \beta_5 = 0 \quad \text{versus} \quad H_a : \text{at least one of } \beta_4 \text{ and } \beta_5 \text{ is not zero.}$$

For the turkey data, the $P$-value for this test is .6156, so that there is no evidence against the simpler model.

Question: The residual sum of squares for the reduced model here is $SSE = 0.8112$. What is the value of the $F$-statistic for the test just carried out and what are its numerator and denominator degrees of freedom?
The estimated coefficients for the model with equal slopes are

\[ \hat{\beta}_0 = -0.488 \quad \hat{\beta}_1 = 0.487 \quad \hat{\beta}_2 = -0.274 \quad \hat{\beta}_3 = 1.920 \]

Thus we estimate that for any fixed age, the average weight of Georgia turkeys is 0.274 pounds more than that of Virginia turkeys, and 1.92 pounds less than that of Wisconsin turkeys.

The three estimated regression lines are (see Figure 3)

<table>
<thead>
<tr>
<th>Origin</th>
<th>( \hat{y} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>(-0.488 + 0.487x)</td>
</tr>
<tr>
<td>V</td>
<td>(-0.761 + 0.487x)</td>
</tr>
<tr>
<td>W</td>
<td>(1.157 + 0.487x)</td>
</tr>
</tbody>
</table>

Figure 3: Parallel regression lines for turkey weights.
7.3.3 Analysis of Covariance

In a situation where the assumption of parallel lines is plausible, we can test for no difference between groups by testing whether all the intercepts are the same. Assuming the parallel lines model,

\[ Y = \beta_0 + \beta_1 x + \beta_2 d_1 + \beta_3 d_2 + \varepsilon, \]

we wish to test

\[ H_0 : \beta_2 = \beta_3 = 0 \quad \text{versus} \quad H_a : \text{at least one of } \beta_2 \text{ and } \beta_3 \text{ is not zero}. \]

Here we generally for the \( F \) statistic taking the parallel lines model as the full model and the null model \( Y = \beta_0 \beta_1 x + \varepsilon \) as the reduced model.

For the turkey data, the \( P \)-value for this test is practically zero (\( P \)-value = 3.5 × \( 10^{-6} \)). Thus, controlling for age, there is strong evidence of a difference in mean turkey weights between the three states.

8 Correlation

The most commonly used measure of the association between two numerical variables is the correlation coefficient. The correlation is simply the covariance of the two variables divided by both their standard deviations.

- For numerical variables \( X \) and \( Y \), the population version of the correlation coefficient is

\[ \rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \]

- The sample version is

\[ r = \frac{\text{Cov}(X, Y)}{s_X s_Y} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{s_X s_Y}, \]

also called Pearson's product moment correlation.

The correlation coefficient has the following properties:
-1 \leq r \leq 1.

- \textbf{\( r = 0 \)} indicates no linear association.

- \textbf{\( r < 0 \)} indicates negative (linear) association, i.e., large values of \( X \) are associated with small values of \( Y \) (and small values of \( X \) with large values of \( Y \)).

- \textbf{\( r > 0 \)} indicates positive (linear) association, i.e., large values of \( X \) are associated with large values of \( Y \) (and small values of \( X \) with small values of \( Y \)).

- The closer \(|r|\) is to one, the stronger the linear association. \(|r| = 1\) indicates perfect linear association, i.e., all the points lie on a straight line.

- \textbf{\( r^2 \)} gives the proportion of the variation in \( y \) that is explained by straight-line dependence on \( x \) (and vice-versa).

\textbf{Note:} \( r \) may fail to describe a \textit{nonlinear} association well, no matter how strong.